Model Thinking

Course Notes by Rainer Groh

Email: rainer.groh@bristol.ac.uk

Contents

Abstract

The following document is a collection of lecture notes that I took while taking the course "Model Thinking" by Scott E. Page, Professor of Complex Systems, Political Science, and Economics at the University of Michigan, Ann Arbor. Any errors and omissions in these notes are naturally entirely my own, and all credit for the content goes to Scott. Needless to say, these lecture notes are by no means a substitute for his awesome course. Rather, consider this document as a primer to thinking with qualitative and quantitative models; either to think more clearly, make better decisions or be more informed about the world around you. If you enjoy these lecture notes, then I highly recommend that you try out Scott's course. It's free and can be found at Coursera.

1 Introduction

We live in a complex world with diverse people, firms, and governments whose behaviours aggregate to produce novel, unexpected phenomena. We see political uprisings, market crashes, and a never-ending array of social trends. How do we make sense of it all? The short answer is: models.

Evidence shows that people who think with models consistently outperform those who don't. And moreover, people who think with lots of models outperform people who use only one.

Why do models make us better thinkers? Models help us to better organize information—to make sense of that firehose or hairball of data (choose your metaphor) bombarding us on the Internet. Models improve our ability to make accurate forecasts. They help us make better decisions and adopt more effective strategies. They can even improve our ability to design institutions.

These lecture notes are a starter toolkit of models. We start with models about tipping points; move on to cover models that explain the wisdom of crowds; models that show why some countries are rich and some are poor; and finally, models that help us unpack the strategic decisions of firms and politicians. The models covered provide a foundation for future social studies, but most imporantly, give you a huge leg up in life.

We'll start with a gentle and general introduction and then get into the thick of things.

1.1 Why Model?

There are various reasons why we want to create models about the world. Here are four:

1. To be an intelligent citizen of the world—a multi-disciplinary liberal arts thinker.

2. To think clearer—models are a crutch to weed out logical inconsistencies.

3. To use and understand data. There is now a firehose of data online that we should attempt to turn into useful knowledge. Models allow us to structure data to tease out new knowledge and perhaps arrive at some wisdom in the long run.

4. To decide, strategise and design. Models allow us to structure data, which then allows us to make better decisions.

Model thinking is eclectic, and insights can come from a lot of different fields. Often the really big insights come from combining ideas from different areas. Ideally you want to choose the most powerful ideas/models from the big disciplines of human thought: physics, biology, chemistry, mathematics, economics, history, psychology, sociology, political science, law, *etc.*

What is needed to teach and understand a model?

1. What is the model? What are the underlying assumptions? How does the model work? What are the results/outcomes it predicts? What are possible applications and implications?
2. Technical details—understanding the math involved and practising with example problems.
3. Fertility—most models are developed for one purpose, but where else does this model work? Can I cross-pollinate with other models/ideas?

1.2 Intelligent Citizen of the World

Models make you a more intelligent citizen in this complex world. All models are simplifications and based on abstractions. This implies that all models are somewhat wrong, but some are certainly useful. Always remember: models are the map of the territory, not the territory itself. Because of this, models should always be validated against experiments/reality. The good models then endure and can help us do/design things in better ways. In fact, models are the new lingua franca of business, politics, the academy, *etc.* They enable people to do a better job at their chosen vocation.

In a way, models are lenses through which we can observe the world in a specific way. This is why a multidisciplinary view is important. Life is full of opposites and seeming contradictions, *e.g.* two heads are better than one (get a second opinion), yet too many chefs ruin the broth (too much input creates confusion).

So...

- What you need is lots of accurate models. The "lots" aspect provides discrimination between opposing drivers, and the "accuracy" allows for good prediction. The models that do best are formal models, the ones based on the big ideas of science, *e.g.* biology, chemistry, physics, maths. If models are tools for thinking, then we need lots of tools in our toolbox.
- Models should ideally influence you how to think BUT not tell you what to do. This is because the map is not necessarily the territory.
- Models make us humble—they make us see the full complexity and multi-dimensionality of a problem, and make us realise how much we have to leave out to have a model that is useful, *i.e.* a model that we can actually use to make predictions. An all-encompassing model would be so complicated as to be impossible to use.

1.3 How Models Make Us Clear Thinkers

When thinking about a problem, a lot of clarity can simply arise by going through the process of writing a model. This forces us to think about the underlying assumptions, what factors need to be accounted for to model the process under consideration, what the likely outcomes could be, and what the model should certainly not predict (the range of outcomes).

This is how we would go about writing down a model:

1. Name the parts. Example: choosing a restaurant. What matters? Names of restaurants, individual people, how much money is available, time, preferences. What doesn't matter? Clothes I am wearing.
2. Determine the relationship between the parts and how these play out, *i.e.* work through the logic. When doing this, you often figure that the logic plays out very different from what intuition would suggest.

3. Explore inductively by running lots of parametric studies to study the possible range of outcomes. Once you have the model you can change/add/subtract features and explore their effect.

4. Explore the possible types of outcome: equilibrium, cycle, random, complex. For example, the demand of oil probably depends on the size of an economy, and since that grows relatively smoothly around 2%/year, the demand for oil should be quite predictably sloping up. But the price of oil also depends on people's subjective judgements, human psychology, over- and underreactions, and should thus be more complex than a nice linear curve.

5. Identify logical boundaries. There are opposite proverbs for almost anything. There are conditions for which either one of two opposite scenarios holds true, and models allow us to determine when a certain scenario is more likely to be valid.

6. Communicate with others. We can break down the complexities of a process into its relevant dependent parts, and by exploring their interactions, communicate quite clearly what we think will happen. For example, the process of voting in a democracy could be broken down into voters and candidates, and each voter's preference is likely to arise as a result of a candidate's likeability and his/her policy. So depending on how these two factors overlap we can communicate more clearly why we believe someone will vote in a specific way.

1.4 Using and Understanding Data with Models

Models can also help us to unpack, use and understand data in better ways:

1. Find basic patterns in the data, *e.g.* constant, linear or cyclic relationships.

2. Prediction. Either based on deductive or on inductive reasoning, *i.e.* deriving fundamental models or statistical regressions based on data.

3. Produce bounds on likely outcomes. In the long term we can't predict for sure, but perhaps we can provide realistic bounds.

4. Rertrodiction. We might not have collected enough data in the past, so we use a model based on current data to predict what might have happened in the past.

5. Predict other things. Say you have an unemployment model, and if this model gives you a reasonable inflation rate as well, then this gives you extra confidence in the model. The best models can predict something unexpected beyond their initial specification.

6. Inform data collection. In the previous section we said models require naming of the relevant parts. So if you know the relevant parts that inform the model, then you know what data to collect.

7. Estimate hidden parameters. For example, if people are getting sick but you don't know how the disease spreads or how virulent it is, but you have a graph of how many people are getting the disease over time. By fitting this data to a model, you can now estimate what the virulent parameter may be.

8. Calibration. Get tons and tons of data to calibrate a model. Basically one step away from machine learning.

1.5 Using Models to Decide, Strategize and Design

How do models help us to make better decisions?

1. Good real time decision aids. When we have complexity, as we often do in life, with lots of interconnected parts (*e.g.* banks) we can use a model to decide how to act (*e.g.* bail certain banks out or not), because we can quantify their interdependency to some extent (*e.g.* Wells Fargo and AIG are systemically more important than Lehman. If AIG fails, then the whole system falls apart).

2. Comparing different options.

3. Counterfactuals. We can only run the world once, but with models we can at least run abstractions of the world more than once. But remember, these are just models (always wrong, sometimes useful).

4. Identify and rank levers. Basically a parametric study where we investigate the effect of one parameter. For example, we wouldn't want Germany to fail because it would greatly effect other countries. The point is to find the most important parameter, *i.e.* the greatest lever.

5. Experimental design. How to design experiments by revealing things about the underlying mechanism. Essentially, help us to design experiments to make better policies.

6. Institutional design. Let's say we have θ which is the environment, a set of technologies, or people's preferences. Then we have $f(\theta)$ which is the outcomes we want to achieve given our technologies/environment/*etc.* But we don't get the outcome straight away, because we need to engineer a mechanisms to get there, and these mechanism (*e.g.* markets) aren't perfect. So we need to design these mechanisms well (markets don't always work, *i.e.* we need specific constraints). Models will inform on how to design these mechanisms. Should we have a market here, a democracy, *etc.*?

7. How to choose specific policies *e.g.* choosing a CO_2 cap-and-trade policy.

2 Segregation and Peer Effects

Consider this widespread empirical phenomenon: groups of people who hang out together tend to look alike, speak alike, think alike, *etc.* This tendency of individuals to associate with similar others is known as *Homophily.*

But why is this the case? Why is there segregation in Detroit? Or why do smokers tend to stick together?

Let's create a model of how these segregation patterns develop using Schelling, Granovetter and Standing Ovation models. These models are called agent-based models and work as follows:

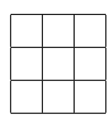
1. Take a bunch of people, companies, *etc.*
2. Map their ways of behaviour and rules that they follow.
3. Once you have agents that are following specific rules, you get a macro level outcome. What is this outcome? We may logically think that rational people behave in certain ways, and a model will surprise us when the effects are in fact opposite.

2.1 Schelling's Segregation Model

Schelling set out to study the following empirical phenomenon: There is widespread racial and income segregation in most parts of the developed world. Why do Whites/Blacks/Latinos/Asians segregate into communities in large cities? Why do the rich and poor segregate by income?

Of course, one explanation is that people are fundamentally racist, and there is no need for a model. To test alternative hypotheses, we can create an agent-based model: people $\rightarrow$ behaviours $\rightarrow$ aggregate all the behaviours $\rightarrow$ look at the outcomes.

Imagine a 3×3 checkerboard as shown below. Some people (represented by squares) are rich and others are poor. Schelling's introductory question now is: given that a certain proportion of people are like me, should I stay here or should I move? Say if 3 out of 8 of my neighbours are like me, do I stay or should I move? So depending on the ratio of people like me, I decide to move or stay, and everyone uses this same heuristic.



Microincentives don't equal macrooutcomes

So let's take a grid of $51 \times 51 = 2601$ squares where 77% of the squares, *i.e.* ≈ 2000 squares in total, are populated by one of two different groups (red or green), and the rest of the squares are empty. Initially the squares are randomly filled by either red or green, so that the percentage of like-coloured squares around each individual square is 50% (see Figure 1 for a version of this model in the free software NetLogo).

We now set the similarity-wanted percentage at 30%, meaning that each square will look at the 8 squares around it, and if at least 30% of the surrounding squares are of the same colour it will remain in place, and otherwise move to an adjacent unpopulated black square. So with a similarity-wanted percentage of 30%, each person in one of the checkerboard squares will stay put if 30% of the squares around them are populated by similar people, and will leave if the percentage

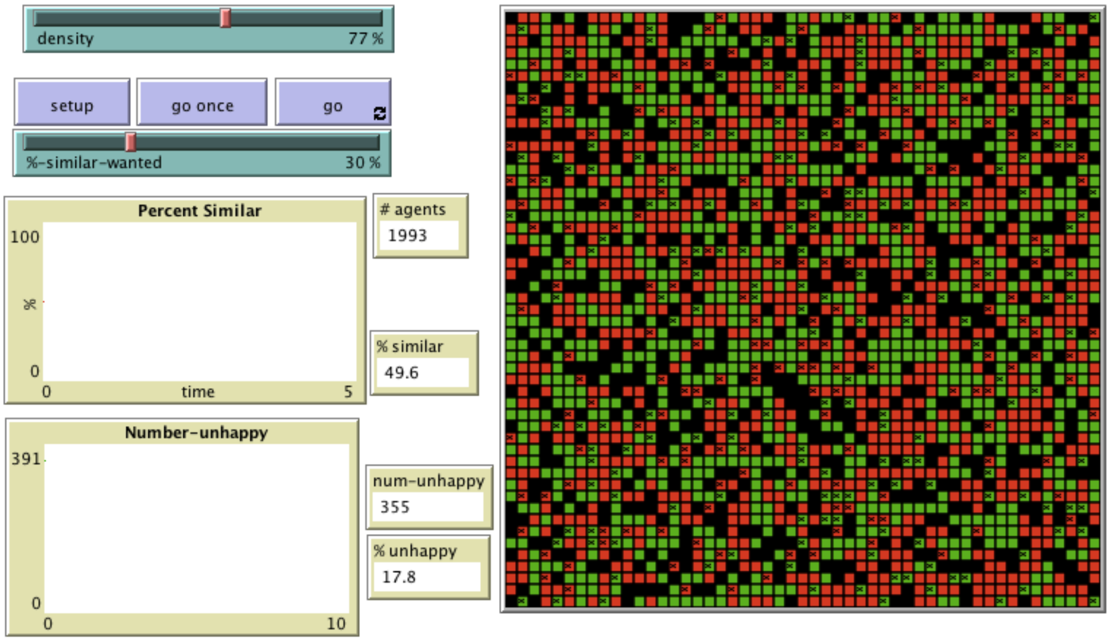


Figure 1: Initial segregation model population in NetLogo. The red and green squares depict two different groups and the black squares are empty spaces where anyone can move to.

falls below this. In the beginning, anyone will have around 50% similar neighbours out of the 8 neighbours directly surrounding, just because people are randomly distributed.

The second metric is how many percent of the total population are unhappy (the percentage of the population for which the similarity percentage is not met). As shown in Figure 1, currently 17.8% of the entire population is unhappy, *i.e.* the 30% similarity is not met. If we now start the simulation, the individual squares will move around the checkerboard until an equilibrium state is reached. Of course, the program wants to minimise this percentage of unhappy people and this drives people around the checkerboard.

If you run this experiment on a computer with a 30% similarity-wanted percentage, people shuffle around the checkerboard until the system goes to a 0% unhappy equilibrium state, at which point the average similarity of neighbours is actually around 70%. The counterintutive finding here is that individuals only wanted 30% similarity around them, but in the aggregate we got 70% similarity. So in a rather tolerant group that only wants a third of the population to be like them, if you enforce that rule and reach a state where everyone is happy such that there is no further driver for more re-shuffling, you actually get 70% similarity (see Figure 2).

This means that the macro and the micro are not the same! Lots of people enforcing 30% at the micro doesn't lead to 30% at the macro. Even worse, if you start with a 40% similarity-wanted percentage you get 80% similarity overall in the aggregate. The most counter-intuitive finding is if we start with a random distribution (on average $\approx 50\%$ of cells around each cell are the same) and we enforce a 50% similarity-wanted percentage. Even though on average (in the aggregate) this

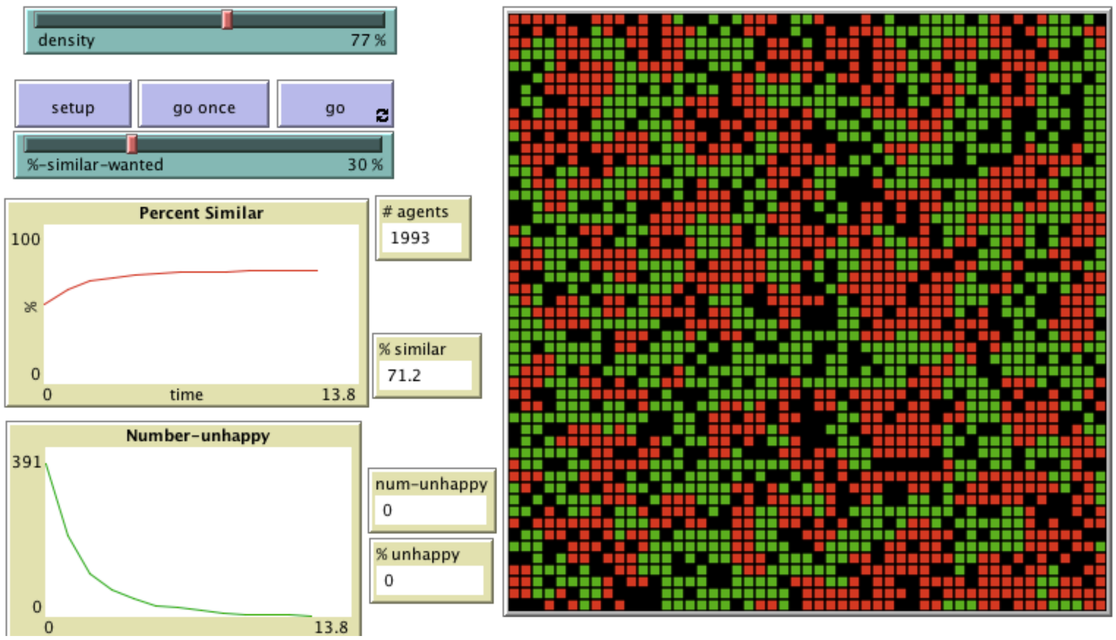


Figure 2: A similarity-wanted percentage of 30% for each individual in the checkerboard actually leads to 71.2% similarity on average in the aggregate. This is because to get every single cell to achieve 30% similarity some will need to have considerably more similarity—indeed, some cells have 100% similarity around them.

random distribution has a similarity of 50%, this is not true for each individual cell, meaning that 42% are still initially unhappy (see Figure 3).

This 42% of unhappy cells will now drive the re-shuffling process until everyone is happy. At this equilibrium state we get a whopping 88% similarity in the aggregate. Figure 4 clearly shows the extent of this segregation with lots of different islands of similar cells. If we think about it, 50% isn't even that intolerant—in this case of two different options, it's the definition of equality. Yet we have learned that trying to enforce this 50% similarity for every single cell is actually very hard to do, and the only way to achieve this is to have some cells with more than 50% similarity, and this aggregate network effect creates an insane amount of segregation.

If you crank the similarity-wanted percentage up to 80%, then you can't even get to an equilibrium because people are always just moving around avoiding to be around anyone unlike them. So the notion that people are insanely racist, and this is why people segregate, doesn't bear out in Schelling's model. The overarching moral of the story is that microincentives don't equal macrooutcomes.

It does show, however, that the slightest preference one way can lead to dramatically pronounced outcomes. Say we increase the similarity-wanted percentage to 66%—for every oppositely coloured cell there has to be two cells of the same colour—then the average similarity on the whole jumps up to 98%! That is pretty much the definition of perfect segregation (see Figure 5).

Tipping points

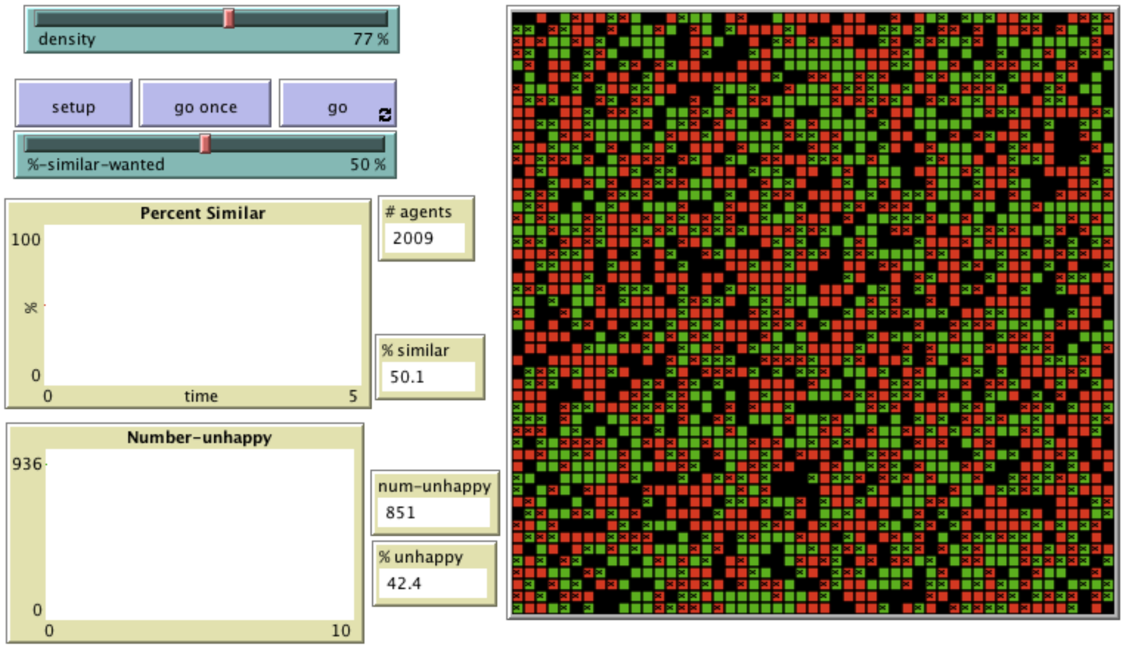


Figure 3: Initial segregation model population in NetLogo.

The segregation predicted by Schelling's model is due to complex interactions in networks. However, two simple mental constructs can help us make sense of the mechanisms at play. Both these mental models fall in the category of tipping points.

1. An exodus tip occurs when one person leaving causes the similarity percentage of another person to drop below a threshold such that this person leaves as well. So someone moving out, leads to another person moving out.

2. A genesis tip is occurs when someone moves in and lowers the similarity-wanted percentage such as to cause someone of the opposite characteristic to move out. So one person moves in and this leads to someone else moving out.

2.2 Measuring Segregation

The counterintuitive finding of the Schelling model is that even if people are relatively tolerant, they still end up being segregated. How do we actually measure real-world segregation so that we can collect data to inform our models?

One option is the index of dissimilarity. First, you sum the total number of people in each individual group, *e.g.* poor and rich people. For example, if we have $B = 150$ rich people and $Y = 90$ poor people throughout a city, we now want to compare the global ratio of $90/150 = 3/5$ across the entire city to similar ratios for suburbs. For example, if we have $b = 5$ rich and $y = 3$

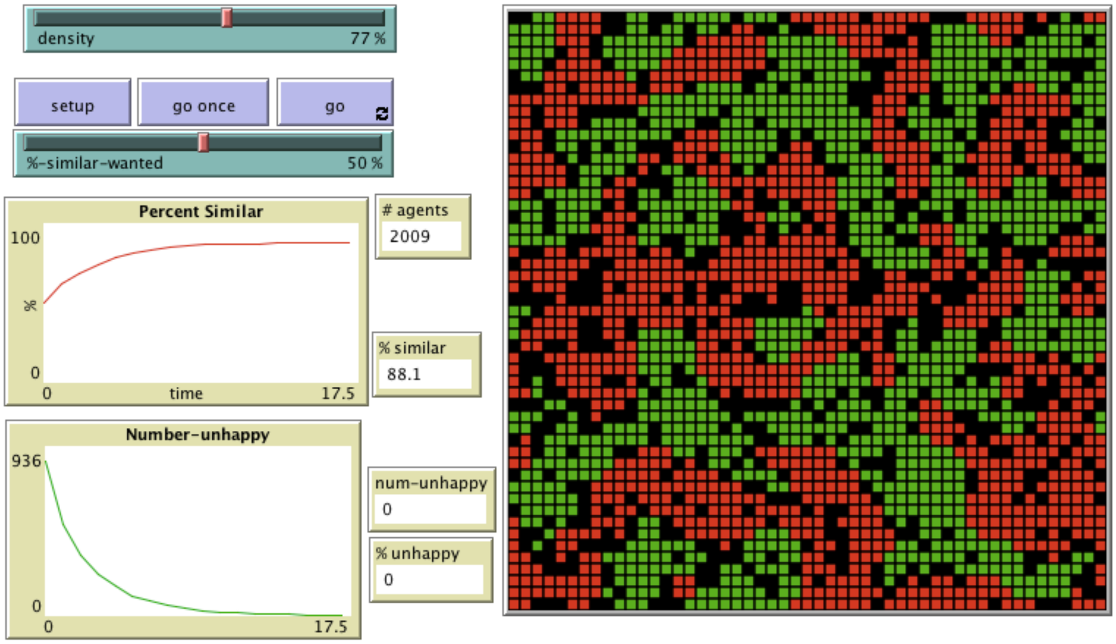


Figure 4: A similarity-wanted percentage of 50% for each individual in the checkerboard actually leads to 88.1% similarity on average in the aggregate.

poor living in a specific district of the city we would calculate,

$$\left|\frac{b}{B} - \frac{y}{Y}\right| = \left|\frac{5}{150} - \frac{3}{90}\right| = 0. \tag{1}$$

So because this block represents the 3/5 split of poor to rich throughout the entire city, there is no segregation in this district. Alternatively, if we have a district with 10 rich people and no poor people we have

$$\left|\frac{10}{150} - \frac{0}{90}\right| = 1/15, \tag{2}$$

and the other way around

$$\left|\frac{0}{150} - \frac{10}{90}\right| = 1/9, \tag{3}$$

and if we have a 50/50 split

$$\left|\frac{5}{150} - \frac{5}{90}\right| = 1/45. \tag{4}$$

So let's say that over the entire city we have 6 blocks of 50/50, 12 blocks of 100/0, and 6 blocks of 0/100, with each block having 10 people, then

$$6\frac{1}{45} + 6\frac{1}{9} + 12\frac{1}{15} = \frac{72}{45}. \tag{5}$$

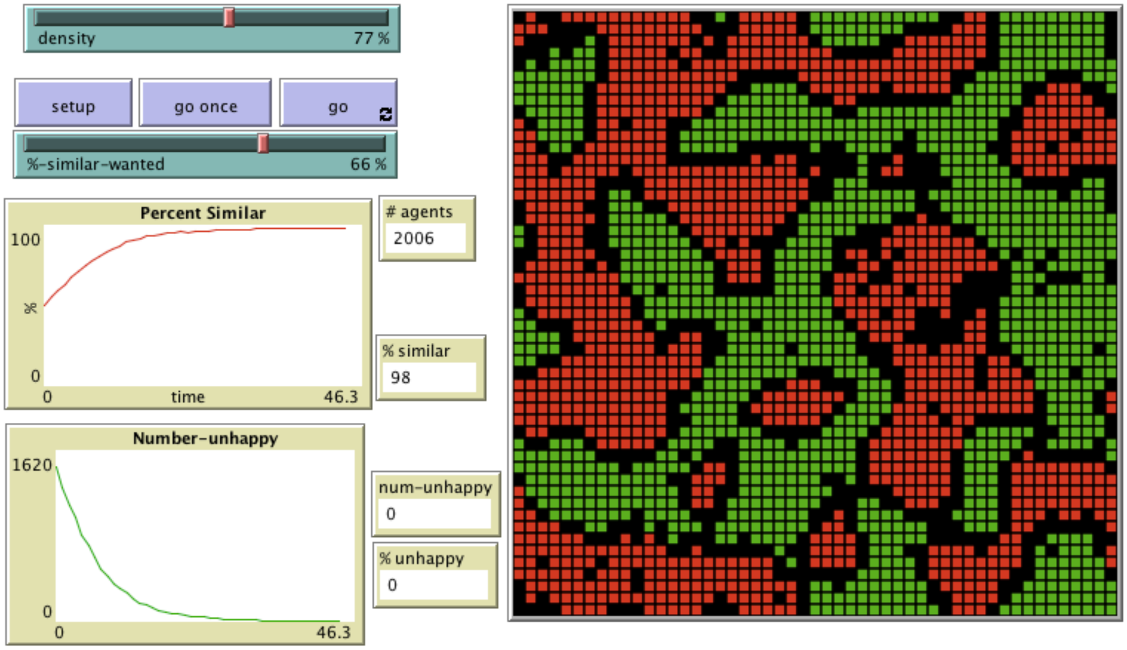


Figure 5: A similarity-wanted percentage of 66% for each individual in the checkerboard actually leads to 98% similarity on average in the aggregate.

Now the question becomes: is this bad or good? Once you have a measure, you want to go to the extremes in order to put some bounds on your measurements, and then compare where you lie within those bounds. So if we had 8 districts of 50/50, with each block having 10 people, we would have $4 \times 10 = 40$ rich and 40 poor. Each block has 5/40 for rich and poor such that

$$\left| \frac{5}{40} - \frac{5}{40} \right| = 0. \tag{6}$$

So yes, obviously if we have a perfect 50/50 mix everywhere then there can be no segregation. Let's now say 4 blocks are all poor and 4 blocks are all rich. Again you have 40 rich and 40 poor in total. But the individual absolute values for each block now are

$$\left| \frac{0}{40} - \frac{10}{40} \right| = 1/4. \tag{7}$$

So now if we add all of them up for each block $4 \times 1/4 + 4 \times 1/4 = 2$. This means we have 2 for perfect segregation, and 0 for none. And by dividing by 2 we can recalibrate from the range of 0 to 2 to the range of 0 to 1. Thus, $72/45/2 = 72/90 = 80\%$ segregated, which is quite segregated.

So we have learned that we can construct a very simple measure to compare how segregated an area is by race, income, *etc.*

2.3 Granovetter's Model

On many occasions it is the tail of the distribution, the extremes, that drive what happens. Phenomena governed by the tails require a proverbial shift in thinking from "The dog wags its tail" to "The tail wags the dog". Events driven by extremes are very often the case in contagion cases where a small group drives something from the micro to the macro. For example, noone predicted the Orange Revolution in the Ukraine or the Arab Spring movement. So what drives these unpredictable things?

One way of modelling this is a Granovetter model. Take N individuals, each of which has a threshold at which point they will join the movement, *e.g.* each individual requires 50 or 100 others, *etc.* to join the movement as well. How does the outcome vary as we vary the thresholds?

A related question is, if we have a group of friends and anyone can start wearing a purple hat, when do you start wearing it? Let's say the possible thresholds for five people are $0, 1, 2, 2, 2$, where 0 doesn't care what other people do and the person with 1 needs one friend and the people with 2 need two people to buy the hat as well. So because we have one guy buying the hat anyways (the 0 guy) he will definitely get a hat and influence the guy with 1 to buy one too. Now we have two people with the hat and so the other three people buy the hat too and everyone has one.

Another example: $1, 1, 1, 2, 2 \rightarrow$ Nothing happens. Nobody has the threshold of 0 and so nothing happens.

Another example: $0, 1, 2, 3, 4 \rightarrow$ All end up with the hat. So basically, the people that really didn't want the purple hat (3 and 4) still had to buy the hat because a minority (the person with 0) started a contagion.

Even if you look at the averages, the $1, 1, 1, 2, 2$ group has a lower average threshold than the $0, 1, 2, 3, 4$ group. But the average (a Gaussian measure) doesn't matter here, it's the tail *a.k.a.* the extremes, the one 0 in the second group, that drives the effect. In general, for events driven by the extremes (power laws), the arithmetic mean is a meaningless measure.

This means collective action is more likely to happen if the thresholds are lower, and most importantly, the more variation you have in the thresholds (*e.g.* a single 0) the more likely contagion will occur. Thus, in predicting these phenomena—arguably the big events in history such as revolutions—the average sentiment is not enough. We need to know the distribution to be able to predict contagions/collective action, *etc.* The more heterogeneity in a specific group, the more likely collective action becomes.

2.4 Standing Ovation Model

The Standing Ovation Model builds off of the Granovetter model. When a performance ends, you have a short window to decide if you are going to stand up. And when other people around you do stand up, you need to decide very quickly if you stand up too. Standing ovations are a nice domain to think about when considering how people follow certain rules. Standing ovation is a sort of peer/group effect and can also be a piece of information. If a knowledgable person stands up, then you might deduce that the show was good.

Let's write the following model. Your threshold to stand is T, and an objective (non-personal) judgement of quality of the show is Q. The signal you get though is $S = Q + E$ where E is an error term. This error is your subjective filter of the show, including your personal taste, expertise, *etc.* If $S > T$ then you stand, but you also stand if more than $X\%$ of the people in the audience stand. So we can increase the probability of standing if either Q increases (better quality show), or T reduces (lower threshold), or X reduces (more easily influenced by others).

What does it mean for X to be big? You need a ton of people to stand, *i.e.* you are very secure of your opinions and have low social conformity. What about small X? People quickly jump onto a bandwagon—they have high social conformity.

Let's say we have 1000 people with $T = 60$ and $Q = 50$. Because $50 < 60$ noone stands up. But if E is between -15 and 15, then the signal goes from 35 to 65 and some people stand up. If E is between -50 to 50 then our signal goes from 0 to 100, and thus 40% of people stand up because they are over the $T = 60$ threshold. Now if 40% of people stand up then it is likely that we will have a standing ovation because people tend to conform to such a large group. So if $Q < T$ then you are more likely to get a standing ovation when E is a large variation. What causes E to be a large variation? A diverse or unsophisticated audience, or a multidimensional and complex performance.

Let's now extend this model a bit further to the advanced standing ovation model. People at the front can't see the rest of the audience, but everyone else can see them. This means they are not influenced by anyone, but can influence everyone else. People at the back can't be seen by anyone, but see everyone else. Hence, they can observe what is going on more accurately but noone will react to what they do. Also adding groups or dates increases the likelihood of standing ovations, because if one of the people in the group stands up, it is likely that everyone in the group will stand up too.

In summary, what makes a standing ovation?

- High quality show
- Low threshold to stand up
- Larg peer effects
- High variance in tastes, opinion, expertise in the audience
- Celebrities, high profile, social influencers up front
- Big groups and dates

Of course these things are equally applicable to collective action, people doing up their houses in neighbourhoods, *e.g.* you give some people a lot of money to rebuild their house and then other people will want to follow, investing their own cash to do so.

2.5 The Identification Problem

How do you figure out if something happened because people sorted to be with people that are like themselves, or because of peer effects/group dynamics? For example, sorting is manifest in segregation in cities and sorting of students in schools into peer groups, whereas peer effects manifest in the northern USA saying "pop", the eastern and western USA "soda", and the southern USA "coke".

However, some questions are not as clear-cut as this. Why do people in a specific areas increasingly vote for the same party? A sorting explanation is that democrats move to live with democrats. A peer-effect explanation is that people are influenced by their neighbours. Similarly, there are regions where the number of hospice days is the same over a large areas of the country. Do good doctors move into the same areas, or do peers influence each other to take advantage of the system/free-ride?

Another problem is that it is very difficult to tell after the fact if sorting or peer-effects are at play, because the outcomes tend to be the same. So what you need to do is find evidence of the process and you need to have dynamic data, not just a particular snapshot in time, to be able to distinguish between the two. To then truly figure out which one of the two effects you are looking at, you need multiple models as a reference to be able to figure out what is happening.

3 Modelling Aggregation

3.1 Aggregation

In the social sciences, many models are written as *agents + behaviours = outcomes*, where the +-sign represents the aggregation of two different entities. For more interesting/complex phenomena, aggregation is often more complicated than arithmetic aggregation, *i.e.* via $1 + 1 = 2$. Often, aggregation is nonlinear, for example, tolerant people in the micro can still lead to macro-level segregation (as per Schelling's segregation model).

In fact, aggregation can lead to counter-intuitive effects when properties emerge as a result of interactions. One issue with reductionist science is that when you study individual agents or particles they may behave one way in isolation, but when they are all interconnected we can observe some new phenomenon. For example, a single water molecule can be easily understood in terms of all the chemical properties. But from understanding a single water molecule we cannot appreciate the property of wetness, which only arises when multiple molecules are bonded together and our hand slices them apart. The wetness emerges as an emergent property of the interaction of many water molecules.

Aggregation of action allows us to:

1. Predict points and understand data.
2. Understand complicated patterns and outcomes that arise from simple rules. Indeed, very simple rules can lead to very complex outcomes, just taking binary on/off rules you can have equilibrium, periodicity, chaos and complexity.
3. Work through logic. If you like A more than B, and B more than C, then how do these preferences aggregate?

3.2 Central Limit Theorem

A simple and effective model for aggregation. Each agent makes an independent decision and we will measure the choices made via a probability distribution, which we can use to make predictions about what will happen.

The central limit theorem states that if I add up a bunch of independent (not influenced by each other) decisions, then we end up with a Gaussian distribution, *i.e.* choices aggregate around the mean. For example, flipping a coin twice and counting the different possibilities of getting heads: 0 heads 2 tails; 1 head 1 tail; 2 heads 0 tails. So we have 0H, 1H, 2H with probability 1/4, 1/2, 1/4 which is a little bell curve probability distribution.

If we have N different outcomes then the mean is $N/2$ and we can fit a bell curve to that. If we know that the mean is not $N/2$ then we can use a binomial distribution, in which case the mean is $p \times N$. In this case, the distribution can still be plotted using a Gaussian but with a different mean value.

Another concept is the standard deviation (std.dev.), which measures how far the distribution of the outcomes is spread out from the mean. The stats decay exponentially in terms of the deviation:

- ± 1 std.dev. includes 68% of possible outcomes
- ± 2 std.dev. includes 95% of possible outcomes

- ± 3 std.dev. includes 99.75% of possible outcomes

If the mean is 100 and the std.dev. is 2, then 95% of the time we will get an outcome between 96 and 104.

For a Gaussian distribution with mean equal to $N/2$, the std.dev. $= \frac{\sqrt{N}}{2}$. So for example, if $N = 100$, then the mean is 50 and the std.dev. is $\frac{\sqrt{100}}{2} = 5$. So 95% of the time, we will be between 40 and 60, and 99.75% of the time we will be between 35 and 65. One case where this example holds is for flipping one hundred coins and counting heads. If the mean is not $N/2$ but $p \times N$ instead, then we have std.dev. $= \sqrt{p(1-p)N}$, which means that if $p = 1/2$ we recover $\frac{\sqrt{N}}{2}$ again.

Example: The Boeing 747 has 380 seats and we have a 90% passenger show-up rate. An airline sells 400 tickets to ensure that the flight ends up full. The airline assumes that this probability is based on independent decisions (not true for big groups or bad weather, where the event affects more than one person).

Mean number of people turning up $= 400 * 0.9 = 360$, which is less than 380 so the airline should be fine on average. But let's look at the distribution. Std.dev. $= \sqrt{.9 * .1 * 400} = \sqrt{36} = 6$. So we have a bell curve with a mean of 360 and a std.dev. of 6. Which means 68% of the time there will be between 354 and 366 passengers, 95% of the time there will be between 348 and 372 passengers, and 99.75% of the time there will be between 342 and 378 passengers. So 99% of the time the airline won't overbook!

One way to interpret the central limit theorem is that we have a bunch of random variables which we assume to be independent and with a finite variance, *i.e.* they are bounded by some fundamental constraints. Under those circumstance, when we add everything up we will get a bell curve.

A lot of the predictability of the world results from random variables that are bounded by some constraint so that the central limit theorem applies. But of course if choices are not independent then we can get much more aggregation and feedback loops (interdependence). Under these circumstances black swans may occur and it may be all but impossible to predict what occurs.

3.3 Six Sigma

Six sigma quality control arises from the normal distribution to produce components with fewer variations.

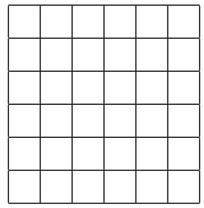
So to recap: in a normal distribution the distribution lies within 68% for one standard deviation from the mean, 95% for 2 standard deviations and 1 in 3.4 million for 6 standard deviations.

For example, if I have 500 pounds of banana sales with a standard deviation of 10 pounds, then six std.dev. are 60 bananas. So if I store 560 pounds of bananas, then even if I have a 1 in 3.4 million event, I will have stored enough.

Or say I require a 500-560 mm metal thickness. So the mean is 530 mm. So now I want 500 and 560 to be within 6 sigmas. So if my six sigma is 30, then my std.dev. is $30/6 = 5$ mm. Hence, if I can get my std.dev. tolerance down to 5 mm I will be within the range for all statistical purposes. Of course, getting down to six sigma is where the rubber hits the road, as this is the tough engineering challenge.

3.4 Game of Life

The Game of Life is a very simple model of aggregation. It's one type of cellular automata problems. It's not about a specific application (global warming, finance, *etc.*) but a toy model. So it's a very basic model that shows the complexities of life and illustrates how hard it is to deduce the micro from the macro.



Let's introdce the Game of Life, using a grid of squares as shown above. We know that each non-border cell has 8 neighbours. Let's assume that each cell can either be on or off, and let's create the following rules. If a cell is currently off it can only come on if it has exactly three squares around it on. If the square is currently on and there are less than 2 neighbours on around it, the square turns off. If there are more than three neighbours on around any square, then the square also turns off. So, in summary

- If square is off, it needs exactly three neighbours on to go on too.
- If square is on, it needs two or three neighbours on to stay on.

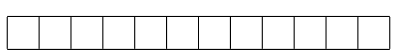
This example of very simple rules can give rise to a lot of complexity. In fact, we can get the following:

1. Blinkers going back and forth between two states, or cycles through numerous states
2. Growth
3. Everything just dies off
4. Randomness
5. Gliding movement

The moral of the story is that units that follow very basic rules can aggregate to give rise to very intricate phenomena, *i.e.* emergence from complexity. So the Game of Life nicely reinforces the concept of evolution that there can be bottom-up design from the interaction of very basic rules, without a great designer controlling things from the top. This is self-organisation. Patterns without a designer.

3.5 Cellular Automata

The Game of Life is a particular cellular automata model, where simple rules can lead to amazingly complex outcomes. So what has to be true about these cellular models to get a specific outcome? Assume a one-dimensional grid of squares that can either be on or off. The Game of Life was a grid with 8 neighbours. In its one-dimensional line form, each cell only has two neighbrous. This



is a simpler version and we can exhaustively study each of the rules, *i.e.* what behaviour does each produce. The possible outcomes are: equilibrium, alternation, chaos (randomness) and complexity (structures).

So basically what you do is take three cells in series and given the 8 combinations of cells on/off you decide what happens to the middle cell. You then just move through time for that rule set. The profound idea behind this is that we can get everything by simple yes or no questions, and everything we see around us could arise from simple binary rules.

Physicist John Wheeler sums this up nicely:

> "It from bit, otherwise every it, every particle, every field of force, even the space time continuum itself, derives its function, its meaning, its very existence entirely, even if in some contexts indirectly, from the apparatus solicited answers, to yes or no questions, binary choices. Bits, it from bit, symbolizes the idea that every item of the physical world has at its bottom, a very deep bottom in most instances, an immaterial source of explanation that which we call reality arises in the last analysis from the posing of yes no questions. And the registering of equipment evoked responses. In short, that all things physical are information theoretic in origin and that this is a participatory universe."

A way of classifying these rules is via Langton's lambda. Basically Langton counted the number of rules that produce an "on" and then divided by the total number of possibles "ons", that is 8. So $\lambda = 0/8$ will just produce death and $\lambda = 8/8$ will produce stuff everywhere. If you are in the middle, around $\lambda = 4/8$ and $5/8$ you are more likely to get something interesting. Chaos is produced by $\lambda = 2 - 6$ and complexity only for $\lambda = 3 - 5$. So chaos and complexity are produced by intermediate levels of interdependence.

Summary:

1. Simple rules combine to produce almost anything
2. It from bit
3. Complexity and randomness require a goldilocks of interdependence

3.6 Preference Aggregation

The central limit theorem is about aggregating numbers and actions. The Game of Life and cellular automata aggregate rules. Now let's talk about aggregating preferences.

How do we represent preferences? One way to quantify preferences is to force actors to make a comparison and look at the revealed actions. This will create a ranking for different classifications. But note that these rankings require judging similarities, and hence judging occurs by comparing features that are representative of both entities being compared. Precisely, this caveat of measuring preferences by similarities creates a problem with transitivity. If $A > B$ and $B > C$ then $A > C$. But a lot of the time people don't follow this rule for preferences.

The general preference model would just compare two choices at a time for a larger set. So for example, writing $A > B$ and $B > C$, then $A > C$ has 8 different possibilities, because I can rearrange this in 2^3 ways. But this system may break transitivity. So a better way might be to just sort as in $A > B > C$ in one go. In this case I only have 6 possibilities.

Suppose I have a bunch of people with different preferences, how do their preferences add up? What are the general preferences of society? It gets tricky, of course, when people have different preferences, *i.e.* different orders of $A > B > C$.

- Person 1: $A > B > C$
- Person 2: $B > A > C$
- Person 3: $A > C > B$

There is some diversity here, but clearly option C is the worst. Also, two people prefer A over B and so the general preference is $A > B > C$. Now consider,

- Person 1: $A > B > C$
- Person 2: $B > C > A$
- Person 3: $C > A > B$

There is no clear winner now. But let's boil it down to a pair-based vote.

- C vs A: 2:1
- C vs B: 1:2
- A vs B: 2:1

This breaks the transitive property again (because we just cycled the preferences). So the gist is that every individual made a rational consistent choice about preference that was transitive, but when we aggregate the preferences we break rationality/transitivity. This leads to a general paradox known as:

Condorcet Paradox: Even if each person behaves rationally, the collective may not. Meaning, that if we aggregate preferences, we may not get what the majority actually wants.

4 Decision Making

We now turn to decision models of how people make or should make decisions. We are ultimately doing this for two reasons:

1. Normative reasons: Use models to come up with prescriptions that allow us to make better decisions.

2. Positive reasons: Use models to predict the behaviour of others, and why people make the decisions they do.

Fundamentally, there are two classes of problems:

1. Multi-criterion problems: We need to weigh multiple options against each other. Consider deciding between two different cars. How do I quantify the choices to decide what to buy? One way is to write down a bunch of criteria like fuel consumption, comfort, *etc.* and then rank them. Or we could have a spatial model, whereby a design space of two or more variables is sketched, and we then try to measure the distance of the different options from our ideal.

2. Under uncertainty: Risk-benefit analysis in terms of probability, *e.g.* decision trees, or Value of Information. This latter option attempts to measure how much a piece of information is really worth to us in making a decision.

4.1 Multi-criterion Decision Making

The idea is that there are a lot different dimensions and you are trying to decide which choice makes you happiest.

Qualitative decision: Should I buy this house or the house down the street? There are specific features like square feet, #bedrooms, #bathrooms, lot size, location, condition *etc.* We can make a table and fill in the feature numbers for either choice. Then we count the number of features that either option wins and add them up. So if it is $4:2$ for house 1, then this house wins. You could also weigh the alternatives, *i.e.* give each feature a multiplicative factor to weigh its importance. What happens if in your gut you don't like your option? Perhaps you are missing a criterion (like style of house) or are not weighing a feature correctly?

Caveat: We don't want the model to tell us what to do, we want it to help us make better decisions. Equally, if we see someone make a decision, given the criteria, we can try to figure out why he/she made the decision.

4.2 Spatial Choice Models

Rather than comparing different criteria, as we did for the multi-criteria decision model, we will decide on an ideal point and then compare the different options to that ideal.

Example: The Ideal Burger—2 cheese slices, 2 patties, 2 tomatoes, 4 tbsp of ketchup, 4 tbsp of mayo and 4 pickles. This is our ideal point in six-dimensional space. So let's compare this against the Big Mac and the Whopper. The Big Mac scores $2, 2, 0, 3, 4, 6$. So how much do I like it? Let's take the absolute value of the difference between Big Mac and the Ideal: $0, 0, 2, 1, 0, 2$. If we add up we get 5. Similarly, for the Whopper we have $2, 1, 2, 3, 4, 4$ and so the absolute value of the difference is $0, 1, 0, 1, 0, 0$ which equals a total of 2. Hence, the distance of the Whopper from the Ideal is only 2, while the Big Mac is 5. So we'd rather buy the Whopper.

If we were mapping this to voting, we could have social policy (liberal/conservative) and economic policy (liberal/conservative) so we can figure out normatively what to vote. Positively, I can also figure out about what others value. So if we see a friend eating a Big Mac, we don't know his/her ideal point but we can compare the differences between the Big Mac and the Whopper. These burgers are the same for cheese, ketchup and mayo, but very different for pickles and tomatoes. So he/she might not like tomatoes or prefer pickles.

4.3 Probability

Multicriterion decision making and spatial decision making are related to Amos Tversky's notion of comparison from features and comparison by distance, *i.e.* judging how similar things are by comparing how much they share and how far apart they are.

For decision making under uncertainty we need something else: probability.

Axioms:

1. The probability of a single event is always between 0 and 1
2. The sum of all possible outcomes is equal to 1
3. If A is a subset of B, then the probability of A occurring is less then the probability of B occurring.

Three types of probability:

1. *Classical*: Academic probability where you can write down in a pure mathematical sense what the probability is, *e.g.* rolling die. This is deductive.
2. *Frequency*: Count the number of specific outcomes, *e.g.* how often does a specific outcome occur? Monte Carlo analysis would be a good example—counting things up in a historic time series. This is inference.
3. *Subjective*: Cases where we need to guess, *i.e.* assign subjective probabilities. We can use Bayesian reasoning using base rates to give us a fair guess.

So in the cases where we don't have the option of classical- or frequency-based approaches, we need to resort to subjective reasoning, but this is best supported by a model.

4.4 Decision Trees

Decision trees are useful when there are a lot of contingencies—lots of options. We can also use them to infer how someone views the world.

Example 1: 60% chance of making 3pm train for \$200, or otherwise take 4pm train for \$400. What to do?

If we buy, then we have 60% chance of making train at \$200, and 40% chance of paying $200 + 400 = \$600$. If we dont buy, then we know we pay \$400. What's the better choice? $0.6 * 200 + 0.4 * 600 = 120 + 240 = \$360 < \$400$. Buy!

Example 2: We can win a scholarship for \$5000. There are 200 applicants that write a 2 page essay, and 10 finalists that write a 10 page essay. How do we make the choice if it is worth applying?

What are the costs of writing the two-page essay? Let's say \$20. What are the costs of writing the ten-page essay? Let's say \$40. The chance of going through the first round is 5% and the further win probability is 10%.

- Winning the second stage entails a 90% chance of losing \$60, and a 10% chance of winning \$5000 with a \$60 upfront cost. Hence, the net value is $0.1*(5000-20-40)+0.9*(-20-40) =$ \$440. So writing essay 2 costs us \$40 but the payoff is \$440, so we should definitely write essay 2, when faced with the option.

- For essay 1, we have a 5% chance of getting to the \$440 and a 95% chance of wasting \$20. So the net value is $.05 * 440 - 0.95 * 20 =$ \$3.

Because \$3 > \$0 we should consider applying. An alternative way of calculating the same thing would be to do the math in one swoop: $0.05 * 0.1 * (5000 - 20 - 40) + 0.05 * 0.9 * (0 - 20 - 40) + 0.95 * (-20) =$ \$3.1.

Inferring probabilities: We have a friend who tells us that he has a risky investment with a payoff of \$50,000 for a \$2,000 investment. Basically, he is thinking that: $50p - 2(1 - p) > 0 \Rightarrow 52p > 2 \Rightarrow p = 4\%$. So our friend is assuming that there is a greater than 4% that this investment will pan out.

Infer payoffs: We have got a standby ticket taking us back home for Christmas, and are still at our hotel. The airline says we have a 1/3 chance of making the flight. If we stay home, then the payoff is zero. If we go to the airport, then there is a 1/3 chance of going home and a 2/3 chance of not. How much do we value going home? $0.33(V - c) - 0.66c > 0 \Rightarrow V > 3c$, so the value of seeing our parents has to be greater than three times the cost.

4.5 Value of Information

Decision trees allow us to figure out what to do under uncertainty. Let's say we know the probability of an outcome. How much would that information of knowing the outcome be worth to us?

Roulette wheel: A roulette wheel has 38 things we can bet on. The odds of winning are 1/38. What is the value of information whether our number wins? If we can win \$100 dollars and we win 1/38th of the time, then knowing when our number comes up is \$100/38.

What is the value of information of the winning number? The number that wins is arbitrary and so knowing any number that wins is of course \$100.

Car example: Buy a car now or rent for \$500? There is a 40% chance that the company offers a \$1000 cash back offer starting next month. How much would we pay to know for sure that there will be a cash back program? Calculate the value without the information, then calculate the value with the information, and finally calculate the difference.

Buy now: full price. Buy in a month: cash back happens: $0.4(1000 - 500) = 200$. Cashback doesn't happen: $-0.6 \times 500 = -300$. This gives us $-$\$100. With information, if someone tells us what to do, *i.e.* there is no uncertainty, then there are only two options. 40% of the time we make \$500, *i.e.* $0.4(1000 - 500) =$ \$200, and 60% of the time we won't buy so we don't save anything. This means we get $0.4 \times 500 - 0 =$ \$200. So the value of information is $200 - (-100) =$ \$300.

5 Modelling People

"Imagine how difficult physics would be if electrons could think." — Murray Gell-Mann.

5.1 Thinking Electrons

Step back and consider how we model people? If we want to understand organisations, then we need to understand the parts. We have belief structures, objectives, and emotions which tend to make things a bit messy. On top of that we are all different. The combination of messy and different is hard (impossible?) to model.

Models:

1. Rational: based on the assumption that people behave such as to optimise for their goals. There is an objective function (market share, votes, profits, *etc.*). People then optimise to maximise, *e.g.* your utility is a function of leasure and consumption, which both follow diminishing returns (square root is a good fit) so we might write the utility function as $U = \sqrt{CL}$ and then we figure out what $\max(U)$ is given the costs on C and L. Of course, we don't all sit around calculating these functions, but economists assume that in aggregate we do as a first-order approximation.

2. Behavioural: model people as close as possible to how they really behave, *i.e.* include their biases and heuristics. Don't assume they are rational, observe how they behave (and in fact they behave irrational in predictive ways), and then use those observations to predict.

3. Rule-based: use psychology to figure out certain rules that people seem to be using and how this predicts that they will act. The Schelling model of segregation was a classic, simple rule-based model.

All people models are a combination of these three, and each model will provide slightly different predictions of how different people behave.

5.2 Rational Actor Model

Has come under a lot of critcism because of biases and heuristics literature. It is still a useful tool to have in your pocket.

1. Assume people have an objective. A goal or purpose.

2. Given the objective, people try to optimise the outcome.

For example, maximise profits, revenue, market share, happiness, voting behaviour, *etc.*

Example: maximise revenue (r) = price (p) $*$ quantity (q), where we assume that $p = 50 - q$. Therefore $r = 50q - q^2 \rightarrow dr/dq = 50 - 2q = 0 \rightarrow q = 25$ to maximise.

When we assume rationality people often assume selfishness. This is not true. Example, if you find \$100 on the street while walking with a friend, the rational thing to do is pocket it all. But if you really care about your friend then you might share. So there is nothing intrinsic about selfishness in the rationality model, it depends on your utility function. If the utility function is altruism then you would share with your friend. Rationality just defines behaviours that are in alignment with maximising what you value, and what you value can of course be highly variable.

Mathematically, if your utility is a function of consumption and donation, and there is a diminishing rate of return we might say that happiness (h) = $\sqrt{cd}$, *i.e.* some donation (d) and some consumption (c) will make you happy. If donation and consumption are related by $d = 40 - c$ then $h = \sqrt{c(40-c)}$ and so the optimal consumption and donation are equal at $c = d = 20$. This would be the rational choice. All donation would be irrational in this case.

Example: If we are interacting with other people, then the maximisation becomes more difficult because we need to account for the other actors too. Do to this, we also assume rational behaviour. For person 1, it doesn't really matter what he does on a Friday night. He/she gets some happiness out of staying home and going to the city. But person 2 is happier when he/she does something with person 1. Now person 2 could assume rationality and say person 1 prefers the city and try to meet him there, rather then taking the bet of trying to meet him at home.

Table 1: Payoff function for two-player game

	Home	City
Home	1,1	1,0
City	2,0	2,4

So when do we typically see rational behaviour? When

- the stakes are high
- there are repeated decisions
- we have groups making decisions (caveat: there can also be group think)
- the choices are easy to make

Even though we know that these are idealised assumptions, rationality gives us a great benchmark of what would be optimal. Second, it is often easy to compete and so you can objectively figure out a benchmark, *i.e.* how far from rationality are people actually behaving? If we assume irrationality there are 1000 ways this could manifest itself and so it's basically impossible to write up a good predictive model. Finally, in the aggregate, people's irrationalities might cancel out and so overall rationality might be a good assumption.

5.3 Behavioural Models

In the rational model, you have objective functions and you make an optimal choice under this rationality.

But there are recurring observations from psychology that we make systematic errors in rational thinking. There is also some neurological evidence to support this.

In "Thinking, Fast and Slow", Danny Kahnemann makes the point that we have two different types of thinking systems, one fast that is intuitive, and one slow that uses rigorous thinking.

Example 1: Prospect theory You have two options, you can either gain \$400 for sure, or alternatively 50% of the time you gain \$1000 or 50% of the time you get nothing. Most people choose option A, especially if we crank up the amounts to \$400 million.

But if we turn this around to losses, *i.e.* losing \$400, then most people choose the second option. That is, they take a chance to lose nothing or potentially more. This is called loss aversion in that we feel more pain in losing something then we feel joy in gaining the same amount.

Example 2: Hyberbolic Discounting Option A is \$1000 today, and option B is \$1100 in one week. Most people choose option A. Option A is \$1000 in a year and option B is \$1100 in a year and one week. Then most people choose option B. This means people discount the immediate future more then the far future, and this is why it is called hyperbolic discounting (from the shape of the graph). If I am on a diet, I can't put the chocolate cake off right now there in front of me, but if someone asks me if I will eat one next week, of course I will say that I will be able to resist.

Example 3: Status Quo Bias Two options, check a box to contribute to the pension fund or check a box not to contribute to the pension fund. Which box would you use as an institution to maximise pension contributions? The second one, because if you make the choice for someone to begin with and then ask them to opt out, they opt out less frequently then when asked to opt in.

Example 4: Anchoring Bias Prime someone with a number, and then ask them to guess something. The size of the first number will influence their next guess. For example, if you ask someone what the last two numbers of their cell phone number are, and then ask them to guess how many countries in Africa are in the UN, the first two numbers will influence their guess.

Of course, there are a ton of these biases. Some people are critical of these because they are WEIRD: observations based on western, educated, intelligent, rich, industrialised subjects. The big question is which of these biases hold cross-culturally? People also learn from their biases, so how strong are these biases in repeated interactions and when do they go away?

It's computationally difficult to incorporate these biases in models, hence lots of social models break down when trying to account for them. So you need a two track analysis: start with a simple model to figure out what influences the problem, and then figure out what biases influence this rational model as a second-order effect (which biases are most influential in the case we are consdering).

5.4 Rule-based Models

People follow very basic rules. For example, in the Schelling model, people decided to move or stay depending purely on how many like-minded individuals lived around them.

There are four different variations: the rules can be fixed or adaptive, and the choice can be decision-based (sole agent) or a game (multiple agents). In games, for example, my payoff depends on what other people do, and so having adaptive rules might be better, but if I am a genius and the others not so clever, then a fixed rule may be best.

Fixed: Fixed rule, *e.g.* take the straight line to travel between two points, *i.e.* the most direct route. Fixed strategy, *e.g.* divide evenly, tit-for-tat, *etc.*.

Adaptive:

- Gradient-based method: people start off doing something and then tinker around with the recipe, moving in the correct direction to reach the optimum, that is, up a hill. In a random method, you would randomly add other stuff/take away stuff to haphazardly optimise.
- Adaptive strategy: Either an adaptive response where I watch another person and then think what is the best rational response, or mimicry, where I just copy what other people are doing well.

Sometimes optimal rules are very simple, and in that case, it would make sense to use a rule-based approach. Simple rules can often be exploited though, if other agents in a game realise what your rule is and then just exploit it (classic case of a blind rule that wasn't updated when reality changed).

Rules are great because they are easy to model, they can capture a lot of the optimality effects in a very simple rule, but they are somehow ad hoc and are exploitable.

5.5 When Does Behaviour Matter?

Different types of models: rational, behavioural, and rule-based. Does it matter which model you use? Part of the time the reason we model is to figure out how people behave, *i.e.* are they rule-based, rational or irrational?

Example: Two sided market Buyers are willing to buy between \$0 and \$100. Sellers are willing to sell between \$50 and \$150. If you are a buyer, then you will bid slightly beneath your threshold to save money, and a seller will ask for slightly above his/her threshold to make more money.

The relevant buyers will be willing to buy between \$50 and \$100. And the relevant sellers are willing to sell between \$50 and \$100 as well. This is the intersection of the sets above. So the median in this range is \$75 and so we could argue that the average selling/buying price will be \$75, rationally speaking. If you think about behavioural analysis, people might just throw out a bunch of random numbers because they don't actually know what the optimum is. In this case you can argue that the deviations from \$75 just average out. A rule-based model can also explain this using a zero intelligence agent. A buyer chooses a random value less than his/her threshold, and if you're a seller, you pick a random value more than your value. If I am a buyer with value \$40, I might buy it for \$20, and if I am a seller with a value of \$60 I might sell it for \$63. If you run this model lots of times you get pretty much \$75 on average. So it turns out that when you have markets and institutions, behaviour does not matter as much.

Example: Race to the bottom game Pick a number between 0 and 100. Who, ever is closest to 2/3 of the mean wins. A rational person would bid the Nash equilibrium of 0. If say everyone chooses 6, then the mean is 6, so 2/3 of 6 is 4. But everyone knows it's 4 so if everyone now chooses 4, then 2/3 of the mean is $8/3 = 2.666$, and so we get a race to the bottom until everyone guesses 0.

If you are super biased, then you would probably guess 50 (maybe you don't get the 2/3 thing). A rule-based approach might be, well everyone will guess 50 so I'll guess $2/3 * 50 = 33$. Now some people will take it one step further and say $2/3 * 33 = 22$. And other take it one step further $2/3 * 22 = 14$. So we are creeping down to the equilibrium value of 0. So what people do is a combination of all of this. They start with a heuristic choice (say 50), and then apply the rule a couple of times.

Let's say we have two rational people (R) and one irrational (X). Suppose I am part of the rational group, what will the irrational person pick? If they pick X what do we pick? Our number has to be $2/3(R+R+X)/3 = 2(R+R+X)/9$, now solve for R: $9R = 2(2R+X) \to 5R = 2X \to R = 2X/5$. So if the irrational person chooses 50, then R becomes 20. So if the rational people truly want to be rational, and account for the fact that an irrational person is in the audience, then they will not guess 0 but 20 instead (in this example).

The simple lesson is this, rational behaviour is a really good benchmark but including behavioural biases will inform us of some of the differences. If the differences are big, then we can decide if we need to revise our model.

6 Categorical and Linear Models

6.1 Categorical Models

Categorical models are a very simple class of model that allow us to make sense of data by binning the data into individual buckets. Putting stuff in categories allows us to reduce variation in the data, *i.e.* find things which are similar.

For example, consider the initial public offering of Amazon back in the 1990's. As an investor at that time, is it worth the money? This depends if at the time, you saw Amazon as a delivery company (low margins), or as an IT service company (large margins). So based on what box you put Amazon in, informs your decision whether you will invest or not.

A lot of the time we also just categorise or lump because it makes our life easier. Consider the following example about different foods to eat:

Item	Calories
Pear	100
Cake	250
Apple	90
Banana	110
Pie	350

How much variation is there in this data? The mean calorific content is 180. If we subtract the mean from everything we get: 80, 70, 90, 70, 170; then square that to get: 6400, 4900, 8100, 4900, 28900; and add everthing up to 53,200. This total figure (sum of the squares between each item and the mean) is known as the total variation.

Now, what are the obvious categories: fruit and dessert. Let's repeat the total variation calculation for these two categories. Fruit calories: 90, 100, 110. Mean is 100; subtract mean, square, then sum: $100 + 0 + 100 = 200$. Similarly, dessert calories: 250, 350. Mean is 300; subtract mean, square, then sum: $2500 + 2500 = 5000$. So, just by categorising we have been able to massively reduce the variation in the data.

How much have we reduced the variation? Fruit variation + dessert variation $= 200 + 5000 = 5200$. How much of the variation did the categorisation explain? $(53200-5200)/53200 = 90.2\%$. So $R^2 = 90.2\%$ of the variation is just down to categorising into fruit and dessert. Note that, $R^2 = 1$ means that all of the variation is explained, whereas $R^2 = 0$ explains nothing.

We can also push the categories a bit further to get a finer grained appreciation of the data: fruit, dessert, veg, grain. One of the differences between experts and non-experts is that experts categorise correctly and place things into the correct categories. But remember, when interpreting categories and the effects of certain categories, correlation doesn't neccesarily mean causation.

6.2 Linear Models

Assume that there is an independent variable x, and some other variable y that is a function of x, $y(x)$. We assume that there is a particular relation between x and y which is linear, $y = mx + b$.

Example: Linear model of cost of TV: $\text{cost} = 15 * \text{inches} + 100$. This simple equation can tell us a couple of interesting things. For this equation, does y increase or decrease in x, and what is the rate of this decrease/increase? In this case y increases with increase in x and the rate is \$15/inch. What does a 30 inch TV cost? $15 * 30 + 100 = \$550$.

So we can use this model to make very simple predictions. Often the model is not entirely accurate, but even then it can still inform decision making. In a lot of cases, experts do not do any better than very simple linear models. The bottomline is you want the model to inform your judgement, rather than tell you exactly what to do.

6.3 Fitting Lines to Data

How do you draw the best possible line?

R^2 is the % of the variation explained. For example, if we have linearly distributed data on a plot and we draw a horizontal line to express the mean, we would get loads of variation about that line, hence low R^2.

Example: Kids' grade and shoe size:

Grade (x)	Size (y)
1	1
2	5
4	9

Here we do not care about the variation in grade, but about variation in shoe size, *i.e.* only the y values. The mean is: $(1+5+9)/3 = 5$ and the variation is $(1-5)^2 = 16$, $(5-5)^2 = 0$, $(9-5)^2 = 16$, with a total variation of $16+0+16 = 32$. If we take the line $y = 2x$, how far is this from the actual data?

- $(x = 1, y = 1 * 2 = 2)$ so that variation is $(2-1)^2 = 1$
- $(x = 2, y = 2 * 2 = 4)$ so that variation is $(4-5)^2 = 1$
- $(x = 4, y = 2 * 4 = 8)$ so that variation is $(8-9)^2 = 1$

So the total variation is 3. So now our R^2 (comparison of the mean to the prediction) is: $1-3/32 = 91\%$.

How to draw the best curve, *i.e.* the one that maximises R^2? We know that our equation is $y = mx + b$, which means that each data point variation is $(m*1+b-1)^2$, $(m*2+b-5)^2$, and $(m*4+b-9)^2$. The total error is (squaring and adding): $21m^2+14mb+3b^2-94m-30b+81$. Using calculus we can differentiate this equation with respect to m and b, and solve the following two simultaneous equations ($42m+14b-94 = 0$ and $14m+6b-30 = 0$) to get $b = -1$ and $m = 18/7$.

What do we do with a bunch of variables? Let's say Y = test scores, Q = IQ, T = teacher ability, Z = class size. We run a multi-variable regression analysis to determine the factors of the equation: $Y = a + b*Q + c*T + d*Z$. From intuition we would probably expect d to be $-$, b $+$ and c $+$. But without a model we can't really say which one of these factors is the most influential. For example, based on a regression of 78 studies, class size had a positive sign 4 times, negative 13 times, and no effect 61 times. So you would expect it to have no, if slightly negative correlation. We thought d was gonna be $-$, but it turns out it's probably not that negative. So in multiple variable models, we can figure out which variables have the largest effect and in which direction.

6.4 Non-linear

The problem with linear models is that the world is typically non-linear. Purely from a statistical point of view, there are infinitely more non-linear curves than linear ones. So what are the techniques from linear modelling that we can use for non-linear functions?

1. Approximate non-linearity with a succession of linear curves.
2. Break data into quadrants and then fit a curve that is linear in each of the quadrants taking into account the slopes of the other quadrants, *i.e.* a spline curve.
3. Include non-linear terms. Say the data looks like $\sqrt{x}$, then of course we can replace the regression with $y = m * \sqrt{x} + b$ or $y = m * z + b$ where $z = \sqrt{x}$. Hence, the generalisation is $y = m * z + b$ where $z \to f(x)$.

6.5 Big Coefficient

If we have a simple linear regression model: $y = a_1 * x_1 + a_2 * x_2 + b$, we can look at the two coefficients a_1 and a_2 and check which one is bigger.

So what you do is, you collect a bunch of data and run a multi-variable regression model to figure out which of the coefficients have the biggest bang for the buck, *i.e.* which variable describes the underlying data the most. We can then feed back and change variables. This massive regression approach kind of mutes the utility of an *a priori* defined theoretical model because we can just collect data about a phenomenon and run a series of regression models with arbitrary parameters until our model fits.

But such big-data approaches do not obviate the need for less statistical models, because understanding the data does not mean we understand where it came from. So just observing that something is so, does not mean we understand why it is so. Regression model provide no sense of explanation. Second, correlation does not imply causation. Data sampling is epistemologically tricky because we can have spurious correlations. Lastly, linear models tell us the sign and magnitude within the data range. We can only extrapolate so far from the available date without running into serious accuracy problems. This is especially so when we have feedback loops, *i.e.* over time the system changes. Under these circumstances, extrapolating linearly *ad infinitum* is obviously dangerous.

So big coefficient thinking is important but we have to be aware of its limitations. It might be important to identify the best "bang for your buck variables" and then optimise, but this precludes the chance of something entirely different and better (non-marginal) that you are not aware of because no hint of it shows up in your current data. This clash is big coefficient versus new reality thinking. Optimise versus new paradigm.

Big coefficient thinking is our small-scale safe marginal thinking of today. The interstate highway system cost $25 billion in 1956 which is $207 billion today. Who would spend the money now?? This is the new reality, creating an entirely new system. And in creating NEW systems, this is where models come in, how do we think outside the box? By using models.

7 Tipping Points

We shall now look at highly non-linear models, where a small change can cause a disproportionally big effect, *e.g.* a system that tips at a critical point. What we want to understand is what a tipping point is, and how to differentiate between phenomena that seem like tipping points but are in fact not.

The classic proverbial example is the straw that broke the camel's back. One more piece of straw has an unproportionally large effect than any other straw before it. Before the final straw, each straw might have only marginally influenced the height of the camel, but that last marginal straws breaks it's back and the camel collapses.

The housing market in 2008 was probably a tipping point because there was a discrete and fundamental change to the system, *i.e.* the state of the system before wasn't the same as after. Hockeystick growth is not a tipping point because nothing fundamentally changes in the underlying system. These are more exponential growths that are qualitative of the original system, there is no small change that leads to a large effect.

We will consider two famous models for tipping points: percolation models and the SIS (susceptible, infected, susceptible) model. Then we'll make a distinction between direct tips, where a particular action tips the system, and contextual tips, where a change in the system's environment forces the system to tip. We can also differentiate tips between the four classes (static stable, periodic, chaotic, complex), *i.e.* from static stable to periodic, or periodic to random, or random to complex.

7.1 Percolation Model

Take a grid of squares either filled in or not. The rule is that you can only jump from filled-in to filled-in box. So to go from the top of the box to the bottom of the box you need to be able to progress from filled-in box to filled-in box all the way down. This model is based on the physical phenomenon of water percolating through soil.

Let P be the probability that a square is filled. Given P, is it possible to percolate from top to bottom? For a big graph, as long as $P < 59.2\%$ nothing happens. But as $P \geq 59.2\%$ then we can get percolation. Why does it tip? Because adding one new square at a time slowly but steadily approaches the tipping point, until at some point, it becomes really likely that you make it to the bottom. So 20% to 21% doesn't have a large effect but going from 58% to 59% to 60% makes a huge difference.

This means there is a phase transition at 59% where the "water" spreads throughout the entire space, rather than arresting. This also means that there is a density for which fire in a forest is really likely to spread (*i.e.* $> 59\%$).

Banks: You can do the same thing in stress tests of banks. If a bank fails, it won't be able to pay its loans to other banks, so if these banks don't get that money, do they then fail too? And does the failure then percolate further? So as banks become more interconnected, failure more easily percolates and the system might be poised to tip. Note that for less interconnected systems a tipping point might be absolutely impossible under all conditions, whereas for more interconnectedness you might not see the tipping point until some critical condition is reached.

Information Percolation: How quickly do rumours spread? Well, you can use the same model. If a rumour is juicy enough, the likelihood is high that it will spread. The relationship between juiciness of rumour and the likelihood that it will spread is not nice and linear. It has a tipping point, it basically doesn't spread until some critical value, at which point everyone hears about it. The reason why you get this is because we are all interconnected in a network. In fact, the internet has connected us more than ever, increasing the likelihood of rumours spreading.

Innovation: A lot of the time people are trying to figure something out but can't, and then all of a sudden loads of people come with a solution or a new innovation. You can also think about this in terms of the process of innovation in general. The two points connected by percolation are the assumptions and the proven statement; in engineering, it is the problem specification and the solution. So as information and technology accumulates, the likelihood that you can percolate from point A to B increases, until at some point you reach the critical point and there are multiple paths to actually percolate.

7.2 Contagion I—Diffusion

This model comes from epidemiology and is known as the SIS model (susceptible - infected - and susceptible again). Also, SIR models exist (susceptible - infected - recovered).

We want to show that this model produces a tipping point, *i.e.* the model leads to a basic reproduction number. If this number is bigger than one, everyone gets the disease, if it is less than one noone gets the disease. A simpler model is the diffusion model, whereby everyone just gets it. For example, W_t is the # of people in a group of N that have the "wobblies" at time t, and $N - W_t$ is the # of people without. τ is the transmission rate. So if two people meet what is the likelihood that the disease gets transmitted?

Let's say that the probability of someone having the wobblies is W_t/N and the probability of not having it is $(N - W_t)/N$. So the probability of transmitting between two random people is $\tau * W_t/N * (N - W_t)/N$. Let's add another variable, the contact rate c, where c is the probability that two people will meet. So the number of meetings you get is the total number of people, times the probability that two people will meet, times the original transmissions probability: $N * c * \tau * W_t/N * (N - W_t)/N$. So the total number of infected people at time $t + 1$ is:

$$W_{t+1} = W_t + N * c * \tau * \frac{W_t}{N} \frac{(N - W_t)}{N} = \text{old} + \text{new} \tag{8}$$

What does this tell us about the diffusion of the disease? It is basically a classic cumulative "S"-shaped curve that starts slow, then picks up, and converges to a value where everyone is infected, *i.e.* $W_t = N$. So in the beginning there are few people who have it, so it can't spread very fast. Later there are few people to spread it to, and so diffusion slows down. In the middle when 50% have it, contagion is high.

But the point is, there is NO tipping point. It's just the natural exponential diffusion of the process. There is deceleration and acceleration, but there is no fundamental underlying change of the system. This diffusion S-curve, or hockeystick curve is not a tipping point (Facebook didn't tip, it diffused).

7.3 SIS Model

The difference between the SIS and diffusion model is that in the SIS model, after someone has been infected, they can recover, and then be infected again. So we would expect some back and forth oscillations. For example, once you have had the flu, you can recover but then get a mutation.

$$W_{t+1} = W_t + N * c * \tau * \frac{W_t}{N} \frac{(N - W_t)}{N} - a * W_t \tag{9}$$

Note the extra $-a * W_t$ term. These are the people that become cured, *i.e.* a is the rate of cure. So if people get better quicker than they can get sick, then the disease can't spread.

By rearranging we get,

$$W_{t+1} = W_t + W_t * \left(c * \tau \frac{(N - W_t)}{N} - a \right) \tag{10}$$

Suppose we are early on in the disease. This means W_t is small and $(N - W_t)/N$ is of the order of 1. So

$$W_{t+1} = W_t + W_t * (c * \tau - a) \tag{11}$$

and the disease won't spread if:

- $c * \tau - a < 0$, but will spread if
- $c * \tau - a > 0$.

The number $R_0 = c * \tau / a$ is known as the basic reproduction number, and if $R_0 > 1$ the disease spreads, and if $R_0 < 1$ the disease doesn't spread. So this means we have a tipping point when the system changes from not spreading to spreading at $R_0 = 1$. Here are some typical reproduction numbers for diseases:

- Measles: $R_0 \approx 15$
- Mumps: $R_0 \approx 5$
- Flu: $R_0 \approx 3$

There might be a ton of diseases out there with an $R_0 < 1$ and the reason why we don't hear about them is because they don't spread.

How do we stop this thing from spreading? Vaccines. How many do you need to vaccinate? Let $V = \%$ of people vaccinated. So the new basic reproduction rate is $r_0 = R_0 * (1 - V)$. We want $R_0 * (1 - V) < 1$, so $V > 1 - 1/R_0$. So how many people do we need to vaccinate to prevent diseases from spreading?

- Measles: $1 - 1/15 = 14/15 = 93.3\%$
- Mumps: $1 - 1/5 = 4/5 = 80\%$

There is a clear tipping point here. If we vaccinate 79% of the population against mumps, then it'll spread and everyone will get it. BUT vaccinate 81% of the people and it won't and no-one will!!

So the model has informed us about a policy. Vaccination has NO, absolutely NO effect unless you go over the critical threshold. This is equally applicable to the spread of rumours, information, *etc.*

7.4 Classifying Tips

So we've seen two different models that lead to tips: percolation models and diffusion models. Now, we will provide some formal definitions and classification of tipping points.

- **Direct**: When I change one variable it causes the variable to tip. For example, killing of Archduke Franz Ferdinand tipped the system into allegiances and war.

- **Contextual**: Changes in some parameter cause the underlying characteristics of the system to change, which then leads to a tip. In the percolation model, we had a change in the density of the trees, which changed the system and then led to a tip.

One way to classify tipping points is using the phase space diagram of dynamical systems, *i.e.* velocity versus position, where velocity = 0 defines an equilibrium point. BUT if the sign of the velocity to either side of the equilibrium point is such as to cause a movement away from the equilibrium point for a small perturbation, then we have an unstable equilibrium point, *a.k.a.* a tipping point. Vice versa, an attractor is a behaviour that attracts the system, *e.g.* a stable equilibrium point, periodic oscillations or chaos.

Sitting on top of an unstable equilibrium point is a direct tip, a small change in the variable, *i.e.* position, leads to a big movement. Often, we focus on the direct tip (the straw that broke the camel's back, the spark that caused the forest fire, the assassination that started the war), but it is much less the last straw that is to blame but the STATE the system was in to begin with.

Let's say you had a stable equilibrium point and as you change a parameter the stable equilibrium point vanishes. All of a sudden, just by changing the parameters of the system (*e.g.* number of trees, number squares filled in; basically, going over 59.2% in the percolation model, or going over the basic reproduction number in the SIS model), all hell will break loose because a previously stable configuration is no longer stable. This is a contextual tip, brough on by a change in the environement.

Finally, a system can tip within classes (equilibrium, periodic, chaos, complex) or between these classes (from equilibrium to periodic to chaos to complex).

7.5 Measuring Tips

How do we measure tips? We know that there are two types of tips, direct and contextual, where the variable itself causes a tip and where a change in the environment causes a tip. These can also be within a class (like equilibrium, periodic, chaotic and complex) and across these classes.

In an active tip we have a system on top of an unstable equilibrium peak, there is a 50% chance that it will go either way, left or right. But once it has occurred we know 100% that it is either going to be on the right or the left. So in the beginning, there was uncertainty to where it was going to go and then that uncertainty turned into certainty. So we'll measure tipping by the reduction in uncertainty.

Initially there are a whole bunch of outcomes that can occur, but once the tip occurs, there is only one option or a bunch of different options.

Diversity index: Suppose we have 4 options each with a 1/4 chance of coming true (A,B,C,D): $P(A) + P(B) + P(B) + P(D) = 1$. Now we calculate that two of the same outcomes occur $P(A)^2 + P(B)^2 + P(C)^2 + P(D)^2 = \sum_{i=A}^{D} P(i)^2$. So if $P(A) = P(B) = P(C) = P(D) = 1/4$, then $4(.25^2) = 1/4$. The diversity index is then defnined to be $1/\sum(P(i)^2) = 1/.25 = 4$. This means there are 4 different types.

Now if $P(A) = 1/2$, $P(B) = 1/3$, $P(C) = 1/6$, then $\sum P(i)^2 = (1/4)+(1/9)+(1/36) = 14/36 \Rightarrow$ Diversity index $= 36/14 = 2\ 4/7$ which is less then three. So basically there are slightly less than 3 possible things that can happen.

The way we use this metric is to say: originally, there were 2 4/7 places the system could go, and once it's done there is only 1 definite place, *i.e.* it is 100% in the tipped configuration.

Entropy: Entropy $= -\sum P(i) * \log_2 P(i) = -4 * (1/4 \log_2 1/4) = -4 * (1/4 * (-2)) = 2$. What entropy tells us is the number of bits of information we need to know to identify what the outcome is. Example, for 4 options A,B,C,D I can always identify where it'll go depending on two questions. Will it be A or B. If yes, you know. If no, then you ask will it be C or D? So based on two questions and two bits of information you know where the system will go.

Example: Consider two options, A and B, of 50% chance.

- Diversity index $= \frac{1}{(.5)^2+(.5)^2} = 2$
- Entropy $= -2 * (0.5 * \log_2 .5) = 1$ (one question)

What happens after the tip?

- Diversity index: $1/(1) = 1$
- Entropy: $-1 * (1 * \log_2 1) = 0$

So we could say there were 2 options which boiled down to one (diversity index). Or we could say that previously we needed 1 bit of information and then we needed none, *i.e.* it became certain.

So the way to think about tips is a change in the likelihood of outcomes, *i.e.* a decrease or increase in the uncertainty (NOT a kink, a kink doesn't mean it's a tip. It could just be diffusion).

8 Economic Growth

Why are some countries rich and others are poor? We will answer this question using four models:

- Exponential growth: rate at which money grows
- Economic growth using a primitive economy growth model: the model shows that there are limits to economic growth because without innovation growth stops
- Solow's growth model: incorporates innovation and shows that it is important because of multiplier effects
- Extensions: what enables growth to continue over time?

First, let's answer the fundamental question: growth of what? Happiness? That is rather hard to measure. We need an objective quantifiable metric.

Gross domestic product: total market values of all goods and services sold/consumed/ produced within an economy. We also need to account for inflation to measure the real GDP.

The Second World War created a massive bust in GDP and then some nice postwar growth afterwards. After 2008, we had negative growth, *i.e.* contraction. Can China grow forever and become 50x the size of the US? No, look at Japan, which sustained really high growth rates and then it fell of a clip as Japan caught up with more developed nations.

Exponential growth in a bank account just goes up forever. This is the power of compounding. But economic growth is often asymptotic, because the innovations you come up with, depreciate over time and so you converge to an asymptote unless you come up with new stuff.

Does money make you happier? Going from \$20,000 to \$60,000 annual salary makes someone less happy than going from \$1,000 to \$10,000. Lifting people out of poverty is clearly a massive boost to happiness, but towards the tail, it is not as much.

8.1 Exponential Growth

Why are countries rich and some are poor? Let's start with a simple model: exponential growth = compounding. The exponential growth rate will explain why compounding is so important. We can then use the *Rule of 72* to compute what the difference between a 8% and 2% growth rate is.

Compounding: $X = x_0(1+r)^{\text{period}}$ so over a ten year period at $r = 5\%$ growth, the multiplication of money is 1.05^{10}.

Let's do the same for GDP: $GDP = GDP_0(1+r)^{\text{period}}$. So why is the growth rate r so important? Let's look at Table 2, which shows that over a 100 year period a mere 4% difference in growth rate made a HUGE difference. This leads to the general rule known as the *Rule of 72*.

Table 2: Exponential growth over 100 years

Year:	0	1	10	35	100
2%:	1000	1020	1219	2000	7245
6%:	1000	1060	1791	7686	339302

Rule of 72: Divide 72 by the growth rate (in %) and the resulting number will tell you approximately the time (in years) for the compounding quantity to DOUBLE.

- for 2%: $= 72/2 = 36$ years (took 35 years in Table 2 above).
- for 6%: $= 72/6 = 12$ years

Because we have an exponential difference in outcome between different growth rates, economists focus on growth rate so much.

Infinitesimal period: Let's say we break our yearly rate down from years into days, hours and seconds. So essentially the compounding period becomes a smaller and smaller entity.

- 1 year: $X = x_0 \left(1 + r\right)^1$
- 1 day: $X = x_0 \left(1 + \frac{r}{365}\right)^{365}$
- 1 hour: $X = x_0 \left(1 + \frac{r}{365*24}\right)^{365*24}$ *etc.*

in the limit this becomes $X = x_0 \left(1 + \frac{r}{n}\right)^{nt} \equiv x_0 e^{rt} = 2.71^{rt} x_0$

So this formula will give you the rate at which the economy grows. It doesn't grow yearly, but continuously. So this formula shows the exponential, continuous growth explicitly.

8.2 Basic Growth Model

Example: workers pick coconuts, which they can either eat or sell to buy machines, which in turn can pick coconuts faster. But machines of course wear out (depreciate). Let's assume the following variables:

- $L_t = \#$ workers at time t
- $M_t = \#$ machines at time t
- $O_t =$ Output of coconuts at time t
- $E_t = \#$ coconuts consumed at time t
- $I_t = \#$ coconuts invested into machines at time t
- $s =$ coconut savings rate
- $d =$ machine depreciation rate

Assumption 1: output is increasing and concave in labor and machines, *i.e.* $O_t = \sqrt{L_t M_t}$. So this is diminishing return to scale. We asymptotically get less and less from more workers and machines.
Assumption 2: Output is consumed or invested, *i.e.* $O_t = E_t + I_t$ and $I_t = sO_t$.
Assumption 3: Machines depreciate, *i.e.* $M_{t+1} = M_t + I_t - dM_t$.

Let's now simplify. Let $L_t = 100 \to O_t = 10\sqrt{M_t}$. In the real world you would have to create a labour market of people going to work (or not) depending on the incentives present. Also, assume $d = 0.25$ and $s = 0.3$.

Year 1: 4 machines, so $O_t = 10\sqrt{4} = 20$.
Investment: $I_t = 0.3O_t = 0.3 * 20 = 6$.
Eat: 14

Depreciation: $0.25M_t = 0.25 * 4 = 1$.

Year 2 machines $= 4 + 6 - 1 = 9$. Output $= 10\sqrt{9} = 30$.
Investment: $I_t = 0.3O_t = 0.3 * 30 = 9$.
Eat: 21
Depreciation: $0.25M_t = 0.25 * 9 \approx 2$.

Year 3 machines $= 9 + 9 - 2 = 16$. Output $= 10\sqrt{16} = 40$.
Investment: $I_t = 0.3O_t = 0.3 * 40 = 12$.
Eat: 28
Depreciation: $0.25M_t = 0.25 * 16 = 4$.

Year 4 machines $= 16 + 12 - 4 = 24$. Output $= 10\sqrt{24} \approx 50$.

So say we have 400 machines. Then Output $= 10\sqrt{20} = 200$.
Investment: $I_t = 0.3O_t = 0.3 * 200 = 60$.
Depreciation: $0.25M_t = 0.25 * 400 = 100$.

New number of machines is now $400 - 40 = 360$! So we now have less machines and so the economy will shrink because it produces less coconuts. So it seems we have overshot the capabilities of the economy. So there must be a natural level where everything is in balance, *i.e.* in equilibrium.

So what is the equilibrium? At equilibrium the number of machines will stay the same. So equilibrium occurs when investment = depreciation.

Investment $= 0.3 * 10\sqrt{M_t} = 3\sqrt{M_t}$
Depreciation $= 0.25M_t$

So $3\sqrt{M_t} = M_t/4 \Rightarrow 12\sqrt{M_t} - M_t = 0 \Rightarrow M_t = 144$ machines. At 144 machines we have an output of $O_t = 10\sqrt{144} = 120$ coconuts. For $O_t = 120$ we have investment $I_t = 120 * 0.3 = 36$ and depreciation $D_t = 144/4 = 36$. Thus, we are at quilibrium!

So the ironic thing is that the growth model is basically no longer a growth model at equilibrium. This is because our output was concave and depreciation was linear. So to get more growth we need technological innovation so that our output shifts up, maybe no longer a function of $\sqrt{M_t}$.

8.3 Solow Growth

In the previous model growth stopped at equilibrium. It showed that without innovation or an increase in labour, growth will stop as long as depreciation and investment stay the same.

So we shall include one more variable to explain why our economies don't just stop growing. This one more variable is technology.

- L_t = labour at time t
- K_t = Capital at time t
- A_t = Technology at time t
- O_t = Output at time t

So, $O_t = A_t K_t^\beta L_t^{(1-\beta)}$. If $\beta = 0.5$ we have the square root function we had before, $\sqrt{L_t K_t}$. Depending how capital intensive a technology is, β will change accordingly.

So before we had output $O_t = 10\sqrt{M_t}$, investment $I_t = 3\sqrt{M_t}$, depreciation $D_t = M_t/4$, so that equilibrium $0.25M_t = 3\sqrt{M_t} \Rightarrow M_t = 144$.

So what innovation would do is: output $O_t = 10A_t\sqrt{M_t}$. Let's say everything is twice as productive as before with $A_t = 2$, *i.e.* output $O_t = 20\sqrt{M_t}$. Then investment $I_t = 0.3 * 20\sqrt{M_t} = 6\sqrt{M_t}$ and depreciation $D_t = M_t/4$, so the equilibrium is $6\sqrt{M_t} = M_t/4 \Rightarrow M_t = 496$, which produces an output of $O_t = 20\sqrt{496} = 480$, which is significantly higher than 144 previously. In fact, even though productivity doubled, output actually quadrupled. Of course, this is the long run output at equilibrium and not an instantaneous increase.

So with innovation we get a multiplier effect. First, labour and capital become more productive (output higher) and there are incentives to invest more capital (# of machines higher). So the effect is multiplicative because in the output equation the number of machines is multiplied by the technology.

In the basic growth model, growth stops and the equilibrium is driven by the savings rate. In Solow's model, growth can continue because innovation creates a multiplier effect.

8.4 Will China Continue to Grow?

Can China continue their 8%-10% levels of growth? How is this possible in the first place? How do you maintain it? Massive amounts of innovation! Let's see why it is hard to keep up high rates of growth.
Assume: Depreciation = 0.1, Savings rate = 0.2, Output = $100\sqrt{M_t}$ with $M_t = 3600$.
Output = 100*60 = 6000
Investment = $.2 * 100\sqrt{3600}$ = 20*60 = 1200
Depreciation = .1*3600 = 360

So we have 840 new machines = 4440 working machines.
Output = $100 * \sqrt{4400}$ = 6700
Investment = $.2 * 100\sqrt{4400}$ = 20*67 = 1340
Depreciation = .1*4400 = 440

So we have 5340 working machines. Before we had an output of 6000 and now 6700 which is a growth rate of 11%. Continuing:
Output = $100 * \sqrt{5340}$ = 7300
Investment = $.2 * 100\sqrt{5340}$ = 20*73 = 1460
Depreciation = .1*5340 = 530

So we have 6270 working machines. Now we went from 6700 to 7300 which is a growth of 9%. Whats going to happen at 10,000 machines?
Output = $100 * \sqrt{10000}$ = 10,000
Investment = $.2 * 100\sqrt{10,000}$ = 20*100 = 2000
Depreciation = .1*10,000 = 1000

So we have 11,000 working machines. So now output has gone from 10,000 to $100\sqrt{11,000} = 10{,}500$. So this only gives us a growth rate of 5% even though we are adding 1000 new machines. This is the curse of large numbers!

What's going to happen at 22,500 machines?
Output = $100\sqrt{22500}$ = 15,000
Investment = $.2 * 100\sqrt{22,500}$ = 20*150 = 3000
Depreciation = .1*22,500 = 2250

So we have a net of 750 machines with 23250 working machines. Next period our output will be $100\sqrt{23250}$ = 250. This 250 is only a 1-2% increase in output.

We started at 10% growth, growing at a fairly fast clip but then drastically slowed down. This is because of the asymptotic characteristics of output (the square root) and the linear relationship of depreciation. So we need innovation to bump up the growth rate. Which means we need to stop putting money into capital and assets, and rather invest in innovation.

Bottomline: If you have more money than labour you can invest in assets and this will keep growth high. But when the invested capital approaches the state of the technology you need to bump up technology.

8.5 Why Do Some Countries Not Grow?

What's in our model? Labour L_t, capital K_t, technology A_t and output O_t. What is not in here? Culture, inequality. Remember, output the output equals: $O_t = A_t K_t^{\beta} L_t^{(1-\beta)}$.

Growth requires a strong central government to protect capital and investment but that government cannot be controlled by a select group. You need the government to protect people's capital and asset rights, and also their rights if they come up with new technologies. The catch is the government can't be too strong. Often governments are run by a couple of people that can't help themselves (the context is such) that they start to siphon money out of the economy for self-enrichment.

Increases in innovation means that less labor is required. In the short term this creates costly unemployment but in the long term it is beneficial as it sustains growth. So growth requires creative destruction; when you build the tractor you wipe out an entire industry. Equally, tapes replace vinyl, CD replace tapes, and MP3s replace CDs. Each one goes through a wave growth from low to high to low, and the overall consumption from one peak to the next goes up.

The problem is when creative destruction doesn't create new jobs in the new industry, *e.g.* Craigslist replacing newspapers because ad revenues are decreasing. Online technology just needs less labour. The people using Craigslist write the ads themselves. If newspapers had a strong lobbying power with the government they might have prevented classified ads on the internet. So this might have saved loads of jobs BUT it would have prevented creative destruction, new technology and long-term growth.

Can we apply this model in other cases? We can apply it to things like our own production, our own personal GDP. It tells us that we can work hard (assets and capital) but if we don't invest in new skills (technology) we will stagnate. Continued growth requires continued innovation.

8.6 Piketty's Growth Model

The world can be linear, diminishing returns (asymptotic) or accelerating (compounding). Piketty applies compounding to inequality.

Before the Great Depression the richest 10% held 40-50% of the economy in the US, that fell down to 35% after WWII and now it's back up at 50% again. The same is true in Europe, and even in new modern economies like Singapore the same phenomenon is happening.

Piketty says that there is a rate of return on capital, R. And there is a growth rate of the economy, g. Typically the growth rate of capital is ALWAYS greater than the growth rate of the economy. The only time this was different was in the post WWII period.

- growth in wealth = wealth * R
- growth in economy = GDP * g

If $R > g$ then the rate of wealth compounding is greater than that of the economy. Growth in the economy typically trickles down to the well-being of people in general. Growth in wealth is local as it only makes individual people rich. So even if the economy is growing at a decent clip at 2% the rich will get richer.

BUT the point that Picketty misses is: are this top 10% the SAME people, yes or no? Turns out that in capitalism you should have the ability come to riches but also to fall off the pedestal. In this case, yes the rich might be getting richer but they are different people.

So the true story might be a bit more complex. First of all, the rich pay more taxes. Second, they consume more. Third, they donate. Fourth, LUCK and UNLUCK = stupidity rate.

So a more accurate model might be (R - tax - consumption - donation - stupidity) $> g$.

9 Diversity and Innovation

9.1 Problem Solving and Innovation

In this section we'll talk about how individuals and teams go about problem solving. How does diversity play a role in problem solving? How does someone's idea combine with others to create good ideas. Hence, the role of diversity and recombination in problem solving.

General Problem: take an action a, and try to establish a function F that gives the value of that action, $F(a)$. For example, a health care policy and how good that solution is.

To do this, we will use the metaphor of a landscape. The landscape has many local maxima and one global maxima. So if you are stuck on a local maximum, how do we get to the global maximum?

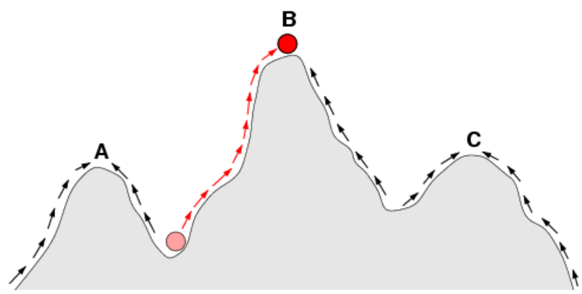


Figure 6: A multi-hill fitness landscape (Photo credit: Wikipedia).

The first part is perspective, which is basically how you pose the problem, *i.e.* encoding. Then out of that encoding you create a landscape. Different perspectives give different landscapes. The second part is heuristics. Heuristics are how you move across the landscape. Climbing up a hill is one heuristic for maximisation, random search is another. The next bit is how individuals interact as a team to get better solutions. Diversity basically gives a broader perspective, *i.e.* a different landscape, but also different heuristics and therefore you have a greater chance of hitting upon a global maximum.

Teams also allow for recombination, by combining the solution to one problem with a solution to another problem to get a better solution. A big driver to innovation is often the recombination of different solutions to smaller subproblems.

Perspectives → Heuristics → Diversity → Recombination

9.2 Perspectives and Innovation

Perspective is basically how you represent a problem—the model you have in your head. It's the design landscape on which you are trying to find a maximum—the optimum solution. We need to formalise this landscape though (a model), in order to maximise.

First, we formally define what a perspective is. A perspective is a representation of the set of all possible solutions. Once we have the encoding of this set of solutions, we can give a value to each of these solutions and find a maximum. One way to think about perspectives is coordinate systems in terms of base vectors. You can use polar coordinates to represent points in space or use Cartesian coordinates instead. In both cases you are representing the same quantity (a point in space) but sometimes one method is easier or superior over the other. For example, a straight line is easier described in Cartesian coordinates but an arc is easier described in polar coordinates.

So we can appreciate that by changing the perspective of a problem we can come up with new solutions. For example, by representing the elements in terms of molecular weight, Mendeleyev realised that there was a lot of structure to chemical elements. On the other hand, an alphabetic representation wouldn't get us a lot of structure. In fact, Mendeleyev had holes in his periodic table at the time, but these gaps were then filled in over time. In fact, they actually provided little hints for what elements to look for.

Another example, how do you organise a bunch of applicants. You can do that with GPA, work ethic or creativity. So depending of what you are looking for, each one of these might be fine. If we go back to the landscape metaphor and assume it is rugged with a bunch of peaks, then this landscape will have many local optima (a peak). This means that if you are blindly hill climbing you can get stuck on local optima. So a good perspective is actually one that only yields one global optimum (or at least very few local optima). This is convexity, and means that you can't get stuck on local optima.

These best perspectives in terms of framing an optimisation problem are called *Mt Fuji* perspectives, because Mt Fuji in Japan looks like one big pyramid (a single peak). An example is the shovel landscape, *i.e.* optimising the size of a shovel. How many pounds of coal can I shovel as a function of the shovel's size? Zero dimensions essentially means shovel-stick only, and we can't shovel anything. Then as we go from spoon to normal shovel, we can shovel more and more. But as we make the shovel bigger beyond the optimum shovel size, we start to have problems lifting the shovel, which means the amount which we can shovel goes down until eventually we can't shovel anything anymore (massive shovel). This makes hill climbing super easy to find the optimum.

Another example: *Sum to Fifteen*. Cards numbered 1-9 face-up on a table. There are two players which take turns in selecting cards. You win if you hold exactly three cards that sum to fifteen. This game is the same as *Tic Tac Toe*. But you will only see this analogy if you place the cards into a magic square and not in a straight line. Arranged in a straight line you will never appreciate that *Sum to Fifteen* and *Tic Tac Toe* are basically the same game.

This is the **Savant Existence Theorem**: For any problem there exists a perspective that creates a Mt Fuji landscape (in fact many do). The reason this is true is that if you put the very best solution in the middle and the very worst ones at the ends and line up the ends, you can create a Mt Fuji. This is of course with hindsight when you already have the solution, but the point is that it is possible. The flip side is that there are a ton of bad alternatives.

So this is what insightful people do: they find ways to frame problems, *i.e.* provide perspectives that lead to optimal solutions. This is very much *negative entropy* thinking because a random perspective isn't going to be very helpful, but if you apply some clever thinking (energy) you can create some beneficial structure out of the whole thing. In fact, the best perspective encapsulates a good deal of information, hence the *negative entropy* metaphor.

So the underlying takeaway is: if stuck and with no solution, try to frame the problem in a different perspective. Represent reality in a different way. Of course, this is not easy, otherwise everyone would be doing it.

9.3 Heuristics

How you go about finding solutions to a problem once you have framed it in a specific perspective?

- One is **hill climbing**. Just go up in a positive slope direction. If you reach a point where all directions are down you have reached a local optimum.
- Another is **do the opposite**. Think of what the existing solution is and do the opposite. For example, take Priceline: here is how much I want to pay to stay at your hotel, rather than the hotel telling you how much you will have to pay. The contrarian approach.
- Another one is **big rocks first**. Suppose you have to take the following test: Put a bunch of rocks in a bucket. If you start with the little rocks first, then the big rocks won't fit in nicely afterwards. If you put the big rocks in first, then you can arrange the small ones in after. So, big rocks first! Successful people know how to put the big, important things first. Deal with the important things first.

No free lunch theorem: All algorithms that search the same number of solutions with the goal of locating the maximum value of a function defined on a finite set, they all perform exactly the same when averaged over all possible functions. Basically, that means that no heuristic is the "god heuristic" that is best all the time. All heuristics have particular best use cases. For example, the problems you see as a person or in management, the "big rocks first" heuristic is a good algorithm to use. Put another way, if we know nothing about the problem then no perspective is better than another.

The bottom line is, unless you know something about the problem being solved, no algorithm or heuristic performs better than any other. You are just groping in the dark if you know nothing about the problem. You basically have to define your landscape properly first. For example, there are some things where little rocks first is better, earthquake proofing walls of the Inca's for example.

Figure 7: A earthquake proofing wall in Cusco, Peru. Note the small rocks at the bottom of the wall.

One insight from the free lunch theorem is that you can combine different heuristics together and perhaps come up with a better heuristic than either heuristic on its own. For example, search

North, West, East and South individually, or search North-East, North-West, South-East, South-West, and combine the two to search in 8 rather than 4 directions.

So given that two people may construct two different perspectives of the same problem with different peaks and troughs, different heuristics might work better for these two perspectives/people. So heuristics are perspective dependent and we need to look at both to know what the best combination is.

Diverse perspective + diverse heuristics lead teams to arrive at better solutions.

9.4 Teams and Problem Solving

When you solve a problem, you first come up with a perspective of how to solve the problem and then you use heuristics to search among the possible solutions given by the perspective. The heuristics allow you to find the best solutions.

Now we will combine perspectives and heuristics to show that teams are best at finding the best solutions. It doesn't need to be a team at one instance in time, but can be a series of incremental solutions over time.

If we want to choose the best candy bar, we can have a masticity landscape and a caloric landscape. Let's say the caloric landscape produces peaks A,B,C (random characteristics) and the masticity perspective produces peaks A,B,D,E,F. The caloric landscape is better because it has fewer peaks (closer to a Mt Fuji). So one way of checking how good you are at solving a problem is to check how many possible optima your combination of perspective and heuristic gives. Let's now incorporate the "value" of those peaks

A	B	C	D	E	F
10	8	6	6	2	4

We can now ask what is the average value of a peak in the caloric landscape? Average of (ABC) is average of 10, 8, and 6, which is 8. For masticity: average of (ABDEF) is average of 10, 8, 6, 4 and 2 which is 6. So again, caloric is a better landscape.

But now let's look at what a team would do. The intersection of the two landscapes in terms of peaks is A and B. So if the first person using the caloric landscape hill climbs and gets stuck at B, then the masticity person won't be able to help. But if the caloric person gets stuck at C, then the masticity person will be able to say no choose A or B instead. So now they can only get stuck at the intersection A,B which has an average value of 9. So the more landscapes and heuristics we have, the more likely we are to find the global optimum. An argument for a multi-model thinker!

So this explains why over time we find better and better solutions. Part of it is that we are getting smarter and smarter, but part of it is also that we are looking at the problem with new landscapes and heuristics.

Big claim: team can only get stuck on a solution that is a local optimum for every member of the team. That means that the team has to be better than the people in it. So what we want for effective problem solving is loads of different people with different perspectives and different heuristics, *i.e.* diversity.

So what's missing? This seems highly stylised. The first one is communication. We have assumed that we have perfect communication. There are typically problems with communication and ego that get in the way. The other thing we have assumed is no possibility of error in evaluating solutions.

9.5 Recombination

Recombination means combining different perspectives and heuristics to create ever more and better perspectives and heuristics. This is the avenue to more and more innovation.

IQ test: Fill the series 1 2 3 5 _ 13. Answer: 8. Or fill the series 1 4 _ 9 16 25. Answer: 9. Or harder ones, fill 1 2 6 _ 1806. This last one looks much harder but it is actually just a combination of the last two, squaring and subtracting. (2-1)=1^2, (6-2) = 2^2, (42-6)= 6^2 and (1806-42) = 42^2.

The beautiful thing about combinations is that they create exponential possibility for new ideas. How many ways to pick three objects from 10 different ones? $\frac{10!}{3!7!} = 10 * 9 * 8/6 = 120$. So if I have 10 ideas then I have 120 combinations of picking three of them. If I have a deck of cards with 52 cards and I just want to select 20 of them $\frac{52!}{20!32!} = 126$ trillion! So every time someone has an idea we have a huge explosion of possible combinations.

This is the basis for **recombinant growth**. For example, take a car. It is a recombination of a number of previous solutions to other problems.

Exaptation: Birds primarily developed feathers to keep warm. But then feathers got exapted for flying. They were eventually used in another context. So once something is developed, it gets used for a bunch of other things even if it wasn't meant for that purpose.

But of course to get exaptation and recombinant ideas we needed communication. This is where the university was great because we developed a place where all these ideas could be shared and recombined. This is also why Steve Jobs wanted people to bump into each other in corridors.

So innovation comes from diversity (perspectives and heuristics) to create new ideas and then recombination to combine them in novel ways.

10 Markov Processes

The premise of Markov models is really simple. A system can be in a number of different states, and entities that comprise this system move between these states according to certain transition probabilities. Markov models now compute how likely it is for something to move from one state to another.

Example: students in a lecture can be alert or bored. Assume that they move from being alert to being bored with probability p and from bored to alert with probability q. A Markov model will tell us how this system of bored and alert students evolves over time.

There is a **Markov Convergence theorem** which states that as long as a couple of very mild assumptions hold true (finite number of states, transition probabilities don't change and all states can be accessed from all other states) then the system will tend towards an equilibrium over time.

The reason why Markov processes are a useful way of thinking about the world is that they can provide a lot of understanding from a very basic model.

10.1 A Simple Markov Model

Let's work through the simple transition example from alert to bored students and vice versa. To begin, we need to make some assumptions. As you give the students a verbal cue 20% of alert students become bored, while 25% of bored students become alert.

Let's do the calculations by hand first. Say we have 100 alert students and 0 bored students. This means we move to 80 alert and 20 bored after the verbal cue. Let's say we can repeat this over and over again without the probabilities changing. Then after another cue of the 80 alert students 16 will become bored, and of the 20 bored 5 will become alert. So we end up with 69 alert and 31 bored. Repeating this over and over again is too cumbersome.

An easier way is to use the Markov Transition Matrix:

Probabilities	Alert (t)	Bored (t)
Alert (t+1)	0.8	0.25
Bored (t+1)	0.2	0.75

If I multiply this matrix by a vector $[1; 0]$ I am basically calculating the probability of a population of all alert students becoming bored/alert.

$$\begin{bmatrix} 0.8 & 0.25 \\ 0.2 & 0.75 \end{bmatrix} \cdot \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} = \begin{Bmatrix} 0.8 \\ 0.2 \end{Bmatrix} \tag{12}$$

This gives us $[0.8; 0.2]$. I can take this vector now and multiply it again by the matrix and get $[0.69; 0.31]$. In the next period we get $[0.63, 0.37]$. If we do this *ad infinitum* where do we end up? Well if we turn this around and start with all bored students then the starting vector is $[0; 1]$, which gives us $[0.25, 0.75]$. Taking this vector and multiplying again we get $[.45, .55]$. Then $[.5, .5]$. Then $[.53, .47]$. So it sort of looks like it goes to some equilibrium, and that there is convergence. Note that the transition probabilities are assumed to stay constant.

So how do we find this equilibrium? In general, our vector is $[p; 1-p]$. So after I multiply and calculate the process I want to get the same vector $[p; 1-p]$ again. Hence,

$$\begin{bmatrix} 0.8 & 0.25 \\ 0.2 & 0.75 \end{bmatrix} \cdot \begin{Bmatrix} p \\ 1-p \end{Bmatrix} = \begin{Bmatrix} p \\ 1-p \end{Bmatrix} \tag{13}$$

In linear algebra, this type of problem $\boldsymbol{A}\vec{x} = \lambda\vec{x}$ is called an eigenvalue problem, where $\vec{x}$ is the eigenvector of the matrix $\boldsymbol{A}$ with associated eigenvalue λ. For this easy 2×2 problem we can solve the problem using basic Algebra (for larger problems this is harder and an eigenproblem solver is the way to go):

$$\begin{aligned} .8p + .25(1-p) &= p \\ 16p + 5 - 5p &= 20p \\ p &= \frac{5}{9} \approx 0.556 \end{aligned}$$

Or equally

$$\begin{aligned} .2p + .75(1-p) &= 1-p \\ 4p + 15 - 15p &= 20 - 20p \\ p &= \frac{5}{9} \approx 0.556 \end{aligned}$$

So, $p = 0.556$ is the equilibrium state where 55.6% of the people are alert. Now, this is not a stationary equilibrium. People are still churning, moving back and forth between bored and alert, but the proportion of people alert *vs* people bored does not change. This is why this is called a statistical equilibrium point.

10.2 Markov Democritization

A little reminder: In Markov models we have predefined states, and certain entities move between those states according to constant transition probabilities. We can apply this model to democracy because countries (entities) can be free (first state), not free (second state) or partly free (third state). Let's start with a simple model to start.

Democracies and non-democracies Assumption: Each decade 5% of democracies become dictatorships and 20% of dictatorships become democracies.

Markov Transition Matrix

$$\begin{bmatrix} 0.95 & 0.2 \\ 0.05 & 0.8 \end{bmatrix} \tag{14}$$

The equilibrium state is given by $0.95p + 0.2(1-p) = p$. So $95p + 20 - 20p = 100p \Rightarrow p = 4/5 = 0.8$. So the surprising thing is that even though 95% of democracies stay democracies, we only get 80% democracies in the long run. This is another example of micro and macro differences (just like Schelling's model for segregation), in that the micro (95%) does not necessarily drive the macro (80%).

Free, not free, partly free Based on data, countries can be free, partly free, or not free. Each decade 5% of free and 15% of not free countries become partly free, 5% of not free and 10% of partly free countries become free, and 10% of partly free countries become not free.

Markov Transition Matrix

So the equilibrium is given by:

$$\begin{bmatrix} 0.95 & 0.1 & 0.05 \\ 0.05 & 0.8 & 0.15 \\ 0 & 0.1 & 0.8 \end{bmatrix} \begin{Bmatrix} p \\ q \\ 1-p-q \end{Bmatrix} = \begin{Bmatrix} p \\ q \\ 1-p-q \end{Bmatrix} \tag{15}$$

	Free (t)	Partly Free (t)	Not Free (t)
Free (t+1)	0.95	0.1	0.05
Partly Free (t+1)	0.05	0.8	0.15
Not Free (t+1)	0	0.1	0.8

So, $0.95p + 0.1q + 0.05(1 - p - q) = p$ and $0.05p + 0.8q + 0.15(1 - p - q) = q$ which gives p = 62.5% and q = 25%. Which means 62.5% remain free, 25% are partly free and 12.5% are not free. So again, even though 90% of democracies remain free locally, this means that if the transition probabilities remain constant, then we end up with only 62.5% countries free in the long run. Unless the transition probabilities change, we are unlikely to get close to 100% free countries.

10.3 Markov Convergence Theorem

So why is it that Markov models converge to an equilibrium if the transition probabilities stay the same? Remember this is a statistical equilibrium, *i.e.* the proportion of entities within states stays the same, but there is churn within the system.

Let's take a Markov transition matrix:

$$\begin{bmatrix} 0.8 & 0.25 \\ 0.2 & 0.75 \end{bmatrix} \tag{16}$$

Which means we have $0.8p + 0.25(1 - p) = p \Rightarrow p = 5/9$ for equilibrium.

The underlying assumptions of Markov processes are:

1. Finite number of states
2. Fixed transition probabilities
3. Can eventually get from any state to another. Maybe not right away, but there must be some way of transitioning between any two states.
4. Not a simple cycle, *i.e.* it doesn't just periodically cycle between two or three, *etc.* states.

Markov convergence states that, given these four assumptions, a Markov process converges to an equilibrium distribution which is unique. This means there is one and only one unique equilibrium.

The first thing this implies is that the **initial conditions don't matter**, *i.e.* no sensitivity to initial conditions. **History also doesn't matter**. If it's a Markov process we are going to that one unique equilibrium over time. Also **intervening to change the states doesn't matter** unless you are fundamentally changing the transition probabilities. So the Markov process is very deterministic because small changes basically don't make a difference. Equilibria in a Markov process are very strong attractors.

Does this mean that we should not have redistribution policies? Even though the Markov process says initial conditions and history don't matter, it could take a long time to reach the equilibrium. The only way interventions and history can matter in a Markov process is when the transition probabilities change. Either actively by interventions or naturally over a finite time span. So we need to focus on policies that change the transition probabilities.

So the bottom line is, changing states is temporary if the underlying transition probabilities remain the same. We will eventually always reach the same equilibrium. Changing transition probabilities can be more permanent and so this is where we should focus energy if we want to change the system.

10.4 Exapting the Markov Process

How can we use the Markov model in contexts we have never thought of? A Markov process is: fixed states with fixed probabilities of transitioning between states.

Simple, standard applications

Voter Turnout Take a set of voters V at time t, and a set of non-voters N. Taking the eigenvector of the Markov transition matrix allows us to find the percentage of non-voters and voters at equilibrium. Of course this is a statistical equilibrium because there will still be static churn at equilibrium. So this method won't tell us who votes what.

School enrolment Kids that go to school and kids that don't go to school. Again just assign transition probabilities and populate the matrix, and calculate the equilibrium.

Outside the box application Given a transition matrix:

	A(t)	B(t)
A(t+1)	0.8	0.25
B(t+1)	0.2	0.75

If we think about this matrix, then we can see that each column represents the likelihood that, given a certain state, we will either stay in that state or transition into another. What else can we use this for?

Identify Writers Figure out who wrote a book. Take a book by an anonymous author and find the percentage of cases where the word "for" is followed by "the record", "example", "the sake of" and compute a transition matrix. You then compare that transition matrix against a predefined transition matrix of authors you have in a database.

Medical Diagnoses Given someone has received treatment for a disease, there are typically a series of reactions. If the treatment is successful you may go through the following: pain, depression, pain, success. If not successful it might look like this: depression, mild pain, no pain, failure. What this means is that if I give someone a treatment and they show a specific reaction, then I can assign probabilities if the treatment is working or not.

Lead up to a war Suppose two countries are in a state of tension leading to embargoes, which then leads to war. You can now ask, historically, given these transitions, what is the likelihood that we will see a war? By gathering a bunch of transition data you could at least put a figure on the situation if the current state looks like it will lead to war.

11 Lyapunov Functions

Lyapunov functions map models to outcomes in the following way: if a Lyapunov function can be constructed to describe a system/model, then, by definition, that system/model goes to equilibrium.

In general, systems can be in equilibrium, oscillate periodically, be chaotic, or exhibit complex behaviour. If we come up with a Lyapunov function then we know for sure that the system goes to equilibrium and we can also know how quickly the system converges to that equilibrium state.

Physics: Let's say we have a system for which each change in speed necessarily means that the speed decreases. Hence, at some point that system must reach the zero-speed mark and come to rest.

Economics: Let's say people trade, and every time they trade, the overall happiness-level in the system goes up. Of course there is also a maximum level of happiness. So each time people trade, happiness goes up and at some point it reaches a ceiling, at which point happiness is in equilibrium.

Formal: $F(x)$ is a Lyapunov function if:

- Assumption 1: $F(x)$ has a maximum value
- Assumption 2: There exists a $k > 0$ such that if $x_{t+1} \neq x_t$ then $F(x_{t+1}) > F(x_t) + k$. If both these are true, then at some point $x_{t+1} = x_t$.

This is related to **Zeno's Paradox**: Say I want to leave a room. The first day I go half the way, the next day another half of the way, then another half of the way. Then I am basically going 1/2, 1/4, 1/8 *etc.* of the distance. This seems to imply that I am converging on leaving the room but never really leaving it. Thus, the paradox is that you can continually half the distance between a point and yourself, but never really get there. The way to get around Zeno's paradox is to say that when you move you move by a value of at least k, where k is constant and finite. This means that our movements are now defined by a Lyapunov function and we will, by definition, reach our destination.

There is also a bonus to Lyapunov functions. We can figure out how fast the process will reach the equilibrium state, because this rate of convergence depends on the minimum increment size k. The hard part, however, is figuring out the Lyapunov function $F(x)$ itself. Once you have the Lyapunov function it is relatively easy to do the calculations. Sometimes getting the function is straightforward, but oftentimes it is very hard.

Summary of Lyapunov process: A process has a maximum/minimum, and every time the system moves, it moves towards the maximum/minimum by a finite step. Then at some point the system must reach an equilibrium state.

11.1 Organisation of Cities

Puzzle: Why do cities organise with no central planning? For example, there are no massive queues at restaurants, no massive queues at coffee shops, no (or rare) break down of traffic. How does this work without central planning? How are the right number of people at the right places at the right time?

Basically, how does self-organisation work? Suppose there are five things you have to do during the week: cleaners, C; grocery, G; deli, D; book store, B; fish market, F. And you can choose which day to go. There are five days *Mon-Fri* that you can go to one of these places; each on one of the five days. For example, the order might be (C, G, D, B, F). Suppose there are a number of people that have to hit the same five places in five days. Ideally, you want to select your choices so that there is relatively little going on/avoid crowded places.

So if people switch two locations so as to avoid crowds, we can fit a Lyapunov function and show that the system goes to an equilibrium. If I go to the cleaners on Monday and it's always crowded, then I just switch with something else.

Assume the following sets: (C, G, D, B, F); (G, C, D, B, F); (C, D, G, F, B); (C, B, F, G, D); and (C, F, D, B, G). Let's try to formulate a Lyapunov function for this process:

Attempt 1: Total number of people at each location. This doesn't work because 5 people go to each location.

Attempt 2: Total of number of people that the five people meet each week. For example, for the first person: $3 + 0 + 2 + 2 + 1 = 8$. Suppose first person switches C and F. Then we get $0+0+2+2+0 = 4$. What about the other people? Well, the three people at C on Monday won't be seeing person 1 anymore, so there is a total $4 * 2 = 8$ less meetings.

Is Attempt 2 a Lyapunov function? Does it have a minimum? Yes, 0. If I move to meet less people does the total number of meetings go down? Yes, because if I meet less people then they also don't meet me. So the total number of meetings has to drop. In this example, the drop k is actually easy because if I meet one less person, then that person is not meeting me and so the total k is 2. So what we get in the end is that noone is running into anyone else during the week.

11.2 Exchange Economies and Negative Externalities

If we can construct a Lyapunov function, then the system will definitely go to equilibrium. If we can't come up with one, then we don't know. It might, but it might not.

Exchange Markets Places where people bring things to trade. Is that system going to an equilibrium, or are people just going to be trading in a mess?

Assumption 1: Each person brings a wagon full of stuff.

Assumption 2: People trade with others but only if each gets an increase in happiness by some amount k.

What are possible Lyapunov functions?

Attempt 1: Total happiness of the people. Does this satisfy the two Lyapunov function requirements (maximum and increase towards maximum with each move)? Yes, there needs to be a maximum total happiness level, and yes, because each trade was assumed to make both sides happy. So by definition each trade will make everyone happier. So yes, the system goes to an equilibrium.

Where can't we fit a Lyapunov function? Negative externalities! For example, consider this fictitious scenario:
North Korea: Trades nuclear weapons for oil.
Iraq: Trades oil for nuclear weapons.
USA: Not involved.

North Korea: happier. Iran: happier. USA: less happy. So total happiness didn't necessarily go up here because France, UK, Germany also won't be happier by this trade arrangement. So we can't put a Lyapunov function on this process because some countries might not be involved in the North

Korea-Iraq trade, but still face negative externalities. Note, that if the externalities were positive, then we could define a Lyapunov function. Other examples where this occurs: political coalitions, mergers, alliances.

11.3 Time to Convergence and Optimality

Let's clean up two details:

1. Can we say how long it will take for the process to go to equilibrium?
2. Does the process always stop at the maximum or the minimum?

Lyapunov function definition: if the process stops, then the process is at a maximum/minimum. If it doesn't stop then it needs to take a step towards the maximum/minimum.

1. Suppose $F(x_1) = 100$, $k = 2$ and max $= 200$. Then the number of periods has to be less then or equal to 50. This is because the process might go to equilibrium in one step, but it will certainly go there in 50 steps, because each step needs to go at least a distance of 2. So the closer the state is to the maximum and the larger k is, the tighter the bound we can put on how long it will take the process to reach equilibrium. (k is big and maximum/minimum is small).
2. Short answer: No. Generally, the process does not stop at the global maximum because the space can be rugged with multiple local optima. Let's go back to the preference model:

   person 1 prefers: $A > B > C$

   person 2 prefers: $B > C > A$

   person 3 prefers: $C > A > B$

   Suppose we are in a situation where person 1 has B, person 2 has C and person 3 has A, and they can trade. Can they make everyone happier by trading? Person 1 wants person 3's A, but person 3 doesn't want person 1's B, so person 3 says no. Hence, person 1 can't make a trade. Person 2 wants person 1's B but person 1 doesn't want person 2's C and so no trade. Person 3 wants person 2's C but person 2 doesn't want person 3's A. So again no trade. So we have a situation that is supoptimal because we can't get to the global optimum by pairwise trades. Hence, it is possible to define a Lyapunov function for a process and stop at a suboptimal point.

11.4 Fun and Deep

There remains a lingering question: If we have a system, can we make sure that the system goes to equilibrium? We will do this in a fun way and in a deep way.

Fun: Chairs and Offices Assume a company is moving to a new office. How does it allocate chairs and offices? To allocate the chairs, let's just randomly assign chairs and then people can trade if need be. This is a good idea because people won't be trading all day. Let the Lyapunov function be the happiness of owning a chair, and assume that there is a maximum level of happiness. If we assume some small cost of trading, then the total happiness goes up after each trade.

Sounds like trading is a good idea, so let's also trade for offices. In this case trading is a terrible idea. Why? Offices are different than chairs because of negative externalities. You have to account for the people next to the offices that are being traded because these people are affected as well by trading and moving.

Deep: Can we always know if a process stops? There is a famous problem known as the *Collatz problem*, summarised by "half or three plus one". The problem goes like this: pick a number, if even divide by two, if it is odd then times by 3 and plus one. Stop if you reach one.

Example: Let's start with 19, go to $3 * 19 + 1 \rightarrow 58/2 \rightarrow 29$ *etc.*
Let's start with 5: $3 * 5 + 1 \rightarrow 16/2 \rightarrow 8/2 \rightarrow 4/2 \rightarrow 2/1 \rightarrow 1$.
Let's start with 7: $3*7+1 \rightarrow 22/2 \rightarrow 11*3+1 \rightarrow 34/2 \rightarrow 17*3+1 \rightarrow 52/2 \rightarrow 26/2 \rightarrow 13*3+1 \rightarrow 40/2 \rightarrow 20/2 \rightarrow 10/2 \rightarrow 5*3+1 \rightarrow 16/2 \rightarrow 8/2 \rightarrow 4/2 \rightarrow 2/2 \rightarrow 1$.

If we start with 27 this process takes aaaages. The point being, that for some things we can figure relatively quickly if there is an equilibrium and we can fit a Lyapunov function, but for others it is much more complicated to figure it out.

11.5 Lyapunov or Markov

Let's contrast Lyapunov functions with Markov functions because both go to equilibria. So what are the differences?

Lyapunov: $F(x)$ needs to have a maximum or minimum, and if it changes state x then it needs to get closer to the maximum or minimum by a value k.

Markov: finite states, fixed transition probabilities, can get from any state to another and it's not a simple cycle between two or multiple states. Given those assumptions, the system goes to a stochastic unique equilibrium. There can still be churn in the system, but the proportion of states remains the same.

These two processes are fundamentally different though because a Markov process is path (history) independent. It doesn't matter what the initial conditions are, because when the transition probabilities are fixed we will always end up in the same place. A Lyapunov function *is* path dependent. The equilibrium you end up on depends on where you started. This implies that there can't be just one unique equilibrium. In a Lyapunov process it's also a fixed point where the system stops, it's not a stochastic process as in the Markov process where the system continues to churn.

What have we learned about Lyapunov functions?

1. If you can construct a Lyapunov function then the system goes to equilibrium. Trading chairs in an office, we could construct a Lyapunov function. But for trading offices we couldn't because of negative externalities.

2. You can compute the maximum number of cycles to achieve the equilibrium state (you can bound the time).

3. The equilibrium need not be unique or efficient.

4. The reason a system doesn't go to an equilibrium and the reason you can't construct a Lyapunov function is because there are negative externalities which influence the system in the opposite direction (it's not all benefit). The reason we can't decide if the Collatz problem (divide evens by two and multiply odds by three and add one) goes to an equilibrium because "multiply by three and add one" is a "negative externality" that points in the opposite direction of "divide by two".

12 Coordination and Culture

12.1 What Is Culture and Why Do We Care?

Culture itself has hundreds of definitions. Culture is such a complex topic that a simple definition cannot capture its entirety. So let's look at a couple of definitions so we can look at how the definition of culture has changed over time. In fact, culture matters a lot for the survival of nations and economies.

Tyler (1871): The complex whole which includes knowledge, belief, art, law, morals and customs. Complex entirety of human existence which varies across different countries.

Boaz (1911): The totality of mental and physical reactions and activities that characterise behavioural responses to environment, others, and to himself/herself. There needs to be some consistency to mental and physical reactions of specific people across the globe.

Trilling (1955): When we look at people to the degree of abstraction which the idea of culture implies, we cannot but be touched and impressed by what we see, we cannot help but be awed by something mysterious at work, some creative power that seems to transcend any particular act or habit or quality that may be observed. To make *a coherent life*, to confront the terrors of the outer and the inner world, to establish the ritual and art, *the pieties and duties which make possible the life of the group and the individual*—these are culture, and to contemplate these various enterprises which constitute a culture is inevitably moving.

In order for different groups to functions they need a unifying culture that creates implicit trust, norms and mental shortcuts about how others will behave. And of course not all groups do this in the same way.

Differences: The Ultimatum Game
Player 1: offers player 2 an equal split of $10
Player 2: has the ability to accept or reject the offer
If player 2 accepts: both players get the split
If player 2 rejects: both get $0

What is the minimum split that player 1 needs to offer so that both get something? And do different people from different cultures play the game in the same way? No, consider these two examples. Lamalera: Indonesian whale hunters—offered $5.70. Machigenga: Amazonians who lack personal names—offered $2.60

Another way we can measure cultural differences is with survey data. There are two axes that seem to express most differences:

1. survival values *vs* self-expression values

2. traditional values *vs* secular rational values

In this manner, the world splits up quite nicely into Protestant Europe, English-speaking world, Islamic, Catholic Europe, Confucian, *etc.* (see Figure 8 below).

But of course in these geographic areas not all people are the same. In fact, the differences within geographic areas are often greater than differences between geopgraphic areas. In biology, you could say "there is no great blue heron", because what we mean by the "Great Blue Heron" is an entire population of herons, but there are still going to be very specific genetic, behavioural, *etc.* differences between individual herons.

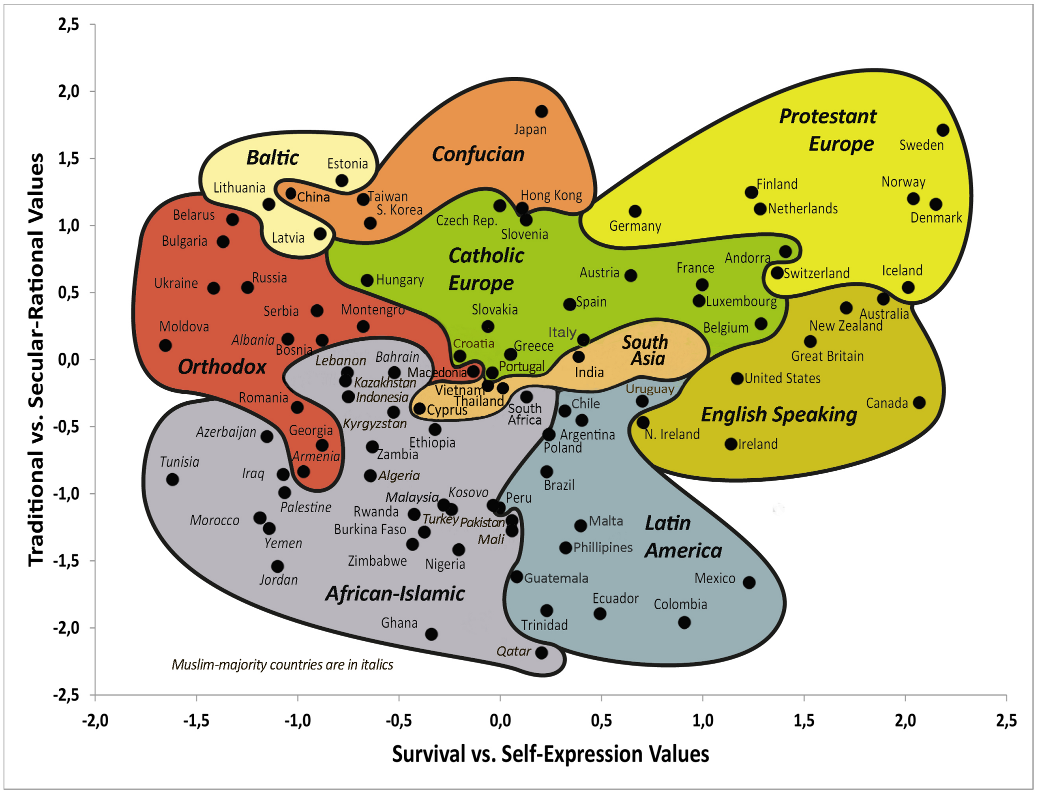


Figure 8: Inglehart-Welzel cultural map.

There are other ways to characterise culture:

- power distance/inequality
- uncertainty avoidance/tolerance
- individualism/collectivism
- masculinity/femininity
- confucianism/dynamism

For example, the US is low on power distance (high inequality), very high on individualism, medium on masculinity and low on uncertainty avoidance. France is higher on power distance, lower on individualism, lower on masculinity and much higher on uncertainty avoidance.

But of course you can't capture everything in culture just using 5 metrics, which is precisely why there are so many definitions of culture. For example, El Salvador and South Korea will categorise almost identically using the 5 categories above, but anybody who has been to these two countries will know that El Salvador and South Korea are vastly different.

Why do we care? The reason we care is that:

> "Virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time. It can be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence." — Kenneth Arrow

So a lot of economic backwardness can be explained by a lack of trust, and trust is a cultural dimension.

The problem with this statement is that social capital and trust have to be "measurable" to be scientific, they cannot just be buzzwords. One way to measure trust is using surveys, *e.g.* do you claim unentitled benefits, do you pay public transport, do you pay your taxes, do you keep money you found? Or generally speaking, would you say people can be trusted? With this last question the percentage is 70% in Sweden, 33% in Italy and 10% in Turkey. But the most important thing here is to determine if this is a causal relationship, does trust create more wealth? Because it might be that once you have a high GDP you might start trusting instead, *i.e.* the causal relationship is the other way around.

12.2 Pure Coordination Game

To make sense of cultural difference we will construct a very simple model called the pure coordination model. Let's start with some examples:

1. Ketchup question: where do you store your ketchup? In the fridge or in the cupboard? Most people in the US store it in the fridge, but most people in the UK store it in the cupboard. This is a coordination game, because you put the ketchup where the people sharing your living space (siblings, mother, father) will find it too. So in fact, it is probably handed down to you from your parents, and is therefore cultural.

2. Electric plugs: another coordination game—everyone in one country wants to coordinate to one plug, but there may be differences across countries.

3. Driving on the left and right: there is an incentive to drive on the same side of the road throughout the world, so that when you go to a different country you don't cause car crashes. It is also easier to produce cars with steering on one side only. Again, we have coordination, but this time cross-country. The countries that tend to drive on the left are islands (Britain, Japan) and they have fewer boarders with other countries that could cause coordination issues.

So we want to take this coordination theme and turn it into a game. The **Pure Coordination Game**: We have two people, Player 1 and Player 2, and each one can put the ketchup in the fridge or in the cupboard. If they put the ketchup in the same place, then both players get a payoff of 1, but if they put it in different places then they get no payoff.

Now let's increase the complexity of the game to N number of people, which can choose between two different options. If a person finds himself/herself surrounded by people of the opposite ketchup-choice, then he/she will most likely switch to coordinate or purely to fit in. So people change their behaviour because they want to coordinate/match (similar to segregation driven by preferences). So does this process actually lead to a specific outcome without churning (statistical equilibrium)?

We can answer this with a Lyapunov function. Remember a Lyapunov function is a function $F(x)$ with a minimum or maximum, which converges closer to the minimum or maximum by a fixed amount every time it moves. This process then stops at some point, when the function reaches the minimum or maximum. So if the Lyapunov function is the number of coordinations, then every time someone switches preference when he/she is surrounded by a different ketchup-preference, then the number of coordinations goes up by 2. The process doesn't have to stop with everyone having the same preference, instead you can get segregation.

So the Lyapunov function tells us that depending on the social structure, either everyone will do the same thing or we will get segregated blocks.

Is there a difference between **Coordination or standing ovation?** The difference is that in the coordination game there is a measurable difference in payoffs—no one would choose not to coordinate, whereas in the standing ovation model there could be more "psychological" incentives of "it's ok to differ". So in the coordination model there is a huge physical cost to be different, whereas in the standing ovation model, the cost is mostly psychological.

Inefficient coordination We can also observe many behavioural patterns that are interesting but do not seem to be optimal. For example, consider the **Maui - Des Moines Game**. We have a nice beautiful beach destination (Maui) and a nice city in Iowa (Des Moines). But Des Moines can probably be considered to be not as nice as Maui as a tourist destination. Suppose you and your partner win a vacation, but you need to go to the same place. You go to the airport but you can't remember which place to go. If you both go to Maui, then you get a payoff of 2. Whereas if you both go to Des Moines, then you get a payoff of 1. If you go to different places you get no payoff. So obviously both of you will go to Maui.

BUT this is not always the case. For example, we can replace Des Moines with the Imperial System of Measurement, and Maui with the Metric System of Measurement, which pretty much has the same payoff. Both are good and valid systems, but the metric system is more rational and easier to understand. So the reason the US doesn't move to the metric system, is because everyone is already using the imperial system—you are stuck in a local optima and you would need to expend a lot of energy to get out of it (convince a lot of people to get a movement going).

The same is true for greeting by bowing or shaking hands. When you meet someone, both parties should do the same to prevent poking-out eyes. So is shaking hands better than bowing? Shaking hands means you can look into the eyes, there is physical contact and you can check for sweaty palms, firm grip, *etc.* But, you are put in contact with the other persons germs, which in the modern world, after a long flight, could make you prone to a cold. So perhaps 10 years ago shaking was better, but now with travel arrangements bowing has surpassed shaking? Perhaps, this doesn't matter though because we are stuck in the shaking local maximum.

12.3 Emergence of Culture

Why do cultures differ? Extend the pure coordination game to multiple games, because in every culture there are a lot of different coordination games being played:

- Wear shoes inside the house?
- Cross the street when the "Don't Walk" sign is flashing?
- Read the newspaper at breakfast?
- Hug friends?

- Interrupt when talking?

Each of these is a pure coordination game: Do you get dirt on your feet or take off your shoes? Do you get run over by a car? Are you being rude at the breakfast table?

The complexity of this multiple-game approach blows up pretty quickly. If you simply assume 20 coordination games with two answers, then we get $2^{20} > 1$ million different cultures. So we get different cultures because everyone is trying to coordinate on loads of different aspects.

Axelrod's culture model A set of features (coordination games) with a set of certain traits (possible ways of behaving) and any person is just a combination of features with associated traits. So you start off with a grid of people with certain feature-trait combinations, then each person on the grid will look to their neighbour, and by assigning a probability, that person will change a certain trait of a random feature given what the people around them are doing. If we have someone who is very different from us, then we choose not to interact with them at all.

Let's say that before meeting, a leader has the following arbitrary traits: 53211, and the follower these traits: 51331. The sharing percentage is: 40% (2/5). But perhaps after meeting the follower switches to: 53311.

If you set something like this up on a computer, you see the formation of different cultures even when the sharing percentages are relatively low. Each culture is surrounded by other cultures which essentially protects their unique identity.

So Axelrod's model gives us the notion of thick boundaries: people near each other will either be exactly the same or differ by a lot, and the dividing line is a very deep trench. If the deep trench wasn't there, then people would interact and the cultures would merge.

12.4 Coordination and Consistency

So culture arises as a result of multiple coordination games. But we can also add some other features such as

1. Coherence and consistency: knowing one trait of a specific group can allow predictions about other traits.

2. Heterogeneity: even if an entire population can be quantified as a particular culture, this does not mean that everyone within that culture will behave the same way. Rather, the aggregation of all the behaviours leans in one direction.

We can make some assumptions to incorporate this:

1. Values have "meaning".

2. People desire consistency.

3. How much innovation/error do we need in terms of people seeking consistency in order to get the cultural differences of Axelrod's model? Given the heterogeneity in most cultures, we should expect a fair amount of error.

So the bottom line is that the coordination model can be extended to incorporate consistency and errors as well.

1. Coordination rule: Before meeting, the leader has traits 53211 and the follower the traits 51331. After meeting, the follower changes to 53311, *i.e.* the follower changed some random trait that was non-overlapping with the leader.

2. Consistency: 5144 becomes 5544. So pick two attributes and set the value of the second equal to the value of the first. Suppose you hug your friends but not your parents, then some form of cognitive dissonance is likely, where you question why you are not hugging your parents.

What do you expect to get? A consistent coordinated model. Indeed, this is what does happen if you run a model like this, but it takes very long to converge.

An unexpected outcome is that small errors lead to substantial population level heterogeneity. So small random errors, like not hugging your parents but your friends, leads to an unproportionally big spread in behaviour within a culture.

Why this big spread? If you have a culture where every trait for every feature is the same, then you would think that one small error or innovation would just go away. But equally, you have a non-zero likelihood that this error/innovation will spread to other features of that person, or to the same feature of other people in the population. Given enough people in a population, these small errors become significant—heterogeneity spreads out.

Can we model this mathematically? Without error, the state of everything being equal is a stable sink, it attracts the system and then just stays there. But with error included, you can actually transition away from this sink and end up at any one of the possible states. Hence, we can model this via a Markov process with a giant transition matrix, and given different error rates, compute the equilibrium.

So what we have learned in this section is that coordination explains a lot of what creates culture and cultural differences. We can coordinate on "wrong" actions, we can emerge our preferences over time, and third we can coordinate with ourselves to create consistency.

So the bottom line is that culture is multiple and consistent coordination. Also, small amounts of innovation/error can lead to large amounts of heterogeneity. This means that culture can be modelled as a combination of Lyapunov and Markov models. The Lyapunov model explains the maximisation/minimisation in the absence of errors to reach a specific state, and small errors that turn that process into a Markov process as all possible states are now obtainable by starting from any other state.

13 Path Dependence

Path dependence implies that what happens in this very moment depends intricately on the preceding moments—the path we took to get here. In a Markov process history does not matter at all, but we need models that allow us to incorporate historic effects.

Path dependence The keyboard on your computer is probably the typical QWERTY outline, and the fact this is so is because there was path dependence to previous typewriters as this arrangement caused less jamming of keys.

So the bottom line of path dependence is that the outcome probabilities depend upon the sequence of past outcomes. Note that the outcome is not necessarily path dependent but the probability of the outcome is path dependent.

Where does this apply? Some simple examples.

- Technology: AC *vs* DC electricity, gas *vs* electric cars
- Common law: how the law evolves is dependent on what the law was before
- Institutional choices: do you have a single-payer system or multiple-payer system
- Economic success: the current economy depends on previous outcomes. Ann Arbor is the largest public university, while a couple of miles away in Jackson we have the largest four-walled prison. These were choices made in the past but this had a massive effect on how the cities developed. For example, the population of Jackson continued to decrease over the last decades but the population of Ann Arbor is continually increasing. This was because after WWII, the GI Act brought loads of veterans to Ann Arbor to study.

When people talk about path dependence they often mean increasing returns. So when Ann Arbor built the university and hospitals, the hope was that there was going to be a virtuous cycle of people moving to Ann Arbor, which then attracted more people, and so on. BUT this is not true. Path dependence is not the same as increasing returns!

Path dependence is related to chaos, namely extreme sensitivity to initial conditions, *i.e.* the path taken depends on the initial conditions.

In Markov processes, history doesn't matter but there is statistical churning from one state to another. In path dependence models, you also have an evolving system churning from state to state but this time history matters. Which assumption is broken to invalidate the Markov process? The fixed transition probabilities.

13.1 Urn Models

To describe path dependence we will use a model based on urns, and this model will allow us to distinguish between models that are path dependent and those that are "phat" dependent—the order of events doesn't matter but the set of things does. They will also helps us distinguish between processes that are path dependent over a specific period but converge to the same equilibrium in the long run, and those that are purely path dependent.

Urn model - Bernoulli Consider an urn that contains balls of various colours. The outcome we measure is the colour of ball selected and this outcome has a specific probability associated to it. The simplest urn model is the Bernoulli model (which describes many casino games), where you have a fixed number of balls in the urn. For example, if an urn has blue and red balls then:

U = {B blue, R red}
P(red) = R / (B+R)
Note, that the probability stays fixed and the probabilities are independent, *i.e.* the probabilities of certain outcomes do not change over time.

Urn model - Polya Process Start with U = {1 red, 1 blue}. Select a ball, and add a new ball that is the same colour as the ball selected. Then repeat.

What is going to happen is that the probabilities change because the number and distribution of balls is changing—path dependent. The more red balls you select the more likely it becomes that on the next round you will select a red ball again.

1. Result 1: Any probability of red balls is an equilibrium and equally likely. We could get 4%, 23%, or 99% red balls. Any of those outcomes is equally likely.

2. Result 2: Any history of B blue and R red is equally likely. So the probability of $RRRB$ is the same as $BRRR$. The reason this is interesting is because if you know the frequency of B and R balls you can not infer the order because each order is equally likely.

Why is the Polya process important? Suppose the Polya process is a model for fashion—choosing a red or blue shirt. You buy a red shirt, now your friend will be influenced to buy a red shirt. Now someone else comes in and he thinks "What is more popular?", and now he is more likely to buy a red shirt. The same model can be used for PC *vs* Mac, *etc.*

Urn model - Balancing Start with U = {1 Blue, 1 Red}. Select and add a new ball that is the opposite colour of the colour you selected. So the more red balls you select the more likely it becomes that on the next round you will select a blue ball.

This results in a balancing process that converges to equal percentages of the two colours of balls. Where does this apply? It applies where you want to keep equal constituencies happy. So if you invest in the Southern USA then you'll increase the probability that next time you'll invest in the Northern USA to keep everything equal. The Olympic selection process of balancing games over the globe is similar.

Important Distinction: So in a Polya process anything can happen and you can get any outcome of probabilities between 0 and 1 with equal chance, while in the balancing process you get a fixed equilibrium of 1/2 equal probabilities. So in the Polya process the equilibrium is extremely path dependent, but in the balancing process it is not. In the balancing process, a short period before the equilibrium might be path dependent but the final equilibrium is not.

- Path dependent outcomes: colour of the balls in a given period depends on the path taken

- Path dependent equilibrium: percentage of red balls in the long run depends on the path taken

	Path dependent outcome	Path dependent equilibrium
Polya	x	x
Balancing	x	1/2

This distinction is important because in the balancing process history can matter at each moment but not have any impact in the long run. Conisder two examples. The US was always going to be a sea-to-sea country but the history that got you there was path dependent. Another example was the railroads. The tracks were always going to be laid connecting the cities but the particular sequences of Carnegies and Rockefellers laying these tracks didn't matter.

Another distinction:

- Path dependence: outcome probabilities depend upon the sequence of past outcomes
- Phat dependence: outcome probabilities depend upon past outcomes but not their order

The Polya process is PHAT because the outcome probabilities do not depend upon the order in which outcomes occurred. They only depend on the set of outcomes that occurred.

For thirty rounds the possible number of paths is $2^{30} > 1$ billion, while the total number of states or sets is $\{0, 1, 2, 3 \ldots 30\} = 31$. So if something is path dependent then a billion possibilities matter, but if something is set dependent then only 31 possibilities matter.

Urn process - sway process Start with U = {1 blue, 1 red}. Select one ball and in period t, add a ball of same colour as the selected ball and also add $2^{(t-s)} - 2^{(t-s-1)}$ of the colour chosen in each period $s < t$.

The interesting thing about this process is that early decisions take on more and more importance for the final outcome. If this is the case, then you can get full path dependence. For example, early mover advantage, old laws influencing new laws, *etc.*

13.2 Mathematics on Urn Models

Polya Process: Start with U = {1 blue, 1 red}. Select one ball, then add a new ball that is the same colour as the ball selected. Let's prove result 1 and result 2 mentioned in the previous section:

Result 2: Any history of B blue and R red balls is equally likely. Let's start with three blue and one red:
$P(RBBB) = 1/2 * 1/3 * 2/4 * 3/5 = 1/20$
$P(BBBR) = 1/2 * 2/3 * 3/4 * 1/5 = 1/20$
$P(BRBB) = 1/2 * 1/3 * 2/4 * 3/5 = 1/20$
$P(BBRB) = 1/2 * 2/3 * 1/4 * 3/5 = 1/20$
Or equally 2 red and 2 blue:
$P(RBRB) = 1/2 * 1/3 * 2/4 * 2/5 = 1/30$
$P(BBRR) = 1/2 * 2/3 * 1/4 * 2/5 = 1/30$
$P(RRBB) = 1/2 * 2/3 * 1/4 * 2/5 = 1/30$
$P(BRBR) = 1/2 * 1/3 * 2/4 * 2/5 = 1/30$

The denominator always stays the same because it's just $2 * 3 * 4 * 5 \ldots$, and the numerator also stays the same because the order of the numbers just changes.

Result 1: Any probability of red balls is an equilibrium and is equally likely.
$P(BBBB) = 1/2 * 2/3 * 3/4 * 4/5 = 1/5$
$P(BBBR) = 1/2 * 2/3 * 3/4 * 1/5 = 1/20$ but could have gotten $P(BBRB)$, $P(BRBB)$, $P(RBBB)$ at equal probability (from result 2 above), so the probability is actually $4/20 = 1/5$.
$P(RRRB) =$ same as above
$P(RRBB) = 1/30$ but this probability is the same as $P(RBRB)$, $P(BRBR)$, $P(BBRR)$, $P(RBBR)$, $P(BRRB)$ all at the same probability, and so we actually have $6/30 = 1/5$.

So let's now go to a much harder case:
$P(50B) = 1/2 * 2/3 * 3/449/50 * 50/51 = 1/51 = P(50R)$
$P(49B\ 1R) = 1/2 * 2/3 * 3/449/50 * 1/51 = 1/50 * 51$, but there are 50 ways of creating the

$P(49B\ 1R)$ because the red can appear in one of 50 places. Therefore, $P = 1/51$.
$P(47B\ 3R) = 1/2 * 2/3 * 3/447/48 * 1/49 * 2/50 * 3/51 = 1 * 2 * 3/(48 * 49 * 50 * 51)$

Well, the first red ball could have occurred in 50 different places, the second ball in 49 different places and the third ball in 48 different places. Furthermore, I could have arranged each of these three balls in 3! places (1,2,3; 1,3,2; 2,1,3; 2,3,1; 3,1,2; 3,2,1), and hence I need to multiply my probability by $48 * 49 * 50/(3 * 2 * 1)$ to get 1/51 overall!

We can of course repeat this for any $P(xB\ yR)$.

13.3 Path Dependence and Chaos

Path dependence means that the sequence of previous events effects the long term outcomes including equilibrium.

In a Markov process the long term equilibrium never depends on the previous events—history doesn't play a role because the system always wants to converge to the same equilibrium. Let's recap the Markov assumptions:

- finite states with fixed transition probabilities between all states
- all states are obtainable from any other state
- we are not constricted to a simple periodic cycle.

Given these assumptions, a Markov process converges to a stochastic equilibrium which is unique. Markov processes are not path dependent because the transition probabilities are fixed. For a Polya process, for example, the transition probabilities changed because we keep adding balls. So history matters when the transition probabilities change.

Now let's relate path dependence to chaos. We will do this using a recursive function. The outcome at time t, is $x(t)$ and with an outcome function $F(x)$ which spits out new outcomes given the previous one, for example, $F(x) = x + 2 \to 1, 3, 4, 7, 9, 11 \ldots$ Chaos is extreme sensitivity to initial conditions. This means that in a chaotic system, if initial points x and x' are very close starting points—they differ just by a tiny amount—after many iterations of the outcome function, the outcomes differ by arbitrary but large amounts.

Example - Tent Map Let X be a number in the interval $(0, 1)$. We define the outcome function

$$F(x) = 2X \text{ if } X < 1/2 \tag{17}$$

$$F(x) = 2 - 2X \text{ if } X > 1/2 \tag{18}$$

So $F(x)$ ramps up to 1 between $0 < x \leq 1/2$ and then ramps down to 0 between $1/2 \leq x < 1$. Let's start with two points very close to each other and observe what happens over the iterations (see Table 3 and the plot in Figure 9).

Note that this is not path dependence because everything really just depends on the value of one initial point. So chaos in its classical definition means extreme sensitivity to initial conditions, and path dependence means what happens on the way has an effect on the outcome.

- Independent: Outcome doesn't depend on starting point or what happens along the way
- Initial conditions: Outcome depends on starting state

Table 3: Chaos in numbers

	X = .4321	X = .4322
F(x)	.4321	.4322
F(x)	.8642	.8644
F(x)	.2716	.2712
F(x)	.5432	.5424
	... at 14th period ...	... at 14th period ...
F(x)	.2368	.5824

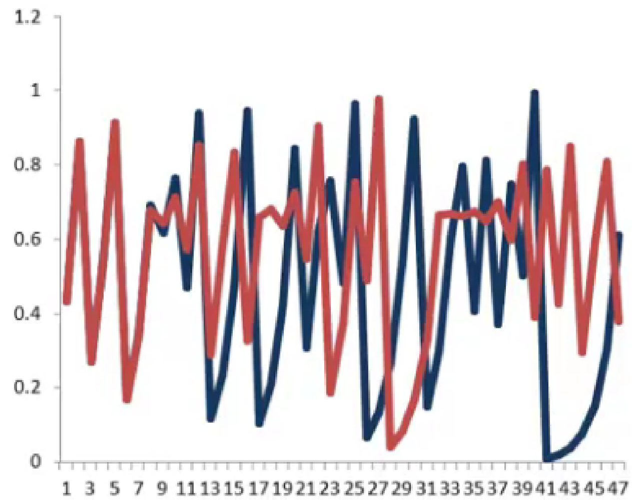


Figure 9: Different histories seeded by small variations in initial conditions—the hallmark of a chaotic system.

- Path dependent: Outcome probabilities depend upon the sequence of past outcomes
- Phat dependence: Outcome probabilities depend upon past outcomes but not their order.

When historians think of past events they think the sequence matters—not only the events and not only the initial conditions. So they side with things being path dependent. Why? Independence doesn't make sense because then everything would just be random without causality. Chaos doesn't make sense because then everything would be determined by the initial conditions. Why is it path rather than phat? Why does the sequence matter?

It's because the early events have the largest influence and therefore the sequence matters. Let's look at the US as a counterfactual: 1814 women have right to vote, 1823 civil war ending slavery, 1880 finish transcontinental railroad, 1894 find gold in California, 1923 buy Midwest from France, 1967 defeat British in war of independence. It is likely that American society would look entirely different if history had played out like this, *e.g.* everyone might be speaking French.

13.4 Path Dependence and Increasing Returns

Increasing returns: The more I have something, the more I am going to get/accumulate or the more other people want it. The more people get QWERTY typewriters, the more other people want QWERTY typewriters to get consistency.

In the Polya process we get increasing returns because the more you select one ball, the more likely it is to get the same colour because you are adding that colour back to the urn. So is this process an example of increasing returns (are they the same thing) or are they just related? Is increasing returns logically the same as path dependent equilibria?

Example - Gas *vs* Electric cars Both gas and electric cars originally had the same increasing returns, because the more people would drive them, the more other people would want them. But gas won out because its increasing returns were stronger.

U = {5 blue, 1 red}—where blue is gas. Select one and if red, then add 1 blue & 1 red. If blue, then add 10 blue.

So now I have increasing returns in both blue and red, but because the returns in blue are much stronger than those in red, I am always going to get blue dominating in the long run. So this process has increasing returns, but it is certainly not path dependent because

- the equilibrium is always the same
- not even in the short term are there large variations in outcome

So increasing returns don't directly give rise to path dependence. So can you get path dependence without increasing returns? Yes!

Example - Symbols U = {1 blue, 1 red, 1 green, 1 yellow}. Select one; if red then add 1 green; if green then add 1 red; if blue thenadd 1 yellow; if yellow then add 1 blue. So there are no increasing returns here. BUT if I aggregate red and green balls into one set, and blue and yellow into one set, then I see that this is just the Polya process on these sets, and it will produce path dependence.

So increasing returns doesn't mean it is path dependent, and you can get path dependence without increasing returns. Thus, these two concepts are logically distinct.

Beyond Urns Most path dependence does not arise from Urn-like models but from social interactions. Particularly, externalities, *i.e.* how choices by one person affect other people. Public projects are BIG. They are likely to bump into each other, and therefore create externalities, *e.g.* roads, universities, sewers.

Consider projects {A,B,C,D,E} with each project having a value of 10 and creating externalities as shown in Table 4. So the diagonal of the table is the value of each project, and the off-diagonal terms are the externalities.

Table 4: Table of project values and externalities.

	A	B	C	D	E
A	10	-20	5	-10	10
B	-20	10	-10	30	-10
C	5	-10	10	0	0
D	-10	30	0	10	0
E	10	-10	0	0	10

Value of A: 10
Value of AB: 10+10-20 = 0
Value of AC: 10+10+5 = 25
Value of ACD: 10+10+10+5-10+0 = 25

Value of B: 10
Value of BA: 0
Value of BC: 10+10-10 = 10
Value of BD: 10+10+30 = 50

So if I start out with A, I might do AC, but if I start with B, I might do BD. Thus, the externalities create path dependence, and the choice of the first project defines what project I will do later. In fact, if most externalities are positive then path dependence is less likely than if externalities are negative.

13.5 Path Dependence or Tipping Points

Path dependence: what happens on the way affects the outcome, *i.e.* outcome probabilities depend upon the sequence of past outcomes.

Let's define two distinctions:

1. path dependent outcomes: outcome in a given period, the outcome depends on the path
2. path dependent equilibrium: the equilibrium in the long run depends on the path

When relating this to tipping points it's the path dependent equilibrium class that we want to focus on. We previously defined two types of tips: active tips and contextual tips. Active tips were changes in the control parameter causing a tip in that parameter, whereas contextual tips were changes in another parameter causing a tip in the control parameter.

What's the difference between path dependence and tipping points? Tips were a single instance in time where the equilibrium changed drastically. Path dependence was what happens along the path that effects the final equilibrium, *i.e.* the accumulation of what happens. For tips it wasn't accumulation but a single event.

Polya process - draw four balls - tip or path dependent? One way we can decide if this is a tip is to use the diversity index, which we previously introduced for tips. Diversity index: if I draw four balls out of the urn, I can get the probabilities listed in Table 5.

Table 5: Probabilities of drawing different number of red balls in a Polya Process.

Nr of red balls	0	1	2	3	4
Probability	1/5	1/5	1/5	1/5	1/5

This means the diversity index is $= 1/\sum(P_i)^2 = 1/(5*(1/5)^2) = 5$.

Now let's change the system by adding a bit of initial information. Suppose the first ball is red so that we have two red and one blue ball (added one red). Now the chance of getting three reds in a row is: $2/3*3/4*4/5 = 2/5$. The chance of getting two red and one blue is: $2/3*3/4*1/5$ but we can have 3 ways of getting this (RRB, BRR, RBR). So the probability is 3/10. The chance of getting two blue and one red is: $2/3 * 1/4 * 2/5$ and multiplying by 3 we have $2/10$. Chance of all three blue is: $1/3 * 2/4 * 3/5 = 1/10$.

So under these new conditions we have

P(4R)	P(3R)	P(2R)	P(1R)
4/10	3/10	2/10	1/10

The diversity index now is $\frac{1}{(16+9+4+1)/100} = 10/3$. An abrupt tip would mean that the diversity index went from 5 to something like 1.2. But it only went from 5 to 3.33 which means something happened along the way, but it wasn't abrupt enough to be classified as a tip. So the difference between tipping points and path dependence is one of degree. In path dependence we move from one state to some other state in a gradual way, but in tipping points the change is much more abrupt.

14 Networks

Networks have become very popular, especially because of the Internet. All this data from networks is providing us with information about social trends, political views, *etc.*

Why are networks so interesting to study? They provide very clear visual information about a phenomenon when plotted out, *e.g.* clusters, trees, groups, segregation, important nodes/influencers. Second, networks provide us with information on how entities interact, *e.g.* how many liberal bloggers are citing conservative bloggers and vice versa.

What are we going to look at in this section?

- the logic of networks. What rules do people use to form networks?
- the structure of networks. How many nodes are there? How are they connected?
- the functionality of networks. What does the network do, especially emergent functionality?

14.1 Structure of Networks

Structure of a network: nodes and edges. Nodes are individual entities which are connected by edges. The edges can either be directed or undirected. An undirected edge does not have a direction, *e.g.* all the states of the USA. A directed edge means that a node points to another node, but not necessarily the other way around, *e.g.* fashion influence.

There are four measures of structure:

1. Degree: how many edges each node has on average. One edge at a node means that the node has a degree of one. Four edges connected to a node means a degree of four. The degree of the network is the degree of each node divided by the total number of nodes.

   So, the average degree of the network = 2 * edges/nodes (because each edge has two nodes connected to it).

   All nodes connected to a specific node are called the neighbours of that node. This results in the following theorem: *The average degree of neighbours of nodes will be at least as large as the average degree of the network.* This implies that most people's friends are more popular than they are themselves. Consider the following example network in Figure 10.

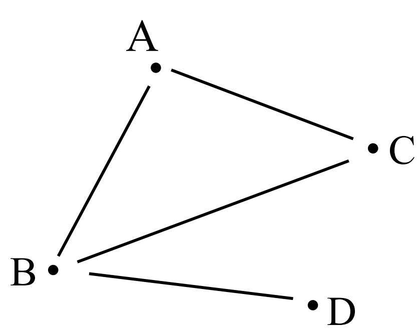


Figure 10: An example network.

The degree of the four nodes and their neighbours is:

- degree(A) = 2, average degree of neighbours (B,C) = 2.5
- degree(B) = 3, average degree of neighbours (A,C,D) = 1.66
- degree(C) = 2, average degree of neighbours (A,B) = 2.5
- degree(D) = 1. average degree of neighbours (B) = 3

The average degree of the network is 2, while the average degree of the neighbours is $9.66/4 = 2.4$. So the degrees-of-neighbours average is larger than the network degree.

The network degree is important because it tells you about the density of connections. For example, it could tell you about how people are connected, which could be a proxy for social capital. It could also measure the speed of diffusion throughout the network.

2. Path length: how far it is from each node to another, calculated from number of edges times length. The average path length is the average path length between all nodes, which equals $\sum$ (edges*length) / nodes.

   Note, one way to represent a Markov process is with a directed network where the transition probabilities are equal to the length of the edges.

   Path length is important because it tells us about how far people are apart in an organisation, and the likelihood that information will travel.

3. Connectedness: whether the graph is connected to itself. Can you get to any node from any other node along the edges? In disconnected networks information can't flow, and this can either be good (disease) or bad (library).

   Connectedness is important because it can tell us if the network is a Markov process (getting from any state to every other), the capabilities of the network, how interconnected things are and how detrimental a failure will be.

4. Clustering coefficient: how tightly clustered the edges are. It is the percentage of triples of nodes that have edges between all three nodes.

   So is A connected to B and C, B connected to A and C, and C connected to A and B? For example, four nodes have four different triangles between them—do these exist? If we have one of them then our clustering coefficient is 1/4. If we have two of the triangles then we have a clustering coefficient of 2/4. Clustering means that we can have networks with the same number of nodes and edges, but with different clustering coefficients.

   Clustering can tell us something about redundancy and robustness if one node gets knocked out. It's also a measure of social capital, and finally innovation adoption because it can go around the loops of the triangles (feedback loops).

   But, the most powerful measure is actually just looking at a picture of the network because it most intuitively shows all of these measures. Oftentimes the measures can be relatively similar in terms of their magnitude, but pictures of the networks are vastly different. A picture is worth at least 3-4 statistics, we would have to look at all these measures together to detect the difference in networks to the same degree as in a picture.

14.2 Logic of Network Formation

We want to think of nodes in a network as agents that can make decisions about what other nodes to connect to, and these decisions create structure in the network. Furthermore, these microdecisions will create macro outcomes—emergent properties.

We'll focus on three simple ways that networks can form:

1. Random—stochastic formation of networks

   N Nodes with P probability that two nodes are connected. It turns out that you get a contextual tipping point. For large N, the network almost always becomes connected when $P > 1/(N-1)$.

   The problem with these random networks is that, in reality, most networks do not look random. There is some structure to them.

2. Small worlds—some nodes initially connect to other nodes near them, and then these nodes randomly connect to other nodes further away. Social networks look like this.

   People have some percentage of "local" (people you live with, work with) or "clique" (people that you met travelling) friends and some percentage of random friends.

   If you start of with a set of nodes arranged in a circle, with each node only connected to its neighbour, then the clustering coefficient starts off at 0.5. Then if you start rewiring by creating random connections between nodes, what you see is a drop in the clustering coefficient and also a drop in the average path length.

   So as people have more random friends we get less clustering and also a smaller social distance between different people.

3. Preferential attachment—used to describe the internet. You are more likely to connect to nodes that are more connected. This microlevel rule creates emergent macro properties.

   Nodes arrive on the scene, and the probability that the node connects to an existing node is proportional to the node's degree (the number of connections it already has). So basically, if a new node arrives it will want to connect to the node that is most connected.

   The first thing that happens is that there are a handful of nodes that develop with loads of connections, and the majority of nodes will have very few. Basically, a long tail distribution.

   Statistically, the specifics of the network will look different when run over multiple iterations but the distribution of the network nodes (some nodes with high degree, a lot of nodes with low degree) will be the same. What does this mean? The exact network you will get is path dependent, but the distribution you get is not.

14.3 Network Function

When a network has some structure to it, some form of functionality will emerge. So given the structure (clustering, path length, node degree, *etc.*), what function can the network perform?

When nodes are interacting, they are mostly thinking about themselves, but when these preferences interact on the microlevel, they can lead to emergent properties or functionalities on the macro level.

6 degrees of separation Stanley Milgram asked 296 people from Nebraska to get a letter to a stockbroker in Boston. On average it took 6 steps to get there. Duncan Watts did the same as an email experiment with 48,000 senders and 19 targets in 57 countries. The average number of connections was six again. Why is it six? People definitely don't start with the aim of having 6 degrees of separation to any random person. This property just emerges.

Random clique network: each person has C clique friends and R random friends. Also we'll use the concept of a K-neighbour, which is a neighbour that is at least K steps away from you.

So in our random network of C clique friends and R random friends, the C clique friends are all 1-neighbours because we know everyone in the clique. Furthermore, our individual random friends in other cliques are also 1-neighbours. So 1-neighbours $= C + R$. The 2-neighbours are our clique friends' random friends, our random friends' random friends, and other random friends' clique friends.

So 2-neighbours $= CR + RR + RC$. My clique friends' clique friends are just my clique friends. So $CC = C$.

Continuing this, my 3-neighbours are $RRR + RRC + RCR + CRR + CRC$ and $RCC = RC$, $CCR = CR$ and $CCC = C$.

So let's now say that $C = 140$ and $R = 10$, which is a realistic distribution for most people.

- 1-neighbours $= 140 + 10 = 150$
- 2-neighbours $= 14 * 10 + 140 * 10 + 10 * 10 = 2900$
- 3-neighbours $= 10*10*10+10*10*140+10*140*10+140*10*10+140*10*140 = 239,000$

239,000 is a lot of people. This means that people get jobs, find friends, partners, *etc.* through weak ties. Not necessarily your first ties. 3-neighbours are your roommate's brothers' friend, or your mother's coworkers daughter, which are not that far removed from you, but loads of them exist and so they are much more likely to get you a job. This translates into the strength of weak ties.

Another functionality that emerges is robustness. You only have very few nodes with many connections. These are the important nodes in the network which need to be taken out to destroy it. Because these nodes are dwarfed in number by the less-connected nodes, the likelihood of knocking out a highly connected node is very low.

This also means that if you want to intervene in the network to create some change, then you want to target the highly connected nodes—the nodes that are most central to the network. So for example, if you think about vaccinations, you would want to vaccinate the highly connected people, the teachers, bus drivers, *etc.*

15 Path Randomness and Random Walk Models

Sometimes outcomes depend on ability, and sometimes on a bit of luck. Typically it is a combination of these two factors. To model luck we can use random walk models. In a random walk model the cumulative value of something depends on a sequence of random variables.

We'll start out with a definition of randomness and from that we will develop a model that includes factors of skill and luck. This should answer the question of how much of the outcome should be attributed to luck and how much to skill.

Examples: John Steinbeck and Jesse Owens—Were these guys good or were they lucky? Of course, both were skilful because they needed skill to write and run, respectively. But of course, there was also a fair amount of luck involved in being at the right place at the right time, knowing the right people, *etc.*

Then we will consider random walk models. We'll consider binary random walks where we just take steps of $+1$ or -1, and then we will consider normal random walks where the step size is taken from the normal Gaussian distribution. This second model can be used to explain the efficient market hypothesis, which states that price movements in the stock market are random walks.

15.1 Sources of Randomness

Let's unpack a little of what we mean by randomness. It can have different conceptual underpinnings. We need to start with a probability distribution, where we plot the possible outcomes versus the probability of their occurrence. More commonly we get a bell curve, power law or a long tail distribution.

So when considering randomness we need to ask ourselves two questions:

1. What is the probability distribution of the randomness?
2. Where does this randomness come from?

When we write randomness in models we use a small perturbation variable, ϵ, such that our state variables are perturbed to $X + \epsilon$. The source of this ϵ can come from noise. Secondly, it can come from simple measurement error or human error, *i.e.* differences in measuring dimensions, variations in behaving, *etc.* Third, we have uncertainty because in any project things may change as you are moving along, *e.g.* materials get more expensive/cheaper, or the underlying factors change, *etc.* Fourth, we have complexity, which can create emergent phenomena from interactions of the constituents and these are typically very hard to predict precisely. Thus, it is easier to account for complexity using the same randomness parameter ϵ. Finally, we have capriciousness; people are difficult to predict, they are crazy and so we throw in an error term to account for their crazyness.

15.2 Skill and Luck

What is the role and skill in performance? We could think of luck as a form of randomness. We could simply state that outcome = skill + luck. It would be nice to construct a model that gives us an overview of how much can be attributed to luck and how much to skill. For example,

$$\text{Outcome} = a \times \text{Luck} + (1 - a) \times \text{skill}, \quad \text{for a in the range (0,1).} \tag{19}$$

So now variable a measures the position along the luck-skill continuum.

How do you determine variable a? If a person/company continually produces good outcomes, then you would infer that a should probably be large. If we have $a = 1/2$, equal amounts of luck and skill, then we have

$$O = 1/2 \times L + 1/2 \times S, \tag{20}$$

where S is the skill level and L the luck level. Because L is random we should expect a large variation in outcome depending on the size of L. If $a = 0.1$ then we have $O = .1 \times L + .9 \times S$ and now the marginal effect of the $.1 \times L$ term is much smaller compared to the $.9 \times S$ term, assuming S and L are of similar magnitudes.

Why do we care about this? We want to assess outcomes, *i.e.* recognise if the outcome is because of luck or skill. If it is luck then there will be a lot of reversion to the mean. Hence, it allows us to maker better predictions. It also allows us to align with reality and provide good advice/guidance/feedback to our peers. Finally, because we understand reality better, we can arrive at a fairer allocation of resources and areas to exploit to get repeatable outcomes.

We can split most human endeavours into two domains:

1. skill-dominated domains
2. luck-dominated domains

Ideally, you want to be competing in skill-dominated domains because this is really where you can get consistent results when you have an advantage in skill. But, there is a catch. **The paradox of skill** . . .

When you have the very best people competing, the difference in their skills levels will be very close. So the winner will ultimately be determined by luck. This also means that outcomes for the highly skilled can be quite distinct over many events, even though the skill levels are very close. This means that there is a lot of skill involved in getting into the top 5, but a lot of luck when you finish #1.

15.3 Random Walks

Random walk models are derived from flipping a coin, where each outcome (heads or tails) has a probability of 1/2. This concept can be used to create a binary random walk.

Binary random walk Let's set $X = 0$, and each period we flip a fair coin. If it is heads we add 1, and if it is tails then we subtract 1. There are three surprising mathematical results from this model.

1. After N (even #) of flips you should be expect to be at 0.
2. For any number K, a random walk will pass both $-K$ and $+K$ an infinite number of times. What this means is that the process will never take off and diverge. This result is quite counterintuitive because K is ANY number.
3. For any number K, a random walk will have a streak of K heads, (and K tails) an infinite number of times. So streaks will happen and do happen and are not something mythical. Again, this is counterintuitive because this means that any streak will occur an infinite number of times over an infinitely long period. Consider this simple thought experiment. Consider flipping a coin 16 times. The chance of getting 16 heads is $(0.5)^{16} = 65,536$. This is very unlikely to happen to any single person, but if there are 100,000 people in the casino playing this game, then it is not surprising that it might happen once.

There is another phenomenon that is implicit in these 3 results: regression to the mean. A group that did well for a short time, should be average in the long run. A good example is the "Good to Great" companies that were described in Collins' 2001 book but then underperformed over the next 10 years. This is related to the "No Free Lunch" theorem, that stated that no single metric of viewing a problem to reach a maximum is better than any other. So the metrics determined by Collins in his book may have been optimal for the time, but then the conditions changed and those heuristics no longer worked. No single or set of metrics is better all the time.

Another phenomenon that can be debunked with random walks is the idea of a hot hand. Statistics show that the probability of making a free throw after missing the first is 75%. The probability of making the second free throw after the first is also 75%. So the idea of the hot hand is a myth. It's all just a random walk, just with a higher than 50-50% probability.

15.4 Normal Random Walks and Wall Street

Instead of steps being $+1$ and -1, in a normal random walk model the steps can be any value and these values come from a normal distribution. The normal random walk can be used to establish the efficient market hypothesis, which states that any current price reflects all available information, and if this is true then any changes in the price should be random. Therefore, it's impossible to beat the market.

Normal random walk Set $X = 0$, each period draw a number from a normal distribution with mean 0 and standard deviation 1 and add this number to X. After a certain period you would expect some walks to be really high, others to go really low, but most to be centred somewhere around the 0 mark. So, some big winners, some big losers but most centred around zero.

If we extend this to Wall Street we can see historically that the returns aren't quite normally distributed. There are more days were nothing really happens, and more days were loads happens, and a bit less were something moderate happens. This modified distribution can then be established and used in a random walk.

The explanation for why the stock market is a random walk, is to say that a stock goes up or down given the information about its performance tomorrow. But that information is generally widely available and so gets priced in very quickly. As a result, if the flow of information is unimpeded and information is immediately reflected in stock prices, then tomorrow's price change will reflect only tomorrow's news (which is uncertain) and will be independent of the price change today. Because tomorrow's news is by definition unknown and unpredictable, the resulting price changes must be unpredictable (random walk).

So is this true? For example, one thing that happens is the January effect in that the stock market rallies in January. So people should rationally expect more in December anticipating this effect and invest, which would drive the market up. But this is not generally true. Other critiques include that there is way too much fluctuation in individual stocks to be described by a random walk, there are consistent winners over 30-40 years (Berkshire Hathaway). A another strong critique of the random walk hypothesis is that the stock market indexes are generally correlated with long-term GDP, suggesting that in the long term there is a correlation to the underlying economic metrics of the companies that make up the stock market.

15.5 Finite Memory Random Walks

The idea behind a memory random walk is that the value at time t depends on the cumulative value of the history before time t. The idea of a finite memory random walk is that the value depends only on the previous N periods. For example,

$$V_t = X_t + X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4} \tag{21}$$
$$V_{10} = X_{10} + X_9 + X_8 + X_7 + X_6 \tag{22}$$
$$V_{11} = X_{11} + X_{10} + X_9 + X_8 + X_7 \tag{23}$$

What are these X_t's? They could be new employees, products, *etc.*, essentially any random variable.

Sports The finite memory random walk model can be used to explain some phenomena in sports. We can imagine someone retiring and someone new being drafted, such that the value of the sports team depends on a particular period.

Let's say we have 28 teams, with the value of each team depending on five players:

$$V_t = X_t + X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4} \tag{24}$$

The champion is the team with the highest V_t. Let's run this for 28 years (which can be done in a spreadsheet). The surprising findings are that the number of distinct winning teams over three random runs (16,12 and 16 different winning teams) is surprisingly close to what actually happens in the NBA, NFL, NHL and MLB over the same period (8,18,14,18 and different winning teams). Similarly, counting the team with the most championship wins (5,8,5 in total) is surprisingly close to the NBA, NFL, NHL and MLB (8,4,5,5 in total). Finally, looking at longest streaks the model always produces 5, whereas the actual data is (3,2,3,3), so not so close. But overall 3/4 metrics have come out pretty close.

The bottom line is that this simple model has predicted that we should expect a lot of distinct winners, a team that wins a good number of the championships (dominant team) and an absence of streaks, which is what we see in the data.

In general, random walks show the following:

- there is a distinction between luck and skill
- there is paradox of skill at high skill levels
- there is regression to the mean in random walks
- the stock market has some randomness to it in the short term
- if we have a random walk it is not path dependent. So if we see some process happening we can ask is it a Markov process, path dependence, Lyapunov process or a random walk?

16 Colonel Blotto Game

16.1 No Best Strategy

Colonel Blotto Game Imagine that you and I both have a bunch of troops, and we have to divide them along some fronts to fight against each other. Whoever wins the most fronts wins the entire game. So the game is

- 2 players each with T troops
- N fronts ($T >> N$) to put troops
- Both allocate troops across fronts
- Determine winner by # of fronts won

It's a simple game at the outset but which can lead to very complicated outcomes. Overall it's a zero sum game, meaning there is a winner and a loser. In real life you want to stay out of zero sum games because you have to work really hard. This is because the other person is also trying really hard to win the single payoff = super competitive.

Example: T = 100, N = 3

	Front 1	Front 2	Front 3
Player 1	34	33	33
Player 2	40	40	20

I evenly allocate my troops. What does the other player think? Well, he might assume that I will just equally allocate my troops and then focus his troops on two fronts to win those two and lose the third, thus still winning overall. So equal allocation is not a good strategy.

The problem is: Any strategy can be defeated.

Player 2's strategy from before can be beat with 20, 50, 30

	Front 1	Front 2	Front 3
Player 1	20	50	30
Player 2	40	40	20

But this strategy can then be beat with 50,0,50

	Front 1	Front 2	Front 3
Player 1	20	50	30
Player 2	50	0	50

So there is no obvious best strategy to play.

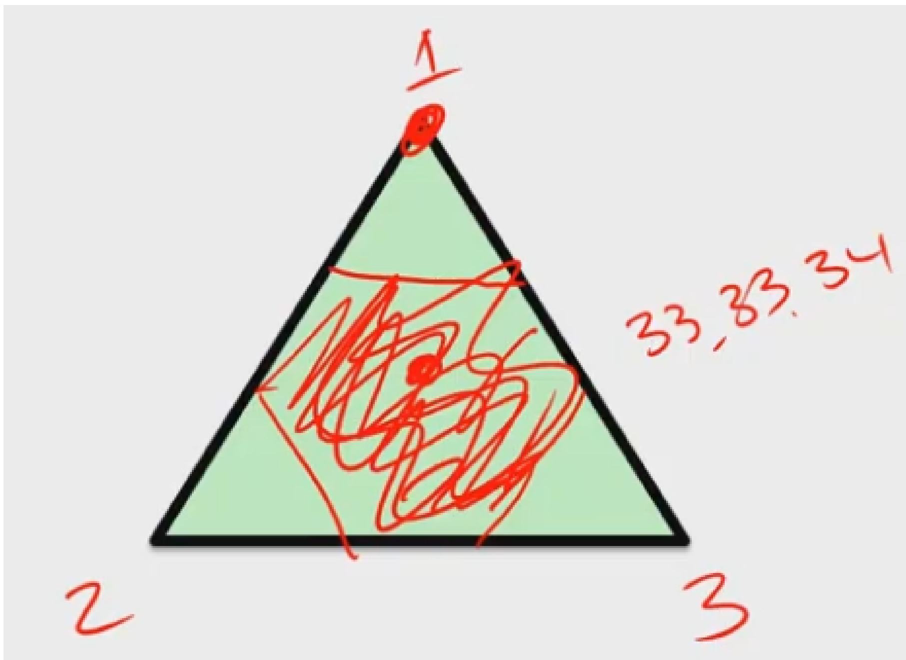
Here is another insight: You do not need all troops to win.

	Front 1	Front 2	Front 3
Player 1	0	41	21
Player 2	40	40	20

Now I am only using 62 troops out of the available 100 but still winning.

So what happens in this game in the long run? Do we get equilibrium, cycle, complexity, or chaos? It is a mixture of equilibrium and chaos.

Consider the triangle below. Putting all your troops at the vertices is obviously foolish. Equally allocating troops is also not a great strategy as shown above. But you can still win with this equal allocation strategy. What you can show is that the best strategy is randomly choosing an allocation within a central hexagon (shaded red in the figure). If both players choose strategies within the central hexagon, then the winner is random. Thus, the outcome of the Colonel Blotto game is a random equilibrium.



Is winning the Colonel Blotto game skill or luck? With equal number of troops it is probably luck, but with unequal troops, skill starts to creep in.

So Colonel Blotto is a game of strategic mismatch with no best strategy because every strategy is beaten by another, and you don't need all troops to win. For equal troops between players the best strategy is to choose one with relatively equal allocation between fronts and then the outcome is random. For unequal troops, skill becomes more important.

16.2 Applications of Colonel Blotto

Wider applicability of the Colonel Blotto game. One of the reasons we construct models is that they are fertile. A model in one setting can be applied in another setting.

The seeting in the Colonel Blotto game is troops on fronts. But that same model can be used for other forms of competition.

Reminder: Blotto means that 2 players with T troops need to allocate resources on N fronts ($T >> N$). Whoever wins more fronts, wins the game. In this description, the military setting is obvious. Does it apply anywhere else?

- Electoral College: To become president you do not need to win the most votes but the most states. So where do I allocate my troops, that is money, time, helpers, *etc.*, to win the most states? For example, California is staunchly Democratic and so neither party spends much time or money there. What they do is spend time and money in the swing states of the Midwest and Florida.

- Terrorism: There are a whole bunch of places where a terror organisation can attack, *e.g.* airports, bus stations, stadium, water systems, financial places, *etc.* How do we allocate our resources to prevent the most terrorism?

- Trials: Lawyers clash on over different lines of defence (arguments) and the resources they can spend is the time spent preparing for each line of defence. So how much time do I spend on each argument to win the most number of arguments?

- Hiring: Applicants compete over relevant skills and abilities, and they need to decide to what level each skill is worth developing. So the person that gets hired is not necessarily the person with the best skill overall, but the person that ticks the most boxes of the criteria that were asked for.

- Sports: For example, tennis players compete on the serve, return of serve, ground strokes, net play, *etc.*, and they need to decide how much ability they should develop in each area of the game. This explains why some players have an advantage against certain players, but disadvantages against others. Some of their skills beat some of the other person's skills but not all of them.

So the Colonel Blotto game is much like a giant game of trumps...

16.3 Troop Advantages

What does it mean to have a troop advantage?

Troop Advantage: Suppose there are three fronts and I have 180 troops and you only have 100 troops. If I allocate 60,60,60 there is not a lot that you can do. So if you want to win one front you need 61 on one front but that leaves only 39 for the rest, and so you can't win the crucial second front. So the bottom line is, that if you have more fronts you are at a definite advantage.

But what happens if we increase the number of fronts? Is it still as useful? As the number of fronts increases, I need a larger relative resource advantage to guarantee victory. So the advantage of resource superiority decreases as the number of fronts increases.

Let's say we have two sides, A and B, with troop numbers $A = 150$ and $B = 96$. If we have three fronts and A just allocates 50, 50, 50 then B can never win because allocating 51 only leaves 45. Let's say we have five fronts and B allocates 30, 30, 30, 30, 30 then B can win because he can allocate 32, 32, 32, 0, 0. Hence, the greater the number of fronts, the more the resource superiority needs to be win to guarantee victory.

Key insight: If you're the weak player, you want to add dimensions, fronts, battlefields, *etc.* to increase your chance of winning. So for a sporting event, I need to add tricks. If it's business, then we need to add new packaging, colours, *etc.* If you're an army, you add unforeseen targets for attack.

Multiple player Blotto Let's say we have the following three players with 5 fronts and 20 troops.
Player 1: 4 6 2 6 2
Player 2: 5 0 8 7 0
Player 3: 6 5 1 5 3

Player 2 beats player 1, player 3 beats player 2, and player 1 beats player 3. So $1 > 3 > 2 > 1$ and we get a cycle. In Colonel Blotto we thus expect to see cycles. In effect, Blotto games become a more sophisticated version of rock, paper, scissors.

16.4 Blotto and Competition

Let's think about firms, sports teams, *etc.* competing, and how Blotto applies to competition. There are two types of competition. Firms can have market share, *i.e.* different companies taking a certain share of the total market. You can also think of sports teams having win-loss records where everyone plays everyone. So the two ways of thinking about competition is in terms of market share or win-loss records. Can a model help us to understand this difference?

Potential Models: Random, skill plus luck, finite memory model, Blotto model. What are the different things that each of these models say?

- Random: here we would expect equal wins, lots of regression to the mean and no time dependency. If we look at mutual funds this is what goes on. There is a lot of regression to the mean and the fund that is good this year is not necessarily good the next year.
- Skill plus luck: we would expect unequal wins, *i.e.* consistent winners and semi-consistent rankings. But we would not expect a lot of time dependency. If we look at market shares in industry or sports competition then this might be a good model.
- Finite memory random walk: here we are going to have unequal wins and semi-consistent rankings, loads of time dependency, and most importantly, we will have movement from top to bottom because the amount of luck *vs* skill changes over time. This is because after some period advantages/disadvantages in the history no longer matter. So you would expect some regression to the mean and more movement.
- Blotto with equal troops: outcomes look the same as random but you will see lots of maneuvering, *e.g.* random trading, because the optimal strategy is to be random and unpredictable.
- Blotto with unequal troops: outcomes same as skill-luck because you have an advantage over the other party, but at the microlevel you should still observe lots of manoeuvering. In American Football there is a salary cap, and so you see teams that have better management and better players but you still see loads of manoeuvring which is indicative of Blotto.
- Blotto with unequal troops and limited movement: outcomes same as finite walk, lots of manoeuvring and cycles, *i.e.* we should see $A > B > C > A$.

So how do we determine if it is Blotto or Skill-luck?

- Dimensionality: if players are making high dimensional strategic decisions then it's more like Blotto. Marathon running and javelin throwing are more skill-luck. Tennis is more Blotto.

- Zero sum: if actions are only good relative to other actions (*e.g.* invading from the east) then it may be more like Blotto.

Let's apply this thinking to a Presidential Election:

- Random: There are economic shocks and these determine who becomes president
- Luck-skill: Candidates vary in their skill—communication, persuasion, *etc.*
- Random walk: it's a whole series of random events, *i.e.* world events, domestic events, economic events, *etc.*, which all add up
- Blotto: allocation of troops across states with some parties having more troops than others.

Is any one of these right? No, it is a combination of all these. Therefore by combining all these frameworks we have a lens through which to observe a process, and then we can understand what the underlying dynamics are. By combining models we get a much richer experience of what is going on.

17 Prisoners' Dilemma

The most famous game in Game Theory is the Prisoners' Dilemma. So why is it so interesting and how can you apply it to a variety of instances?

Prisoners' Dilemma: There are two players. They can either cooperate or they can defect. If they cooperate both players get a payoff of 4, and if they both defect then both get a payoff of 2. If either of them defects, the person that defected will get a payoff of 6. So there exists an incentive for either of them to defect. This creates tension for what is good collectively, namely to cooperate, and what is good individually, namely to defect.

	C	D
C	4,4	0,6
D	6,0	2,2

What is a prisoners' dilemma and what is not?

	C	D
C	4,4	0,9
D	9,0	2,2

This is not a good prisoners' dilemma because if you play it multiple times you could just have the two players alternate between 0,9 and 9,0 and then the longterm average outcome is (9+0)/2 = 4.5 which is greater than 4. So this game is an example of weak alternation.

So when people typically write the prisoners' dilemma in a formal manner they typically write it as such

	C	D
C	T,T	0,F
D	F,0	R,R

where $T > R$ and $F > T$ and $2 * T > F$. So why is this so interesting? The efficient outcome is to get T, T so that both players coordinate to get the highest collaborate payoff.

In fact, this goes along the lines of reasoning of the Pareto efficiency, which is the solution that makes everybody collectively best off. So if we make one person better off from this optimum, at least one other person will be worse off. BUT, this is also true of $0, F$ and $F, 0$, because if we go to R, R or T, T or the counter $F, 0$ and $0, F$, then one player will always be better off by getting some non-zero payoff while the other will always be worse off. So the only outcome that is not a good outcome in terms of Pareto efficiency is R, R.

However, there is another way of looking at the different choices, namely Nash equilibrium. If both players are at T, T, then either of them should be thinking to get to $F, 0$ and $0, F$ because they will be better off. So both players end up defecting and end up at R, R.

So there are 3 efficient outcomes but there is only 1 equilibrium outcome. This essentially means that the incentives are not aligned with what we want to achieve socially. It is this disconnect which is so interesting, because we want the outcomes to be aligned with our social incentives (we want to get T, T but we get R, R).

Applications: The domains where this is interesting is, for example,

- *Arms control*: cooperation is education and defection is bombs. Both countries want to spend money on education, but they can't help themselves and spend money on arms, and so everyone is worse off.

- *Price competition*: cooperation is high prices and defection is price cutting. From the companies' perspective you essentially want some type of cartel-like collusion where competitors agree to high prices, rather than undercutting each other to get more customers (of course the consumers are better off).

- *Technological adoption*: cooperation is no ATM machine, defection is installing ATM machines. This is an example of adopting incremental technology. Both banks need to spend capital to install ATMs, and in isolation, installing ATMs will provide them with some additional profits. However, when all other banks install them too, they end up all paying the capital expenditure and end up competing away the advantage. And because of the capital outlay or increased efficiency which led to lower prices, they end up making less money collectively. This was similar to the adoption of better and better looms in the textile industry (the consumer won and everyone made less money), or everyone standing on their tippy toes at a concert. Sometimes this is known as the Red Queen Effect (run faster and faster just to stay put).

- *Political campaigns*: collaboration is purely positive ads, and defection is running negative ads. If both only go positive, then both will seem like saints. But then one of them decides to publish negative statements about the other to get an advantage, the other retaliates, they go back to 50-50 chance of winning, but both appear tarnished.

Self-interest Game

	C	D
C	4,4	0,6
D	6,0	7,7

In this case the incentives are aligned with the outcomes because both players would rather defect if the other is coordinating. And in fact, if they do both defect, then they end up at the Pareto efficiency point. Outcome $7, 7$ in this case is the only Pareto efficient point because deviating from it causes both players to be worse off. But it is also the Nash equilibrium, and so outcomes are aligned with incentives.

17.1 Seven Ways to Cooperation

In the previous section we saw that in a prisoners' dilemma game, collectively the Pareto efficiency lies at cooperating but the Nash equilibrium lies at both defecting (at a lower payoff). This means that the incentives don't line up with the best outcome.

	C	D
C	4,4	0,6
D	6,0	2,2

So how do we get cooperation in a prisoners' dilemma, when it is not in either of the prisoners' interest?

Let's start with a simpler model. We have two agents and if they cooperate they create a benefit to the other of value b, and incur a cost individually of c, with $b > c$. Socially, we would therefore like to cooperate because the benefit to the other is greater than the cost to me, but individually, we would not like to cooperate because this incurs a cost. So this game captures the essence of the prisoners' dillemma, because there exists a tension between cooperation and defection.

Direct reciprocity: If we play this game many times, then maybe it is time for me to be selfish now, and then I can allow the other person to be selfish later. This is the tit-for-tat strategy, where we cooperate as long as the other is cooperating, but defect once the other person has defected. Let's assume the following parameters:

P = probability we meet again
Payoff for deviating: 0
Payoff for cooperating: $-c + Pb$, where c is the cost incurred and b the benefit.

If $-c + Pb > 0$, then $P > c/b$, and so as long as the probability of meeting is greater than the cost divided by the benefit, then cooperation should emerge in the prisoners' dilemma.

Assume you are in a big city. Then noone will let you pass in the queue of a grocery store, no matter the difference between items in your carts, because the probability is low that you will meet again. But in a small city the probability of bumping into the same person is higher, and so you would expect to see more frequent cases of reciprocity.
LA vs Iowa city:
Benefit: 10 and Cost: 2, so c/b = 1/5.
LA: P = 1/1000 and Iowa city: P = 1/2.
So LA: 1/1000 < 1/5 ($P < c/b$), whereas in Iowa city 1/2 > 1/5 ($P > c/b$).

Reputation—indirect reciprocity: Instead of directly meeting again, the person that was nice develops a reputation of being nice because the other person starts tell others about him/her.

Let q = probability reputation is known.
Payoff for deviating: 0
Payoff for cooperating: $-c + qb$.

We need $-c + qb > 0$ to get cooperation and so $q > c/b$. So this means that if the reputation is high enough, then we should see cooperation.

Network reciprocity: Let's suppose we have a set of cells in a network. Will the cells cooperate with each other? In a regular network we have k neighbours, we get cooperation when $k < b/c$. So this means that as you get more connected you have more incentives to defect.

Our payoff is: $kb - kc$
A boundary defector defects when surrounded by collaborators: $(k - 1)b$
So we need: $k(b - c) > (k - 1)b \Rightarrow k > b/c$ to get cooperation.

When we have reputation, then denser ties are better, but when we have network reciprocity denser ties are worse because there is more of an incentive to defect when everyone around you is collaborating. So the type of network you want depends on the mechanism that drives collaboration.

Group selection: Group selection states that selection acts on groups of people, for example through cultures being more fit, rather than individuals. And so groups of collaborators can win out.

If we have two groups with a certain percentage of cooperators, then within a group defectors do better because they are surrounded by cooperators, and so have incentives to reap the benefits without any costs. But between groups, groups that have more cooperators are more likely to win, because they are united to fight the other group.

Kin Selection: Different members of a species have different amounts of relatedness. Let's assume that players are related and that you care about other people based on the their relatedness, r, then $rb > c$. So you would expect that within a group with a lot of relatedness, you would see a lot of cooperation.

Laws/prohibitions: You can just make things illegal. For example, it is not legal to talk on your cell phone while driving because it is bad collectively. Also, you can't go about settling scores by murder because collectively this is bad for society.

Incentives: For example, 24 hours after a snow fall, your sidewalk has to be shovelled, otherwise you'll pay a fine.

18 Collective Action and Common Pool Resource Problems

We'll now extend the core idea of the prisoners' dilemma, *i.e.* the idea that every individual is incentivised to defect but that this is not good for the collective. We will extend it to multiple agents and look at how this differs from a the classic problem invilving a pair of agents.

Collective Action: In a collective action problem I make a choice to contribute, but this contribution comes at a cost.

Let x_i be the action of a person i for the collective good. We'll assume that the payoff to person i is

$$p = -x_i + \beta \sum_{j=1}^{N} x_j \tag{25}$$

where β is a number between 0 and 1, and x_j is the effort of others (either 0 or 1) to the collective good. So the payoff is the negative of my effort (the cost to me for my effort) plus the sum of all people in the collective multiplied by some discounting factor of the effort of others.

Example: 10 people and $\beta = 0.6$. Suppose I am x_1 and $x_2 = x_3 = \cdots = x_{10} = 1$. So my payoff is going to be:

- If I don't cooperate: 0 + .6*9*1 = 5.4
- If I do cooperate: -1 + .6*10*1 = 5

So in this case it is in my interest to not cooperate, but for everyone else in the collective it is beneficial for me to cooperate. So this is similar to the prisoners' dilemma problem.

Where does this apply? Carbon emissions: it is beneficial for the collective to reduce carbon emissions, but it places a cost on me personally and so I don't actually want to do it. In fact, I reap the benefits if everyone cooperates and I am one of the few that does not. This leads to the free rider problem—a problem where it is in my interest to not go along with the collective.

Common pool resource problem: Imagine a common resource in the "commons", let's say cod in the sea, and we are all best off if we stick to an agreement where everyone just fishes a little so that the cod can replenish. Under these circumstances it is in my personal interest to break that agreement and fish just a little more so that I can reap the benefits. The problem is that, if more and more people do this, everyone will be worse off including myself.

We'll model this as follows: x_i is the amount consumed by person i, X is the total amount consumed and the amount available in the next period is $C_{t+1} = (C_t - X)^2$.

Example: $C_1 = 25$ and $X = 20$. $C_2 = (25 - 20)^2 = 25$. So we have a nice equilibrium where replenishment is equal to the amount we are removing. If we just increase the amount taken out by 1, we already get a massive difference (the square makes it so). $C_3 = (25 - 21)^2 = 16$. So even just living slightly above your means can lead to a massive collapse as these deficits add up.

18.1 No Panacea

What we saw previously is that there are certain things we can do about the two-agent prisoners' dilemma problem (kin and group selection, reputation, reciprocity, *etc.*). For common resource problems and collective action problems there is no panacea, and we have to look at the situation individually.

In collective action problems, it is in my interest to defect if everyone is collaborating—the free rider. And in a common resource problem, it is again in my interest to remove a little more resource from the commons than everyone else, but if this behaviour spreads then the entire thing can collapse.

Here is the canonical collective action problem: payoff to person i is:

$$p = -x_i + \beta \sum_{j=1}^{N} x_j \tag{26}$$

and by studying the model we could come up with institutional mechanisms to deal with this problem. The issue is that there is a lot missing in this model, *i.e.* not all features of the real world are included. So the question is, do we need to know more?

In the collective action problem, we want to know that people are contributing, whereas in the common resource problem we want to make sure that people are not over-harvesting.

The particulars:

- *Cow grazing on the commons*: if one set of cattle overgraze, then there is no grass left for the community and all the cattle die. How do you overcome this? The problem is overgrazing. So let's create a rotation scheme, where on Monday one person gets the commons, and on Tuesday someone else gets the commons. So by branding cows and rotating you can solve this problem.

- *Lobster fishing*: we could think that we should just do the same—select fishermen and then rotate, or maybe different people can fish in different parts of the coast. There is a fundamental difference though—in the cattle problem you know exactly what the amount of resources is, you can see how much grass is left, which means you can pull cows back when resources go low. In the lobster problem, you don't have the luxury of monitoring the resource, and so you need a different approach.

- *Drawing water from a stream*: this is an asymmetric problem because the people upstream are effecting the people downstream. So we need to focus most of our efforts and policies on the upstream person to make sure he/she collaborates.

So the bottom line is that there is no panacea. It is not enough to just look at the mathematical model and based on this alone come up with different solutions to the collective action and common resource problem. Instead we need to know the particulars of each situation.

18.2 Mechanism Design

One of the reasons we use models is that they can help us build better institutions. Based on a certain model we can determine a set of actions that people can take, what the payoffs and costs are, and then design these rules into an institution. So essentially when designing institutions we want to design the rules of the game, the actions people can take, and the payoffs associated with the actions. This is essentially designing the right incentive structure to get the outcomes that we want, and prevent those we do not.

When constructing these mechanism there are typically two features of social context that we need to overcome:

- *Hidden action*: you often don't see the actions that people take, you only see the outcomes. So even though the actions are hidden you would like to create incentive structures that nudge people to take the action you want.

- *Hidden information*: if it is not the action that is hidden, it is often the information, like who are the hard workers or safe drivers. Again, knowing who is who helps us to create incentive systems, *i.e.* rules, to overcome the problems that we are facing.

Getting these incentives right is hard. "Imagine how difficult physics would be if electrons could think"—Murray Gell-Mann. Because of this it is often best to start with a model that assumes rationality, and then include some psychological biases afterwards.

18.3 Hidden Action and Hidden Information

When designing mechanisms we are essentially designing incentives structures to get the outcomes we want. So we are trying to induce people to do the right type of effort.

Hidden action: So the first thing we deal with when designing mechanisms is hidden actions—how people behave when in fact we can just see the outcomes. Most of the time we can't monitor people all the time, and nor should we want to, and therefore we want to create incentives so that people behave how we want them to.

Sometimes these are called moral hazard problems. Because I am not being watched I could slack off or cheat. So how do we overcome the moral hazard?

Suppose I have some workers that can take some effort 0 or 1. I can't see this effort but I can observe the outcome which is either good or bad. Let's say that the probability that an outcome is good given that the effort is 1 is 100%. The probability that the effort is good given that effort was 0 (slacking off) is p. I want to induce the effort level of 1, but this is difficult because it is costly to do so for the employee.

I want to create a contract that pays M if the outcome is good, and 0 if the outcome is bad. This is called incentive compatible, where it makes sense to put forward the effort. A simple model for this would be:

Effort 1: $M - c$
Effort 0: pM

Incentive compatible: $M - c \geq pM$ which means $M \geq c/(1-p)$. This means I have to pay them at least the cost divided by $(1-p)$, the probability that they were lazy.

In words, the equation $M - c \geq pM$ means that it makes sense for a worker to put in effort if the amount of money received minus the cost of effort is greater than or equal to the probability that if they slack off, the outcome will be good.

Comparative statics: $M \geq c/(1-p)$ is what I must pay people to get good outcomes. So what happens as c and p change?

1. c: if c goes up, then M goes up. This makes sense because if effort is costlier, then I need to pay people more for the effort.

2. p: if c goes up, then M also goes up. There is more incentive to slack off because the probability to get a good outcome anyway is high.

The cool thing is, we now know what the action is because it is revealed once I pay people enough.

Hidden information: The second hidden property is information about the agents in your system/organisation, *e.g.* high-ability workers, risky drivers, *etc.*

Ability of workers: High (H) or Low (L). H-worker cost per hour of effort is c. L-worker cost per hour of effort is $C > c$. What are the incentives for these two types of workers?

Propose the contract: pay M, but first you must work for K hours.

Incentive compatible:

- A high quality worker will be willing to work as long as the eventual salary is greater than the cost of the tryout: $M > Kc$
- A low quality worker will not be willing work the same number of hours because the eventual payoff will be less than the cost: $M < KC$.

This means we want to force the people to work for $K = M/C$ hours, and whoever agrees is a high-quality worker.

Comparative statics:

- M: As M gets bigger, K gets bigger. As the job pays more you have to make them work longer.
- C: If C goes up, then K goes down. As the cost of the tryout is greater for the applicants they will be less willing to work for long periods.

These are sometimes called costly signalling models. The high ability workers had to signal that they are willing to work long hours. So you use the signal to determine the hidden information, in a way the signal of working the hours in the tryout becomes a surrogate for the hidden information.

18.4 Auctions

We want to use models to help us design institutions and also to decide which institutions we can trust. For example, we can use models to design auctions.

In auctions we can have seal-bid rules, second-bid rules or calling-out rules. The overarching objective is to get as much money out of your item as possible.

Ascending Bid: Individuals keep calling out prices until noone increases the price anymore. There are three ways to behave:

1. *Rational*: only bid up to your personal value you are willing to pay.
2. *Psychological*: accounting for the behaviour of others and exploiting their mistakes or foibles, *e.g.* just bidding for the thrill of winning.
3. *Rule following*: as the name says, I follow a specific rule, *e.g.* when and how much to bid. You would still only bid up to the value you are comfortable with, but the rule might inform the increment.

Let's assume people are rational. The outcome is that the highest bidder gets the item at the second highest value because he pays the price of where the previous person dropped out.

Second Price Auction: Each person submits a bid and the highest bidder gets it at the second highest bid.

1. *Rational*: you are incentivised to bid as closely as possible to what you judge the true value of the item to be. This is because the net amount between what you value the item to be and what you end up paying is optimal, no matter what the other people are doing.

2. *Psychological*: we don't know if people will tell the truth. Maybe we can assume they will overbid because people really want the item. Nevertheless, even if they do, you should vote your true value because the net result between what you value and what you pay is optimal.

3. *Rule following*: maybe I want to shade my bid 10%, or overbid, or underbid.

So the outcome of this auction is that the highest value bidder gets the item at the second highest price. Same outcome as above.

Sealed Bid Auction: Each person submits a bid. The highest bigger gets it at the highest bid.

1. *Rational*: shade a little off the price. But the higher you bid, the more likely you are to win. So you need to balance these two forces.

   If V = Value of item, and B = Your Bid, then the net surplus you get is $(V - B)$. If your probability of winning is P for your bid of B, then the winnings are $P(V - B)$. If we have randomly distributed bids between 0 and 1 then the probability of winning is just $P = B$, because if you bid 1 you have a 100% of winning and if you bid 0.5 you have a 50% chance of winning. Therefore your winnings are $B(V - B) = BV - B^2$ and you want to maximise this. Taking derivatives with respect to B and equating to 0 to find the maximum:

   $dW/dB = V - 2B = 0$ such that optimal bid is $B = V/2$. If you think the other person is bidding his/her true value then you want to be bidding half your value.

   But if I am bidding $V/2$, then so should the other person. This means that if the other person is also bidding $V/2$ then my probability of winning goes up to $2B$, because the other person is now bidding half her value and my chance of winning goes up. So now the chance of winning is $2B(V - B) = 2BV - 2B^2$. Taking derivatives and setting to zero we have:

   $dW/dB = 2V - 4B$ such that the optimal bid is $B = V/2$.

   So this is great because if we are both bidding half of what we believe the true value to be, then we will both be doing the optimal thing. It is not a race to the bottom!

2. *Psychological*: might shade to an even number or an uneven number.

3. *Rule following*: shade by a fixed percentage.

So the outcome is that the highest value bidder will get the item at half their value. In fact, if I get it at half my value, and because bids are randomly distributed, the second highest bidder's expected bid is probably at half of what I bid. Because they bid at half their expected value, I basically won the auction at the expected value of the second highest bidder (similar to above but not quite the same).

Revenue Equivalence Theorem: With rational bidders, a wide class of auction mechanisms, including

- sealed bid
- second price

- ascending bid

produce identical expected outcomes, under the assumption of rationality, which is often a good benchmark but…

Of course, we may not have purely rational bidders. What could happen then? In a sealed bid auction people could start to project on to other people trying to figure out how they would behave and create all sorts of weird rules. In a second price auction, people might simply get confused about the rules. In an ascending auction people rarely underbid and often get in a frenzy just because they want to win. So if we are trying to maximise profits of an auction house, then if we have sophisticated rational actors we might stick to a sealed auction, but if we have unexperienced actors we might want to work them into a frenzy in an ascending auction.

18.5 Public Projects

How can we use mechanisms to figure out if we should fund public projects? So should we use collective money to fund a public project that everyone can use?

Example of an office coffee machine: Suppose three people are willing to contribute money to a communal coffee machine

Person 1 is willing to pay \$40
Person 2 is willing to pay \$50
Person 3 is willing to pay \$20

The cost of the coffee machine is \$80 and the collective pot of money contains \$110, which is great.

The problem is that the individual contributions are hidden information, noone knows them before the fact—they are private. People value a coffee machine, but we don't know how much they value the coffee machine. So the challenge is how we get people to reveal their hidden information—incentive compatibility.

Pivot mechanism: You only have to pay the minimal amount you would have to contribute for the project to be viable. For the example above, if other people will contribute \$60 without you, then you pay \$20. If other people would pay \$70 without you would have to pay \$10. This creates an incentive for you to tell the truth, because you will only pay the marginal amount.

Formally, each person claims values V_1, V_2, V_3, *etc.* If $V_1 + V_2 + V_3 >$ Cost, then we do the project and person 1 must pay $\max\{cost - V2 - V3, 0\}$. So either person 1 pays the difference left over, or if nothing is left over he/she pays nothing.

How does this pan out in a situation?

Cost = \$80, Value = \$40

Sum of others	Claimed value	Pay	Net
30	40	Not enough	to buy
40	40	40	0
45	40	35	5
90	40	0	40

So in this case only one time does the project not get done, and one time person 1 doesn't even have to pay.

What if person 1 cheats and only claims $30?

Sum of others	Claimed value	Pay	Net
30	30	Not enough	to buy
40	30	Not enough	to buy
45	30	Not enough	to buy
90	30	0	40

This time the project only happens once. In fact, if we look at the third row, in the first case, person 1 would have come $5 ahead by getting $40 of value for $35 but in the second case when he shaves, he gets nothing. So by shaving his value down he actually gets less. So if you underclaim, the project is not likely to get done.

What if you overclaim?

Sum of others	Claimed value	Pay	Net
30	50	50	-10
40	50	40	0
45	50	35	5
90	30	0	40

This time he is worse off again because in the first row he loses $10 and for all other rows he is even.

So what this pivot mechanism does is induce people to tell the truth. This is what we call incentive compatibility—each person has an incentive to reveal his/her true value. Let's go back to the example. The true value for everyone if buying a $90 machine is as follows:

Person 1: $30
Person 2: $30
Person 3: $30

Person 1: $90 - $30 - $30 = $30
Person 2: $90 - $30 - $30 = $30
Person 3: $90 - $30 - $30 = $30

So if we used the pivot mechanism to reveal people's true value and we got to the above situation, then everyone is paying the right amount. But this doesn't always happen. Let's say the cost is $80 and

Person 1: $40
Person 2: $50
Person 3: $20

Person 1: $80 - $50 - $20 = $10
Person 2: $80 - $40 - $20 = $20

Person 3: \$80 - \$40 - \$50 = 0

The total revenue is only \$30 so we haven't raised enough money. So we need some other incentives to get the people to raise the money.

The four things you want in a mechanism are efficiency, people always joining, incentive compatibility and balance. You can't get all four, so you need to sacrifice one of the four.

19 Replicator Dynamics

Replicator dynamics are interesting because they are used in three different disciplines:

1. *Psychology*: to model learning

2. *Economics*: to model populations of people learning

3. *Ecology*: to model evolution

The basic idea is, imagine that there is a set of types $i = \{1, 2, 3, \ldots, N\}$ and each type has a specific payoff $\pi(i)$. Additionally, different proportions of a population $Pr(i)$ are using (successfully or unsuccessfully) these types. What you then expect is that certain proportions of a population start copying or learning from the most successful types.

In terms of learning, there is a big population out there that is using a set of different strategies, say S_1, S_2, S_3, *etc.* A proportion of the population will be using S_1, S_2 and S_3 maybe at 40%, 40% and 20%. But if you now look at the payoff of these strategies and it turns out that the payoffs are 5, 4 and 6, then the question becomes, which strategy do I use? Is it the more popular/common one, or the one that is most effective, *i.e.* rational? In replicator dynamics we account for both of these factors.

Rather than thinking about strategies, we can replace this with phenotypes in biology, *e.g.* length of tongue. If we have phenotypes P_1, P_2 and P_3 which take 50%, 30% and 20% of the population but have a fitness of 5, 4 and 6, we can again ask, which one of the phenotypes is the best one? The most common or the most fit? So in evolution we see the same dynamics and factors as in learning, and so we can use the same replicator models.

Fisher's theorem: The rate of adaptation is proportional to the variation in that population. This means that you want a lot of variation in order to adapt faster to the optimal strategy/phenotype/characteristics/*etc.* This runs counter to six sigma thinking which states that more variation is worse because it leads to more errors.

19.1 Replicator Equation

Let's look at replicator dynamics in terms of learning. There are a set of types $i = \{1, 2, 3, \ldots, N\}$, which are actions or strategies, with an associated payoff for each of $\pi(i)$ and a specific proportion in the population of $Pr(i)$.

So how do people learn? One strategy would just be to copy what other people are doing. If you copy, then you would copy taking into account the proportion of the strategy in the wider population. Another way would be to hill climb, and this is intuitively connected to the payoff of each strategy rather than the proportion.

Let's suppose we are rational. Then we would choose the highest payoff. But if we suppose that we are more sociological minded, then we would copy someone else, especially what most people are doing, because they wouldn't be doing this if it wasn't working.

Replicator dynamics accounts for both. The way we do this is it to apply a weight to $\pi(i)$, the payoff, and $Pr(i)$, the proportion. Possible weight equations to use are:

weight $= \pi(i) + Pr(i)$
weight $= \pi(i) \times Pr(i)$

The weight we will use is the latter equation, because for addition, if the payoff is 0 and we have some non-zero proportion of the population using a certain type, then people would still be using it, and vice versa. This isn't balancing the two forces. In the second case, if either of the two factors is 0, then the weight is 0 and noone would be using it.

The equation we get from this is:

$$Pr_{t+1}(i) = \frac{Pr_t(i) \times \pi(i)}{\sum_{j=1}^{N} Pr_t(j) \times \pi(j)}. \tag{27}$$

The probability that you play at time $t+1$ is the ratio of your weight at time t and the sum of all weights of all participants at time t.

Example: Payoffs are (2,4,5) and proportions are (1/3,1/6,1/2). So: $w_1 = 1/3 * 2 = 4/6$, $w_2 = 1/6 * 4 = 4/6$ and $w_3 = 1/2 * 5 = 13/6$, with a sum $w_1 + w_2 + w_3 = 23/6$. $Pr(1) = 4/6/23/6 = 4/23 = Pr(2)$; $Pr(3) = 15/6/23/6 = 15/23$. Note that $Pr(1) + Pr(2) + Pr(3) = 1$ as it should be.

We now want to apply this to games. For example the shaker-bower game:

	S	S
S	2,2	0,0
B	0,0	1,1

$Pr = (1/2, 1/2)$ and $\pi = (1/2*2+1/2*0, 1/2*1+1/2*0) = (1, 1/2)$, where shaking goes first. So $w_S = 1/2 * 1 = 1/2$ and $w_B = 1/2 * 1/2 = 1/4$. So in the next period: $Pr_S = 1/2/3/4 = 2/3$ and $Pr_B = 1/4/3/4 = 1/3$. So we started with an equal number of shaker and bowers, after one period we have 2/3 proportion of shakers. If we run this enough times we would end up with all shakers.

So now we have two models pointing us in the same direction. A rational actor would obviously choose shaking straight away because the payoff is higher, but a replicator model that adds some sociology to it gives us the same answer. So this strengthens our initial result from rationality.

Let's play another game, the SUV-Compact game:

	SUV	C
SUV	2,2	2,0
C	0,2	3,3

From rationality you would assume that people just choose the compact because it has the highest payoff. It is also an equilibrium because the more people use compacts the more certain I am I will get the highest payoff.

Let's start out with $Pr(i) = (1/2, 1/2)$. $\pi(i) = (1/2*2+1/2*2, 1/2*0+1/2*3) = (2, 3/2)$. So $w_{SUV} = 1/2*2 = 1$ and $v_C = 1/2*3/2 = 3/4$. $P_{SUV} = 1/(1+3/4) = 4/7$; $P_C = 3/4/(1+3/4) = 3/7$.

So what we get in this game is that we get more SUVs in the long run. So we are more likely to get an outcome of 2,2. Hence evolution leads to a suboptimal outcome.

19.2 Fisher's Theorem

Replicator dynamics is the idea that the proportion of people playing a particular action at time $t+1$ depends on the proportion playing at time t and the payoff. We can now move from applications of learning, where there was a payoff to each action, to one of evolution, where there is a fitness.

So in the replicator model there is a set of types $i = 1, 2, 3, \ldots, N$ and the fitness of each type is $\pi(i)$ with a proportion at time t of $Pr(i)$. We can now think through the logic of how many of a certain type there will be in the next population, which depends on the fitness of the trait and proportion of the trait in the population.

One way to think about this is using a fitness wheel, which is just the classic "spin-a-wheel" sort, with all the different types on the wheel as individual wedges. The size of a wedge determines its fitness and the number of wedges determines the proportion of that trait in the population.

Fisher's theorem: This theorem allows us to combine the following models

1. *Model 1*: there is no cardinal bird. The variation in the cardinal species is so high, that you can't define one unique set of traits that define the cardinal.

2. *Model 2*: rugged landscape—the idea that when you are hill climbing to optimise, the design space is typically multi-peaked meaning that you can get stuck in local optima.

3. *Model 3*: replicator dynamics.

and also the role that variation plays in adaptation. So given that there is no one unique trait in a species and that we can have these different traits on different locations on a multi-peak design space, what replicator dynamics will do is to choose those traits that are most common at the highest fitness level.

Fisher's theorem then states that higher variance leads to faster rates of adaptation. The main idea is that the more variation there is the more likely it is that I can copy something that is better than what I already have.

Example—low variation: 1/3 with fitness 3, 1/4 with fitness 4 and 1/3 with fitness 5, giving an average fitness of 4. Let's calculate the weights: $w_1 = 1/3*3 = 1$; $w_2 = 1/3*4 = 4/3$; $w_3 = 1/3* 5 = 5/3$. Sum weights: 1 + 4/3 + 5/3 = 12/3. $Pr(1) = 1/12/3 = 3/12$; $Pr(2) = 4/3/12/3 = 4/12$; $Pr(3) = 5/3/12/13 = 5/12$. So the new average is $3 * 3/12 + 4 * 4/12 + 5 * 5/12 = 41/6$. So the average fitness has gone up by 1/6.

Example—medium variation: 1/3 with fitness 2, 1/4 with fitness 4 and 1/3 with fitness 6, giving an average fitness of 4 but with higher variance. Let's calculate the weights: $w_1 = 1/3 * 2 = 2/3$; $w_2 = 1/3 * 4 = 4/3$; $w_3 = 1/3 * 6 = 6/3$. Sum weights: 2/3 + 4/3 + 6/3 = 12/3. $Pr(1) = 2/3/12/3 = 2/12$; $Pr(2) = 4/3/12/3 = 4/12$; $Pr(3) = 6/3/12/13 = 6/12$. So the new average is 2*2/12 + 4*4/12 + 6*6/12 = 4 4/6. So the average fitness has gone up by 4/6.

Example—high variation: 1/3 with fitness 0, 1/4 with fitness 4 and 1/3 with fitness 8, giving an average fitness of 4 but with even higher variance. Let's calculate the weights: $w_1 = 1/3 * 0 = 0$; $w_2 = 1/3*4 = 4/3$; $w_3 = 1/3*8 = 8/3$. Sum weights: 0 + 4/3 + 8/3 = 12/3. $Pr(1) = 0/12/3 = 0$; $Pr(2) = 4/3/12/3 = 4/12$; $Pr(3) = 8/3/12/13 = 8/12$. So the new average is 0*0 + 4*4/12 + 8*8/12 = 6 2/3. So the average fitness has gone up 2 2/3 = 16/6.

Summary: So clearly the highest gain is with the highest variation, and this seems to increase monotonically.

First: $(3-4)^2 + (4-4)^2 + (5-4)^2 = 2$
Second: $(2-4)^2 + (4-4)^2 + (6-4)^2 = 8$
Third: $(0-4)^2 + (4-4)^2 + (8-4)^2 = 32$

So both the variation and the gain increased by a factor of 4 each time, meaning that a change in variation leads to a proportional amount of change in adaptability. In fact, this is Fisher's

Fitness	Gain	Variation
(3,4,5)	1/6	2
(2,4,6)	4/6	8
(0,4,8)	16/6	32

fundamental theorem: The change in average fitness due to selection (if we have replicator dynamics) will be proportional to the variance.

More variation being better runs counter to six sigma thinking where you are saying you want less variation to reduce errors. So this is a classic example where two models pull you in a different direction, meaning you need to consider both and then decide which effect is the dominant one.

20 Multi-model Thinker

20.1 Variation or Six Sigma

Fisher's fundamental theorem says the more variation, the faster you can adapt. This runs counter to six sigma, which states that if you have an optimal design you want low amounts of variation from that optimum.

So this is an example of opposite proverbs: never too old to learn *vs* you can't teach an old dog new tricks; the pen is mightier than the sword *vs* actions speak louder than words, *etc.*

So how do we make sense of these things when they contradict? We need to go back and study the assumptions that underlie the model. We do this to figure out in which scenarios the assumptions are valid. This is obvious if you think about the no free lunch theorem—there is never one variable of optimisation that will work in all cases, and neither will any model.

One way to think about this is with a landscape. If the landscape is fixed and I am sat on the optimum, then obviously I don't want to move. In production and manufacturing, the things we are producing are fixed (for now) and so we want to manufacture these with the least variation.

However, if the underlying landscape moves, as is the case in nature, where the ecological environment is always changing, the sitting at an optimum doesn't mean much when the environment changes and you are no longer at that optimum. Now you want to adapt as quickly as possible to the new optimum, and the way to do that is with large variation.

Fisher's theorem is about dynamics, nature churning, *etc.* This means that if we know that the landscape is evolving we need to foster variation (nature, business, marketing strategy, product innovation, *etc*).

Six sigma is about a stable equilibrium. This means that if we know the landscape is fixed, then we just want to optimise and remove all variation (getting airplanes of the ground, preventing clinical mistakes, *etc*).

20.2 Prediction

Let's go back to the categorical models and linear models we looked at before and use them for prediction.

Suppose I wanted to predict the height of the Saturn V rocket. There are multiple ways of thinking about this analogously. The rocket is like a water tower, the statue of liberty, the Empire State Building, *etc.* So what I am doing is using categories to try and predict or guess a property about something else or related. Essentially, the categories allow us to make better predictions because they reduce the variance of all possible options. If I take desserts and categorise them into fruits and cake, I will probably be better able to predict the calories of a specific dessert you give me, simply because I know that on average fruit has less calories than cake. Of course when different people use different categories they will make different predictions, and when we combine different categories we are more likely to make a good prediction.

Alternatively, we can think in terms of a linear model. In a linear model we assume that the outcome is a linear combination of a number of different properties multiplied by weighting factors. For example, the calories in a subway are just a sum of all the constituents (there is no weighting factor here).

In both cases we measure our model in terms of the percentage of variation away from the mean explained by the model. We'll now extend that to predictions of the crowd, so a bunch of categorical or linear models acting together.

20.3 Diversity Prediction Theory

We want to think of individual people making predictions based on some model: categorical, linear, Markov, Lyapunov. Then how do crowds of people make predictions, and by combining all these different models, can we make better predictions? Also, what does it mean for a crowd to make a mistake, and what does it mean for an individual to make a mistake?

This leads to the diversity prediction theorem, which tries to predict how accurate crowds will be at predicting. The accuracy of the crowd is obviously dependent on the accuracy of individuals, but counterintuitively, it also depends on the diversity of viewpoints and models in the crowd.

Crowd accuracy = individual accuracy + diversity

Example: Amy predicts 10, Belle predicts 16, Carlos predicts 25, giving an average prediction of 17. But the actual value to be predicted is 18. How accurate are these people then? To determine this we use variations:

Amy: $(10-18)^2 = 64$
Belle: $(16-18)^2 = 4$
Carlos: $(25-18)^2 = 49$

This gives an average variation of 39. So how accurate was the crowd with its average prediction of 17? Taking variations again we get: $(17-18)^2 = 1$. So the crowd is by far better than any individual in it. This is the wisdom of crowds. Where does this come from? Diversity... and we will calculate diversity by comparing the prediction of each person to crowd's mean prediction:

Amy: $(10-17)^2 = 49$
Belle: $(16-17)^2 = 1$
Carlos: $(25-17)^2 = 64$
And the average variation (diversity): 38
Crowd's error: 1
Average error: 39

Diversity prediction theorem: Hence, we can see that: Crowd's error = Average error - Diversity.

This is not only true for this canonical example, but a mathematical identity called the diversity prediction theorem. It is always true. Formally speaking we write this is follows:

$$(c-\theta)^2 = \frac{1}{n}\sum_{i=1}^{n}(s_i-\theta)^2 - \frac{1}{n}\sum_{i=1}^{n}(s_i-c)^2 \tag{28}$$

where c is the average prediction of the crowd, θ is the true value, s_i are the individual guesses of person i, and n is the total number of people. If you take this expression and expand everything out, you will see that loads of terms cancel out and we are left with the identity above.

As an example, consider a cow weight guessing contest. There is data in the literature where the crowd error in guessing a 1100 lb cow was 0.6. But the average error of the individuals was actually 2956 (not that high because $50^2 = 2500$ so individuals can reasonably guess the weight of a cow). But because the diversity was also 2955.4 this meant that the crowd error dropped to 0.6. So the reason why the crowd was so accurate was because the individual error was reasonable but we had loads of diversity too, which was the crucial thing.

Let's break this down a bit more by going back to the equation:

Crowd's error = Average error - Diversity.

You can only have the "wisdom of crowds" if two things are true.

1. The diversity needs to be high because this subtracts from the average error (otherwise we would have the madness of crowds)
2. The average individual error needs to be high or moderately high, because if it were small, then individuals wouldn't need the crowd to make better predictions.

In a punchline, the wisdom of crowds comes from reasonably smart people with loads of diversity (high to moderate individual error, high diversity). Conversely, the madness of crowds comes from like-minded people being all wrong (high individual error, low diversity).

So the crucial question is, how do you get that diversity? By having people that are using different models, different variables, different weightings, different viewpoints. This diversity is what leads to the wisdom of crowds.

20.4 Many Model Thinker

So let's recap what we want to model and what types of models we have learned about:

1. Intelligent citizen of the world
   - Growth models showed that countries can get really high growth rates just by investing in capital, but at some point, you hit the point of diminishing returns, and to get past the frontiers of what is known you need innovation. Innovation also has another effect that it now makes sense to invest even more in capital.
   - Colonel Blotto game: Made us understand how in some strategic decisions it makes sense to add more dimensions, especially as an underdog.
   - Markov Models: Processes with fixed transition ratios always converge to a specific statistic equilibrium, which means interventions only create lasting effects if we can change those transition probabilities. Intervening in the state doesn't really matter, which also suggests that history doesn't matter.
   - Segregation models: when things aggregate, the outcome can be vastly different. You can't deduce the macro from the micro.
2. Clearer thinker
   - Tipping points: If you see a kink in the graph it might just be exponential growth and not a tip. Tips occur when the likelihood of something happening changes drastically over time. This can either be due to changes in the underlying variable or because of changes in the environment.
   - Path dependence: Gradual changes when a process unfolds, rather than the sudden changes in tips.

In general, by using SRI models, percolation models, Markov models, tipping points, *etc.* we can figure out why some processes look linear and why other have an S-curve or hockey-stick curve.

3. Understand and use data

    - Categories and linear models which can be combined into prediction models by using the wisdom of crowds.
    - Markov models: finding out who the author of a book is.

4. Decide, strategies and design

    - Prisoners' dilemma, collective action problems.
    - Mechanism design: how do we design institutions and contracts so that we get what we want. We need to align the incentives with people's behaviour, so-called incentive compatibility. Sometimes we can't get the ideal solution because there are competing drivers, and sometimes it doesn't really matter what we do, anything will work.

In any of these models you always need to be aware that people can either be rational, they can use rules or they can be psychologically biased. We need to be aware of these quirks, which often means starting from a rationally based model and then checking how this would break down because of biases.

So the bottom line is that models thinking is a crutch for better thinking!