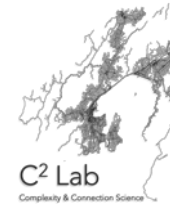


From the complexity of stories
to questions about cause and effect.

Markus Luczak-Roesch | @mluczak



<http://complexity.sim.vuw.ac.nz>



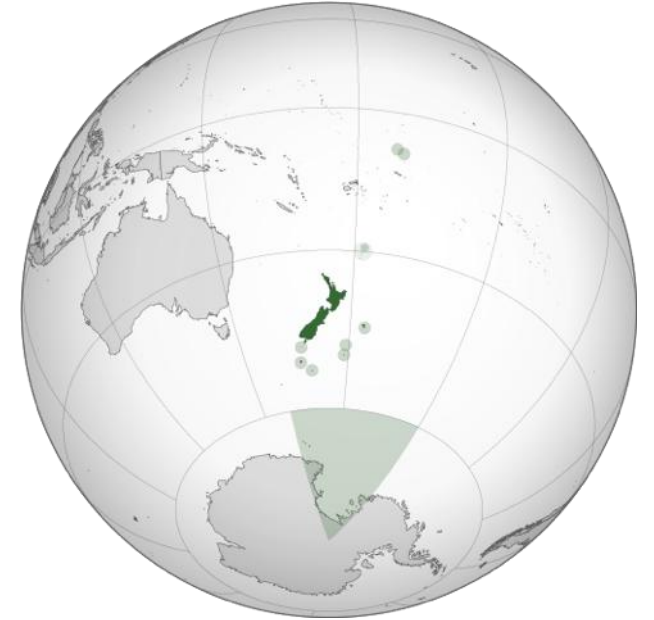
About me: A life in three worlds



- 2008-2010 graduate research associate at FU Berlin (DE), Project Corporate Semantic Web
- 2010-2013 lecturer at FU Berlin (PhD thesis submitted Aug 2013, defended Jan 2014), thesis title: Usage-dependent Maintenance of Structured Web Data Sets



- 2013-2016 senior research fellow at University of Southampton (UK), project SOCIAM – The Theory and Practice of Social Machines (<http://sociam.org>)



- since July 2016 faculty at Victoria University of Wellington (NZ)
- since 2019 associate investigator at Te Pūnaha Matatini

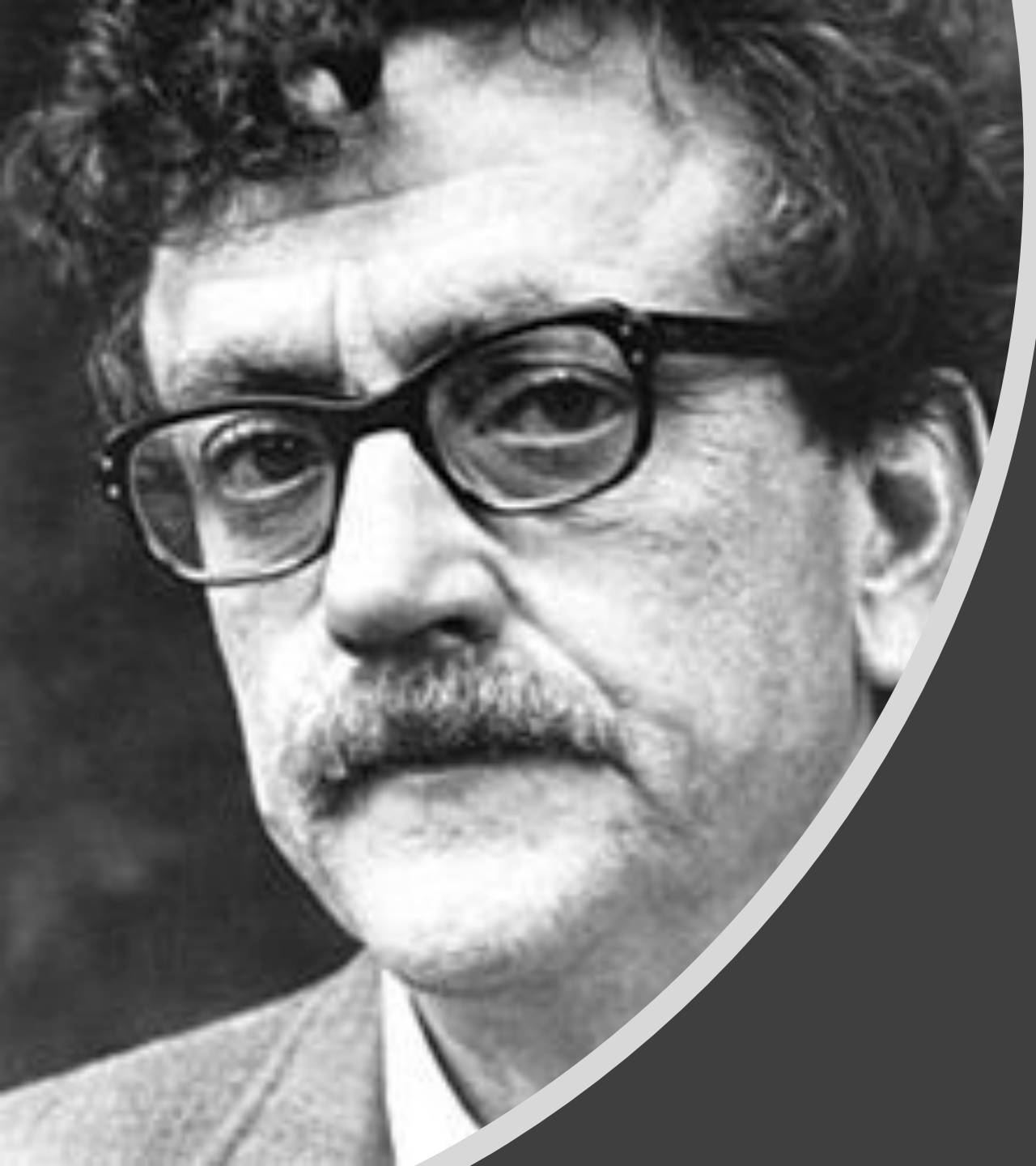
In this lecture I will talk about

- literature,
- philosophy,
- number theory,
- complex networks, and
- things I haven't even fully understood myself yet.

Acknowledgements

- Adam Grener
- Ronald Fischer
- Joahnnes Karl
- Emma Fenton
- Tom Goldfinch
- Isabel Parker
- Kieron O'Hara
- Ramine Tinati
- Te Pūnaha Matatini
- The team at the Complexity and Connection Science Lab
- My wife Johanna and our wonderful children Levi, Yael and Mili





“I have tried to bring scientific thinking to literary criticism and there has been very little gratitude for this.”

Kurt Vonnegut -
https://www.youtube.com/watch?v=GOGr_u_4z1Vc

G

B

E

I

Good fortune



B

E

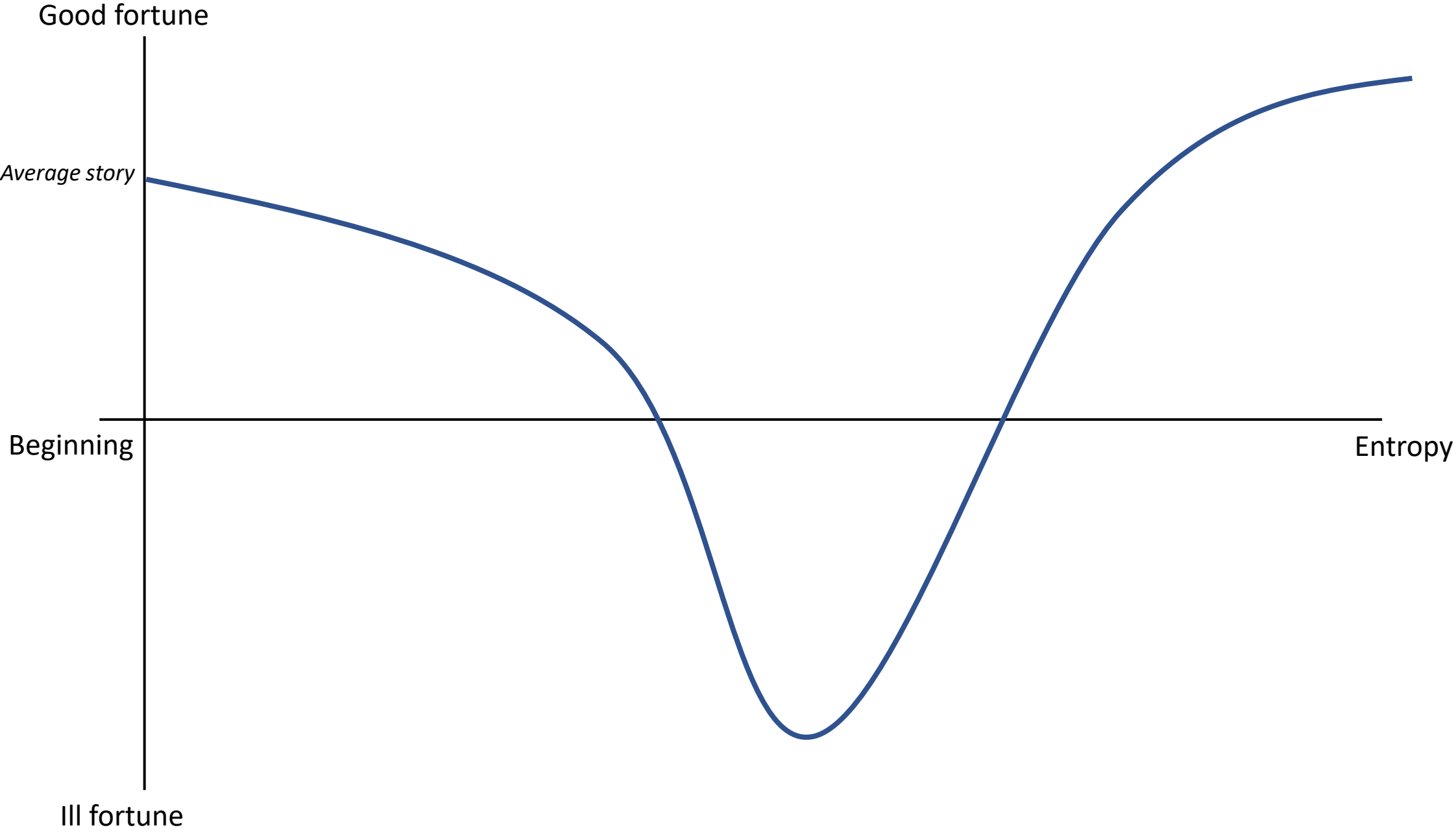
Ill fortune

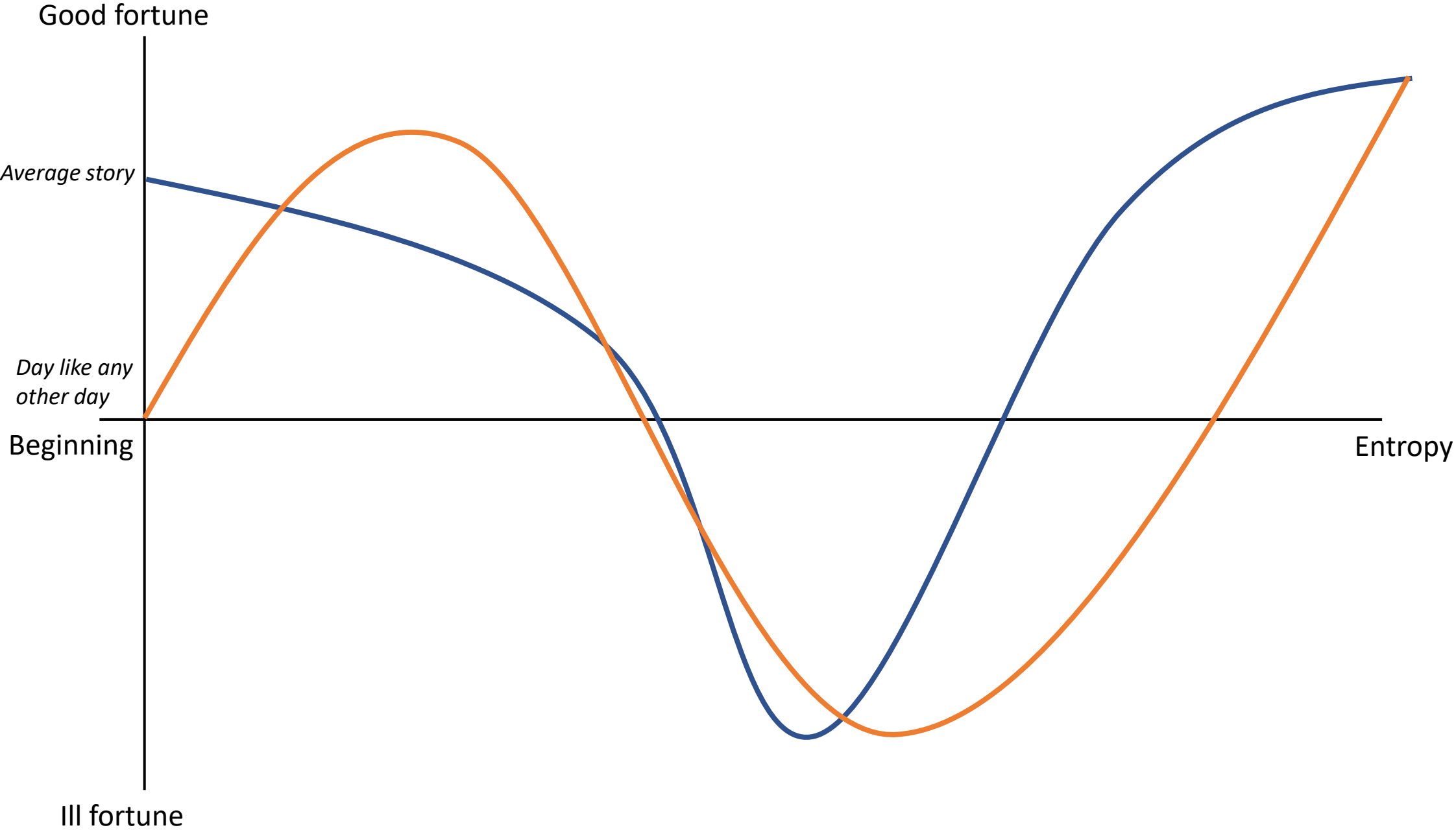
Good fortune

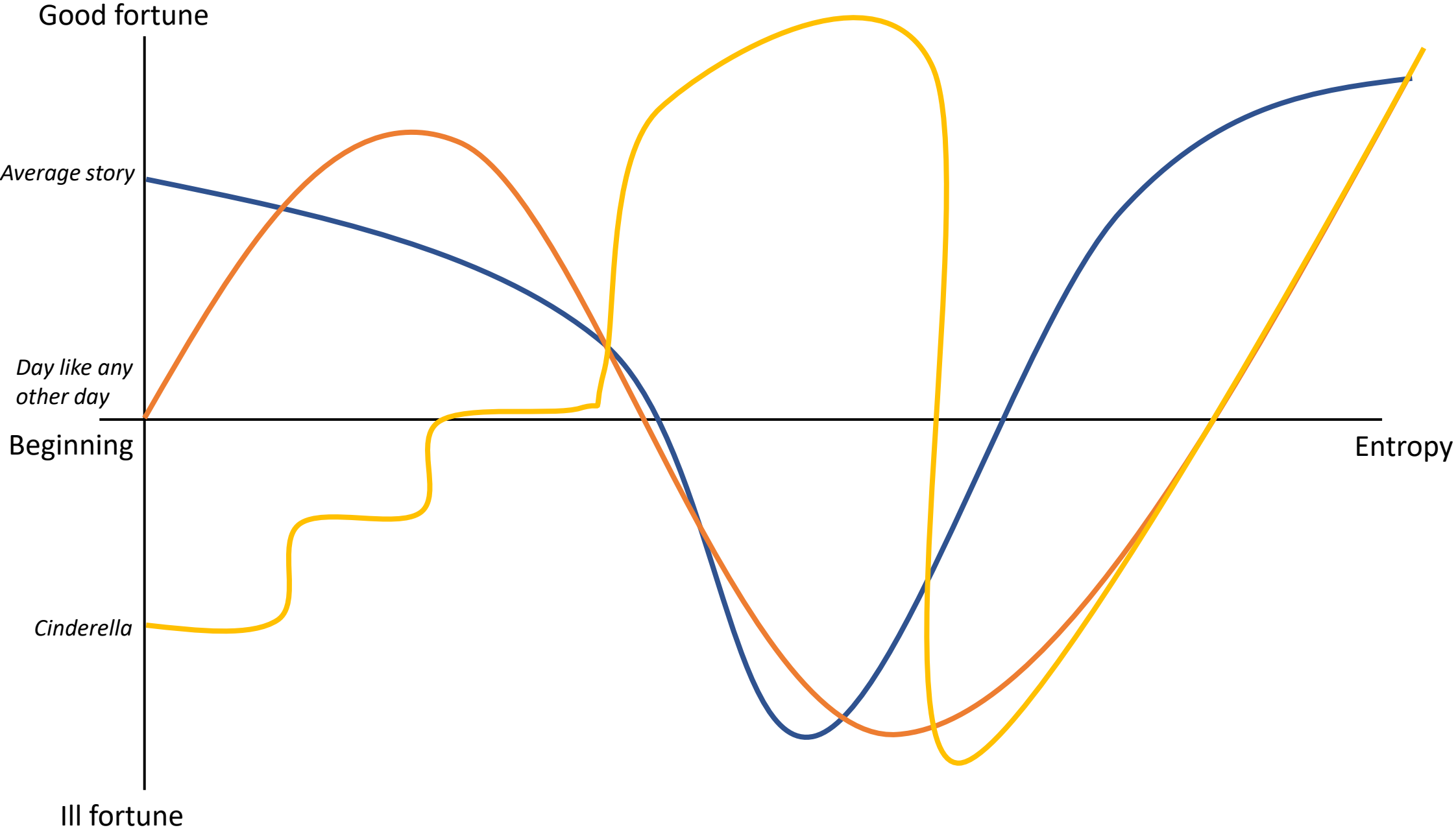
Beginning

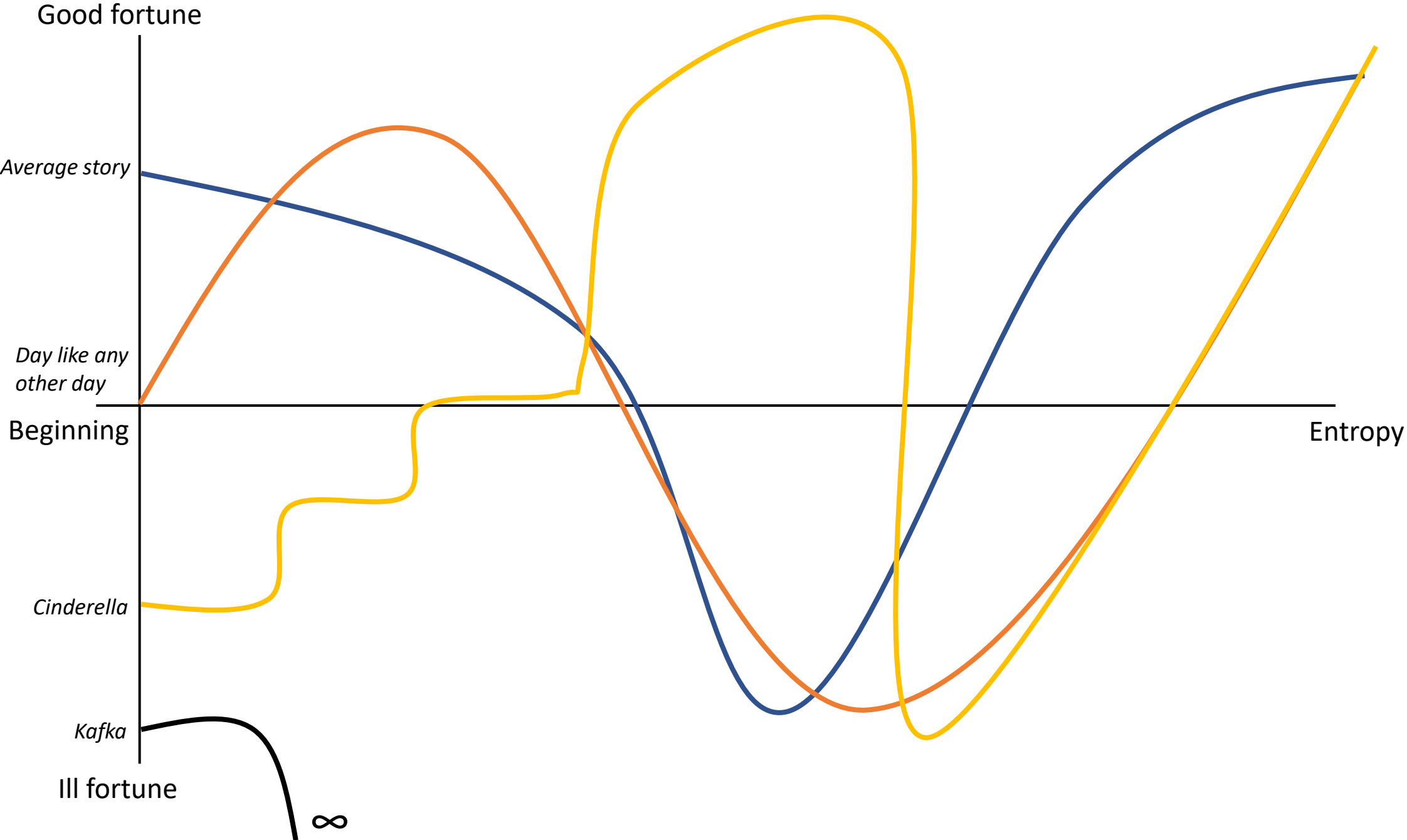
Entropy

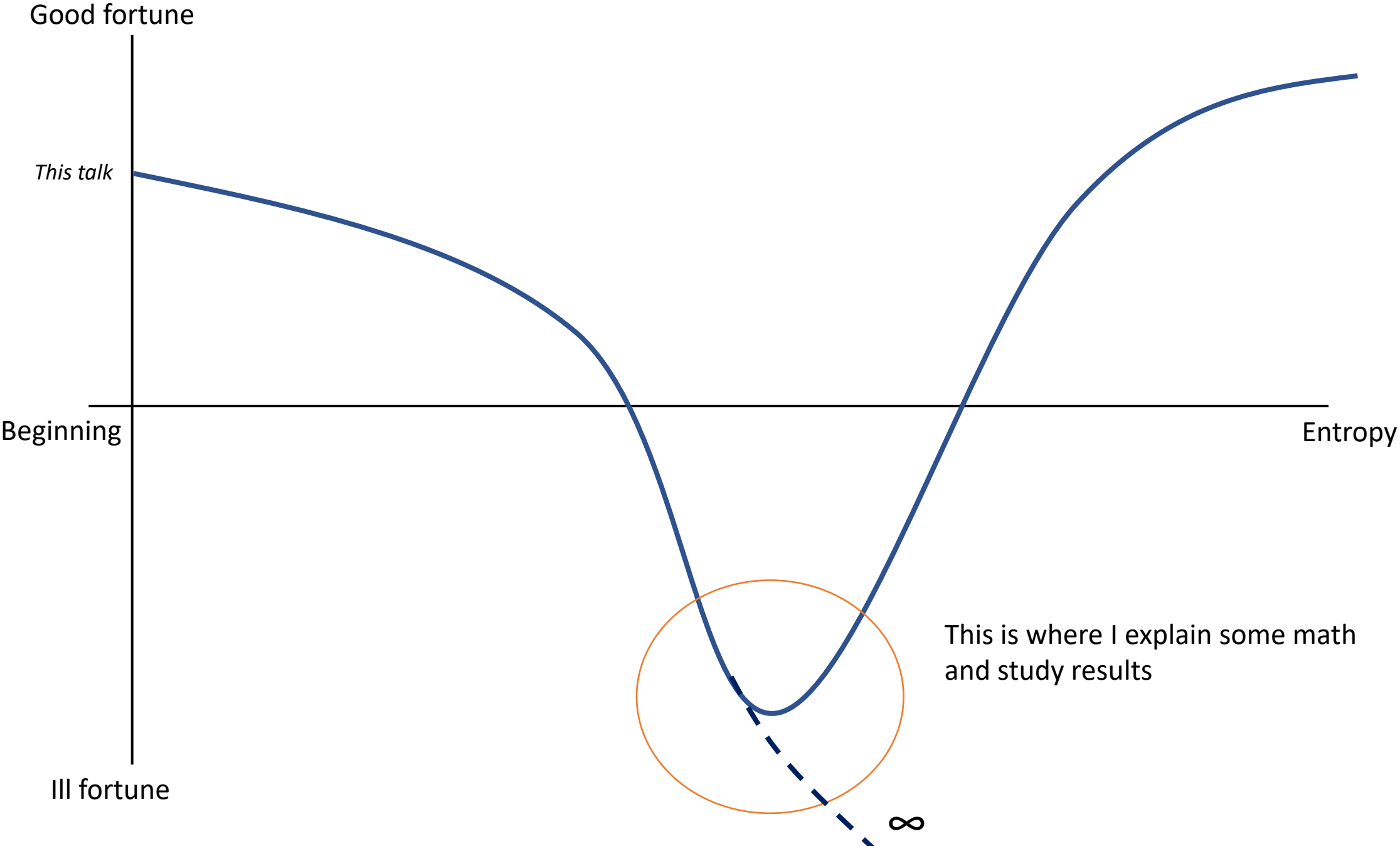
Ill fortune



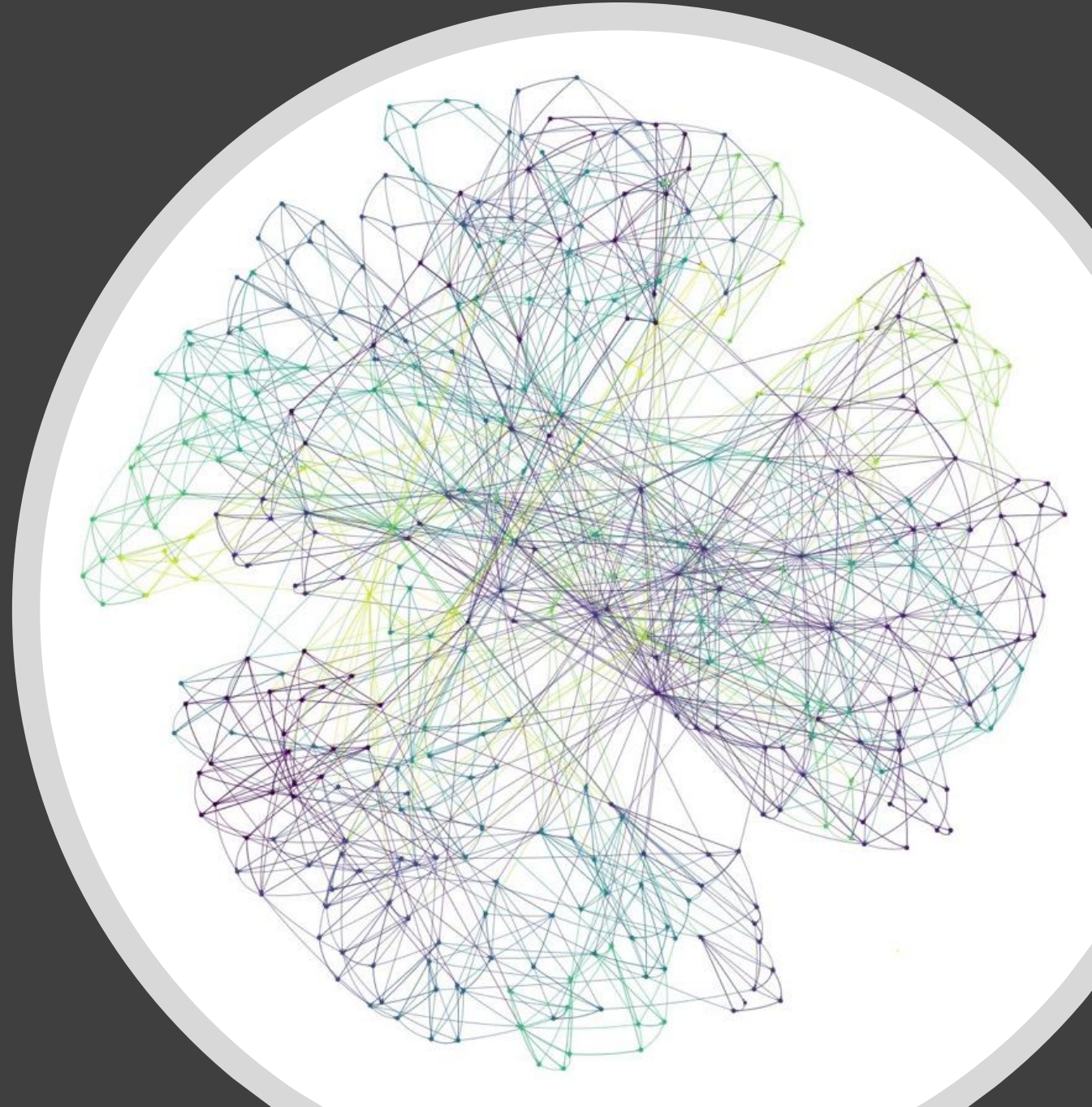




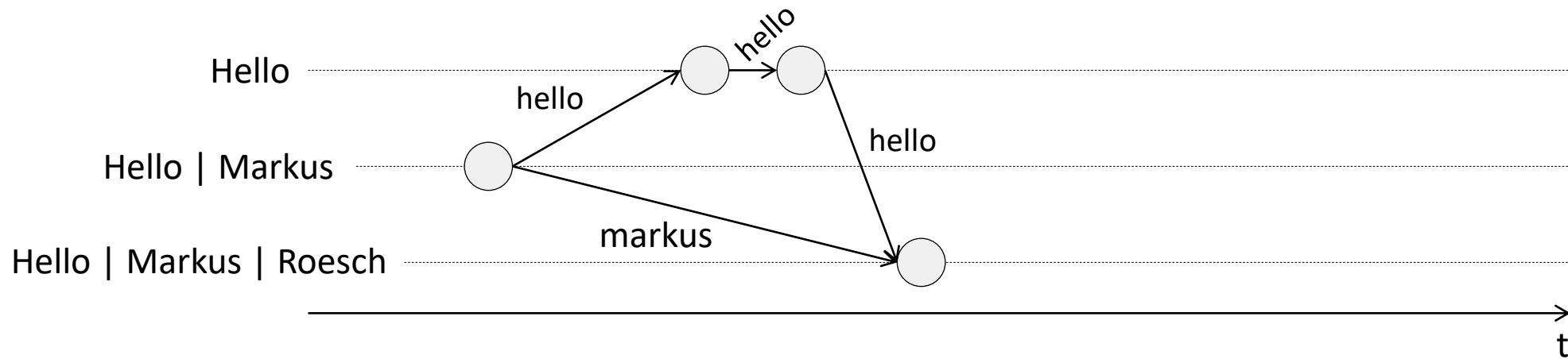




Stories are
complex systems



Complexity arises from the actions and interactions of the **many things that are going on in our world at the same time**. We try to capture this complexity and make it *easily* understandable using “Transcendental Information Cascades”.



What are Transcendental
Information Cascades?
From **Π** to Popper to primes



“My goal is simple. It is a complete understanding of the universe, why it is as it is and why it exists at all.”

Stephen William Hawking
January 1942 - 14 March 2018



“My goal is simple. It is a complete understanding of the universe, why it is as it is and why it exists at all.”

Stephen William Hawking
January 1942 - 14 March 2018



“My goal is simple. It is a complete understanding of the universe, why it is as it is and why it exists at all.”

Stephen William Hawking

January 1942 - 14 March 2018 →

Albert Einstein's birthday

→ π day ($\pi = 3.14\dots$)

← Coincidence

Coincidence

- events that are temporally related but have no observed causal relationship
- C.G. Jung coined the term “synchronicity” for cases of “acausal but meaningful coincidences”
 - Underpins his theory of the collective unconscious
 - Obscure theory of parapsychology or fundamental to our understanding of the mind?



Diaconis and Mosteller on coincidence^[9]

- directions for a general theory of coincidences are
 - hidden cause
 - psychology
 - Multiple Endpoints and the Cost of "Close"
 - The Law of Truly Large Numbers

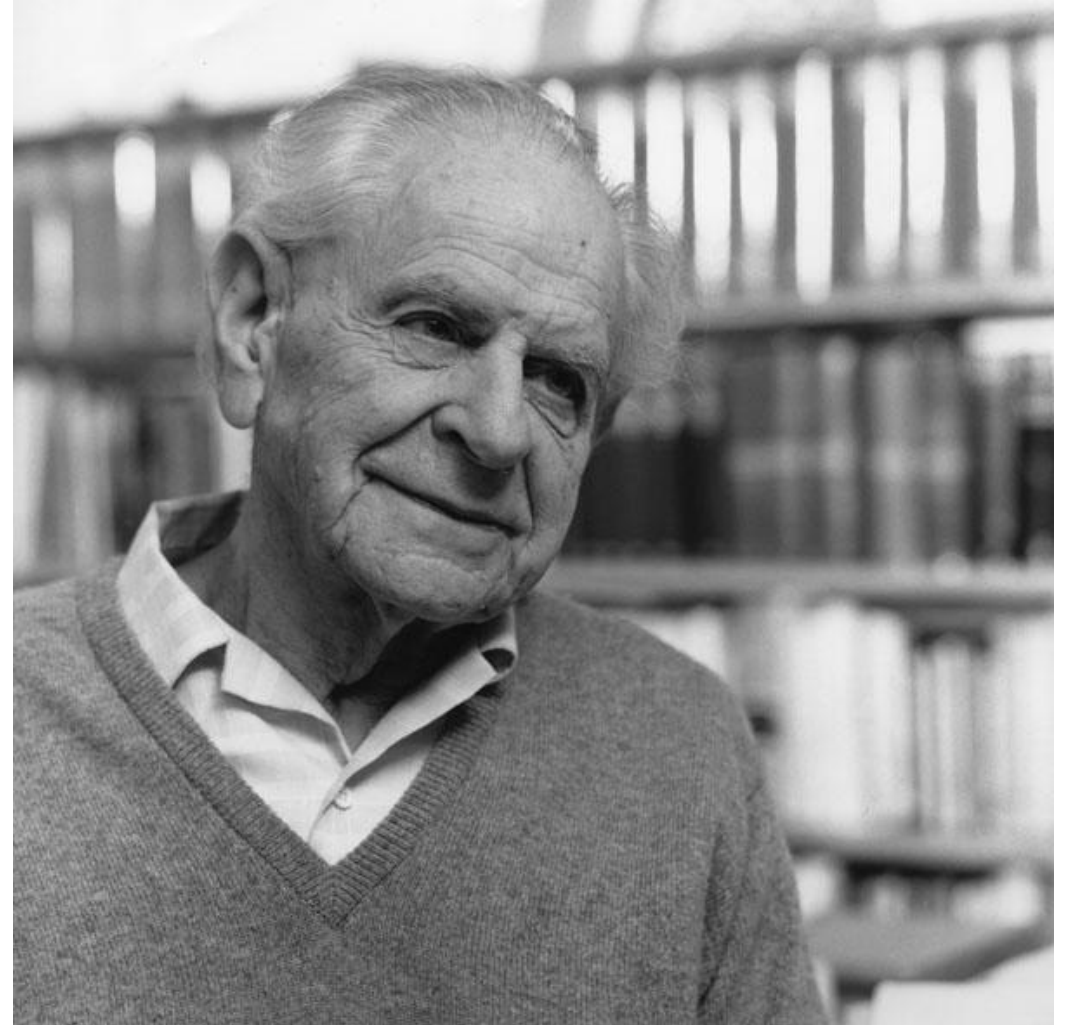
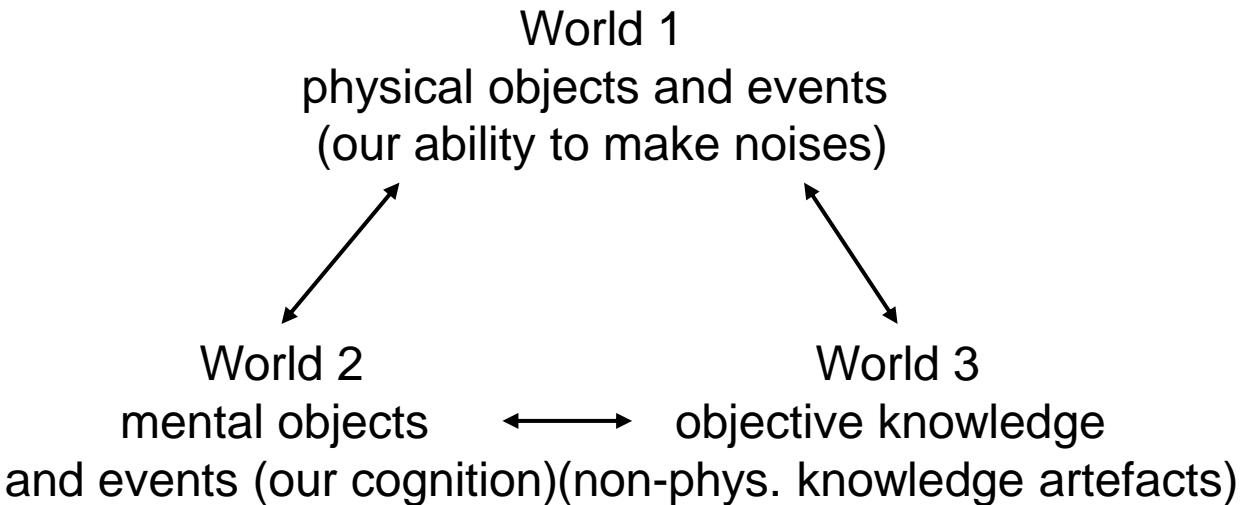
[9] Diaconis, P. and Mosteller, F., 2006. Methods for studying coincidences. In Selected Papers of Frederick Mosteller (pp. 605-622). Springer, New York, NY.

Diaconis and Mosteller on coincidence^[9]

- directions for a general theory of coincidences are
 - hidden cause
 - ~~psychology~~
 - ~~Multiple Endpoints and the Cost of "Close"~~
 - The Law of Truly Large Numbers
- but they emphasized
 - “[...] we are handicapped by lack of empirical work. We do not have a notion of how many coincidences occur per unit of time [...]”

[9] Diaconis, P. and Mosteller, F., 2006. Methods for studying coincidences. In Selected Papers of Frederick Mosteller (pp. 605-622). Springer, New York, NY.

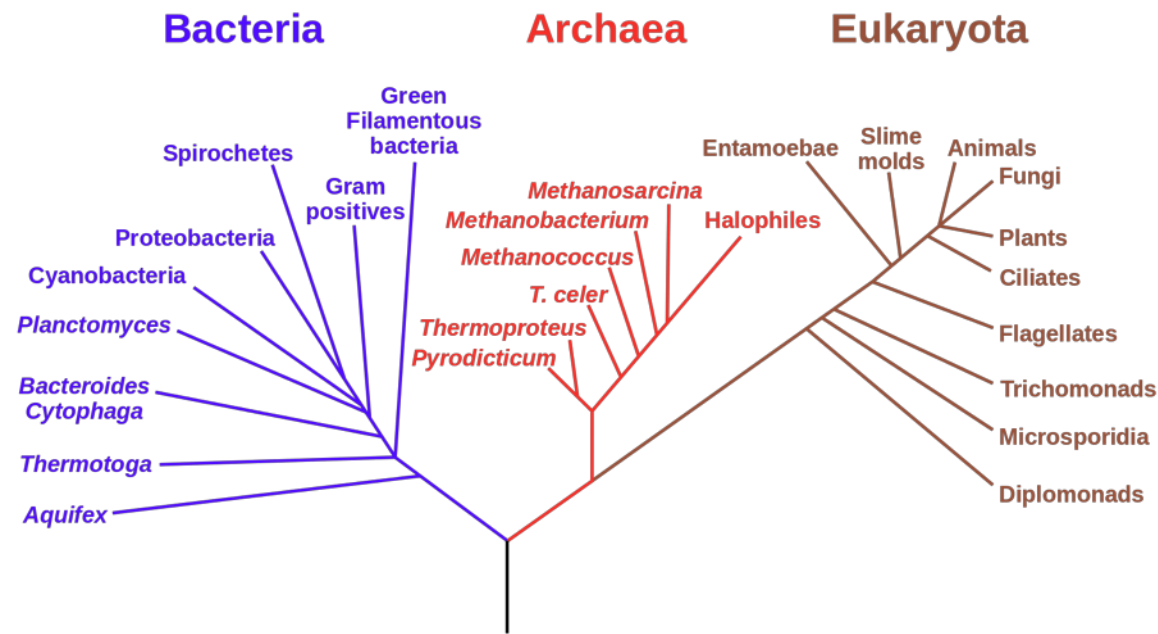
A story of three worlds to unpack acasuality^[1]



[1] Popper, K., 2013. Knowledge and the Body-Mind Problem: In defence of interaction. Routledge. (original lecture in 1969)

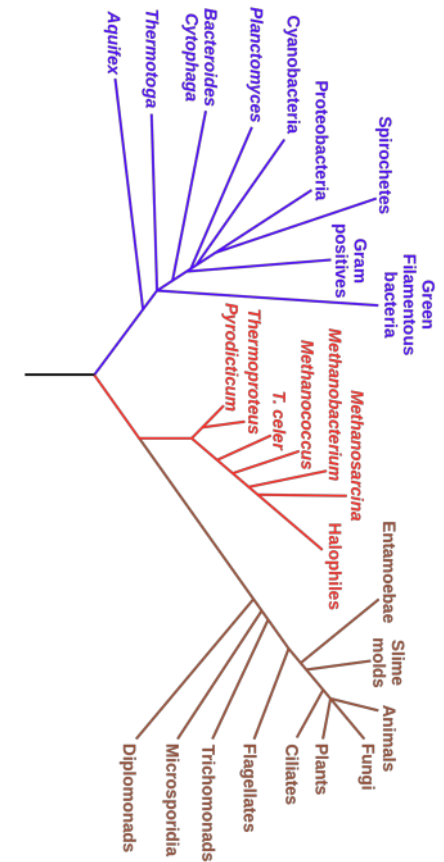
Evolutionary emergence

Phylogenetic Tree of Life

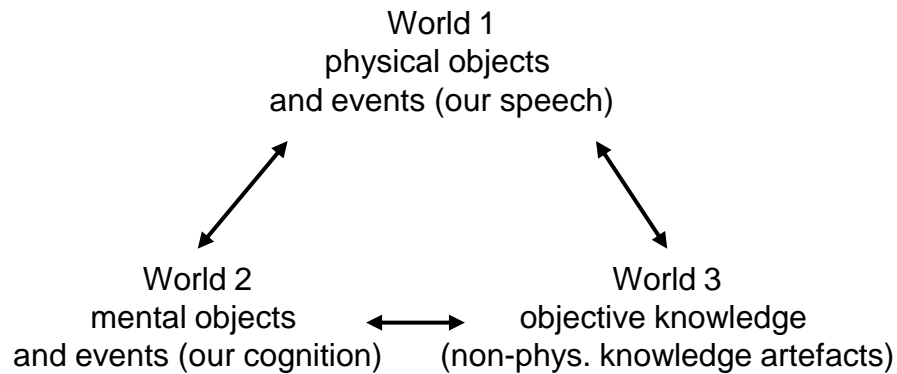


Evolutionary emergence

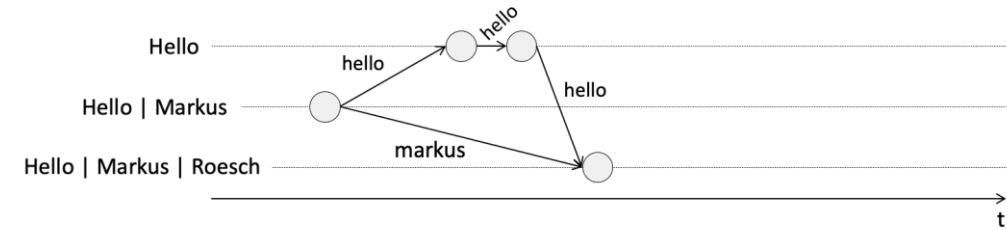
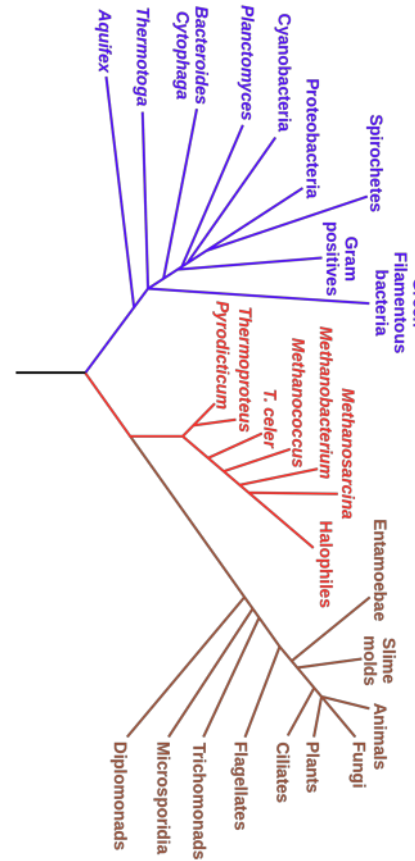
- Popper suggests the view that evolution means “ascending”
in the genetic tree misses the point
 - what’s optimal under the conditions at one point may become suboptimal when the conditions change
 - instead of “ascending into higher forms” it is “increased variety” that should be regarded



Evolutionary emergence and TICs



“a transcendental method in Kant's sense of attempting to understand the conditions of knowledge itself”



prime

composite

•• 2

••• 3

4

••
••

••••• 5

6

•••
•••

••••••• 7

8

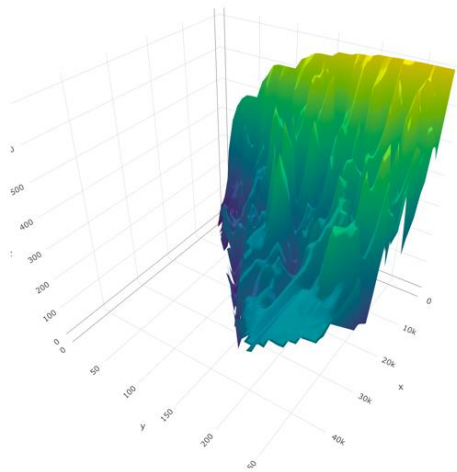
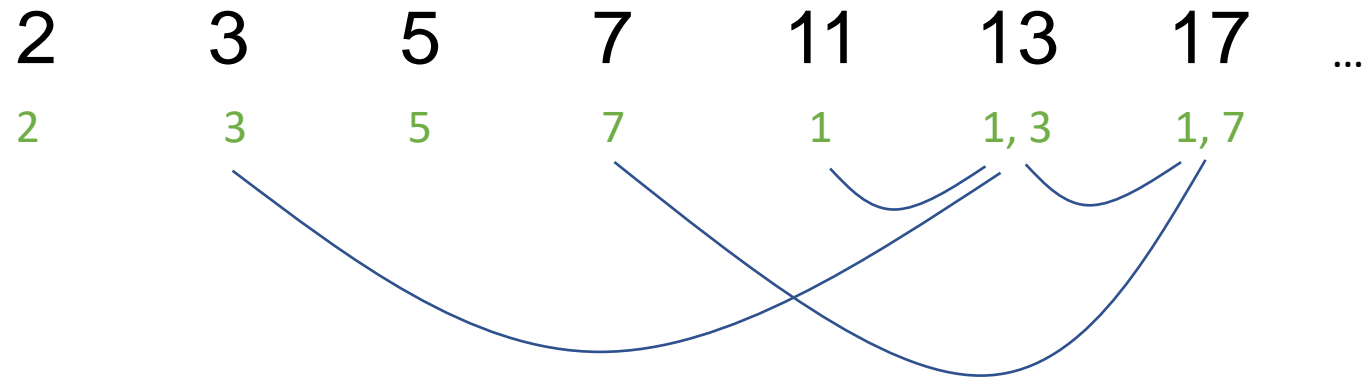
•••
•••

9

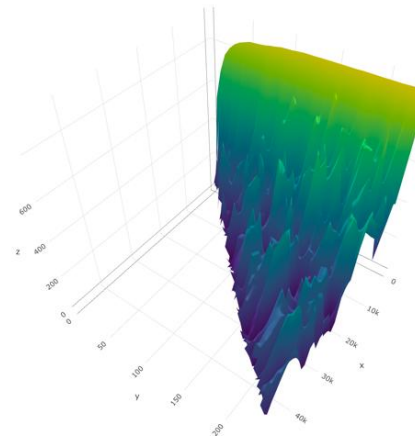
•

Randomly
meaningful

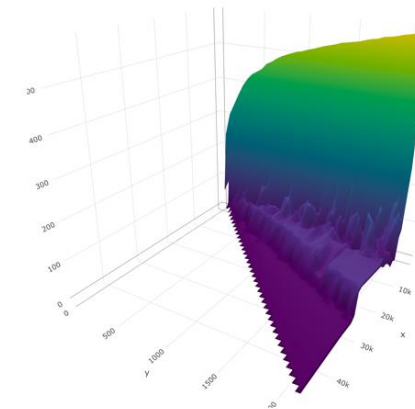
Primes, meaningfully non-random but acausal.



Primes

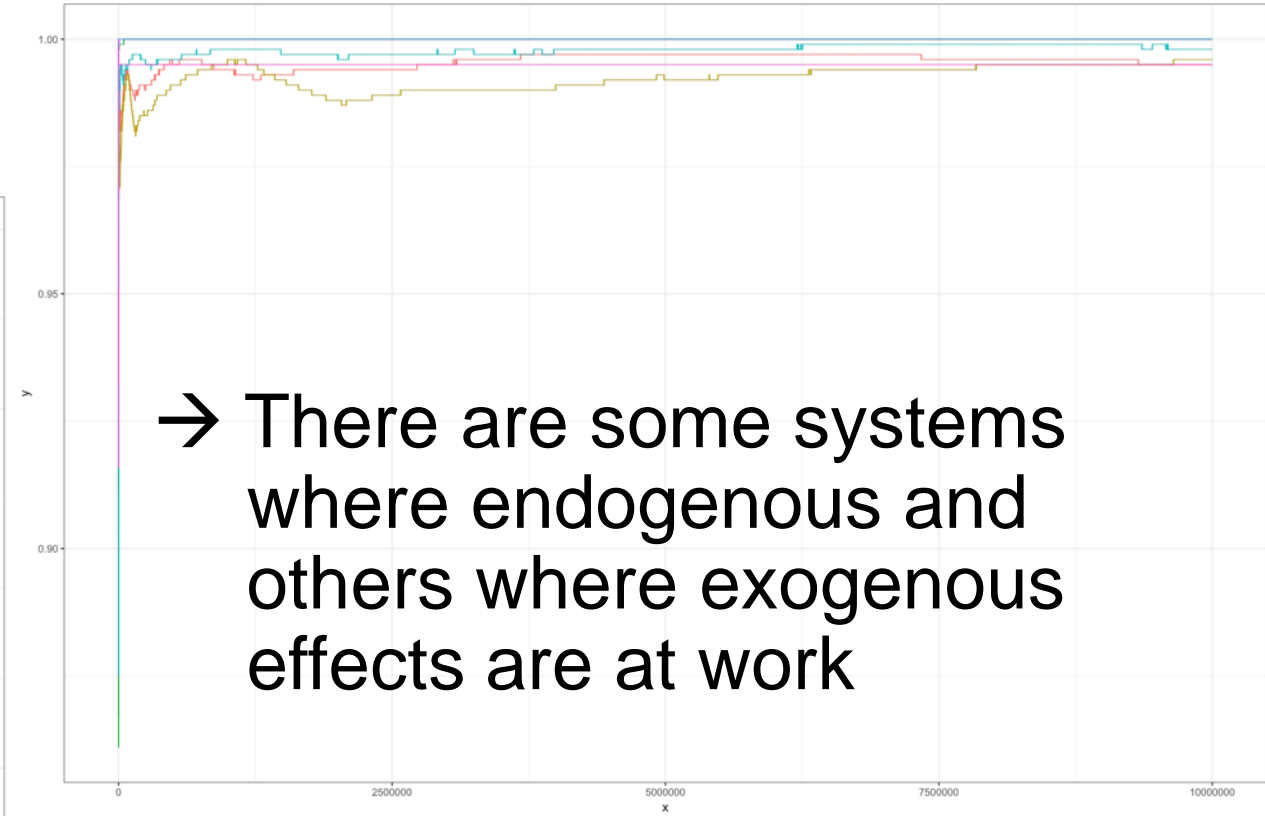
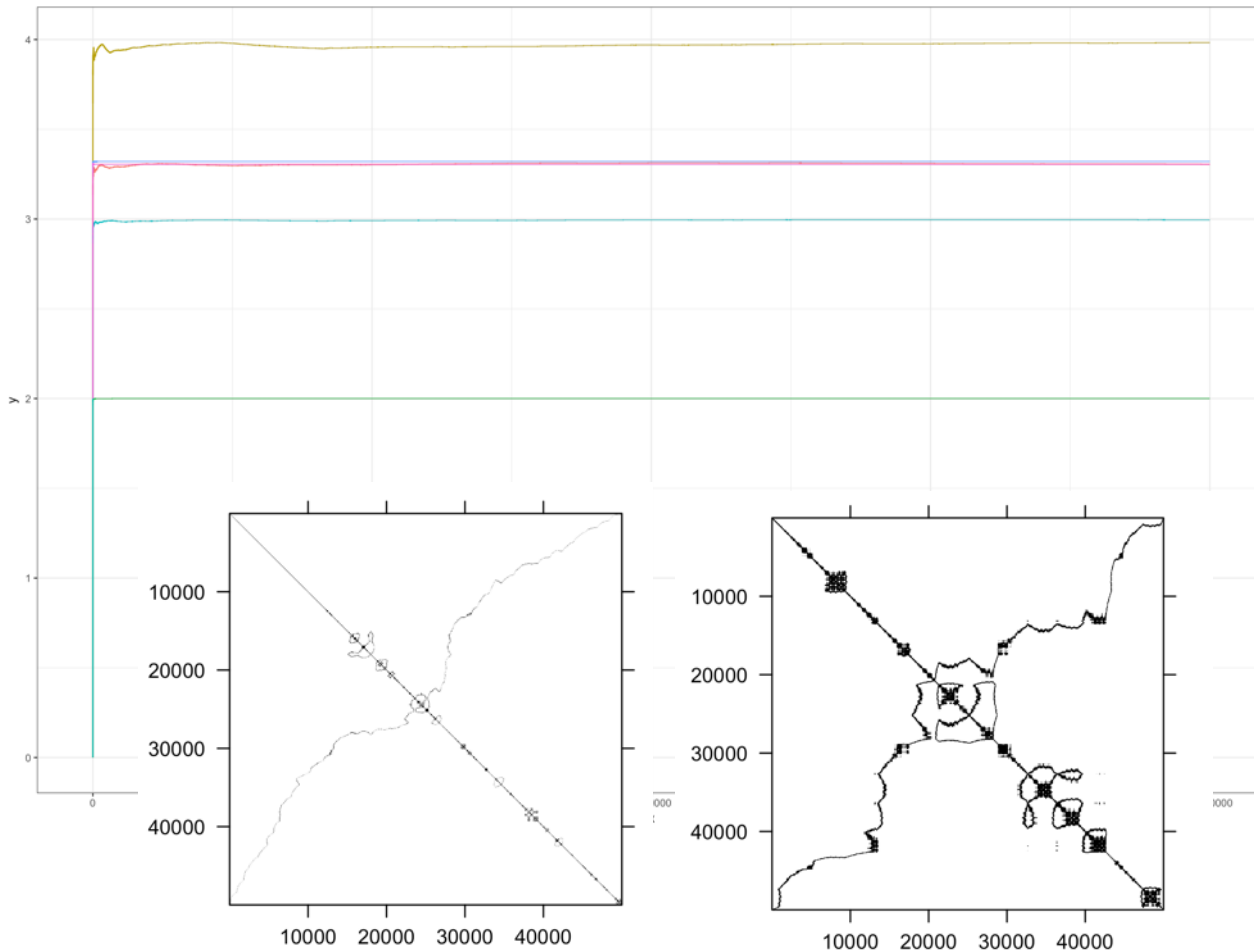


Primes in random order

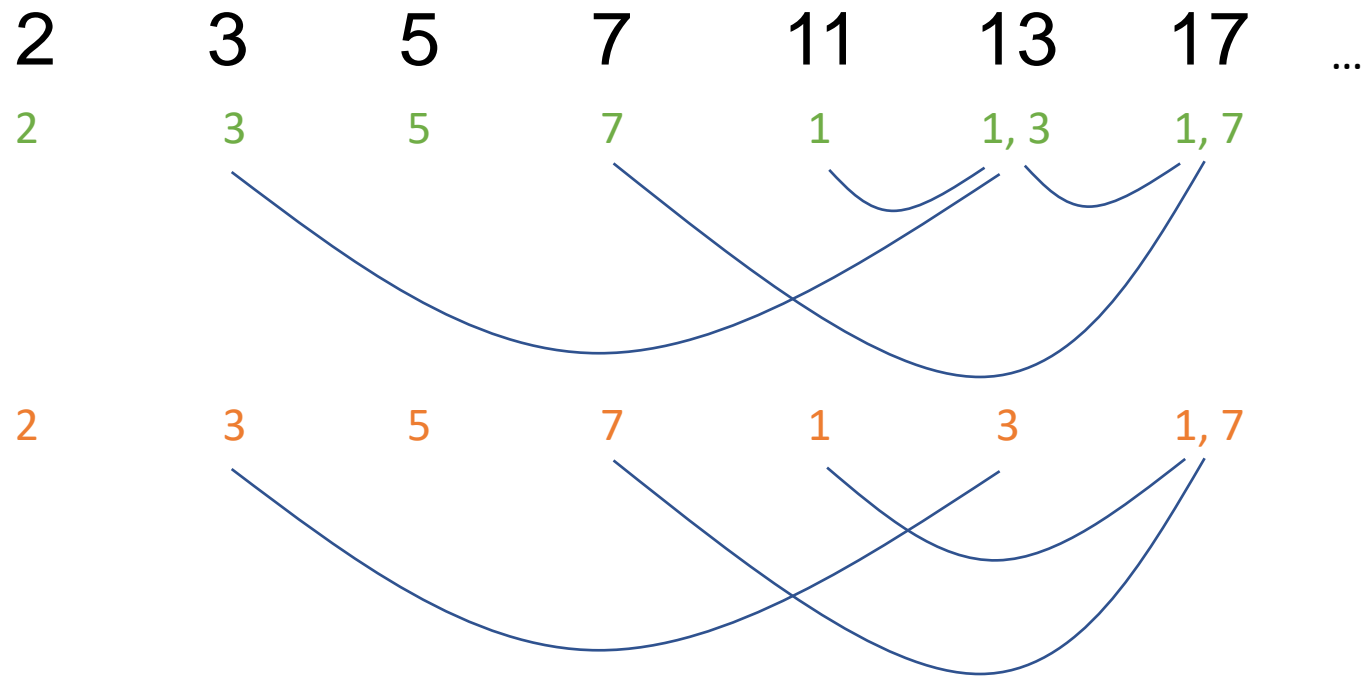


Random numbers

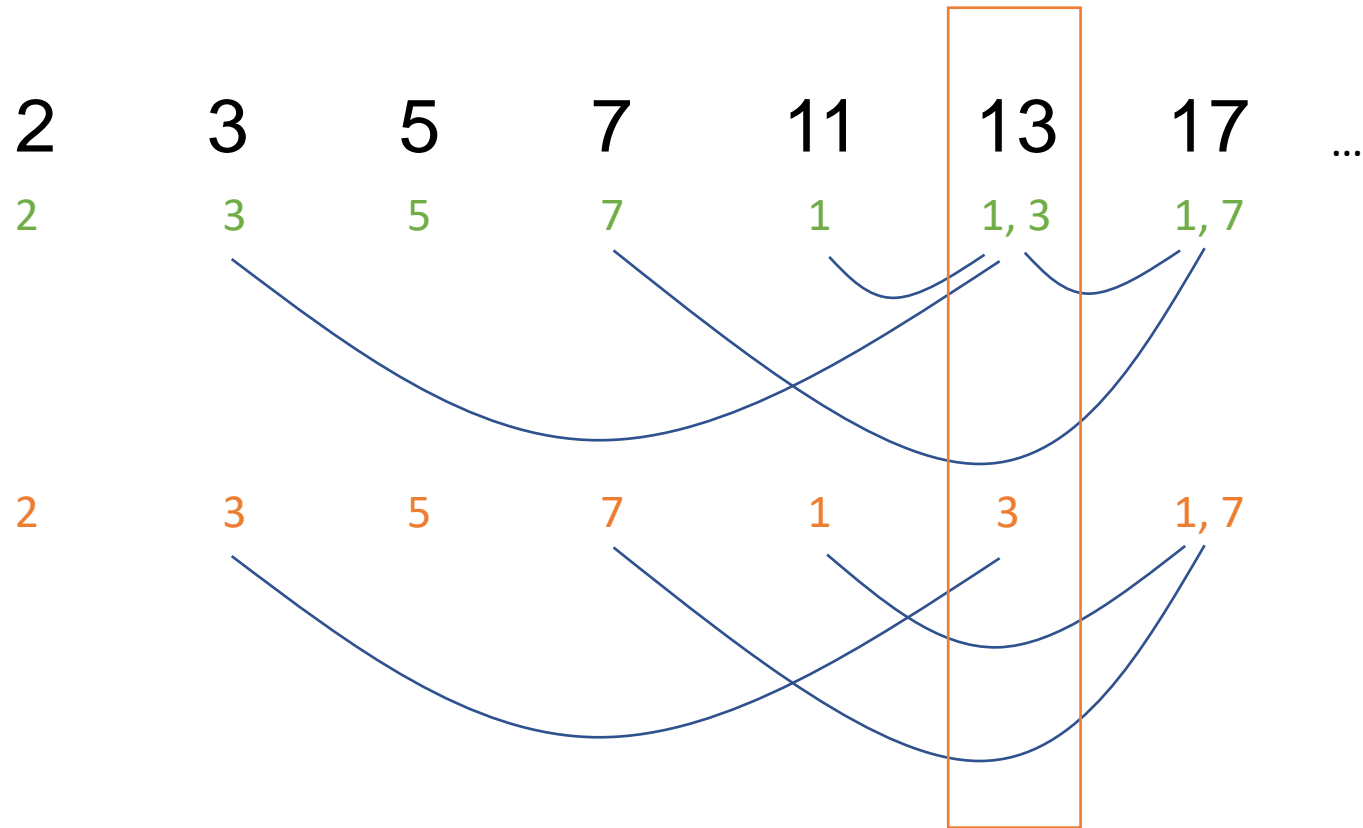
Changing the base makes the pattern get stronger



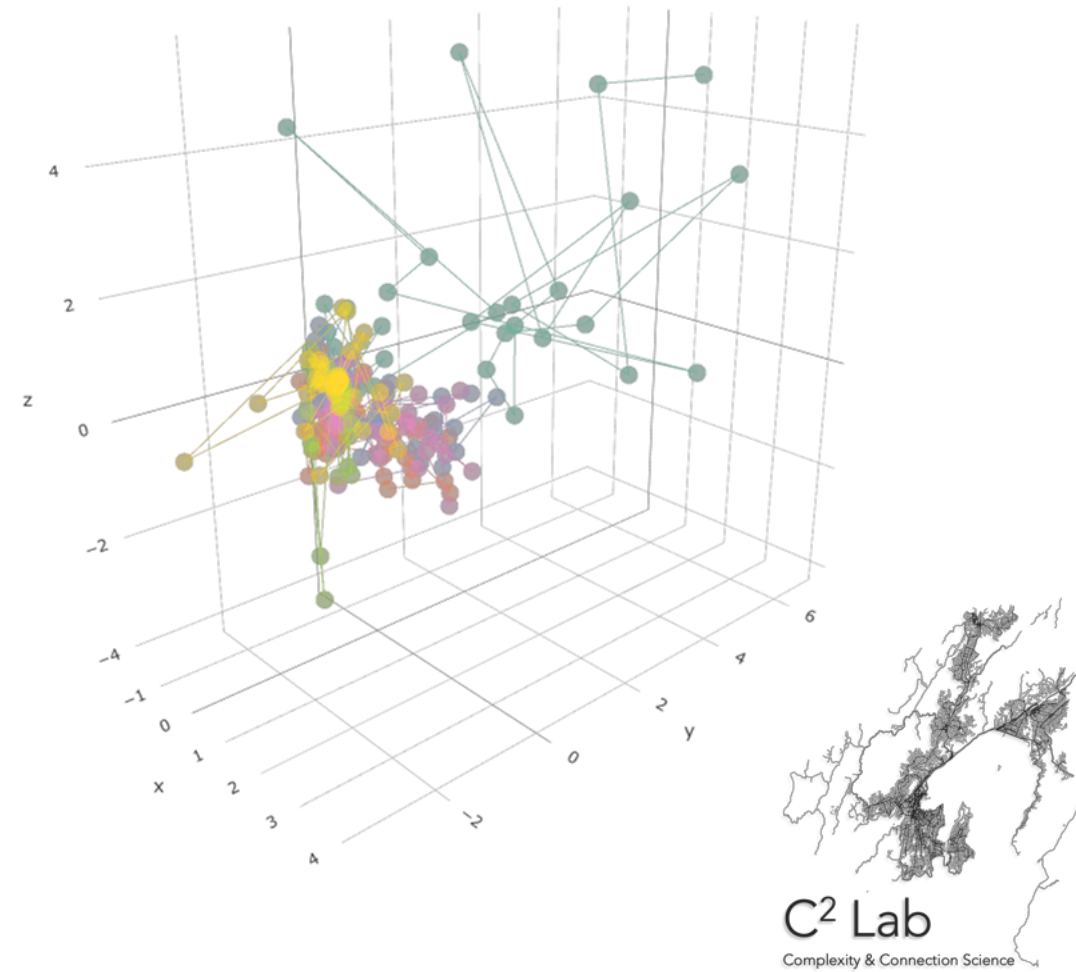
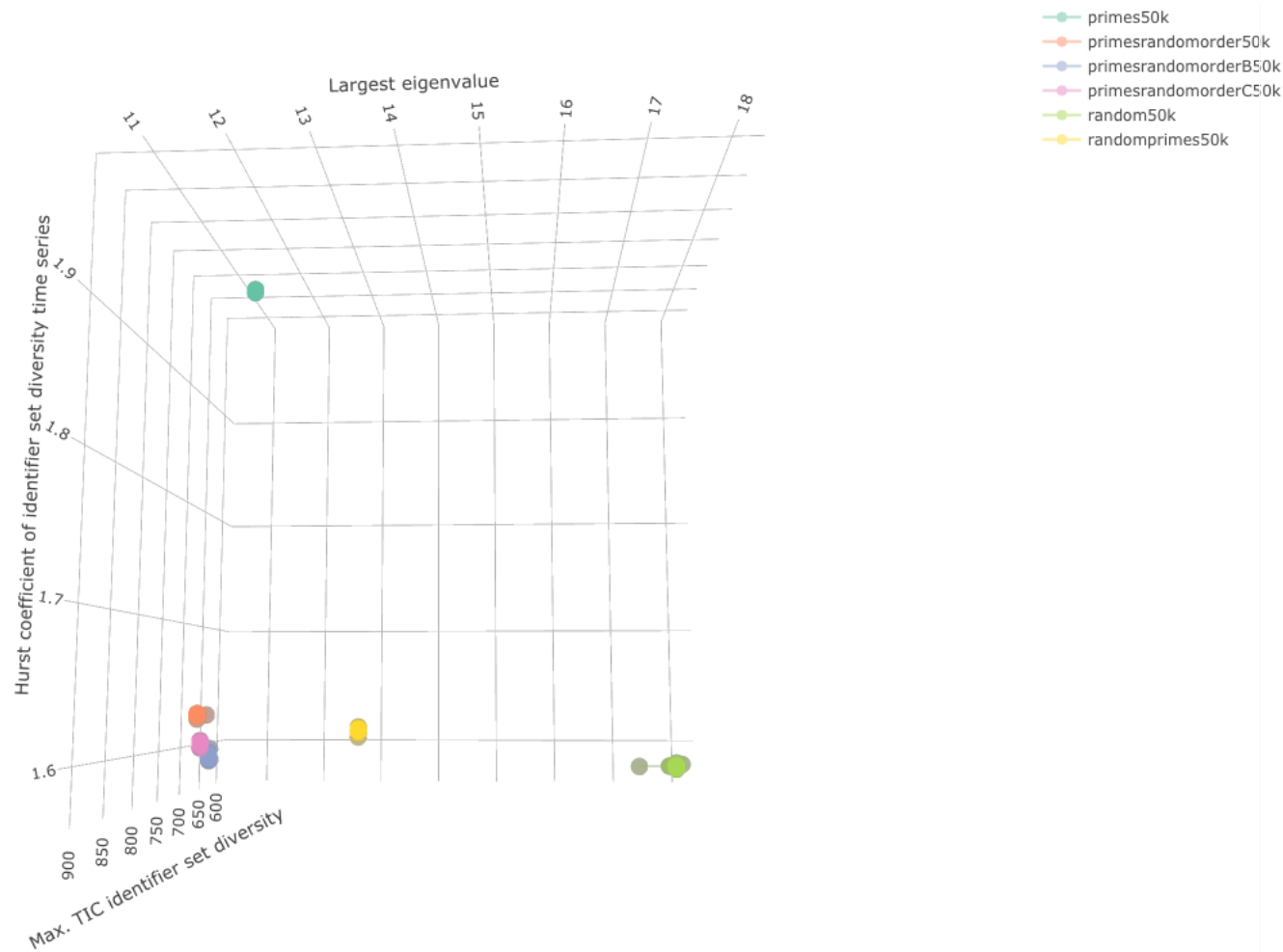
What if we would imagine something that happened did not?



What if we would imagine something that happened did not?



Characterising significant statistical insignificance



Messrs. CHAPPELL and CO. beg to announce that they have made arrangements with
MR.
CHARLES DICKENS

FOR
ONE FAREWELL READING

(THE ONLY READING THAT MR. DICKENS WILL EVER GIVE IN NOTTINGHAM)

ON
THURSDAY EVENING, FEB. 4th, 1869

WHEN HE WILL READ HIS
DOCTOR MARIGOLD

AND THEN

From Dickens to Data Science

All fourteen of Dickens's completed novels were published serially in weekly or monthly instalments.



Cadby? Yes.
 Flora, Mr F's Aunt? Yes.
 Punks? Yes.
 Miss Wade? No.
 Lagnier? No.
 Cavalitto? Carry through.
 The Meagles? No.
 Pet and Gowran? No. (Next N?)
 Daniel Doyce? Slightly.
 Plornish Family? No.

Mr and Mrs Chadband? No.
 Allan Woodcourt? Yes. Return.
 Skimpole - family? Yes.
 Boythorn. - About him, but not himself.
 Mr Jarndyce. Yes - And his love for Esther to be now brought out.
 George - and Bagnets? No. Next N?

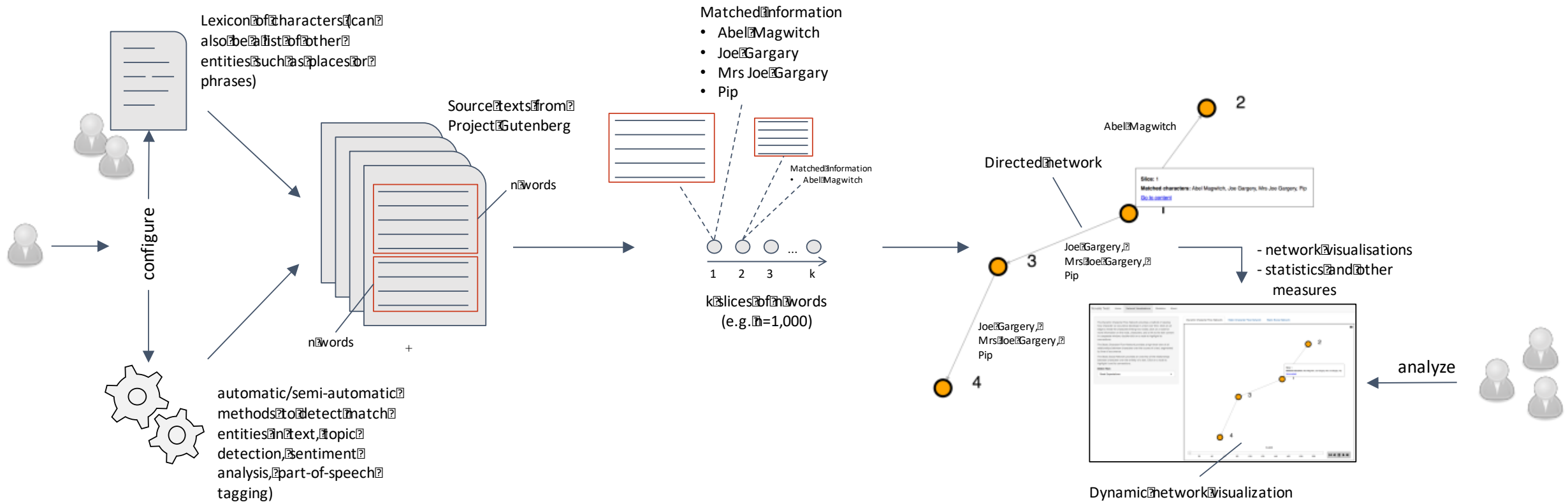
Mr and Mrs Chadband? No.
 Allan Woodcourt? Yes. Return.
 Skimpole? - family? Yes.
 Boythorn. - About him, but not himself.
 Mr Jarndyce. Yes - And his love for Esther to be now brought out.
 George - and Bagnets? No. Next N?

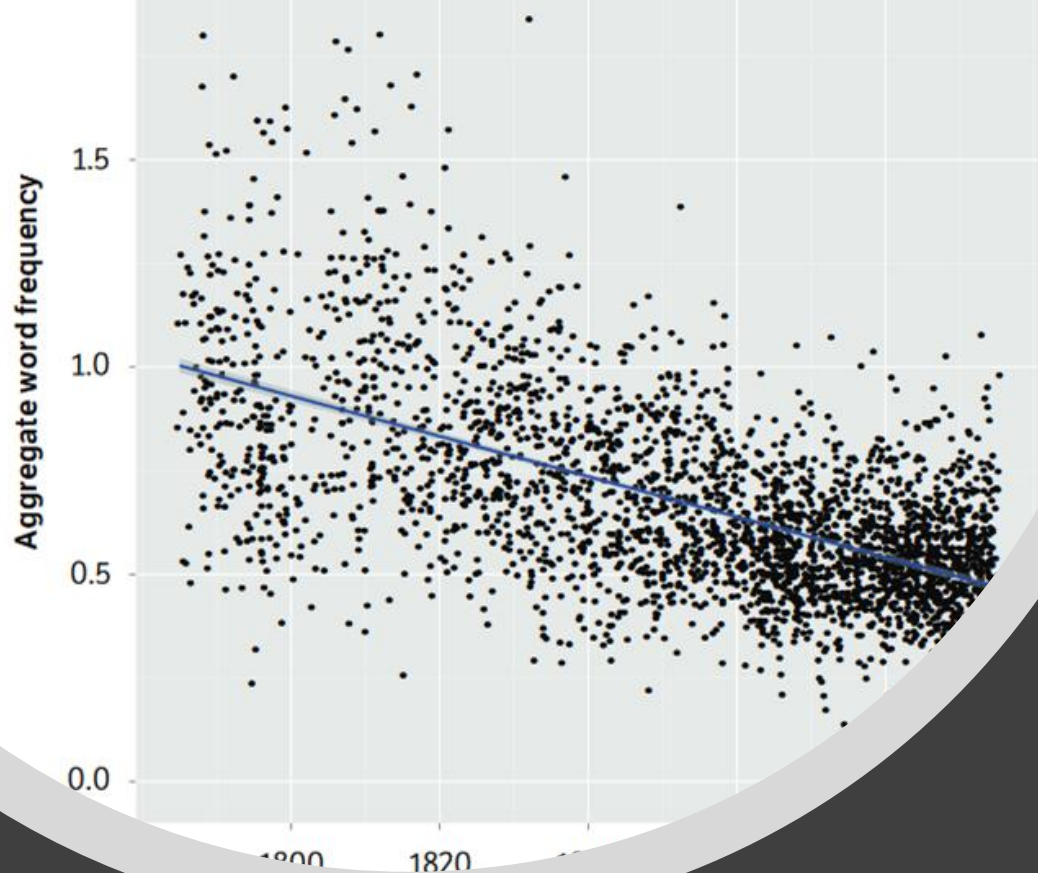
Dickens' Working Notes for His Novels. Edited by Harry Stone, U of Chicago Press, 1987.

Managing Characters

"I have endeavoured in the progress of this Tale, to resist the temptation of the current Monthly Number, and to keep a steadier eye upon the general purpose and design." Preface to Martin Chuzzlewit (1844)

Transcendental Information Cascades applied to English literature

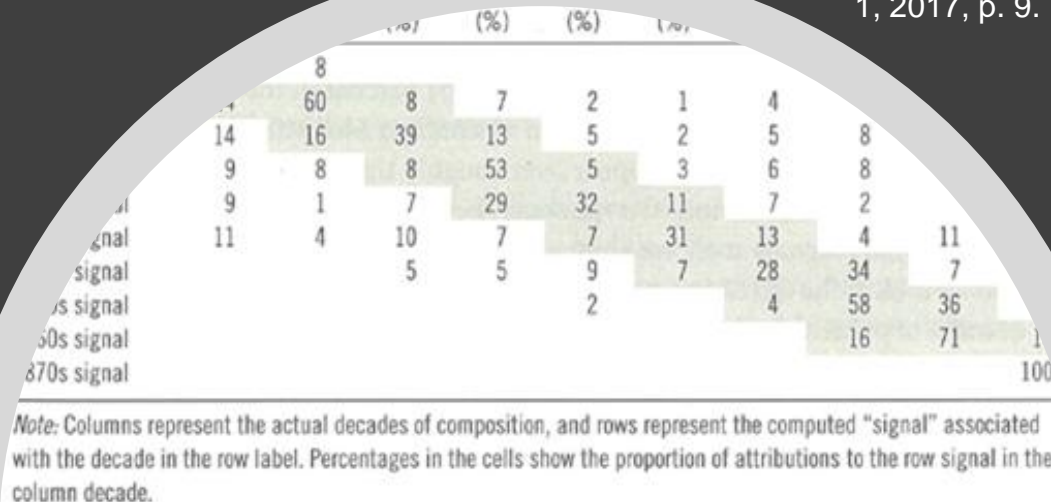




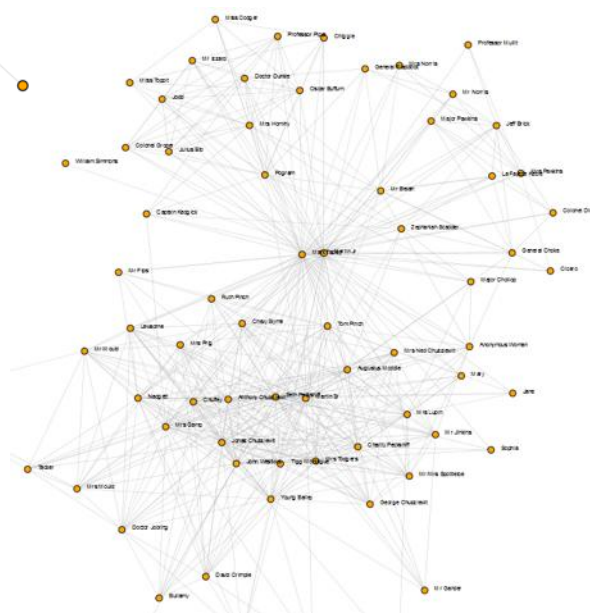
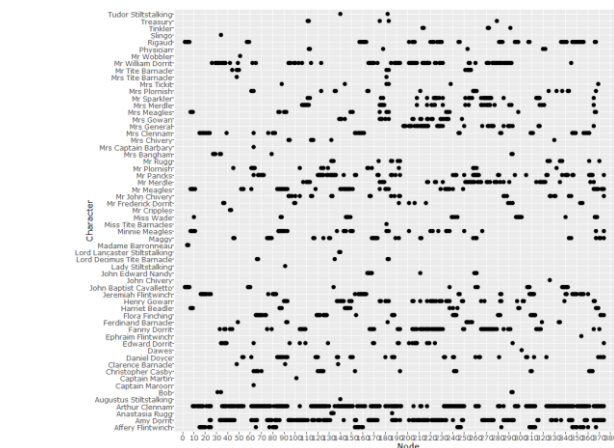
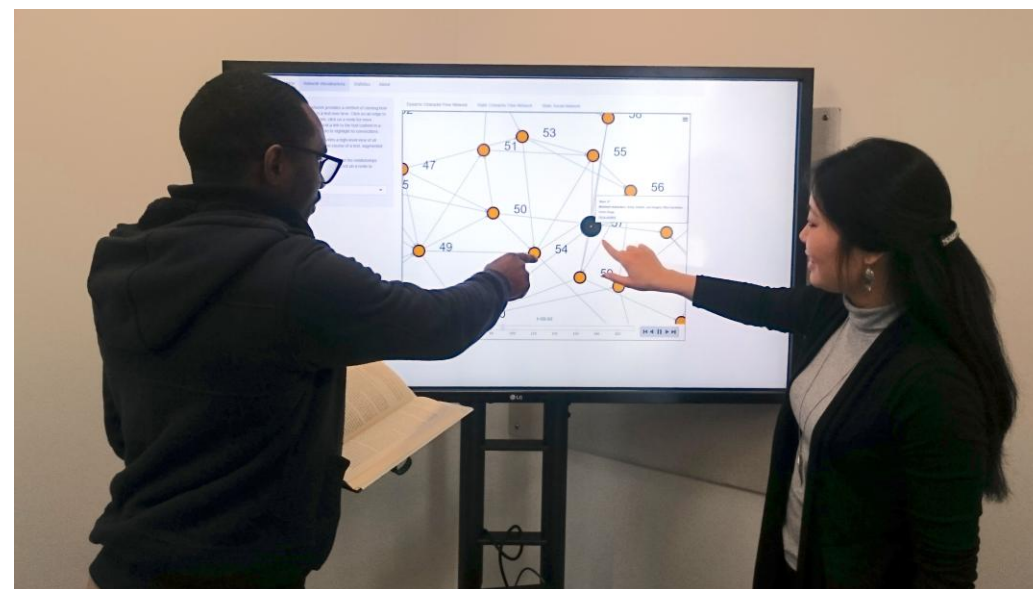
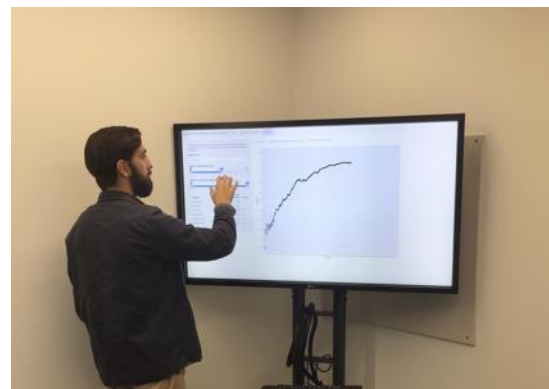
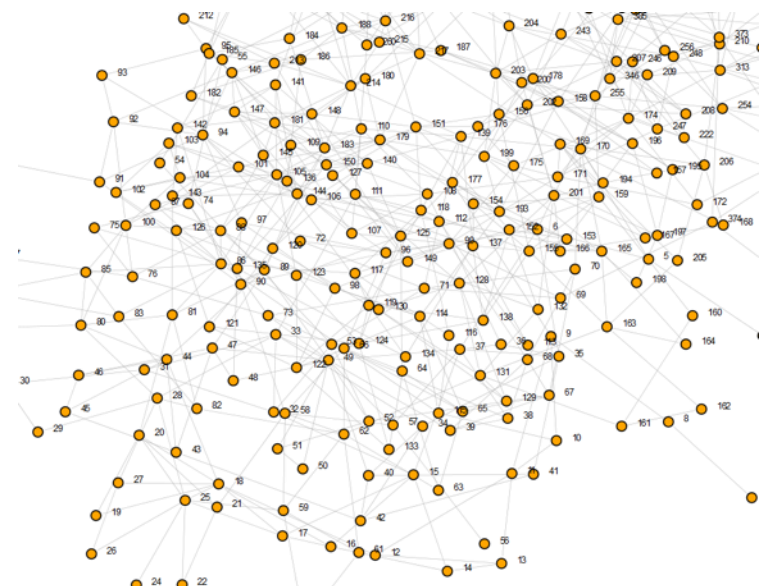
“Not-So-Distant Reading”

“The point, then, is that data-driven approaches are **not just doing the same thing better or at a larger scale**. They are **doing a different thing altogether**: interacting with the objects of the world. Traditional literary criticism, on the contrary, interacts with the past, with tradition. While one falsifies theories, the other develops from them. The figure of one is the data visualization. And the figure of the other is narrative.”^[12]

Rosenthal, Jesse. “Narrative Against Data.” *Genre*, vol. 50, no. 1, 2017, p. 9.



“Not-So-Distant Reading”

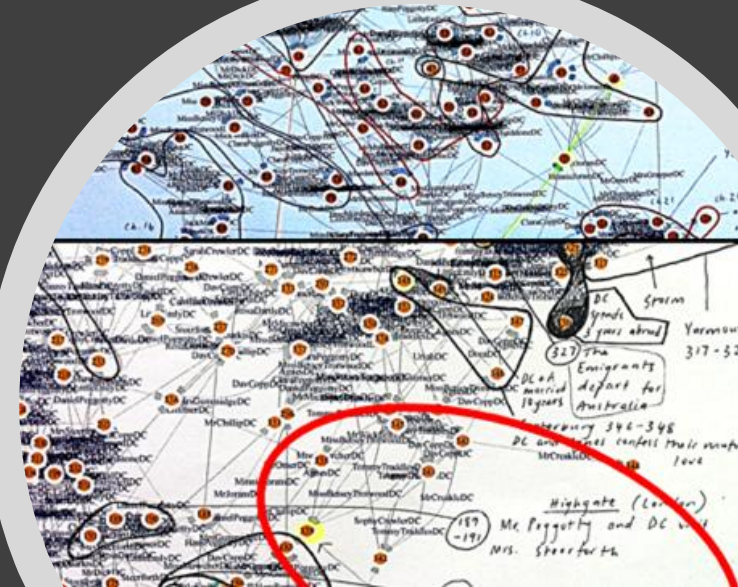


“Not-So-Distant Reading”

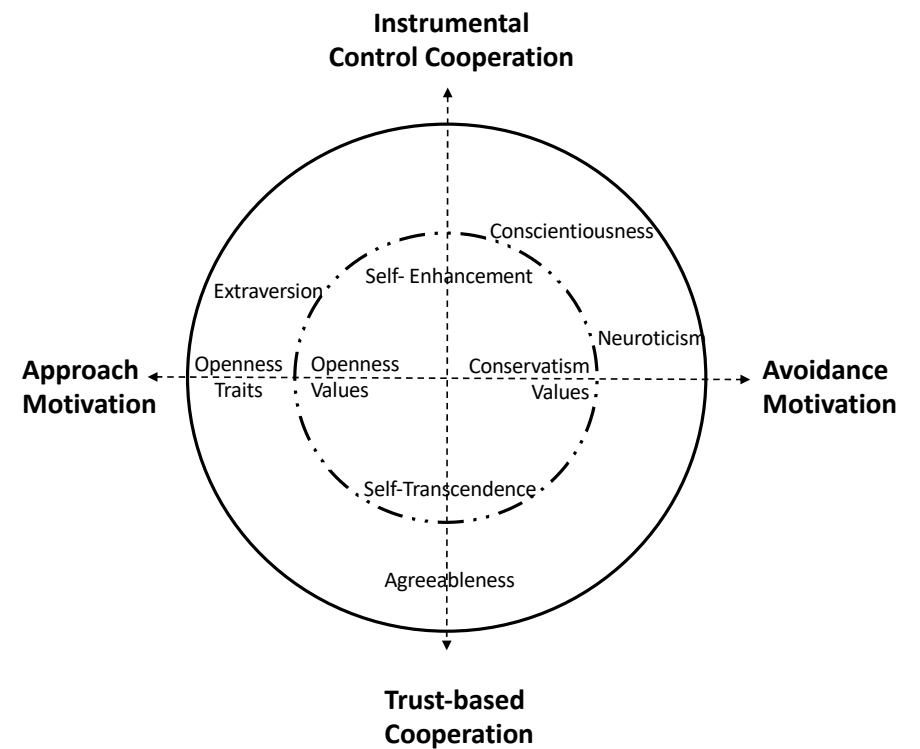
Although our method draws on foundational methodologies within digital humanities (hypothesis-testing, quantitative analysis), it explores the seeming contradiction between ‘distant’ and ‘close’ reading.

Luczak-Roesch, M., Grener, A. and Fenton, E., 2018. Twenty Thousand Leagues Above the Book: An Interactive Visual Analytics Approach to Literature. In Proceedings of the International Conference on Supporting Group Work (GROUP), ACM. DOI: 10.1145/3148330.3154507.

Luczak-Roesch, M., Grener, A. & Fenton, E. (2018). Not-so-distant reading: A dynamic network approach to literature. *it - Information Technology*, 60(1), pp. 29-40. Retrieved 1 Mar. 2018, from doi:10.1515/itit-2017-0023



Stories are about complex systems



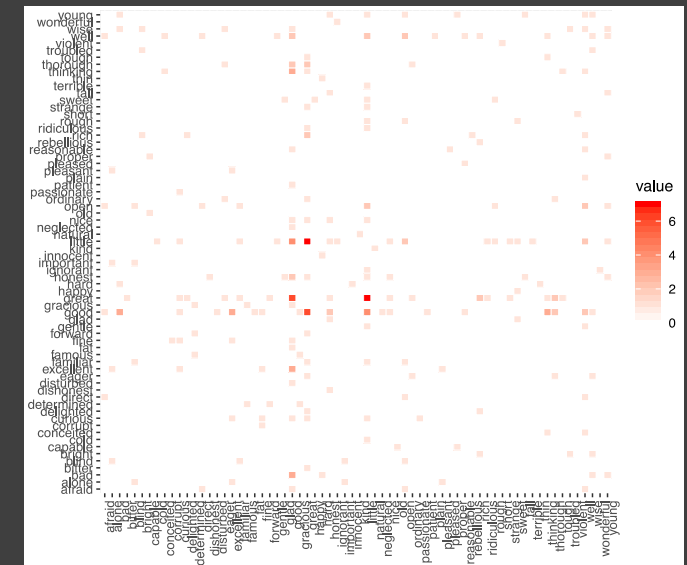
The psycholexical approach

“a great challenge i had to face was traveling to Africa fro a month. i was part of a group from school. in XXX i built mud homes for a family in a village called XXX. this was a challenge physically and mentally. the heat and nature of the work was a **great** challenge physically. i was not used to the intensity of the heat and when working long hours i became very **tired**. mentally it was **hard** to see the living conditions human beings were exposed too. I was also told of their story, the family had been struck with aids and had lost a few members, leaving behind 3 young children to be cared for by the grandmother. i addressed this challenge by remembering i was there to do my part and help in whatever way i could, i reminded myself that i am very privileged and i was very happy and lucky to be able to help these people by giving them a safe new home. i also took in everything that was around me and noticed how happy the family was, this gave me a lot of modivation to push through and do a good job”

The Big Five Inventory-2 (BFI-2)

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

1 Disagree strongly	2 Disagree a little	3 Neutral; no opinion	4 Agree a little	5 Agree strongly
<i>I am someone who...</i>				
1. ___ Is outgoing, sociable.				31. ___ Is sometimes shy, introverted.
2. ___ Is compassionate, has a soft heart.				32. ___ Is helpful and unselfish with others.
3. ___ Tends to be disorganized.				33. ___ Keeps things neat and tidy.
4. ___ Is relaxed, handles stress well.				34. ___ Worries a lot.
5. ___ Has few artistic interests.				35. ___ Values art and beauty.
6. ___ Has an assertive personality.				36. ___ Finds it hard to influence people.
7. ___ Is respectful, treats others with respect.				37. ___ Is sometimes rude to others.
8. ___ Tends to be lazy.				38. ___ Is efficient, gets things done.
9. ___ Stays optimistic after experiencing a setback.				39. ___ Often feels sad.
10. ___ Is curious about many different things.				40. ___ Is complex, a deep thinker.
11. ___ Rarely feels excited or eager.				41. ___ Is full of energy.
12. ___ Tends to find fault with others.				42. ___ Is suspicious of others' intentions.
13. ___ Is dependable, steady.				43. ___ Is reliable, can always be counted on.
14. ___ Is moody, has up and down mood swings.				44. ___ Keeps their emotions under control.
15. ___ Is inventive, finds clever ways to do things.				45. ___ Has difficulty imagining things.
16. ___ Tends to be quiet.				46. ___ Is talkative.
17. ___ Feels little sympathy for others.				47. ___ Can be cold and uncaring.
18. ___ Is systematic, likes to keep things in order.				48. ___ Leaves a mess, doesn't clean up.
19. ___ Can be tense.				49. ___ Rarely feels anxious or afraid.
20. ___ Is fascinated by art, music, or literature.				50. ___ Thinks poetry and plays are boring.
21. ___ Is dominant, acts as a leader.				51. ___ Prefers to have others take charge.
22. ___ Starts arguments with others.				52. ___ Is polite, courteous to others.
23. ___ Has difficulty getting started on tasks.				53. ___ Is persistent, works until the task is finished.
24. ___ Feels secure, comfortable with self.				54. ___ Tends to feel depressed, blue.
25. ___ Avoids intellectual, philosophical discussions.				55. ___ Has little interest in abstract ideas.
26. ___ Is less active than other people.				56. ___ Shows a lot of enthusiasm.
27. ___ Has a forgiving nature.				57. ___ Assumes the best about people.
28. ___ Can be somewhat careless.				58. ___ Sometimes behaves irresponsibly.
29. ___ Is emotionally stable, not easily upset.				59. ___ Is temperamental, gets emotional easily.
30. ___ Has little creativity.				60. ___ Is original, comes up with new ideas.





The problem with our understanding of personality is that we see it through contemporary eyes

In other words: authors/speakers are long dead, fictive characters never lived, and humans are bad at assessing personality from text

1982

objective rules. Although we have continuously sought to find an objective procedure that would approximate the judgments being made by investigators, all of these efforts have failed. As one example, in reducing a set of 1666 trait adjectives to a subset roughly half that size, 25 indices of word length, grammatical form, difficulty level, interpretive ambiguity, and response extremeness were used in a stepwise multiple regression analysis to predict one investigator's suitability judgments. The variable "root term versus all other forms" correlated .32 with these judgments, whereas the addition of "difficulty level" from the Norman data raised the multiple correlation to .34. An index of interpretive ambiguity raised the multiple correlation to .35, and one of response extremeness only sent it up to .36. None of the other objective indices provided any significant increments in the multiple correlation. Clearly, then, most of the variance in these judgments—which tend to be correlated above .80 among various pairs of investigators—is not predictable from a linear combination of those particular indices. Even though the quest continues, it seems unlikely that a completely objective procedure will be devised, unless it includes as predictors consensus ratings of "slanginess" and "awkwardness," and incorporates some measure of semantic redundancy between a given term and other more common (less awkward, less slangy) ones.

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. *Advances in personality assessment*, 1, 203-234.

→2018

5.2 Difficulties with Measuring IAA

As we showed in Section 5.1, the agreement across all annotation layers is comparably low. There are several reasons for that. Indeed, emotion annotation is highly subjective, but it is not the only subjective category. The cause and target of the emotion are not always clearly recognizable in the text and are also subjective categories (two annotators may find two different causes for the same emotion), hence the low agreement scores across all categories. The only exception are *experiencer* annotations, which are the most reliable among all annotations and match the substantial agreement scores of character annotation (the only type of entities that can be involved in an experiencer relation).

We illustrate the difficulties the annotators face when annotating emotions with roles with the following example: "they had never seen . . . what was really hateful in his face; . . . they could only express it by saying that the arched brows and the long emphatic chin gave it always a look of being lit from below . . ." All annotators agree on the character ("they") and the emotion ("hateful" expressing disgust). Similarly, both annotators agree that the disgust is related to properties of the face which is described, however, one annotator marks "his face" as target, the other marks the more specific but longer "the arched brows and the long emphatic chin gave it always a look of being lit from below" as cause.

If we abstract away from the text spans, both annotators agree that the emotion of disgust has something to do with "his face", however they disagree on the target annotation and the cause annotation. So, though conceptually, the annotations by two people are similar, this is not captured by our calculation of inter-annotator agreement.

Kim, E., & Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1345-1359).

It seems very hard, so let's do it.

- Selection of 25 English novels
- Earliest: The unfortunate traveller by Thomas Nashe (1594)
- Latest: Edwin Drood by Charles Dickens (1870)
 - Robinson Crusoe, Oliver Twist, Evelina, Nicholas Nickleby, Pickwick, Frankenstein, The Old Curiosity Shop, Barnaby Rudge, Great Expectations, Martin Chuzzlewit, Dombey and Son, David Copperfield, Bleak House, Hard Times, Little Dorrit, A Tale of Two Cities, Great Expectations, Our Mutual Friend, Mystery of Edwin Drood, Emma, Mansfield Park, Northanger Abbey, Persuasion, Pride and Prejudice, Sense and Sensibility
- Dickens + Austen = 21

Fischer, R., Karl, J. A., Luczak-Roesch, M., Fetvadjeiev, V. H., & Grener, A. Tracing Personality Structure in Narratives: A Computational Bottom-Up Approach to Unpack Writers, Characters, and Personality in Historical Context. *European Journal of Personality*.

“The quick brown Fox jumps over the not so lazy Dog. He was happy after that.”

→ Fox:quick, Fox:happy, Dog:not_lazy

Vocabulary differences reflect the different styles of authors as seen by humanities scholars

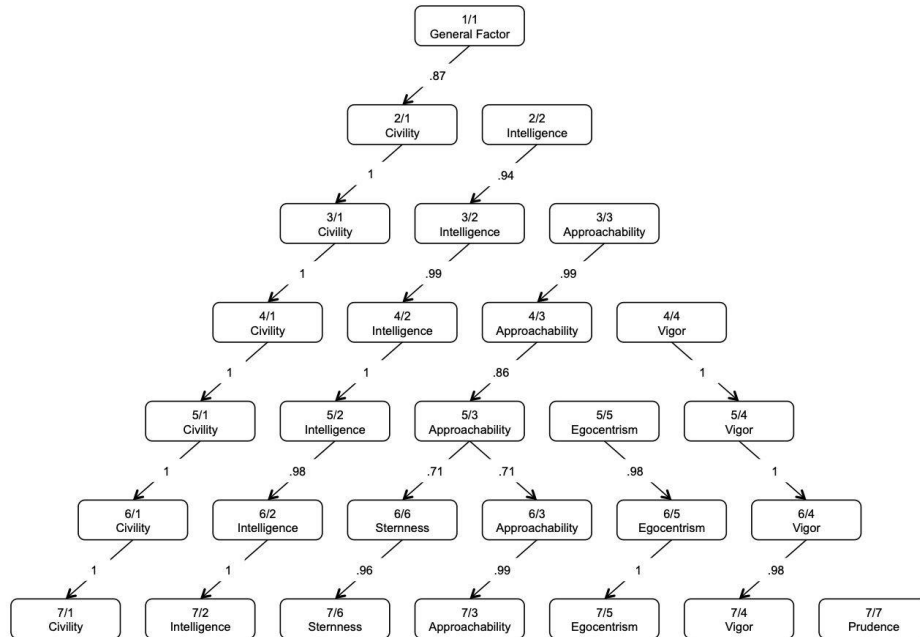


Figure 5. Factor cascades for the Austen novels, using the 1,710 dictionary. Correlations above .50 are shown.

- Austen → smaller trait vocabulary, introspection of individuals
- some conceptual overlap with five factor scales
- but primary focus on social- normative aspects
- even Practical Intelligence factor has content dealing with social orientation, such as philanthropic and charitable

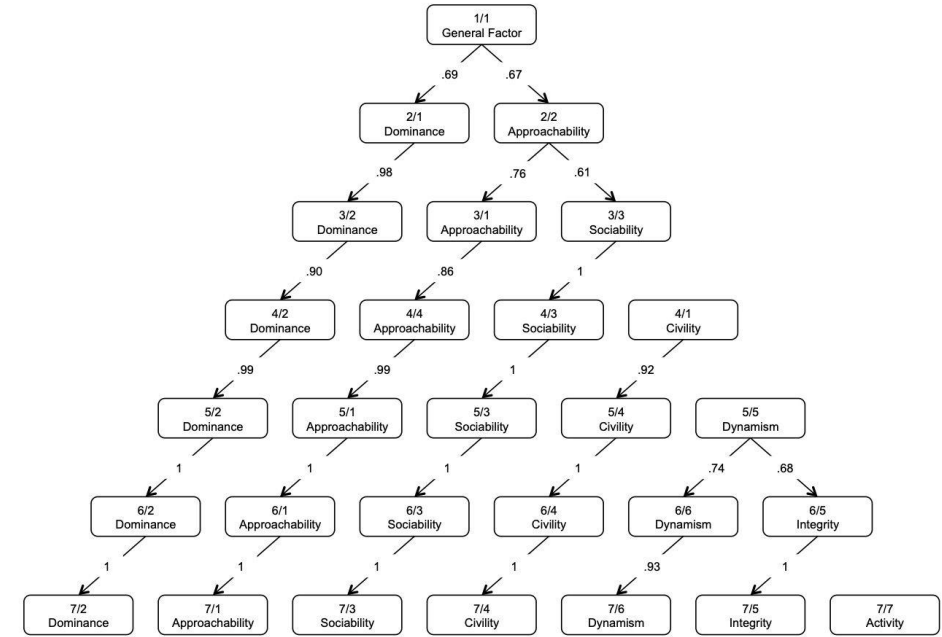


Figure 6. Factor cascades for the Dickens novels, using the 1,710 dictionary. Correlations above .50 are shown.

- Dickens → broad trait vocabulary, complex description of the social world of large groups of characters
- some relations to combinations of markers of Big Five
- but own idiosyncratic content revolving around themes of social relations such as arrogance, dominance, sociability, and civility

But...

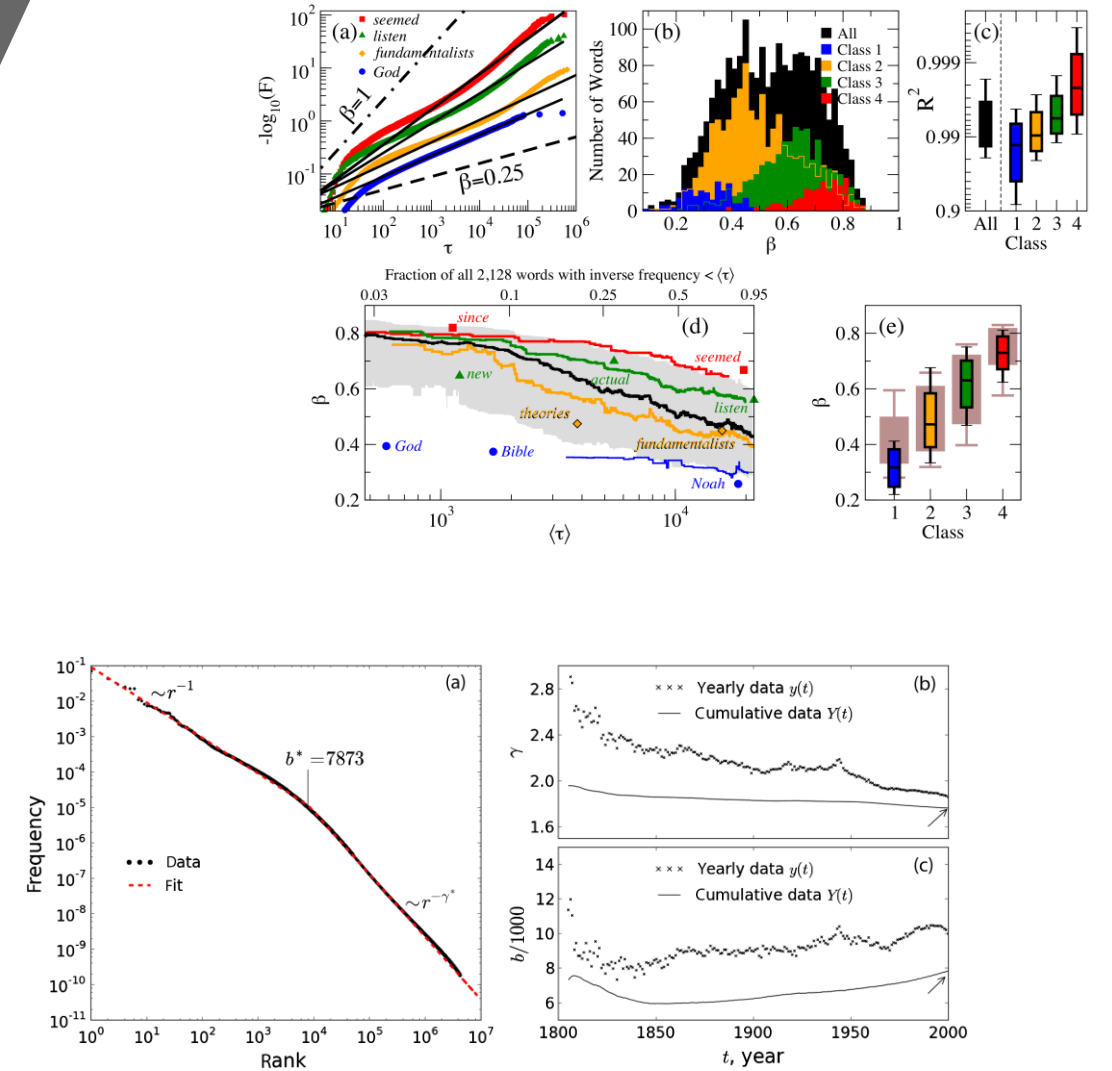
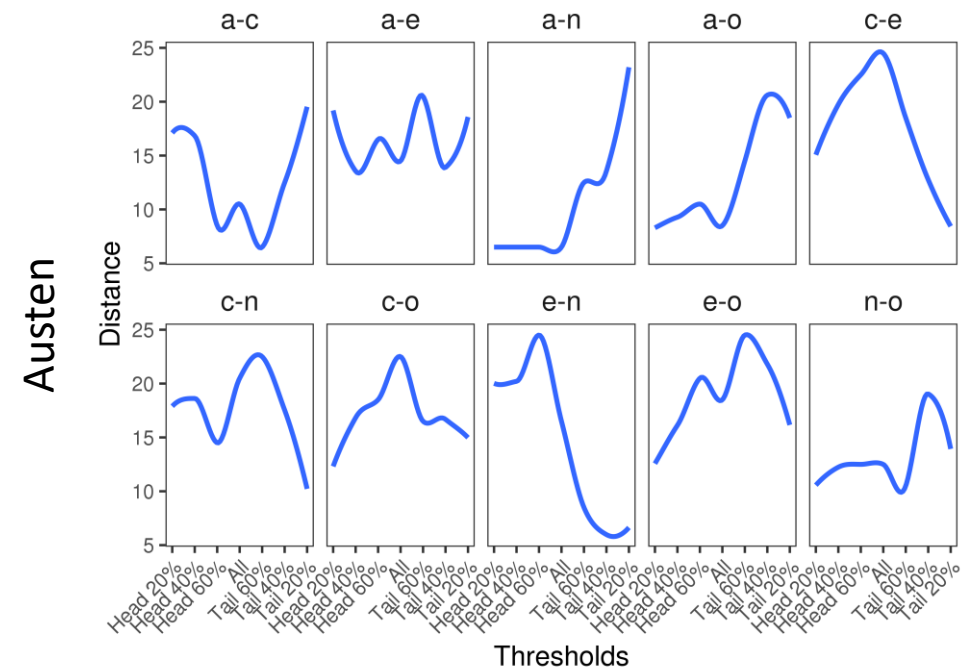
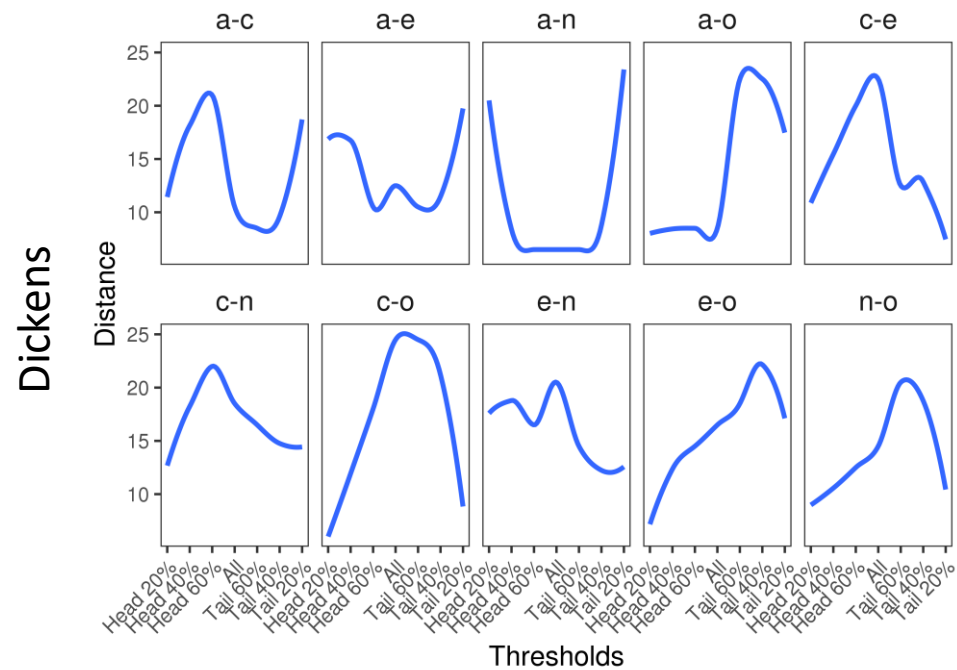


FIG. 1. The rank-frequency distribution shows double-scaling behavior (Zipf's plot). (a) Rank-frequency distribution for the English database $Y(2000)$ (solid line) and a ML fit of Eq. (1) (dashed line). (b,c) parameters γ and b obtained from ML fits of Eq. (1) to yearly $y(t)$ (x symbols) and accumulated $Y(t)$ (solid line) database. Arrows indicate the values of the parameters γ^* and b^* obtained for the fit in (a). Results are shown for the time range $t \in [1805, 200]$ in which data are most reliable; accumulation starts in $t_0 = 1520$.

Unstable structures



Why?

RESEARCH

REPORT

COGNITIVE SCIENCE

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,^{1*} Joanna J. Bryson,^{1,2*} Arvind Narayanan^{1*}

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology.

Dominant words rise to the top by positive frequency-dependent selection

Mark Pagel^{a,b,1}, Mark Beaumont^c, Andrew Meade^a, Annemarie Verkerk^{a,d}, and Andreea Calude^e

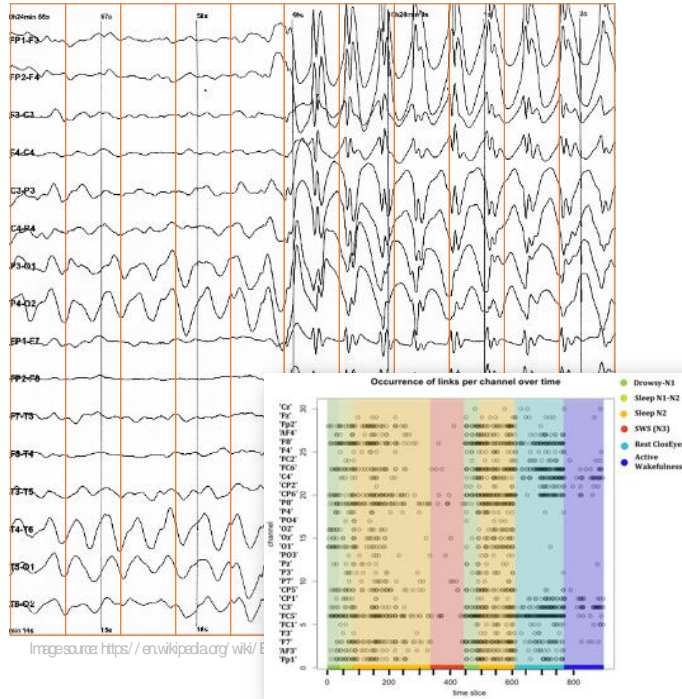
^aSchool of Biological Sciences, University of Reading, Whiteknights, RG6 6UR Reading, United Kingdom; ^bSanta Fe Institute, Santa Fe, NM 87501; ^cSchool of Biological Sciences, University of Bristol, BS8 1TW Bristol, United Kingdom; ^dDepartment of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, 07745 Jena, Germany; and ^eDepartment of General and Applied Linguistics, University of Waikato, 3240 Hamilton, New Zealand

Edited by Barbara H. Partee, University of Massachusetts at Amherst, Amherst, MA, and approved February 25, 2019 (received for review October 3, 2018)

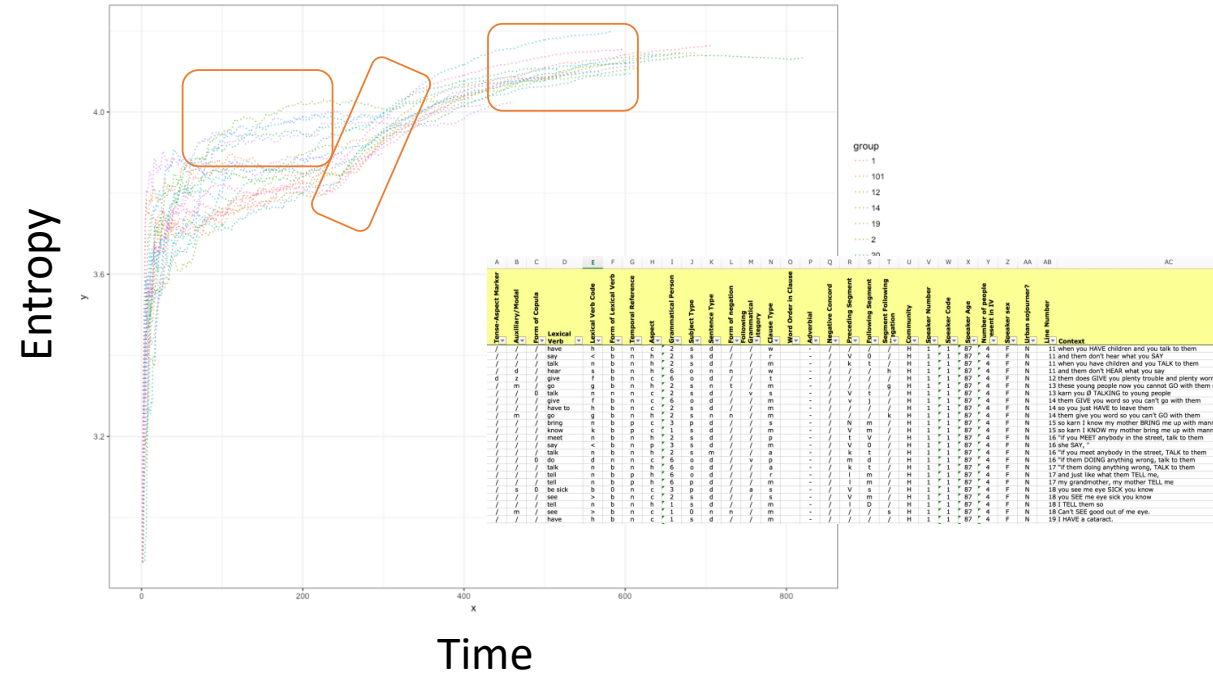
A puzzle of language is how speakers come to use the same words for particular meanings, given that there are often many competing alternatives (e.g., “sofa,” “couch,” “settee”), and there is seldom a necessary connection between a word and its meaning. The well-known process of random drift—roughly corresponding in this context to “say what you hear”—can cause the frequencies of alternative words to fluctuate over time, and it is even possible for one of the words to replace all others, without any form of selection being involved. However, is drift alone an adequate explanation of a shared vocabulary? Darwin thought not. Here, we apply models of neutral drift, directional selection, and positive frequency-dependent selection to explain over 417,000 word-use choices for 418 meanings in two natural populations of speakers. We find that neutral drift does not in general explain word use. Instead, some form of selection governs word choice in over 91% of the meanings we studied. In cases where one word dominates all others for a particular meaning—such as is typical of the words in the core lexicon of a language—word choice is guided by positive frequency-dependent selection—a bias that makes speakers disproportionately likely to use the words that most others use. This bias grants an increasing advantage to the common form as it becomes more popular and provides a mechanism to explain how a shared vocabulary can spontaneously self-organize and then be maintained for centuries or even millennia, despite new words continually entering the lexicon.

stochastic effects—no selection need be involved. Applied to language (11, 12), random drift can be used to study changes in the frequencies with which speakers use various words for a given meaning, such as “sofa,” versus “couch” or “settee.” Drift’s importance in population studies, then, is that its mathematical expression provides a precise null expectation against which stronger claims, such as those that Darwin and Müller made, can be assessed (11, 12).

For example, in language, a common observation is that when the number of speakers who use a word is plotted against that word’s rank-order position in a list of words sorted by frequency (e.g., Fig. 1 A–C), sharply down-sloping curves arise that can be described by the form $f(k) = ak^{-\beta}$, where $f(k)$ is the observed number of speakers who use a word, and k is its rank order position (1, 2, ..., k) (13). Studies in linguistic settings have shown that drift can produce curves with these shapes (12, 14–16), even the extreme example in Fig. 1C where, among competing alternatives, one word has risen to the top, dominating all others. On the other hand, while drift can in principle produce any monotonically declining curve, some outcomes of drift are more probable than others (17). So, the real question becomes not whether drift can produce outcomes such as those in Fig. 1 A–C, but whether mechanisms other than drift provide more likely explanations. This is the challenge that claims of selection in language must meet.

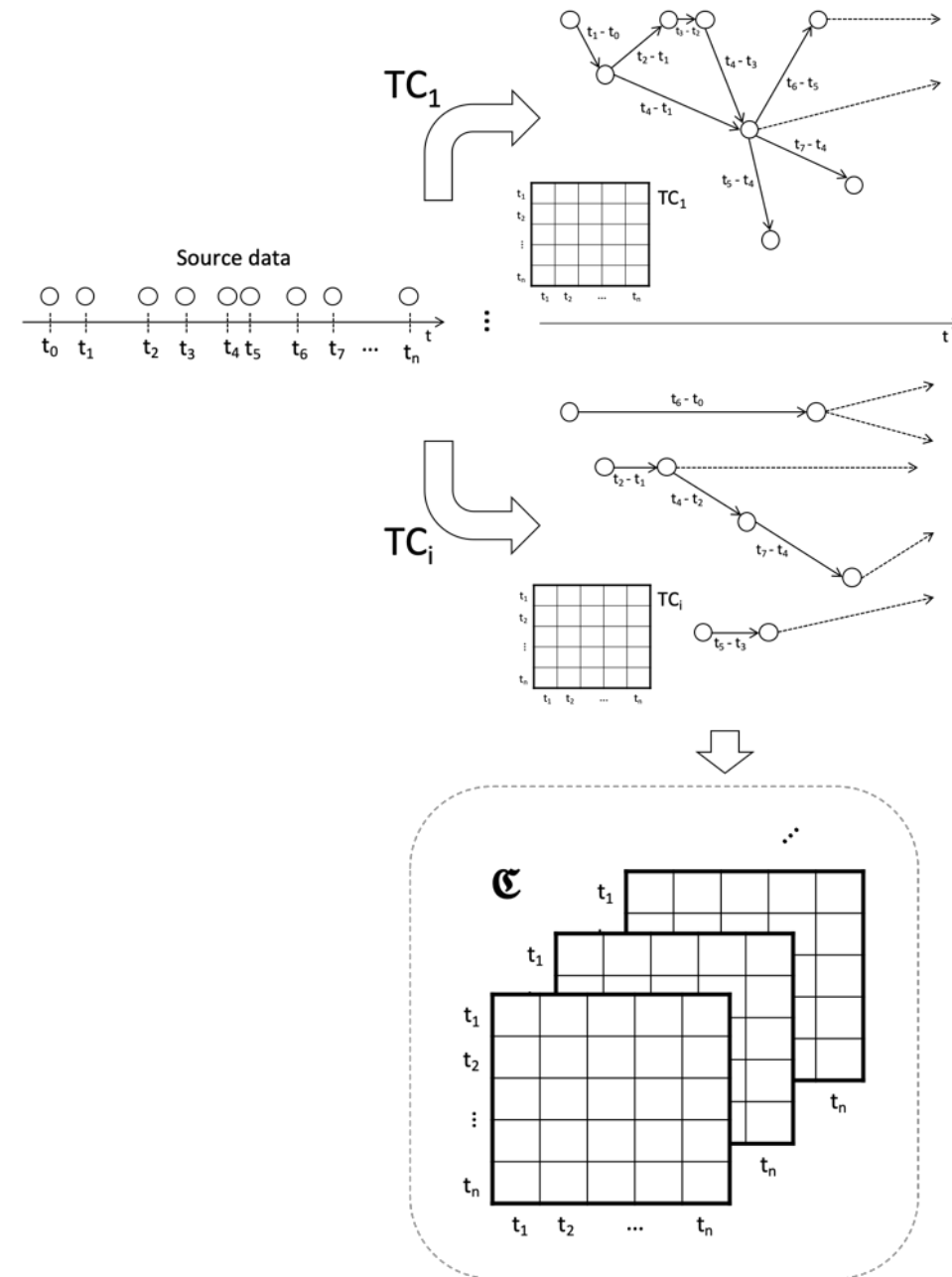


Stories are told in complex ways



Bringing different views together

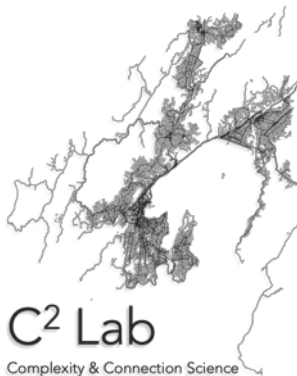
- There are many different views to the same system (many possible tokenisations)
- Usually these are studied independently or need different analytical frameworks (because of different data types)
- TIC are an analytical middle-ground (transform any kind of data into the same kind of temporal network before analysis)



“If it’s beautiful but a little bit wrong, I am ok with that.”

Steven Strogatz on the MindScape Podcast -

<https://www.preposterousuniverse.com/podcast/2019/04/08/episode-41-steven-strogatz-on-synchronization-networks-and-the-emergence-of-complex-behavior/>





**“If it’s beautiful but a
little bit wrong,
I am ok with that.”**

C² Lab
Complexity & Connection Science

Markus Luczak-Roesch | @mluczak



<http://complexity.sim.vuw.ac.nz>