Complexity Science Hub

Decoding the codebook of life

Eddie Lee, PhD

Data Science in Fundamental Physics

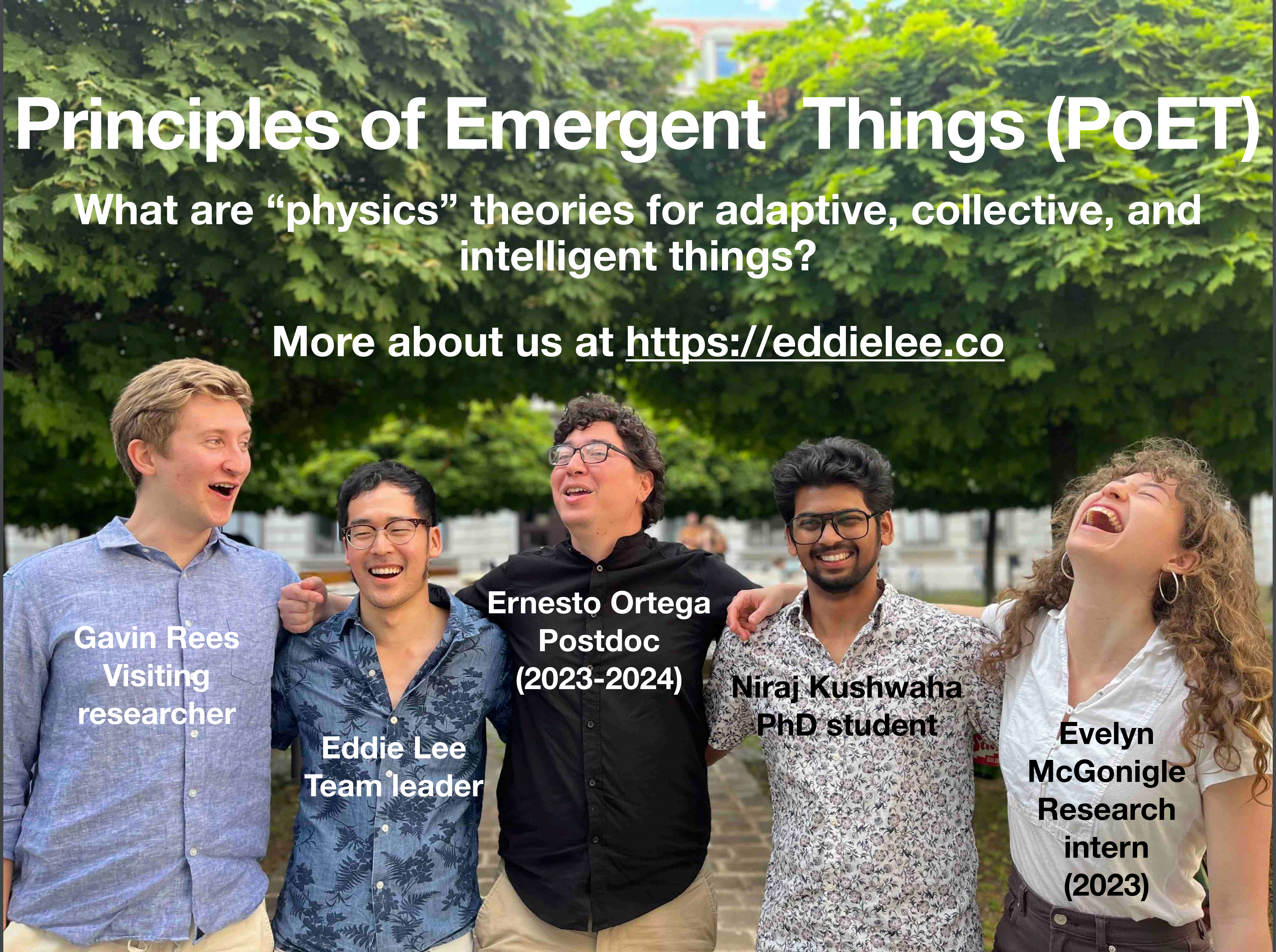
https://eddielee.co

June 6, 2024

Principles of Emergent Things (PoET)

What are “physics” theories for adaptive, collective, and intelligent things?

More about us at https://eddielee.co



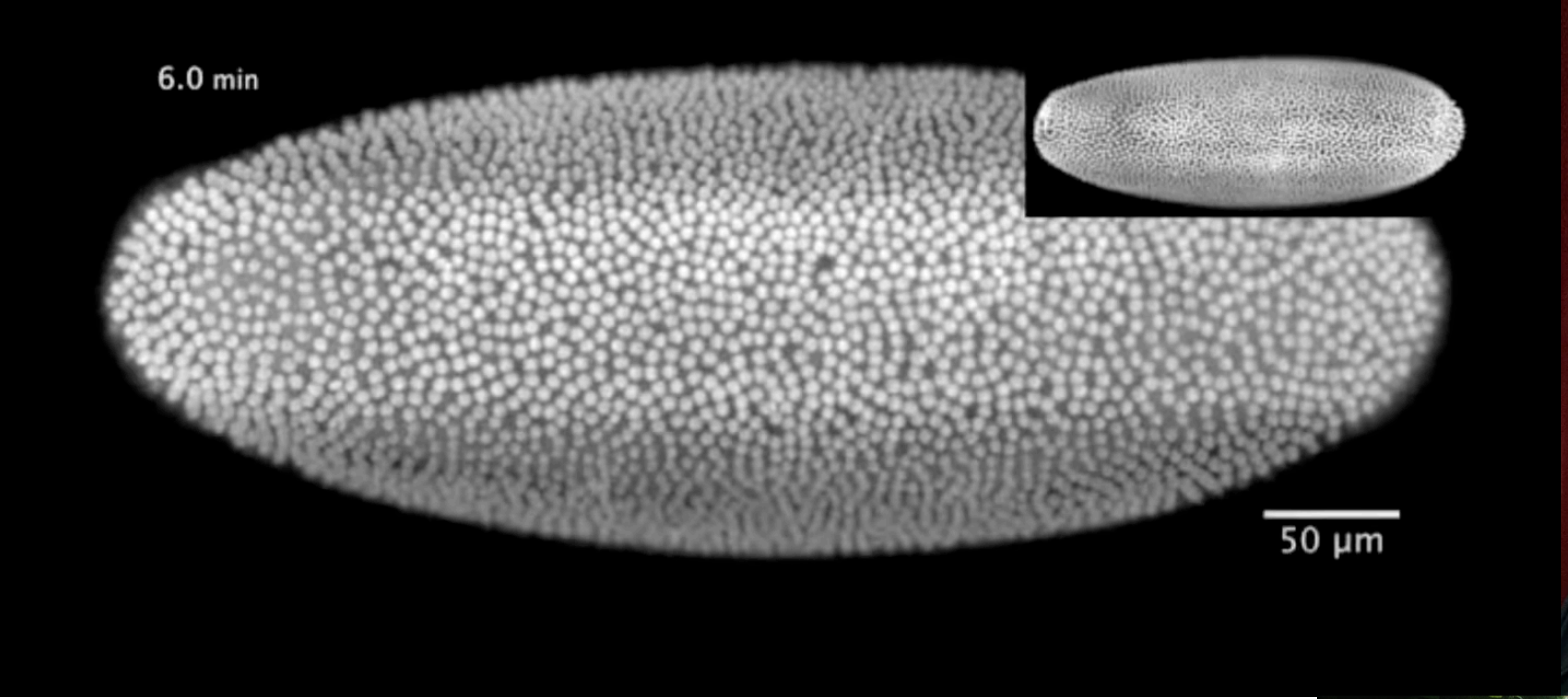
Gavin Rees
Visiting
researcher

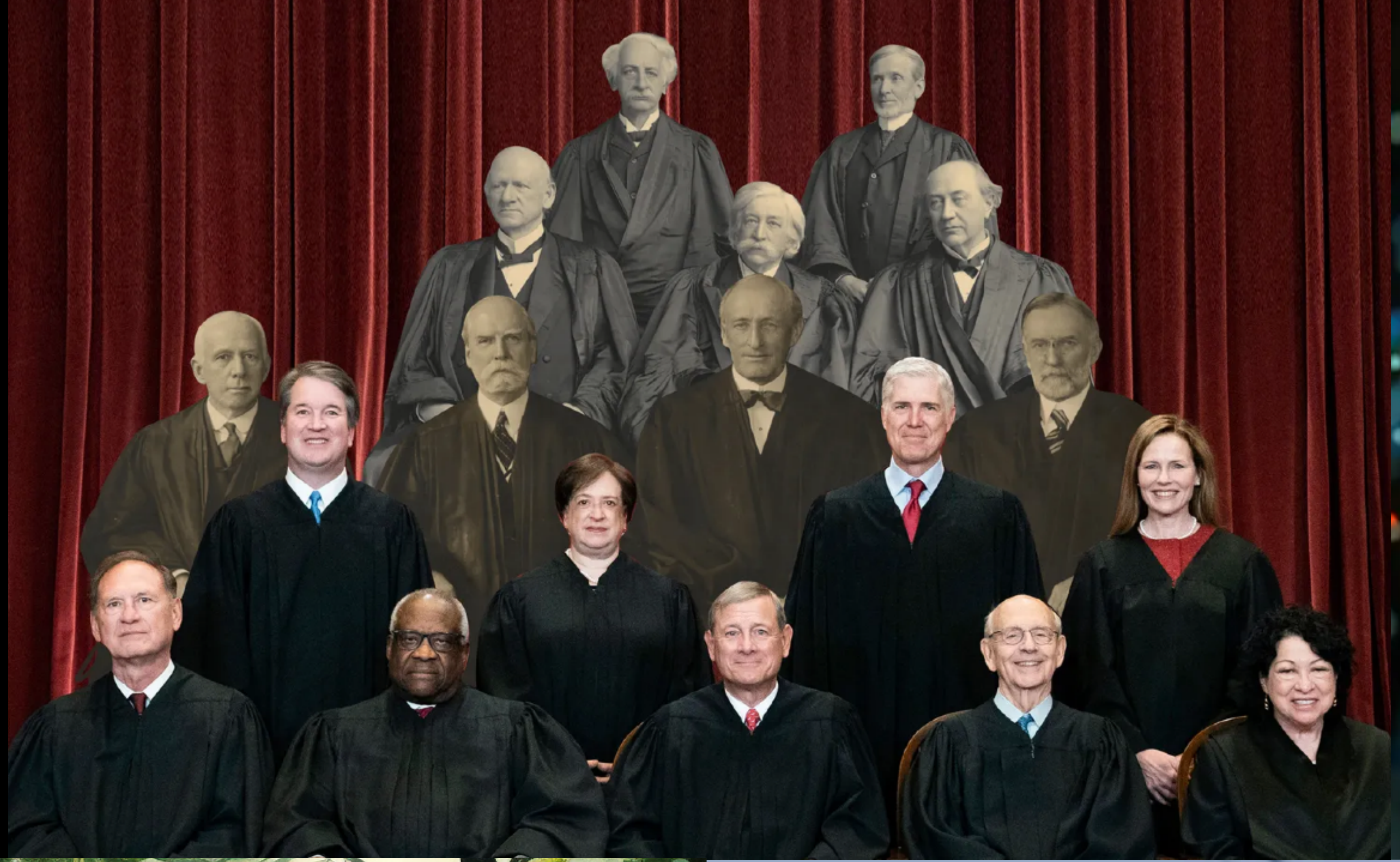
Eddie Lee
Team leader

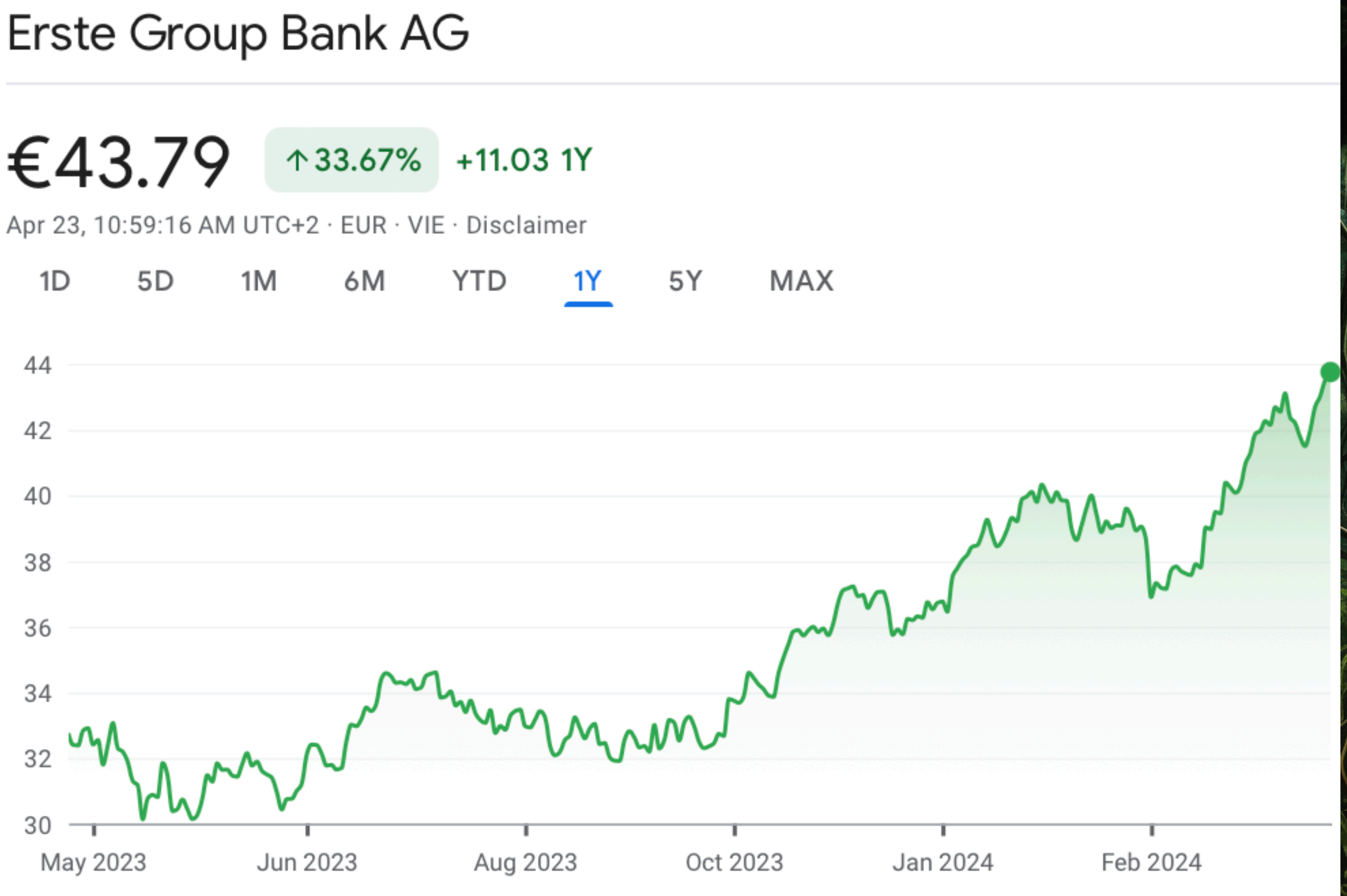
Ernesto Ortega
Postdoc
(2023-2024)

Niraj Kushwaha
PhD student

Evelyn
McGonigle
Research
intern
(2023)



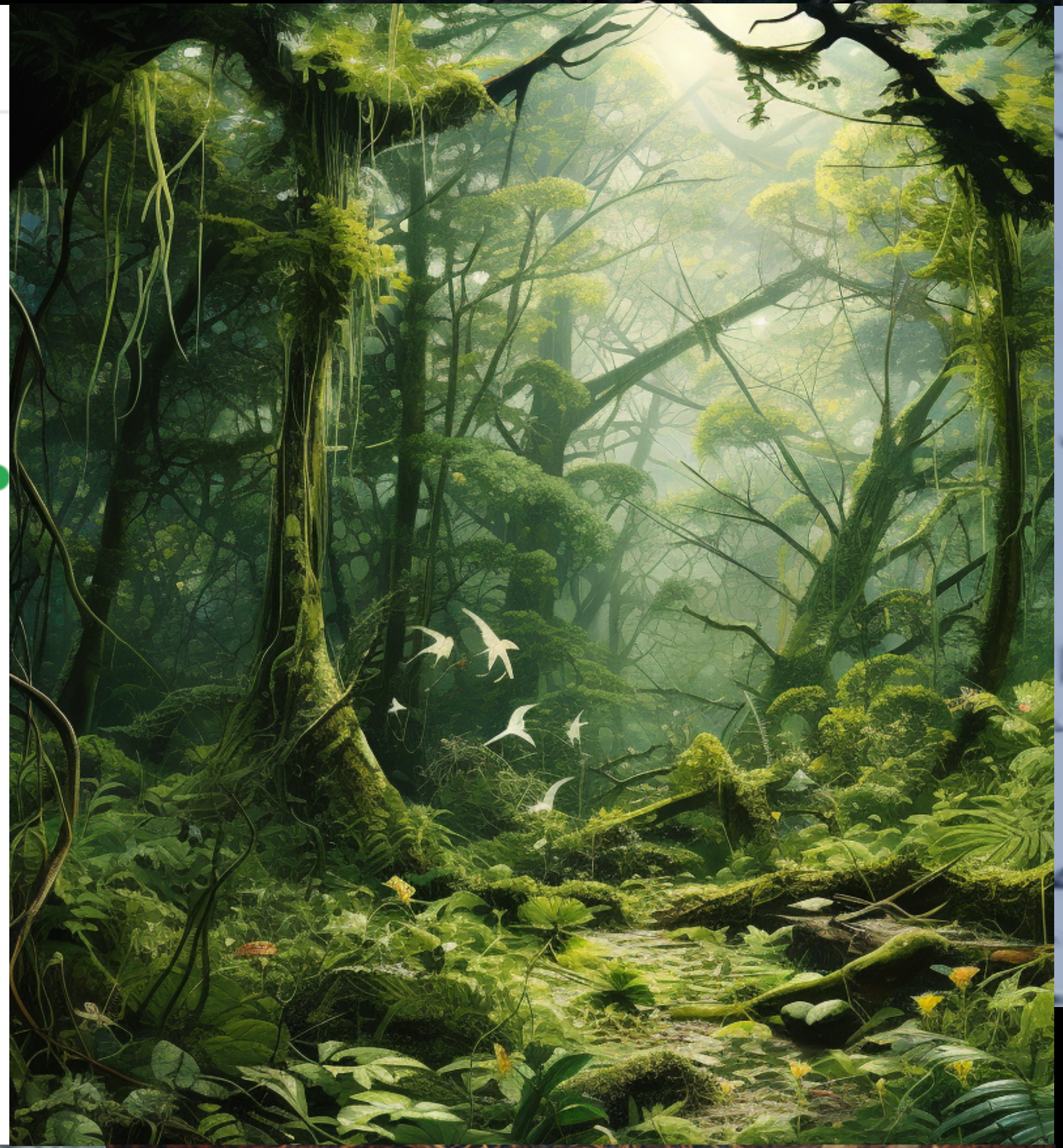


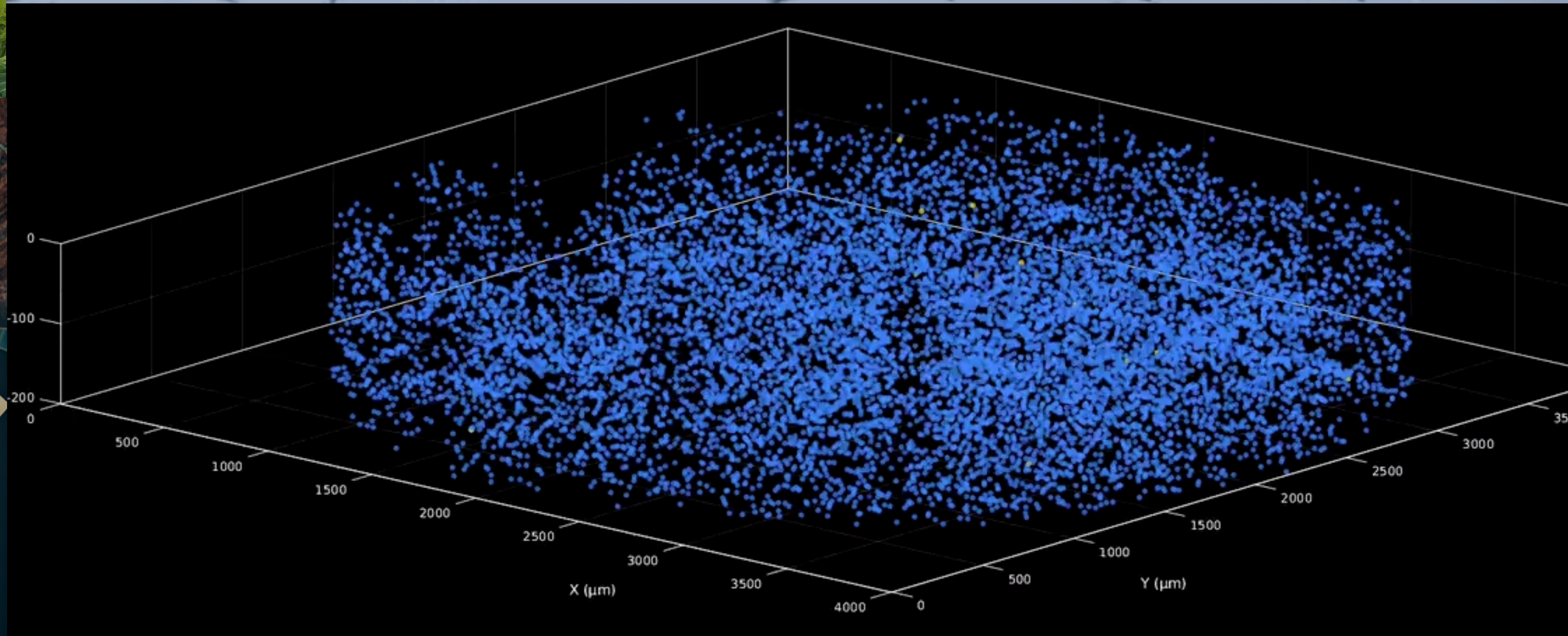


Erste Group Bank AG

€43.79 ↑33.67% +11.03 1Y

Apr 23, 10:59:16 AM UTC+2 · EUR · VIE · Disclaimer



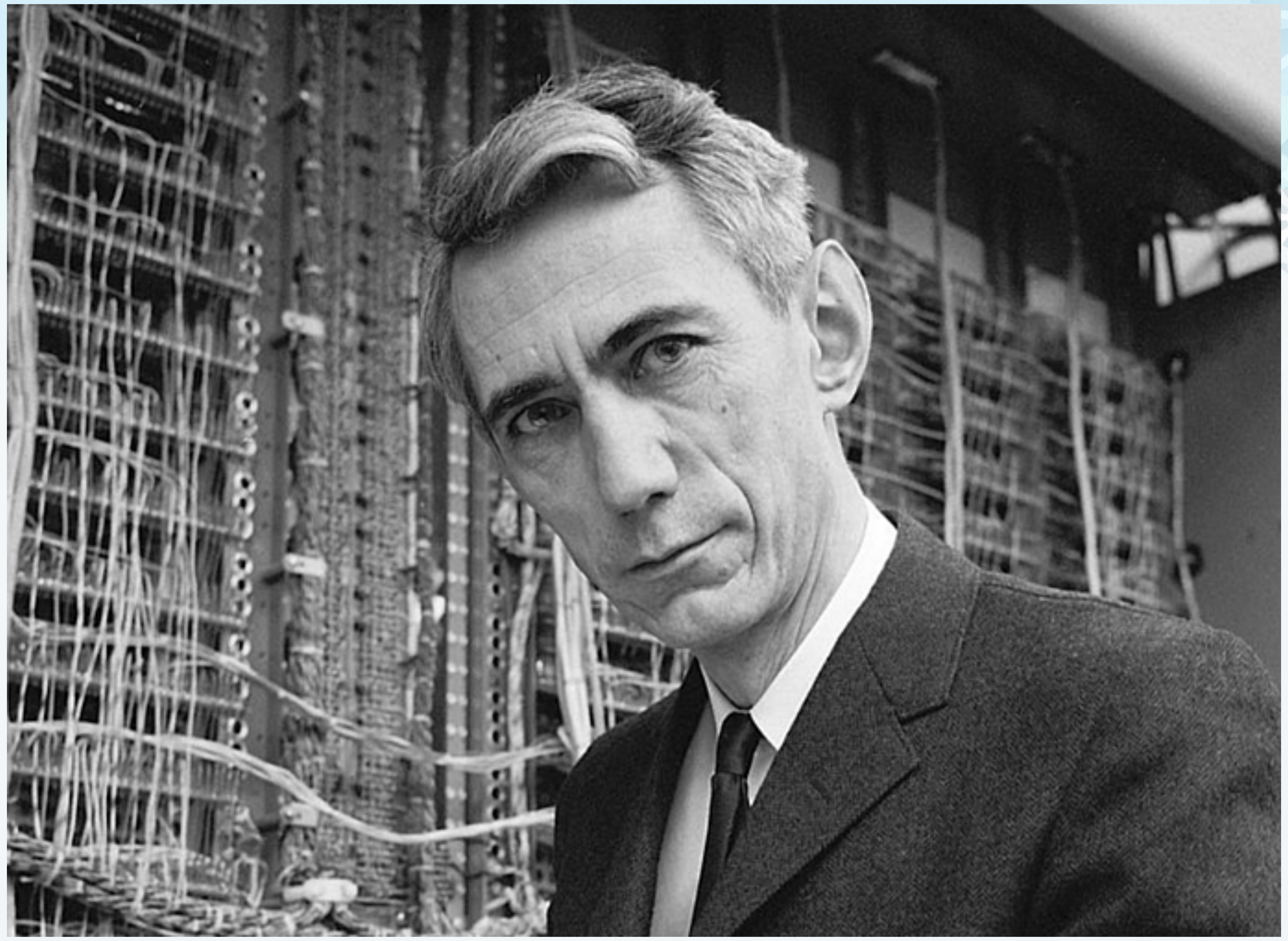


Claude Shannon & information theory

- Juggler and unicyclist
- Worked at Bell Labs
- Contributions to cryptography and artificial intelligence
- How do we characterize the information passing through a communication channel?
- A maximum (physical) rate at which information can be communicated through a channel
- Information entropy

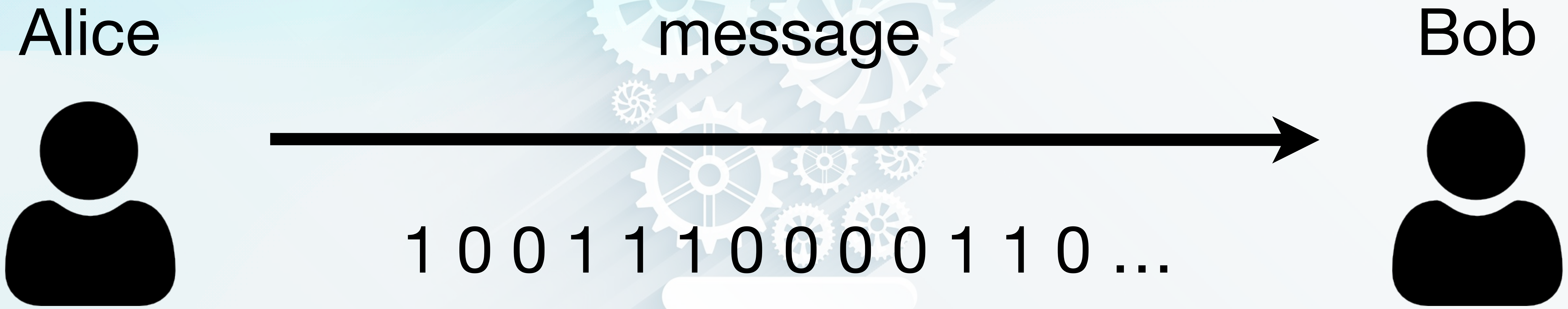
Claude Shannon & information theory

- Juggler and unicyclist
- Worked at Bell Labs
- Contributions to cryptography and artificial intelligence
- How do we characterize the information passing through a communication channel?
- A maximum (physical) rate at which information can be communicated through a channel
- Information entropy

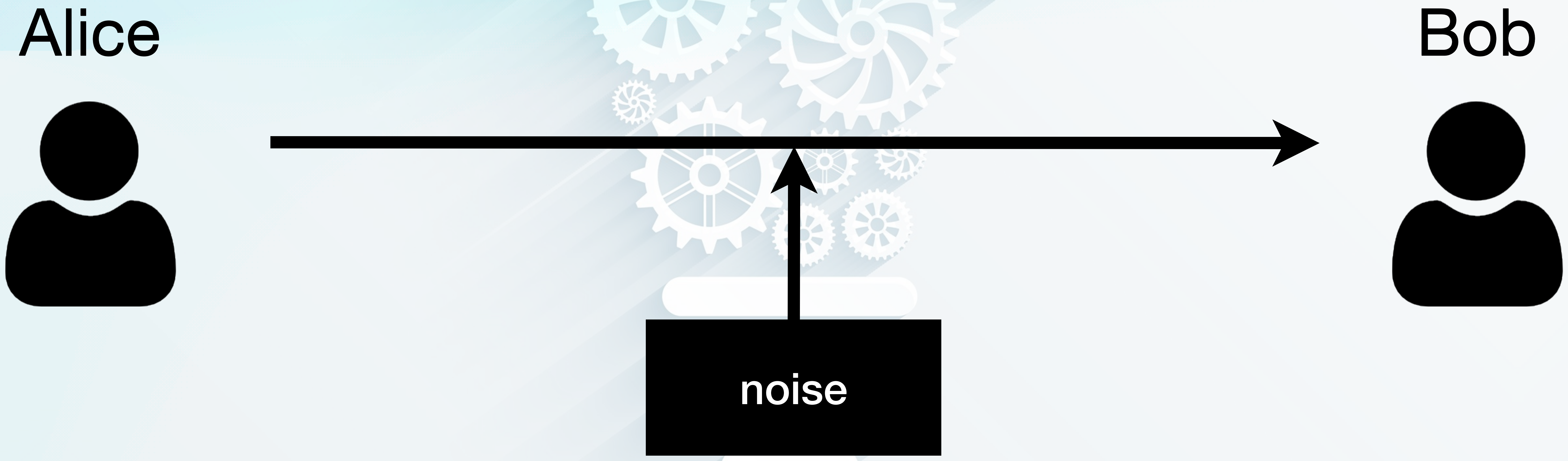


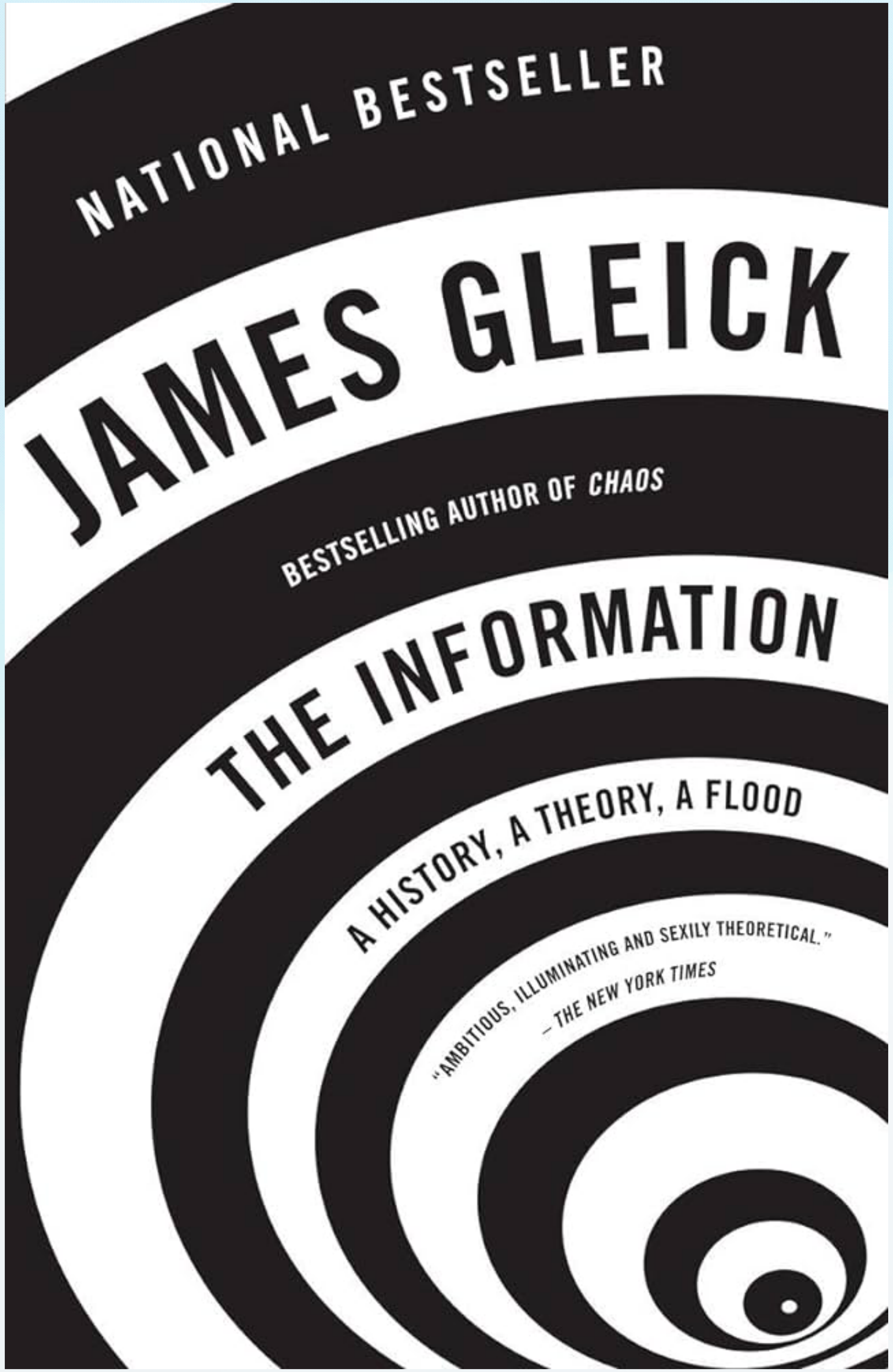
1916-2001

Information content of a message



Information content of a message




NATIONAL BESTSELLER
JAMES GLEICK
BESTSELLING AUTHOR OF *CHAOS*
THE INFORMATION
A HISTORY, A THEORY, A FLOOD
"AMBITIOUS, ILLUMINATING AND SEXILY THEORETICAL."
—THE NEW YORK *TIMES*

4 projects at various stages of development

inference

multiscale data analysis

computation

innovation & obsolescence

Sociology is the “science of institutions, their genesis and their functioning.”

–Émile Durkheim

Lee, E. D., Broedersz, C. P., Bialek, W. "Statistical Mechanics of the US Supreme Court." Journal of Statistical Physics, 160(2):275–301 (2015).

Lee, E. D. "Partisan Intuition Belies Strong, Institutional Consensus and Wide Zipf's Law for Voting Blocs in US Supreme Court." Journal of Statistical Physics 173(6):1722–1733 (2018).

Lee, E. D., Katz, D. M., Bommarito II, M. J., Ginsparg, P. H. Sensitivity of collective outcomes identifies pivotal components. Journal of The Royal Society Interface 17, (2020).

Lee, E. D. & Cantwell, G. T. Valence and interactions in judicial voting. Philosophical Transactions of the Royal Society A 20230140 (2024).

Rees, G. & Lee, E.D. Four dimensions of US legislative voting. In progress.

Image from *Cornell Chronicle*

US Supreme Court (SCOTUS)

- Highest US court
- President nominates, Senate confirms
- 9 justices at most
- Selects own docket
- Decisions by majority vote
- “natural court” is a set of justices who sat on the bench together

Image from *Cornell Chronicle*

Maxent modeling with the US Supreme Court

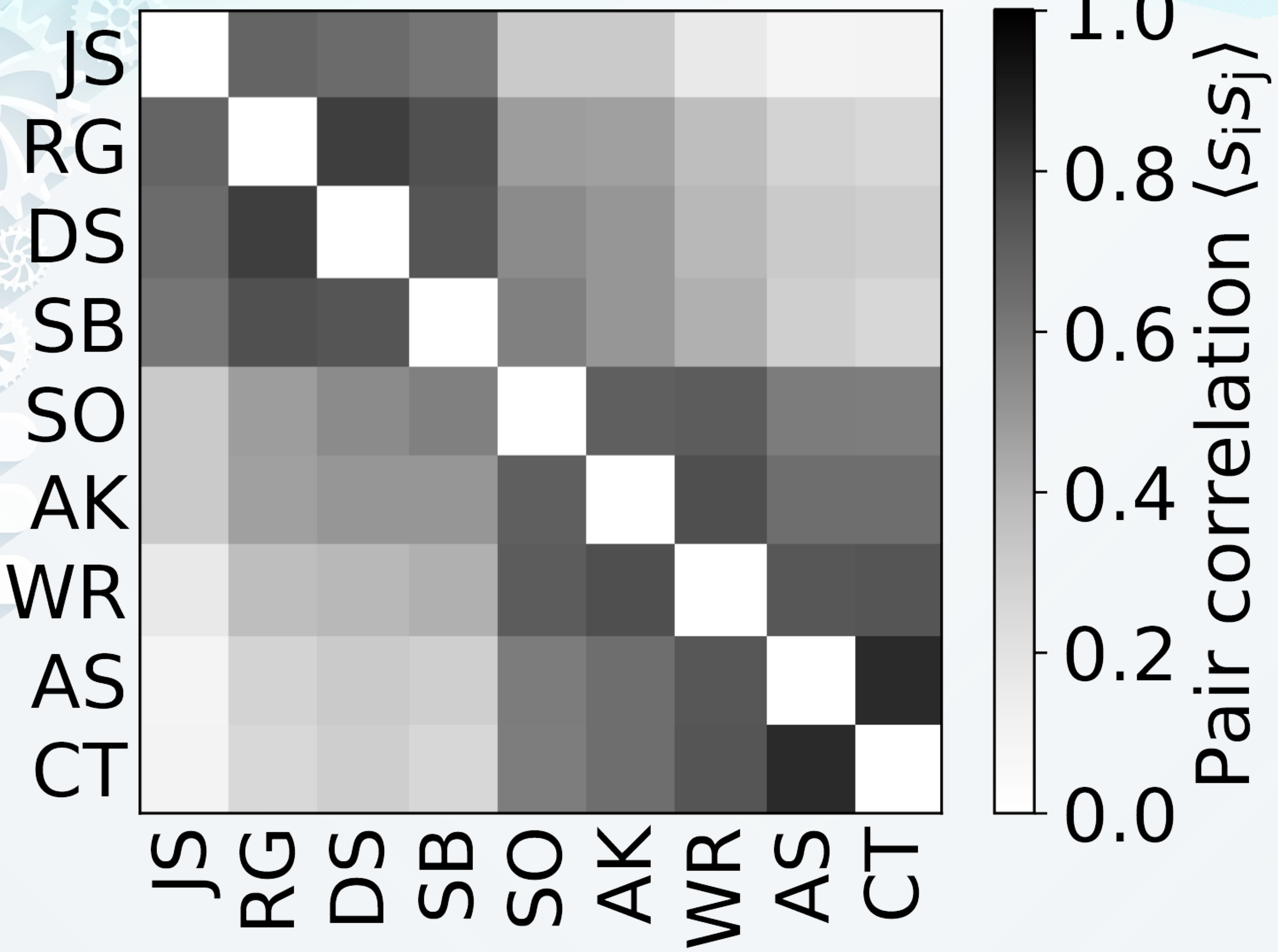
$$\begin{aligned}\vec{s} &= \{\text{yea}, \text{nay}, \text{nay}, \text{nay}, \text{nay}, \text{nay}, \text{yea}, \text{yea}, \text{yea}\}\\ &= \{1, -1, -1, -1, -1, -1, 1, 1, 1\}\end{aligned}$$

Correlations imply interesting behavior

$$\vec{s} = \{\text{yea}, \text{nay}, \text{nay}, \text{nay}, \text{nay}, \text{nay}, \text{yea}, \text{yea}, \text{yea}\}$$

$$= \{1, -1, -1, -1, -1, -1, 1, 1, 1\}$$

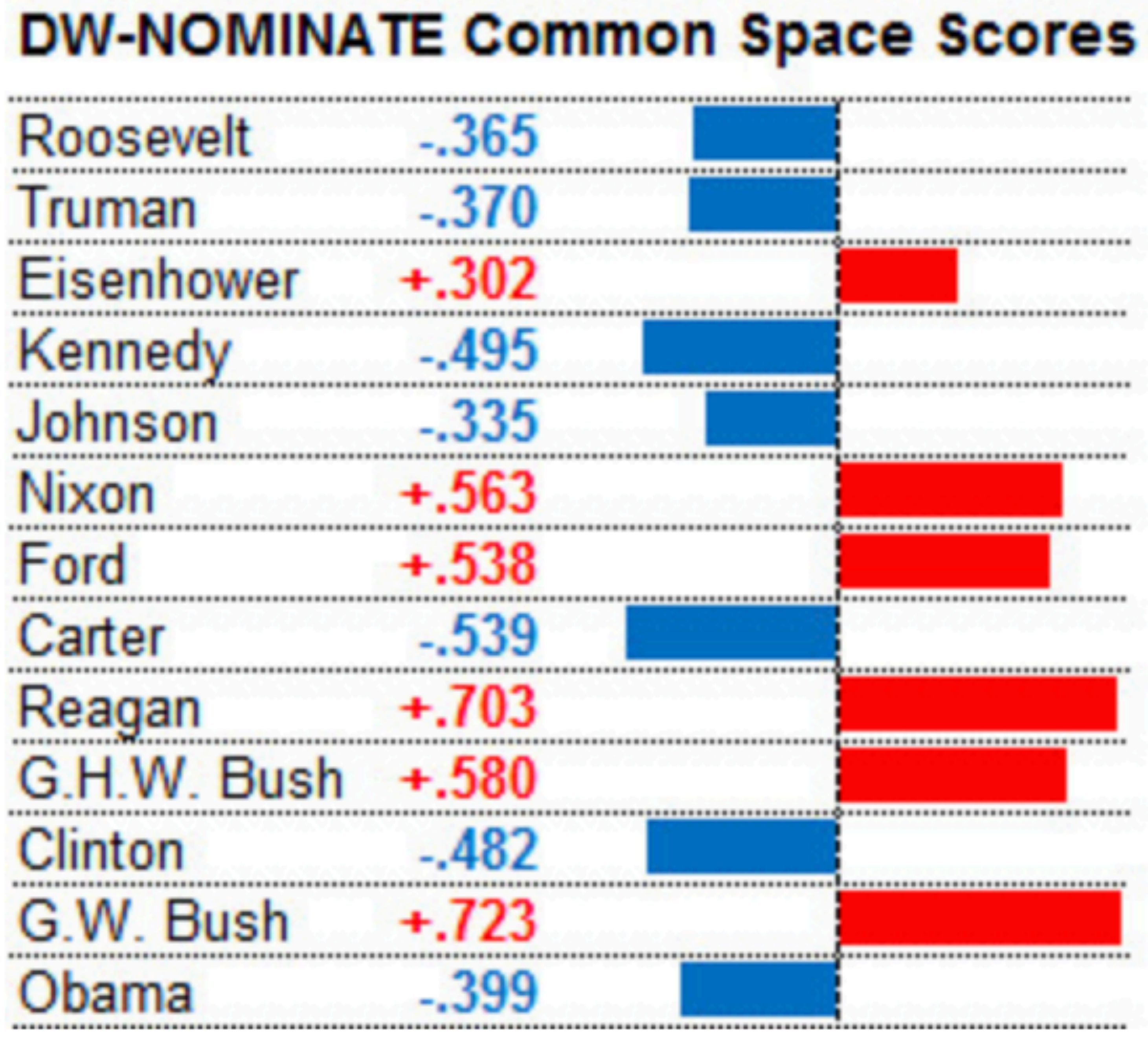
$$\langle s_i s_j \rangle = \sum_s p(\vec{s}) s_i s_j$$



Lee, Broedersz, Bialek, *JSoP* (2015)

(D)W-Nominate

- Spatial voting model originating in 1980s
- Assumption of independent voters maximizing utility along different political issues
- Equivalent to a kernel regression technique, Gaussian processes with radial basis function solved by maximizing the posterior
- Very parameter heavy, a problem with sparse voting data



DW-NOMINATE Common Space Scores for Recent Presidents

Roosevelt	-.365
Truman	-.370
Eisenhower	+.302
Kennedy	-.495
Johnson	-.335
Nixon	+.563
Ford	+.538
Carter	-.539
Reagan	+.703
G.H.W. Bush	+.580
Clinton	-.482
G.W. Bush	+.723
Obama	-.399

FiveThirtyEight
New York Times

Parameter counting with N=36 voters

Taking all justices from 1946-2016

K=8,737
D=10

[number of voters] x [number of dimensions] + [number of votes] x [number of dimensions] x 2

~ 100,000 parameters

Maximum entropy principle

- there are an infinity of models, this chooses a unique model that is also minimal as a quantitative form of Occam’s razor

Maximum entropy principle

- there are an infinity of models, this chooses a unique model that is also minimal as a quantitative form of Occam’s razor
- maximize the entropy while fitting a limited set of important features of the system (Lagrangian multipliers from multivariable calculus)

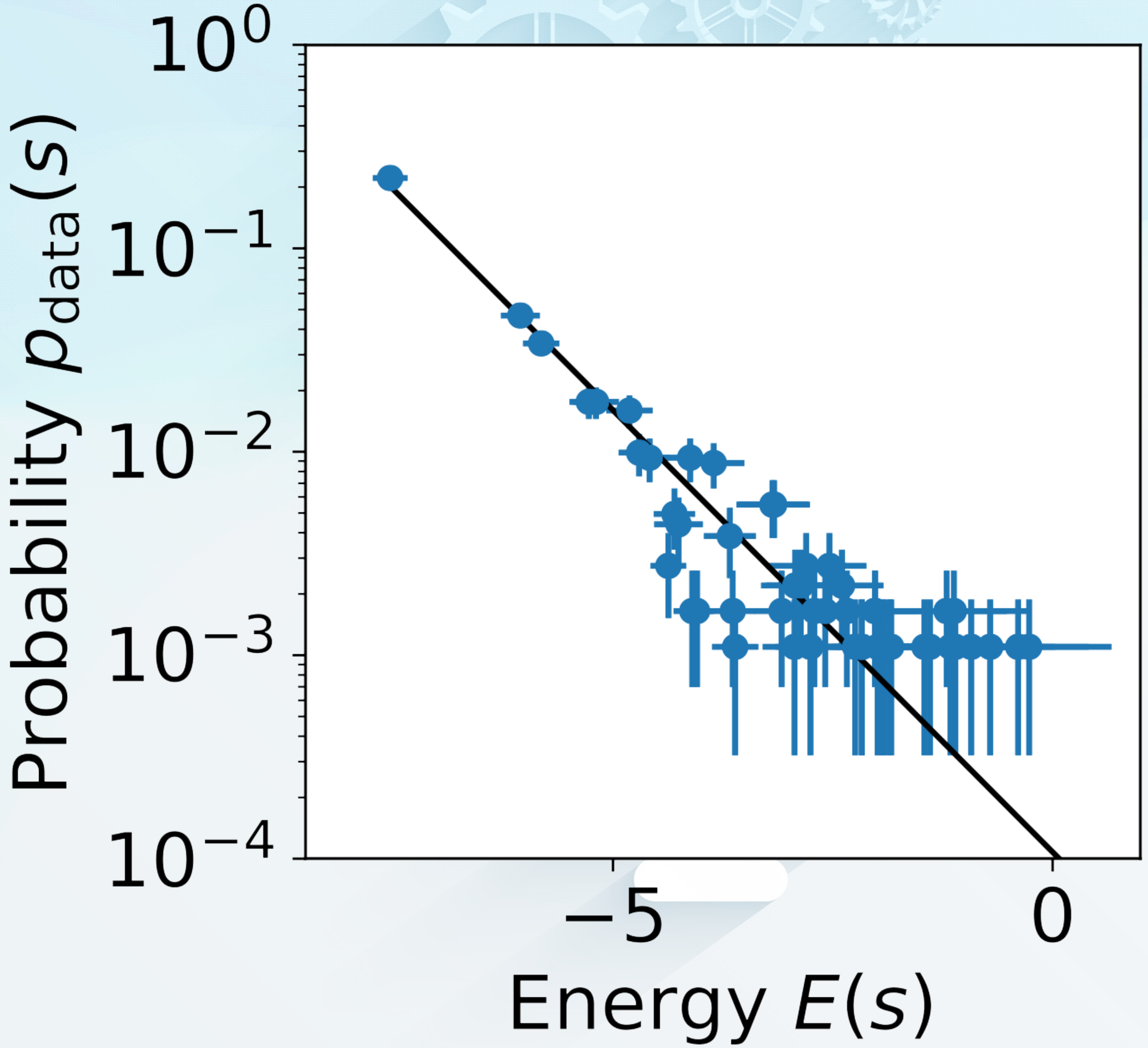
Maximum entropy principle

- there are an infinity of models, this chooses a unique model that is also minimal as a quantitative form of Occam's razor
- maximize the entropy while fitting a limited set of important features of the system (Lagrangian multipliers from multivariable calculus)
- leads to a "Boltzmann" probability distribution (a.k.a. exponential family, maximum entropy, restricted Boltzmann machines)

$$p(s) = \frac{e^{-E(s)}}{Z}$$

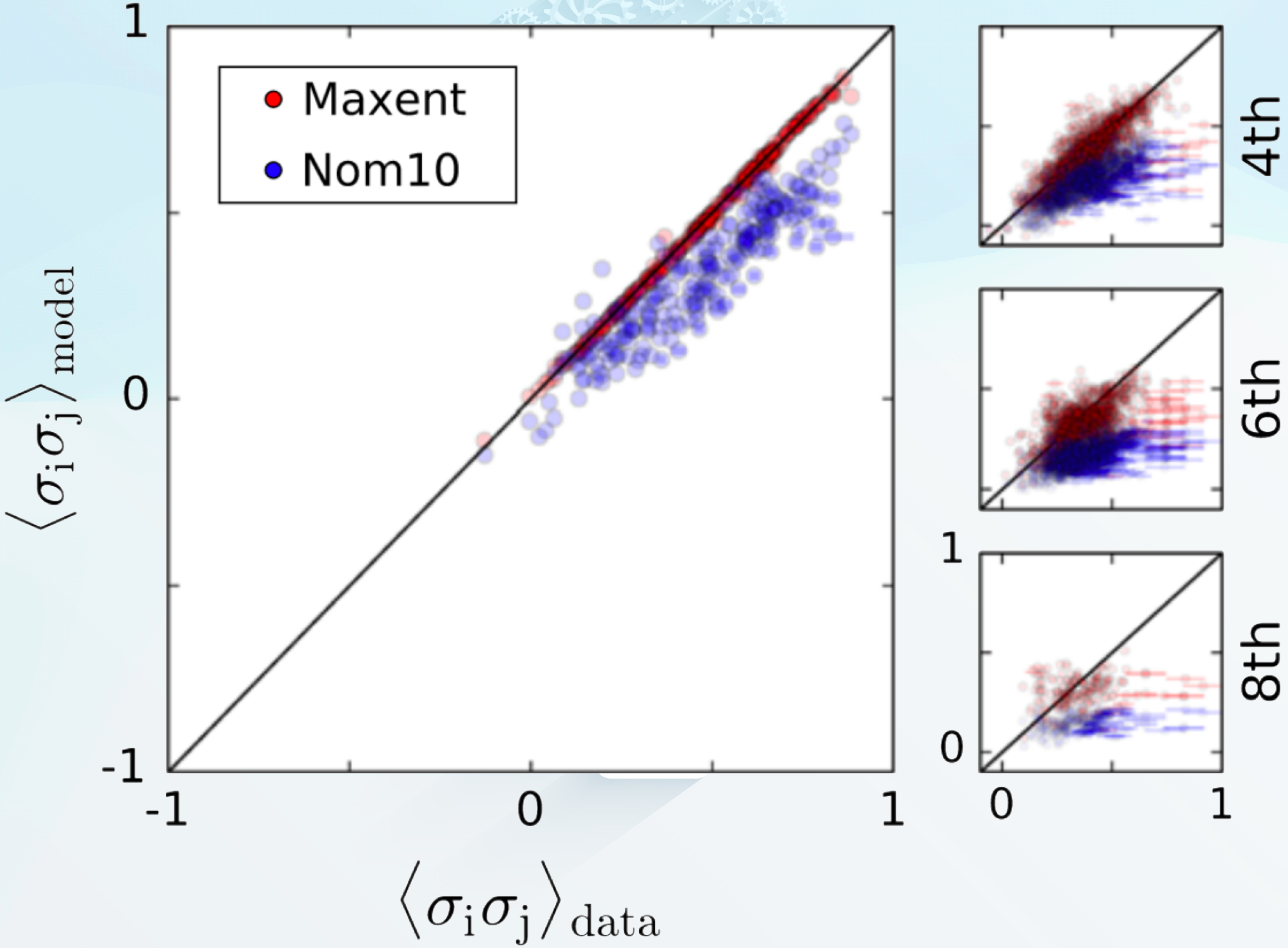
accounting for interactions, 630 parameters

Model captures entire distribution



Lee, Broedersz, Bialek, *JSoP* (2015)

Maxent performs better than W-Nominate



a powerful (elucidating and transparent) way of building models with interactions

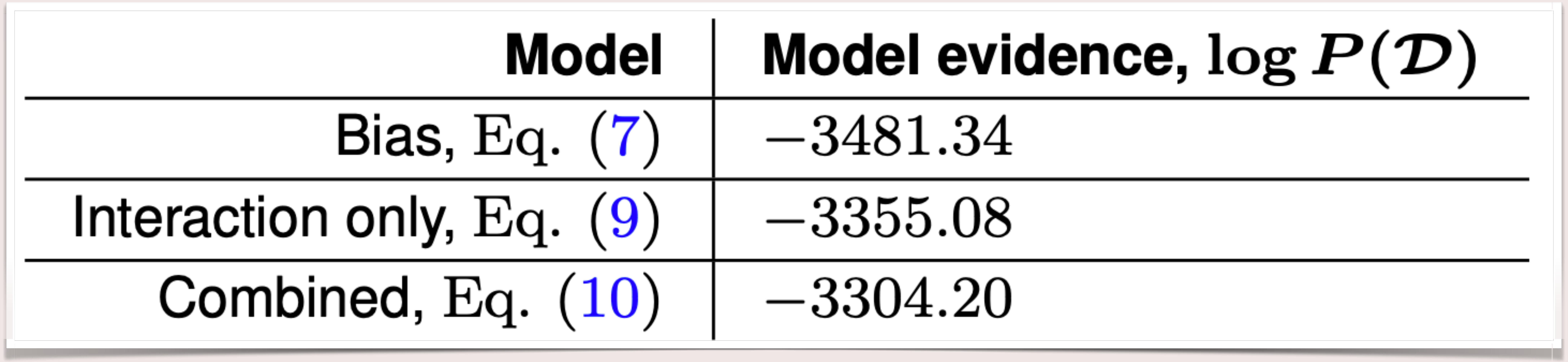
Valence and interactions in judicial voting

Edward D. Lee, Complexity Science Hub
George Cantwell, Cambridge University

Philosophical Transactions Royal Society A

Lee, Broedersz, Bialek (2015)
Lee (2018)
Lee, Katz, Bommarito, Ginsparg (2020)

Testing the model on the Second Rehnquist Court

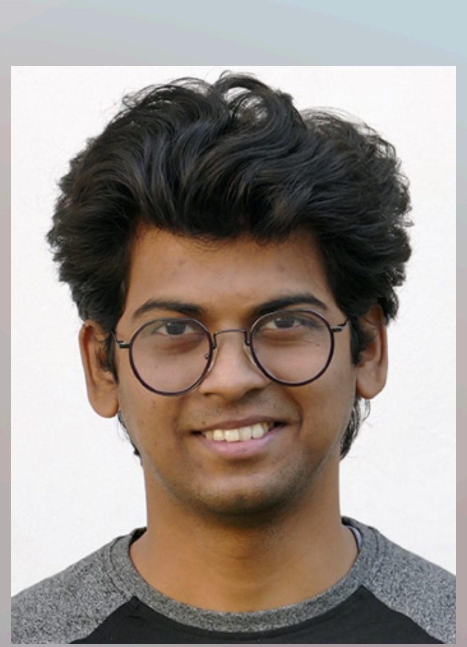


Model	Model evidence, $\log P(\mathcal{D})$
Bias, Eq. (7)	-3481.34
Interaction only, Eq. (9)	-3355.08
Combined, Eq. (10)	-3304.20

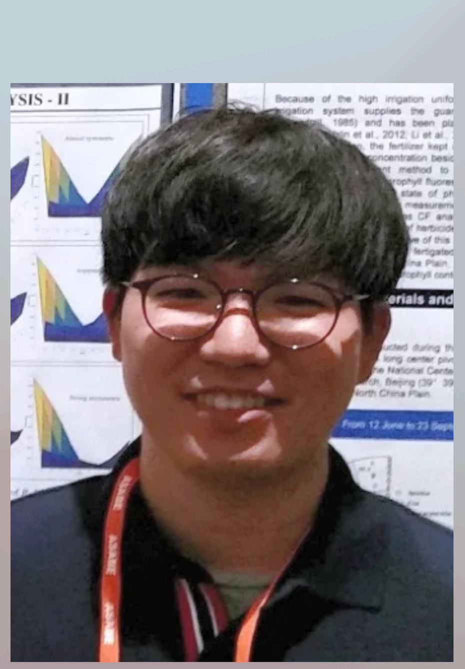
Next steps with analysis and experiments

- identify coalitions
- identify important players (“Partisan intuition” in 2018)
- identify sensitive points in the system that would tip the balance (“Sensitivity of collective outcomes identifies pivotal voters” in 2019; “Discovering sparse control strategies in neural activity” in 2022)
- go beyond one-dimensional ideology (Rees & Lee, in progress)

Cascades of conflict activity



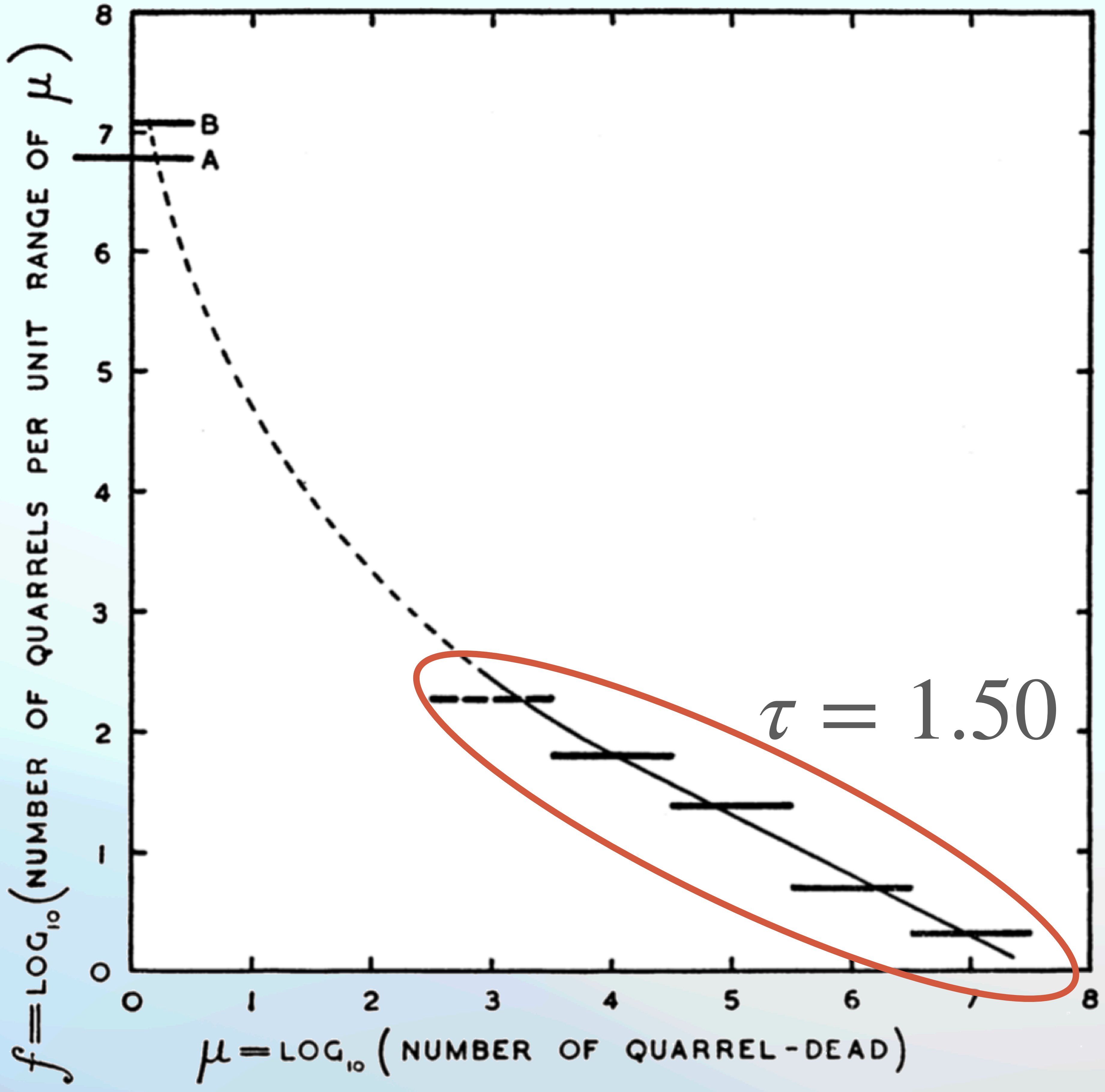
Niraj Kushwaha



Woi Oh Sok

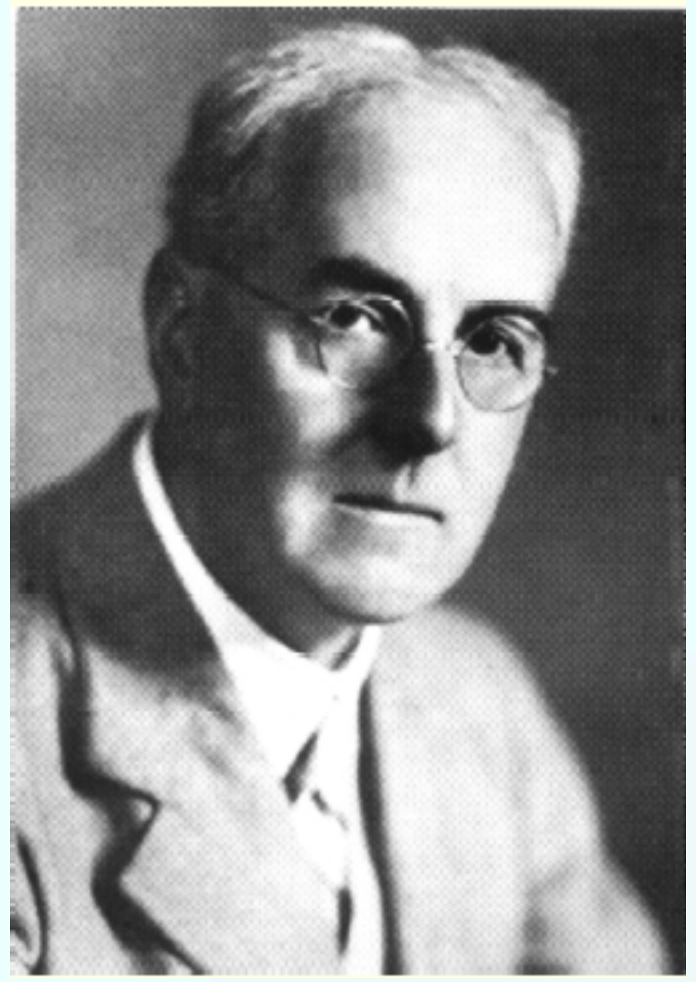
Richardson's law for fatalities

Interstate wars from 1820–1945



Normalized frequency $P(F)$

Number of fatalities F



Lewis Fry Richardson
(1881-1953)

Power law distribution

$$P(F) \sim F^{-\tau}$$

Richardson, *Nature* (1948)

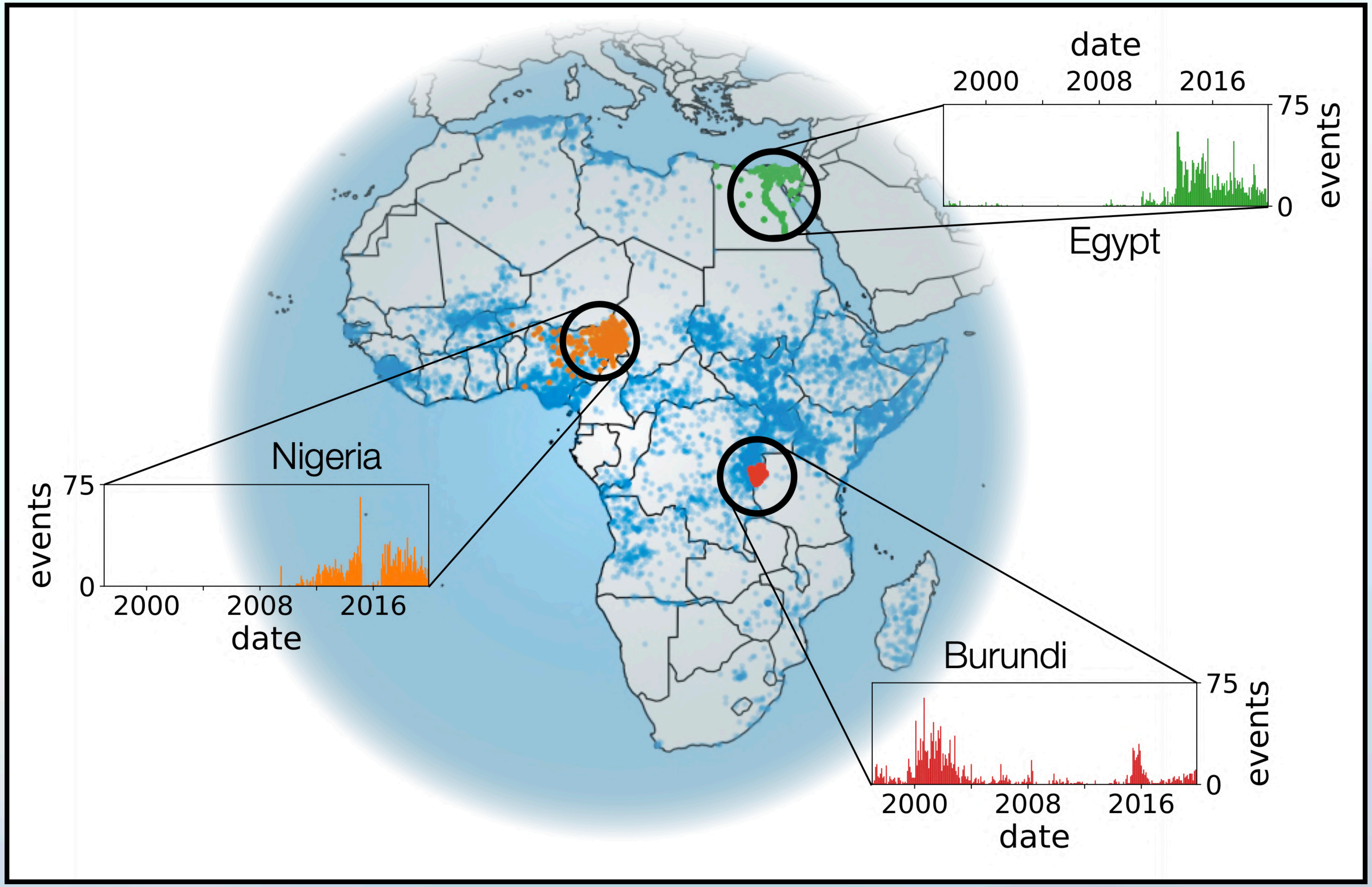
World War II

Leningrad (Sept. 1941–Jan. 1944)
SOVIET UNION
Kiev
to Stalingrad (Aug. 1942–Jan. 1943)
HUNGARY
SWITZ.
55°N
45°N

Libyan Civil War

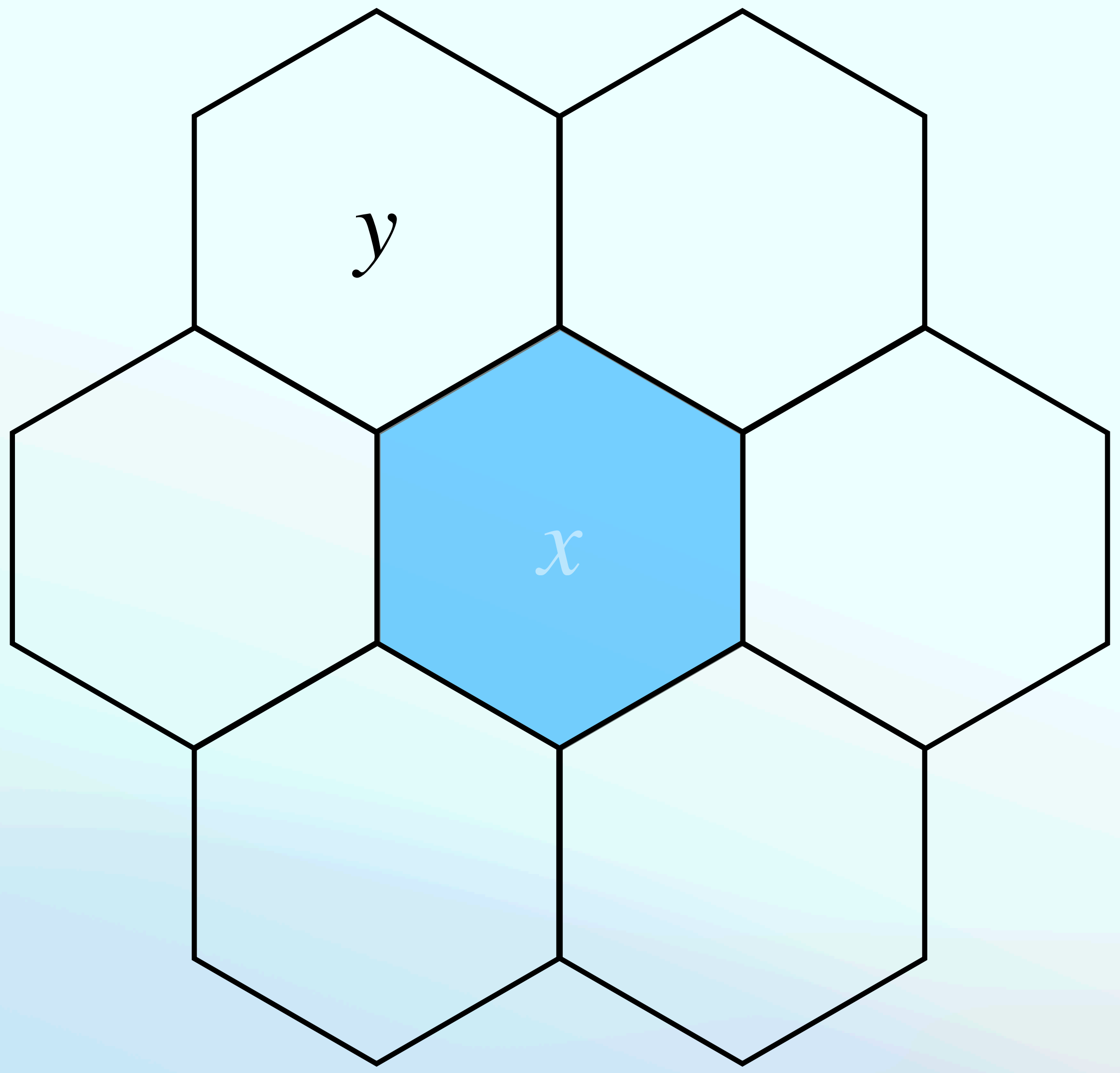
Armed Conflict & Location Event Data Project

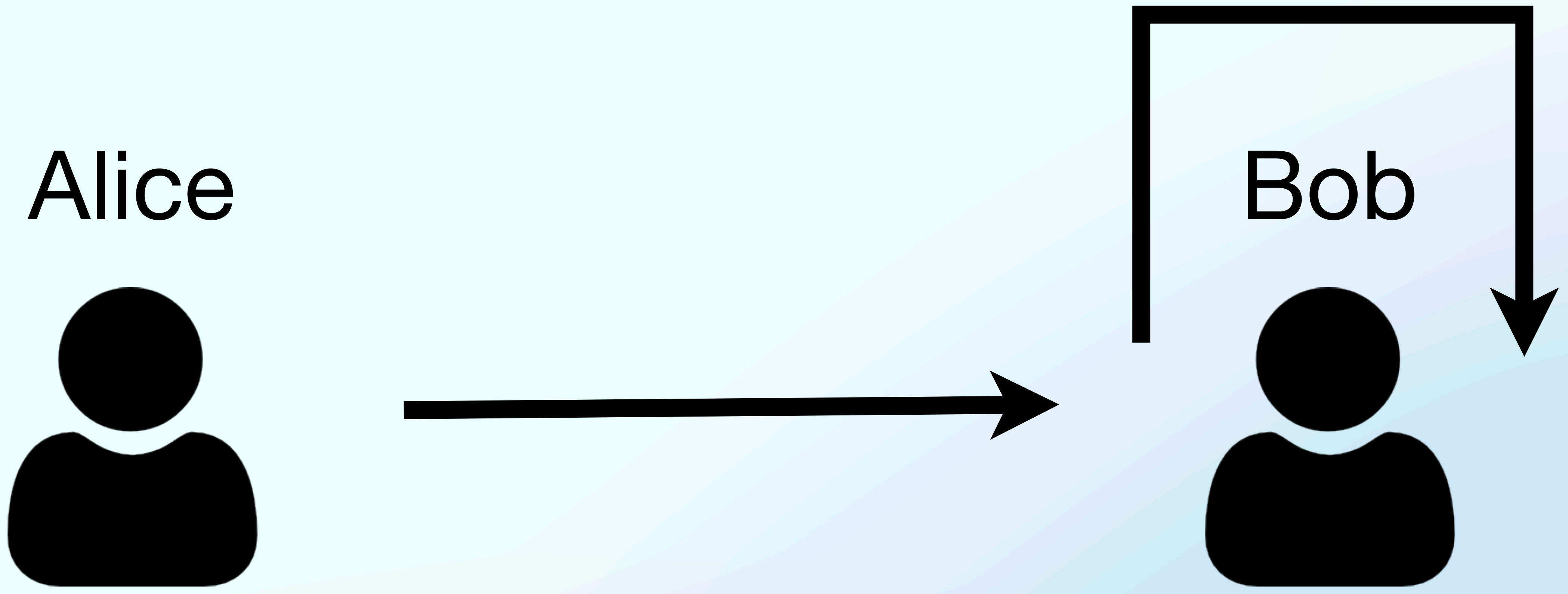
- 27 years (1997–today)
- Spans Africa (~8,000 km)
- 400,000 events (reports, fatalities, day, location)
- >700,000 direct fatalities



Raleigh, Linke, Hegre, Karlsen, J Peace Res (2010)
Figure from Kushwaha & Lee, PNAS Nexus (2023)

Inferring relationships between sites i and j





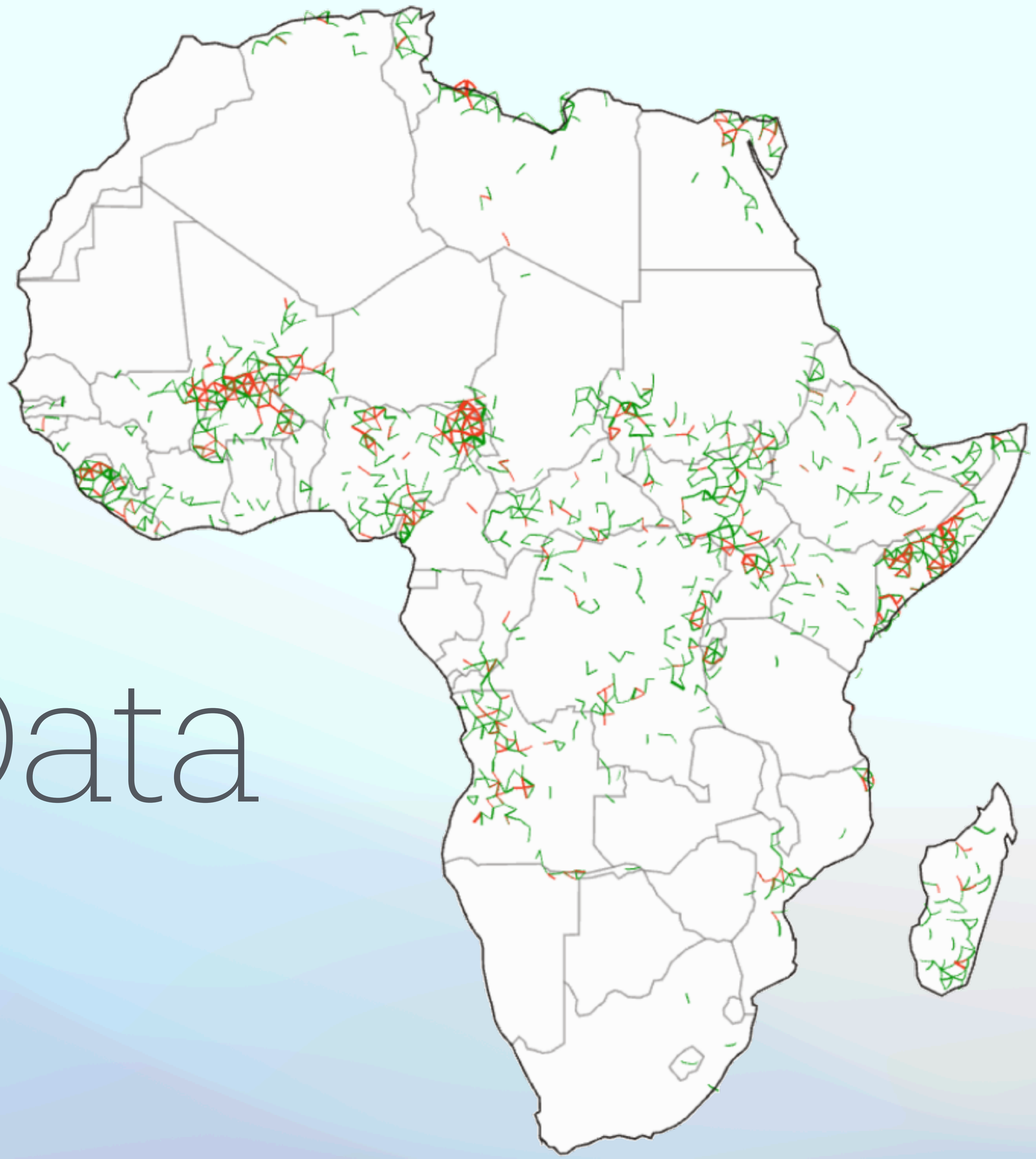
🟩 Causal link (uni-directional)

🟥 Causal link (bi-directional)

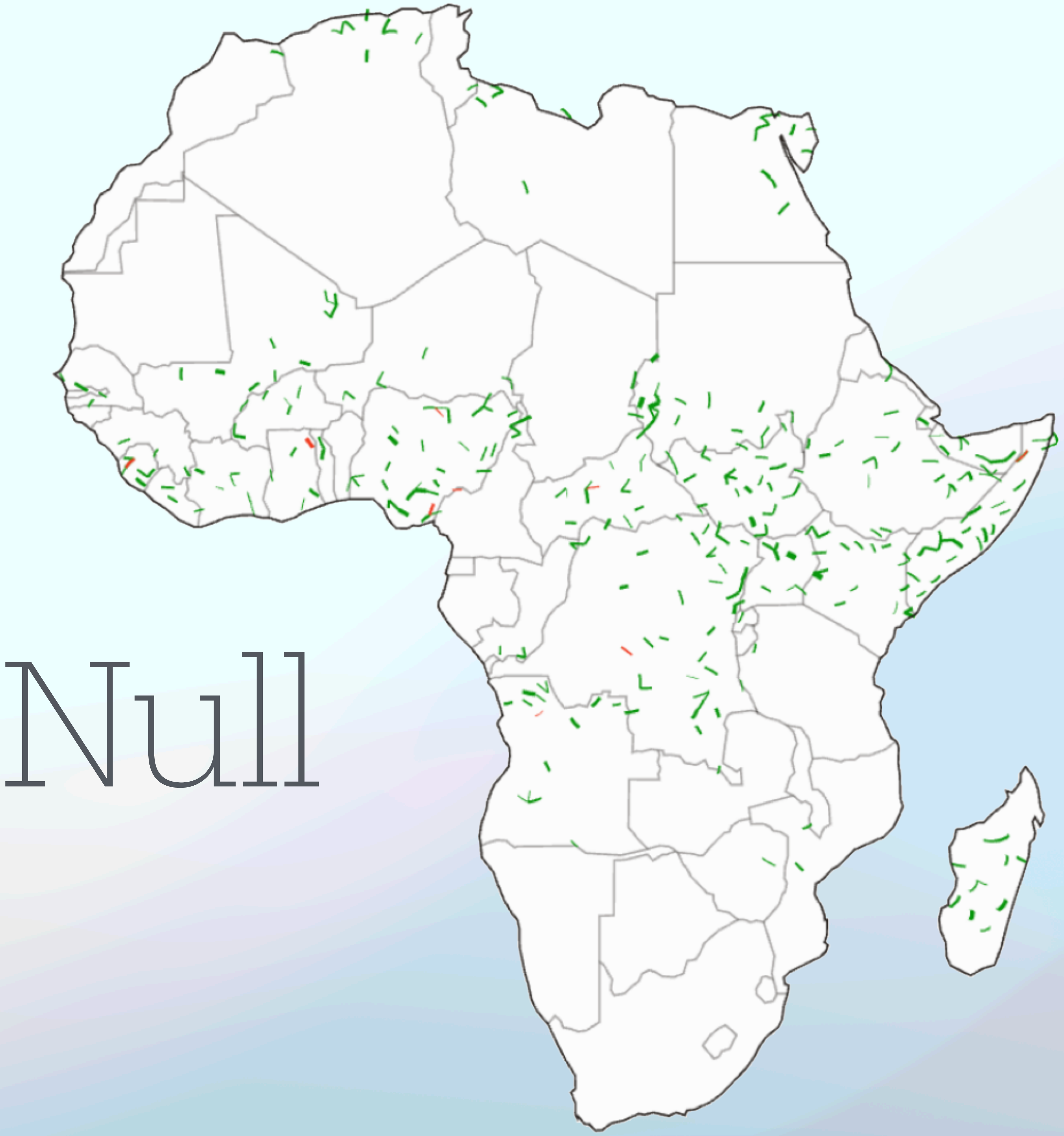
● Conflict event

Kushwaha & Lee, *PNAS Nexus* (2023)

Conflict network



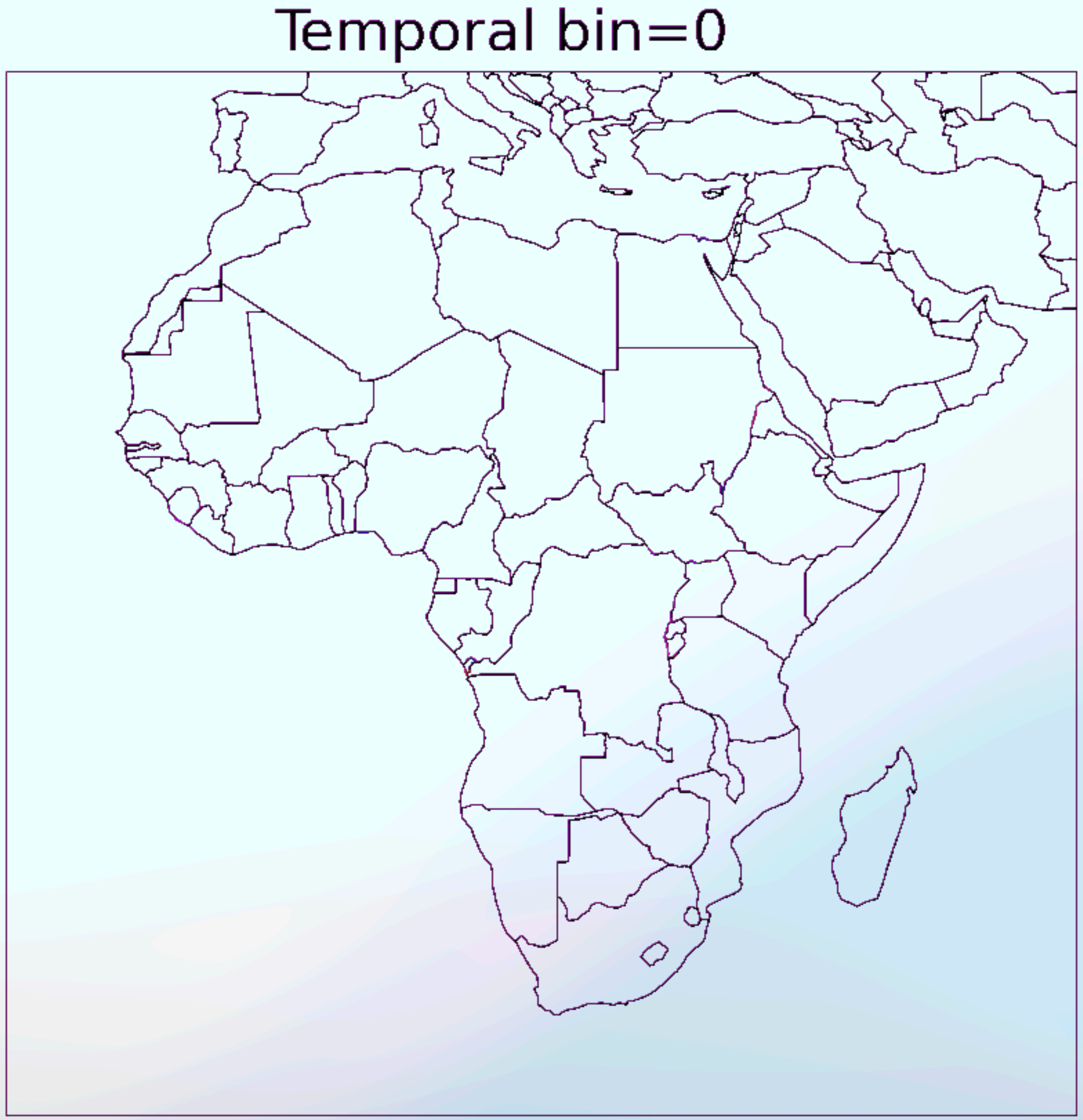
Data



Null

Kushwaha & Lee, *PNAS Nexus* (2023)

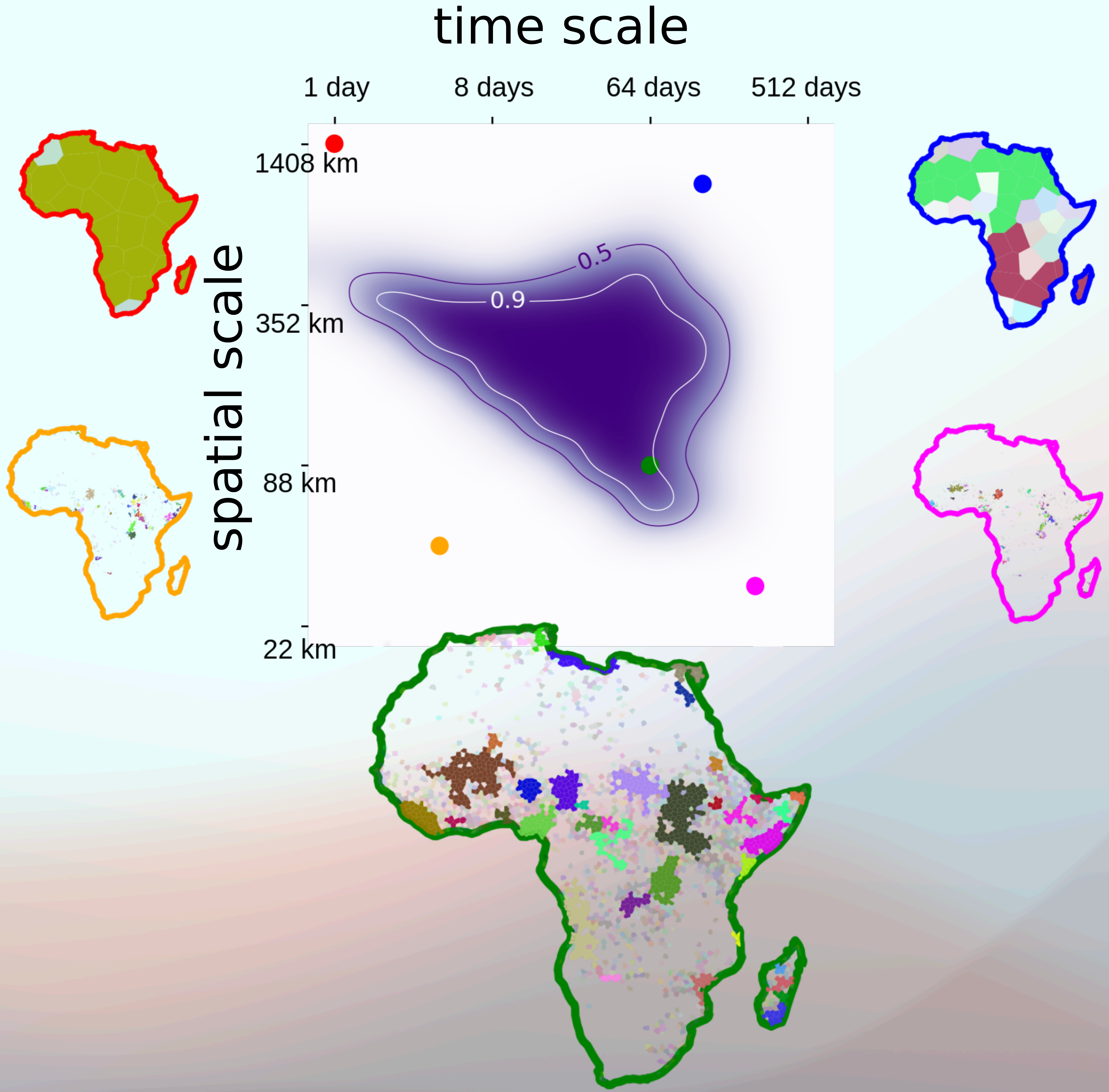
Conflict avalanches



(01 Jan 1997 - 05 Mar 1997)

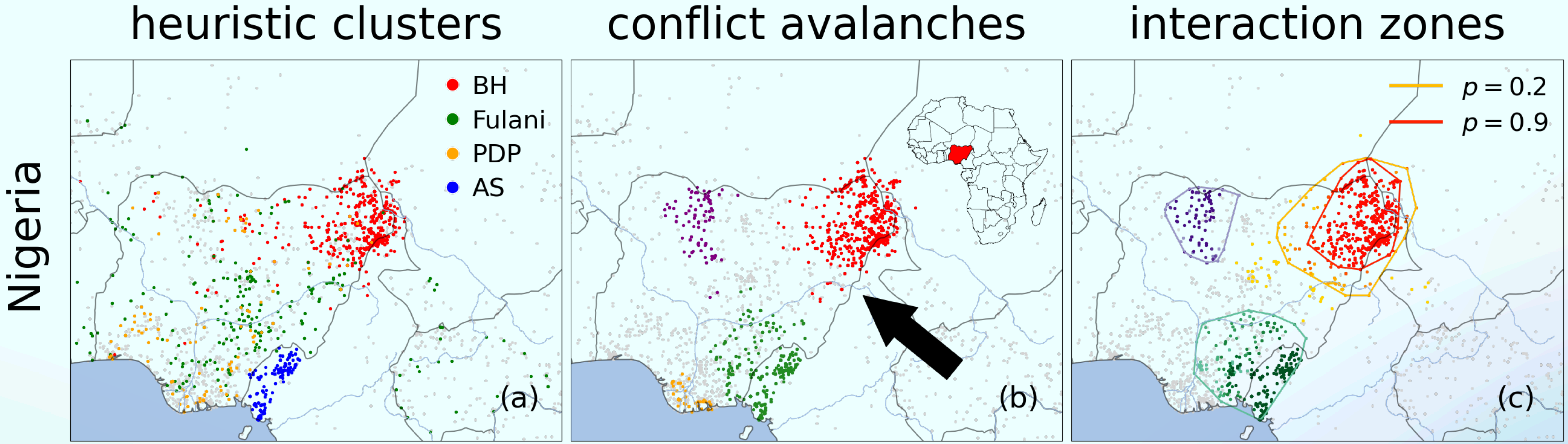
Kushwaha & Lee, *PNAS Nexus* (2023)

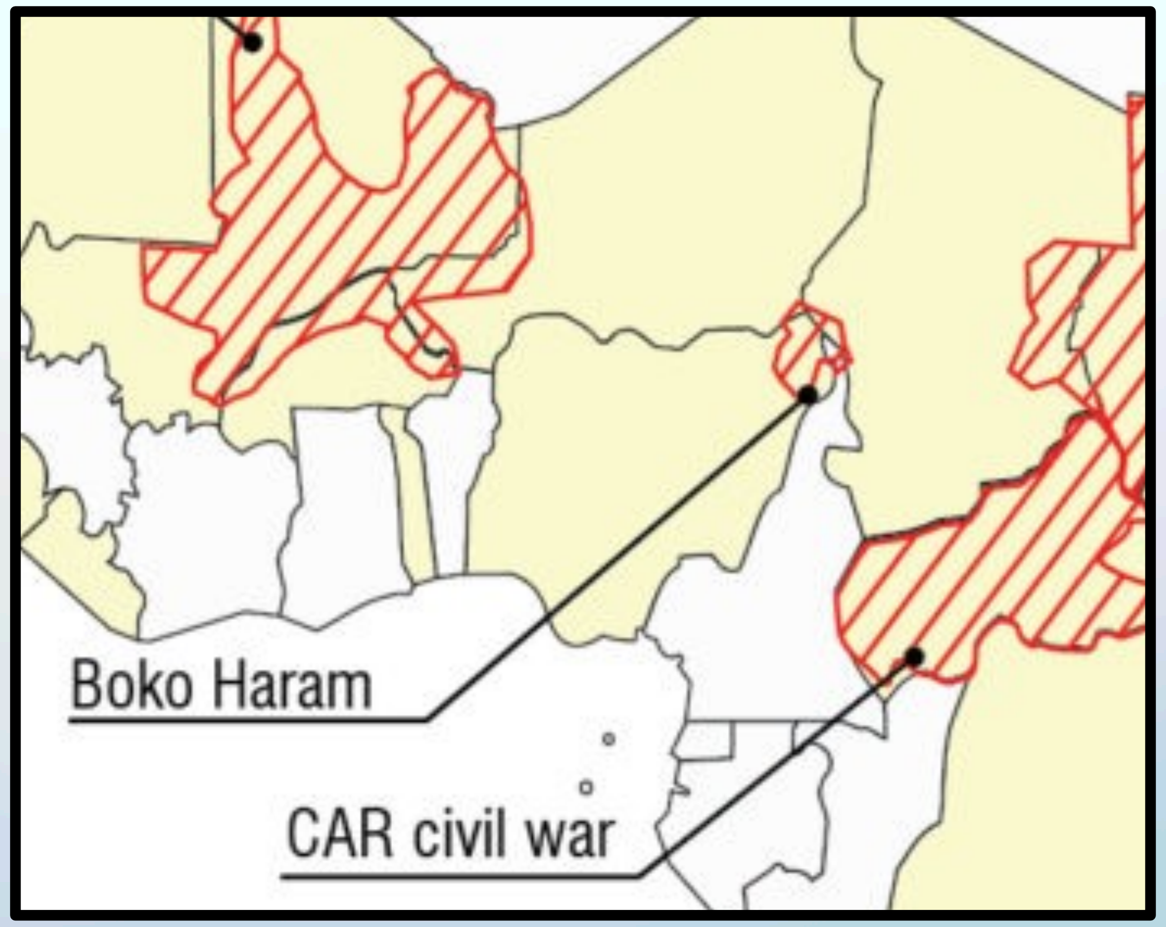
Discovering the mesoscale



Kushwaha & Lee, *PNAS Nexus* (2023)

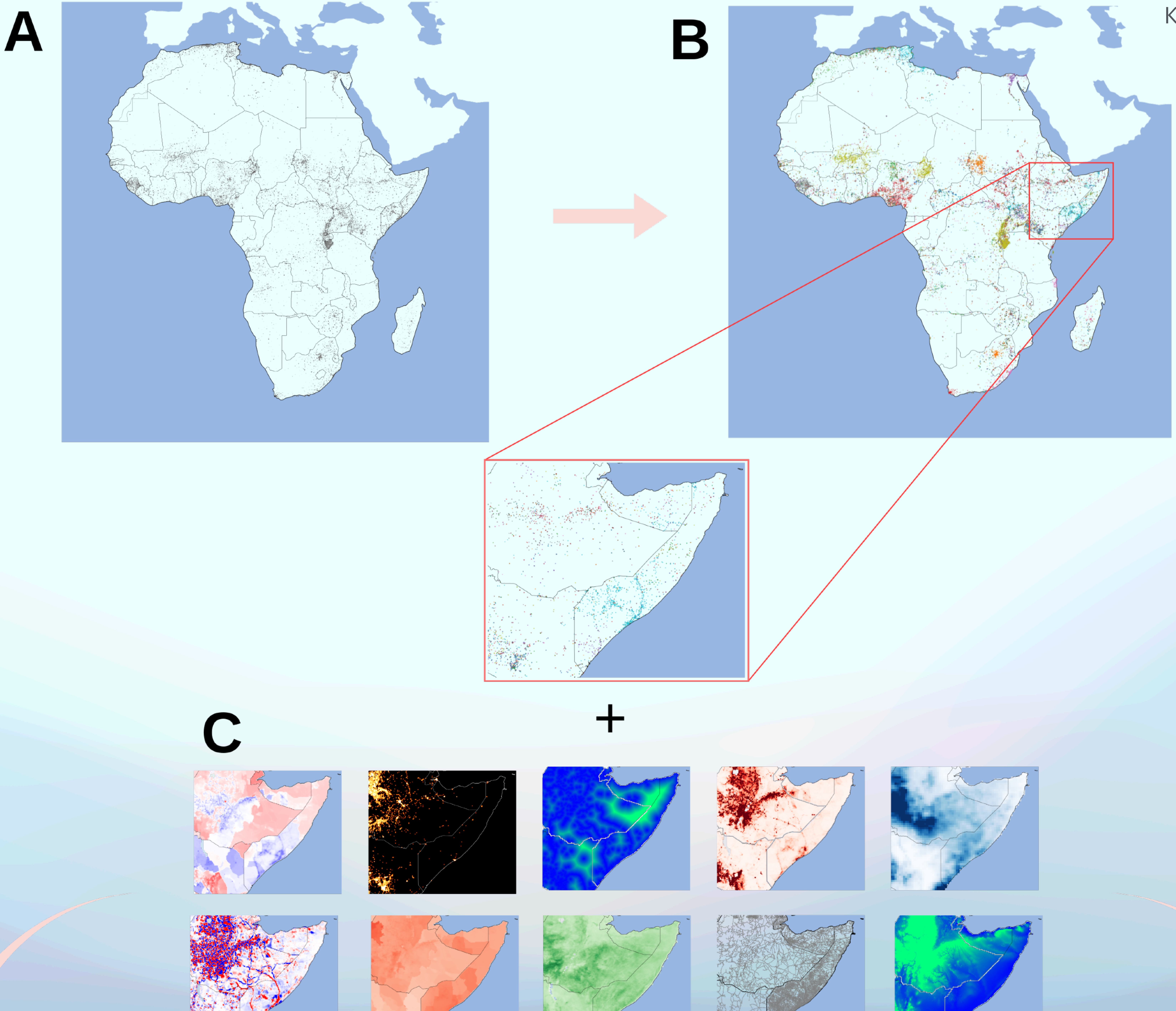
Systematic clusters discover mechanism



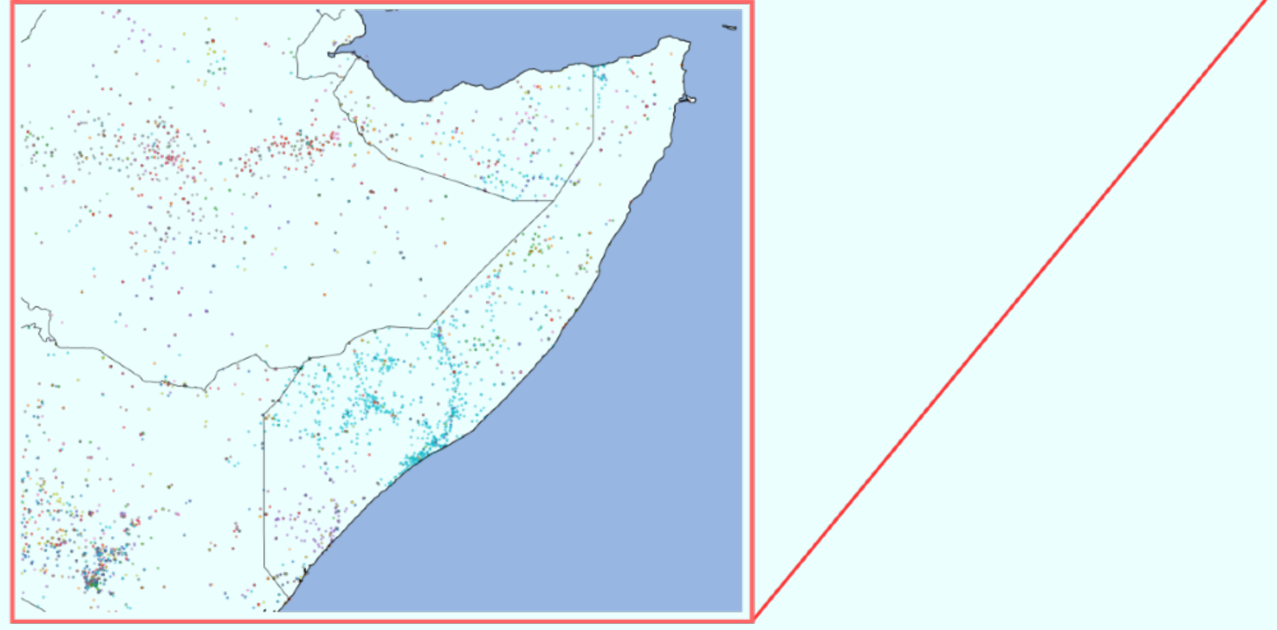


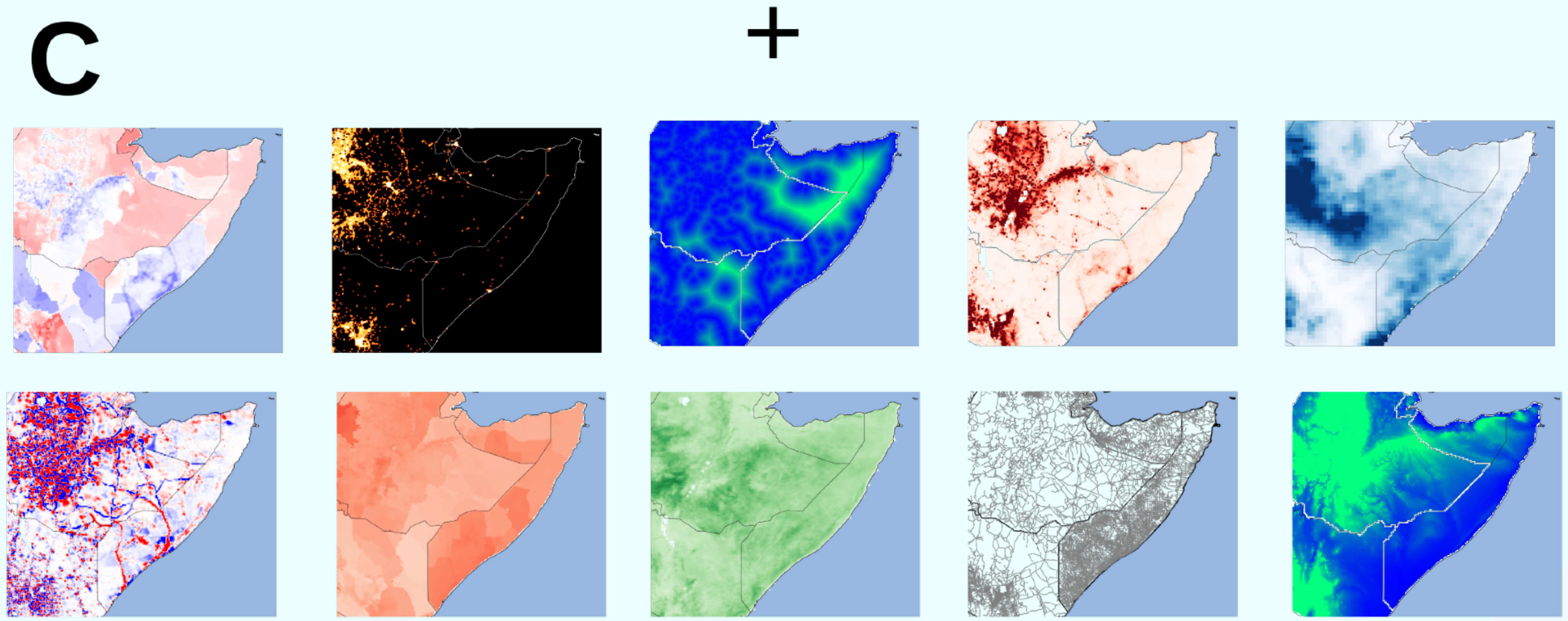
Kushwaha & Lee, *PNAS Nexus* (2023)

Kushwaha and Lee (in progress)



Kushwaha and Lee (in progress)





.

.

.

Climatic variables

Temperature
Precipitation
NDVI

Geographic variables

Distance from inland water bodies
Distance from coatline
Elevation

Population variables

Population count
Population density

Avalanche variables

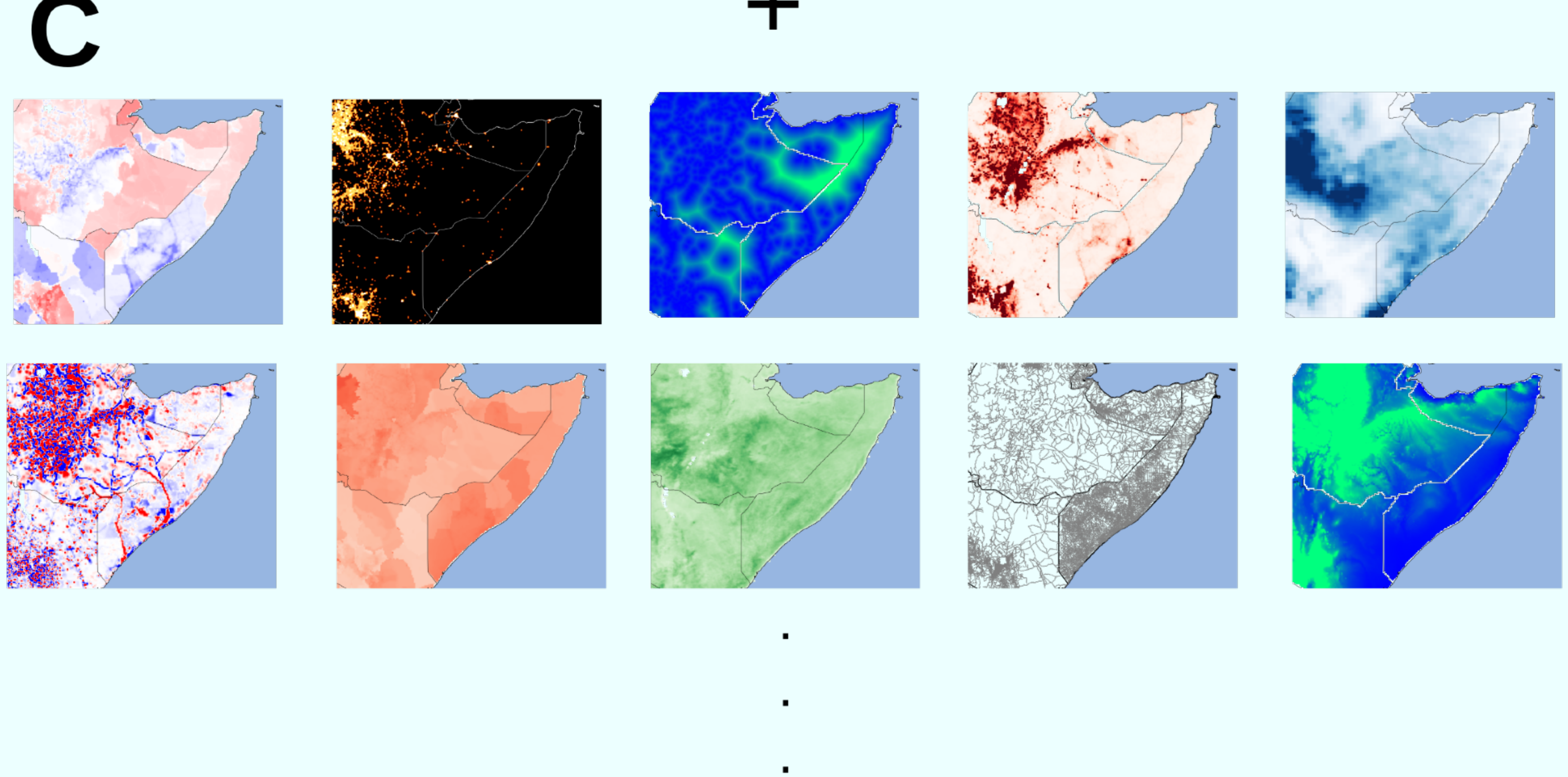
Total number of fatalities
Total number of news reports
Total duration

Economic variables

Demographic variables

Infrastructural variables

Kushwaha and Lee (in progress)



Climatic variables

Temperature
Precipitation
NDVI

Geographic variables

Distance from inland water bodies
Distance from coatline
Elevation

Population variables

Population count
Population density

Avalanche variables

Total number of fatalities
Total number of news reports
Total duration

Economic variables

GDP
GDP per capita
HDI
% increase in GDP
% increase in GDP per capita
% increase in HDI

Demographic variables

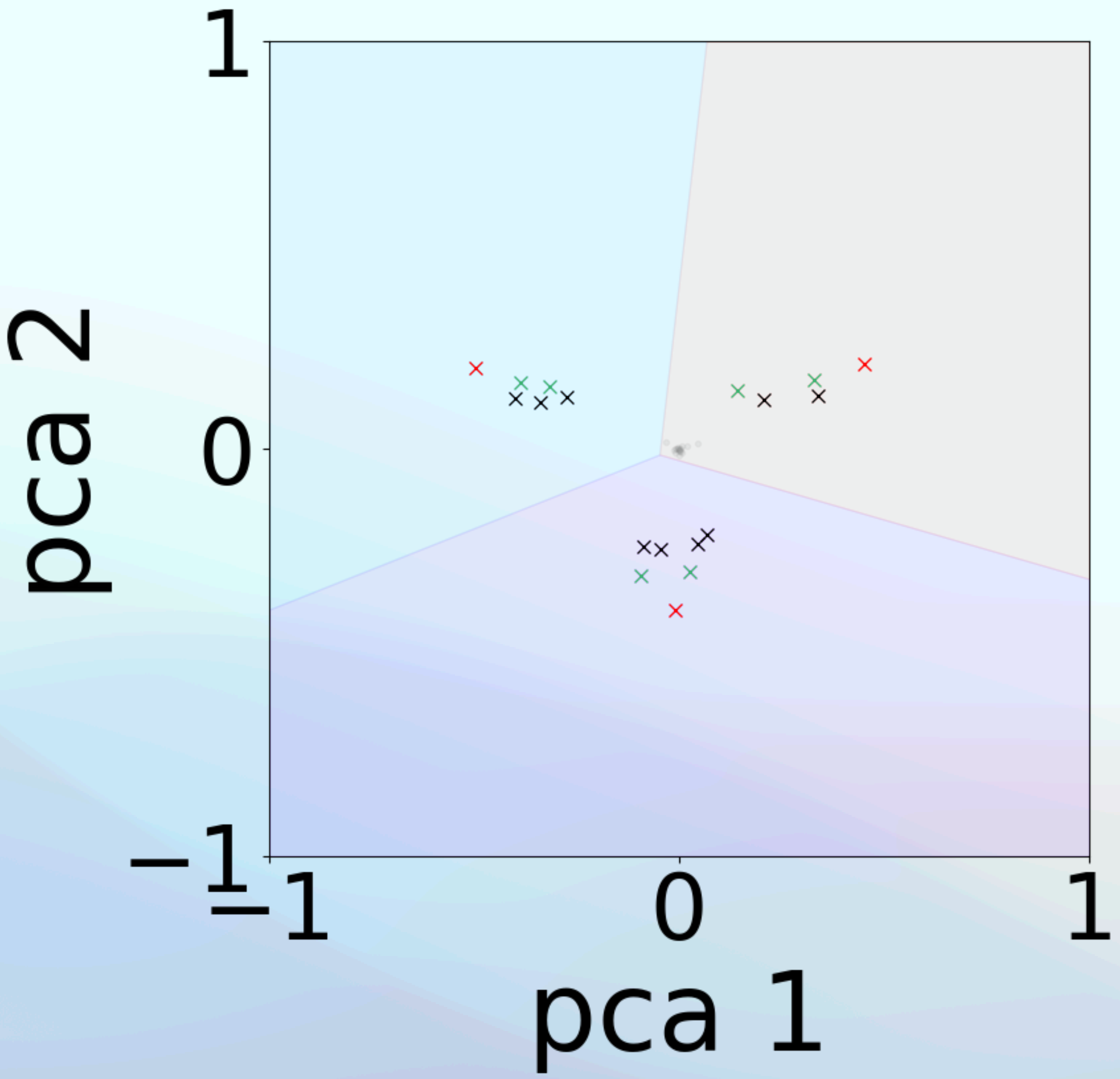
Net migration
Birth rate
Death rate
Interacting ethnic groups

Infrastructural variables

Cellular phone per 100 people
Electric consumption
Shortest distance to roads
Night light

Kushwaha and Lee (in progress)

Triangle of madness



Archetype	Characteristics
Major unrest	High population riparian or coastal plains with good infrastructure
Local conflicts	Average population regions with no other discernible important characteristics
Sporadic and spillover conflicts	Low population with bad infrastructure and economy

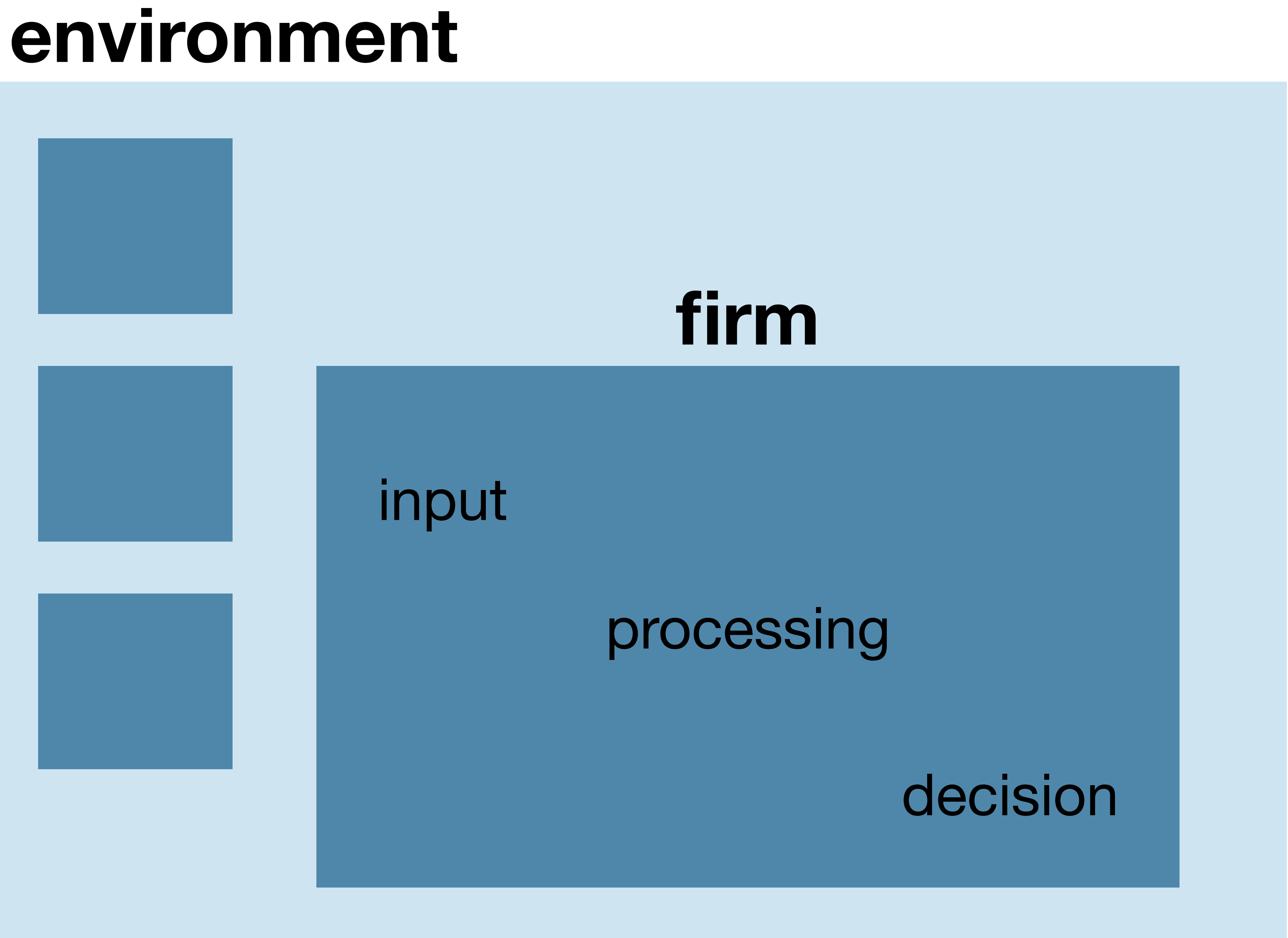
information consumption and firm size

Eddie Lee Complexity Science Hub
Alan Kwan Hong Kong University
Anjali Bhatt Harvard Business School
Rudi Hanel Complexity Science Hub
Frank Neffke Complexity Science Hub

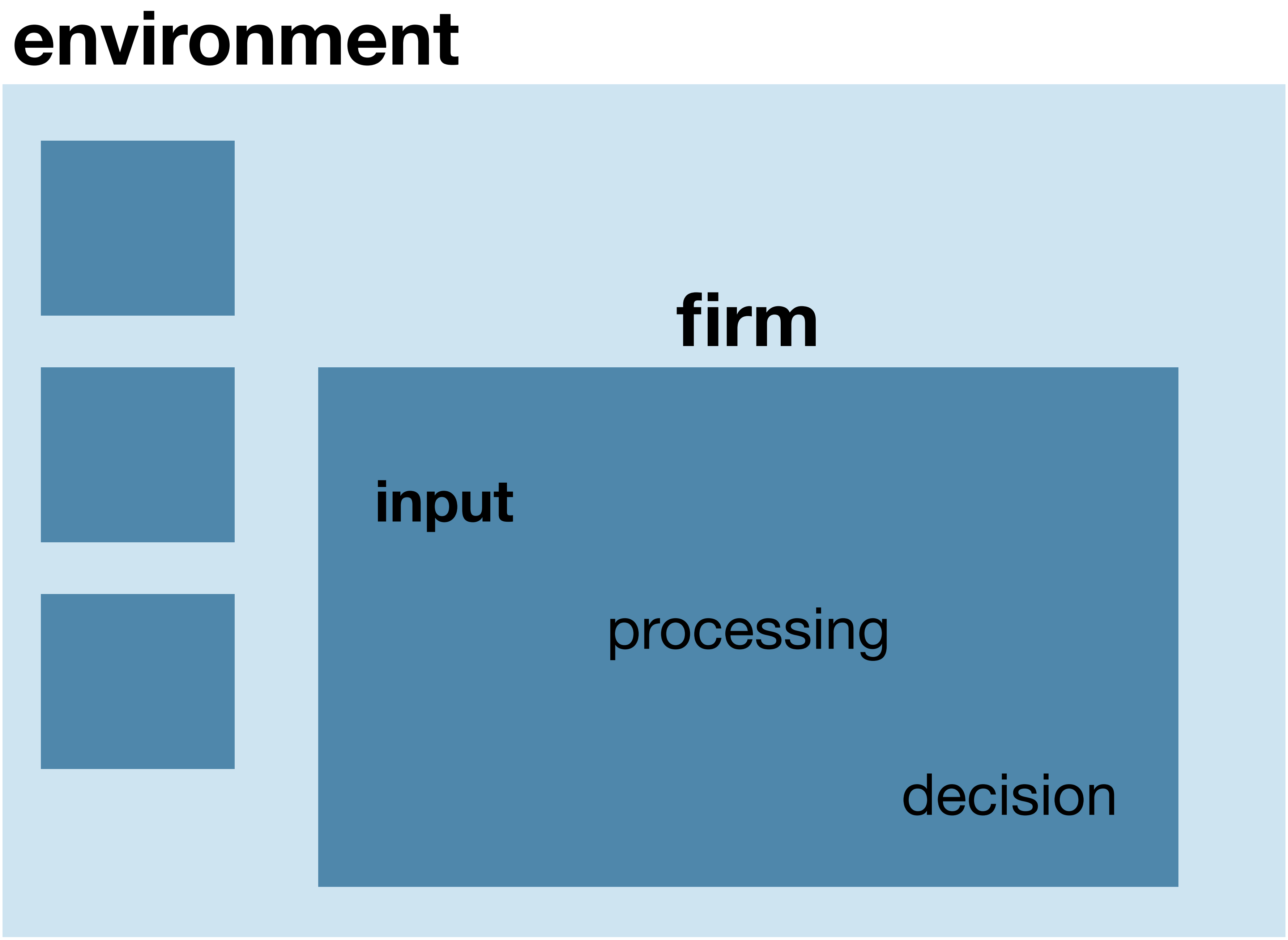
Brenner, Bialek, & de Ruyter van Steveninck, *Neuron* (2000)
Endres & Wingreen, *PNAS* (2008)
Lee, Flack & Krakauer, preprint (2022)
Fleig & Balasubramanian, preprint (2023)

Stentor coeruleus

firms are information entities



employees consume information



data details

- publishers spanning technology, marketing, legal, biotech, manufacturing, and a wide range of business services
- 1 billion records —> 100 million records per day, filtering by to distinguish between sports, adult, and entertainment websites versus a set of articles from well-known business publishers
- two-week period between the dates of June 10 and June 23, 2018
- filters applied toward the content and visitors suggests that information consumption events observed in this study contain the subset of observations most likely to come from work-related visitors and work-related content

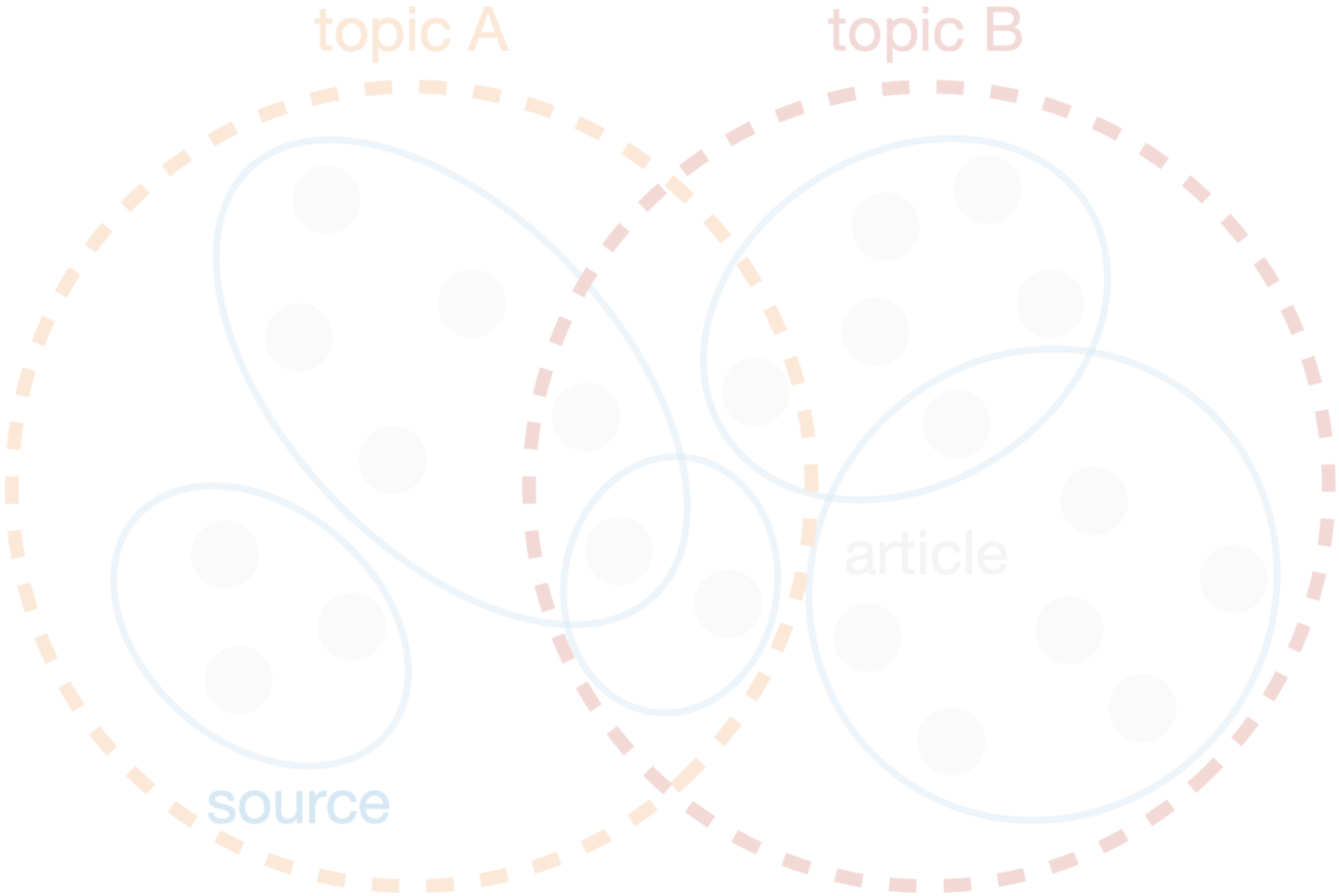
Kwan & Zhu (2019), preprint
Kwan & Zhu (2020), preprint
Kwan, Liu, & Matthies (2022), preprint

Sources like

Bloomberg News
Wall Street Journal
Forbes
Business Insider
CBSi
1105Media
ITCentral Station
Questex
etc.

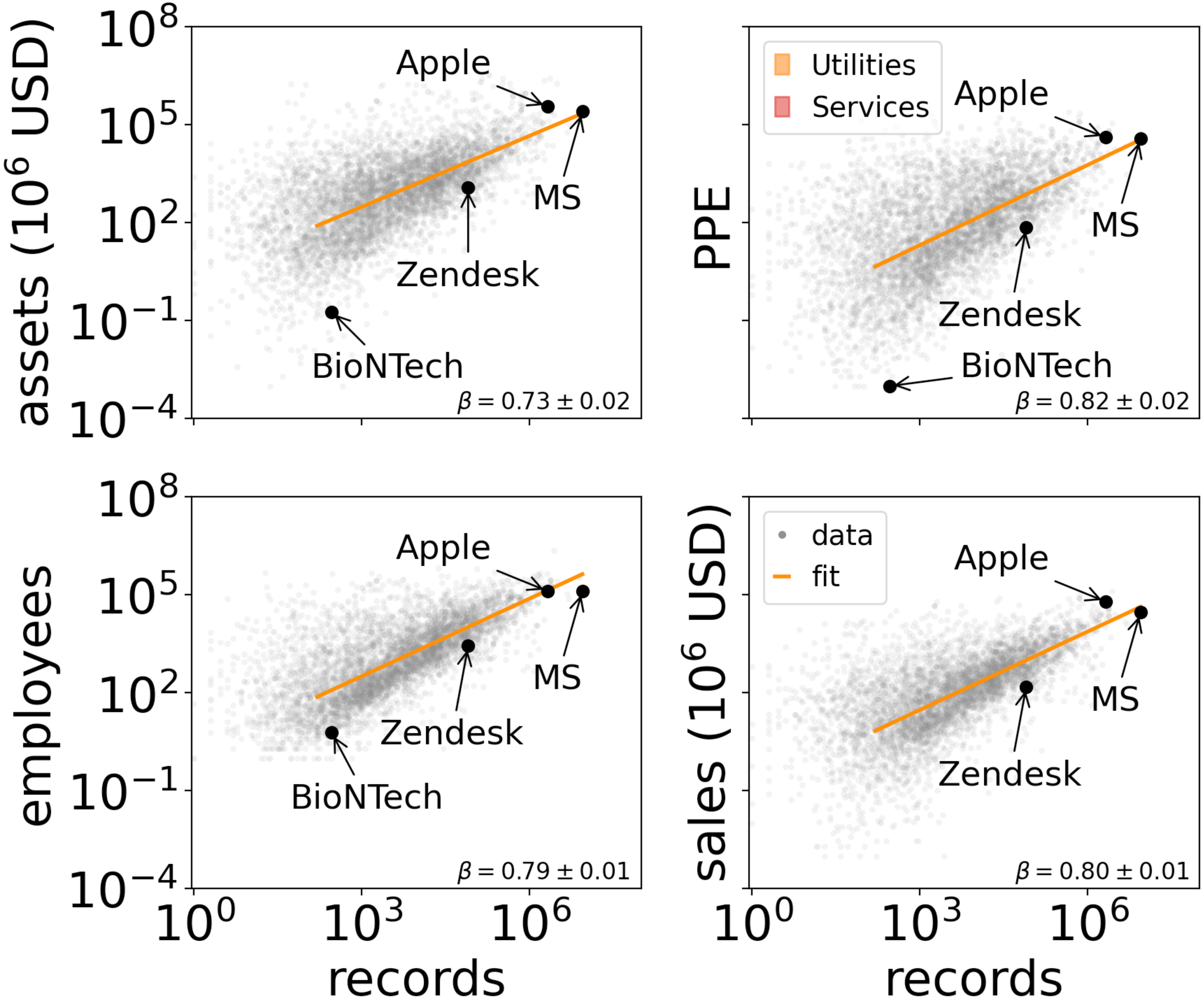
Business-relevant topics like

Social media
Discipline
Vacations
3D animation software
Trade notes
Blu-ray
Globalization
etc.

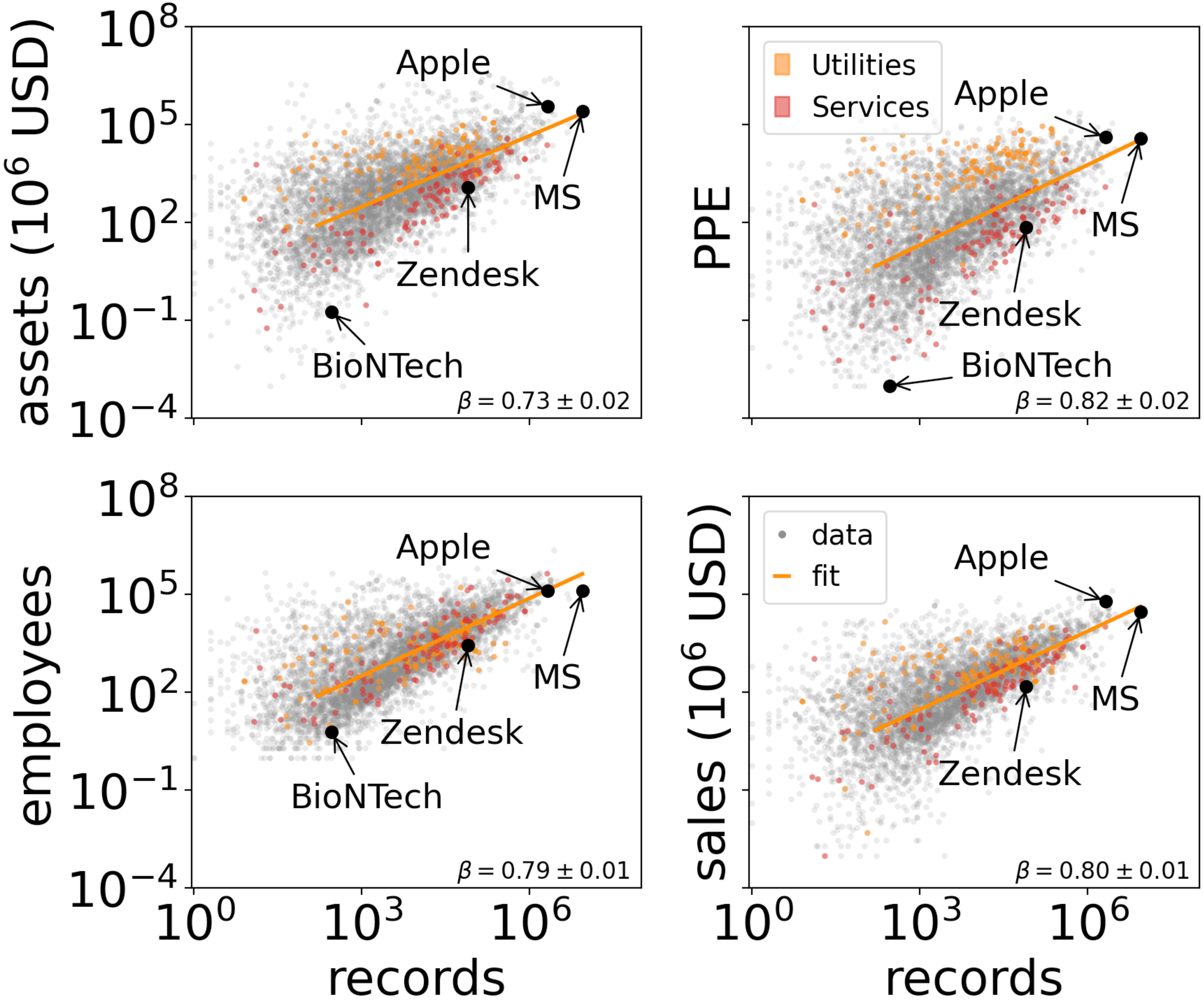


information economy of scale

Stanley et al. (1996), *Nature*
Axtell (2001), *Science*
Bettencourt et al. (2007), *PNAS*
Gabaix (2009), *Annual Review of Economics*
Zhang, Kempes, & West (2022), preprint
etc.

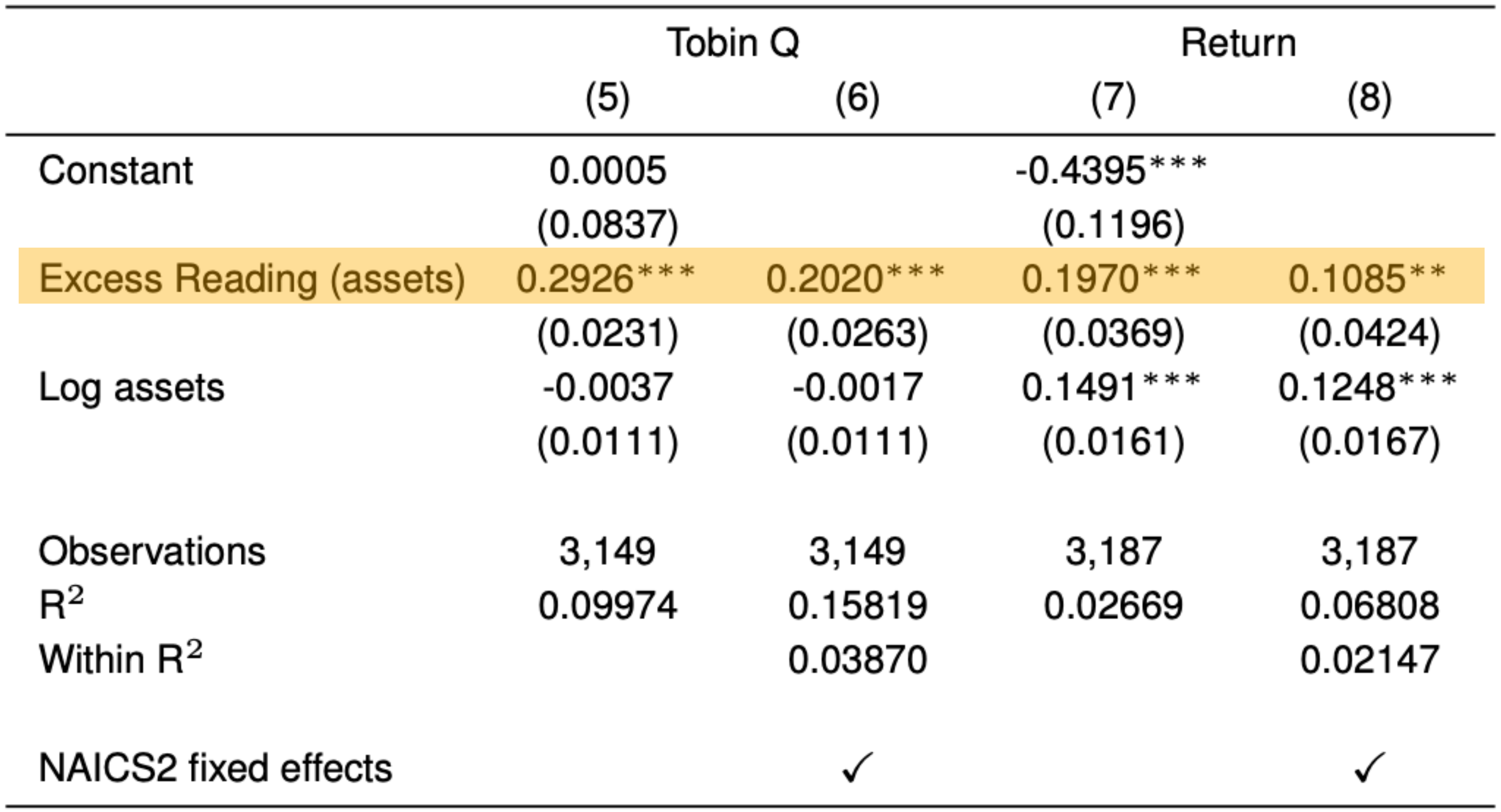


assets (10^6 USD)
10^8
10^5
10^2
10^{-1}
10^{-4}
Apple
MS
Zendesk
BioNTech
$\beta = 0.73 \pm 0.02$
PPE
🟧 Utilities
🟥 Services
Apple
MS
Zendesk
BioNTech
$\beta = 0.82 \pm 0.02$
employees
10^8
10^5
10^2
10^{-1}
10^{-4}
Apple
MS
Zendesk
BioNTech
$\beta = 0.79 \pm 0.01$
10^0
10^3
10^6
records
sales (10^6 USD)
● data
— fit
Apple
MS
Zendesk
$\beta = 0.80 \pm 0.01$
10^0
10^3
10^6
records



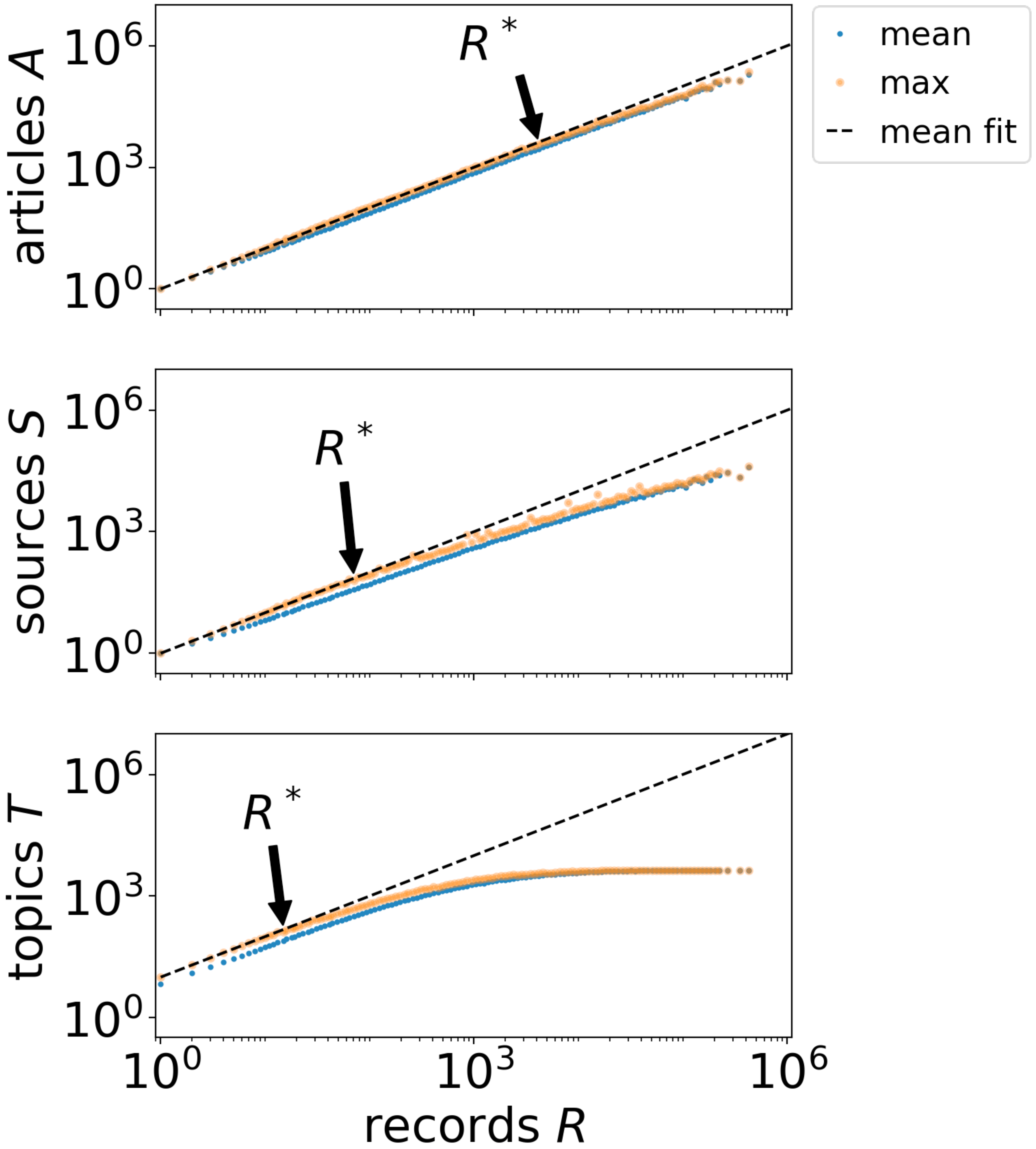
assets (10^6 USD)
10^8
10^5
10^2
10^{-1}
10^{-4}
Apple
MS
Zendesk
BioNTech
$\beta = 0.73 \pm 0.02$
PPE
Utilities
Services
Apple
MS
Zendesk
BioNTech
$\beta = 0.82 \pm 0.02$
employees
10^8
10^5
10^2
10^{-1}
10^{-4}
Apple
MS
Zendesk
BioNTech
$\beta = 0.79 \pm 0.01$
10^0
10^3
10^6
records
sales (10^6 USD)
data
fit
Apple
MS
Zendesk
$\beta = 0.80 \pm 0.01$
10^0
10^3
10^6
records

Deviations from baseline strongly correlated with future valuation and contemporaneous returns

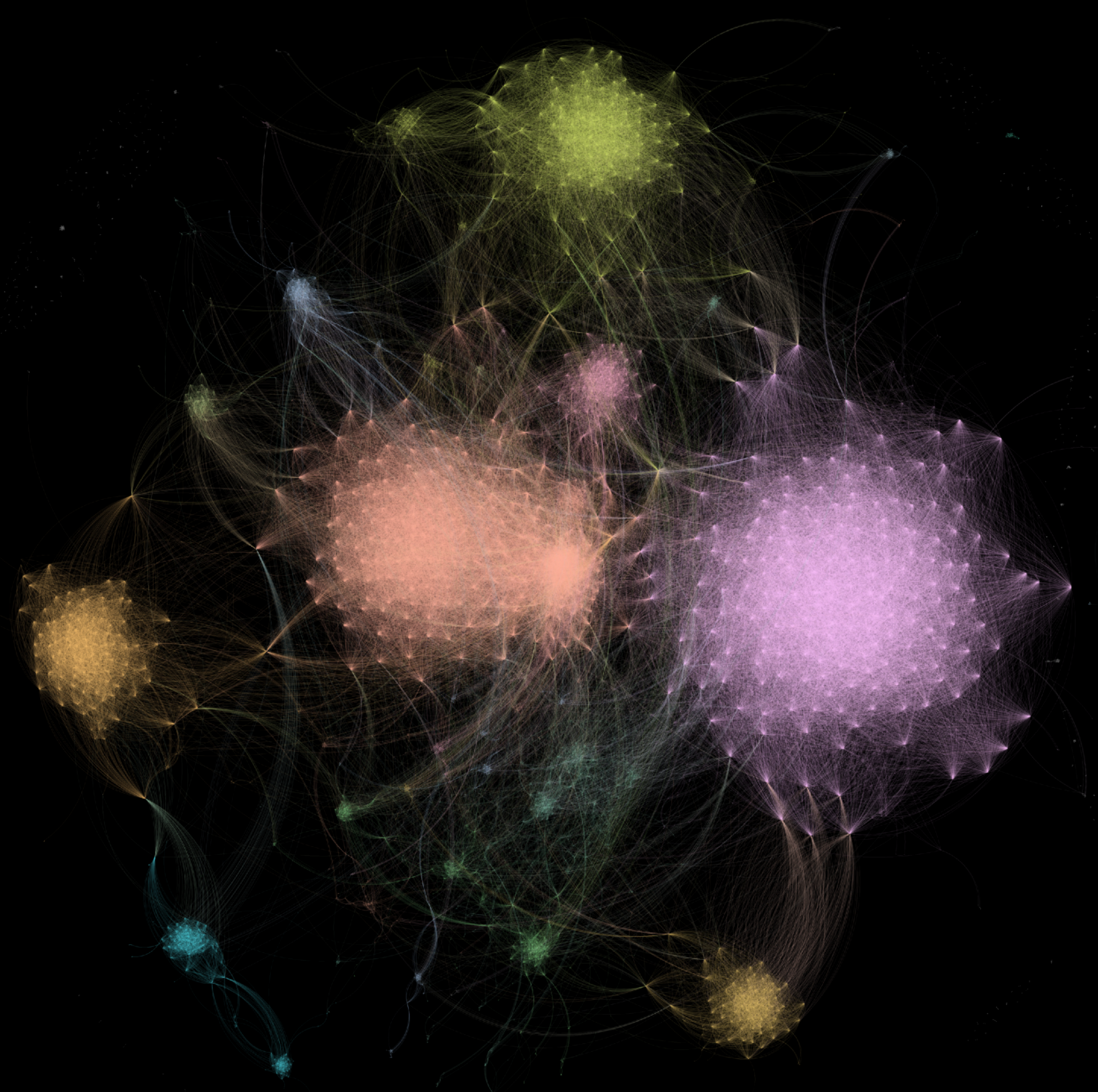


	Tobin Q		Return	
	(5)	(6)	(7)	(8)
Constant	0.0005		-0.4395***	
	(0.0837)		(0.1196)	
Excess Reading (assets)	0.2926***	0.2020***	0.1970***	0.1085**
	(0.0231)	(0.0263)	(0.0369)	(0.0424)
Log assets	-0.0037	-0.0017	0.1491***	0.1248***
	(0.0111)	(0.0111)	(0.0161)	(0.0167)
Observations	3,149	3,149	3,187	3,187
R^2	0.09974	0.15819	0.02669	0.06808
Within R^2		0.03870		0.02147
NAICS2 fixed effects		✓		✓

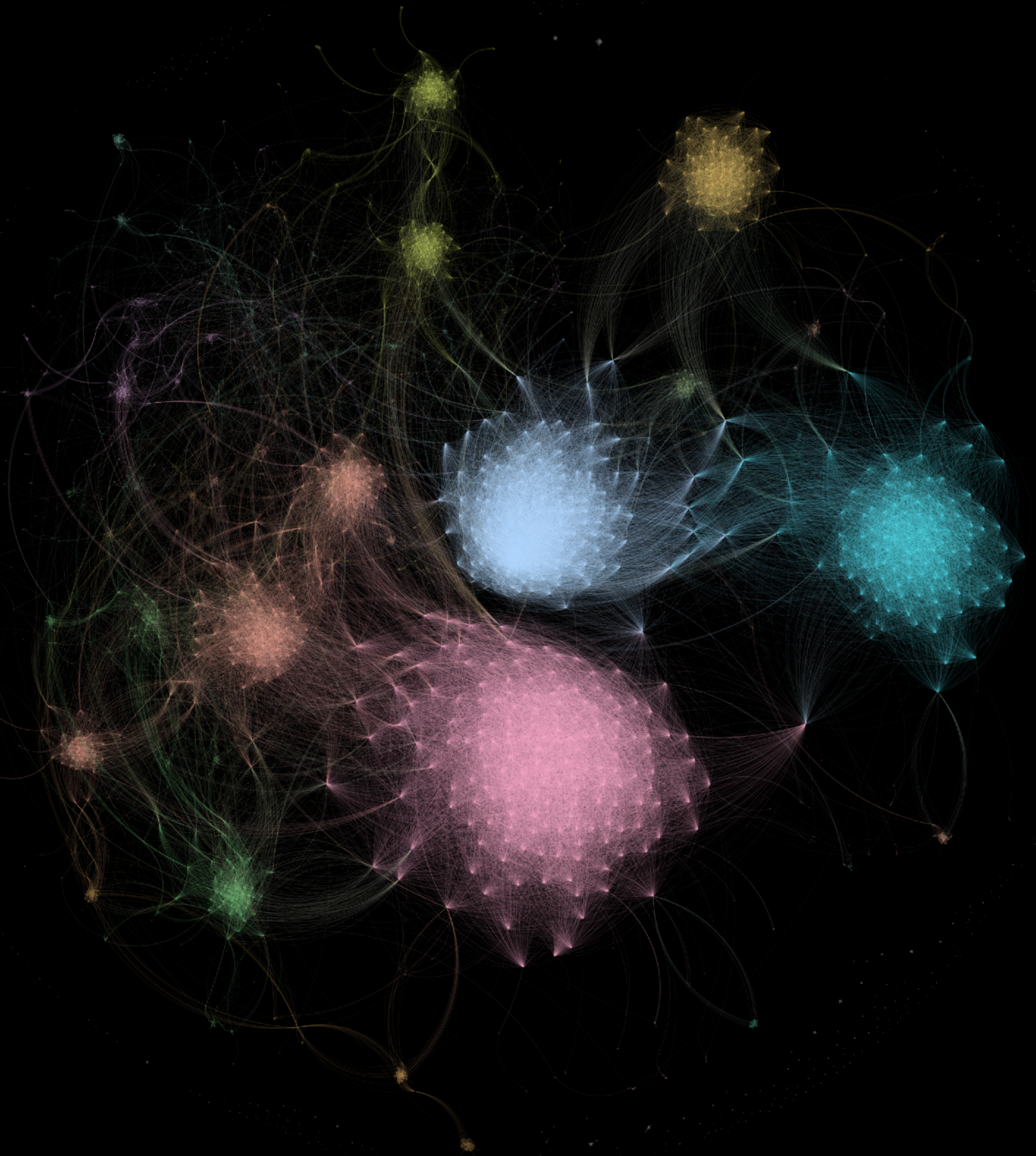
limited diversity of reading



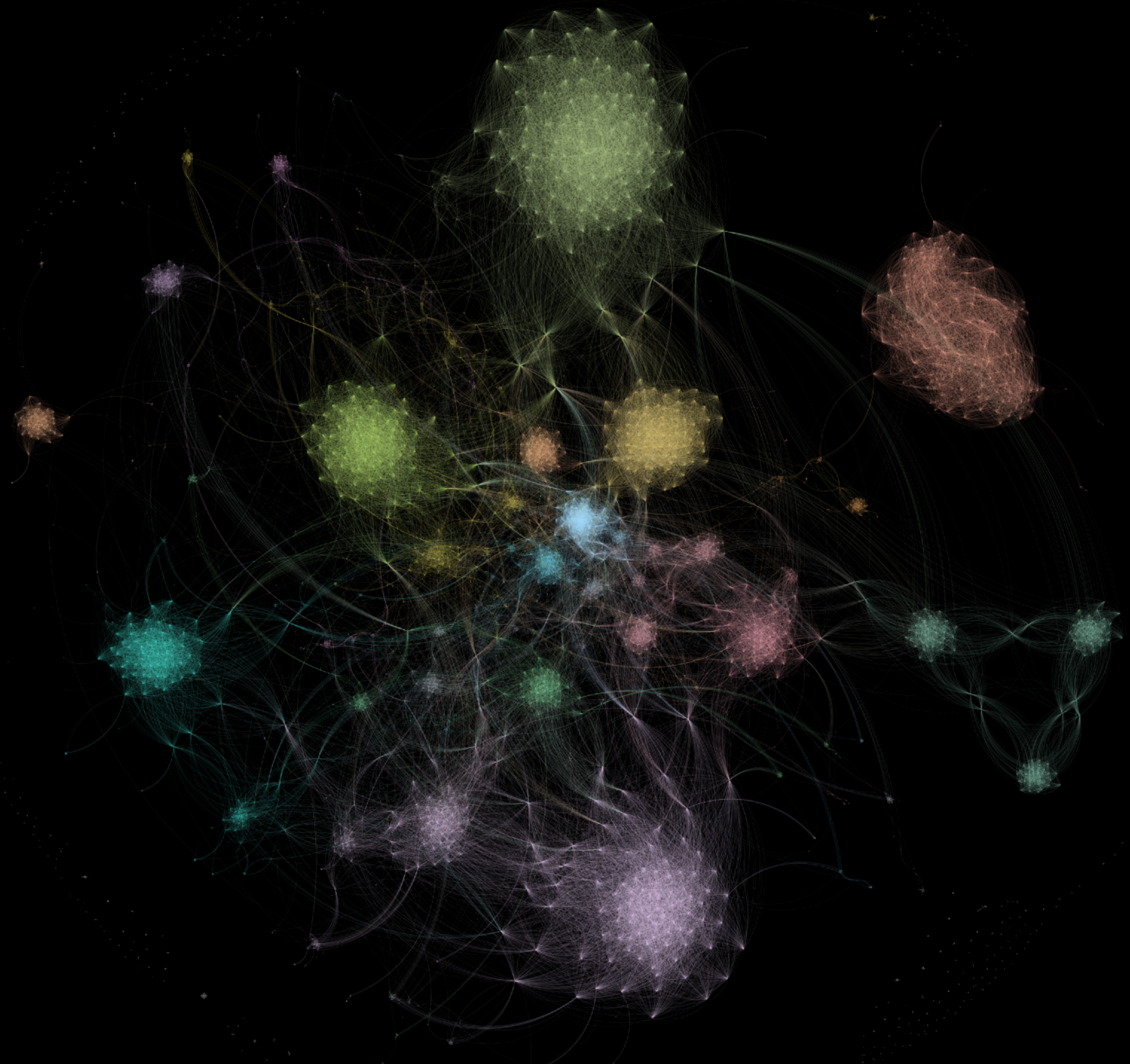
insurance

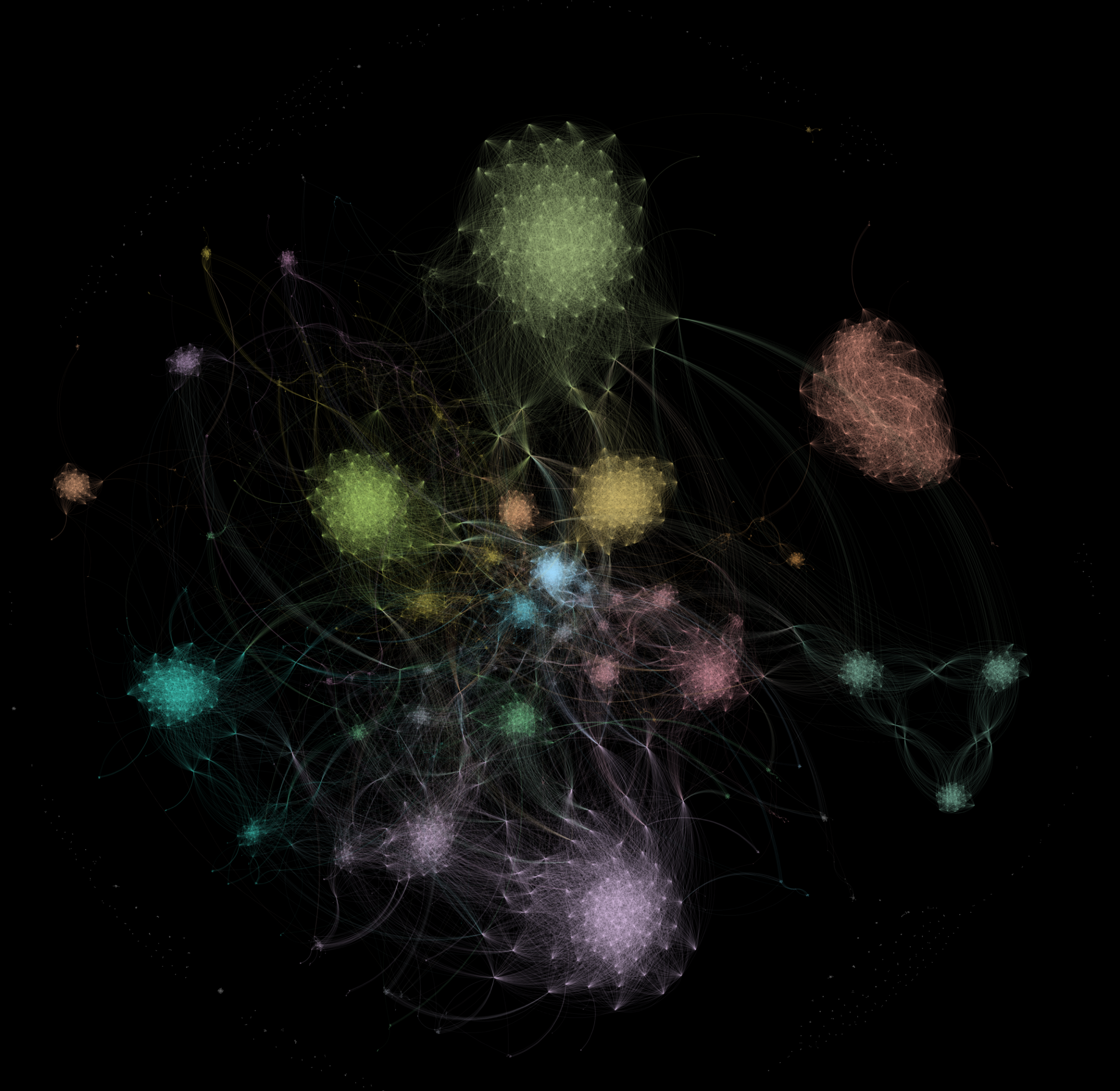


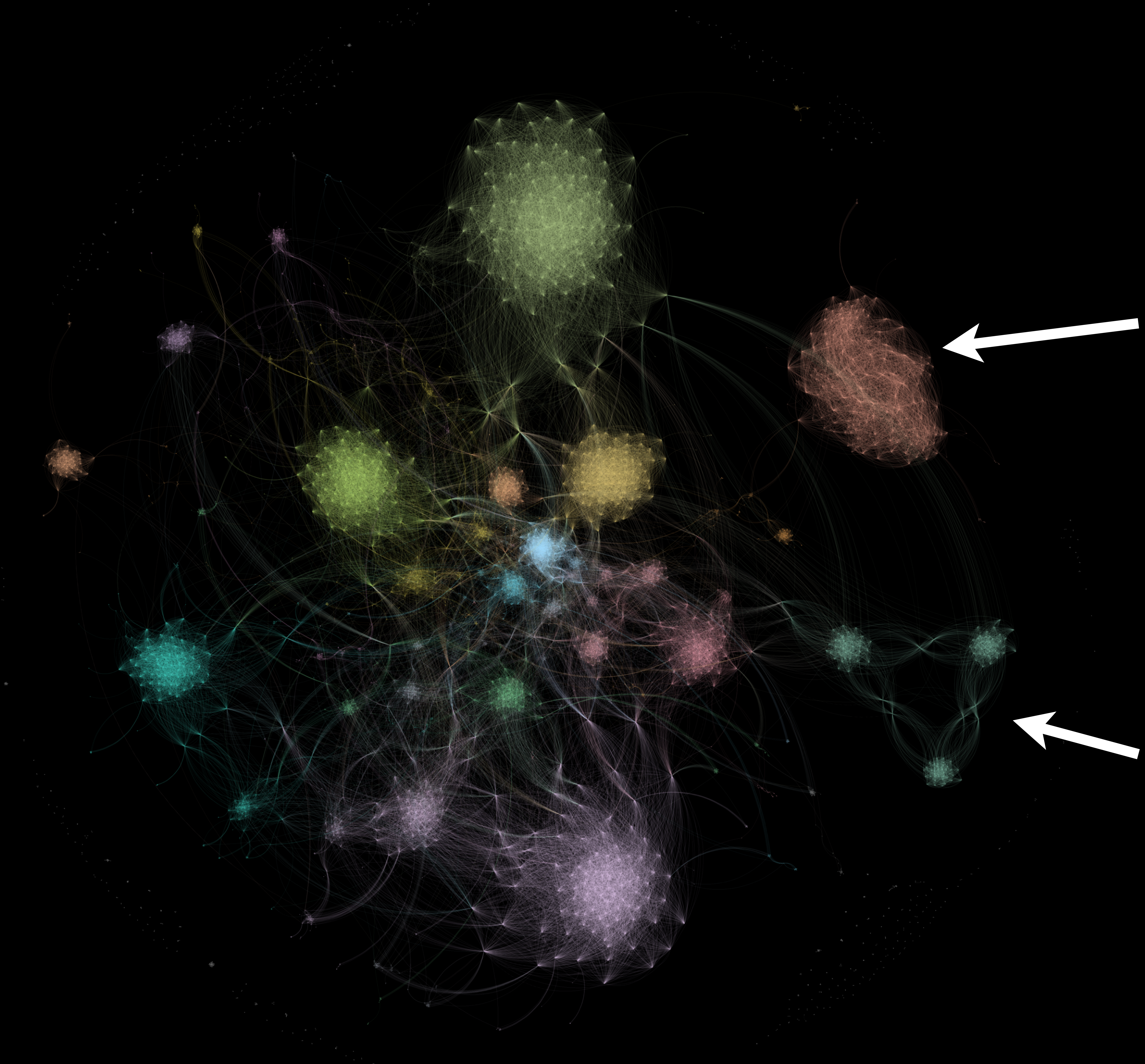
clothing



automobile



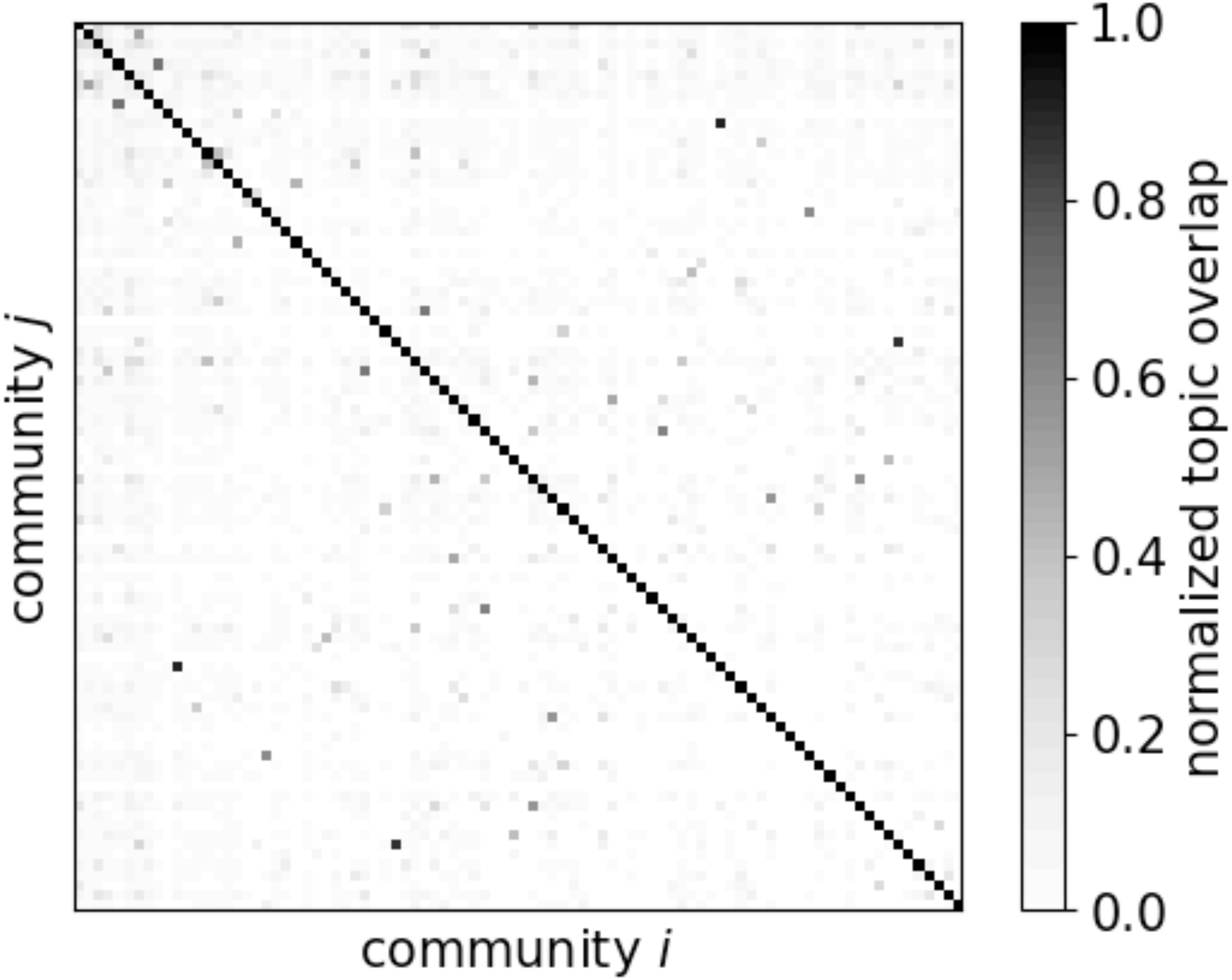




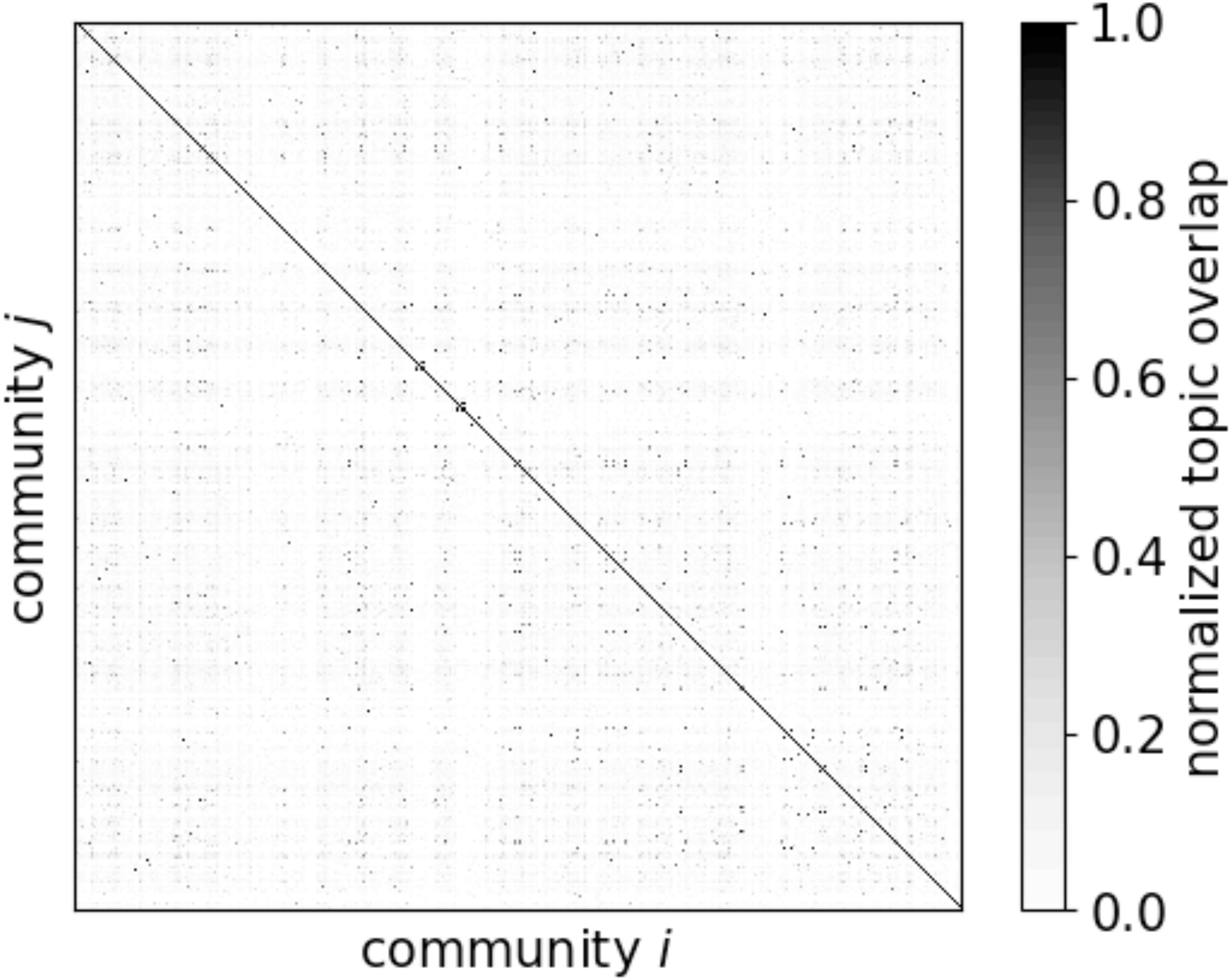
community A with
normalized topic vector
$\vec{t}_A$

community B with
normalized topic vector
$\vec{t}_B$

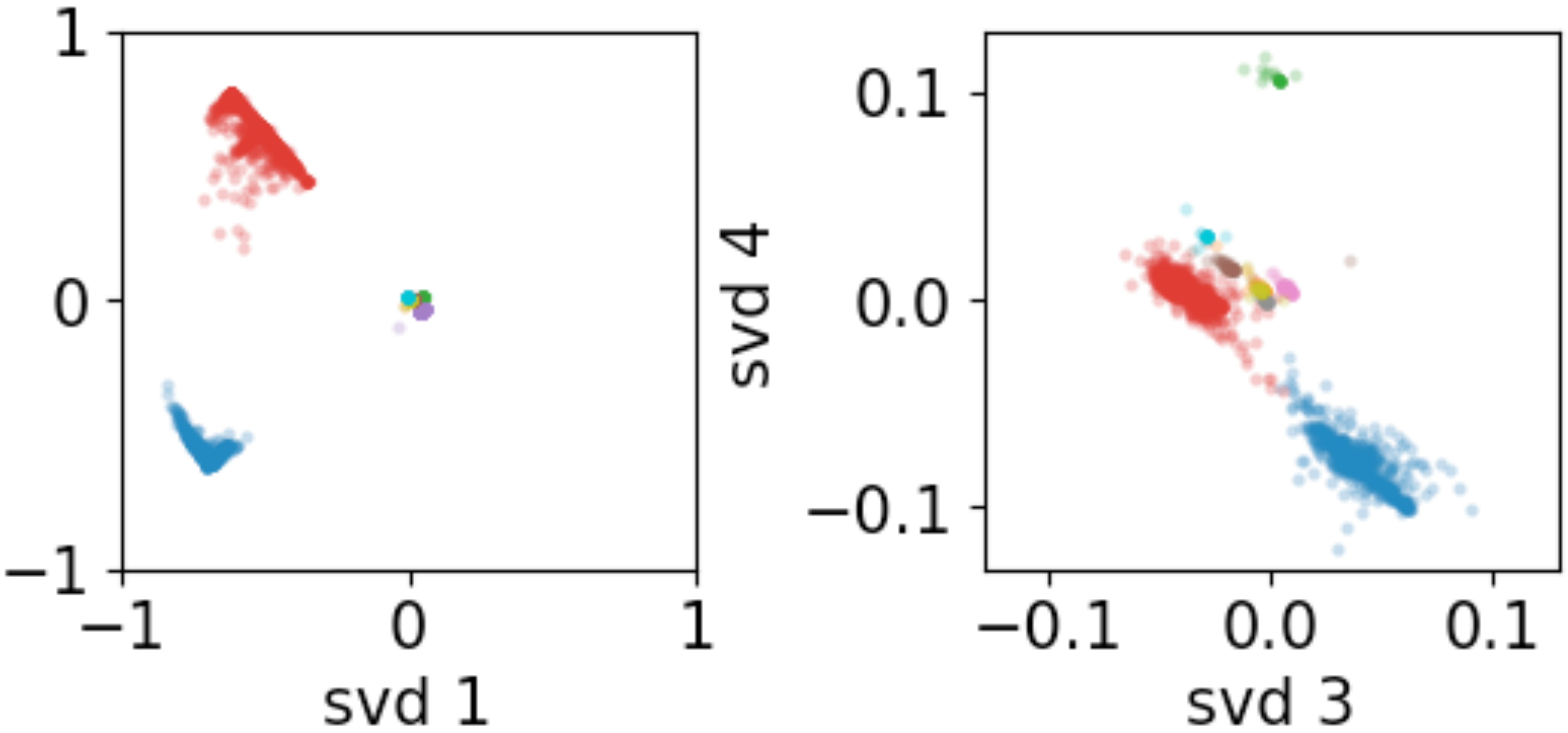
Similarity matrix of “departments” within an auto manufacturing firm



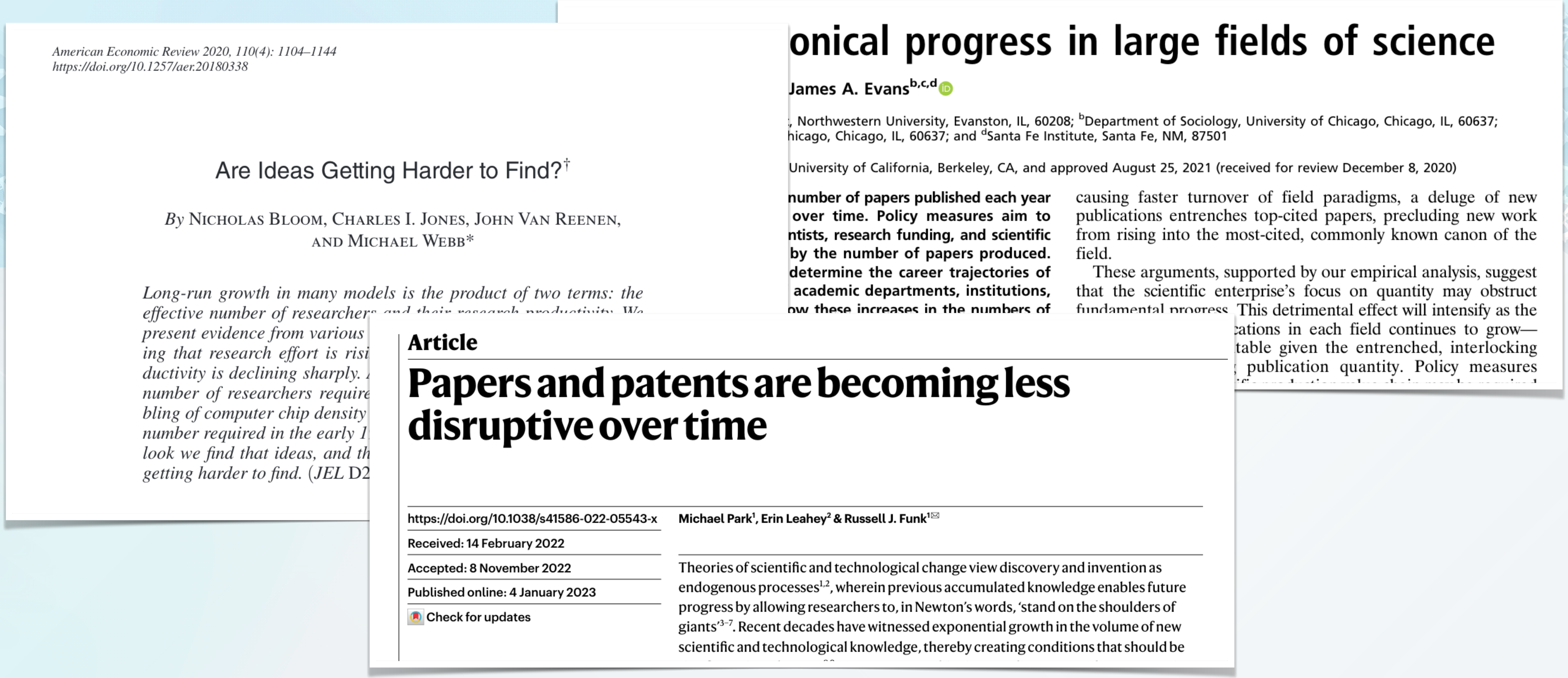
Similarity matrix of "departments" across many firms



Types of “departments”



Is science becoming less innovative?



American Economic Review 2020, 110(4): 1104–1144
https://doi.org/10.1257/aer.20180338

Are Ideas Getting Harder to Find?†

By Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb*

Long-run growth in many models is the product of two terms: the effective number of researchers and their research productivity. We present evidence from various ing that research effort is risi ductivity is declining sharply. number of researchers require bling of computer chip density number required in the early 1 look we find that ideas, and th getting harder to find. (JEL D2

onical progress in large fields of science

James A. Evans[b,c,d]

, Northwestern University, Evanston, IL, 60208; [b]Department of Sociology, University of Chicago, Chicago, IL, 60637; hicago, Chicago, IL, 60637; and [d]Santa Fe Institute, Santa Fe, NM, 87501

University of California, Berkeley, CA, and approved August 25, 2021 (received for review December 8, 2020)

number of papers published each year over time. Policy measures aim to ntists, research funding, and scientific by the number of papers produced. determine the career trajectories of academic departments, institutions, ow these increases in the numbers of

causing faster turnover of field paradigms, a deluge of new publications entrenches top-cited papers, precluding new work from rising into the most-cited, commonly known canon of the field.

These arguments, supported by our empirical analysis, suggest that the scientific enterprise's focus on quantity may obstruct fundamental progress. This detrimental effect will intensify as the cations in each field continues to grow— table given the entrenched, interlocking publication quantity. Policy measures

Article

Papers and patents are becoming less disruptive over time

https://doi.org/10.1038/s41586-022-05543-x

Received: 14 February 2022

Accepted: 8 November 2022

Published online: 4 January 2023

Check for updates

Michael Park[1], Erin Leahey[2] & Russell J. Funk[1]

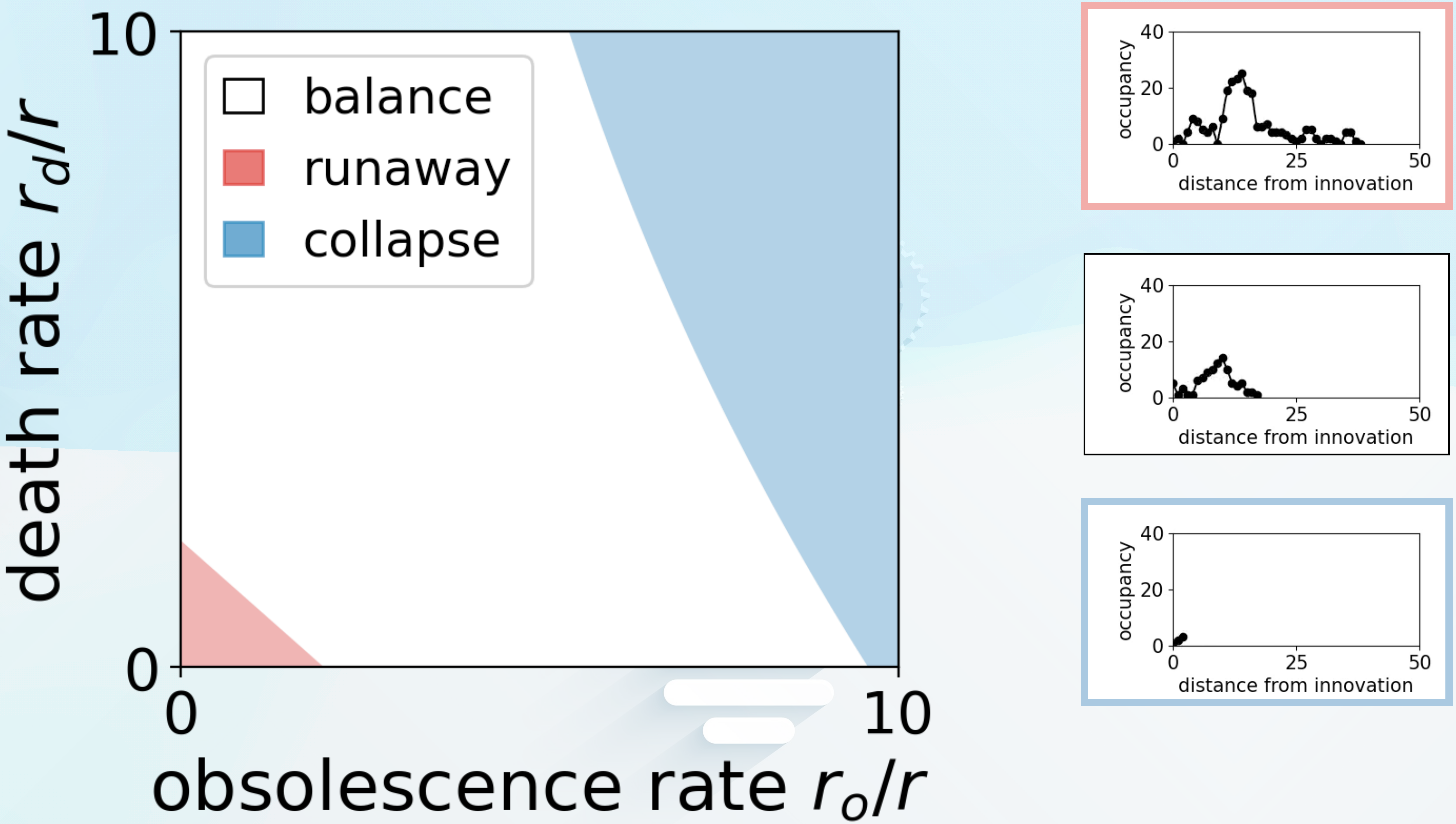
Theories of scientific and technological change view discovery and invention as endogenous processes[1,2], wherein previous accumulated knowledge enables future progress by allowing researchers to, in Newton's words, 'stand on the shoulders of giants'[3–7]. Recent decades have witnessed exponential growth in the volume of new scientific and technological knowledge, thereby creating conditions that should be

"idea"
obsolescence
innovation
relatedness

Iwai (1984)
Kauffman (2000)
Silverberg & Verspagen (2005)
Valverde & Solé (2007)
Hidalgo et al. (2007)
Sharma et al. (2023)

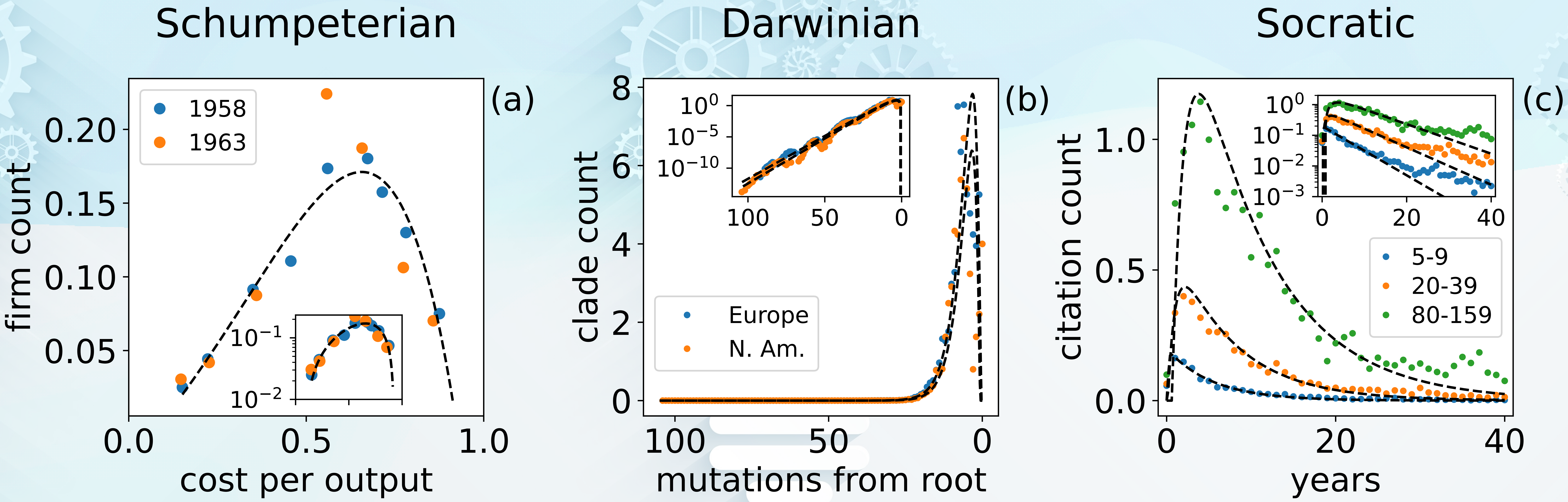
agent i
agent k
agent m
agent j
obsolescence
innovation

Lee, Kempes, West, *PNAS* (2024)



Lee, Kempes, West, *PNAS* (2024)

Predicting frequency of innovation agents across systems



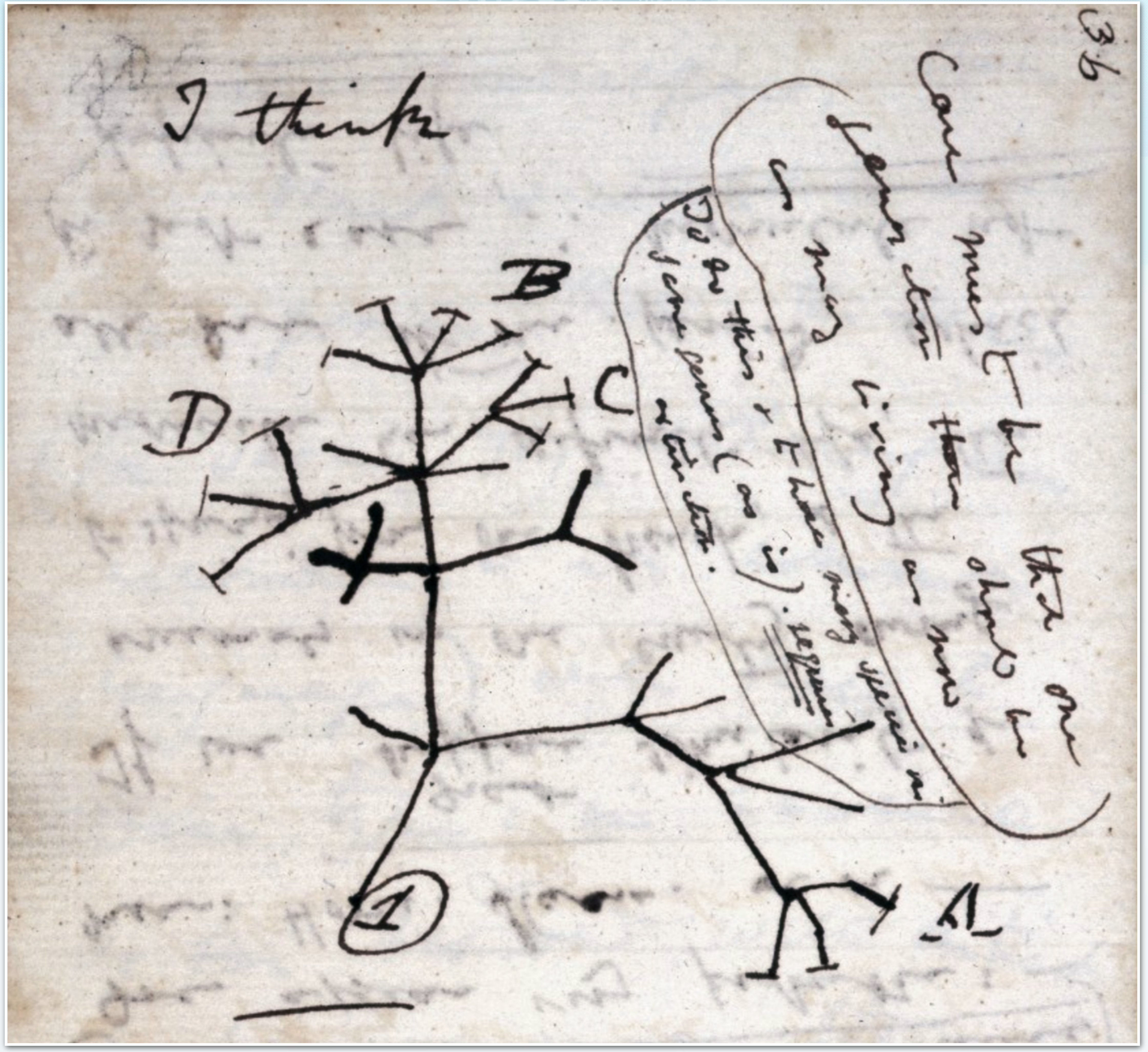
Schumpeterian

Darwinian

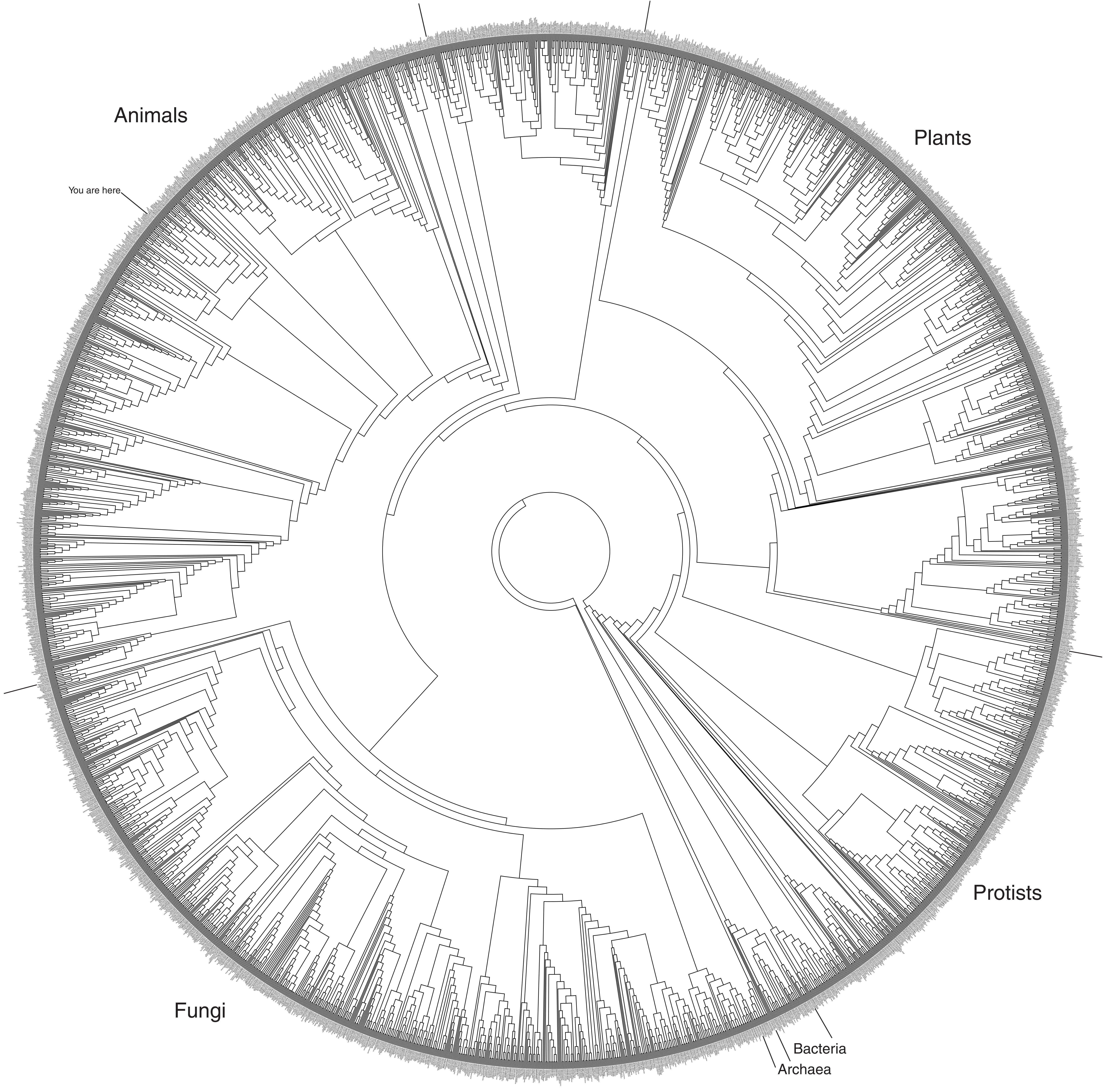
Socratic

Lee, Kempes, West, *PNAS* (2024)

Space of the possible

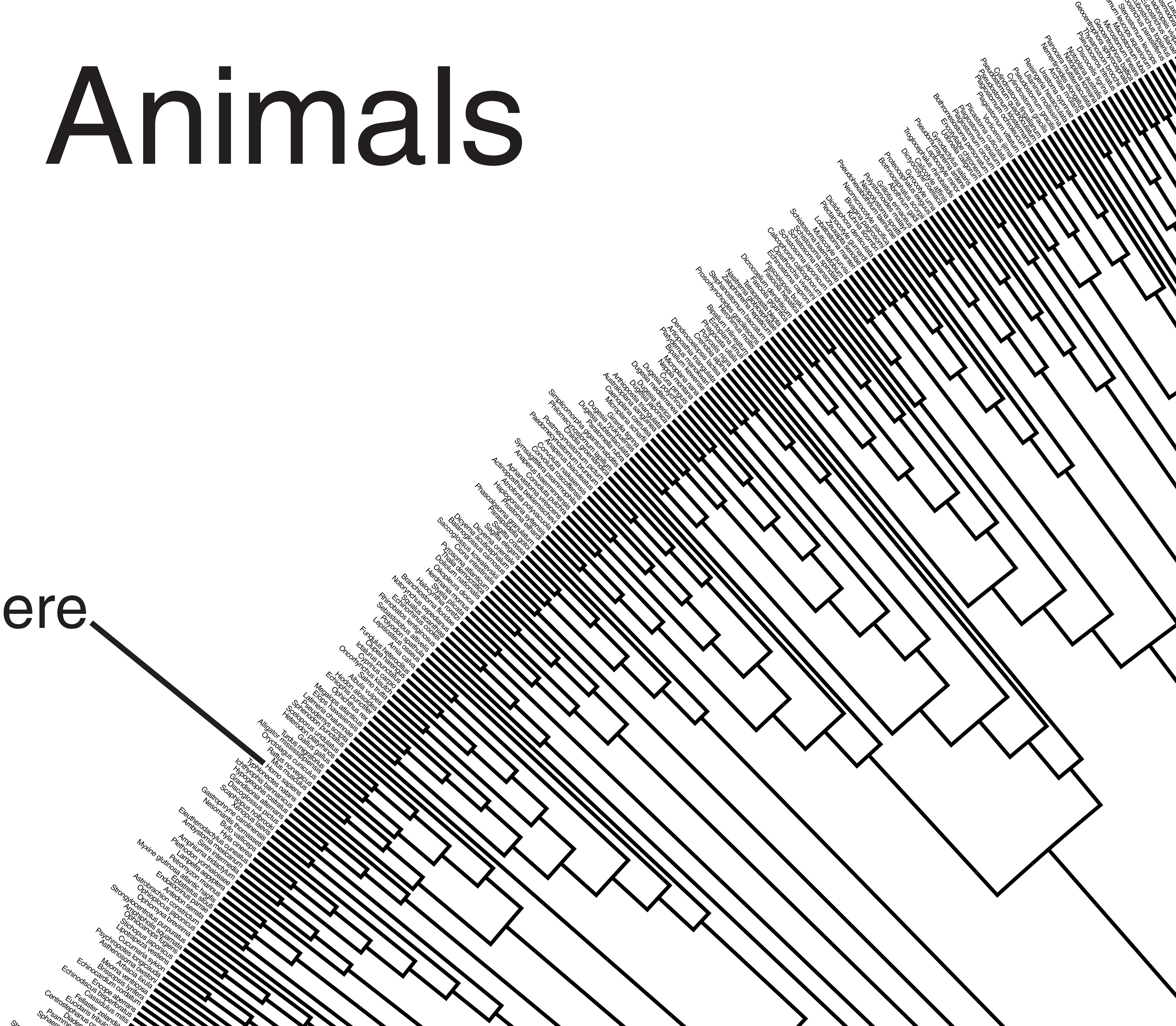

36
I think
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is). requires extinction.
B
C
D
1
A

Ernesto Ortega



Hillis Lab, UTA

Animals



You are here

Hillis Lab, UTA

Ceremonial Burial
Temple 40.1

Horseback Riding
Cavalry 2.1.2.20

Pottery
Granary 60.1
Hanging Gardens 300

Alphabet

The Wheel
Chariot 4.1.2.40

Masonry
Palace 200
City Walls 120.2
Great Wall 300
Pyramids 300

Bronze Working
Phalanx 1.2.1.20
Colossus 200

Mysticism
Oracle 300
Doubles effect of Temples

Code of Laws
Courthouse 80.1

Writing
Diplomat 0.0.2.30
Library 80.1

MapMaking
Trireme 1.0.3.40 (2)
Lighthouse 200

Mathematics
Catapult 6.1.1.40

Currency
Marketplace 80.1

Iron Working
Legion 3.1.1.20

Monarchy
Monarchy govt.

Literacy
Great Library 300

Astronomy (Mysticism)
Copernicus' Observatory 300

Construction
Aqueduct 120.2
Colosseum 100.4
Fortresses

Trade (Code of Laws)
Caravan 0.1.1.50

Feudalism (Masonry)

The Republic
Republic govt.

Philosophy (Mysticism)

Navigation
Sail 1.1.3.40 (3)
Magellan's Expedition 400

Bridge Building (Iron Working)
Roads on rivers

Engineering (The Wheel)

Chivalry (Horseback Riding)
Knights 4.2.2.40

Democracy
Democracy govt.

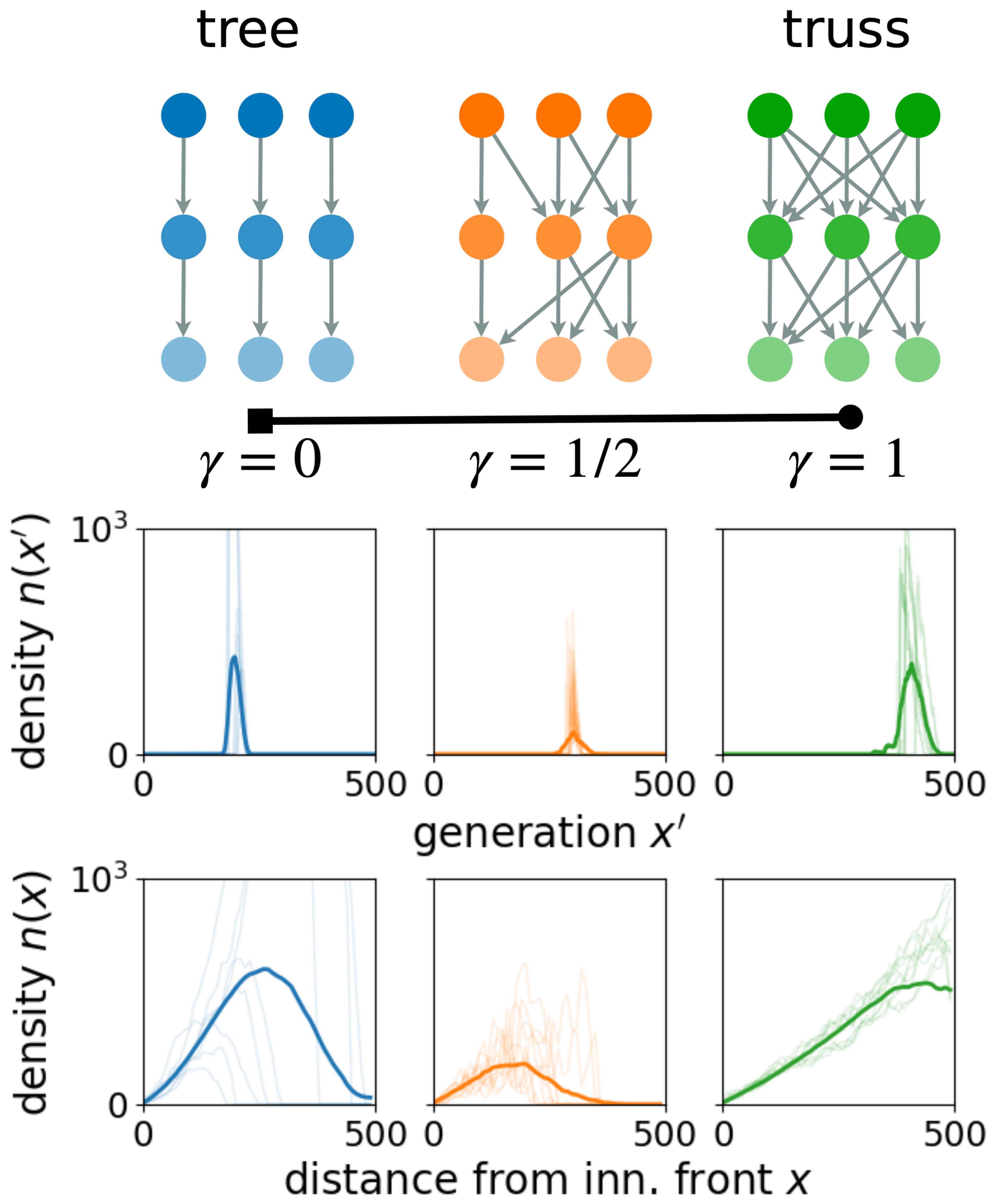
Medicine (Trade)
Shakespeare's Theatre 400
Prevents Plague

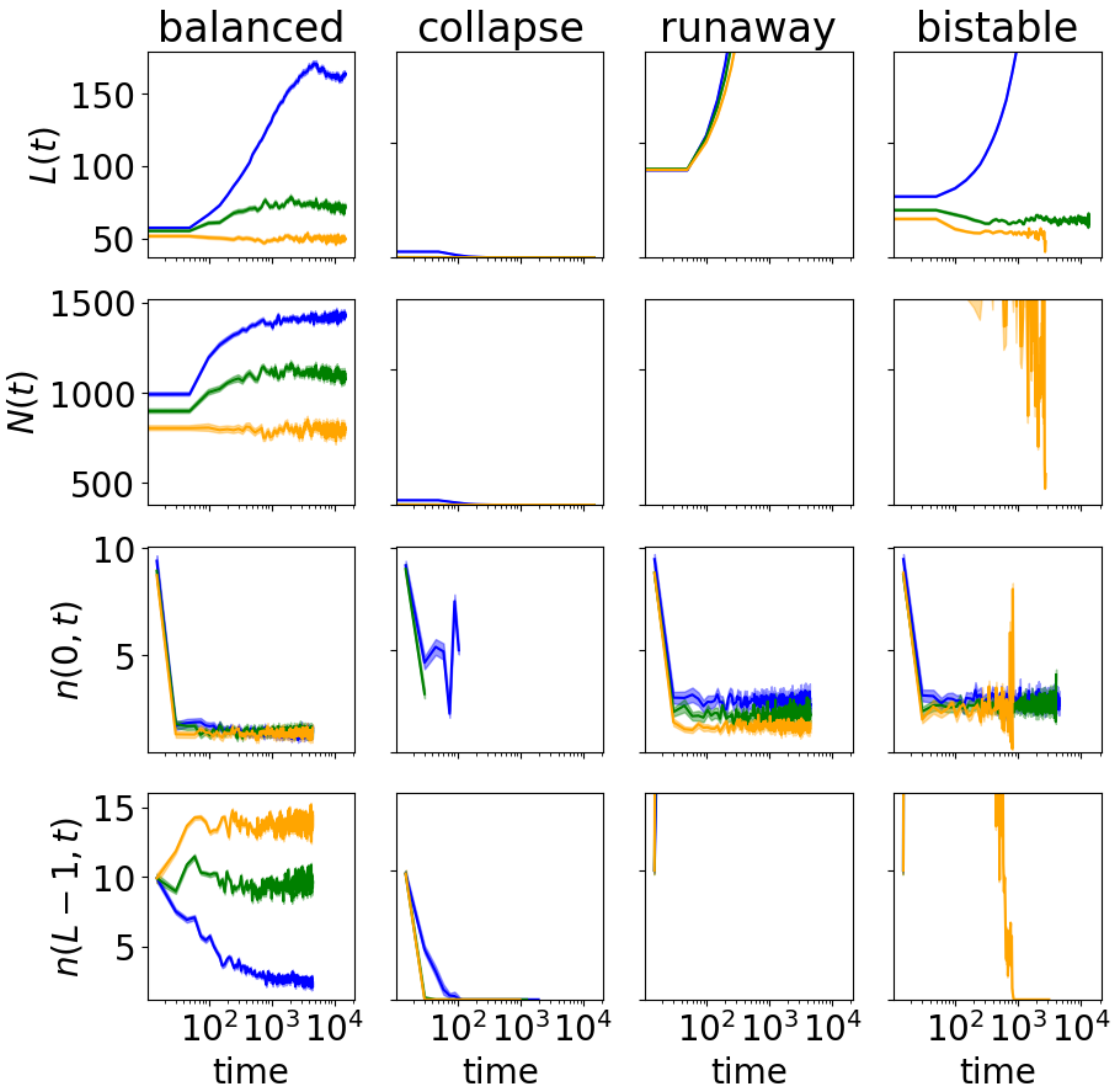
University (Mathematics)
University 160.3

Physics (Mathematics)

Invention (Literacy)

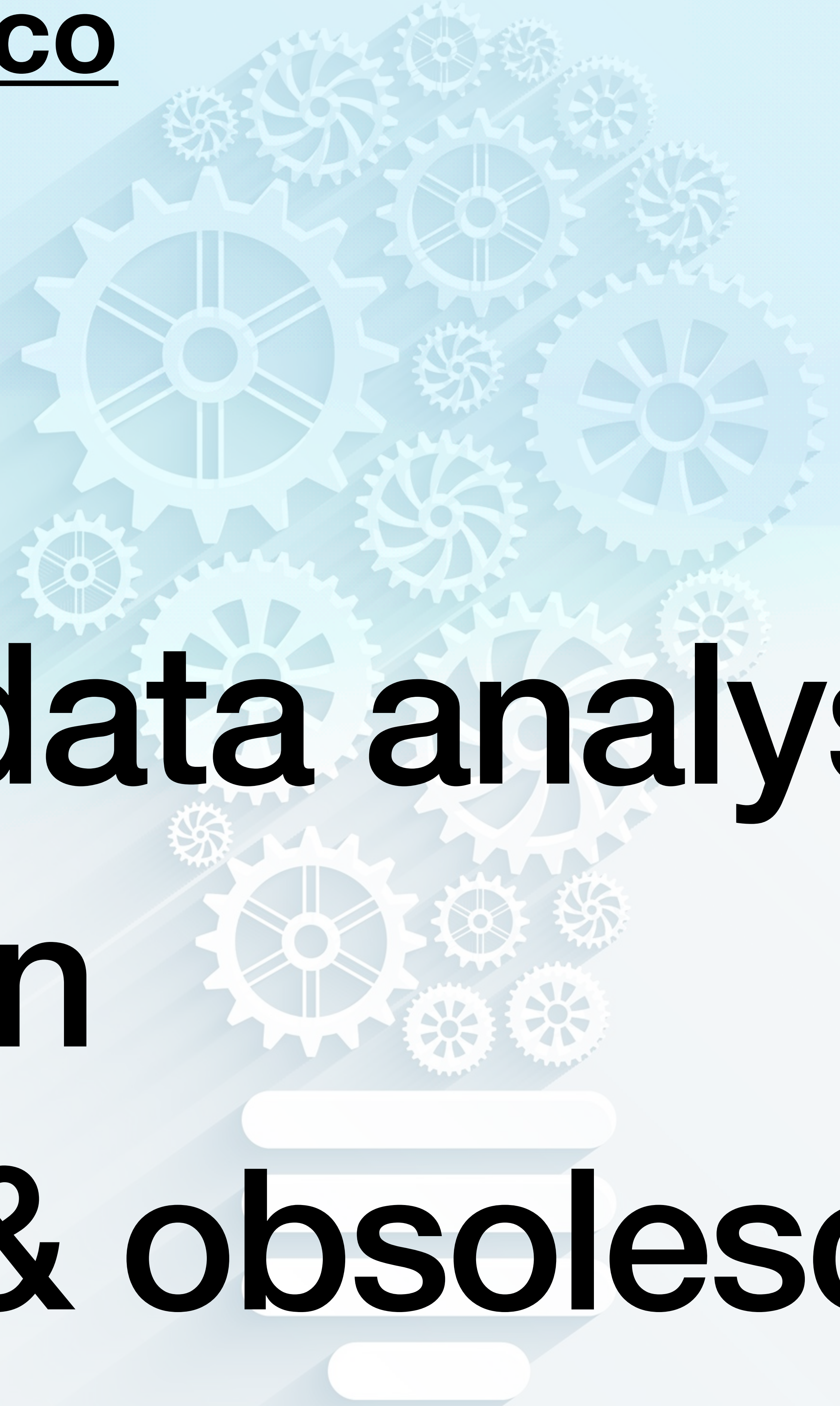
Numerics and analytics





Questions?

More at https://eddielee.co



1. inference
2. multiscale data analysis
3. computation
4. innovation & obsolescence

Gavin Rees

Alan Kwan

Frank Neffke

Rudi Hanel

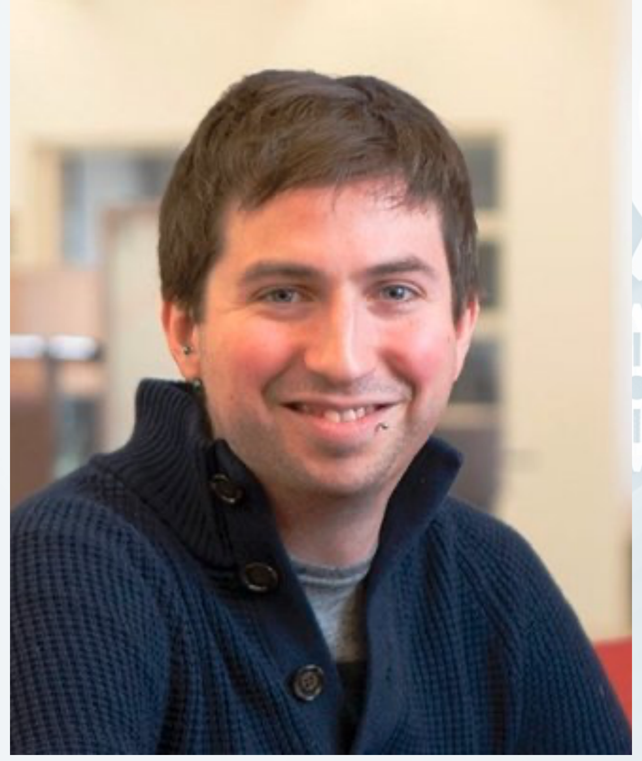
Bryan C. Daniels

David C. Krakauer

Daniel M. Katz

Anjali Bhatt

Ernesto Ortega



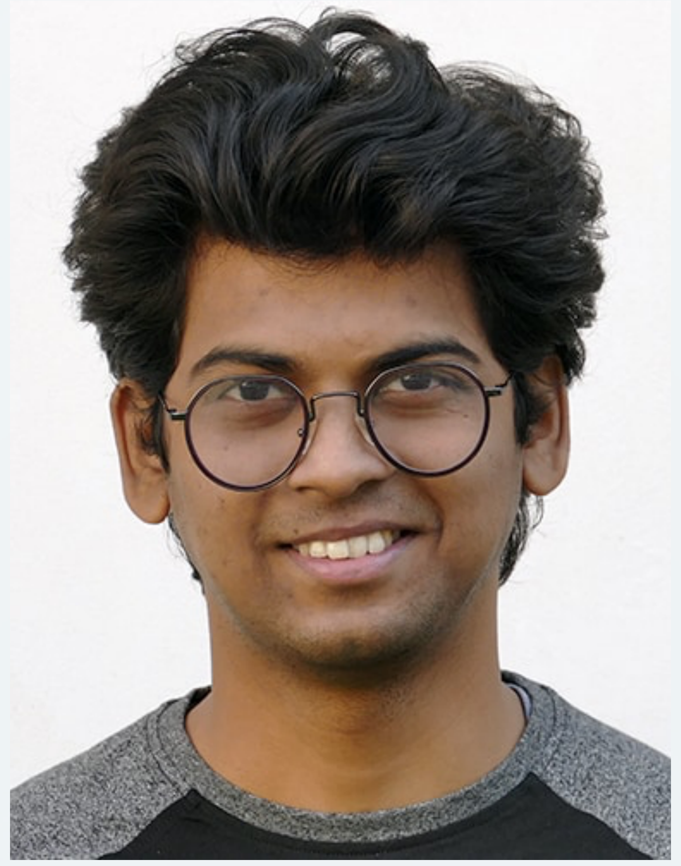
George Cantwell

Geoffrey B. West

Chase P. Broedersz

Chris R. Myers

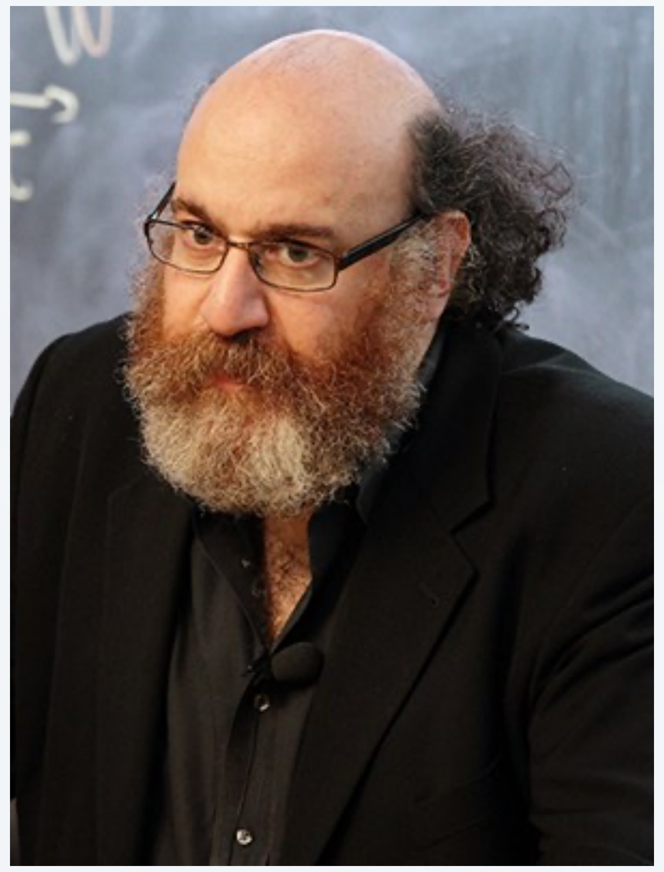
Michael J. Bommarito



Niraj Kushwaha

Woi Oh Sok

Chris P. Kempes



William Bialek

Jessica C. Flack

Paul H. Ginsparg