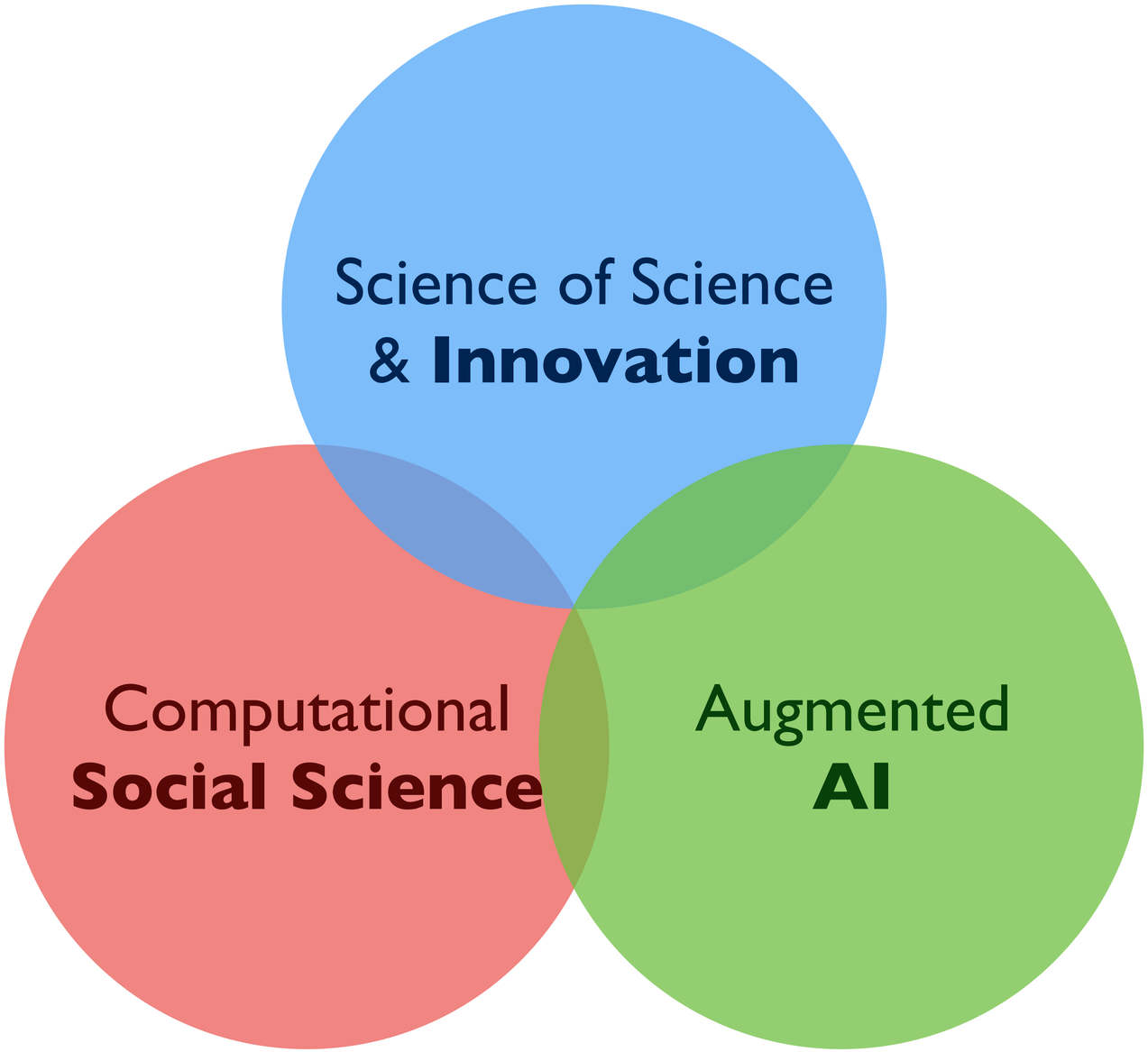
Designing Diversity

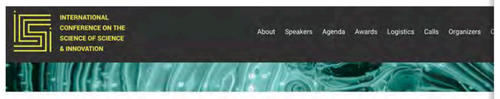
for

Sustained Innovation

james evans

KNOWLEDGE LAB

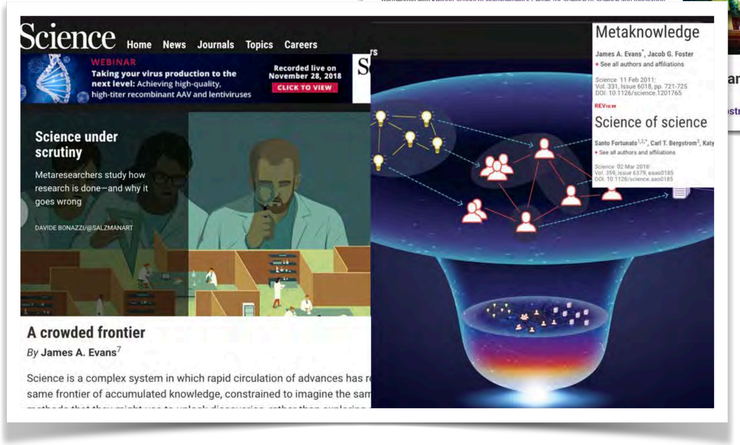

Science of Science
& **Innovation**
Computational
Social Science
Augmented
AI

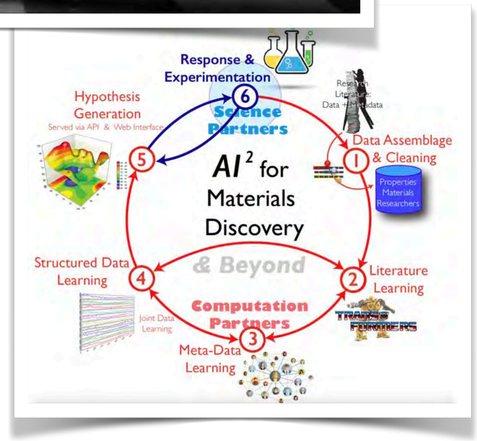


ant deadlines

stracts: February 28 - March 22, 2022

Critical Technologies Network

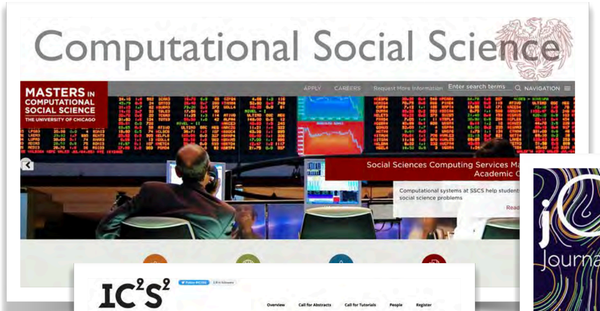




Science of Science & **Innovation**

Computational **Social Science**

Augmented **AI**

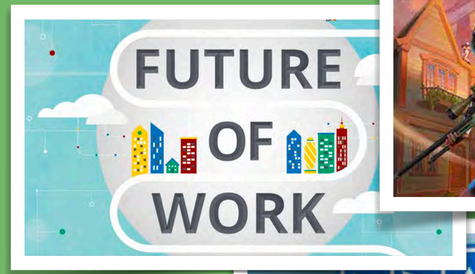


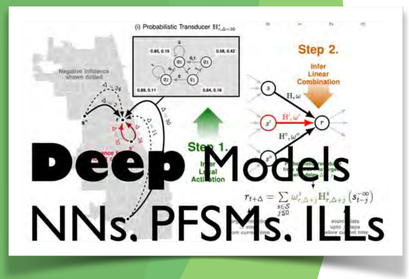
Plurality

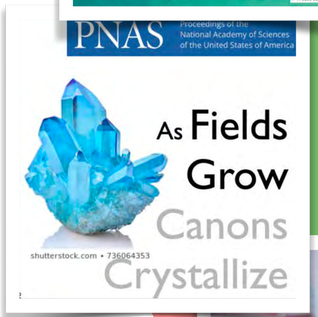
Institute for TRUSTED INTELLIGENCE in society

Science of Science
& **Innovation**

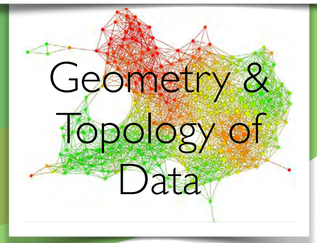
PRINCETON UNIVERSITY PRESS *Knowing*







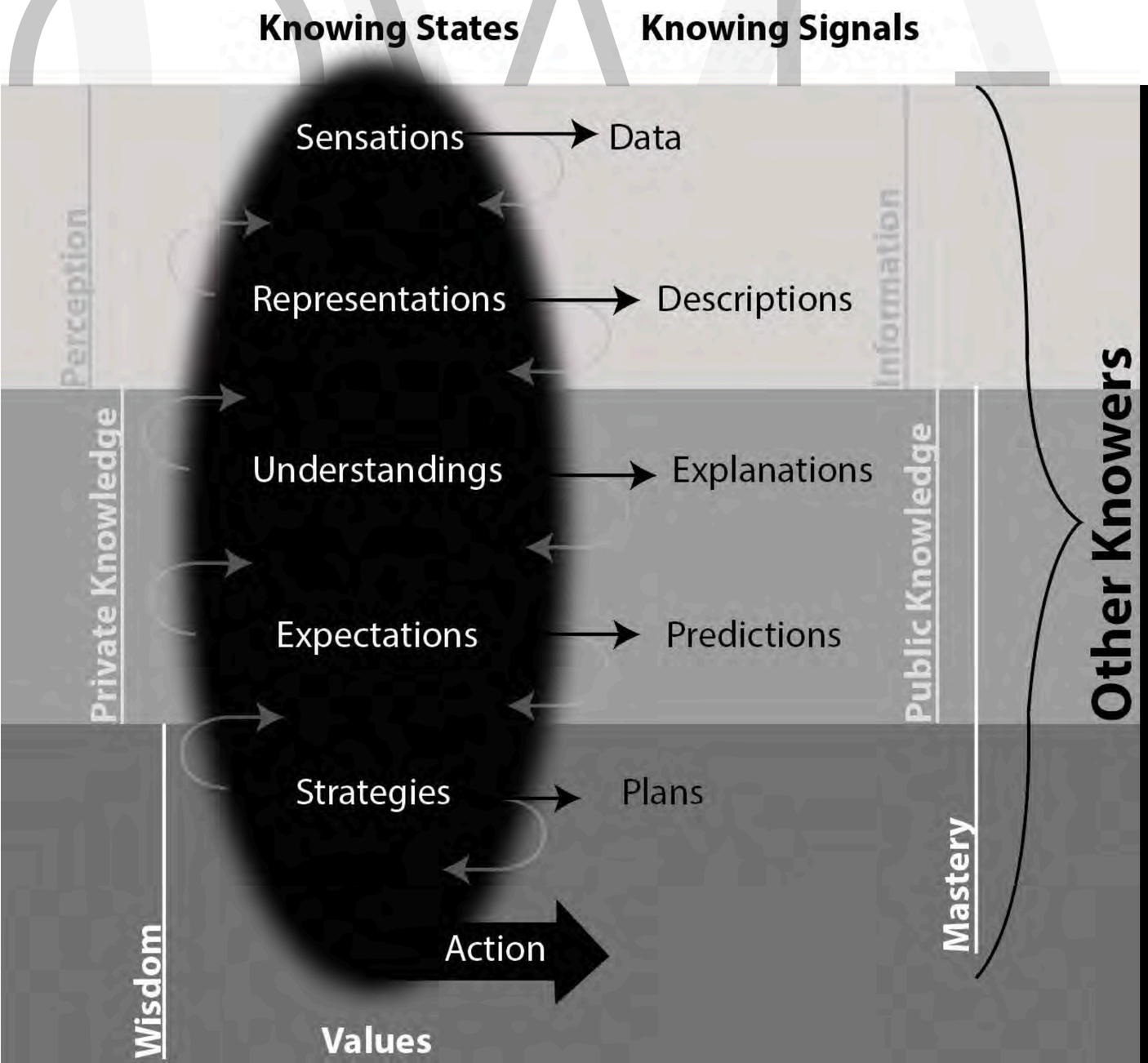
Thinking with Deep Learning

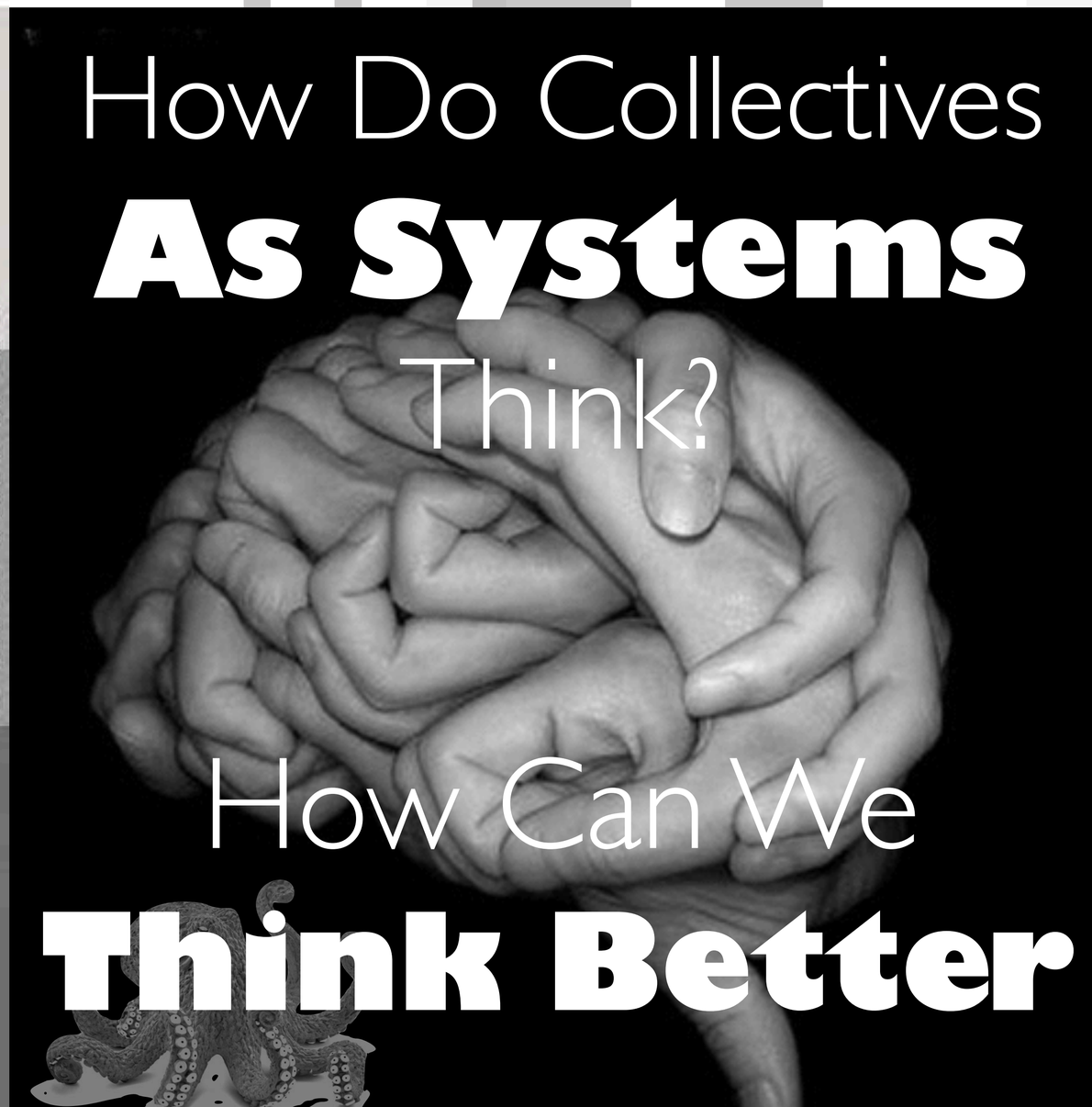


Computational
Social Science

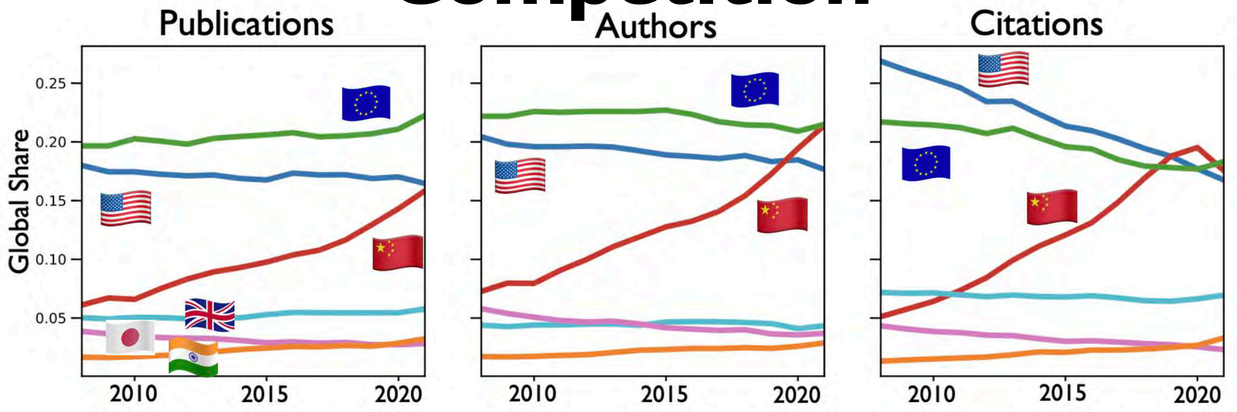
Augmented
AI

A Complex System View of Knowledge

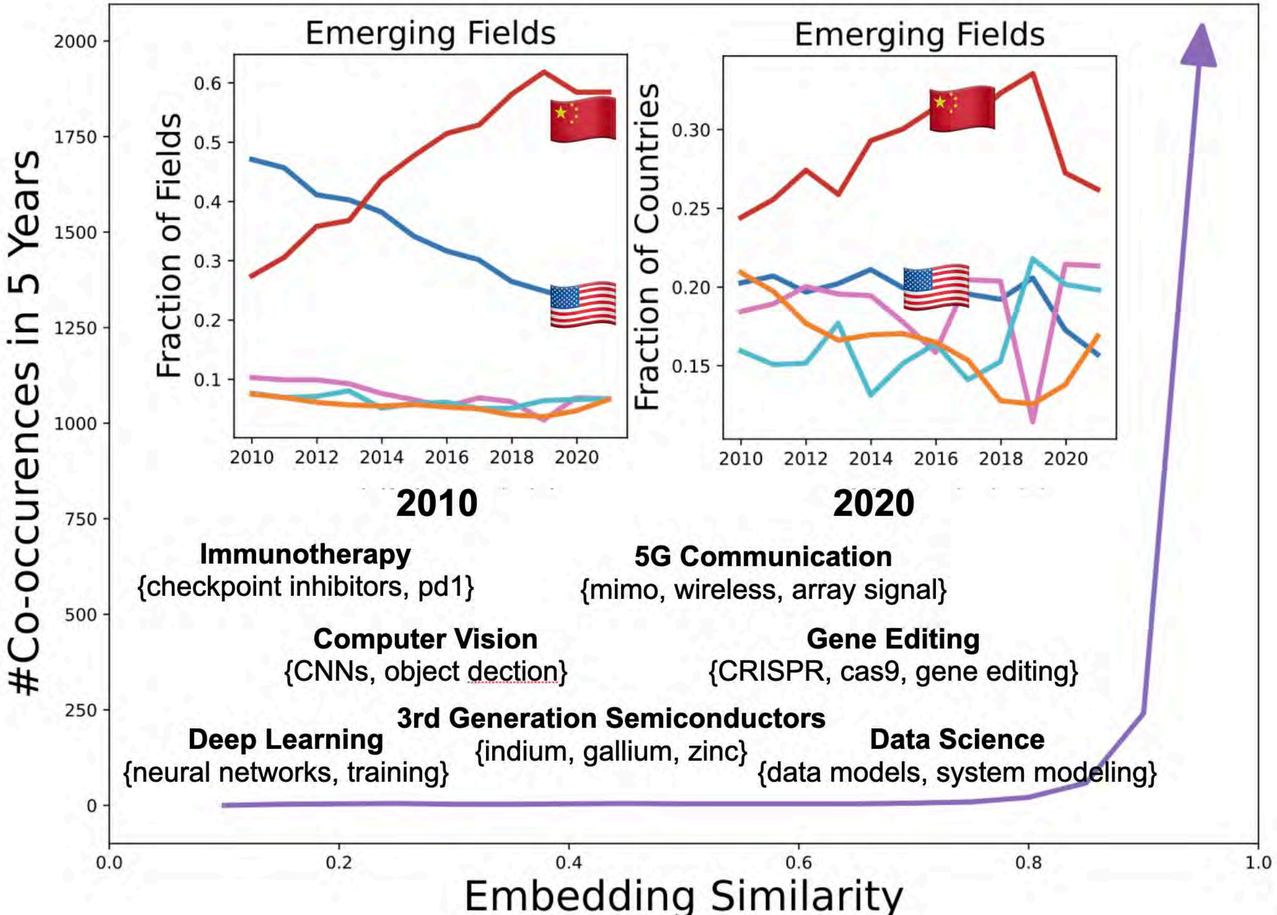




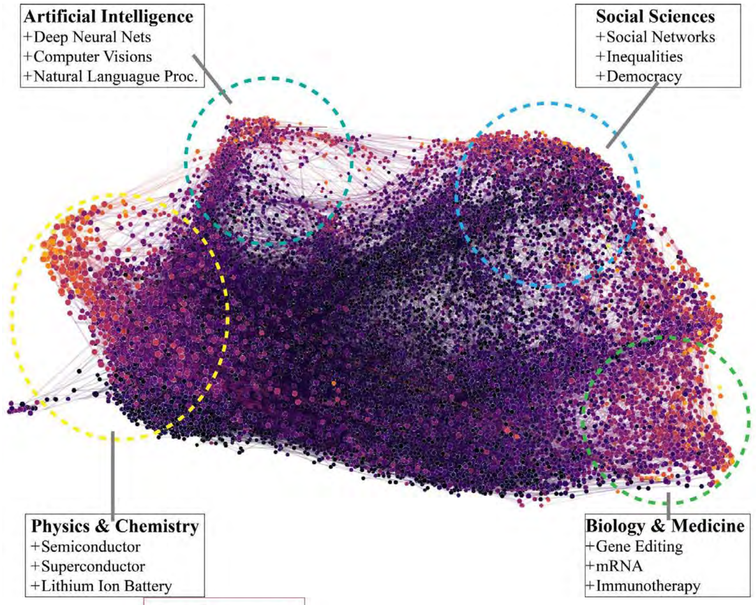
Competition



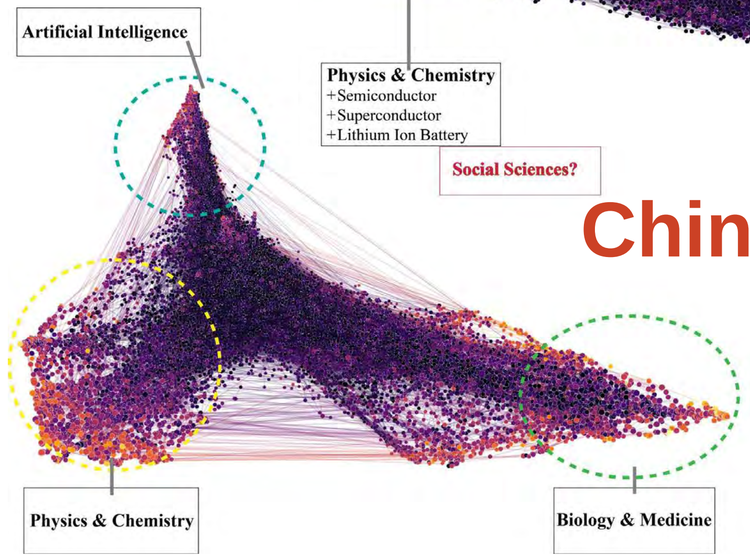
Collaboration

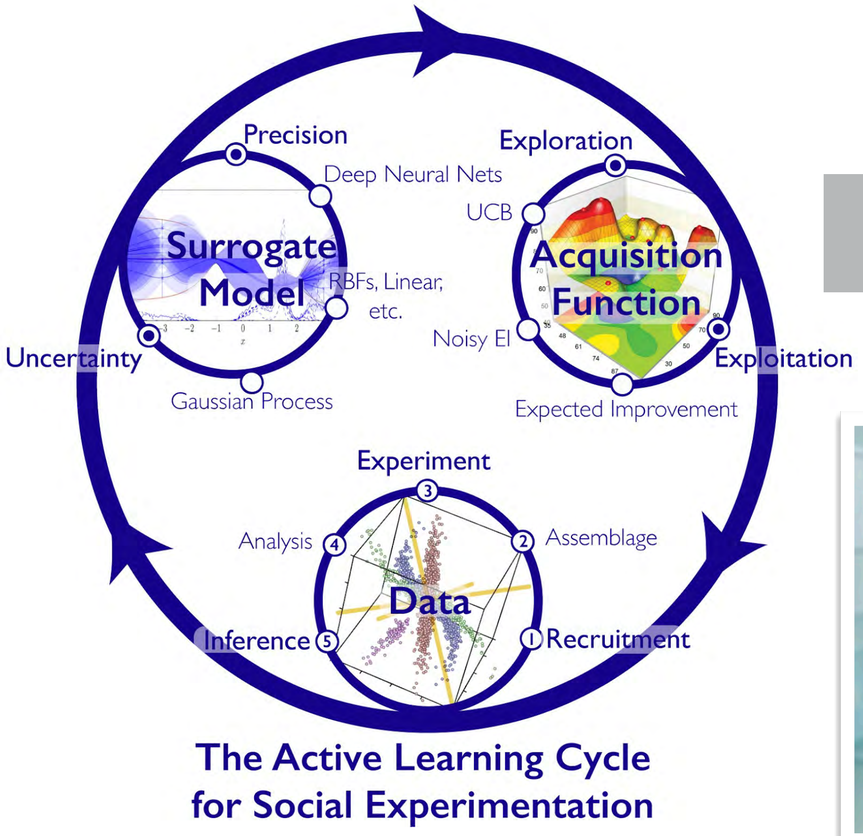


U.S.A.



China

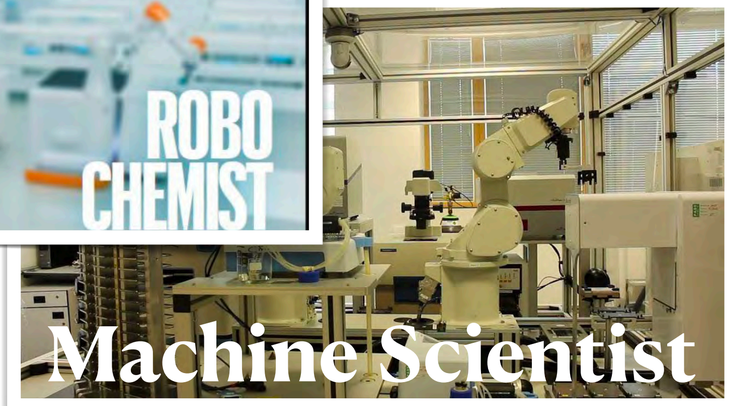




The Active Learning Cycle for Social Experimentation

Understand

Innovation



Simulate/Automate

Innovation

GEN 1 CAPITAL

Meta

amazon

sensetime

Tencent 腾讯

Alibaba.com

Wisdom of Crowds

Vox Populi

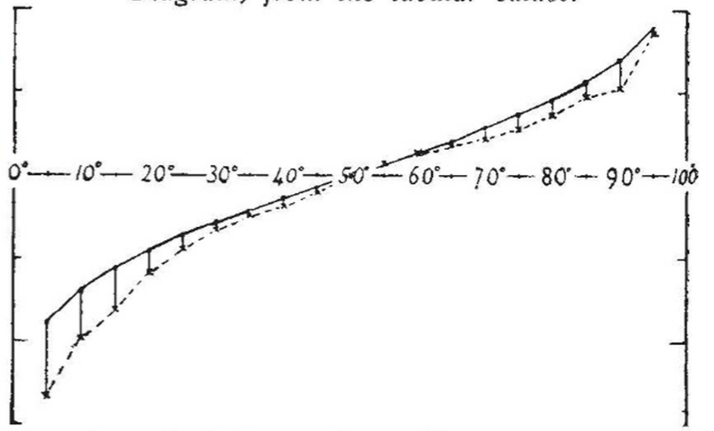
West of England
Fat Stock and Poultry Exhibition

Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

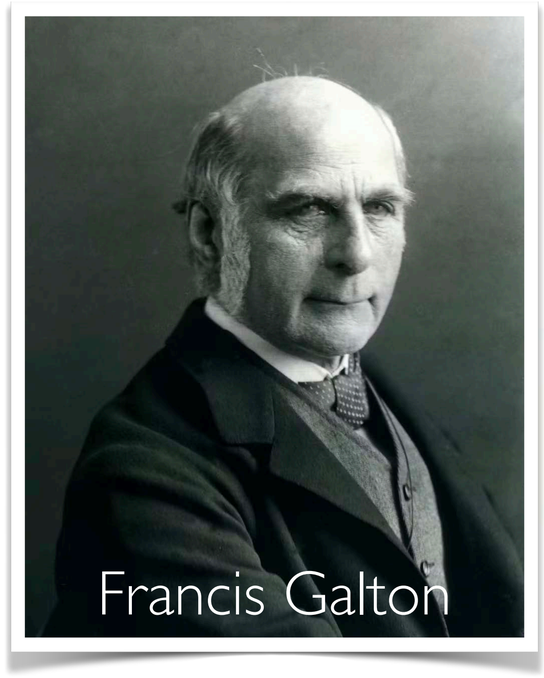
Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e =37	Excess of Observed over Normal
°				
5	1074	−133	−90	+43
10	1109	− 98	−70	+28
15	1126	− 81	−57	+24
20	1148	− 59	−46	+13
q_1 25	1162	− 45	−37	+ 8
30	1174	− 33	−29	+ 4
35	1181	− 26	−21	+ 5
40	1188	− 19	−14	+ 5
45	1197	− 10	− 7	+ 3
m 50	1207	0	0	0
55	1214	+ 7	+ 7	0
60	1219	+ 12	+14	− 2
65	1225	+ 18	+21	− 3
70	1230	+ 23	+29	− 6
q_3 75	1236	+ 29	+37	− 8
80	1243	+ 36	+46	−10
85	1254	+ 47	+57	−10
90	1267	+ 52	+70	−18
95	1293	+ 86	+90	− 4

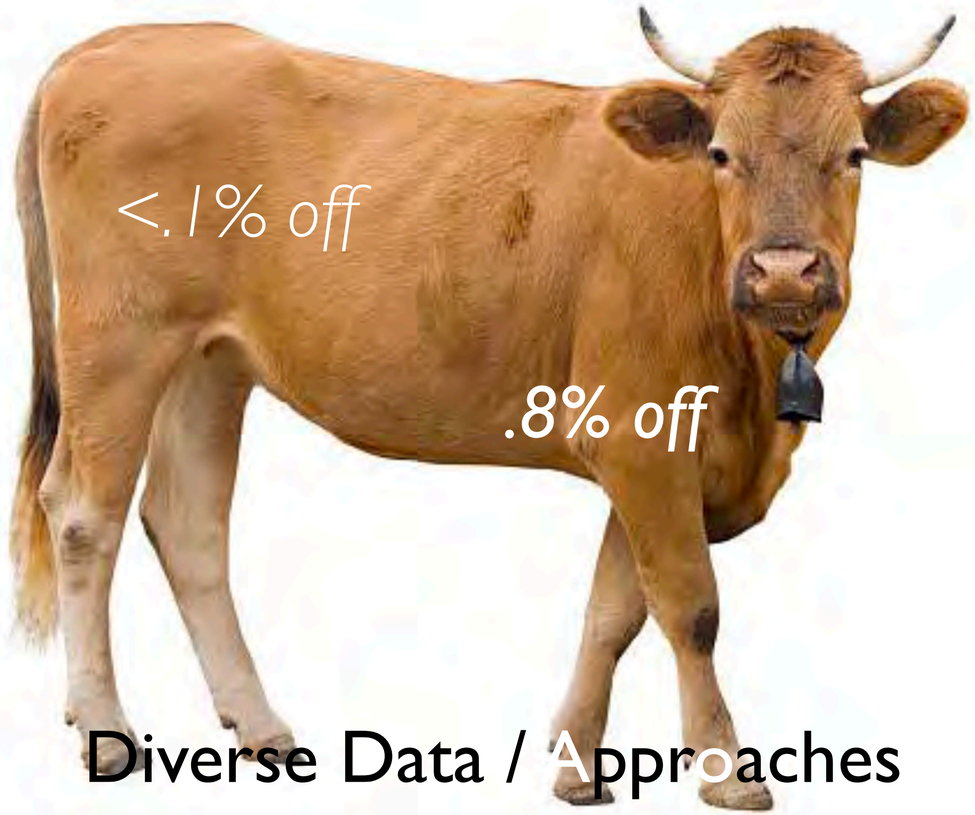
q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

Diagram, from the tabular values.



The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed





Diverse Data / Approaches

Diversity

Cognitive

Perceptual / Experiential

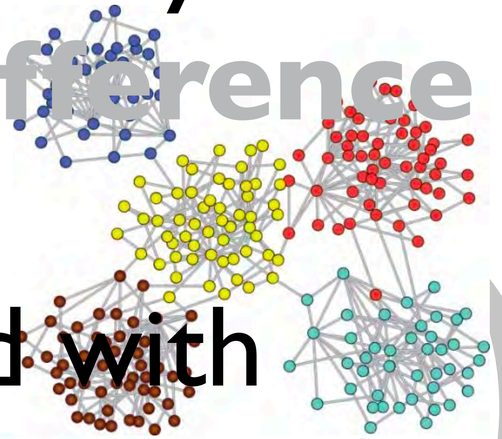
Identity

Wisdom of
Science Technology Business
Innovating
Crowds

Chapters

1. Innovative Scientific Crowds characterized by
Collective Abduction & Bridged Difference

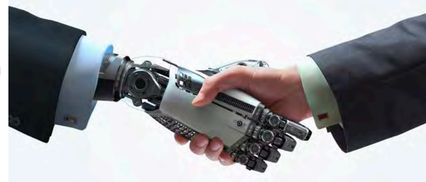
SFI

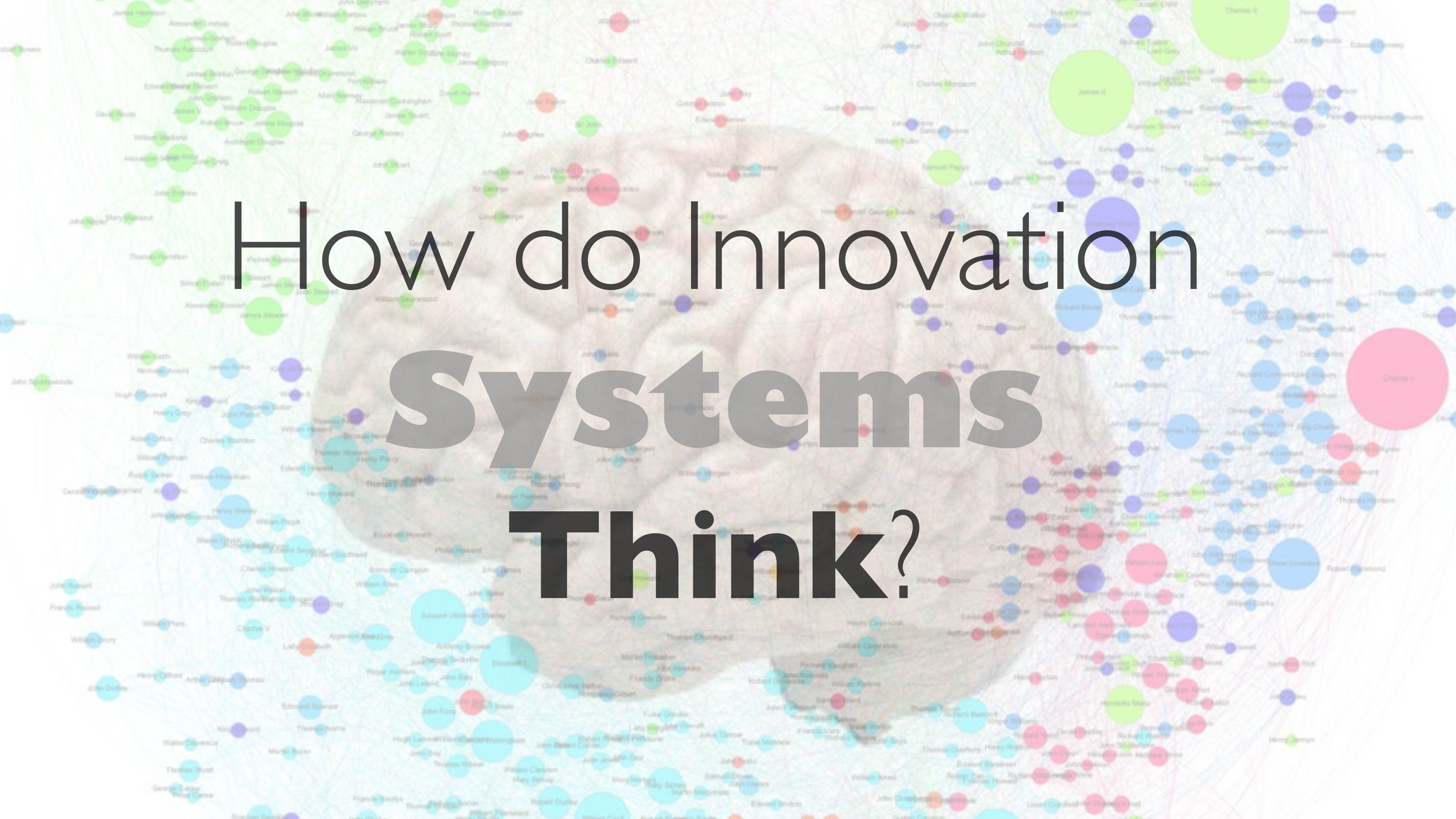


2. Connection and Conformity is associated with
Reduced Replication & Impeded Innovation

AI

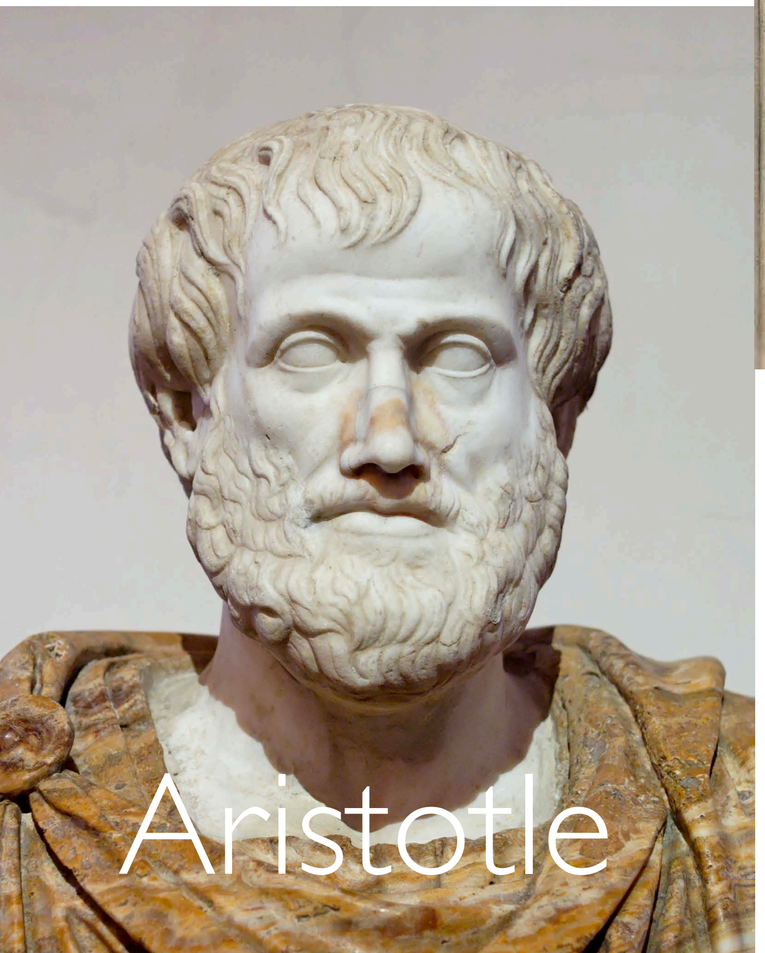
3. AI that Designs Diversity Beyond Human Limits to
Accelerate and Punctuate Advance

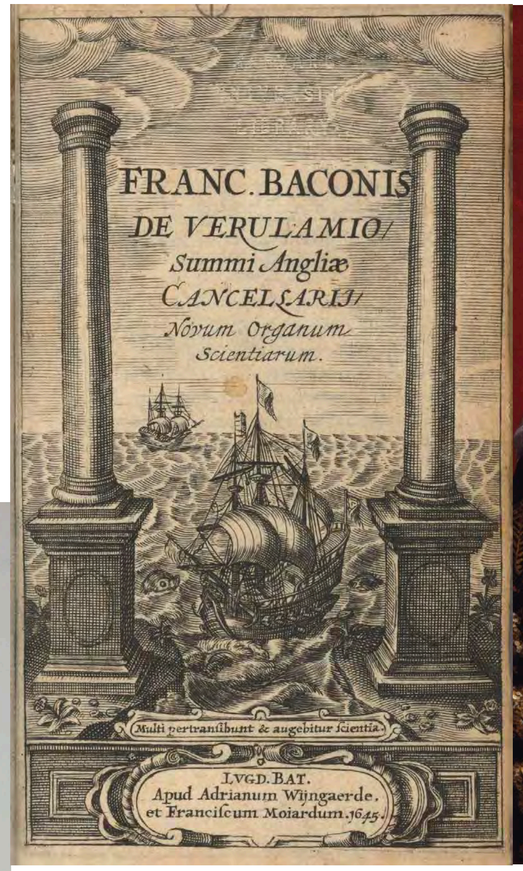



How do Innovation
Systems
Think?

Deduction

Logically extend from facts/axioms



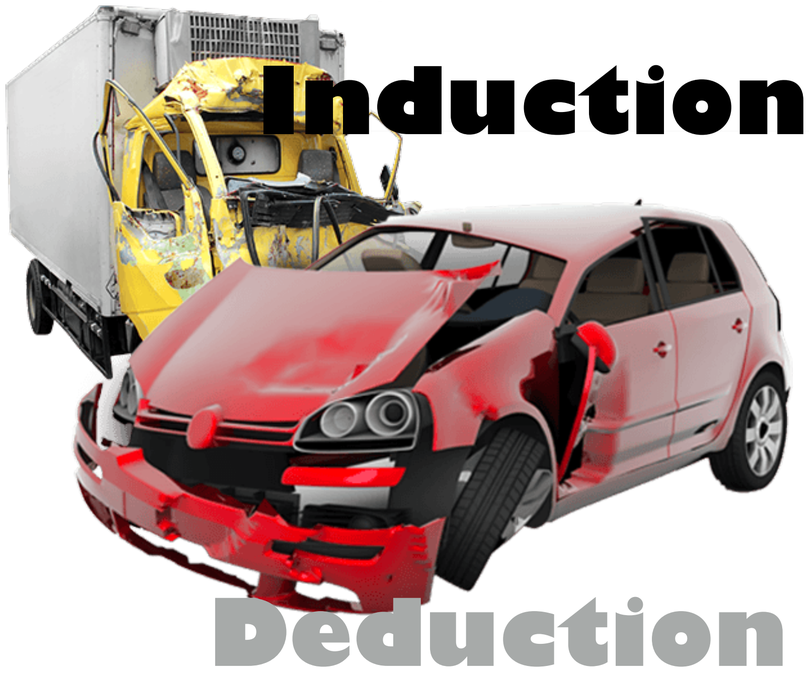


Imaginatively generalize from observations

Induction

Abduction

surprise



PubMed

20M PubMed articles (1865-2015)

15,000 MeSH term Concepts (e.g., PCR, hypertension, DNA, testosterone)

APS physics™

.5M APS articles (1880-2015)

80,000 PACS code Concepts (e.g., neutron star core, lie algebras, polarization)

UNITED STATES PATENT AND TRADEMARK OFFICE

uspto

1.5M US Patents

45,000 USPC subclasses (e.g., arc lamp, electrolytic condenser, paper, button)

Contents

Concepts, Phenomena, Tools

Unlikely to have been successfully combined

Unlikely to have been previously imagined

Contexts

Journals, Conferences, Coversations

To Identify Surprise

requires a

AI-Driven

Sufficient Model

Digital Double

Almost Everything

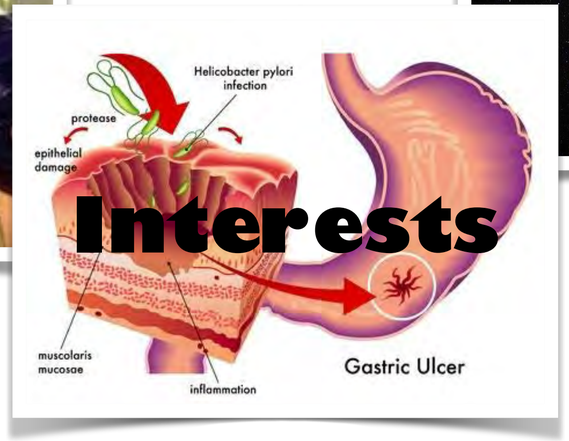
Surprising (improbable) outcomes

$$Novelty \propto \frac{1}{P(I|A,S)}$$

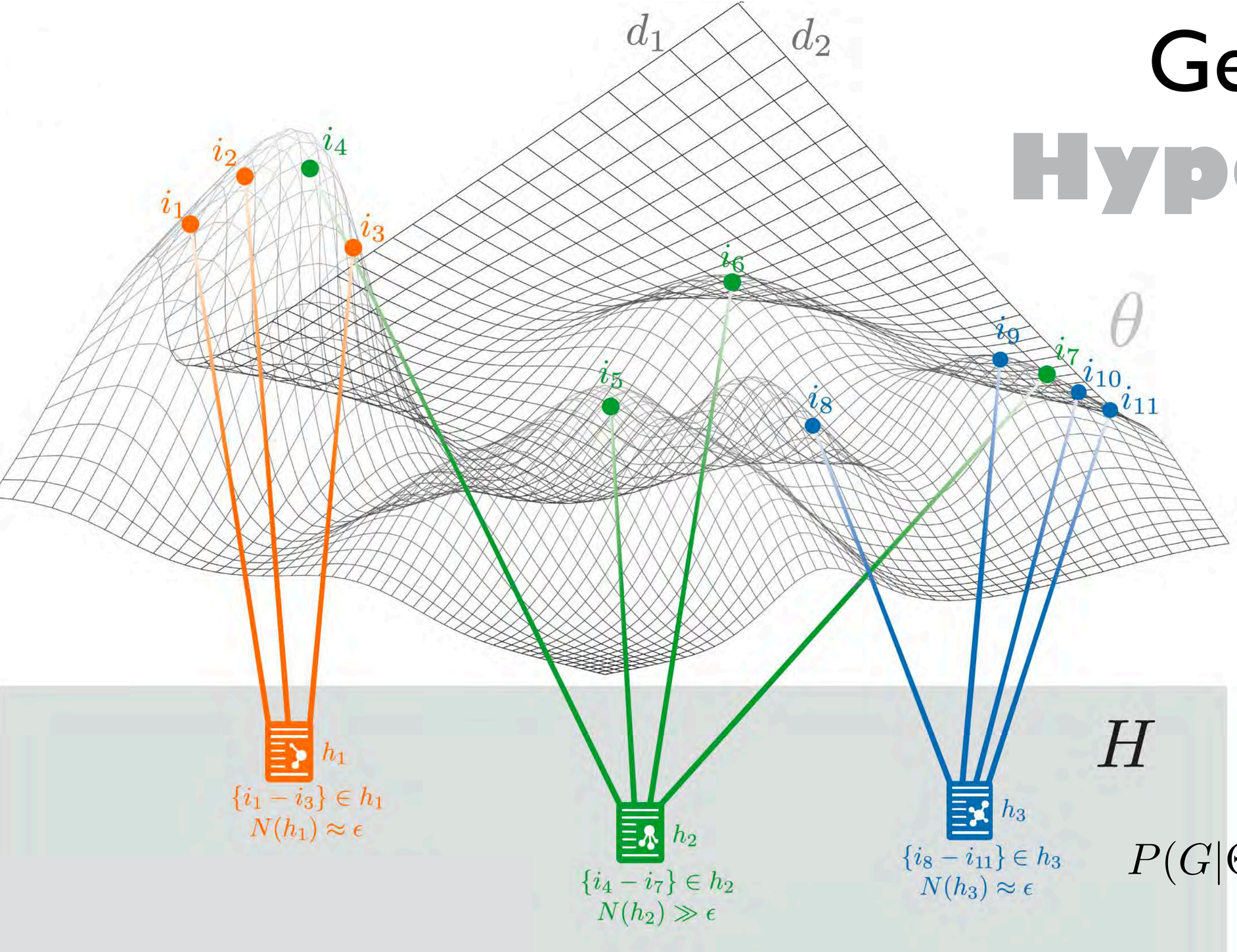
Decreasing Function

Dumbphone

Heliobacter pylori



Oumuamua



Generative Hypergraph Model

1.

$$\lambda_h = \sum_d \prod_{i \in h} \theta_{id} \times \prod_{i \in h} r_i$$

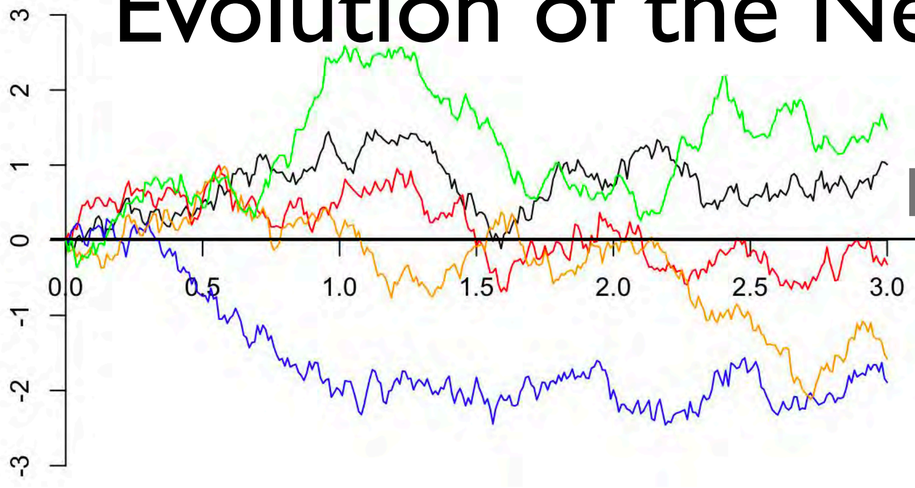
2.

$$X_h \sim Poisson(\lambda_h)$$

3.

$$P(G|\Theta, R) = \prod_{h \in H} P(x_h|\Theta, R)$$

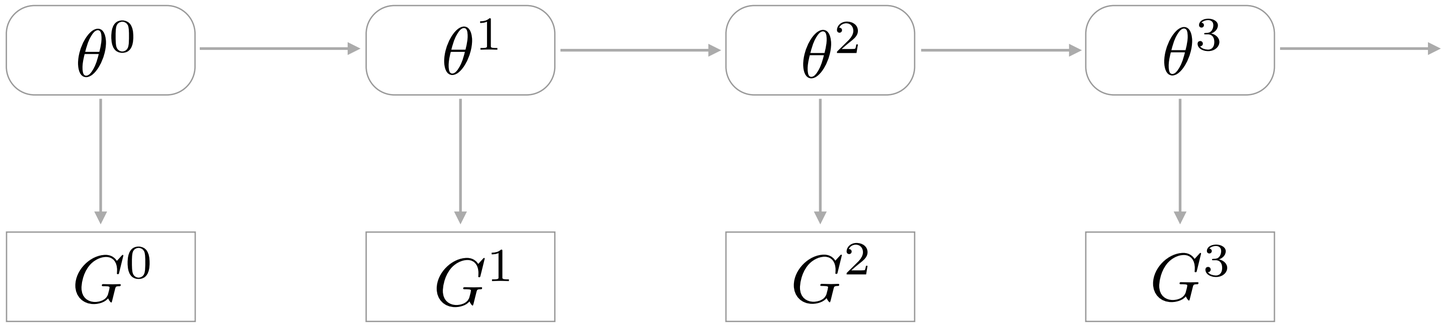
Evolution of the Network



Hidden Markov Model

Brownian Motion

Time



G^t: observed network at time *t*

θ^t: latent positions of the elements at time *t*

Complete Model

- Log-likelihood function

$$l(\theta_1, ..., \theta_T) = \log P(G_1, ..., G_T | \theta_1, ..., \theta_T)$$

$$= \sum_{t=1}^{T} [\log P(\theta^t | \theta^{t-1}) + \log P(G^t | \theta^t)]$$

Gaussian Density for Brownian motion

Poisson Density for Hyperedge Section

$$= \sum_{t=1}^{T} [\sum_i \sum_k (\theta_{ik}^t - \theta_{ik}^{t-1})^2 / 2\sigma^2 + \sum_{h \in G^t} (x_h \log \sum_k \prod_{i \in h} \theta_{ik}^t - \sum_k \prod_{i \in h} \theta_{ik}^t)]$$

- Impossible to optimize!

Incomputable

2^N possible combinations

Maximal Likelihood Estimate

Theorem

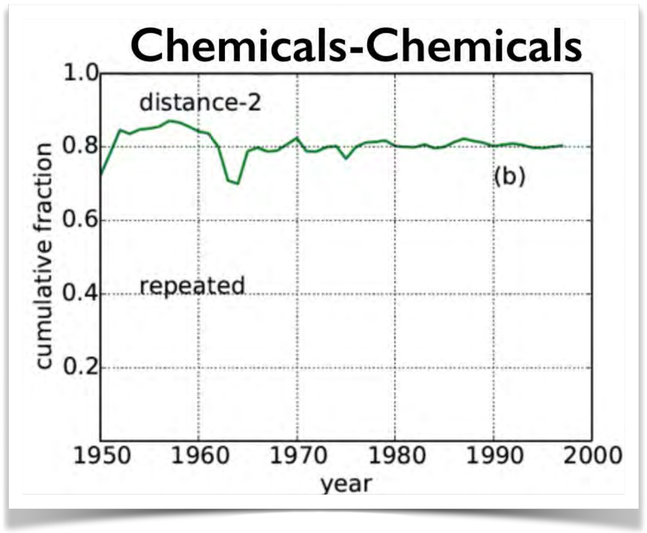
Let $f(\theta, t) = \sum_{d=2}^{D} S_d^t$ and $t \sim randint(1, T)$, then

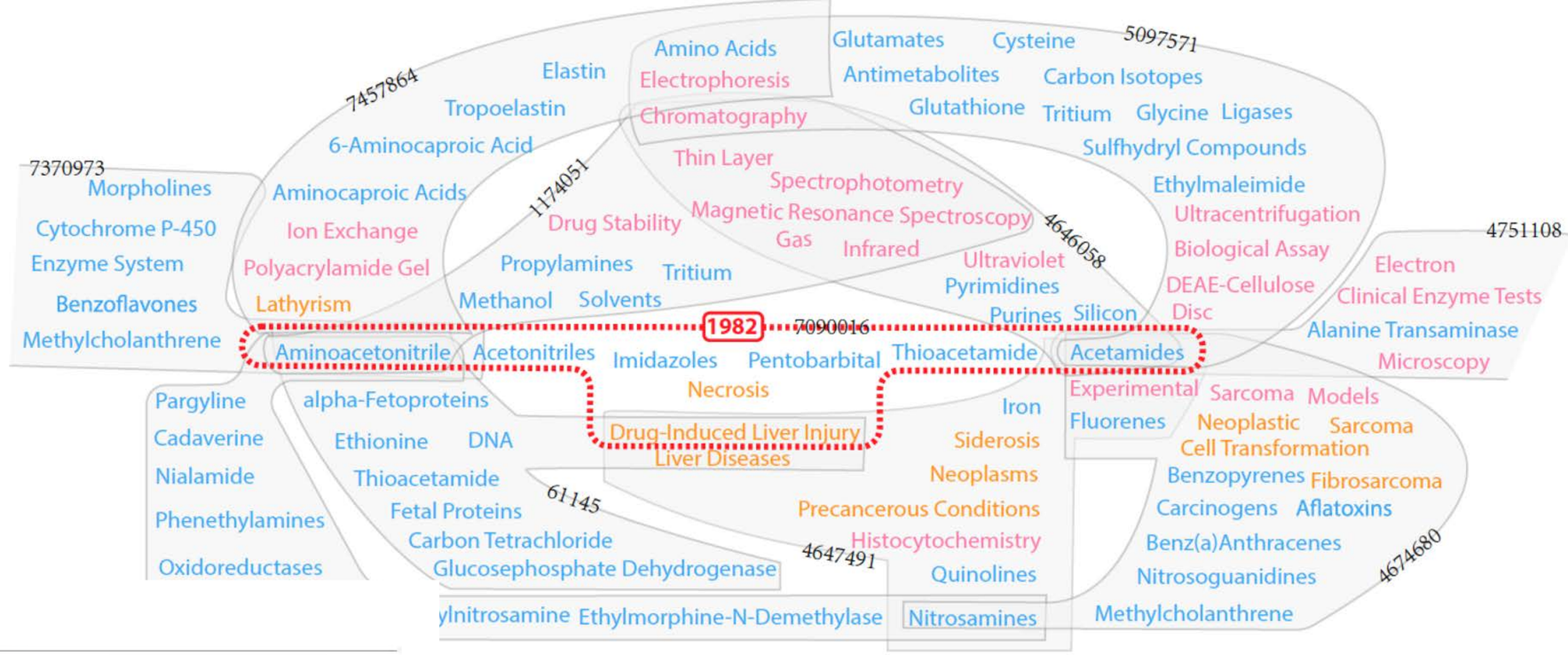
$$E[\nabla f(\theta, t)] = \nabla l(\theta)$$

Corollary

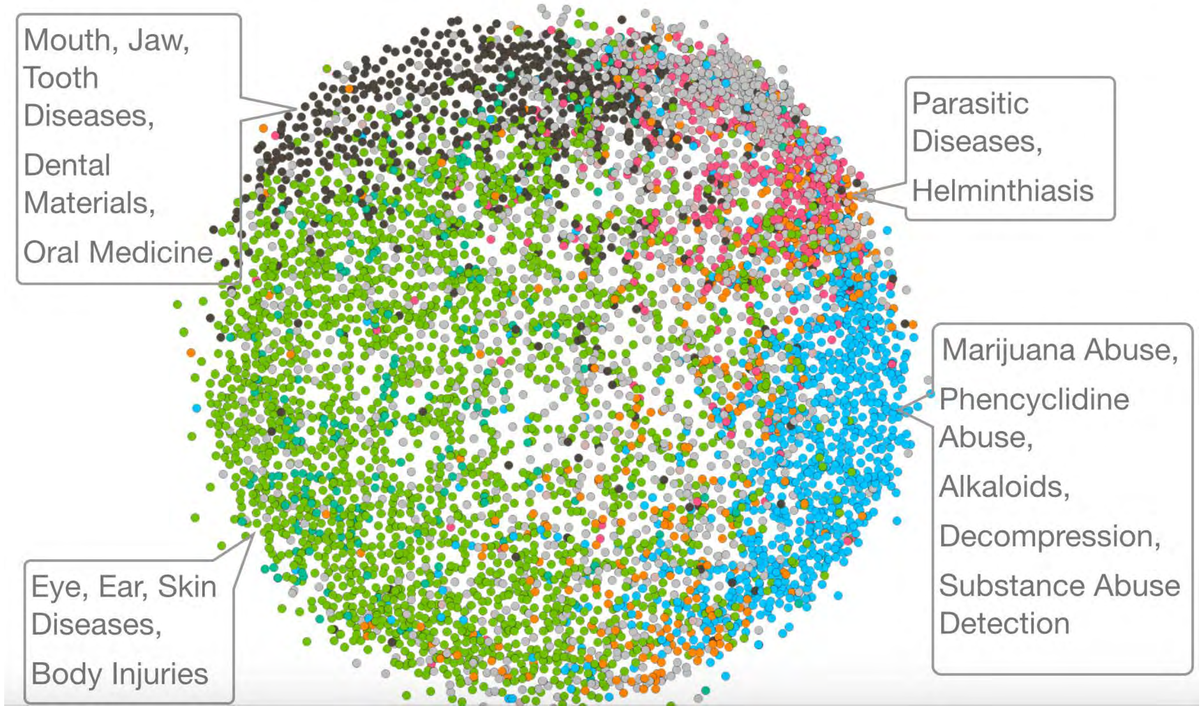
$\hat{\theta}$ will converge to the maximal likelihood estimate.

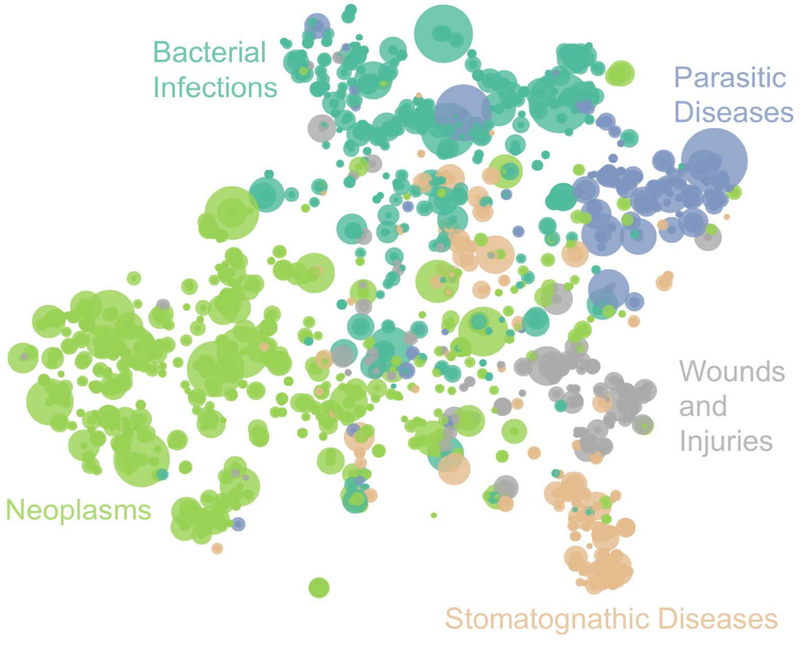
Choice: Negative Sampling vs. Contrastive Noise





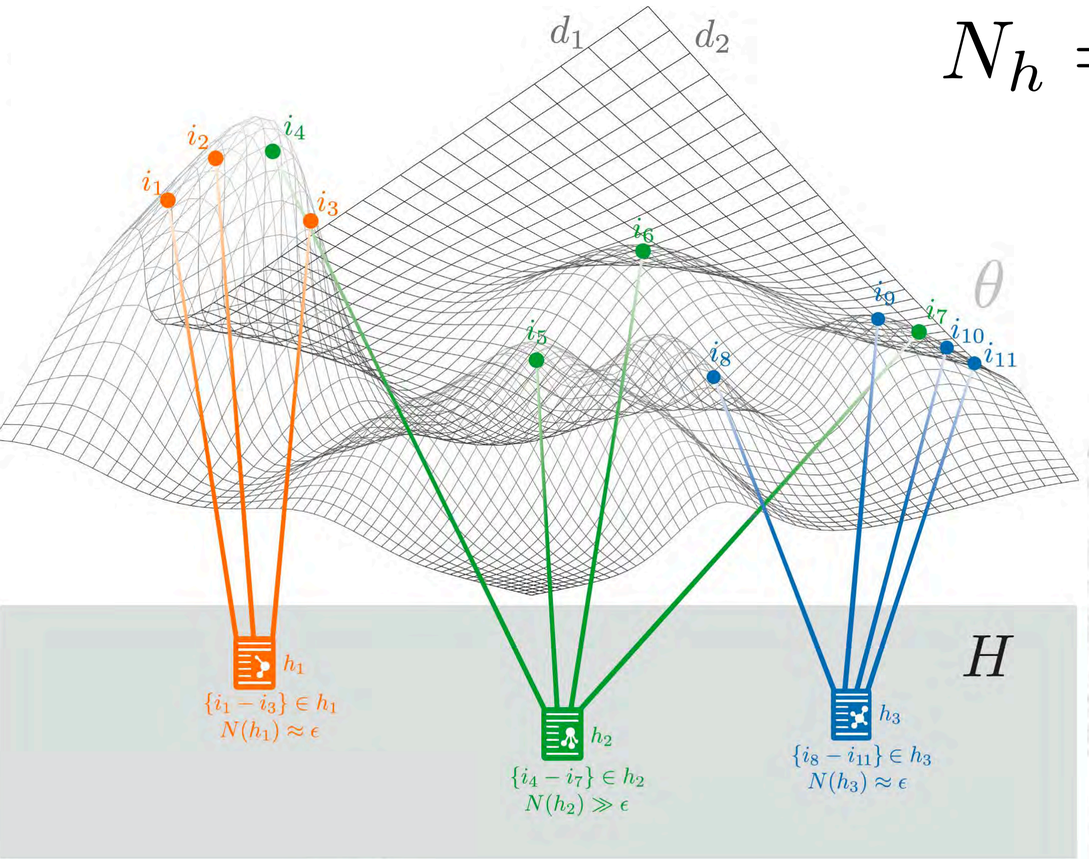
Community Structure

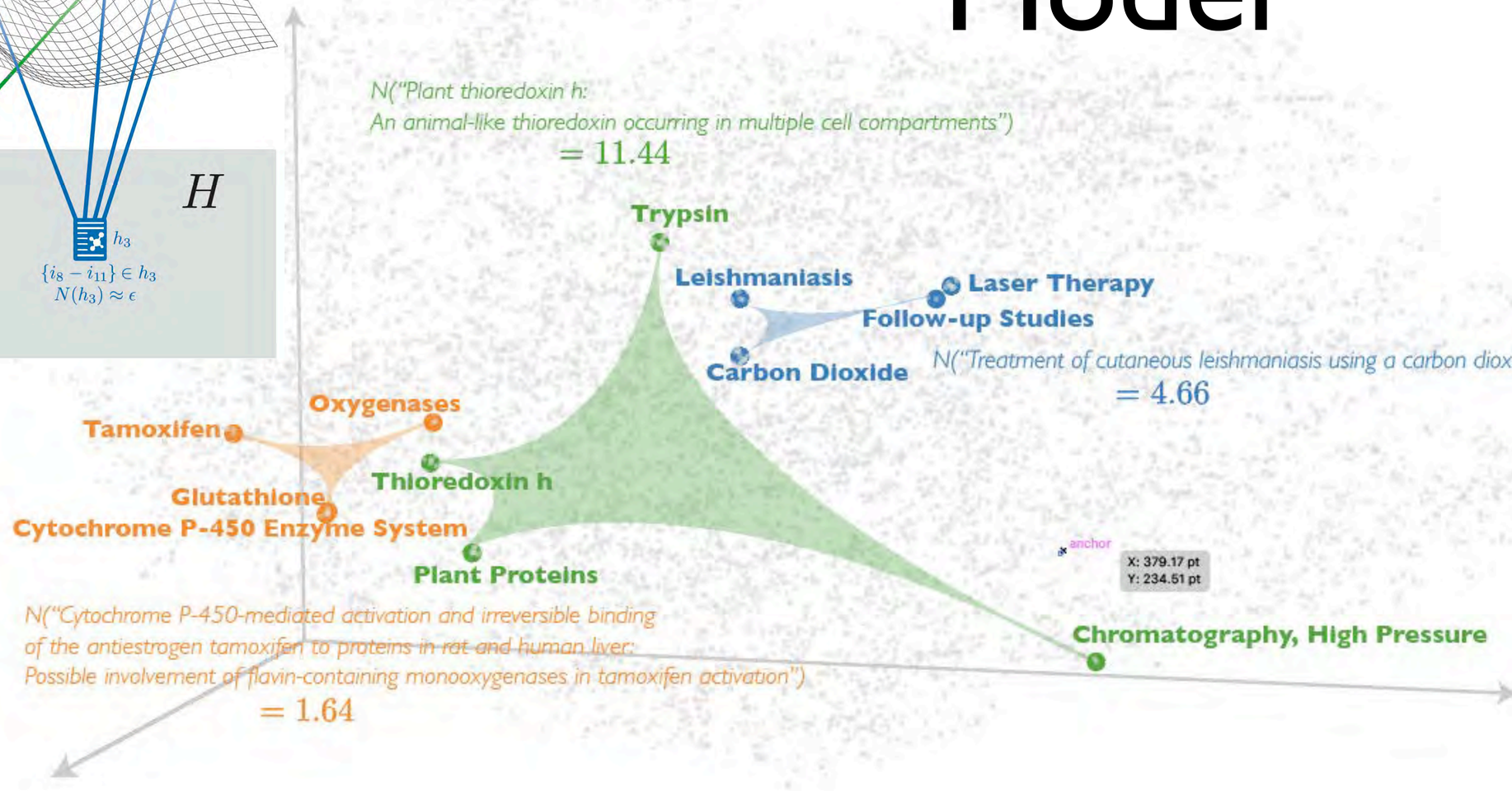




Static Measure Model

$$N_h = -\log \sum_d \prod_{i \in h} \theta_{id}$$

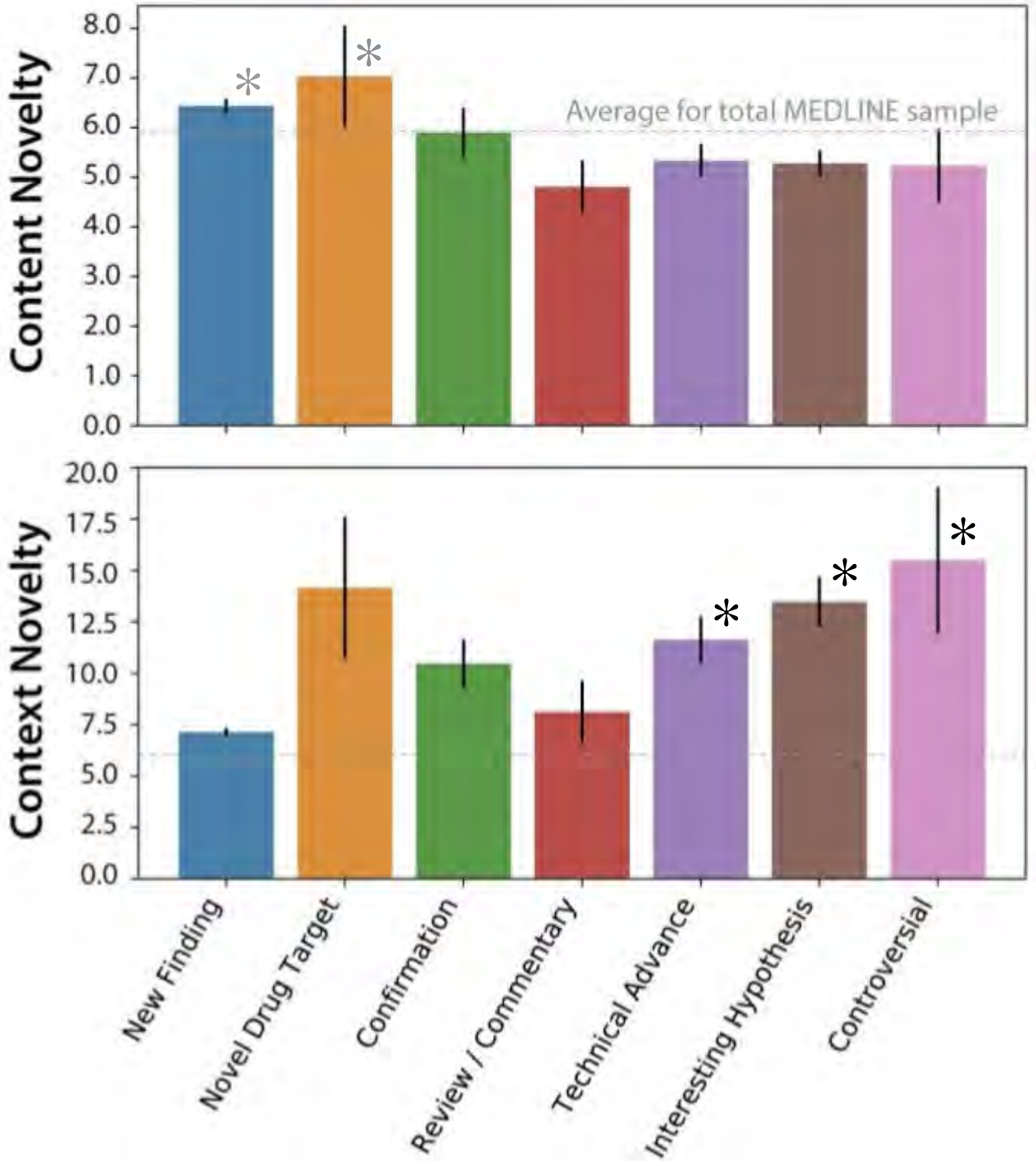




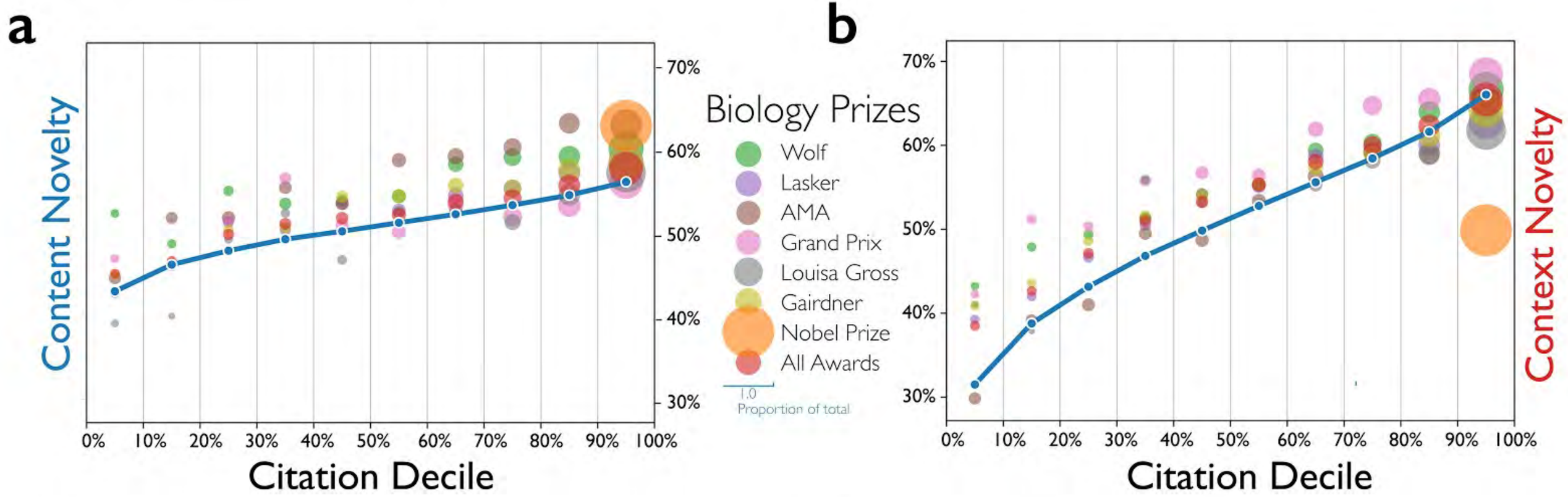
Faculty 1000

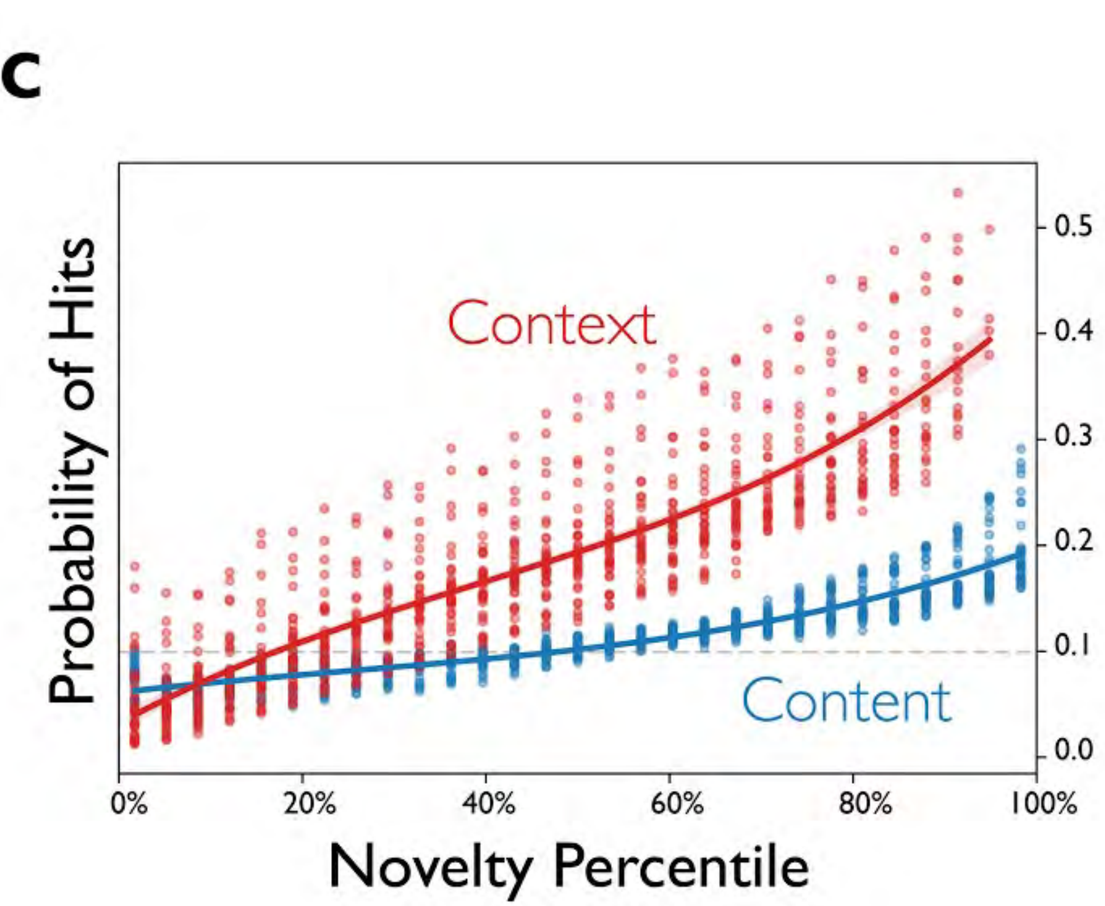
Content novelty is most associated with **New findings** and **Drug targets** emerge from within-domain investigation.

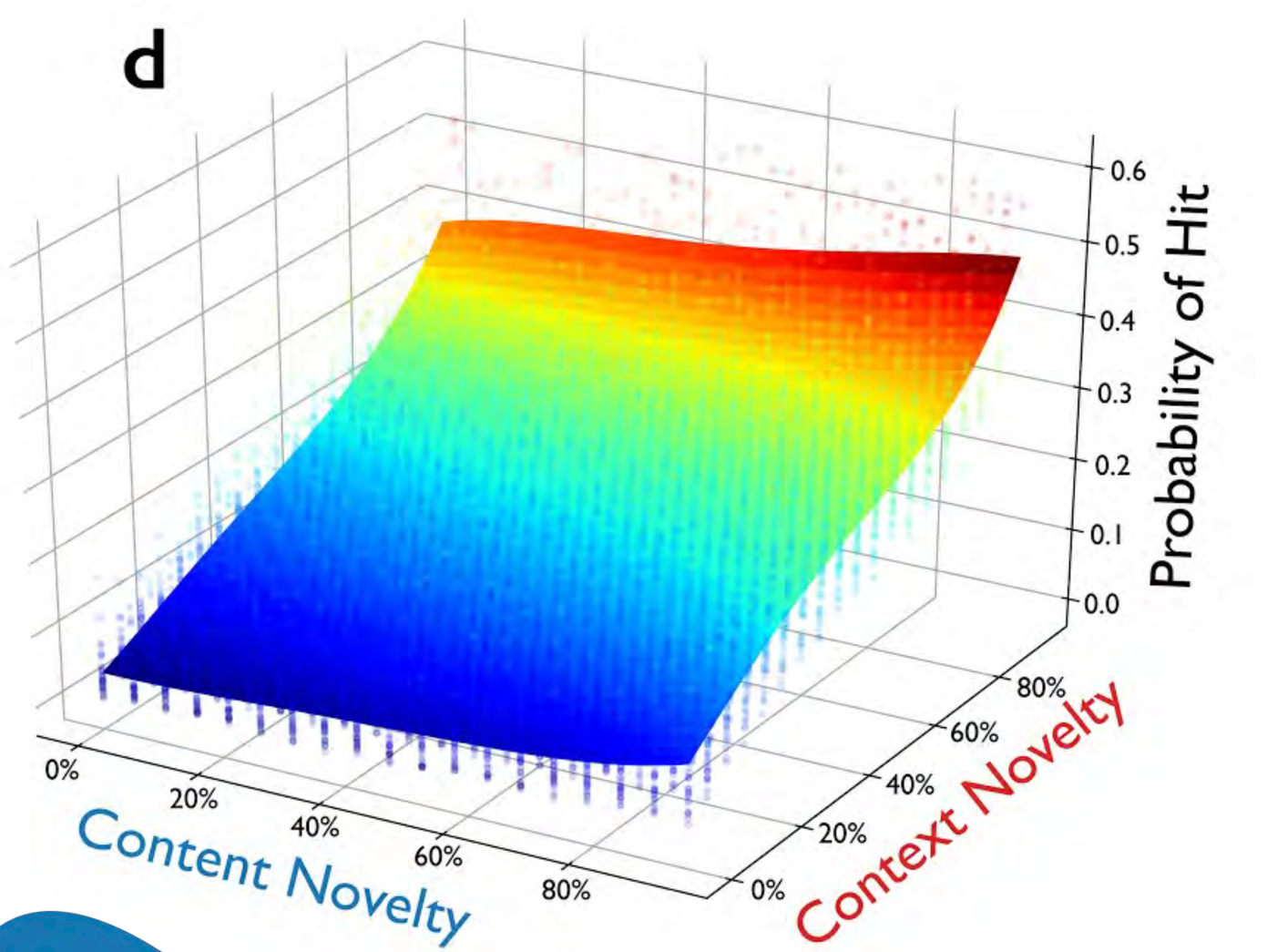
Context novelty is most associated with **Controversial**, **Interesting hypothesis** and **Technical advance** emerge from cross-domain investigation.



PubMed


a
b
Content Novelty
Context Novelty
Citation Decile
Citation Decile
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
30%
40%
50%
60%
70%
Biology Prizes
Wolf
Lasker
AMA
Grand Prix
Louisa Gross
Gairdner
Nobel Prize
All Awards
1.0
Proportion of total

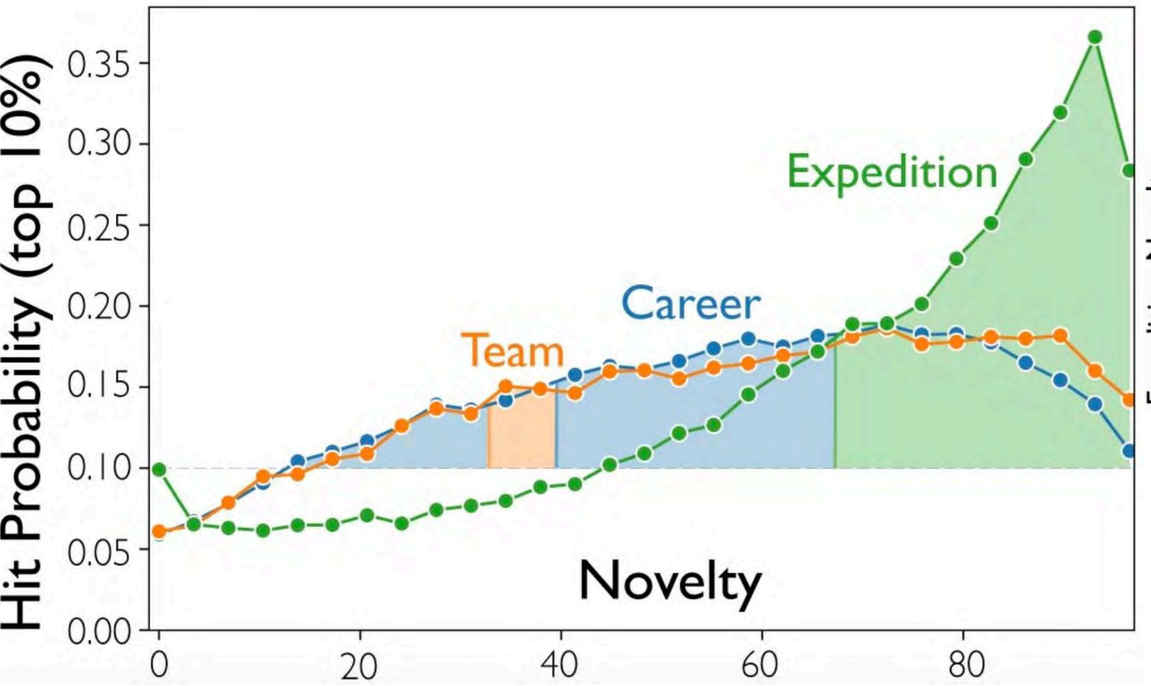


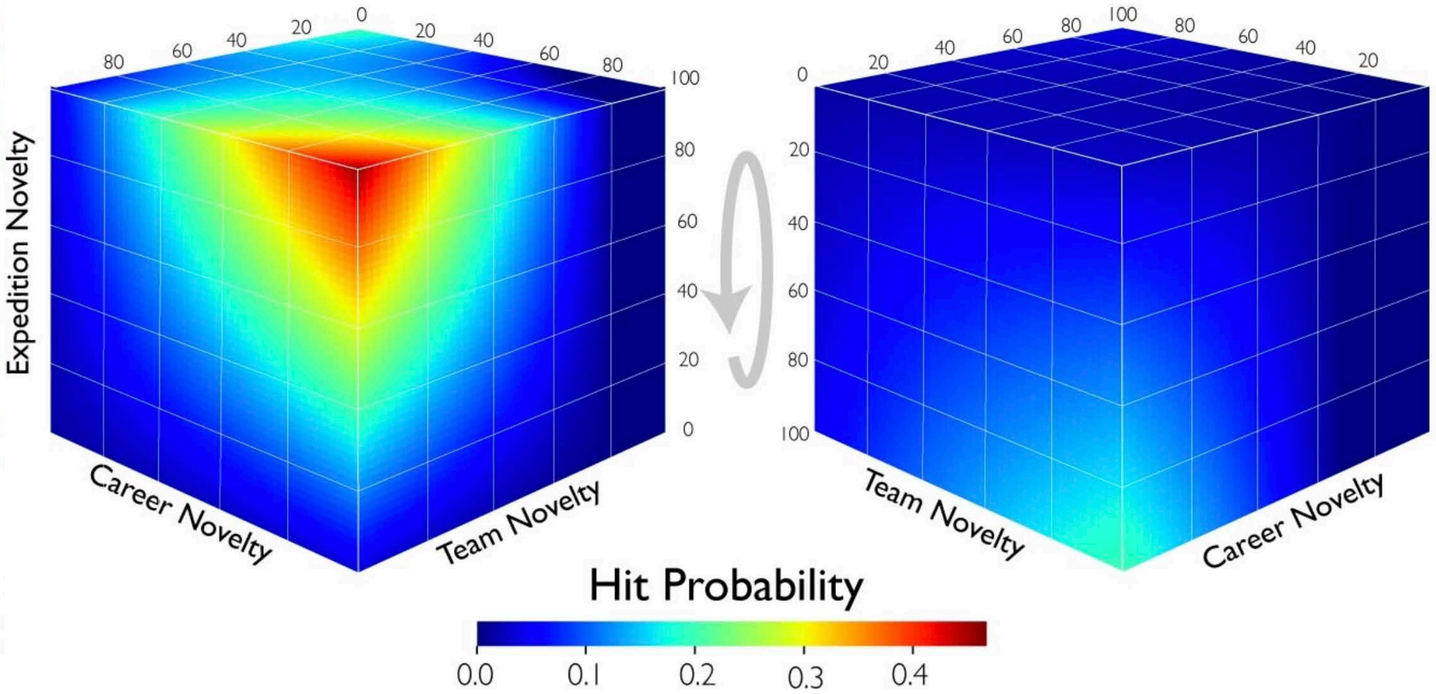


PubMed

Disconnection through Disciplines

Career, Team, & Expedition Novelty



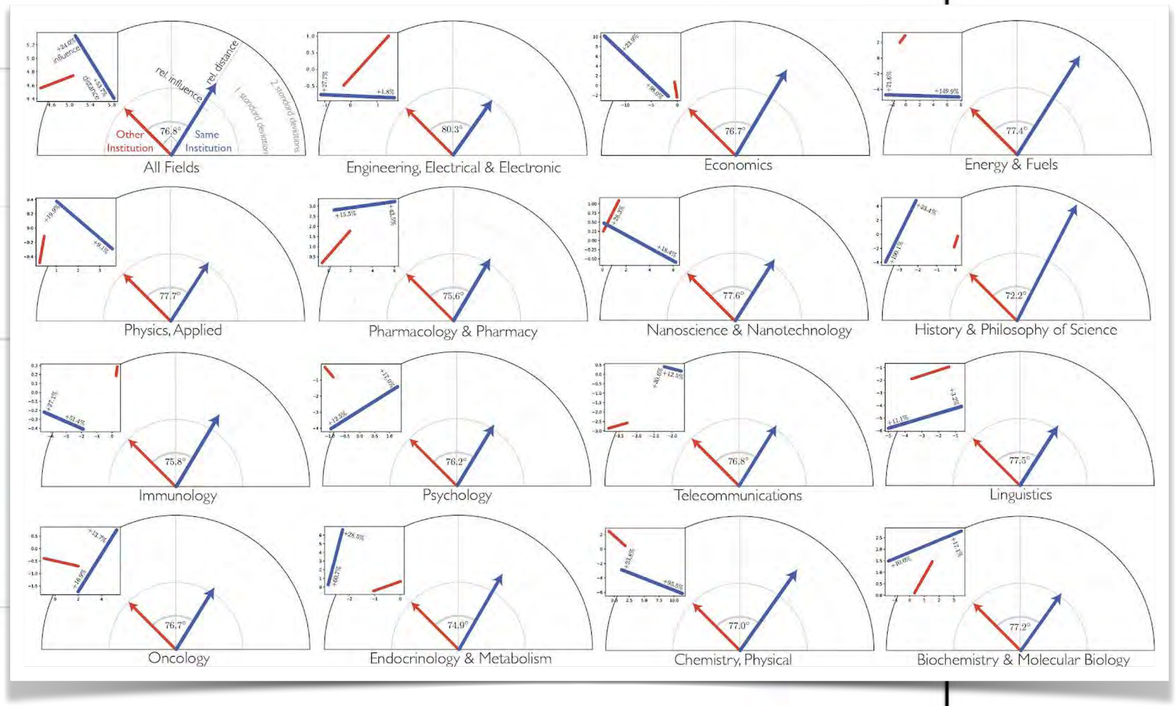


Radical Insight comes in conversation with Outsiders / Aliens

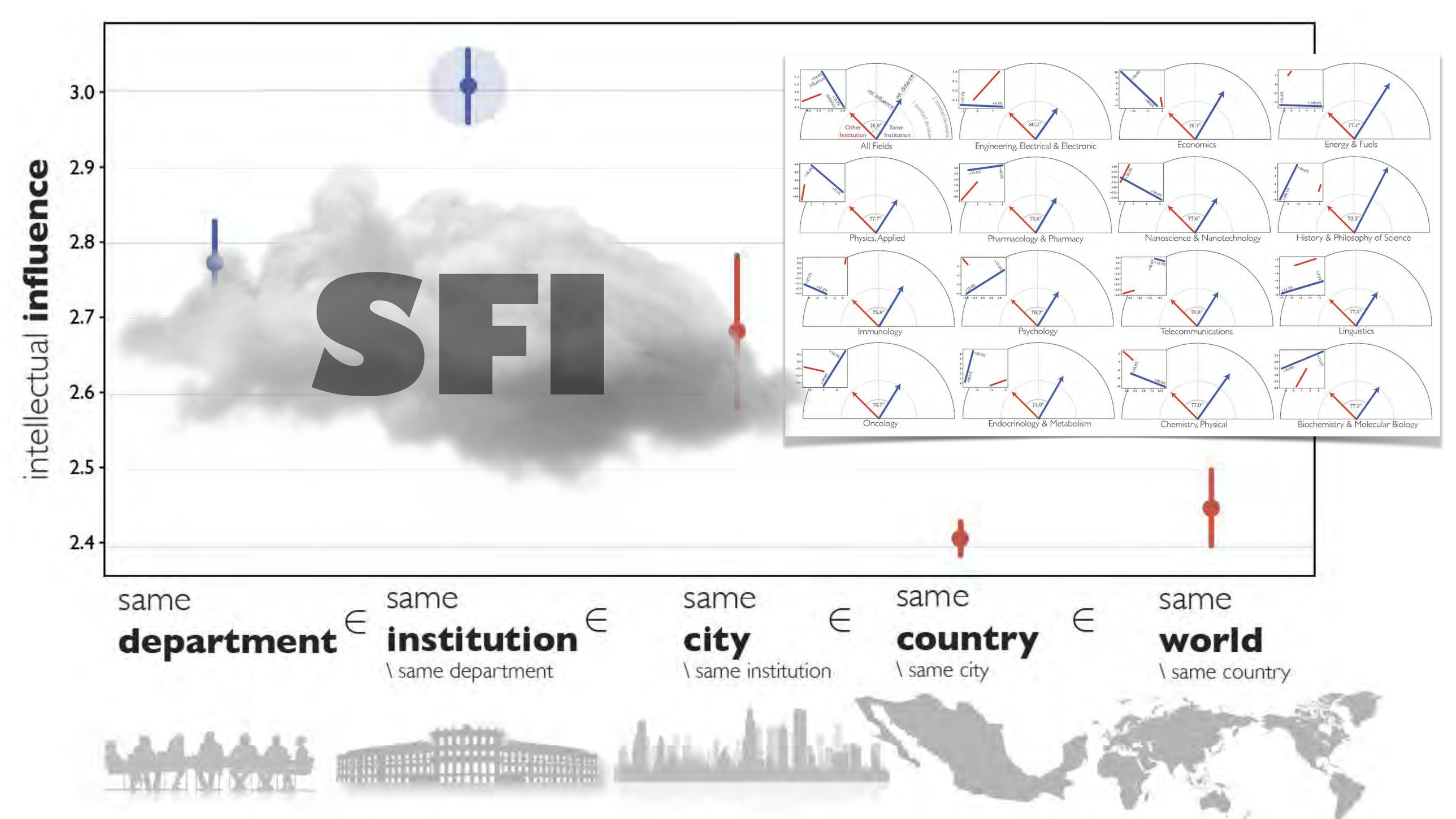
by accident

intellectual **influence**

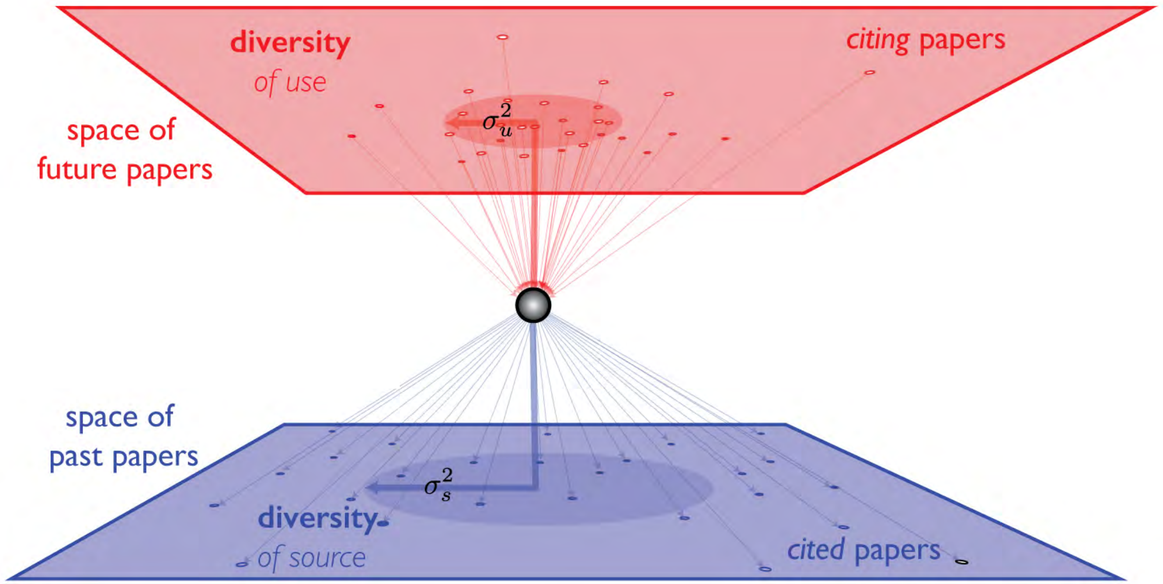
3.0
2.9
2.8
2.7
2.6
2.5
2.4

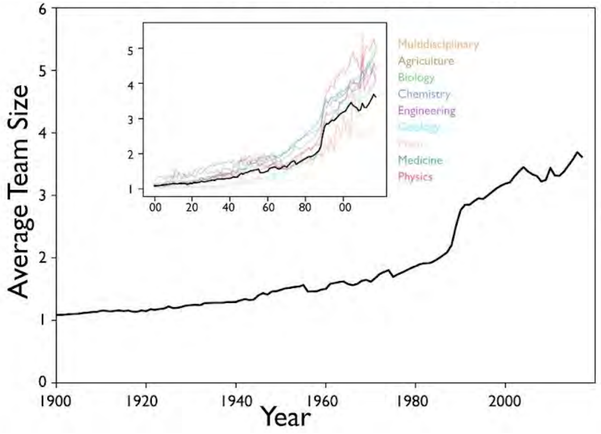


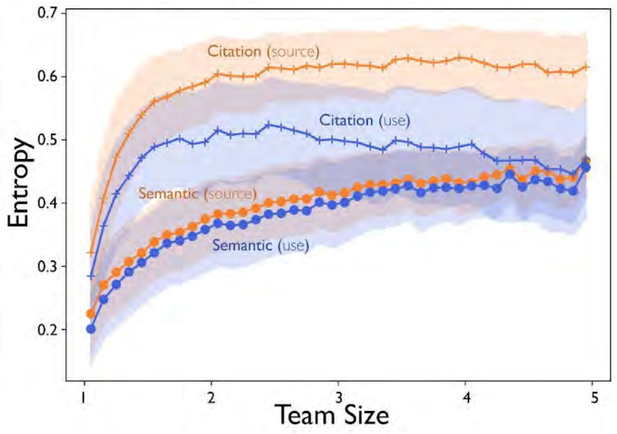
same **department** ∈ same **institution** \ same department ∈ same **city** \ same institution ∈ same **country** \ same city ∈ same **world** \ same country

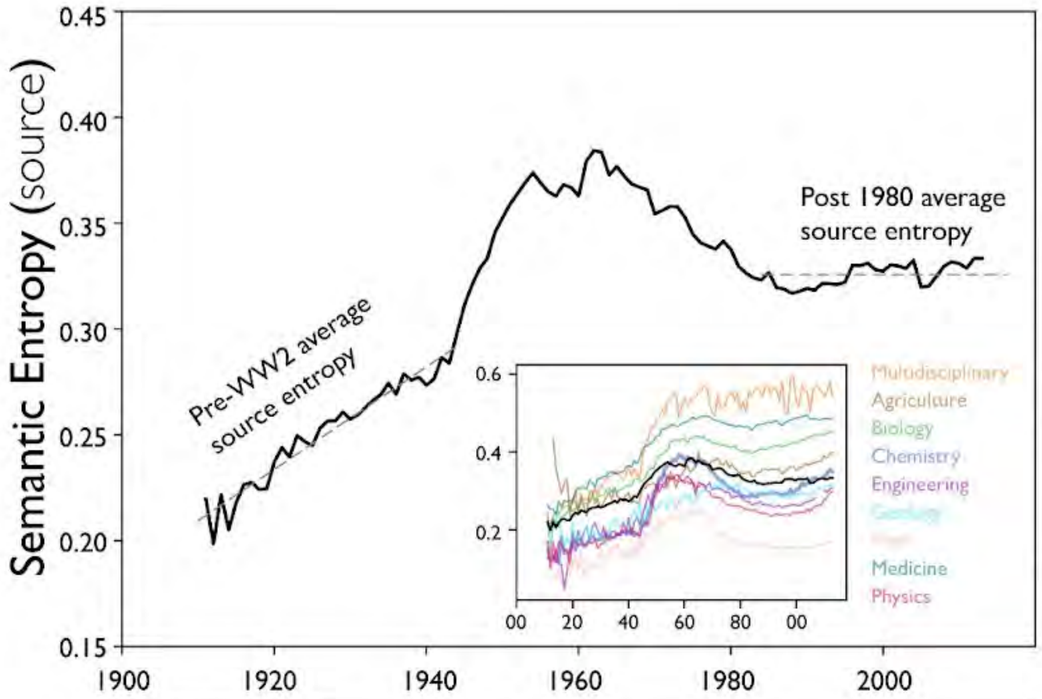

intellectual influence
3.0
2.9
2.8
2.7
2.6
2.5
2.4
SFI
All Fields
Engineering, Electrical & Electronic
Economics
Energy & Fuels
Physics, Applied
Pharmacology & Pharmacy
Nanoscience & Nanotechnology
History & Philosophy of Science
Immunology
Psychology
Telecommunications
Linguistics
Oncology
Endocrinology & Metabolism
Chemistry, Physical
Biochemistry & Molecular Biology
same department ∈ same institution \ same department ∈ same city \ same institution ∈ same country \ same city ∈ same world \ same country

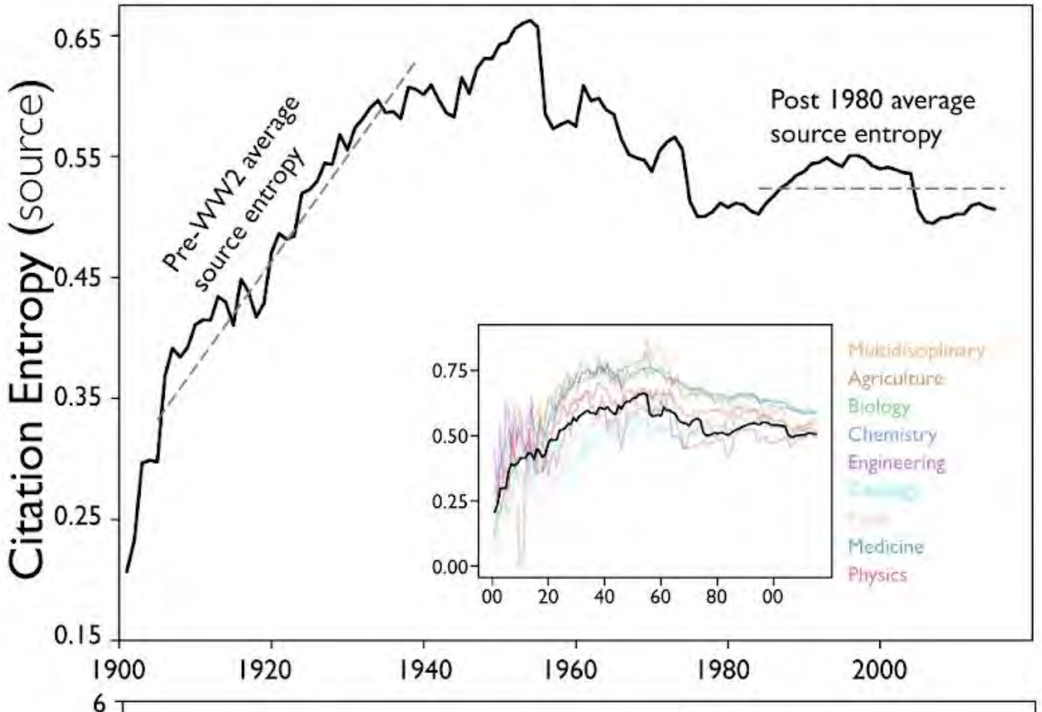
Broad reach of ***modern*** science





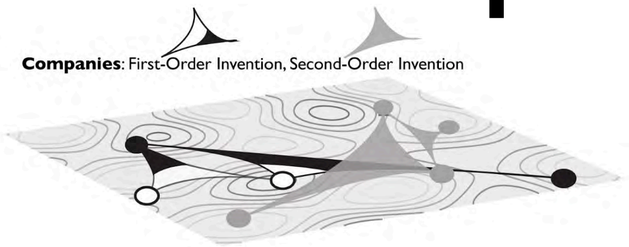


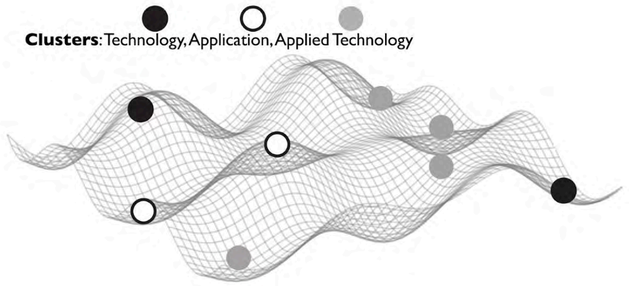


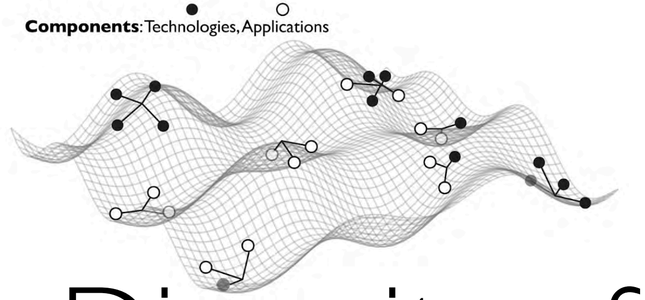


Venture-funded

Start-Ups

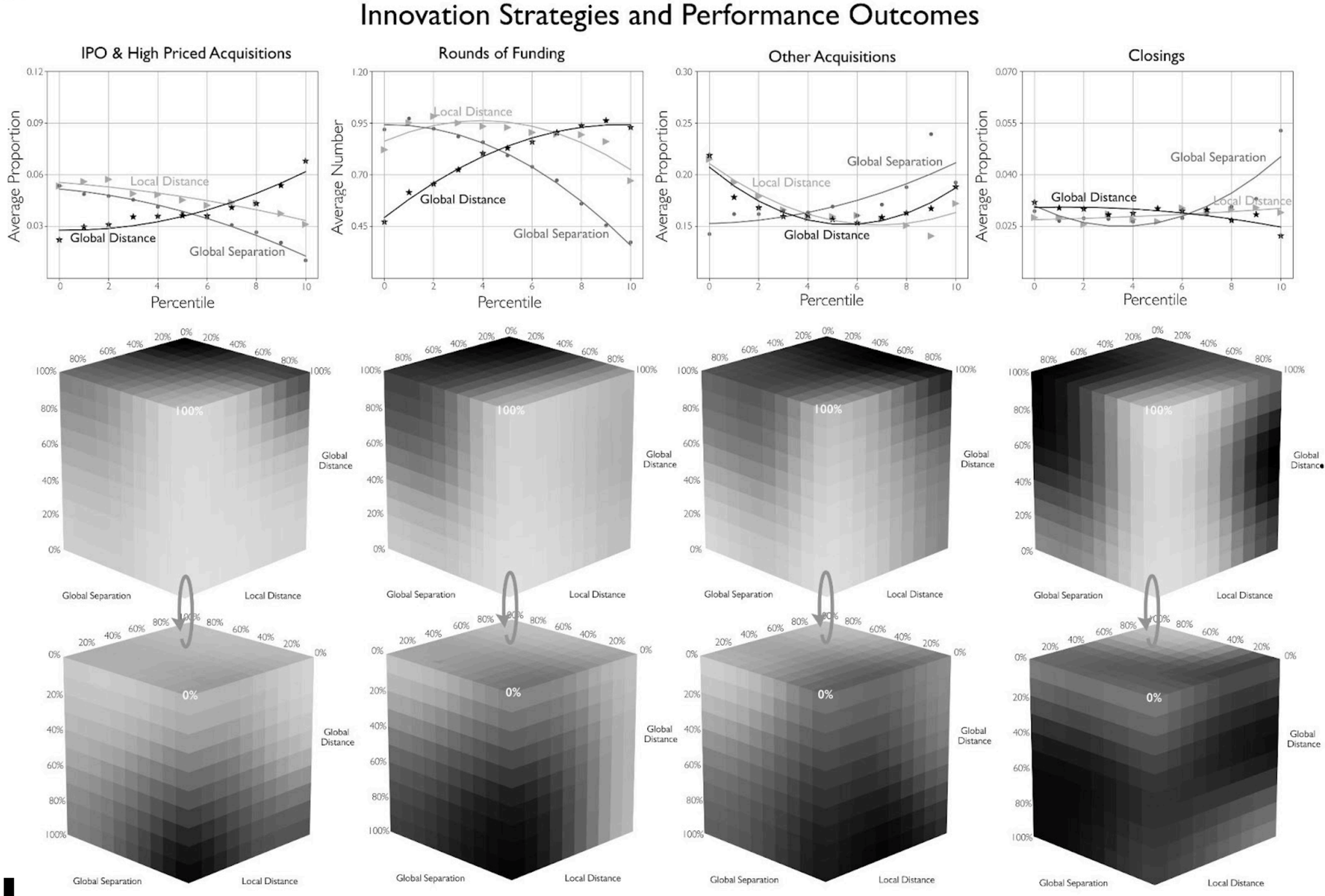






Diversity of

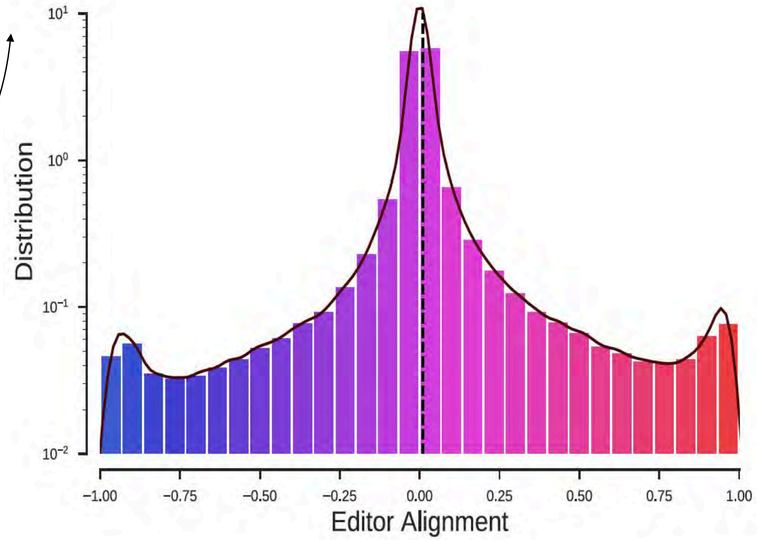
Applied Tech.



rapid invention

153

4



prior
update
posterior

$$P(p) = Beta(a, b)$$

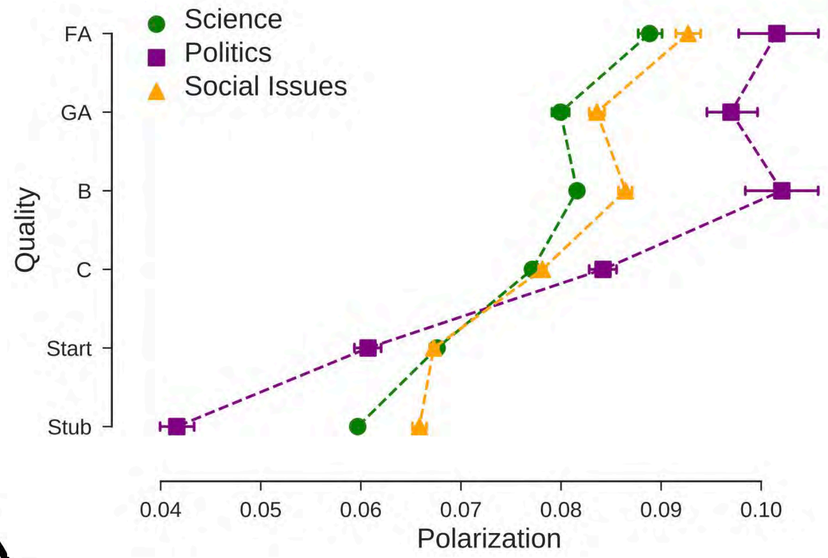
$$p|X \sim Beta(a + X, b + K - X)$$

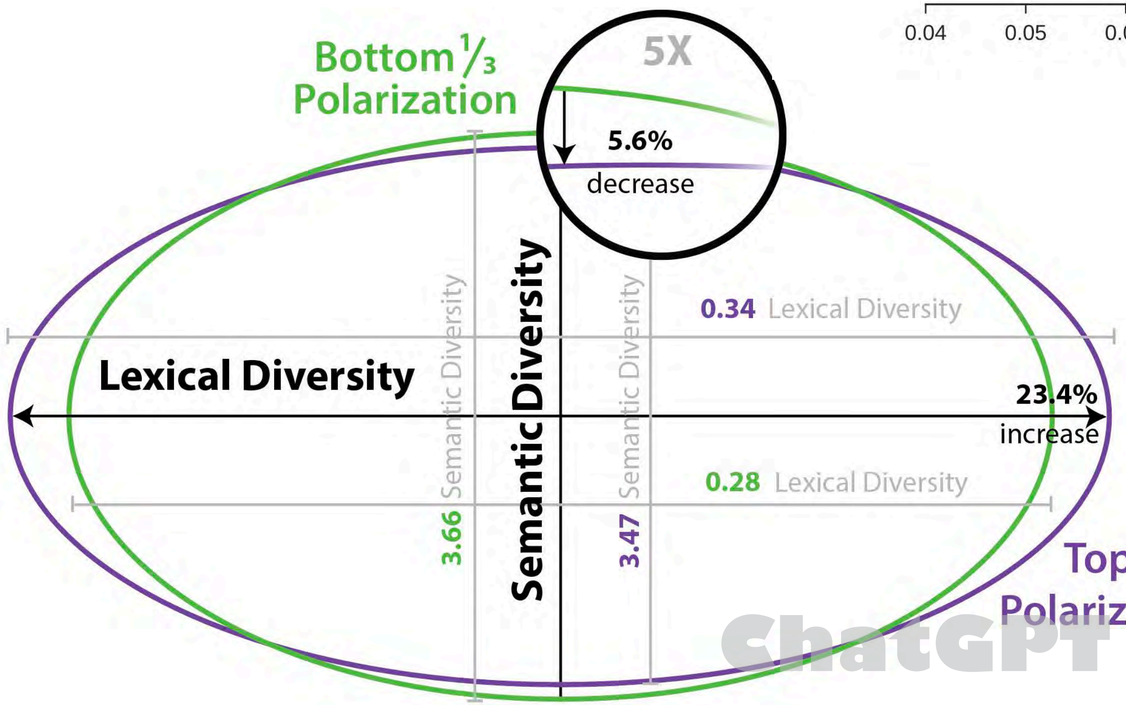
$$E[p|X] = (X + a)/(K + a + b)$$

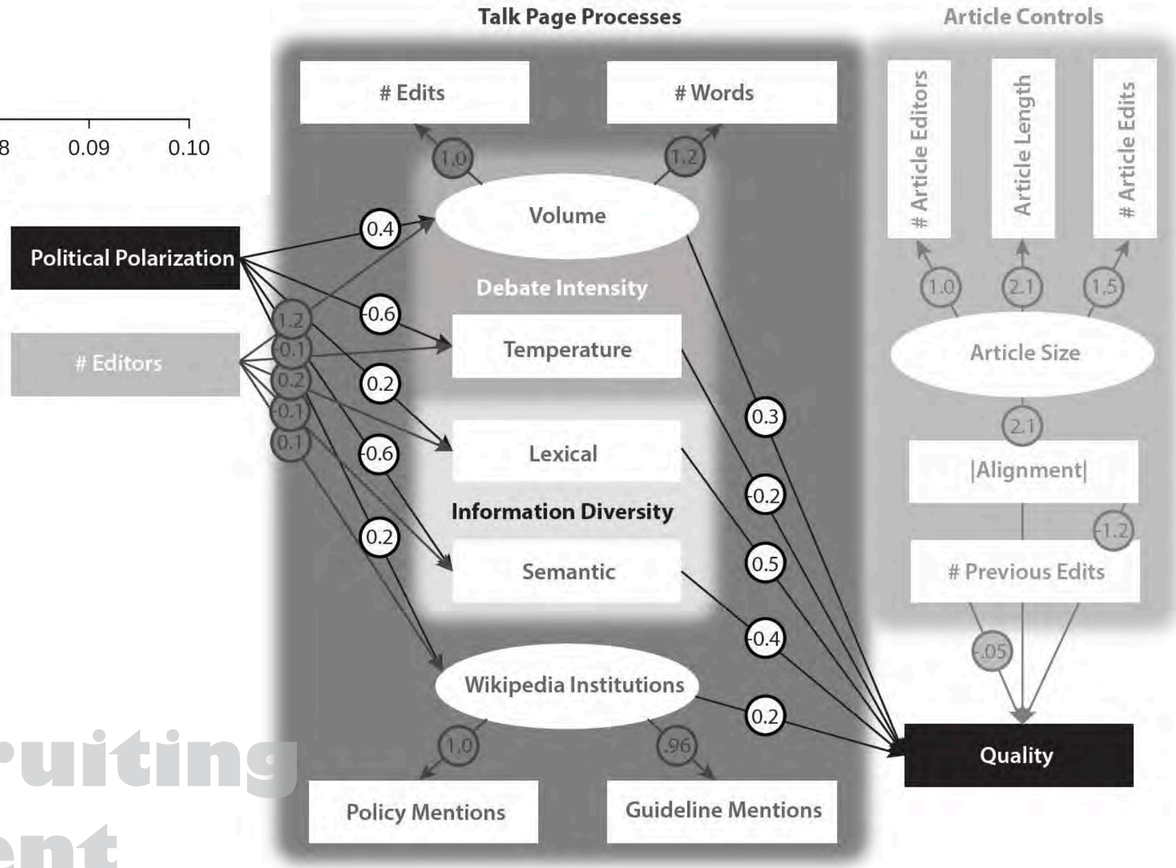
X = bytes to conservative pages; K= bytes to political pages

Improvement Through Polarized Reconnection

Encyclopedic Monopoly





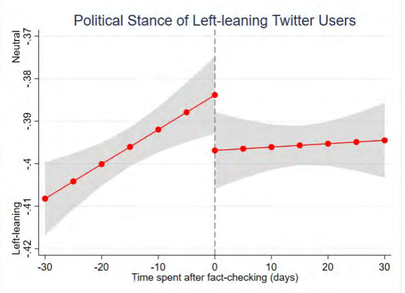


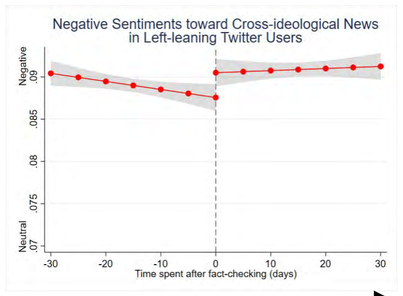
ChatGPT Recruiting Experiment

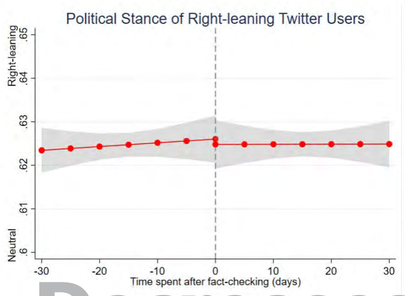
Misinformation & Fact-Checking

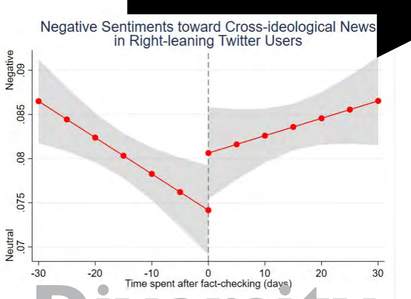
on Twitter

Vigilante Diversity

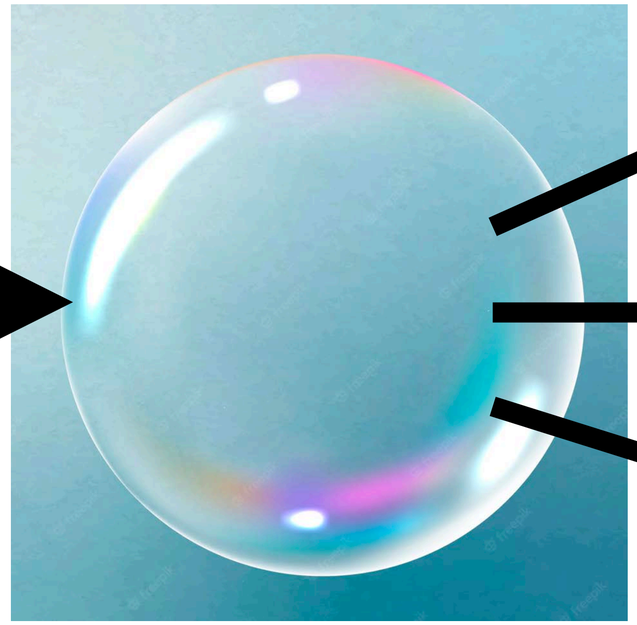






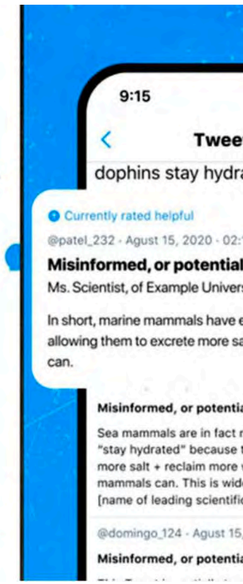


Decreases Diversity



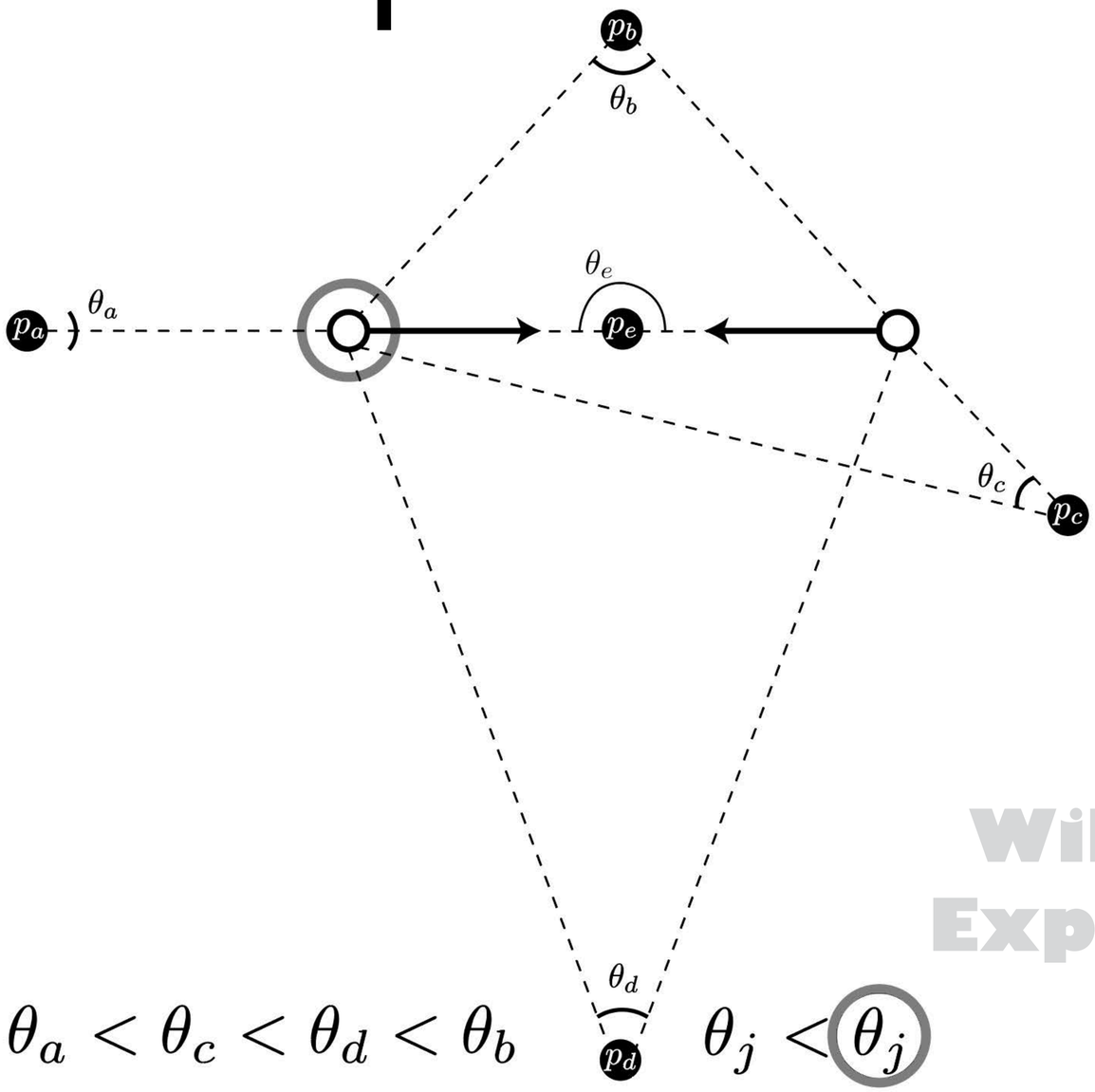
Balanced, Curated Diversity

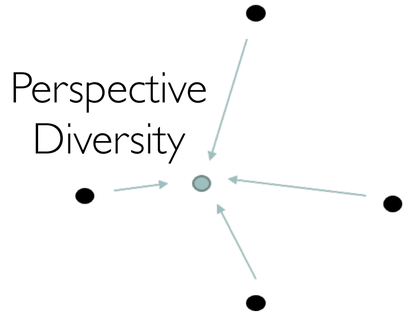
Welcome to Birdwatch

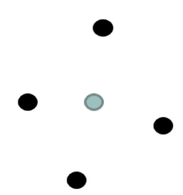


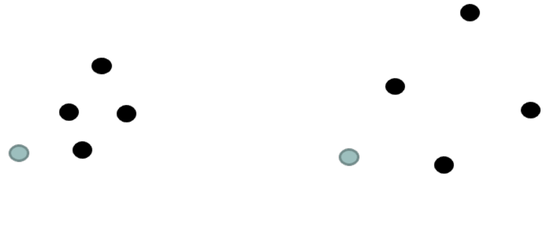
Increases Diversity

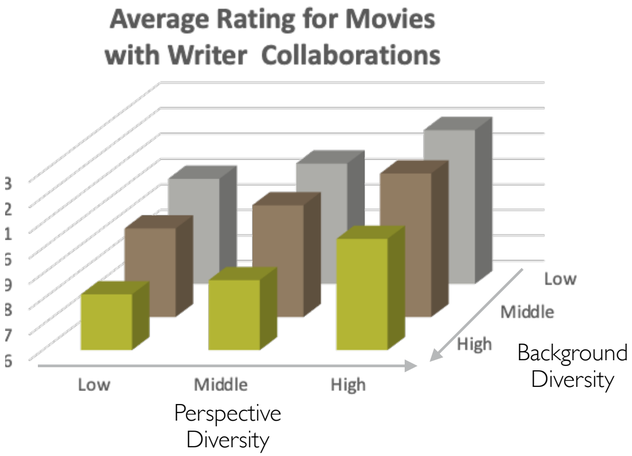
Diversity of **Perspectives**

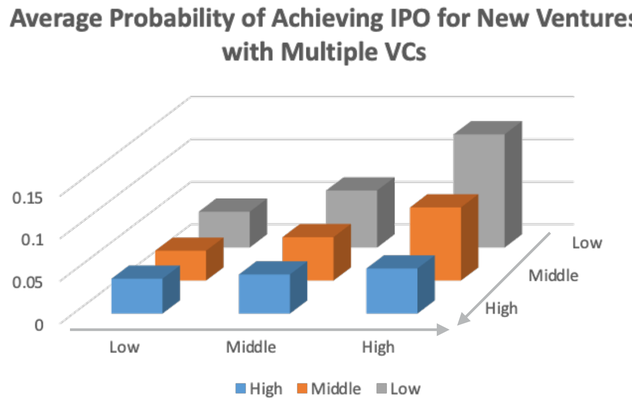












Wikipedia Experiment

Optimal **Diversity**

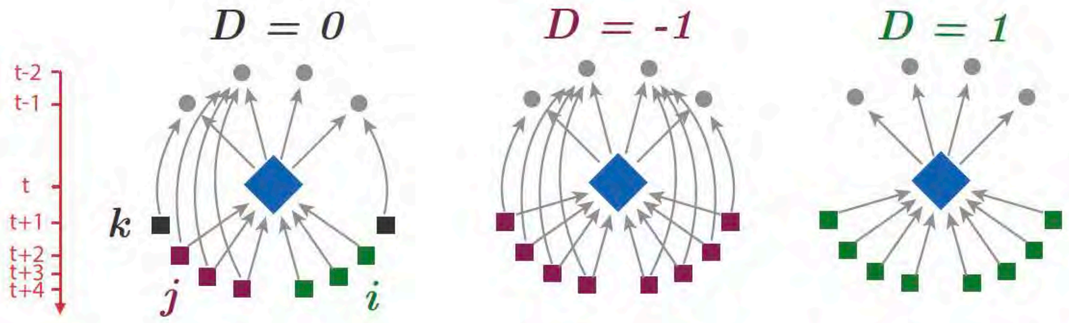
Danish Experiment

Across innovators, teams, and fields
Dense (Social) Connections
Collective Certainty
Novel Discovery/Invention
Distort
& Slow

Large, Tall, Smooth Teams Dampen Disruption

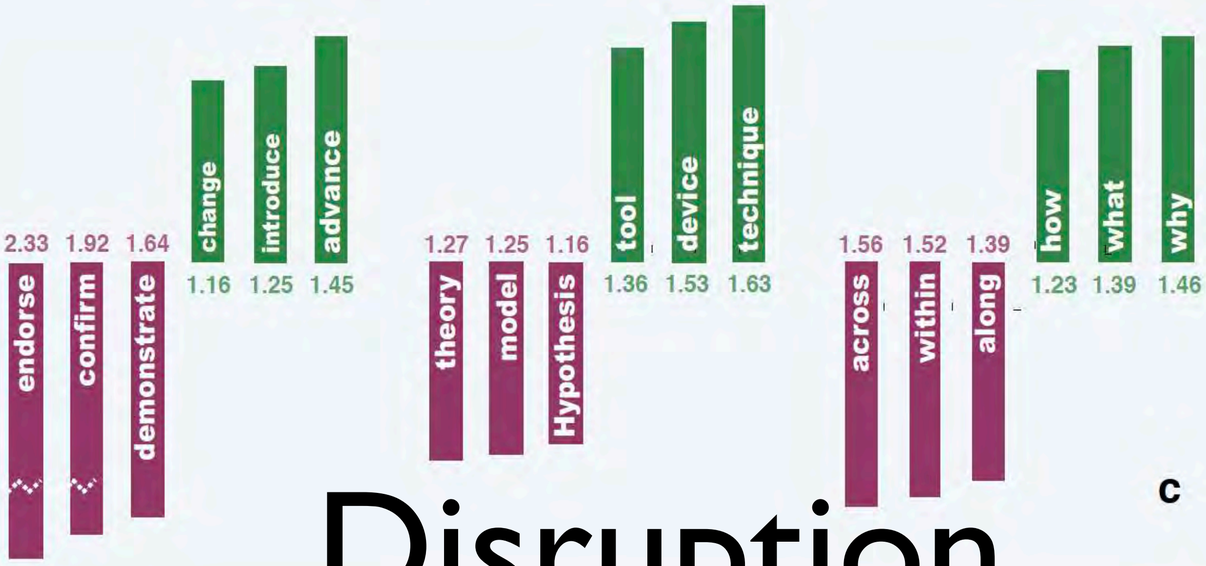
Small, Flat, Sparse Teams Amplify It

Disruption

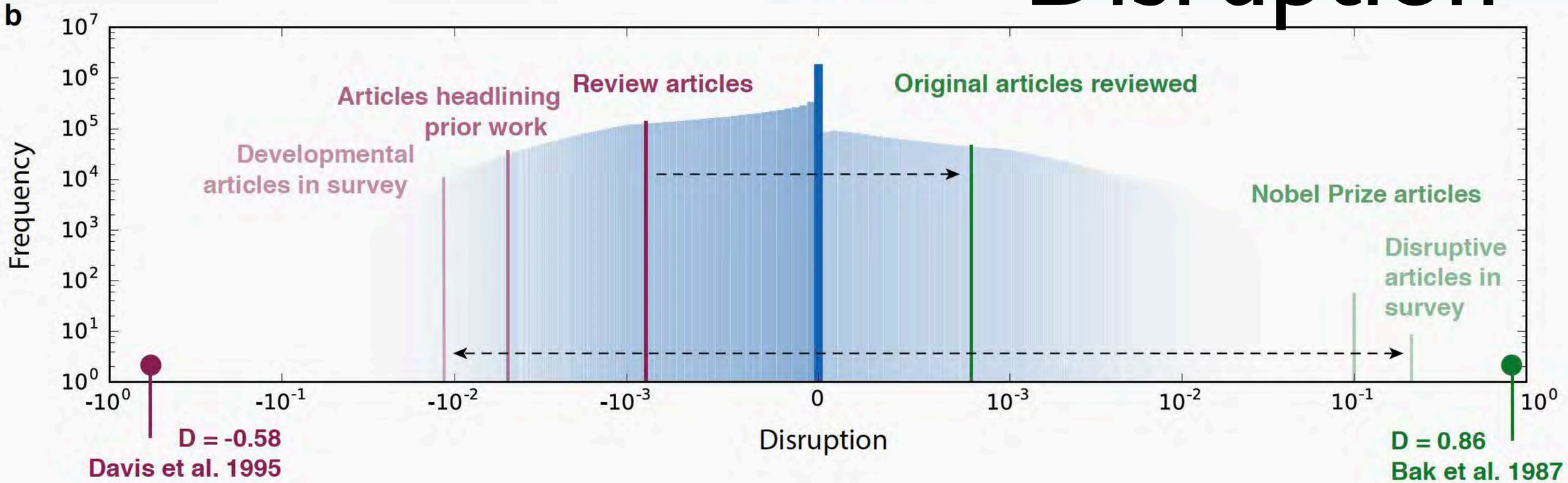


$$Disruption:\ D = p_i - p_j = \frac{N_i - N_j}{N_i + N_j + N_k}$$

a



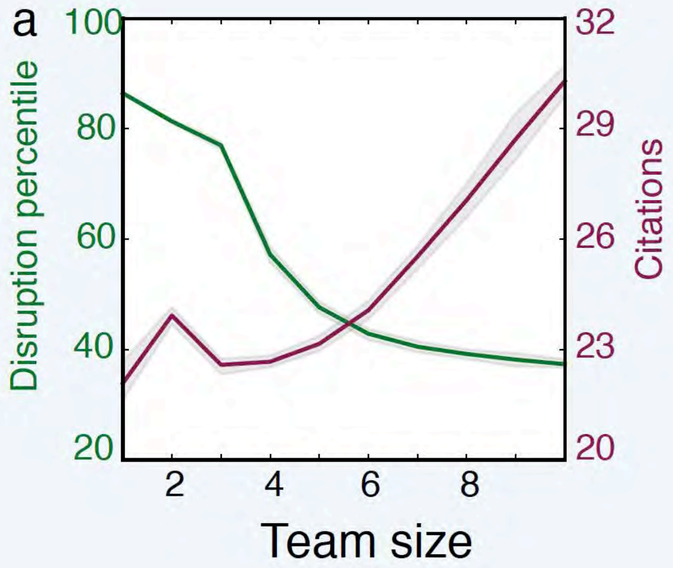
Disruption distribution of articles

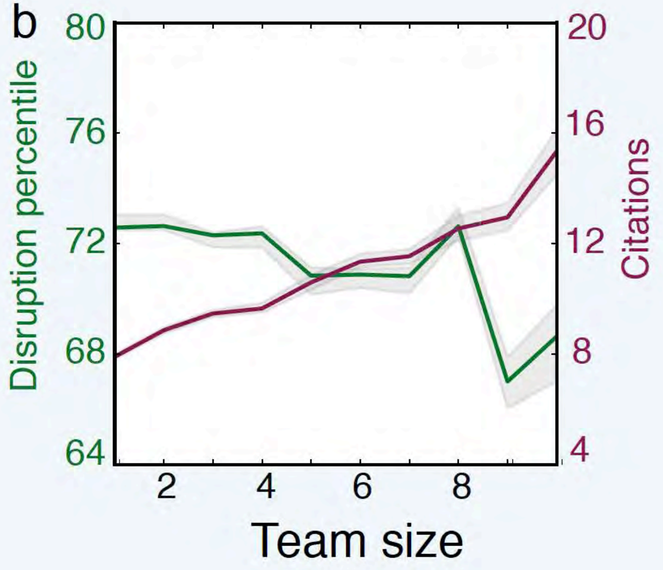


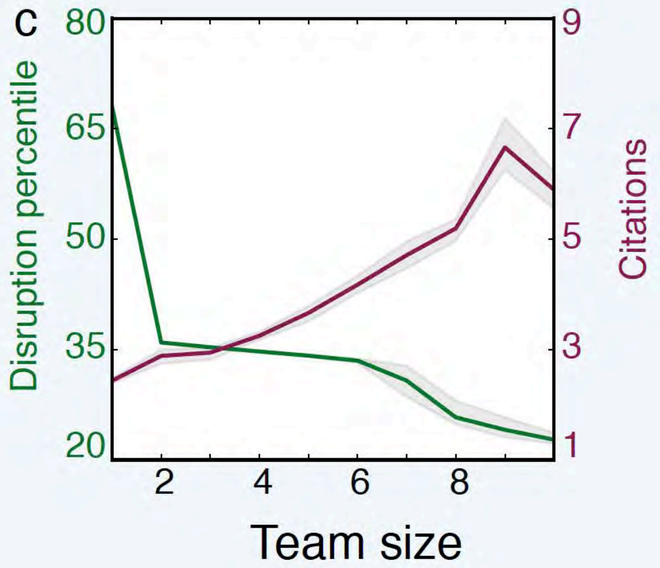
Articles

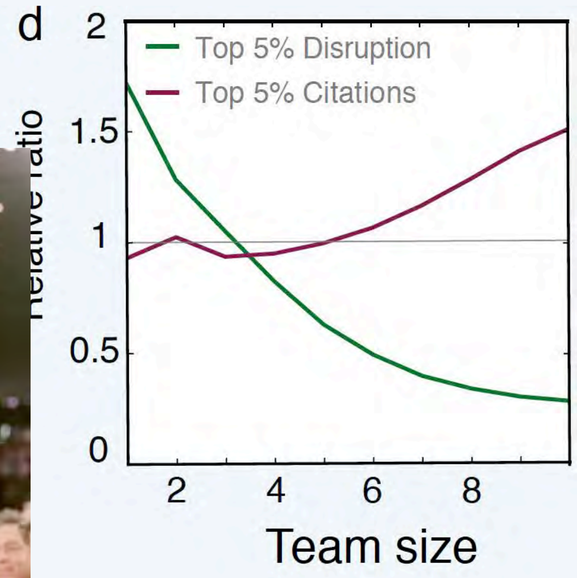
Patents

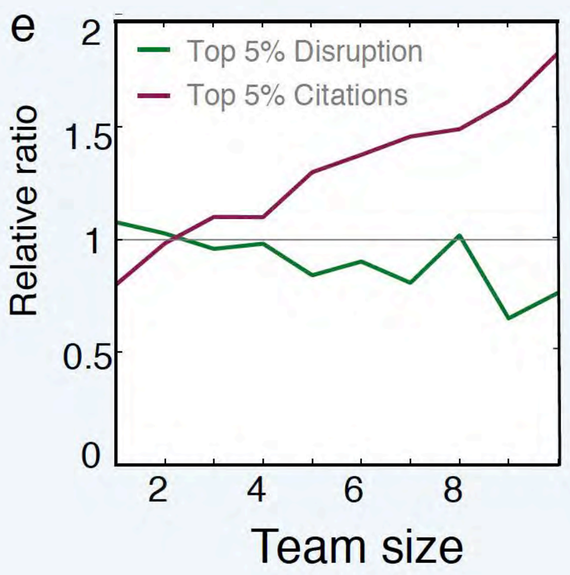
Software

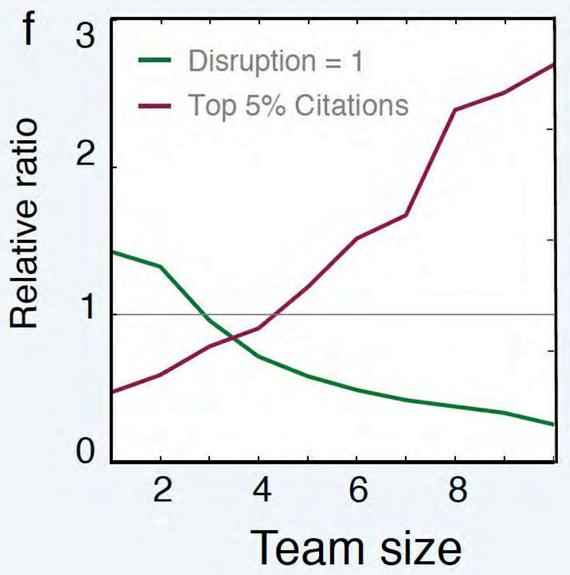




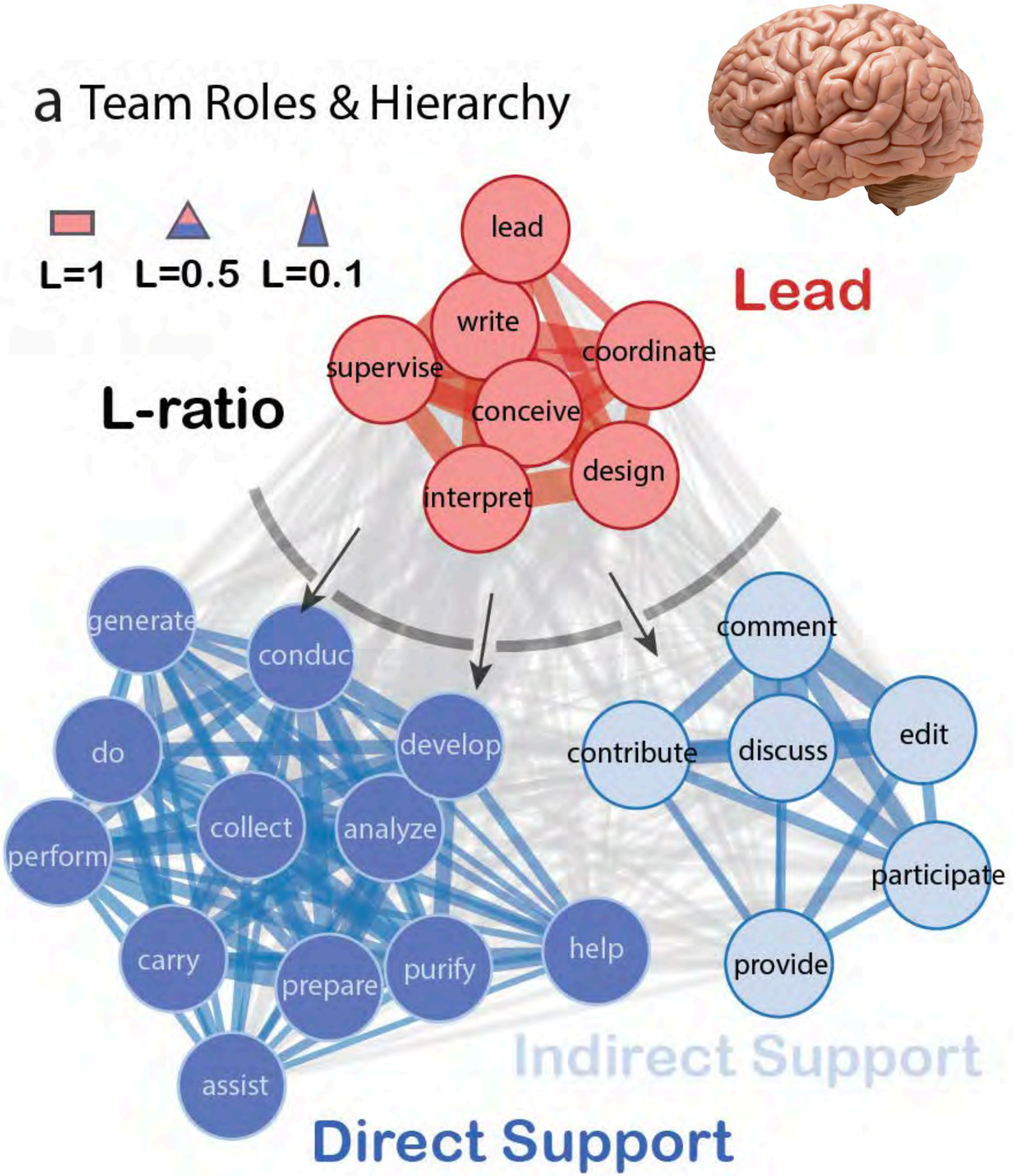




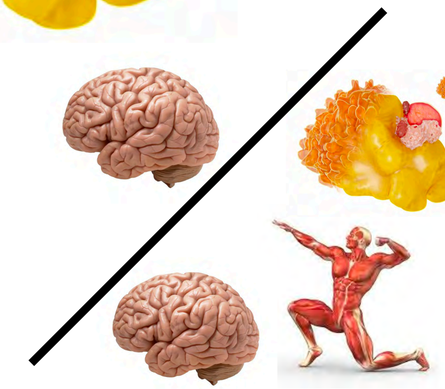




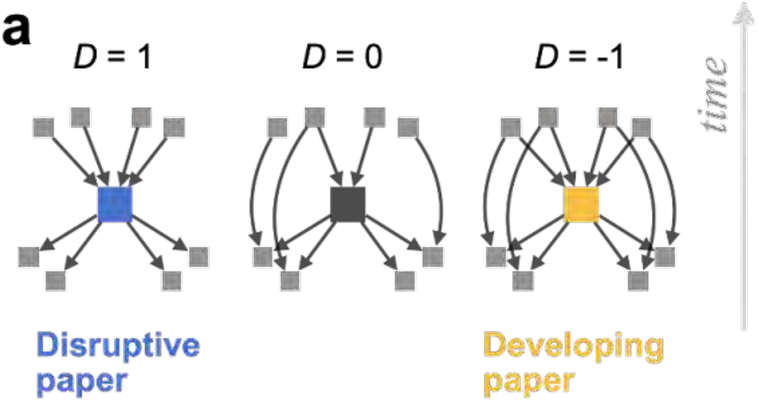
Team Structure

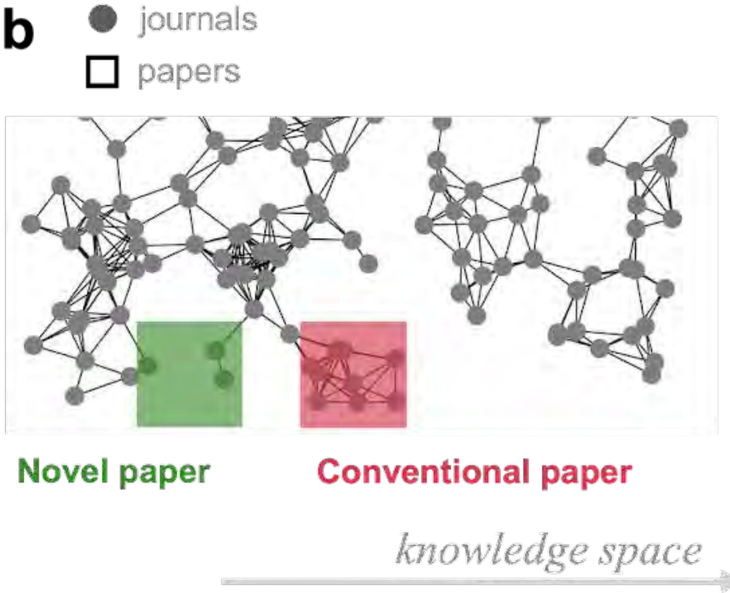


89,575



Team Structure Matters for Innovation

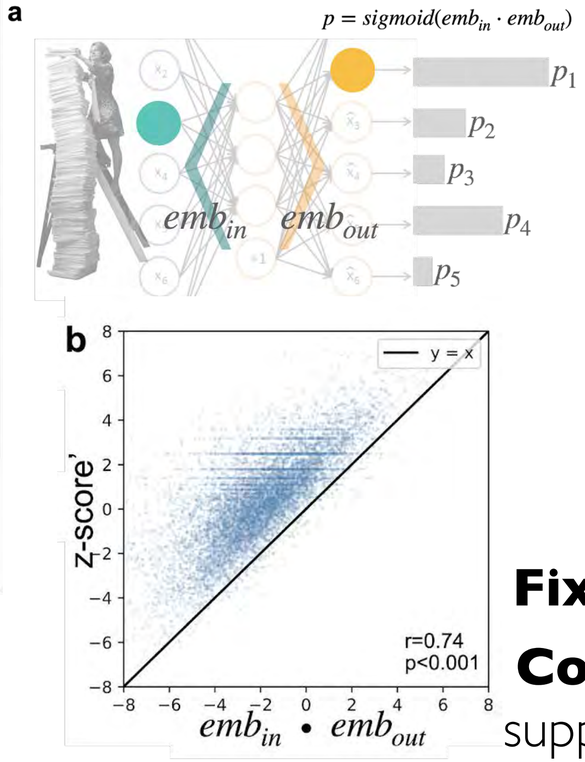


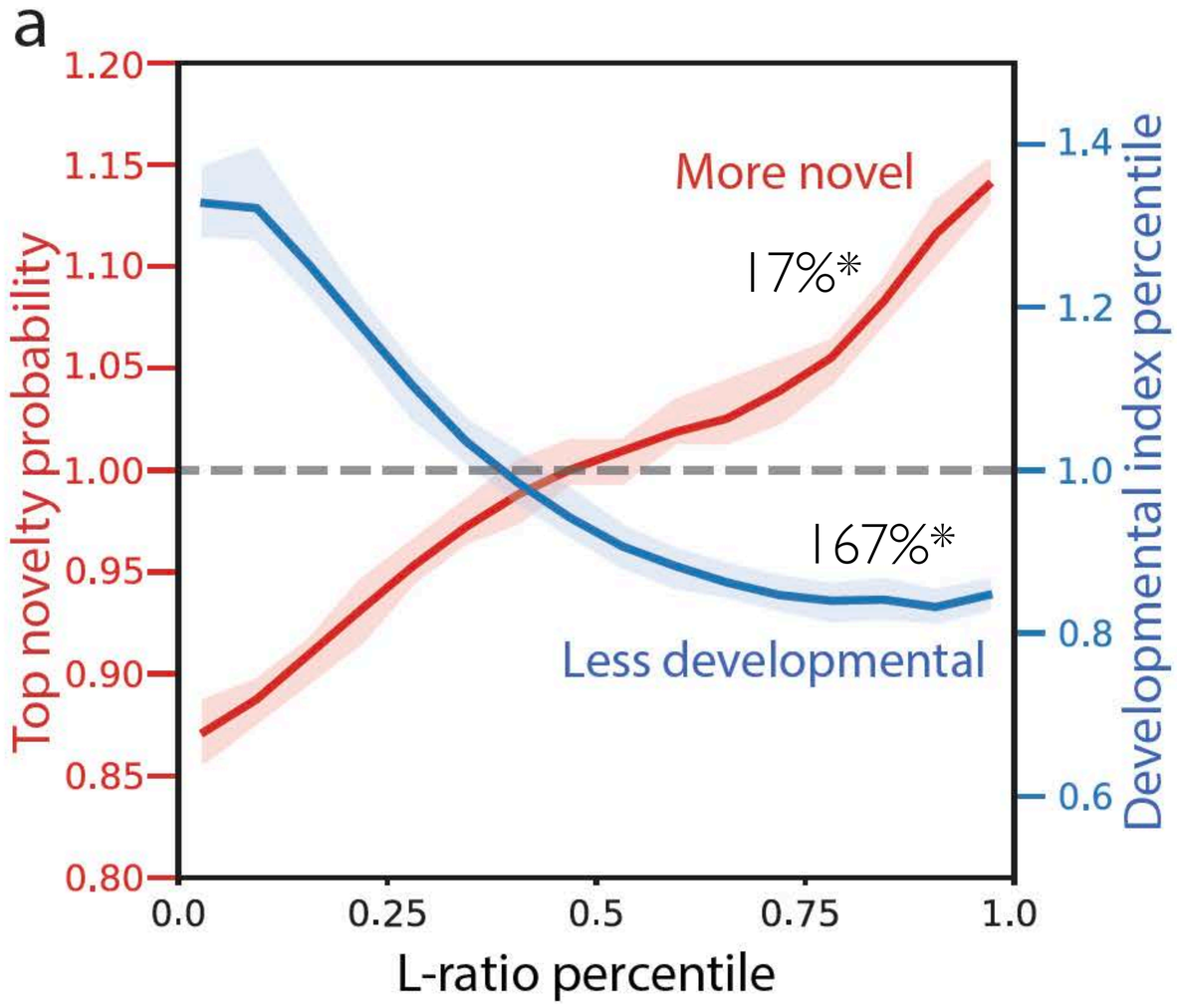


$$\textit{Disruption: } D = p_i - p_j = \frac{N_i - N_j}{N_i + N_j + N_k}$$

$$\textit{Novelty: } z_{ij} = \frac{obs_{ij} - exp_{ij}}{\sigma_{ij}} \quad MCMC$$

Pointwise mutual information (PMI) between two items = inner product of two vectors representing items within a latent semantic space

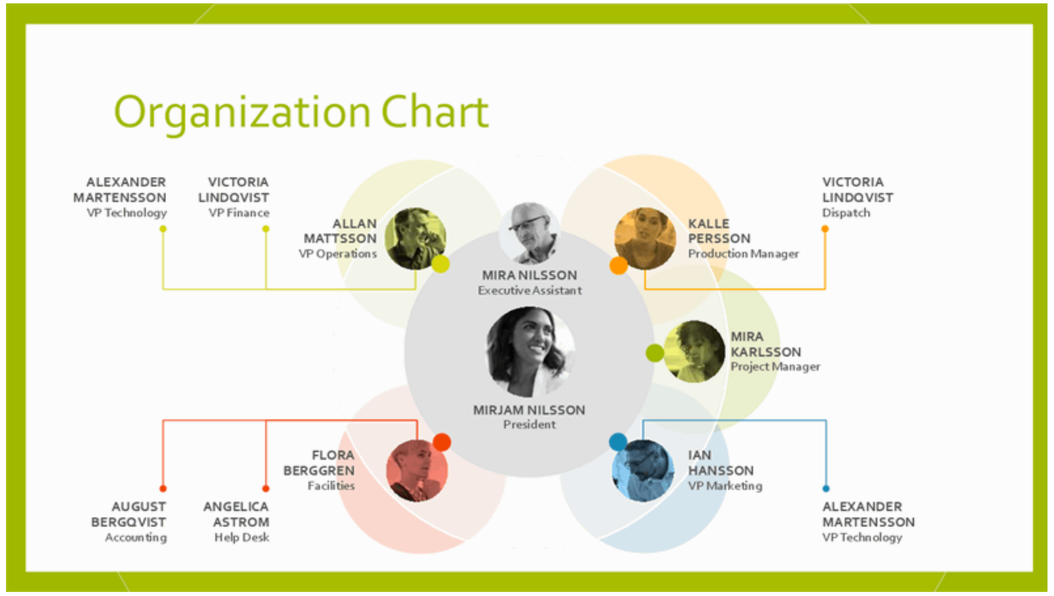


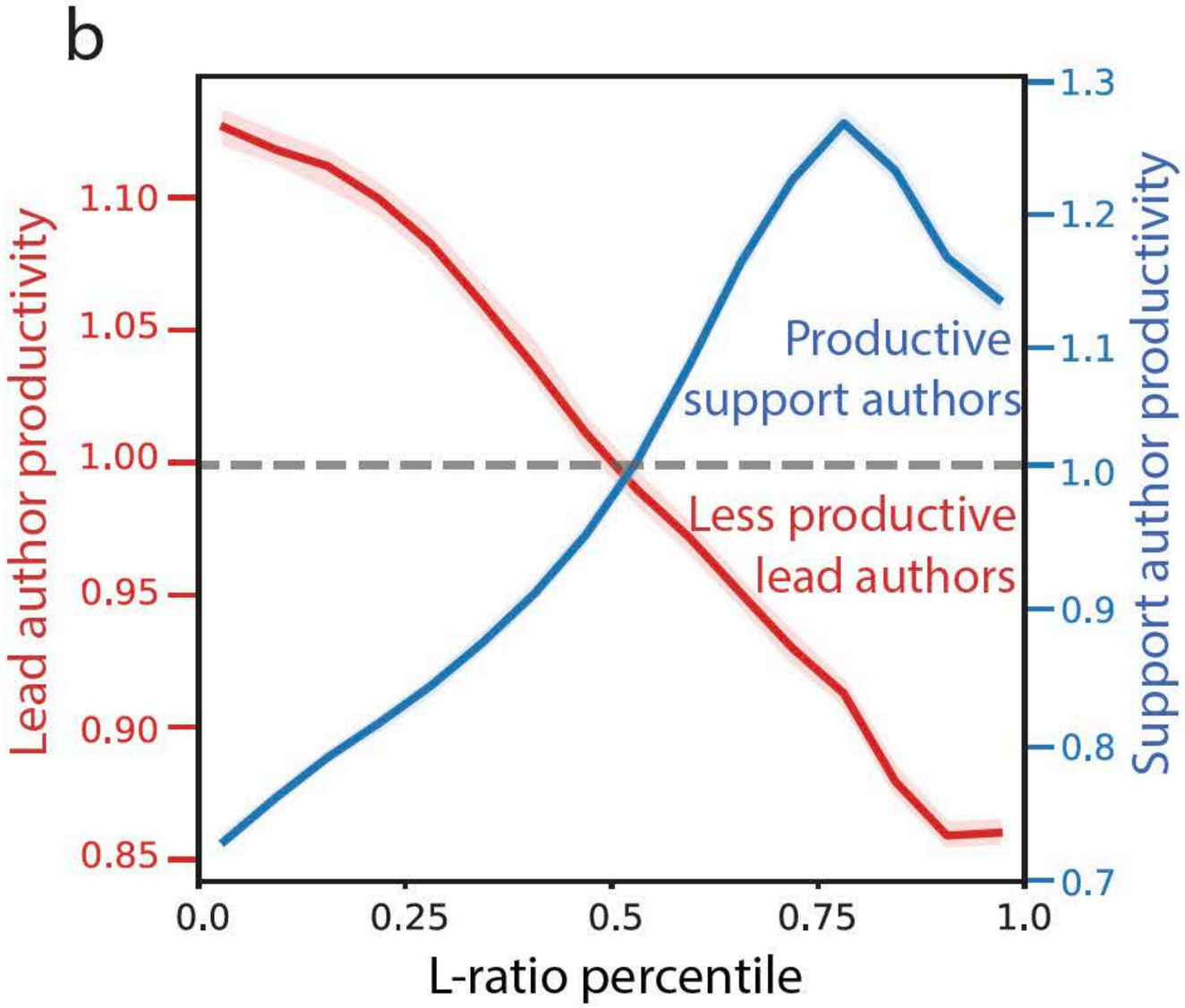


Fixed Effects by author & by field

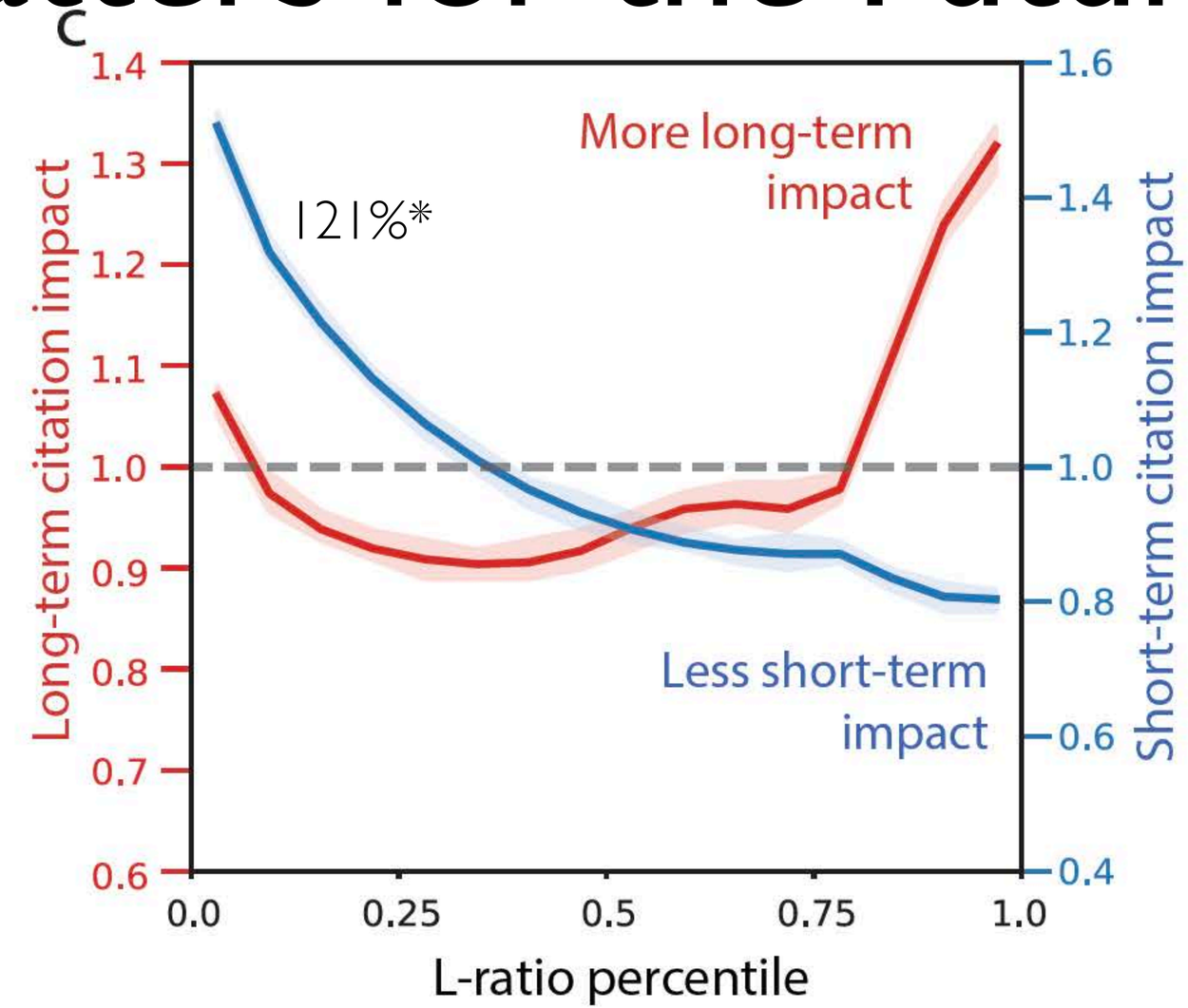
Controlling for: team size, mean, stdev & max career age, supported by funding agencies, # & award amount of funding

Team Structure Matters for Credit





Team Structure Matters for the Future



Fixed Effects by author & by field

Controlling for: team size, mean, stdev & max career age, supported by funding agencies, # & award amount of funding

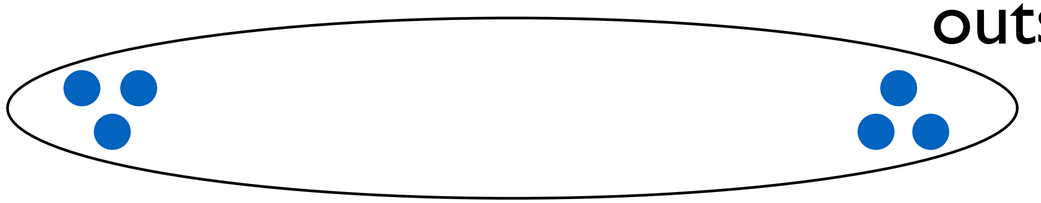
Activating Search

Collective Composition Matters

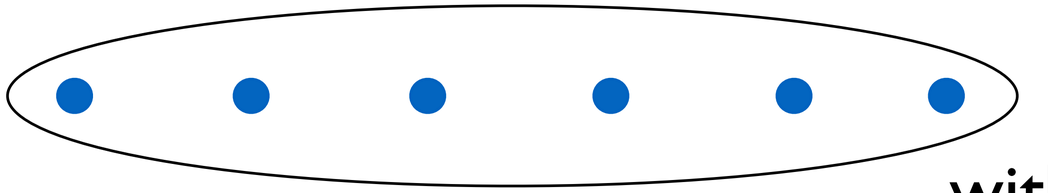
Concept/Tech **Diameter** matters…

But so do **Holes** in the structure

Diameter = Largest Hole



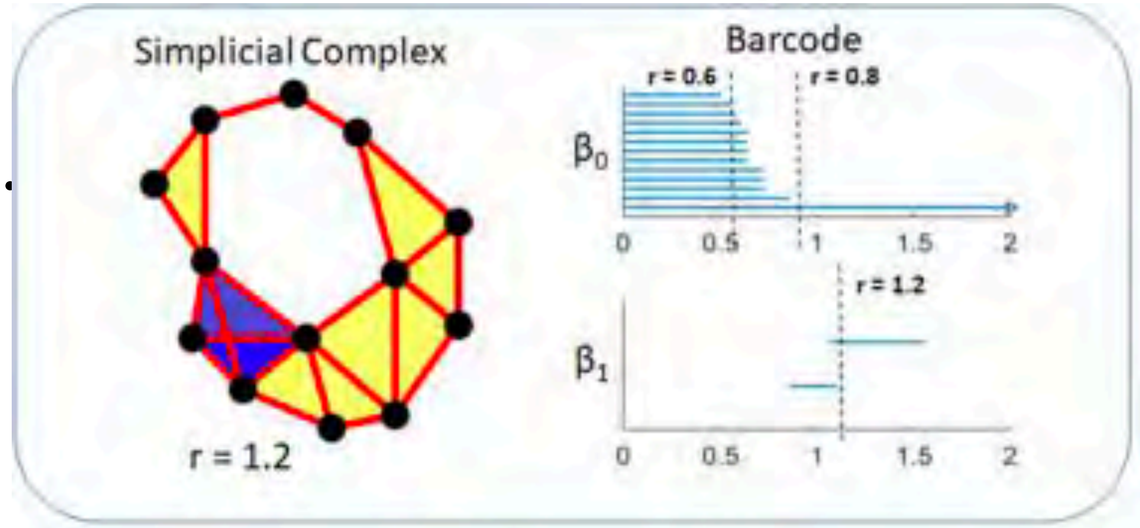
outside

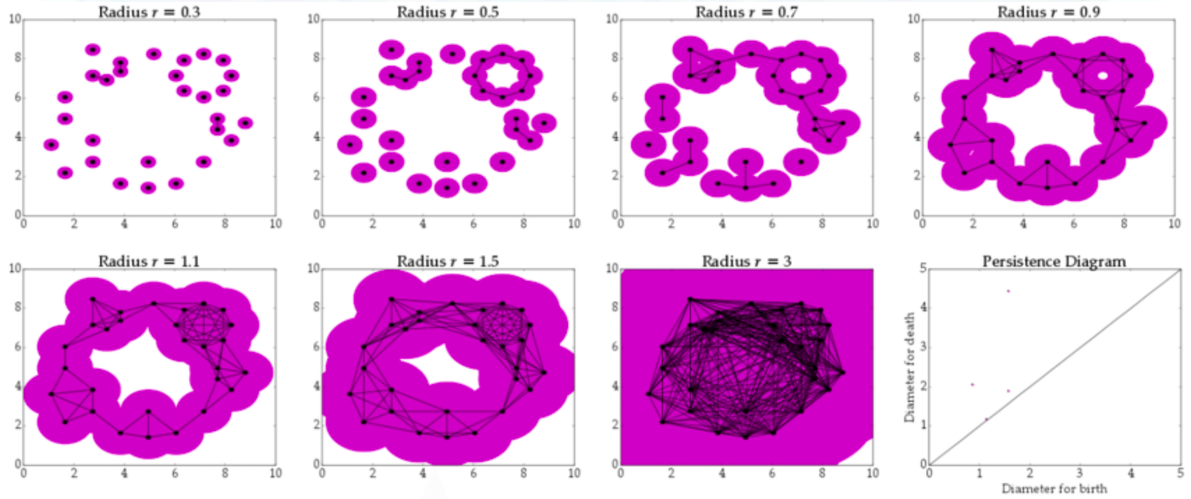


within

Diameter = Largest Hole * N

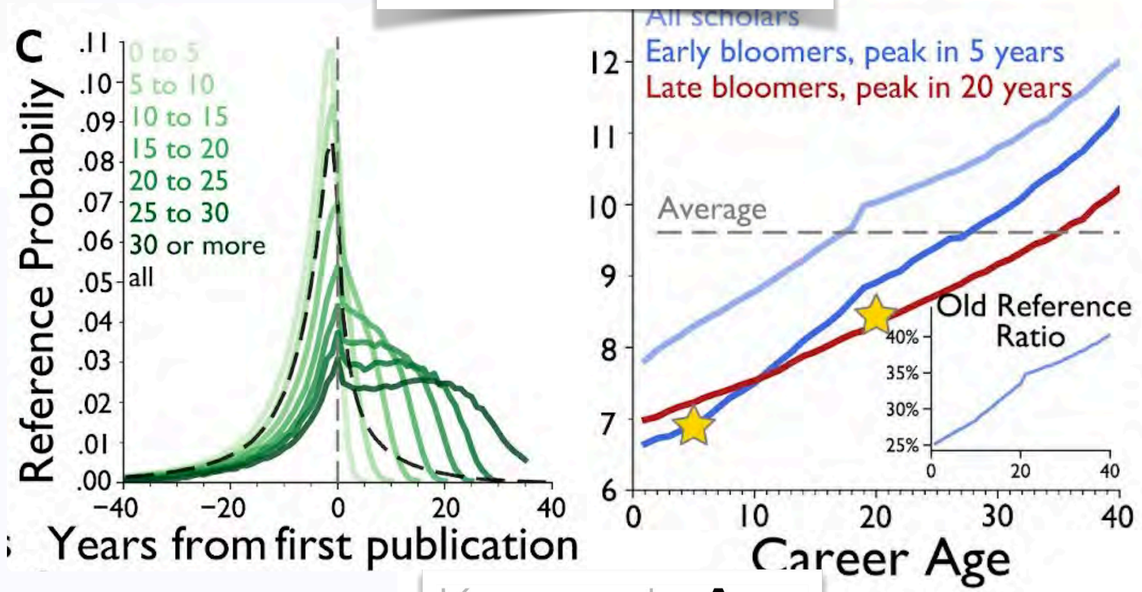
Persistent Homology



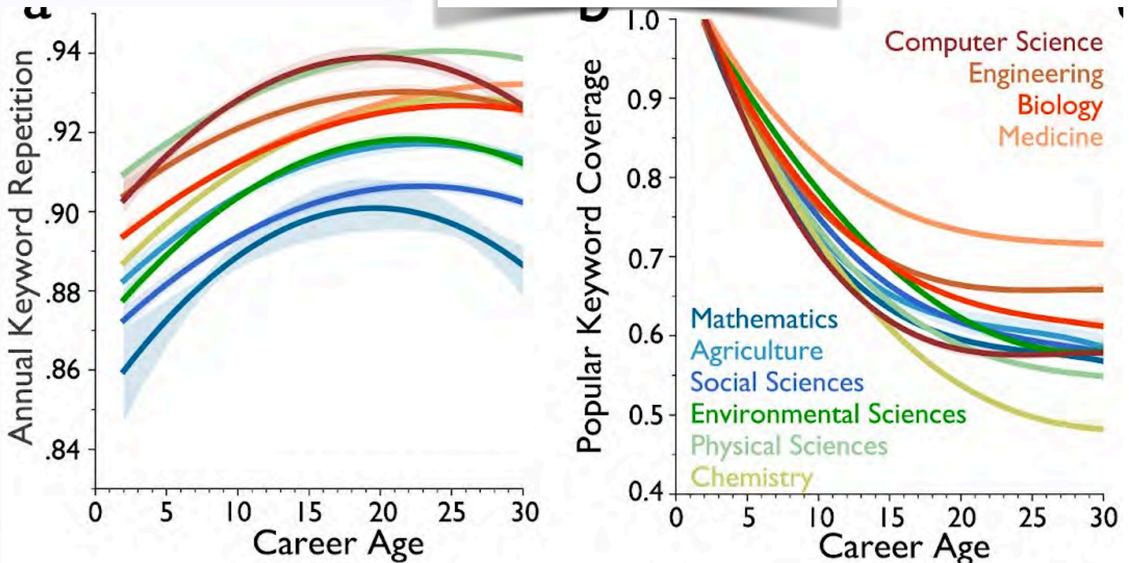


Aging & Innovation

References **Age**

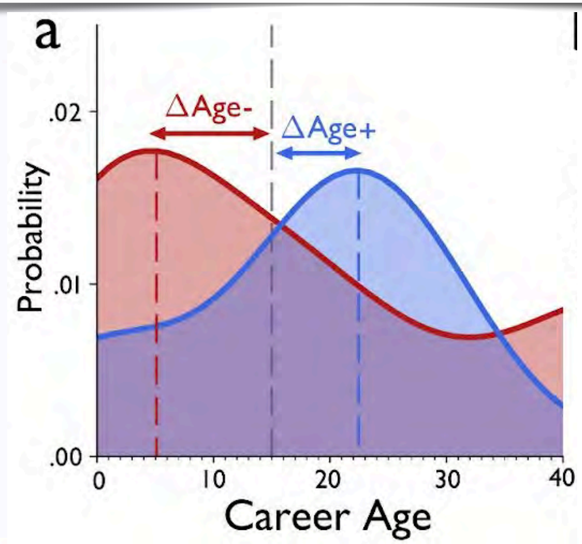


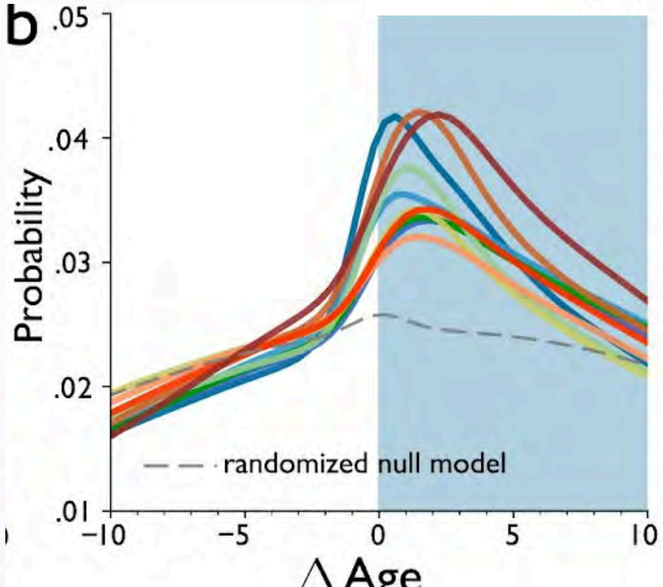
Keywords **Age**

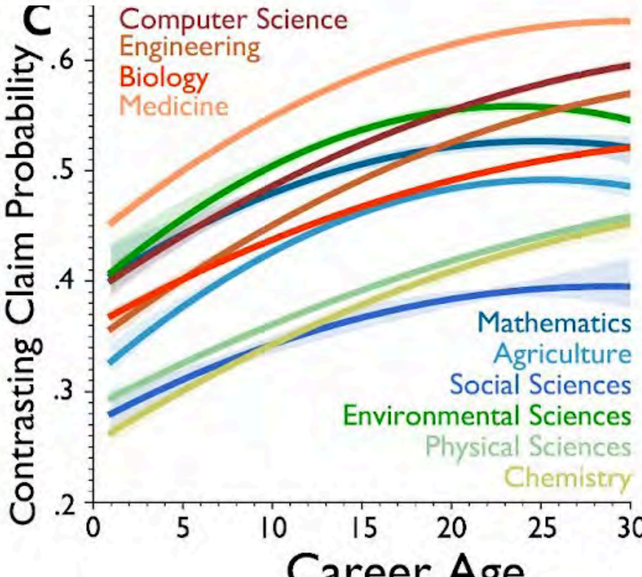


Aging & Innovation

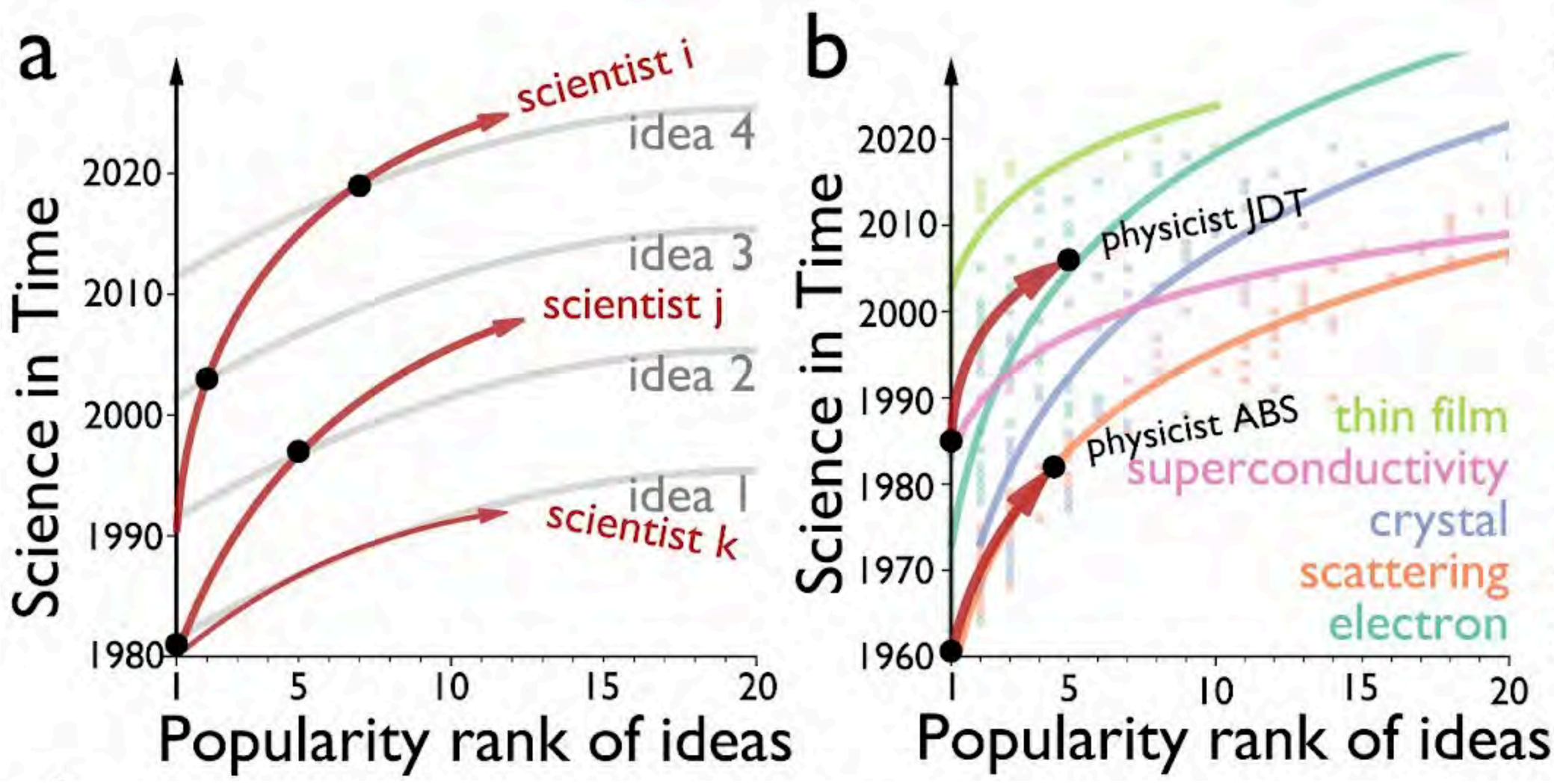
Criticism **Accelerates**







Scientists, their best **Ideas**, and the unfolding **Frontier**



Fields Age

Level-1

$$RefAge = \pi_0 + \pi_1 Age + \pi_2 TS + e$$

Level-2

$$\pi_0 = \beta_{00} + r_0$$
$$\pi_1 = \beta_{10} + r_1$$
$$\pi_2 = \beta_{20} + r_2$$

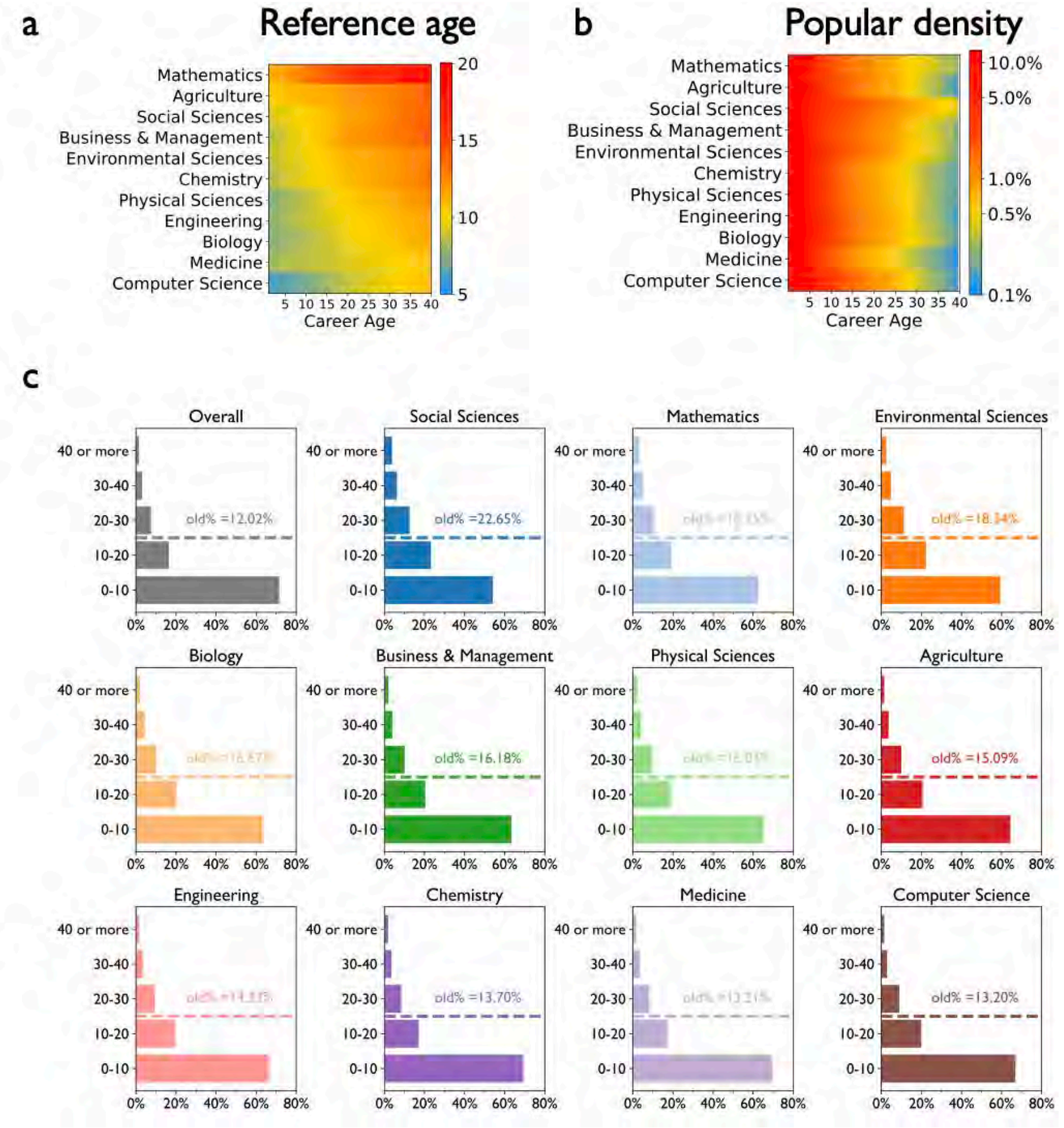
Level-3

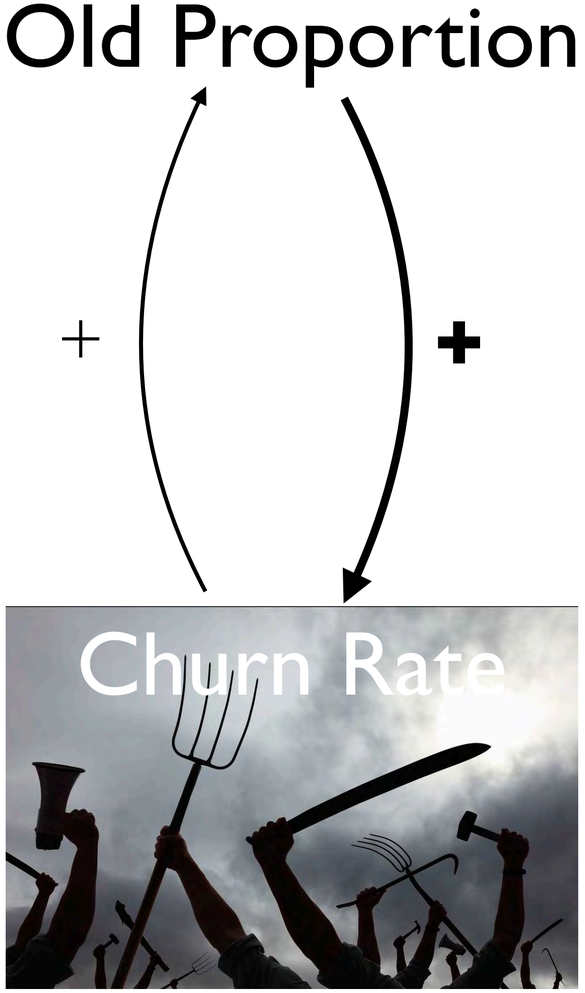
$$\beta_{00} = r_{000} + r_{001}(MeanAge) + r_{002}(MeanTS) + u_{00}$$
$$\beta_{10} = r_{100} + r_{101}(MeanAge) + r_{102}(MeanTS) + u_{10}$$
$$\beta_{20} = r_{200} + u_{20}$$

Scientist age (+) **reference age**

Mean age (+) **reference age**

holding constant mean age within field and mean team size.

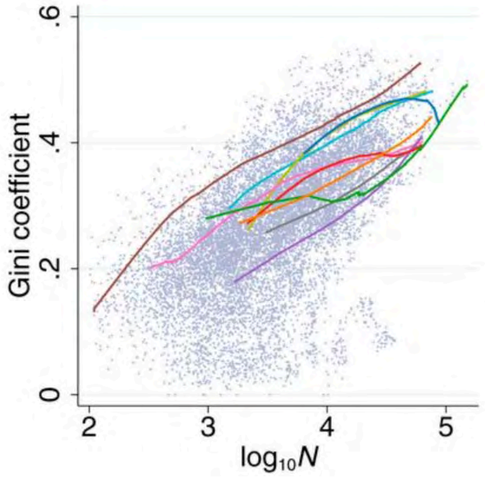


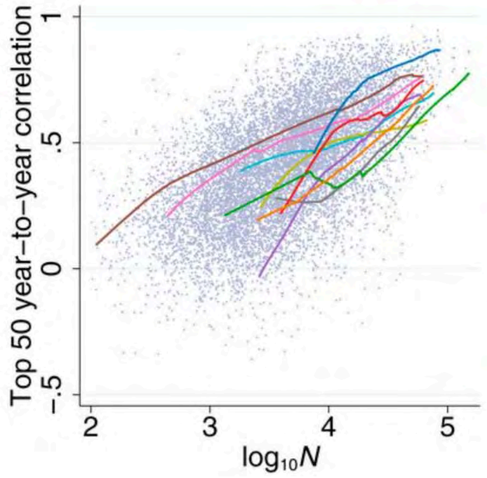


Independent Variables	Old Proportion Predicts Churn Rate		Churn Rate Predicts Old Proportion	
	SSR *F*-test		SSR *F*-test	
Field	**lags=1 year**	**lags=2 year**	**lags=1 year**	**lags=2 year**
Mathematics	9.67***	4.76**	8.17***	5.55***
Geology	16.76***	7.74***	0.36	4.69**
Economics	11.45***	9.16***	0.61	1.59
Psychology	18.66***	6.65***	0.32	0.96
Materials science	14.29***	4.34**	5.49**	3.65**
Physics	9.08***	2.19	4.05**	2.27
Chemistry	11.09***	2.45*	1.29	0.52
Biology	13.82***	4.01**	17.07***	8.31***
Medicine	17.27***	5.07***	0.1	0.57
Engineering	16.38***	4.98**	2.84*	1.8
Computer science	30.53***	13.33***	4.19**	0.75

* p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

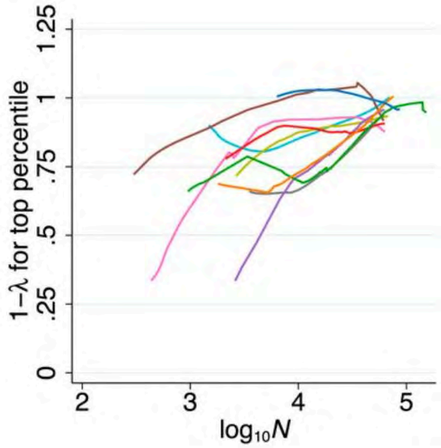
PNAS
Proceedings of the National Academy of Sciences of the United States of America

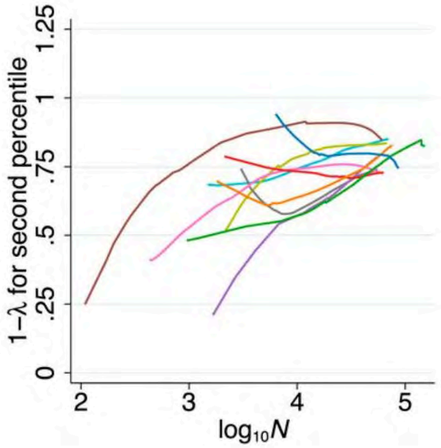


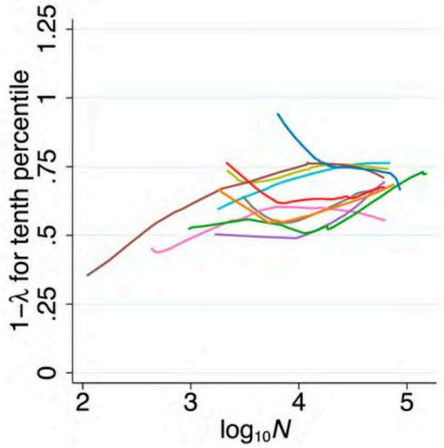


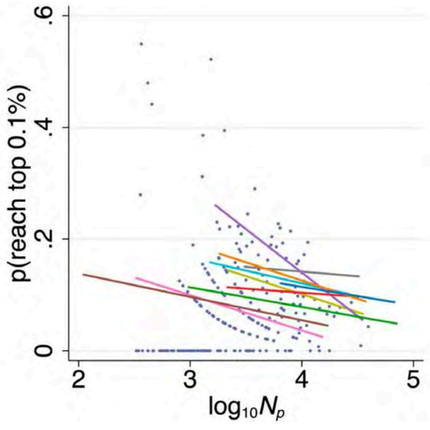
As Fields Grow

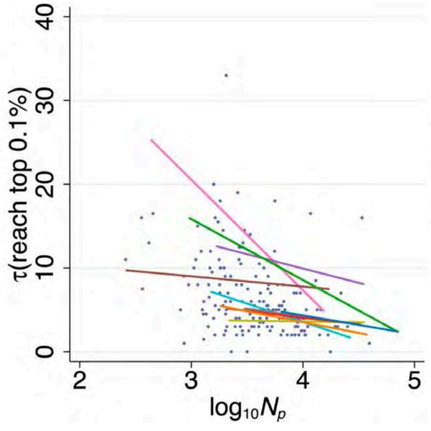
Canons Crystallize

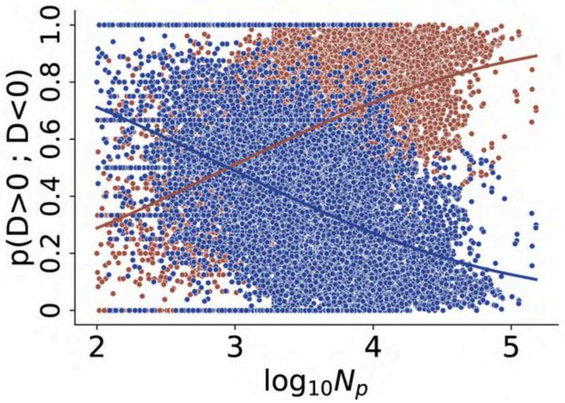


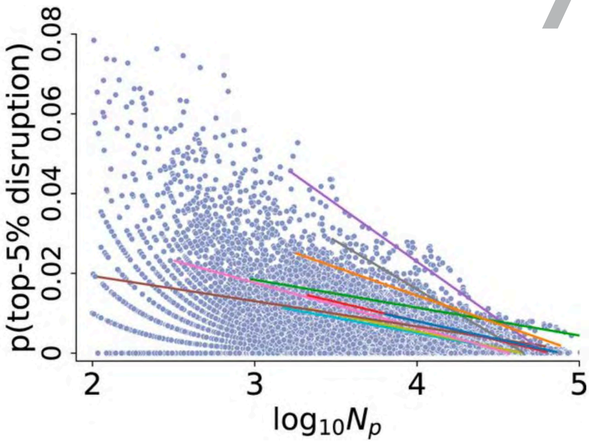












Diversity Collapse
In other Domains

- Cultural objects and the Interwebz
- Language extinction with social contact
- Biological extinction with ecological contact

(NOT division of labor, and increasing specialization as in cities.)

Centralized communities & the Replication Crisis

Stem Cell Research Papers Are Retracted

By ANDREW POLLACK JULY 2, 2014

EMAIL

Two scientific papers that initially electrified

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

News & Comment > News > 2014 > October > Article

NATURE | NEWS

Psychologists strike a blow for reproducibility

Thirty-six labs collaborate to check 13 earlier findings.

Ed Yong

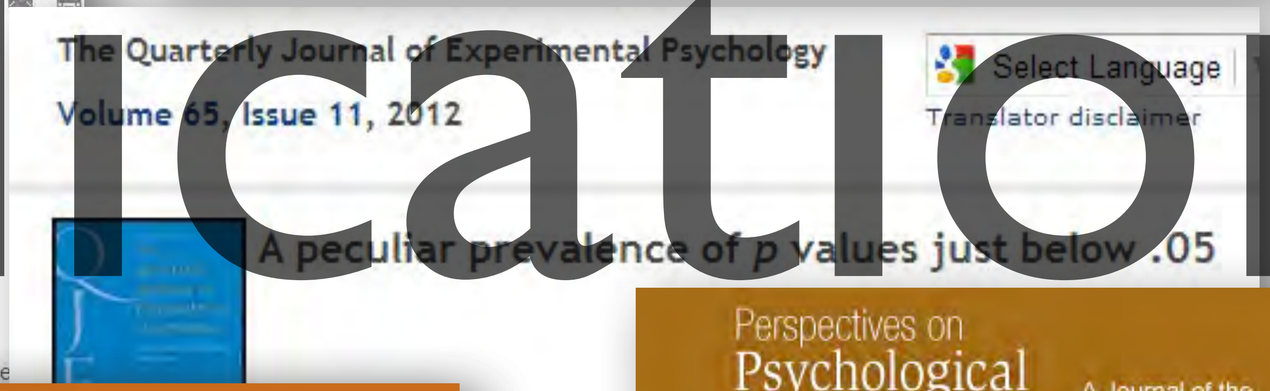
26 November 2013

Rights & Permissions

A large international group set up to test the reliability of psychology experiments has succe

repr

effec



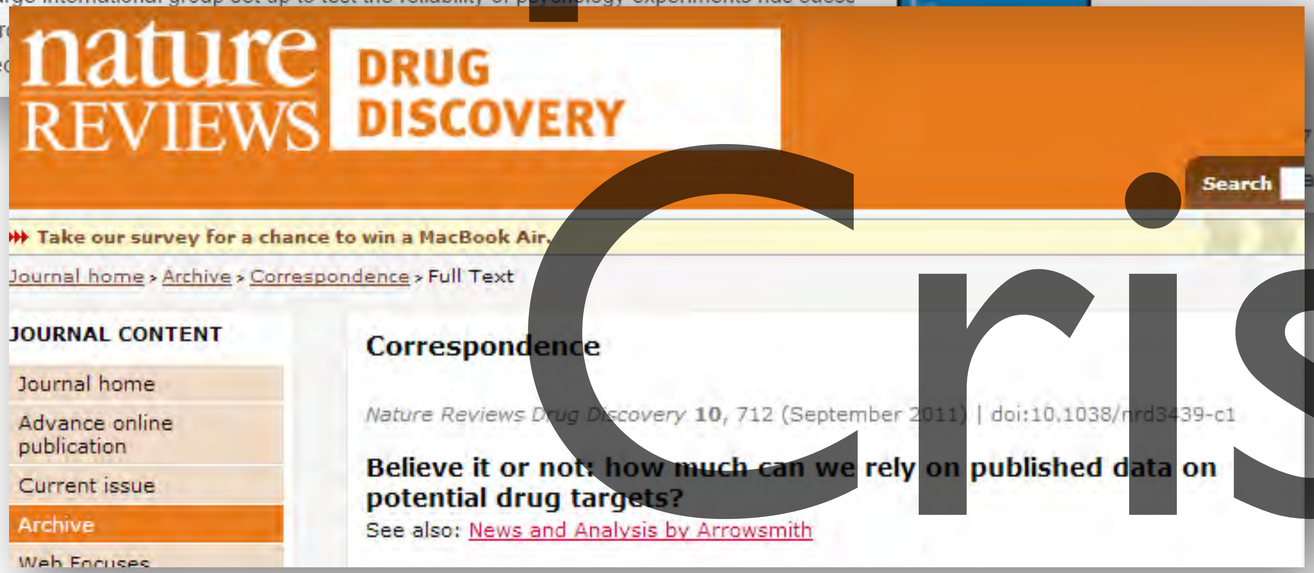
The Quarterly Journal of Experimental Psychology

Volume 65, Issue 11, 2012

Select Language

Translator disclaimer

A peculiar prevalence of *p* values just below .05



nature REVIEWS DRUG DISCOVERY

Search

Take our survey for a chance to win a MacBook Air.

Journal home > Archive > Correspondence > Full Text

JOURNAL CONTENT

Journal home

Advance online publication

Current issue

Archive

Web Focuses

Correspondence

Nature Reviews Drug Discovery **10**, 712 (September 2011) | doi:10.1038/nrd3439-c1

Believe it or not: how much can we rely on published data on potential drug targets?

See also: News and Analysis by Arrowsmith

711335

alande[b*]

Perspectives on Psychological SCIENCE

A Journal of the Association for Psychological Science

Home | All Issues | Subscribe | RSS | Email Alerts

Home

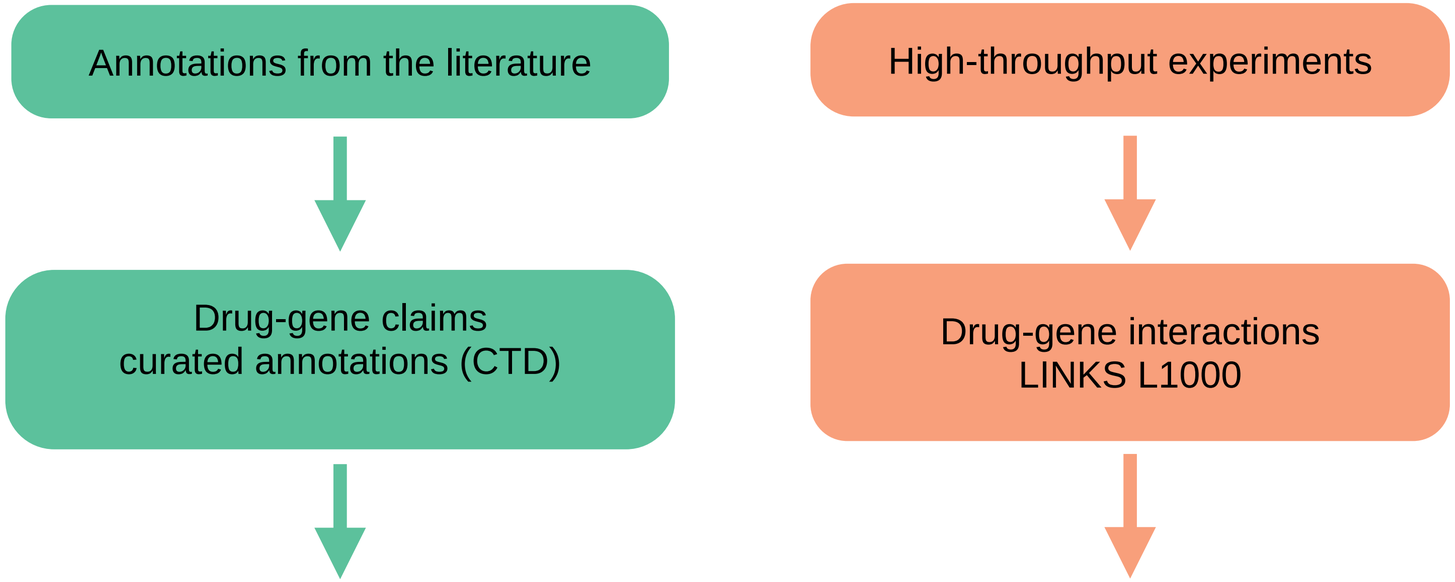
Editors' Introduction to the Special Section on Replicability in Psychological Science

A Crisis of Confidence?

Harold Pashler[1] and

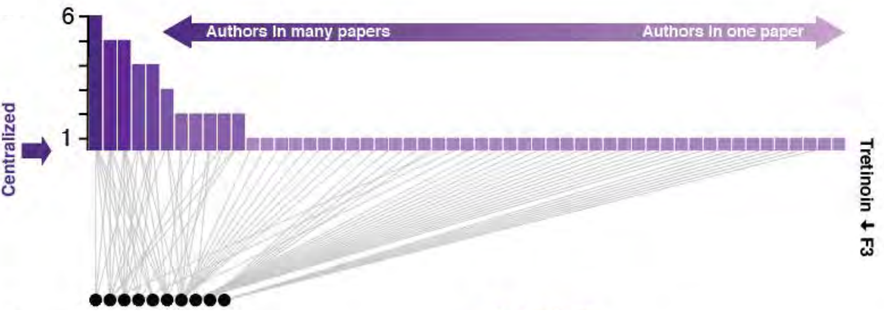
Eric–Jan Wagenmakers[2]

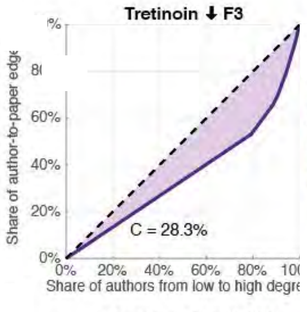
Aligning published claims with high-throughput experiments

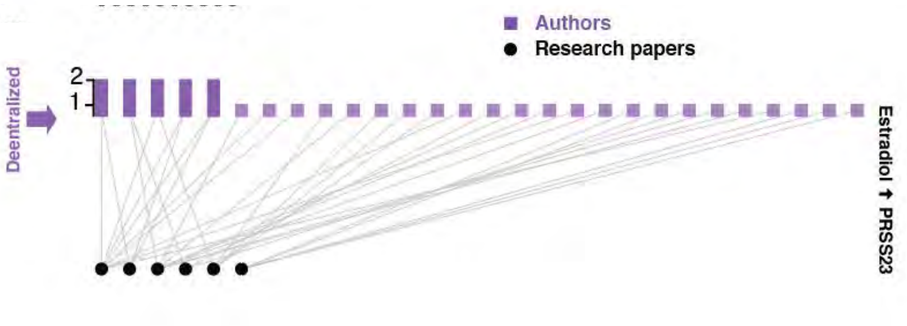


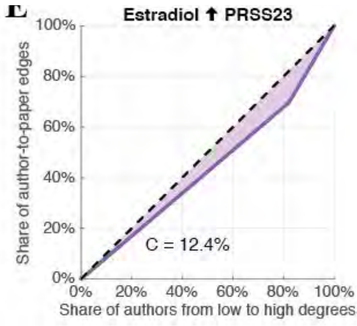
Drug	*Gene*	*Interaction*	*Support # papers*	*Oppose # papers*	*Papers IDs*	*Effect size*	*Statistical Significance*	*Variability cell, dose, time*
Benzo(a)pyren	*SLC22A3*	*Decreases expression*	*5*	*0*	*21224254, 19188863*	*+1.646*	*Y*	*0.783*
Sulforafan	*NQO1*	*Increases expression*	*16*	*0*	*1001256; …*	*+8.405*	*Y*	*1.719*
Estradiol	*PRSS23*	*Increases expression*	*6*	*2*	*21224254; 10022439*	*+10.524*	*Y*	*2.097*
Tretinoin	*MYC*	*Decreases expression*	*12*	*0*	*10023685; …*	*–2.654*	*N*	*11.848*

Centralization of Scientific Communities

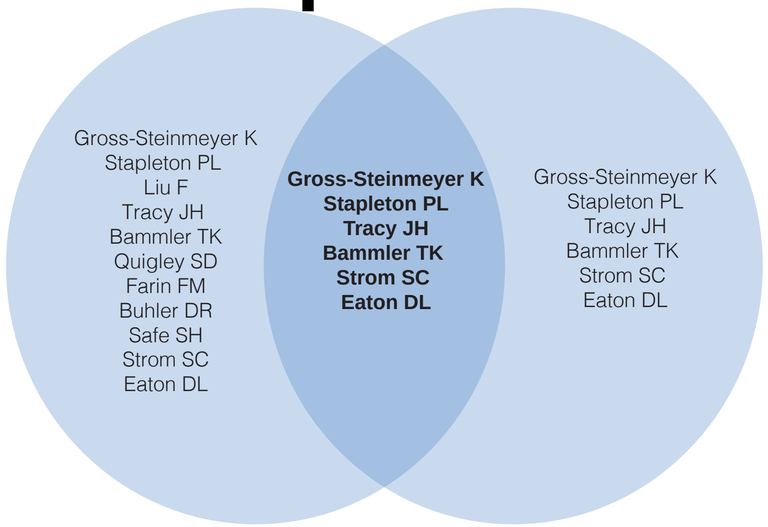






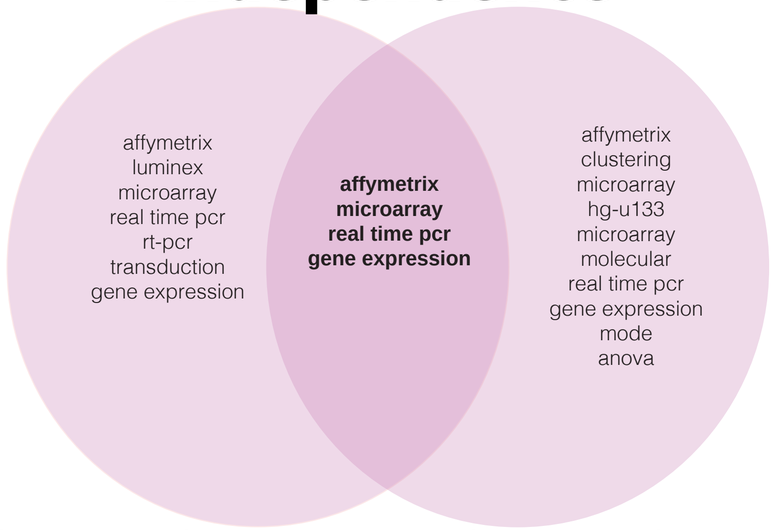


Social independence



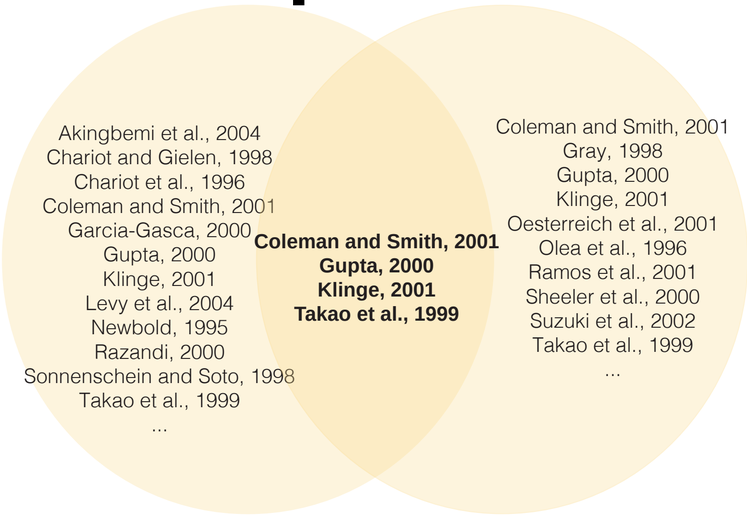
Gross-Steinmeyer K
Stapleton PL
Tracy JH
Bammler TK
Strom SC
Eaton DL

Methodological independence



affymetrix
microarray
real time pcr
gene expression

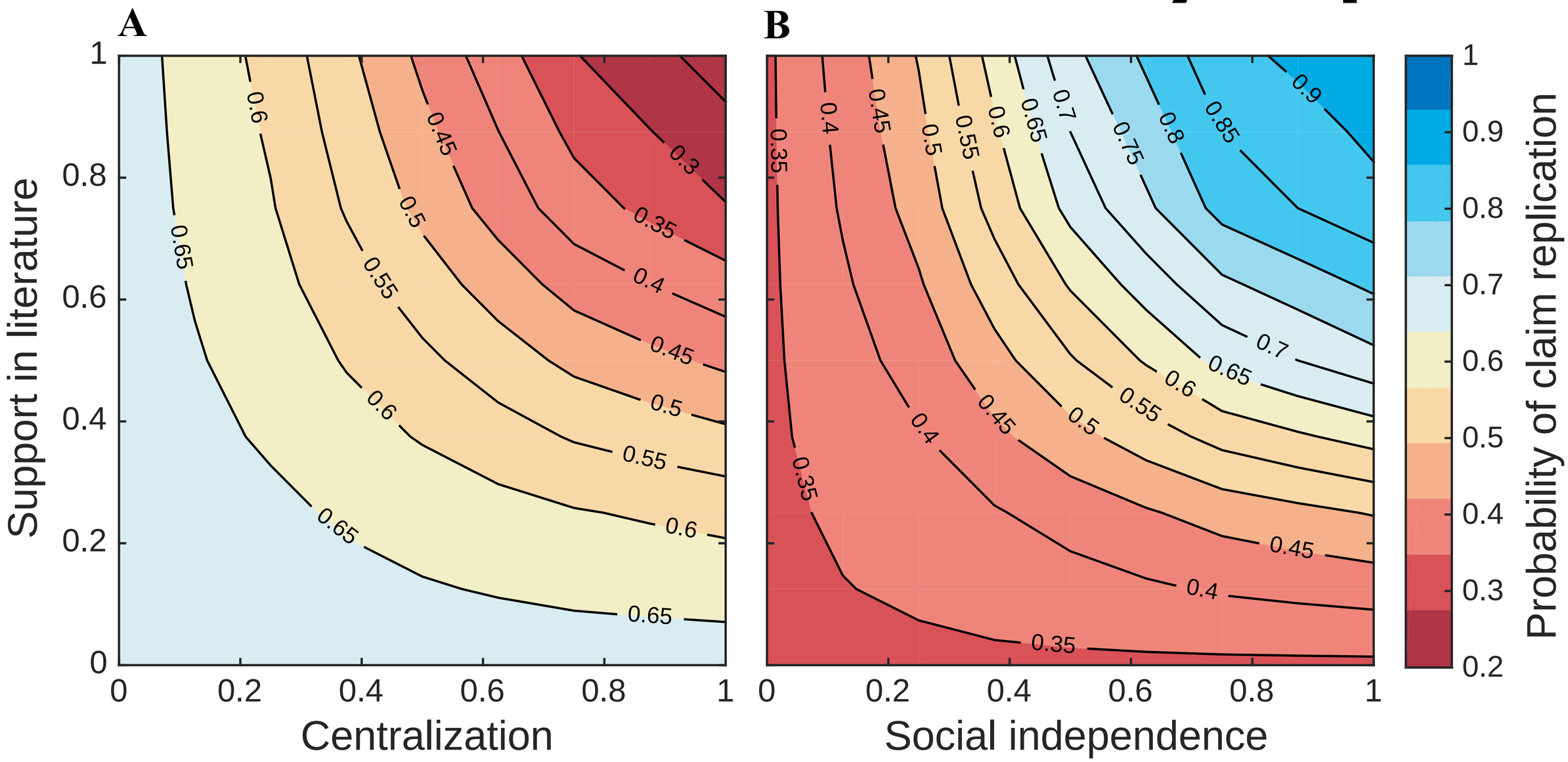
Prior Knowledge independence



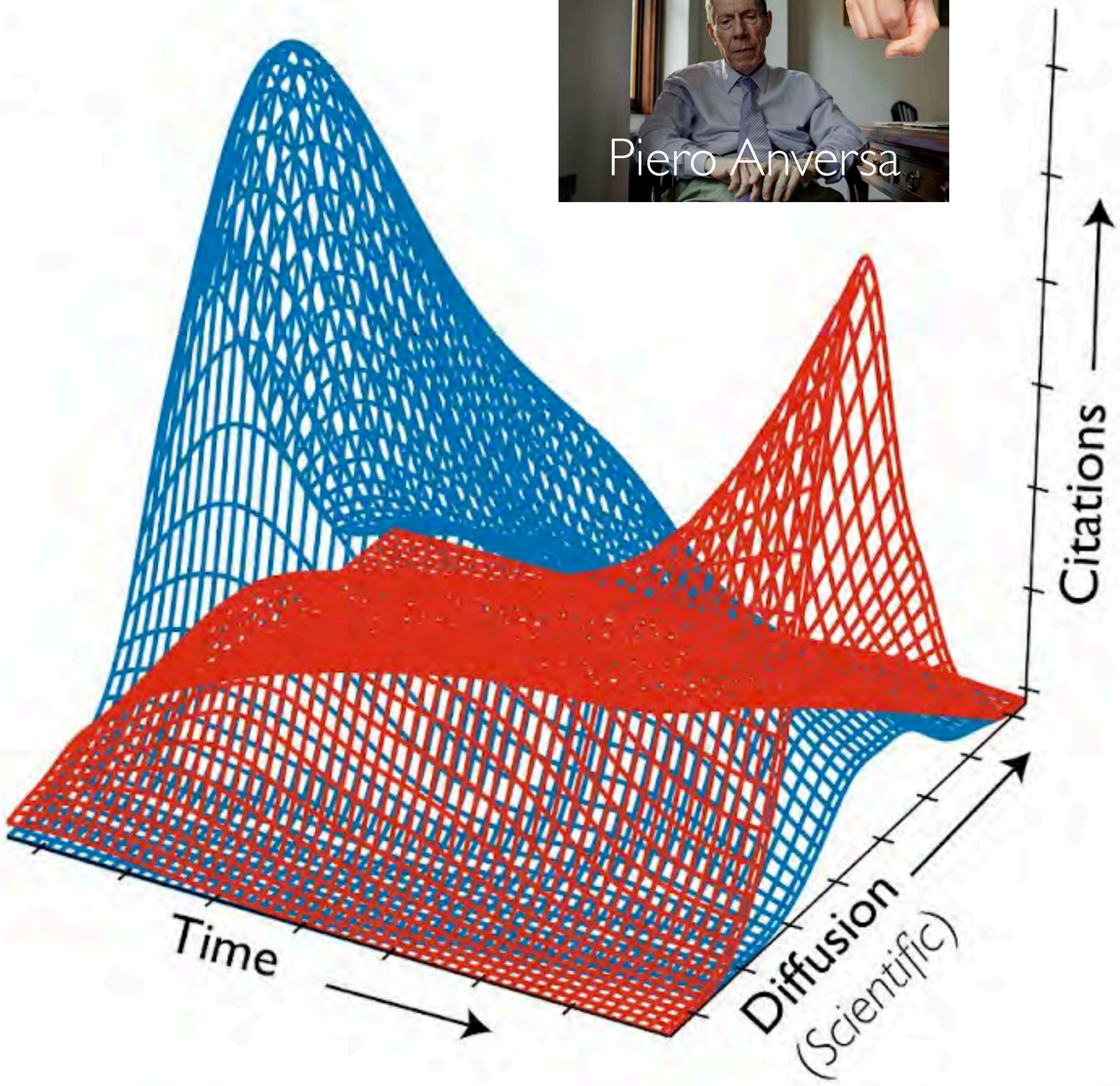
Coleman and Smith, 2001
Gupta, 2000
Klinge, 2001
Takao et al., 1999

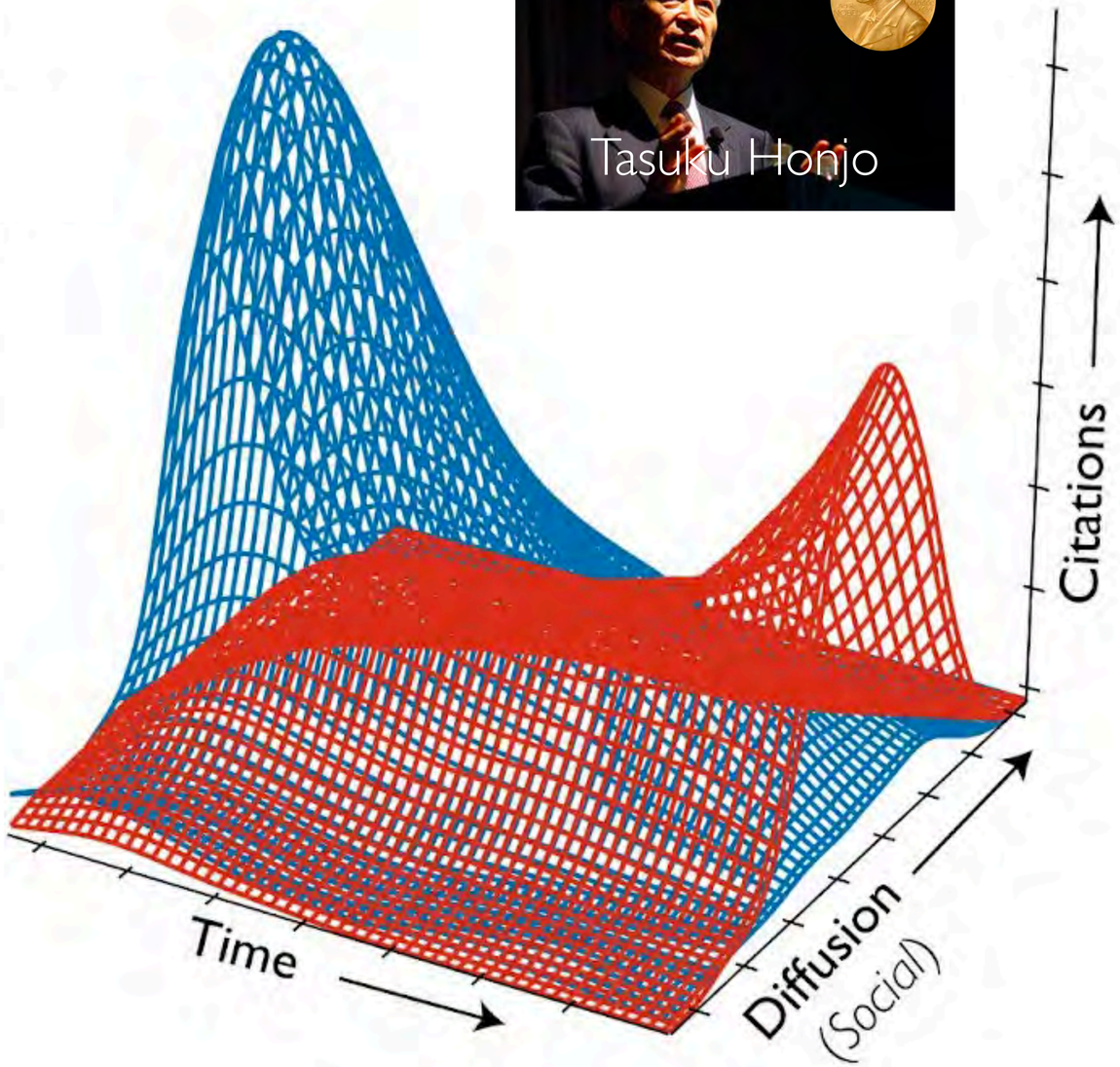
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Findings from Decentralized Communities More Likely Replicate



Locate (**& Burst**) Academic Bubbles





PMID:11777997 - Cardiac Regeneration

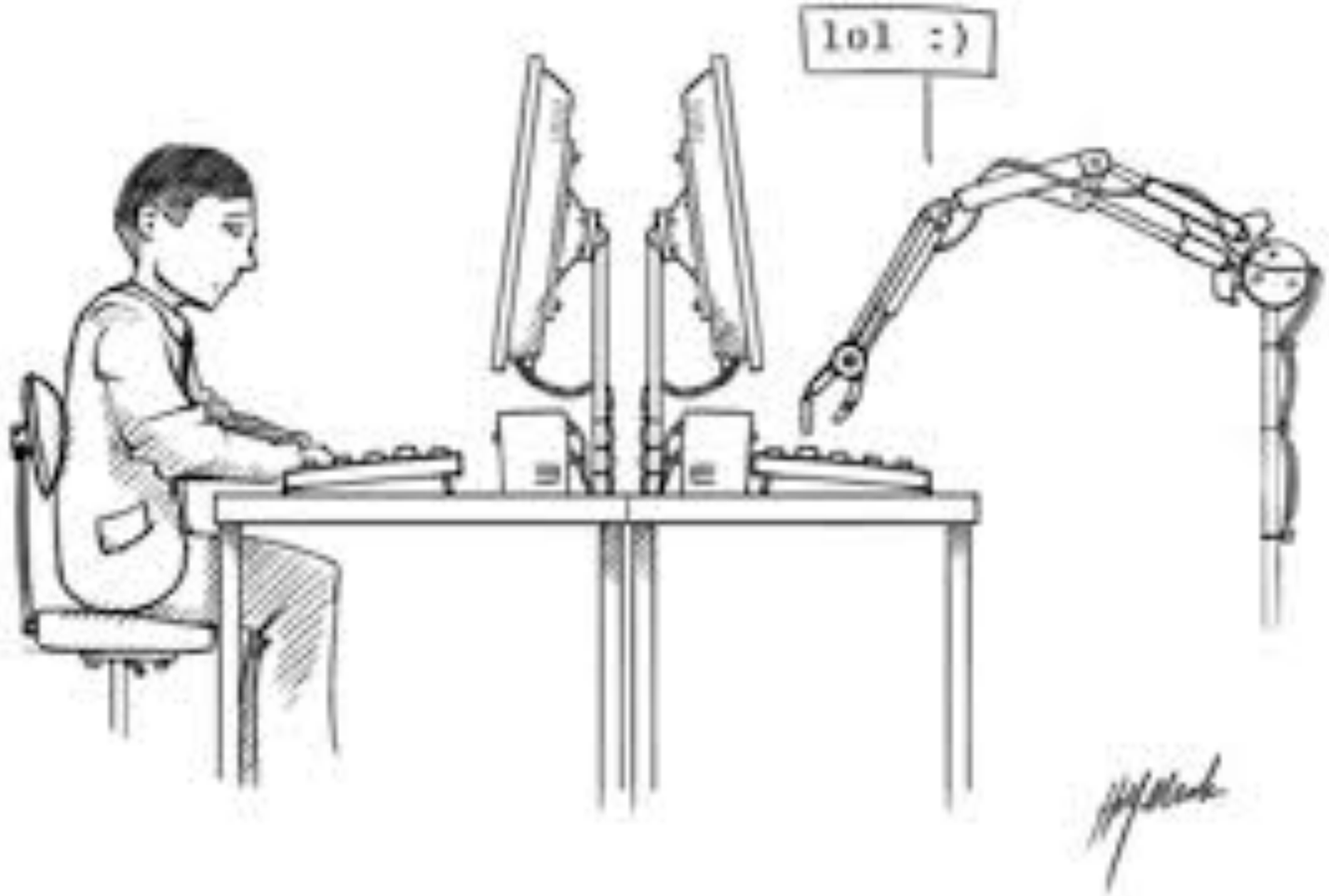
PMID:11015443 - Cancer Immunotherapy

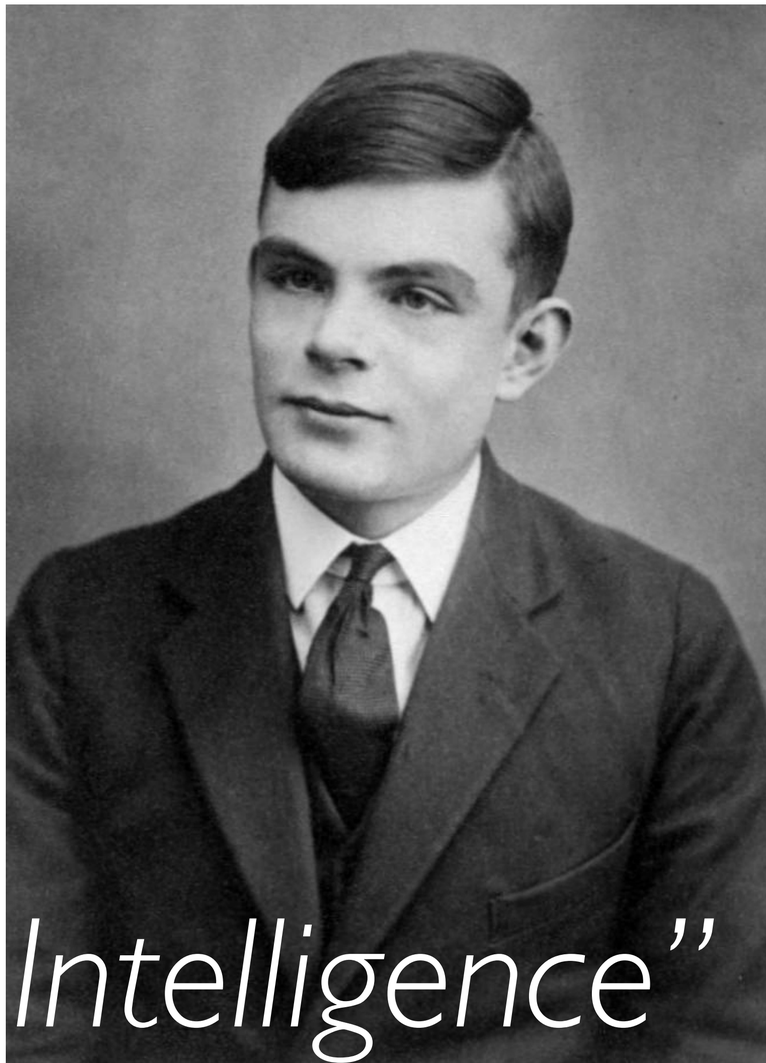
Can we design diversity to increase collective imagination

Hybrid Vigor

Jamshid Sourati
Postdoc

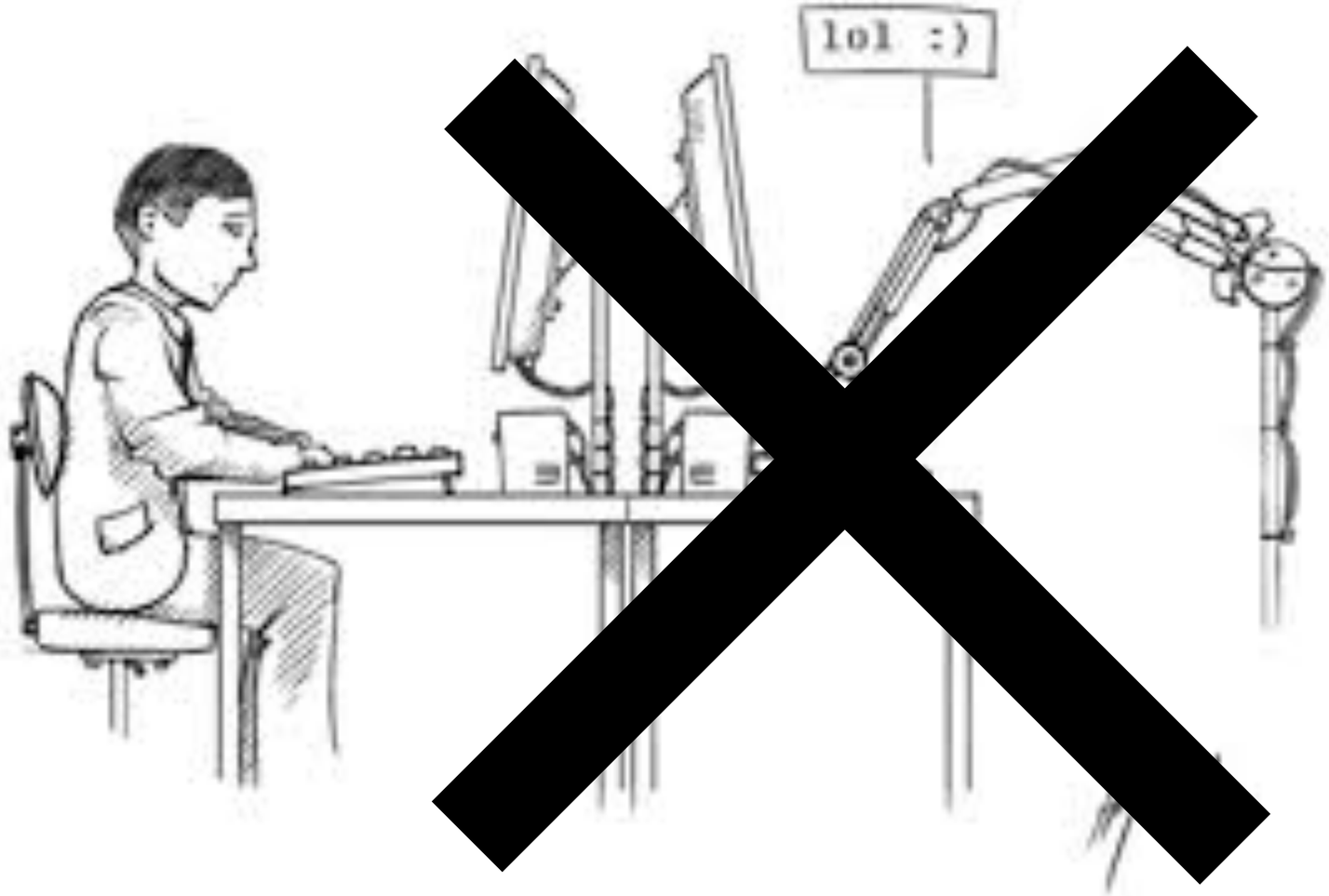
Alan Turing's **'Immitation Game'**

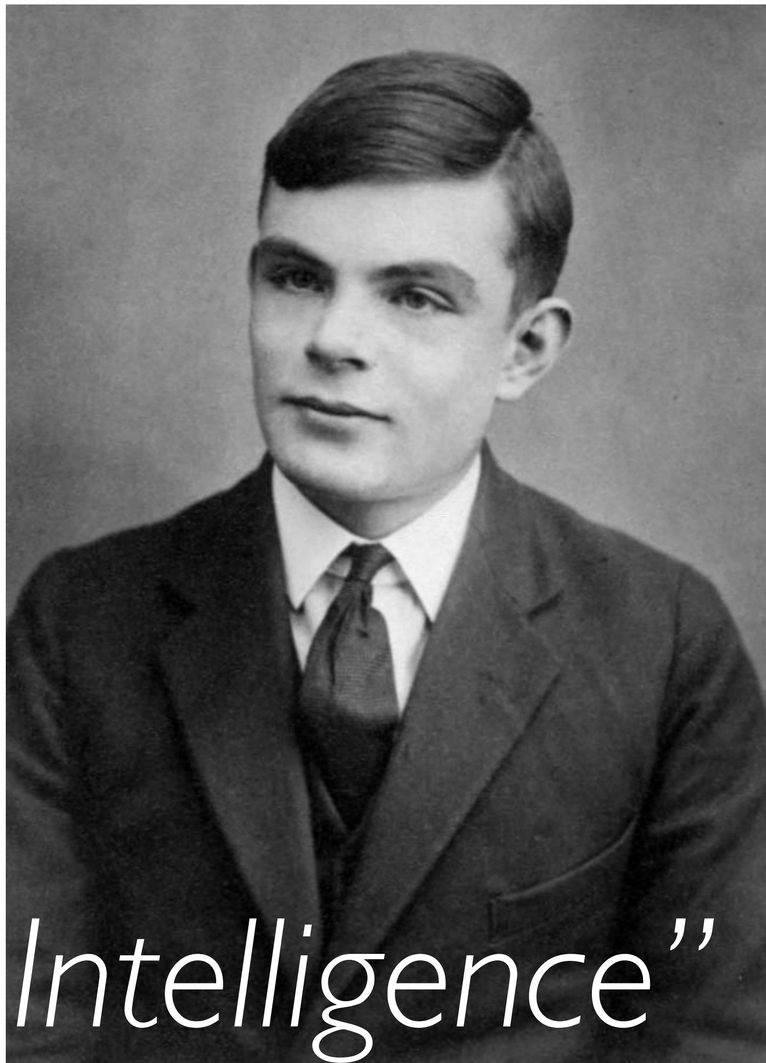




in *"Computing Machinery & Intelligence"*

Alan Turing's **'Immitation Game'**





in *"Computing Machinery & Intelligence"*

Alien Intelligences
Think Differently

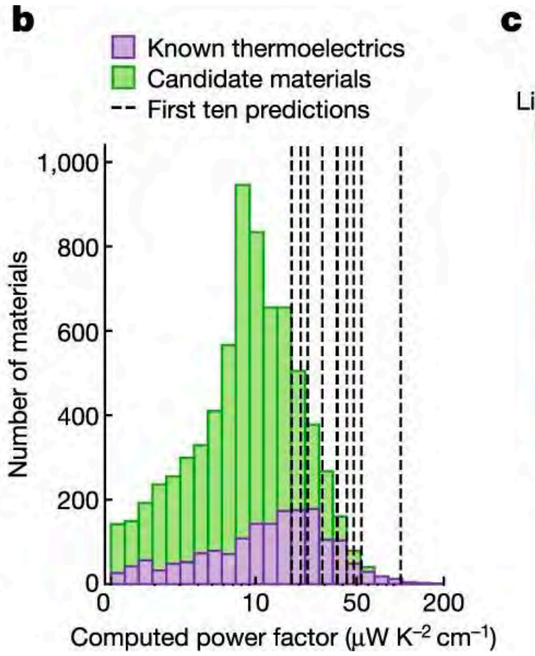
nature
International journal of science

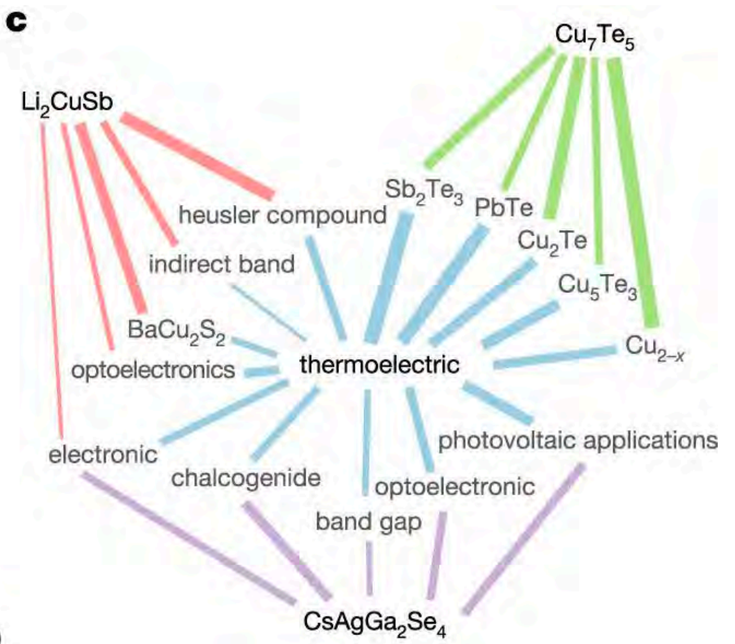
Letter | Published: 03 July 2019

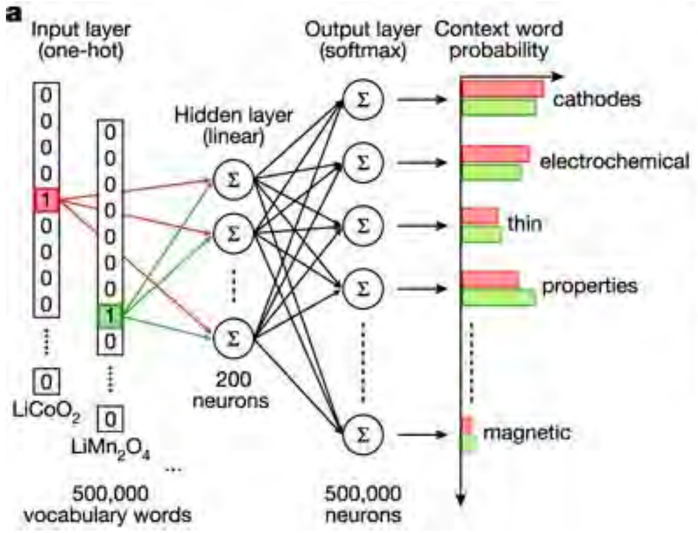
Unsupervised word embeddings capture latent knowledge from materials science literature

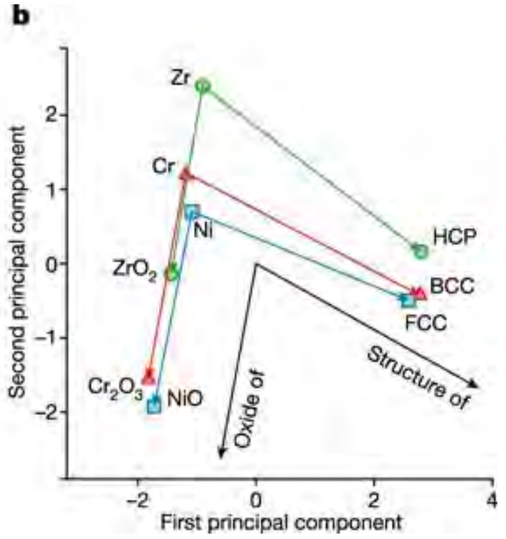
Vahe Tshitoyan ✉, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder ✉ & Anubhav Jain ✉

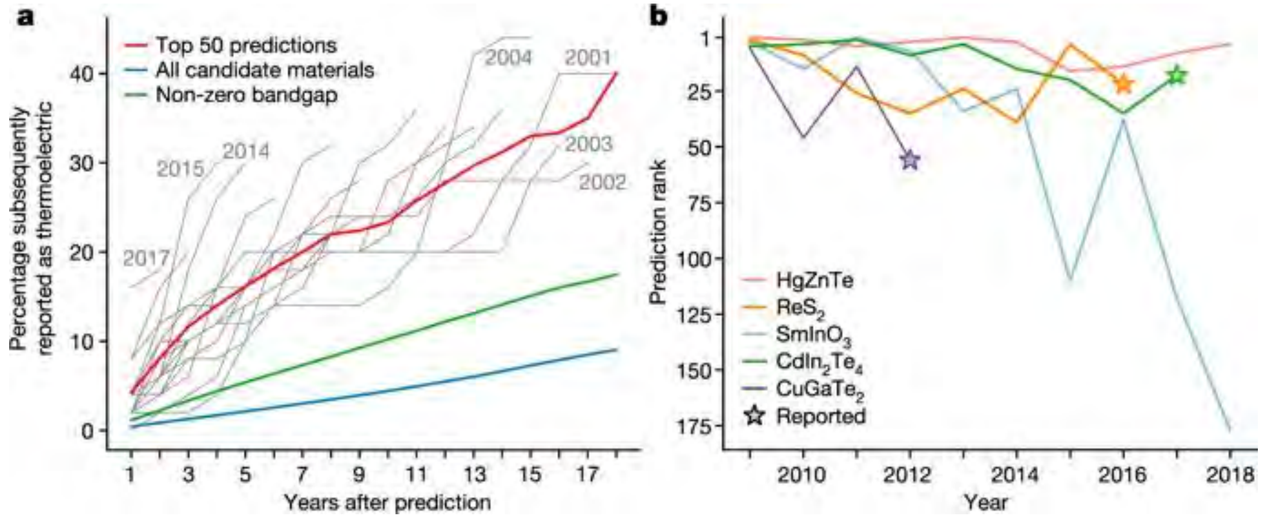
trics





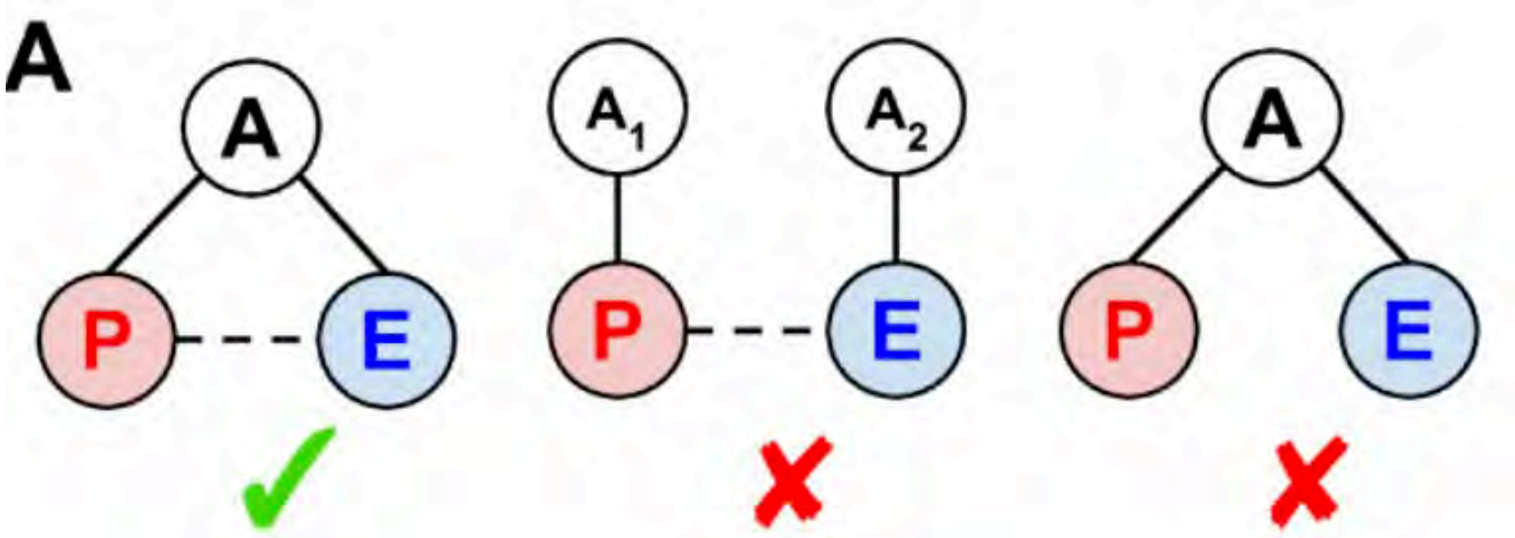




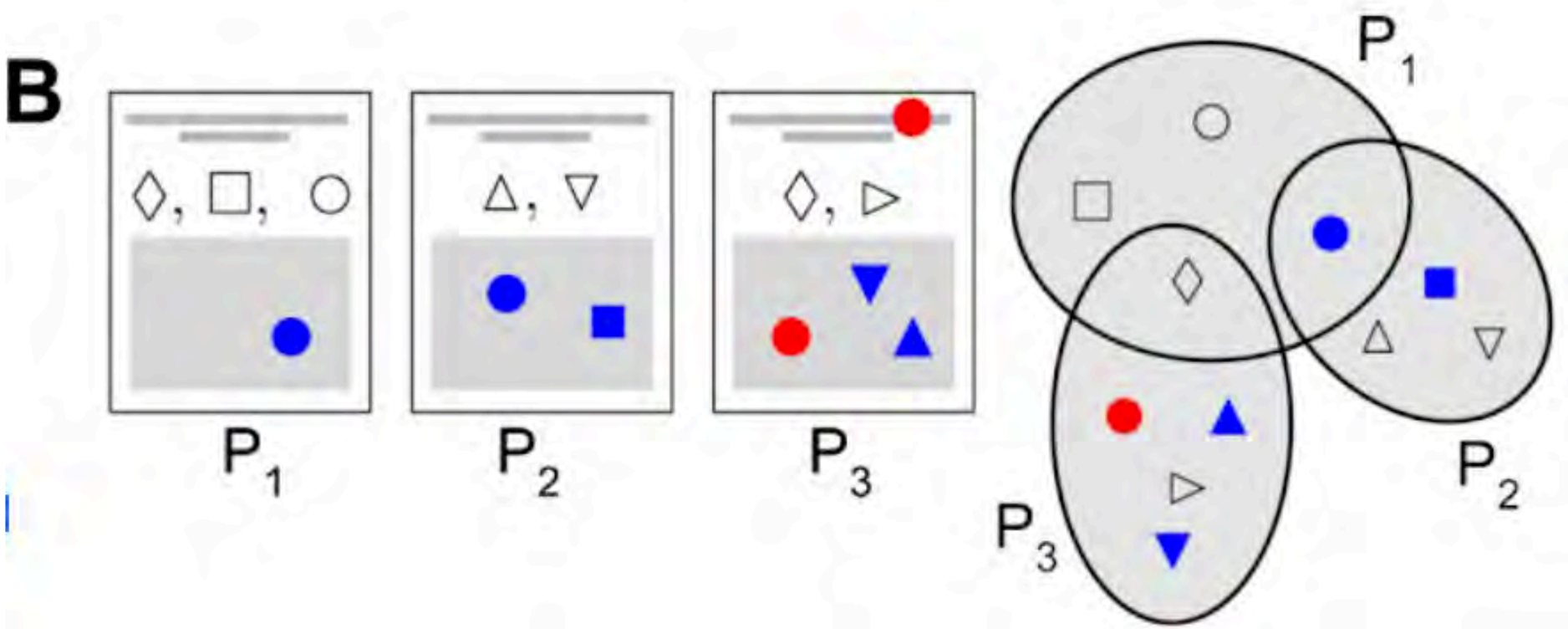


40% Precision

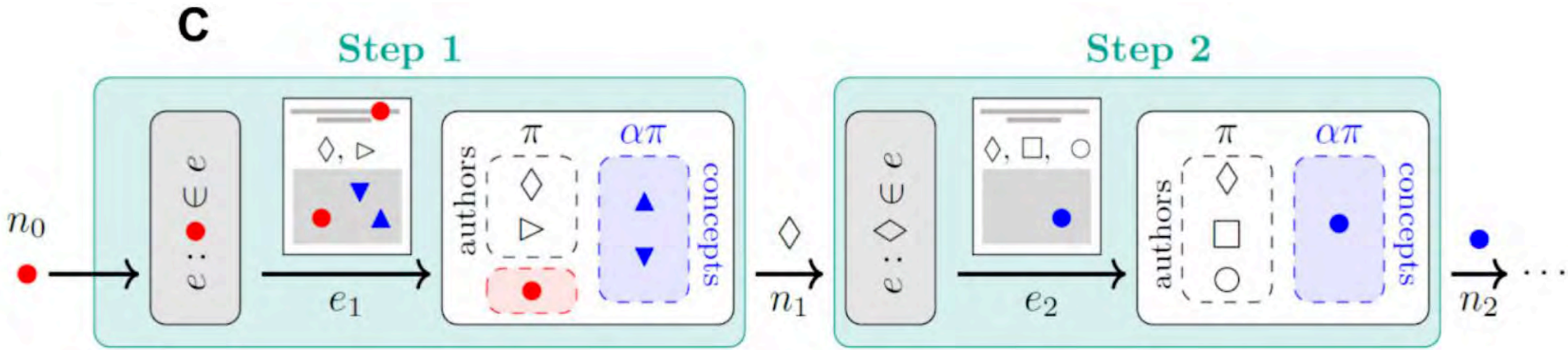
What Inferences are Most Likely **Cognitively?**



What Inferences are Most Likely **Cognitively?**



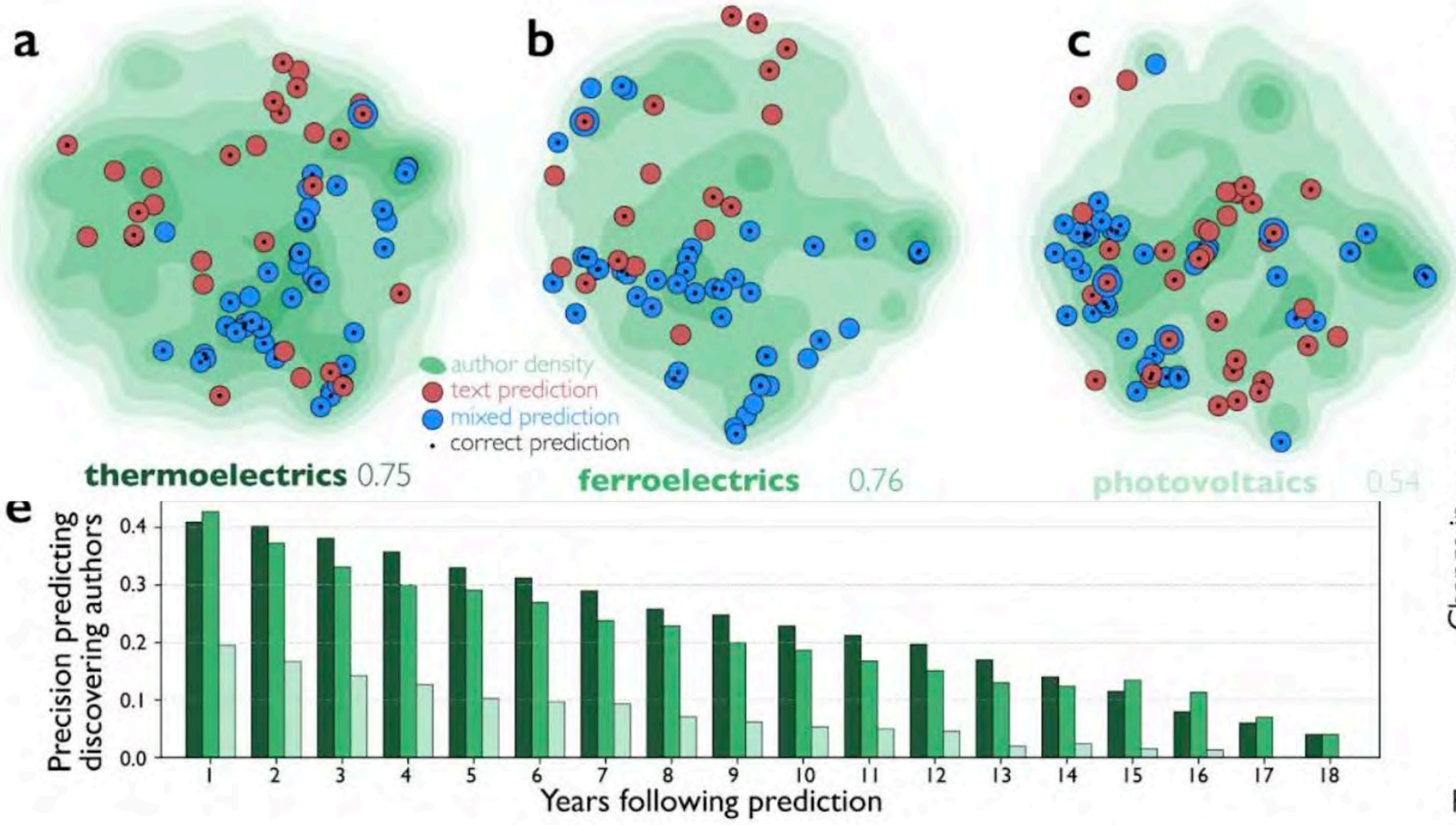
What Inferences are Most Likely **Cognitively?**

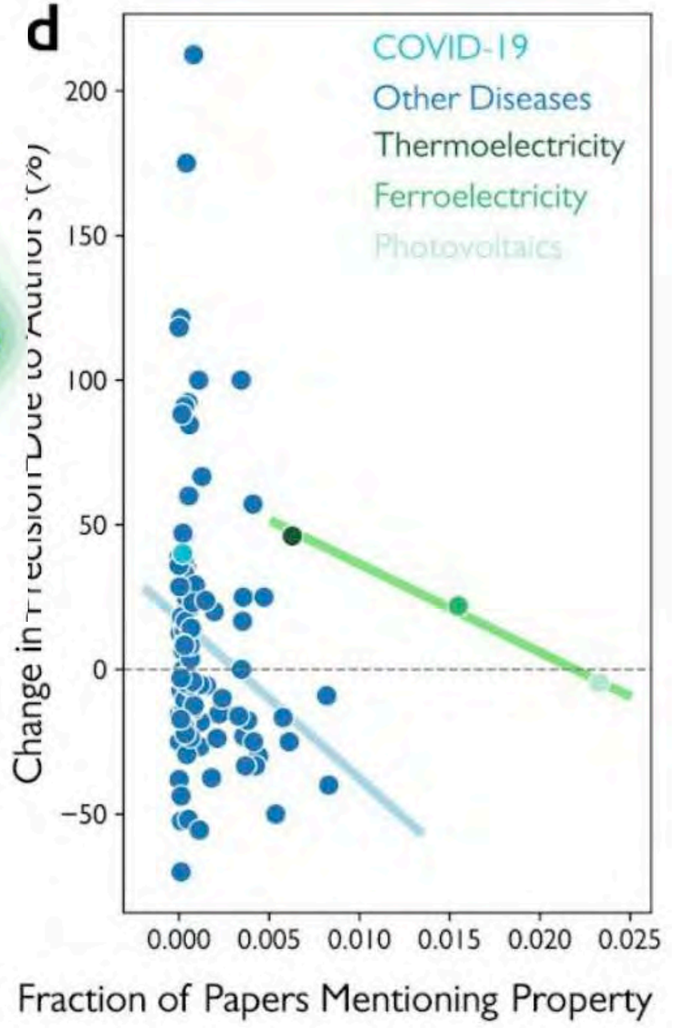


$$n_i \sim \pi\mathcal{U}\left(e_i \text{ authors} \cup \{\bullet\}\right) + \alpha\pi\mathcal{U}\left(e_i \text{ entities}\right)$$

$\mathcal{U}$: uniform distribution

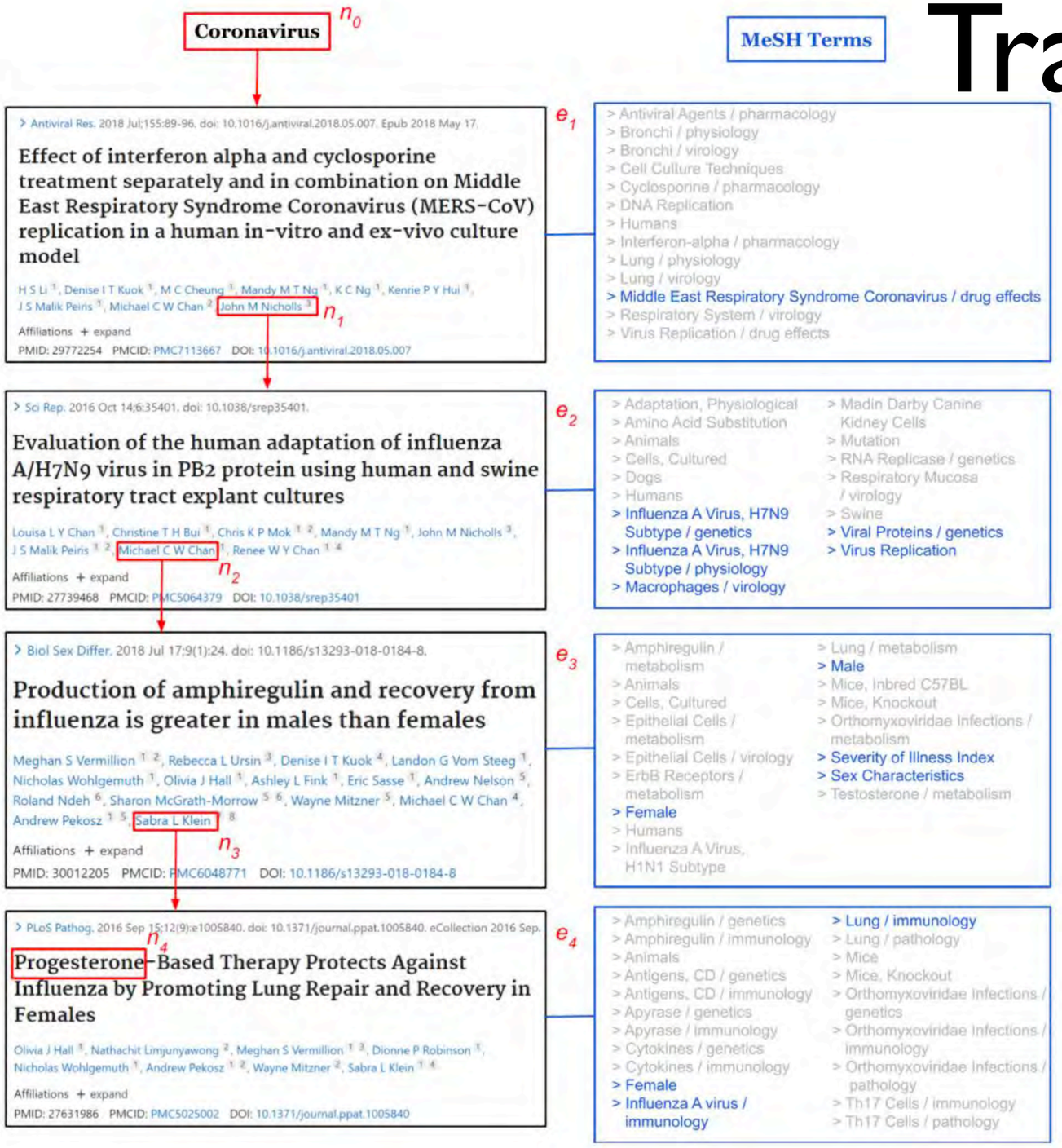
Why/When?

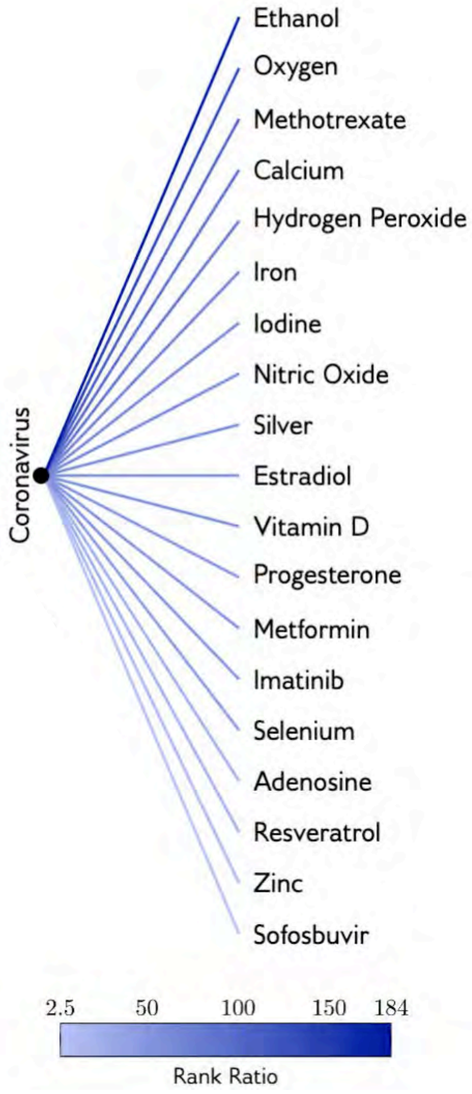




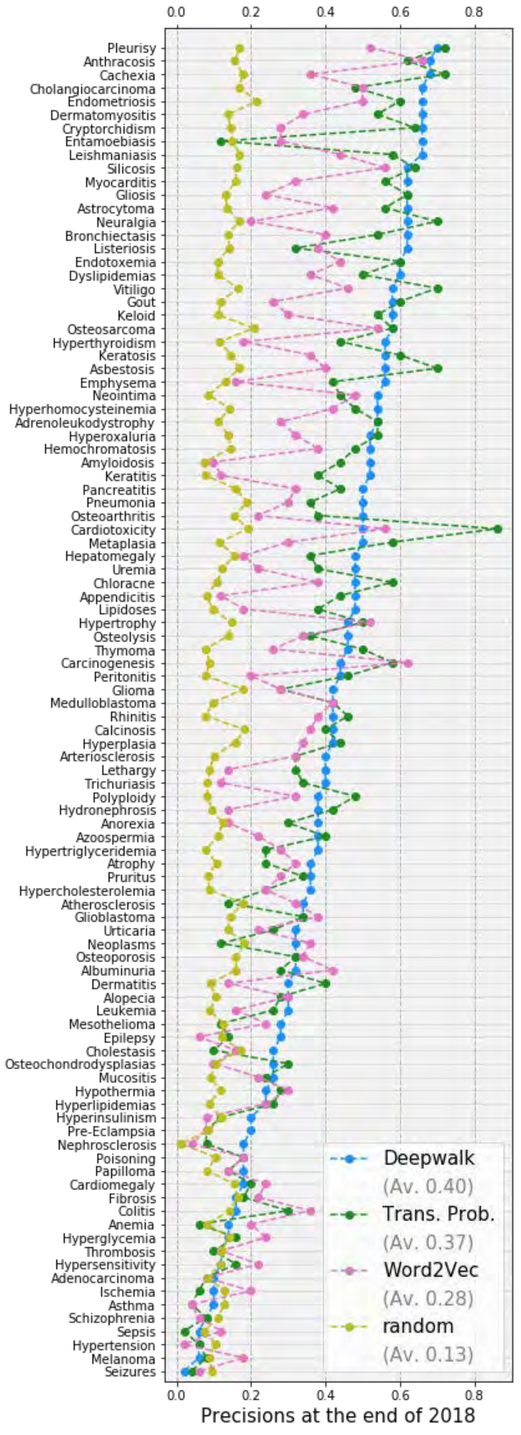
Predicting Cognitively Available Inference

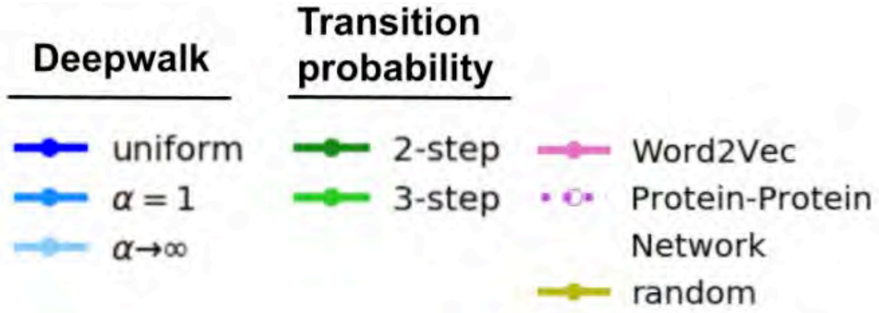
Tracing Discovery In Science

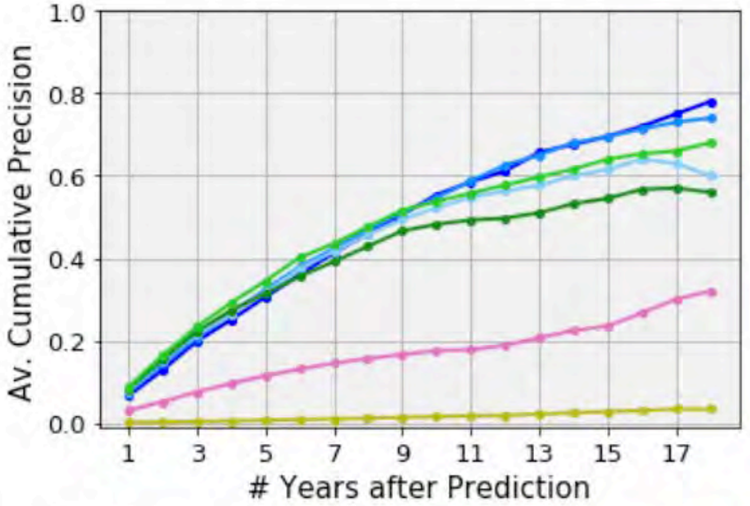


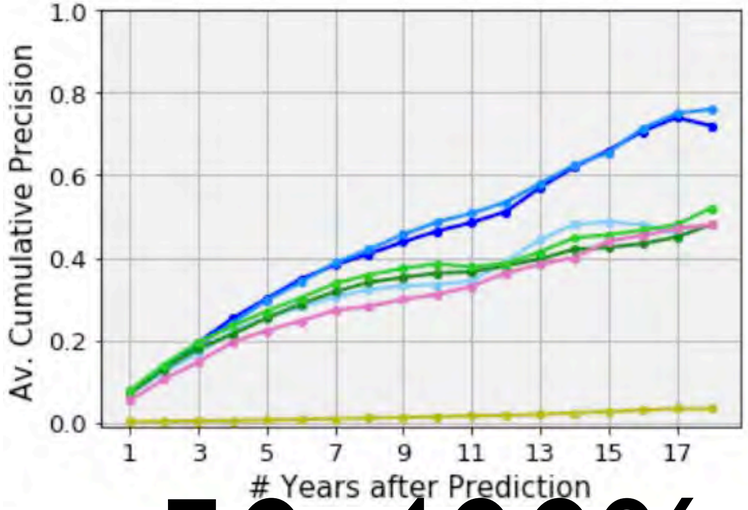


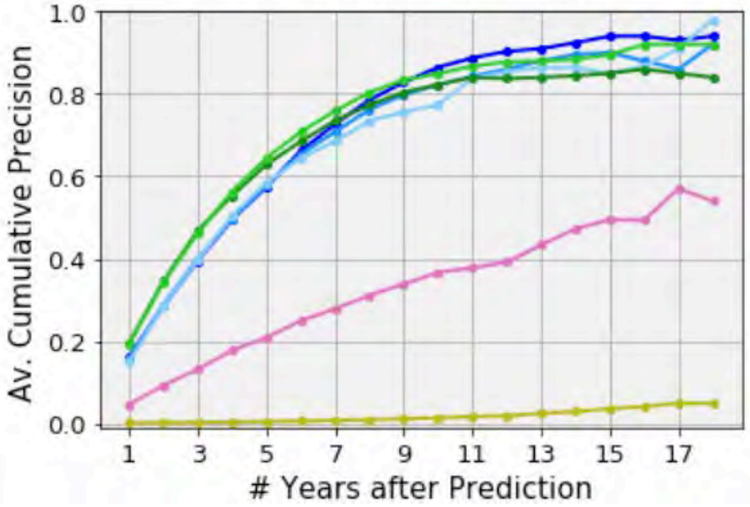
100 Drugs; 450 Diseases



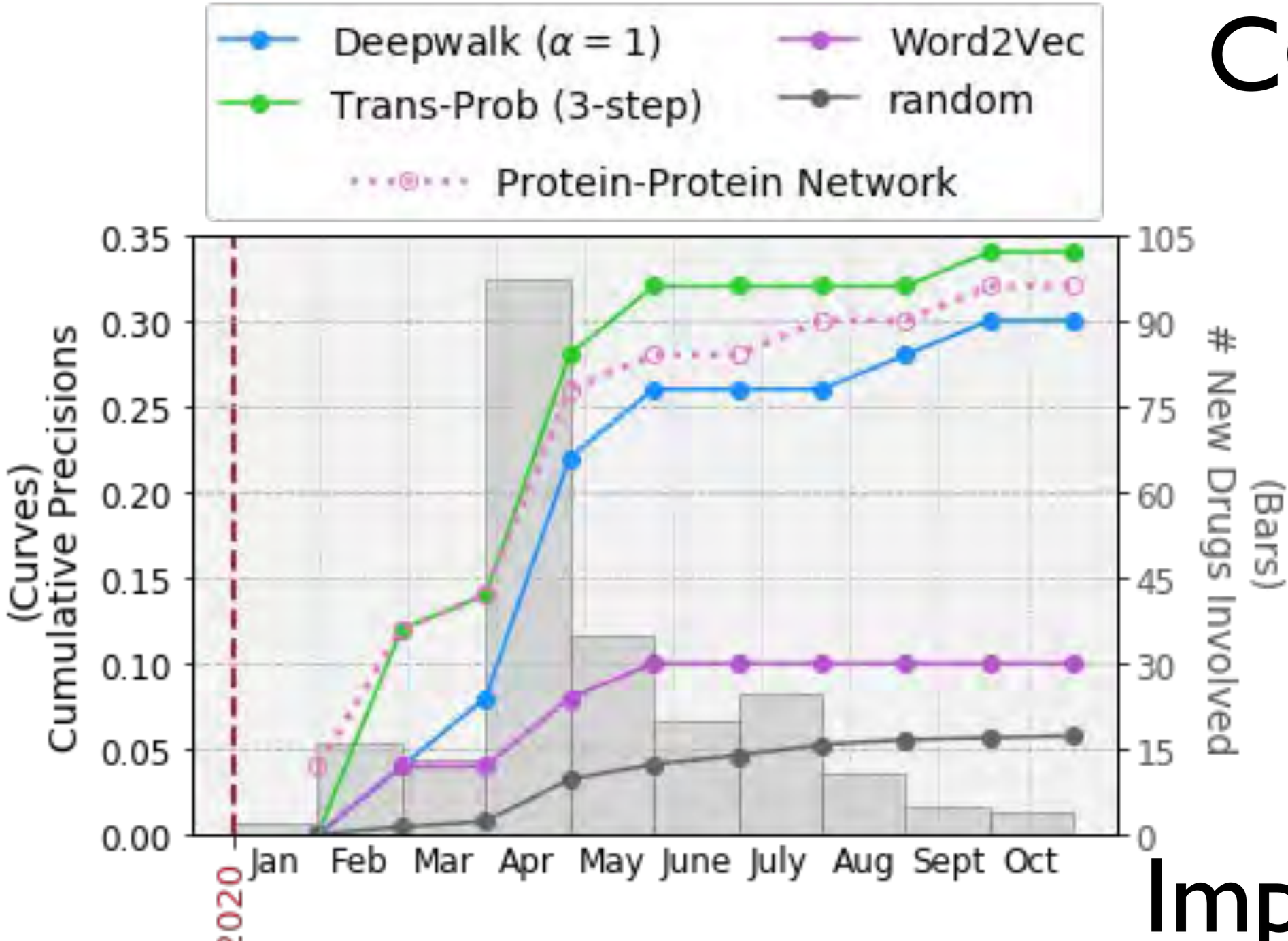








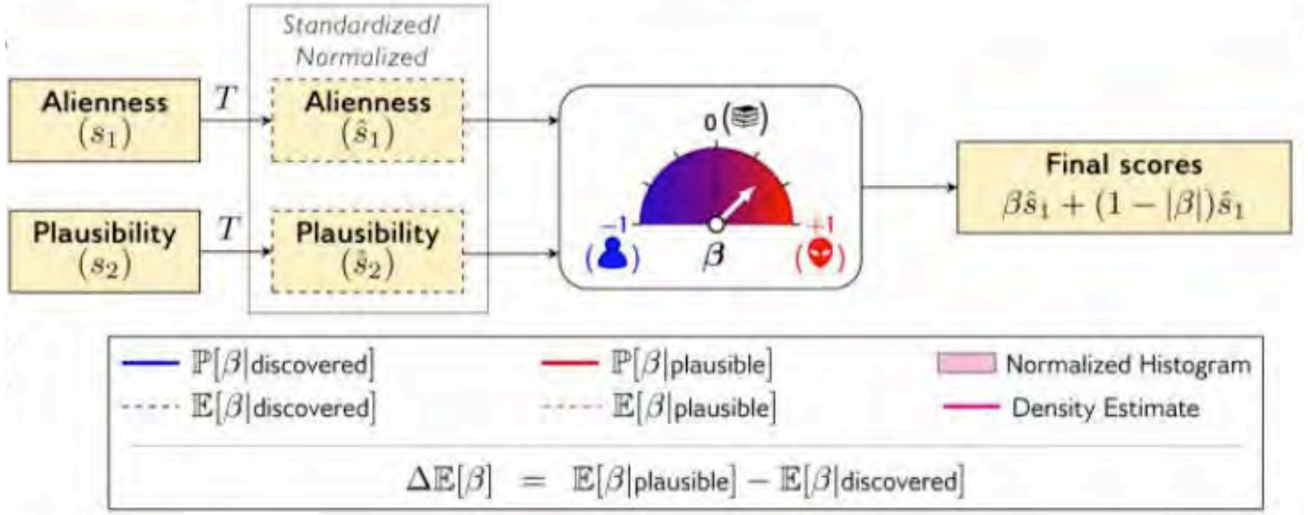
50-400% Improvement

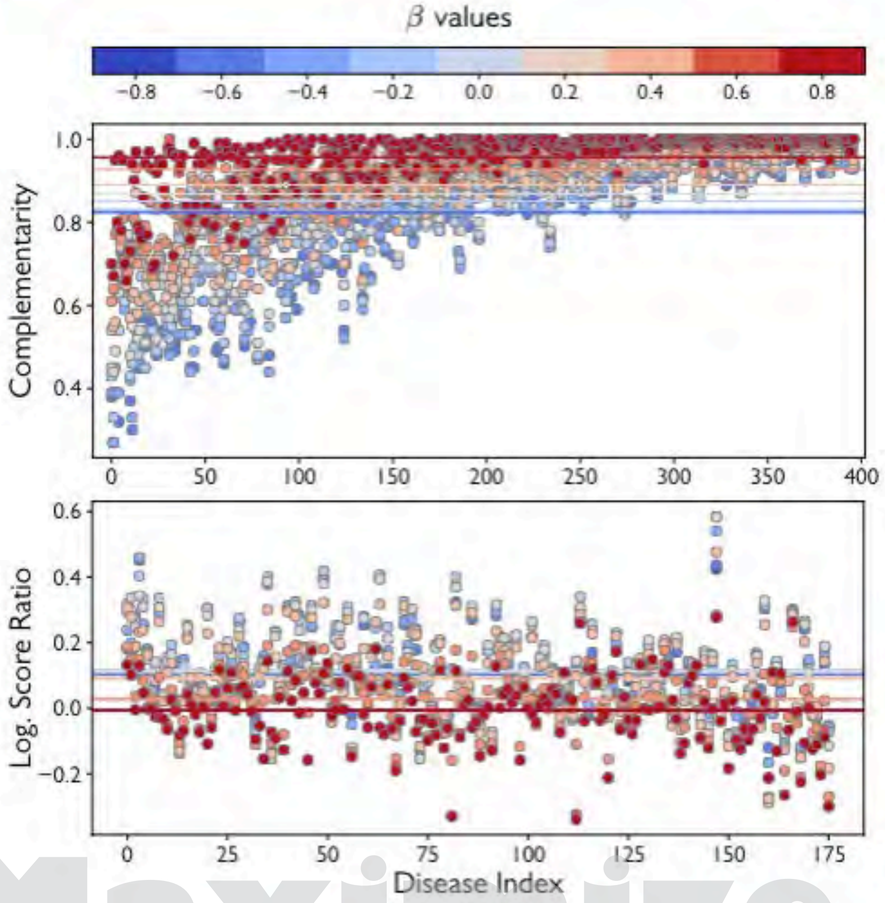


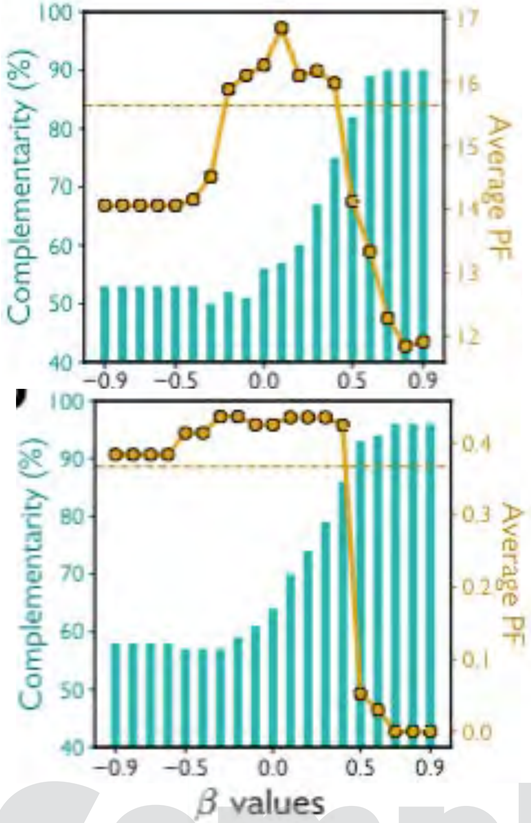
COVID-19 Search

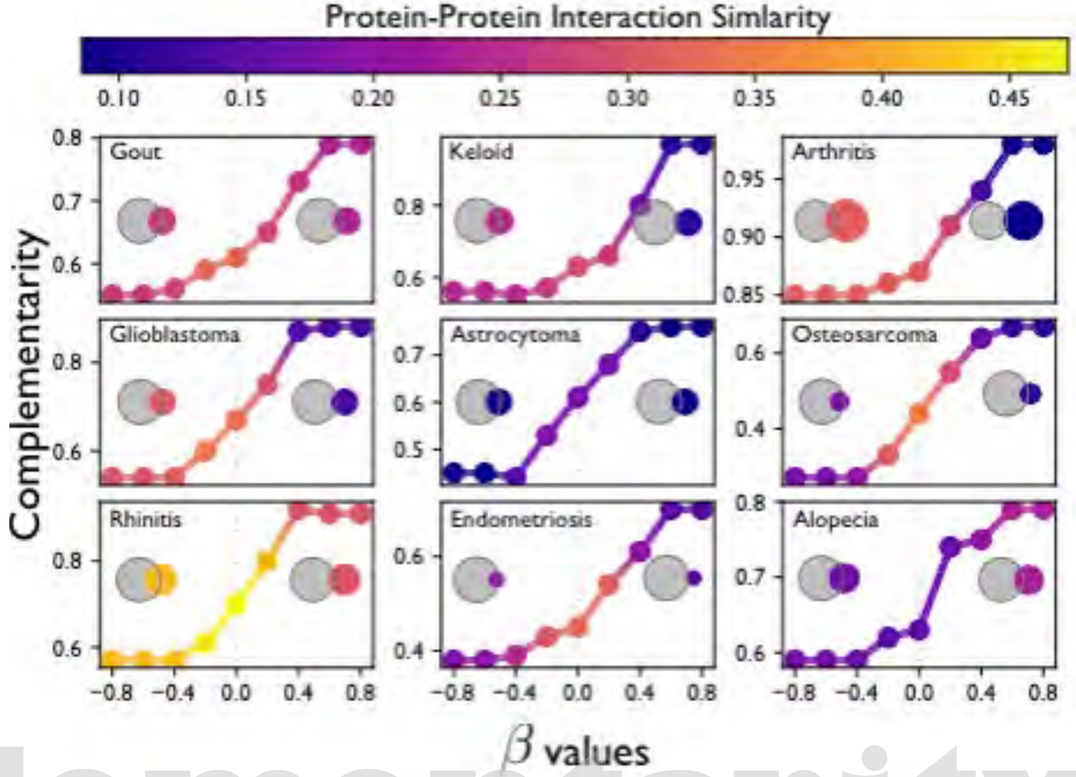
400% Improvement

By Avoiding Scientists we can **Predict** what they *won't* **Discover**



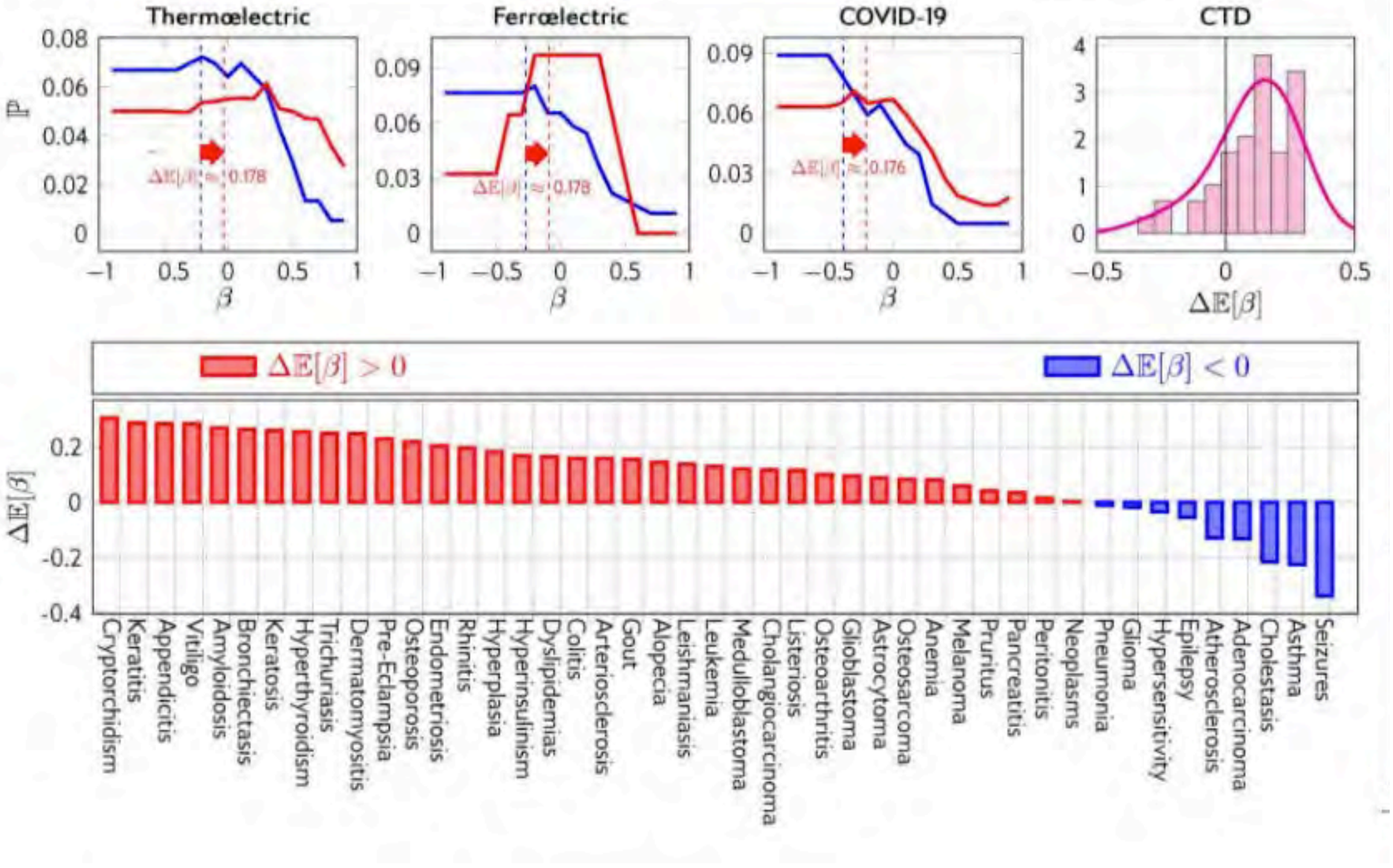


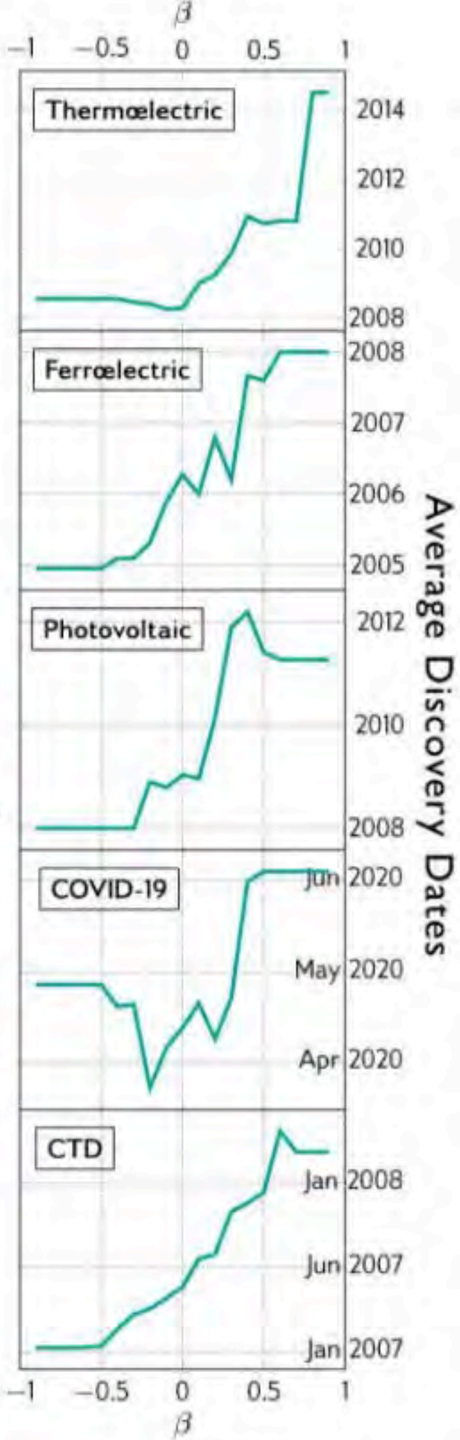




Maximize Complementarity

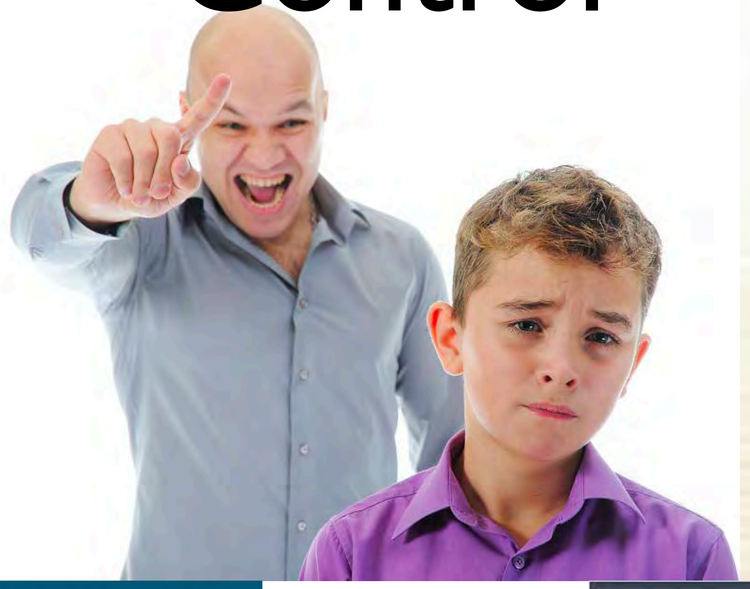
(and its better)





Training Logic

Control

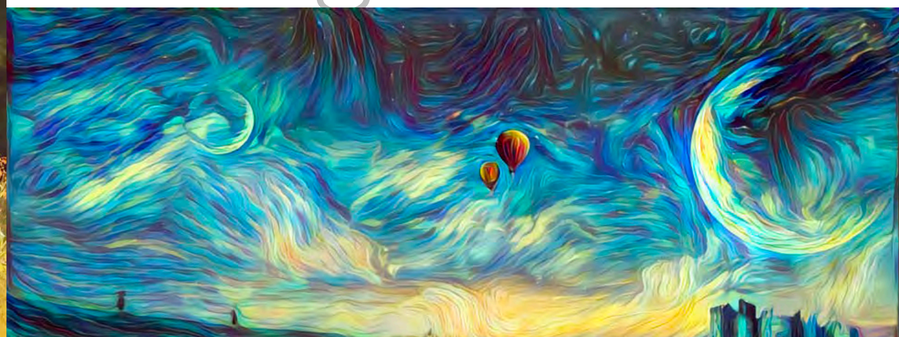


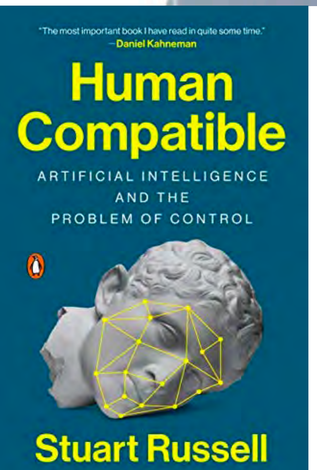
Caregiving

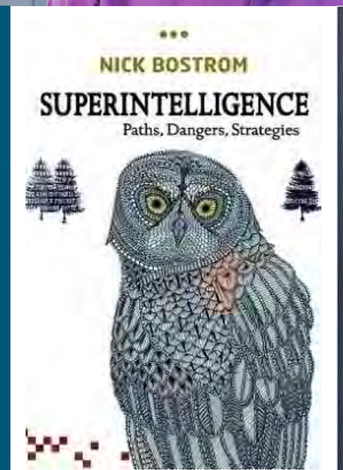
Alien Capacities

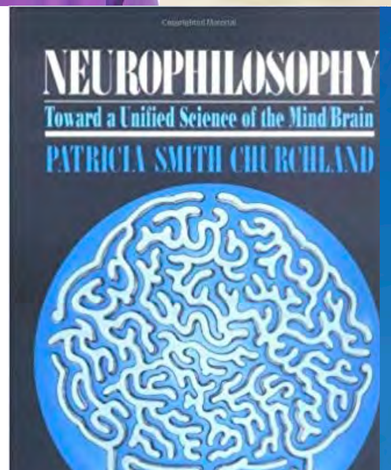
Objectives

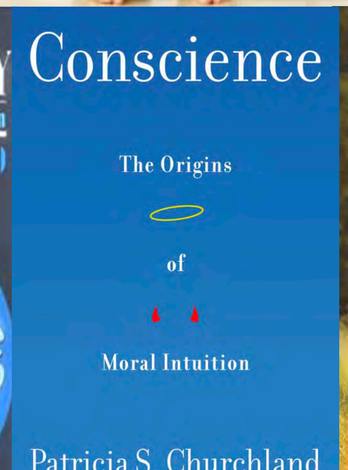
Imagination

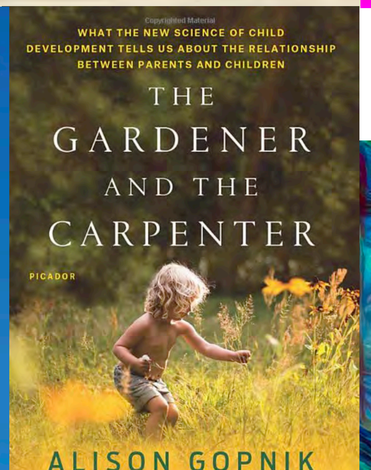








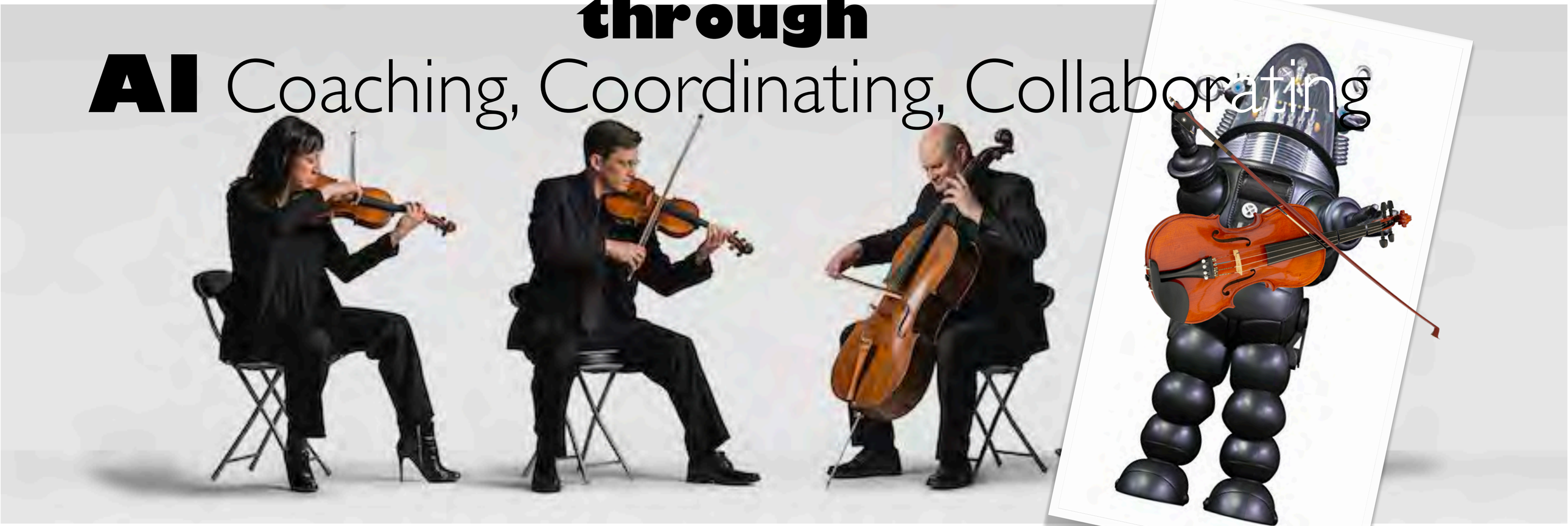




Build Ensembles of Diverse Intelligences

through

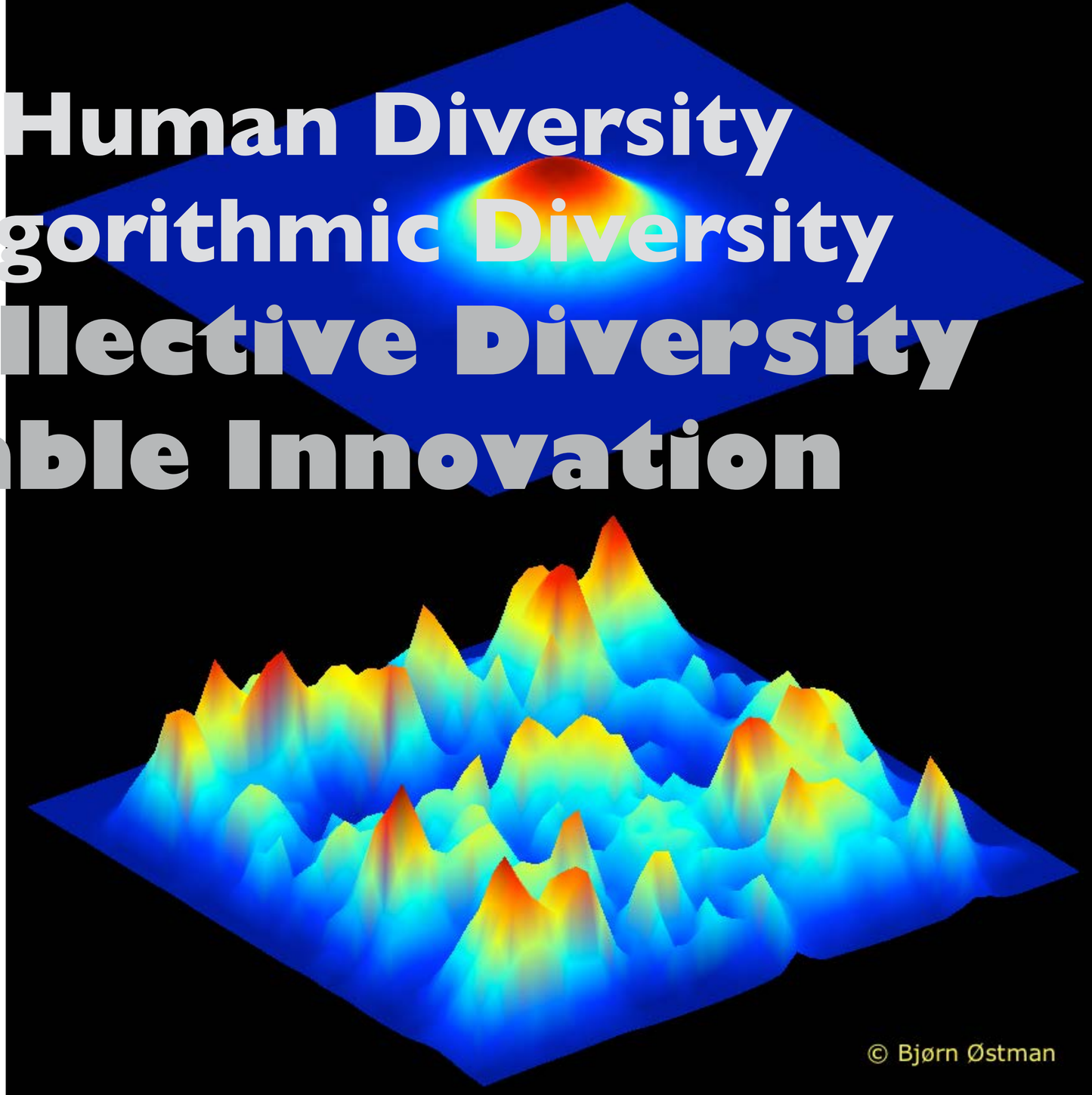
AI Coaching, Coordinating, Collaborating



To Advance Science, Technology, Enterprise,

& **the regulation of other AI**s

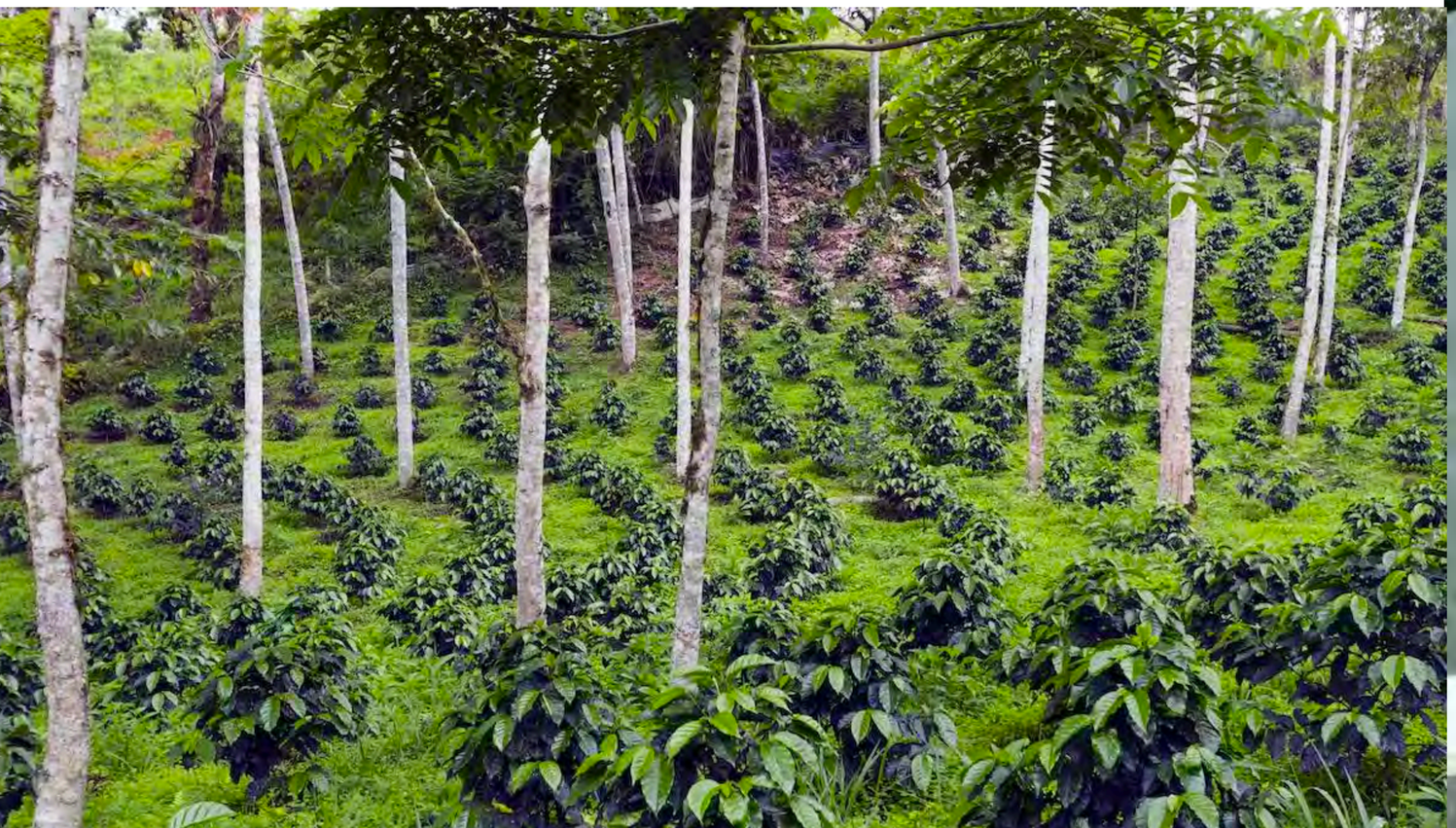
Conserving Human Diversity
Cultivating Algorithmic Diversity
Designing Collective Diversity
for Sustainable Innovation



- Innovation in Science, Tech, Enterprise & Culture are **High-Dimensional & Rugged** requiring
- Social, Cultural, Institutional, & Methodological **Disconnection**

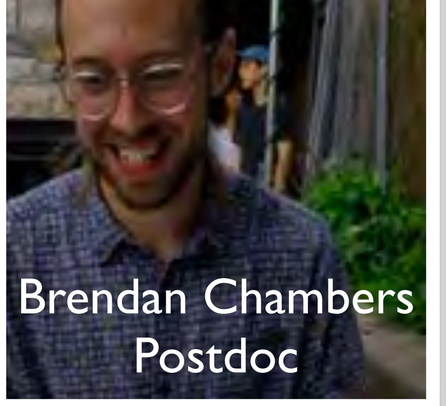
SANTA FE INSTITUTE

Sustainable Bridging



Jamshid Sourati
Postdoc

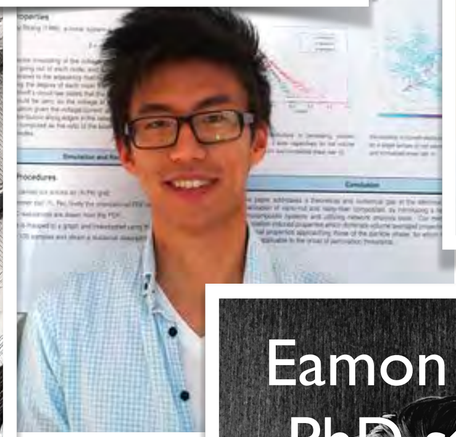
NSF


Brendan Chambers
Postdoc

Haizi Yu
Postdoc

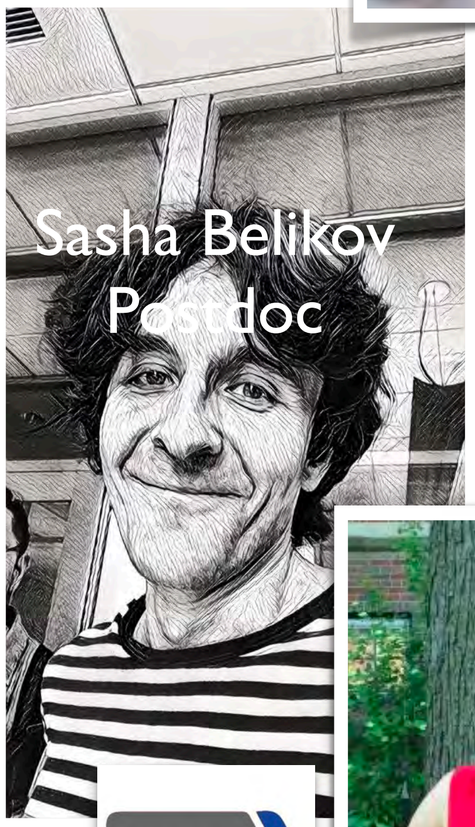
DARPA

jumptrading

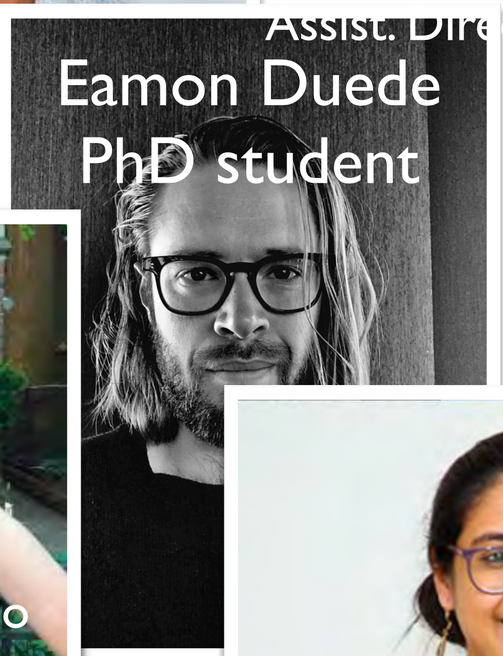


Caitlin Lewis
Assist. Director

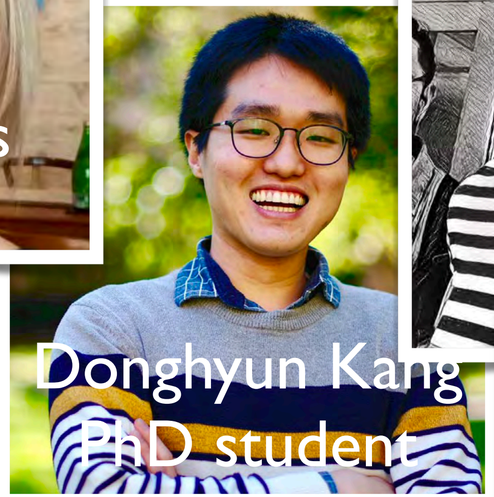
Fengli Xu
Postdoc


Sasha Belikov
Postdoc

Molly Lewis
Postdoc


Eamon Duede
PhD student

Bhargav Desikan
Masters student


Donghyun Kang
PhD student

John
Templeton
Foundation

NIH

Clara del Junco
Postdoc

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE