DESIGNING DISCOVERY

Hypergraph Models of Innovation for Science & Technology

KNOWLEDGE LAB

James Evans
Feng (Bill) Shi

KNOWLEDGE LAB

Big Data, Machine Learning and Intelligent Crowdsourcing enables us to:

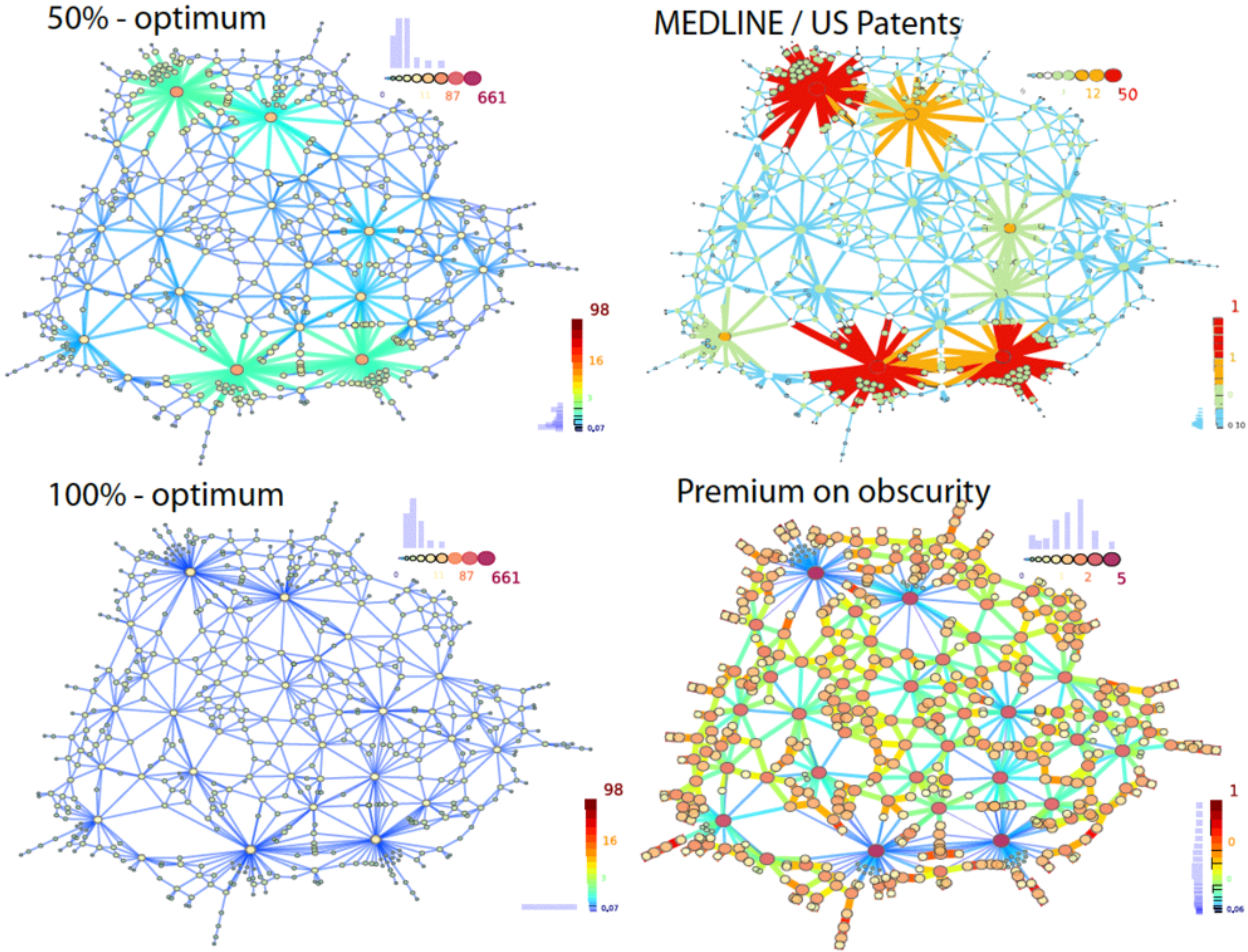
1. Trace
2. Understand
3. Discover
4. Improve…the scientific and scholarly process

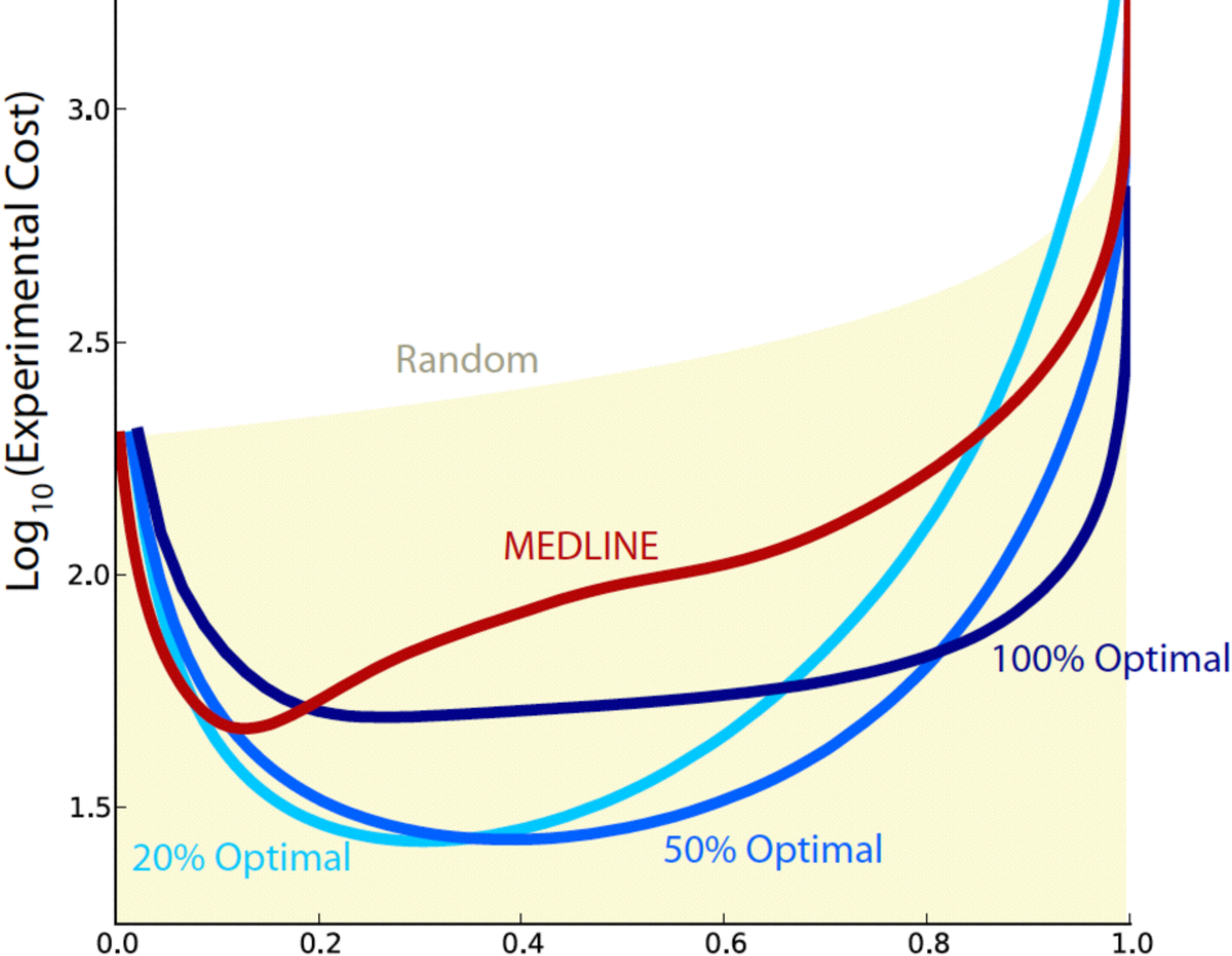
computationally enhanced

Science of Science

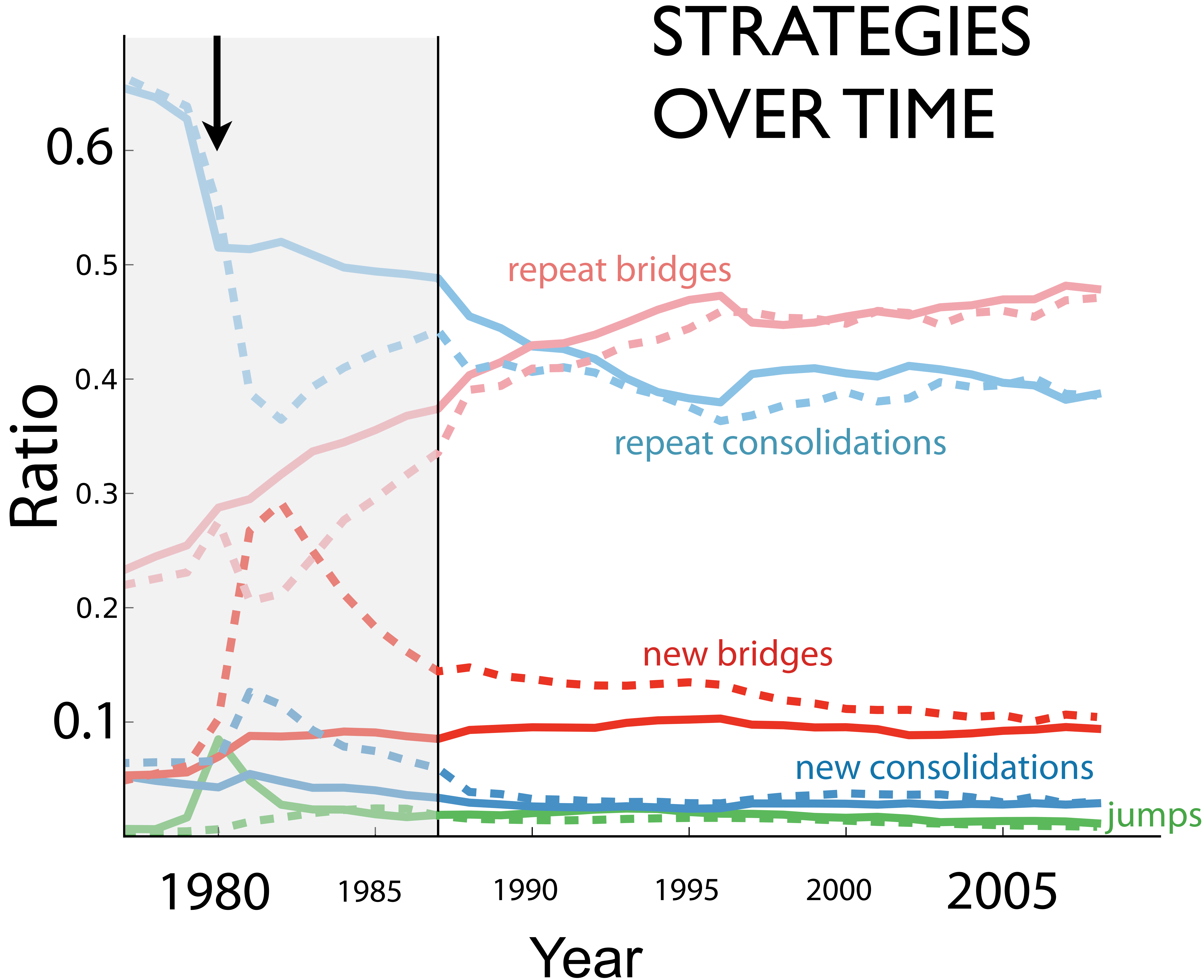
How Science Thinks

jevans@uchicago.edu


50% - optimum
87
661
98
16
0,07
MEDLINE / US Patents
12
50
1
1
100% - optimum
87
661
98
16
0,07
Premium on obscurity
2
5
1
0
0,06

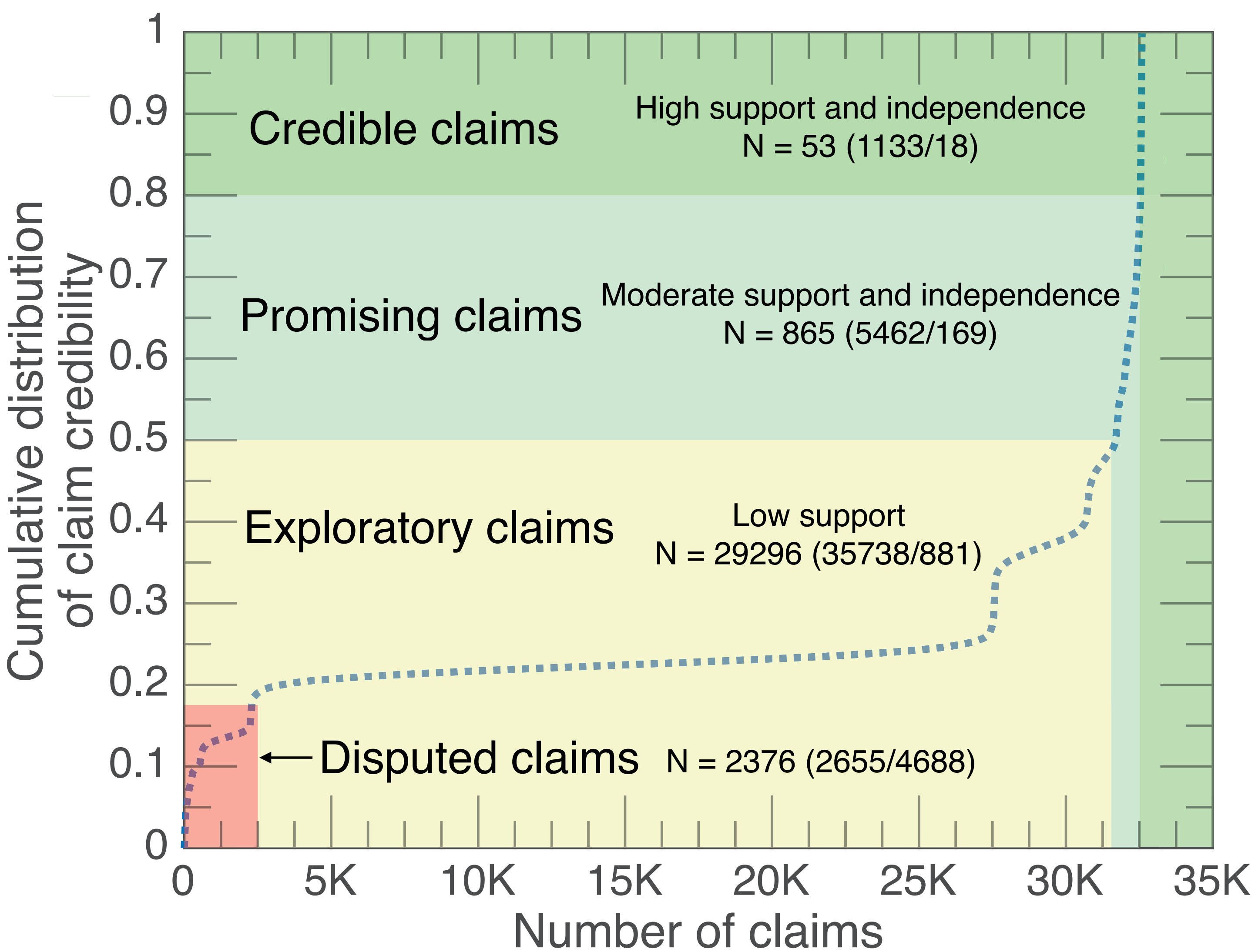


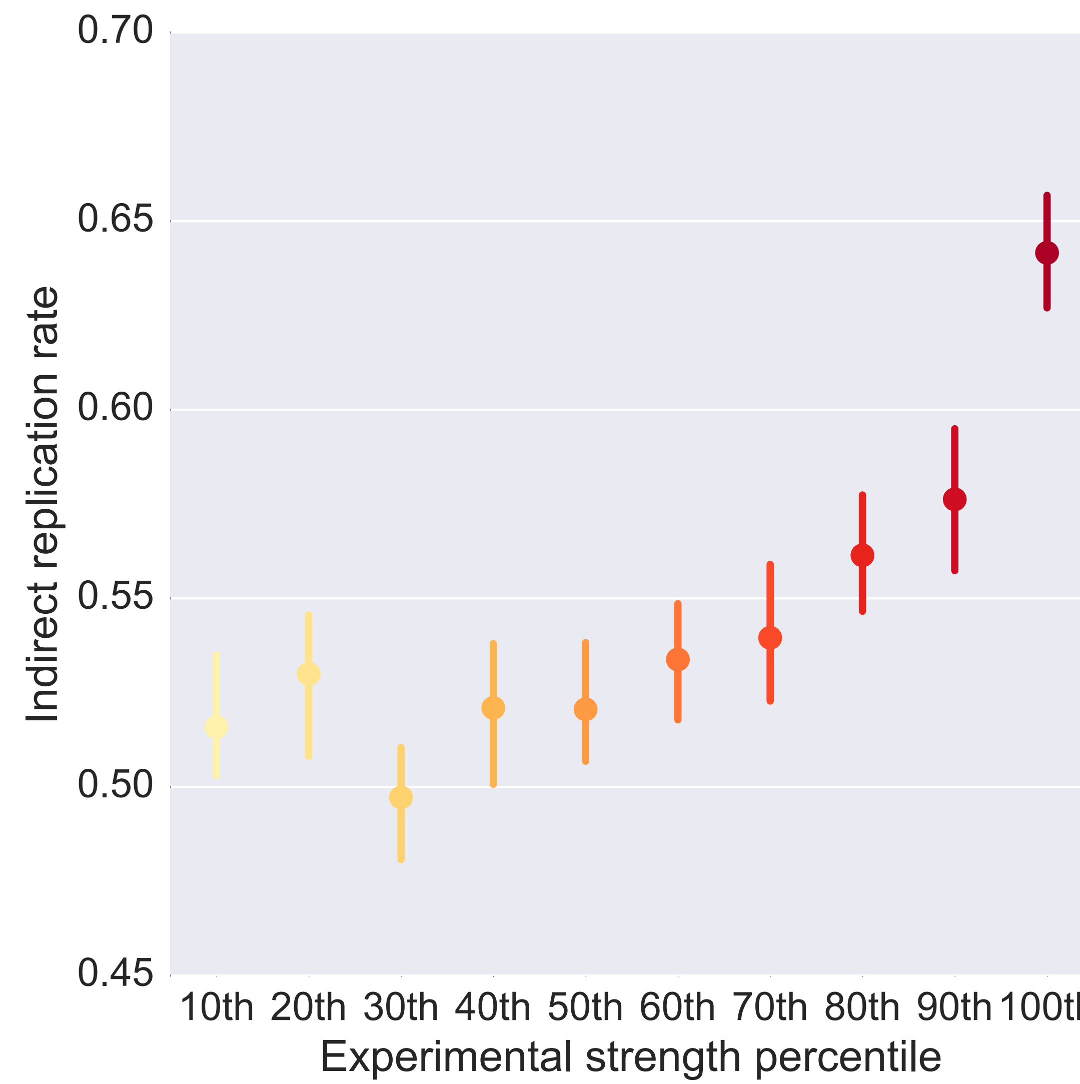
Log10(Experimental Cost)
3.0
2.5
2.0
1.5
Random
MEDLINE
100% Optimal
50% Optimal
20% Optimal
0.0
0.2
0.4
0.6
0.8
1.0

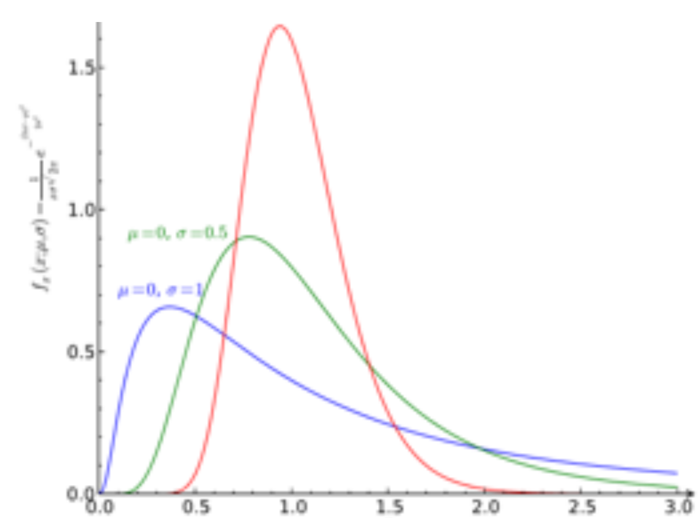


STRATEGIES
OVER TIME
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
repeat bridges
repeat consolidations
new bridges
new consolidations
jumps
1980
1985
1990
1995
2000
2005
Year

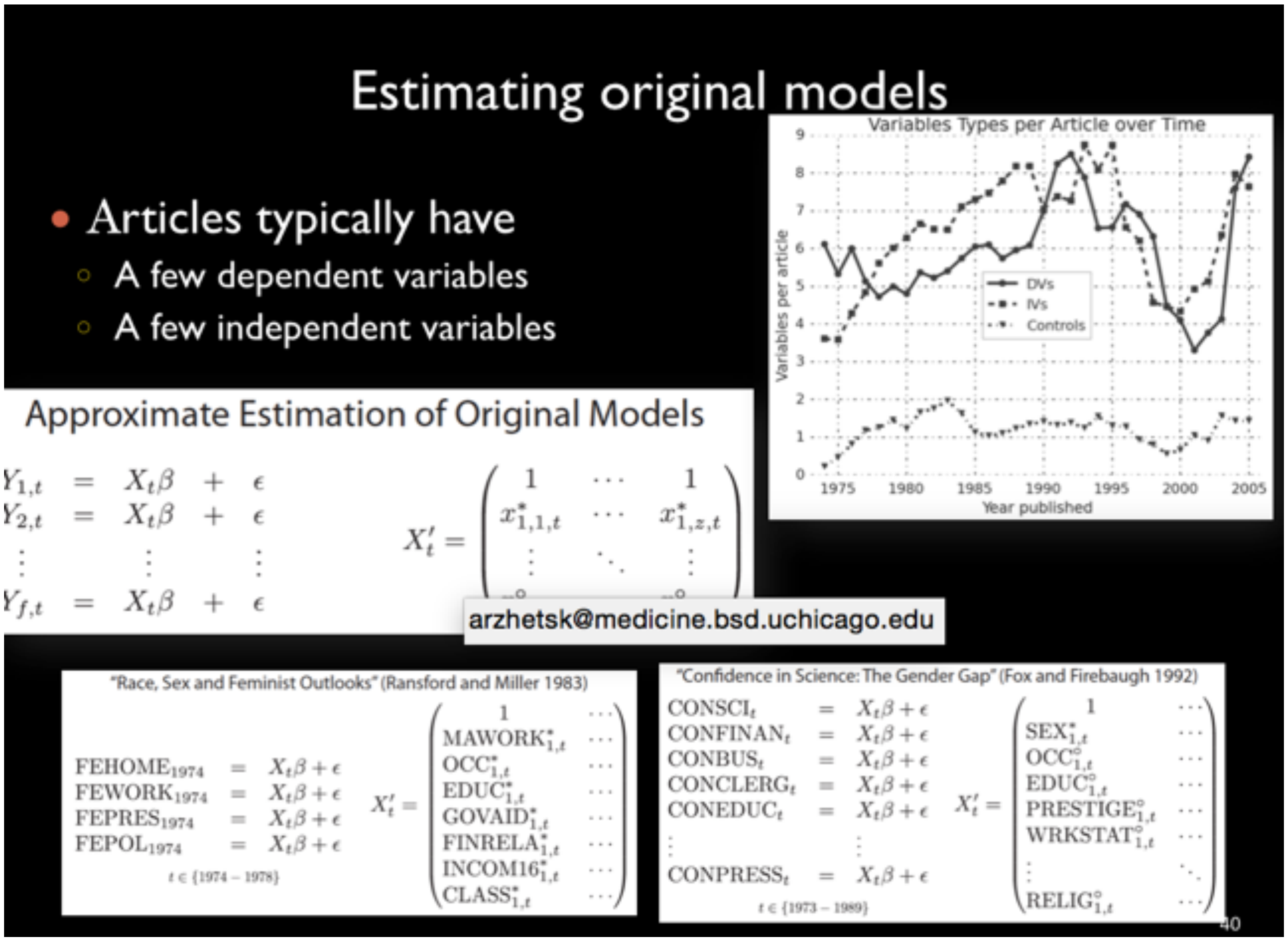
High-throughput quasi-replication

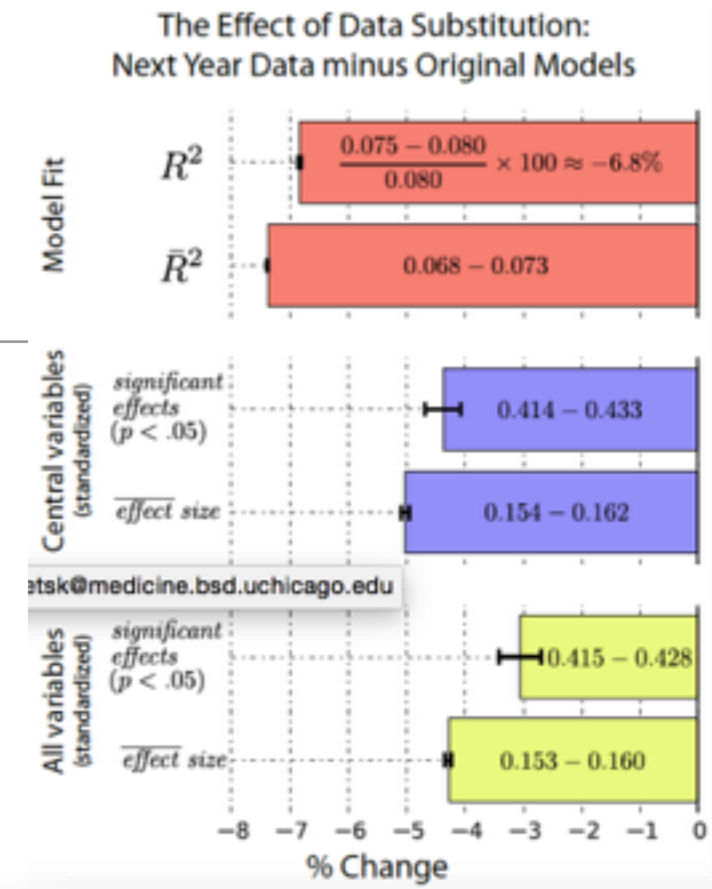


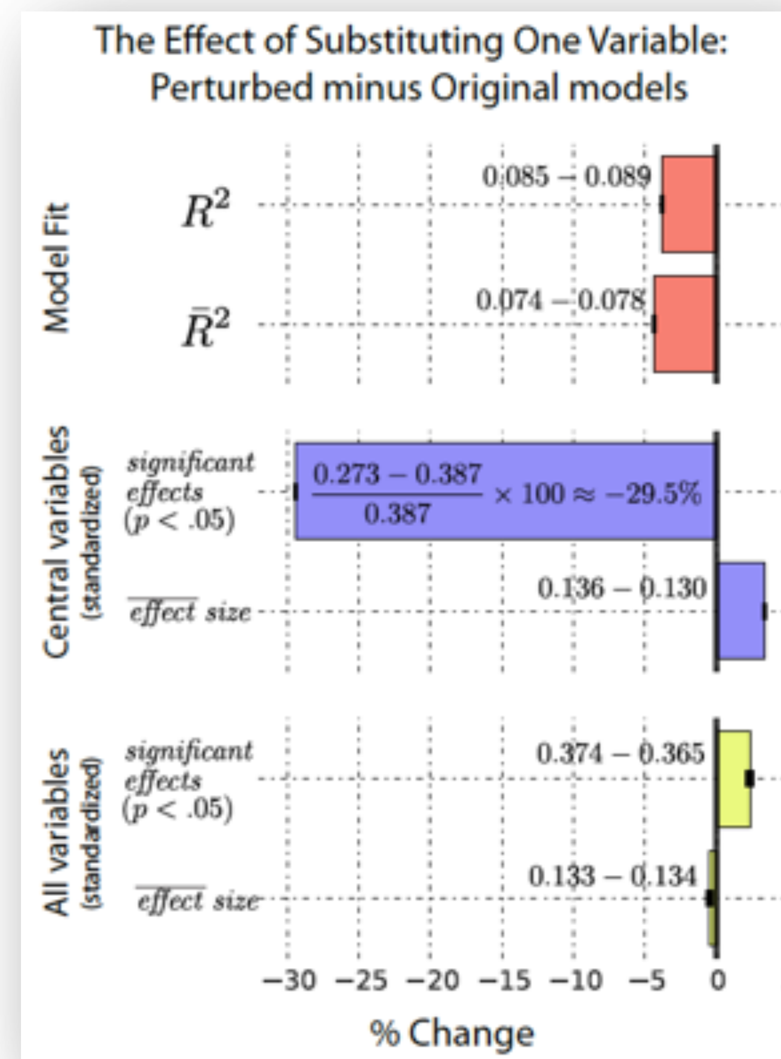




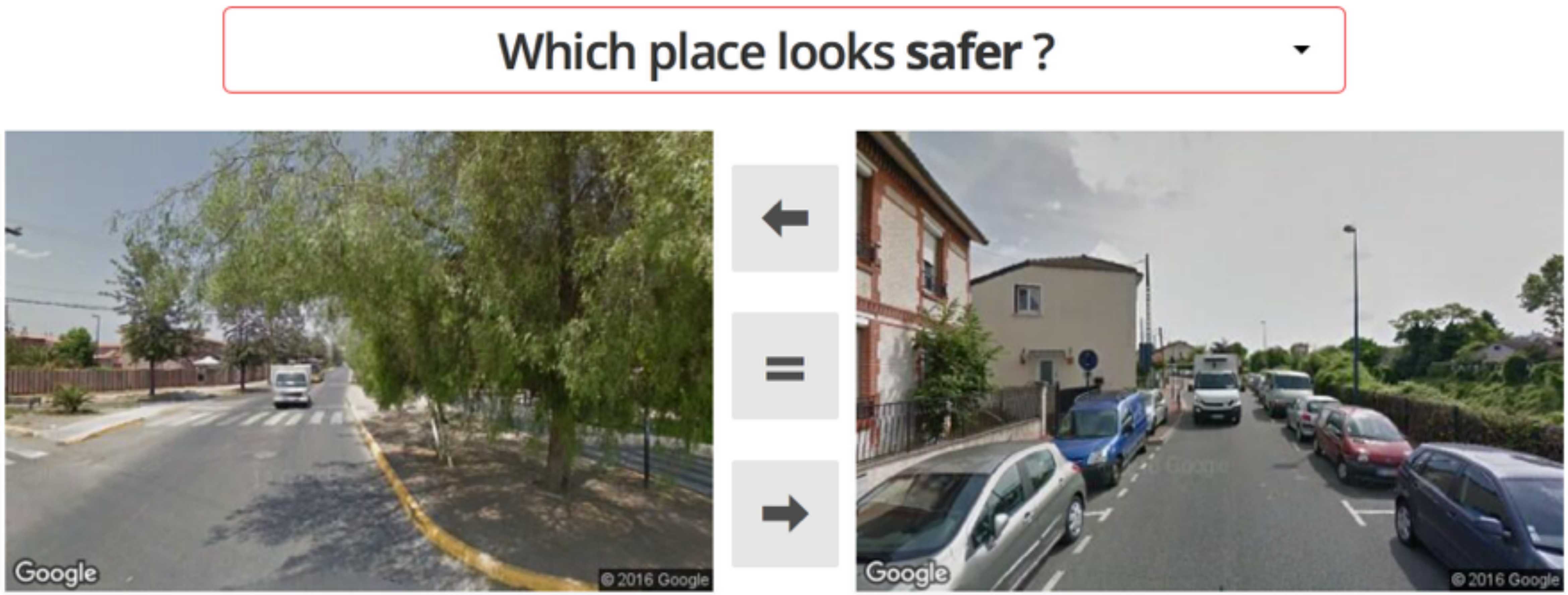
…and in the social sciences





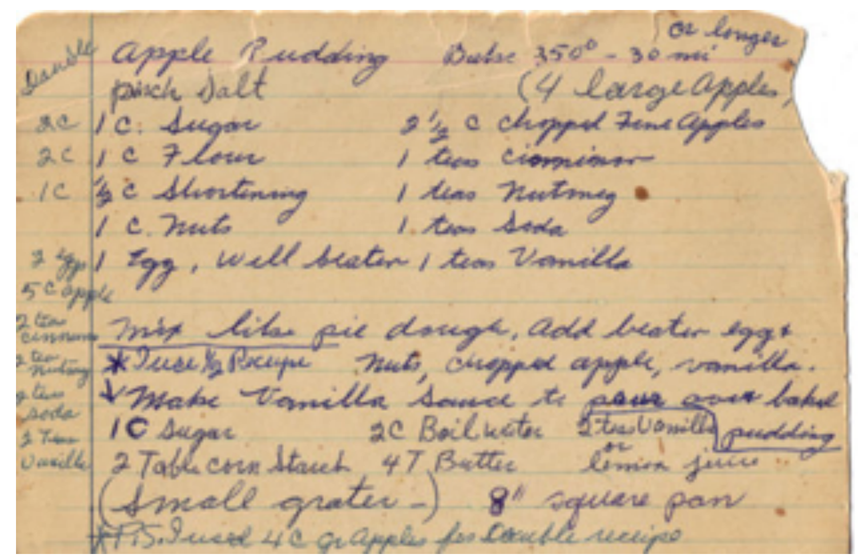


Active Learning for Intelligent Survey Design



PREDICTING & GENERATING SCIENTIFIC SUCCESS

- Predict **combination of concepts** in future discoveries & inventions



- Predict **level of impact** for future discoveries

KNOWLEDGE REPRESENTATION

Representation **Extraction** **Inference**

- Collocation Network / Adjacency Matrix

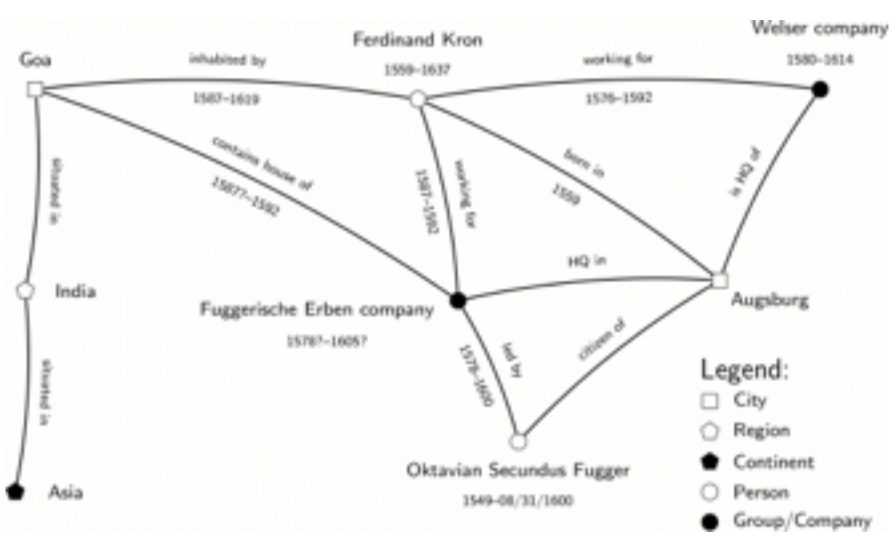
Inexpensive

Over

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & A_{2n} \\ \vdots & & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}$$

KNOWLEDGE REPRESENTATION

Representation	**Extraction**	**Inference**
• Collocation Network / Adjacency Matrix	*Inexpensive*	*Over*
• Semantic Graph or Hypergraph	*Expensive*	*Under*



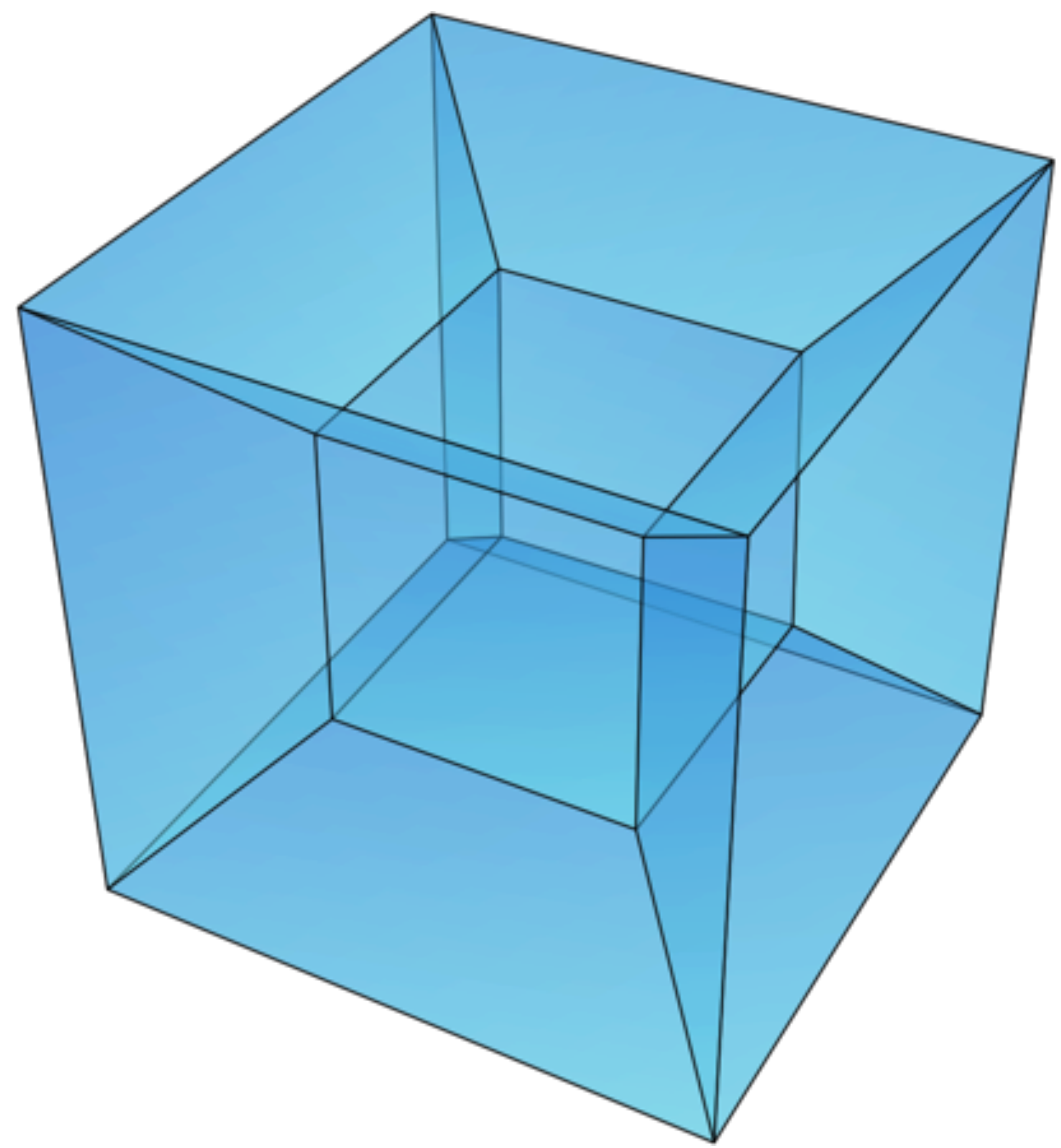
KNOWLEDGE REPRESENTATION

Representation	**Extraction**	**Inference**
• Collocation Network / Adjacency Matrix	*Inexpensive*	*Over*
• **Collocation Hypergraph / Adjacency Tensor**	*Inexpensive*	*Exact*
• Semantic Graph or Hypergraph	*Expensive*	*Under*

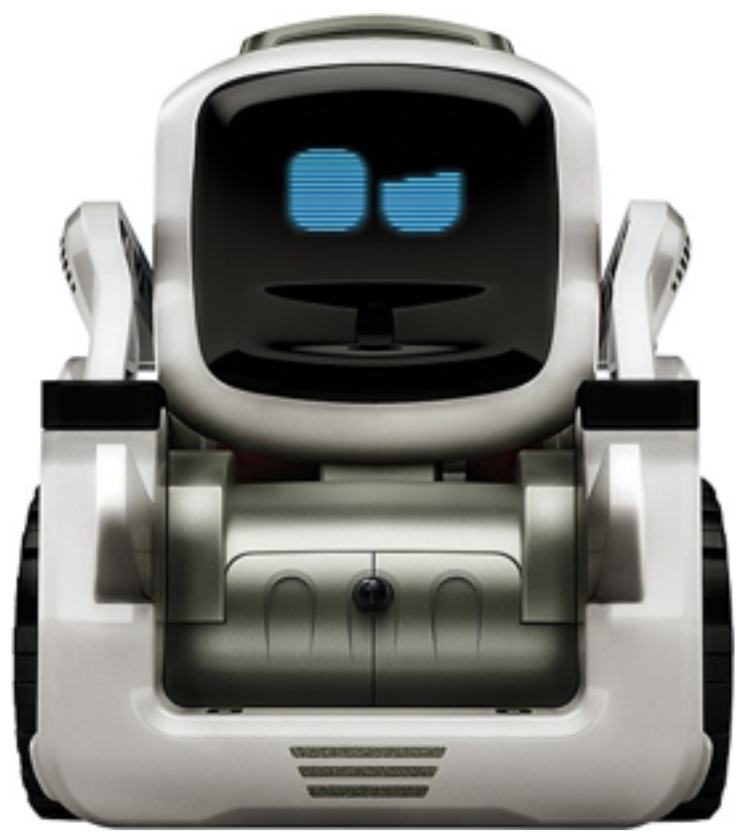
Graph vs. Hypergraph Matrix vs. Tensor

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & A_{2n} \\ \vdots & & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}$$

- Hypergraph CAN ALSO be rendered as a simple, 2-mode matrix where each set connects to each concept, but this removes all informational geometry
- Hypergraph rendered as a hyper-matrix or tensor retains its geometry, convexity, etc.



MODELING OPPORTUNITY



Automatically generate promising discoveries

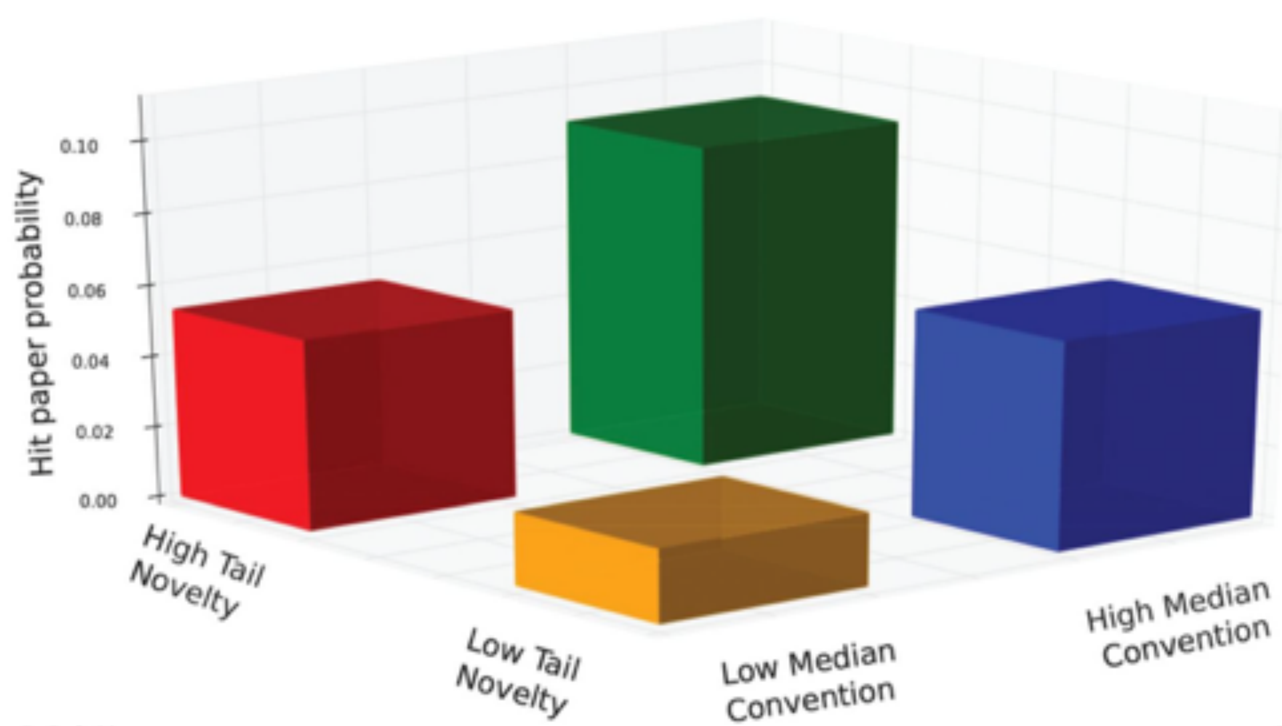
(or feed recipes to scientist chefs)

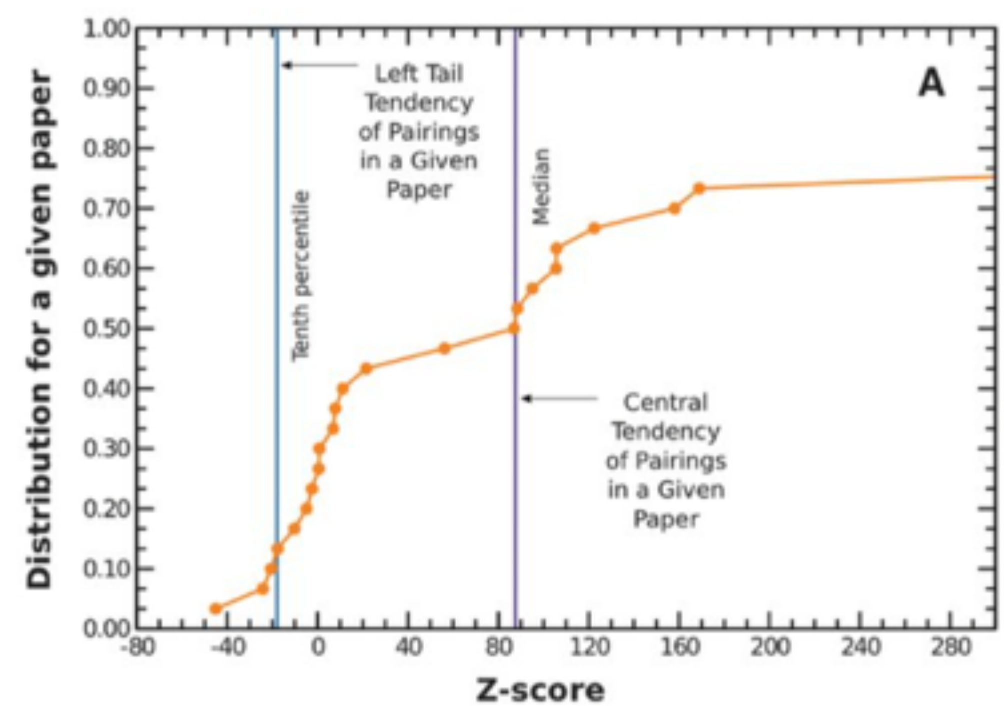
RELATED APPROACHES

REPORT

Atypical Combinations and Scientific Impact

Brian Uzzi[1,2], Satyam Mukherjee[1,2], Michael Stringer[2,3], Ben Jones[1,4,*]





CONCEPTS = CITED JOURNALS

COMBINATIONS = PAIRWISE FREQUENCY DISTRIBUTION

PREDICTION $< 10\%$

OUR PROJECT

CONCEPTS = CONCEPTS

CONTEXTS = JOURNALS

COMBINATIONS = COMPLETE COMBINATION

PREDICTION > 40%

Mixed-Membership, High-Dimensional Block Model

1 $(\theta_{11}, \theta_{12})$

2 $(\theta_{21}, \theta_{22})$

3 $(\theta_{31}, \theta_{32})$

Propensity that this combination will turn into a paper:

$$\lambda = (\theta_{11}\theta_{21}\theta_{31} + \theta_{12}\theta_{22}\theta_{32})r_1 r_2 r_3$$

popularity of node i

Number of papers on this combination: $X \sim Poisson(\lambda)$

Generative Model for Hypergraph

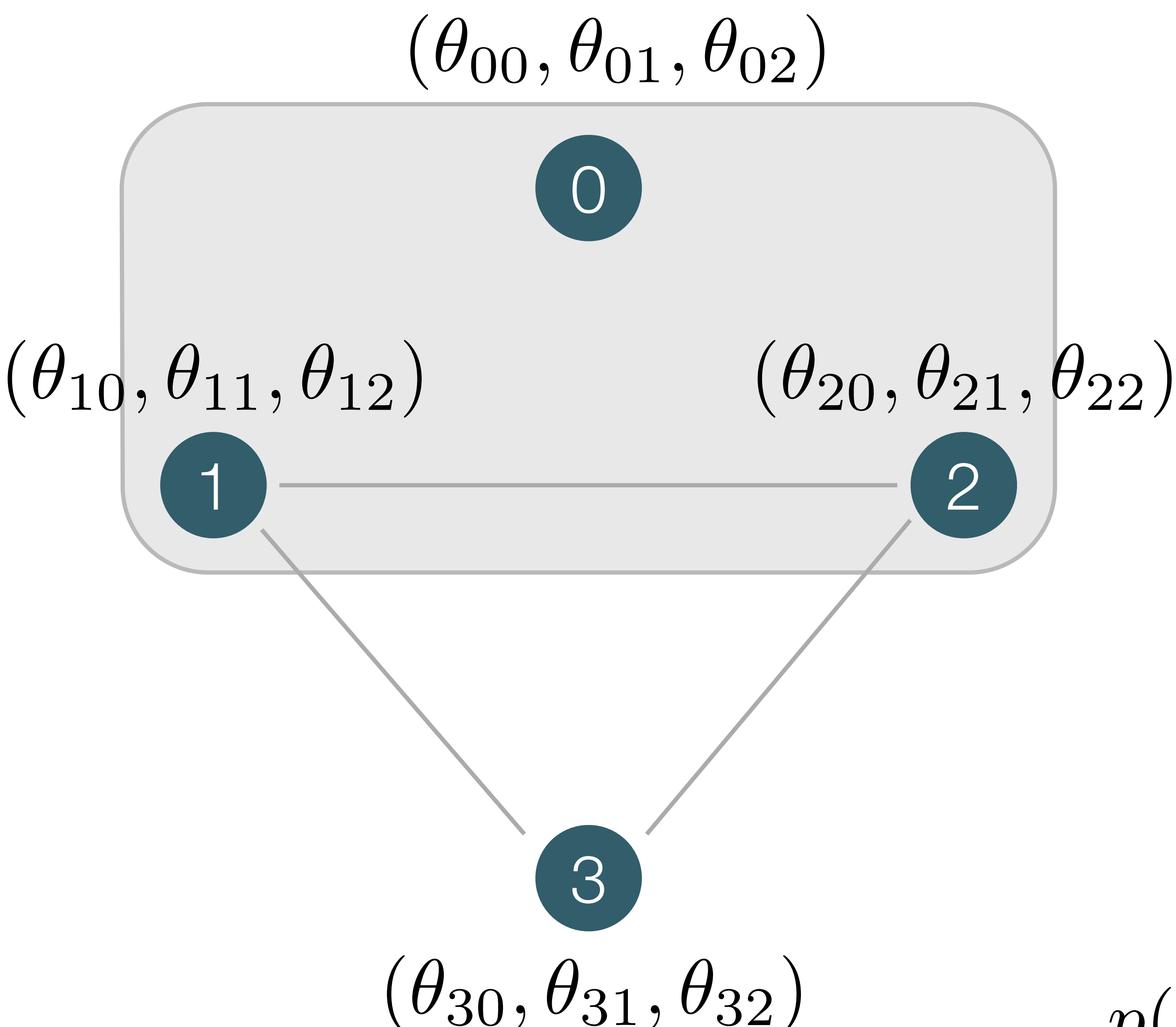
- For any combination h of nodes
- Calculate propensity $\lambda_h = \sum_k \prod_{i \in h} r_i \theta_{ik}$
- Draw the number of hyperedges of h from

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

- Likelihood to generate the hypergraph

$$p(G|\theta, r) = \prod_{h \in H} p(x_h|\theta, r)$$

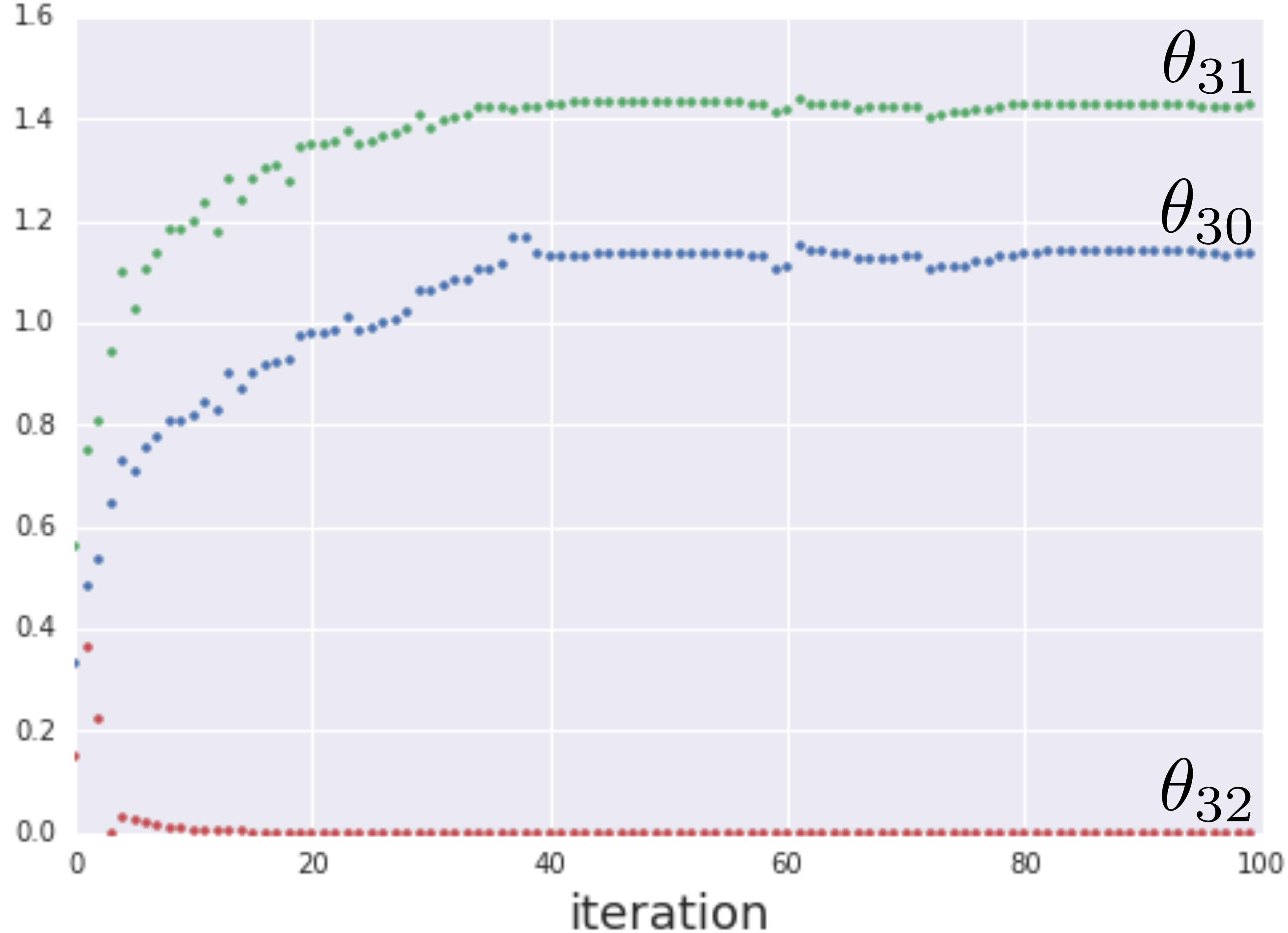
Toy Example



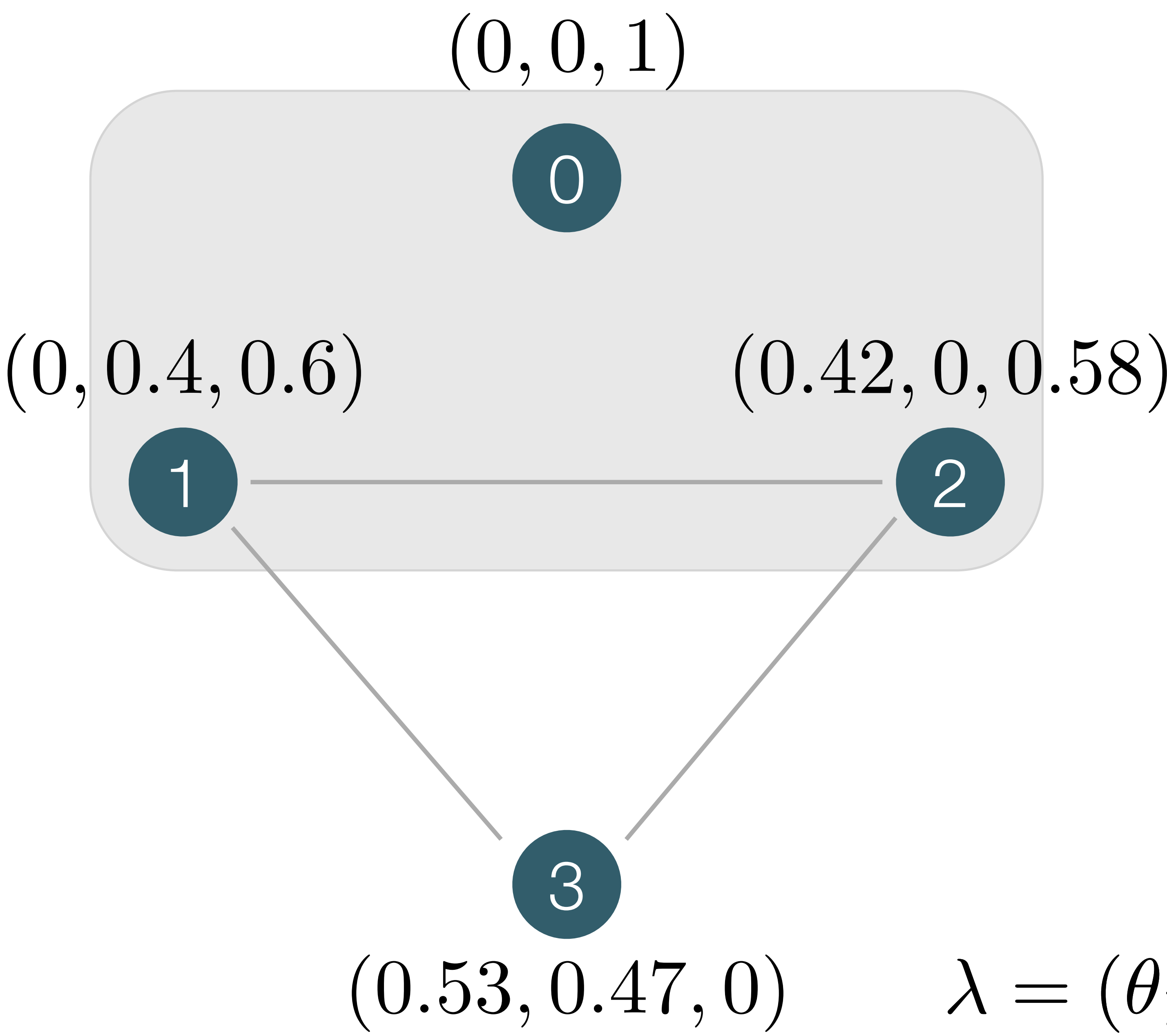
h	X_h
$\{0, 1, 2\}$	1
$\{1, 2\}$	1
$\{1, 3\}$	1
$\{2, 3\}$	1
$\dots$	0

$$p(G|\theta, r) = \prod_{h \in H} p(x_h|\theta, r)$$

Toy Example — Convergence



Toy Example

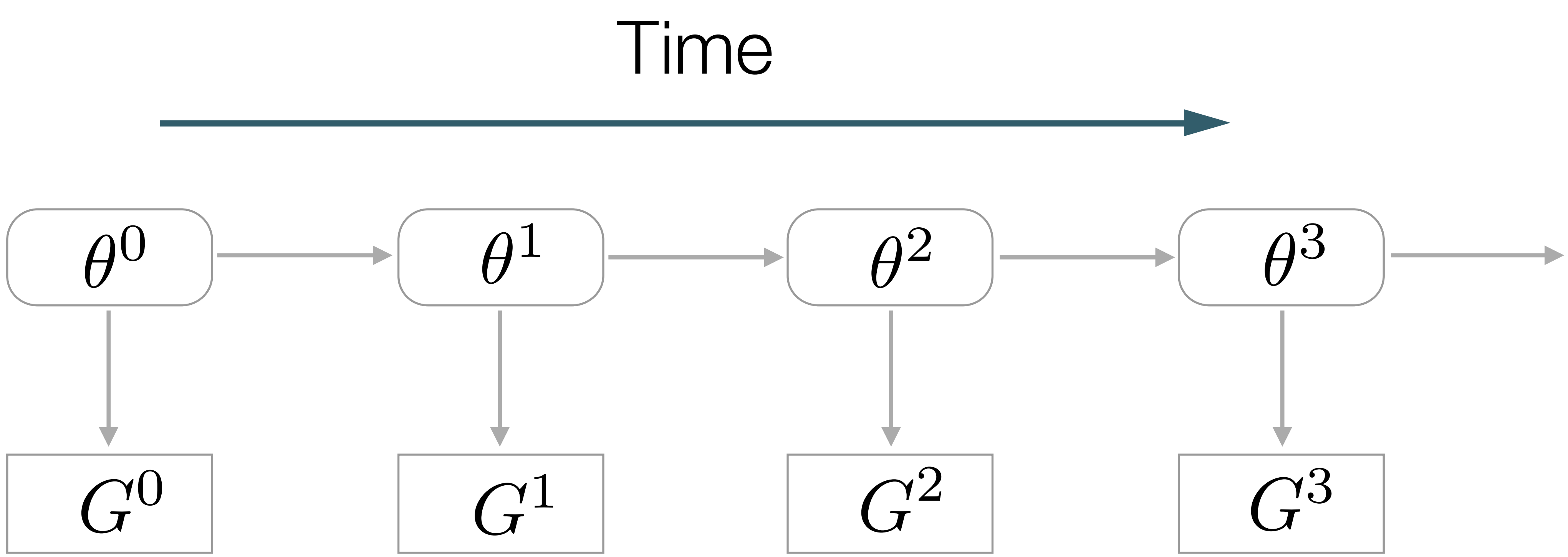


h	X_h
$\{0, 1, 2\}$	1
$\{1, 2\}$	1
$\{1, 3\}$	1
$\{2, 3\}$	1
$\ldots$	0

$$\lambda = (\theta_{11}\theta_{21}\theta_{31} + \theta_{12}\theta_{22}\theta_{32})r_1 r_2 r_3$$

Evolution of the Network

Hidden Markov Model



G^t: observed network at time t

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

Complete Model

- Log-likelihood function

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

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

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

- Impossible to optimize!

Incomputable

2^N possible combinations

Maximal Likelihood Estimate

Algorithm

- Generate *t* from 1,….,T uniformly at random
- For $d = 2, ..., D$, pick a random set H_d^t of combinations of order d from G^t.
- Calculate $S_d^t = \sum_{h \in H_d^t} [x_h \log \sum \prod \theta_{ik}^t - \sum \prod \theta_{ik}^t]$
- Approximate $\nabla l(\theta)$ by $\nabla(\sum_d S_d^t)$
- Update $\hat{\theta} = \hat{\theta} + \eta \nabla l(\hat{\theta})$

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.

PubMed

APS
physics™

PubMed

APS physics™

UNITED STATES
PATENT AND TRADEMARK OFFICE

uspto

Datasets

20M PubMed articles (1865 to 2015)

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

.5M APS articles (1880-2015)

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

1.5M US Patents

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

available at www.sciencedirect.com

journal homepage: www.elsevier.com/locate/peptides

NAP protects hippocampal neurons against multiple toxins

Ilona Zemlyak [a,b], *Nathan Manley* [b], *Robert Sapolsky* [b], *Illana Gozes* [a,*]

[a] *Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Israel*
[b] *Department of Biological Sciences, Stanford University, Stanford, USA*

ARTICLE INFO

Article history:
Received 1 July 2007
Received in revised form
6 August 2007
Accepted 7 August 2007
Published on line 11 August 2007

Keywords:
NAP
Kainic acid
Oxygen-glucose deprivation
Sodium cyanide
Neuronal death

ABSTRACT

The femtomolar-acting protective peptide NAP (NAPVSIPQ), derived from activity-dependent neuroprotective protein (ADNP), is broadly neuroprotective in vivo and in vitro in cerebral cortical cultures and a variety of cell lines. In the present study, we have extended previous results and examined the protective potential of NAP in primary rat hippocampal cultures, using microtubule-associated protein 2 (MAP2) as a measure for neuroprotection. Results showed that NAP, at femtomolar concentrations, completely protected against oxygen-glucose deprivation, and cyanide poisoning. Furthermore, NAP partially protected against kainic acid excitotoxicity. In summary, we have significantly expanded previous findings in demonstrating here direct neuroprotective effects for NAP on vital hippocampal neurons that are key participants in cognitive function in vivo.

available at www.sciencedirect.com

journal homepage: www.elsevier.com/locate/peptides

NAP protects hippocampal neurons against multiple toxins

Ilona Zemlyak [a,b], *Nathan Manley* [b], *Robert Sapolsky* [b], *Illana Gozes* [a,*]

[a] *Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Israel*
[b] *Department of Biological Sciences, Stanford University, Stanford, USA*

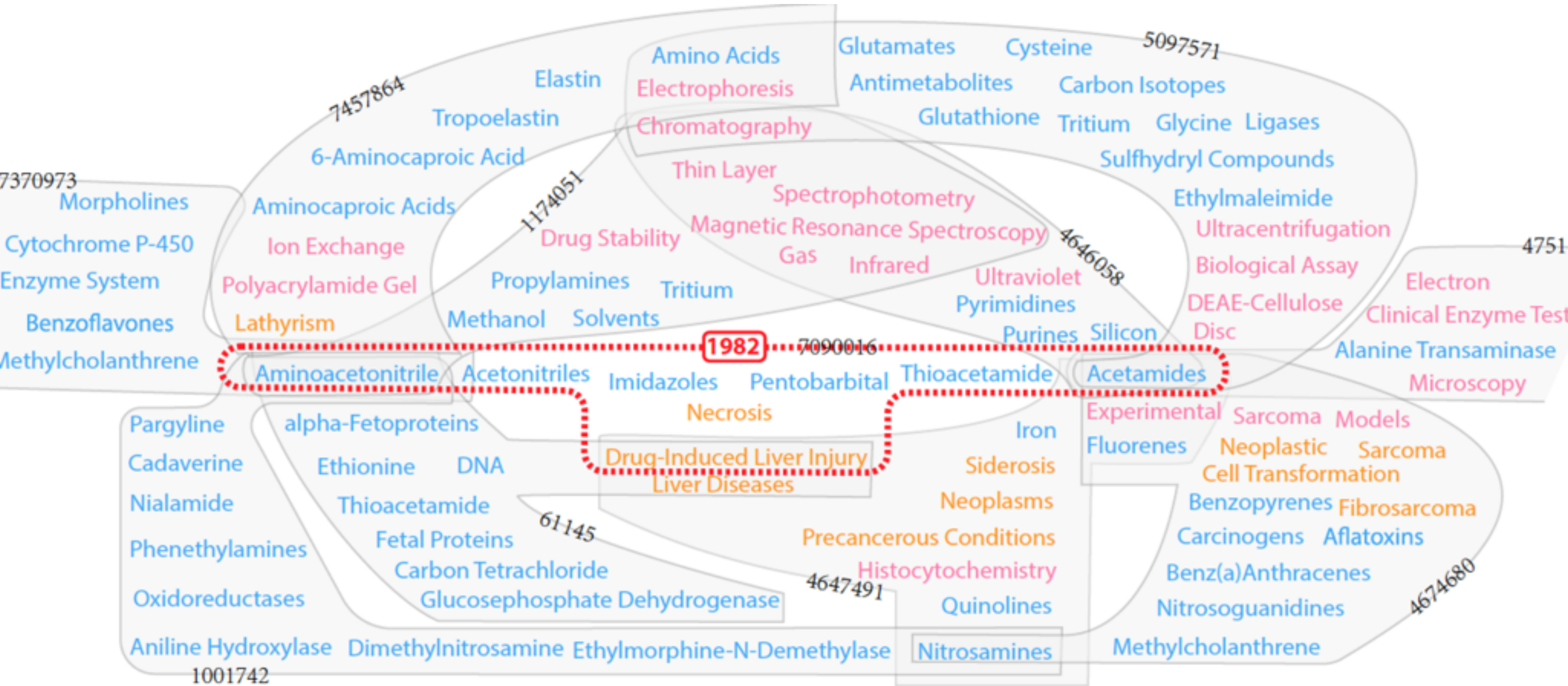
A R T I C L E I N F O

Article history:
Received 1 July 2007
Received in revised form
6 August 2007
Accepted 7 August 2007
Published on line 11 August 2007

Keywords:
NAP
Kainic acid
Oxygen-glucose deprivation
Sodium cyanide
Neuronal death

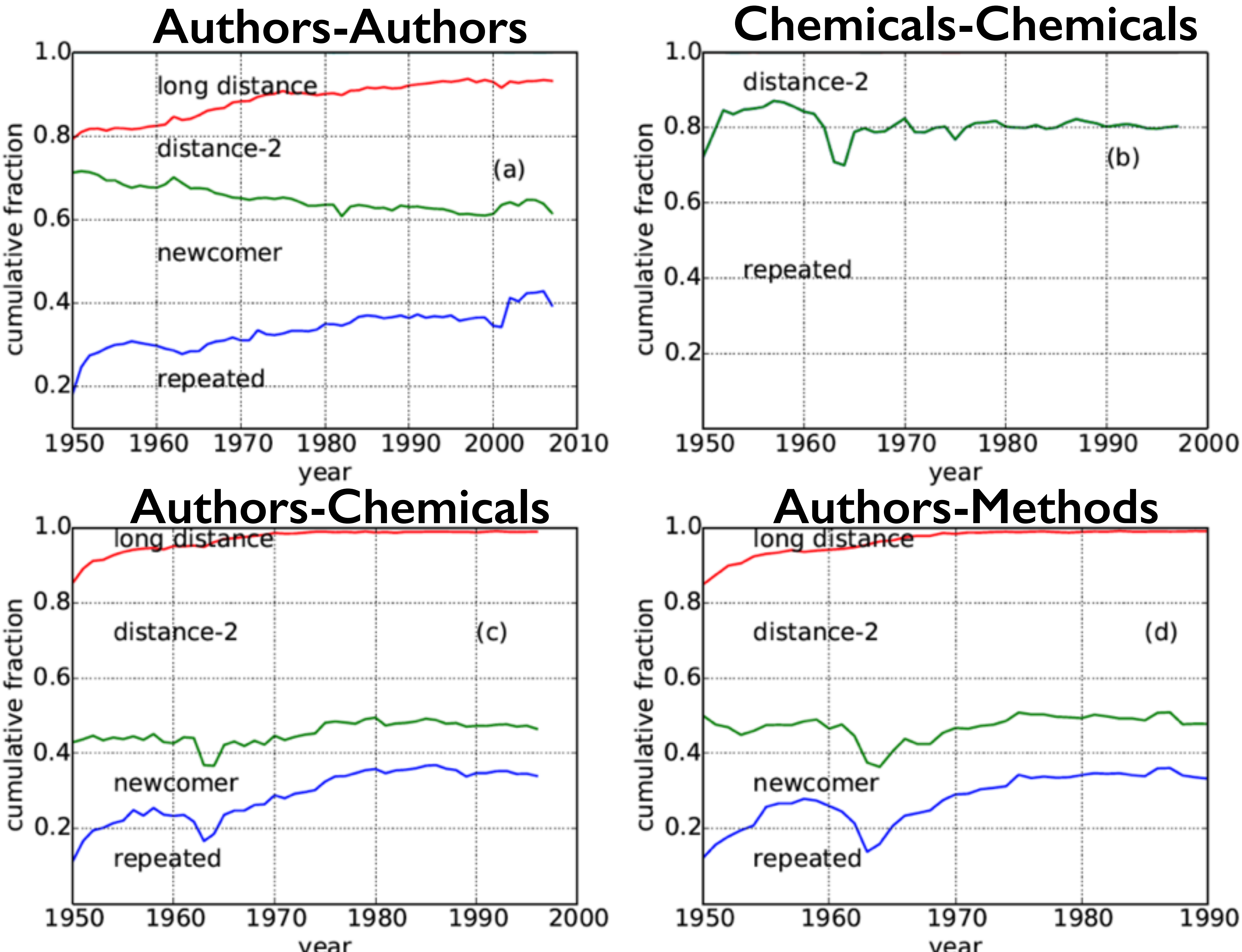
A B S T R A C T

The femtomolar-acting protective peptide NAP (NAPVSIPQ), derived from activity-dependent neuroprotective protein (ADNP), is broadly neuroprotective in vivo and in vitro in cerebral cortical cultures and a variety of cell lines. In the present study, we have extended previous results and examined the protective potential of NAP in primary rat hippocampal cultures, using microtubule-associated protein 2 (MAP2) as a measure for neuroprotection. Results showed that NAP, at femtomolar concentrations, completely protected against oxygen-glucose deprivation, and cyanide poisoning. Furthermore, NAP partially protected against kainic acid excitotoxicity. In summary, we have significantly expanded previous findings in demonstrating here direct neuroprotective effects for NAP on vital hippocampal neurons that are key participants in cognitive function in vivo.

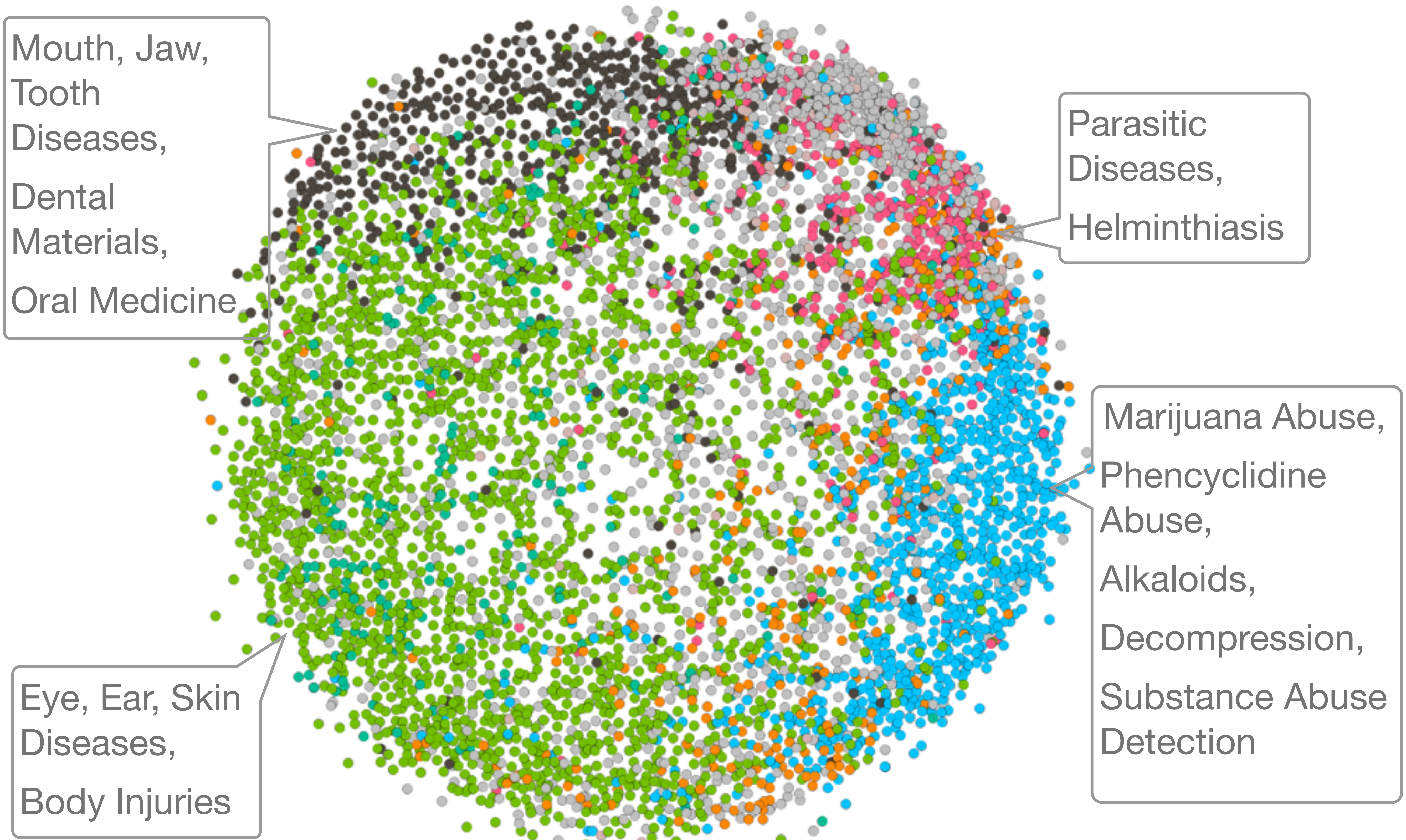


7457864
Elastin
Tropoelastin
6-Aminocaproic Acid
Aminocaproic Acids
Ion Exchange
Polyacrylamide Gel
Lathyrism
Amino Acids
Electrophoresis
Chromatography
Thin Layer
Spectrophotometry
Magnetic Resonance Spectroscopy
Gas
Infrared
Ultraviolet
Glutamates
Cysteine
5097571
Antimetabolites
Carbon Isotopes
Glutathione
Tritium
Glycine
Ligases
Sulfhydryl Compounds
Ethylmaleimide
Ultracentrifugation
Biological Assay
DEAE-Cellulose
Disc
1174051
Drug Stability
Propylamines
Tritium
Methanol
Solvents
4646058
Pyrimidines
Purines
Silicon
7370973
Morpholines
Cytochrome P-450
Enzyme System
Benzoflavones
Methylcholanthrene
4751
Electron
Clinical Enzyme Test
Alanine Transaminase
Microscopy
1982
7090016
Aminoacetonitrile
Acetonitriles
Imidazoles
Pentobarbital
Thioacetamide
Acetamides
Necrosis
Experimental
Sarcoma
Models
Pargyline
alpha-Fetoproteins
Iron
Fluorenes
Neoplastic
Sarcoma
Cadaverine
Ethionine
DNA
Drug-Induced Liver Injury
Liver Diseases
Siderosis
Cell Transformation
Nialamide
Thioacetamide
Neoplasms
Benzopyrenes
Fibrosarcoma
61145
Phenethylamines
Fetal Proteins
Precancerous Conditions
Carcinogens
Aflatoxins
Carbon Tetrachloride
Histocytochemistry
Benz(a)Anthracenes
4647491
4674680
Oxidoreductases
Glucosephosphate Dehydrogenase
Quinolines
Nitrosoguanidines
Aniline Hydroxylase
Dimethylnitrosamine
Ethylmorphine-N-Demethylase
Nitrosamines
Methylcholanthrene
1001742

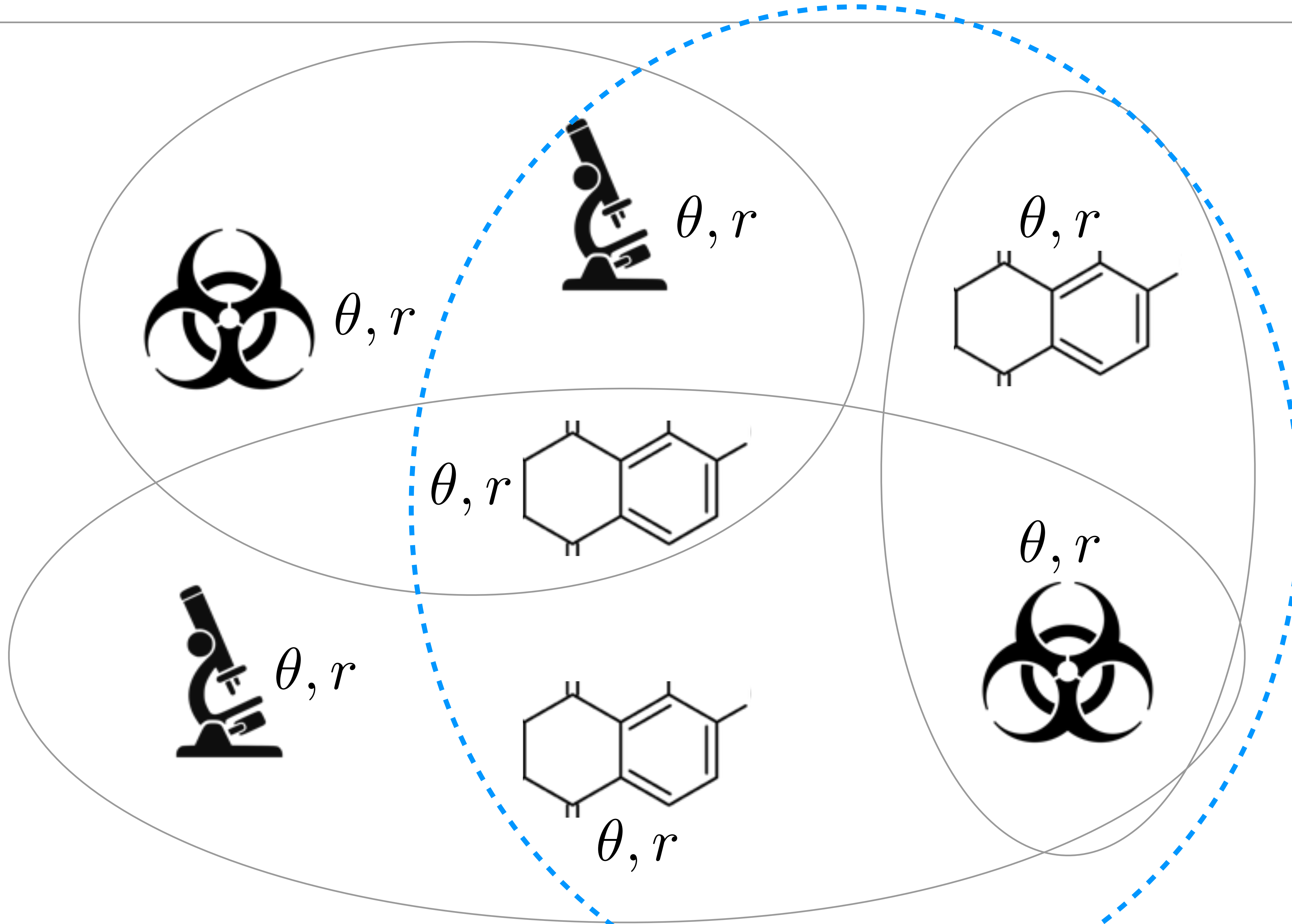
Network Distance



Community Structure



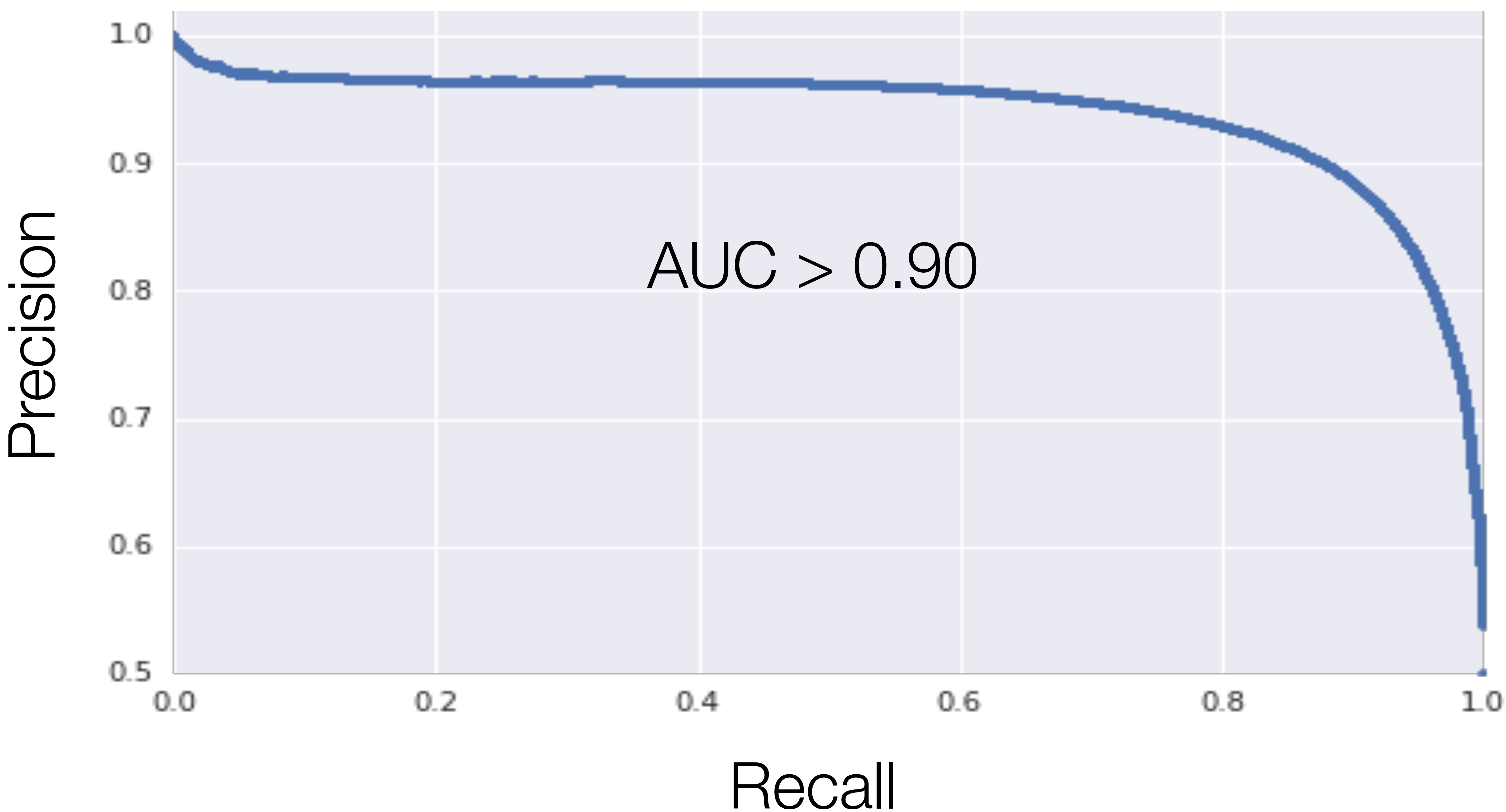
Predict New Hyperedges



$$\lambda = (\theta_{11}\theta_{21}\theta_{31} + \theta_{12}\theta_{22}\theta_{32})r_1 r_2 r_3$$

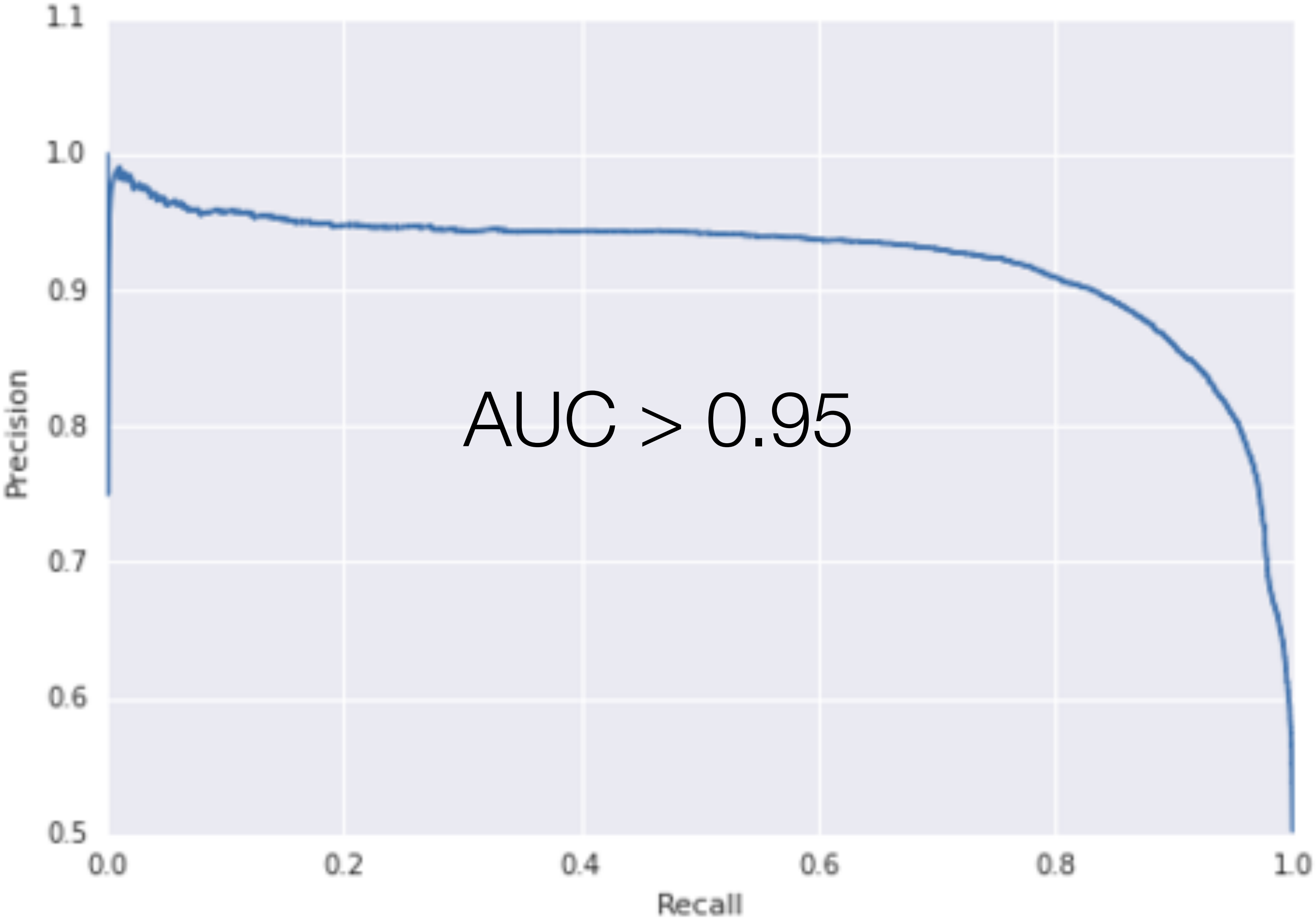
Predicting Papers

- Precision: out of all predicted papers, how many actually happen
- Recall: out of all future papers, how many are predicted to happen



Predicting Papers

- Precision: out of all predicted papers, how many actually happen
- Recall: out of all future papers, how many are predicted to happen

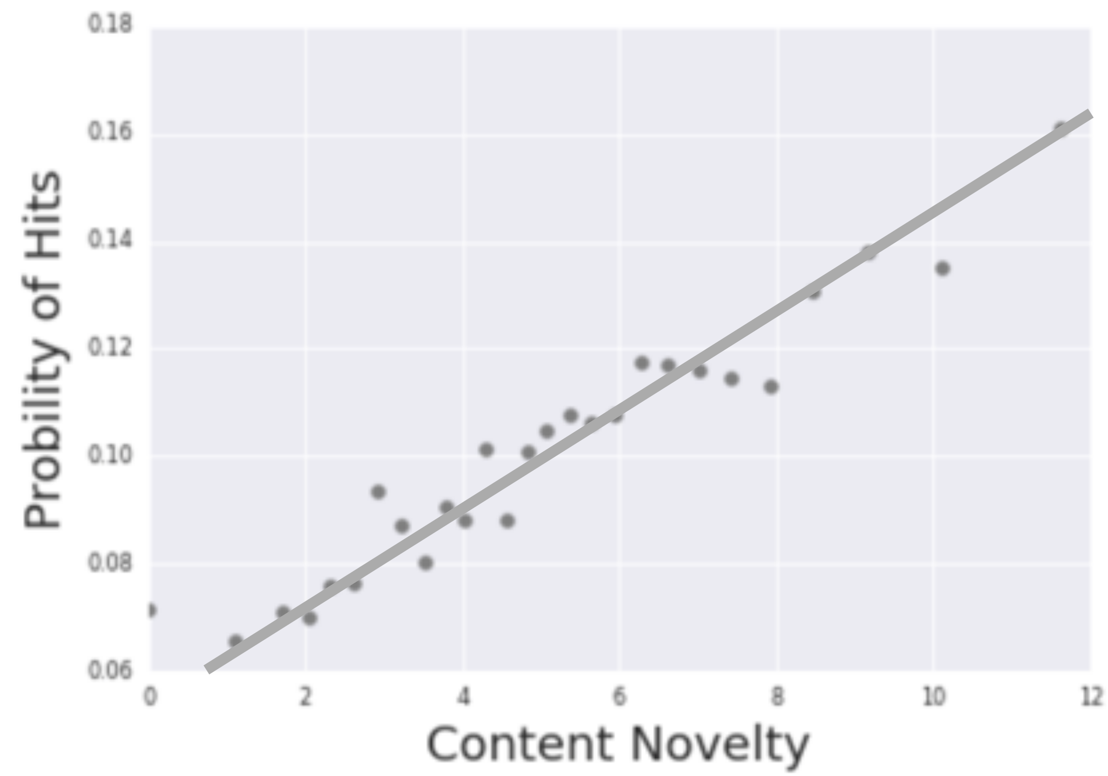


Precision

Recall

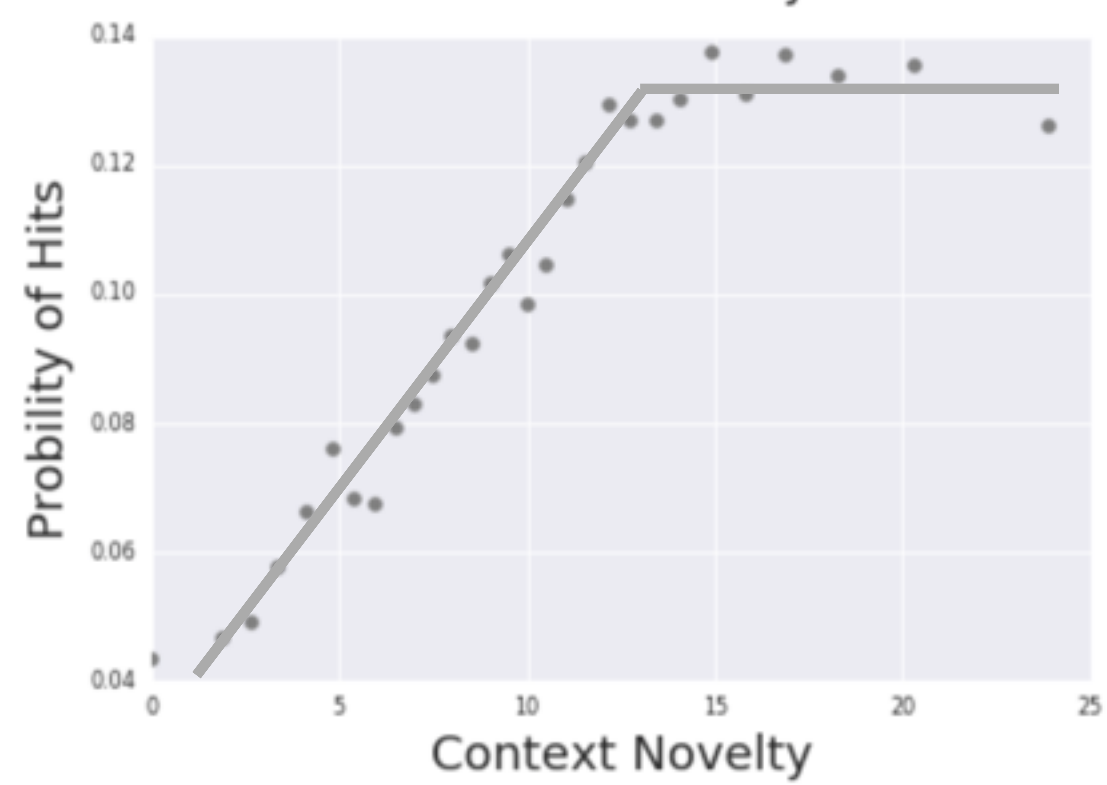
Novelty and Impact

Searching Broadly



Linear

Not Citing too Broadly

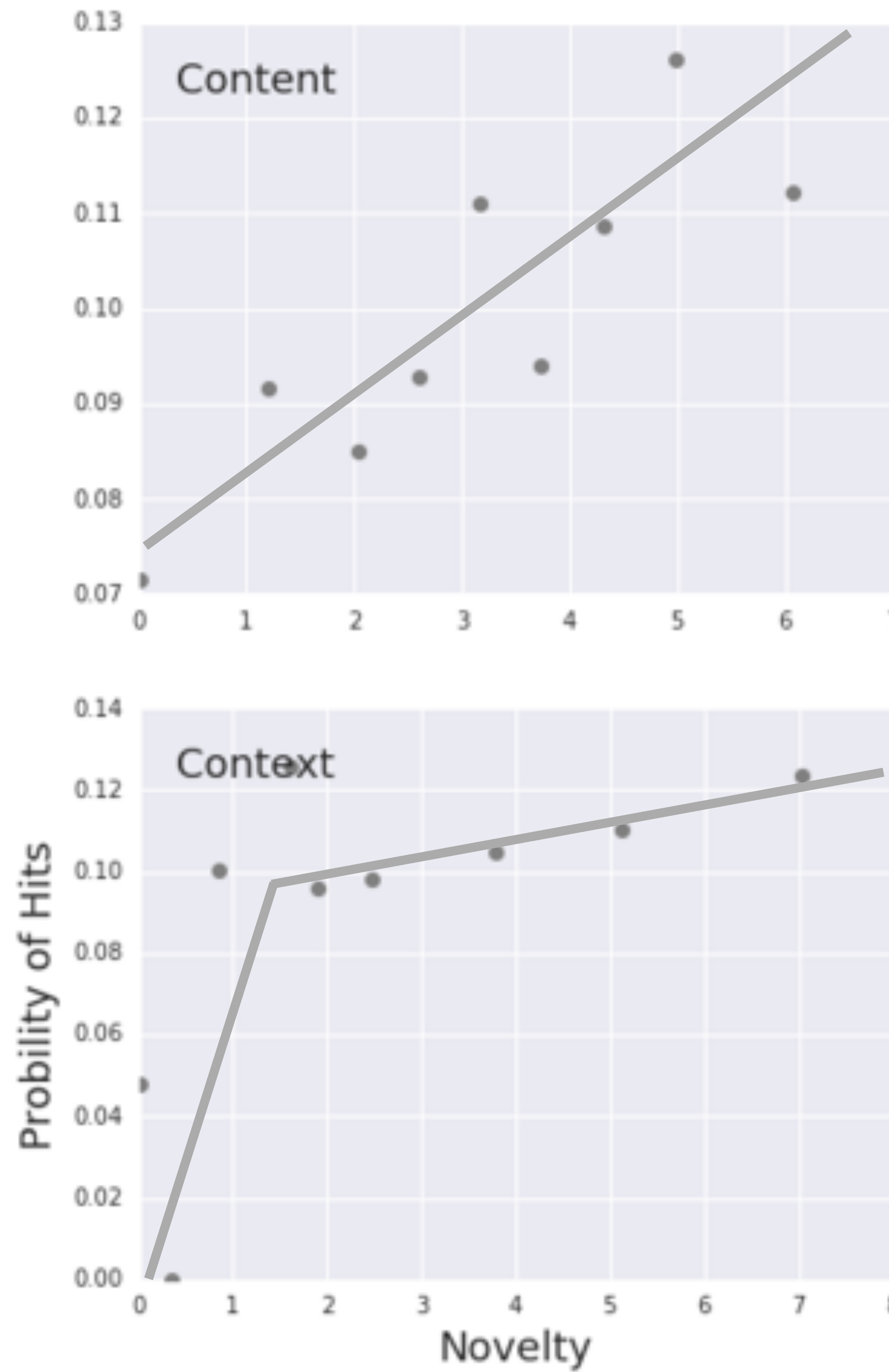


Threshold

Novelty and Impact

Searching Broadly

Not Citing too Broadly

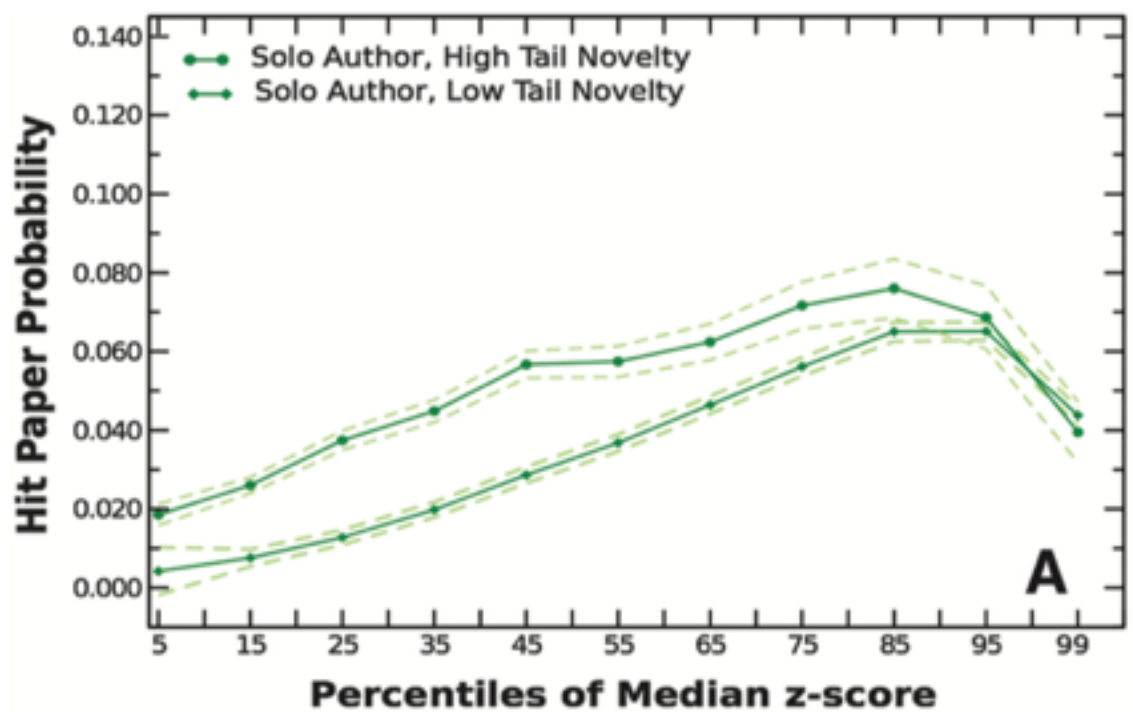


Content vs. Context

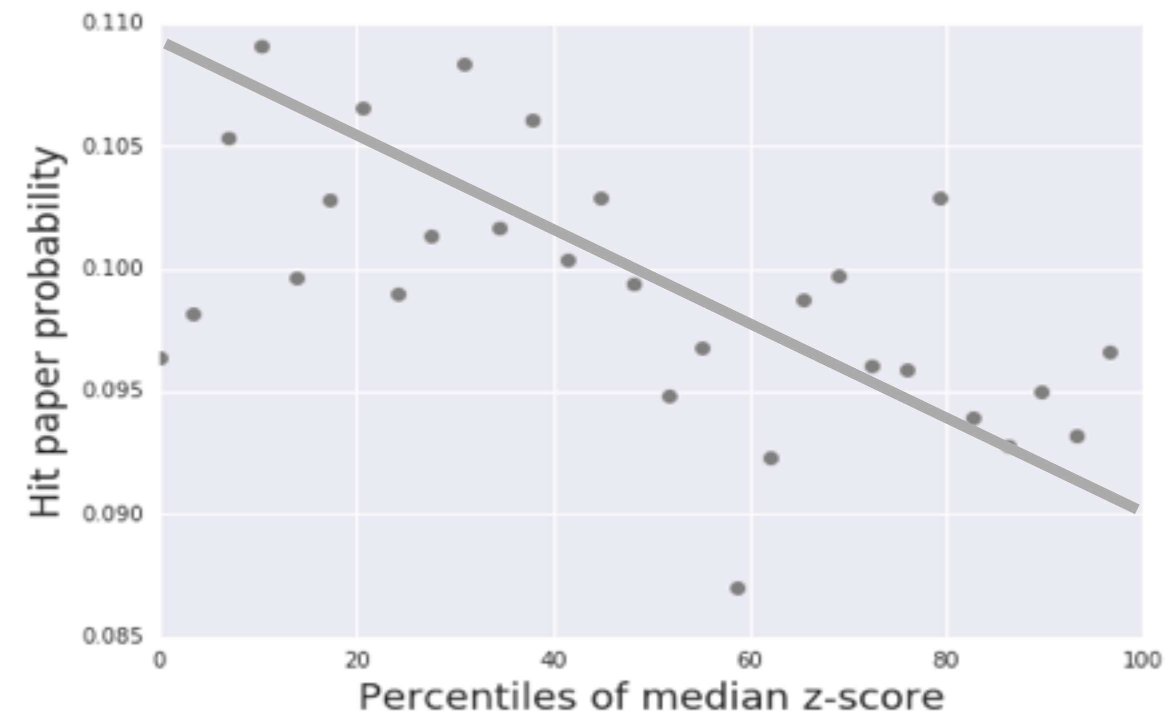
Content Correlates at **<.1** with Context

Context does NOT proxy for Content

Context vs. Content



Impact increases with conventionality of journal combinations [Uzzi, et al.]

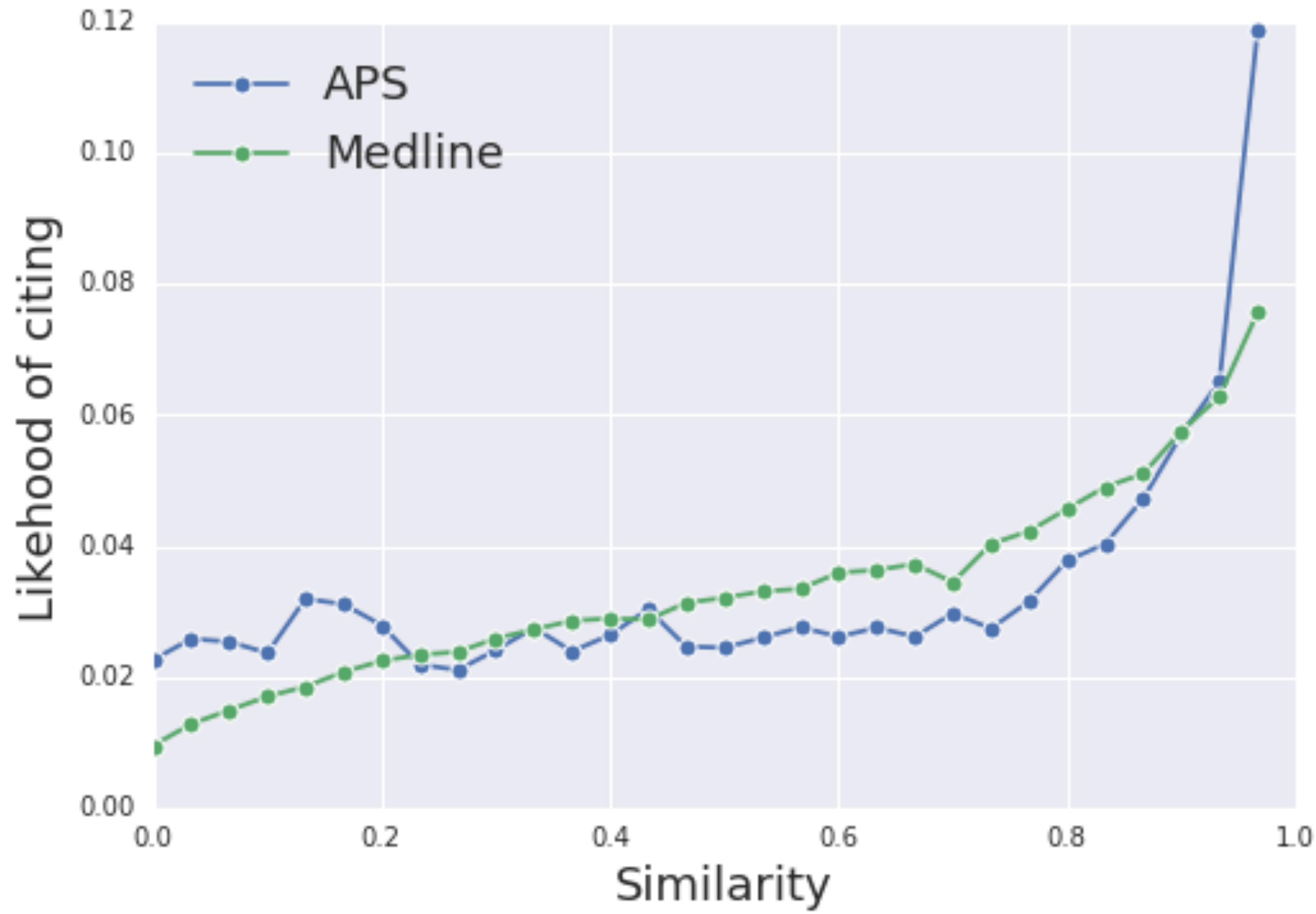


Impact (weakly) decreases with conventionality of MeSH term combinations.

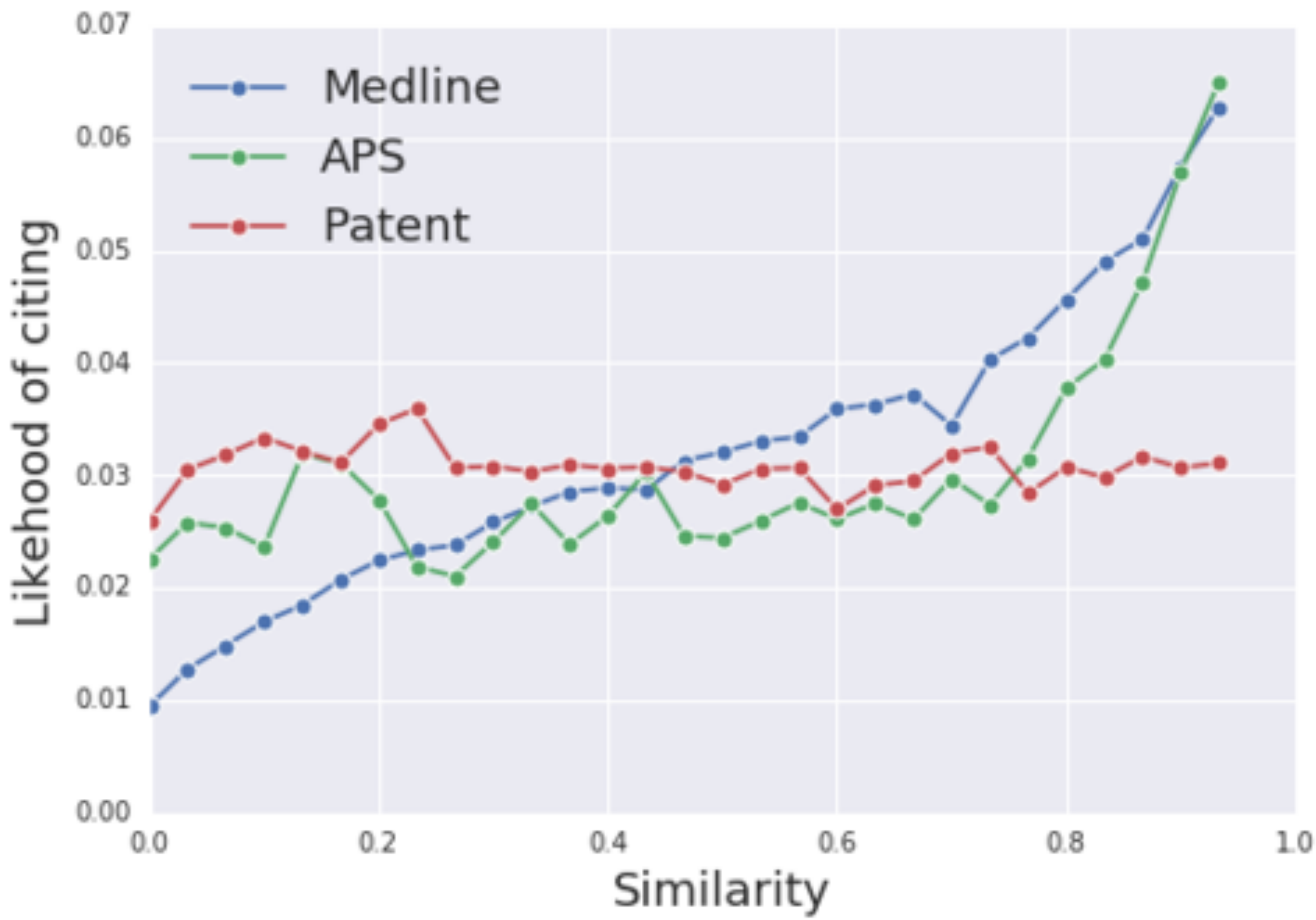
Same methodology [Uzzi, et al.], but contradictory results?

1. Technically: Not enough MeSH terms per paper to calculate median and tail.
2. Conceptually: Reference list is intended to situate a paper in the literature
3. Mismatch between content and context?

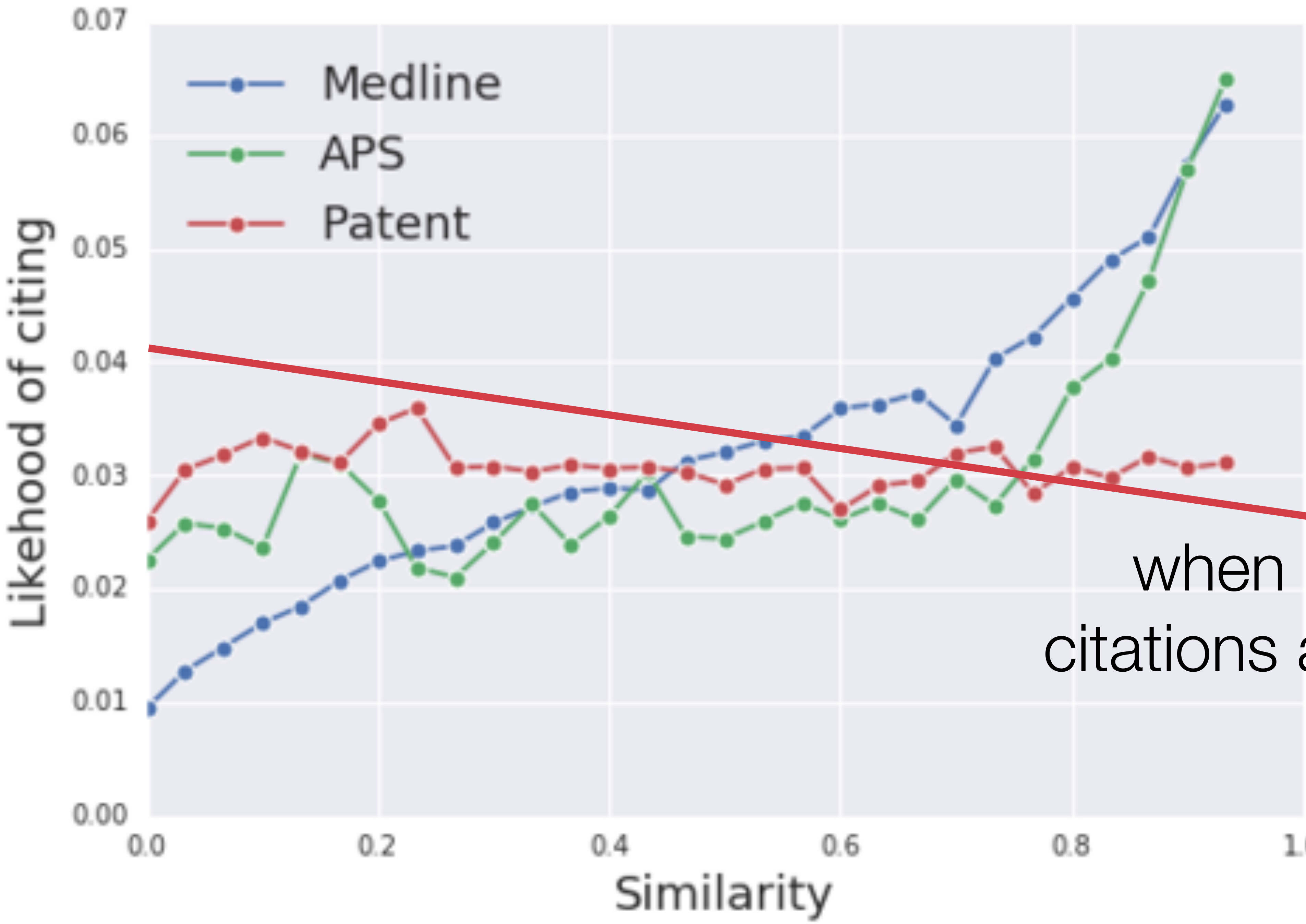
Humble Innovation:
Novel Exploration and Conservative Claims



Audacious Invention:
Novel Exploration and Outsized Claims



Audacious Invention: Novel Exploration and Outsized Claims



when examiner citations are removed

Content vs. Context

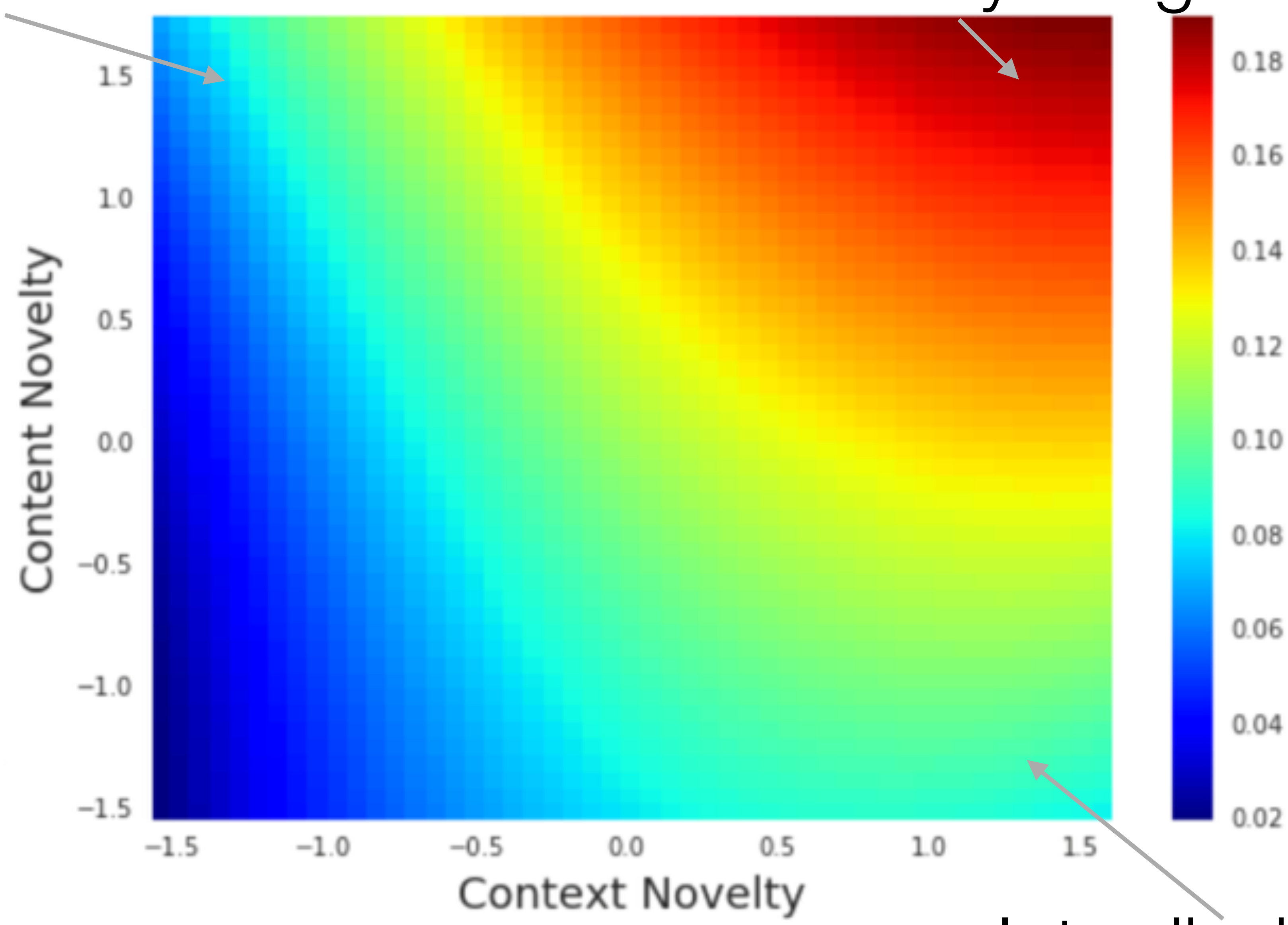
Content Correlates at **~.1** with Context

Context does NOT proxy for Content

Novelty and Impact

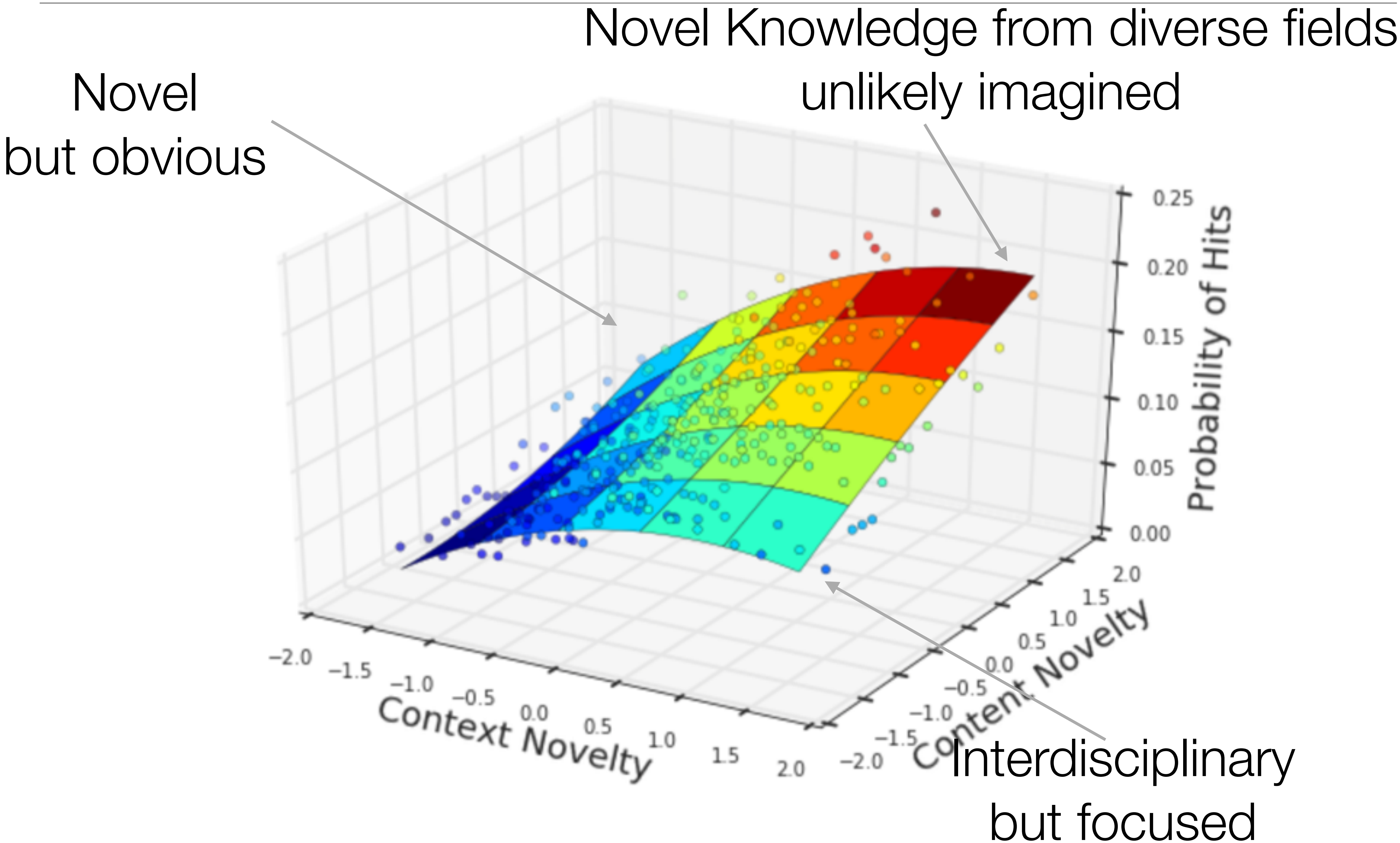
Novel Knowledge from diverse fields
unlikely imagined

Novel
but obvious



Interdisciplinary
but focused

Novelty and Impact



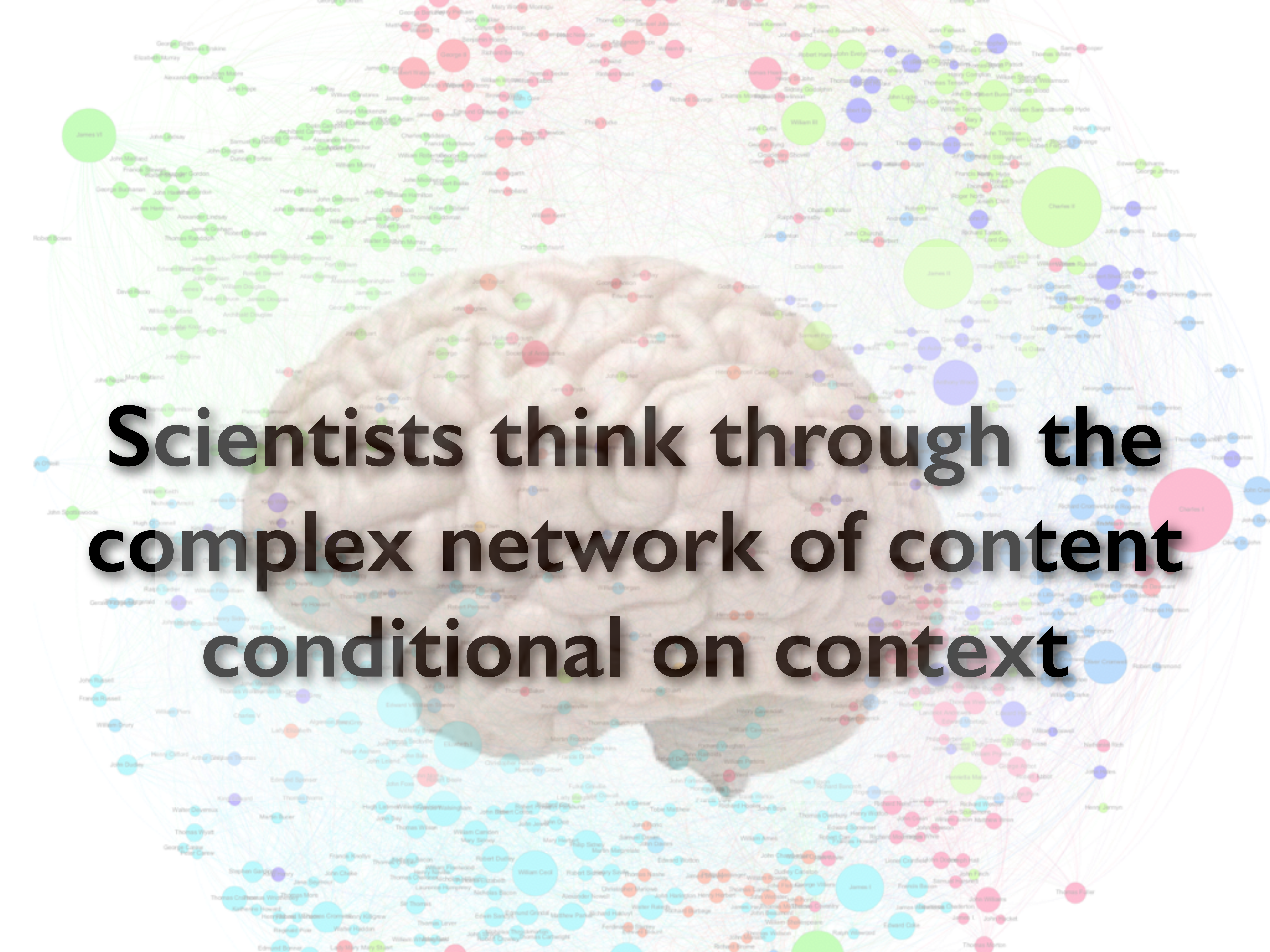
Doubling Sensitivity with High Dimensionality

- 12-13% - pairwise context (Uzzi's method)
- **25%** - high dimensional context (our method)
- 10% - pairwise content (Uzzi's method)
- **20%** - high dimensional content (our method)

Critical for estimating the effects for sparse signal—content similarity

Content & Context

- ~ 0.0 correlation between content and context novelty
- **30%** of hit probability captured by hypergraph of context + context
- Not linearly additive, but SUBSTANTIAL marginal effect



Scientists think through the complex network of content conditional on context

Science Think **Differently** from ***Scientists*** Think

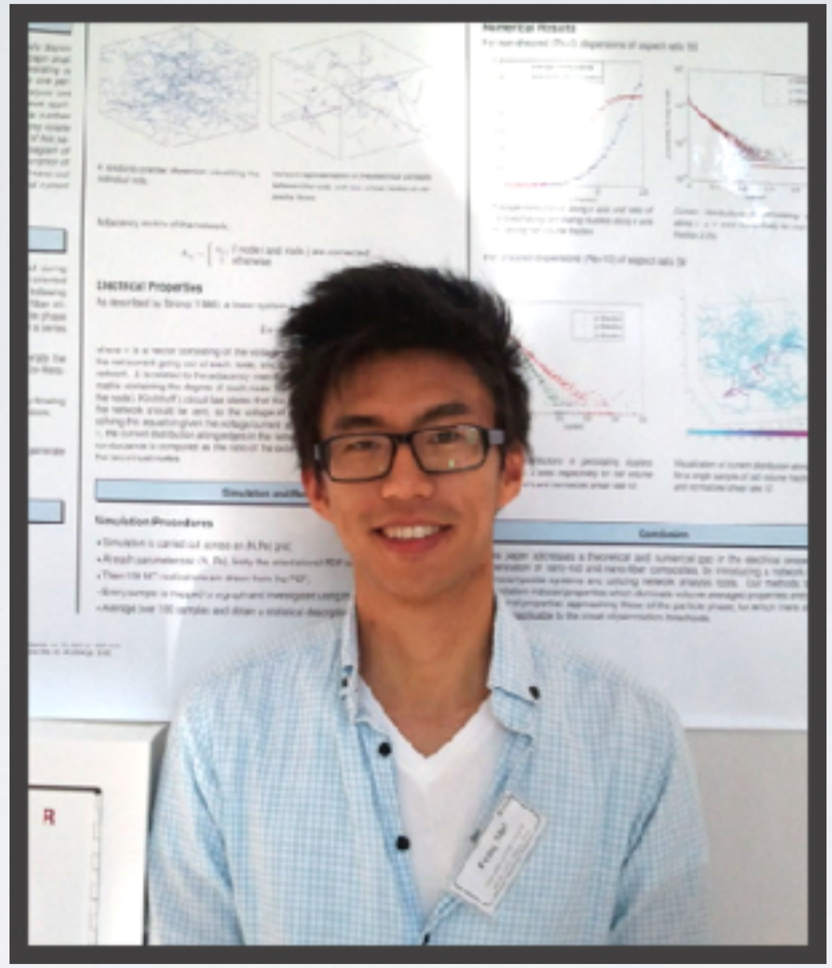
Science thinks like a **Global Bayesian**

…by conditioning success/impact on affirmation of global priors

Scientists have *much* weaker priors

…but succeed by **appearing to build on the shoulders of their audience**

Negative crowd-sourcing - finding combinations unlikely to have been imagined nearly doubled the likelihood of success



Bill Shi