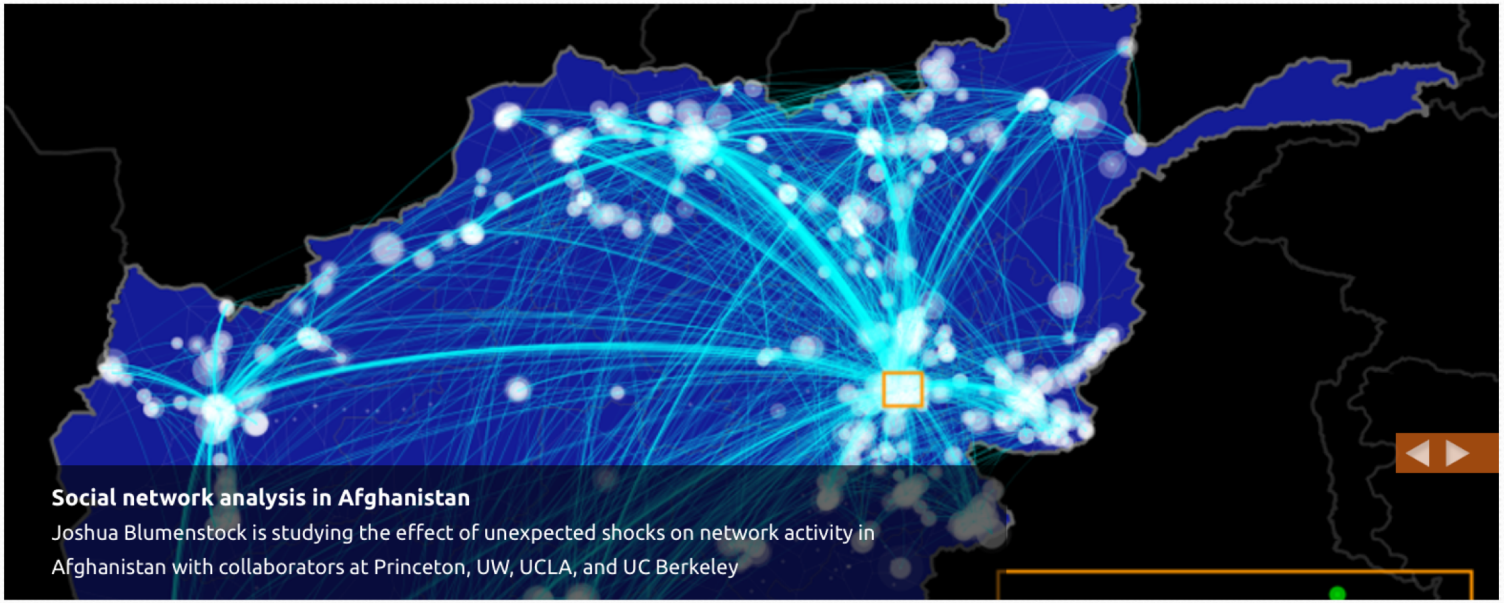
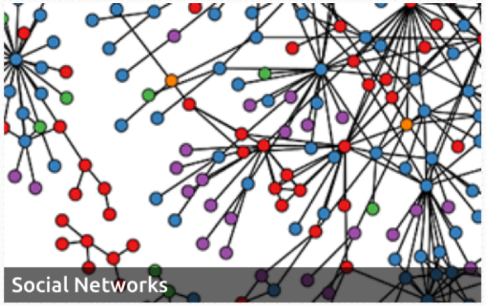
INFX 371
Applied Machine Learning

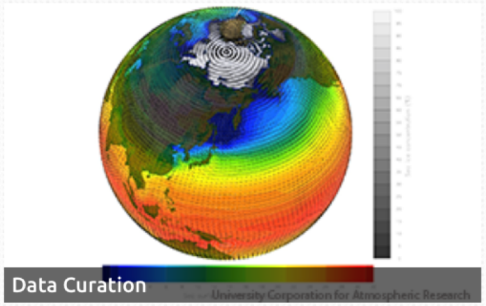
Spring 2018

Introductions



Research Focus Areas





GORDON AND BETTY MOORE FOUNDATION

DATA-DRIVEN DISCOVERY

Data Science Environments

What We Do

Overview

Over the course of the last decade many disciplines have evolved from recording observations in laboratory notebooks to the use of instruments capable of digitally recording many gigabytes of data in a day. This abundance of data provides unprecedented opportunities for discovery. Tapping its potential requires the application of sophisticated new computational techniques operating on large scale storage, computational and network resources. Since its creation in 2008, the eScience Institute has worked to create the intellectual and physical infrastructure needed to meet this challenge.

At the core of the eScience Institute are individuals who have proven track records in developing and applying advanced computational methods and tools to real world problems. Their task is to seek out and engage researchers across disciplines where eScience approaches are likely to have the greatest impact. To ensure that researchers have access to the necessary physical infrastructure, the Institute has undertaken coordinated planning and support for advanced local and remote computational platforms. This includes developing relationships with commercial and non-commercial service providers as well as the development of shared facilities on campus. This support extends to assistance in the preparation of select proposals where we are able to focus resources, improving their chances for success.

Also in... What We Do

Appliance Gallery
Find and use the eScience Institute's virtual machines equipped with software useful for specific applications.
Campus Compute & Storage
Learn about what UW is doing to support scalable scientific computing on campus
Consulting & Services
From algorithm development to database creation to cloud computing, we can help.
Projects
Explore some of our current collaborations with research scientists.
Relevant Courses
View a list of courses offered in eScience disciplines.
SQLShare Success Stories
Tools
Whether it's database management, visualization, or developer tools, learn about tools we can help you use.

Latest eScience News

Data Science Incubation Program - Winter 2016
2 hours 4 min ago

Ben Marwick On How Computers Broke Science

UW Data Science Seminar

ANALYSIS, VISUALIZATION & DISCOVERY

The **Data Science Seminar** is a university-wide effort bringing together thought-leading speakers and researchers across campus to discuss topics related to data analysis, visualization and applications to domain sciences. The seminar is typically held on **Wednesdays 3:30-4:30pm**. Unless otherwise noted, the location for Winter Quarter 2016 is **Johnson 102**.

All talks are free and open to the public.

Upcoming Speakers

JAN 6

Geometric graph-based methods for high dimensional data
Andrea L. Bertozzi
Professor, Department of Mathematics, UCLA

JAN 27

Inferring Complex Behavioral Mechanisms in Difficult Places
Carter Butts
Professor, Department of Sociology, UC Irvine

FEB 24

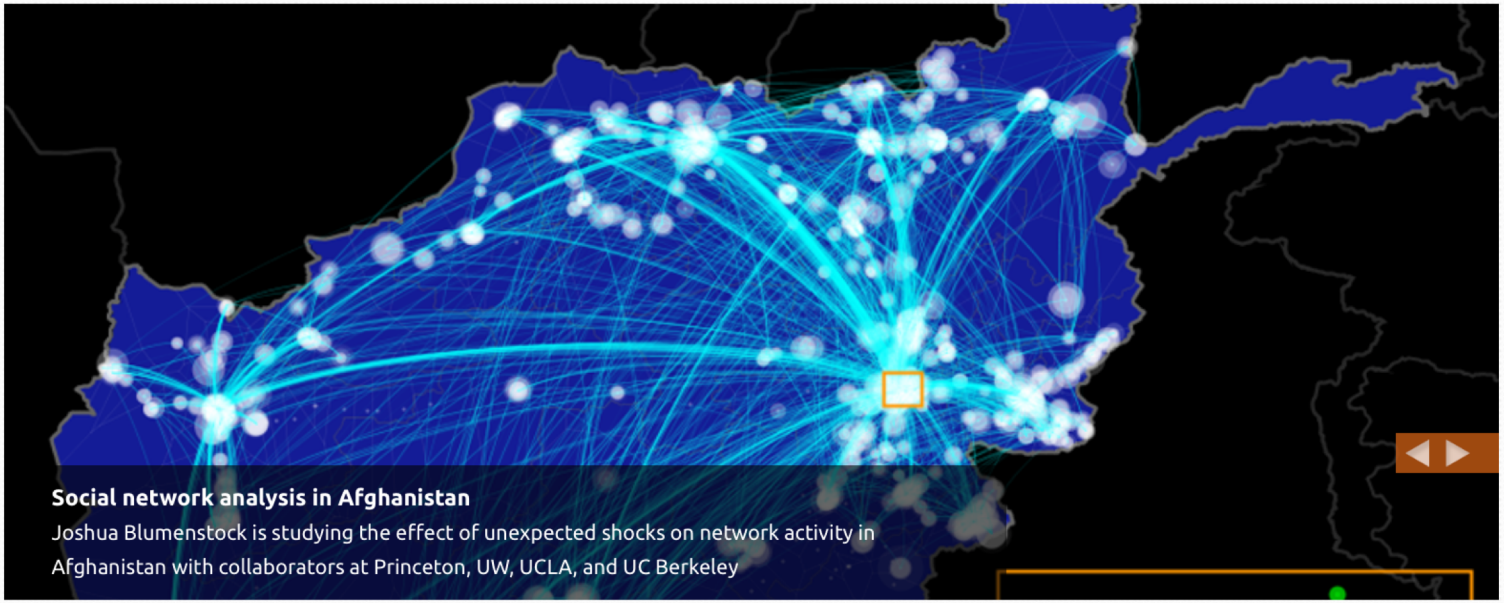
Turning the Virtual Tables: Social Media, Opposition, and Government Responses in Russia and Venezuela
Joshua Tucker
Professor, Department of Politics, NYU

FEB 25

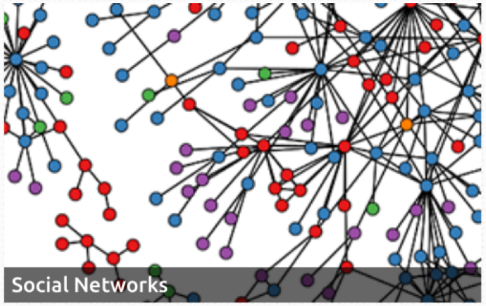
Machine Learning on Images: Combining Passive Microwave and Optical Data to Estimate Snow Water Equivalent in Afghanistan's Hindu Kush
Jeff Dozier
Professor, School of Environmental Science & Management, UC Santa Barbara

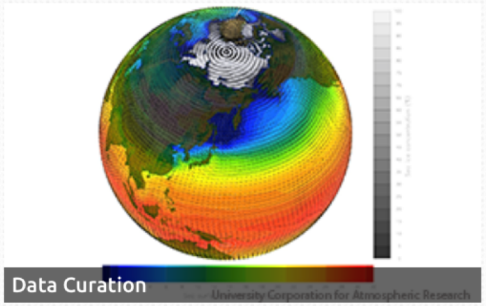
APR 20

Seeing Networks Change
Martin Krzywinski
Genome Sciences Centre, BC Cancer Agency

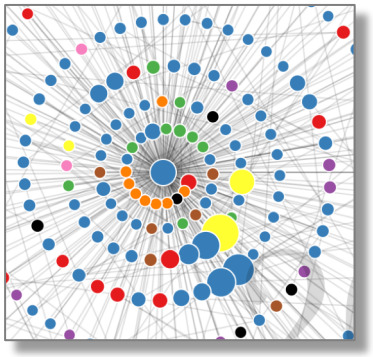


Research Focus Areas

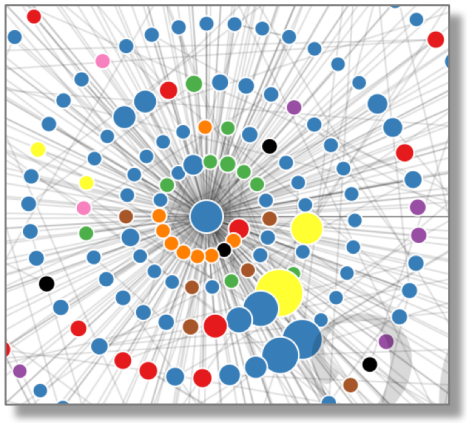




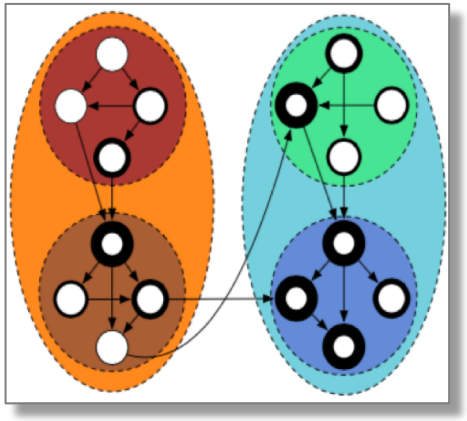
Science of Science



JW



Knowledge Science



Knowledge Engineering

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

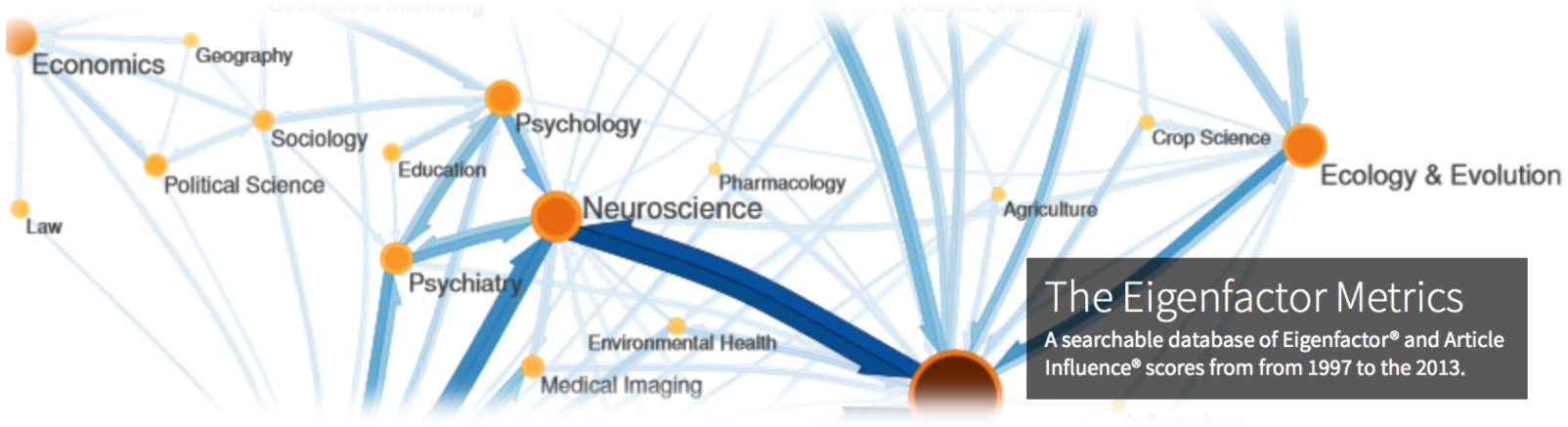
Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

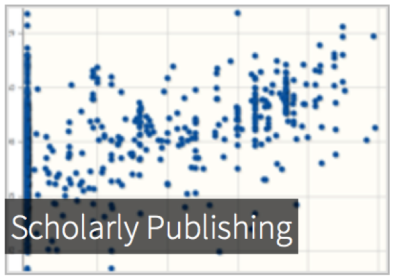
Competing Interests: The author has declared that no competing interests exist.

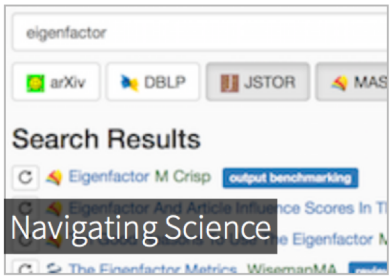
DOI: 10.1371/journal.pmed.0020124

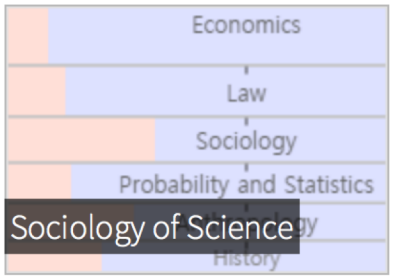

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE
Home
About
Projects
Jobs
Research
Press
Semantic Scholar
Cut through the clutter.
Home in on key papers, citations, and results.
Search
Try:
Pedro M. Domingos
Deep Learning
Penn Treebank



RESEARCH AREAS







NEWS

23 Nov.

JEVIN WEST ON MEGAJOURNALS IN THE *CHRONICLE OF HIGHER EDUCATION*

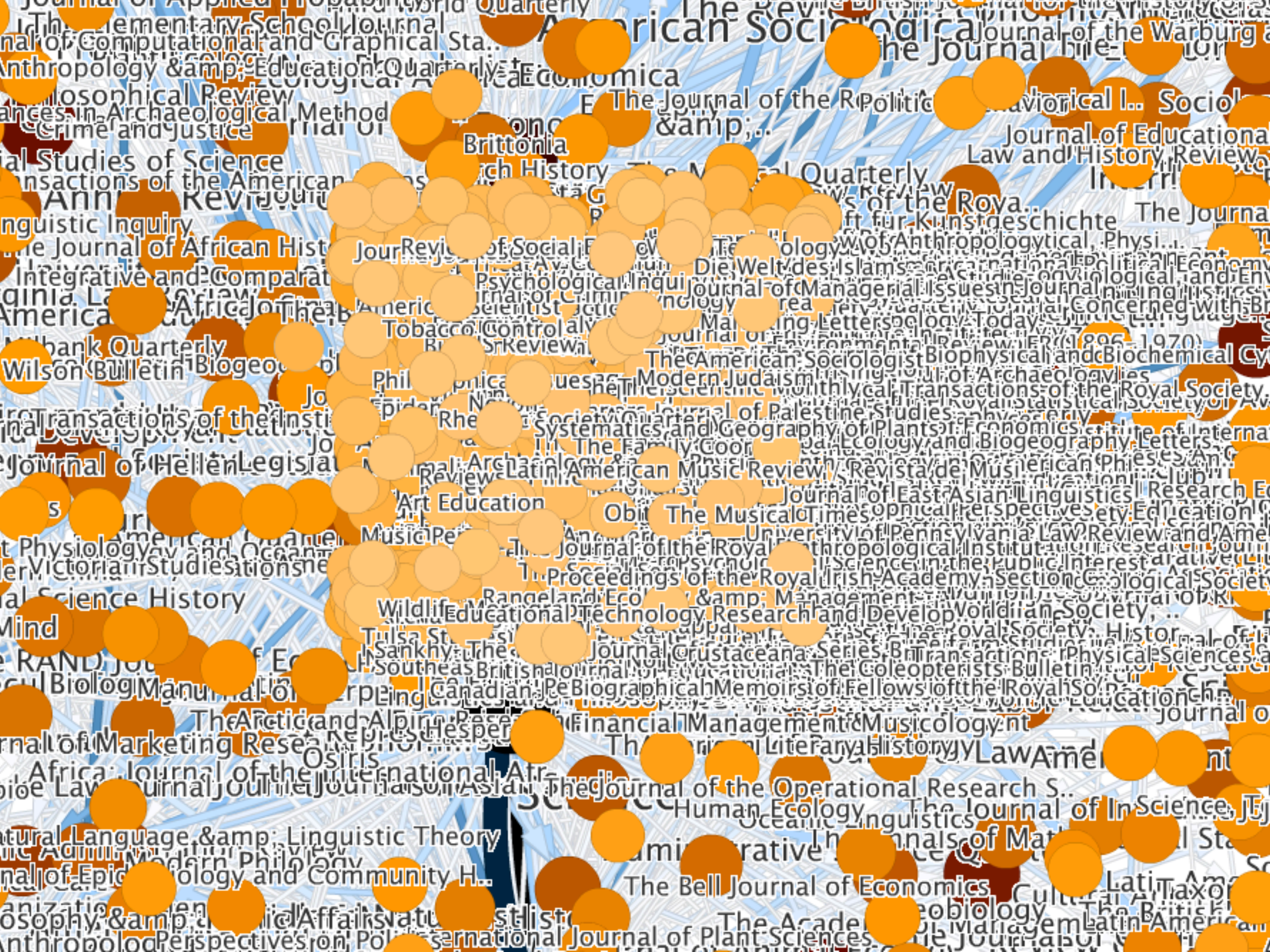
Jevin West discusses the rise of the megajournal and our open access cost effectiveness tool in the *Chronicle of Higher Education*.

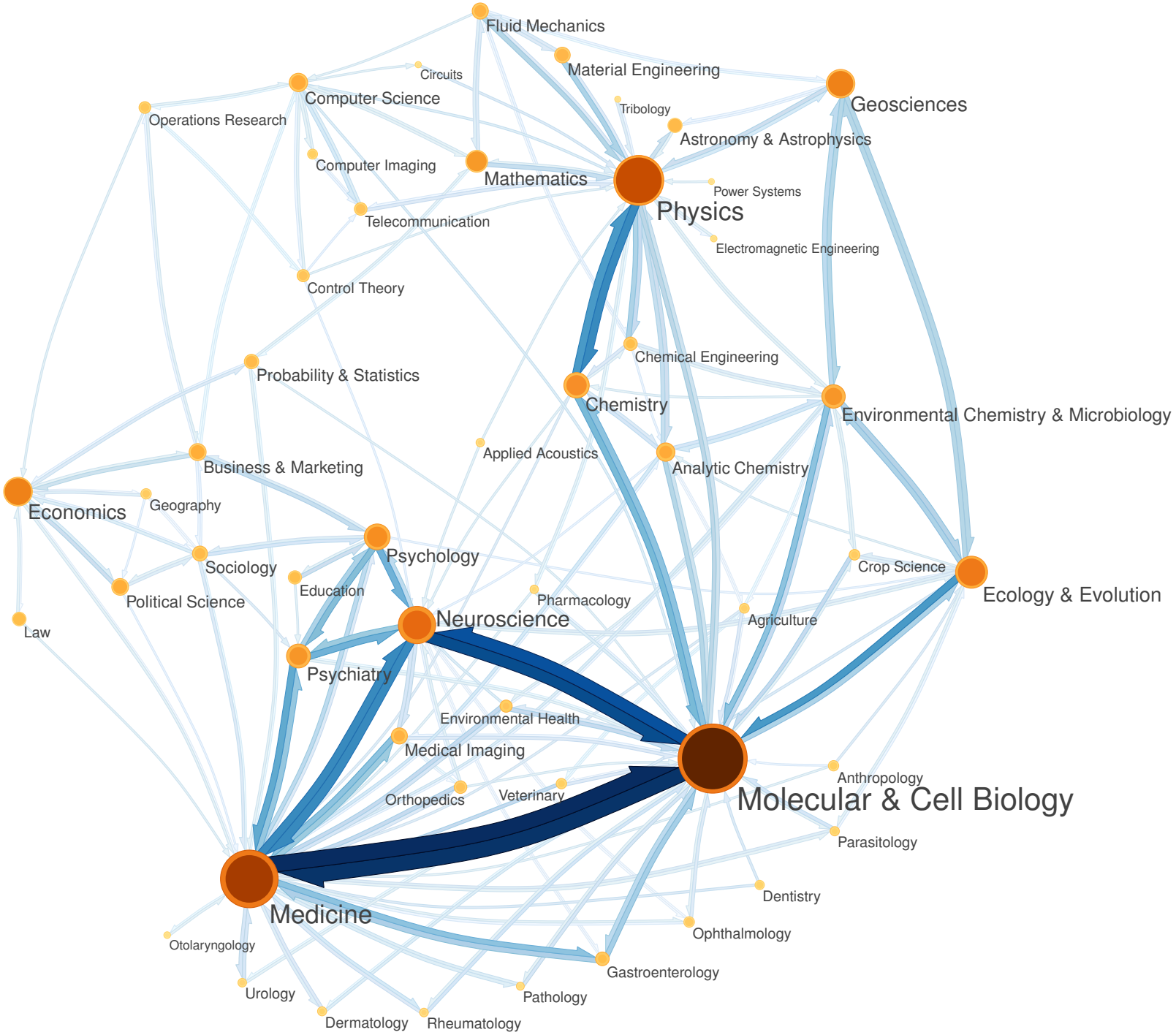
---

23 Nov.

EIGENFACTOR TEAM PLACES SECOND IN MICROSOFT RESEARCH'S WSDM CUP

The WSDM Cup Challenge asked teams to use 30GB of data from the Microsoft Academic Graph to rank the

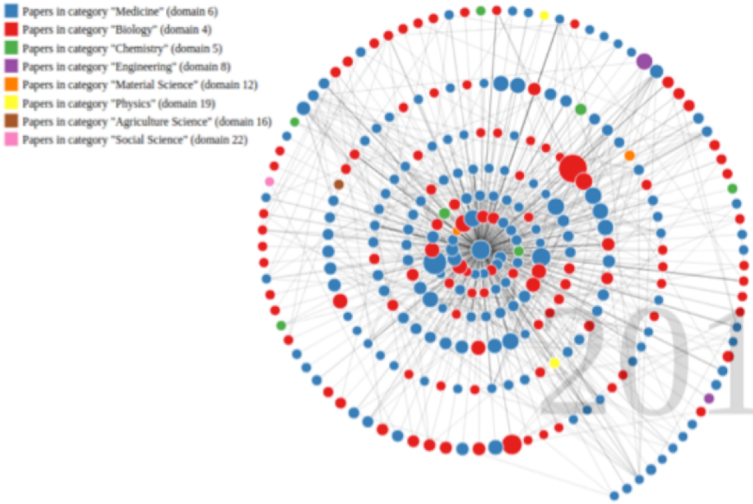

American Sociological
The Elementary School Journal
Journal of the Warburg
Anthropology & Education Quarterly
Economica
Philosophical Review
Crime and Justice
Brittonia
Journal of Educational
Law and History Review
Social Studies of Science
Linguistic Inquiry
The Journal of African History
Die Welt des Islams
Psychological Inquiry
Journal of Managerial Issues
Tobacco Control
Milbank Quarterly
The Wilson Bulletin
The American Sociologist
Biophysical and Biochemical Cy
Modern Judaism
Journal of Palestine Studies
Systematics and Geography of Plants
The Family Coordinator
Latin American Music Review
Revista de Música
Journal of Hellenic
Journal of East Asian Linguistics
Art Education
The Musical Times
University of Pennsylvania Law Review
Victorian Studies
Proceedings of the Royal Irish Academy. Section C
Social Science History
Educational Technology Research and Development
Mind
Journal Crustaceana Series
The Coleopterists Bulletin
Biographical Memoirs of Fellows of the Royal Society
Financial Management
Musicology
Journal of Marketing Research
Osiris
The Journal of the Operational Research S..
Human Ecology
Natural Language & Linguistic Theory
Modern Philology
Journal of Epidemiology and Community H..
The Bell Journal of Economics
Perspectives on Politics
International Journal of Plant Sciences


Fluid Mechanics
Material Engineering
Circuits
Computer Science
Geosciences
Operations Research
Tribology
Astronomy & Astrophysics
Computer Imaging
Mathematics
Power Systems
Physics
Telecommunication
Electromagnetic Engineering
Control Theory
Chemical Engineering
Probability & Statistics
Chemistry
Environmental Chemistry & Microbiology
Applied Acoustics
Business & Marketing
Analytic Chemistry
Economics
Geography
Psychology
Crop Science
Sociology
Education
Political Science
Pharmacology
Ecology & Evolution
Law
Neuroscience
Agriculture
Psychiatry
Environmental Health
Medical Imaging
Anthropology
Orthopedics
Veterinary
Molecular & Cell Biology
Parasitology
Dentistry
Medicine
Ophthalmology
Otolaryngology
Gastroenterology
Urology
Pathology
Dermatology
Rheumatology

JW

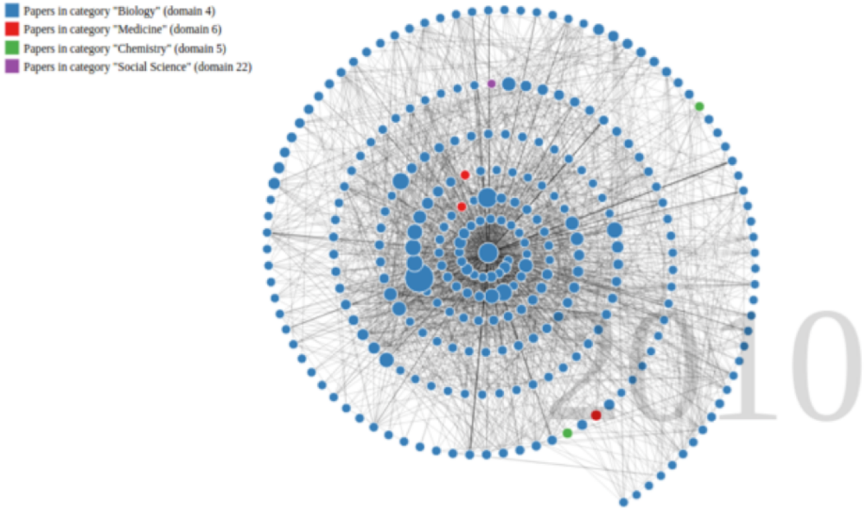
Visualizing Interdisciplinarity

Jason Portenoy



A more sparse network indicates fewer citations between papers shown in the network. This could be a result of the central scholar having impact across a wider set of academic communities.

A denser network means that the papers that cite the central author also tend to cite each other.

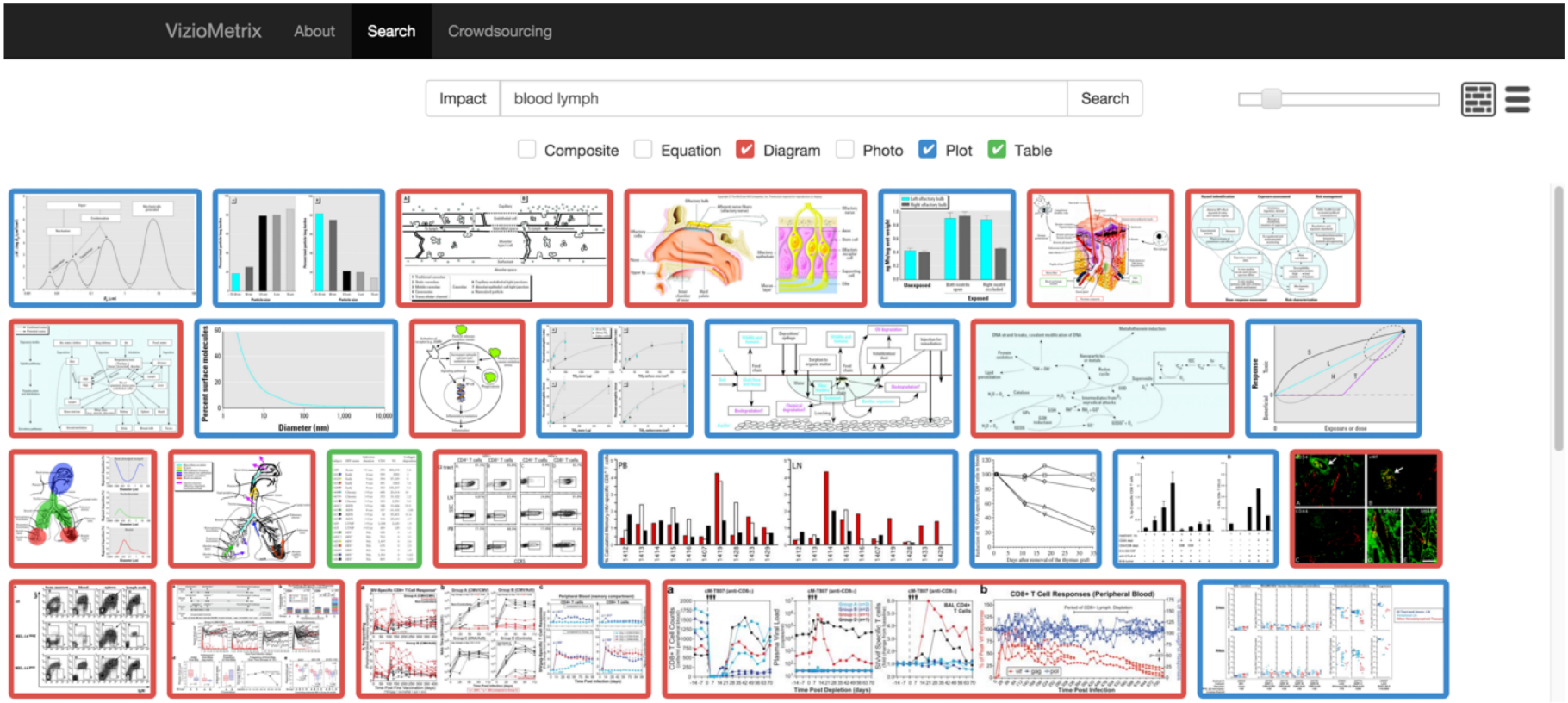


scholar.eigenfactor.org

JW

Figure-Centric Search Engine

viziometrics.org



The Graph

The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between publications, as well as authors, institutions, journal and conference "venues," and fields of study.

The Data

This data is available as a set of zipped text files stored in Microsoft Azure blob storage and available via HTTP. The file size (zipped) is ~30GB and may be downloaded here.

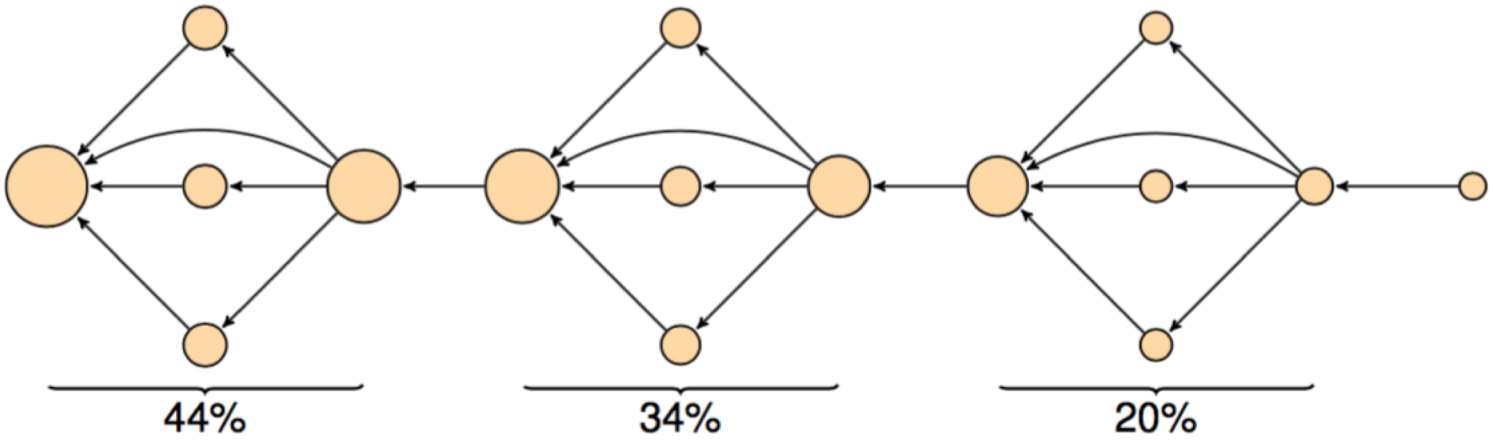
The Challenge

The goal of the Ranker Challenge is to assess the query-independent importance of scholarly articles, using data from the Microsoft Academic Graph.

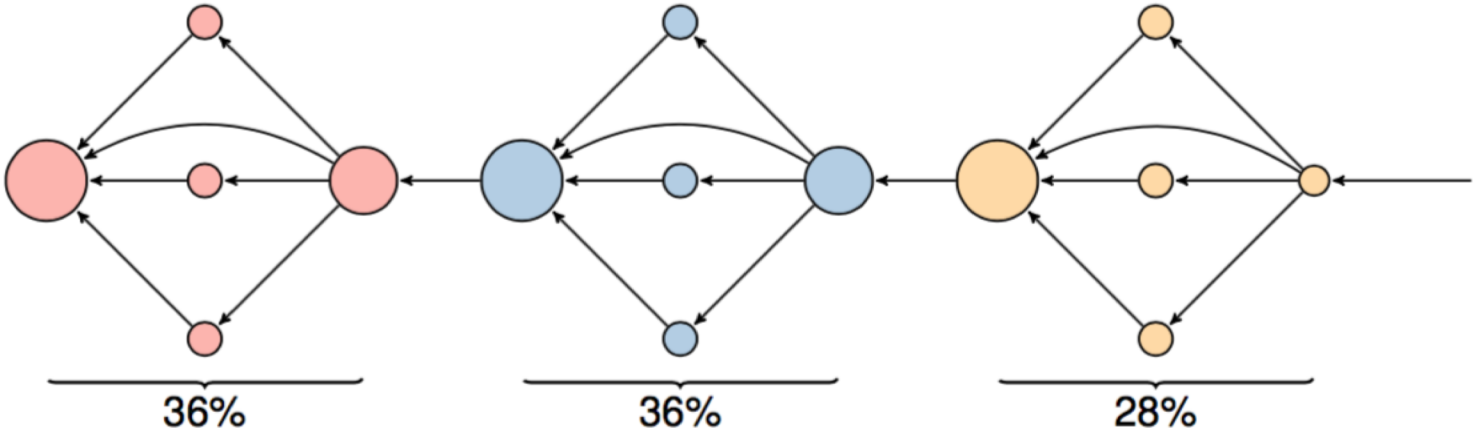
JW

Ranking on time-directed networks

PageRank

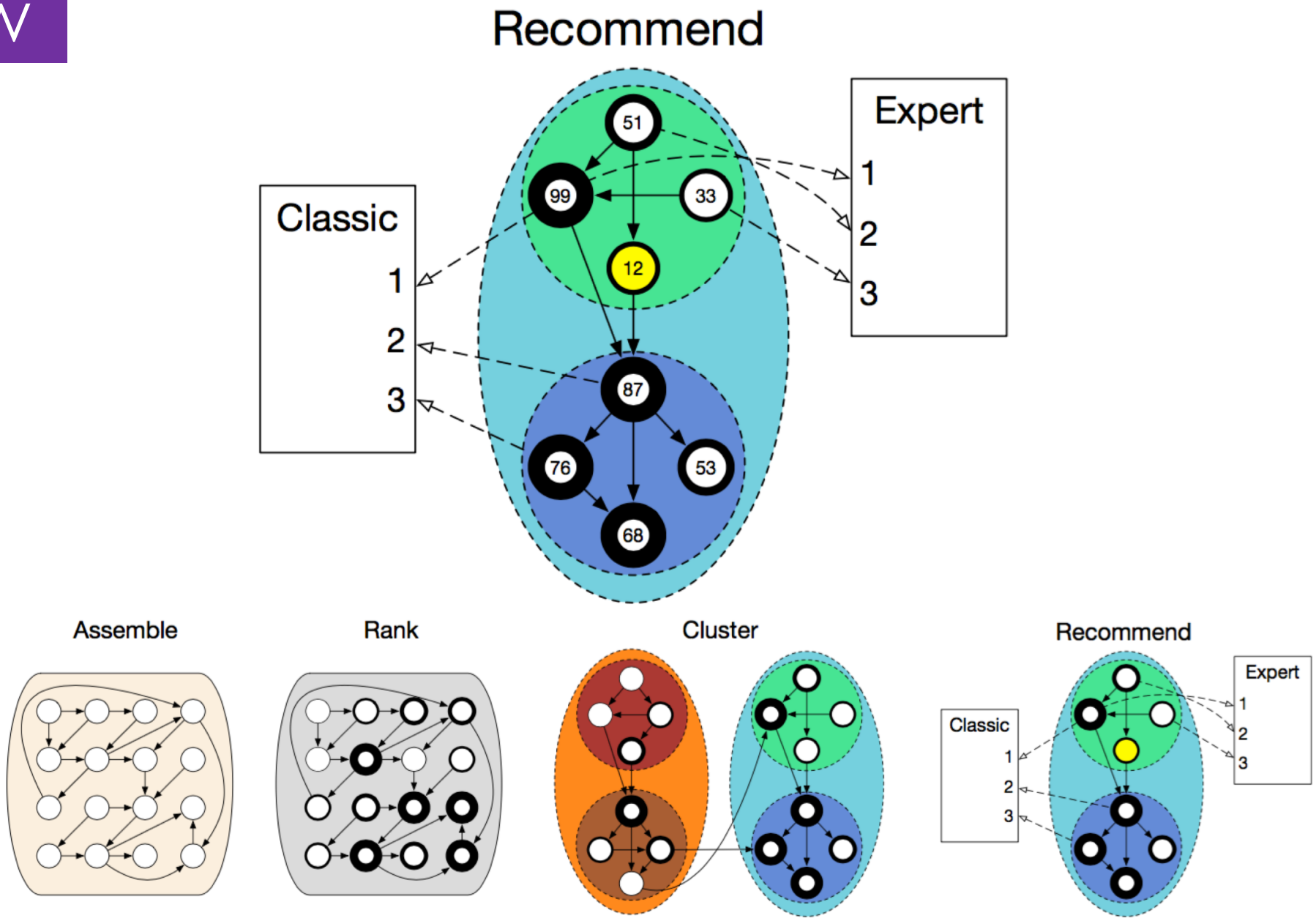


ALEF



Time

JW



West, Wesley-Smith, Bergstrom (2016) A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE, Transactions on Big Data* (in press)

Babel | Home | About | API

babel.eigenfactor.org

oren etzioni

Ian Wesley-Smith

DBLP | JSTOR | MAS | PLOS | PubMed

s

l Methods For Analyzing Speedup Learning Experiments O Etzioni **satisfaction programs**

Face And Computer-Mediated Communities Amitai Etzioni, Oren Etzioni 1998 **resources sustained**

ument Clustering O Zamir **document clustering**

Communities: Virtual Vs. Real A Etzioni 1996 **implications internet**

l Methods For Analyzing Speedup Learning Experiments. O Etzioni 1993 **scheduling problems**

l Methods For Analyzing Speedup Learning Experiments O Etzioni 1993 **generating abstractions**

Get Related Web Document Clustering: A Feasibility Demonstration O Zamir 1997 **document clustering**

Get Related Web Document Clustering: A Feasibility Demonstration. O Zamir 1997 **browsing large**

Get Related Sound And Efficient Closed- World Reasoning O Etzioni **proving problem**

Get Related Appears In Comm. OfACM O Etzioni **scalable comparison-shopping**

« Previous | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Next »

Papers related to Statistical Methods For Analyzing Speedup Learning Experiments O Etzioni **satisfaction programs**

Get Related Automatically Configuring Constraint Satisfaction Programs: A Case Study S Minton 1995 **satisfaction programs**

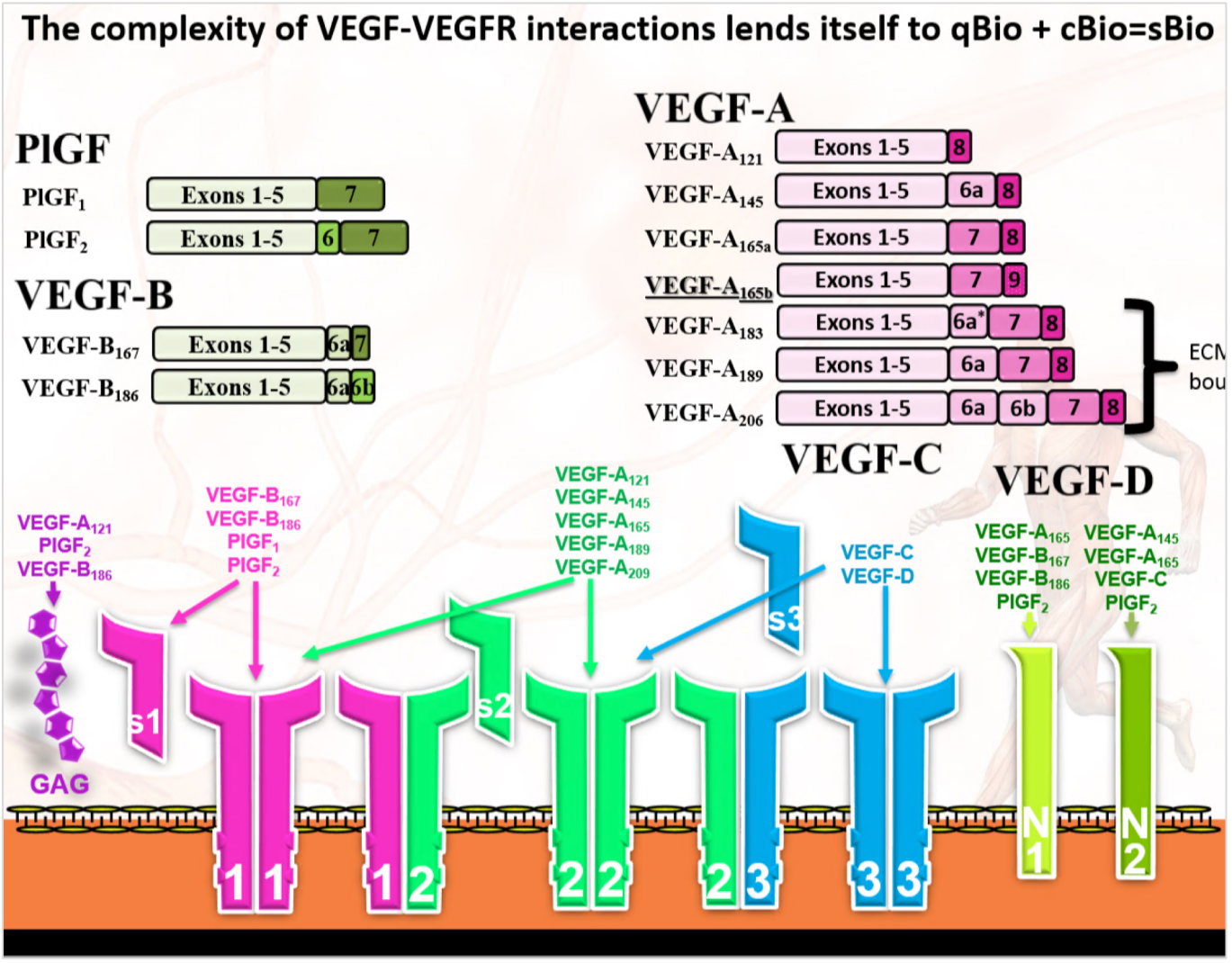
Get Related Abstraction Via Approximate Symmetry T Ellman 1992 **satisfaction programs**

Get Related Integrating Heuristics For Constraint Satisfaction Problems: A Case Study S Minton 1992 **satisfaction programs**

Get Related An Analytic Learning System For Specializing Heuristics S Minton 1992 **satisfaction programs**

Get Related Automated Synthesis Of Constrained Generators W Braudaway 1988 **satisfaction programs**

Mining the literature



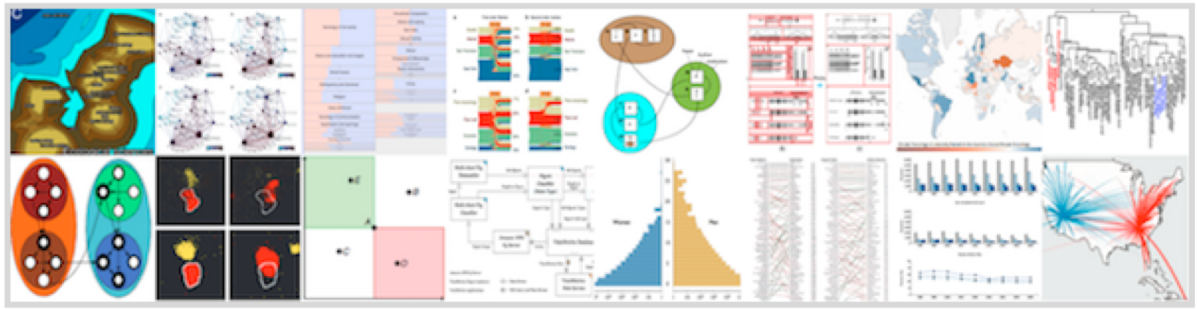
In collaboration with P. I. Imoukhuede, University of Illinois

http://jevinwest.org

JW

Jevin D. West

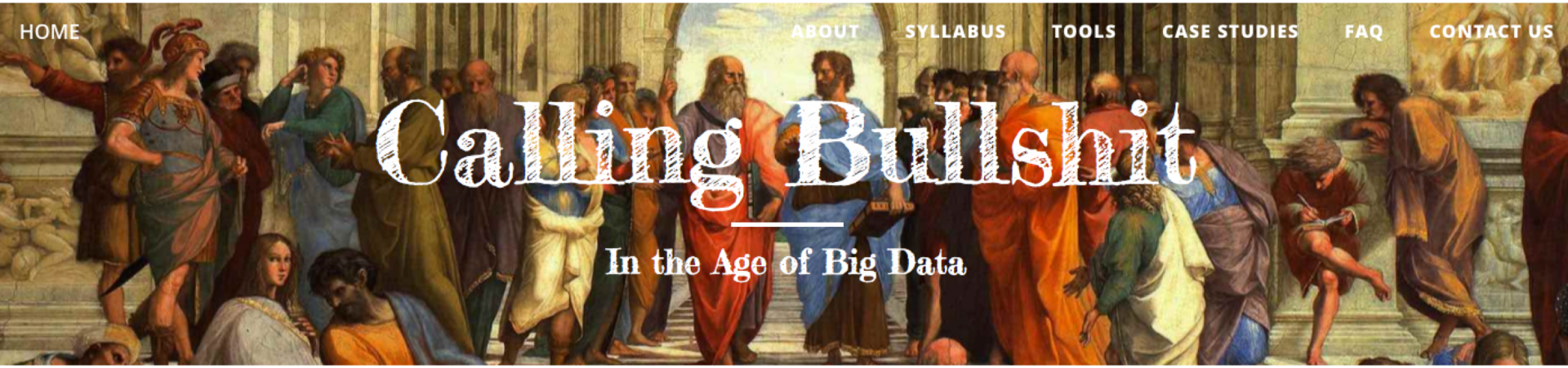
Home | Research | Publications | Presentations | Teaching | Bio | CV | News



Science of Science

Science is the greatest of human inventions. It has solved and continues to solve many of societies most pressing questions in human health, planetary wellness and economic viability. But one of Science's new challenges is the well being of Science itself. The reproducibility crisis, misaligned incentives and evaluations of scientists, literature overload, publication bias, and out-of-date publishing models are just a few of the maladies of Science. Turning the microscope on Science itself - the *Science of Science* - is the focus of my research.

How is this different than the sociology and history of science, science policy or scholarly communication? Overlaps exist and methods are borrowed from these established disciplines, but the difference is the scale and kind of data, the methods and tools from data science, and the amalgam of these disciplines under one roof. It is difficult to understand literature overload or the reproducibility crisis if one does not examine, in parallel, what drives scientists to publish, what technologies they use to disseminate their findings, and their established norms for publishing.

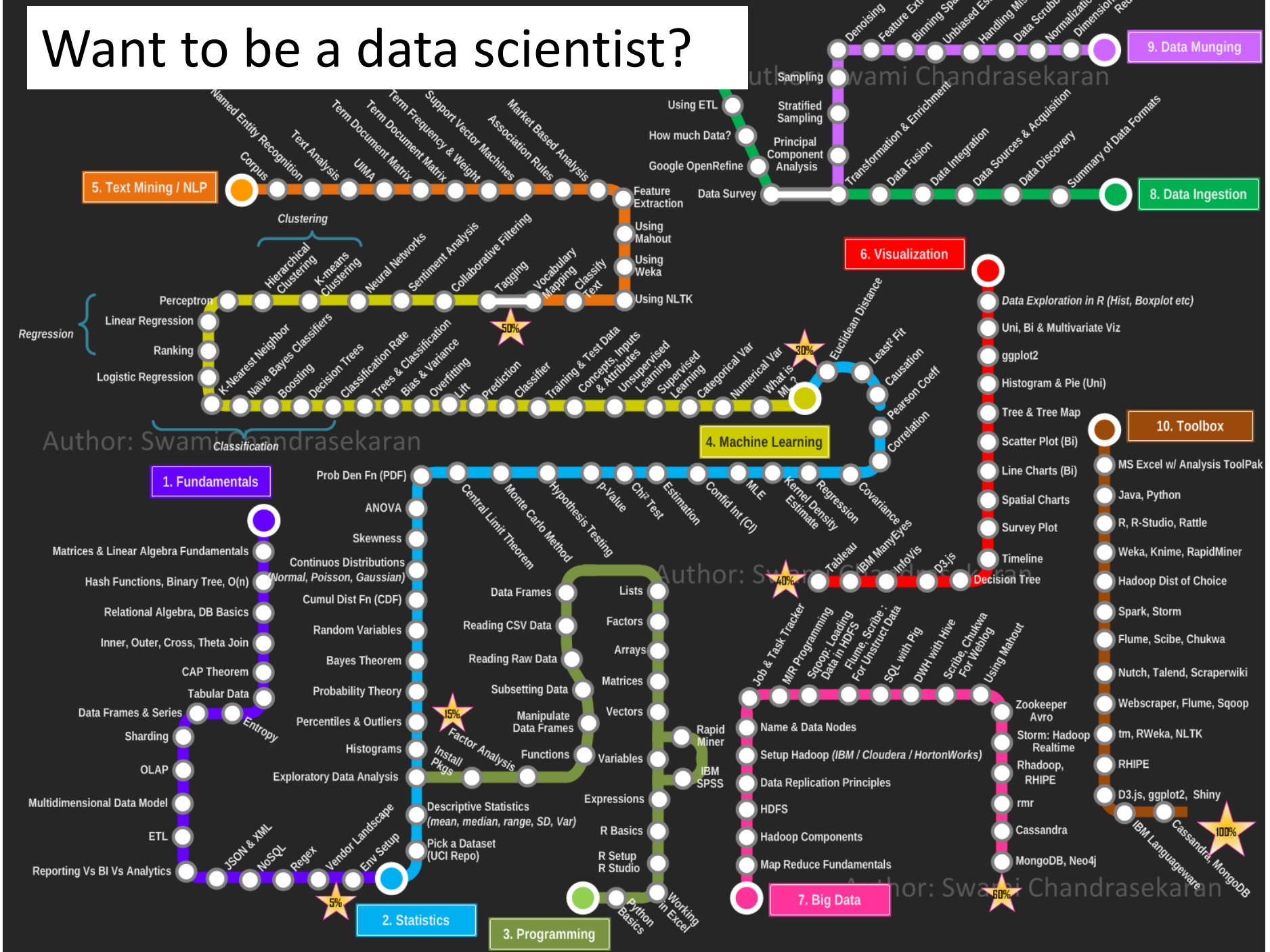


The world is awash in bullshit. Politicians are unconstrained by facts. Science is conducted by press release. So-called higher education often rewards bullshit over analytic thought. Startup culture has elevated bullshit to high art. Advertisers wink conspiratorially and invite us to join them in seeing through all the bullshit, then take advantage of our lowered guard to bombard us with second-order bullshit. The majority of administrative activity, whether in private business or the public sphere, often seems to be little more than a sophisticated exercise in the combinatorial reassembly of bullshit.

Fall 2017

http://callingbullshit.org

Want to be a data scientist?



What is machine learning?

What is econometrics?

Hal Varian (Google)

- **Machine learning**, data mining, predictive analytics, etc. all use data to predict some variable as a function of other variables.
  - May or may not care about insight, importance, patterns
  - May or may not care about inference---how y changes as some x changes
- **Econometrics**: Use statistical methods for prediction, inference, causal modeling of economic relationships.
  - Hope for some sort of insight, inference is a goal
  - In particular, causal inference is goal for decision making

Class

- An *Introduction* to Econometrics and Machine Learning (if you want additional challenges, please come talk to me)
- Mix of conceptual and practical
- No required textbook but consider getting one (see syllabus for recommendations)
- 5 problem sets, 3 quizzes, participatory assignments (no final)
- Feel free to call me Jevin; email me directly at jevinw@uw.edu because I miss canvas messages
- TA Introductions

Class Rules

- Bring Laptop with software installed and running
- Check canvas for updates on readings and assignments
- Contact me through email
- Coding Assignments
  - Comment Code
  - Readable variable naming (for graders and yourself)
  - Write functions
- I will not debug your code

My Expectations of You

- Respectfulness to your fellow students
  - Limited use of non-class technology
- Preparedness and honesty
  - Timely attendance (with exceptions)
  - Individual work (in a team environment)
- Willingness to work
  - Fast moving class with lots to learn
  - ~2 hrs outside of class for every hour in class

Your expectations of Me

- Introduce you to Econometrics and Applied ML
  - Toolsets and concepts
  - Prepare you for next steps (jobs!)
- Customized Education Experience
  - Use current technologies and case studies
  - *Adaptive class* that responds to needs (speeds up, slows down)
- A philosophy of success rather than filtering
  - I want you to all succeed, that is my goal; it is not to filter

Learning Objectives (Week 1)

1. Install Python
2. Become familiar with Python (Jupyter Notebooks, numpy, scipy, pandas)
3. Explore an example data set using Python

Homework

- Problem Set #1 (**due Tuesday, April 5**)
  - Make sure to turn in two files: (1) html version of your finished notebook and (2) your jupter notebook code file (.ipynb)
- See canvas (home page) for required readings

Problem Set #1

- Due April 5th at 10:30 am
- Make sure to turn in Tutorial Jupyter notebook (participatory points)
- Good coding practice
  - Write your name, date and class section
  - Comment Code
  - Readable variable naming (for graders and yourself)
  - Write functions
  - Turn in to canvas (1) html version of your notebook and (2) jupyter notebook code version (.ipynb)
  - Questions: talk to Maria and Srijan