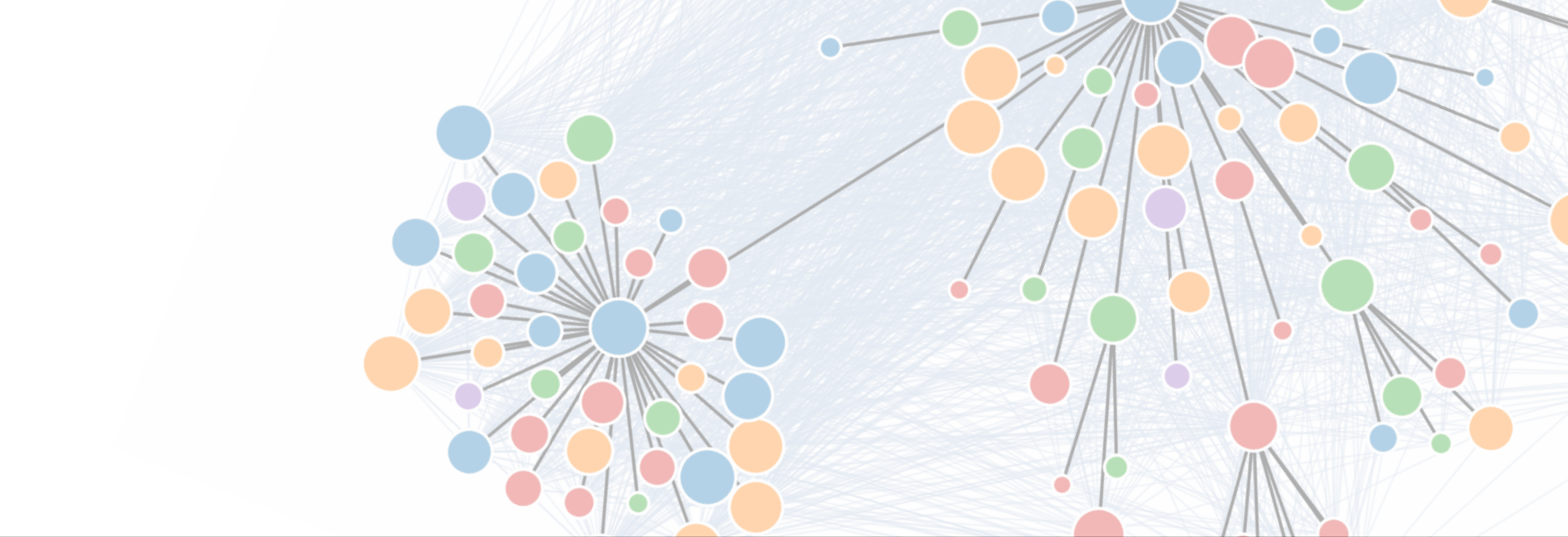


Researcher productivity and data-driven predictions in the science of science

Aaron Clauset
@aaronclauset
Computer Science Dept. & BioFrontiers Institute
University of Colorado, Boulder
External Faculty, Santa Fe Institute

the desire to predict science pervades society

what will be discovered?

by whom, when, and where?

the desire to predict science pervades society

what will be discovered?

by whom, when, and where?

individuals	what questions are interesting, impactful, fundable?
publishers, funders	what manuscripts or projects will be most impactful?
hiring committees	which applicant will perform best? which will make most valuable contributions?
society	how can tax and other dollars be invested to make technological, biomedical, and scientific advances?

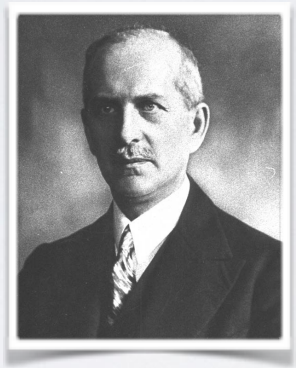
how *predictable* are scientific discoveries?

simple question with a 150+ year history

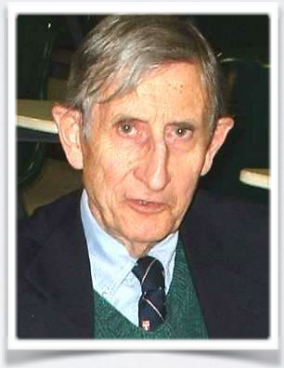
how *predictable* are scientific discoveries?

simple question with a 150+ year history, e.g.:

Bolesław Prus
(1847-1912)



Florian Znaniecki
(1882-1958)



Freeman Dyson

Steven Weinberg
(Nobel Physics, 1979)

- mainly conceptual, focusing on goals and general approaches

  (Weinberg: "to explain the world") (Dyson: "birds and frogs")

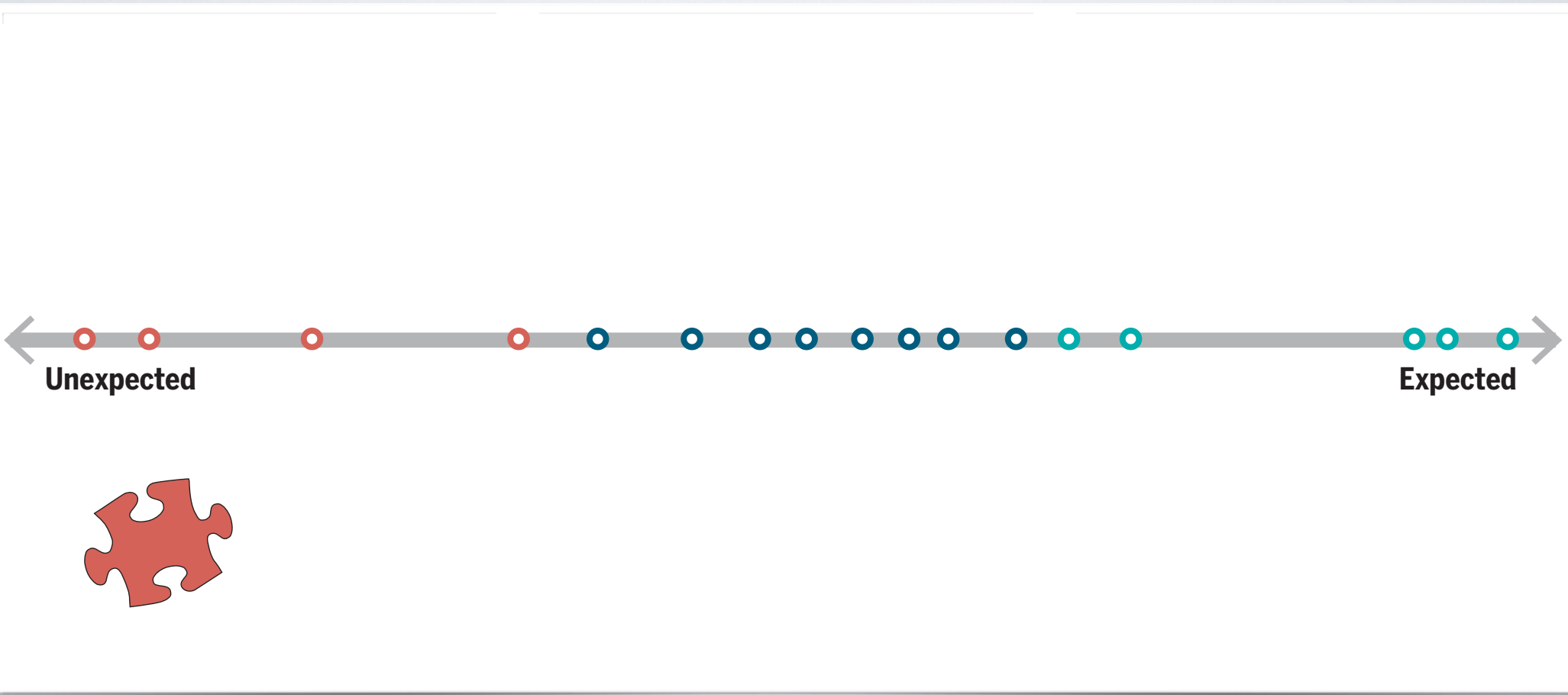
- **progress toward a genuine "science of science" was slow**

  hard to get good data

  judgement of experts seemed good enough

a conceptual framework for predictability:

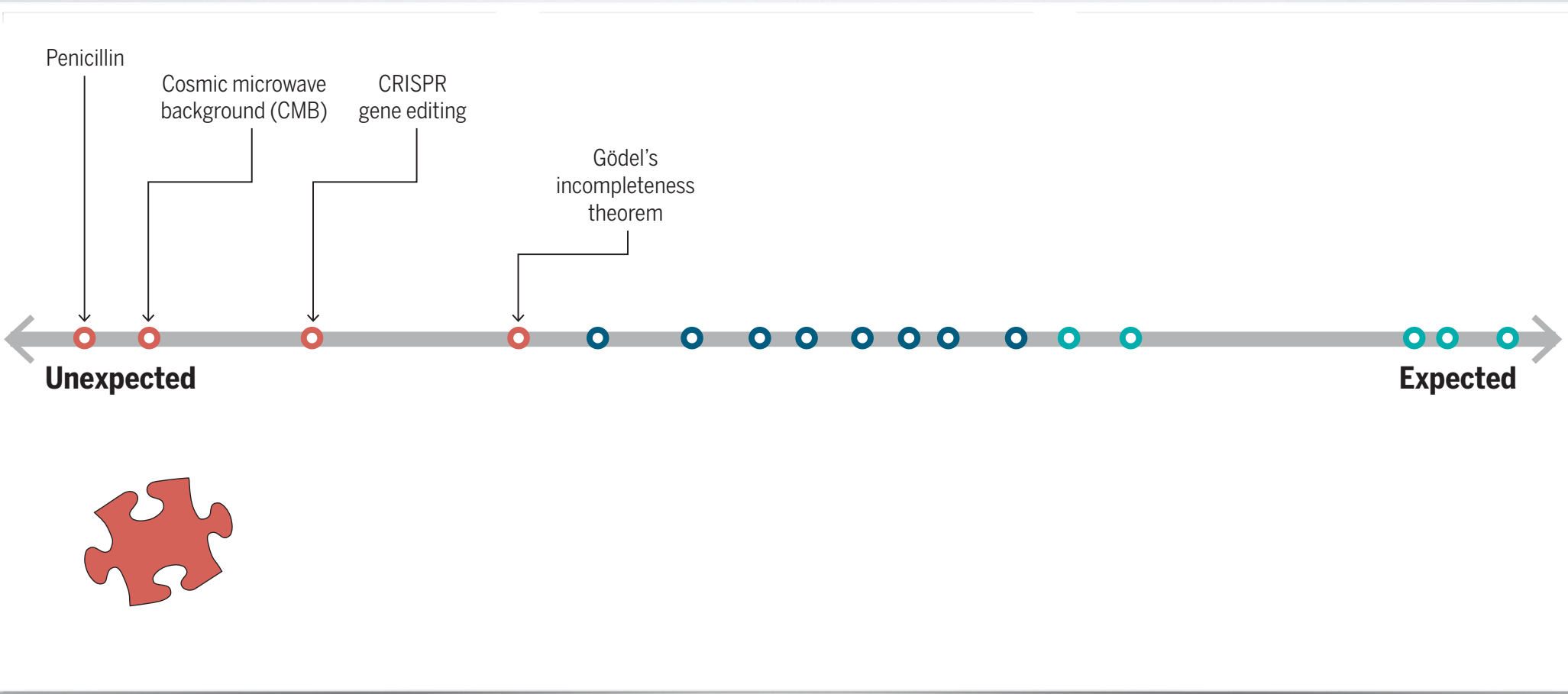
predictability depends on scientific context



Clauset, Larremore & Sinatra, *Science* **355**, 477-480 (2017)

unexpected discoveries

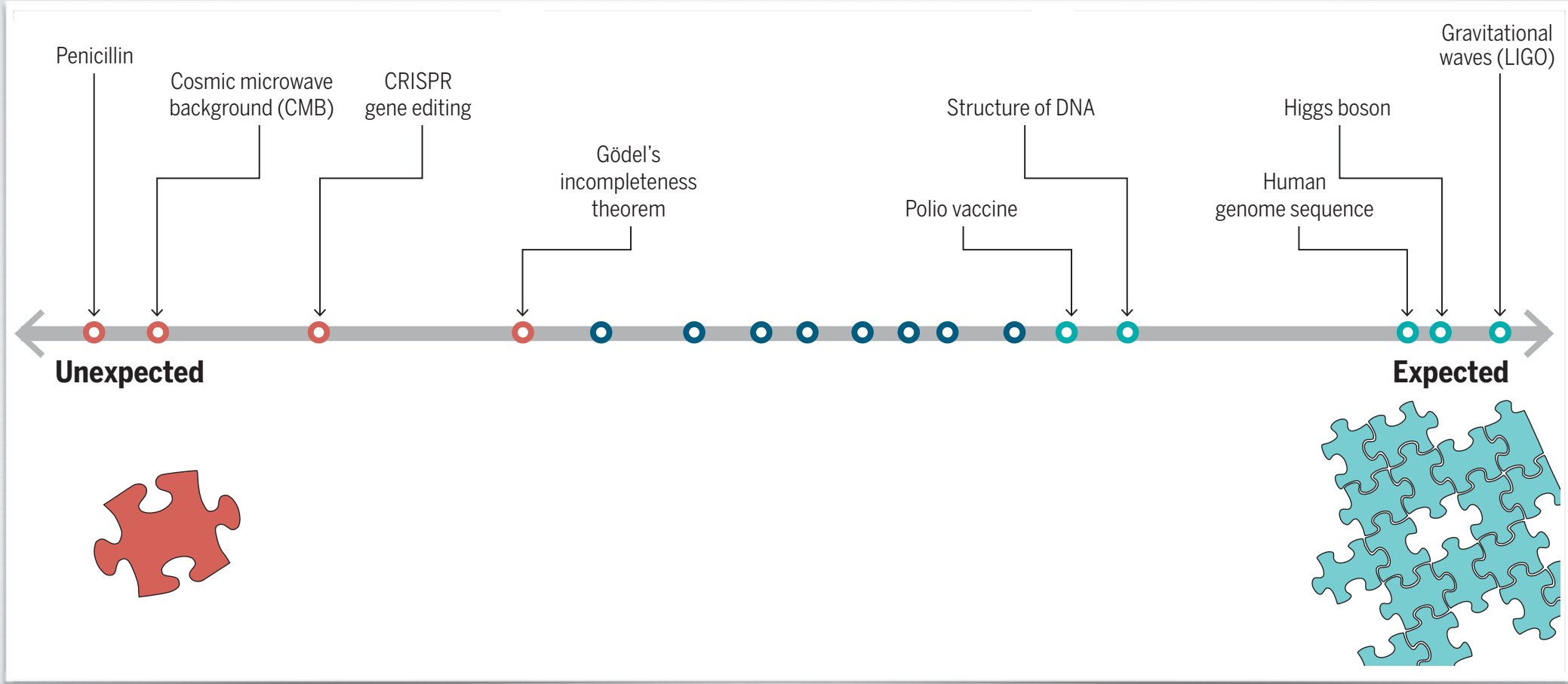
changes the way we understand the world, or finds novel use elsewhere



Clauset, Larremore & Sinatra, *Science* **355**, 477-480 (2017)

expected discoveries

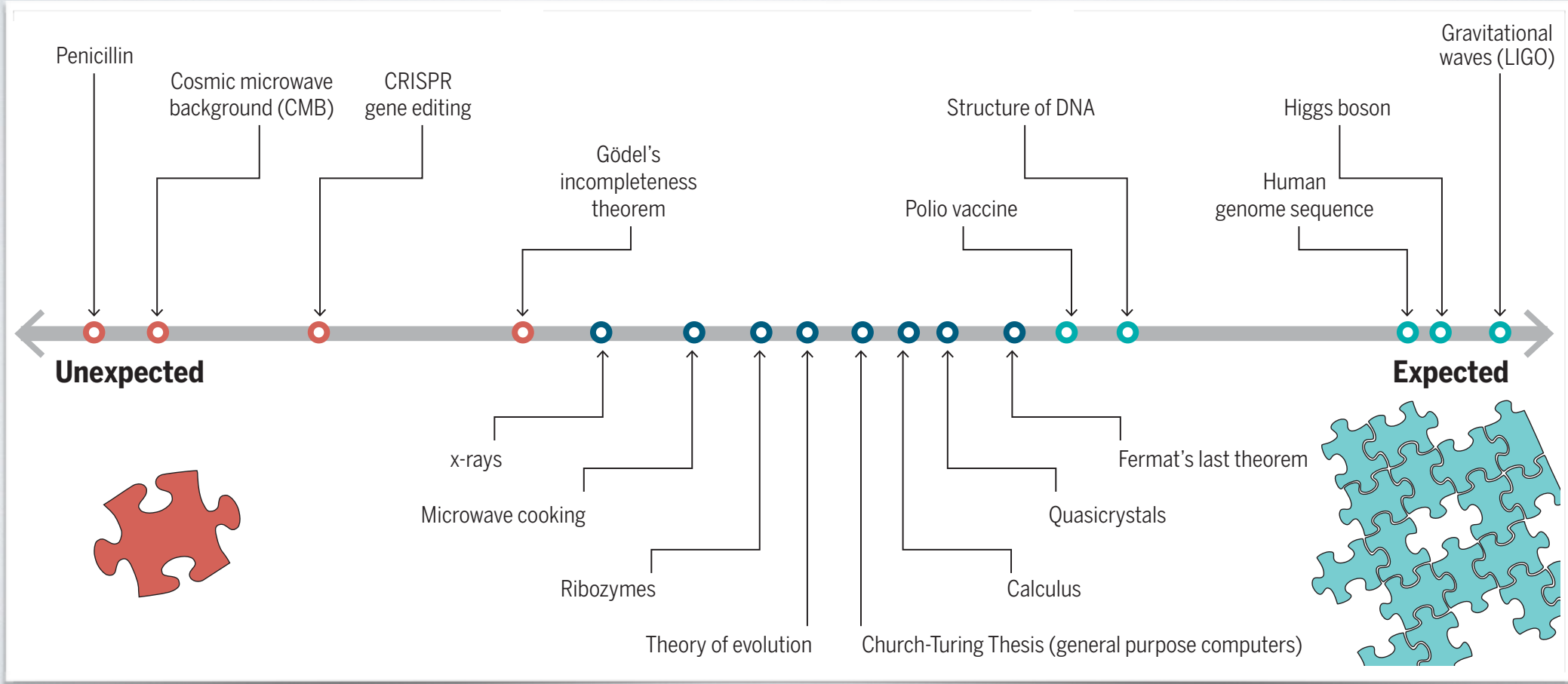
accumulation of theory and evidence, fits with other ideas



Clauset, Larremore & Sinatra, *Science* **355**, 477-480 (2017)

"normal" discoveries

some elements surprising, but fits partly within existing ideas



Clauset, Larremore & Sinatra, *Science* **355**, 477-480 (2017)

a simple conceptual framework for predictability:

predictability correlates with *past / related evidence*

unexpected discoveries *rearrange* our understanding of the world, & correlate poorly with the past

data most useful for better predicting only modestly unexpected discoveries (the middle)

Clauset, Larremore & Sinatra, *Science* **355**, 477-480 (2017)

a modern *science of science*

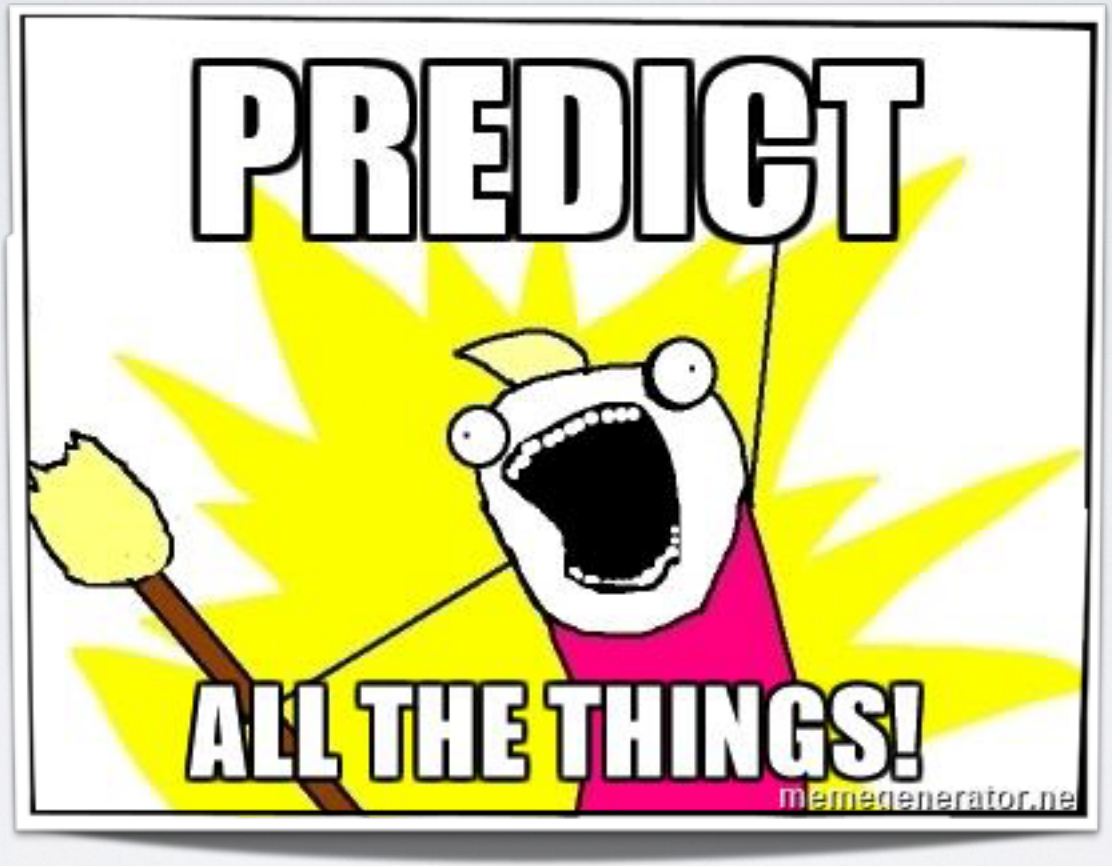
1. **abundant data and computation**

   Google Scholar, PubMed, Web of Science, arXiv, JSTOR, OCRID, EasyChair, etc.
   supercomputers, cloud computing, etc.

2. **interdisciplinary community**

   sociologists, physicists, computer scientists, biologists, economists, statisticians, etc.

3. **surely all this data must enable better predictions of future discoveries!**


PREDICT
ALL THE THINGS!
memegenerator.net

can we predict researcher productivity?

EASY **the canonical narrative of life-long productivity**

1. rapid rise to an early peak
2. decline or flattening

EASY the canonical narrative of life-long productivity

1. rapid rise to an early peak
2. decline or flattening

Publication rates in psychology, 1986 ✓

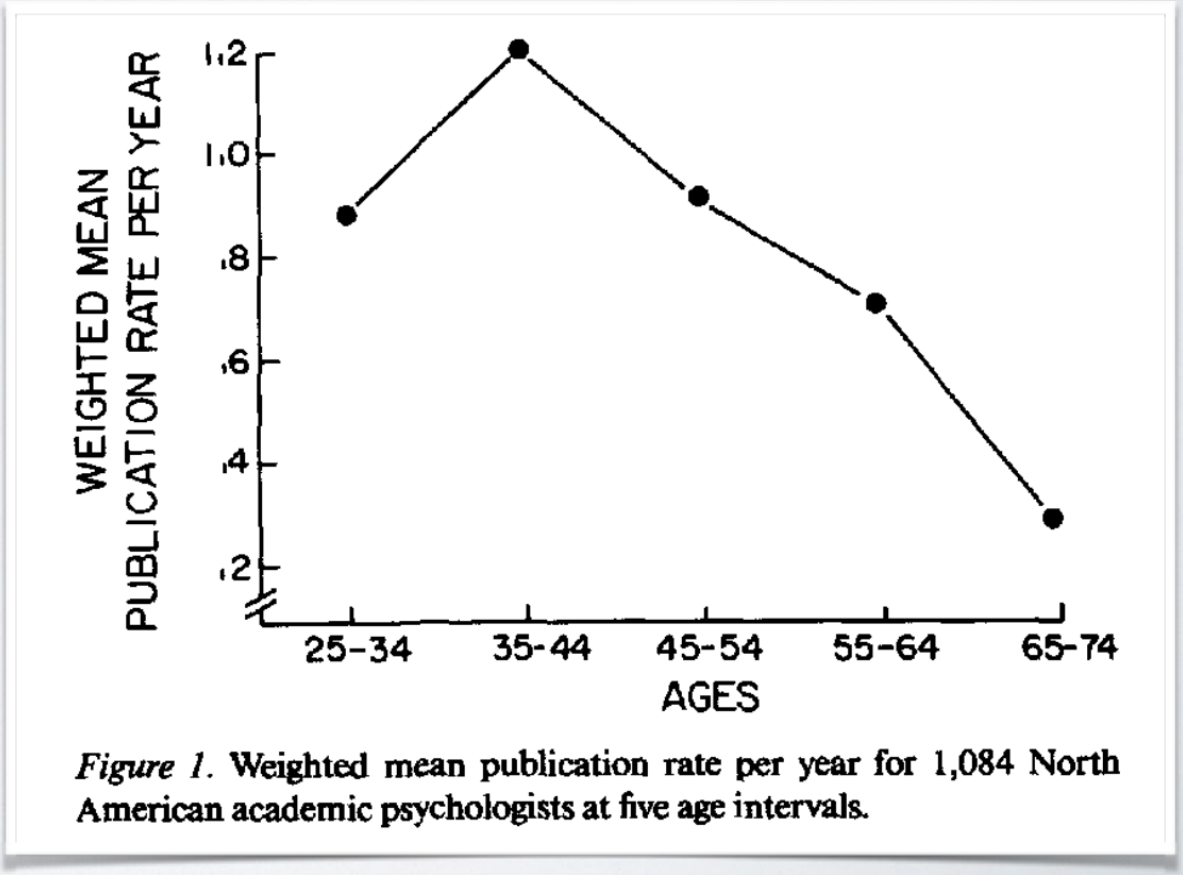


Figure 1. Weighted mean publication rate per year for 1,084 North American academic psychologists at five age intervals.

Horner, et al. *Psychology and Aging* **1** (4), 319 (1986)

EASY the canonical narrative of life-long productivity

1. rapid rise to an early peak
2. decline or flattening

Publication rates in psychology, 1986 ✓

. . . in Russian science & math, 1954 ✓

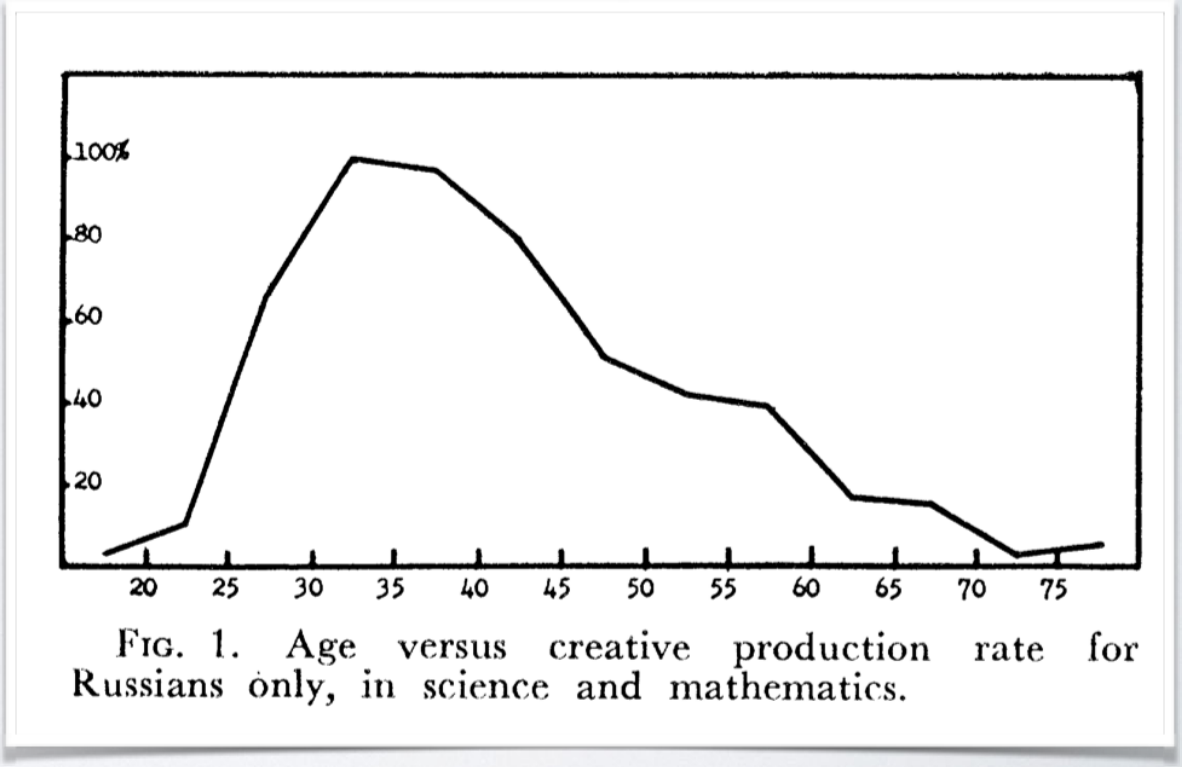


FIG. 1. Age versus creative production rate for Russians only, in science and mathematics.

Lehman, *The Scientific Monthly* **78**, 321-326 (1954)

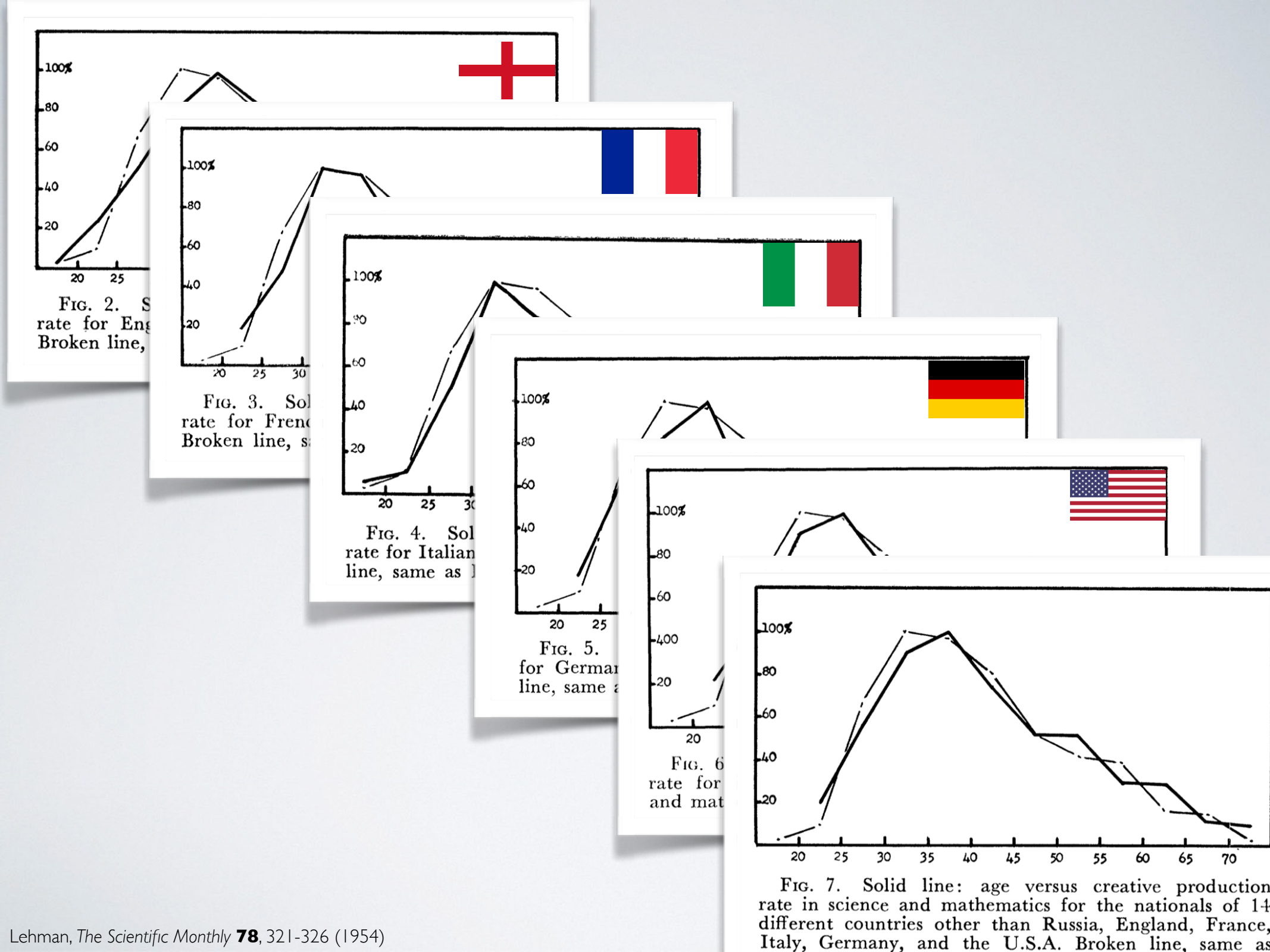


FIG. 2. S
rate for Eng
Broken line,

FIG. 3. Sol
rate for Frenc
Broken line, s

FIG. 4. Sol
rate for Italian
line, same as

FIG. 5.
for Germa
line, same a

FIG. 6
rate for
and mat

FIG. 7. Solid line: age versus creative production rate in science and mathematics for the nationals of 14 different countries other than Russia, England, France, Italy, Germany, and the U.S.A. Broken line, same as

Lehman, *The Scientific Monthly* **78**, 321-326 (1954)

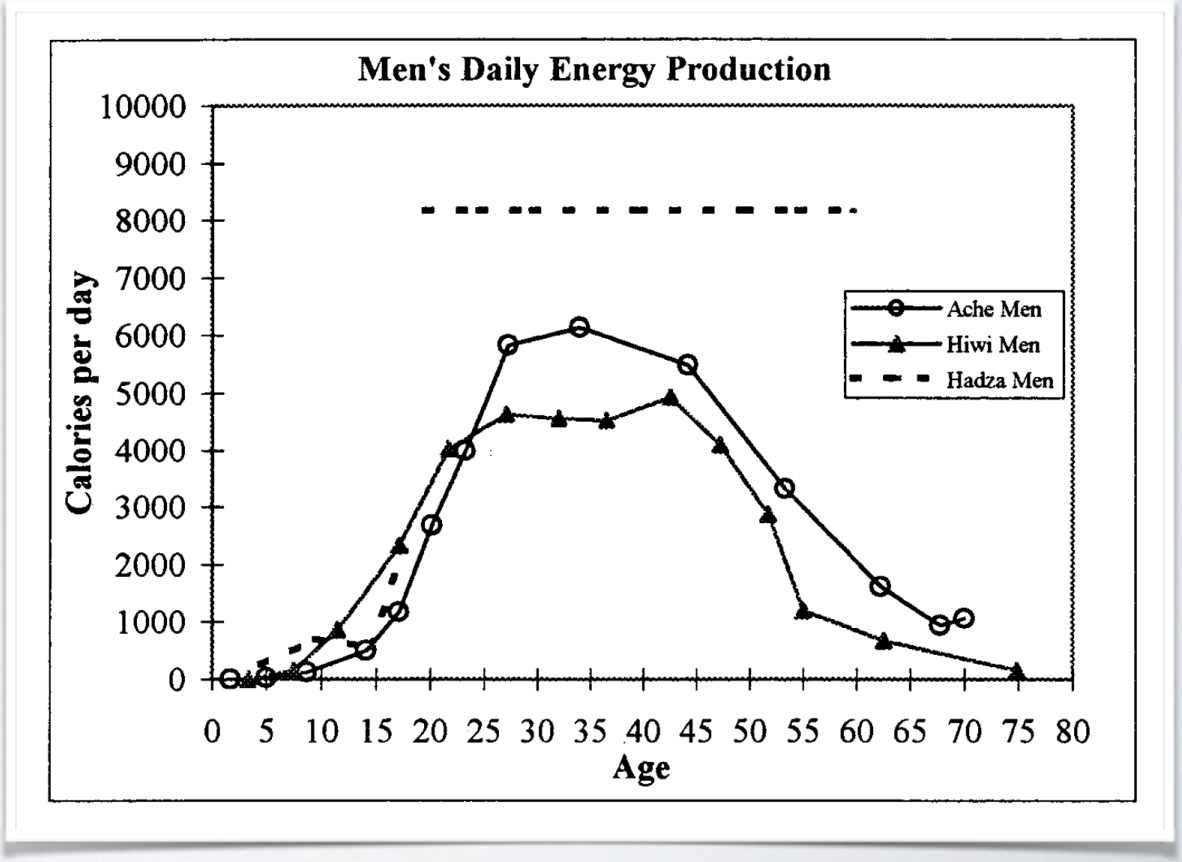
EASY the canonical narrative of life-long productivity

1. rapid rise to an early peak
2. decline or flattening

Publication rates in psychology, 1986 ✓

. . . in ~~Russian~~ science & math, 1954 ✓

. . . hunter-gather groups ✓



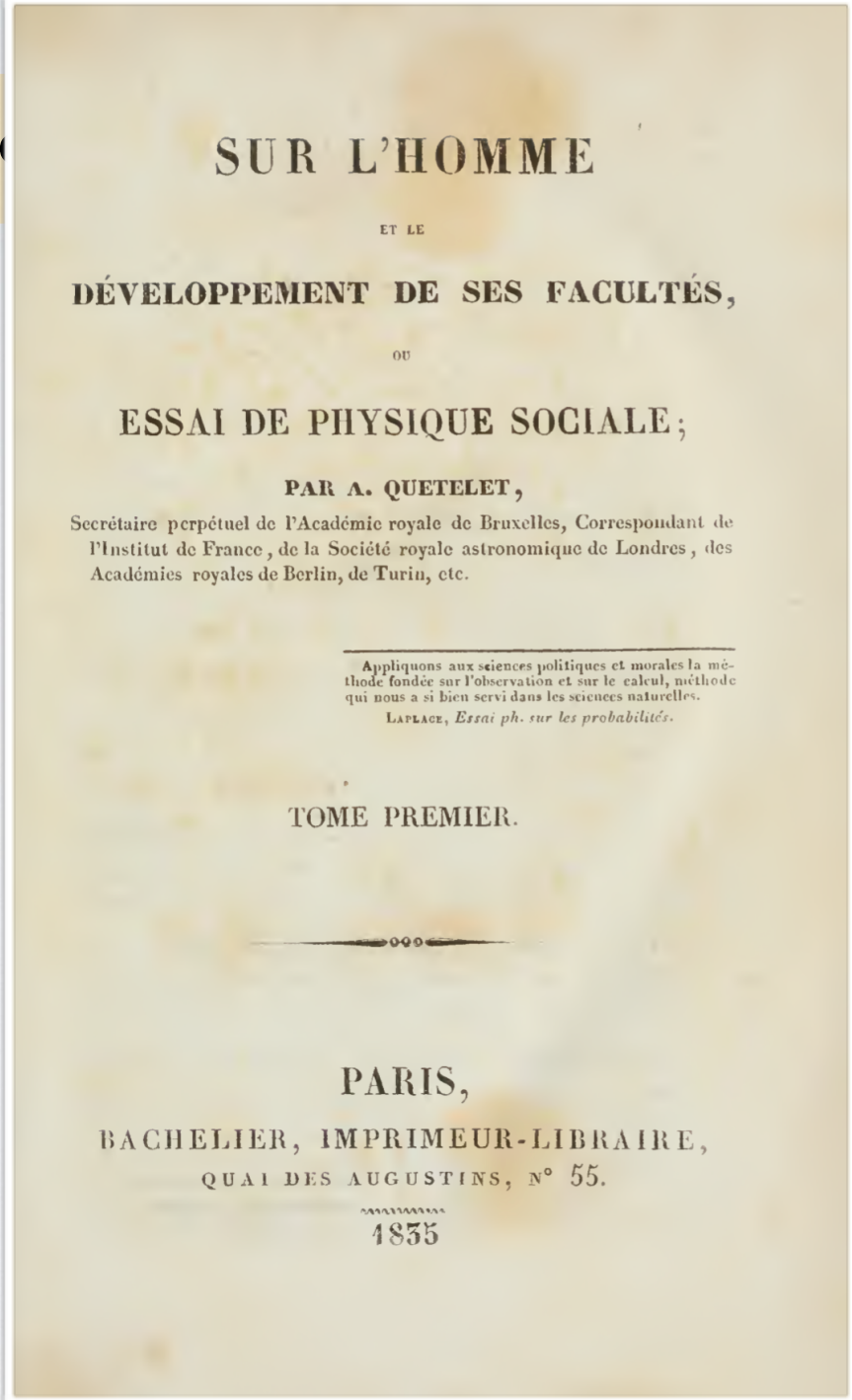
Kaplan et al., *Evol. Anth.: Issues, News, and Rev.*, **9**, 156-185 (2000)

EASY **the canonical narrative of lif**

1. rapid rise to an early peak
2. decline or flattening

Publication rates in psychology, 1986 ✓
. . . in ~~Russian~~ science & math, 1954 ✓
. . . hunter-gather groups ✓
. . . French & Philly criminals, 1835 ✓
. . . French artists, 1835 ✓
. . . Many others, 1950s - present ✓

60+ years of evidence = very predictable



SUR L'HOMME

ET LE

DÉVELOPPEMENT DE SES FACULTÉS,

OU

ESSAI DE PHYSIQUE SOCIALE;

PAR A. QUETELET,

Secrétaire perpétuel de l'Académie royale de Bruxelles, Correspondant de l'Institut de France, de la Société royale astronomique de Londres, des Académies royales de Berlin, de Turin, etc.

Appliquons aux sciences politiques et morales la méthode fondée sur l'observation et sur le calcul, méthode qui nous a si bien servi dans les sciences naturelles.

LAPLACE, *Essai ph. sur les probabilités.*

TOME PREMIER.

PARIS,

BACHELIER, IMPRIMEUR-LIBRAIRE,

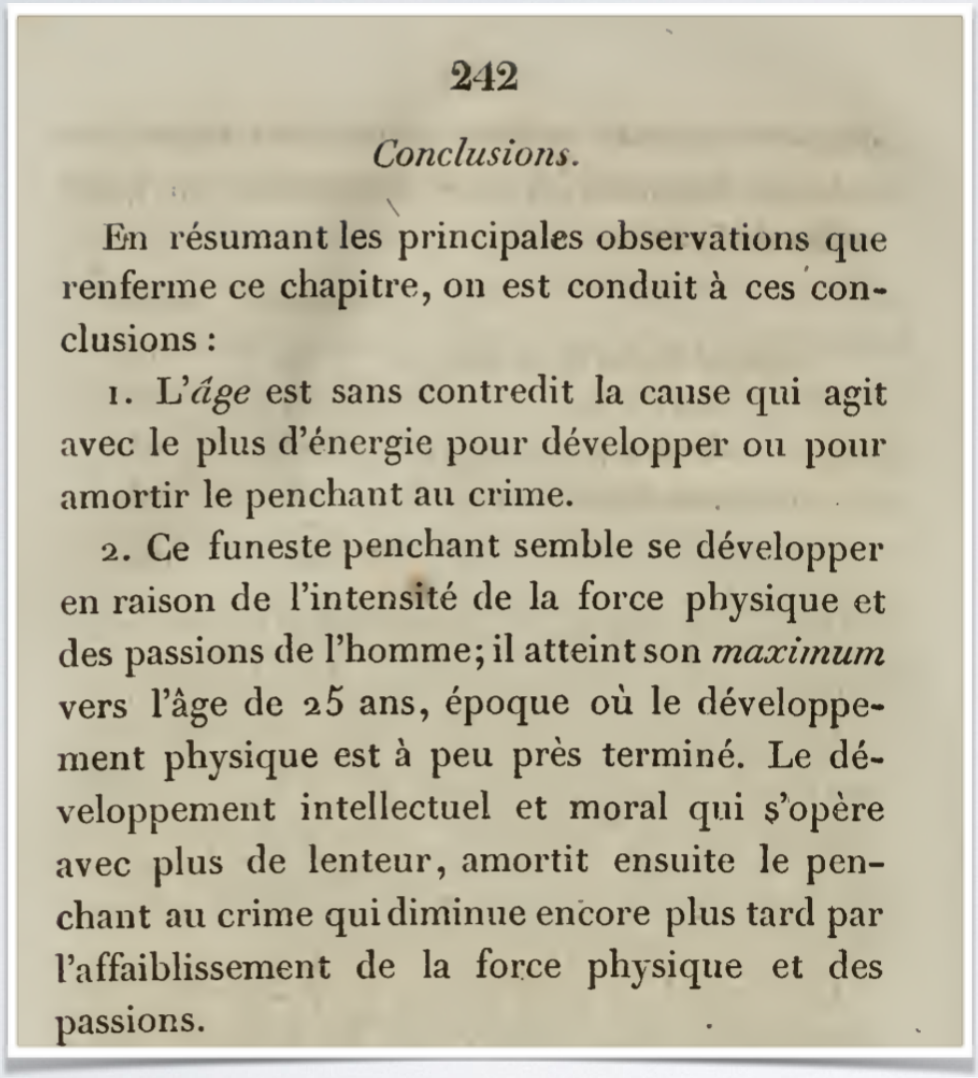
QUAI DES AUGUSTINS, N° 55.

1835

Quetelet, 1835.

NATURE DES CRIMES.	MOINS de 16 ans.	16 à 21 ans.	21 à 25 ans.	25 à 30 ans.	30 à 35 ans.	35 à 40 ans.	40 à 45 ans.	45 à 50 ans.	50 à 55 ans.	55 à 60 ans.	60 à 65 ans.	65 à 70 ans.	70 à 80 ans.	80 et au-dess.
Viol sur des enfans au-dessous de 15 ans...	4	120	71	96	73	39	34	45	22	18	26	17	21	2
Vols domestiques...	54	965	845	766	528	351	249	207	112	56	61	34	14	»
Autres vols.....	332	2479	2050	2292	1716	1249	1016	707	433	263	190	98	65	10
Viol et attentat à la pudeur.....	9	155	156	148	99	38	40	27	9	5	3	1	2	»
Parricide.....	6	13	12	13	6	3	2	1	4	2	»	»	»	»
Blessures et coups...	6	180	300	359	219	129	101	95	55	35	23	10	7	1
Meurtre.....	15	139	198	275	172	103	84	49	48	30	25	17	9	»
Infanticide.....	1	40	99	134	76	44	30	8	7	1	8	4	2	»
Rébellion.....	5	67	129	156	115	51	51	35	29	16	16	5	5	»
Vol sur chem. public.	21	80	111	149	107	60	62	46	22	21	8	6	4	»
Assassinat.....	10	90	144	203	183	100	104	89	53	32	24	13	15	1
Blessures envers un ascendant.....	2	47	64	73	72	40	30	16	8	2	1	»	»	»
Empoisonnement....	5	6	17	30	27	15	20	12	6	2	5	4	1	»
Faux témoignage et subornation.....	2	23	46	48	44	42	42	35	23	15	15	11	7	»
Faux divers.....	8	86	202	276	312	244	207	185	129	78	75	28	28	2

Quetelet, 1835.



242

Conclusions.

En résumant les principales observations que renferme ce chapitre, on est conduit à ces conclusions :

1. L'*âge* est sans contredit la cause qui agit avec le plus d'énergie pour développer ou pour amortir le penchant au crime.

2. Ce funeste penchant semble se développer en raison de l'intensité de la force physique et des passions de l'homme; il atteint son *maximum* vers l'âge de 25 ans, époque où le développement physique est à peu près terminé. Le développement intellectuel et moral qui s'opère avec plus de lenteur, amortit ensuite le penchant au crime qui diminue encore plus tard par l'affaiblissement de la force physique et des passions.

This fatal inclination appears to be developed on account of the intensity of the physical strength and passions of man; **It reaches its peak at the age of 25**, when physical development is nearly complete. The intellectual and moral development which takes place more slowly, then **cushions the penitent** to the crime, which **diminishes even later** by the weakening of physical strength and passions.

Google Translate

Quetelet, 1835.

EASY **the canonical narrative of life-long productivity**

1. rapid rise to an early peak
2. decline or flattening

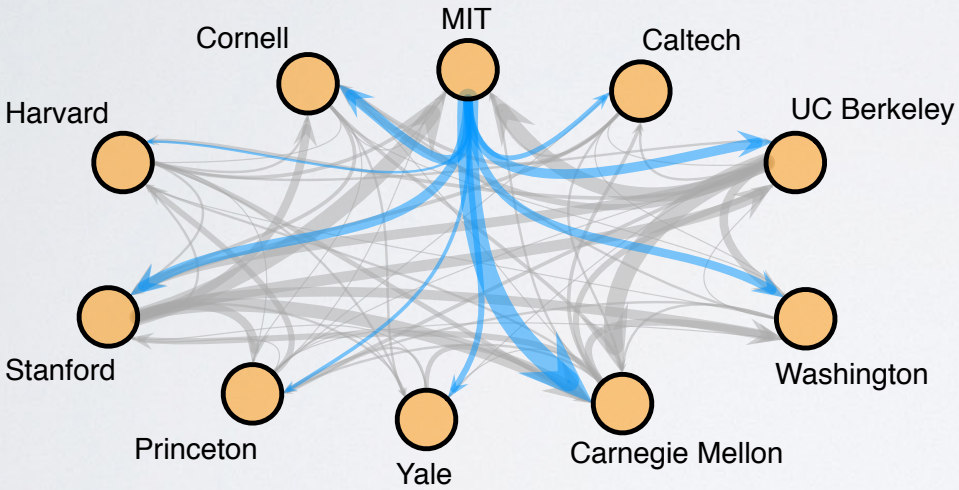
Publication rates in psychology, 1986 ✓
. . . in ~~Russian~~ science & math, 1954 ✓
. . . hunter-gather groups ✓
. . . French & Philly criminals, 1835 ✓
. . . French artists, 1835 ✓
. . . Many others, 1950s - present ✓

what about today?

scientific productivity among computer scientists

hiring data on all CS faculty in North America

all published papers for same faculty



Colorado Faculty Hiring Project

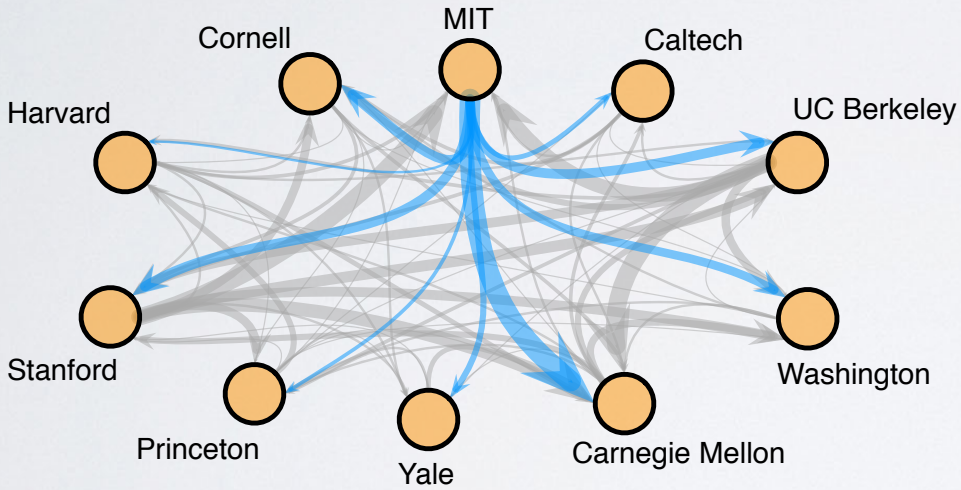
dblp
computer science bibliography

Clauset et al., *Science Advances* **1**, e1400005 (2015)

scientific productivity among computer scientists

hiring data on all CS faculty in North America

all published papers for same faculty



Colorado Faculty Hiring Project

WARNING
messy data

dblp
computer science bibliography

can't trust raw productivity counts

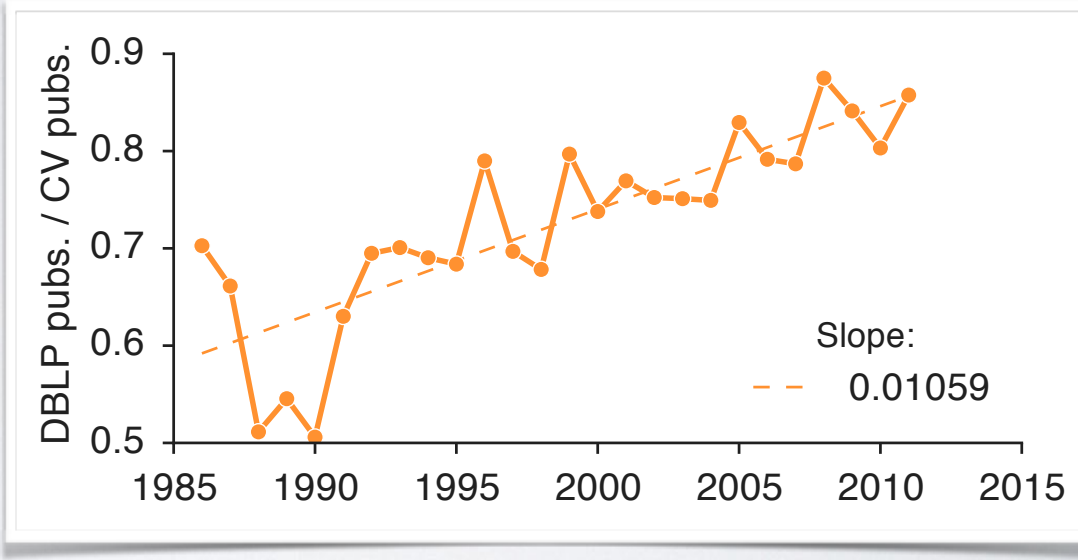
DBLP coverage varies over time

- for 10% of faculty, download and parse their CVs manually
- compare against DBLP records

DBLP coverage varies over time

- for 10% of faculty, download and parse their CVs manually
- compare against DBLP records
- systematic change in coverage:

  indexing all venues
  scope of computer science
  digitization / accessibility



- **correction** : adjust pub counts upward based on coverage trend

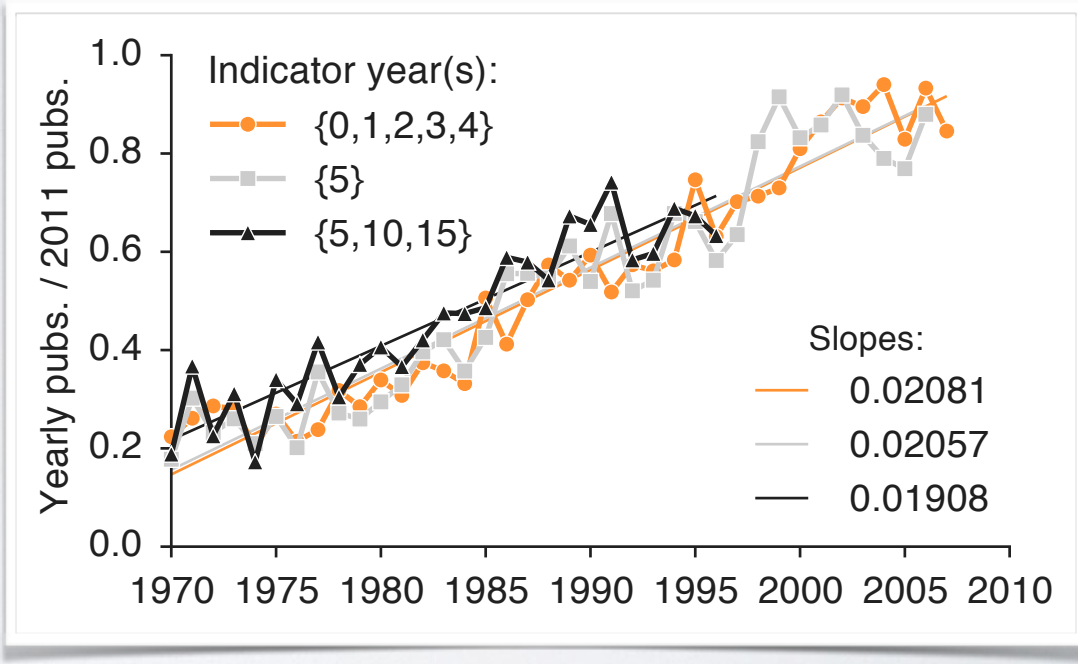
all faculty are getting more productive

- even after adjustment,
  publication rates still rising

all faculty are getting more productive

- even after adjustment, publication rates still rising
- consistent rate of growth:

  +1 paper per year / decade
- independent of which "indicator" year we compare across careers



- **correction** : adjust for coverage + adjust for growth

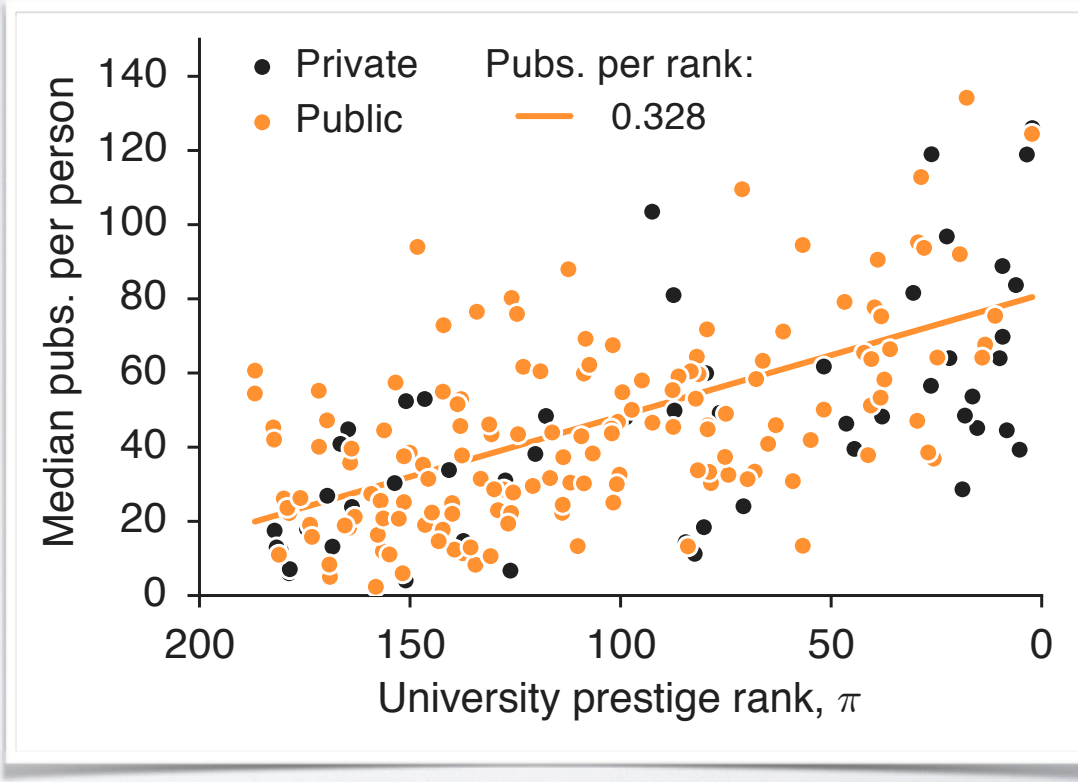
elite departments also more productive

- median papers in 10 years after first hire

elite departments also more productive

- median papers in 10 years after first hire
- greater prestige = greater productivity

  +3.3 pubs / year per 100 rank difference
- no difference between publics vs. privates



life-long productivity in computer science

- correct for coverage
- correct for general rate growth
- average publications vs. years post-hire

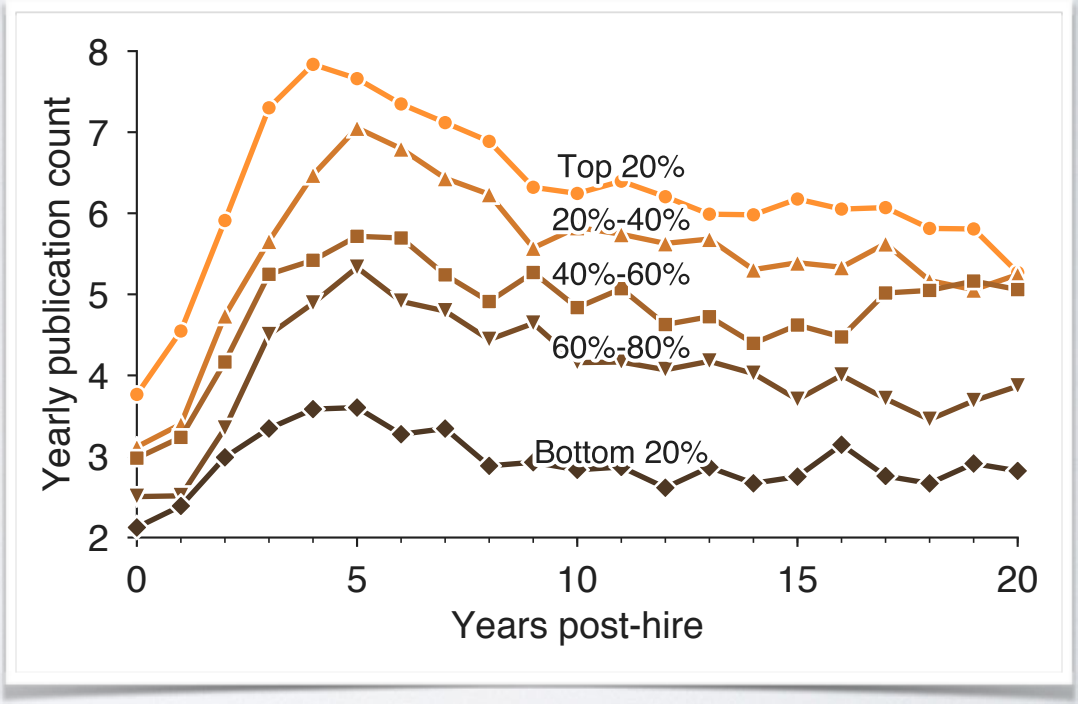
life-long productivity in computer science

- correct for coverage
- correct for general rate growth
- average publications vs. years post-hire

  consistent shape!

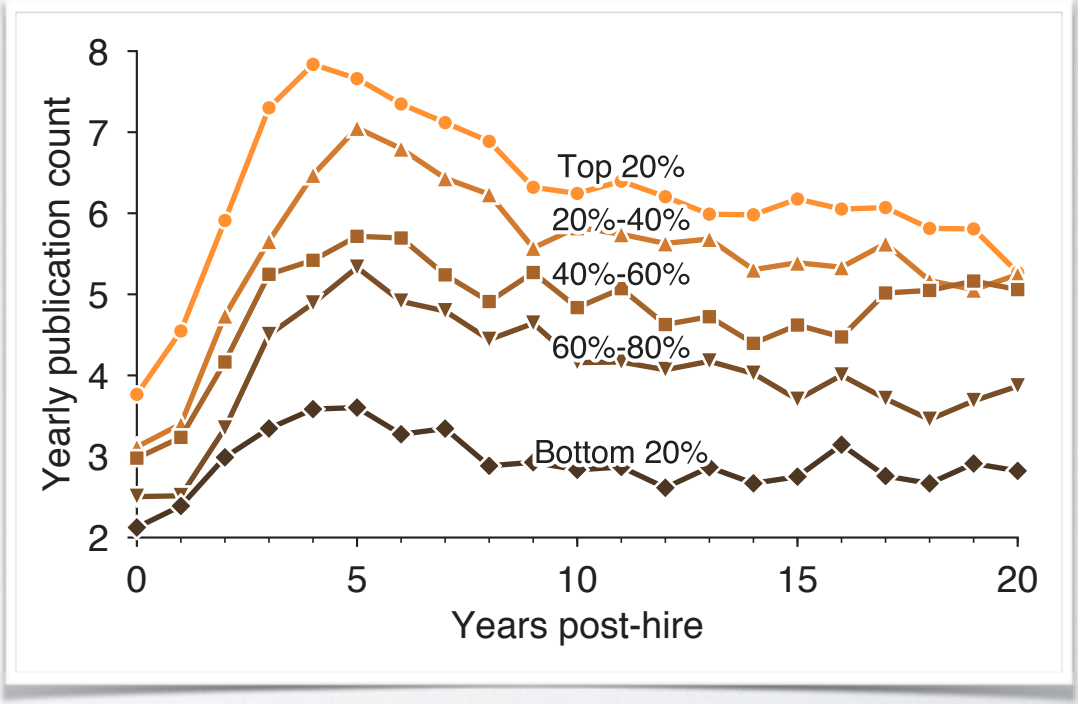
  shifts upward with prestige

  looks familiar!



EASY the canonical narrative of life-long productivity

1. rapid rise to an early peak
2. decline or flattening



Publication rates in psychology, 1986 ✓

. . . in ~~Russian~~ science & math, 1954 ✓

. . . hunter-gather groups ✓

. . . French & Philly criminals, 1835 ✓

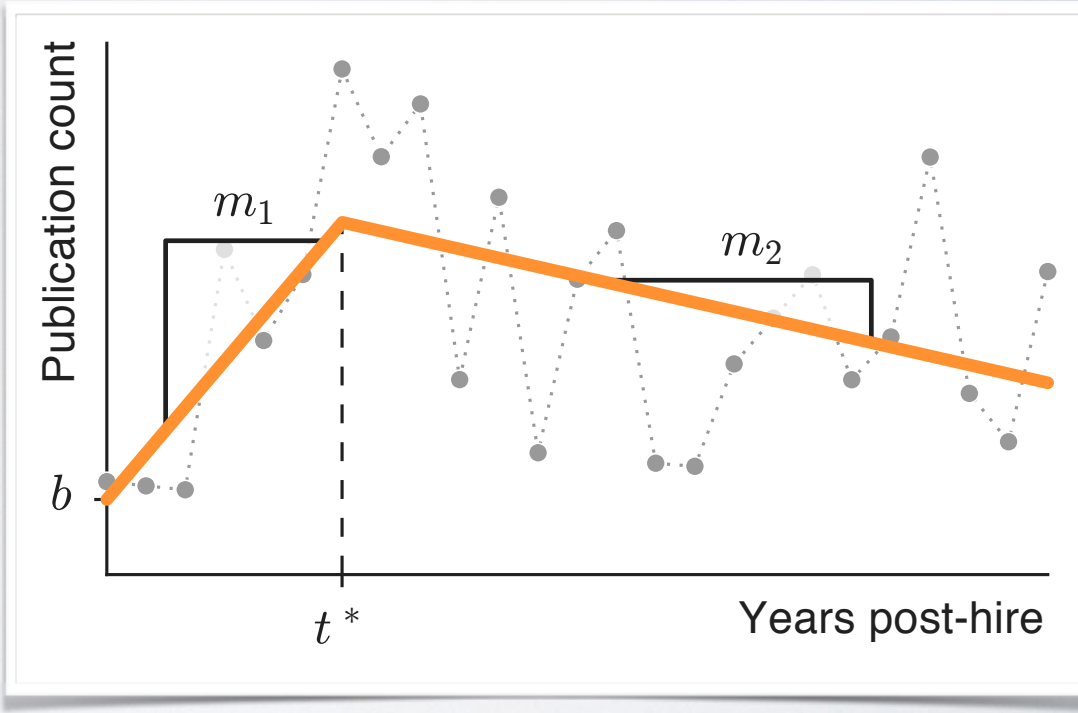
. . . French artists, 1835 ✓

. . . Many others, 1950s - present ✓

. . . Computer Science faculty ✓

does the average trajectory reflect *individual* behavior?

1. rapid rise to an early peak at t^*
2. decline or flattening



$$f(t) = \begin{cases} b + m_1 t & 0 \le t \le t^* \\ b + m_1 t^* + m_2 (t - t^*) & t \ge t^* \end{cases}$$

does the average trajectory reflect *individual* behavior?

1. rapid rise to an early peak at t^*
2. decline or flattening

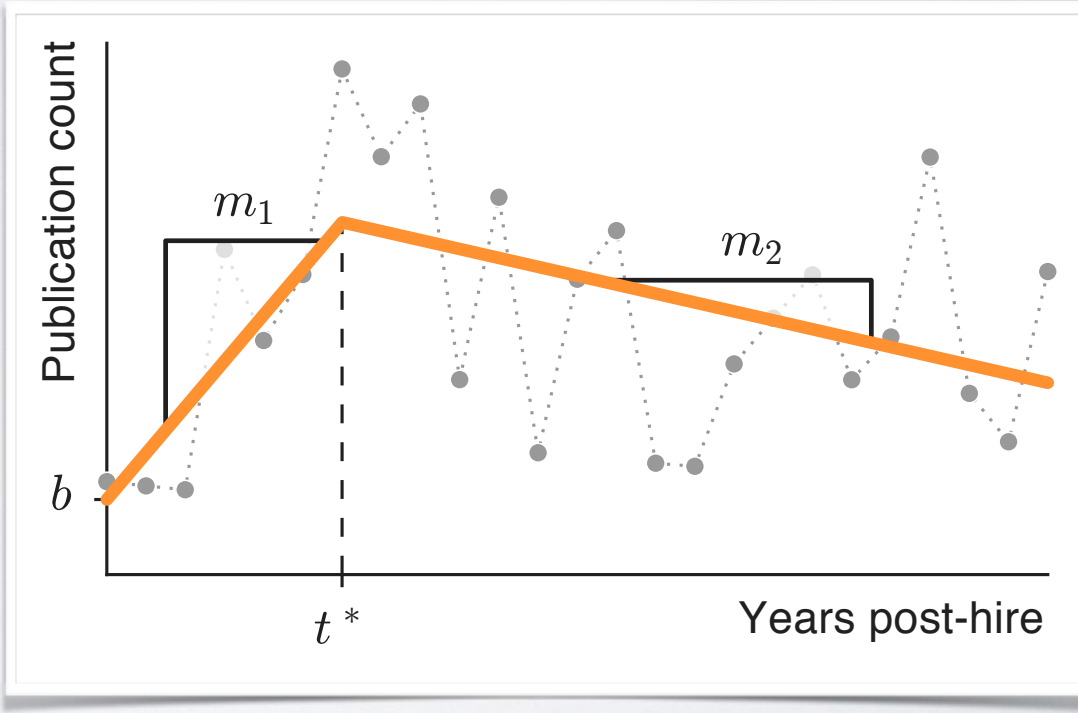
4 conditions on parameters:

"rise" $m_1 > 0$

"early peak" $t^* \leq \frac{1}{2}|t_f - t_0|$

"decline" $m_2 \leq 0$

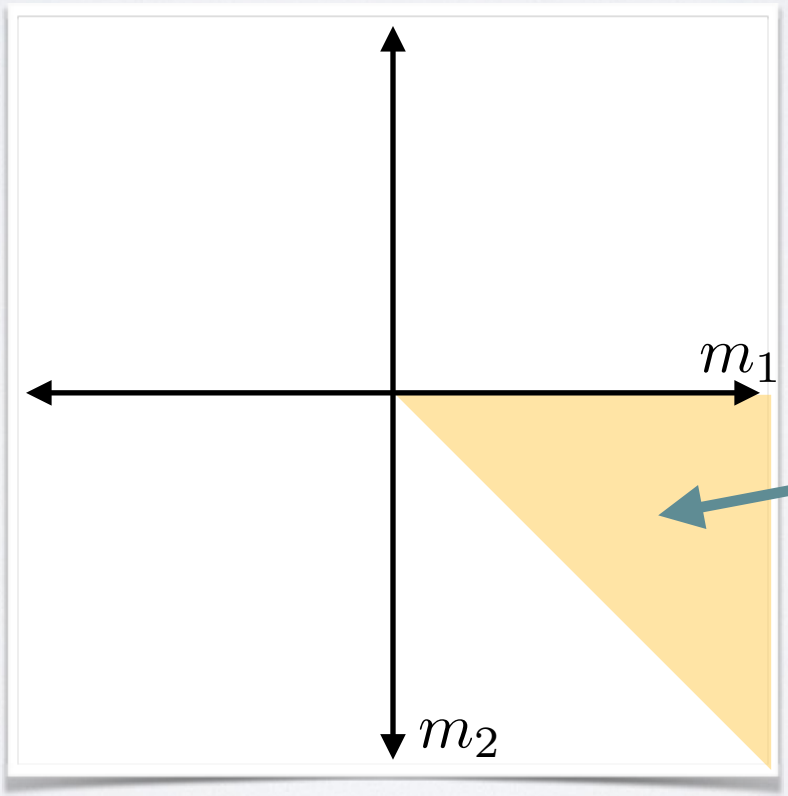
"rapid" $|m_1| > |m_2|$



$$f(t) = \begin{cases} b + m_1 t & 0 \leq t \leq t^* \\ b + m_1 t^* + m_2(t - t^*) & t \geq t^* \end{cases}$$

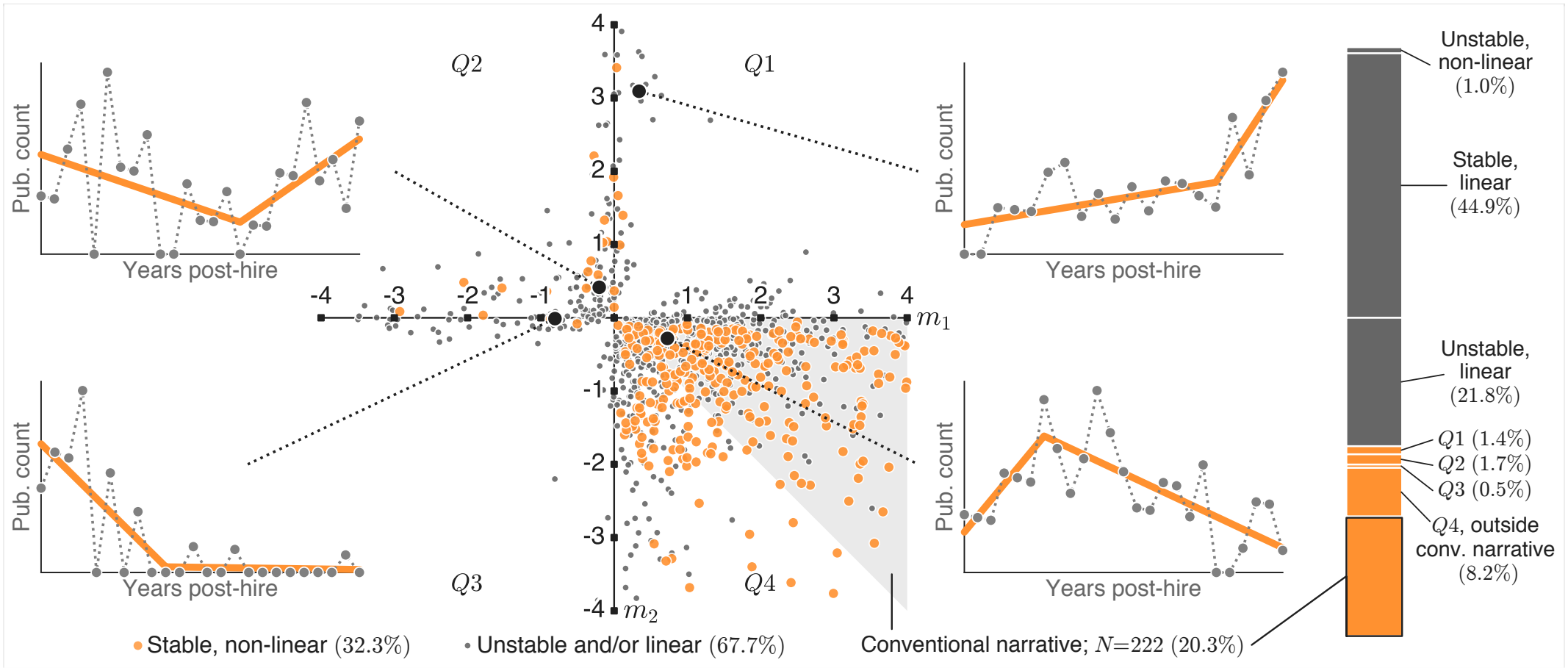
trajectories for individuals

- fit model to each individual
- exclude individuals better fit by a simple line $f(t) = b + m_1 t$
- plot scatter in the (m_1, m_2)-plane



Use model comparison to control complexity and bootstrap perturbation to determine stability of quadrant assignment

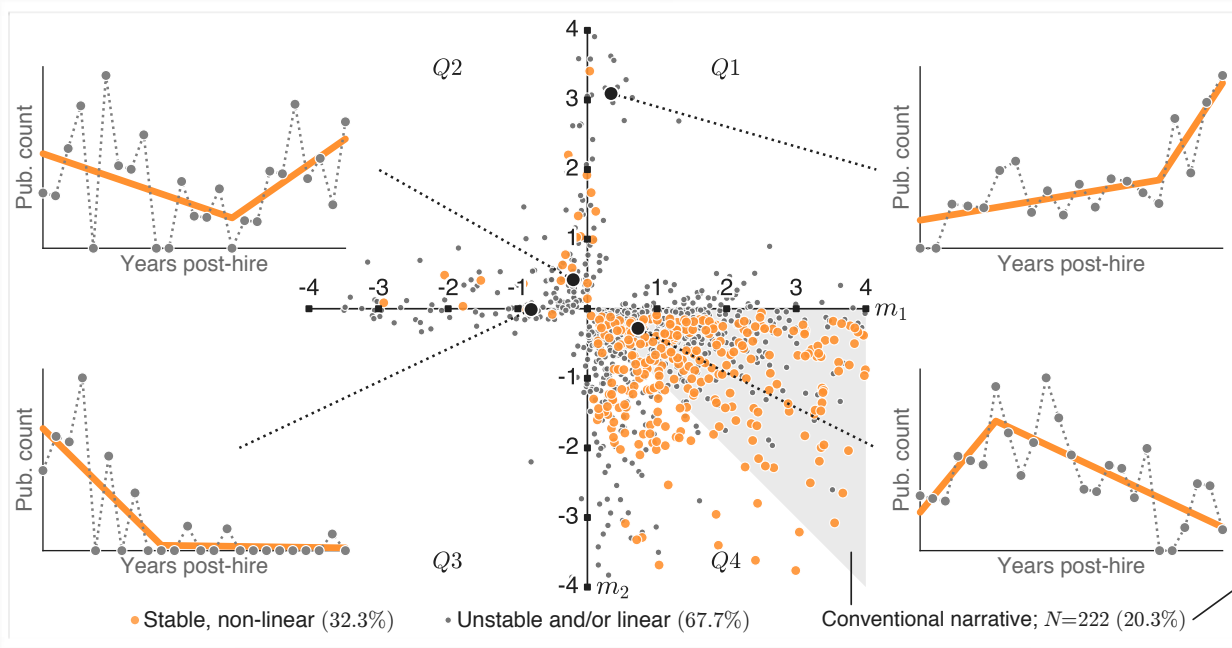
trajectories for individuals



trajectories for individuals

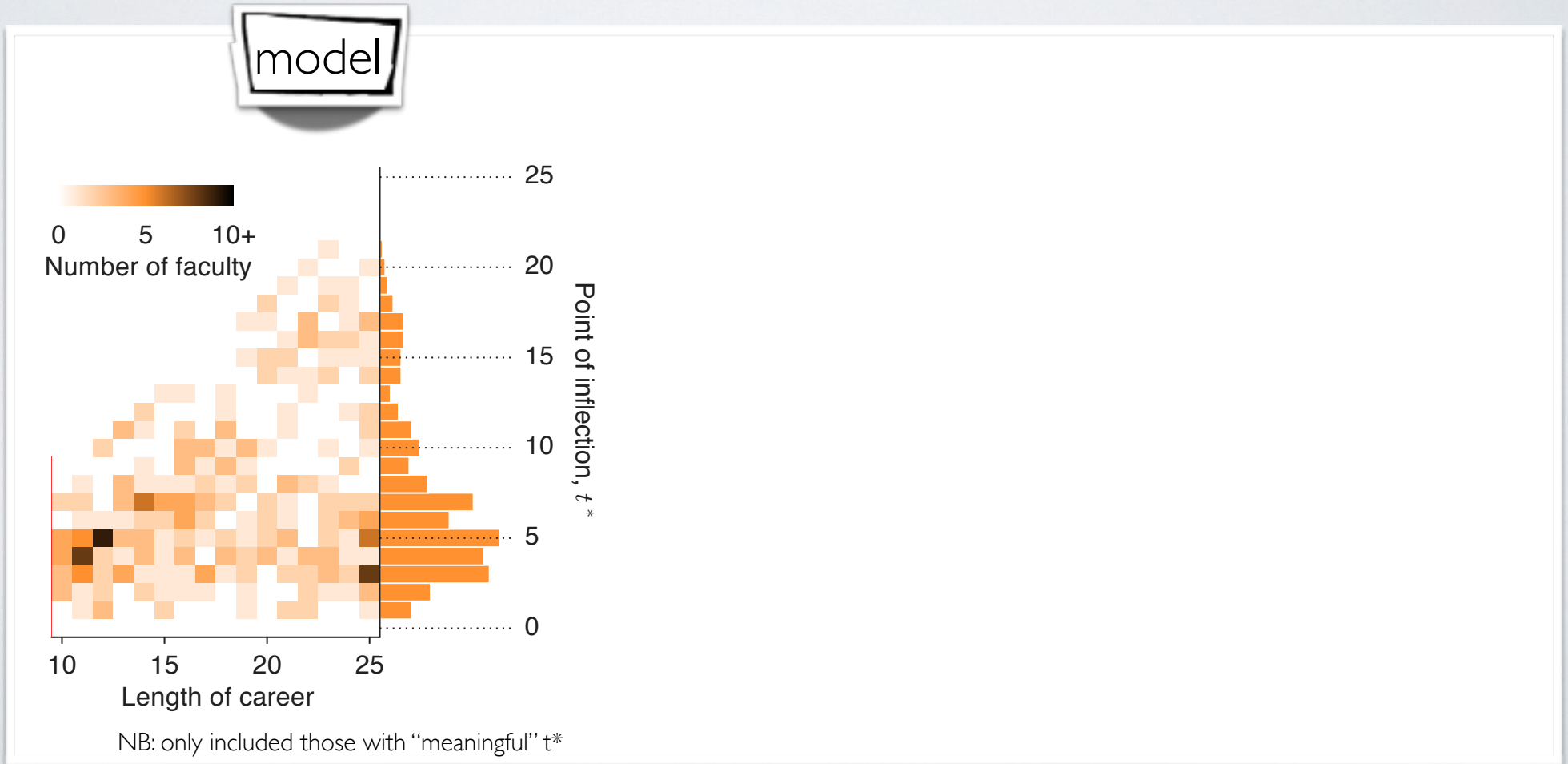
- 68% have unstable or linear fits
- only 20% follow canonical narrative
- men and women *statistically indistinguishable*, across quadrants and within canonical octant

- initial slope m_1 higher at more prestigious departments:
  1.21 p/yr vs. 0.75 p/yr
- intercept level correlates with past productivity = continuity (postdoc, prestige, etc.)



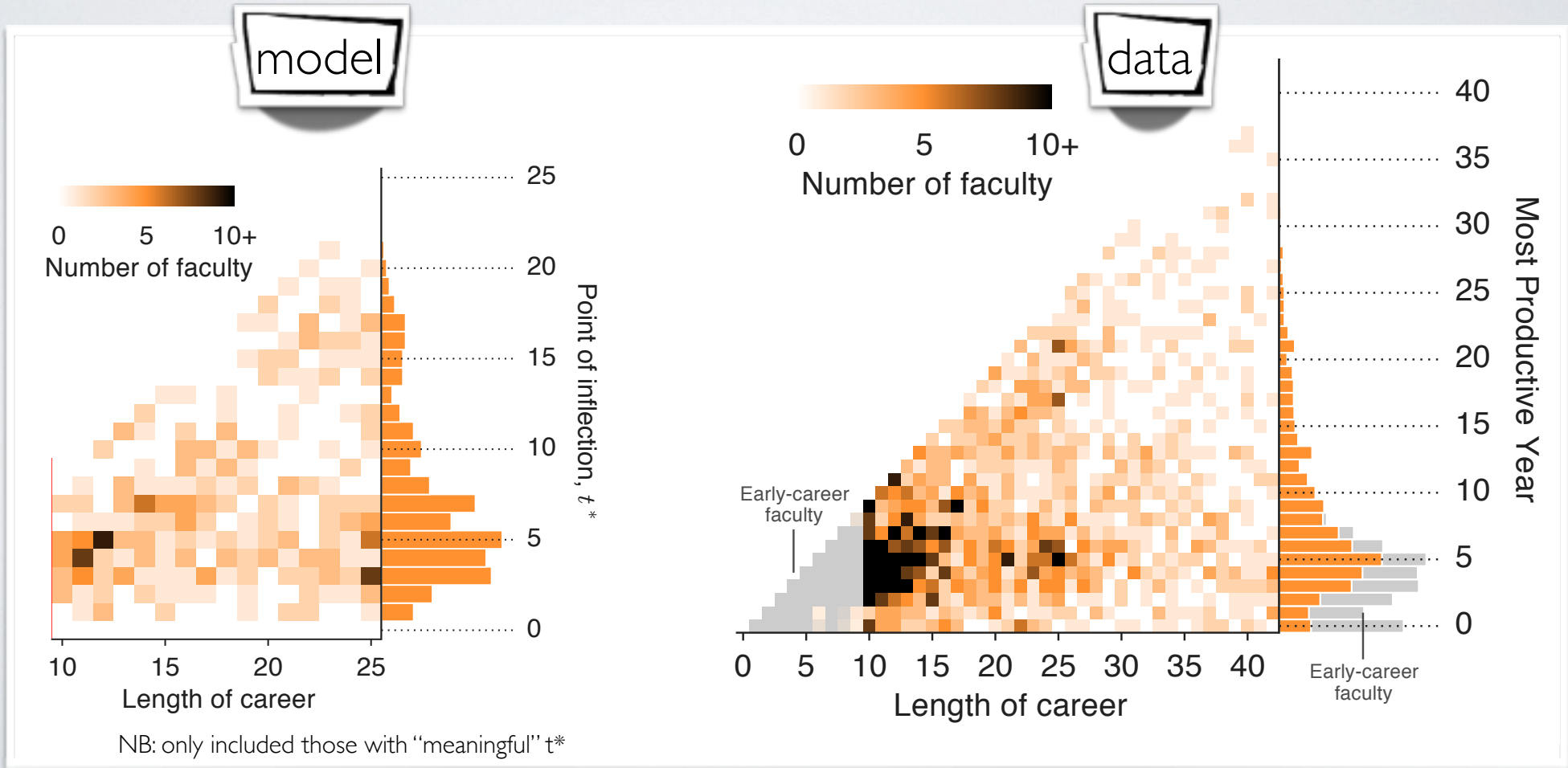
what about the change-point t^*?

- distribution of change points (model)



what about the change-point t^*?

- distribution of change points (model)
- distribution of peak productivity (data)

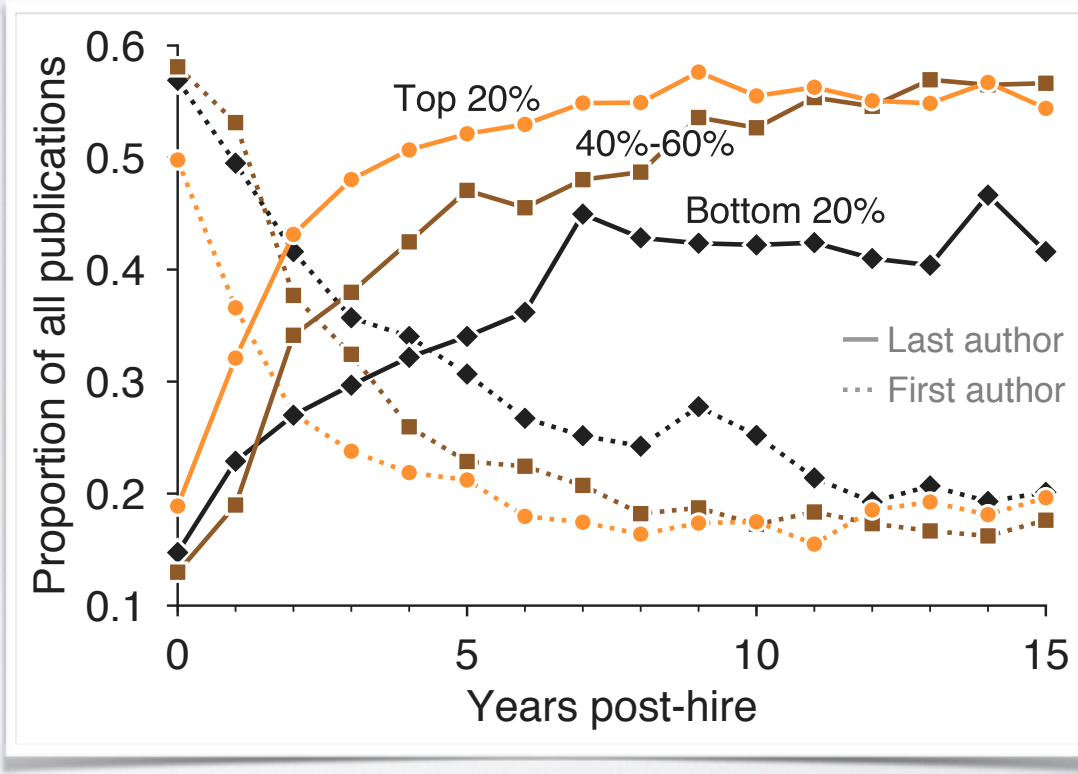


okay, maybe author order is predictable?

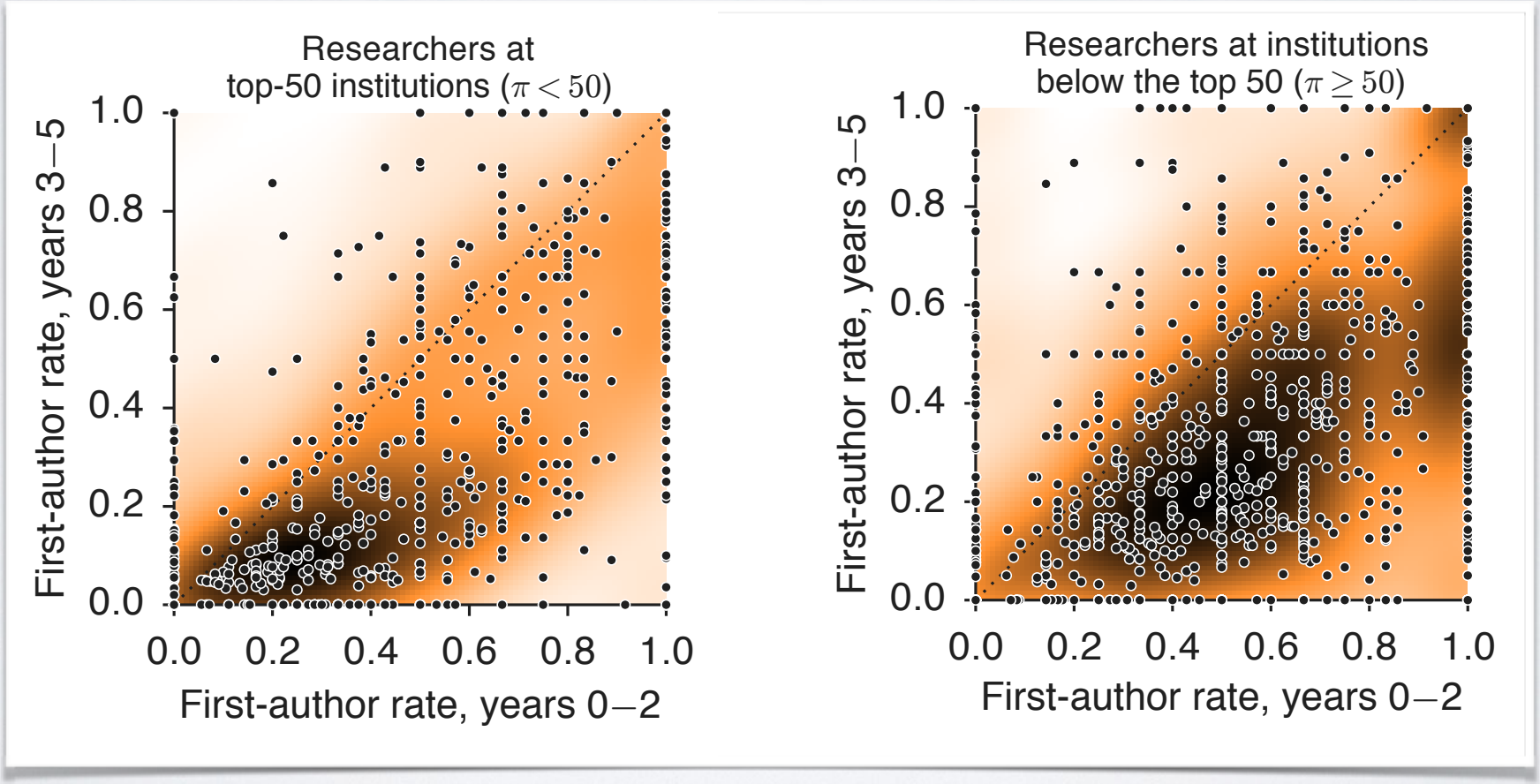
- first- and last-author rates
- (exclude alphabetical author orders, as in Theory of CS)

okay, maybe author order is predictable?

- first- and last-author rates
- (exclude alphabetical author orders, as in Theory of CS)
- consistent and reasonable pattern!
- transition to last-author role happens ≈ 2 years early for top-50 departments



but even first- and last-author rates are diverse



top 50 departments

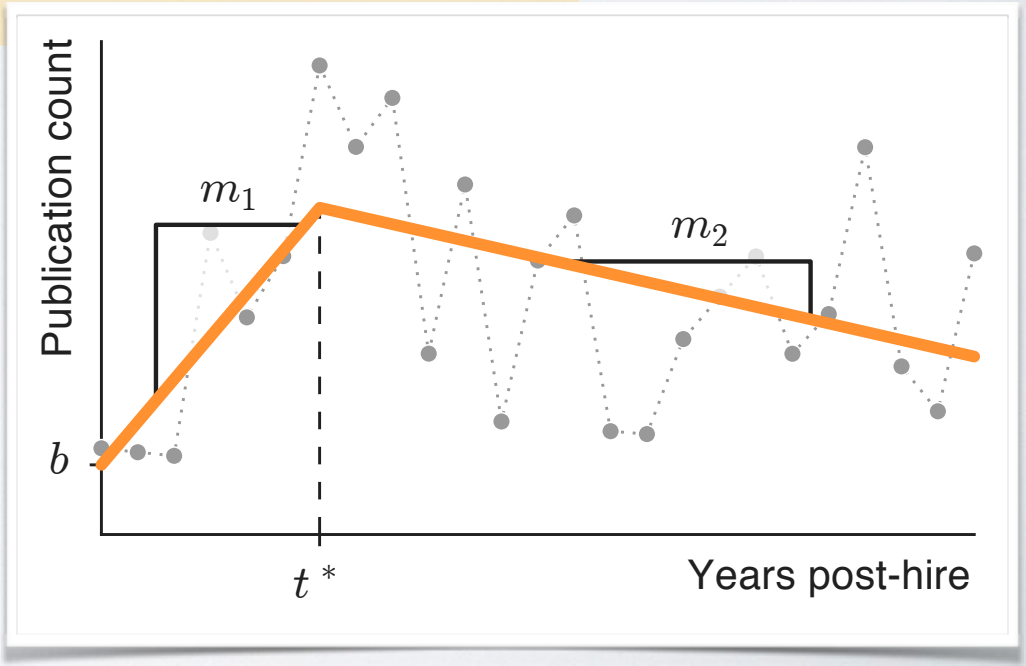
departments 50+

what we've learned

- the canonical narrative is *not* representative
- holds only for $1/5$ faculty — *most* faculty are different
- enormous diversity (unpredictability) in productivity patterns

future work

- does prestige *cause* productivity? how could we test this?
- what predicts model features? (tenure…)
- what might a minimal model look like?



predicting discoveries

predicting discoveries

some aspects of science are *highly predictable*

- most citation counts, institution of origin, maximum impact, etc.
- interdisciplinary research is harder to publish & fund
- under-represented groups (women, minorities) receive less funding

other aspects appear *fundamentally unpredictable*

- productivity over a career, timing of biggest discovery, etc.
- likely long-term impact of proposed project or manuscript

predicting discoveries

some aspects of science are *highly predictable*

- most citation counts, institution of origin, maximum impact, etc.
- interdisciplinary research is harder to publish & fund
- under-represented groups (women, minorities) receive less funding

other aspects appear *fundamentally unpredictable*

- productivity over a career, timing of biggest discovery, etc.
- likely long-term impact of proposed project or manuscript

bibliographic information is *abundant but crude*, and is a *lagging indicator* of scientific innovation

could we make better predictions with *more* data?

- contents of papers, preprints, workshops, research team communication, rejected manuscripts or proposals, peer reviews, post-publication reviews

risks of automation

citations and publications prone to *feedback loops*

- rich-get-richer dynamic
- can amplify inequalities
- if opportunities for future success allocated by markers of recent success
- can create self-fulfilling predictions, which narrow innovation

action item: what measures of success are not susceptible to feedback loops?

risks of automation

citations and publications prone to *feedback loops*

- rich-get-richer dynamic
- can amplify inequalities
- if opportunities for future success allocated by markers of recent success
- can create self-fulfilling predictions, which narrow innovation

action item: what measures of success are not susceptible to feedback loops?

selection by automatic prediction of future "impact"

- at request of universities or publishers
- can induce herding behavior (e.g., automated stock traders)
- need fairness, accountability and transparency (FAT) in ML
- need humans in the loop

action item: how do we build FAT ML principles into systems?

the future could be bright

science is a large and diverse ecosystem

science of science could *expand or contact* it

what lessons can we learn from ecology and evolutionary theory? design principles of robustness, diversifying selection, stabilizing feedback, etc.

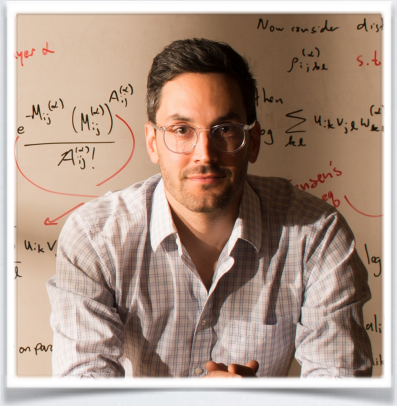
if discovery is inherently unpredictable, better to cultivate a diverse scientific ecosystem than try to automate its prediction

SPECIAL SECTION PREDICTION

ESSAY

Data-driven predictions in the science of science

Aaron Clauset,[1,2]* Daniel B. Larremore,[2] Roberta Sinatra[3,4]



Daniel B Larremore
(Colorado)

Roberta Sinatra
(Central Eur. U.)

Samuel F Way
(Colorado)

Allison C Morgan
(Colorado)

The misleading narrative of the canonical faculty productivity trajectory

Samuel F. Way[a,1], Allison C. Morgan[a], Aaron Clauset[a,b,c,2], and Daniel B. Larremore[a,b,c,1,2]

Way, et al., *Proc. Natl. Acad. Sci. USA* **114**, E9216-E9223 (2017)
Clauset, Larremore & Sinatra, *Science* **355**, 477-480 (2017)