

# Prediction and its limits for scientific discovery

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



# pervasive desire to predict science

---

- ▶ **what will be discovered?**
- ▶ **by whom, when, and where?**



# pervasive desire to predict science

- ▶ what will be discovered?
- ▶ by whom, when, and where?



	individuals	what questions are useful, 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?

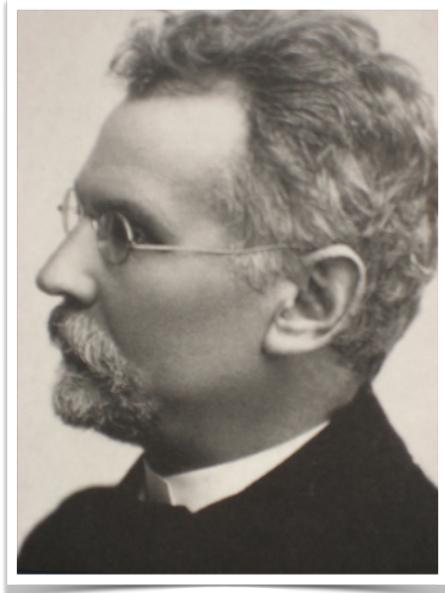
# *how predictable are scientific discoveries?*

---

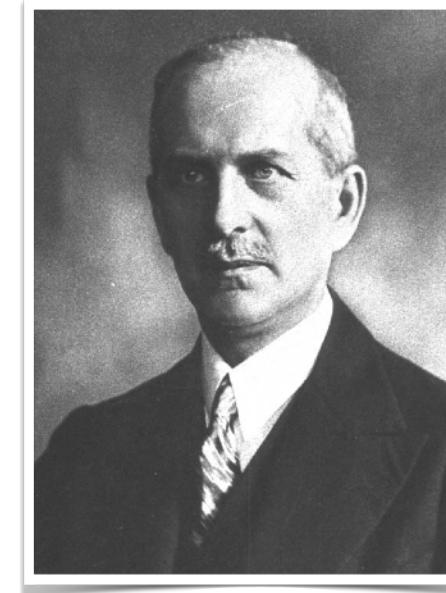
- ▶ **simple question with a 150+ year history**

# how predictable are scientific discoveries?

## ► simple question with a 150+ year history



Bolesław Prus  
(1847-1912)



Florian Znaniecki  
(1882-1958)



Freeman Dyson  
(1923-2020)



Steven Weinberg  
(1933-)



Harriet Zuckerman  
(1937-)

...

- philosophy, physics, sociology...
- 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

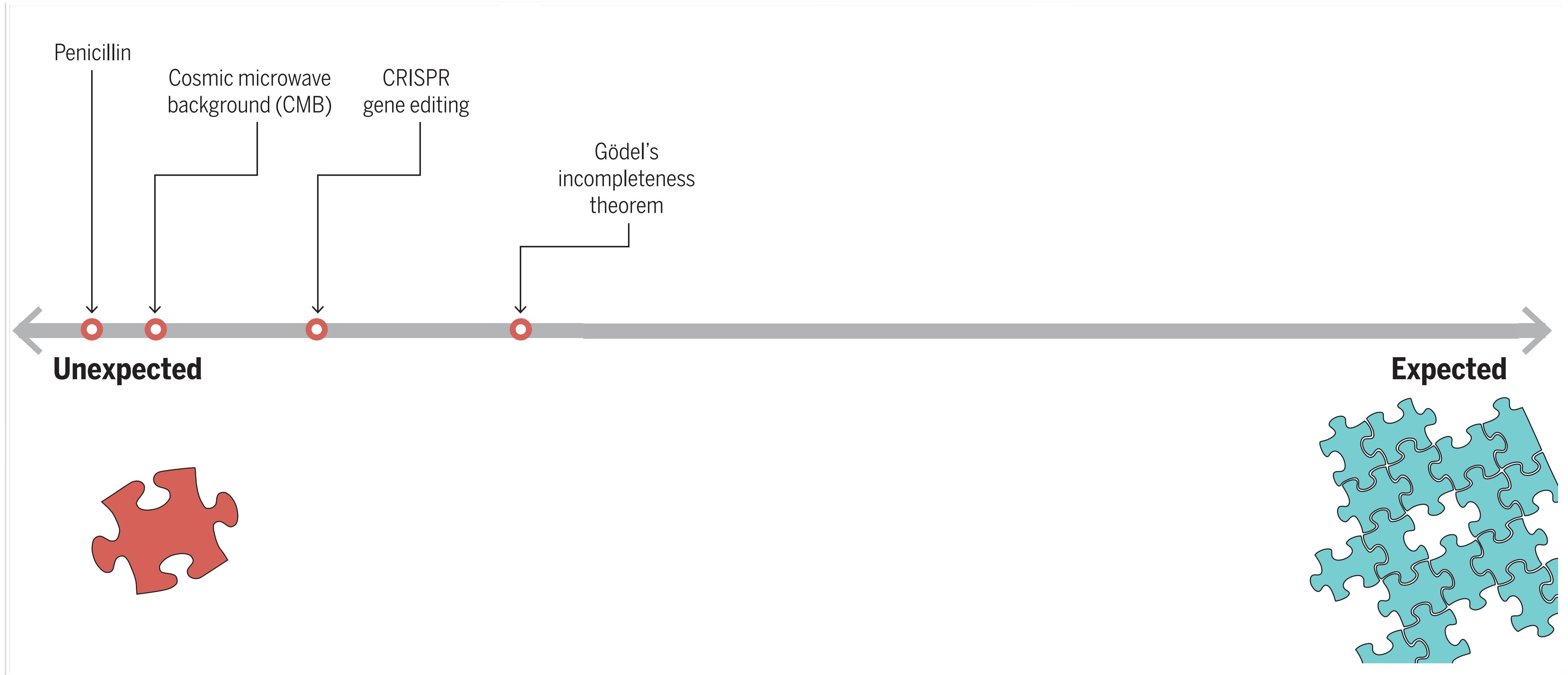
# predictability depends on context



# predictability depends on context

## unexpected discovery

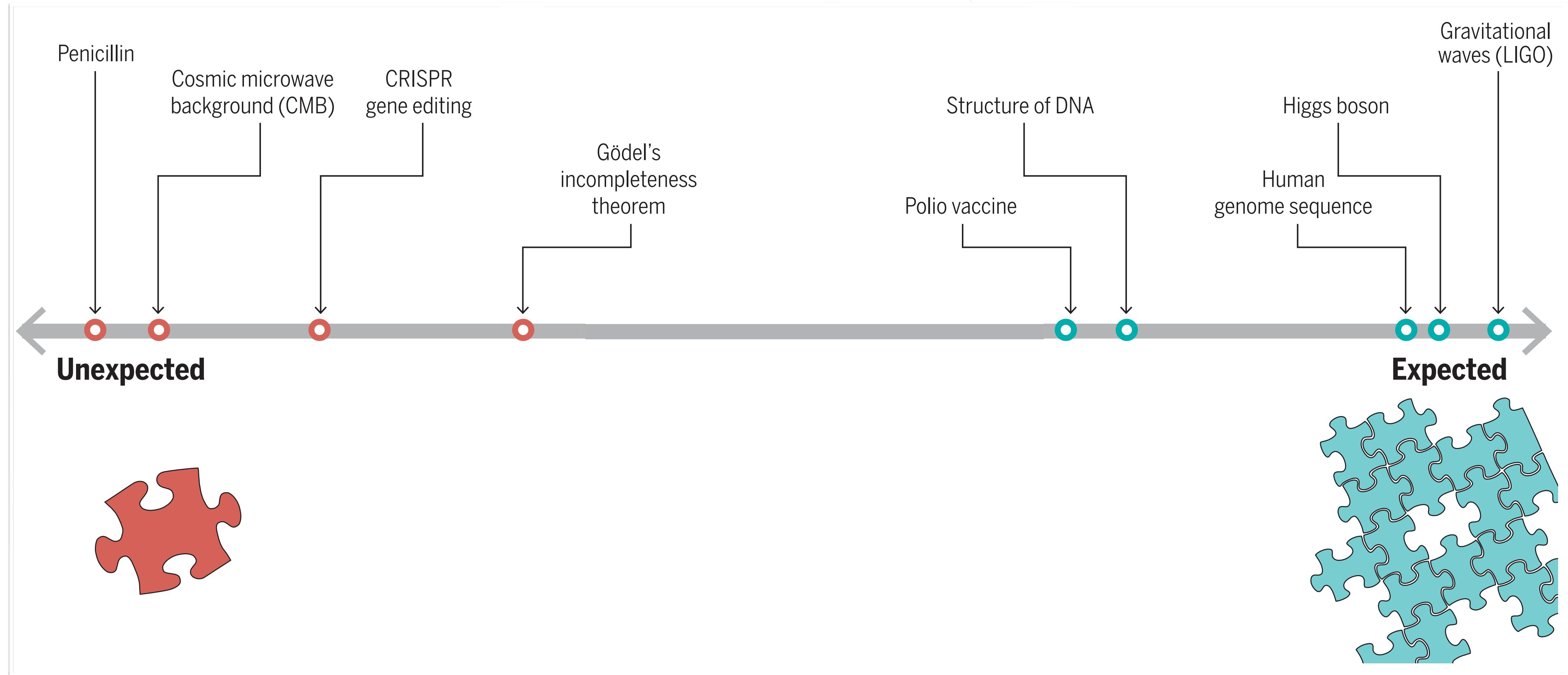
changes the way we understand the world, or finds novel use elsewhere



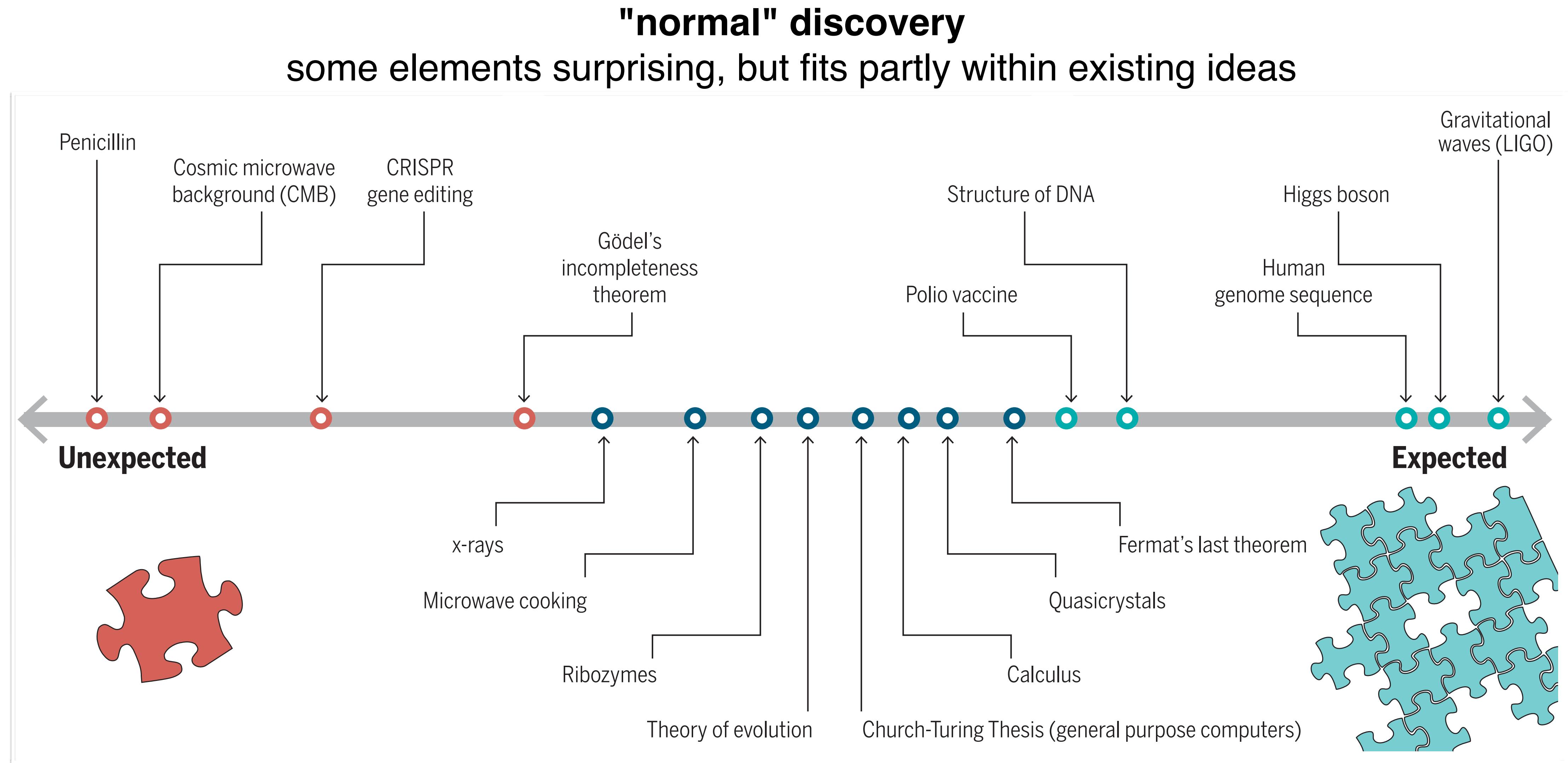
# predictability depends on context

**expected discovery**

accumulation of theory and evidence, fits with other ideas



# predictability depends on context



# a modern science of science

---

predicting discovery

# a modern science of science

predicting discovery

▶ abundant data

- (1) publications + citation networks,  
(2) people, (3) funding
- Google Scholar, PubMed, Web of Science, arXiv,  
JSTOR, ORCID, EasyChair, NIH, NSF, patents,  
CVs, etc.

▶ abundant computation

▶ growing interdisciplinary community

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

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

APS Data Sets for Research

WEB OF SCIENCE™



arXiv.org



PubMed

# a modern science of science

predicting discovery

► **surely all this data must enable better predictions of future discoveries?**



APS Data Sets for Research

WEB OF SCIENCE™



Google  
Scholar

ORCID

Connecting Research  
and Researchers

arXiv.org



PubMed

# a modern science of science

predicting discovery

- ▶ surely all this data must enable better predictions of future discoveries?
- ▶ yes, but...



APS Data Sets for Research

WEB OF SCIENCE™



Google  
Scholar

ORCID

Connecting Research  
and Researchers

arXiv.org



PubMed

# a modern science of science

predicting discovery

- ▶ surely all this data must enable better predictions of future discoveries?
- ▶ yes, but...

- the data are crude + biased + noisy + incomplete : *they don't directly measure knowledge or progress*
- what things are predictable and what things are not?



APS Data Sets for Research

WEB OF SCIENCE™



ORCID

Connecting Research  
and Researchers

arXiv.org



PubMed

# productivity over a career

---

the canonical narrative (50+ years of evidence):

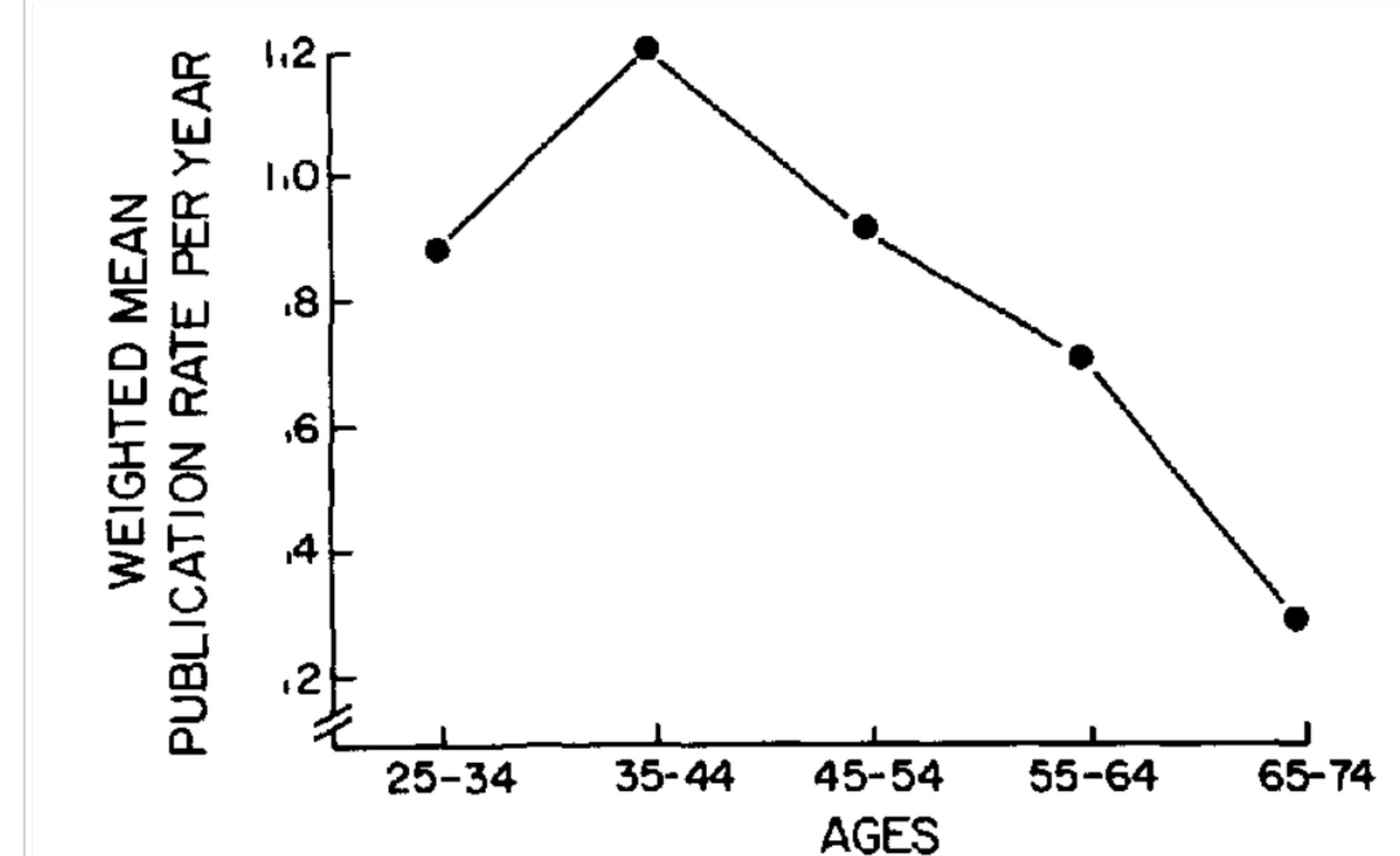
- ▶ **rapid rise to an early peak**
- ▶ **decline or flattening**

# productivity over a career

the canonical narrative (50+ years of evidence):

- ▶ **rapid rise to an early peak**
- ▶ **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.*

# productivity over a career

the canonical narrative (50+ years of evidence):

- ▶ **rapid rise to an early peak**
- ▶ **decline or flattening**

publication rates in psychology, 1986 ✓  
. . . in Russian science & math, 1954 ✓

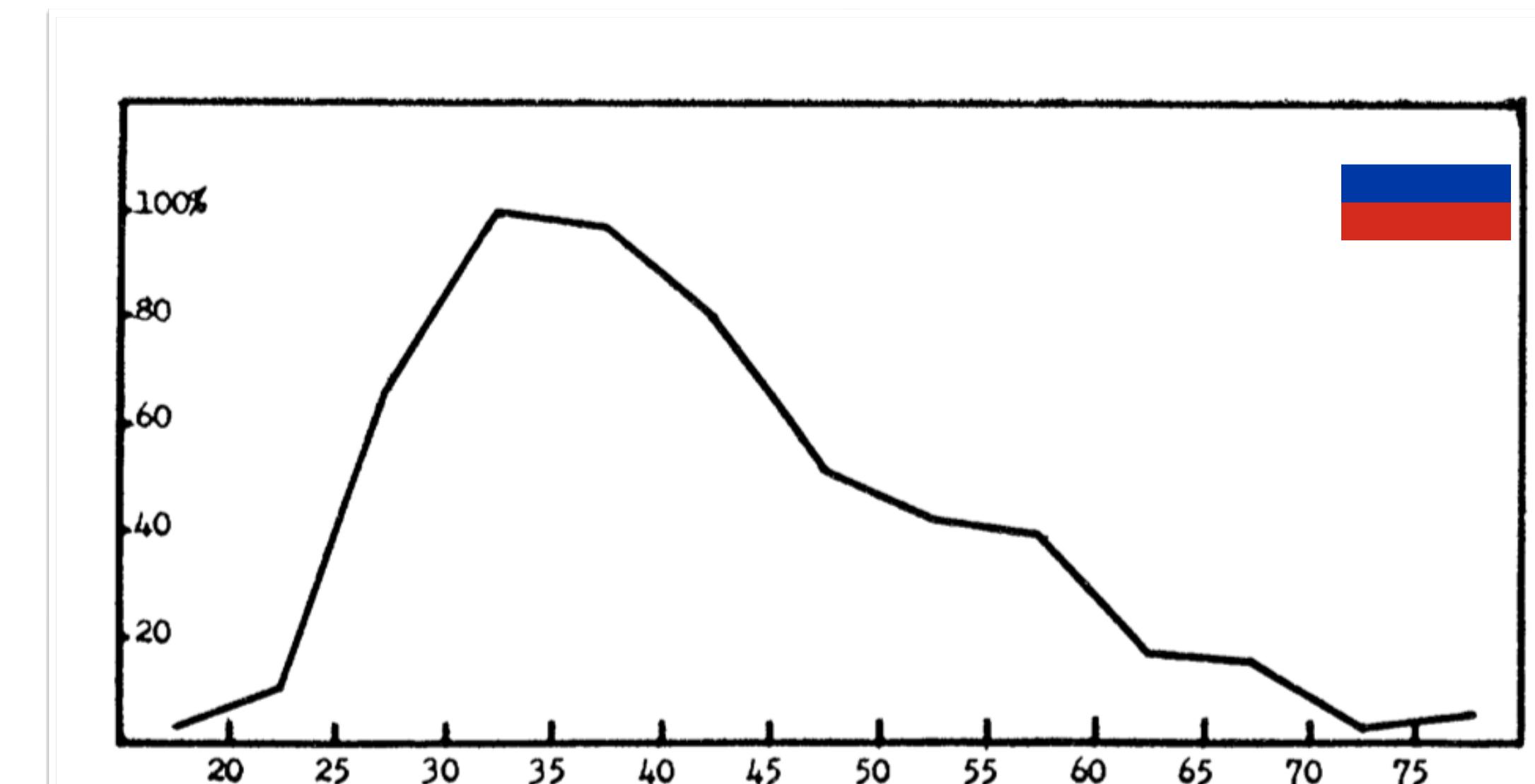


FIG. 1. Age versus creative production rate for Russians only, in science and mathematics.

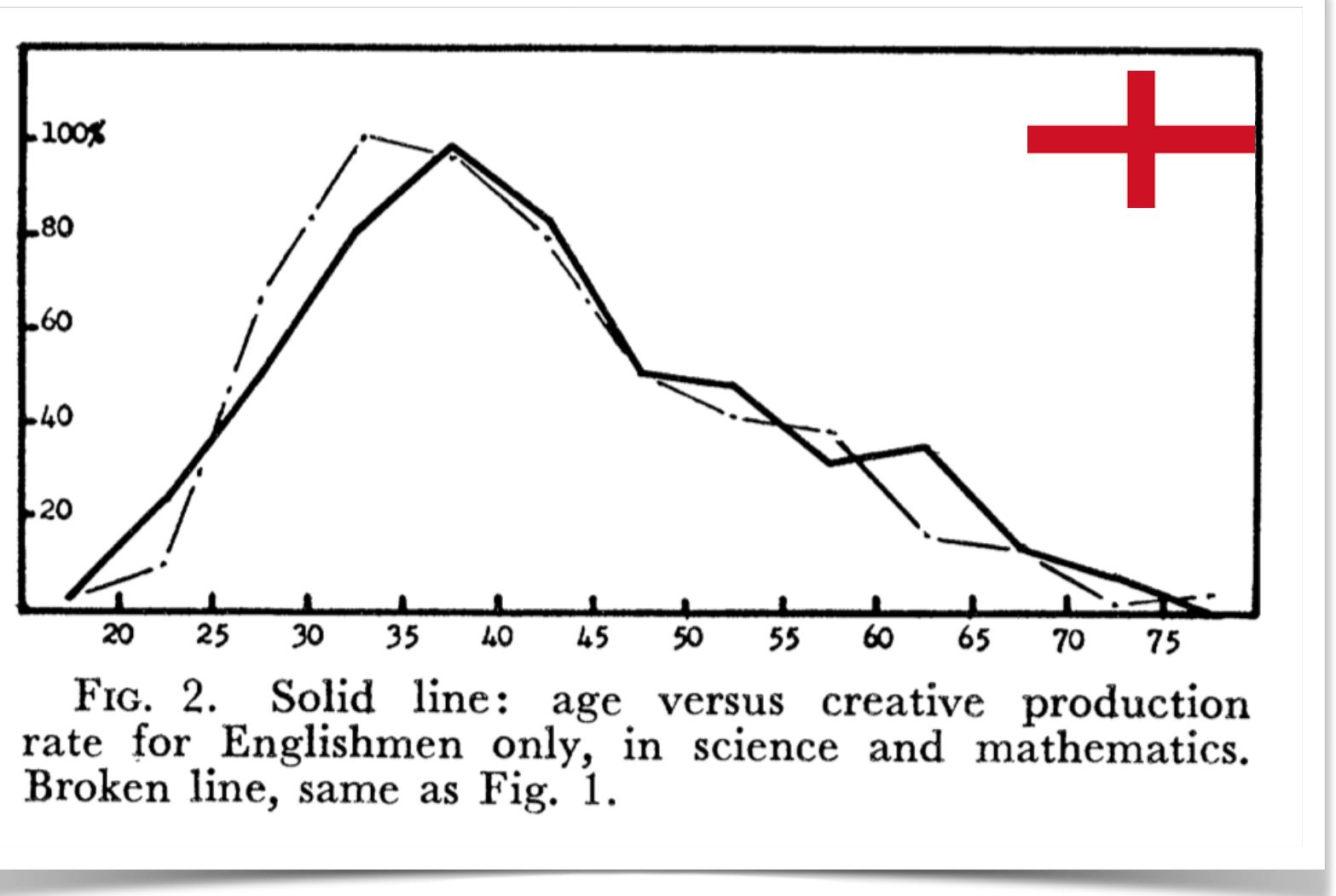


FIG. 2. Solid line: age versus creative production rate for Englishmen only, in science and mathematics. Broken line, same as Fig. 1.

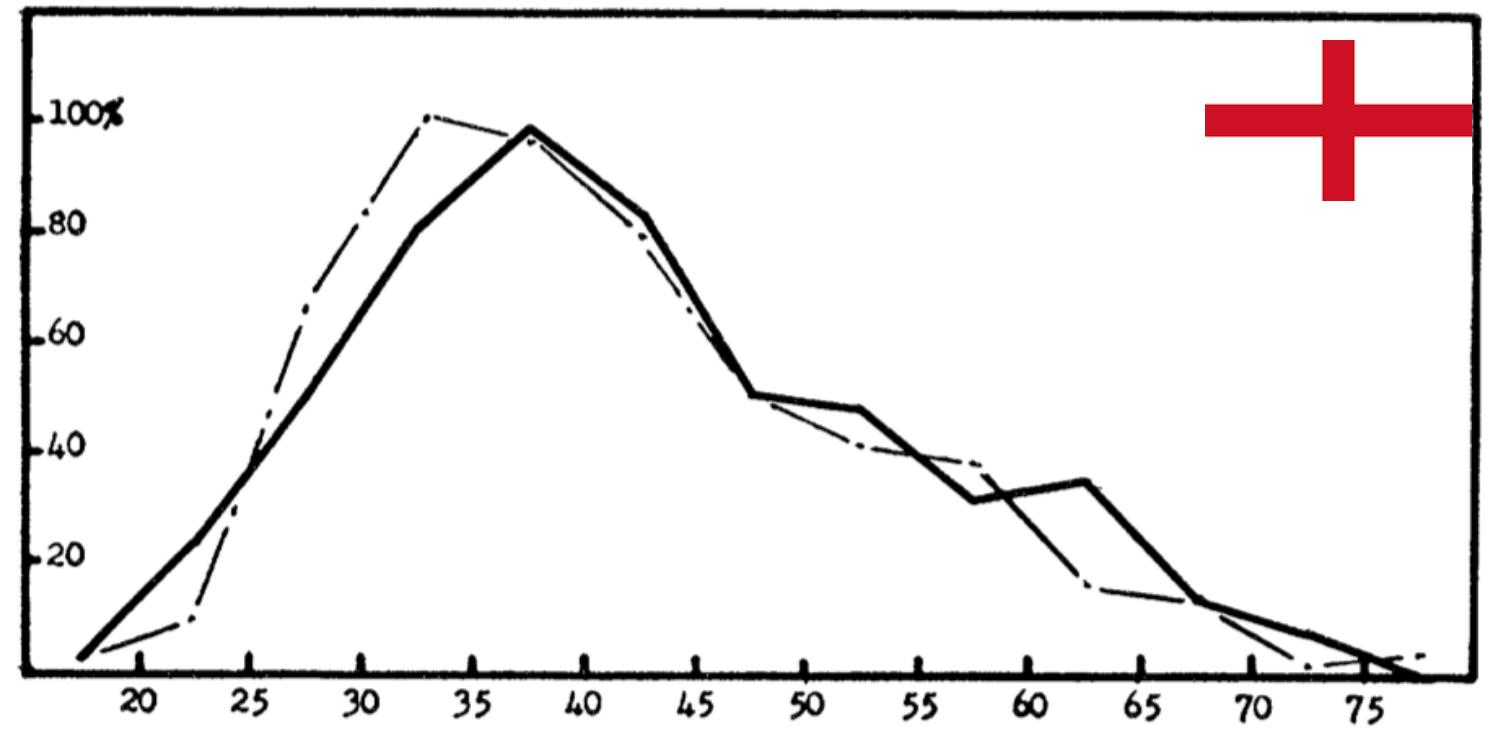


FIG. 2. Solid line: age versus creative production rate for Englishmen only, in science and mathematics. Broken line, same as Fig. 1.

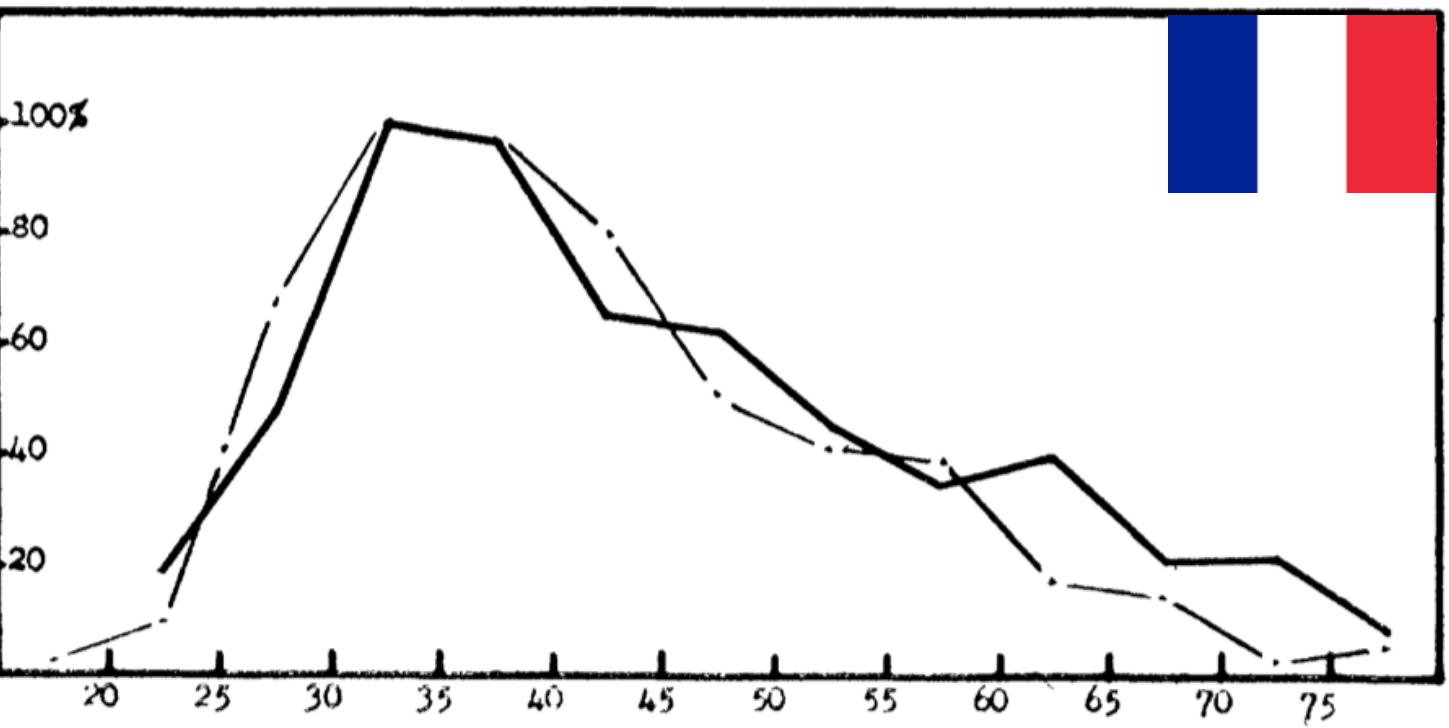


FIG. 3. Solid line: age versus creative production rate for Frenchmen only, in science and mathematics. Broken line, same as Fig. 1.

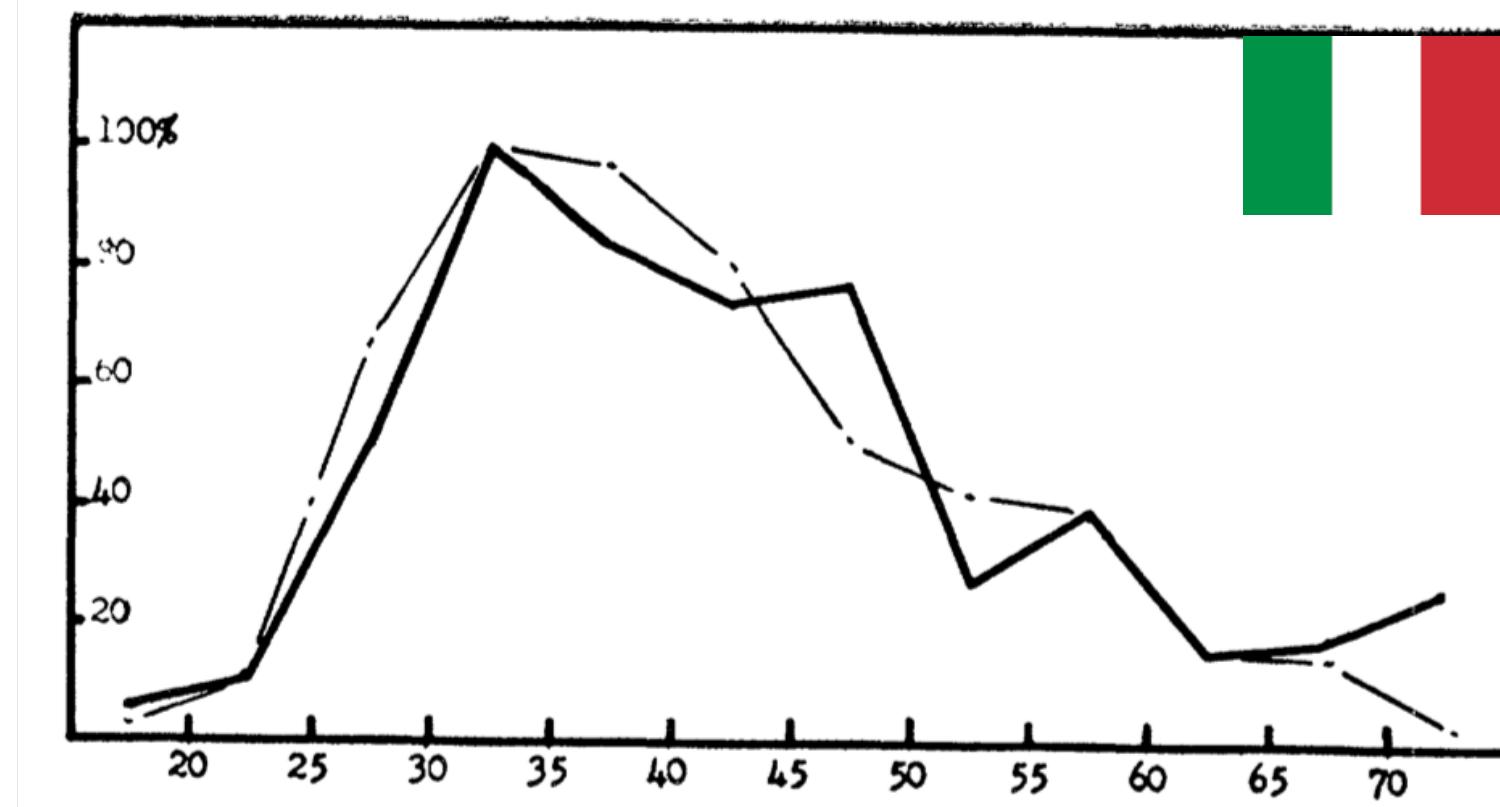


FIG. 4. Solid line: age versus creative production rate for Italians only, in science and mathematics. Broken line, same as Fig. 1.

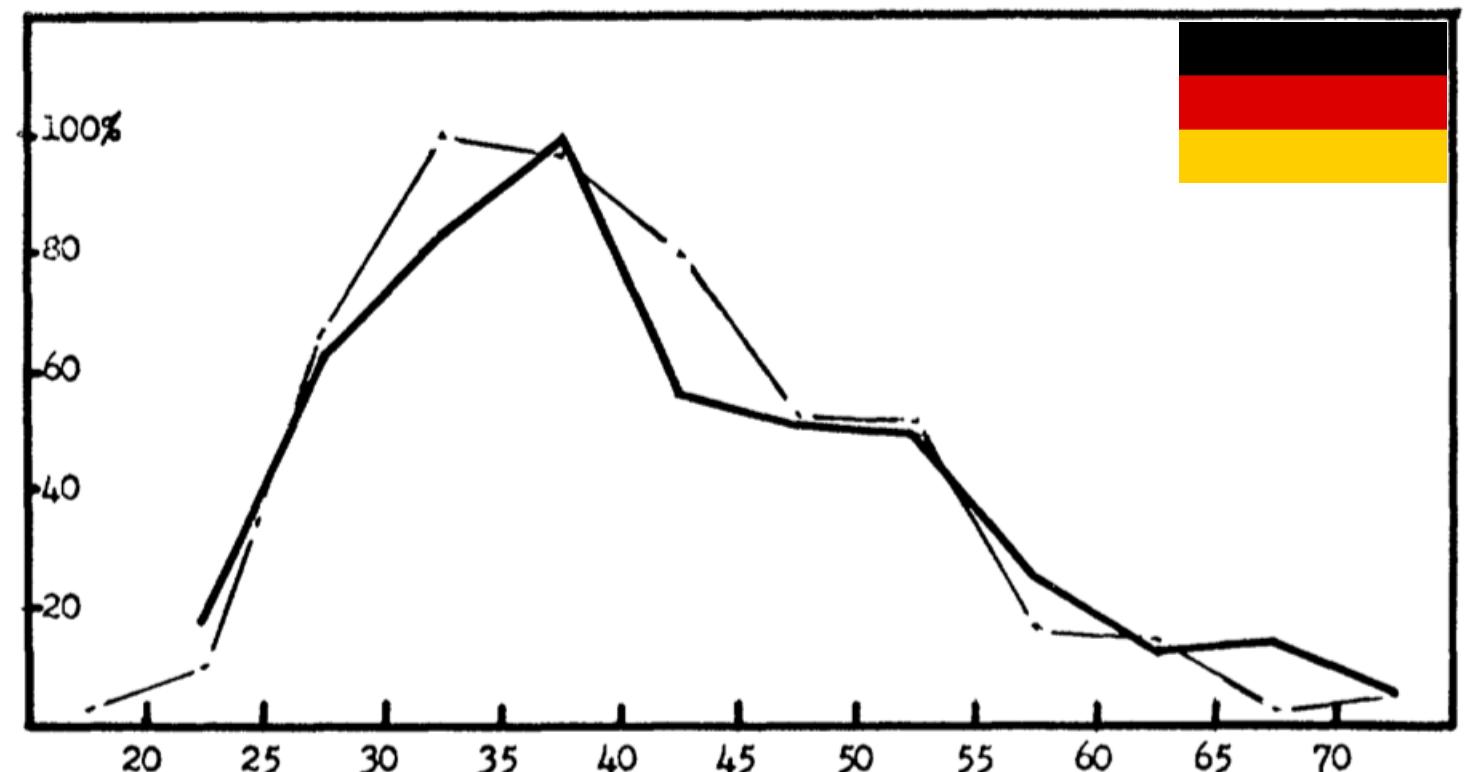


FIG. 5. Solid line: age versus creative production rate for Germans only, in science and mathematics. Broken line, same as Fig. 1.

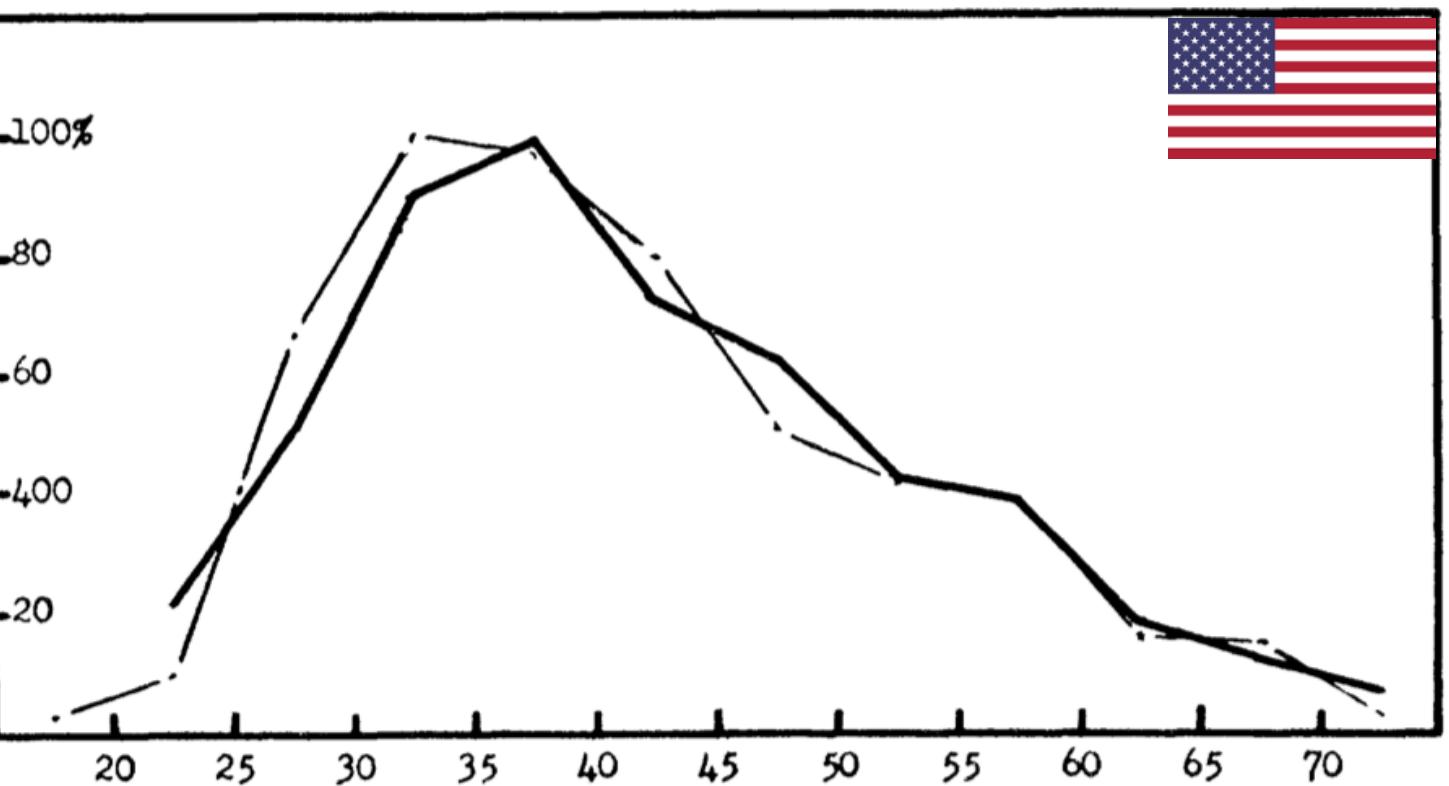


FIG. 6. Solid line: age versus creative production rate for individuals from the U.S.A. only, in science and mathematics. Broken line, same as Fig. 1.

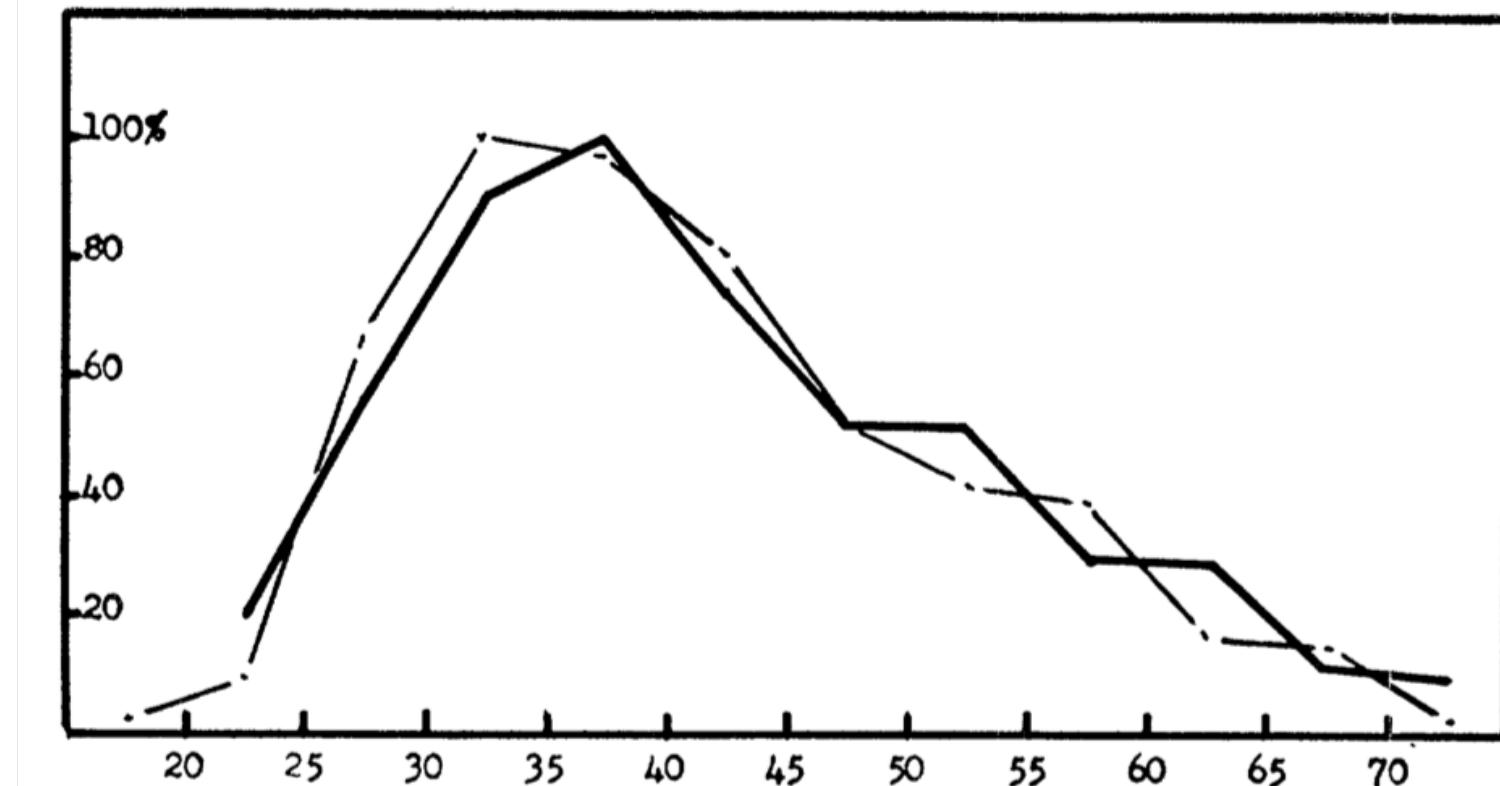


FIG. 7. Solid line: age versus creative production rate in science and mathematics for the nationals of 14 different countries other than Russia, England, France, Italy, Germany, and the U.S.A. Broken line, same as Fig. 1.

# productivity over a career

the canonical narrative (50+ years of evidence):

- ▶ **rapid rise to an early peak**
- ▶ **decline or flattening**

publication rates in psychology, 1986 ✓

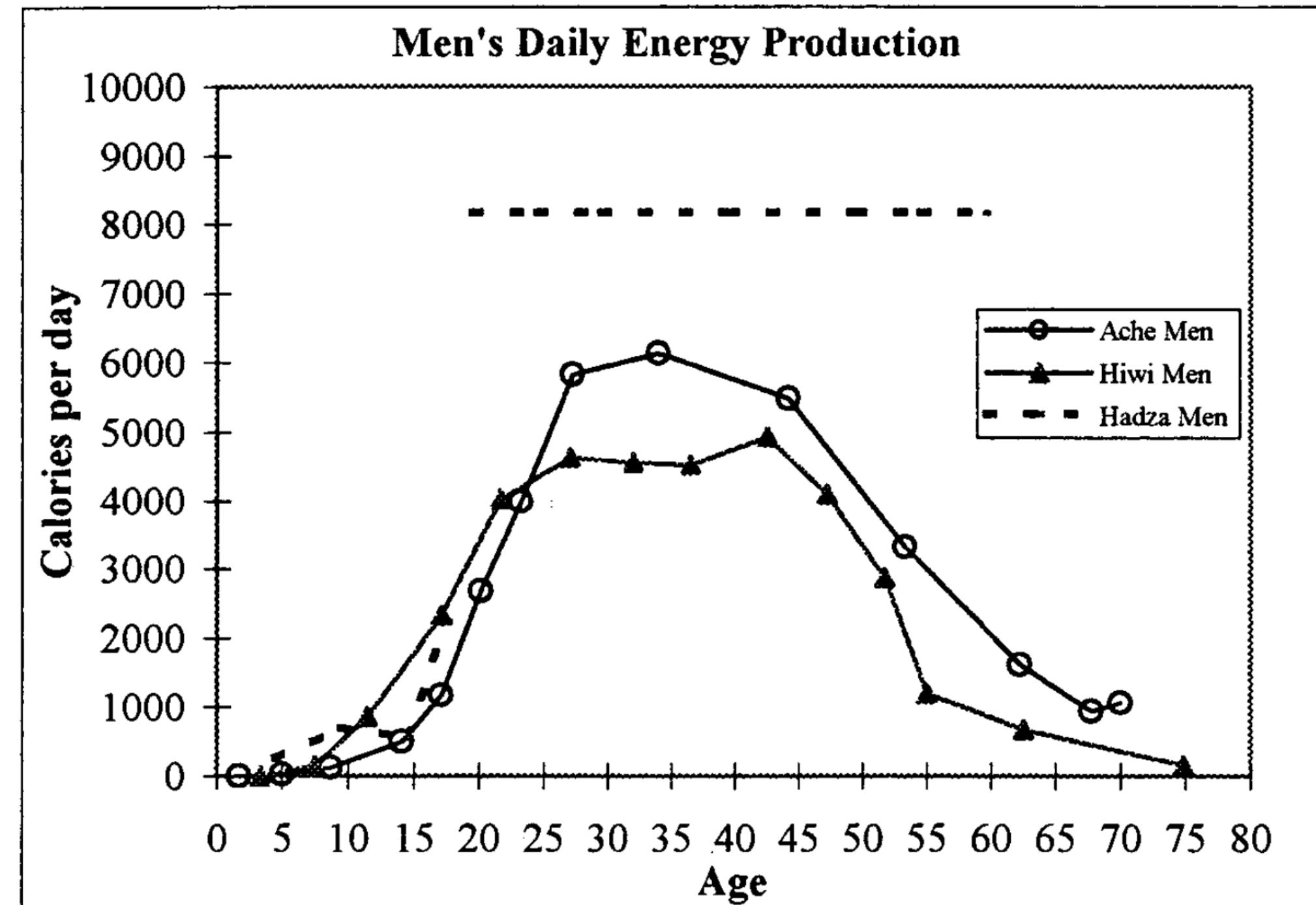
... in Russian science & math, 1954 ✓

... hunter-gather groups ✓

... French & Philly criminals, 1835 ✓

... French artists, 1835 ✓

... many others, 1950s - present ✓



# productivity over a career

the canonical narrative (50+ years of evidence):

- ▶ **rapid rise to an early peak**
- ▶ **decline or flattening**

publication rates in psychology, 1986 ✓

... in Russian science & math, 1954 ✓

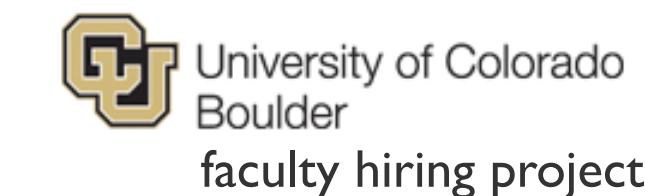
... hunter-gather groups ✓

... French & Philly criminals, 1835 ✓

... French artists, 1835 ✓

... many others, 1950s - present ✓

... **computer scientists** 🧑‍💻 🧑‍💻 ✓

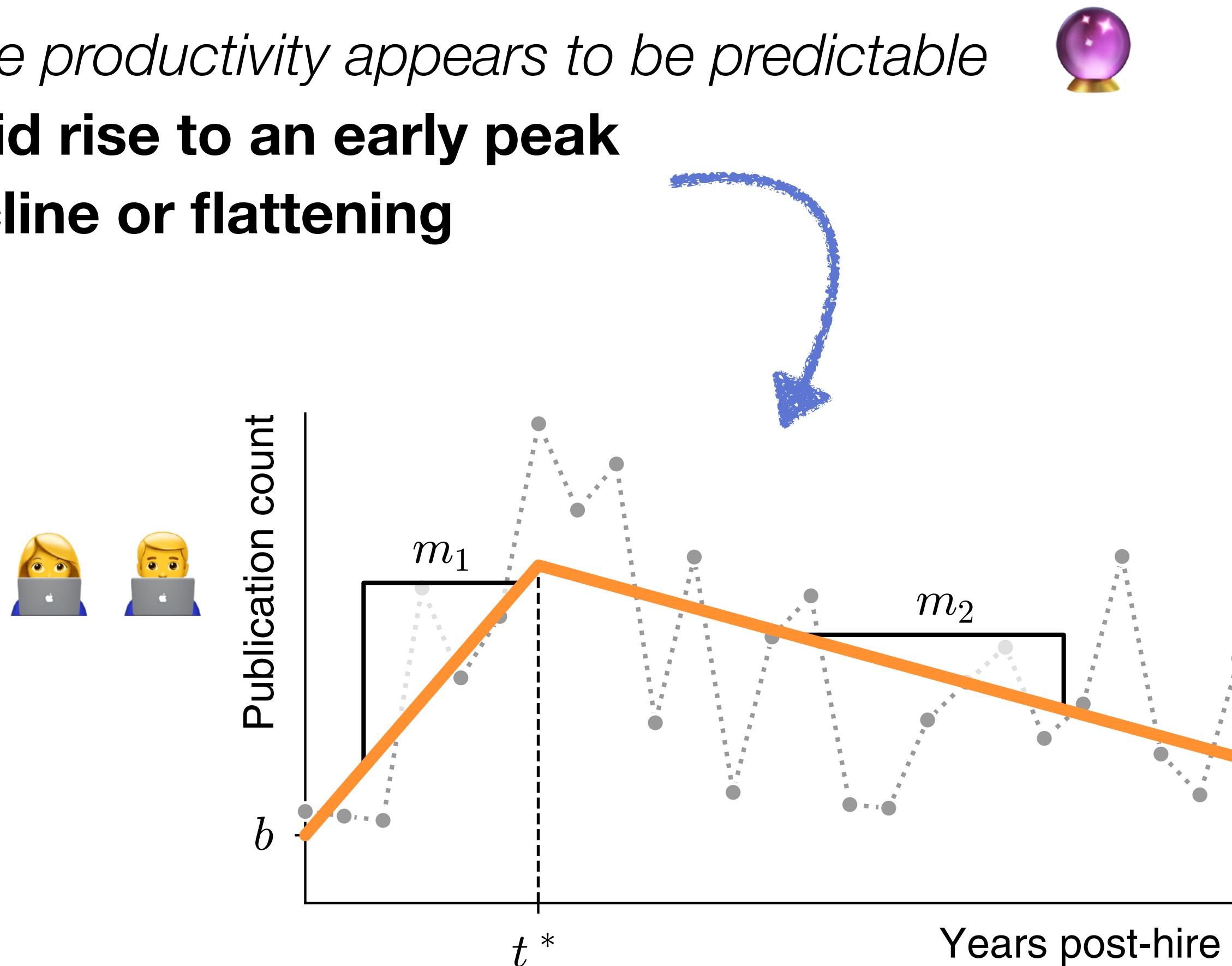


$n = 2453$  early career computer science faculty

# productivity over a career

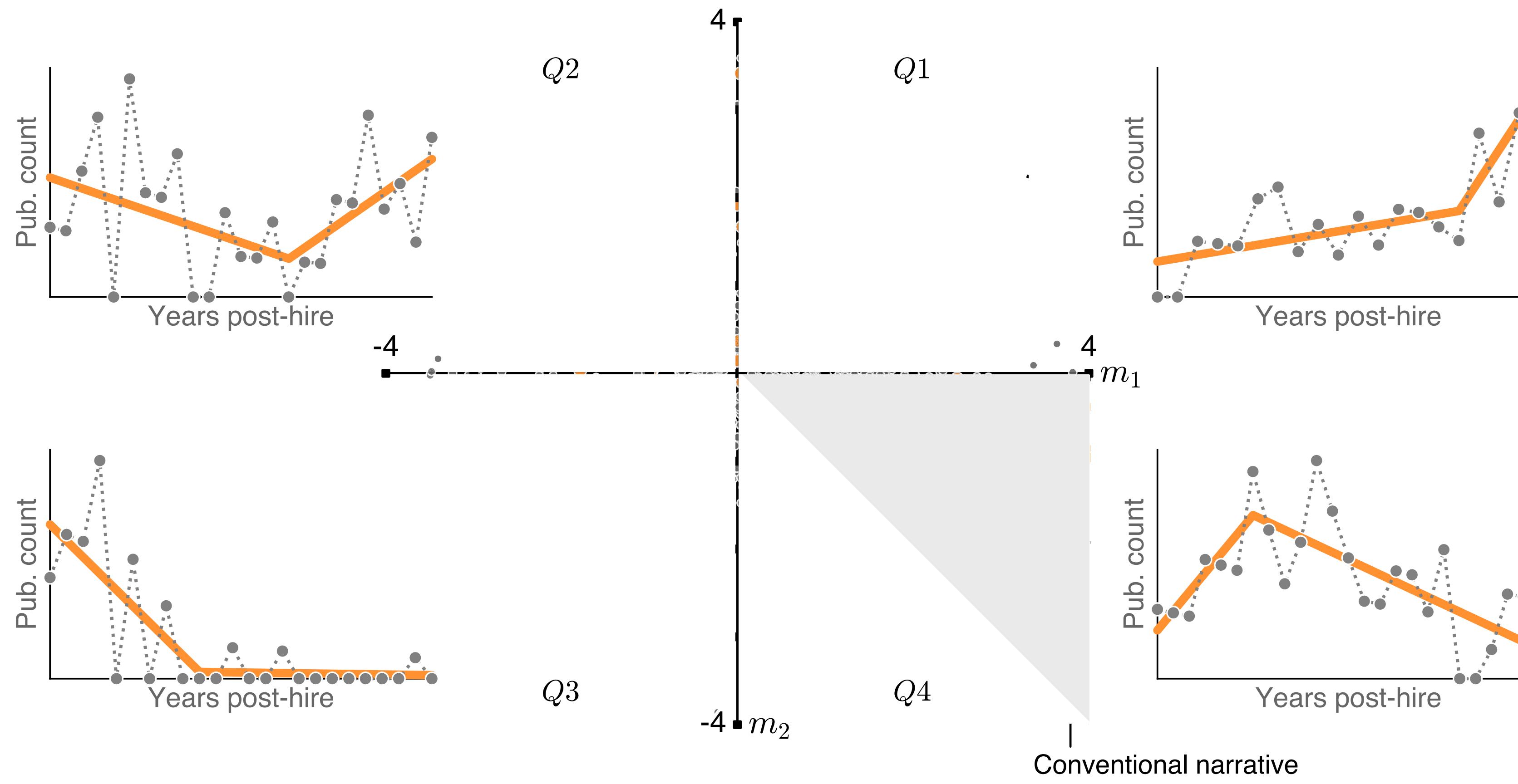
average productivity appears to be predictable

- ▶ rapid rise to an early peak
- ▶ decline or flattening



# productivity over a career

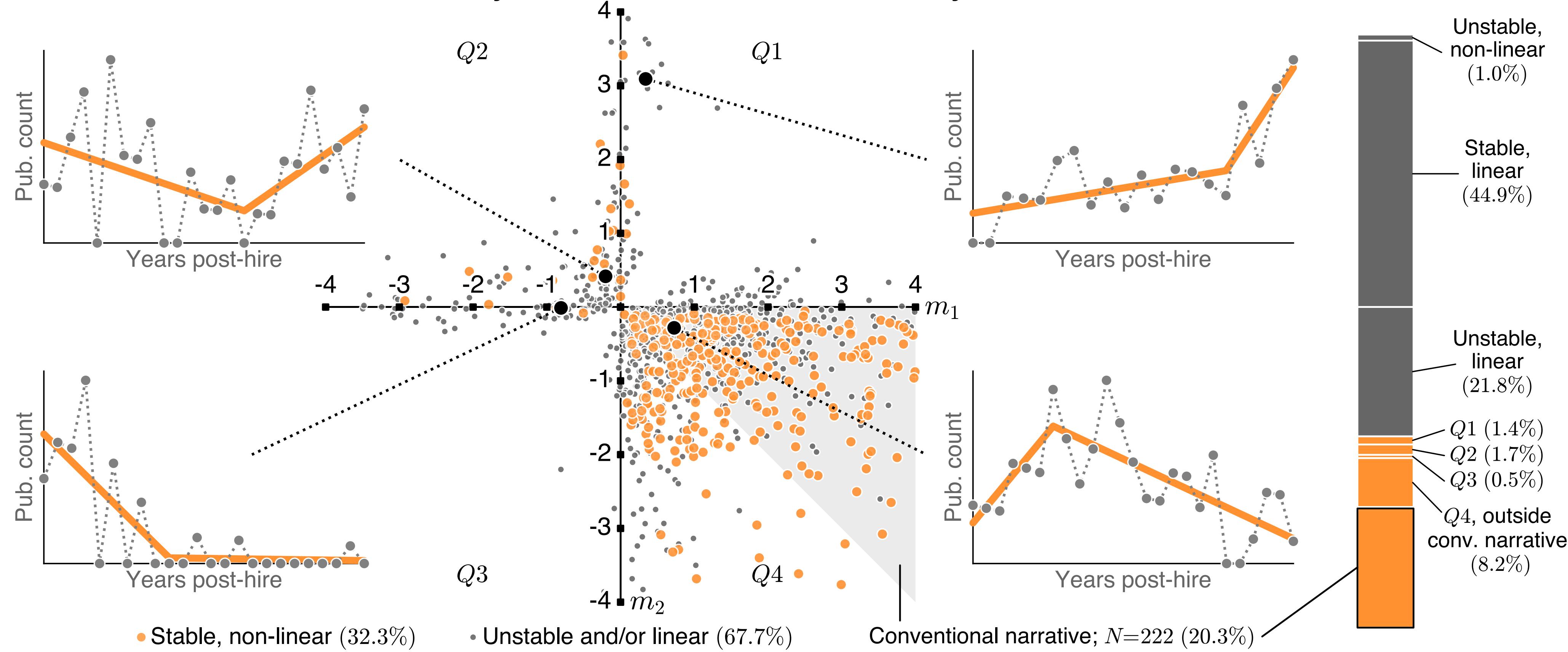
average productivity appears to be predictable



# productivity over a career

average productivity appears to be predictable — **except it's not**

▶ the conventional narrative only holds for 20.3% of faculty



# timing of big discoveries

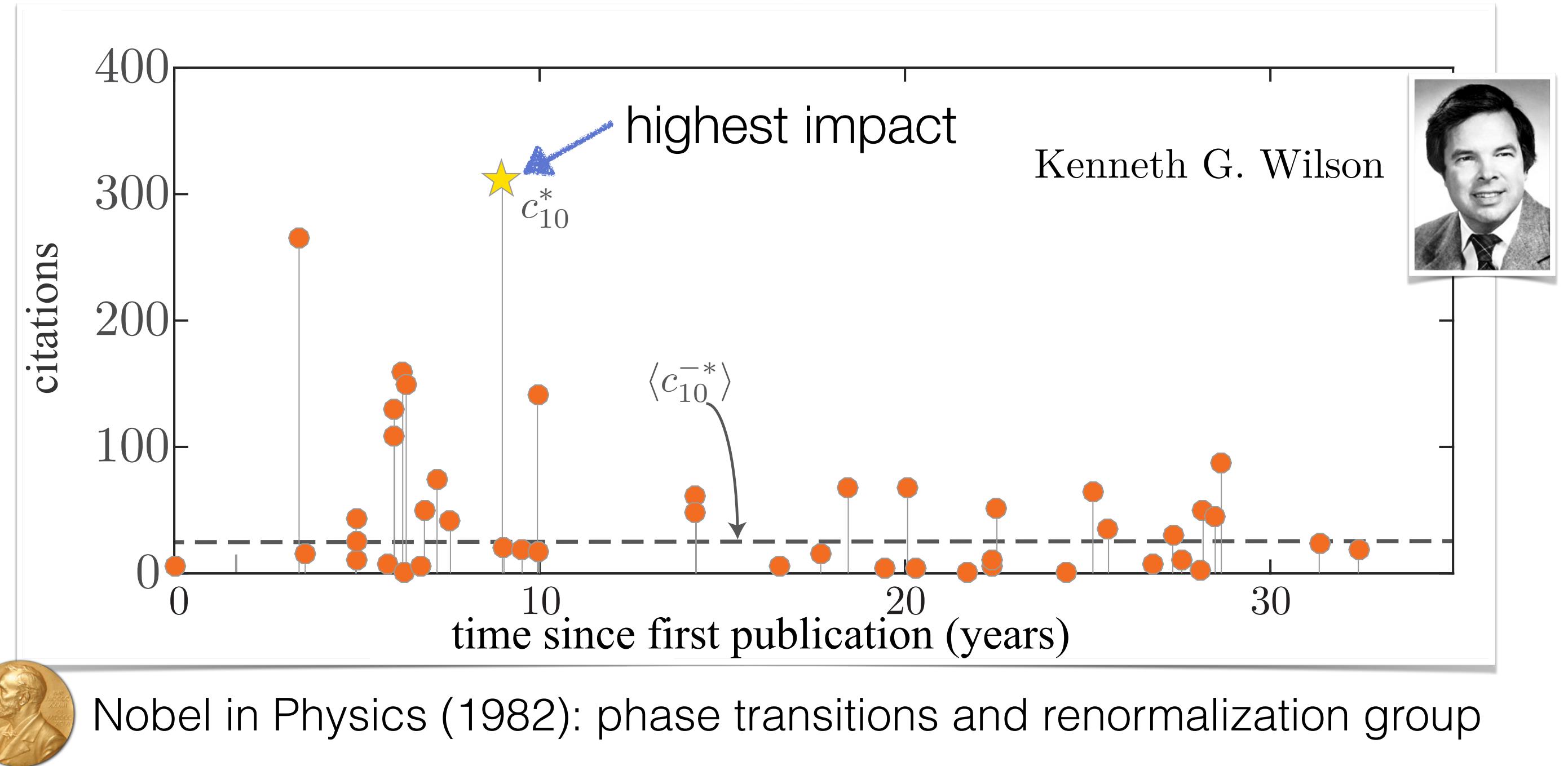
---

conventional narrative: *scientific creativity peaks early*

# timing of big discoveries

conventional narrative: *scientific creativity peaks early* ✓

APS Data Sets for Research



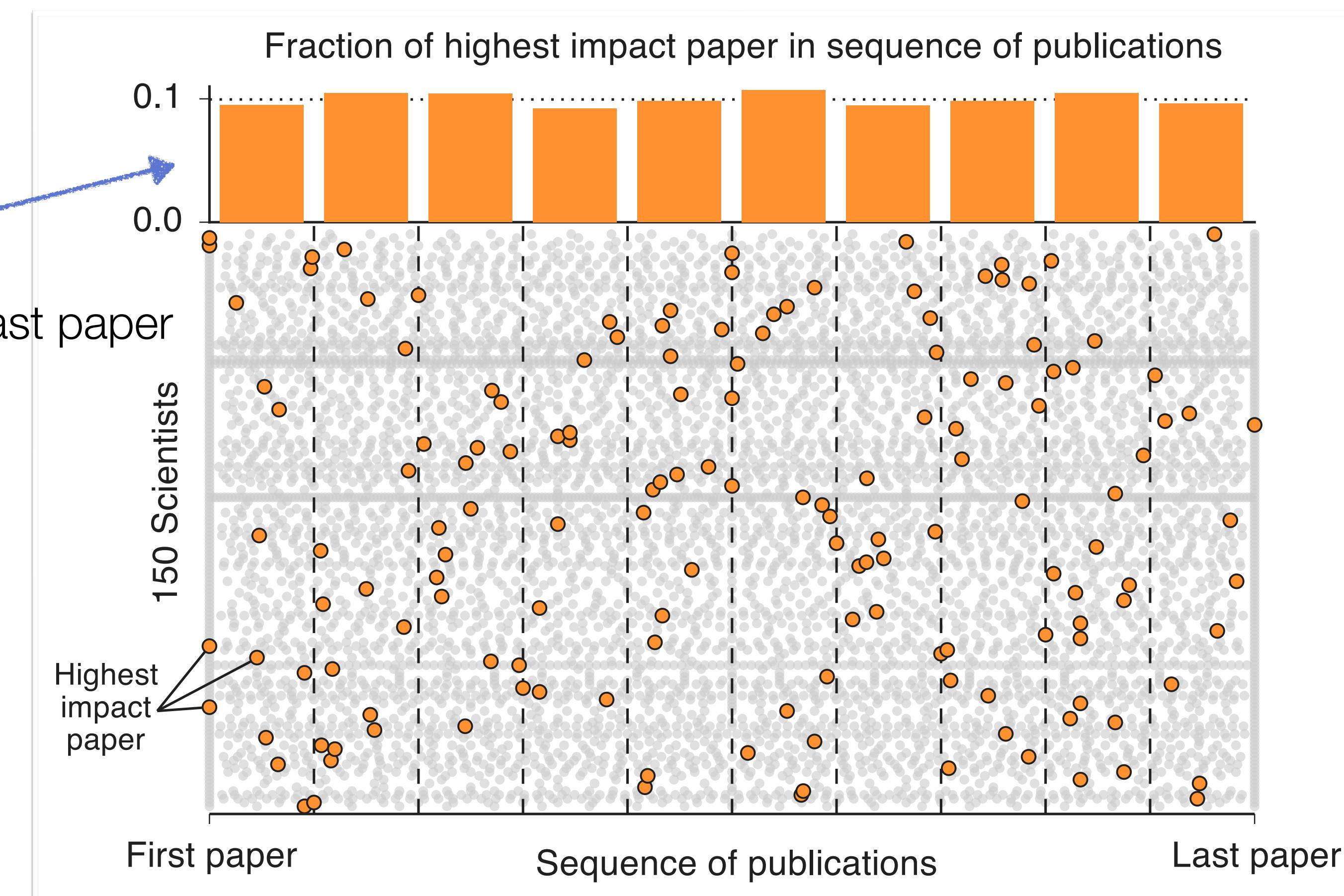
# timing of big discoveries

conventional narrative: ~~scientific creativity peaks early~~ — **except it doesn't**

▶ all publications, ordered first to last



highest impact:  
equally likely to be first or last paper



# predicting discoveries

some aspects of science are ***highly predictable***



- ▶ most citation counts, institution of origin, maximum impact, etc.
- ▶ aggregate trends like CPU speed, solar cell efficiency, battery cost, etc.
- ▶ interdisciplinary research is harder to publish & fund
- ▶ under-represented groups (women, non-whites) receive less funding, attention, etc.



# predicting discoveries

some aspects of science are ***highly predictable***



- ▶ most citation counts, institution of origin, maximum impact, etc.
- ▶ aggregate trends like CPU speed, solar cell efficiency, battery cost, etc.
- ▶ interdisciplinary research is harder to publish & fund
- ▶ under-represented groups (women, non-whites) receive less funding, attention, etc.

other aspects appear ***fundamentally unpredictable***



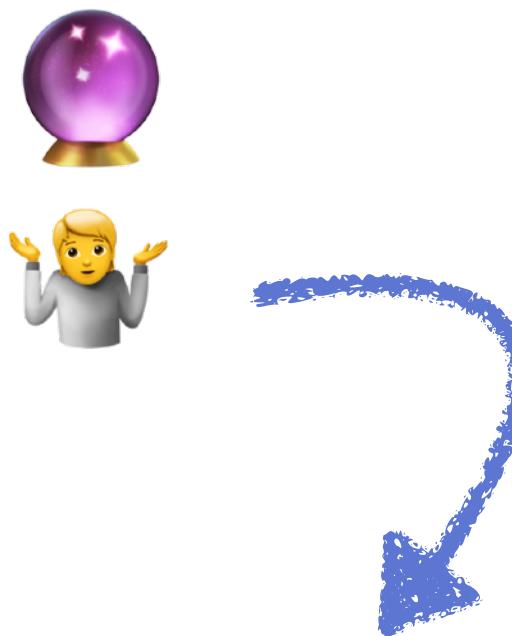
- ▶ productivity over a career, timing of biggest discovery, etc.
- ▶ long-term impact of proposed project or manuscript
- ▶ what discoveries are not being made because of our focus on predictability?
- ▶ predicting discovery is just. plain. hard. (even for humans)



# predicting discoveries

some aspects of science are ***highly predictable***

other aspects appear ***fundamentally unpredictable***



▶ the data are ***crude + biased + noisy + incomplete***



they don't directly measure knowledge or progress

▶ poor understanding of ***mechanisms*** that drive scientific discovery



social and scientific, individual and structural

why are some things predictable, and others not?

▶ predicting ***new discoveries*** is a form of ***extrapolation = hard***



even expert humans struggle! should we expect dumb machines to do better?

# looking forward

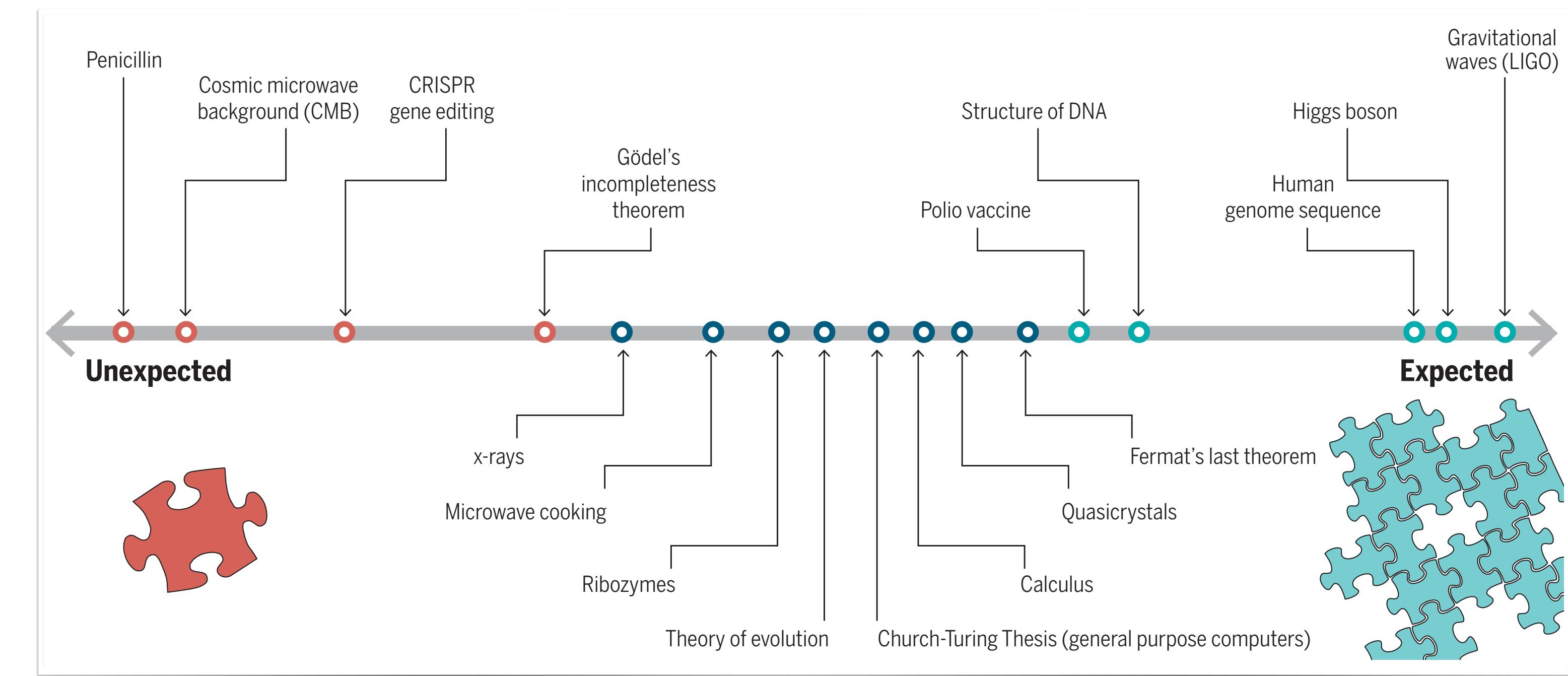
- ▶ science is a large and diverse ecosystem 
- ▶ this diversity is a key part of its continued success
- ▶ machine learning could expand or contract it 
- ▶ can we adapt diversity ideas from ecology and evolutionary theory?  
design principles of robustness, diversifying selection, stabilizing feedback, etc. 
- ▶ if discovery is inherently *unpredictable*, better to cultivate a diverse scientific ecosystem than try to automate its prediction 



"novel discoveries are valuable precisely because they have never been seen before, while data-driven prediction techniques can only learn about what's been done in the past"



# a role for machine intelligence

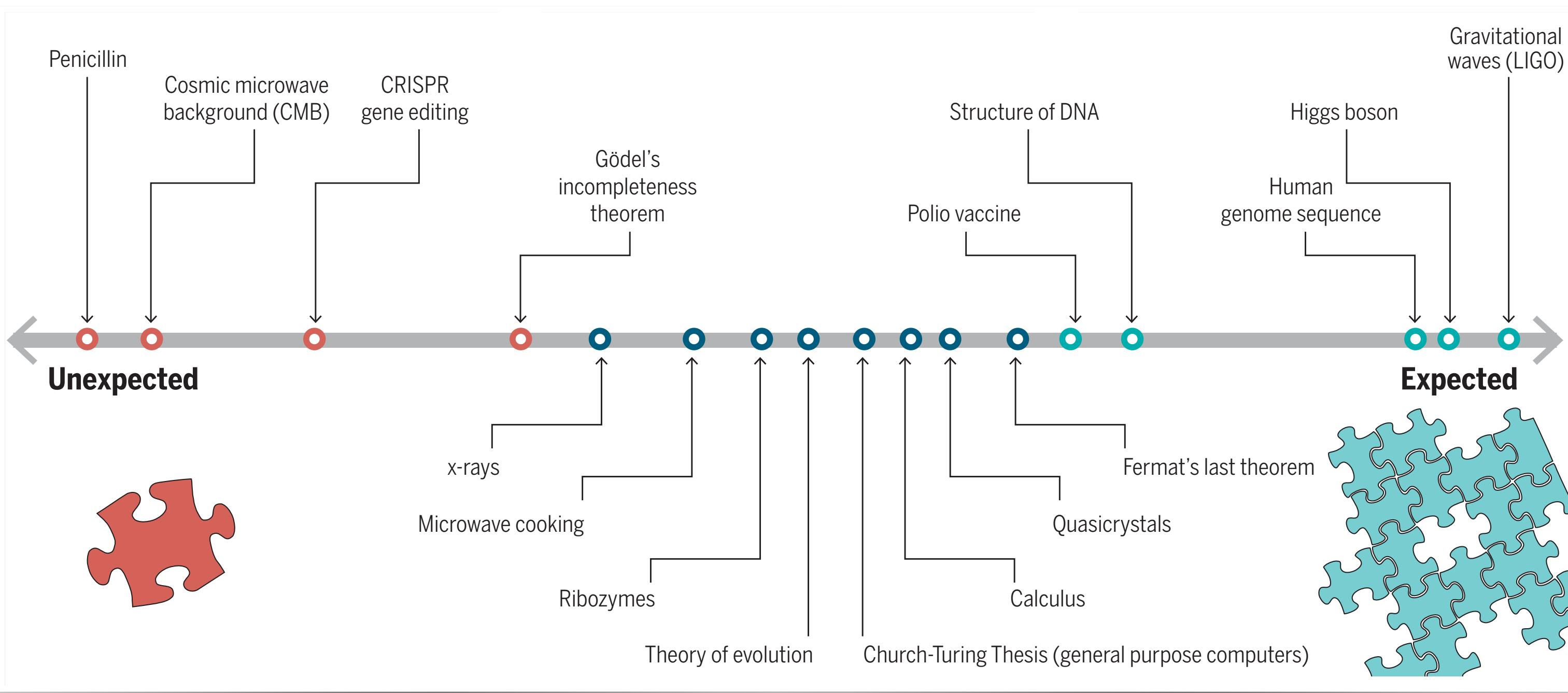


# a role for machine intelligence



- ▶ knowledge production ("science") is a complex *social* system

- ▶ **probably not automatable** 🤔  
machines: *interpolation*  
science : *extrapolation*



# a role for machine intelligence



▶ knowledge production ("science") is a complex social system

▶ **probably not automatable** 🤔  
machines: *interpolation*  
science : *extrapolation*

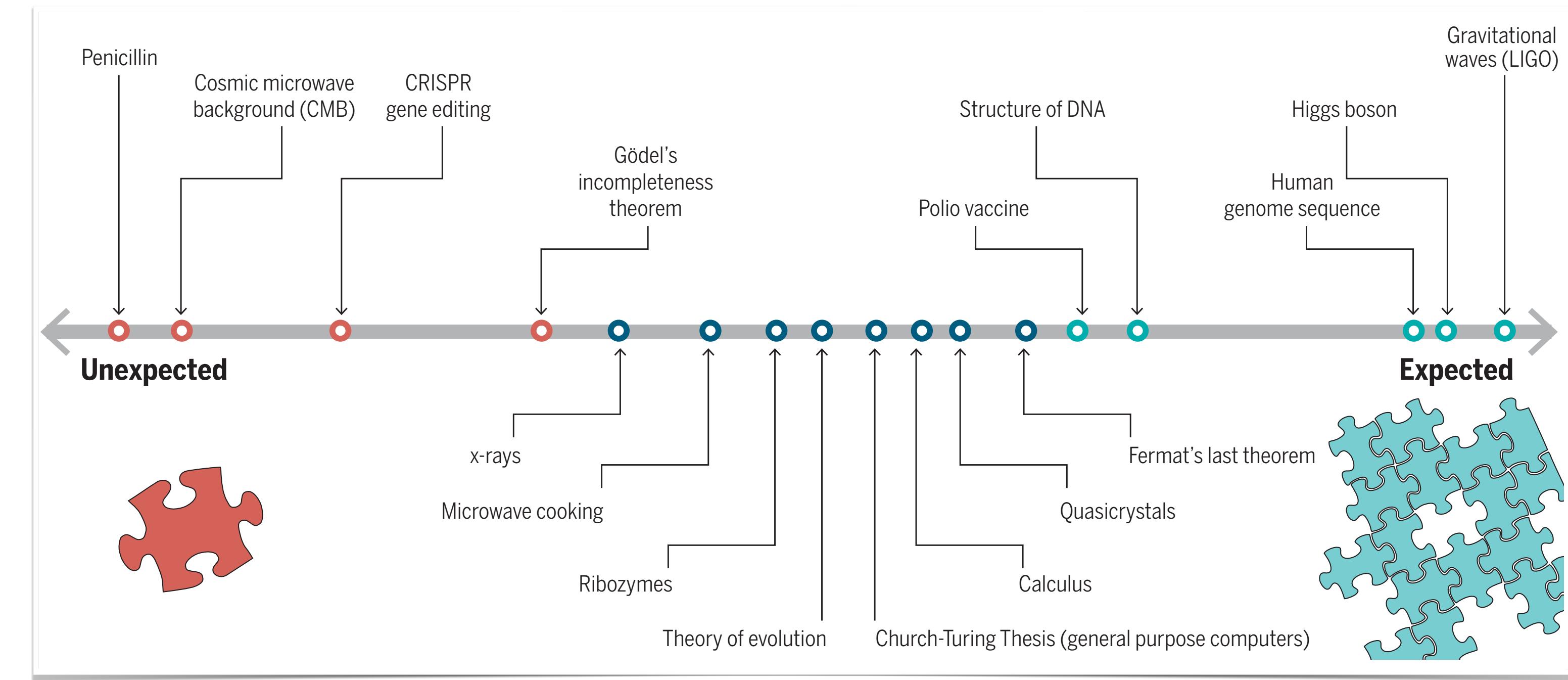
▶ current AI requires huge amount of human pampering (training data, tuning, maintenance, improvement)

🤖 current AI is "dumb" = no model of mind, no physical intuition, no understanding, no *thinking*

▶ machines don't know *what* questions to ask = most useful for "expected" discoveries

**but that's okay.**

*science among the machines will be a grand story of collaboration*



# a role for machine intelligence



## ▶ science is probably not automatable

machines: *interpolation*

science : *extrapolation*

## ▶ solution = collaboration

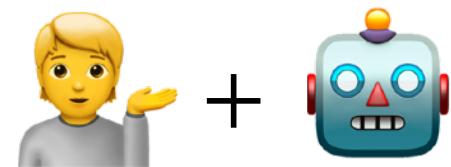


humans : design, build, decide, interpret, extrapolate

machines : collect, scale, calculate, estimate, interpolate

## ▶ hybrid approaches

will extend human control of natural and artificial processes in seemingly magical ways and  
*it will change humans in the process*



## ▶ a secret:

this is history! not the future.

every revolutionary technology has been a super power that changes humans:



language , writing , mathematics , government , science , computers ...

# looking forward (again)

- ▶ science is a large and diverse ecosystem    
  - ▶ this diversity is a key part of its continued success
  - ▶ machine learning could expand or contract it 
  - ▶ can we adapt diversity ideas 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 



"novel discoveries are valuable precisely because they have never been seen before, while data-driven prediction techniques can only learn about what's been done in the past"



SPECIAL SECTION

PREDICTION

ESSAY

## Data-driven predictions in the science of science

Aaron Clauset,<sup>1,2\*</sup> Daniel B. Larremore,<sup>2</sup> Roberta Sinatra<sup>3,4</sup>

Science 355, 477-480 (2017)



Prof. Aaron Clauset  
(Colorado)



Prof. Daniel B. Larremore  
(Colorado)



Prof. Roberta Sinatra  
(ITU Copenhagen)



## The misleading narrative of the canonical faculty productivity trajectory

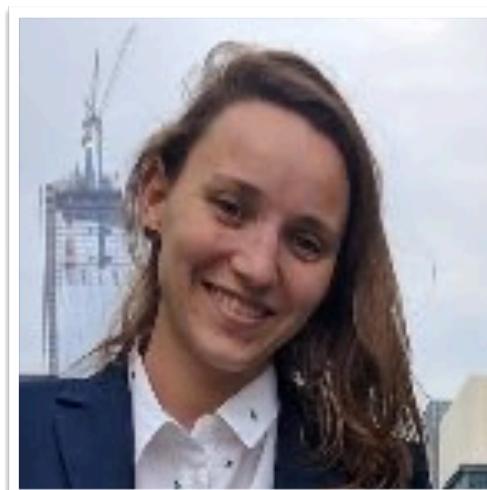
Samuel F. Way<sup>a,1</sup>, Allison C. Morgan<sup>a</sup>, Aaron Clauset<sup>a,b,c,2</sup>, and Daniel B. Larremore<sup>a,b,c,1,2</sup>

<sup>a</sup>Department of Computer Science, University of Colorado, Boulder, CO 80309; <sup>b</sup>BioFrontiers Institute, University of Colorado, Boulder, CO 80303; and <sup>c</sup>Santa Fe Institute, Santa Fe, NM 87501

PNAS 114 (44) E9216 (2017)



Dr. Samuel F. Way  
(Colorado)



Dr. Allison C. Morgan  
(Colorado)

see also:

Morgan et al. "The unequal impact of parenthood in academia." *Science Advances* 7 (2021)

Way et al., "Productivity, prominence, and the effects of academic environment." *PNAS* 116 (2019)

Morgan et al., "Prestige drives epistemic inequality in the diffusion of scientific ideas." *EPJ Data Science* 7 (2018)

Way et al., "Gender, productivity, and prestige in computer science faculty hiring networks." *Proc. WWW* (2016)

Clauset et al., "Systematic inequality and hierarchy in faculty hiring networks." *Science Advances* 1 (2015)





# Questions?