Predicting bench-to-bedside translation

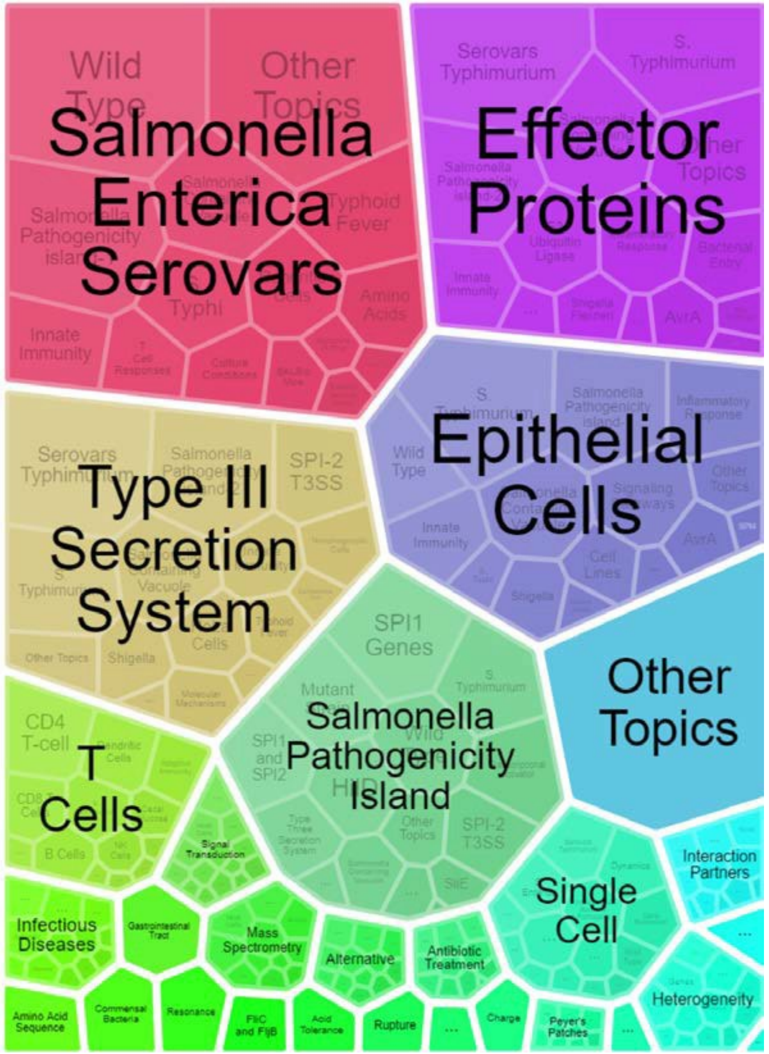
Council of Councils
May 18, 2018

George Santangelo, Ph.D.
Director, Office of Portfolio Analysis
DPCPSI/OD/NIH

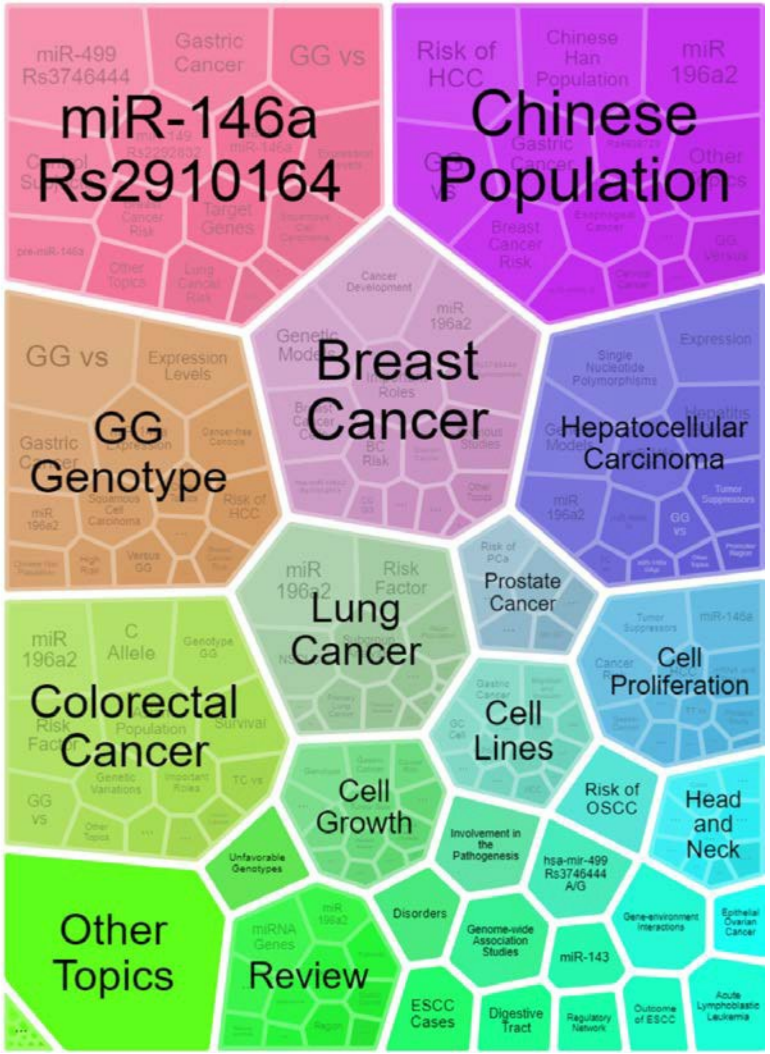
National Institutes of Health
Office of Portfolio Analysis

Which of these areas of research is most likely to translate from bench to bedside?

Salmonella pathogenesis
810 publications
2006 to 2017



Cancer biomarkers
475 publications
2006 to 2017



NIH National Institutes of Health
Office of Portfolio Analysis

Predicting which advances in scientific knowledge are most likely to translate from bench to bedside

Scientific advances can take decades to translate into improvements in human health

Can this time interval be shortened?

- Clinical trials and guidelines (CT/CGs) are attempts to improve human health (shots on goal)
- We can define translation as the citation of a biomedical research publication by a CT/CG
    - Predicated on the idea that the citation occurs because some aspect of the work is of value to the citer
- The goal of this project is to provide decision-makers with information about the likelihood that one or more publications will be cited by a CT/CG

Predicting which advances in scientific knowledge are most likely to translate from bench to bedside

Scientific advances can take decades to translate into improvements in human health

Can this time interval be shortened?

Can we identify particular data profiles associated with publications that have a high likelihood of translation?

— We built a machine learning model to explore this possibility

- Generated data profiles that incorporate the citation dynamics of all papers in PubMed
  - These data profiles include NLM-assigned Medical Subject Heading (MeSH) terms, almost all of which are located within one of three major branches in the MeSH ontology (Human, Animal, or Molecular/Cellular)
- Trained the model to recognize data profiles that have the hallmarks of translation
- Applied this algorithm to all publications in PubMed and assigned each a score (estimated likelihood of translation)

Historical pattern of bench-to-bedside translation

~25% of all papers published in 1995 received a CT/CG citation by 2014

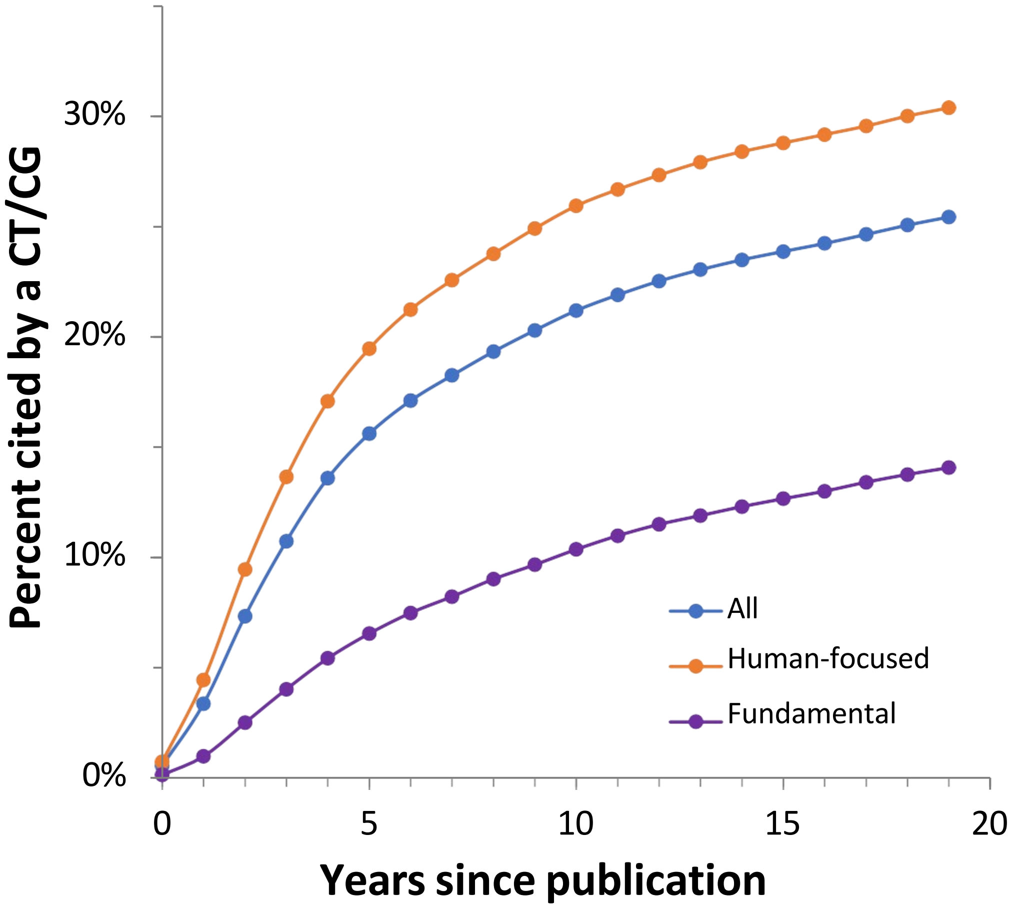
Distinct patterns for papers in different MeSH categories:

Fundamental publications (0% Human MeSH)

- CT/CG citations accumulate at a slower rate and reach a lower plateau

Human-focused publications (100% Human MeSH)

- CT/CG citations accumulate at a faster rate and reach a higher plateau (roughly twice that of fundamental publications)



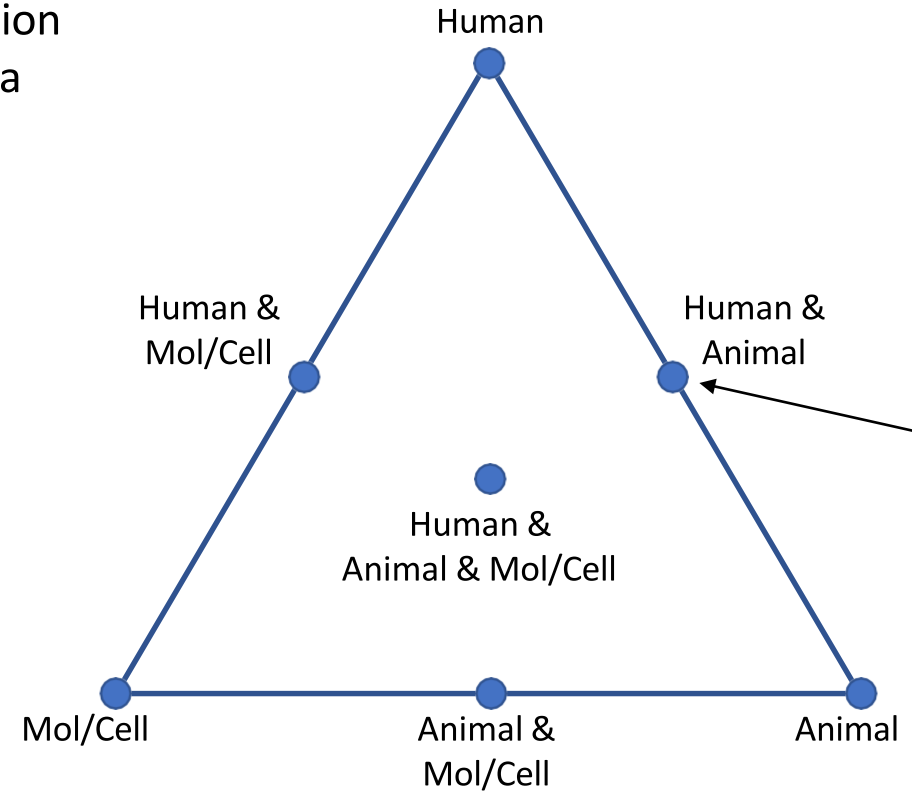
National Institutes of Health
Office of Portfolio Analysis

Griffin Weber's triangle of biomedicine*: Pro and Con

Pro:

— Uses MeSH categories to position publications at the vertices of a trilinear graph:

- Human
- Animal
- Molecular/Cellular



Con:

— Uses binary counting (only eight positions)

- Sample publication #1
  MeSH: 1 Human, 3 Animal
- Sample publication #2
  MeSH: 5 Human, 3 Animal
- Both of these publications are placed at the same Human/Animal position

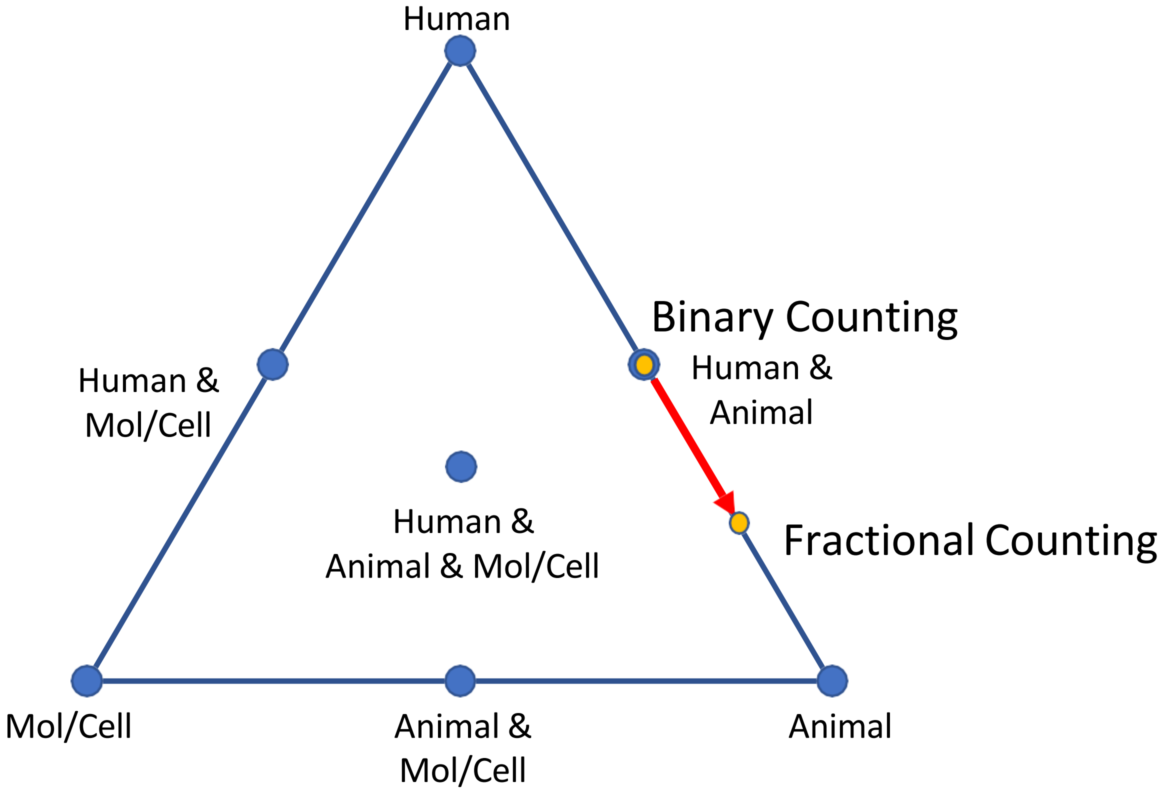
Though fractional counting is computationally intensive, it solves this problem

* Weber G.M. Identifying translational science within the triangle of biomedicine *Journal of Translational Medicine* 11(126) doi:10.1186/1479-5876-11-126 (2013)

Introducing fractional counting to Griffin Weber's triangle of biomedicine

Sample publication #1

MeSH terms { **1 Human**
3 Animal



Human

Human & Mol/Cell

Binary Counting

Human & Animal

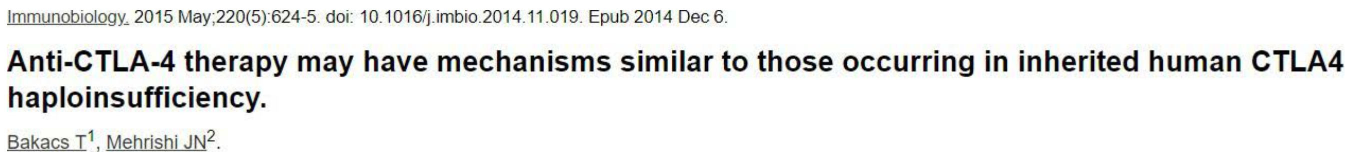
Human & Animal & Mol/Cell

Fractional Counting

Mol/Cell

Animal & Mol/Cell

Animal



Immunobiology. 2015 May;220(5):624-5. doi: 10.1016/j.imbio.2014.11.019. Epub 2014 Dec 6.

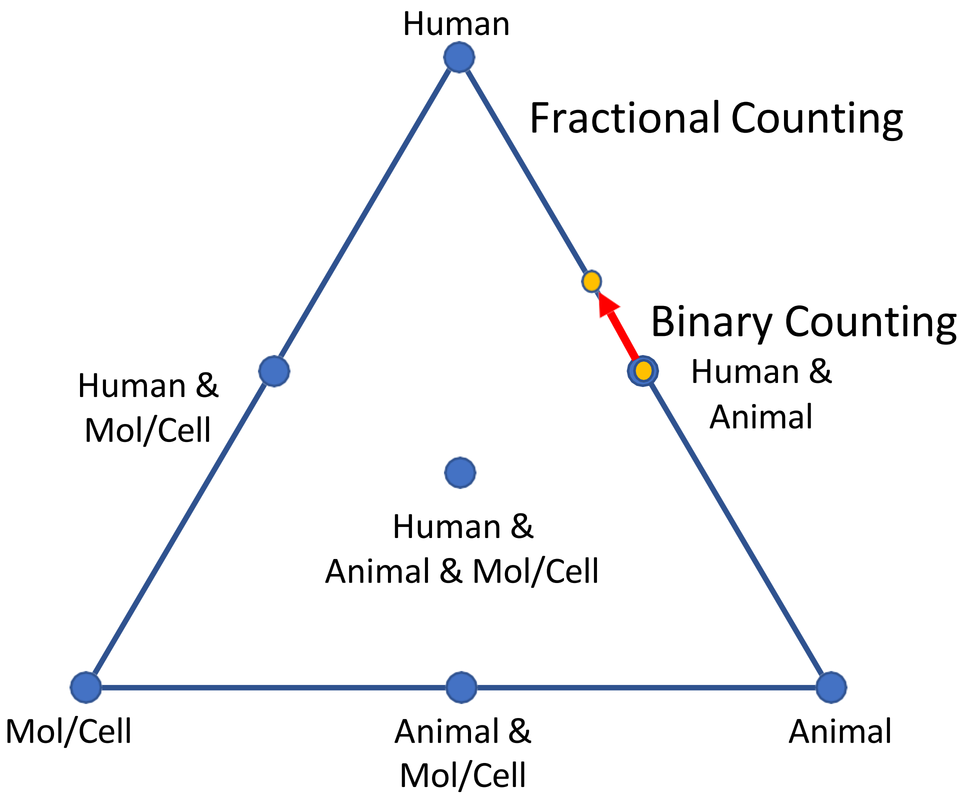
Anti-CTLA-4 therapy may have mechanisms similar to those occurring in inherited human CTLA4 haploinsufficiency.

Bakacs T[1], Mehrishi JN[2].

Introducing fractional counting to Griffin Weber’s triangle of biomedicine

Sample publication #2

MeSH terms {
5 Human
3 Animal



Human

Fractional Counting

Binary Counting

Human &
Mol/Cell

Human &
Animal

Human &
Animal & Mol/Cell

Mol/Cell

Animal &
Mol/Cell

Animal

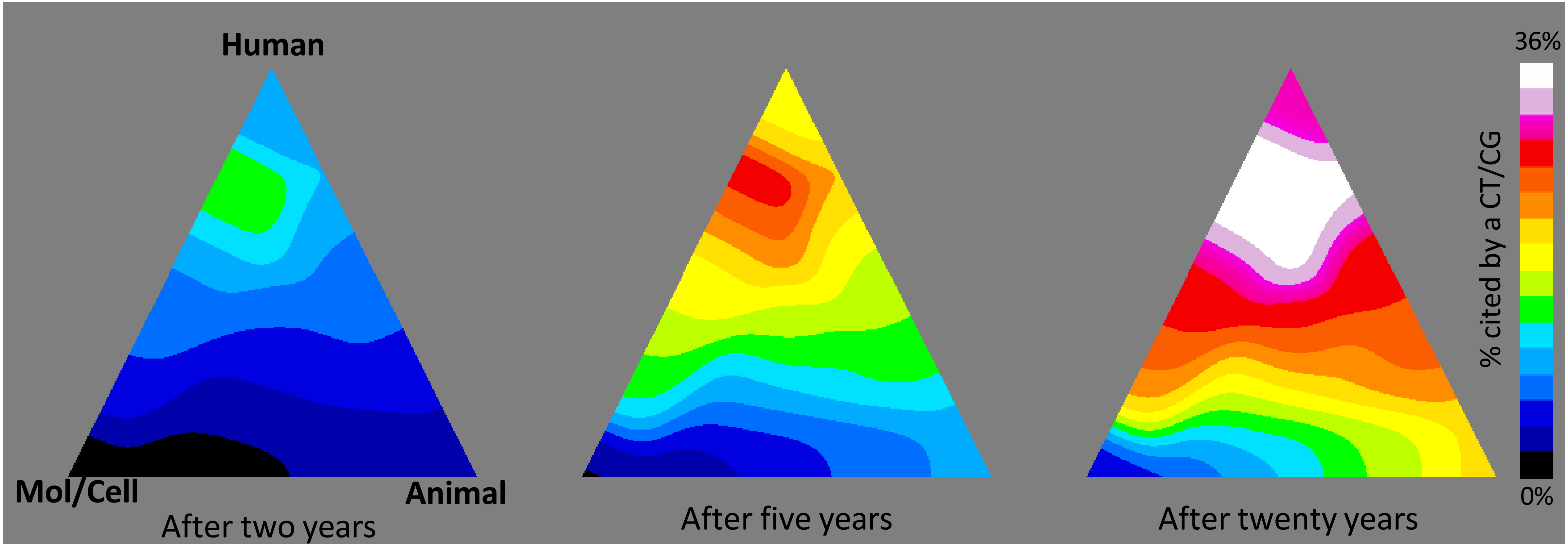
Biomed Pharmacother. 2016 Oct;83:1247-1252. doi: 10.1016/j.biopha.2016.08.050. Epub 2016 Aug 24.

Xenogeneic cell-based vaccine therapy for colorectal cancer: Safety, association of clinical effects with vaccine-induced immune responses.

Seledtsova GV[1], Shishkov AA[1], Kaschenko EA[1], Seledtsov VI[2].

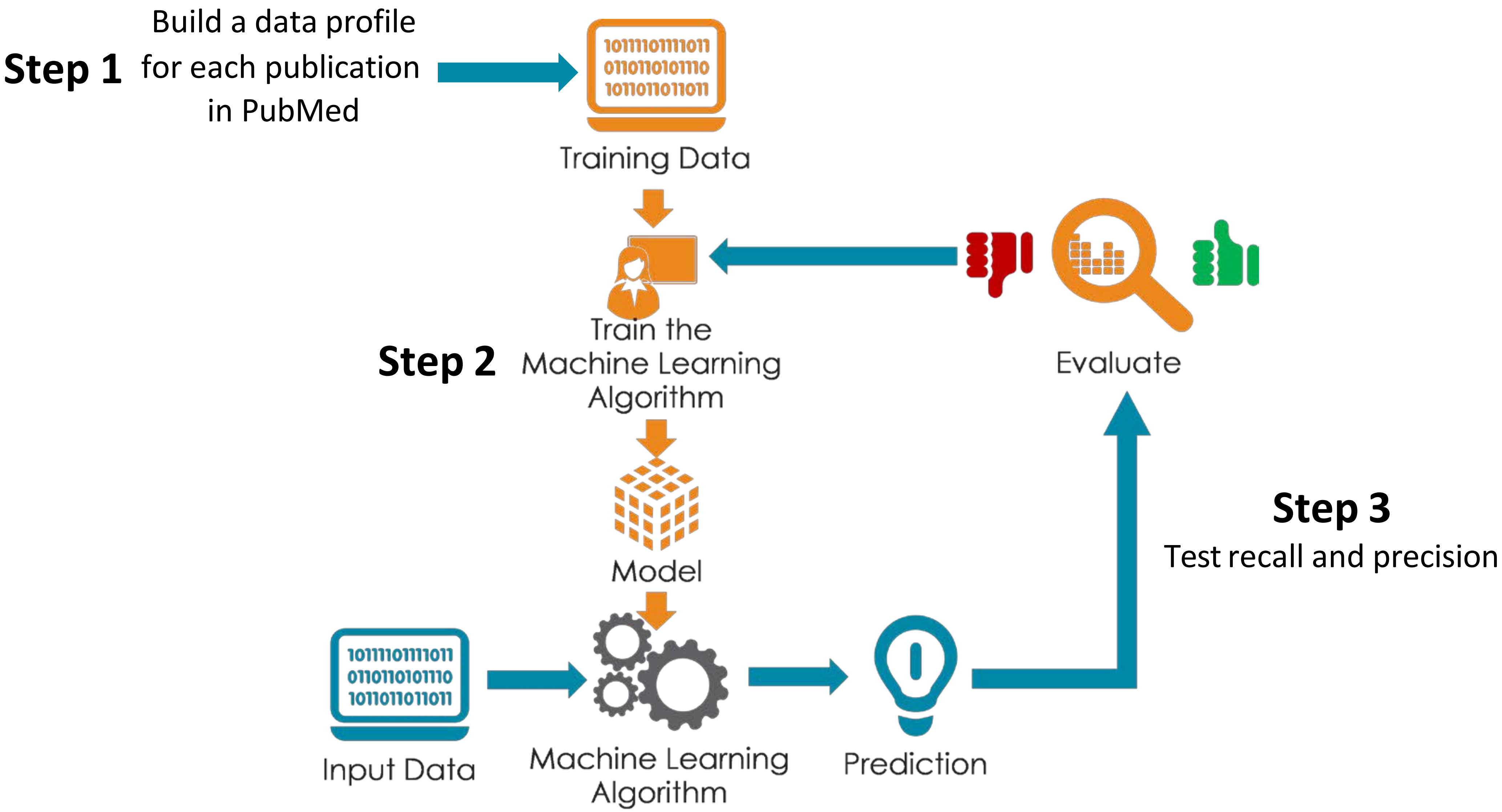
MeSH profile of publications cited by CT/CGs over time

All PubMed publications in 1995; fractional counting



Can the data profiles of publications and their citing networks be used to predict future translation?

Machine learning: quantifying translational potential at scale



Step 1 Build a data profile for each publication in PubMed

Training Data

Step 2 Train the Machine Learning Algorithm

Evaluate

Model

Step 3

Test recall and precision

Input Data

Machine Learning Algorithm

Prediction

Using machine learning to predict translation

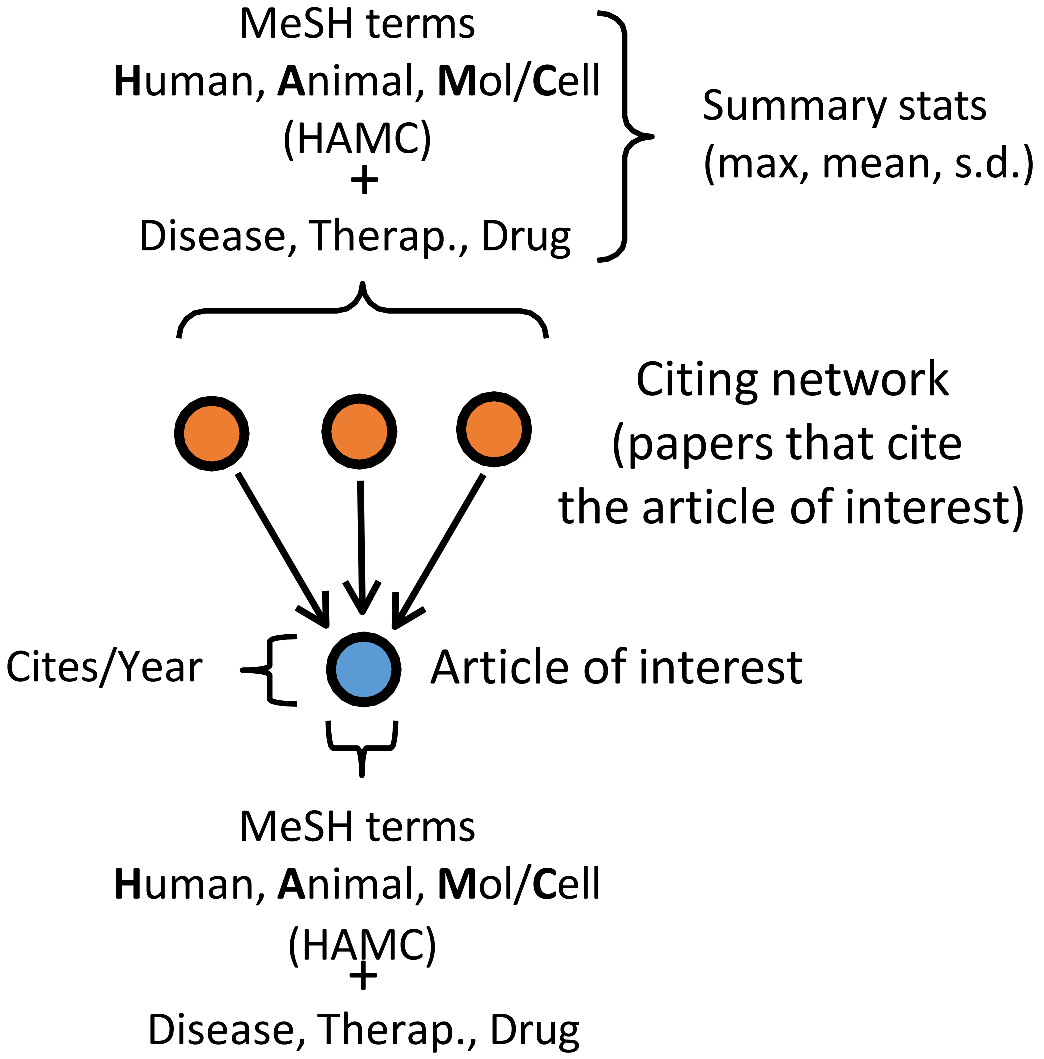
Step 1

Build data profiles

Create a training set of
data elements associated with…

— Each article of interest
  - Citations per year
  - MeSH categories (HAMC)
  - Modifying MeSH terms
    - Disease
    - Therapeutic/Diagnostic Approaches
    - Chemicals/Drugs

— Each corresponding citing network
  - MeSH categories (HAMC)
  - Modifying MeSH terms
    - Disease
    - Therapeutic/Diagnostic Approaches
    - Chemicals/Drugs
  - Summary stats (max, mean, s.d.)



MeSH terms
Human, **A**nimal, **M**ol/**C**ell
(HAMC)
+
Disease, Therap., Drug

Summary stats
(max, mean, s.d.)

Citing network
(papers that cite
the article of interest)

Cites/Year

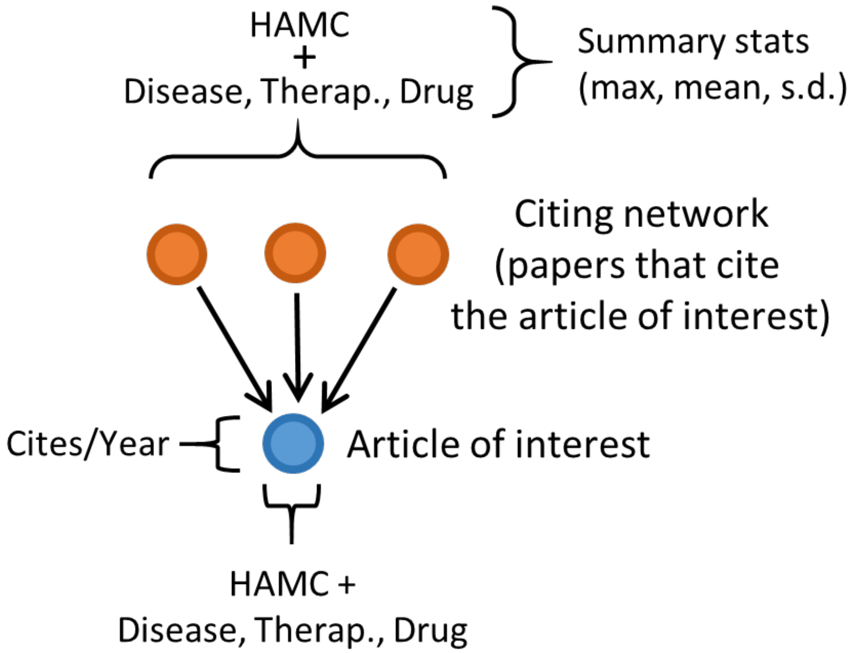
Article of interest

MeSH terms
Human, **A**nimal, **M**ol/**C**ell
(HAMC)
+
Disease, Therap., Drug

Using machine learning to predict translation

Step 2

Train the algorithm



HAMC + Disease, Therap., Drug

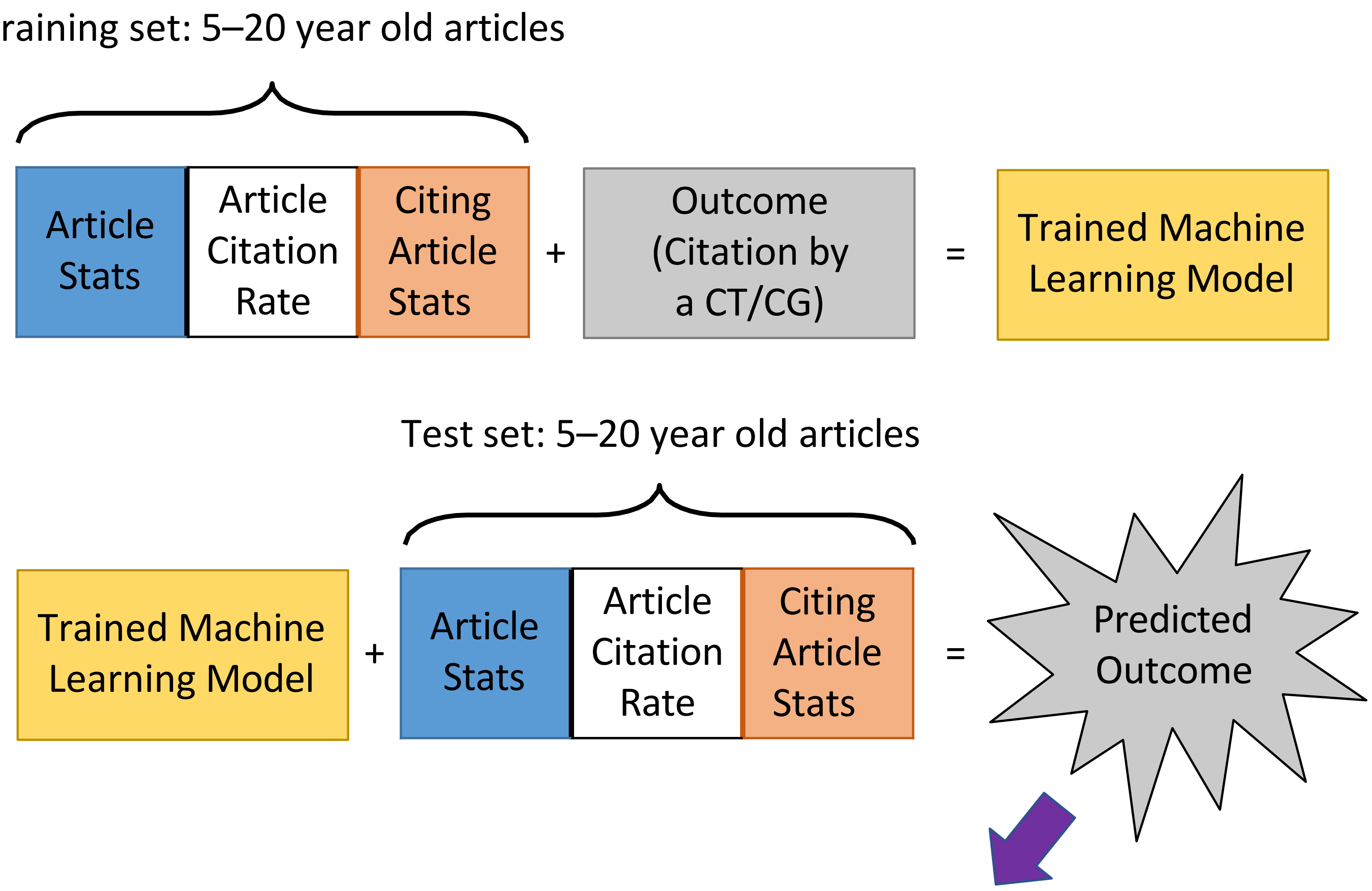
Summary stats (max, mean, s.d.)

Citing network (papers that cite the article of interest)

Cites/Year

Article of interest

HAMC + Disease, Therap., Drug



Training set: 5–20 year old articles

Article Stats | Article Citation Rate | Citing Article Stats + Outcome (Citation by a CT/CG) = Trained Machine Learning Model

Test set: 5–20 year old articles

Trained Machine Learning Model + Article Stats | Article Citation Rate | Citing Article Stats = Predicted Outcome

Approximate **P**otential to **T**ranslate (**APT**) scores

(estimated likelihood of translation)

Using machine learning to predict translation

Step 3

Test recall and precision

— **Limited**

Use data profiles two years after publication

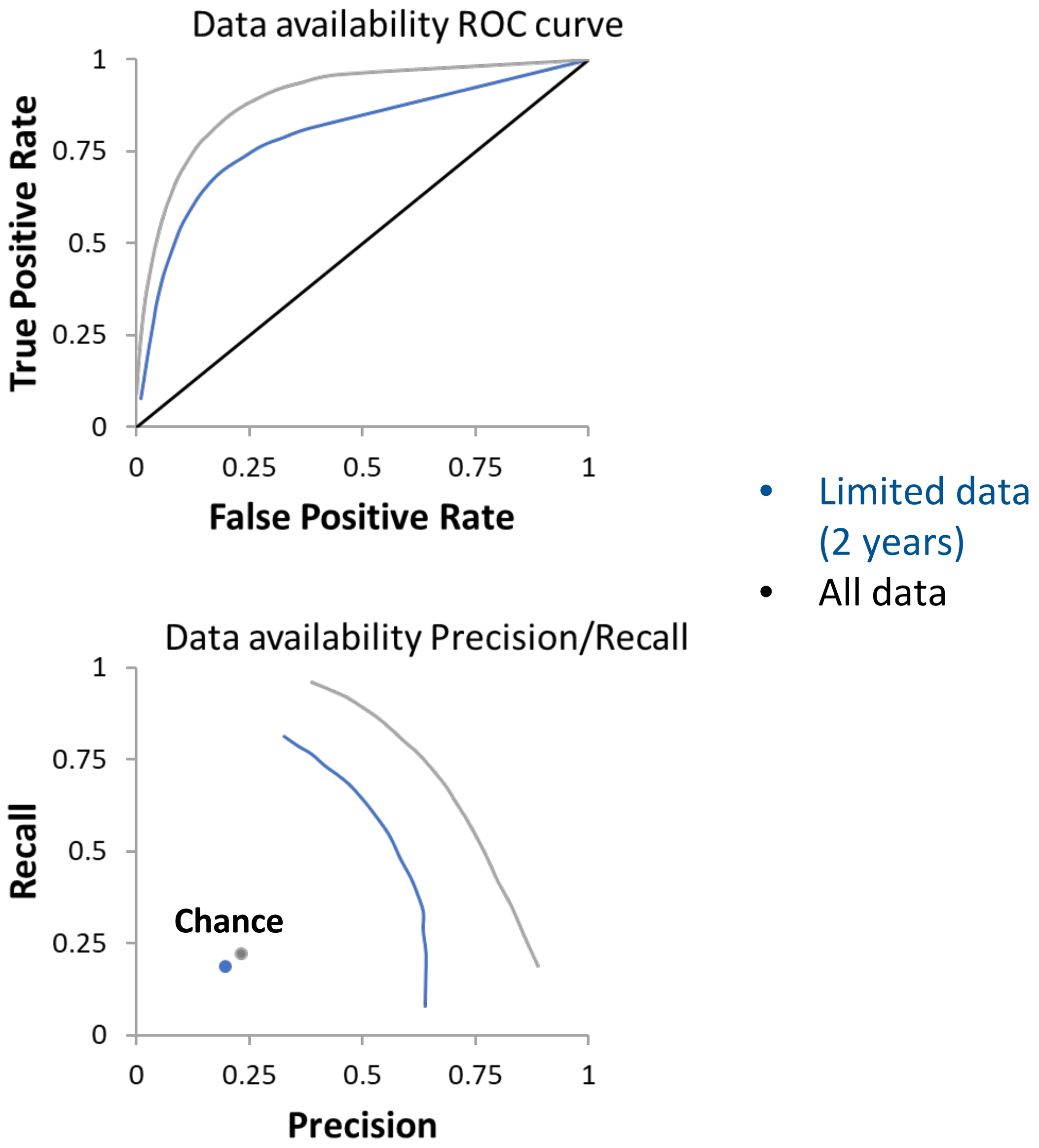
- 84% accuracy

— **Expanded**

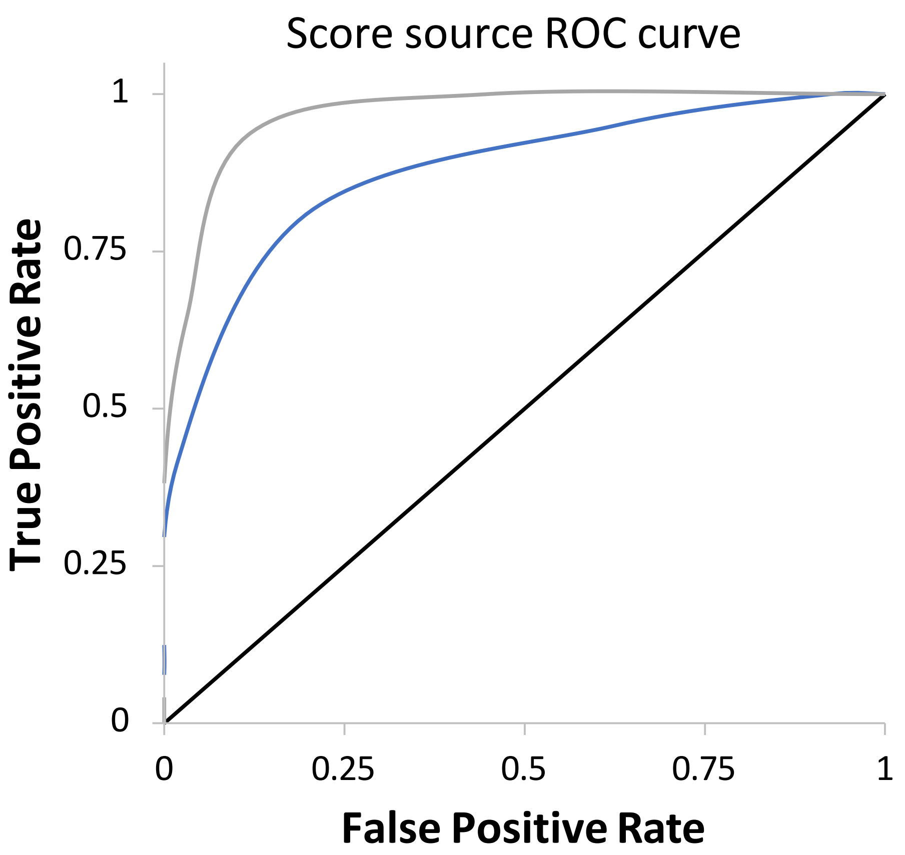
Use data profiles for all years after publication

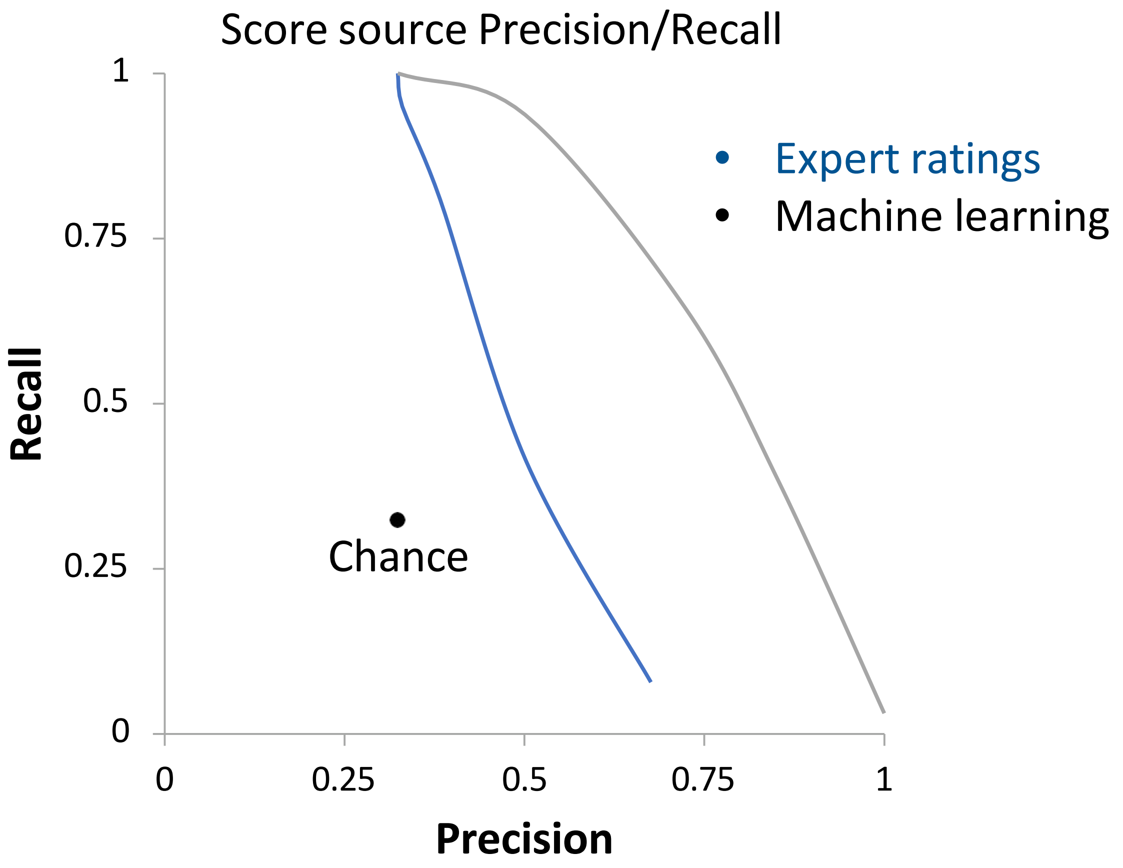
- 85% accuracy

Predictions made two years after publication are almost as accurate as those made when including data available after many more years have elapsed



Machine learning predictions are at least as accurate as expert review

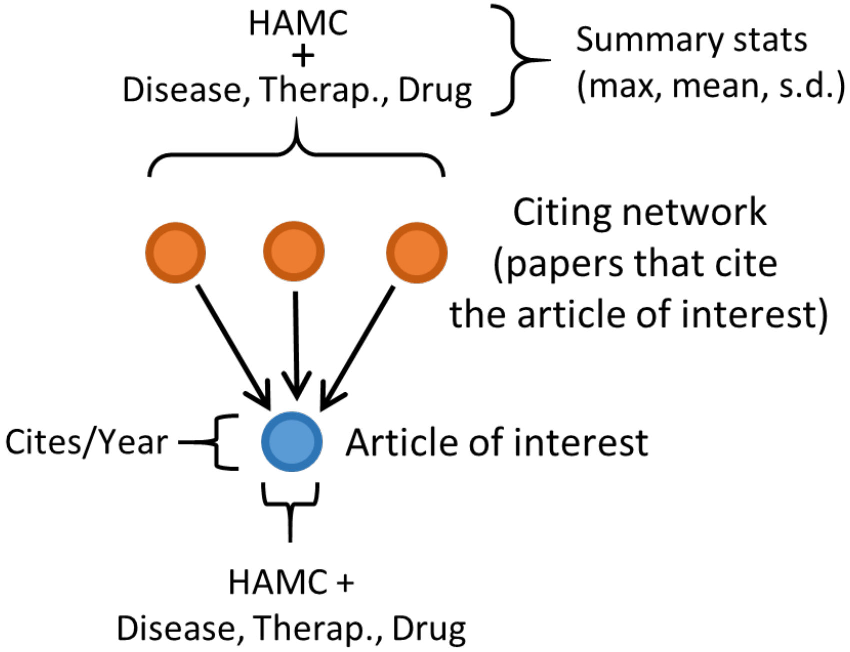




Can we identify particular data profiles that have a high likelihood of translation?

The ranked importance of variables that determine APT scores differs between human-focused and fundamental articles

Rank	All articles	Human-focused articles (100% H MeSH)	Fundamental articles (0% H MeSH)
1	Cites/Year (P)	Cites/Year (P)	**Human (mean)**
2	Human (mean)	Drug (s.d.)	**Human (s.d.)**
3	Human (s.d.)	Disease (s.d.)	Cites/Year (P)
4	Disease (mean)	Therap. (mean)	Disease (mean)
5	Human (max)	Human (mean)	Disease (s.d.)



HAMC + Disease, Therap., Drug

Summary stats (max, mean, s.d.)

Citing network (papers that cite the article of interest)

Cites/Year

Article of interest

HAMC + Disease, Therap., Drug

Proof of concept: can we use citation "genetics" to discover what types of citing papers increase an article's APT score?

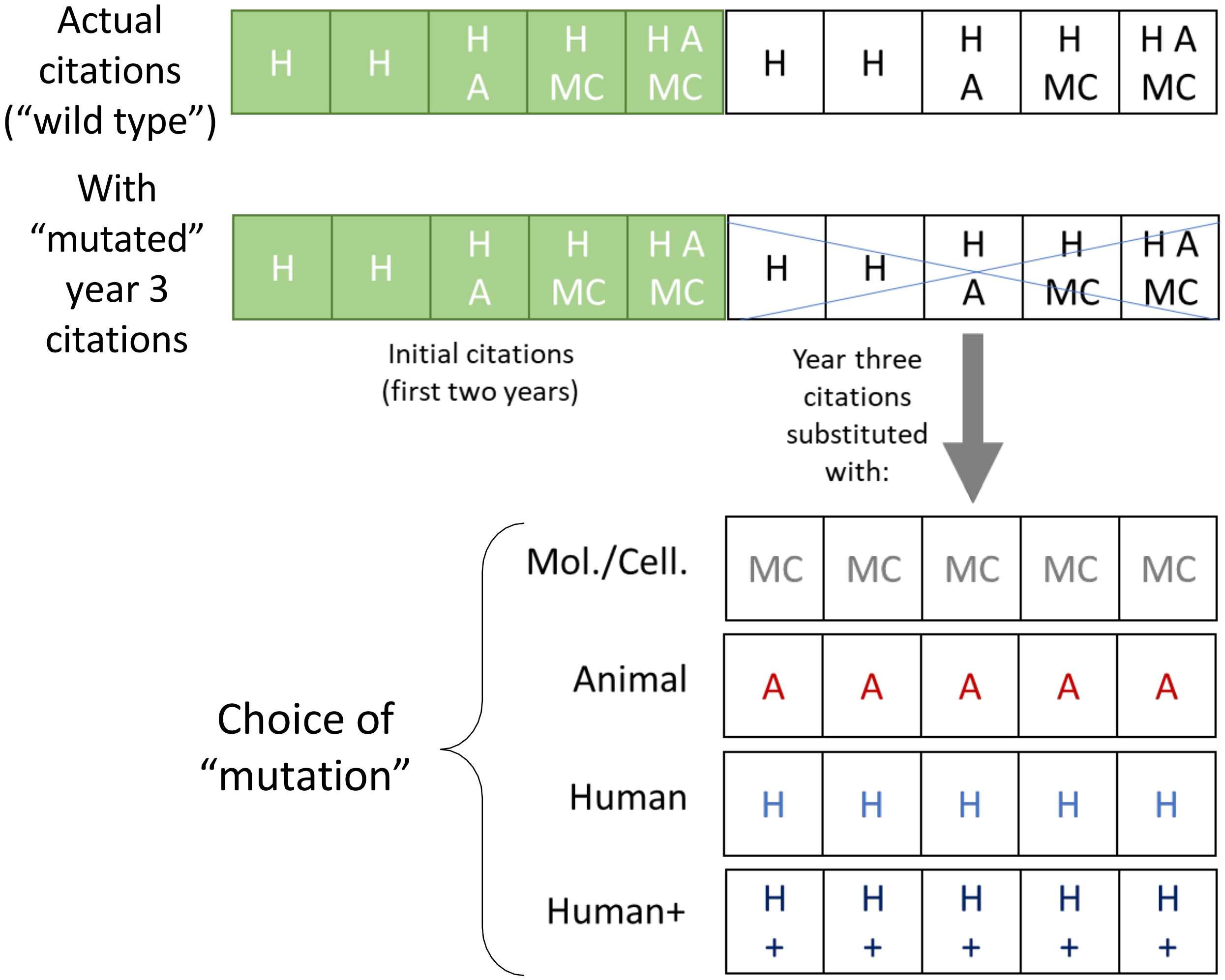
Articles of interest in this experiment were those with an APT score of 25% after two years that increased to an APT score of 50-95% after three years

"Mutate" the citing papers in year three to include only the following MeSH terms:

- Human
- Animal
- Mol/Cell
- Human plus Disease, Therapeutic/Diagnostic, and Chemical/Drug (H+)

Test with the trained machine learning algorithm

How does engineering the data profiles of citing papers impact APT scores?



Actual citations ("wild type")	H	H	H A	H MC	H A MC	H	H	H A	H MC	H A MC
With "mutated" year 3 citations	H	H	H A	H MC	H A MC	~~H~~	~~H~~	~~H A~~	~~H MC~~	~~H A MC~~

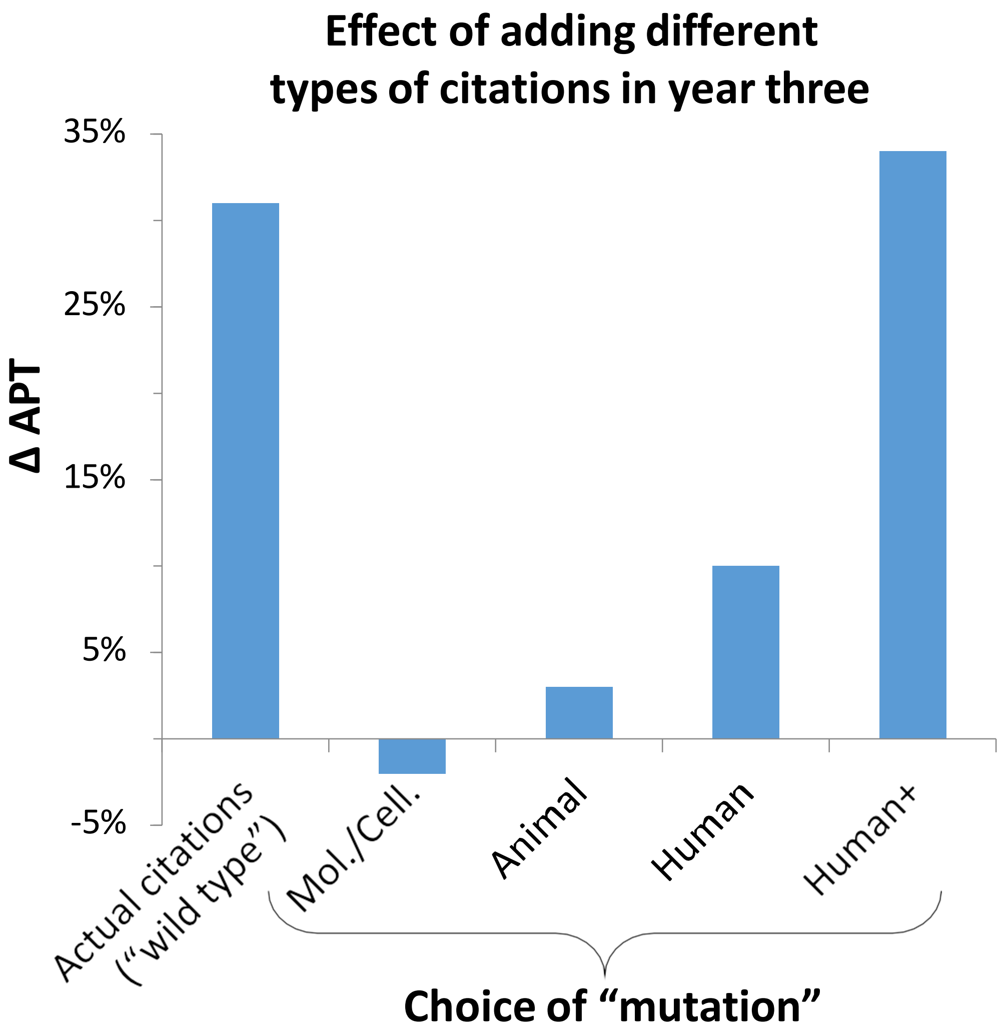
Initial citations (first two years)

Year three citations substituted with:

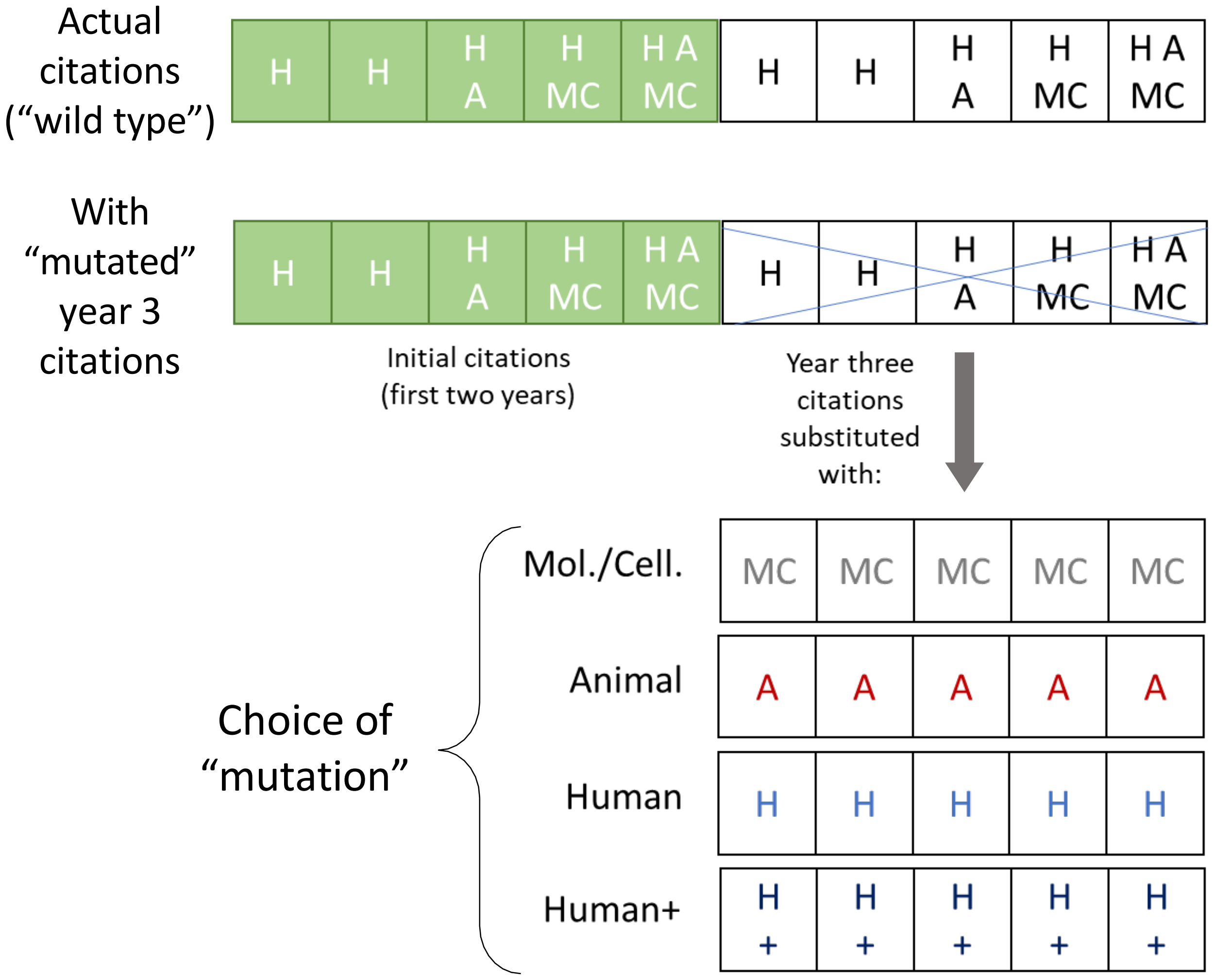
Choice of "mutation"

Mol./Cell.	MC	MC	MC	MC	MC
Animal	A	A	A	A	A
Human	H	H	H	H	H
Human+	H +	H +	H +	H +	H +

The strongest predictor of translation for all articles is citation by papers with Human+ MeSH terms



Effect of adding different types of citations in year three



Actual citations (“wild type”)	H	H	H A	H MC	H A MC	H	H	H A	H MC	H A MC
With “mutated” year 3 citations	H	H	H A	H MC	H A MC	H	H	H A	H MC	H A MC

Initial citations (first two years)

Year three citations substituted with:

Choice of “mutation”

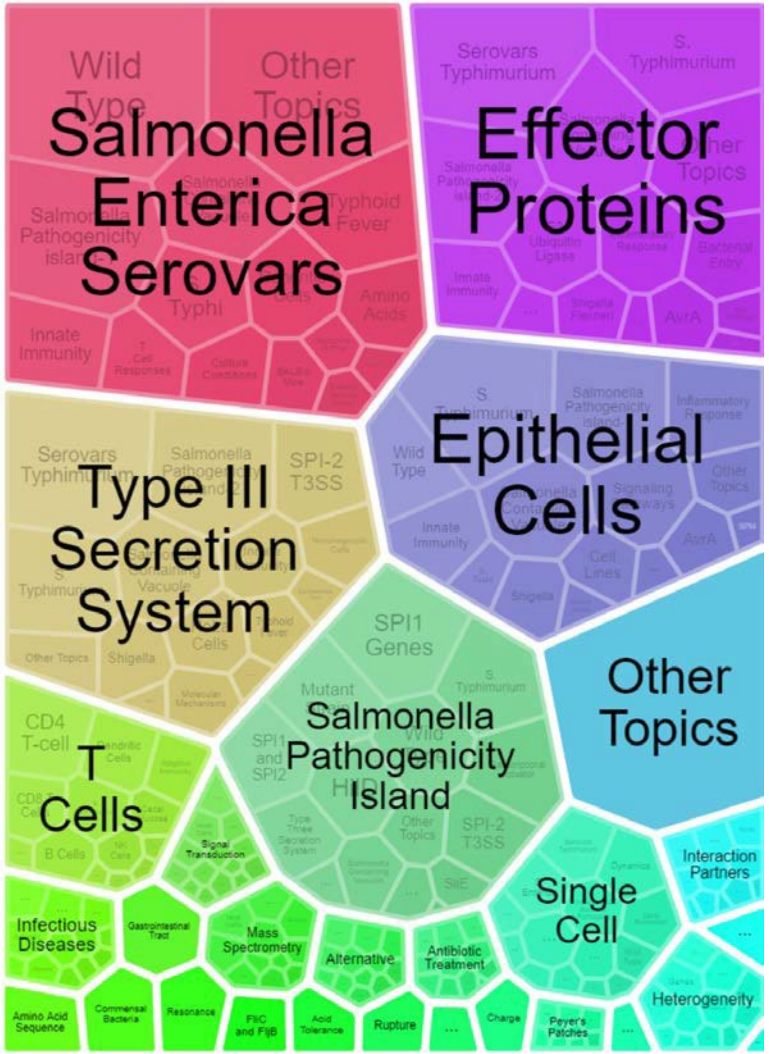
Mol./Cell.	MC	MC	MC	MC	MC
Animal	A	A	A	A	A
Human	H	H	H	H	H
Human+	H +	H +	H +	H +	H +

NIH National Institutes of Health
Office of Portfolio Analysis

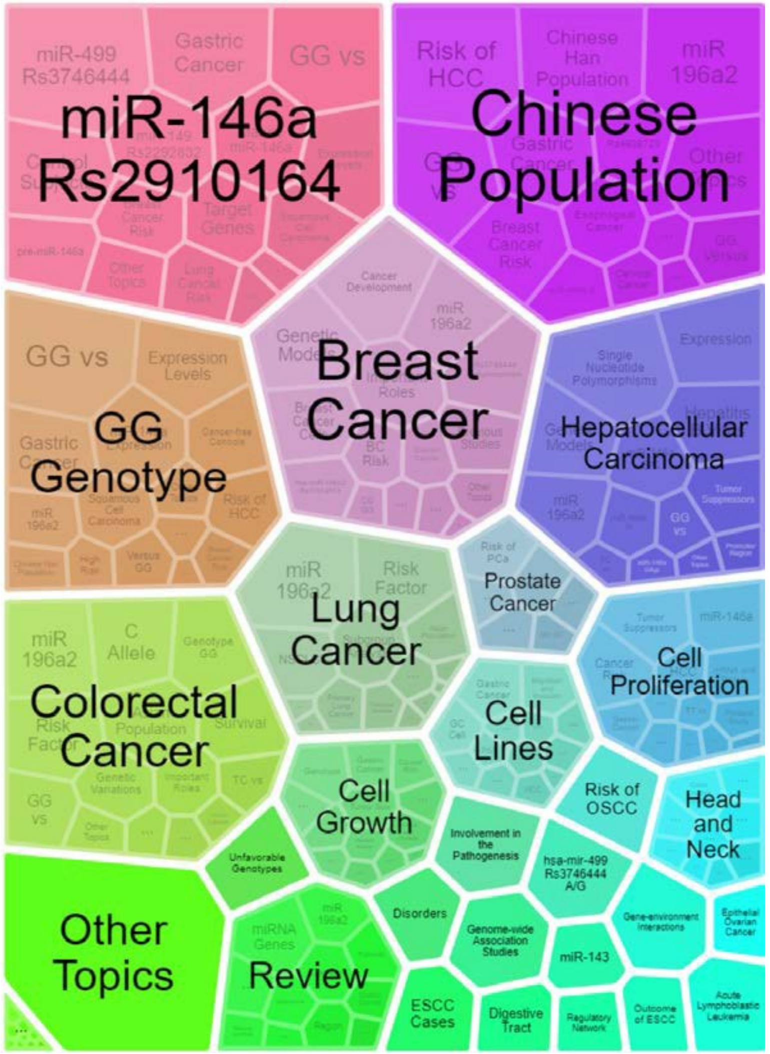
Can we use APT scores in the real world to identify emerging areas of translational research?

Which of these areas of research is most likely to translate from bench to bedside?

Salmonella pathogenesis
810 publications
2006 to 2017

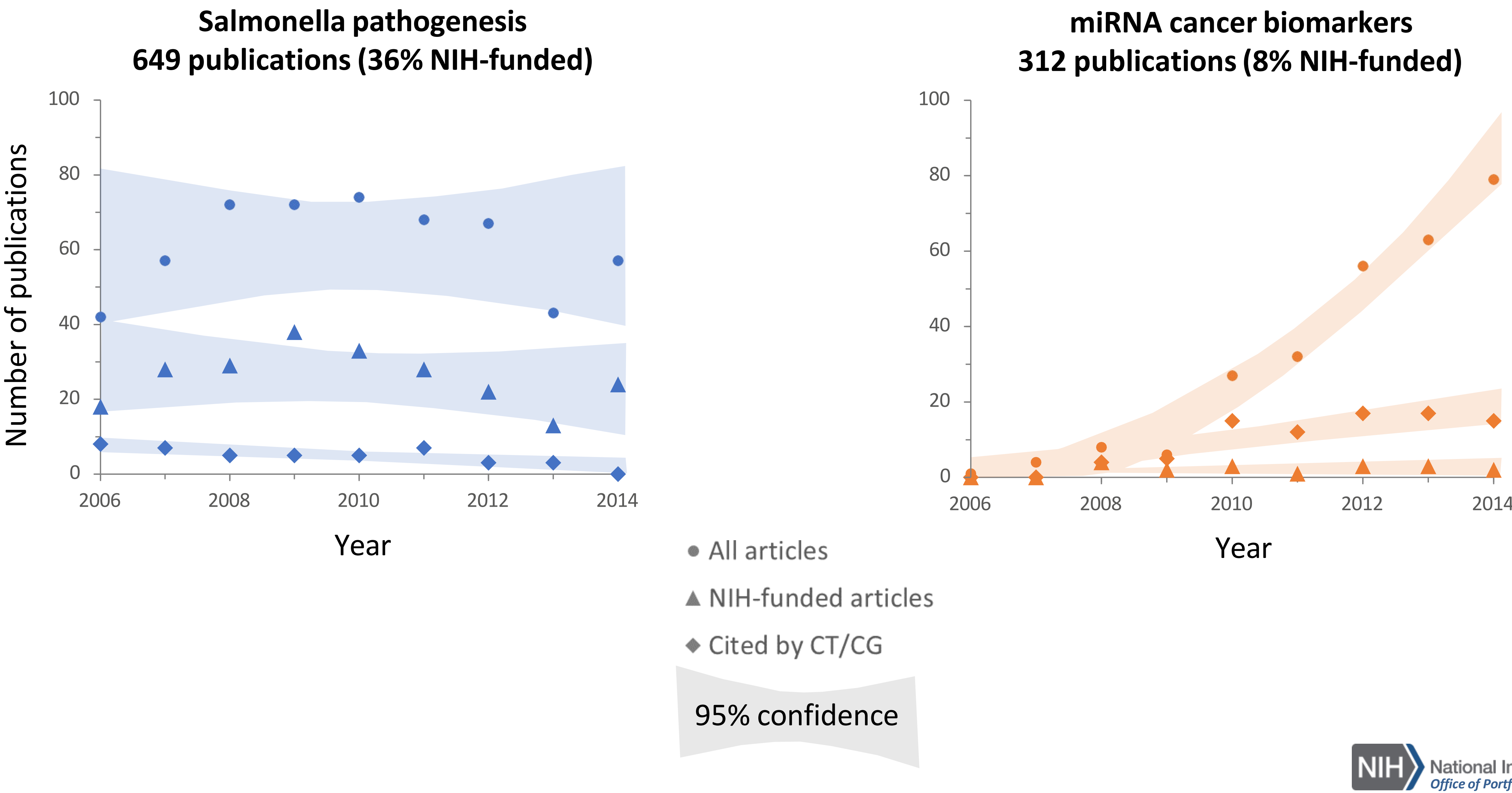


Cancer biomarkers
475 publications
2006 to 2017



NIH National Institutes of Health
Office of Portfolio Analysis

Two examples: Salmonella pathogenesis and miRNA cancer biomarkers



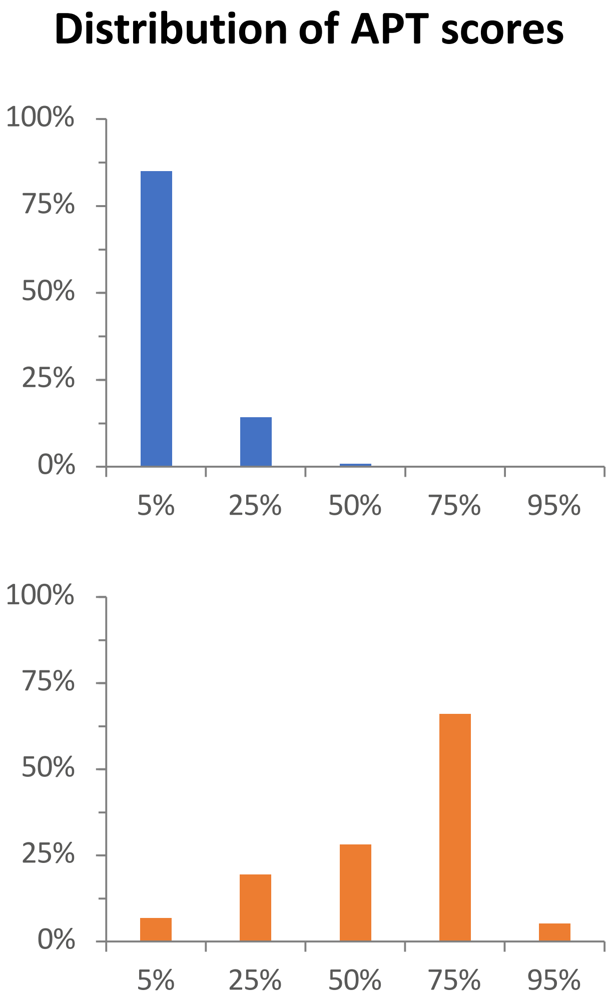
Salmonella pathogenesis
649 publications (36% NIH-funded)

miRNA cancer biomarkers
312 publications (8% NIH-funded)

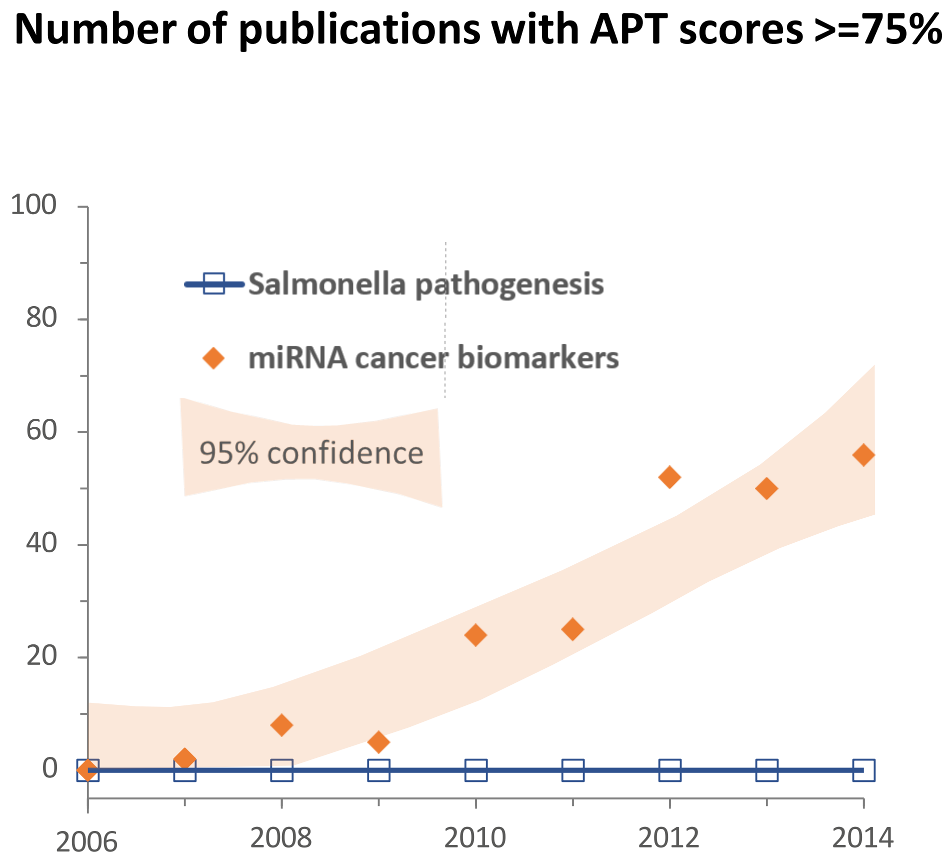
Two examples: Salmonella pathogenesis and miRNA cancer biomarkers

Salmonella pathogenesis
2006 to 2014
649 publications

miRNA cancer biomarkers
2006 to 2014
312 publications



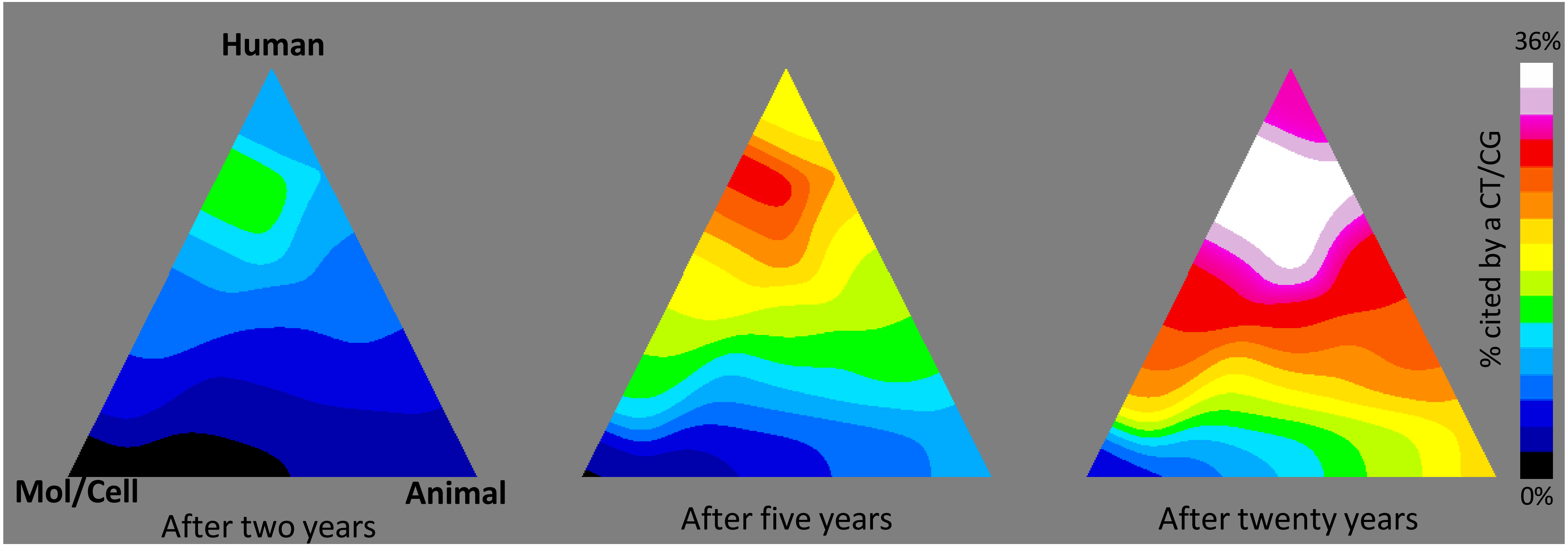
Distribution of APT scores



Number of publications with APT scores >=75%

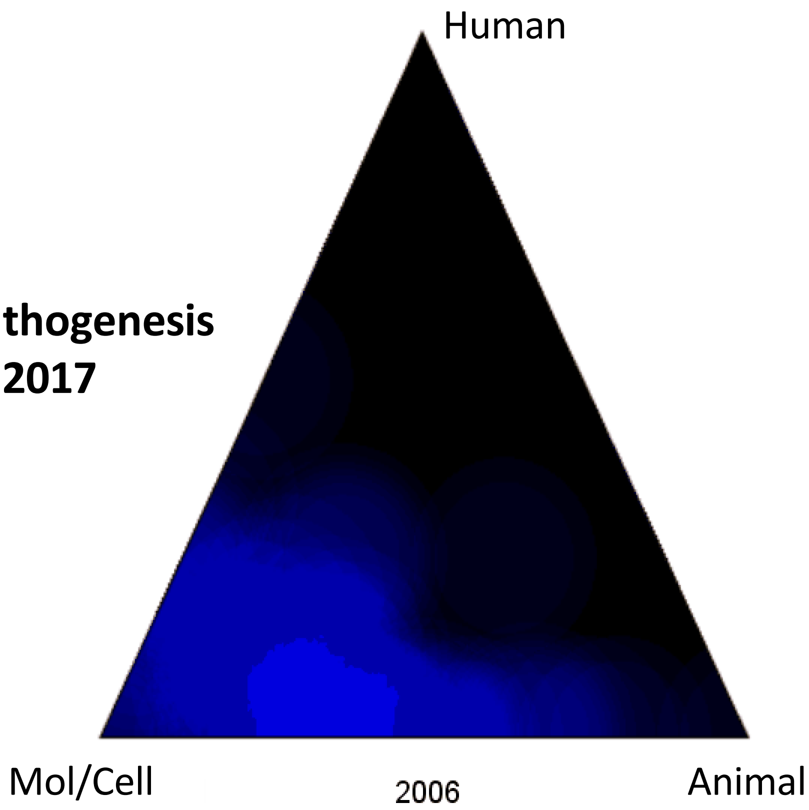
MeSH profile of publications cited by CT/CGs over time

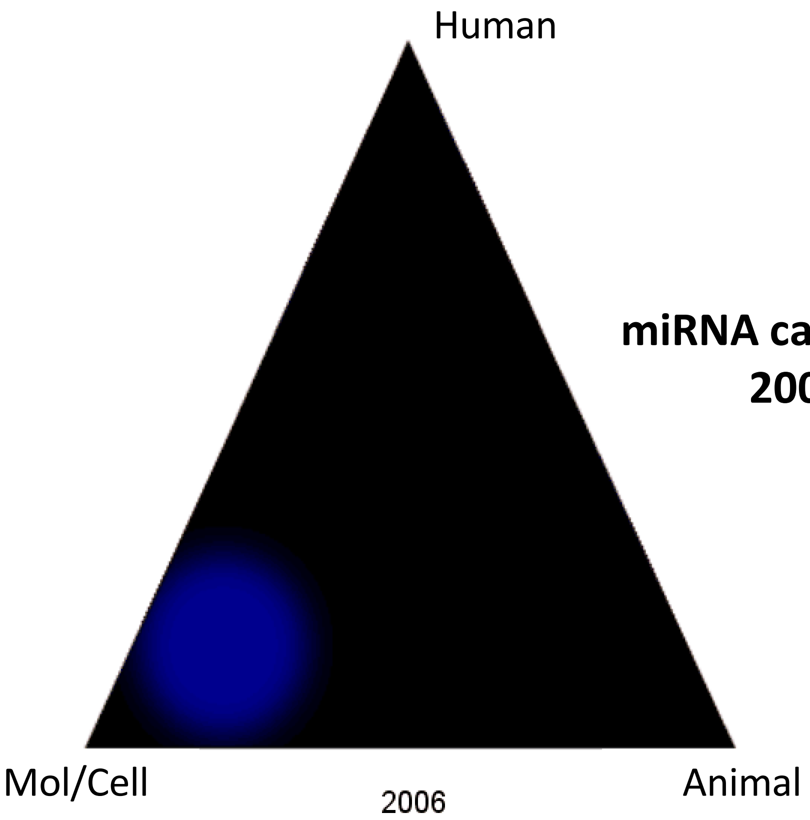
All PubMed publications in 1995; fractional counting



Visualizing the pattern of emergent translational science

Salmonella pathogenesis
2006 to 2017





miRNA cancer biomarkers
2006 to 2017

National Institutes of Health
Office of Portfolio Analysis

Summary and Conclusions

— As expected, fundamental research articles take longer to be cited by a clinical trial or guideline (CT/CG)

- Twice as many human-focused as fundamental research articles eventually receive a clinical citation
- The cohort of fundamental research articles cited by a CT/CG grows steadily over the first ten years after publication

— Information conveyed by the scientific community within two years after publication suffices for our machine learning model to predict the likelihood of citation by a CT/CG (**A**pproximate **P**otential to **T**ranslate, or APT score)

— APT scores are as least as good as subject matter experts in predicting clinical impact

— MeSH profiles of the papers citing a research article are predictive of its clinical impact

- Citation by Molecular/Cellular-focused papers **decrease** an article's APT score
- Citation by human-focused papers that have modifying MeSH terms (Disease, Therapeutic/Diagnostic, and Chemicals/Drugs) **increase** an article's APT score

— APT scores can be used to identify emerging areas of translational research

Acknowledgments

OPA Analysts/Data Scientists

Ian Hutchins
Paula Fearon
Travis Hoppe
Rob Harriman
Patricia Forcinito
Aviva Litovitz
Matt Perkins
Rebecca Meseroll
Abbey Zuelke
Esha Sinha

OPA Software Developers

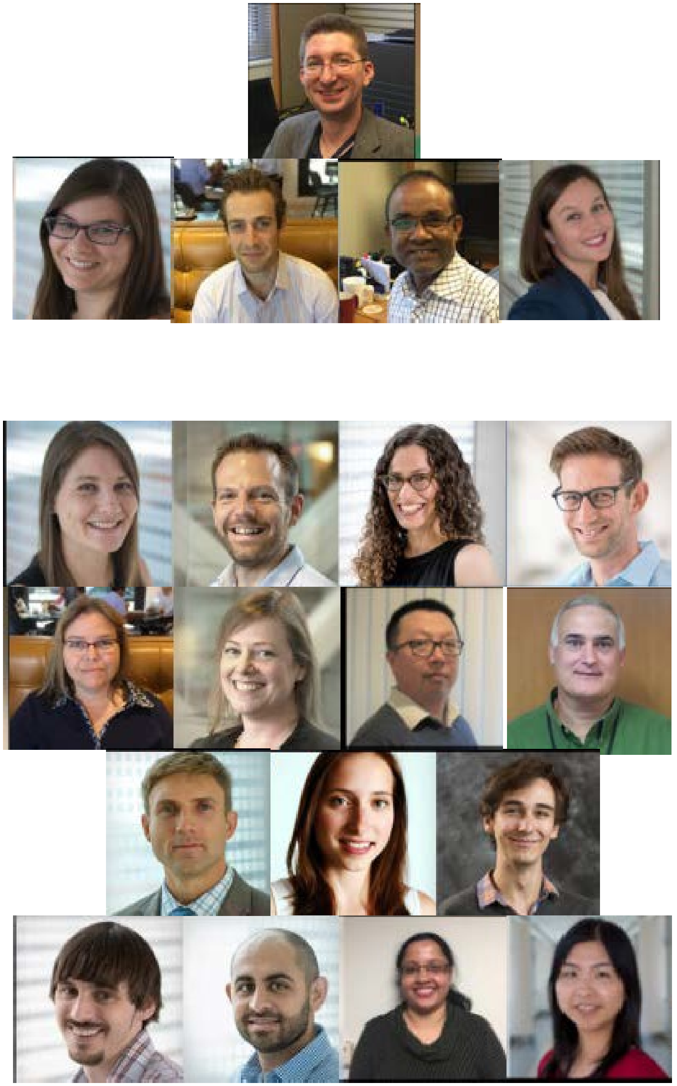
Kirk Baker
Matt Davis
Payam Meyer
Hao Yu
Ehsan Haque
Tom Xu
Shannon Davis
Brad Busse

OPA IT Specialist

Chuck Lynch

OPA Admin Support

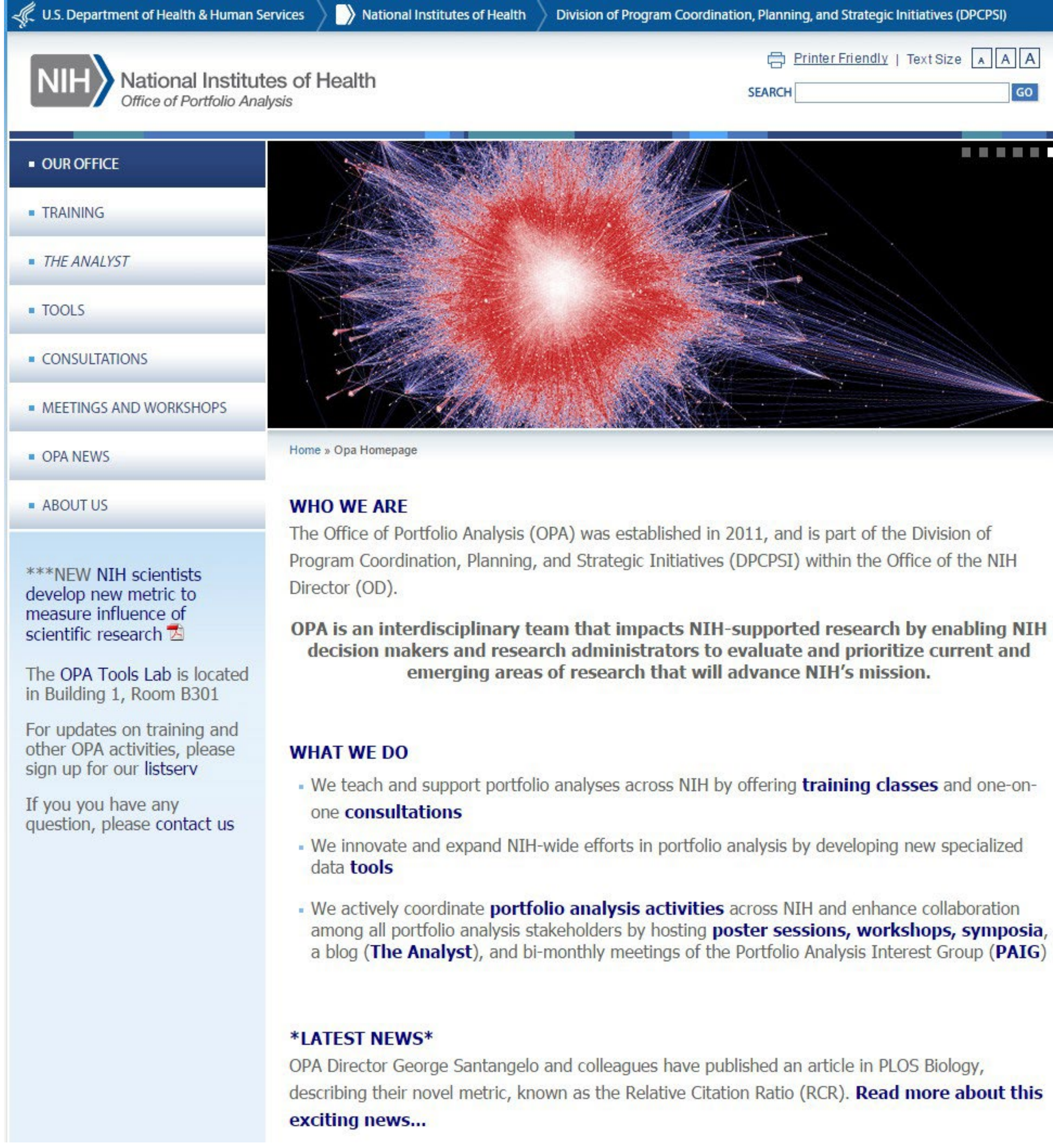
Sharon Chaney



LEXICAL INTELLIGENCE

über RESEARCH

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



U.S. Department of Health & Human Services › National Institutes of Health › Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI)

NIH National Institutes of Health
Office of Portfolio Analysis

***NEW NIH scientists develop new metric to measure influence of scientific research

The OPA Tools Lab is located in Building 1, Room B301

For updates on training and other OPA activities, please sign up for our listserv

If you you have any question, please contact us

WHO WE ARE

The Office of Portfolio Analysis (OPA) was established in 2011, and is part of the Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI) within the Office of the NIH Director (OD).

OPA is an interdisciplinary team that impacts NIH-supported research by enabling NIH decision makers and research administrators to evaluate and prioritize current and emerging areas of research that will advance NIH's mission.

WHAT WE DO

- We teach and support portfolio analyses across NIH by offering **training classes** and one-on-one **consultations**
- We innovate and expand NIH-wide efforts in portfolio analysis by developing new specialized data **tools**
- We actively coordinate **portfolio analysis activities** across NIH and enhance collaboration among all portfolio analysis stakeholders by hosting **poster sessions, workshops, symposia**, a blog (**The Analyst**), and bi-monthly meetings of the Portfolio Analysis Interest Group (**PAIG**)

LATEST NEWS

OPA Director George Santangelo and colleagues have published an article in PLOS Biology, describing their novel metric, known as the Relative Citation Ratio (RCR). **Read more about this exciting news...**