Professional Development Workshop: Irresponsible Conduct of Research: What Happens When Things Go Wrong

Data Management, Publication/Authorship, and Peer Review

Thomas B. Knudsen, PhD
Developmental Systems Biologist
US EPA, National Center for Computational Toxicology
knudsen.thomas@epa.gov
ORCHID 0000-0002-5036-596x

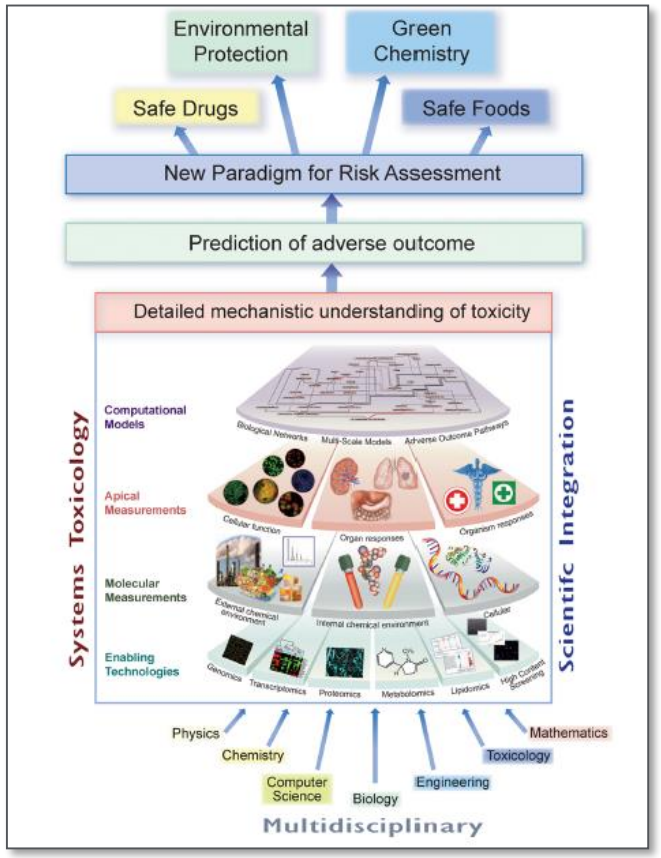
DISCLAIMER: The views expressed are those of the presenter and do not reflect Agency policy.

San Diego California

June 22–26, 2019

Teratology Society's 59th Annual Meeting

Teratology in the era of ‘big-data’



Sturla et al. (2014) Chem Res Toxicol 27

- Biological systems are complex (and the data equally so) requiring integration across disciplines;
- the technologies we use and biological questions we ask continue to be increasingly dependent on data science;
- rewards (and risks): mining for hidden correlations difficult or impossible to extract from smaller individual studies;
- responsible conduct of research is critical to your professional development in the era of data-rich biology.

‘Big-data’ connectivity

- ***Replication*** is the cornerstone of good science; however, this becomes increasingly difficult as the technologies used, and the biological questions being asked, become more and more complex.

- ***Where we are:*** unprecedented amounts of data from genomic sciences, epidemiological studies, high-throughput screening, and high-content imaging now provide the opportunity to profile the toxicological landscape.

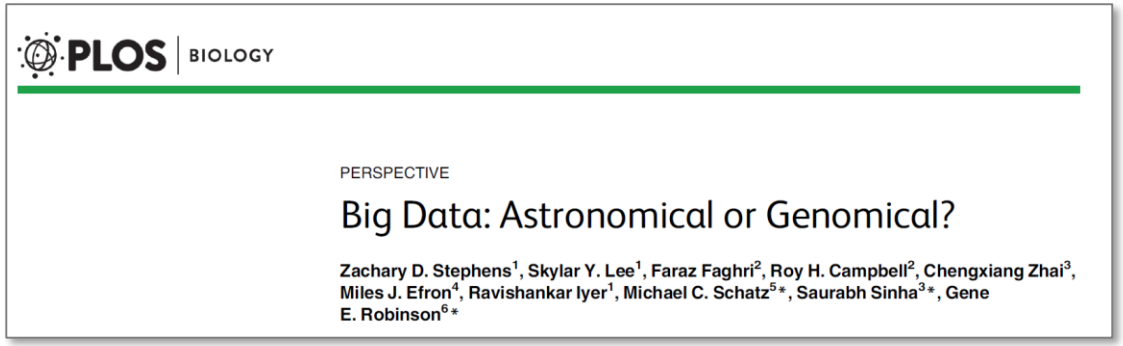
- ***Where we need to be:*** open access to structured and unstructured data, facilitate data-sharing, and promote analytical-transparency to enable replication and extension of results from independently collected datasets.

https://www.the-scientist.com May, 2019

- Innovations in machine learning and virtual-reality technologies for biomedical research applications:
  - *in silico* brain
  - algorithms for cancer treatment
  - data-diving in biological sciences
  - drug development (and toxicity testing).
- Today's fastest supercomputers compute at the petascale (10^{15} calculations per second);
- next frontier, exascale, computes 10^{18} calculations per second and mimics the 'speed of life';
- the promise for revolution in biology comes with the need for reproducibility (GIGO).

Genomics is a “four-headed beast”:

data acquisition, storage, distribution, and analysis pose great demands


PLOS BIOLOGY

PERSPECTIVE

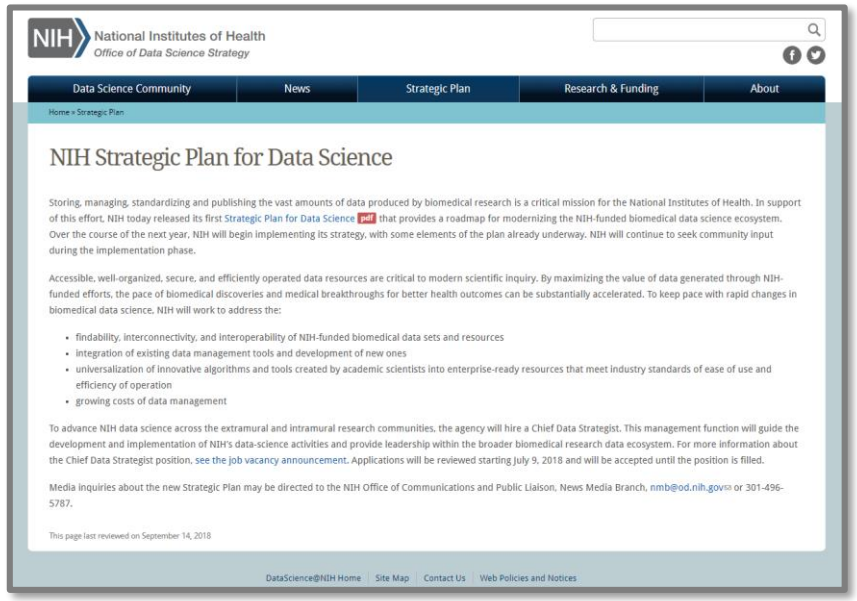
Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

Data science brings significant career opportunities for biologists working with AI technologies and should be part of undergraduate and graduate training.

- up to 2B human genomes could be sequenced by 2025, and scRNAseq technologies sequencing thousands of individual cells in a single sample;
- genomics data will soon exceed Astronomy, YouTube and Twitter combined; these datsets need to be compared in huge numbers for precision medicine to be effective;
- as corporate data stores will grow to an inconceivable dimension, a clear data strategy is necessary to make sure data resources can be used, shared and moved efficiently.

NIH's strategic plan for data science



NIH's Big Data to Knowledge (BD2K) program launched in 2014 led to a strategic plan to ensure data-science activities and products remain nimble and secure but adhere to 'FAIR' principles

Findable — unique identifiers that are searchable

Accessible — retrieved by open systems, secure authentication

Interoperable — standardized vocabularies

Reusable — adequately described to a new user

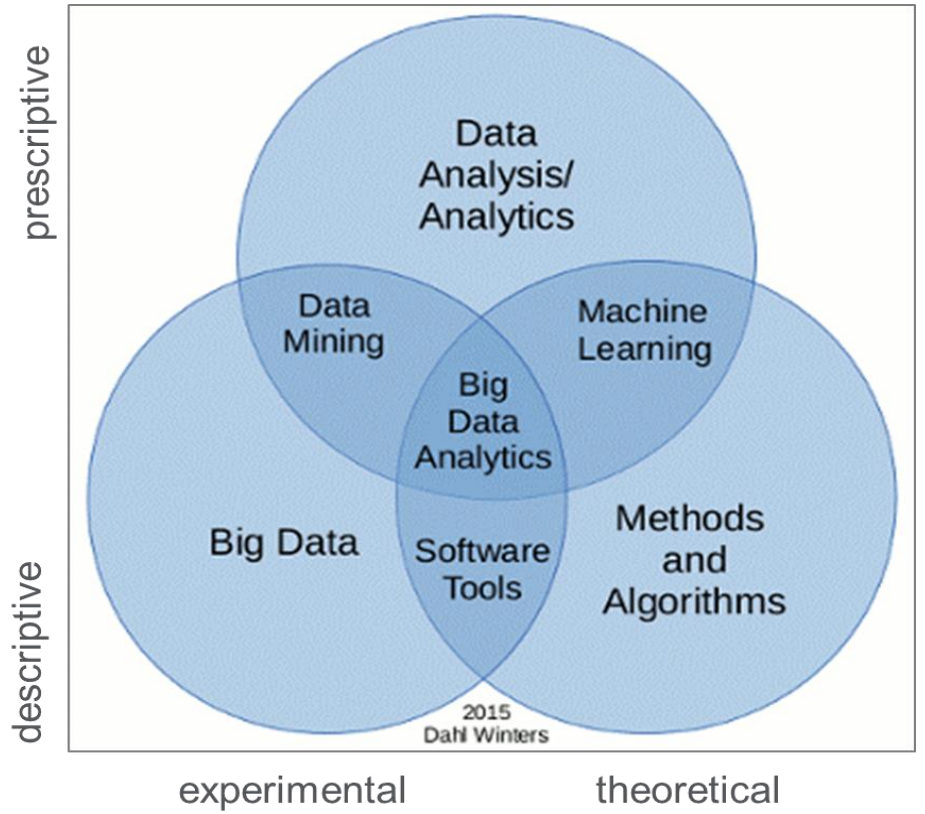
https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

‘Big-Data’ Analytics: *the bigger picture*

- ***What:*** mine massive amounts of data for hidden correlations that are difficult to extract from smaller individual studies.

- ***Why:*** generate novel testable hypotheses about biological systems that can be applied to systems toxicology.

- ***How:*** dealing with big-data analytics requires *4Vs* in integrative data science: Volume, Velocity, Variety, Veracity.



“The culture baggage of biology that privileges data generation over all other forms of science is holding us back.”

- Larry Hunter, U Colorado – Denver

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

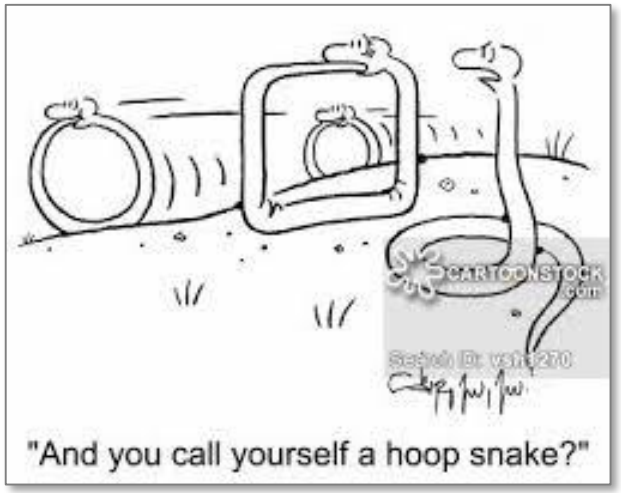
Geir Kjetil Sandve[1,2*], Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

PLOS Computational Biology (2013) vol 9: e1003285

To sustain a scientifically robust outcome in the big-data era:

- *formulate the complex problem under investigation in a clear and lucid manner*
- *identify new solutions and limitations in the methodologies*
- *indicate what the work is trying to achieve*
- *provide a perspective of the work consistent with the nature of a journal.*

- Reasons for breakdown:
  - competing demands on investigator's time
  - publication pressure to get the research out while still relevant
  - omission of experimental details to meet page limitations
  - privacy in human studies, intellectual property rights
  - sociological issues (lack of credit, inadequate training)
  - …



When things go wrong: *publication fraud and human error*

- *Science* writer J Bohannon set up a sting with an overtly bogus manuscript on a cancer wonder drug submitted to multiple journals:

  - writing as “Ocorrafoo Cobange” of the “Wassee Institute of Medicine in Asmara” (neither of which exist), he created a fictitious database of molecules, lichens, and cancer cell lines;

  - then queried the database to randomly address the concept ‘*Molecule X from lichen species Y inhibits the growth of cancer cell Z*’ using a computer program;

  - hundreds of fatally flawed manuscripts touting the wonder drug were then submitted as bait to 304 open-access journals and 167 predatory publishers;

  - by the time Science went to press (October, 2013), 157 journals accepted the bogus paper (16 despite damning reviews) and 98 rejected it.

- Retraction Watch (https://retractionwatch.com) tracks 500-600 papers yearly with human error or intentional misconduct (18K in total).

EDITORS … *"a personal view of some reasons for desk-rejection"*

- Out of scope: study overall does not align with journal priorities
- Writing style: poor verbal quality or weak organization
- Cross-checking: "ithenticate" finds excessive overlap (>27%) with published literature.
- Animals: ARRIVE checklist [https://www.elsevier.com/__data/promis_misc/622936arrive_guidelines.pdf]
- Study design: descriptive, regulatory, mechanistic?
- Scale: small sample sizes (n), input parameters (k), or output parameters (p)
- Data quality: incomplete methodology or representation (figures, tables)
- Outdated referencing: context should extend knowledge from the past 5-6 years
- ….

Relevance to 'big-data': editors can encourage harmonization of discovery-based or hypothesis-driven approaches to toward AOP databases for which quantitative mechanistic relationships can be made.

Final Thoughts

- Open data-sharing is a cornerstone of systems toxicology to enable replication and extension of results from independently collected datasets.

- Remember the *4Vs* in integrative data science: Volume, Velocity, Variety, Veracity as best practices in publishing studies that generate or utilize ‘big-data’.

- Must incentivize ways to make sense of ‘big-data’ on a broader scale without falling prey to a meaningless mass of interconnected data linkages.

- Best practices for reproducible computational research (authors), encourage studies utilizing extant data (editors), and promote scientific openness (publishers).