Reproducible Computational Research
Why, What, How

Vasant Honavar
Artificial Intelligence Research Laboratory
Computer Science Graduate Program
Bioinformatics and Genomics Graduate Program
Neuroscience Graduate Program
Center for Big Data Analytics and Discovery Informatics
Pennsylvania State University

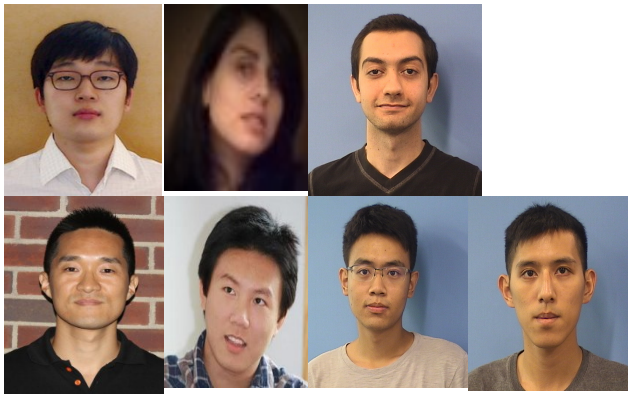
vhonavar@ist.psu.edu
http://faculty.ist.psu.edu/vhonavar

Research Interests

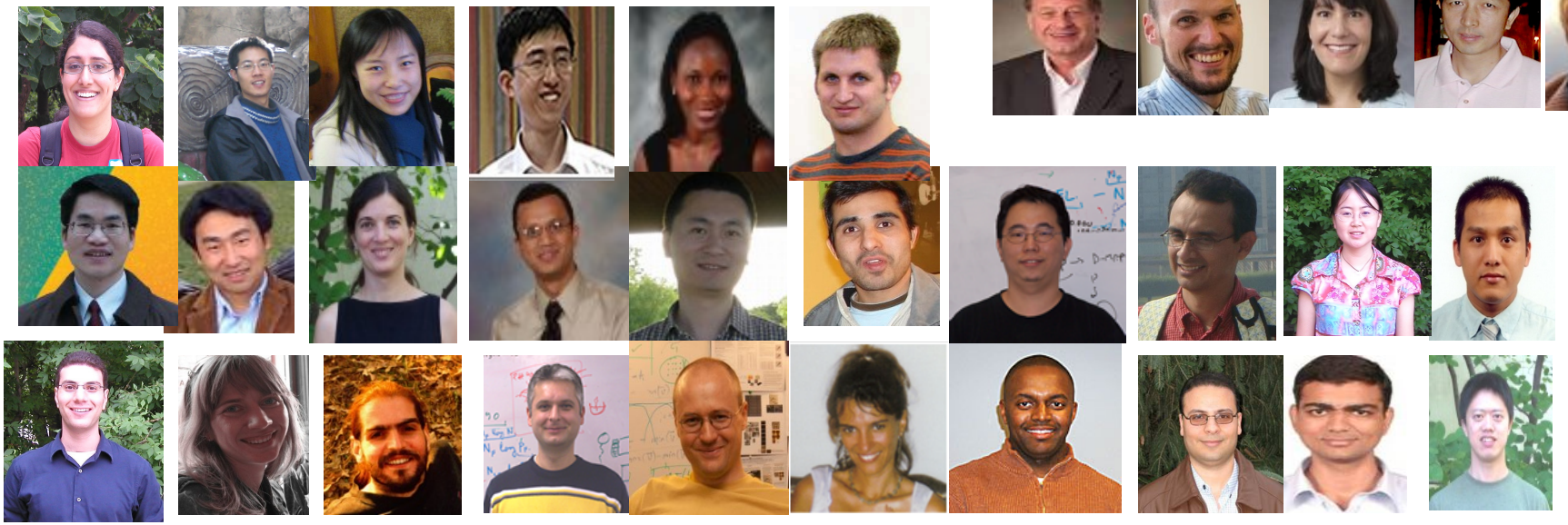
- Machine learning: Statistical, information theoretic, linguistic and structural approaches to machine learning; learning predictive relationships from sequential, graph-structured, multi-relational, multimodal, partially specified, partially labeled, distributed data, linked data
- Causal Inference: Causal inference from disparate experimental and observational studies, causal inference from relational data, causal inference from temporal data
- Knowledge Representation and Inference: Logical, probabilistic, and decision-theoretic knowledge representation and inference; federated knowledge bases; selective information sharing; federated services; representing and reasoning about qualitative preferences
- Applied Informatics
  - Bioinformatics: Macromolecular structure and function, analysis, inference, modeling, and prediction of macromolecular (protein-protein , protein-RNA, and protein-DNA) interaction networks and interfaces, immune networks, etc.
  - Health Informatics: Predictive and causal modeling of health outcomes from patient (health records, genomics, socio-economic, environmental) data
  - Brain Informatics: Modeling and analysis of structure and dynamics of brain networks from fMRI data
- Algorithmic Discovery:
  - Algorithmic abstractions of scientific domains
  - Representations of scientific artifacts (experiments, data, models, assumptions, hypotheses, theories …)

Artificial Intelligence Research Laboratory

Current Ph.D. Students



Recent Ph.D. Graduates (2005 – 2017)



Collaborators

First: A story from the trenches

- Andorf, Carson, Drena Dobbs, and Vasant Honavar. "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach." *BMC bioinformatics* 8.1 (2007): 284.

Protein annotation using machine learning: Background

- Exponential increase in protein sequences
- Experimental determination of structure and function lags behind
- Automated methods for protein function annotation
  - Allow high-throughput annotation of thousands of sequences
  - Increase the risk of error propagation
- Potential sources of errors
  - noisy training data
  - error in the algorithm
  - mislabeled data
  - poor performance of algorithm
  - simple clerical errors
  - human error

Focus: Protein Kinases

- Protein Kinases are among
  - The most well-studied proteins
  - The most popular drug targets
- Two broad classes (some have dual specificity)
  - Serine/Threonine kinases
  - Tyrosene kinases
- Protein serine/threonine phosphorylation regulates virtually every signaling pathway in the eukaryotic cell
- Tyrosine phosphorylation modulates key biological events associated with development and disease
  - cancer, diabetes, and inflammation
- Accurate annotation extremely important

Data set: Human and Mouse Protein Kinases

Gene ontology annotations (www.geneontology.org)

GO:0003674 : molecular_function (121801)
- GO:0003824 : catalytic activity (41632)
- GO:0016740 : transferase activity (13210)
- GO:0016301 : kinase activity (5613)
- GO:0004672 : protein kinase activity (3415)
- GO:0004674 : protein serine/threonine kinase activity (2077)
- GO:0004713 : protein-tyrosine kinase activity (771)

Data retrieved in 2007 [Andorf et al., BMC Bioinformatics, 2007]

Data set: Human and Mouse Protein Kinases

- Initial goal: predicting protein kinase subclasses using machine learning
- Machine learning algorithms
  - Naïve Bayes: Amino acid composition
  - NB(*k*): Extension of Naïve Bayes to *k*th order Markov model
  - SVMs using these data representations
  - A hybrid algorithm that combines the above with annotation transfer based on sequence homology (BLAST)
- Initial Question: how effective are these methods on classifying kinases?

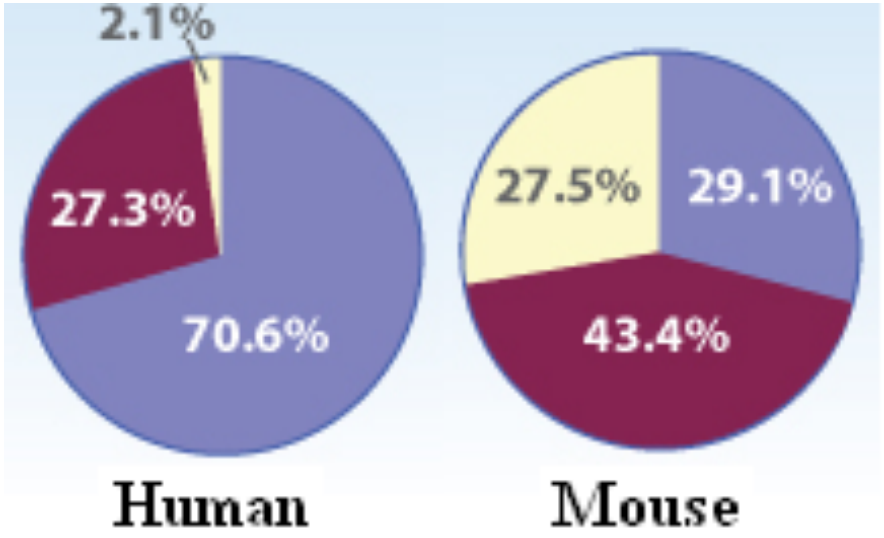
Experiment 1

- Data: 244 mouse and 330 human protein sequences
  - GO families GO0004674 (Serine/Threonine Kinase)
  - GO0004713 (Tyrosine Kinase)
- Reference class labels: annotations returned by AmiGO
  - 71 mouse and 233 human proteins are labeled with GO0004674
  - 106 mouse and 90 human proteins are labeled with GO0004713
  - 67 mouse and 7 human proteins had both labels
- Train classifier on human data and test on human data
- Train classifier on human data and test on mouse data

Experiment 1: Results

- Classifier trained on human data and tested on human data (cross validation)
  - 89.1% accuracy with a 0.85 correlation coefficient
  - Good! ☺
- Classifier trained on human data and tested on mouse data
  - 15.1% accuracy and a -0.42 correlation coefficient
  - Bad! ☹
- Result surprising because
  - Human and mouse kinases share common origin (homologues)
- Question: How can we explain these results?

First observation: Discrepancy in Distribution



Distributions of Functions in AmiGO Annotations

- Ser/Thr Kinase
- Tyr Kinase
- Dual Specificity

Second Observation: Evidence codes

- 211 of the 244 mouse protein kinases had a RCA (inferred from reviewed computational analysis) evidence code
- Of the 33 mouse proteins that did not have a RCA evidence code, 28 were classified correctly by the classifier trained on human data
- Question: What is special about the 211 mouse proteins with GO function labels with RCA evidence code?

Third Observation: Source of RCA annotations

- Annotations returned by AmiGO came from the Mouse Genome Informatics Database (MGI)
- The MGI annotations came from the Fantom2 (Functional Annotation of Mouse) Database
- Each of the 211 mouse proteins had at least one RCA from FANTOM Consortium and the RIKEN Genome Exploration Research Group (Okazaki et al, Nature, 420, 563-573, 2002)
- Are there other independent annotations for these proteins?
  - Fortunately Yes - UniProt

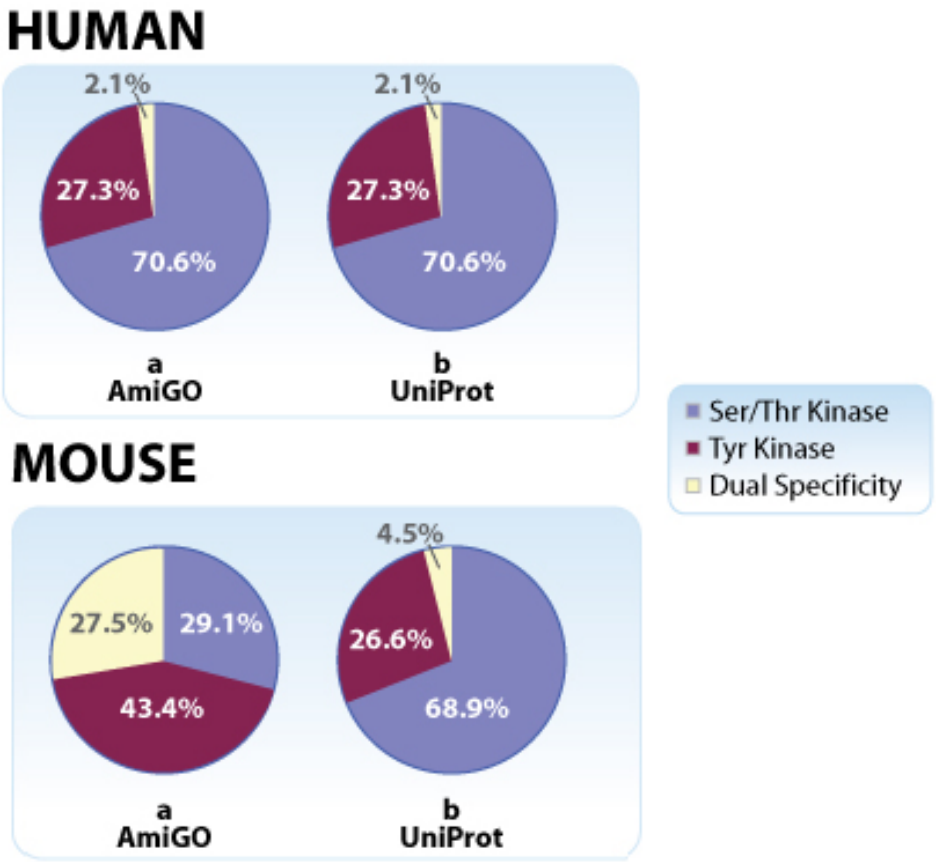
Fourth Observation: Inconsistency between UniProt and AmiGO

- AmiGO RCA annotations for 201 of the 211 mouse proteins were inconsistent with UniProt annotations

KINASE FAMILY	AmiGO Ser/Thr	AmiGO Tyr	AmiGO dual specificity
UniProt Ser/Thr	10	105	35
UniProt Tyr	54	0	3
UniProt dual specificity	0	4	0

- A search of the Mouse Kinome Database shows that 154 of the 244 mouse kinases have a human ortholog with sequence similarity greater then 90%!
- Why does machine learning fail on this problem?

Fifth Observation: Distribution of Annotations



Comparison of the Distributions of Functions in AmiGO and UniProt Annotations

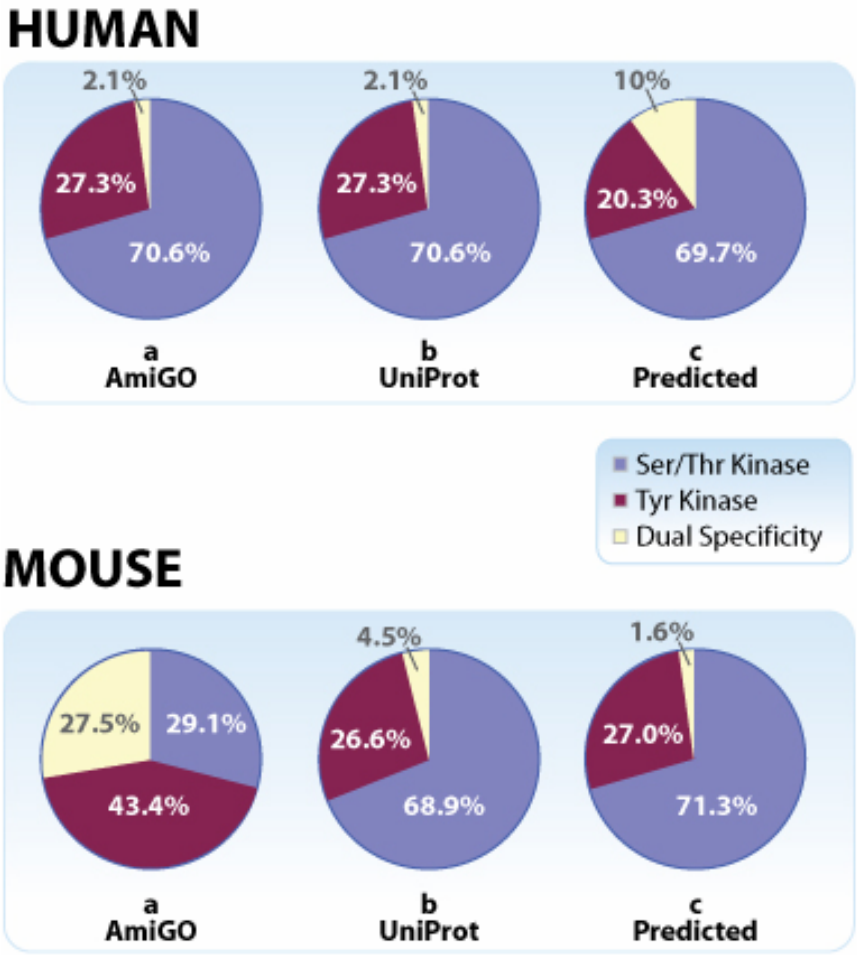
Story so far

- When reference annotations are from AmiGO:
  - Classifier trained on human kinases and tested on human kinases – good
  - Classifier trained on human kinases and tested on mouse kinases – bad
- AmiGO RCA and UniProt annotations inconsistent
- Questions:
  - Could the AmiGO RCA annotations be incorrect?
  - How does the classifier trained on human and tested on mouse perform when the reference annotations are from UniProt?

Experiment II

- Use UniProt labels instead of AmiGO labels as reference
- Train classifier using Human proteins and test on mouse proteins
- Test accuracy on mouse proteins: 97%!
- 205 of the 211 proteins that were mislabeled with respect to AmiGO reference labels were correctly labeled with respect to UniProt reference labels

Sixth Observation: Distribution of Annotations



Comparison of the Distributions of Functions in AmiGO, UniProt, and Predicted Annotations

Tentative Conclusions

- There is no reason to expect that the relative distribution of the Ser/Thr kinsases and Tyr kinases in human and mouse would be dissimilar
- The machine learning approach used is sound, and found effective in other macromolecular sequence classification tasks
- Could it be the case that the annotations returned by AmiGO for the 211 mouse protein kinases (nearly 95% of the 244 mouse protein kinases) are incorrect?

Following up

- To the best of our knowledge, the problematic mouse kinase annotations with RCA evidence code
  - Came from Okazaki et al, Nature, 420, 563-573, 2002
  - Were propagated to MGI through the Fantom2 (Functional Annotation of Mouse) Database
  - And from MGI to AmiGO
- Examination of GO annotation is often the first step in many high throughput studies e.g., gene expression analysis
- Question: How far did these annotations propagate?

Following up

- 136 rat protein kinase annotations from AmiGO had:
  - ISS - inferred based on sequence or structural similarity- evidence code
  - Functions assigned based on some of the 201 potentially incorrectly annotated mouse proteins
    - 94 Ser/Thr kinase proteins mislabeled as either a Tyr kinase or dual specific
    - 42 Tyr kinase proteins mislabeled as a Ser/Thr kinase or a dual specific
- 201 mouse and 136 rat protein kinase annotations are probably incorrect!
- Not to mention annotations of other kinases and analyses that relied on these erroneous annotations!

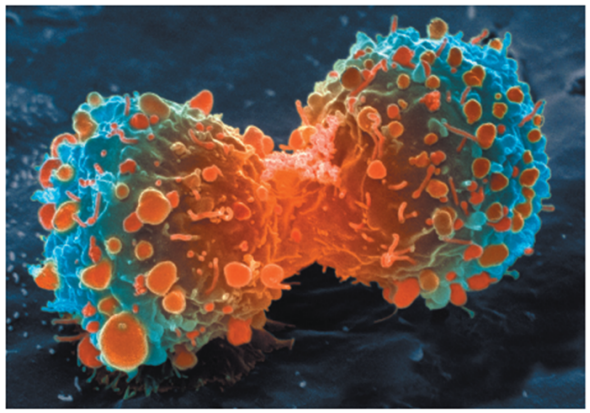
Conclusions: Detecting Annotation Errors Using Machine Learning

- The apparent *failure* of a machine learning approach helped us discover potential errors in annotations
- Our discovery further underscores the need for better procedures for
  - Multiple checks for consistency of annotations – especially in the case of annotations with RCA and ISS evidence codes
  - Better methods for tracking propagation of annotations across databases
  - Reproducible (and correctible) computational workflows
- The erroneous mouse kinase annotations were traced to errors in annotation scripts used and have since been fixed by the MGI ☺

Reproducibility crisis

COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will

investigators must reassess their approach translating discovery research into gre clinical success and impact.

Many factors are responsible for the h failure rate, notwithstanding the inh ently difficult nature of this disease. C tainly, the limitations of preclinical to

47/53 "landmark" publications could not be replicated

[Begley, Ellis Nature, 483, 2012]

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Error prone

Biologists must realize the pitfalls of work on massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues Jonathan F. Russell.

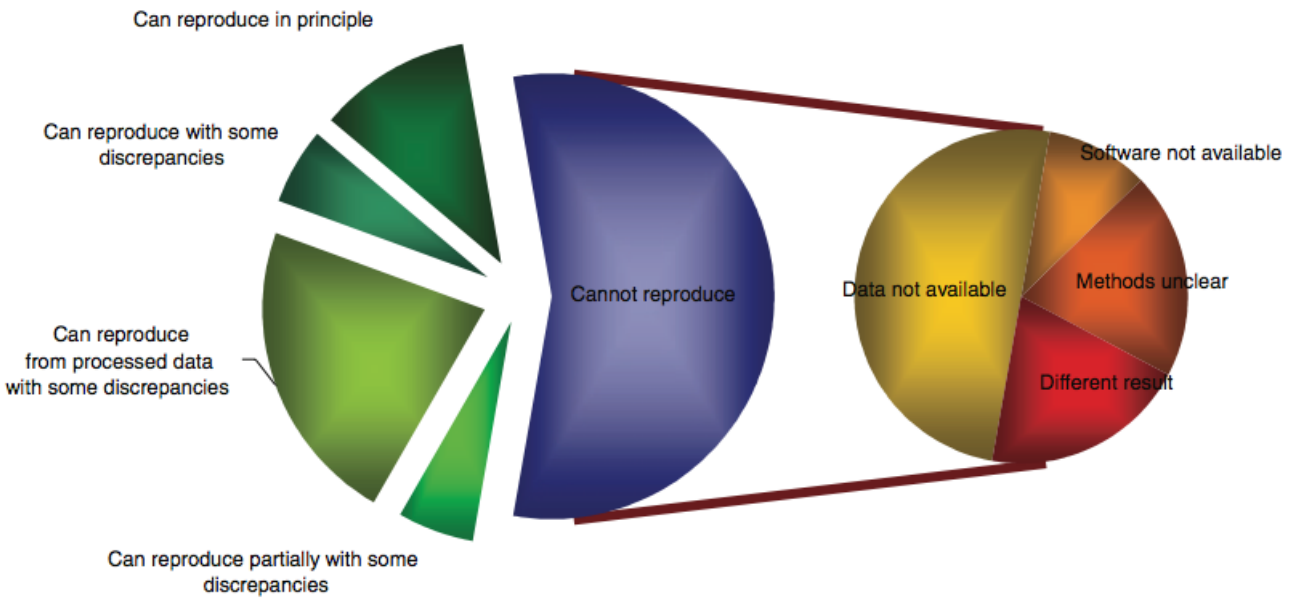
The case for open computer programs

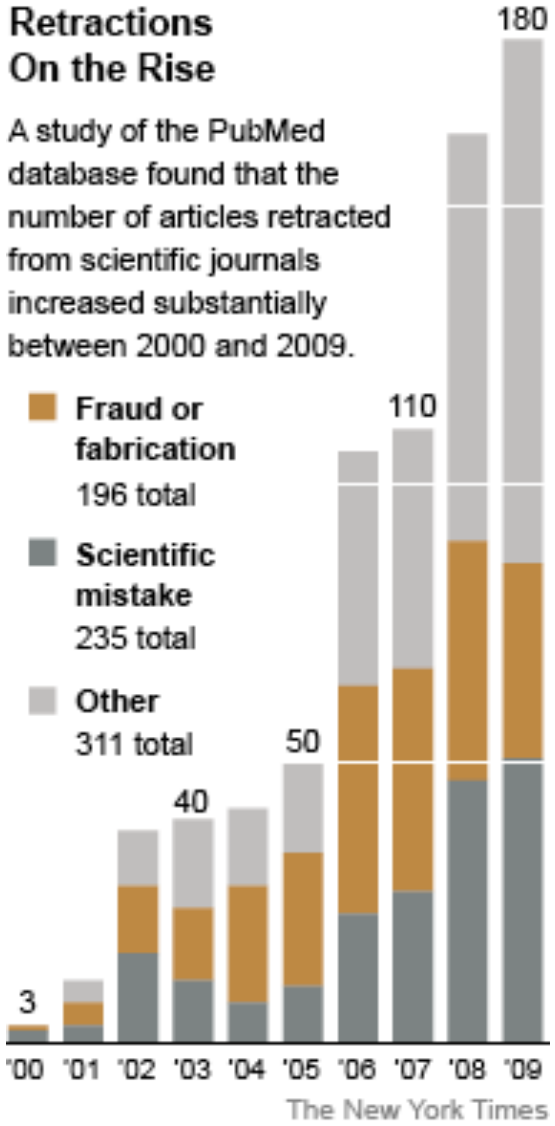
Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won't stand up.

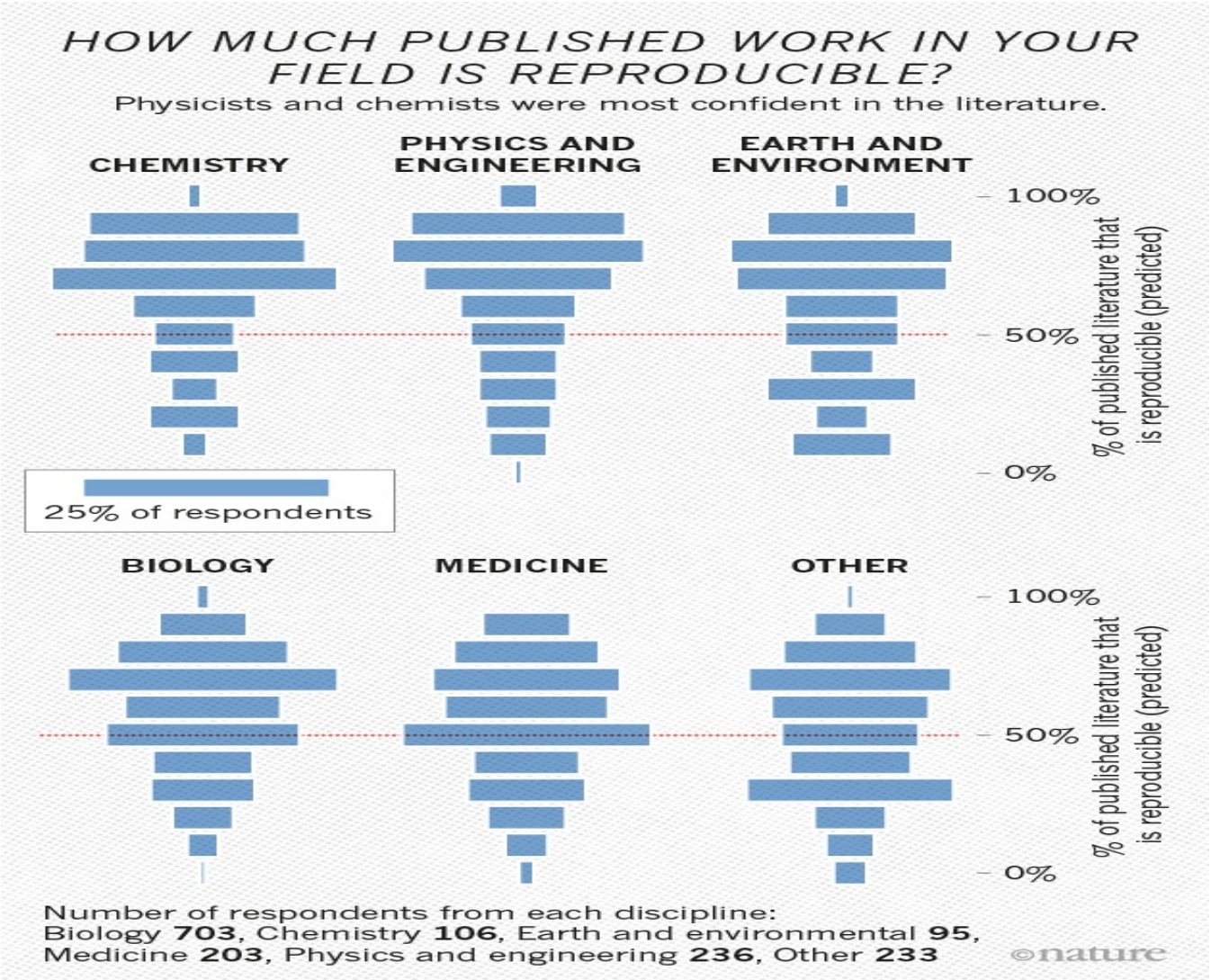
Know when your numbers are significant

Reproducibility crisis



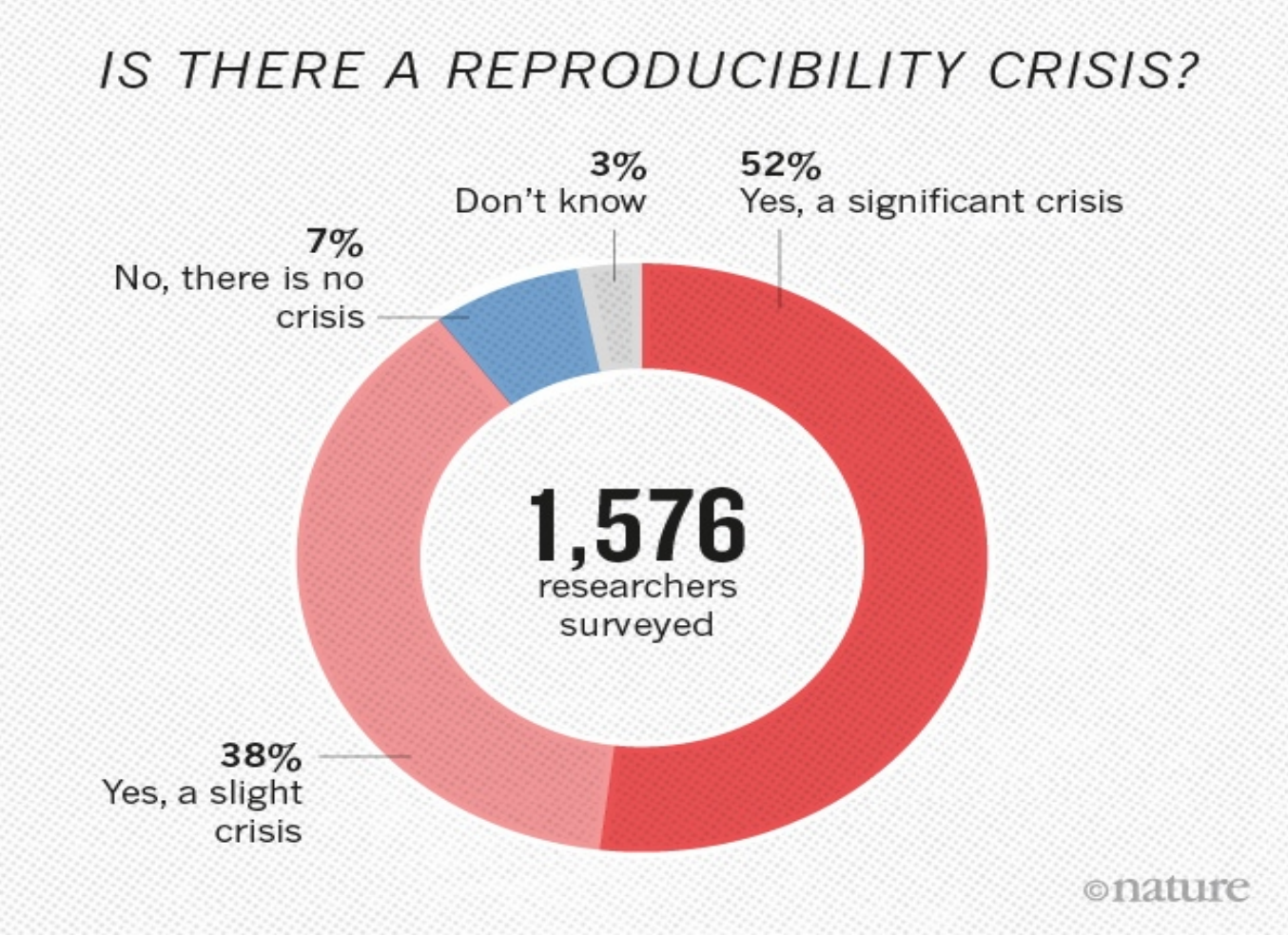


- More retractions:
- >15x increase in the last 10 years
- At current rate, by 2045 as many papers published as retracted



*

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature,533(7604), 452-454.



*

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature,533(7604), 452-454.

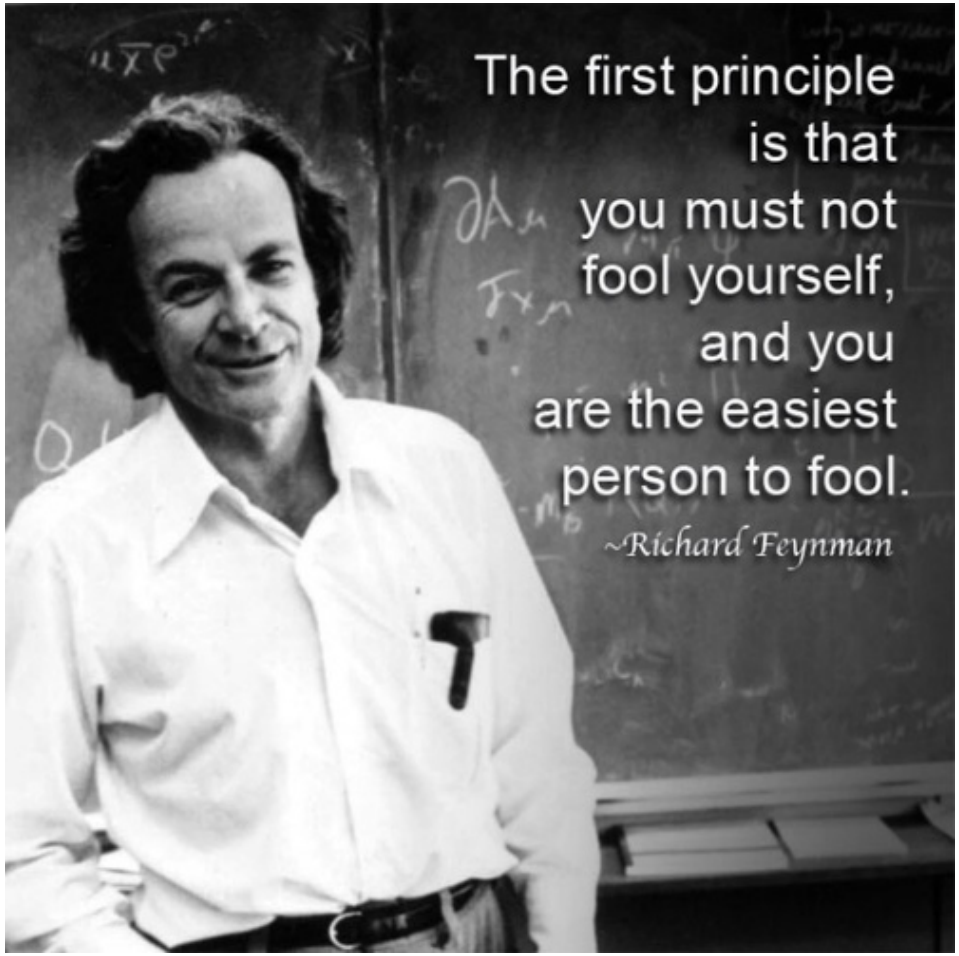
On scientific practice

- Science is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into testable laws and theories.
- The success and credibility of science are anchored in the willingness of scientists independent testing and replication by other scientists. This requires the complete and open exchange of data, procedures and materials.

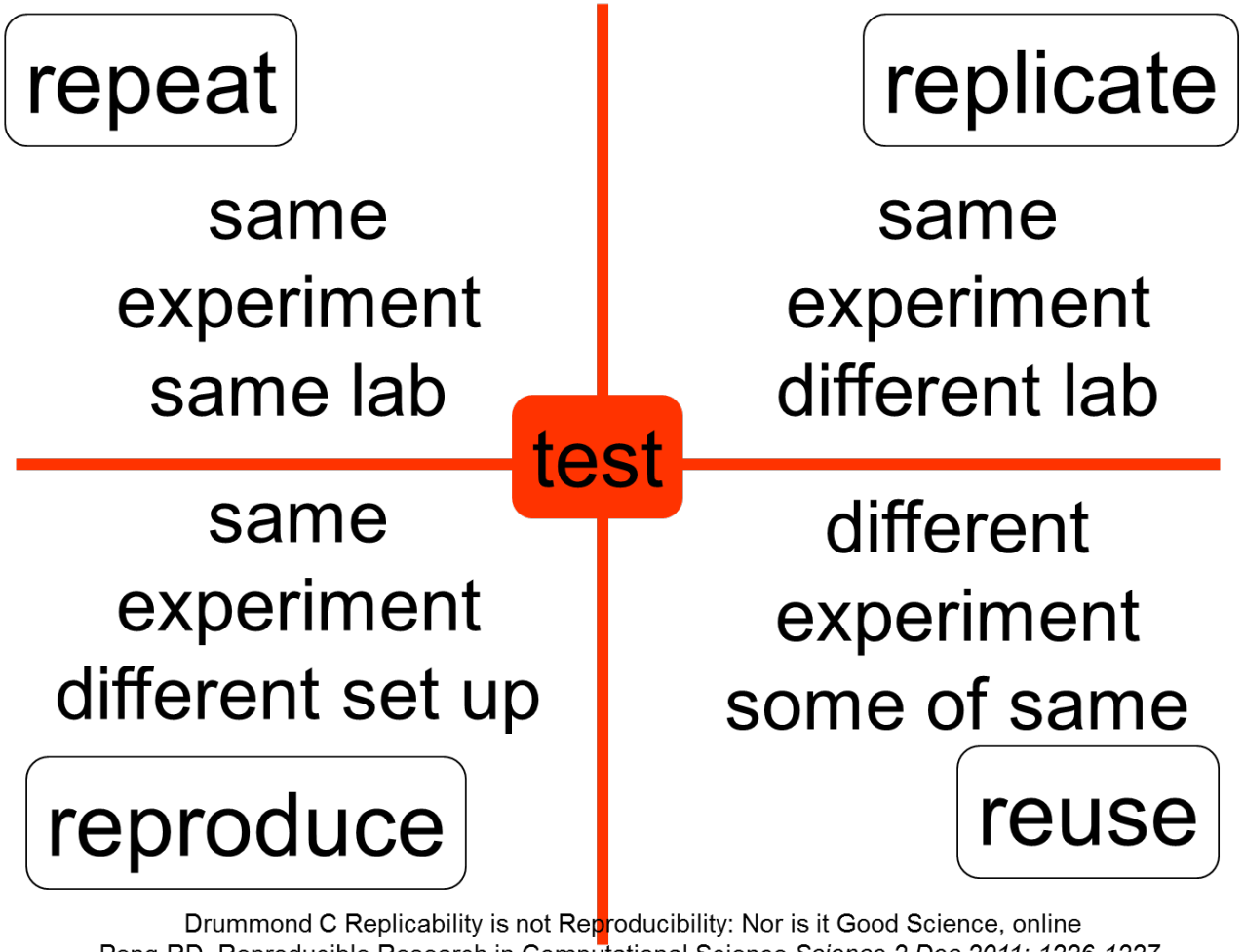
Scientific method

- Empirically testable
- Replicable
- Objective
- Transparent
- Falsifiable
- Logically consistent

On scientific practice



Pillars of the scientific method



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

Pillars of the scientific method

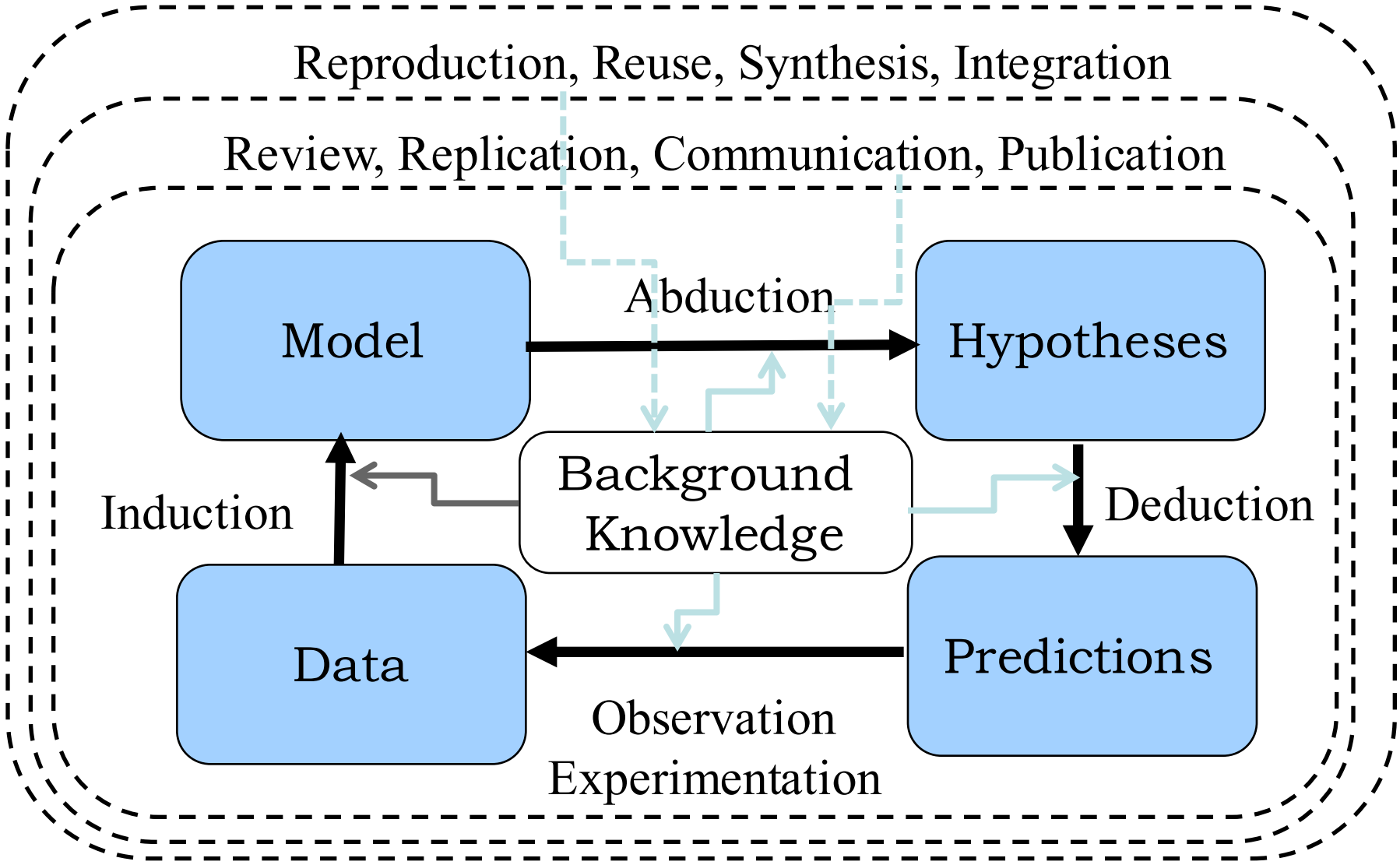
- **Replication** – independent researchers going out and collecting new data to verify scientific findings – considered the scientific gold standard.
- **Reproduction** – independent researchers analyze the same data and produce the same result. Focus on transparency of data analysis.

Peng, Roger D. (2011) "Reproducible Research in Computational Science." Science 334.6060: 1226–1227.

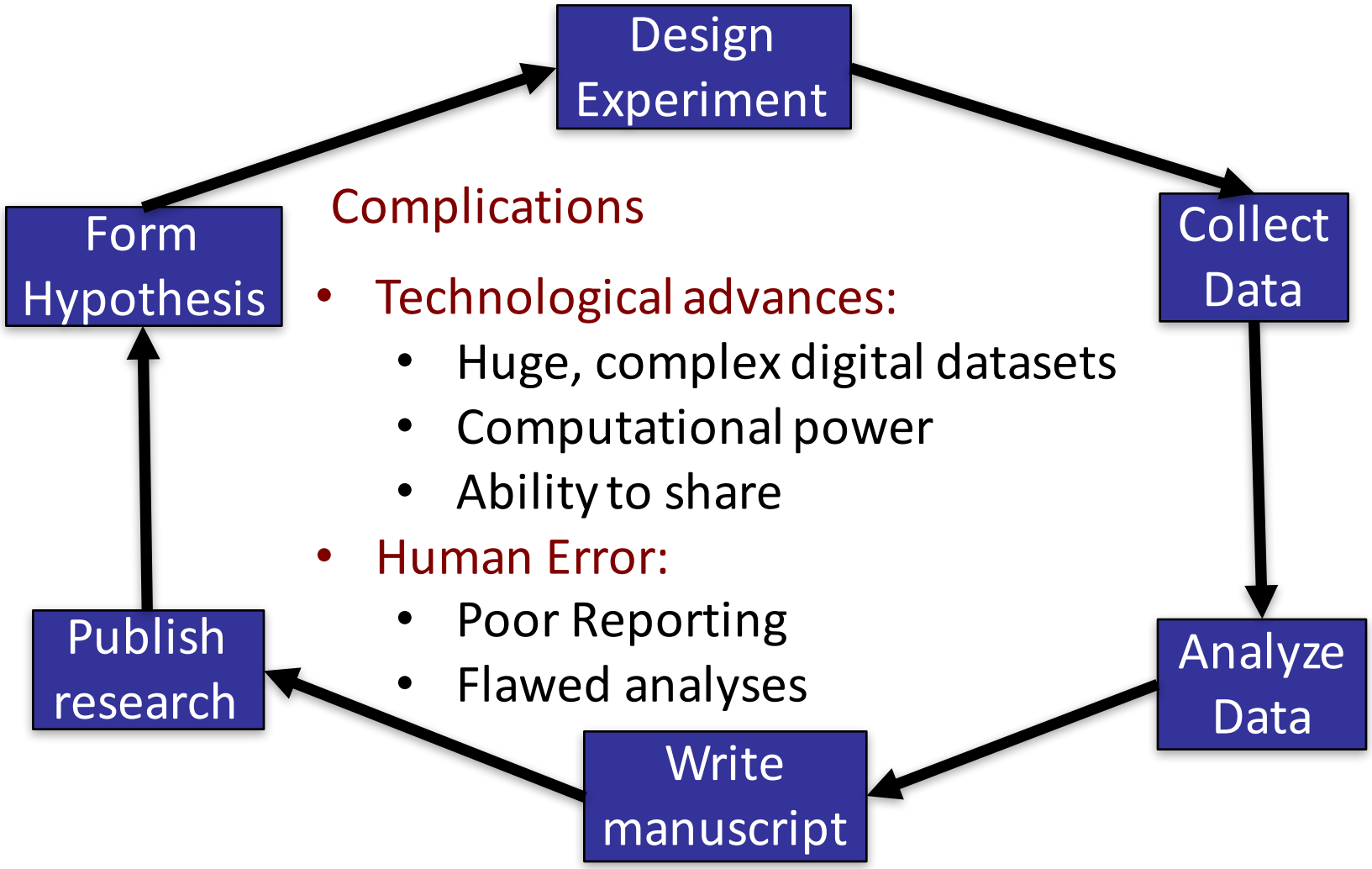
Many dimensions

- Samples
- Measurements
- Experiment design
- Data
- Statistical considerations
- Analysis
- Scientific misconduct

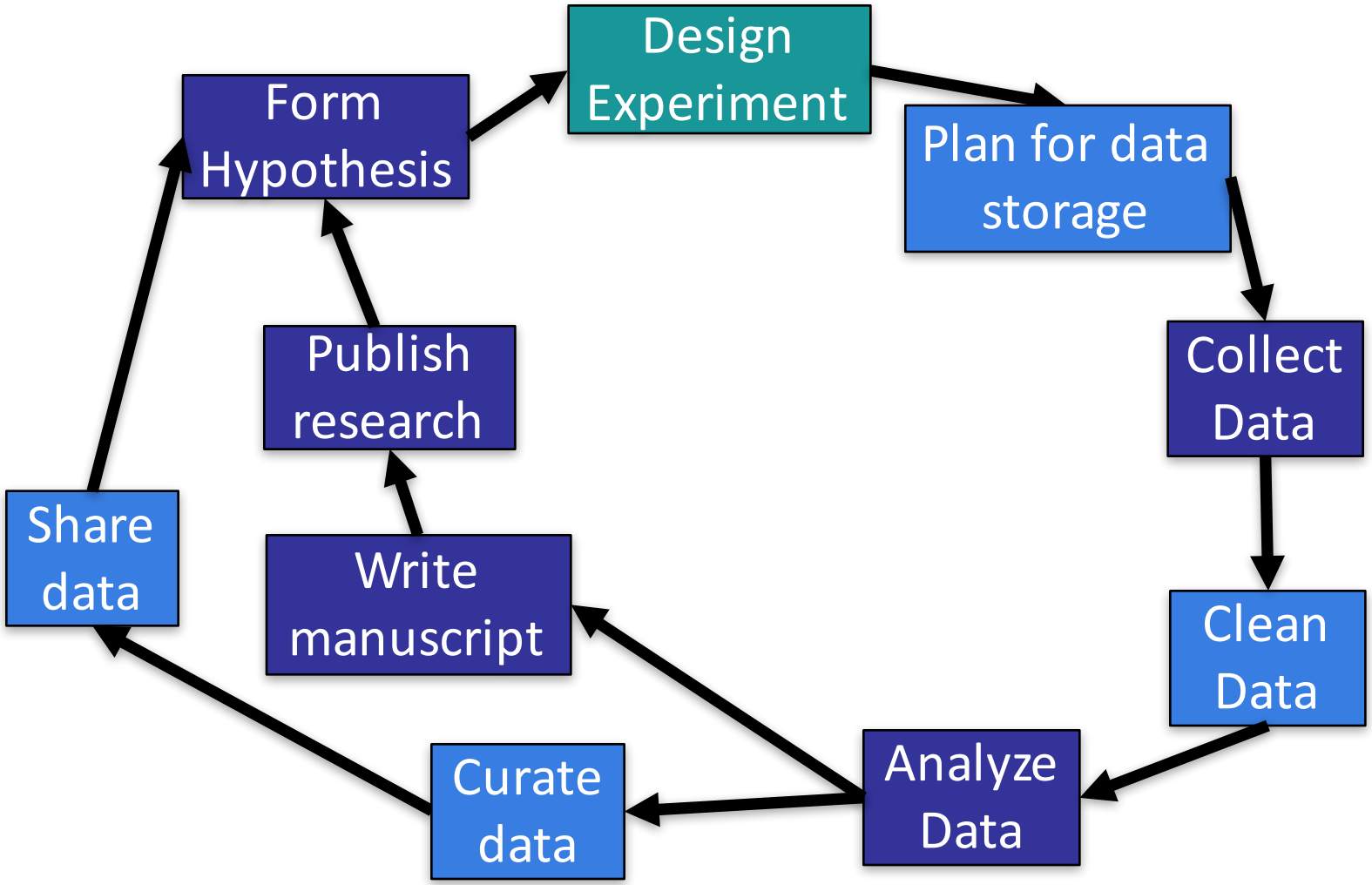
Science as we know it



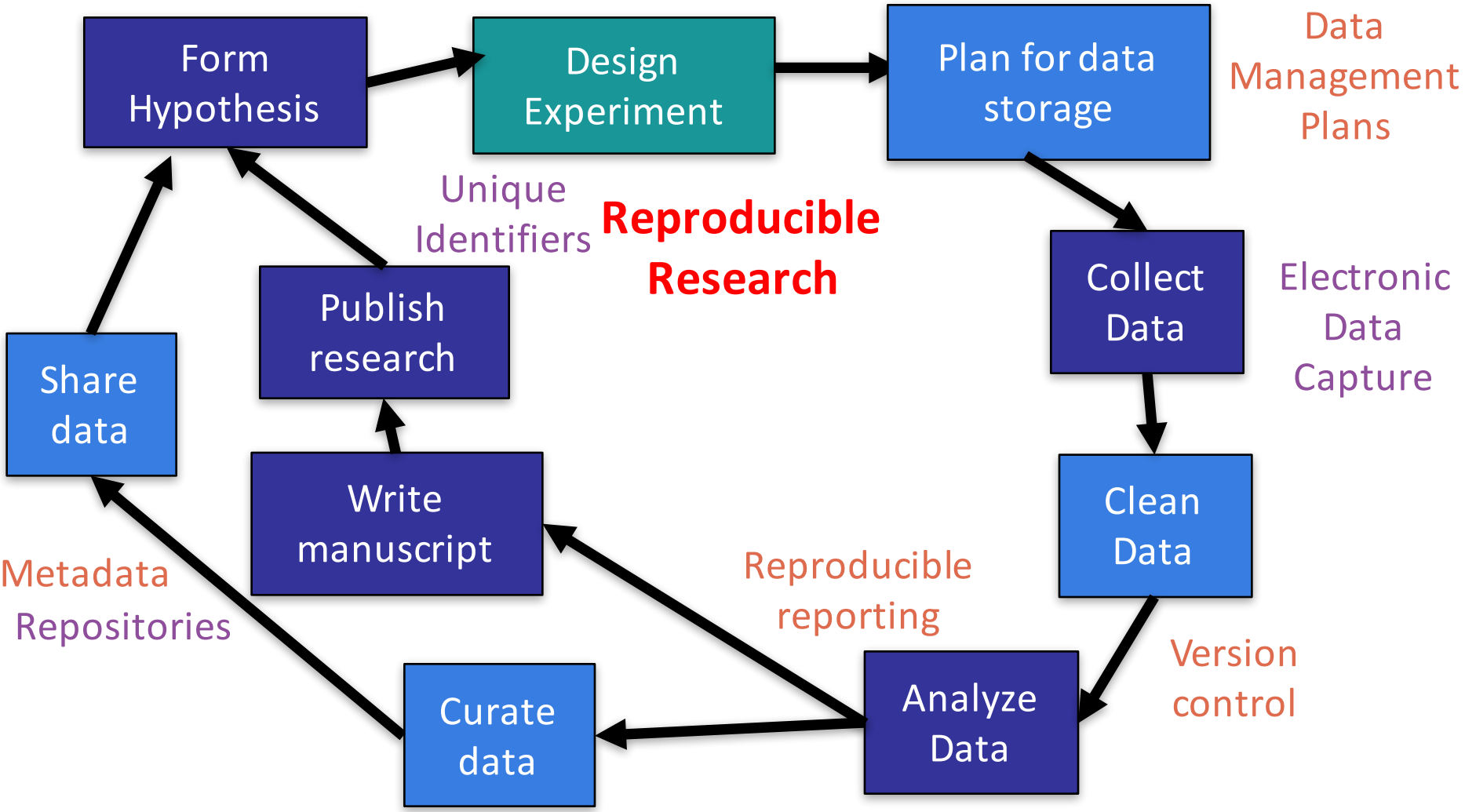
Traditional scientific research cycle



Modern scientific research cycle



Reproducible scientific research cycle



Reproducible Research: Relation to scientific method

Steps of a scientific method:

- Define a question
- Form an explanatory hypothesis
- Test the hypothesis by performing an experiment and collecting data in a reproducible manner
- Analyze the data
- Interpret the data and draw a conclusion
- Publish results
- Validate (reproduce) against the findings of other researchers

Crawford S, Stucki L (1990), "Peer review and the changing research record", "J Am Soc Info Science", vol. 41, pp. 223–228

The steps related to the Reproducible Research are in red

Reproducible Computational Research

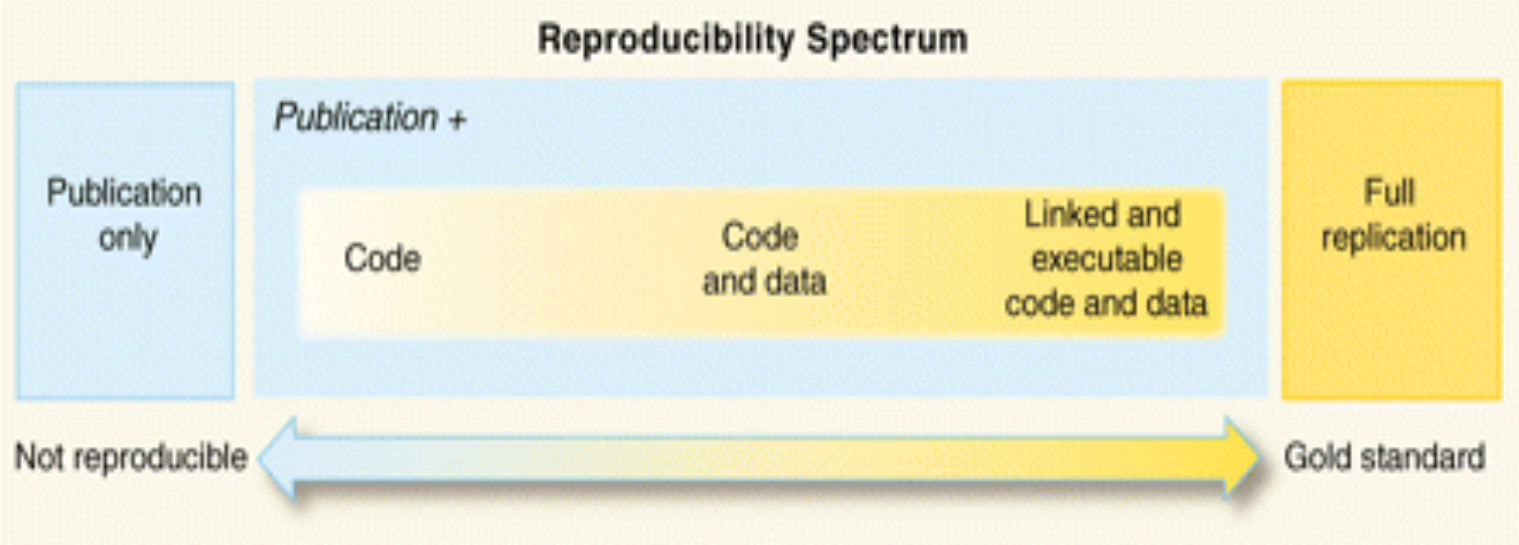
- Reproducible Research (RR) aims to complement scientific articles with everything required to independently reproduce the results described therein:
- Everything includes:
  - Experimental design
  - Data
  - Analysis workflow
  - Computer codes
  - A precise description of how the code was applied to the data (parameter choices etc.)

Delescluse, Matthieu, et al. "Making neurophysiological data analysis reproducible: Why and how?" Journal of Physiology-Paris 106.3 (2012):159-170.

Reproducibility Spectrum

"The published paper is only an advertisement of the scholarship; it is not the scholarship itself. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

Jonathan Buckheit and David Donoho, paraphrasing Jon Claerbout



*

Peng, Roger D. (2011) "Reproducible Research in Computational Science." Science 334.6060: 1226–1227.

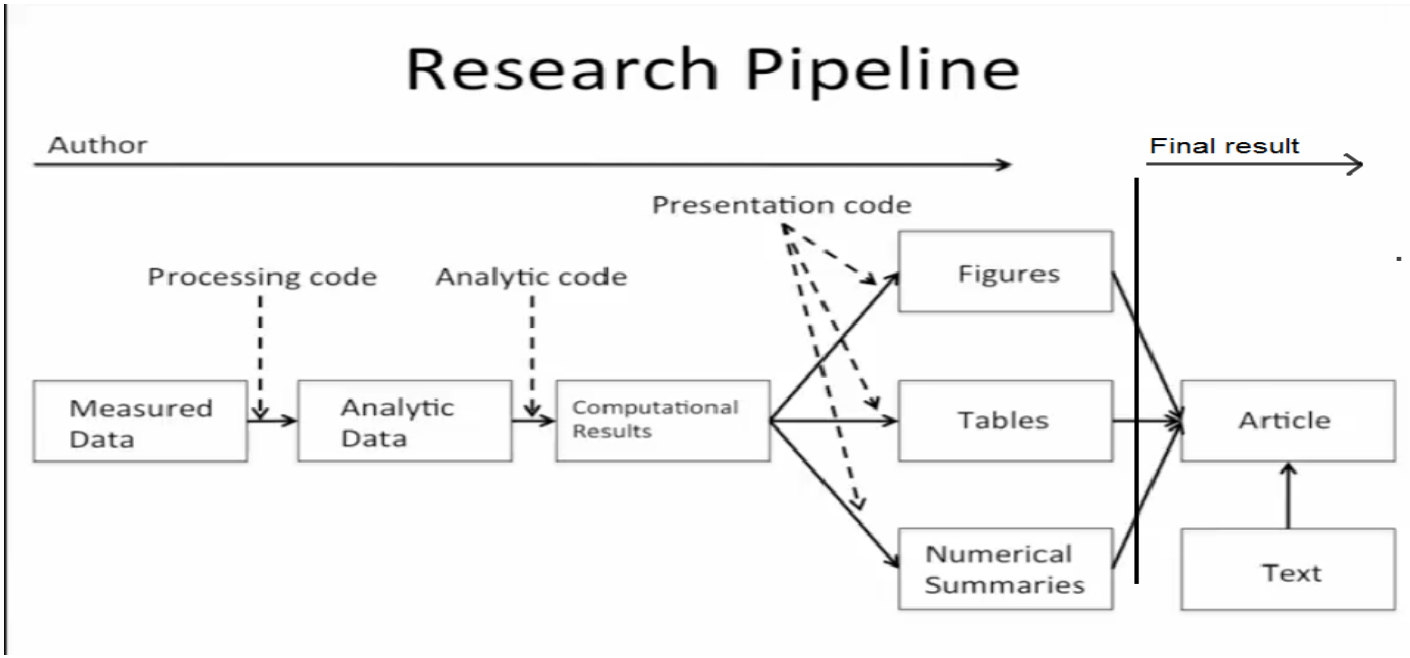
Basic criteria for reproducibility

A study is reproducible if

- The experiment design is fully described
- The data are made available
- The statistical methods are fully described and computer code is made available
- Documentation for both data and methods made available; and
- Standard methods of distribution are employed.

Reproducibility is especially important for studies that have low probability of full replication within the relevant timeframe for a variety of reasons

Reproducible research pipeline



Source: http://www.biostat.jhsph.edu/~rpeng/research.html

Challenges to reproducibility

- High throughput measurement technologies
- Emergence of big data
- Secondary use of research data in contexts other than originally intended
- Integrative analyses of disparate data of uncertain quality and provenance
- Tools of uncertain quality
- Low Signal-to-noise
- Hyper-competitive research environment

Reproducibility crisis: A computational perspective

- Reluctance to make data available
- Poor documentation of data
- Effort needed to regenerate data from description provided
- Lack of automation (e.g., interactive use of excel spreadsheets in data analysis)
- Poor scientific practices
  - Incomplete description of key steps (e.g., for generating a plot)
  - data dredging (p hacking)
  - failure to report negative finding
  - inclusion/exclusion of samples to obtain a desired conclusion
    ….

Reproducibility crisis: A computational perspective

- Lack of availability of code:
  - Lack of transparency
  - Reluctance to make codes public
  - Additional time and efforts to improve code quality, documentation, etc.
  - Poor software engineering skills
- Poor quality code
  - Errors
  - Poor software engineering practices
  - Poor coding style
  - Poor documentation

Reproducibility crisis: A computational perspective

- Poorly described data processing steps
  - Normalization
  - Feature engineering
  - Feature selection
  - Dimensionality reduction
- Poorly described computational experiments
  - Lack of details about user specified parameters
  - Random number used
  - Training and test sets
  - Performance measures

Reproducibility crisis: A computational perspective

- Infrastructure
  - Data resources
  - Computational resources
  - Dependence on libraries
  - Dependence on Operating System
  - Dependence on system conditions (load, etc.)

Current status: Reproducible research in machine learning

- Increasing availability of benchmark data
- Comparison of a new machine learning method against competing methods mandatory
- Publication of code increasingly mandatory
- Increasing automation of computational experiments
- Increasing pressure on authors for transparency (parameter tuning, data pre-processing, etc.)

Current status: Reproducible research in biostatistics

Authors should provide all data code in order to reproduce all results, images and tables with:

- README file
- Consistent coding style and documentation
- Test data sets
- Simulations and random numbers
- General advice

Peng, R. D. (2009). Reproducible research and biostatistics. Biostatistics,10(3), 405-408.

Reproducibility delayed is reproducibility denied

- Making a research project reproducible at the completion of the project, just as documenting software after you are done writing code, is not the best way
- Document everything as you go
- This means
  - Data
  - Experimental details
  - Code
  - Preprocessing steps
  - Parameters
  - ….

Reproducible computational research: Personal incentives

- Published findings can be verified
- Alternative analyses conducted
- Challenge uninformed criticisms (“put up or shut up”)
- Expedite exchange of ideas among investigators
- Better visibility of research
- More citations and higher impact
- Increased trust in research quality (outside academia, e.g., industry, public)
- Engaging the broader community in improving data, methods, tools… (analogous to open source software)

Reproducible computational research: Tools

Recommended programs to use to achieve reproducibility:

- Latex (Tex editor)
- Version control systems - Git software systems
- Make – pipeline

Literate programming concept (Knuth).

Reproducible computational research: Tools

- Matlab not recommended
  - Proprietary toolboxes
- Open source alternatives
  - Octave
  - Scilab
  - Sagemath

Reproducible computational research: Tools

R programming language:

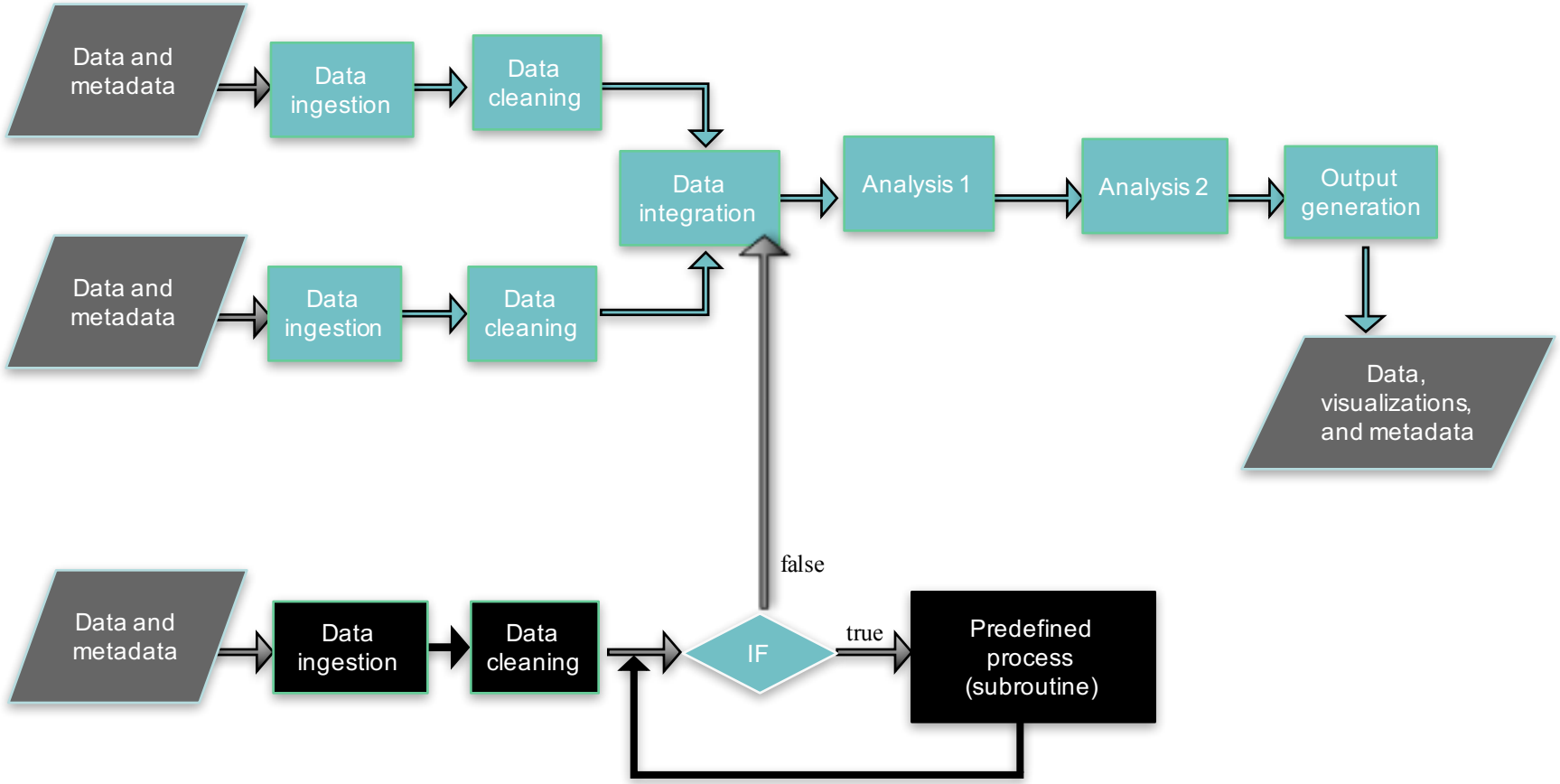
- R studio – development environment for R programming language
- Graphic packages, such as ggplot2
- Packages as knitr or rmarkdown – literate programming support

Reproducible computational research: Tools

Python programming language:

- Many open scientific libraries available – scipy, numpy, etc.
- IPython notebook
- Sumatra package – save parameter values, code state, output results and files

Reproducible computational research: Tools



Reproducible computational research: Tools

Computational workflows in which

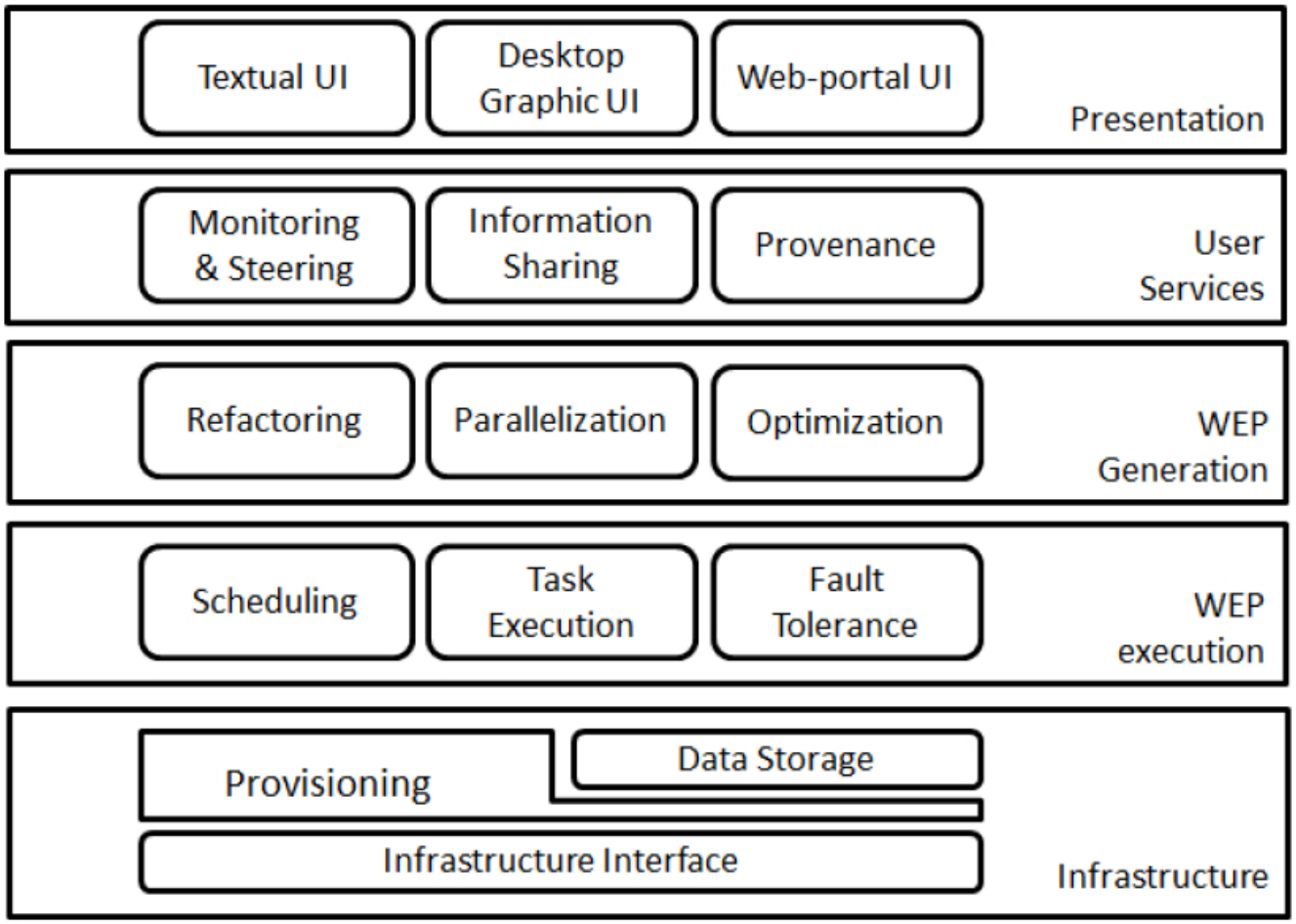
- Each step can be implemented using programming language of one's choice
- The inputs and parameters of each step are formally recorded
- Each step can be executed on appropriate computational platform
- It is possible to reuse of individual steps as well as the overall workflow

Reproducible computational research: Tools

Workflows provide

- Abstraction and encapsulation
- Ease of use
- Single access point for multiple analyses
- Reproducibility
- Reuse and adaptation
- Documentation
- Training

Scientific Workflow Management Systems



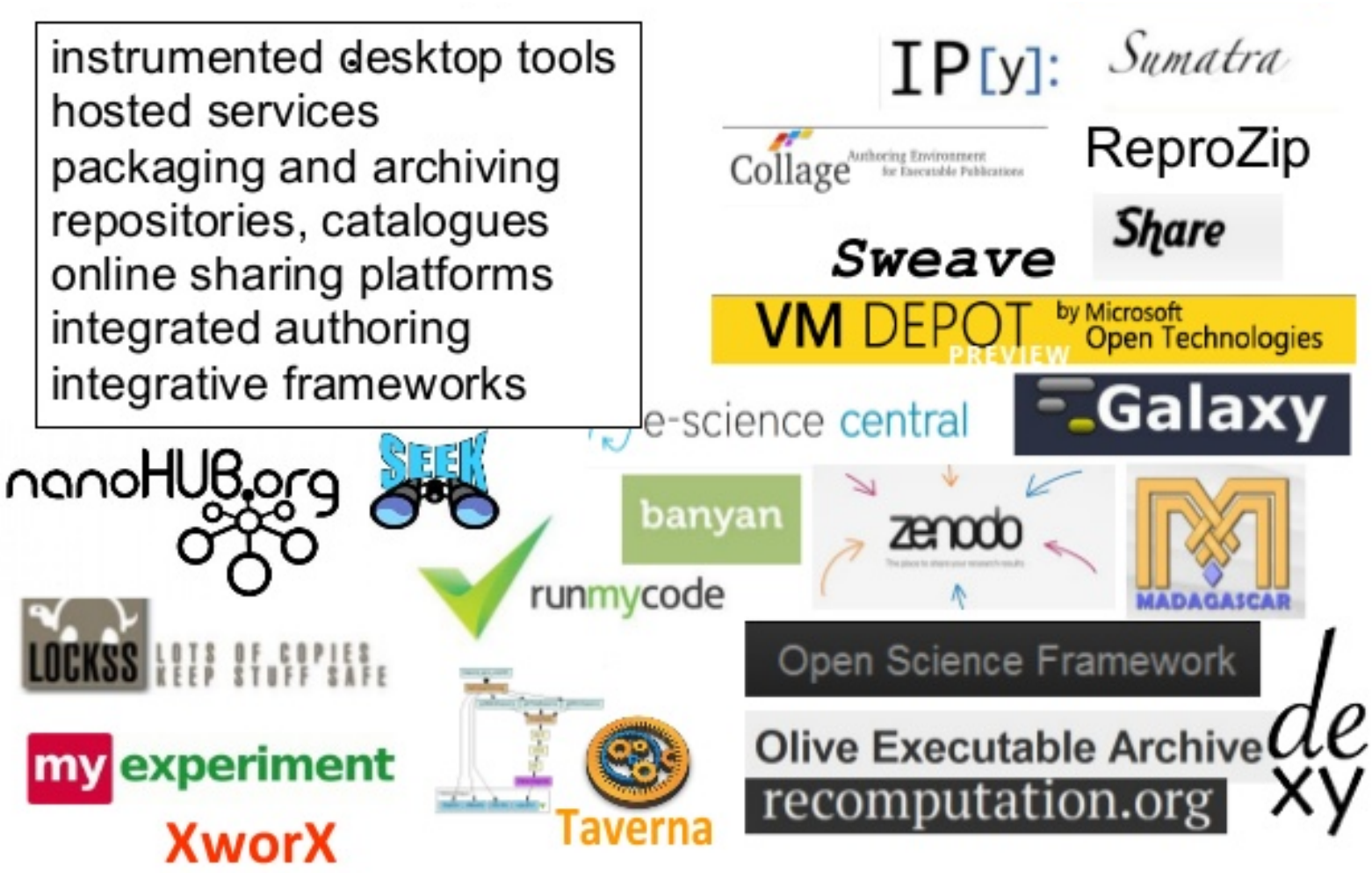
Reproducible computational research: Tools

Several mature scientific workflow systems available:

- Kepler (built on top of Ptolemy)
- Pegasus
- Taverna
- Galaxy (initially focused on genomics)
- and many more….

Reproducible computational research: Tools

instrumented desktop tools
hosted services
packaging and archiving
repositories, catalogues
online sharing platforms
integrated authoring
integrative frameworks



*

Reproducible Computational Research: Resources

Courses:

- Data science specialization (www.coursera.org) (John Hopkins University) – course 5 Reproducible research
- Research Methods: An Engineering Approach (www.edx.org) (Wits University)
- Research Data Management and Sharing (www.coursera.org) (The University of North Carolina at Chapel Hill & The University of Edinburgh)

Reproducible Computational Research: Resources

Software tools for RR:

- Software carpentry (www.Software-carpentry.org) – basic computing skills for researchers
- Bootcamps - one or two day long courses – teaching coding and professional skills for researchers – like the one that you are attending
- Courses - www.coursera.org, www.edx.org, www.udacity.org - for programming skills in R, Python, etc.

Reproducible Computational Research: Resources

Books:

- Stodden, V., Leisch, F., & Peng, R. D. (Eds.) (2014). *Implementing Reproducible Research*. CRC Press
- Gandrud, C. (2013). *Reproducible Research with R and R Studio*. CRC Press
- Subramanian, G. (2015). *Python Data Science Cookbook*. Packt Publishing Ltd.

Thank you!