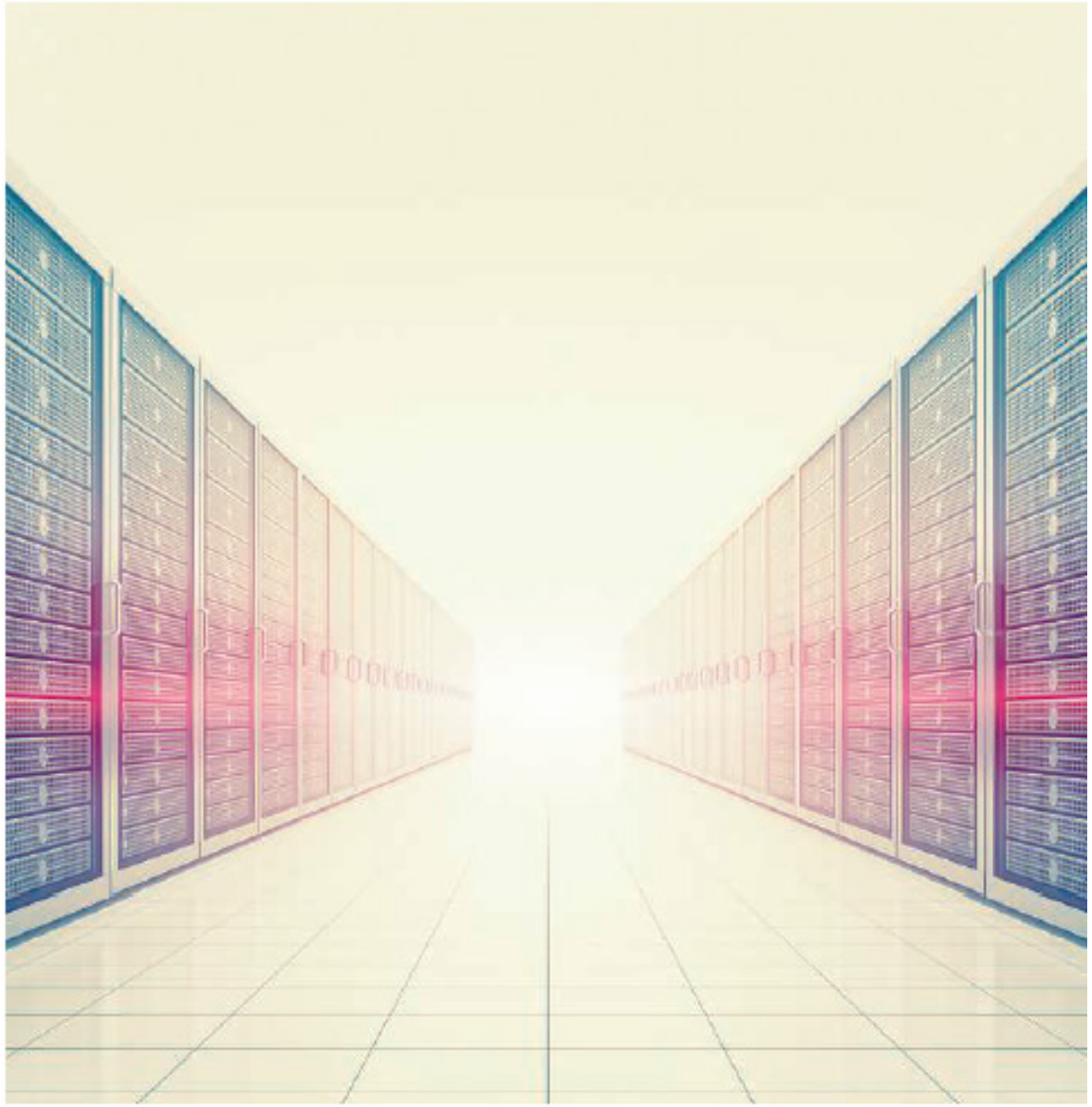
Visualizing data integrity and management through the lens of transparency

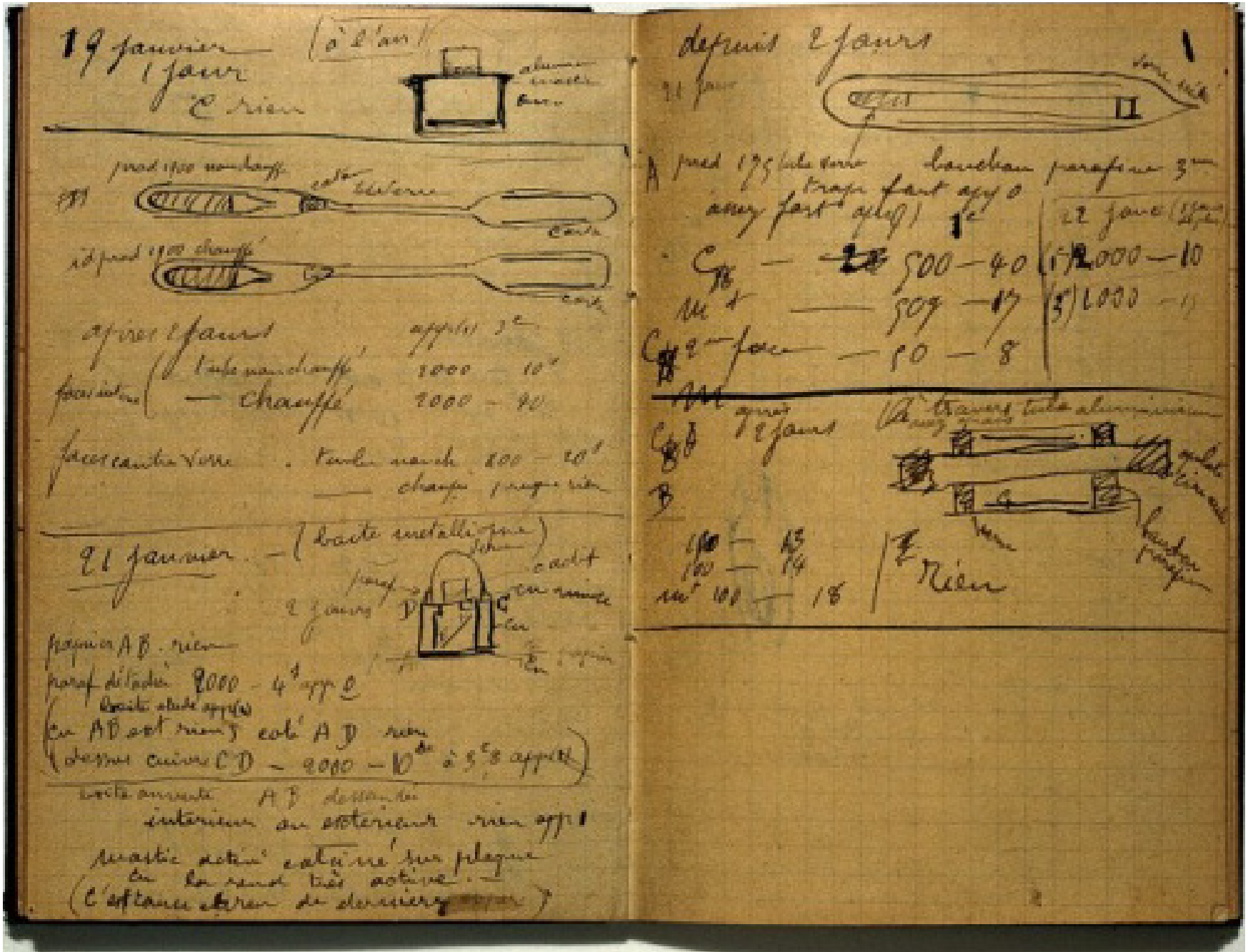
C. S. Raman
UMB School of Pharmacy
15 September 2016



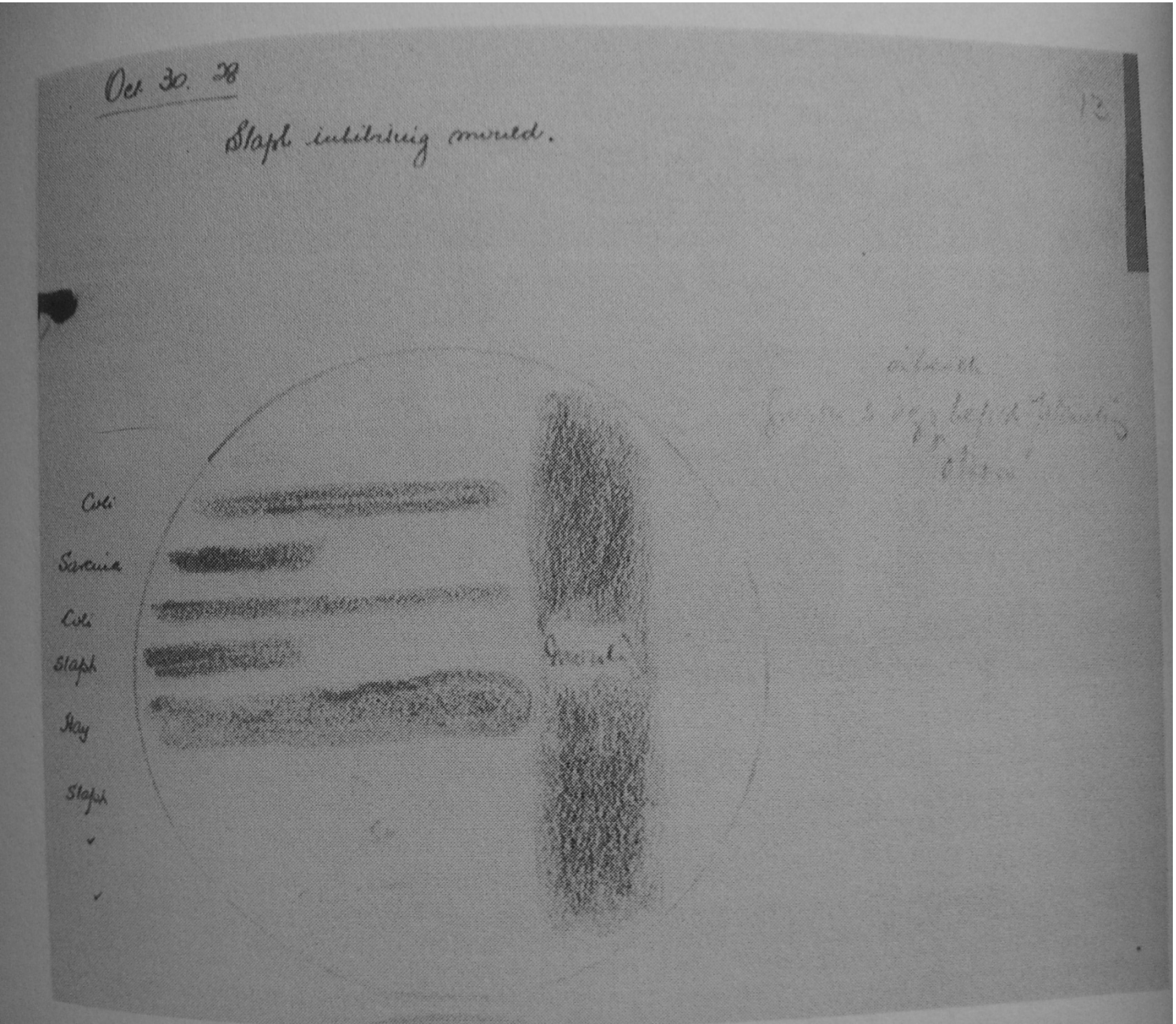
"...interpretations come and go, but data are forever."

Marcia McNutt
Editor-in-Chief
Science *Journals*

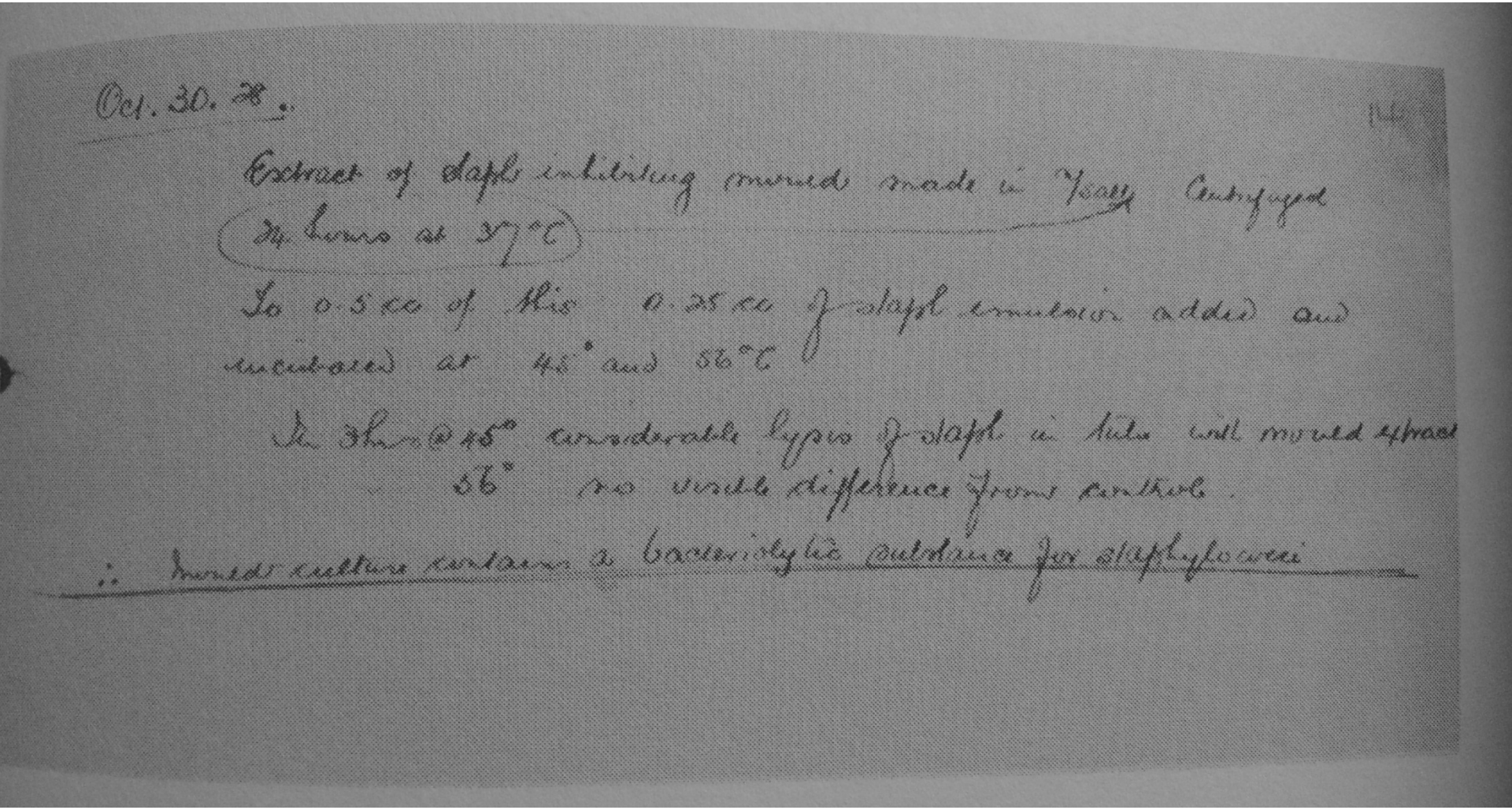
Marie Curie's radioactive notebook (1899 – 1902)



Alexander Fleming's data recordings: 1928



Alexander Fleming's data recordings: 1928



Oct. 30. 28.

Extract of staph inhibiting mould made in Yeast Autolysate 24 hours at 37°C

To 0.5 cc of this 0.25 cc of staph emulsion added and incubated at 45° and 56°C

At 3 hrs @ 45° considerable lysis of staph in tube with mould extract 56° no visible difference from control.

∴ Mould culture contains a bacteriolytic substance for staphylococci

Life is mostly composed of the elements carbon, hydrogen, nitrogen, oxygen, sulfur, and phosphorus. Although these six elements form the backbone nucleic acids, proteins, and lipids and thus the bulk of living matter, it is theoretically conceivable that some other elements in the periodic table could serve the same functions. Here we describe a bacterium, strain GFAJ-1 of the Halomonadaceae, isolated from Mono Lake, California, that is able to substitute arsenic for phosphorus to sustain its growth. Our data show evidence for arsenate in macromolecules that normally contain phosphate, most notably nucleic acids and proteins. Exchange of one of the major bio-elements may have profound evolutionary and geochemical importance.

RETRACTED

SCIE

A Bacterium ... Arsenic Inste...

Felisa Wolfe-Simon,[1,2]* Jodi Sw...
Shelley E. Hoeft,[2] Jennifer Pett...
Paul C. W. Davies,[1,7] Ariel D. A...

NATURE

Methods: Face up to false positives

Daniel MacArthur

Nature **487**, 427–428 (2012) doi:10.1038/487427a

'Scientists and journals must work together to ensure that eye-catching artefacts are not trumpeted as genomic insights'

'hunting for biological surprises without due caution can easily yield a rich crop of biases and experimental artefacts, and lead to high-impact papers built on nothing more than systematic experimental 'noise'.'

Science

AAAS

12 December 2014

RETRACTED

When conta[illegible]anges minds: An experi[illegible]t on transmission of supp[illegible]or gay equality

Michael [illegible] and Donald P. Green[2]

Ca[illegible]gle (20 minute) conversation change minds on d[illegible]sive social issues, such as same-sex marriage?

"LaCour has not produced the original survey data from which someone else could independently confirm the validity of the reported findings." --Marcia McNutt, Editor-in-Chief, May 2015

Global data integrity crisis

nature 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.

Believe it or not: how much can we rely on published data on potential drug targets?

See also: News and Analysis by Arrowsmith

Florian Prinz[1], Thomas Schlange[2] & Khusru Asadullah[3]

Bayer Healthcare

Global data integrity crisis

"…we will soon issue an international good practice for regulatory authorities and inspectors that can help reduce incidents of incomplete presentation of data by manufacturers or deliberate data falsification."

March 2015

Dr Margaret Chan, Director-General, World Health Organization

https://goo.gl/LHQKdW

FDA's draft guidance on data integrity

https://goo.gl/4sAhPR

April 2016

Theater makeup artist turned image forensics expert

Jana Christopher

Ensuring data integrity in science

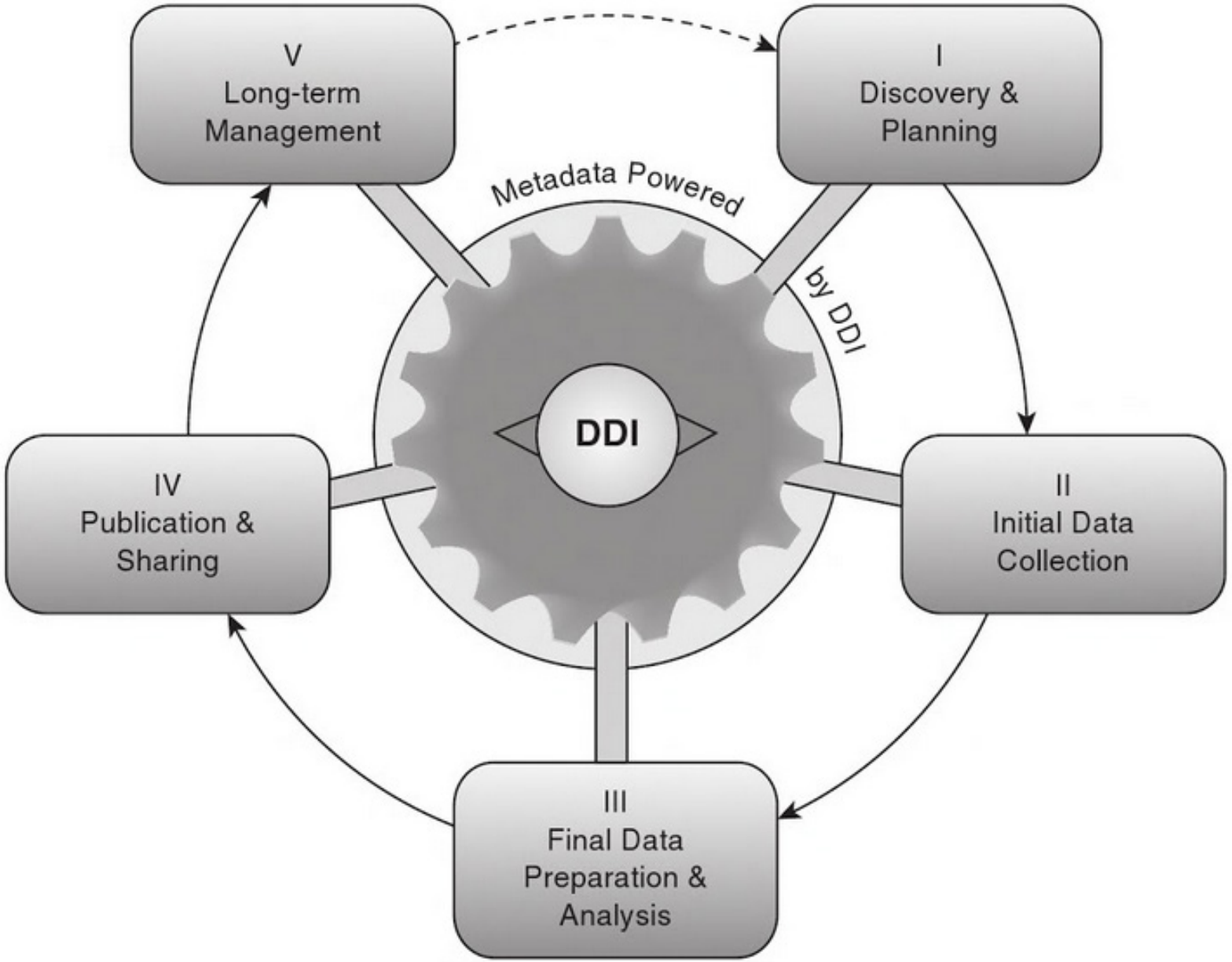
Framework proposed by Committee on Science, Engineering, and Public Policy (COSEPUP) of the National Academy of Sciences (NAS), National Academy of Engineering, and Institute of Medicine (2010).

- Urgent attention required on two fronts:
  (a) "general practice of science" and
  (b) "personal behaviors of scientists"

- "Reinforce clarity and transparency to build and maintain trust in science."

- "Teach scientists to describe experiments, data, and calculations fully so that other scientists can replicate the research"

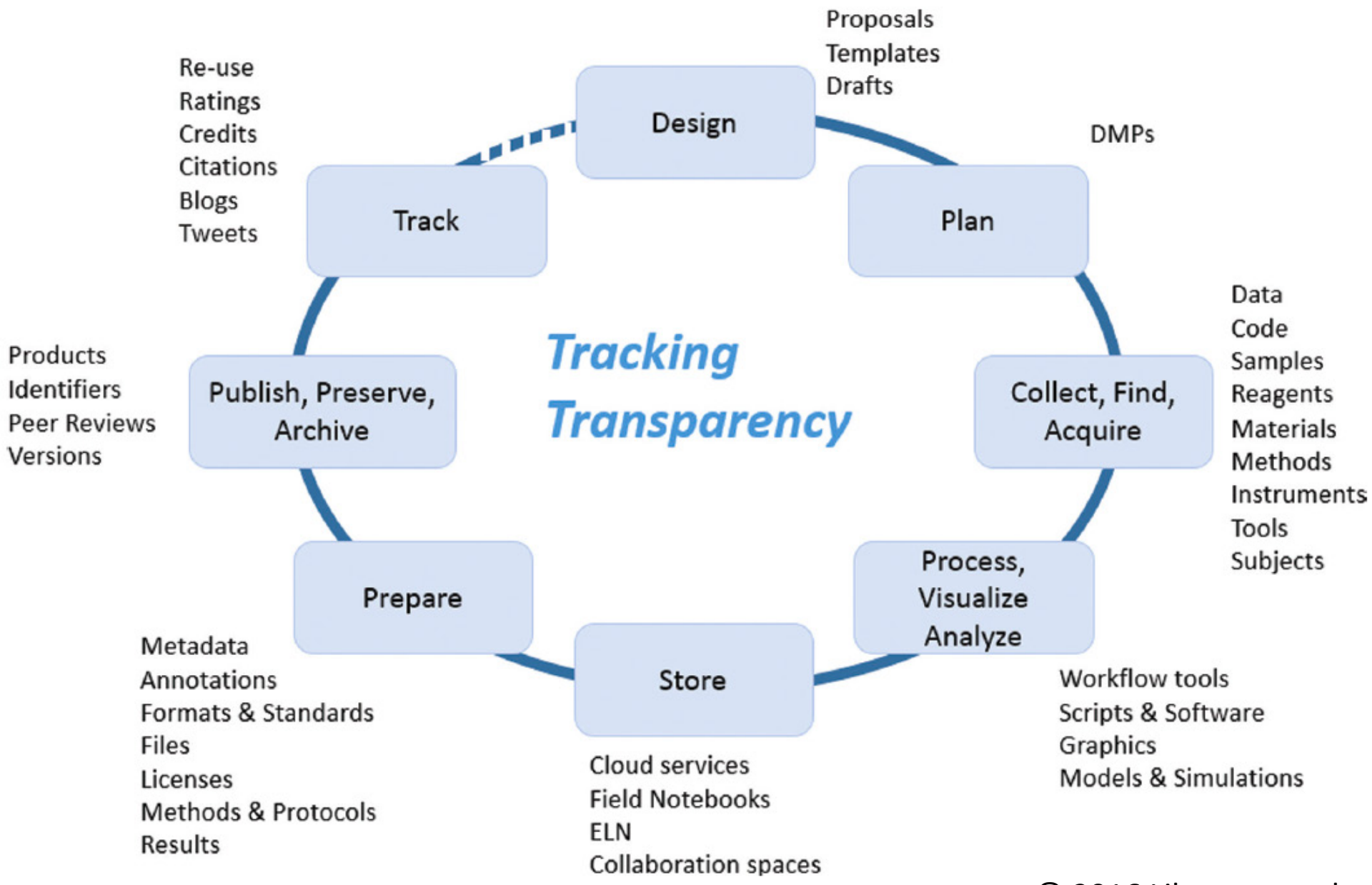
Research data lifecycle: Data documentation initiative



Source: DDI (2013)

Urgent need for establishing explicit analytical workflows

Research data lifecycle: Transparency



Key resource: Office of Research Integrity RDM site

https://ori.hhs.gov/education/products/rcradmin/topics/data/open.shtml

Got data?

Data published in papers represents a small fraction of all *useful* data generated in labs

peer review; stable; citable; usually unstructured

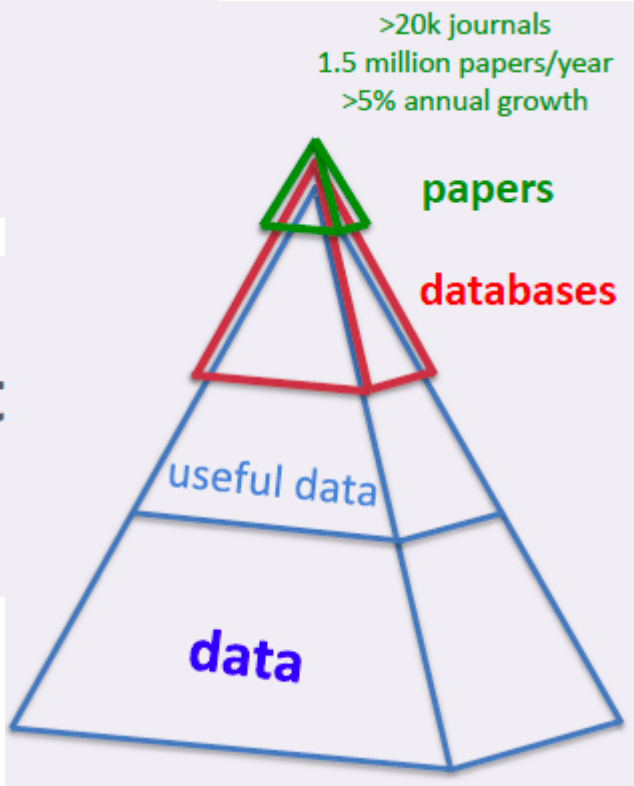
Data deposited in databases represents a bigger fraction of the data

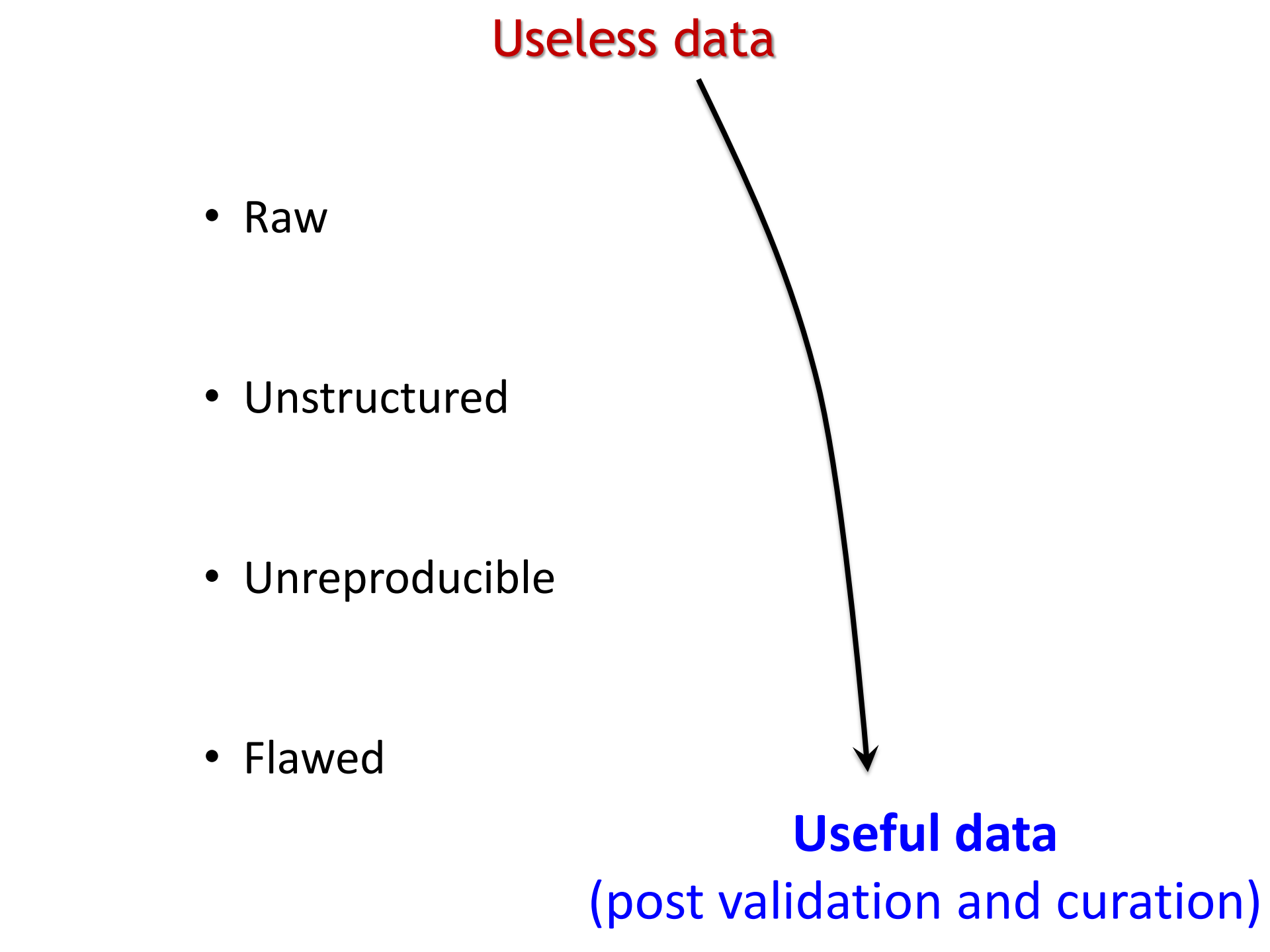
usually curation; (stable); citable; structured

Data deposited in repositories may capture a large fraction of the dat

some curation; often unstructured

>> validation, citability, stability




Useless data
• Raw
• Unstructured
• Unreproducible
• Flawed
Useful data
(post validation and curation)

If you plan on using Adobe Photoshop®…

- Changing brightness and contrast is allowed ONLY when you apply it equally across the entire image and equally to all controls.
- Do NOT alter contrast to make your data disappear (e.g. masking)
- Do NOT enhance or emphasize one region of your image while leaving others unchanged
- Do NOT use touch up tools (e.g. cloning, healing, etc) on your data
- Do NOT use features that deliberately obscure manipulations
- Be prepared to supply the Journal editors with original data on request

Enhancing transparency and including source data

SOURCE DATA

TRANSPARENT PROCESS

OPEN ACCESS

Published online: September 13, 2016

Article

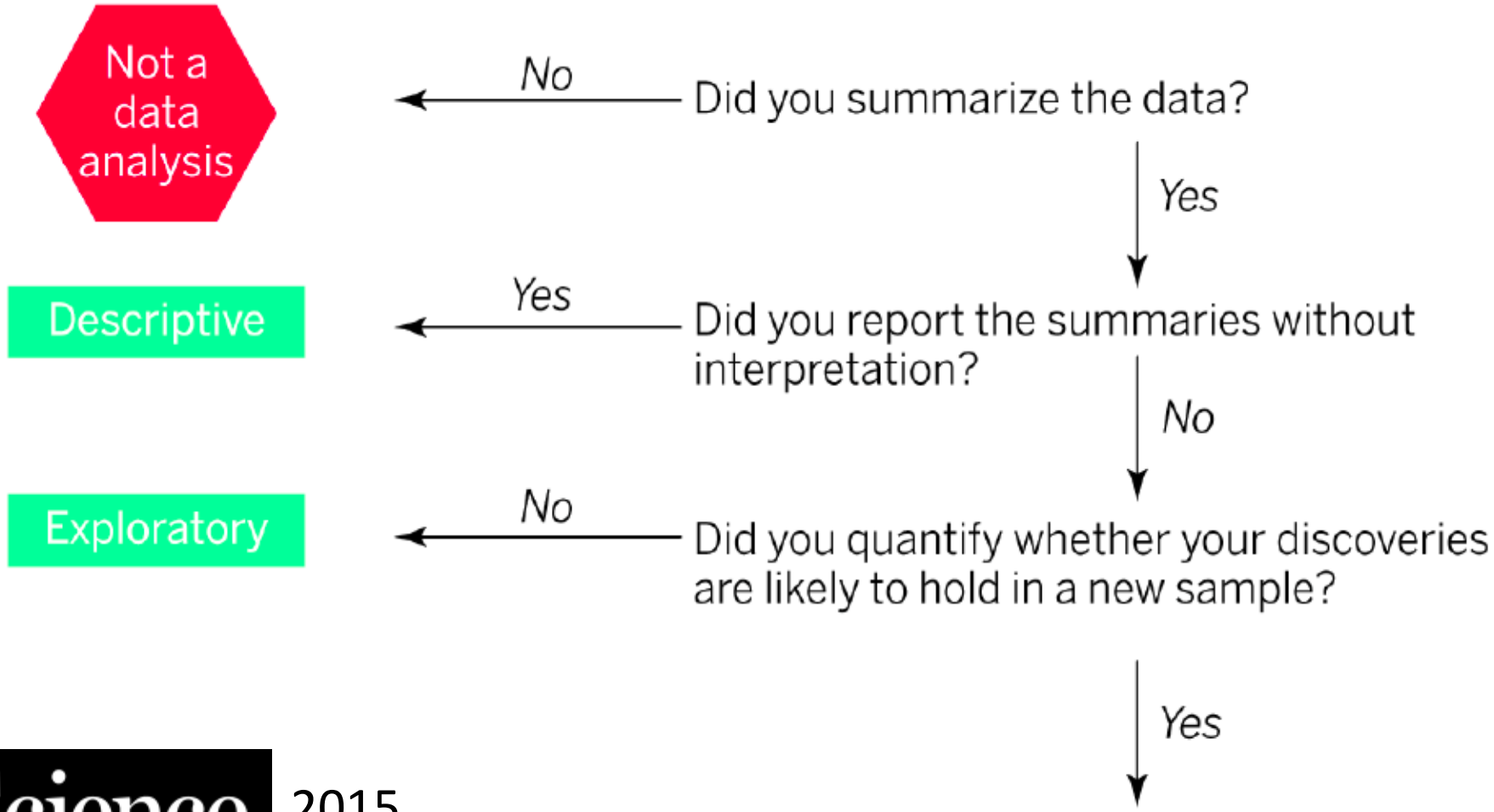
THE EMBO JOURNAL

FBW7 suppression leads to SOX9 stabilization and increased malignancy in medulloblastoma

Aldwin Suryo Rahmanto[1,†], Vasil Savov[2,†], Andrä Brunner[1,‡], Sara Bolin[2,‡], Holger Weishaupt[2,‡], Alena Malyukova[1], Gabriela Rosén[2], Matko Čančer[2], Sonja Hutter[2,3], Anders Sundström[2], Daisuke Kawauchi[3], David TW Jones[3], Charles Spruck[4], Michael D Taylor[5], Yoon-Jae Cho[6], Stefan M Pfister[3,7], Marcel Kool[3], Andrey Korshunov[3,8], Fredrik J Swartling[2,*,§] & Olle Sangfelt[1,**,§]

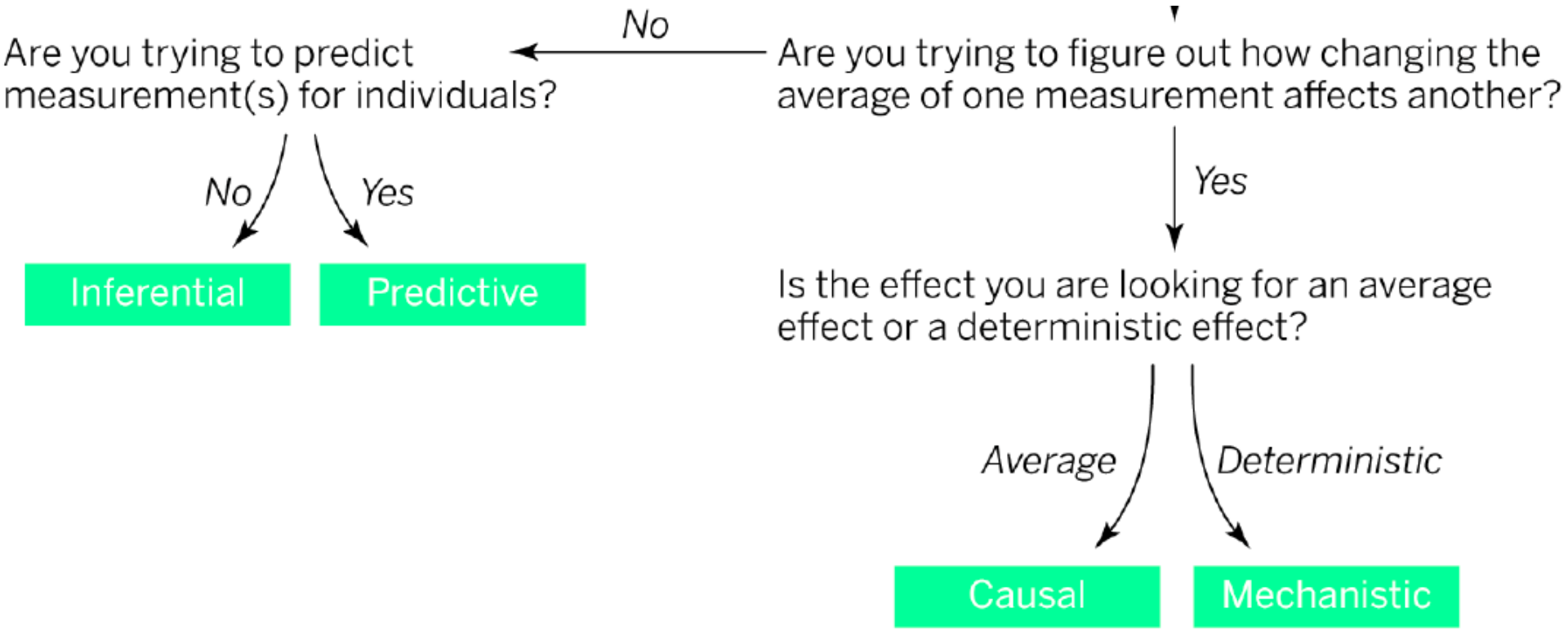
Data analysis: know thy question

"Mistaking the type of question being considered is the most common error in data analysis" Jeff Leek, JHU SPH



Science 2015

Data analysis: know thy question



Science 2015

Data analysis: know thy question

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	"correlation does not imply causation"
Exploratory	Inferential	"data dredging"
Exploratory	Predictive	"overfitting"
Descriptive	Inferential	"n of 1 analysis"

Science 2015

FAIR principles for data stewardship

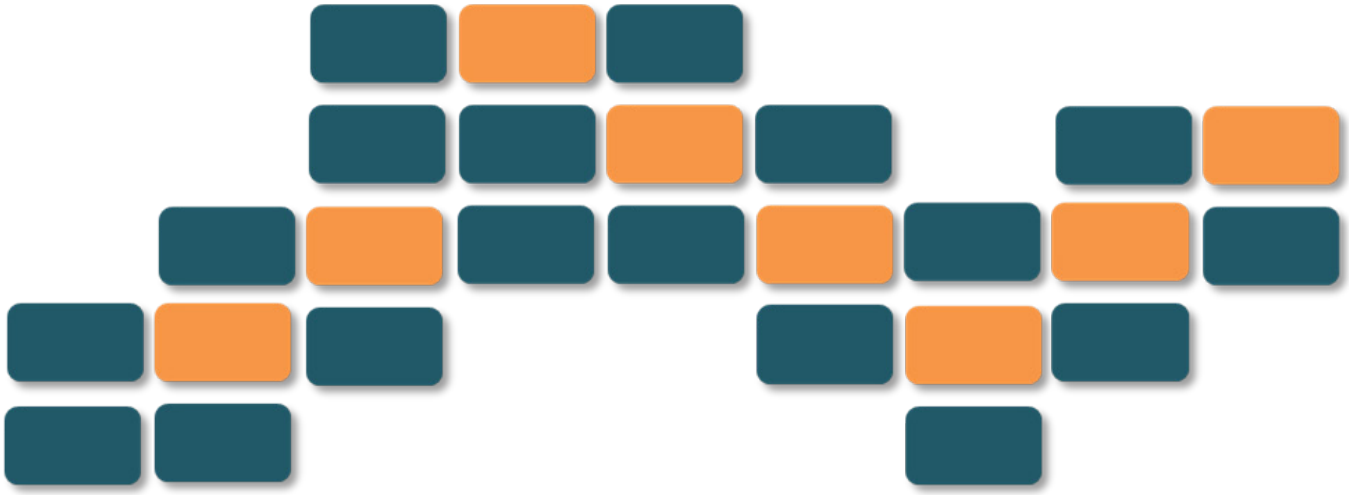
Find

Access

Interoperate

Re-use

Data



SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18

FAIR principles for data stewardship

Equally important to good scholarship is the publication of non-data research objects. Explicit analytical workflows, for example, are essential to most forms of knowledge generation. Publication of these according to FAIR principles is essential to ensure transparency of the work as well as maximal use to the community. The key to working with data is to realize that the human touch, the urge to annotate tables with footnotes and cram multiple elements and data types into every cell of a table, gets in the way of computation, automation and scaling up. And this impedes the usefulness of your work for other people. All research objects should be findable, accessible, interoperable and reusable (FAIR) both for machines and for people.

NATURE GENETICS | VOLUME 48 | NUMBER 4 | APRIL 2016

biosharing.org

standards, databases, policies

A curated, informative and educational resource on inter-related data standards, databases, and policies in the life, environmental and biomedical sciences

Find

Recommendations

Standards and/or databases recommended by journal or funder data policies.

Discover

Collections

Standards and/or databases grouped by domain, species or organization.

Learn

Educational

About standards, their use in databases and policies, and how we can help you.

FAIRDOM

Findable

Accessible

Interoperable

Reusable

FAIRDOM helps you to be in control of collecting, managing, storing, and publishing your data, models, and operating procedures.

The Dataverse Project

Open source research data repository software

Top 10 tips for research data management (RDM)

- Imagine the worst case scenario
- Check UMB and funder RDM policy
- Identify information sources and people who can help you
- Where are the useful checklists and workflows?
- Data Management Plan (DMP)
- Think about what to archive and how to describe and cite it
- Check out suitable data archives
- Does your dataset require special software?
- What are the RDM-associated costs at each stage?
- Who owns your data and can you make it publicly available?

Data management plan

Repositories for appropriately consented data

nature genetics

No impact without data access

A considerable proportion of the usefulness and interest of research publications in our field comes from the data and associated metadata. We therefore insist that data be available for peer reviewers to see and readers to use. Authors should use public permanent repositories designed for appropriately consented data.

Repositories for appropriately consented data

VOLUME 46 | NUMBER 9 | SEPTEMBER 2014 | NATURE GENETICS

Data use under the NIH GWAS Data Sharing Policy and future directions

Dina N Paltoo[1,10], Laura Lyman Rodriguez[2,10], Michael Feolo[3,10], Elizabeth Gillanders[4], Erin M Ramos[2], Joni L Rutter[5], Stephen Sherry[3], Vivian Ota Wang[2], Alice Bailey[2], Rebecca Baker[1], Mark Caulder[5], Emily L Harris[6], Kristofor Langlais[1], Hilary Leeds[7], Erin Luetkemeier[1], Taunton Paine[1], Tamar Roomian[2,9], Kimberly Tryka[3], Amy Patterson[1] & Eric D Green[2] for the National Institutes of Health Genomic Data Sharing Governance Committees[8]

In 2007, the US National Institutes of Health (NIH) introduced the Genome-Wide Association Studies (GWAS) Policy and the database of Genotypes and Phenotypes (dbGaP) to facilitate 'controlled' access to GWAS data based on participants' informed consent. dbGaP has provided 2,221 investigators access to 304 studies, resulting in 924 publications and significant scientific advances. Following on this success, the 2014 Genomic Data Sharing Policy will extend the GWAS Policy to additional data types.

Repositories for appropriately consented data

VOLUME 47 | NUMBER 7 | JULY 2015 | NATURE GENETICS

The European Genome-phenome Archive of human data consented for biomedical research

Ilkka Lappalainen[1], Jeff Almeida-King[1], Vasudev Kumanduri[1], Alexander Senf[1], John Dylan Spalding[1], Saif ur-Rehman[1], Gary Saunders[1], Jag Kandasamy[1], Mario Caccamo[1,5], Rasko Leinonen[1], Brendan Vaughan[1], Thomas Laurent[1], Francis Rowland[1], Pablo Marin-Garcia[1,5], Jonathan Barker[1], Petteri Jokinen[1], Angel Carreño Torres[2], Jordi Rambla de Argila[2], Oscar Martinez Llobet[2], Ignacio Medina[1], Marc Sitges Puy[2], Mario Alberich[2], Sabela de la Torre[2], Arcadi Navarro[2–4], Justin Paschall[1] & Paul Flicek[1]

The European Genome-phenome Archive (EGA) is a permanent archive that promotes the distribution and sharing of genetic and phenotypic data consented for specific approved uses but not fully open, public distribution. The EGA follows strict protocols for information management, data storage, security and dissemination. Authorized access to the data is managed in partnership with the data-providing organizations. The EGA includes major reference data collections for human genetics research.

Whither statistics?

NATURE

Error prone

Nature **487**, 406 (2012) doi:10.1038/487406a

'biologists fail to design experiments properly, and so submit underpowered studies that have an insufficient sample size and trumpet chance observations as biological effects.... Researchers ...must agree on standards that will protect against avoidable errors. '

Nature Reviews Neuroscience

Power failure: why small sample size Undermines the reliability of neuroscience

Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson & Munafò

Nature Reviews Neuroscience **14**, 365-376 (2013) doi:10.1038/nrn3475

'the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful.'

"There are three kinds of lies: lies, damned lies, and statistics" Mark Twain

"If your experiment needs statistics, you ought to do a better experiment." Ernest Rutherford (Works in fields with high signal-to-noise)

"If you are going to analyze your data using statistical methods, plan the methods carefully, do the analyses seriously, and report the data, methods, and results completely." Harvey J. Motulsky (2015)

10 simple rules: effective statistical practice

Rule 1: Statistical Methods Should Enable Data to Answer Scientific Questions

Rule 2: Signals Always Come with Noise

Rule 3: Plan Ahead, Really Ahead

Rule 4: Worry about Data Quality

Rule 5: Statistical Analysis Is More Than a Set of Computations

Rule 6: Keep it Simple

Rule 7: Provide Assessments of Variability

Rule 8: Check Your Assumptions

Rule 9: When Possible, Replicate!

Rule 10: Make Your Analysis Reproducible

Blind analysis requires serious consideration

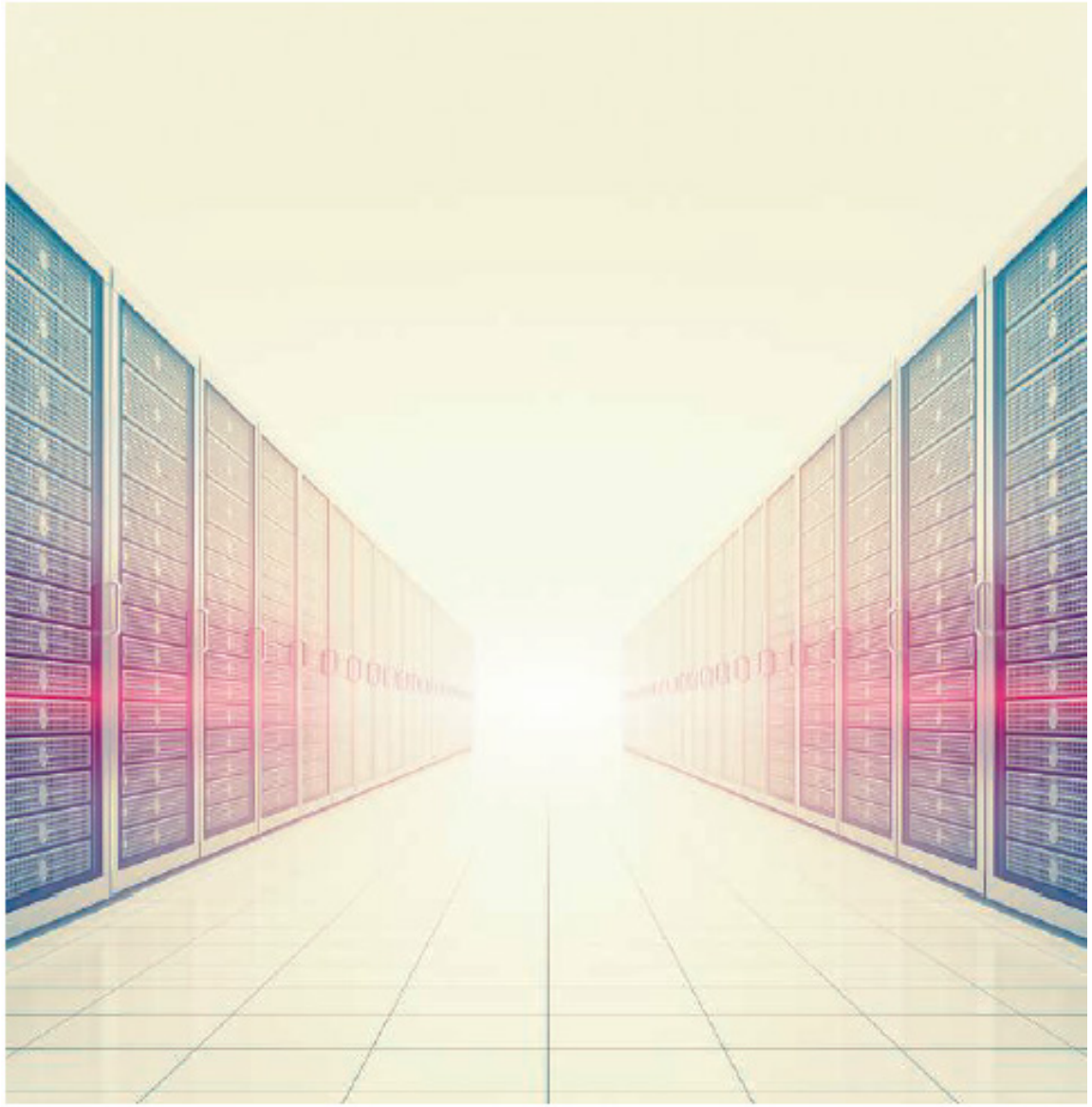
Illustration by Dale Edwin Murray

Confirmation bias

"The tendency to look for and perceive evidence consistent with our hypotheses and to deny, dismiss or distort evidence that is not."

"[P]erhaps the best known and most widely accepted notion of inferential error to come out of the literature on human reasoning" Evans, J., (1990) Bias In Human Reasoning: Cause and Consequences.

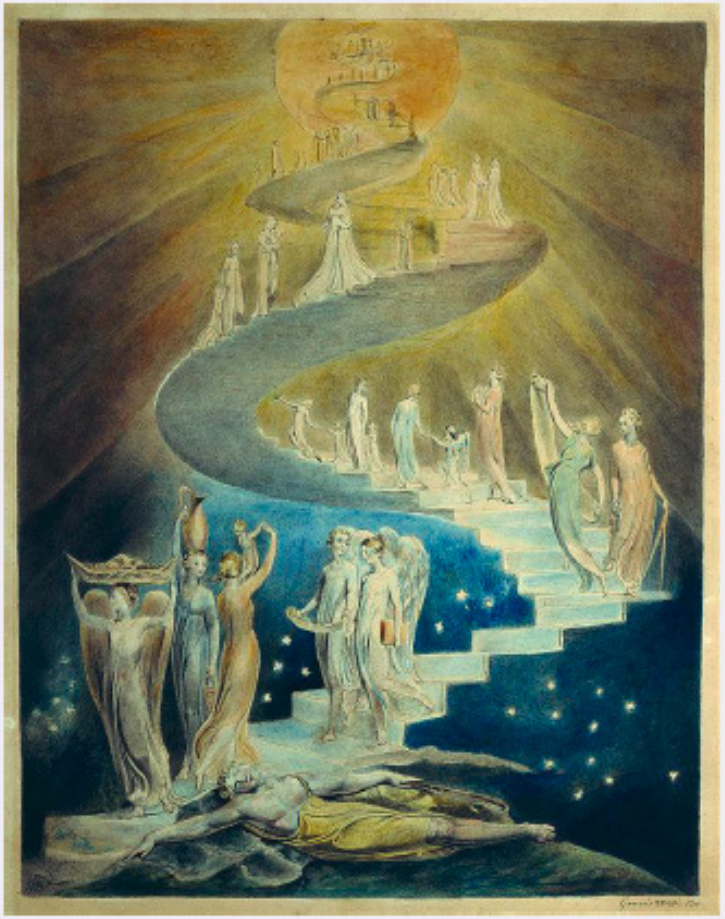
Even "[g]ood scientists are not immune from confirmation bias. They are aware of it and avail themselves of procedural safeguards against its pernicious effects." (Lilienfeld, S., Sci. Am. 303, 18 (2010). Fudge Factor: A Look at a Harvard Science Fraud Case)



"...interpretations come and go, but data are forever."

Marcia McNutt
Editor-in-Chief
Science *Journals*

Open data



Jacob's Dream - William Blake

Opening the Gates…

Gate of Hell - William Blake