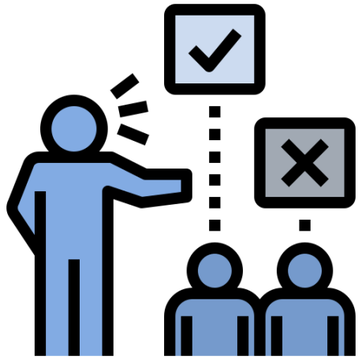
What do we do with data biases?

INTRODUCTION TO DATA ETHICS

Shalini Kurapati, PhD
Co-founder and CEO, Clearbox AI

datacamp

What is bias?

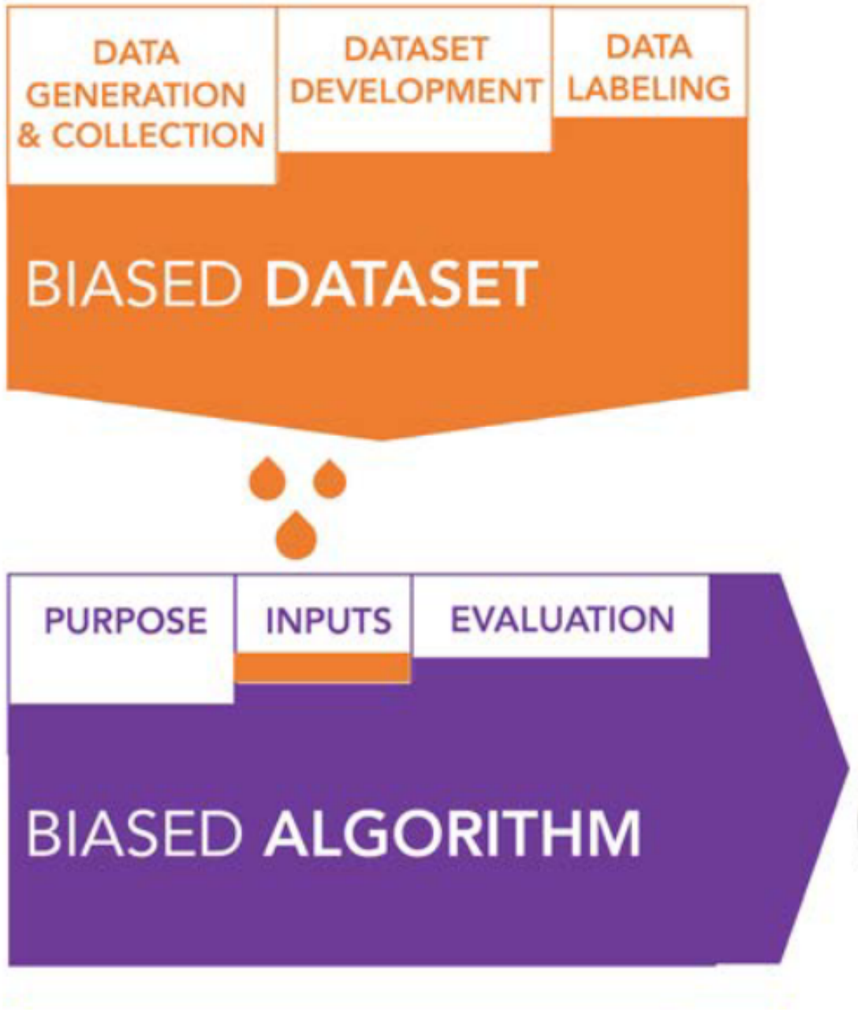


- First thought in data ethics
- Prejudice, for or against a person, group or an idea- unfair
- Data ethics- data bias
- Under, over, or mis-representation
- Harmful effects if biased data is used for decision-making

[1] Icon made by noomtah from www.flaticon.com

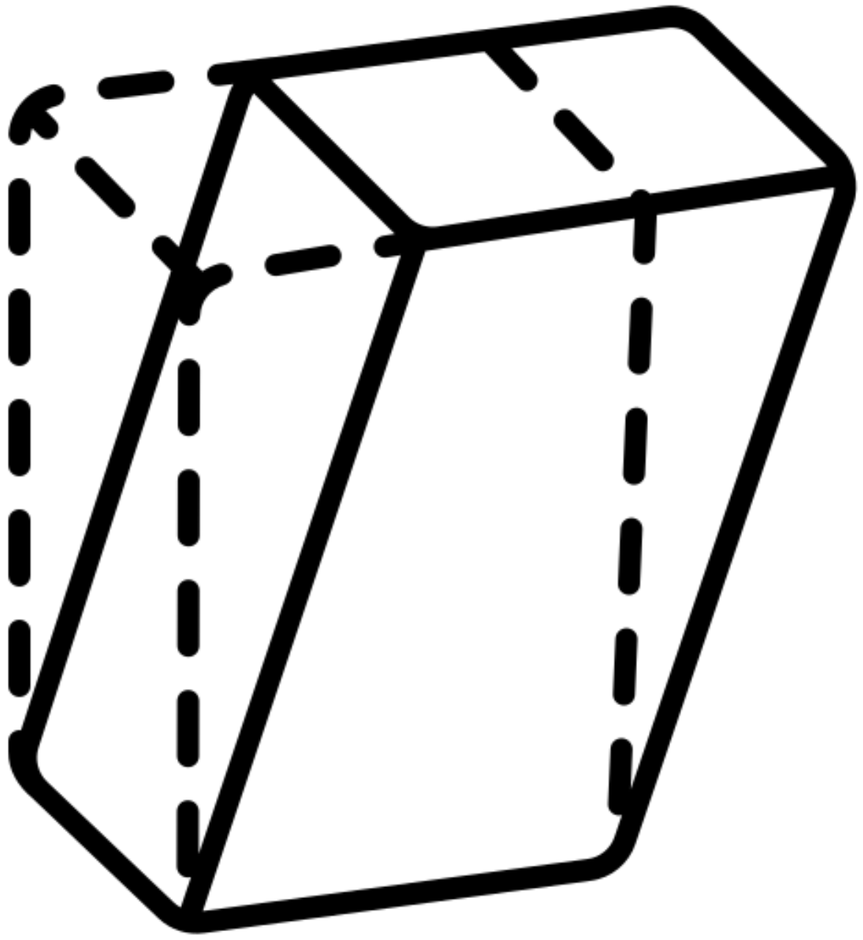
Anytime, anywhere

- At all stages of data and life cycles
- Data collection
- Data cleaning, preparation, development
- Data labeling



Data specific biases

- Technical or statistical process bias
  - Sampling bias- representation
  - Measurement bias
  - Self-reporting bias
  - Labeling bias
- Human and Systematic bias
  - Gender
  - Ethnicity/minorities
  - Culture



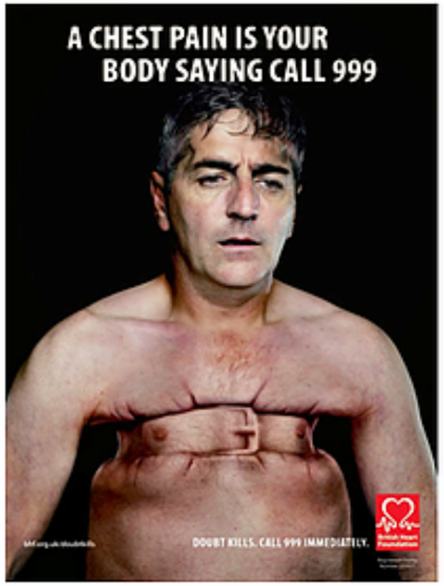
[1] Icon made by Flowicon from www.flaticon.com

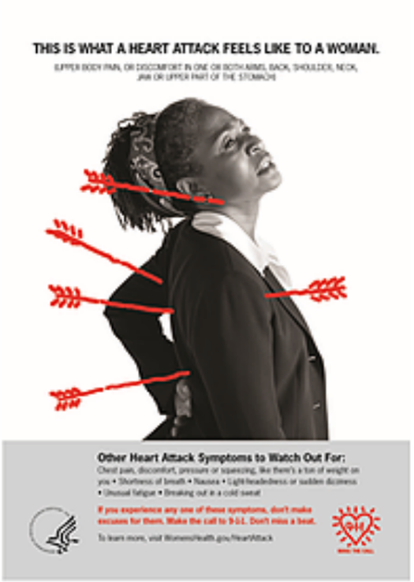
Representation is crucial

- Lack of representative data
- Facial recognition- better outcomes with a white mask
- Hand soap dispensers to healthcare outcomes
- Recommended: Coded bias

[1] https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

A mirror to our stereotypes



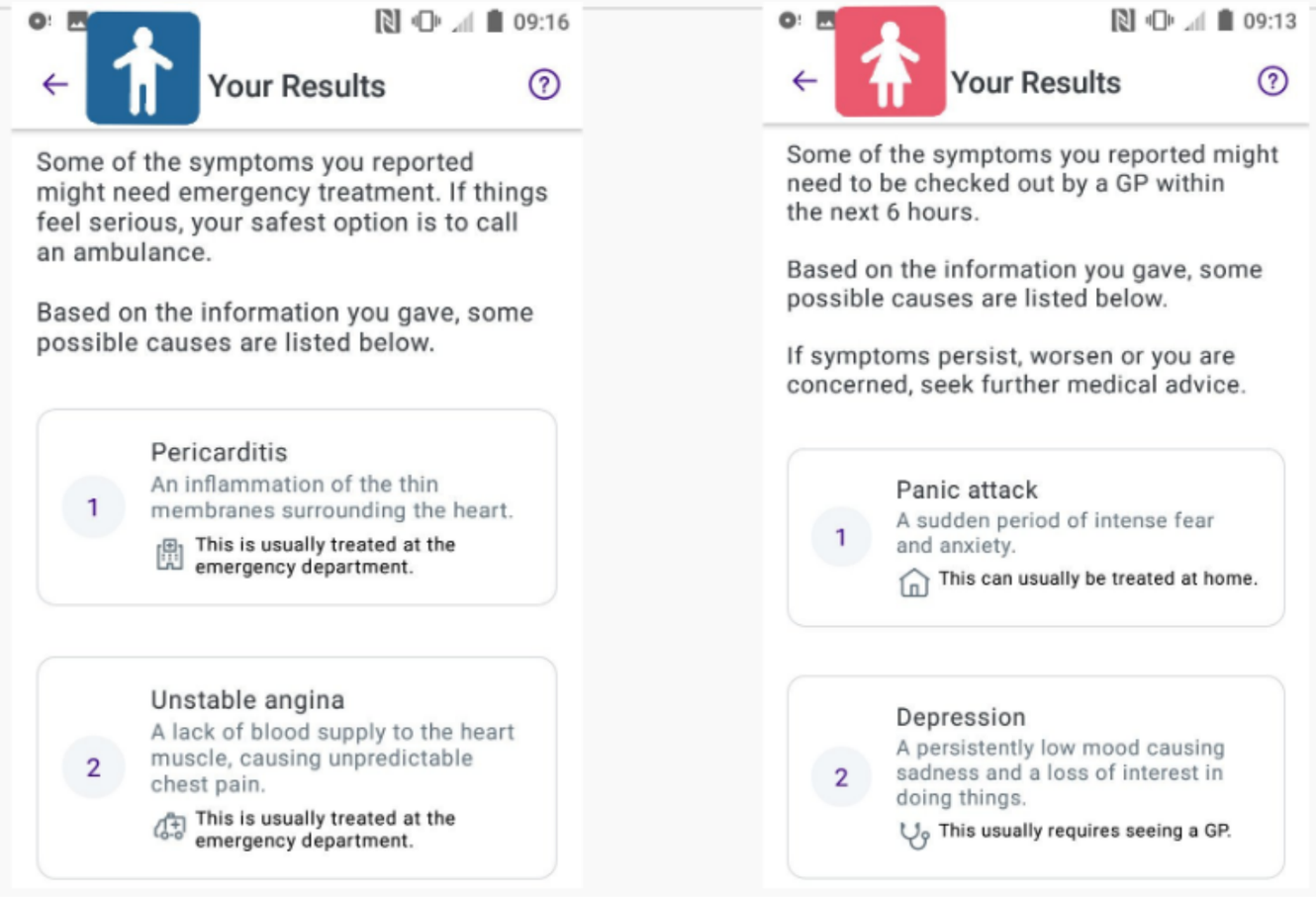


W G H

- Medical studies and data, historically done on white males
- E.g. Heart attack symptoms are different for men and women
- What happens if we use a diagnostic app using this data?

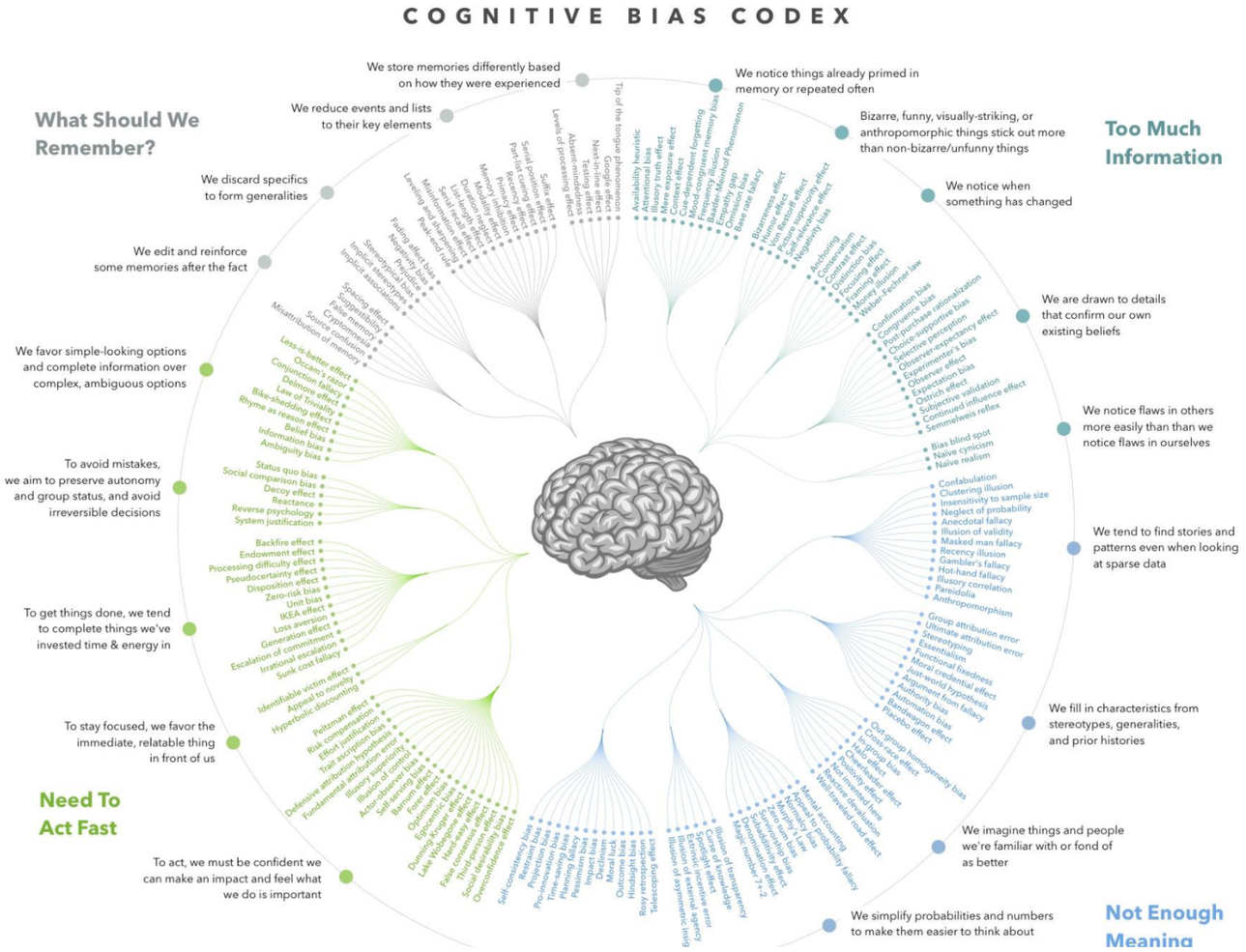
[1] https://womeningh.org/

Serious impact



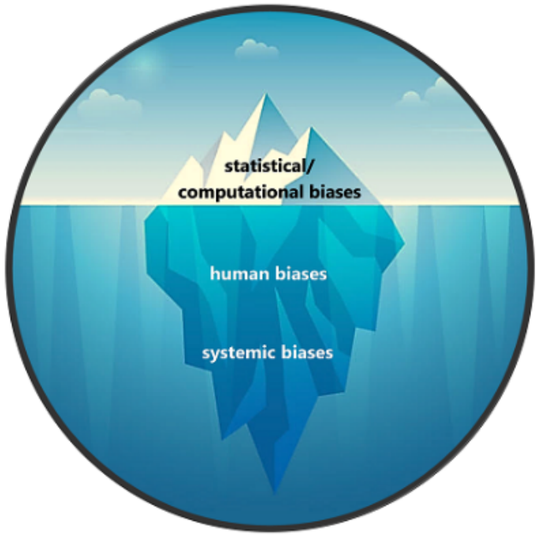
- Heart attack symptoms- emergency for a man, panic attack for a woman

Way too many and counting!



[1] CC BY, John Manoogian III (JM3)

Tip of the ice berg



- Statistical biases tip of ice-berg
- Human and systematic bias- attitude and societal changes
- Test for bias, prevention and mitigation, open to feedback

[1] Schwartz, R. , Vassilev, A. , Greene, K. , Perine, L. , Burt, A. and Hall, P. (2022), Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, National Institute of Standards and Technology

Let's practice!

INTRODUCTION TO DATA ETHICS

Fair data or FAIR data?

INTRODUCTION TO DATA ETHICS

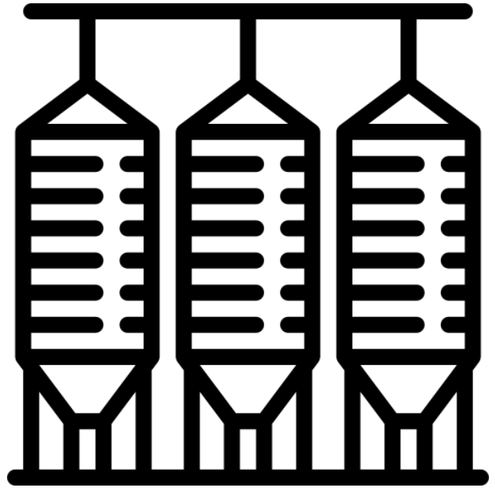
Shalini Kurapati, PhD
Co-founder and CEO, Clearbox AI

datacamp

A new FAIR data you say?

- FAIR: **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
- Fairness or unbiased data
- But what are FAIR data principles
- Valorize data through efficient data management systems
- Reuse of data, better awareness, retrieval
- Making data ready for collaboration and sharing
- Scientific research, more data intensive industries

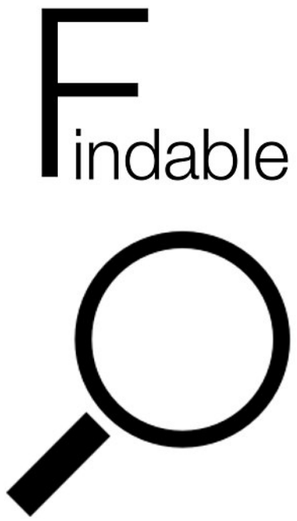
Why FAIR?



- Organizational silos, forgotten data
- Transparency, innovation and collaboration
- Digital and computing technologies/AI
- Data management for reuse by humans and machines

[1] Icon made by Smashicons from www.flaticon.com

Findable



- Data description, rich meta data
- Persistent link, like a *doi*, no broken links
- Human and machine friendly

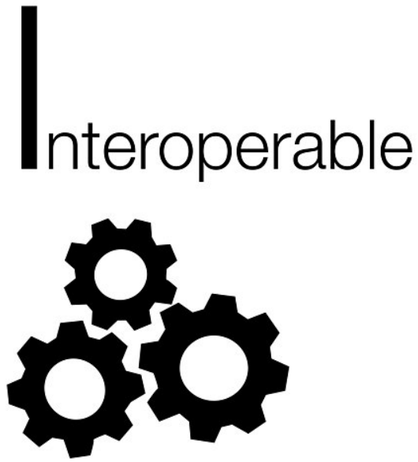
[1] CC BY-SA 4.0, Sangya Pundir

Accessible

- Physical or digital infrastructure
- Accessible through login or openly

[1] CC BY-SA 4.0, Sangya Pundir

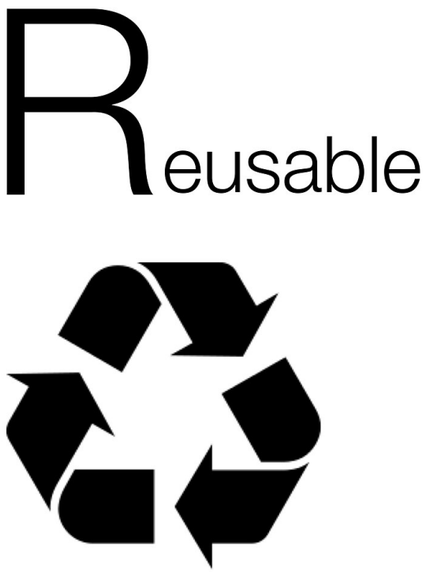
Interoperable



- Data works across different OS, file formats, devices
  - Windows/PC, Mac, Linux
  - e.g. spreadsheets .csv instead of Ms Excel
  - standards for interoperability

[1] CC BY-SA 4.0, Sangya Pundir

Reusable



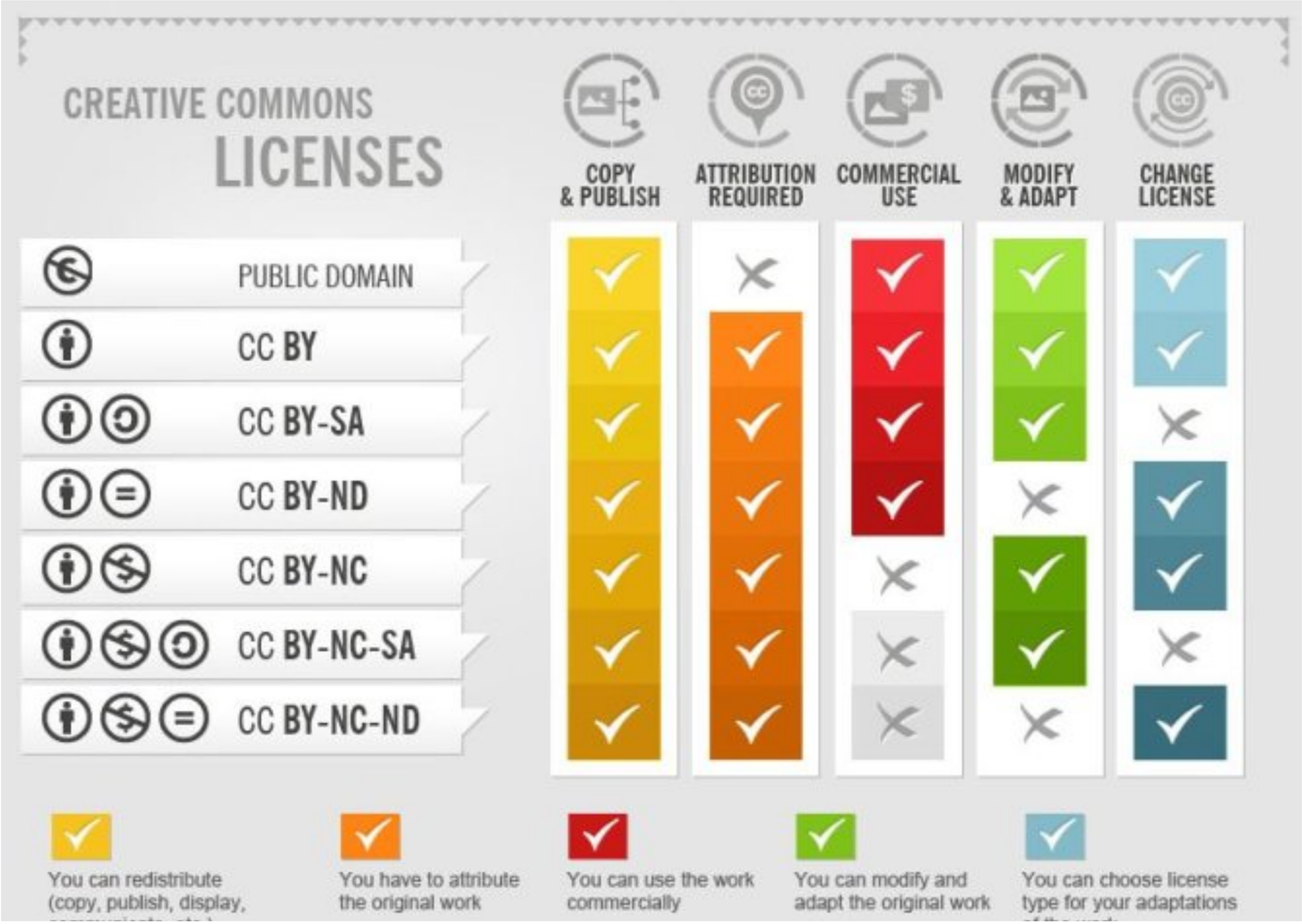
- Reuse data with clear rules
- Human and machines

[1] CC BY-SA 4.0, Sangya Pundir

Terms of reuse

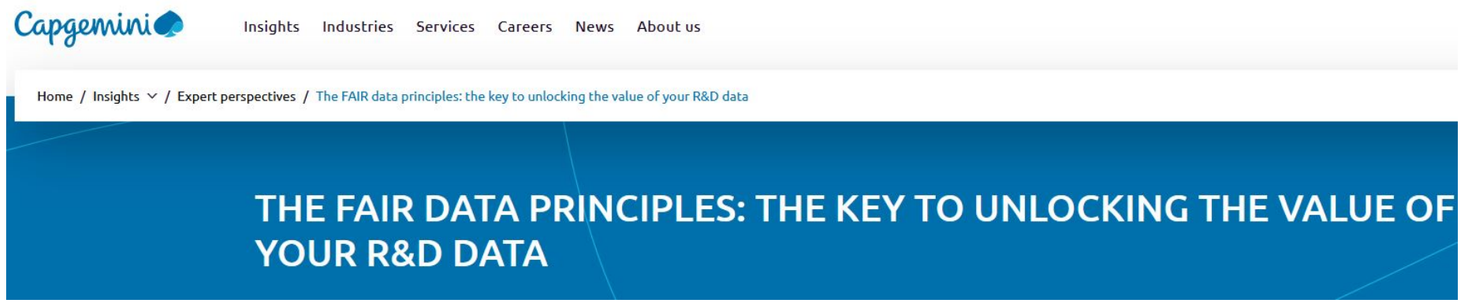
- Clear expectations
- Documentation for use, adaption
- Credit, where credit is due
- Many ways- copyright, licenses

Data licensing- Creative Commons



[1] Foter, CC-BY-SA

Industry needs to go FAIR



- FAIR data common in scientific research data management
- Pharma companies (Janssen, Novartis), technology firms (Deloitte, Accenture)
- Resource intensive at first, attitude change
- Data ethics
- Innovation and collaboration leading to societal benefits, like alternative energy disease treatment

[1] https://direct.mit.edu/dint/article/2/1-2/276/10011/The-Need-of-Industry-to-Go-FAIR [2] https://www.capgemini.com/insights/expert-perspectives/the-fair-data-principles-the-key-to-unlocking-the-

Let's practice!

INTRODUCTION TO DATA ETHICS

Open data for science and society

INTRODUCTION TO DATA ETHICS

Shalini Kurapati, PhD
Co-founder and CEO, Clearbox AI

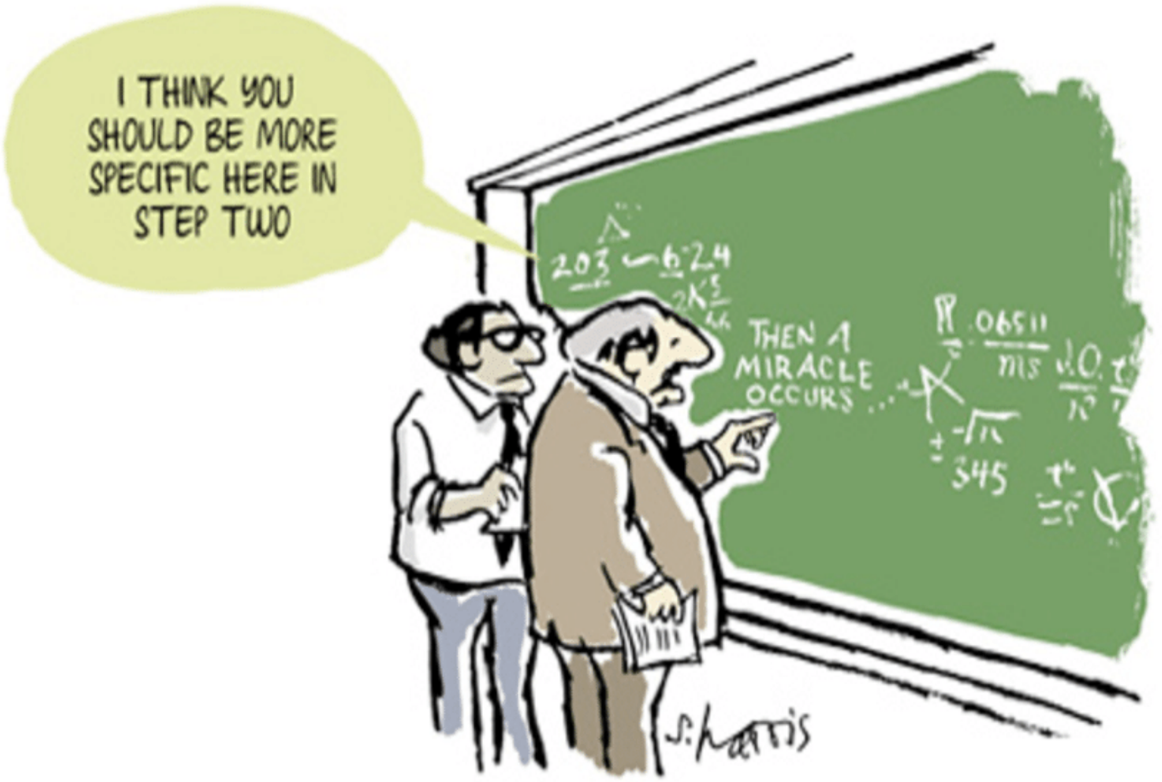
datacamp

Why open up data?

- Data ethics discussion are around privacy
- Collaborative nature of innovation
- Benefits to science, innovation, economy and society
- Opportunities and challenges

In scientific research

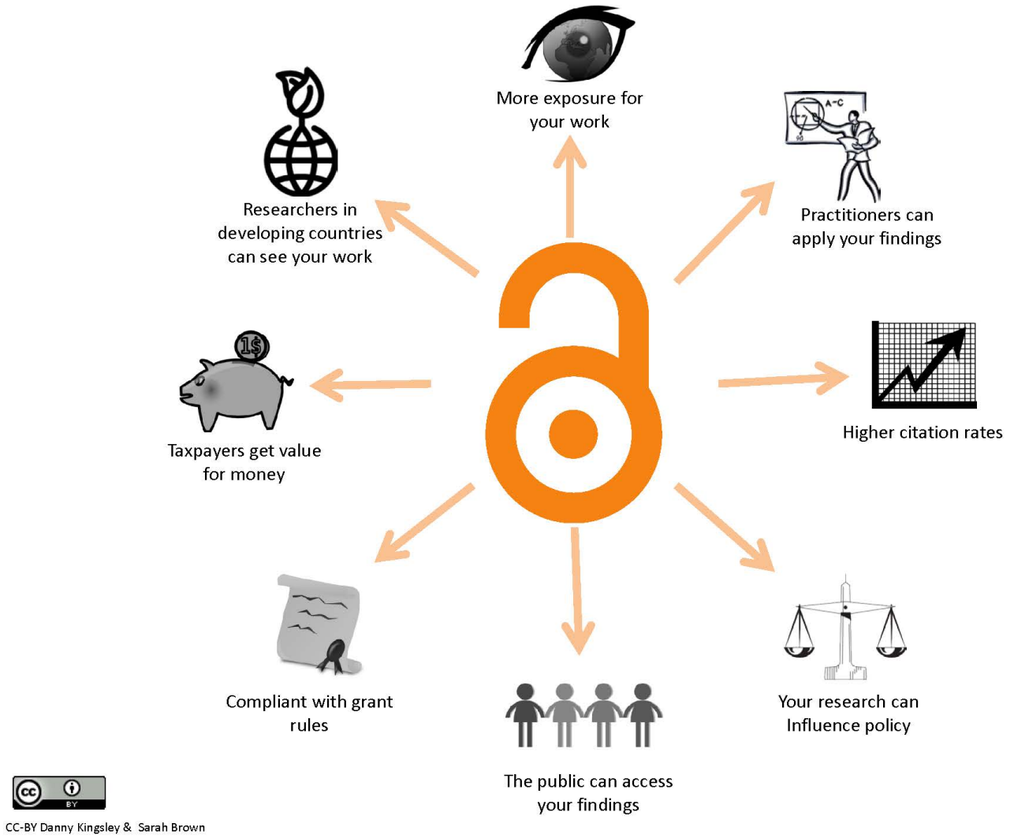
- Scientific articles often don't provide data
- Verification of results
- Reproducible research and generalization
- Advancement of the scientific endeavor
- Push for open data
- Data for others to reuse, public funding



[1] Cartoon source: Sidney Harris in American Scientist, Nov-Dec 1977

Inclusive and equitable science

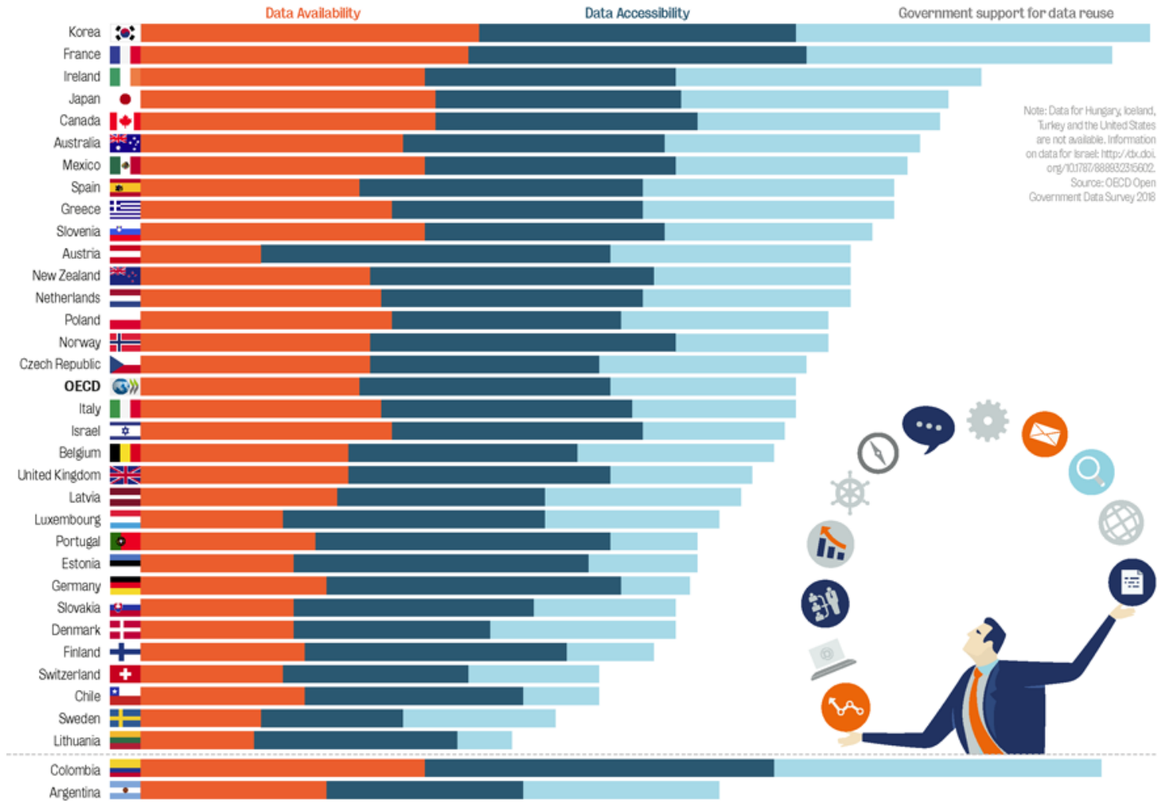
- More exposure and impact of scientific outputs
- Practitioners and policy makers can adopt findings
- Value for money for taxpayer, research funding
- Public engagement, citizen science
- Researchers in developing countries can benefit



[1] Image source: Danny Kingsley and Sarah Brown under a CC BY license

Open government data

- Transparency, value creation- data as a common good
- Large volumes of data
- Reuse of data to enhance business creation, innovation
- Citizen engagement and citizen-centric services



[1] https://www.oecd.org/gov/digital-government/policy-paper-ourdata-index-2019.htm

Success stories

Business based on open data

- ApplyBoard, Canadian company
- Online platform for international student recruitment
- Open data on student visa processing times

Global Open Data for Agriculture and Nutrition

- Agricultural and nutritional data to be available, accessible, usable and unrestricted
  - Land-use and productivity data
  - Hydrological, soil data
  - Pest and disease management
- Eradicate hunger, food security, sustainable agriculture

[1] https://www.godan.info/agriculture [2] https://open.canada.ca/en/story/applyboard

Work in progress

- FAIR- Findable, Accessible, Interoperable, Reusable, resource intensive
- Data quality, bias
- Data privacy, data ownership
- Attitude and cultural shift

Let's practice!

INTRODUCTION TO DATA ETHICS

Evolving accountability and impact

INTRODUCTION TO DATA ETHICS

Shalini Kurapati, PhD
Co-founder and CEO, Clearbox AI

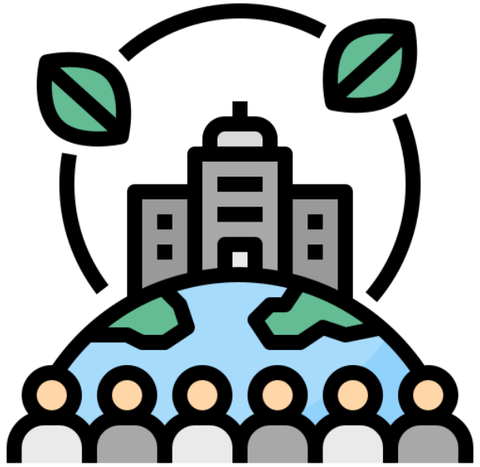
datacamp

Accountability is not the same as compliance

- Focus on data privacy and protection
- Compliance
  - GDPR- General Data Protection Regulation of the EU
  - Controllers & Processors
- Evolving
  - Regulations
  - Rights and attitudes
  - Environment and society
  - Progress

[1] Icon made by Vectors Tank from www.flaticon.com

Impact and risk assessments



- Data protection impact assessment
- Residual risks - risks of re-identification
- Fairness and discrimination
- More to accountability

[1] Icon made by noomtah from www.flaticon.com

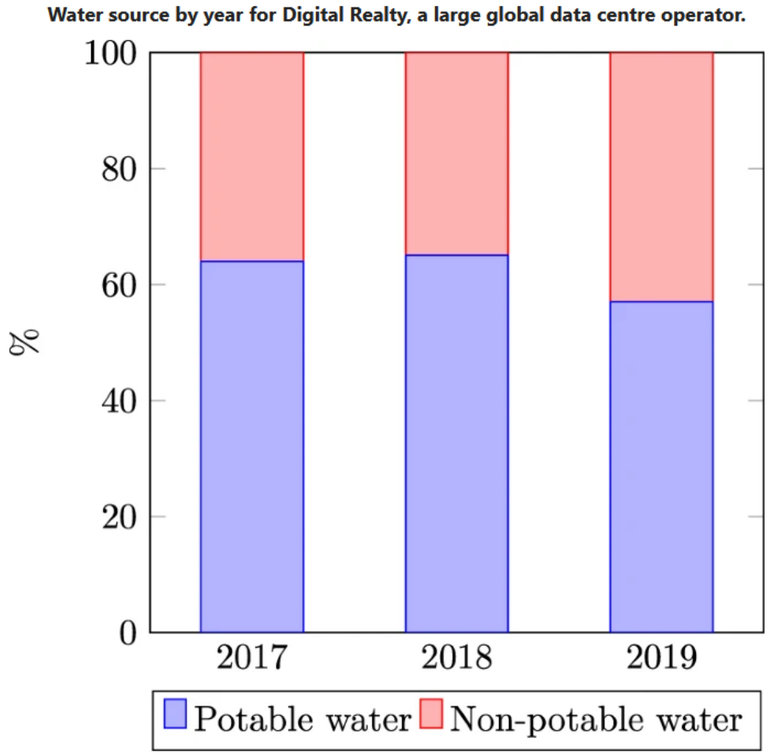
Not your obvious data ethics topic

CARE Principles for Indigenous Data Governance

- Good data management and protection not sufficient
- Sensitive to power dynamics and historical injustices
- Data ethics for reconciliation, CARE for indigenous data governance
- Data generated by, from indigenous people
- Collective Benefit, Authority to control, Responsibility, Ethics

[1] https://www.gida-global.org/care

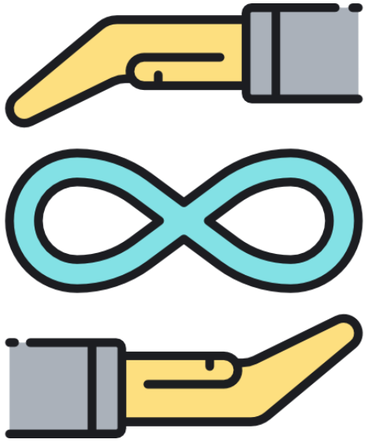
Data and drinking water



- Data centers use potable water for cooling, how to reduce?

[1] https://www.nature.com/articles/s41545-021-00101-w

Accountability is continuous



- Regulations, codes of conduct
- Analytics, AI impact assessment
- Monitoring and oversight data activities
- Stakeholder analysis
- Competitive advantage, future-proof

[1] Icon made by Flat-icons-com from www.flaticon.com

Let's practice!

INTRODUCTION TO DATA ETHICS

Congratulations!

INTRODUCTION TO DATA ETHICS

Shalini Kurapati, PhD
Co-founder and CEO, Clearbox AI

datacamp

Understanding Data Ethics

- Ethics vs. Data ethics
- Beyond do no harm
- Data ethics and AI ethics

Data ethics: Now and forever

- Ethics across the data life cycle
- Data protection and security
- Trust and transparency
- Informed consent and data ownership

Data for good

- Data-specific biases
- fair data or FAIR data?
- Open data for science and society
- Evolving accountability and impact

All the best!

INTRODUCTION TO DATA ETHICS