Good practice in human data handling and analysis

IMforFUTURE meeting 2018, Zagreb

Lennart Karssen

26–28 March 2018

PolyΩmica
’s-Hertogenbosch, NL
l.c.karssen@polyomica.com

PolyΩmica

Outline

Introduction

About me

- Dutch
- 2008: PhD in experimental atomic physics, Utrecht University, NL
- 2009: Teaching physics @ National University of Rwanda
- 2009: Unix/Linux consultant @ Snow B.V., NL
- 2010: Postdoc bioinformatics, data manager, sysadmin @ genetic epidemiology group, Erasmus MC, Rotterdam, NL
- 2013: PolyΩmica
- 2018: PolyKnomics

snow
UNIX SPECIALISTS

Erasmus MC
Universitair Medisch Centrum Rotterdam

Image source: Wikipedia, snow.nl, erasmusmc.nl

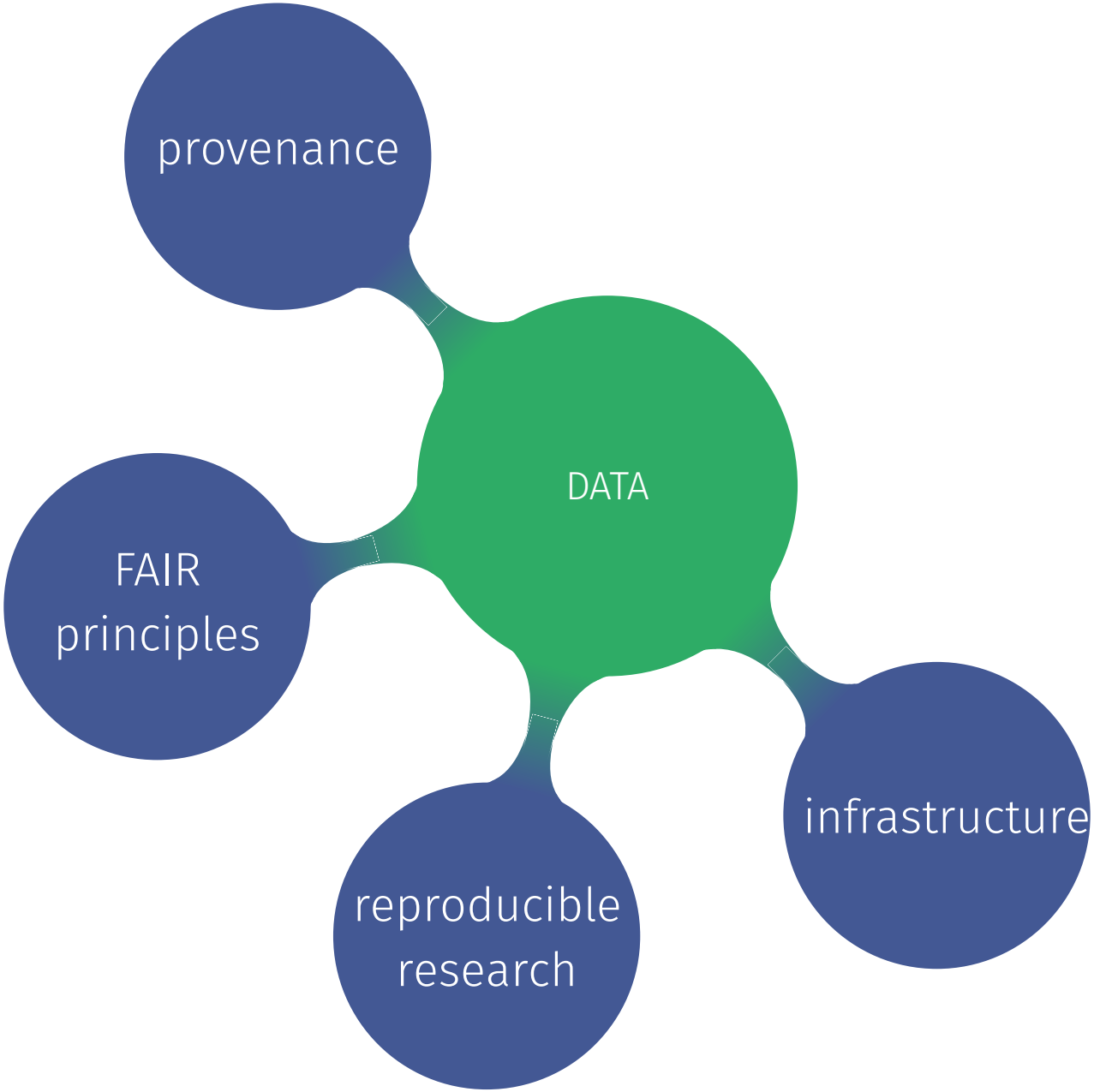
About PolyΩmica

- Research
  - omics analysis: genomics, glycomics,
  - development of statistical methods
  - scientific software development
  - EU-funded projects like IMforFUTURE, PainOmics
- (Scientific) consulting
  - experimental design
  - analyses
  - compute/storage infrastructure
- Education
  - (statistical) genetics
  - epidemiology
  - scientific computing
  - ...

http://www.polyomica.com

Data stewardship

What is data stewardship?



What is data stewardship?

- Lots of data generated: in the lab, on computers: Big Data/Data Tsunami

What is data stewardship?

- Lots of data generated: in the lab, on computers: Big Data/Data Tsunami
- What about
  - the quality of the data?
  - analysis?
  - the reproducibility of research?
  - re-use of new and existing data?

What is data stewardship?

- Lots of data generated: in the lab, on computers: Big Data/Data Tsunami
- What about
  - the quality of the data?
  - analysis?
  - the reproducibility of research?
  - re-use of new and existing data?

In short

The past, present and future quality and availability of data

Typical data stewardship questions

- Where to store the data for analysis?
- Where to archive?
- What to archive?
- Where to run analyses?
- Who pays?
- Shouldn't forget security/privacy!

Aim of data stewardship

Aim for

- proper collection
- annotation
- archiving
- long-term care

of data.

Aim of data stewardship

Aim for

- proper collection
- annotation
- archiving
- long-term care

of data. Not just traditional data, also

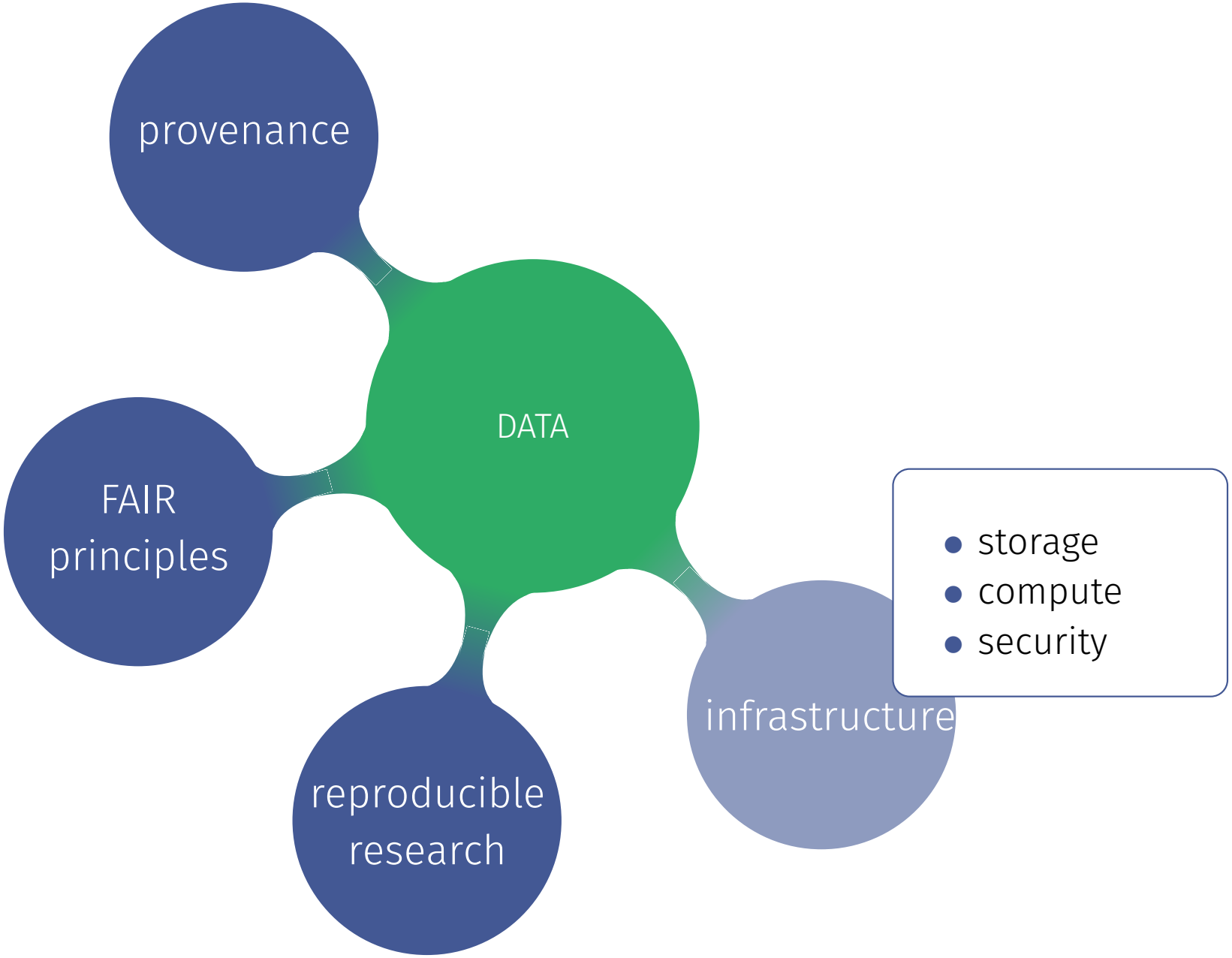
- algorithms
- tools
- workflows

leading to the data

Driving forces

- Funding agencies require data management plan
- Funding agencies require integration with existing data
- Publishers require access to (some of) the data

What is data stewardship?



Some numbers

Disk space usage of common data sets

Data set	Size
Lipidomics GWAS: 117 lipids, 20M SNPs, 3.5k people	122 GB
1kG imputed data: 30M SNPs, 3.5k people	500 GB
UK Biobank imputed data: 90M SNPs, 500k people	2.1 TB
Aligned NGS data (BAM files): 1.3k people	7.2 TB

More numbers

Computation time

Data set	Time
Lipidomics GWAS: 117 lipids, 20M SNPs, 3.5k people	10 d
1kG imputation: 30M SNPs, 3.5k people	1 w
Aligning NGS data (creating BAM files): 1.3k people	weeks

More numbers

Computation time

Data set	Time
Lipidomics GWAS: 117 lipids, 20M SNPs, 3.5k people	10 d
1kG imputation: 30M SNPs, 3.5k people	1 w
Aligning NGS data (creating BAM files): 1.3k people	weeks

Missing information:

How many CPUs/compute nodes were used?

More numbers

Computation time

Data set	Time
Lipidomics GWAS: 117 lipids, 20M SNPs, 3.5k people	10 d
1kG imputation: 30M SNPs, 3.5k people	1 w
Aligning NGS data (creating BAM files): 1.3k people	weeks

Missing information:

How many CPUs/compute nodes were used?

Transferring data over a network

1 TB via 1 Gbps (gigabit per second) network: $\approx$ 2.2 h

More numbers

Computation time

Data set	Time
Lipidomics GWAS: 117 lipids, 20M SNPs, 3.5k people	10 d
1kG imputation: 30M SNPs, 3.5k people	1 w
Aligning NGS data (creating BAM files): 1.3k people	weeks

Missing information:

How many CPUs/compute nodes were used?

Transferring data over a network

1 TB via 1 Gbps (gigabit per second) network: $\approx$ 2.2 h

theoretically...

More numbers: Costs of server-class hardware

The "mini cluster" of one of our clients

Year	Item	Price (€)
2009	new compute + storage server	30k
2010	add compute + storage server	21k
2013	add compute nodes + extra storage	34k
2016	replace compute + storage server	11k
2017	replace/increase storage	11k
2018	(?) increase storage + compute	~25k

More numbers: Costs of server-class hardware

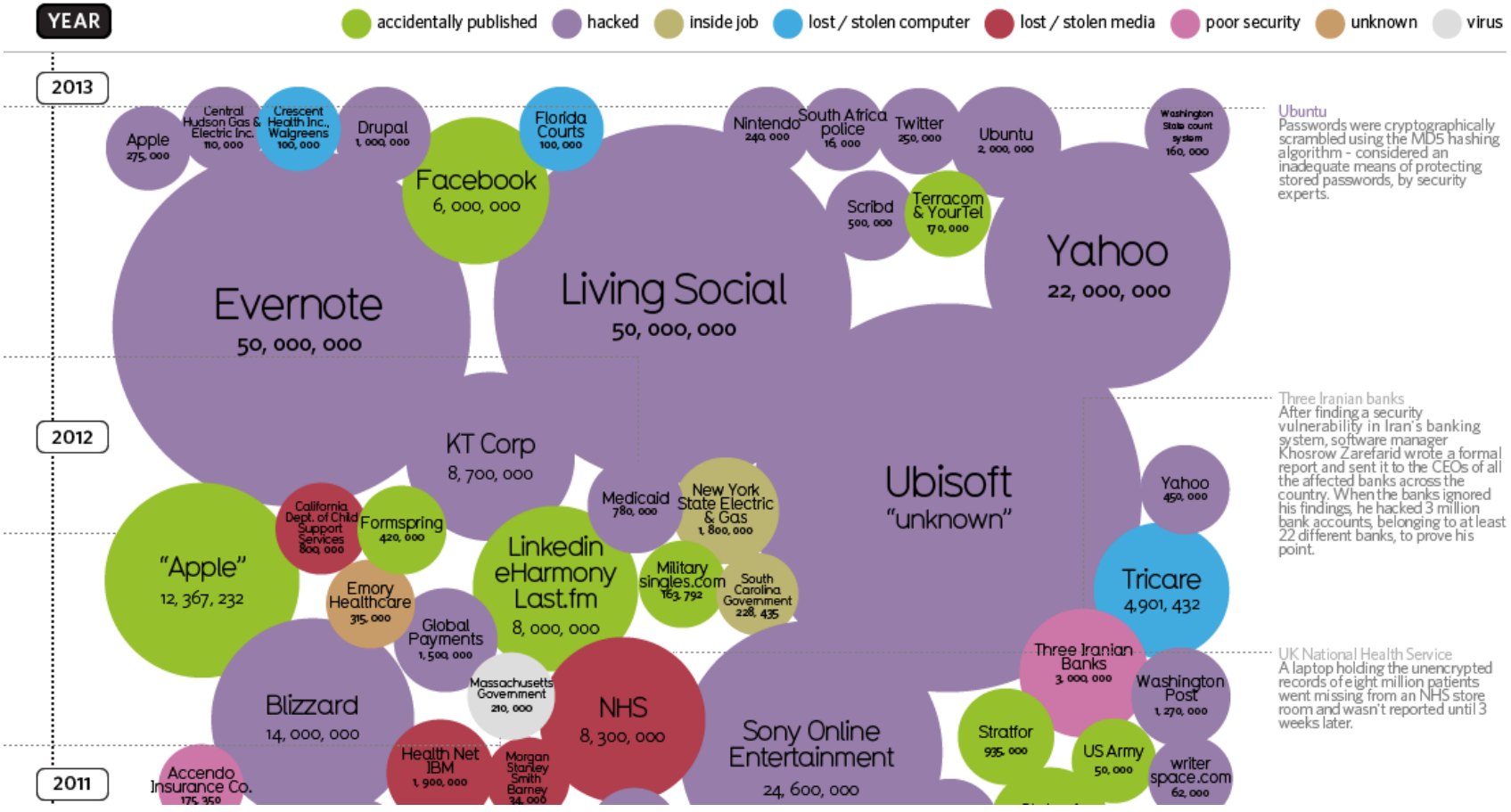
The "mini cluster" of one of our clients

Year	Item	Price (€)
2009	new compute + storage server	30k
2010	add compute + storage server	21k
2013	add compute nodes + extra storage	34k
2016	replace compute + storage server	11k
2017	replace/increase storage	11k
2018	(?) increase storage + compute	~25k

Excluding

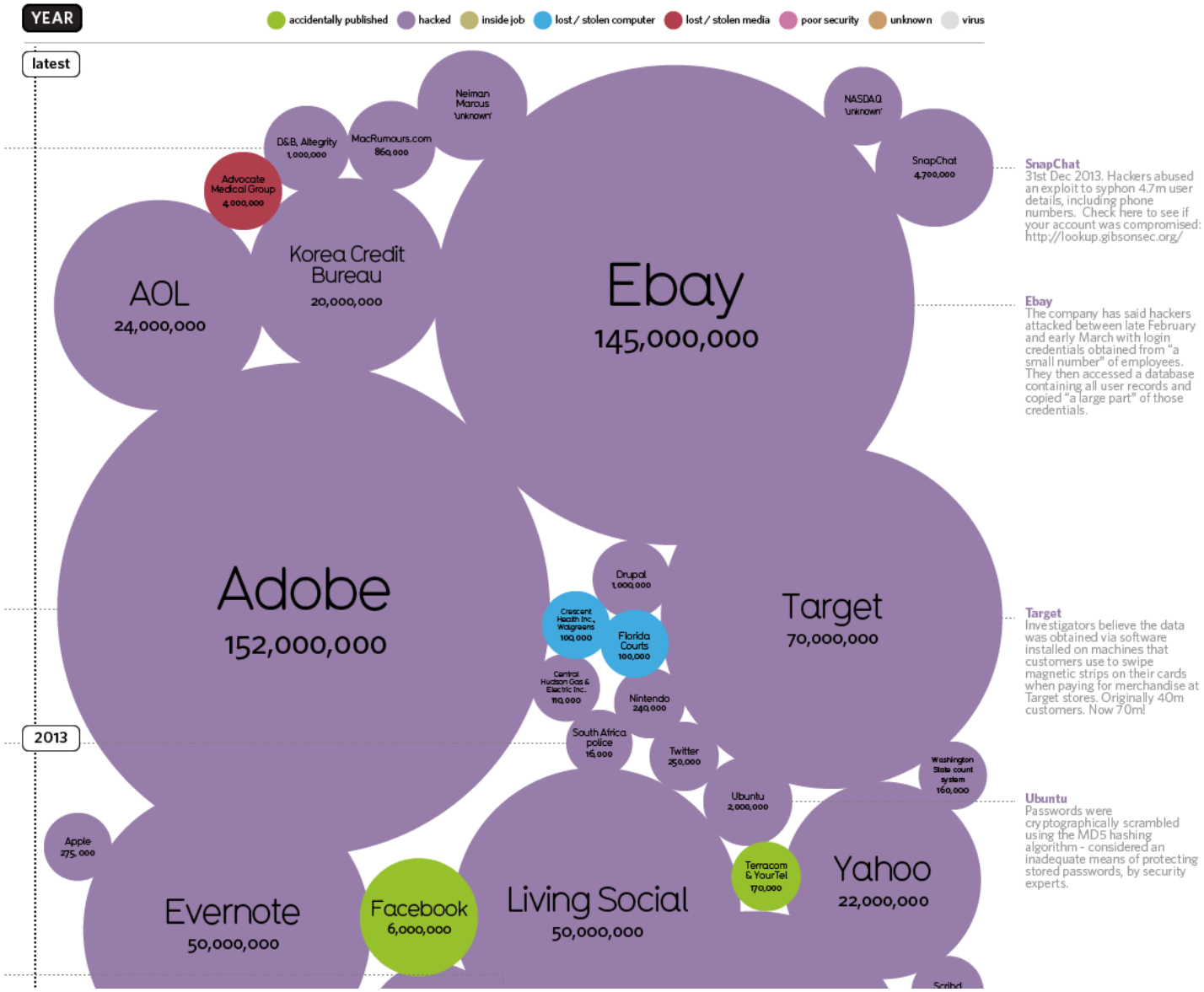
- installation costs
- maintenance costs
- data centre costs

Security: data breaches

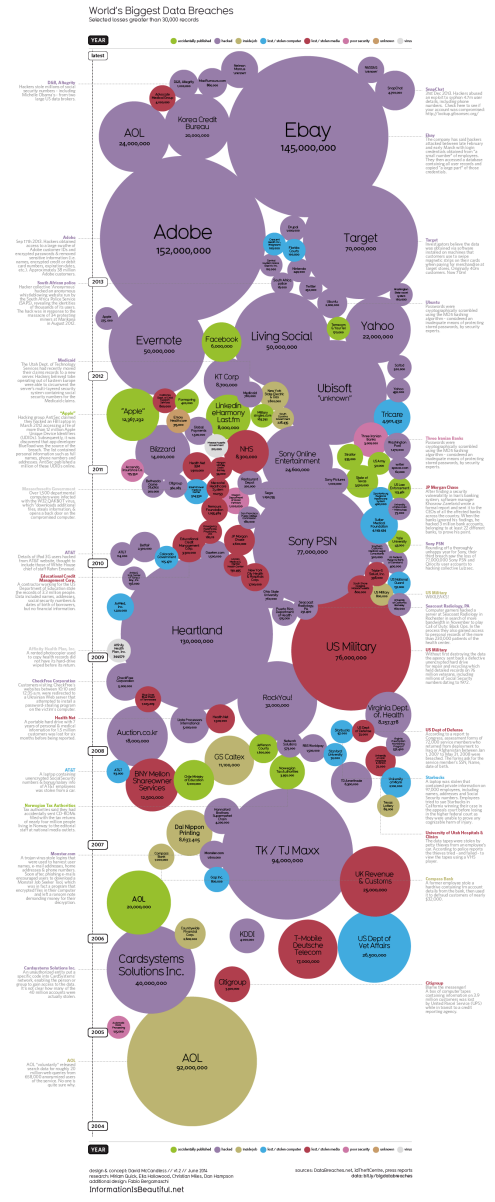


© Information is Beautiful.net

Security: data breaches



Security: data breaches



Security: data breaches Facebook/CA

Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

How Trump Consultants Exploited the Facebook Data of Millions

A HURRICANE FLATTENS FACEBOOK

Security: tips

- Secure connections: SSH, HTTPS
- Don't share accounts or passwords
- Use different, long, complicated passwords for each server/service/site
  - $\rightarrow$ use a password manager (e.g. KeePass, KeePassX, LastPass, ...)
- Use SSH keys for logging in to remote servers
- Keep names, addresses, post codes, etc. and the link to sample IDs separate in a safe location

Security: tips

- Secure connections: SSH, HTTPS
- Don't share accounts or passwords
- Use different, long, complicated passwords for each server/service/site
  - $\rightarrow$ use a password manager (e.g. KeePass, KeePassX, LastPass, ...)
- Use SSH keys for logging in to remote servers
- Keep names, addresses, post codes, etc. and the link to sample IDs separate in a safe location

Actually...

Unless you are the data manager: hand over to the data manager and delete it!

Security: tips (contd.)

- Store data only on servers or PCs at work

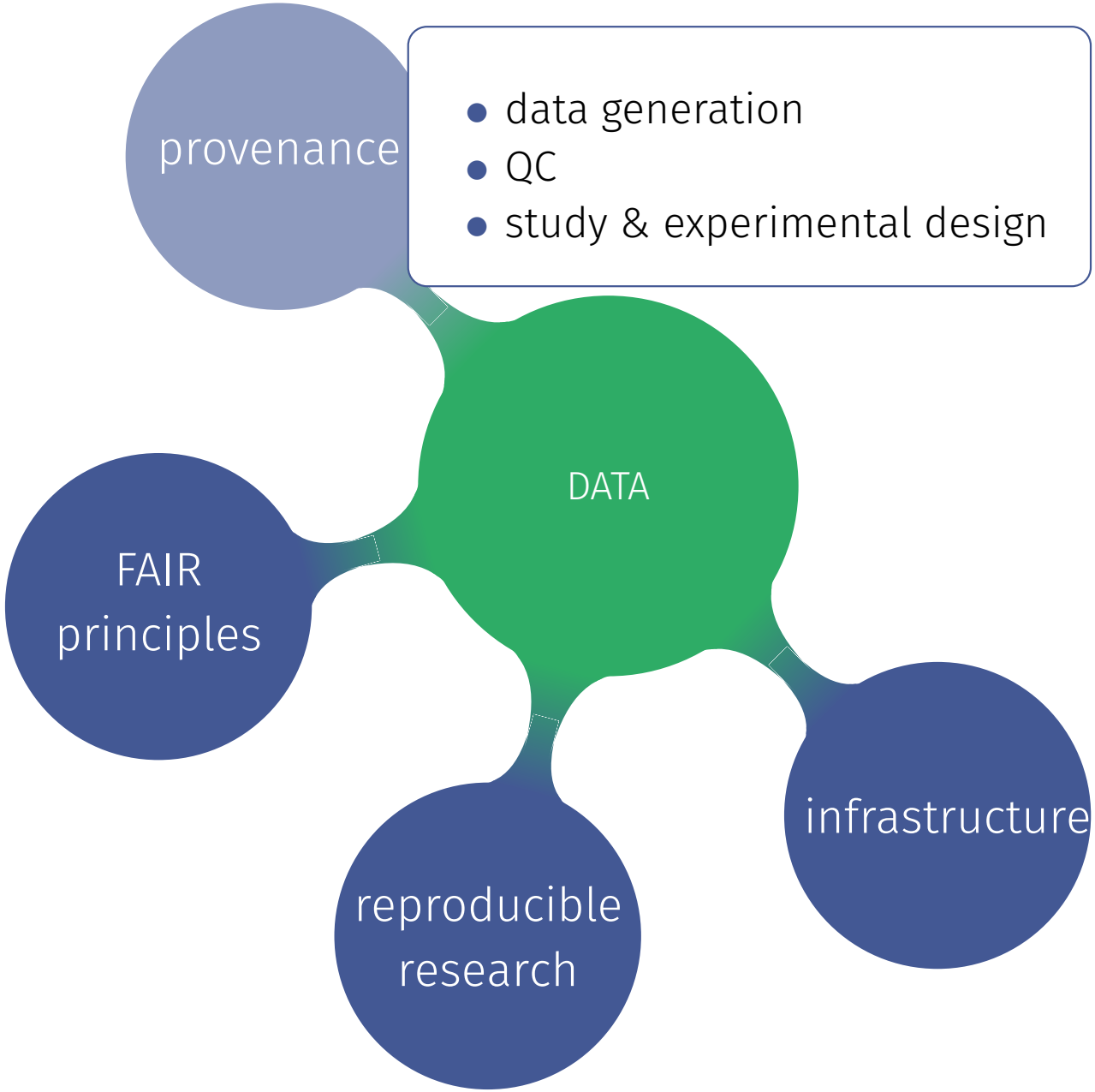
Security: tips (contd.)

- Store data only on servers or PCs at work

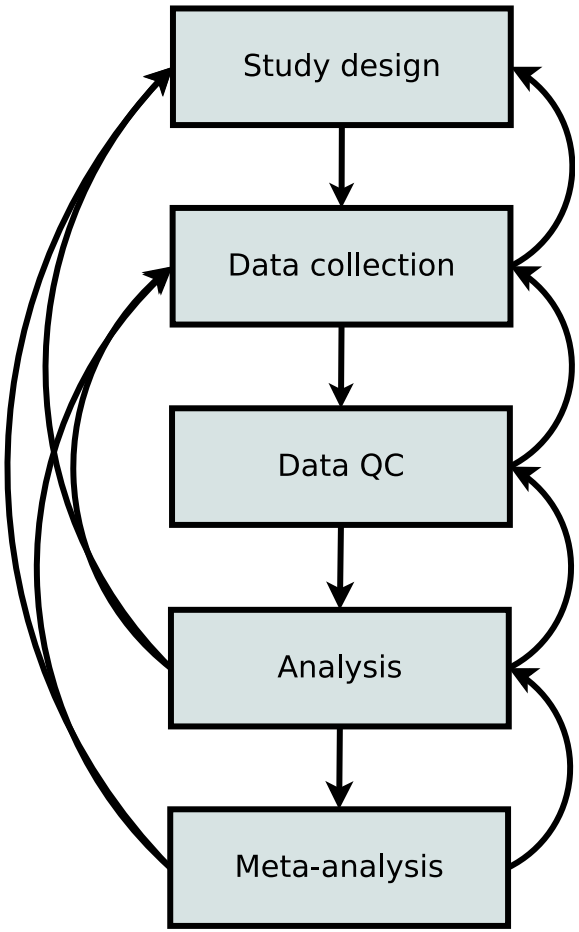
Please, not on...

- USB sticks
- External harddisks
- Laptops
- DropBox/MS OneDrive/Google Drive
  - You loose control:
    - delete != gone
    - you may give them the right to use and/or modify the data
    - service may shut down
  - ”There is not such thing as the cloud. Only somebody else’s computer”
- (unless sufficiently encrypted)

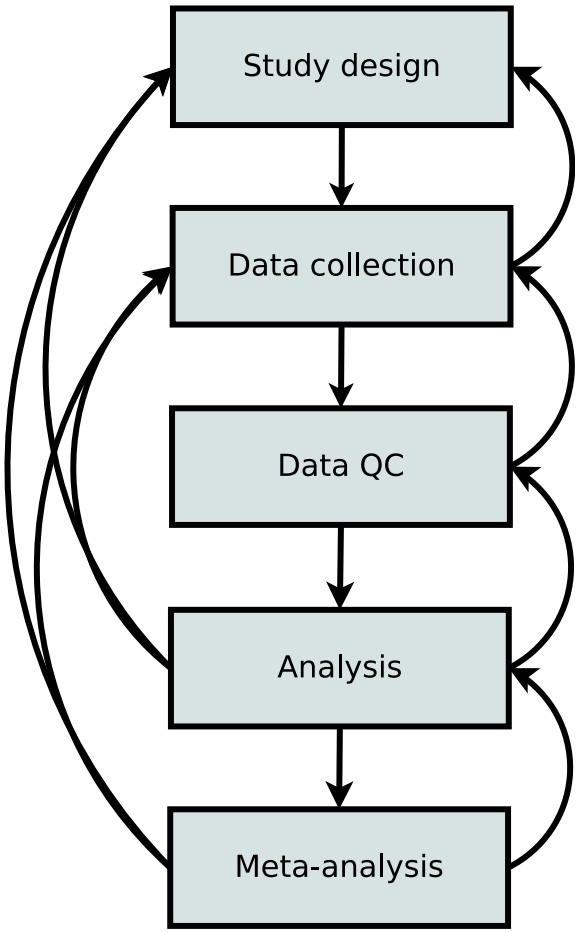
What is data stewardship?



A typical workflow in "omics epidemiology"



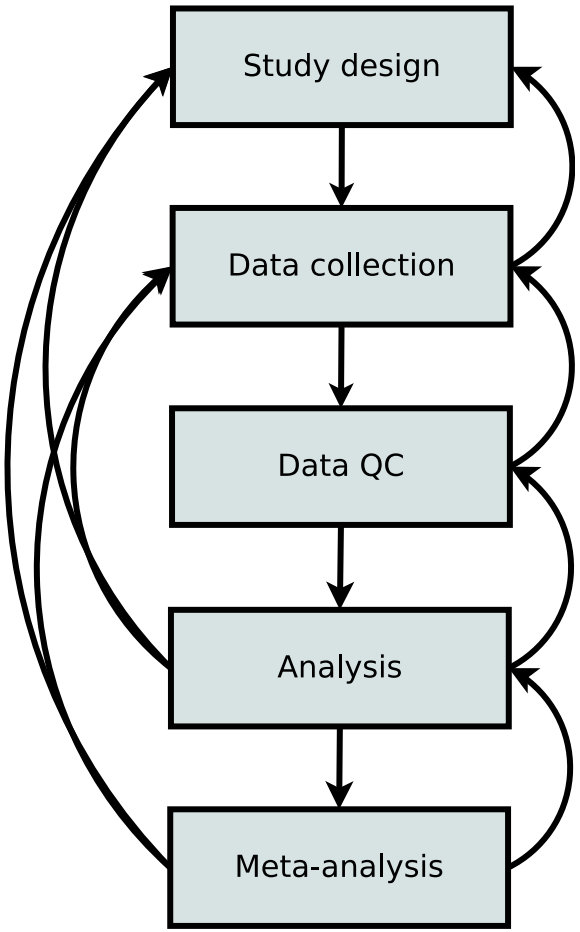
A typical workflow in "omics epidemiology"



Study design

- What data to collect? Where? From whom? For how long?
- Staff

A typical workflow in "omics epidemiology"



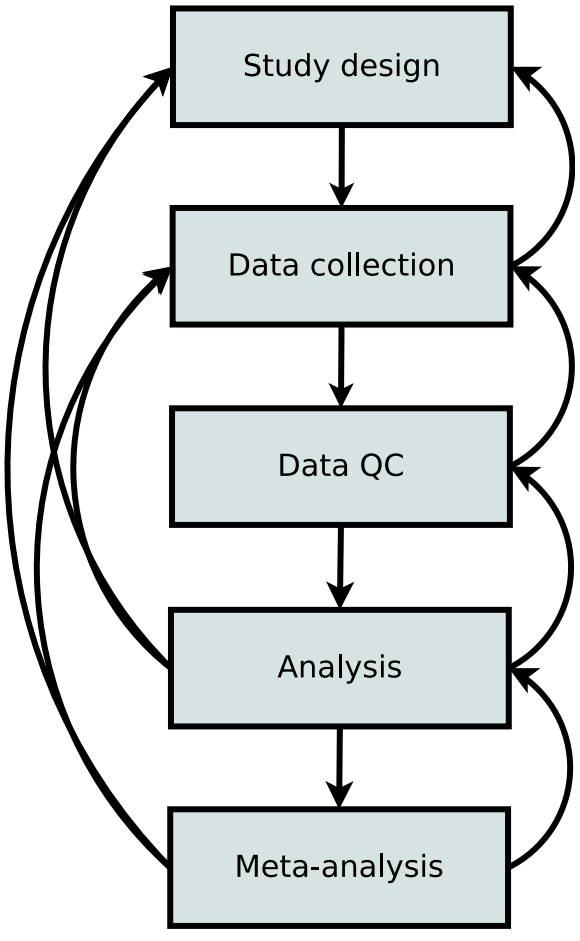
Study design

- What data to collect? Where? From whom? For how long?
- Staff

Data collection

- Plasma, MRI, blood pressure, questionnaires, ...
- Medical doctors, biologists, technicians, data manager, *you*?

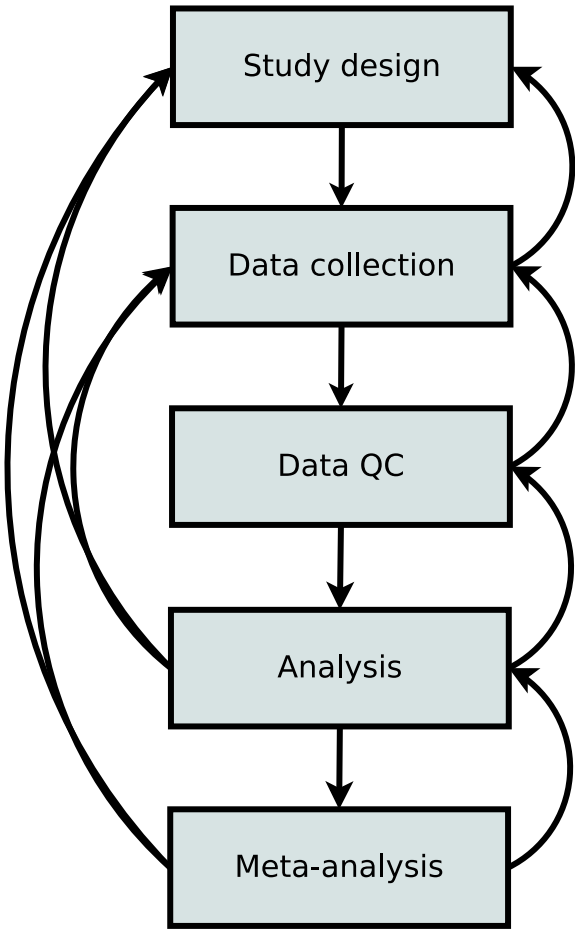
A typical workflow in "omics epidemiology"



Quality Control

- Remove outliers and other crazy values, check validity
- Units of measured values (!)
- Requires expert knowledge of the data
- Post-docs, data manager, *you*?

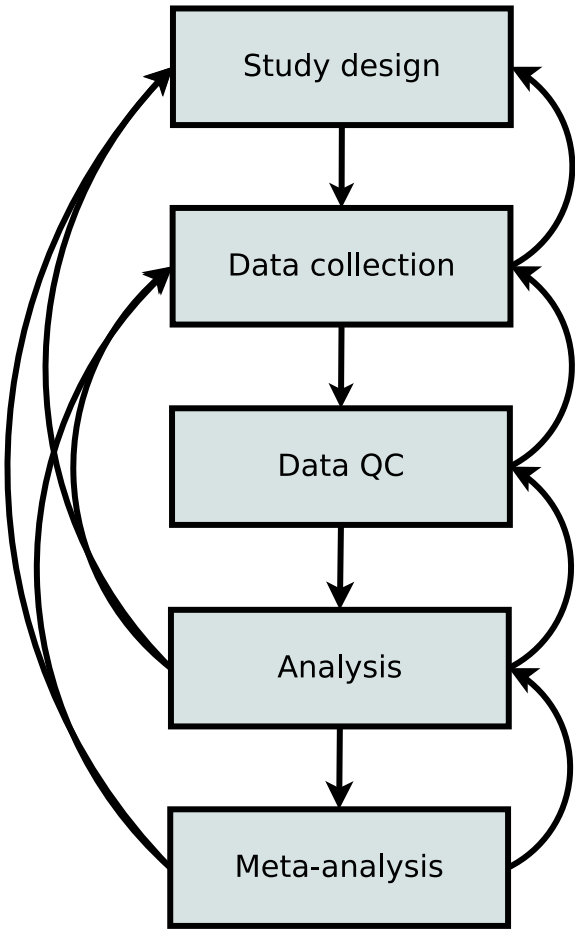
A typical workflow in "omics epidemiology"



Analysis

- use the data for research
- post-docs, PhD students, staff, *you*

A typical workflow in "omics epidemiology"



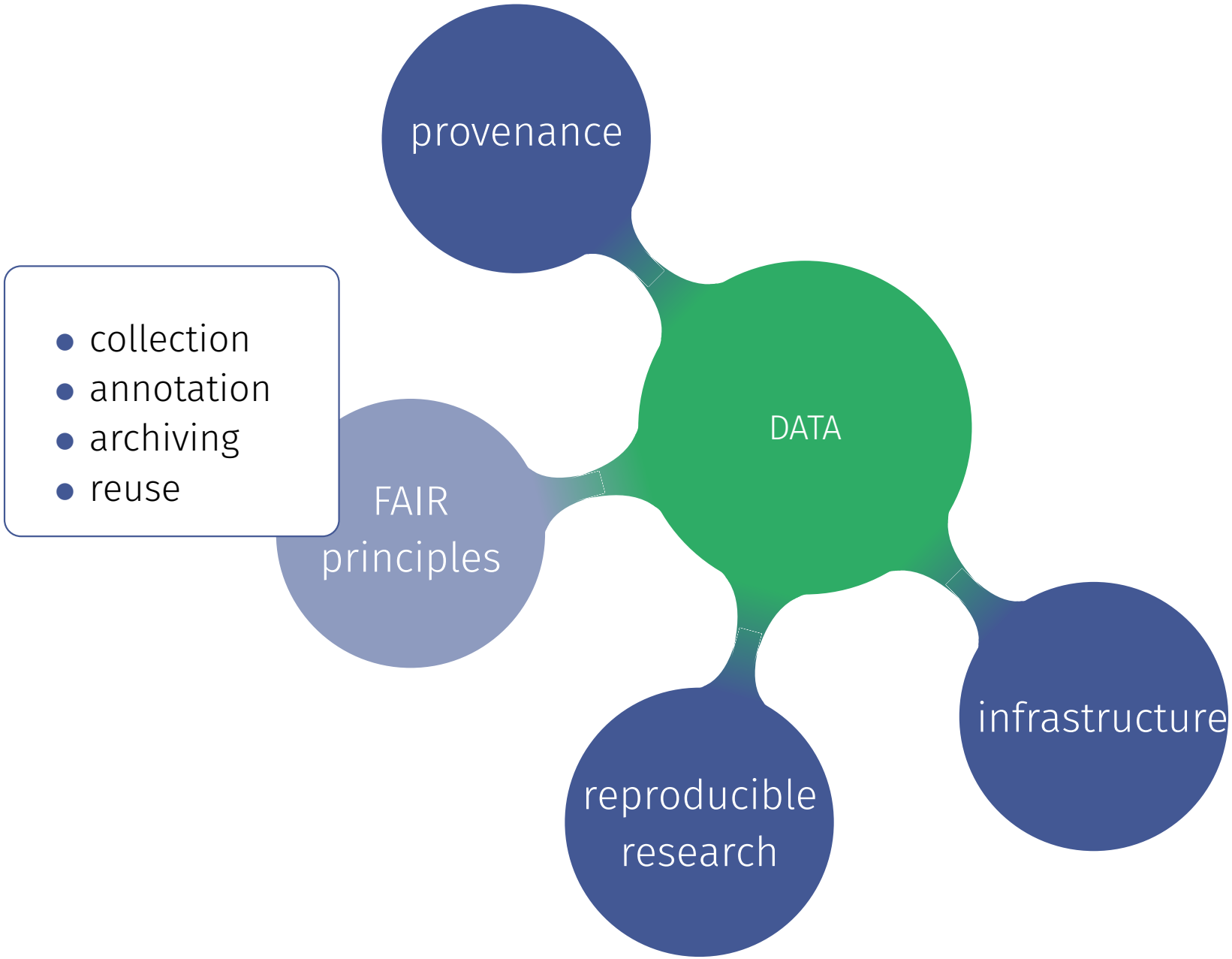
Analysis

- use the data for research
- post-docs, PhD students, staff, *you*

Meta-analysis

- combine data from various cohorts; harmonization
- post-docs, PhD students, *you*

What is data stewardship?



FAIR data

- **F** - Findable
- **A** - Accessible
- **I** - Interoperable
- **R** - Reusable

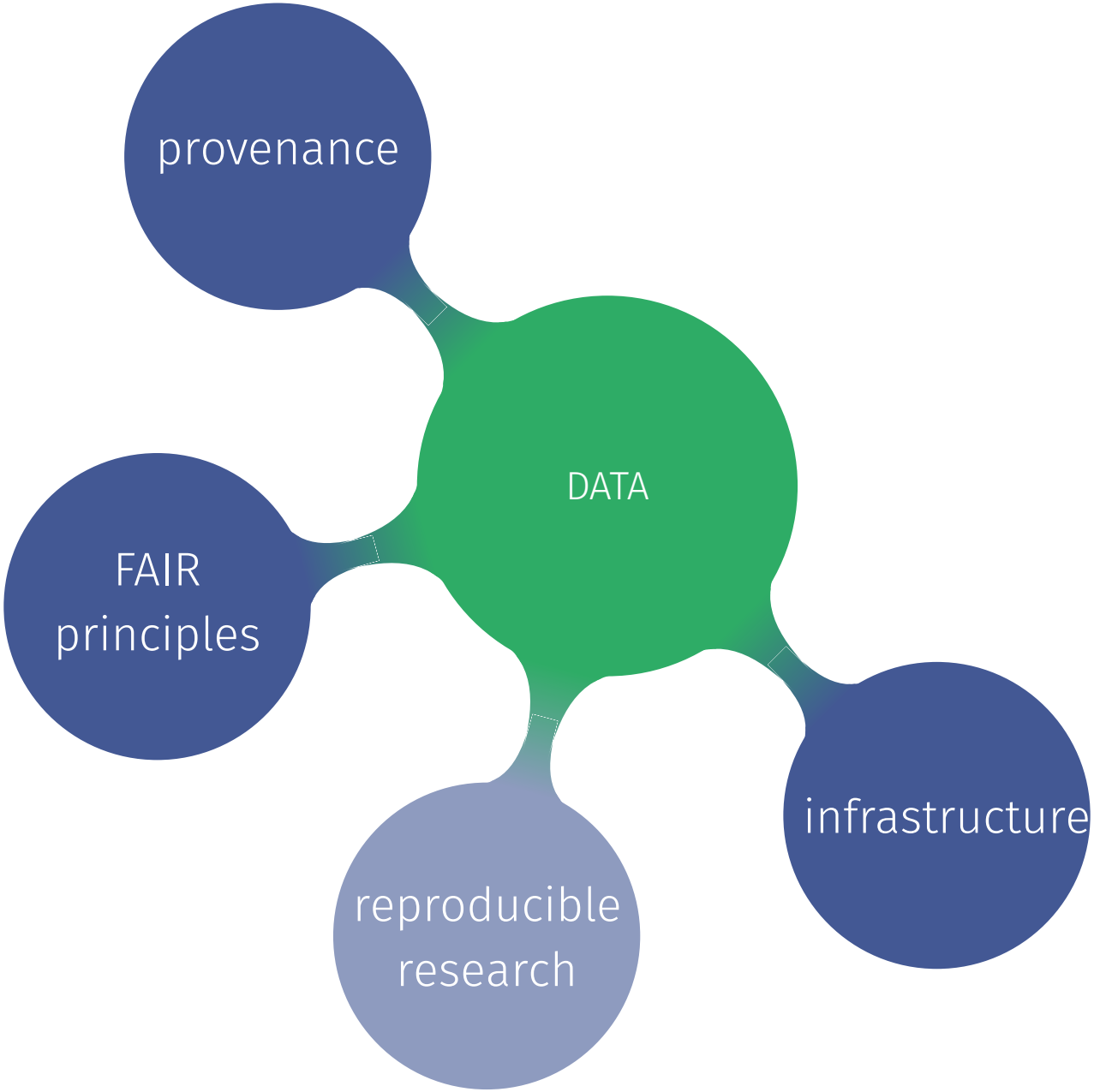
doi:10.1038/sdata.2016.18

"Good data management is **not a goal in itself**, but rather is the key conduit leading to **knowledge discovery** and **innovation**, and to subsequent **data and knowledge integration** and **reuse** by the community after the data publication process"

FAIR data (cont'd)

- emphasis on machine readability
- rich meta data
- open protocols/file formats
- links/references
- license

What is data stewardship?



Reproducible research

- Transparency (recent claims of data manipulation!)
- Whatever you do should be reproducible for
  - yourself
    - rerun with different parameters
    - after *X* months when writing a paper/thesis
  - your successor(s)
  - your supervisors
  - journals/your peers

Reproducible research

- Transparency (recent claims of data manipulation!)
- Whatever you do should be reproducible for
  - yourself
    - rerun with different parameters
    - after *X* months when writing a paper/thesis
  - your successor(s)
  - your supervisors
  - journals/your peers

How?

- Documentation
- Version control
- Literate programming

Ethics

(Recent) History

- 1939–1945: Nazi medical war crimes
  - Outcome: Nuremberg Code → First international code of research ethics
- 1948: UN Universal Declaration of Human Rights

Image source: Wikipedia

(Recent) History

1930–1972: Syphilis study at Tuskegee (Alabama, USA)

- $\sim$ 600 African-American men ($\sim$ 400 cases, $\sim$ 200 controls)
- No informed consent
- some procedures (e.g. spinal taps): ”special free treatment”
- continued even after treatment (penicillin) was found in the 1940s
- Outcomes: national policies, medical-ethical committees

(Recent) History

1944–1974: Cold war human radiation experiments

- > 400 experiments to determine the effects of radiation on human health; calibration of instruments for radiation detection
- Most studies involved minimal risks, appropriate informed consent
- However, some subject suffered injuries in studies that offered no prospect of direct benefit, or were presented as standard practice

1979 Belmont Report

Three principles essential to ethical conduct of research with humans

- Respect for persons
  - individuals should be treated as autonomous agents
  - persons with diminished autonomy are entitled to additional protections
- Beneficence
  - do no harm
  - maximise possible benefits, minimise possible harms
- Justice
  - individuals and groups should be treated fairly and equitably w.r.t. burdens and benefits of the research

What to keep in mind

- Consider:
  - vulnerable populations needing specific protection (e.g. children, prisoners)
  - risks associated in participation in research and appropriate protections
  - inclusion/exclusion criteria
- Informed consent
- Medical-ethical review board

Open questions for our field

Once we have blood/DNA...

- Broad consent?
- Explicitly ask permission for each research question?
- How about unforseen questions?
- How about DNA vs. privacy?
- Opt out?

Open questions for our field

Once we have blood/DNA...

- Broad consent?
- Explicitly ask permission for each research question?
- How about unforseen questions?
- How about DNA vs. privacy?
- Opt out?

Open data

- Should clinical trial data be open?
- Should biobank data be open?
- (and what does 'open' mean?)

EU General Data Protection Regulation (GDPR)

- enforceable from 25 May 2018
- single set of rules for all EU member states
- if consent is used as basis for processing:
  - must be explicit for data collected, and the purposes data is used for
  - consent for children must be given by the child's parent or custodian, and verifiable
  - data controllers must be able to prove "consent" (opt-in)
  - consent may be withdrawn
- data breach: data controller has to notify the authorities without undue delay
- right to erasure
- data portability: transfer personal data from one electronic processing system to and into another
- much more...

Conclusions/wrapping up

Why should YOU care about data stewardship?

Privacy

- Personal data from real people
- Genetic data: also affects their offspring
- Moreover, data leaks:
  - are bad publicity
  - may affect other people's willingness to participate
  - ...

Why should YOU care about data stewardship?

Privacy

- Personal data from real people
- Genetic data: also affects their offspring
- Moreover, data leaks:
  - are bad publicity
  - may affect other people's willingness to participate
  - ...

Cost of creating and maintaining a data set

- Sample collection (interviews, biosamples, ...)
- Curation
- Storage (not just digital, but freezers, cell lines, ...)

Why should YOU care about data stewardship?

High-quality research

- Understand your data
  - What was measured?
  - How was it measured?
  - What are reasonable values? Consult experts! Use units!
  - Understand data formats (e.g. VCF vs. plink)
- Understand your software: what does it do?
- Provenance
  - Where did the data come from?
  - Authoritative source?
  - What QC was done?

Conclusions

- Take good care of the data you work with (now and in the future)
- Understand your data: provenance, ethics
- Document what you do
- Talk to people → integration
- When you apply for grants, don't forget budget for:
  - data manager
  - bioinformatics/IT infrastructure

Thank you!

Any questions?

Thank you!

Any questions?