Data Literacy
Lecture 02
Managing Data

Kay Nieselt & Philipp Hennig

29 October 2019

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science
Department of Computer Science

Overview of lecture

- Data organization: data management plan
- data formats and conversion
- metadata
- reusage (FAIR data)
- storage and preservation
- big data

Recall: PPDAC

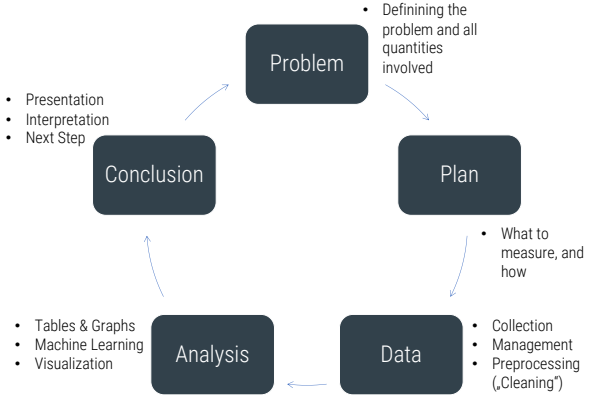
adapted from Spiegelhalter, 2019

Warm up: Do you recall what PPDAC was?

Recall: PPDAC

adapted from Spiegelhalter, 2019

Warm up: Do you recall what PPDAC was?

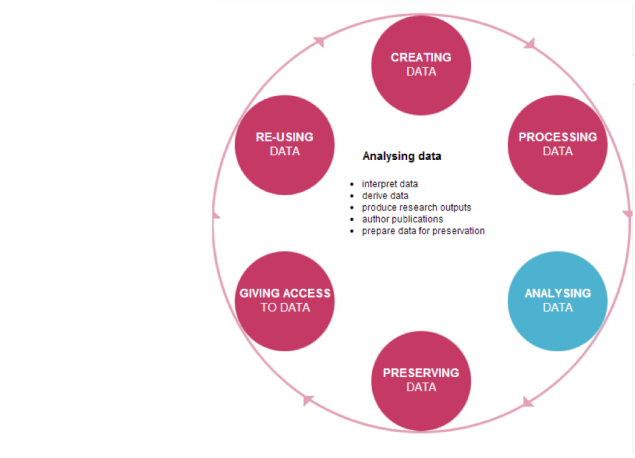


What is data management?

Actions that contribute to effective *storage*, *preservation* and *reuse* of data and documentation (metadata) throughout the research lifecycle

Image from `https://www.ukdataservice.ac.uk/manage-data`

Data management covers all elements of the data cycle



Data Management Plan

- Formal document
- Outlines what you will do with your data during & after you complete your research
- Ensures your data is safe for the present & the future

Funders like DFG have set up guidelines how to handle research data

From a Nature cover:

Everyone needs a data-management plan

They sound dull, but data-management plans are essential, and funders must explain why.

Keep your research data organized with a management plan. Credit: Jasper Juinen/Bloomberg/Getty

PDF version

RELATED ARTICLES

Data management made simple

High-quality science requires high-quality open data infrastructure

Don't let Europe's open-science dream drift

Empty rhetoric over data sharing slows science

SUBJECTS

DFG: see `https://www.dfg.de/en/research_funding/proposal_review_decision/applicants/research_data/index.html`)

Components of a General DMP

1. Information about data & data format
2. Metadata content and format
3. Policies for access, sharing and re-use
4. Long-term storage and data management
5. Maybe also: budget

Data Formats

Description of data to be produced:

- Experimental
- Observational
- Raw or derived
- Physical collections
- Models and their outputs
- Software
- Images
- Etc…

Data Formats

1. How will data be collected
   - when
   - where
2. How will data be analyzed/processed
   - software
   - algorithms
   - models
   - workflows

Programming Languages

Primary data science programming languages today are:

- Python: bread-and-butter programming language; interpreted language, contains libraries and features (e.g. regular expressions);

Programming Languages

Primary data science programming languages today are:

- Python: bread-and-butter programming language; interpreted language, contains libraries and features (e.g. regular expressions);
- Perl: used to be popular, no longer much in use

Programming Languages

Primary data science programming languages today are:

- Python: bread-and-butter programming language; interpreted language, contains libraries and features (e.g. regular expressions);
- Perl: used to be popular, no longer much in use
- R: programming language (mainly) of statisticians; visualisation; useful for exploration (versus Python useful for production)

Programming Languages

Primary data science programming languages today are:

- Python: bread-and-butter programming language; interpreted language, contains libraries and features (e.g. regular expressions);
- Perl: used to be popular, no longer much in use
- R: programming language (mainly) of statisticians; visualisation; useful for exploration (versus Python useful for production)
- Matlab: fast and efficient matrix operations; proprietary; there is also GNU Octave as an alternative

Programming Languages

Primary data science programming languages today are:

- Python: bread-and-butter programming language; interpreted language, contains libraries and features (e.g. regular expressions);
- Perl: used to be popular, no longer much in use
- R: programming language (mainly) of statisticians; visualisation; useful for exploration (versus Python useful for production)
- Matlab: fast and efficient matrix operations; proprietary; there is also GNU Octave as an alternative
- Java/C/C++: mainstream programming languages; language for Big Data systems; Hadoop and Spark are based on Java and C++, resp.

Programming Languages

Primary data science programming languages today are:

- Python: bread-and-butter programming language; interpreted language, contains libraries and features (e.g. regular expressions);
- Perl: used to be popular, no longer much in use
- R: programming language (mainly) of statisticians; visualisation; useful for exploration (versus Python useful for production)
- Matlab: fast and efficient matrix operations; proprietary; there is also GNU Octave as an alternative
- Java/C/C++: mainstream programming languages; language for Big Data systems; Hadoop and Spark are based on Java and C++, resp.
- Mathematica: symbolic math; proprietary

Programming Languages

Primary data science programming languages today are:

- Python: bread-and-butter programming language; interpreted language, contains libraries and features (e.g. regular expressions);
- Perl: used to be popular, no longer much in use
- R: programming language (mainly) of statisticians; visualisation; useful for exploration (versus Python useful for production)
- Matlab: fast and efficient matrix operations; proprietary; there is also GNU Octave as an alternative
- Java/C/C++: mainstream programming languages; language for Big Data systems; Hadoop and Spark are based on Java and C++, resp.
- Mathematica: symbolic math; proprietary
- Excel: spreadsheet program; proprietary; a large amount of (hidden) functionality

Notebooks

What is the primary delivery of a data science project?

Notebooks

What is the primary delivery of a data science project?
No,
not the program,

Notebooks

What is the primary delivery of a data science project?
No,
not the program,
not the data set,

Notebooks

What is the primary delivery of a data science project?
No,

- not the program,
- not the data set,
- not the report, ..

Notebooks

What is the primary delivery of a data science project?
No,

not the program,
not the data set,
not the report, ..
... it should be a computable notebook

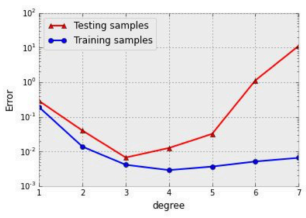
Notebooks

Figure from Skiena "The Data Science Manual"

Notebooks tie together:
Data, code, computed results, written analysis
Because projects should be:

- *reproducible:* when rerun produce the same result
- *tweakable:* allow change of parameters or algorithms
- *documented:* in notebooks text, code and visualisations can be combined

In [40]:
```
degrees = range(1, 8)
errors = np.array([regressor3(d) for d in degrees])
plt.plot(degrees, errors[:, 0], marker='^', c='r', label='Testing samples
plt.plot(degrees, errors[:, 1], marker='o', c='b', label='Training sample
plt.yscale('log')
plt.xlabel("degree"); plt.ylabel("Error")
  = plt.legend(loc='best')
```



By sweeping the degree we discover two regions of model performance:

- **Underfitting** (degree < 3): Characterized by the fact that the testing error will get lower if we increase the model capacity.
- **Overfitting** (degree > 3): Characterized by the fact the testing will get higher if we increase the model capacity. Note, that the training error is getting lower or just staying the same!.

Notebooks and pipelines

Notebooks make it easier to maintain data pipelines, the sequence of processing steps from start to finish.
You may have to redo your analysis from scratch (in fact you should expect this), so build your code accordingly.

Image from: Postmedia, https://business.financialpost.com/commodities/energy/pipeline-relief-on-the-horizon-for-oilpatch-with-surprise-moves-to-boost-capacity-through-existing-lines

Notebooks

Notebooks

Jupyter/IPython Notebook is based on a set of open standards for interactive computing. It uses an open document format based on JSON.

Data/File formats

all these standards

Properties:

- easy to parse
- human readable (depends on the data)
- widely used

Data/File formats

Standards

- CSV: for tables like spreadsheets (mainly for numbers and statistics)
- XML: for structured non-tabular data
- SQL: for multiple related tables
- JSON: Javascript Object Notation for APIs

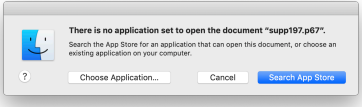
But also:

- Moving Images: MOV, MPEG-4
- Audio: WAVE, MP3
- Images: SVG, TIFF, JPG 2000
- Text: PDF/A, txt, OpenDocument Text

Data formats

Data formats that offer the best chance for long-term access are:

- Non-proprietary (also known as open), and
- Unencrypted & uncompressed



Data formats

Information About Data Types & Data Format(s)

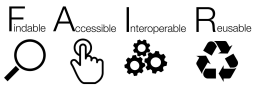
- Existing data
  - If existing data are used, what are their origins?
  - Will your data be combined with existing data?
  - What is the relationship between your data and existing data?
- How data will be managed in short-term
  - Version control
  - Backing up
  - Security & protection
  - Who will be responsible

FAIR data versus fair data

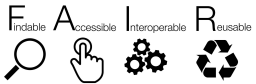
Fairness versus FAIR Data Principles

Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18

FAIR = **F**indable **A**ccessible **I**nteroperable **R**eusable



FAIR data

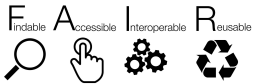


FAIR = **F**indable **A**ccessible **I**nteroperable **R**eusable
Formulated altogether 15 principles and 14 metrics.

Findable:

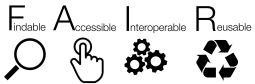
F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

FAIR = **F**indable **A**ccessible **I**nteroperable **R**eusable



Accessible:
A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

FAIR = **F**indable **A**ccessible **I**nteroperable **R**eusable



Interoperable:

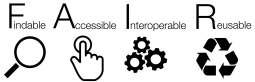
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

FAIR data

FAIR = **F**indable **A**ccessible **I**nteroperable **R**eusable



Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

In 2019 the Global Indigenous Data Alliance (GIDA) released the CARE Principles:

Collective benefit, *Authority* to control, *Responsibility*, and *Ethics* to ensure data guidelines address historical contexts.

Read more here: `https://www.gida-global.org/care`

Understand data also in 20 years

Metadata = data reporting / documentation
Thus: Metadata is data about data. It describes the content, quality, condition, and other characteristics of a dataset.

Who	created the data?
What	is the content of the data?
When	were the data created?
Where	is it geographically?
How	were the data developed?
Why	were the data developed?

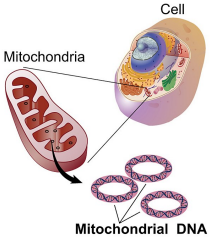
Metadata in real life

Image from: Wikipedia

Aim: Collect and set up a platform to manage and analyse all publicly available human mitochondrial genomes

Data collection: 48,600 genomes available

International consortium was founded about 3 years ago.



Background information: Mitochondrial DNA is the small circular chromosome found inside mitochondria. These organelles found in cells have often been called the powerhouse of the cell. Mitochondrial DNA are passed almost exclusively from mother to offspring through the egg cell.

Metadata in real life

The actual data: about 16,569 base pairs encoding 37 genes

Here are the first couple of hundred bases:

```
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
```

Metadata in real life

The actual data: about 16,569 base pairs encoding 37 genes

Here are the first couple of hundred bases:

GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA

Metadata (entries of the database) has grown to contain 72 attributes (e.g. accession_id, sex, age, language, population, geographic info, publication, technical info, ...) An excerpt from the metadata

	A	B	C	D	
	Attribute	Description / Comments	Possible answers (nothing else allowed)	Example
Population	population	Describe the sample with a meaningful identification. We are aware of the difficulty of defining a population, but we want to try to describe the samples as detailed as possible. Please use either ethnolinguistic or geographic criteria. Use broad descriptors for medical cohorts. Refer to the terms used in the original publication if possible. Consider avoiding denigratory terms or culturally sensitive terms. Please use english descriptors.		Dutch
Publication	doi	digital object identifier of the publication		10.1126/sciadv.15
	author	**ONLY first author** of the paper, following the PubMed citation style. Surname followed by first capital letter of given name(s), separated by space. Several given names are not separated.		Haak W
	publication_date	Date of publication, (year)	integer	2016
	title	Title of the publication.		Ancient mitochond
	journal	In which journal was the data published?		Science
	publication_type	What kind of publication is it?	paper / peerPrint / direct submission to genbank / direct submission to mitoDB / article	paper
	publication_status	published or not?	published / protected / private / in press / in preparation / submitted	published
	publication_comments	can link information to a newer version		
Technical Info	mt_sequence	Sequence itself, given as separate fasta file, or included in the excel upload file. The sequence has to follow the GenBank submission guidelines. The sequence must not contain '-'.	String, not aligned	ATGCCATG...
	percentage_N	Percentage of the genome reconstructed	double	80
	completeness	How complete is the sequence? We set a threshold for modern data to 99%, for ancient data to 98% Caclulation: total length of sequences mapped against rCRS / number of N's	double	99
	tissue_sampled	Which tissue was sampled?		bone
	sampling_date	Date when the sample was sampled.	integer	2014
	sequencing_platform	Sequencing platform that was ues to sequence the samples.	Illumina / 454 / sanger / nanopore / pacbio / affymetrix	Illumina

Access and sharing

Data management plan should describe policies for access, sharing, and reuse of data.

- Property issues
- Citation (how should data be cited if used)
- Access

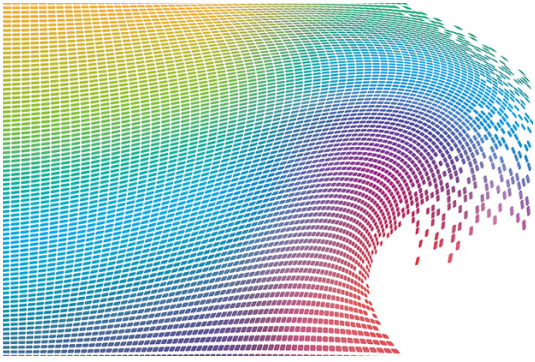
Access and sharing

- What data should be archived? (Mainly raw data)
- Where should data be archived?
  - General repositories: e.g.

    Dryad (`https://datadryad.org/stash/`
  - Community standards

Big data

Illustration: Carl DeTorres, via IEEE Spectrum



(Read more in ‘The DNA Data Deluge’ by MC Schatz and B Langmead, IEEE Spectrum 2013).

Big data: 3 Vs

Generally we speak of *Big Data*, if it is

- large and
- heterogeneous and complex and
- generated within a very short period of time

Big data: 3 Vs

Generally we speak of *Big Data*, if it is

- large and
- heterogeneous and complex and
- generated within a very short period of time

Big data: data is a combination of Volume, Variety, Velocity (3 Vs)

> **"Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it." [Edd Dumbill, O'Reilly]**

Big data: 3 Vs

Generally we speak of *Big Data*, if it is

- large and
- heterogeneous and complex and
- generated within a very short period of time

Big data: data is a combination of Volume, Variety, Velocity (3 Vs)

> **"Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it." [Edd Dumbill, O'Reilly]**

Big data: 3 Vs

Generally we speak of *Big Data*, if it is

- large and
- heterogeneous and complex and
- generated within a very short period of time

Big data: data is a combination of Volume, Variety, Velocity (3 Vs)

> **"Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it." [Edd Dumbill, O'Reilly]**

Note: actually sometimes big data is referred to as the 5Vs:

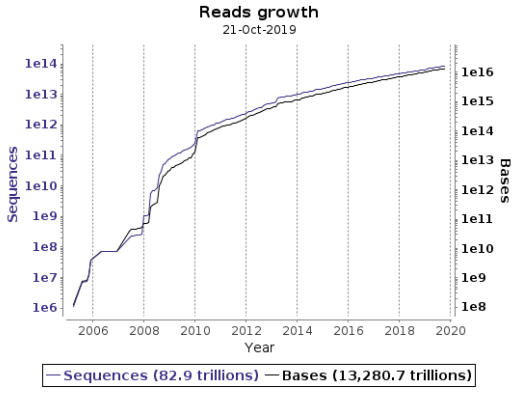
- Veracity/Validity - how reliable is the data
- Value - how much value can be gained from the data

Big data: examples

- Twitter: Until 2014: $1.7 \cdot 10^{10}$ tweets archived by the largest library of the world, the Library of Congress. This has been stopped since January 1, 2018.
- Large Hadron Collider: $2 \cdot 10^{17}$ bytes analysed for the detection of the Higgs particles.
- Sloan Digital Sky Survey: image material with more than $3 \cdot 10^{12}$ pixels (data all publicly available)
- AT&T has $242 \cdot 10^{15}$ bytes (petabytes) of data traffic every day in its networks

Big data: examples

Another example from bioinformatics / genomics:



Dramatic cost reduction and simultaneous x-fold speed increase have led to an explosion of raw data.
Today: more than 13 petabytes of bases.

Managing big data

”small data”:

- flat files
- RDBMs (MySQL, PostGreSQL, Oracle)

New solutions for saving big data:
Highly performant, scalable, distributed access

- Data formats: hierarchical data format (HDF), latest version is HDF5 (.hf5, .h5, .he5), faster than SQL
- Hadoop Distributed File System (Apache)
- NoSQL databases

Why data management:

Manage your data for yourself:

- Keep yourself organized (recall Philipp Hennig's suggested folder structure)
- Use notebook environment
- Track your analyses / research to be reproducible
- Control versions of data (use git)
- Quality control your data
- Make backups
- Format data for re-use
- Prepare data to share it

Summary

- Data management is an important part of the data life cycle.
- A data management plan saves time and ensures that data are relevant and useful for others.
- Major components of a DMP are:
  - Information about data & data format
  - Metadata content and format
  - Access & sharing
- Specific data formats and systems have been developed for big data

References

Some of the material of this lecture has been taken from the following sources:

- Whitmire, Amanda L. (2014). Research Data Management Curriculum. Oregon State University Libraries. Retrieved Oct 26, 2018 from: http://guides.library.oregonstate.edu/grad521Lectures.
- Steven S Skiena (2017) The Data Science Design Manual, Chapter 3.