University of BRISTOL

22/02/2017

Applied Data Science

Data Wrangling

Dr. Niall Twomey

niall.twomey@bristol.ac.uk

University of Bristol

Recap and outlook

Lec1 Data formats

- Data types in Python (`list`, `dict`, `set`)
- Serialisation (CSV, JSON, HDF5)

Lec3 Databases

- Relational: SQL
- NoSQL: MongoDB
- Graph: Neo4j

Recap and outlook

Lec1 Data formats

- Data types in Python (`list`, `dict`, `set`)
- Serialisation (CSV, JSON, HDF5)

Lec3 Databases

- Relational: SQL
- NoSQL: MongoDB
- Graph: Neo4j

Lec4 **Data wrangling**

- (Neo4j: continued from last week)
- Demonstration
- Cleaning data
- Missing data

Neo4j

Last week

- Created nodes
- Created relationships
- Performed simple queries
- Neo4j demo

https://neo4j.com/graphgists/

https://neo4j.com/graphgist/34b3cf4c-da1e-4650-81e4-3a7107336ac9

https://neo4j.com/graphgist/64507854-323f-478e-a3c3-2b0e6ab694eb

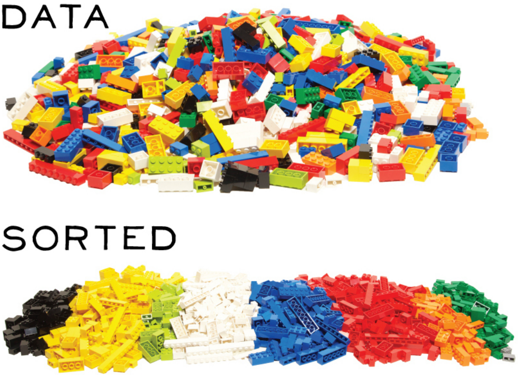
Recurring theme in data science lectures...

- There is no single correct recipe for data science
  - Data storage
  - Database selection
  - Data wrangling
- Match appropriate technique to the task at hand
- Familiarity is essential to select suitable methods

- Data wrangling: detect, correct, remove or otherwise deal with corrupted or inaccurate records

http://egallic.fr/R/intro_carto/images/lego_data_1.jpg

- Data wrangling: detect, correct, remove or otherwise deal with corrupted or inaccurate records



http://egallic.fr/R/intro_carto/images/lego_data_1.jpg

Demo 1: `pandas`

Demo 2: Data Wrangler

http://pandas.pydata.org/

http://casas.wsu.edu/datasets/

http://casas.wsu.edu/datasets/twor.summer.2009.zip

http://vis.stanford.edu/wrangler/

Preprocessing

- We have seen some techniques in data wrangling in the demos
- Here are some general strategies:
  - Type screening
  - Range check
  - Illegal values
  - Robust checks: regular expressions
    - Phone numbers
    - Email addresses
    - Dates
  - Multi-column validation
  - Check unique/distinct values
  - Remove duplicates (need to define match criteria)
- Caution with NoSQL databases (including MongoDB, Neo4j)

`http://www.regexlib.com/`

Random samples from database

- SQL

```
SELECT column FROM table ORDER BY RAND() LIMIT 1
```

Random samples from database

- SQL

```
SELECT column FROM table ORDER BY RAND() LIMIT 1
```

- MongoDB

```
db.collection.aggregate({$sample: { size: 1 }})
```

Random samples from database

- SQL

```
SELECT column FROM table ORDER BY RAND() LIMIT 1
```

- MongoDB

```
db.collection.aggregate({$sample: { size: 1 }})
```

- Neo4j

```
MATCH (n) RETURN n SKIP <random_number> LIMIT 1
```

Missingness (I)

- Data can be missing for several reasons
  - Some people will feel uncomfortable filling out questionnaires fully (*e.g.* salary)
  - Some parts of questionnaires is not relevant (*e.g.* census)
  - Communication link may have disconnected
  - Data may be censored (*e.g.* health applications)
  - Data may be corrupted
  - Information may simply not be known

Missingness (I)

- Data can be missing for several reasons
  - Some people will feel uncomfortable filling out questionnaires fully (*e.g.* salary)
  - Some parts of questionnaires is not relevant (*e.g.* census)
  - Communication link may have disconnected
  - Data may be censored (*e.g.* health applications)
  - Data may be corrupted
  - Information may simply not be known
- Regardless of reason for missing data, it is important to deal with missingness

Missingness (II)

- Assumptions:
  - Set of mandatory columns
  - Have access to (validated) external information (optional; see fusion lecture next week)
- Possibly methods of dealing with missing data:
  - Model-based approaches
  - Nearest neighbour
  - Mean/median/mode imputation
  - ‘Missing data’ indication feature

Ghahramani, Zoubin, and Michael I. Jordan. ”Supervised learning from incomplete data via an EM approach.” Advances in neural information processing systems (1994): 120-120.

Smola, Alexander J., S. V. N. Vishwanathan, and Thomas Hofmann. ”Kernel Methods for Missing Variables.” AISTATS. 2005.

Zheng, Fei, and Geoffrey I. Webb. ”Tree augmented naive Bayes.” Encyclopedia of Machine Learning. Springer US, 2011. 990-991.

Summary

- The ‘garbage in, garbage out’ mantra is particularly relevant in data science
- Data always needs to be evaluated and validated before analysis
- Evaluating the quality of data cleaning is difficult since gold standard databases normally do not exist

- Missing data is an unavoidable reality
- Many applications will benefit from techniques to deal with missingness
- Experimentation is often required to determin the most appropriate solution

Resources

- McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython." O'Reilly Media, Inc.", 2012.
- Kotsiantis, S. B., D. Kanellopoulos, and P. E. Pintelas. "Data preprocessing for supervised leaning." International Journal of Computer Science 1.2 (2006): 111-117.
- Blog article on data readiness levels: `inverseprobability.com/2017/01/12/data-readiness-levels`
- For time series data (*e.g.* physiological signals) see digital filters: `https://en.wikipedia.org/wiki/Digital_filter`
- Applied data science github page (will be updated with the content of today's lecture this afternoon): `https://github.com/njtwomey/ADS`