Inf2 – Foundations of Data Science
S2 Week 7: Software engineering for data science

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS
OF
DATA
SCIENCE

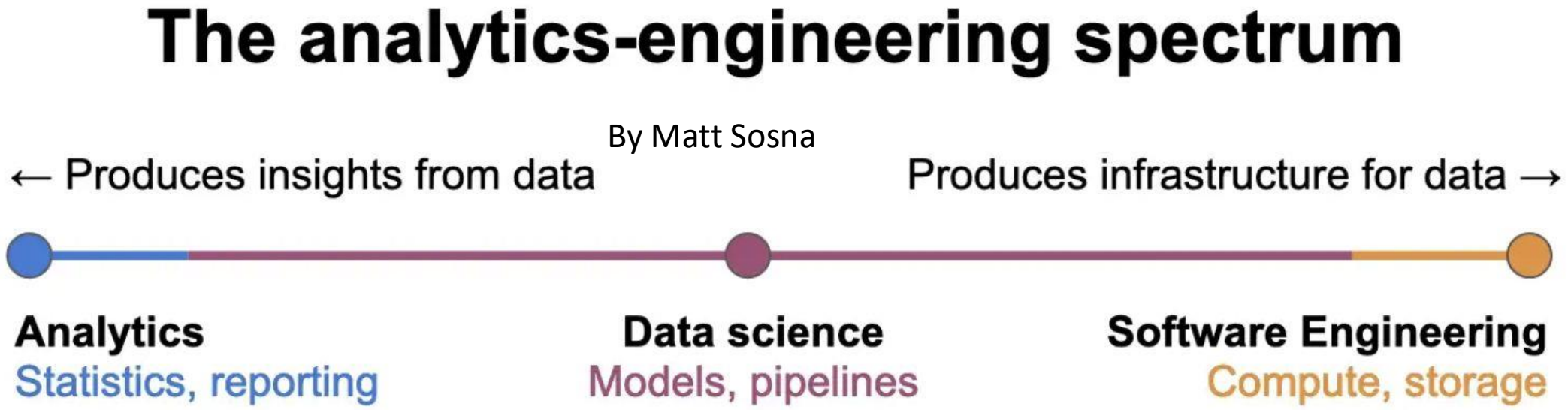
Announcements

- There will be a mock exam provided
- Week 8 – writing workshop
- Week 9 – presentation on student homelessness data
- Week 10 - (TBC) revision
- Office hour Monday at 4pm
- Questions on project now?

Overview

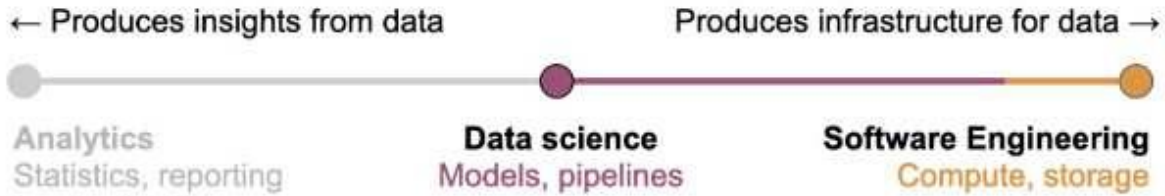
- What is a "data scientist" and what scale of data science?
- Why is software engineering needed for data science?
- Reproducibility
- Data management and code management
- Efficiency and scaling

What is a data scientist?



Matt Sosna, How to enter data science, https://mattsosna.com/DS-transition-1

Data science job for US electric car maker

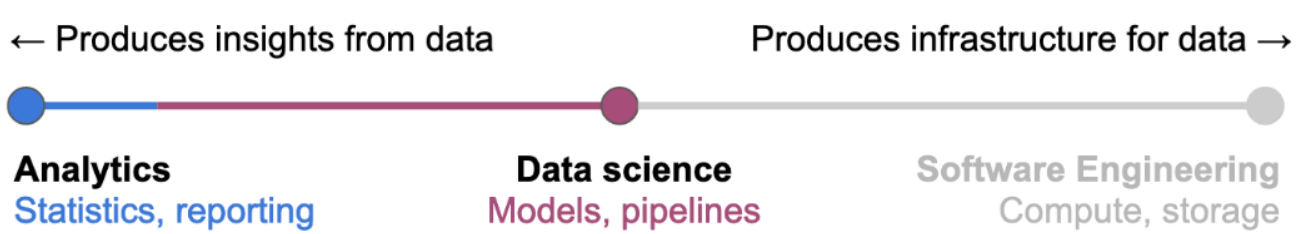


Required Skills

- Extensive experience writing software with Python
- Experience with multiple data architecture paradigms (e.g. SQL, NoSQL, Kafka, Spark)
- Experience and interest in frontend development, preferably with the Javascript React framework
- Knowledge of various data communication protocols (e.g. REST API, Websockets)
- Able to work under pressure while collaborating and managing competing demands with tight deadlines
- A passion and curiosity for data and data-driven decision making
- Experience with open source machine learning libraries and frameworks (Tensorflow, Keras, etc)
- Familiarity with continuous integration pipelines (Docker, Jenkins, Kubernetes)
- Drive to introduce a predictive model to a production environment
- Success building and tuning image classification models
- MS in engineering, physics, mathematics, or equivalent.
- 3 - 5 years relevant experience.
- Have high attention to details.
- Be a team player and have the ability to collaborate well across diverse functional groups
- Strong verbal and written communication skills to manage and communicate the health and integrity of the data and systems.
- Experience in high volume manufacturing is a plus

Matt Sosna, How to enter data science, https://mattsosna.com/DS-transition-1

Data science job for local search, business and ratings site



We Are Looking For

- 3+ years of experience as a data scientist or MS/PhD and 2+ years of industry experience in a quantitative role.
- Fluency with SQL and Python or R for data analysis.
- Solid understanding of statistical inference, experimental design and analysis.
- Enthusiasm for clean code and sharing reproducible results.
- Communication skills to work with partners on engineering, product and business teams.
- An eye for great data visualization with Matplotlib, Plotly, ggplot, or Tableau.
- If you don't have 2+ years of industry experience in a quantitative role, please take a look at our College Data Scientist roles instead!

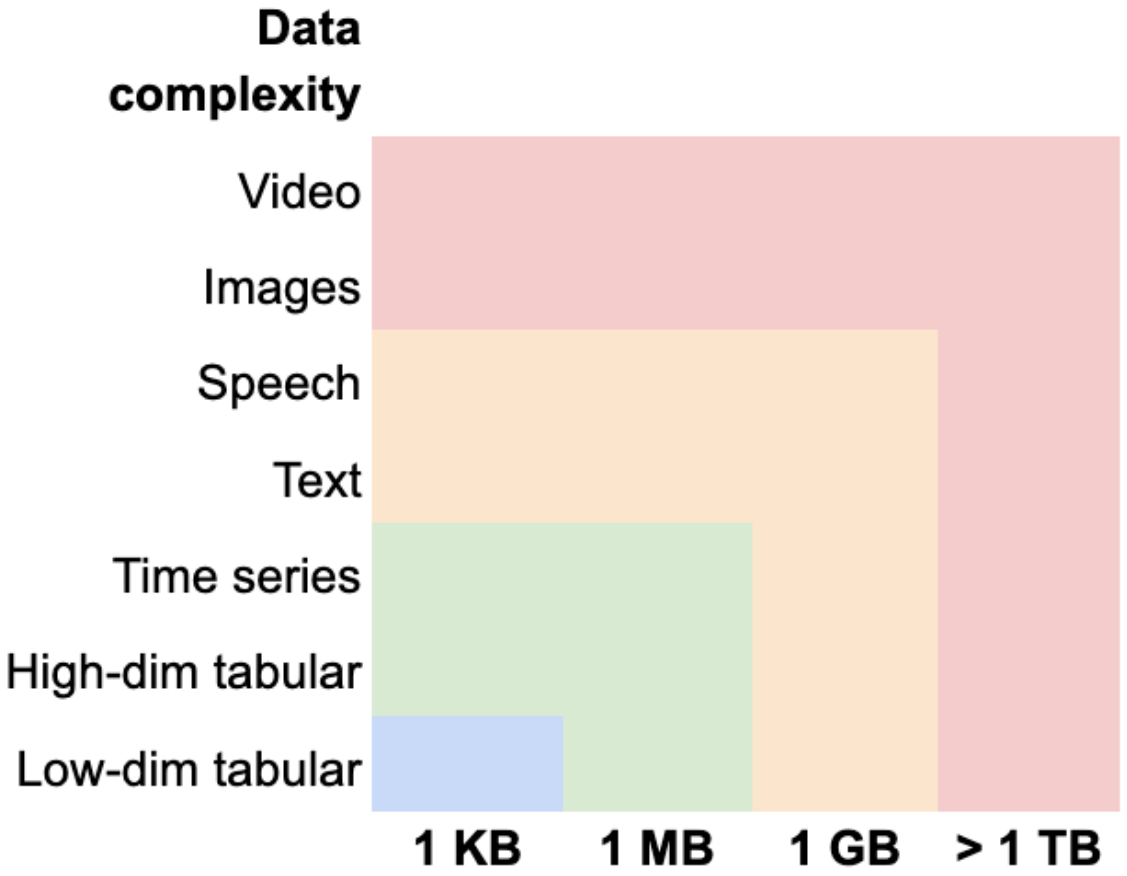
What You Will Do

- Define key metrics to track Yelp's performance and inform product decisions.
- Assess and frame questions from partners into actionable deliverables.
- Design, execute, and analyze complex experiments impacting millions of users.
- Devise and evaluate models for diverse business needs, such as identifying growth opportunities, personalizing user experience, and matching consumers to businesses.
- Own analyses start-to-finish and communicate key insights to stakeholders.
- Share your technical skills to develop and maintain high-quality, reusable analysis tools.

Matt Sosna, How to enter data science, https://mattsosna.com/DS-transition-1

How many cores? I.e. what scale of data science

"If there's no need for something complex, you'll likely find yourself forgoing the deep learning packages and instead cranking out linear and logistic regressions. (Which if done correctly, are incredibly valuable and shouldn't be frowned upon. Answering the question correctly is better than using a cool technique just because it's cool.)"



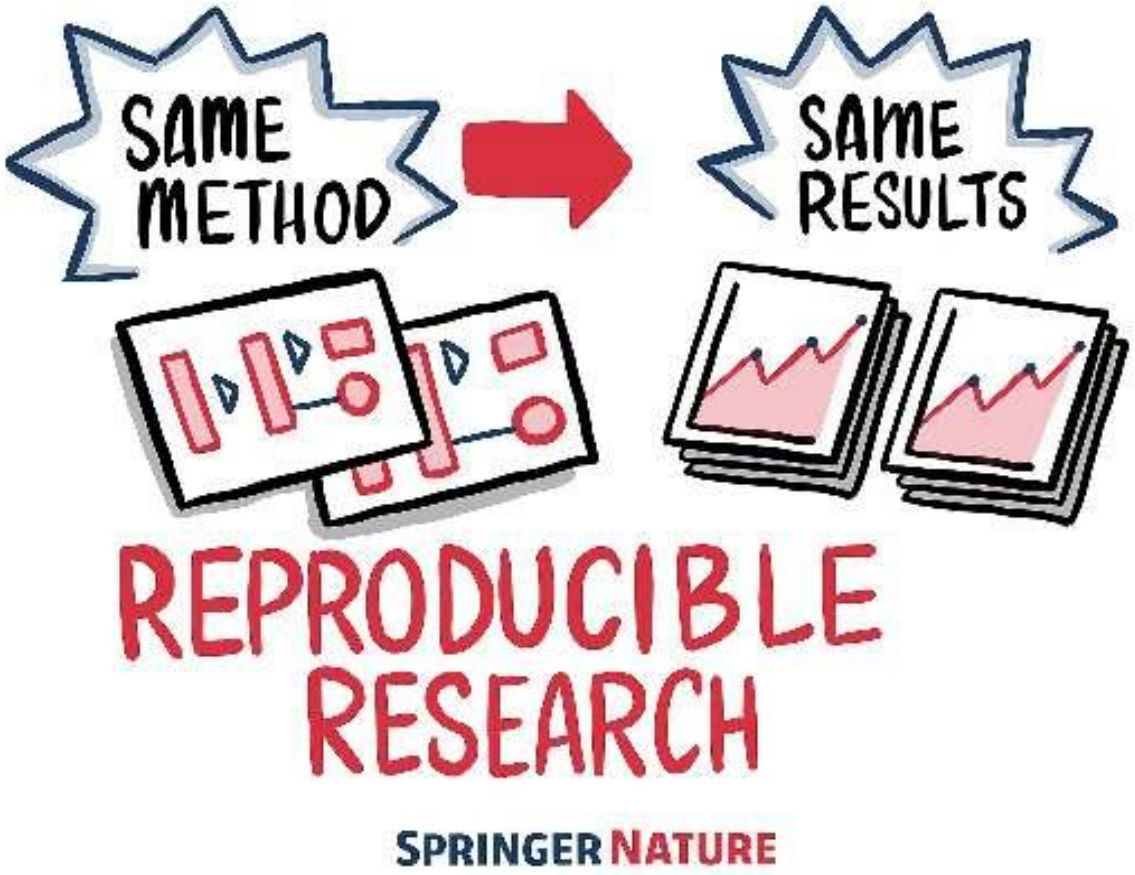
Matt Sosna, How to enter data science, https://mattsosna.com/DS-transition-1

Why is software engineering im portant for data science?

https://forms.office.com/e/NbdUEqmx9M

Reproducibility

Reproducible research in science



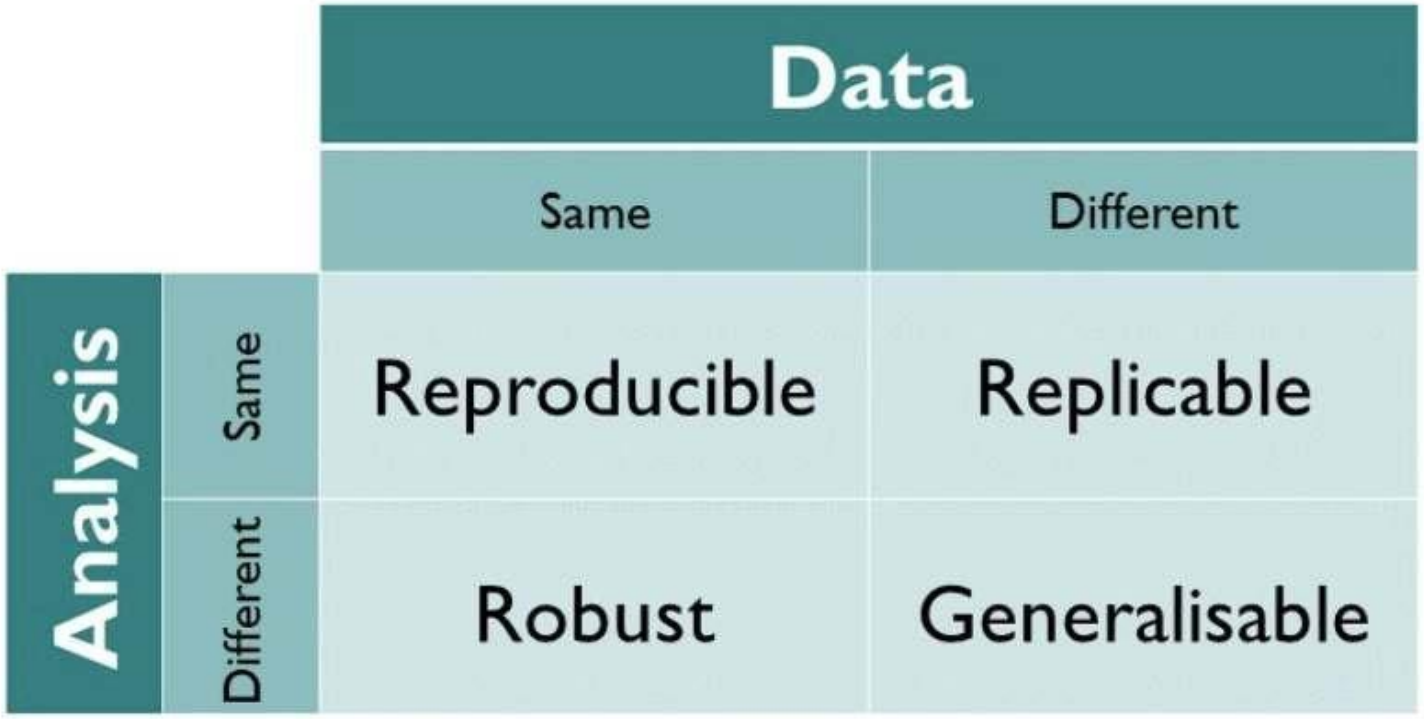
For others and the "future you"

Credit: Ludic Group LLP from Kirstie Whitaker's presentation at,
Scientific Data 2017. CC-BY 4.0. DOI: 10.6084/m9.figshare.5577340.v1.

What is reproducibility in data science?

- I give you the data and code to re-run my analysis, and you can reproduce my results
- I can re-run my analysis in 3 months' time
  - You from 3 months ago doesn't answer email!

Beyond reproducibility



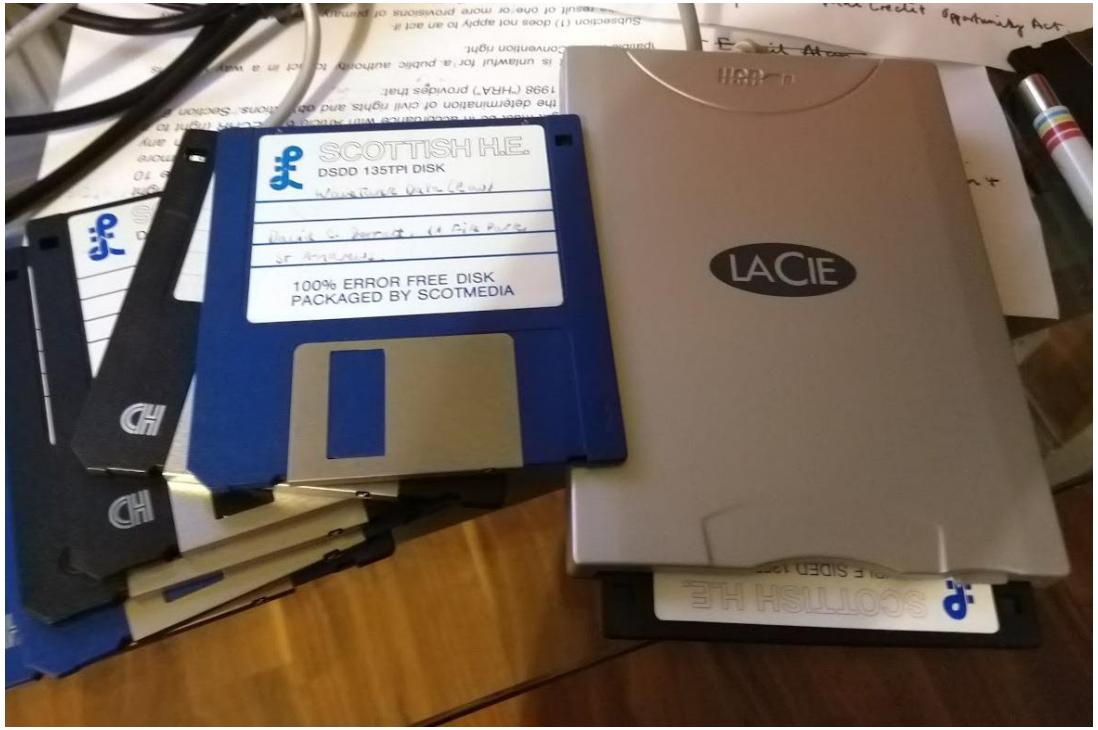
- Ideally our results generalise
- Reproducibility is necessary...
- … but not sufficient
  - we can be reproducing a nonsensical analysis!

The Turing Way Community. (2021, November 10). The Turing Way: A handbook for reproducible, ethical and collaborative research. Zenodo. http://doi.org/10.5281/zenodo.3233853

Barriers and facilitators of reproducibility

- Data
  - Lost or stored
  - Changes
- Code
  - Not robust to changes in data
  - Buggy
  - Library changes
  - Notebooks
- Documentation
- Any others?

Lost or unshared data



Some responses when asking for data in scientific papers

- 2004 (data collected in 1990s)
  - "I got these data 10 years ago, two operating systems ago, when mass storage was based on magneto-optic disks, that I do not use any more.
- 2021 (data collected in 2005)
  - . . . I think that I do have the data somewhere but I may not have access to it at the moment. . . .
- 2021 (data collected in 1995)
  - . . . Sadly the data (if it still exists) is on a dusty floppy disc somewhere – so I wouldn't be able to find or access

Avoiding losing data – sharing via repositories

Scottish Government
Riaghaltas na h-Alba
gov.scot

data.gov.scot

A new service in development

Find information published by Scottish Government on:

- data and evidence on COVID-19
- Scotland's gender equality index
- wealth in Scotland

EBRAINS

TOOLS AND RESOURCES

Data and Knowledge

Online solutions to facilitate sharing of and access to research data, computational models and software

Edinburgh DataShare

INFORMATION SERVICES

Edinburgh DataShare

What is Edinburgh DataShare?

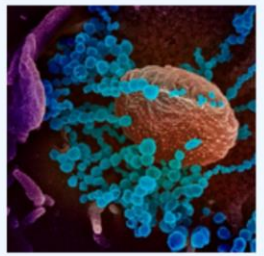
Edinburgh DataShare is a digital repository of research data produced at the University of Edinburgh, hosted by Information Services. Edinburgh University researchers who have produced research data associated with an existing or forthcoming publication, or which has potential use for other researchers, are invited to upload their dataset for sharing and safekeeping. A persistent identifier and suggested citation will be provided.

zenodo

Search

Upload Communities

Featured communities



Chicago COVID-19 Response

This repository community collects research outputs and i
efforts in Chicago. Users are encouraged to upload their re
discovery of information. Although Open Access articles ar

Curated by: saragon

Changing data

- Suppose you're analysing a dataset that's being updated constantly
- You edit the code during the week.
- The result you get on Friday with the live dataset is different from the result you got on Monday
  - Is the change because of the code or the data?

Code – robustness to changing data

Data format of marks downloaded from Gradescope in 2022:

```
Action,   Mark, SID, CWID
To Grade, 72,   124, 1
…
```

Data format in 2023

```
Mark, SID, CWID
72,   124, 1
…
```

2022 analysis of data

```
df = pd.read_csv('marks.csv')
df = df.iloc[:,1:3]
print(df['Mark']).mean()
```

What happens if we run this code on the 2023 data?

Buggy code

- Usual software engineering practices
  - Clean code
  - Code review
  - Testing
- Specific to data science:
  - Testing against simplified data
  - Visualisation
  - Checking results make sense
- Sharing code and data
  - Both revealing and intimidating!

Credit: Ludic Group LLP from Kirstie Whitaker's presentation at, Scientific Data 2017. CC-BY 4.0. DOI: 10.6084/m9.figshare.5577340.v1.

Library changes

Original

```
import seaborn as sns
…
sns.barplot(data=df,
        x='Food', y='Area',
        palette=colors)
```

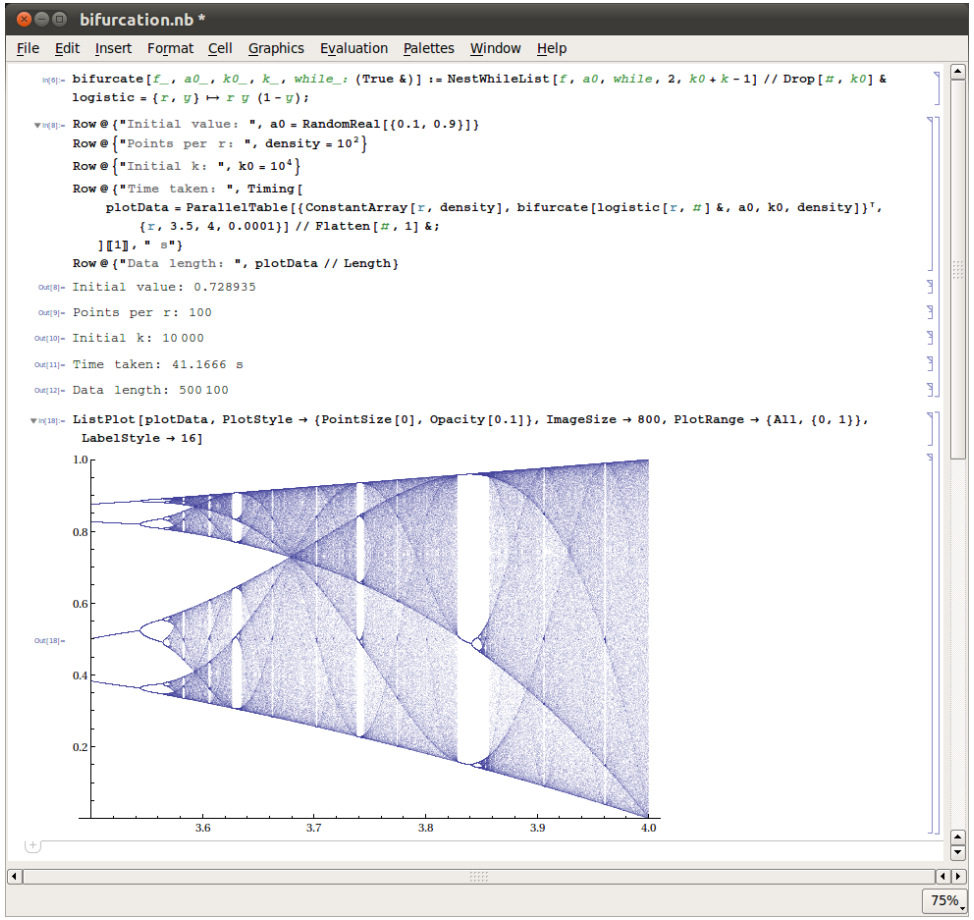
12 months later

```
import seaborn as sns
…
sns.barplot(data=df,
        x='Food', y='Area',
        palette=colors)
```

MatplotlibDeprecationWarning: Unable to determine Axes...

Notebooks – 1988…

Science Museum, Creative Commons A Wikipdedia ttribution-NonCommercial-ShareAlike 4.0 Licence



Wikipedia

…2015

jupyter

The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

A problem with notebooks

```
[13] a = 5
[14] b = a + 6
[15] print(b)
11
```

```
[17] a = 10
[14] b = a + 6
[15] print(b)
11
```

- Make sure notebooks are run from start to end for reproducibility!!
- For large notebooks with functions or large code blocks:
  - Pull out functions into Python modules and import from Notebook

Data management and code management

Data versus code

- Version control for code (e.g. using Github) is a good thing but does it work for data?
- VC systems deal well with “small” text files – kilobytes rather than megabytes, and definitely not gigabytes (Wilson et al., 2017)
- Thus they are good for code but not so good for data, especially large datasets
- Thus we need different solutions for storing and keeping track of changes to data and code

Data recommendations (1)

- Save the raw data
  - Don’t overwrite with a better, cleaned-up version
  - Protect, e.g. with file permissions
- If downloading from a database, record details used to obtain data (exact query, date of retrieval, version of DB on that date)
- Back-up the data

Data recommendations (2)

- Save and share a clean version of the data
  - Cleaned
  - Open data format, e.g. CSV, JSON, YAML, XML
  - Meaningful variable names and file names
  - Metadata about meaning of columns (if not clear from original dataset)
- Make sure you the cleaned dataset is “Tidy” (see lectures on data)
  - One observation per row
  - One variable per column
  - Unit not stored with numbers – store in metadata or separate column
  - Ideally unique ID for each observation
- If generating data, share in a repository

Code recommendations

- Use version control – git is currently dominant
- Documentation
- Specifying versions – this can be done using Conda environments and .req files

jupyter=1.0.0

matplotlib=2.2.3

numpy=1.15.0

pandas=0.23.4

scikit-learn=0.19.1

scipy=1.1.0

seaborn=0.9.0

python-graphviz=0.8.4

Jupyter notebooks and version control

- Problem: Jupyter notebooks are written in JSON, so the diffs stored in version control systems are not very intelligible
  - R markdown doesn't suffer from this problem
- Workaround: Use tooling for diffing & merging Jupyter notebooks,
- e.g. Git integrations in VSCode or nbdime
- Clear output manually before committing

```
$ diff a.ipynb b.ipynb
76,77d75
<     "plt.rc('axes', grid=False)\n",
<     "plt.rc('axes', facecolor='white')\n"
90c88
<       "image/png": "iVBORw0KGgoAAAANSUhEUg
AAAAAlwSFlz\nAAAWJQAAFiUBSVIk8AAAIABJREFUeJ
lY\n1waDyDZg8MX+zMU2F4Mx1x8PwWAwxmBjg4yNi2B
N+9z/o9zzynprvq1D6nqqtqr1prbRNFEQghhBBCCCGE
BCiPdQ5CKEEEIIIYQQQggh3kORixBCCCGEEIIIYR4D0
EEEIIIYQQQggh3kORixBCCCGEEIIIYR4D0UuRixBCC
Qggh3kORixBCCGEEIIIYR4D0UuRixBCCGEEIIIYR4D
ELChCIXIYQQQirDGPOmKaFj3BhzkMNx/H/G\nmG3GmP
BSJe+DCDMjAH7L4TjeAmA+gLc5\nHEMRGNcDqJi3AVg
DjobzZwBuBvBxR/dP\nsvERADcC+LTrgRBCCCFEHx
T7IRRdFf\nRlH0K1EUXe96LIQQQgjRB0UuRixBCCGEEI
```

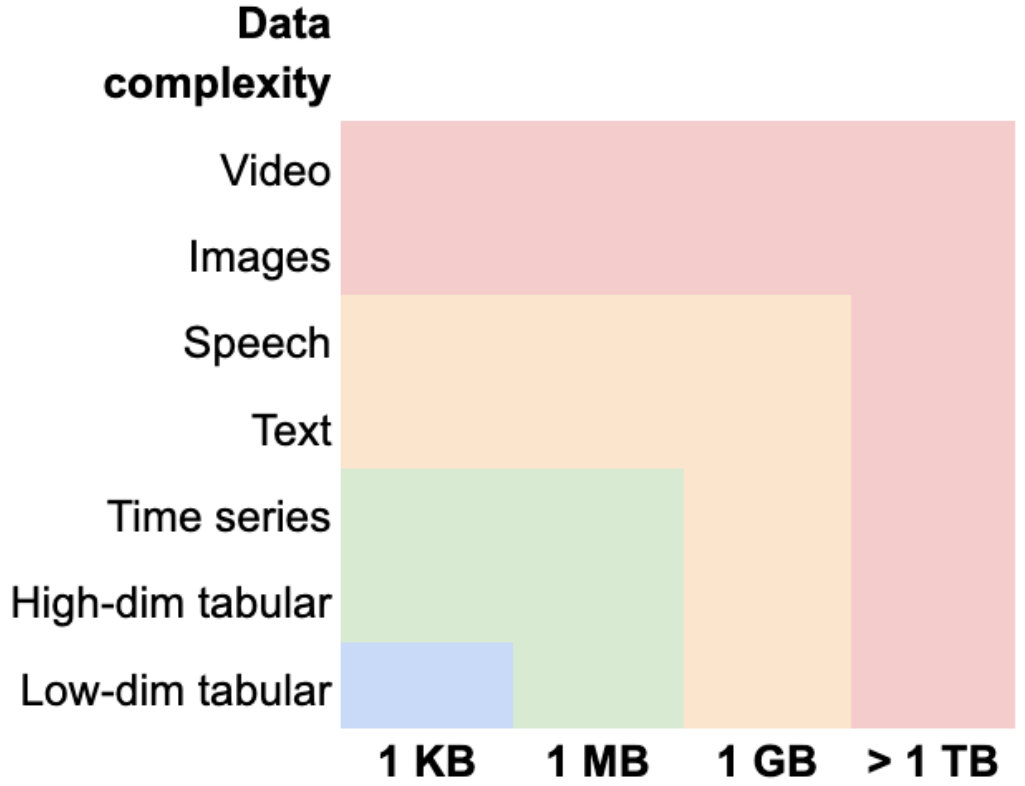
Recommendation – *Significance* magazine

- Published by the Royal Statistical Society, the American Statistical Association and the Statistical Society of Austrailia
- Wide range of articles about data science and stats at an accessible level
- https://significancemagazine.com/issues/

Efficiency and scaling

https://pola.rs/

- Pandas is great for small-ish tabular data...
- … what happens if the data grows?
  1. Use a faster python library https://pola.rs/
  2. Store data in a relational DB and retrieve and process using SQL



Retrieving and filtering data

Python

```
df = pd.read_csv('student_animals')
students = df.loc[df.animal=='Walrus',
      ['name', 'animal']]
students = students.rename(columns=
  {'animal': 'favourite_animal',
   'name': 'student_name'},)
```

SQL

```
SELECT name AS student_name,
      animal AS favorite_animal
FROM students
WHERE animal = 'Walrus';
```

Example from
https://mattsosna.com/DS-transition-4/

- Run SQL query from python and then process in python
- Advantage: if 2 million students but only 300 have a Walrus as their favourite animal, we reduce the load time hugely

Joins (AKA merges in Pandas)

Python/pandas merge

```
# Query 1
pd.merge(users, sql_pros,
        how='inner',
        on='id')


# Query 2
pd.merge(users, sql_pros,
        how='outer'
        on='id')
```

SQL join

```
 -- Query 1: don't drop any rows in 'users'
   SELECT *
     FROM users AS u
LEFT JOIN sql_pros AS sp
    USING (id);

-- Query 2: don't drop any rows in either table
   SELECT *
     FROM users AS u
FULL JOIN transactions AS t
       ON u.id = t.user_id;
```

Example from https://mattsosna.com/DS-transition-4/

Groupby/Aggregation

Pandas

```
df.groupby('name',
        As_index=False)\
   .mean()['name', 'score']
```

SQL

```
SELECT s.name,
        AVG(s.score) AS avg_score
    FROM students AS s;
```

Summary

- What is a "data scientist" and what scale of data science?
- Why is software engineering needed for data science?
- Reproducibility
- Data management and code management
- Efficiency and scaling
- **Don't lose sight of what data you have and what analysis you're doing**
  - **you could be doing the wrong thing very quickly**