Starting out in Data Science

Tim Vivian-Griffiths

Outline of this presentation

- My experiences of starting out in the field
- What I'd like to have known at the start
- **Disclaimer:**
  - Many points here are my own *opinions*
  - Positive opinions will be specific
  - Negative opinions will be descriptive
  - Feedback and questions welcome
- **Main aims:**
  - Tips on how to use data science methods in a current role or research
  - How to start out in a new data science career

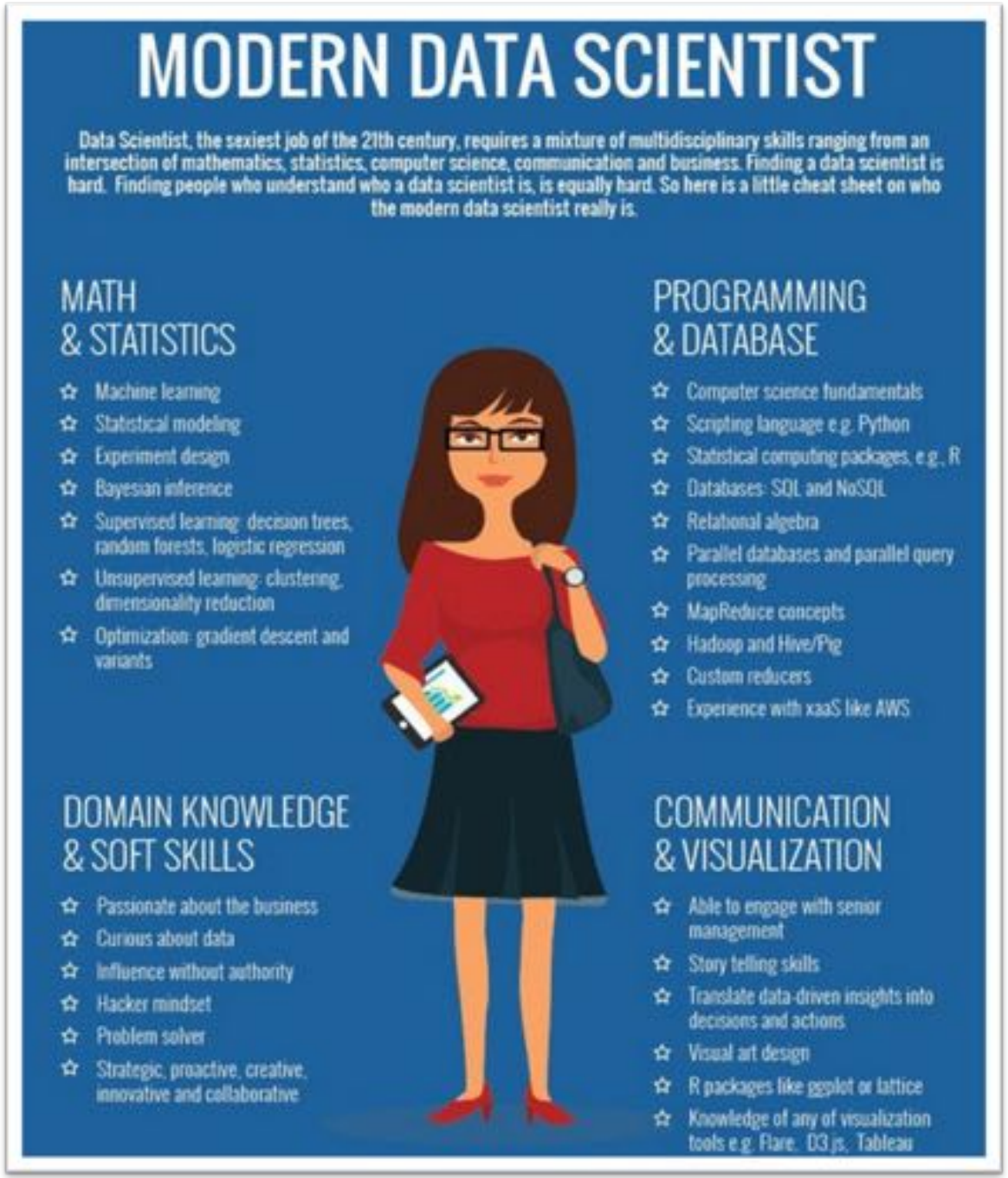
About myself: Data Scientist at

- Previously: Bartending and Waitering
- Re-educate at 29
  - Psychology Diploma at Cardiff University
- **No experience of coding before 30**
  - Online tutorials
  - MRes. Birmingham University
  - Open University Maths
  - Ph.D Cardiff University

What is a Data Scientist?

- Use as a possible checklist
- **Please** take with a pinch of salt!
- This is a DS **TEAM**
- No one can do **ALL** of this

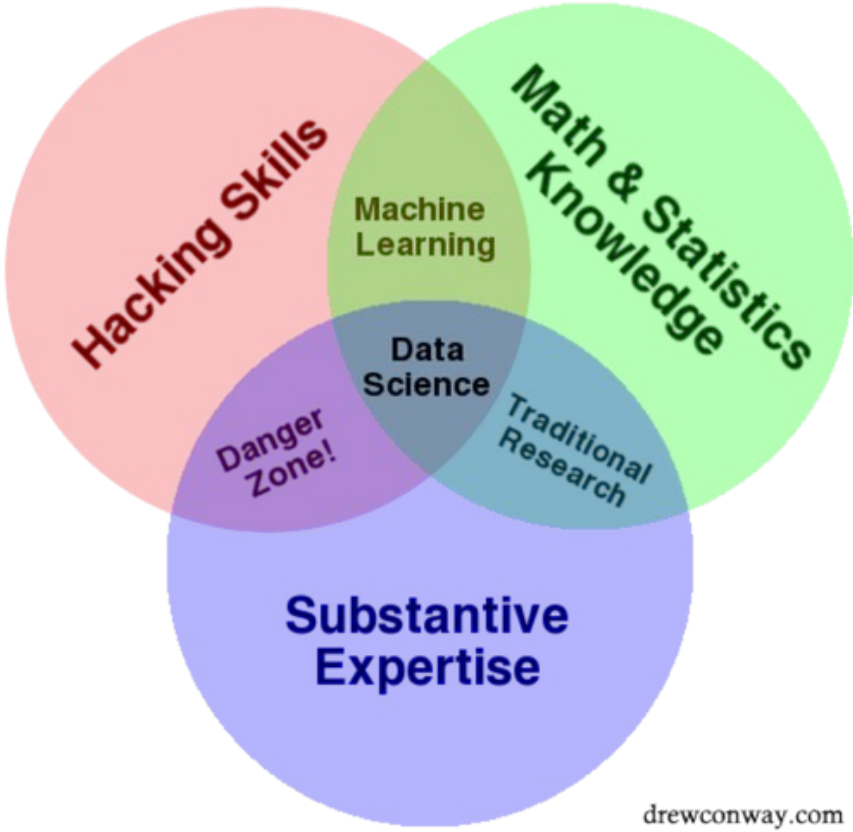
- Everyone can use a spell checker



Sexiest job??

Disciplines of Data Science

1. Communication and Soft-Skills:
   - Can be developed and improved at any stage
2. Mathematics:
   - Statistics
   - Calculus
   - Linear Algebra
3. Coding:
   - How to customize your work
   - Use tools developed by others
   - You don’t need to be a software engineer



Communication and Soft-Skills

- “If you can't **explain** it **simply**, you don't understand it well enough”
  - Albert Einstein
- Can be improved at any time – practice when you can

http://dilbert.com/strip/2018-04-03

What you can do right now

- Explain your work/research
  - Colleagues
  - Friends/family
- Explore Visualisations
- Work on presentations
- Apply the “**So what?**” principle
  - This is what people will remember
- Read blogs/listen to podcasts

Avoid unforced errors:

Mathematics

The Open University

- Very thorough material
- Distance learning with regular meetups
- Costs money
- Longer term commitment
- Can use older tech / 80s-tastic videos
- Not all topics relevant to my work

KHAN ACADEMY

- From basics to advanced
  - *"Kindergarten to calculus"*
- From mission statement:
  - Practice Exercises
  - Instructional Videos
  - Personalized Learning Dashboard

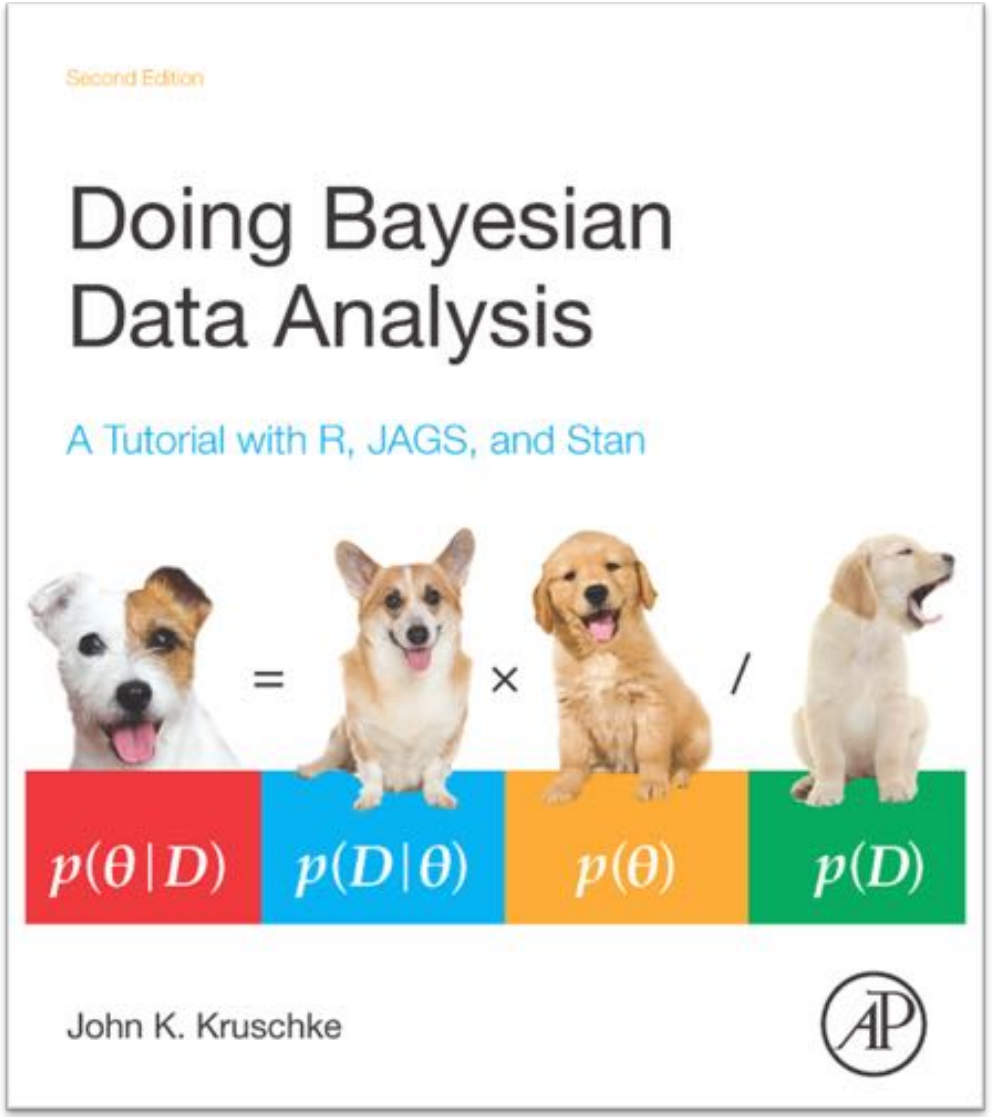
Note on Bayesian Statistics:

- Probability
- Linear Algebra/Matrix operations
- Calculus

$$p(D) = \int \mathrm{d}\theta\, p(D|\theta)\, p(\theta)$$
$$\approx \frac{1}{N} \sum_{\theta_i \sim p(\theta)}^{N} p(D|\theta_i)$$

PyMC3

https://github.com/pymc-devs/pymc3



Coding

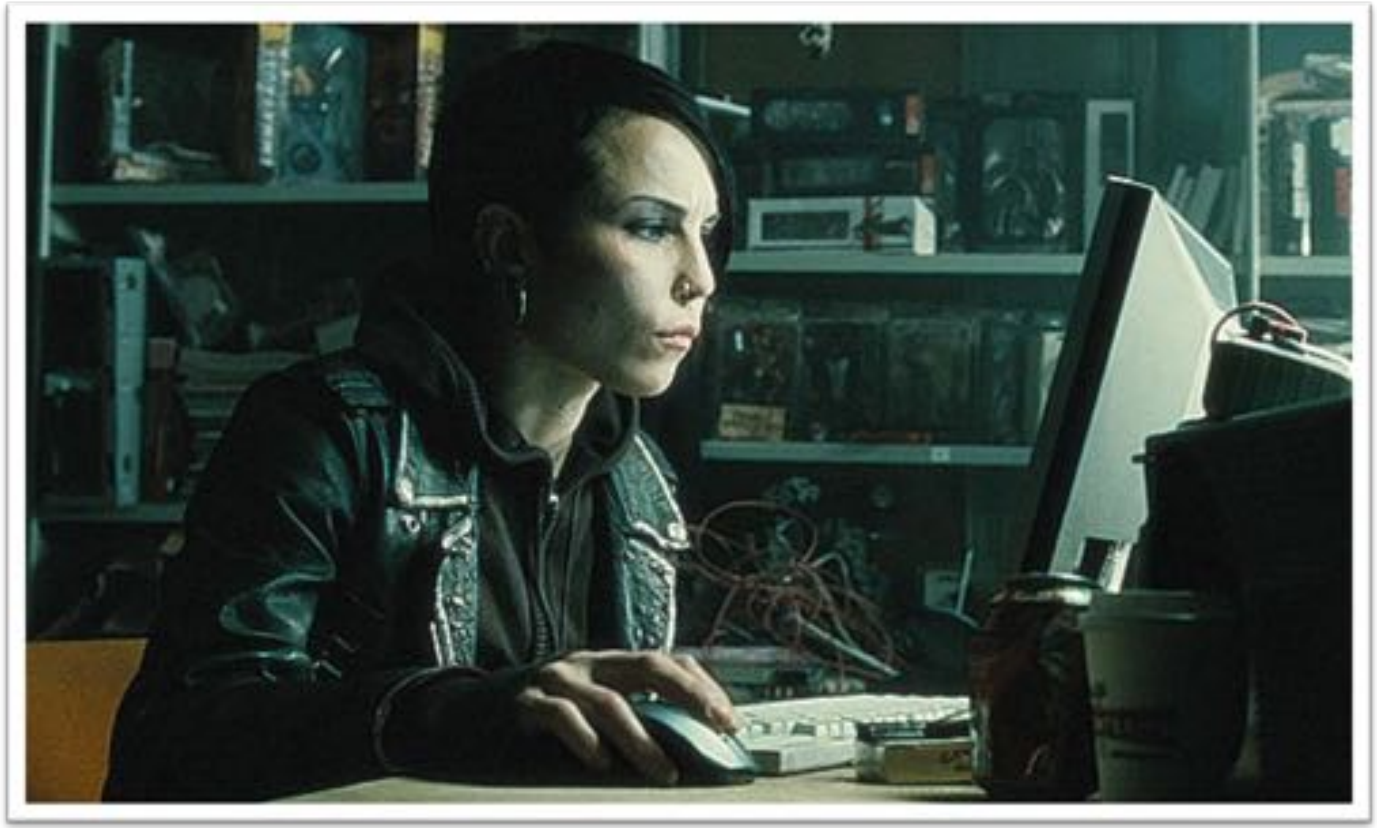
Benefits of Coding

- **Incredibly flexible:**
  - You are are less constrained by what others *think* you need to do
- **Very adaptable:**
  - Easy to repeat established processes
    - Create modules
    - Create packages
  - Test your code:
    - Pytest
    - Hypothesis
    - testthat
- **Deploy models**
  - API creation
  - Flask
    - https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-i-hello-world

What is coding?

@racoconn - Twitter

Hack Right NOW!



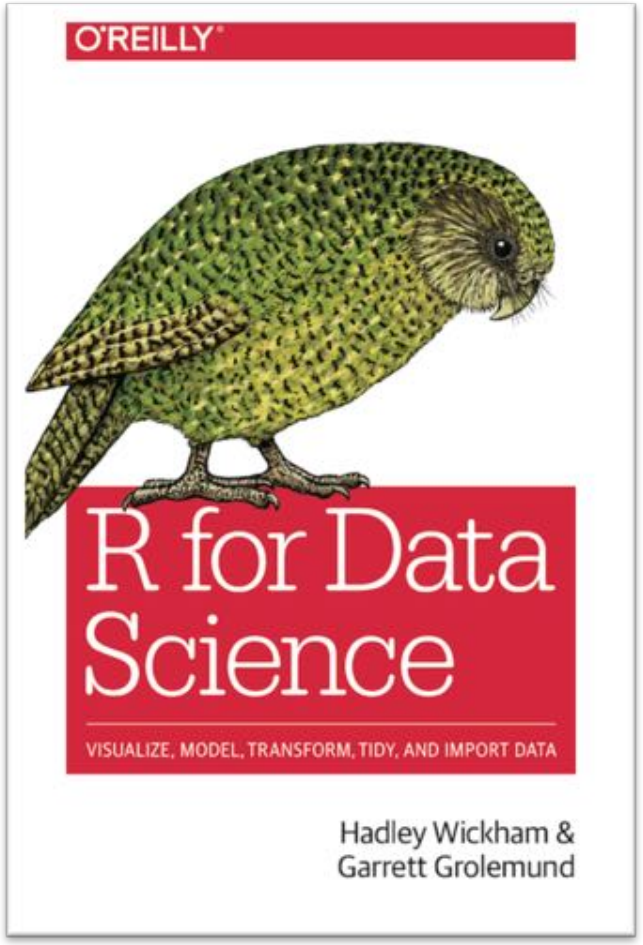
https://hackertyper.net/

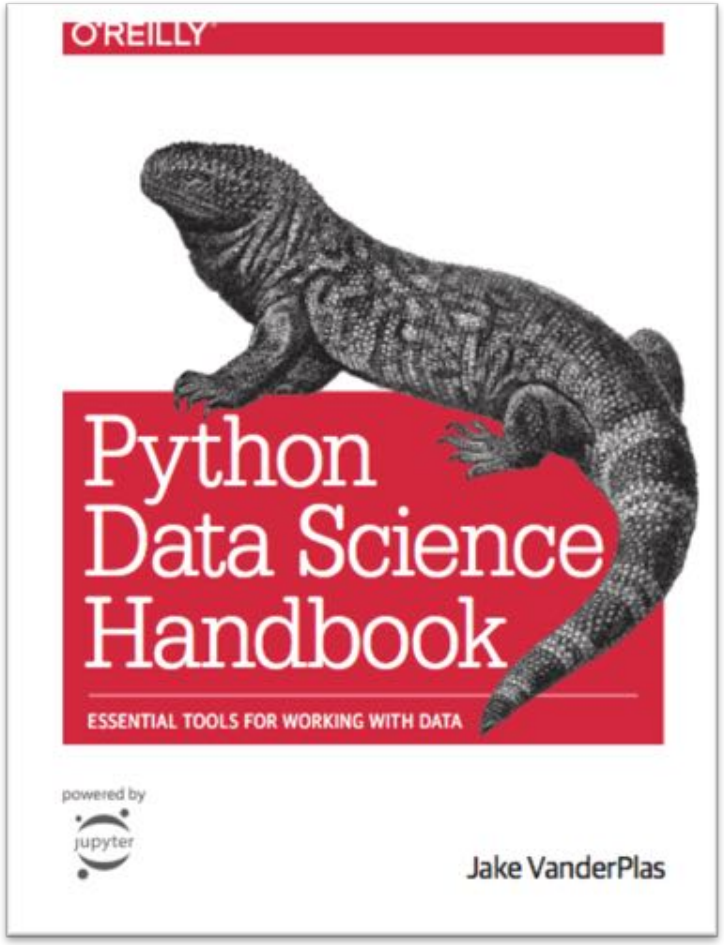
Where to start?

- Video tutorials / browser based exercises
- **No setup required**
- Subscription based:
  - $29 month: Monthly plan
  - $25 month: Annual plan
- Variety of topics for Python and R

- Intro to both languages
- Data preprocessing/cleaning
- Machine Learning
- Data Visualisation

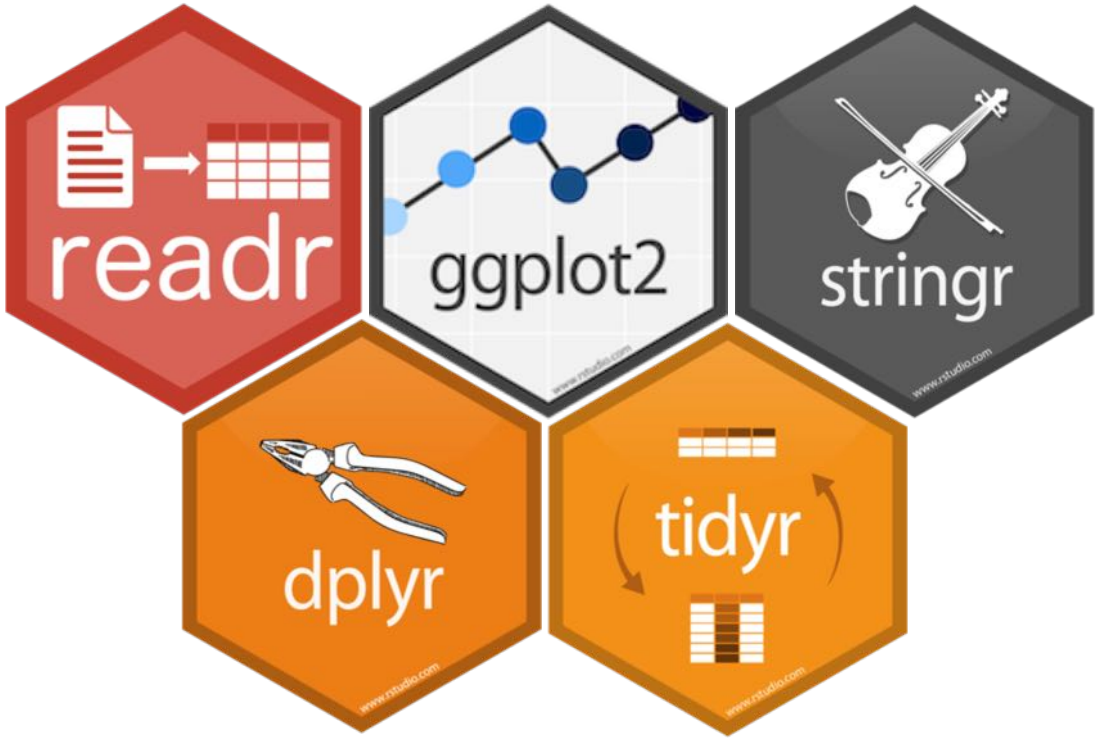
Moving on:


O'REILLY
R for Data Science
VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA
Hadley Wickham &
Garrett Grolemund

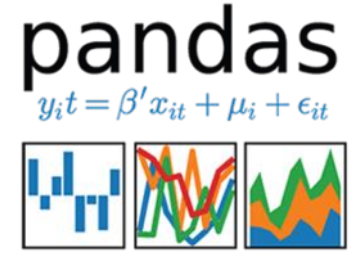

O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by
jupyter
Jake VanderPlas

Language specific advice:

- **R:** base and **tidyverse**



- **Python:** Scientific Stack



- **Matplotlib**:
  - https://realpython.com/python-matplotlib-guide/
- **Seaborn:**
  - https://seaborn.pydata.org/

Text Editors:

Version Control

- Excellent software to track development of code
- Allows checkpoints, reference tags and *rollback* ability
- Create *branches* to isolate and trial new developments/features
- Vital for teamwork

Learning resources

DataCamp

Udemy – Jason Taylor

- Git Complete:
  - The definitive, step-by-step guide to Git

A note on Big Data

- Hadoop and Spark
  - Both of these can be **hard**
- Is your data really big data?
  - 1000s?
  - 100s of 1000s?
  - Millions?
  - Billions?
- Tools are available:
  - Python and R
  - You don’t have to be an engineer
  - You don’t **have** to learn Scala
    - But it’s great for functional programming
      - Alvin Alexander

- First learn:
  - Pandas
  - SQL

Online courses/MOOCs

- What to watch out for:
  - Reviews
    - Check for any pushiness from creators
  - How often is it updated?
    - Check for forums
  - Cost:
    - With subscriptions – try monthly first
  - **Too much too soon**
    - Especially in introduction courses
    - Reliance on boilerplate notebooks
    - “Get code on GitHub”

Common Platforms:

- Wide variety of DS courses
- Subscription service
- Free to audit
  - Limited access
- Popular courses:
  - Applied data science with Python
  - Data Science, Johns Hopkins (R)
  - Advanced Machine Learning
  - Functional programming with Scala
  - Deep Learning
- Videos and exercises

- Selection of free and paid courses
- **Nanodegree**
  - Data Analyst
  - Machine Learning
  - Artificial Intelligence
  - Natural Language Processing
  - Self-driving car engineer
  - Deep Learning
- Can cost more
  - Has good reviews
- Videos and exercises

Common Platforms:

- Pros:
  - Can be great intro courses
  - Good for beginners
  - Cheap (always buy on sale!!!)
- Cons:
  - Almost all video based
  - Don’t expect to **passively learn**
- Check:
  - Reviews
  - Date of last update

- Top down approach
- Gets great feedback
- Easy to set up GPU environment

Paperspace

Very new addition:

- Platform for data science competitions
- Just started online training
  - https://www.kaggle.com/learn/overview
- I've just seen this!
- You don't **HAVE** to do MOOCs

Cloud Computing

A CLOUD GURU

- Subscription model
- But – you can start on Udemy
  - Remember the sales!

Deep Learning

- Cutting edge
  - Your libraries will change over time
- Don’t focus on TensorFlow in the beginning – my opinion
  - Keras is incredibly accessible – “Deep Learning for humans”
  - PyTorch is gaining in popularity – fast.ai

“Uncool” stuff (but really the most important)

SQL – Get your data

- “Bread and butter” of Data Science
- You don’t need to be a DBA
- Kaggle – that’s the picture!

Data Preparation and Cleaning

- Feature engineering for machine learning
  - Udemy – 4.6 out of 5 stars
  - Soledad Galli – Gave a PyData London talk
- “You’re not a data scientist until you can do this”
  - Paraphrasing of a review

Some final points:

- Don’t let anyone put you off
- You can learn at your own pace
- Don’t be afraid to ask for help
  - But don’t tolerate bullying
- If I can do this, you can too

You don’t need to be a Data Scientist

- You don’t have to do the “sexiest job” of the 21th Century

You don’t need to be a Data Scientist

- Biology
- Chemistry
- Physics
- Genetics
- Zoology
- Psychology
- Medicine
- Engineering
- Mathematics
- Marketing
- Sales
- Town Planning
- Logistics
- Transportation
- Finance
- Aviation
- Project management
- Recruitment

Just use data – there’ll be plenty!

Thanks for listening

- Any questions?