



DATA SCIENCE

SYD DAT 8

Week 1 Lesson 2- Basics of Data Science & Git
Thursday 25th May 2017

1. What is it that data scientists do? Day to Day
2. What do they need to succeed
3. Data science project
4. Git. What is Git. 1. version control. 2. collaboration
5. Git. The terminal
6. Git. Their repositories, basic actions
7. Git. Open source, collaboration etc.
8. Python: More manipulation with pandas etc. More reading dat in etc.
Self coding. After having retrieved the content.
9. Discussion
10. Homework

ABOUT ME

Pre-Data Science Life

- Early career was rooted in mechatronic engineering and business.
- Moved into strategic marketing, where my job was to imagine new industrial technology solutions
- Travelled
- “Discovered” machine learning and decided to go all in.



ABOUT ME

New Data Science Life

- Early data jobs contracting in reporting.
- Read about machine learning, networked, did online courses, attended meetups, competed in Kaggle
- Consulted in data science under my own businesses
- Developed a machine learning based customer insight tool
- Joined Servian as a consultant specialising in machine learning. My main clients being Commonwealth Bank, Optus, Westpac and some government organisations.



DATA SCIENCE PART TIME COURSE

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?



Michael E. Driscoll
@medriscoll

 Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu
[@peteskomoroch](#)

RETWEETS FAVORITES
35 26

1:57 PM - 17 Jul 2012

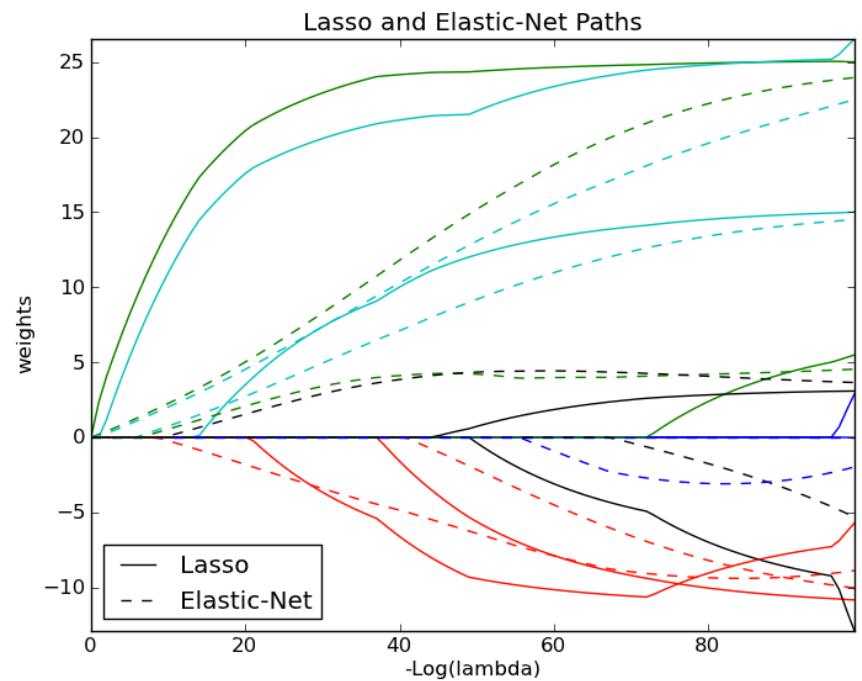
WHAT IS DATA SCIENCE?

“those who use both **data** and
science to create something
new”

September 2011 D.J. Patil in Building Data Science Teams

MODELS AND METHODS FOR DATA

“Model building is **complex** because it requires **combining information** from **exploring** the data and information from sources **external** to the data such as subject matter theory and other sets of data”



COMPUTING WITH DATA

“Data analysis projects today rely on databases, computer and network hardware, and computer and network software. A collection of models and methods for data analysis will be used only if the collection is implemented in a computing environment that makes the models and methods sufficiently efficient to use”



WHY NOW?

“We have massive amounts of data about many aspects of our lives, and, simultaneously, an abundance of inexpensive computing power. Shopping, communicating, reading news, listening to music, searching for information, expressing our opinions—all this is being tracked online”



But our offline lives too:

Datafication - “taking all aspects of life and turning them into data.”

- Mobile sensors
- Social media connections
- Likes of real world objects
- Medical records
- Memes



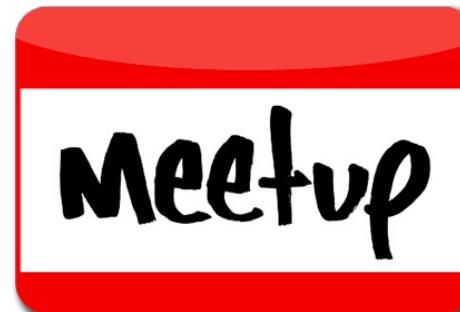
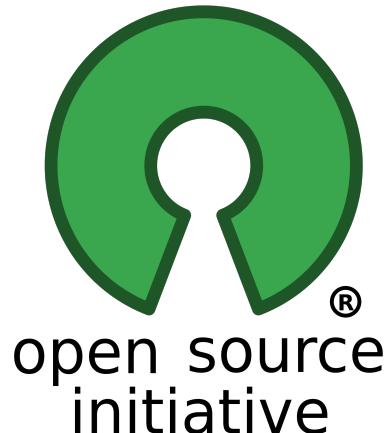
COMMUNICATION

“Data science is the practice of turning tools and raw data into something that non-data scientists might care about.” - Sandy Ryza



OPEN SOURCE & PEDAGOGY

- Github
- Public data sets
- Blogs
- Academia & corporate research



“Education in data science does many things. It trains statisticians. But just as important it trains non-statisticians, conveying how valuable data science is for learning about the world.”

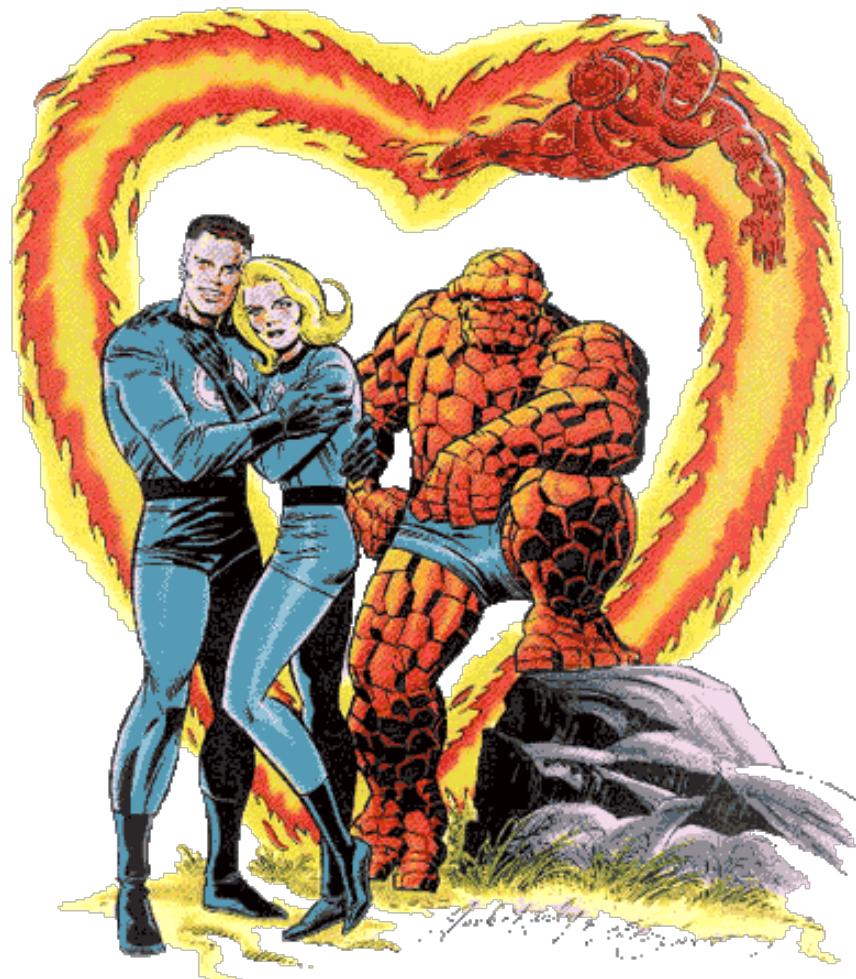
WHAT MAKES A GOOD DATA SCIENCE TEAM

A GOOD MIX

Ideal data science skills can be spread across team.

Some literacy of each skill is necessary.

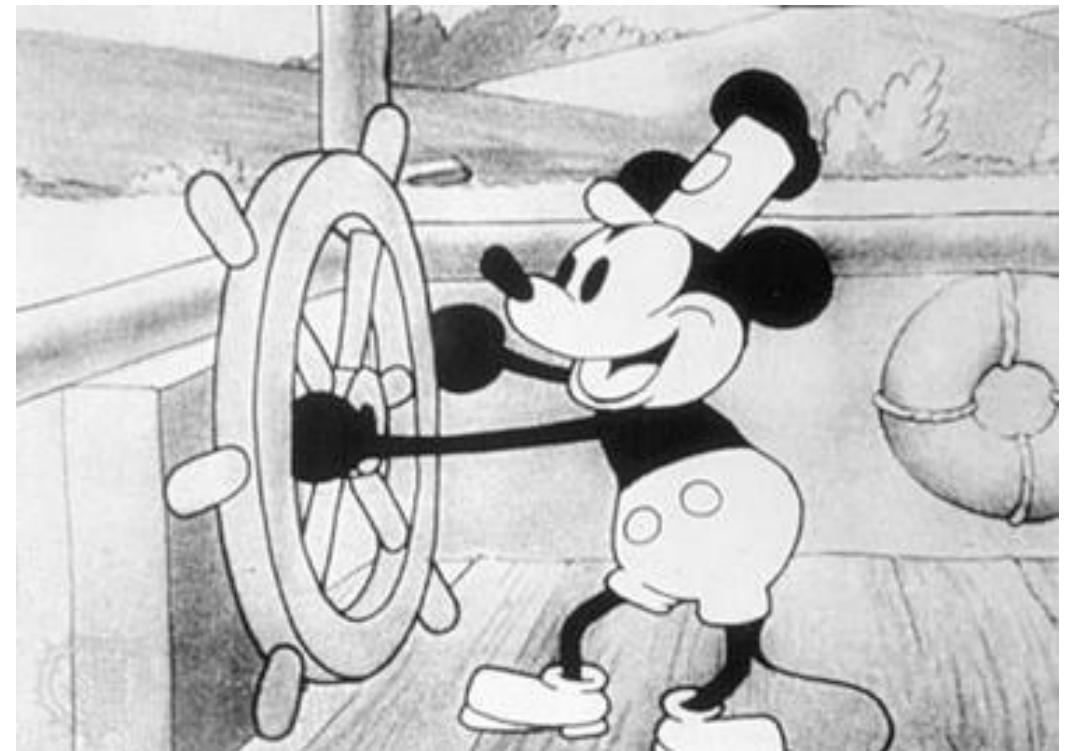
Typically small teams of 2-4



SENIOR SPONSORSHIP

Data science is sales

- Steering committee needs to be well informed and receptive.
- Business value should be articulated repeatedly.



NETWORKING WITHIN THE ORGANISATION

- Domain expertise
- Ideas
- Alignment to business needs
- Finding internal customers for data science



IT CO-OPERATION

Mission of IT departments is
fundamentally different to data science.

control and governance

vs

access, freedom and use

IT CO-OPERATION

Mission of IT departments is fundamentally different to data science.

control and governance

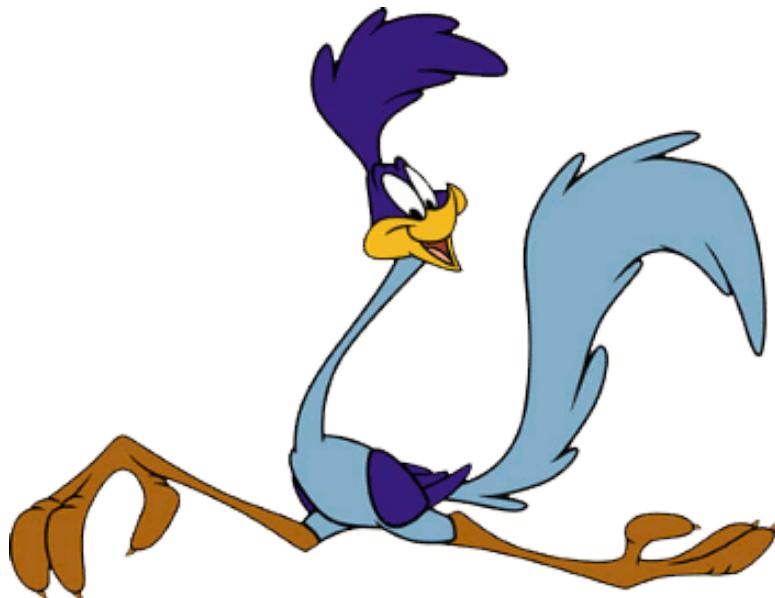
vs

access, freedom and use

A good relationship is essential for timely access to resources.



AGILE

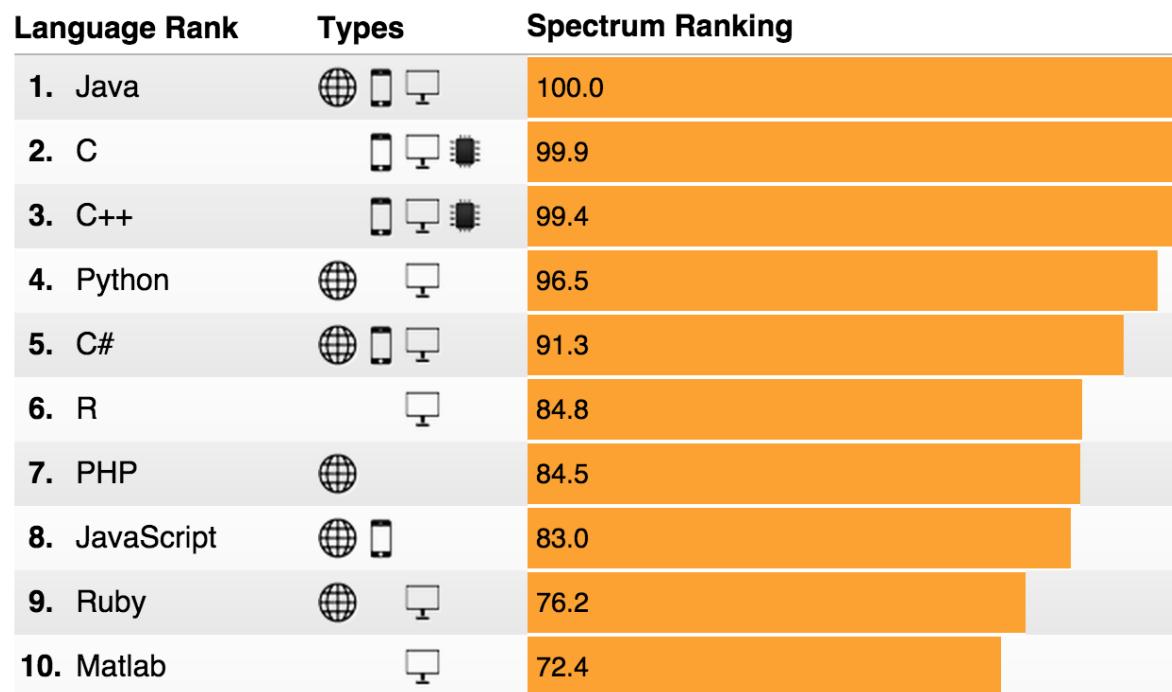


Agile methodologies are well suited to data science.

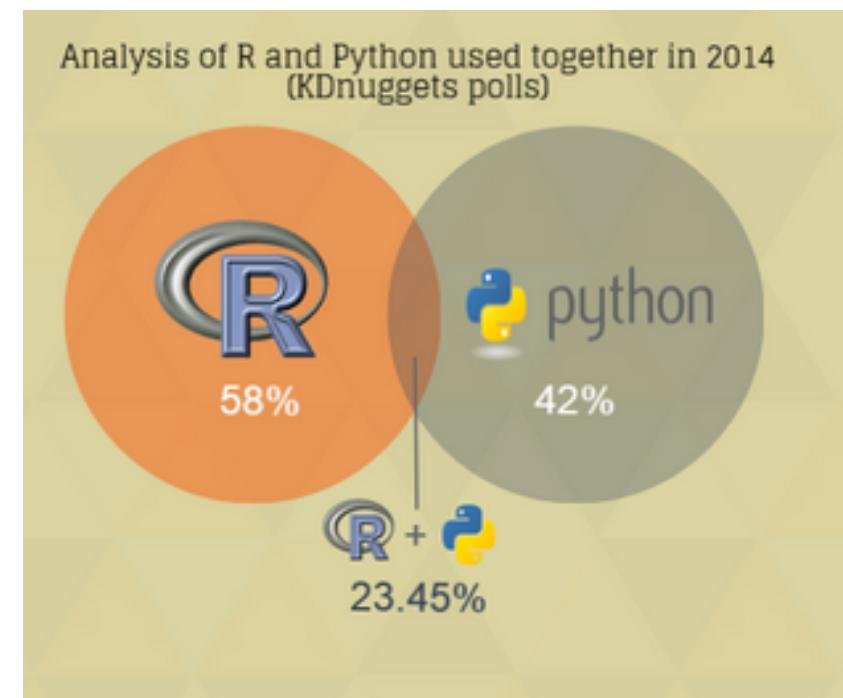
- Incremental development
- Responsive to new information / insights
- Time-boxed goals
- Structure for cooperating with stakeholders

TOOLS OF DATA SCIENCE

PYTHON vs R

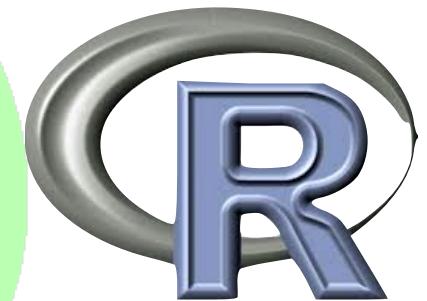
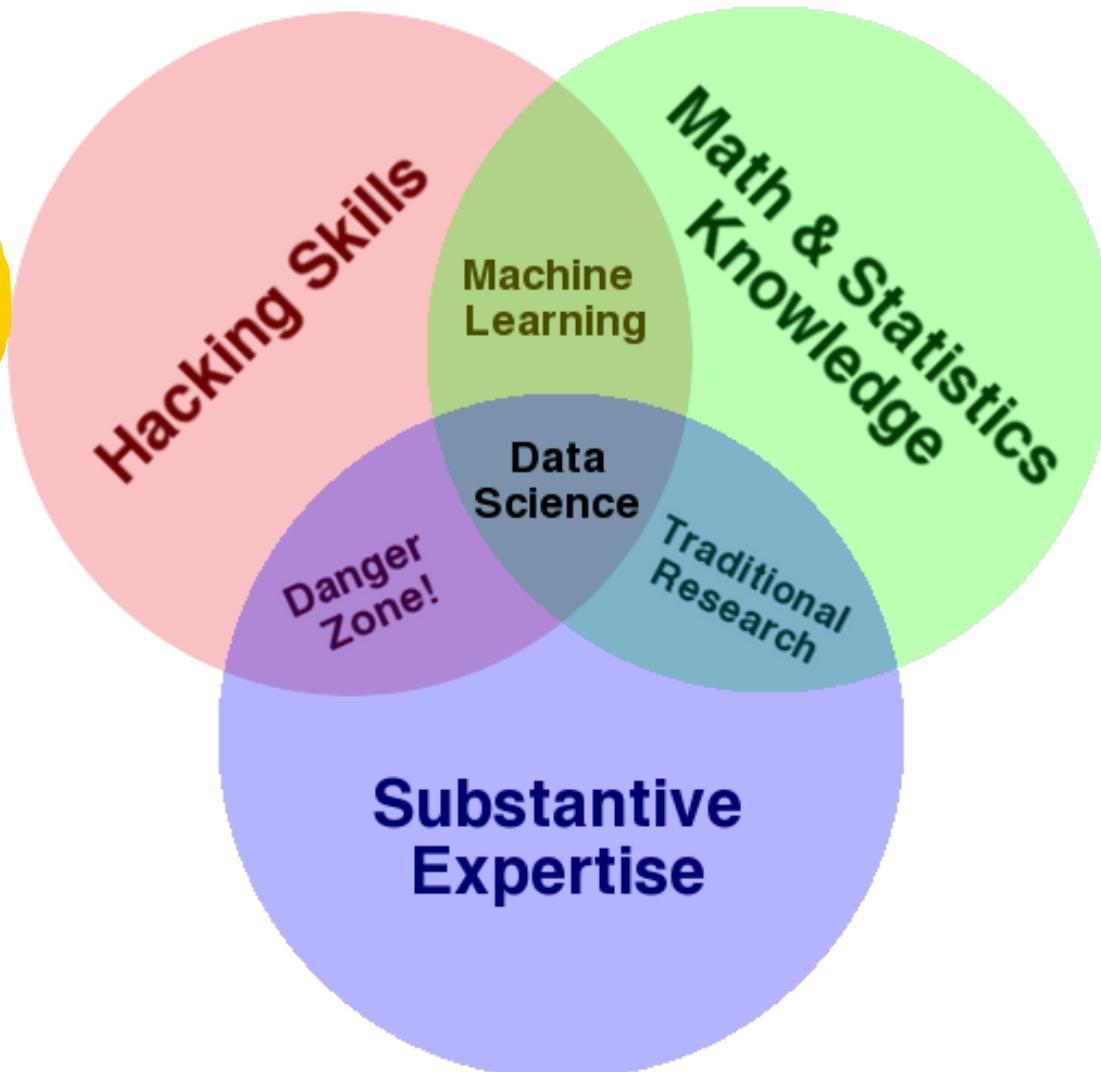
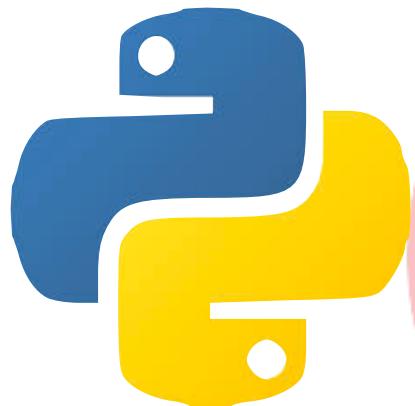


IEEE Spectrum Survey 2015



DataCamp Infographic 2015

PYTHON & R



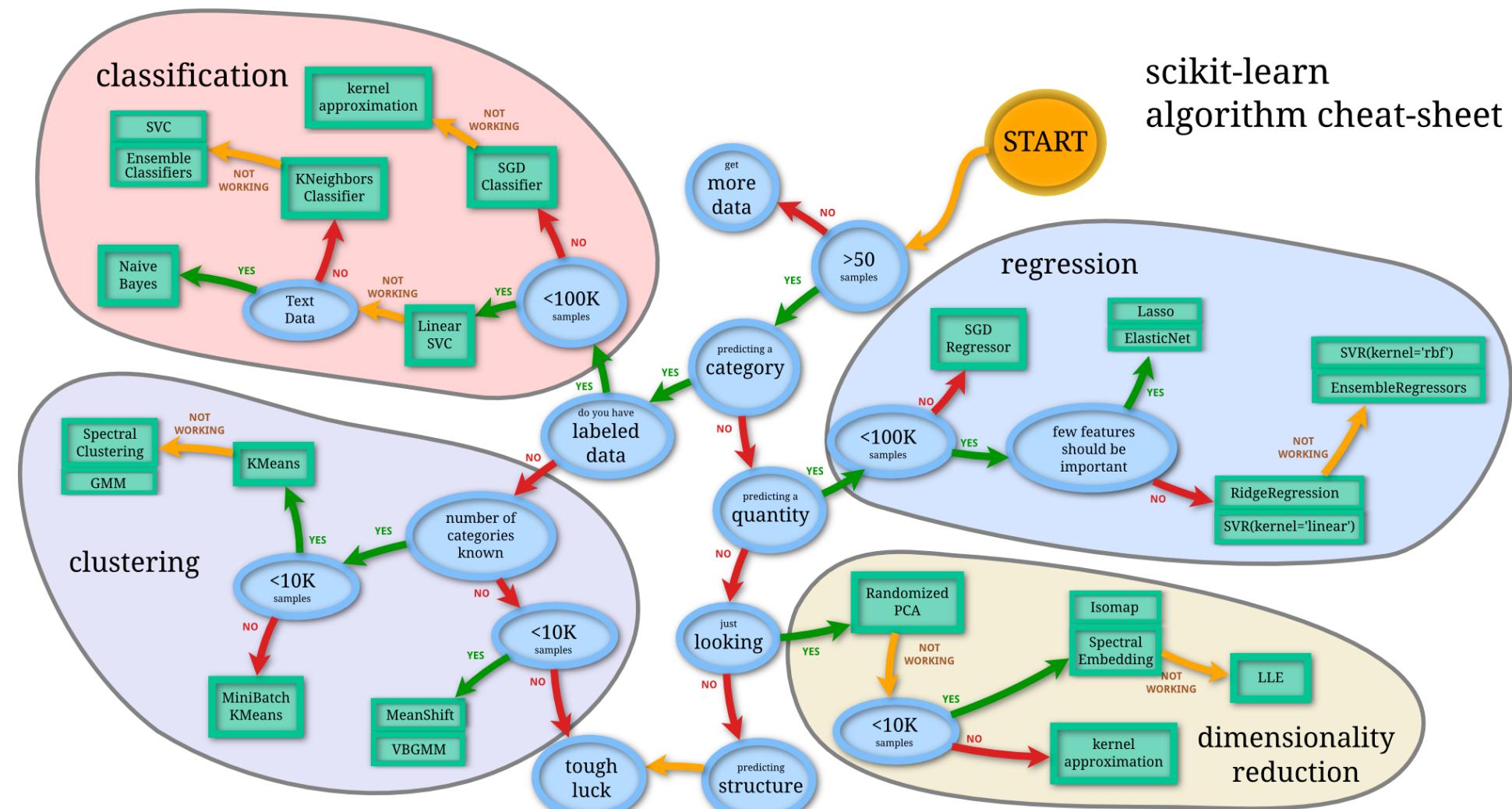
- › Created by Guido Van Rossem in 1991 and emphasizes productivity and code readability
- › Version 3 (but 2.7 is still very popular)
- › “Python is an interpreted, object-oriented, high-level programming language with dynamic semantics”



- Batteries Included: Large collection of built in libraries e.g. SciKit, Pandas, Theano, etc
- Simple and clean syntax
- General purpose language: lots of people outside of data science will be able to work with it



scikit-learn algorithm cheat-sheet



“Data analysis projects today rely on databases, computer and network hardware, and computer and network software. A collection of models and methods for data analysis will be used only if the collection is implemented in a computing environment that makes the models and methods sufficiently efficient to use”



WHEN IT GETS BIGGER



cloudera

DATA SCIENCE PART TIME COURSE

USING DATA SCIENCE PACKAGES

- Packages are libraries of code written to solve a particular set of problems
- In Python there are many related to data science including Pandas, SciKit Learn, Numpy
- These are installed and managed with PIP (Pip Installs Packages)

```
pip install some-package-name
```

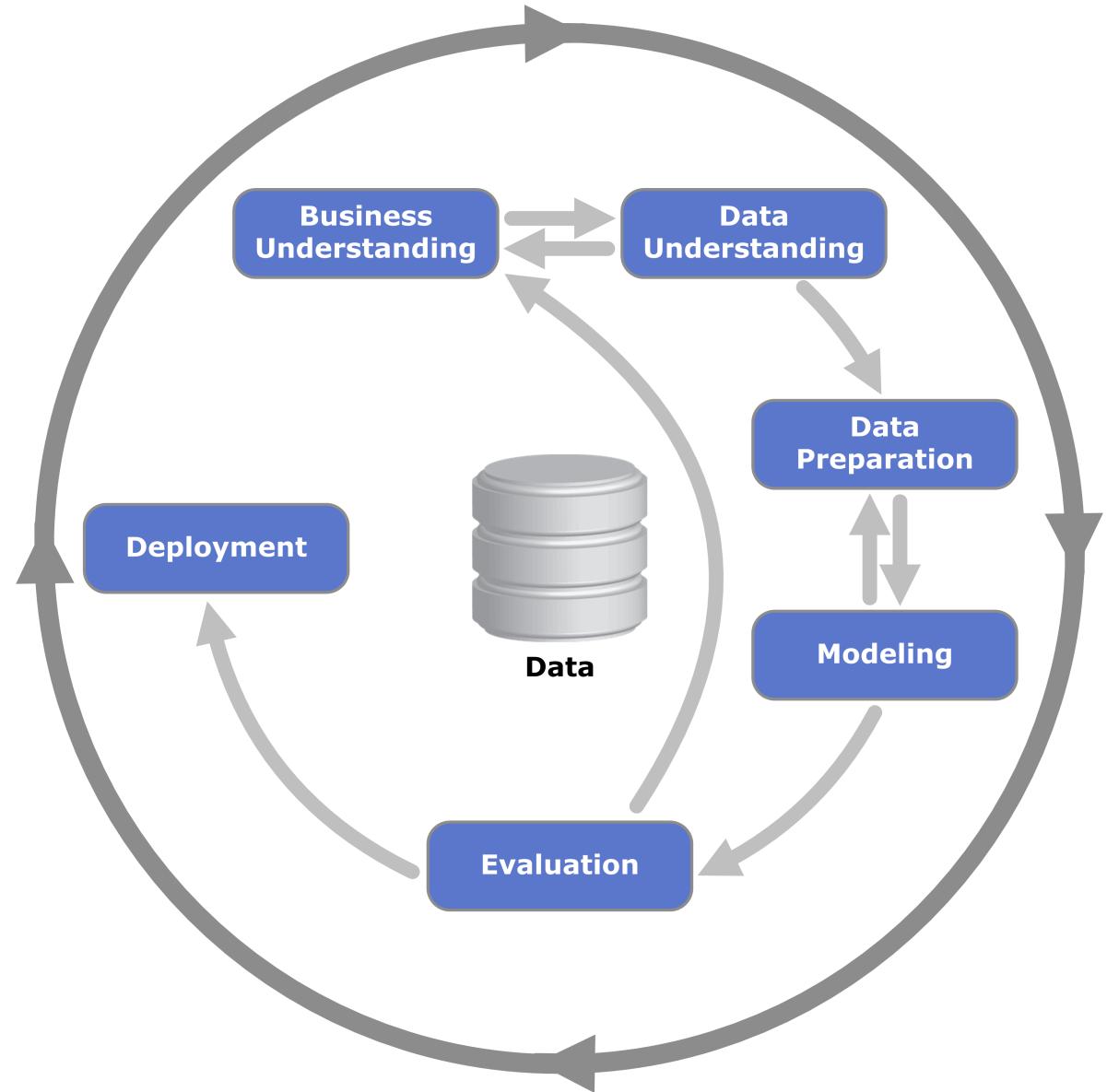
- pandas: manipulate data
- SciPy / NumPy: scientific computing and numerical calculations
- scikit-learn: use machine learning methods
- matplotlib: visualise data
- statsmodels: perform statistical tests
- BeautifulSoup: read in XML and HTML data
- iPython: interactive programming

DATA SCIENCE PART TIME COURSE

WHAT ARE THE STEPS IN A DATA SCIENCE PROJECT?

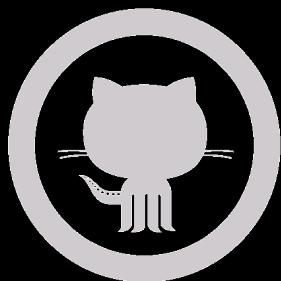
MODELING PROCESS

CRISP-DM
CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING



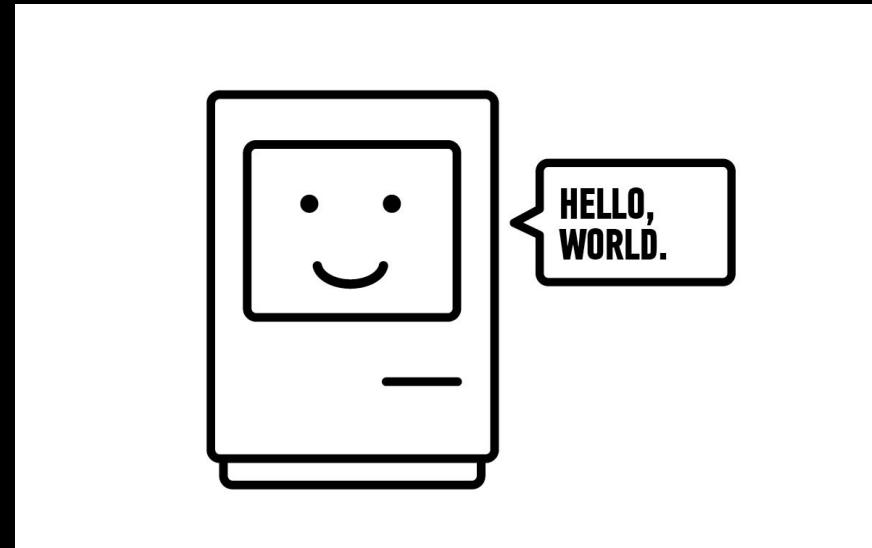
DATA SCIENCE PART TIME COURSE

GIT



DATA SCIENCE PART TIME COURSE

GIT LAB



DATA SCIENCE PART TIME COURSE

PANDAS LAB

DATA SCIENCE - Week 1 Day 2

DISCUSSION TIME

Homework

- › Finish any exercises from today's lab
- › Read Introduction to Statistical Learning