GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 6

Week 1 Lesson 2- Basics of Data Science & Git
Wednesday 12th October

AGENDA 2

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

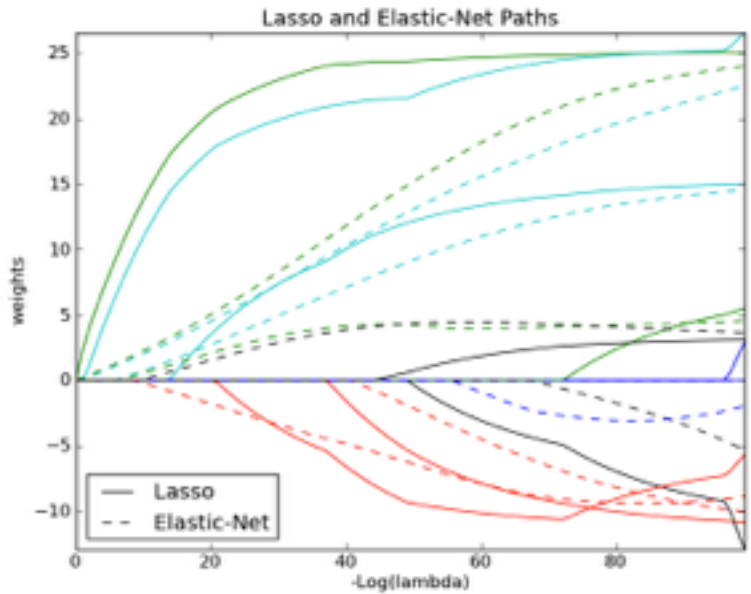
DATA SCIENCE PART TIME COURSE

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

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.

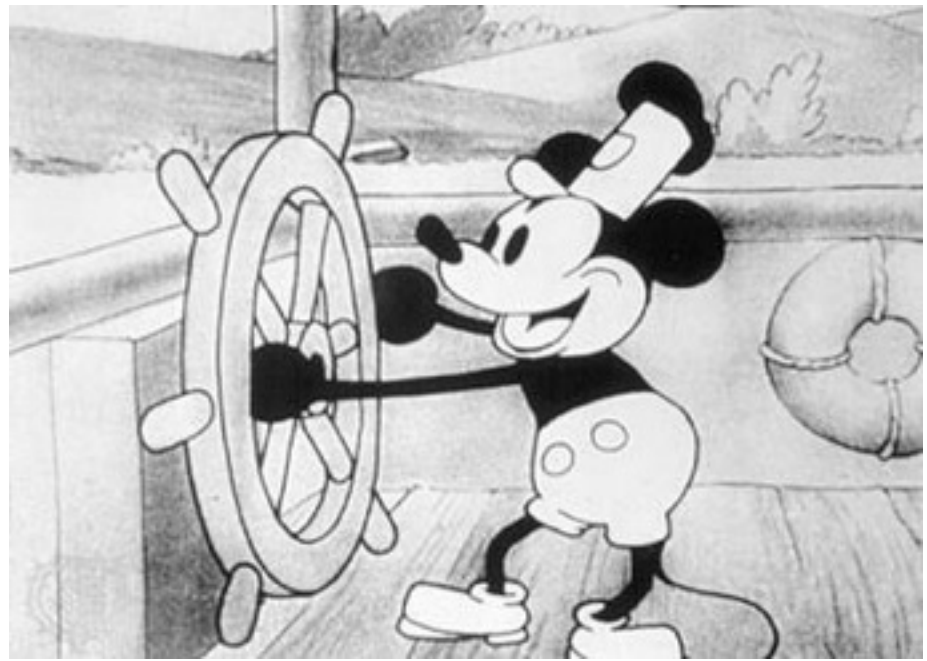
Some literacy of each skill is necessary.

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.

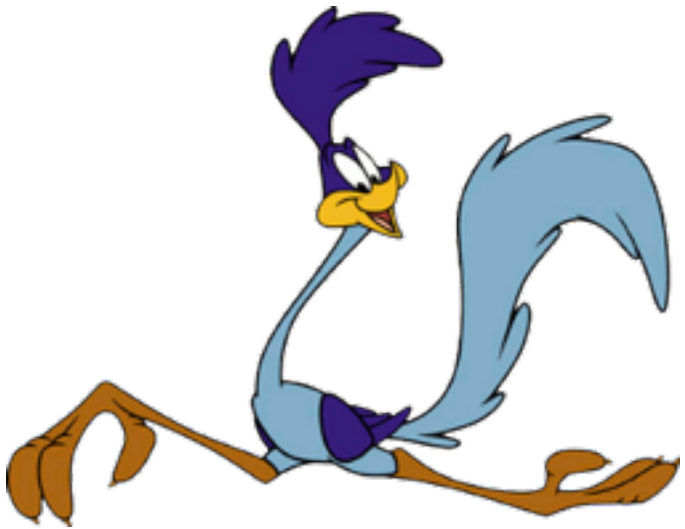
control and governance

vs

access, freedom and use

A good relationship is essential for timely access to resources.

AGILE



Agile methodologies are well suited to data science.

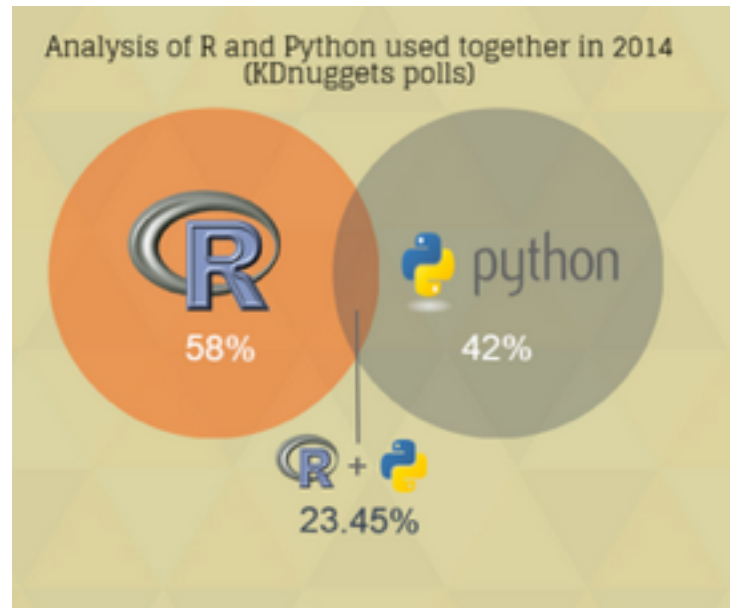
- Incremental development
- Responsive to new information / insights
- Time-boxed goals
- Structure for cooperating with stakeholders

TOOLS OF DATA SCIENCE

PYTHON vs R

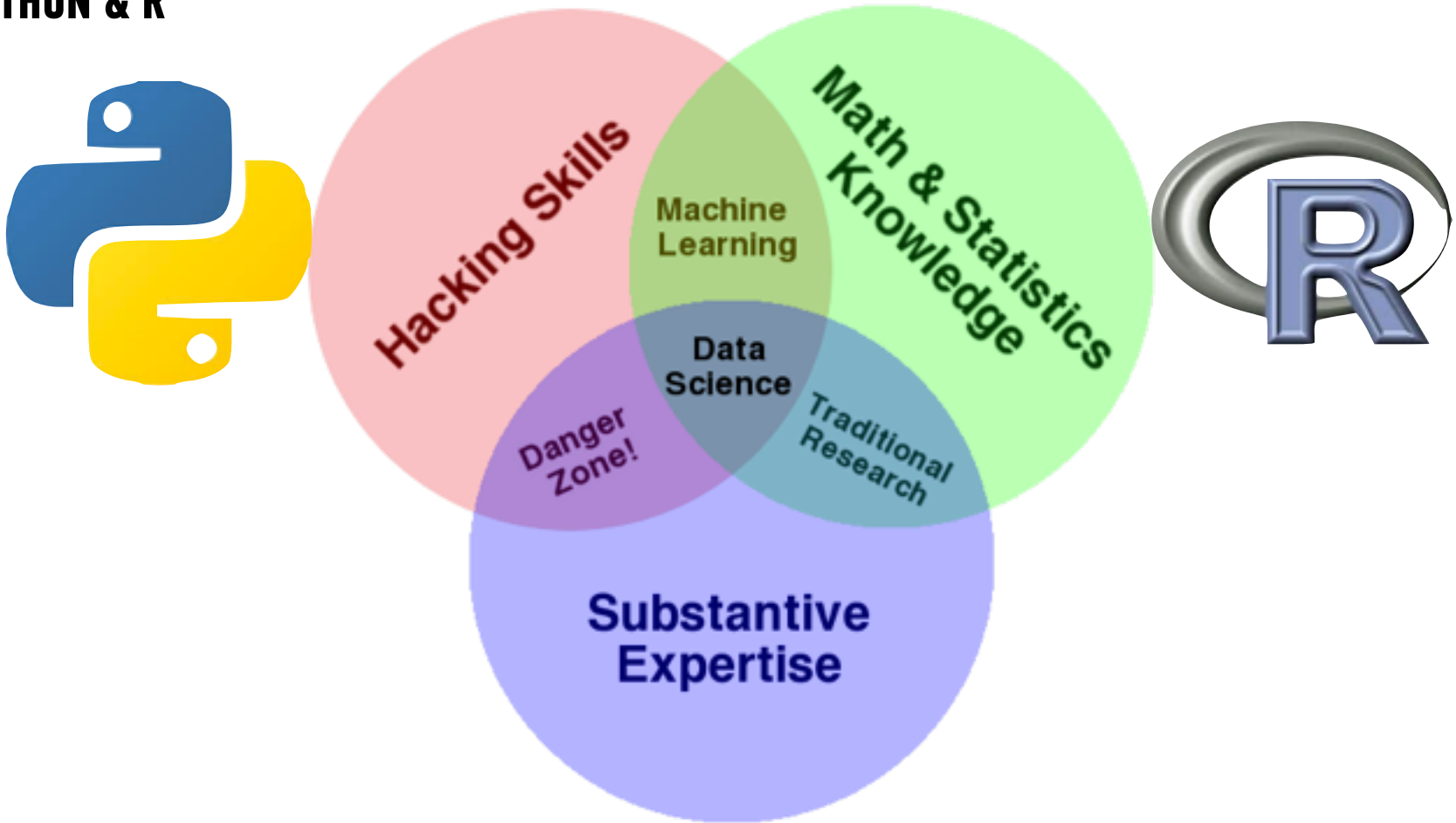
Language Rank	Types	Spectrum Ranking
1. Java		100.0
2. C		99.9
3. C++		99.4
4. Python		96.5
5. C#		91.3
6. R		84.8
7. PHP		84.5
8. JavaScript		83.0
9. Ruby		76.2
10. Matlab		72.4

IEEE Spectrum Survey 2015



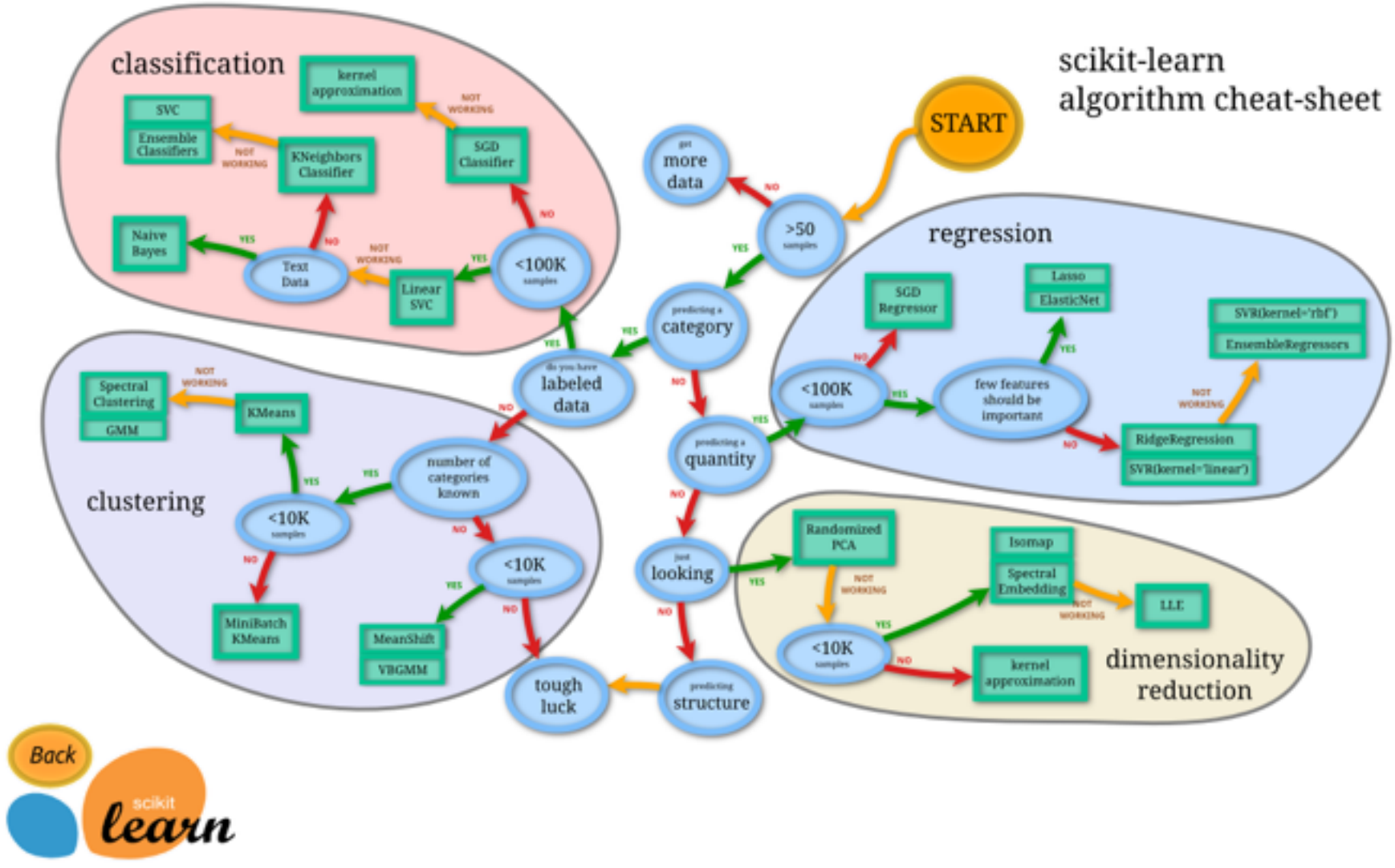
DataCamp Infographic 2015

PYTHON & R


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
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
START
get more data
NO
>50 samples
YES
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
predicting a category
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
do you have labeled data
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
MeanShift
VBGMM
NO
predicting a quantity
YES
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
<10K samples
YES
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

"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

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
- Beautiful Soup: read in XML and HTML data
- iPython: interactive programming

WHAT ARE THE STEPS IN A DATA SCIENCE PROJECT?

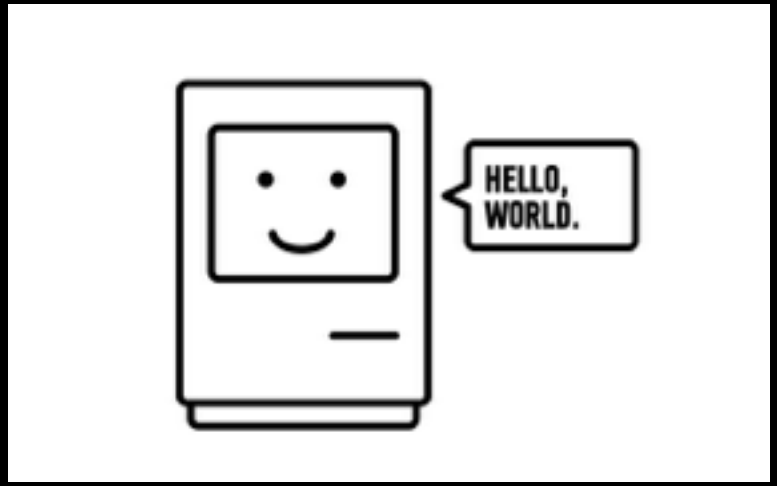
MODELING PROCESS

CRISP-DM
CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

GIT

DATA SCIENCE PART TIME COURSE

GIT LAB



DATA SCIENCE PART TIME COURSE

PANDAS LAB

DISCUSSION TIME

DISCUSSION TIME

Prework

- Readings
  - Metacademy Learning Plan
  - Data Science Handbook
  - An Introduction to Statistical Learning

DISCUSSION TIME

Homework

- Homework1.ipynb (in homework folder of the git repository SYD_DAT_6)
  - Due next Friday
  - I will review within 7 days
  - Counts to letter of completion