REUMass Amherst 2015 Data Science Bootcamp

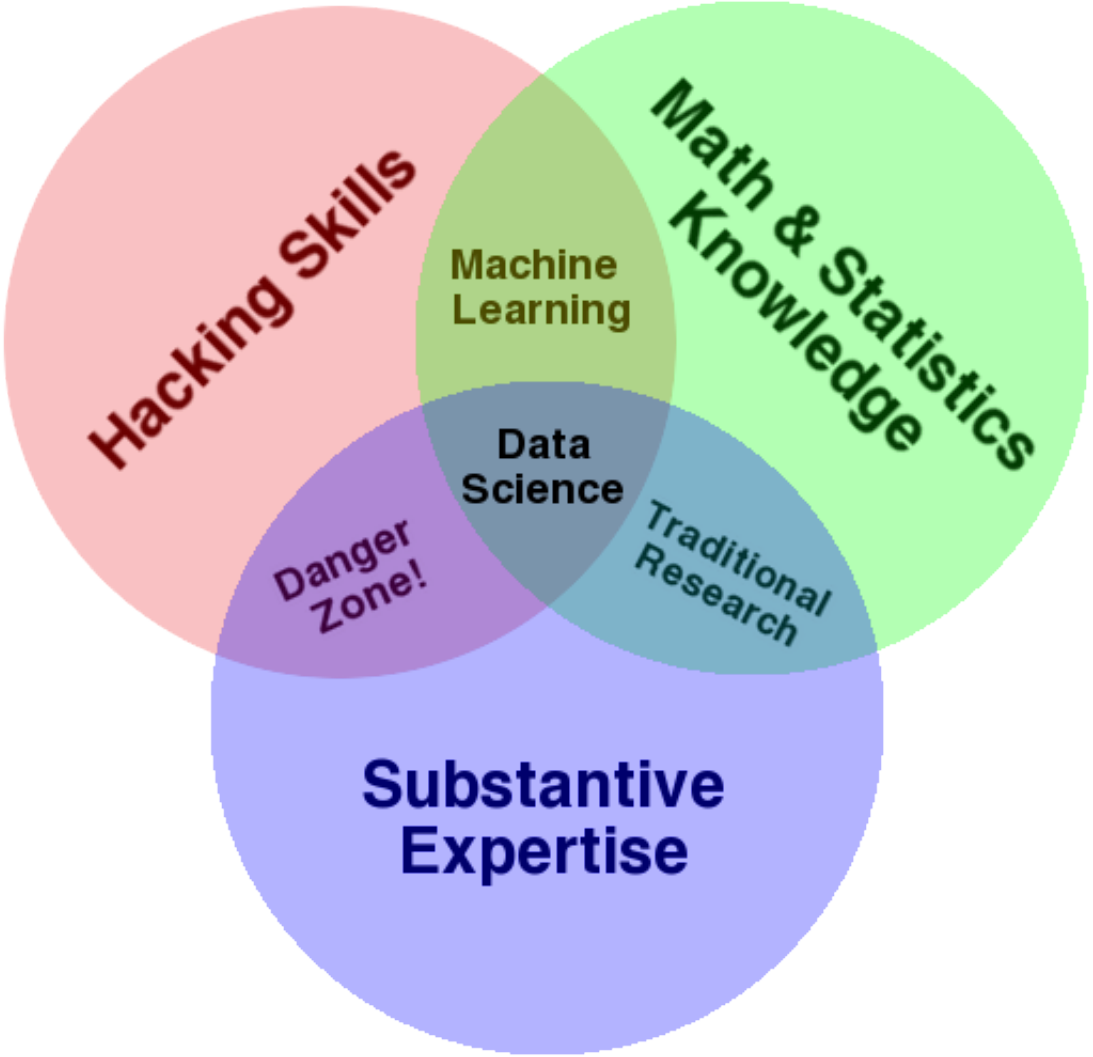
Day 1: Introduction

Prof. Ben Marlin
marlin@cs.umass.edu

Plan for Day 1:

- Introduction to Data Science
- Python Crash Course

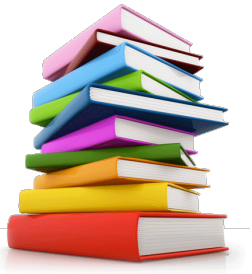
Introduction to Data Science


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

REUMASS Amherst Data Science Bootcamp 2015

Applications



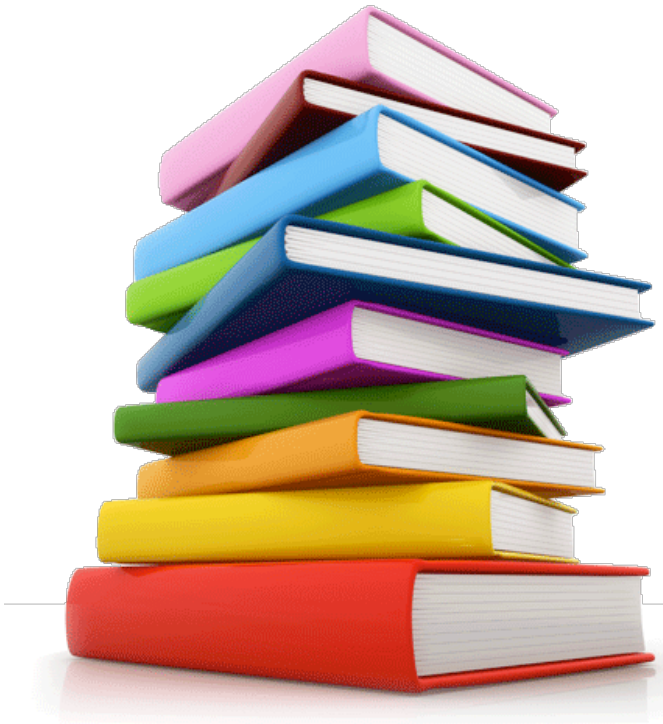



Analytics


Infrastructure



Types of Data:



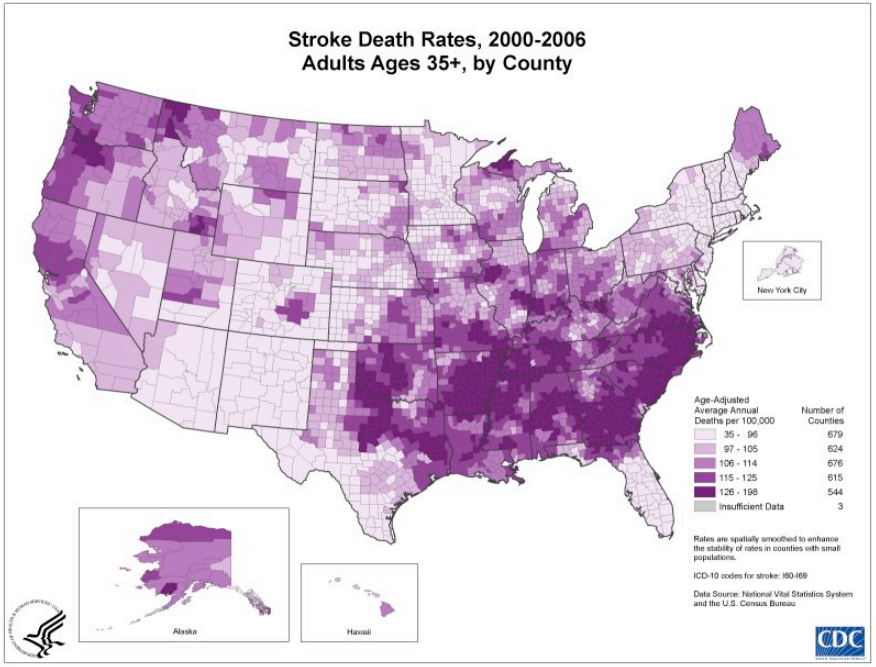
Types of Data:

ex020903.log - Notepad

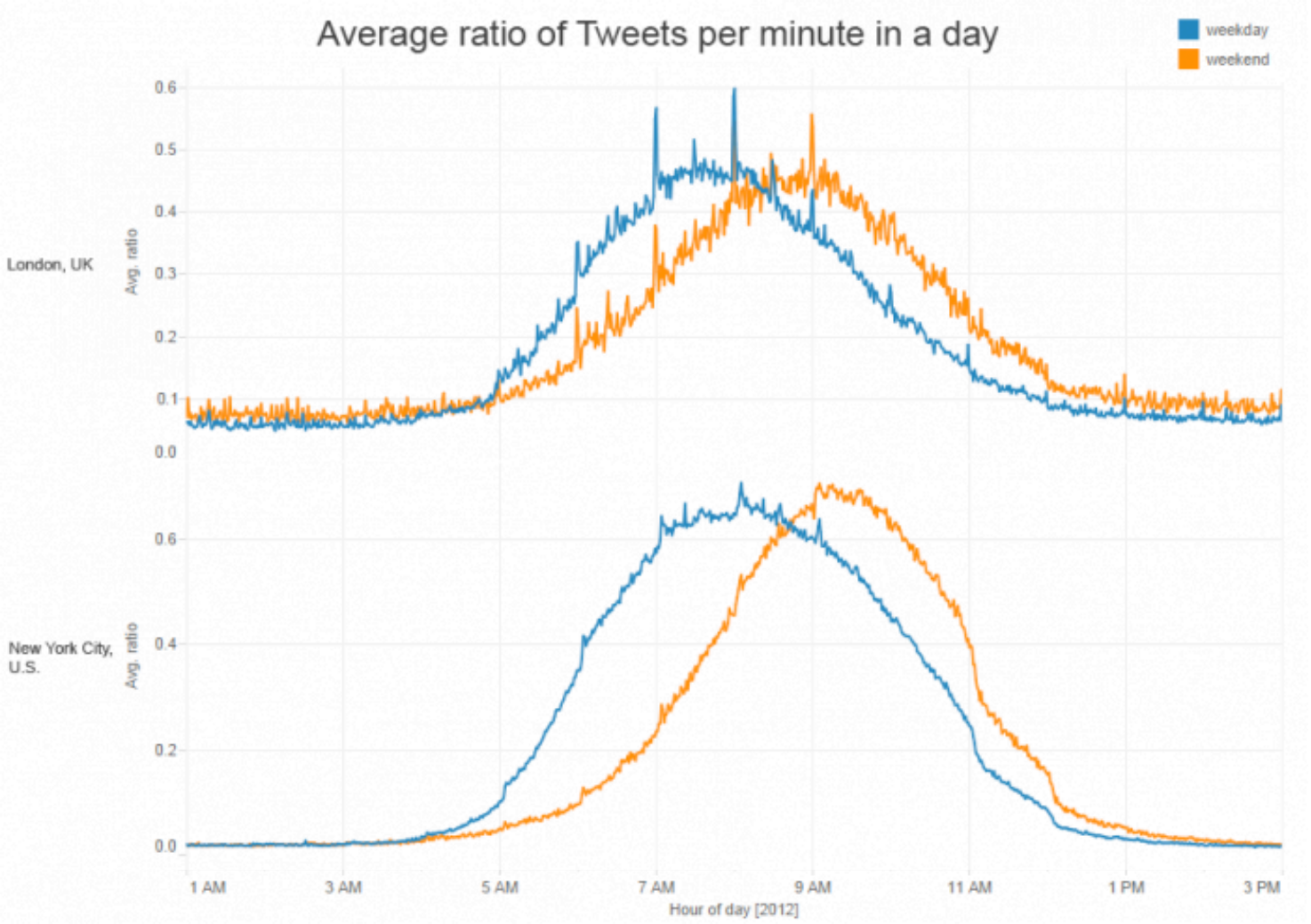
File Edit Format View Help

```
:09:13 24.161.88.191 - W3SVC175 WSH120 80 GET /headerhome.js - 200 0 939 283
:09:13 24.161.88.191 - W3SVC175 WSH120 80 GET /wcstyles.css - 200 0 10213 282
:09:13 24.161.88.191 - W3SVC175 WSH120 80 GET /images/27pct-banner.gif - 200
:09:13 24.161.88.191 - W3SVC175 WSH120 80 GET /images/subway2.jpg - 200 0 114
:09:13 24.161.88.191 - W3SVC175 WSH120 80 GET /images/ligature_spacer6.gif -
:09:13 24.161.88.191 - W3SVC175 WSH120 80 GET /images/wclogov3t.gif - 200 0 2
:09:17 24.161.88.191 - W3SVC175 WSH120 80 GET /whatwedo.html - 200 0 7073 382
:09:17 24.161.88.191 - W3SVC175 WSH120 80 GET /header.js - 200 0 1145 292 0 H
:09:17 24.161.88.191 - W3SVC175 WSH120 80 GET /wcstyles.css - 200 0 10213 295
:09:17 24.161.88.191 - W3SVC175 WSH120 80 GET /images/ligature_spacer6.gif -
:09:17 24.161.88.191 - W3SVC175 WSH120 80 GET /images/whitespacer.gif - 200 0
:09:17 24.161.88.191 - W3SVC175 WSH120 80 GET /images/wclogo100w.gif - 200 0
:09:17 24.161.88.191 - W3SVC175 WSH120 80 GET /footer.js - 200 0 1083 292 0 H
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /wcjournal.html - 200 0 3990 39
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /header.js - 200 0 1145 293 0 H
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /wcstyles.css - 200 0 10213 296
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /images/whitespacer.gif - 200 0
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /images/ligature_spacer6.gif -
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /images/btn-subscribe.gif - 200
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /images/wc-journal-logo-250x50.
:09:26 24.161.88.191 - W3SVC175 WSH120 80 GET /footer.js - 200 0 1083 293 0 H
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /whoweare.html - 200 0 3292 396
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /header.js - 200 0 1145 292 0 H
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /wcstyles.css - 200 0 10213 295
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /images/whitespacer.gif - 200 0
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /images/ligature_spacer6.gif -
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /images/wclogo100w.gif - 200 0
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /images/weidman2-100x125-bw.jpg
:09:32 24.161.88.191 - W3SVC175 WSH120 80 GET /footer.js - 200 0 1083 292 0 H
:09:43 24.161.88.191 - W3SVC175 WSH120 80 GET /contactus.html - 200 0 2406 39
:09:43 24.161.88.191 - W3SVC175 WSH120 80 GET /header.js - 200 0 1145 293 0 H
:09:43 24.161.88.191 - W3SVC175 WSH120 80 GET /wcstyles.css - 200 0 10213 296
```

Types of Data:



Types of Data:



Types of Data:

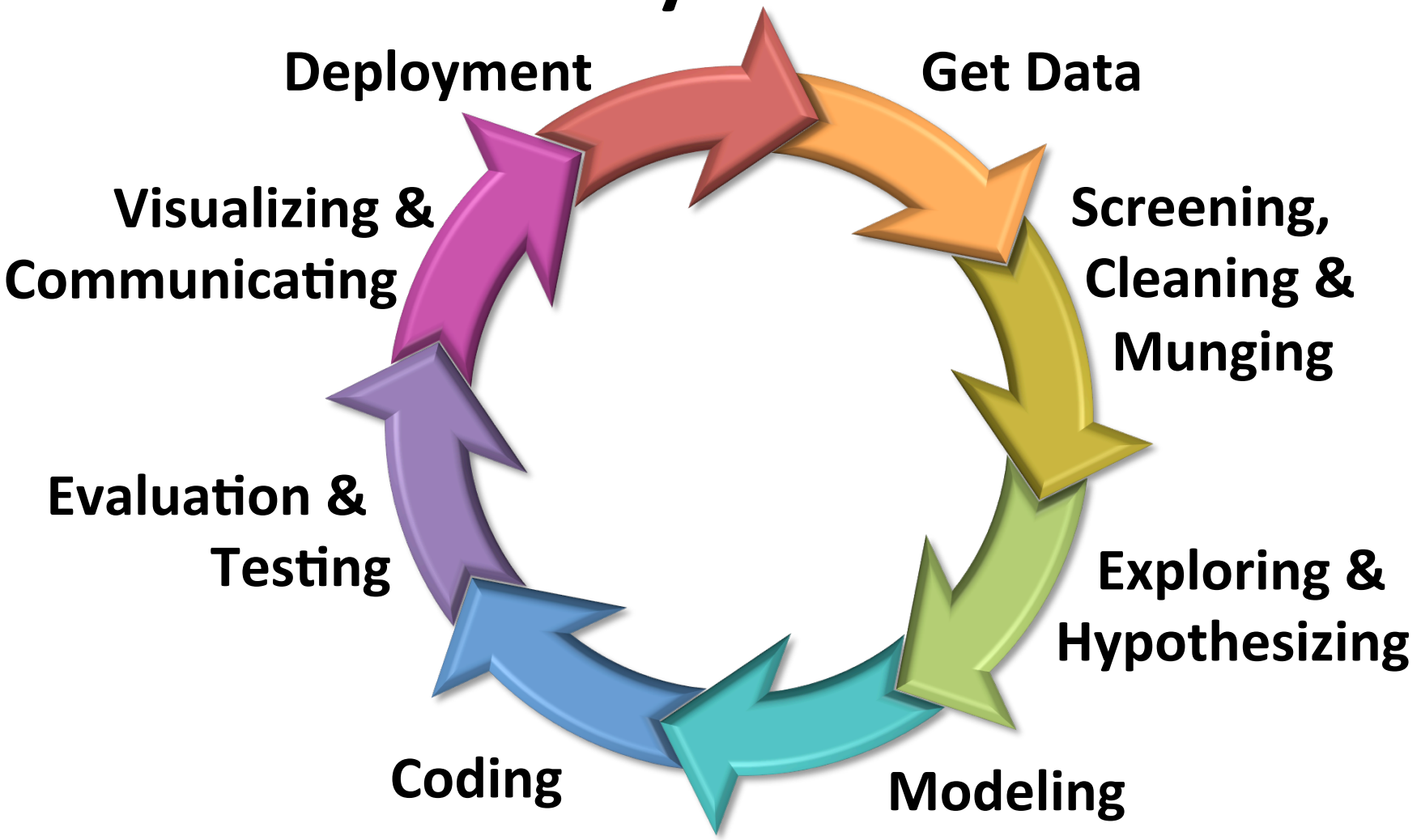






Types of Data:

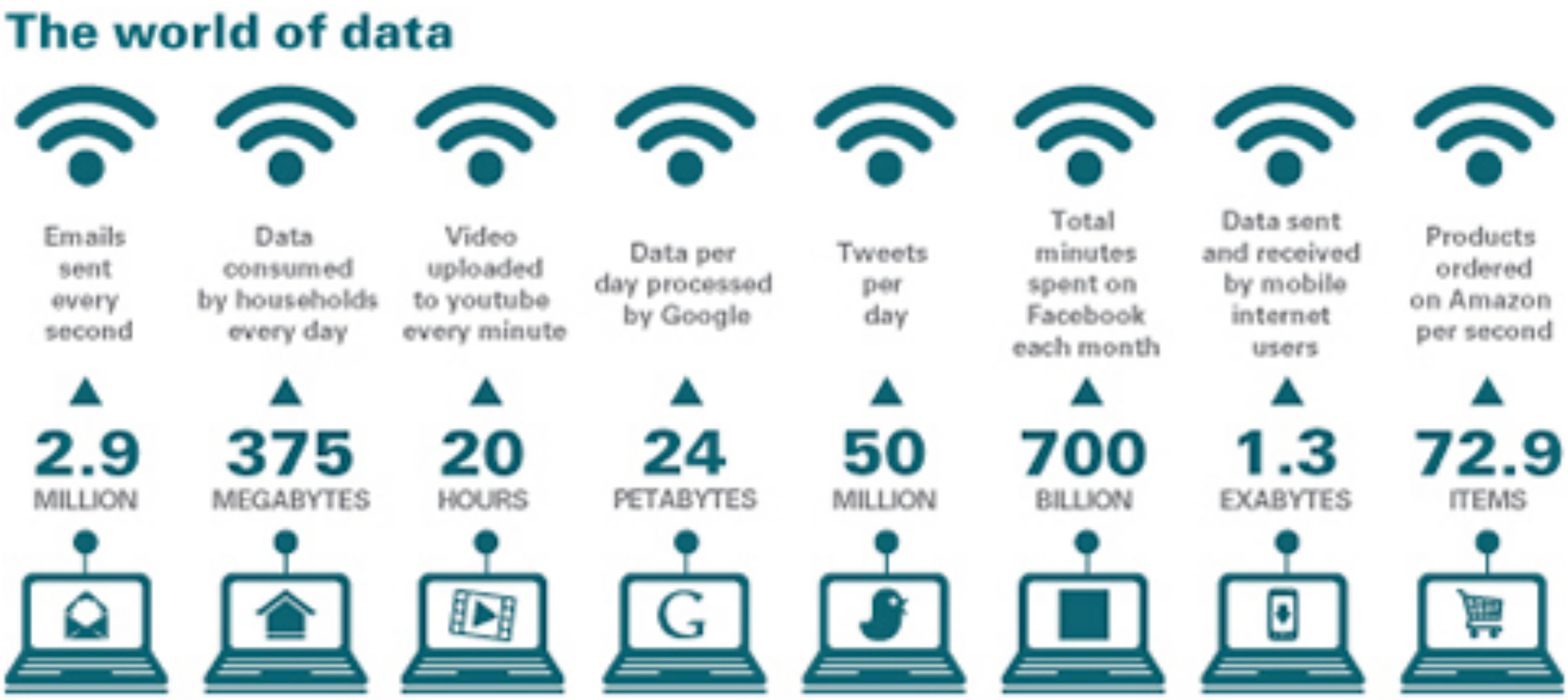
Data Science Life Cycle:



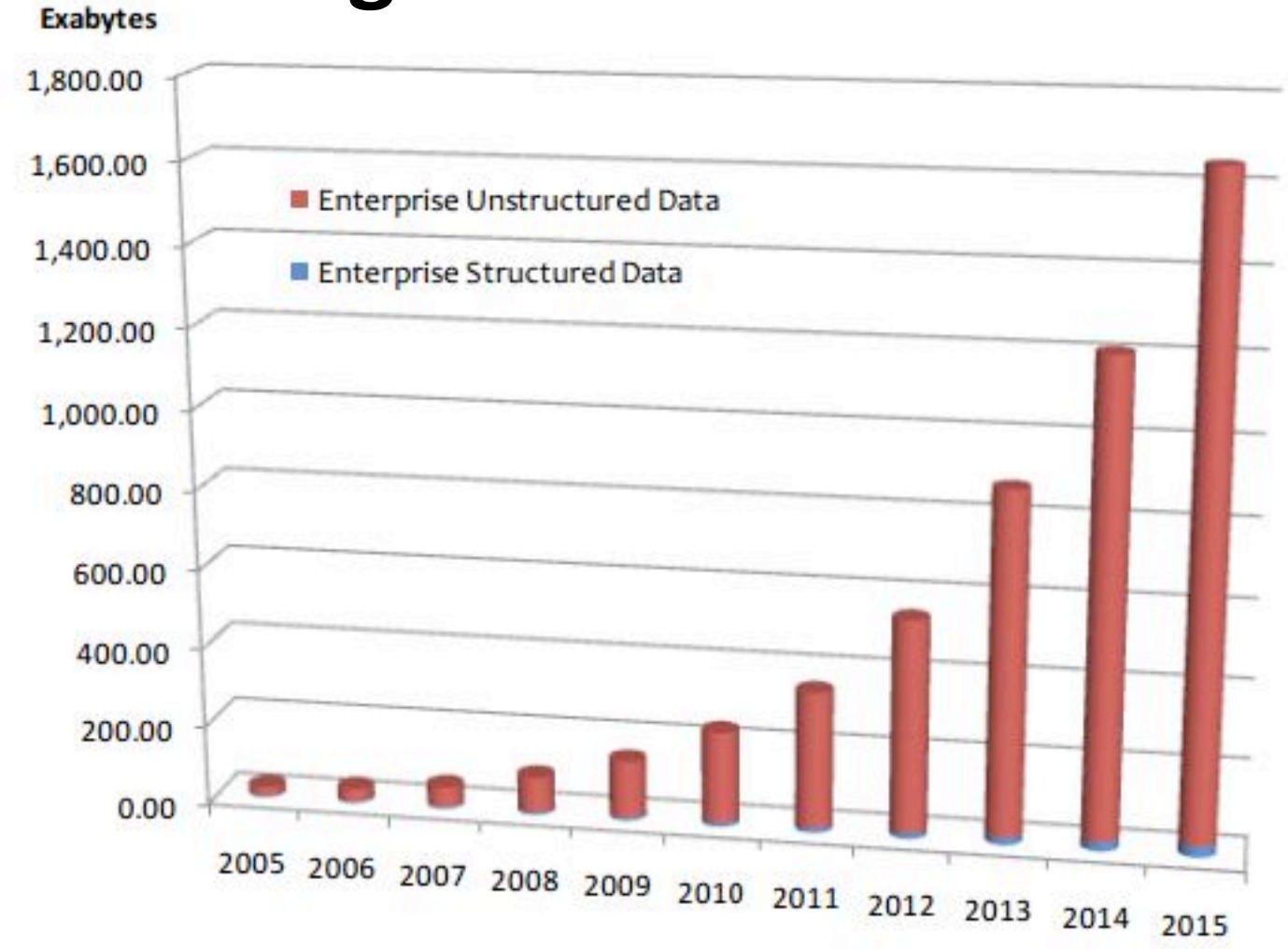
Types of Analyses/Models:

- **Descriptive:** *Quantitatively describing the main features of a data set.*
- **Predictive:** *Extracting patterns from data to make predictions about unknown or future events.*
- **Causal:** *Using data to infer cause and effect relationships between events.*

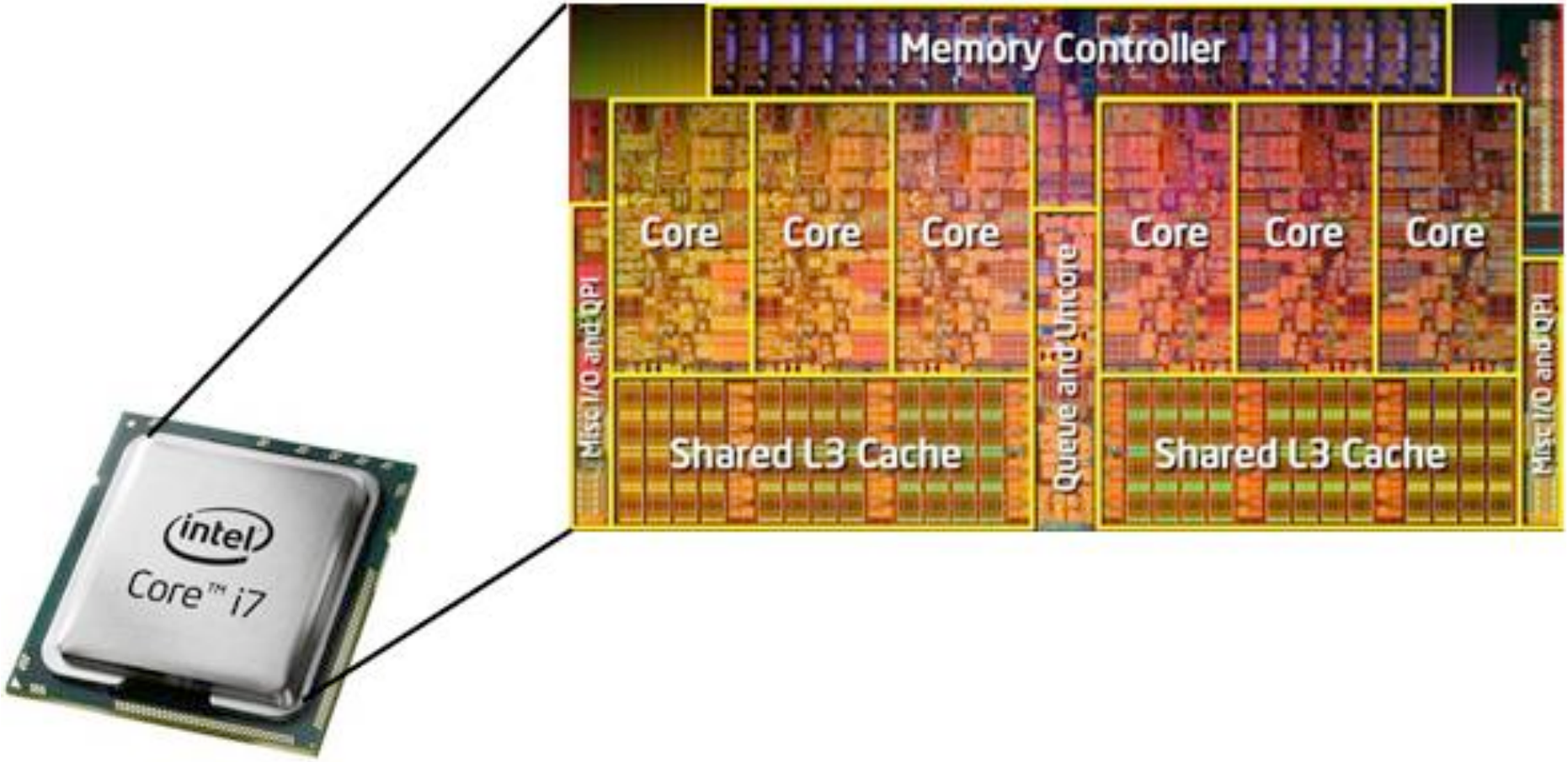
Big Data:


The world of data
Emails sent every second
2.9 MILLION
Data consumed by households every day
375 MEGABYTES
Video uploaded to youtube every minute
20 HOURS
Data per day processed by Google
24 PETABYTES
Tweets per day
50 MILLION
Total minutes spent on Facebook each month
700 BILLION
Data sent and received by mobile internet users
1.3 EXABYTES
Products ordered on Amazon per second
72.9 ITEMS
Source: A collaboration between Good, Oliver Munday and IGM: "The World of Data"

Even More Big Data:



Data Science Infrastructure: Hardware



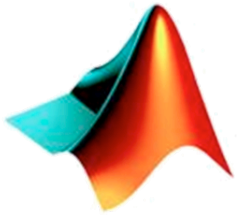
Data Science Infrastructure: Hardware

Data Science Infrastructure: Hardware

Data Science Infrastructure: The Cloud

Data Science Infrastructure: Databases

Data Science Infrastructure: Programming

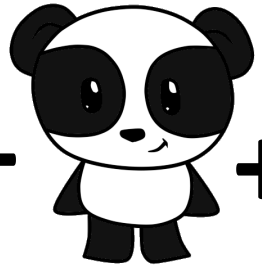


MATLAB

+

WEKA
The University of Waikato

+



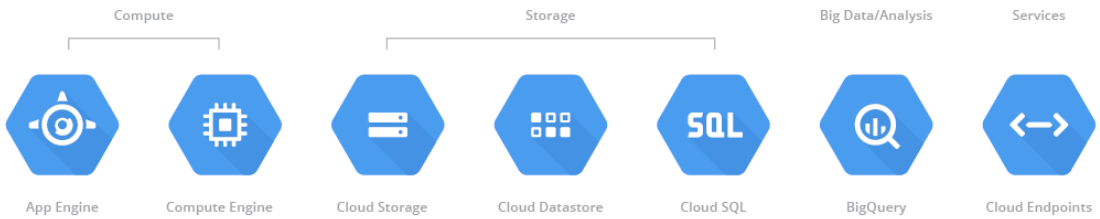
+

Data Science Infrastructure: The Cloud(s)

Microsoft Azure

amazon web services™ | EC2

Google Cloud Platform



Python Crash Course

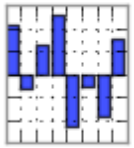
Why Python?

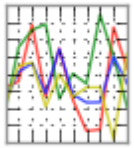
SciPy

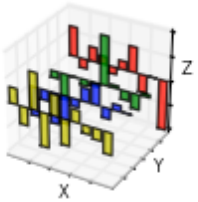
NumPy

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$







scikit learn

matplotlib

IP[y]:
IPython

Python Basics: Python is …

- Interpreted, not compiled.
- Dynamically typed.
- Strongly typed.
- Declarative (with functional elements).
- Object oriented.
- Syntactically very compact.

Python Basics: Variables and Assignment

- All variables are *names* for objects.
- All instances of basic types and data structures are objects.
- Assignment (ie: A=B) sets or changes the object the name refers to.

Python Basics: Control Flow & Functions

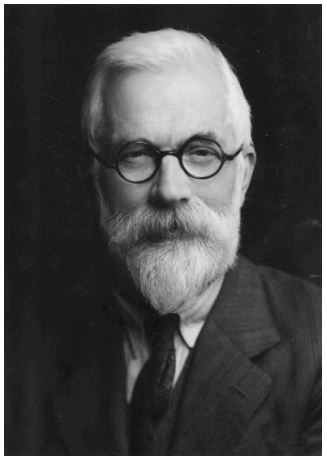
- *If/elif/else*, *for*, *while*, *break*, *continue*
- *def* is used to define functions.
- Code blocks are set off with a *:* (colon) and scope is defined by indentation.
- Math and logic mostly as you expect.

Python Basics: Data Types and Containers

- int, float, boolean (*True, False*)
- Strings (any quotes)
- Tuples and Lists (for sequences)
- Dictionaries *(key-value pairs)*
- Arrays and Matrices *(via NumPy/SciPy)*

Exercises: Fisher's Iris Data

Data set of properties of Iris varieties used by Fisher to learn an Iris classification model.



Iris Virginica

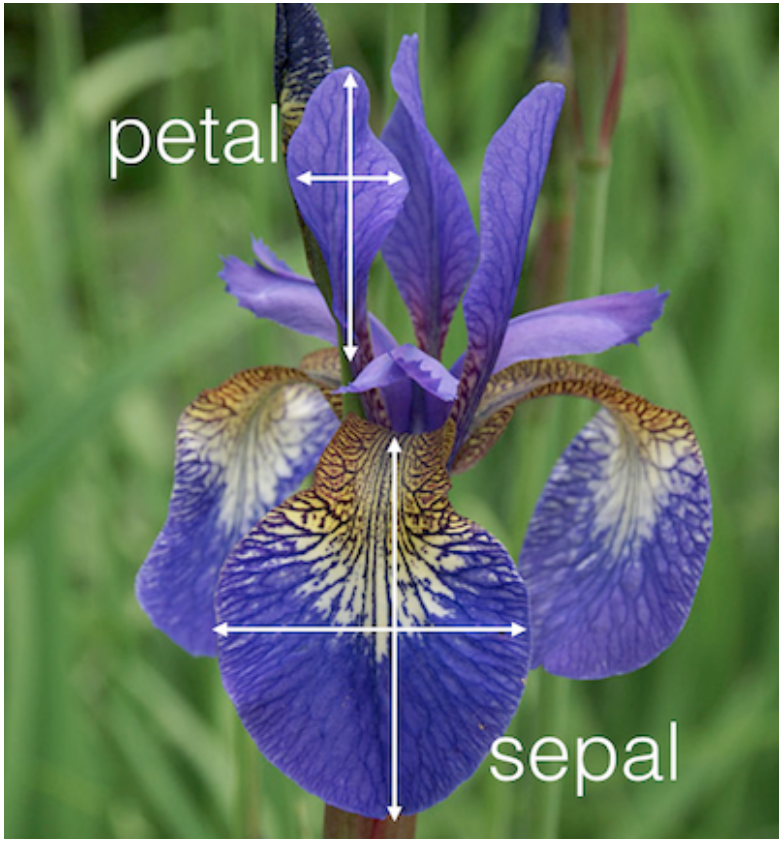
Iris Versacolor

Iris Setosa

Exercises: Fisher's Iris Data

sepal length	sepal width	petal length	petal width	Type
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
⋮	⋮	⋮	⋮	⋮

Represented in Python as one Numpy array for the features, and one for the labels.



Exercises: Fisher's Iris Data

Labels
Classes
Targets

Features

sepal length	sepal width	petal length	petal width	Type
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
⋮	⋮	⋮	⋮	⋮

Data
Cases

Exercises: Fisher's Iris Data

- Q1: How many data cases are there?
- Q2: How many examples of each type are there?
- Q3: What is the average value of each of the features over the whole data set?
- Q4: What is the average value of each of the features per type?