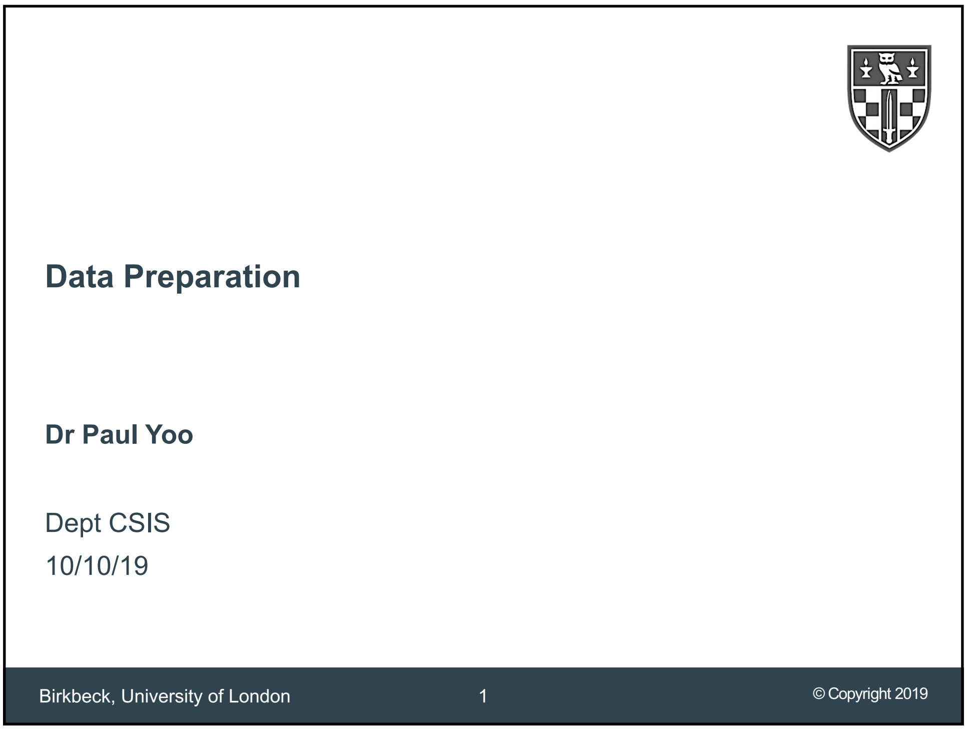
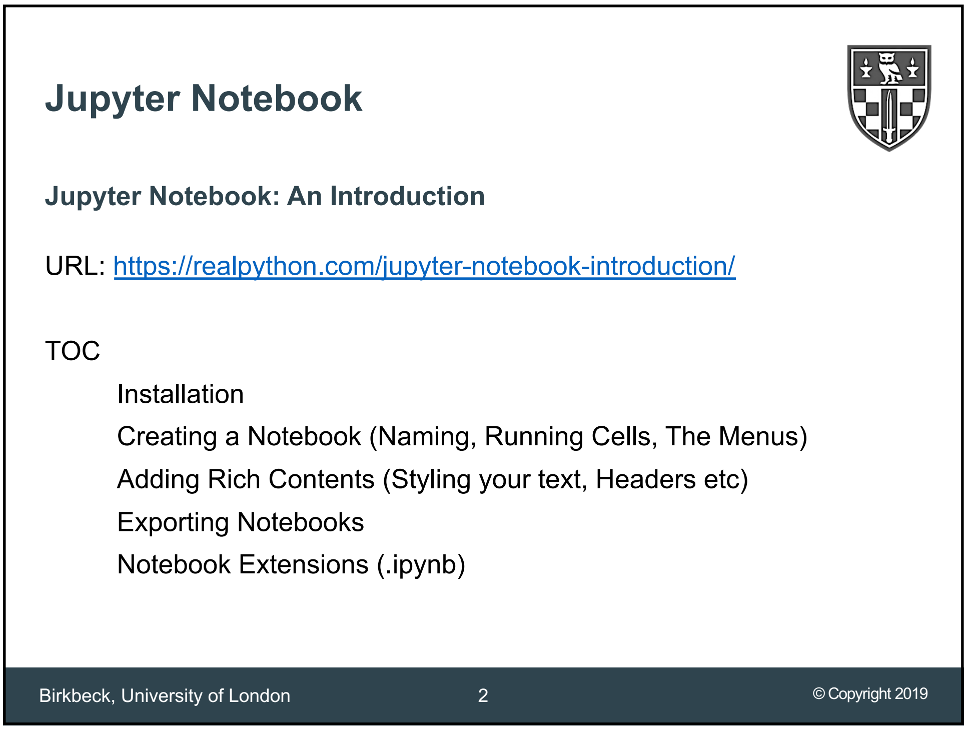


Data Preparation

Dr Paul Yoo

Dept CSIS
10/10/19



Jupyter Notebook

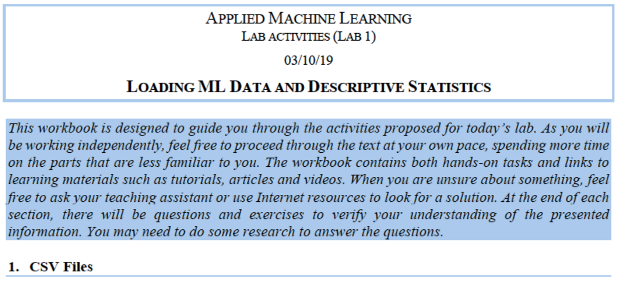
Jupyter Notebook: An Introduction

URL: https://realpython.com/jupyter-notebook-introduction/

TOC

- Installation
- Creating a Notebook (Naming, Running Cells, The Menus)
- Adding Rich Contents (Styling your text, Headers etc)
- Exporting Notebooks
- Notebook Extensions (.ipynb)

Lab Questions

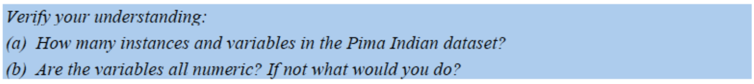


APPLIED MACHINE LEARNING
LAB ACTIVITIES (LAB 1)
03/10/19

LOADING ML DATA AND DESCRIPTIVE STATISTICS

This workbook is designed to guide you through the activities proposed for today's lab. As you will be working independently, feel free to proceed through the text at your own pace, spending more time on the parts that are less familiar to you. The workbook contains both hands-on tasks and links to learning materials such as tutorials, articles and videos. When you are unsure about something, feel free to ask your teaching assistant or use Internet resources to look for a solution. At the end of each section, there will be questions and exercises to verify your understanding of the presented information. You may need to do some research to answer the questions.

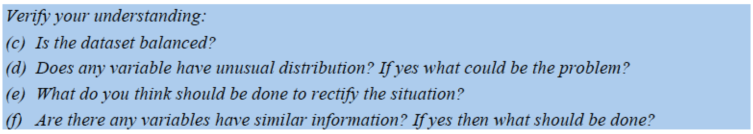
1. **CSV Files**



Verify your understanding:

(a) How many instances and variables in the Pima Indian dataset?

(b) Are the variables all numeric? If not what would you do?



Verify your understanding:

(c) Is the dataset balanced?

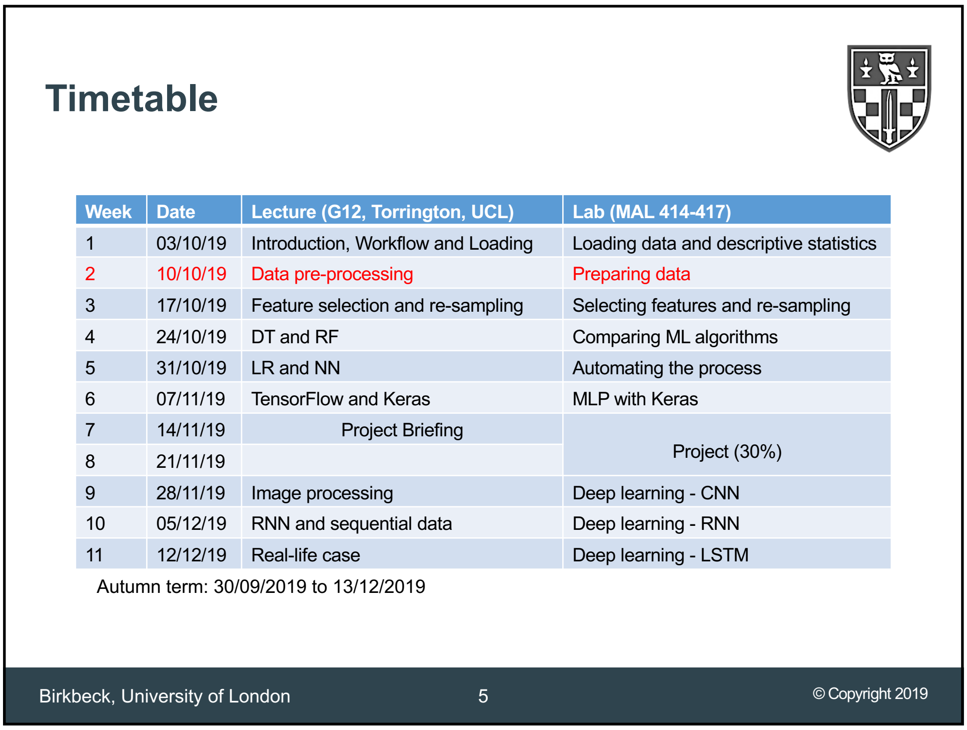
(d) Does any variable have unusual distribution? If yes what could be the problem?

(e) What do you think should be done to rectify the situation?

(f) Are there any variables have similar information? If yes then what should be done?

Module Auditors

The module is currently oversubscribed..

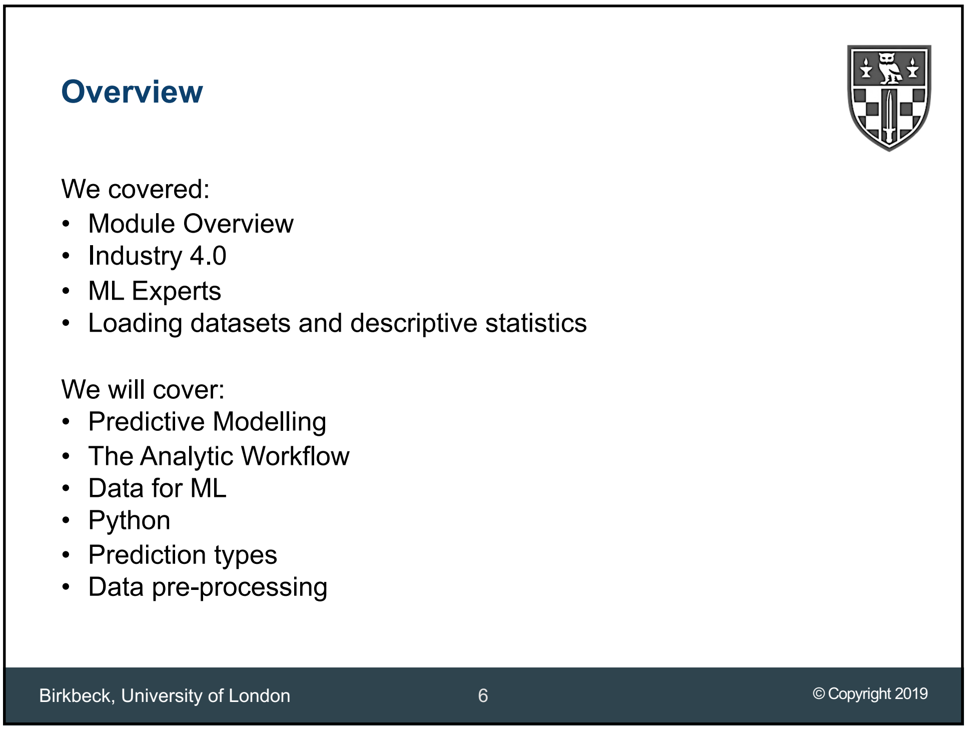


Timetable

Week	Date	Lecture (G12, Torrington, UCL)	Lab (MAL 414-417)
1	03/10/19	Introduction, Workflow and Loading	Loading data and descriptive statistics
2	10/10/19	Data pre-processing	Preparing data
3	17/10/19	Feature selection and re-sampling	Selecting features and re-sampling
4	24/10/19	DT and RF	Comparing ML algorithms
5	31/10/19	LR and NN	Automating the process
6	07/11/19	TensorFlow and Keras	MLP with Keras
7	14/11/19	Project Briefing	Project (30%)
8	21/11/19		
9	28/11/19	Image processing	Deep learning - CNN
10	05/12/19	RNN and sequential data	Deep learning - RNN
11	12/12/19	Real-life case	Deep learning - LSTM

Autumn term: 30/09/2019 to 13/12/2019

Birkbeck, University of London 5 © Copyright 2019



Overview

We covered:

- Module Overview
- Industry 4.0
- ML Experts
- Loading datasets and descriptive statistics

We will cover:

- Predictive Modelling
- The Analytic Workflow
- Data for ML
- Python
- Prediction types
- Data pre-processing

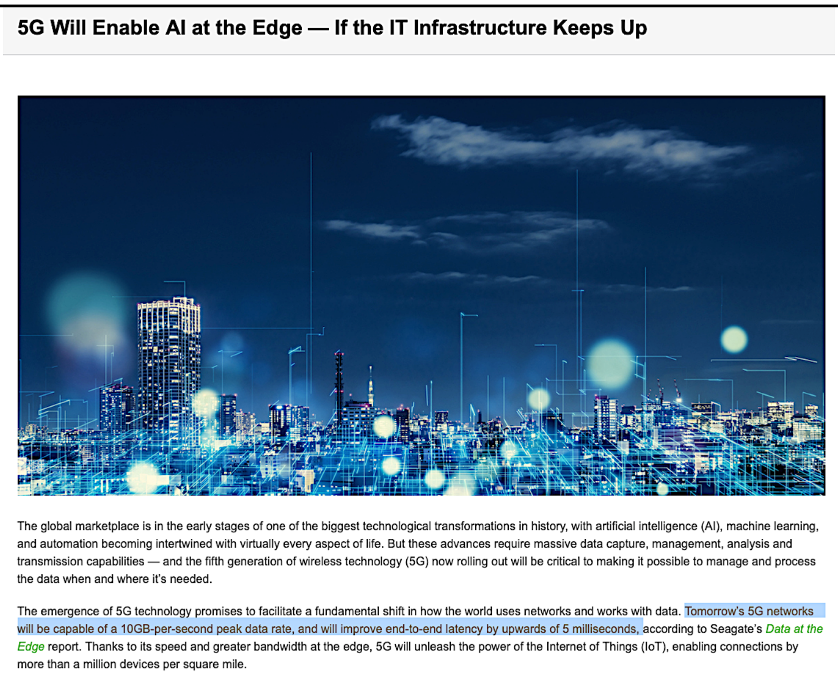
Birkbeck, University of London 6 © Copyright 2019

Discussion

What is 5G?

In small groups discuss what you think 5G is and their impacts on ML are.

You have **5 minutes** and then we will discuss your answers.



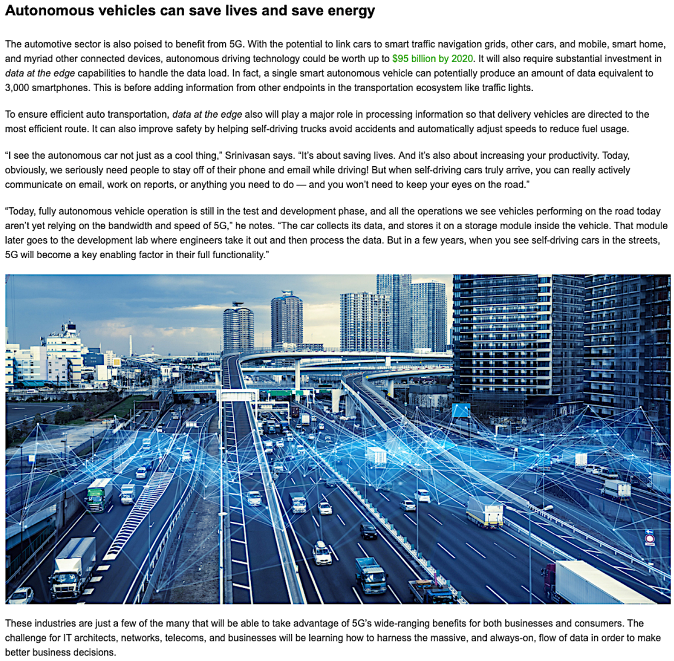
5G Will Enable AI at the Edge — If the IT Infrastructure Keeps Up

The global marketplace is in the early stages of one of the biggest technological transformations in history, with artificial intelligence (AI), machine learning, and automation becoming intertwined with virtually every aspect of life. But these advances require massive data capture, management, analysis and transmission capabilities — and the fifth generation of wireless technology (5G) now rolling out will be critical to making it possible to manage and process the data when and where it's needed.

The emergence of 5G technology promises to facilitate a fundamental shift in how the world uses networks and works with data. Tomorrow's 5G networks will be capable of a 10GB-per-second peak data rate, and will improve end-to-end latency by upwards of 5 milliseconds, according to Seagate's *Data at the Edge* report. Thanks to its speed and greater bandwidth at the edge, 5G will unleash the power of the Internet of Things (IoT), enabling connections by more than a million devices per square mile.

Source: https://blog.seagate.com/enterprises/5g-enable-ai-edge-it-infrastructure/

4G takes tens of ms. For a vehicle moving at 100 Kmph it is a too long time, and can get fatal. 5G brings down latency to < 1ms.



Autonomous vehicles can save lives and save energy

The automotive sector is also poised to benefit from 5G. With the potential to link cars to smart traffic navigation grids, other cars, and mobile, smart home, and myriad other connected devices, autonomous driving technology could be worth up to $95 billion by 2020. It will also require substantial investment in *data at the edge* capabilities to handle the data load. In fact, a single smart autonomous vehicle can potentially produce an amount of data equivalent to 3,000 smartphones. This is before adding information from other endpoints in the transportation ecosystem like traffic lights.

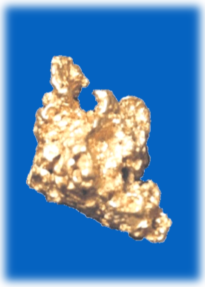
To ensure efficient auto transportation, *data at the edge* also will play a major role in processing information so that delivery vehicles are directed to the most efficient route. It can also improve safety by helping self-driving trucks avoid accidents and automatically adjust speeds to reduce fuel usage.

“I see the autonomous car not just as a cool thing,” Srinivasan says. “It’s about saving lives. And it’s also about increasing your productivity. Today, obviously, we seriously need people to stay off of their phone and email while driving! But when self-driving cars truly arrive, you can really actively communicate on email, work on reports, or anything you need to do — and you won’t need to keep your eyes on the road.”

“Today, fully autonomous vehicle operation is still in the test and development phase, and all the operations we see vehicles performing on the road today aren’t yet relying on the bandwidth and speed of 5G,” he notes. “The car collects its data, and stores it on a storage module inside the vehicle. That module later goes to the development lab where engineers take it out and then process the data. But in a few years, when you see self-driving cars in the streets, 5G will become a key enabling factor in their full functionality.”

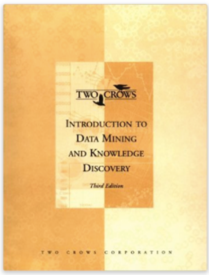
These industries are just a few of the many that will be able to take advantage of 5G’s wide-ranging benefits for both businesses and consumers. The challenge for IT architects, networks, telecoms, and businesses will be learning how to harness the massive, and always-on, flow of data in order to make better business decisions.

Predictive Modelling



“Most of the big payoff [in data mining] has been in predictive modeling.”

– Herb Edelstein



TWO CROWS

Introduction to Data Mining and Knowledge Discovery

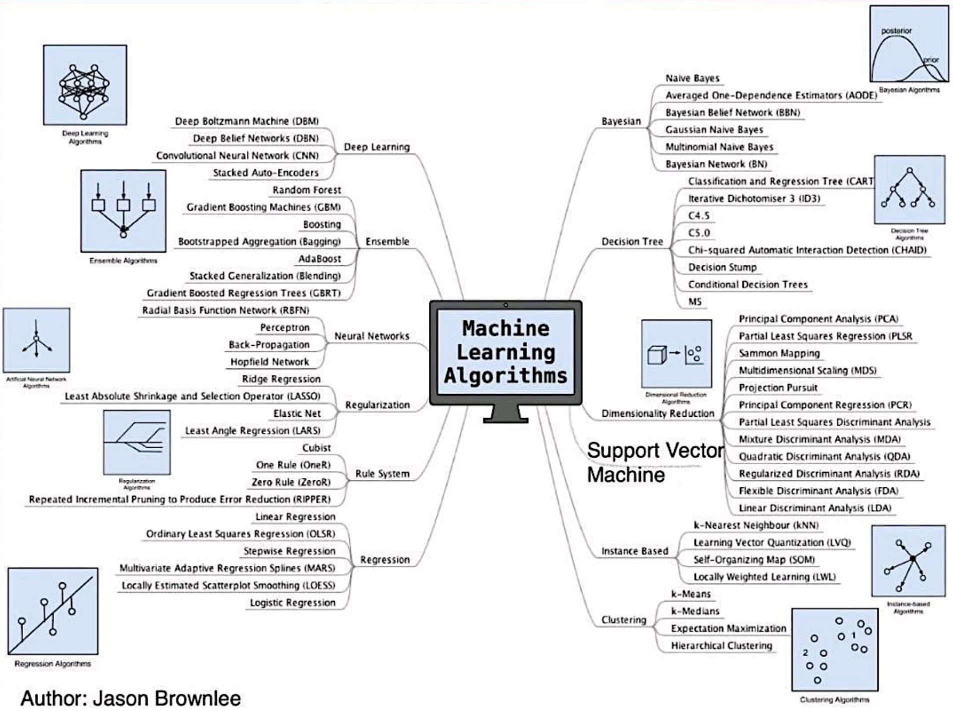
Third Edition

Two Crows Corporation

This module focuses on a specific sub-field of machine learning called predictive modeling.

Predictive Modelling ML Steps

1. **Define Problem:** Investigate and characterise the problem in order to better understand the goals of the project.
2. **Analyse Data:** Use descriptive statistics and visualisation to better understand the data you have available.
3. **Prepare Data:** Use data transforms in order to better expose the structure of the prediction problem to modeling algorithms.
4. **Evaluate Algorithms:** Design a test harness to evaluate a number of standard algorithms on the data and select the top few to investigate further.
5. **Improve Results:** Use algorithm tuning and ensemble methods to get the most out of well-performing algorithms on your data.
6. **Present Results:** Finalise the model, make predictions and present results.

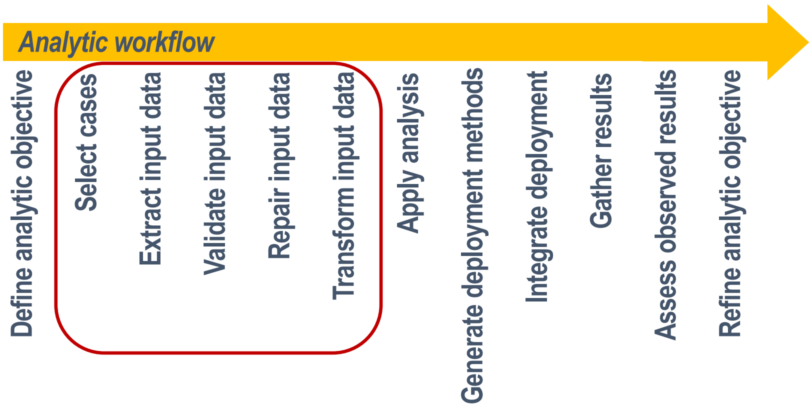


Machine Learning Algorithms

- Deep Learning
  - Deep Boltzmann Machine (DBM)
  - Deep Belief Networks (DBN)
  - Convolutional Neural Network (CNN)
  - Stacked Auto-Encoders
- Ensemble
  - Random Forest
  - Gradient Boosting Machines (GBM)
  - Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Stacked Generalization (Blending)
  - Gradient Boosted Regression Trees (GBRT)
- Neural Networks
  - Radial Basis Function Network (RBFN)
  - Perceptron
  - Back-Propagation
  - Hopfield Network
- Regularization
  - Ridge Regression
  - Least Absolute Shrinkage and Selection Operator (LASSO)
  - Elastic Net
  - Least Angle Regression (LARS)
- Rule System
  - Cubist
  - One Rule (OneR)
  - Zero Rule (ZeroR)
  - Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
- Regression
  - Linear Regression
  - Ordinary Least Squares Regression (OLSR)
  - Stepwise Regression
  - Multivariate Adaptive Regression Splines (MARS)
  - Locally Estimated Scatterplot Smoothing (LOESS)
  - Logistic Regression
- Bayesian
  - Naive Bayes
  - Averaged One-Dependence Estimators (AODE)
  - Bayesian Belief Network (BBN)
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Bayesian Network (BN)
- Decision Tree
  - Classification and Regression Tree (CART
  - Iterative Dichotomiser 3 (ID3)
  - C4.5
  - C5.0
  - Chi-squared Automatic Interaction Detection (CHAID)
  - Decision Stump
  - Conditional Decision Trees
  - M5
- Dimensionality Reduction
  - Principal Component Analysis (PCA)
  - Partial Least Squares Regression (PLSR
  - Sammon Mapping
  - Multidimensional Scaling (MDS)
  - Projection Pursuit
  - Principal Component Regression (PCR)
  - Partial Least Squares Discriminant Analysis
  - Mixture Discriminant Analysis (MDA)
  - Quadratic Discriminant Analysis (QDA)
  - Regularized Discriminant Analysis (RDA)
  - Flexible Discriminant Analysis (FDA)
  - Linear Discriminant Analysis (LDA)
- Support Vector Machine
- Instance Based
  - k-Nearest Neighbour (kNN)
  - Learning Vector Quantization (LVQ)
  - Self-Organizing Map (SOM)
  - Locally Weighted Learning (LWL)
- Clustering
  - k-Means
  - k-Medians
  - Expectation Maximization
  - Hierarchical Clustering

Author: Jason Brownlee

The Analytic Workflow



Analytic workflow

- Define analytic objective
- Select cases
- Extract input data
- Validate input data
- Repair input data
- Transform input data
- Apply analysis
- Generate deployment methods
- Integrate deployment
- Gather results
- Assess observed results
- Refine analytic objective

Source: SAS.com

How can you find data for ML?

There are many ways from which you can get data.

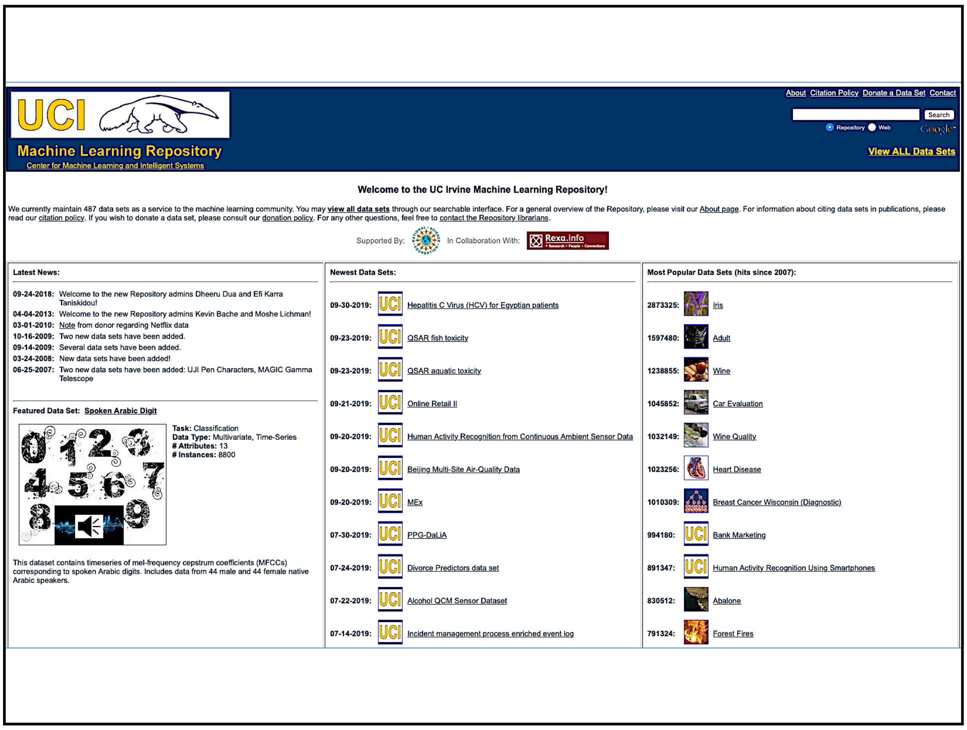
- From your company which provides you data for their task
- From an investor for who you are making something
- From online scrapping
- From different sites or repositories

UCI Machine Learning repository

http://archive.ics.uci.edu/ml/index.php

- Small – fit into memory and model them in reasonable time
- Well behaved – don't need to do a lot of feature engineering
- Benchmarks – many people have used them

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contact
Search
● Repository ○ Web
Google
View ALL Data Sets

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 487 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians.

Supported By: In Collaboration With: Rexa.info

Latest News:

09-24-2018:	Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!
04-04-2013:	Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010:	Note from donor regarding Netflix data
10-16-2009:	Two new data sets have been added.
09-14-2009:	Several data sets have been added.
03-24-2008:	New data sets have been added!
06-25-2007:	Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: Spoken Arabic Digit

Task: Classification
Data Type: Multivariate, Time-Series
# Attributes: 13
# Instances: 8800

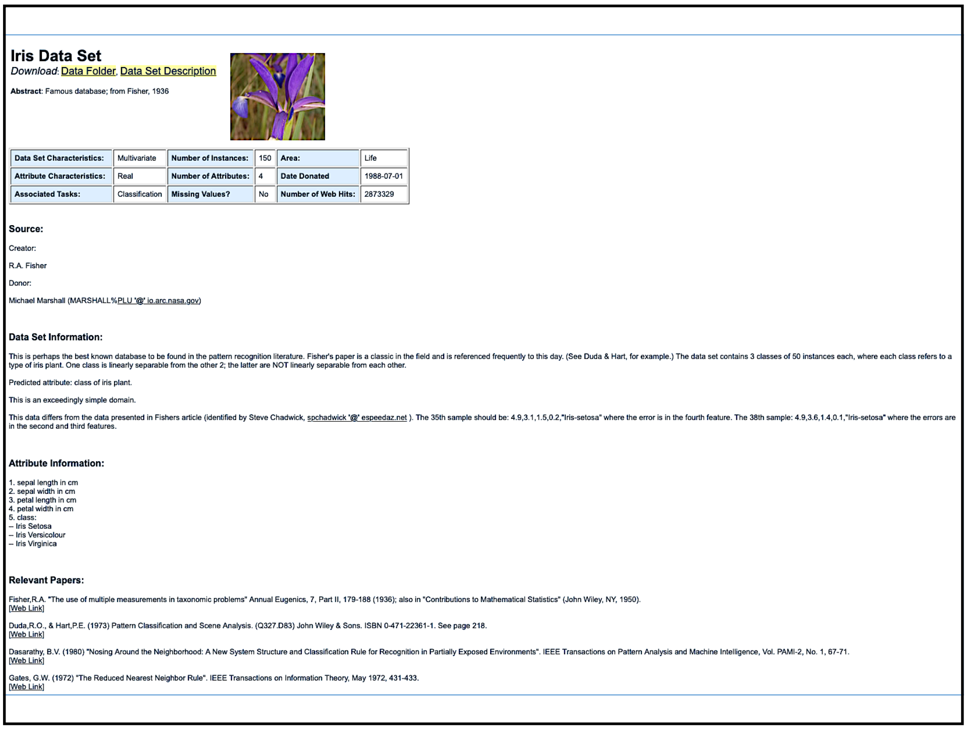
This dataset contains timeseries of mel-frequency cepstrum coefficients (MFCCs) corresponding to spoken Arabic digits. Includes data from 44 male and 44 female native Arabic speakers.

Newest Data Sets:

09-30-2019:	Hepatitis C Virus (HCV) for Egyptian patients
09-23-2019:	QSAR fish toxicity
09-23-2019:	QSAR aquatic toxicity
09-21-2019:	Online Retail II
09-20-2019:	Human Activity Recognition from Continuous Ambient Sensor Data
09-20-2019:	Beijing Multi-Site Air-Quality Data
09-20-2019:	MEx
07-30-2019:	PPG-DaLiA
07-24-2019:	Divorce Predictors data set
07-22-2019:	Alcohol QCM Sensor Dataset
07-14-2019:	Incident management process enriched event log

Most Popular Data Sets (hits since 2007):

2873325:	Iris
1597480:	Adult
1238855:	Wine
1045852:	Car Evaluation
1032149:	Wine Quality
1023256:	Heart Disease
1010309:	Breast Cancer Wisconsin (Diagnostic)
994180:	Bank Marketing
891347:	Human Activity Recognition Using Smartphones
830512:	Abalone
791324:	Forest Fires



Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	2873329

Source:

Creator:

R.A. Fisher

Donor:

Michael Marshall (MARSHALL%PLU '@' io.arc.nasa.gov)

Data Set Information:

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Predicted attribute: class of iris plant.

This is an exceedingly simple domain.

This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick '@' espeedaz.net). The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature. The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
– Iris Setosa
– Iris Versicolour
– Iris Virginica

Relevant Papers:

Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
[Web Link]

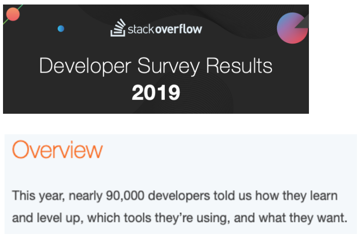
Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.
[Web Link]

Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Exposed Environments". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 1, 67-71.
[Web Link]

Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule". IEEE Transactions on Information Theory, May 1972, 431-433.
[Web Link]

Python

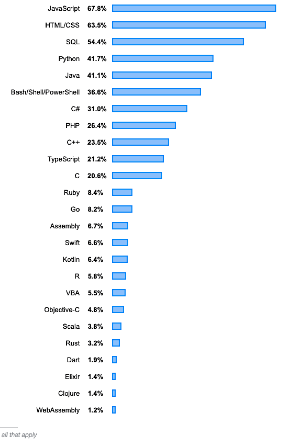
It is consistently appearing in the top 10 programming languages in surveys on StackOverflow.



Overview

This year, nearly 90,000 developers told us how they learn and level up, which tools they're using, and what they want.

URL: https://insights.stackoverflow.com/survey/2019



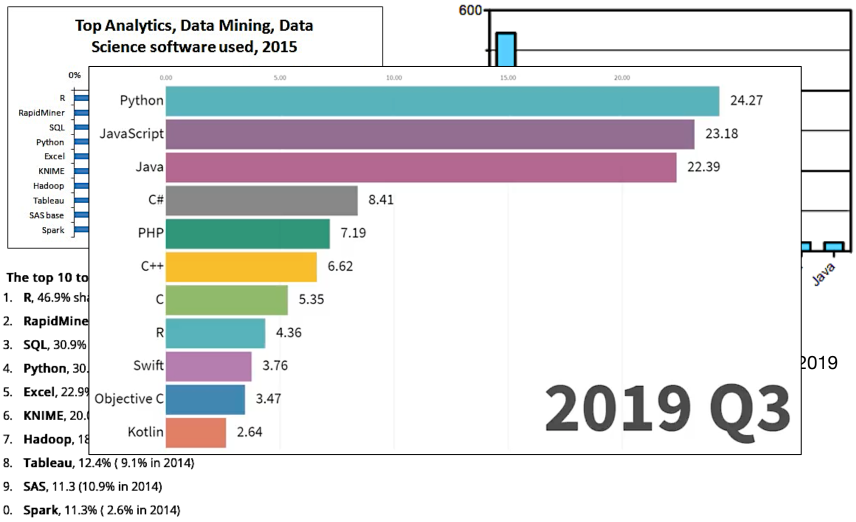
JavaScript	67.8%	
HTML/CSS	63.5%	
SQL	54.4%	
Python	41.7%	
Java	41.1%	
Bash/Shell/PowerShell	36.6%	
C#	31.0%	
PHP	26.4%	
C++	23.5%	
TypeScript	21.2%	
C	20.6%	
Ruby	8.4%	
Go	8.2%	
Assembly	6.7%	
Swift	6.6%	
Kotlin	6.4%	
R	5.8%	
VBA	5.5%	
Objective-C	4.8%	
Scala	3.8%	
Rust	3.2%	
Dart	1.9%	
Elixir	1.4%	
Clojure	1.4%	
WebAssembly	1.2%	

87,354 responses; select all that apply

Birkbeck, University of London 19 © Copyright 2019

KDD Nuggets tool survey in 2015

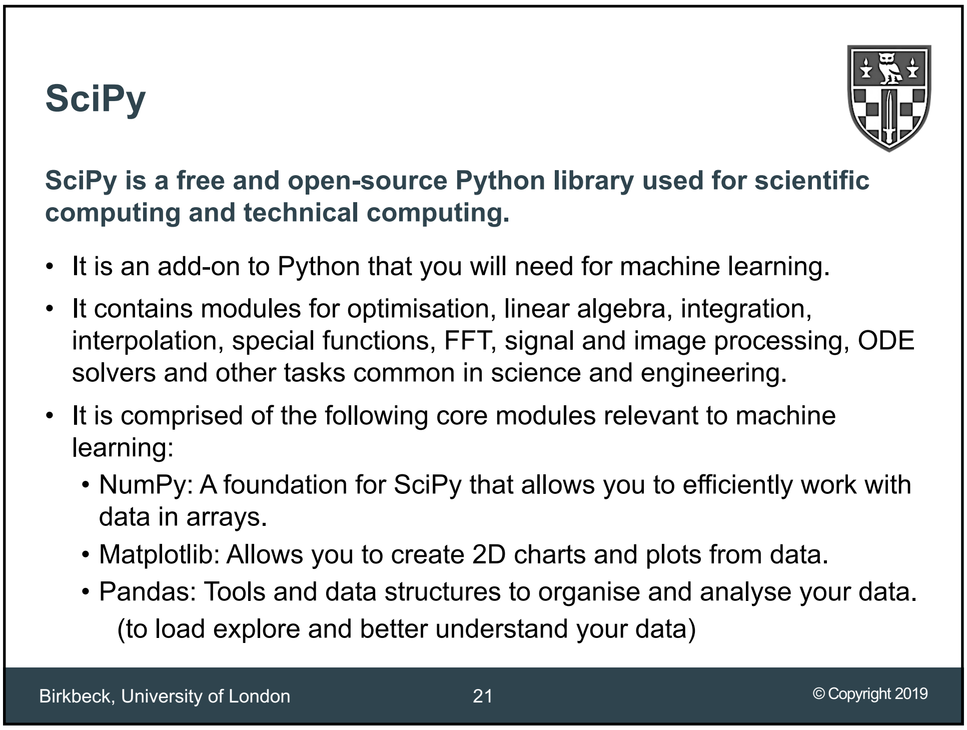
Kaggle platform survey in 2011



The top 10 to

1. **R**, 46.9% sh
2. **RapidMine**
3. **SQL**, 30.9%
4. **Python**, 30.
5. **Excel**, 22.9
6. **KNIME**, 20.
7. **Hadoop**, 18
8. **Tableau**, 12.4% (9.1% in 2014)
9. **SAS**, 11.3 (10.9% in 2014)

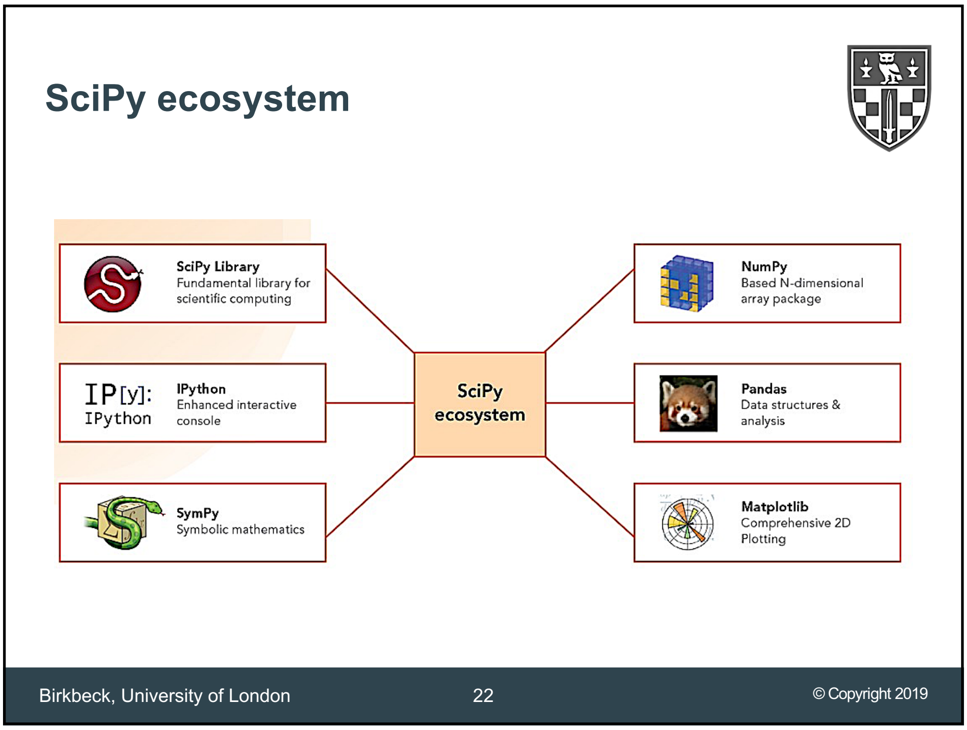
0. **Spark**, 11.3% (2.6% in 2014)



SciPy

SciPy is a free and open-source Python library used for scientific computing and technical computing.

- It is an add-on to Python that you will need for machine learning.
- It contains modules for optimisation, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.
- It is comprised of the following core modules relevant to machine learning:
  - NumPy: A foundation for SciPy that allows you to efficiently work with data in arrays.
  - Matplotlib: Allows you to create 2D charts and plots from data.
  - Pandas: Tools and data structures to organise and analyse your data.
    (to load explore and better understand your data)



SciPy ecosystem

SciPy ecosystem

SciPy Library
Fundamental library for scientific computing

IPython
Enhanced interactive console

SymPy
Symbolic mathematics

NumPy
Based N-dimensional array package

Pandas
Data structures & analysis

Matplotlib
Comprehensive 2D Plotting

scikit-learn

The scikit-learn library is how you can develop and practice ML in Python.

- scikit = SciPy + toolkit
- It is built upon and requires the SciPy.
- ML algorithms for classification, regression, clustering and etc.
- Tools for evaluating models, tuning parameters and pre-processing data.

Python Installation

Python 3.7.2

- Python Beginners Guide

  https://wiki.python.org/moin/BeginnersGuide/Download
- python --version
- pip - Python package management tool
- *pip install jupyter scipy numpy matplotlib pandas sklearn tensorflow theano keras seaborn subprocess.run graphviz pydot*
- *Anaconda 2019.03 for Windows Installer (Python 3.7 version)*

Questions?

paul@dcs.bbk.ac.uk