COMP47460

Introduction to Machine Learning (Blended Delivery)

Aonghus Lawlor
Derek Greene

School of Computer Science
Autumn 2018

Overview

- Module Outline
- Practical Details
- Machine Learning
  - Common Applications
  - Supervised v Unsupervised Learning
  - Representing Data as Features

Module Outline: Topics Covered

- Introduction and Fundamentals
- Supervised Learning
  - Classification: KNNs, Decision trees, Naive Bayes
  - Regression analysis
- Unsupervised Learning Algorithms
  - k-Means, Hierarchical clustering
- Working with Data
  - Dimensionality reduction, feature selection
- Evaluation and Methodologies
  - Statistical testing
  - Evaluating performance of ML systems
- Further Topics: Ensembles, Recommender systems

Practical Details

The course material will be made available on Moodle following the timetable of COMP47490: Tuesdays & Thursdays

There are 2 Tutorials/Practicals (one in Oct. and one in Nov), We have split them in two sessions each. You can attend either the first or second session of each tutorial/practical (they will be the same).

	Date	Location	SessionA	SessionB
Tutorial 1	Friday 19th Oct	B1.06 CompSci	13:00-14:20	14:30-16:00
Practical 1	Friday 26th Oct	H1.49 SCH	13:00-14:20	14:30-16:00
Tutorial 2	Friday 9th Nov	E2.16 SCE	13:00-14:20	14:30-16:00
Practical 2	Friday 16th Nov	H1.49 SCH	13:00-14:20	14:30-16:00

Notes, assignments, and additional material will be available on CS Moodle page for COMP47460 (https://csmoodle.ucd.ie)

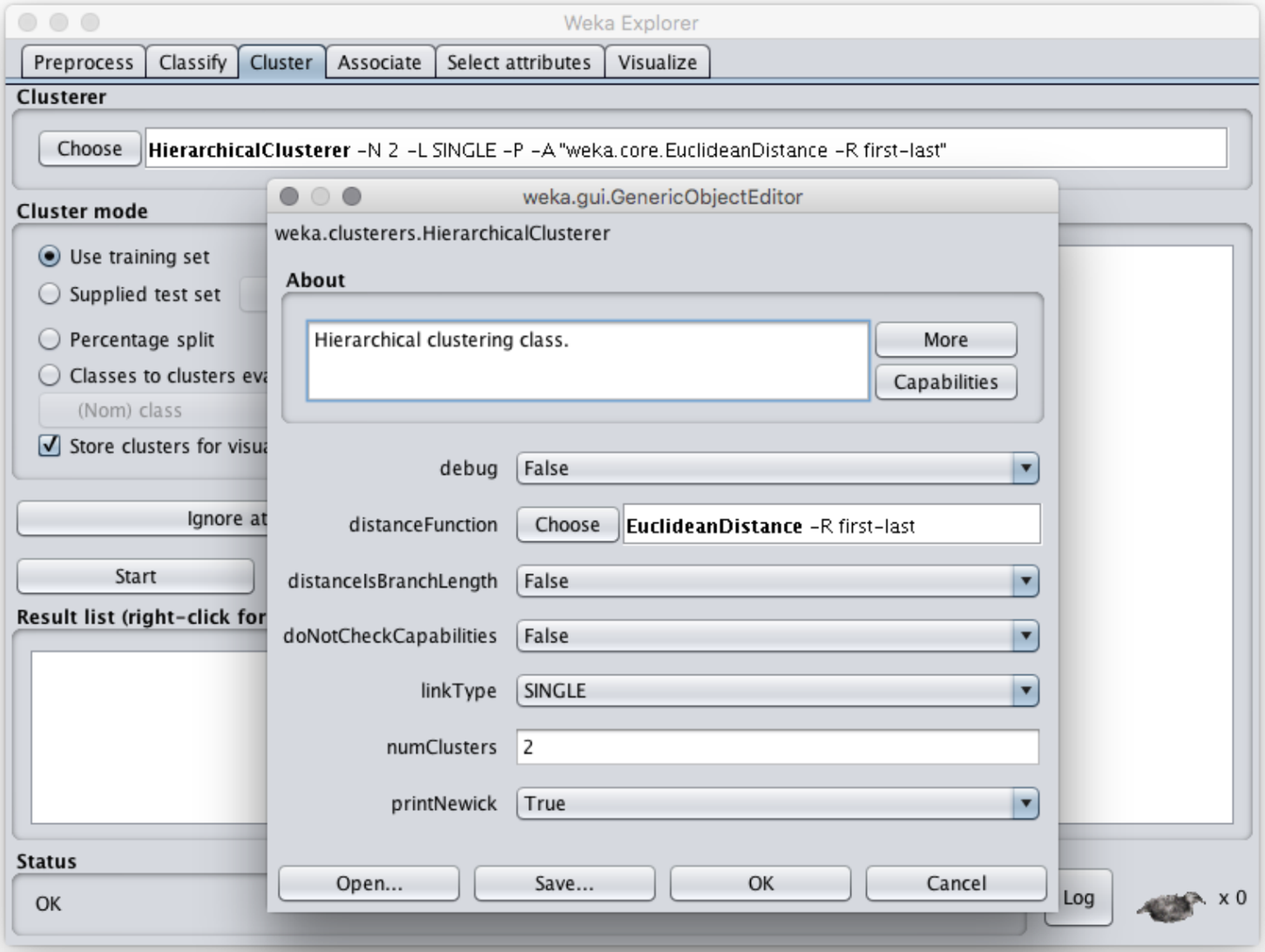
Moodle is currently open for registration via self-enrolment.
Password: **`mlo2018`**

Check that your surname, firstname and student ID number are all correct.

For all module queries contact aonghus.lawlor@ucd.ie

Practical Details

Tutorials require laptop with Java *Weka Toolkit* - Version 3.8 Stable



http://www.cs.waikato.ac.nz/ml/weka

Practical Details

Module marks are based on assignments + final theory exam:

20%	Assignment 1: Weka + Report
20%	Assignment 2: Weka + Report
60%	End of Semester Theory Exam

Note:

- ! All assignment deadlines are hard deadlines.
  1-5 days late: 10% deduction from overall mark
  6-10 days late: 20% deduction from overall mark
  Not accepted after 10 without extenuating circumstances
- ! Plagiarism will be treated seriously. Any evidence of plagiarism in an assignment or exam will result in a 0 mark.

Practical Details

CS grading scheme applies for this module. Pass mark is 40%.

Grade	Min	Max
A+	95	100
A	90	95
A-	85	90
B+	80	85
B	75	80
B-	70	75
C+	65	70
C	60	65
C-	55	60
D+	50	55
D	45	50
D-	40	45

Grade	Min	Max
E+	35	40
E	30	35
E-	25	30
F+	20	25
F	15	20
F-	10	15
G+	8	10
G	5	8
G-	2	5
NG	0	0

Plagiarism and UCD Computer Science

- Plagiarism is a serious academic offence
- [Student Code, section 6.2] or [UCD Registry Plagiarism Policy] or [CS Plagiarism policy and procedures]
- Our staff and demonstrators are proactive in looking for possible plagiarism in all submitted work
- Suspected plagiarism is reported to the CS Plagiarism subcommittee for investigation
- Usually includes an interview with student(s) involved
  - 1st offence: usually 0 or NG in the affected components
  - 2nd offence: referred to the University disciplinary committee
- Student who enables plagiarism is equally responsible

  http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf

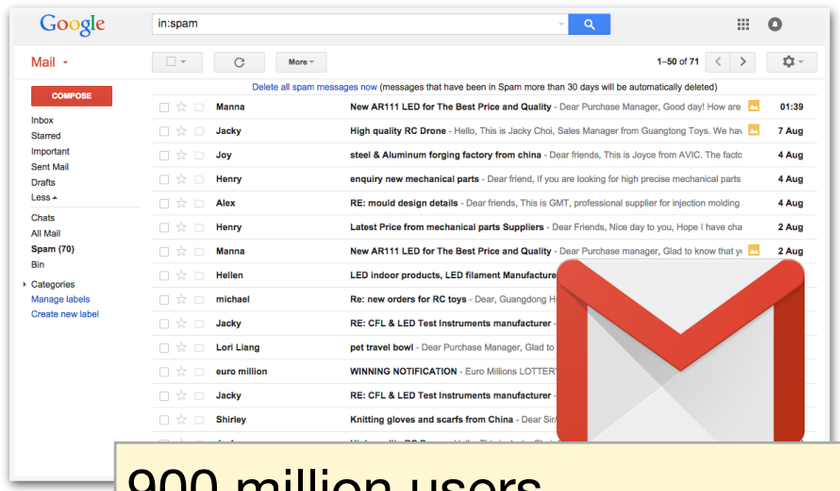
  http://www.ucd.ie/registry/academicsecretariat/docs/student_code.pdf

  http://libguides.ucd.ie/academicintegrity

Why Study Machine Learning?

- Explosion in rich, complex data to analyse - online and offline.
- Significant recent progress in algorithms and theory.
- Computational power is now available.
- Industry demand - Data scientists, Data engineers…
- New applications in many disciplines - Medicine, engineering, humanities…

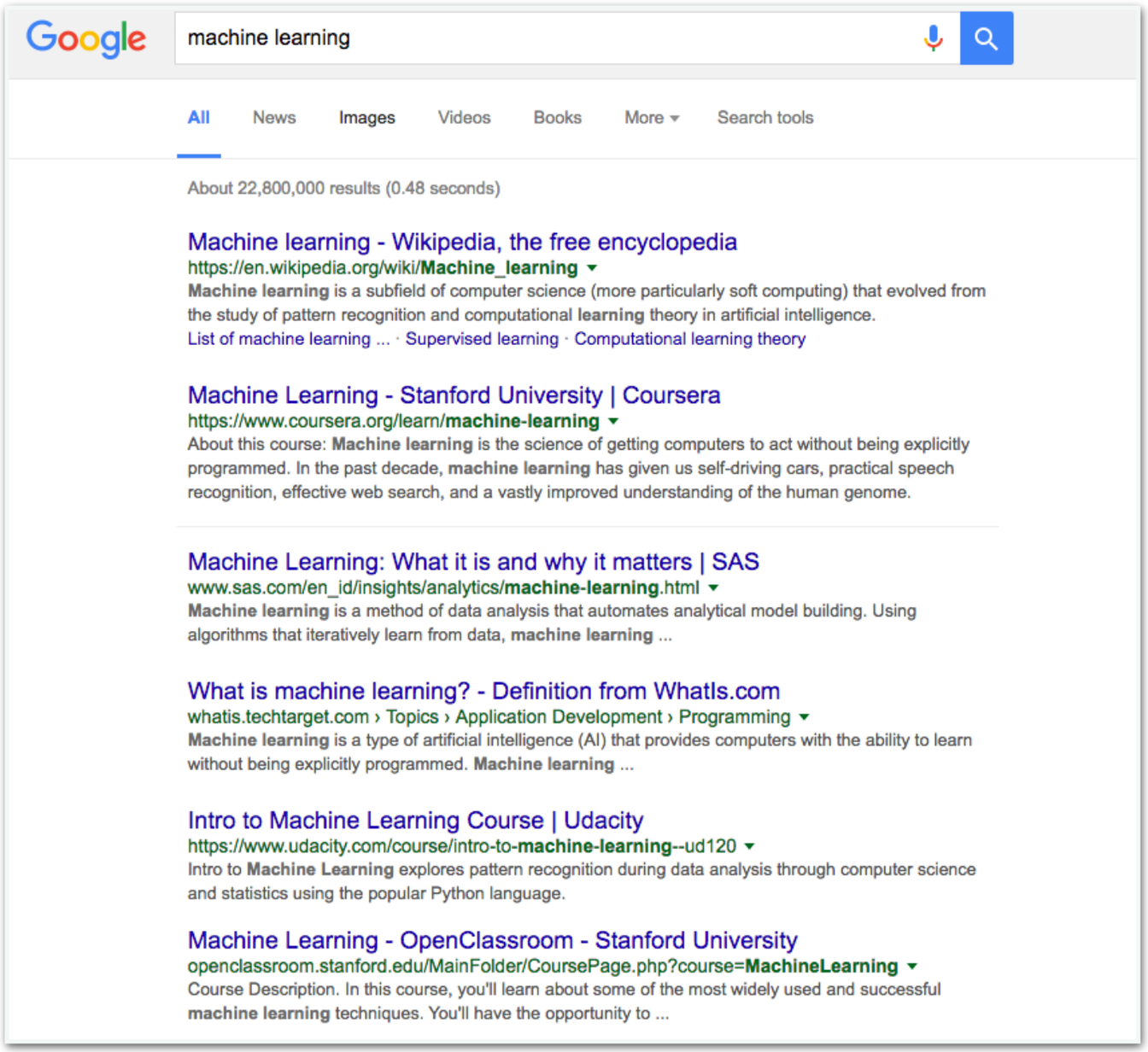
316 million monthly active users
500 million tweets per day
5 billion user sessions per day



900 million users
Handles 2+ trillion mails per year

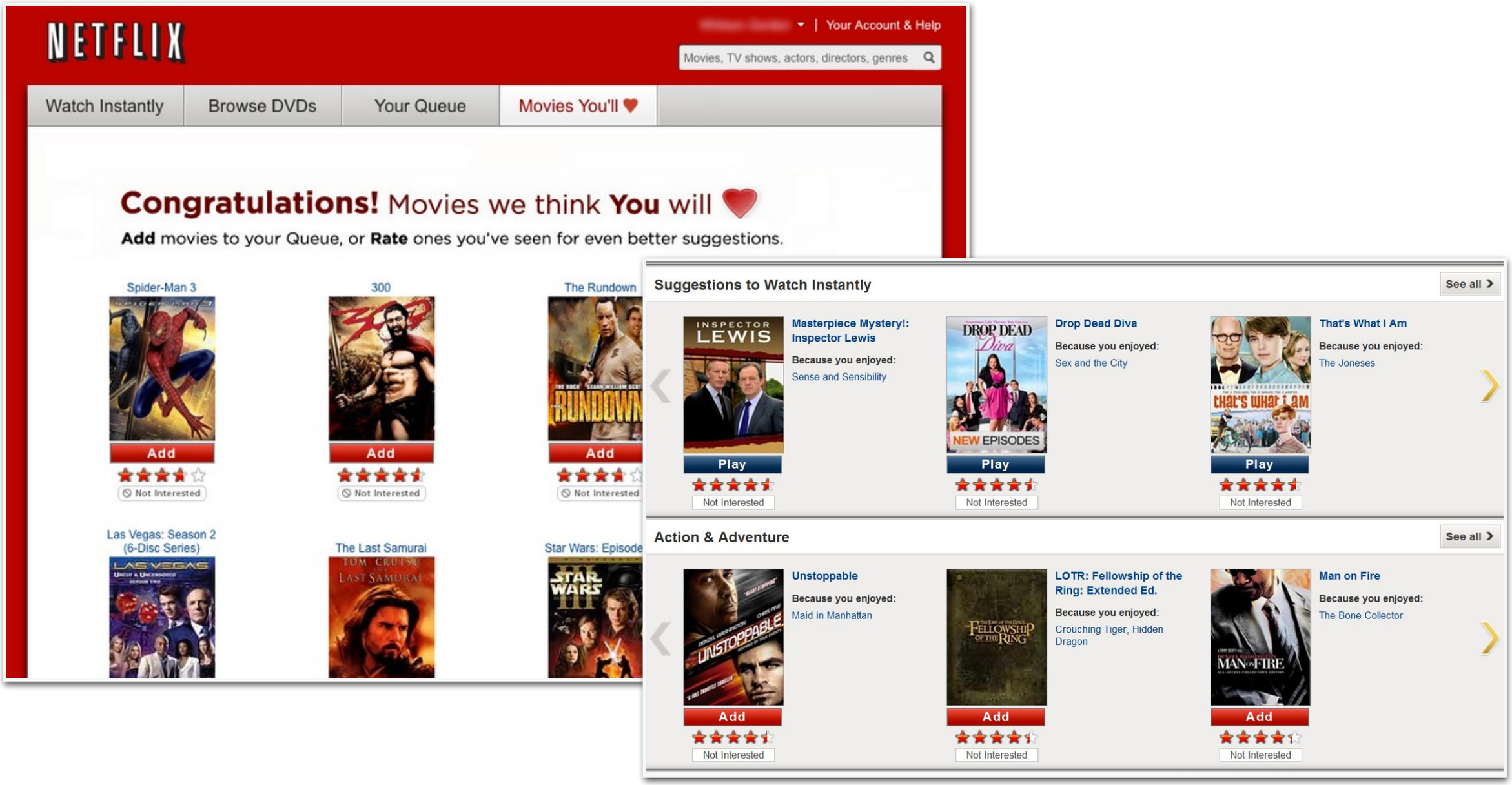
Application: Web Search

Submit a query to a search engine, it finds pages relevant to the query, and returns them ranked by relevance.



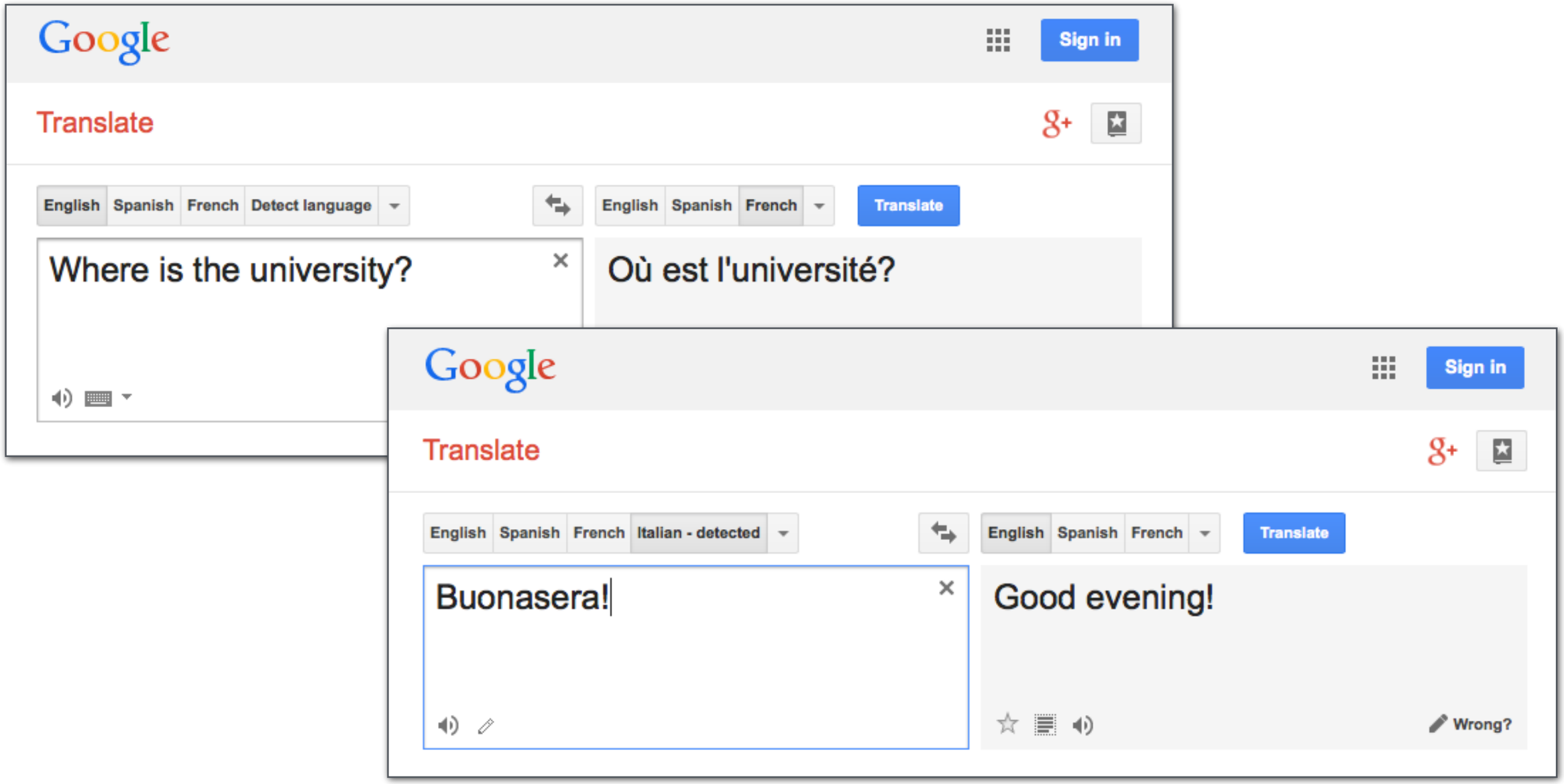
Application: Movie Recommendation

Netflix provides personalised recommendations for movies you might like, based on previous user ratings.



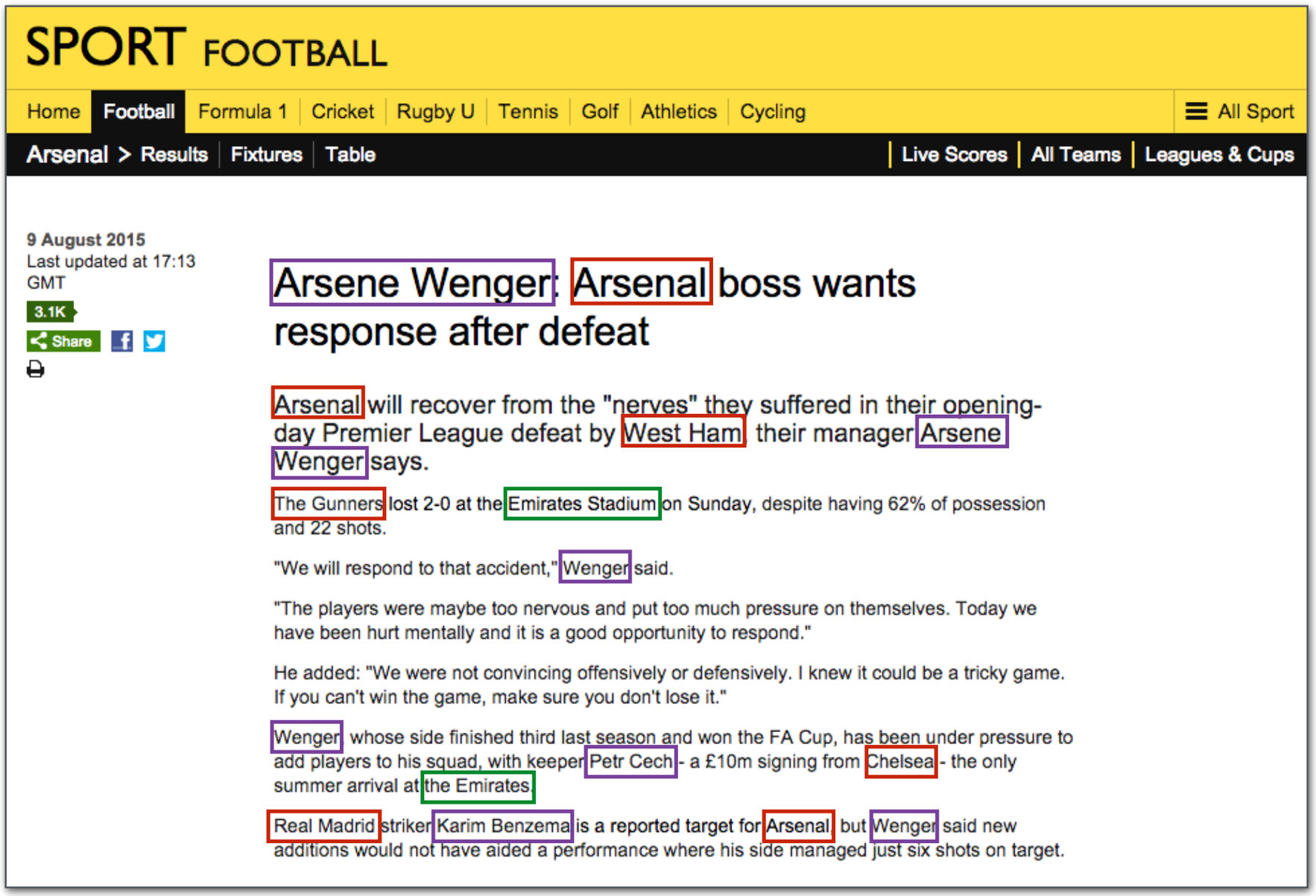
Application: Machine Translation

Use examples of translated documents to learn how to translate between the two languages.



Application: Entity Recognition

Automatically extracted named entities (e.g. people, places, organisations) from text documents (e.g. news articles).



Tag colours:

Person

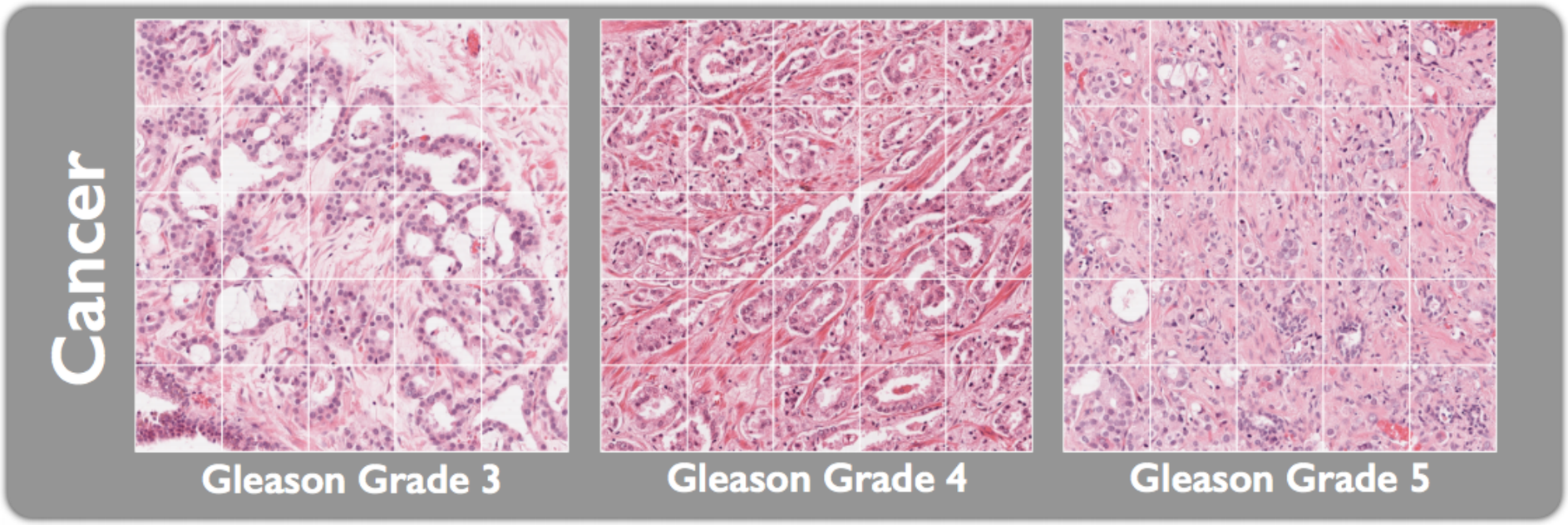
Location

Organisation

Applications in Medicine

Machine Learning provides tools and support solving diagnostic and prognostic tasks in a variety of medical domains, including…

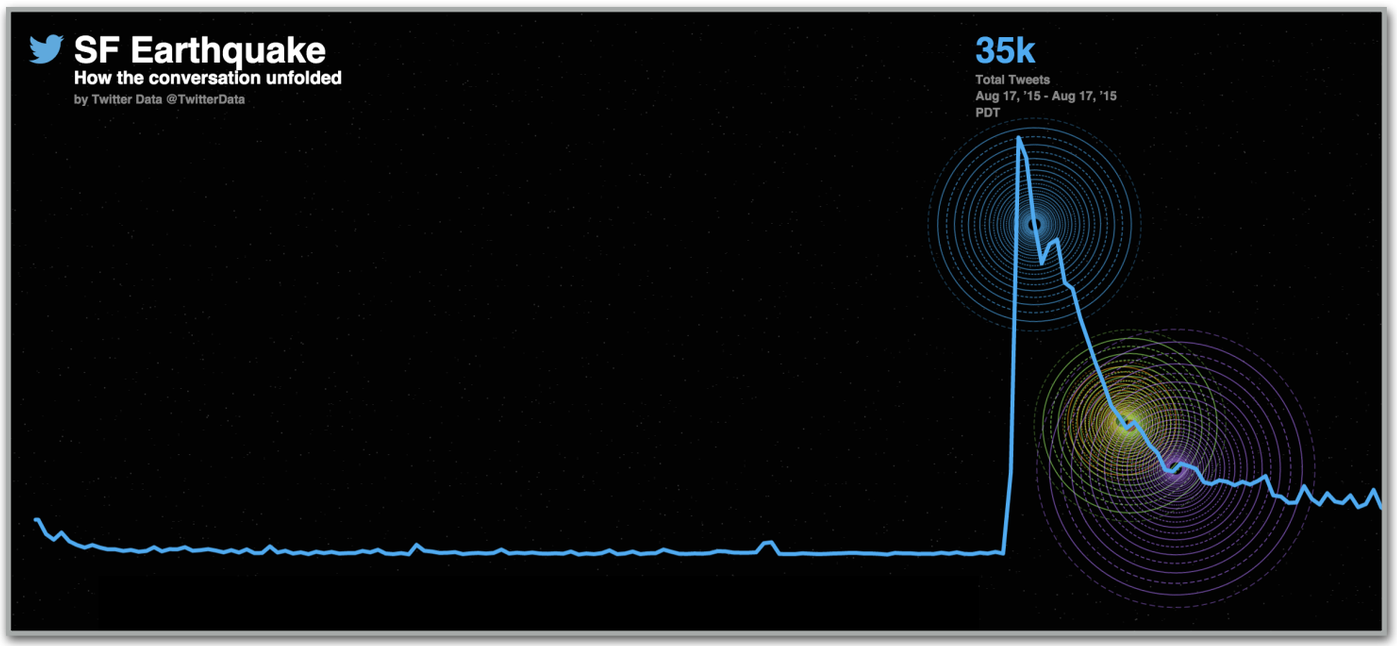
- Disease diagnosis based on previous correct cases
- Prediction of disease progression
- Medical image analysis and understanding
- Hospital information systems



Application: Anomaly Detection

Algorithms to find patterns in data that do not conform to a model of "normal" behaviour in a system. In some systems, these are rare events. In other systems, these are unexpected bursts of activity.

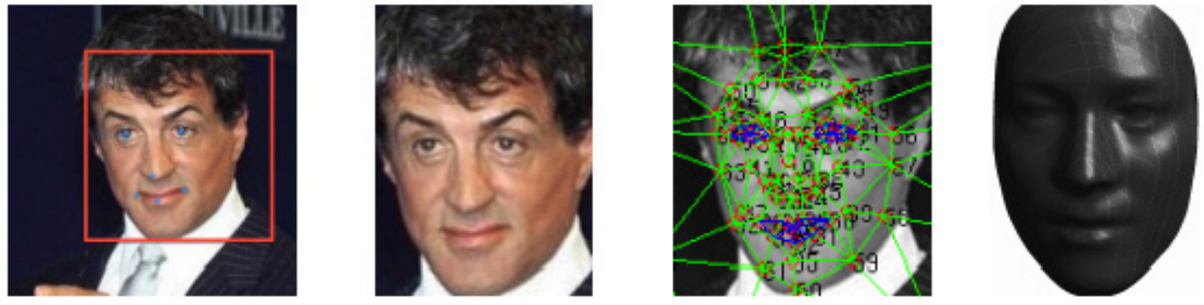
- *Cybersecurity*: Spike in number of false login attempts.
- *Payment systems*: Unusual number of failed or incomplete payments made.
- *Fraud detection*: Unexpected transactions in financial networks.
- *Event detection*: Sudden spike in volume of social media posts.



Application: Face Recognition

Facebook tags photos by comparing them to profile pictures.

"We currently use facial recognition software that uses an algorithm to calculate a unique template based on someone's facial features, like the distance between the eyes, nose and ears. This template is based on your profile pictures and photos you've been tagged in on Facebook."



In 2013, Facebook revealed its users have uploaded > 250 billion photos, and are uploading 350 million new photos each day.

Using other cues (e.g. hair style, clothing), allows Facebook to accurately identify people, even when their face is obscured.

http://www.fastcolabs.com/3028414/how-facebooks-machines-got-so-good-at-recognizing-your-face

http://arstechnica.com/?p=695873

Application: Autonomous Vehicles

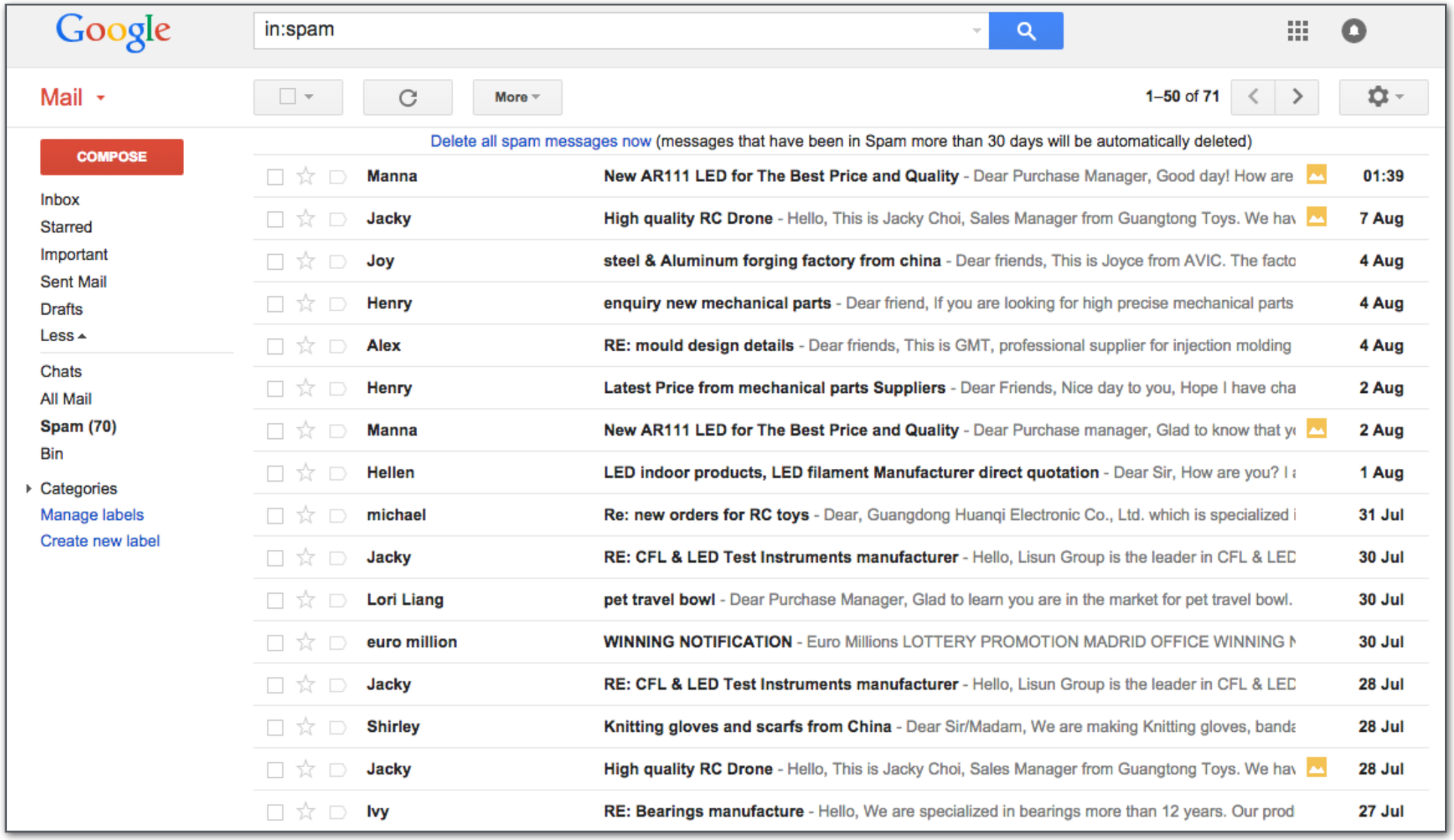
- Car manufacturers and researchers are exploring the potential of self-driving cars. Involves analysis of large volumes of sensor data, categorised using ML approaches combined with human labelling.
- 2004: Autonomous cars tried to navigate a 150 mile desert DARPA race. None of the 21 teams finished.
- 2015: Google driverless test vehicles have driven nearly 1 million miles, with no accidents caused by a self-driving car. Prototypes launched on public roads.

http://googleblog.blogspot.ie/2015/05/self-driving-vehicle-prototypes-on-road.html

http://www.nature.com/news/autonomous-vehicles-no-drivers-required-1.16832

Application: Spam Classification

Apply a learning algorithm to automatically classify incoming emails into *spam* or *non-spam*, based on previous examples of legitimate and spam email.



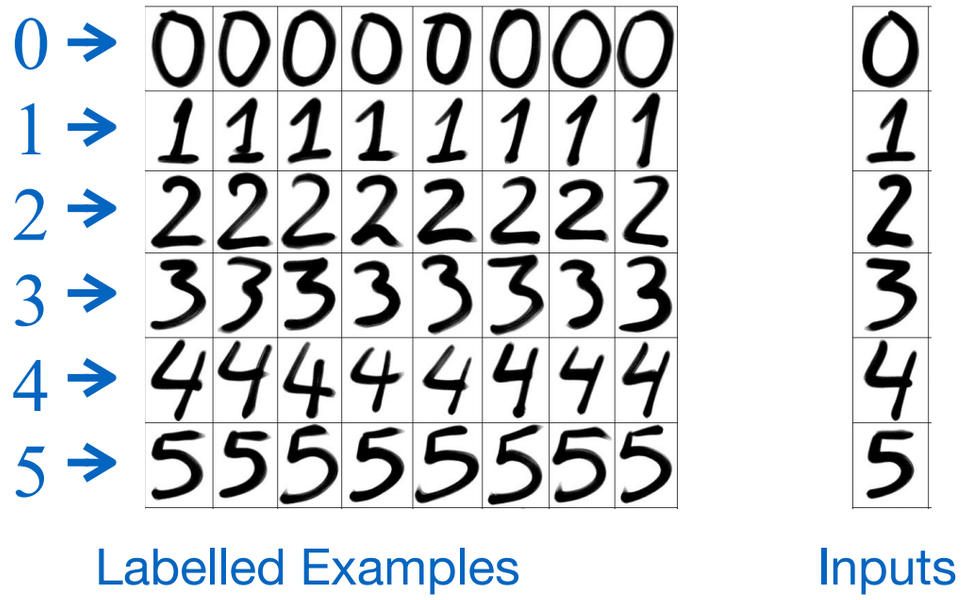
Supervised v Unsupervised Learning

- Supervised Learning:
  An algorithm that learns a function from examples of its inputs and outputs. It requires manually-labelled example data to learn the correct answer for a given query input.
  - e.g. Classification, Regression algorithms

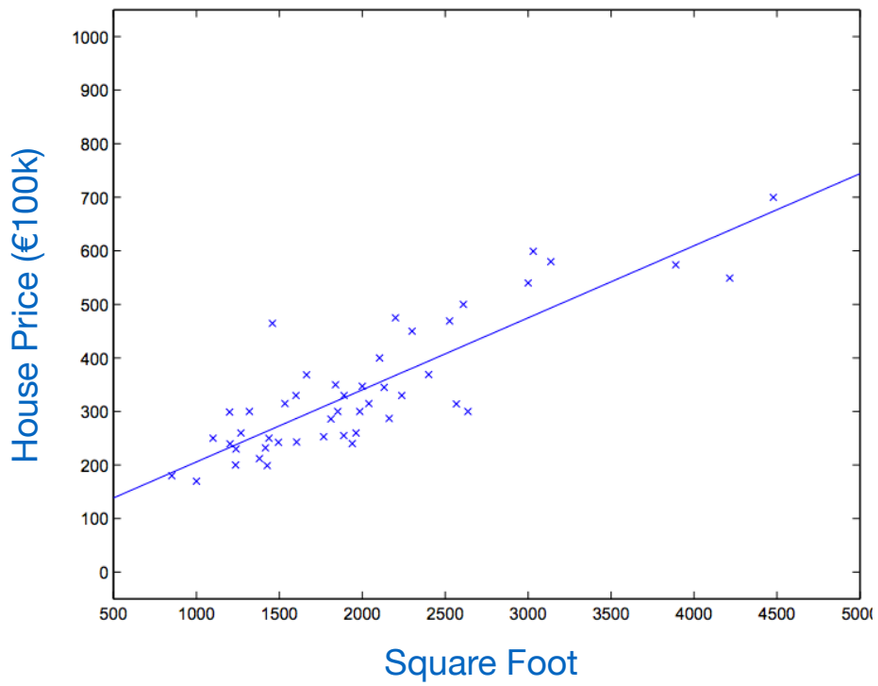
- Unsupervised Learning:
  An algorithm that finds patterns in data when no manually labelled examples are available as inputs. More focused on data exploration and knowledge discovery.
  - e.g. Clustering, Graph partitioning algorithms

Supervised Learning

- Classification:
  Examples represented by a set of features, which help decide the *class* to which a new query input belongs (i.e. output is a label)



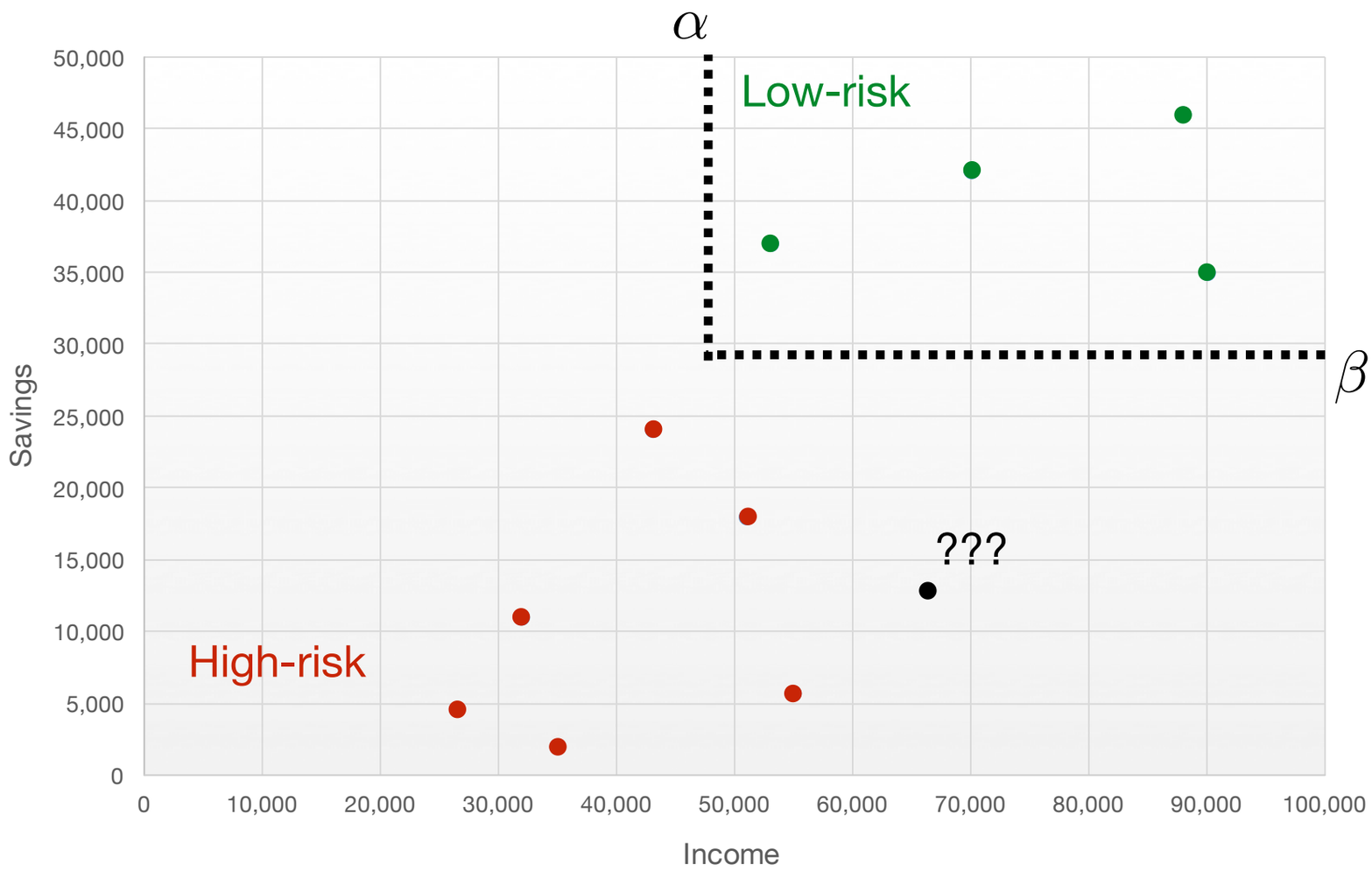
- Regression:
  Examples characterised by a set of features, which help decide the value of a continuous output variable (i.e. output is a number)



Typical Classification Task

Example:
Credit Scoring

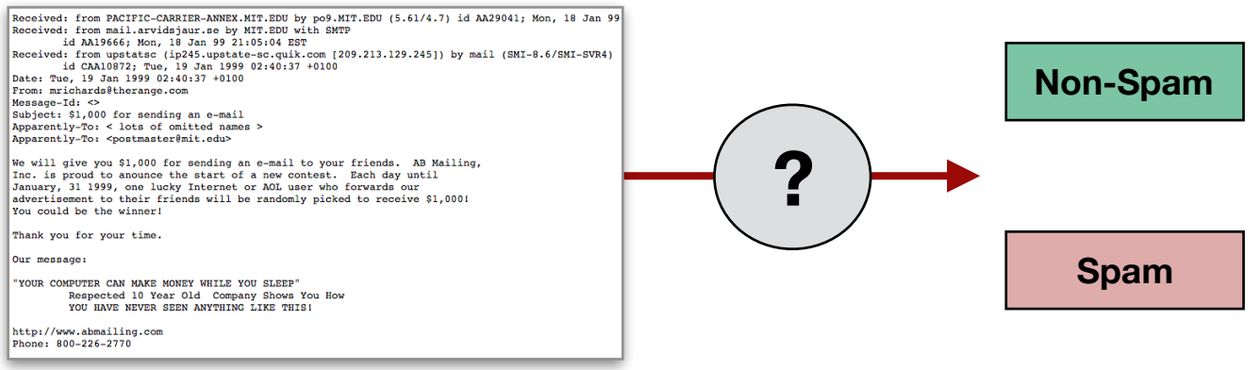
Manually classify customers into two categories (low-risk and high-risk) based on savings and income data.



Q. Can we train an algorithm to learn to automatically classify new customers as either low-risk or high-risk?
i.e. can we learn α and β?

Classification Tasks

- Binary Classification:
  Assign an input to one of two possible target class labels.



- Multiclass Classification:
  Assign an input to one of *M* different target class labels.

Classification Tasks

- **Evaluation:** Standard approach for classification tasks is to split the set of examples into a *training set* and the *test set.*
- Training set: Examples provided to the classifier to build a model of the data. Each example has been manually assigned a class label.
- Test set: Examples held back from the classifier, which are used to evaluate the accuracy of the classifier. Test examples are completely separate from the training set.
- Why not just train on all the data?
  - The test set is used to evaluate how well the model built by the classifier will generalise to new input examples.
  - Using the training data will give us over-optimistic results!

Classification Algorithms

- Many different learning algorithms exist for classification (e.g. k-nearest neighbour, decision tree, neural network, support vector machine).
- Problem dimensions will often determine which classification algorithm will be practically applicable, due to processing, memory, and storage constraints.

1. Number of input examples *N*.
   → Sometimes millions of input examples.
2. Number of feature (dimensions) *D* representing each input example.
   → Often 10-1000, but sometimes far higher.
3. Number of target classes *M*.
   → Often small (binary), but sometimes far higher.

Representing Data

- Examples are represented by one or more features, which can be distinguished by the type and number of values they can take.
  - Binary: Takes only two values - a boolean True/False decision
    e.g. married={True,False}, test_result={Pass,Fail}
  - Categorical (Nominal): A feature that takes values from two or more categories, with no intrinsic ordering to the categories.
    e.g. blood_group={A,B,AB,O}, nationality={French,Irish,Italian}
  - Ordinal: Similar to a categorical variable, but there is a clear ordering of the variables.
    e.g. grade={A,B,C,D,E,F}, dosage={Low,Medium,High}
  - Continuous: Numeric measurements, with or without a fixed range for the values.
    e.g. temperature, weight, height, latitude, longitude etc.

Typical Classification Task

- Training set with N=10 examples (customers). Each is described by D=5 features: 3 continuous, 2 categorical
- Each example has one of two class labels = {High-risk,Low-risk}

Example	*Income*	*Savings*	*Married*	*Gender*	*Age*	*Class*
1	35,000	2,000	Y	M	32	High-risk
2	51,000	18,000	N	M	34	High-risk
3	70,000	42,000	Y	F	41	Low-risk
4	26,500	4,500	N	M	22	High-risk
5	32,000	11,000	N	F	25	High-risk
6	53,000	37,000	N	F	39	Low-risk
7	88,000	46,000	Y	M	48	Low-risk
8	55,000	5,700	N	M	55	High-risk
9	90,000	35,000	Y	F	61	Low-risk
10	43,000	24,000	Y	M	33	High-risk

Q. To which class does this new customer belong?

Example	*Income*	*Savings*	*Married*	*Gender*	*Age*	*Class*
X	66,000	13,000	Y	M	44	???

Typical Classification Task

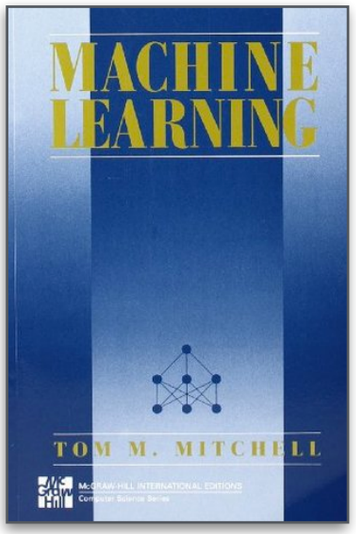
- Training set with N=10 examples (fruit). Each is described by D=4 features: 3 continuous, 1 categorical
- Each has one of three class labels = {Apple,Pear,Orange}

Example	*Height*	*Width*	*Taste*	*Weight*	*Class*
1	60	62	Sweet	186	Apple
2	70	53	Sweet	180	Pear
3	55	50	Tart	152	Apple
4	76	40	Sweet	152	Pear
5	68	71	Tart	207	Orange
6	65	68	Sour	221	Apple
7	63	45	Sweet	140	Pear
8	55	56	Sweet	154	Apple
9	76	78	Tart	211	Orange
10	60	58	Sour	175	Apple

Q. To which class does this new fruit belong?

Example	*Height*	*Width*	*Taste*	*Weight*	*Class*
X	63	68	Sweet	168	???

Additional Reading

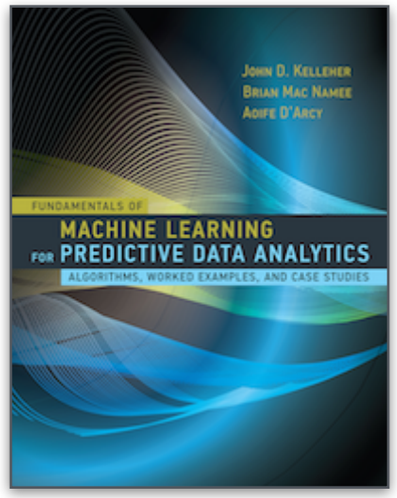


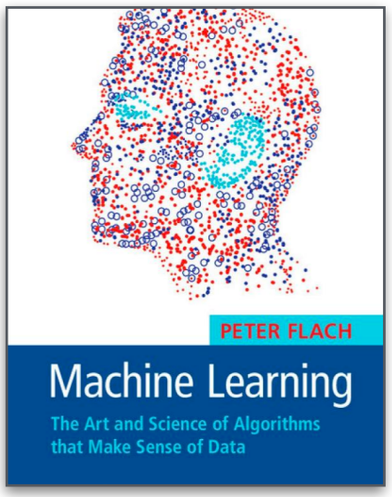
Machine Learning
McGraw-Hill 1997

Tom M. Mitchell

Fundamentals of Machine Learning for Predictive Data Analytics

John D. Kelleher, Brian Mac Namee, Aoife D'Arcy





Machine Learning: The Art and Science of Algorithms that Make Sense of Data

Peter Flach

Data Mining: Practical Machine Learning Tools and Techniques, 3rd Ed

Ian H. Witten, Eibe Frank, Mark A. Hall