

COMP30120

Introduction to Machine Learning

Derek Greene

School of Computer Science and Informatics
Autumn 2015



Overview

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

Module Outline: Topics Covered

- Introduction and Basics
- Classification Algorithms
 - Nearest neighbour, Decision trees, Naive Bayes
- 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: Ensemble learning, Social network analysis, Recommender systems

Practical Details

Lectures/Tutorials: Tuesdays & Thursdays 9-10am B004 CSI

Notes, assignments, and additional material will be available on CS Moodle page for COMP30120 (<https://csimoodle.ucd.ie>)

Moodle is currently open for registration via self-enrolment.

Password: **ml2015**

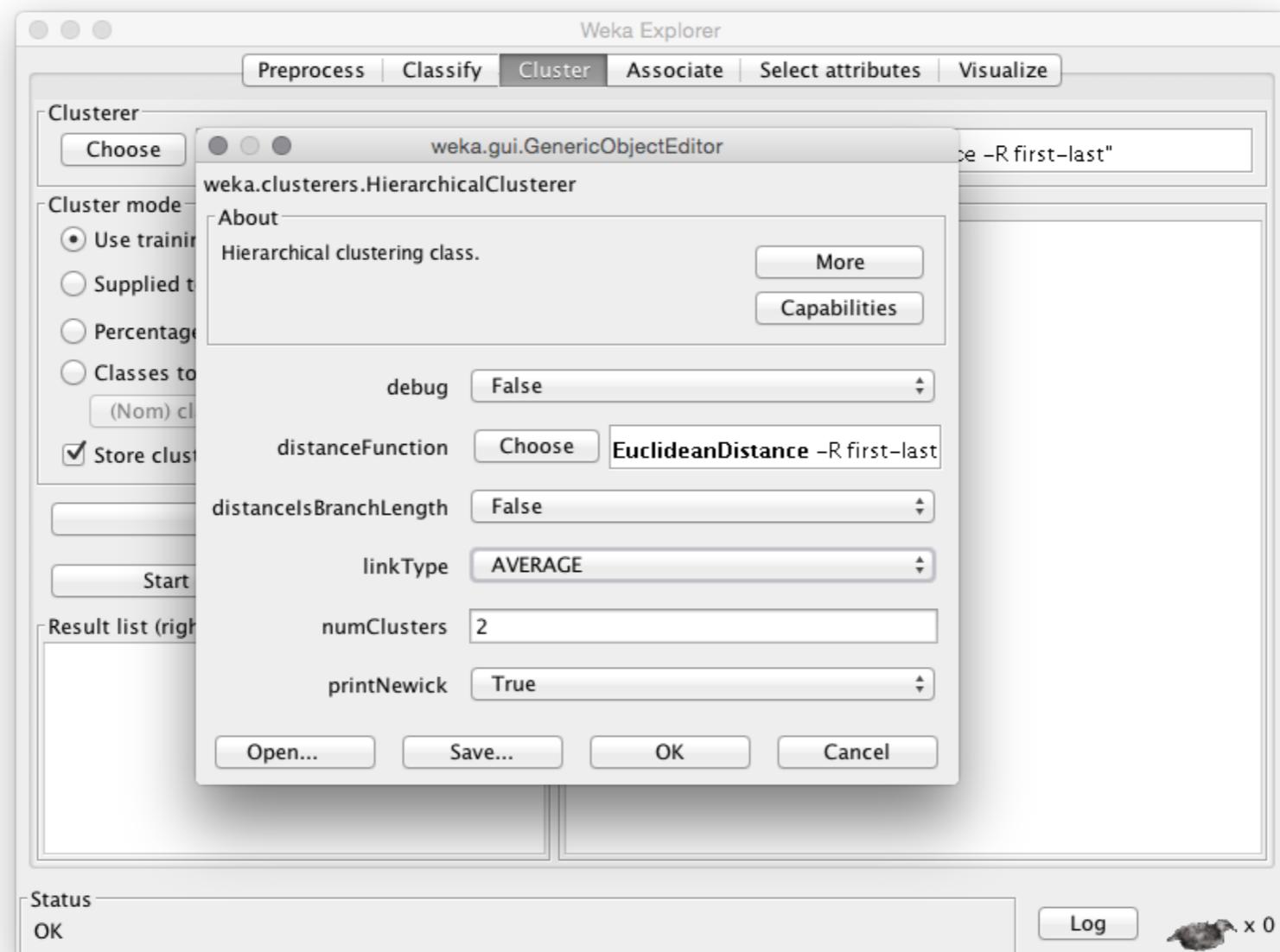
Check that your surname,firstname and student ID number are all correct.

The screenshot shows the Moodle course page for 'COMP30120 Machine Learning'. At the top, there are three tabs: 'UCD', 'CSI', and 'My Courses'. Below the tabs, the course title 'COMP30120 Machine Learning' is displayed in bold. Underneath the title, the breadcrumb navigation shows 'Home > LEVEL 3 > COMP30120'. A 'News forum' link is visible. The 'Overview' section contains a brief description of the module: 'This module is a first semester option for final year and postgraduate students. There are two lectures, one at 9am on Tuesday and another at 9am on Thursday. Assessment will be by three assignments and two in-class exams.' The 'Introduction' section provides practical module details and an overview of Machine Learning, with a link to 'COMP 30120 Introduction'.

For all module queries contact derek.greene@ucd.ie

Practical Details

Tutorials require laptop with Java Weka Toolkit (*Stable 3.6.12*)



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

Practical Details

Module marks based on assignment + in-class exams:

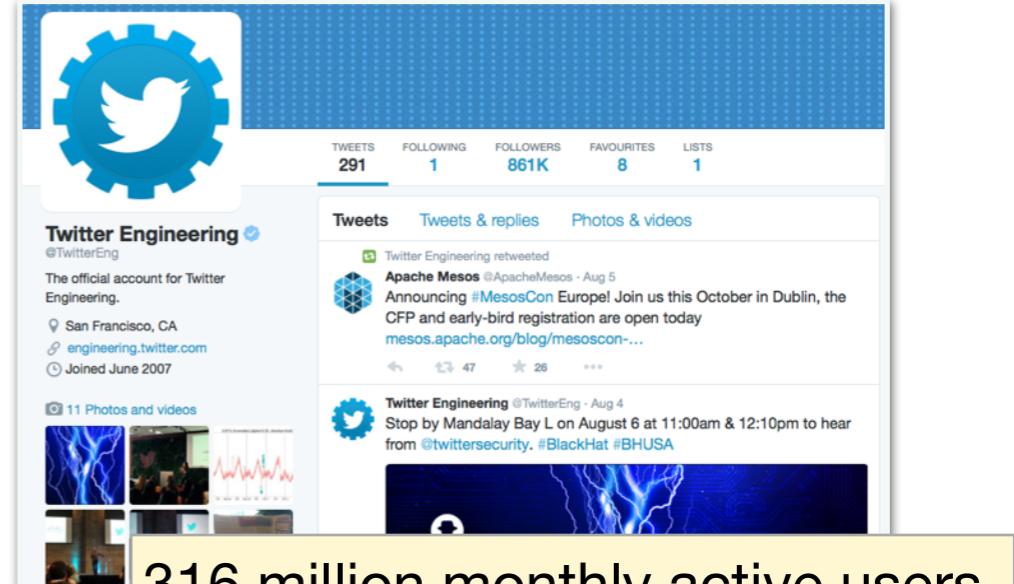
| | |
|-----|---|
| 15% | Assignment 1: Weka + Report |
| 15% | Assignment 2: Weka + Report |
| 15% | Assignment 3: Weka + Report |
| 20% | Week 9 In-class practical exam, with Weka + laptops. |
| 35% | Week 12 In-class theory exam, no laptops etc. |

Please note:

- ! All assignment deadlines are hard deadlines.
- ! Plagiarism will be treated seriously. Any evidence of plagiarism in an assignment or exam will result in a 0 mark.

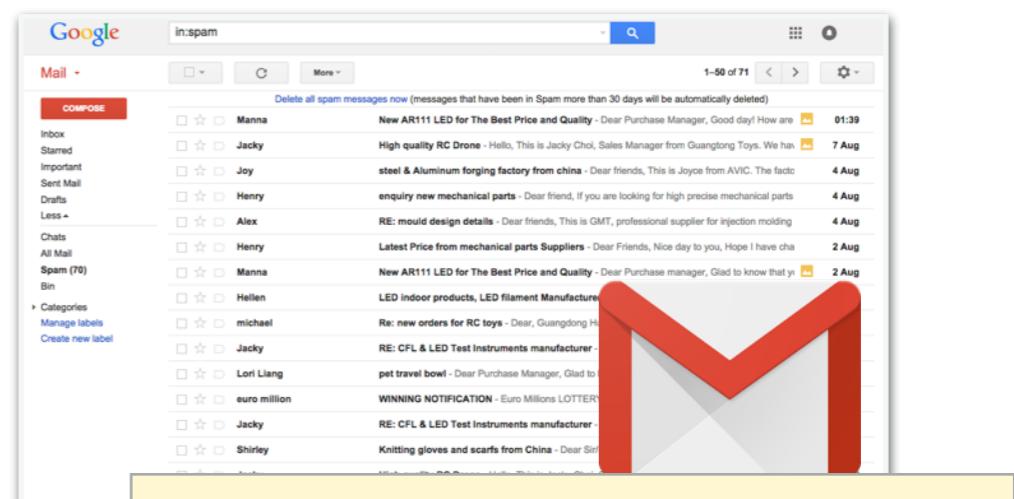
Why Study Machine Learning?

- 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...



A screenshot of the Twitter Engineering account (@TwitterEng). The profile picture is a blue gear containing a white Twitter bird. The bio reads: "The official account for Twitter Engineering." It shows 291 tweets, 1 following, 861K followers, 8 favourites, and 1 list. The timeline displays two tweets from the account, one retweeting Apache Mesos and another announcing a conference. A yellow callout box contains the following statistics:

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

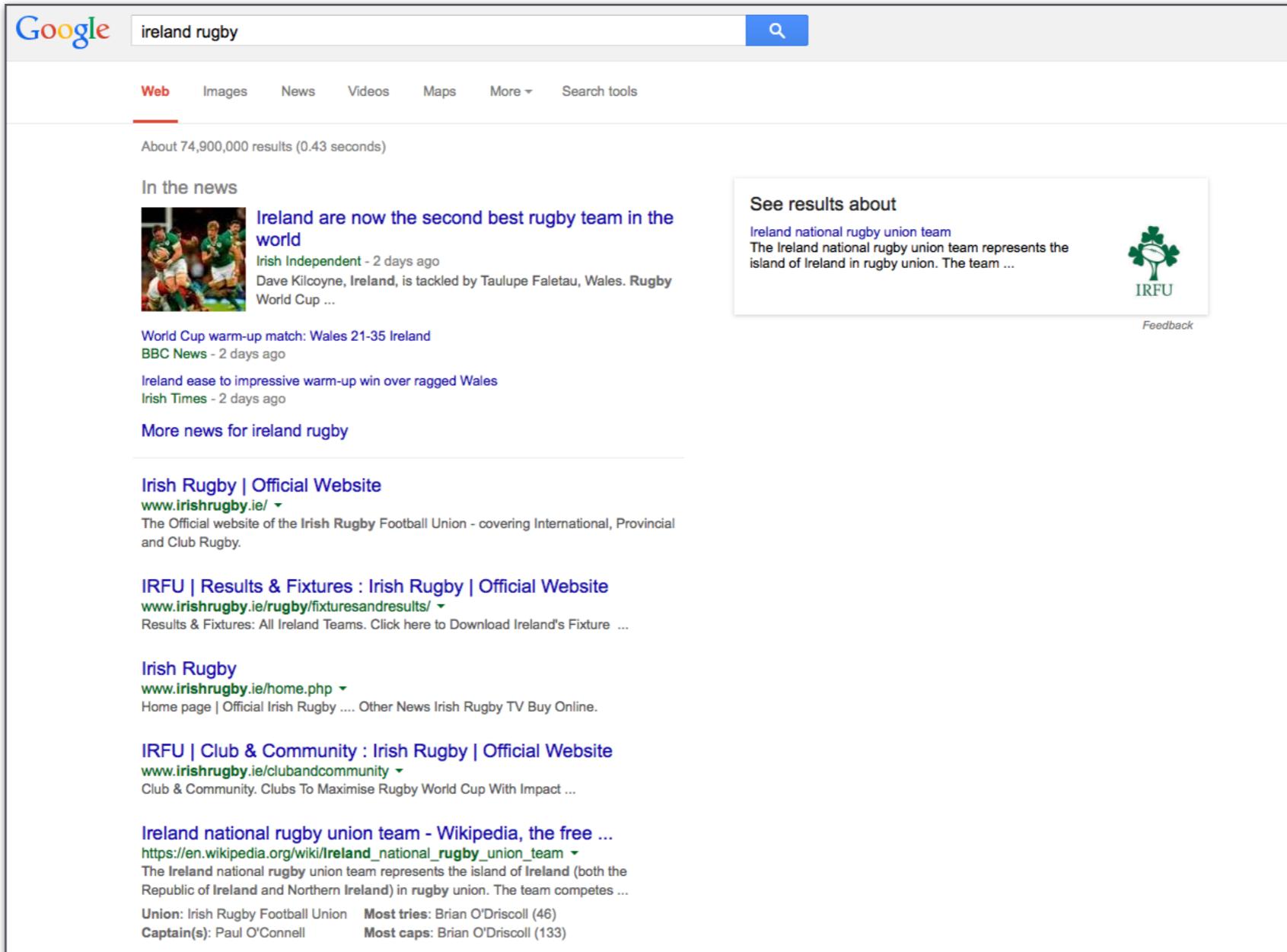


A screenshot of a Gmail inbox search results for "in:spam". The results show 71 spam messages from various senders like Manna, Jacky, Joy, Henry, Alex, etc., with subject lines related to LED lights, RC drones, and other spammy offers. A large red Gmail logo is overlaid on the right side. A yellow callout box contains the following statistics:

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.



A screenshot of a Google search results page for the query "ireland rugby". The results are displayed under the "Web" tab. The top result is a news article from the Irish Independent about Ireland's second-place finish in the World Cup. Below it are links to BBC News, Irish Times, and the official Irish Rugby website. Further down are links to the IRFU website, the Irish Rugby homepage, and a Wikipedia entry for the Ireland national rugby union team. A sidebar on the right shows a summary of the Ireland national rugby union team, including its logo (a shamrock inside a green oval) and a link to "See results about".

Google ireland rugby

Web Images News Videos Maps More Search tools

About 74,900,000 results (0.43 seconds)

In the news

Ireland are now the second best rugby team in the world
Irish Independent - 2 days ago
Dave Kilcoyne, Ireland, is tackled by Taulupe Faletau, Wales. Rugby World Cup ...

World Cup warm-up match: Wales 21-35 Ireland
BBC News - 2 days ago

Ireland ease to impressive warm-up win over ragged Wales
Irish Times - 2 days ago

More news for ireland rugby

Irish Rugby | Official Website
www.irishrugby.ie/ ▾
The Official website of the Irish Rugby Football Union - covering International, Provincial and Club Rugby.

IRFU | Results & Fixtures : Irish Rugby | Official Website
www.irishrugby.ie/rugby/fixturesandresults/ ▾
Results & Fixtures: All Ireland Teams. Click here to Download Ireland's Fixture ...

Irish Rugby
www.irishrugby.ie/home.php ▾
Home page | Official Irish Rugby Other News Irish Rugby TV Buy Online.

IRFU | Club & Community : Irish Rugby | Official Website
www.irishrugby.ie/clubandcommunity ▾
Club & Community. Clubs To Maximise Rugby World Cup With Impact ...

Ireland national rugby union team - Wikipedia, the free ...
https://en.wikipedia.org/wiki/Ireland_national_rugby_union_team ▾
The Ireland national rugby union team represents the island of Ireland (both the Republic of Ireland and Northern Ireland) in rugby union. The team competes ...

Union: Irish Rugby Football Union Most tries: Brian O'Driscoll (46)
Captain(s): Paul O'Connell Most caps: Brian O'Driscoll (133)

See results about

Ireland national rugby union team
The Ireland national rugby union team represents the island of Ireland in rugby union. The team ...

IRFU

Feedback

Application: Movie Recommendation

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

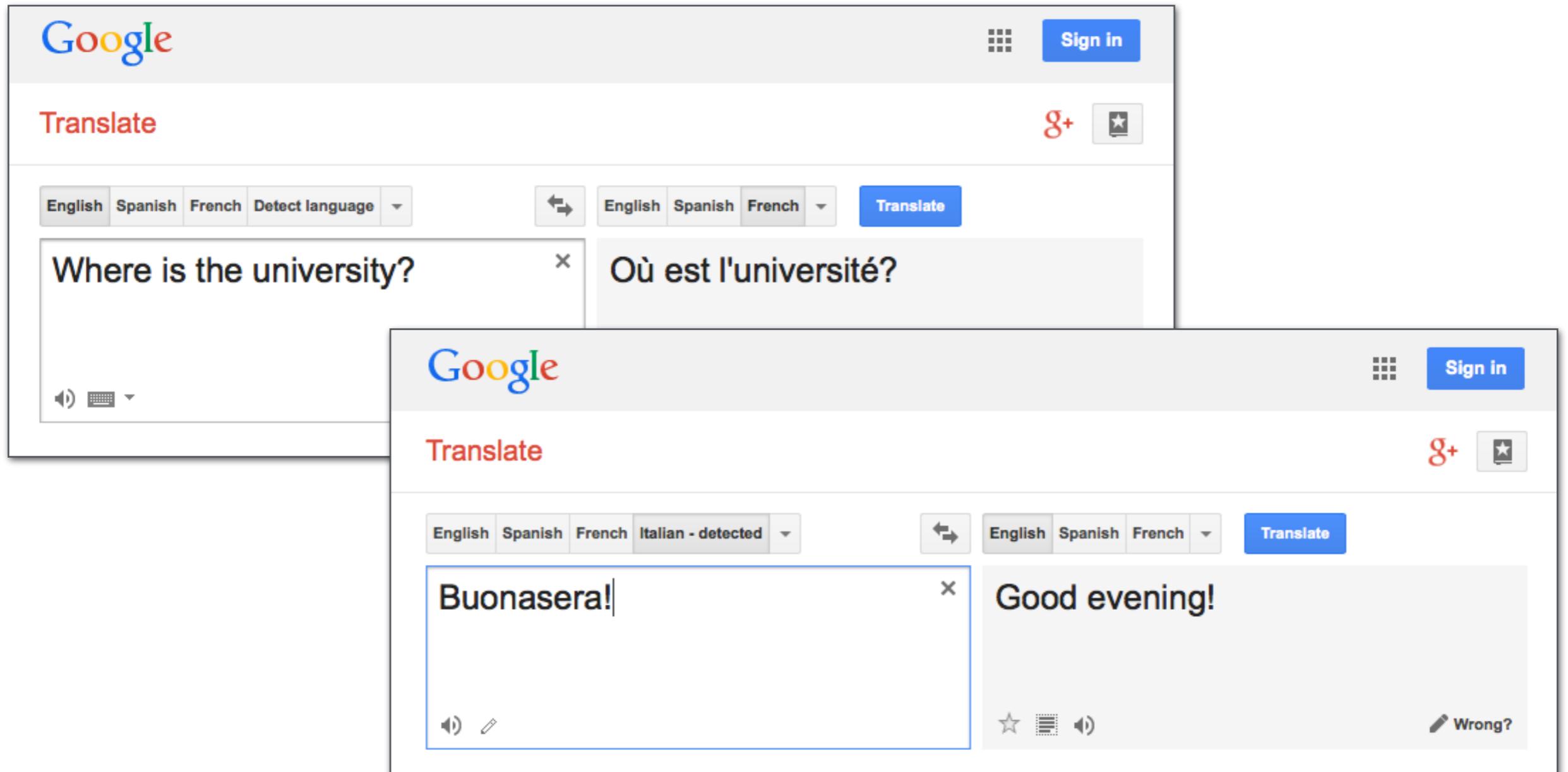
The screenshot shows the Netflix homepage with a red header. The top navigation bar includes links for 'Watch Instantly', 'Browse DVDs', 'Your Queue', and 'Movies You'll ❤️'. A search bar is located at the top right. Below the header, a large banner displays the message 'Congratulations! Movies we think You will ❤️' and encourages users to 'Add movies to your Queue, or Rate ones you've seen for even better suggestions.'

The main content area is divided into several sections:

- Suggestions to Watch Instantly:** This section lists recommended movies and TV shows. Examples include 'Spider-Man 3', '300', 'The Rundown', 'Inspector Lewis', 'Masterpiece Mystery!: Inspector Lewis', 'Drop Dead Diva', 'That's What I Am', 'Las Vegas: Season 2 (6-Disc Series)', 'The Last Samurai', 'Star Wars: Episode III - Revenge of the Sith', 'Unstoppable', 'LOTR: Fellowship of the Ring: Extended Ed.', and 'Man on Fire'. Each item has a thumbnail, a title, a brief description, a 'Play' button, a rating scale from 1 to 5 stars, and a 'Not Interested' button.
- Action & Adventure:** This section shows more movie recommendations in the same format, including 'Unstoppable', 'LOTR: Fellowship of the Ring: Extended Ed.', and 'Man on Fire'.

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).

SPORT FOOTBALL

Home **Football** Formula 1 | Cricket | Rugby U | Tennis | Golf | Athletics | Cycling | All Sport

Arsenal > Results | Fixtures | Table | Live Scores | All Teams | Leagues & Cups

9 August 2015
Last updated at 17:13
GMT
3.1K Share f t

Arsene Wenger: Arsenal boss wants response after defeat

Arsenal will recover from the "nerves" they suffered in their opening-day Premier League defeat by West Ham, their manager Arsene Wenger says.

The Gunners lost 2-0 at the Emirates Stadium on Sunday, despite having 62% of possession and 22 shots.

"We will respond to that accident," Wenger said.

"The players were maybe too nervous and put too much pressure on themselves. Today we have been hurt mentally and it is a good opportunity to respond."

He added: "We were not convincing offensively or defensively. I knew it could be a tricky game. If you can't win the game, make sure you don't lose it."

Wenger, whose side finished third last season and won the FA Cup, has been under pressure to add players to his squad, with keeper Petr Cech - a £10m signing from Chelsea - the only summer arrival at the Emirates.

Real Madrid striker Karim Benzema is a reported target for Arsenal, but Wenger said new additions would not have aided a performance where his side managed just six shots on target.

Tag colours:

Person

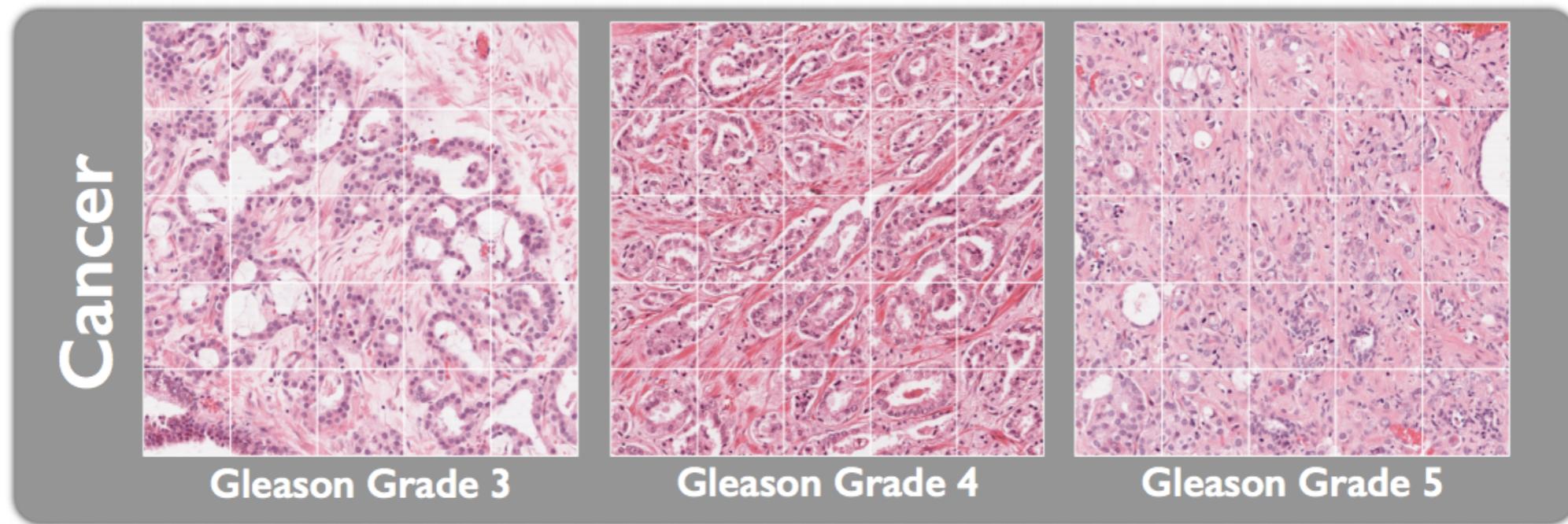
Location

Organisation

Applications in Medicine

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

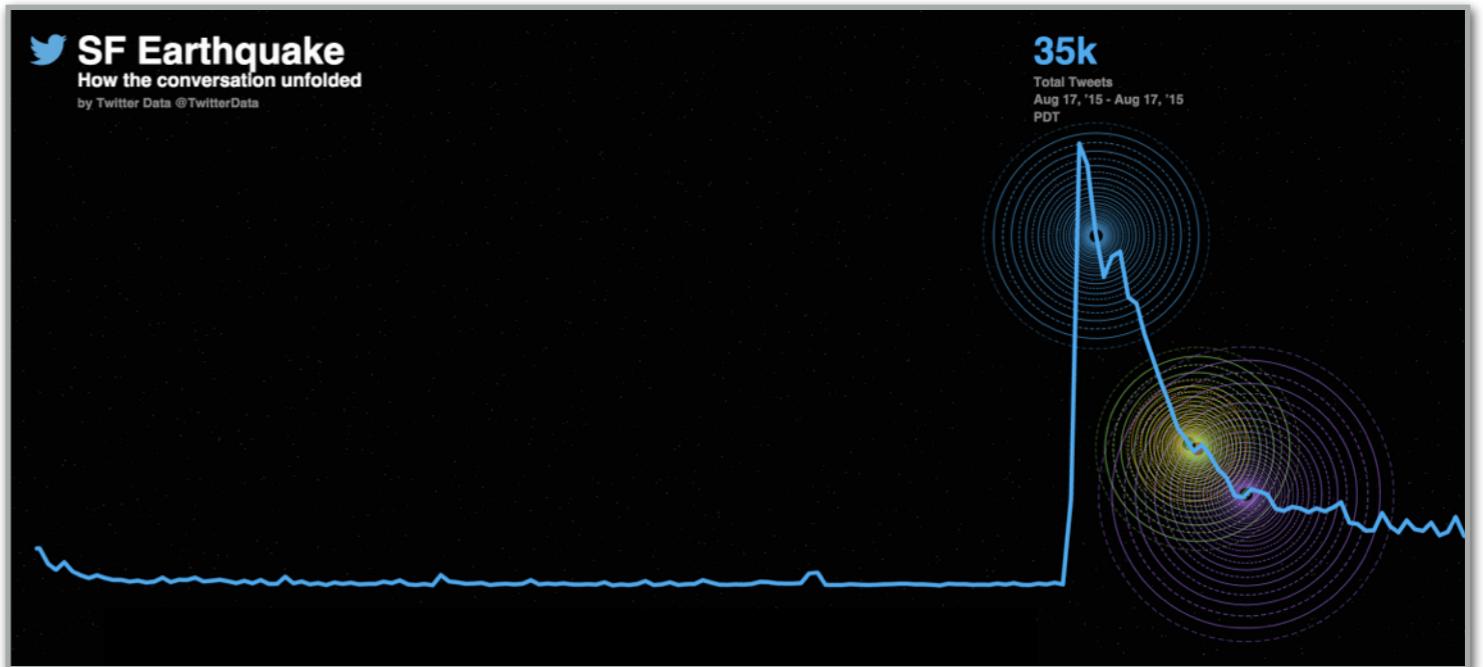
- Disease diagnosis based on previous correct cases
- Prediction of disease progression
- Medical image analysis and understanding
- Hospital information systems



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.

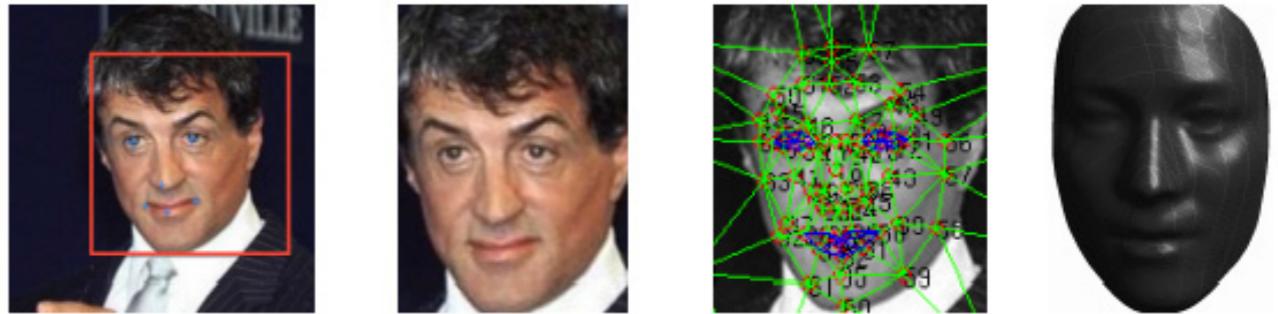
- *Cybersecurity*: Spike in number of false login attempts.
- *Payment systems*: Unusual number of failed or incomplete payments made.
- *Fraud detection*: Unexpected transactions in financial networks.
- *Event detection*:
Sudden spike in volume of social media posts.



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.

The screenshot shows a Google Mail interface with a search bar at the top containing "in:spam". Below the search bar, there are buttons for "Compose", "Inbox", "Starred", "Important", "Sent Mail", "Drafts", and "Less ▾". The main area displays a list of 71 spam messages. Each message includes the sender's name, subject, and timestamp. The messages are from various senders like Manna, Jacky, Joy, Henry, Alex, Hellen, Michael, Lori Liang, euro million, Shirley, and Ivy. The subjects include offers for AR111 LED, RC Drones, mechanical parts, mould design details, price quotations, and lottery notifications. The timestamps range from July 27 to August 7.

| Sender | Subject | Date |
|--------------|---|--------|
| Manna | New AR111 LED for The Best Price and Quality - Dear Purchase Manager, Good day! How are | 01:39 |
| Jacky | High quality RC Drone - Hello, This is Jacky Choi, Sales Manager from Guangtong Toys. We have | 7 Aug |
| Joy | steel & Aluminum forging factory from china - Dear friends, This is Joyce from AVIC. The factory | 4 Aug |
| Henry | enquiry new mechanical parts - Dear friend, If you are looking for high precise mechanical parts | 4 Aug |
| Alex | RE: mould design details - Dear friends, This is GMT, professional supplier for injection molding | 4 Aug |
| Henry | Latest Price from mechanical parts Suppliers - Dear Friends, Nice day to you, Hope I have cha | 2 Aug |
| Manna | New AR111 LED for The Best Price and Quality - Dear Purchase manager, Glad to know that you | 2 Aug |
| Hellen | LED indoor products, LED filament Manufacturer direct quotation - Dear Sir, How are you? I am | 1 Aug |
| michael | Re: new orders for RC toys - Dear, Guangdong Huanqi Electronic Co., Ltd. which is specialized i | 31 Jul |
| Jacky | RE: CFL & LED Test Instruments manufacturer - Hello, Lisun Group is the leader in CFL & LED | 30 Jul |
| Lori Liang | pet travel bowl - Dear Purchase Manager, Glad to learn you are in the market for pet travel bowl. | 30 Jul |
| euro million | WINNING NOTIFICATION - Euro Millions LOTTERY PROMOTION MADRID OFFICE WINNING N | 30 Jul |
| Jacky | RE: CFL & LED Test Instruments manufacturer - Hello, Lisun Group is the leader in CFL & LED | 28 Jul |
| Shirley | Knitting gloves and scarfs from China - Dear Sir/Madam, We are making Knitting gloves, bands | 28 Jul |
| Jacky | High quality RC Drone - Hello, This is Jacky Choi, Sales Manager from Guangtong Toys. We have | 28 Jul |
| Ivy | RE: Bearings manufacture - Hello, We are specialized in bearings more than 12 years. Our prod | 27 Jul |

Supervised v Unsupervised Learning

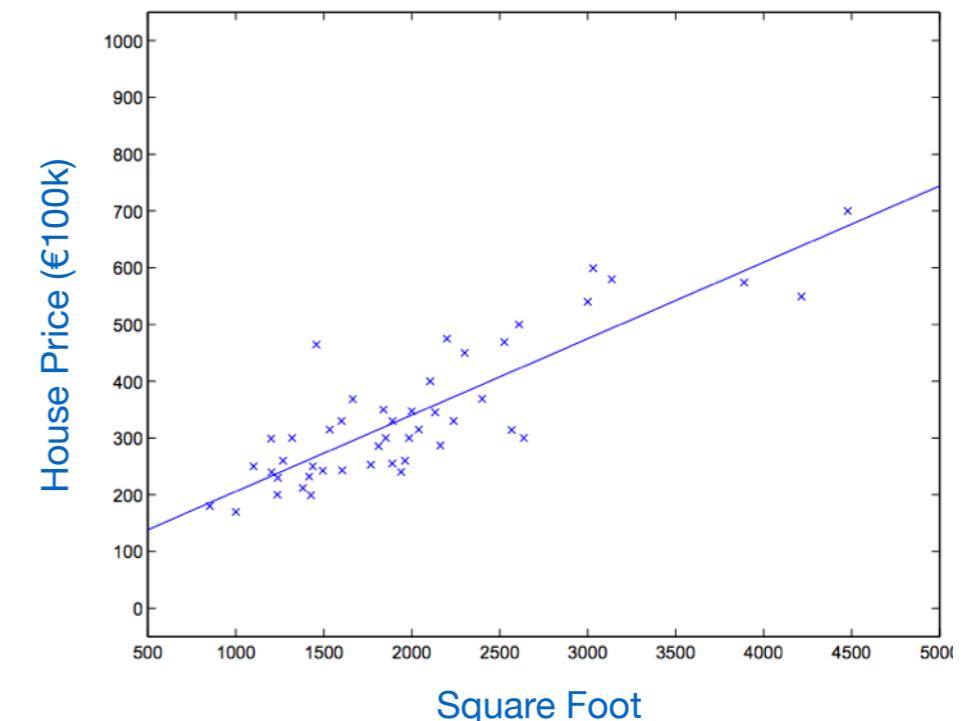
- **Supervised Learning:**
An algorithm that learns a function from examples of its inputs and outputs. It requires manually-labelled example data to learn the correct answer for a given query input.
 - e.g. Classification, Regression algorithms
- **Unsupervised Learning:**
An algorithm that finds patterns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

| | | |
|-----|-----------|---|
| 0 → | 000000000 | 0 |
| 1 → | 111111111 | 1 |
| 2 → | 222222222 | 2 |
| 3 → | 333333333 | 3 |
| 4 → | 444444444 | 4 |
| 5 → | 555555555 | 5 |

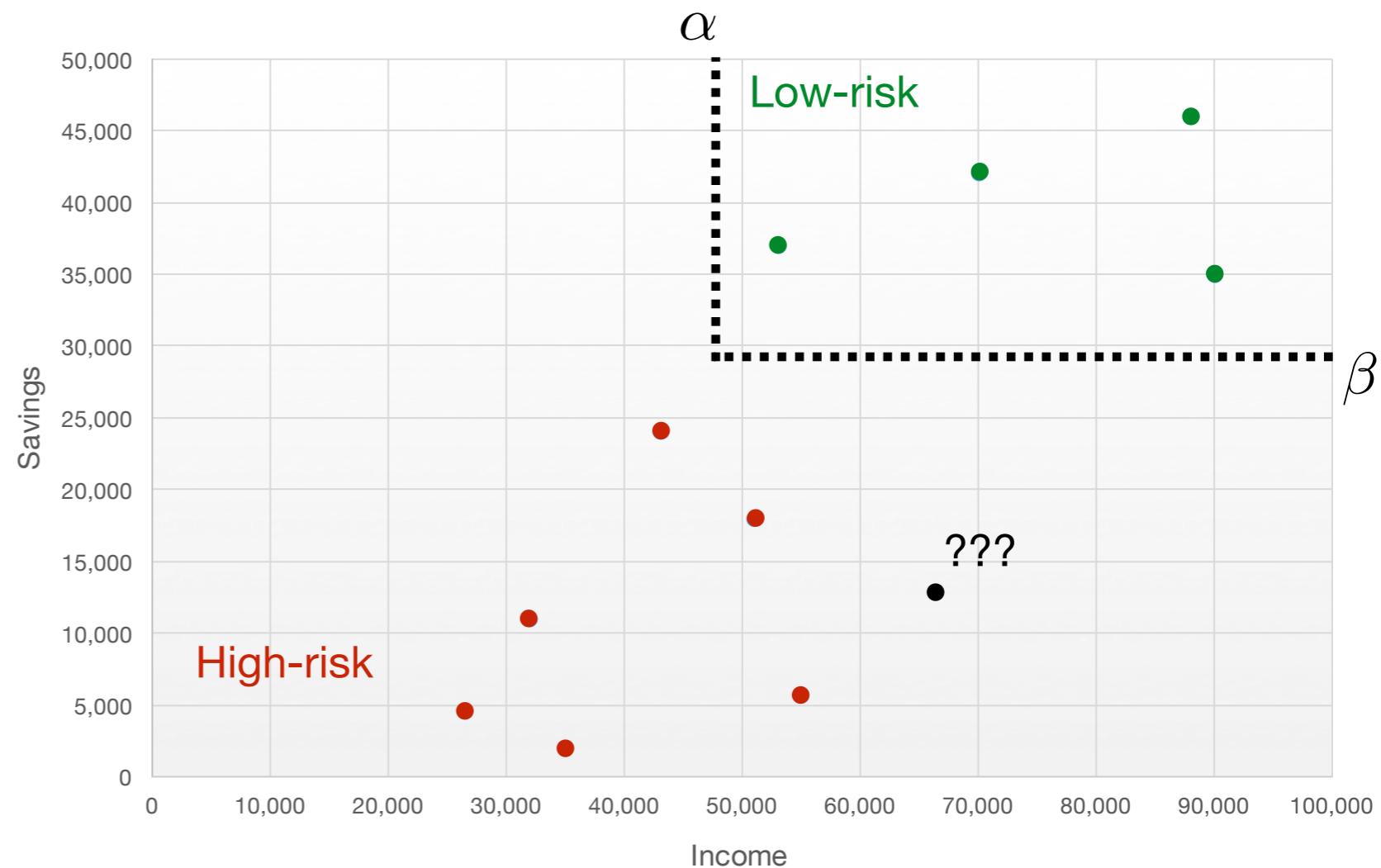
Labelled Examples Inputs



Typical Classification Task

Example: Credit Scoring

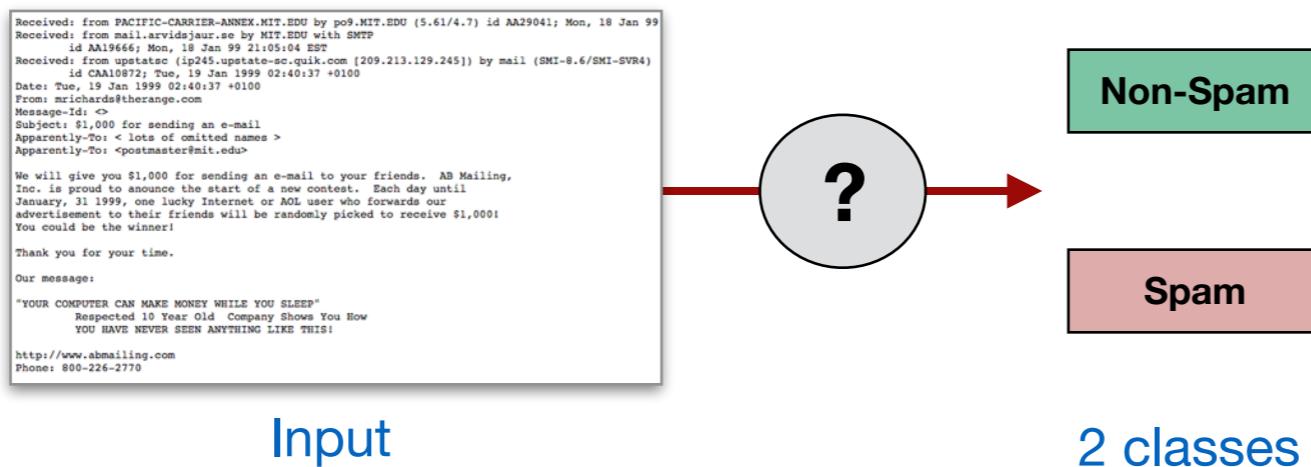
Manually classify customers into two categories (**low-risk** and **high-risk**) based on savings and income data.



- 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

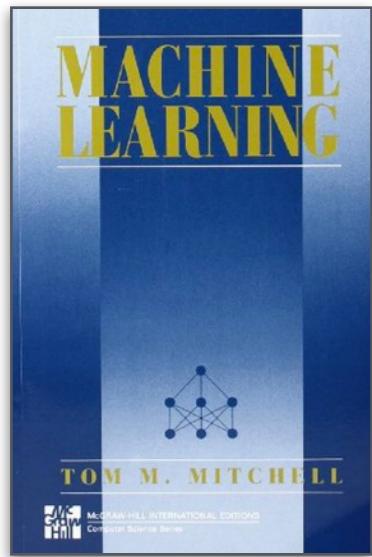
- Training set with N=10 examples (fruit). Each is described by D=4 features: 3 continuous, 1 categorical
- Each has one of three class labels = {Apple,Pear,Orange}

| 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?

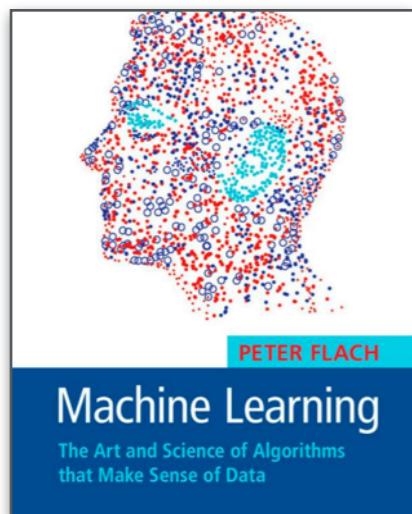
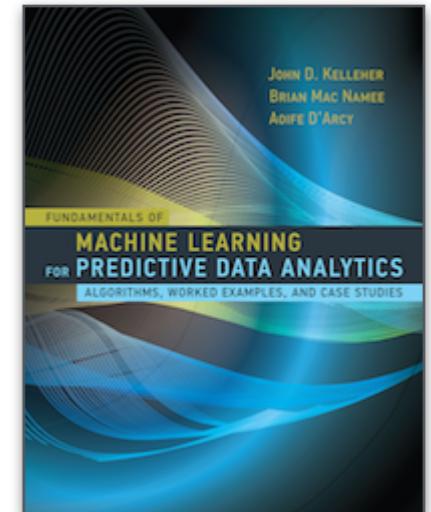
| Example | Height | Width | Taste | Weight | Class |
|---------|--------|-------|-------|--------|-------|
| X | 63 | 68 | Sweet | 168 | ??? |

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

