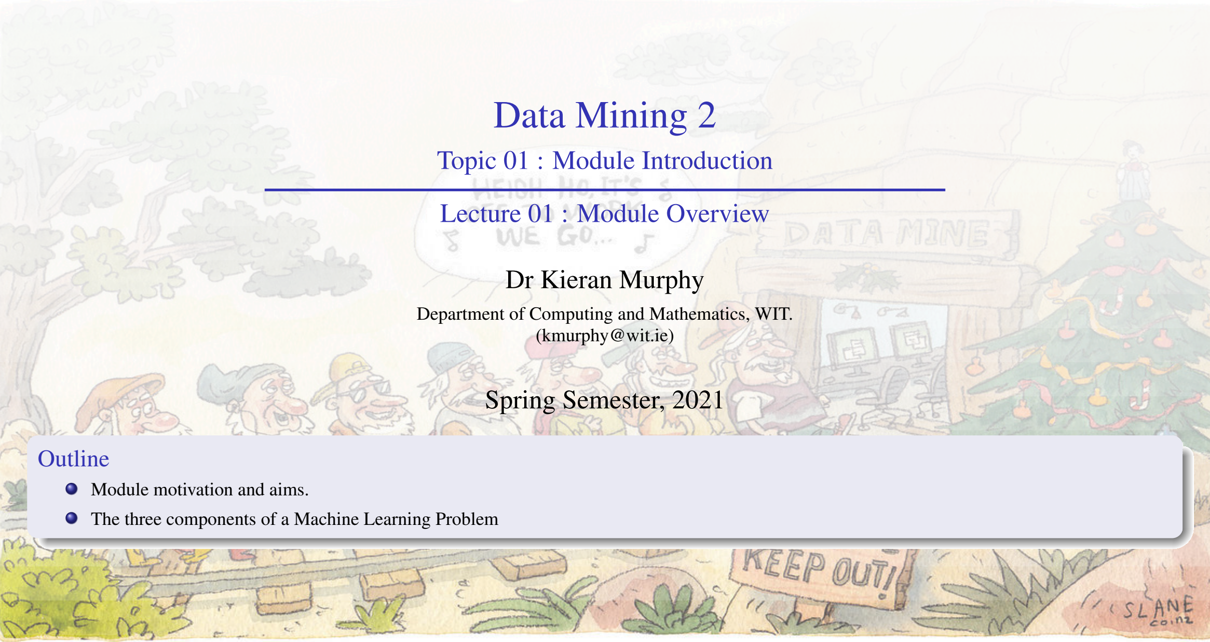


Data Mining 2

Topic 01 : Module Introduction

Lecture 01 : Module Overview

Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(kmurphy@wit.ie)

Spring Semester, 2021

Outline

- Module motivation and aims.
- The three components of a Machine Learning Problem

Outline

What is Data Mining ?

We are drowning in data but starving for knowledge!

Necessity is the mother of invention $\Rightarrow$ Data Mining $\approx$ Automated analysis of massive data sets.

Definition 1 (Data Mining)

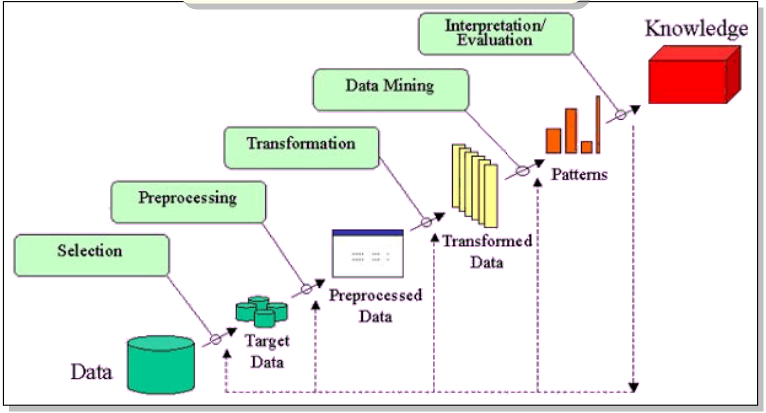
The non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data in large data repositories

non trivial	— obvious knowledge is not useful (we already know it)
implicit	— hidden difficult to observe knowledge
previous unknown	— if known then, why go to this effort?
potentially useful	— actionable easy to understand

Data Mining vs Knowledge Discovery in Data (KDD)

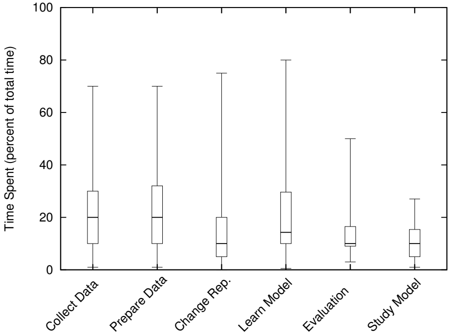
- Data mining and KDD are often used interchangeably.
- Actually data mining is only a part of the KDD process.

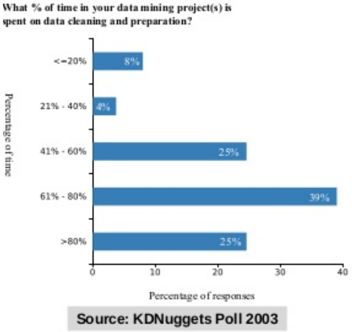
Knowledge Discovery in Data (KDD) Process



See A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)

Data Mining (Model Building) is less than half of Data Mining

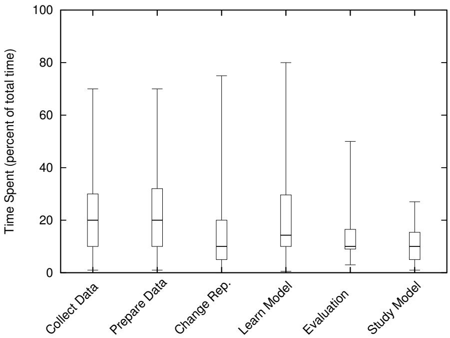


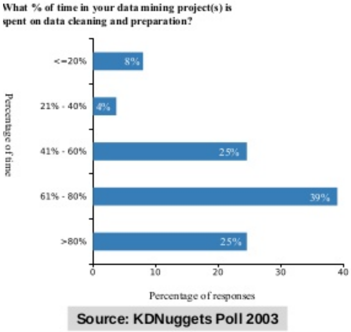


- Boxplots: median is 20% on collecting data, 20% on preparing data, and 10% on changing data representation — all before starting on model.
- Bar chart — data cleaning and preparation consumes at least 80% of project time for 25% of the participants, and 61% to 80% for another 39%.

See Study on the Importance of and Time Spent on Different Modeling Steps, 2012

Data Mining (Model Building) is less than half of Data Mining



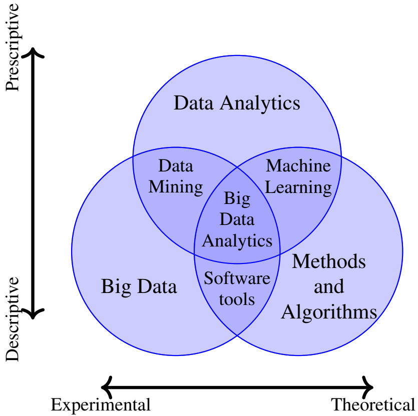


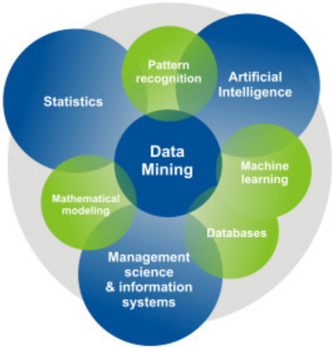
- Boxplots: median is 20% on collecting data, 20% on preparing data, and 10% on changing data representation — all before starting on model.
- Bar chart — data cleaning and preparation consumes at least 80% of project time for 25% of the participants, and 61% to 80% for another 39%.

See Study on the Importance of and Time Spent on Different Modeling Steps, 2012

Related Disciplines — Data Mining vs Data Analytics vs Data Science[*]

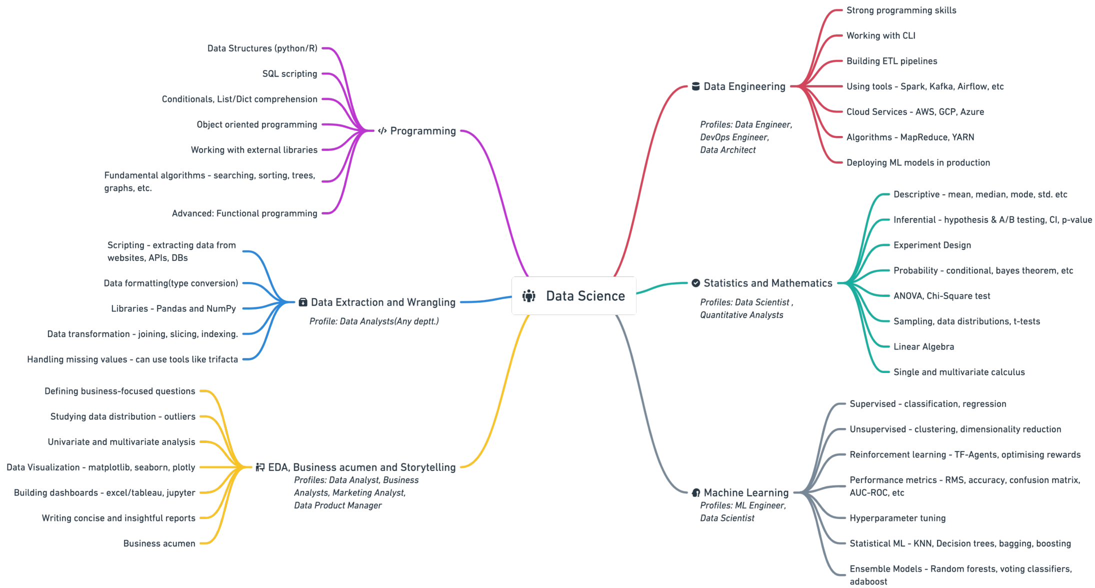
- Data Mining is about finding the patterns in a data set, and using these patterns to make predictions.
- Data Science is a field of study which includes everything from Big Data Analytics, Data Mining, Predictive Modelling, Data Visualisation, Mathematics, and Statistics.





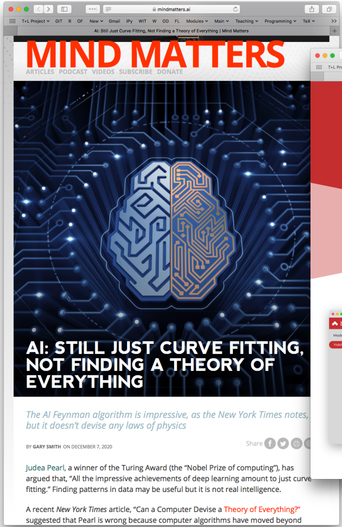
*In other words, have we titled this module correctly? Probably not, and it should be called Data Analytics 2 or Data Science 2

Data Science Mind Map

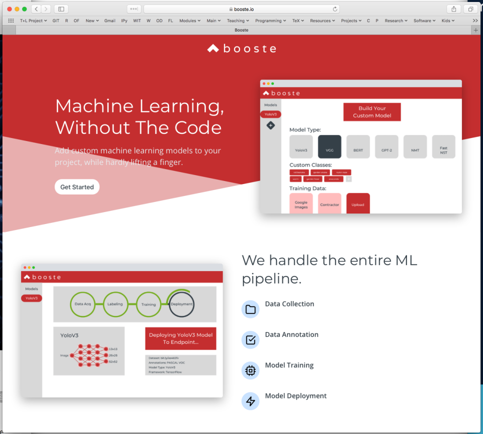


Data Science in 2021

There are still some skeptics . . .



. . . lower barriers and models as assets . . .

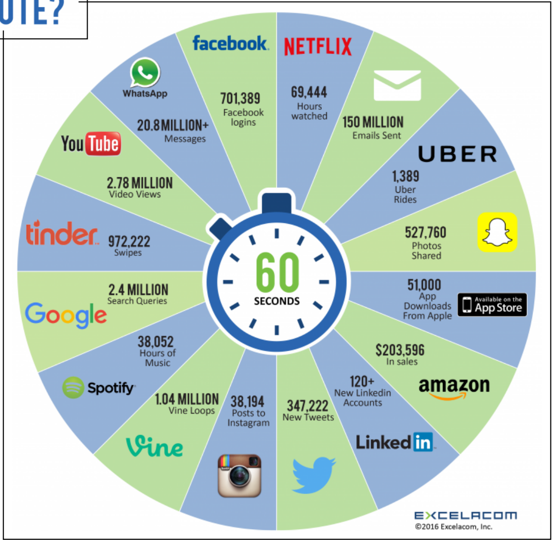


. . . MLOps

How Much Data?

By Month

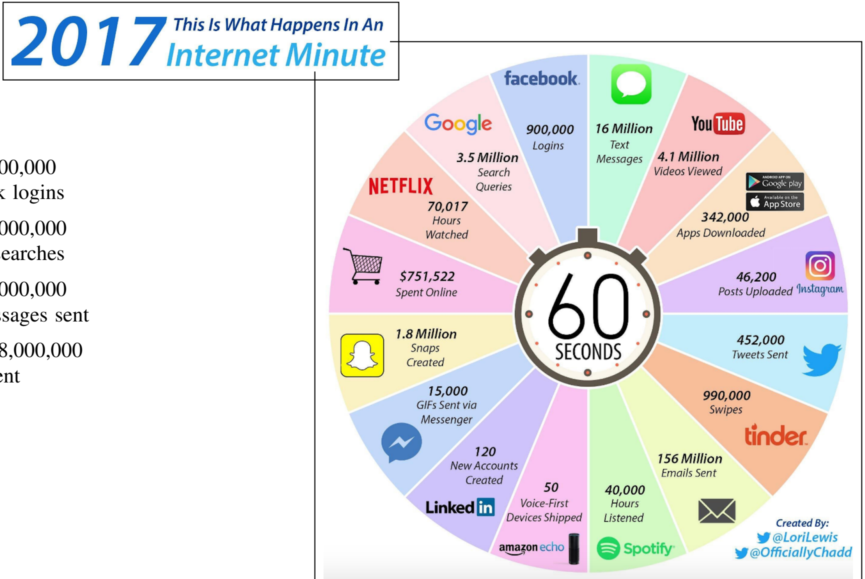
- 30,754,000,000
  Facebook logins
- 105,235,200,000
  Google searches
- 912,038,400,000
  WhatsApp messages sent
- 6,577,200,000,000
  emails sent
- 3.044.980.512
  Hours watched on Netflix



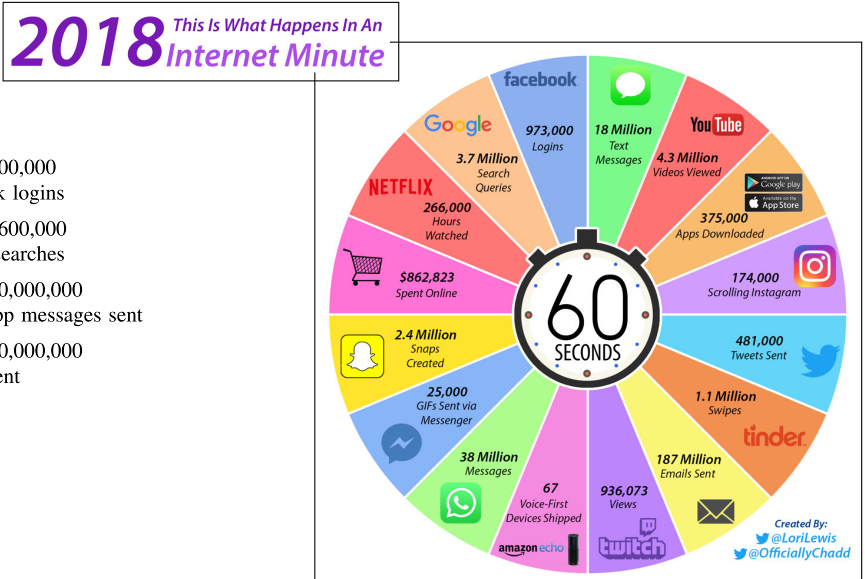
How Much Data?

By Month

- 39,463,200,000 Facebook logins
- 153,468,000,000 Google searches
- 701,568,000,000 Text messages sent
- 6,840,288,000,000 emails sent



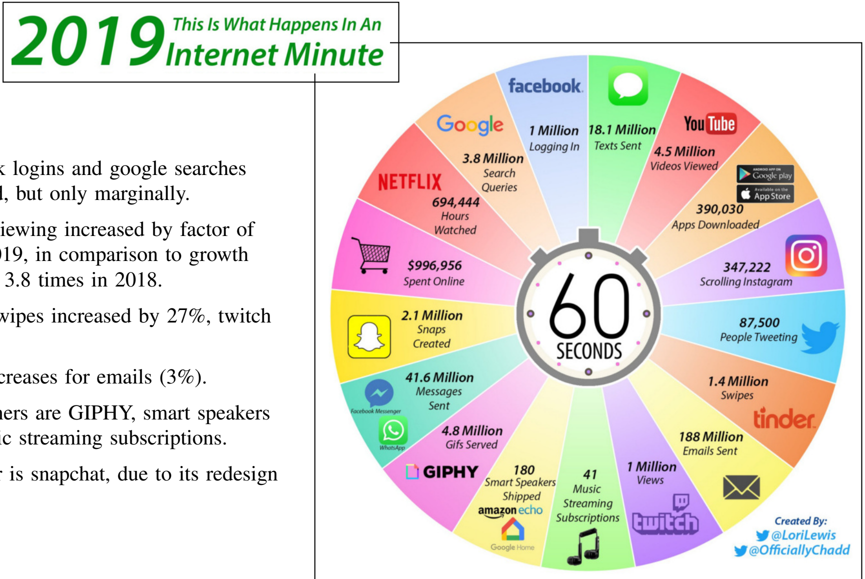
How Much Data?



By Month

- 42,033,600,000 Facebook logins
- 162,237,600,000 Google searches
- 1,641,600,000,000 WhatsApp messages sent
- 8,078,400,000,000 emails sent

How Much Data?



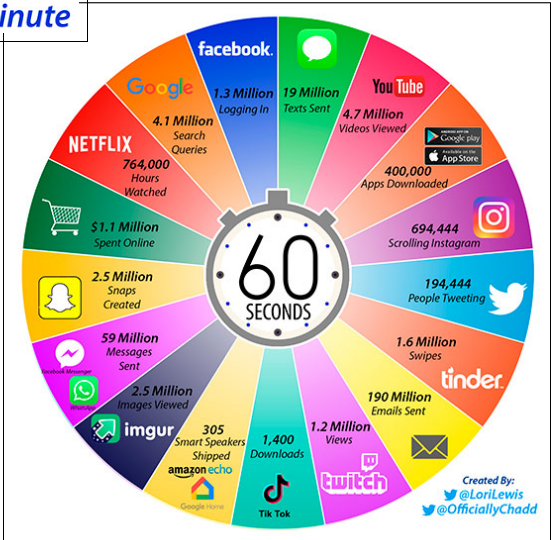
By Month

- Facebook logins and google searches increased, but only marginally.
- Netflix viewing increased by factor of 2.6 in 2019, in comparison to growth factor of 3.8 times in 2018.
- Tinder swipes increased by 27%, twitch by 20%.
- Small increases for emails (3%).
- Big winners are GIPHY, smart speakers and music streaming subscriptions.
- Big loser is snapchat, due to its redesign issues.

How Much Data?

By Month

- Instagram doubled !!
- Online shopping and Netflix both increased by only ≈10% !
- Facebook logins up by 30% — greater “news” consumption.
- Twitter more than doubled — what happened here?
- Smart speakers increased by 70%.
- Tinder swipes increased by 14%
- Number of emails sent nearly static.
- New additions — Tic Toc
- While SMS only increased. by 5%, messaging increased by 44%.



Technologies

Resources

- All lecture slides, handouts and datasets: GitHub — datamining2-202021.github.io/live
- All activities: quizzes and assignments: Moodle — moodle.wit.ie/course/view.php?id=171304

Delivery

- Scheduled zoom sessions, links in Moodle.
  - Lecture session can tend to get get very non-interactive so to help avoid this please have video on and please stop me at any point to ask questions or to tell me I've overlook a chat message etc.
  - Lectures and lab sessions may be recorded — in the sessions that I record I will post links to slack an hour or so after the session.
- Chat over zoom/slack:
  - Suggest using slack chat — if only for the stupid reason that it pings when a message comes in. But also, messages are viewable post the zoom call.

What can you do to mitigate online delivery limitations?

- Prepare — review material in advance of the sessions, install/download the software/datasets.
- Interact — yes, this is rich coming for an introvert mathication, but we live in strange times.
- Time management — give tasks a serious/focused effort, but when stuck ask for help.

Assessment Structure — 100% Continuous Assessment

Covering skills

- Data Wrangling + Feature Engineering (pandas and friends)
- NLP, Text processing (regex)
- Model building and optimisation (skilearn, tensorflow, . . .)

Breakdown

- Metric:
  - 20% Student engagement + 80% Demonstration of skills/understanding
- Activities:
  - Moodle quizzes based on analysing datasets / model building / etc.
  - Data science problems with mixture of Kaggle style grading and traditional grading.

Calandar

- Week 14/15 end of semester individual review interview (zoom).
- $\underbrace{\text{4 weeks + slow week + 4 weeks + Easter break (2 weeks) + 3 weeks}}_{\text{12 teaching weeks}}$ +3 weeks for CA

Outline

Three Components of a Machine Learning Problem

It is easy to get lost among the multitude of choices one needs to make when given data mining problem. A good decomposition is the following:

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

[†] A Few Useful Things to Know about Machine Learning, Domingos, 2012.

3 Components — Representation

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search

Representation refers to formulating the problem as a machine learning problem — typically a classification problem, a regression problem or a clustering problem.

- How do we represent the input?
- What features to use?
- How do we learn additional features?
- With each type of problem, we have multiple subtypes.
  For example which classifier? a decision tree, a neural network, a support vector machine, a hyperplane that separates the two classes etc.

3 Components — Evaluation

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search

Evaluation refers to an objective function or a scoring function, to distinguish good models from a bad model.

- For a classification problem, we need this function to know if a given classifier is good or bad. A typical function can be based on the number of errors made by the classifier on a test set, using precision and recall.
- For a regression problem, it could be the squared error, or likelihood. Do we include regularisation? etc

3 Components — Optimisation

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search

Optimisation is concerned with searching among the models in the language for the highest scoring model.

- How do we search among all the alternatives?
- Can we use some greedy approaches, branch and bound approaches, gradient descent, linear programming or quadratic programming methods.