

# 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

# 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.

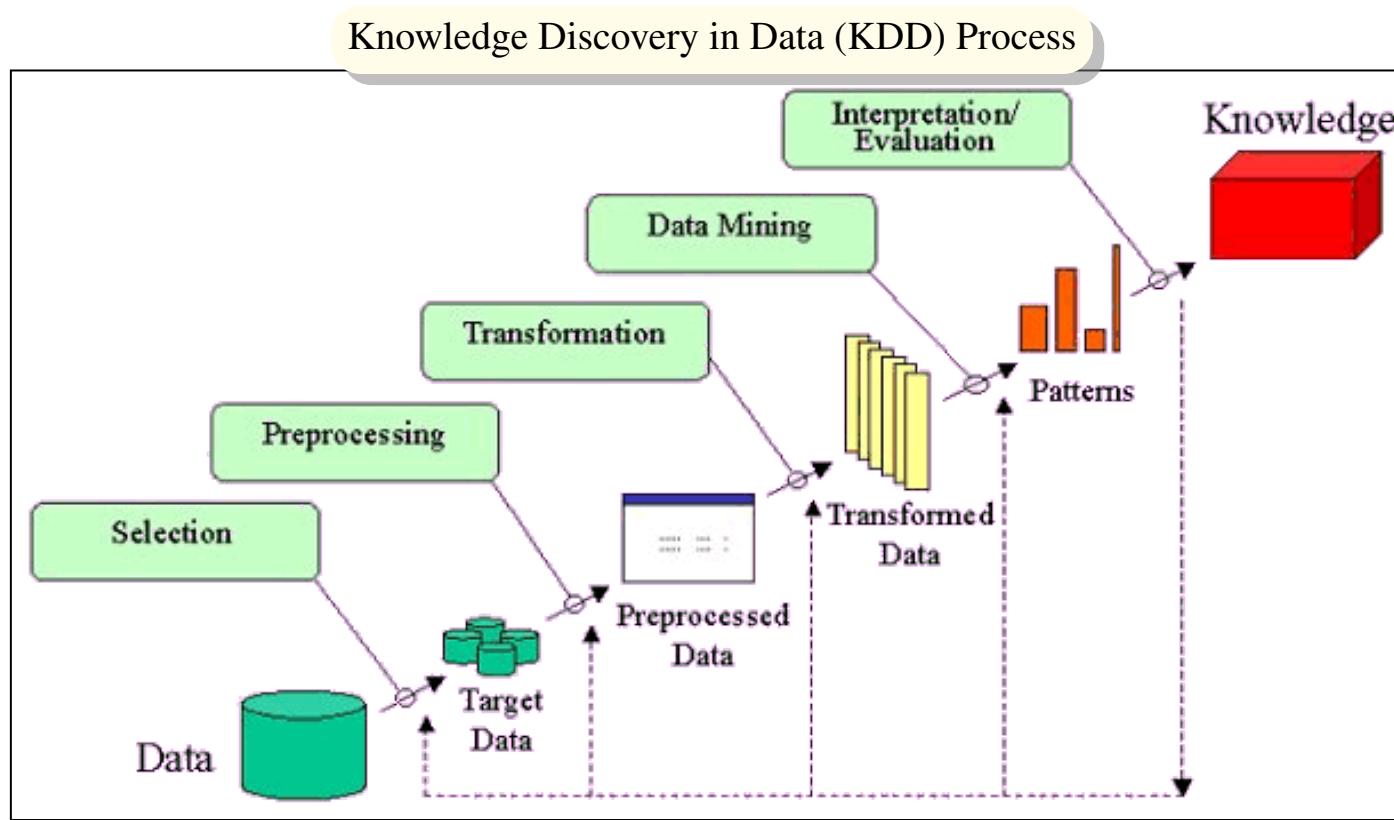
## Definition 1 (Data Mining)

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

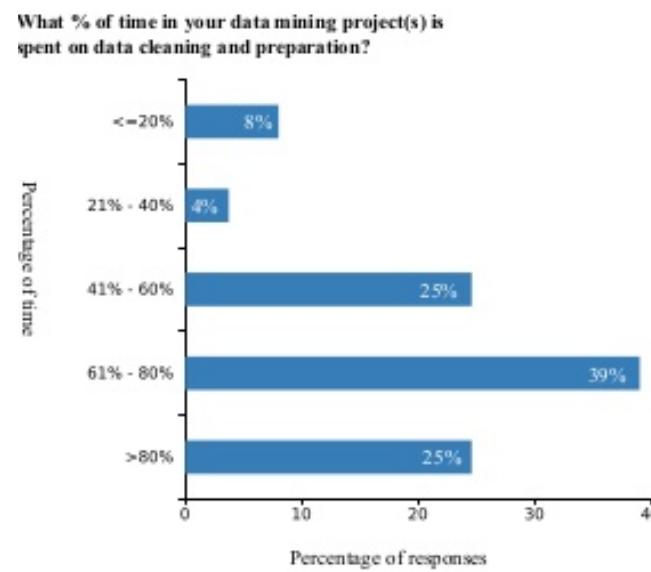
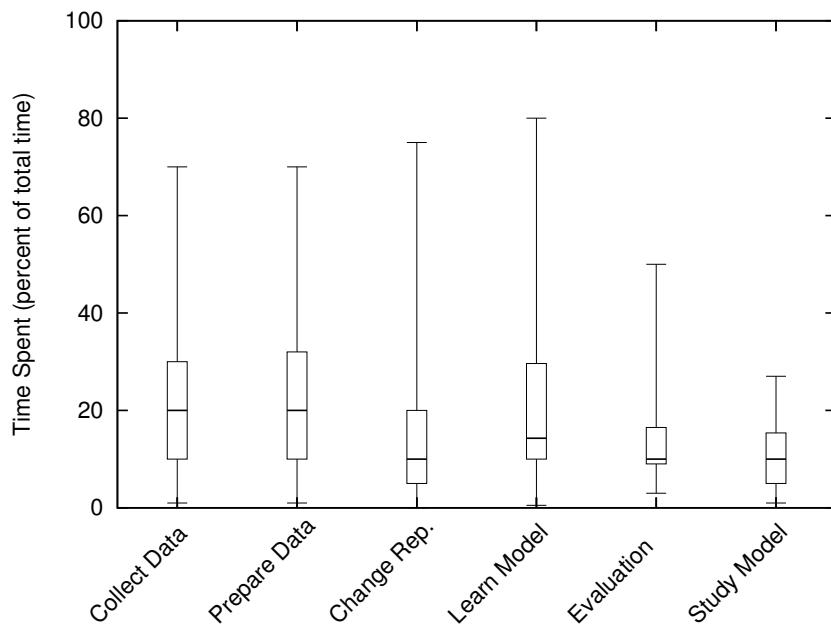
# Data Mining vs Knowledge Discovery in Data (KDD)

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



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



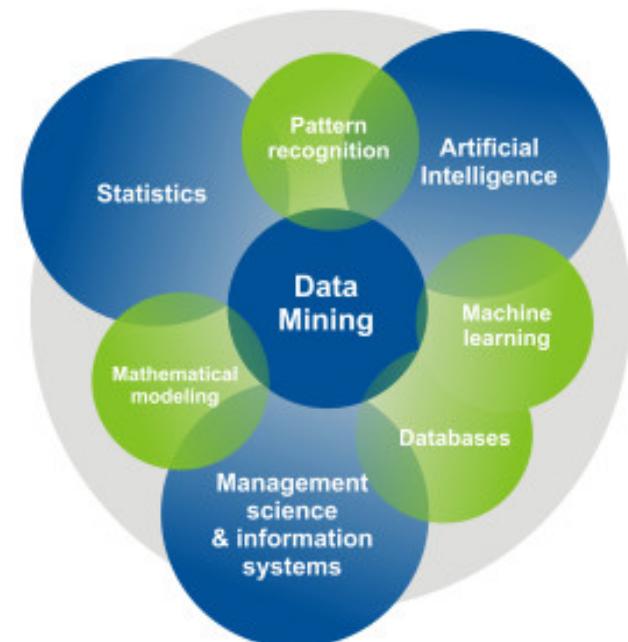
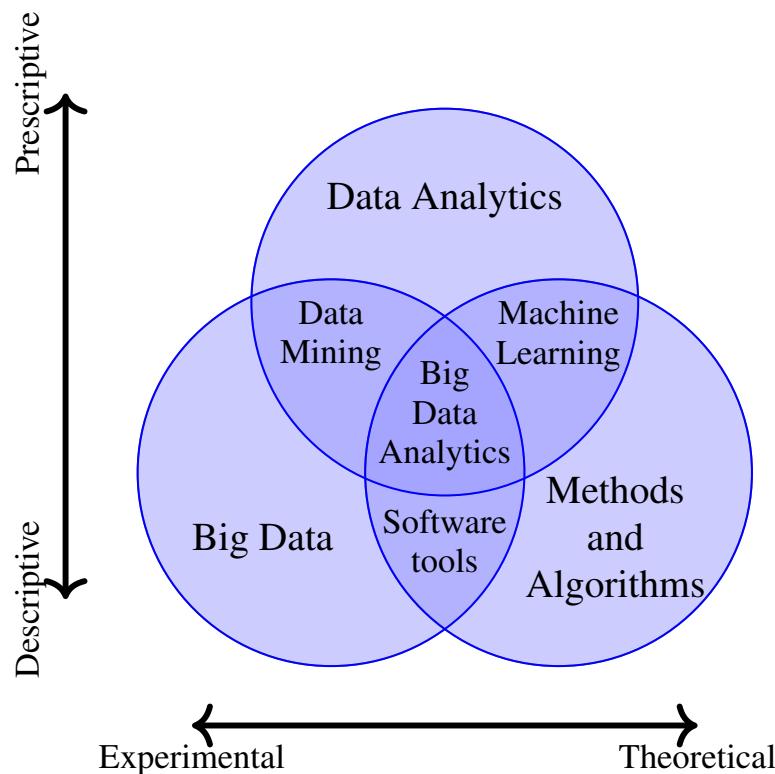
Source: KDNuggets Poll 2003

- 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](#)

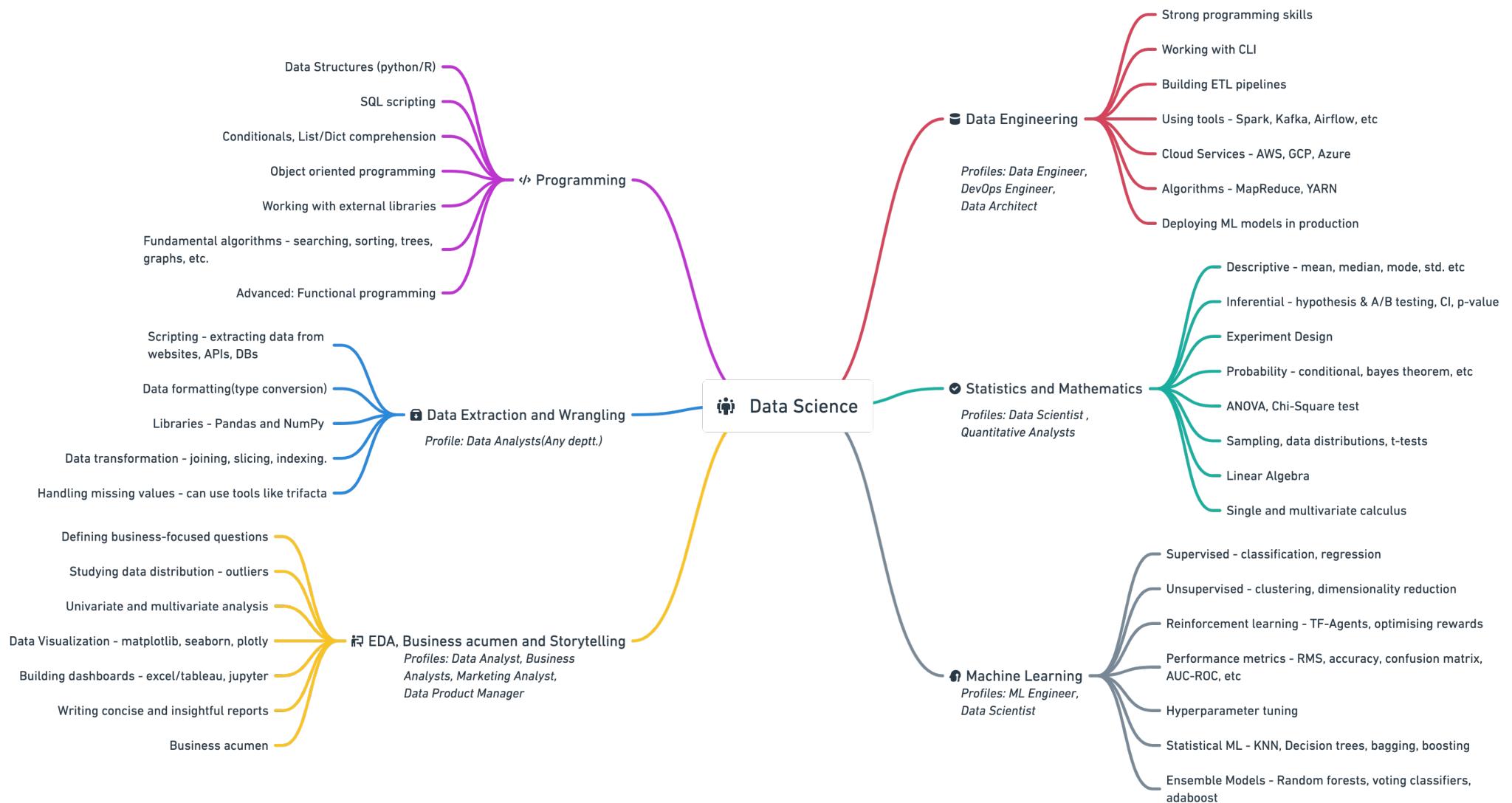
# Related Disciplines — Data Mining vs Data Analytics vs Data Science<sup>†</sup>

- 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



What? Why? and How?      What?

# Data Science in 2021

There are still some skeptics ...

The screenshot shows the homepage of mindatters.ai. At the top, there's a navigation bar with links like T+L Project, GIT, R, OF, New, Gmail, IPy, WIT, W, OD, FL, Modules, Main, Teaching, Programming, TeX, etc. Below the navigation, a large red banner with the text "MIND MATTERS" and sub-links for ARTICLES, PODCAST, VIDEOS, SUBSCRIBE, and DONATE. The main content features a large image of a brain with a circuit board pattern overlaid. A prominent headline reads "AI: STILL JUST CURVE FITTING, NOT FINDING A THEORY OF EVERYTHING". Below the headline, a quote from the New York Times is displayed: "The AI Feynman algorithm is impressive, as the New York Times notes, but it doesn't devise any laws of physics". At the bottom, there's a section about Judea Pearl's Turing Award win and a recent New York Times article.

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

This screenshot shows the booste.io platform. It features a red header with the text "Machine Learning, Without The Code" and a subtext: "Add custom machine learning models to your project, while hardly lifting a finger." Below this is a "Get Started" button. To the right, there's a "Build Your Custom Model" interface with sections for "Model Type" (YoloV3, VGG, BERT, GPT-2, NMT, Fast NST), "Custom Classes" (with three buttons: "Automated", "Manual", and "Custom"), and "Training Data" (with three buttons: "Google Images", "Contractor", and "Upload"). Below this, there's a smaller screenshot of the platform showing a pipeline: Data Acq → Labelling → Training → Deployment, with a YoloV3 model example. The pipeline has four green circles indicating progress.

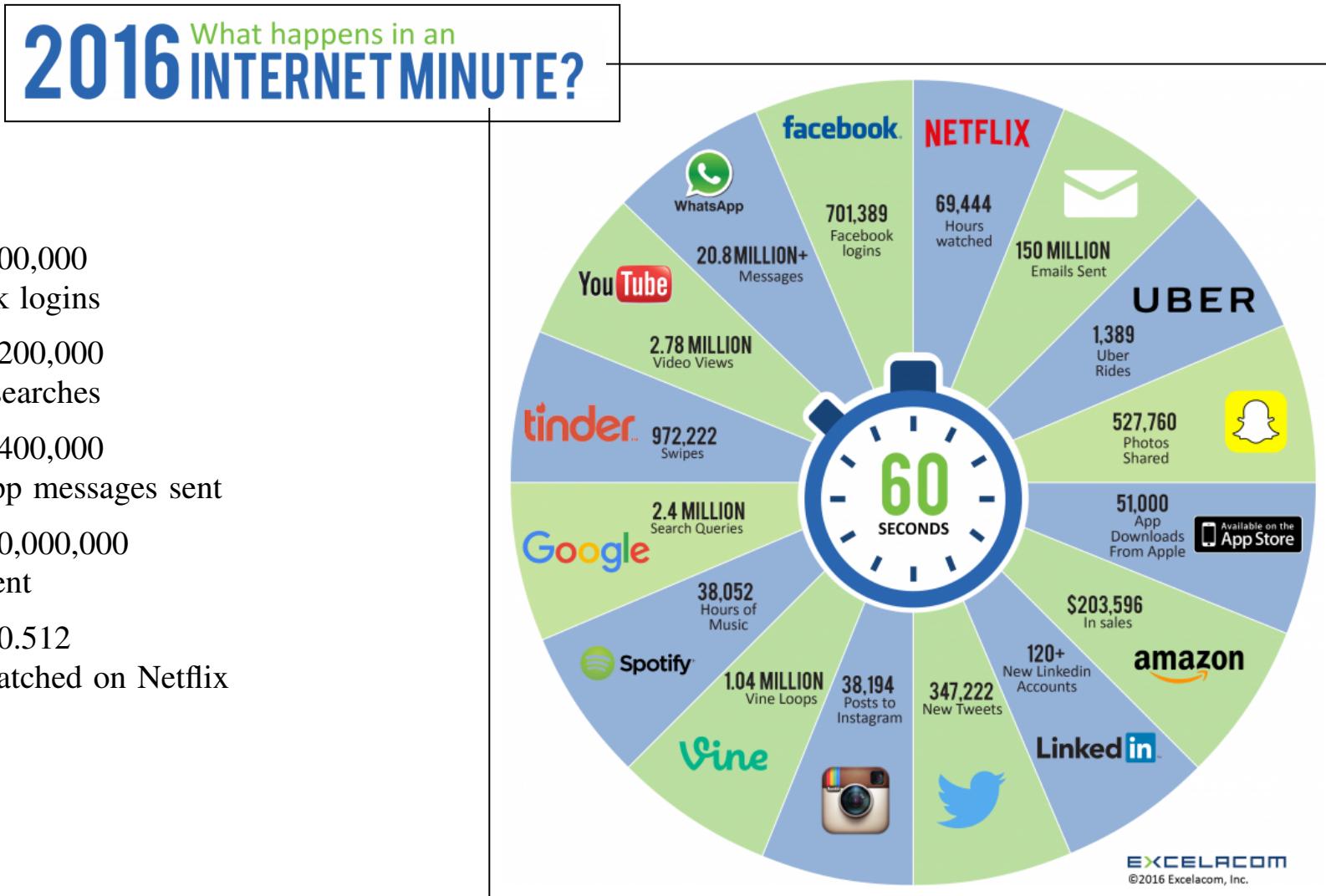
We handle the entire ML pipeline.

- Data Collection
- Data Annotation
- Model Training
- Model Deployment

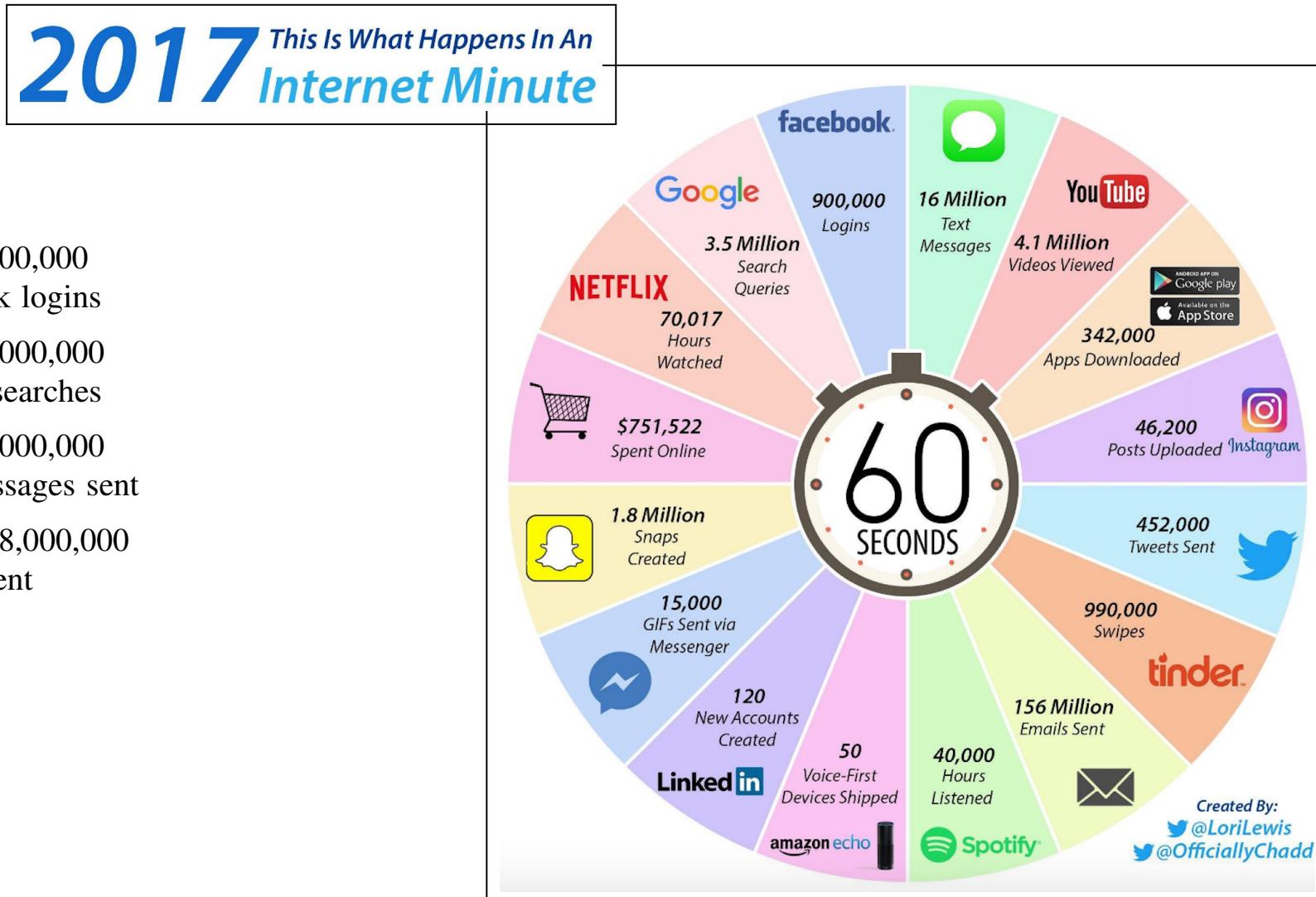
... MLOps

The screenshot shows the MLflow.org website. The header includes links for T+L Project, GIT, R, OF, New, Gmail, IPy, WIT, W, OD, FL, Modules, Main, Teaching, Programming, TeX, Resources, Projects, C, P, Research, Software, Kids, etc. The main content area has a dark blue background with a blue wavy network graphic. The text "An open source platform for the machine learning lifecycle" is centered. On the right, there's a "Latest News" sidebar with links to MLflow 1.13.1, 1.13.0, and 1.12.1 releases, and an announcement for PyTorch and MLflow integration. At the bottom, there are icons for MLflow's compatibility with various frameworks and its scalability with Apache Spark.

# How Much Data?



# How Much Data?



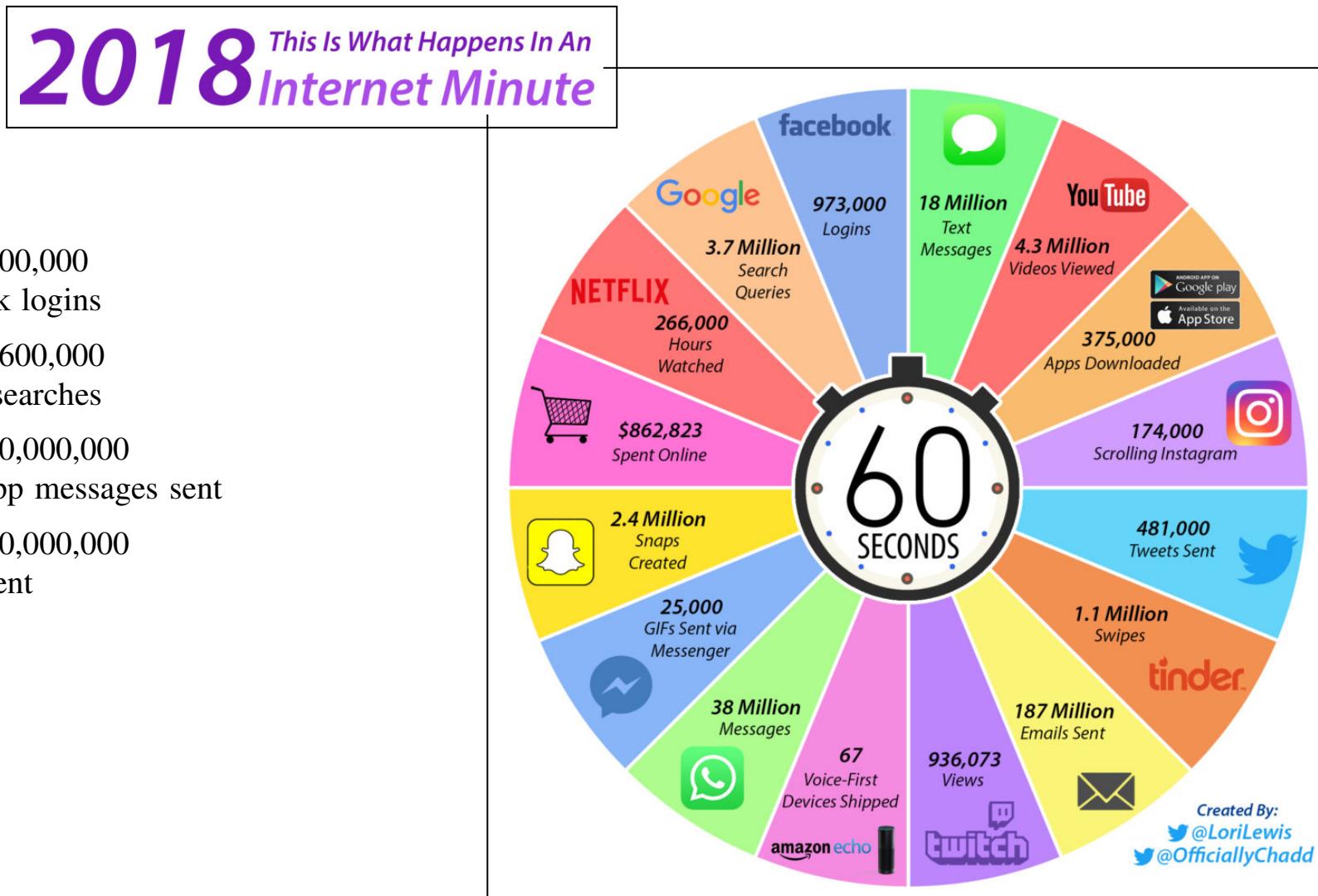
## By Month

- 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

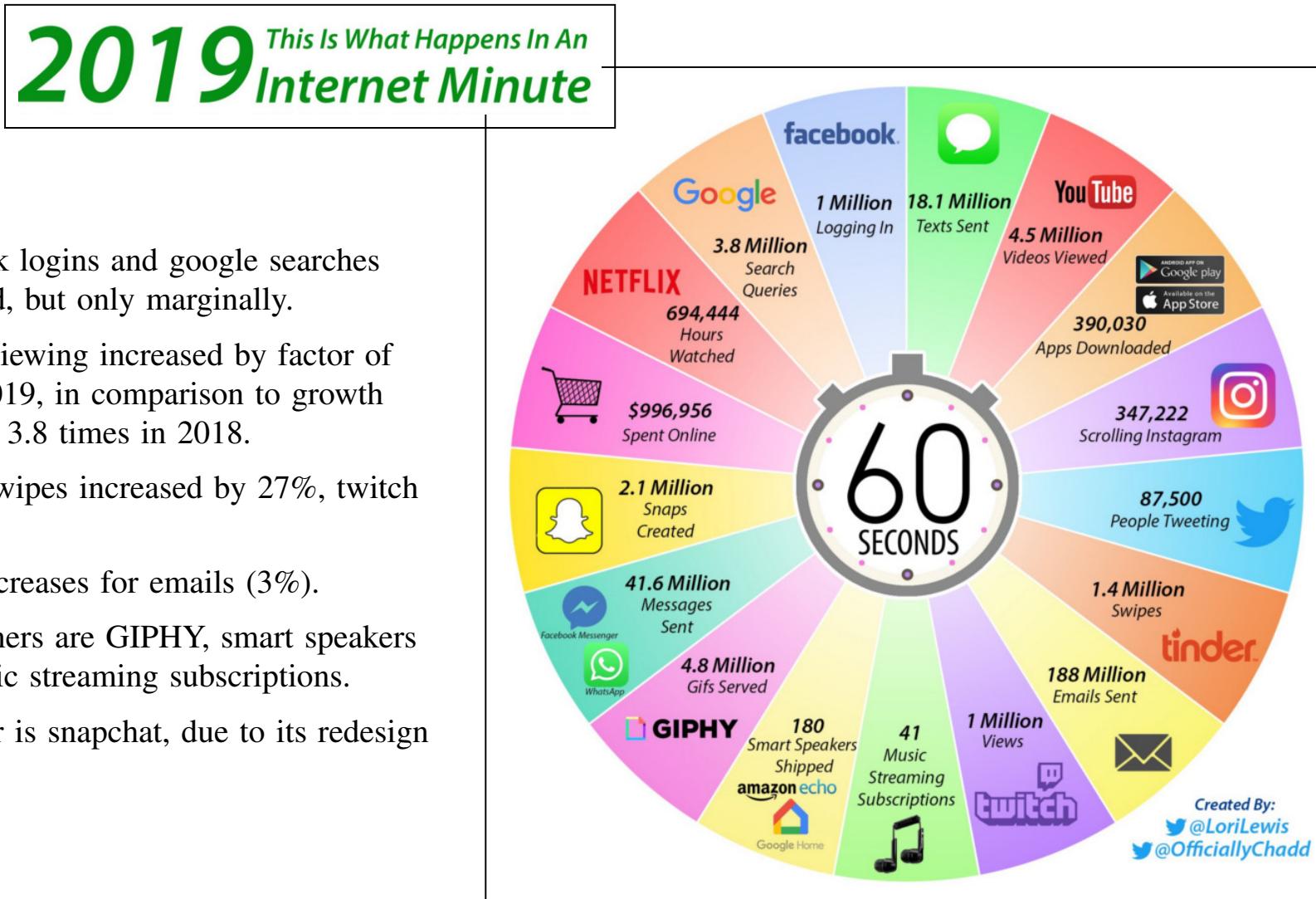
# How Much Data?

## By Month

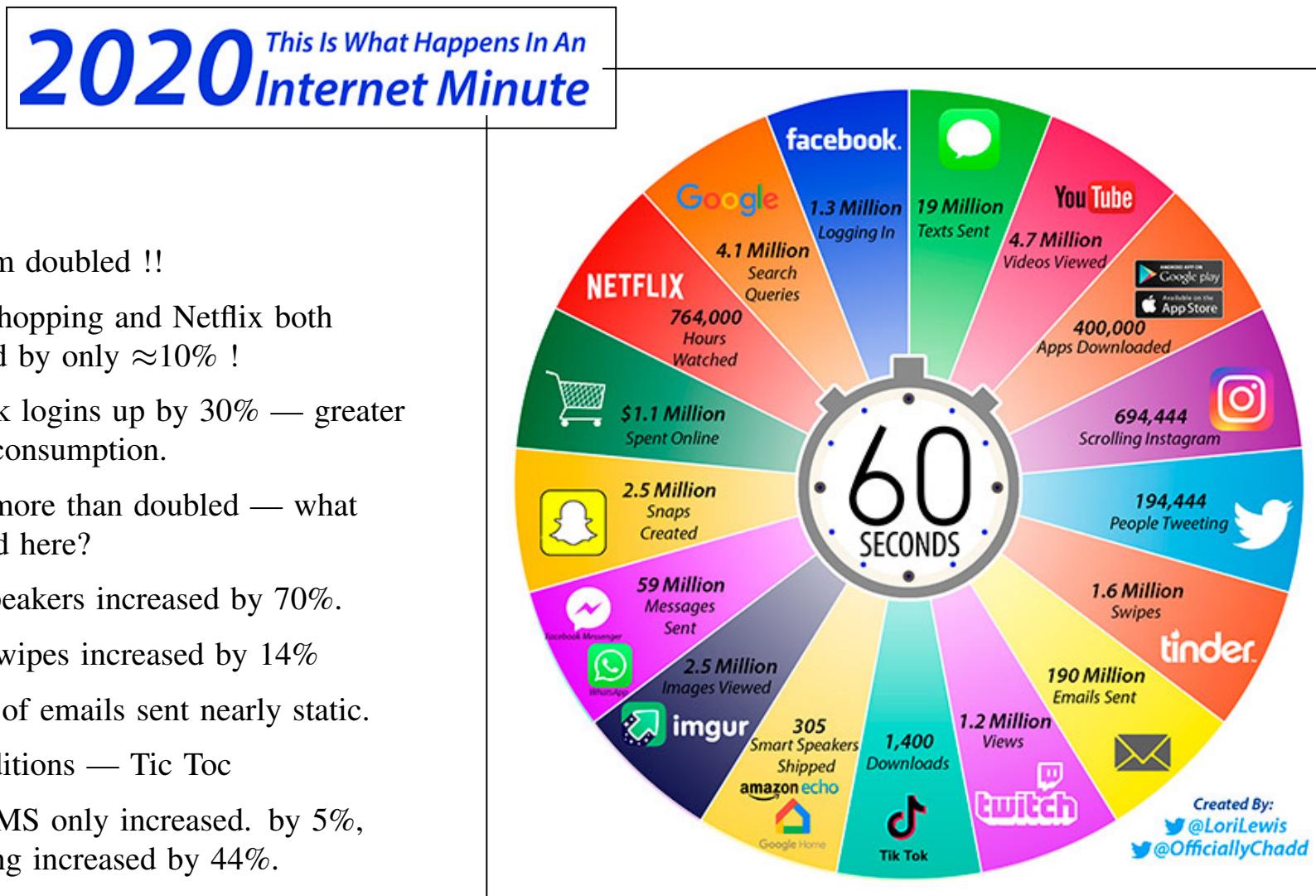
- 42,033,600,000 Facebook logins
- 162,237,600,000 Google searches
- 1,641,600,000,000 WhatsApp messages sent
- 8,078,400,000,000 emails sent



# How Much Data?



# How Much Data?



# Technologies

## Resources

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

## Delivery

- Scheduled zoom sessions, links in Moodle.
  - Lecture session can tend to 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 overlooked 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 matriculation, 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).
- 4 weeks + slow week + 4 weeks + Easter break (2 weeks) + 3 weeks + 3 weeks for CA  
12 teaching weeks

# 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		

---

<sup>†</sup>A Few Useful Things to Know about Machine Learning, Domingos, 2012.

## 3 Components — Representation

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