COMP9321

Data Services Engineering

Term 1, 2019

Week 1 Lecture 1

Course Overview

COMP9321 2019T1

UNSW
SYDNEY

UNSW
SYDNEY

1.5 kg cherries
juice of 1 lemon
30 g icing sugar
150 g butter
80 g caster sugar
½ tablespoon water
200 g plain flour
1 egg
zest of 1 lemon
2 tablespoons cornflour

1 Place cherries in medium saucepan and place over heat. Cover. After the cherries lose considerable juice, which may take a few minutes, remove from heat. In a small bowl, mix the sugar and cornstarch together. Pour this mixture into the hot cherries and mix well. Add the almond extract, if desired, and mix. Return the mixture to the stove and cook over low heat until thickened, stirring frequently. Remove from the heat and let cool. If the filling is too thick, add a little water, too thin, add a little more cornstarch.
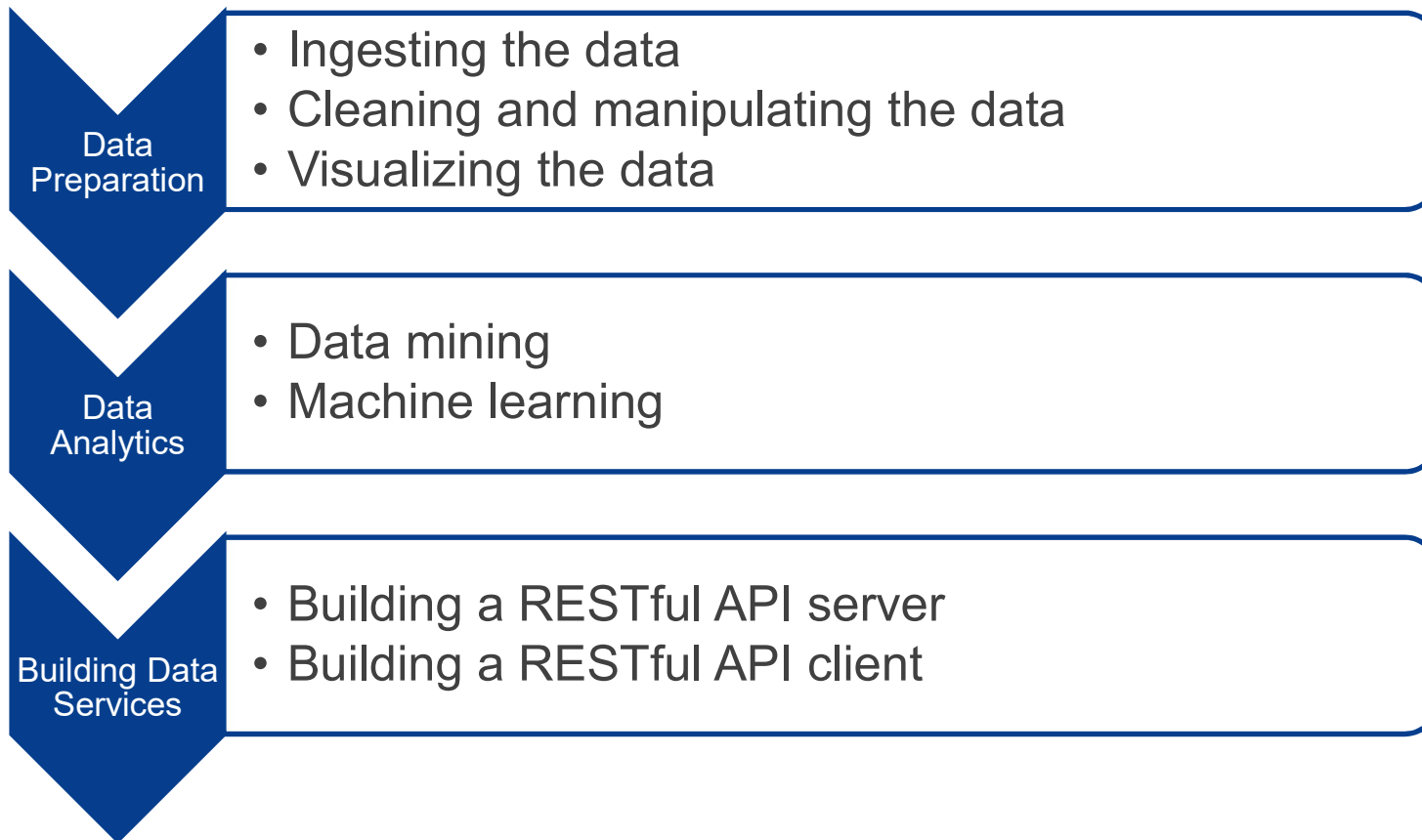
2 Preheat the oven to 375 degrees F.

3 Use your favorite pie dough recipe. Prepare your crust. Divide in half. Roll out each piece large enough to fit into an 8 to 9-inch pan. Pour cooled cherry mixture into the crust. Dot with butter. Moisten edge of bottom crust. Place top crust on and flute the edge of the pie. Make a slit in the middle of the crust for steam to escape. Sprinkle with sugar.

4 Bake for about 50 minutes. Remove from the oven and place on a rack to cool.

UNSW
SYDNEY

Course Structure

Data Preparation
- Ingesting the data
- Cleaning and manipulating the data
- Visualizing the data

Data Analytics
- Data mining
- Machine learning

Building Data Services
- Building a RESTful API server
- Building a RESTful API client

UNSW
SYDNEY

Teaching Team

- *Lecturer-in-Charge (LiC)*
  - Lina Yao (lina.yao@unsw.edu.au)
  - Office: K17 401I
  - Web: http://www.linayao.com/

- *Course Administrator*
  - Chaoran Huang (chaoran.huang@unsw.edu.au)

- *Tutors*
  - Alireza Tabebordbar   a.tabebordbar@unsw.edu.au
  - Chaoran Huang   chaoran.huang@unsw.edu.au
  - Mohammadali Yaghoubzadehfard   m.yaghoubzadehfard@unsw.edu.au
  - Shayan Zamanirad   shayan.zamanirad@unsw.edu.au
  - Xiaocong Chen   xiaocong.chen@unsw.edu.au

- *Course Web site*

  - http://www.cse.unsw.edu.au/~cs9321/

UNSW
SYDNEY

COMP9321 Evolution

Previously known as Web applications engineering.

What was taught and why needed revision

- How to build Web sites using Java
- Standardised frameworks for Web apps (plenty)

Many Web apps are now data-oriented or utilise data heavily
  –functionality requires combining or processing complex data from multiple
   sources

So COMP9321 became Data Service Engineering:

- How to work with data (Data processing)
- How to make the design and implementation of data-oriented application easy
  (i.e., an approach/technique)

UNSW
SYDNEY

So what is this course about?

Data Services Engineering

Data = is the problem we want to deal with, understanding the problems and
possible ways to work with Data (e.g., "get" data, "publish" data, discover or
manage multiple data sources, etc).

Services = is the proposed solution/design approach to make our problem
"manageable".

Engineering = (best practices, weighing options, we will think about these all
throughout, at least try to) - obtain conceptual ideas as well as practical skills

Course Aims

This course aims to introduce the student to core concepts and practical skills for engineering the data in service-oriented data-driven applications. Specifically, the course aims to answer these questions:

- *How to access and ingest data from various external sources?*

- *How to process and store the data for applications?*

- *How to curate (e.g. Extract, Transform, Correct, Aggregate, and Merge/Split) and publish the data?*

- *How to visualize the data to communicate effectively?*

- *How to apply available analytics to the data?*

- *How to utilize recent recommender systems to help making decisions?*

Fundamentally, we will look at these questions through the lens of 'service-oriented' software design and implementation principles. At each topic, we will learn some core concepts, and how to implement the concepts in software through services.

UNSW
SYDNEY

Assumed Knowledge

Before commencing this course, we will assume that students have:

- completed one programming course (expected to be in Python)
- basic data modelling and relational database knowledge
- basic linear algebra knowledge

These are assumed to have been acquired in the following courses:

For Postgrad - COMP9021 and COMP9311

For Undergrad – (COMP1531 or COMP2041) and COMP3311

UNSW
SYDNEY

Assessment

Assessment:

- 50% exam: individual assessment. Format TBA.
- 40% on assignment work(10%+10%+20%)
- 10% on 7 online quizzes (WebCMS3-based quiz system, 'open' test)

Final Mark = quizzes + assignments + exam

Note:

   to pass this course you have to get **at least 25/50** marks in **final exam.**

   <u>**Plagiarism checking**</u> will be run for each assignment

UNSW
SYDNEY

Assignments

We have two individual assignments and one group project.

Assignment 1: Data ingestion and manipulation (*individual*):

- 10 marks

- Release today, due on the end of week 3.

Assignment 2: Data publication as a RESTful service API (*individual*):

- 10 marks

- Release on the Tuesday of week 3, due on the end of week 5.

Assignment 3: Data analytics project (*group*):

- 20 marks

- Release on the Tuesday week 5, due on the end week 8.

- Demo on week 9, schedule will be released after submission

Labs

- A self-guided lab exercise is released (roughly) every week.

- You can do them in your own time, but you are encouraged to try and complete as much as you can during the class.

- Use the forum. Share what you have learned/found.
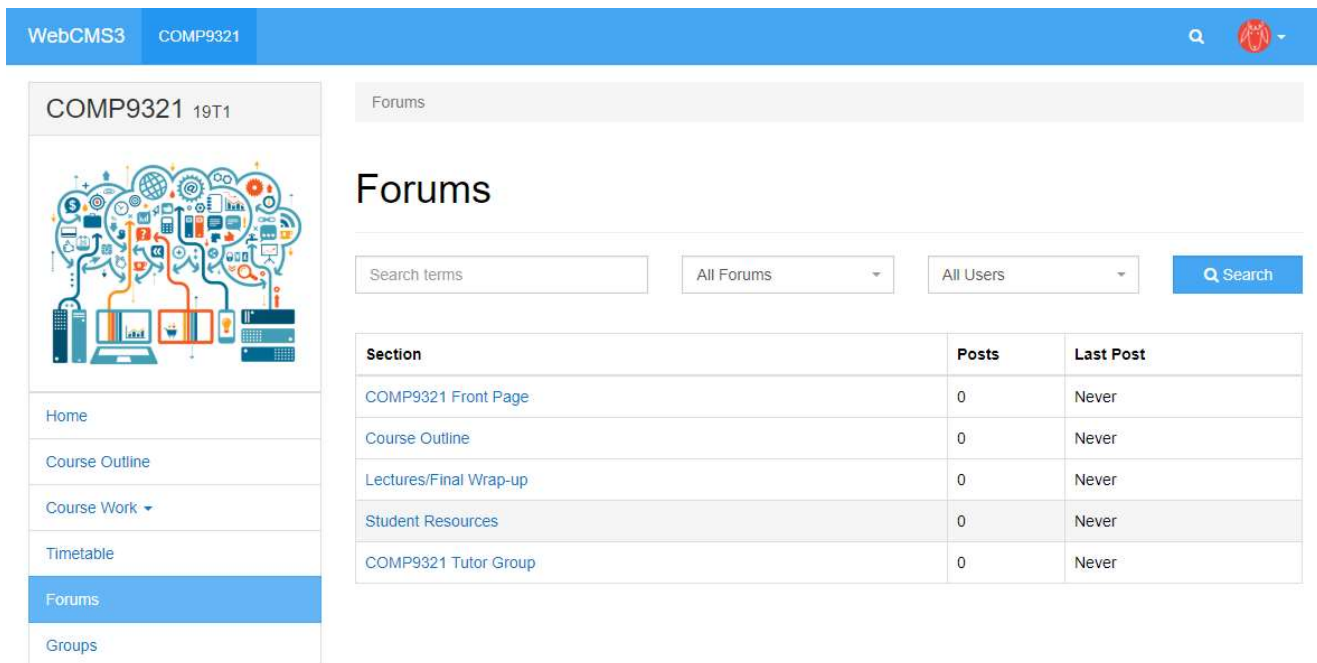 **Nice jokes are also welcomed. Penalties apply to bad ones.**

Lab consultations schedule(from Week 2 to Week 10)

Day	Time	Location	Tutor
Monday	18:00 - 20:00	Lyre	Xiaocong
Tuesday	18:00 - 20:00	Lyre	Chaoran
Wednesday	18:00 - 20:00	Lyre	Alireza
Thursday	18:00 - 20:00	Lyre	Shayan
Friday	16:00 – 18:00	Lyre	Mohammad-Ali
Friday	18:00 - 20:00	Lyre	Mohammad-Ali

Forums

If you have any questions related to the course materials, ask on the forums. We will check the forums regularly.

If you have another questions or urgent request, send email to me or Chaoran. You should know the email address.

Supplementary Exam Policy

Supp Exam is only available to students who:
- DID NOT attend the final exam

AND
- Have a good excuse for not attending

AND
- Have documentation for the excuse

Submit special consideration within 72 hours (via myUNSW with supporting docs)

Everybody gets **exactly one chance** to pass the final exam. For CSE supplementary assessment policy, follow the link in the course outline.

Plagiarism

!Important

Please check: https://student.unsw.edu.au/conduct

Plagiarism is defined as using the words or ideas of others and presenting them as your own.

UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties **as severe as being excluded from further study at UNSW**.

GitHub private repositories are now *free* and *unlimited*.

This is not a first year/semester course which means if you plagiarize, it will lead to *at least level 3 academic misconduc*t. In short, *0 marks for this course* and *recorded by university* even you drop the course.

Questions?

Ask now or email

cs9321@cse.unsw.edu.au

Introduction

COMP9321 2019T1

Sexy Job

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"

"The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."
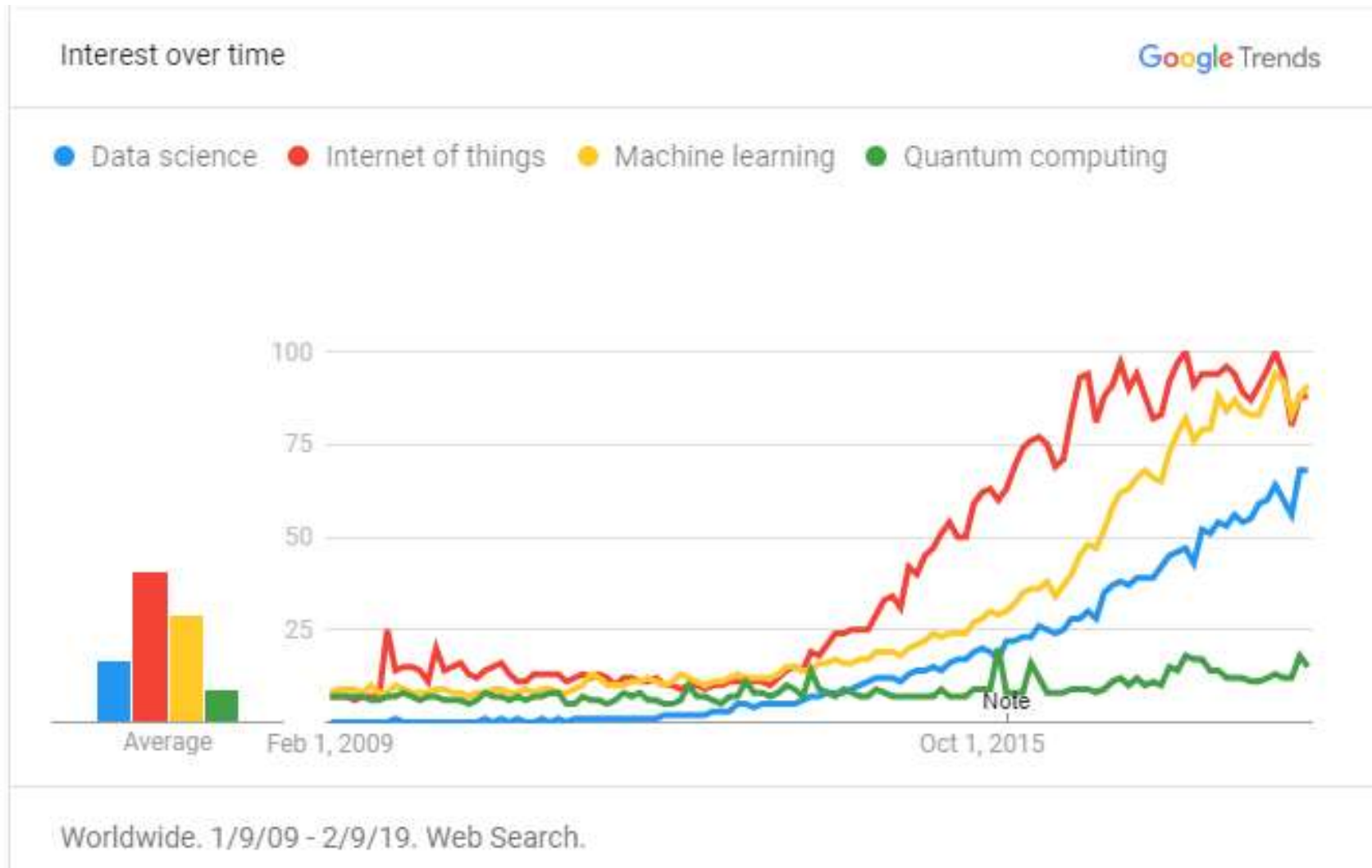
Hal Varian, Google's Chief Economist

Data is Popular
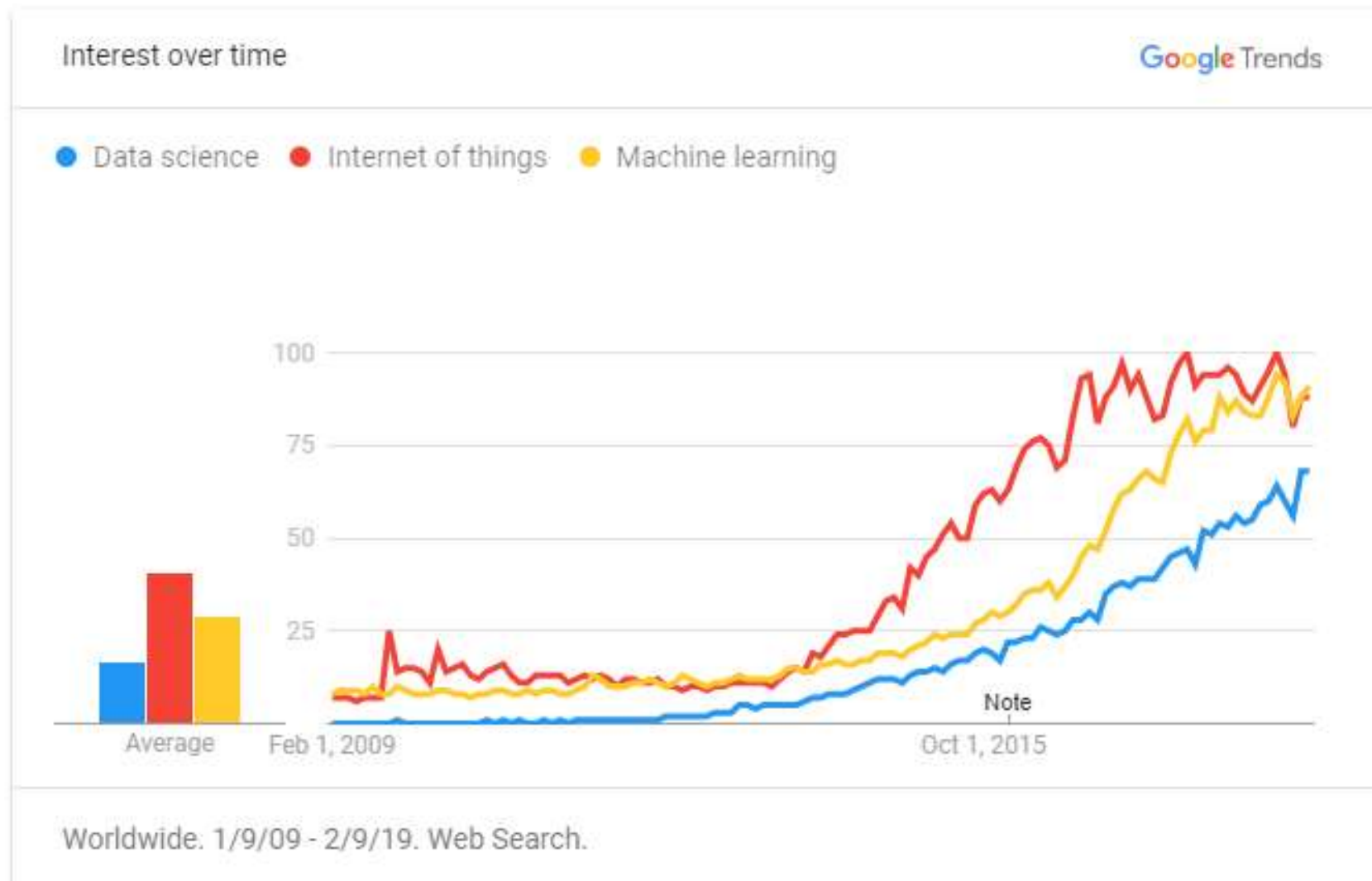
Data is Popular?
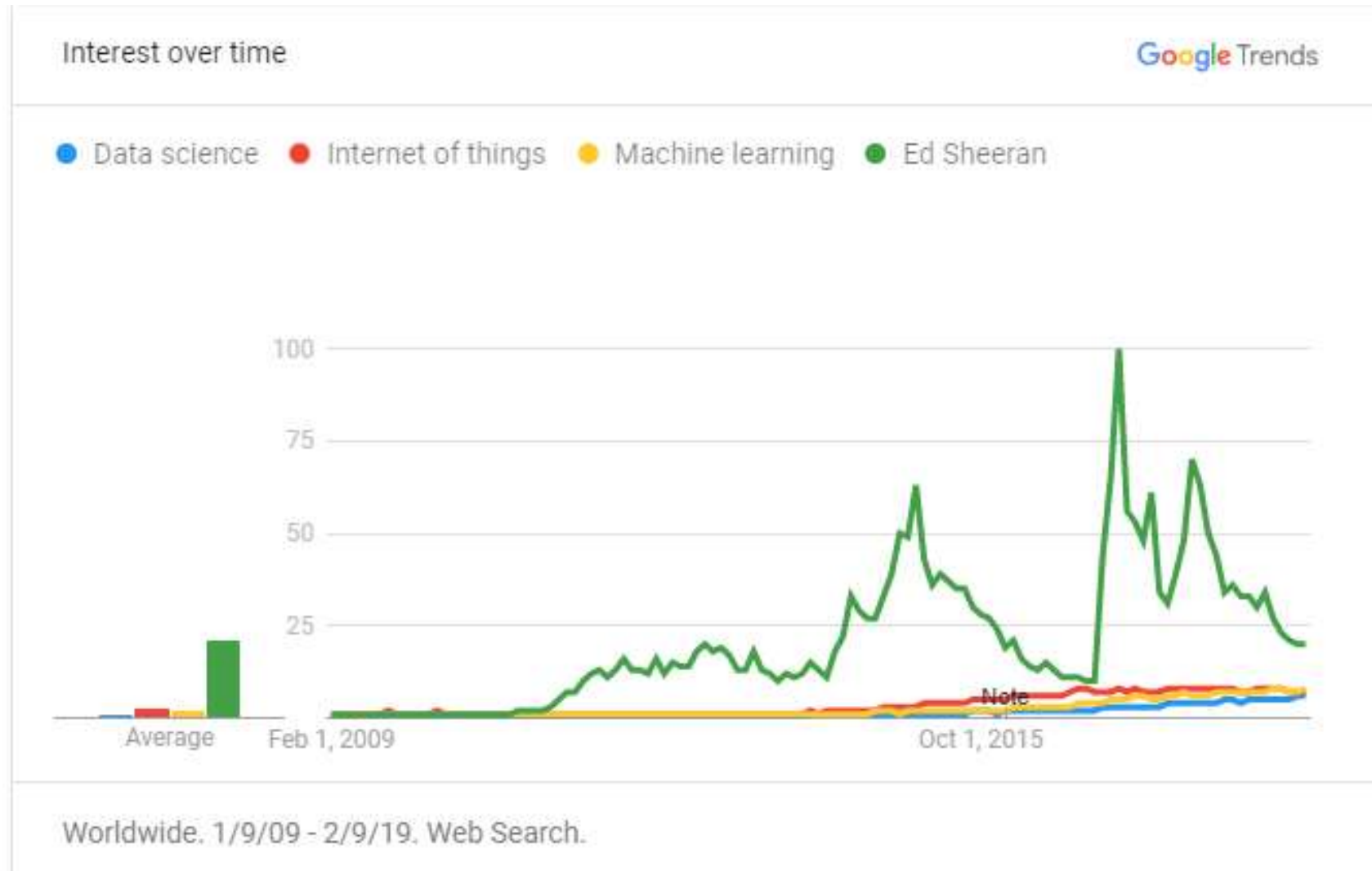
Data is Popular

Data is valuable

Extracting information

Data-driven deals, selected

	Target company (Date)	Value of deal, $bn	Business
facebook	Instagram (2012)	1.0	Photo sharing
facebook	WhatsApp (2014)	22.0	Text/photo messaging
Alphabet	Waze (2013)	1.2	Mapping and navigation
IBM	The Weather Company (2015)	2.0	Meteorology
IBM	Truven Health Analytics (2016)	2.6	Health care
intel	Mobileye (2017)	15.3	Self-driving cars
Microsoft	SwiftKey (2016)	0.25	Keyboard/artificial intelligence
Microsoft	LinkedIn (2016)	26.2	Business networking
ORACLE	BlueKai (2014)	0.4	Cloud data platform
ORACLE	Datalogix (2014)	1.0	Marketing

Source: Company reports, estimates

UNSW SYDNEY

Data is Massive

"There are **2.5 quintillion bytes (EB)** of data created **each day** at our current pace, but that pace is **only accelerating** with the growth of the Internet of Things (IoT)." [1]

» 2.7 Zetabytes of data exist in the digital universe today.[2]

» Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.[3]

» Akamai analyzes 75 million events per day to better target advertisements.[3]

» Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.[4]

» In 2008, Google was processing 20,000 terabytes of data (20 petabytes) a day.[5]

1. Forbes, *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*
2. Wikibon, *The Rapid Growth in Unstructured Data*
3. Wikibon, *Taming Big Data*
4. SAS, *Big Data Meets Big Data Analytics*
5. TechCrunch, *Google Processing 20,000 Terabytes A Day, And Growing*

UNSW
SYDNEY
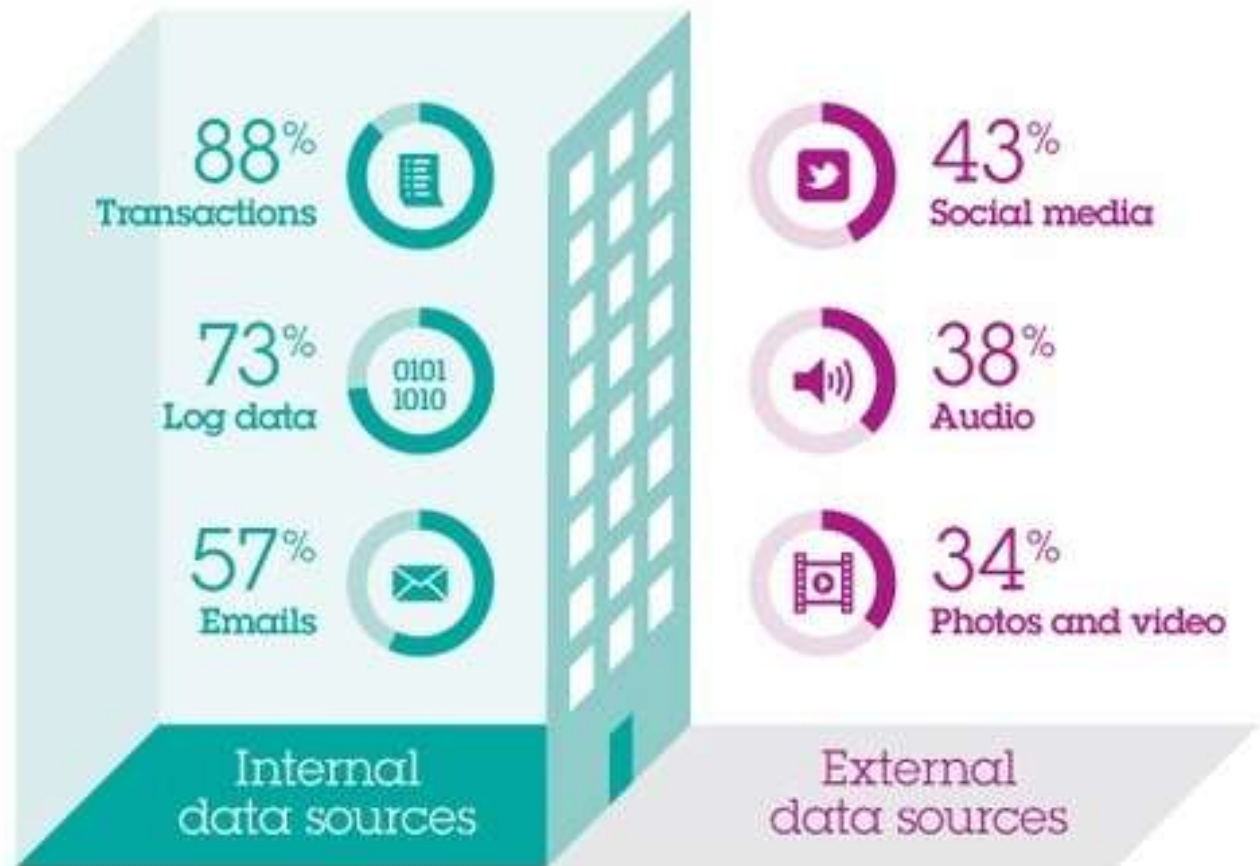
Where does data come from?

Where does data come from?
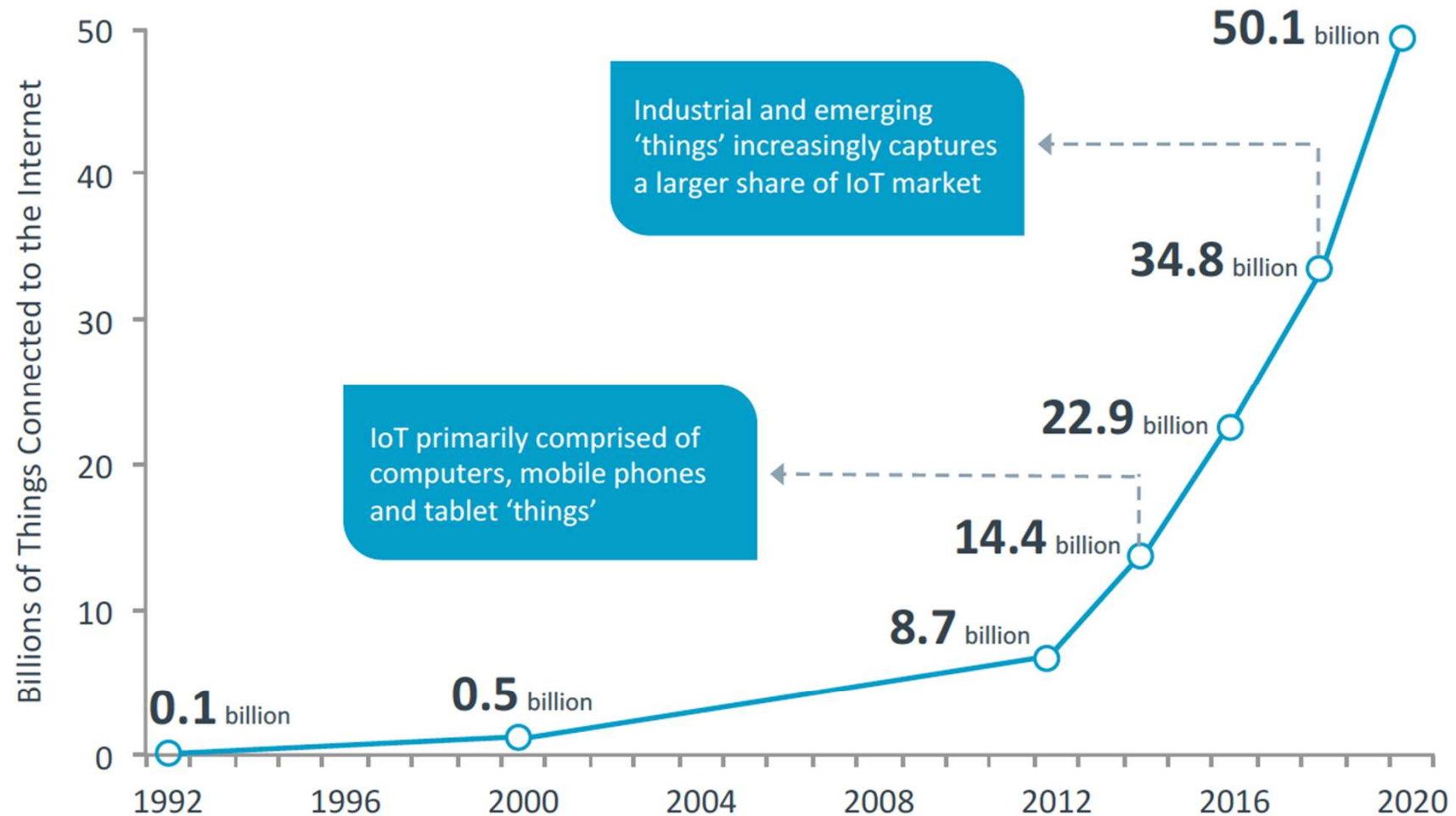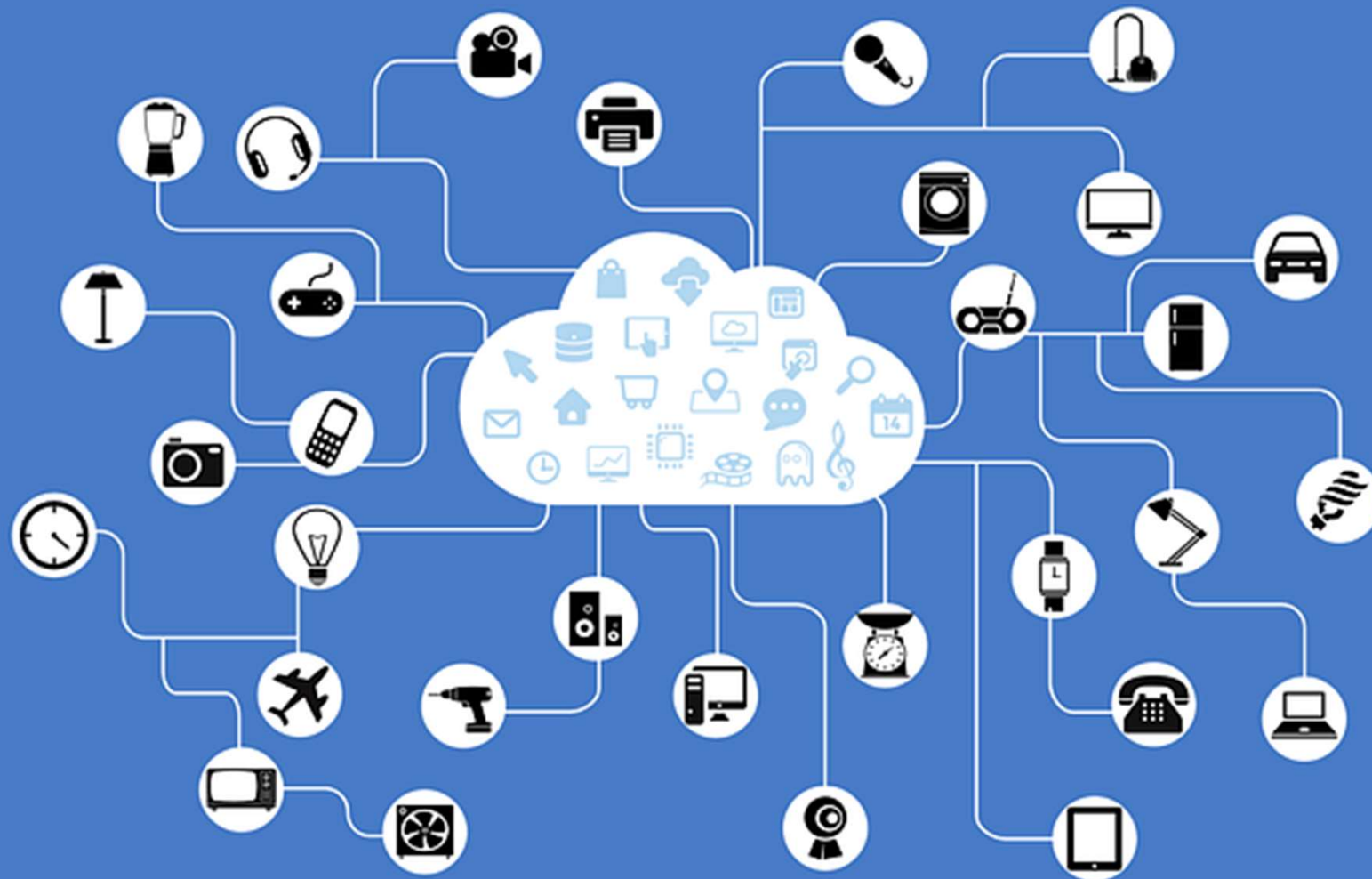
Projecting the 'Things' Behind the Internet of Things

From 2014-2020, IoT grows at an annual compound rate of 23.1% CAGR

Billions of Things Connected to the Internet

50.1 billion

Industrial and emerging 'things' increasingly captures a larger share of IoT market

34.8 billion

22.9 billion

IoT primarily comprised of computers, mobile phones and tablet 'things'

14.4 billion

8.7 billion

0.1 billion

0.5 billion

1992   1996   2000   2004   2008   2012   2016   2020

The connected world

UNSW
SYDNEY

Even more data…
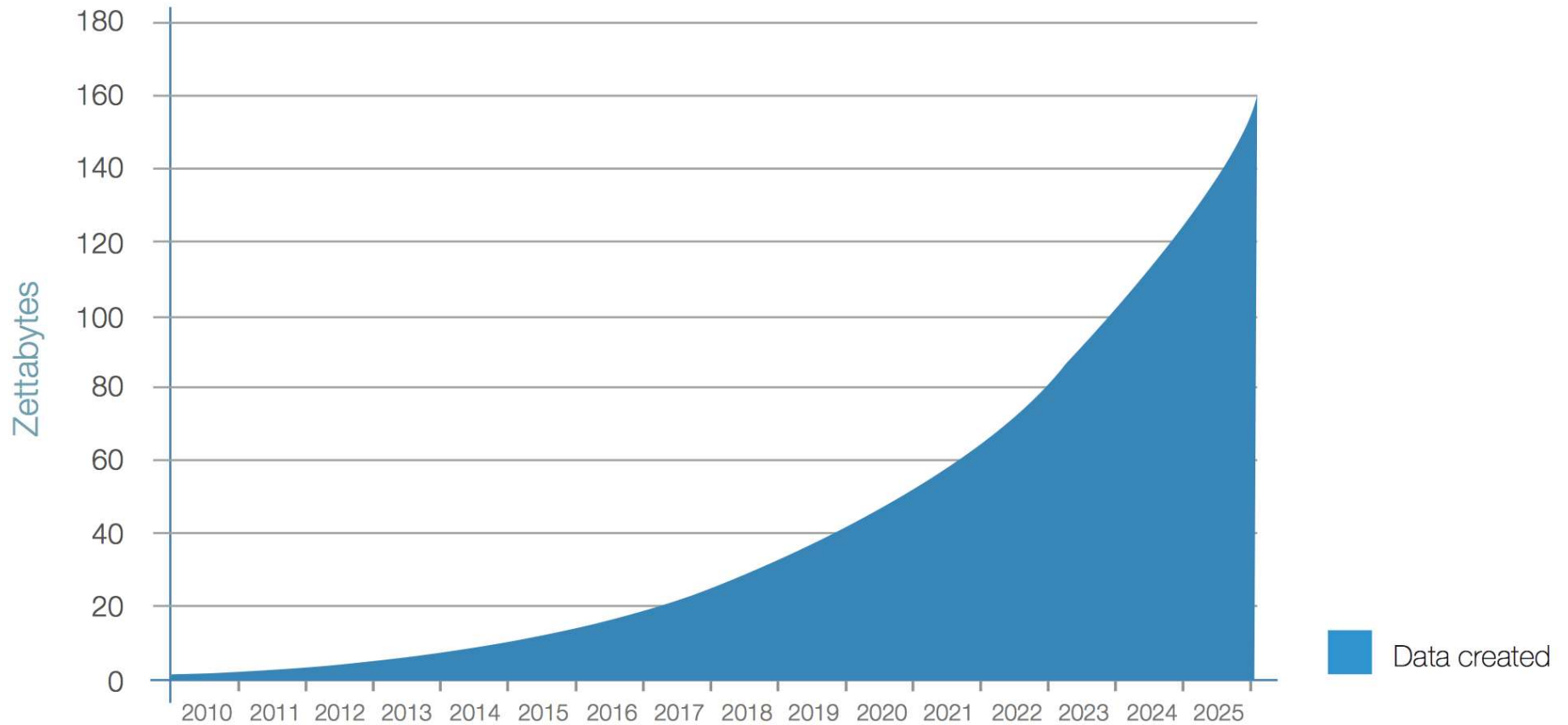
Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

What can we do with the data?

Life Sciences

Clinical research is a slow and expensive process, with trials failing for a variety of reasons. Advanced analytics, artificial intelligence (AI) and the Internet of Medical Things (IoMT) unlocks the potential of improving speed and efficiency at every stage of clinical research by delivering more intelligent, automated solutions.

Banking

Financial institutions gather and access analytical insight from large volumes of unstructured data in order to make sound financial decisions. Big data analytics allows them to access the information they need when they need it, by eliminating overlapping, redundant tools and systems.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

UNSW SYDNEY

What can we do with the data?

Manufacturing

For manufacturers, solving problems is nothing new. They wrestle with difficult problems on a daily basis - from complex supply chains, to motion applications, to labor constraints and equipment breakdowns. That's why big data analytics is essential in the manufacturing industry, as it has allowed competitive organizations to discover new cost saving opportunities and revenue opportunities.

Health Care

Big data is a given in the health care industry. Patient records, health plans, insurance information and other types of information can be difficult to manage – but are full of key insights once analytics are applied. That's why big data analytics technology is so important to heath care. By analyzing large amounts of information – both structured and unstructured – quickly, health care providers can provide lifesaving diagnoses or treatment options almost immediately.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

What can we do with the data?

Government

Certain government agencies face a big challenge: tighten the budget without compromising quality or productivity. This is particularly troublesome with law enforcement agencies, which are struggling to keep crime rates down with relatively scarce resources. And that's why many agencies use big data analytics; the technology streamlines operations while giving the agency a more holistic view of criminal activity.

Retail

Customer service has evolved in the past several years, as savvier shoppers expect retailers to understand exactly what they need, when they need it. Big data analytics technology helps retailers meet those demands. Armed with endless amounts of data from customer loyalty programs, buying habits and other sources, retailers not only have an in-depth understanding of their customers, they can also predict trends, recommend new products – and boost profitability.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

UNSW
SYDNEY

Also

Spam/False Information Detection
Credit card fraud detection
Recommendation systems
Human activity recognition/prediction
Machine translation
Face/Scene recognition
Image caption
Self-driving cars

a cat is sitting on a toilet seat
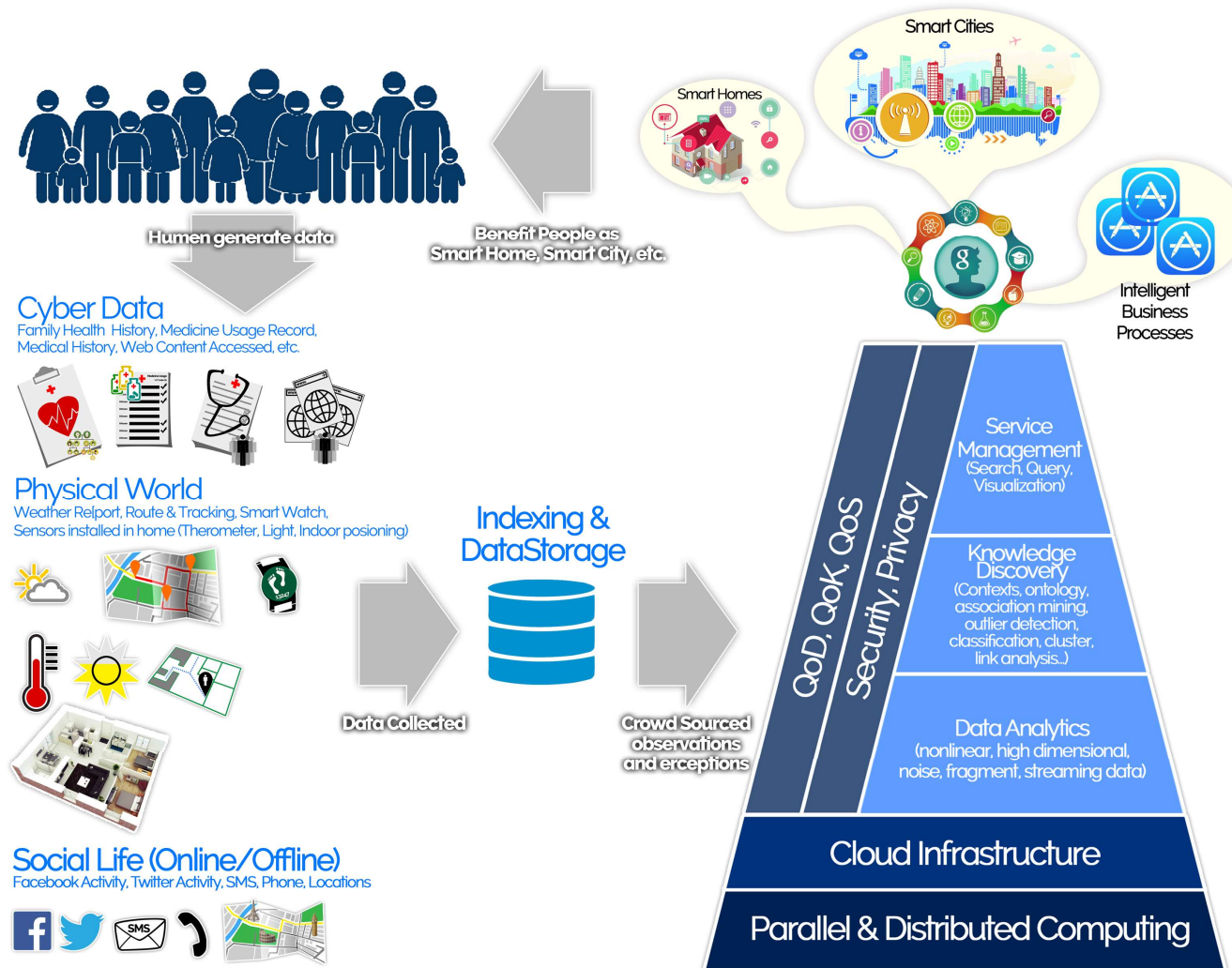logprob: -7.79
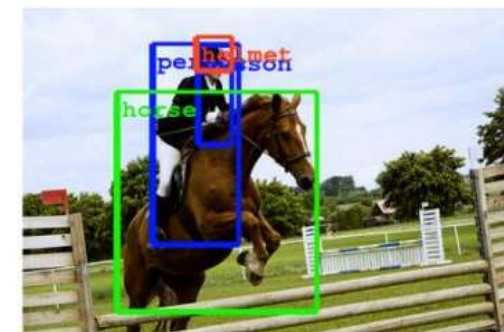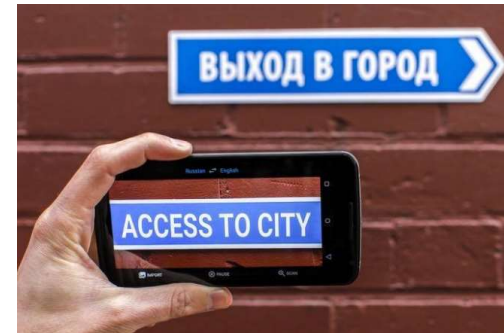
Unraveling Power of Deeply Connected World

- Produce a treasure trove of big data
    - » data that can help cities predict accidents and crimes
- Give doctors real-time insight into information from pacemakers or biochips
    - » enable optimized productivity across industries through predictive maintenance on equipment and machinery
- Create true smart homes with connected appliances
- Provide critical communication between self-driving cars
- …

UNSW SYDNEY

Artificial Intelligence (AI)

A typical AI perceives its environment and takes actions that maximize its chance of successfully achieving its goals. The overall research goal of AI is to create technology that allows computers and machines to function in an intelligent manner.

Classical AI research focus on the knowledge representation and knowledge engineering. Learning is a fundamental concept of AI research, it aims to let AI gain 'know' how to achieve its goals.

AI-powered IoT Analytics

Smart Cities

Smart Homes

Human generate data

Benefit People as Smart Home, Smart City, etc.

Intelligent Business Processes

Cyber Data
Family Health History, Medicine Usage Record, Medical History, Web Content Accessed, etc.

Physical World
Weather Relport, Route & Tracking, Smart Watch, Sensors installed in home (Therometer, Light, Indoor posioning)

Indexing & DataStorage

Data Collected

Crowd Sourced observations and erceptions

Social Life (Online/Offline)
Facebook Activity, Twitter Activity, SMS, Phone, Locations

QoD, QoK, QoS Security, Privacy

Service Management
(Search, Query, Visualization)

Knowledge Discovery
(Contexts, ontology, association mining, outlier detection, classification, cluster, link analysis...)

Data Analytics
(nonlinear, high dimensional, noise, fragment, streaming data)

Cloud Infrastructure

Parallel & Distributed Computing

UNSW SYDNEY

… and more

- **Diagnostic analysis**
- **Predictive analysis**
- **Find relation between unknown elements/events**
- **Prescriptive analysis**
- **Monitoring an event as it happens**

while it cannot

- **Predict a definitive future**
- **Imputation of new data source**
- **Find a creative solution to a business problem**
- **Find solution to a not so well-defined problem**
- **Data management/Simplify data for a new data source**

UNSW
SYDNEY

Looks promising…Yet how?

What do you do with all this data?

» Too much data to search through it manually or processing in traditional ways…

But there is valuable information in the data:

» How can we use it for fun, profit, and/or the greater good?

Boosting in computing power helps.

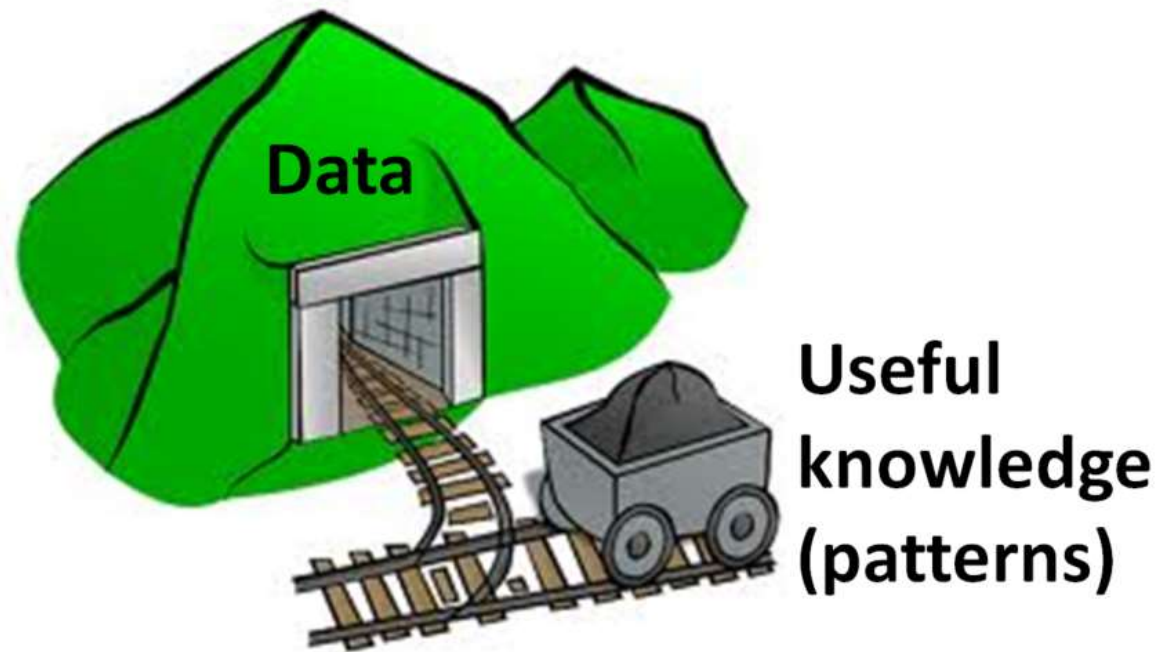» *Data mining* and *machine learning* are key tools we use to make sense of very large datasets.

Tool: Data Mining (DM)

Automatically extract useful knowledge from large datasets
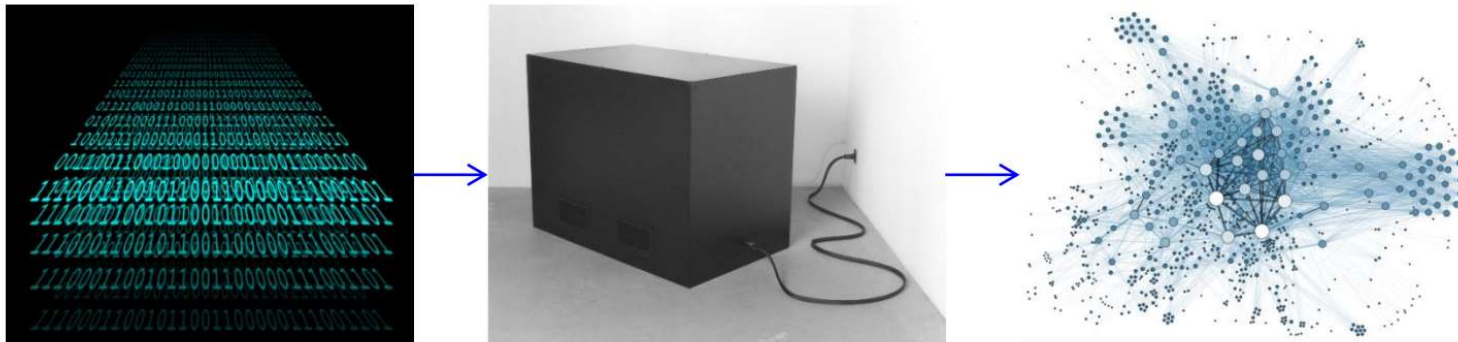
Usually, to help with human decision making

Tool: Data Mining (DM)

Data

Useful knowledge (patterns)

Tool: Machine Learning (ML)

Using computer to automatically detect patterns in data and use these to make predictions or decisions.

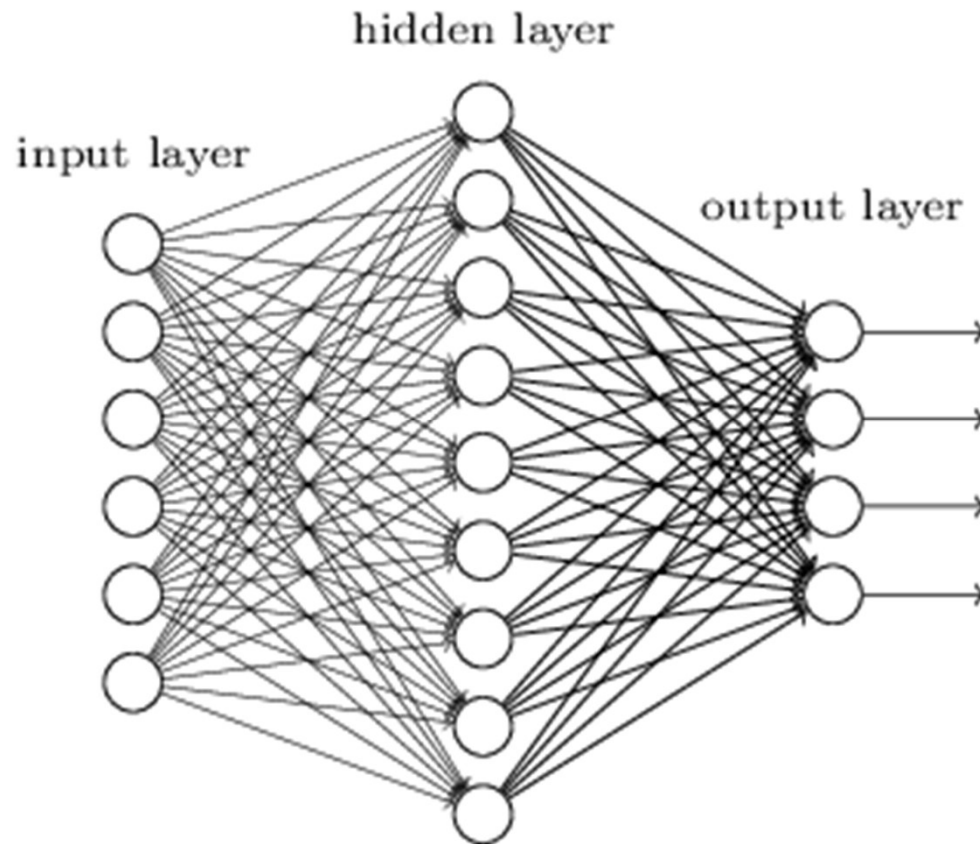Most useful when: – We want to automate something a human can do. – We want to do things a human can't do (look at 1 TB of data)

Tool: Machine Learning (ML)

From this → to this →
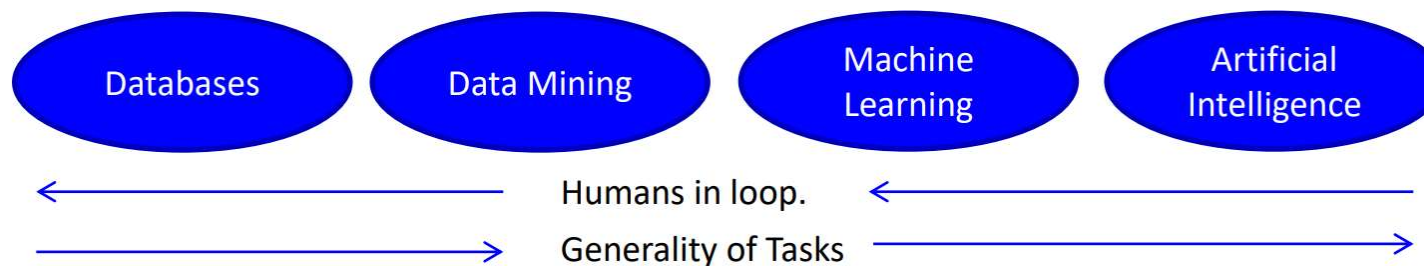
in seconds

UNSW
SYDNEY
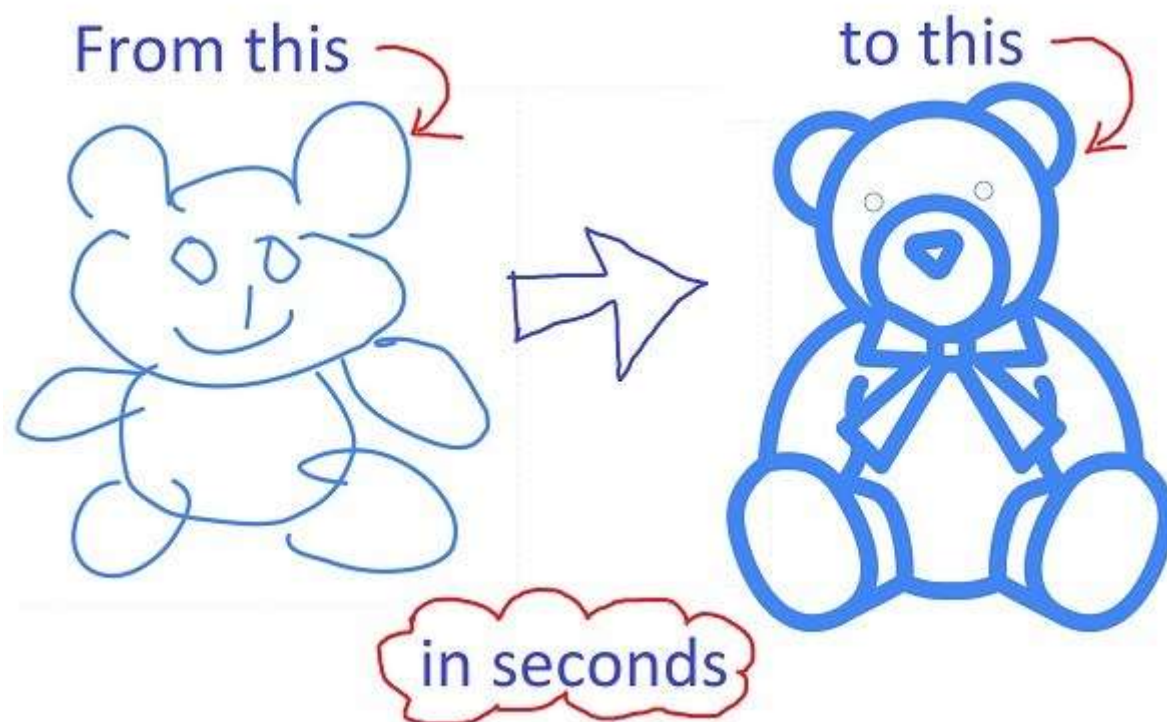
Tool: Deep Learning (DL)

Deep learning is part of machine learning. Deep learning use widely on computer vision, speech recognition, natural language processing and so on.

DM vs. ML

Data mining and machine learning are very similar:
  » Data mining often viewed as closer to databases.
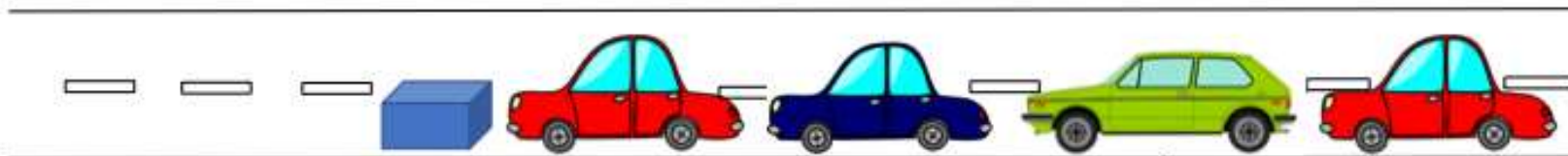 » Machine learning often viewed as closer AI.


Both are similar to statistics, but more emphasis on:
  » Large datasets and computation.
  » Predictions (instead of descriptions).
  » Flexible models (that work on many problems).

DM vs. ML

Data Mining vs. Machine Learning Illustration:

Data Mining: It takes time to batch and cluster data. If an obstacle is encountered, it cannot be dealt with in real time. It needs to wait for an analysis to be initiated by a person.
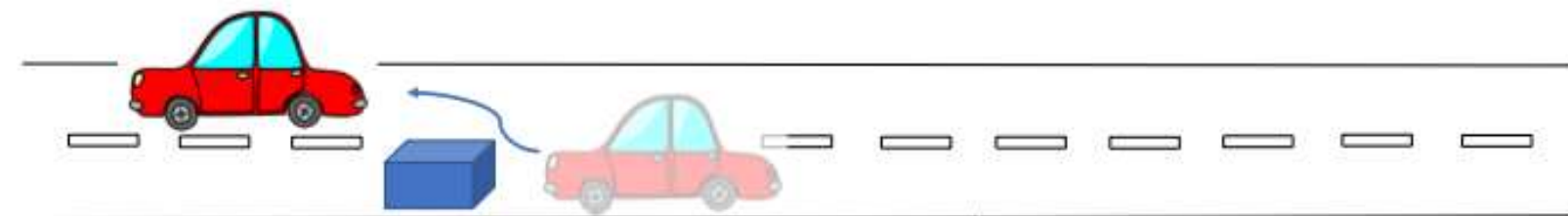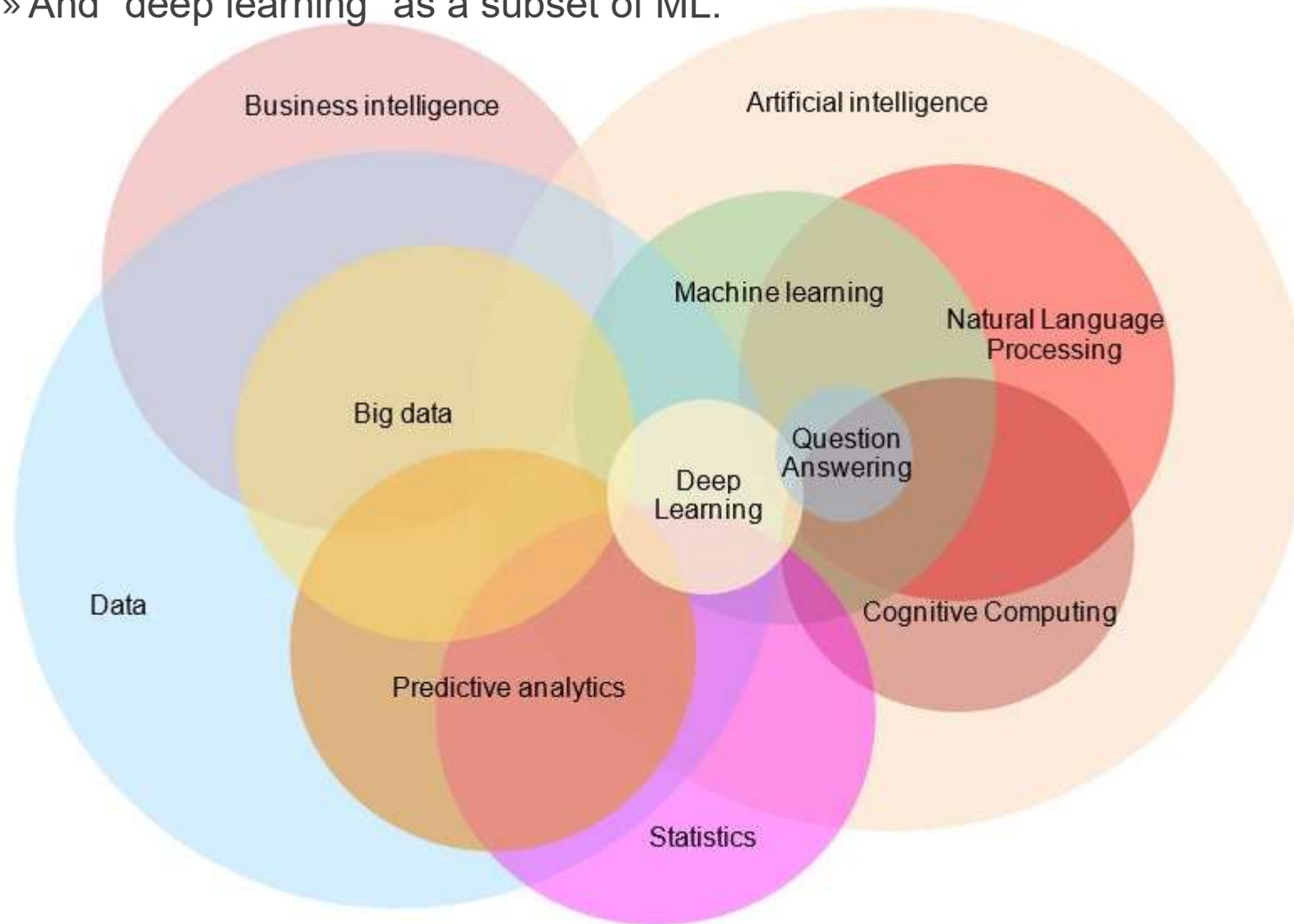
Machine Learning: The machine detects what is relevant to the task. If an obstacle is encountered, and analysis can be done inline, without a human stepping into the loop.

Deep Learning vs. ML vs. AI

Traditional we've viewed ML as a subset of AI.

» And "deep learning" as a subset of ML.

ML vs. Data Science (DS)

Machine Learning and Statistics are part of data science.

The word learning in machine learning means that the algorithms depend on some data, used as a training set, to fine-tune some model or algorithm parameters.

Data science is much more than machine learning though. Data, in data science, may or may not come from a *machine* or mechanical process (survey data could be manually collected, clinical trials involve a specific type of small data) and it might have nothing to do with *learning* as I have just discussed. But the main difference is the fact that data science covers the whole spectrum of data processing, not just the algorithmic or statistical aspects.

Data life cycle

So what is *Data Services?*

Data services are web services used to handle the programming logic for data virtualization in a cloud-hosted data storage infrastructure.

Many uses of this term involve services that are also called "data as a service" (DaaS) – these are Web-delivered services offered by cloud vendors that perform various functions on data.

The category of data services is quite broad.

  » Data services can help with the aggregation of data from various parts of an architecture, or in the creation of a central data center repository.
  » Data services may deal with data in transit, or with storage.
  » Data services could also perform various types of analytics on big data sets.

How do data services work?

When combined with data virtualization, data services provide an abstraction layer from the details of stored data.

Data services automate the work of locating heterogeneously-stored data and provide developers and data analysts with simple programmatic tools to find and extract the data they need with little effort.

In an application, data services **act as a middleware**, independently finding and delivering data the application requests.

Data services are essentially web services for data.

What are the benefits of data services?

Data services give IT more flexibility in how and where it stores data. By making it **easy to find and deliver** data from anywhere, IT can choose storage that is **cost-effective and convenient to maintain**.

For example, data services make it **feasible** for organizations to store data in the cloud or to use a hybrid cloud for data storage.

Once created, data services are **reusable**, making it possible for the organization to save a great deal of time on future development.

Since developers have fewer data-related programming tasks to complete, new IT initiatives can be **deployed rapidly**, making the organization more agile.

UNSW
SYDNEY

In this course

Week	Lectures	Labs	Assignment	Quiz
1	Course Introduction, Data access and Data Ingestion		Assignment1 Release	
2	Data Cleansing and manipulation	Access data		Quiz 1
3	Data Visualization	Data Cleaning	Assignment 1 Due, Assignment 2 Release	Quiz 2
4	Building a Data service	Data Visualization		Quiz 3
5	RESTful web Clients	Build a RESTful service	Assignment 2 Due, Assignment 3 Release.	Quiz 4
6	Data Analytics Overview	Build a RESTful client		Quiz 5
7	Data Analytics Applied Techniques and Tools	Sci-kit learn introduction		Quiz 6
8	Introduction to Recommender Systems	Classification & Clustering	Assignment 3 Due.	Quiz 7
9	Demo week	Build a Simple Rec Sys		
10	Final wrap-up	Consultation		

UNSW
SYDNEY

An Example for Data Service Engineering

How Does My Fridge Do That?

You are leaving the home (sense user)

There's no milk in fridge (sense object)

Use this information to make a decision (process)

Inform user of decision (communicate)

UNSW
SYDNEY

How Does My Fridge Do That?

You are leaving the home (sense user)

• What type of sensor?

• Distinguish between parent and child

• Identify reason for leaving home

• Identify other contexts (e.g., store hours)

There's no milk in fridge (sense object)

Use this information to make a decision (process)

Inform user of decision (notify)

How Does My Fridge Do That?

You are leaving the home (sense user)

There's no milk in fridge (sense object)

• What type of sensor?

• Is milk needed?

• No milk or "little" milk? (prediction)

Use this information to make a decision (process)

Inform user of decision (notify)

How Does My Fridge Do That?

You are leaving the home (sense user)

There's no milk in fridge (sense object)

Use this information to make a decision (process)

• Where is processor?

• What are the rules?

• Fixed rules versus dynamic rules (learning)

Inform user of decision (notify)

Data access and Data Ingestion

Data ingestion is the process of obtaining and importing data for immediate use or storage in a dataset.

Data can be streamed in *real time* or ingested *in batches*.
- » real time: data item is imported as it is emitted by the source
- » in batches: data items are imported in discrete chunks at periodic intervals of time

Challenges:

Data sources may be
- » numerous
- » massive
- » in various format

Thus difficult to
- » ingest data at a reasonable speed
- » process it efficiently

Data Cleansing and manipulation

Data cleansing is also referred as s data scrubbing

the act of detecting and removing and/or correcting *dirty data* in a dataset.

which includes data that is *incorrect*, *out-of-date*, *redundant*, *incomplete*, or *formatted incorrectly*.

After cleansing, a data set should be consistent with other similar data sets in the system.

Data cleaning **differs** from *data validation*

» Validation means data is *rejected* from the system at entry and

» Validation is performed *at the time of entry*

The aim is to ensure *data quality:*

Validity, Accuracy, Completeness, Consistency, Uniformity

UNSW
SYDNEY

Data Cleansing and manipulation

Data cleansing includes

  » Data auditing

  » Parsing

  » Data transformation

  » Duplicate elimination

  » Post-processing and controlling

Data manipulation is the process

  of changing data to make it

  easier to read or be *more organized*.

Includes:

  Updating, Adding, Removing, Sorting, Selection,
  Merging, Shifting, Aggregation … …

A "Dirty Job"

UNSW
SYDNEY

Data Visualization

Data visualization refers to the techniques used to
   communicate data by encoding it as *visual* objects.

"main goal of data visualization is to *communicate information clearly and effectively* through graphical means."

<div align="right">*Vitaly Friedman (2008)*</div>

allows us visual access to huge amounts of data

help understand *trends*, *patterns*, and to make *correlations*

typically instruments for *reasoning* about *quantitative information*

RESTful API

Application Programming Interface (API) is a set of subroutine definitions, communication protocols and tools for building software. It may be for a web-based system, operating system, database system, computer hardware or software library.

RESTful API is an API that uses HTTP request to GET/PUT/POST/DELETE data. RESTful API is based on the representational state transfer(REST) technology.

Source: https://searchmicroservices.techtarget.com/definition/RESTful-API
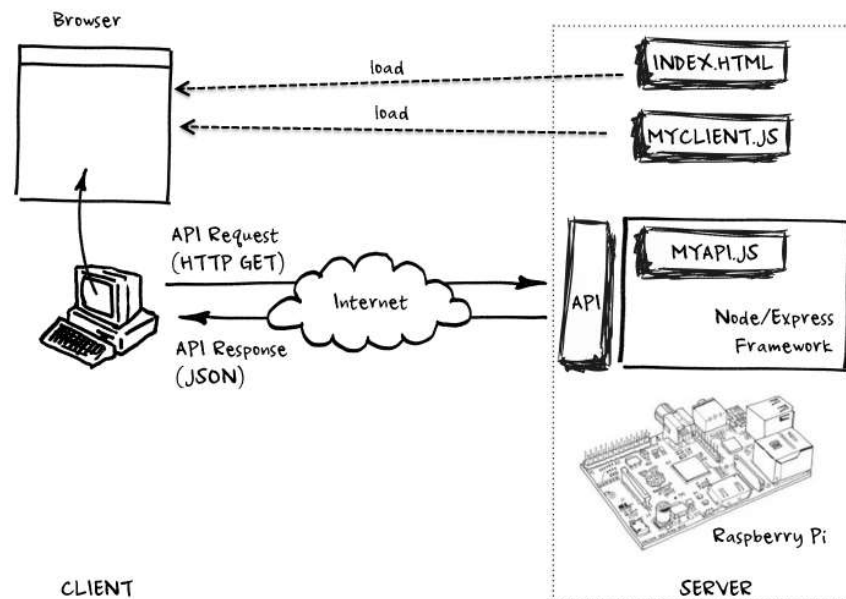
Building a Data service

Here we are talking about *RESTful web data services*

Representational State Transfer (REST) is a software architectural style that defines a set of *constraints* to be used for creating *web services,* and such services is called **RESTful Services.**

RESTful services provide interoperability between systems on the Internet

and allow the requesting systems to access and manipulate *textual representations* of web resources by using a *uniform* and *predefined* set of **stateless** operations

REST constrains include

Client–server architecture, Statelessness,

Cacheablilty, Layered system, Uniform interface
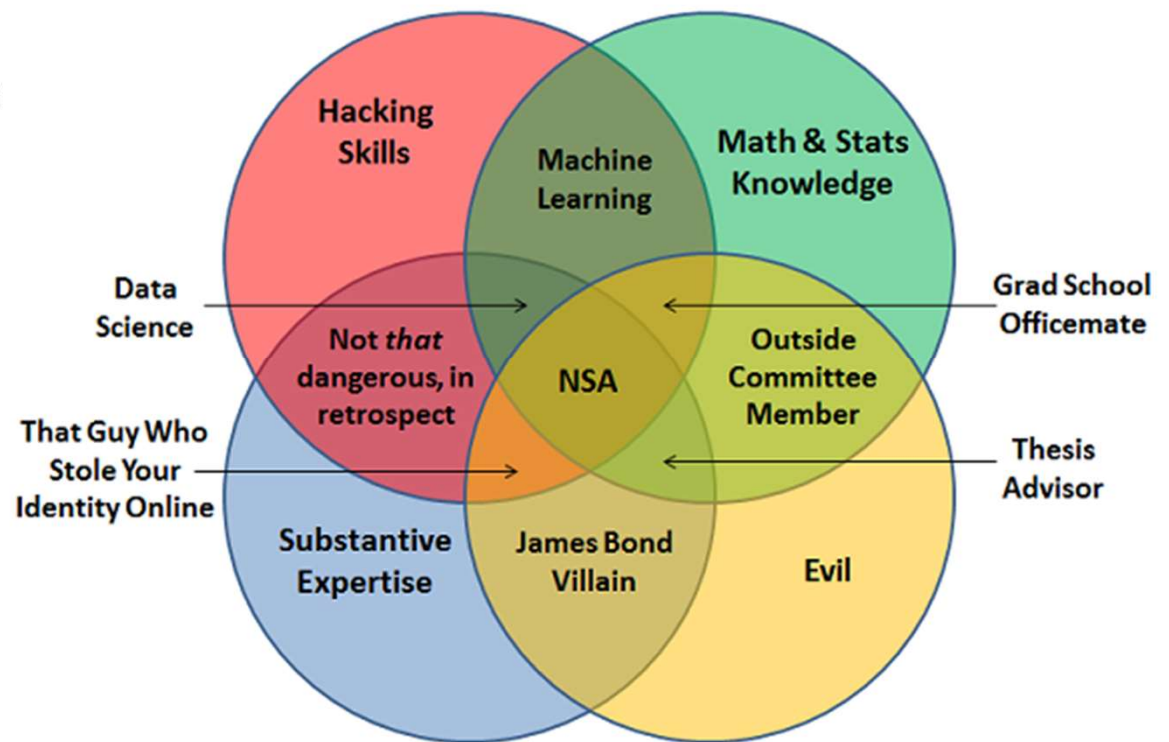
and optionally *Code on demand*

Data Analytics

Data analytics is the process of examining datasets in order to
draw conclusions about the information they contain.

includes
   » exploratory data analysis
   » confirmatory data analysis

OR
   » quantitative data analysis
   » qualitative data analysis

Data Analytics

Business Intelligence collects, integrates, analyzes data using reports and dashboards to support decision making

Advanced Analytics uses sophisticated techniques to discover insights, make predictions and generate recommendations using data/text mining, deep learning/neural networks, machine learning, reinforcement learning and artificial intelligence

Data Integration

Ingests, transforms, integrates and delivers structured data to a scalable data warehouse platform

Data Engineering

Develops and maintains large-scale data processing systems for preparing structured and unstructured data for analytic modeling

Data Science

Builds analytic models that determine strength of patterns and relationships, quantifies cause-and-effect and measures model goodness of fit

UNSW SYDNEY

Data Analytics

In this course, we cover

» Basic concepts of machine learning and data mining

» Regression

» Basic artificial neural networks

» Prediction

» Classification

» Clustering

Introduction to Recommender Systems

Recommender Systems are software tools and techniques providing suggestions for *items* to be of use to a *user*.

└─ relate to various decision-making processes

what items to buy
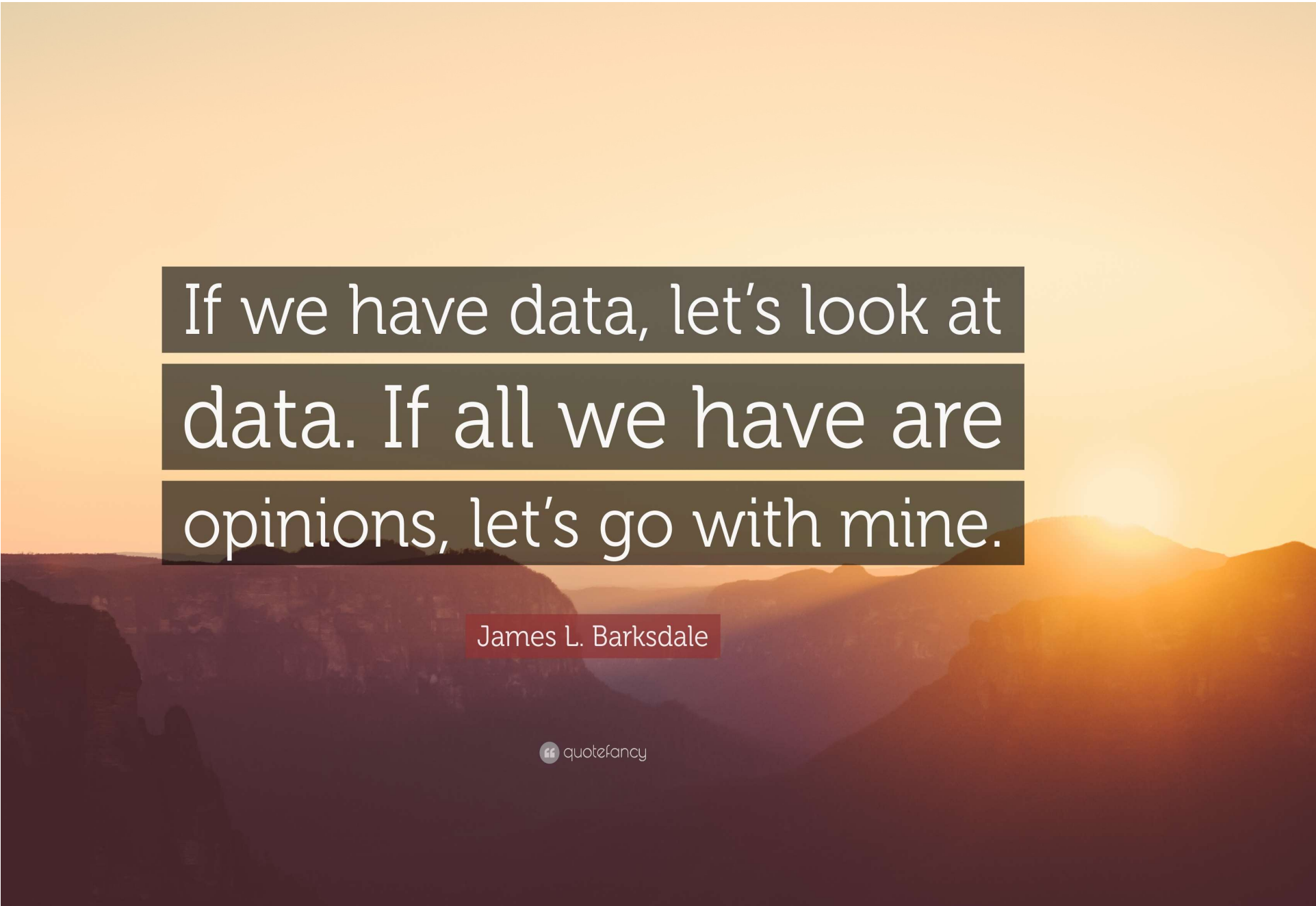
what music to listen to

what online news to read

…

Items:

» general term used to denote what the system recommends

Paradigms:

» *Collaborative Filtering*
» *Content-based Filtering*
» *Knowledge-Based Recommendations*
» *Hybridization Strategies*

UNSW
SYDNEY

Is data all we need?

> If we have data, let's look at data. If all we have are opinions, let's go with mine.

James L. Barksdale

quotefancy