CDS501: PRINCIPLES & PRACTICES OF DATA SCIENCE & ANALYTICS

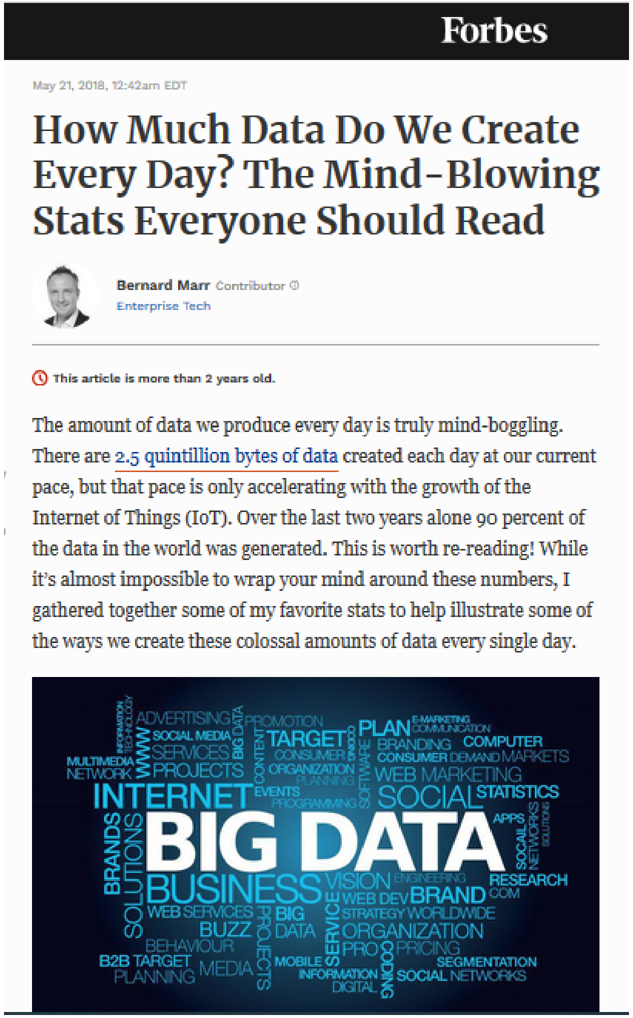
Chapter 1: Data Science Process

Outline

- Introduction
- What is Data Science?
- Applications of Data Science
- Stages of Data Science Project

How Much Data Do We Create Every Day?

Introduction



Forbes

May 21, 2018, 12:42am EDT

How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read

Bernard Marr Contributor
Enterprise Tech

This article is more than 2 years old.

The amount of data we produce every day is truly mind-boggling. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone 90 percent of the data in the world was generated. This is worth re-reading! While it's almost impossible to wrap your mind around these numbers, I gathered together some of my favorite stats to help illustrate some of the ways we create these colossal amounts of data every single day.

306.4 billion emails are sent everyday.

500 million Tweets are made everyday.

95 million photos and **videos** are shared every day on Instagram.

1.7MB of data is created every second by every person during 2020.

2.5 quintillion bytes of data are produced by humans every day.

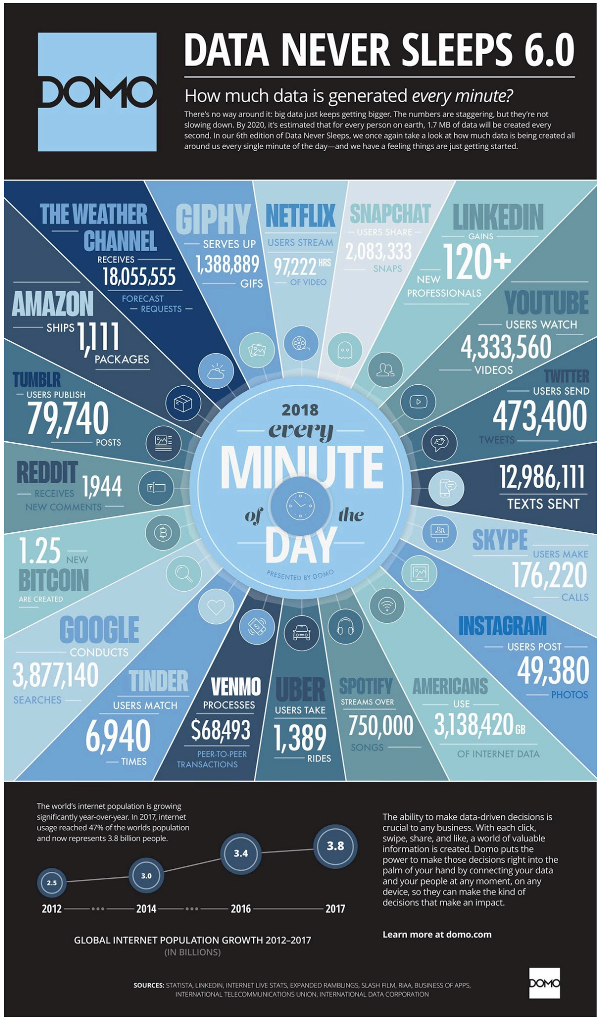
Introduction

1 quintillion = 1 000 000 000 000 000 000

A million million million

2.5 quintillion = 2 500 000 000 000 000 000

Introduction


DATA NEVER SLEEPS 6.0

How much data is generated *every minute?*

There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.

DOMO
THE WEATHER CHANNEL RECEIVES 18,055,555 FORECAST REQUESTS
GIPHY SERVES UP 1,388,889 GIFS
NETFLIX USERS STREAM 97,222 HRS OF VIDEO
SNAPCHAT USERS SHARE 2,083,333 SNAPS
LINKEDIN GAINS 120+ NEW PROFESSIONALS
YOUTUBE USERS WATCH 4,333,560 VIDEOS
TWITTER USERS SEND 473,400 TWEETS
12,986,111 TEXTS SENT
SKYPE USERS MAKE 176,220 CALLS
INSTAGRAM USERS POST 49,380 PHOTOS
AMERICANS USE 3,138,420 GB OF INTERNET DATA
SPOTIFY STREAMS OVER 750,000 SONGS
UBER USERS TAKE 1,389 RIDES
VENMO PROCESSES $68,493 PEER-TO-PEER TRANSACTIONS
TINDER USERS MATCH 6,940 TIMES
GOOGLE CONDUCTS 3,877,140 SEARCHES
1.25 NEW BITCOIN ARE CREATED
REDDIT RECEIVES 1944 NEW COMMENTS
TUMBLR USERS PUBLISH 79,740 POSTS
AMAZON SHIPS 1,111 PACKAGES
2018 every MINUTE of the DAY
PRESENTED BY DOMO
2.5
3.0
3.4
3.8
2012
2014
2016
2017

The world's internet population is growing significantly year-over-year. In 2017, internet usage reached 47% of the worlds population and now represents 3.8 billion people.

GLOBAL INTERNET POPULATION GROWTH 2012–2017
(IN BILLIONS)

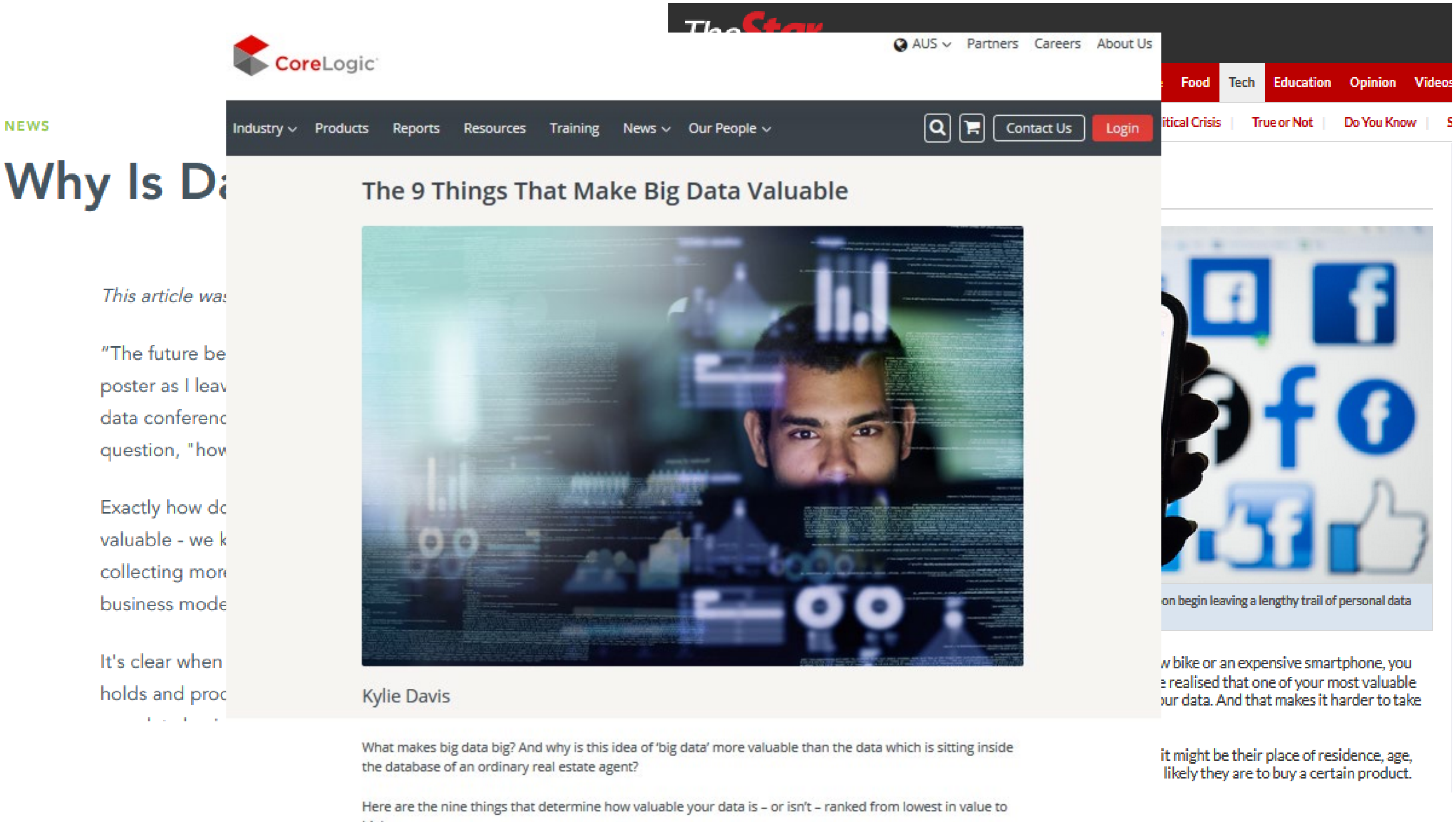
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

Learn more at domo.com

SOURCES: STATISTA, LINKEDIN, INTERNET LIVE STATS, EXPANDED RAMBLINGS, SLASH FILM, RIAA, BUSINESS OF APPS, INTERNATIONAL TELECOMMUNICATIONS UNION, INTERNATIONAL DATA CORPORATION

- Social media
- Online video sharing platform
- Media service provider
- Online retailer
- Mobile payment service

Introduction



NEWS

Why Is Da

This article was

"The future be
poster as I leav
data conferenc
question, "how

Exactly how dc
valuable - we k
collecting more
business mode

It's clear when
holds and proc

The 9 Things That Make Big Data Valuable

Kylie Davis

What makes big data big? And why is this idea of 'big data' more valuable than the data which is sitting inside the database of an ordinary real estate agent?

Here are the nine things that determine how valuable your data is – or isn't – ranked from lowest in value to

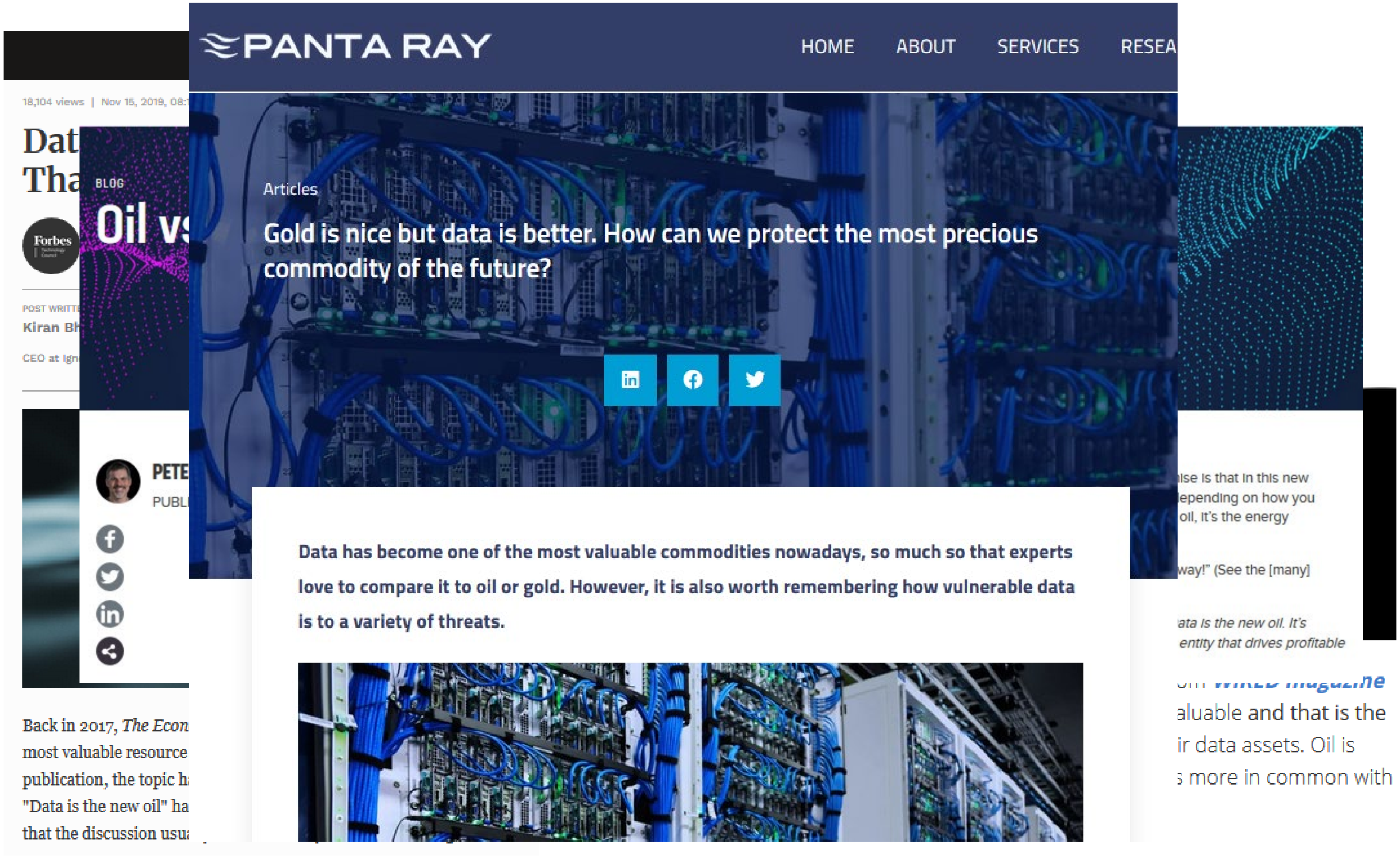
on begin leaving a lengthy trail of personal data

w bike or an expensive smartphone, you
e realised that one of your most valuable
our data. And that makes it harder to take

it might be their place of residence, age,
likely they are to buy a certain product.

Data is valuable

Introduction



18,104 views | Nov 15, 2019, 08:

Dat
Tha

Forbes

POST WRITTE
Kiran Bh
CEO at Ign

Back in 2017, *The Econ*
most valuable resource
publication, the topic h
"Data is the new oil" ha
that the discussion usu

BLOG

Oil vs

PETE
PUBL

PANTA RAY

Articles

Gold is nice but data is better. How can we protect the most precious commodity of the future?

Data has become one of the most valuable commodities nowadays, so much so that experts love to compare it to oil or gold. However, it is also worth remembering how vulnerable data is to a variety of threats.

ise is that in this new
depending on how you
oil, it's the energy

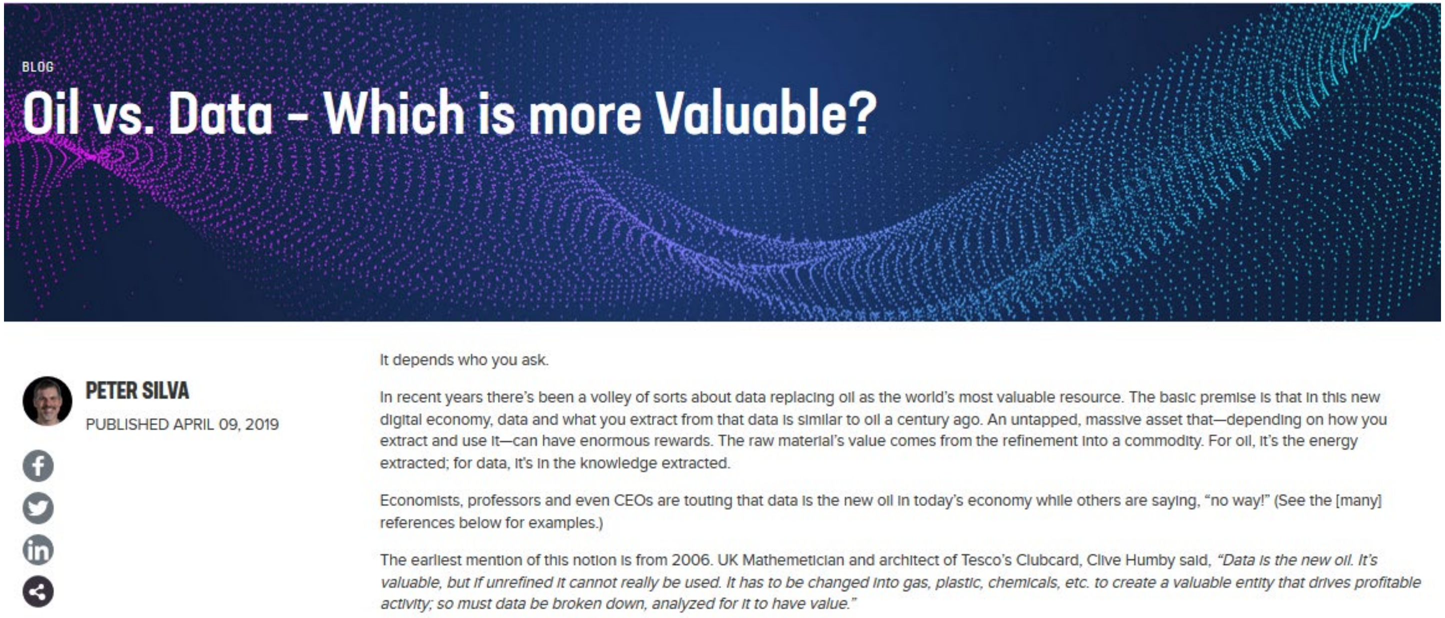
way!" (See the [many]

ata is the new oil. It's
entity that drives profitable

WIRED magazine
aluable and that is the
ir data assets. Oil is
s more in common with

Data is valuable

Introduction



BLOG

Oil vs. Data - Which is more Valuable?

PETER SILVA

PUBLISHED APRIL 09, 2019

It depends who you ask.

In recent years there's been a volley of sorts about data replacing oil as the world's most valuable resource. The basic premise is that in this new digital economy, data and what you extract from that data is similar to oil a century ago. An untapped, massive asset that—depending on how you extract and use it—can have enormous rewards. The raw material's value comes from the refinement into a commodity. For oil, it's the energy extracted; for data, it's in the knowledge extracted.

Economists, professors and even CEOs are touting that data is the new oil in today's economy while others are saying, "no way!" (See the [many] references below for examples.)

The earliest mention of this notion is from 2006. UK Mathemetician and architect of Tesco's Clubcard, Clive Humby said, *"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."*

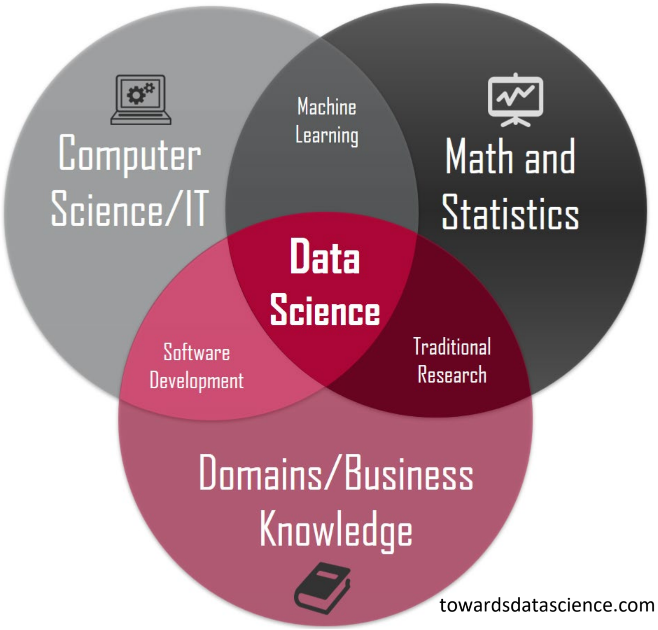
Data is unrefined and cannot really be used

Data needs to be changed into useful form

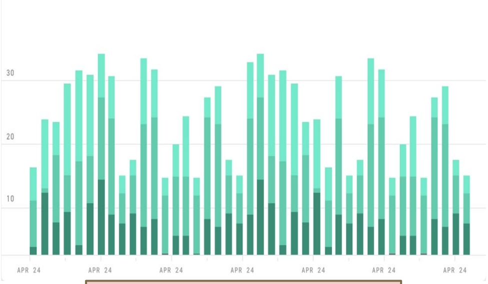
Manage, Analyze and Model

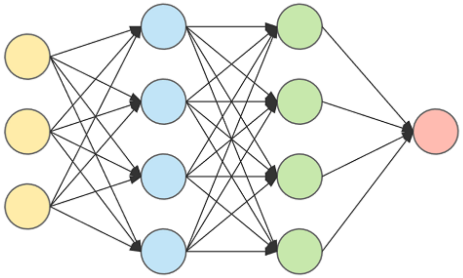
What is Data Science?

- Data science is a process of extracting knowledge and insight, and transforming hypothesis to actionable predictions from a huge and diverse set of data through managing, analyzing and modelling using various statistical and computing methods



towardsdatascience.com





Visualization

Graphical representation of data to understand and communicate information

Data Science

Machine Learning

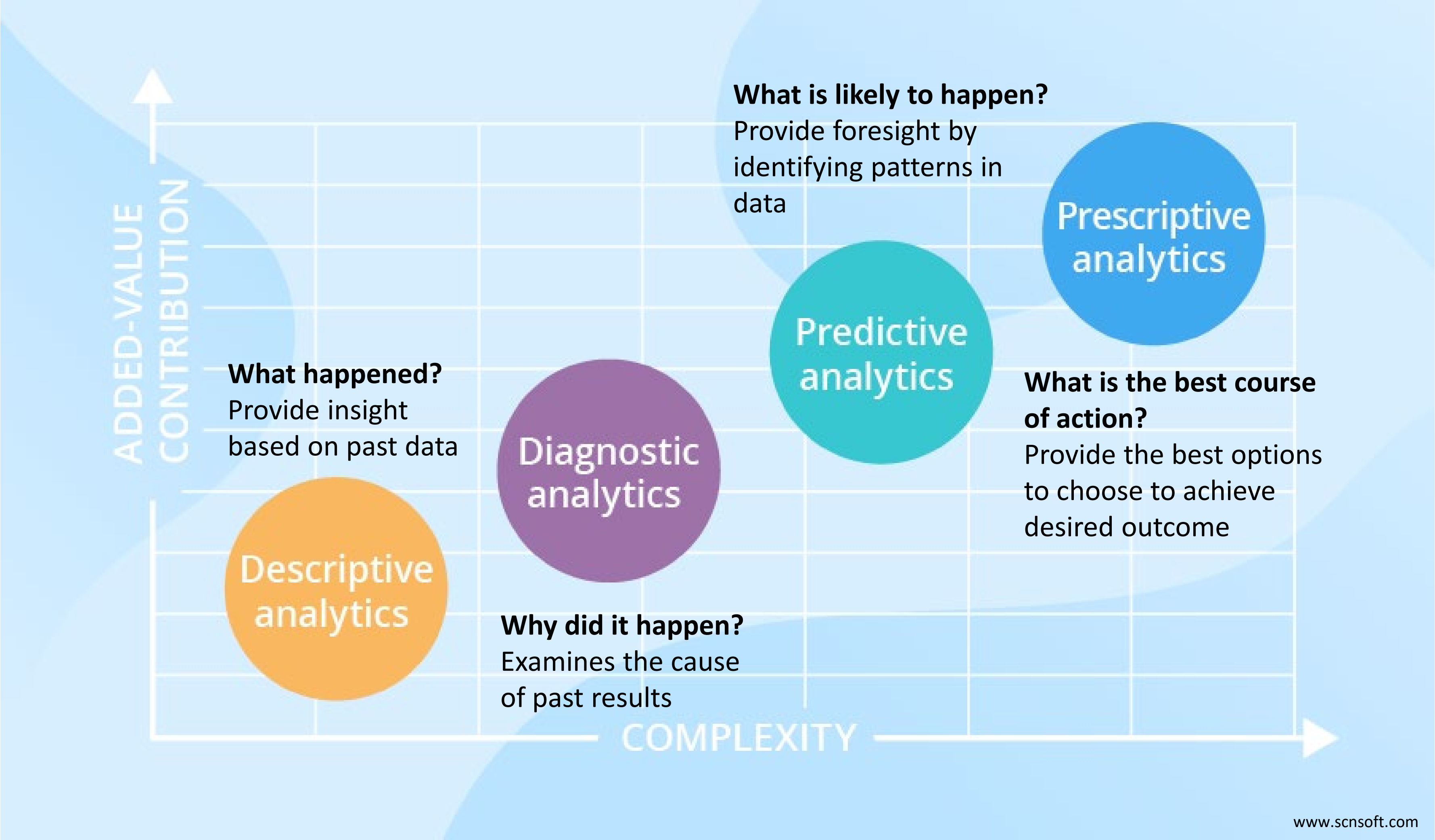
Subset of AI to learn and identify patterns in data without human assistance

Statistics

$$s^2 = \frac{\sum (X-\bar{X})^2}{N-1}$$

Methods of analyzing and interpreting data

Types of Data Analytic



What happened?
Provide insight based on past data

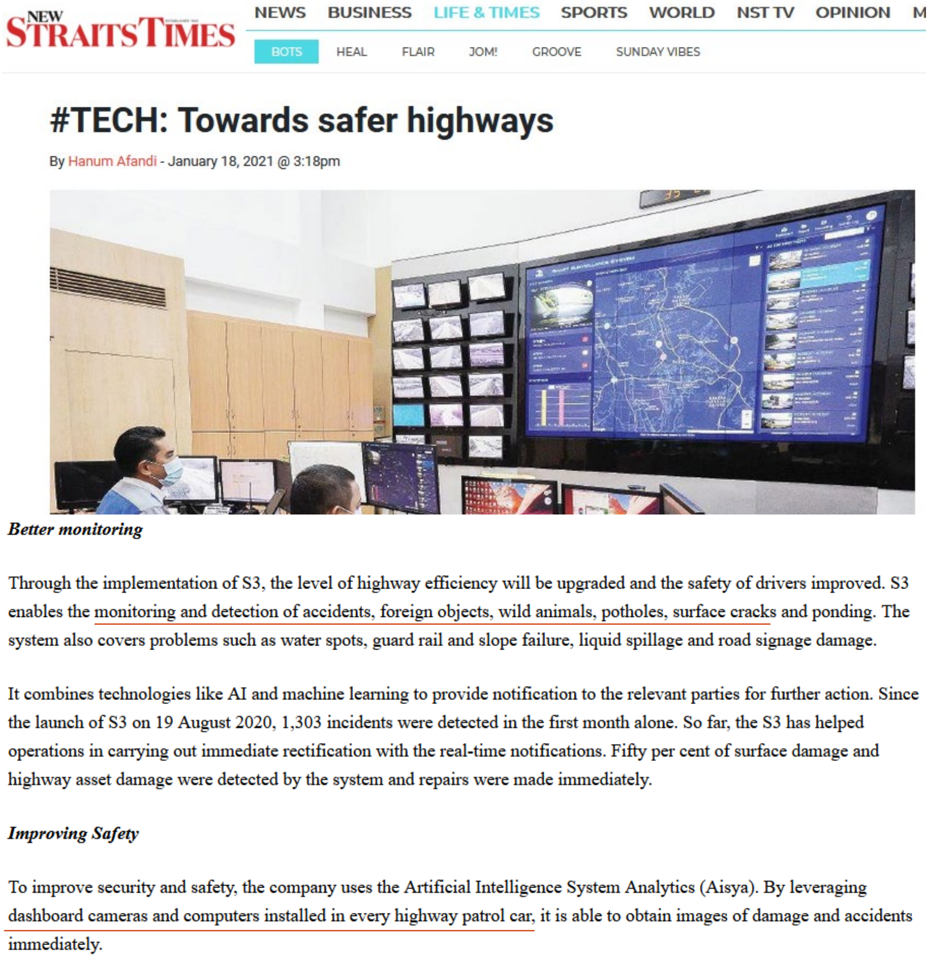
Why did it happen?
Examines the cause of past results

What is likely to happen?
Provide foresight by identifying patterns in data

What is the best course of action?
Provide the best options to choose to achieve desired outcome

www.scnsoft.com

Applications of Data Science



NEW STRAITS TIMES
NEWS BUSINESS LIFE & TIMES SPORTS WORLD NST TV OPINION M
BOTS HEAL FLAIR JOM! GROOVE SUNDAY VIBES

#TECH: Towards safer highways

By Hanum Afandi - January 18, 2021 @ 3:18pm

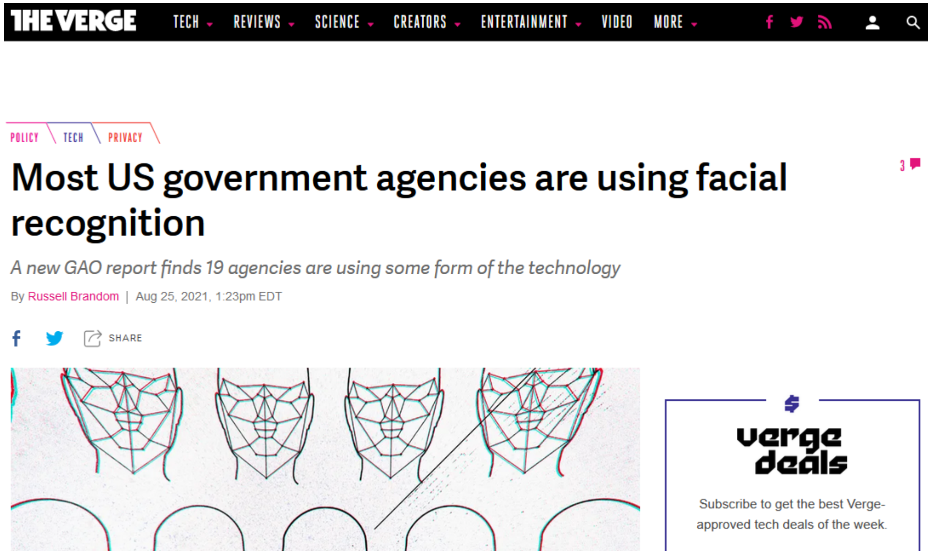
Better monitoring

Through the implementation of S3, the level of highway efficiency will be upgraded and the safety of drivers improved. S3 enables the monitoring and detection of accidents, foreign objects, wild animals, potholes, surface cracks and ponding. The system also covers problems such as water spots, guard rail and slope failure, liquid spillage and road signage damage.

It combines technologies like AI and machine learning to provide notification to the relevant parties for further action. Since the launch of S3 on 19 August 2020, 1,303 incidents were detected in the first month alone. So far, the S3 has helped operations in carrying out immediate rectification with the real-time notifications. Fifty per cent of surface damage and highway asset damage were detected by the system and repairs were made immediately.

Improving Safety

To improve security and safety, the company uses the Artificial Intelligence System Analytics (Aisya). By leveraging dashboard cameras and computers installed in every highway patrol car, it is able to obtain images of damage and accidents immediately.



THE VERGE
TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE
POLICY TECH PRIVACY
3
SHARE

Most US government agencies are using facial recognition

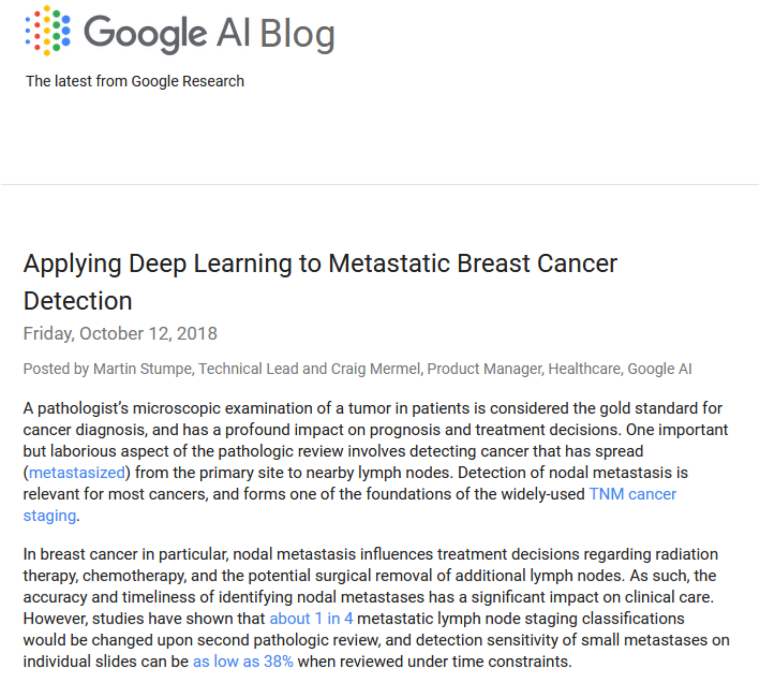
A new GAO report finds 19 agencies are using some form of the technology

By Russell Brandom | Aug 25, 2021, 1:23pm EDT

verge deals
Subscribe to get the best Verge-approved tech deals of the week.

Computer Vision

Applications of Data Science



Google AI Blog

The latest from Google Research

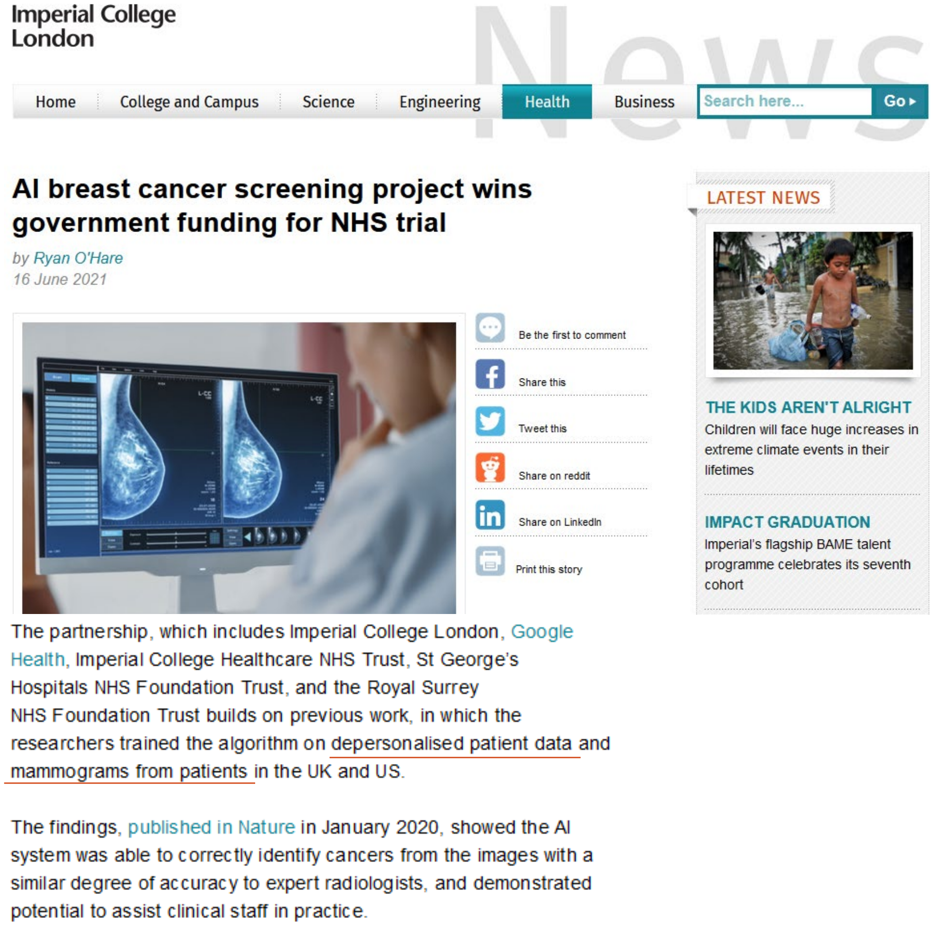
Applying Deep Learning to Metastatic Breast Cancer Detection

Friday, October 12, 2018

Posted by Martin Stumpe, Technical Lead and Craig Mermel, Product Manager, Healthcare, Google AI

A pathologist's microscopic examination of a tumor in patients is considered the gold standard for cancer diagnosis, and has a profound impact on prognosis and treatment decisions. One important but laborious aspect of the pathologic review involves detecting cancer that has spread (metastasized) from the primary site to nearby lymph nodes. Detection of nodal metastasis is relevant for most cancers, and forms one of the foundations of the widely-used TNM cancer staging.

In breast cancer in particular, nodal metastasis influences treatment decisions regarding radiation therapy, chemotherapy, and the potential surgical removal of additional lymph nodes. As such, the accuracy and timeliness of identifying nodal metastases has a significant impact on clinical care. However, studies have shown that about 1 in 4 metastatic lymph node staging classifications would be changed upon second pathologic review, and detection sensitivity of small metastases on individual slides can be as low as 38% when reviewed under time constraints.



AI breast cancer screening project wins government funding for NHS trial

by Ryan O'Hare

16 June 2021

The partnership, which includes Imperial College London, Google Health, Imperial College Healthcare NHS Trust, St George's Hospitals NHS Foundation Trust, and the Royal Surrey NHS Foundation Trust builds on previous work, in which the researchers trained the algorithm on depersonalised patient data and mammograms from patients in the UK and US.

The findings, published in Nature in January 2020, showed the AI system was able to correctly identify cancers from the images with a similar degree of accuracy to expert radiologists, and demonstrated potential to assist clinical staff in practice.

LATEST NEWS

THE KIDS AREN'T ALRIGHT

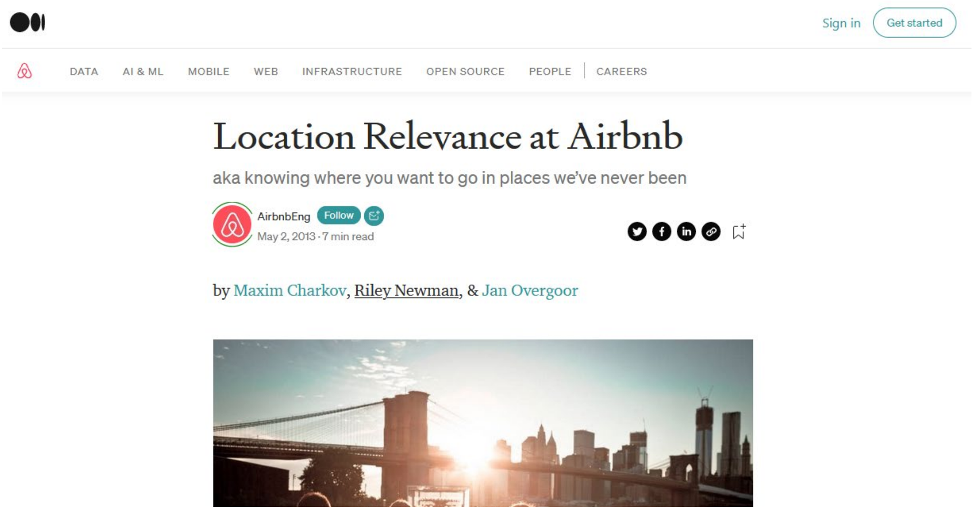
Children will face huge increases in extreme climate events in their lifetimes

IMPACT GRADUATION

Imperial's flagship BAME talent programme celebrates its seventh cohort

Healthcare Services

Applications of Data Science

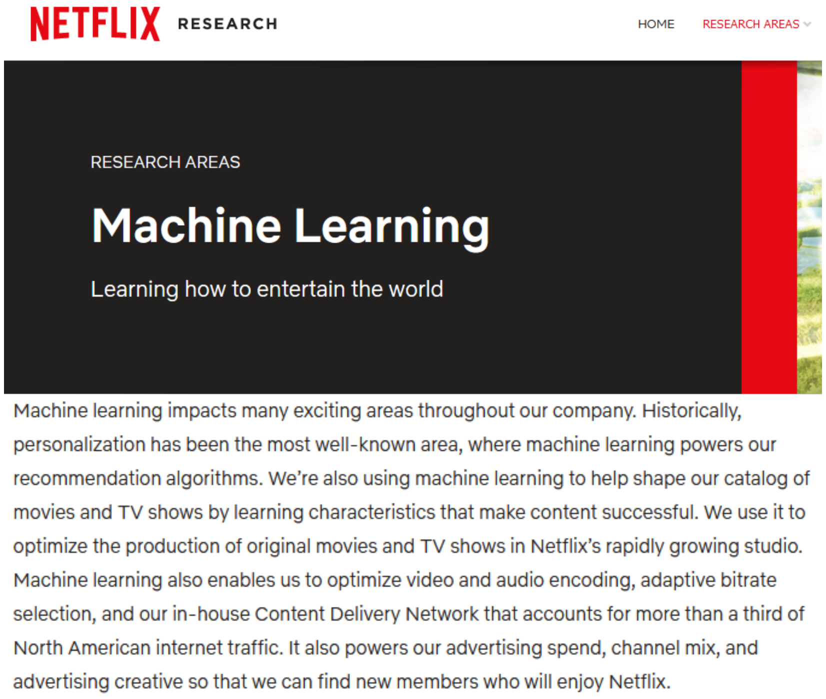


Location Relevance at Airbnb

aka knowing where you want to go in places we've never been

AirbnbEng Follow
May 2, 2013 · 7 min read

by Maxim Charkov, Riley Newman, & Jan Overgoor



RESEARCH AREAS

Machine Learning

Learning how to entertain the world

Machine learning impacts many exciting areas throughout our company. Historically, personalization has been the most well-known area, where machine learning powers our recommendation algorithms. We're also using machine learning to help shape our catalog of movies and TV shows by learning characteristics that make content successful. We use it to optimize the production of original movies and TV shows in Netflix's rapidly growing studio. Machine learning also enables us to optimize video and audio encoding, adaptive bitrate selection, and our in-house Content Delivery Network that accounts for more than a third of North American internet traffic. It also powers our advertising spend, channel mix, and advertising creative so that we can find new members who will enjoy Netflix.

Recommendation Systems

Applications of Data Science

- Is there any relationship between weather and sales?
- Is weather is influencing the sales of your shop?

Humidity

Temperature

Sales

Quantity Sold

Applications of Data Science

- Can we detect a person with depression disorder?

Social Media Activity

Person with and without Depression

Medical Data

The Roles in Data Science Projects

Project sponsor

- Person who wants the result
- Decide the project success/failure
- Keep them updated

Client

- The end user
- Has interest in your model
- Who will be using your model

Data scientist

- Design the project steps
- Apply the process
- Pick the statistic and machine learning techniques that will be used

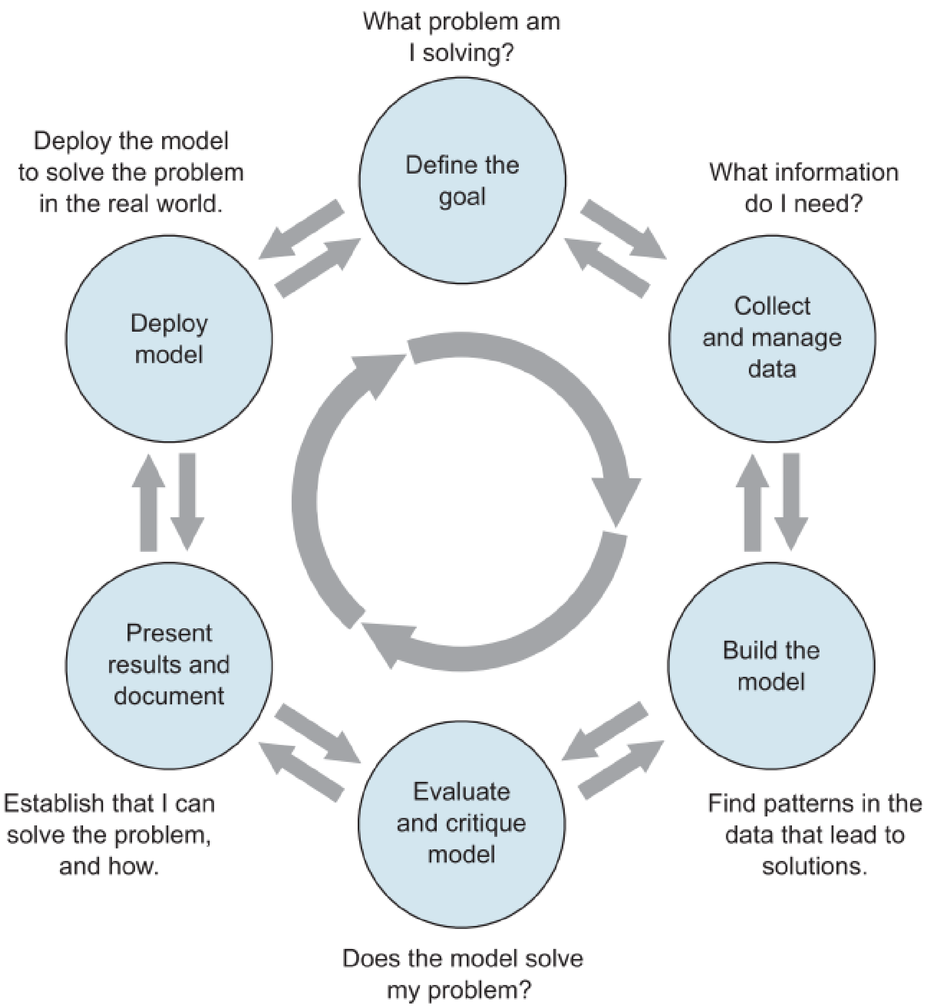
Data architect

- Responsible for data and its storage
- Manage data warehouse
- e.g. database administrator

Operations

- Responsible for acquiring data and deliver the final results
- Responsible for deployment

Stages of a Data Science Project


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Defining the Goal

- Discuss and work with stakeholders/sponsor to understand and identify business problems
- An online shop is making losses
    - Products – which products are not selling well?
    - Customers – how to identify customers who are more likely to buy?
    - Fraudulent orders – how to identify fraudulent orders?

Defining the Goal

- Define a specific, quantifiable and achievable goal
- "The detection accuracy rate must be at least 85%"

Defining the Goal

- Specific goal allows stopping condition and acceptance criteria to be defined
- Otherwise the project will go unbounded

Data Collection and Management

- The most time-consuming step in the process
- Most important and crucial step

Data Collection and Management

- Identify data that is relevant to the question
- What are the attributes that are related to the target?
- Do you have the attributes that are related to the target?

Data Collection and Management

- Number of products in the order – unusually large orders could be fraudulent
- Type of products in the order – wary of orders that are uncommonly purchased together.
- Billing/Shipping address of the order – not a real location or not a residential location
- Number of receiving orders in a timeframe – placing multiple orders at the same time

Data Collection and Management

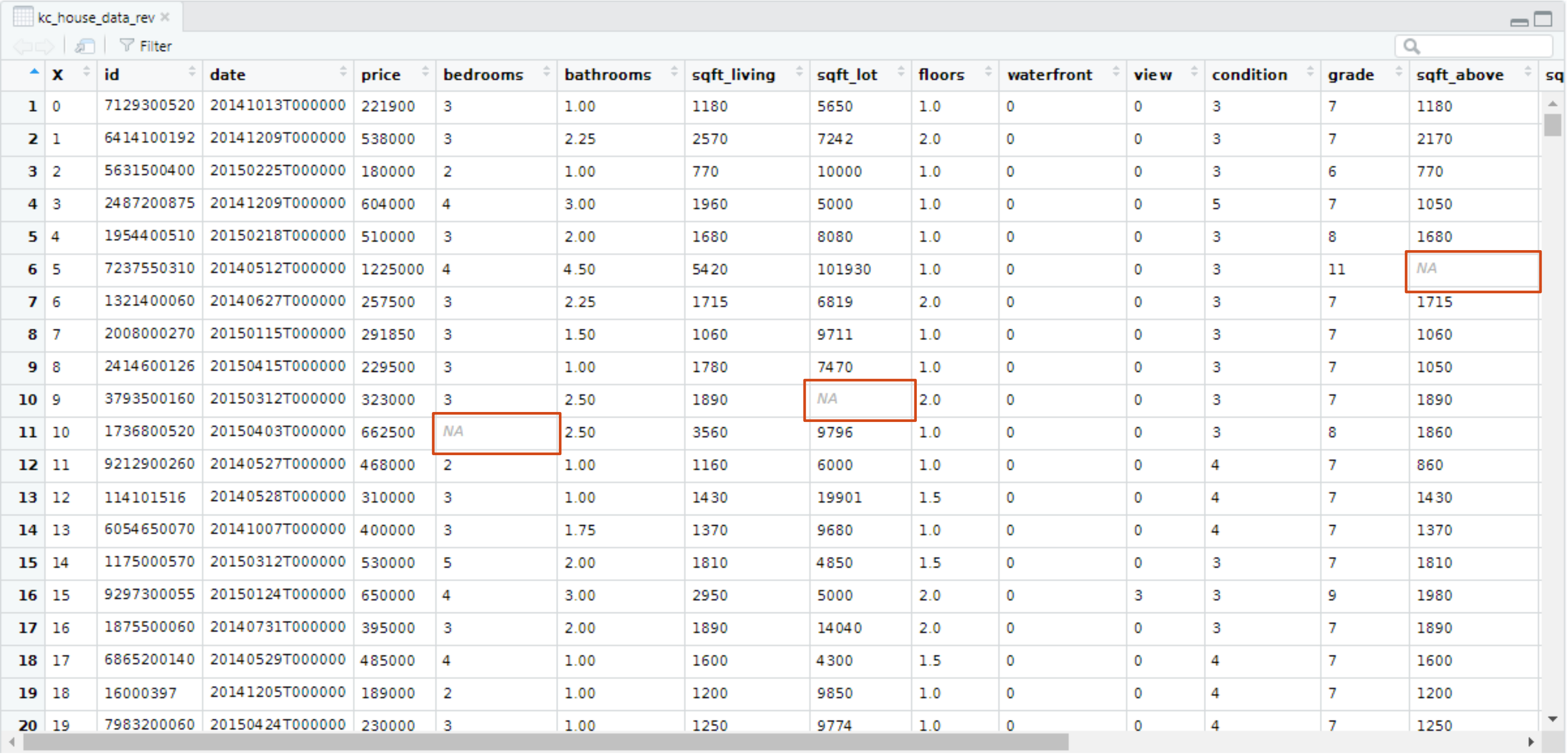
- Do you need additional attributes to address problem?
- Do you have sufficient examples or not?

Data Collection and Management

- Remove redundant and irrelevant examples
- The collected data often needs to be cleaned from missing values and outliers,

Data Collection and Management

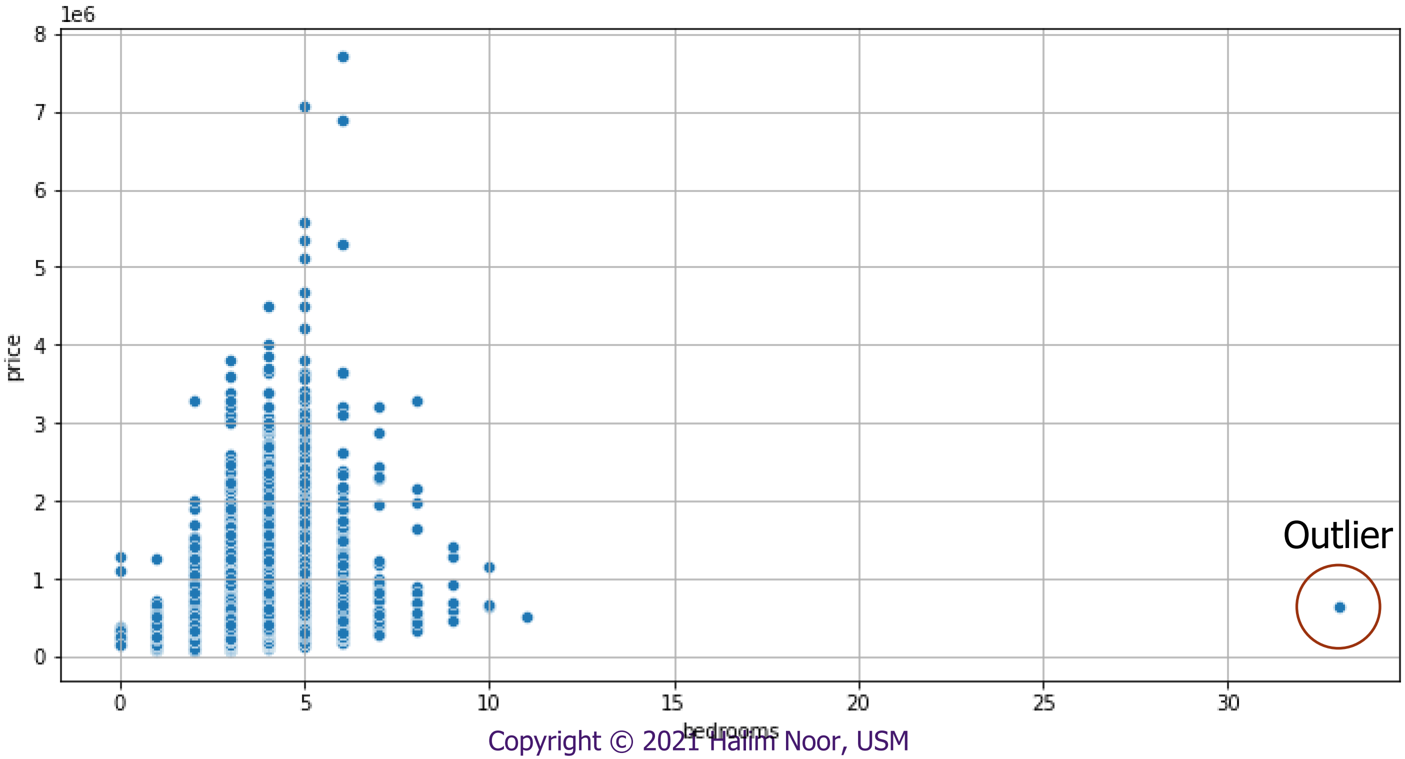
- An attribute of an example (row) has no value
- Are there missing values? How do handle them?



	X	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sq
1	0	7129300520	20141013T000000	221900	3	1.00	1180	5650	1.0	0	0	3	7	1180	
2	1	6414100192	20141209T000000	538000	3	2.25	2570	7242	2.0	0	0	3	7	2170	
3	2	5631500400	20150225T000000	180000	2	1.00	770	10000	1.0	0	0	3	6	770	
4	3	2487200875	20141209T000000	604000	4	3.00	1960	5000	1.0	0	0	5	7	1050	
5	4	1954400510	20150218T000000	510000	3	2.00	1680	8080	1.0	0	0	3	8	1680	
6	5	7237550310	20140512T000000	1225000	4	4.50	5420	101930	1.0	0	0	3	11	NA	
7	6	1321400060	20140627T000000	257500	3	2.25	1715	6819	2.0	0	0	3	7	1715	
8	7	2008000270	20150115T000000	291850	3	1.50	1060	9711	1.0	0	0	3	7	1060	
9	8	2414600126	20150415T000000	229500	3	1.00	1780	7470	1.0	0	0	3	7	1050	
10	9	3793500160	20150312T000000	323000	3	2.50	1890	NA	2.0	0	0	3	7	1890	
11	10	1736800520	20150403T000000	662500	NA	2.50	3560	9796	1.0	0	0	3	8	1860	
12	11	9212900260	20140527T000000	468000	2	1.00	1160	6000	1.0	0	0	4	7	860	
13	12	114101516	20140528T000000	310000	3	1.00	1430	19901	1.5	0	0	4	7	1430	
14	13	6054650070	20141007T000000	400000	3	1.75	1370	9680	1.0	0	0	4	7	1370	
15	14	1175000570	20150312T000000	530000	5	2.00	1810	4850	1.5	0	0	3	7	1810	
16	15	9297300055	20150124T000000	650000	4	3.00	2950	5000	2.0	0	3	3	9	1980	
17	16	1875500060	20140731T000000	395000	3	2.00	1890	14040	2.0	0	0	3	7	1890	
18	17	6865200140	20140529T000000	485000	4	1.00	1600	4300	1.5	0	0	4	7	1600	
19	18	16000397	20141205T000000	189000	2	1.00	1200	9850	1.0	0	0	4	7	1200	
20	19	7983200060	20150424T000000	230000	3	1.00	1250	9774	1.0	0	0	4	7	1250	

Data Collection and Management

- Values that are significantly differ from others
- Are there outliers? How do we handle them? Should we remove it or keep it?



Data Modeling

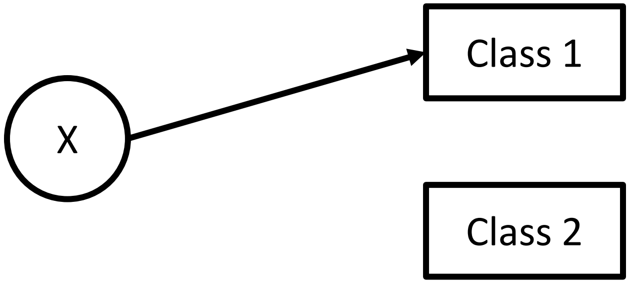
- To confirm our insight and hypothesis about the data
  - Number of rooms and House prices – positive relationship
- Fit the data using linear model
  - If slope is positive = positive relationship
  - If slope is negative = negative relationship

Data Modeling

- A model is an approximation of the data that describes the relationship between the attributes
- A model can be used to make predictions

Data Modeling

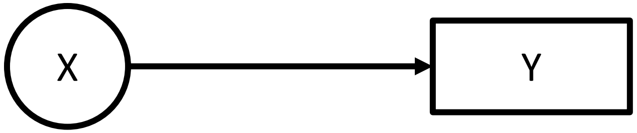
- Classification – deciding if something belongs to one category or another
- Placed order and Fraudulent or Not Fraudulent



X is an order, fraudulent order (Class 1) or valid order (Class 2)
X is social media activity, depression (Class 1) or not (Class 2)

Data Modeling

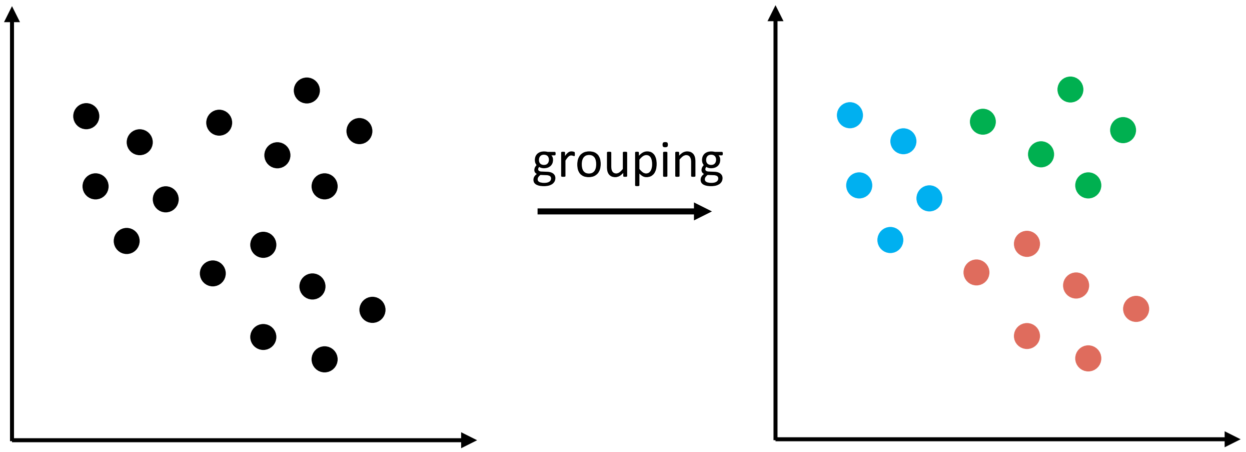
- Regression (scoring) – estimating a numeric value



X is house attributes, house price (Y)
X is weather, sales (Y)

Data Modeling

- Clustering – grouping items into most similar groups



Model Evaluation

- Determine if the model meets the goals
- Is it accurate enough for your needs? Does it perform better than whatever estimate that is currently being used?
- Do the results of the model make sense in the context of the problem domain?
- If no, then repeat the Modeling step (or the steps before it)

Presentation and Documentation

- Document the model for those who are responsible for using, running, and maintaining the model once it has been deployed
- Presentation must not be technical
- Presentation must make sense to the human brain and easy to understand – use visualization
- Highlight the most interesting findings or recommendation (if any)

Model Deployment

- Ensure the model run smoothly
- Model can be updated when needed
- Monitor the performance of the model
- Why the model’s decision is being overridden frequently?
- Is the model incomplete?

Summary

- Data science is a process of extracting knowledge or insight from data
- Data science project involves many roles and skills – back-and-forth between data scientist and project stakeholders
- Project goal must be specific, measurable and quantifiable

End