CPSC 100

Computational Thinking

Artificial Intelligence (cont'd)

Instructor: Firas Moosvi
Department of Computer Science
University of British Columbia

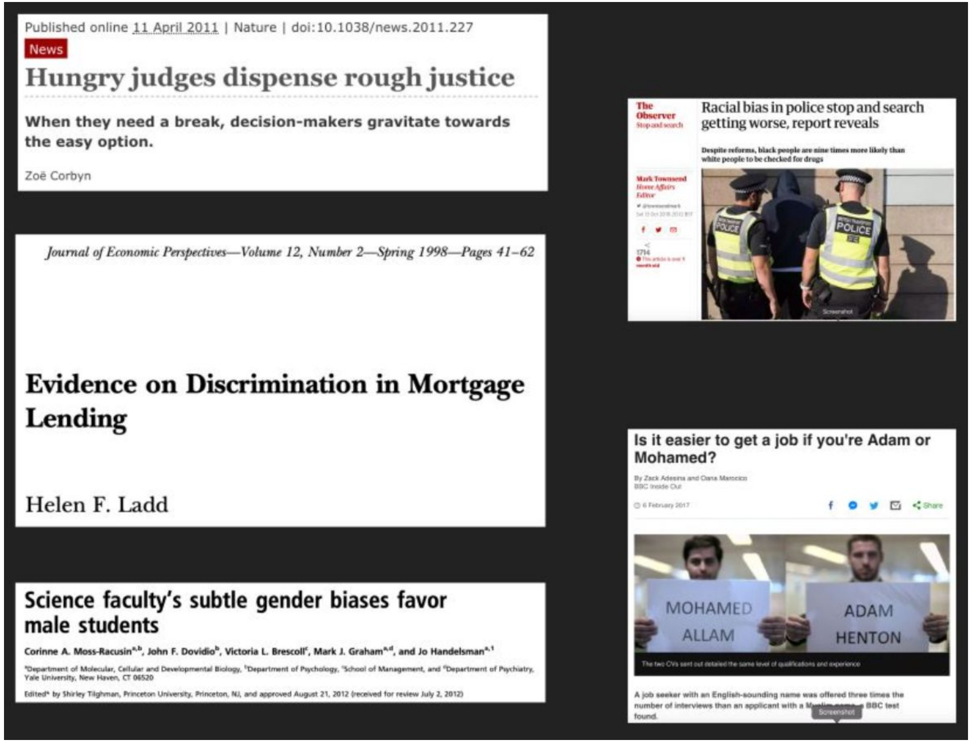
Agenda

- Finish off “Bias and Fairness” in the AI section
- Start Data Mining section!
- Introduction to Clustering

Bias and Fairness

Human bias

Bias in people refers to our tendency to take quick decisions based on little information

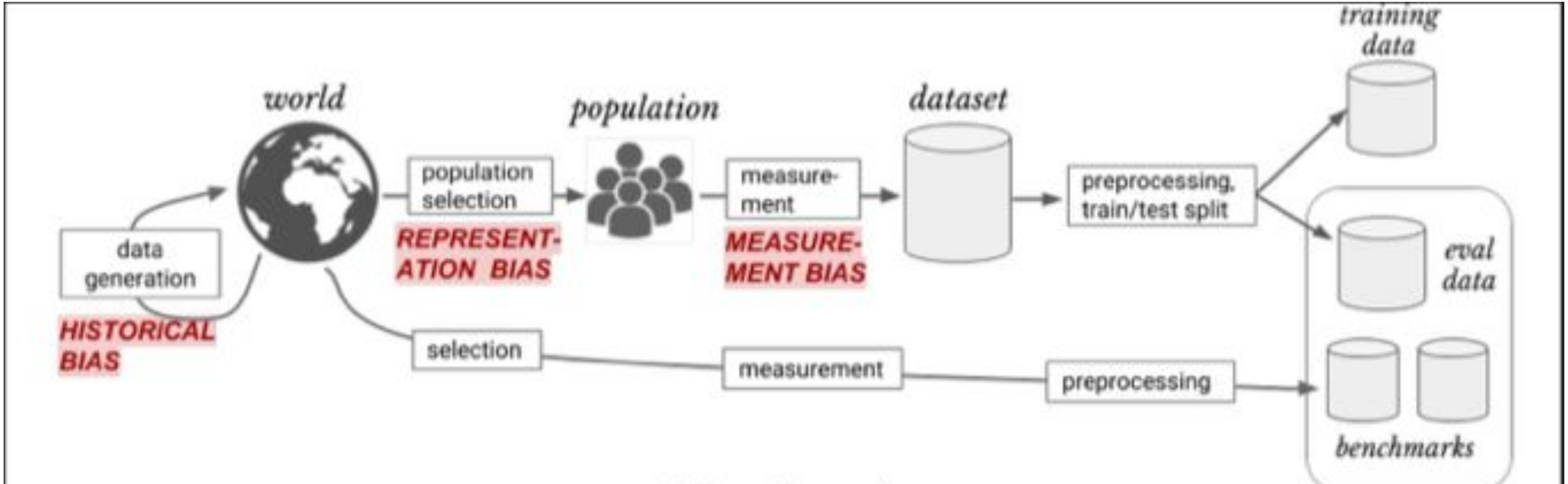


Why worry about bias in algorithms

Decisions made by a ML algorithm are:

- Cheap
- Scalable
- Automated
- Self-reinforcing
- Seemingly objective
- Often lacking appeals processes
- Not just predicting but also causing the future

Sources of bias in ML algorithms



(a) Data Generation

Representation bias

Representation bias arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population. [1]

1. **The sampling methods only reach a portion of the population.** For example, datasets collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones. Similarly, medical data for a particular condition may be available only for the population of patients who were considered serious enough to bring in for further screening.

2. **The population of interest has changed or is distinct from the population used during model training.** Data that is representative of Boston, for example, may not be representative if used to analyze the population of Indianapolis. Similarly, data representative of Boston 30 years ago will likely not reflect today's population.

Measurement bias

Measurement Bias arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group- or input-dependent noise that leads to differential performance.

1. **The measurement process varies across groups.** For example, if a group of factory workers is more stringently or frequently monitored, more errors will be observed in that group. This can also lead to a feedback loop wherein the group is subject to further monitoring because of the apparent higher rate of mistakes (Barocas and Selbst, 2016;

2. **The quality of data varies across groups.** Structural discrimination can lead to systematically higher error rates in a certain group. For example, women are more likely to be misdiagnosed or not diagnosed for conditions where self-reported pain is a symptom (Calderone, 1990). In this case, "*diagnosed* with condition X" is a biased proxy for "has condition X."

3. **The defined classification task is an oversimplification.** In order to build a supervised ML model, some label to predict must be chosen. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure. Consider the prediction problem of deciding whether a student will be successful (e.g., in a college admissions context). Fully capturing the outcome of 'successful student' in terms of a single measurable attribute is impossible because of its complexity. In cases such as these, algorithm designers resort to some available label such as 'GPA' (Kleinberg et al., 2018), which ignores different indicators of success achieved by parts of the population.

Historical bias

Historical bias arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

Example: image search In 2018, 5% of Fortune 500 CEOs were women (Zarya, 2018). Should image search results for "CEO" reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for "CEO" to display a higher proportion of women.

Fairness in algorithms

Increasing attention on algorithms being **fair**, not just accurate

Fairness can be measured as:

- demographic (or statistical) parity: population percentage should be reflected in the output classes
- Equality of false negatives or equalized odds: constant false-negative (or both false-negative and true-negative) rates across groups.
- Equal opportunity: equal True Positive Rate for all groups
- Other metrics...

Accuracy and fairness tend to be at odds with each other.

Algorithms can be audited to test their fairness.

Are we ethically required to sacrifice accuracy for fairness? To what extent?

When the metric becomes the target (Goodhart's Law)

"When a measure becomes a target it ceases to be a good measure"

Metrics introduced in the [British public healthcare system] (e.g. waiting time in ER) caused people to game it:

- Cancelled scheduled operations to draft extra staff to ER
- Required patients to wait outside the ER, e.g. in ambulances
- Put stretchers in hallways and classified them as "beds"
- Hospital and patients reported different wait times

Big Data is significantly changing college applications

- Universities are given higher ranking for things such as receiving more applications, being more selective, and having more students accept their offers (while tuition is not considered)
- This even pushed some mid-tier universities to reduce the number of offer letter sent out, especially to good students who they think will not accept. Can affect applications to "safety schools"

Can you think of more examples?

Can you think of ways to avoid this trap?

Ethics of pricing algorithms

Algorithms are currently used to adjust prices based on:

- Willingness of buyer
- Availability

Uber surge pricing:

- In 2014, terrorists attacked a café in Sidney, holding 10 customers and 8 employees hostage for 16 hours
- During this time, people from the surrounding areas were evacuated. Transportation was disrupted.
- Uber prices adapted by increasing the rate to a minimum of 100$
- In general, underserved (poorer) areas get worse rates under current pricing policy
- Drivers are also subjected to different pricing/waiting times
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4331080

Is Uber morally obliged to avoid such pricing disparities?

From trading to gambling addiction?

Case study: Robinhood Trading app

Designed to make trading more accessible and equitable. Robinhood's mission statement is "To democratize finance for all”

Concerns around the most “dopamine-inducing” features (source):

- Green confetti to celebrate transactions.
- A constant update of stock related articles.
- A colorful, eye-catching interface.
- Emoji phone notifications.
- One-click trading for instant gratification.
- Free stocks in the shape of lottery tickets.

Technology that will be in widespread use

Tailored solutions for specific tasks, not general intelligence

Prototypes that work today in labs & narrow deployments

Some examples:

- Non-text input modalities (vision; speech)
- Consumer modeling (recommendation; marketing)
- Cloud services (translation; question answering; AI-mediated outsourcing)
- Transportation (automated trucking; some self-driving cars)
- Industrial robotics (factories; some drone applications)
- AI knowledge work (logistics planning; radiology; legal research; call centers)
- Policing & security (electronic fraud; cameras; predictive policing)

Technologies that won't take off as quickly

Overall, areas in which

- major entrenched **regulatory regimes** need to be navigated
- there exist **social/cultural barriers** to adoption
- the **human touch** is crucial
- substantial **new hardware** would need to be developed

Some **examples**:

- childcare, healthcare, eldercare
- education
- coaching, counselling
- consumer robots beyond niche applications
- semantically rich language understanding

Superhuman Intelligence

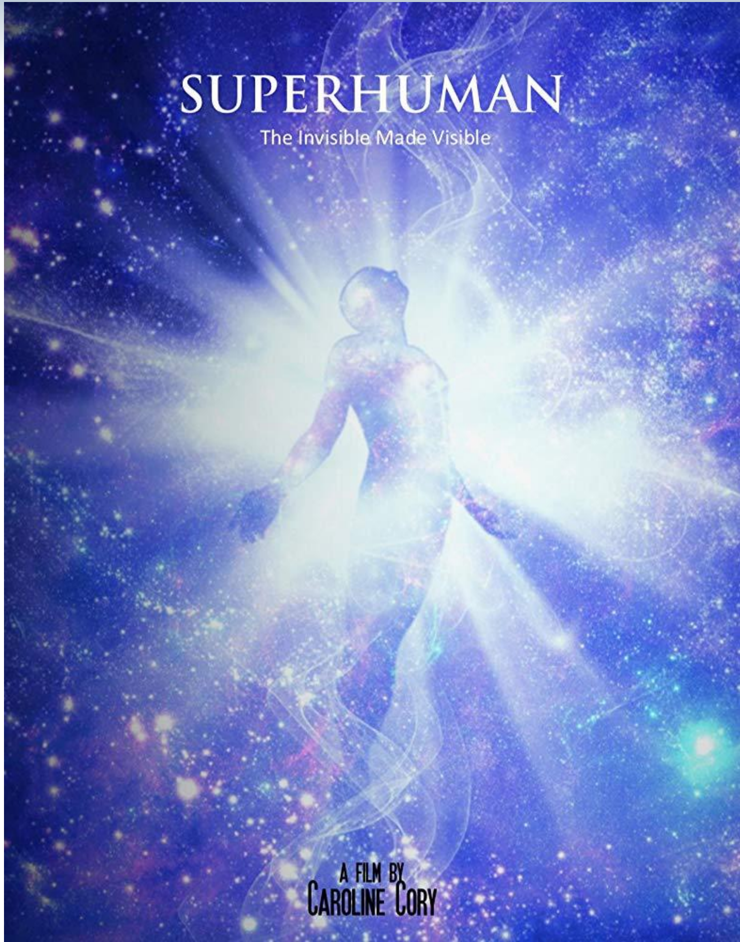
AI systems will increasingly be capable of reaching **human-level performance**

Superhuman intelligence isn’t such a foreign, scary thing

– governments, corporations, NGOs exhibit behavior much more sophisticated and complex than that of any individual

Many important problems need superhuman intelligence; AI can help

– improved **collective decision making**
– more efficient use of **scarce resources**
– addressing **underserved communities**
– **climate change**; other societal challenges



Ethics of AI

Will a new technology:

disempower **individuals vs corporations**?

⇒ user modeling; **data mining**; fostering addictive behaviors; developmental effects on children

disempower **individuals vs governments**?

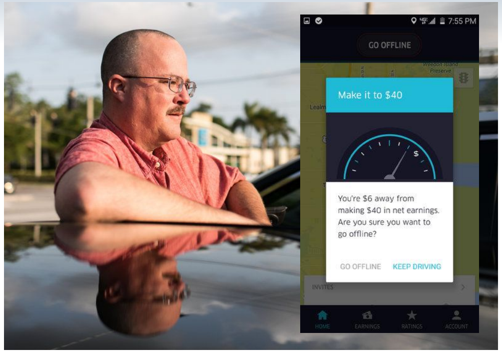
⇒ facilitate disinformation (deep fakes; bots masquerading as people; filter bubbles); enable qualitatively new military or security tactics

take **autonomous actions** in a way that obscures responsibility

⇒ autonomous weapons; self-driving cars; loan approval systems

disproportionately affect **vulnerable/marginalized groups**

⇒ automated decision making tools trained in ways that may encode existing biases



Social Impact

How will AI technologies **transform society**?

Will there be a **social backlash** against AI?

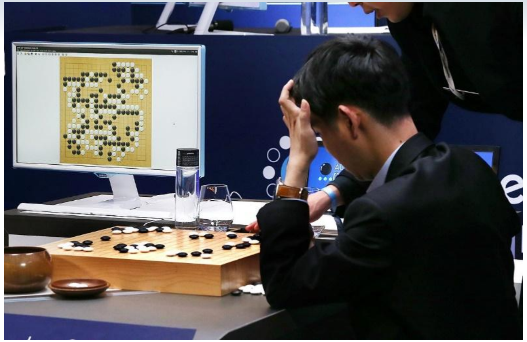
- If so, what will be considered AI?

This **generation of children** will grow up taking for granted many technologies that strike us as magical

How will **human relationships** change in the presence of always-available social agents?

As we are increasingly **augmented by AI**, what are our inherent cognitive/emotional/motivational limitations, beyond which augmentation won't help?

Ethical consideration of advances in AI



Is it wrong to create machines capable of making human labor obsolete? Will humans become demoralized by the presence of vastly more intelligent robots?

How can we ensure that intelligent robots will not be put to an evil purpose by a malevolent human? How can we ensure they do not adopt malevolent purposes themselves?

Is it morally acceptable to create "personal" (self-conscious) AI?

CPSC 100

Computational Thinking

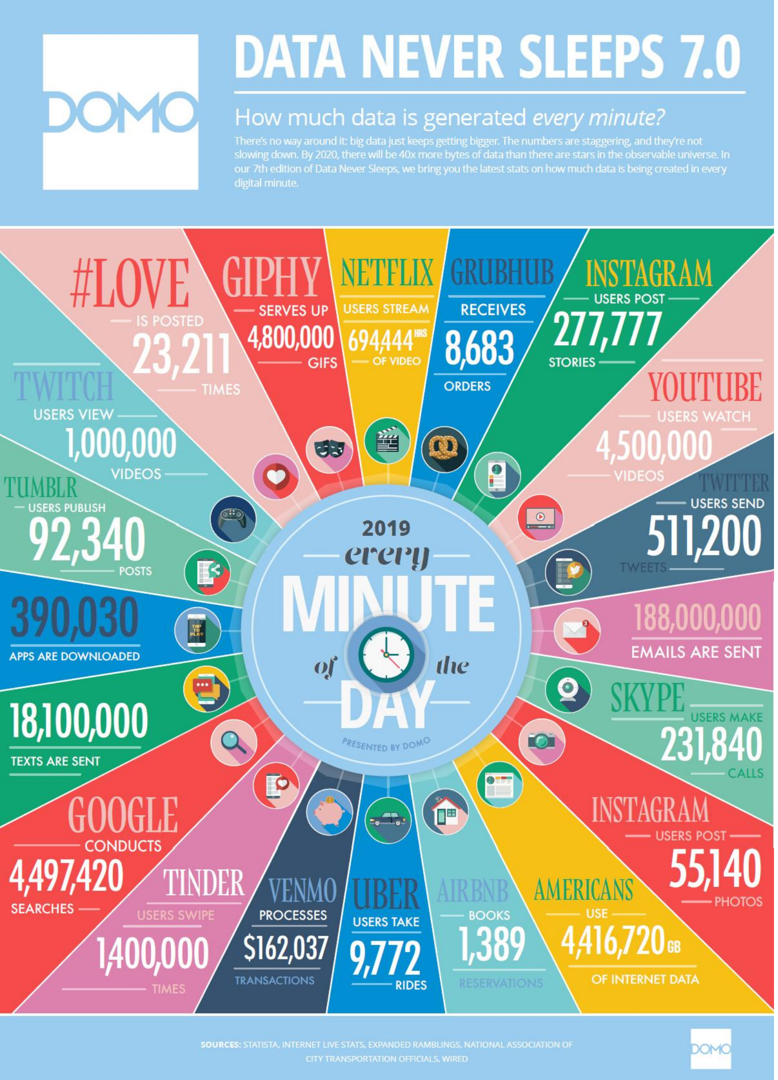
Data Mining

Instructor: Firas Moosvi
Department of Computer Science
University of British Columbia

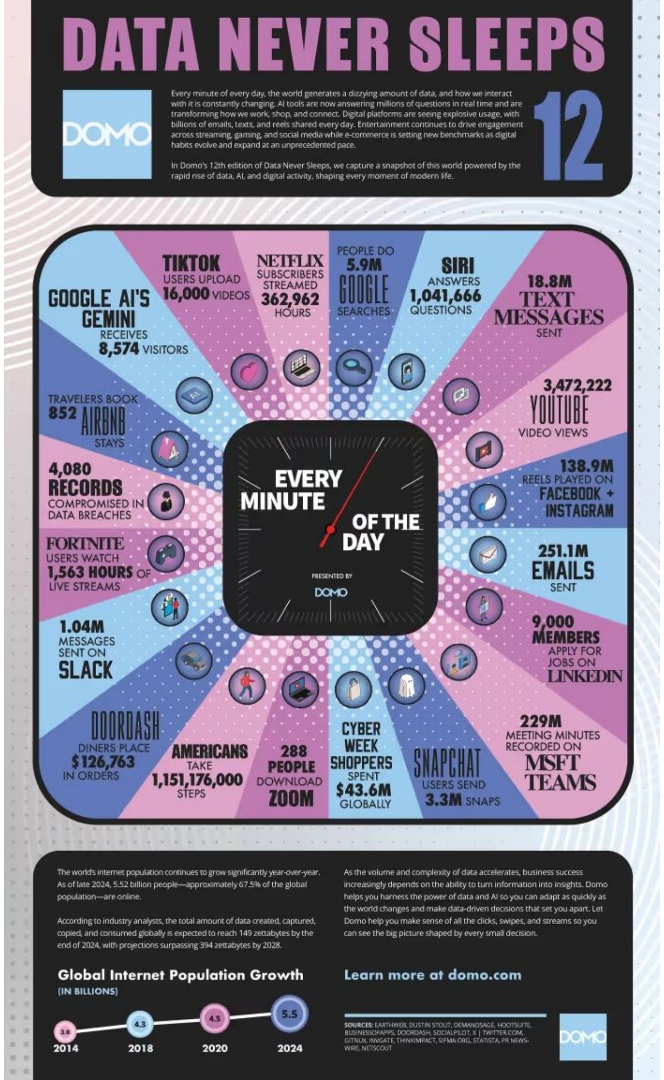
Learning Goals

After this week's lecture, you should be able to:

- Explain what is data mining, and why it is important to be aware of
- Differentiate between the uses of clustering and classification techniques and give examples of each.
- Define and identify the measures of quality for clustering algorithms.
- Convert between different data units (KB, MB, TB, etc…)
-
-

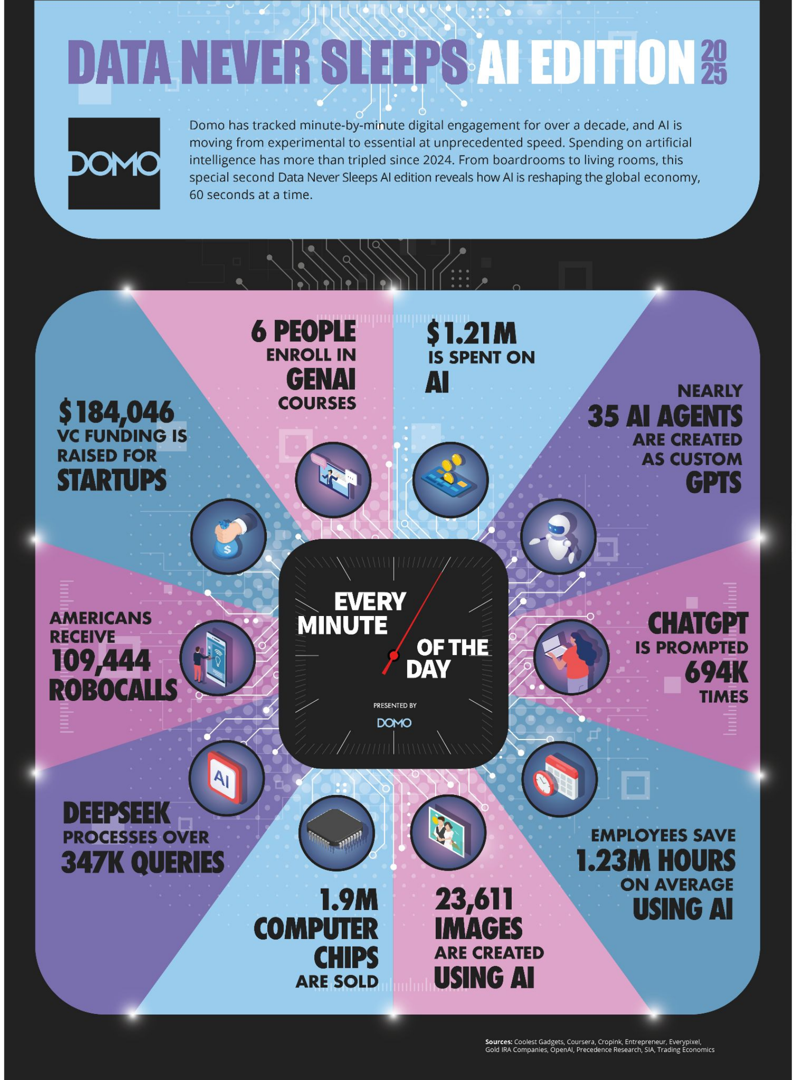


DATA NEVER SLEEPS 7.0
DOMO
How much data is generated every minute?
There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute.
#LOVE IS POSTED 23,211 TIMES
GIPHY SERVES UP 4,800,000 GIFS
NETFLIX USERS STREAM 694,444 HRS OF VIDEO
GRUBHUB RECEIVES 8,683 ORDERS
INSTAGRAM USERS POST 277,777 STORIES
TWITCH USERS VIEW 1,000,000 VIDEOS
YOUTUBE USERS WATCH 4,500,000 VIDEOS
TUMBLR USERS PUBLISH 92,340 POSTS
TWITTER USERS SEND 511,200 TWEETS
390,030 APPS ARE DOWNLOADED
188,000,000 EMAILS ARE SENT
18,100,000 TEXTS ARE SENT
SKYPE USERS MAKE 231,840 CALLS
2019 every MINUTE of the DAY
PRESENTED BY DOMO
GOOGLE CONDUCTS 4,497,420 SEARCHES
INSTAGRAM USERS POST 55,140 PHOTOS
TINDER USERS SWIPE 1,400,000 TIMES
VENMO PROCESSES $162,037 TRANSACTIONS
UBER USERS TAKE 9,772 RIDES
AIRBNB BOOKS 1,389 RESERVATIONS
AMERICANS USE 4,416,720 GB OF INTERNET DATA
SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED



DATA NEVER SLEEPS 12
DOMO
Every minute of every day, the world generates a dizzying amount of data, and how we interact with it is constantly changing. AI tools are now answering millions of questions in real time and are transforming how we work, shop, and connect. Digital platforms are seeing explosive usage, with billions of emails, texts, and reels shared every day. Entertainment continues to drive engagement across streaming, gaming, and social media while e-commerce is setting new benchmarks as digital habits evolve and expand at an unprecedented pace.
In Domo's 12th edition of Data Never Sleeps, we capture a snapshot of this world powered by the rapid rise of data, AI, and digital activity, shaping every moment of modern life.
GOOGLE AI'S GEMINI RECEIVES 8,574 VISITORS
TIKTOK USERS UPLOAD 16,000 VIDEOS
NETFLIX SUBSCRIBERS STREAMED 362,962 HOURS
PEOPLE DO 5.9M GOOGLE SEARCHES
SIRI ANSWERS 1,041,666 QUESTIONS
18.8M TEXT MESSAGES SENT
TRAVELERS BOOK 852 AIRBNB STAYS
3,472,222 YOUTUBE VIDEO VIEWS
4,080 RECORDS COMPROMISED IN DATA BREACHES
138.9M REELS PLAYED ON FACEBOOK + INSTAGRAM
EVERY MINUTE OF THE DAY
PRESENTED BY DOMO
FORTNITE USERS WATCH 1,563 HOURS OF LIVE STREAMS
251.1M EMAILS SENT
1.04M MESSAGES SENT ON SLACK
9,000 MEMBERS APPLY FOR JOBS ON LINKEDIN
DOORDASH DINERS PLACE $126,763 IN ORDERS
AMERICANS TAKE 1,151,176,000 STEPS
288 PEOPLE DOWNLOAD ZOOM
CYBER WEEK SHOPPERS SPENT $43.6M GLOBALLY
SNAPCHAT USERS SEND 3.3M SNAPS
229M MEETING MINUTES RECORDED ON MSFT TEAMS
The world's internet population continues to grow significantly year-over-year. As of late 2024, 5.52 billion people—approximately 67.5% of the global population—are online.
According to industry analysts, the total amount of data created, captured, copied, and consumed globally is expected to reach 149 zettabytes by the end of 2024, with projections surpassing 394 zettabytes by 2028.
As the volume and complexity of data accelerates, business success increasingly depends on the ability to turn information into insights. Domo helps you harness the power of data and AI so you can adapt as quickly as the world changes and make data-driven decisions that set you apart. Let Domo help you make sense of all the clicks, swipes, and streams so you can see the big picture shaped by every small decision.
Global Internet Population Growth
(IN BILLIONS)
3.0 2014
4.3 2018
4.5 2020
5.5 2024
Learn more at domo.com
SOURCES: EARTHWEB, DUSTIN STOUT, DEMANDSAGE, HOOTSUITE, BUSINESSOFAPPS, DOORDASH, SOCIALPILOT, X | TWITTER.COM, GITNUX, INVGATE, THINKIMPACT, SIFMA.ORG, STATISTA, PR NEWSWIRE, NETSCOUT

Source


DATA NEVER SLEEPS AI EDITION 2025
DOMO
Domo has tracked minute-by-minute digital engagement for over a decade, and AI is moving from experimental to essential at unprecedented speed. Spending on artificial intelligence has more than tripled since 2024. From boardrooms to living rooms, this special second Data Never Sleeps AI edition reveals how AI is reshaping the global economy, 60 seconds at a time.
EVERY MINUTE OF THE DAY
PRESENTED BY DOMO
6 PEOPLE ENROLL IN GENAI COURSES
$1.21M IS SPENT ON AI
NEARLY 35 AI AGENTS ARE CREATED AS CUSTOM GPTS
CHATGPT IS PROMPTED 694K TIMES
EMPLOYEES SAVE 1.23M HOURS ON AVERAGE USING AI
23,611 IMAGES ARE CREATED USING AI
1.9M COMPUTER CHIPS ARE SOLD
DEEPSEEK PROCESSES OVER 347K QUERIES
AMERICANS RECEIVE 109,444 ROBOCALLS
$184,046 VC FUNDING IS RAISED FOR STARTUPS
AI
Sources: Coolest Gadgets, Coursera, Cropink, Entrepreneur, Everypixel, Gold IRA Companies, OpenAI, Precedence Research, SIA, Trading Economics

Source

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

Source

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

IDC

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

95m

photos and videos are shared on Instagram

Instagram Business

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

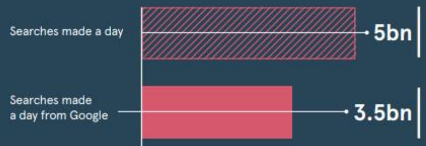
3.9bn

people use emails

4TB

of data produced by a connected car

Intel



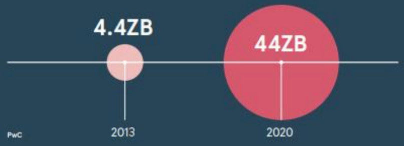
Smart Insights

28PB

to be generated from wearable devices by 2020

Statista

ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC

Source RACONTEUR

Why Should I Care about Data Mining?

By 2020, about 1.7MB of new data will be created every second for every human on the planet.

By then, our accumulated digital universe of data will grow from 4.4 zettabytes today to ~44 zettabytes, or
44 *trillion* gigabytes.

A trillion gigabytes is a **zettabyte** ,
1 000 000 000 000 000 000 000 bytes
10^{21} bytes
2^{70} bytes

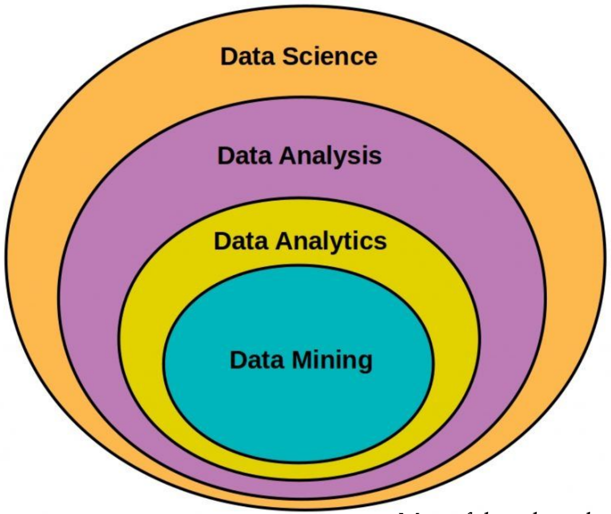
Data Mining –Definition

Data mining is the process of looking for patterns in large data sets

There are many different kinds for many different purposes

In depth exploration of some tasks:

- Clustering (k-means and DBScan)
- Classification (decision trees)
- Also look at kNN



Map of data disciplines

Clustering Definition

Clustering is partitioning a set of items into subgroups so as to ensure certain measures of quality (e.g., "similar" items are grouped together)

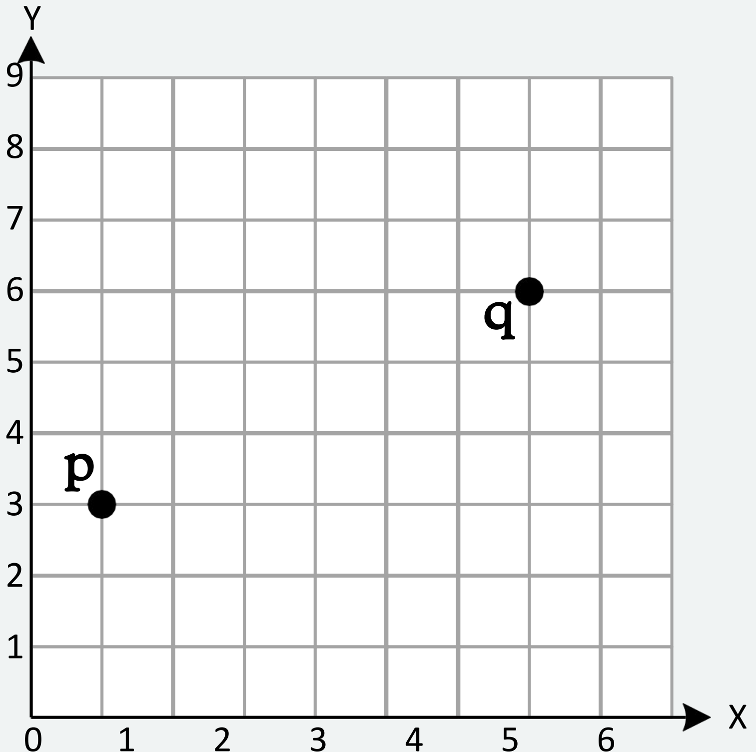
"Similar", for our purposes is measured by the Euclidean distance between the points

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

p, q = Two points in Euclidean n-space

$q_i - p_i$ = Euclidean vectors

n = n-space

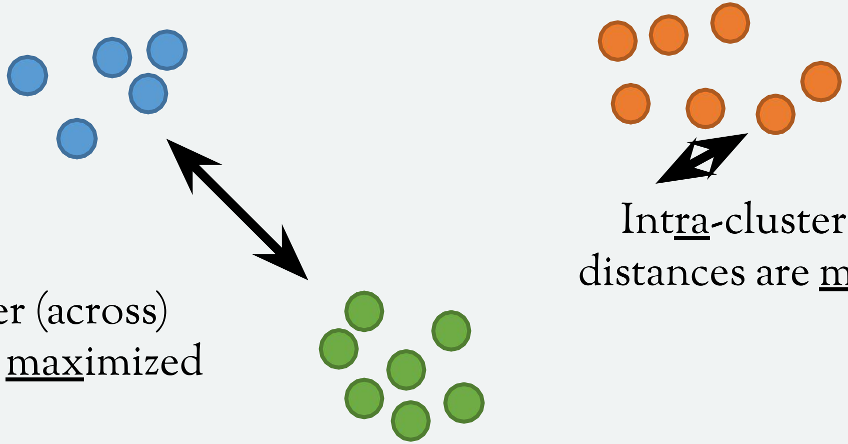


KMeans Algorithm - Watch Video before lab!

What is Good Clustering?

Items in the *same* cluster are similar (*intra*-cluster)

Items in *different* clusters are different (*inter*-cluster)



Intra-cluster (within) distances are minimized

Inter-cluster (across) distances are maximized

Measuring Clustering Quality Possible Criteria to Use

Intra-class (intra-cluster) similarity: points within a cluster contain are close to each other (or at least to their closest neighbours)

Inter-class (inter-cluster) dissimilarity: points in two different clusters are far from each other (or at least to their closest neighbours in other clusters)

Size similarity : clusters have similar size

Wrap up