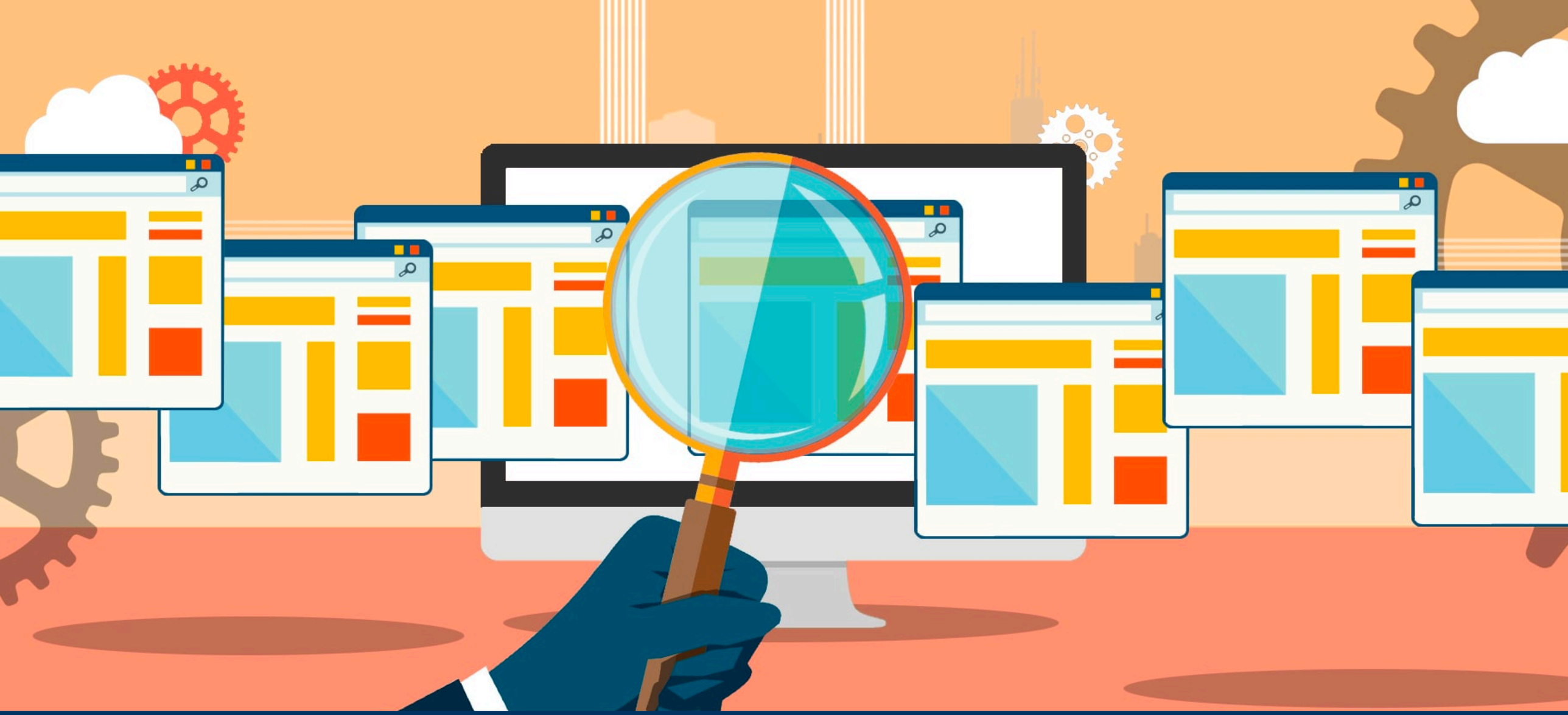


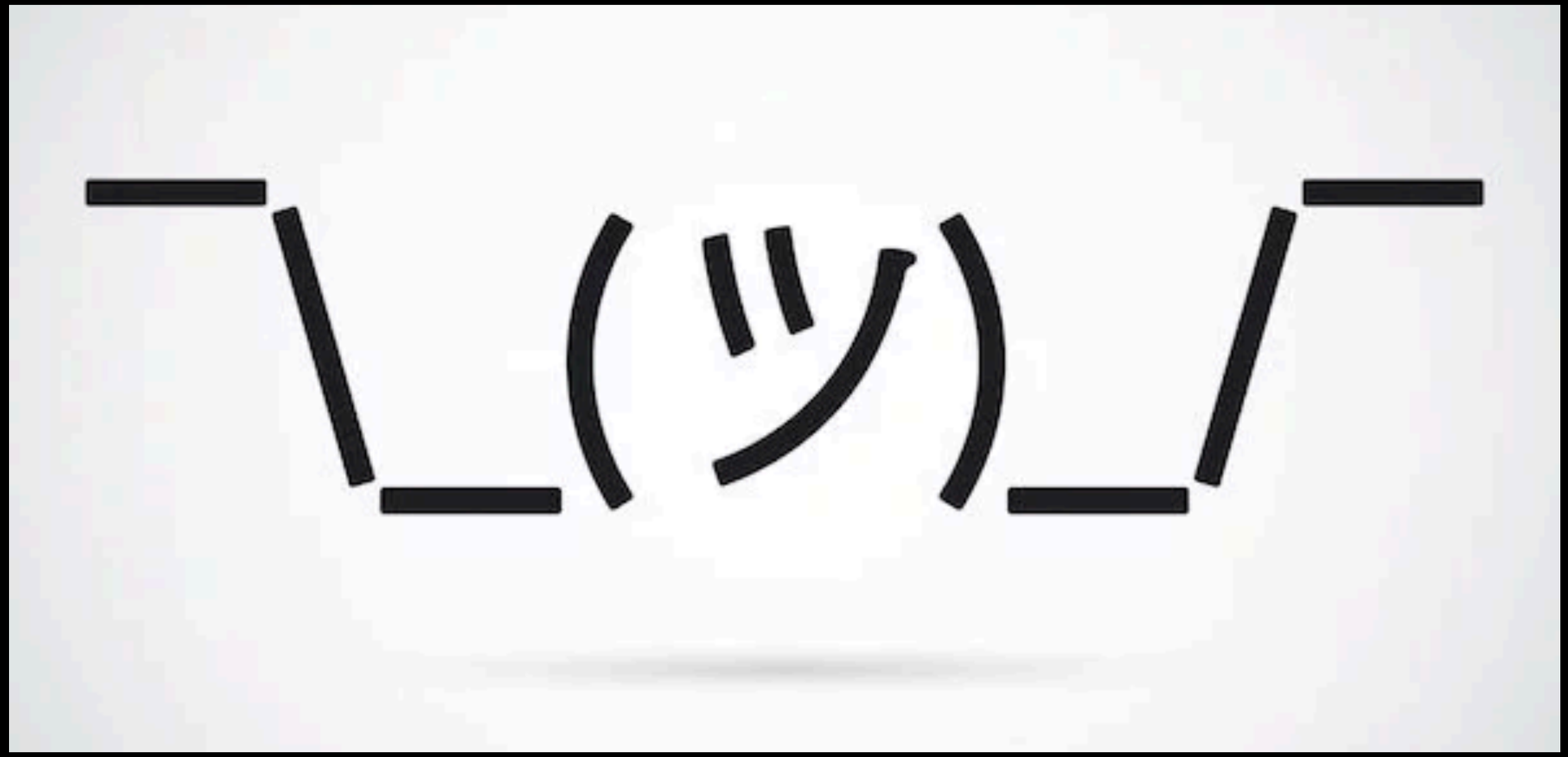
Data Science and Motivations

INST414 - Data Science Techniques

This Lecture's Learning Objectives

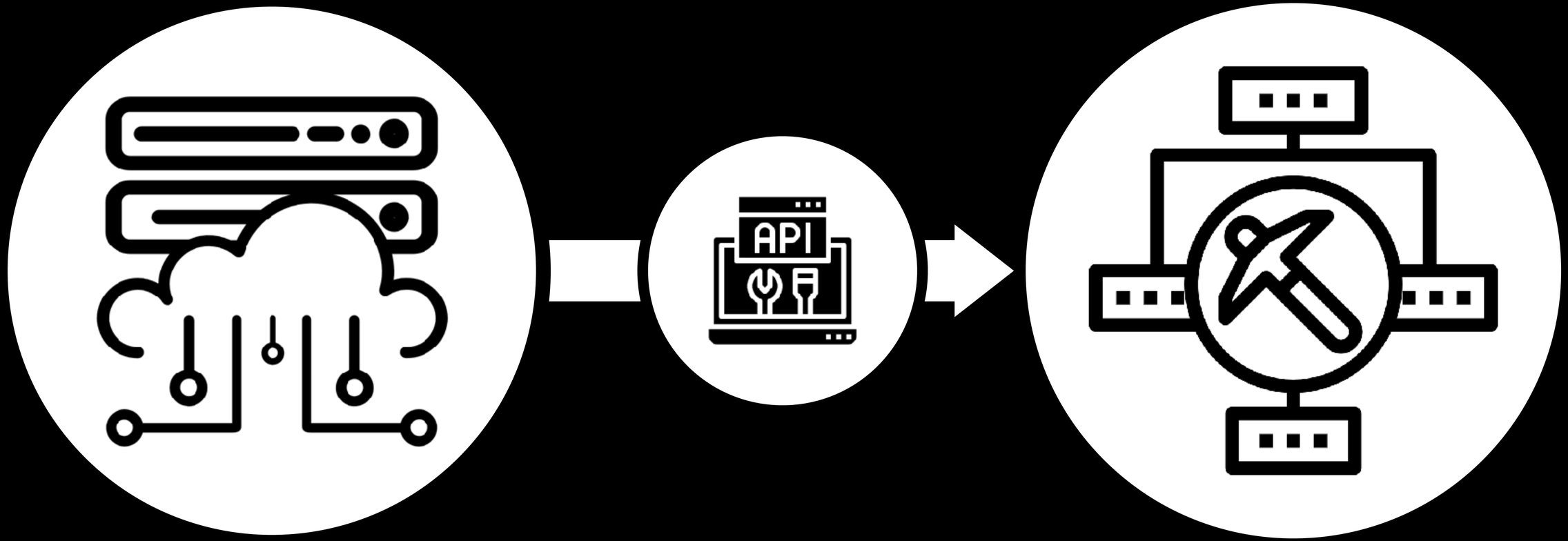
Define data science and differentiate it from other fields of study

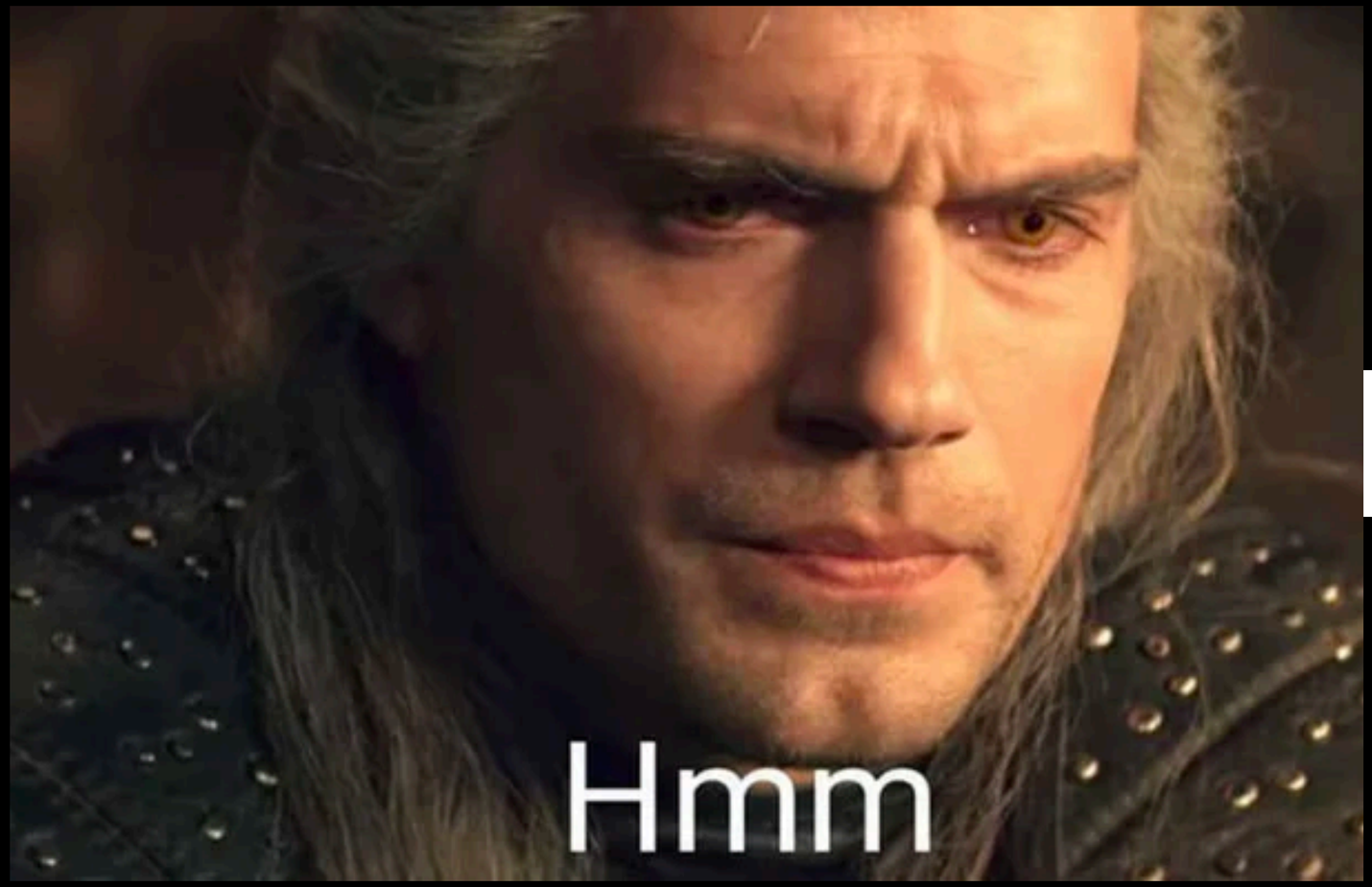
Define and differentiate three kinds of data we will analyze



What is Data Science?

This class: Mining + Analysis on Real Data



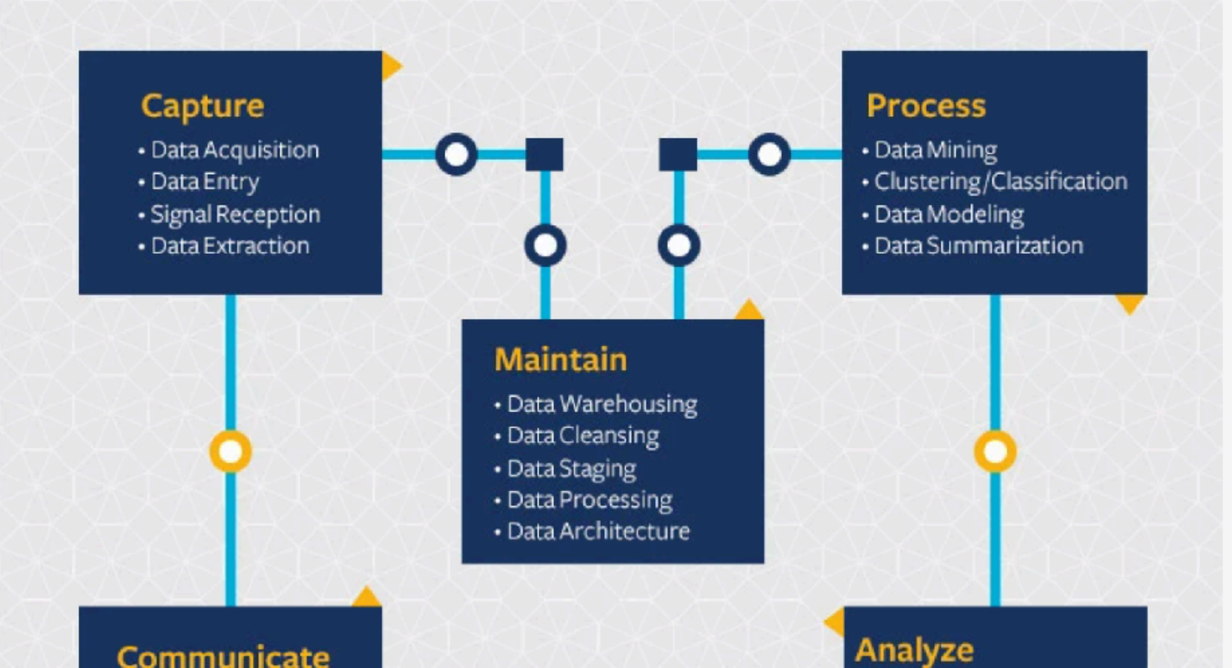


What is “Data Science”?

What is Data Science?

Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.

The Data Science Life Cycle

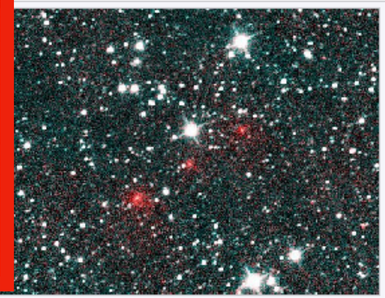


Data science

Contents [hide]

Not to be confused with information science.

Data science is an interdisciplinary academic field [1] that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data.[2] Data science also integrates domain knowledge from the underlying application domain (e.g., natural sciences, information technology, medicine).[3] Data science is multifaceted and can be described as a science, as a research paradigm, as a research method, as a discipline, as a workflow, and as a profession.[4]



The existence of Comet NEOWISE (here depicted as a series of red dots) was discovered by analyzing astronomical survey data acquired by a space telescope, the Wide-field Infrared Survey Explorer.

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyse actual phenomena" with data.[5] It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge.[6] However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[7][8]

A **data scientist** is someone who creates programming code and combines it with statistical knowledge to create insights from data.[9]

Foundations [edit]

Data science is an interdisciplinary field[10] focused on extracting knowledge from typically large data sets and applying the knowledge and insights from that data to solve problems in a wide range of application domains.[11] The field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science, statistics, information science, mathematics, data visualization, information visualization, data sonification, data integration, graphic design, complex systems, communication and business.[12][13] Statistician Nathan Yau, drawing on Ben Fry, also links data science to human–computer interaction: users should be able to intuitively control and explore data.[14][15] In 2015, the American Statistical Association identified

What Is Data Science?

View Free Analytics Services

Explore Free Analytics Offers
View free offers for Analytics services in the cloud

Check out Analytics Services
Innovate faster with the most comprehensive set of Analytics services

Browse Analytics Trainings
Get started on Analytics training with content built by AWS experts

Read Analytics Blogs
Read about the latest AWS Analytics product news and best practices

What is data science?

Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

What is data science?

Learn how data science can unlock business insights and accelerate digital transformation and enable data-driven decision making

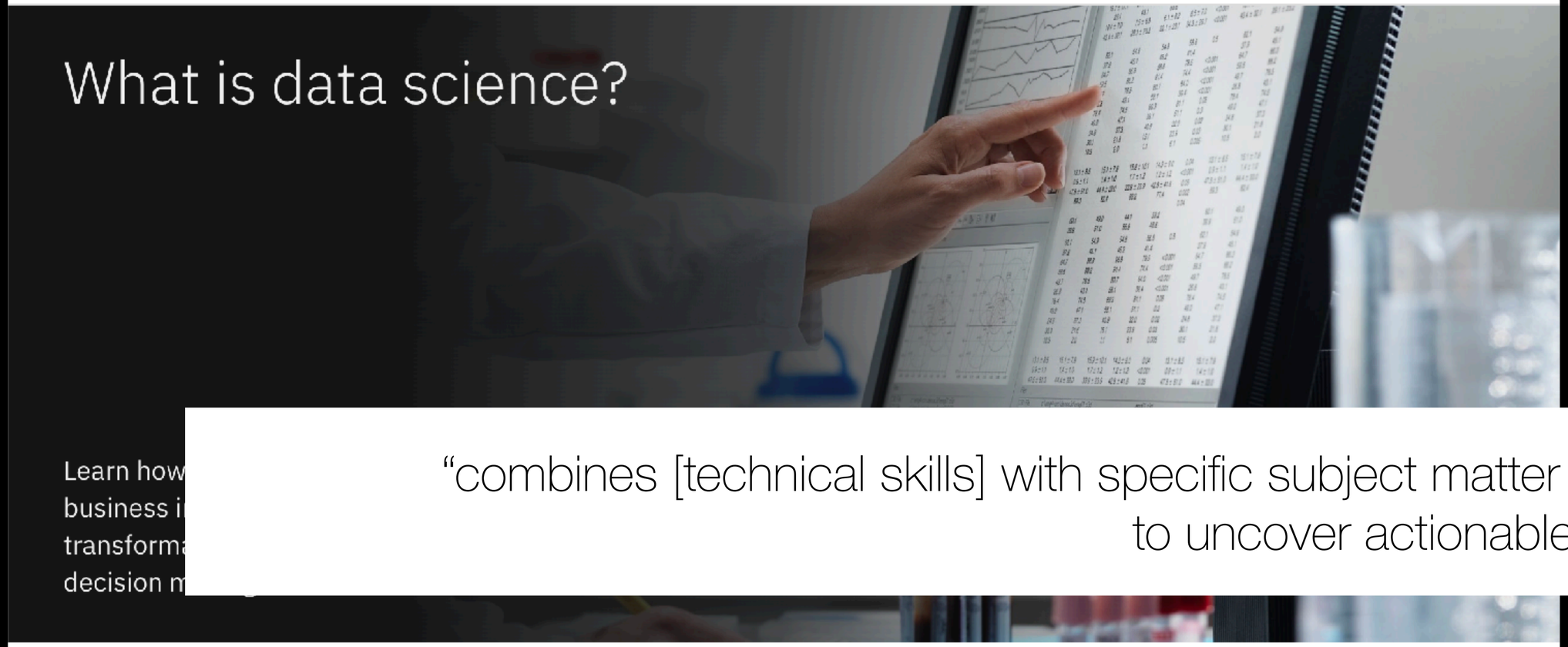
What is data science?

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning.

The accelerating volume of data sources, and subsequently data, has made data science is one of the fastest growing field across every industry. As a result, it is no surprise that

What do these definitions have in common?

What is data science?



Learn how
business i
transforma
decision m

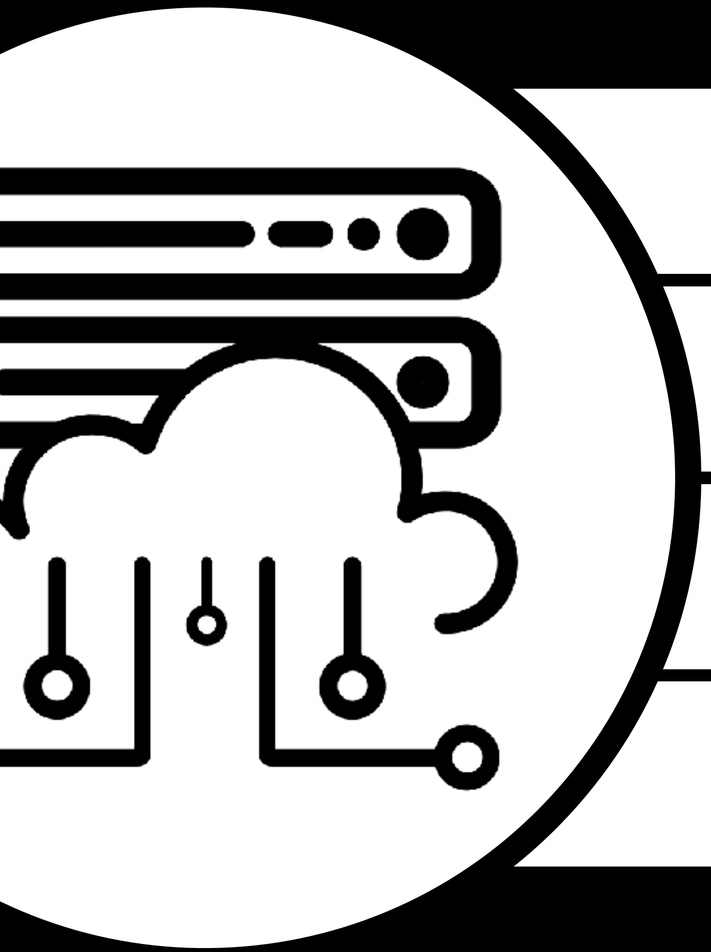
"combines [technical skills] with specific subject matter expertise to uncover actionable insights"

What is data science?

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning.

The accelerating volume of data sources, and subsequently data, has made data science is one of the fastest growing field across every industry. As a result, it is no surprise that

Data Science – Discovering insights



Valid: The insight should hold on new data with some certainty

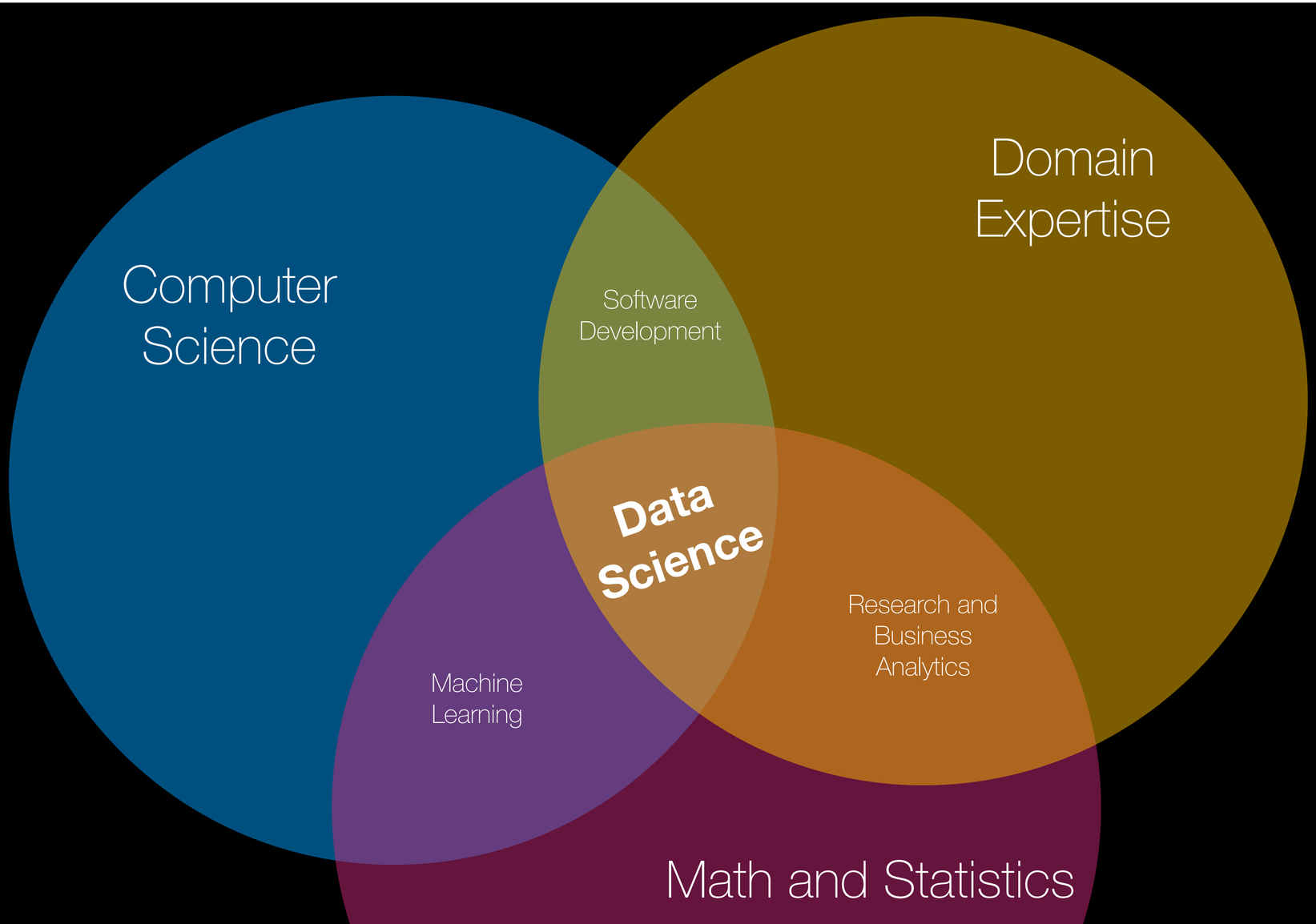
Useful: The insight should be actionable in some way

Unexpected: The insight is not obvious to the system

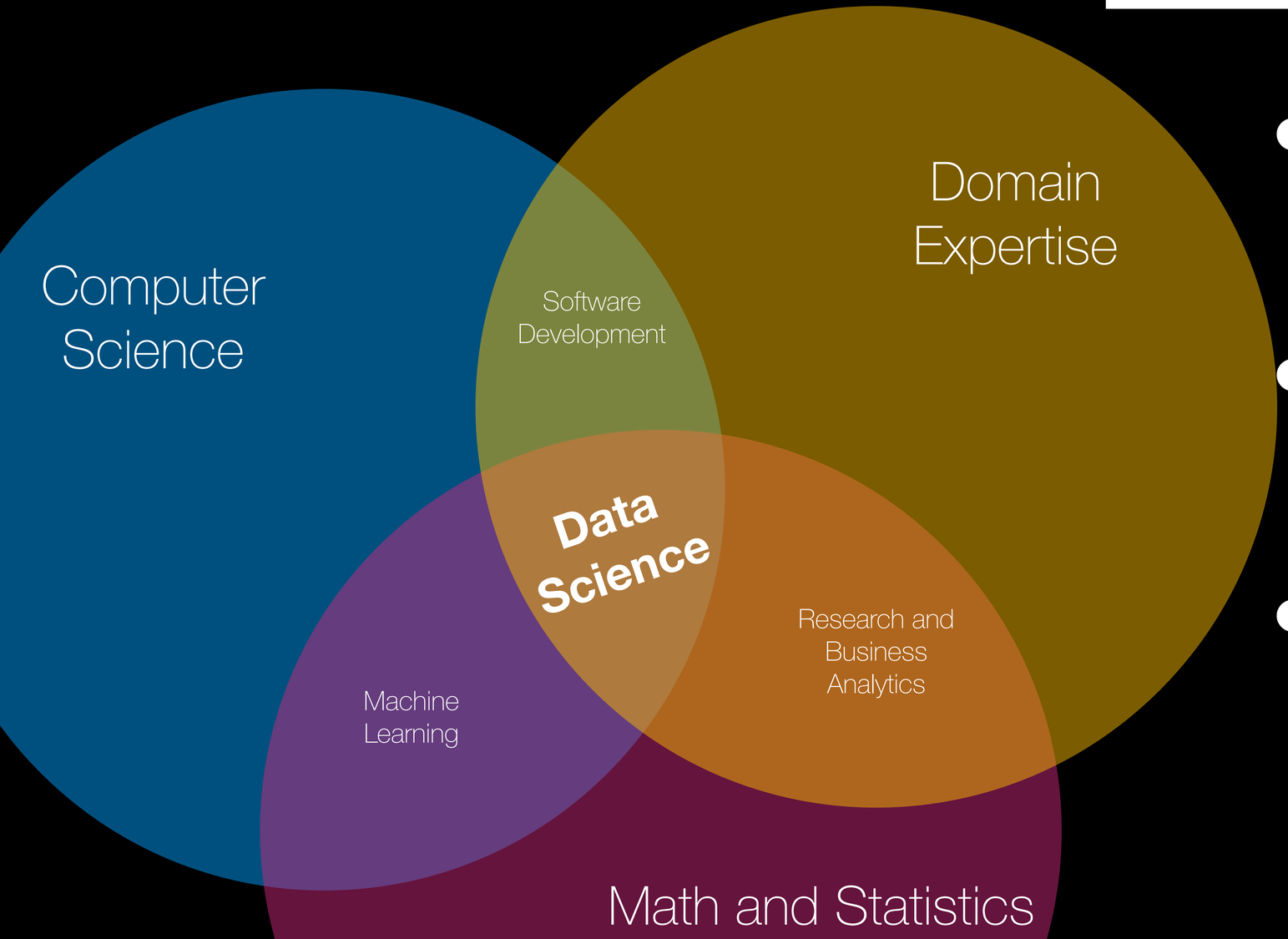
Understandable: Humans should be able to interpret the pattern or model providing the insight

What differentiates data science from other disciplines?

Data Science and Other Disciplines



Data Science ≠ Machine Learning



- Data science *includes* machine learning
- But also extracting and visualizing insights
- Not all insights require sophisticated models

What We Learned From Apple

nytimes.com/2021/01/27...

Apps | Jordan's Style | https://inf.ethz.ch/... | tome - Synology... | Other Bookmarks

PERSONAL TECH | What We Learned From Apple's New Privacy L...

By Brian X. Chen

Jan. 27, 2021 Updated 12:53 p.m. ET

We all know that apps collect our data. Yet one of the few ways to find out what an app does with our information involves reading a privacy policy.

Let's be real: Nobody does that.

So late last year, Apple introduced a new requirement for all software developers that publish apps through its App Store. Apps must now include so-called privacy labels, which list the types of data being collected in an easily scannable format. The labels resemble a nutrition marker on food packaging.

These labels, which began appearing in the App Store in December, are the latest attempt by tech designers to make data security easier for all of us to understand. You might be familiar with earlier iterations, like the padlock symbol in a web browser. A locked padlock tells us that a website is trusted, while an unlocked one suggests that a website can be malicious.

The question is whether Apple's new labels will influence the choices people make. "After they read it or look at it, does it change how they use the app or stop them from downloading the app?" asked Stephanie Nguyen, a research scientist who has studied user

Opinion | The GameStop folly

washingtonpost.com/opinio...

Apps | Jordan's Style | https://inf.ethz.ch/... | tome - Synology... | Other Bookmarks

Sections

The Washington Post

Democracy Dies in Darkness

cody@bunta.in

Opinion by **Helaine Olen**

Columnist

Jan. 26, 2021 at 6:11 p.m. EST

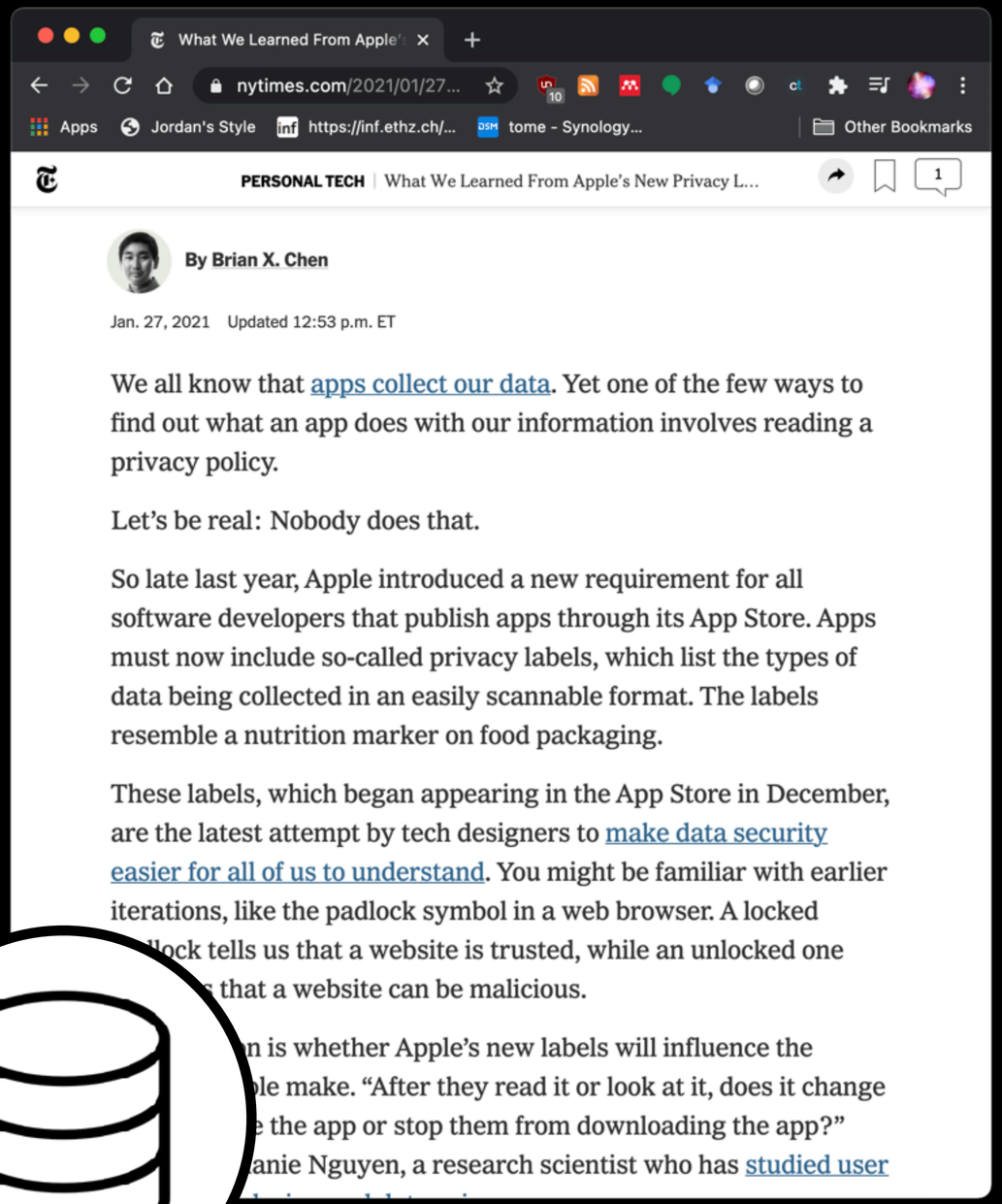
Add to list

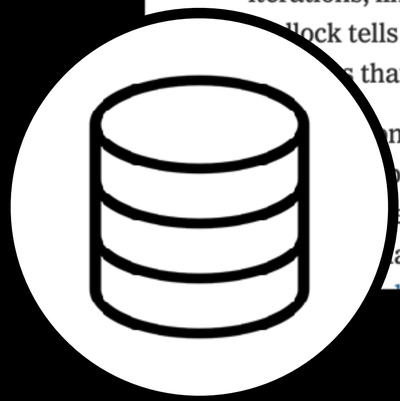
Perhaps the most surprising news this week centers around GameStop, a store that we don't hear about much anymore. The video game retailer's prospects have been dimming for years: Why go to a mall, particularly amid a pandemic, when almost everything you need is available online? Sales are falling and stores are closing. In December, the company told investors it would close 1,000 locations by the end of March. Unsurprisingly, the Wall Street short-sellers are circling.

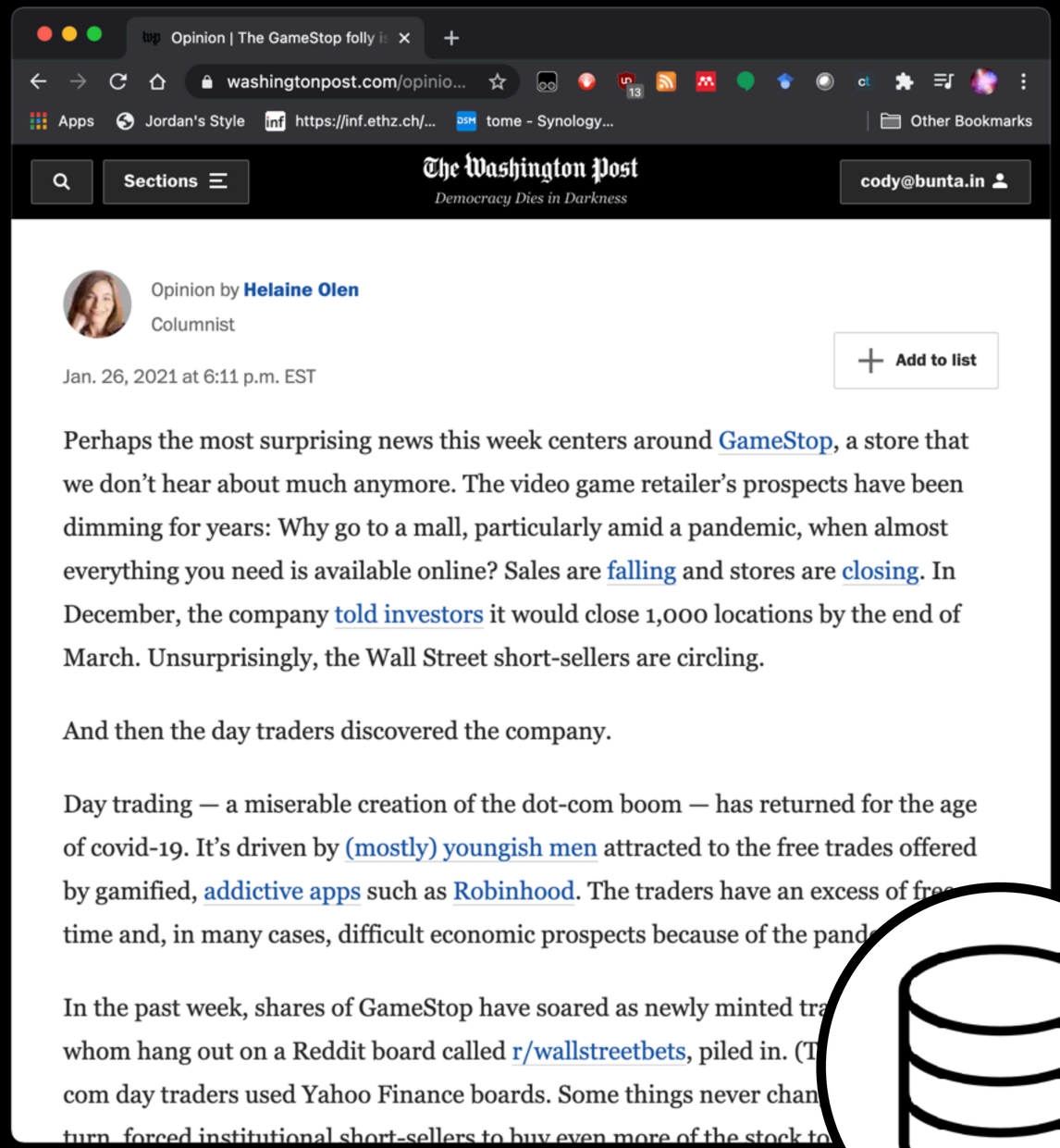
And then the day traders discovered the company.

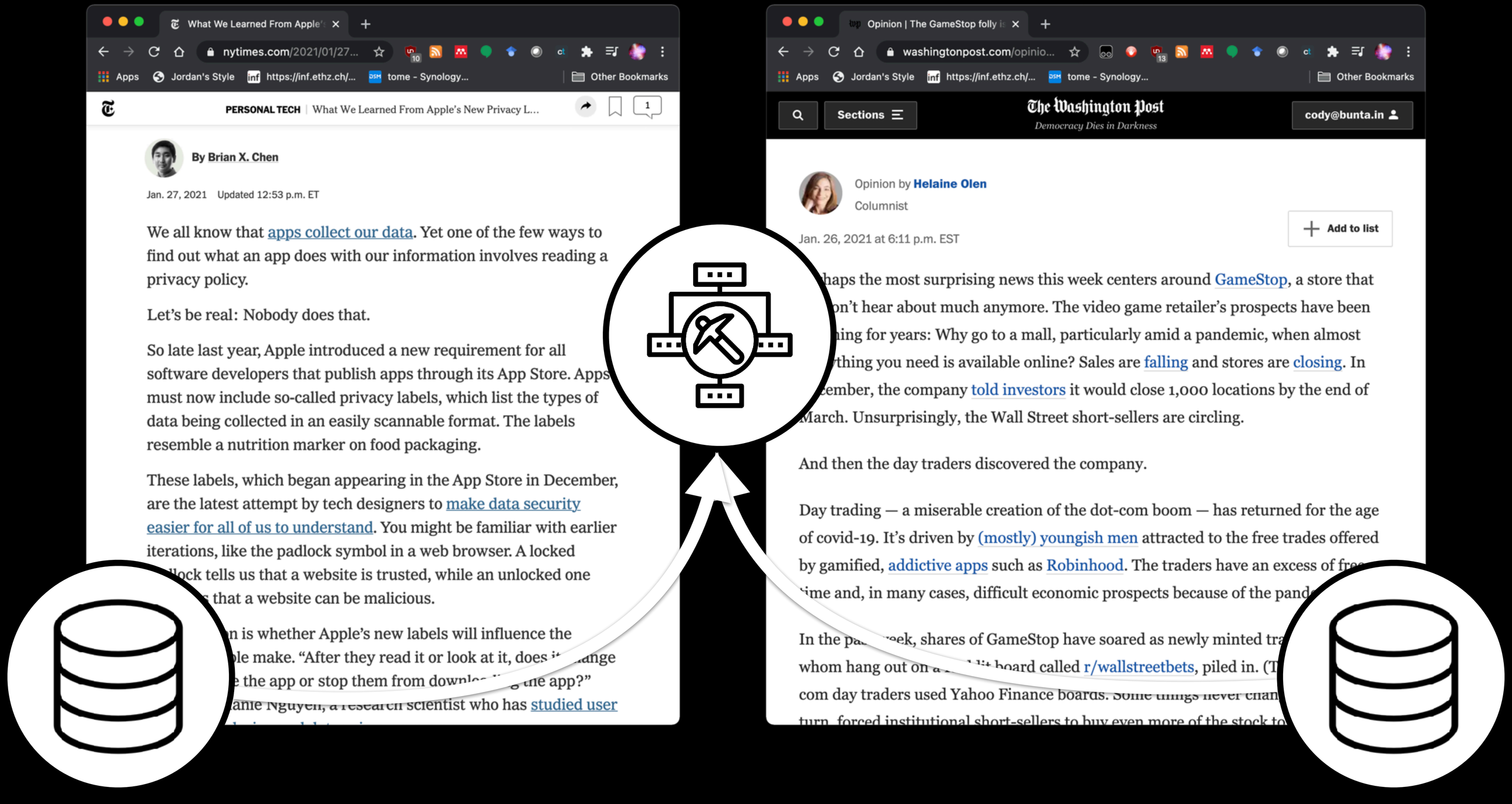
Day trading — a miserable creation of the dot-com boom — has returned for the age of covid-19. It's driven by (mostly) youngish men attracted to the free trades offered by gamified, addictive apps such as Robinhood. The traders have an excess of free time and, in many cases, difficult economic prospects because of the pandemic.

In the past week, shares of GameStop have soared as newly minted traders, many of whom hang out on a Reddit board called r/wallstreetbets, piled in. (The original dot-com day traders used Yahoo Finance boards. Some things never change.) This, in turn, forced institutional short-sellers to buy even more of the stock to avoid bigger

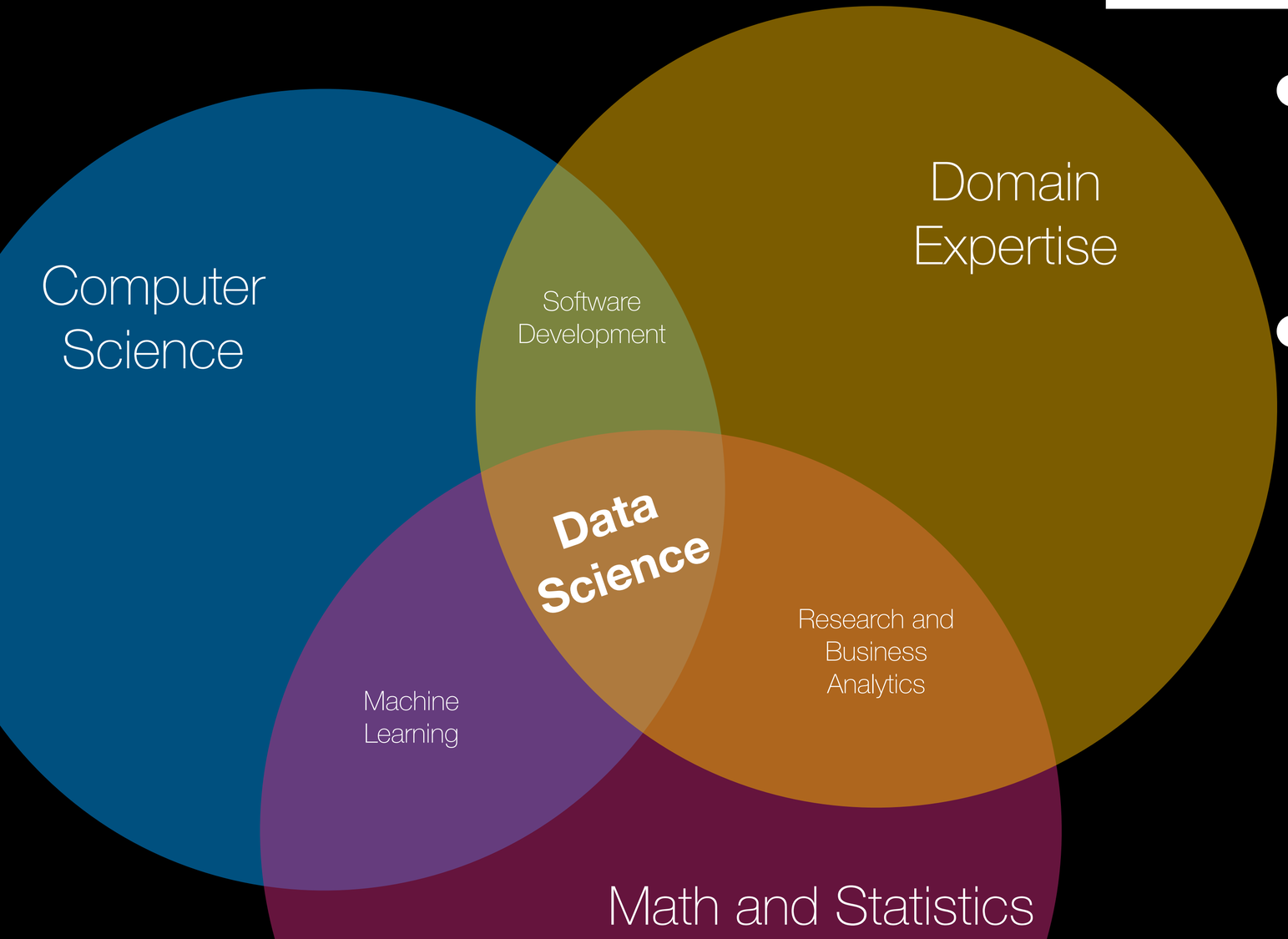

What We Learned From Apple
nytimes.com/2021/01/27...
Apps
Jordan's Style
https://inf.ethz.ch/...
tome - Synology...
Other Bookmarks
PERSONAL TECH | What We Learned From Apple's New Privacy L...
By Brian X. Chen
Jan. 27, 2021 Updated 12:53 p.m. ET
We all know that apps collect our data. Yet one of the few ways to find out what an app does with our information involves reading a privacy policy.
Let's be real: Nobody does that.
So late last year, Apple introduced a new requirement for all software developers that publish apps through its App Store. Apps must now include so-called privacy labels, which list the types of data being collected in an easily scannable format. The labels resemble a nutrition marker on food packaging.
These labels, which began appearing in the App Store in December, are the latest attempt by tech designers to make data security easier for all of us to understand. You might be familiar with earlier iterations, like the padlock symbol in a web browser. A locked
lock tells us that a website is trusted, while an unlocked one
that a website can be malicious.
n is whether Apple's new labels will influence the
le make. "After they read it or look at it, does it change
the app or stop them from downloading the app?"
anie Nguyen, a research scientist who has studied user



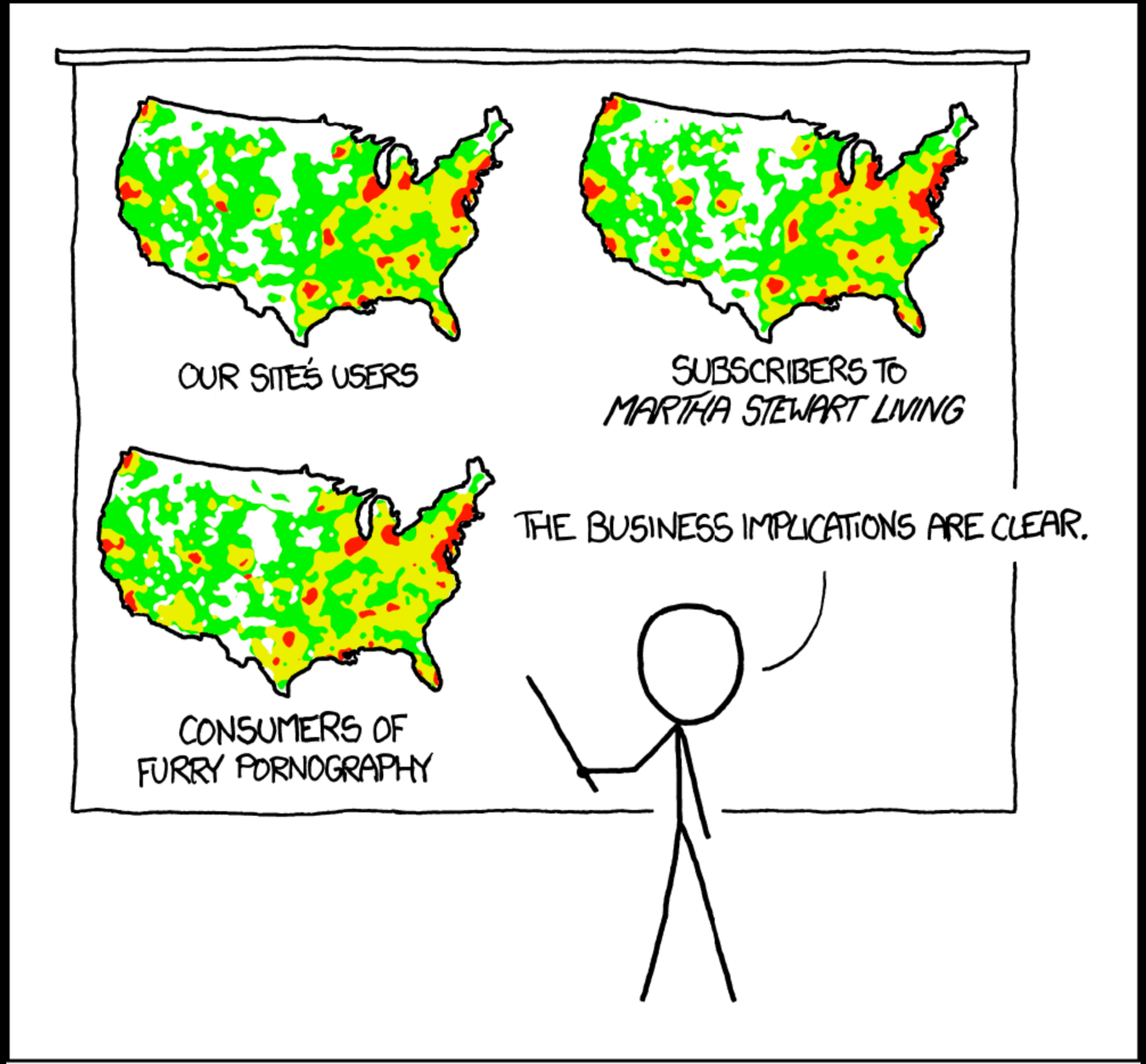

Opinion | The GameStop folly
washingtonpost.com/opinio...
Apps
Jordan's Style
https://inf.ethz.ch/...
tome - Synology...
Other Bookmarks
Sections
The Washington Post
Democracy Dies in Darkness
cody@bunta.in
Opinion by Helaine Olen
Columnist
Jan. 26, 2021 at 6:11 p.m. EST
Add to list
Perhaps the most surprising news this week centers around GameStop, a store that we don't hear about much anymore. The video game retailer's prospects have been dimming for years: Why go to a mall, particularly amid a pandemic, when almost everything you need is available online? Sales are falling and stores are closing. In December, the company told investors it would close 1,000 locations by the end of March. Unsurprisingly, the Wall Street short-sellers are circling.
And then the day traders discovered the company.
Day trading — a miserable creation of the dot-com boom — has returned for the age of covid-19. It's driven by (mostly) youngish men attracted to the free trades offered by gamified, addictive apps such as Robinhood. The traders have an excess of fre
time and, in many cases, difficult economic prospects because of the pand
In the past week, shares of GameStop have soared as newly minted tra
whom hang out on a Reddit board called r/wallstreetbets, piled in. (T
com day traders used Yahoo Finance boards. Some things never chan
turn, forced institutional short-sellers to buy even more of the stock to


What We Learned From Apple
nytimes.com/2021/01/27...
Apps
Jordan's Style
https://inf.ethz.ch/...
tome - Synology...
Other Bookmarks
PERSONAL TECH | What We Learned From Apple's New Privacy L...
By Brian X. Chen
Jan. 27, 2021 Updated 12:53 p.m. ET
We all know that apps collect our data. Yet one of the few ways to find out what an app does with our information involves reading a privacy policy.
Let's be real: Nobody does that.
So late last year, Apple introduced a new requirement for all software developers that publish apps through its App Store. Apps must now include so-called privacy labels, which list the types of data being collected in an easily scannable format. The labels resemble a nutrition marker on food packaging.
These labels, which began appearing in the App Store in December, are the latest attempt by tech designers to make data security easier for all of us to understand. You might be familiar with earlier iterations, like the padlock symbol in a web browser. A locked
lock tells us that a website is trusted, while an unlocked one
that a website can be malicious.
n is whether Apple's new labels will influence the
le make. "After they read it or look at it, does i
ange
the app or stop them from downlo
the app?"
anie Nguyen, a research scientist who has studied user
Opinion | The GameStop folly
washingtonpost.com/opinio...
Apps
Jordan's Style
https://inf.ethz.ch/...
tome - Synology...
Other Bookmarks
Sections
The Washington Post
Democracy Dies in Darkness
cody@bunta.in
Opinion by Helaine Olen
Columnist
Jan. 26, 2021 at 6:11 p.m. EST
Add to list
haps the most surprising news this week centers around GameStop, a store that
on't hear about much anymore. The video game retailer's prospects have been
ing for years: Why go to a mall, particularly amid a pandemic, when almost
thing you need is available online? Sales are falling and stores are closing. In
cember, the company told investors it would close 1,000 locations by the end of
March. Unsurprisingly, the Wall Street short-sellers are circling.
And then the day traders discovered the company.
Day trading — a miserable creation of the dot-com boom — has returned for the age of covid-19. It's driven by (mostly) youngish men attracted to the free trades offered by gamified, addictive apps such as Robinhood. The traders have an excess of fre
ime and, in many cases, difficult economic prospects because of the pand
In the pa
eek, shares of GameStop have soared as newly minted tra
whom hang out on a
dit board called r/wallstreetbets, piled in. (T
com day traders used Yahoo Finance boards. Some things never chan
turn, forced institutional short-sellers to buy even more of the stock to

Data Science ≠ Math/Statistics



- Data science definitely *uses* statistics
- Simple counting may not reveal many useful insights
- And can be misleading

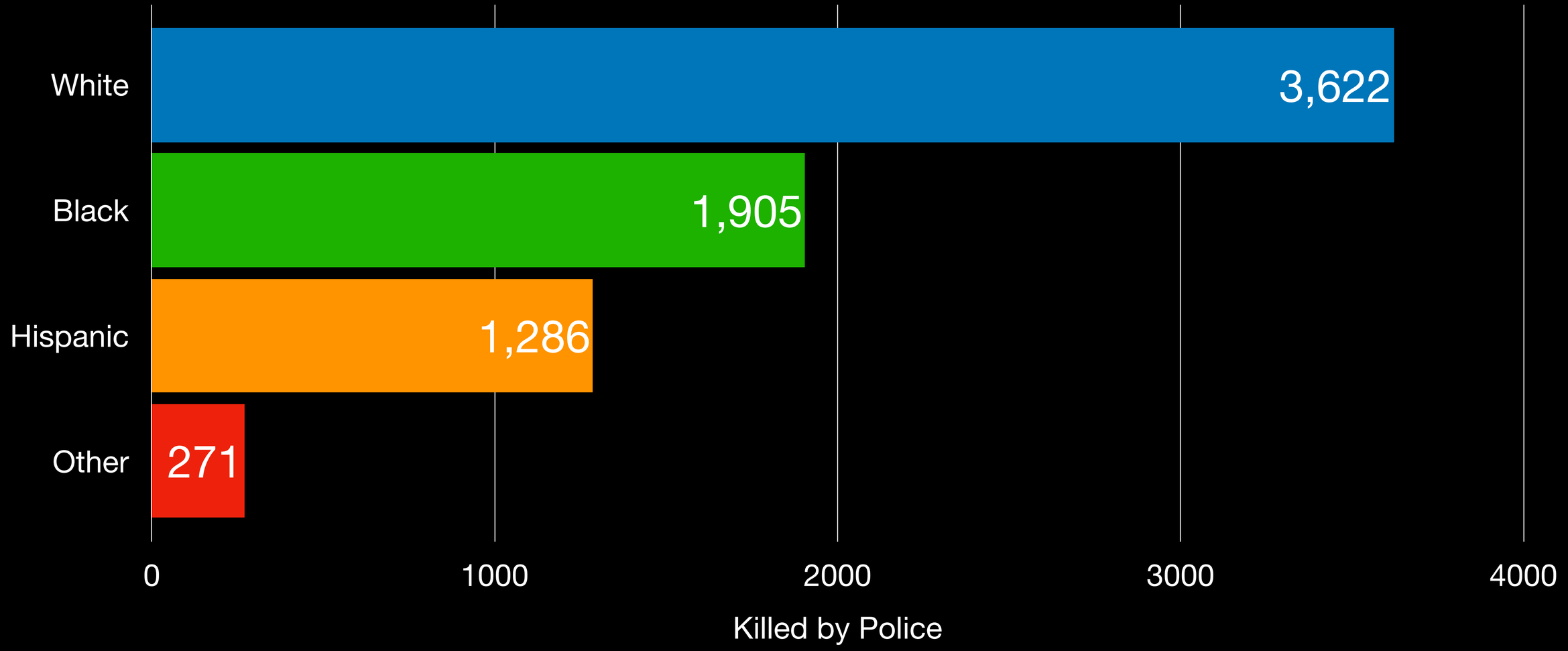


PET PEEVE #208:

GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

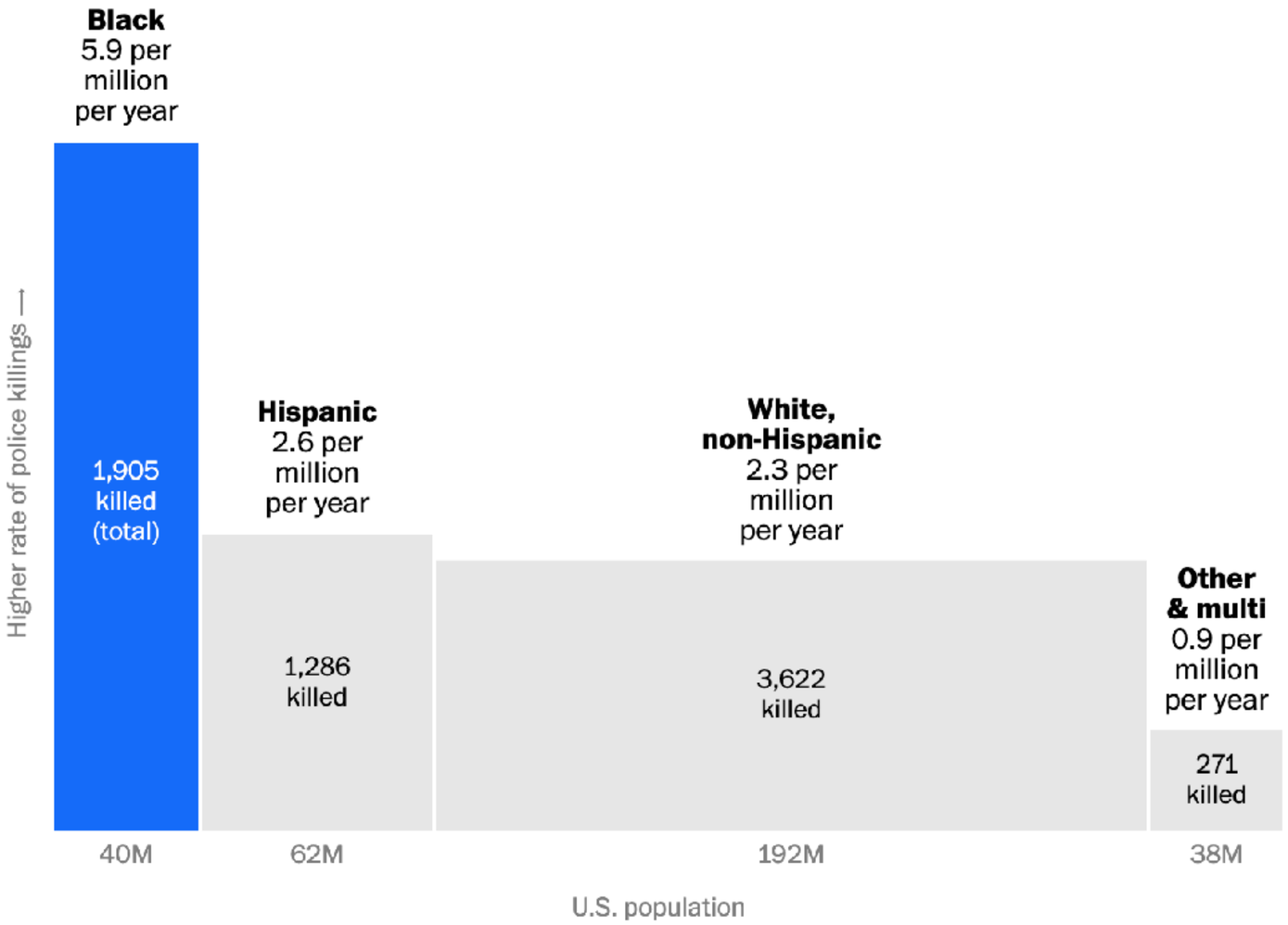
You should first normalize by population

Deaths by Police in the US by race



The Washington Post
Democracy Dies in Darkness

Black Americans are killed at a much higher rate than White Americans

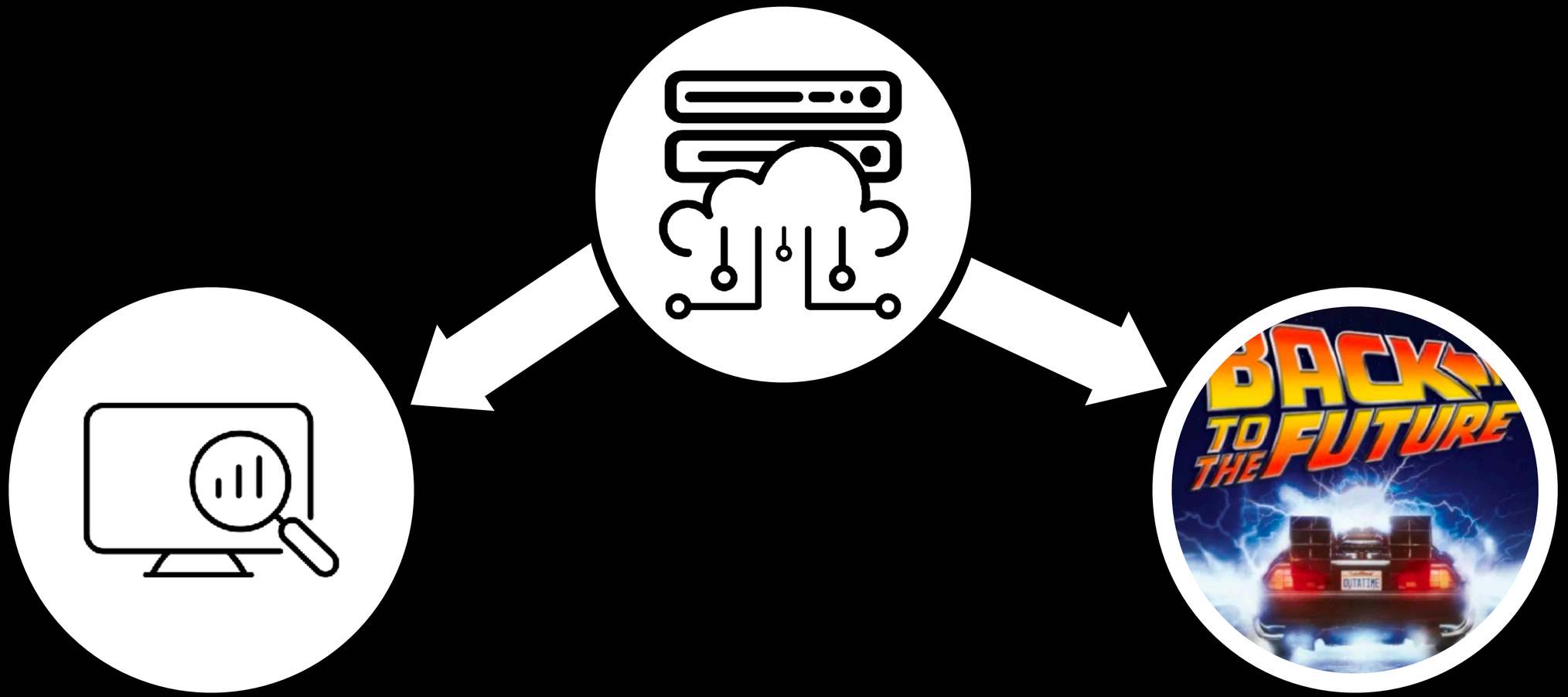


Despite more White Americans being killed by police…

A Black American is far more likely to be killed than White Americans

Requires domain expertise

Data Science Tasks



Description

Prediction

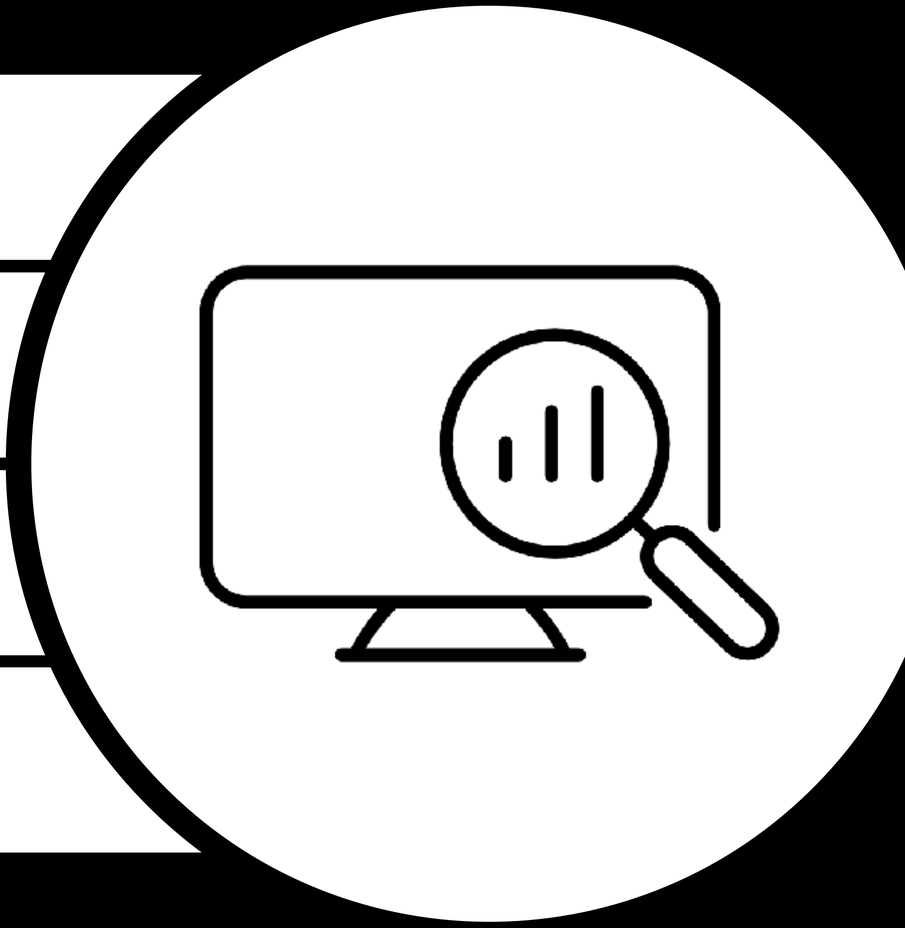
Data Science Tasks – Description

Finding similar instances

Extracting clusters of data

Describing the data's structure

Identifying important elements in the data



Data Science Tasks – Prediction

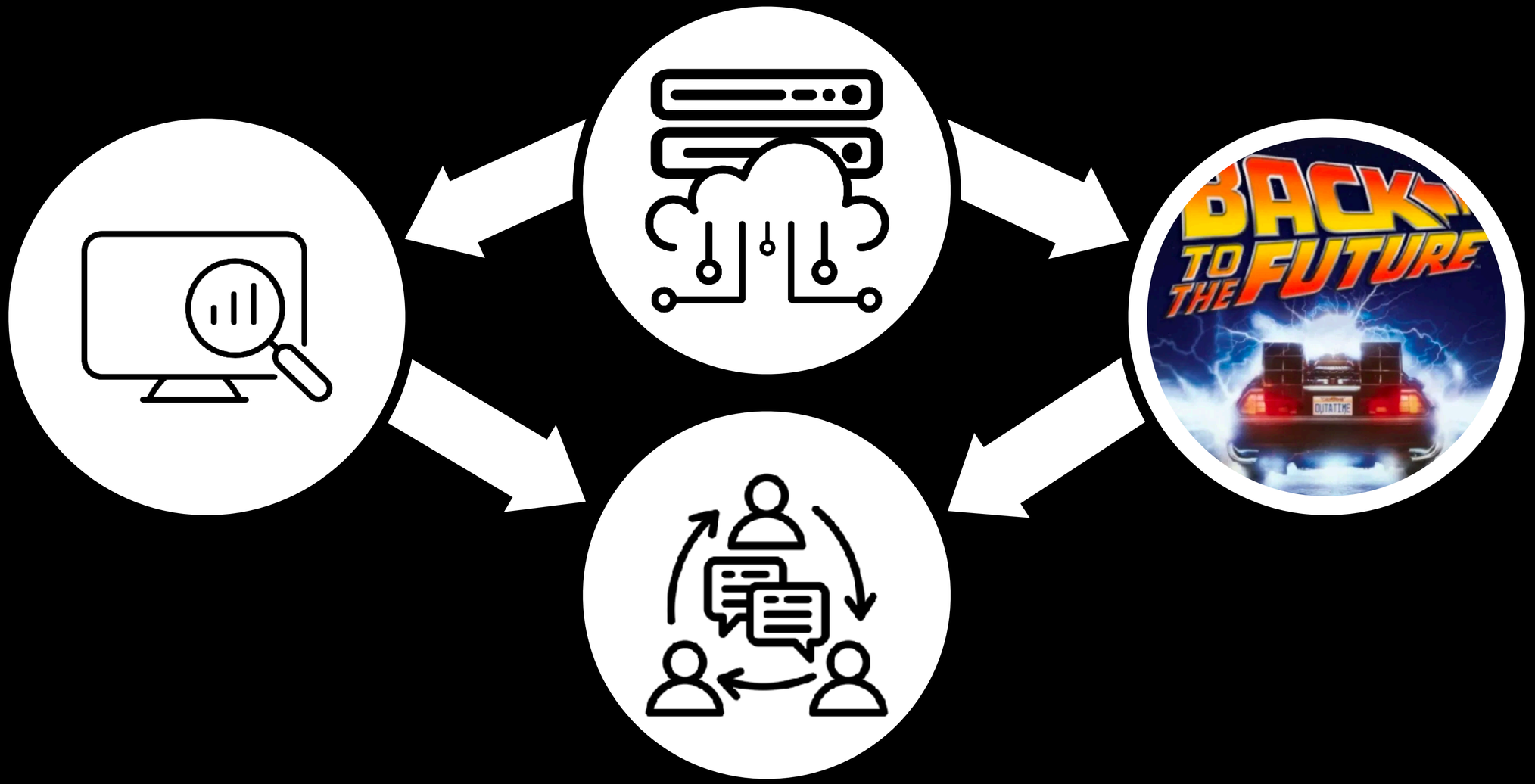
Predicting future values of a metric

Predicting a user's rating for an item

Predicting ads that are likely to be clicked

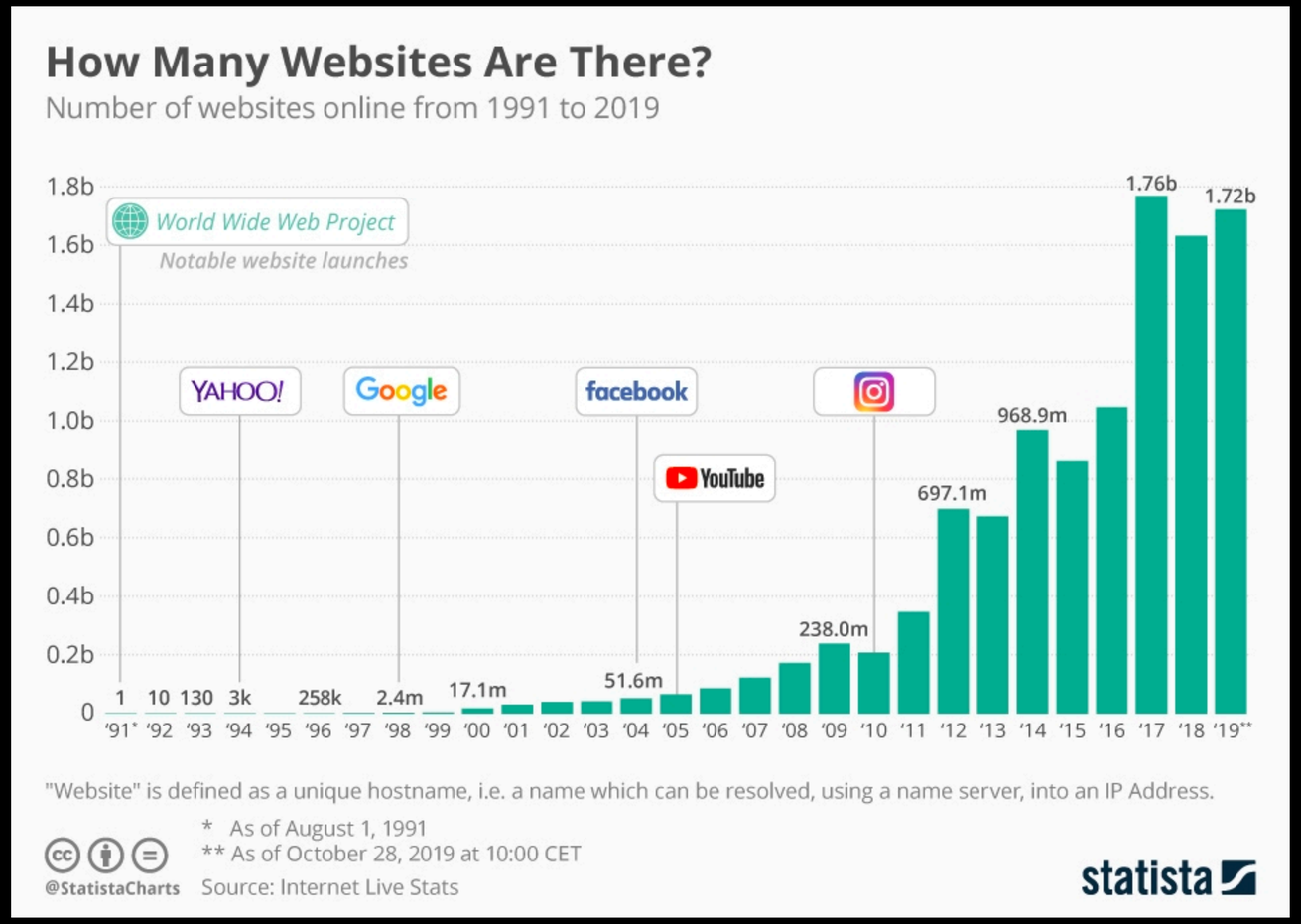
Recommending new items

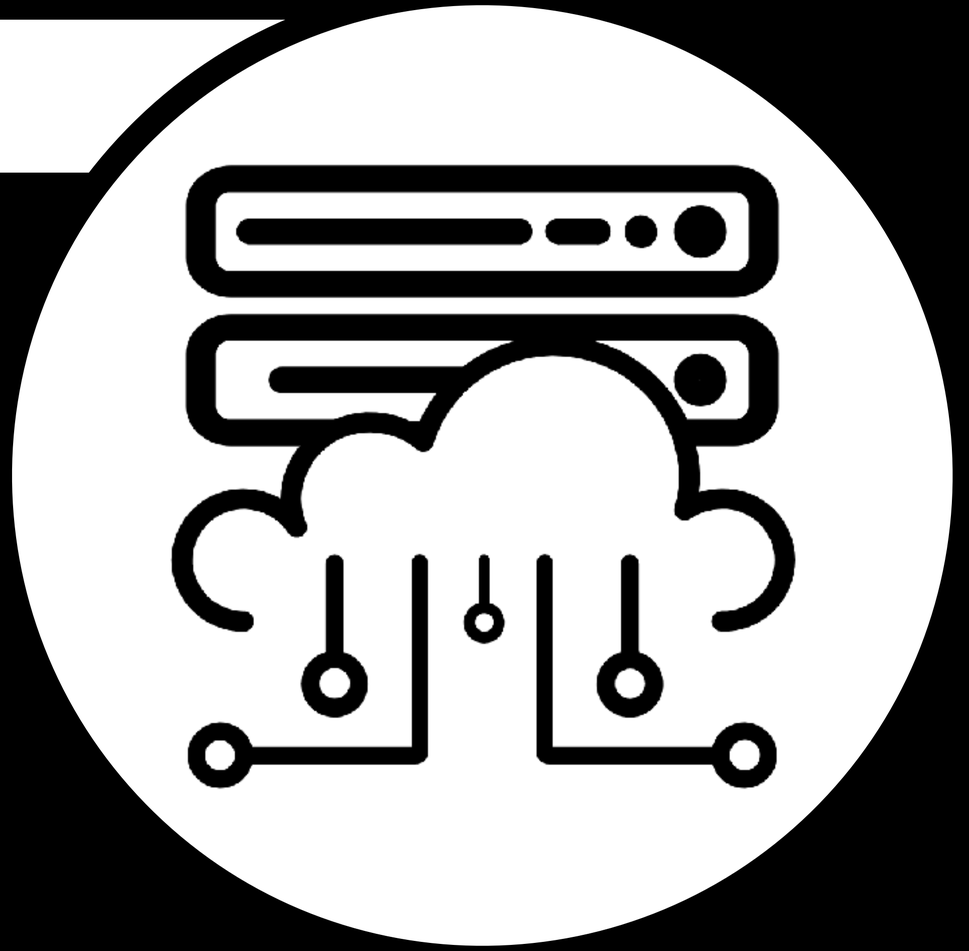
Data Science Tasks



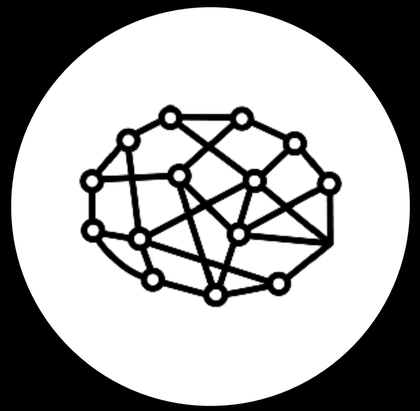
Techniques from data science need to intersect many other disciplines

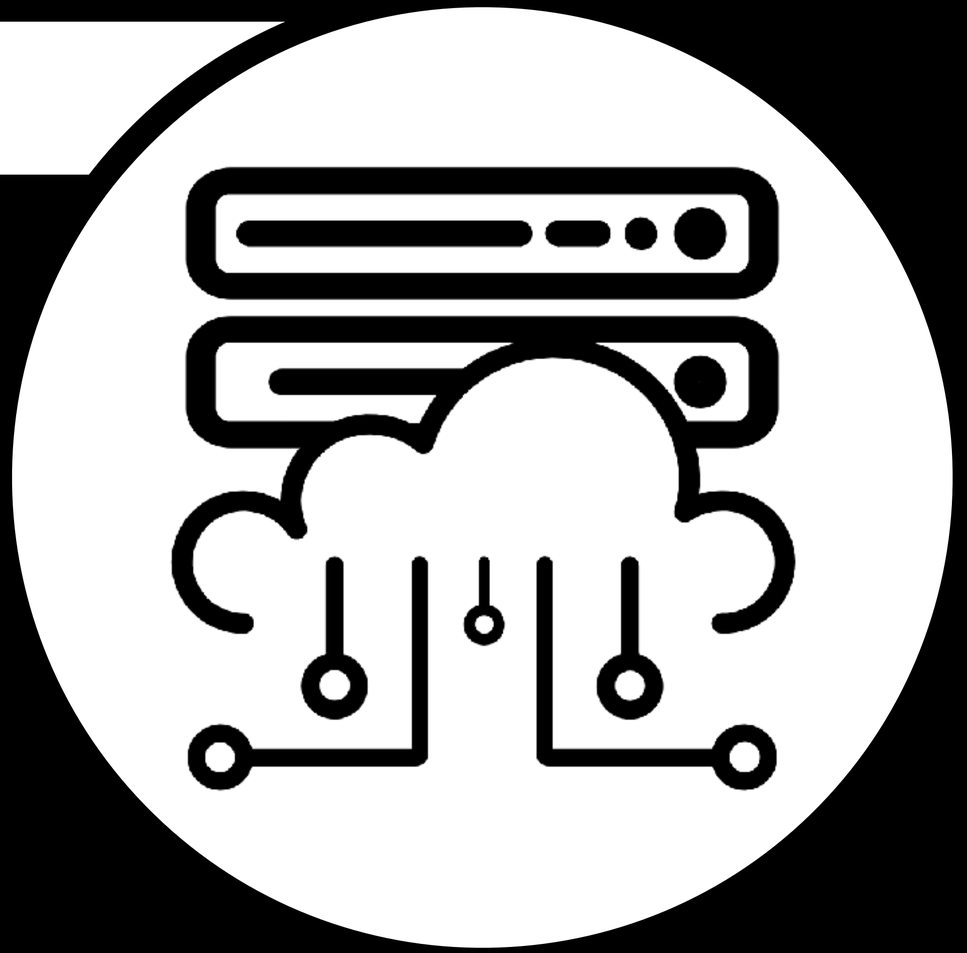
Data can be BIG





Data comes in many different forms





What We Learned From Apple's

nytimes.com/2021/01/27...

Apps | Jordan's Style | https://inf.ethz.ch/... | tome - Synology... | Other Bookmarks

PERSONAL TECH | What We Learned From Apple's New Privacy L...

Jan. 27, 2

We all
find ou
privac

Let's b

So late
softwa
must r
data b
resem

These
are the
easier
iterati
padloc
sugges

The qu
choice
how th
asked

WeRateDogs®
13.3K Tweets

Follow

WeRateDogs®
@dog_rates

Your Only Source For Professional Dog Ratings Instagram and Facebook ➜
WeRateDogs partnerships@weratedogs.com

links and things ➜ campsite.bio/weratedogs Joined November 2015

18 Following 8.9M Followers

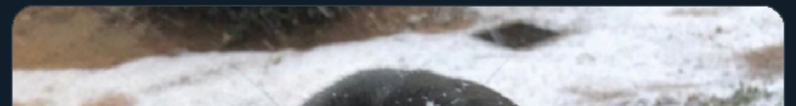
Followed by Iain | he/him/his, Soya Park, and 54 others you follow

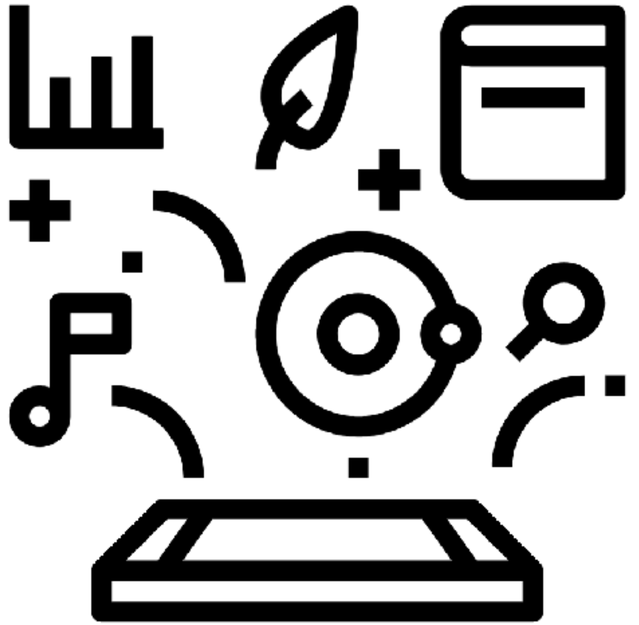
Tweets | Tweets & replies | Media | Likes

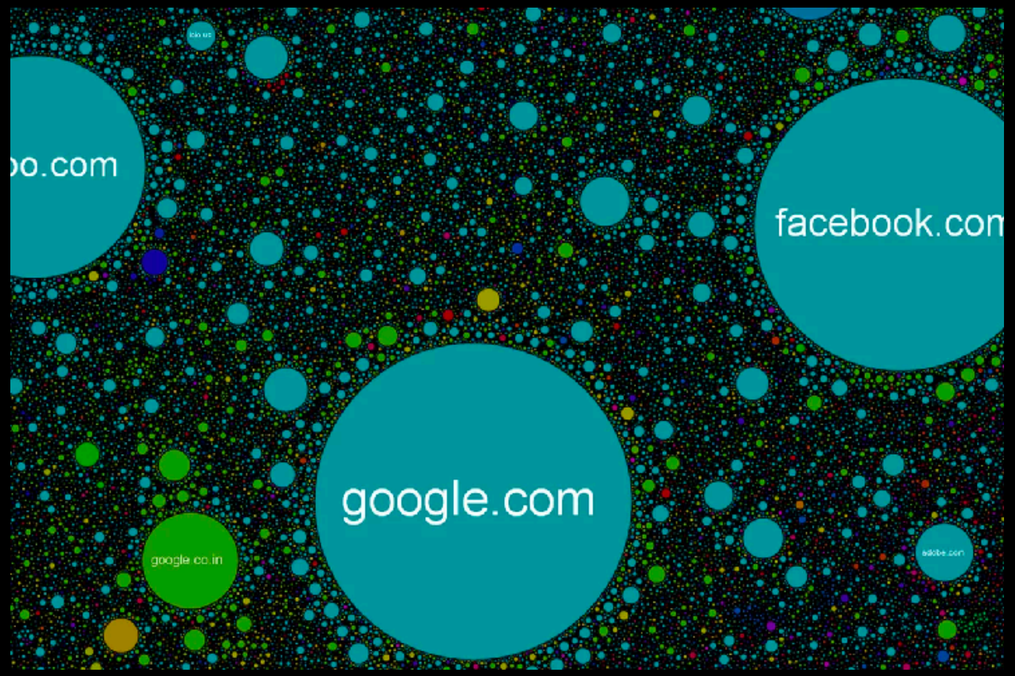
Pinned Tweet

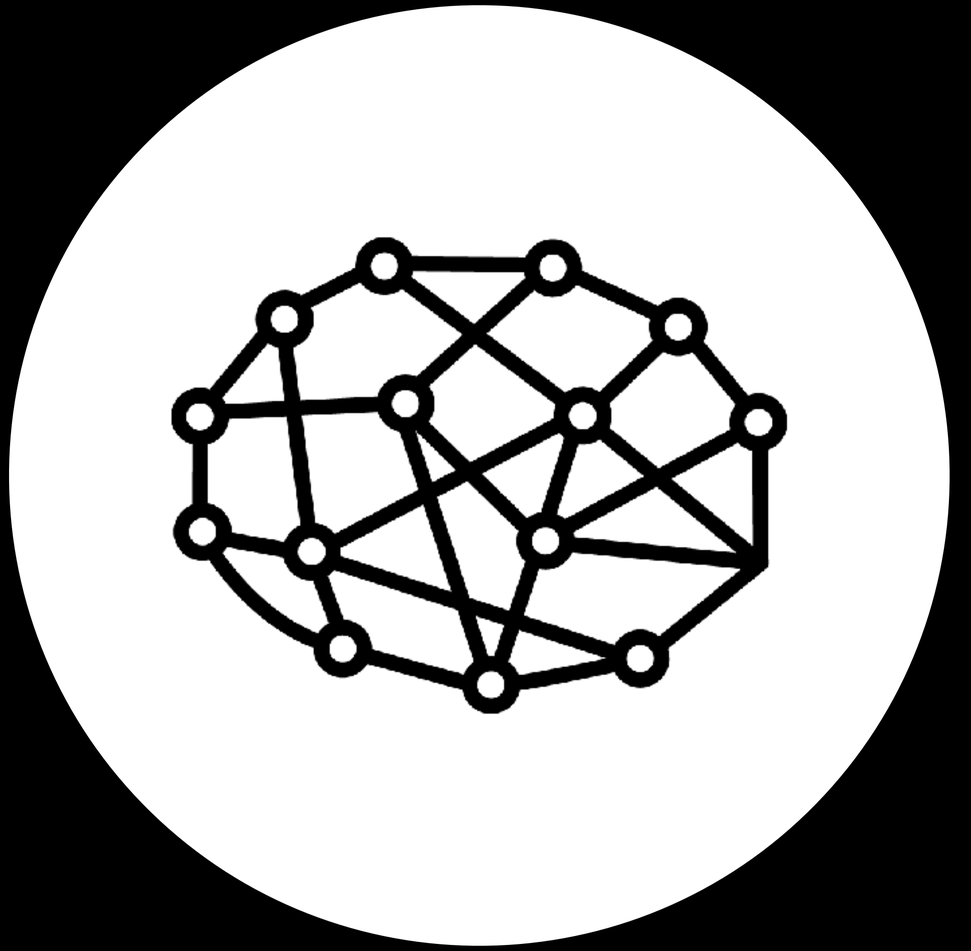
WeRateDogs® @dog_rates · Dec 8, 2020

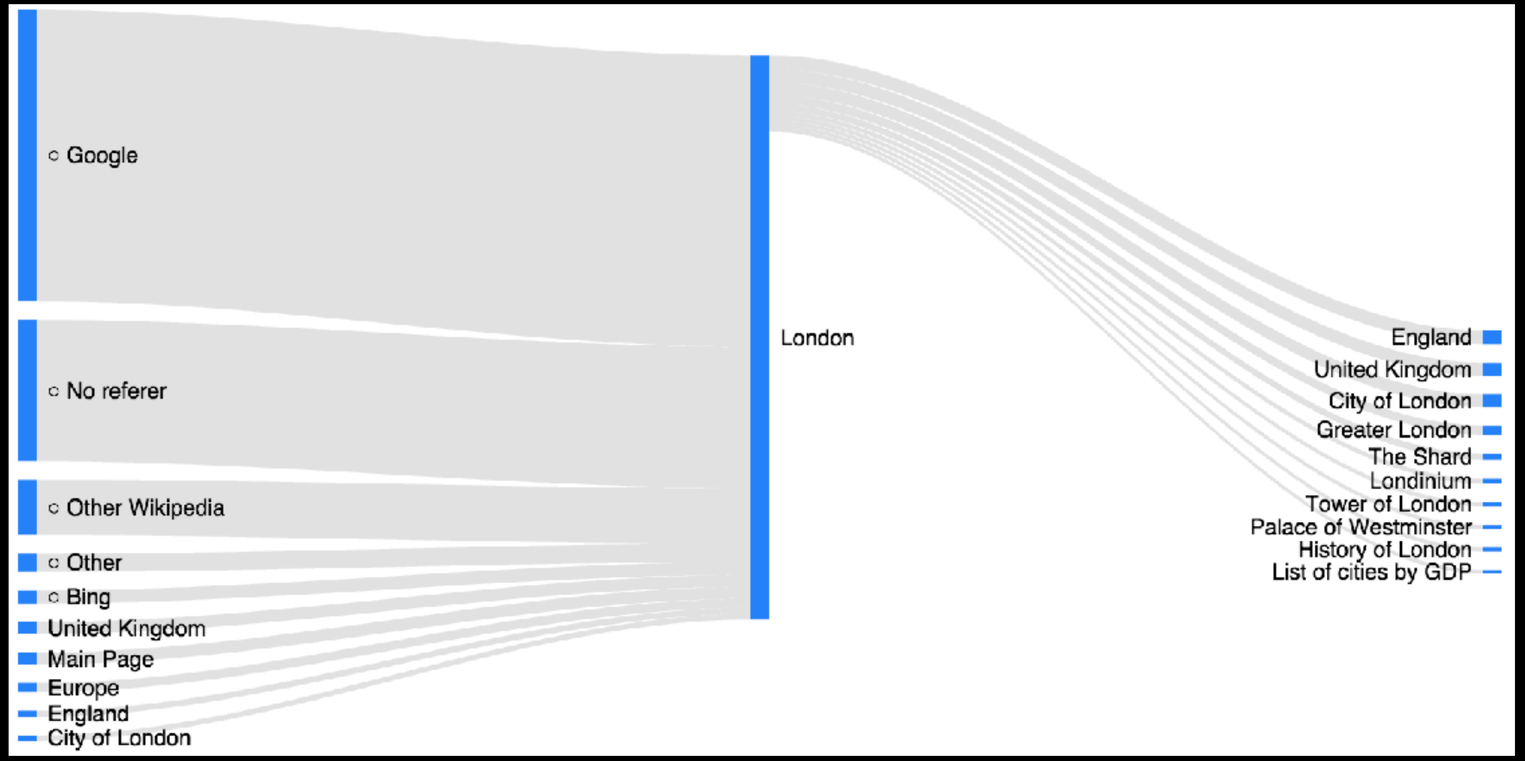
This is Ollie. He's just here to get your attention. If you're in the US and would like us to send a dog pic to your phone every single day, text "🐶" to 213-212-6731

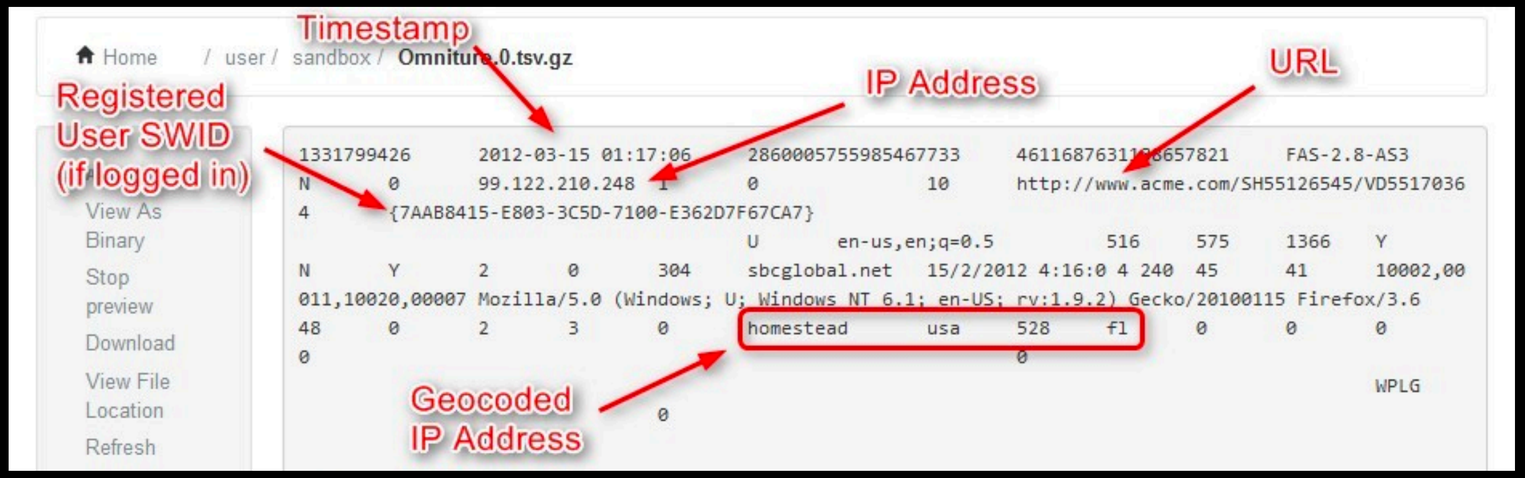





facebook.com
google.com
google.co.in



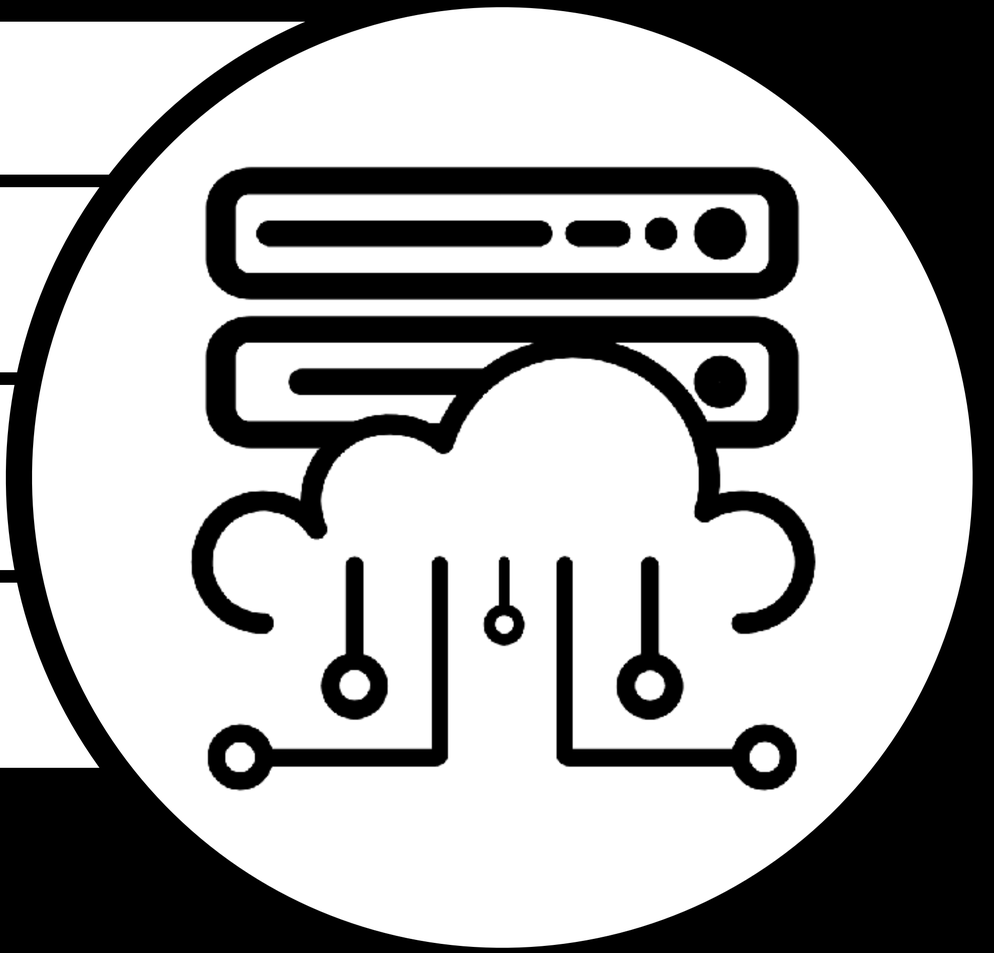

Google
No referer
Other Wikipedia
Other
Bing
United Kingdom
Main Page
Europe
England
City of London
London
England
United Kingdom
City of London
Greater London
The Shard
Londinium
Tower of London
Palace of Westminster
History of London
List of cities by GDP


Home / user / sandbox / Omniture.0.tsv.gz
Timestamp
IP Address
URL
Registered User SWID (if logged in)
View As Binary
Stop preview
Download
View File Location
Refresh
1331799426 2012-03-15 01:17:06 2860005755985467733 4611687631108657821 FAS-2.8-AS3
N 0 99.122.210.248 1 0 10 http://www.acme.com/SH55126545/VD5517036
4 {7AAB8415-E803-3C5D-7100-E362D7F67CA7}
U en-us,en;q=0.5 516 575 1366 Y
N Y 2 0 304 sbcglobal.net 15/2/2012 4:16:0 4 240 45 41 10002,00
011,10020,00007 Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
48 0 2 3 0 homestead usa 528 fl 0 0 0
0 0
WPLG
0
Geocoded IP Address


LOG

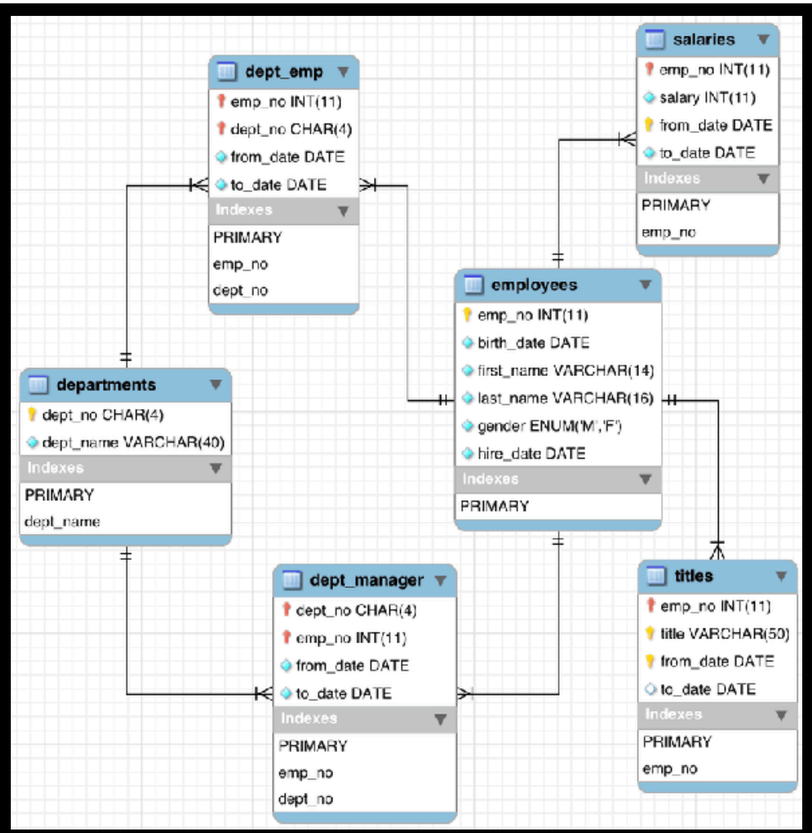
Data varies in structure

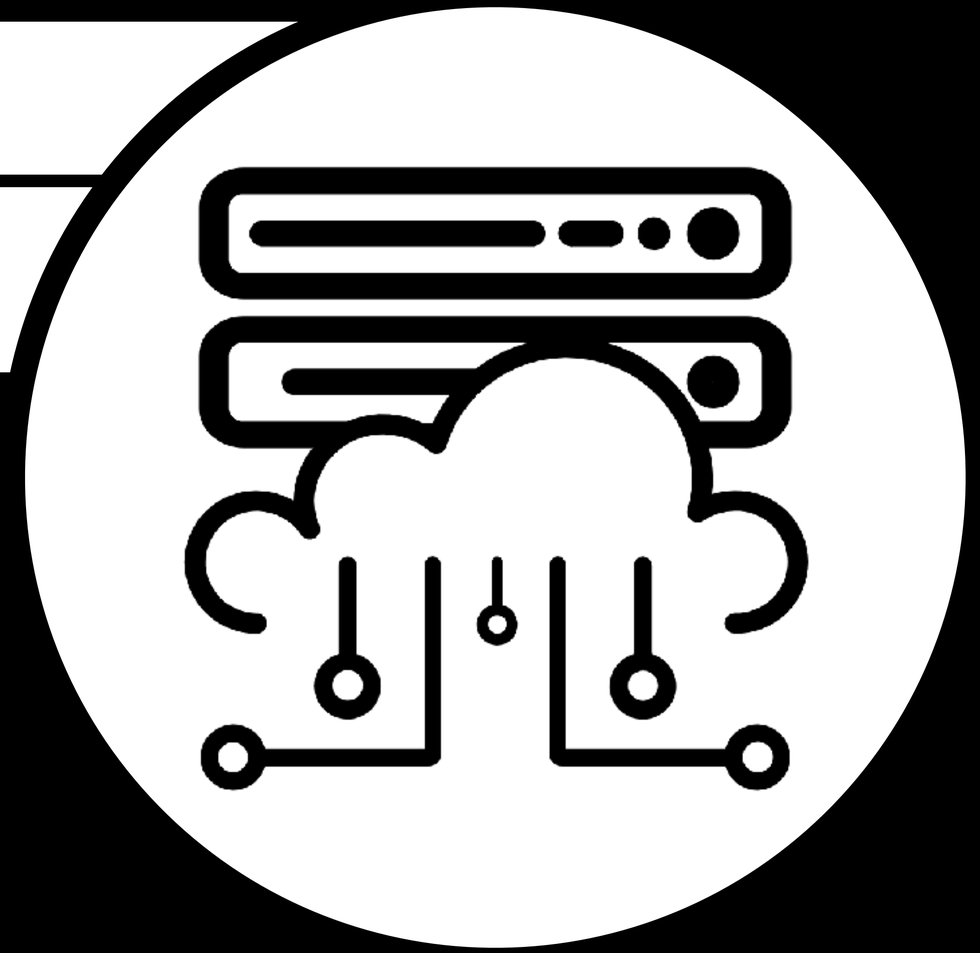
- Structured data (databases)
- Unstructured data (raw text, images)
- Semi-structured data (JSON, XML, CSV)



Data varies in structure

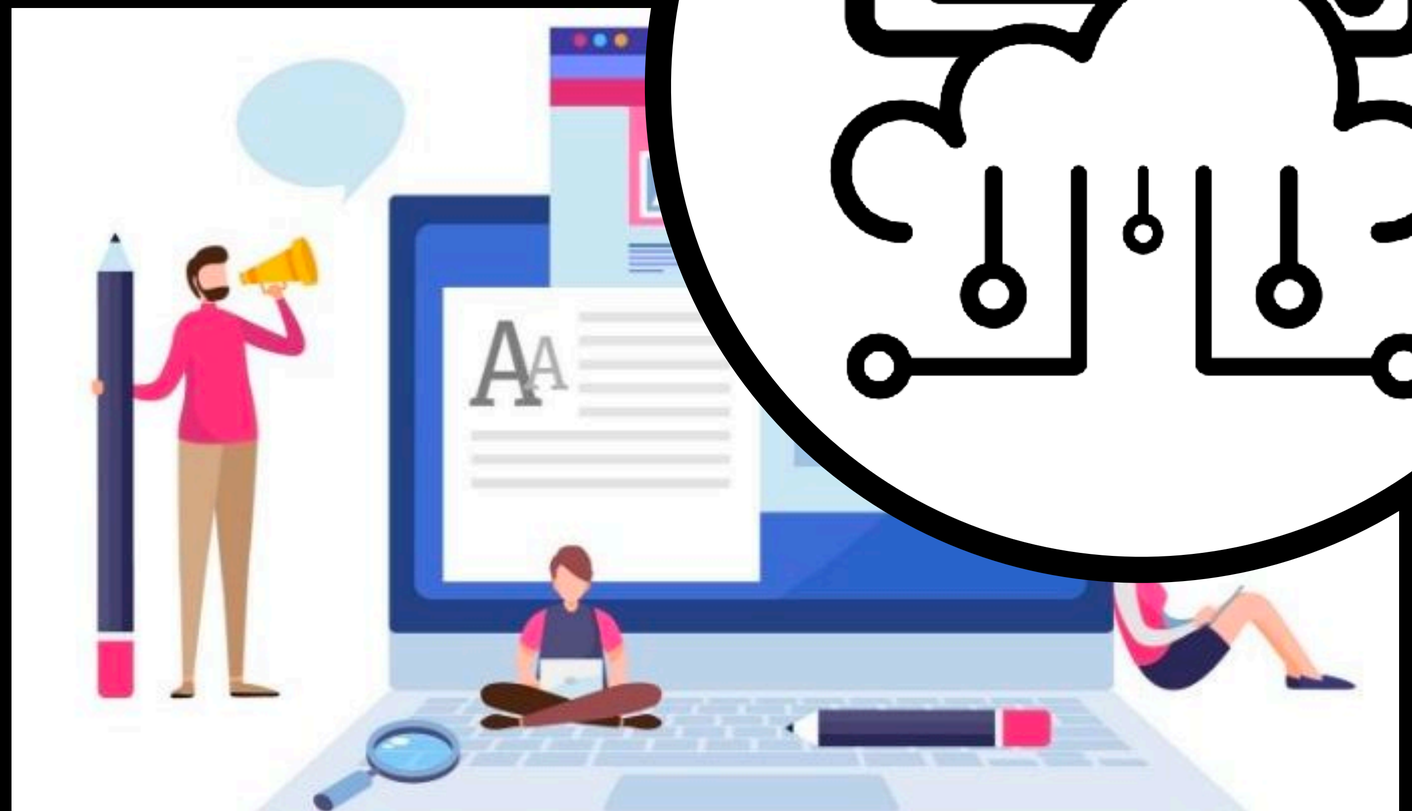
Structured data (databases)





Data varies in structure

Unstructured data (raw text, images)

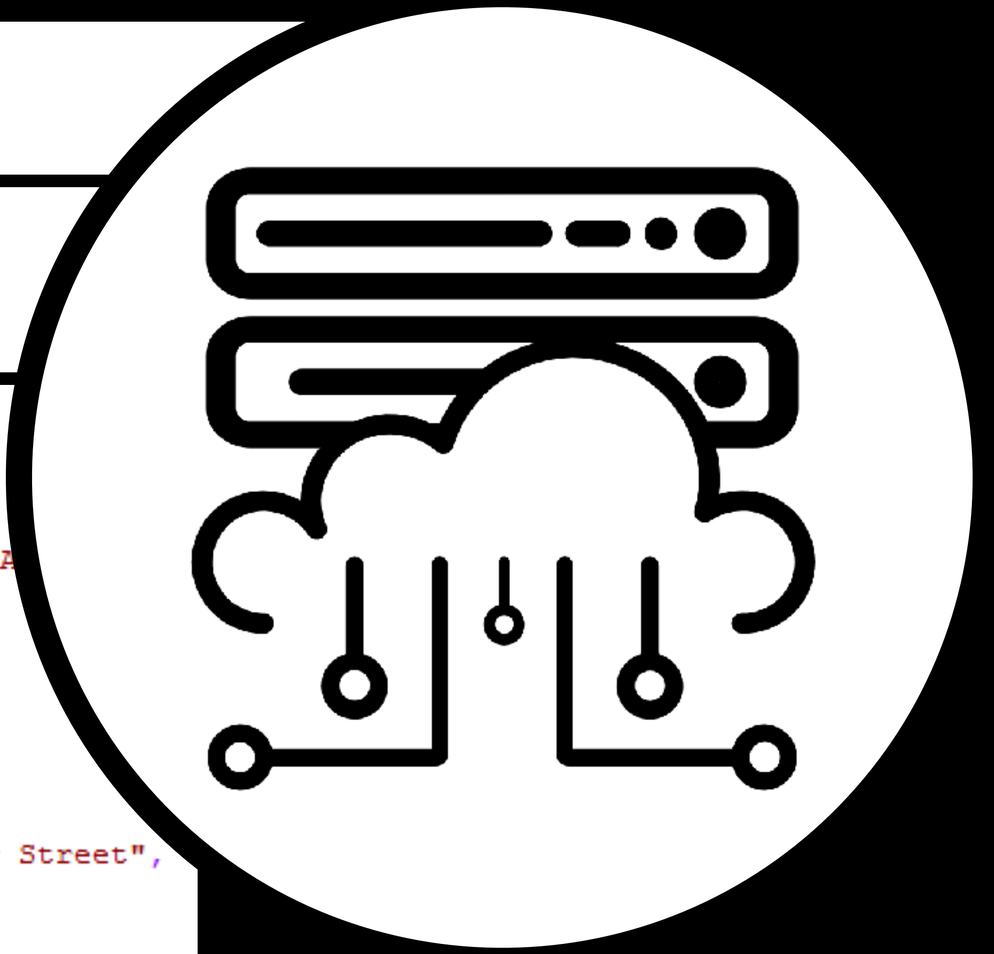


Data varies in structure

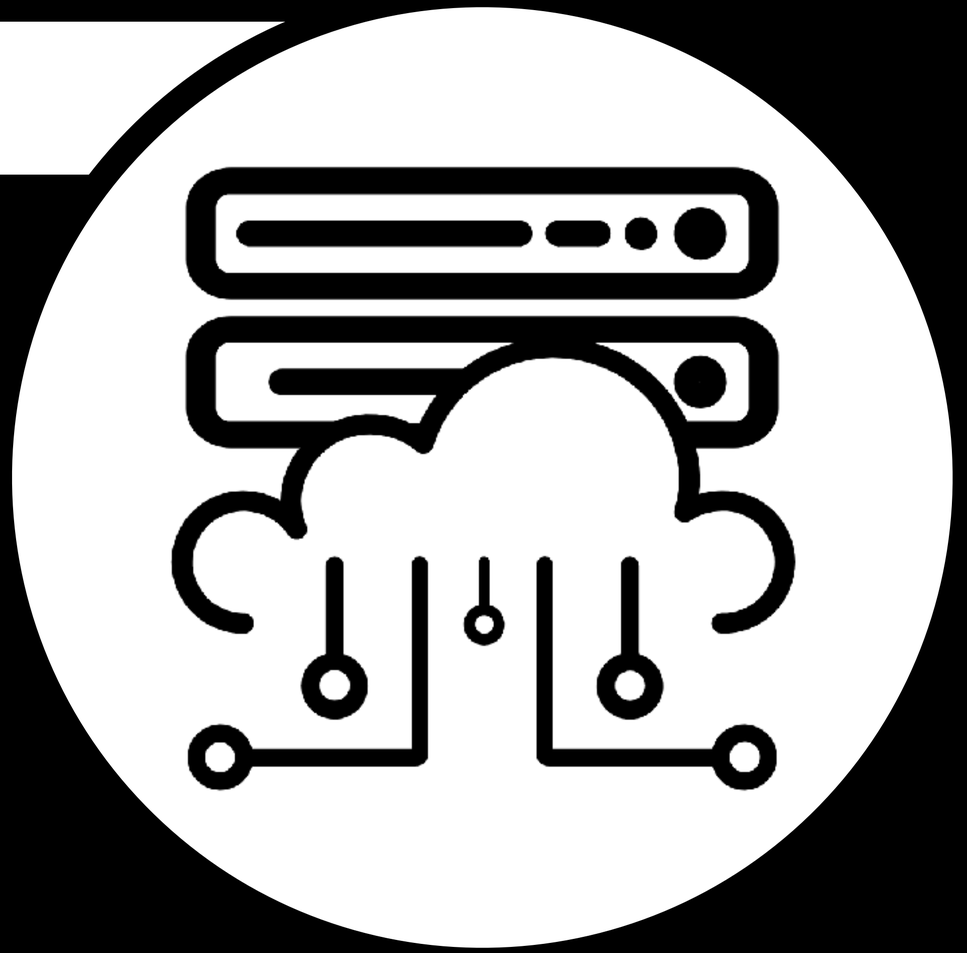
Semi-Structured data (JSON, XML, CSVs)

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!-- Edited by XMLSpy® -->
- <CATALOG>
  - <CD>
      <TITLE>Empire Burlesque</TITLE>
      <ARTIST>Bob Dylan</ARTIST>
      <COUNTRY>USA</COUNTRY>
      <COMPANY>Columbia</COMPANY>
      <PRICE>10.90</PRICE>
      <YEAR>1985</YEAR>
    </CD>
  - <CD>
      <TITLE>Hide your heart</TITLE>
      <ARTIST>Bonnie Tyler</ARTIST>
      <COUNTRY>UK</COUNTRY>
      <COMPANY>CBS Records</COMPANY>
      <PRICE>9.90</PRICE>
      <YEAR>1988</YEAR>
    </CD>
  - <CD>
      <TITLE>Greatest Hits</TITLE>
      <ARTIST>Dolly Parton</ARTIST>
      <COUNTRY>USA</COUNTRY>
      <COMPANY>RCA</COMPANY>
      <PRICE>9.90</PRICE>
      <YEAR>1982</YEAR>
    </CD>
  - <CD>
      <TITLE>Still got the blues</TITLE>
      <ARTIST>Gary Moore</ARTIST>
      <COUNTRY>UK</COUNTRY>
      <COMPANY>Virgin records</COMPANY>
      <PRICE>10.20</PRICE>
      <YEAR>1990</YEAR>
    </CD>
  - <CD>
```

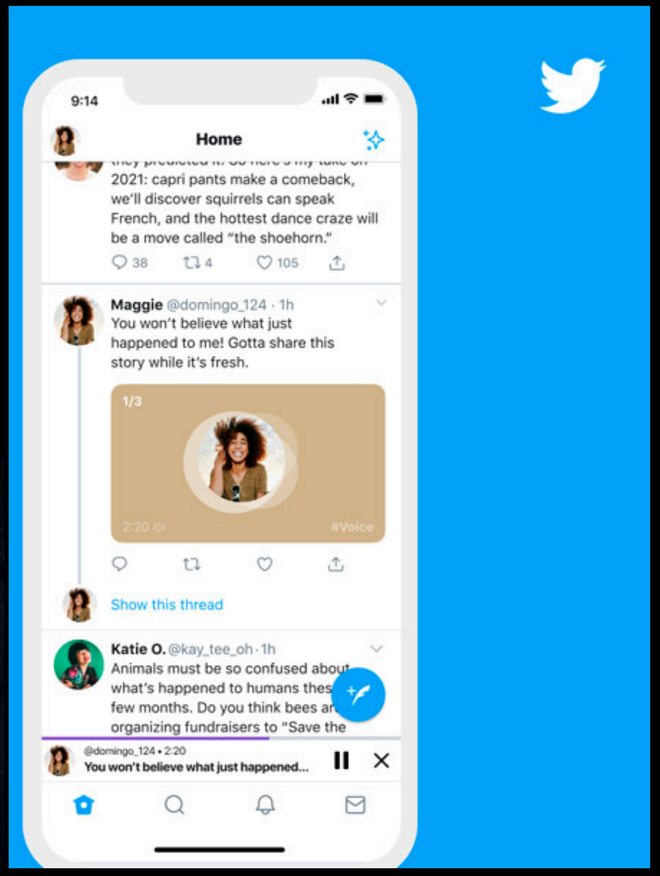
```
{
    "orders": [
        {
            "orderno": "748745375",
            "date": "June 30, 2088 1:54:23 A
            "trackingno": "TN0039291",
            "custid": "11045",
            "customer": [
                {
                    "custid": "11045",
                    "fname": "Sue",
                    "lname": "Hatfield",
                    "address": "1409 Silver Street",
                    "city": "Ashland",
                    "state": "NE",
                    "zip": "68003"
                }
            ]
        }
    ]
}
```

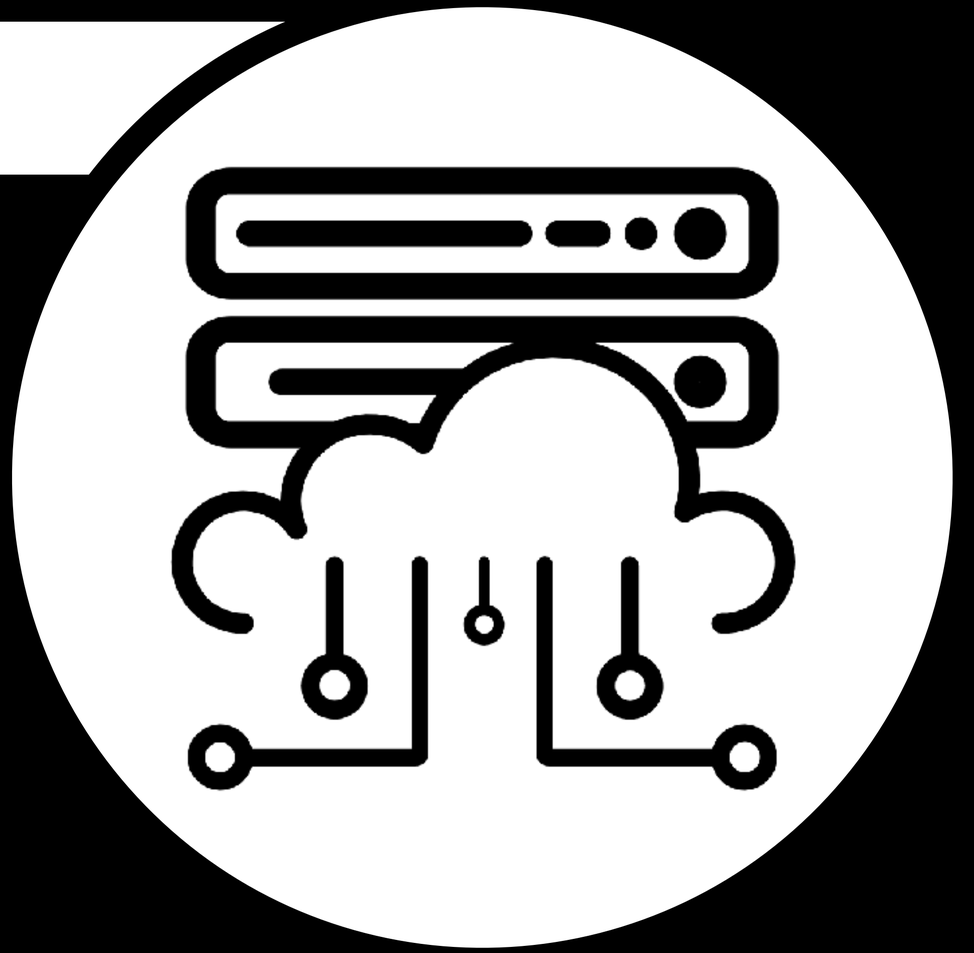


Data sources are dynamic, constantly evolving



Data can come in never-ending *streams*



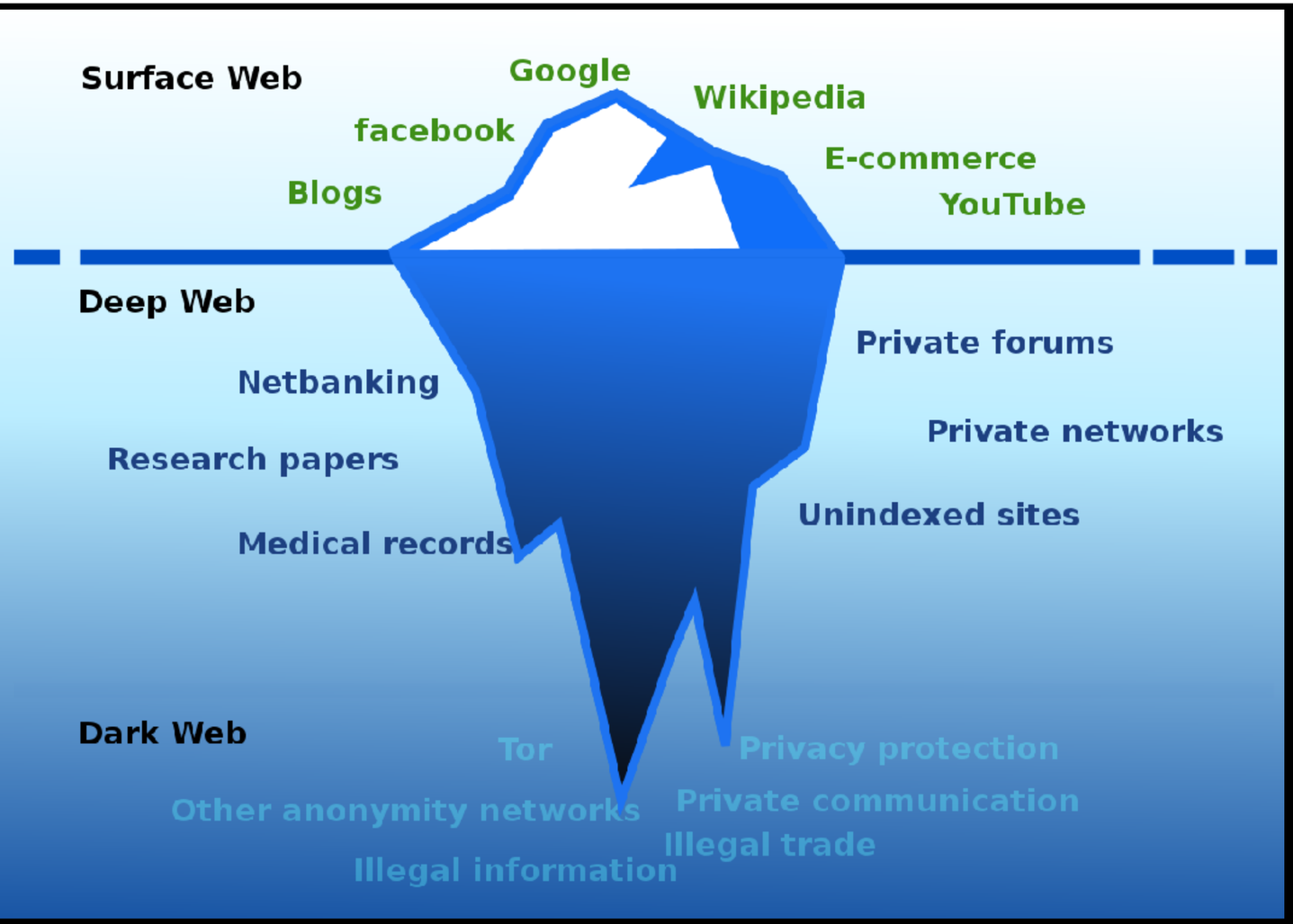


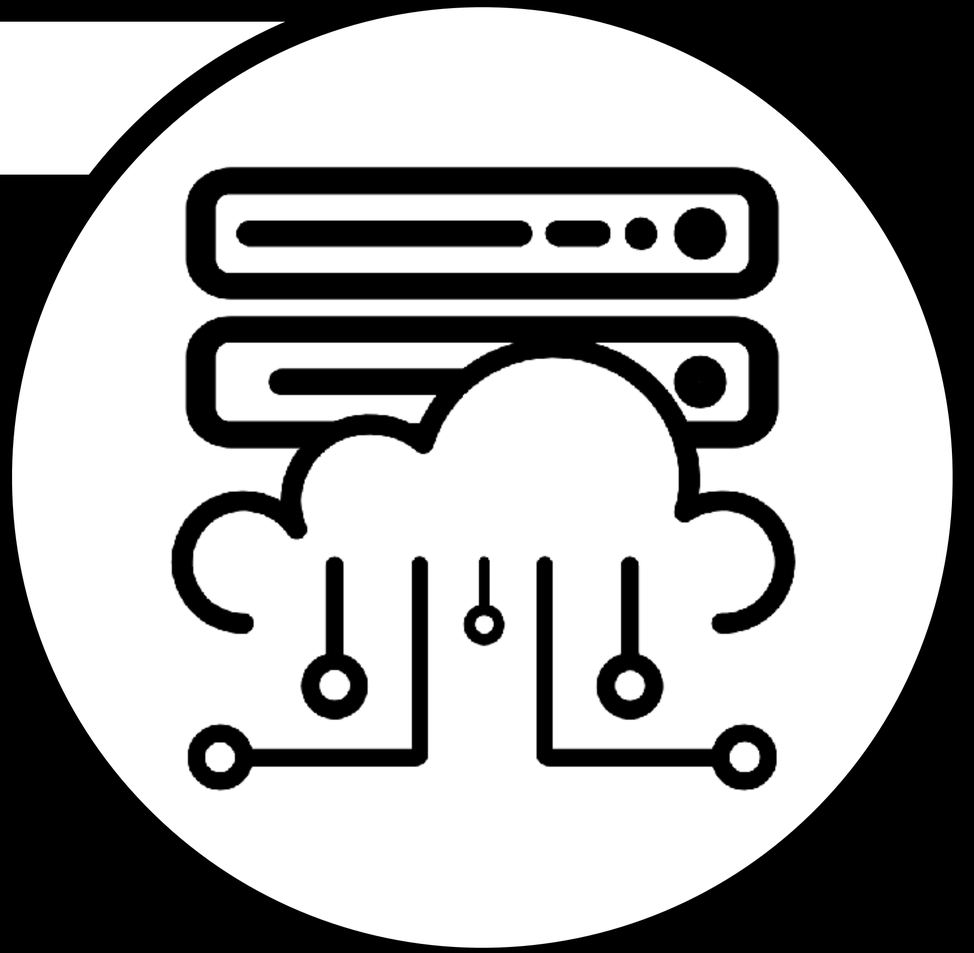
Why might streams make things different?

What is the average in an infinite stream of numbers?

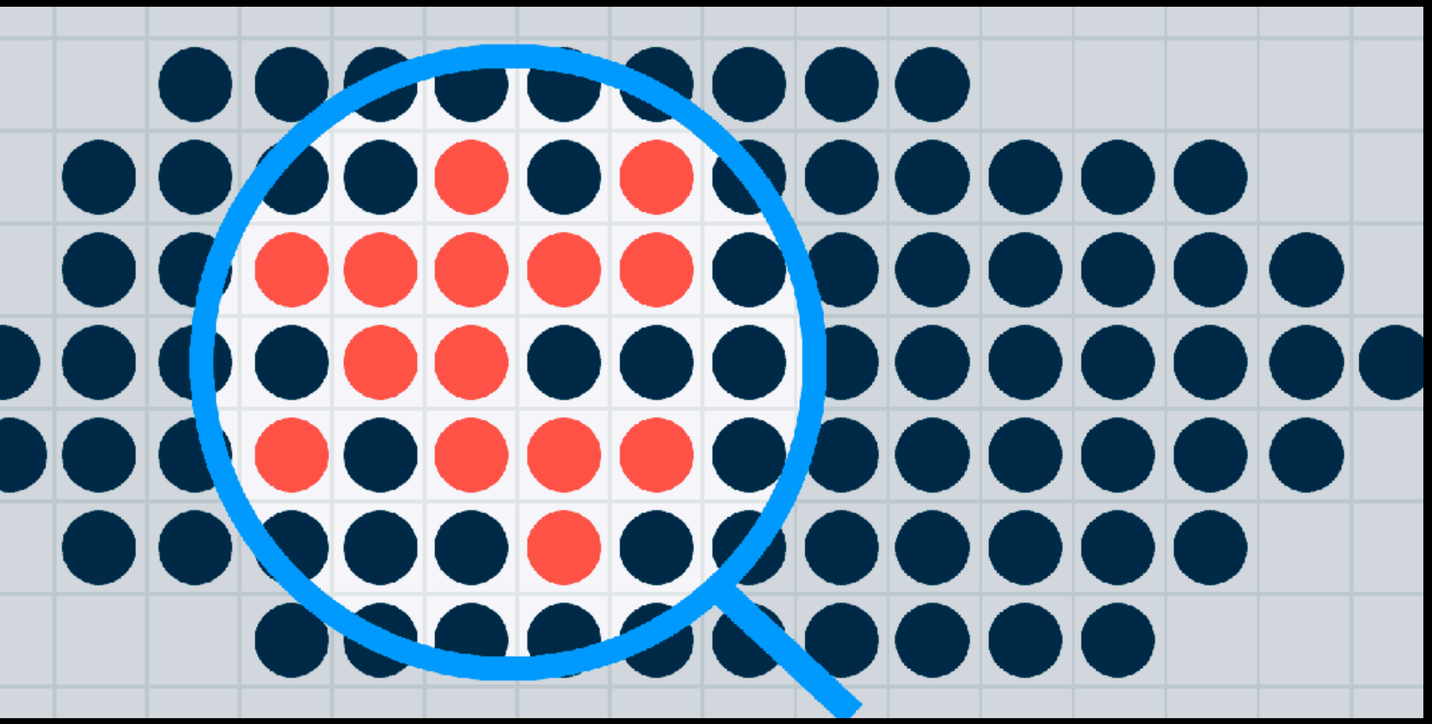
How do you create a sample from an infinite stream?

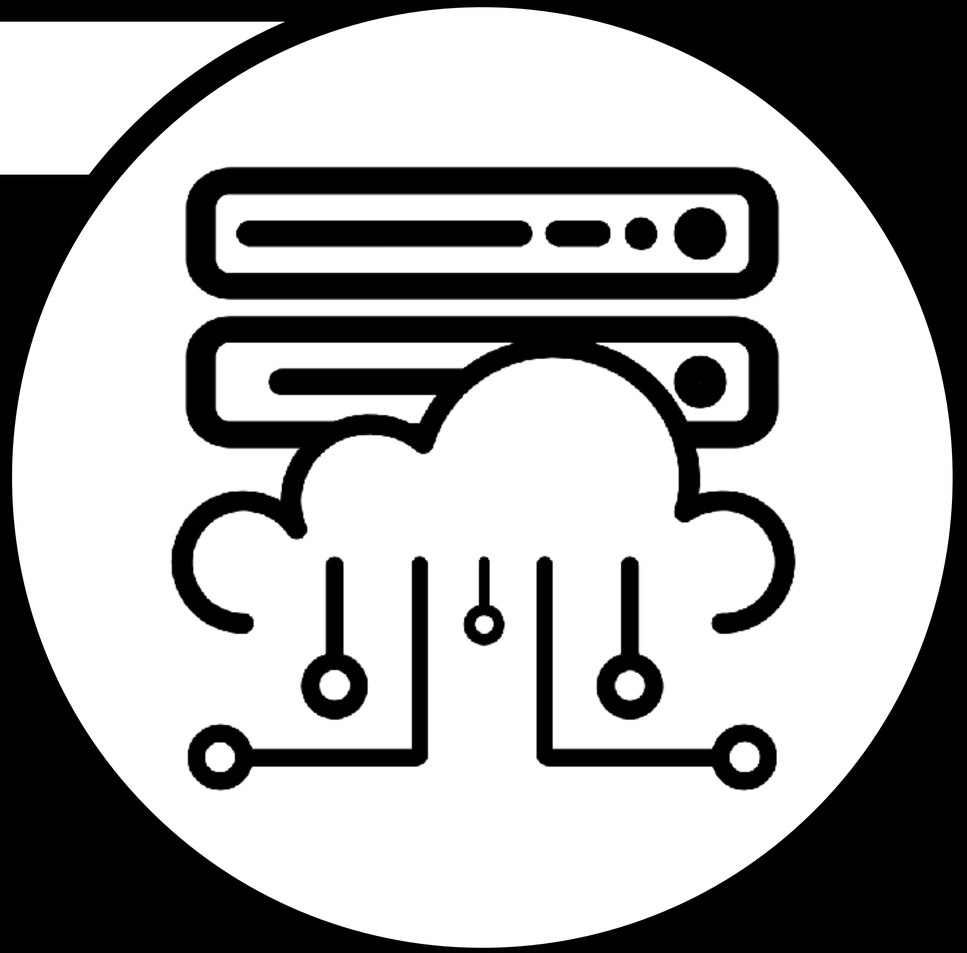
A lot of data is not directly accessible





Data can be biased



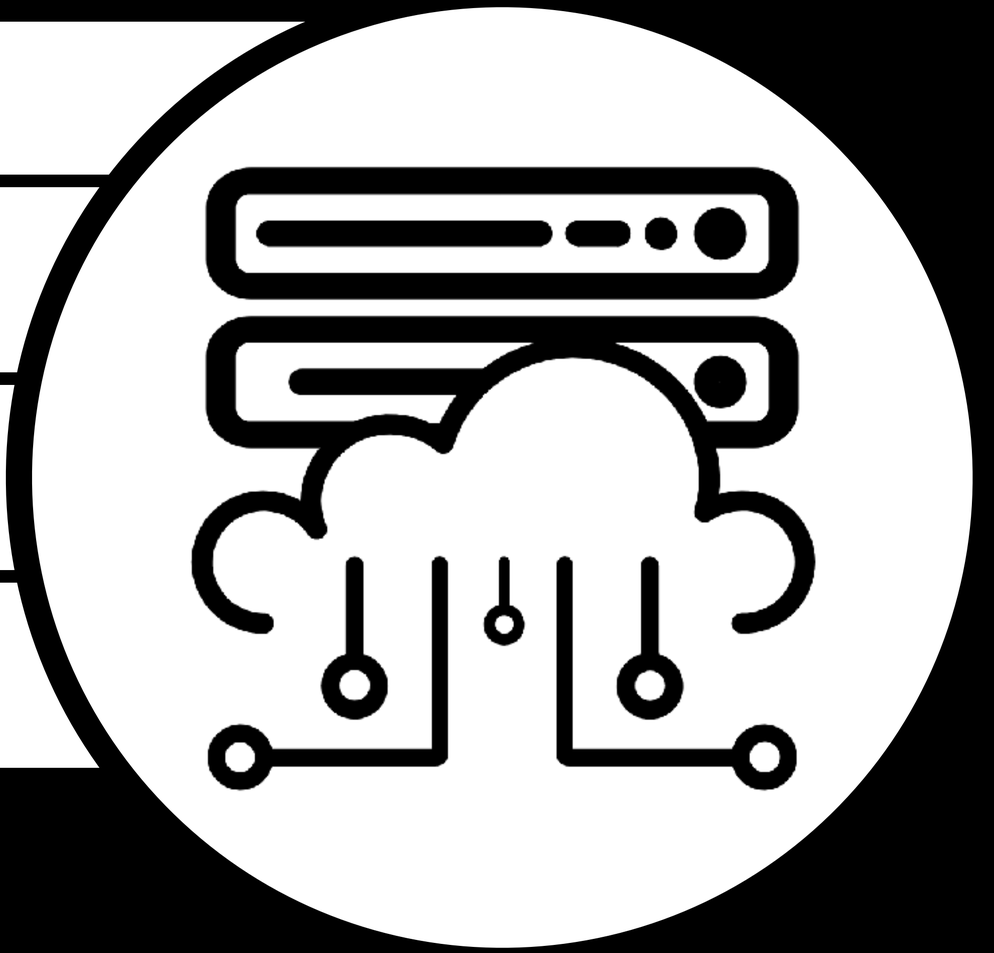


Finding insights in large-scale data is difficult

Too much data. 99% of data is useless to 99% of people

Limited coverage of data sources

Limited keyword-oriented query interfaces



What is the Commons Worth? Estimating the Value of Wikimedia Imagery by Observing Downstream Use

Kristofer Erickson*
University of Leeeds
Leeds, UK
smcke@leeds.ac.uk

Felix Rodriguez Perez†
Independent scholar
Glasgow, UK
felixdrp@gmail.com

Jesus Rodriguez Perez‡
University of Glasgow
Glasgow, UK
rpsoft@gmail.com

ABSTRACT

The Wikimedia Commons (WC) is a peer-produced repository of freely licensed images, videos, sounds and interactive media, containing more than 45 million files. This paper attempts to quantify the societal value of the WC by tracking the downstream use of images found on the platform. We take a random sample of 10,000 images from WC and apply an automated reverse-image search to each, recording when and where they are used 'in the wild'. We detect 54,758 downstream uses of the initial sample and we characterise these at the level of generic and country-code top-level domains (TLDs). We analyse the impact of specific variables on the odds that an image is used. The random sampling technique enables us to estimate overall value of all images contained on the platform. Drawing on the method employed by Heald et al (2015), we find a potential contribution of USD $28.9 billion from downstream use of Wikimedia Commons images over the lifetime of the project.

CCS CONCEPTS

• **Social and professional topics** → **Copyrights**; *Economic impact*; Computer supported cooperative work;

KEYWORDS

Wikimedia Commons, peer production, images, economic valuation, Creative Commons, public domain, curation, open licensing

ACM Reference Format:
Kristofer Erickson, Felix Rodriguez Perez, and Jesus Rodriguez Perez. 2018. What is the Commons Worth? Estimating the Value of Wikimedia Imagery by Observing Downstream Use. In *OpenSym '18: The 14th International Symposium on Open Collaboration, August 22–24, 2018, Paris, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3233391.3233533

*Associate Professor of Media and Communication, School of Media and Communication, University of Leeds.
†Independent scholar, Glasgow UK.

1 INTRODUCTION

Established in 2003, the Wikimedia Commons (WC) is a significant volunteer-led repository of free-to-use public domain images. As of March 2018 it contained 45,583,565 files, of which 43,039,140 were images [3]. Every illustration or photograph contained in the WC – referred to in copyright law as a 'work' – is available on a free and open basis. This is because either the original term of copyright protection in the work has expired, or the creator of the work has made it available under an open license. As of March 2018 the most commonly used open license on the WC was CC-BY-SA 3.0, which allows use for any purpose, including commercially, as long as the user provides credit to the original author of the work and continues to offer it under the same open license. Other commonly used licenses on the WC allow free use without the viral share-alike clause or the attribution requirement. This feature makes the WC very different from commercial image libraries where copyright law normally forbids unauthorised use and distribution of works.

Given the size and scope of the WC, there has been surprisingly limited empirical investigation of its economic and societal impact. Indeed, much of the cross-disciplinary scholarly work available has tended to use the WC as a valuable site for data-mining and other experimental research, or as a case study in collective governance [15] [5]. Searching for scholarly articles on the topic of the WC is also hindered by the fact that many scholarly scientific papers contain citations to illustrations and images available on the WC, vastly increasing the amount of false positives in search results.[1] The WC is clearly an important resource for science and humanities researchers. But does it have a wider societal impact, and if so, can we attempt to quantify the size of its potential influence?

This paper attempts to characterise the downstream use of image files contained on the WC by performing an automated reverse-image search on a sample of 10,000 randomly-selected image files. We record information about the images prior to the search (image size, quality, license parameters) as well as information about the URLs where images appear (quantity of downstream uses, domain

Table 1: Summary statistics for main variables drawn from 10,000 image pages on Wikimedia Commons

Variable	MIN	MAX	MEAN	SD
Any external use	0	1	.348	.476
Any non-commercial	0	1	.304	.46
Any commercial	0	1	.267	.442
Total uses	0	395	5.48	19.78
Total commercial	0	331	2.99	11.722
Total non-commercial	0	129	2.49	8.93
Age of image (years)	0	14	4.4	2.995
Image size (square)	12	8074.5	1324.8	947.3
Format non-jpeg	0	1	.057	.233
Uploader's own work	0	1	.47	.499
Quality image	0	1	0	.062
Originated on Flickr	0	1	.15	.356
Used on Wikipedia	0	1	.171	.376
PD licenses	0	1	.234	.423
Attribution licenses	0	1	.161	.368
Viral licenses	0	1	.582	.493

Majority of content is never used

What is “Data Science”?

Ultimately about extracting **useful** insights

We focus on web data, but data is from many places

This Lecture's Learning Objectives

Define data science and differentiate it from other fields of study

Define and differentiate three kinds of data we will analyze

What questions do you have?

Prof. Cody Buntain | @codybuntain | cbuntain@umd.edu