CDS6214

Data Science Fundamentals

Lecture 2

Data Science Pipeline | Sources | Storage

data
science
pipelines

Data Science "Pipeline" Models

American data visualization expert

Ben Fry's Model (Data Visualization Process)

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

Source: https://bcourses.berkeley.edu/courses/1377158/pages/cs-194-16-introduction-to-data-science-fall-2015

Data Science "Pipeline" Models

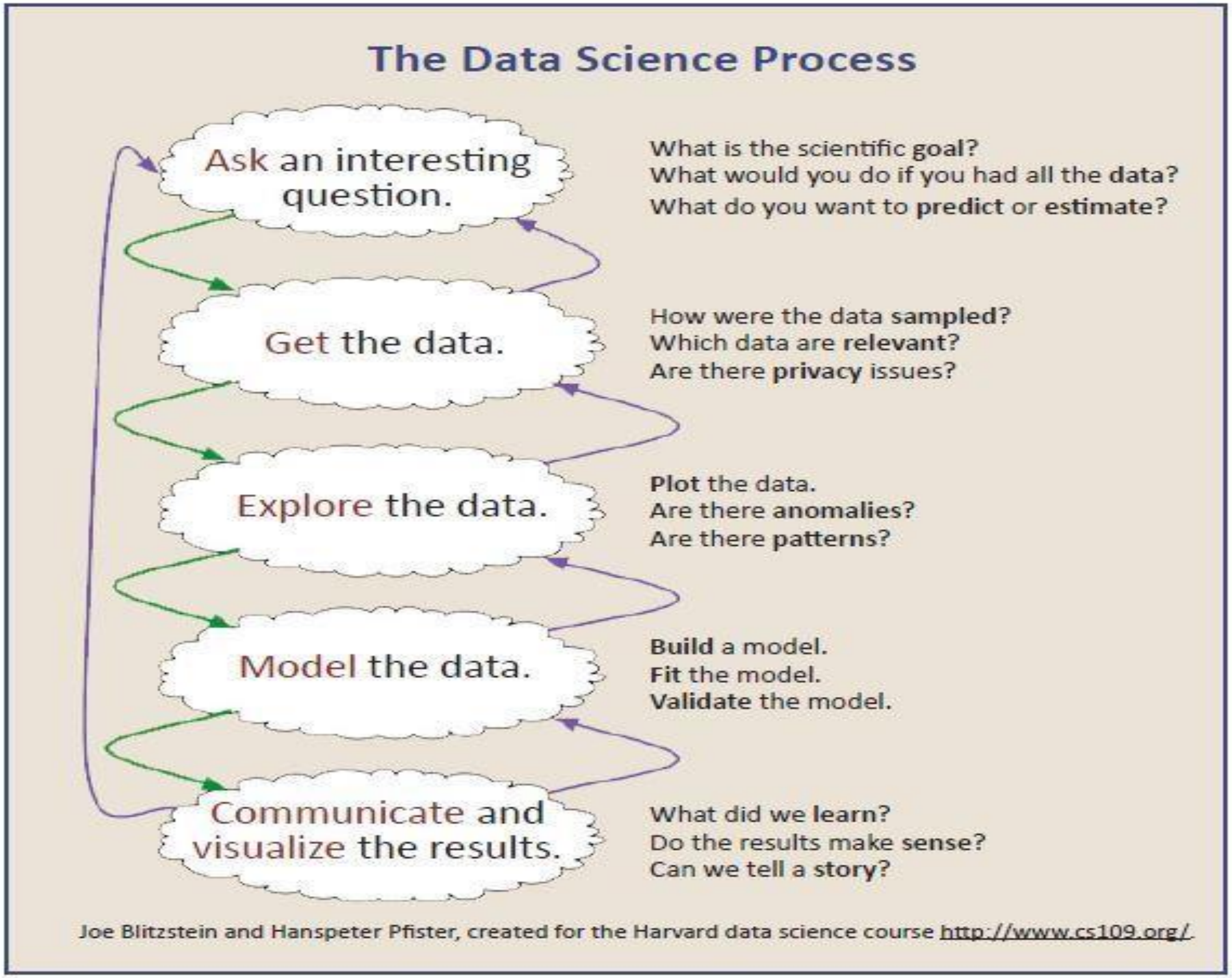
Data scientist and Co-founder of Cloudera

Jeff Hammerbacher's Model

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

Source: https://bcourses.berkeley.edu/courses/1377158/pages/cs-194-16-introduction-to-data-science-fall-2015

Data Science Process



The Data Science Process

Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

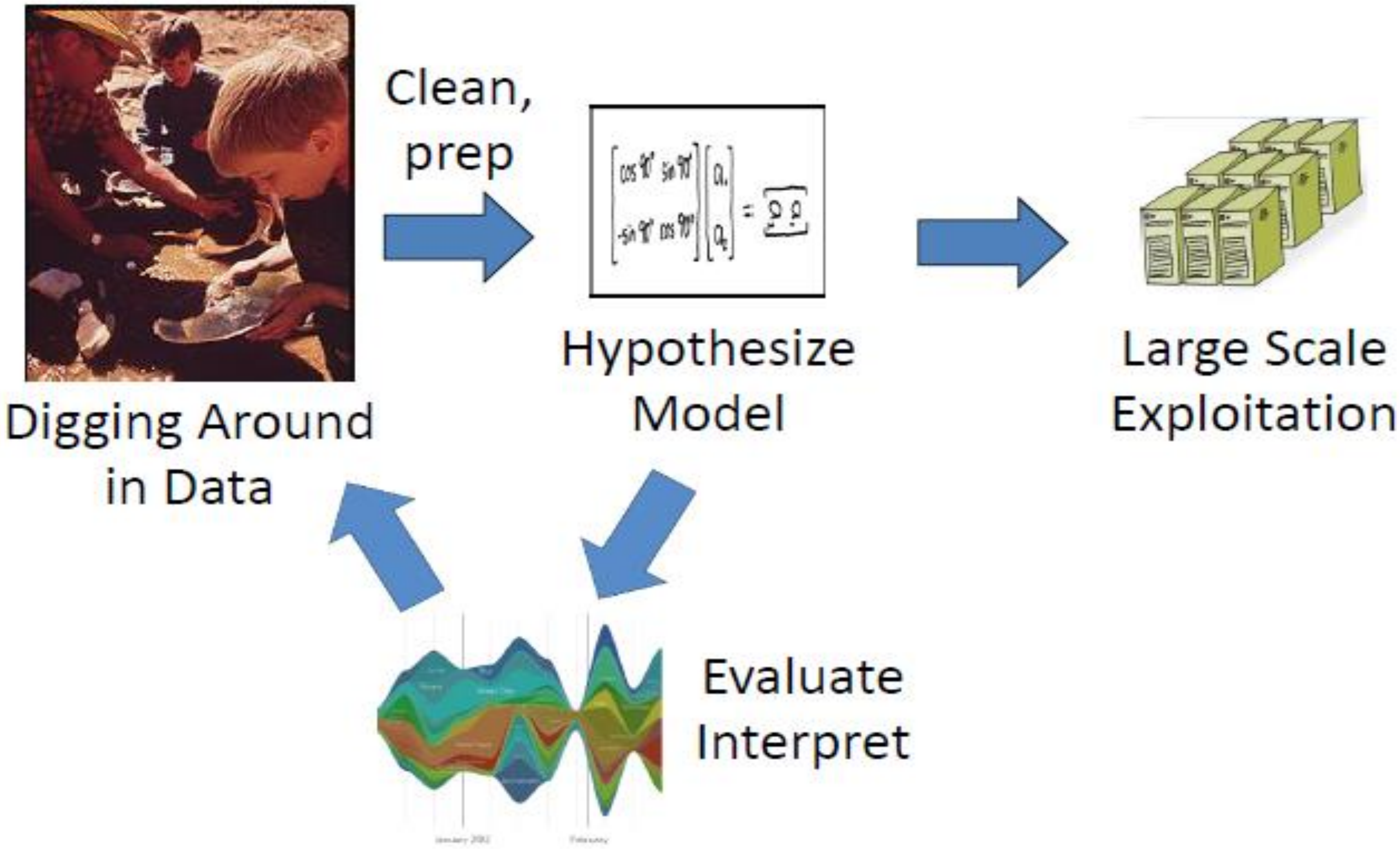
Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://www.cs109.org/.

Process in-a-nutshell



Digging Around in Data

Clean, prep

Hypothesize Model

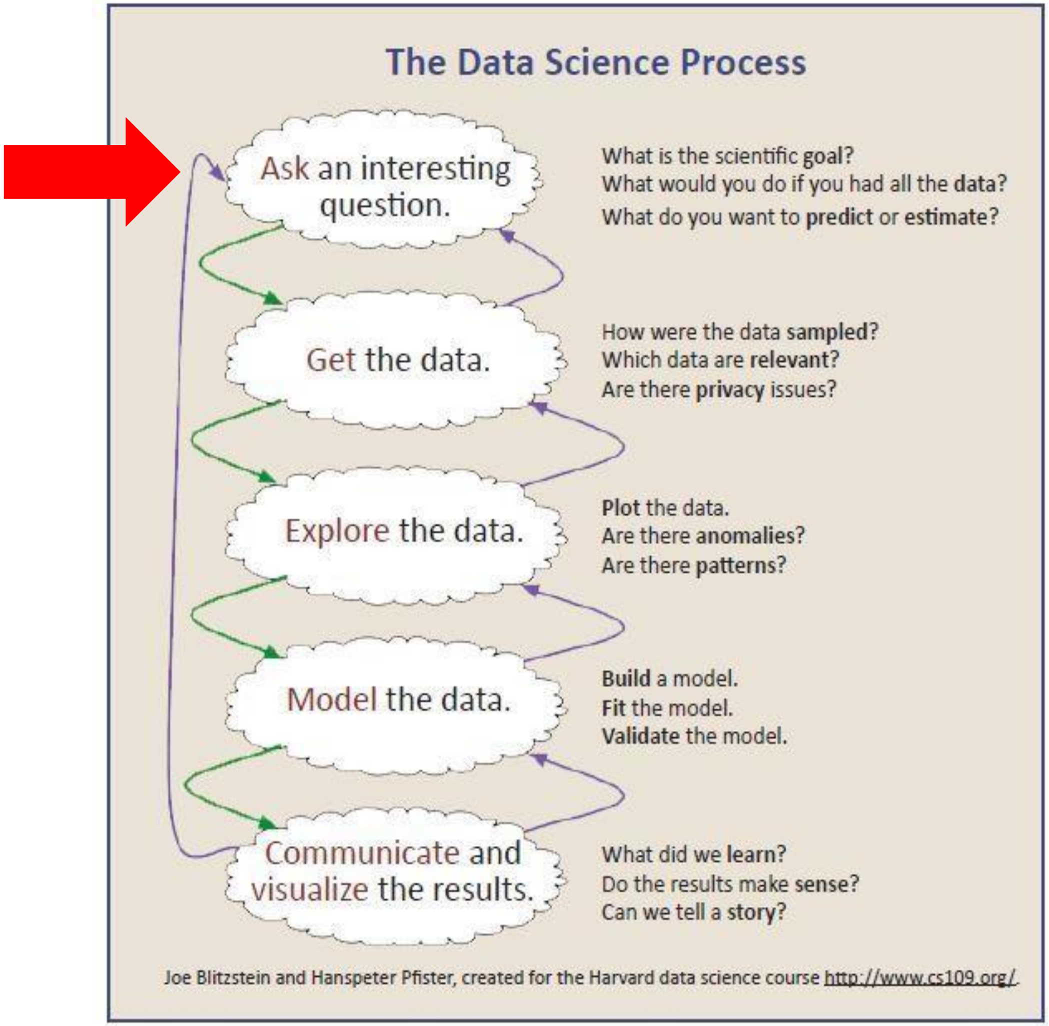
Large Scale Exploitation

Evaluate Interpret

Source: https://bcourses.berkeley.edu/courses/1377158/pages/cs-194-16-introduction-to-data-science-fall-2015

data science pipelines

Data Science Process



The Data Science Process

Ask an interesting question.

What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?

Get the data.

How were the data sampled?
Which data are relevant?
Are there privacy issues?

Explore the data.

Plot the data.
Are there anomalies?
Are there patterns?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we learn?
Do the results make sense?
Can we tell a story?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://www.cs109.org/.

What should we do next?

- Identify the question
- Collect and pre-process the data
- Explore and analyze the data
- Model the data
- Infer and visualize results

The Question

There are 6 basic types of questions that can be asked:

1. **Descriptive**: A ***descriptive*** question is one that seeks to summarize a characteristic of a set of data.

e.g. What is the mean number of servings of fresh fruits and vegetables per day?

2. **Exploratory**: An ***exploratory*** question is one in which you analyze the data to see if there are patterns, trends, or relationships between variables.

e.g. What is the relationship between a range of dietary factors and viral illnesses?

The Question (2)

3. **Inferential**: An ***inferential*** question would be a restatement of the proposed hypothesis as a question and would be answered by analyzing a different set of data. The proposed hypothesis is usually derived from an exploratory question

> *e.g. Given the proposed hypothesis, that the habit of eating at least 5 servings of fresh fruit and vegetables per day is linked with fewer viral illnesses per year based on a sample population of US adults. Is this hypothesis also true for Asian population?*

4. **Predictive**: A ***predictive*** question would be one where you ask what are the set of predictors / factors for a particular behaviour.

> *e.g. What type of people will eat a diet high in fresh fruits and vegetables during the next year?*

The Question (3)

5. **Causal**: A ***causal*** question asks about whether changing one factor will change another factor, on average, in a population.

e.g. Will an increase in consumption of fresh fruits and vegetables reduce the frequency of contracting viral illnesses?

6. **Mechanistic**: A ***mechanistic*** question points to *how* a factor affects the outcome.

e.g. How a diet high in fresh fruits and vegetables leads to a reduction in the number of viral illnesses.

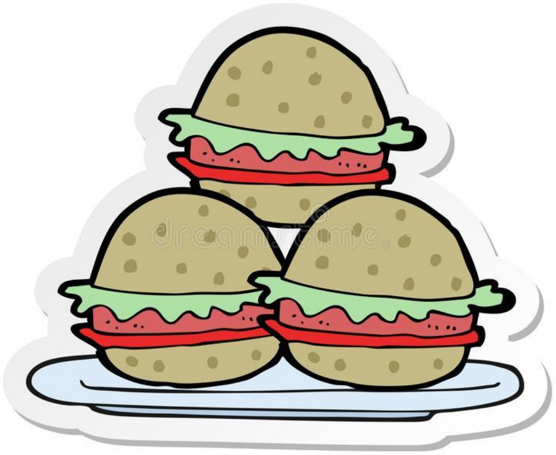
Characteristics of a Good Question

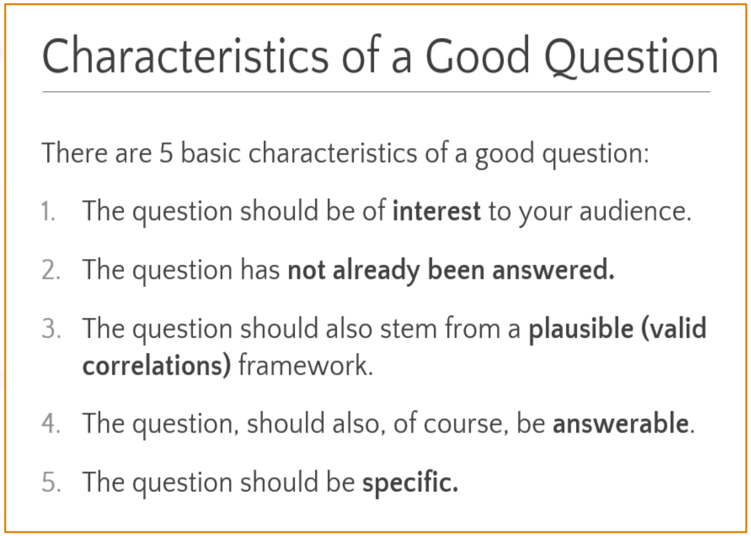
There are 5 basic characteristics of a good question:

1. The question should be of **interest** to your audience.
2. The question has **not already been answered.**
3. The question should also stem from a **plausible (valid correlations)** framework.
4. The question, should also, of course, be **answerable**.
5. The question should be **specific.**

Example

Is there a relation between the number of chicken burgers consumed per month and obesity among teens ?





Characteristics of a Good Question

There are 5 basic characteristics of a good question:

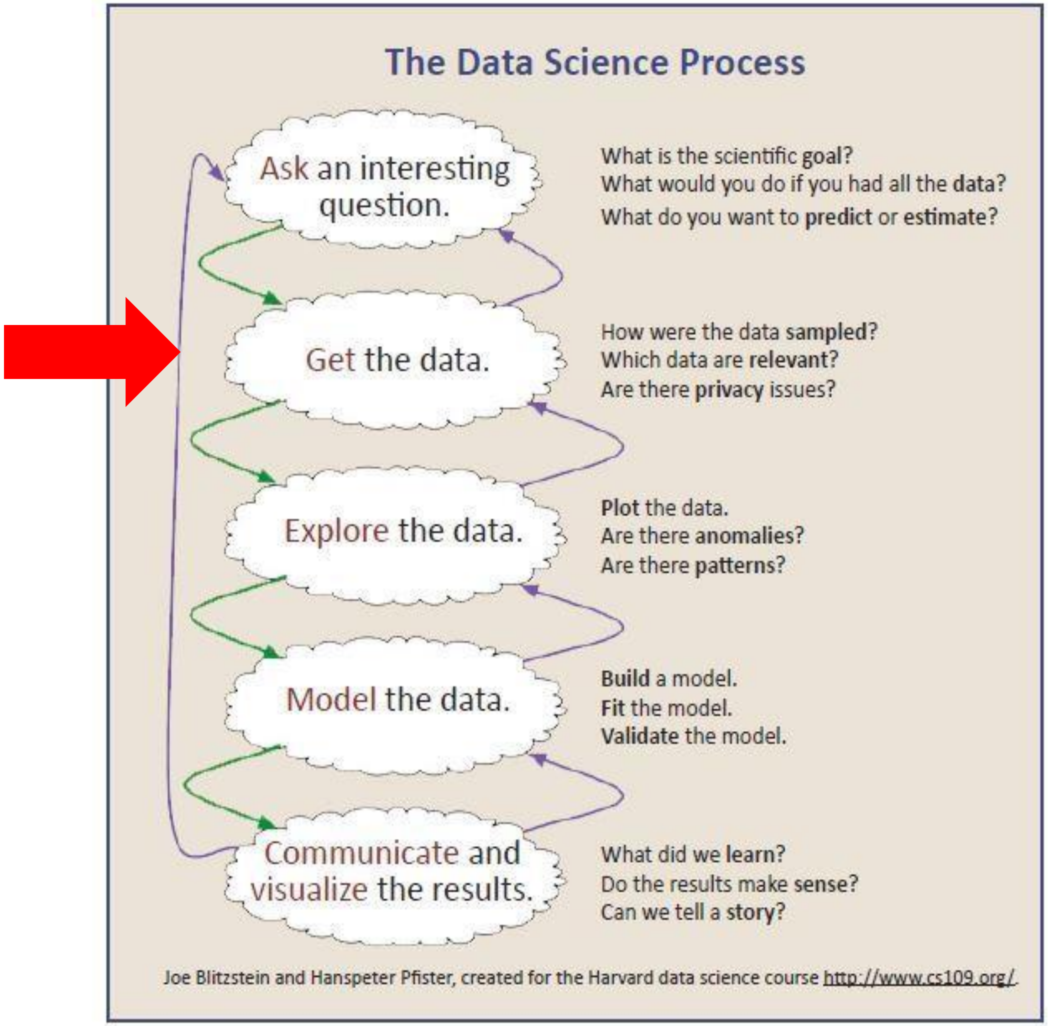
1. The question should be of **interest** to your audience.
2. The question has **not already been answered.**
3. The question should also stem from a **plausible (valid correlations)** framework.
4. The question, should also, of course, be **answerable**.
5. The question should be **specific.**

Translating a Question into a Data Problem

- Every question must be operationalized as a **data problem** that leads to a result.
- Specifically, the questions asked should lead to **interpretable results.**
- Ensure that the data available to answer your question provide reasonably specific **measures** of the factors required to answer your question.
- **Potential problems:**
    - Confounding factor is a potential problem when your question asks about the relationship between factors.
    - Inappropriate data is used. The result is not interpretable because the underlying way in which the data was collected led to a biased result.

data science pipelines

Data Science Process



The Data Science Process

Ask an interesting question.

What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?

Get the data.

How were the data sampled?
Which data are relevant?
Are there privacy issues?

Explore the data.

Plot the data.
Are there anomalies?
Are there patterns?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we learn?
Do the results make sense?
Can we tell a story?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://www.cs109.org/.

What should we do next?

- Identify the question
- Collect and pre-process the data
- Explore and analyze the data
- Model the data
- Infer and visualize results

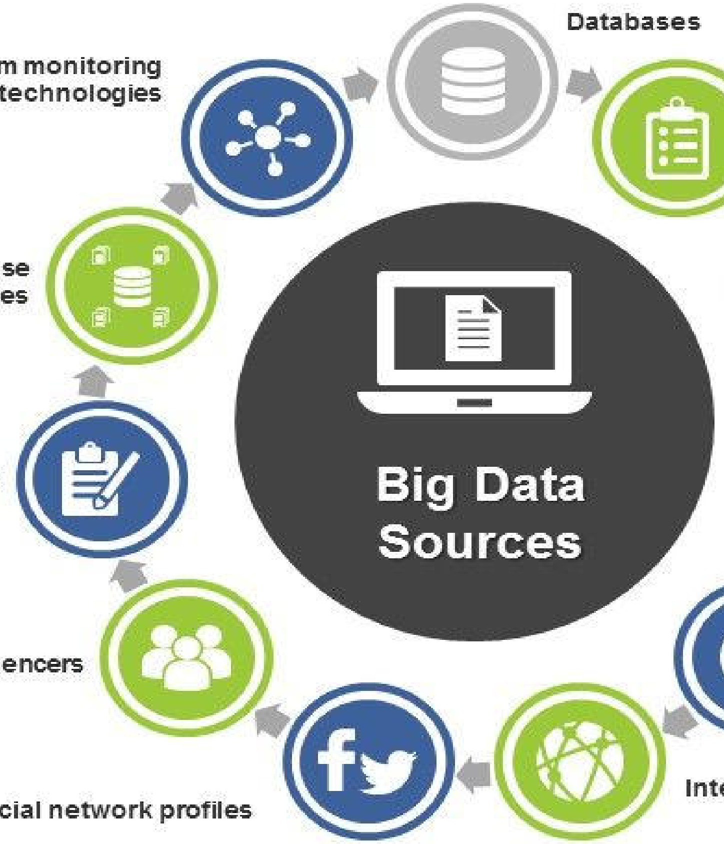
What are the possible data sources?

- Generate the Data (surveys, interviews, etc.)
- Download the Data as files
- Connect to existing databases
- Call a web service API
- Setup equipment such as AV units and sensors
- Search for the Data (open data, search engines, industries/institutions)

What are the possible data sources?

- Generate the Data (surveys, interviews, etc.)
- Download the Data as files
- Connect to existing databases
- Call a web service API
- Setup equipment such as AV units and sensors
- Search for the Data (open data, search engines, industries/institutions)

Big Data Sources



Databases

Legacy Documents

Media

Cloud

Web

Internet of Things

Social network profiles

Social Influencers

Activity Generated Data

Data warehouse appliances

Network and in-stream monitoring technologies

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

Data Collection

- Process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.
- Data can be *qualitative* or *quantitative*. Any difference between the two in terms of:
  a) Content?
  b) Methods of Collection?
  c) Example?

Common Problems in Data Collection

1. Irrelevant or duplicate data collected
2. Pertinent data omitted
3. Erroneous or misinterpreted data collected
4. Too little data acquired from client
5. Poor documentation
6. Conflicting data
7. Handwriting
8. Language barrier
9. Insufficient time

Common Data Files

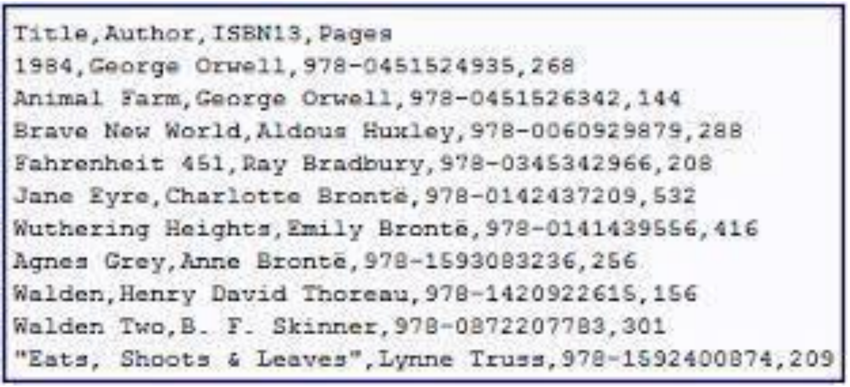
Text, binary etc. – Some require more processing than others

- Syslog files
- Spreadsheets (xls, csv)
- PDF files
- Image files
- Raw text files
- Formatted text files

Common Data Format

- **Delimited values: Comma separated values, Tab separated values**
- Markup languages: HTML5 / XML, JSON
- Ad-hoc formats: Graph edge lists, voting records, device-captured signals, server logs, etc.

What is the **relation** between rows and columns?



```
Title,Author,ISBN13,Pages
1984,George Orwell,978-0451524935,268
Animal Farm,George Orwell,978-0451526342,144
Brave New World,Aldous Huxley,978-0060929879,288
Fahrenheit 451,Ray Bradbury,978-0345342966,208
Jane Eyre,Charlotte Brontë,978-0142437209,532
Wuthering Heights,Emily Brontë,978-0141439556,416
Agnes Grey,Anne Brontë,978-1593083236,256
Walden,Henry David Thoreau,978-1420922615,156
Walden Two,B. F. Skinner,978-0872207783,301
"Eats, Shoots & Leaves",Lynne Truss,978-1592400874,209
```

Tabular Data

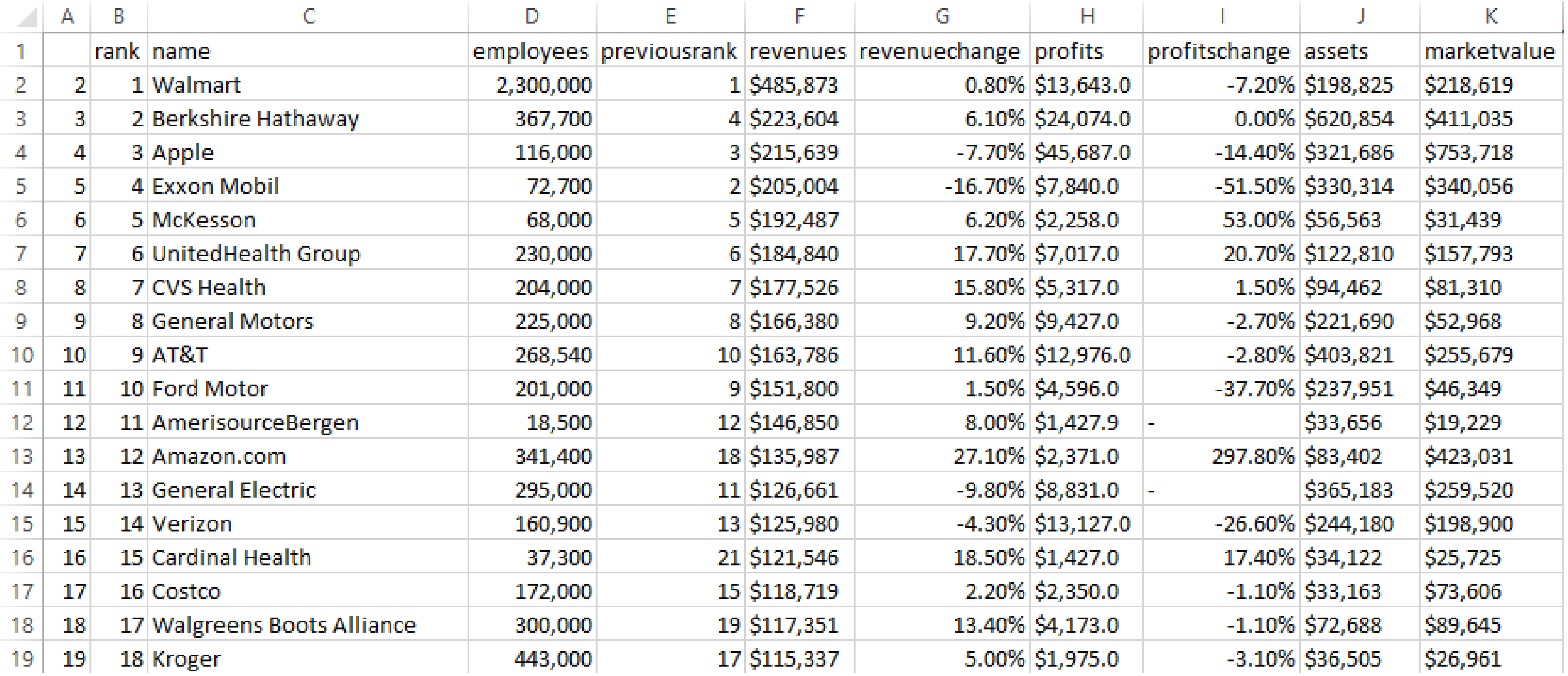
What is a table?

- A **table** is a collection of **rows** and **columns**
- Each row has an **index**
- Each column has a **name**
- A **cell** is specified by an (index, name) pair
- A cell may or may not have a **value**

Often stored as text files in **CSV** or **TSV** format.

Tabular Data – Example

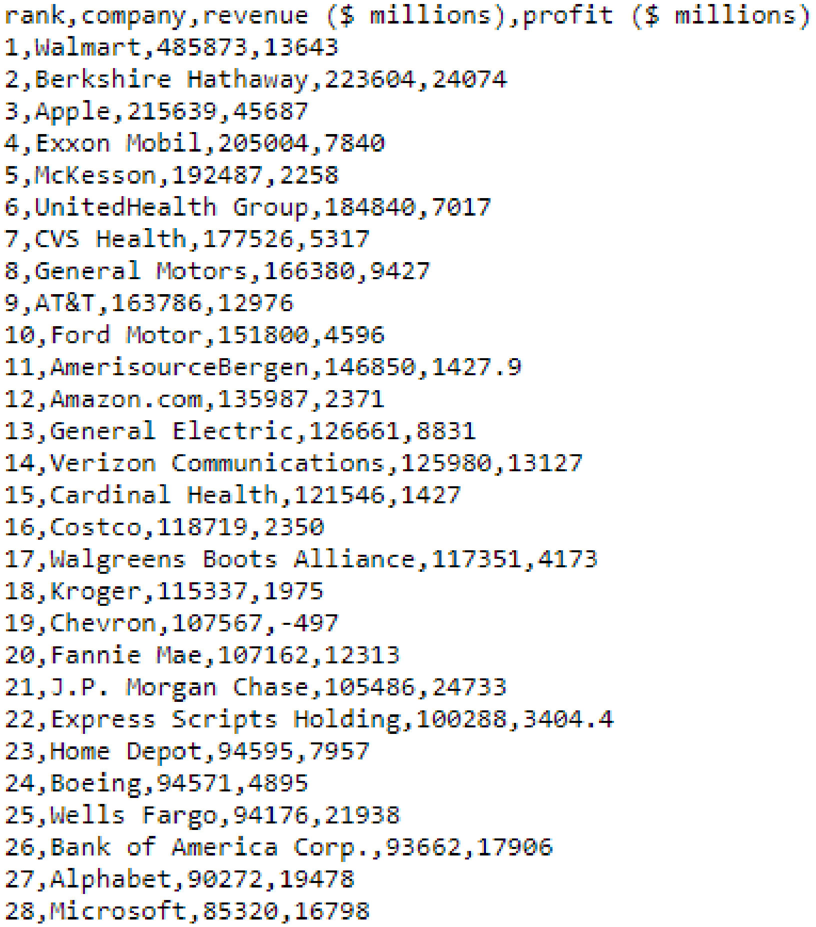
Fortune 500 Companies 2017 Data (open in spreadsheet editor)



	A	B	C	D	E	F	G	H	I	J	K
1		rank	name	employees	previousrank	revenues	revenuechange	profits	profitschange	assets	marketvalue
2	2	1	Walmart	2,300,000	1	$485,873	0.80%	$13,643.0	-7.20%	$198,825	$218,619
3	3	2	Berkshire Hathaway	367,700	4	$223,604	6.10%	$24,074.0	0.00%	$620,854	$411,035
4	4	3	Apple	116,000	3	$215,639	-7.70%	$45,687.0	-14.40%	$321,686	$753,718
5	5	4	Exxon Mobil	72,700	2	$205,004	-16.70%	$7,840.0	-51.50%	$330,314	$340,056
6	6	5	McKesson	68,000	5	$192,487	6.20%	$2,258.0	53.00%	$56,563	$31,439
7	7	6	UnitedHealth Group	230,000	6	$184,840	17.70%	$7,017.0	20.70%	$122,810	$157,793
8	8	7	CVS Health	204,000	7	$177,526	15.80%	$5,317.0	1.50%	$94,462	$81,310
9	9	8	General Motors	225,000	8	$166,380	9.20%	$9,427.0	-2.70%	$221,690	$52,968
10	10	9	AT&T	268,540	10	$163,786	11.60%	$12,976.0	-2.80%	$403,821	$255,679
11	11	10	Ford Motor	201,000	9	$151,800	1.50%	$4,596.0	-37.70%	$237,951	$46,349
12	12	11	AmerisourceBergen	18,500	12	$146,850	8.00%	$1,427.9	-	$33,656	$19,229
13	13	12	Amazon.com	341,400	18	$135,987	27.10%	$2,371.0	297.80%	$83,402	$423,031
14	14	13	General Electric	295,000	11	$126,661	-9.80%	$8,831.0	-	$365,183	$259,520
15	15	14	Verizon	160,900	13	$125,980	-4.30%	$13,127.0	-26.60%	$244,180	$198,900
16	16	15	Cardinal Health	37,300	21	$121,546	18.50%	$1,427.0	17.40%	$34,122	$25,725
17	17	16	Costco	172,000	15	$118,719	2.20%	$2,350.0	-1.10%	$33,163	$73,606
18	18	17	Walgreens Boots Alliance	300,000	19	$117,351	13.40%	$4,173.0	-1.10%	$72,688	$89,645
19	19	18	Kroger	443,000	17	$115,337	5.00%	$1,975.0	-3.10%	$36,505	$26,961

Tabular Data (csv)

Fortune 500 Companies 2017 Data (CSV format)



```
rank,company,revenue ($ millions),profit ($ millions)
1,Walmart,485873,13643
2,Berkshire Hathaway,223604,24074
3,Apple,215639,45687
4,Exxon Mobil,205004,7840
5,McKesson,192487,2258
6,UnitedHealth Group,184840,7017
7,CVS Health,177526,5317
8,General Motors,166380,9427
9,AT&T,163786,12976
10,Ford Motor,151800,4596
11,AmerisourceBergen,146850,1427.9
12,Amazon.com,135987,2371
13,General Electric,126661,8831
14,Verizon Communications,125980,13127
15,Cardinal Health,121546,1427
16,Costco,118719,2350
17,Walgreens Boots Alliance,117351,4173
18,Kroger,115337,1975
19,Chevron,107567,-497
20,Fannie Mae,107162,12313
21,J.P. Morgan Chase,105486,24733
22,Express Scripts Holding,100288,3404.4
23,Home Depot,94595,7957
24,Boeing,94571,4895
25,Wells Fargo,94176,21938
26,Bank of America Corp.,93662,17906
27,Alphabet,90272,19478
28,Microsoft,85320,16798
```

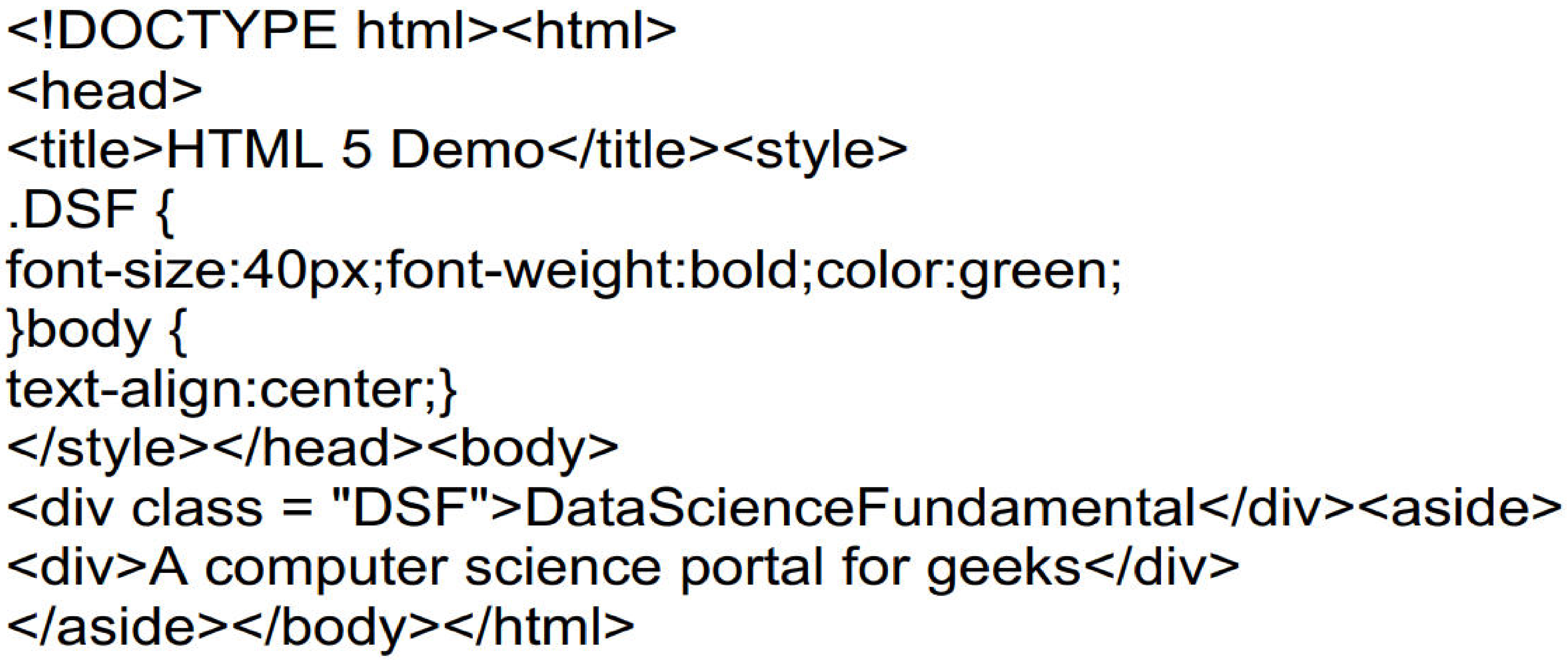
Common Data Formats

- Delimited values: Comma separated values, Tab separated values

- **Markup languages: HTML5 / XML, JSON**

- Ad-hoc formats: Graph edge lists, voting records, device-captured signals, server logs, etc.

Example of markup language:HTML5



```
<!DOCTYPE html><html>
<head>
<title>HTML 5 Demo</title><style>
.DSF {
font-size:40px;font-weight:bold;color:green;
}body {
text-align:center;}
</style></head><body>
<div class = "DSF">DataScienceFundamental</div><aside>
<div>A computer science portal for geeks</div>
</aside></body></html>
```

Source:https://www.geeksforgeeks.org/html5/

Example of markup language: XML

XML: Generalizes HTML and specifies data **structure**. XML schema can be applied later to interpret XML data and specify **data types**. Here is a sample XML-encoded **data**:

[Example]

```
<location>
  <latitude>37.78333</latitude>
  <longitude>122.4167</longitude>
</location>
```

When stored without a schema, the numerical data are stored as **strings**.

Source: https://bcourses.berkeley.edu/courses/1267848

JSON (JavaScript Object Notation)

- A lightweight data-interchange format.
- It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999.
- Completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others.
- Consists of two components:
  - **A collection of name/value pairs**. In various languages, this is realized as an object, record, struct, dictionary, hash table, keyed list, or associative array.
  - **An ordered list of values.** In most languages, this is realized as an array, vector, list, or sequence.

JSON – Example Tweet Format

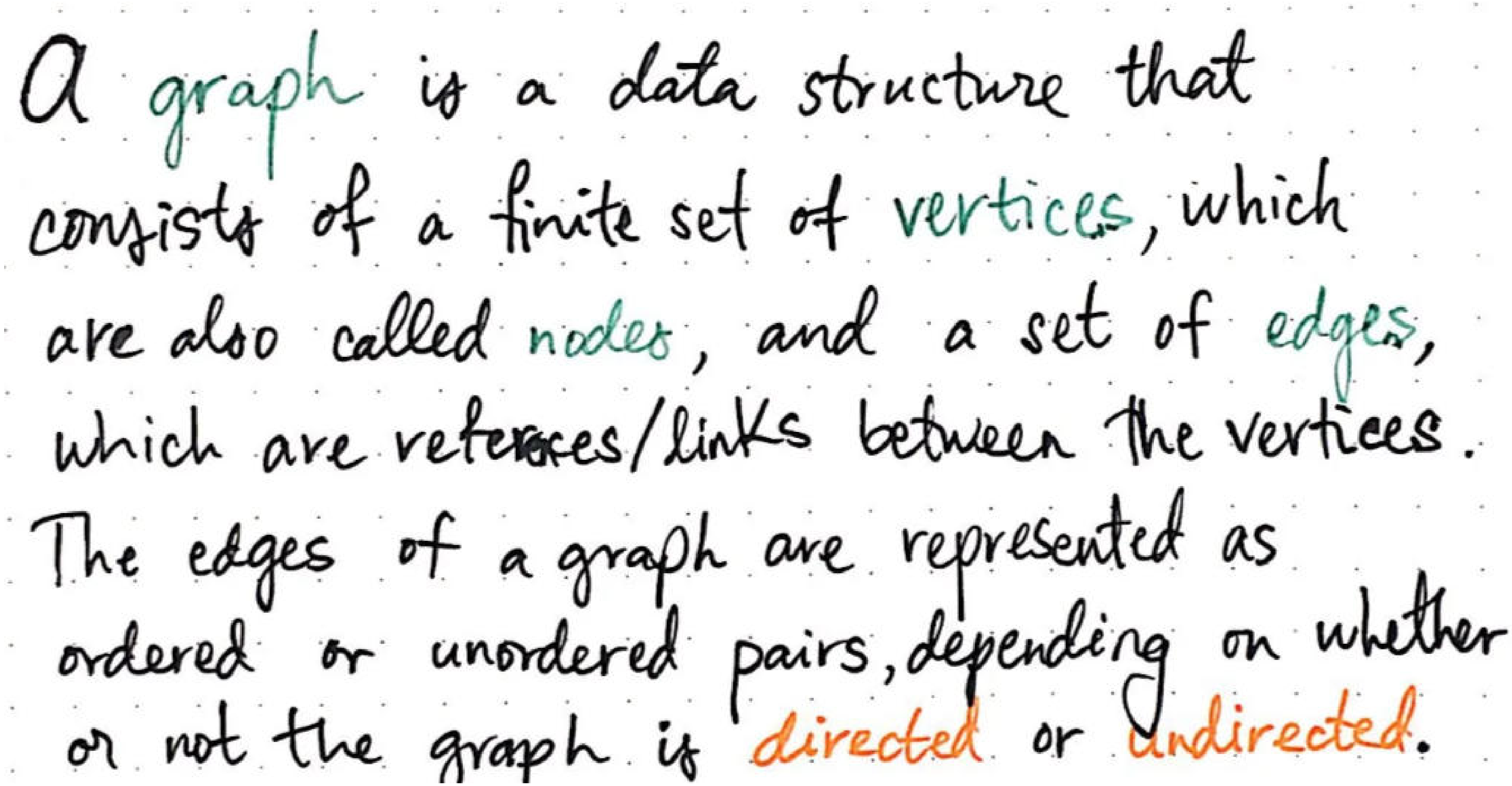
```
{"location":[
                 {"latitude":37.78333, "longitude":122.4167 }
 ]}
```

Common Data Formats

- Delimited values: Comma separated values, Tab separated values
- Markup languages: HTML5 / XML, JSON,
- **Ad-hoc formats: Graph edge lists, voting records, device-captured signals, server logs, etc.**

Example – Graph Edge Lists

Let's look at definition of a graph extracted exactly as it is from a website:



A graph is a data structure that consists of a finite set of vertices, which are also called nodes, and a set of edges, which are references/links between the vertices. The edges of a graph are represented as ordered or unordered pairs, depending on whether or not the graph is directed or undirected.

https://medium.com/basecs/from-theory-to-practice-representing-graphs-cfd782c5be38

Example – Graph Edge Lists

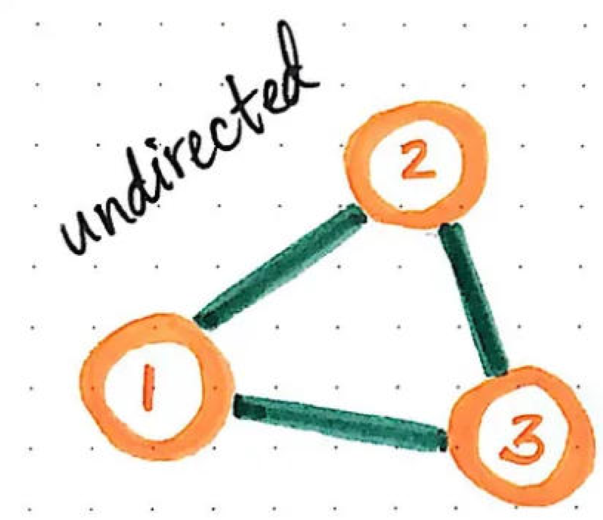


Figure A

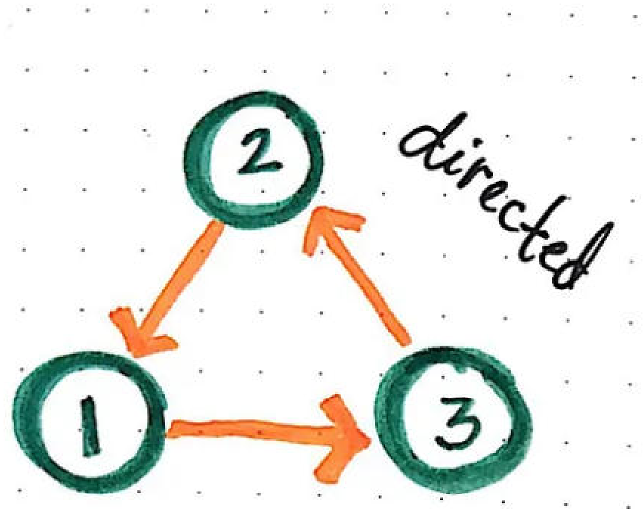
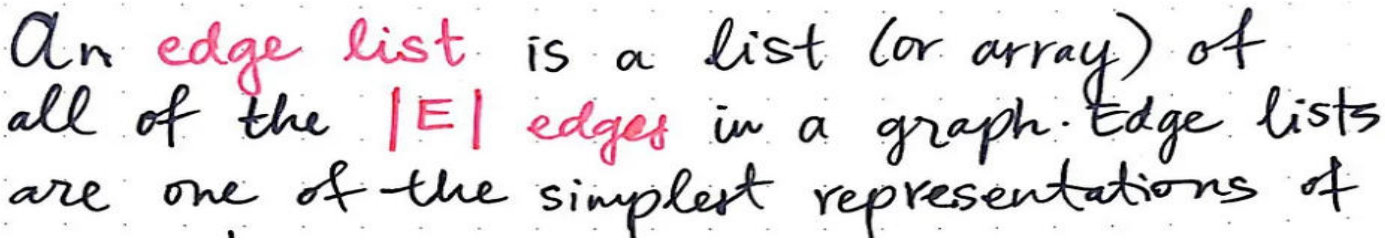


Figure B

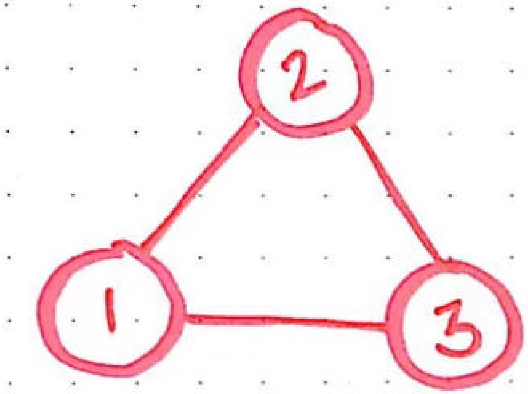
For each figure, can you identify the set of vertices and edges ?

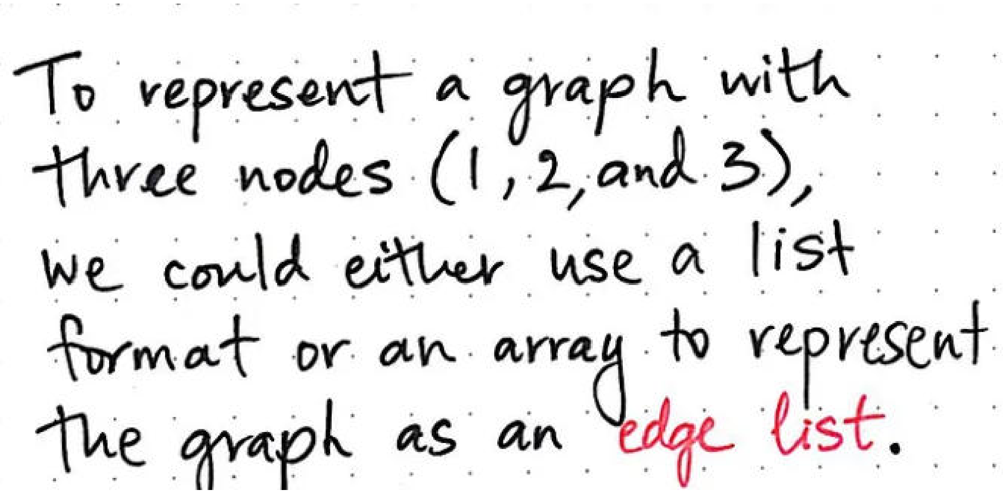
https://medium.com/basecs/from-theory-to-practice-representing-graphs-cfd782c5be38

Example – Graph Edge Lists



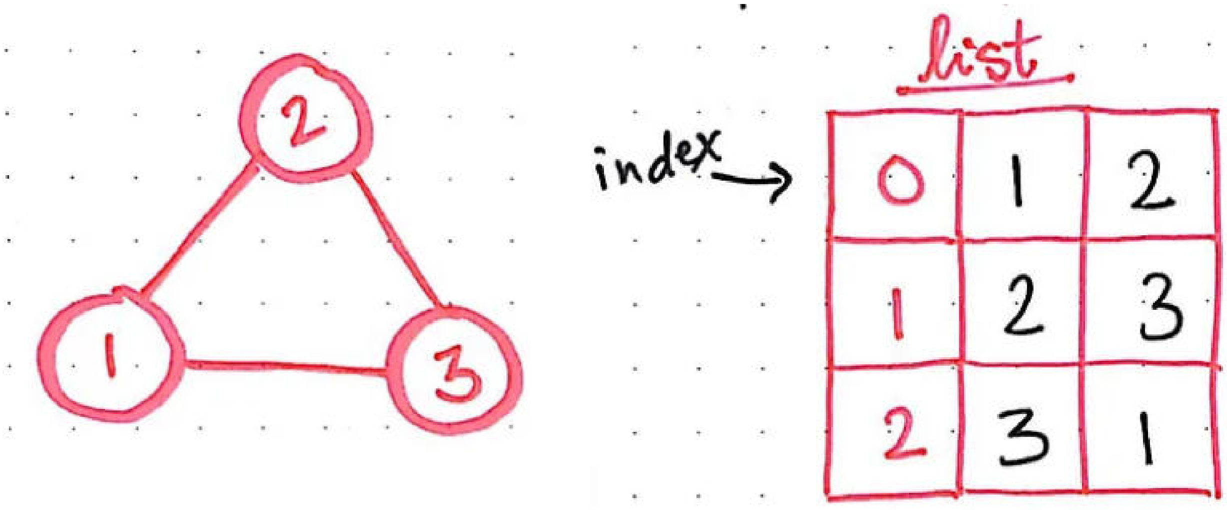
An edge list is a list (or array) of all of the |E| edges in a graph. Edge lists are one of the simplest representations of





To represent a graph with three nodes (1, 2, and 3), we could either use a list format or an array to represent the graph as an edge list.

Example – Graph Edge Lists



0	1	2
1	2	3
2	3	1

How do you represent the above graph as an array ?

https://medium.com/basecs/from-theory-to-practice-representing-graphs-cfd782c5be38

Example – Apache Web Log

Processes, usually daemons, create logs

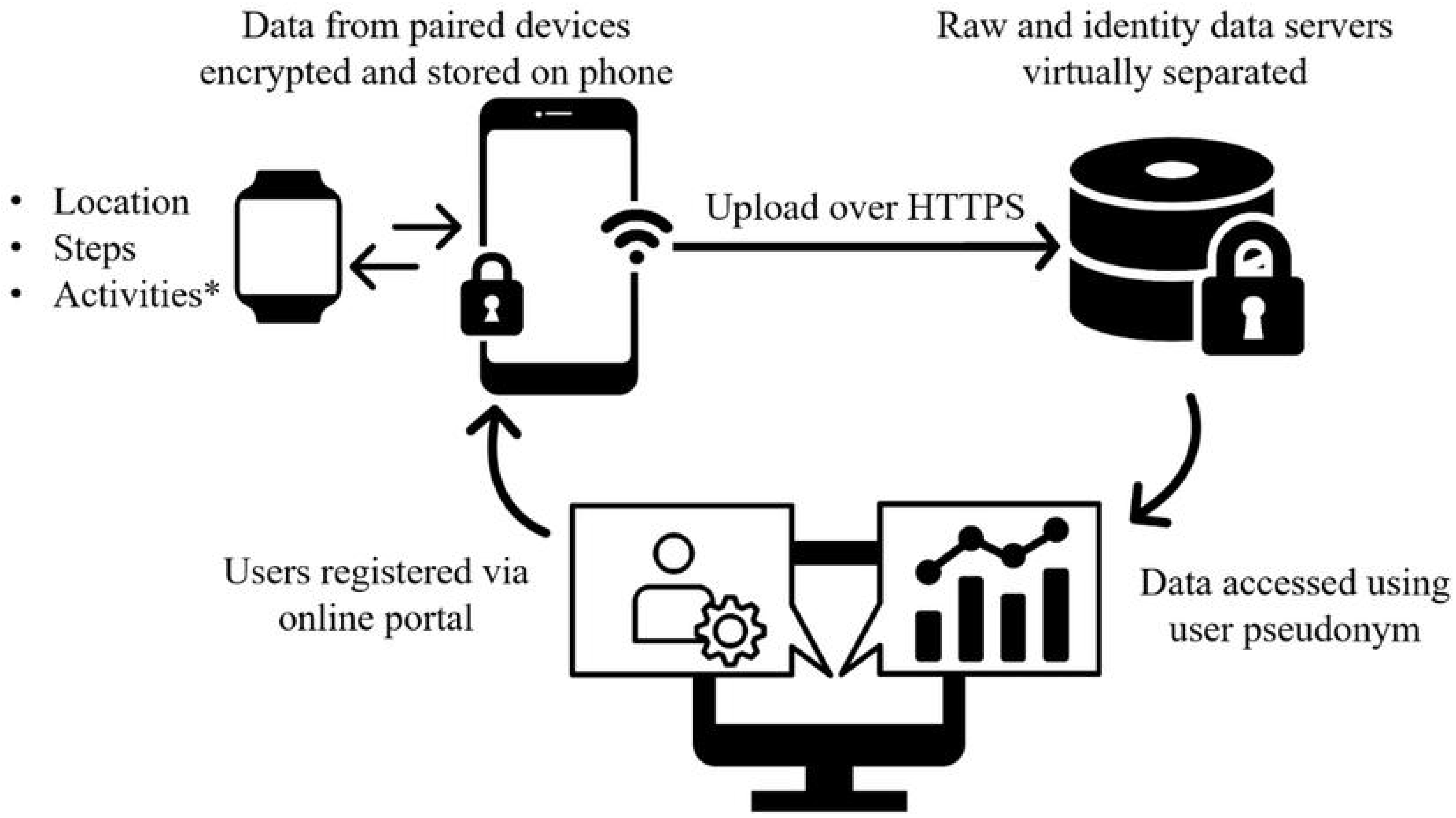
e.g., `httpd, mysqld, syslogd`

```
66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html
HTTP/1.1" 200 11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"

111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET / HTTP/1.1"
200 10801 "http://www.google.com/search?q=log+analyzer&ie=utf-
8&oe=utf-8 &aq=t&rls=org.mozilla:en-US:official&client=firefox-a"
"Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7)
Gecko/20070914 Firefox/2.0.0.7"

111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET /style.css
HTTP/1.1" 200 3225 “"http://www.loganalyzer.net/" "Mozilla/5.0
(Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914
Firefox/2.0.0.7"
```

Example of setup for data collection for wearable and mobile technology



Data from paired devices encrypted and stored on phone

Raw and identity data servers virtually separated

- Location
- Steps
- Activities*

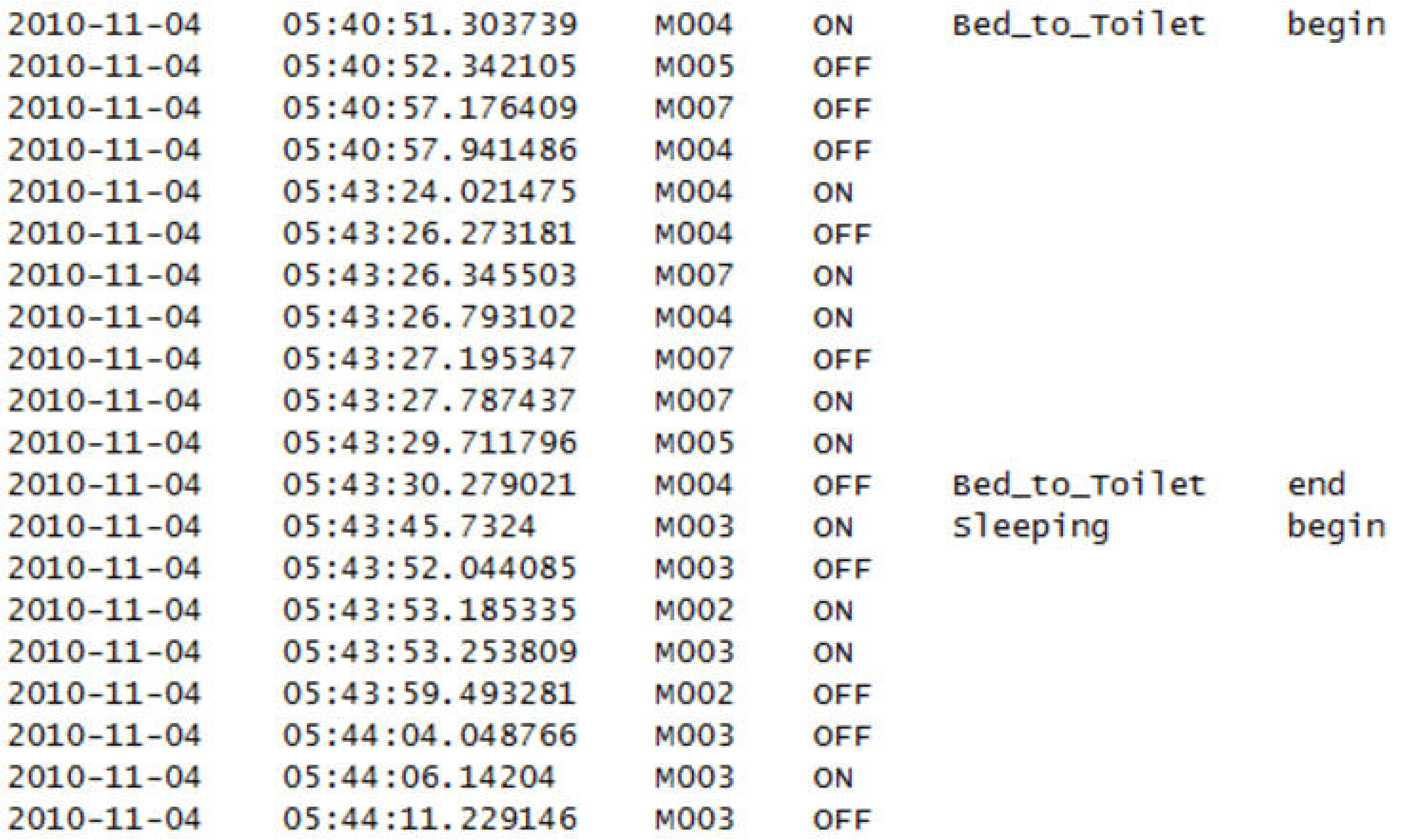
Upload over HTTPS

Users registered via online portal

Data accessed using user pseudonym

https://www.researchgate.net/figure/Data-collection-setup-data-is-collected-using-sensors-on-board-a-smartwatch-and_fig5_328379775

Time based Data – Device Measurements



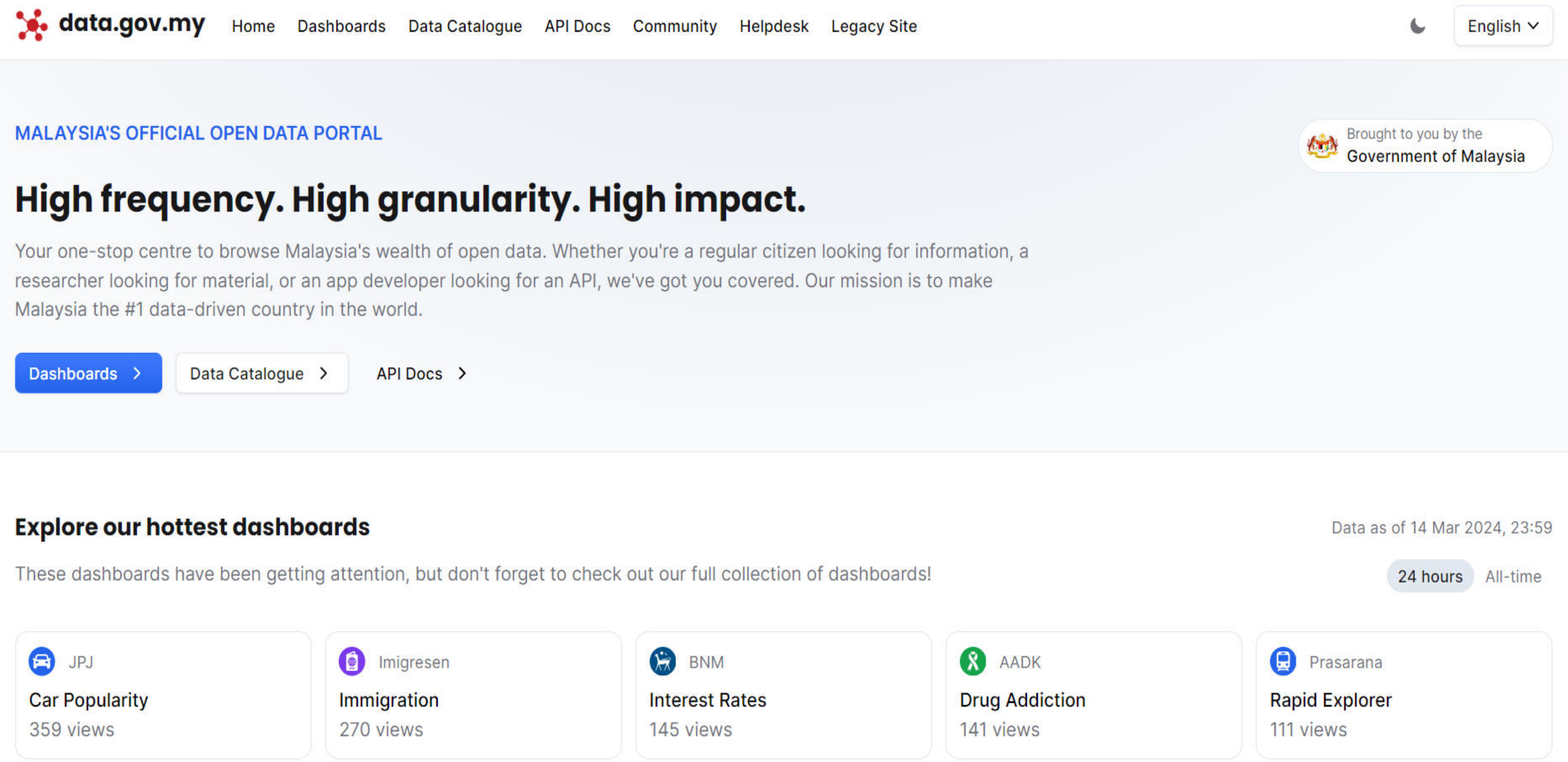
```
2010-11-04    05:40:51.303739    M004    ON     Bed_to_Toilet    begin
2010-11-04    05:40:52.342105    M005    OFF
2010-11-04    05:40:57.176409    M007    OFF
2010-11-04    05:40:57.941486    M004    OFF
2010-11-04    05:43:24.021475    M004    ON
2010-11-04    05:43:26.273181    M004    OFF
2010-11-04    05:43:26.345503    M007    ON
2010-11-04    05:43:26.793102    M004    ON
2010-11-04    05:43:27.195347    M007    OFF
2010-11-04    05:43:27.787437    M007    ON
2010-11-04    05:43:29.711796    M005    ON
2010-11-04    05:43:30.279021    M004    OFF    Bed_to_Toilet    end
2010-11-04    05:43:45.7324      M003    ON     Sleeping         begin
2010-11-04    05:43:52.044085    M003    OFF
2010-11-04    05:43:53.185335    M002    ON
2010-11-04    05:43:53.253809    M003    ON
2010-11-04    05:43:59.493281    M002    OFF
2010-11-04    05:44:04.048766    M003    OFF
2010-11-04    05:44:06.14204     M003    ON
2010-11-04    05:44:11.229146    M003    OFF
```

Sample raw and activity annotated sensor data.Sensors IDs starting with M are motion sensors

https://www.researchgate.net/figure/Sample-raw-and-activity-annotated-sensor-dataSensors-IDs-starting-with-M-are-motion _fig2_308843277

Malaysia's Official Open Data Portal

https://data.gov.my/



MALAYSIA'S OFFICIAL OPEN DATA PORTAL

Brought to you by the
Government of Malaysia

High frequency. High granularity. High impact.

Your one-stop centre to browse Malaysia's wealth of open data. Whether you're a regular citizen looking for information, a researcher looking for material, or an app developer looking for an API, we've got you covered. Our mission is to make Malaysia the #1 data-driven country in the world.

Explore our hottest dashboards

Data as of 14 Mar 2024, 23:59

These dashboards have been getting attention, but don't forget to check out our full collection of dashboards!

JPJ
Car Popularity
359 views

Imigresen
Immigration
270 views

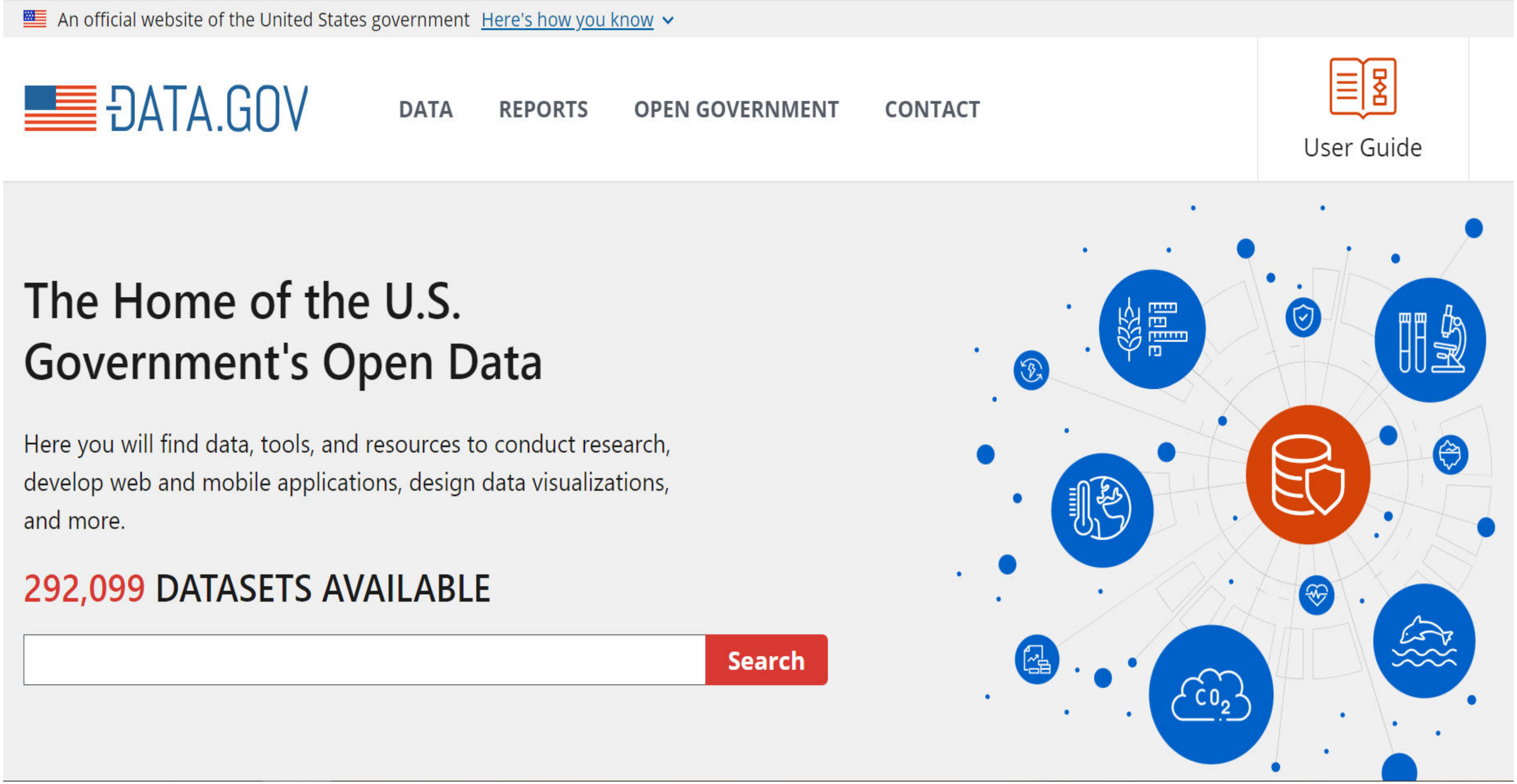
BNM
Interest Rates
145 views

AADK
Drug Addiction
141 views

Prasarana
Rapid Explorer
111 views

The Home of the U.S. Government's Open Data

https://data.gov/



An official website of the United States government Here's how you know

DATA.GOV

DATA REPORTS OPEN GOVERNMENT CONTACT

User Guide

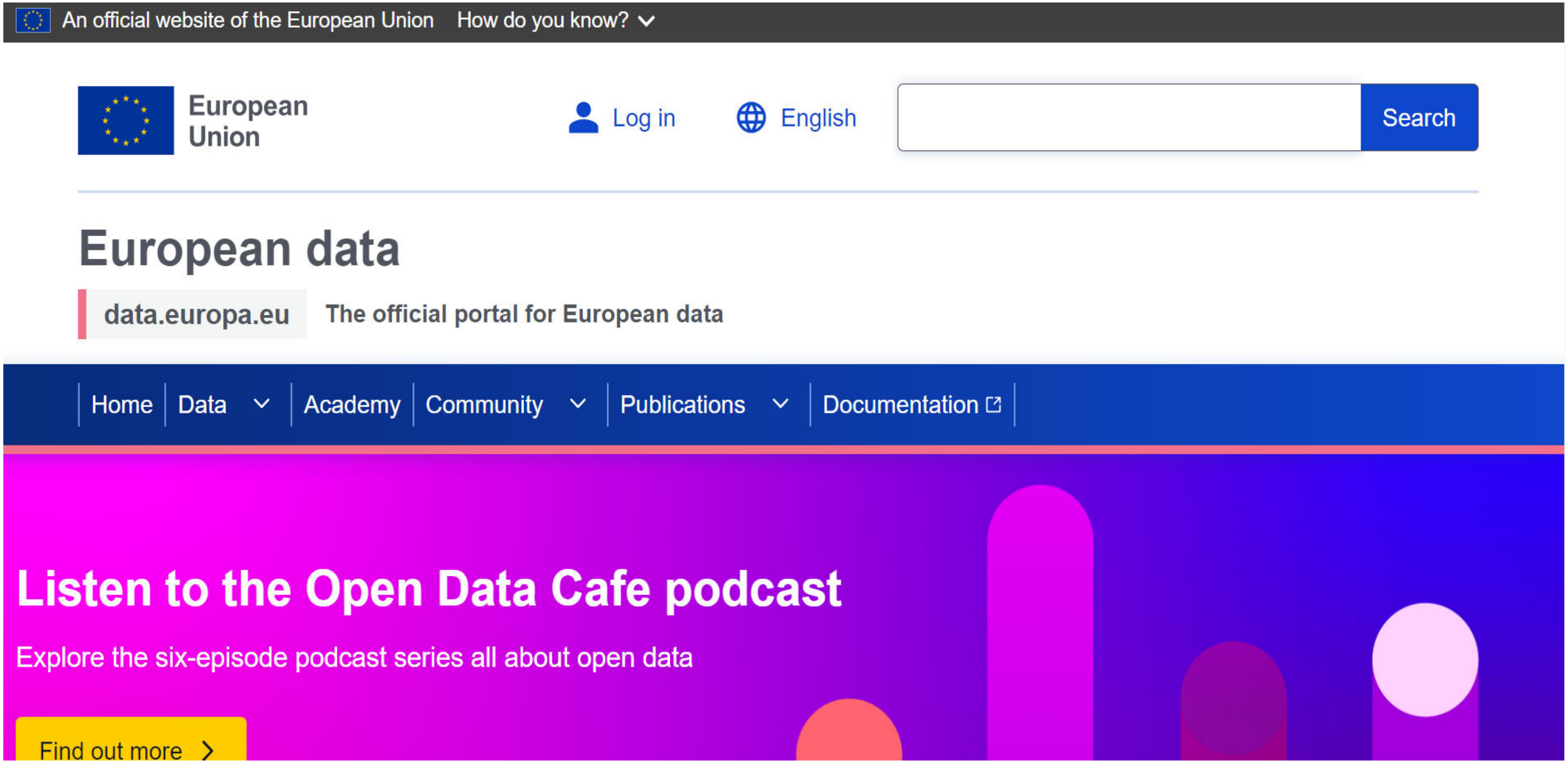
The Home of the U.S. Government's Open Data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

292,099 DATASETS AVAILABLE

The Official Portal for European Data

https://data.europa.eu/en



An official website of the European Union How do you know?

European Union

Log in English Search

European data

data.europa.eu The official portal for European data

Home Data Academy Community Publications Documentation

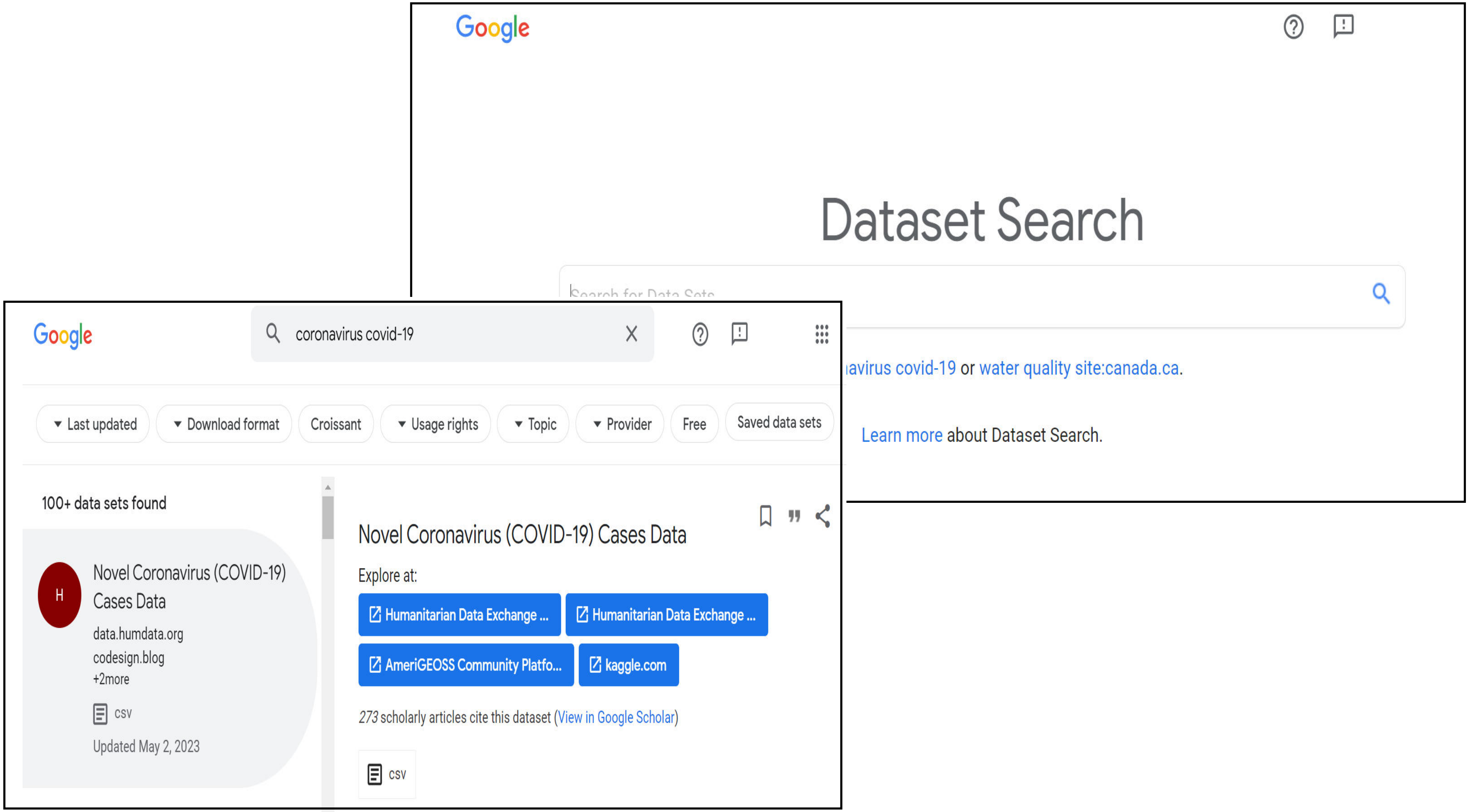
Listen to the Open Data Cafe podcast

Explore the six-episode podcast series all about open data

Find out more

Google Dataset Search

https://datasetsearch.research.google.com/



Google

Dataset Search

Search for Data Sets

avirus covid-19 or water quality site:canada.ca.

Learn more about Dataset Search.

Google

coronavirus covid-19

Last updated
Download format
Croissant
Usage rights
Topic
Provider
Free
Saved data sets

100+ data sets found

Novel Coronavirus (COVID-19) Cases Data
data.humdata.org
codesign.blog
+2more
csv
Updated May 2, 2023

Novel Coronavirus (COVID-19) Cases Data

Explore at:

Humanitarian Data Exchange ...
Humanitarian Data Exchange ...
AmeriGEOSS Community Platfo...
kaggle.com

273 scholarly articles cite this dataset (View in Google Scholar)

csv

types of data store

a repository for persistently storing and managing collections of data which include not just repositories like databases, but also simpler store types such as simple files, emails etc.

By 2025,
an estimated 463 exabytes of data will be generated every single day
(1EB = 1 million TB)

Data Sources: Example of a Web Company

Example: Facebook

- Application databases
- Web server logs
- Event logs
- API server logs
- Ad server logs
- Search server logs
- Advertisement landing page content
- Wikipedia
- Images and video

Structured Data

Semi-structured Data

Unstructured Data

Source: https://bcourses.berkeley.edu/courses/1267848

The (changing) role of Schema

A data **schema** specifies the **structure** and **types** of a data repository, e.g. the types of each column in a table. They may also specify constraints **within** or **between** data fields.

Traditional databases (Relational DB) are **schema-on-write**. You cannot load data into a table without a schema.

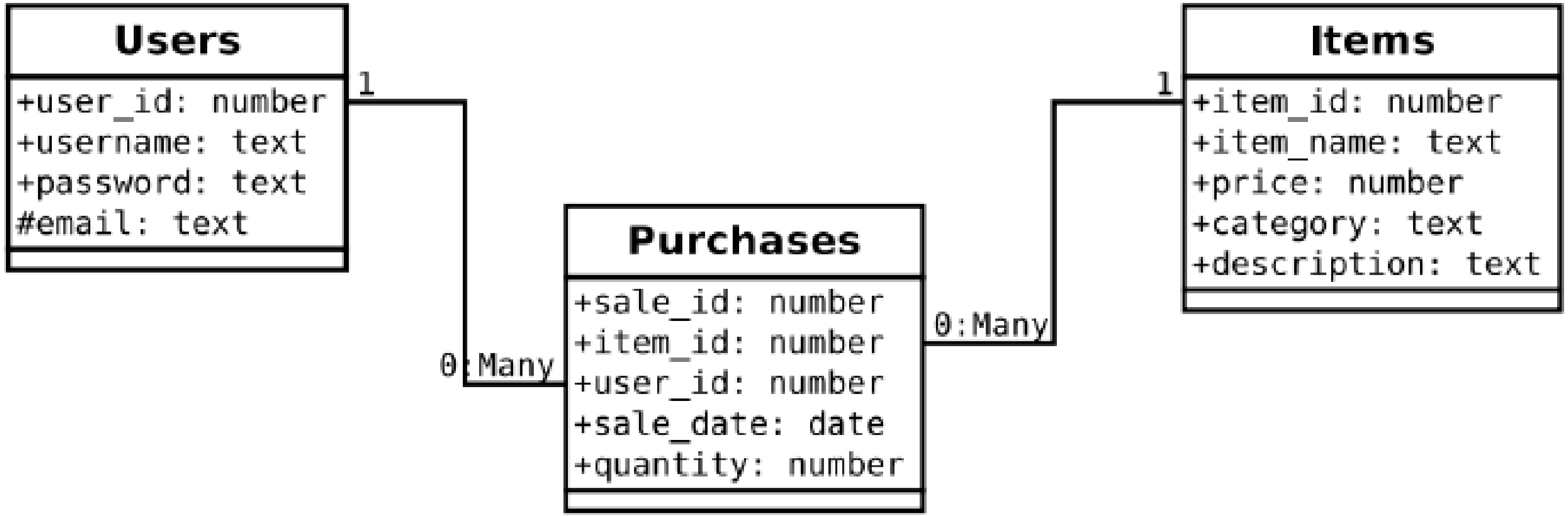
Newer data stores (e.g. noSQL) are **schema-on-read** or **schema-less**: You can defer applying a schema until you read the data, or avoid schema altogether

Source: https://bcourses.berkeley.edu/courses/1267848

Relational DB

- **Relational databases** use the notion of databases separated into tables where each column represents a field and each row represents a record
  - Named after a branch of algebraic set theory known as relational algebra
- Tables can be related or linked with each other with use of foreign keys / common columns
- E.g. of Relational DBs: MySQL, PostgreSQL SQLite3
  - Represent and store data in tables and rows
- **Relational database management systems (RDBMSs)** are the primary technology for storing structured data in web and business applications

Relational DB Schema



Users

+user_id: number
+username: text
+password: text
#email: text

Purchases

+sale_id: number
+item_id: number
+user_id: number
+sale_date: date
+quantity: number

Items

+item_id: number
+item_name: text
+price: number
+category: text
+description: text

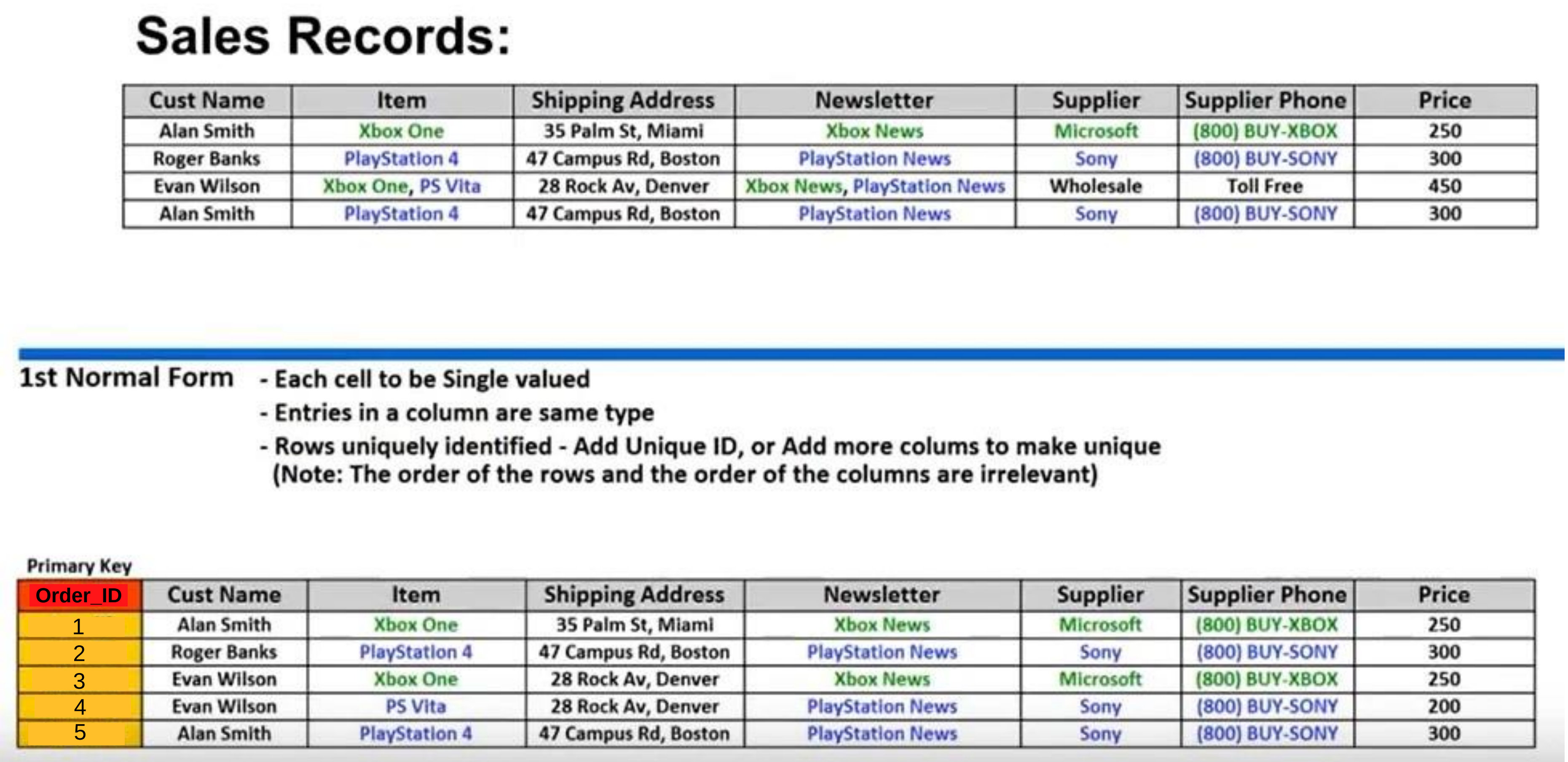
Relational DB

- An important design aspect: **Normalisation** of schema
    - Reduces redundancy (repetitive data), ensure data stored logically
    - Larger tables are divided into smaller tables which are linked using relationships

Normalisation of Schema

- 3 common forms:
  a. **First Normal Form (1NF)**: Eliminate groups of repeating data by creating a new table for each group of related data which is identified by primary key.
  b. **Second Normal Form (2NF)**: If a set of values are the same for multiple records, move them to a new table and link the two tables with foreign key.
  c. **Third Normal Form (3NF):** Fields which do not depend on the primary key of a table must be removed and if necessary be put into another table.

1NF



Sales Records:

Cust Name	Item	Shipping Address	Newsletter	Supplier	Supplier Phone	Price
Alan Smith	Xbox One	35 Palm St, Miami	Xbox News	Microsoft	(800) BUY-XBOX	250
Roger Banks	PlayStation 4	47 Campus Rd, Boston	PlayStation News	Sony	(800) BUY-SONY	300
Evan Wilson	Xbox One, PS Vita	28 Rock Av, Denver	Xbox News, PlayStation News	Wholesale	Toll Free	450
Alan Smith	PlayStation 4	47 Campus Rd, Boston	PlayStation News	Sony	(800) BUY-SONY	300

1st Normal Form

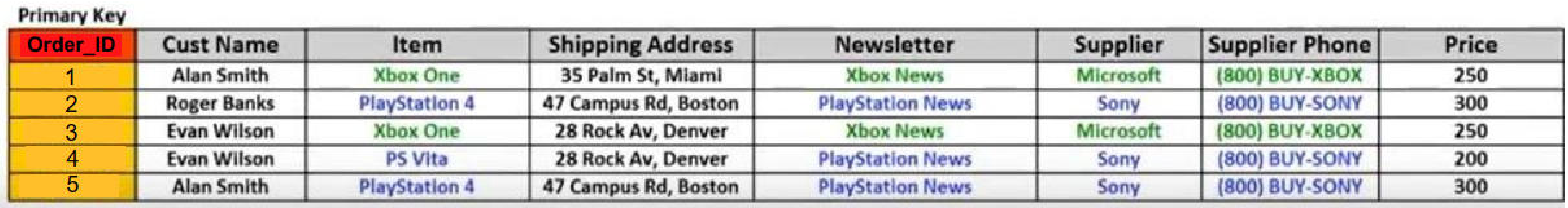
- Each cell to be Single valued
- Entries in a column are same type
- Rows uniquely identified - Add Unique ID, or Add more colums to make unique
  (Note: The order of the rows and the order of the columns are irrelevant)

Primary Key

Order_ID	Cust Name	Item	Shipping Address	Newsletter	Supplier	Supplier Phone	Price
1	Alan Smith	Xbox One	35 Palm St, Miami	Xbox News	Microsoft	(800) BUY-XBOX	250
2	Roger Banks	PlayStation 4	47 Campus Rd, Boston	PlayStation News	Sony	(800) BUY-SONY	300
3	Evan Wilson	Xbox One	28 Rock Av, Denver	Xbox News	Microsoft	(800) BUY-XBOX	250
4	Evan Wilson	PS Vita	28 Rock Av, Denver	PlayStation News	Sony	(800) BUY-SONY	200
5	Alan Smith	PlayStation 4	47 Campus Rd, Boston	PlayStation News	Sony	(800) BUY-SONY	300

Extracted and modified from source: https://ericdatascience.wordpress.com/2019/06/12/database-normalisation

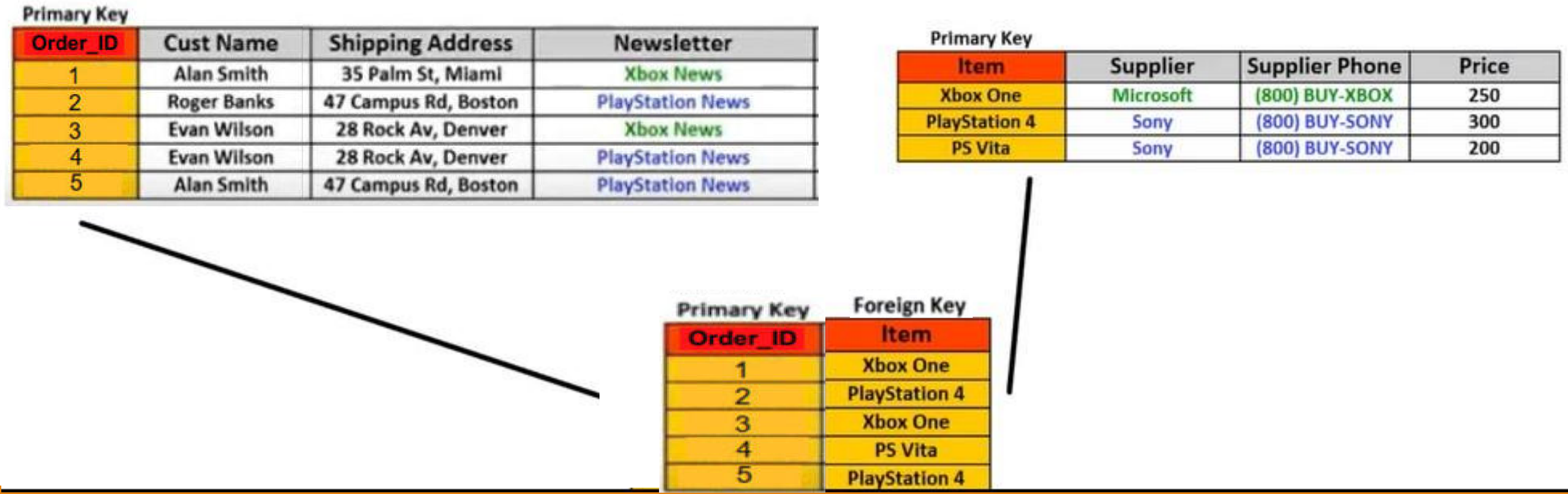
2NF



Primary Key

Order_ID	Cust Name	Item	Shipping Address	Newsletter	Supplier	Supplier Phone	Price
1	Alan Smith	Xbox One	35 Palm St, Miami	Xbox News	Microsoft	(800) BUY-XBOX	250
2	Roger Banks	PlayStation 4	47 Campus Rd, Boston	PlayStation News	Sony	(800) BUY-SONY	300
3	Evan Wilson	Xbox One	28 Rock Av, Denver	Xbox News	Microsoft	(800) BUY-XBOX	250
4	Evan Wilson	PS Vita	28 Rock Av, Denver	PlayStation News	Sony	(800) BUY-SONY	200
5	Alan Smith	PlayStation 4	47 Campus Rd, Boston	PlayStation News	Sony	(800) BUY-SONY	300

2nd Normal Form - All attributes (Non-Key Columns) dependent on the key



Primary Key

Order_ID	Cust Name	Shipping Address	Newsletter
1	Alan Smith	35 Palm St, Miami	Xbox News
2	Roger Banks	47 Campus Rd, Boston	PlayStation News
3	Evan Wilson	28 Rock Av, Denver	Xbox News
4	Evan Wilson	28 Rock Av, Denver	PlayStation News
5	Alan Smith	47 Campus Rd, Boston	PlayStation News

Primary Key

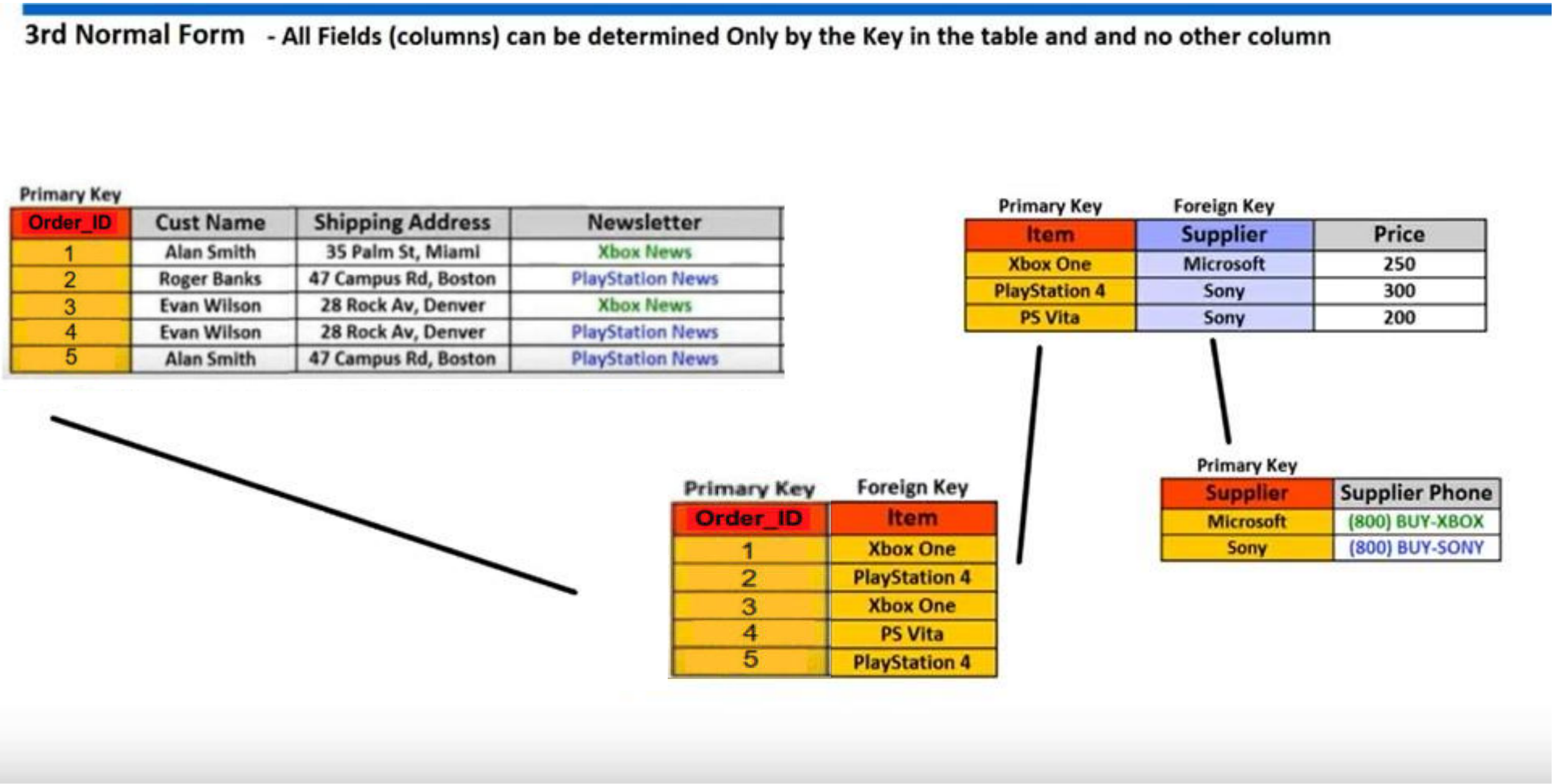
Item	Supplier	Supplier Phone	Price
Xbox One	Microsoft	(800) BUY-XBOX	250
PlayStation 4	Sony	(800) BUY-SONY	300
PS Vita	Sony	(800) BUY-SONY	200

Primary Key / Foreign Key

Order_ID	Item
1	Xbox One
2	PlayStation 4
3	Xbox One
4	PS Vita
5	PlayStation 4

Extracted and modified from source: https://ericdatascience.wordpress.com/2019/06/12/database-normalisation

3NF



3rd Normal Form - All Fields (columns) can be determined Only by the Key in the table and and no other column

Primary Key

Order_ID	Cust Name	Shipping Address	Newsletter
1	Alan Smith	35 Palm St, Miami	Xbox News
2	Roger Banks	47 Campus Rd, Boston	PlayStation News
3	Evan Wilson	28 Rock Av, Denver	Xbox News
4	Evan Wilson	28 Rock Av, Denver	PlayStation News
5	Alan Smith	47 Campus Rd, Boston	PlayStation News

Primary Key / Foreign Key

Item	Supplier	Price
Xbox One	Microsoft	250
PlayStation 4	Sony	300
PS Vita	Sony	200

Primary Key / Foreign Key

Order_ID	Item
1	Xbox One
2	PlayStation 4
3	Xbox One
4	PS Vita
5	PlayStation 4

Primary Key

Supplier	Supplier Phone
Microsoft	(800) BUY-XBOX
Sony	(800) BUY-SONY

Extracted and modified from source: https://ericdatascience.wordpress.com/2019/06/12/database-normalisation

Relational DB

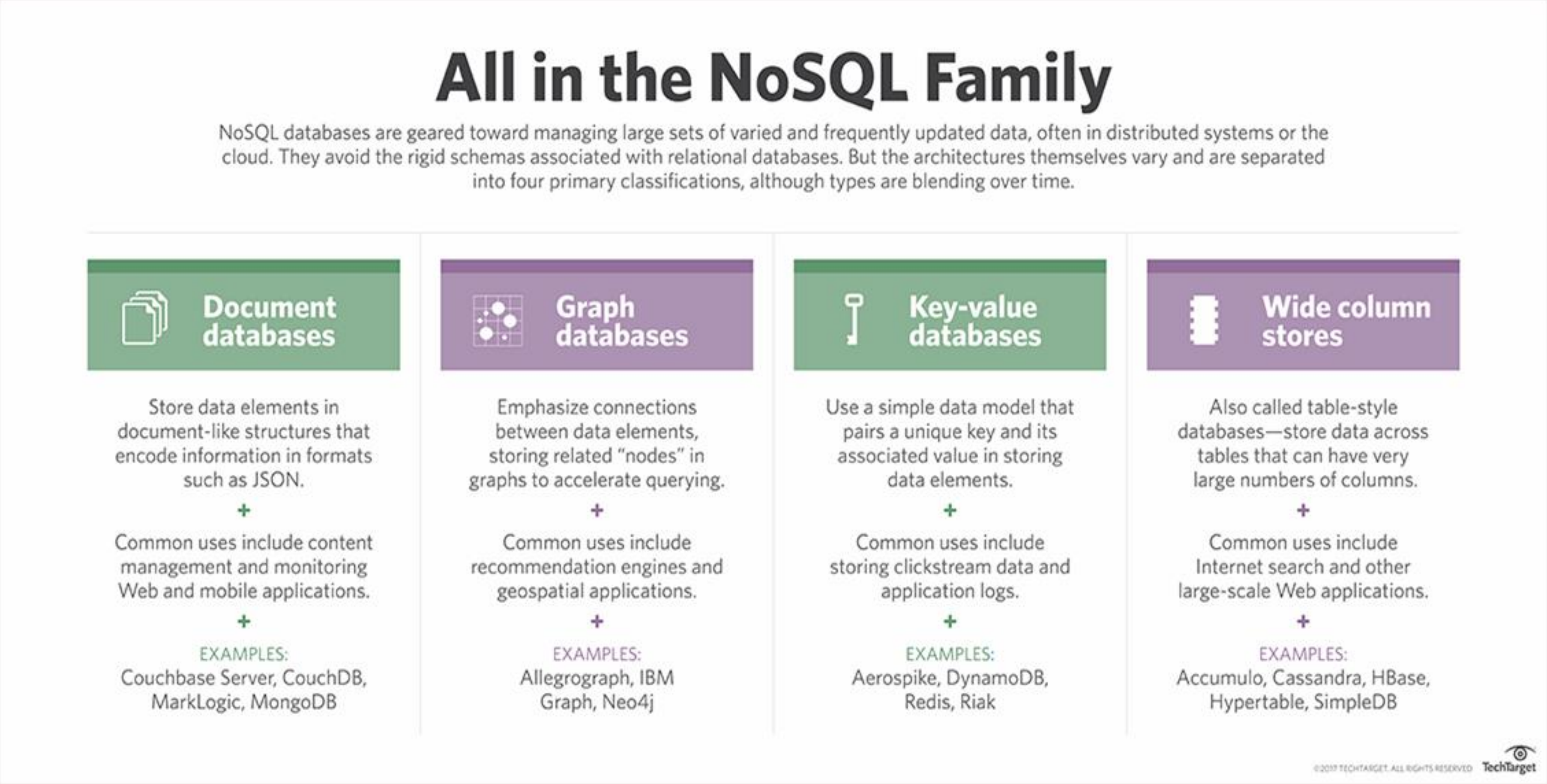
- An important aspect to guarantee reliability of transactions - adherence to the ACID properties:
  a. **Atomicity:** Either all parts of a transaction must be completed or none
  b. **Consistency:** The integrity of the database is preserved by all transactions. DB is not left in invalid state after a transaction
  c. **Isolation**: A transaction must be run isolated in order to guarantee inconsistency in data does not affect other transactions
  d. **Durability**: Changes made by a completed transaction must be preserved or durable

Limitations of Traditional RDBMS

- Traditional RDBMS are not feasible solutions to all data storage problems → obvious limits and difficulties scaling towards Big Data
- Problems include:
  - Slow
  - Scalability issues
  - Unnecessary overhead
  - Poor support for unstructured data
- New technologies emerge to resolve these problems ⇒ NoSQL

Non-Relational DB

- **Non-relational DB** are also known as “**NoSQL DBs**”
- NoSQL databases started gaining popularity in the 2000’s when companies began investing and researching more into distributed databases
- NoSQL DBs represent data in collections of JSON documents E.g. MongoDB
- NoSQL DBs
  - No predefined schema
  - Records can have different fields as necessary (dynamic schema)



All in the NoSQL Family

NoSQL databases are geared toward managing large sets of varied and frequently updated data, often in distributed systems or the cloud. They avoid the rigid schemas associated with relational databases. But the architectures themselves vary and are separated into four primary classifications, although types are blending over time.

Document databases

Store data elements in document-like structures that encode information in formats such as JSON.

+

Common uses include content management and monitoring Web and mobile applications.

+

EXAMPLES:
Couchbase Server, CouchDB, MarkLogic, MongoDB

Graph databases

Emphasize connections between data elements, storing related "nodes" in graphs to accelerate querying.

+

Common uses include recommendation engines and geospatial applications.

+

EXAMPLES:
Allegrograph, IBM Graph, Neo4j

Key-value databases

Use a simple data model that pairs a unique key and its associated value in storing data elements.

+

Common uses include storing clickstream data and application logs.

+

EXAMPLES:
Aerospike, DynamoDB, Redis, Riak

Wide column stores

Also called table-style databases—store data across tables that can have very large numbers of columns.

+

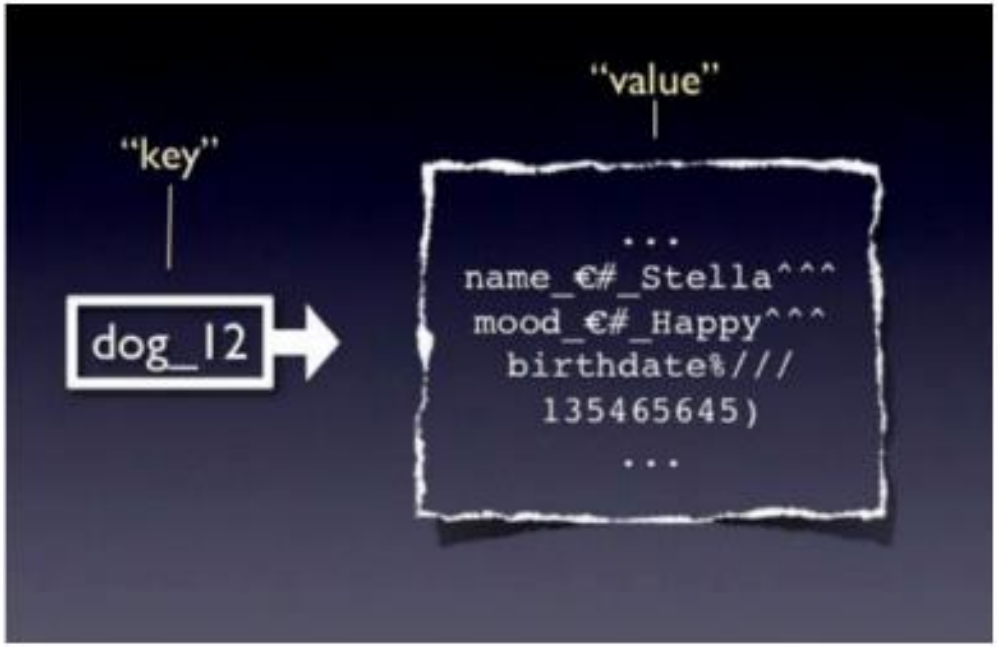
Common uses include Internet search and other large-scale Web applications.

+

EXAMPLES:
Accumulo, Cassandra, HBase, Hypertable, SimpleDB

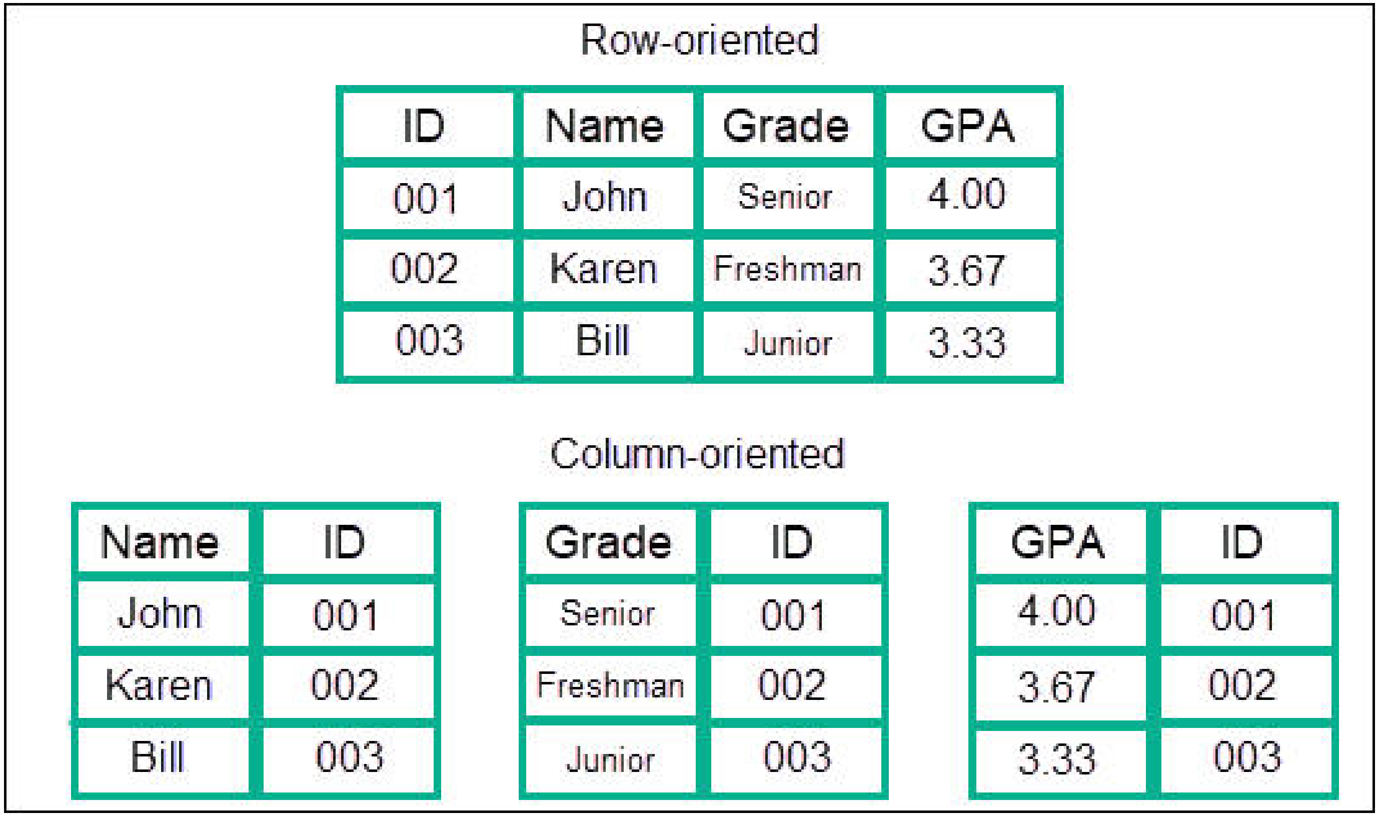
Source: https://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL

Key-value DBs



Key "dog_12": value_name "Stella", value_mood "Happy", etc

Column-based DBs



Row-oriented

ID	Name	Grade	GPA
001	John	Senior	4.00
002	Karen	Freshman	3.67
003	Bill	Junior	3.33

Column-oriented

Name	ID
John	001
Karen	002
Bill	003

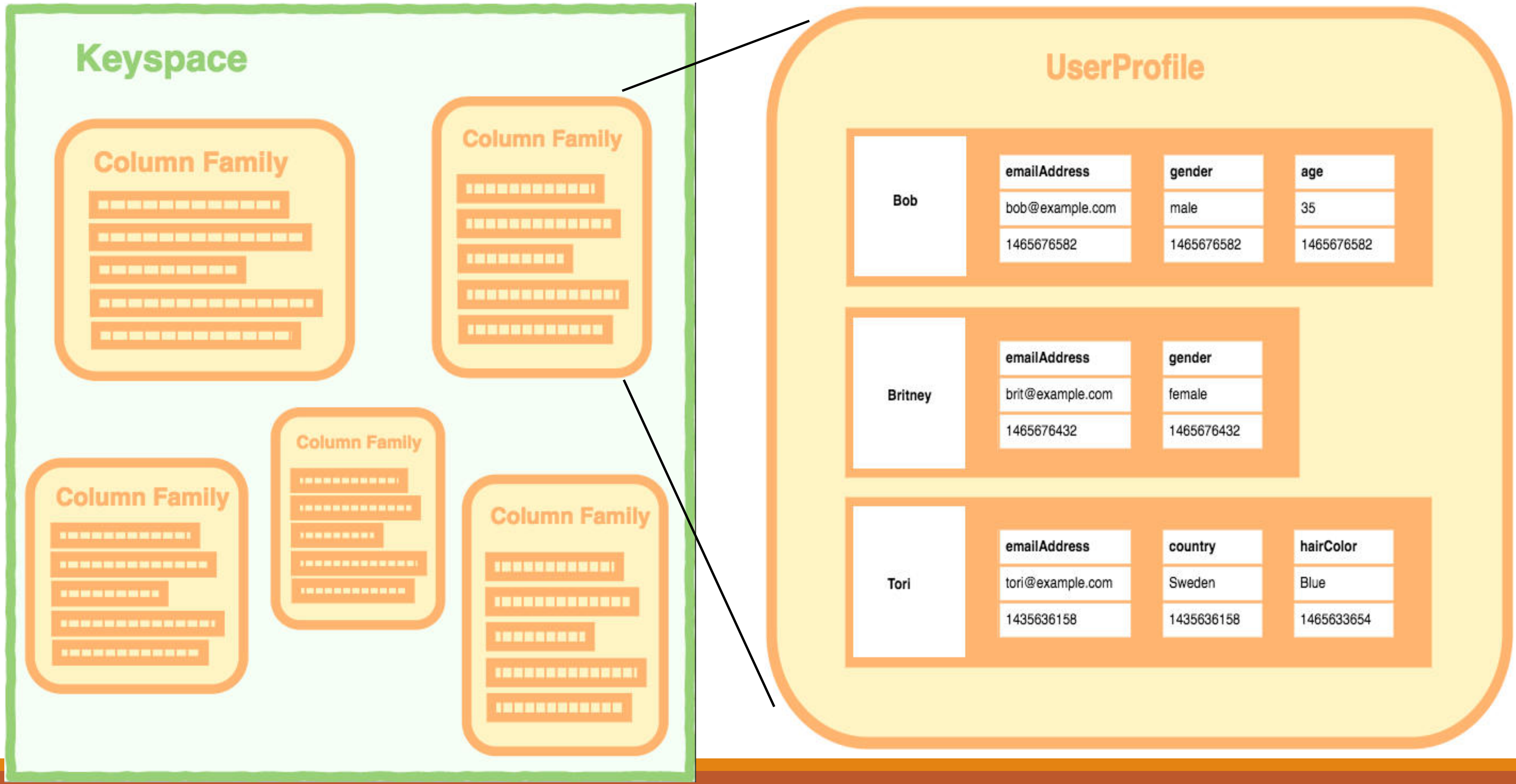
Grade	ID
Senior	001
Freshman	002
Junior	003

GPA	ID
4.00	001
3.67	002
3.33	003

https://www.kdnuggets.com/2021/02/understanding-nosql-database-types-column-oriented-databases.html

Column-based DBs

Use the concept of keyspace. This keyspace contains all the column families, which then contain rows, which then contain columns.



Bob	emailAddress	gender	age
	bob@example.com	male	35
	1465676582	1465676582	1465676582

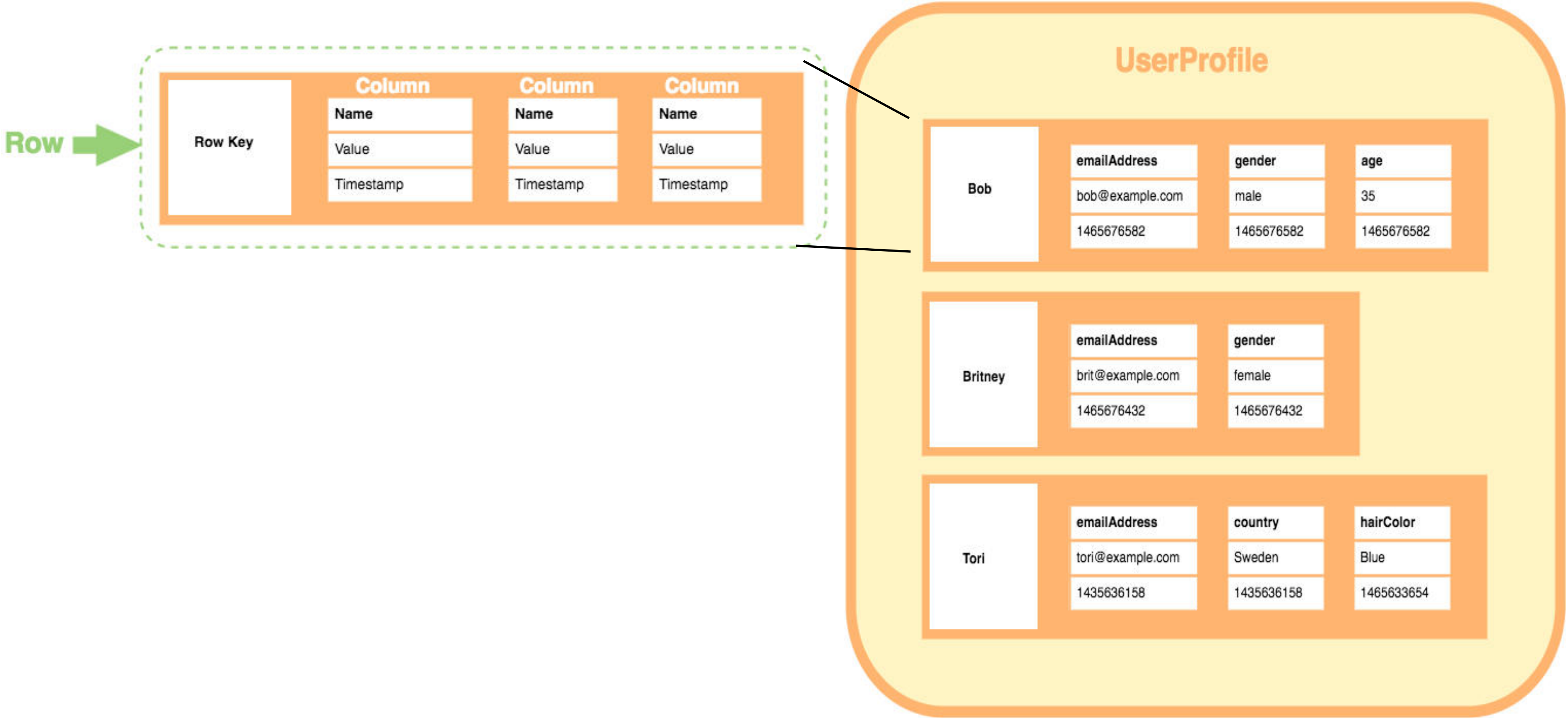
Britney	emailAddress	gender
	brit@example.com	female
	1465676432	1465676432

Tori	emailAddress	country	hairColor
	tori@example.com	Sweden	Blue
	1435636158	1435636158	1465633654

https://www.kdnuggets.com/2021/02/understanding-nosql-database-types-column-oriented-databases.html

Column–based DBs

Use the concept of keyspace. This keyspace contains all the column families, which then contain rows, which then contain columns.



Row Key	Column	Column	Column
	Name	Name	Name
	Value	Value	Value
	Timestamp	Timestamp	Timestamp

UserProfile

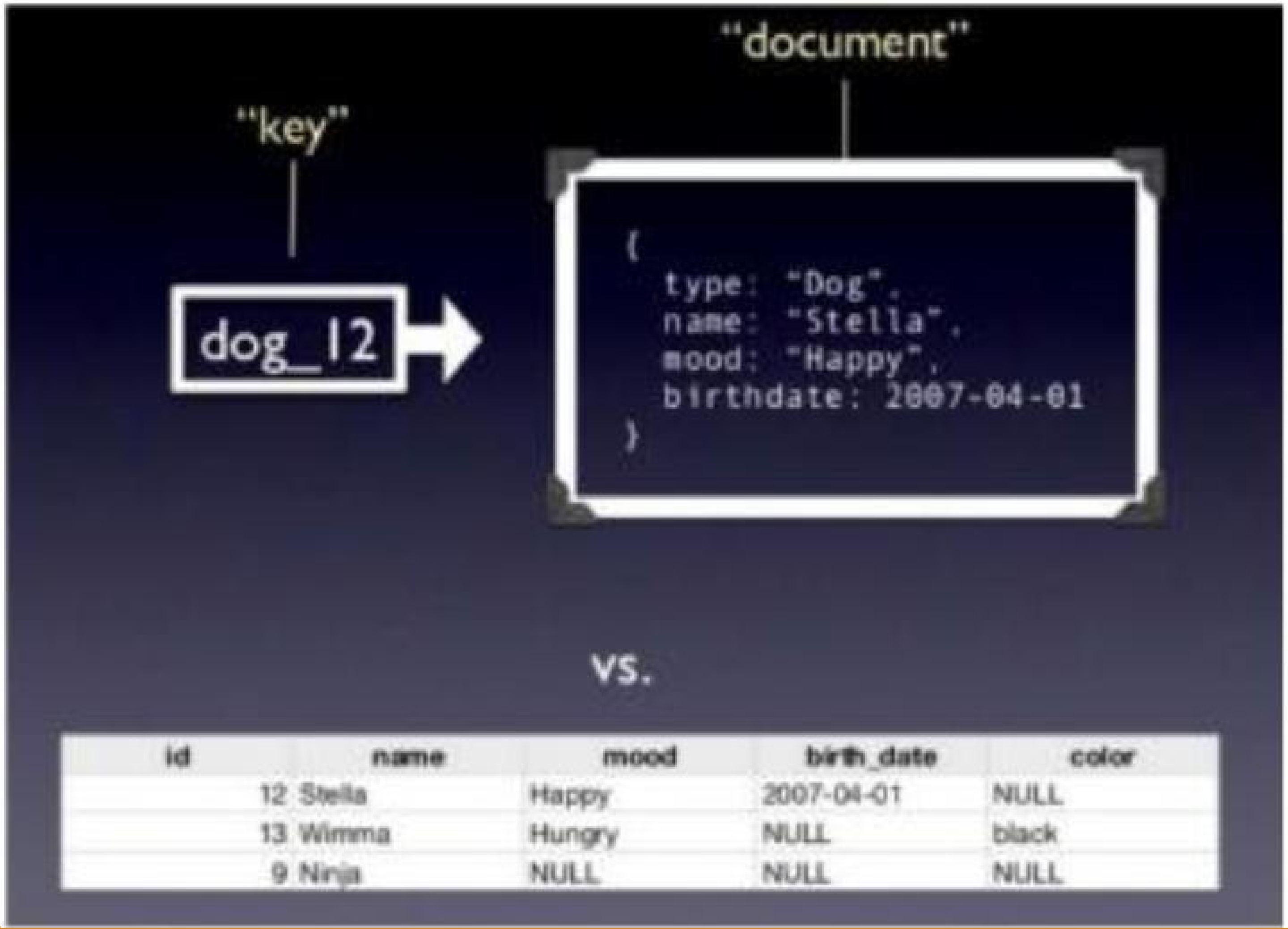
Bob			
	emailAddress	gender	age
	bob@example.com	male	35
	1465676582	1465676582	1465676582

Britney		
	emailAddress	gender
	brit@example.com	female
	1465676432	1465676432

Tori			
	emailAddress	country	hairColor
	tori@example.com	Sweden	Blue
	1435636158	1435636158	1465633654

https://www.kdnuggets.com/2021/02/understanding-nosql-database-types-column-oriented-databases.html

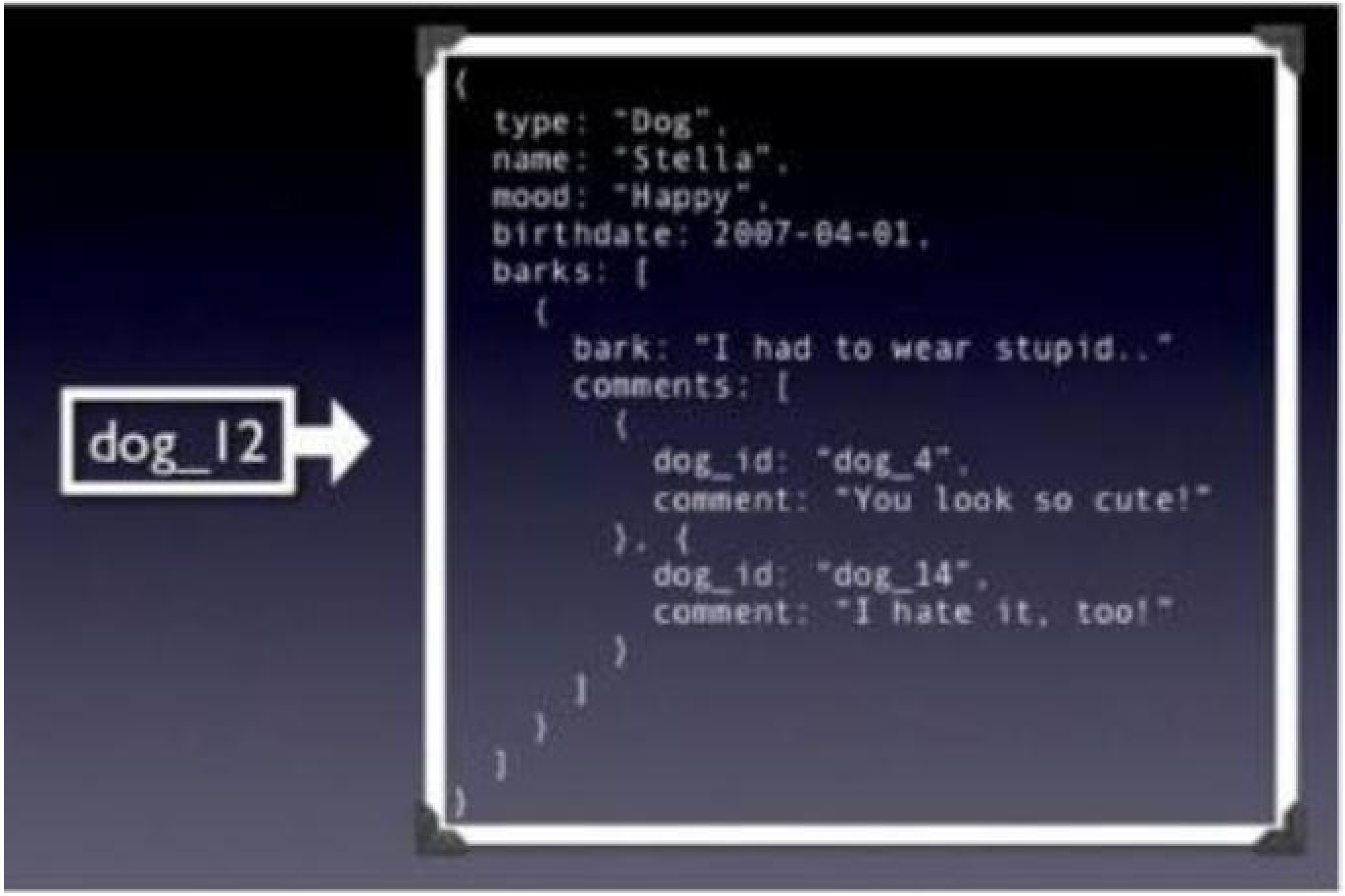
Document-based DBs



```
{
  type: "Dog",
  name: "Stella",
  mood: "Happy",
  birthdate: 2007-04-01
}
```

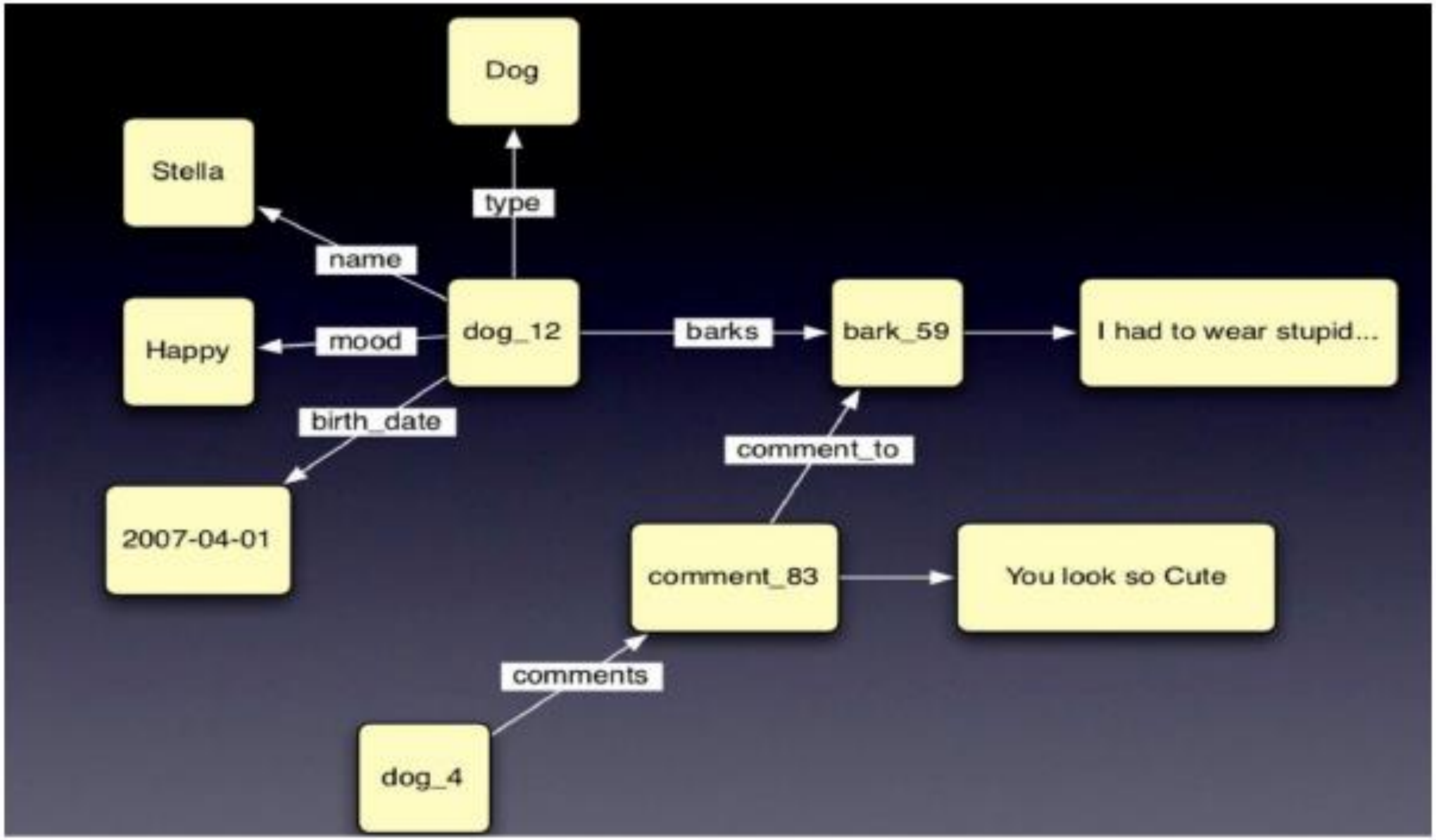
id	name	mood	birth_date	color
12	Stella	Happy	2007-04-01	NULL
13	Wimma	Hungry	NULL	black
9	Ninja	NULL	NULL	NULL

Document-based DBs



```
{
  type: "Dog",
  name: "Stella",
  mood: "Happy",
  birthdate: 2007-04-01,
  barks: [
    {
      bark: "I had to wear stupid.."
      comments: [
        {
          dog_id: "dog_4",
          comment: "You look so cute!"
        }, {
          dog_id: "dog_14",
          comment: "I hate it, too!"
        }
      ]
    }
  ]
}
```

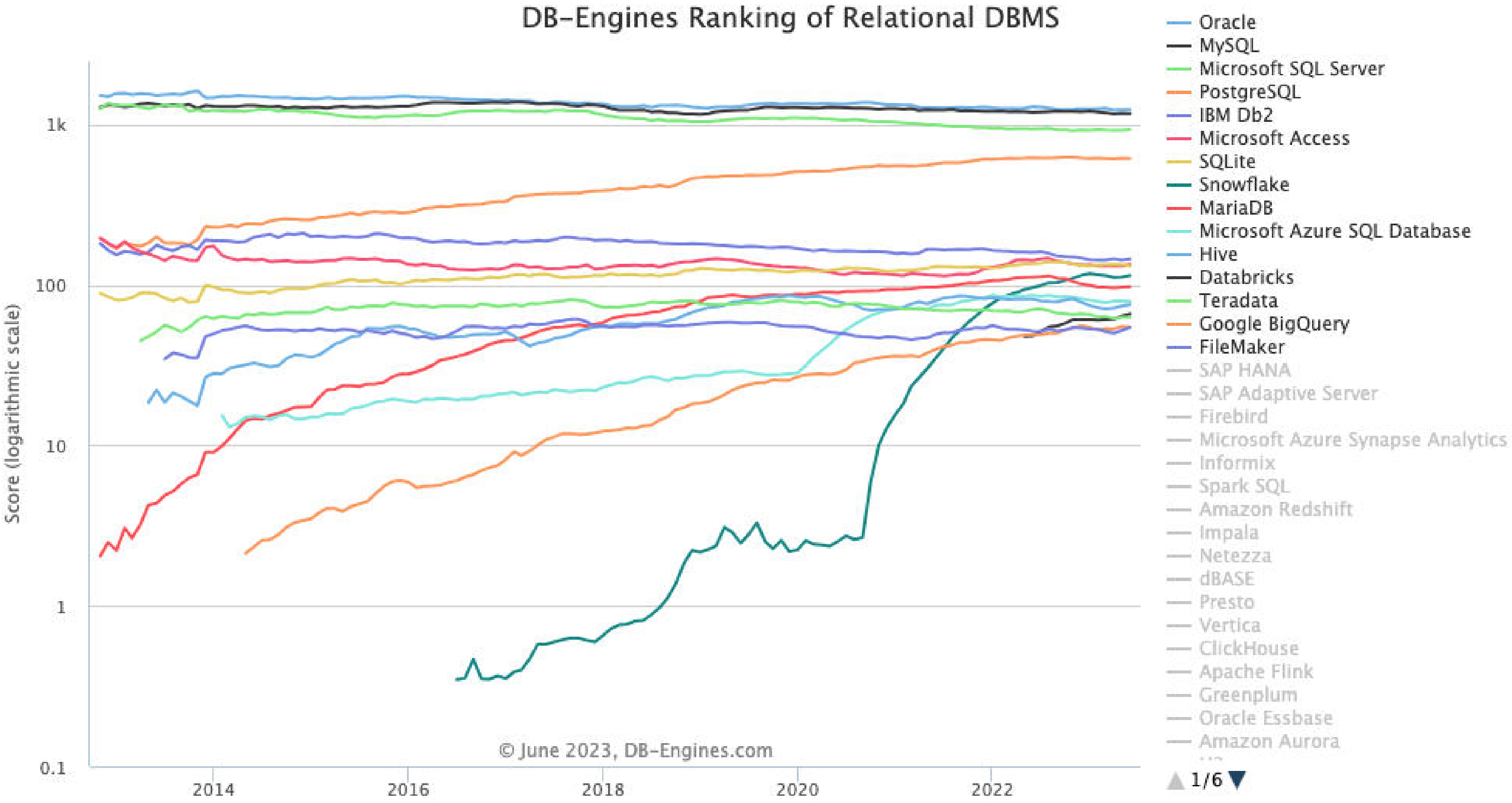
Graph DBs


Dog
Stella
type
name
dog_12
Happy
mood
barks
bark_59
I had to wear stupid...
birth_date
comment_to
2007-04-01
comment_83
You look so Cute
comments
dog_4

Advantages of NoSQL DB

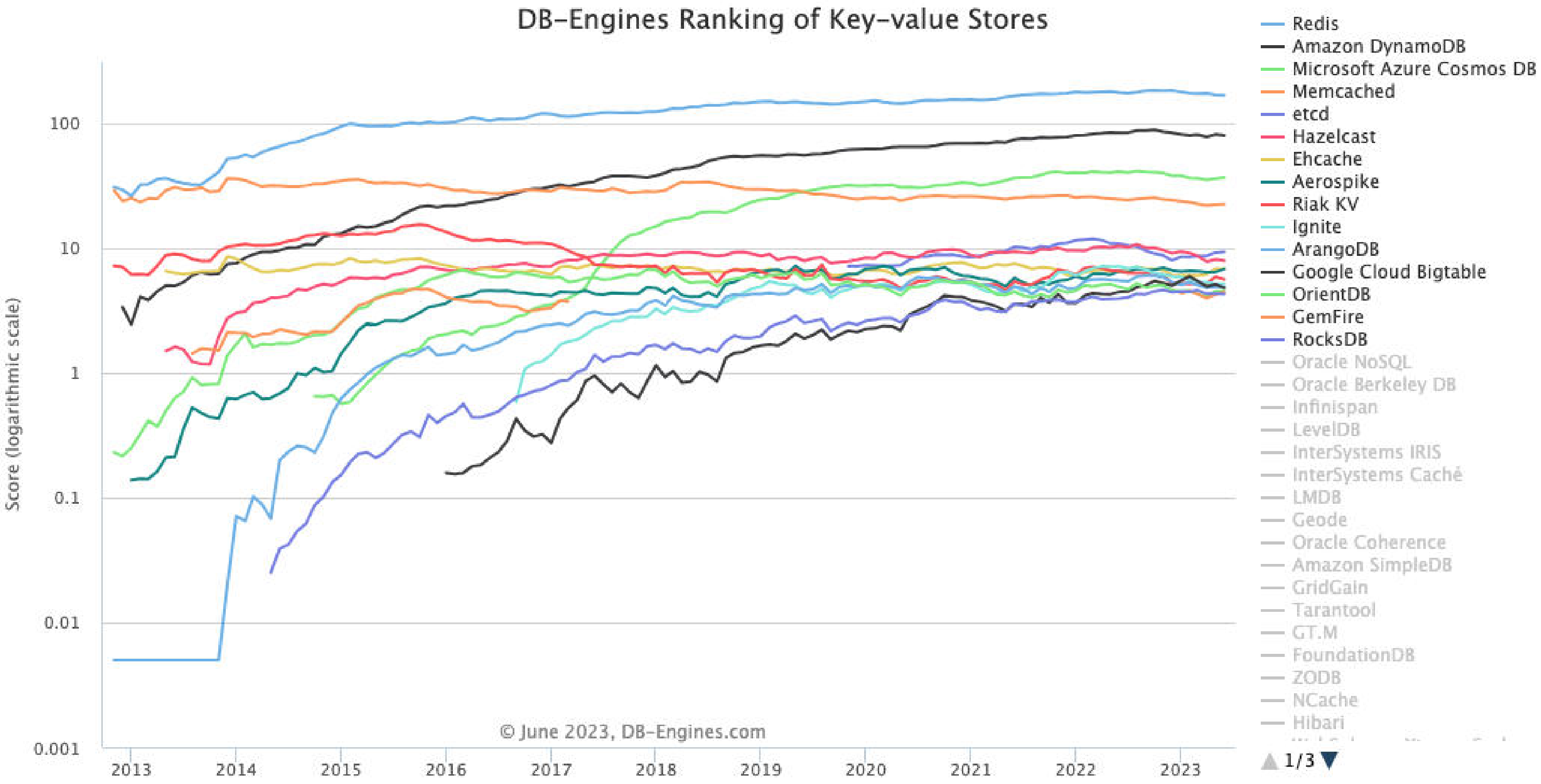
1. NoSQL databases generally process data faster than relational databases
2. NoSQL databases are also often faster because their data models are simpler
3. No schema required: Data can be inserted in a NoSQL database without first defining a rigid database schema. This provides immense flexibility, which ultimately delivers substantial business flexibility.

Comparison of DB Engines



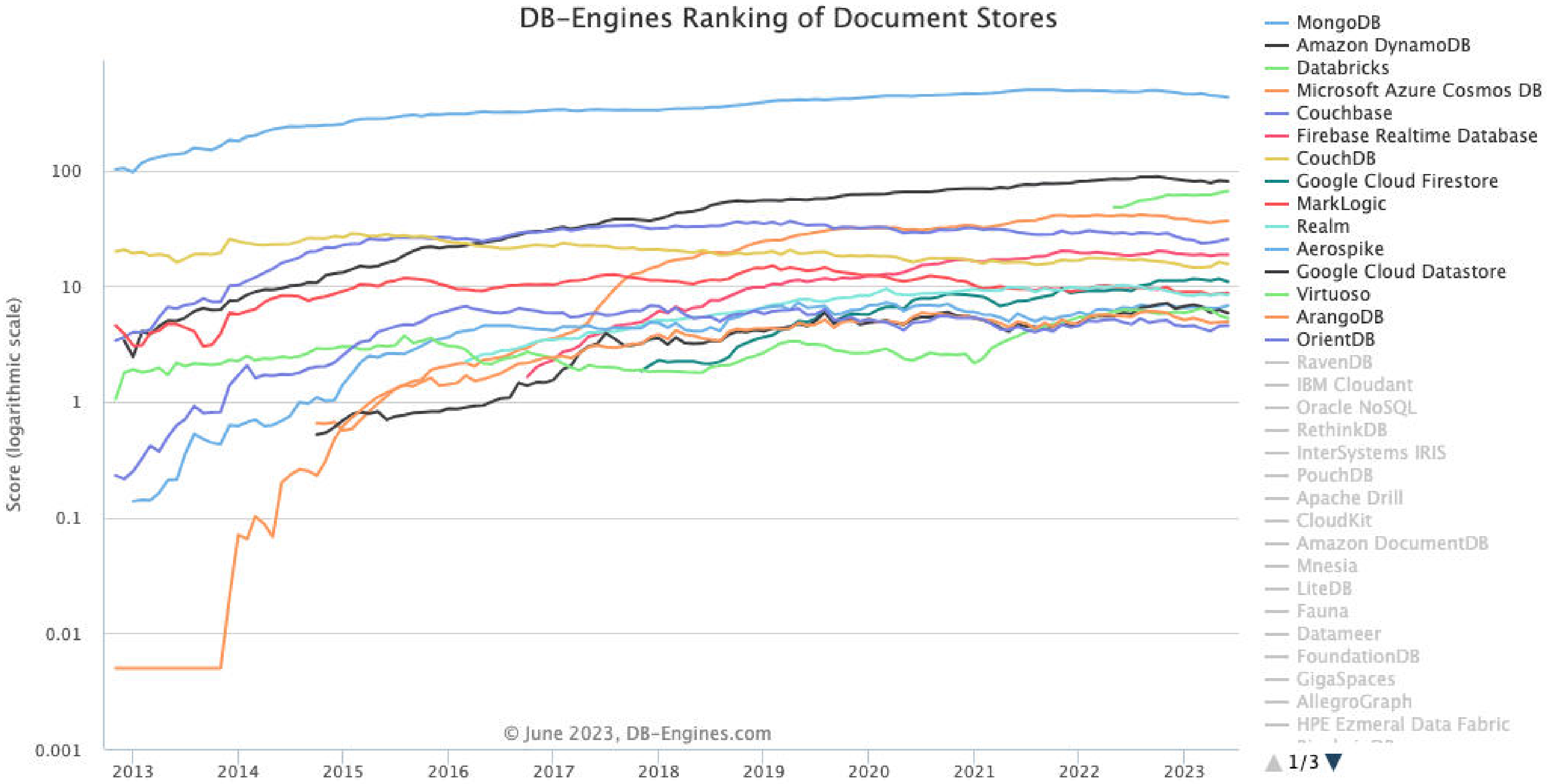
Source: https://www.linkedin.com/pulse/database-engine-ranking-popularity-2023-tamer-khraisha-ph-d-

Comparison of DB Engines



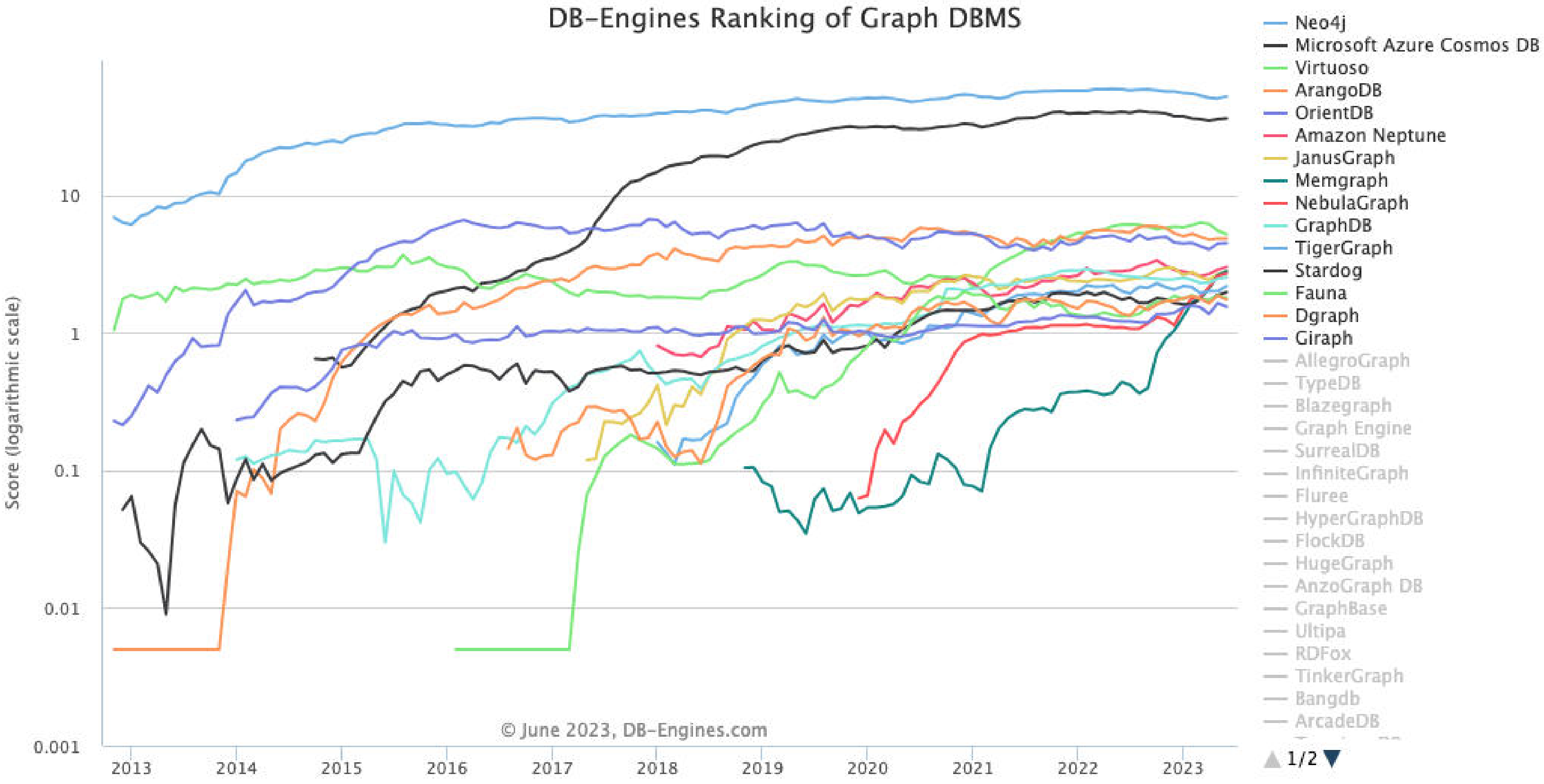
Source: https://www.linkedin.com/pulse/database-engine-ranking-popularity-2023-tamer-khraisha-ph-d-

Comparison of DB Engines



Source: https://www.linkedin.com/pulse/database-engine-ranking-popularity-2023-tamer-khraisha-ph-d-

Comparison of DB Engines



Source: https://www.linkedin.com/pulse/database-engine-ranking-popularity-2023-tamer-khraisha-ph-d-

End of Lecture 2