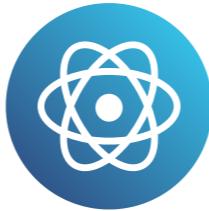


Data sources

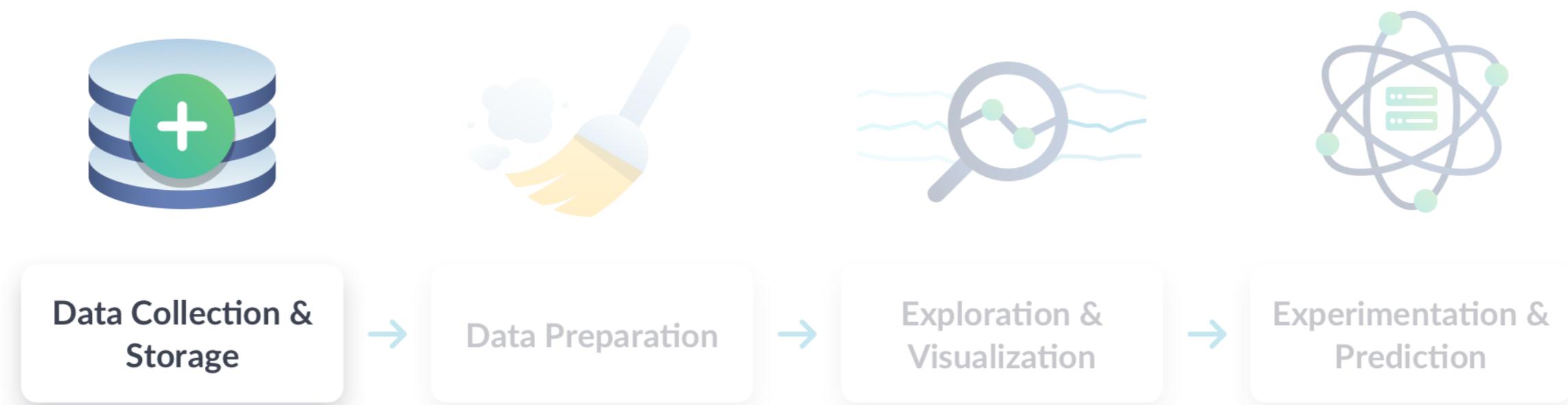
DATA SCIENCE FOR EVERYONE



Sara Billen

Curriculum Manager

The data science workflow



Sources of data

Company data

- Collected by companies
- Helps them make data-driven decisions



Open data

- Free, open data sources
- Can be used, shared, and built-on by anyone



Company data

- Web events
- Survey data
- Customer data
- Logistics data
- Financial transactions



Web data

| event_name | timestamp | user_id |
|----------------|---------------------|---------|
| homepage_visit | 2019-01-01 12:01:01 | 1234 |

- Events
- Timestamps
- User information

Survey data

- Asking people for their opinions
- Methods:
 - Face-to-face interview
 - Online questionnaire
 - Focus group



Net Promotor Score

We appreciate your feedback! X

Thank you for visiting our website. We are always looking for ways to improve your experience. Please take a moment to tell us about your experience.

How likely are you to recommend our website to a friend or colleague?

0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10

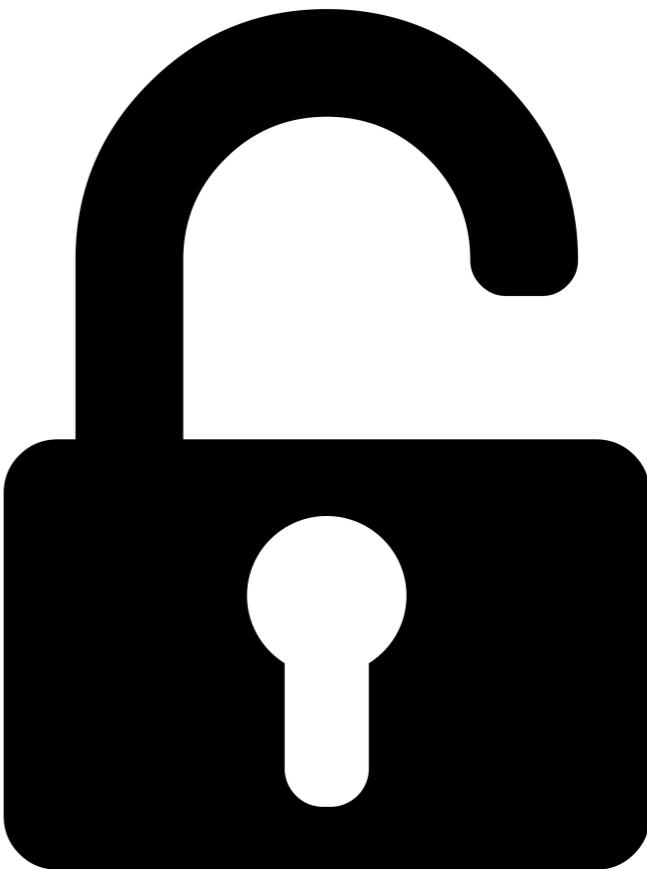
What could we do to improve your experience?

Send Feedback

powered by  QuestionPro

Open data

- Data APIs
- Public records



Public data APIs

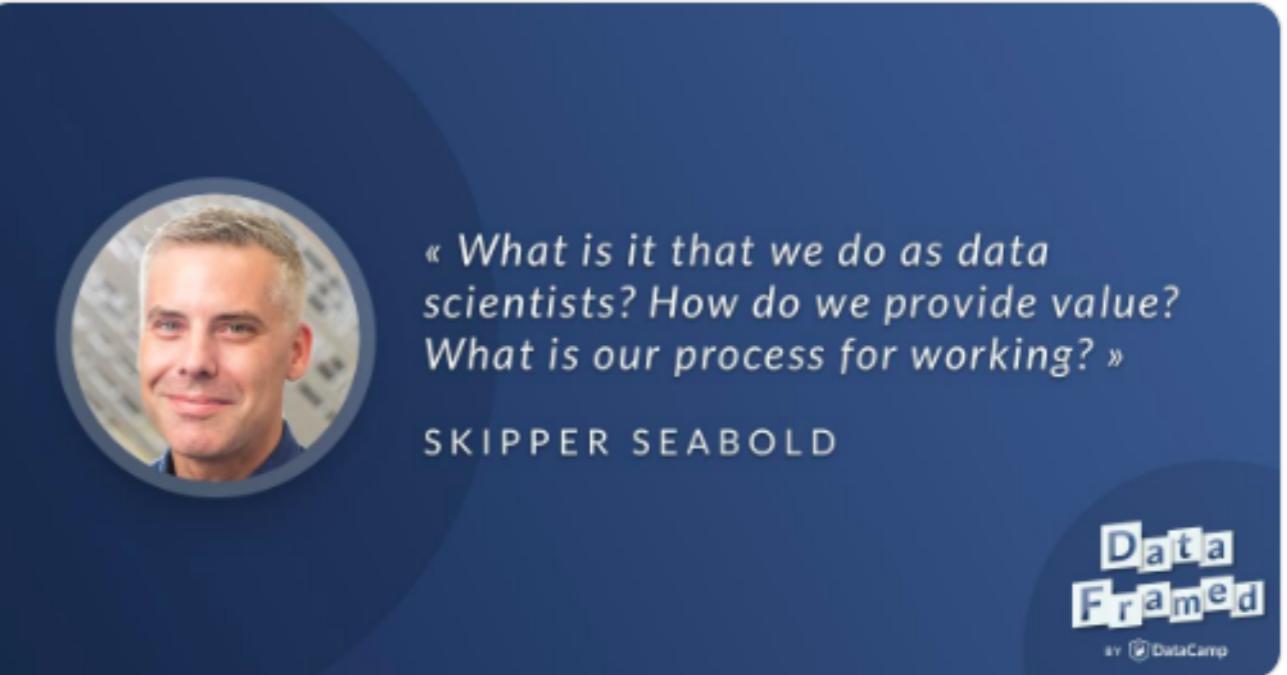
- Application Programming Interface
- Request data over the internet
- Twitter
- Wikipedia
- Yahoo! Finance
- Google Maps
- Many more!

Tracking a hashtag

- All tweets with **#DataFramed** (DataCamp's podcast!)
- Use Twitter API

 **Hugo Bowne-Anderson** @hugobowne · Mar 15

Coming at your ears next Monday -- [@jseabold](#) will break down for you the current and looming credibility crisis in [#datascience](#) on [#DataFramed](#), the [@DataCamp](#) pod.



A circular promotional graphic for the DataFramed podcast. It features a portrait of Skipper Seabold in the center. To the right of the portrait is a quote: « What is it that we do as data scientists? How do we provide value? What is our process for working? ». Below the quote is the name « SKIPPER SEABOLD ». In the bottom right corner of the graphic, there is a logo for « Data Framed » with the text « BY DataCamp » underneath.

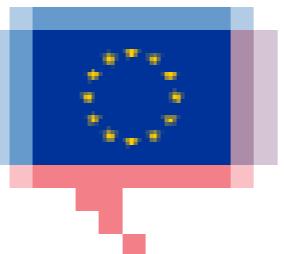
1 4 21 1

Public records

- International organizations
 - e.g.: World Bank, UN, WTO
- National statistical offices
 - e.g.: censuses, surveys
- Government agencies
 - e.g.: weather, environment, population



- For the US, [data.gov](https://www.data.gov)
- For the EU, data.europa.eu



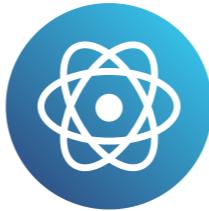
EU **Open Data** Portal

Let's practice!

DATA SCIENCE FOR EVERYONE

Data types

DATA SCIENCE FOR EVERYONE



Sara Billen

Curriculum manager

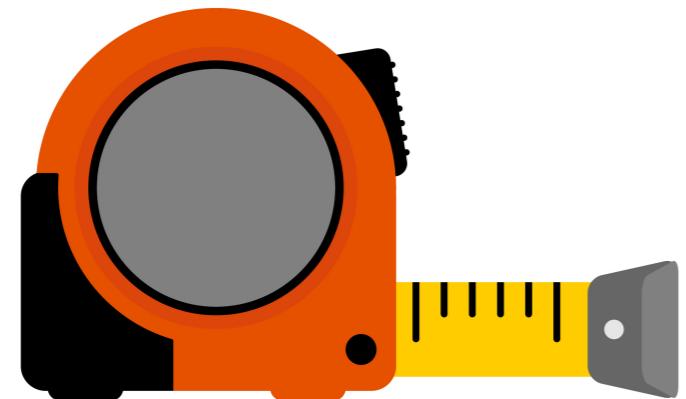
Why care about data types?

- Important later on when:
 - Storing the data
 - Visualizing/analyzing the data

Quantitative vs qualitative data

Quantitative data

- Deals with numbers
- Data can be measured



Qualitative data

- Deals with descriptions
- Data can be observed but not measured



Quantitative data



- Is 60 inches tall
- Has 2 apples in it
- Costs \$1000

Qualitative data

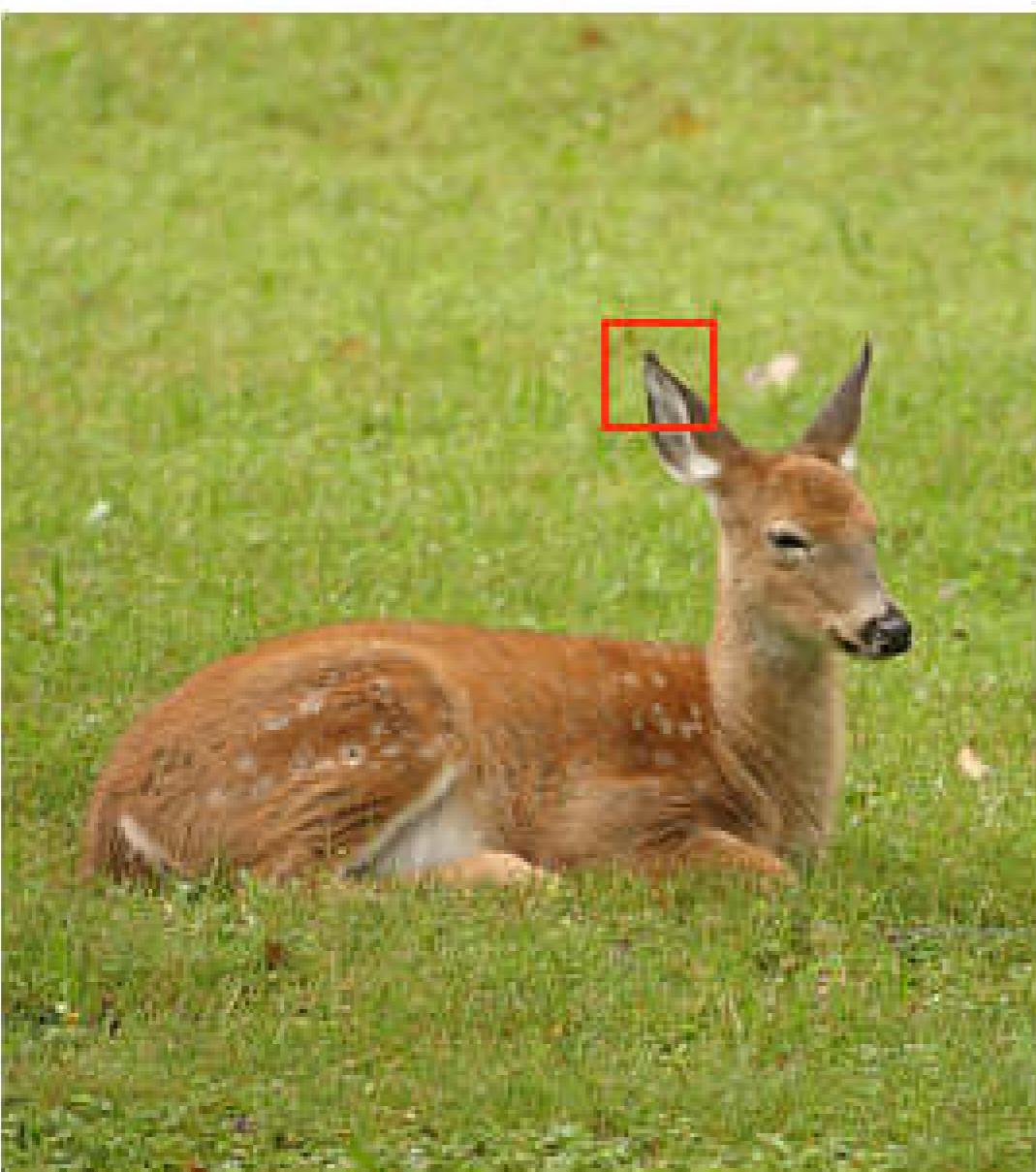


- Is red
- Was built in Italy
- Smells like fish

Other data types

- Image data
- Text data
- Geospatial data
- Network data
- ...

Other data types: Image data



Other data types: Text data

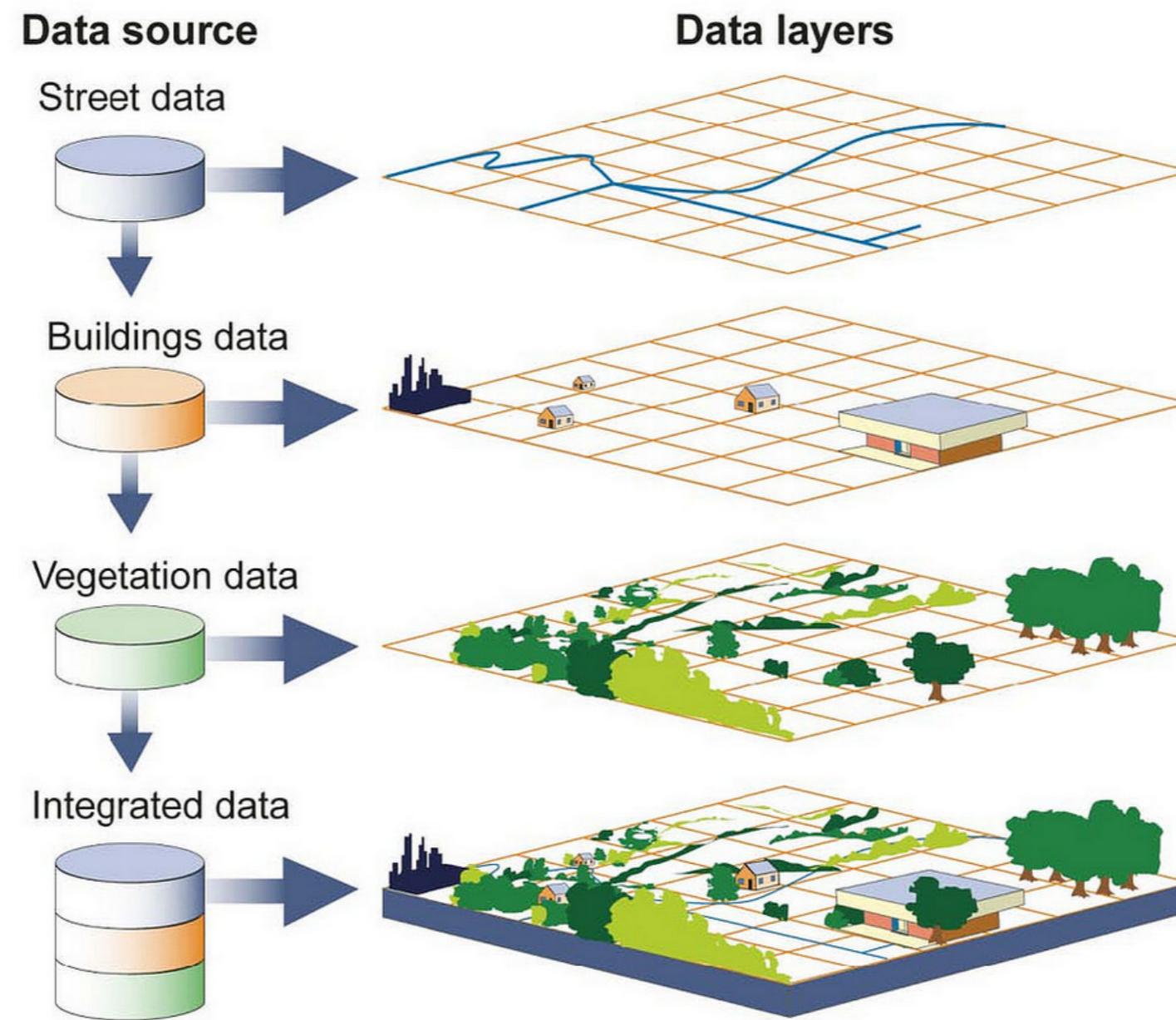
“Great evening, extremely good value”

 Review of [L'Ange 20 Restaurant](#)

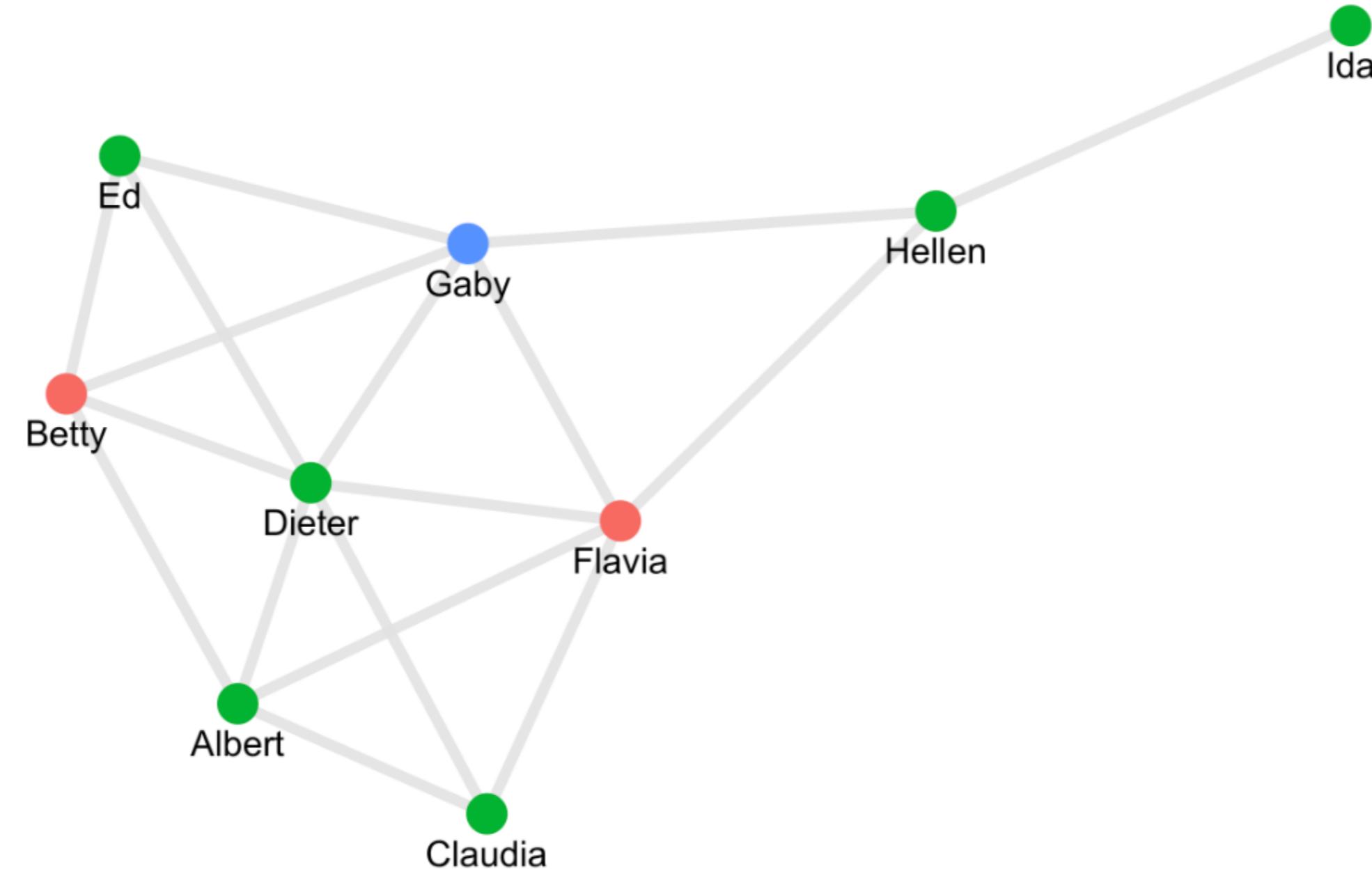
I went to this place with my boyfriend for a special occasion and we were not disappointed. We were greeted warmly by Christopher who guided us through the menu and wine. The food was delicious and I only wish that we could have had room for three courses. The value was excellent compared to other prices we had seen and we found the quality/value and atmosphere hard to match during the rest of our stay.

I had the lamb which I can highly recommend. When we return to Paris we will go back!

Other data types: Geospatial data



Other data types: Network data



Recap

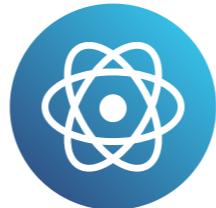
- Quantitative data
- Qualitative data
- Image data
- Text data
- Geospatial data
- Network data

Let's practice!

DATA SCIENCE FOR EVERYONE

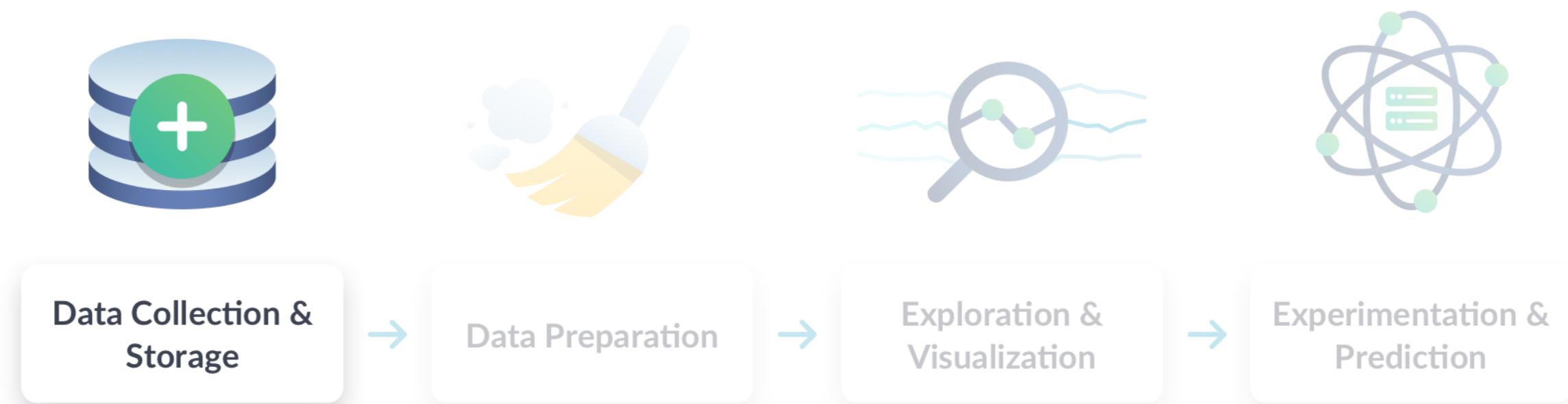
Data storage and retrieval

DATA SCIENCE FOR EVERYONE



Sara Billen
Curriculum Manager

The data science workflow



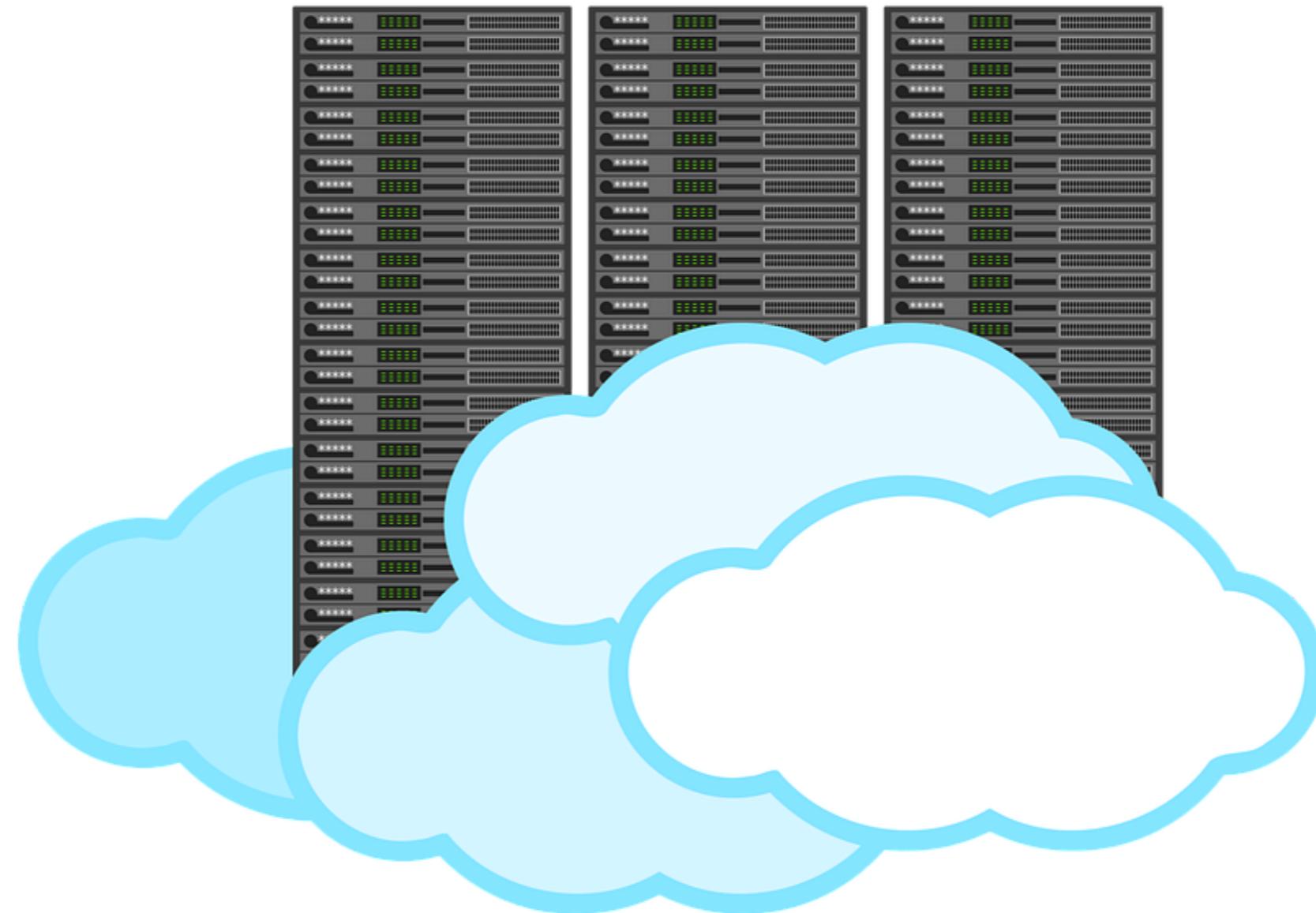
Things to consider when storing data

- Location
- Data type
- Retrieval

Location: Parallel storage solutions



Location: The cloud



Types of data storage

Unstructured

- Email
- Text
- Video and audio files
- Web pages
- Social media

Document Database

Types of data storage

Unstructured

- Email
- Text
- Video and audio files
- Web pages
- Social media

Tabular

| Customer Name | Customer Address | ... |
|---------------|------------------|-----|
| Jane Doe | 123 Maple St. | ... |

Relational Database

Document Database

Retrieval: Data querying



Retrieval: Data querying



| Data Type | Query Language |
|---------------------|----------------|
| Document Database | NoSQL |
| Relational Database | SQL |

Putting it all together: Location



- On-premises cluster
- Cloud provider:
 - Azure
 - AWS
 - Google Cloud

Putting it all together: Data type



Putting it all together: Data type

| Data Type | Storage Solution |
|--------------|---------------------|
| Unstructured | Document Database |
| Tabular | Relational Database |



Putting it all together: Queries



Putting it all together: Queries



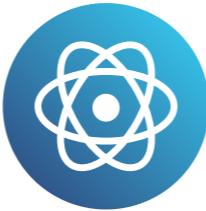
| Data Type | Query Language |
|---------------------|----------------|
| Document Database | NoSQL |
| Relational Database | SQL |

Let's practice!

DATA SCIENCE FOR EVERYONE

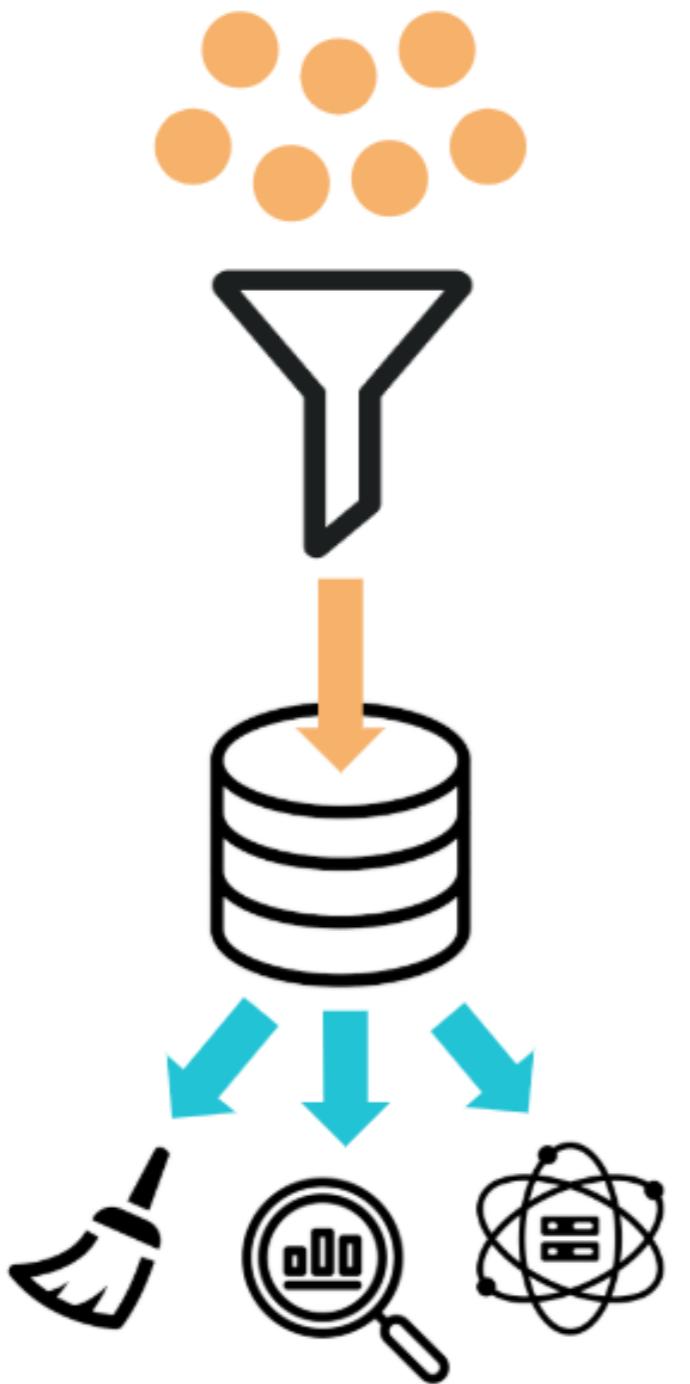
Data Pipelines

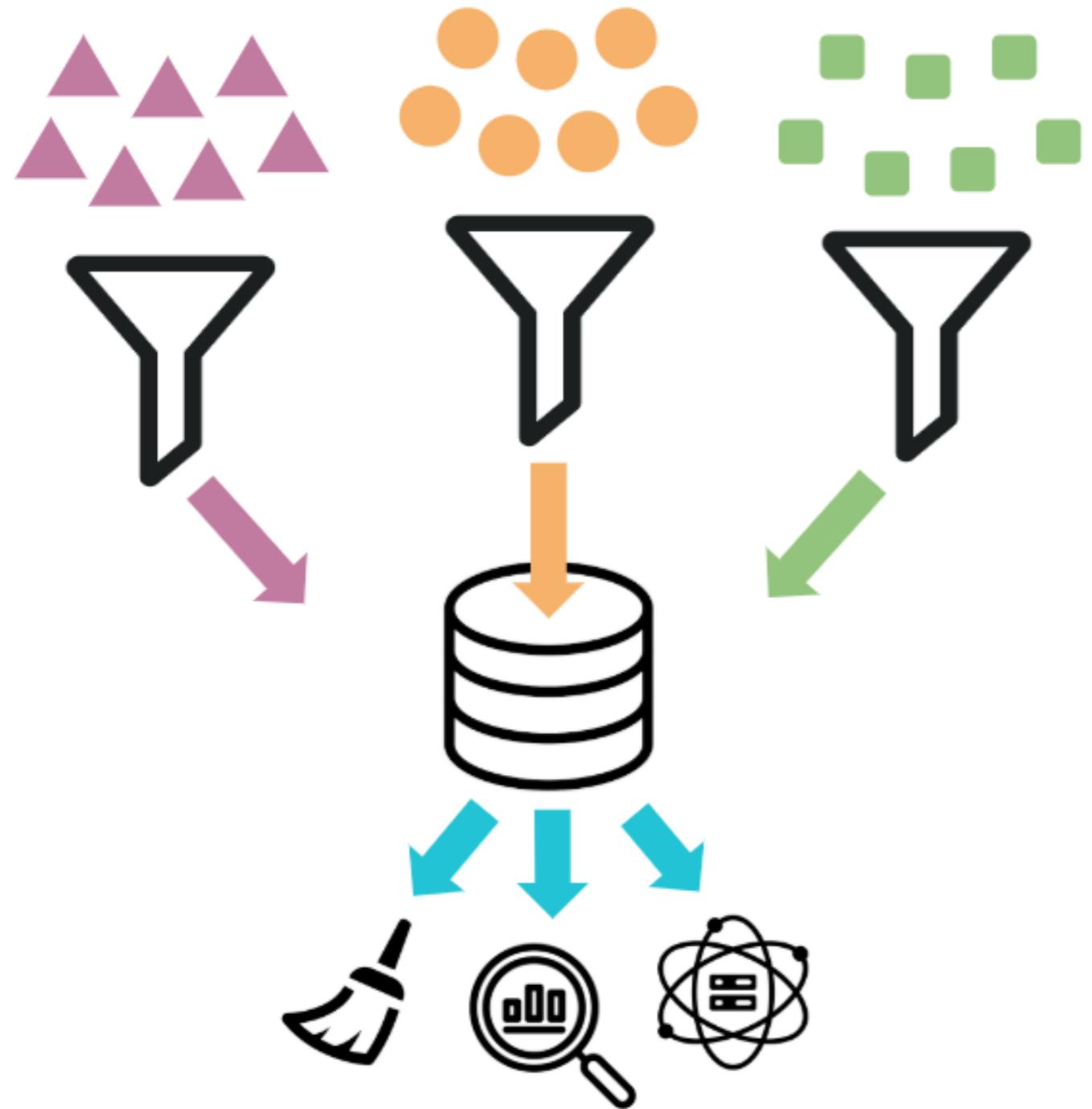
DATA SCIENCE FOR EVERYONE



Sara Billen

Curriculum Manager





How do we scale?

More than one data source:

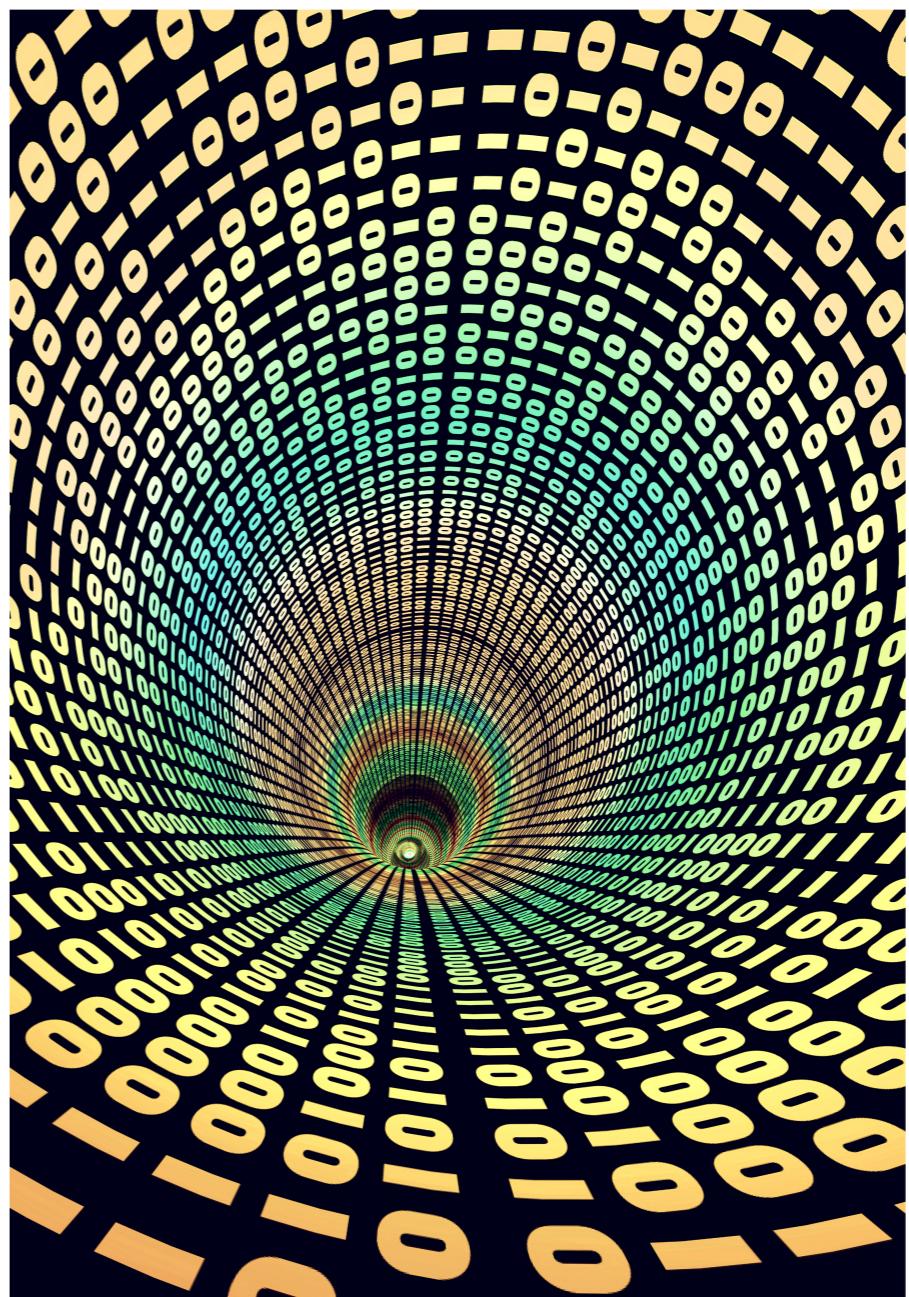
- Public records
- APIs
- Databases

Different data types:

- Unstructured data
- Tabular data
- Real-time streaming data e.g., tweets

What is a data pipeline?

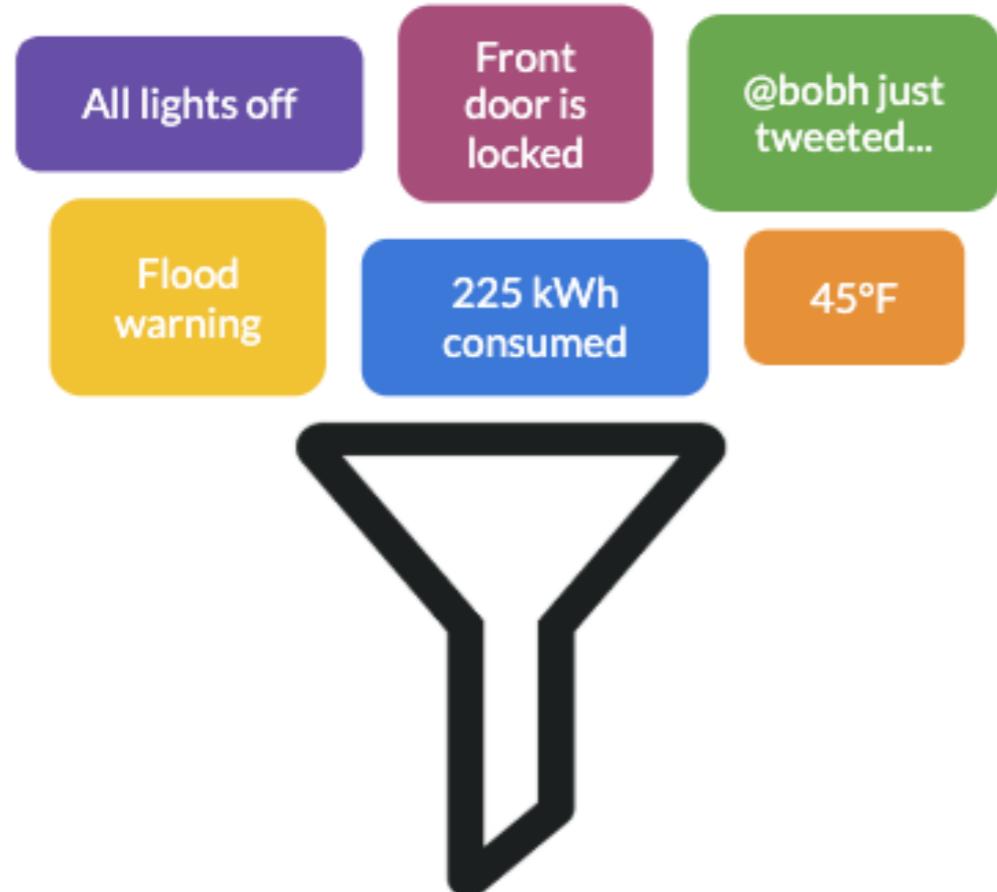
- Moves data into defined stages
- Automated collection and storage
 - *Scheduled hourly, daily, weekly, etc*
 - *Triggered by an event*
- Monitored with generated alerts
- Necessary for big data projects
- Data engineers work to customize solutions
- Extract Transform Load (ETL)



Case study: smart home

| Data | Source | Frequency |
|---------------------|------------------------------|------------------|
| Weather conditions | National Weather Service API | Every 30 minutes |
| Tweets in your area | Twitter API | Real-time stream |
| Indoor temperature | Smart home thermostat | Every 5 minutes |
| Status of lights | Smart light bulbs | Every minute |
| Status of locks | Smart door locks | Every 15 seconds |
| Energy consumption | Smart meter | Weekly |

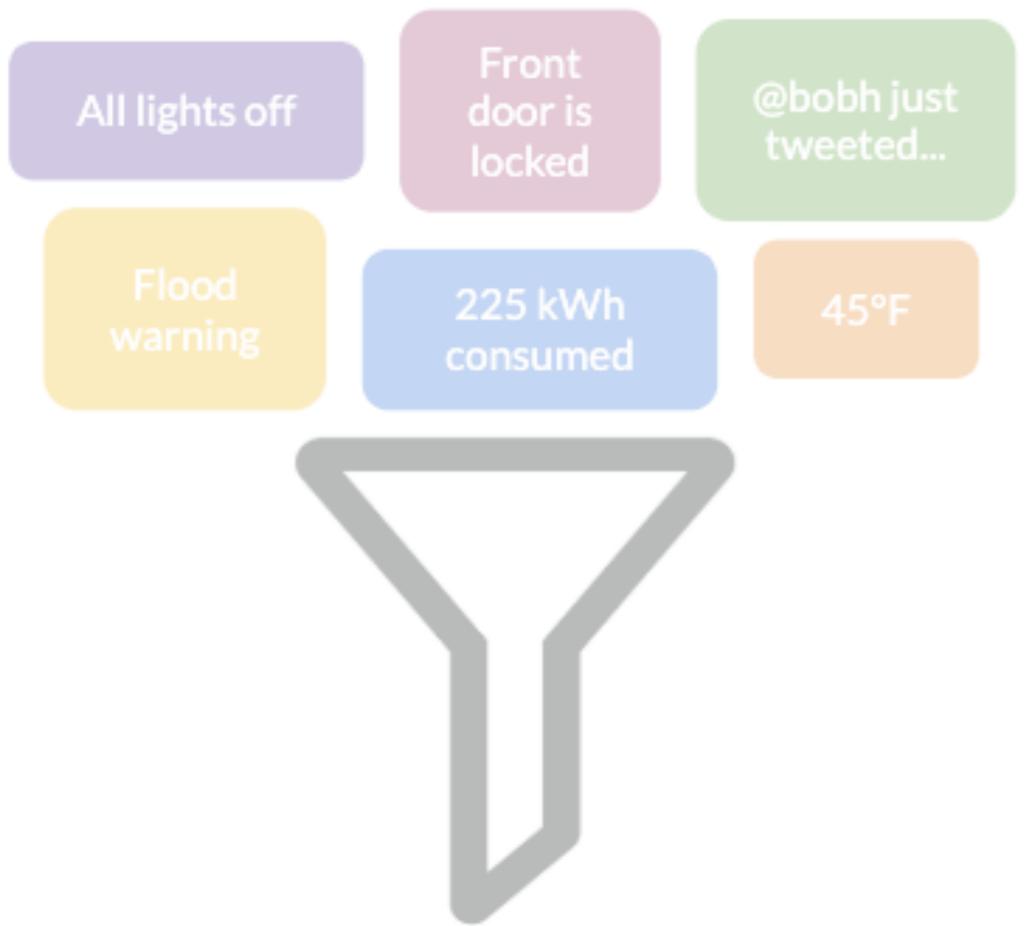
Extract



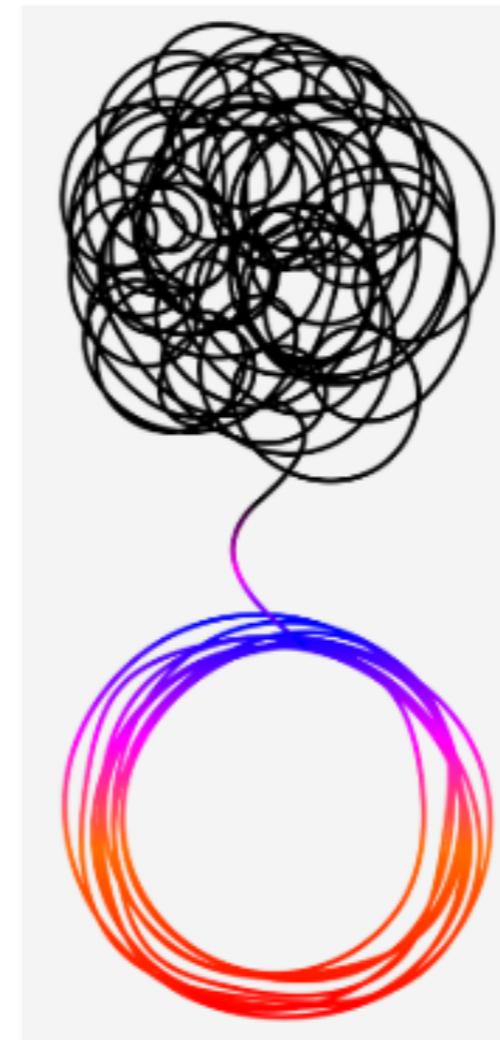
Extract

| Source | Frequency |
|-----------------------|------------------|
| National Weather API | Every 30 minutes |
| Twitter API | Real-time stream |
| Smart home thermostat | Every 5 minutes |
| Smart light bulbs | Every minute |
| Smart door locks | Every 15 seconds |
| Smart meter | Weekly |

Transform



Extract



Transform

Transform

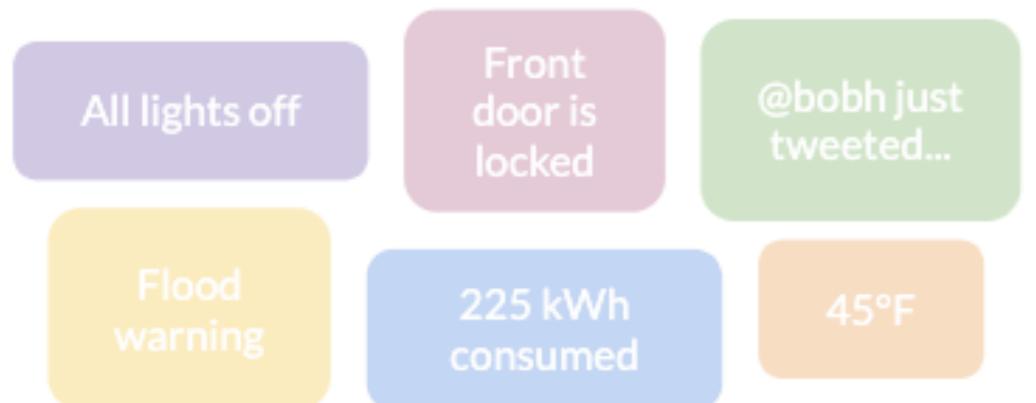
With all the data coming in, how do we keep it organized and easy to use?

Example transformations:

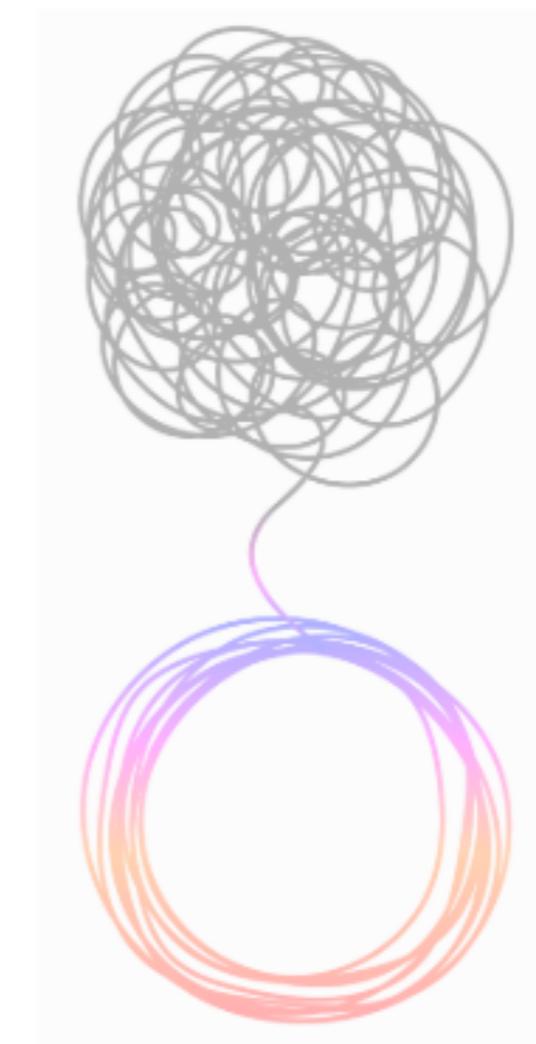
- Joining data sources into one data set
- Converting data structures to fit database schemas
- Removing irrelevant data

Data preparation and exploration does not occur at this stage

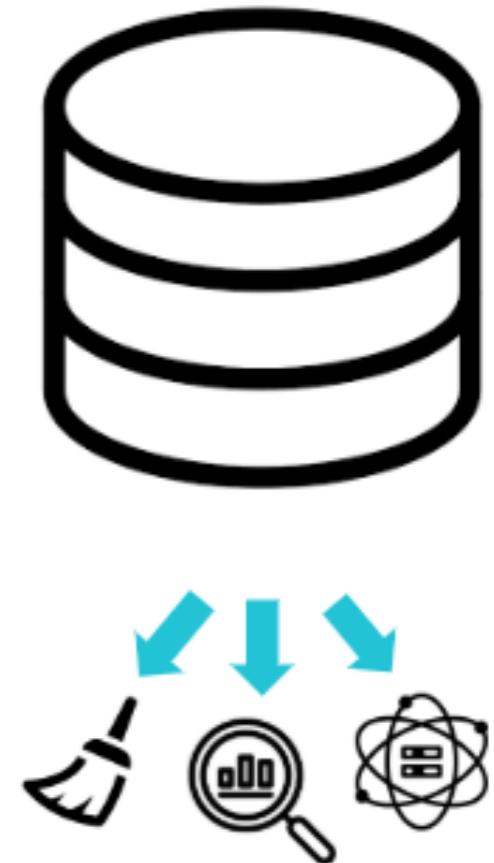
Load



Extract

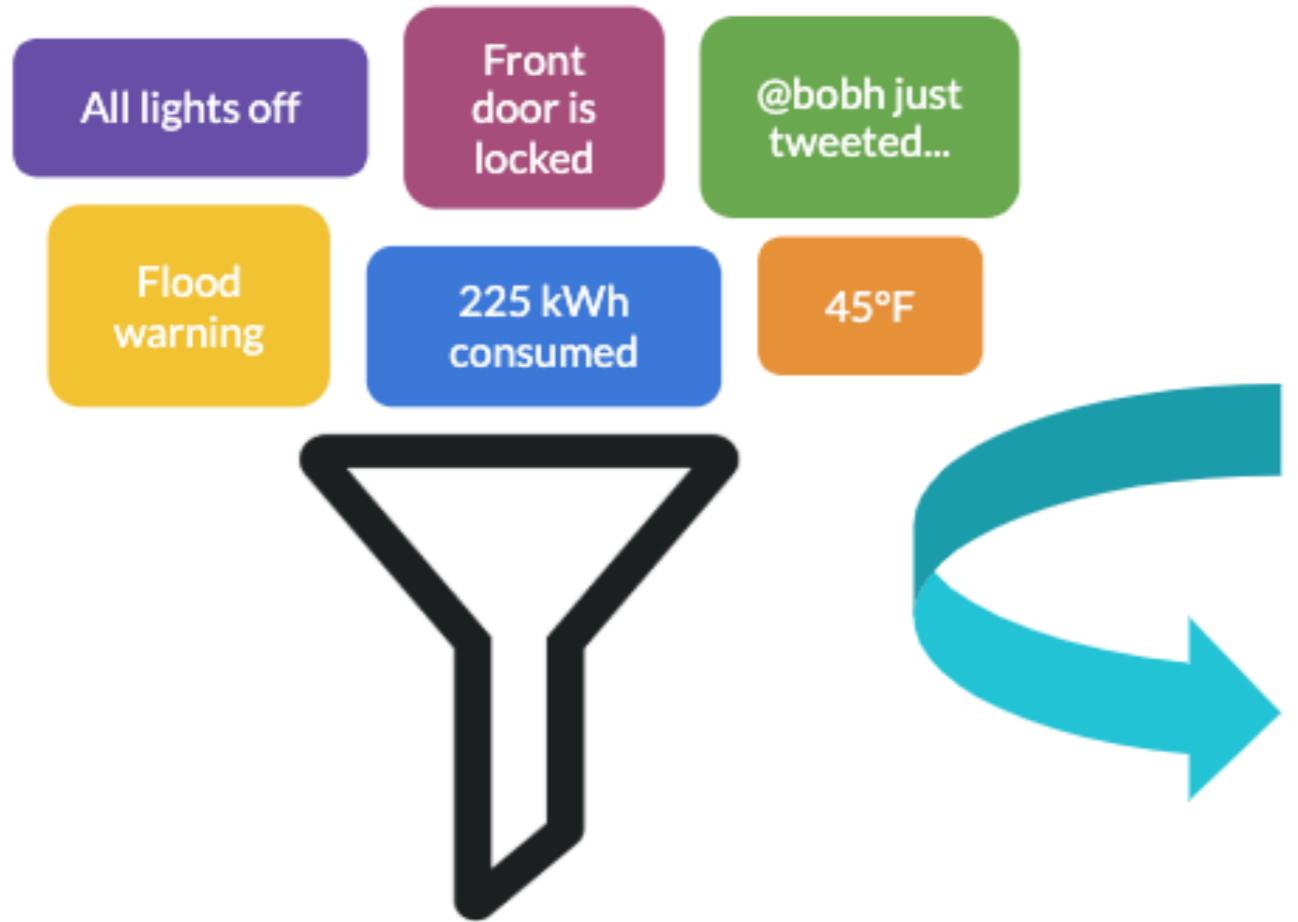


Transform

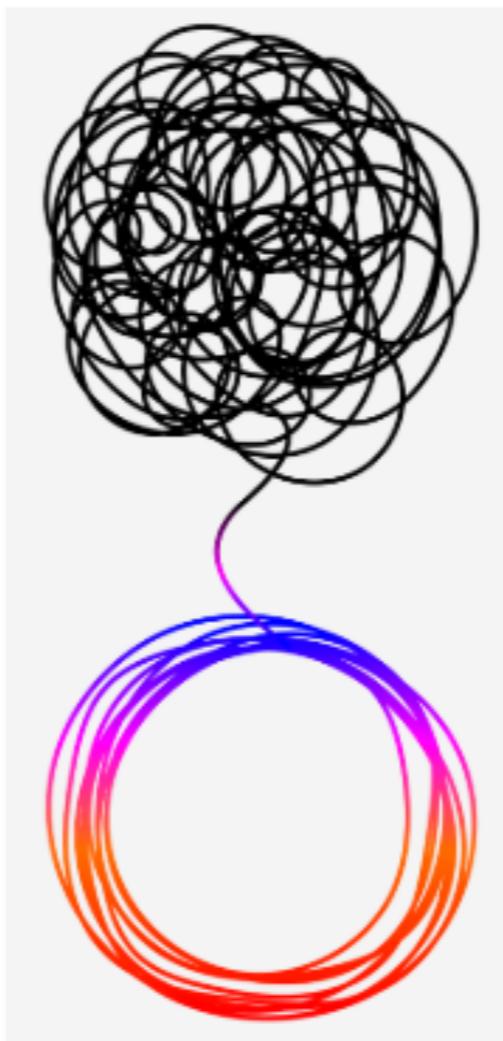


Load

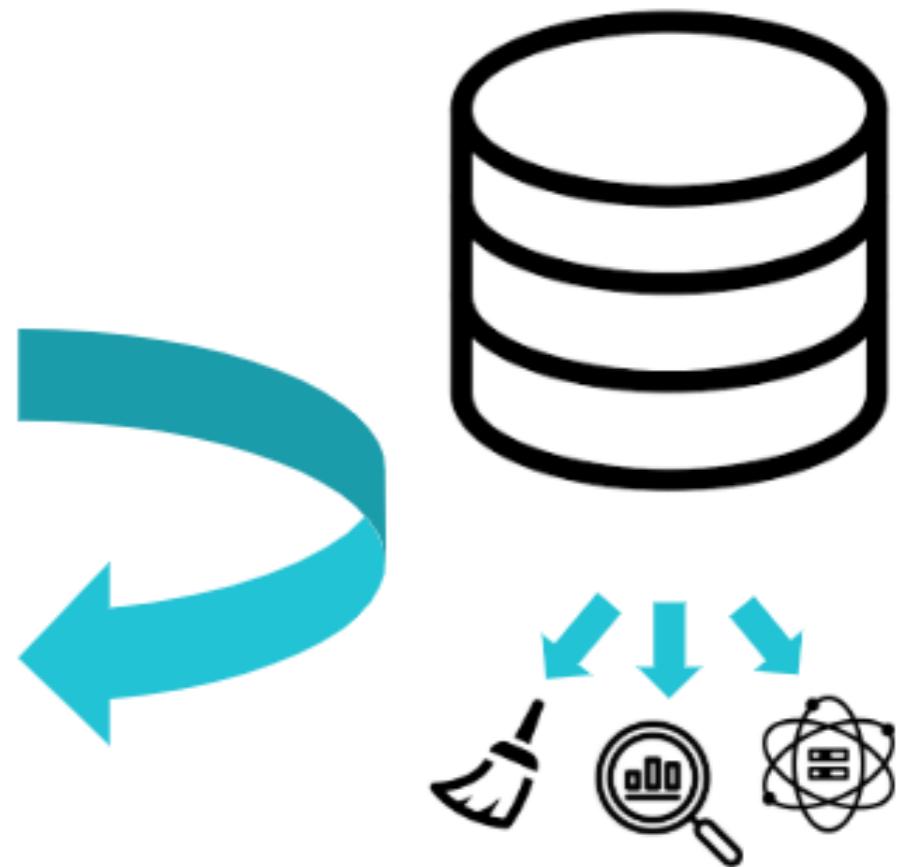
Automation



Extract



Transform



Load

Let's practice!

DATA SCIENCE FOR EVERYONE