What is data science?

Let's ask Google!

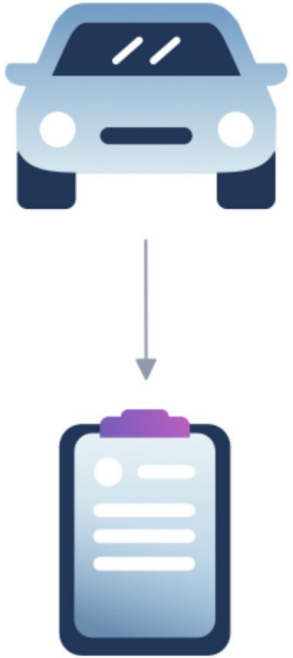
Making data work for you

Use data to better describe the present or better predict the future

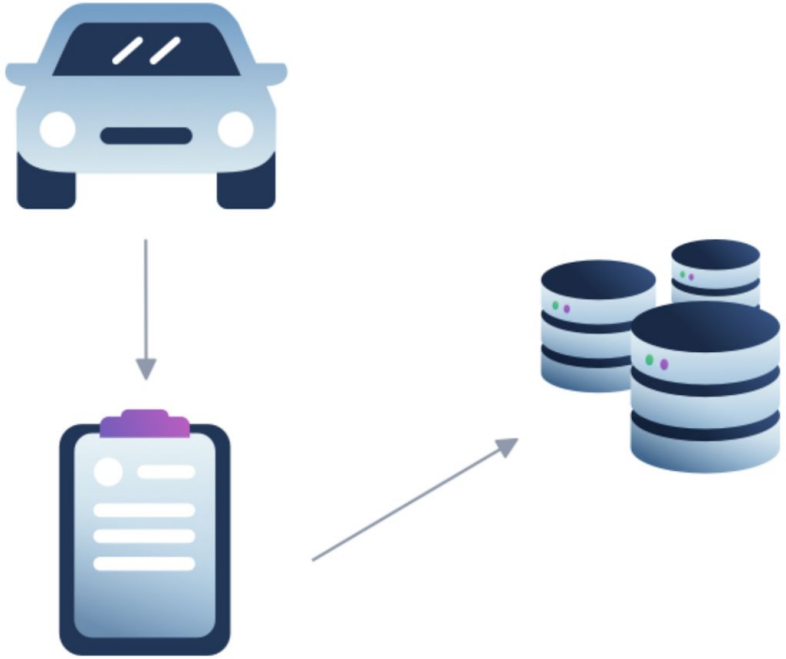
What can data do?

- Describe the current state of an organization or process
- Detect anomalous events
- Diagnose the causes of events and behaviors
- Predict future events

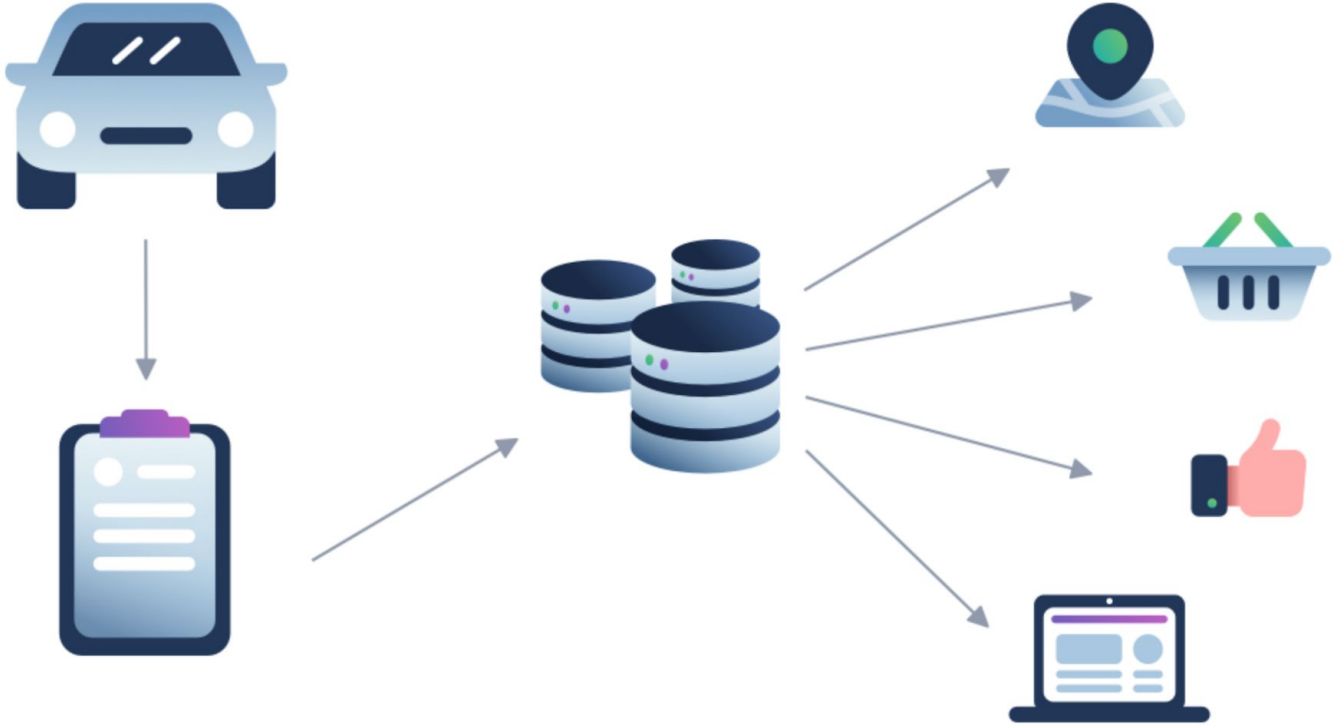
Why now?



Why now?



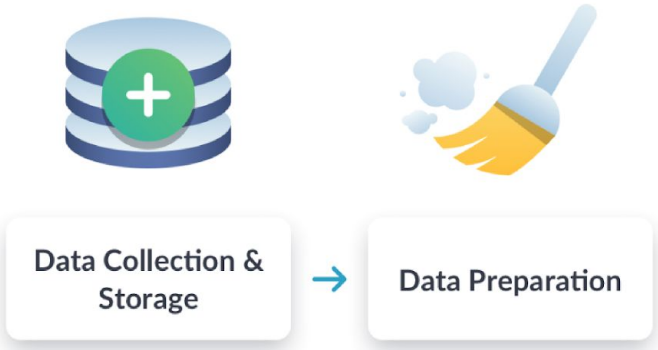
Why now?



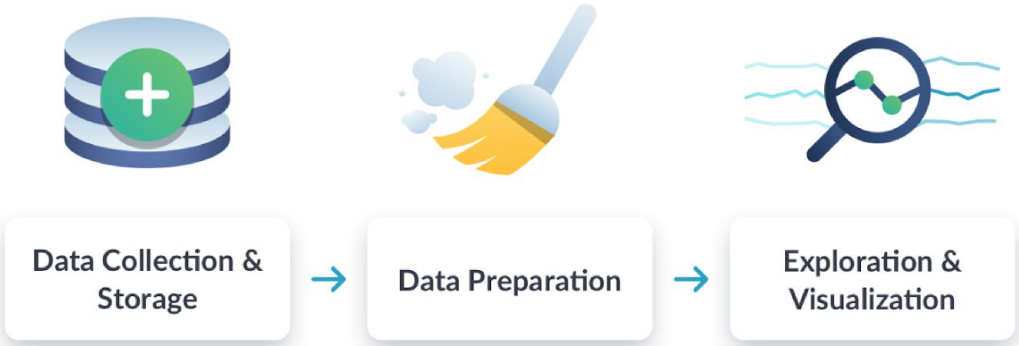
The data science workflow

Data Collection & Storage

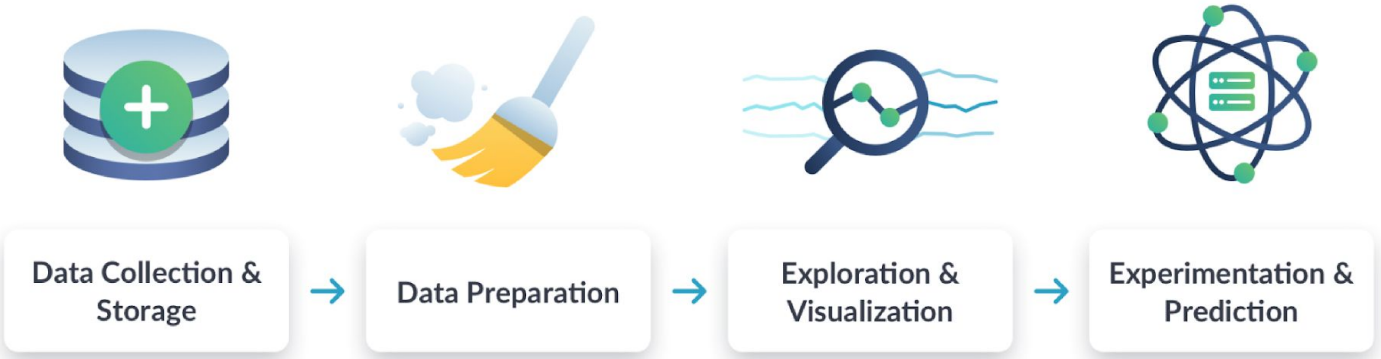
The data science workflow



The data science workflow



The data science workflow



Applications of data science

More case studies

- Traditional machine learning
- Internet of Things (IoT)
- Deep learning

Case study: fraud detection

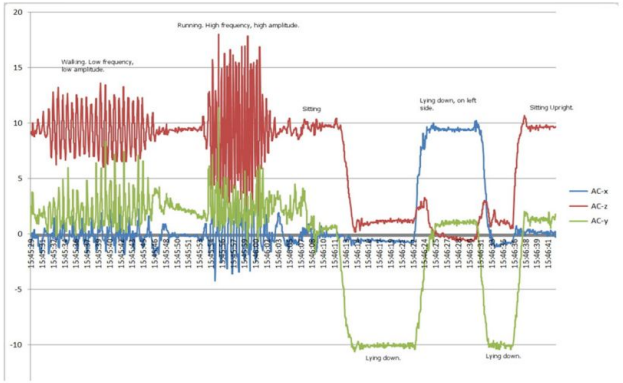
Case study: fraud detection

Amount	Date	Location	...
149.62	2019-05-23	London	...
2.69	2018-10-03	Birmingham	...
378.66	2019-06-15	Liverpool	...
123.5	2019-01-12	London	...
69.99	2018-06-16	Sao Paolo	...
3.67	2019-03-06	Brussels	...
...	...	...	...

What do we need for machine learning?

- A well-defined question
  - *"What is the probability that this transaction is fraudulent?"*
- A set of example data
  - *Old transactions labeled as "fraudulent" or "valid"*
- A new set of data to use our algorithm on
  - *New credit card transactions*

Case study: smart watch

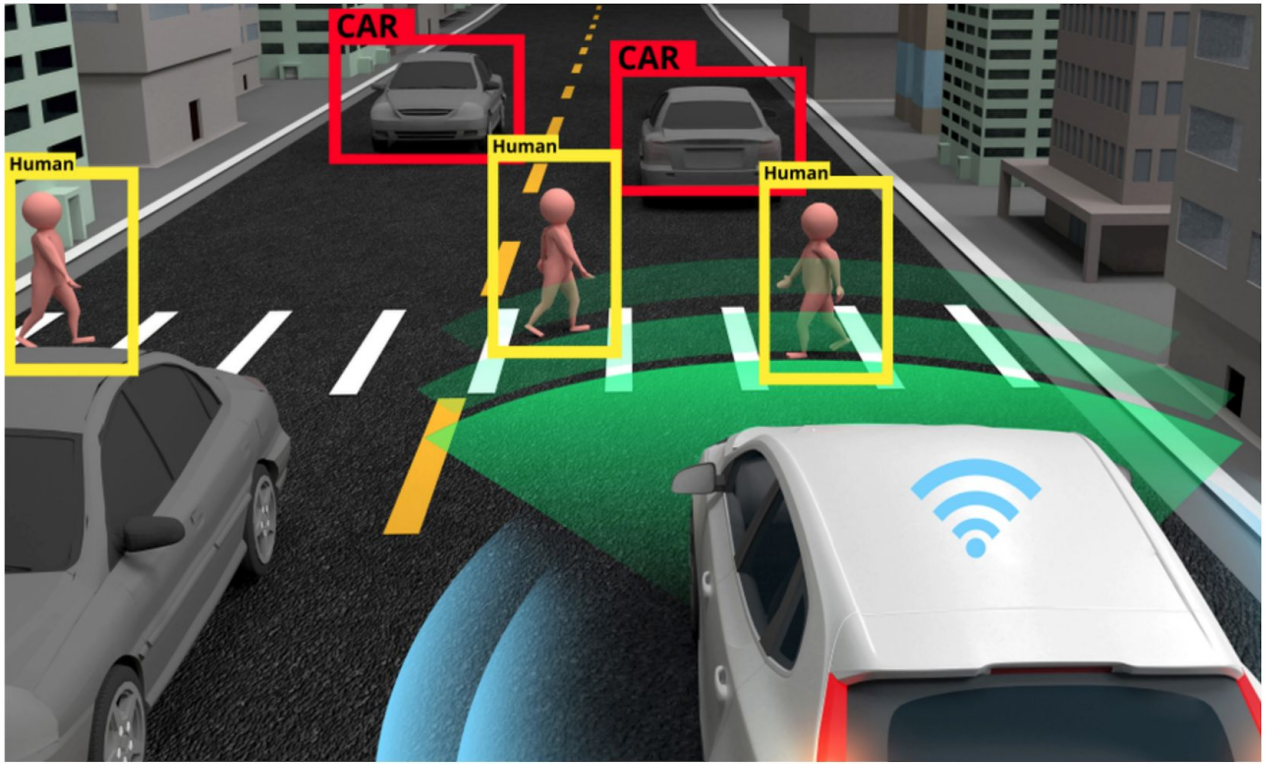


Internet of Things (IoT)

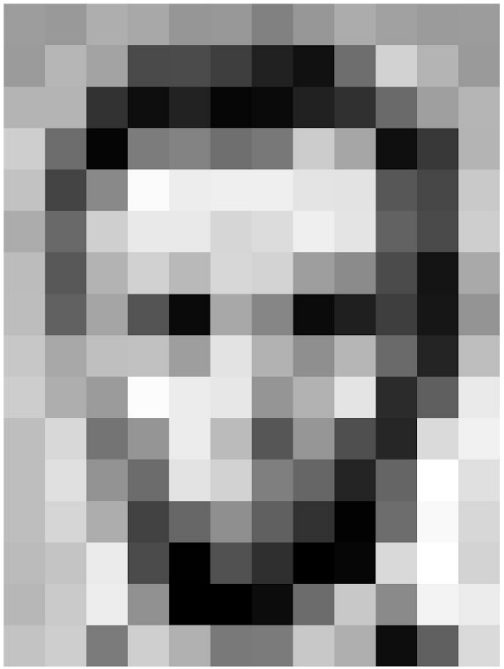
Refers to gadgets that aren't standard computers

- Smart watches
- Internet-connected home security systems
- Electronic toll collection systems
- Building energy management systems
- Much, much more!

Case study: image recognition



Case study: image recognition



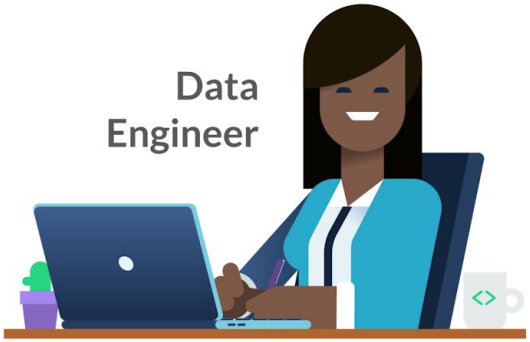
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Deep learning

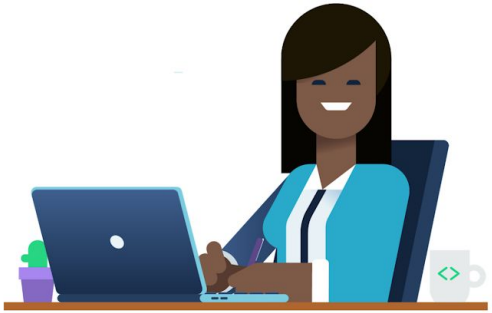
- Many neurons work together
- Requires much more training data
- Used in complex problems
  - Image classification
  - Language learning/understanding

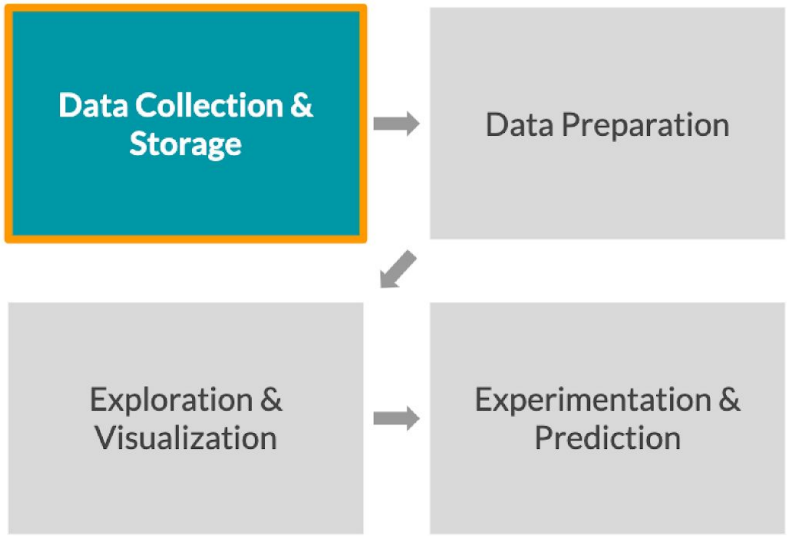
Data science roles and tools



Data engineer

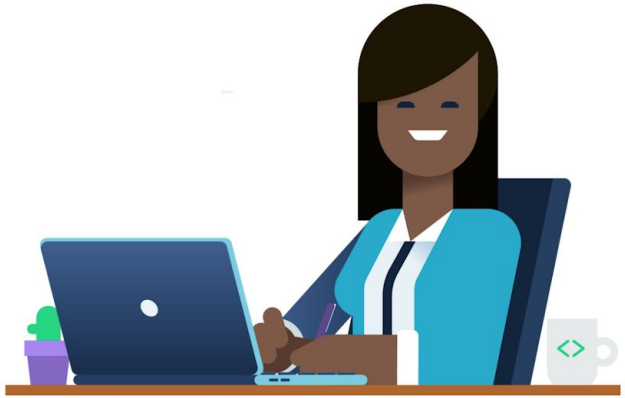
- Information architects
- Build data pipelines and storage solutions
- Maintain data access





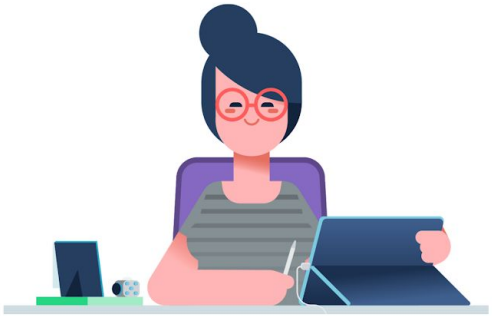
Data engineering tools

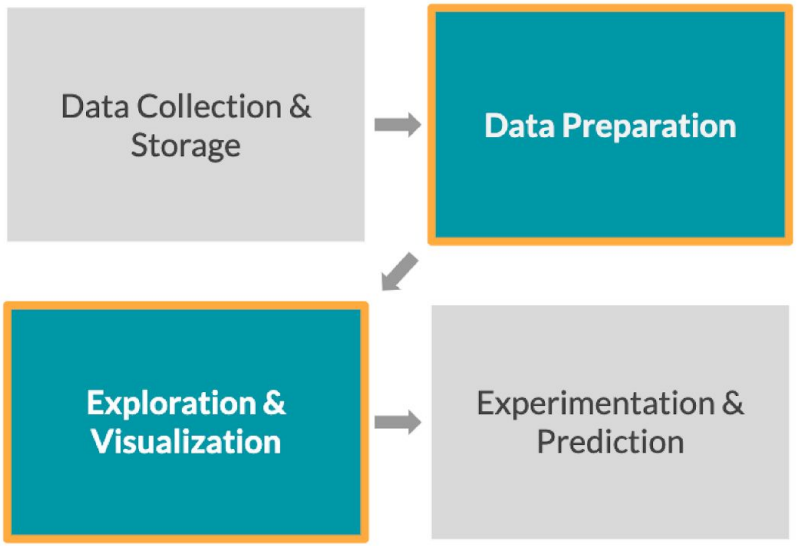
- **SQL**
  - To store and organize data
- **Java**, **Scala**, or **Python**
  - Programming languages to process data
- **Shell**
  - Command line to automate and run tasks
- **Cloud computing**
  - AWS, Azure, Google Cloud Platform



Data analyst

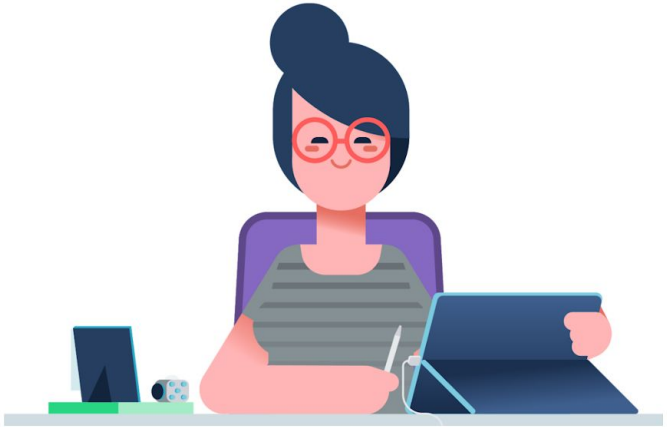
- Perform simpler analyses that describe data
- Create reports and dashboards to summarize data
- Clean data for analysis





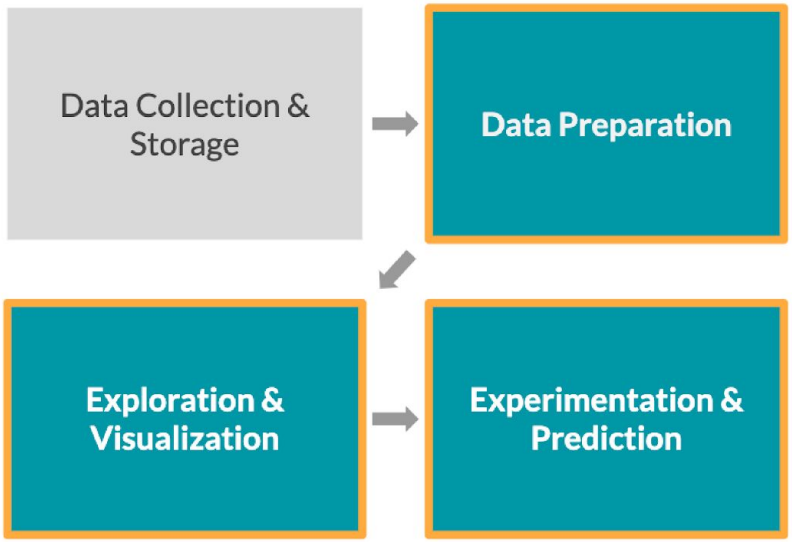
Data analyst tools

- **SQL**
  - Retrieve and aggregate data
- **Spreadsheets (Excel or Google Sheets)**
  - Simple analysis
- **BI tools (Tableau, Power BI, Looker)**
  - Dashboards and visualizations
- *May have:* Python or R
  - Clean and analyze data



Data scientist

- Versed in statistical methods
- Run experiments and analyses for insights
- Traditional machine learning

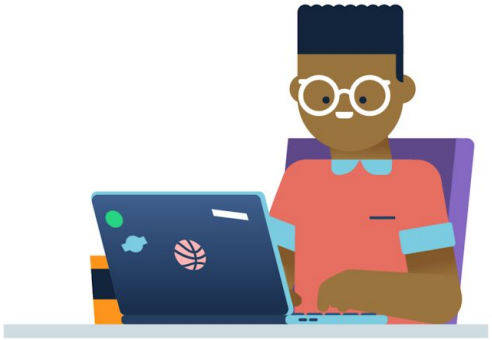


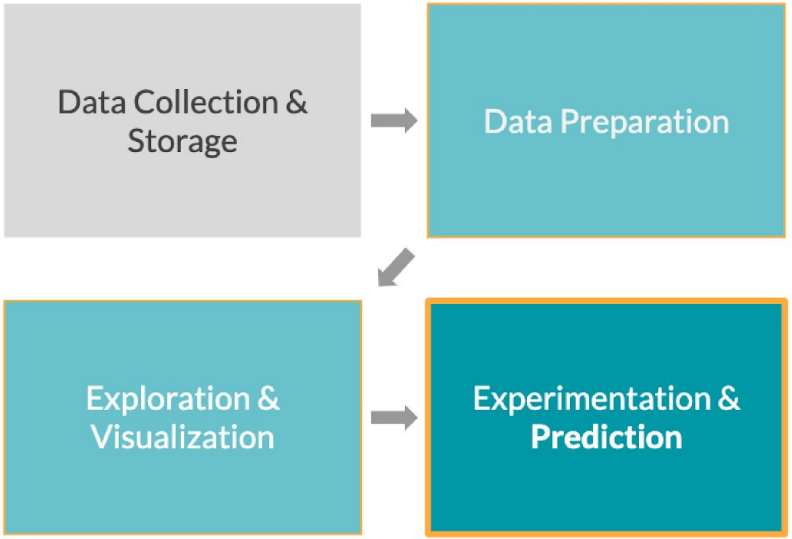
Data scientist tools

- **SQL**
  - Retrieve and aggregate data
- **Python and/or R**
  - Data science libraries, e.g., `pandas` (Python) and `tidyverse` (R)

Machine learning scientist

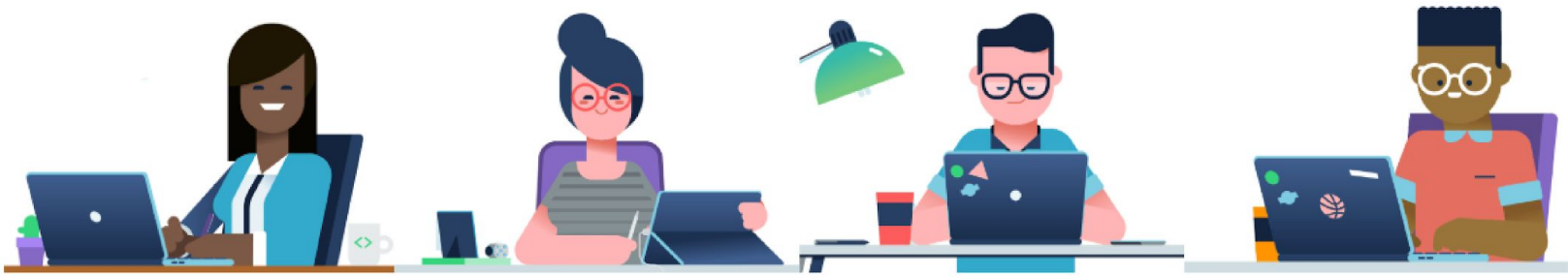
- Predictions and extrapolations
- Classification
- Deep learning
  - Image processing
  - Natural language processing





Machine learning tools

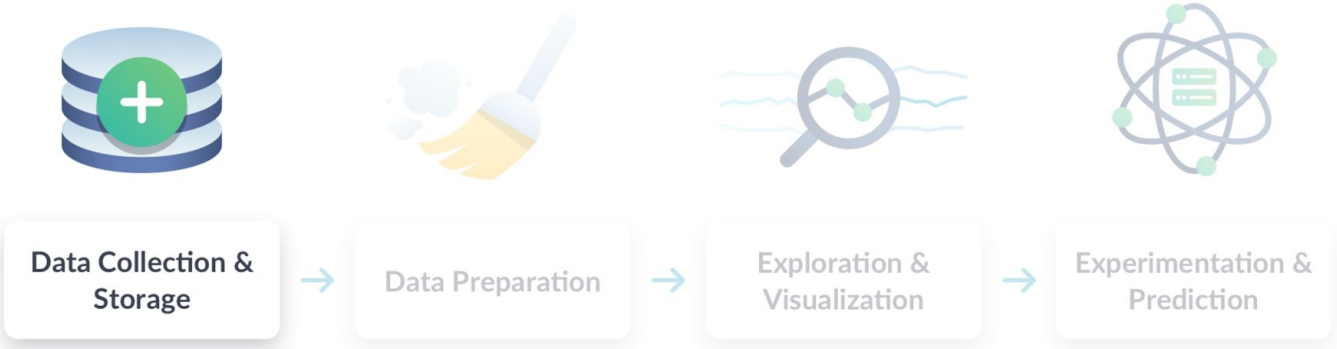
- **Python and/or R**
  - Machine learning libraries, e.g., TensorFlow or Spark



Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

Data sources

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

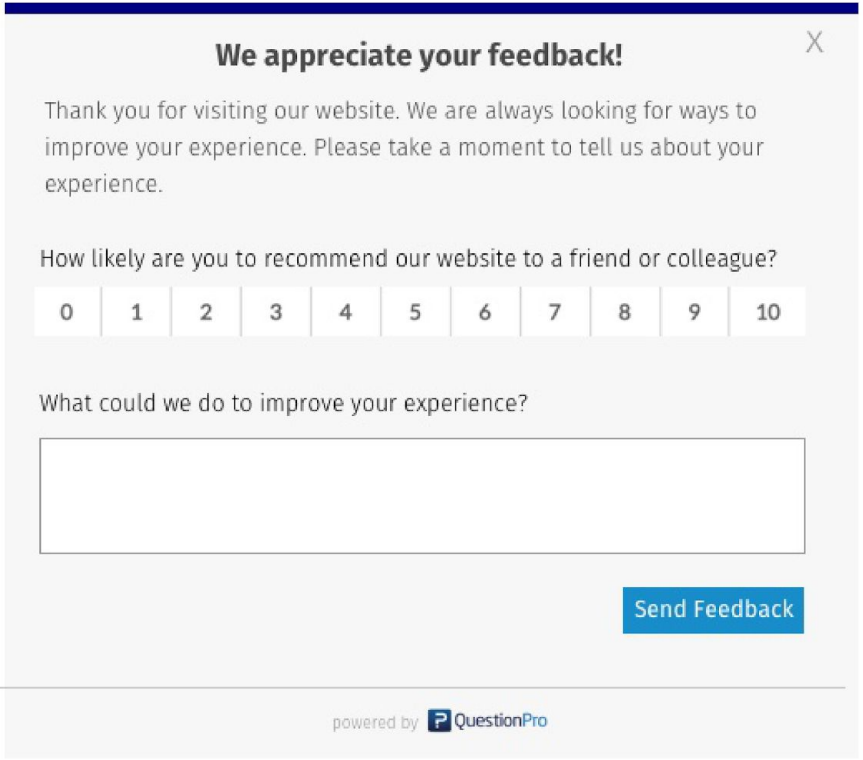
event_name	timestamp	user_id
homepage_visit	2019-01-01 12:01:01	1234

- Events
- Timestamps
- User information

Survey data

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

Net Promotor Score



Open data

- Data APIs
- Public records

Public data APIs

- **A**pplication **P**rogramming **I**nterface
- 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

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**
- For the EU, **data.europa.eu**

Data types

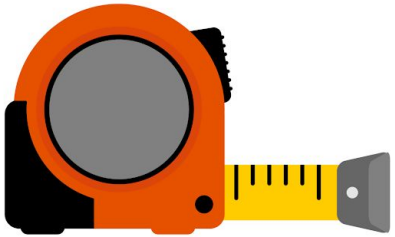
Why care about data types?

- 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

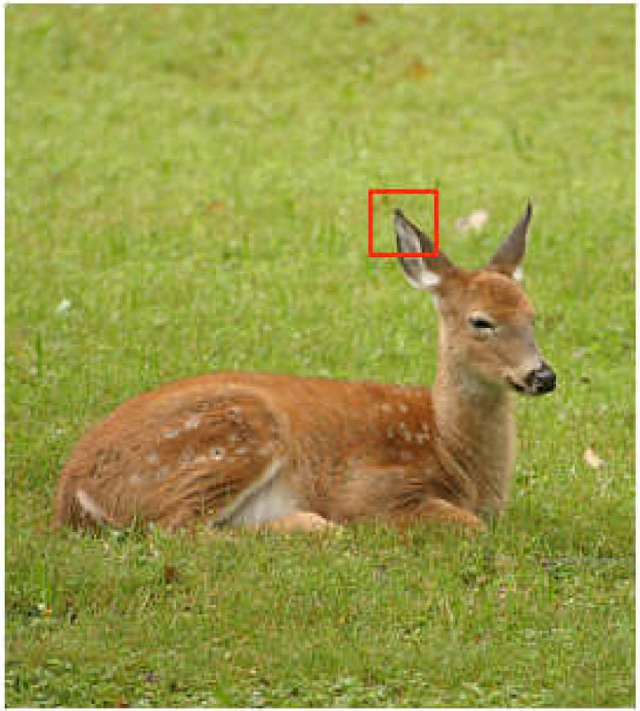
Qualitative data

- 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

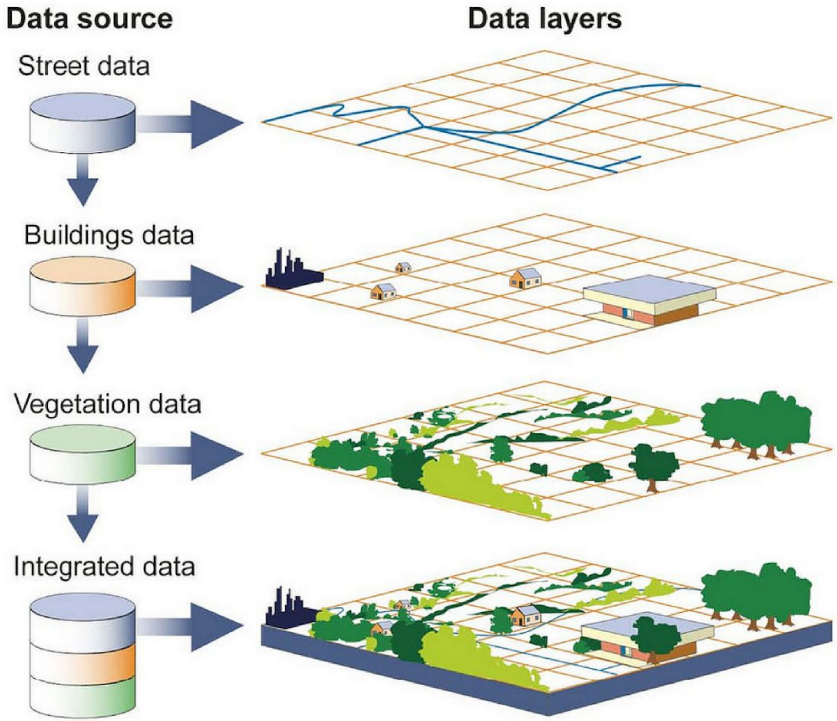
“Great evening, extremely good value”

 Review of L'Ange 20 Restaurant

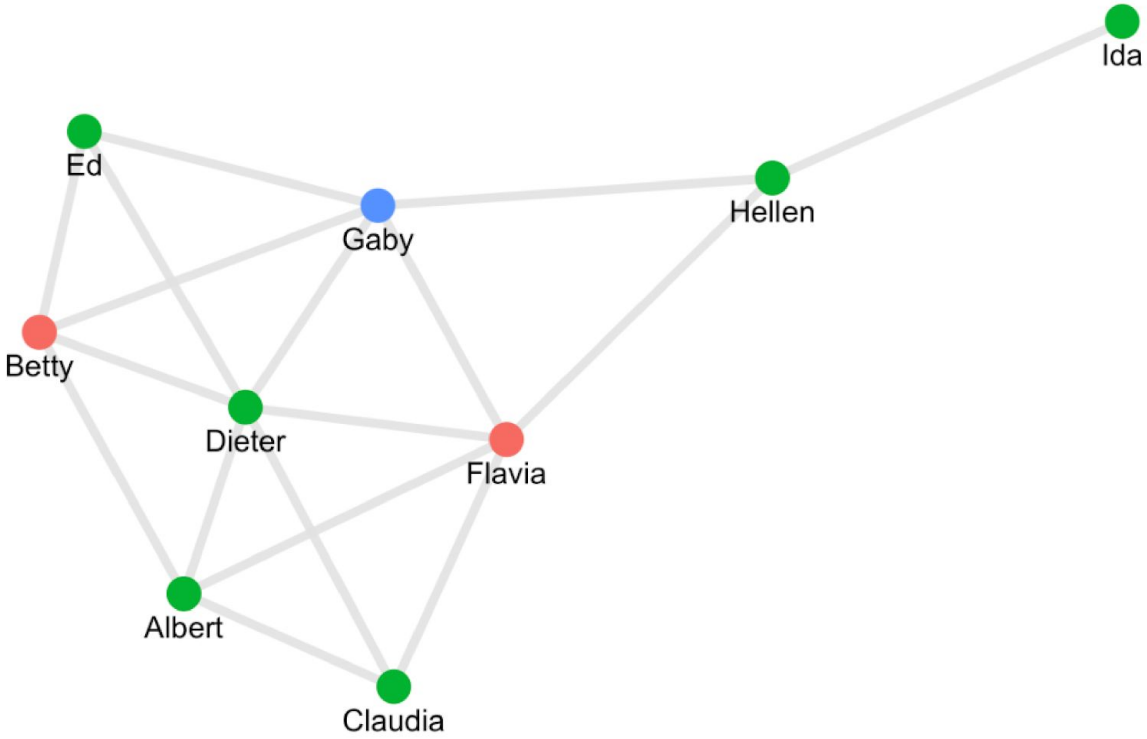
I went to this place with my boyfriend for a special occasion and we were not disappointed. We were greeted warmly by Christopher who guided us through the menu and wine. The food was delicious and I only wish that we could have had room for three courses. The value was excellent compared to other prices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

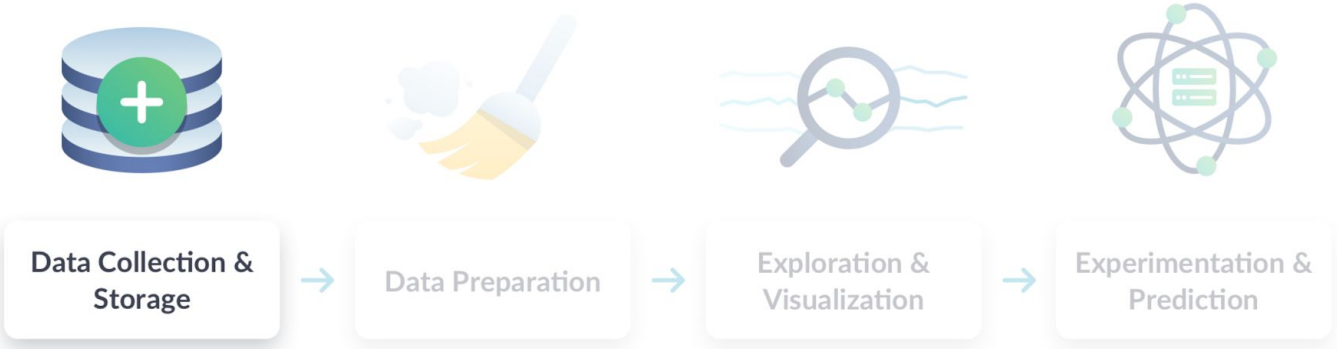


Recap

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

Data storage and retrieval

The data science workflow

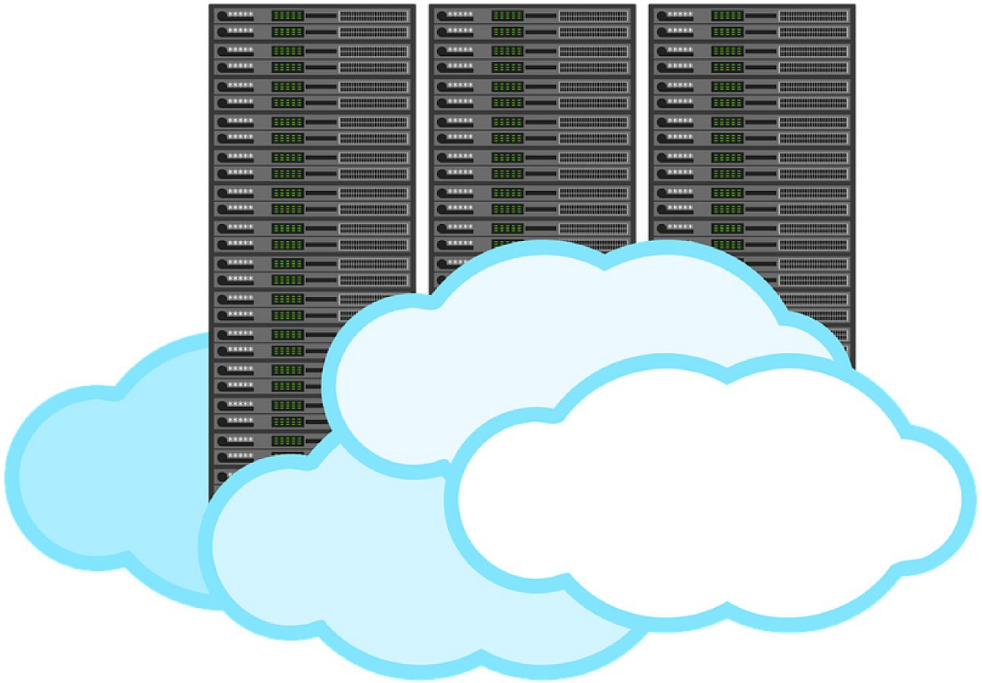


Things to consider when storing data

- 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

Document Database

Tabular

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

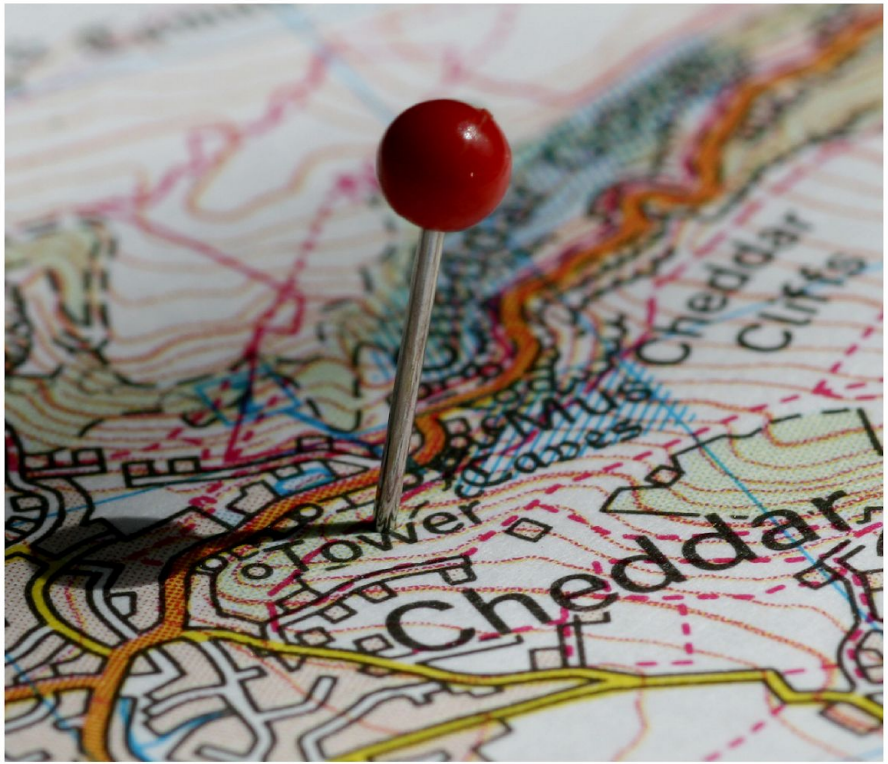
Relational Database

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

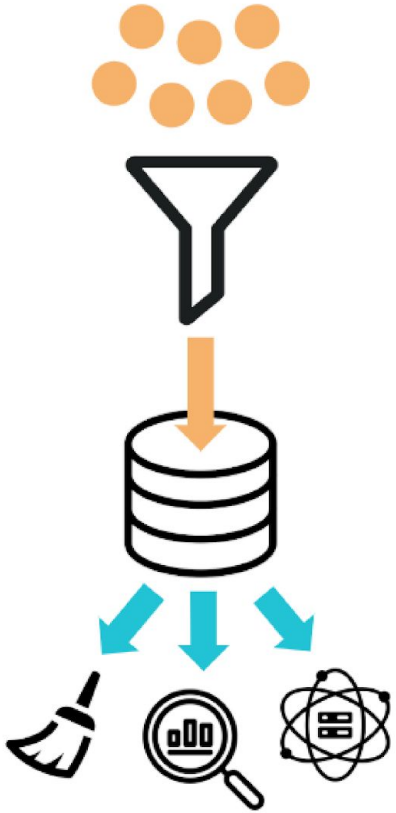
Data Type	Storage Solution
Unstructured	Document Database
Tabular	Relational Database

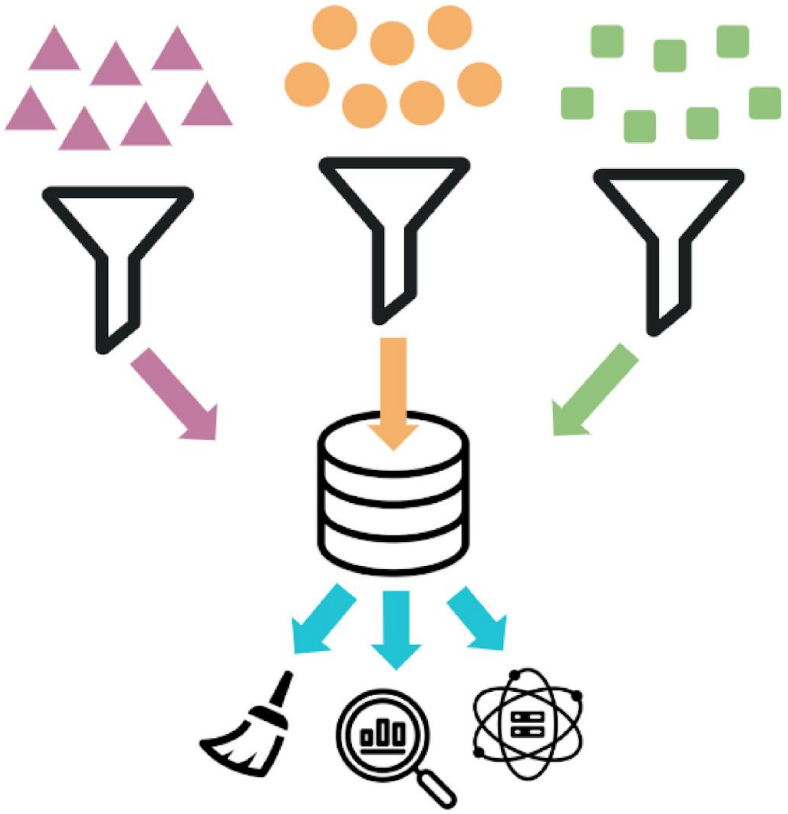
Putting it all together: Queries

Putting it all together: Queries

Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

Data Pipelines





How do we scale?

More than one data source:

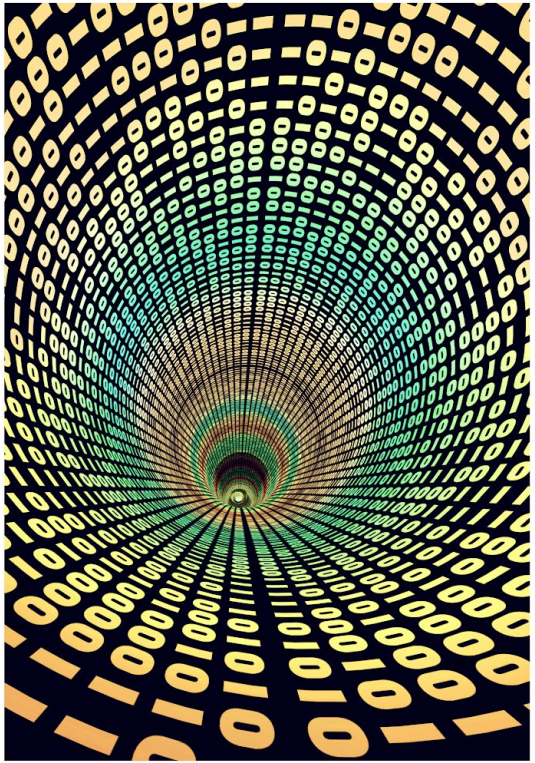
- Public records
- APIs
- Databases

Different data types:

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

What is a data pipeline?

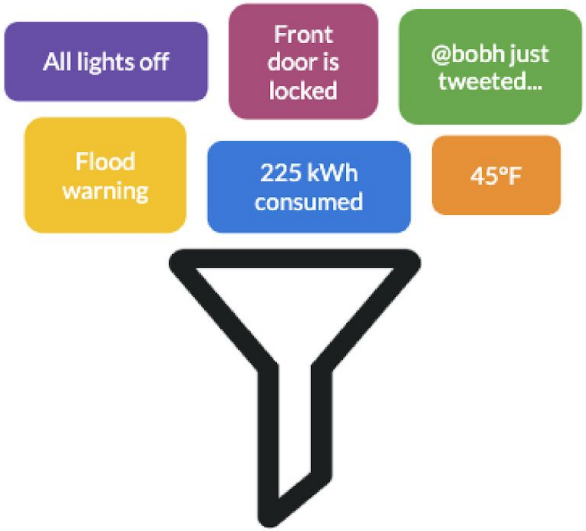
- Moves data into defined stages
- Automated collection and storage
  - *Scheduled hourly, daily, weekly, etc*
  - *Triggered by an event*
- Monitored with generated alerts
- Necessary for big data projects
- Data engineers work to customize solutions
- **E**xtract **T**ransform **L**oad (ETL)



Case study: smart home

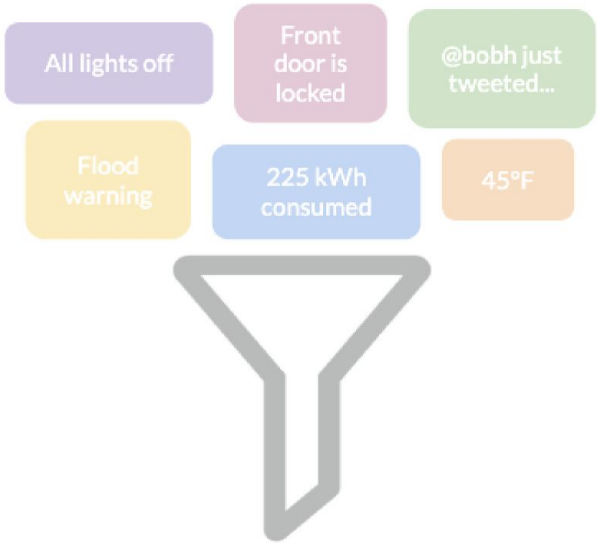
Data	Source	Frequency
Weather conditions	National Weather Service API	Every 30 minutes
Tweets in your area	Twitter API	Real-time stream
Indoor temperature	Smart home thermostat	Every 5 minutes
Status of lights	Smart light bulbs	Every minute
Status of locks	Smart door locks	Every 15 seconds
Energy consumption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

Extract

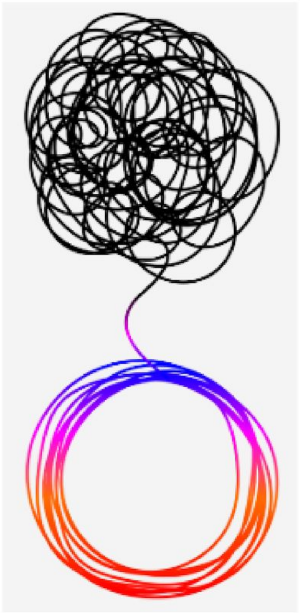


Source	Frequency
National Weather API	Every 30 minutes
Twitter API	Real-time stream
Smart home thermostat	Every 5 minutes
Smart light bulbs	Every minute
Smart door locks	Every 15 seconds
Smart meter	Weekly

Transform



Extract



Transform

Transform

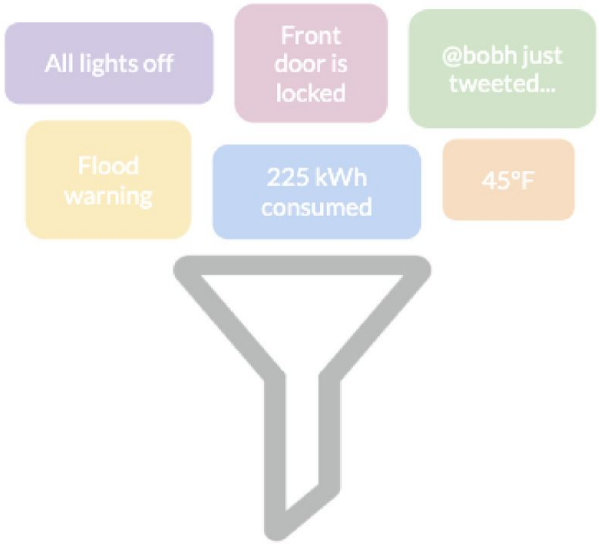
With all the data coming in, how do we keep it organized and easy to use?

Example transformations:

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

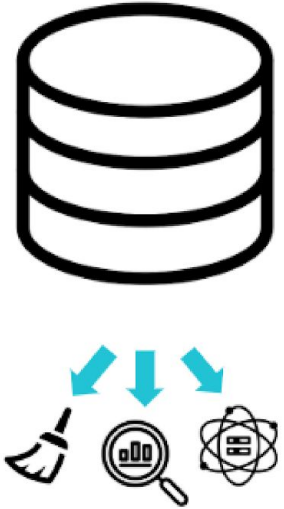
Data preparation and exploration does not occur at this stage

Load



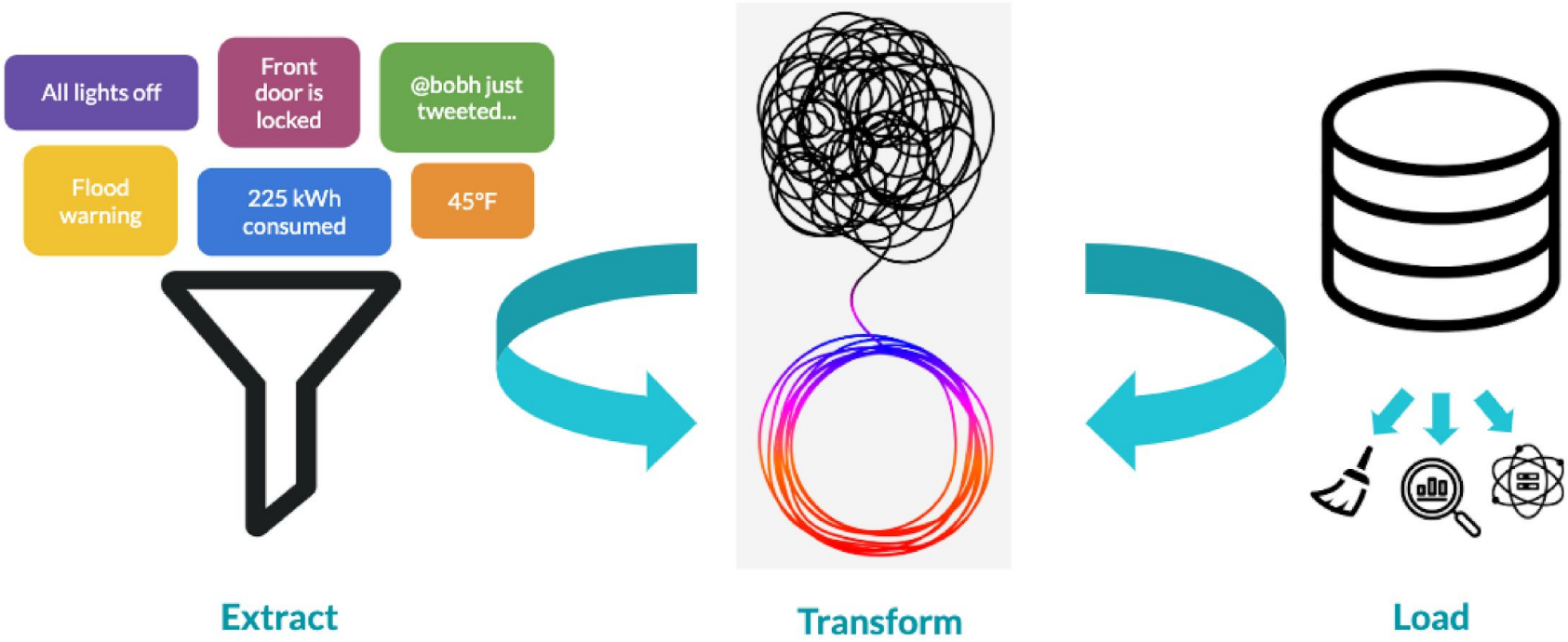
Extract

Transform



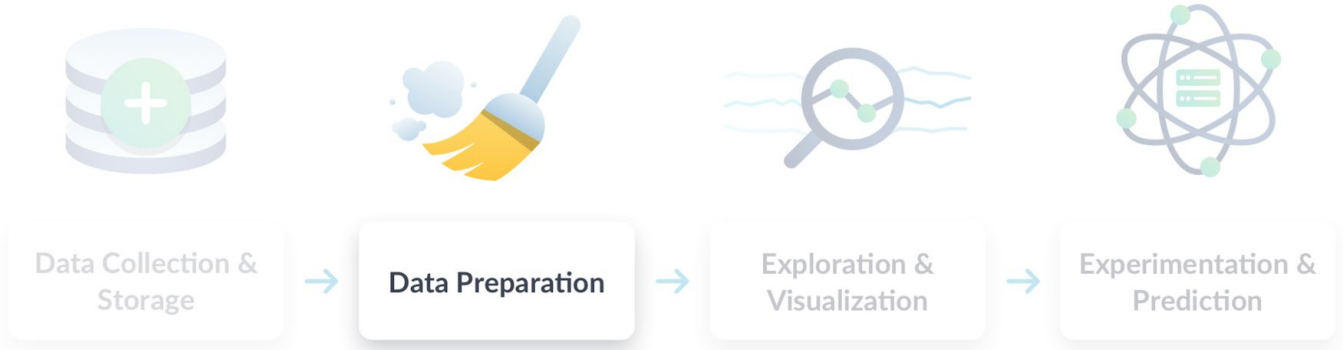
Load

Automation



Data preparation

Data workflow



Why prepare data?

- Real-life data is messy
- Preparation is done to prevent:
  - errors
  - incorrect results
  - biasing algorithms

Let's start cleaning

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

Tidy data

Before

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

Tidy data output

Before

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

After

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Remove duplicates

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Remove duplicates | output

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

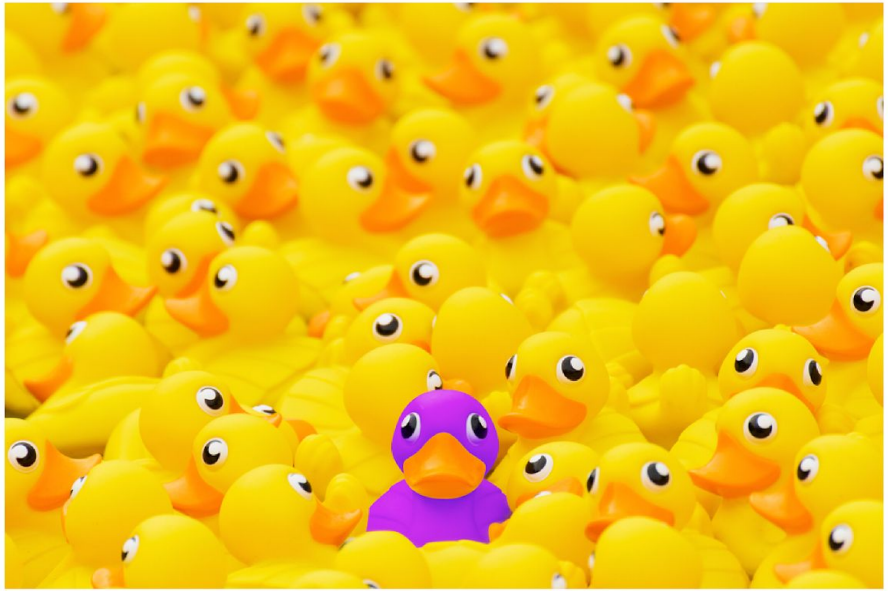
After

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

Unique ID

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"



Unique ID | output

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

Homogeneity

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

Homogeneity | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Homogeneity, again

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Homogeneity, again | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Data types

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Data types | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"USA"
2	Hadrien		1.80	"FR"

Missing values

Before

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"USA"
2	Hadrien		1.80	"FR"

Reasons:

- data entry
- error
- valid missing value

Solutions:

- impute
- drop
- keep

Missing values | output

Before

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"USA"
2	Hadrien		1.80	"FR"

After

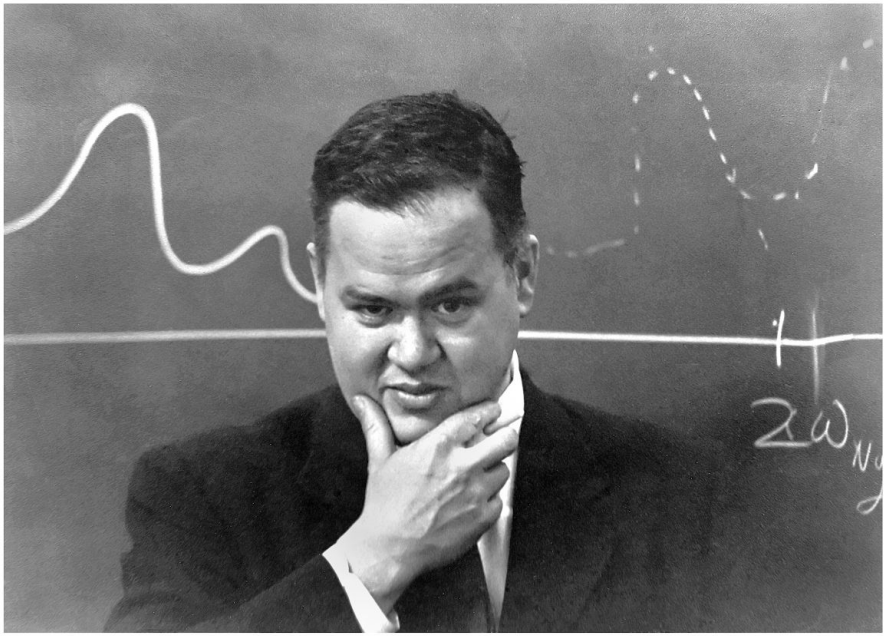
ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"USA"
2	Hadrien	28	1.80	"FR"

Exploratory Data Analysis

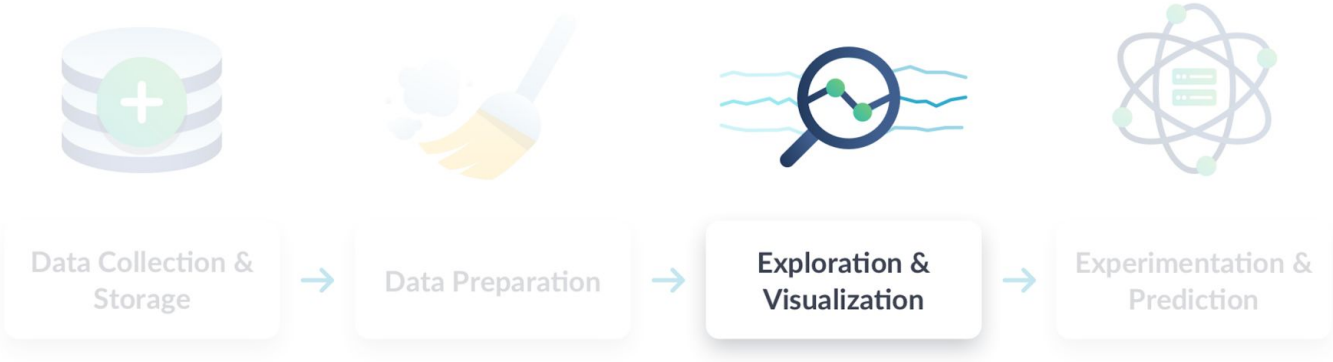
What is EDA?

Exploratory Data Analysis:

- Exploring the data
- Formulating hypotheses
- Assessing characteristics
- Visualizing



Data workflow



Let's dive right in

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

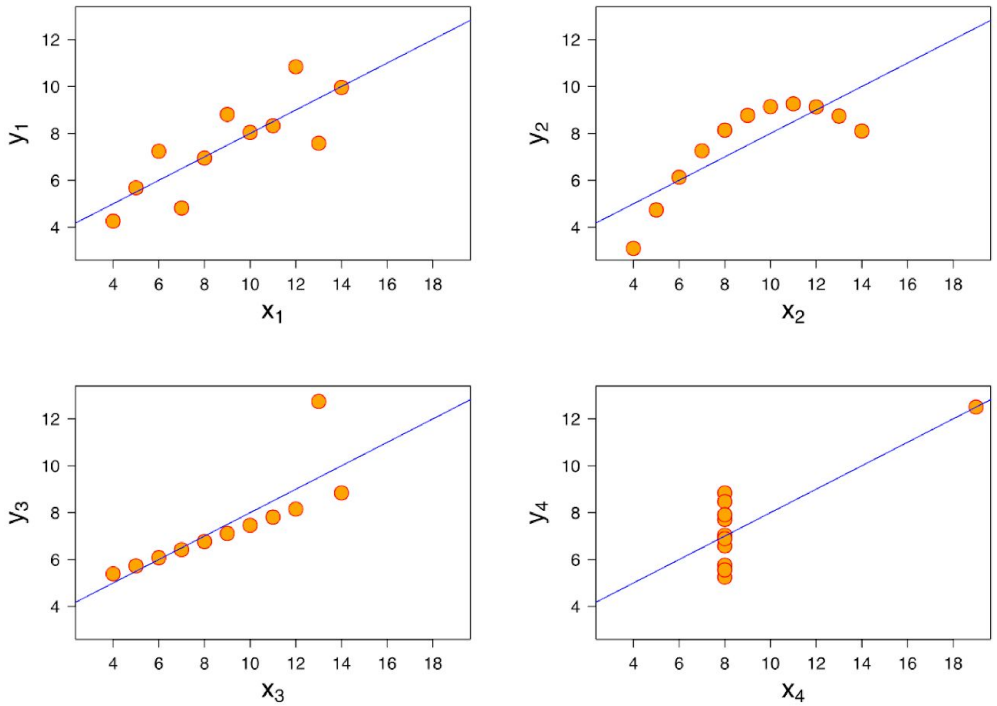
Surprise!

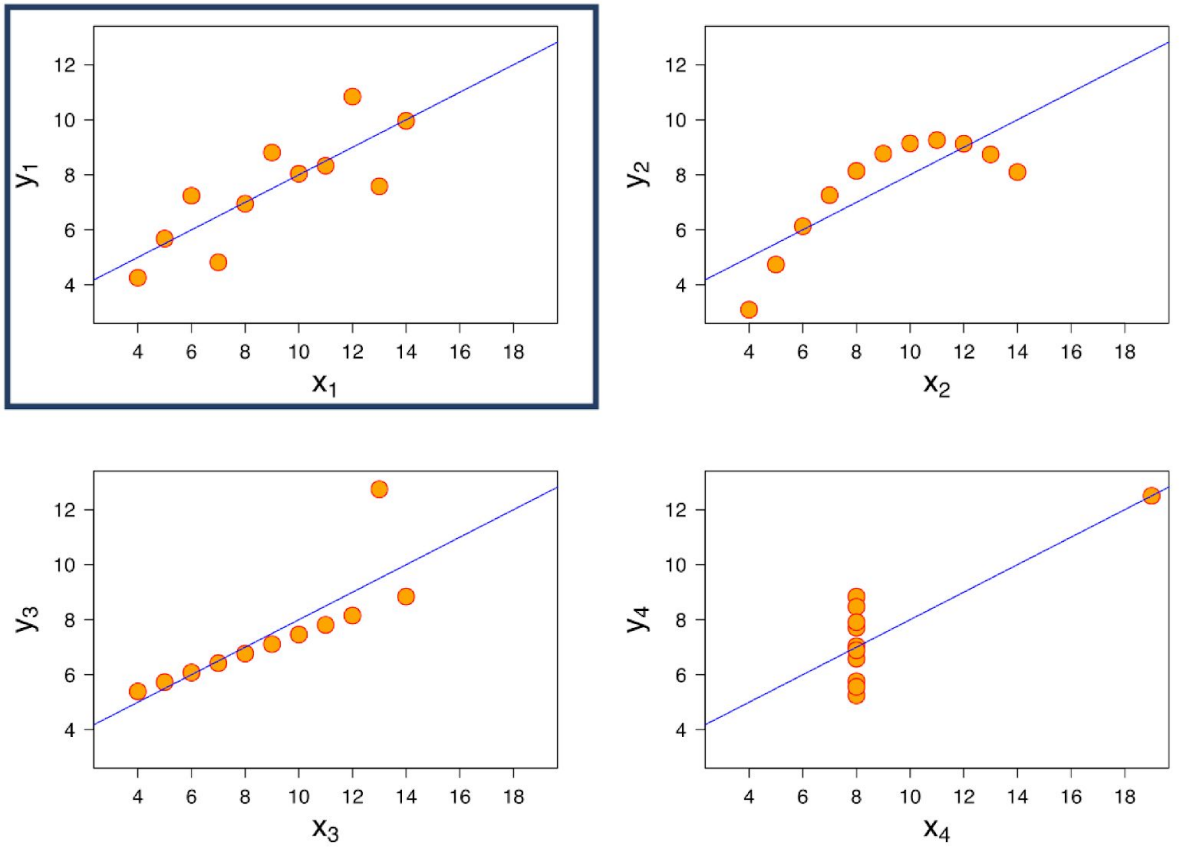
All four datasets display:

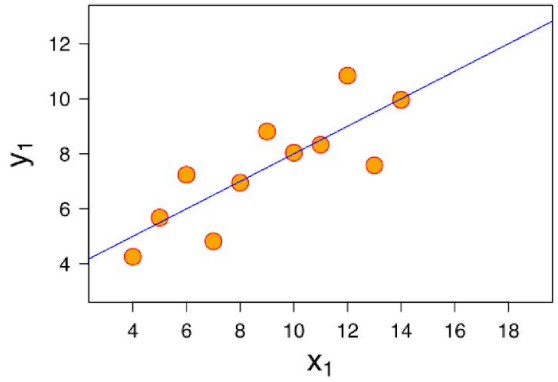
- identical mean and variance for x
- identical mean and variance for y
- identical correlation coefficient
- identical linear regression equation

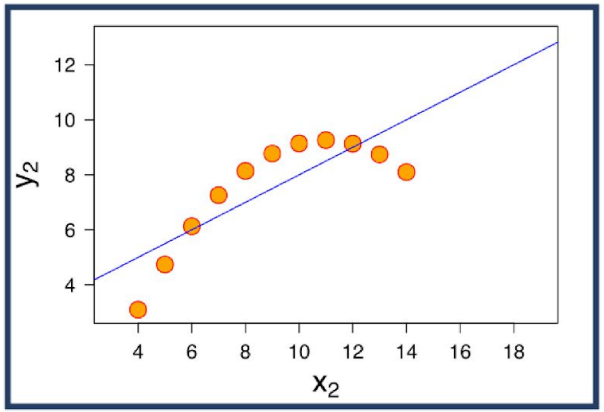
In short: **they look quite similar**

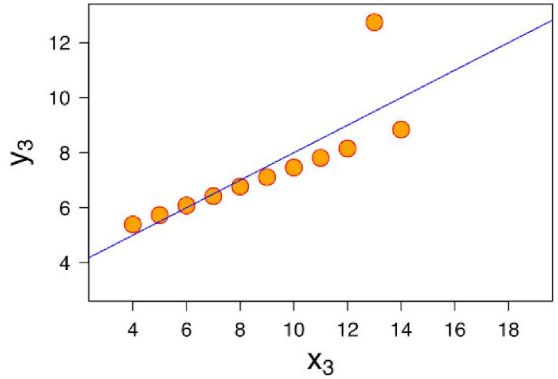
Anscombe's quartet

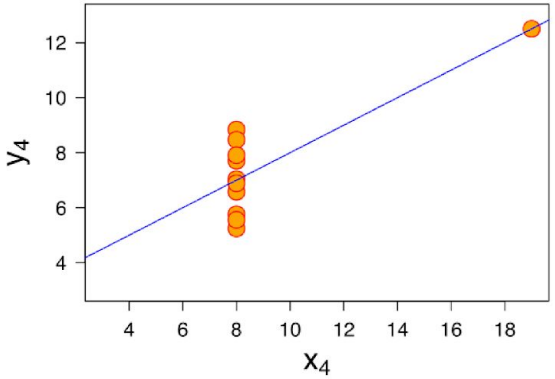


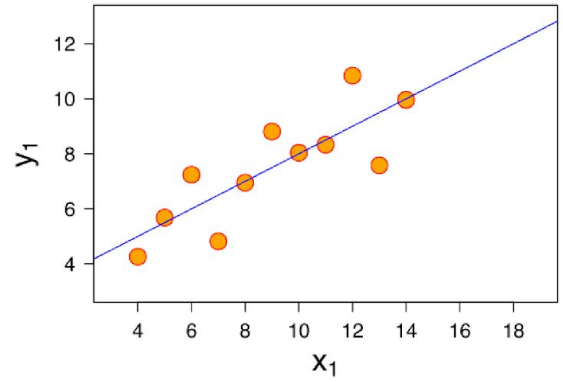


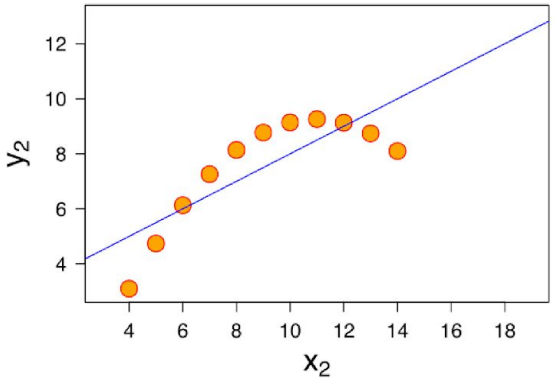


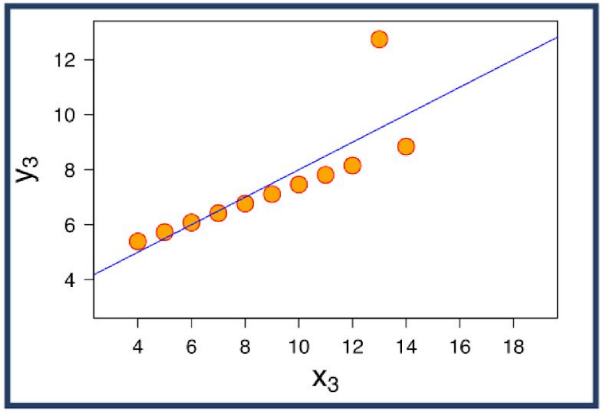


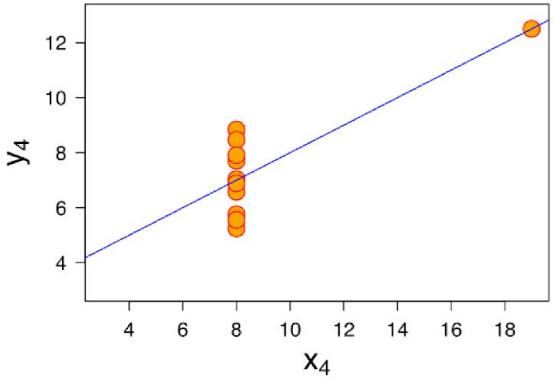


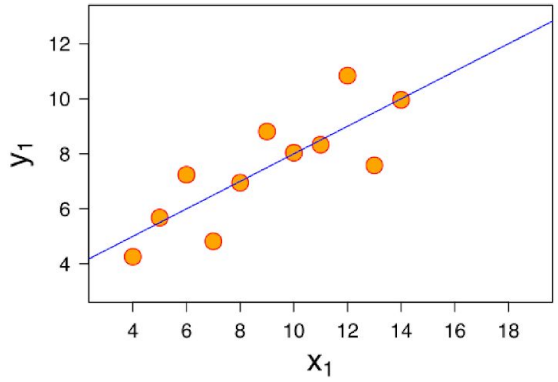


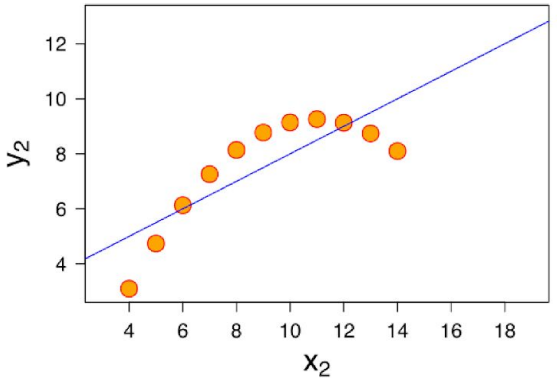


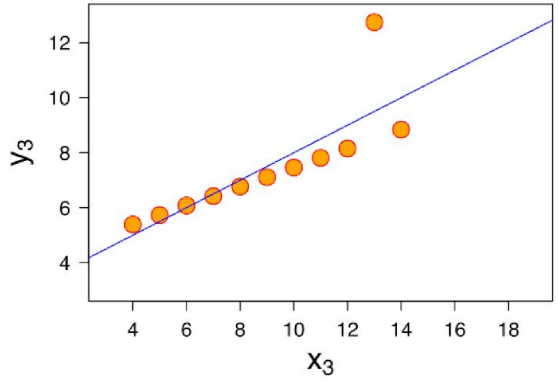


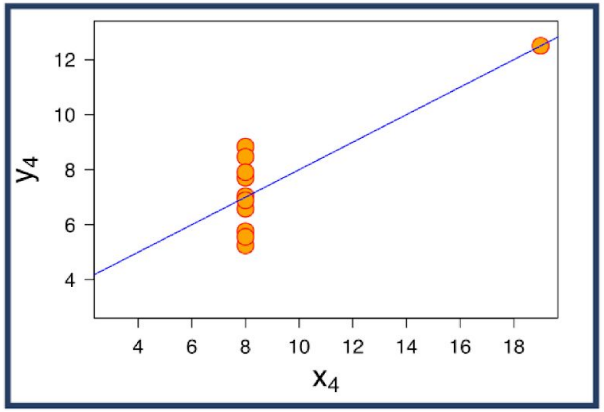












Knowing your data

- Flight Number (number)
- Date (datetime)
- Time (UTC) (datetime)
- Booster Version (text)
- Launch Site (text)
- Payload (text)
- Payload Mass (kg) (number)
- Orbit (text)
- Customer (text)
- Mission Outcome (text)
- Landing Outcome (text)

Previewing your data

Flight	Date	Time (UTC)	Booster Version	Launch Site	Payload
1	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Uni
2	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats
3	2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2+
4	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
5	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

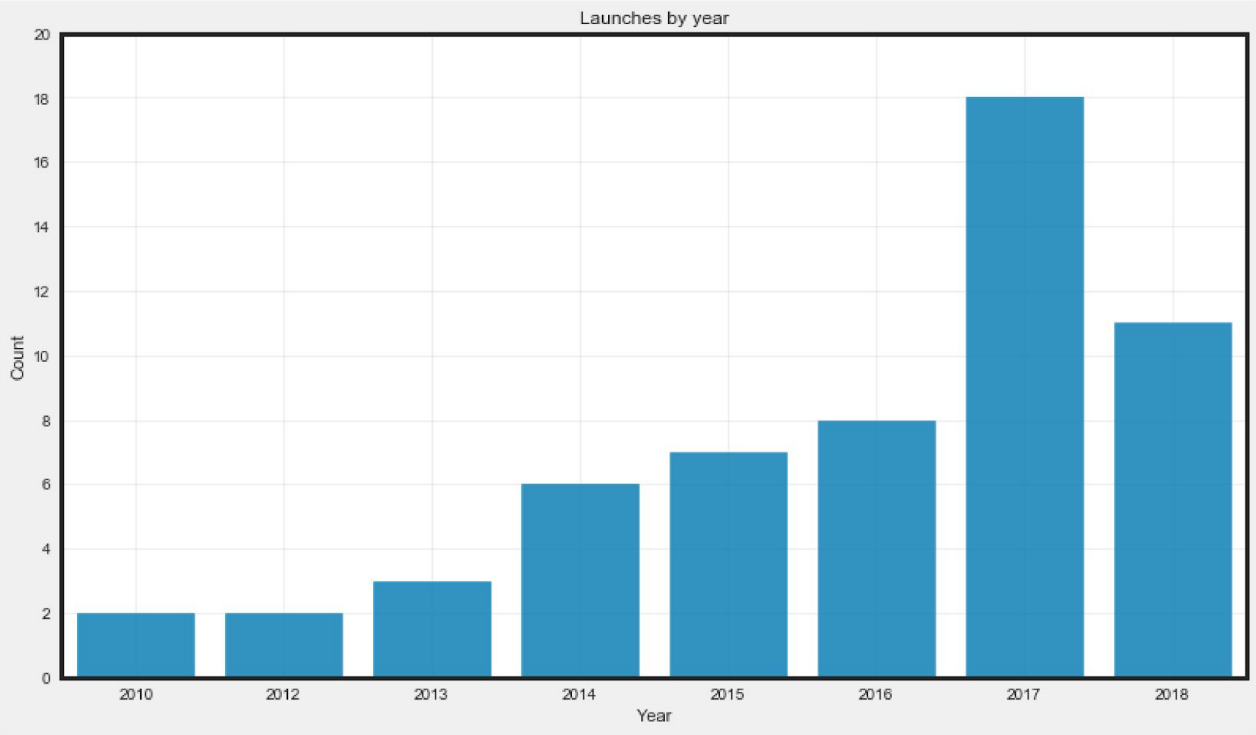
Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
NaN	LEO	SpaceX	Success	Failure (parachute)
NaN	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
525	LEO (ISS)	NASA (COTS)	Success	No attempt
500	LEO (ISS)	NASA (CRS)	Success	No attempt
677	LEO (ISS)	NASA (CRS)	Success	No attempt

Descriptive statistics

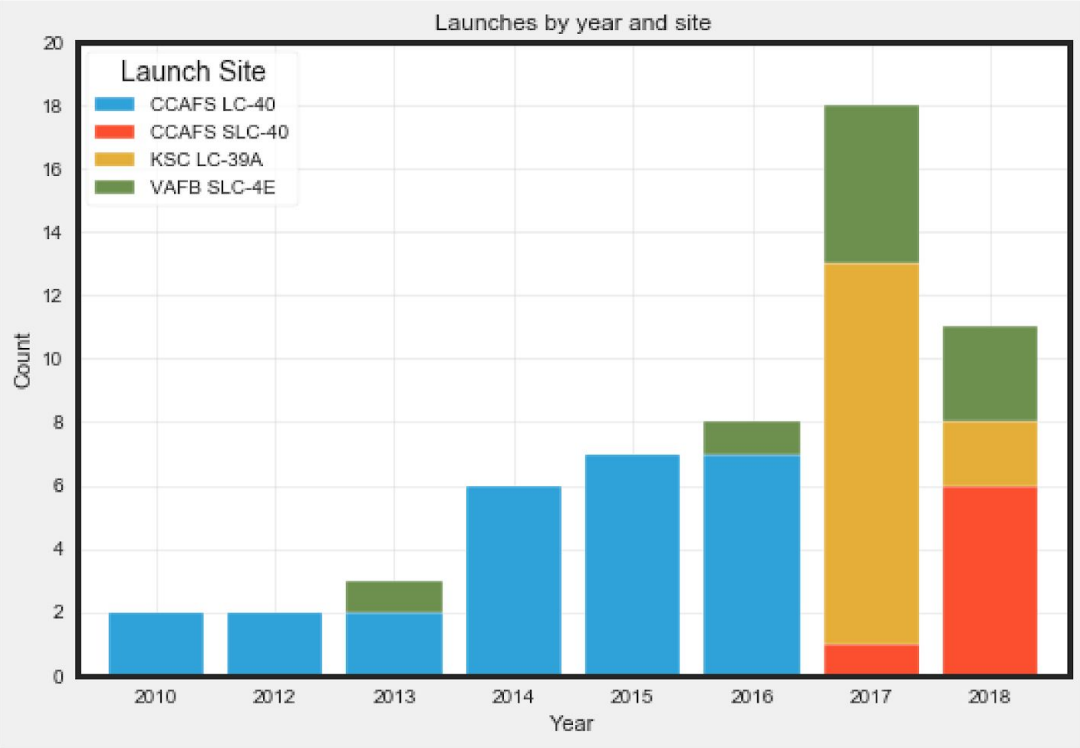
	Flight	Date	Time (UTC)	Booster Version	Launch Site	Payload
count	55	55	55	55	55	55
unique	55	55	53	51	4	55
top	6	2018-03-30	4:45:00	F9 v1.1	CCAFS LC-40	SES-9
freq	1	1	2	5	26	1

	Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
count	53	55	55	55	55
unique	47	8	28	2	12
top	9,600	GTO	NASA (CRS)	Success	No attempt
freq	5	22	14	54	18

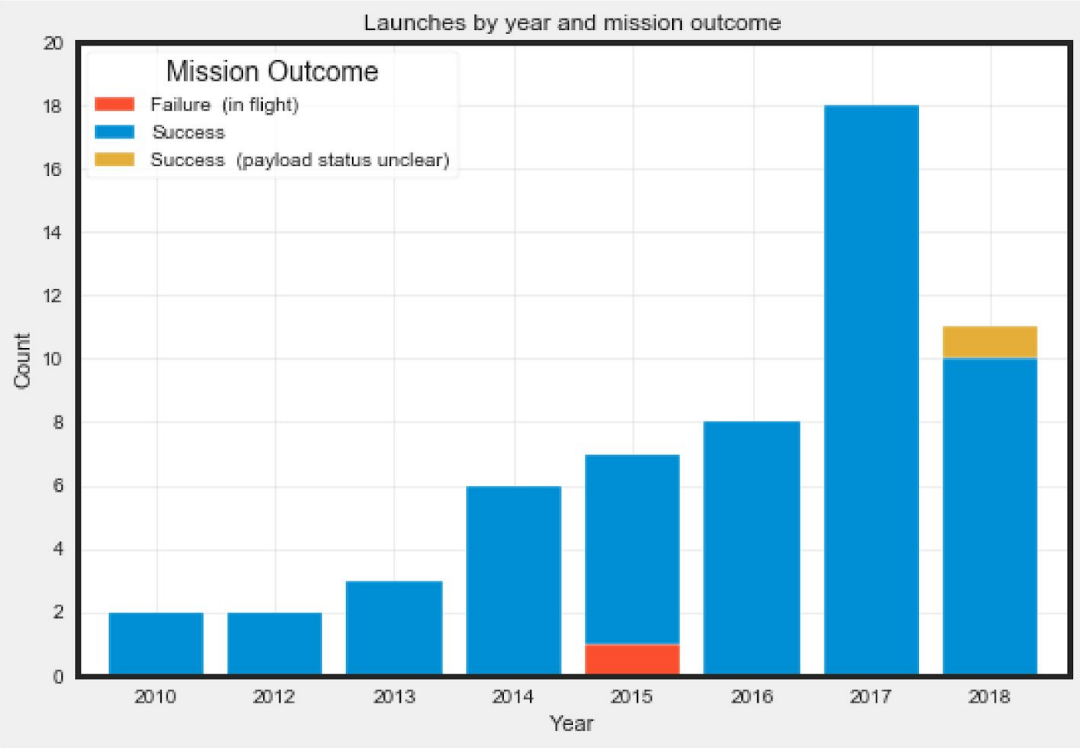
Visualize!



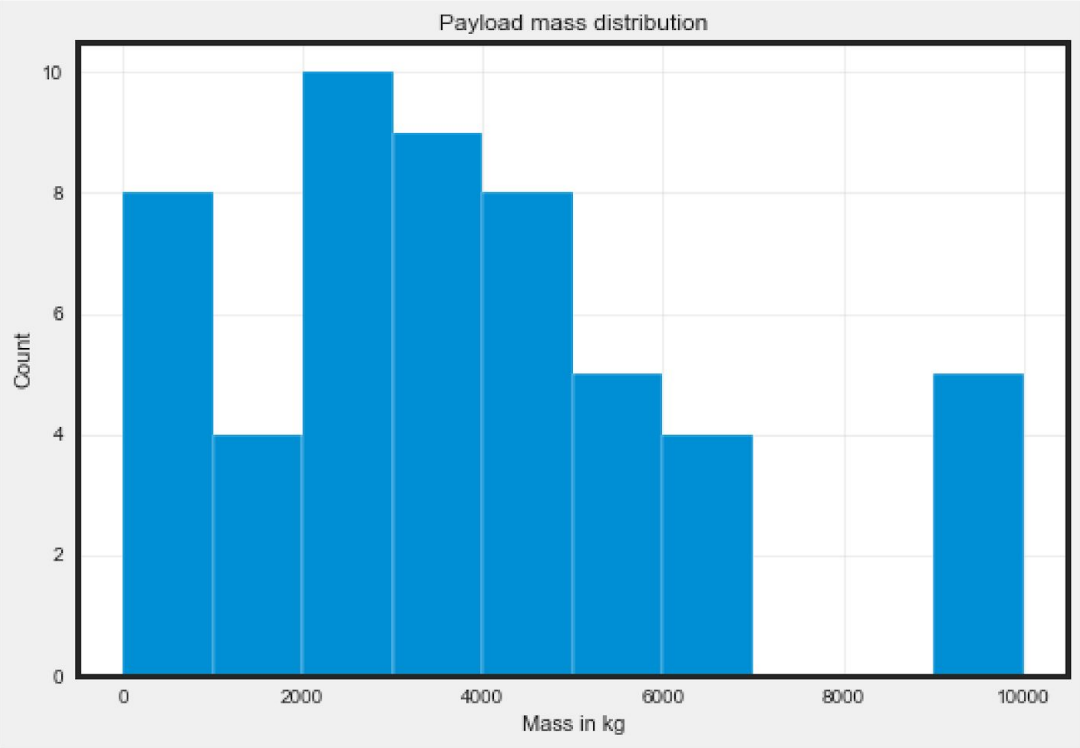
Ask more questions!



Ask more questions!



Outliers

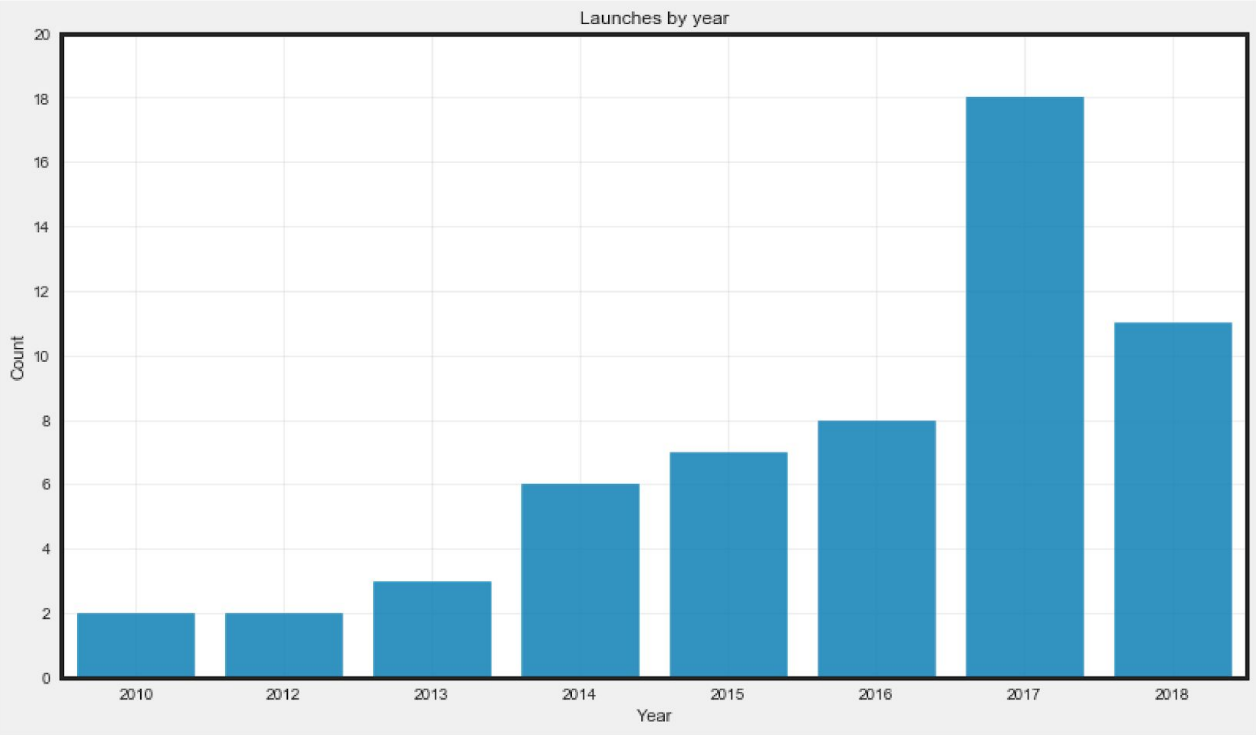


Interactive dashboards

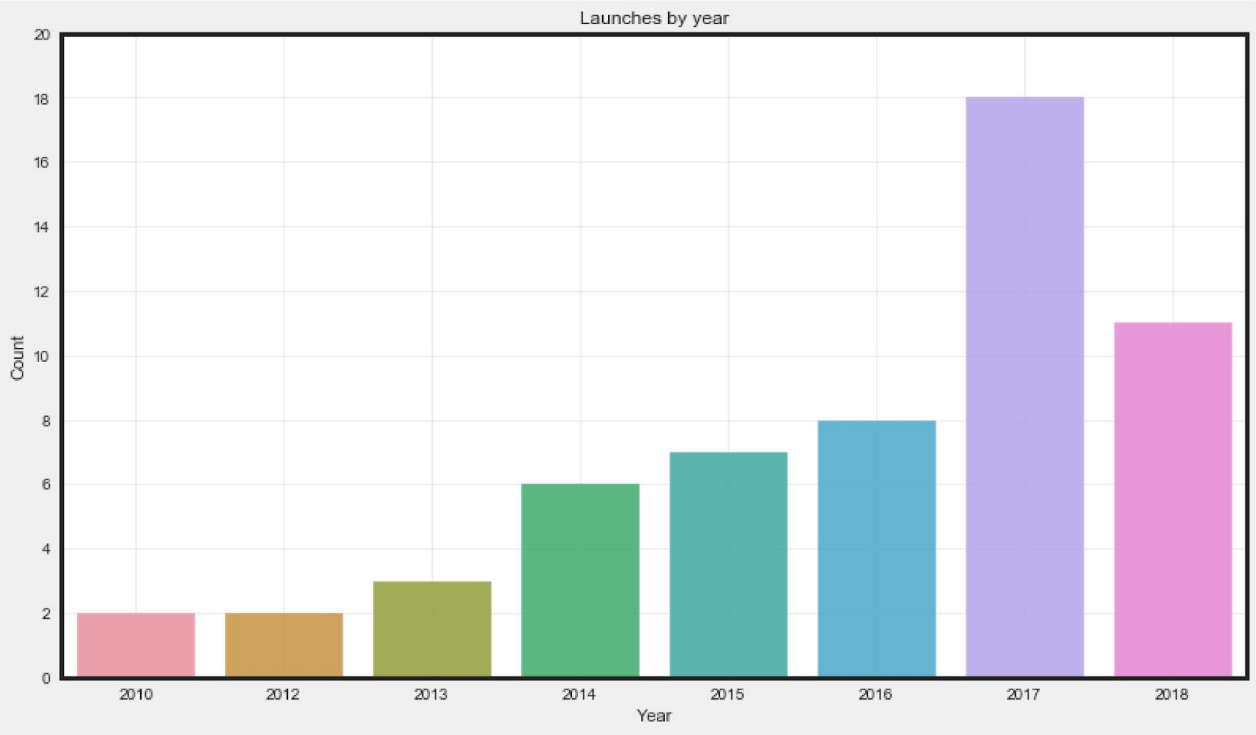
One picture...

> One picture is worth a thousand words...
>
> ...if the picture makes sense.

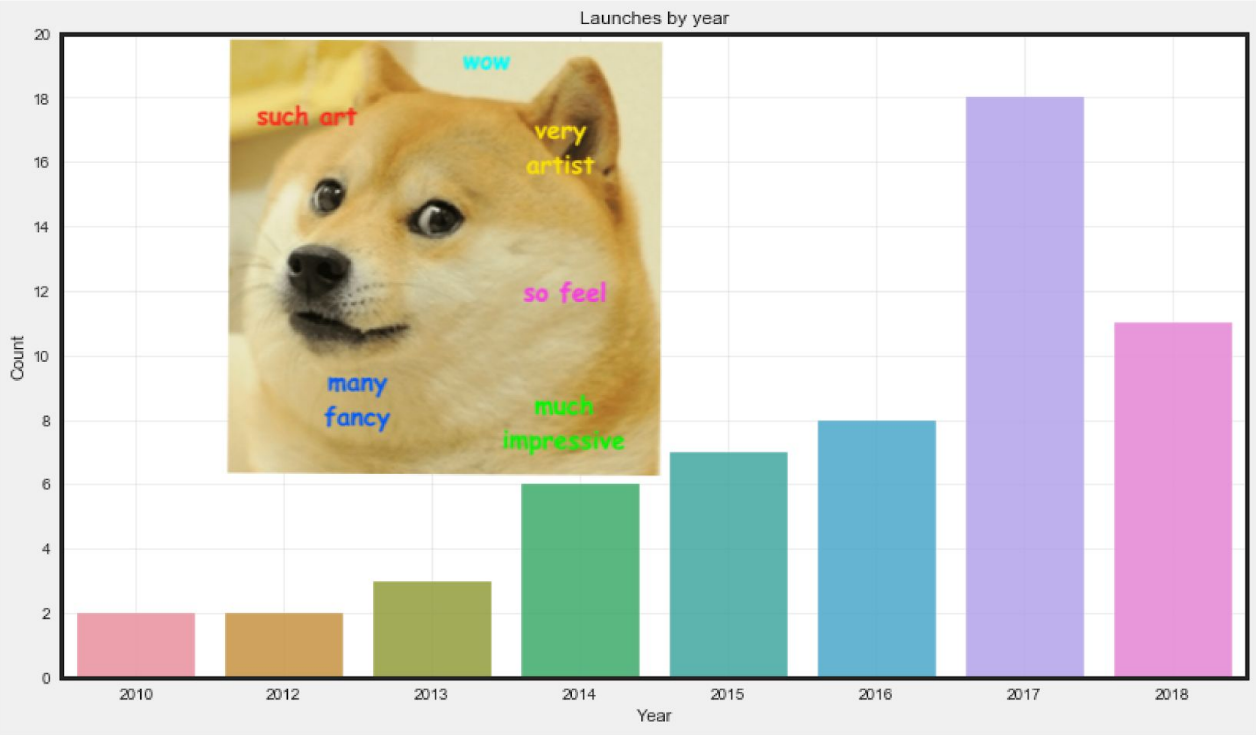
Use color purposefully



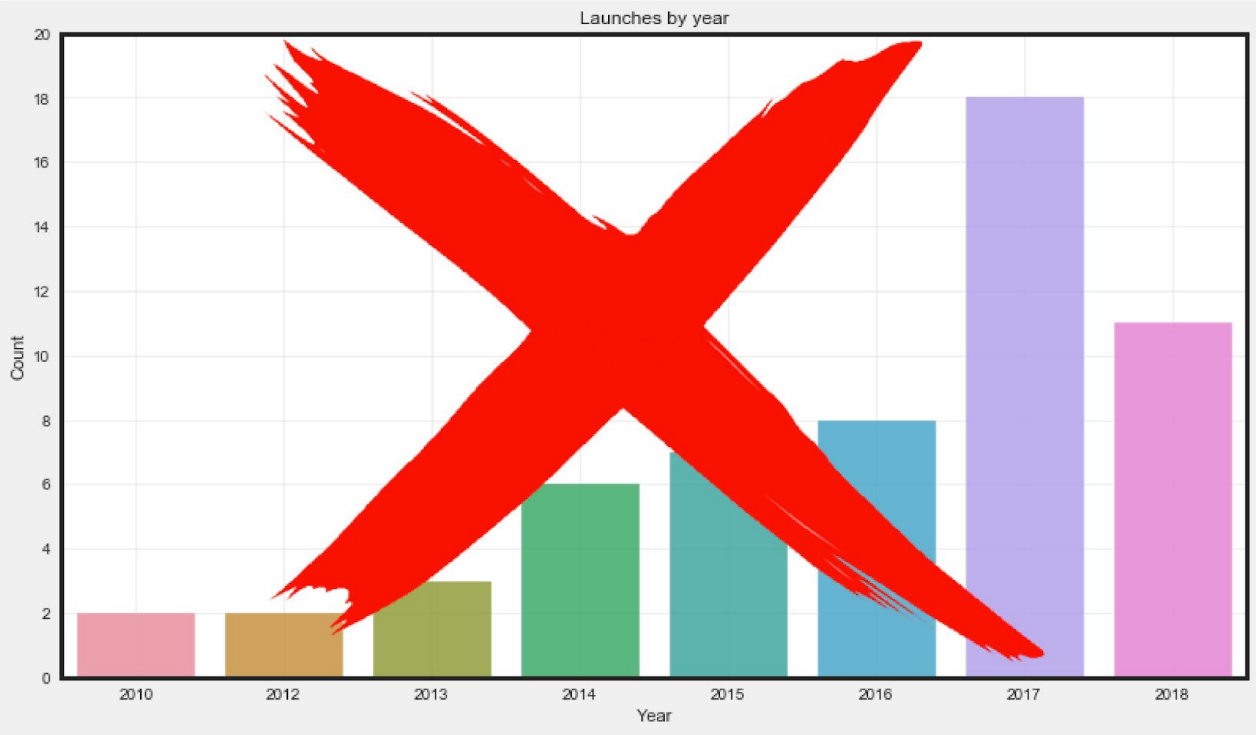
Use color purposefully



Use color purposefully



Use color purposefully



Colorblindness

- Red and green is the most common (but not the only one)
- Information and simulators online
- Existing color palettes accessible to everyone

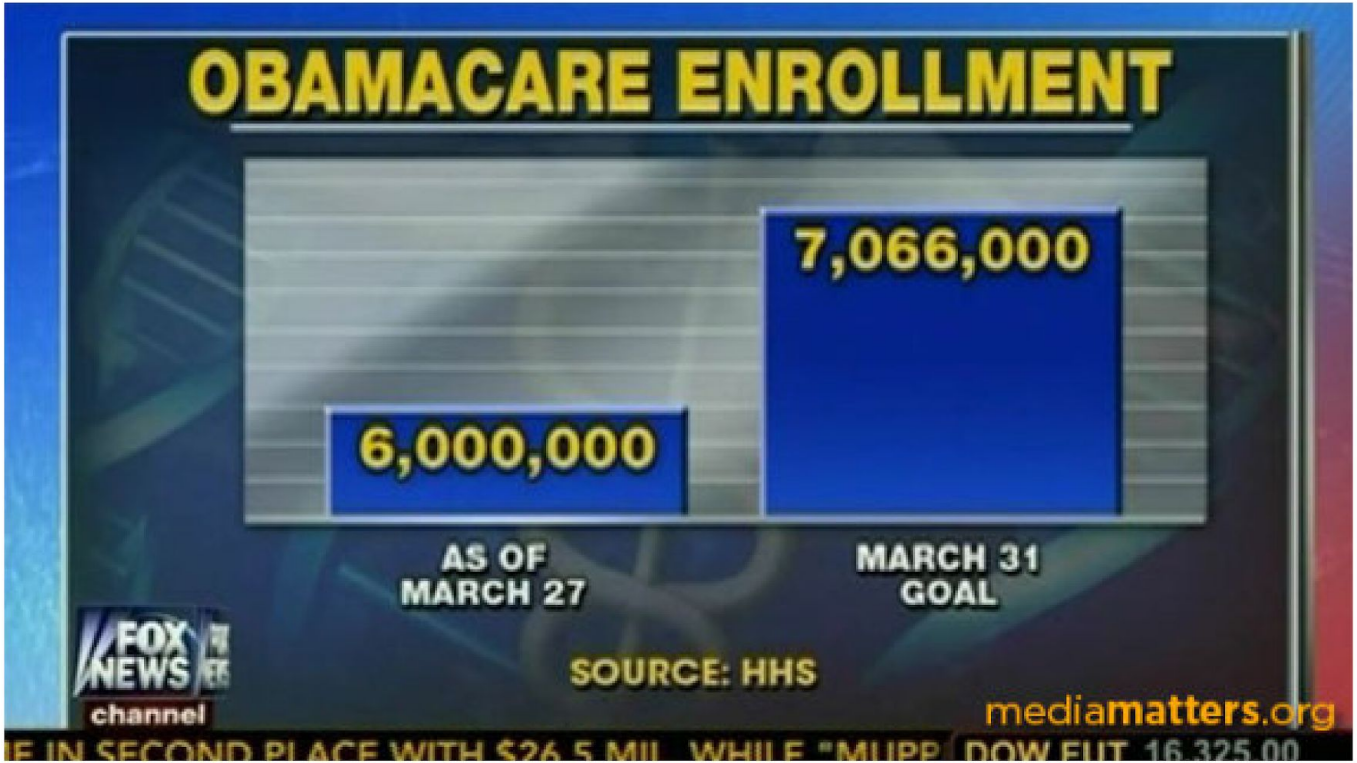
Readable fonts

- sans-serif

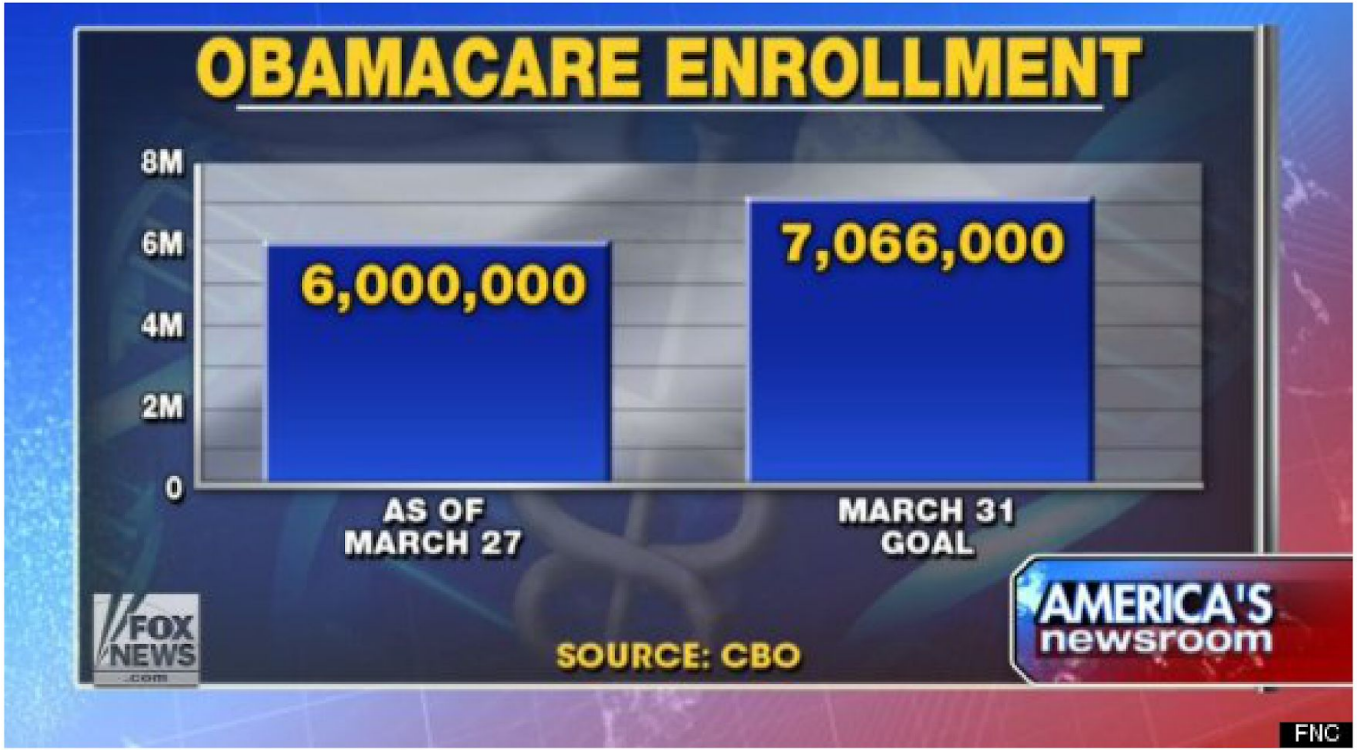
Label, label, label

- title
- x axis label
- y axis label
- legend

Axes

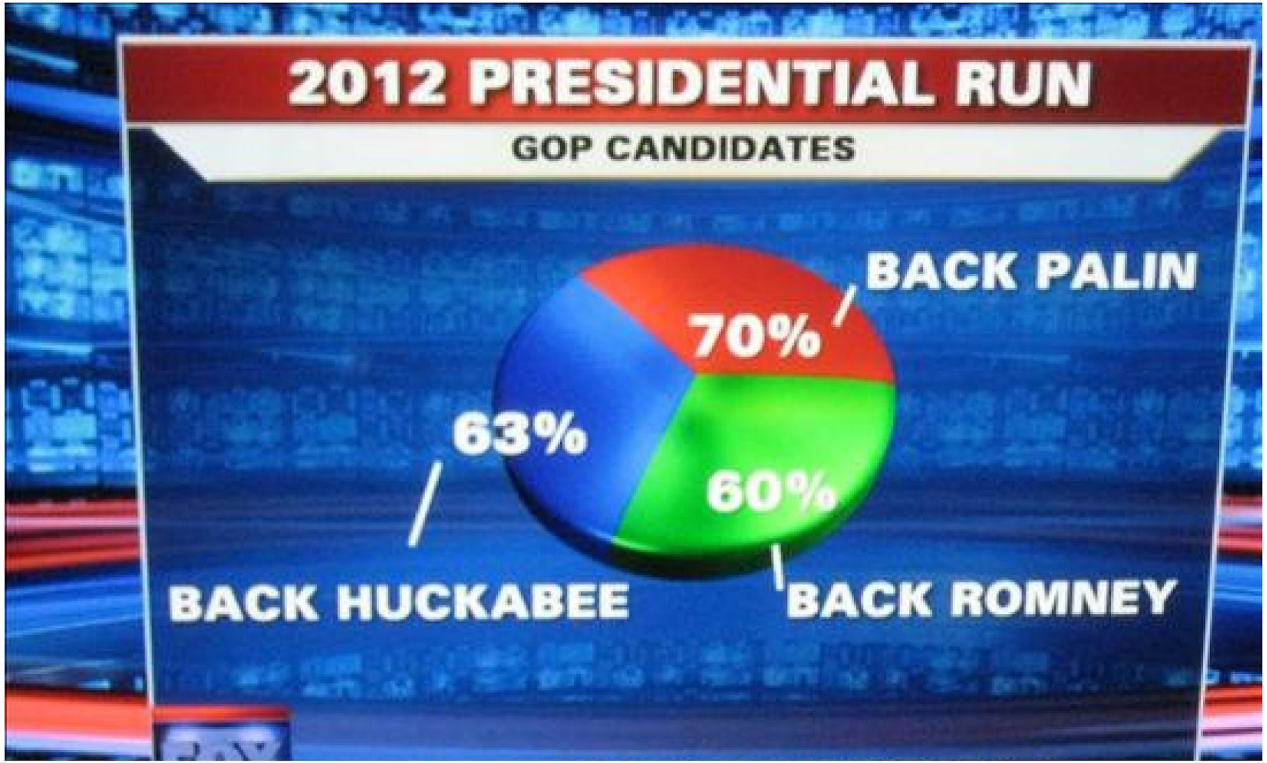


Axes



And the award goes to...

Honorable mention

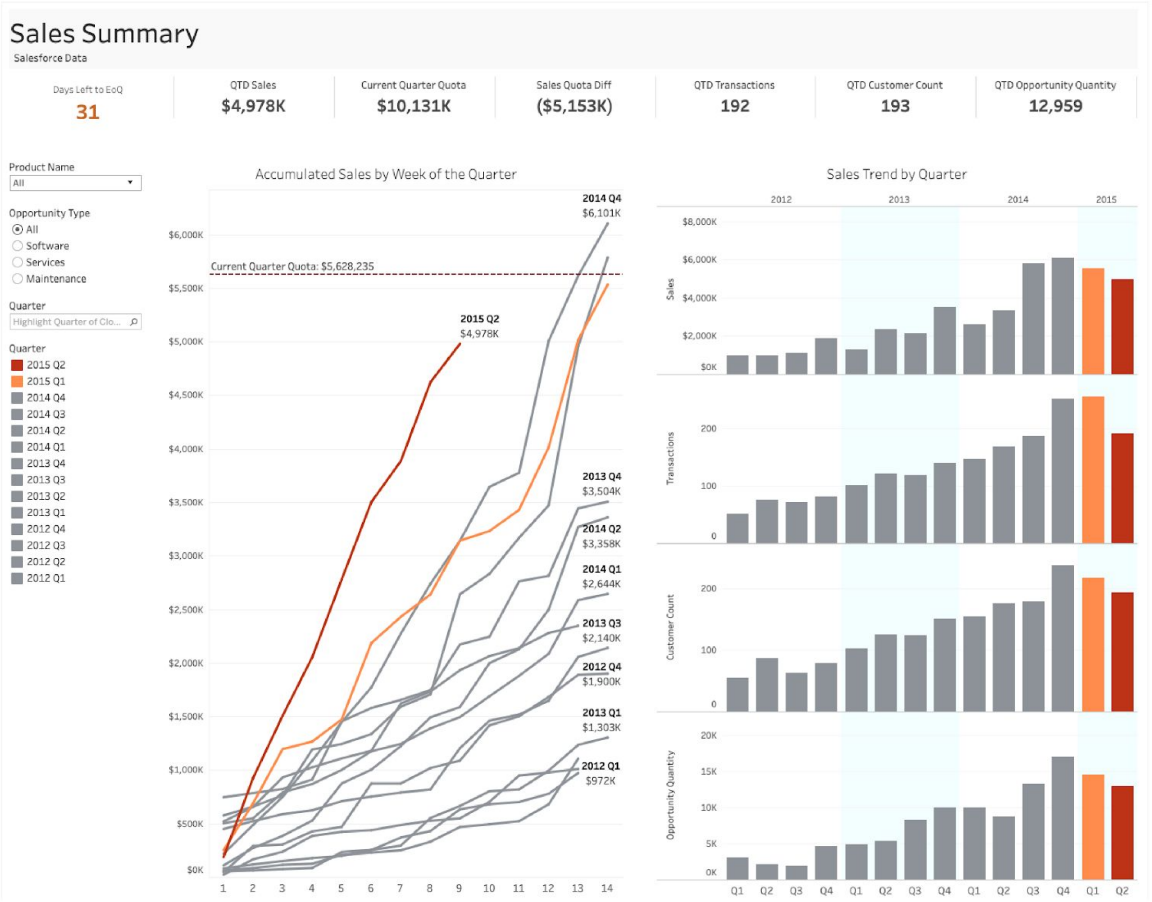


Question

$$1\ picture = 1000\ words$$

$$1000\ pictures = ?$$

A dashboard!



BI tools

tableau

looker

Power BI

Next level