

# What is data science?

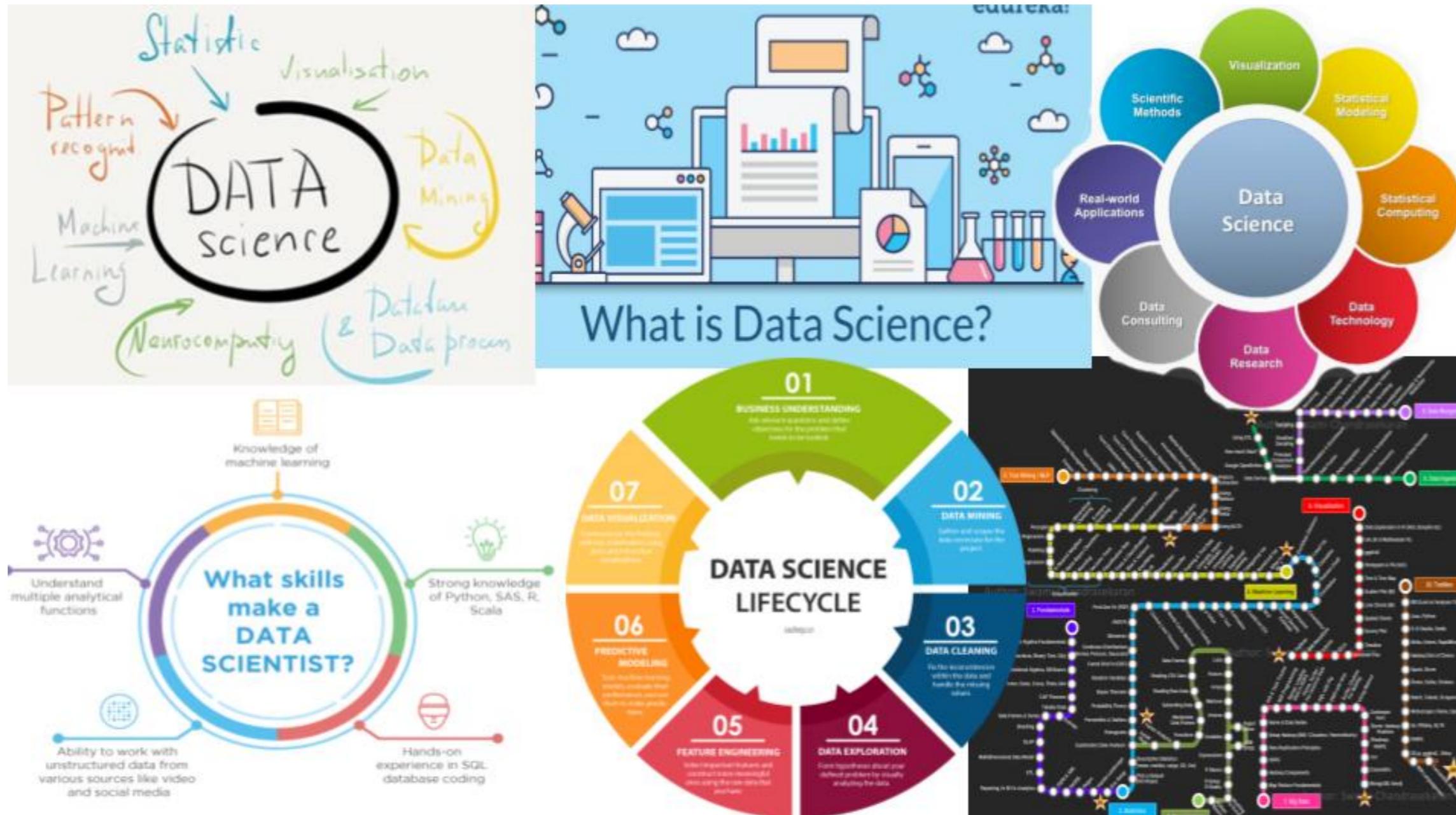
DATA SCIENCE FOR EVERYONE



Lis Sulmont

Curriculum Manager, DataCamp

# Let's ask Google!



# Making data work for you

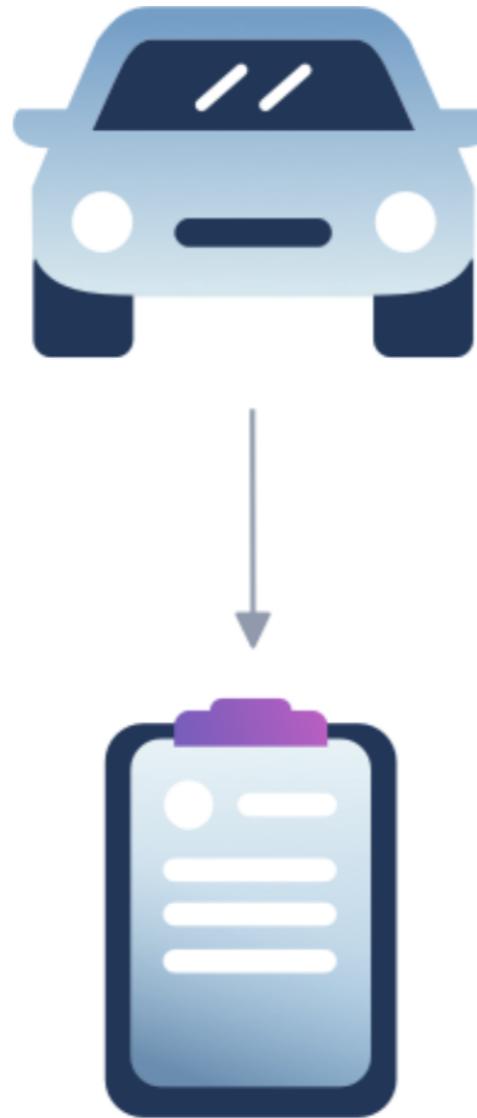


**Use data to better describe the present or better predict the future**

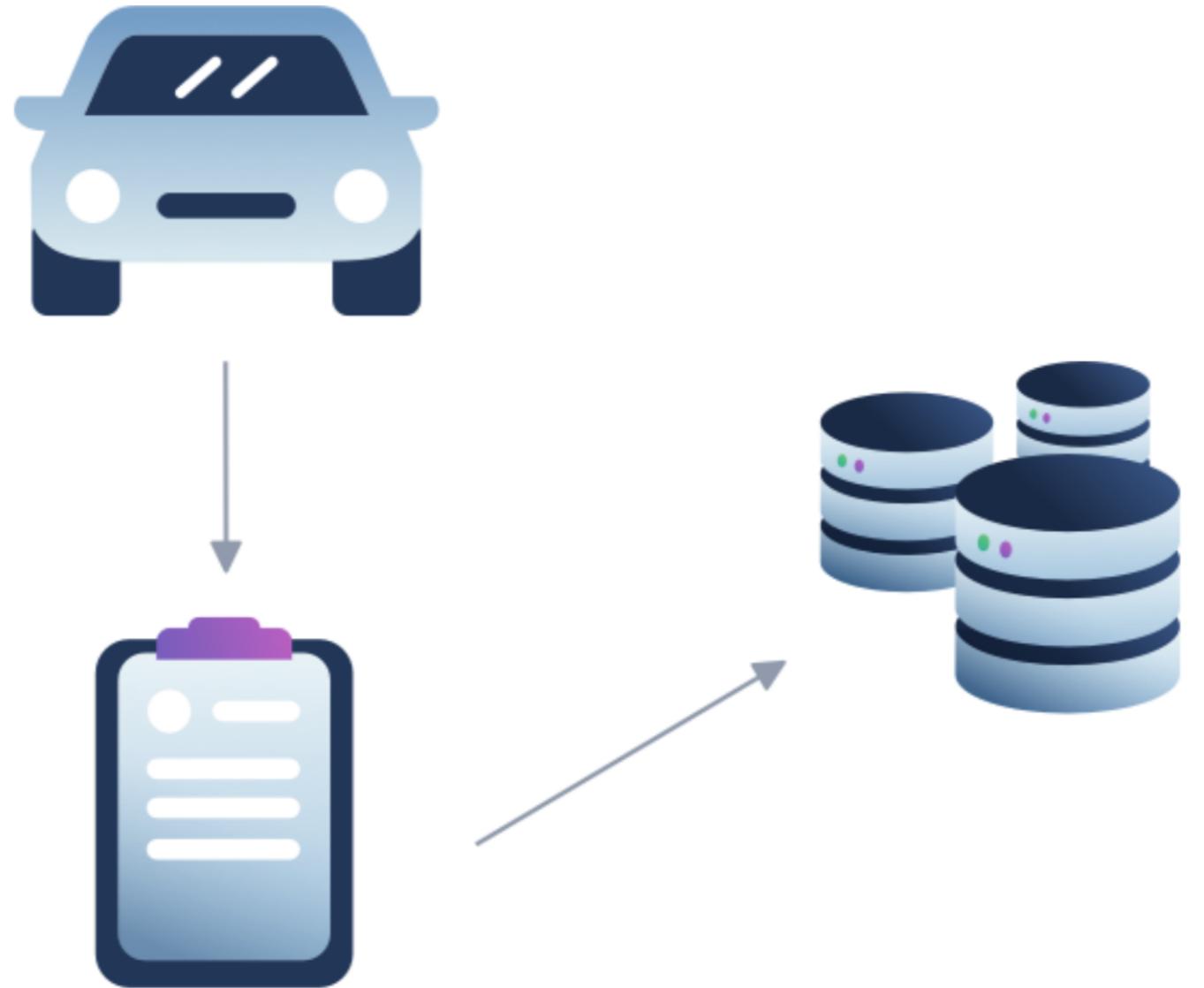
# What can data do?

- 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?

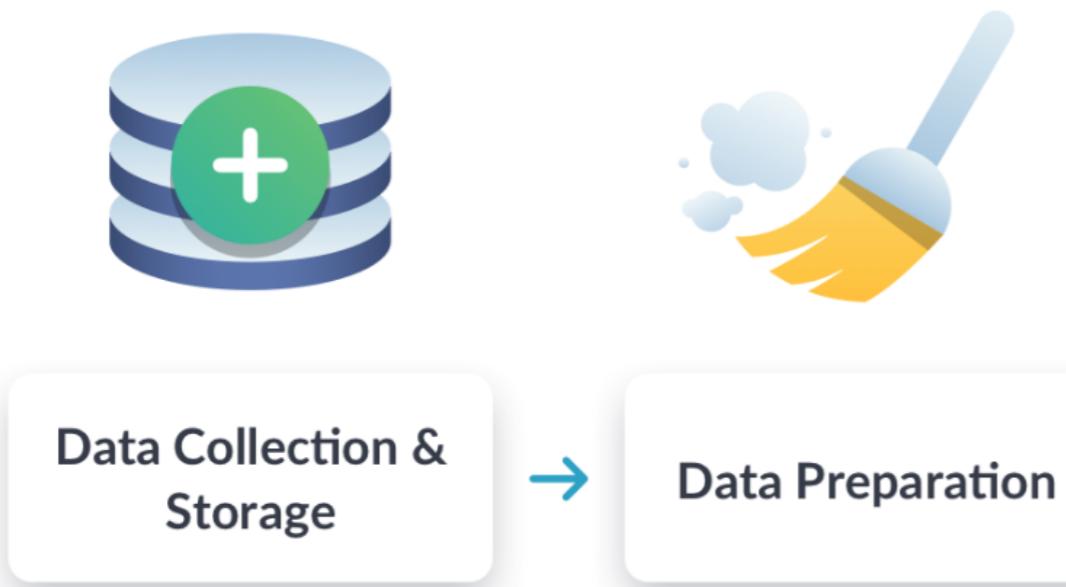


# The data science workflow

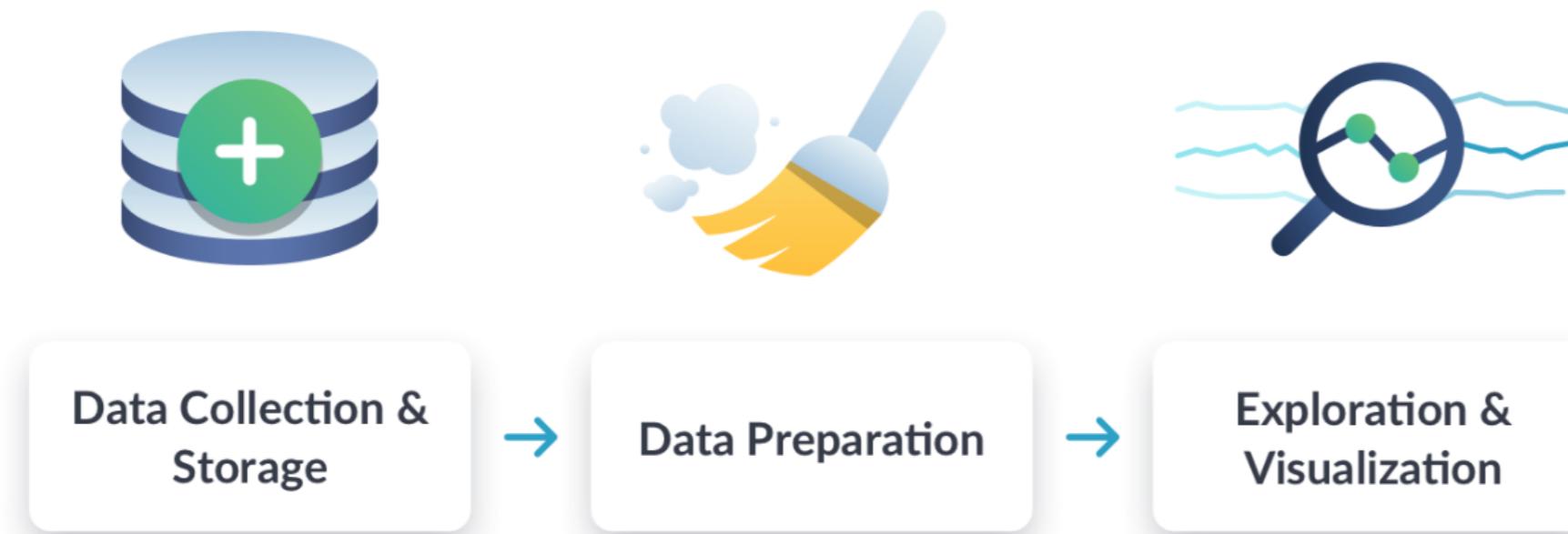


Data Collection &  
Storage

# The data science workflow



# The data science workflow



# The data science workflow



# **Let's practice!**

**DATA SCIENCE FOR EVERYONE**

# Applications of data science

DATA SCIENCE FOR EVERYONE



Lis Sulmont

Curriculum Manager, DataCamp

# More case studies

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

# Case study: fraud detection



# Case study: fraud detection

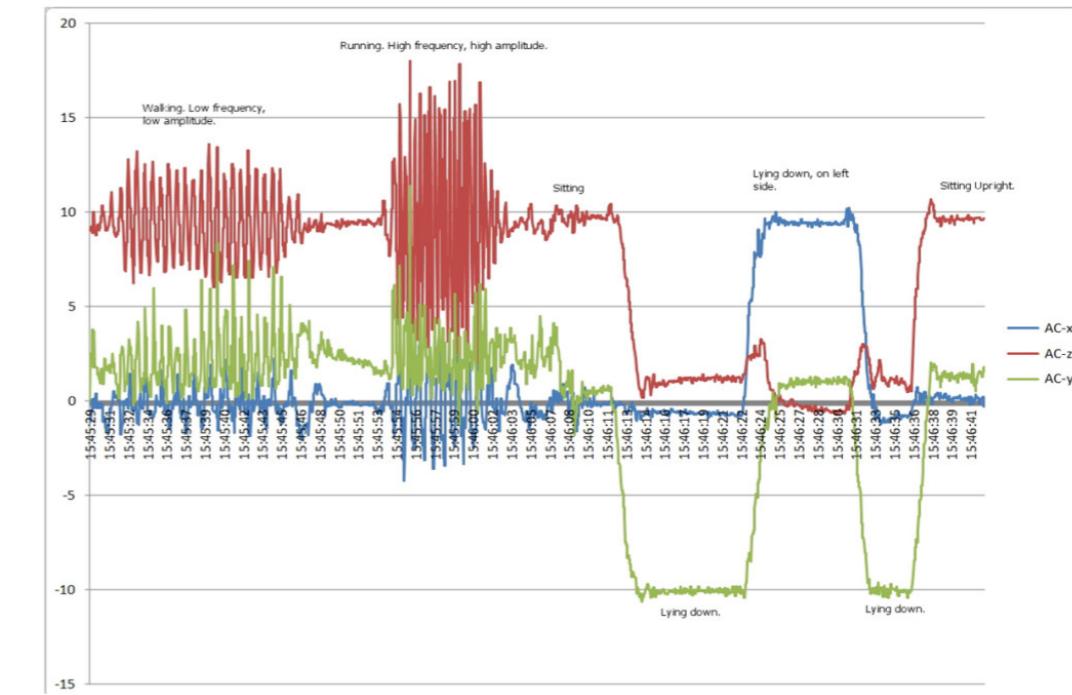
Amount	Date	Location	...
149.62	2019-05-23	London	...
2.69	2018-10-03	Birmingham	...
378.66	2019-06-15	Liverpool	...
123.5	2019-01-12	London	...
69.99	2018-06-16	São Paolo	...
3.67	2019-03-06	Brussels	...
...	...	...	...



# 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

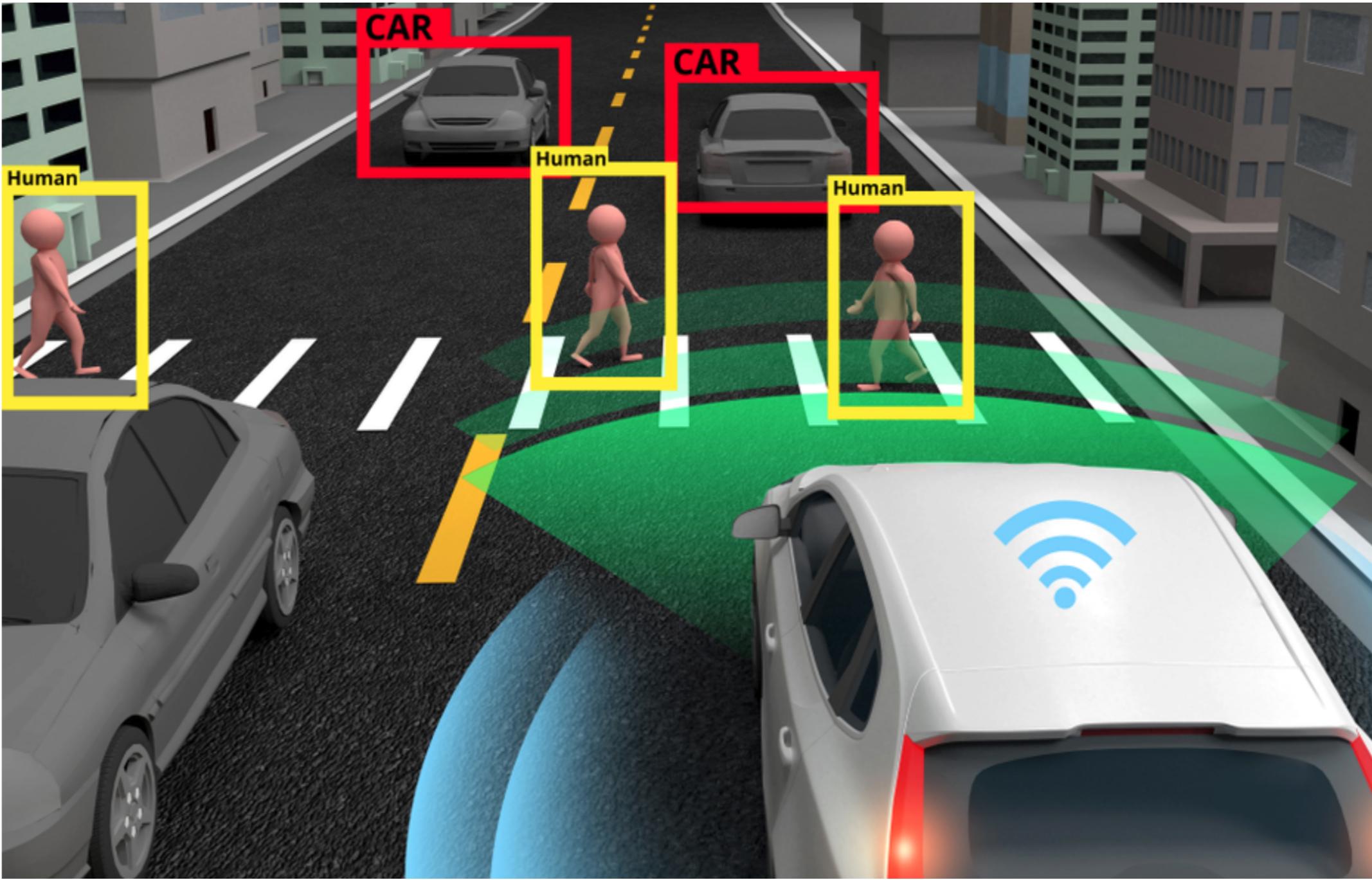


# Internet of Things (IoT)

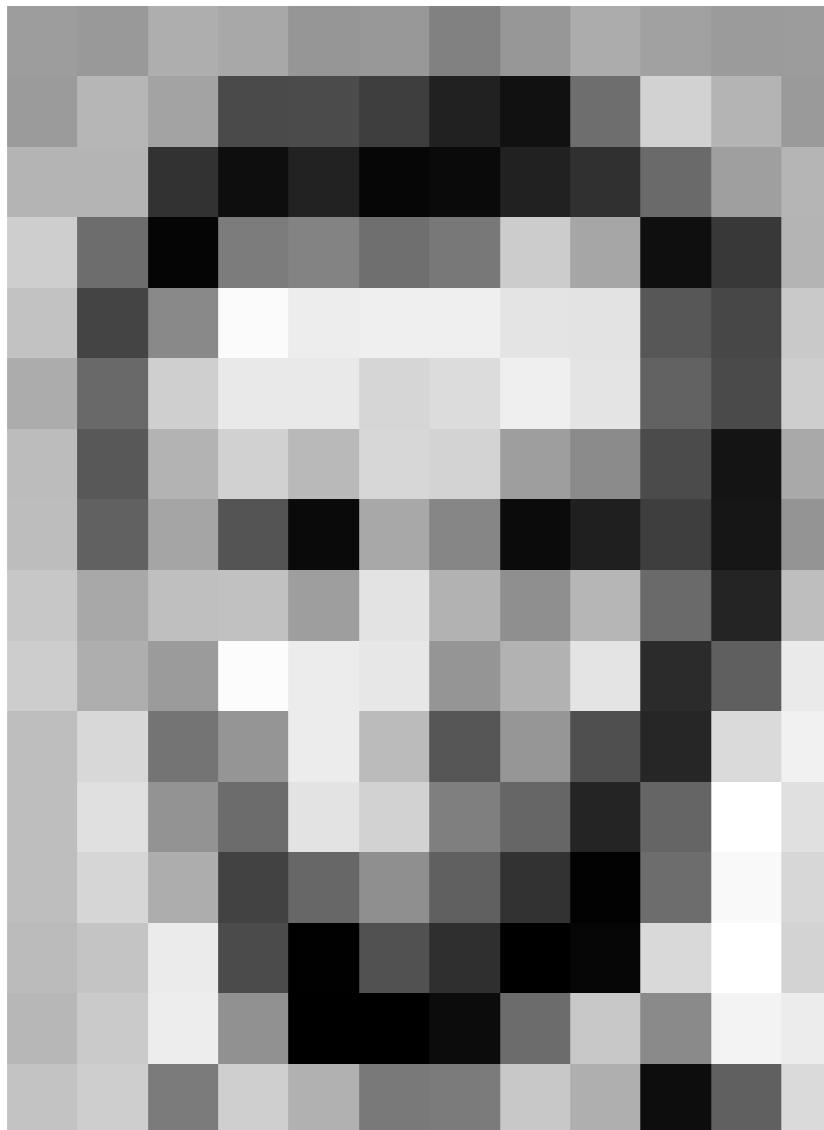
Refers to gadgets that aren't standard computers

- 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	165	166
155	182	168	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	197	251	237	239	239	228	227	87	71	201
172	165	207	239	239	214	220	239	228	98	74	206
188	88	179	209	185	215	211	198	139	75	20	169
149	97	165	84	10	168	134	11	31	62	22	148
199	168	191	163	158	227	178	143	182	105	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	168	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	236	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
196	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	168	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	197	251	237	239	239	228	227	87	71	201
172	165	207	239	239	214	220	239	228	98	74	206
188	88	179	209	185	215	211	198	139	75	20	169
149	97	165	84	10	168	134	11	31	62	22	148
199	168	191	163	158	227	178	143	182	105	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	168	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	236	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
196	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

# **Let's practice!**

**DATA SCIENCE FOR EVERYONE**

# Data science roles and tools

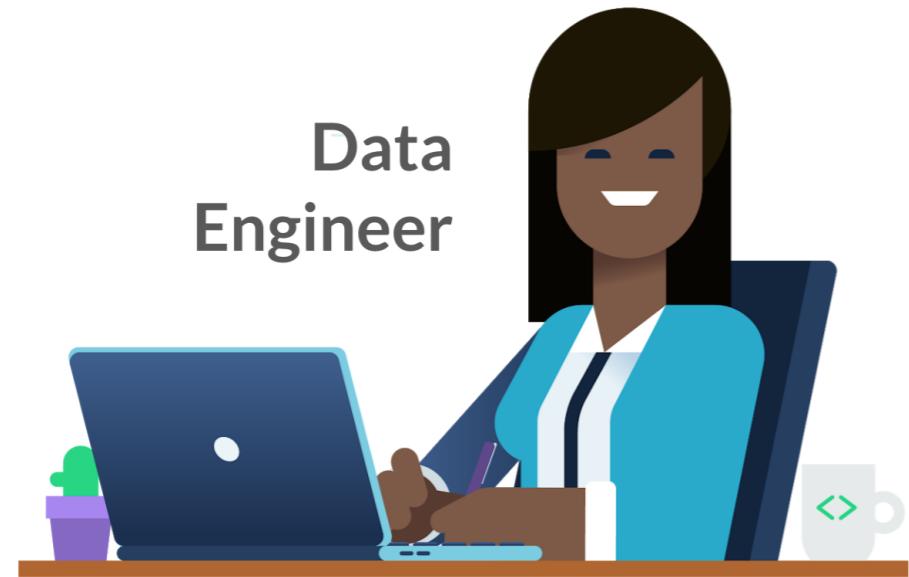
DATA SCIENCE FOR EVERYONE



Lis Sulmont

Curriculum Manager, DataCamp

Data  
Engineer



Data  
Analyst



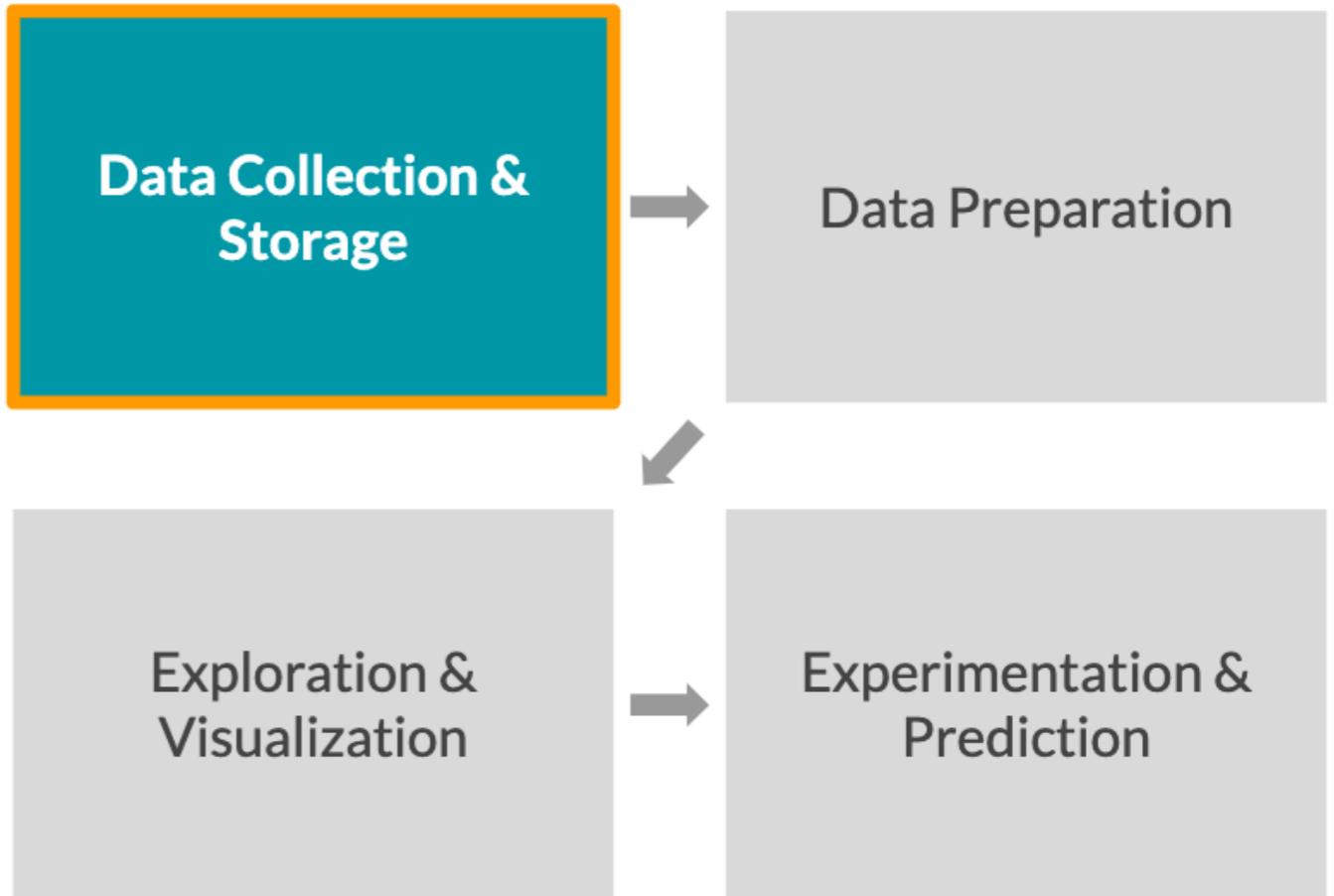
Data  
Scientist

Machine Learning  
Scientist



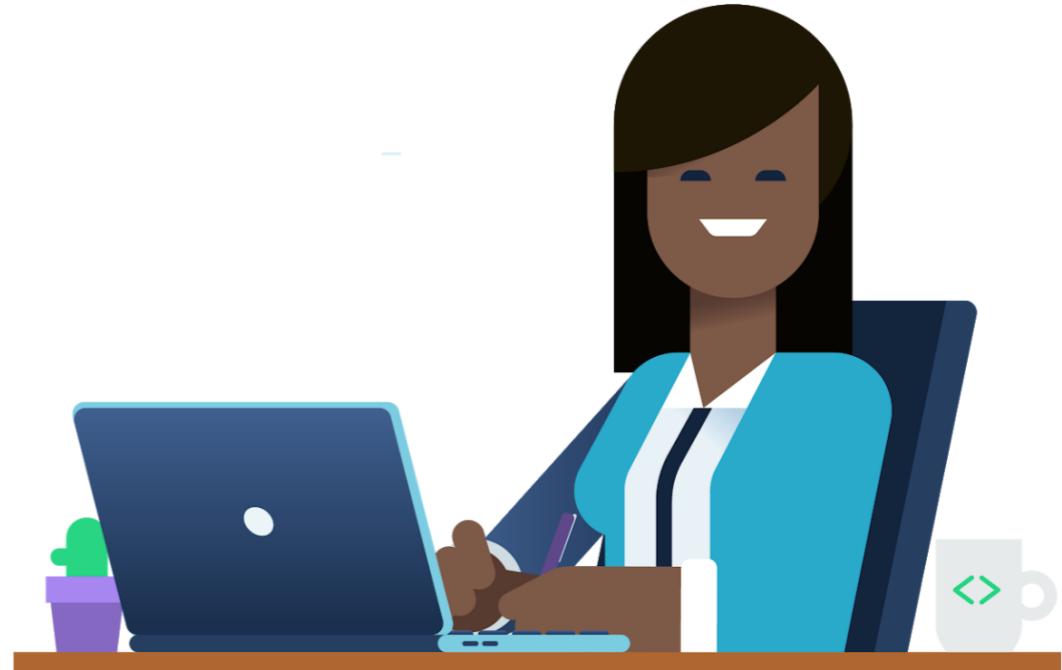
# Data engineer

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



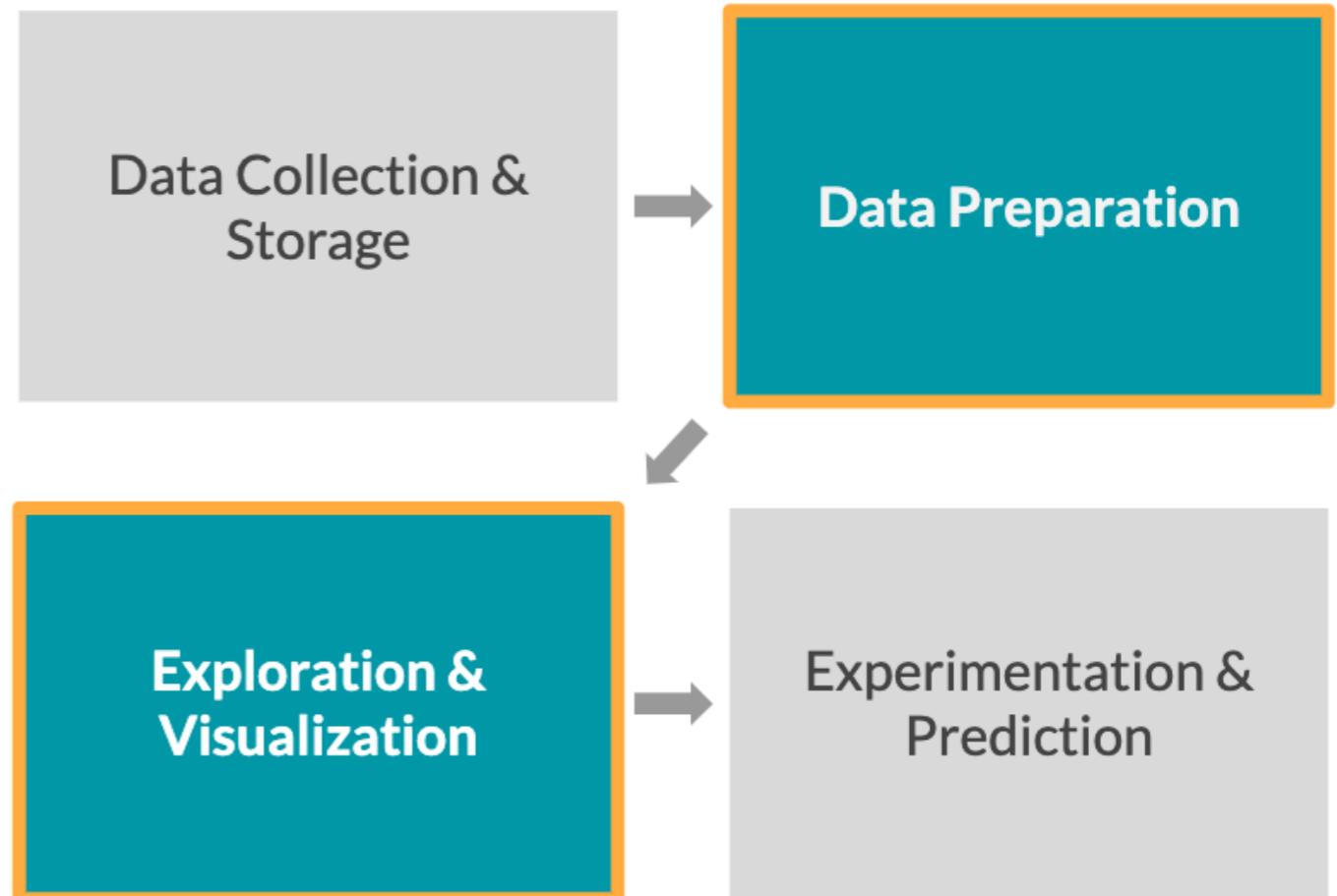
# Data engineering tools

- **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



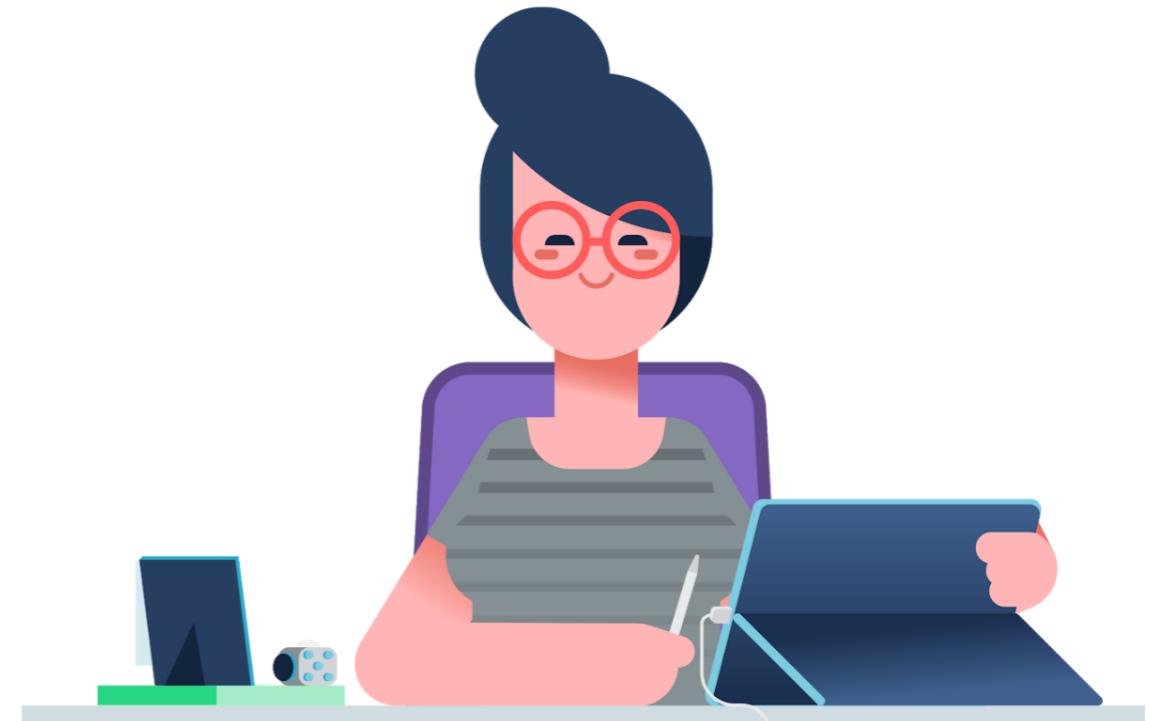
# Data analyst

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



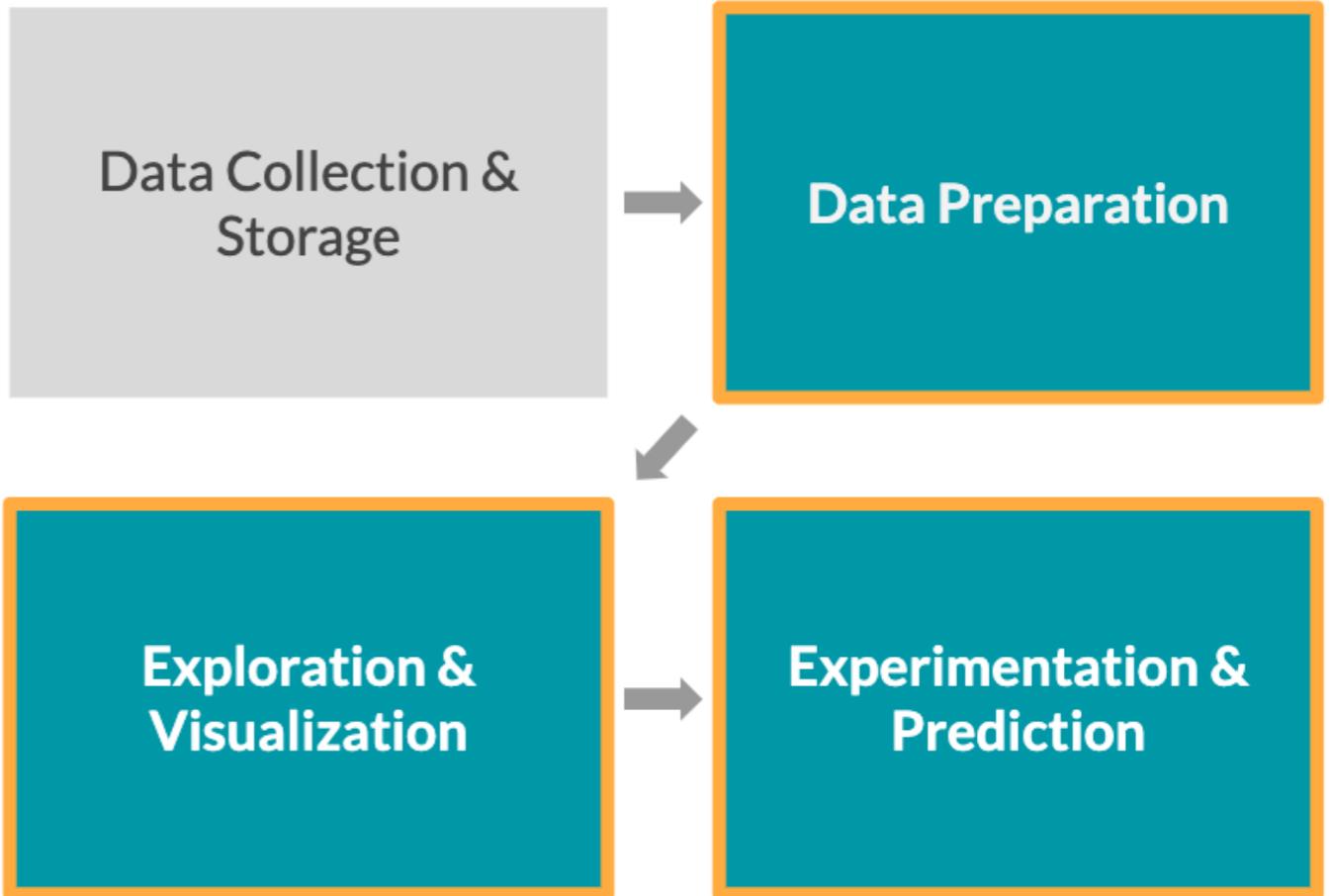
# Data analyst tools

- **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



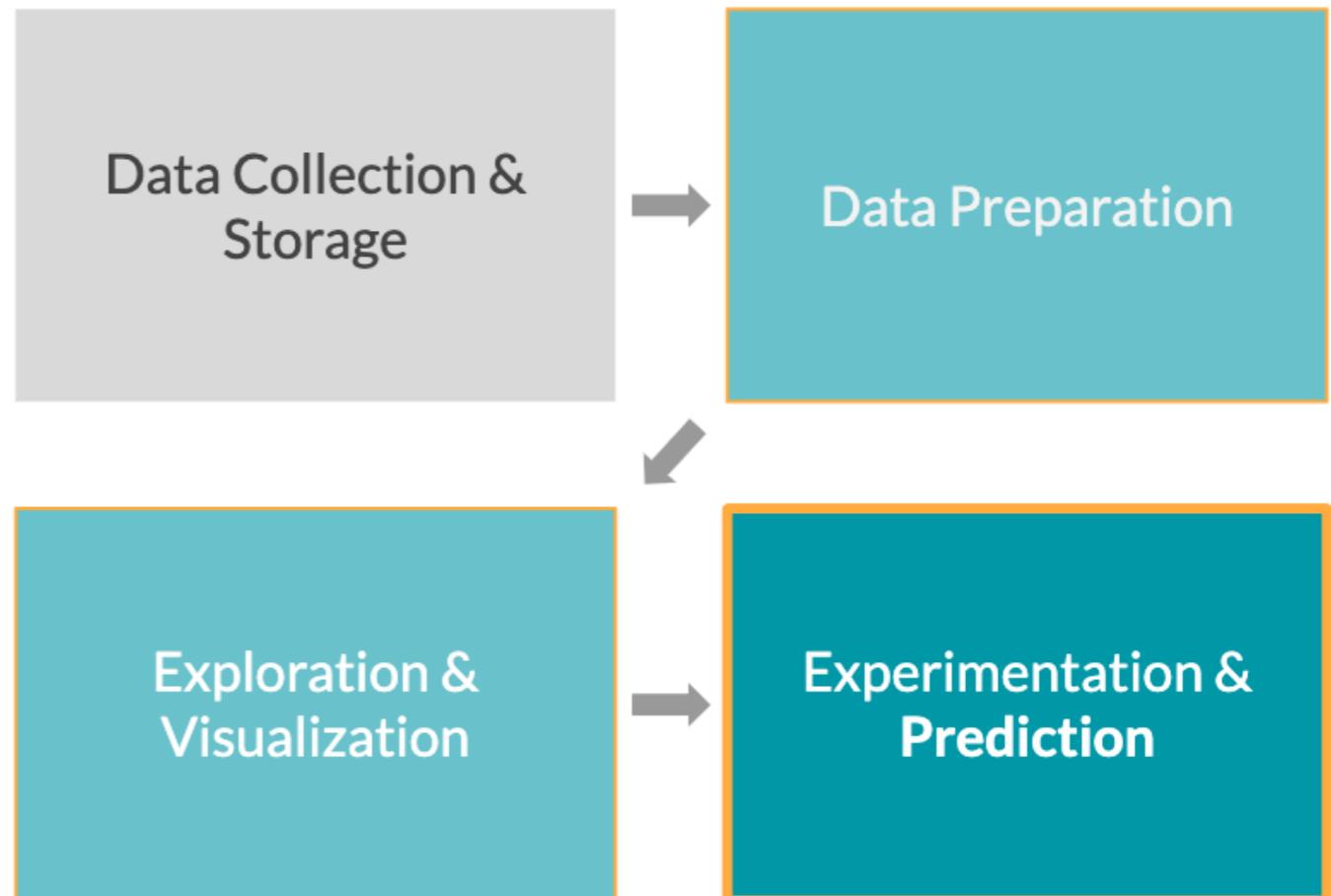
# Data scientist tools

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



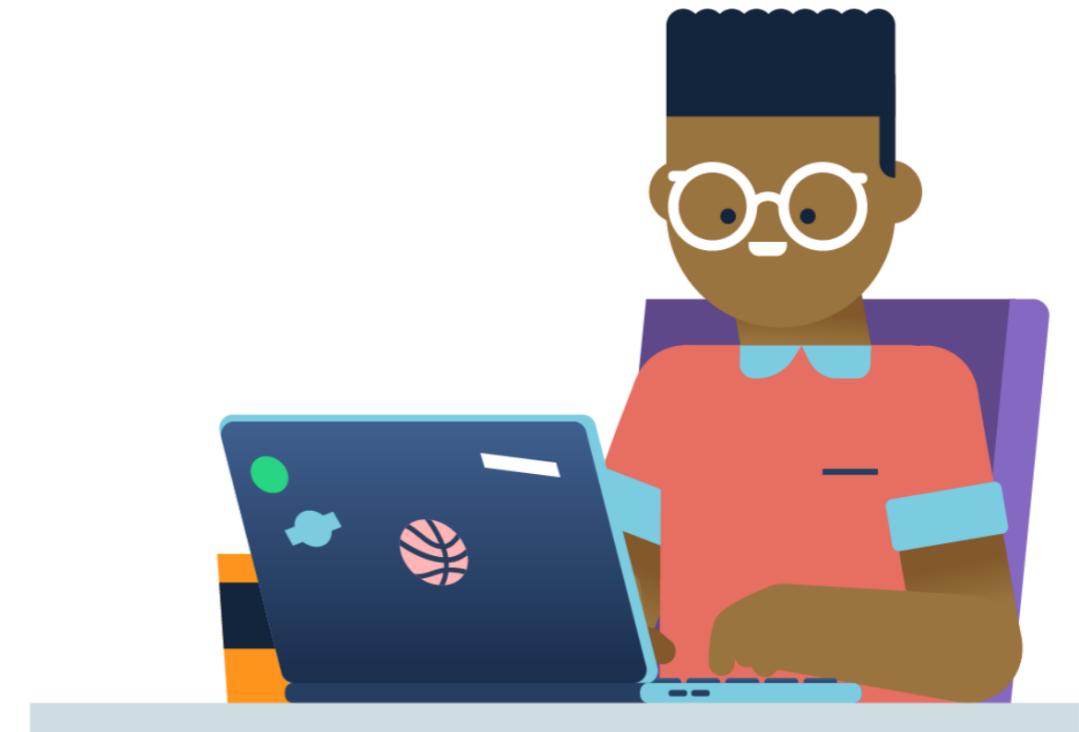
# Machine learning scientist

- 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

# **Let's practice!**

**DATA SCIENCE FOR EVERYONE**