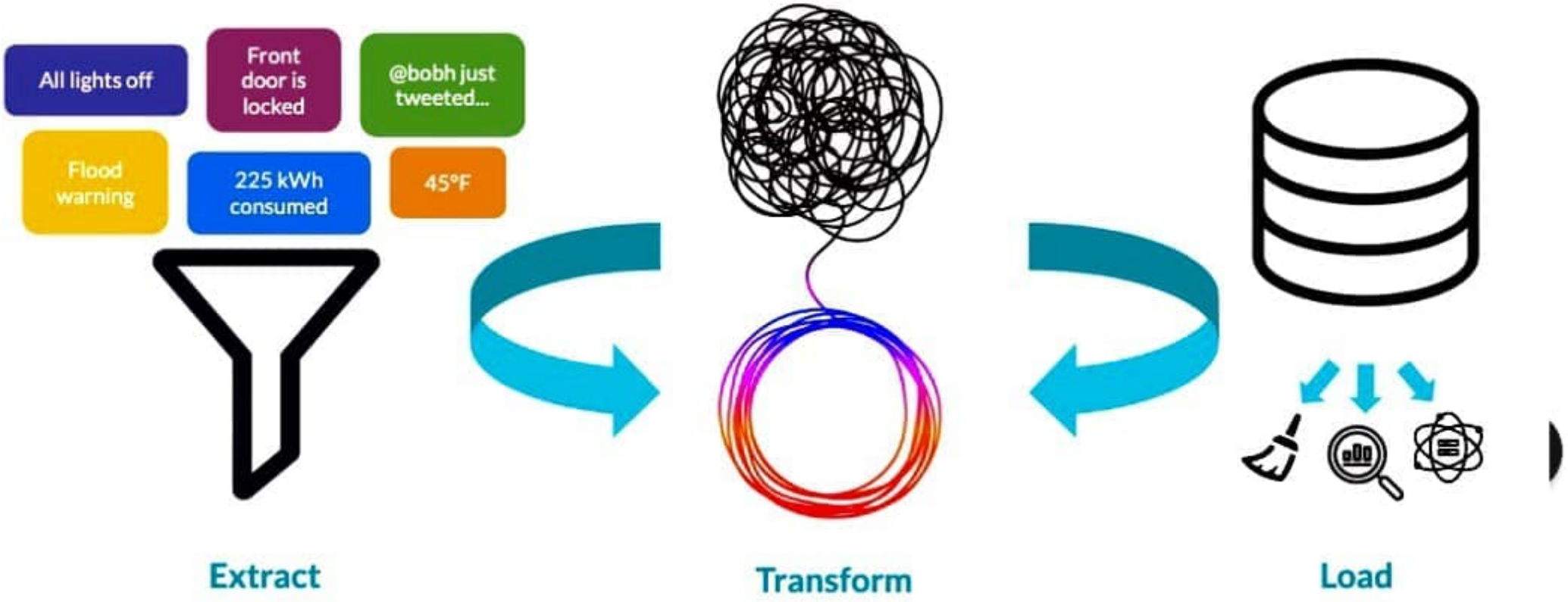
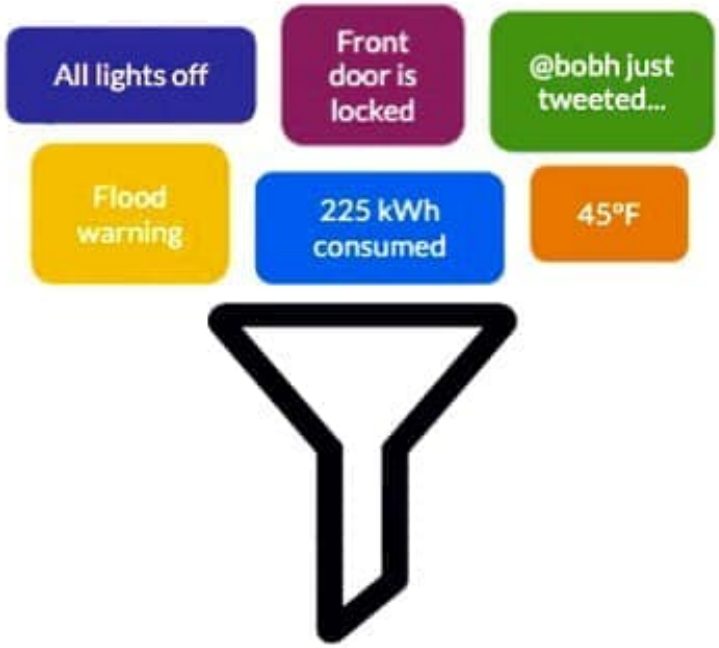
Automation


All lights off
Front door is locked
@bobh just tweeted...
Flood warning
225 kWh consumed
45°F
Extract
Transform
Load

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

Extract	Transform	Load
Collect the latest songs listened to by users on the mobile app	Group and arrange user listening data by musician	Store listening data into a database used by machine learning scientists to generate personalized playlists
Use the Apple store API to get latest download and rating information	Combine data on different users into one dataset for all users	

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

BI tools

tableau

looker

Power BI

Supervised machine learning recap

- Make a prediction based on data
- Data has *features* and *labels*
  - Label: what we want to predict
  - Features: data that might predict the label
- Trained model can make predictions

What is supervised machine learning?

- **Machine learning:** Predictions from data
- ***Supervised machine* learning:** Predictions from data with *labels* and *features*
  - Recommendation systems
  - Diagnosing biomedical images
  - Recognizing hand-written digits
  - Predicting customer churn

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)

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

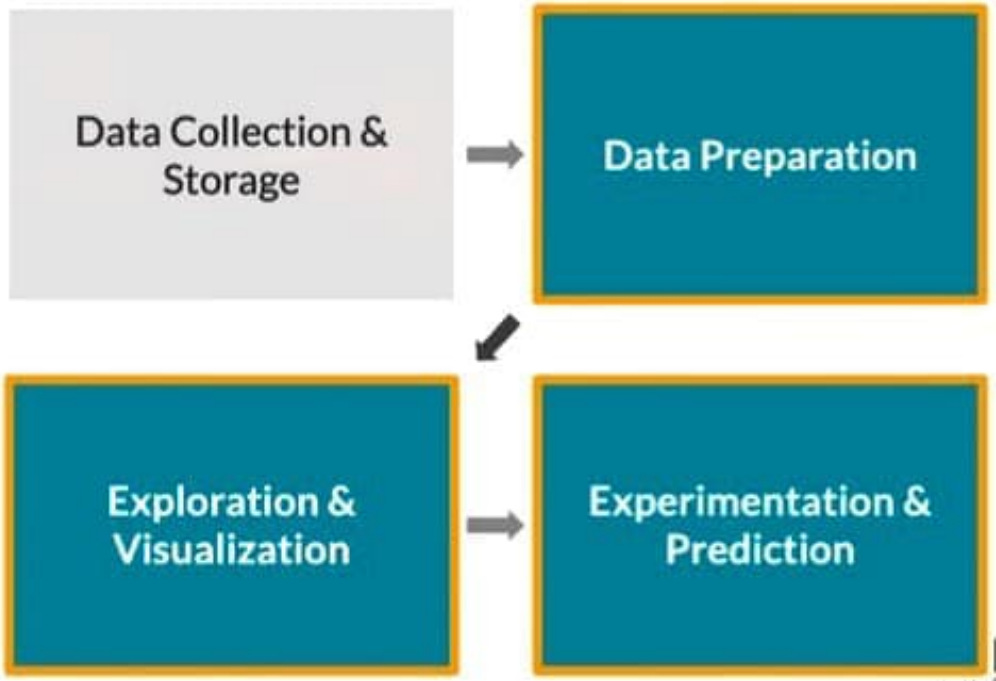
Public data APIs

- Application Programming Interface
- Request data over the internet

- Twitter
- Wikipedia
- Yahoo! Finance
- Google Maps
- Many more!

Data scientist

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



Data Engineer

- Give new team members database access
- Create a new table in the SQL database

Data Analyst

- Update Excel spreadsheet with new graphs
- Create a dashboard for the Marketing team

Data Scientist

- Train an anomaly detection algorithm
- Run a correlation analysis between weather and ice cream sales

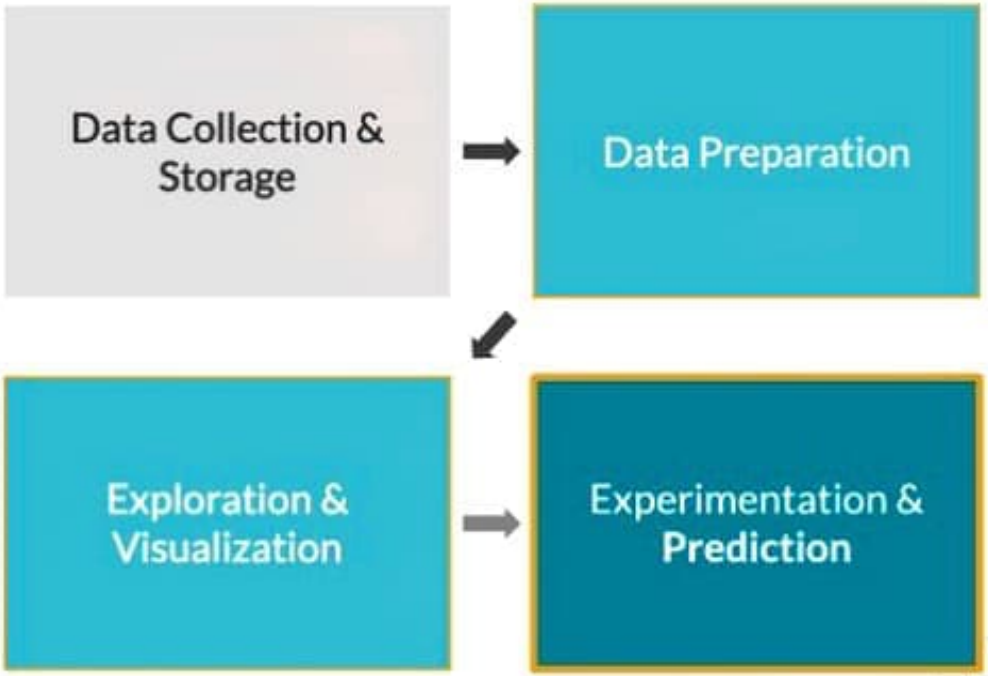
Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala /Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

Machine learning tools

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

Machine learning scientist

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



Data scientist tools

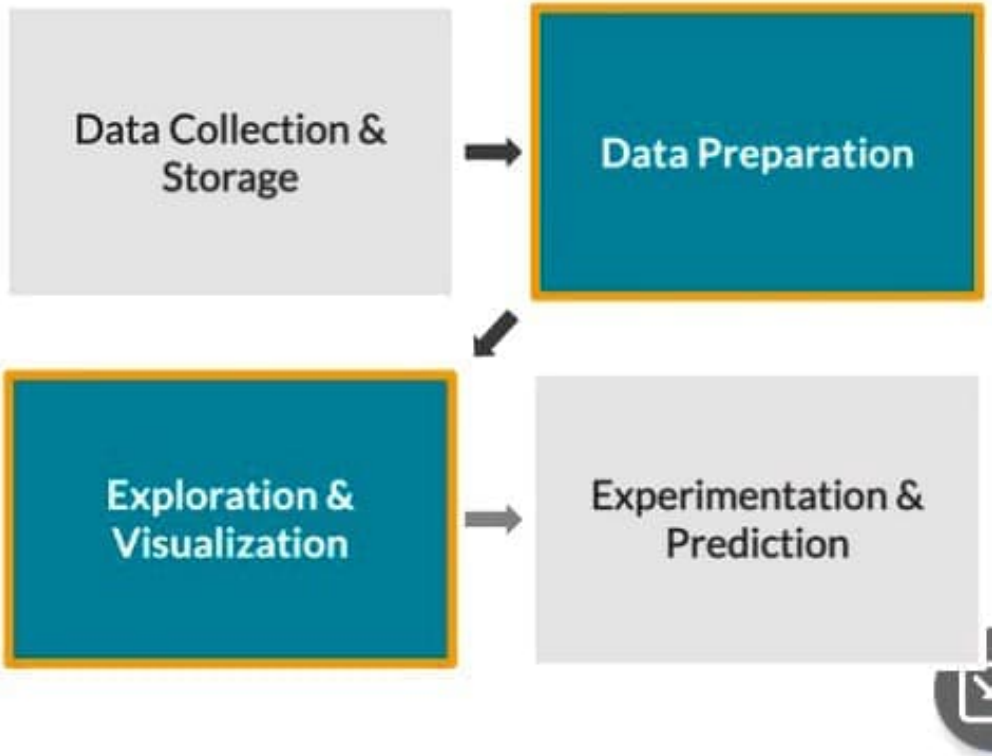
- **SQL**
  - Retrieve and aggregate data
- **Python and/or R**
  - Data science libraries, e.g., `pandas` (Python) and `tidyverse` (R)

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 analyst

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

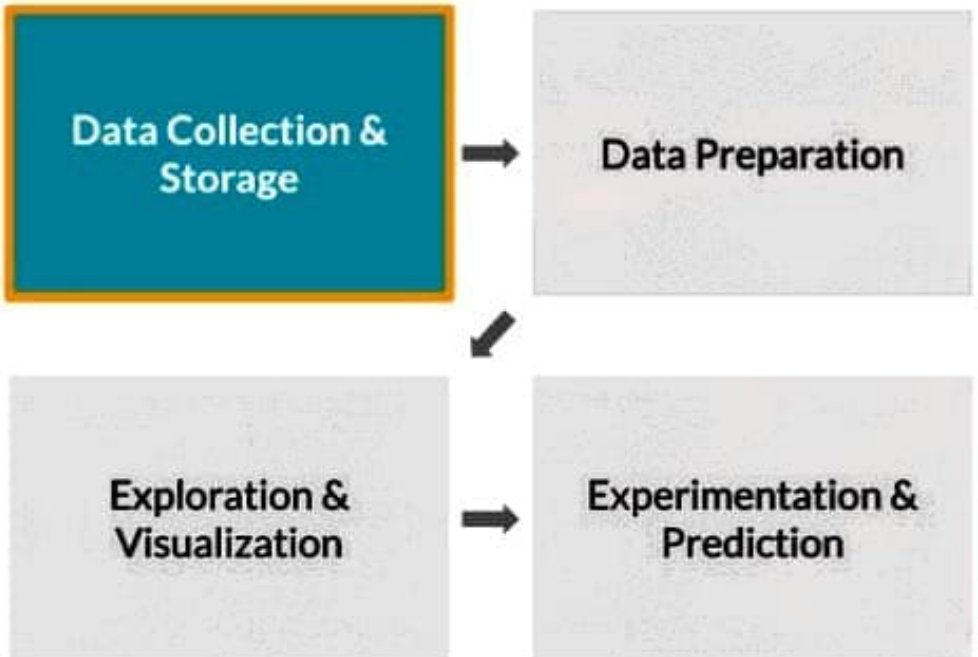


Data engineering tools

- **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 engineer

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



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!