Data Wrangling and Data Analysis
Welcoming

Hakim Qahtan

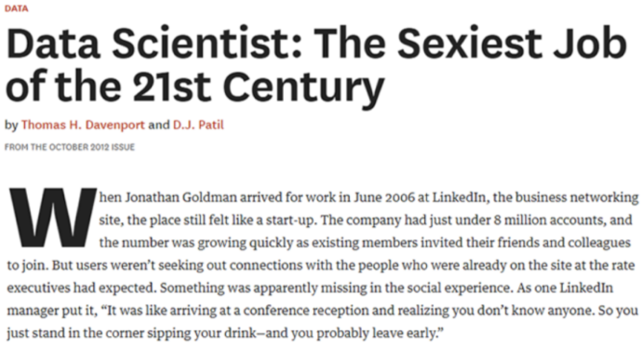
Department of Information and Computing Sciences

Utrecht University

Utrecht University

Data Science - Myth

October 2012 Issue



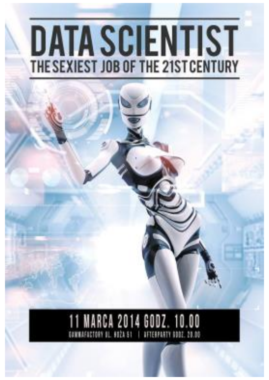
DATA

Data Scientist: The Sexiest Job of the 21st Century

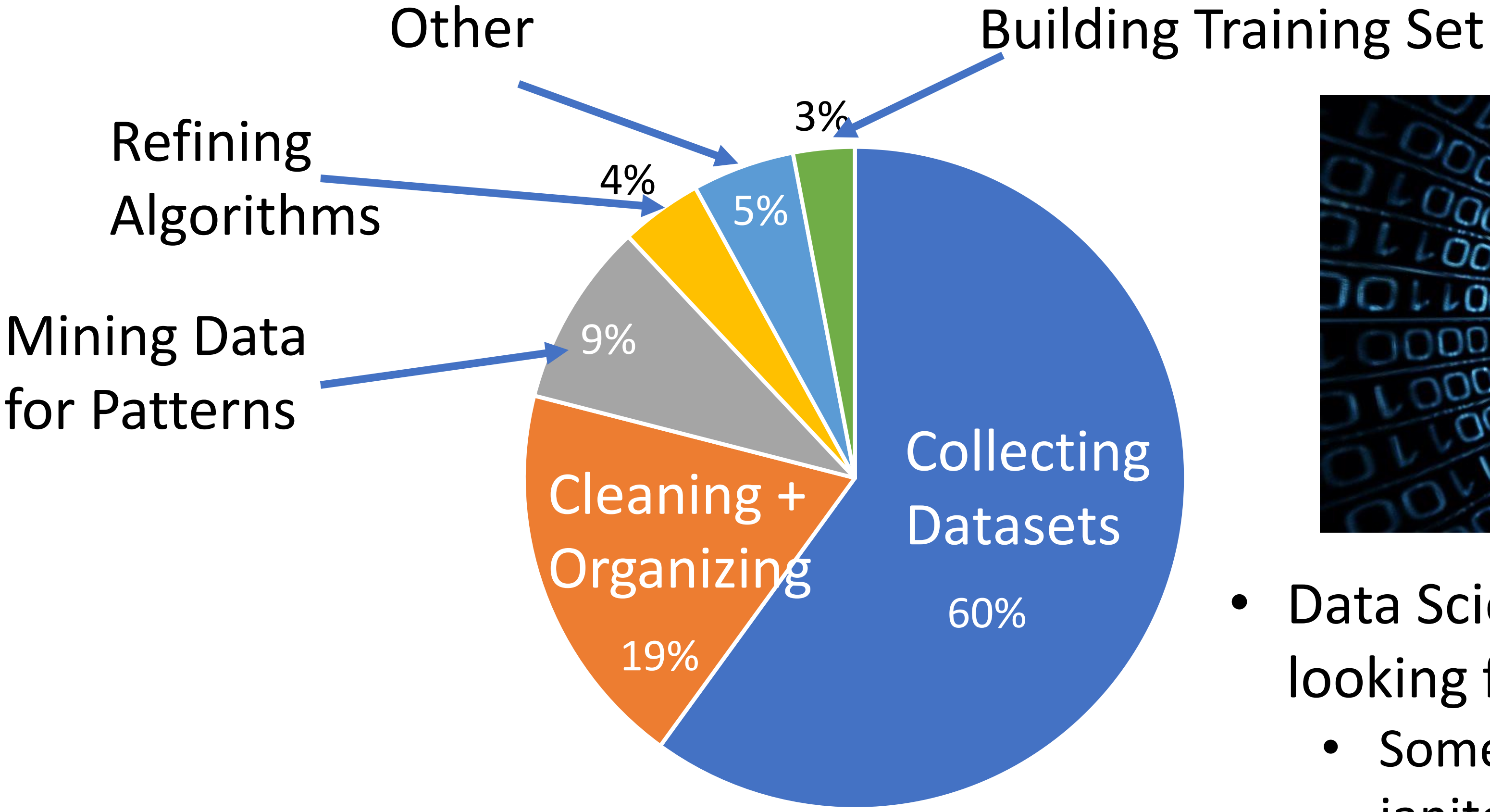
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

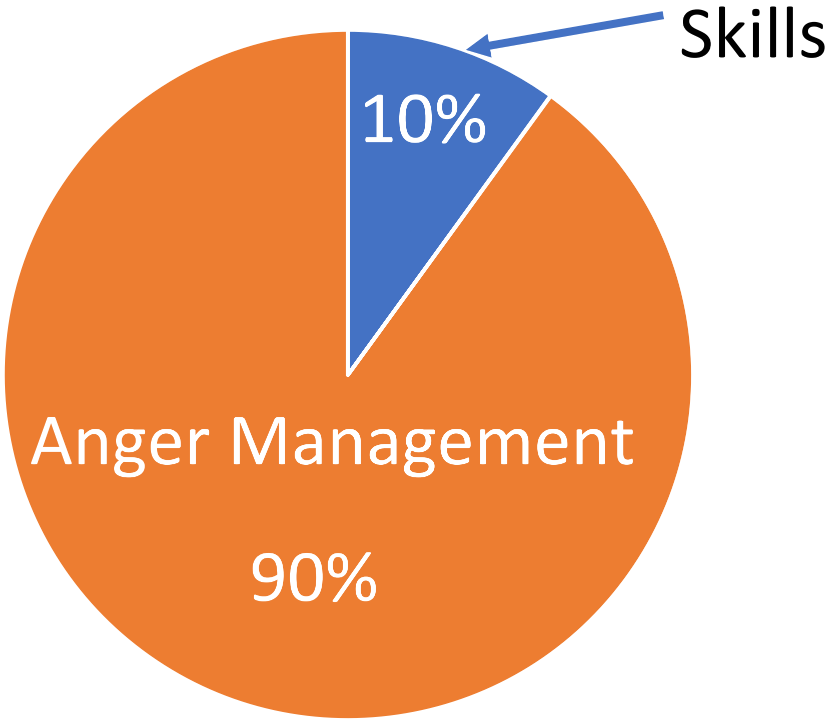

DATA SCIENTIST
THE SEXIEST JOB OF THE 21ST CENTURY
11 MARCA 2014 GODZ. 10.00

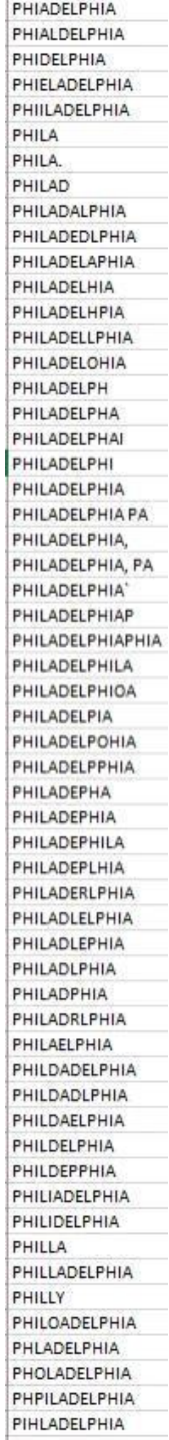
Data Science - Reality



- Data Scientists: 80% of their time looking for and cleaning the data
  - Sometimes, they are called data janitors

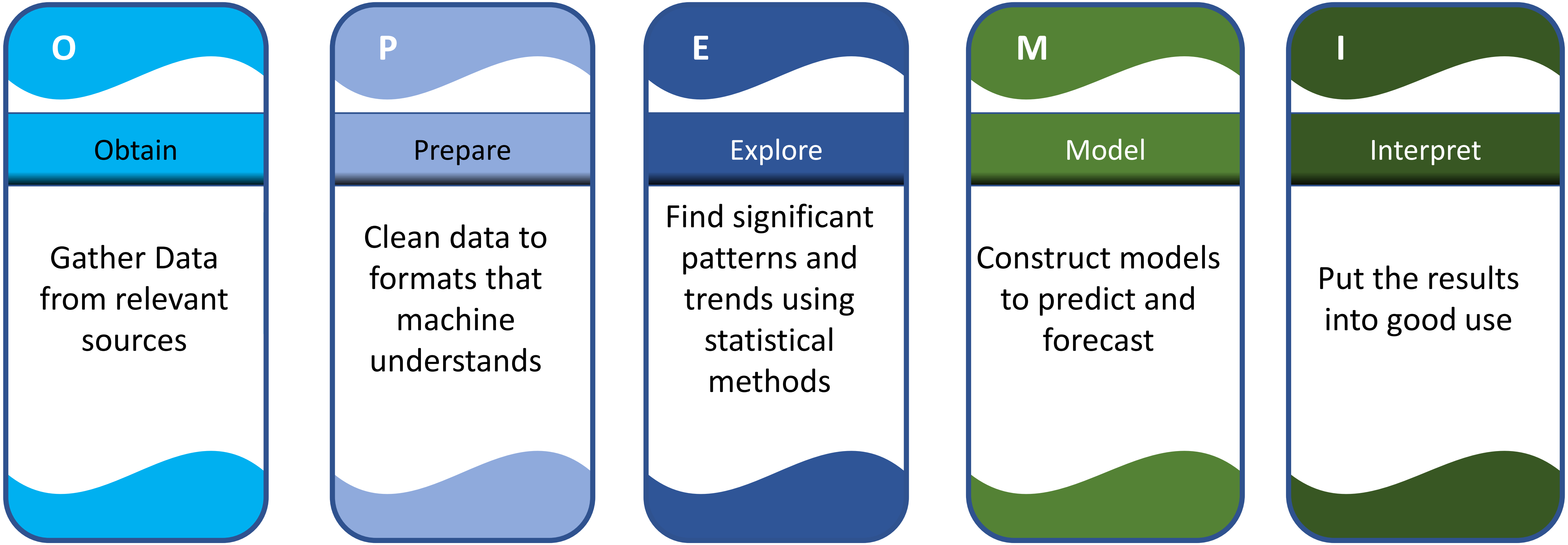
What about Data Wrangling?





PHIADELPHIA
PHIALDELPHIA
PHIDELPHIA
PHIELADELPHIA
PHIILADELPHIA
PHILA
PHILA.
PHILAD
PHILADALPHIA
PHILADEDLPHIA
PHILADELAPHIA
PHILADELHIA
PHILADELHPIA
PHILADELLPHIA
PHILADELOHIA
PHILADELPH
PHILADELPHA
PHILADELPHAI
PHILADELPHI
PHILADELPHIA
PHILADELPHIA PA
PHILADELPHIA,
PHILADELPHIA, PA
PHILADELPHIA`
PHILADELPHIAP
PHILADELPHIAPHIA
PHILADELPHILA
PHILADELPHIOA
PHILADELPIA
PHILADELPOHIA
PHILADELPPHIA
PHILADEPHA
PHILADEPHIA
PHILADEPHILA
PHILADEPLHIA
PHILADERLPHIA
PHILADLELPHIA
PHILADLEPHIA
PHILADLPHIA
PHILADPHIA
PHILADRLPHIA
PHILAELPHIA
PHILDADELPHIA
PHILDADLPHIA
PHILDAELPHIA
PHILDELPHIA
PHILDEPPHIA
PHILIADELPHIA
PHILIDELPHIA
PHILLA
PHILLADELPHIA
PHILLY
PHILOADELPHIA
PHLADELPHIA
PHOLADELPHIA
PHPILADELPHIA
PIHLADELPHIA

Data Science Process



O

Obtain

Gather Data from relevant sources

P

Prepare

Clean data to formats that machine understands

E

Explore

Find significant patterns and trends using statistical methods

M

Model

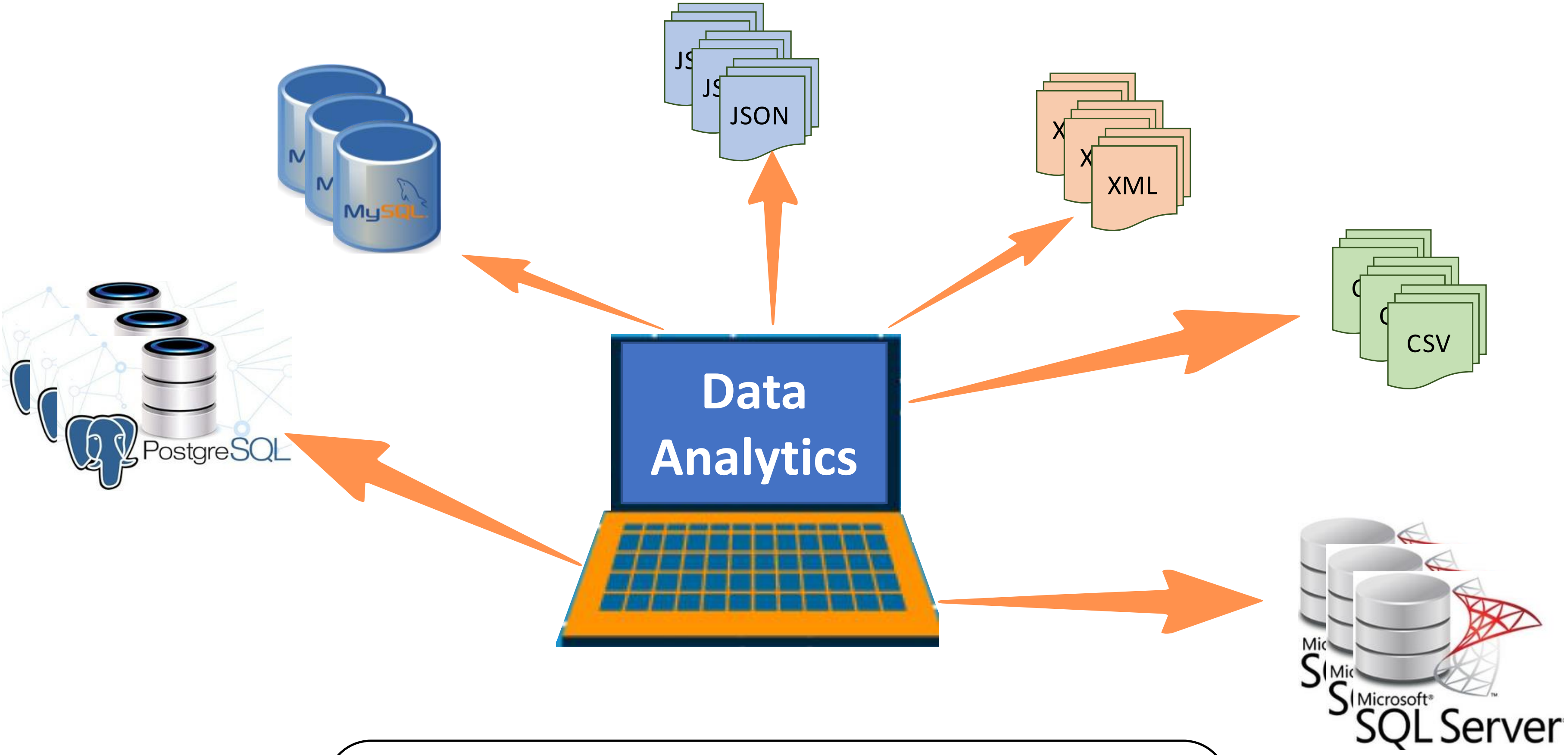
Construct models to predict and forecast

I

Interpret

Put the results into good use

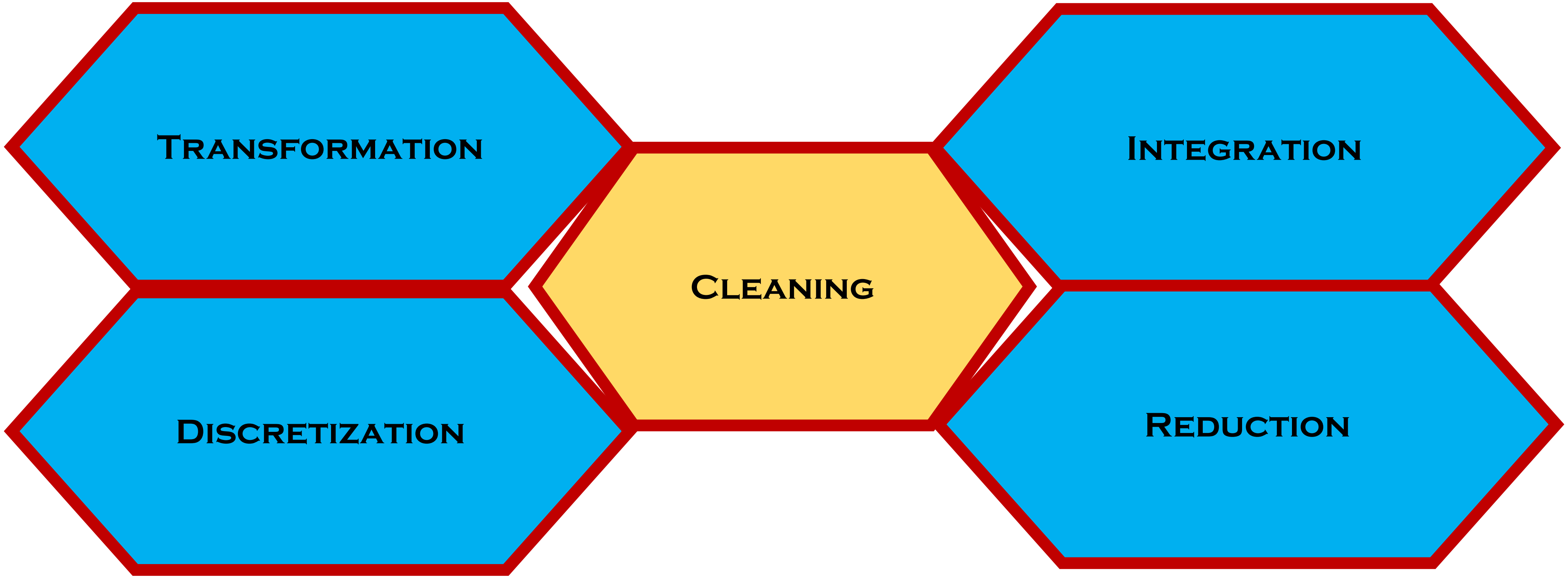
Data Extraction (Obtain or Gather)



- It is recommended to install **PostgreSQL**
- In this course, data extraction will be done using **SQL** queries and **Python** code.

- Data is stored in different formats
- First step is to extract the data required for the analysis.

Data Preparation



Transformation

Integration

Cleaning

Discretization

Reduction

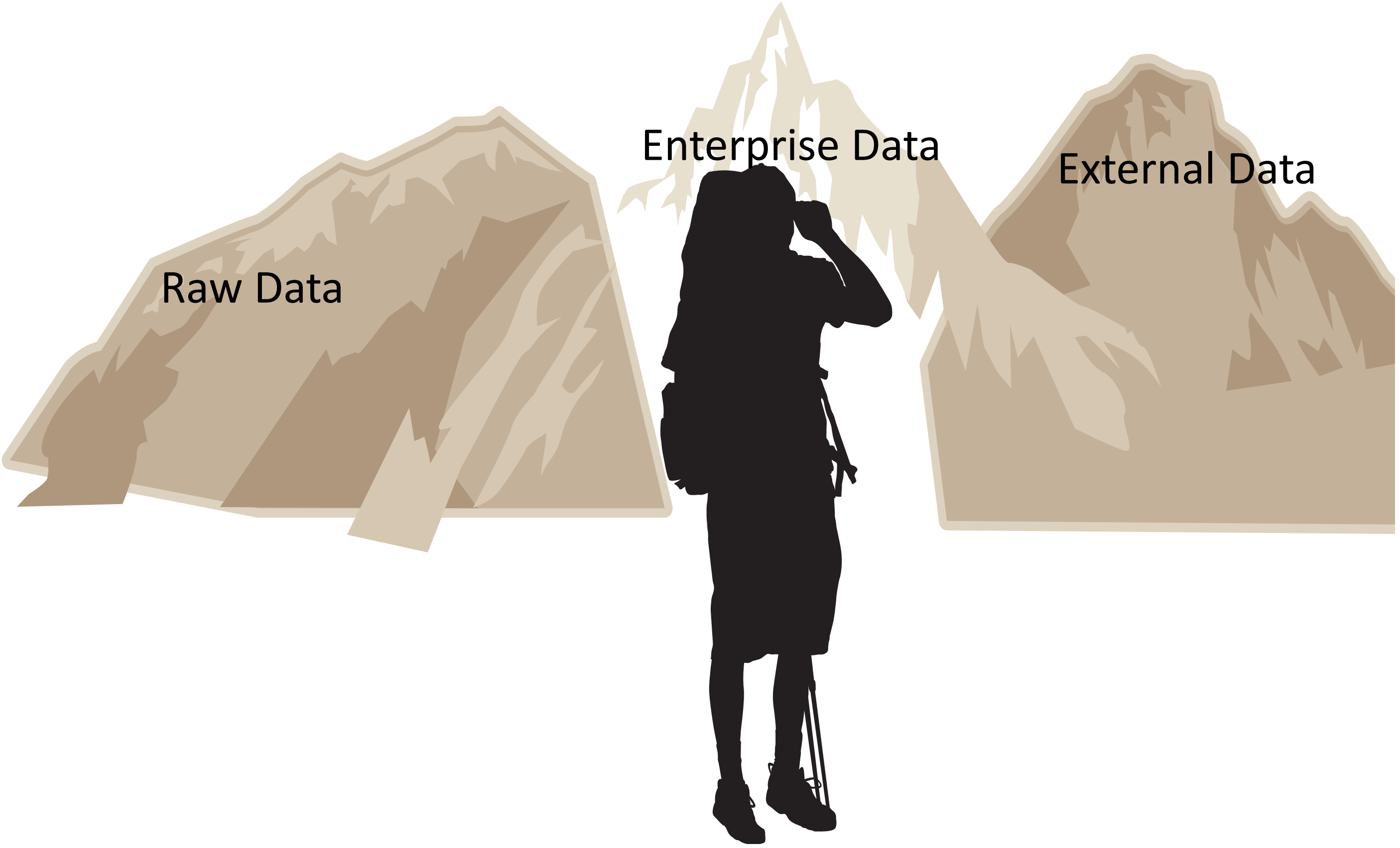
Data Exploration

Visualize and understand all available data

Establish connections to access the data

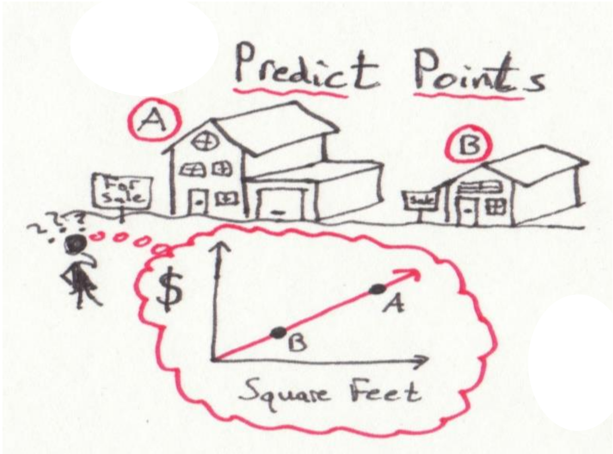
Use appropriate tools to separate the most useful content

Discover hidden insights



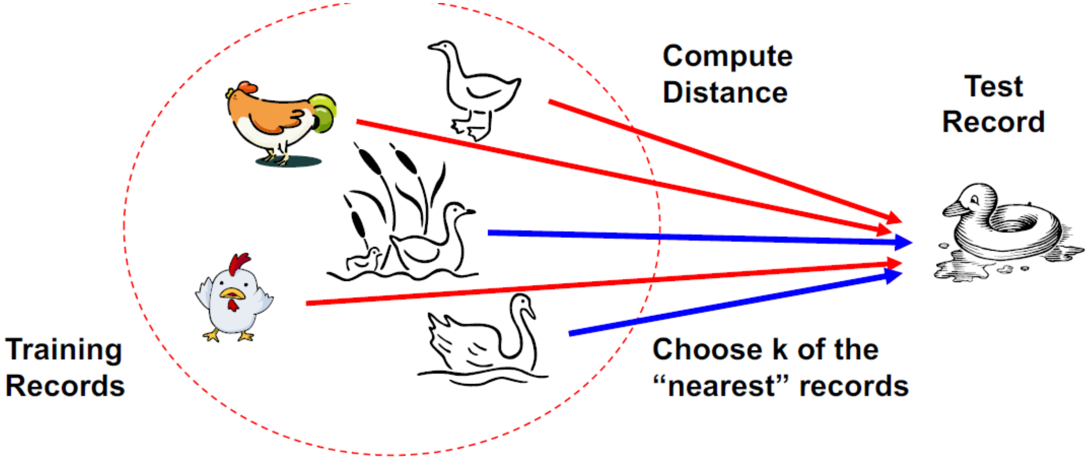
Modeling

- Wide variety of data models
  - Which model better suits your data?
  - What is the problem that you are trying to solve?
- We can use the libraries of Python or R to implement different data analytics models



Predict unknown values based on a set of known values

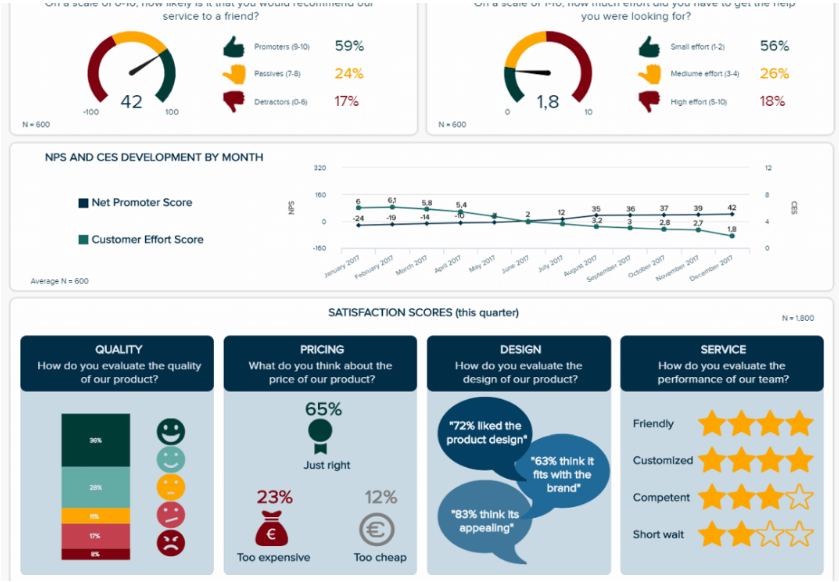
Group similar items together



Classify an item: if it walks like a duck, quacks like a duck, then it is probably a duck

Interpretation

- A good interpretation of the data and the results can help in:
  - Decision-making
  - Anticipating needs
    - Shazam (music identification application)
  - Cost efficiency: cost-reduction opportunities
    - Intel reduces the 19000 test and saved 3Million USD
  - Clear foresight
    - Companies know about their performance from data they collect



Thank you