DATA ANALYSIS WITH PYTHON

TUTORIAL

Introduction to Data Analysis

INTRODUCTION TO DATA ANALYSIS – TUTORIAL

1. What is Data Analysis
2. Example Data Analysis with Python
3. How to use Jupyter Notebooks
4. Intro to Numpy
5. Intro to Pandas
6. Data Cleaning
7. Reading Data SQL, CSVs, APIs, etc.
8. Python

- What is Data Analysis

" A process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making"

WHY PYTHON FOR DATA ANALYSIS?

- Very simple and intuitive to learn
- "Correct" language
- Powerful libraries (not just for Data Analysis)
- Free and open source
- Amazing community, docs and conferences

When to choose R?

- When R Studio is needed
- When dealing with advanced statistical methods
- When extreme performance is needed

THE DATA ANALYSIS PROCESS

- SQL
- Scrapping
- File Formats
  - CSV
  - JSON
  - XML
- Consulting APIs
- Buying Data
- Distributed Databases

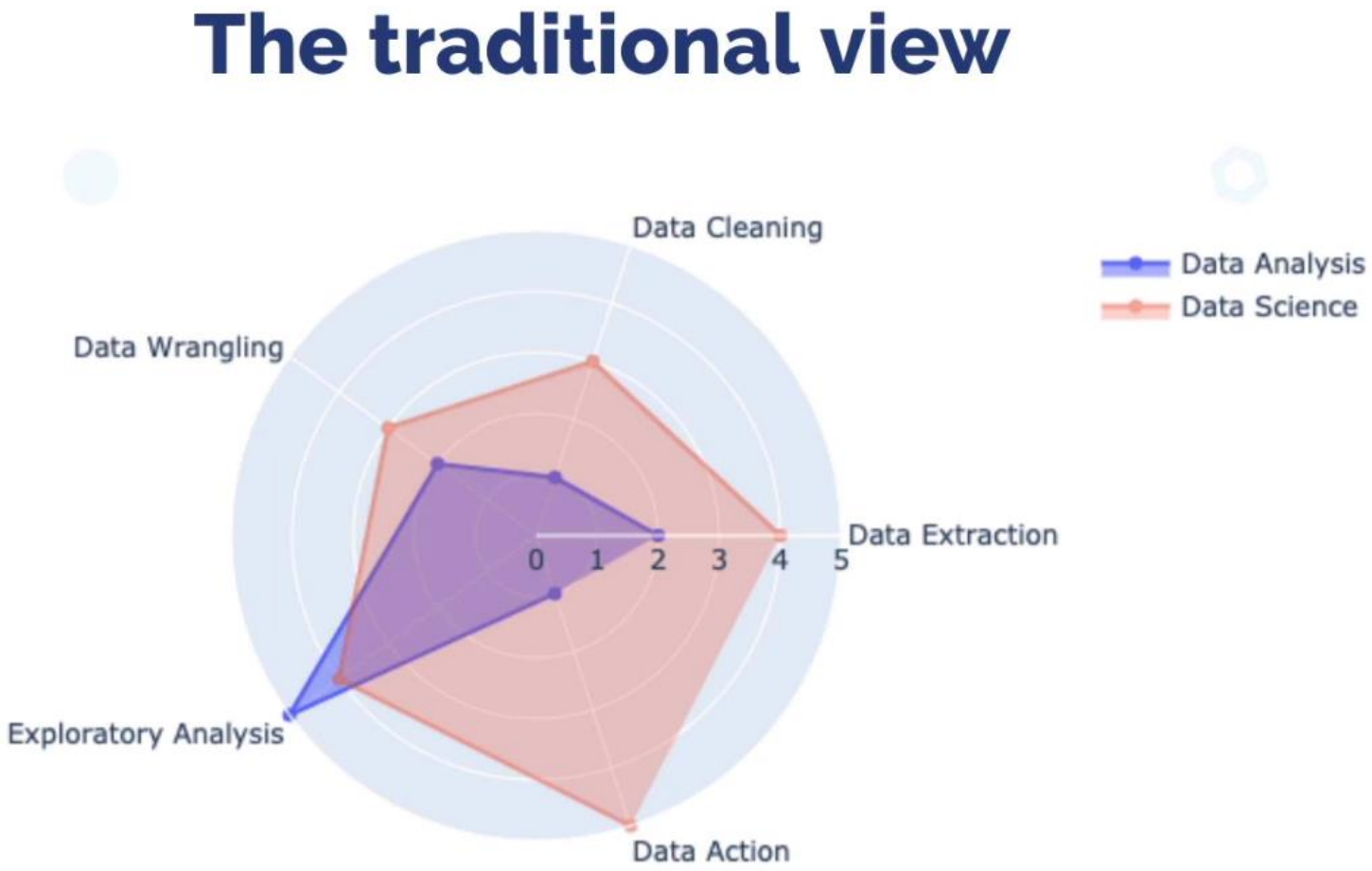
- Missing values and empty data
- Data imputation
- Incorrect types
- Incorrect or invalid values
- Outliers and non relevant data
- Statistical sanitization

- Hierarchical Data
- Handling categorical data
- Reshaping and transforming structures
- Indexing data for quick access
- Merging, combining and joining data

- Exploration
- Building statistical models
- Visualization and representations
- Correlation vs Causation analysis
- Hypothesis testing
- Statistical analysis
- Reporting

- Building Machine Learning Models
- Feature Engineering
- Moving ML into production
- Building ETL pipelines
- Live dashboard and reporting
- Decision making and real-life tests

DATA ANALYSIS VS DATA SCIENCE



PYTHON & PYTHON DATA ECOSYSTEM

- pandas: The cornerstone of our Data Analysis job with Python
- matplotlib: The foundational library for visualizations. Other libraries we'll use will be built on top of matplotlib.
- numpy: The numeric library that serves as the foundation of all calculations in Python.
- seaborn: A statistical visualization tool built on top of matplotlib.
- statsmodels: A library with many advanced statistical functions.
- scipy: Advanced scientific computing, including functions for optimization, linear algebra, image processing and much more.
- scikit-learn: The most popular machine learning library for Python (not deep learning)

THANKS