LECTURE 1

Course Overview

An overview of data science and the data science lifecycle

Data Science, Spring 2024 @ Knowledge Stream

Sana Jabbar

Intro – Sana Jabbar

- BS Telecommunication Engineering @ FAST, Islamabad
- MS Electrical Engineering @ FAST, Lahore
- Since 2018 @ Lums, Research Associate in Computer Vision and Graphics Lab, formerly in Clinical and Translational Lab
- Background: PhD Remote Sensing,
- Research interests
  - Interactive computational tools for earth observation and medicine.
  - Applications in Remote Sensing for Taxation Automation and in medical Diagnosis.

What is Data Science?

Lecture 01

- Intros
- **What is data science?**
- The objective of this course?
- Course Overview
- Data Science Lifecycle

What Is Data Science in this course?

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

What is Data Science?

Definition:

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data.

Joey Gonzalez

Data Science is the application of data centric, computational, and inferential thinking to:

- Understand the world (science).
- Solve problems (engineering).

What is Data Science?

Definition:

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data.

Interdisciplinary Nature:

- Data science combines elements of computer science, mathematics, and domain expertise to solve complex problems.

Structured Data:

- Refers to data that is organized into a specific format, typically consisting of rows and columns, where each piece of information is stored in a well-defined manner.

Unstructured Data:

- Refers to data that lacks a specific, predefined structure, making it more challenging to organize and analyse compared to structured data.

- Intros
- What is data science?
- **The objective of this course?**
- Course Overview
- Data Science Lifecycle

Objective

Lecture 01

Course Objective

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Course Objective

Course Objectives:

- Equip students with the **skills** needed to extract valuable insights from data.

Course Objective

Course Objectives:

- Equip students with the **skills** needed to extract valuable insights from data.

Data Collection

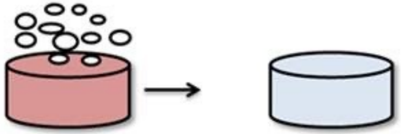
Course Objective

Course Objectives:

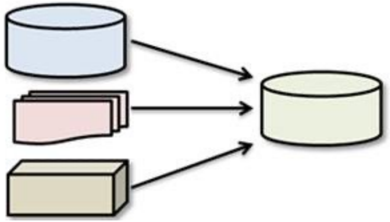
- Equip students with the **skills** needed to extract valuable insights from data.

Data Collection

Data Cleaning



Data Integration



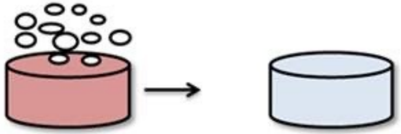
Course Objective

Course Objectives:

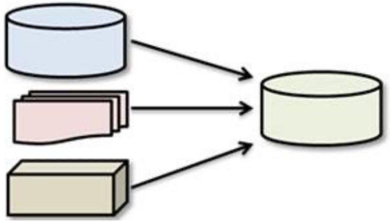
- Equip students with the **skills** needed to extract valuable insights from data.

Data Collection

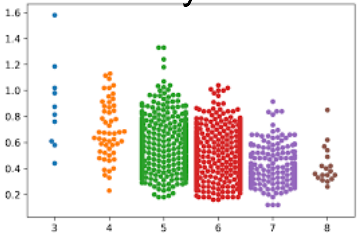
Data Cleaning



Data Integration



Exploratory Data Analysis



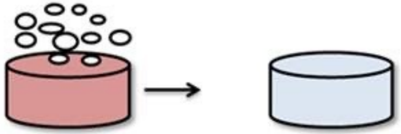
Course Objective

Course Objectives:

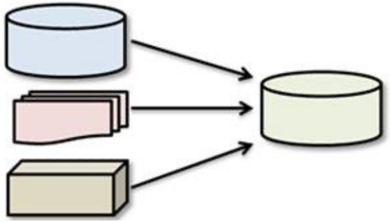
- Equip students with the **skills** needed to extract valuable insights from data.

Data Collection

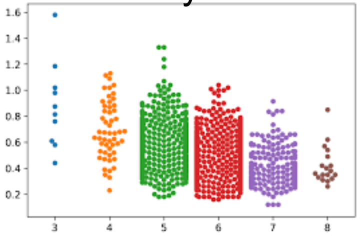
Data Cleaning



Data Integration



Exploratory Data Analysis



Machine Learning

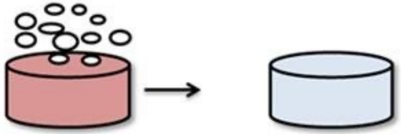
Course Objective

Course Objectives:

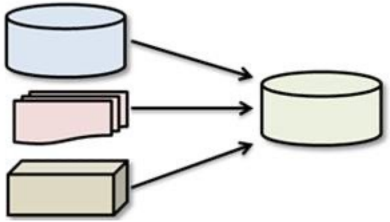
- Equip students with the **skills** needed to extract valuable insights from data.

Data Collection

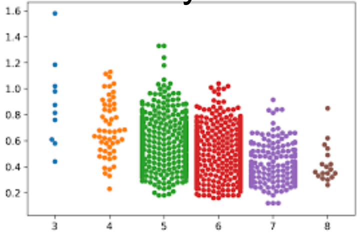
Data Cleaning



Data Integration



Exploratory Data Analysis



Machine Learning

- Prepare students for **real-world** data science challenges.
- Effectively communicate their findings to non-technical stakeholders

Course Objective

Prepare

Prepare students for **data management**, **machine learning**, and **statistics**, by providing the necessary foundation and context.

Enable

Enable students to start careers as data scientists by providing experience working with **real-world data, tools, and techniques**.

Empower

Empower students to apply computational and inferential thinking to address **real-world problems**.

Importance of Data Science

The world is complicated! Decisions are hard.

- Data science drives decision-making across various industries.
- There is a high demand for data scientists in today's job market.
- Data is used everywhere to answer hard questions and make tough decisions:
  - Science
  - Medicine
  - Engineering
  - Sports

Claims about data come up in discussing almost any important issue:

- Instead of ”Alex says," now it’s "the data says."
- It is usually not easy to tell what the data "says"
- **Empower yourself** to participate in the arguments that shape your life and your society

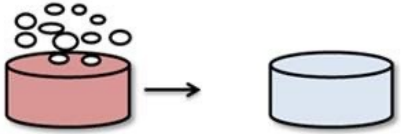
Importance of Data Science

Course Objectives:

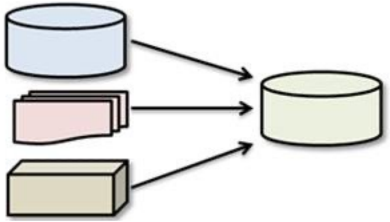
- Equip students with the **skills** needed to extract valuable insights from data.

Data Collection

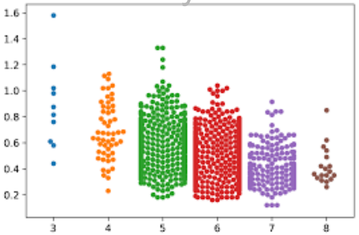
Data Cleaning



Data Integration



Exploratory Data Analysis



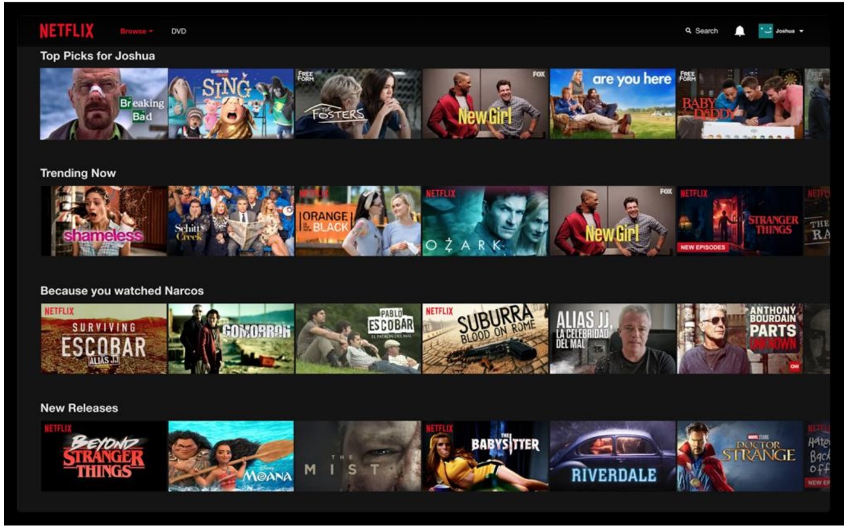
Machine Learning

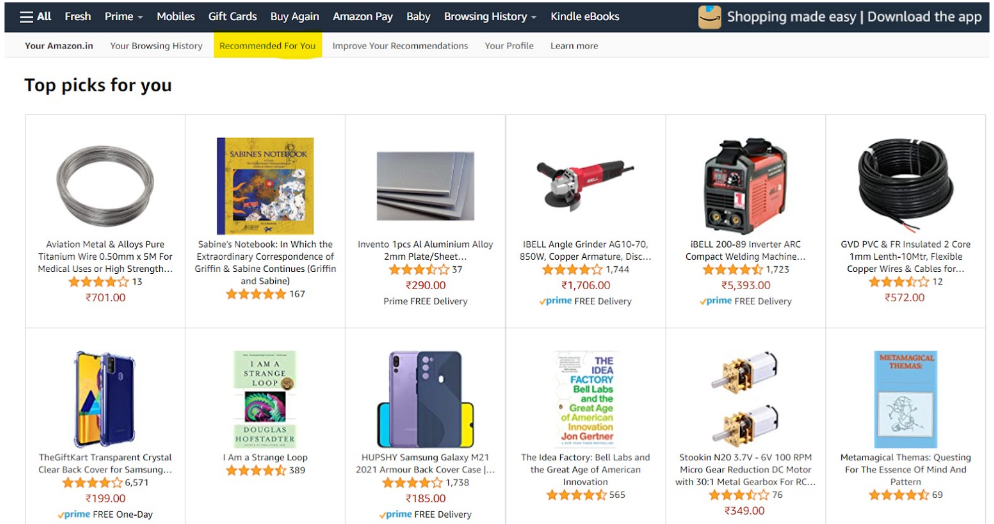
- Prepare them for real-world data science challenges.
- Effectively communicate their findings to non-technical stakeholders

Example:

- Imagine you work for an e-commerce

Recommendation Systems





Top picks for you

Aviation Metal & Alloys Pure Titanium Wire 0.50mm x 5M For Medical Uses or High Strength… 13 ₹701.00

Sabine's Notebook: In Which the Extraordinary Correspondence of Griffin & Sabine Continues (Griffin and Sabine) 167

Invento 1pcs Al Aluminium Alloy 2mm Plate/Sheet… 37 ₹290.00 Prime FREE Delivery

IBELL Angle Grinder AG10-70, 850W, Copper Armature, Disc… 1,744 ₹1,706.00 prime FREE Delivery

iBELL 200-89 Inverter ARC Compact Welding Machine… 1,723 ₹5,393.00 prime FREE Delivery

GVD PVC & FR Insulated 2 Core 1mm Lenth-10Mtr, Flexible Copper Wires & Cables for… 12 ₹572.00

TheGiftKart Transparent Crystal Clear Back Cover for Samsung… 6,571 ₹199.00 prime FREE One-Day

I Am a Strange Loop 389

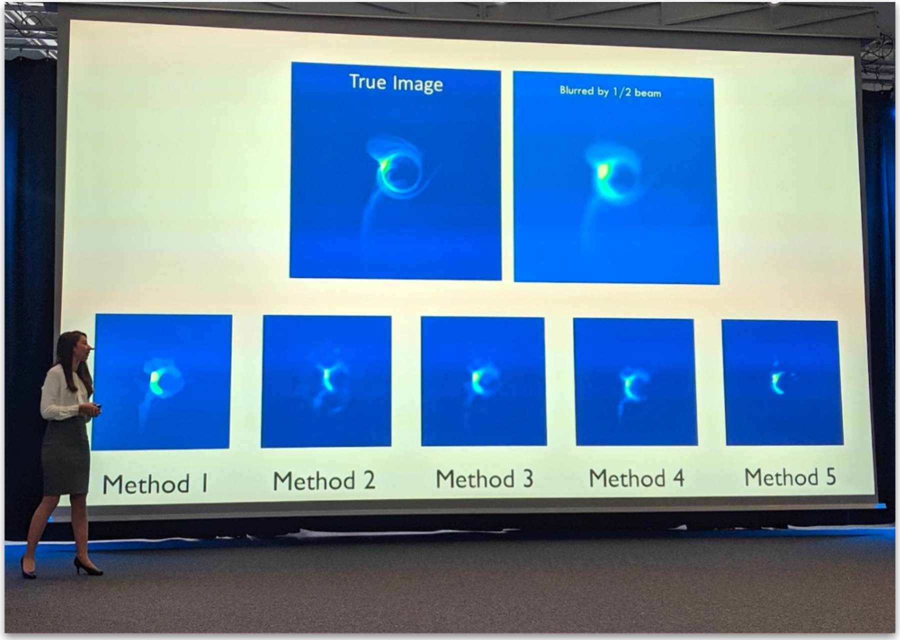
HUPSHY Samsung Galaxy M21 2021 Armour Back Cover Case |… 1,738 ₹185.00 prime FREE Delivery

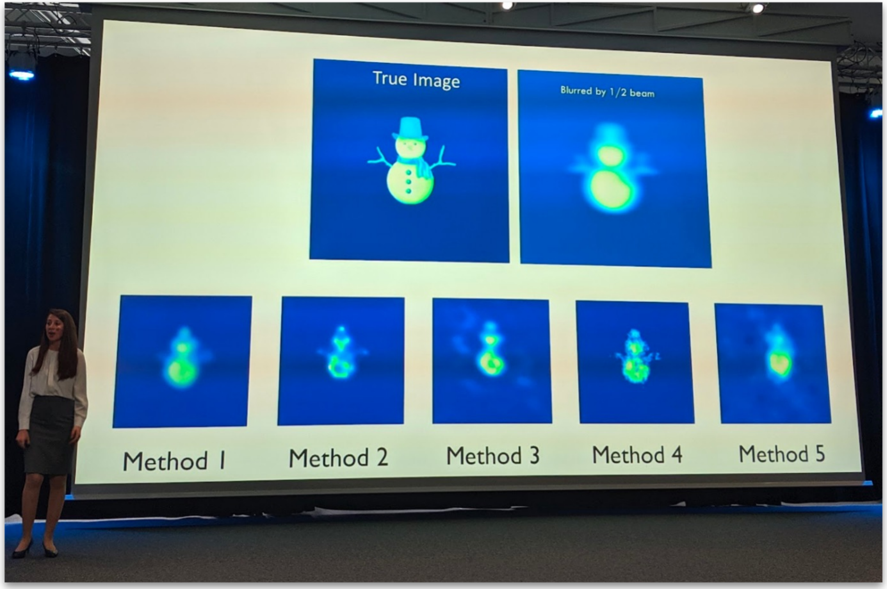
The Idea Factory: Bell Labs and the Great Age of American Innovation 565

Stookin N20 3.7V - 6V 100 RPM Micro Gear Reduction DC Motor with 30:1 Metal Gearbox For RC… 76 ₹349.00

Metamagical Themas: Questing For The Essence Of Mind And Pattern 69

First Image of a Black Hole





Importance of Data Science

Technology Trends

2020s **AI…?**

2010s Data Industry
- Collect and sell information

2000s Internet Industry
- Online retailers and services

1990s Software Industry
- Sold computer software

1980s Hardware Industry
- Sold computers

From Joey Gonzalez.

https://en.wikipedia.org/wiki/AlphaGo_versus_Lee_Sedol

Data!

Knowledge is empowering.

Data science offers **immense potential** to address challenging problems facing society.

The future is in your hands, and I believe:

You will use your knowledge for good.

…I am thrilled to teach Data Science :-)

Importance of Data Science

The world is complicated! Decisions are hard.

Data science is a fundamentally human-centered field that facilitates decision-making by quantitatively balancing tradeoffs.

- To quantify things reliably we must:
  - Find relevant data;
  - Recognize its limitations;
  - Ask the right questions;
  - Make reasonable assumptions;
  - Conduct an appropriate analysis; and
  - Synthesize and explain our insights.

- Apply critical thinking and skepticism at every step

- Consider how our decisions affect others.

Importance of Data Science

The world is complicated! Decisions are hard.

Data science is a fundamentally human-centered field that facilitates decision-making by quantitatively balancing tradeoffs.

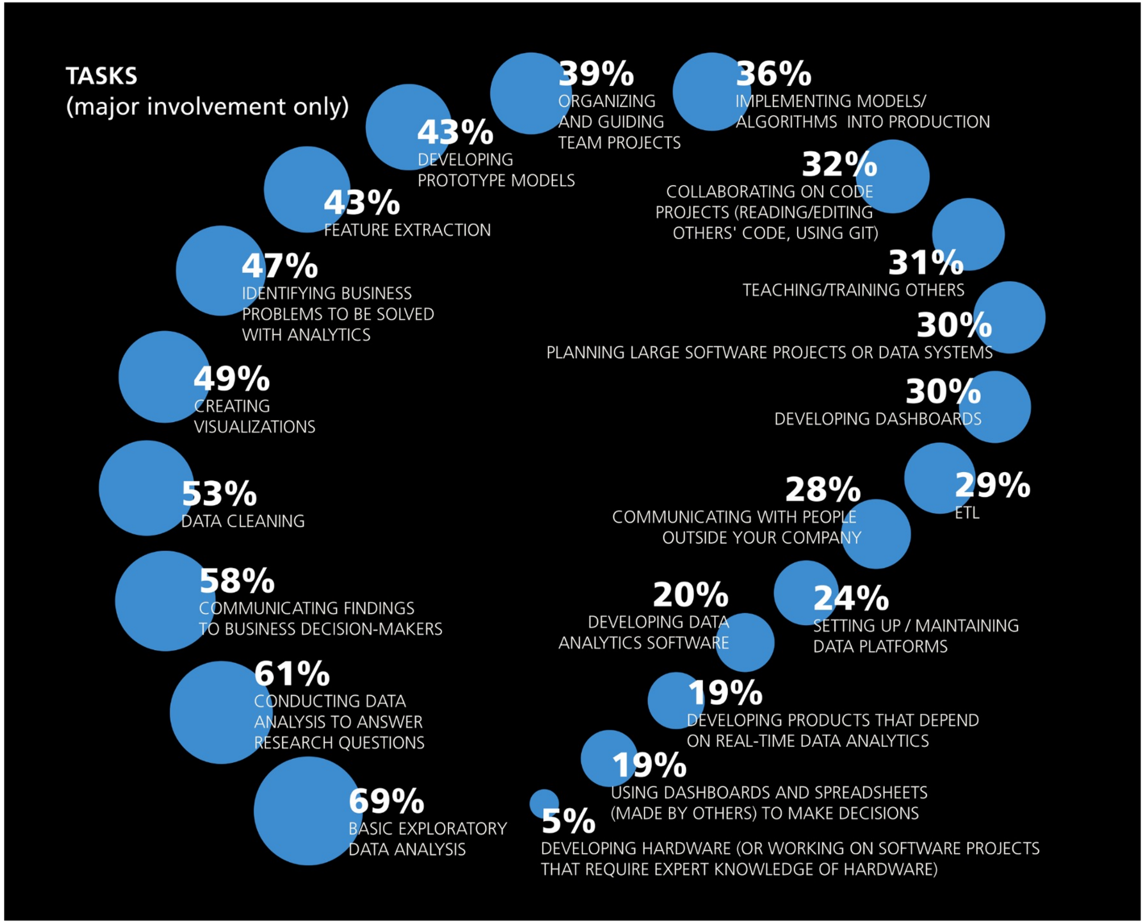
- To quantify things reliably we must:
  - Find relevant data;
  - Recognize its limitations;
  - Ask the right questions;
  - Make reasonable assumptions;
  - Conduct an appropriate analysis; and
  - Synthesize and explain our insights.

- Apply critical thinking and skepticism at every step

- Consider how our decisions affect others.

After this course, you should be able to take data and produce useful insights on the world's most challenging and ambiguous problems.

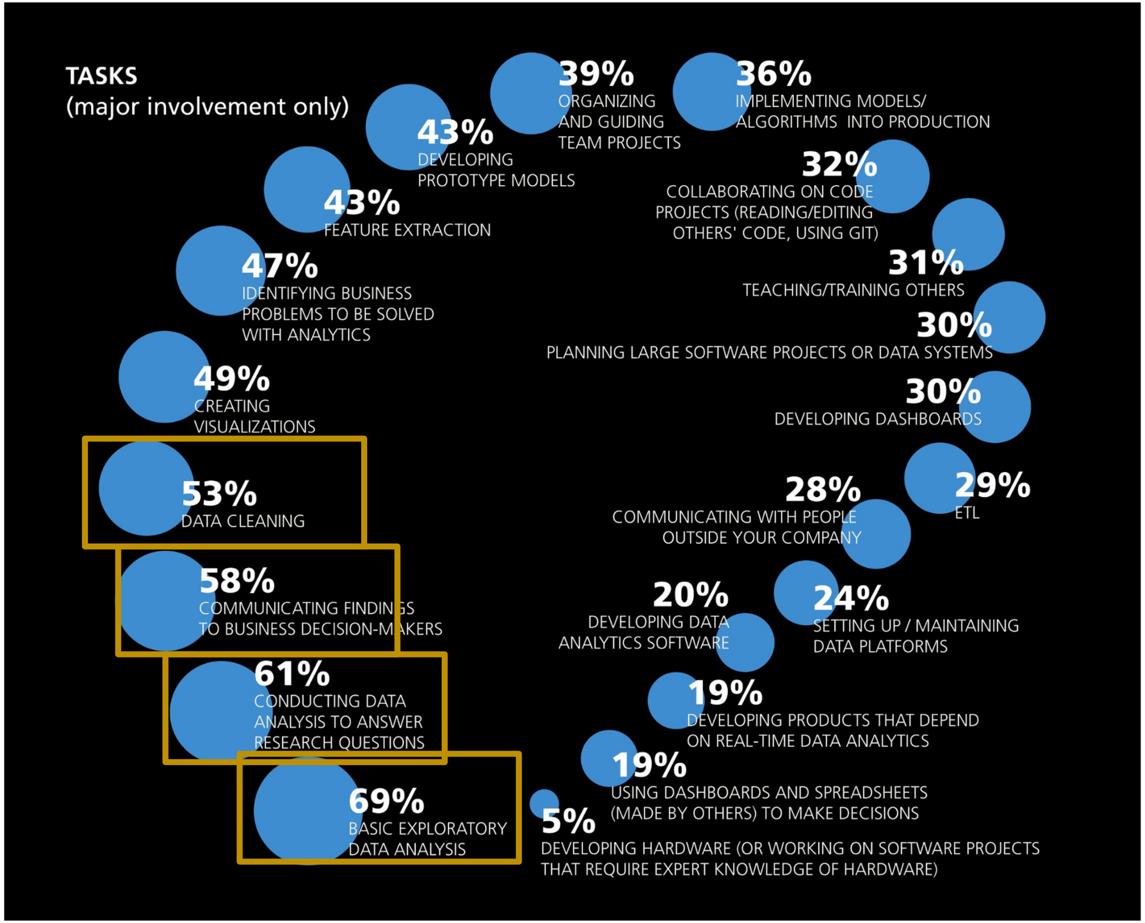
Importance of Data Science



The major tasks that data scientists say they work on regularly.

Based on the results of the 2016 Data Science Salary Survey.

Importance of Data Science



The major tasks that data scientists say they work on regularly.

Based on the results of the 2016 Data Science Salary Survey.

Data Science Requires Engineering and Scientific Insight

Good data analysis is not:

- The simple application of a statistics recipe.
- Simple application of statistical software.

There are many **tools** out there for data science, but they are merely tools.

- **They don't do any of the important thinking!**

"The purpose of computing is insight, not numbers."

R. Hamming. *Numerical Methods for Scientists and Engineers (1962).*

Example Questions in Data Science

Some (broad) questions we might try to answer with data science:

- What show should we recommend to our users to watch?
- In which markets should we focus our advertising campaign?
- Should I send my kids to daycare?
- Is the world getting better or worse?
- What areas of the world are at higher risk for climate change impact in 10 years? 20?
- What should we eat to avoid dying early of heart disease?
- Do immigrants from poor countries have a positive or negative impact on the economy?
- Which university will be the most appropriate for Data science engineering?

- Intros
- What is data science?
- The objective of this course?
- **Course Overview**
- Data Science Lifecycle

Course Overview

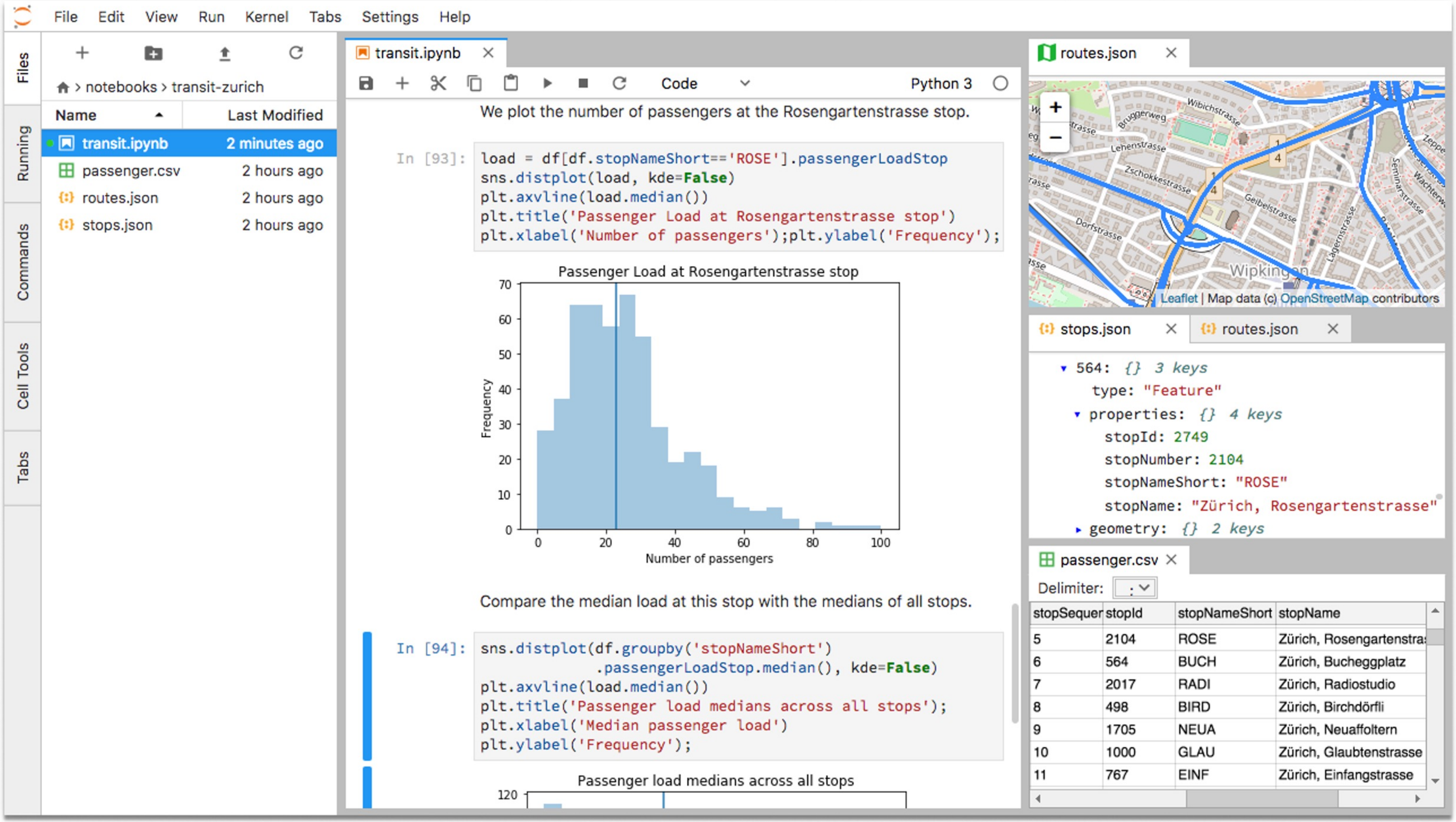
Lecture 01

Tentative List of Topics to be Covered in Data Science

- Pandas and NumPy
- Relational Databases & SQL
- Exploratory Data Analysis
- Regular Expressions
- Visualization
  - matplotlib
  - Seaborn
  - plotly
- Sampling

- Model design and loss formulation
- Linear Regression
- Feature Engineering
- Regularization, Bias-Variance Tradeoff, Cross-Validation
- Gradient Descent
- Data science in the physical world
- Logistic Regression
- Clustering
- PCA

Programming Environment for our Course: Jupyter Notebook



Name	Last Modified
transit.ipynb	2 minutes ago
passenger.csv	2 hours ago
routes.json	2 hours ago
stops.json	2 hours ago

We plot the number of passengers at the Rosengartenstrasse stop.

```
In [93]: load = df[df.stopNameShort=='ROSE'].passengerLoadStop
         sns.distplot(load, kde=False)
         plt.axvline(load.median())
         plt.title('Passenger Load at Rosengartenstrasse stop')
         plt.xlabel('Number of passengers');plt.ylabel('Frequency');
```

Compare the median load at this stop with the medians of all stops.

```
In [94]: sns.distplot(df.groupby('stopNameShort')
                         .passengerLoadStop.median(), kde=False)
         plt.axvline(load.median())
         plt.title('Passenger load medians across all stops');
         plt.xlabel('Median passenger load')
         plt.ylabel('Frequency');
```

```
▾ 564: {} 3 keys
    type: "Feature"
  ▾ properties: {} 4 keys
      stopId: 2749
      stopNumber: 2104
      stopNameShort: "ROSE"
      stopName: "Zürich, Rosengartenstrasse"
  ▸ geometry: {} 2 keys
```

Delimiter: ;

stopSequer	stopId	stopNameShort	stopName
5	2104	ROSE	Zürich, Rosengartenstra:
6	564	BUCH	Zürich, Bucheggplatz
7	2017	RADI	Zürich, Radiostudio
8	498	BIRD	Zürich, Birchdörfli
9	1705	NEUA	Zürich, Neuaffoltern
10	1000	GLAU	Zürich, Glaubtenstrasse
11	767	EINF	Zürich, Einfangstrasse

Data Science Lifecycle

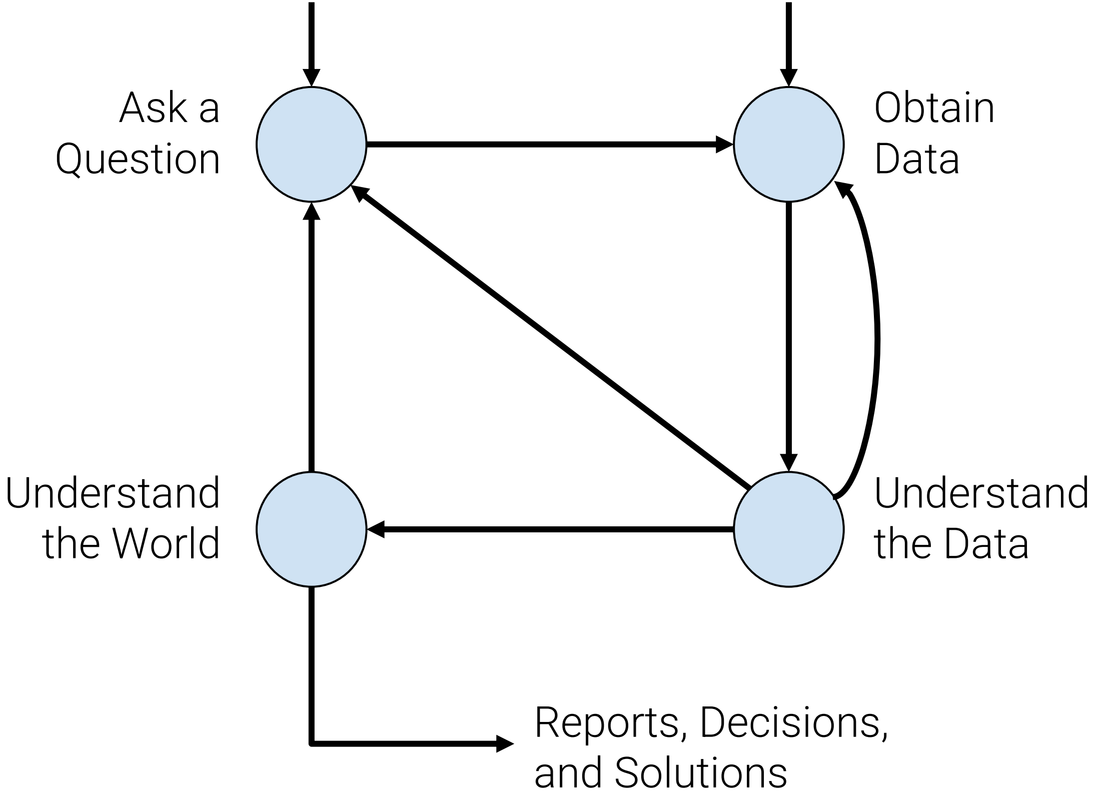
Lecture 01

- Intros
- What is data science?
- The objective of this course?
- Course Overview
- **Data Science Lifecycle**

Data Science Lifecycle

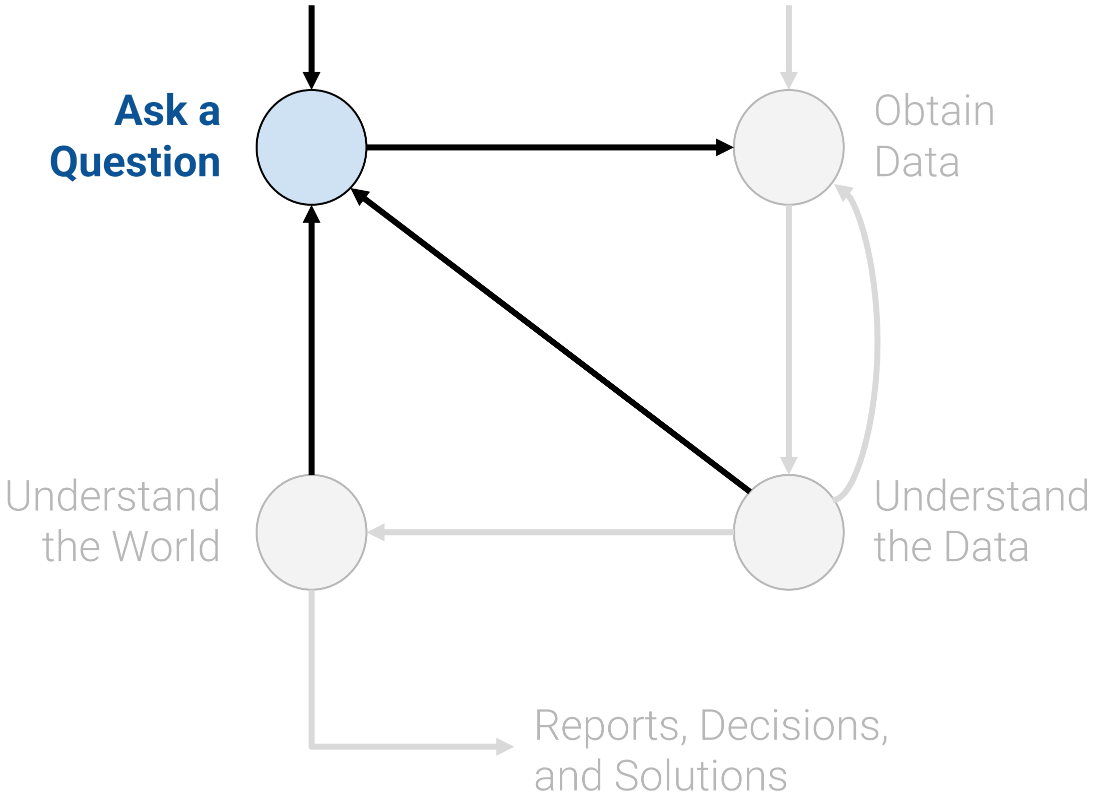
The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!



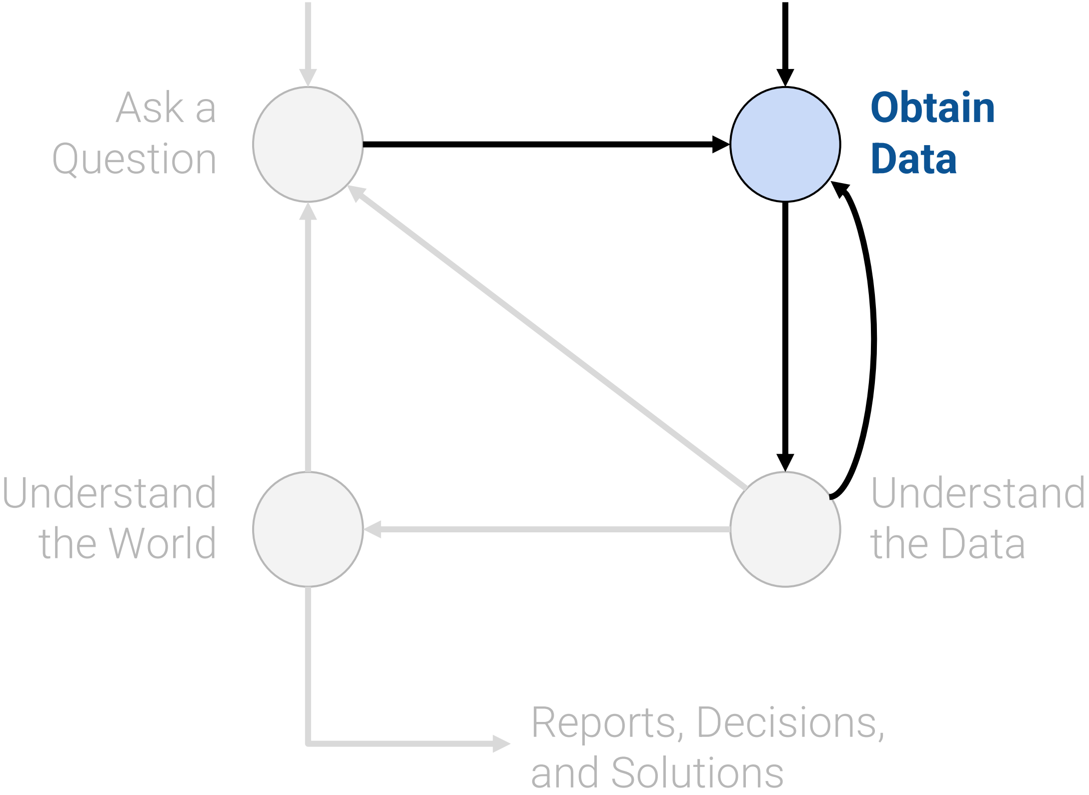
1. Question/Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What hypotheses do we want to test?
- What are our metrics for success?



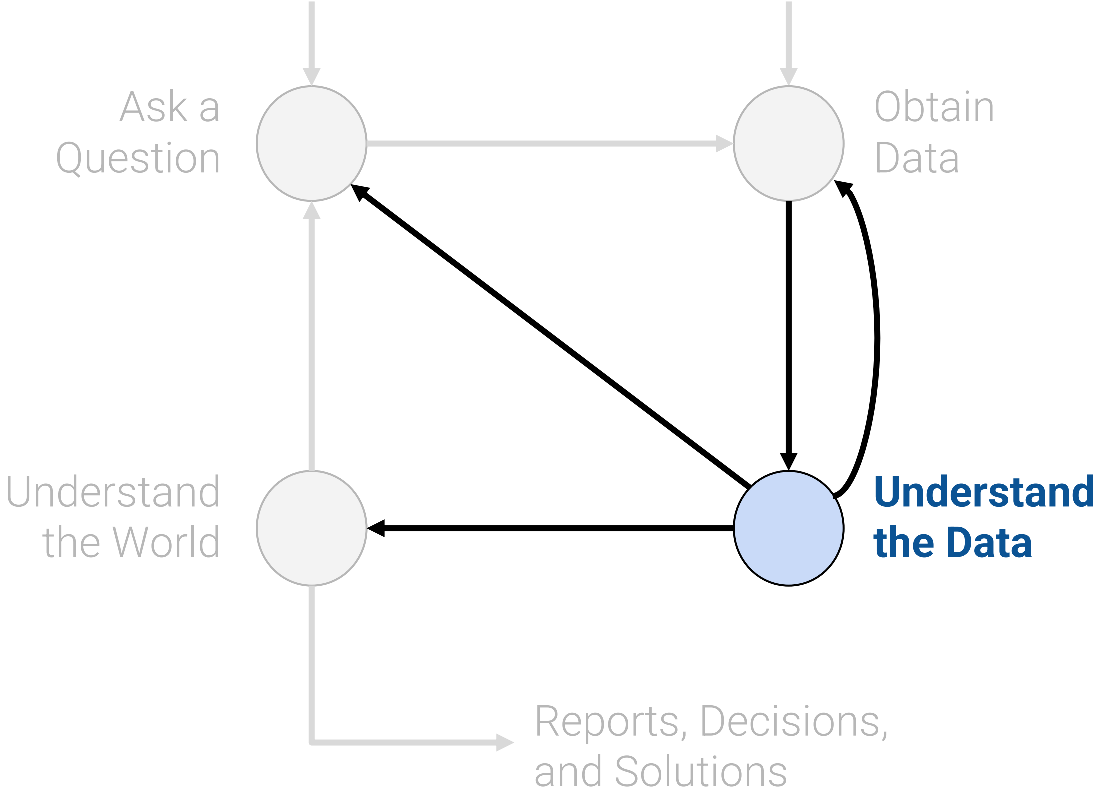
2. Data Acquisition and Cleaning

- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



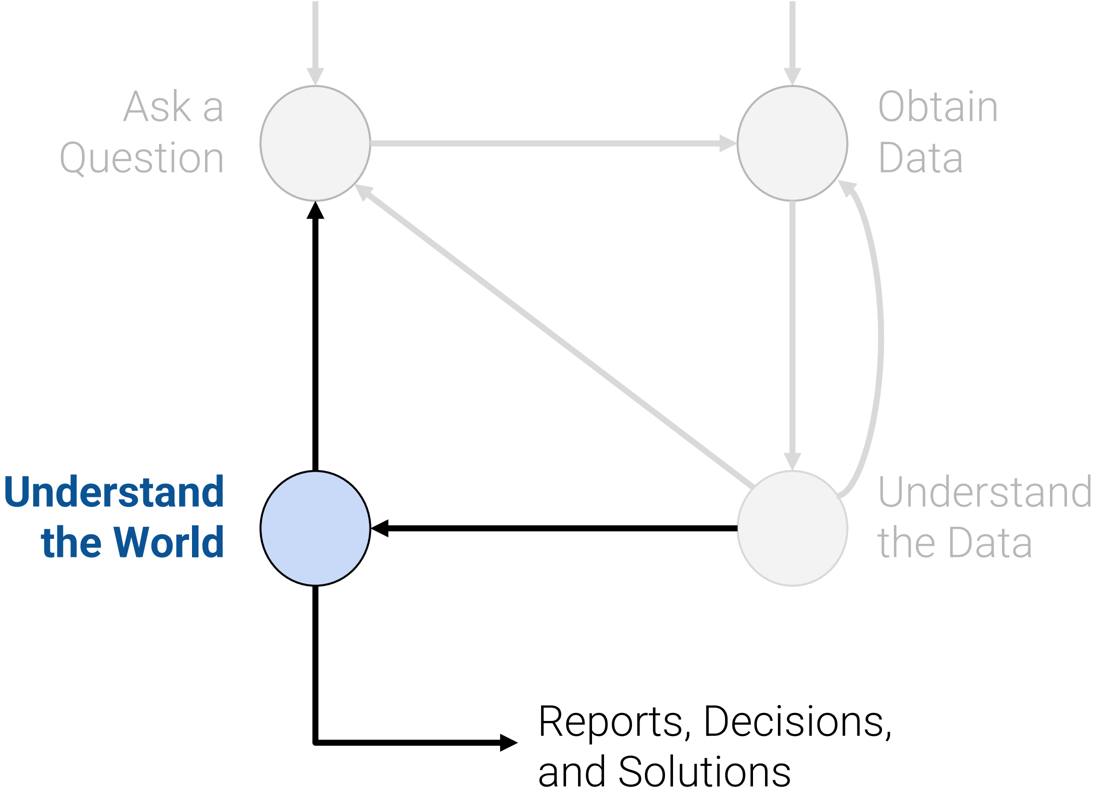
3. Exploratory Data Analysis & Visualization

- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



4. Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



Setup Framework!