



انجمن ریاضی ایران



دانشگاه شهید بهشتی تهران

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001

Data Science Workshop

Amin Anjomshoaa

1986 1987 1988 1989

2002 2003 2004 2005

1983 1984 1985

1982

1981

1980

1979

1978

1977

1976

1975

1974

1973

1972

1971

1970

1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985

1986 1987 1988 1989

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001

2002 2003 2004 2005

2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

2021

2021

2021

2021

2021

2021

2021

2021

2021

Workshop Organization

- Session 1: Data Science foundation
- Session 2: Statistical analysis and modeling methods
- Session 3: Machine learning in practice
- Session 4: Discussion panel

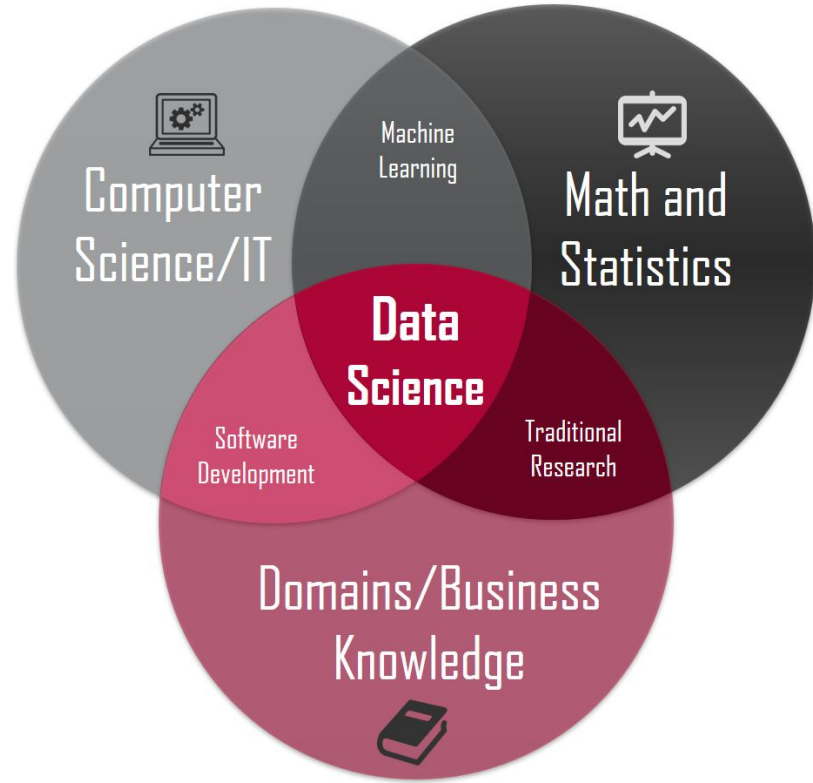
Learning objectives

- Data Science basics
- Big data, data dimension, data integration
- IoT, sensor data streams
- Data Science tools and useful libraries
- Data processing
 - Data Extraction, Transformation, and Loading
 - Exploratory Data Analysis
 - Big Data Processing Frameworks
 - Data Visualization

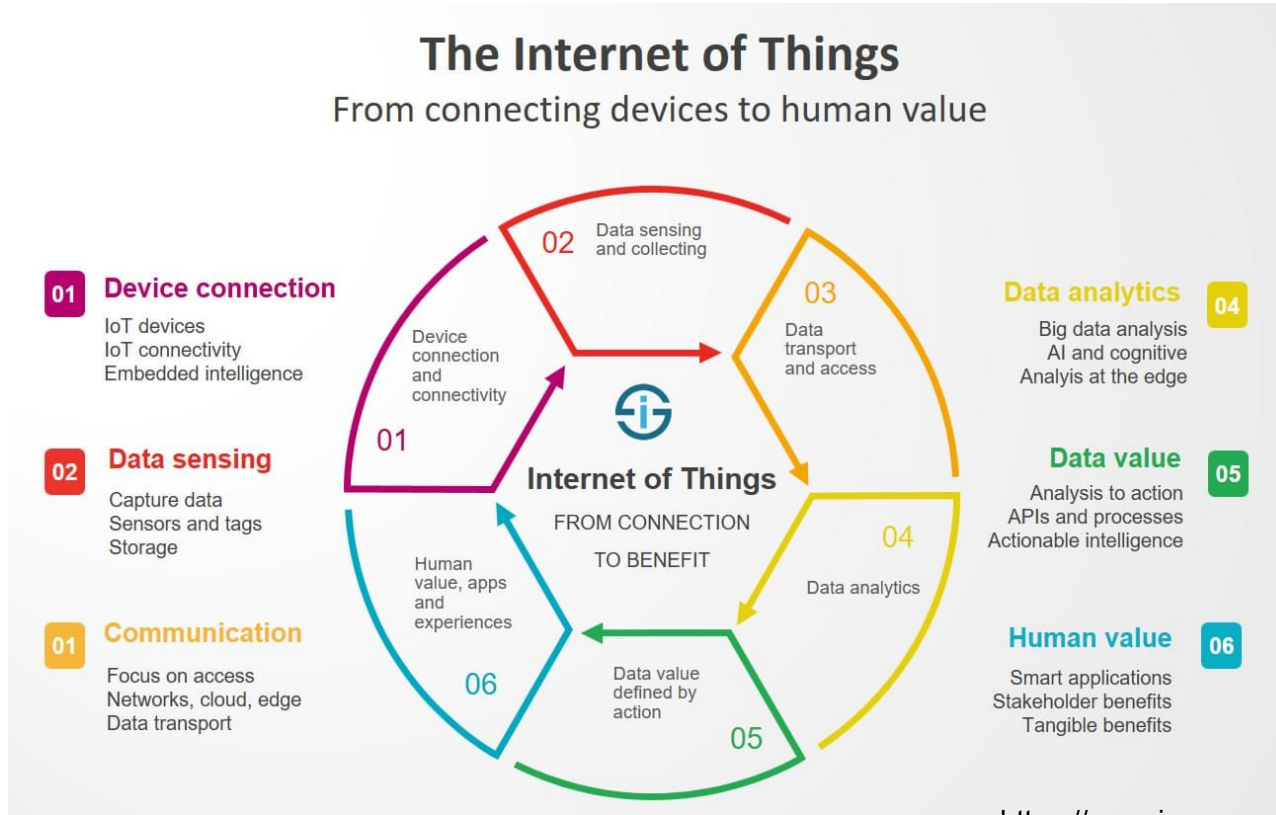
What is Data Science?

Data Science

Data Science is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insight from data.

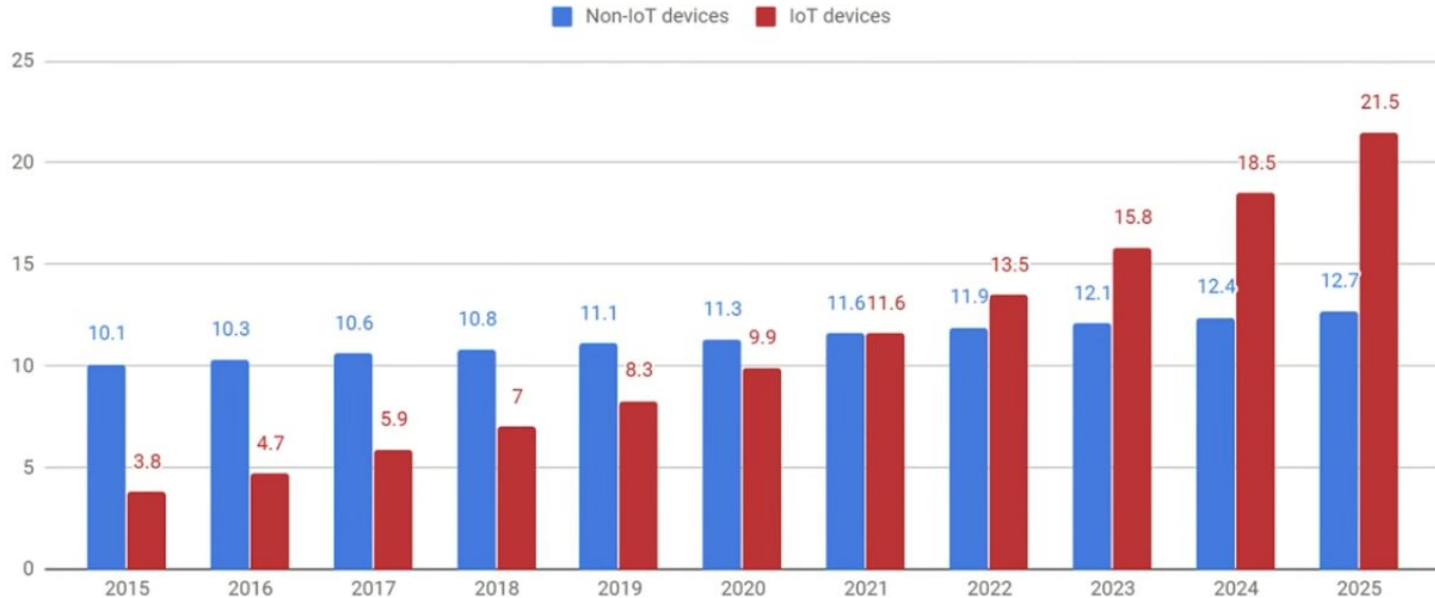


Data Science & Internet of Things (IoT)

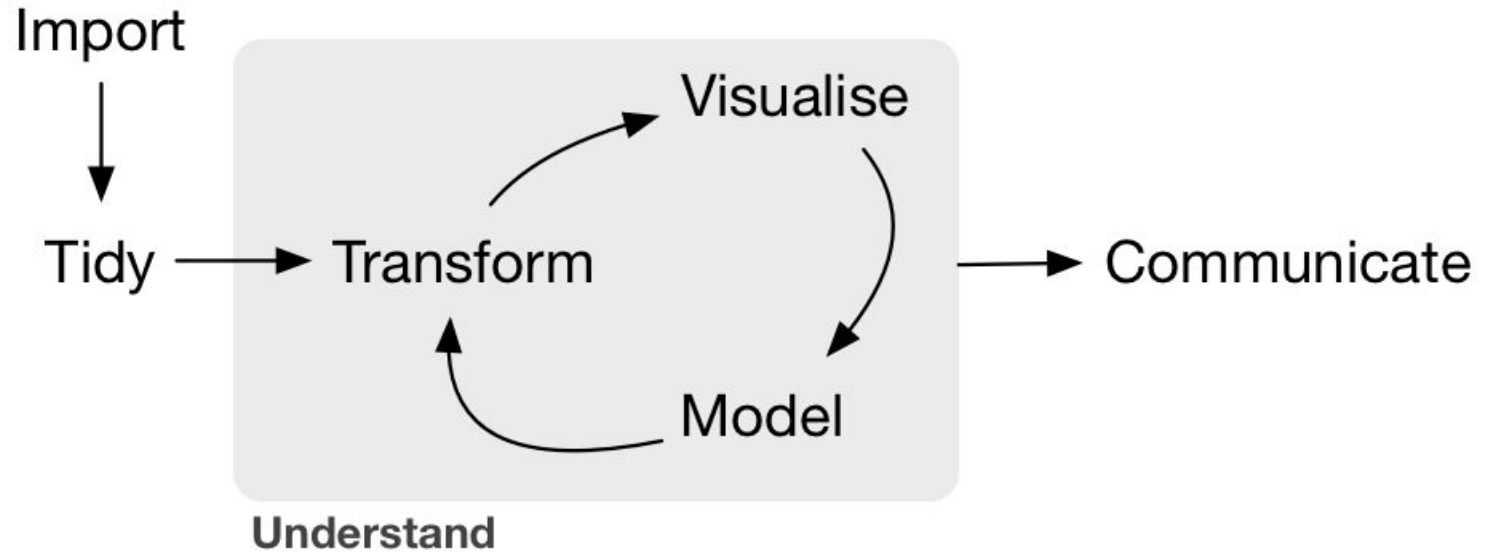


Smart Devices and IoT represents the biggest growth potential

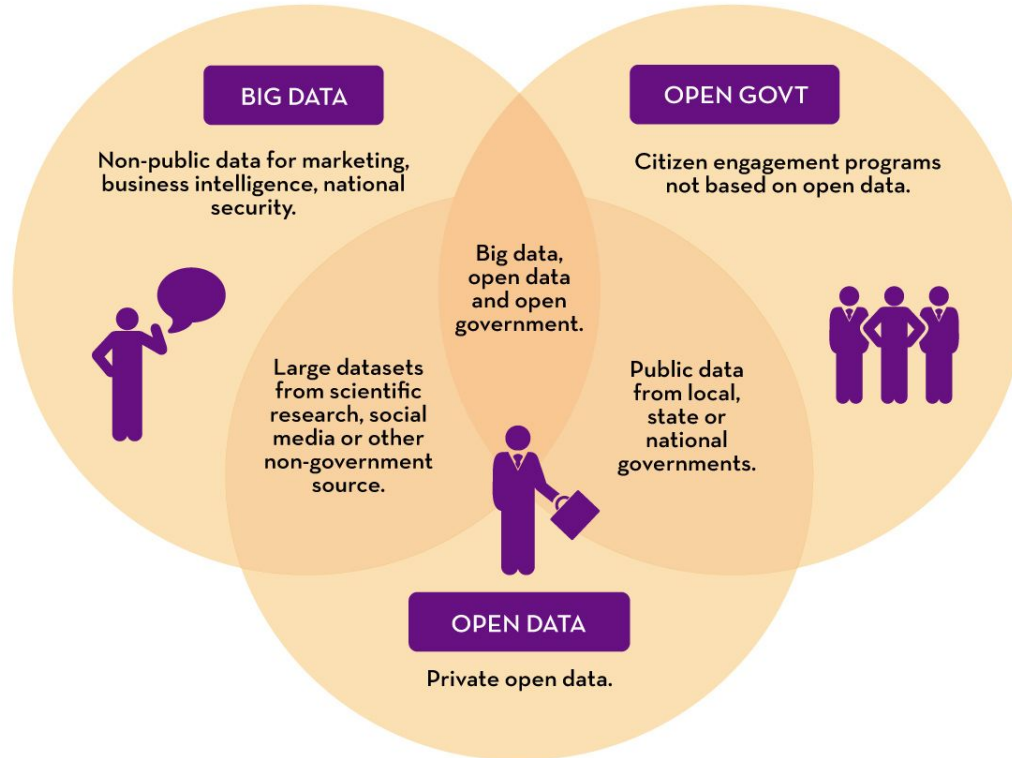
Number of global active connected devices (billions)



Simplified Data Science Process

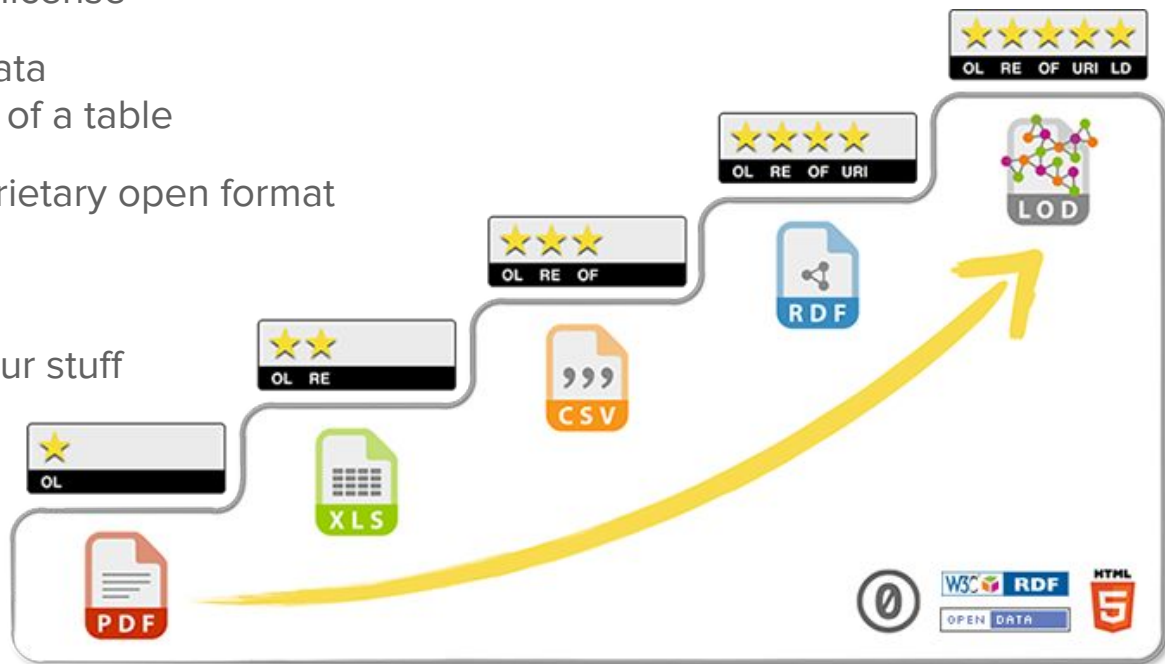


Big Data, Open Data and Open Government



Five-star Open Data

- ★ make your stuff available on the Web
(whatever format) under an open license
- ★★ make it available as structured data
e.g., Excel instead of image scan of a table
- ★★★ make it available in a non-proprietary open format
e.g., CSV instead of Excel
- ★★★★ use URIs to denote things,
so that people can point at your stuff
- ★★★★★ link your data to other data
to provide context



Data Science Foundation

Tensors

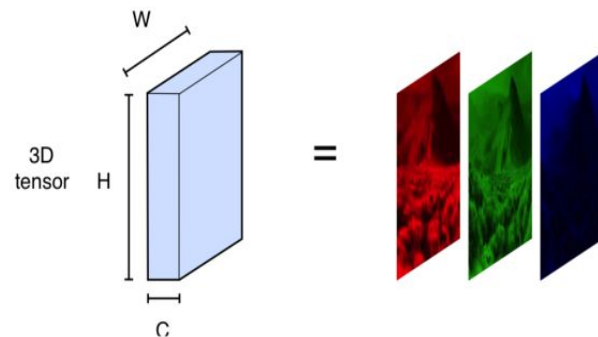
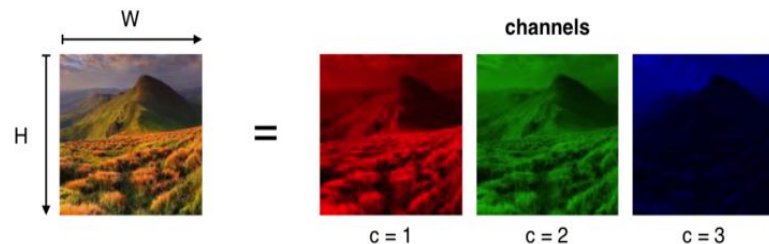
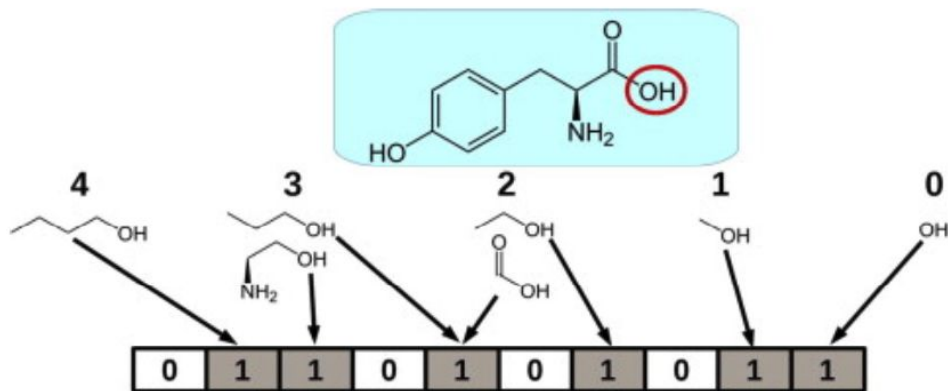
- A tensor is a n-dimensional array with $n \geq 2$
- Scalars are single constant values drawn from the real numbers (0-tensor)
- Vector is a 1-tensor (single dimension)
- A matrix is a tensor of rank 2

Scalar	Vector	Matrix	Tensor
1	$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$	$\begin{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} & \begin{bmatrix} 3 & 2 \end{bmatrix} \\ \begin{bmatrix} 1 & 7 \end{bmatrix} & \begin{bmatrix} 5 & 4 \end{bmatrix} \end{bmatrix}$

Featurization

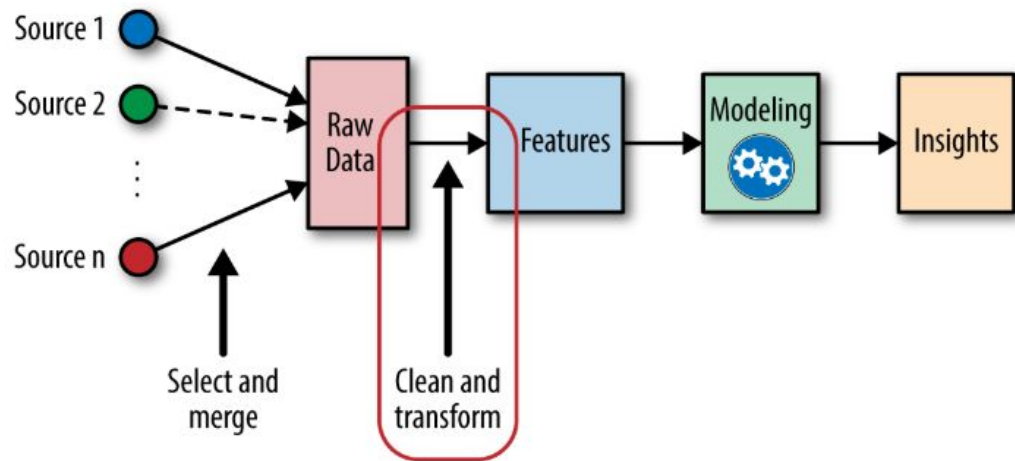
Featurization is a representation of a real-world entity as tensor

$\begin{pmatrix} \text{height} \\ \text{weight} \\ \text{color} \end{pmatrix}$



Features

- Features are individual independent variables that act as the input of prediction models to make predictions.
- New features can be built based on existing features (feature engineering)
- For instance one column of a data set (aka attributes) could represent a feature.
- Number of features is called dimension.



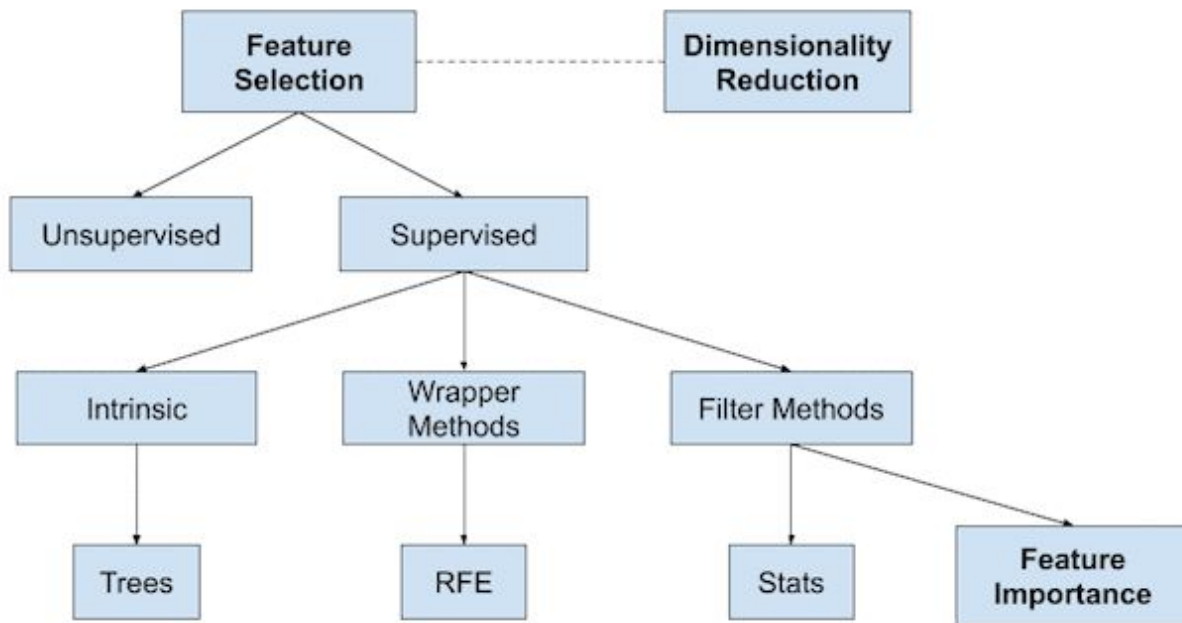
Feature Engineering

- Data Preparation: manipulation and consolidation of raw data from different sources into a standardized format (data augmentation, cleaning, delivery, fusion, ingestion, etc.)
- Exploratory Analysis: identify and summarize the main characteristics in a data set through data analysis and investigation
 - e.g., using data visualizations to understand the data, determine statistical techniques for data analysis, and choose the right features for a model.
- Benchmark: setting a baseline standard for accuracy to which all variables are compared in order to reduce the error rate and improve model's predictability.
 - metrics of benchmarking is decided by data scientists, domain expertise, and business users.

Feature Selection

The process of reducing the number of input variables when developing a predictive model

- to reduce the computational cost of modeling
- to improve the performance of the model.



Methods

- Filtering
 - Single feature evaluation: measure quality of features by different metrics
 - Frequency based
 - remove features according to frequency of features or instances contain the feature
 - Dependency of feature and label (co-occurrence)
 - mutual information (measure the dependence of two random variables)
 - Chi-square statistics (measure dependence of two variables)
 - Subset selection (category distance)
- Wrapping
 - Ranking accuracy using a single feature
 - Subset selection (Sequential forward/backward selection)

Knowledge Extraction Use Case

Showcasing simple data analysis, data visualization, and descriptive analytics

Use case: [Provinces of Iran - Wikipedia](#)

Popular Data Science Tools

Basic Tools/Libraries

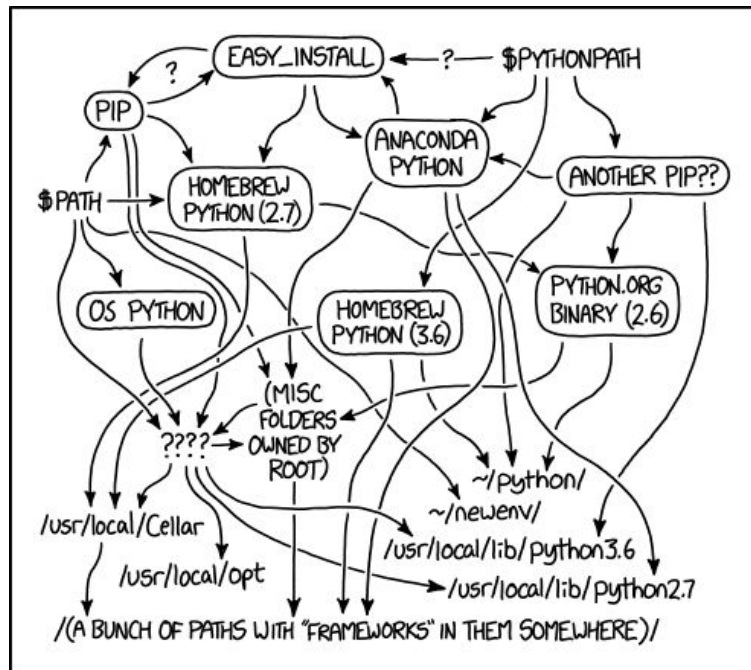
- The foundation rocks are Python and R.
- Jupyter
 - interactive data science and scientific computing
 - Supports a number of programming languages such as Julia, R, Haskell, Ruby, and Python
- Numpy
 - fundamental package for scientific computing with Python
- Matplotlib
 - plotting library for the Python
- Pandas
 - Library for data manipulation and analysis
- Machine learning tools
 - Scikit Learn, PyTorch, TensorFlow, Keras

Setting up the Environment in Python

A virtual environment is an isolated installation of Python and its libraries.

We use them to:

- Experiment with libraries
- Wrap project dependencies
- Avoid conflicting requirements/versions



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Conda

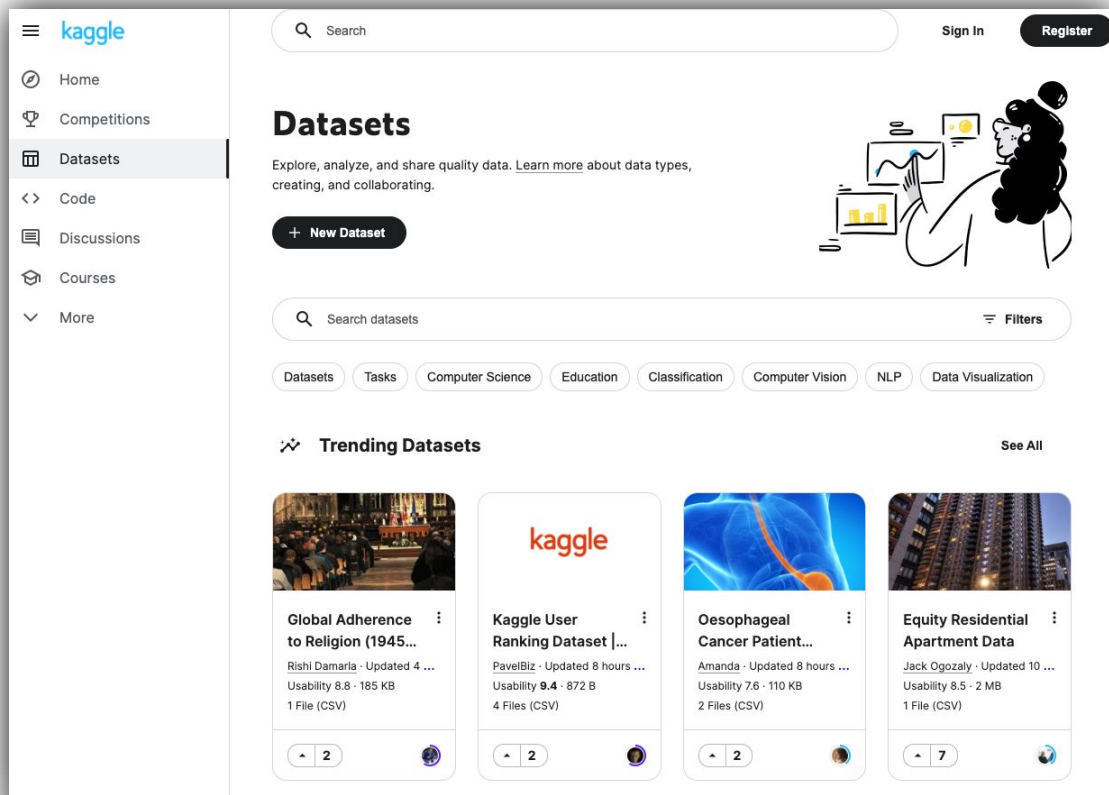
- Conda is an open source package management system and environment management system that runs on Windows, macOS and Linux.
- Supports many languages including language Python, R, Ruby, Lua, Scala, Java, and JavaScript.



Kaggle

Created by Google, is an online platform for Data scientists and Machine Learning enthusiasts.

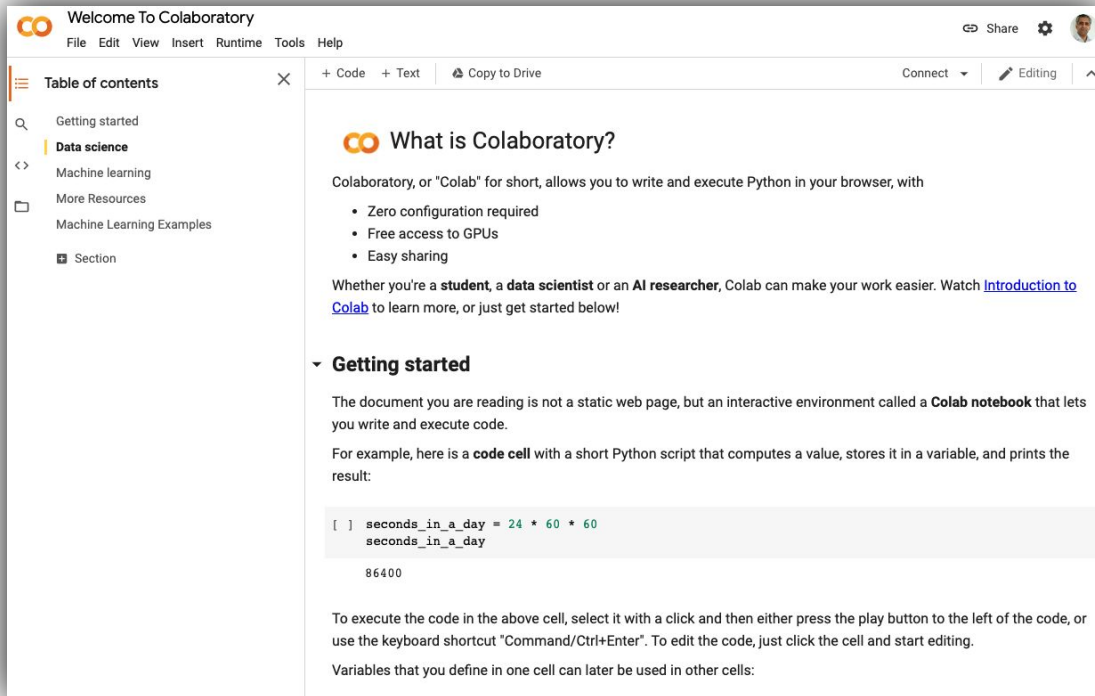
Allows users to find and publish various datasets for data science and machine learning, explore and build models in a web-based data-science environment



Google Colab

Google Colaboratory (also known as Colab) is a free Jupyter notebook environment that runs in the cloud and stores its notebooks on Google Drive.

Allows to write and execute Python in browser, with Zero configuration required, Free access to GPUs, and Easy sharing



Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

Share Settings Profile

Connect Editing

Table of contents

- Getting started
- Data science**
- Machine learning
- More Resources
- Machine Learning Examples
- Section

What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch [Introduction to Colab](#) to learn more, or just get started below!

Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ] seconds_in_a_day = 24 * 60 * 60
    seconds_in_a_day

86400
```

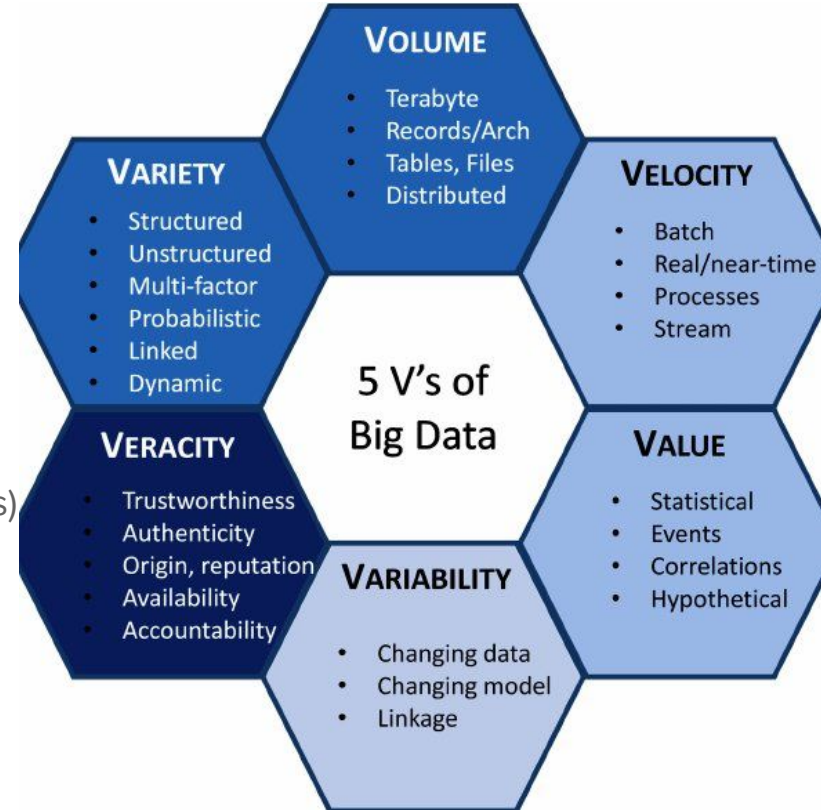
To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter". To edit the code, just click the cell and start editing.

Variables that you define in one cell can later be used in other cells:

Big Data Analytics

Big Data

- Extremely large, complex data sets
 - static/streaming
- Small or large data items
 - Ranging from sensor readings to large high quality satellite images
- Characterized by five-Vs
 - Volume: large size data sets (e.g., sensor readings)
 - Velocity: generation speed of data
 - Variety: complex and heterogeneous formats
 - Veracity: quality of data (data accuracy)
 - Value: ability to transform data into business



Big Data Processing - Batch vs. real-time Processing

Batch processing scenario:

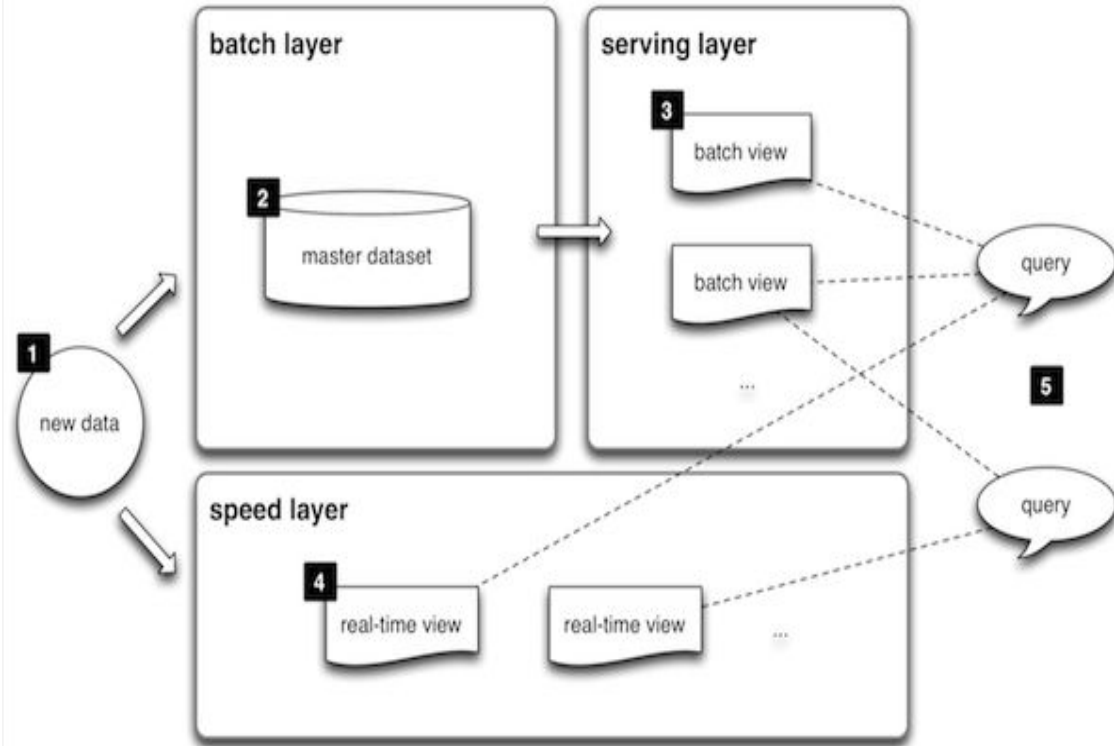


Real-time scenario:

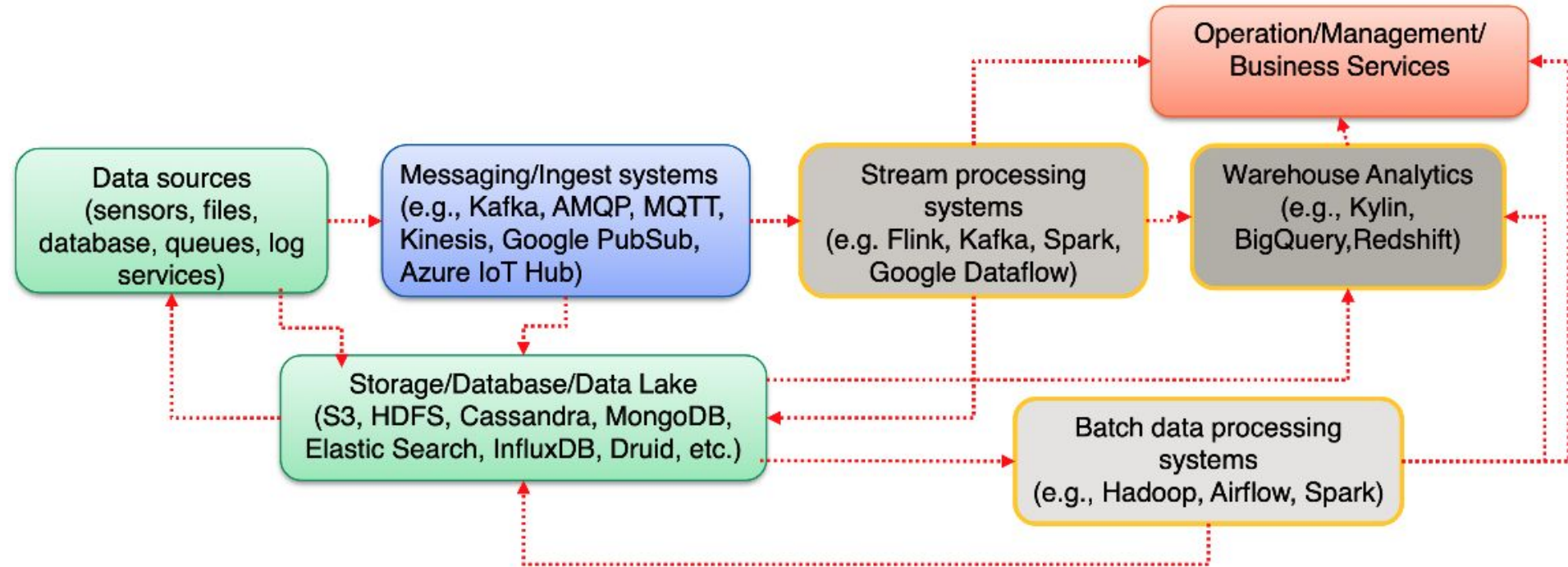


Lambda Architecture Style

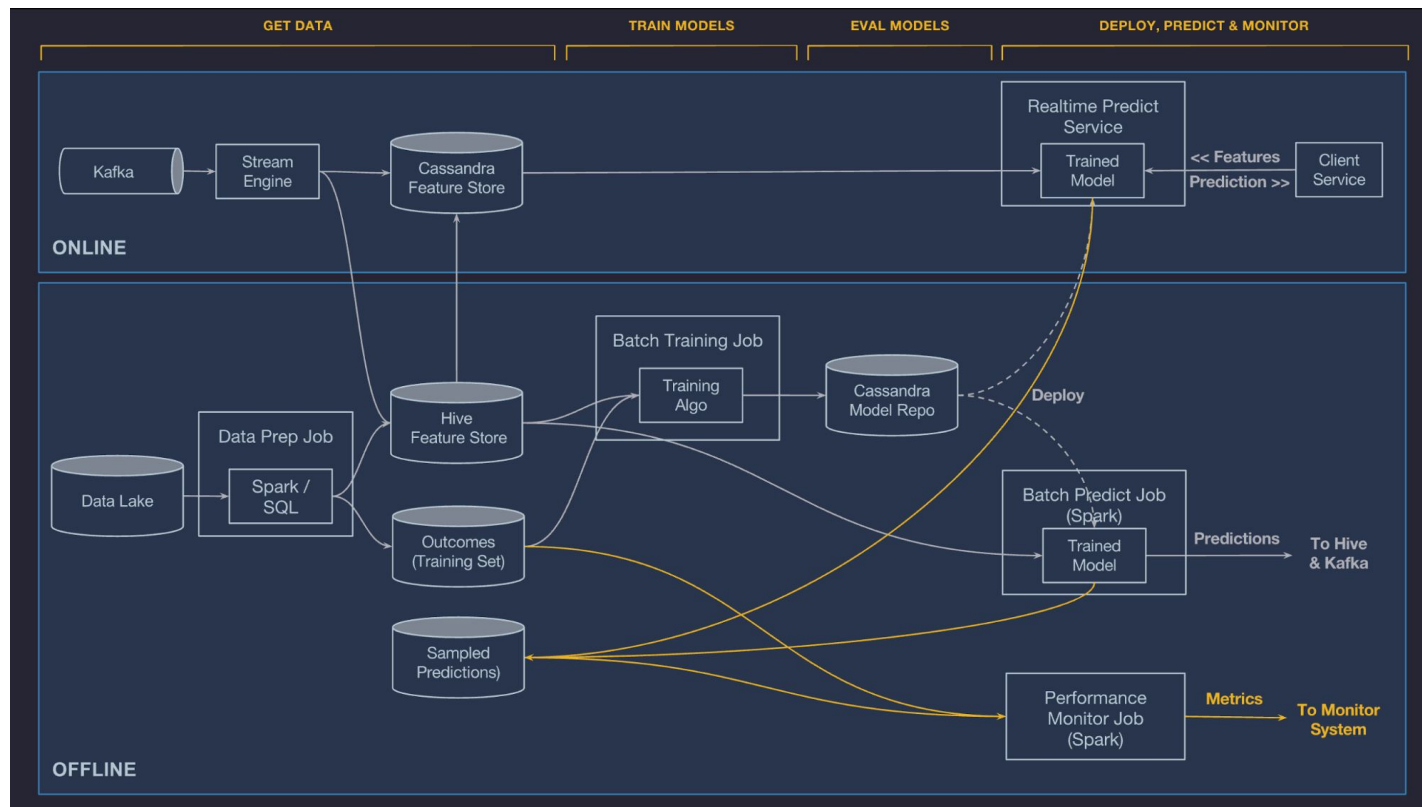
1. All data entering the system is dispatched to both the batch layer and the speed layer for processing.
2. The batch layer has two functions: (i) managing the master dataset (an immutable, append-only set of raw data), and (ii) to pre-compute the batch views.
3. The serving layer indexes the batch views so that they can be queried in low-latency, ad-hoc way.
4. The speed layer compensates for the high latency of updates to the serving layer and deals with recent data only.
5. Any incoming query can be answered by merging results from batch views and real-time views.



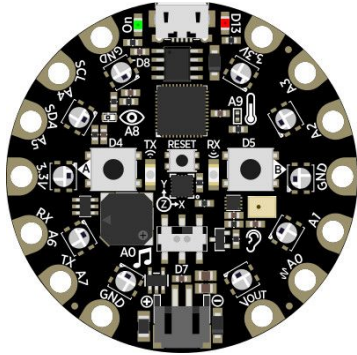
Big Data Analytics Processes and Tools



Michelangelo: Machine Learning Platform



IoT Use Case

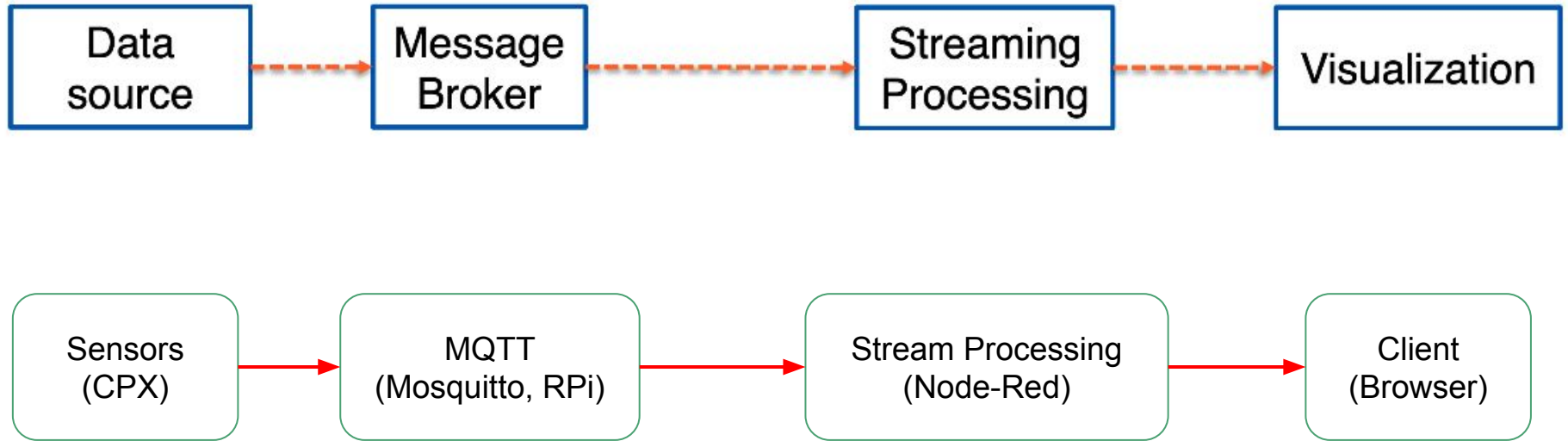


Circuit Playground Express



Raspberry Pi

Scenario



Soft Skills

Why Soft Skills?

In addition to tools and technicalities, you need further skills such as:

- Leadership
- Problem-solving attitude
- Communication skills
- ...

Essential Soft Skills (1/2)

- Critical Thinking
 - Critical thinking is about having a different perspective and the ability to understand what resources are critical to solving the problem.
 - You must know how to look at a problem, frame appropriate questions, and understand how the results will impact the business or target users.
- Curiosity
 - You need to ask questions that are overlooked in general.
- Effective Communication
 - You must have the confidence and skills to put all ideas on the table, discuss and justify all research, theories, and hypotheses, and effectively communicate their findings to technical and non-technical audiences.

Essential Soft Skills (2/2)

- Business Awareness
 - you will need to focus on how a business functions, the financial key points, and what the competition is like.
- Problem Solving Attitude
 - you need is to have the patience and determination to utilize data and make a way to solve the problem in-hand.

Thanks for your attention!
