# Introduction to Data Engineering

## Lecture 1

Rokan Uddin Faruqui

Associate Professor
Dept of Computer Science and Engineering
University of Chittaong, Bangladesh
Email: *rokan@cu.ac.bd*

# Outline

1. Data Engineer vs Data Scientist

# Outline

1. Data Engineer vs Data Scientist

# Data Engineer

- Data Engineers are the link between the management's data strategy and the data scientists that need to work with data.
- What they do is building the platforms that enable data scientists to do their magic.

# Data Engineer

- These platforms are usually used in five different ways:

  1. Data ingestion and storage of large amounts of data
  2. Algorithm creation by data scientists
  3. Automation of the data scientist's machine learning models and algorithms for production use
  4. Data visualization for employees and customers
  5. Most of the time these guys start as traditional solution architects for systems that involve SQL databases, web servers, SAP installations and other ”standard” systems.

# Data Engineer

- But to create big data platforms the engineer needs to be an expert in specifying, setting up and maintaining big data technologies like: **Hadoop, Spark, HBase, Cassandra, MongoDB, Kafka, Redis** and more.
- What they also need is experience on how to deploy systems on cloud infrastructure like at **Amazon** or **Google** or on-premise hardware.

# Data Scientist

- use linear algebra and multivariable calculus to create new insight from existing data.

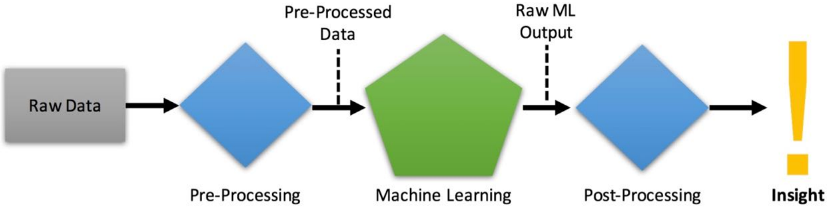
# Data Scientist: Case Study

- An industrial company produces a lot of products that need to be tested before shipping.
- Usually such tests take a lot of time because there are hundreds of things to be tested. All to make sure that your product is not broken.
- Wouldn't it be great to know early if a test fails ten steps down the line? If you knew that you could skip the other tests and just trash the product or repair it.
- That's exactly where a data scientist can help you, big-time. This field is called predictive analytics and the technique of choice is machine learning.

# Data Scientist: Case Study

- You feed an algorithm with measurement data. It generates a model and optimises it based on the data you fed it with.
- That model basically represents a pattern of how your data is looking. You show that model new data and the model will tell you if the data still represents the data you have trained it with.
- This technique can also be used for predicting machine failure in advance with machine learning. Of course the whole process is not that simple.
- The actual process of training and applying a model is not that hard. A lot of work for the data scientist is to figure out how to pre-process the data that gets fed to the algorithms.

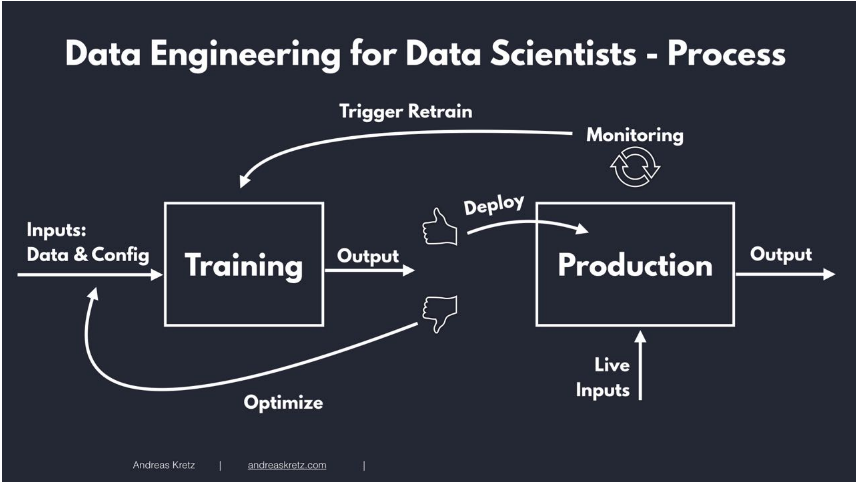
# Machine Learning Pipeline



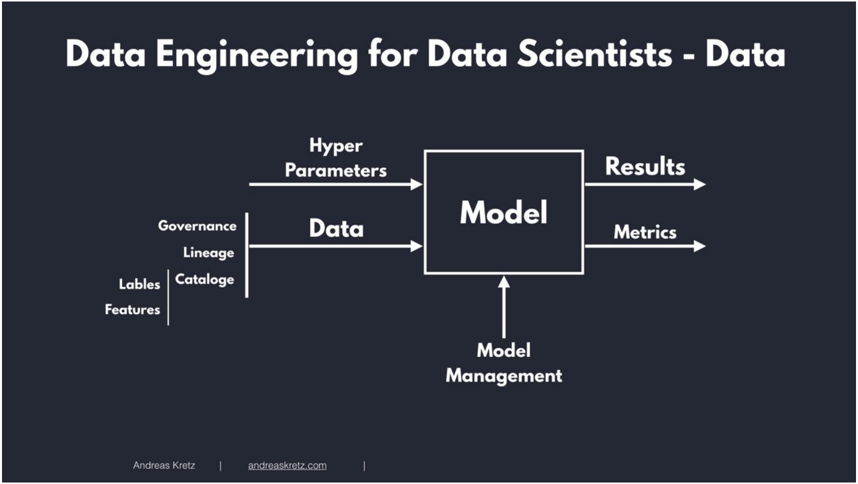
# Data Scientist: Case Study

- In order to train an algorithm you need useful data. If you use any data for the training the produced model will be very unreliable.
- An unreliable model for predicting machine failure would tell you that your machine is damaged even if it is not. Or even worse: It would tell you the machine is ok even when there is a malfunction.
- Model outputs are very abstract. You also need to post-process the model outputs to receive the outputs you desire.

# Machine Learning Workflow



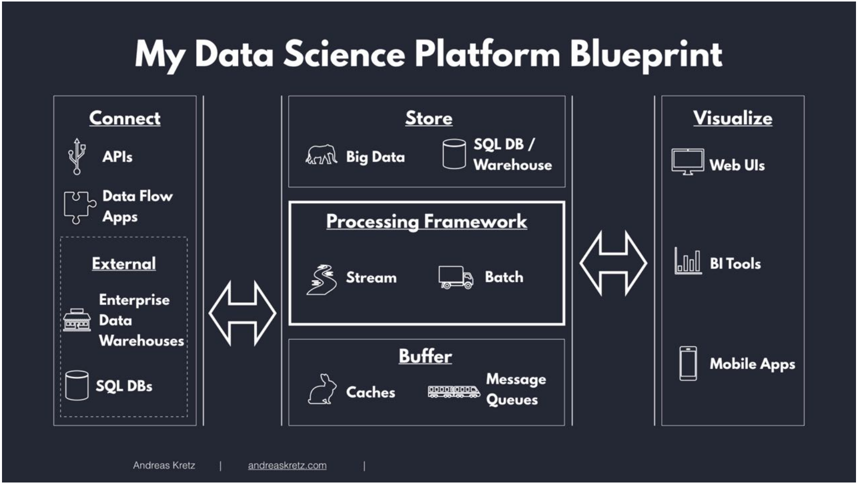
# Machine Learning Model



Data Engineering for Data Scientists - Data

Andreas Kretz | andreaskretz.com |

# Data Science Platform


My Data Science Platform Blueprint
Connect
APIs
Data Flow Apps
External
Enterprise Data Warehouses
SQL DBs
Store
Big Data
SQL DB / Warehouse
Processing Framework
Stream
Batch
Buffer
Caches
Message Queues
Visualize
Web UIs
BI Tools
Mobile Apps
Andreas Kretz | andreaskretz.com |