Introduction to Data-X

Data Science and Machine Learning I

Data-X Plaksha: Lec 1 Fall19

Link to slides: bit.ly/dxp-lec1

Alexander Fred-Ojala
Research Director, Data Lab
SCET, UC Berkeley

PLAKSHA UNIVERSITY

Pantas and Ting Sutardja Center for Entrepreneurship & Technology
Berkeley Engineering

Berkeley UNIVERSITY OF CALIFORNIA

Welcome to Applied Data Science with Venture Applications

Sutardja Center for Entrepreneurship & Technology
Industrial Engineering and Operations Research Department

Teaching Team

Student Presentations

Course History

Prerequisite: Students should have:

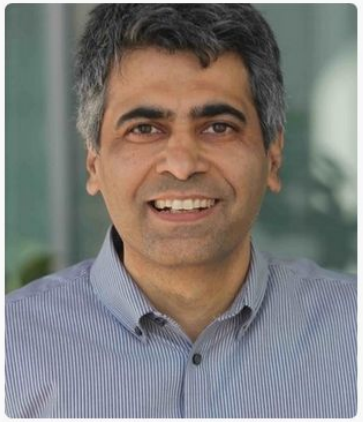
- a working knowledge of Python
- completed a fundamental probability / statistics course
- basic understanding Linear Algebra.

Teaching Team Data-X I

DR. RAVI KOTHARI

Former Chief Scientist, IBM Research India
Professor of CS, Ashoka University
Ph.D. (West Virginia University)

IBM's first Distinguished Engineer from India, Dr. Kothari has deep strategy and innovation experience. His research interests are in Neural Networks, Pattern Recognition, Machine Learning, Big Data, Analytics, Streaming Analytics, Telecom.



DR. IKHLAQ SIDHU

Faculty Director & Chief Scientist
Sutardja Center for Entrepreneurship & Technology, UC Berkeley
Ph.D. & M.S. (Northwestern University)

Dr. Sidhu has 75 patents under his name and is the inventor of seminal technology that is used in internet communications today. He is also the co-creator of Berkeley Method of Entrepreneurship.

Teaching Team Data-X I

Alexander **Fred-Ojala**

- Research Director
  Data Lab, SCET, UC Berkeley
- Co-creator
  UC Berkeley's Data-X course
- Founding Team of 4 companies
  InnoQuant (COO), Auranest (CMO),
  Wheely's (YCombinator alumni), Predli
- Amazon Innovation Fellow
  Awarded the AI fellowship 2018-2019

Co-creator of Data-X

Student Presentations (5-10 volunteers)

1. Name
2. Educational background
3. Professional background
4. Current knowledge of AI / ML / DS

Data-X: Course History

Motivation: STEM edu is broken

- **Missing Perspective**

  Gap between skills taught in academia and useful skills in industry.

  - **Golf analogy**
  - **Coding Bootcamps after graduation**
    - 10x growth over 4 years, $266Mn market*
  - **Need for a new teaching framework!**

* https://www.coursereport.com/reports/2017-coding-bootcamp-market-size-research

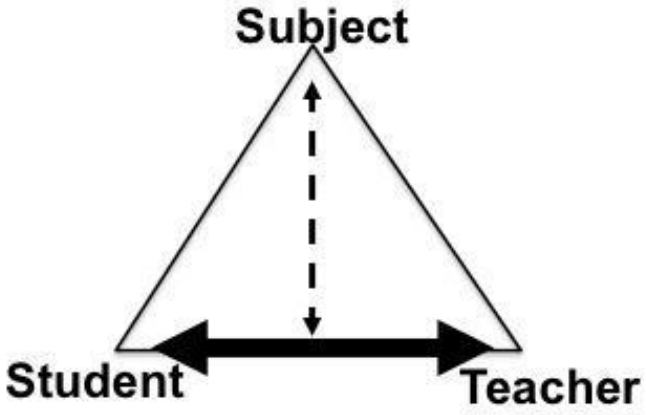
Background

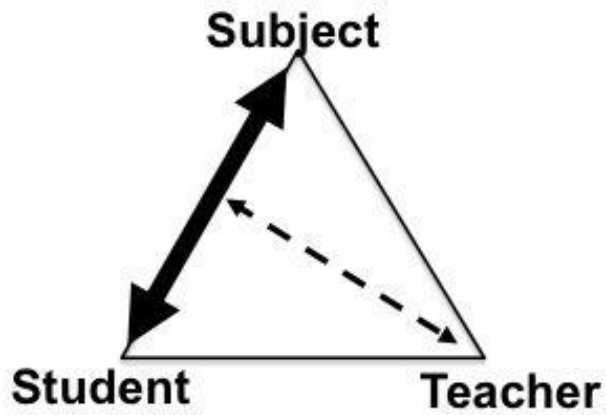
Berkeley Method of Entrepreneurship (BMoE)

Experiential, inductive learning. instead of teaching. Teach mindsets and behaviors linked to effective entrepreneurship.

Pantas and Ting
Sutardja Center
for Entrepreneurship & Technology
Berkeley Engineering

Shift in knowledge transfer



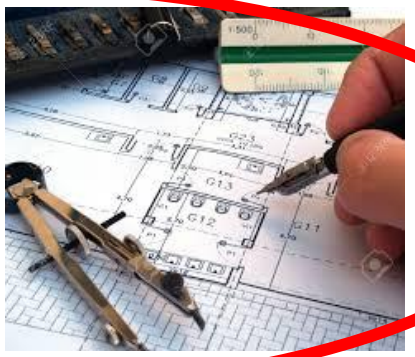


Approach: Course Philosophy

Data-X: Applied Data Science with Venture Applications

Make the Tools

Use State-of-the-Art Open Source Tools



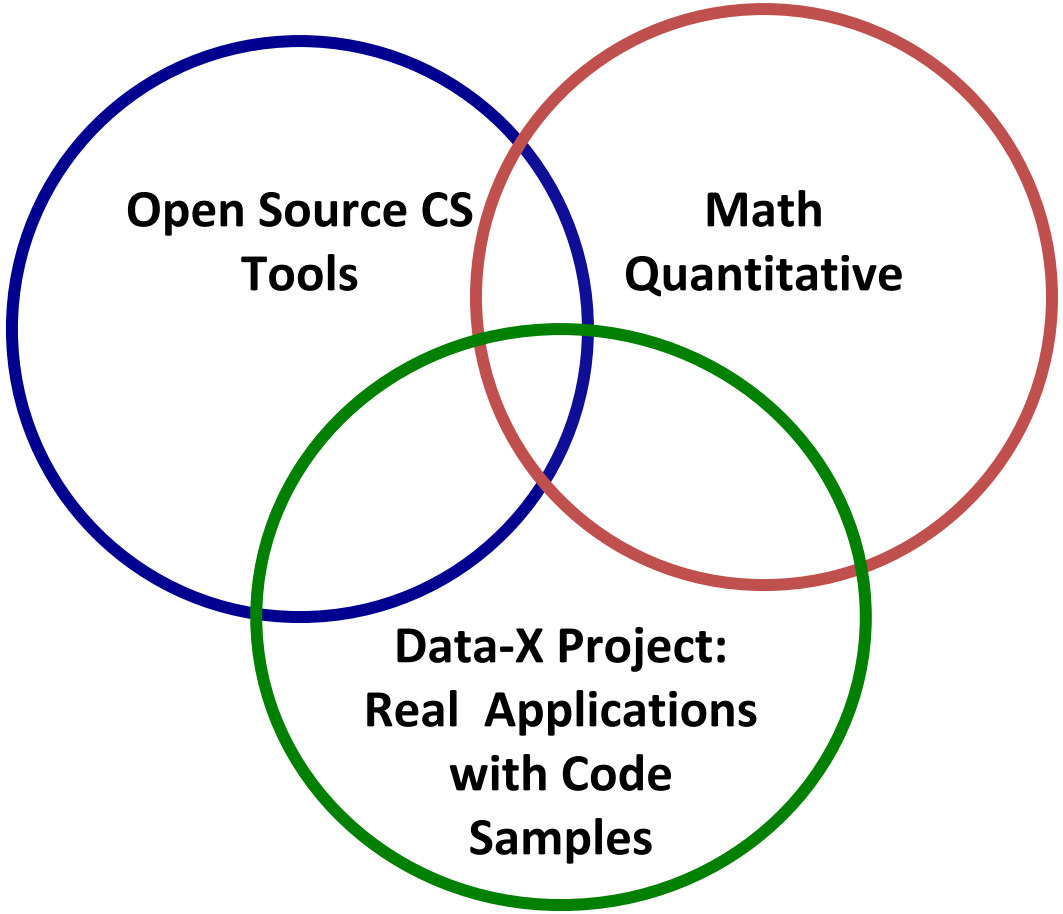
Architect the System

Sell, market, and pitch the product

Most CS / Math | **Data-X** | **Business Schools**

What is in this course

- Course Materials
- Applied Project
- Holistic Perspective: Industry, Social Applications, Customer Driven

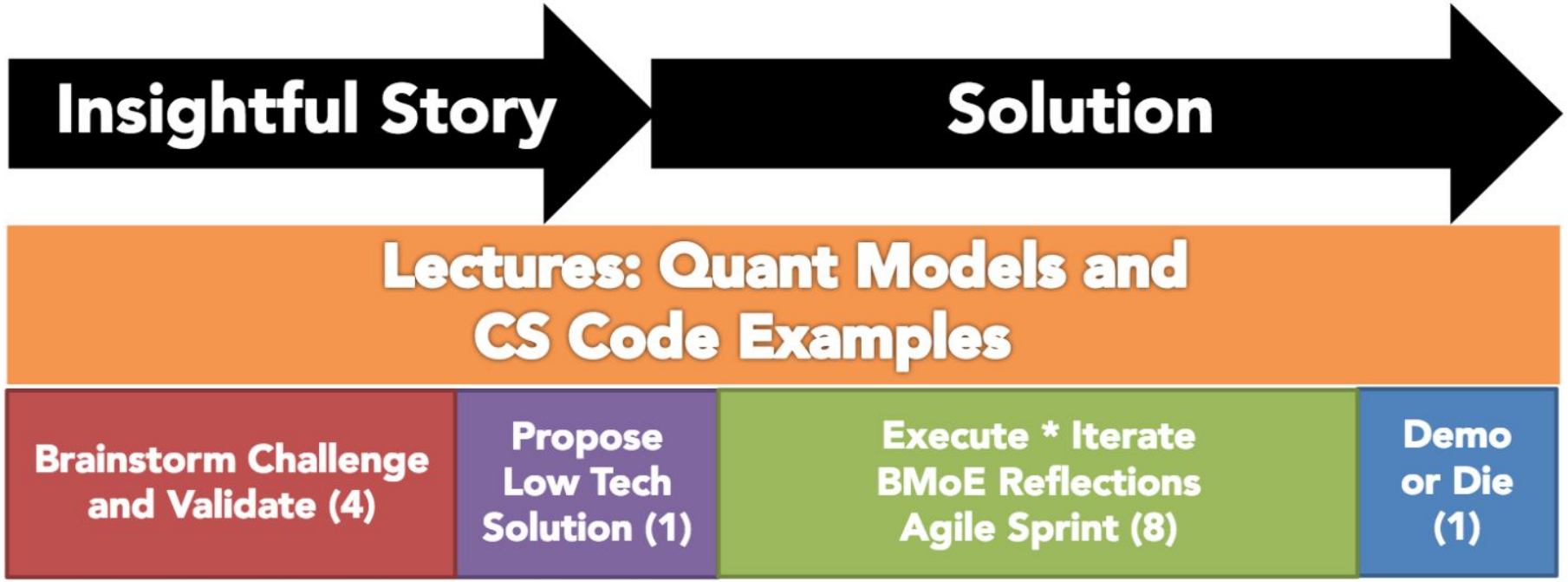


Holistic Perspective: Industry, Social Applications, Customer Driven

Content Overview

- The ML stack most commonly used in creating ML/AI/Data applications
- Application and systems viewpoint of data and ML
- Implementation, architecture, and relevant processes to build real systems
- Connection with relevant mathematical, statistical foundations (optimization, entropy, correlation, LTI, prediction, classification)
- Practical insight into advanced techniques and tools: (eg. CNNs, NLP, scraping, recurrent networks, etc.)
- Application talks: Recommender systems, Spark, Blockchain etc.

Course Overview



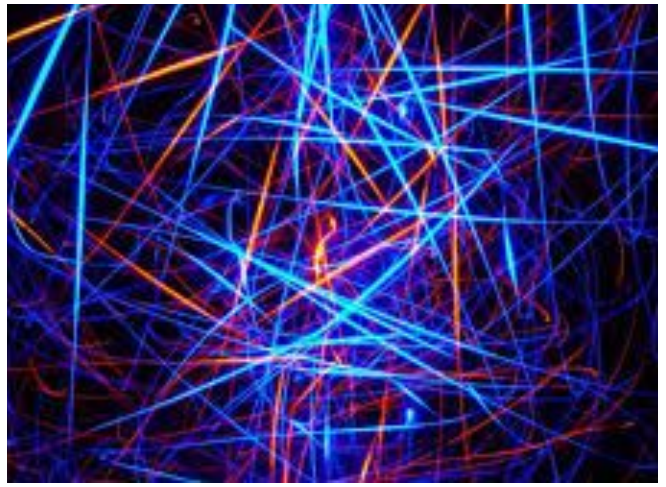
Open-ended, real-world project: Typically 5 students, with available advisor network

Project Types

Business or Consumer Use Case

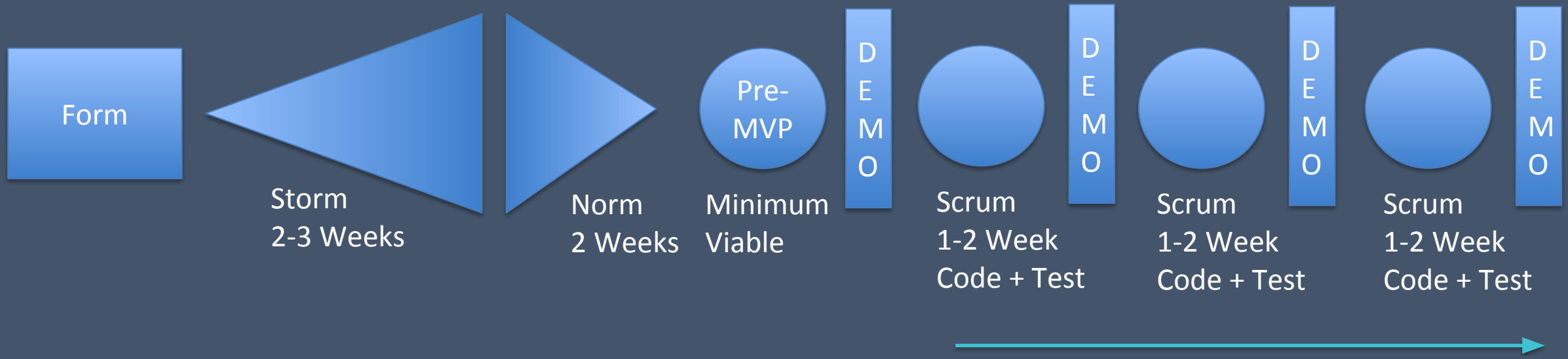
Social Impact

(or improve part of a data pipeline or work towards a research result)



Its Just Cool

Implementation of Behaviors and Process



1. Form, Storm, Norm
2. Minimum Viable Solution
3. Key skeleton components
4. Hypothesis ⟶ Test ⟶ Record
5. Agile Model for Feature Increments (for a changing objective)
6. Agile Analytics

Data-X: Highlights

Data-X

Pantas and Ting
Sutardja Center
for Entrepreneurship & Technology
Berkeley Engineering

Project Summary:

Applied Data Science class in the College of Engineering.

State of the Project:

- 230 students spring 2019
- 600+ alumni students
- 140+ great projects completed
- 10+ published research papers
- 100+ industry experts in network

Many students reach out afterwards and tell us that they landed their jobs as data scientists just because of the class.

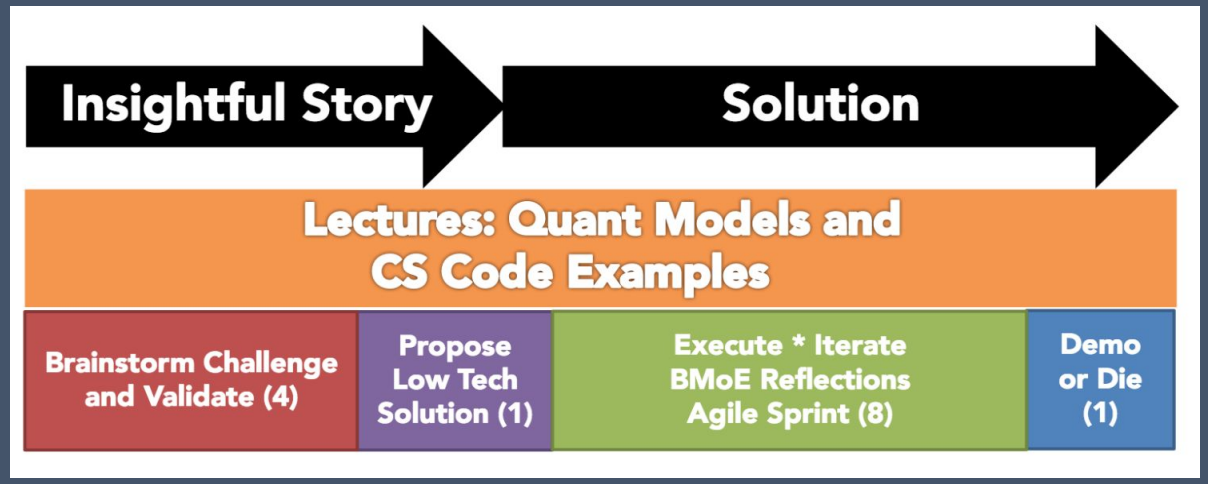
Official website: www.data-x.blog

Research Paper on the teaching methodology:
www.alex.fo/other/dataxieee.pdf

Data-X: Project

- **Totally open-ended!**
  - Collab with mentor
  - Find your own mentor
  - Come up with your own project
- **Groups of 5 (at most 6, at least 4)**
- **First weeks: Don't code.**

  => Lowtech demo.



Find examples at data-x.blog/projects

Project Categories:

Business Use Case

Social Impact

It's Just Cool

System Improvement

Improve Data Pipeline

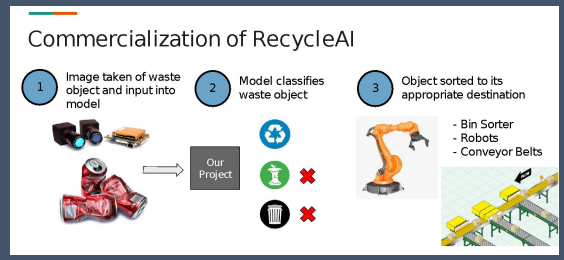
Novel Research

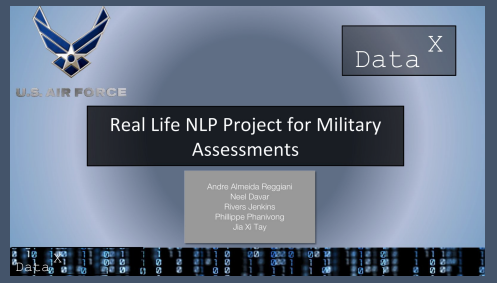
Data-X Project Examples

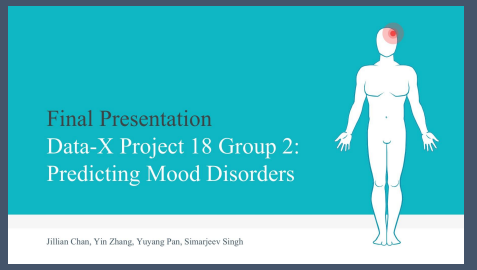
- Automatic detection of fake news
- Prediction of long-term energy prices
- Automatic recycling through image recognition
- AI for crime detection, traffic guidance, medical diagnostics, etc.
- A version of Zillow that is recalculated with the effects of AirBnB income
- Signal processing and pattern analysis to improve earthquake warning systems
- Early Autism Detection
- Secure Health Records stored on a Blockchain

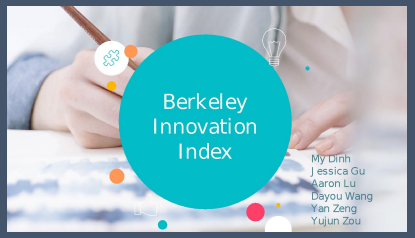
find many, many more at:

www.data-x.blog/projects









Examples of Data-X Company Collaborations Spring 2019

Google: Exploring the potential of Google's Coral edge-TPU hardware (for data security)

Honda: Real-time analytics of smart city and autonomous vehicle data. Smart parking project. Optimal placement of EV chargers.

Ford: How to optimize the load of Ford's Go bike micro mobility service.

Amazon: 3 teams are exploring Alexa projects related to text analysis, shopping optimization, and voice authentication

UN: 4 teams are working with A4Good a UN partner to help them reach their sustainability goals

Plus ~10 startup collaborations

Agenda Day 1

Today:

- Course Introduction High Level Overview of Data-X at Plaksha, and Data applications

By this week:

- Get your Notebook/development environment working
- Develop initial project ideas
- HW assigned by email/Plaksha LMS
- Foundations of applied ML covered

Key Dates:

Class Starts today

Lowtech Demo for Data-X Lab I: Presentation at end of Data-X I

Top project will receive chance to become a UC Berkeley Data Lab project

Data-X Plaksha website

bit.ly/plaksha

- Current Syllabus
- HWs, deadlines, topics
- Recommended Readings
- Grading Rubric
- Contact info
- etc.

More Resources Are Available at:

data-x.blog

Go to data-X.blog for

- Guide to Resources
- Project examples
- Inspiration

SYLLABUS
Edit

Applied Data Science with Venture Applications
IEOR 135/ 290-002

Instructor: Ikhlaq Sidhu
Department of Industrial Engineering & Operations Research

3 Units, Lecture and Lab

We will also be adding more video lectures to the Data-X On-line Tab

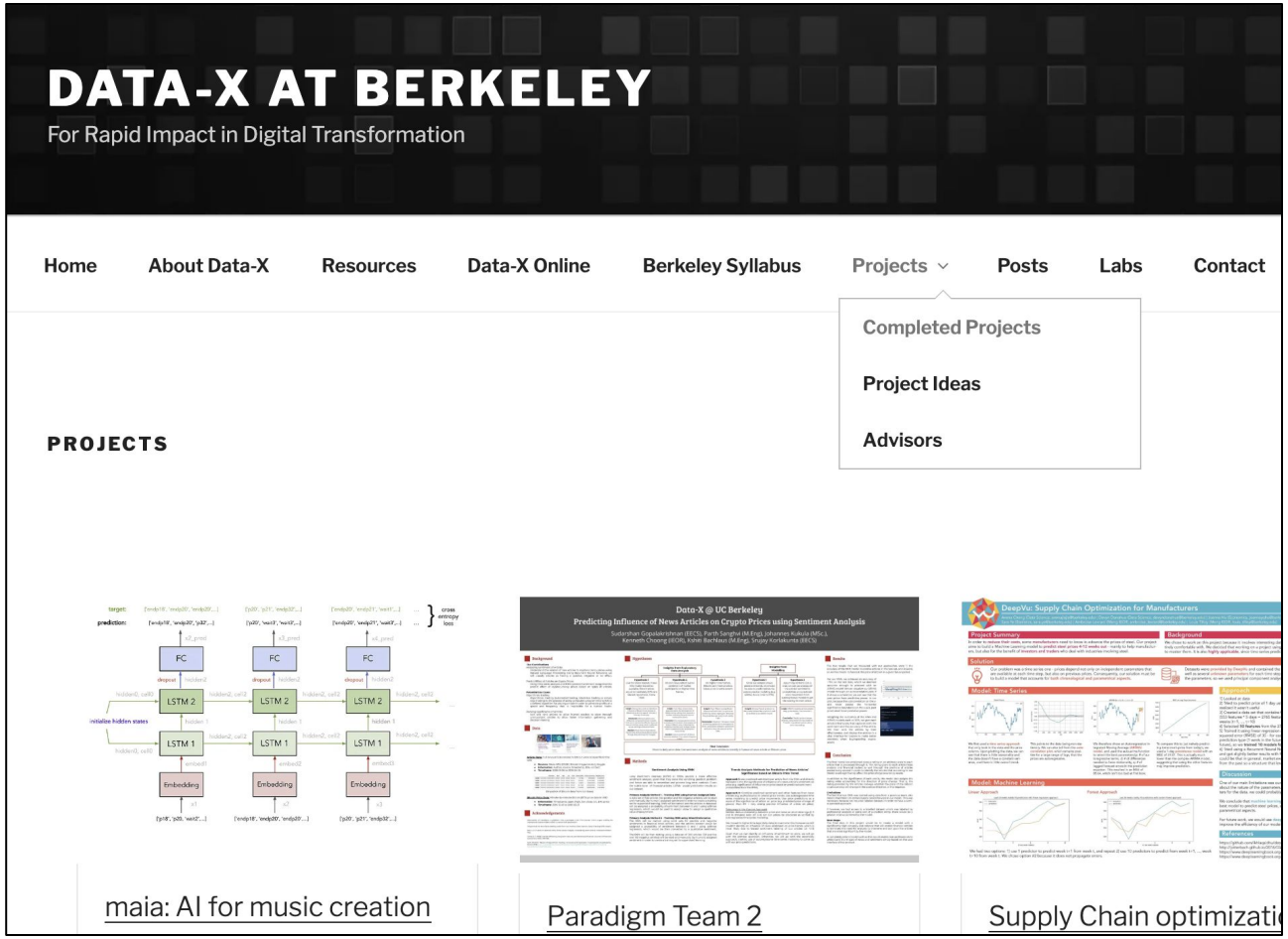
Homework For Week 1

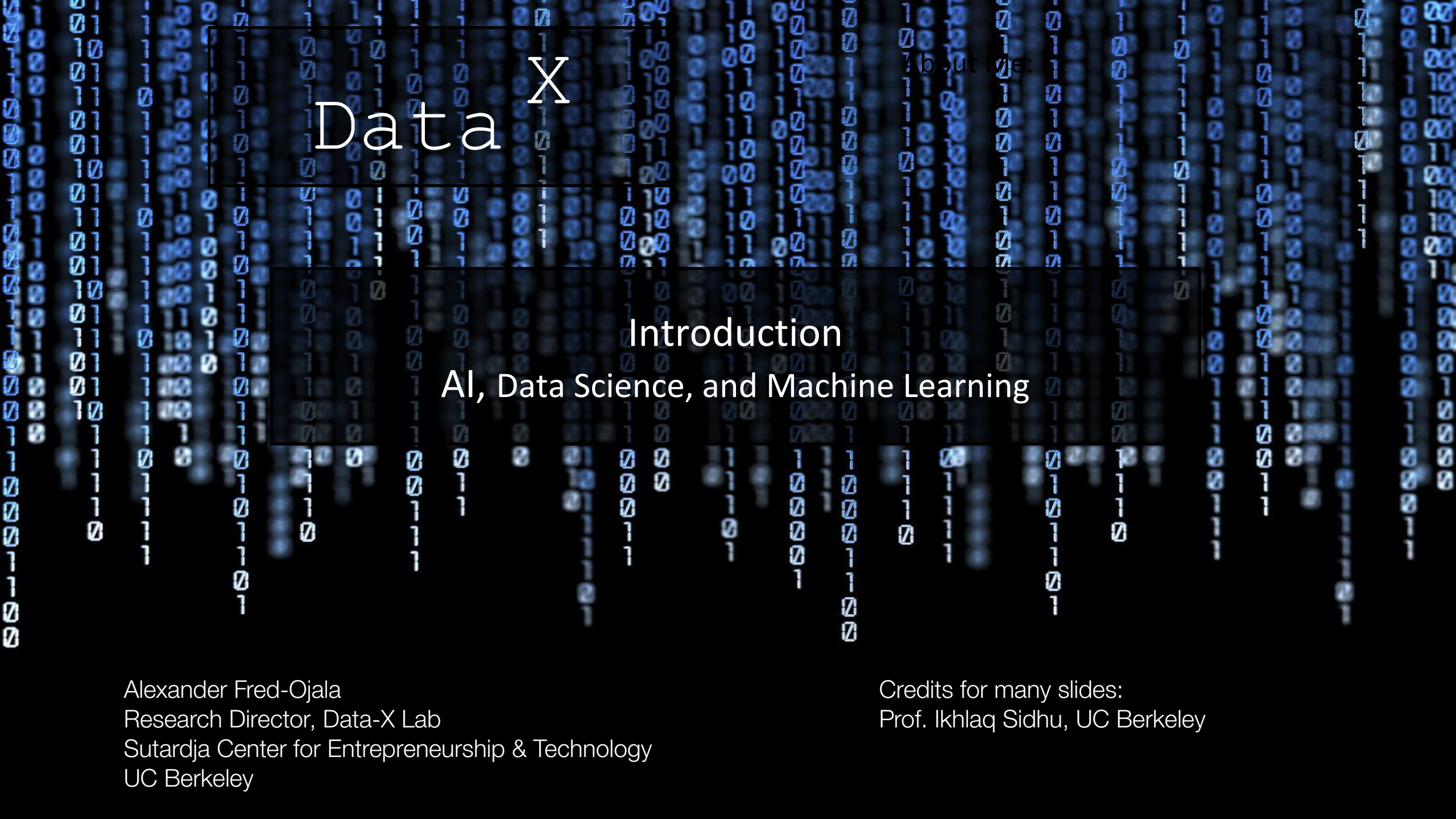
- **<u>HW Part 1: For Your Project – By Next week</u>**
  - Come up with 3 ideas for projects to pursue in Data-X Lab I in 3-5 sentences.
  - A systems or application you will build
  - ***Communicate:*** *WHO the project is for, WHAT will it do, WHY this is needed/valuable.*

- **<u>Homework Part II</u>**
  - Python-based review notebook (only Python concepts week 1)

Project Ideation

- **Past Projects Concepts:**
  - See the Advisor's Tab of data-x.blog
- **Past Projects:**
  - See the archive on the Posts page and on the Labs page of Data-x.blog
- Combine ideas or extend previous work
- **You can also choose to build part of a system,**
  - Ie, just the part that automatically collects data by web scraping, or
  - just the part that makes a decision based on data already available





Data X

About me:

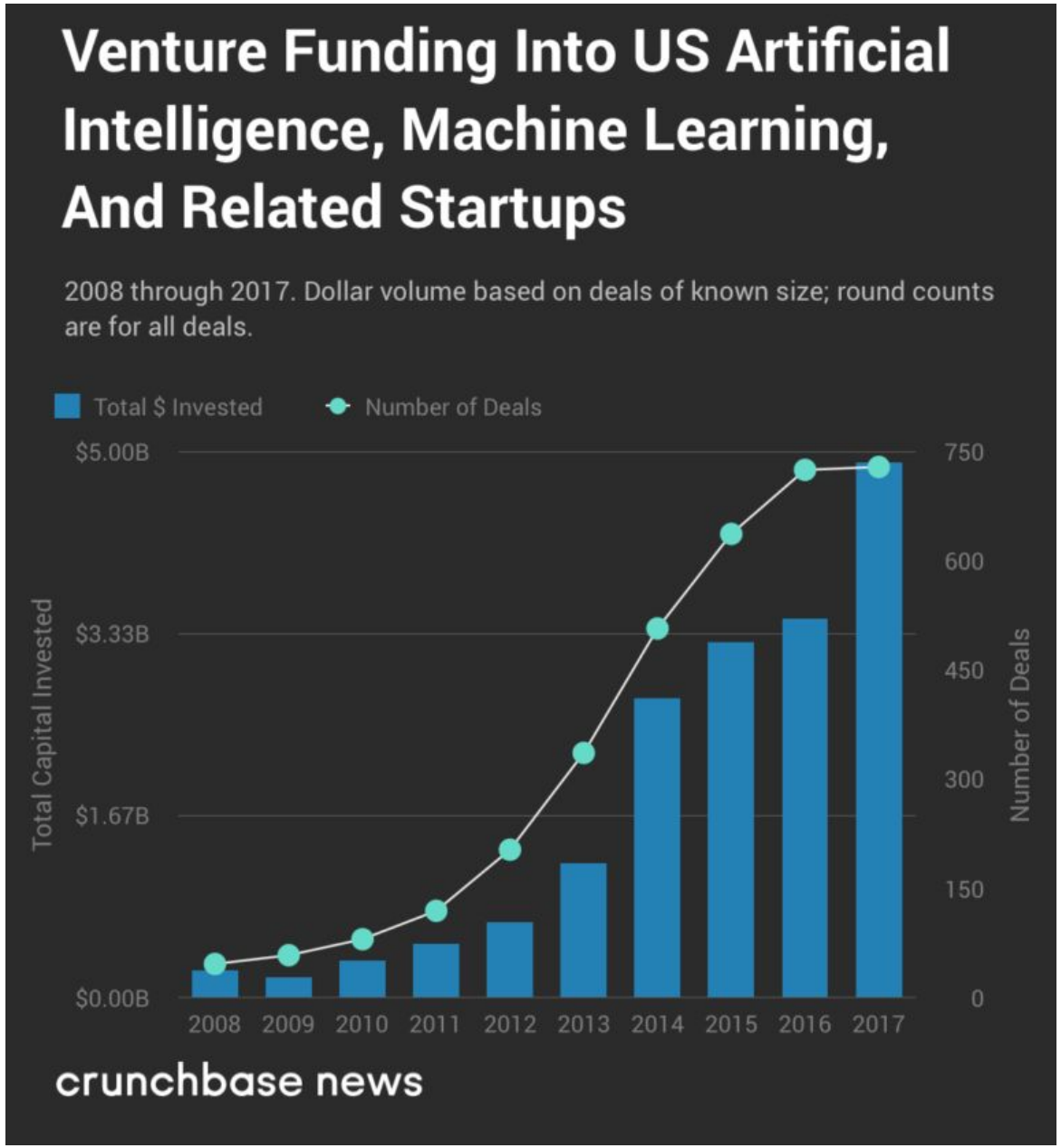
Introduction

AI, Data Science, and Machine Learning

Alexander Fred-Ojala
Research Director, Data-X Lab
Sutardja Center for Entrepreneurship & Technology
UC Berkeley

Credits for many slides:
Prof. Ikhlaq Sidhu, UC Berkeley

AI’s Remarkable Growth

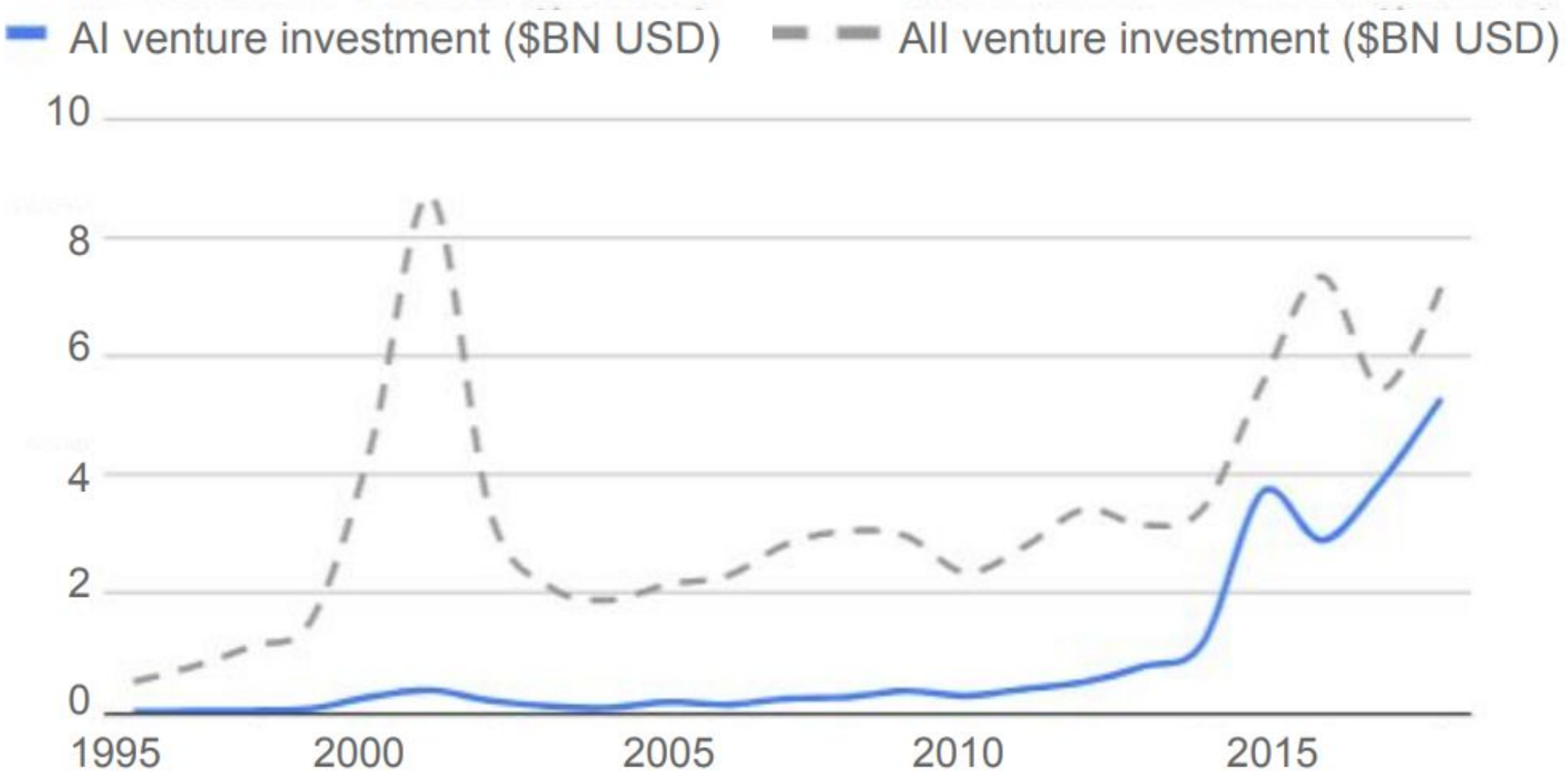


AI by the numbers

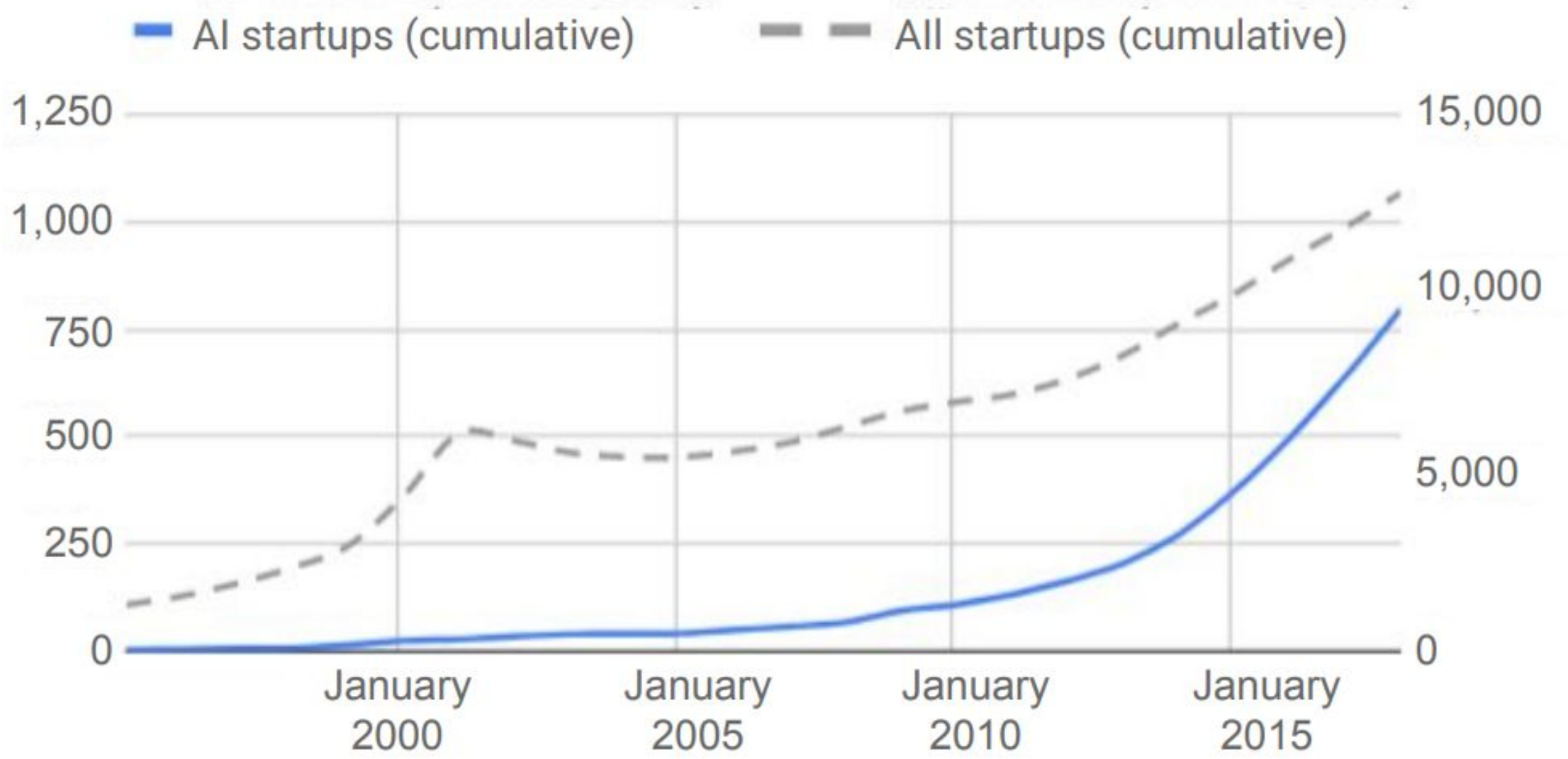
- *14X increase in the number of active AI startups since 2000.*
- *20X increase of investment into AI companies since 2000.*
- *The share of jobs requiring AI skills has grown 4.5X since 2013.*

Source: https://www.forbes.com/sites/louiscolumbus/2018/01/12/10-charts-that-will-change-your-perspective-on-artificial-intelligences-growth/#122c8f204758

AI Investments by VCs (80% of the investments go to AI companies)

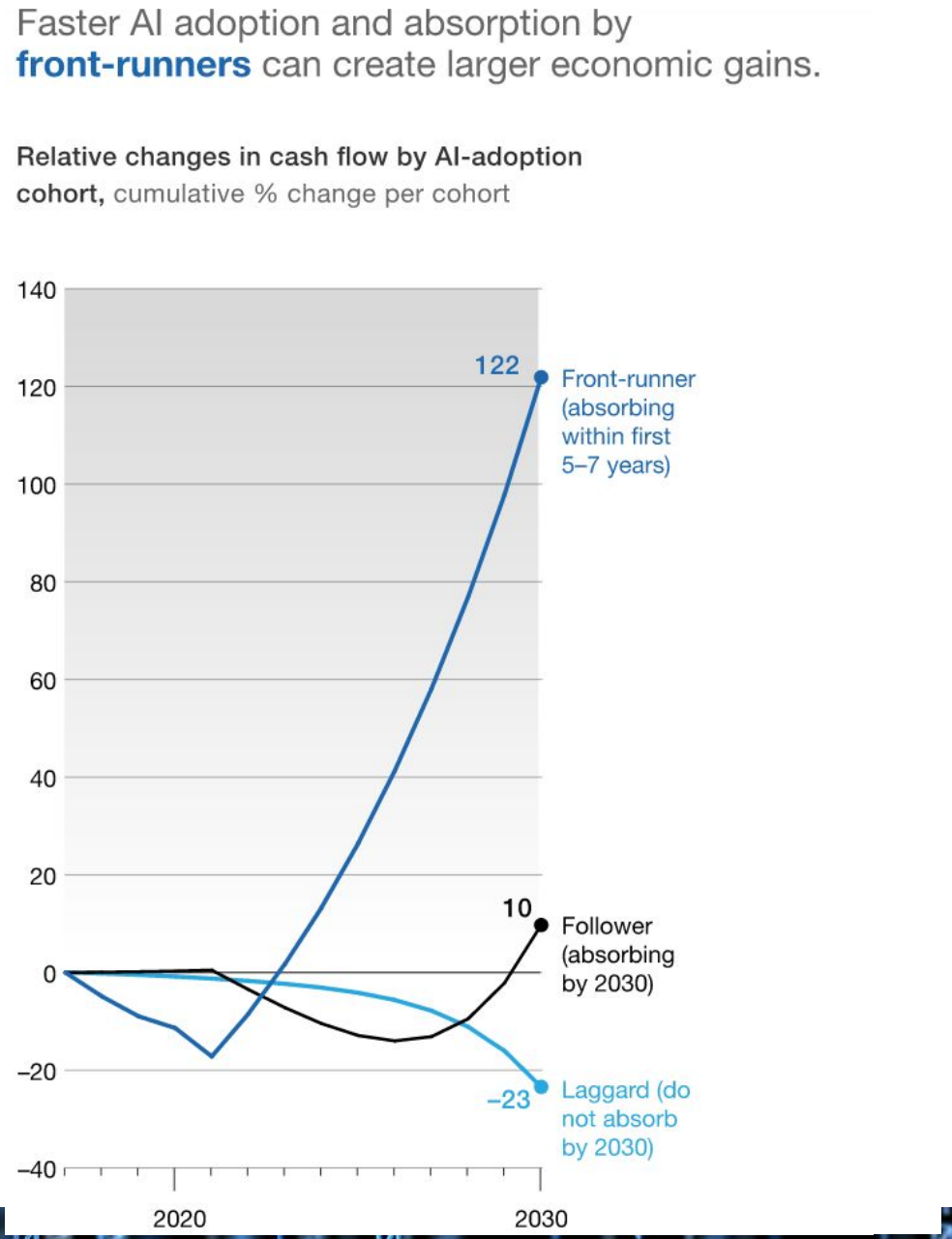


Number of AI startups (75% of all startups say they are AI companies)



AI by the numbers

According to McKinsey AI will add $13 trillion to the global GDP by 2030, and front runners will benefit greatly.



Source: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

Most valuable companies in the world

Most valuable companies in the world

Key Takeaway

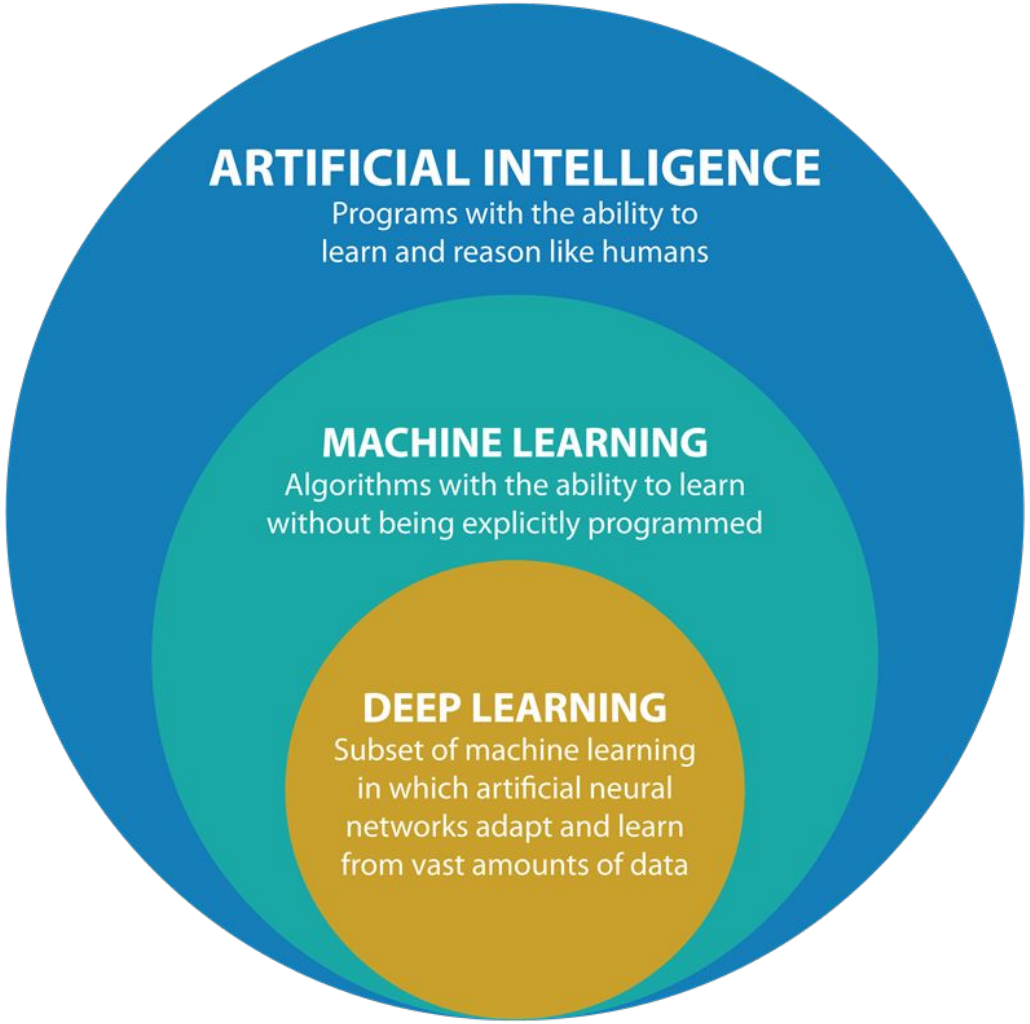
AI is a mature technology --
it delivers a lot of value.

It's important to be educated on what AI can and cannot do --
separate hype from reality

Definitions: What is AI / ML / DS?

Definitions

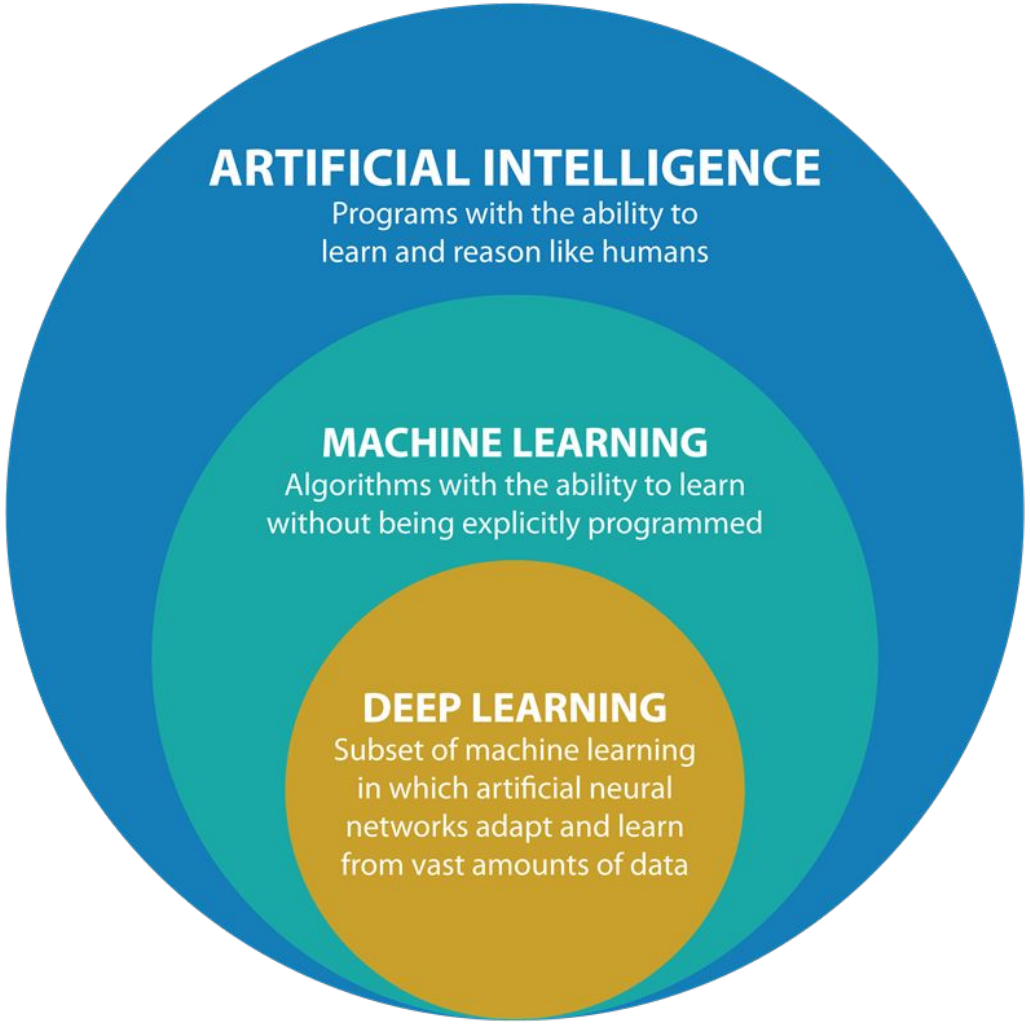
Data Science: A data scientist uses data to answer questions. Very flexible role.



Definitions

Data Science: A data scientist uses data to answer questions. Very flexible role.

Artificial Intelligence (AI): Methods for making computers behave intelligently and solve complex problems.

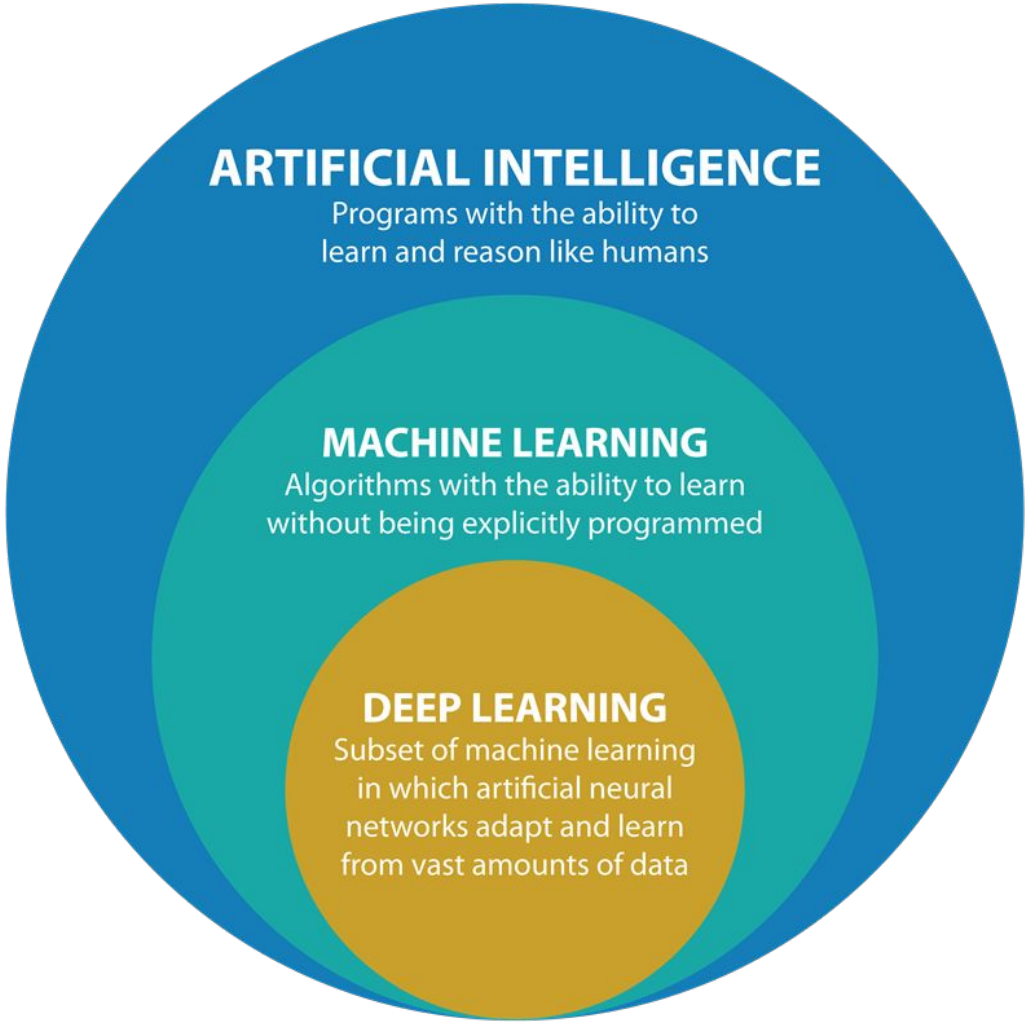


Definitions

Data Science: A data scientist uses data to answer questions. Very flexible role.

Artificial Intelligence (AI): Methods for making computers behave intelligently and solve complex problems.

Machine Learning (ML): Is a subset of AI that uses data in order to train models that can learn without being programmed.



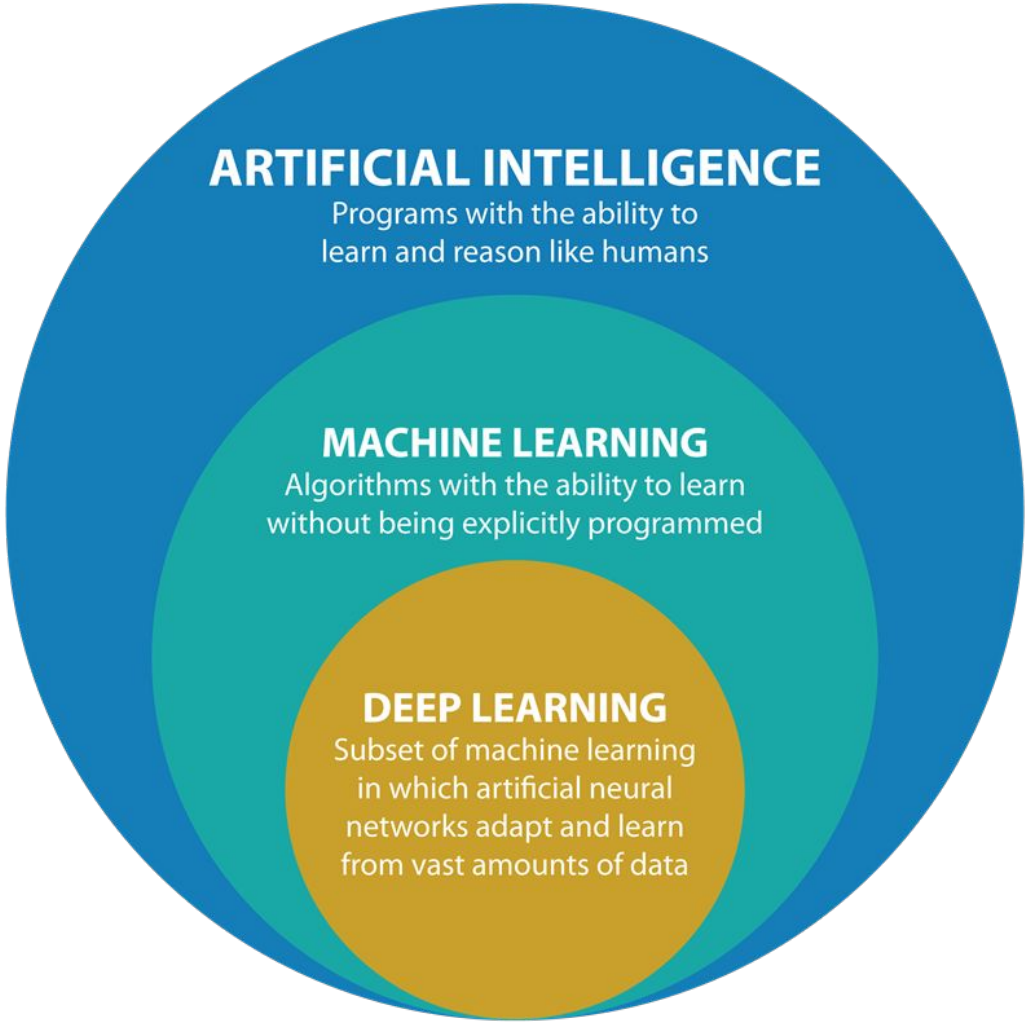
Definitions

Data Science: A data scientist uses data to answer questions. Very flexible role.

Artificial Intelligence (AI): Methods for making computers behave intelligently and solve complex problems.

Machine Learning (ML): Is a subset of AI that uses data in order to train models that can learn without being programmed.

Deep Learning: Deep neural networks and stacked models, a subset of ML.

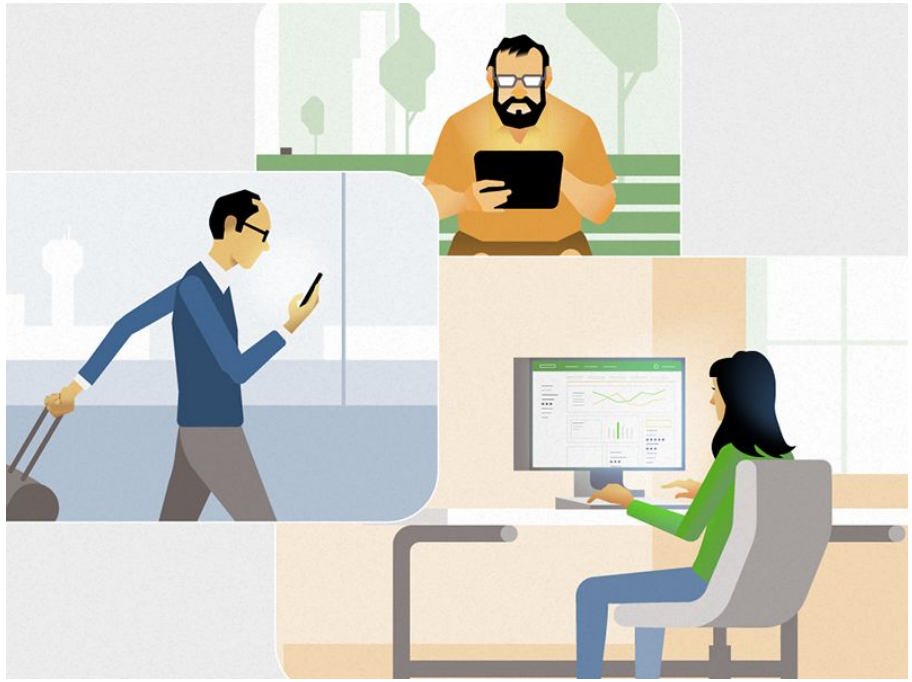


Three General AI Job Titles

1. **Statistician:** Analyzes model uncertainty, interprets results, adds rigor. Great when you want your model to always be correct.

2. **Machine Learning practitioner:** Swift model builder, agile iterations, optimizes for high accuracy. Great in experimentation phase.

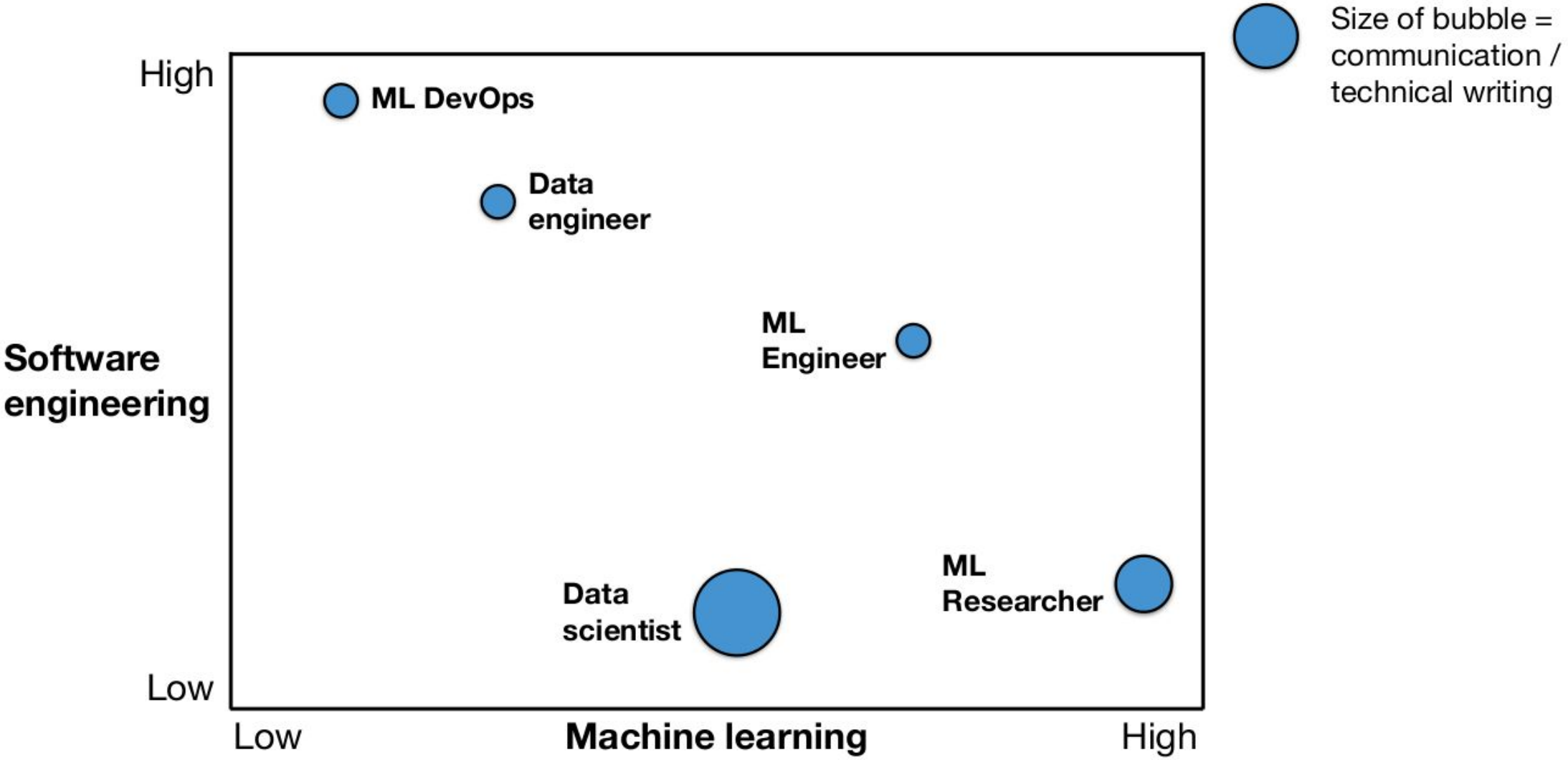
3. **Data analyst:** Explores data and uncovers insight. Great for influencing strategy, decisions, and project direction.



Breakdown of job function by role

Role	Job Function	Work product	Commonly used tools
DevOps Engineer	Deploy & monitor production systems	Deployed product	AWS, etc.
Data engineer	Build data pipelines, aggregation, storage, monitoring	Distributed system	Hadoop, Kafka, Airflow
ML Engineer	Train & deploy prediction models	Prediction system running on real data (often in production)	Tensorflow, Docker
ML Researcher	Train prediction models (often forward looking or not production-critical)	Prediction model & report describing it	Tensorflow, pytorch, Jupyter
Data Scientist	Blanket term used to describe all of the above. In some orgs, means answering business questions using analytics	Prediction model or report	SQL, Excel, Jupyter, Pandas, SKLearn, Tensorflow

What skills are needed for the roles?



Key Takeaway

Terms like AI and Machine Learning plus experts in the field have no specific definition -- there is no standard.

Data & AI expertise requires deep knowledge in software engineering.

An Overview of
Data and AI Applications

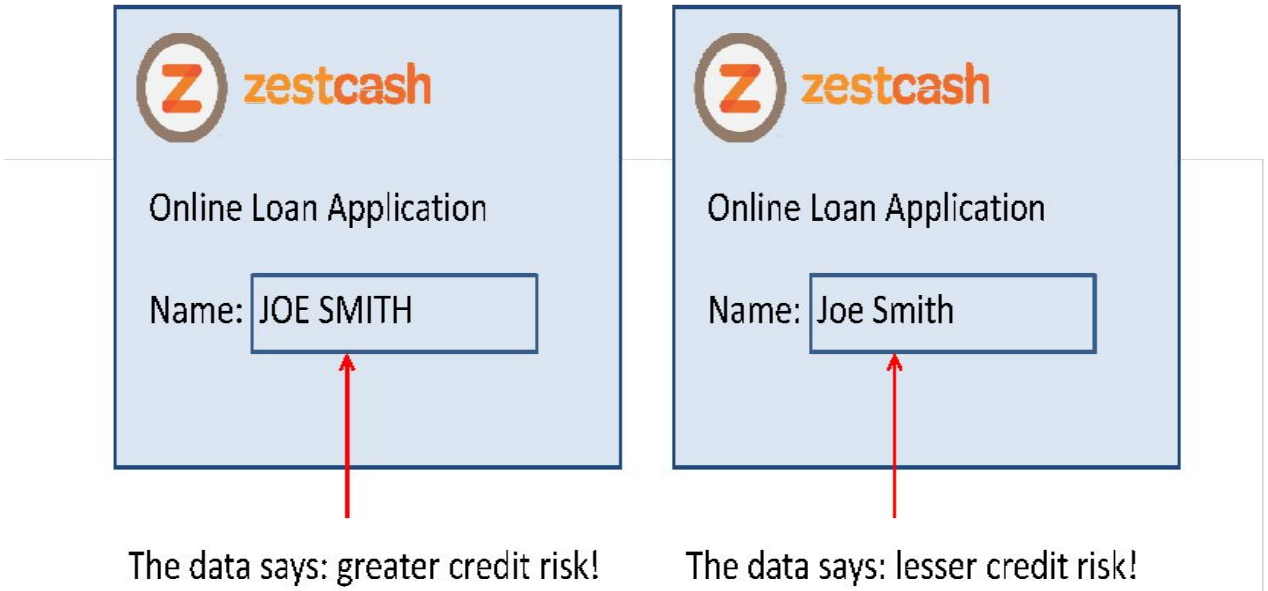
Basic Concept of Working with Data

- Data Wrangling
- In Production

Example: Data and Information is a Competitive Advantage

Real-life Example: ZestCash

- "All data is credit data"



Reference: Shomit Ghose

Another Real Life Example

Woolworths

Customers Who:
Drink lots of milk
Eat lots of red meat

Auto insurance risk: lower!

Customers Who:
Drink spirits
Eat lots of pasta and rice
Buy gasoline at night

Auto insurance risk: higher!

Reference: Shomit Ghose

- Service provider of Gambling and Casinos
- Entry Card
- Pain points
- Intervention

Harrah's Casino: Knowing your customer

Simplest Application Viewpoint

Customer Insight/ Engagement

Operations: Reliable & Predictable

Security & Fraud

Financial Firms

Network Security

Deep Learning Application Examples

Accuracy and usability has increased tremendously since 2012

Self-driving cars

Learns by understanding the world.

Healthcare

Cancer diagnostics, monitoring, personalised medicine.

Financial Forecasting

Time series data, recurrent algorithms

Voice Assistants & Translation

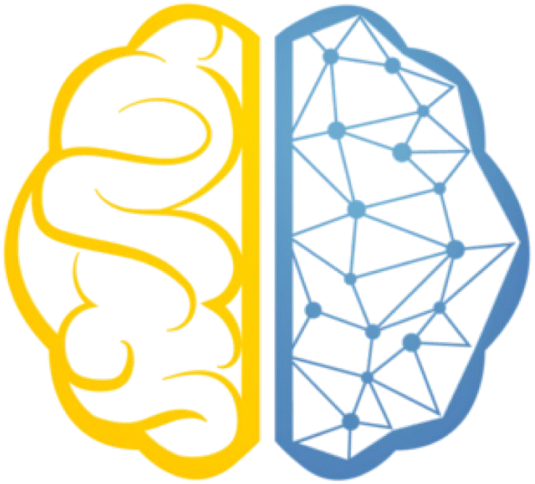
Siri, Google Home, Alexa

Image Recognition

Identify people etc

Sentiment analysis

Extract emotions from text and images



Common Data Science Projects

1. Fraud and anomaly detection
2. Predicting customer value
3. A/B testing
4. Recommendation engines
5. Customer Support automatization
6. Data Integration & Collection

Implementation:
SW Tools / Stack

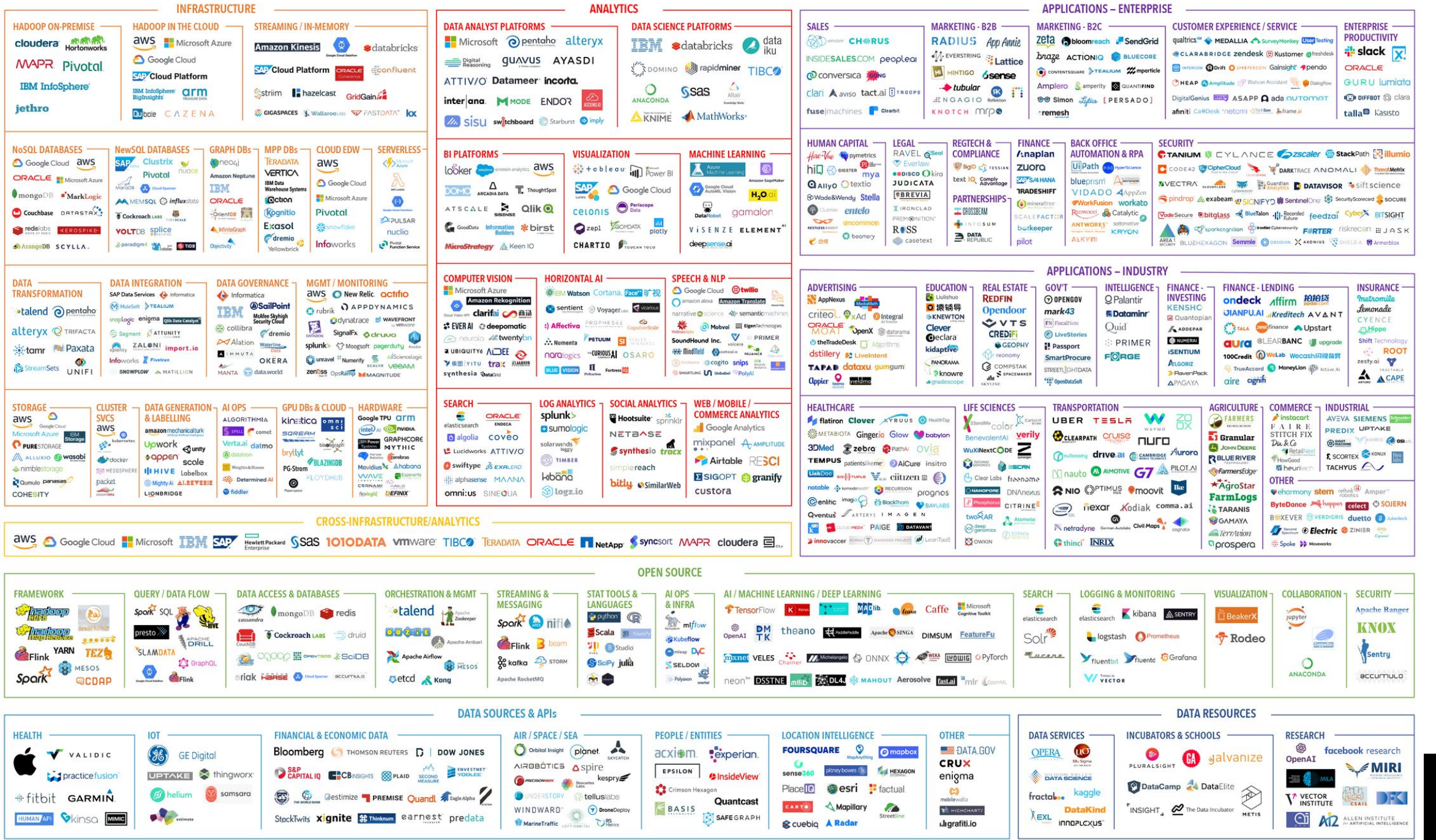
The Most Common Open Source Tools: AI/ML Stack

Start with Python as an interface

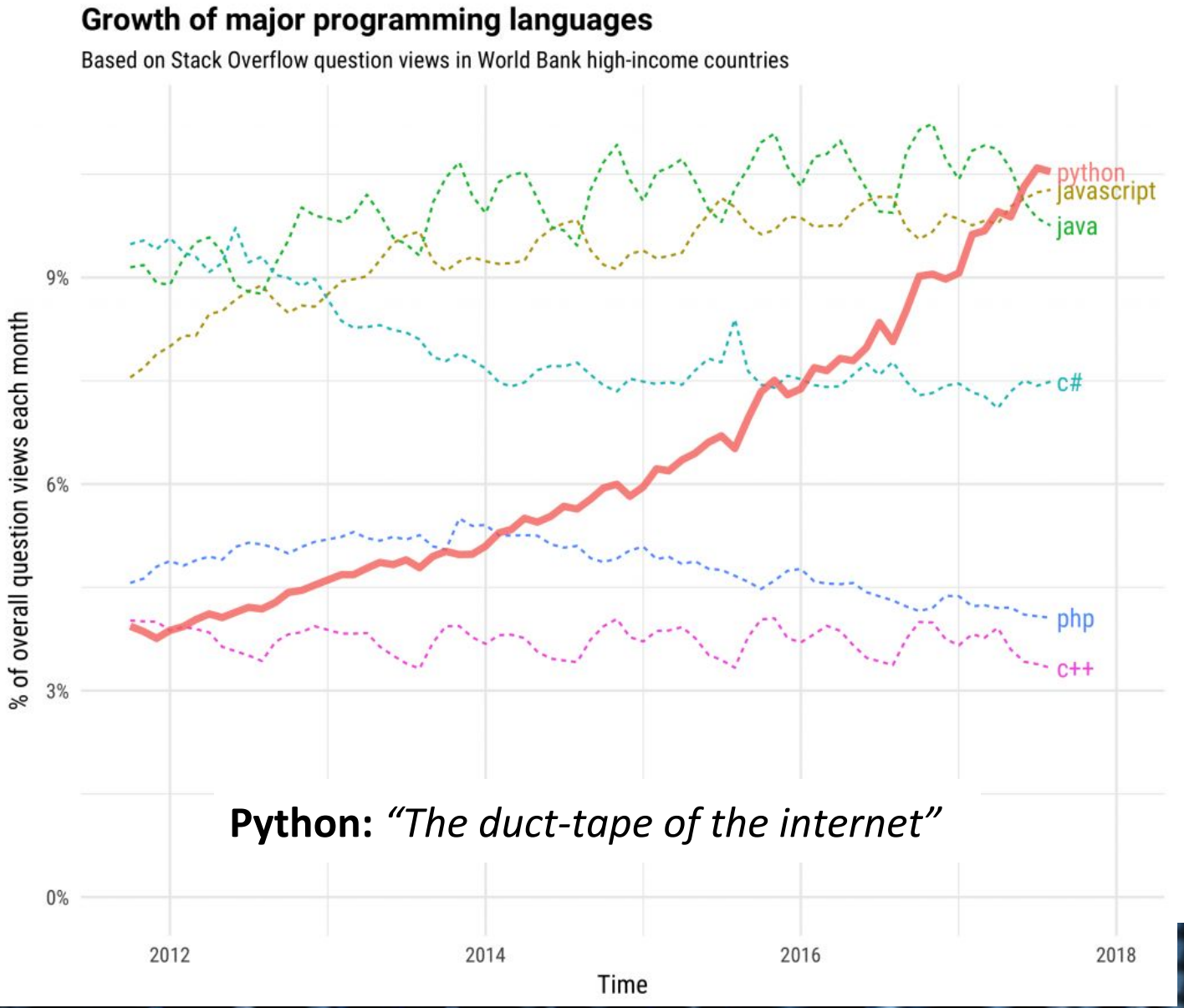
Jupyter Notebooks for prototyping

- **Python:** Core Programming Language
- **NumPy, SciPy:** Working with Arrays
- **Pandas:** Working in Tables, SQL to Pandas
- **Scikit-learn:** ML
- **Matplotlib, Seaborn, Plotly:** Visualizing Data
- **TensorFlow, Keras, Pytorch:** Neural Networks
- **Open-CV:** Image Processing / Computer Vision
- **NLP, word2vec, NLTK:** Natural Language
- **Spark, Kafka:** For large data sets (GB, TB+)
- **Airflow, CI:** Task Management at scale

DATA & AI LANDSCAPE 2019



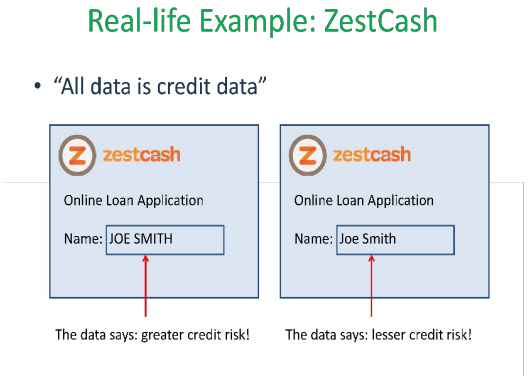
Why Python?



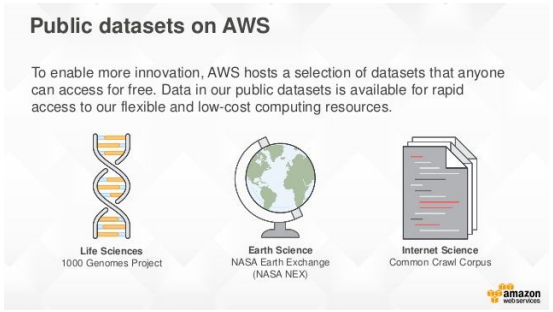
Python: *"The duct-tape of the internet"*

Where Does Data Come From?

Where Does Data Come From?



Your Own Website



Public Data Sets
Stock market, APIs etc.

IOT/Sensors

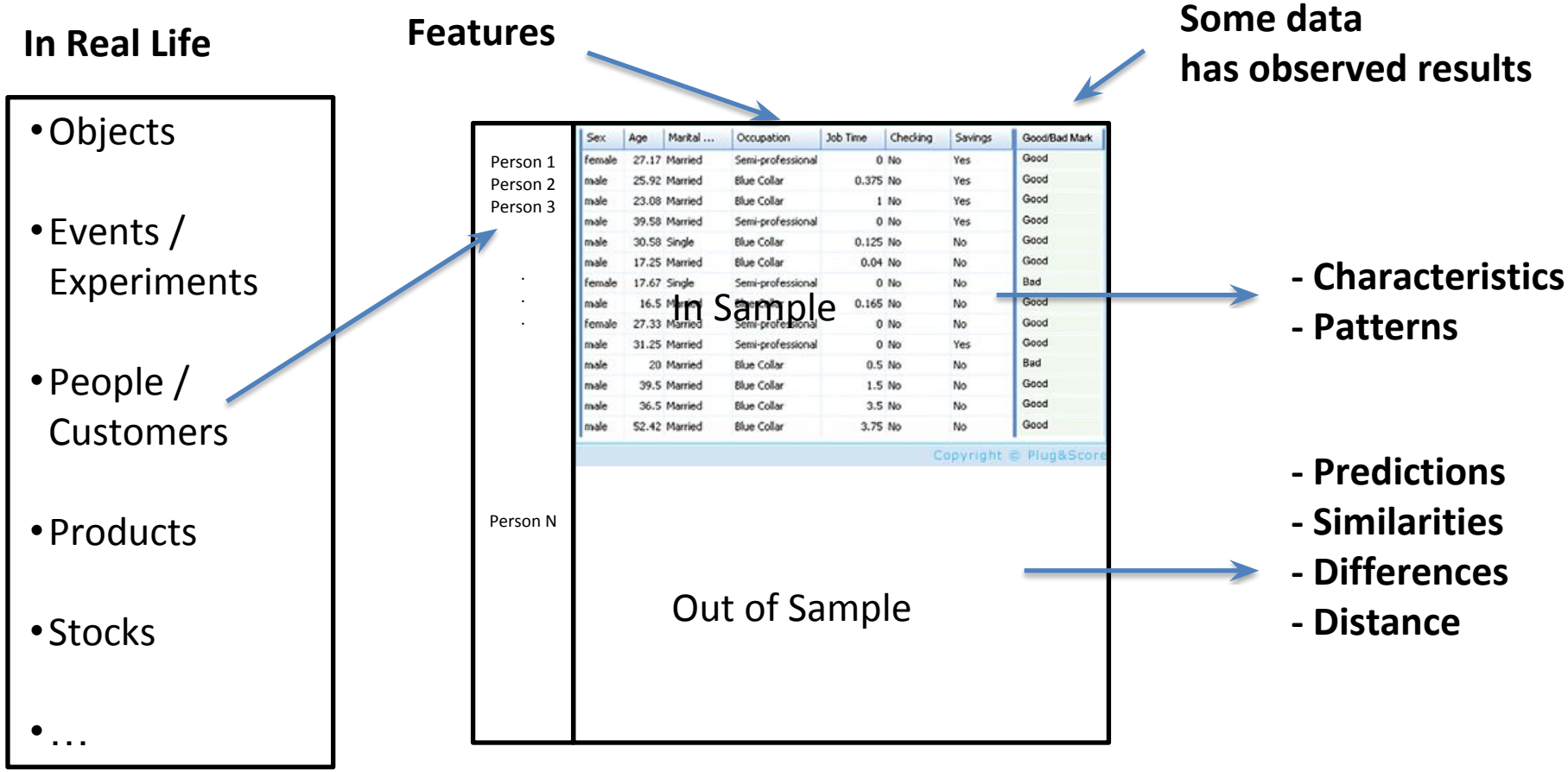
Other Websites

Web Scraping


```python
from bs4 import BeautifulSoup
import requests
page_link ='https://www.website_to_crawl.com'
# fetch the content from url
page_response = requests.get(page_link, timeout=5)
# parse html
page_content = BeautifulSoup(page_response.content, "html.parser")

# extract all html elements where price is stored
prices = page_content.find_all(class_='main_price')
# prices has a form:
#[<div class="main_price">Price: $66.68</div>,
# <div class="main_price">Price: $56.68</div>]

# you can also access the main_price class by specifying the tag of the class
prices = page_content.find_all('div', attrs={'class':'main_price'})
```

Formatting Data

An ML High Level Framework



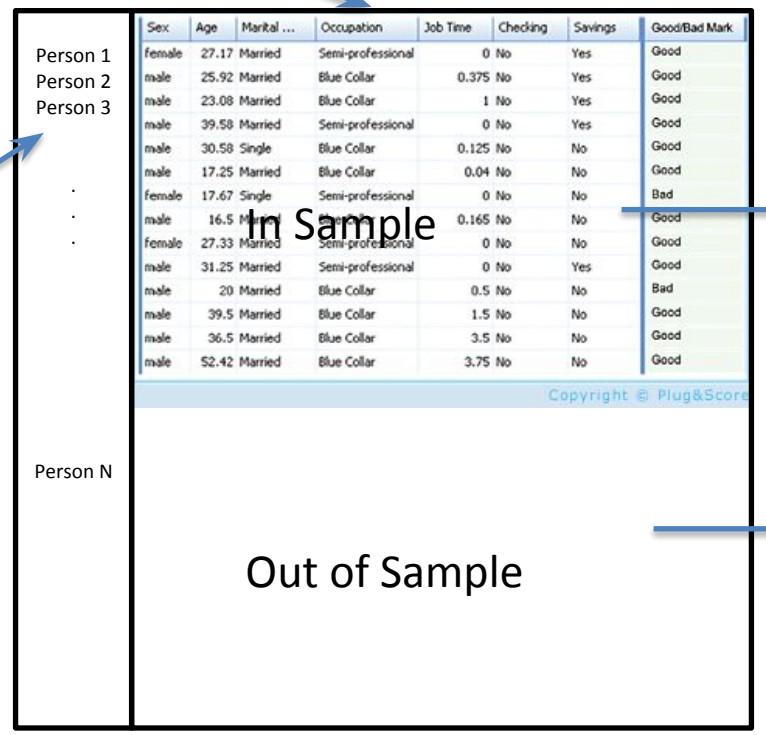
An ML High Level Framework

In Real Life

- Objects
- Events / Experiments
- People / Customers
- Products
- Stocks
- …

Features

Some data has observed results



Sex	Age	Marital …	Occupation	Job Time	Checking	Savings	Good/Bad Mark
female	27.17	Married	Semi-professional	0	No	Yes	Good
male	25.92	Married	Blue Collar	0.375	No	Yes	Good
male	23.08	Married	Blue Collar	1	No	Yes	Good
male	39.58	Married	Semi-professional	0	No	Yes	Good
male	30.58	Single	Blue Collar	0.125	No	No	Good
male	17.25	Married	Blue Collar	0.04	No	No	Good
female	17.67	Single	Semi-professional	0	No	No	Bad
male	16.5	Married	Semi-professional	0.165	No	No	Good
female	27.33	Married	Semi-professional	0	No	No	Good
male	31.25	Married	Semi-professional	0	No	Yes	Good
male	20	Married	Blue Collar	0.5	No	No	Bad
male	39.5	Married	Blue Collar	1.5	No	No	Good
male	36.5	Married	Blue Collar	3.5	No	No	Good
male	52.42	Married	Blue Collar	3.75	No	No	Good

- **Characteristics**
- **Patterns**

- **Predictions**
- **Similarities**
- **Differences**
- **Distance**

CS: Table

Math: Matrix **X**, with N rows – each person
m columns, each feature (age, salary, ..)

$$\mathbf{X} = \begin{bmatrix} -2 & 4 & 7 & 31 \\ 6 & 9 & 12 & 6 \\ 12 & 11 & 0 & 1 \\ 9 & 10 & 2 & 3 \end{bmatrix}$$

A Fundamental Idea: From Table to Score

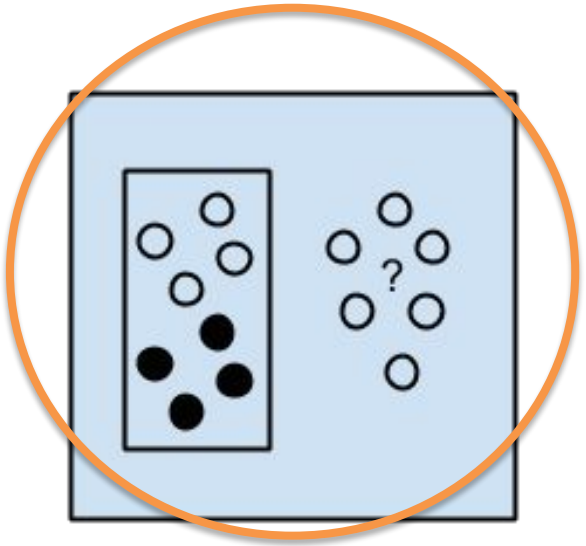
X =

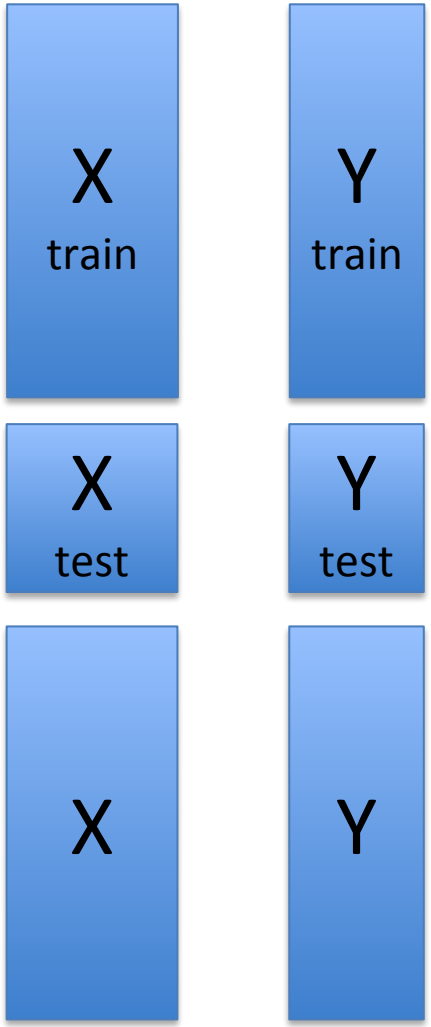
Cust	F1	F2	F3
A	4	2	2
B	4.5	1.5	3
C	3	3	5
D	1	2	2
E	3	1.5	5
F	3.5	3.5	1
..	..	..	..

F(X)

Cust	Credit Score
A	552
B	381
C	760
D	330
E	452
F	678
..	..

Supervised Learning: From Table to Score

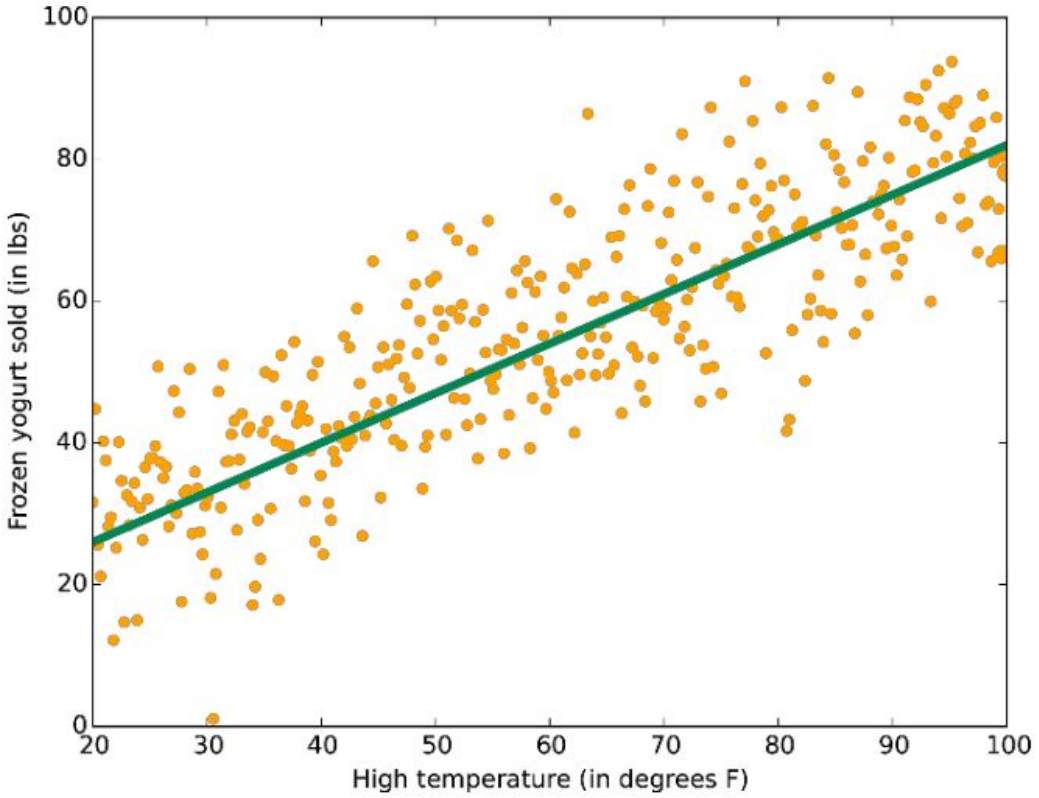




Non-Labeled
Out of Sample

```
#Setting up for Supervised learning
# First clean: use mapping + buckets

# X = matrix of data – e.g 1000 rows
# Y = In sample responses

# Typically we want to split in to
training data and test data

X_train = X[:800] # 800 train samples
Y_train = Y[:800]
X_test = X[800:] # 200 test samples
Y_test = Y[800:]
```

Linear Regression Illustration



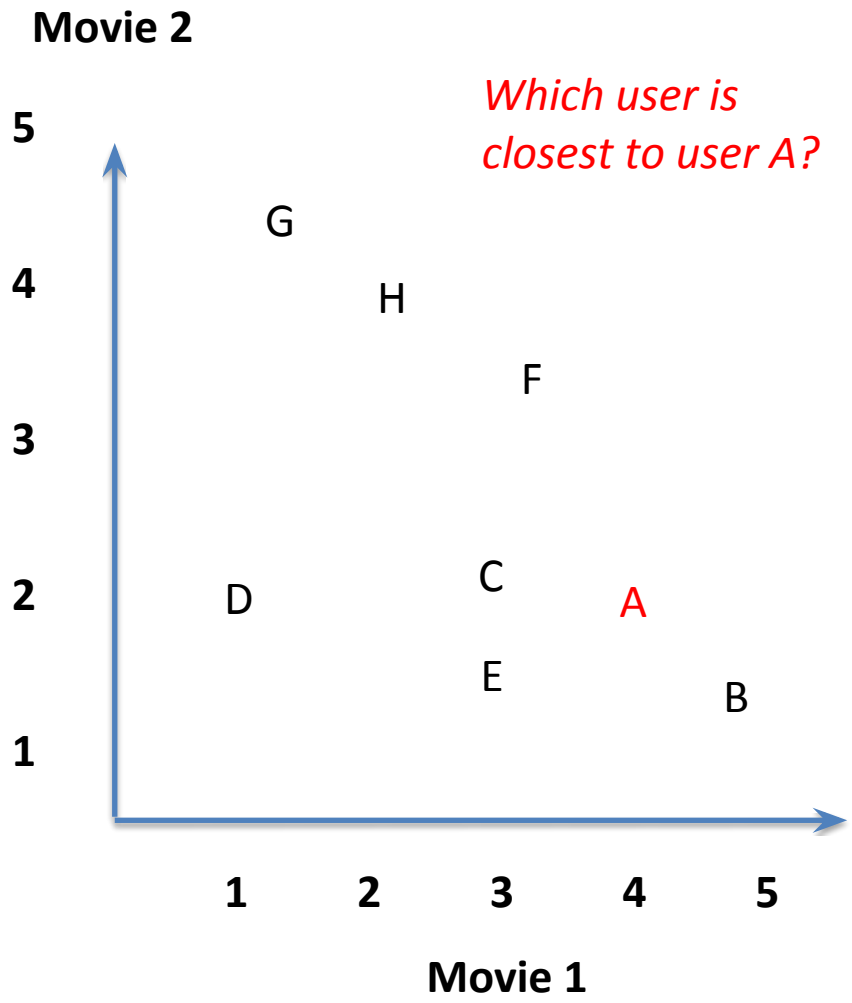

```
#Setting Linear Regression in sklearn
from sklearn import linear_model

model= linear_model.LinearRegression()
model.fit(X_train, Y_train)

Y_pred_train =  model.predict(X_train)
Y_pred_test =  model.predict(X_test)


# Compare Y_pred_test with Y_test for
error.
```

Illustration Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice

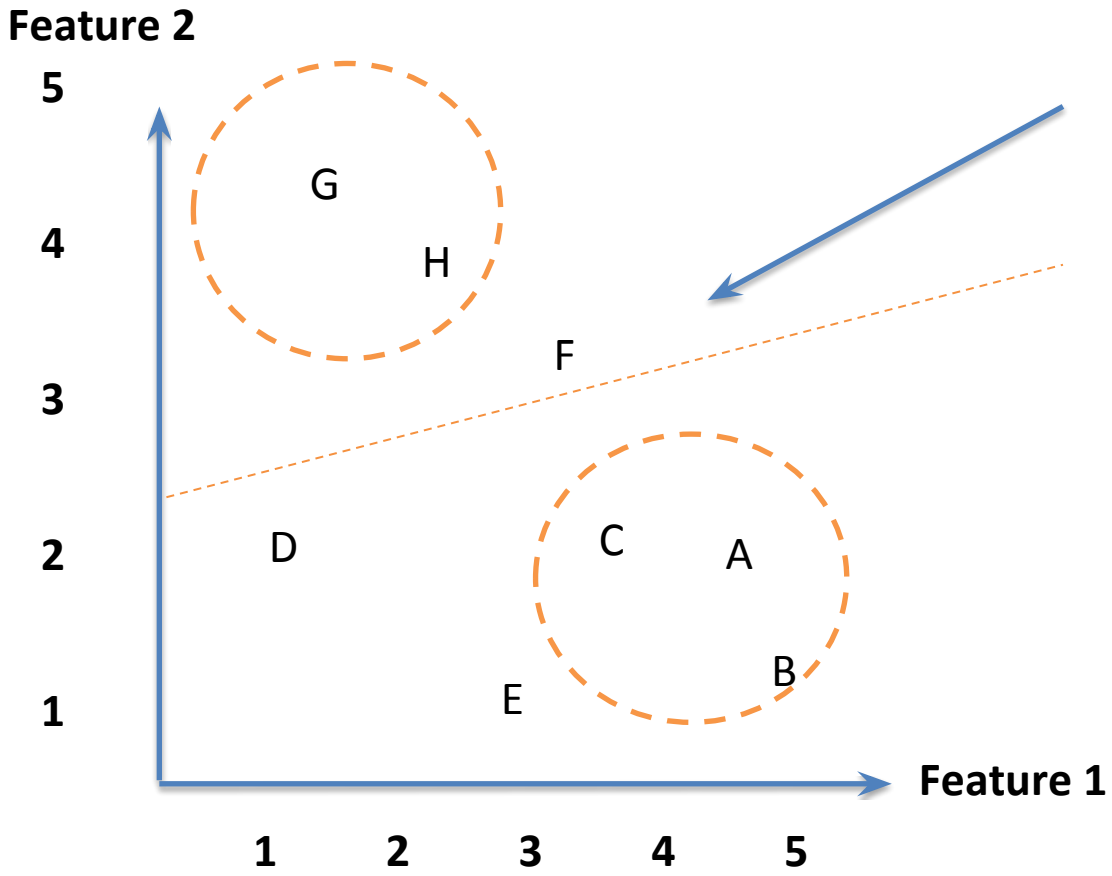
A Fundamental Idea: From Table to N- Dimensional Space

X =

User	F1	F2	F3
A	4	2	2
B	4.5	1.5	3
C	3	2.1	5
D	1	2	2
E	3	1.5	5
F	3.5	3.5	1
..	..	..	..



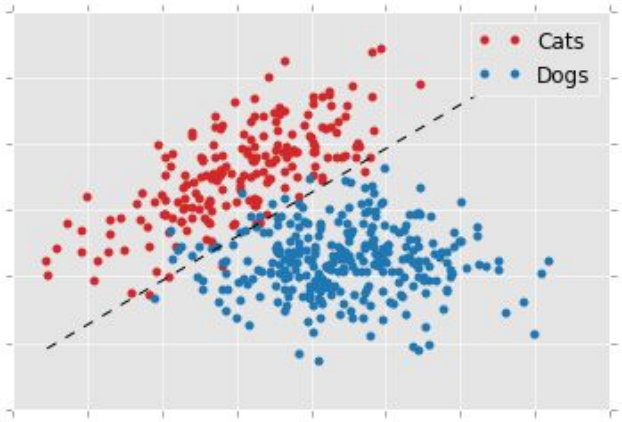
Clustering to Classification



- Target customers?
- Pictures of Cats and Dogs
- Speech recognition
- Recognize Letters: A, B, C..

Traditionally 2 Supervised Tasks: Classification & Predictive Scoring

Extracted Data often in Table Format



Classification:
Cats and Dogs, Speech Recognition
Movie Recommendation

Scoring:
Credit Score, Movie Rating
Heat [illegible]

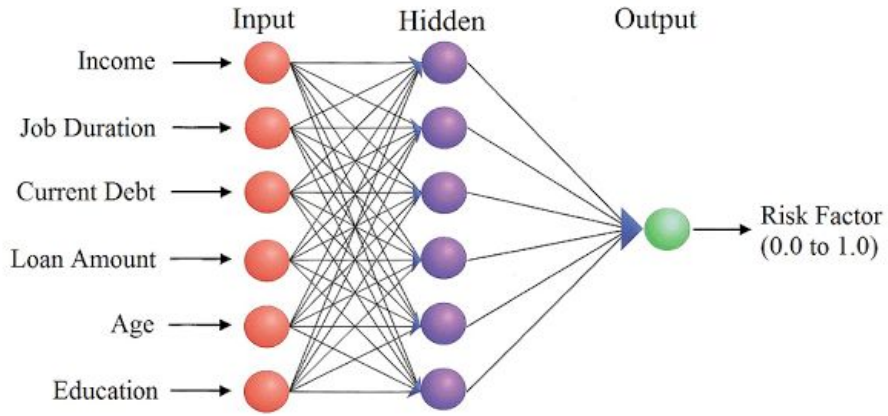
Data X

We have now switched to Neural Networks as Function Approximators

X → ML Algorithms Guess this function F(x) → Y

"Non-deep" feedforward neural network

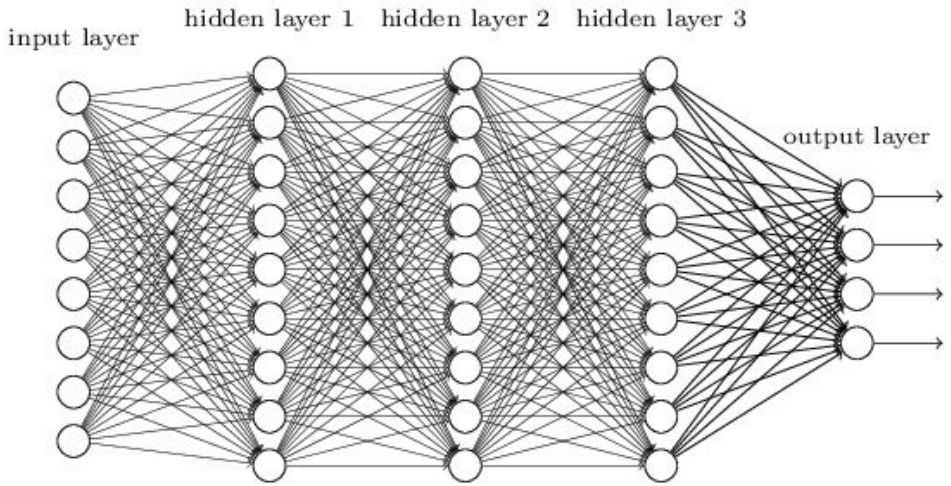
X



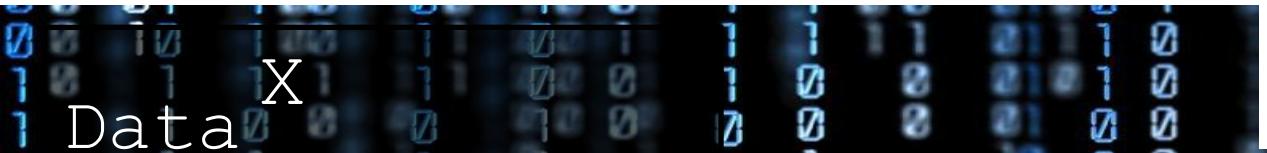
Y

Deep neural network

X



Y



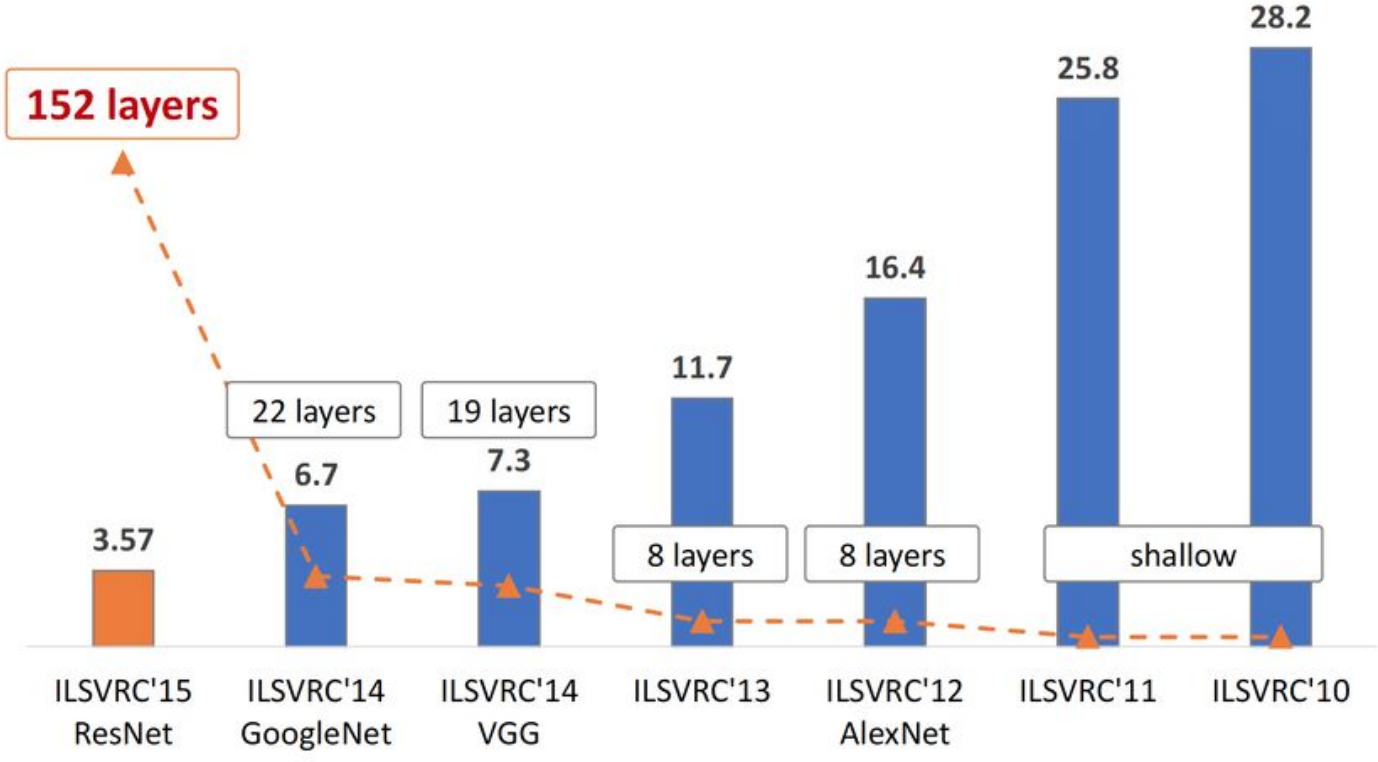
IMAGENET Large Scale Visual Recognition Challenge

Neural nets are now better at classifying images than humans

The Image Classification Challenge:

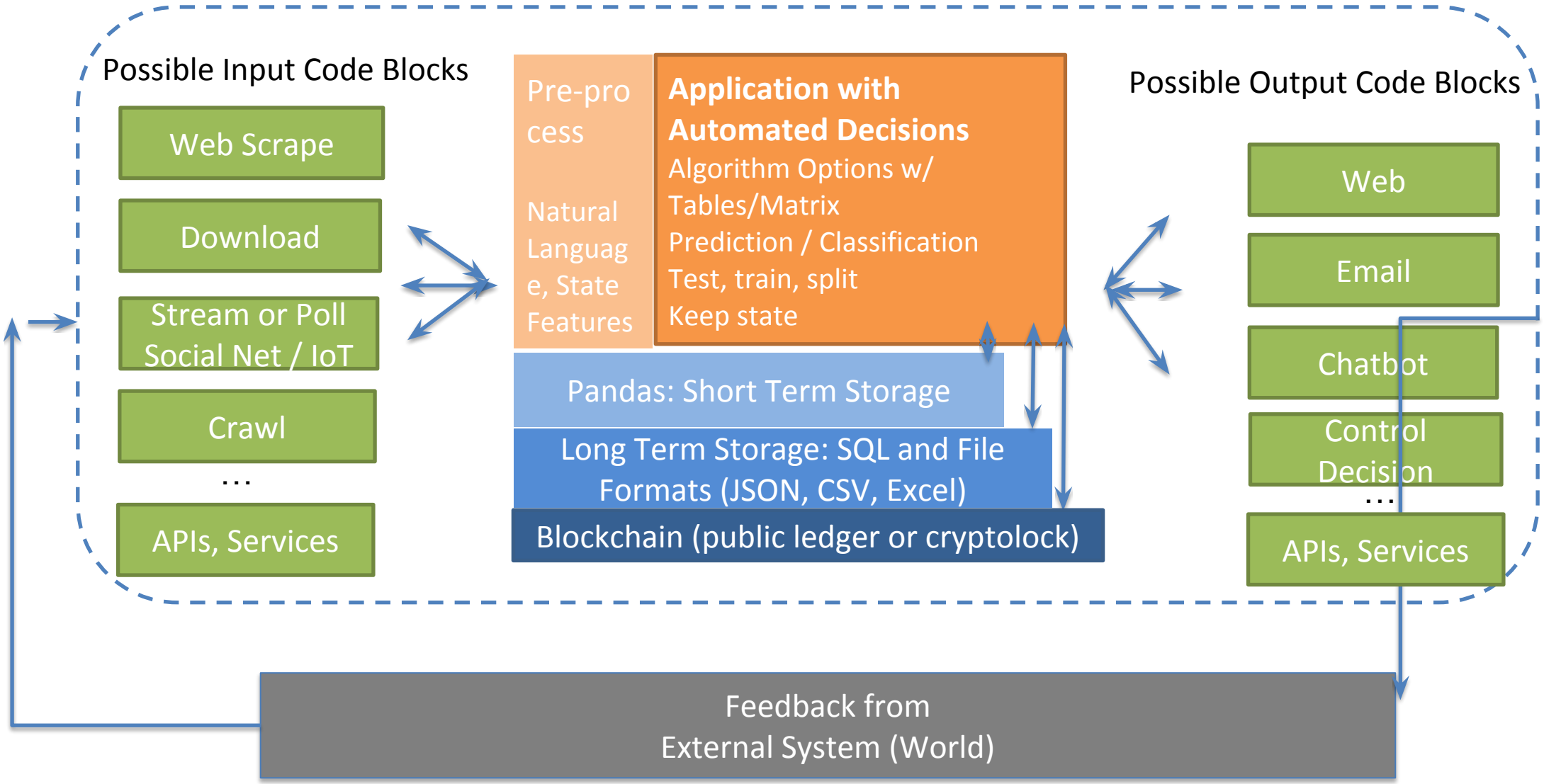
1,000 object classes

1,431,167 images

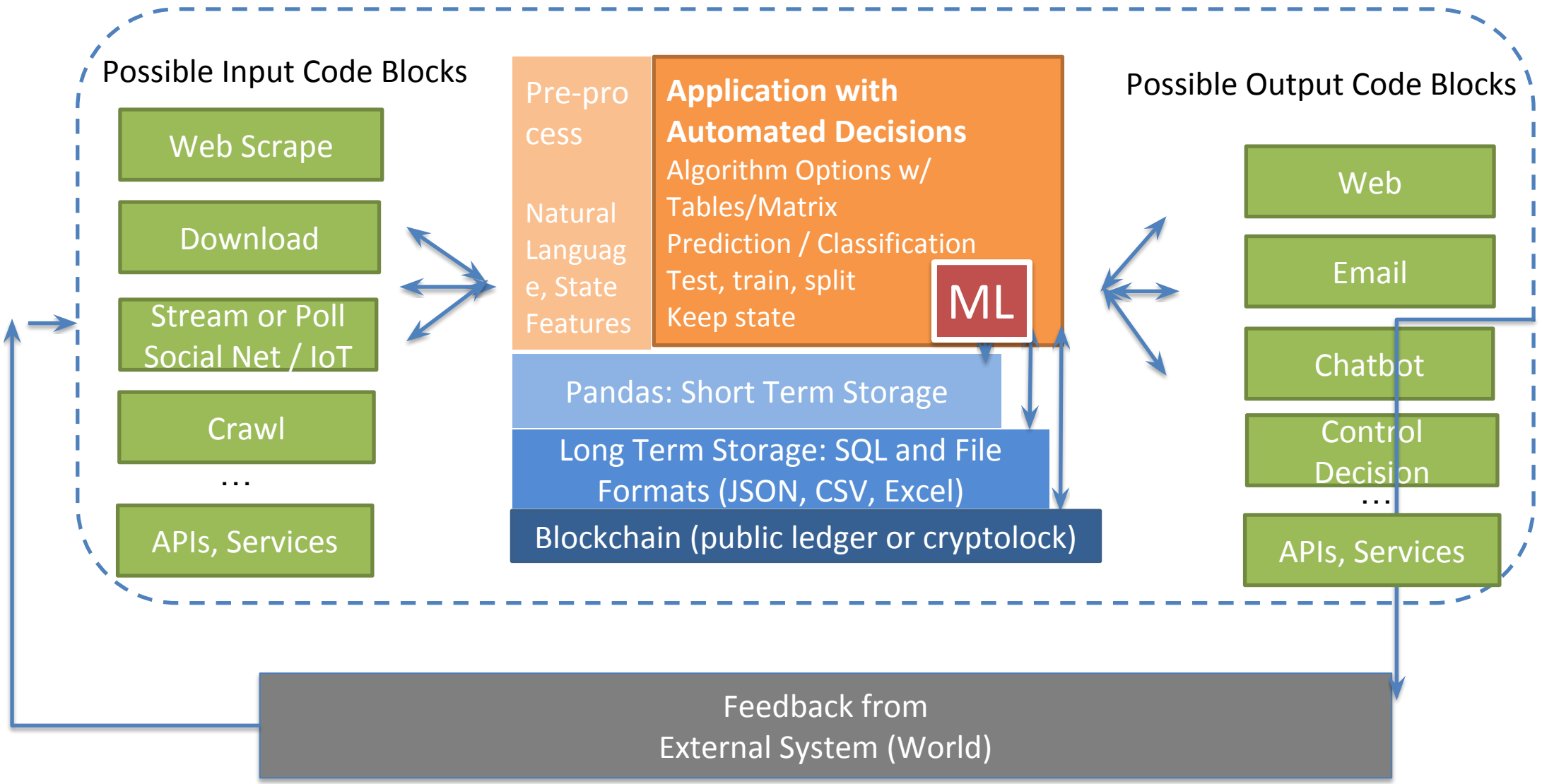


Data X

The Data-X System View



The Data-X System View



Reminder: Homework For Week 1

- **<u>HW Part 1: For Your Project – By Next week</u>**
  - Come up with 3 ideas for projects to pursue in Data-X Lab I in 3-5 sentences.
  - A systems or application you will build
  - ***Communicate:*** *WHO the project is for, WHAT will it do, WHY this is needed/valuable.*

- **<u>Homework Part II</u>**
  - Python-based review notebook (only Python concepts week 1)

End of Section