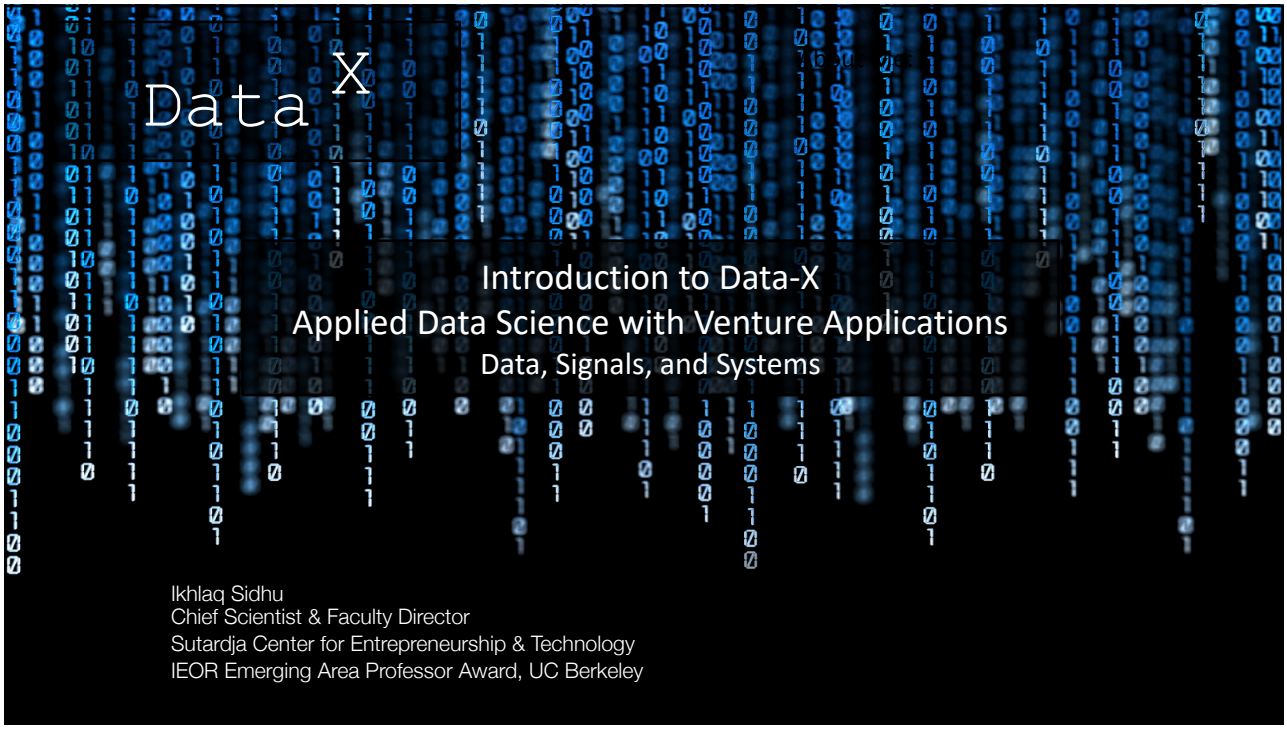


Welcome to Applied Data Science with Venture Applications

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

Course History:

- *Advanced project course for data science systems*
- *Concepts: theory, tools, project.*
- *IEOR 135 – Undergrad, IEOR 290 – Grad*
- *Currently impacted, but we are working on scale and offering every semester*

Prerequisite: Students should have:

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

Welcome to Applied Data Science with Venture Applications

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

Teaching Team

Lead Faculty:
Ikhlaq Sidhu
Arash Nourian

GSIs and Notebook Areas:
Lillian Dong – Fundamental Data Science Tools
Ishaan Malhi – Systems
Zhi Li – Natual Language Processing and Special Topics
Deirdre Quillen – Deep Learning

Project Lead and Mentor :
Nattaphol Vimolchalao

Data X

Agenda on Day 1

Today:

- Course Introduction High Level Overview of Data, and Data-X Project (1:20 min)

By this week:

- Get your Notebook/development environment working
- Develop initial project ideas
- HW assigned
- Answer Questions

Key Dates:
Class Starts today
Final projects: During end of class during reading week
Top project will showcase at Collider Cup event

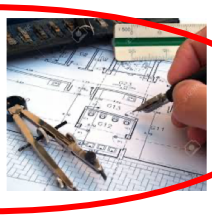
Data X

Course Philosophy

Data-X

Make the Tools

Use State-of-the-Art Open Source Tools



Architect the System

Sell, market, and pitch the product

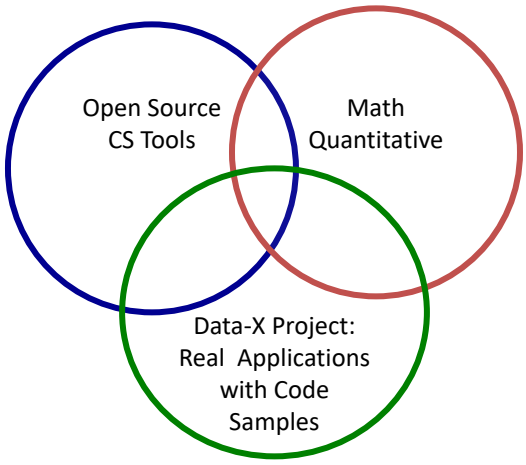
Most CS / Math

Data-X

Business Topics

What is in this course

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



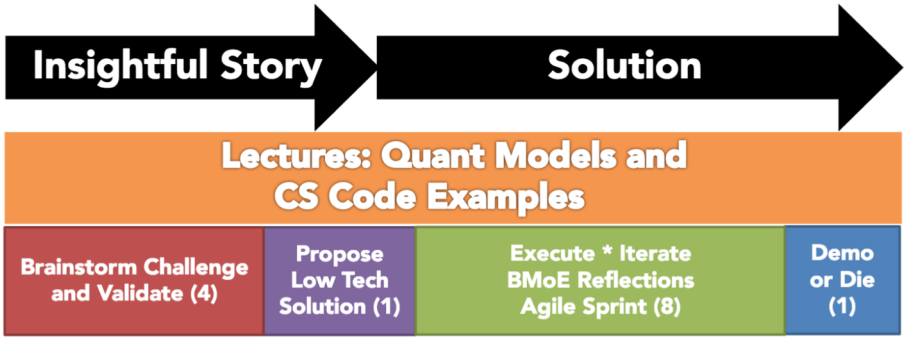
Holistic Perspective: Industry, Social Applications, Customer Driven

What is taught in the class?

- 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.)
- System modeling for data applications
- Application examples: Recommender systems, Blockchain, Spark etc.

Data X

Course Overview

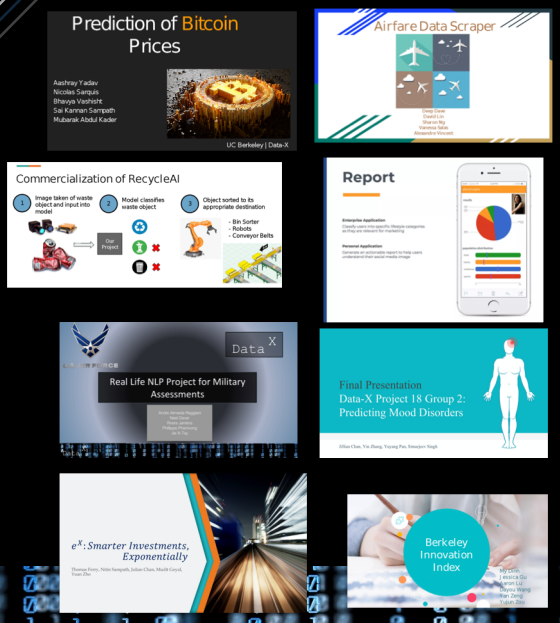


Open-ended, real-world project: Typically 5 students, with available advisor network

Data X

Data-X Project Examples

- 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



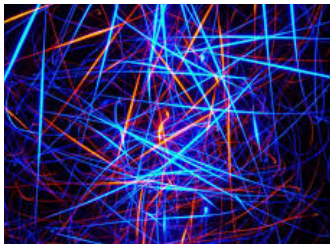
find many, many more at:

Data X www.data-x.blog/projects/

Project Types

Business or Consumer Use Case

Social Impact



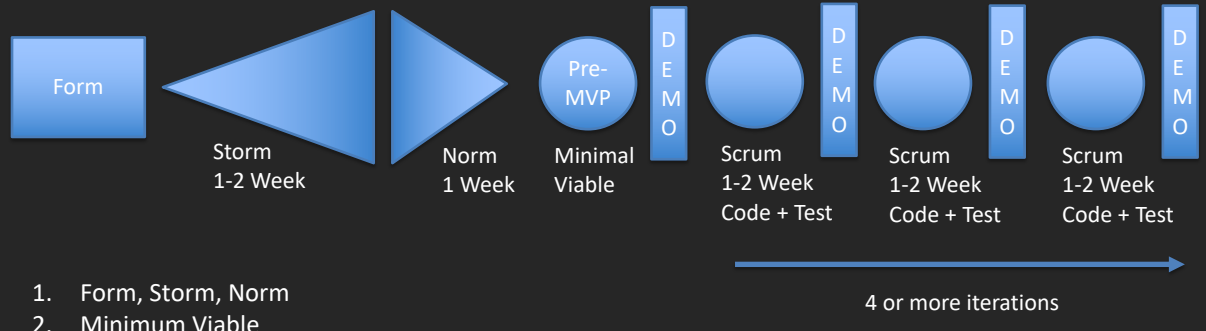
Its Just Cool

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

Project Categories:
- Completely by student team
- Industry proposed projects
- Data-X library

Data X

Implementation Behaviors and Process



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

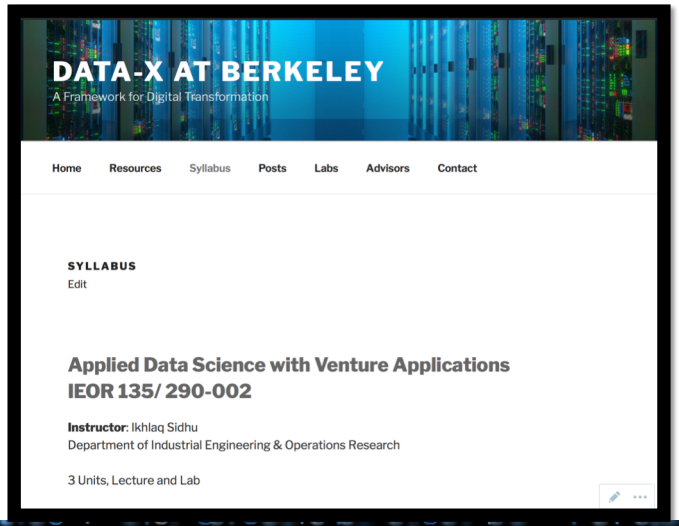
Data X

Many Course Resources Are Already Available at data-x.blog

For students and mentors

See data-x.blog

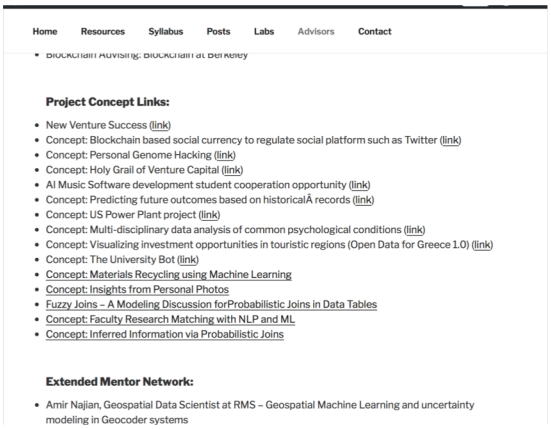
- Lectures and Slides
- Code Samples
- Articles and Readings
- Projects
- Mentors



Data X

Project Ideation

- 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

Homework For Week 1

Project Guidelines: https://data-x.blog/project-guideline/

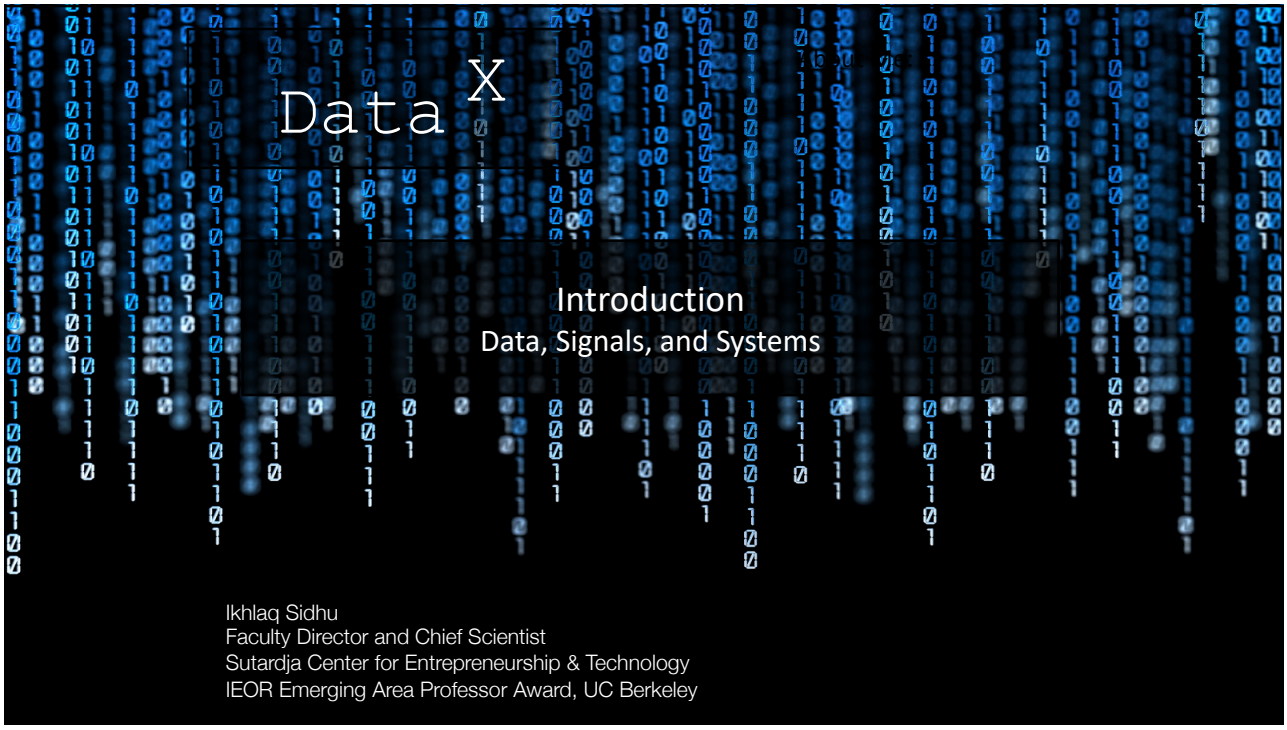
By Next Week

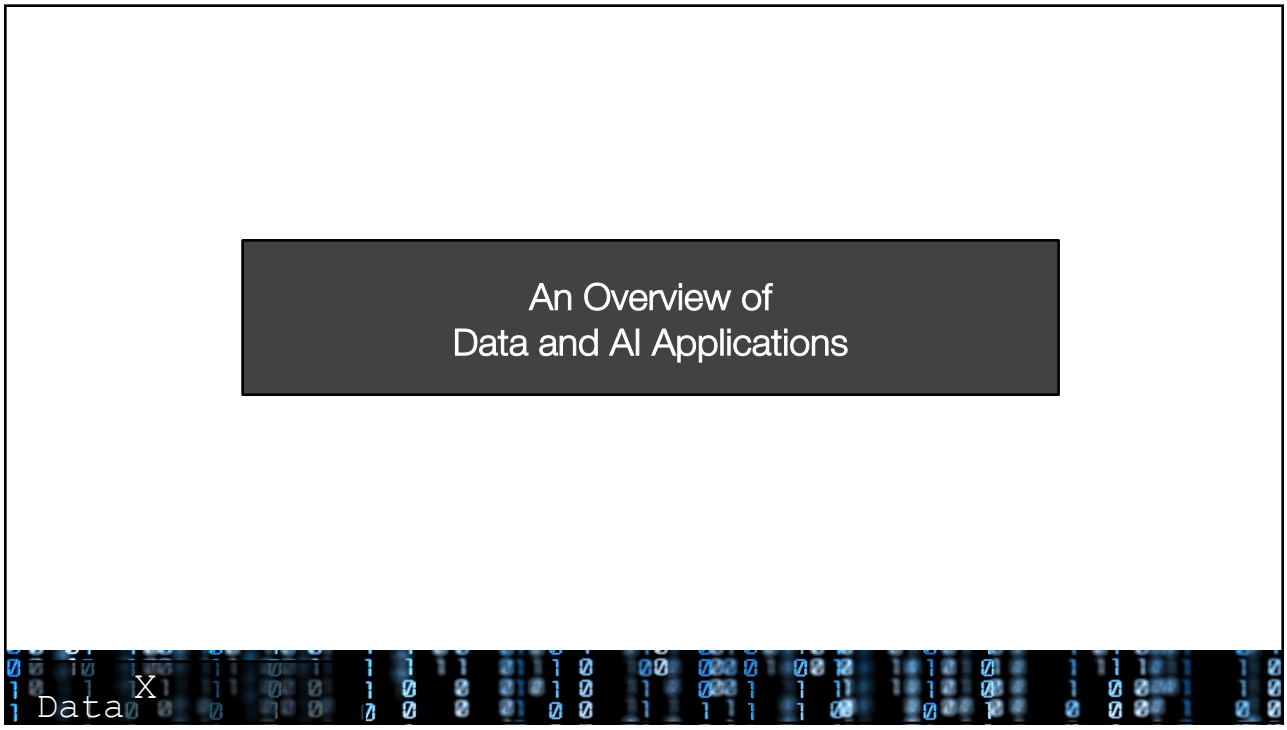
- Project Assignment: write down 3 ideas a project, 1-3 sentences each. Due within one. Submit with a google form here and shared spreadsheet to that everyone can see the concepts submitted. (5 points)
  - A systems or application you will build
  - **Communicate:** WHO the project is for, WHAT will it do, WHY this is needed/valuable.
- Reading: Chapter 1 and Caviar case. Write ½ page to 1 page reflection about which concepts from the chapter and case might be most relevant for your project. You may also critique concepts. Due in one week. (5 points)

Code and Theory HW:

- To be assigned by Arash Nourian

Data X


Data X
Introduction
Data, Signals, and Systems
Ikhlaq Sidhu
Faculty Director and Chief Scientist
Sutardja Center for Entrepreneurship & Technology
IEOR Emerging Area Professor Award, UC Berkeley


An Overview of
Data and AI Applications
Data X

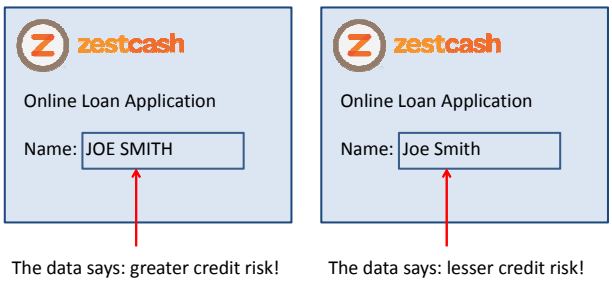
Basic Concept of Working with Data

- Data Wrangling
- In Production

Example: Data and Information is a competitive advantage

Real-life Example: ZestCash

- "All data is credit data"



The data says: greater credit risk!

The data says: lesser credit risk!

Reference: Shomit Ghose

Why: More Simply

Customer Insight/ Engagement	Operations: Reliable & Predictable	Security & Fraud
TARGET, Harrah's	GE, UPS, Walmart	Compliance 360°, Financial Firms, Network Security

Data X

Implementation: SW Tools / Stack

Data X

The Most Common Open Source Tools: AI/ML Stack

Start with Python as an interface

Jupyter Notebooks for prototyping

- Python: The interface
- NumPy, SciPy: Working with Arrays
- Pandas: Working in Tables, SQL to Pandas
- Sklearn: ML
- Matplotlib: Visualizing Data
- TensorFlow, Keras: Neural Networks
- SQL to Pandas
- NLP / NLTK: Natural Language
- Spark: For large data sets (GB, TB+)

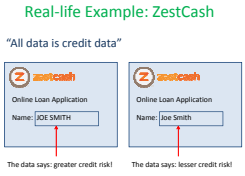
https://www.youtube.com/watch?v=Q0jGAZAdZqM

https://conda.io/docs/user-guide/install/download.html

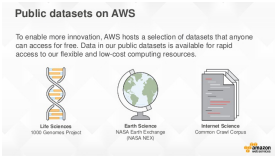
Data X

Where Does Data Come From?

Where Does Data Come From?

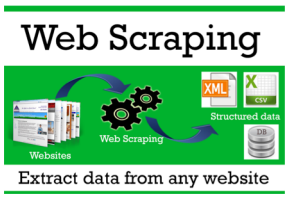


Your Own Web Site



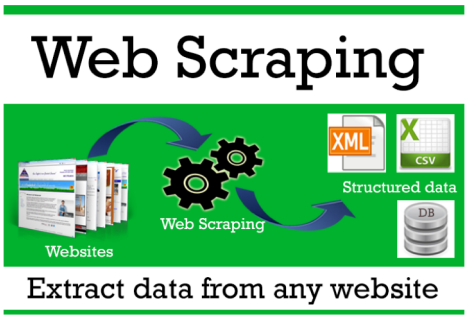
Public Data Sets
Stock market, etc.

IOT/Sensors



Other Web Sites

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'})
```

https://github.com/ikhlaqsidhu/data-x

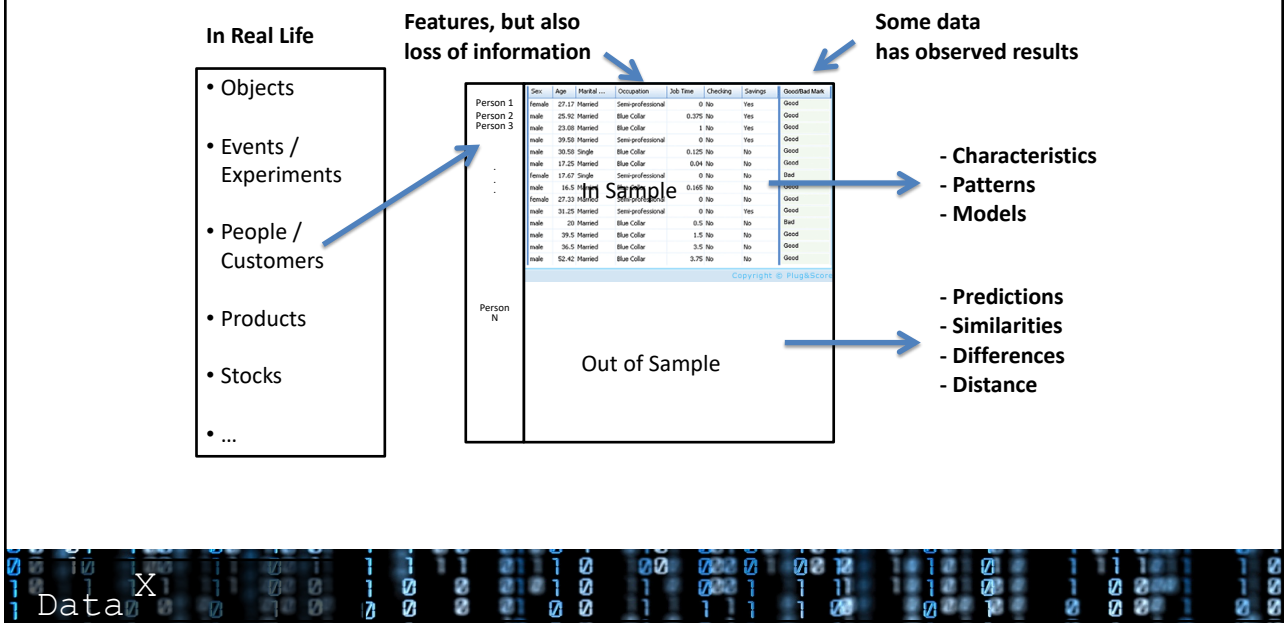
https://github.com/ikhlaqsidhu/data-x/tree/master/03-tools-webscraping-crawling_api_afo

Data X

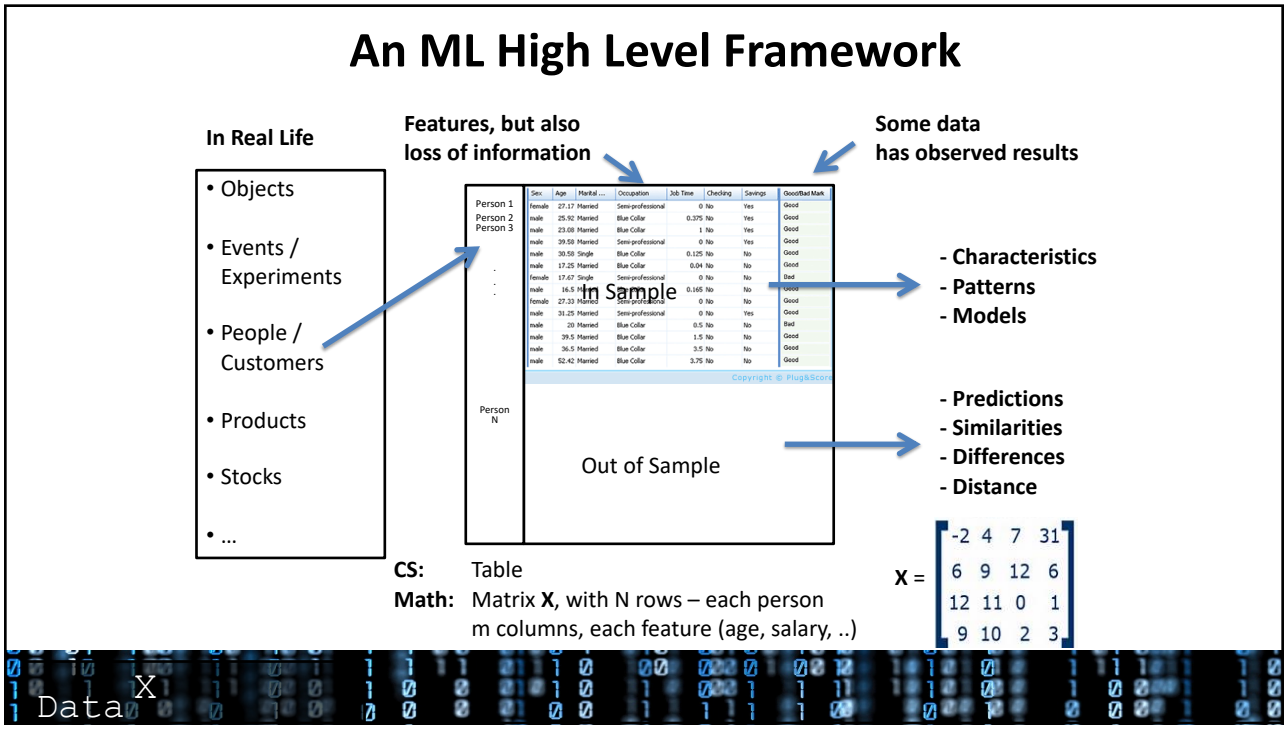
Formatting Data

Data X

An ML High Level Framework



An ML High Level Framework



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
..	..

Data X

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
..	..	..	..

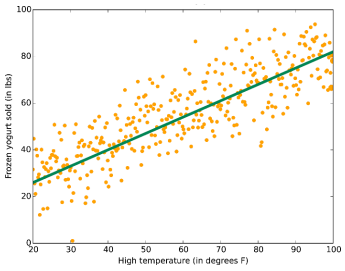
F(X) →

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

```
#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[0:500]
Y_train = Y[0:500]
X_test = X[501:1000]
Y_test = Y[501:1000]
```

Data X

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.
```

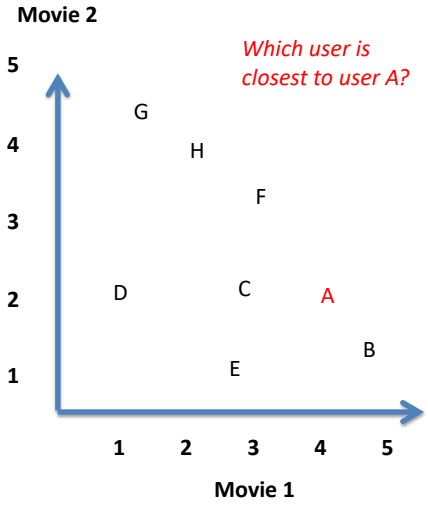
Illustration Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice

Data X

A Fundamental Idea: From Table to N- Dimensional Space

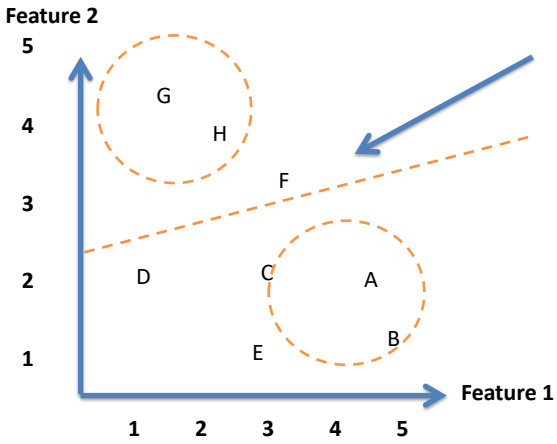
X =

Element	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
..	..	..	..



Data X

Clustering to Classification

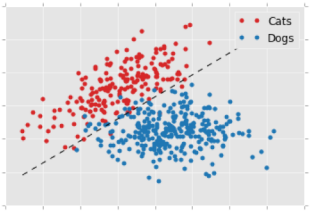


- Target customers?
- Pictures of Cats and Dogs
- Speech recognition
- Recognize Letters: A, B, C..

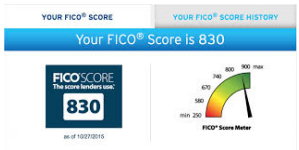
Data X

Traditionally 2 Tasks: Classification & Predictive Scoring

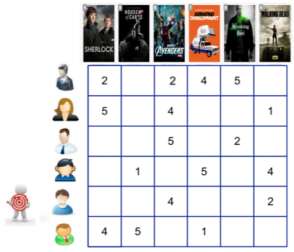
Extracted Data often in Table Format



Classification:
Cats and Dogs, Speech Recognition
Movie Recommendation



Scoring:



The most famous application has been recommendation: "which other user is most like you"

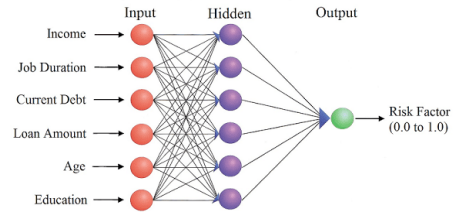
Data X

We have now switched to Neural Networks as Function Approximators

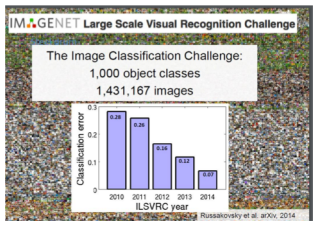
X → ML Algorithms Guess this function F(x) → Y

"Non-deep" feedforward neural network

X

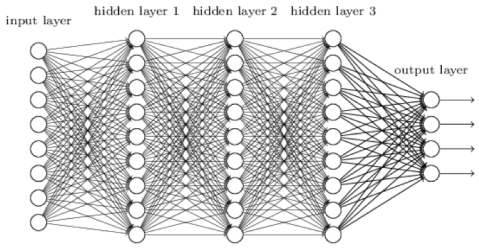


Y



Deep neural network

X



Y

Data X

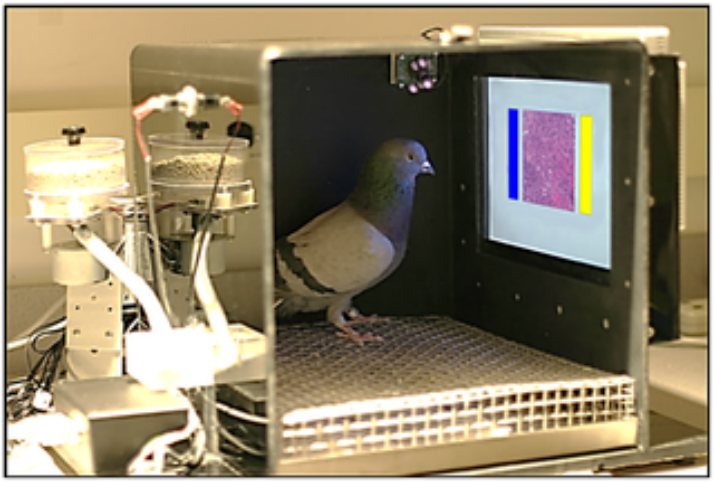


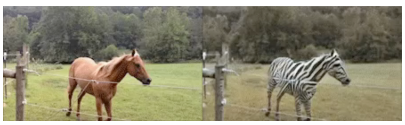
Fig 1. The pigeons' training environment.

Data X

Current Hot Topics in AI/Data

Reinforcement Learning

Unsupervised Image to Image

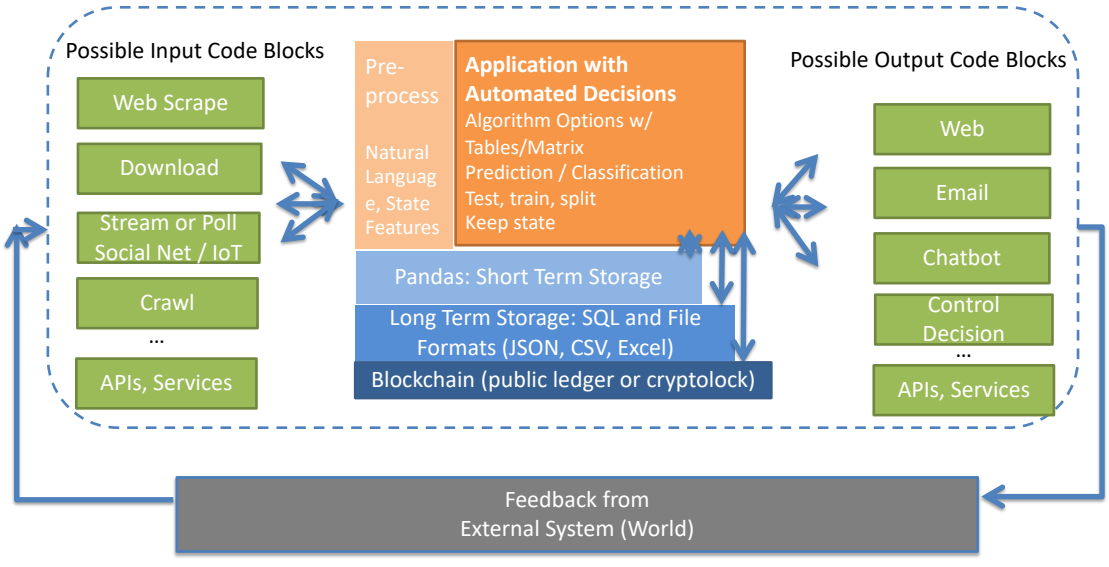


[CycleGAN: Zhu, Park, Isola & Efros, 2017]

Generative Networks and Deep Fakes

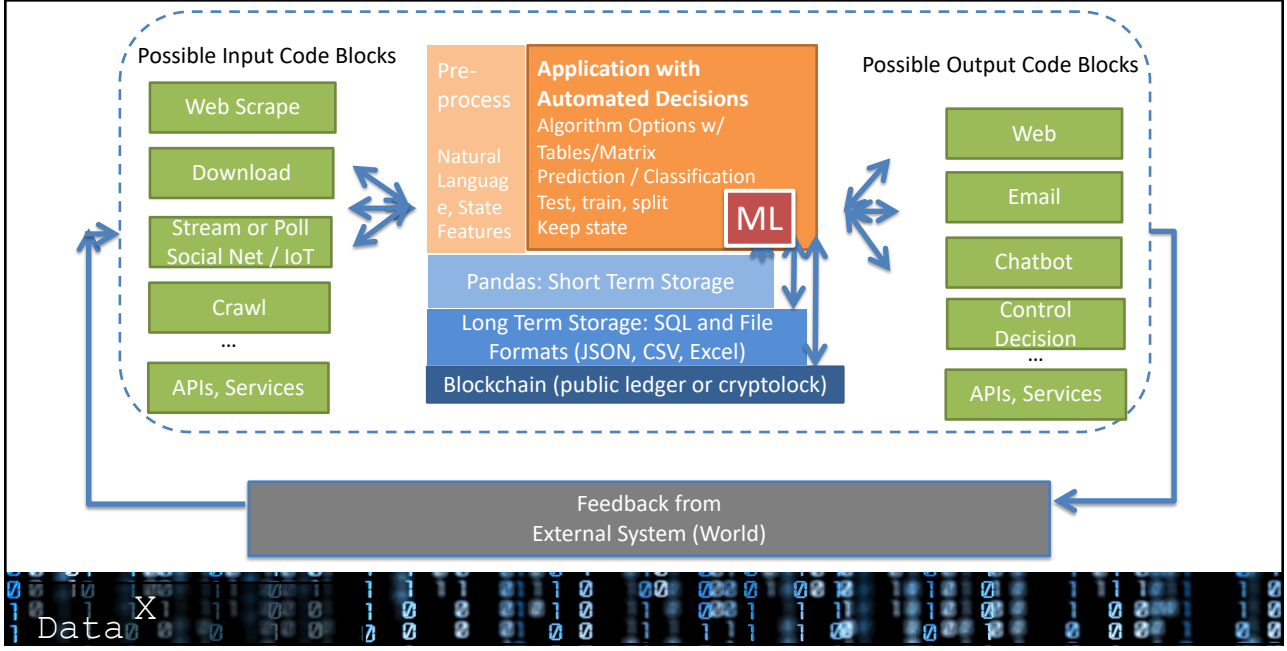
Data X

The Data-X System View



Data X

The Data-X System View: It's more than ML, it's also systems and models



Data X

Our Agenda and Purpose

Project First	Theory, Frameworks System Architecture	Implementation Tools ML Stack, Code	Innovation Behaviors and Process

Questions and/or Discussion to Get Started

Data X

End of Section