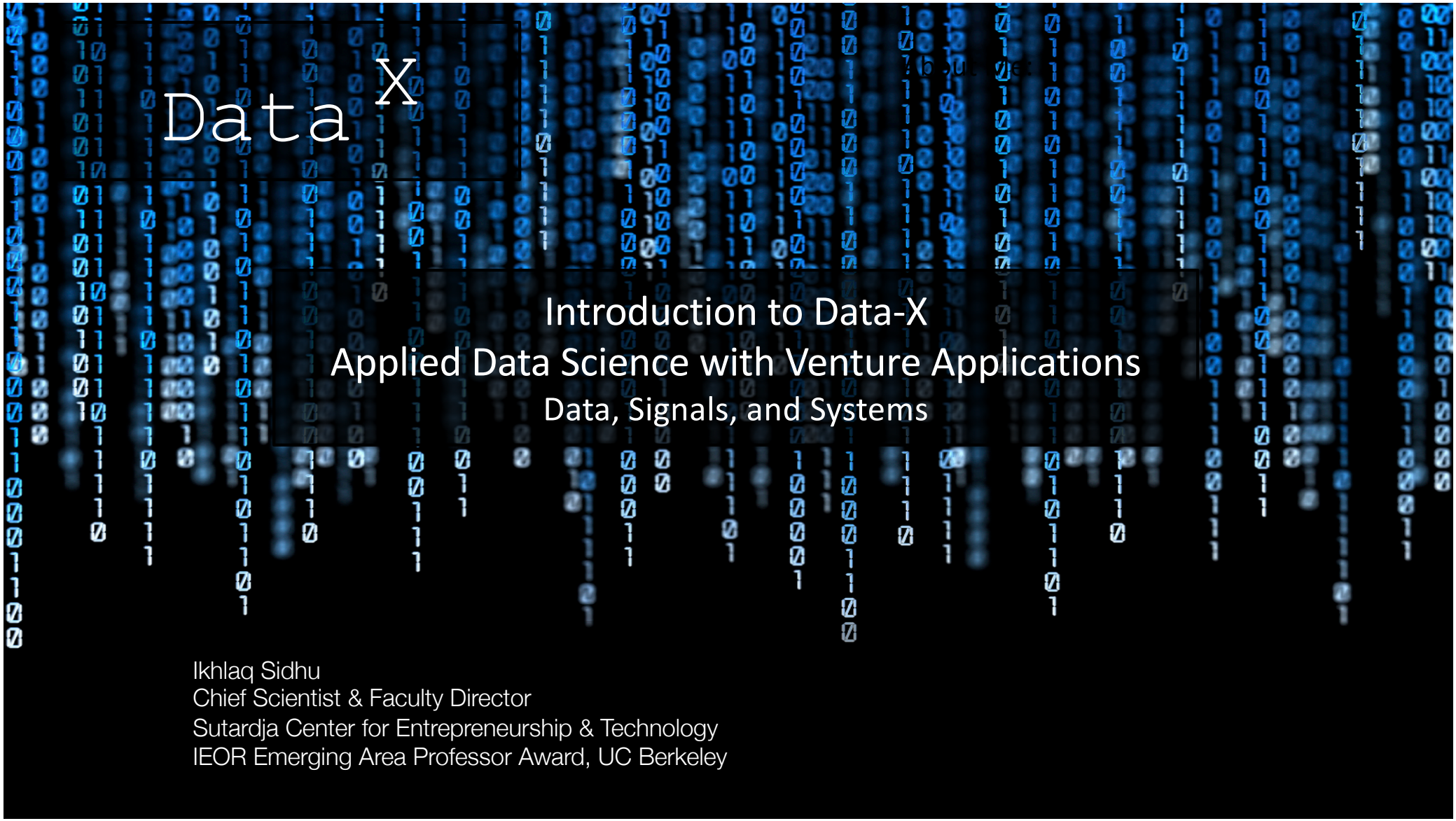


Introduction to Data-X

Applied Data Science with Venture Applications

Data, Signals, and Systems

Ikhlaq Sidhu
Chief Scientist & Faculty Director
Sutardja Center for Entrepreneurship & Technology
IEOR Emerging Area Professor Award, UC Berkeley

Welcome to Applied Data Science with Venture Applications

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

Teaching Team

Course History

Prerequisite: Students should have:

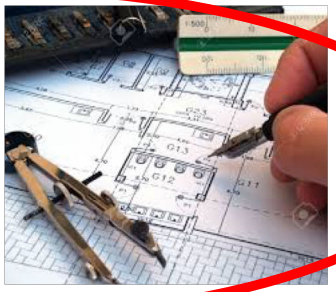
- a working knowledge of Python
- completed a fundamental probability / statistics course
- basic understanding Linear Algebra.

Slowly Converting: Flip Model and More In-class Project Discussion

Course Philosophy

Make the Tools

Use State-of-the-Art Open Source Tools



Architect the System

Sell, market, and pitch the product

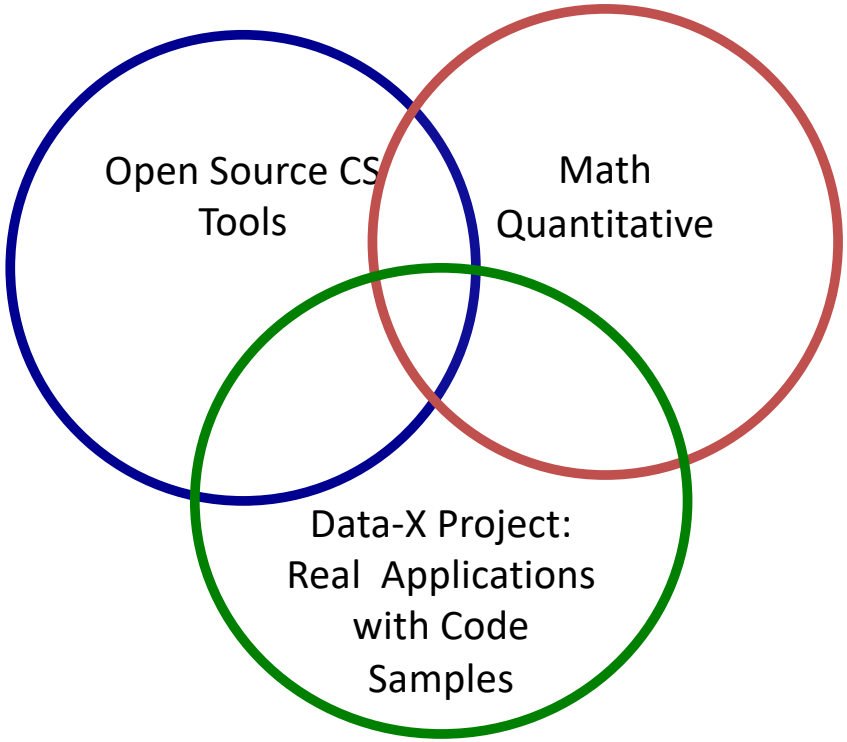
Most CS / Math

Data-X

Business Topics

What is in this course

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

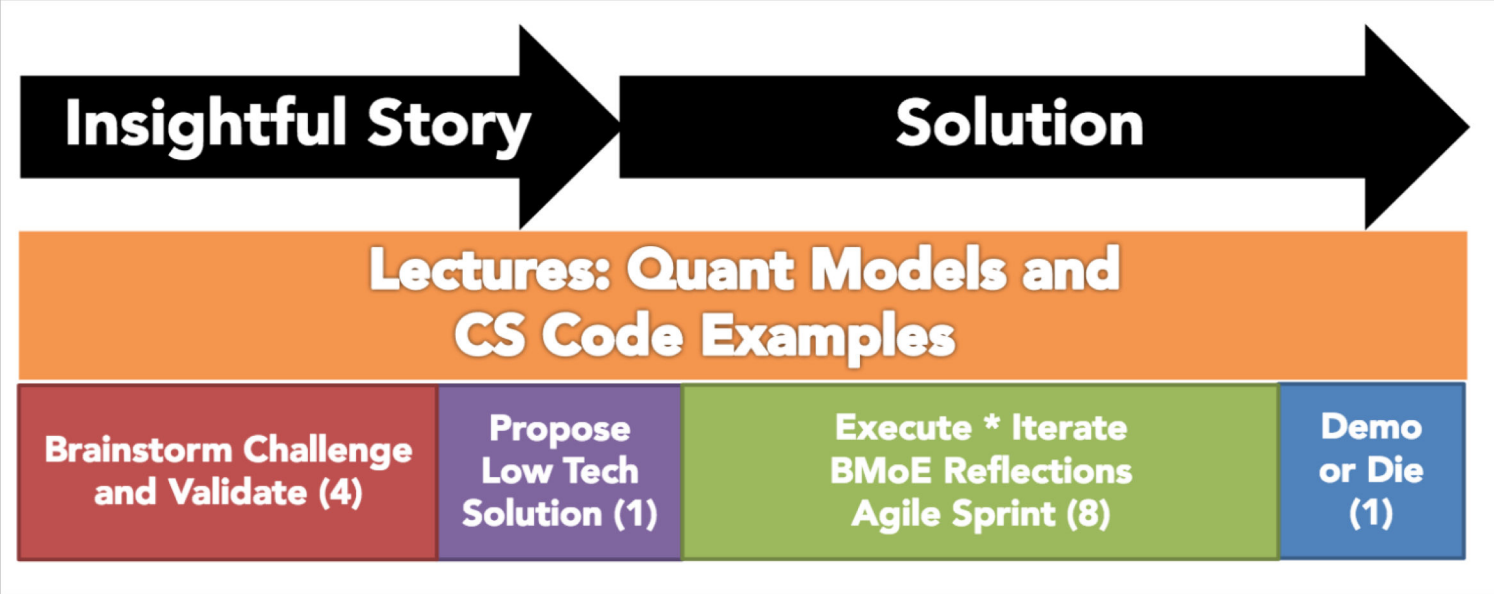


Holistic Perspective: Industry, Social Applications, Customer Driven

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 talks: Recommender systems, Blockchain, Spark etc.

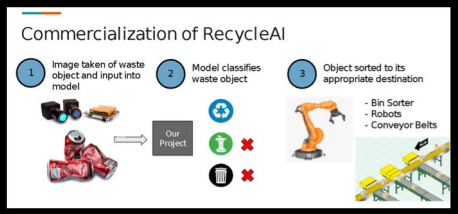
Course Overview

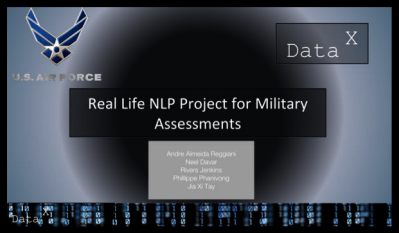


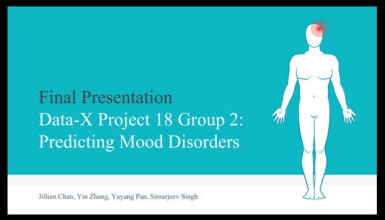
Open-ended, real-world project: Typically 5 students, with available advisor network

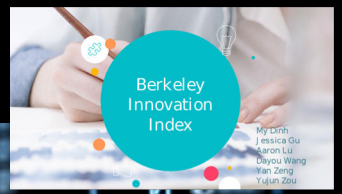
Data-X Project Examples

- 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:

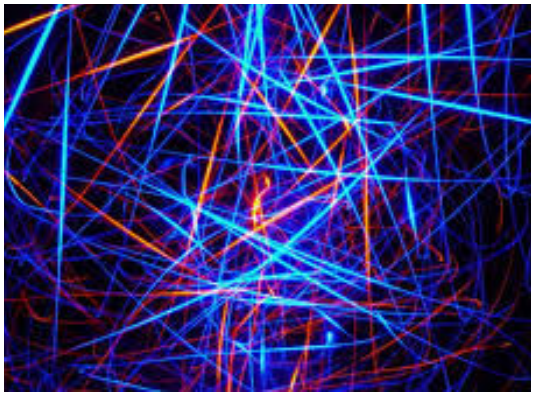
Data X www.data-x.blog/projects

Project Types

Business or Consumer Use Case

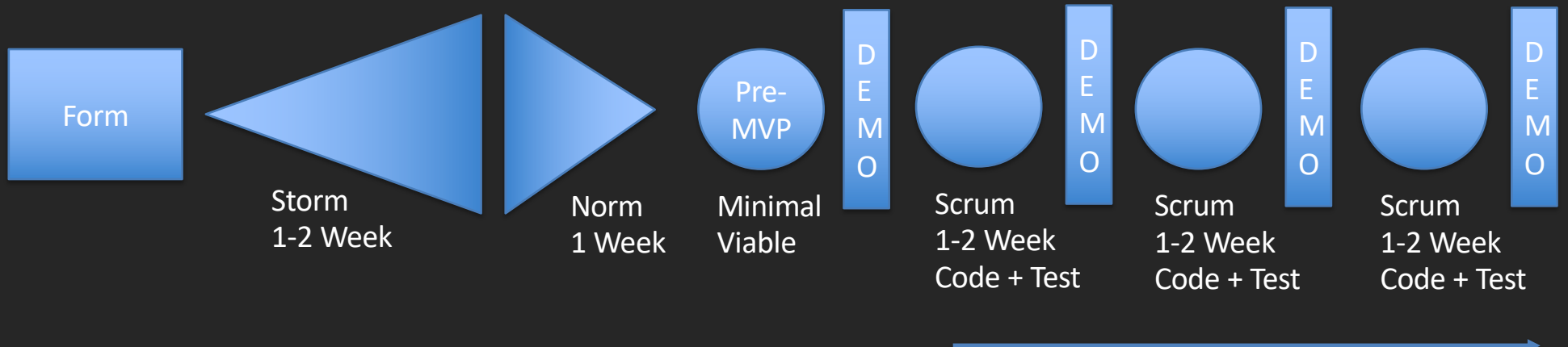
Social Impact

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



Its Just Cool

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

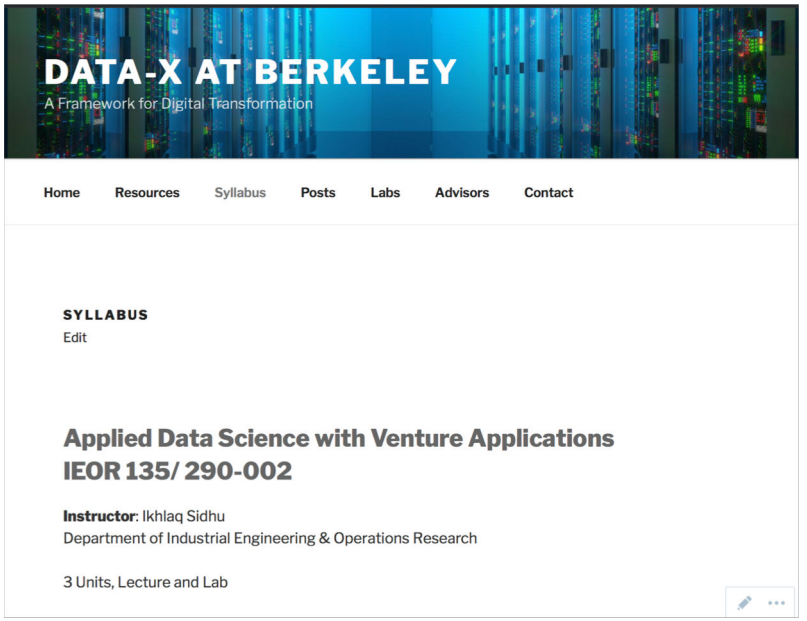
Most Resources Are Available at data-x.blog

1. Go to Data-X.blog
   - Syllabus
   - Instructions for SW Install
   - Link to GitHub with Cookbook Code Samples and Slides

2. Download Instructions to Install Python 3.x Anaconda Environment. For now you only need Anaconda, don't worry about other packages that are not already included.

3. Be able to create your own Jupyter notebook

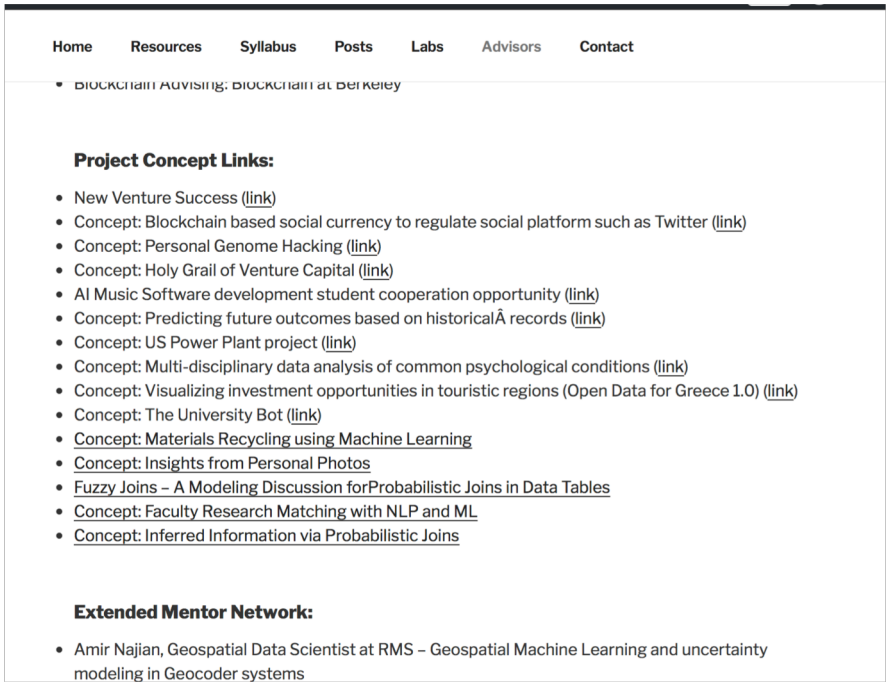
4. Self-Review Python references as needed. See Ref CS01 and as needed BIDS Python Bootcamp.



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

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

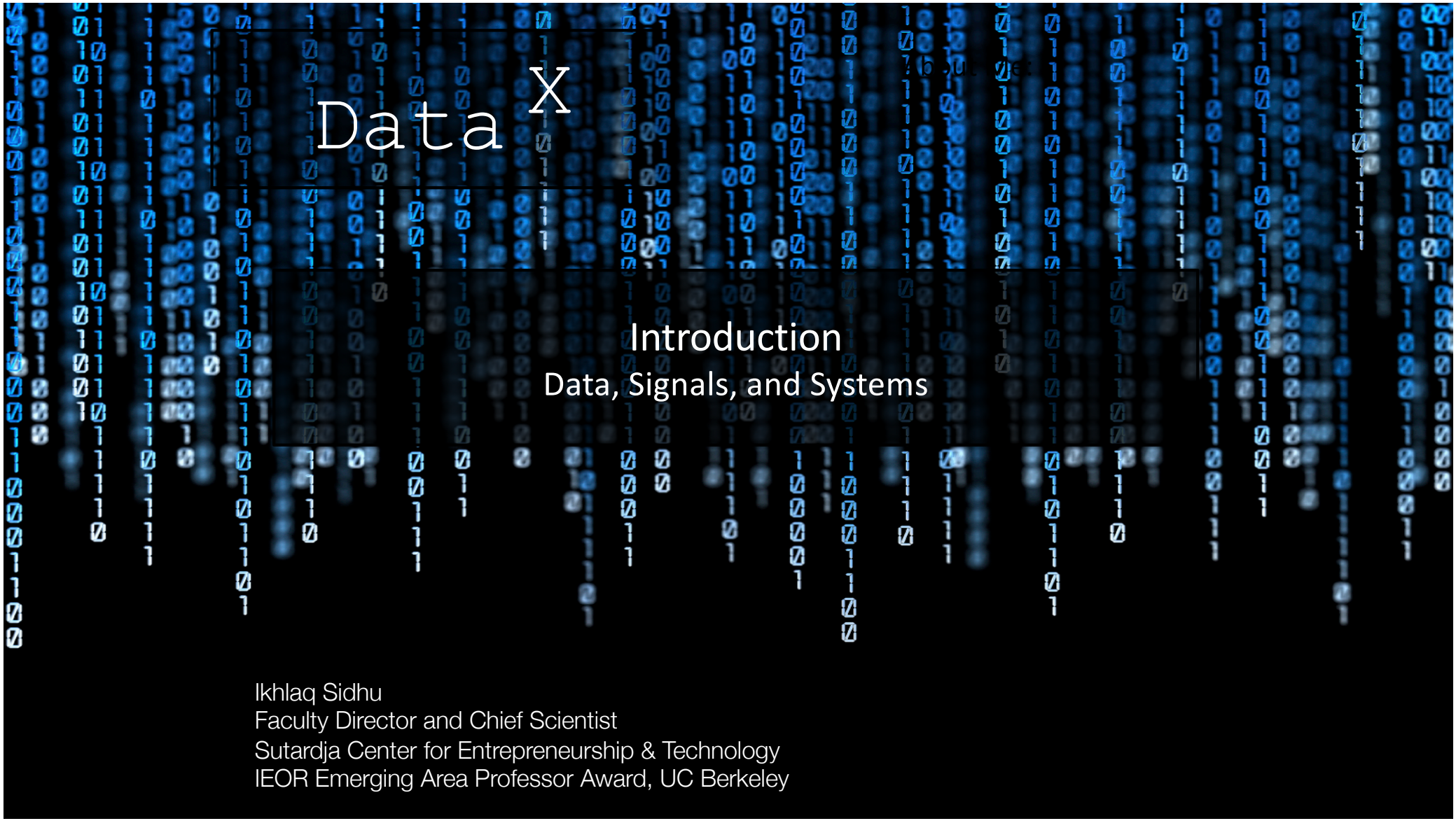


Homework For Week 1

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

- ---

- Homework Part II
- Python-based review notebook (Breakout and Homework, BKHW). To sent by email.



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

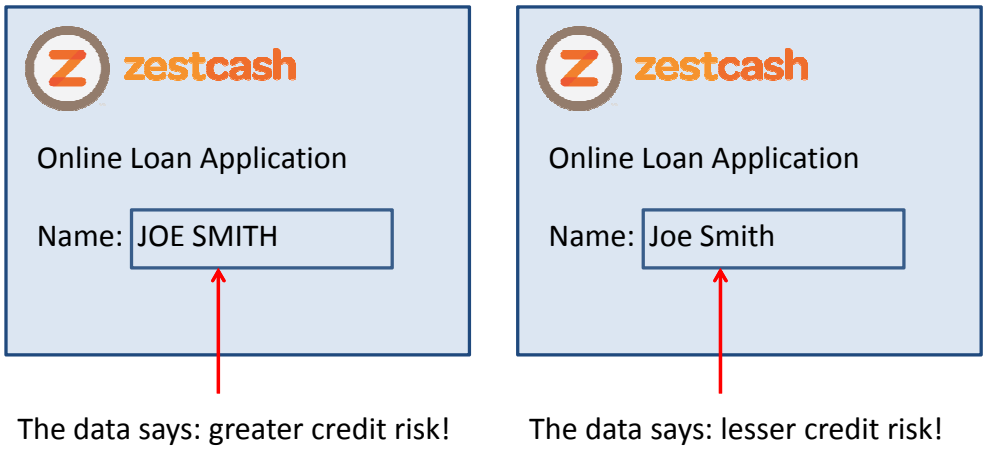
Basic Concept of Working with Data

- Data Wrangling
- In Production

Example: Data and Information is a competitive advantage

Real-life Example: ZestCash

- "All data is credit data"



Reference: Shomit Ghose


• Service provider of Gambling and Casinos
• Entry Card
• Pain points
• Intervention
Harrah's Casino: Knowing your customer
PLAY & WIN
Reference: Supercrunchers

Why: More Simply

Customer Insight/ Engagement

Operations: Reliable & Predictable

Security & Fraud

TARGET

Harrah's

Walmart

Compliance 360

Financial Firms

Network Security

DataX

Implementation:
SW Tools / Stack

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

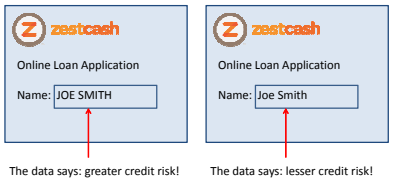
https://conda.io/docs/user-guide/install/download.html

Where Does Data Come From?

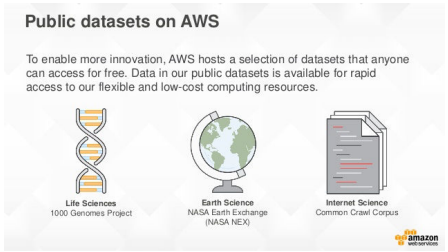
Where Does Data Come From?

Real-life Example: ZestCash

- "All data is credit data"

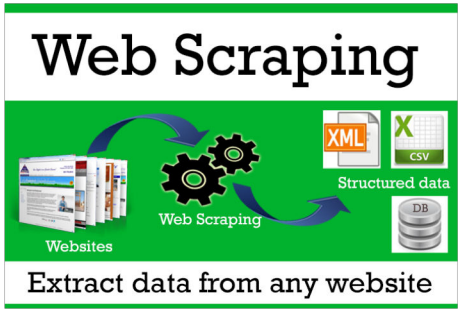


Your Own Web Site



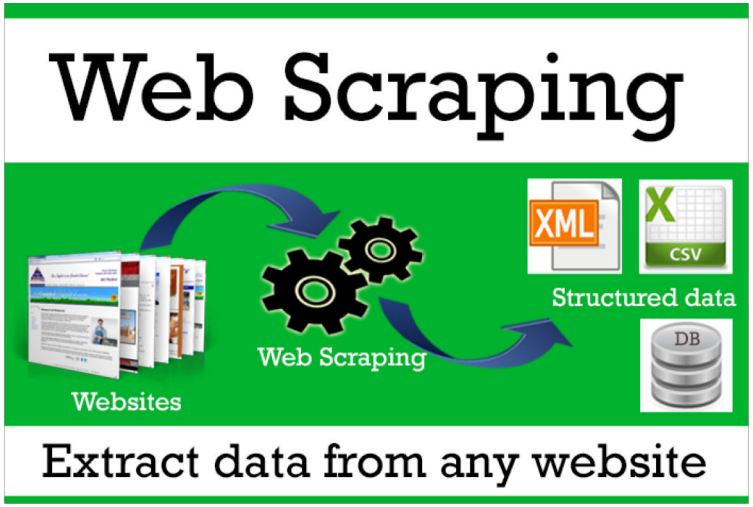
Public Data Sets
Stock market, etc.

IOT/Sensors



Other Web Sites

Web Scraping



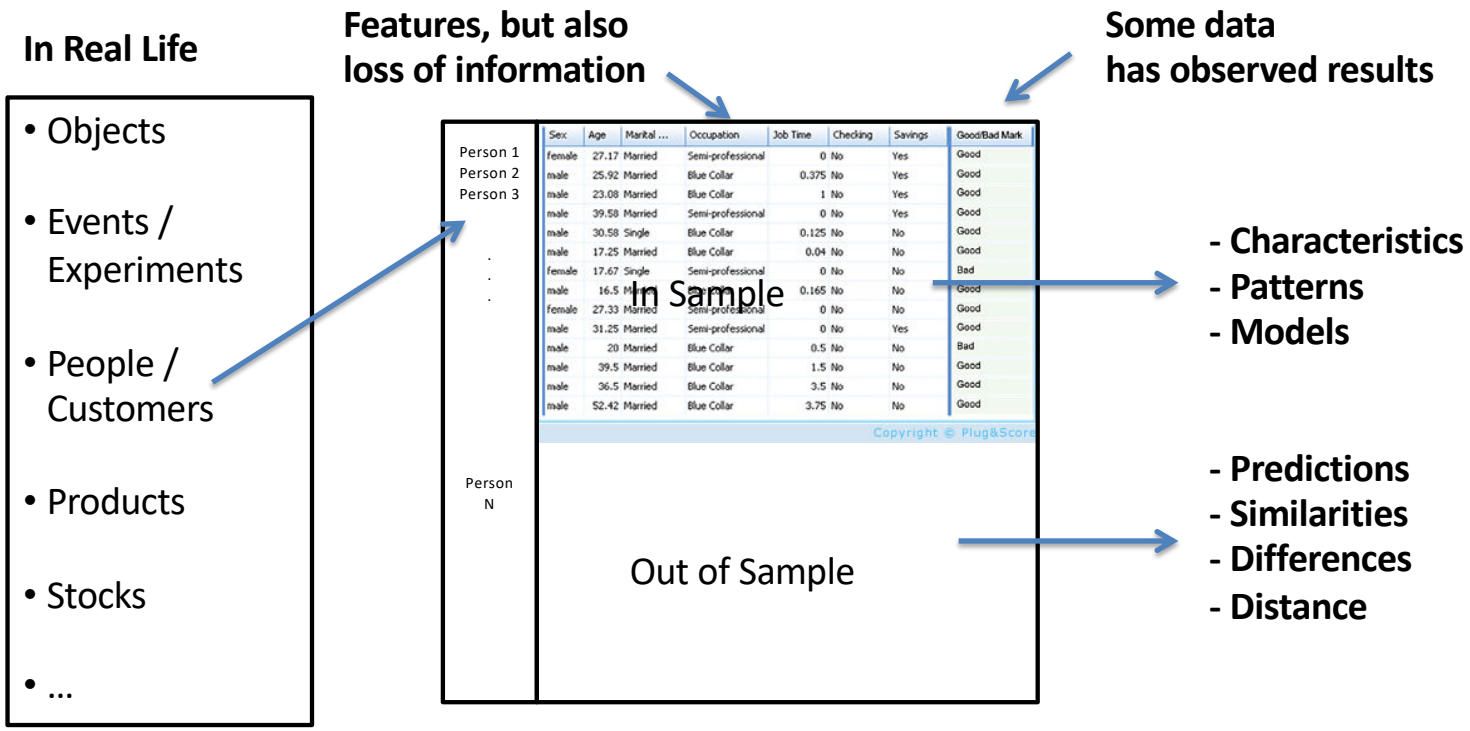

```python
from bs4 import BeautifulSoup
import requests
page_link ='https://www.website_to_crawl.com'
# fetch the content from url
page_response = requests.get(page_link, timeout=5)
# parse html
page_content = BeautifulSoup(page_response.content, "html.parser")

# 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

Formatting Data

An ML High Level Framework



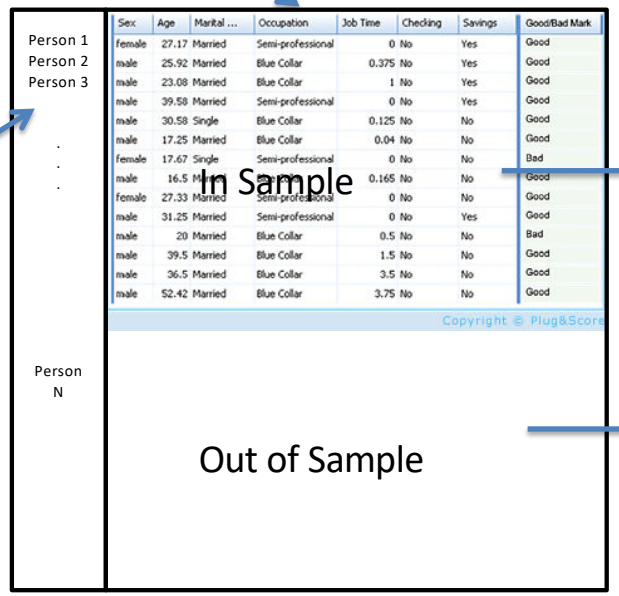
An ML High Level Framework

In Real Life

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

Features, but also loss of information

Some data has observed results



- Characteristics
- Patterns
- Models

- 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
..	..

A Fundamental Idea: From Table to Score

X =

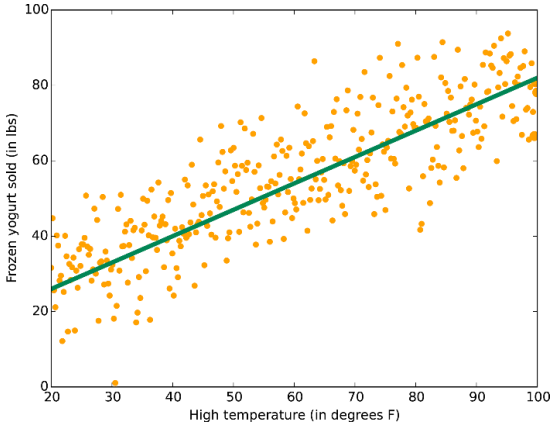
Cust	F1	F2	F3
A	4	2	2
B	4.5	1.5	3
C	3	3	5
D	1	2	2
E	3	1.5	5
F	3.5	3.5	1
..	..	..	..

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]
```

Linear Regression Illustration



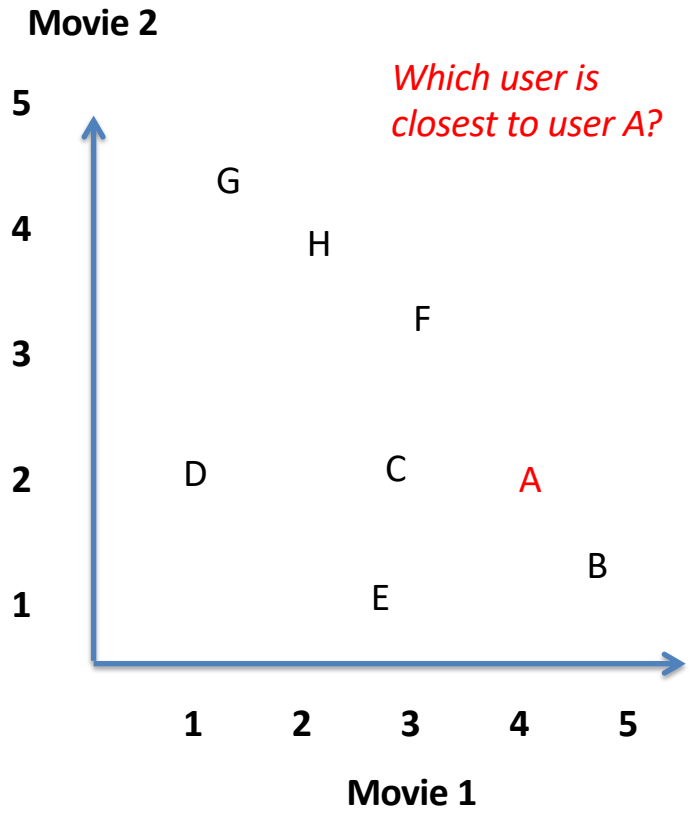

```
#Setting Linear Regression in sklearn
from sklearn import linear_model

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

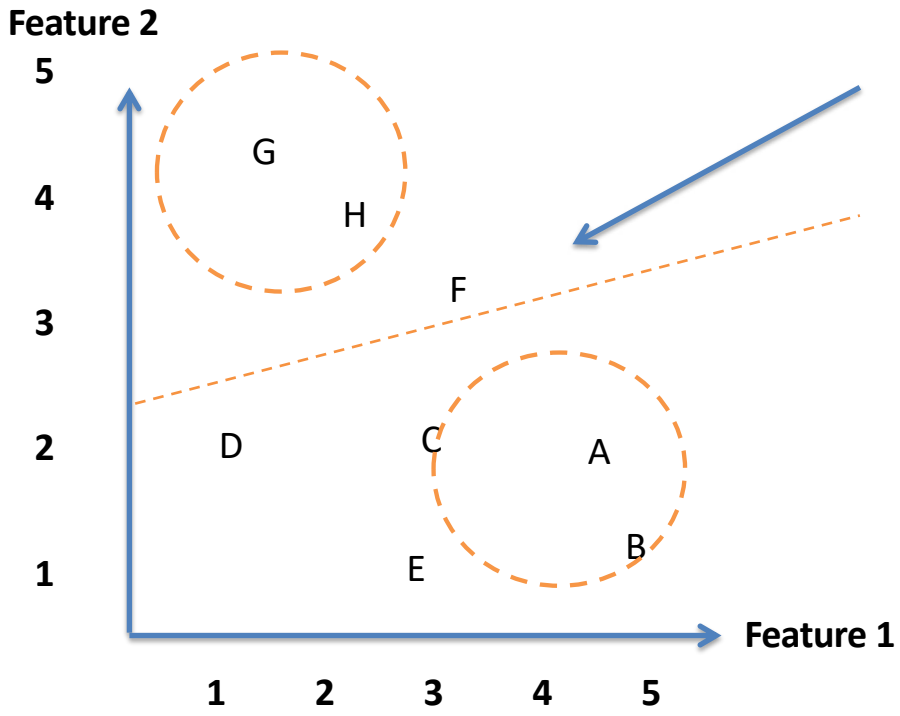
A Fundamental Idea: From Table to N- Dimensional Space

X =

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
..	..	..	..



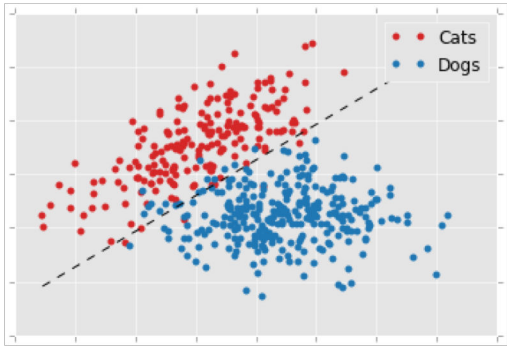
Clustering to Classification



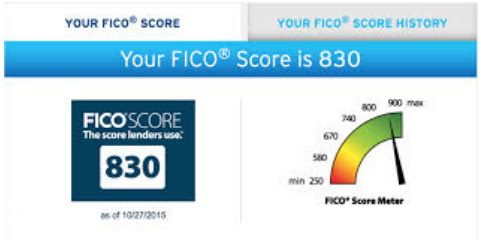
- Target customers?
- Pictures of Cats and Dogs
- Speech recognition
- Recognize Letters: A, B, C..

Traditionally 2 Tasks: Classification & Predictive Scoring

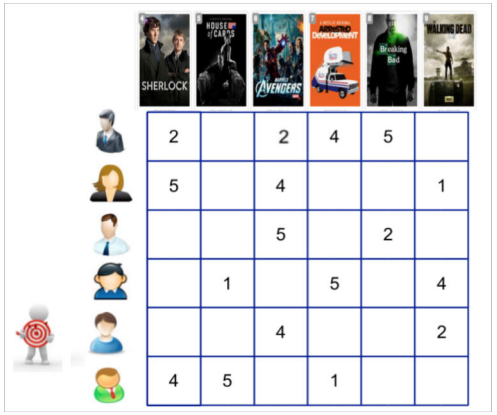
Extracted Data often in Table Format



Classification:
Cats and Dogs, Speech Recognition
Movie Recommendation



Scoring:
Credit [illegible] Movie Rating
Heat [illegible]



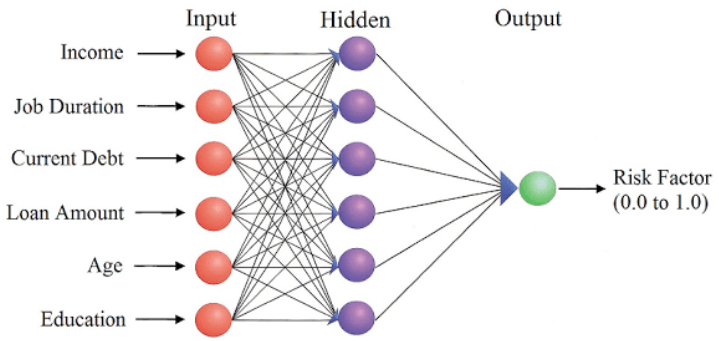
The most famous application has been recommendation: “which other user is most like you”

We have now switched to Neural Networks as Function Approximators

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

"Non-deep" feedforward neural network

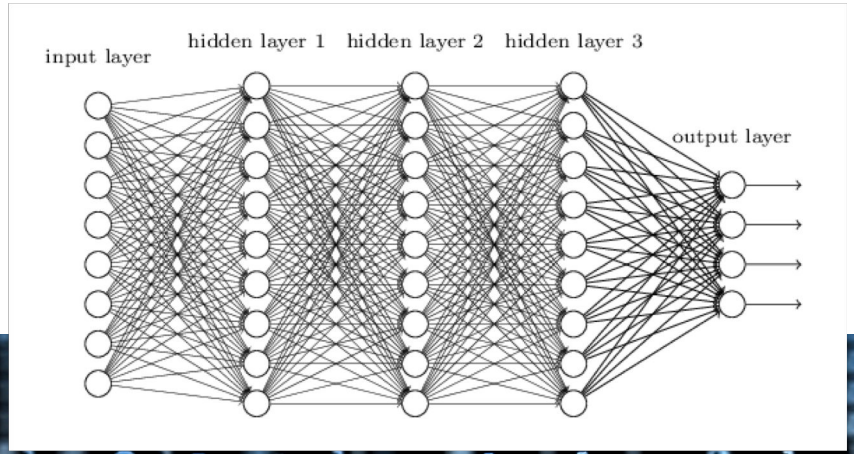
X



Y

Deep neural network

X



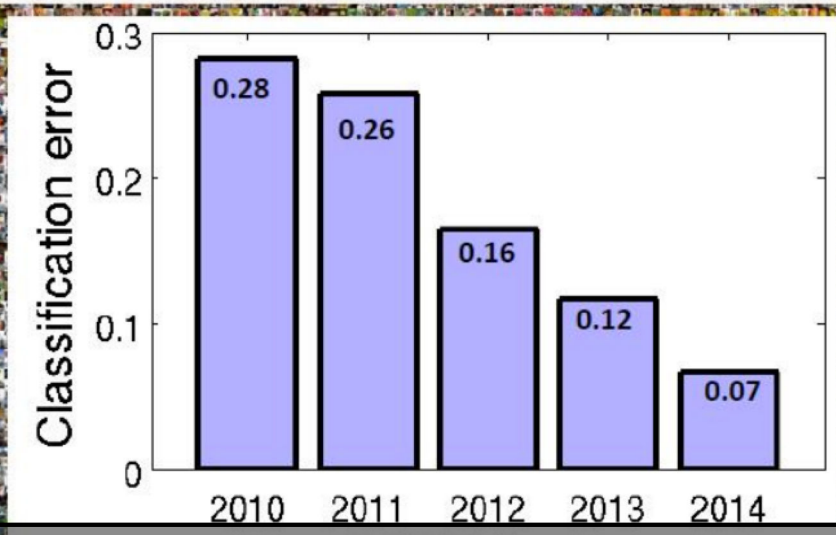
Y

IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:

1,000 object classes

1,431,167 images

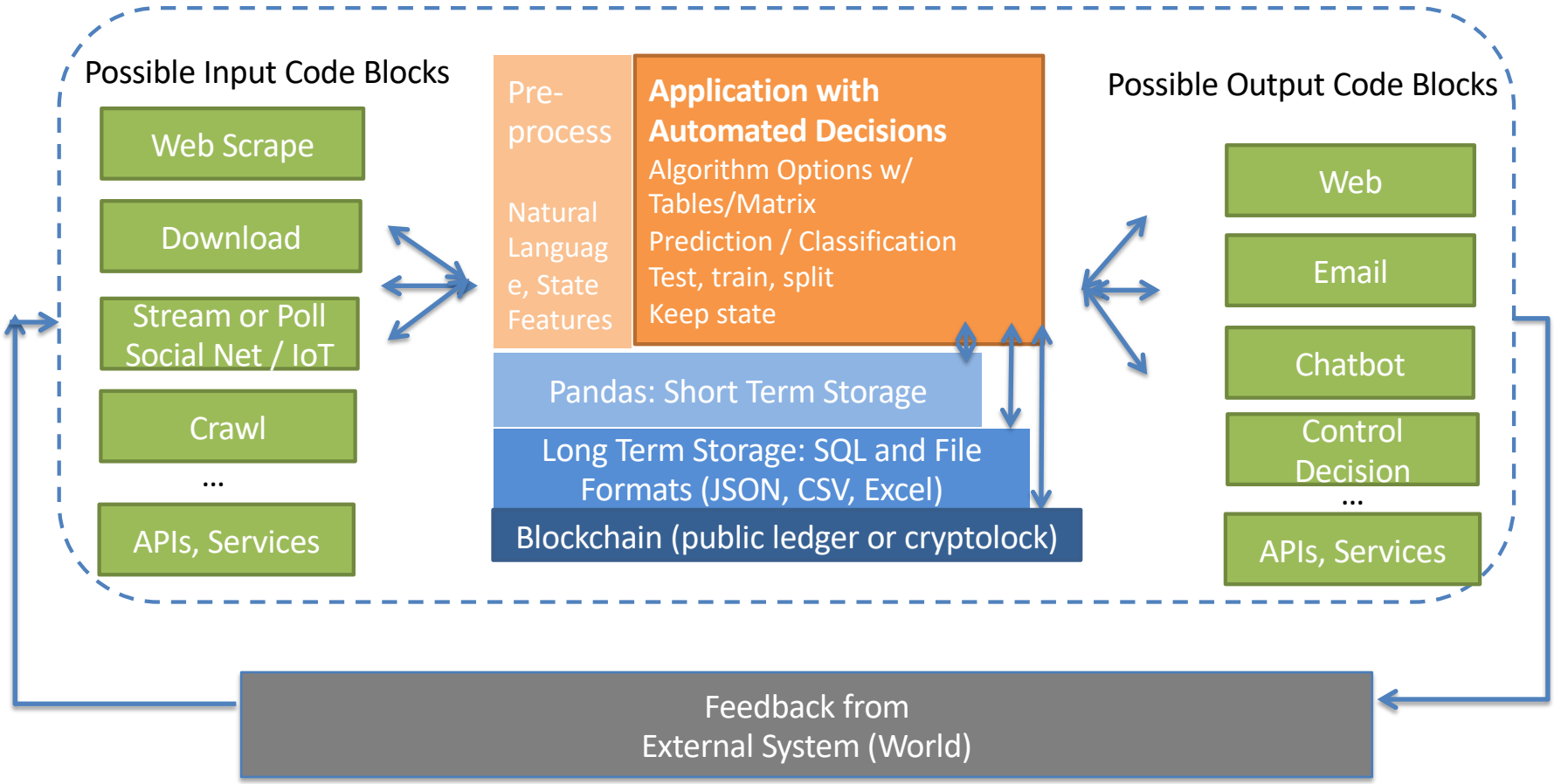


Neural net results are close t human results

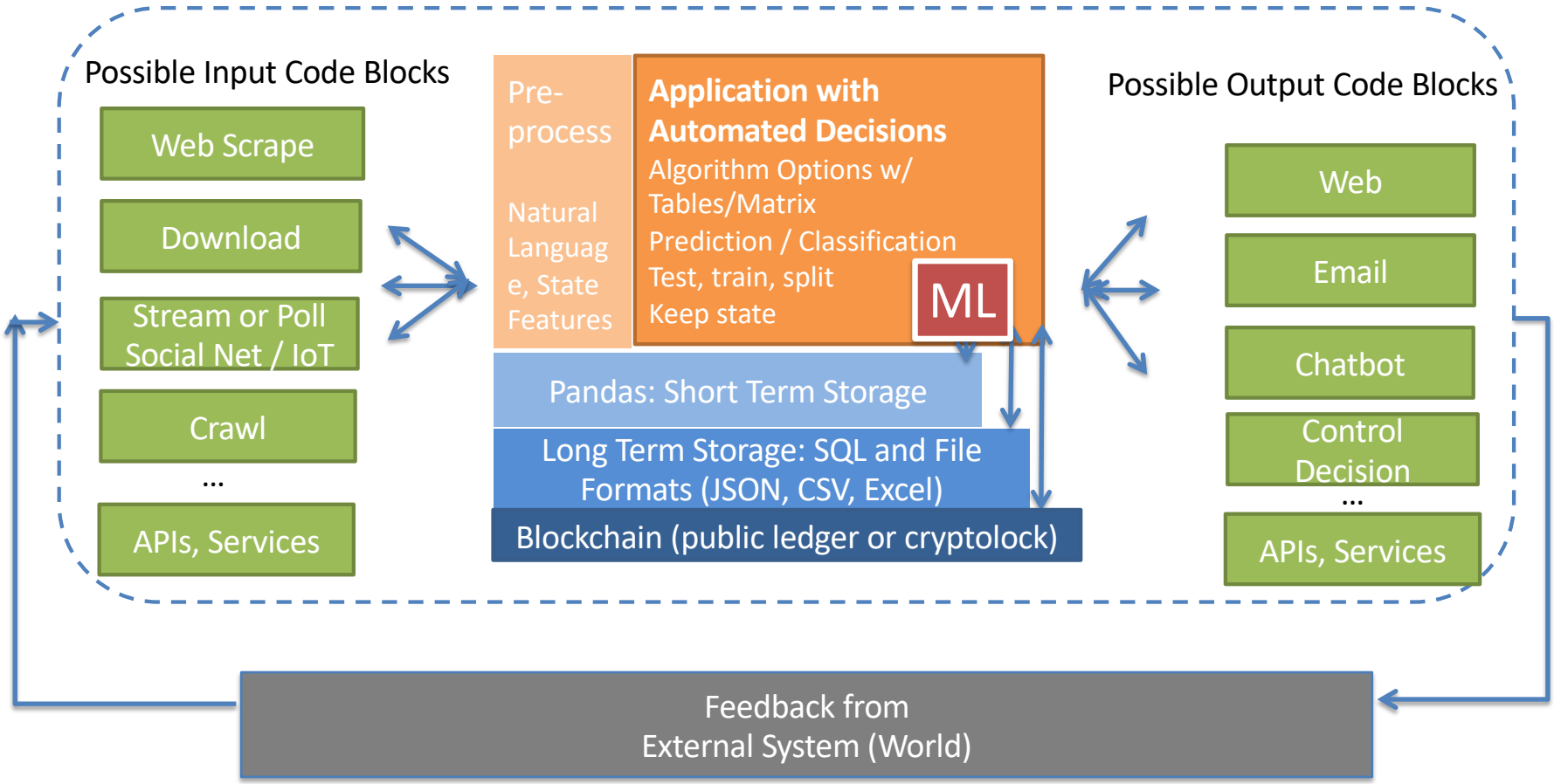
Russakovsky et al. arXiv, 2014

Data

The Data-X System View



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



Reminder: Homework For Week 1

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

- ---

- Homework Part II
- Python-based review notebook (Breakout and Homework, BKHW). To sent by email.

End of Section