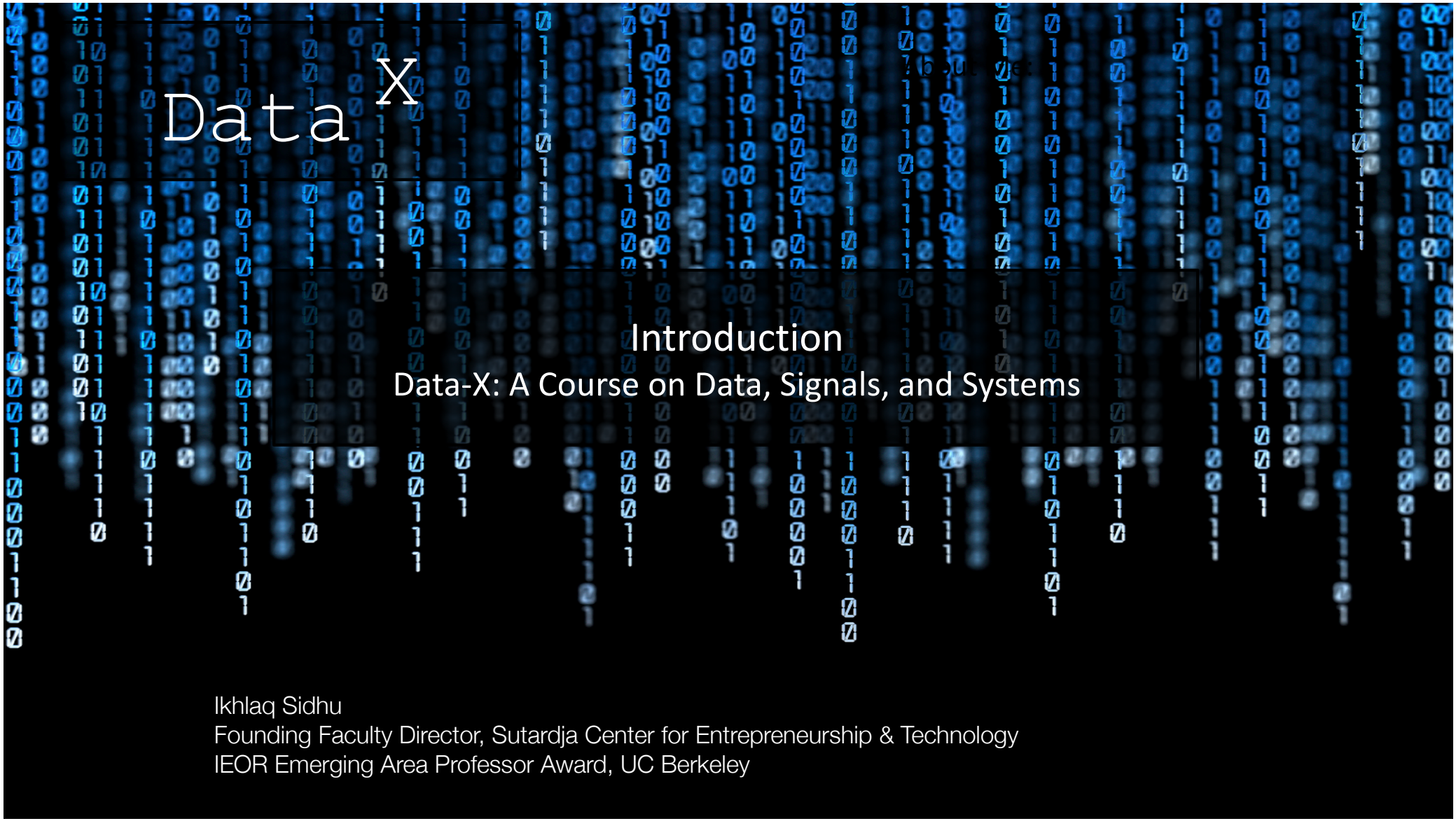


Introduction

Data-X: A Course on Data, Signals, and Systems

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

Welcome to Data-X

- First Class: Introductions (15 min)
  - GSI: TBN, Alexander Fred-Ojala, afo@berkeley.edu (Visiting Scholar), Kevin Bozhe Li, kbl4ew@berkeley.edu,
  - Many others have contributed. Many advisors.

- Pre-requisites:
  1) Working knowledge of Python, 2) Probability and/or Statistics, 3) Know basic matrix multiplication

- What is Data-X (20 min)

- High Level Overview of Data (1.5 hours)

- HW for next Week (15 min)

- Remaining Time – Advisor mixer and/or get help if needed from GSI/teaching team to install your coding environment.

Most Resources Are Available at data-x.blog

By Next Week:

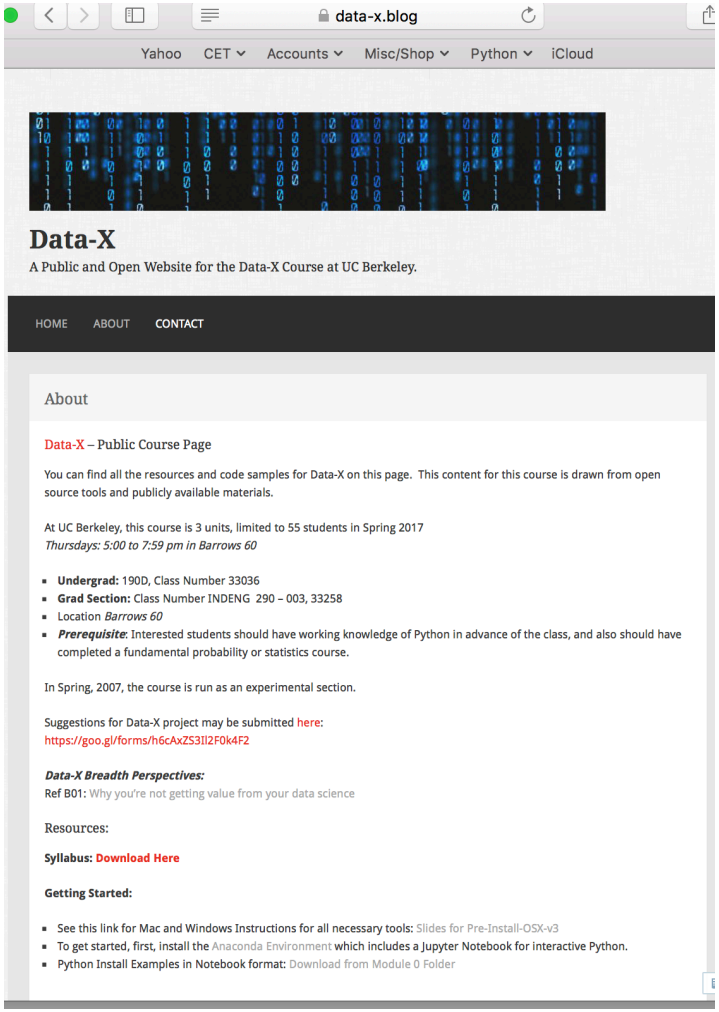
1. Go to Data-X.blog
   - Syllabus
   - Instructions for SW Install
   - Link to GitHub with Cookbook Code Samples and Slides

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

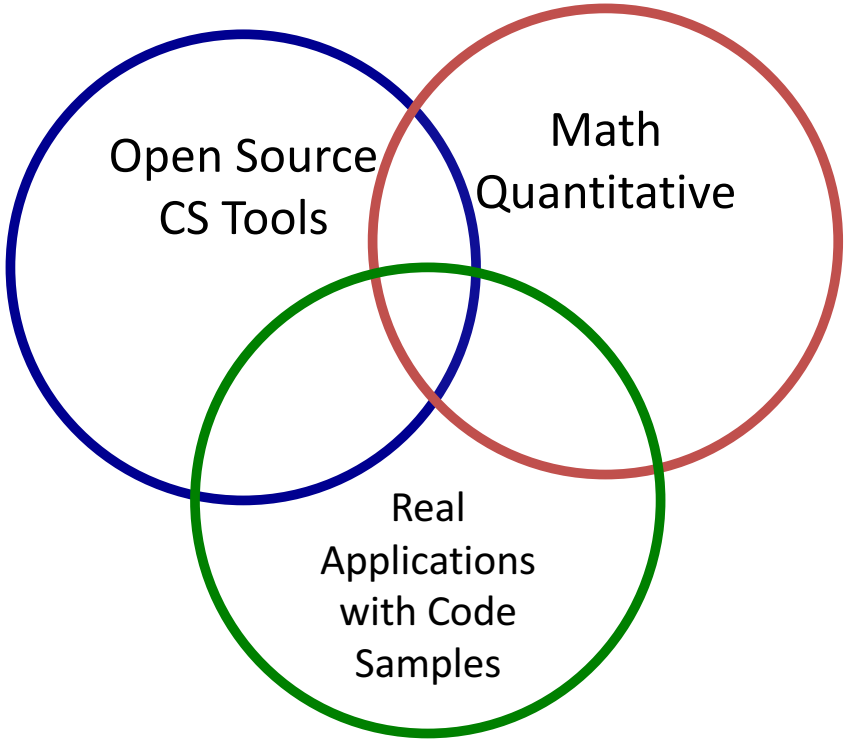
4. Follow Directions to create your own Jupyter notebook and solve problems. Turn in pdf copy electronically. This will be sent out separately

5. Come up with 3 ideas for a group project. Write 1-3 sentences for each. Bring it with you on paper to turn in at next class.



What is Data-X?

- A Course
- Applied Project
- Industry Perspective, Social Applications, Customer Driven



What is in it?

New for fall 2017: More modeling

Common Open Source CS Tools:

- Numpy, SciPy
- Pandas
- TensorFlow, Sklearn
- SQL to Pandas
- NLP / NLTK
- Matplotlib

Often: Working Code First
Fill In Theory After

Quantitative

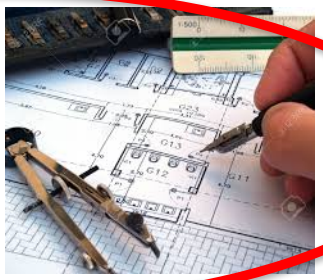
- Prediction: Regression
- ML Classification: Logistic, SVM.. Trees, Forests, Bagging, Boosting,..
- Entropy / Information Topics
- Deep Learning examples, including CCNs
- Correlations
- Markov Processes
- LTI Systems: Fourier, Filters where applicable
- Control Models where applicable

Building Block Code Samples

- Webscraping
- Stock market live download, simple trading
- Convolutional Neural Networks
- Next Word Predictor, Spell Checking
- Recommendation
- Web Crawler
- Chatbot, E-mail
- Social net interfaces including twitter

Data-X: This class will help you combine math and data concepts
The course updates with new tools to stay current. You may learn and use tools not presented in the class project.

What is this class



Make the Tools

Use the Tools (Optimally)

Architect the System

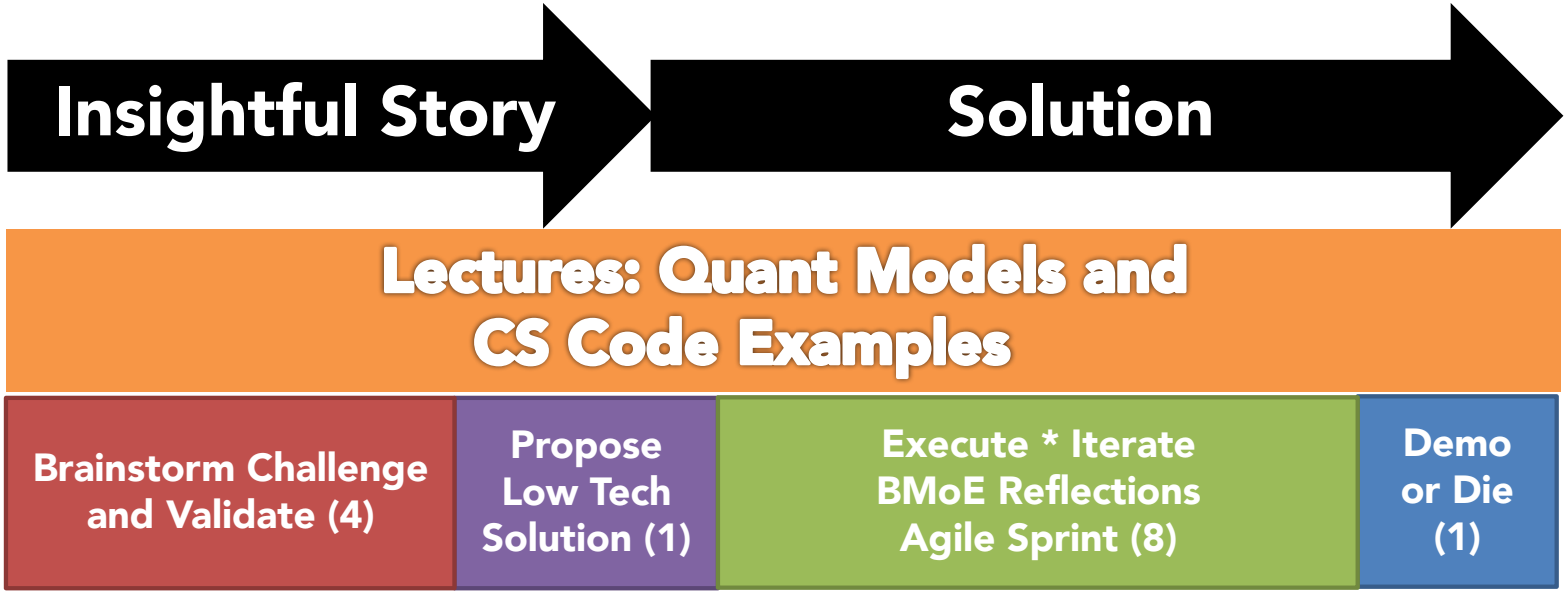
Why and how you build

Most CS

This IEOR Course

Sutardja Center

How the Data-X Course Works:

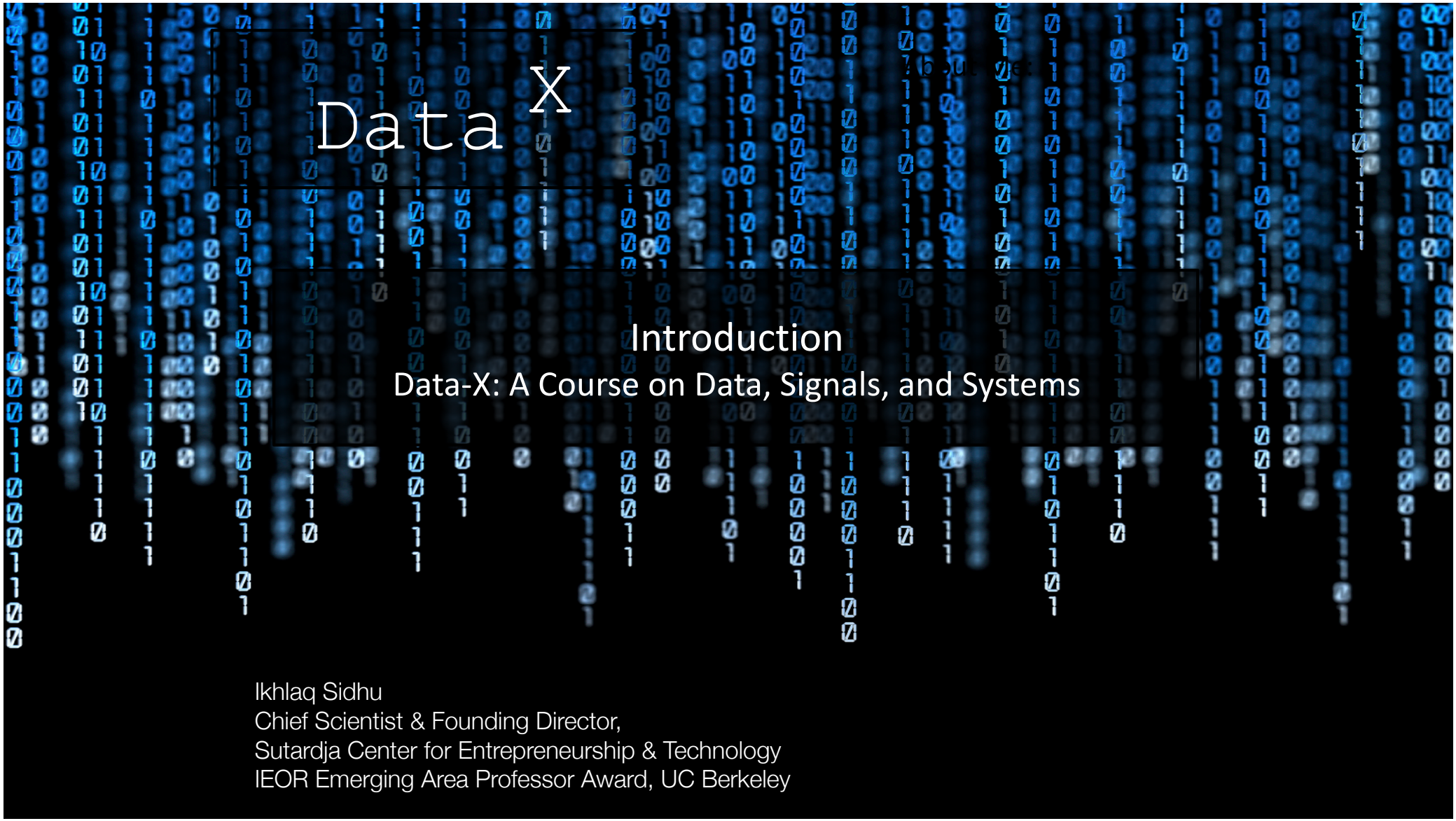


Team: typically 5 students, with available advisor network

Basic Tools to Get Started

- **Available with Anaconda Environment (available for free):**
  - Python, we will use version 3.x, pre-requisite to class
  - NumPy, array processing for numbers, strings, records, and objects
  - Pandas, Powerful data structures and data analysis tools
  - SciPy, Scientific Library for Python
  - Matplotlib, Python 2D plotting library
  - Ipython - Productive Interactive Computing

- **Environment includes:**
  - Jupyter – Interactive web based python
  - Spyder – code development environment with editor



Introduction

Data-X: A Course on Data, Signals, and Systems

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

A High Level Overview of Data

Basic Concept of Working with Data

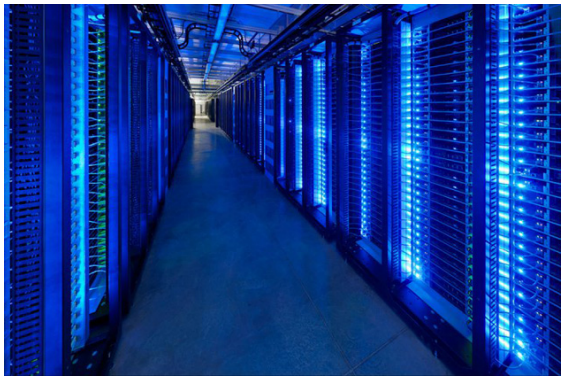
- Data Wrangling
- In Production

Human Interpretation of Data

Human

Large Sets of Data

Insight



How did data become such a big deal?

Scoring Wine

Wine quality =

12.145 +
0.00117 x (winter rainfall) +
0.0614 x (average growing season temperature) –
0.00386 x (harvest rainfall)

Oren Ashenfelter, Princeton. Now used by Christies Auction House

Competitive Advantage in Sports

Money Ball:

How to measure and predict baseball performance

Oakland Athletics baseball team and its general manager Billy Beane

A: Watch and talk with hundreds of players

B: Runs created =

(hits + walks) x Total Bases /

(At Bats + Walks)

Now: Basketball, Football, and soon every other sport

Data X

Customers who viewed this item also viewed these products

Dualit Food XL1500 Processor

$560

Add to cart

Kenwood kMix Manual Espresso Machine

★★★★☆

$250

Select options

Weber One Touch Gold Premium Charcoal Grill-57cm

$225

Add to cart

NoMU Salt Pepper and Spice Grinders

$3

View options

Recommendations based on Algorithms

NETFLIX

amazon.com and you're done.

eHarmony

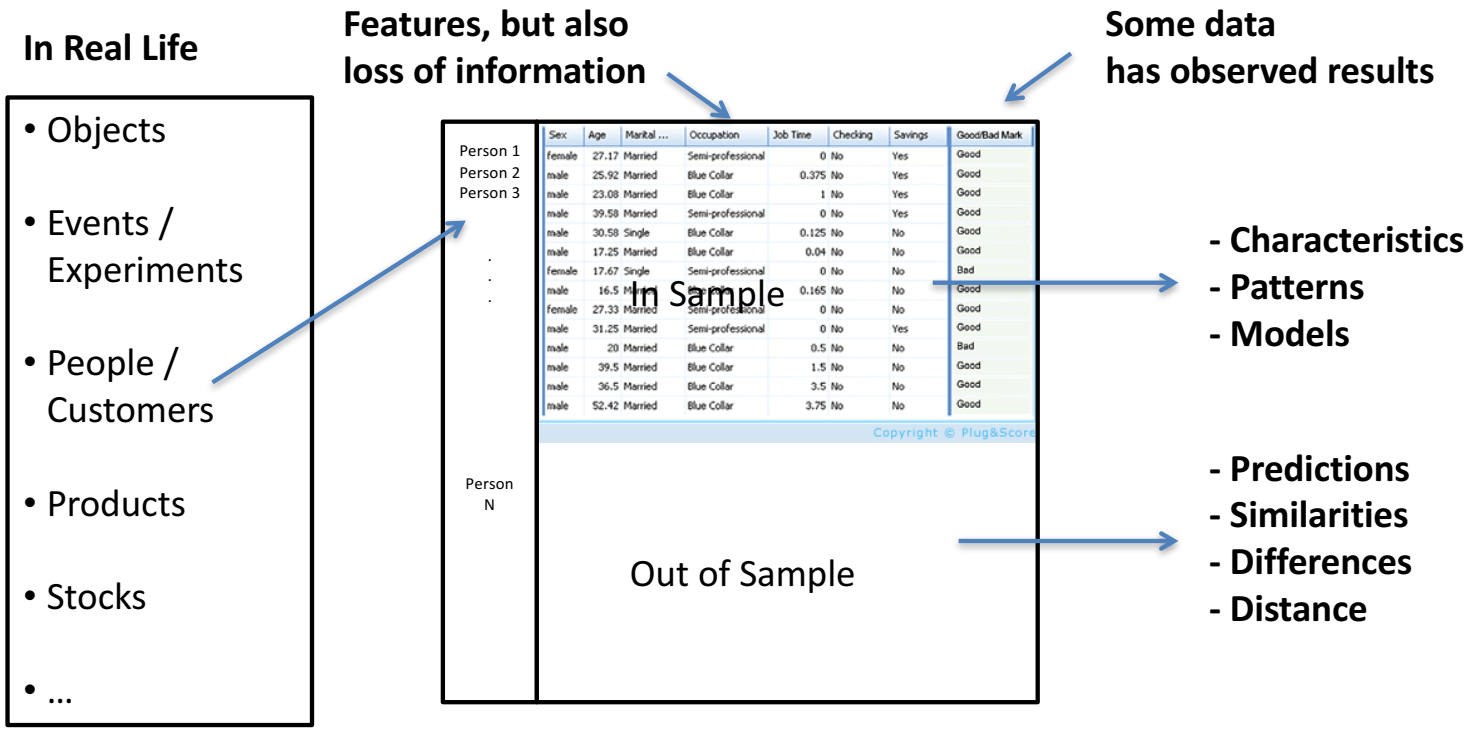
THE HOME DEPOT


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

What has been happening

Context:	Internet Web	Social Nets Recommend	Higher Accuracy Larger training	Control + AI Self learning
	E-Commerce	Ad Driven Fin/Quant	Sharing Economy	?

An ML High Level Framework



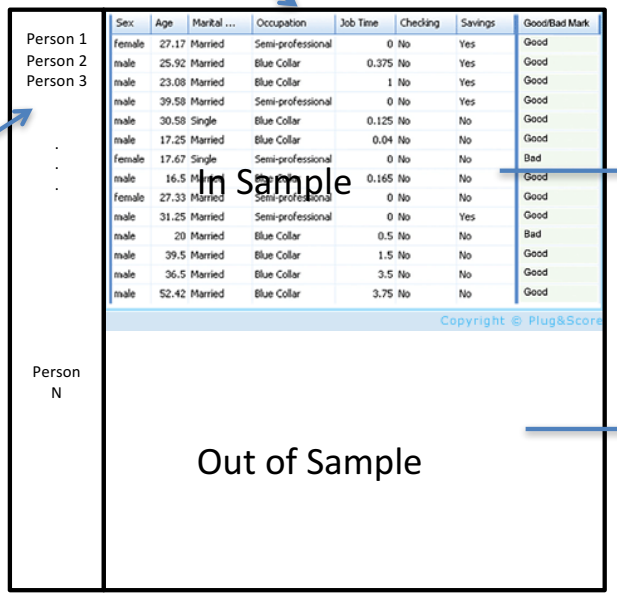
An ML High Level Framework

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

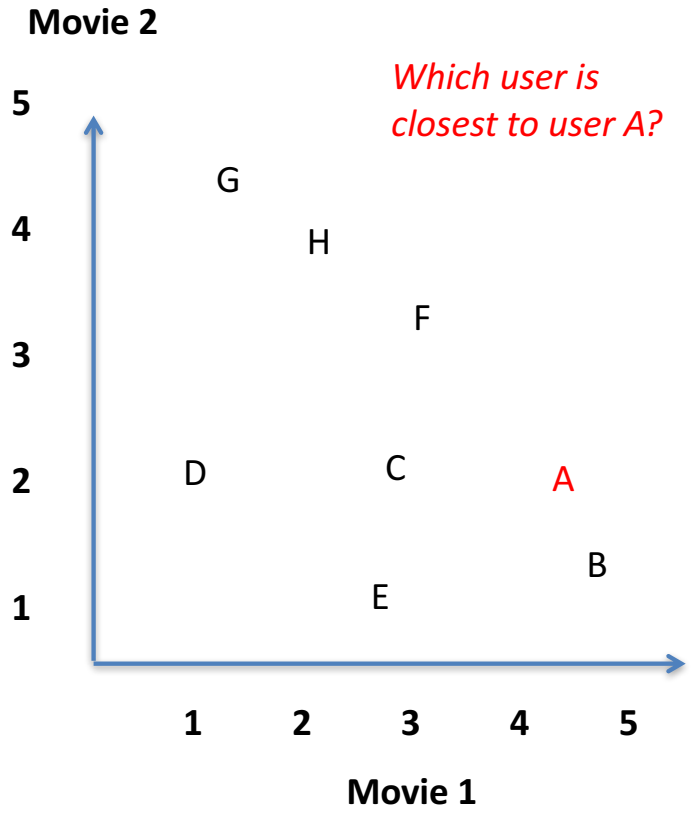
Math: Matrix **X**, with N rows – each person
m columns, each feature (age, salary, ..)

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

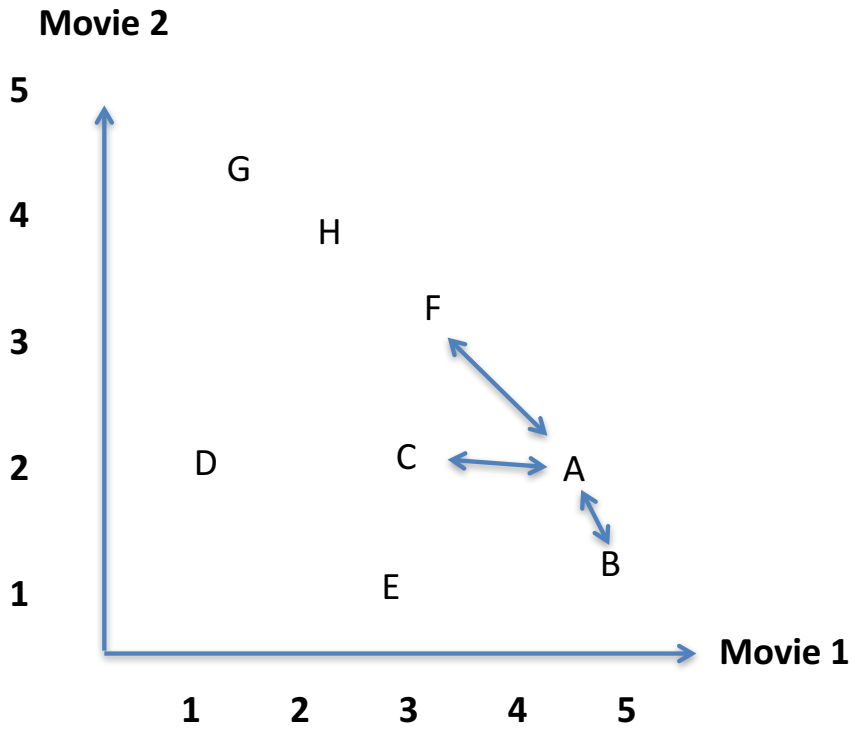
A Fundamental Idea: From Table to N- Dimensional Space

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 by Measuring Distance (Unsupervised)



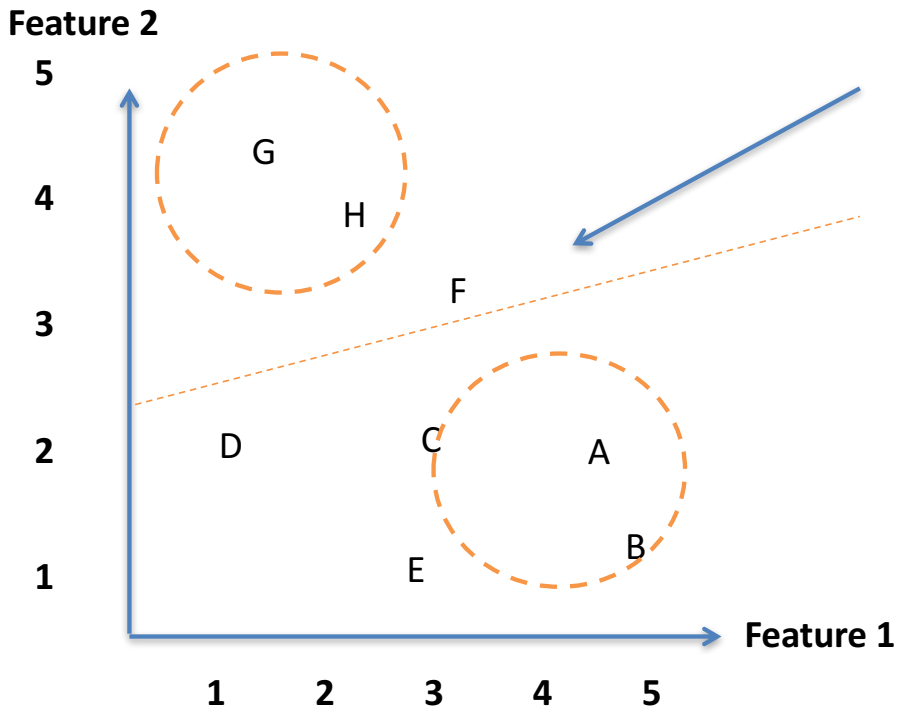
Distance functions

Euclidean: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan: $\sum_{i=1}^{k}|x_i - y_i|^2$

Minkowski: $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Clustering to Classification



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

For Netflix: *70K -> 200K titles (dimensions), 10M plus users (points)*
Factors: Accuracy vs Performance

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
..	..

Machine Learning: Learning from Data

Input Data = Matrix X

Customer 1: [Name, income, x, y, .. Features ..z]
Customer 2: [Name, income, x, y, .. Features ..z]
Customer N: [Name, income, x, y, .. Features ..z]

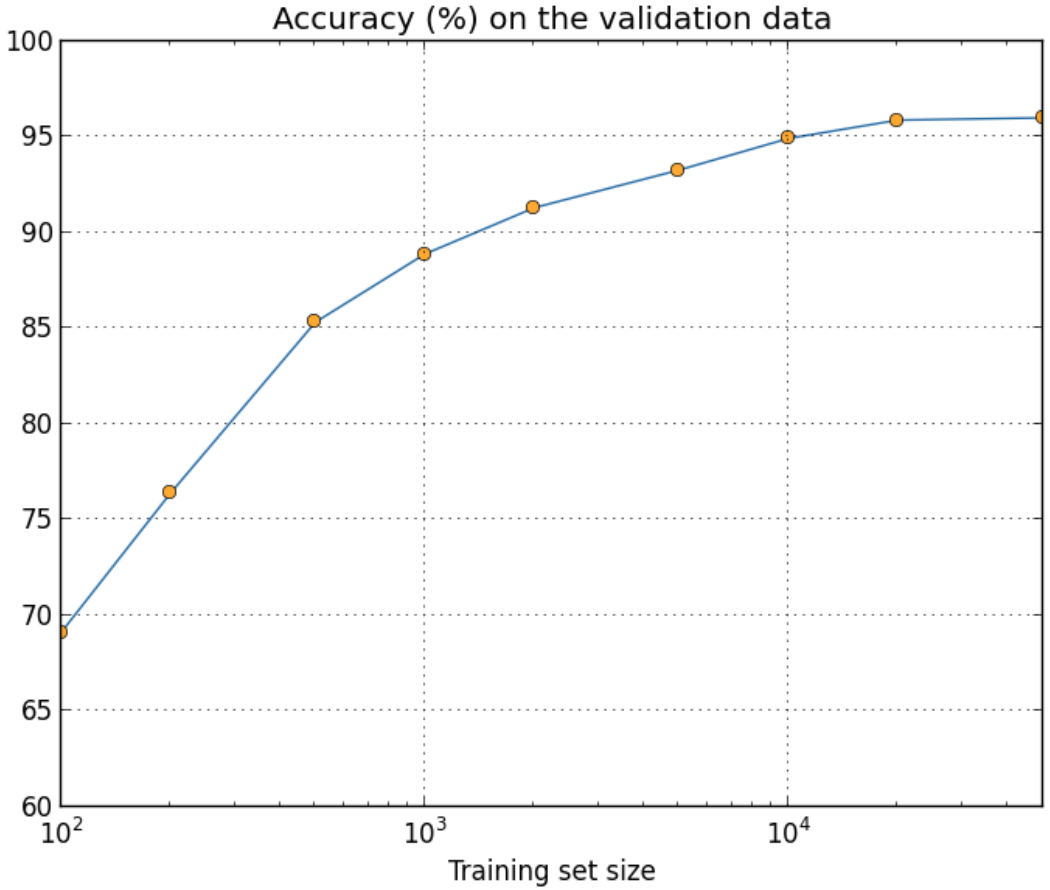
Output Data = Column Vector Y

Customer 1: [20]
Customer 2: [60]
Customer N: [05]

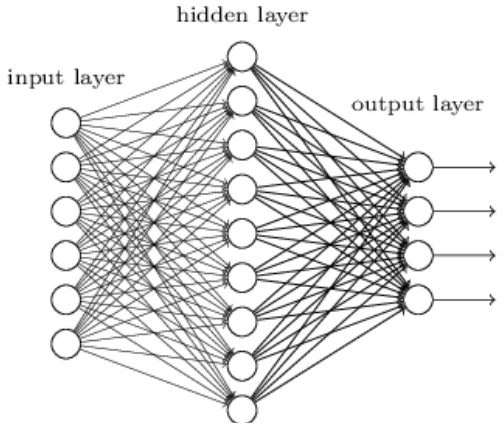
Purchases/year, repaid loan, …

Target: What is F(X) = Y a formula that we don't know

Sample data (training): (x1,y1) (x2,y2) … (xm,ym) we have this

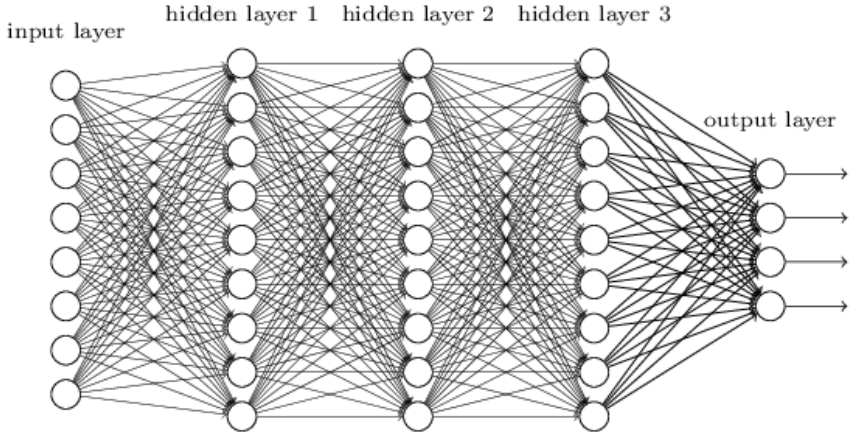

Accuracy (%) on the validation data
100
95
90
85
80
75
70
65
60
10^2
10^3
10^4
Training set size

"Non-deep" feedforward neural network



X Y

Deep neural network



X Y

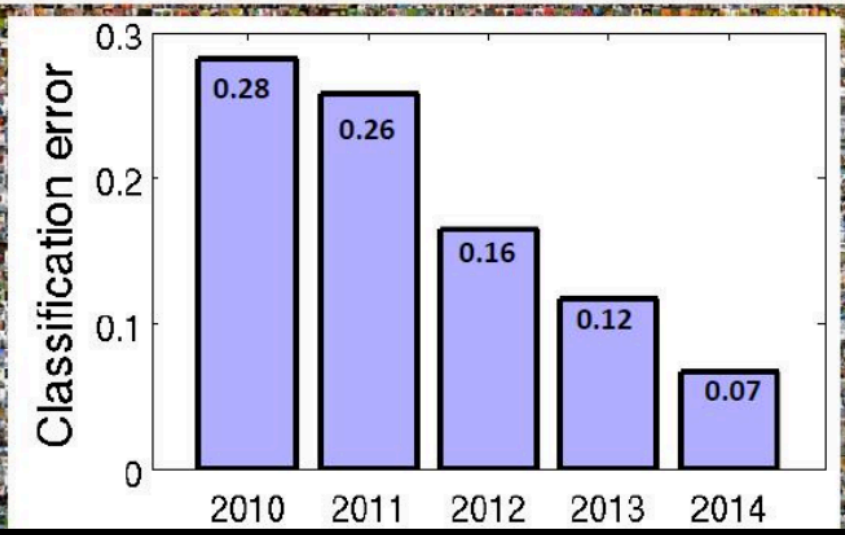
Data X

IMAGENET **Large Scale Visual Recognition Challenge**

The Image Classification Challenge:

1,000 object classes

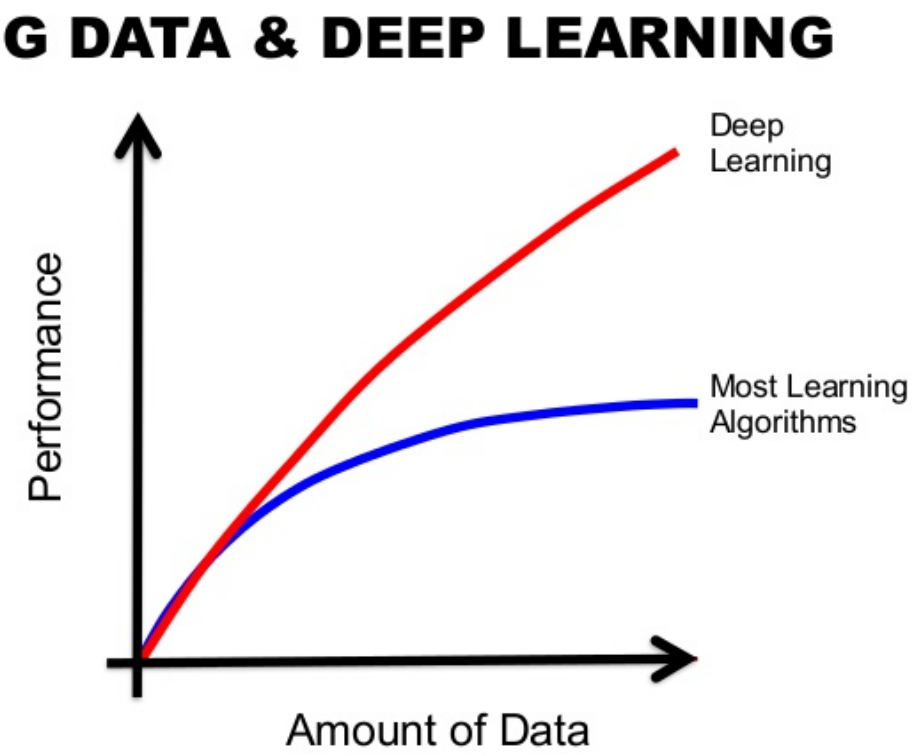
1,431,167 images



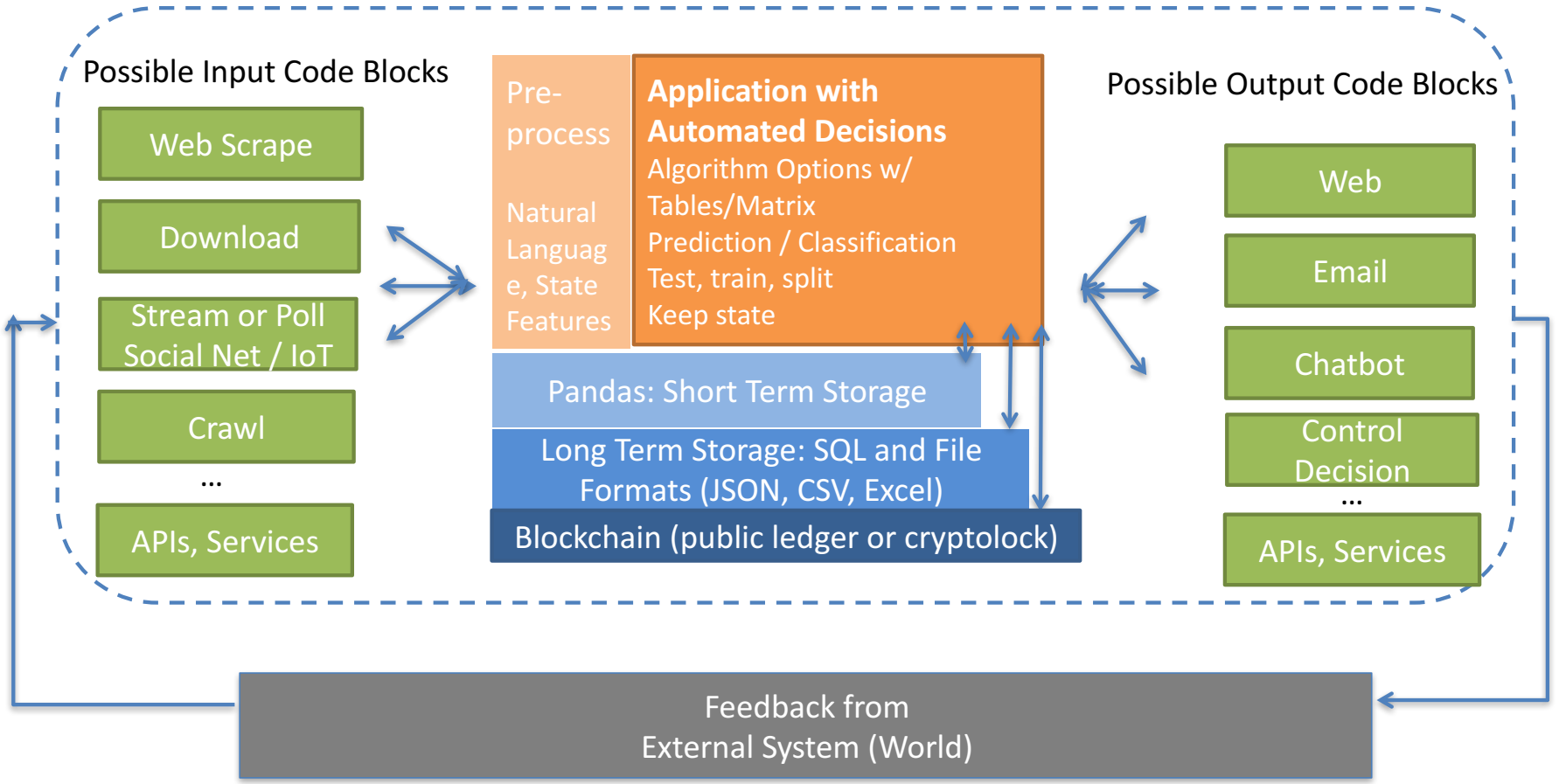
Neural net results are close to human results

Data

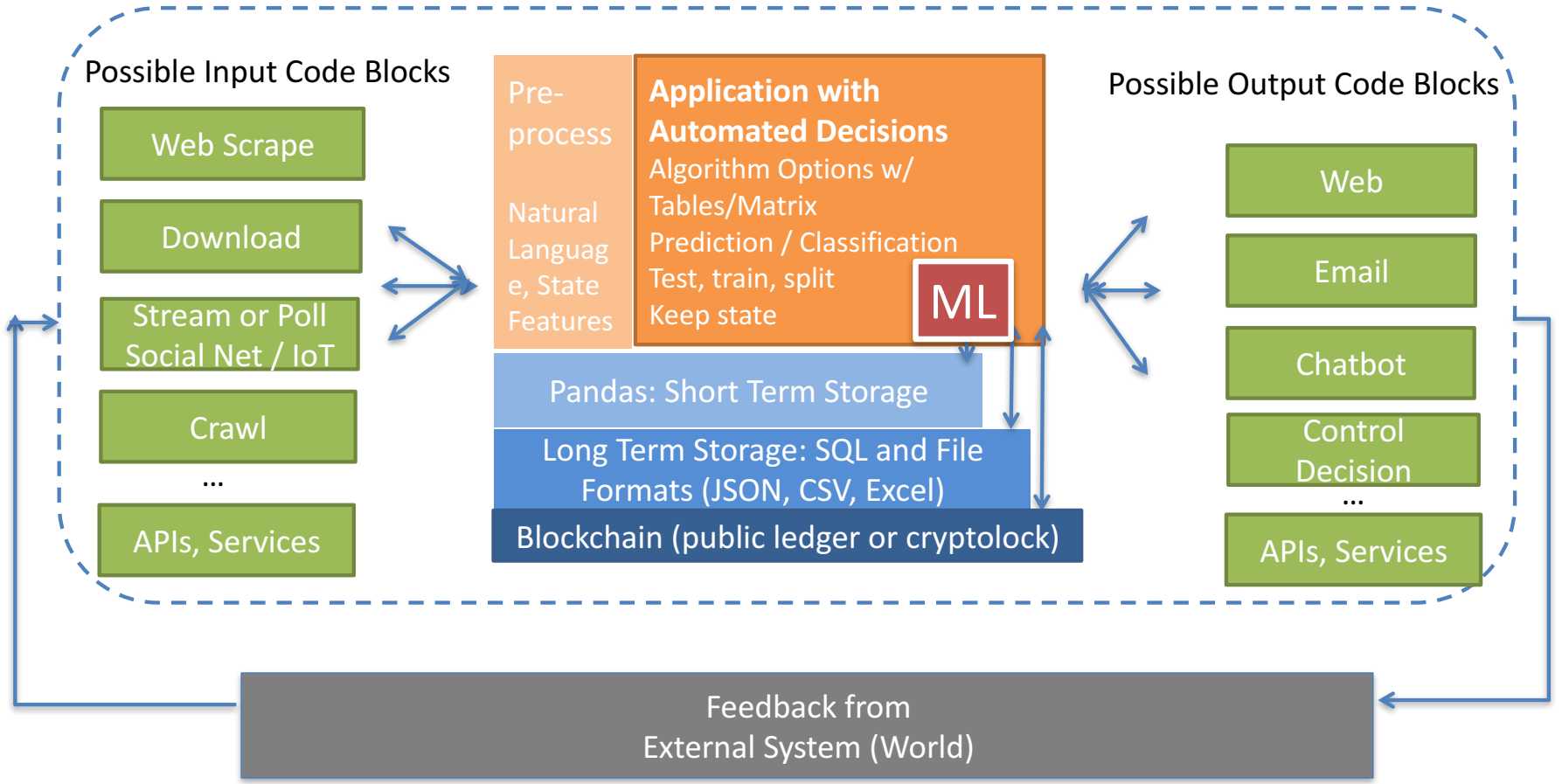
This means
Accuracy



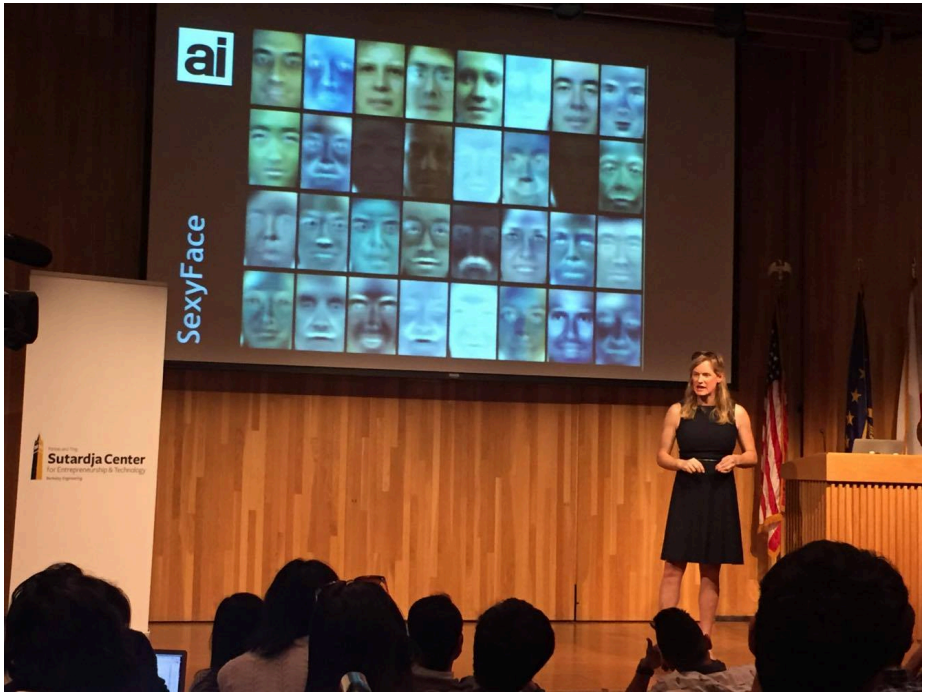
The Data-X System View



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



Vivienne Ming: Face Recognition for Finding Refugees in Camps



Top 8 Business Models Using Data

1. Knowing your customer, better targeting and relationship
   E.g. Target, Disney, Netflix
2. Improving physical product or servicer with complimentary information
   E.g. UPS, FedEx
3. Data-driven reliability or security
   E.g. GE, BMW, Siemens
4. Information Brokers, Arbitrage, and Trading Opportunities
   E.g. Investment funds.
5. Improving the customer journey/experience
   E.g. Harrah's
6. Functional Applications: HR/Hiring, Operations etc.
   E.g. Walmart, Baseball, Sports
7. Efficiency or better performance per dollar cost
   E.g. General IT, SAP, etc
8. Risk Management, regulation, and compliance
   E.g. Compliance 360

Your Project Can also use a Data System for Social Impact

- Financial inclusion
- Health
- Aid to underprivileged
- Joining data for a research purpose
- Justice
- Environment
- …

In this class, we will learn ways to:

* Collect the data about objects

* Combine data sources when needed

* Use tables and databases to store

* Practice making good “features”

* Learn to Analyze; Compute, Classify, Predict

* Visualize some results

Use cookbook applications to get you started on your own applied project in a group.

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.

Same instructions are in HW 1 Assignment

Class Homework Part II

- Download the course materials at https://data-x.blog/
- Review the Getting Started material including the Installation instructions **pre-reqs-install-osx v3.pdf**
- For now:
  - Download and install **Anaconda**
  - Install or be able to launch a Jupyter Notebook

- We will send out Python based Coding Assignment separately, due by the next class session

End of Section