

Data X

## Introduction Data, Signals, and Systems

Ikhlaq Sidhu

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

# Welcome to Applied Data Science with Venture Applications

(Supported by the Data-X Project and Sutardja Center)

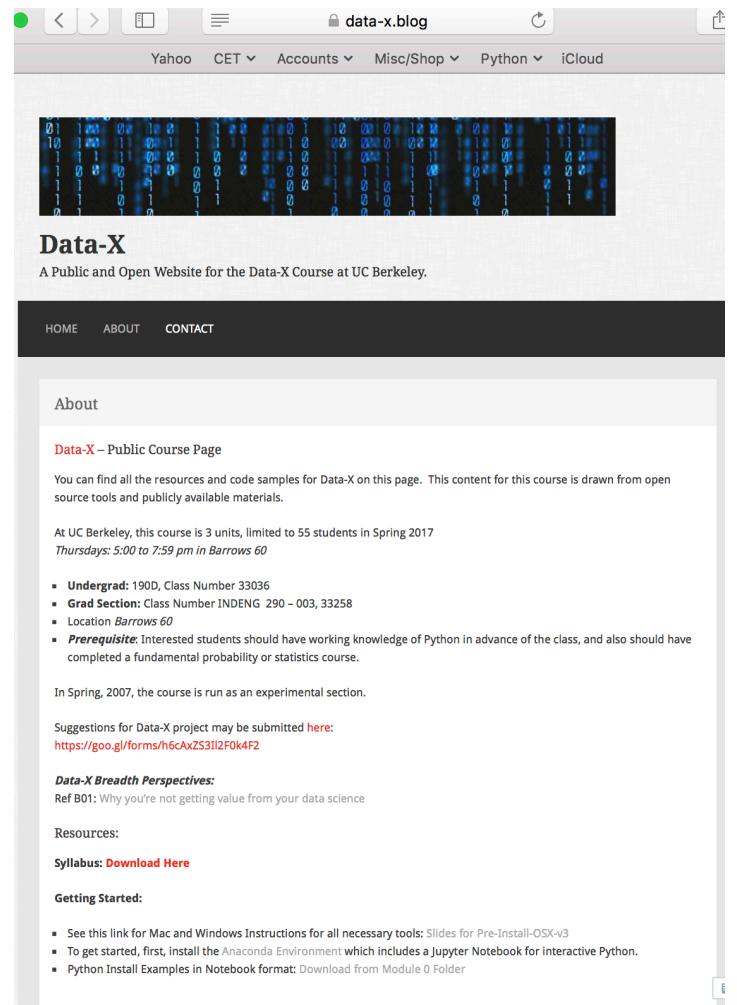
- Teaching Team Introductions (15 min)
  - GSI: TBN, Alexander Fred-Ojala, [afo@berkeley.edu](mailto:afo@berkeley.edu) (Visiting Scholar), Kevin Bozhe Li, kbl4ew@berkeley.edu,
  - Many others have contributed. Many advisors.
- Pre-requisites:  
1) Working knowledge of Python, 2) Probability and/or Statistics, 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.

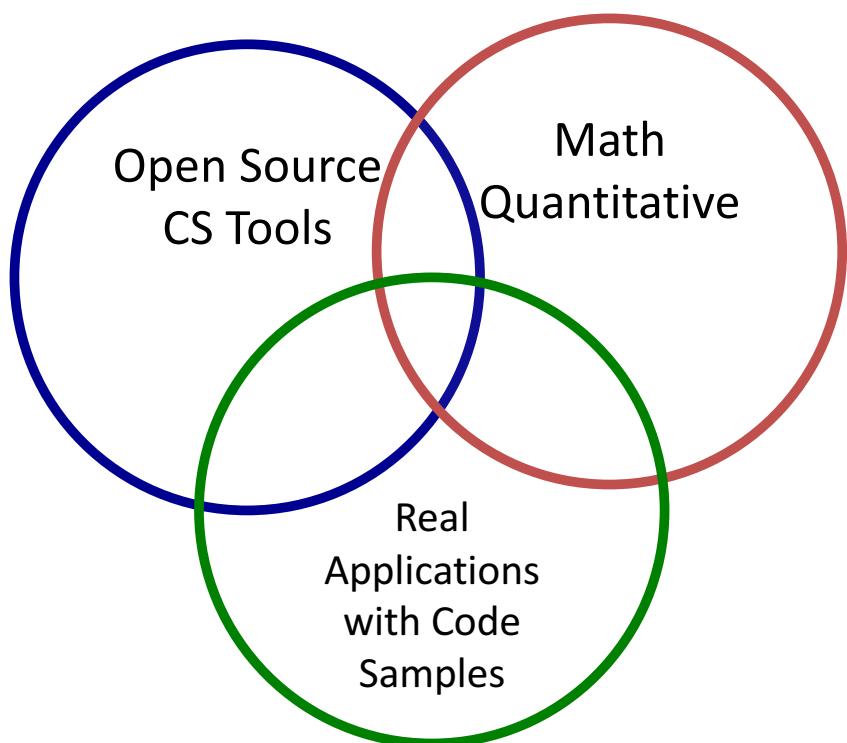


The screenshot shows a web browser window with the URL "data-x.blog" in the address bar. The browser's menu bar includes "Yahoo", "CET", "Accounts", "Misc/Shop", "Python", and "iCloud". The main content area displays the Data-X website. The header features a blue binary background image and the text "Data-X" followed by a subtitle: "A Public and Open Website for the Data-X Course at UC Berkeley". A navigation bar below the header contains links for "HOME", "ABOUT", and "CONTACT". The "ABOUT" section is currently active. It includes a sub-section titled "Data-X – Public Course Page" which provides information about the course being limited to 55 students in Spring 2017. It lists prerequisites such as Undergrad: 190D, Class Number 33036, Grad Section: Class Number INDENG 290 - 003, 33258, and Location Barrows 60. It also mentions a Prerequisite: Interested students should have working knowledge of Python in advance of the class, and also should have completed a fundamental probability or statistics course. The section notes that the course is run as an experimental section in Spring, 2007. It provides a link for suggestions: <https://goo.gl/forms/h6CAxZS3lI2F0k4F2>. Below this, there is a "Data-X Breadth Perspectives" section with a link to Ref B01: Why you're not getting value from your data science. The "Resources" section includes a "Syllabus: Download Here" link and a "Getting Started" section with links to Mac and Windows instructions, Anaconda environment installation, and Python install examples.



# What is Data-X: A Research and Curriculum Project

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



# What is in this class?

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



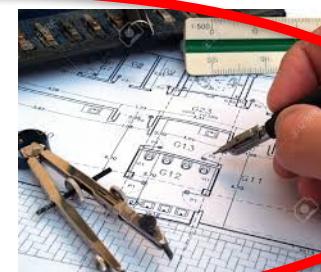
Make the Tools

Most CS



Use the Tools  
(Optimally)

This IEOR Course



Architect the System



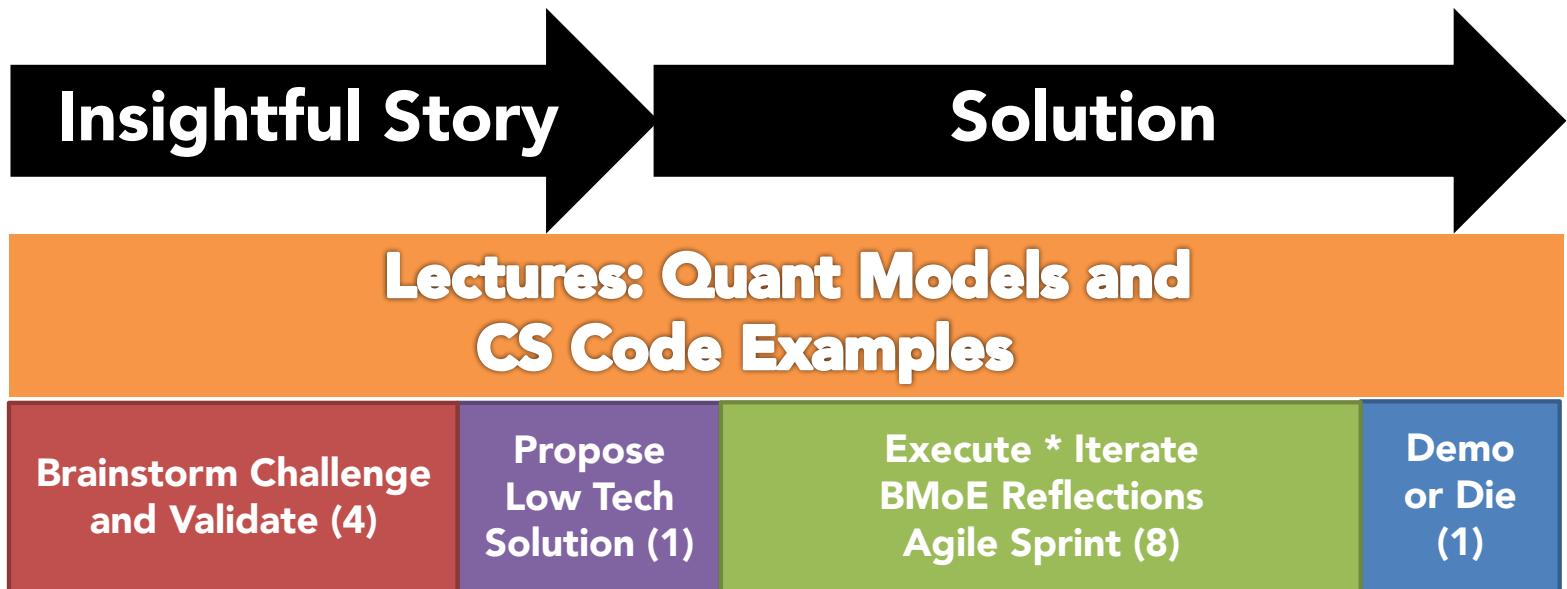
Why and how  
you build

Sutardja Center



Data X

## 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



Data X

## Introduction 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



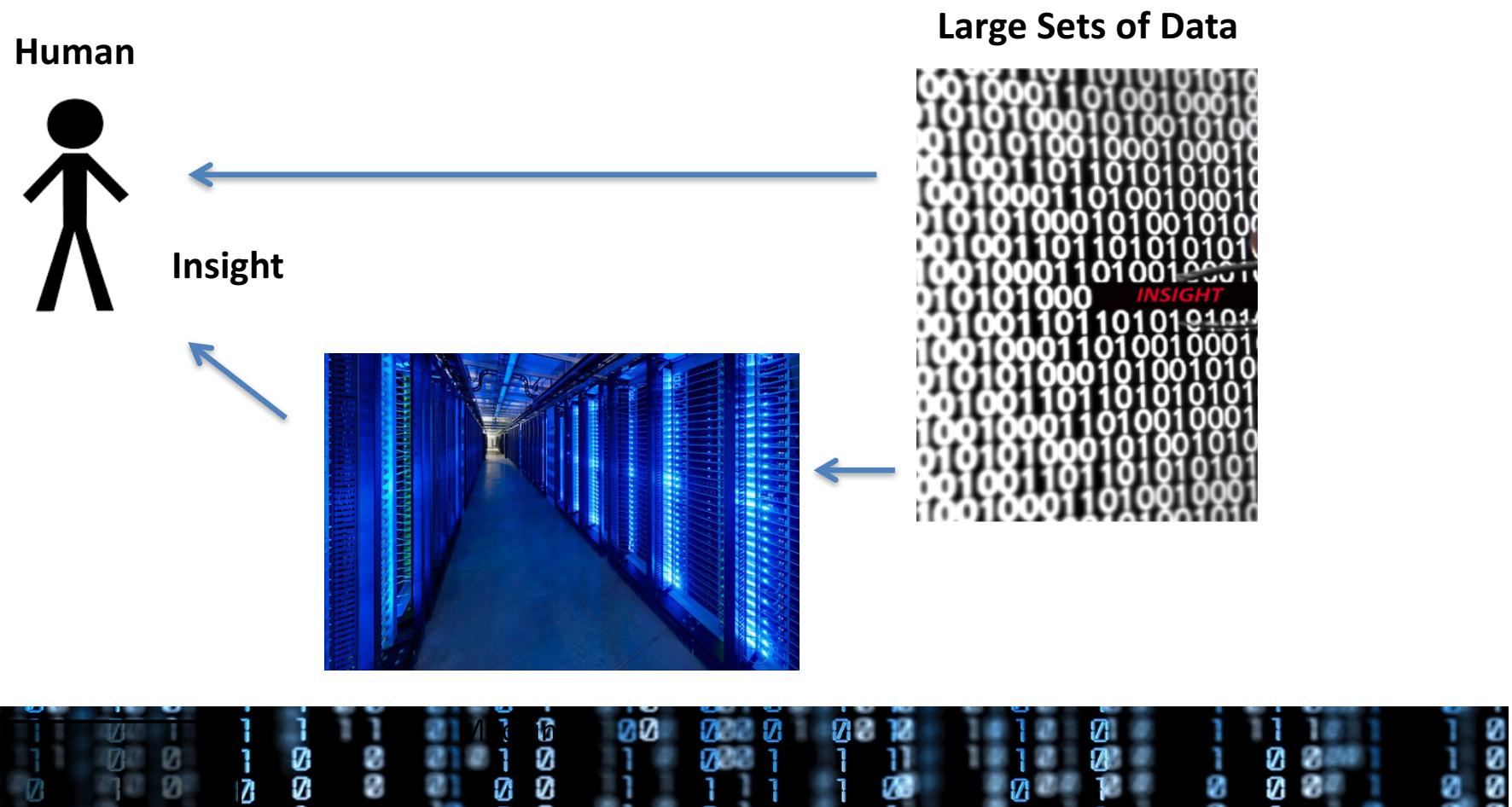
# Basic Concept of Working with Data



- Data Wrangling
- In Production



# Human Interpretation of Data



How did data become such a big deal?



A close-up, low-angle shot of several wine bottles lined up horizontally. The bottles have dark labels and corked tops. The lighting is warm and focused on the necks and shoulders of the bottles, creating a soft glow and casting shadows. A dark rectangular overlay covers the top portion of the image.

# Scoring Wine

Wine quality =

$$12.145 + 0.00117 \times (\text{winter rainfall}) + 0.0614 \times (\text{average growing season temperature}) - 0.00386 \times (\text{harvest rainfall})$$

*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



## 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

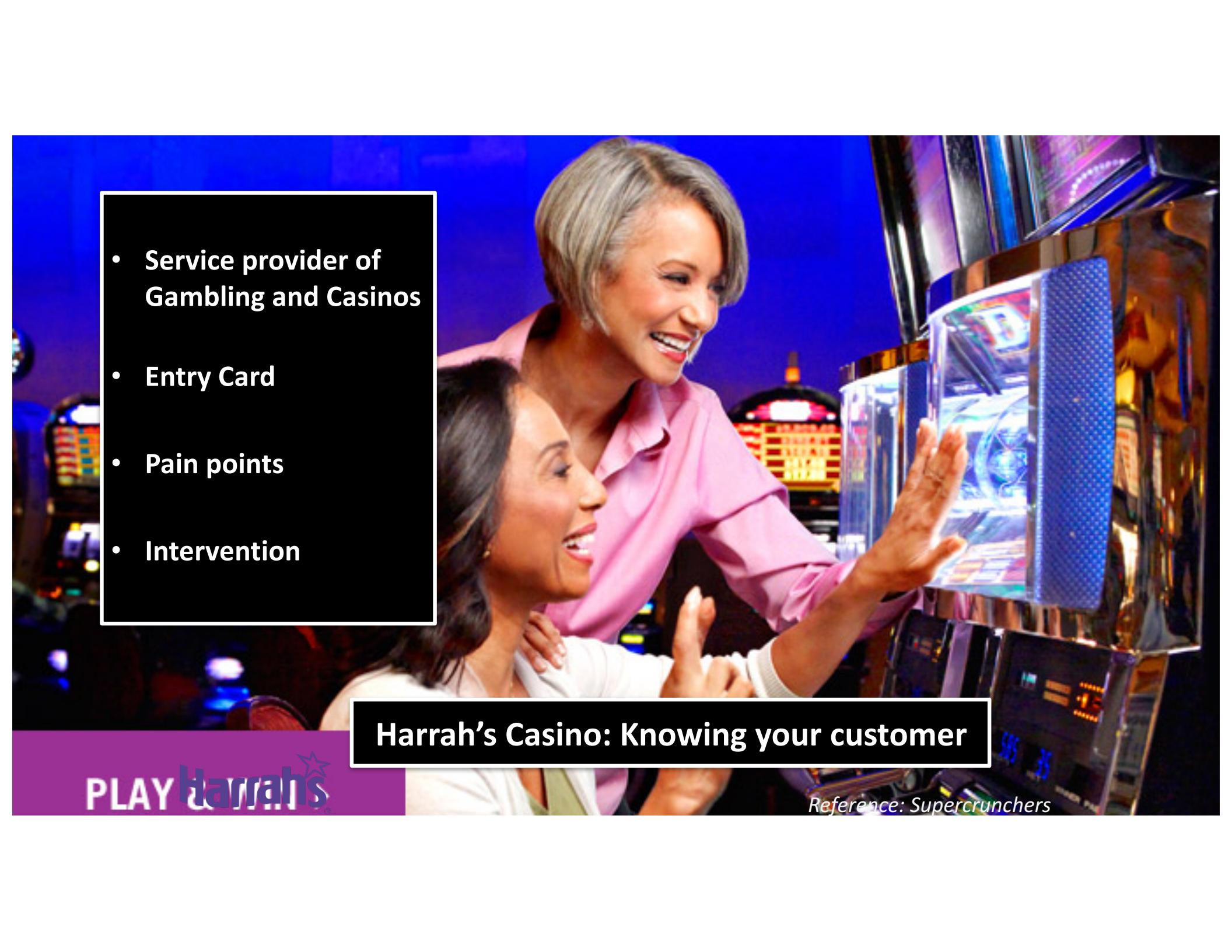
\$3

View options

## Recommendations based on Algorithms



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

A photograph of two women in a casino. One woman, wearing a pink shirt, is smiling and pointing at a slot machine screen. The other woman, wearing a white top, is also smiling and looking at the screen. They are surrounded by bright lights and other slot machines.

Harrah's Casino: Knowing your customer

PLAY Harrah's

Reference: Supercrunchers

# An ML High Level Framework

In Real Life

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

Features, but also loss of information

The diagram illustrates the process of extracting features from real-life entities and organizing them into a dataset. It shows a transition from 'In Real Life' objects to a 'Features, but also loss of information' stage, represented by a table. This table then branches into 'In Sample' and 'Out of Sample' datasets.

**In Sample**

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	30.59	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	Blue Collar	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

Copyright © Plug&Score

**Out of Sample**

Some data has observed results

- Characteristics
- Patterns
- Models

- Predictions
- Similarities
- Differences
- Distance

Data X

# An ML High Level Framework

**In Real Life**

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

**Features, but also loss of information**

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	Blue Collar	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

Copyright © Plug&Score

**Some data has observed results**

- Characteristics
- Patterns
- Models

- Predictions
- Similarities
- Differences
- Distance

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

**CS:** Table

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

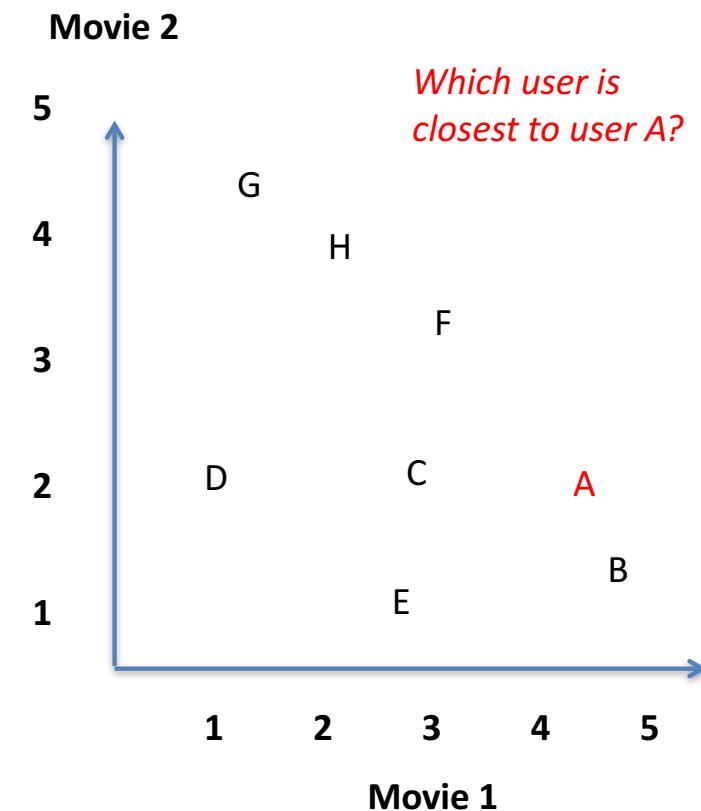


Data  $X$

## A Fundamental Idea: From Table to N- Dimensional Space

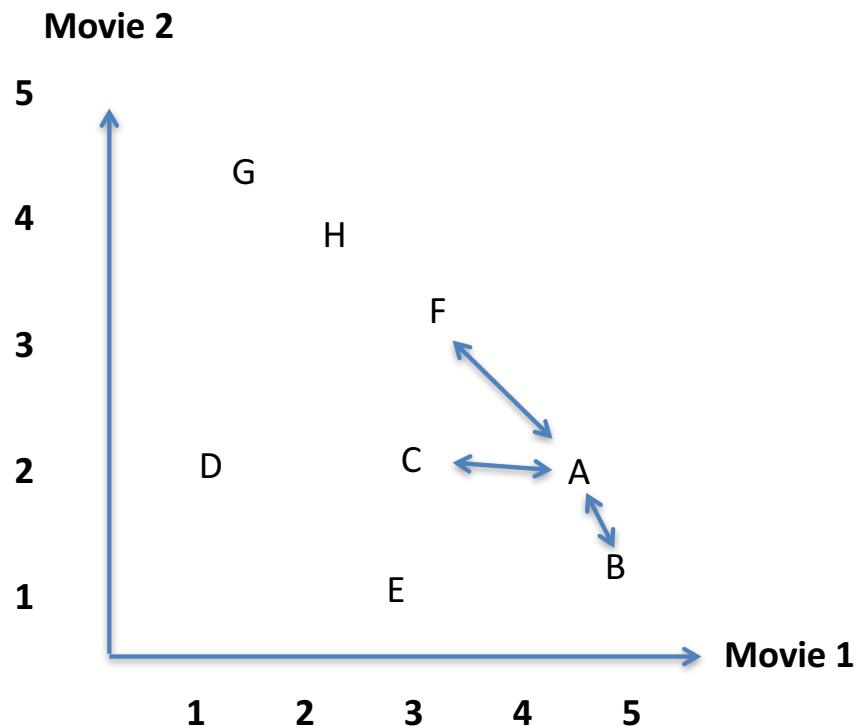
$X =$

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 by Measuring Distance (Unsupervised)



## Distance functions

Euclidean:

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

Manhattan:

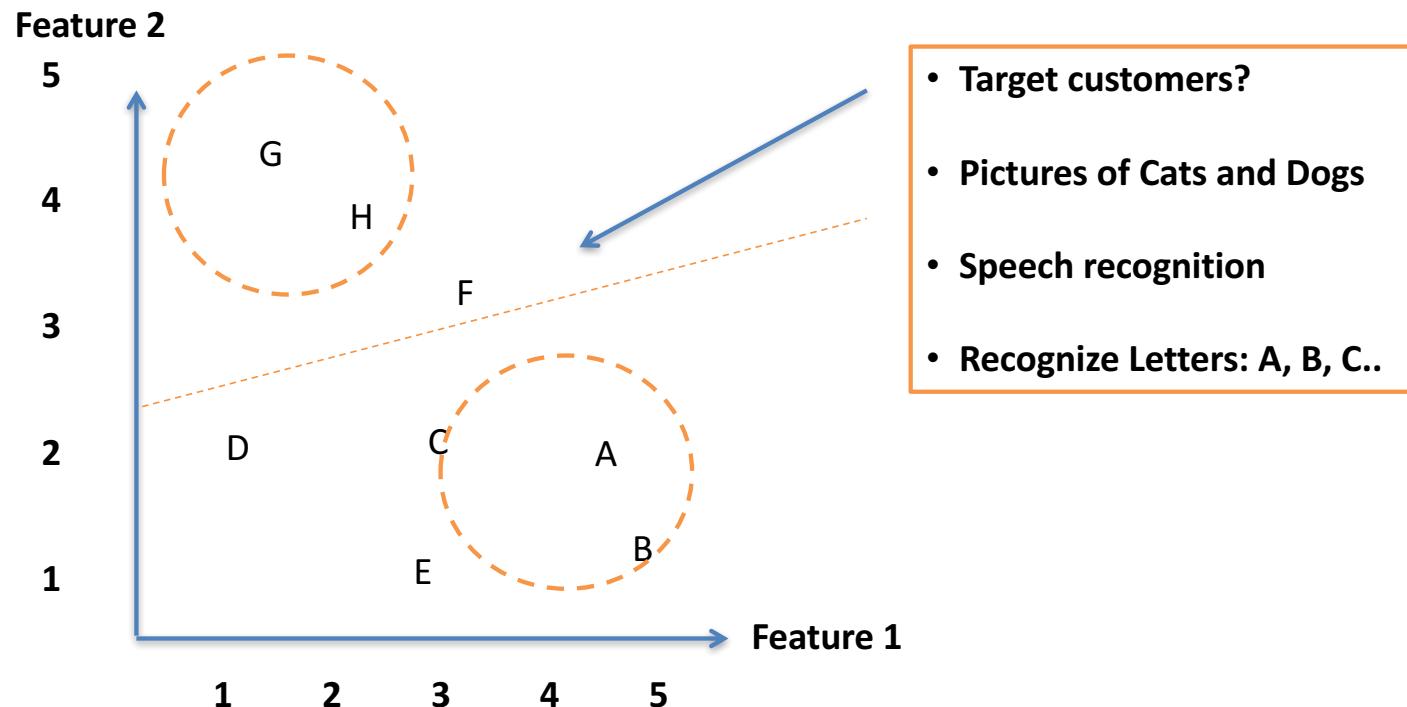
$$\sum_{i=1}^k |x_i - y_i|$$

Minkowski:

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



# Clustering to Classification



For Netflix: 70K  $\rightarrow$  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):  $(x_1, y_1)$   $(x_2, y_2)$  ...  $(x_m, y_m)$**

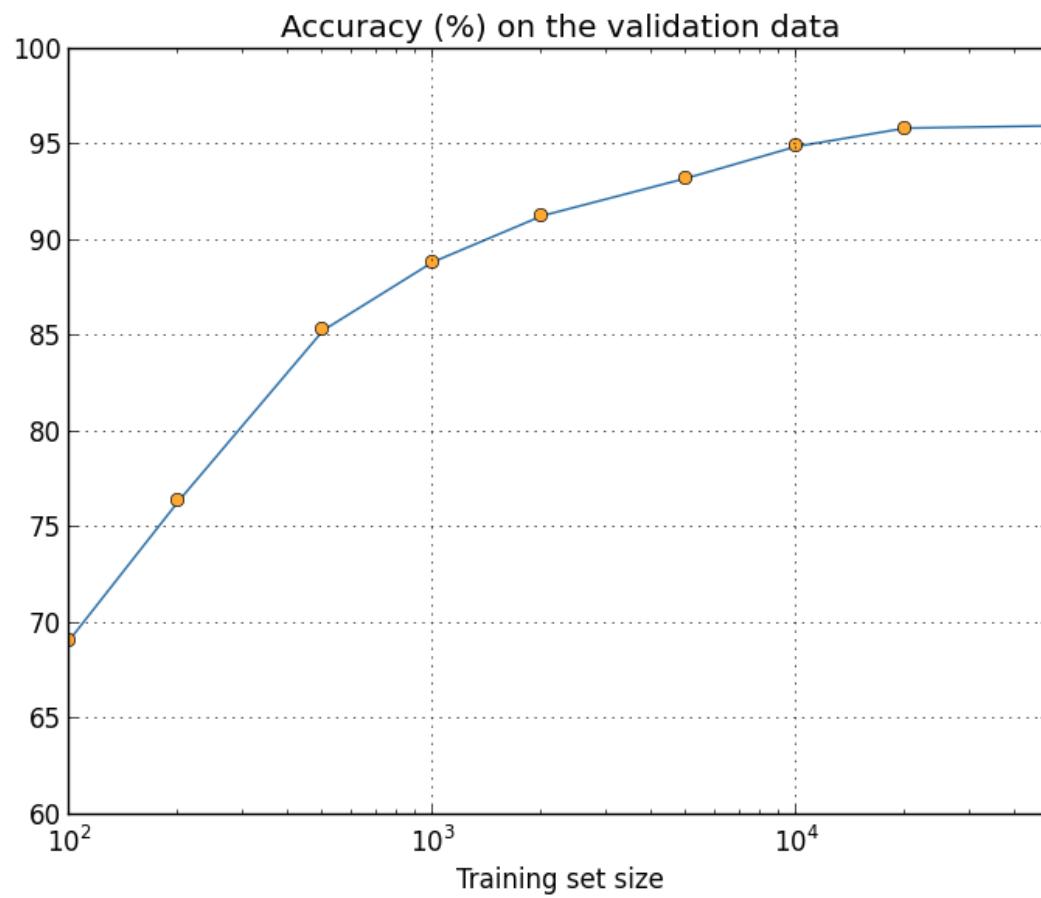
we have this

**Algorithm A  
from H**

**Find  $G(x)$  which is  
approx.  $F(x)$**

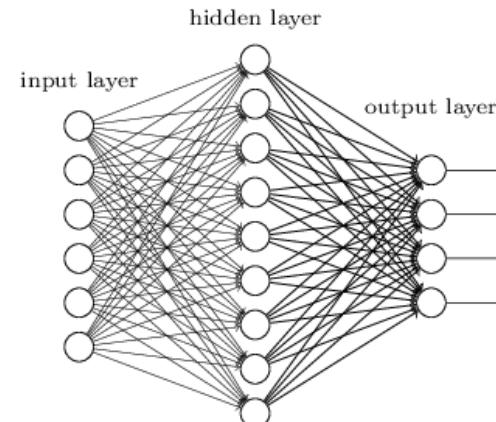
**H: Hypothesis Set:**  
All possible  
algorithms or formulas

Data X



"Non-deep" feedforward  
neural network

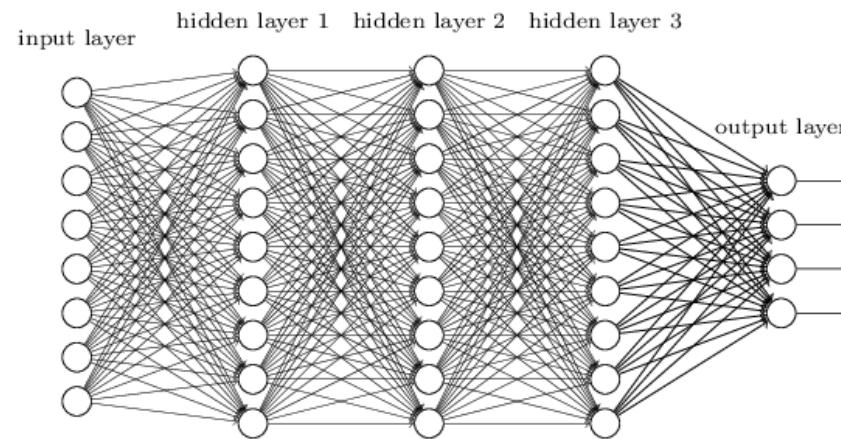
X



Y

Deep neural network

X



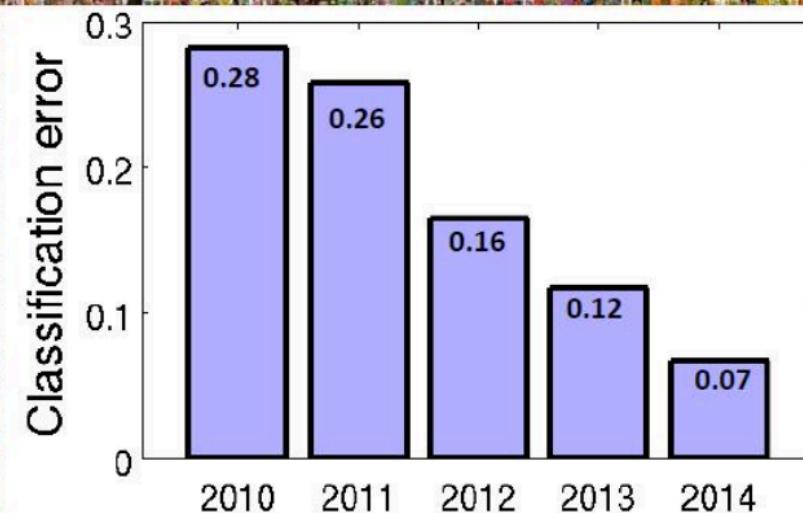
Y



# IMAGENET Large Scale Visual Recognition Challenge

Stanford

The Image Classification Challenge:  
1,000 object classes  
1,431,167 images

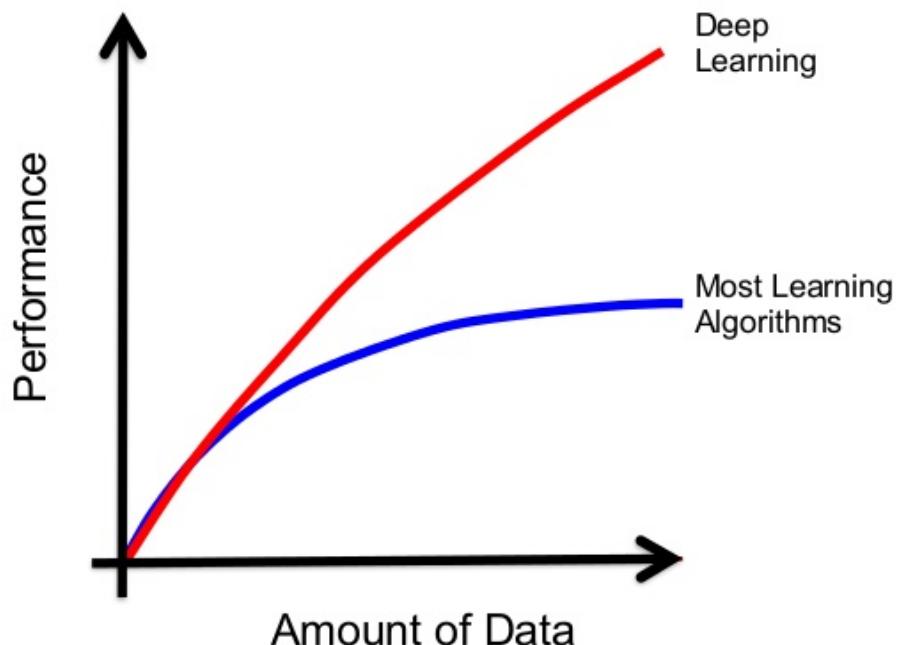


Neural net results are close to human results

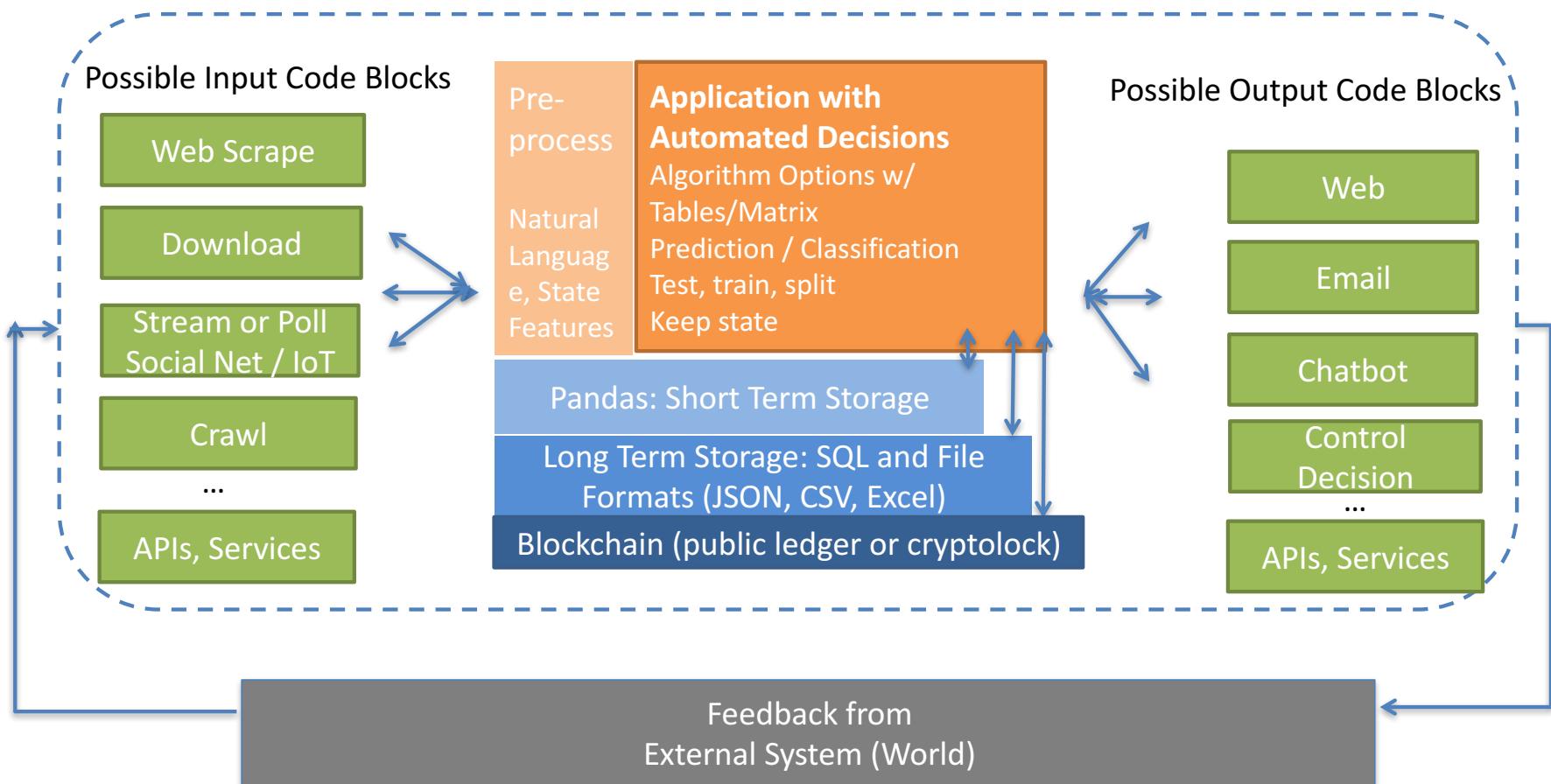
Data X

## **BIG DATA & DEEP LEARNING**

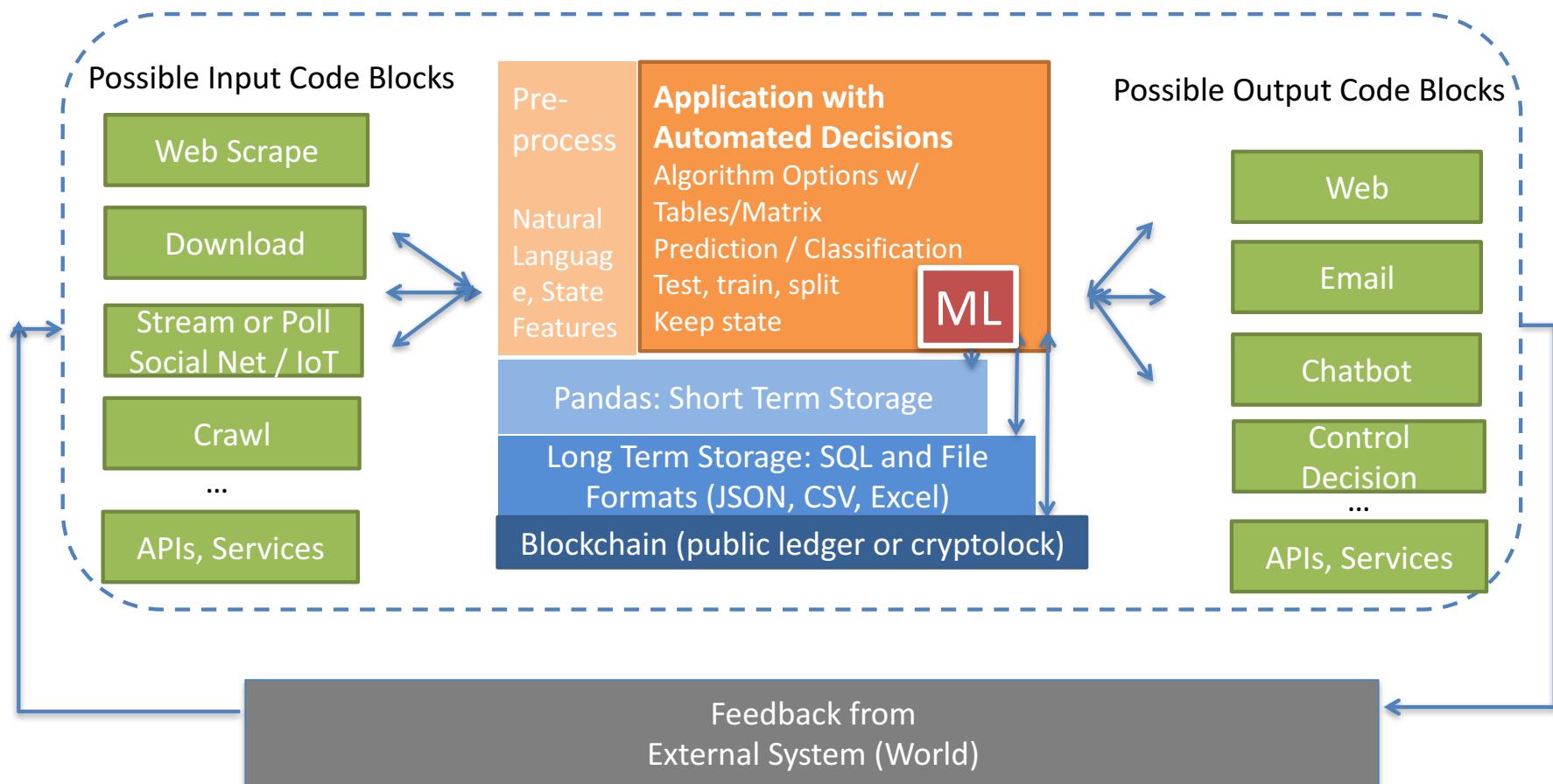
This means  
**Accuracy**



# 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 service 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
- ...



## 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
- Complete and turn in: Python "Review" Breakout and Homework Notebook (BHW).



End of Section

0 0 0 1 0 1 0 1 0 1 1 1 0 0 0 0 0 0 1 0 0 1 0 1 0 1 1 1 0 0  
1 0 1 1 X 1 1 0 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 0 0  
1 Data 0 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 0 0 0 1 0 0