

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

Welcome to Applied Data Science with Venture Applications

Hosted by

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

- Teaching Team Introductions
 - Ikhlaq Sidhu, lead instructor
 - Alexander Fred-Ojala, afo@berkeley.edu (Visiting Scholar)
 - Sana Iqbal , sana_iqbal@berkeley.edu
 - Quinn Thuy Tran, ssquintran@berkeley.edu
 - And many others (coordinators, past students, advisors..)
- Pre-requisites:
1) Working knowledge of Python, 2) Probability and/or Statistics, 3) Know basic matrix multiplication



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 by email/bcourse

Key Dates:

Class Starts today

Final projects: Week of reading Week: Mon-Wed, April 30-May 2, 2018

Top 2 projects will showcase at Collider Cup event on Thursday May 3 or Friday May 4, 2018



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.
5. HW to be emailed by Bcourses lists



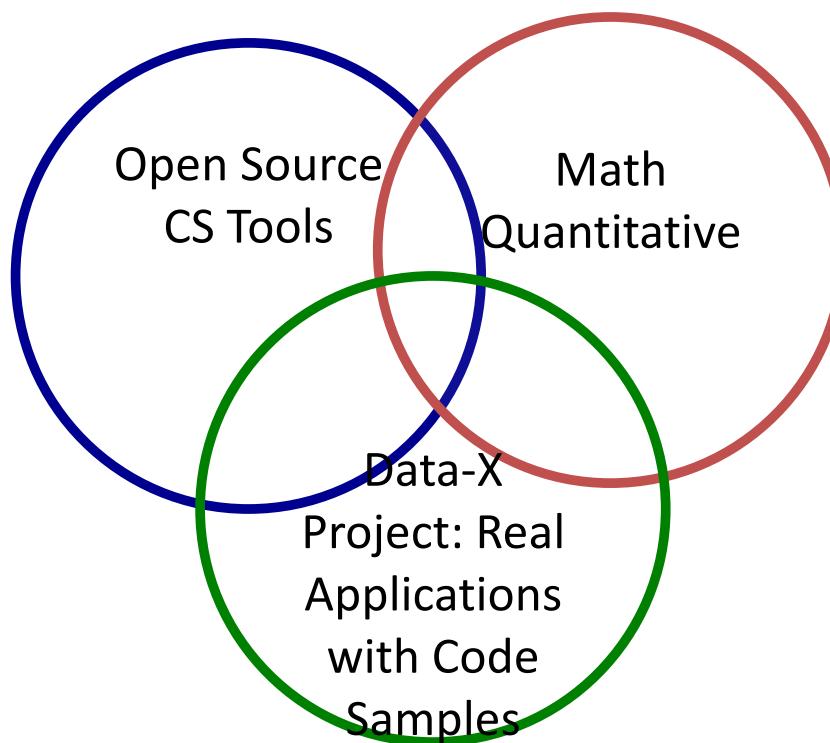
SYLLABUS
[Edit](#)

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

Instructor: Ikhlaq Sidhu
Department of Industrial Engineering & Operations Research
3 Units, Lecture and Lab



What is in this course



Holistic Perspective: Industry, Social Applications, Customer Driven



What is in this class?

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

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 actually in this class?

- The **ML stack** use most commonly used in creating ML/AI/Data applications
- Application and **systems** viewpoint of data and ML
- **Implementation**, architecture, and relevant process to build anything
- Statistical, rule based, and **hybrid** decision systems
- Connection with relevant mathematical foundations (entropy, correlation, spectral, LTI, basic prediction, classification)
- Practical insight into **advanced techniques** and tools: (eg. CNNs, NLP, scraping, recurrent networks, etc.)
- System **modeling** for data applications



Where we will focus:



Make the Tools

Most CS



Use the Tools
(Optimally)

This Course



Architect the System

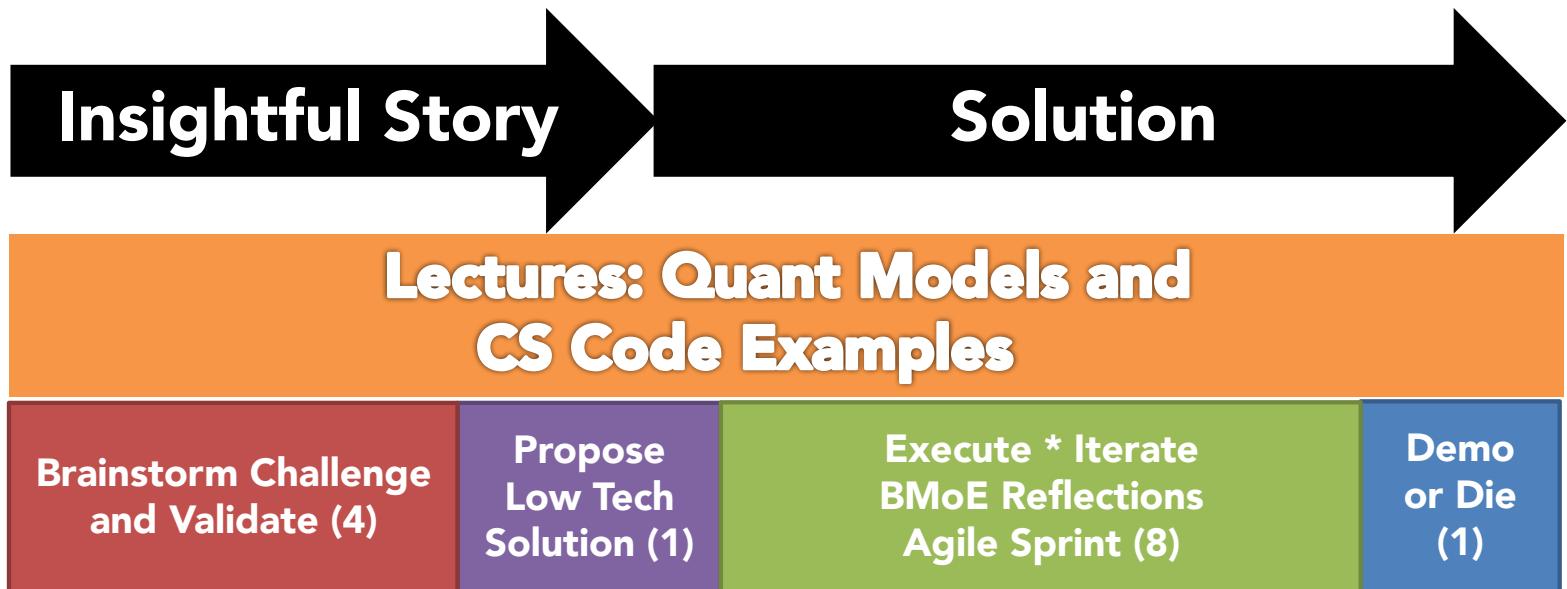


Why and how
you build

Sutardja Center



How the Data-X Course Works:



Team: typically 5 students, with available advisor network



Data X

Introduction Data, Signals, and Systems

Ikhlaq Sidhu
Founding Faculty 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



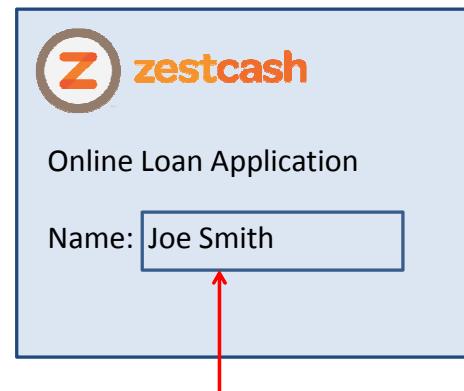
- Data Wrangling
- In Production



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!



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



Harrah's Casino: Knowing your customer

PLAY & WIN ▶

Reference: Supercrunchers

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..
Eg Walmart, Baseball, Sports
7. Efficiency or better performance per dollar cost.
E.G. General IT, SAP, etc
8. Risk Management, regulation, and compliance
Eg. Compliance 360



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 ML framework flow. It starts with 'In Real Life' objects like people and products, which are converted into a dataset with features (like age, sex, occupation) and observed results (like savings). This dataset is divided into 'In Sample' (used for training) and 'Out of Sample' (used for testing/predictions). The observed results lead to characteristics, patterns, and models. The out-of-sample data leads to predictions, similarities, differences, and distances.

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

In Sample

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, ...)

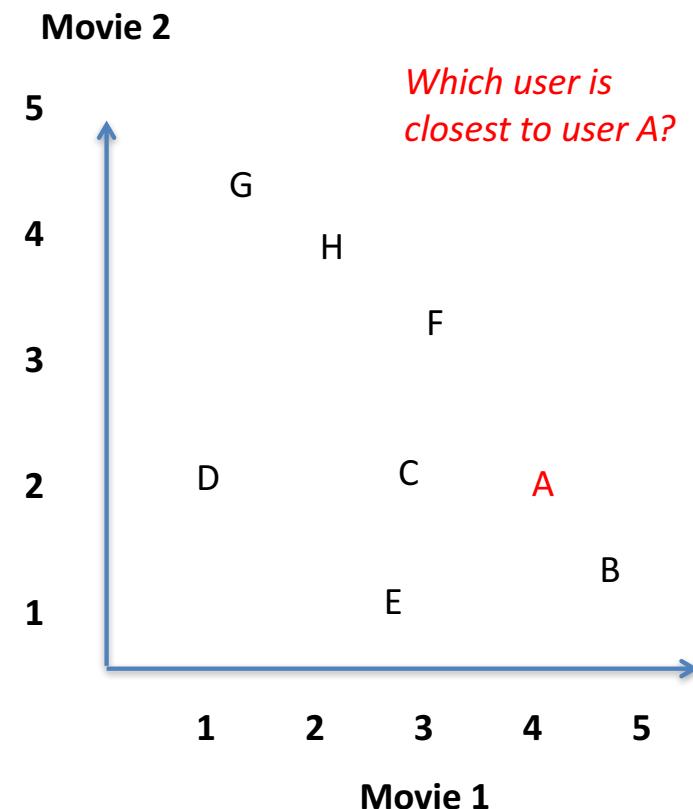


Data X

A Fundamental Idea: From Table to N- Dimensional Space

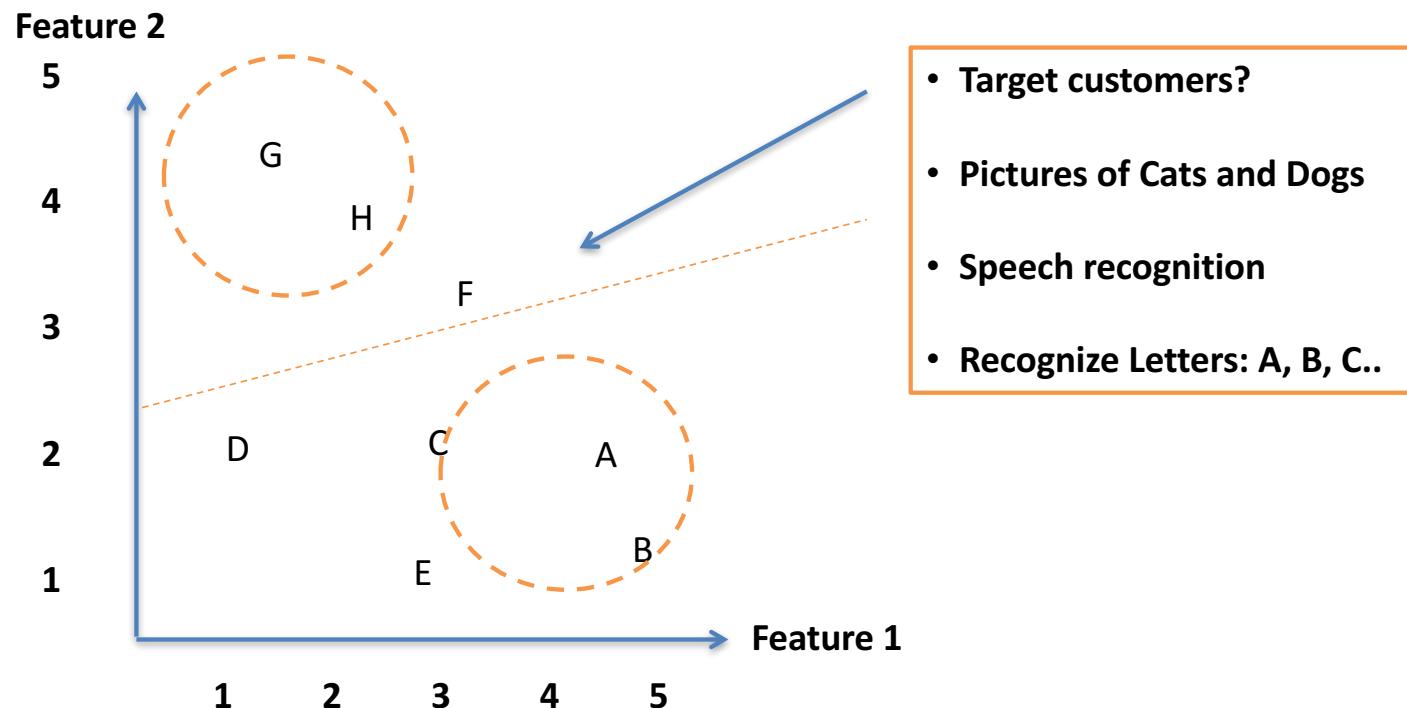
$X =$

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 to Classification



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
..	..



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

H: Hypothesis Set:
All possible algorithms or formulas

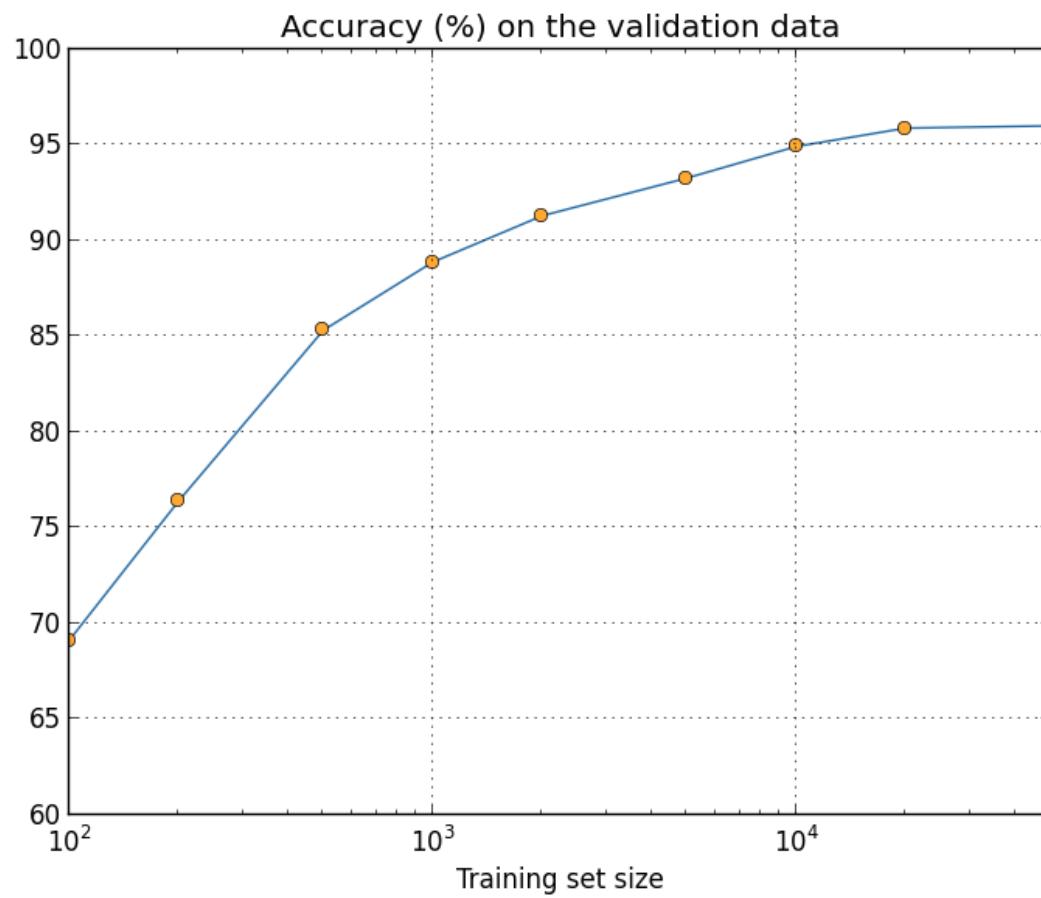
**Algorithm A
from H**

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

a) Supervised ML – as shown

b) Unsupervised learning: training data
Reinforcement learning: done by simulation

Data X



Neural Networks as
Function Approximators

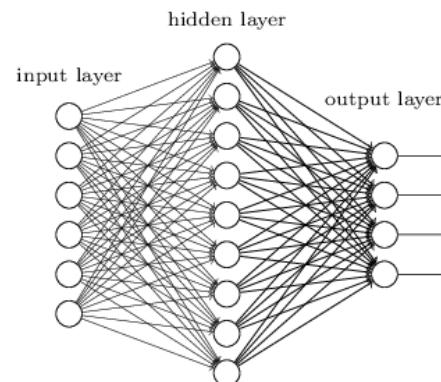


ML Algorithms Guess
this function $F(x)$

Y

"Non-deep" feedforward
neural network

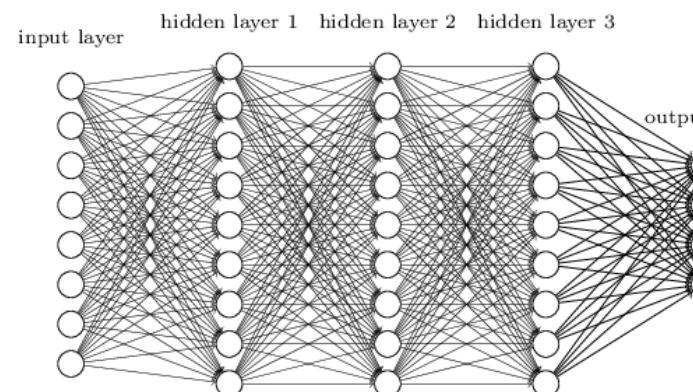
X



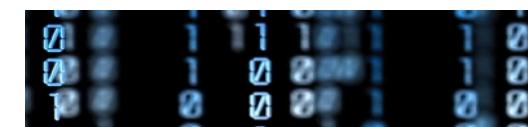
Y

Deep neural network

X



Y



Neural Networks can also
Perform regression

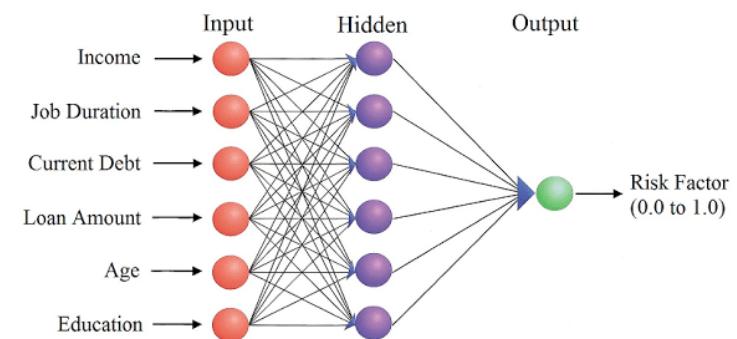
X →

ML Algorithms Guess
this function F(x)

Y →

"Non-deep" feedforward
neural network

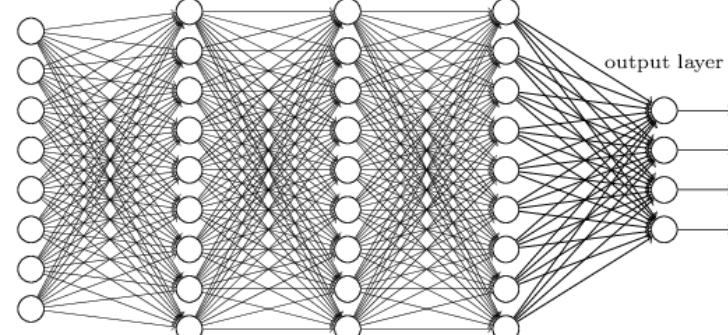
X



Deep neural network

X

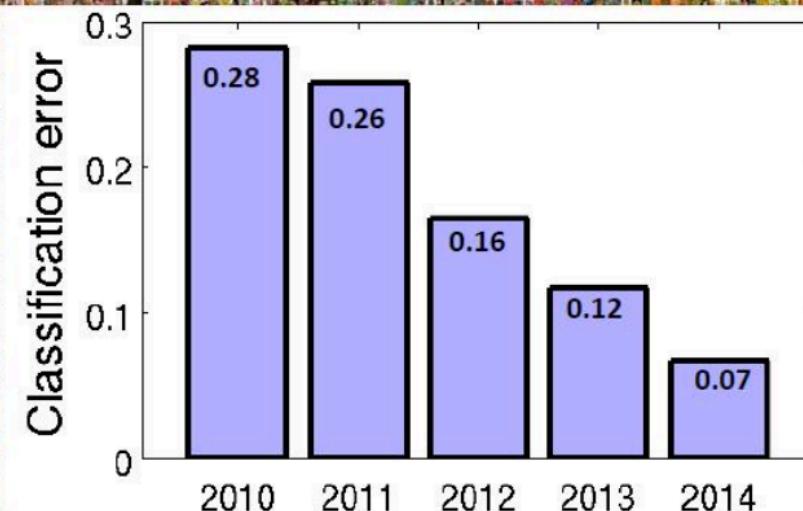
input layer hidden layer 1 hidden layer 2 hidden layer 3



IMAGENET Large Scale Visual Recognition Challenge

Stanford

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

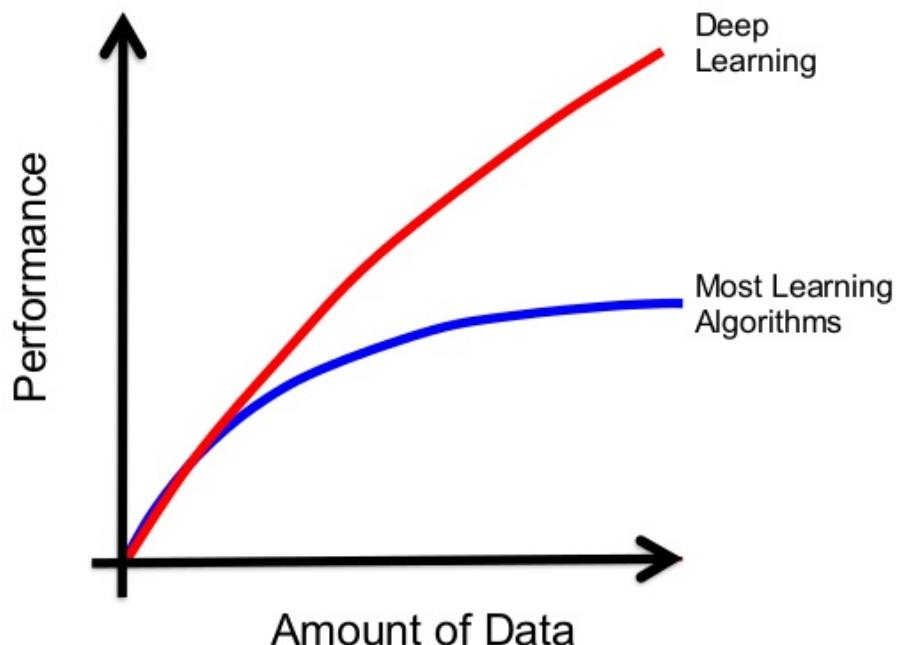


Neural net results are close to human results

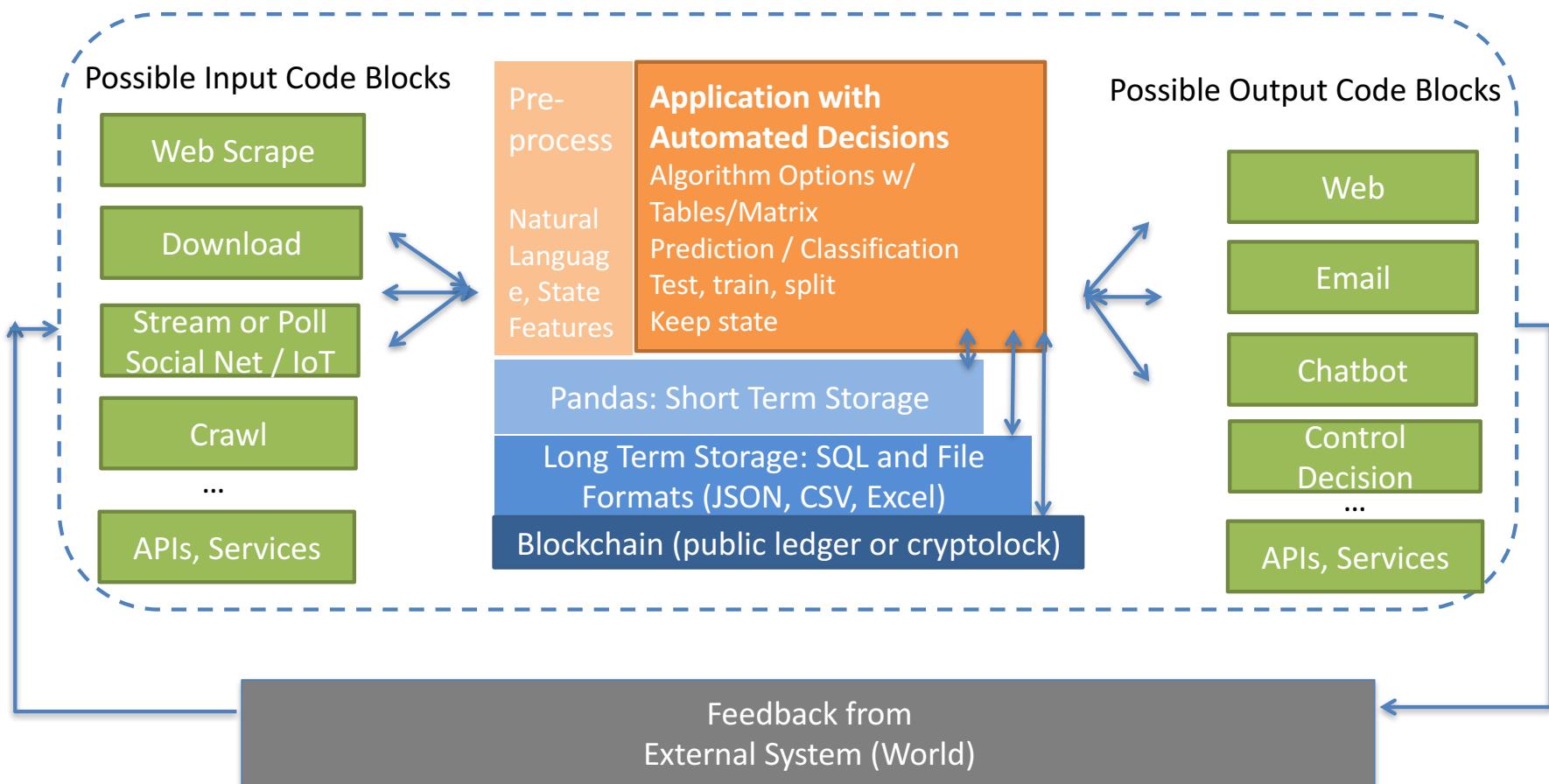
Data X

BIG DATA & DEEP LEARNING

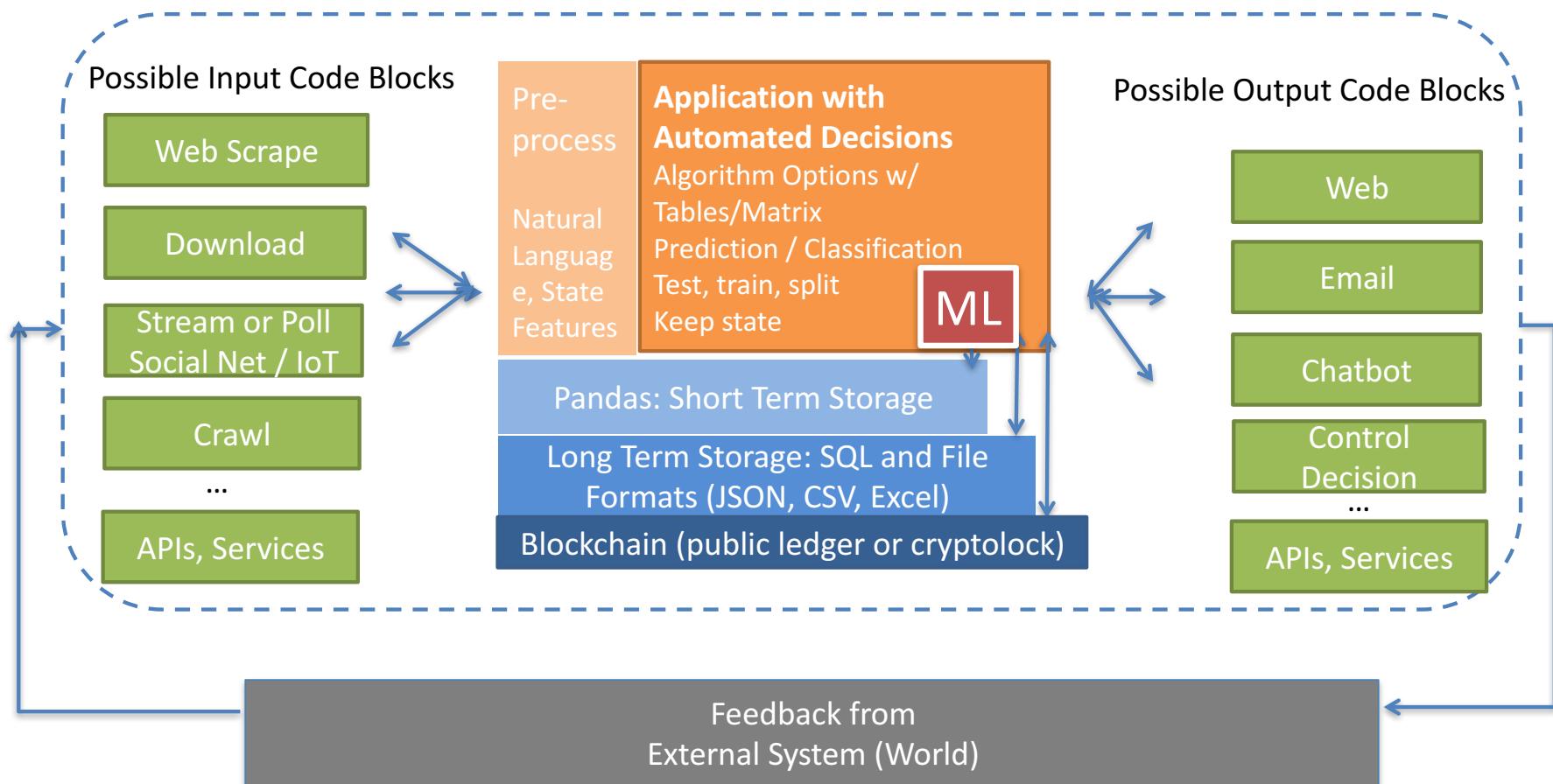
This means
Accuracy



The Data-X System View



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



Project Types

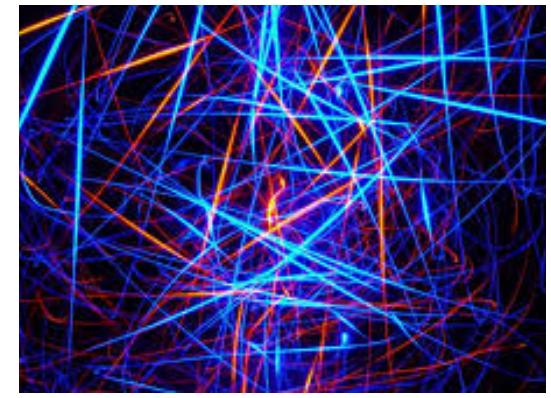


Business or Consumer
Use Case



Social Impact

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



Its Just Cool



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,
 - i.e., just the part that automatically collects data by web scraping, or
 - just the part that makes a decision based on data already available

Home Resources Syllabus Posts Labs Advisors Contact

▼ BLOCKCHAIN ADVISING: BLOCKCHAIN AT BERKELEY

Project Concept Links:

- New Venture Success ([link](#))
- Concept: Blockchain based social currency to regulate social platform such as Twitter ([link](#))
- Concept: Personal Genome Hacking ([link](#))
- Concept: Holy Grail of Venture Capital ([link](#))
- AI Music Software development student cooperation opportunity ([link](#))
- Concept: Predicting future outcomes based on historical records ([link](#))
- Concept: US Power Plant project ([link](#))
- Concept: Multi-disciplinary data analysis of common psychological conditions ([link](#))
- Concept: Visualizing investment opportunities in touristic regions (Open Data for Greece 1.0) ([link](#))
- Concept: The University Bot ([link](#))
- Concept: Materials Recycling using Machine Learning
- Concept: Insights from Personal Photos
- Fuzzy Joins – A Modeling Discussion for Probabilistic Joins in Data Tables
- Concept: Faculty Research Matching with NLP and ML
- Concept: Inferred Information via Probabilistic Joins

Extended Mentor Network:

- Amir Najian, Geospatial Data Scientist at RMS – Geospatial Machine Learning and uncertainty modeling in Geocoder systems



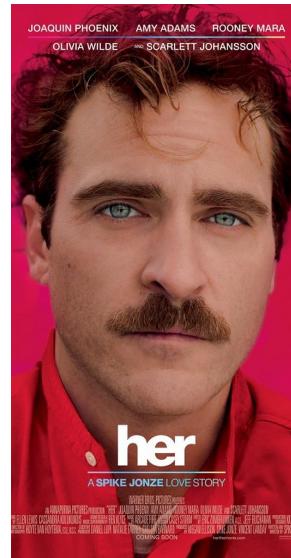
Project Concept: Predicting Energy Prices

- The goal for this project would be to predict future energy prices on a price per kilowatt hour basis for a specific gas fired electric power plant, preferably located in the USA but also open to other countries, assuming fixed competitive supply in its local market over the next 5 years.
- Introduced by Antonio Vitti, Chief Financial Officer and Senior Technology Executive formerly with Merchant Atlas, Inc. See post at data-x.blog



Project Concept: Chatbot with Personality

- How would you go about creating a chatbot that mimics your grandmother.
- Stephen Torres, SCET



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

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 1 1 1 1 0 1 0 1 1 1 0 0
1 Data 0 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1 0 1 1 1 1 1 0 1 0 1 1 1 0 0