CIO Connect Data Science Master Class
From Data Strategy to Implementation
Hong Kong, September 2018

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

Data and AI Approaches

AI, Machine Learning, and Data Science

- What is Machine Learning, Data Science, and AI
- Today's technology in Industry

Ikhlaq Sidhu, content author

Implementation:
SW Tools / Stack

Ikhlaq Sidhu, content author

The Most Common Open Source Tools: AI/ML Stack

Start with Python as an interface
Jupyter Notebooks for prototyping

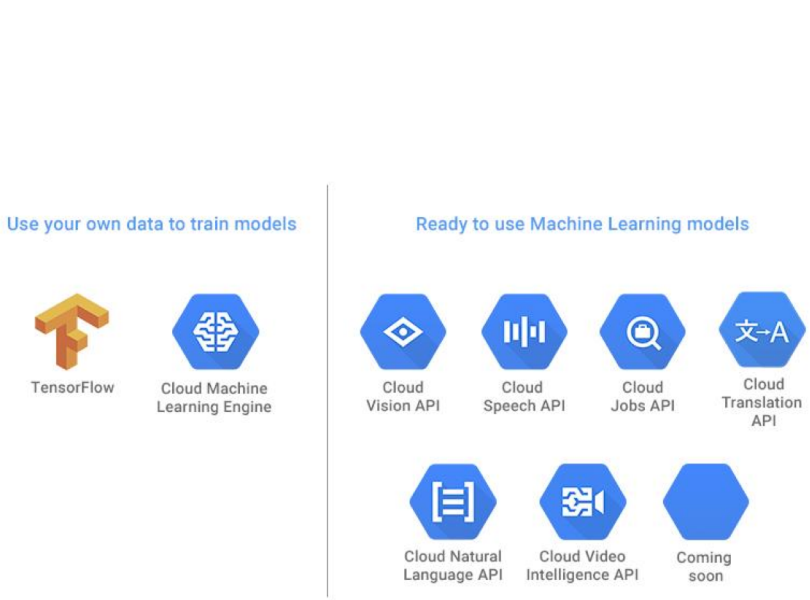
- Python: The interface
- NumPy, SciPy: Working with Arrays
- Pandas: Working in Tables, SQL to Pandas
- Sklearn: ML
- Matplotlib: Visualizing Data
- TensorFlow, Keras: Neural Networks
- SQL to Pandas
- NLP / NLTK: Natural Language
- Spark: For large data sets (GB, TB+)

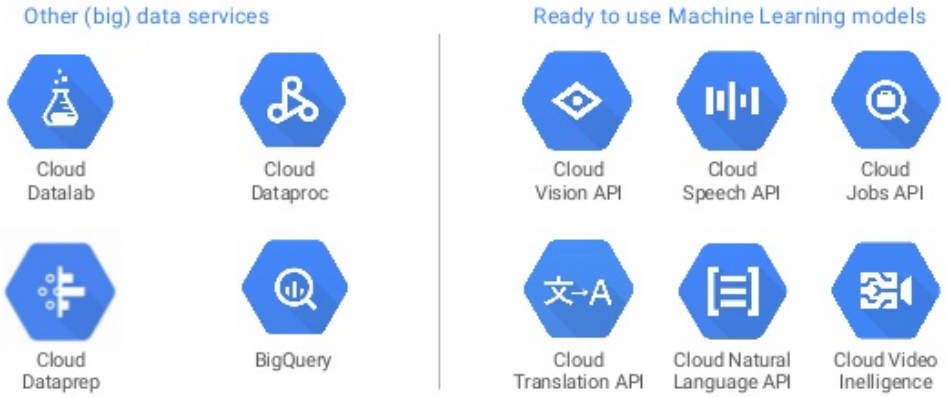
https://www.youtube.com/watch?v=Q0jGAZAdZqM

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

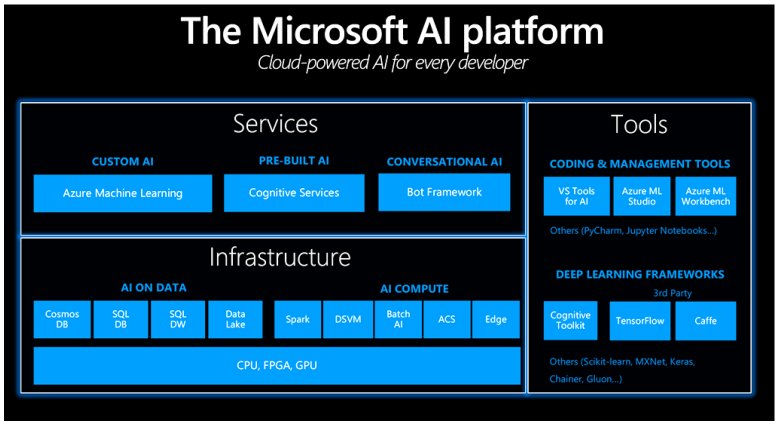
Ikhlaq Sidhu, content author

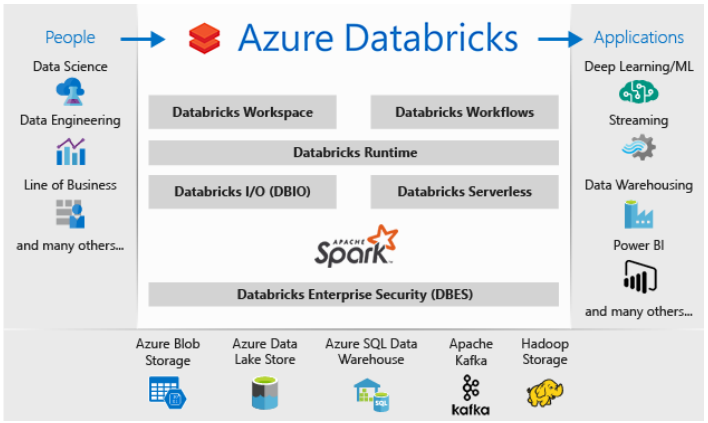
API ML Tools

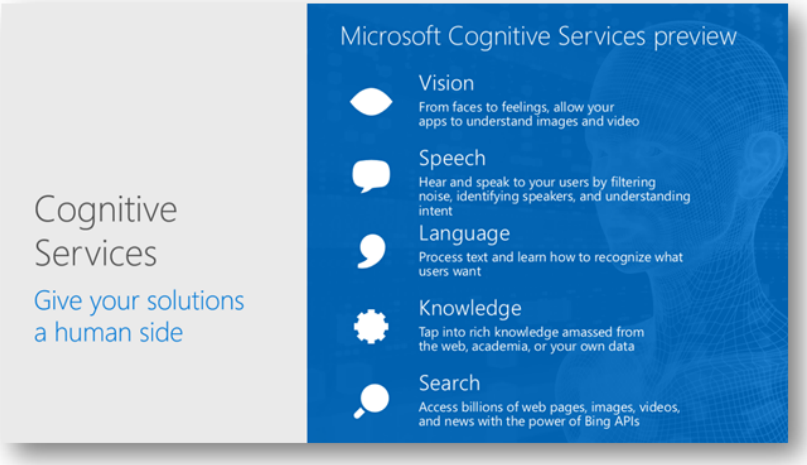


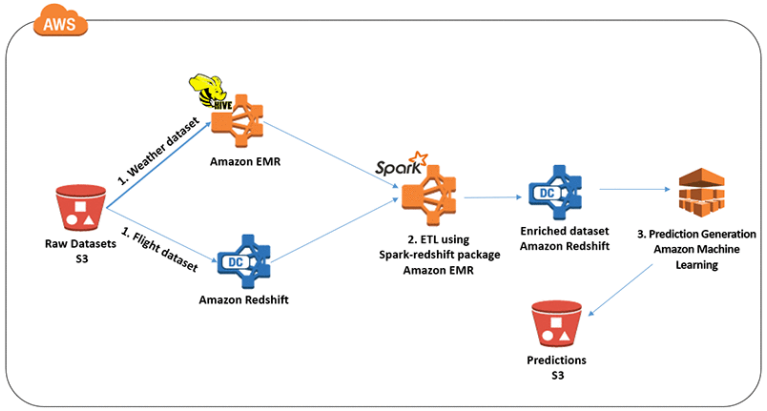


Ikhlaq Sidhu, content author



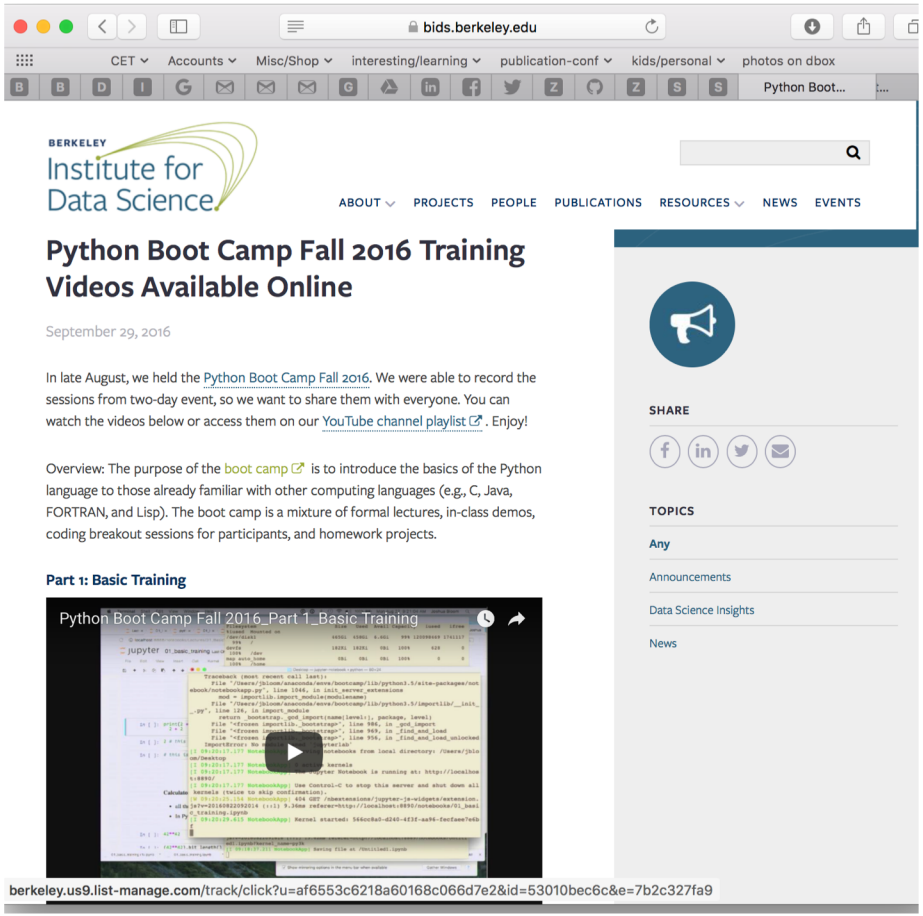


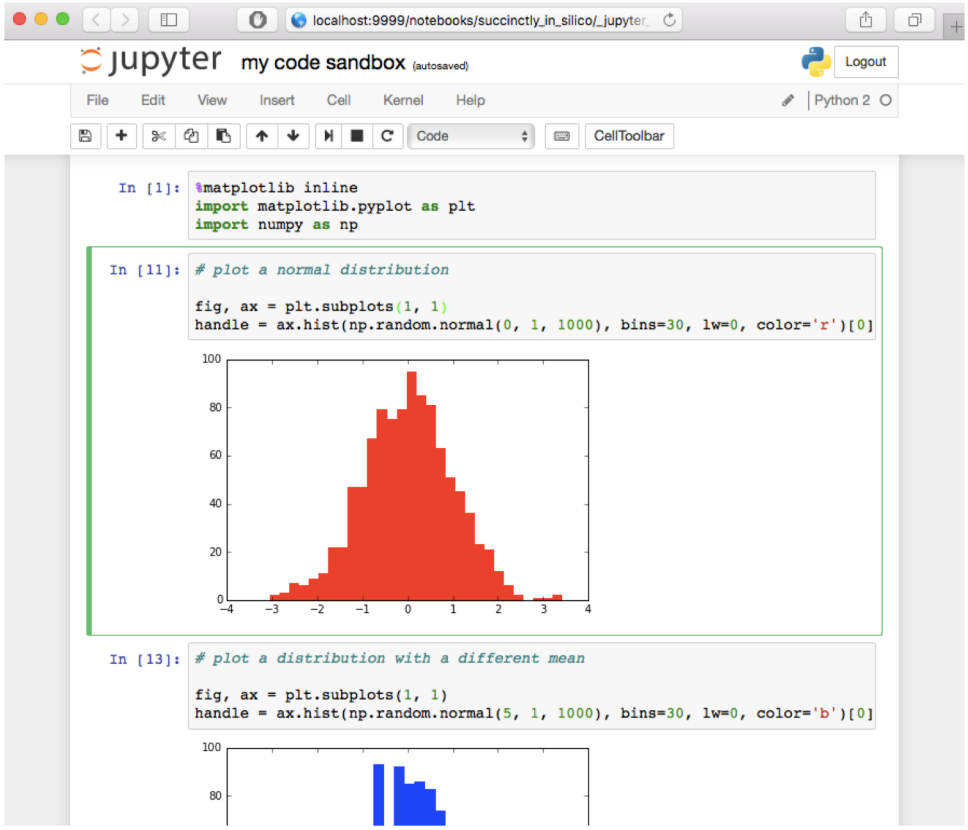




Ikhlaq Sidhu, content author

https://bids.berkeley.edu/news/
python-boot-camp-fall-2016-training-videos-available-online





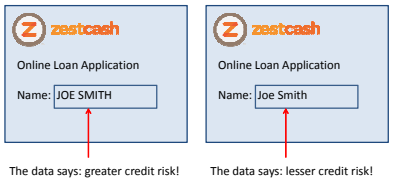
Ikhlaq Sidhu, content author

Where Does Data Come From?

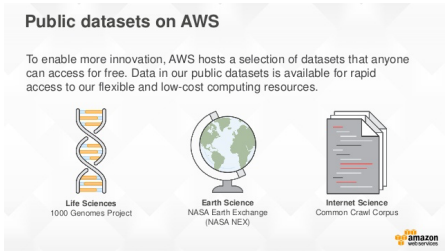
Where Does Data Come From?

Real-life Example: ZestCash

- "All data is credit data"

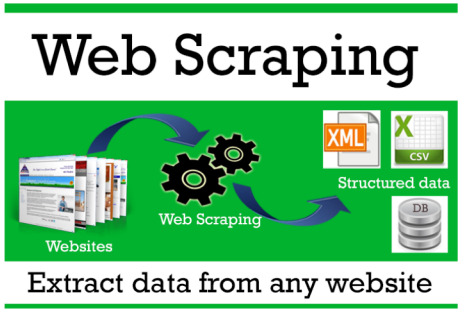


Your Own Web Site



Public Data Sets
Stock market, etc.

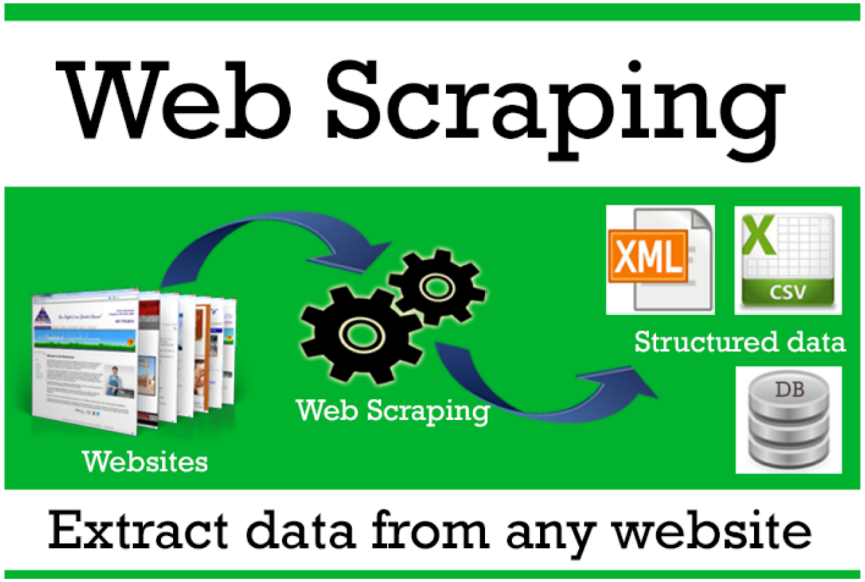
IOT/Sensors



Other Web Sites

Ikhlaq Sidhu, content author

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

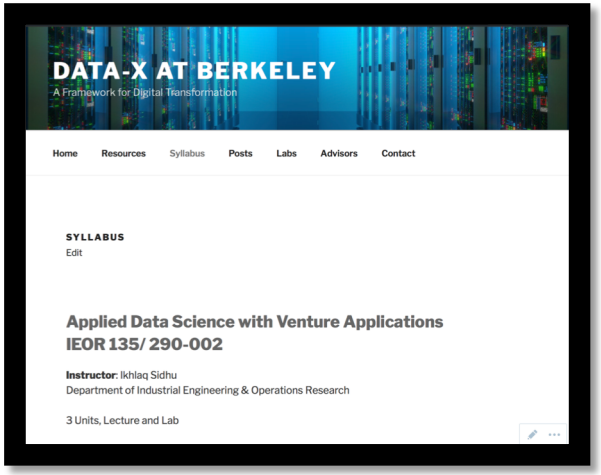
https://github.com/ikhlaqsidhu/data-x/tree/master/03-tools-webscraping-crawling_api_afo

Ikhlaq Sidhu, content author

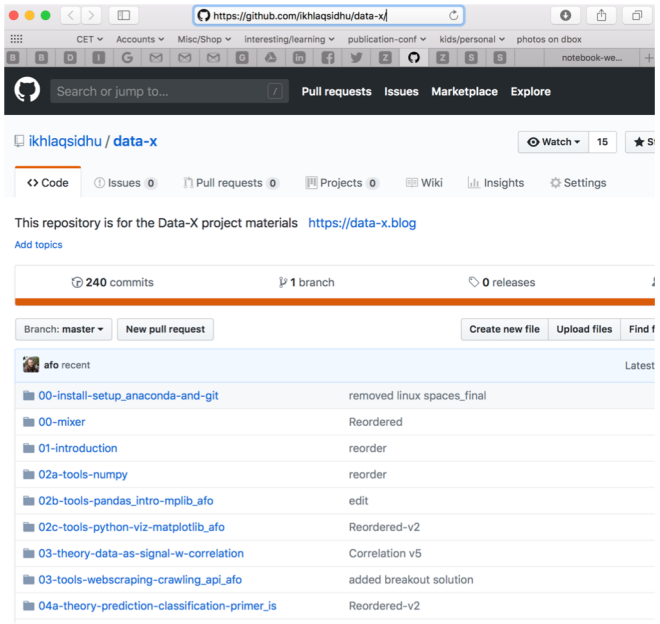
Many Course Resources Are Already Available at

For students and mentors

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



data-x.blog



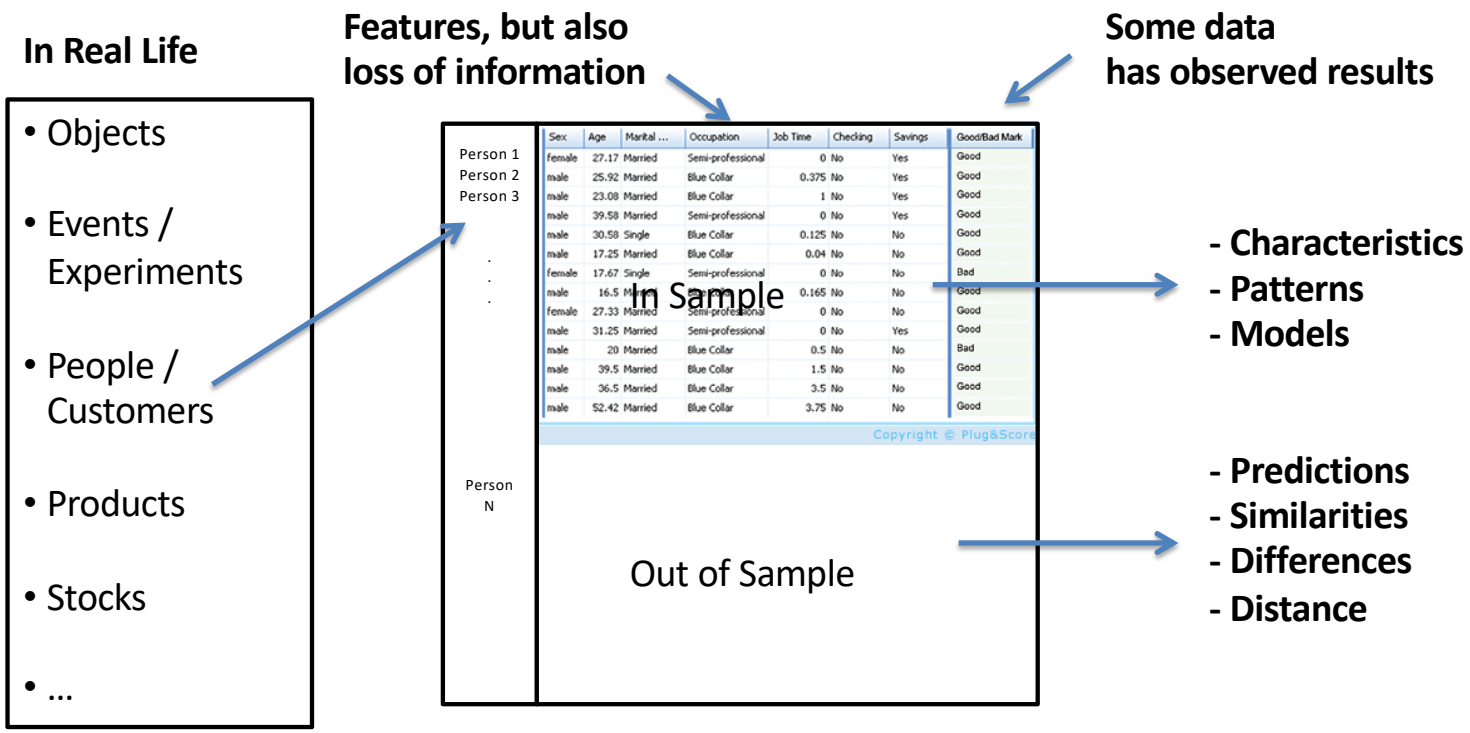
https://github.com/ikhlaqsidhu/data-x/

Ikhlaq Sidhu, content author

Formatting Data

Ikhlaq Sidhu, content author

An ML High Level Framework



Ikhlaq Sidhu, content author

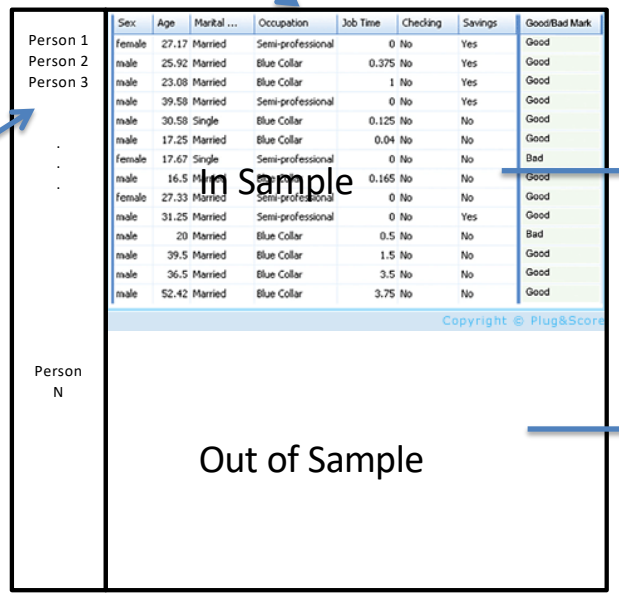
An ML High Level Framework

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, ..)

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

Ikhlaq Sidhu, content author

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
..	..	..	..

X

F(X)

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

Y

Ikhlaq Sidhu, content author

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
..	..	..	..

X

F(X)

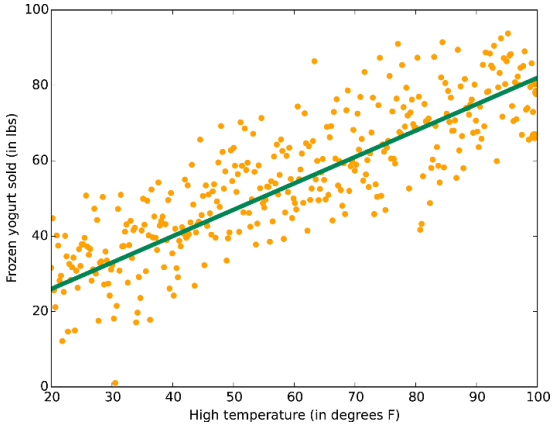
Cust	Credit Score
A	552
B	381
C	760
D	330
E	452
F	678
..	..

Y

```
#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]
```

Ikhlaq Sidhu, content author

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.
```

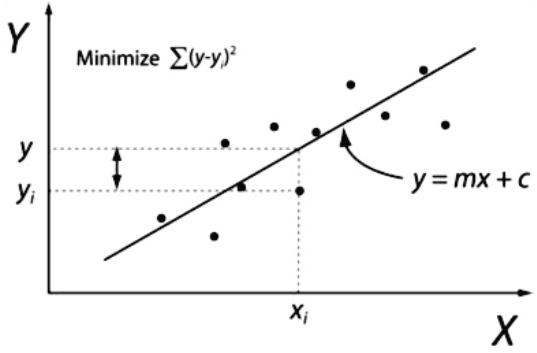
Illustration Source: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice

Prediction

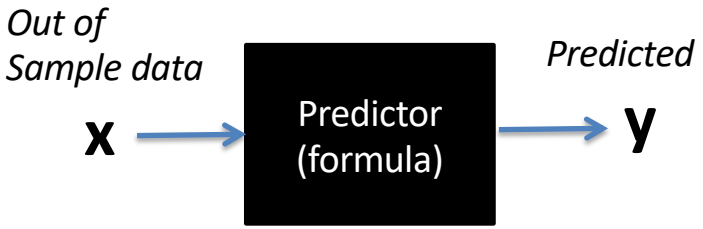
Data We Might Have
(In Sample)

X	Y
2	3
5	9
6	11
8	?
10	?
?	?

Data View



Math View



Our Goal: Working with *out of sample data*

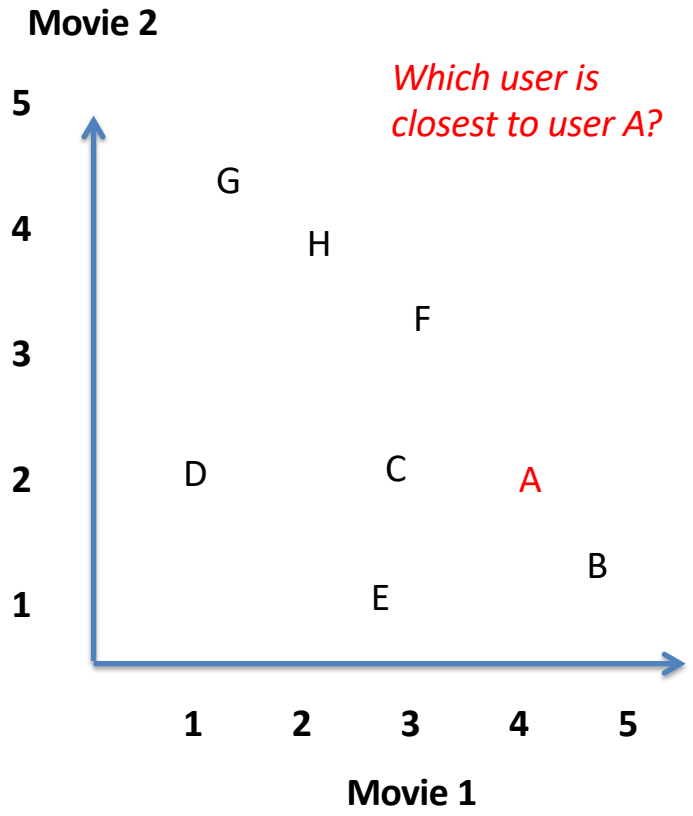
Systems View

Ikhlaq Sidhu, content author

A Fundamental Idea: From Table to N- Dimensional Space

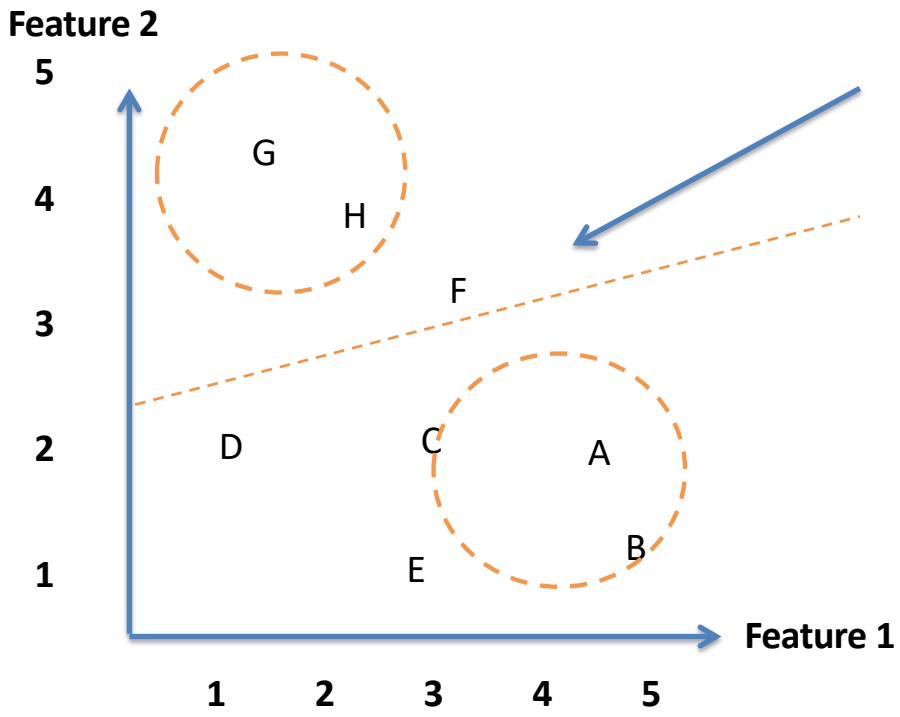
X =

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
..	..	..	..



Ikhlaq Sidhu, content author

Clustering to Classification

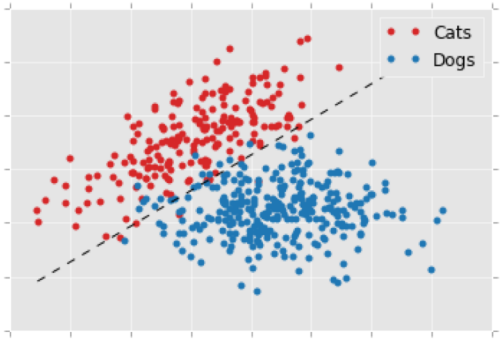


- Target customers?
- Pictures of Cats and Dogs
- Speech recognition
- Recognize Letters: A, B, C..

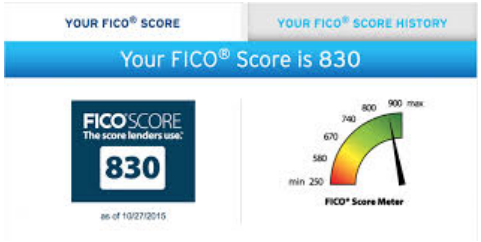
Ikhlaq Sidhu, content author

Traditionally 2 Tasks: Classification & Predictive Scoring

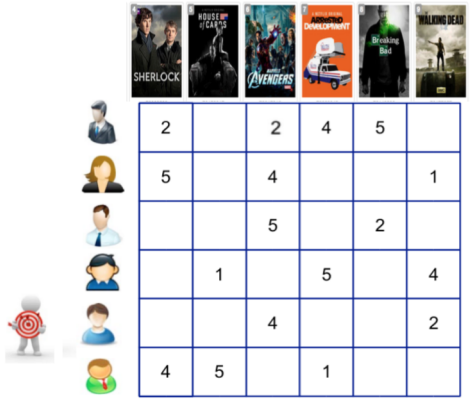
Extracted Data often in Table Format



Classification:
Cats and Dogs, Speech Recognition
Movie Recommendation



Scoring:
Credit Score, Movie Rating
Heath Score, Any Isoquant…



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

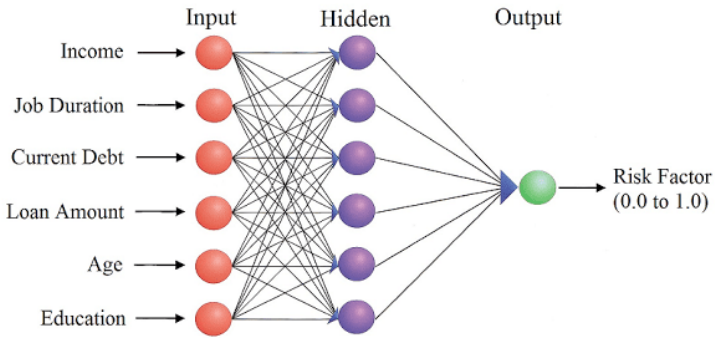
Ikhlaq Sidhu, content author

We have now switched to Neural Networks as Function Approximators

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

"Non-deep" feedforward neural network

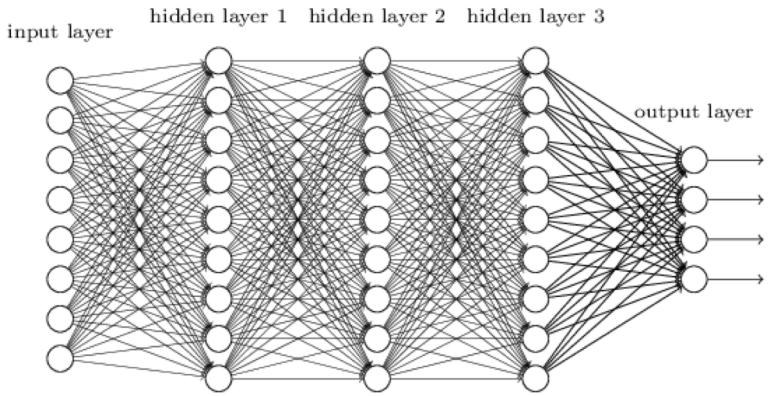
X



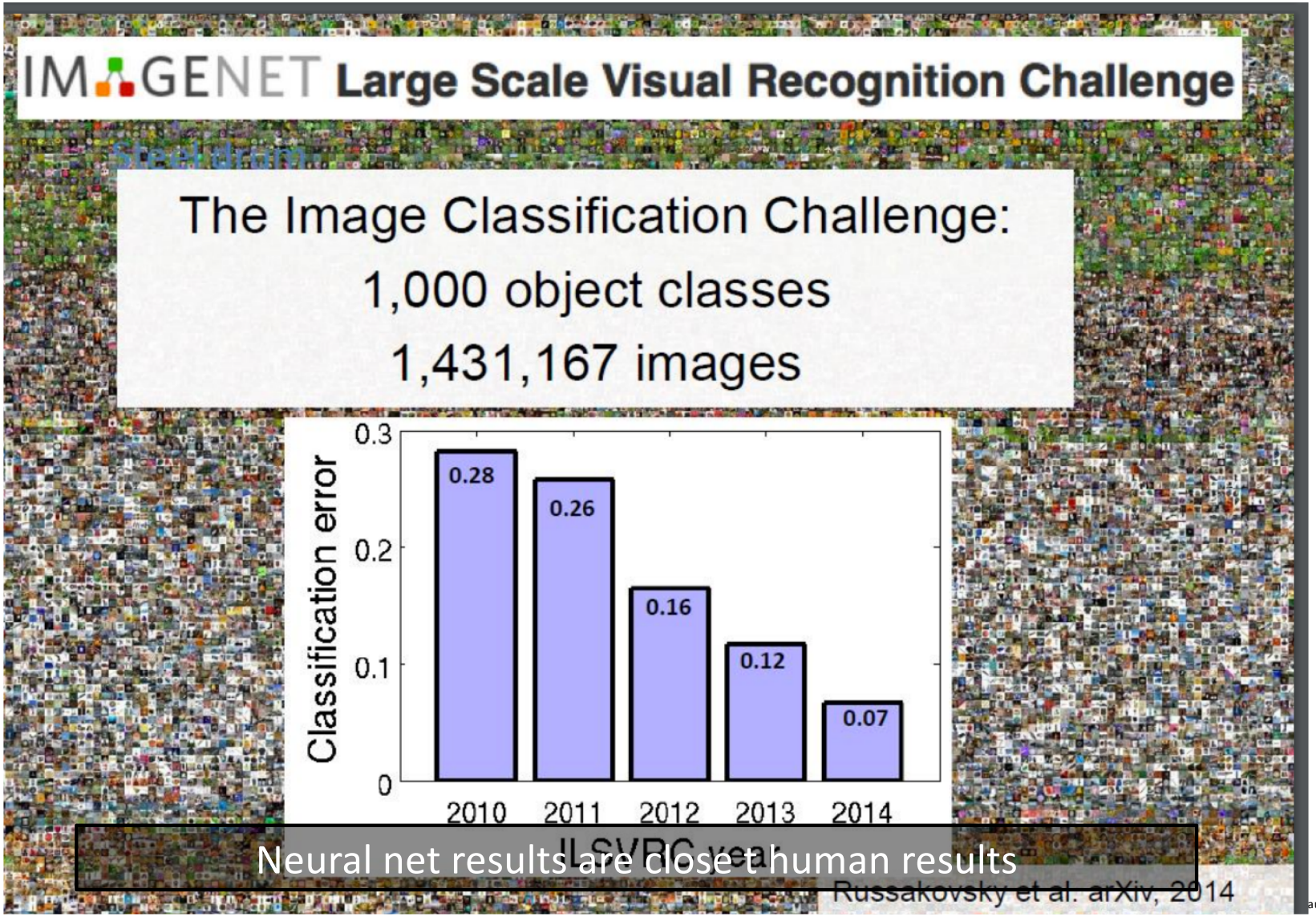
Y

Deep neural network

X



Y

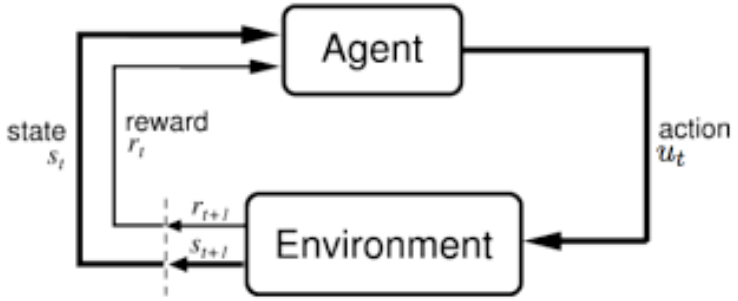

IMAGENET Large Scale Visual Recognition Challenge
The Image Classification Challenge:
1,000 object classes
1,431,167 images
Classification error
0.3
0.2
0.1
0
0.28
0.26
0.16
0.12
0.07
2010
2011
2012
2013
2014
ILSVRC year
Neural net results are close t human results
Russakovsky et al. arXiv, 2014
author

Data and AI Future Directions

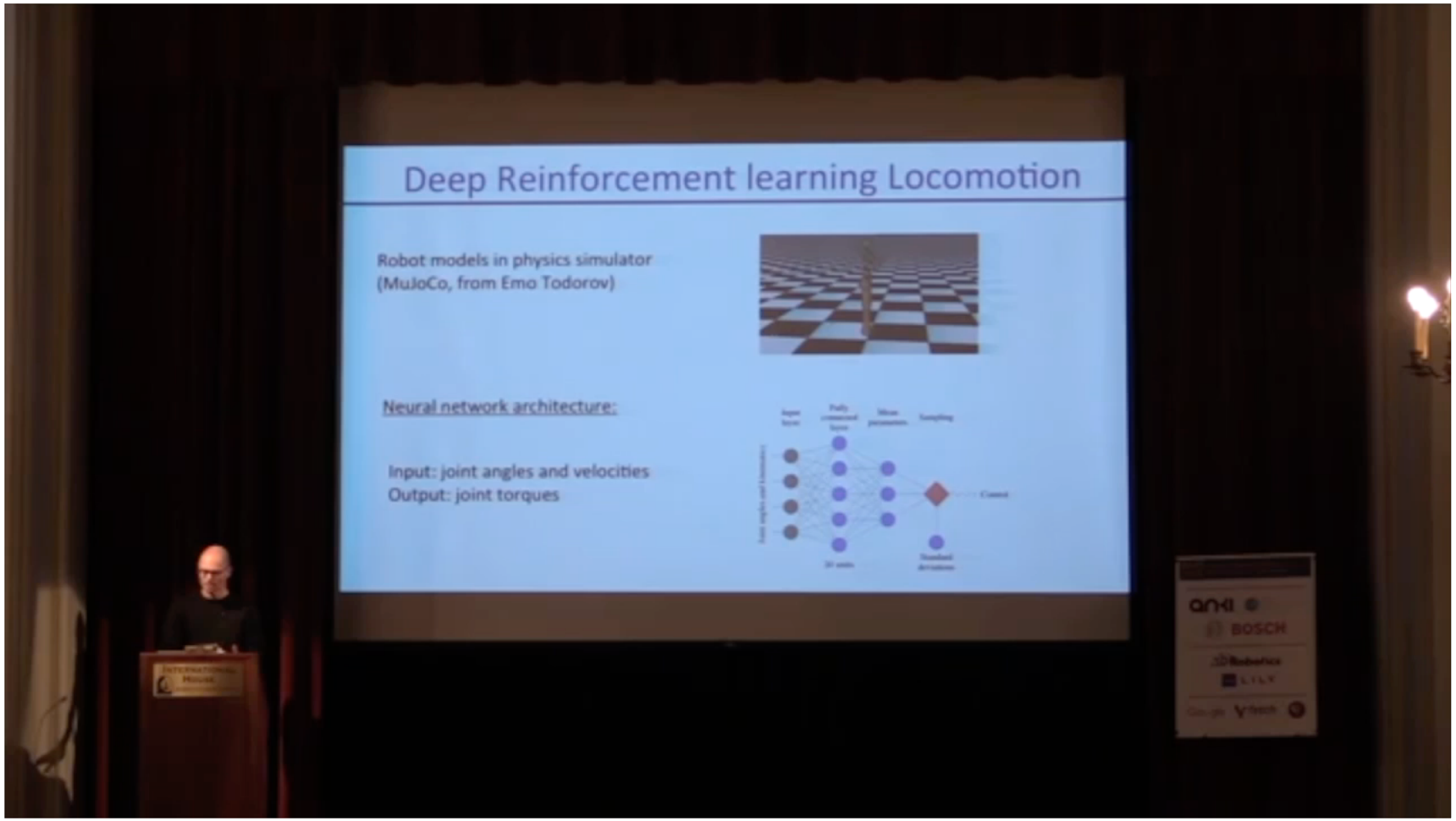
Ikhlaq Sidhu, content author

Peter Abbeel – Deep Reinforcement Learning

Peter Abbeel
Professor at UC Berkeley

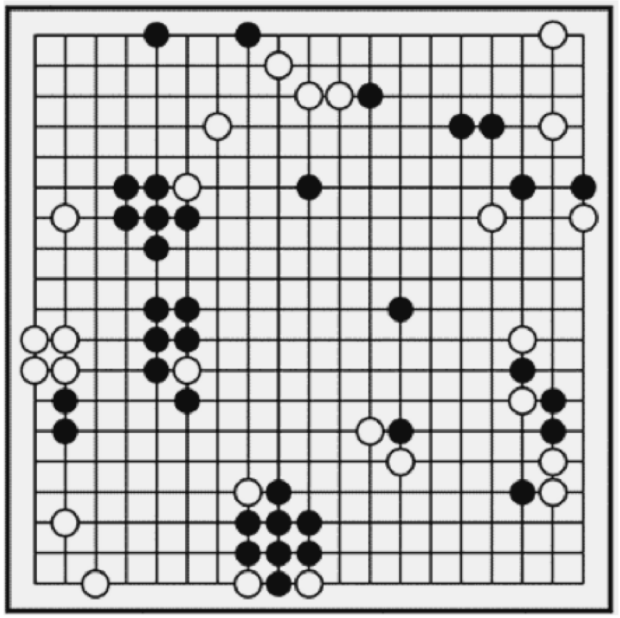


Ikhlaq Sidhu, content author



Ikhlaq Sidhu, content author

Recent AI News



Source: Ken Goldberg, CPAR, People and Robotics Initiative

Ikhlaq Sidhu, content author

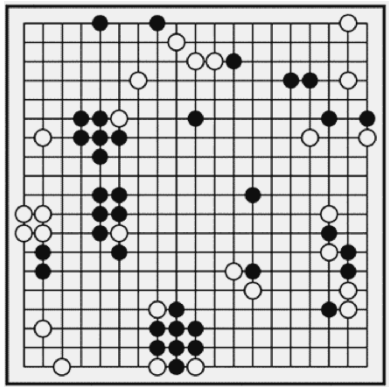
Does this mean AI Can Do Everything Better than Humans

Even then, AI Cannot Solve Real Life Problems Better Than Humans
And in fact, AI Can not even Work without Humans

Ken Goldberg
Leading AI Researcher at Berkeley

Professor and Department Chair, IEOR

William S. Floyd Jr. Distinguished Chair



discrete single agent
fully observed finite

continuous multi-agent
uncertain infinite time horizon

Ken Goldberg UC Berkeley

Ikhlaq Sidhu, content author

AI Systems Only Work because of Human are Part of the System

Google Operations

Massive Data

Result

Feedback By clicks

People Write Web Pages

People at Google Tune the Results

People Click on What They Want

There is no “Intelligence”, “Desire”, or “Existence” in AI without People
There are only people who “invest in, design and operate the machines”

Acknowledgement to Ken Goldberg UC Berkeley

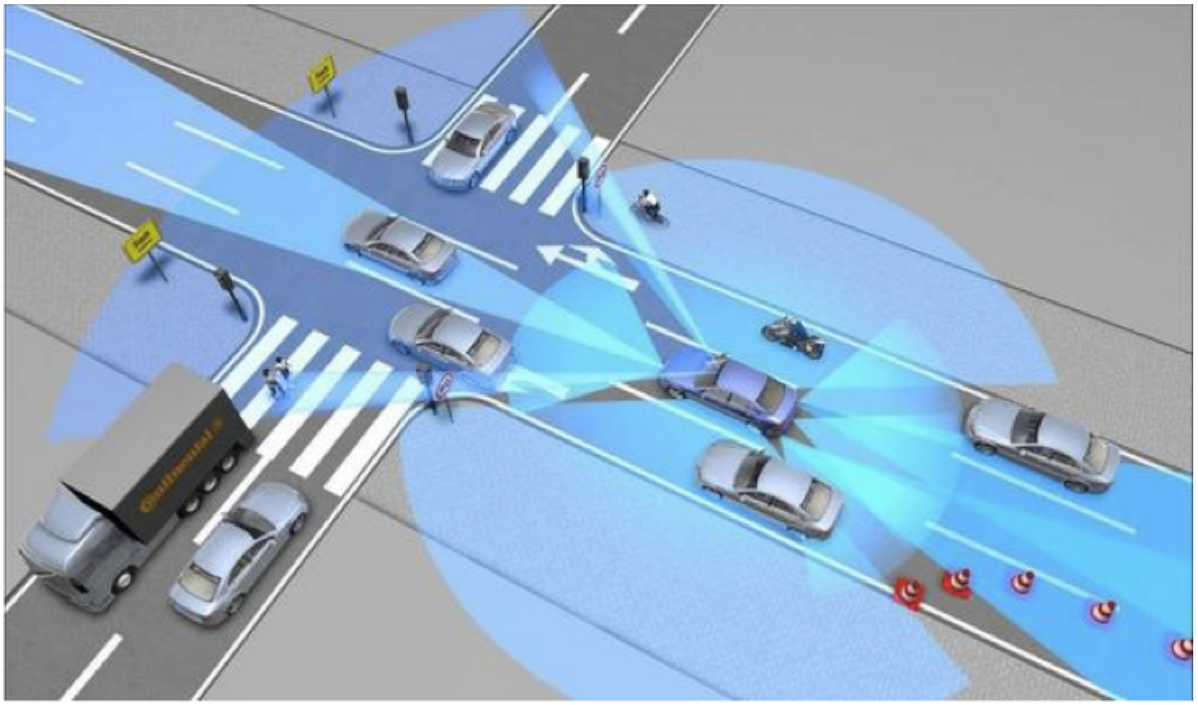
Ikhlaq Sidhu, content author

At Berkeley, we have a lot of research on "How Machines Will Work as Part of Larger Systems that Work with People"

37 faculty

My Drive Home From Berkeley

Autonomous Driving and Driver-Assist



- Communicating intent
- Driver-in-the-loop modeling
- Two-way learning: knowledge transfer between vehicle and driver
- Safety in autonomous and assisted driving

Principal investigators:

Trevor Darrell
UC Berkeley

Anca Dragan
UC Berkeley

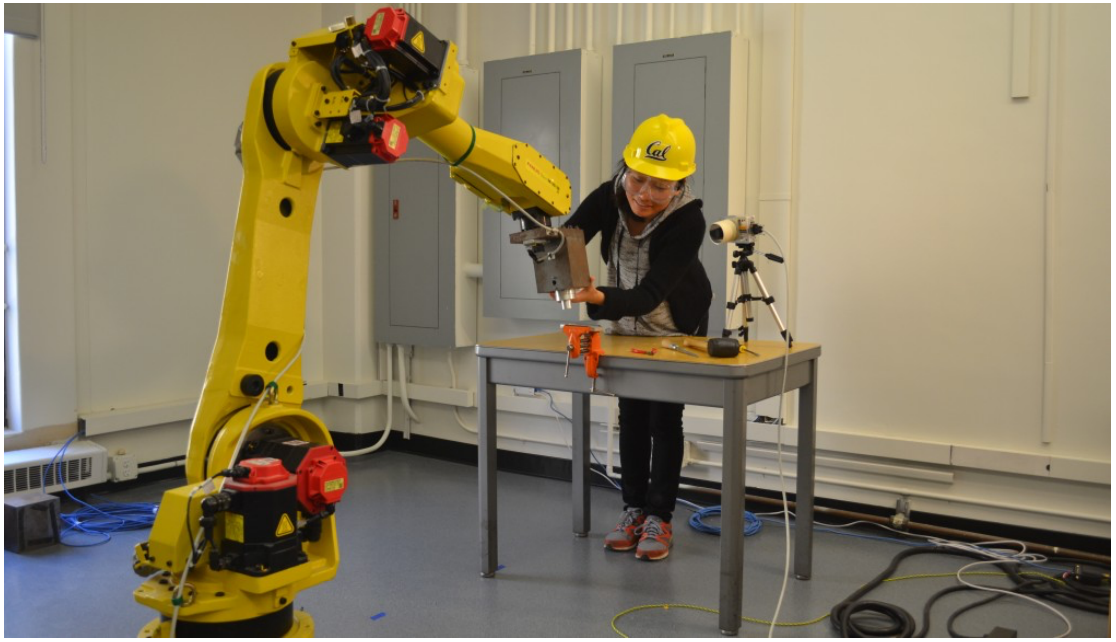
Ken Goldberg
UC Berkeley

Ruzena Bajcsy
UC Berkeley

Francesco Borrelli
UC Berkeley

Source: Ken Goldberg, CPAR, People and Robotics Initiative

Safety in Human-Robot Interaction: Guarantees and Verification



Safety-constrained motion planning for efficiency in factory human-robot interaction

Learning and prediction for safety in HRI

Provably safe human-centric autonomy

Principal investigators:

Claire Tomlin
UC Berkeley

Masayoshi Tomizuka
UC Berkeley

Francesco Borrelli
UC Berkeley

Source: Ken Goldberg, CPAR, People and Robotics Initiative

Most Common Data/AI Research Trends in 2017

- Large-scale machine learning - amounts of data
- Deep learning - recognition, classification
- Reinforcement learning - time sequence, aided by Neural Networks
- Robotics - beyond navigation, to safe interaction
- Computer vision - most prominent perception, better than human
- Natural Language Processing - interacting with people/dialog
- Collaborative systems - autonomous systems w/people + machines using complimentary functions
- Crowdsourcing and human computation – harness human intelligence, uses other AI, vision, ML, NLP, …
- Algorithmic game theory and computational social choice – systems using social computing, incentives, prediction markets, game theory, peer prediction, scoring rules, no regret learning
- Internet of Things (IoT) – using AI to unravel sensory information, interfaces, and protocols
- Neuromorphic Computing – new computing fabrics based on biological models

New Data/AI Systems

http://ai100.stanford.edu/2016-report

Ikhlaq Sidhu, content author

Semantics for AI

AI Umbrella

II: Intelligent Infrastructure

A web on computation, data, and physical entities that make the human environment more supportive, interesting, & safe.

IA: Intelligent Automation

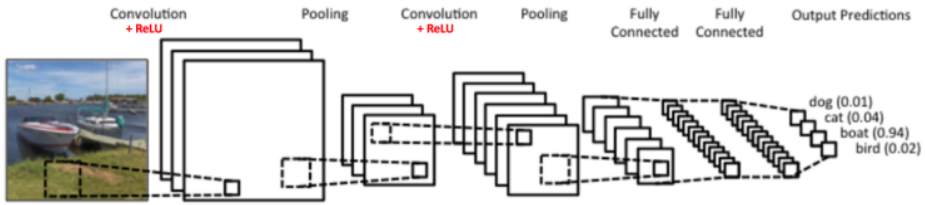
Computation and data are used to create services that augment human intelligence and creativity

Michael Jordan, UC Berkeley

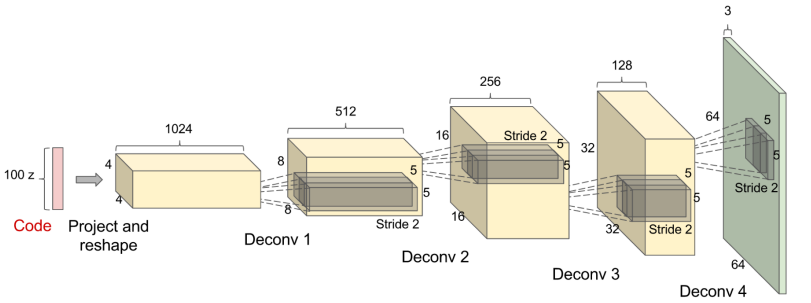
Unsupervised Image to Image

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

Typical CNN converts image to output vector of features



- Ability to generate data *that look real* entails some form of understanding



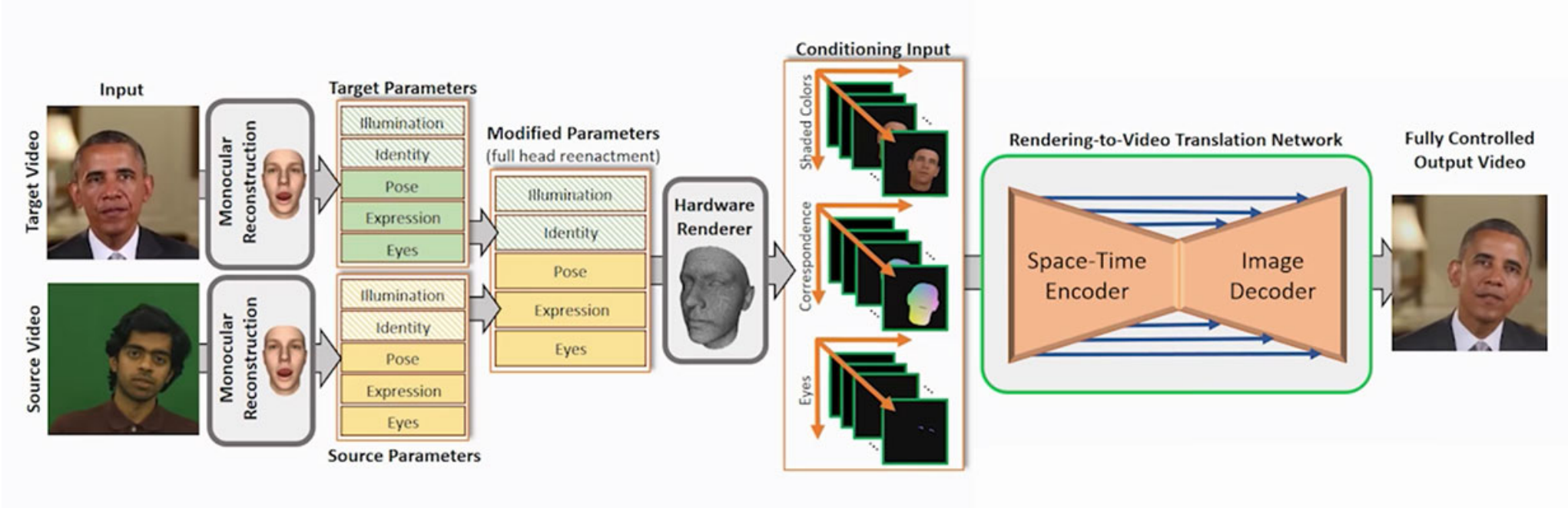
[Radford, Metz & Chintala, ICLR 2016]

IEEE SPECTRUM

Experts Bet on First Deepfakes Political Scandal

Researchers wager on a possible Deepfake video scandal during the 2018 U.S. midterm elections

By **Jeremy Hsu**



Ikhlaq Sidhu, content author

Large Investment and Valuations

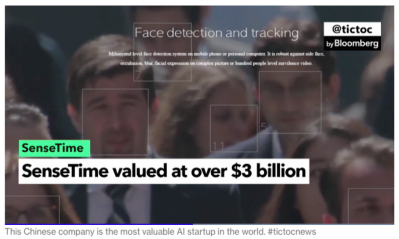
FINANCE • SELF-DRIVING CARS

GM Buying Self-Driving Tech Startup for More Than $1 Billion

China Now Has the Most Valuable AI Startup in the World

Bloomberg News
April 8, 2018, 6:00 PM PDT

- It becomes the world's richest-valued private AI startup
- The company drives China's ambition to dominate global AI



This Chinese company is the most valuable AI startup in the world. #tictocnews

SenseTime Group Ltd. has raised $600 million from Alibaba Group Holding Ltd. and other investors at a valuation of more than $3 billion, becoming the world's most valuable artificial intelligence startup.

'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day

Published time: 1 Sep, 2017 14:08
Edited time: 1 Sep, 2017 14:40
Get short URL

Russian President Vladimir Putin © Alexei Druzhinin / Sputnik

7885 2

Vladimir Putin spoke with students about science in an open lesson on September 1, the start of the school year in Russia. He told them that "the future belongs to artificial intelligence," and whoever masters it first will rule the world.

"*Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world,*" Russian President Vladimir Putin said.

However, the president said he would not like to see anyone "*monopolize*" the field.

"*If we become leaders in this area, we will share this know-how with entire world, the same way we share our nuclear technologies today,*" he told students from across Russia via satellite link-up, speaking from the Yaroslavl region.

TECHNOLOGY NEWS JANUARY 2, 2018 / 11:02 PM / 3 MONTHS AGO

Beijing to build $2 billion AI research park: Xinhua

Reuters Staff 2 MIN READ

BEIJING (Reuters) - Beijing is planning to build a 13.8 billion yuan ($2.12 billion) artificial intelligence development park in the city's west, the official Xinhua news agency reported, as China pushes ahead to fulfill its ambition to become a world leader in AI by 2025.

The AI park will house up to 400 enterprises and have an estimated annual output of 50 billion yuan, Xinhua said, citing a report from authorities in Beijing's Mentougou district.

Ikhlaq Sidhu, content author

Contact:

Ikhlaq Sidhu
Founding Faculty Director, Center for Entrepreneurship & Technology
IEOR Emerging Area Professor, UC Berkeley
sidhu @berkeley.edu, scet.berkeley.edu