Lecture 1
Introduction to machine learning: applications

GEOL 4397: Data analytics and machine learning for geoscientists

Jiajia Sun, Ph.D.

Jan. 15th, 2019

UNIVERSITY of HOUSTON
YOU ARE THE PRIDE
EARTH AND ATMOSPHERIC SCIENCES

Today's agenda

- Data analytics workflow
- Machine learning applications
- Machine learning: what & why
- Machine learning applied to geoscience
- Course overview & Policy

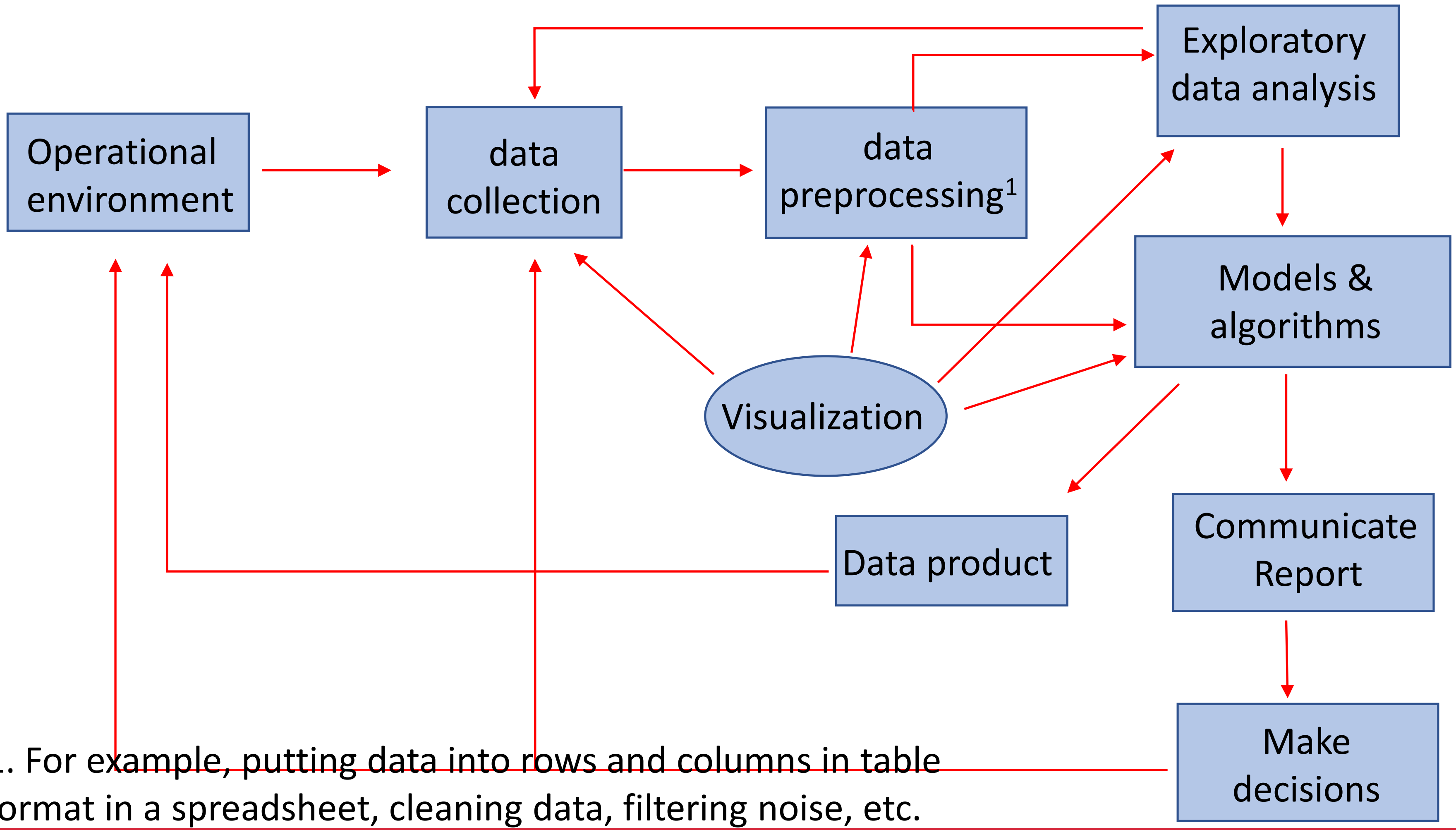
What is data analytics?

- a.k.a data analysis
- A process of converting raw data into actionable intelligence (or information, knowledge, insights)
- For better-informed business decision-making in commercial industries,
- For researchers to answer questions, verify or disprove models, theories and hypotheses.

1.http://searchdatamanagement.techtarget.com/definition/data-analytics
2. https://en.wikipedia.org/wiki/Data_analysis

Data analytics workflow

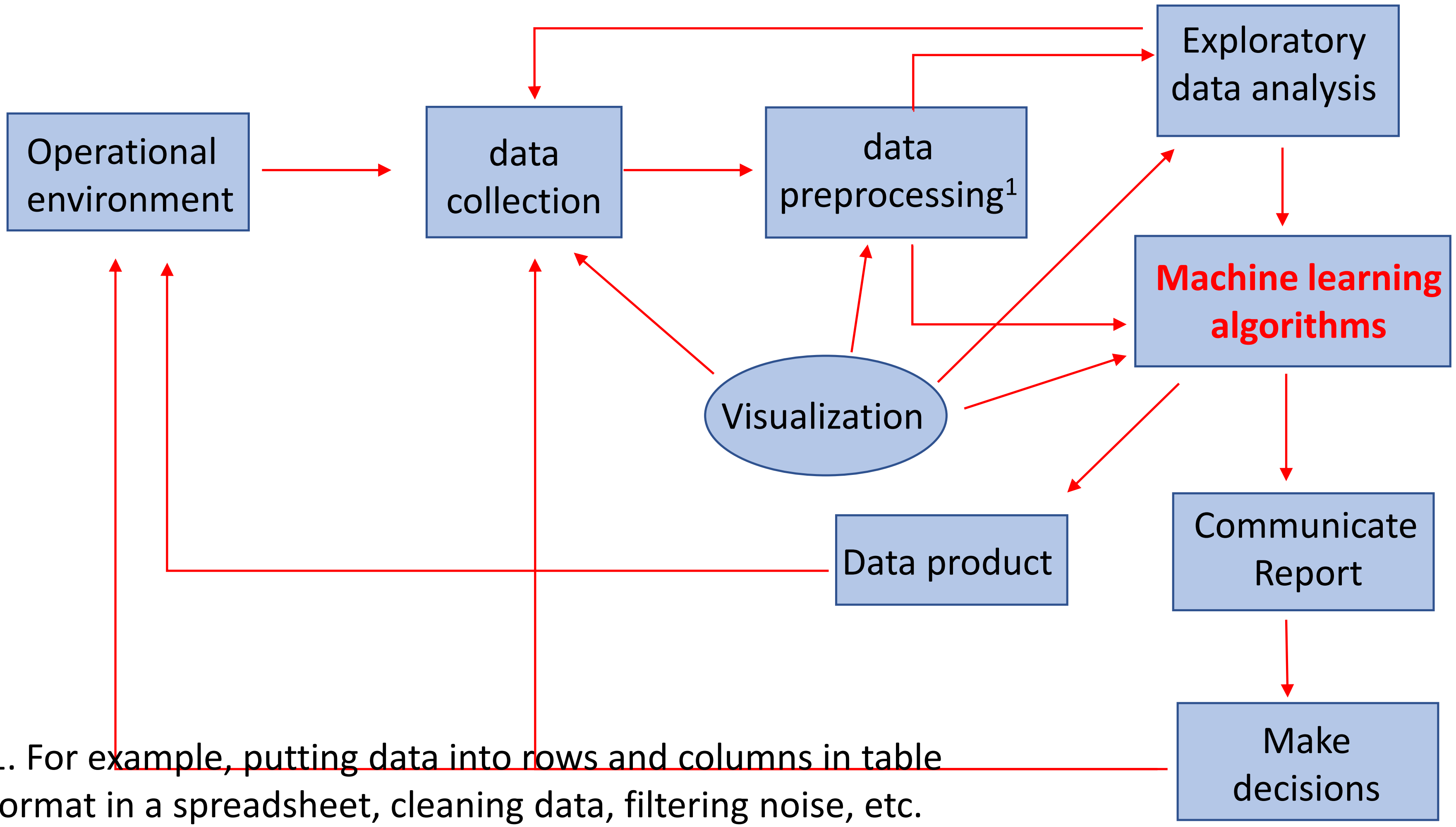
Modified after O'Neil & Schutt (2013, Doing data science)



1. For example, putting data into rows and columns in table format in a spreadsheet, cleaning data, filtering noise, etc.

Machine learning in the context of data analytics

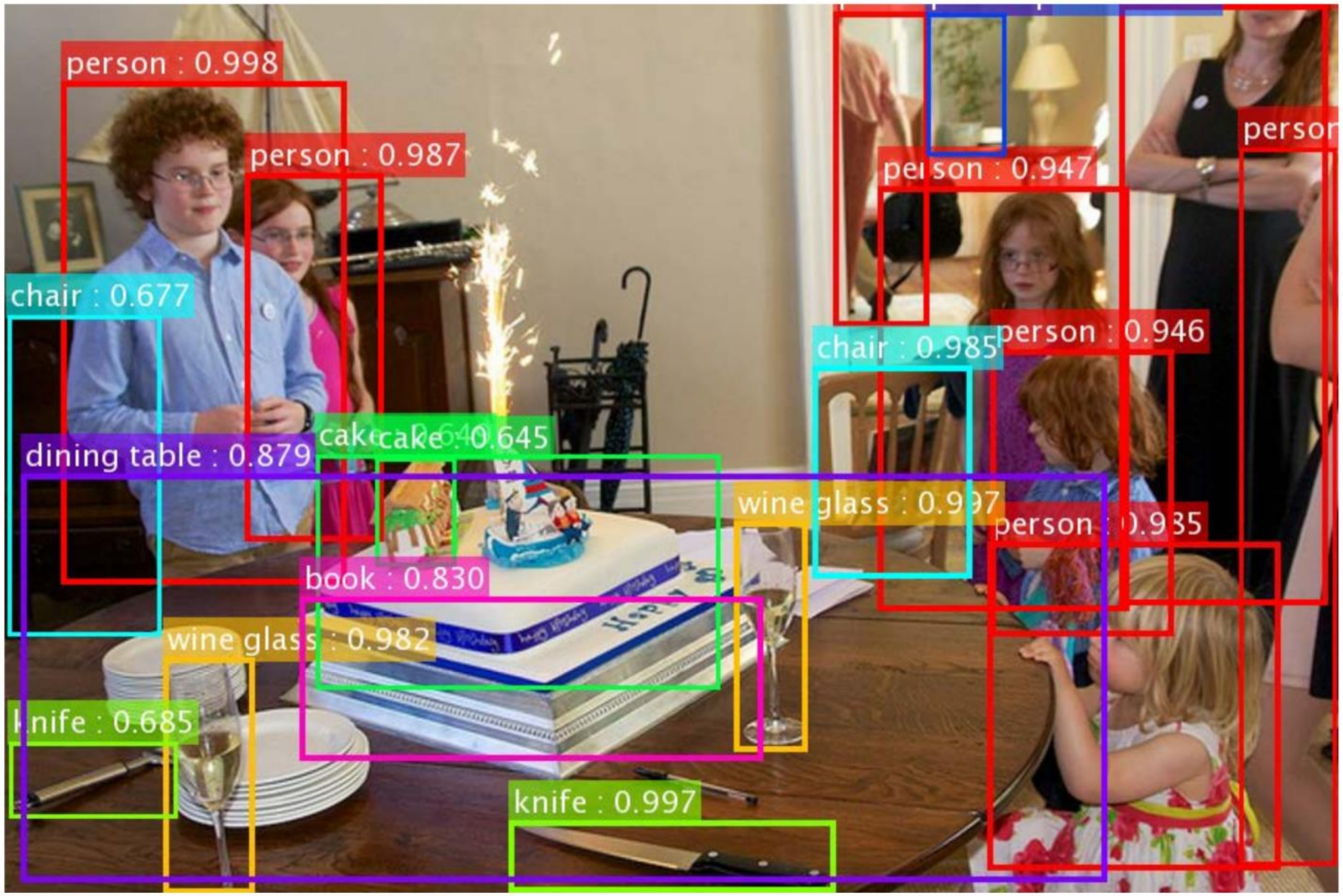
Modified after O'Neil & Schutt (2013, Doing data science)



1. For example, putting data into rows and columns in table format in a spreadsheet, cleaning data, filtering noise, etc.

Machine learning applications

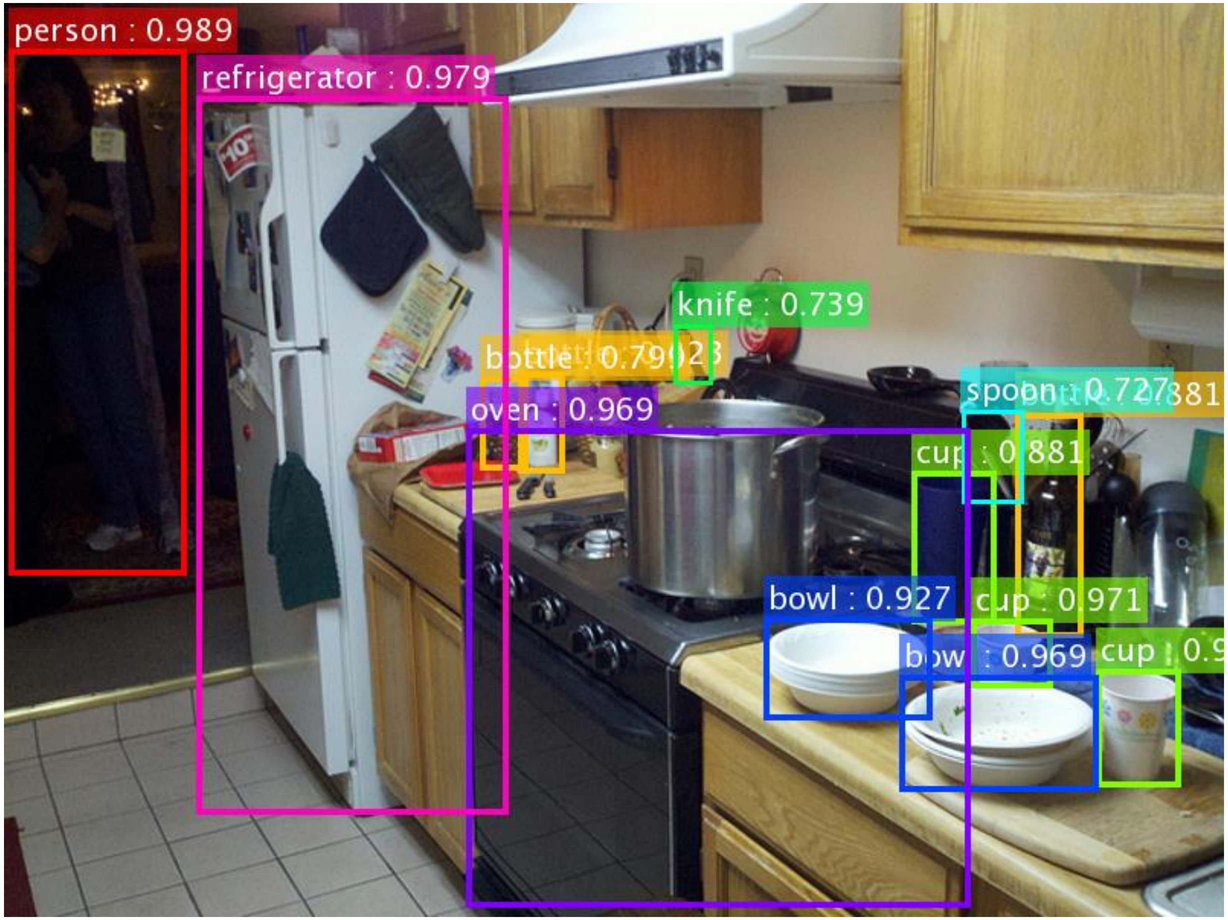
Object detection



ResNet applied to COCO dataset.

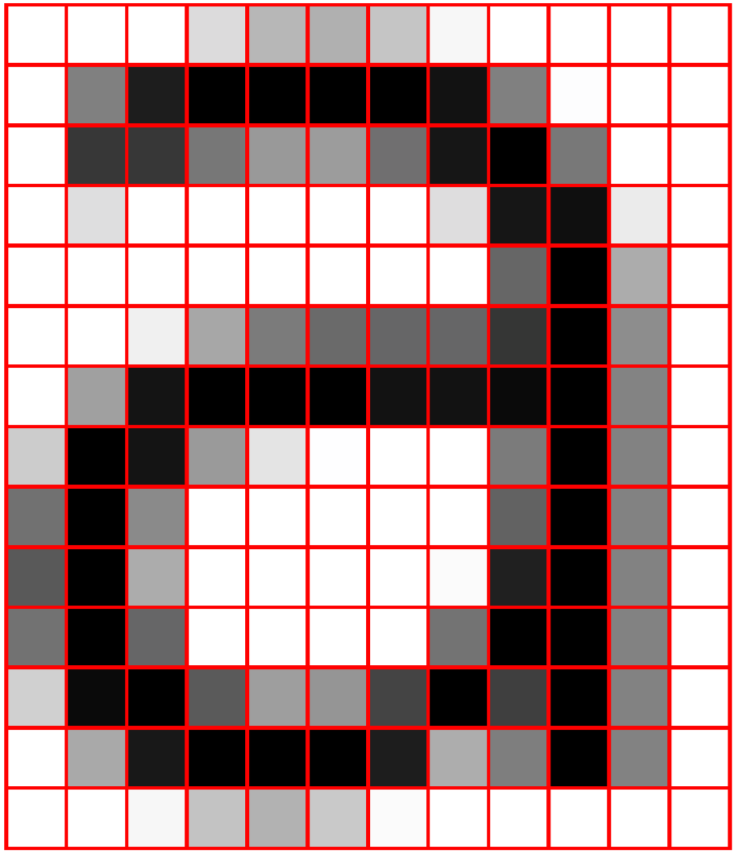
Source: He et al., Deep residual learning for image recognition, CVPR, 2016

Object detection



Source: He et al., Deep residual learning for image recognition, CVPR, 2016

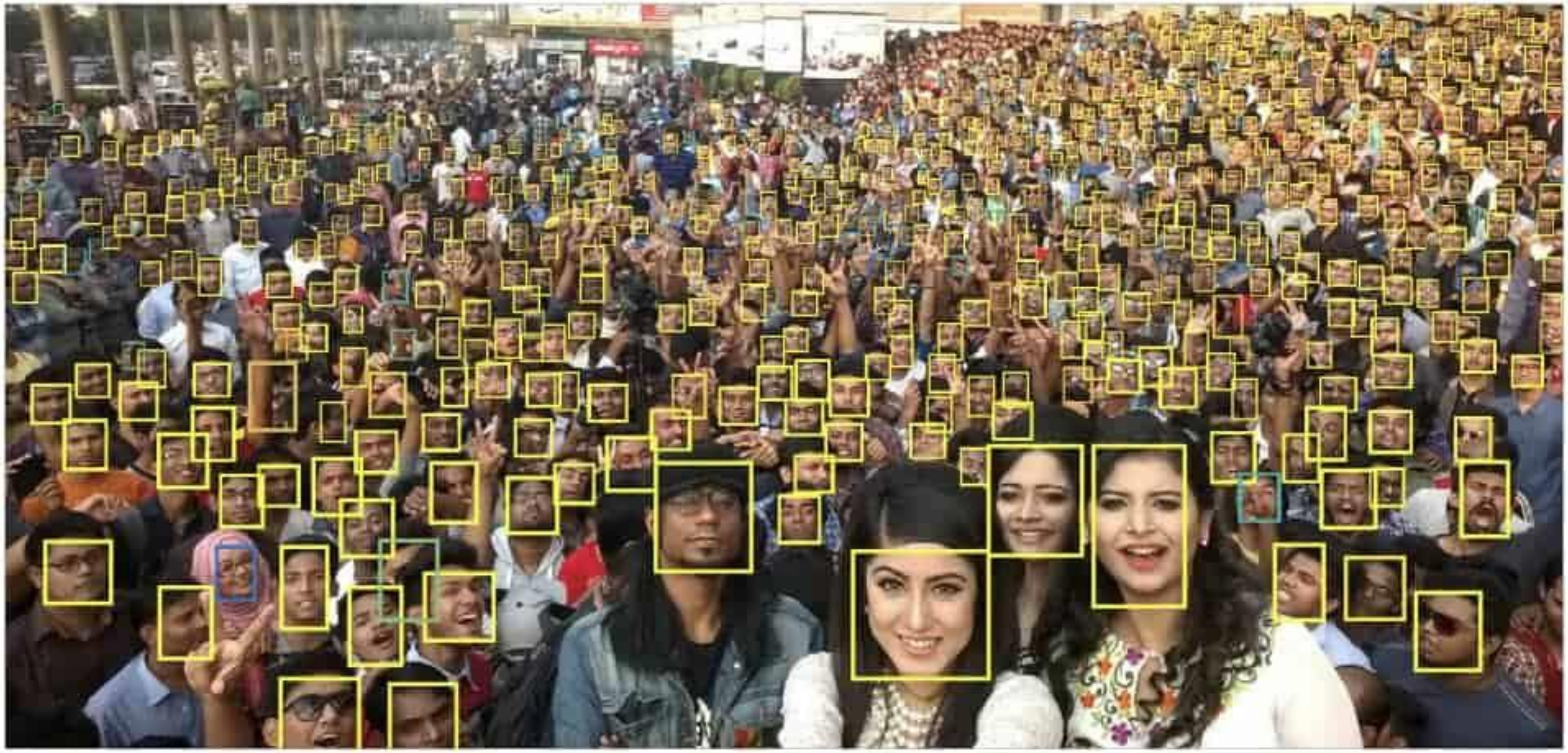
What we see vs. what computers see



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

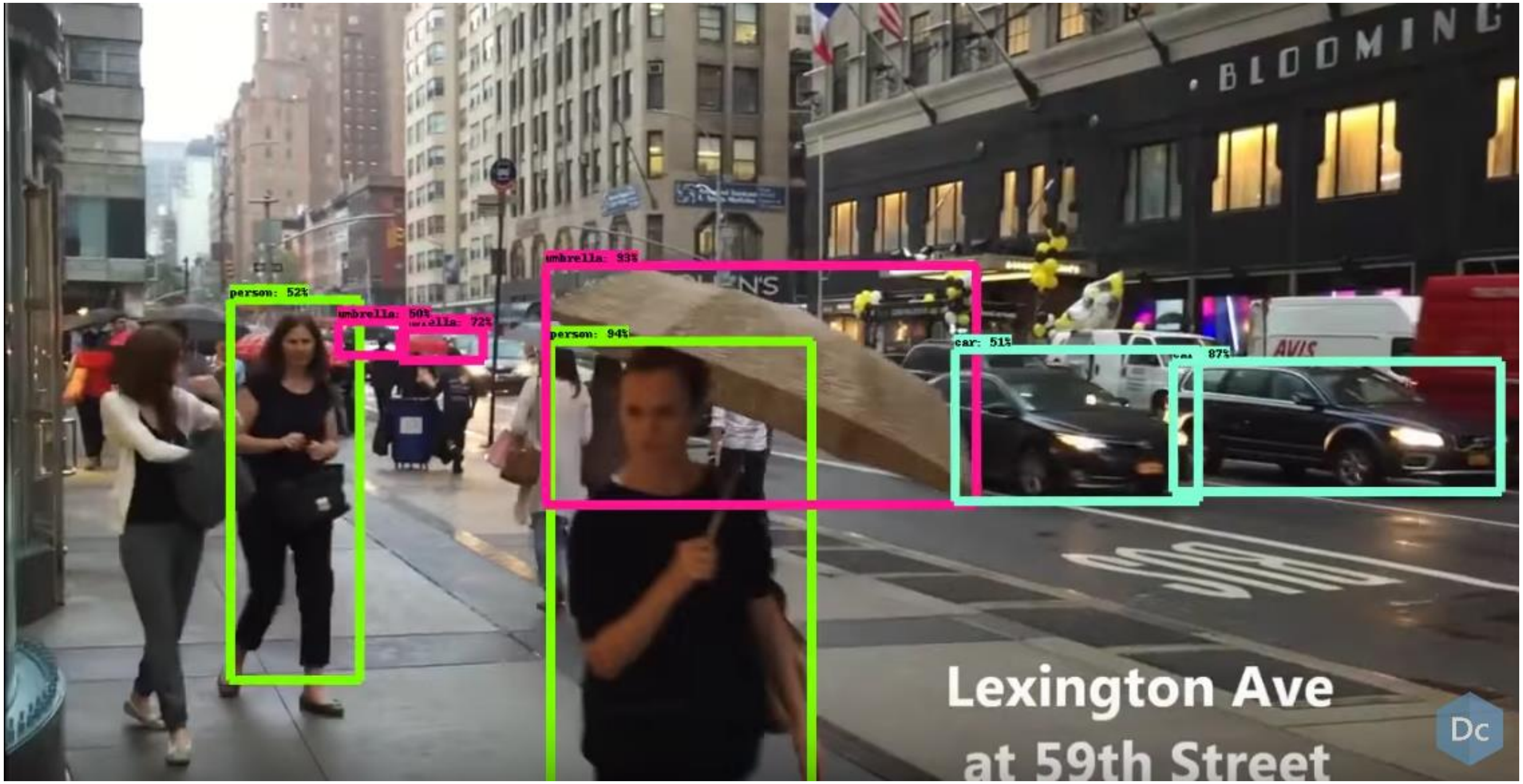
Image source: http://www.huawei.com/en/publications/winwin-magazine/AI/computer-vision-and-the-ai-boom

Detection of objects at different scales



Object detection with face as the only class. Note the existence of large and small faces poses a great challenge here. The authors explore the role of scale invariance, image resolution and contextual reasoning. Source: Hu and Ramanan (2016, https://arxiv.org/abs/1612.04402v1)

Real time object detection



Video online: https://www.youtube.com/watch?v=_zZe27JYi8Y

Real time object detection

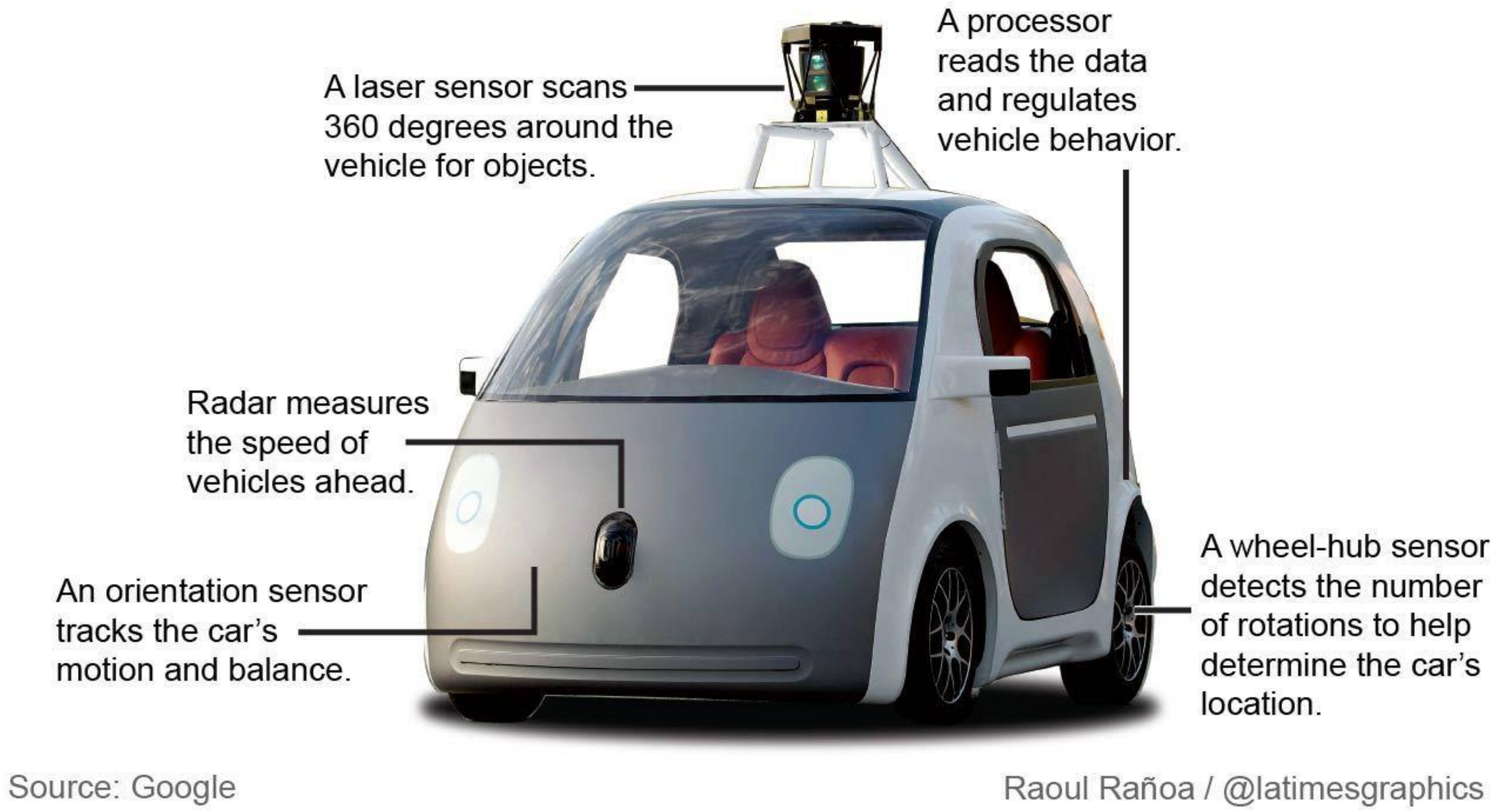
YOLO V2 achieves better results at very high FPS

Video online: https://www.youtube.com/watch?v=VOC3huqHrss&list=RDVOC3huqHrss

Hand written digit recognition

MNIST data set

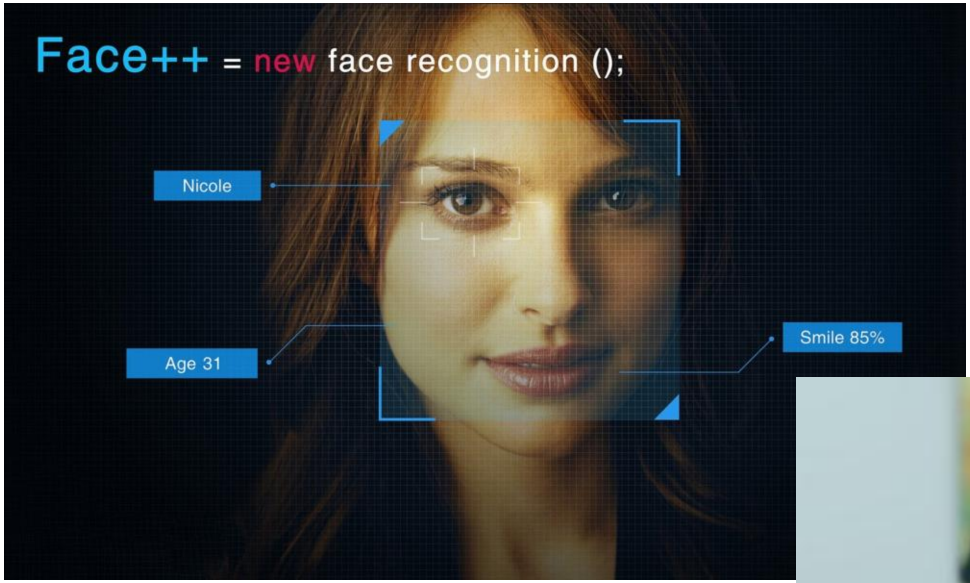
Self-driving car



Voice recognition

Source: https://biostore.co.uk/company/news-articles/voice-recognition-biometrics-making-noise/

Face recognition



Source: https://www.pinterest.com/pin/135600638759424941/?lp=true

Spam filter

Fraud detection

Source: http://tsigroup.com

https://news.google.com/?hl=en-US&gl=US&ceid=US:en

Google News

Search for topics, locations & sources

- Top stories
- For you
- Favorites
- Saved searches
- U.S.
- World
- Local
- Business
- Technology
- Entertainment
- Sports
- Science
- Health

Language & region
English | United States

Settings

Get the Android app

Get the iOS app

Send feedback

Help

Headlines

More Headlines

Theresa May's Brexit deal defeated by record margin: What happens now?

Fox News • one hour ago

- Monumental defeat for Brexit sparks chaos
  CNN • one hour ago
- British Prime Minister Theresa May suffers devastating defeat on key Brexit vote
  Fox News • one hour ago
- Brexit vote debacle puts Theresa May and the UK in a tough, but not catastrophic, corner
  Washington Examiner • 2 hours ago • Opinion
- Reject May's Brexit and Go Back to Voters
  Bloomberg • Today • Opinion

View full coverage

William Barr suggests Mueller report may not be made public — live updates

CBS News • 23 minutes ago

- William Barr Says 'Straight Shooter' Mueller Wouldn't Lead A 'Witch Hunt'
  HuffPost • 5 hours ago

View more

Democrats Jilt Trump on Lunch Talks but Look for Shutdown Exit

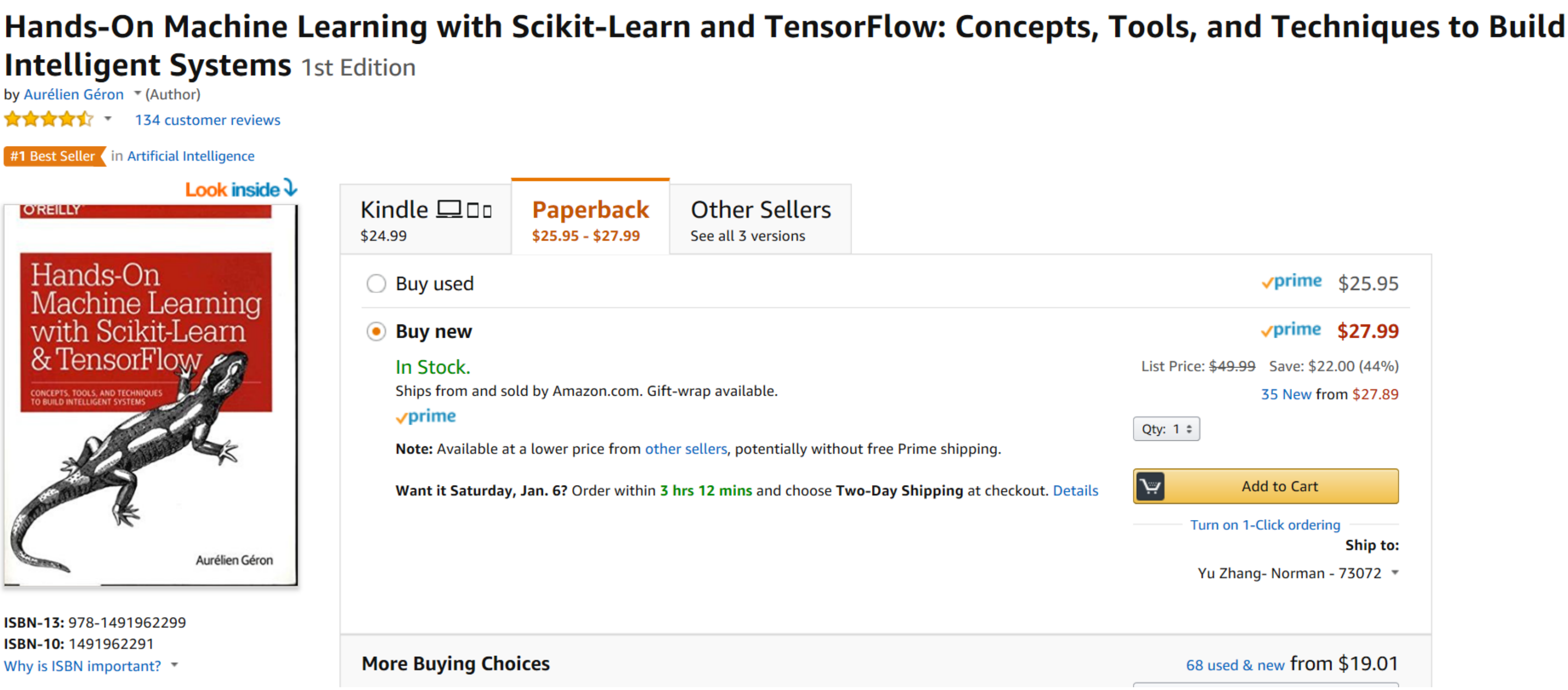
The New York Times • 3 hours ago

- Democrats boycott White House border security meeting
  Fox News • 5 hours ago

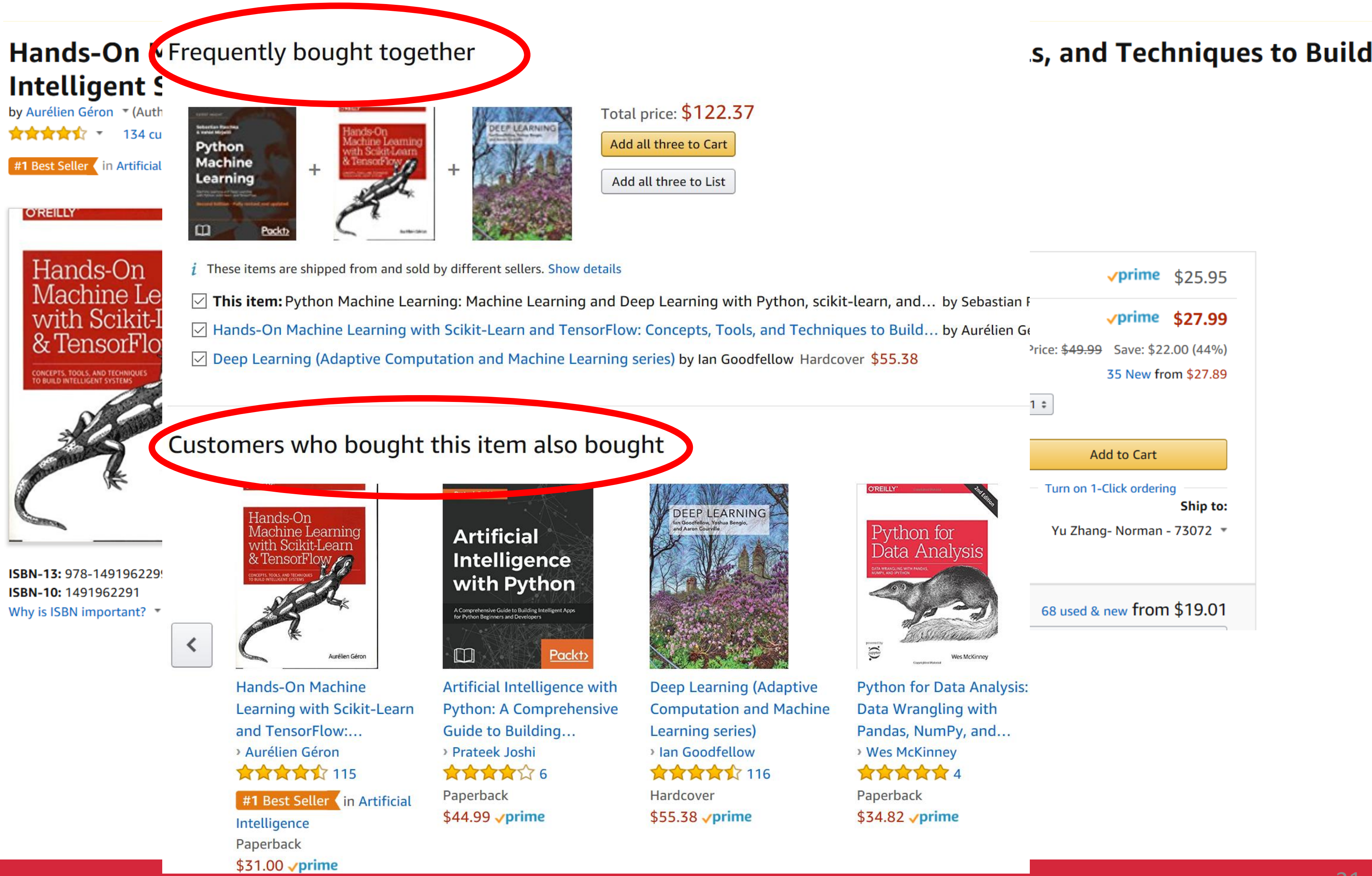
View more

Los Angeles teachers are on strike, exercising a right not enjoyed by most educators

Recommender system



Recommender system



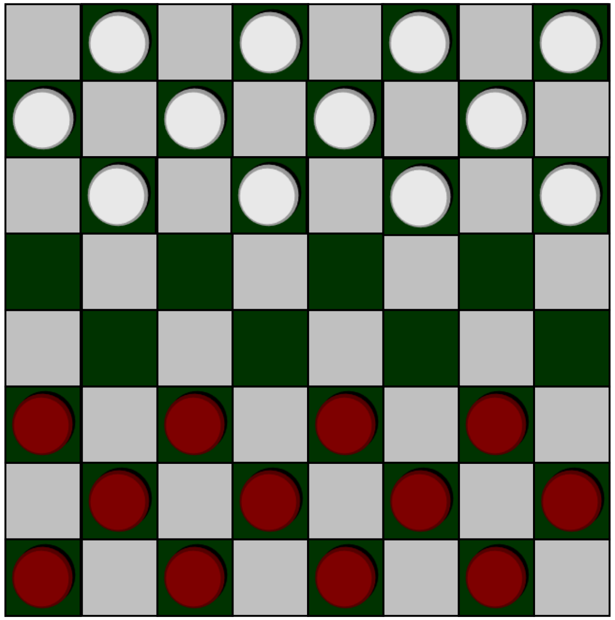
Go game

Image credit: Nature

Image credit: theverge.com

What is machine learning?

- Authur Samuel (1959): [Machine learning is the] field of study that gives computers the ability to learn without being explicitly programmed.



Source: lecture notes of Andrew Ng's Machine learning course on Coursera.com

What is machine learning?

- Tom Mitchell (1997): A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
- E.g. for spam filter,
  - ✓ T: flag emails
  - ✓ E: a lot of emails that were labeled by humans (training data)
  - ✓ P: the ratio of correctly classified emails

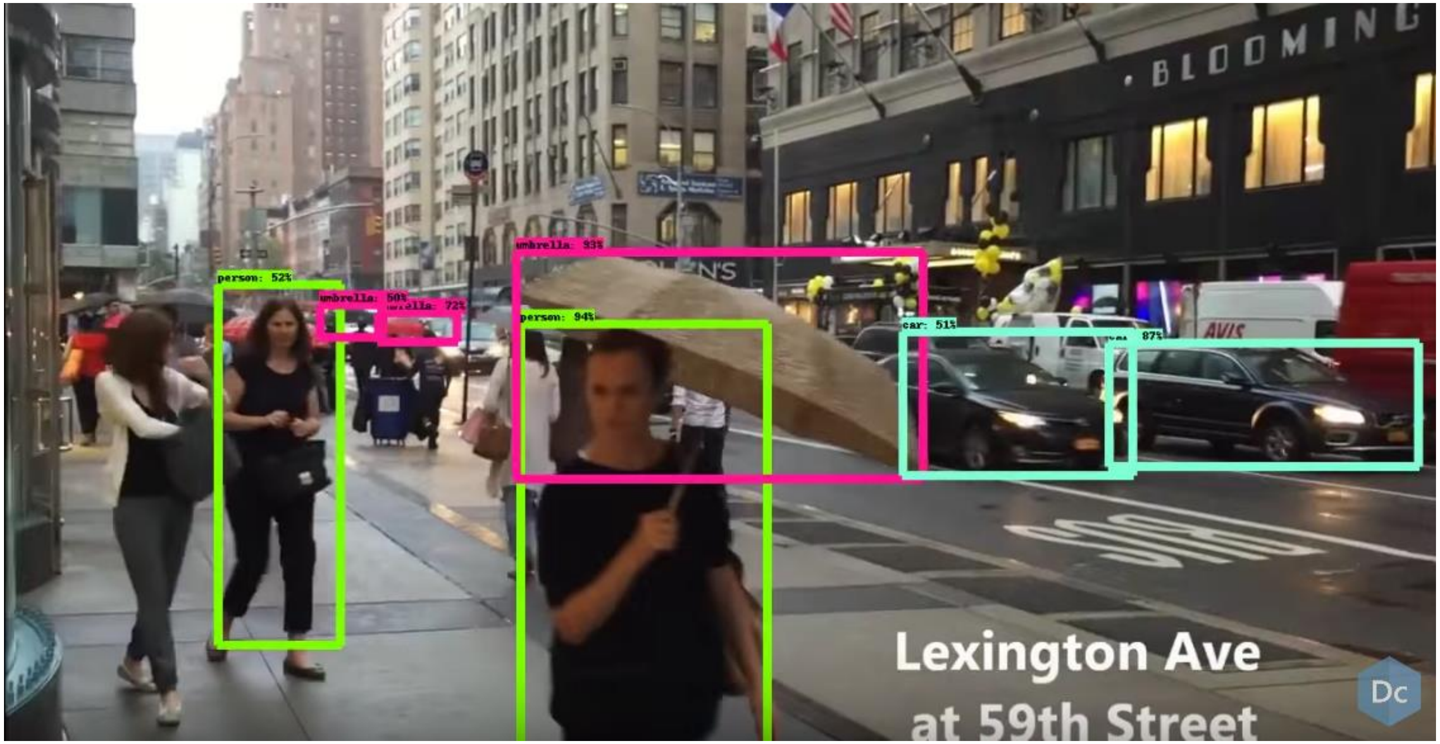
What is not machine learning?

- If you simply download all the codes in github.com, or a copy of Wikipedia, your computer has a lot more data, but it does not suddenly become better at any task .
- Thus, it is NOT machine learning.

Geron, 2017, Hands-on machine learning with Scikit-learn & TensorFlow

Exercise

- What is the T, P, and E for real time object detection?



https://www.youtube.com/watch?v=_zZe27JYi8Y www.cs.cmu.edu/~tom/

Tom Mitchell

E. Fredkin University Professor
Machine Learning Department
School of Computer Science
Carnegie Mellon University

Resume

Tom.Mitchell@cmu.edu, 412 268 2611, GHC 8203
Assistant: Mary Stech, 412 268-6869

What is Machine Learning, and where is it headed?

Video interview (5 min)

- AI, automation, and the future of work
  - Implications of Machine Learning for the workforce, *Science*, December 2017.
  - Governments need better data to track AI impact on jobs, *Nature*, April 2017.
  - 2017 U.S. National Academy report on Information Technology and the Future of Work
  - What Can Machines Learn and What Does It Mean for Occupations and the Economy?, *AEA Papers and Proceedings*, 2018.
- Machine Learning from Verbal User Instruction, video lecture on enabling cell phone users to teach their phones what to do, Simons Institute, Berkeley, February 13, 2017.
- Never Ending Language Learning, video lecture on our computer that is learning to read the web, Brown Univ., Feb. 2014.
- Neural representations of language meaning, video lecture on how the human brain represents word meanings, Berkeley, March 2014.
- When Computers Read, reprise of presentation at the World Economic Forum, Davos, Switzerland, January 2012 (5 minutes).

What is machine learning?

- Kevin Murphy (2012): The goal of machine learning is to develop methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data or other outcomes of interest.

What is machine learning?

- My definition: the field of study that gives computers the ability to learn from data (e.g., discovering patterns and relations among input data), and make predictions.

Why machine learning?

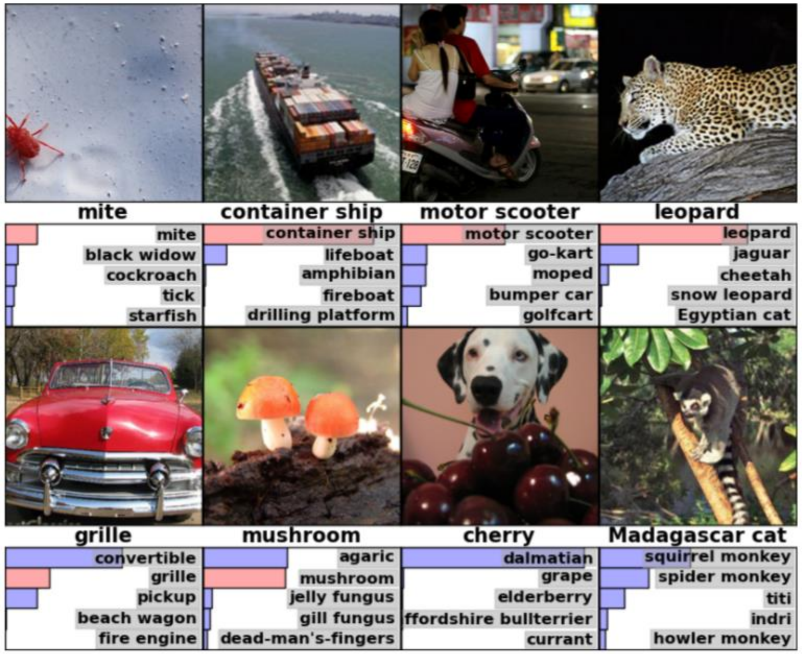
- Some problems cannot be (easily) solved by traditional approach
- For example,

Handwriting recognition

Speech recognition/Natural Language Processing (NLP)

Why machine learning?

- Some problems cannot be (easily) solved by traditional approach
- For example,



Computer vision

(Image credit: Krizhevsky et al, 2012, ImageNet classification)

Go game

Why machine learning?

- Some problems cannot be (easily) solved by traditional approach
- For example,

Autonomous helicopters
(Image credit: Andrew Ng)

"We just did not know how to write a computer program to make this helicopter fly by itself. The only thing that worked was having a computer learn by itself how to fly this helicopter. "
---- Andrew Ng

Why machine learning?

- Huge amounts of data generated every second
  - Web click data, medical records, etc.
- Data volume too large for traditional approach
  - Classification of seismic traces (one of the coding exercise)

Why machine learning?

- Huge amounts of data generated every second
  - Web click data, medical records, etc.
- Data volume too large for traditional approach
  - Classification of seismic traces (one of the coding exercise)
- Big data is one of the reasons why ML took off and became prevalent.

Why machine learning?

- Huge amounts of data generated every second
  - Web click data, medical records, etc.
- Data volume too large for traditional approach
  - Classification of seismic traces (one of the coding exercise)
- Big data is one of the reasons why ML took off and became prevalent.
- Machine learning can discover complex patterns in big data that are not immediately apparent to humans, and help humans gain insights into complex problems

Machine learning applied to geoscience

Facies classification

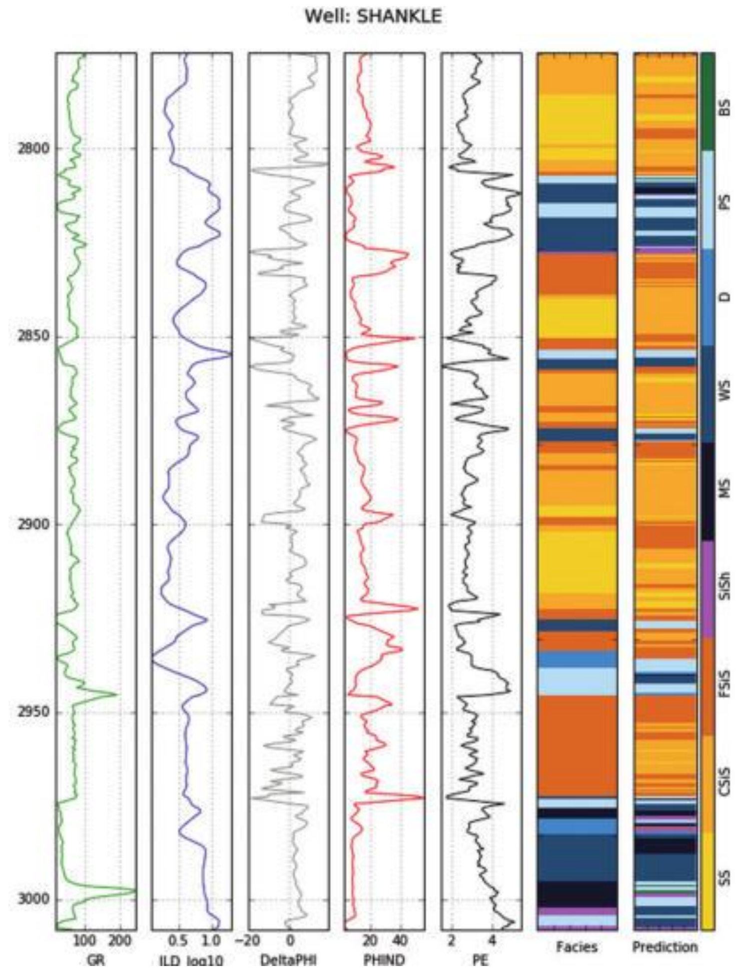
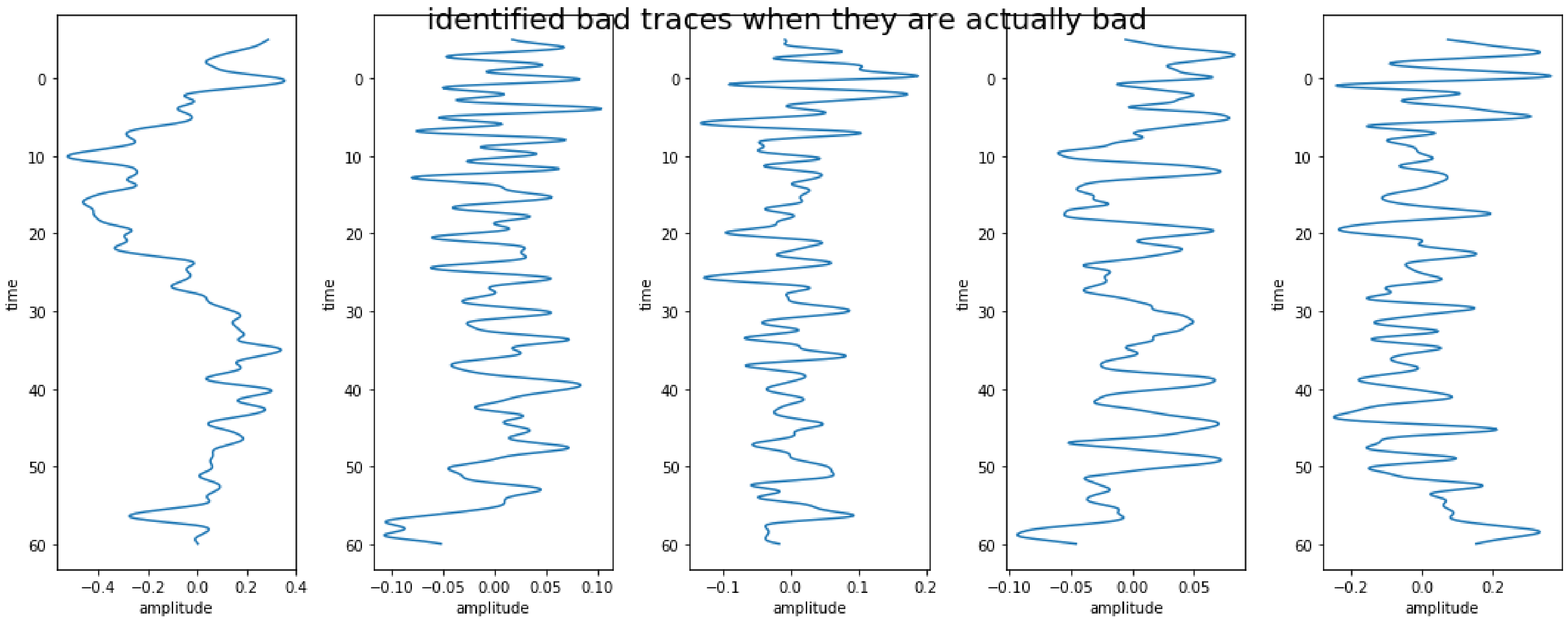


Image source: Hall, 2016, TLE

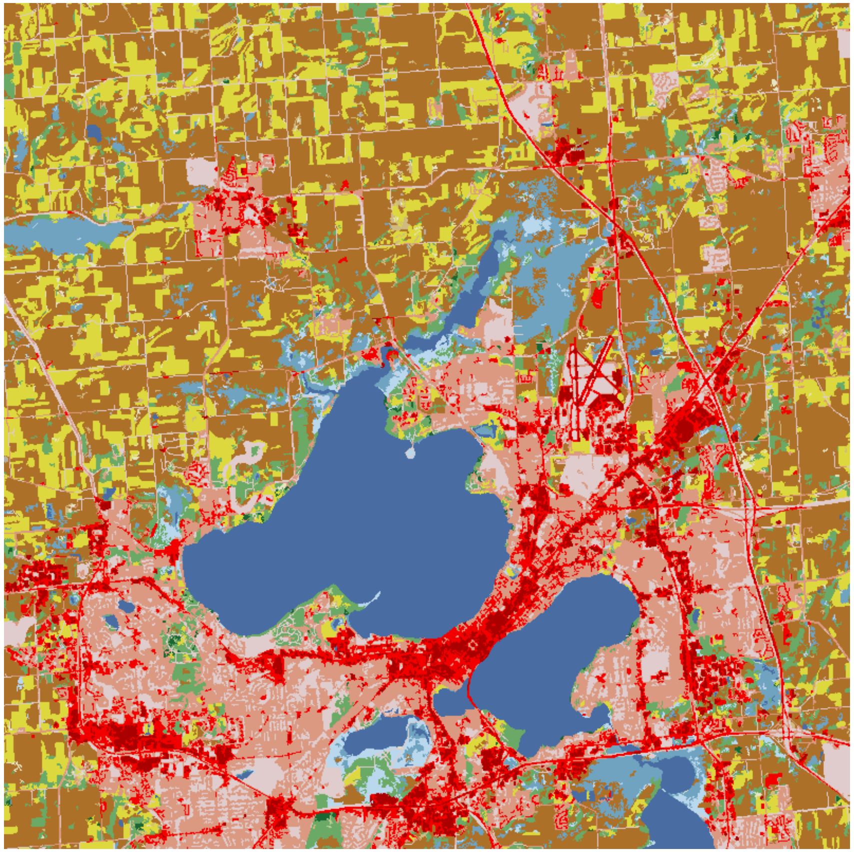
Machine learning applied to geoscience

Seismic traces classification



Machine learning applied to geoscience

Remote sensing image analysis

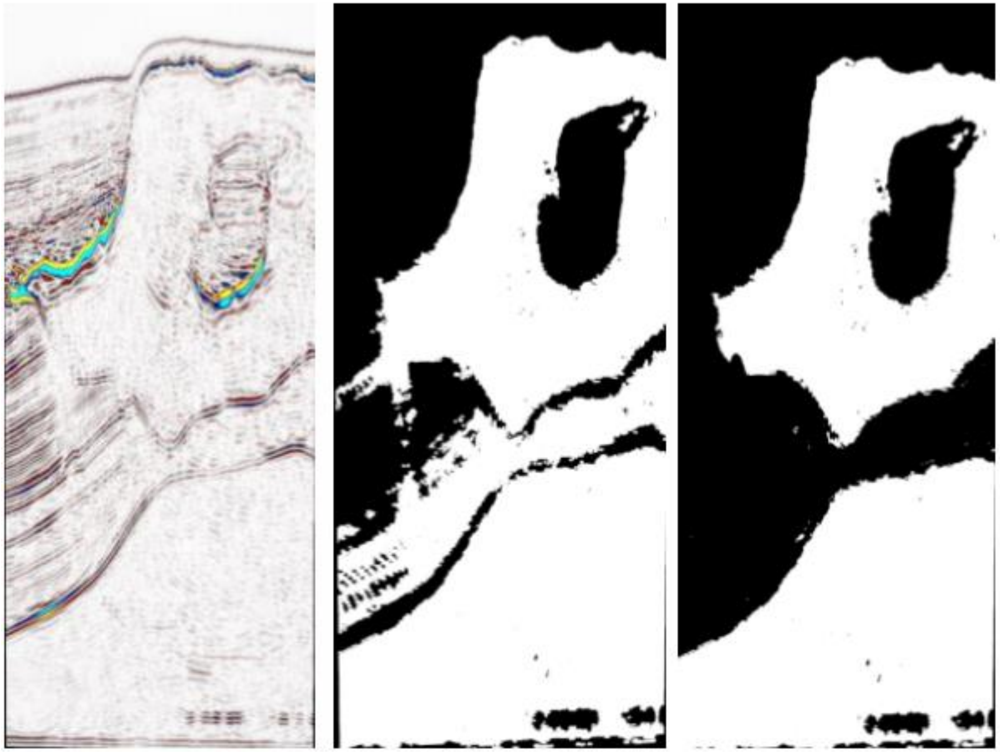


Land cover surrounding Madison, WI. Fields are colored yellow and brown, water is colored blue, and urban surfaces are colored red.

https://en.wikipedia.org/wiki/Land_cover

Machine learning applied to geoscience

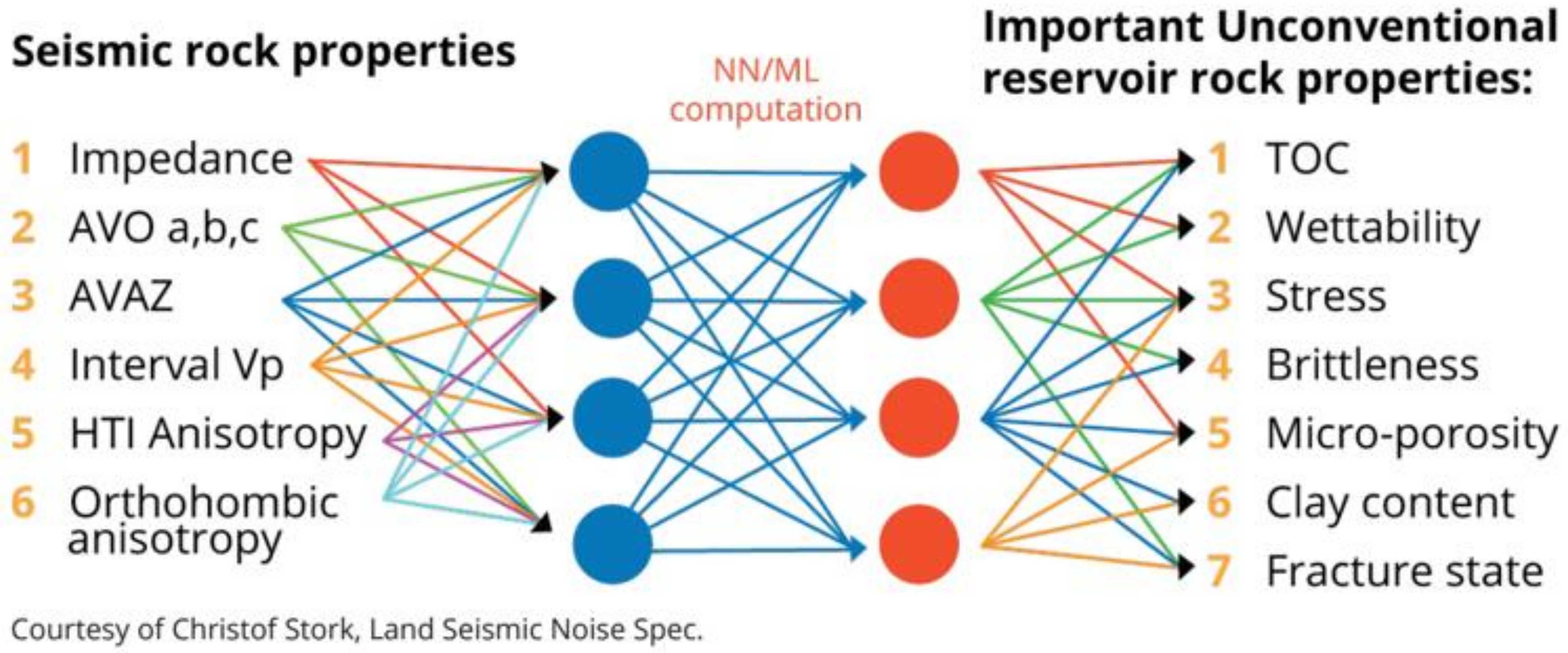
Salt detection



Seismic image　　ML detection　　Post-processing

Guillen et al., 2015

Machine learning applied to geoscience



Machine learning applied to geoscience

LETTER

https://doi.org/10.1038/s41586-018-0438-y

Deep learning of aftershock patterns following large earthquakes

Phoebe M. R. DeVries[1,2]*, Fernanda Viégas[3], Martin Wattenberg[3] & Brendan J. Meade[1]

Aftershocks are a response to changes in stress generated by large earthquakes and represent the most common observations of the triggering of earthquakes. The maximum magnitude of aftershocks and their temporal decay are well described by empirical laws (such as Bath's law[1] and Omori's law[2]), but explaining and forecasting the spatial distribution of aftershocks is more difficult. Coulomb failure stress change[3] is perhaps the most widely used criterion to explain the spatial distributions of aftershocks[4–8], but its applicability has been disputed[9–11]. Here we use a deep-learning approach to identify a static-stress-based criterion that forecasts aftershock locations without prior assumptions about fault orientation. We show that a neural network trained on more than 131,000 mainshock–aftershock pairs can predict the locations of aftershocks in an independent test dataset of more than 30,000 mainshock–aftershock pairs more accurately (area under curve of 0.849) than can classic Coulomb failure stress change (area under curve of 0.583). We find that the learned aftershock pattern is physically interpretable: the maximum change in shear stress, the von Mises yield criterion (a scaled version of the second invariant of the deviatoric stress-change tensor) and the sum of the absolute values of the independent components of the stress-change tensor each explain more than 98 per cent of the variance in the neural-network prediction. This machine-learning-driven insight provides improved forecasts of aftershock locations and identifies physical quantities that may control earthquake triggering during the most active part of the seismic cycle.

neuron may be interpreted as the predicted probability that a grid cell generates one or more aftershocks.

The stress changes and aftershock locations associated with about 75% of randomly selected distinct mainshocks were used as training data; the remaining 25% were reserved to test the trained neural networks. The training and testing datasets both consist of the elements of the stress-change tensor as features and the corresponding labels of either 0, for grid cells without aftershocks, or 1, for grid cells with aftershocks.

We assess the accuracy of the neural-network aftershock location forecasts on the test dataset using receiver operating characteristic (ROC) analysis. ROC curves are widely used to assess the efficacy of diagnostic medical tests. To build these curves, the true positive rate of a binary classifier is plotted against the false positive rate for all possible thresholds of the classifier (see Methods for more details). The area under an ROC curve (AUC) then quantifies the overall performance of a test across all thresholds (Fig. 1). The ROC analysis reveals that the neural-network forecast can explain aftershock locations better than can widely used metrics: the merged AUC value across all slip distributions and grid cells in the test dataset for the neural-network forecast is 0.849, which is larger than that of the classic Coulomb failure stress criterion[3] (AUC = 0.583) resolved on receiver planes parallel to the average orientation of the mainshock fault ($\Delta CFS(\mu = 0.4)$, in which μ is the effective coefficient of friction). Neither classifier has particularly high precision, defined as the percentage of grid cells predicted to be positive that actu-

What is this class about?

- Introduction to machine learning concepts, algorithms and tools for undergraduates who have not been exposed to this subject.

What is this class about?

- Introduction to machine learning concepts, algorithms and tools for undergraduates who have not been exposed to this subject.

- Prerequisites:
  - Willingness to learn
  - Perseverance
  - Calculus

Administrative basics

Instructor: Jiajia Sun

- Email: jsun20@uh.edu
- Office: SR1 127A
- Office phone: 713-743-7380
- Office hours: TuesThur 3:30-5:00 pm, or by appointments

Course materials/announcements

- Blackboard

Administrative basics

- Both lectures and labs
- Lectures: SEC 202
- Labs: SR1 230

Week	Date	Topics	Comments
1	01/15 Tues	Overview of syllabus Lecture: Introduction to Machine learning: applications	
	01/17 Thur	Lecture: Review of linear algebra	
2	01/22 Tues	Lab: Linear algebra in Python	Not graded
	01/24 Thur	Lecture: Introduction to optimization	
3	01/29 Tues	Lab: Gradient descent + Linear regression	Report due on 02/05 at 5:30 pm
	01/31 Thur	Lecture: Introduction to machine learning: concepts	
4	02/05 Tues	Lecture: Logistic regression	
	02/07 Thur	Lab: Logistic regression	Report due on 02/14 at 5:30 pm
5	02/12 Tues	Lecture: Support vector machine	
	02/14 Thur	Lab: Support vector machine	Report due on 02/21 at 5:30 pm
6	02/19 Tues	Lecture: Decision trees	
	02/21 Thur	Lab: Decision trees	Report due on 02/28 at 5:30 pm
7	02/26 Tues	Lecture: Random Forest	
	02/28 Thur	Lab: Random forest	Report due on 03/07 at 5:30 pm
8	03/05 Tues	Lecture: Ensemble learning	
	03/07 Thur	Lab: Ensemble learning	Reprot due on 03/19 at 5:30 pm
9	03/12 Tues	No class due to spring break	
	03/14 Thur	No class due to spring break	
10	03/19 Tues	Review & Recap	
	03/21 Thur	Exam	
11	03/26 Tues	Lecture: Clustering	
	03/28 Thur	Lab: Clustering	Report due on 04/04 at 5:30 pm
12	04/02 Tues	Lecture: Introduction to TensorFlow	
	04/04 Thur	Lab: TensorFlow	Not graded
13	04/09 Tues	Lecture: Introduction to neural networks 1	
	04/11 Thur	Lecture: Introduction to neural networks 2	
14	04/16 Tues	Lab: Deep learning	Report due on 04/23 at 5:30pm
	04/18 Thur	Lecture: Convolutional neural networks 1	
15	04/23 Tues	Guest lecture: Convolutional neural networks 2	
	04/25 Thur	Lab: CNN (optional)	Report due on 05/02 at 5:30 pm
16	04/30 Tues	final presentation??	
	05/02 Thur	final presentation??	
Note	28 class meetings		04/29 last day of class

Course contents

Week 1-3: foundation

- Introduction
- basic Python programming
- review of linear algebra
- intro to optimization

Week 4-8: Supervised learning

- Logistic regression
- Support vector machine
- Decision trees
- Ensemble learning

Week 10: review & exam

Week 11: Unsupervised learning

Week 12-15: Deep learning

Week 16: final

Lab exercises

- 10 in total
- two ungraded
- Programmed in Python in Jupyter Notebook

Skills you can add to your CV

- Programming in Python
- Jupyter Notebook
- Implementing ML using Scikit-Learn library
- Implementing deep learning using TensorFlow
- Keras
- Cloud computing
- Machine Learning
- Deep learning

Grading policy

- Random in-class quizzes: 10%
  - 5 in total, one for 2%
- Exam: 20%
- Lab exercises + report: 60%
  - two ungraded
  - Remaining 8 graded based on coding and writing
- Late policy
  - 2% off per hour.

- Collaboration policy
  - Read student code book, understand ‘collaboration’ vs ‘infraction’
  - Use your judgement

More on collaboration

- Feel free to discuss the lab exercises
- But coding and reports must be done individually
- Acknowledge the help you get from your peer students and Internet.

Missed quizzes and make up work

- If you miss a quiz due to unavoidable circumstances (e.g., health, car accidents), inform the instructor as early as possible, and be prepared to provide relevant records (e.g., a note from doctor, policy report)
- No make-up exam except for rare justifiable circumstances.

First in-class quiz (time permitting)

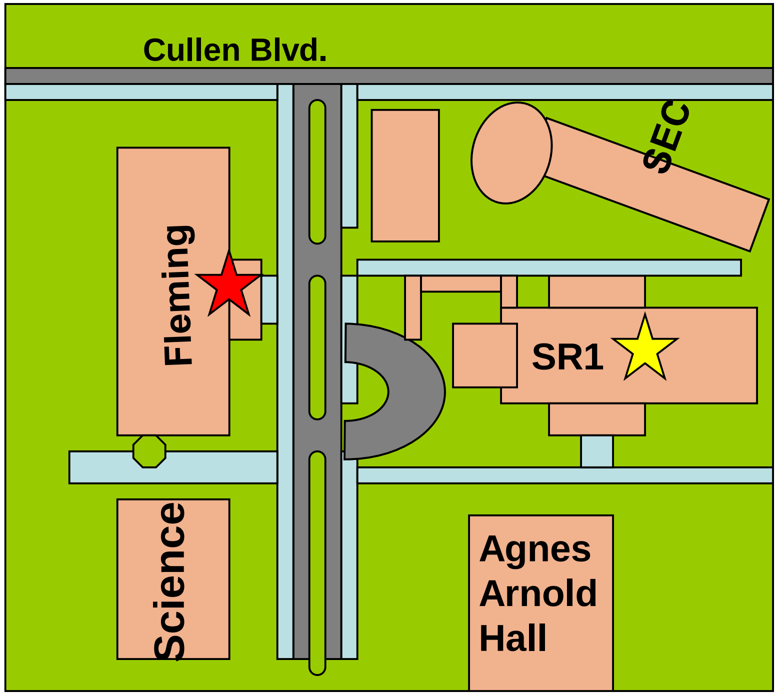
- What do you want to do after graduation?

- What do you expect to learn from this class?

Teaching Assistant

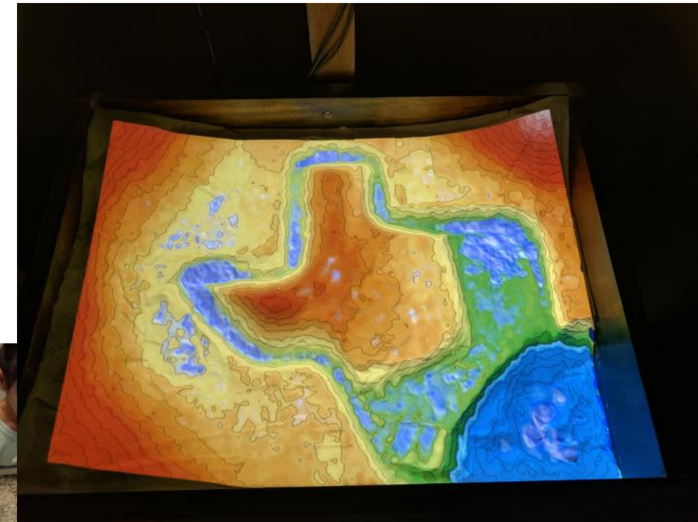
- Xin Zhou
  - xzhou28@central.uh.edu
  - Tues: 2:00 – 4:00 pm at GLC
    Thur: 3:00 – 5:00 pm at GLC

The Geoscience Learning Center (GLC)



GLC

My Office



Fleming 136

M-Th 8:00am-7:00pm

F 8:00am-6:00pm

Staffed by EAS teaching assistants

Coordinators

Dr. Hauptvogel
dwhauptv@central.uh.edu

Dr. Sisson
vbsisson@central.uh.edu

geolearn@nsm.uh.edu

713-893-1420

http://www.geosc.uh.edu/undergraduate/learning-center/index.php

One-on-one tutoring - Rocks - Minerals - Textbooks - Augmented Reality Sand Box - River Modeling Table - Microscopes & Thin Sections - Geophysics Workstation - Computers - Study Space - Google Earth - Videos