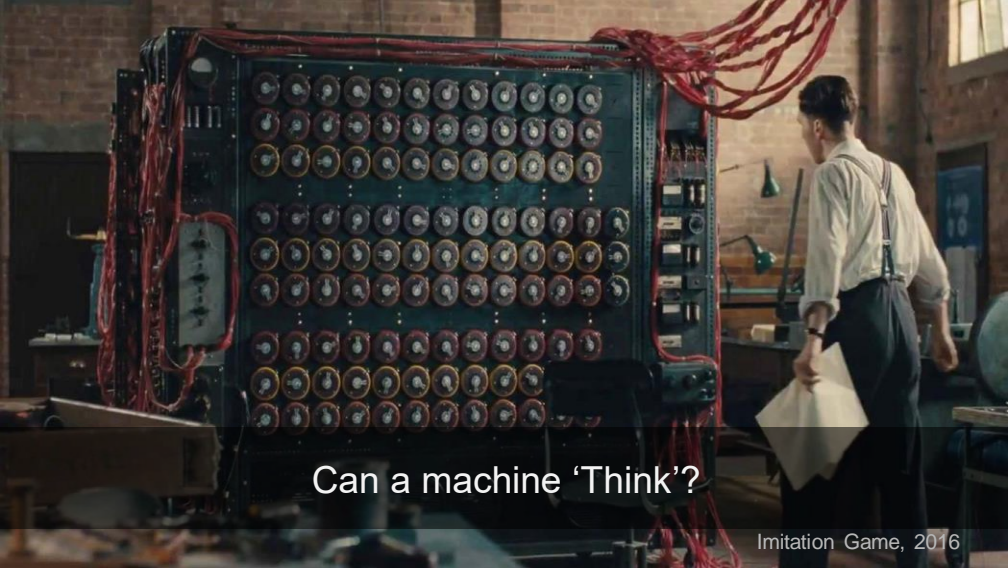


Introduction to Machine Learning



Imitation Game, 2016



Can a machine 'Think'?

Imitation Game, 2016

Alan Turing

"Can machines think?"... The new form of the problem can be described in terms of a game which we call the 'imitation game.'

What is Machine Learning?

Machine Learning Historical Background

Machine Learning Historical Background



From Quest of AI Book

Main problem..

- Classic AI: Symbolic Reasoning
 - No learning
 - Poor handling of uncertainty
 - Hard coding

Machine Learning Historical Background

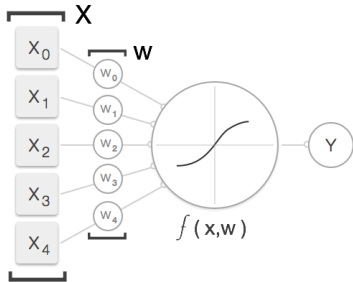
- Born from the ambitious goal of Artificial Intelligence



Dartmouth AI Conferences

Machine Learning Historical Background

- Perceptron: first artificial neuron.



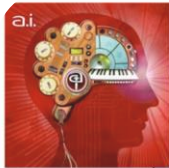
Rosenblatt, source: Wikipedia

Machine Learning:

- Learning
- Poor handling of uncertainty
- Hard coding

Machine Learning Historical Background

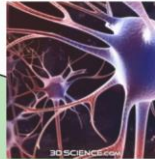
Artificial Intelligence 1960s



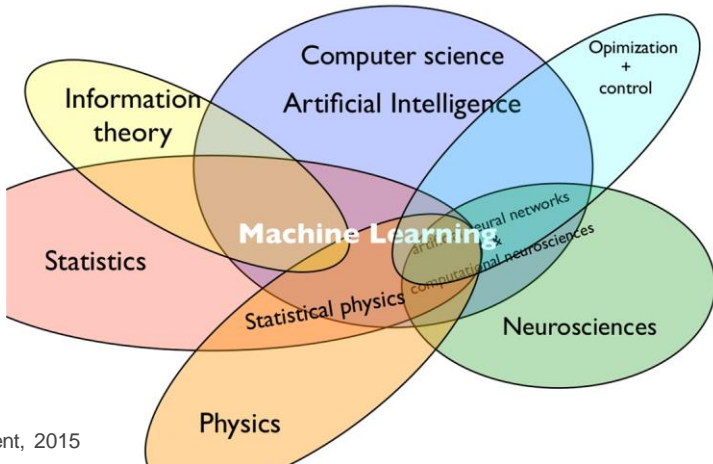
Computer science
Artificial Intelligence
Largely symbolic AI

artificial neural networks

Neurosciences



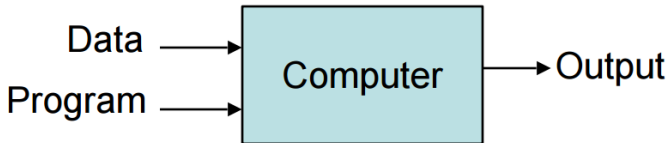
Machine Learning Current View



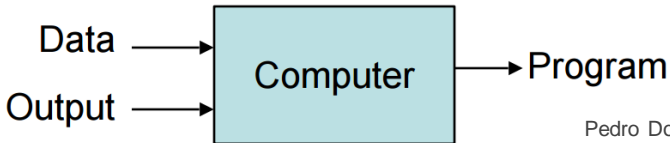
What is Machine Learning?

- “Field of study that gives computers the ability to learn without being explicitly programmed”
 - Arthur Samuel (1959)

Traditional Programming



Machine Learning



Stanford Autonomous Helicopter

Andrew Ng,
Autonomous
Helicopter



Classification

Spam Filtering

+Alex Search Images Maps Play YouTube News Gmail Drive Calendar More -

Google

ham

Alex Smola 0 + Share

1-50 of 15,803

COMPOSE

Inbox (7,180)
Important
Sent Mail
Drafts (61)

<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Southwest Airlines	Your trip is around the corner! - You're all set for your San Jose trip! My Account View My Itinerary Online	2:12 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	DiscountMags.com	\$3.99 Business & Finance Sale... starts now! - Trouble Seeing This Email? View as Webpage STOP these e-r	12:03 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	support, Alex (3)	Your order has shipped... - please send to the address below for an exchange remoteresremotes.com/exchange	7:22 am
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	American Airlines AAdvant.	AAdvantage eSummary - January 2013 - VIEW IN WEB BROWSER >> http://americanairlines.ed10.net/r/JC	1:17 am
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Taesup, Alex, Taesup (3)	Happy new year! - Hi Alex, Thanks for your condolence. I will arrive at Berkeley on 18th (wed) night. So, I car	Jan 11

+Alex Search Images Maps Play YouTube News Gmail Drive Calendar More -

Google

in:spam

spam

Alex Smola 0 + Share

1-50 of 244

COMPOSE

Inbox (7,180)
Important
Sent Mail
Drafts (61)
All Mail

→ Circles

→ [Gmail]
Done (1,006)
[imap]/Drafts
[imap]/Sent
alex.smola@yahoo...

Search people...

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	maee	(Ei&ISTP Index)2013机械与自动化工程国际会议征文: [alex.smola@gmail.com] - 尊敬的老师, 您好: 机械与	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Dear Valued Customers,	Low Interest Rate Loan - Dear Valued Customers, Do you need a loan or funding for any of the following reas-	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	garjeti	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Steven Cooke	Congratulations Alex, \$150 awaits you - Alex: IMPORTANT - NOTICE OF WINNINGS Please make sure yo	Jan 11
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	paper18	【2013-1-15截稿】【2013年机电与控制工程亚太地区学术研讨会APCMCE 2013】[Ei]【香港】【不参-不要	Jan 10
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	First-Class Mail Service	Tracking ID (G)BGD35 849 603 4893 4550 - Fed Ex Order: JN-3339-28981768 Order Date: Thursday, 3 Janua	Jan 10
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	garjeti	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 10
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Candy.Li	中屈,不担当老板的代言人	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Ronan Morgan	Ronan Morgan just sent you a personal message. - LinkedIn Ronan Morgan just sent you a private messag	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	RE/MAX®	2013 Valuable Offer! - Hello Friend, RE/MAX® has issued 2013 valuable property offer in your resident from	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	newsletter	newsletter WWW2013 - Newsletter 6 - See the Portuguese and Spanish version right after the English versio	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	CJCR editor	Chinese Journal of Cancer Research (CJCR) has been indexed by Pubmed and PMC - Click here if this e-mail	Jan 9
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	garjeti (2)	Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOG	Jan 9

Alex Smola,
Introduction ML

Product Recommendation: Imputing Missing Data

Collaborative Filtering

Recently Watched



Top 10 for Alexander



Don't mix preferences on Netflix!

Customers Who Bought This Item Also Bought

Alex Smola,
Introduction ML



Convex Optimization by
Stephen Boyd
★★★★☆ (11)
\$65.78



Point Processes
(Chapman & Hall / CRC
Monographs on S... by
D.R. Cox
\$125.47



Probabilistic Graphical
Models: Principles and
T... by Daphne Koller
★★★★★ (5)
\$71.52

Amazon
books

Netflix Prize

The image shows a screenshot of the Netflix Prize website. At the top, the Netflix logo is visible. Below it, the text "Netflix Prize" is prominently displayed in a large, bold font. To the right of this text, a large, stylized "COMPLETED" stamp is overlaid. Below the main title, there is a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". The background of the page features a dark, abstract design with silhouettes of people and various text elements, including "Movies For You" and "You really liked it...". On the right side of the page, a white box contains a "Congratulations!" message in blue text, followed by a paragraph explaining the prize and a link to the "Leaderboard".

COMPLETED

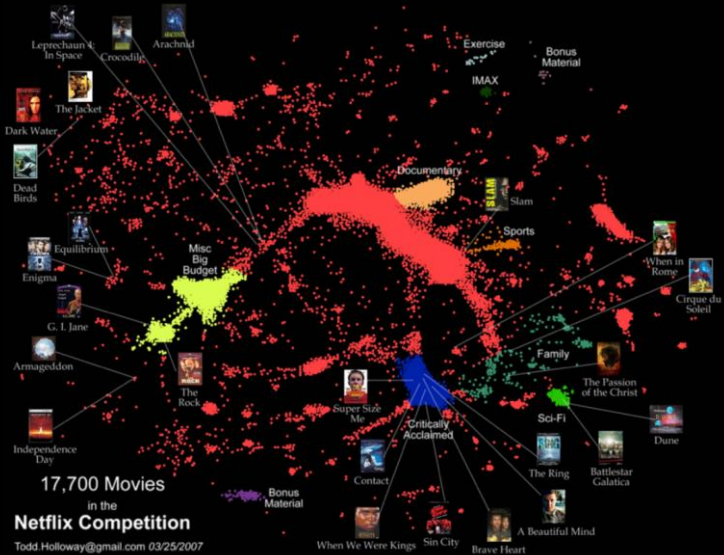
Home Rules Leaderboard Update

Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about [their algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

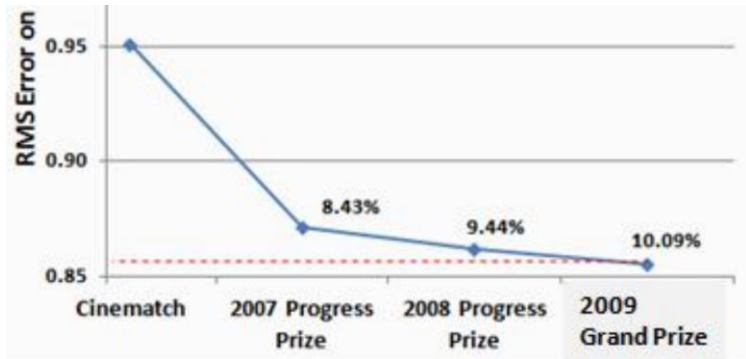
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.



17,700 Movies
in the
Netflix Competition

Todd.Holloway@gmail.com 03/25/2007

Netflix Error time by time



Time Series Prediction

Prediction



Alex Smola,
Introduction ML

tomorrow's stock price

Carnegie Mellon University



A new kind of hedge fund built by a network of data scientists.

[Learn more](#)

NEW DATASET IN 3D 17H 14M 24S

37,329,090,110 PRICE PREDICTIONS

ANNUAL RATE

CAREER EARNINGS

LOGLOSS META MODEL RANK

\$54,000.00	DEPRIVING	\$27.80	0.585	1
\$22,704.00	FUNGIBLE	\$69.87	0.592	2
\$13,668.00	QUIPIKA	\$0.00	0.677	3
\$9,540.00	ALOMOMOLA	\$34.14	0.550	4
\$7,212.00	INCANDESCING	\$12.51	0.518	5
\$5,748.00	VZIKK	\$0.00	0.676	6
\$4,740.00	TUNELITY2	\$0.00	0.679	7
\$4,008.00	BASSET	\$5.28	0.675	8
\$3,456.00	IDLING	\$0.00	0.673	9
\$3,036.00	BIDDOOF	\$1.11	0.546	10
\$2,688.00	SWEETCHIC	\$21.26	0.683	11
\$2,412.00	VINTY	\$0.00	0.677	12
\$2,184.00	BARBARACLE	\$0.89	0.667	13
\$1,992.00	KORM3	\$5.44	0.675	14
\$1,824.00	PLAIDPANDA	\$1.24	0.679	15
\$1,680.00	MUFASA3	\$0.00	0.673	16
\$1,560.00	TEDIUM	\$13.45	0.676	17
\$1,452.00	TEACH	\$177.35	0.683	18
\$1,356.00	AZUMARILL	\$2.49	0.618	19
\$1,272.00	ZANAME	\$0.00	0.674	20
\$1,200.00	NEKUS	\$0.00	0.679	21

Assembling a Super Intelligence

Numerai is not a search for the 'best' model; it is a platform to synthesize many different, uncorrelated models with many different characteristics.

Data scientists compete on [the leaderboard](#) but models are ranked and rewarded based on their contribution to the meta model.

Learn more in *Super Intelligence for the Stock Market*



Howard Morgan

Co-Founder of Renaissance Technologies

KIRAX: 0.65013

CAMBRIJS: 0.85561

Imitating Behavior

Imitation Learning

Alex Smola,
Introduction ML



Drivatar in Forza

FORZA MOTORSPORT | 4

Carnegie Mellon University

Imitation Learning in Games



Avatar learns from
your behavior

Alex Smola,
Introduction M

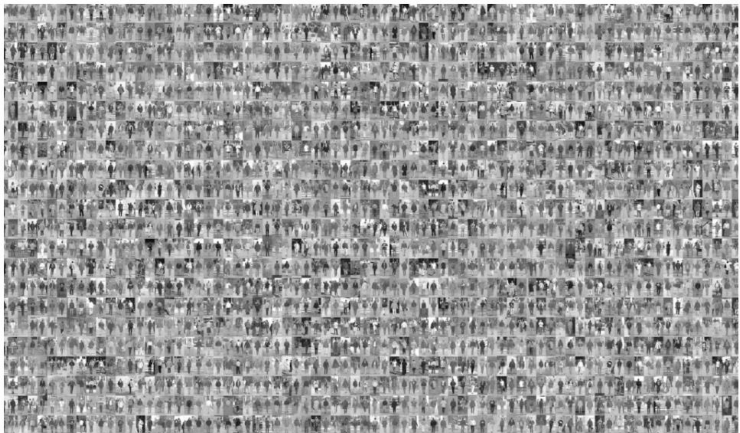
Black & White
Lionsgate Studios

“Hassabis worked as lead AI programmer on the iconic god game [Black &](#)



Machine Learning in Industry

Nando de Freitas,
Introduction ML



Millions of labeled examples are used to build real-world applications, such as pedestrian detection

[Tomas Serre]



Nando de Freitas,
Introduction ML



[Thomas Serre 2012]



Hot Research: Driverless Car



A Tesla driver was caught sleeping on the highway with his car on Autopilot



Dave Smith



May 24, 2016, 11:22 AM

34,432

2



FACEBOOK



LINKEDIN



TWITTER



EMAIL



PRINT



An Open Source Self-Driving Car

Udacity is building an open source self-driving car, and we want your help! Join the effort to create the world's first open source autonomous vehicle. We've broken down the problem into multiple complex challenges, and you or a team can compete to have your solution run in a real self-driving car.

[LEARN MORE](#)[JOIN SLACK](#)[GITHUB](#)

CHALLENGE 1

3D Model for Camera Mount

[VIEW CHALLENGE DETAILS >](#)

CHALLENGE 2

Using Deep Learning to Predict Steering Angles

[VIEW CHALLENGE DETAILS >](#)

CHALLENGE 3

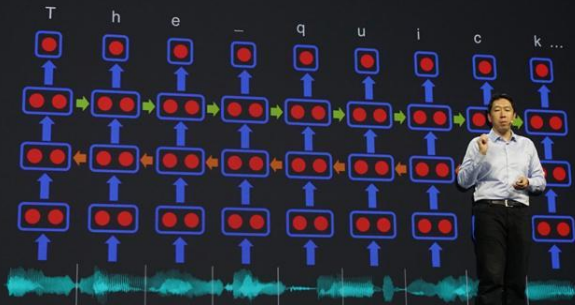
Image-Based Localization

[VIEW CHALLENGE DETAILS >](#)

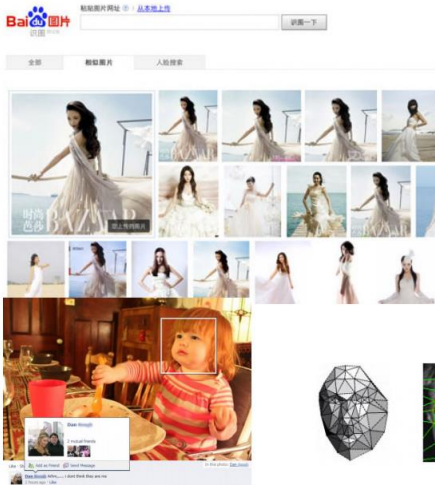
Speech Recognition

Baidu Deep Speech

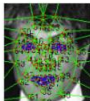
Bi-directional Recurrent Neural Network (BDRNN)



Machines that learn to recognise what they **see** and **hear** are at the heart of Apple, Google, Amazon, Facebook, Netflix, Microsoft, etc.



Biltzstein, Data Sciences Class



Sentiment Analysis

Review sentiment and summarization



my reading was similar to everyone's. she told me she was going to take her time and not rush me out of there. i was there not even 8 minutes she told me i was pregnant then she changed her mind and said i had a miscarriage. im 17 years old i told her she was wrong she then went on and said "i see you and your brother fight alot just know he loves you" i dont even have a brother.

she then told my friend she was going to get stabbed

Was this review helpful? ☒ Yes ☐ No 2
Ask taydube about Fatima's Psychic Studio

[Problem with this review?](#)

Biltzstein, Data
Sciences Class

Paul Bettany did a great role as the tortured father whose favorite little girl dies tragically of disease. For that, he deserves all the credit.

However, the movie was mostly about exactly that, keeping the adventures of Darwin as he gathered data for his theories as incomplete stories told to children and skipping completely the disputes regarding his ideas. Two things bothered me terribly: the soundtrack, with its whiny sound, practically shoving sadness down the throat of the viewer, and the movie trailer, showing some beautiful sceneries, the theological musings of him and his wife and the enthusiasm of his best friends as they prepare for a battle against blind faith, thus misrepresenting the movie completely.

To put it bluntly, if one were to remove the scenes of the movie trailer from the movie, the result would be a non descript family drama about a little child dying and the hardships of her parents as a result.

Clearly, not what I expected from a movie about Darwin, albeit the movie was beautifully interpreted.

Chatbot



Penuh Potensi. Tanpa Pretensi.

Kata.ai menyediakan chatbot yang menguasai Bahasa Indonesia dengan teknologi Natural Language Processing (NLP) untuk meningkatkan customer engagement.

Saya mau pesan tiket Jkt-Bali untuk besok.



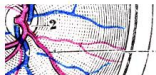
Konsumen Anda

Intent	= flight
Departure time	= tomorrow
Departure loc	= Jakarta
Arrival loc	= Bali



Messaging Platform

Healthcare



Completed • \$100,000 • 661 teams

Diabetic Retinopathy Detection

Tue 17 Feb 2015 – Mon 27 Jul 2015 (18 months ago)

Dashboard

Home

Data

Make a submission

Information

Description

Evaluation

Rules

Prizes

References

Timeline

Forum

Kernels

New Script

New Notebook

Leaderboard

Public

Private

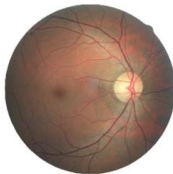
Private Leaderboard

1. Min-Pooling

Competition Details » [Get the Data](#) » [Make a submission](#)

Identify signs of diabetic retinopathy in eye images

Diabetic retinopathy is the leading cause of blindness in the working-age population of the developed world. It is estimated to affect over 93 million people.



The US Center for Disease Control and Prevention estimates that 29.1 million people in the US have diabetes and the World Health Organization estimates that 347 million people have the disease worldwide. Diabetic Retinopathy (DR) is an eye disease associated with long-standing diabetes. Around 40% to 45% of Americans with diabetes have some stage of the disease. Progression to vision impairment can be slowed or averted if DR is detected in time, however this can be difficult as the disease often shows few symptoms until it is too late to provide effective treatment.



Completed • \$200,000 • 192 teams

Second Annual Data Science Bowl

Mon 14 Dec 2015 – Mon 14 Mar 2016 (10 months ago)

Dashboard

Home

Data

Make a submission

Information

Description

Evaluation

Rules

Prizes

About the DSB

Deep Learning Tutorial

Fourier Based Tutorial

Resources

Timeline

Forum

Leaderboard

Public

Private

Private Leaderboard

1. Tencia & Woshialex
2. kunsthart
3. Julian de Wit

[Competition Details](#) » [Get the Data](#) » [Make a submission](#)

Transforming How We Diagnose Heart Disease

We all have a heart. Although we often take it for granted, it's our heart that gives us the moments in life to imagine, create, and discover. Yet cardiovascular disease threatens to take away these moments. Each day, 1,500 people in the U.S. alone are diagnosed with heart failure—but together, we can help. We can use data science to transform how we diagnose heart disease. By putting data science to work in the cardiology field, we can empower doctors to help more people live longer lives and spend more time with those that they love.

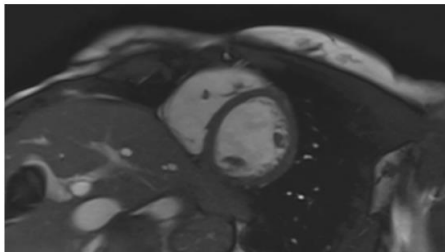
Declining cardiac function is a key indicator of heart disease. Doctors determine cardiac function by measuring end-systolic and end-diastolic volumes (i.e., the size of one chamber of the heart at the beginning and middle of each heartbeat), which are then used to derive the ejection fraction (EF). EF is the percentage of blood ejected from the left ventricle with each heartbeat. Both the volumes and the ejection fraction are predictive of heart disease. While a number of technologies can measure volumes or EF, Magnetic Resonance Imaging (MRI) is considered the gold standard test to accurately assess the heart's squeezing ability.

You only need to download one format of each file.

Each has the same contents but use different packaging methods.

In this dataset, you are given hundreds of cardiac MRI images in [DICOM](#) format. These are 2D cine images that contain approximately 30 images across the cardiac cycle. Each slice is acquired on a separate breath hold. This is important since the registration from slice to slice is expected to be imperfect.

The competition task is to create an automated method capable of determining the left ventricle volume at two points in time: after systole, when the heart is contracted and the ventricles are at their minimum volume, and after diastole, when the heart is at its largest volume.



The volumes at systole, V_S , and diastole, V_D , form the basis of an important clinical measurement known as the [ejection fraction](#):

$$100 * \frac{V_D - V_S}{V_D}.$$



\$1,000,000 • 874 teams

Data Science Bowl 2017

Thu 12 Jan 2017

Merger and Entry Deadline

Wed 12 Apr 2017 (2 months to go)

Dashboard

Home

Data

Make a submission

Information

Rules
about-the-dsb
description
evaluation
prizes
resources
timeline
tutorial

Forum

Kernels

New Script
New Notebook

Leaderboard

[Public Leaderboard](#)[Competition Details](#) » [Get the Data](#) » [Make a submission](#)

Can you improve lung cancer detection?

In the United States, lung cancer strikes 225,000 people every year, and accounts for \$12 billion in health care costs. Early detection is critical to give patients the best chance at recovery and survival.

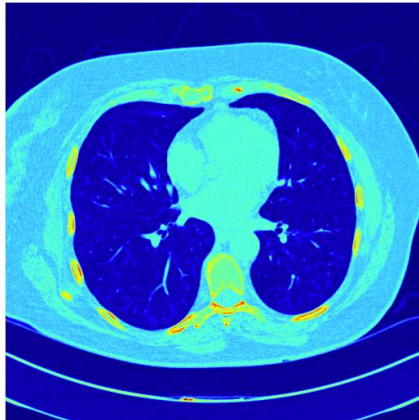
One year ago, the office of the U.S. Vice President spearheaded a bold new initiative, the Cancer Moonshot, to make a decade's worth of progress in cancer prevention, diagnosis, and treatment in just 5 years.

In 2017, the Data Science Bowl will be a critical milestone in support of the Cancer Moonshot by convening the data science and medical communities to develop lung cancer detection algorithms.

Using a data set of thousands of high-resolution lung scans provided by the National Cancer Institute, participants will develop algorithms that accurately determine when lesions in the lungs are cancerous. This will dramatically reduce the false positive rate that plagues the current detection technology, get patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

In this dataset, you are given over a thousand low-dose CT images from high-risk patients in DICOM format. Each image contains a series with multiple axial slices of the chest cavity. Each image has a variable number of 2D slices, which can vary based on the machine taking the scan and patient.

The DICOM files have a header that contains the necessary information about the patient id, as well as scan parameters such as the slice thickness.



Conclusion from Machine Learning Application

When to apply machine learning

- ❑ Human expertise is absent (e.g. *Navigating on Mars*)
- ❑ Humans are unable to explain their expertise (e.g. *Speech recognition, vision, language*)
- ❑ Solution changes with time (e.g. *Tracking, temperature control, preferences*)
- ❑ Solution needs to be adapted to particular cases (e.g. *Biometrics, personalization*)
- ❑ The problem size is too vast for our limited reasoning capabilities (e.g. *Calculating webpage ranks, matching ads to facebook pages*)



Nando de Freitas,
Intro ML

Q: Why now?

Data - User generated content

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, YIMail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, YIMail)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, YIMail)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

Alex Smola,
Introduction ML

flickr



You Tube

DISQUS

yelp

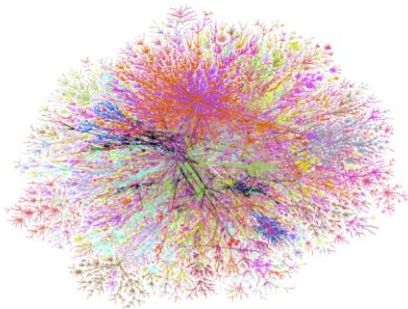
>1B images, 40h video/minute

Carnegie Mellon University

Data

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, YIMail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, YIMail)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, YIMail)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

Alex Smola,
Introduction ML



>10B useful webpages

Data - Identity & Graph

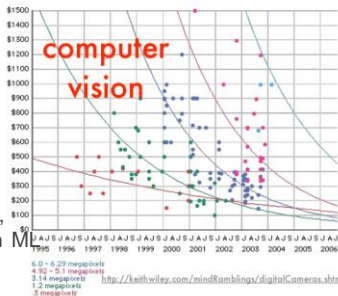
- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, YIMail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, YIMail)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, YIMail)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



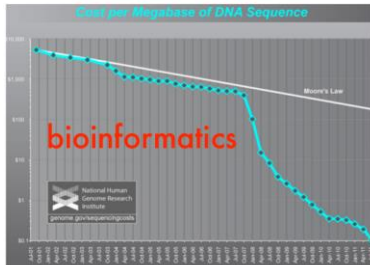
100M-1B vertices

Carnegie Mellon University

Many more sources



Alex Smola,
Introduction ML



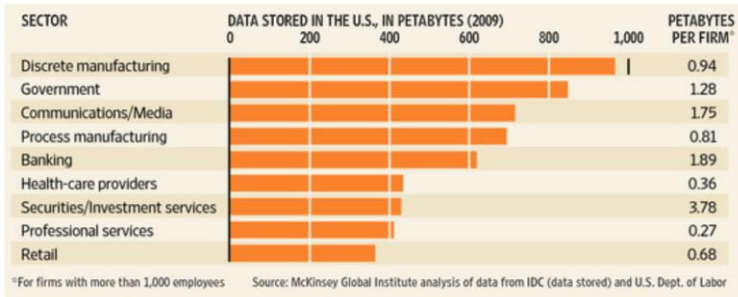
personalized sensors



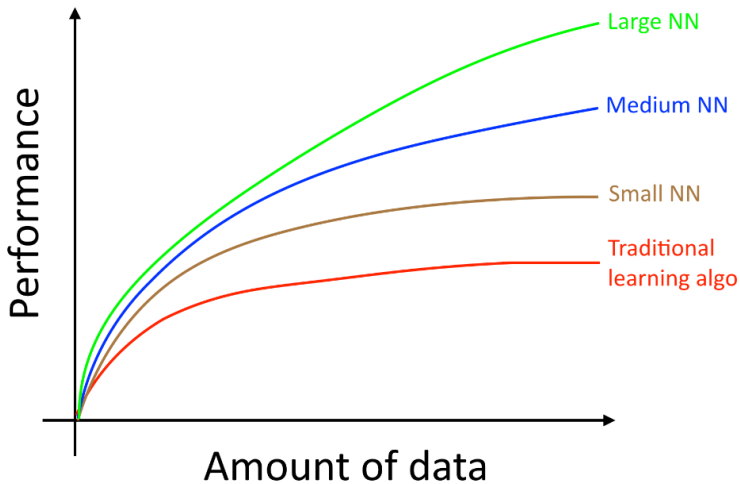
ubiquitous control

nUniversity

Big Data

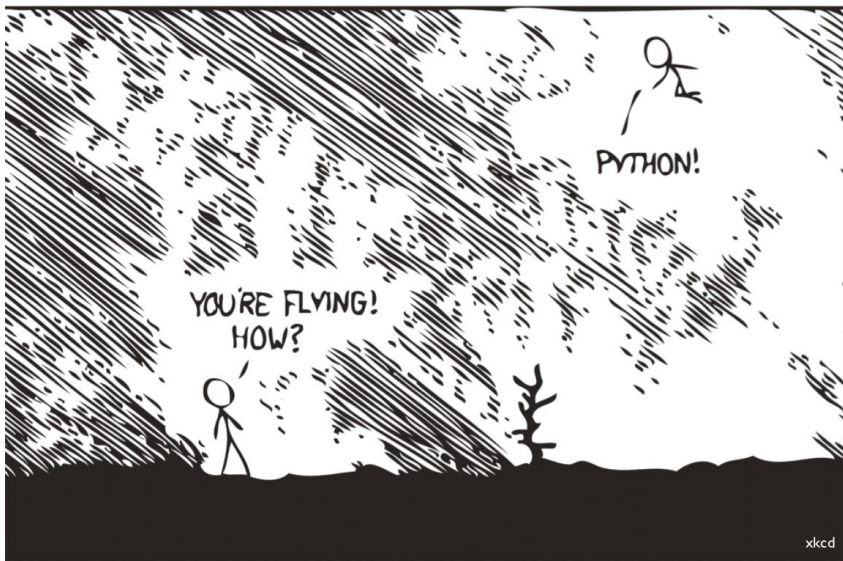


we need Big Learning



Andrew Ng,
Machine
Yearning

Tools



Programming

IP[y]: IPython
Interactive Computing

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Biltzstein, Data
Sciences



machine learning in Python



NumPy



matplotlib



SciPy.org



Sponsored by
ENTHOUGHT