

PACMANN AI

# Machine Learning for Beginners

# PACMANN AI

## Introduction to Machine Learning

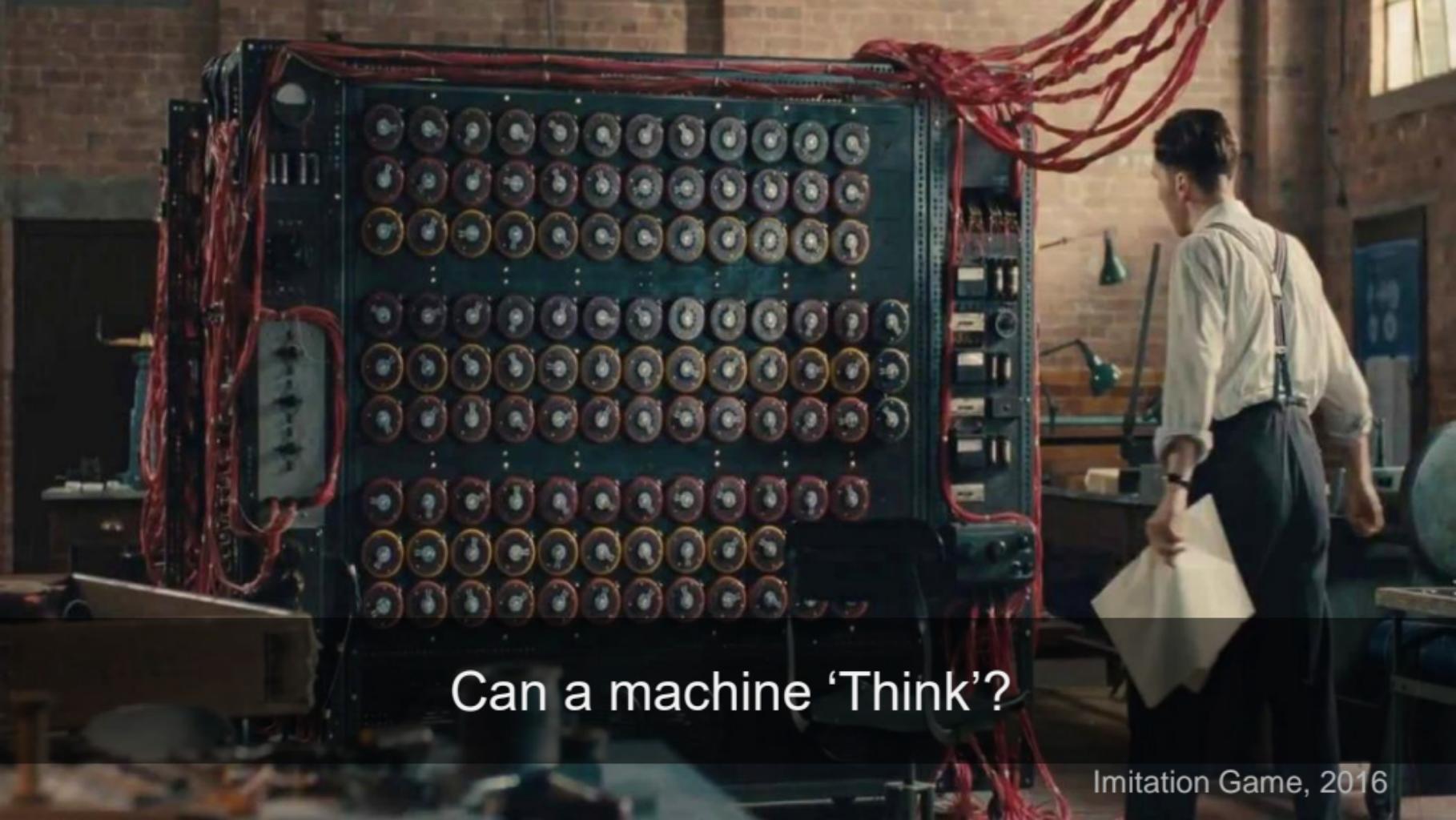
Adityo Sanjaya  
Head of Research PACMANN AI  
Email: adityosanjaya3.14@gmail.com  
+62 85777490099

# Introduction to Machine Learning

Adityo Sanjaya



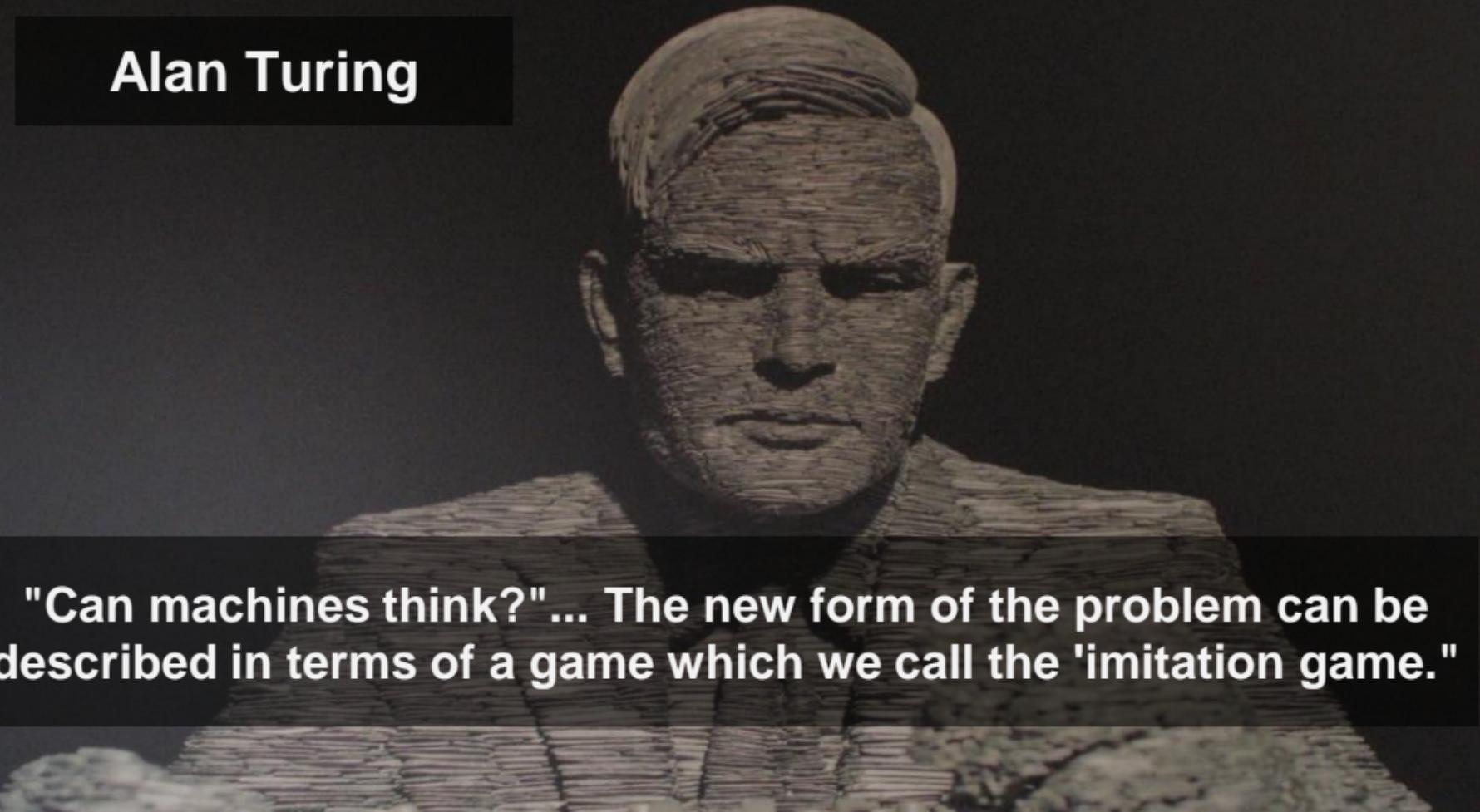
Imitation Game, 2016



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

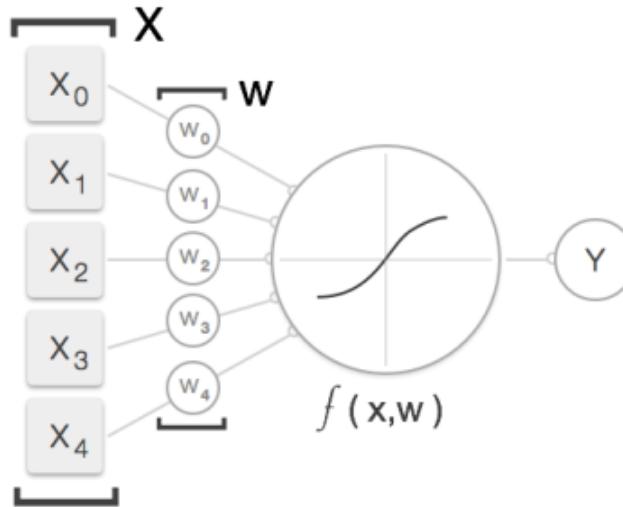
- Born from the ambitious goal of Artificial Intelligence



Dartmouth AI Conferences

# Machine Learning Historical Background

- Perceptron: first artificial neuron.

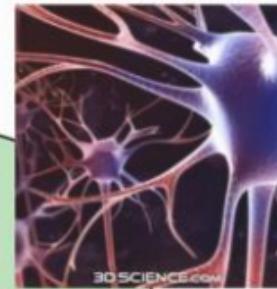
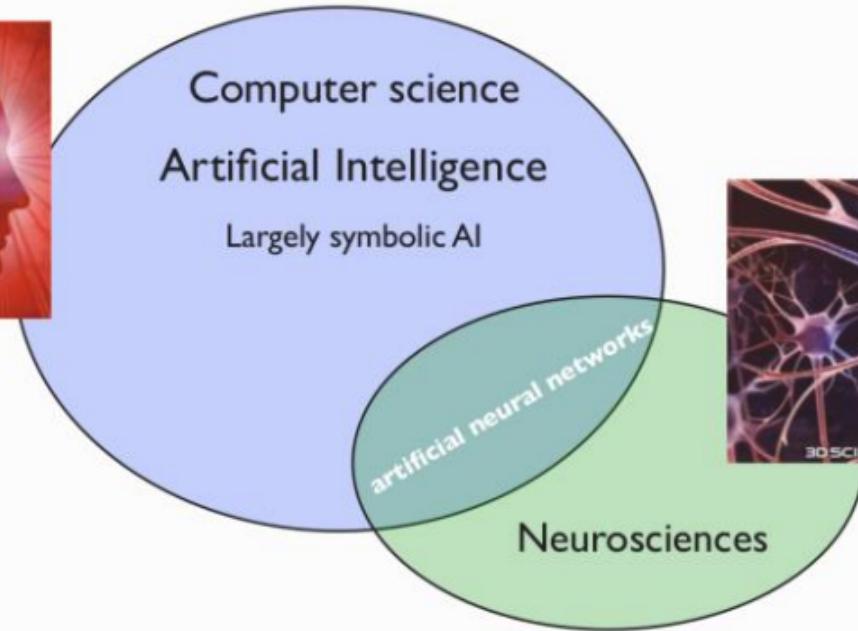


- Machine Learning:
- Learning
  - Poor handling of uncertainty
  - Hard coding

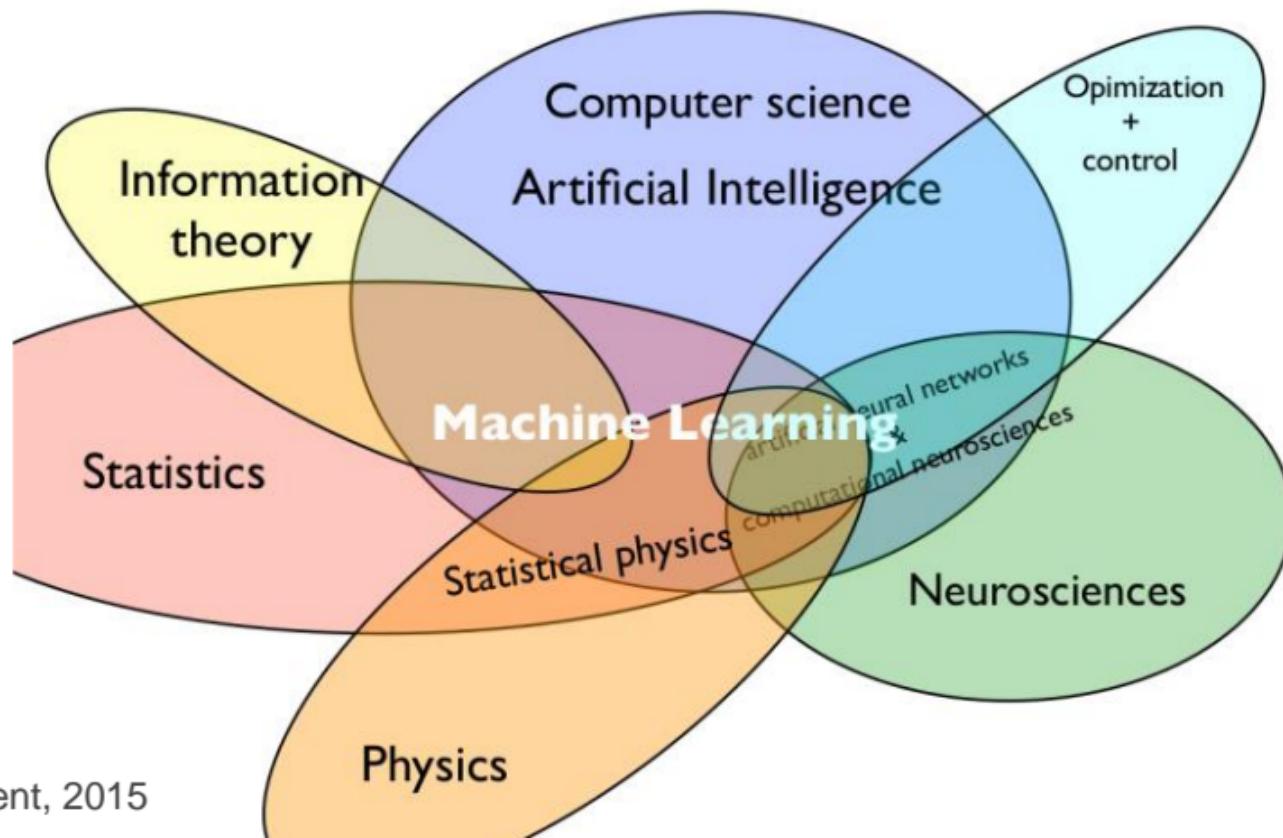
Rosenblatt, source: Wikipedia

# Machine Learning Historical Background

## Artificial Intelligence 1960s



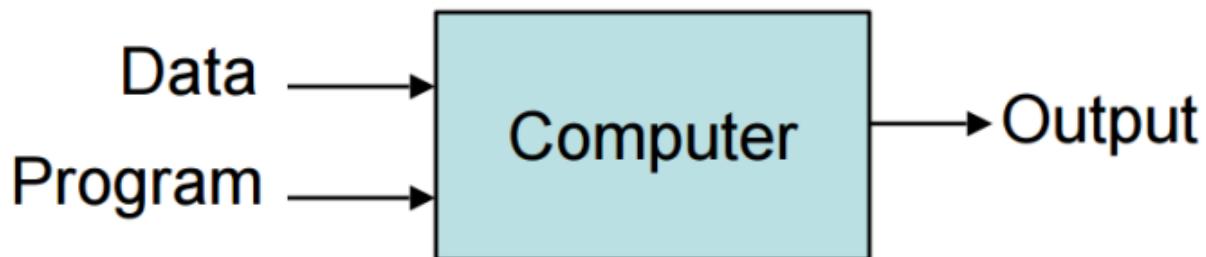
# Machine Learning Current View



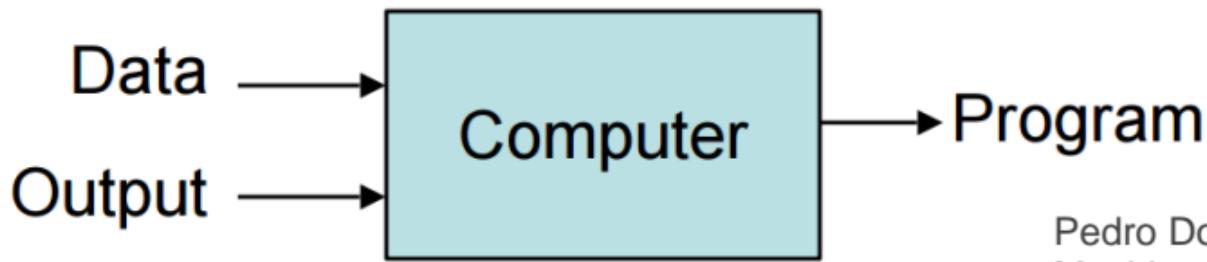
# What is Machine Learning?

- “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



# Motivation

1. The intent of the lecture is **not** to explain **details** of building ML systems.
1. Rather it is an overview of what **can** be accomplished with ML.
1. If it inspires you, then you'll have to take the course and learn a lot of **cool stuff** !



# Machine Learning in The Wild

# Classification

# Spam Filtering

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

Google ham 1–50 of 15,803 < > ⚙

Gmail • COMPOSE

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

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

DiscountMags.com \$3.99 Business & Finance Sale... starts now! - Trouble Seeing This Email? View as Webpage STOP these e-mail 12:03 pm

support, Alex (3) Your order has shipped... - please send to the address below for an exchange remotesremotes.com/exchange 7:22 am

American Airlines AAdvantage AAdvantage eSummary - January 2013 - VIEW IN WEB BROWSER >> http://americanairlines.ed10.net/rJC 1:17 am

Taesup, Alex, Taesup (3) Happy new year! - Hi Alex, Thanks for your condolence. I will arrive at Berkeley on 16th (wed) night. So, I car Jan 11

Alex Smola, Introduction ML

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

Google in:spam spam 1–50 of 244 < > ⚙

Gmail • COMPOSE

Inbox (7,180) Important Sent Mail Drafts (61) All Mail Circles [Gmail] Done (1,006) [Imap]/Drafts [Imap]/Sent alex.smola@yahoo...

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

maeef [EI&JSTP Index]2013机械与自动化工程国际会议论文: [alex.smola@gmail.com] - 歌的老师, 您好: 机械与 Dear Valued Customers, Low Interest Rate Loan - Dear Valued Customers, Do you need a loan or funding for any of the following reasons? Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOGY Steven Cooke Congratulations Alex, \$150 awaits you - Alex: IMPORTANT - NOTICE OF WINNINGS Please make sure yo 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】 【不收-不要】 First-Class Mail Service Tracking ID (G)BGD35 849 603 4893 4550 - Fed Ex Order: JN-3339-28981768 Order Date: Thursday, 3 Janua garjeti Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOGY Candy.Li 中医,不只当老板的代言人 Ronan Morgan Ronan Morgan just sent you a personal message. - LinkedIn Ronan Morgan just sent you a private message RE/MAX® newsletter 2013 Valueable Offer! - Hello Friend, RE/MAX® has issued 2013 valuable property offer in your resident from newsletter WWW2013 - Newsletter 6 - See the Portuguese and Spanish version right after the English version CJC editor Chinese Journal of Cancer Research (CJCR) has been indexed by Pubmed and PMC - Click here if this e-mail garjeti (2) Call for Research Papers - GLOBAL ADVANCED RESEARCH JOURNAL OF ENGINEERING, TECHNOLOGY

# 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  
Techniques by Daphne Koller

★★★★★ (5)

\$71.52

Amazon  
books

# Netflix Prize

The image shows a screenshot of the Netflix Prize website. At the top, there's a red banner with the word "NETFLIX" and a yellow banner below it with "Netflix Prize" and a large "COMPLETED" stamp. Below these are navigation links: Home, Rules, Leaderboard, and Update. The main content area features a dark background with a blurred image of two people looking at a screen. On the right, a white box contains a blue "Congratulations!" heading and text about the prize's goal and the awarding of the \$1M Grand Prize to BellKor's Pragmatic Chaos team on September 21, 2009. It also encourages users to explore the algorithm, leaderboard, and forum. At the bottom, it thanks contributors for improving movie recommendations.

NETFLIX

Netflix Prize

COMPLETED

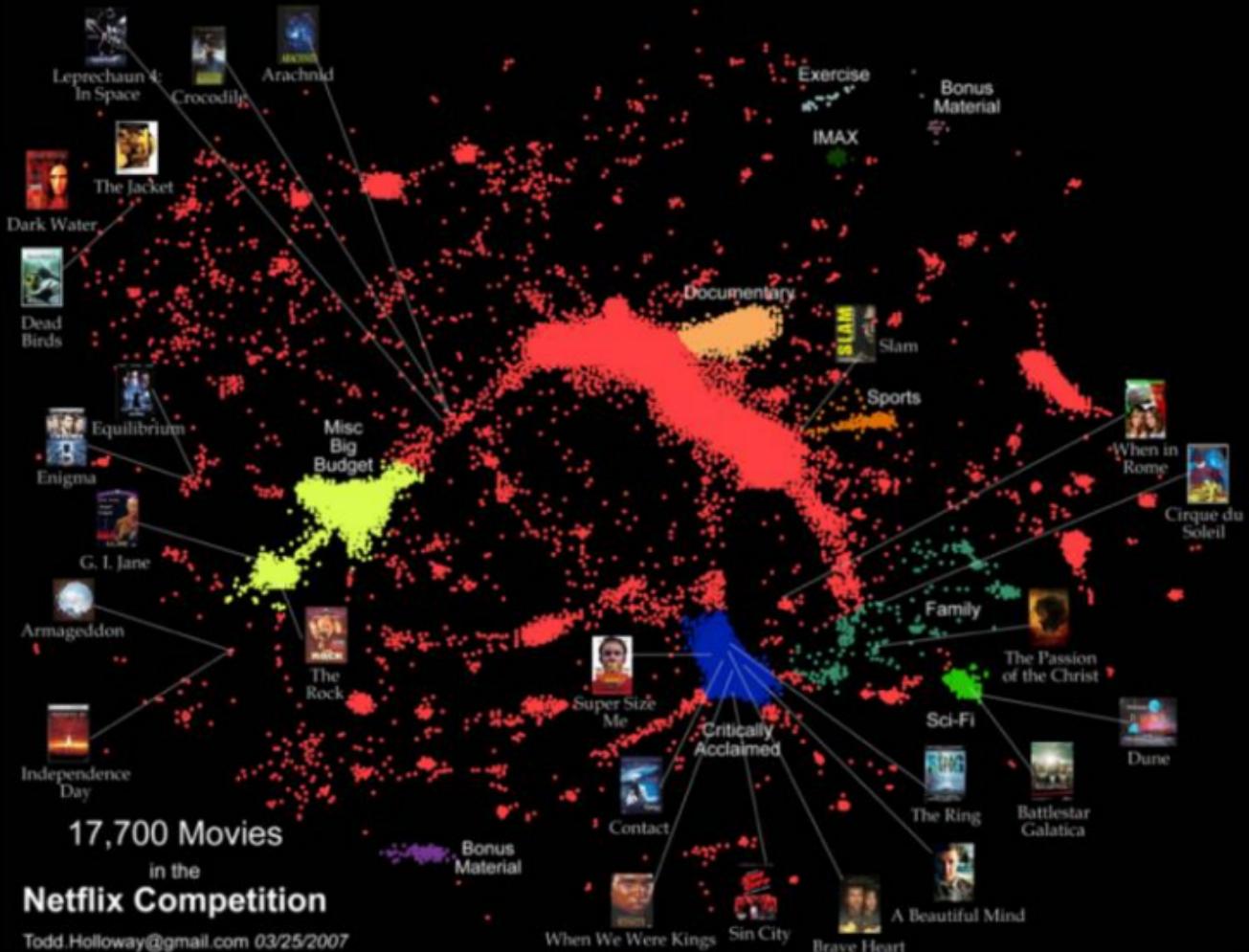
Home Rules Leaderboard Update

Congratulations!

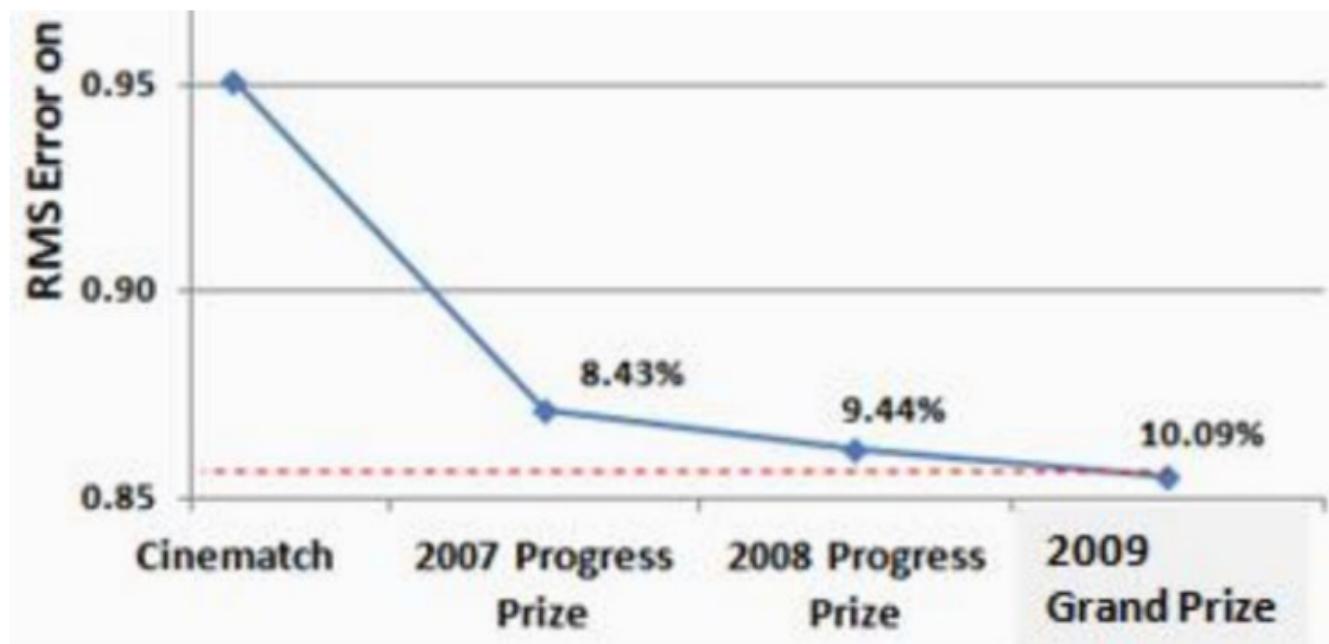
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

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.



# Netflix Error time by time



HBR, Oct 2012

Friday, April 6, 2012

# We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment.

In 2006 we announced the Netflix Prize, a machine learning and data mining competition for movie rating prediction. We offered \$1 million to whoever could improve the accuracy of our existing system and CinemaMatch by 10%. We received many entries, and the competition was a great opportunity to evaluate and quantify the root mean squared error (RMSE) of the predictions. The race was on to beat our RMSE of 0.9625 with the first one of reducing it to 0.8872 or less.

The best performance in the ensemble: Matrix Factorization (which the community generally called SVD, Singular Value Decomposition) and Restricted Boltzmann Machines (RBM). SVD by itself provided a 0.8914 RMSE, while RBM alone provided a competitive but slightly worse 0.8990 RMSE. A linear blend of these two reduced the error to 0.88. To put these algorithms to use, we had to work to overcome some limitations, for instance that they were built to handle 100 million ratings, instead of the more than 5 billion that we have, and that they were not built to adapt as members added more ratings. But once we overcame those challenges, we put the two algorithms into production, where they are still used as part of our recommendation engine.

Links

- Netflix Home
- Netflix Data Science Blog
- Netflix America Latina Blog
- Netflix CTO Blog
- Netflix UK & Ireland Blog
- Netflix India
- Netflix LAT Engineering

RSS Feed

About the Netflix Tech Blog

This is a Netflix blog focused on technology and technology issues. We'll share our perspectives, decisions and challenges regarding the software we build and use to create the Netflix service.

Xavier Amatriain and Justin Basilico, 2012

# Time Series Prediction

# Prediction



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	QUPIKA	\$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	BIDOOF	\$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	TEDIM	\$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*



KIRAK: 0.65013

CAMBRIES: 0.89961

Imitating Behavior

# Imitation Learning



# Imitation Learning in Games



Avatar learns from  
your behavior

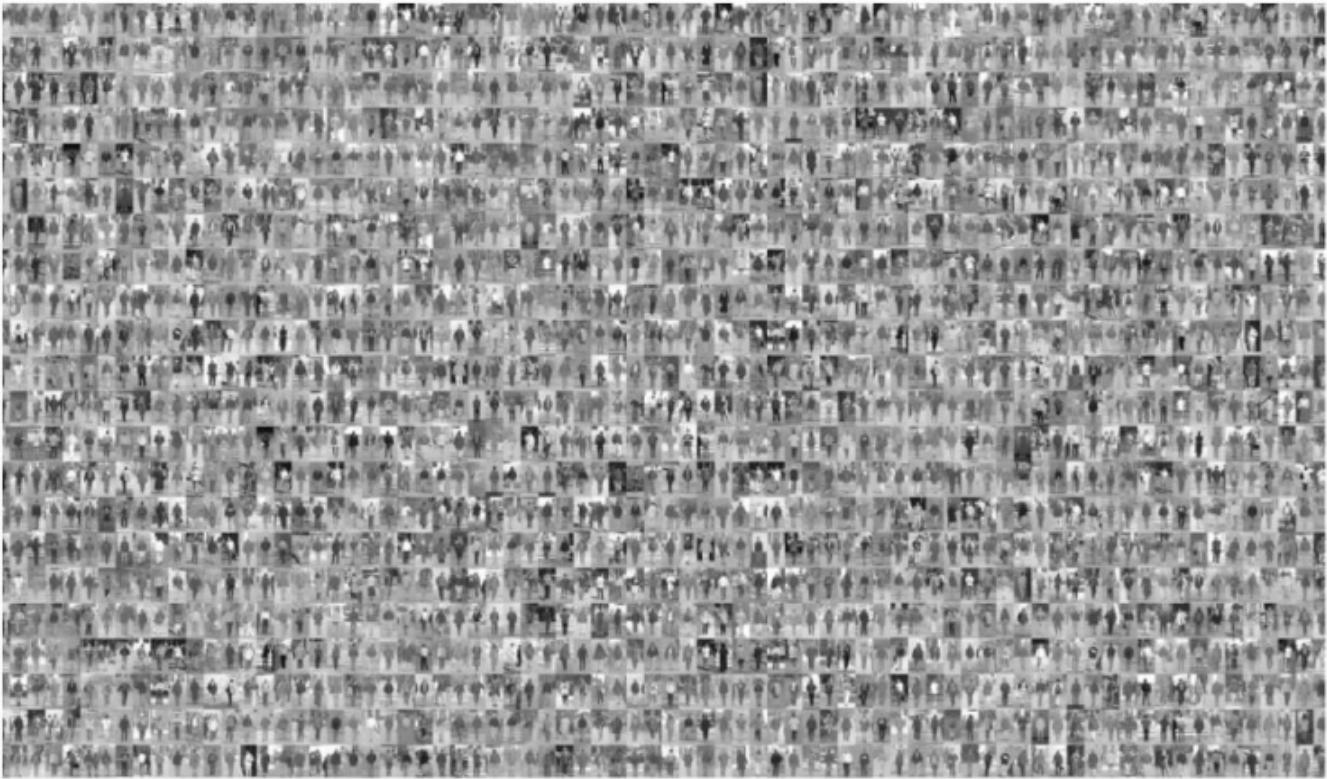
Alex Smola,  
Introduction M

Black & White  
Lionsgate Studios

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



# Machine Learning in Industry



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**

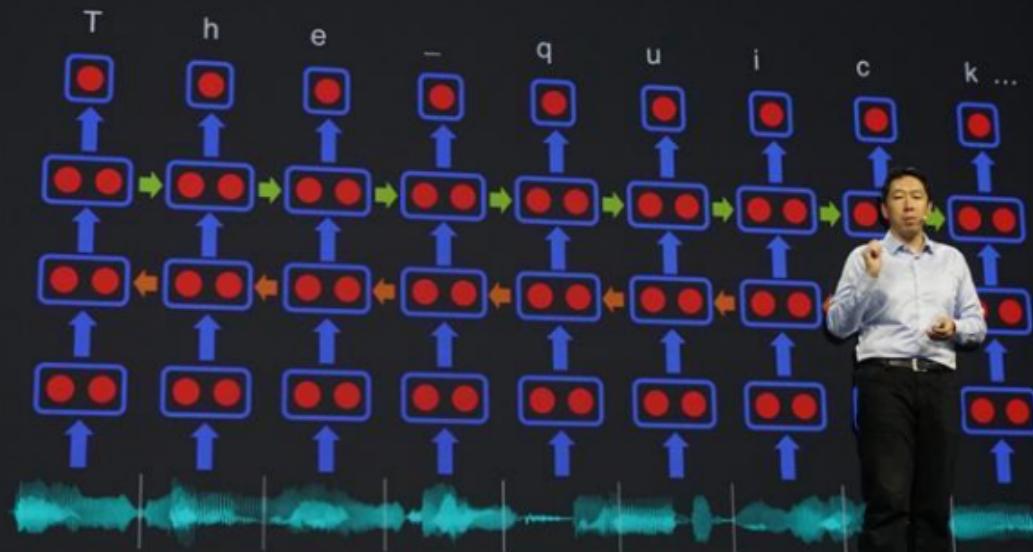
Image-Based Localization

[VIEW CHALLENGE DETAILS >](#)

# Speech Recognition

Baidu Deep Speech

Bi-directional Recurrent  
Neural Network (BDRNN)

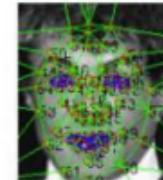


Andrew Ng

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



WORLD'S MOST TRUSTED TRAVEL ADVICE™

my reading was similar to everyones. 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 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.

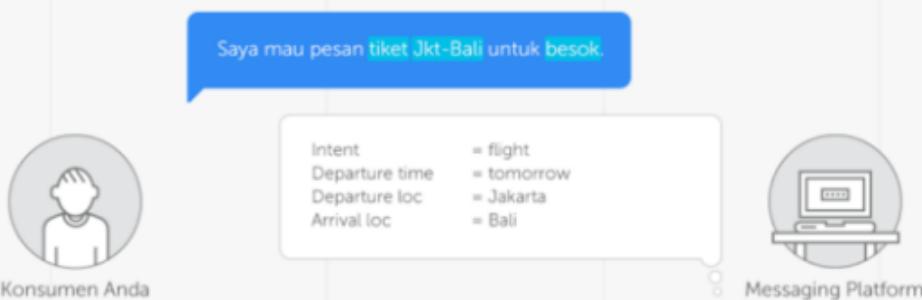
[Kotzias, Denil, Blunsom & NdF, 2014]

# Chatbot

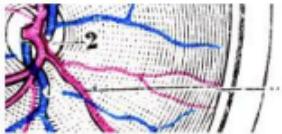


## Penuh Potensi. Tanpa Pretensi.

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



# 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

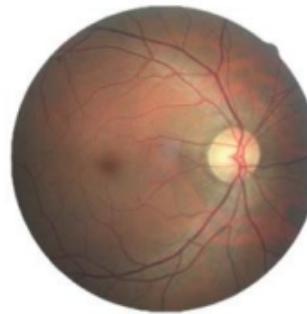
### Private Leaderboard

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 &amp; Woshalex

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.

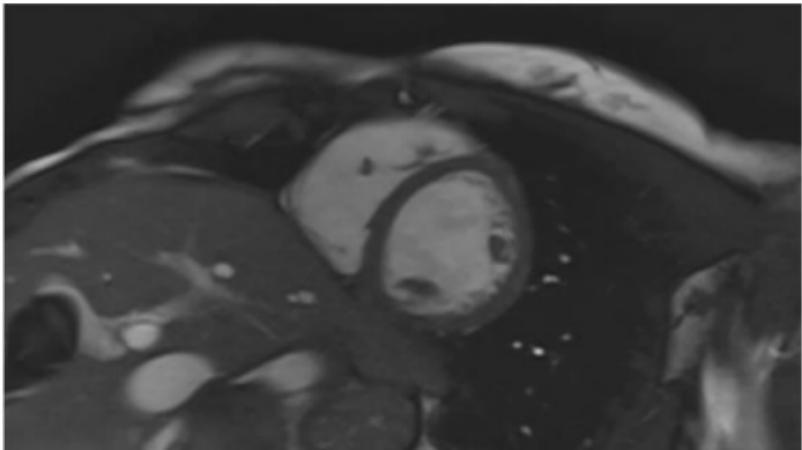
Declining cardiac function is a key indicator of heart disease. Doctors determine cardiac function by measuring end-systolic and end-diastolic volumes (i.e., the size of one chamber of the heart at the beginning and middle of each heartbeat), which are then used to derive the ejection fraction (EF). 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.

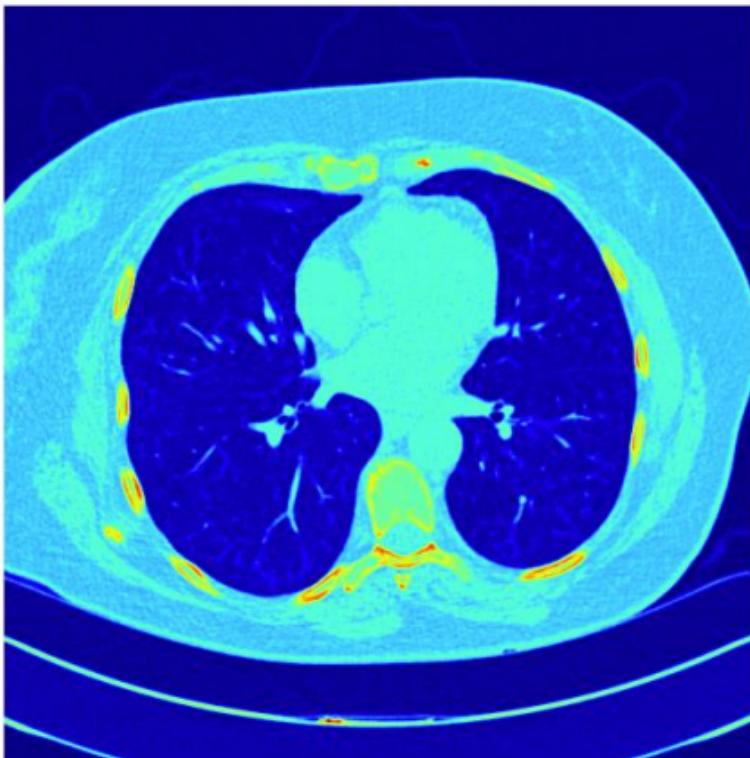
One year ago, the office of the U.S. Vice President spearheaded a bold new initiative, the Cancer Moonshot, to make a decade's worth of progress in cancer prevention, diagnosis, and treatment in just 5 years.

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.



Education

# Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education

Xiaojin Zhu

Department of Computer Sciences, University of Wisconsin-Madison  
Madison, WI, USA 53706  
jerryzhu@cs.wisc.edu

## Abstract

I draw the reader's attention to machine teaching, the problem of finding an optimal training set given a machine learning algorithm and a target model. In addition to generating fascinating mathematical questions for computer scientists to ponder, machine teaching holds the promise of enhancing education and personnel training. The Socratic dialogue style aims to stimulate critical thinking.

## Of Machines

**Q:** *I know machine learning; What is machine teaching?*

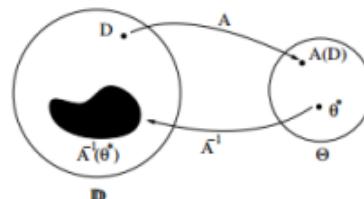
Consider a "student" who is a machine learning algorithm, for example, a Support Vector Machine (SVM) or kmeans clustering. Now consider a "teacher" who wants the student to learn a target model  $\theta^*$ . For example,  $\theta^*$  can be a specific hyperplane in SVM, or the location of the  $k$  centroids in kmeans. The teacher knows  $\theta^*$  and the student's learning algorithm, and teaches by giving the student training examples. Machine teaching aims to design the optimal training set  $D$ .

**Q:** *What do you mean by optimal?*

One definition is the cardinality of  $D$ : the smaller  $|D|$  is, the better. But there are other definitions as we shall see.

**Q:** *If we already know the true model  $\theta^*$ , why bother training a learner?*

The applications are such that the teacher and the learner are separate entities, and the teacher cannot directly "hard wire" the learner. One application is education where the learner is a human student. A more sinister "application" is security where the learner is an adaptive spam filter, and the "teacher" is a hacker who wishes to change the filtering behavior by sending the spam filter specially designed messages. Regardless of the intention, machine teaching aims to maximally influence the learner via optimal training data.



Given a training set  $D \in \mathbb{D}$ , machine learning returns a model  $A(D) \in \Theta$ . Note  $A$  in general is many-to-one. Conversely, given a target model  $\theta^* \in \Theta$  the inverse function  $A^{-1}$  returns the set of training sets that will result in  $\theta^*$ . Machine teaching aims to identify the optimal member among  $A^{-1}(\theta^*)$ . However,  $A^{-1}$  is often challenging to compute, and may even be empty for some  $\theta^*$ . Machine teaching must handle these issues.

**Q:** *Isn't machine teaching just active learning / experimental design?*

No. Recall active learning allows the learner to "ask questions" by selecting items  $x$  and asking an oracle for its label  $y$  (Settles 2012). Consider learning a noiseless threshold classifier in  $[0, 1]$ , as shown below.



To learn the decision boundary  $\theta^*$  up to  $\epsilon$ , active learning needs to perform binary search with  $\log(\frac{1}{\epsilon})$  queries. In contrast, in machine teaching the teacher only needs **two** examples:  $(\theta^* - \frac{\epsilon}{2}, -1)$ ,  $(\theta^* + \frac{\epsilon}{2}, +1)$ . The key difference is that the teacher knows  $\theta^*$  upfront and doesn't need to explore. Note passive learning, where training items  $x$  are sampled *iid* uniformly from  $[0, 1]$ , requires  $O(\frac{1}{\epsilon})$  items.

**Q:** *So the teacher can create arbitrary training items?*

The answer is no. The teacher can only create training items that are consistent with the target model  $\theta^*$ .

## Abstract

I draw the reader's attention to machine teaching, the problem of finding an optimal training set given a machine learning algorithm and a target model. In addition to generating fascinating mathematical questions for computer scientists to ponder, machine teaching holds the promise of enhancing education and personnel training. The Socratic dialogue style aims to stimulate critical thinking.

# Smart Farming



# What Is Smart Farming?

PACMANN AI



There are some problems in agriculture, that farmers don't have any **knowledge** about how to plant better, **manage** their lands effectively. Farmers also have a **limited access** to soft-loan. We think we can **solve** all these problems

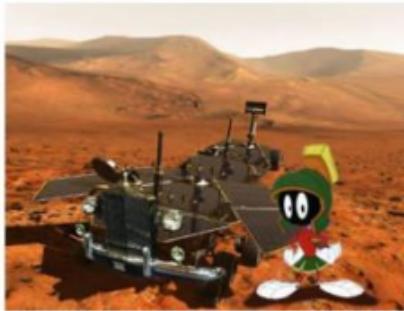
# Credit Rating Map PACMANN AI



## 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, Y!Mail, 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, Y!Local)

Alex Smola,  
Introduction ML

- 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, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



DISQUS



>1B images, 40h video/minute

Carnegie Mellon University

# Data

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, 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, Y!Local)
- 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, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



>10B useful webpages

Alex Smola,  
Introduction ML

# Data - Identity & Graph

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, 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, Y!Local)
- 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, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

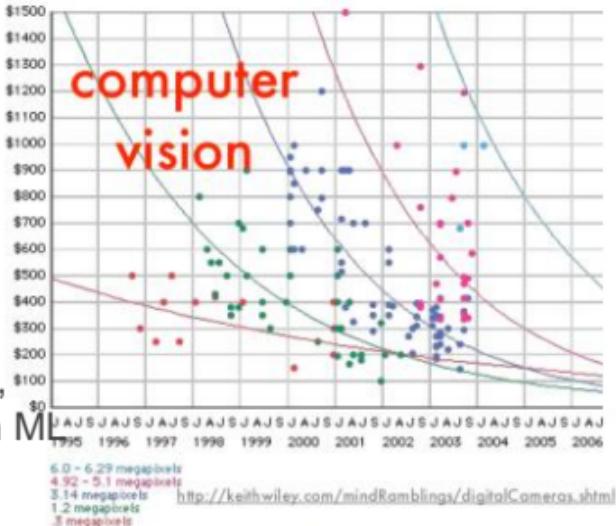
Alex Smola,  
Introduction ML



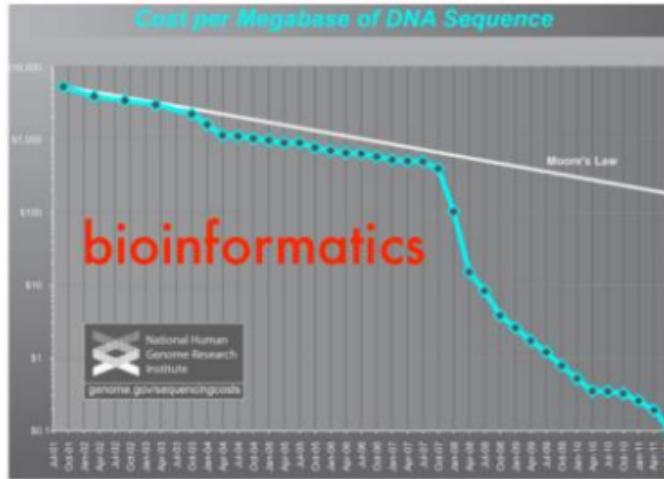
100M-1B vertices

Carnegie Mellon University

# Many more sources



personalized sensors

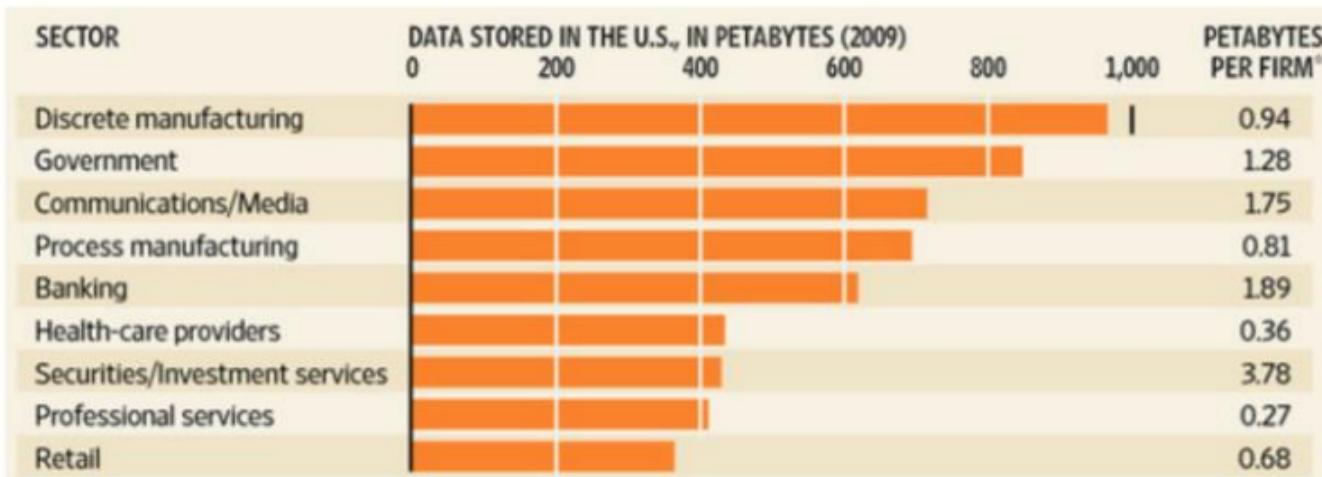


ubiquitous control University

Alex Smola,  
Introduction M

University

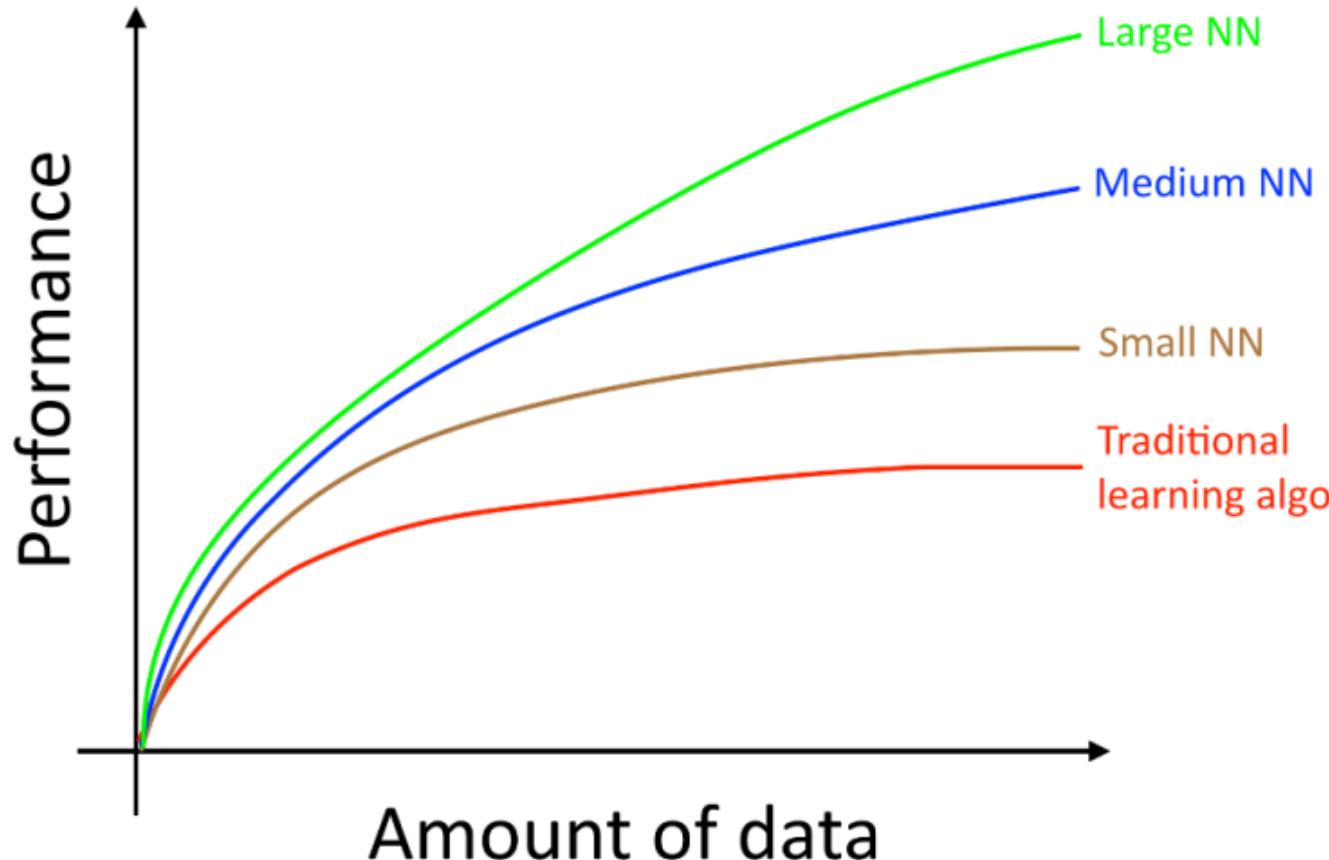
# Big Data



<sup>a</sup>For firms with more than 1,000 employees

Source: McKinsey Global Institute analysis of data from IDC (data stored) and U.S. Dept. of Labor

we need Big Learning



ML Researcher vs Statistician vs Data Analyst vs Data Scientist ?

ML Researcher vs Statistician vs Data Analyst vs Data Scientist ?

It doesn't matter!

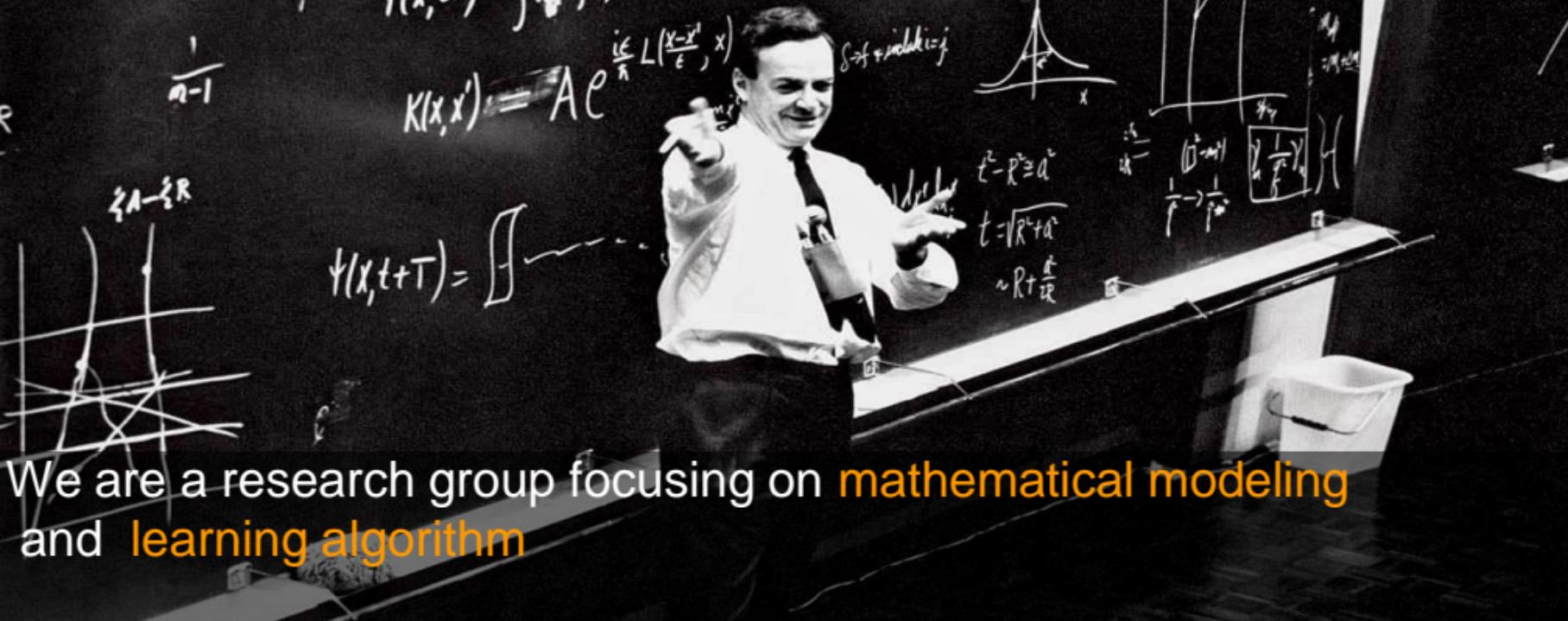
# Hal Varian Explains...

The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data.”** – Hal Varian

Who are we?

# PACMANN

Probably Approximately Correct Machine of Artificial Neural Networks



We are a research group focusing on mathematical modeling  
and learning algorithm

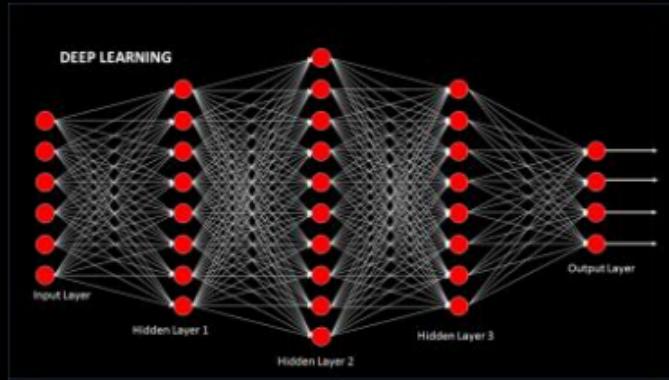
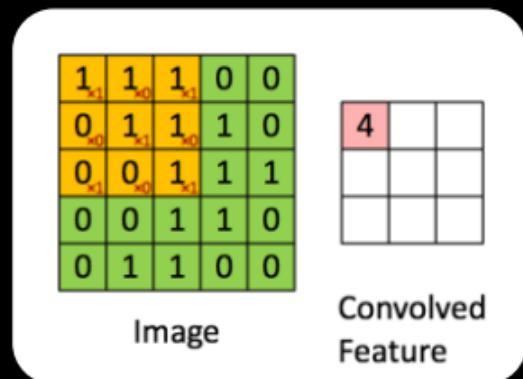
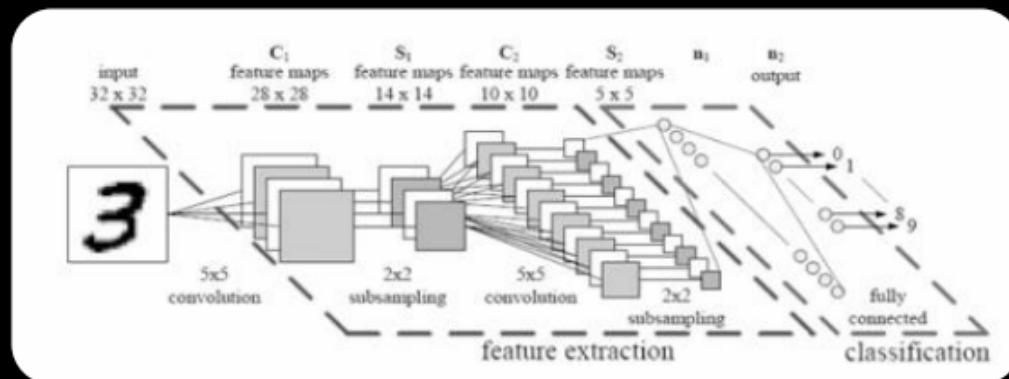


What we do:

We are focusing on Computational Learning Theory as in PAC Learning and Artificial Neural Networks especially Deep Learning.

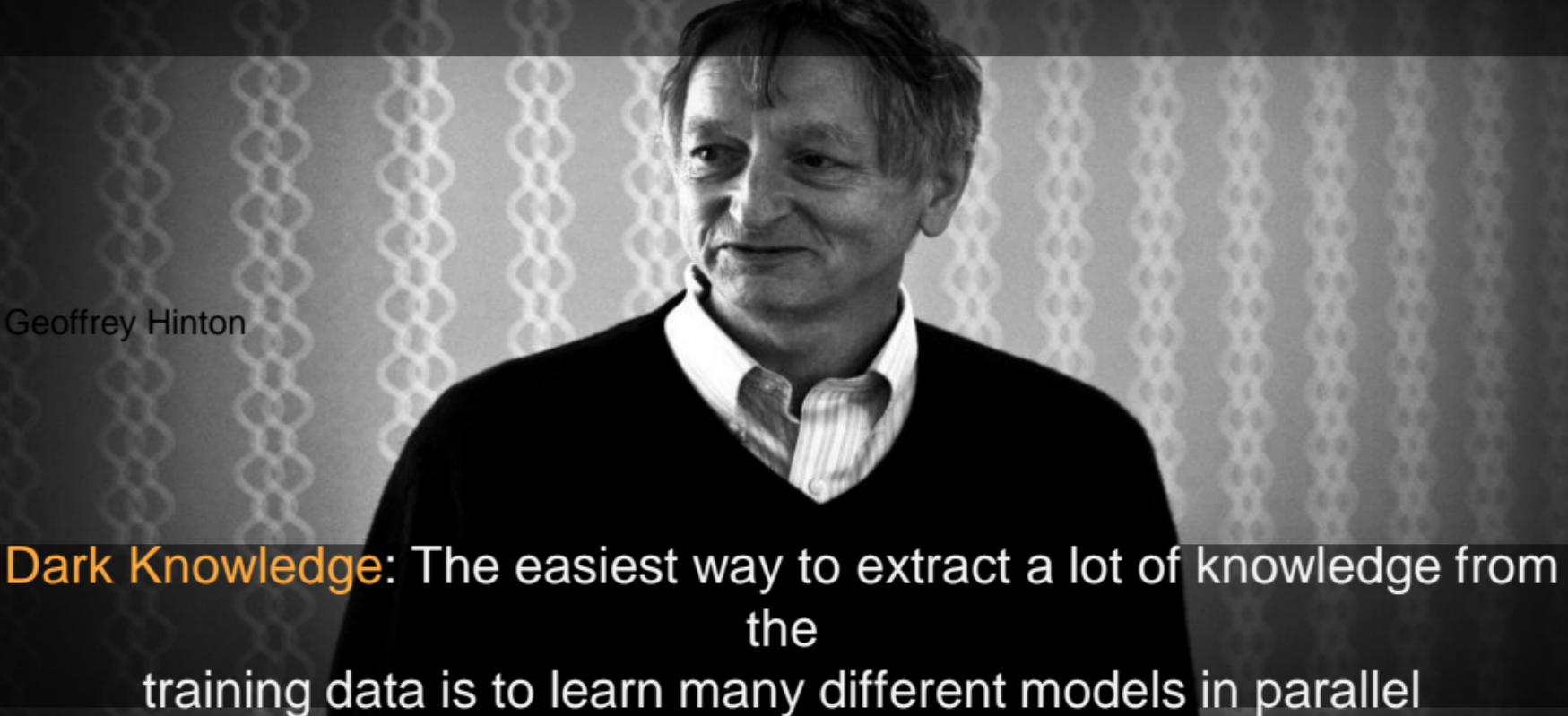
PACMANN  
Research

# PACMANN Research: Image Recognition

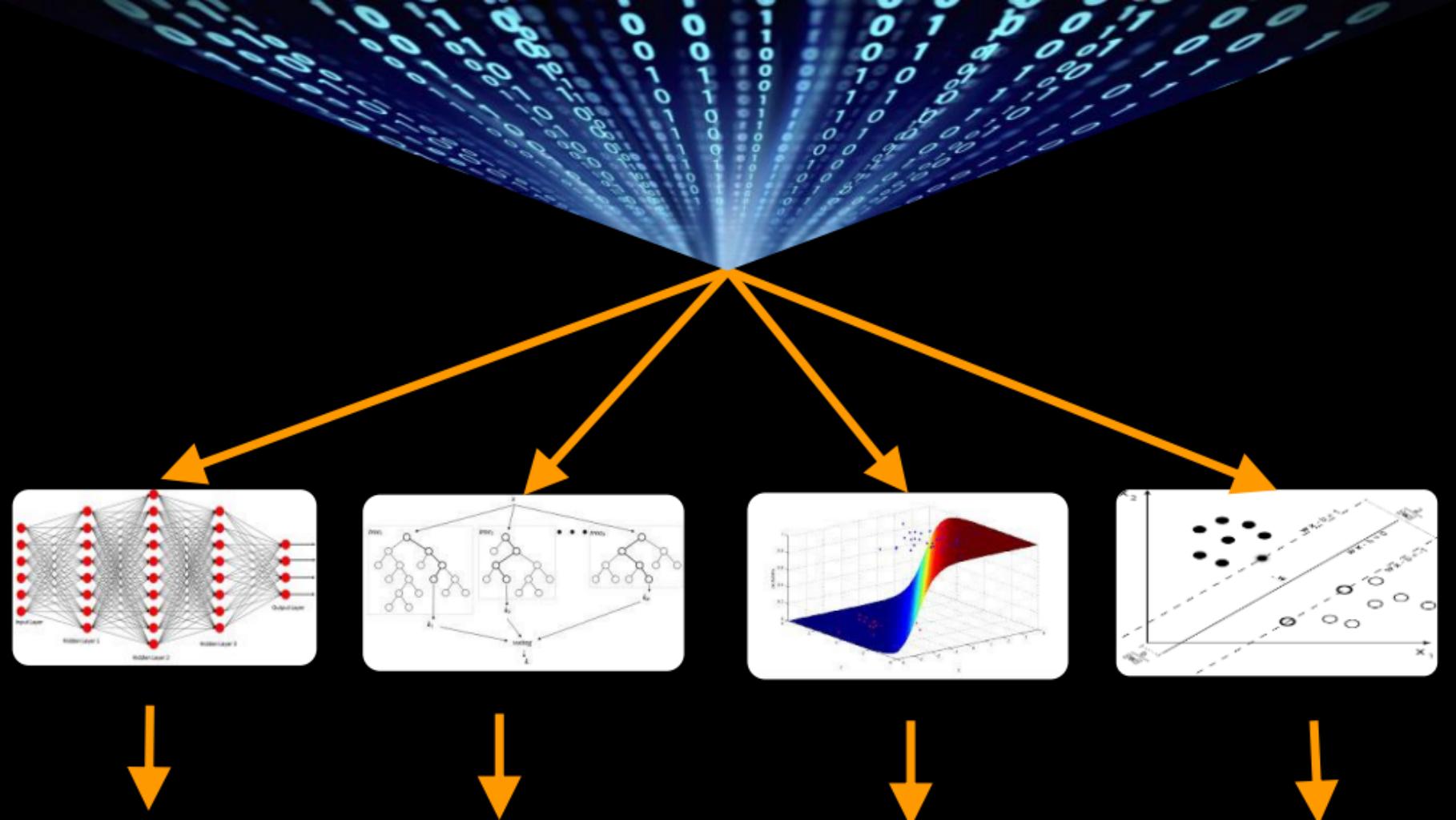


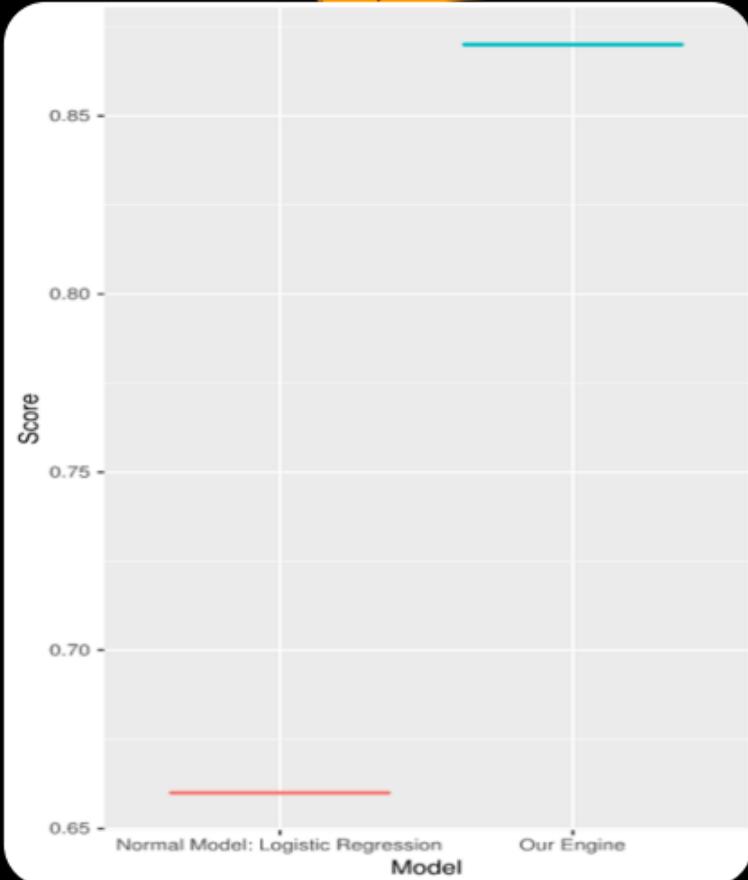
# PACMANN Research: Optimizing Predictive Power

Geoffrey Hinton

A black and white portrait of Geoffrey Hinton, a middle-aged man with short hair, wearing a dark sweater over a striped shirt and tie. He is looking slightly to his left with a faint smile. The background is a light-colored wall with a subtle, repeating geometric pattern.

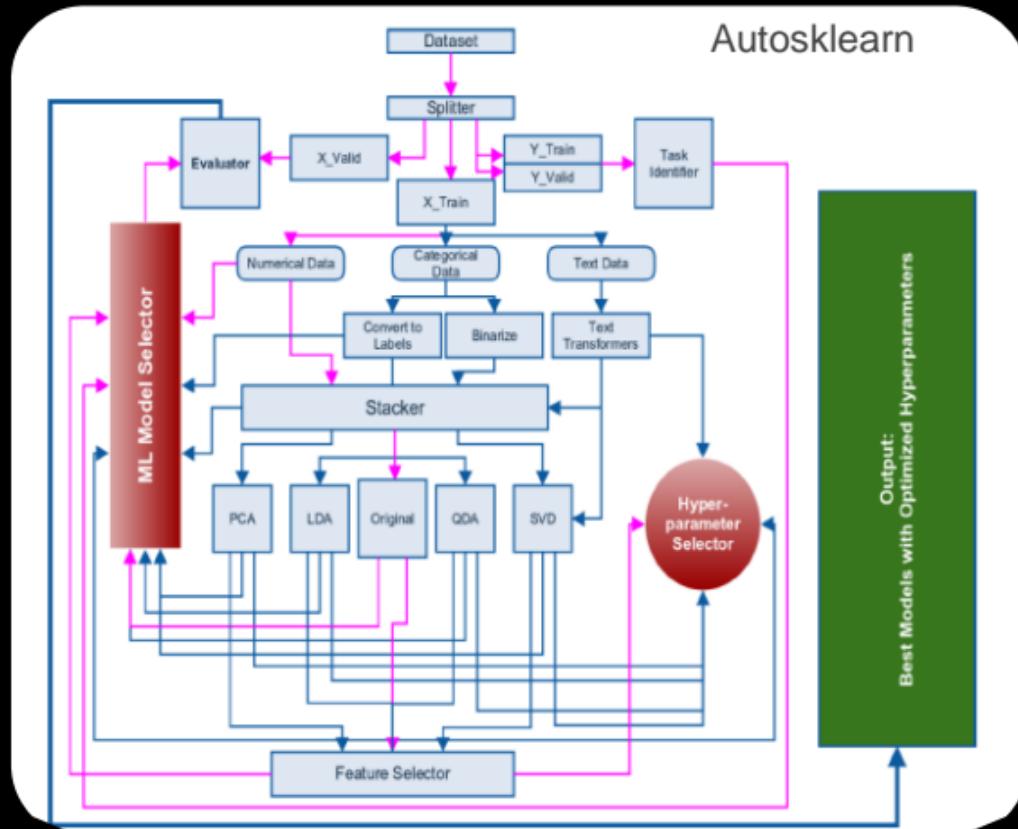
**Dark Knowledge:** The easiest way to extract a lot of knowledge from  
the  
training data is to learn many different models in parallel





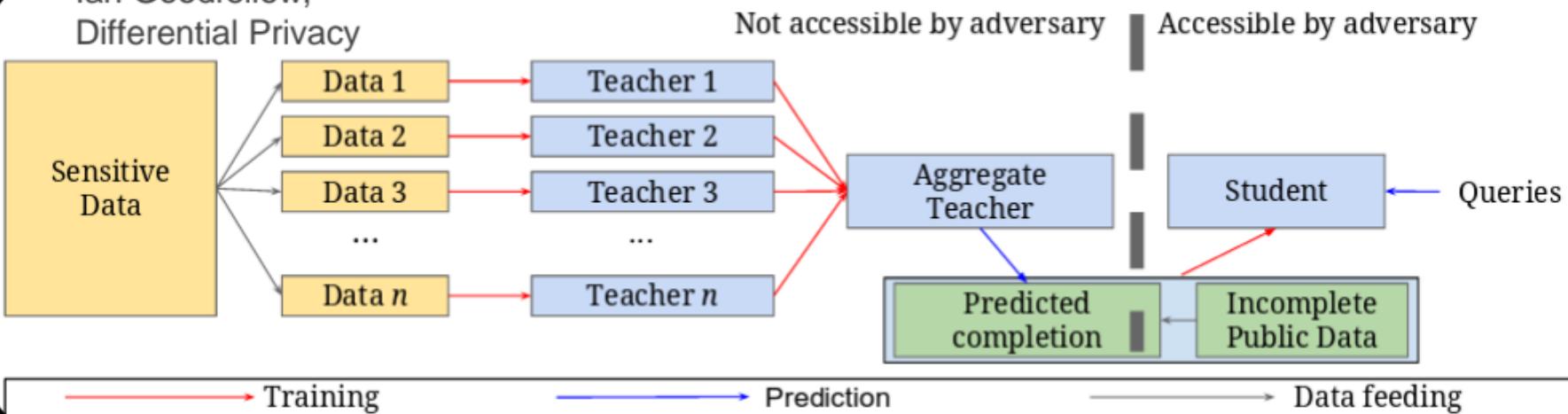
# PACMANN Research:

## Automated Machine Learning



# Differential Privacy

Ian Goodfellow,  
Differential Privacy



Silabus



FMIPA

# MACHINE LEARNING

## For Beginners

Tools:

- Python
- Scikit Learn
- Your own laptop



FREE!

Introduction to ML

KNN & Naive Bayes

Lasso & Ridge Regression

Decision Tree

Study Cases

Python Scikit Learn

Bias-Variance & Resampling

Logistic Regression

Bagging & Random Forest

Intuition of the Models

Linear Regression

Support Vector Machines

Boosting

Advice in Learning ML

30<sup>th</sup> Jan until 3<sup>rd</sup> Feb 2017

09.00-16.00 at D108

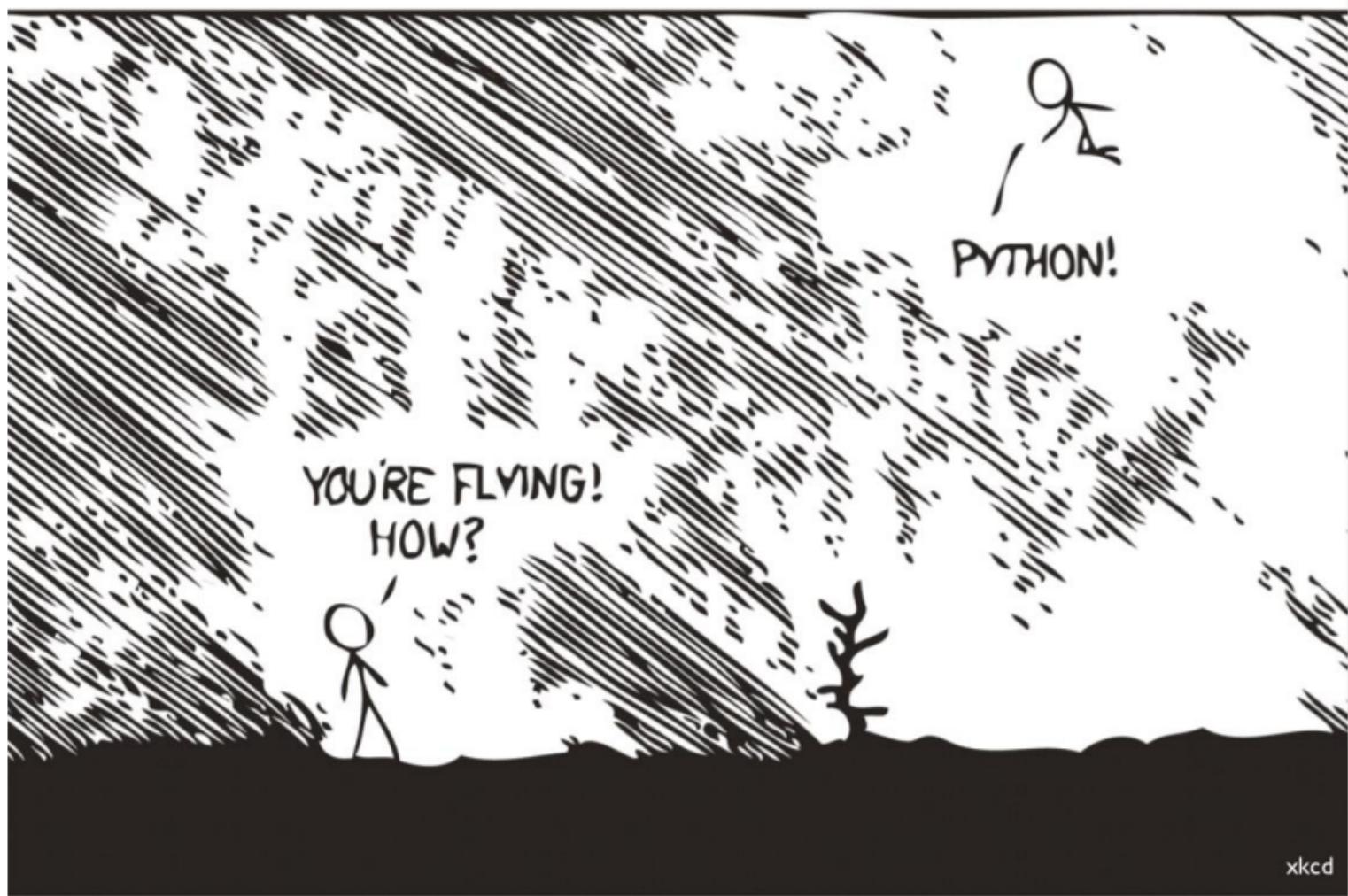
Department of Mathematics FMIPA UI

Registration:

[bit.ly/ML\\_Free\\_Training](http://bit.ly/ML_Free_Training)

CP: 089620615729 (Dhafin)

# Tools

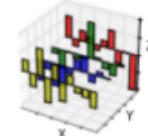
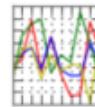
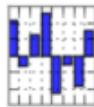


# Programming

IP[y]: IPython  
Interactive Computing

pandas

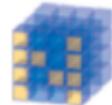
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Biltzstein, Data  
Sciences



NumPy



SciPy.org

Sponsored By  
ENTHOUGHT

matplotlib

# IPython Notebooks

<http://nbviewer.ipython.org/>

Home FAQ IPython Bookmarks

## IPython Notebook Viewer

A Simple way to share your IP[y]thon Notebook as Gists.

Share your own notebook, or browse others'

Enter a gist number or url  Go!

IP(y): Notebook 01 Documenting your Research Journey

File Edit View Insert Cell Kernel Help

Documenting your Research Journey

The purpose of this code is to show how IPython notebooks can be used to document your the GPU and the CPU. We compare the performance of each method using the system I document.

load image

```
In [1]: import PIL  
import PIL.Image  
  
image = PIL.Image.open("clique_barcode.png")  
image_array_rgb = numpy.array(image)  
  
x_original,y_original = numpy.split(image_array_rgb,  
y_original = numpy.concatenate([x_original,  
  
rgba_original = numpy.concatenate([x_original  
  
figsize(6,4)  
  
matplotlib.pyplot.imshow(rgba_original))  
matplotlib.pyplot.title("rgba_original")
```

Probabilistic Programming

Why would I want samples from the posterior, anyways?

We all deal with this question for the remainder of the book, and it is an unfortunate truth to say we can perhaps unwittingly avoid things. For now, let's finish with using posterior samples to answer the follow question: what is the expected number of texts at day 1,  $E[X_1]$ ? Recall that the expected value of a Poisson is equal to its parameter  $\lambda$ ; thus the question is equivalent to  $E[X_1]$ , or the expected value of  $X_1$  at time  $t=1$ .

In the code below, we are calculating the following. Let  $I$  index a posterior sample from the posterior distribution over  $\lambda$  for  $t=1$ . We calculate our  $E[X_1]$  for day 1 using  $\lambda_I = \sum_{i=1}^I \lambda_i$ , and we do  $N_I$ .

import numpy as np  
import numpy.random as npr  
  
# observed texts per day  
# Expected number of text messages received



Non Parametric Regression

Covariance function

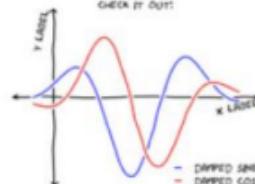
The behavior of individual metabolites from the GFP is governed by the covariance function. The hidden class of functions is a flexible choice.

```
In [1]: from __future__ import print_function  
  
# - Covariance function methods, RBF, Exponential, Matern, etc.  
# - Covariance function derivatives methods, RBF_Deriv1, RBF_Deriv2, Matern_Deriv1, Matern_Deriv2  
  
import os,sys  
os.chdir(os.path.dirname(os.path.abspath(__file__)))  
sys.path.append('..')  
  
import numpy as np  
from rbf import rbf
```

XKCD Plot With Matplotlib

Out [1]:

Check It Out!



Sometimes when showing schematic plots, this is the type of figure I want to display. But drawing it by hand is a bit of a pain. The problem is, matplotlib is a bit too precise. Attempting to duplicate this figure in matplotlib leads to:

Exploring R formula

Let's load that with a new design matrix

```
In [2]: # Using rpy2, please install rpy2 first! My Analysis 1
```



Thank you, question?