Big Data: Hope or Hype?

Professor Padhraic Smyth
Departments of Computer Science and Statistics
University of California, Irvine

ACM Orange County Chapter
University of California, Irvine

November 11th 2015

Outline of Tonight's Talk

- Big data: examples, characteristics, and applications
- Focus on predictive modeling
- Logistic models and deep learning
- Limitations of these approaches
- Concluding comments

Terminology

Large-scale Data Analysis

Data Mining

Data Science

Big Data

Machine Learning

Computational Statistics

……

Terminology

Large-scale Data Analysis

Data Mining

Data Science

Big Data

Machine Learning

Computational Statistics

……

……*Using computer algorithms to analyze data sets that are too large and/or complex for humans to work with*

Hype (or Bordering on it)

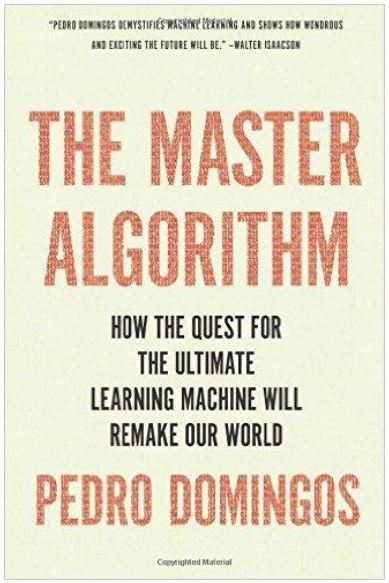
"*Big Data is the headache; deep learning is the solution*,"

VERGE 2015 Conference, Oct 26-29, from venture capitalist Steve Jurvetson, investor in companies such as SolarCity, Tesla Motors and Twitter.

Hype (or Bordering on it)

"*Big Data is the headache; deep learning is the solution*,"

VERGE 2015 Conference, Oct 26-29, from venture capitalist Steve Jurvetson, investor in companies such as SolarCity, Tesla Motors and Twitter.



"All knowledge-past, present, and future– can be derived from data by a single, universal learning algorithm."

"The Master Algorithm is the complete package. Applying it to vast amounts of patient and drug data….is how we will cure cancer."

A Revolution in Data Technology

Graphic from Ray Kurzweil, singularity.com

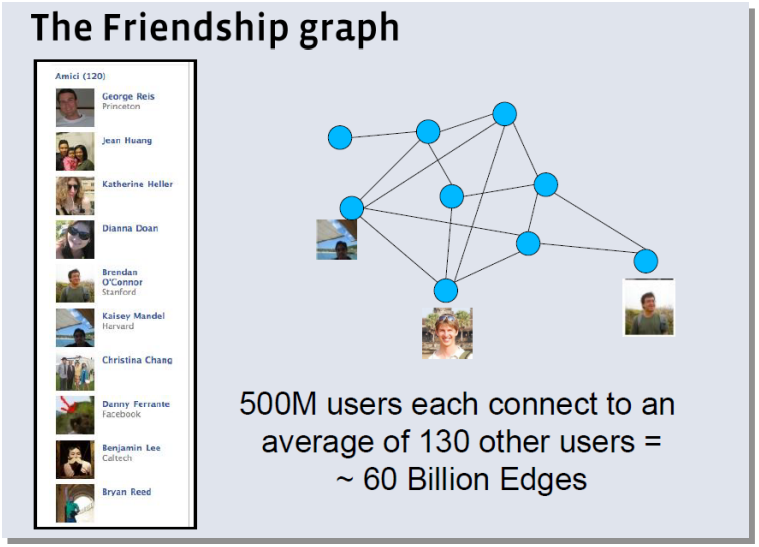
A Paradigm Shift in Data Analysis

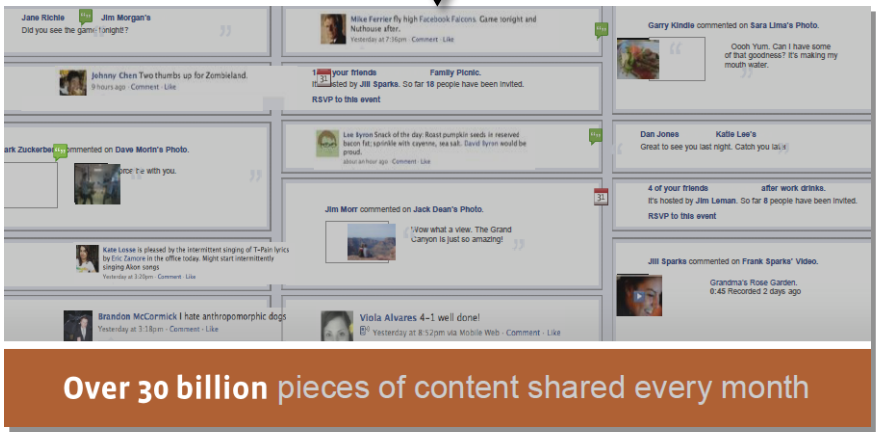
- Technological drivers
  - Sensors (cheap and ubiquitous, e.g., GPS on your phone)
  - Data storage (we are all "data owners")
  - Computational power
  - Data analysis methods (statistics and machine learning)
  - Internet and wireless communication (can collect and share data)

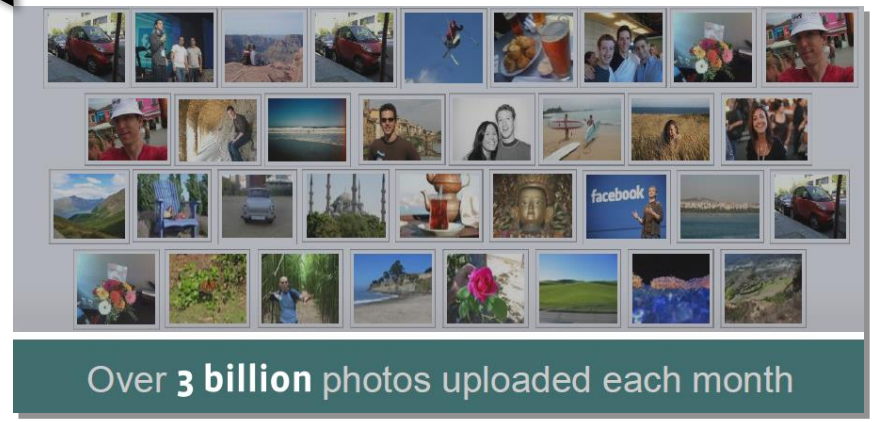
- Convergence…..tremendous demand for data analysis
  - In the sciences, in medicine, in engineering, in business, and more……

- In the past, this demand was met by statisticians
  - Does not scale up – there are way too few statisticians
  - And even statisticians need computers to analyze complex data

facebook







Figures from Lars Backstrom, Facebook, 2011

Example: Detecting Faces and Pedestrians in Images

Figures from Le Cun and Ranzato, ICML 2013 Tutorial

Particle Physics: Large Hadron Collider at CERN

700 Mbytes/second
60 Terabytes/day
20 Petabytes/year

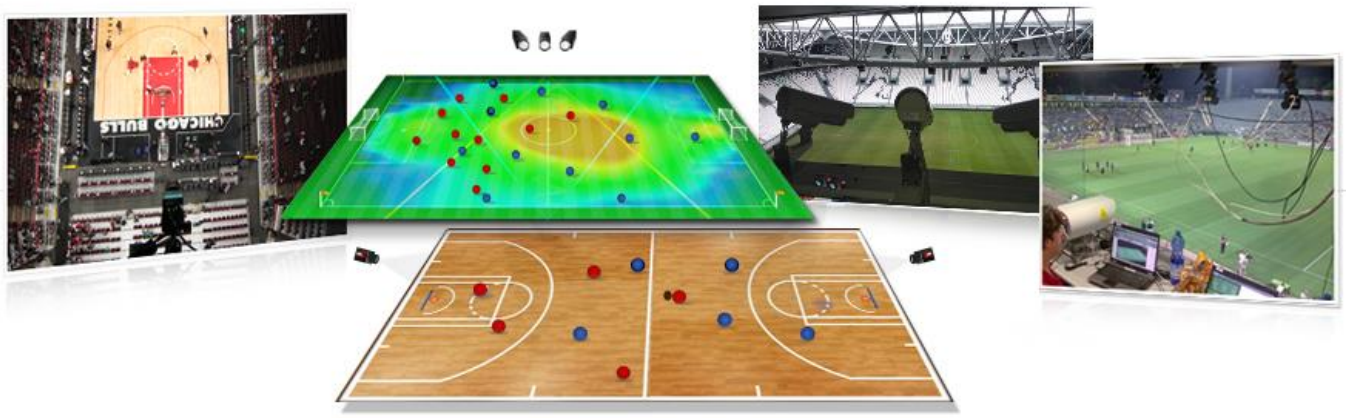
1 Terabyte = 10^{12} bytes
1 Petabyte = 10^{15} bytes

Detecting new types of particles = "Needle in a haystack"

New algorithms for searching massive amounts of data to find unusual patterns

Professor Daniel Whiteson
Department of Physics, UC Irvine

Real-Time Sports Statistics



Graphics from www.stats.com/sportvu/sportvu.asp

Inside the Secret World of the Data Crunchers Who Helped Obama Win

Data-driven decisionmaking played a huge role in creating a second term for the 44th President and will be one of the more closely studied elements of the 2012 cycle

By Michael Scherer @michaelscherer | Nov. 07, 2012 | 273 Comments

Daniel Shea for TIME

"The cave" at President Obama's campaign headquarters in Chicago

Daily Report: At WWDC, Apple Expected to Expand Into Health and Home Monitoring

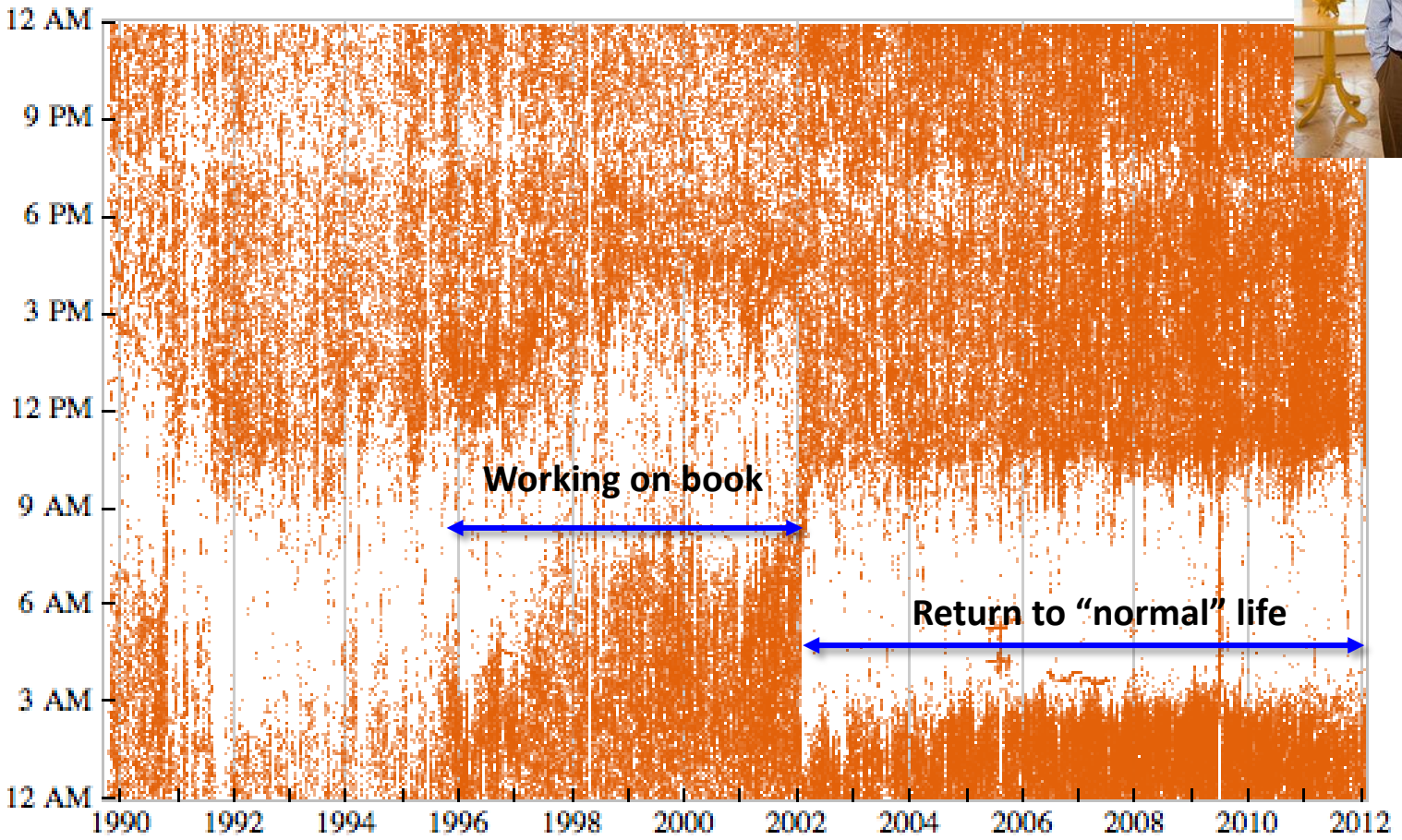
By THE NEW YORK TIMES JUNE 2, 2014 7:14 AM Comment

Apple is unlikely to introduce new devices this week, the things that most excite customers and investors these days. But the company is expected to dive deeper into two new areas: connected health and the so-called smart home, Brian X. Chen reports.

Along with operating system updates for mobile devices and desktop machines, Apple plans to introduce a new health-tracking app at its annual Worldwide Developers' Conference on Monday, according to a person briefed on the product, who spoke on the condition of anonymity because the plans were confidential. The app for mobile devices will track statistics for health or fitness, like a user's footsteps, heart rate and sleep activity.

Email Data

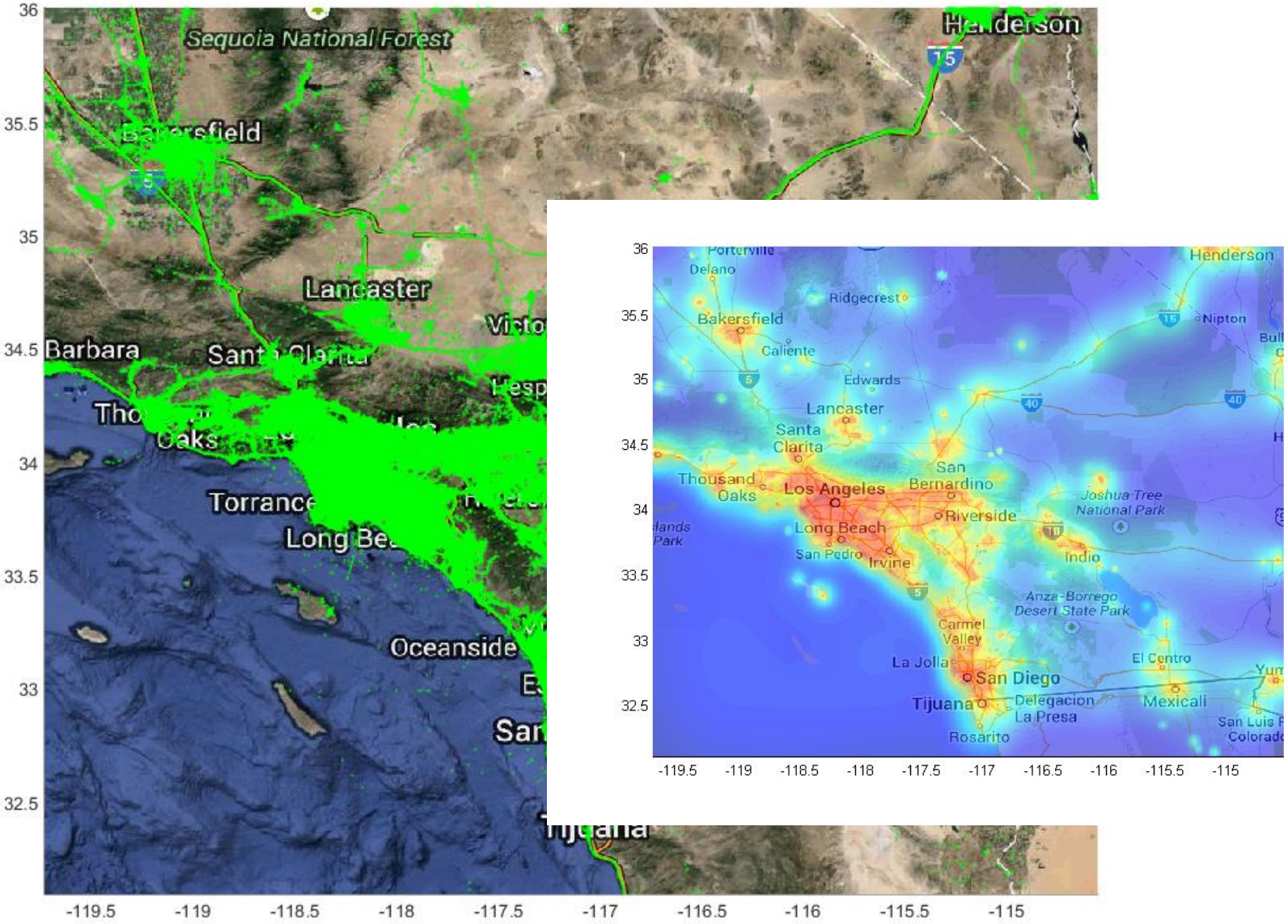
Time plot of 1/3 million emails sent by Stephen Wolfram over 20 years

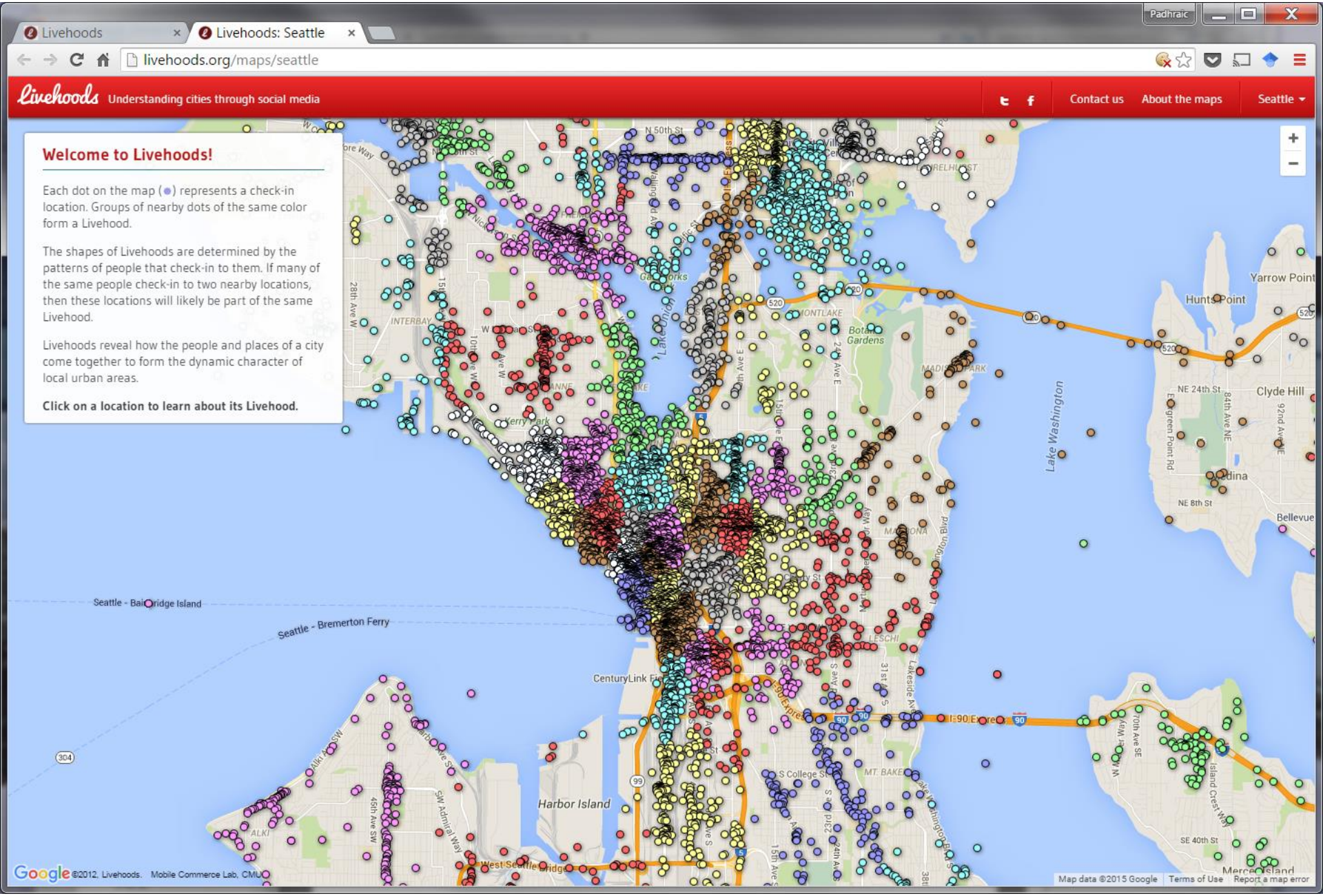


Figures from *The Personal Analytics of My Life*
blog.stephenwolfram.com, March 2012

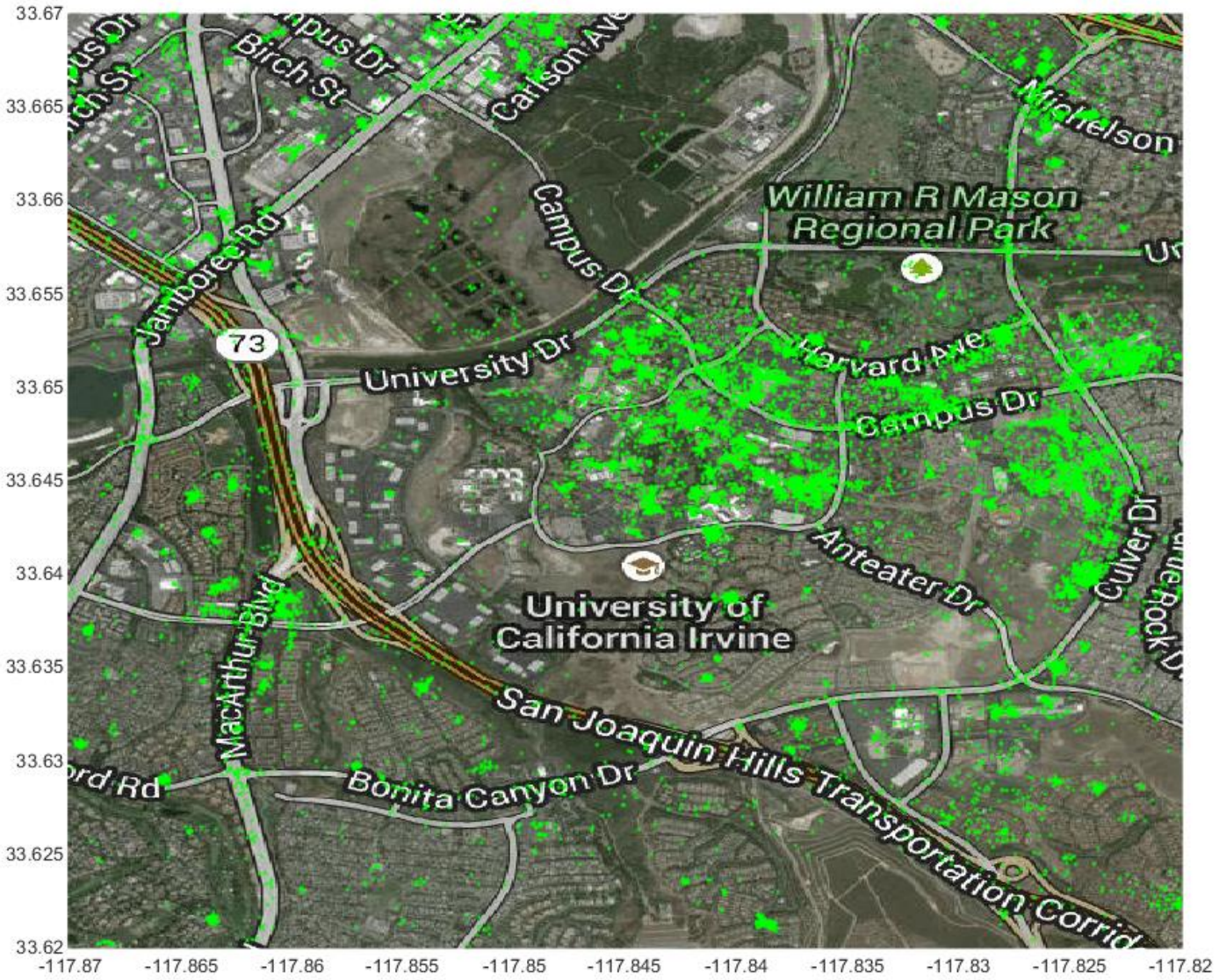
Modeling Human Behavior using Social Media

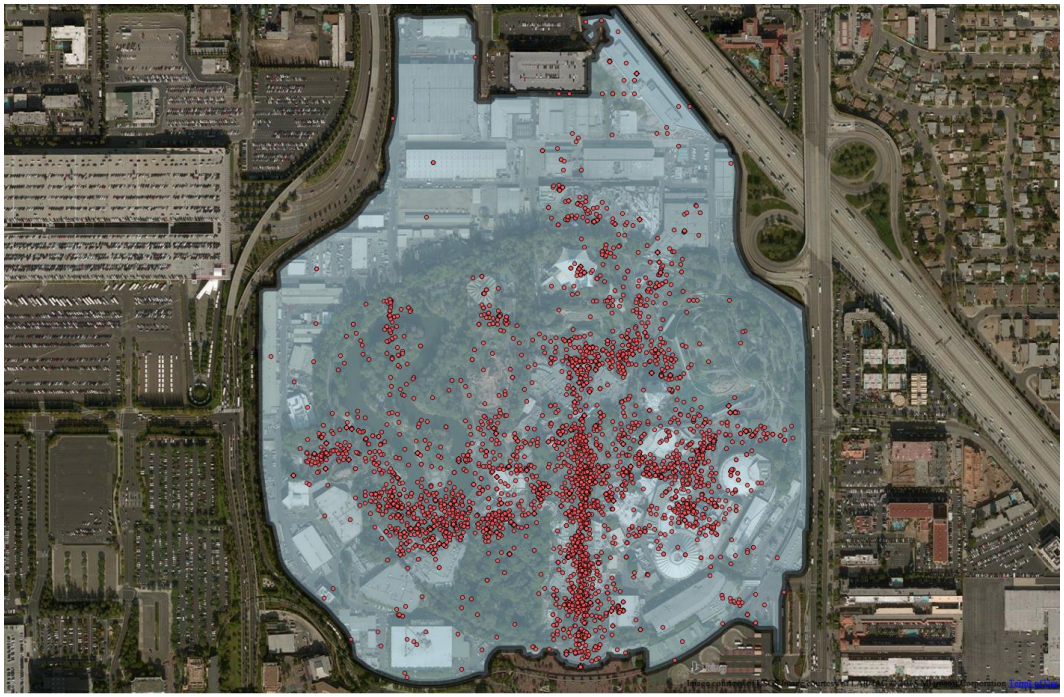
From Lichman and Smyth, ACM SIGKDD 2014



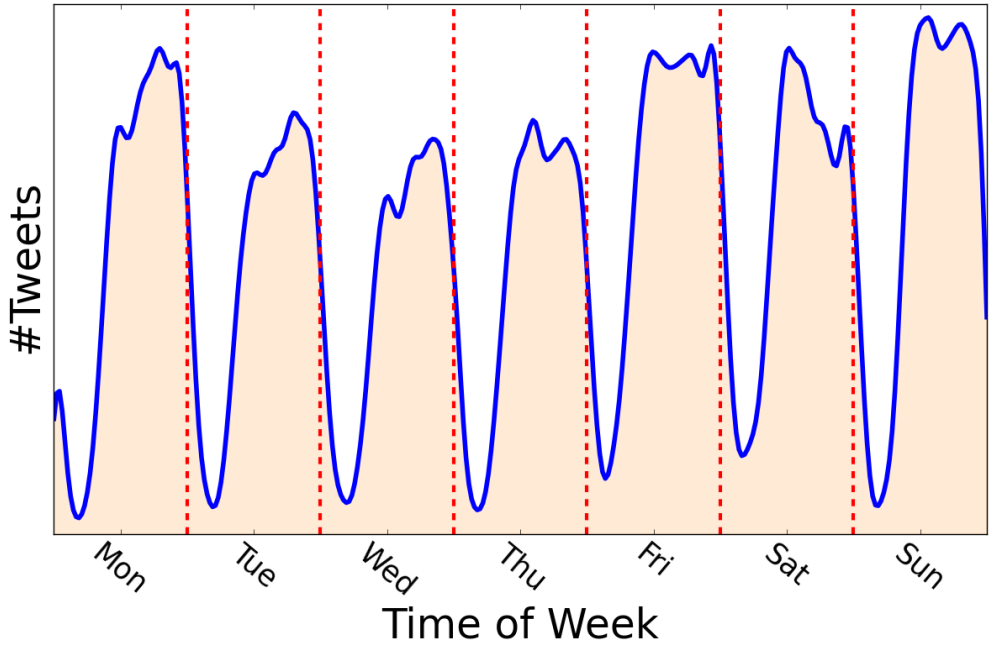

Padhraic
Livehoods
Livehoods: Seattle
livehoods.org/maps/seattle
Livehoods
Understanding cities through social media
Contact us
About the maps
Seattle
Welcome to Livehoods!
Each dot on the map (●) represents a check-in location. Groups of nearby dots of the same color form a Livehood.
The shapes of Livehoods are determined by the patterns of people that check-in to them. If many of the same people check-in to two nearby locations, then these locations will likely be part of the same Livehood.
Livehoods reveal how the people and places of a city come together to form the dynamic character of local urban areas.
Click on a location to learn about its Livehood.
Seattle - Bainbridge Island
Seattle - Bremerton Ferry
Harbor Island
Lake Washington
Yarrow Point
Hunts Point
Clyde Hill
Bellevue
Google ©2012, Livehoods. Mobile Commerce Lab, CMU
Map data ©2015 Google
Terms of Use
Report a map error

Geolocated Tweets around UC Irvine





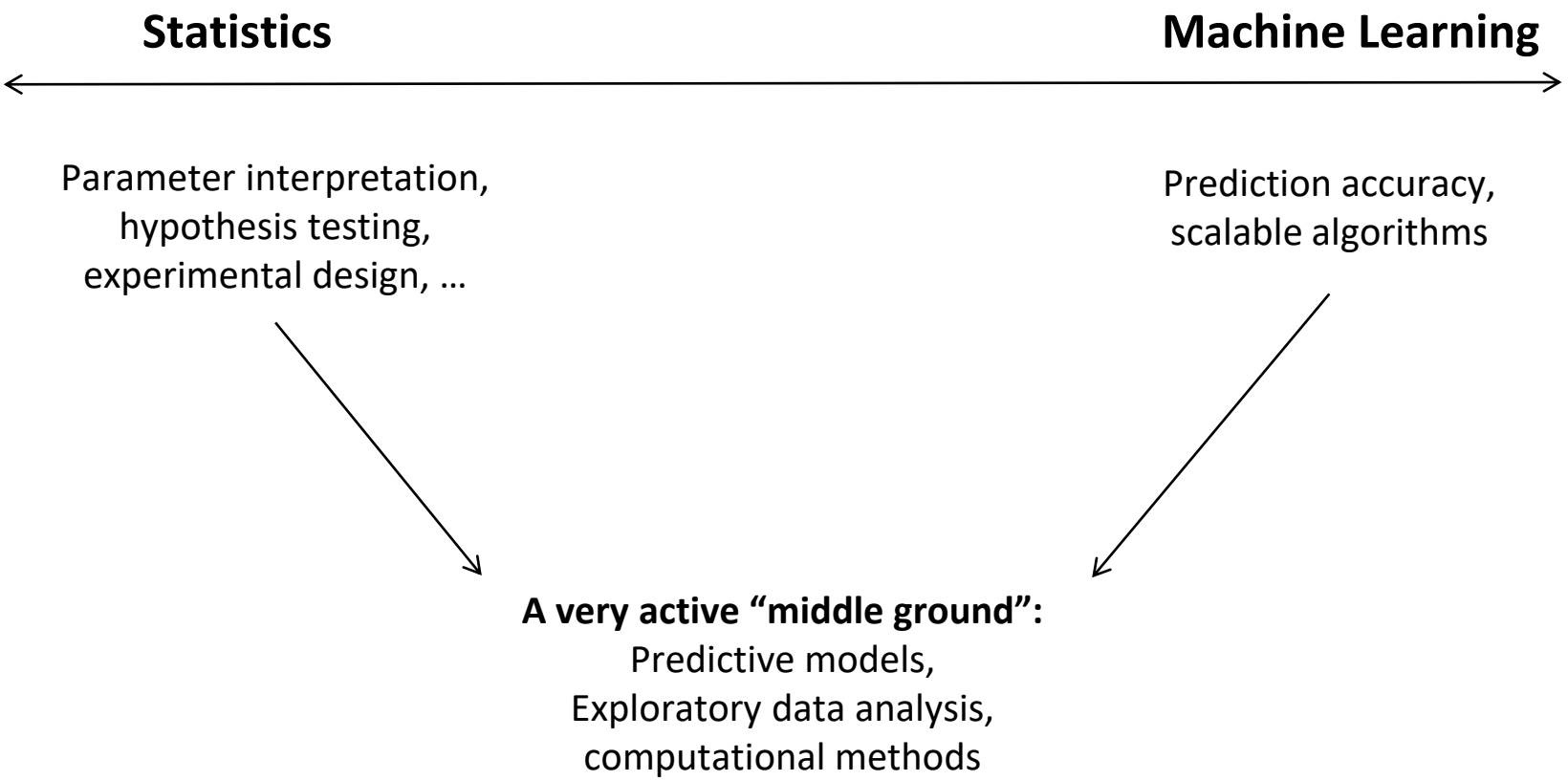
Spatial and Temporal Characteristics of Tweets at Disneyland



Typical Challenges with "Large Data"

- Observational/secondary
  - Collected for some other purposes, e.g., from social media
- Noisy, Biased
  - Measurement mechanisms are often unclear, subject to whims of data owners
- Size
  - Size brings complexity: in data management, in interactive analysis, etc
  - "Big data" is often a combination of many very small data sets (e.g., personalization)
- Complex and Multisource
  - e.g., text data, location data, demographic data: poses challenge in analysis
- Non-Stationary
  - Changing over time: trends, seasonality, etc

The Data Analysis Spectrum



Statistics and Machine Learning

- Similar goals….but different emphases

- Statistics
  - Parameters and models are often of primary importance
    - e.g., which variables in a social science study are associated with behavior
  - Interpretation and understanding are very important

- Machine Learning
  - Predictive accuracy is often the primary goal
    - e.g,. Speech recognition, image recognition
  - Scalability and algorithmic efficiency also a high priority
  - Model interpretation and understanding are often secondary

- **Important point:**
  - **Machine learning methods (typically) have statistical foundations**

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Predictive Modeling

Prediction

Patient ID	Zipcode	Age	….	Test Score	Diagnosis
18261	92697	55		83	1
42356	92697	19		-99	1
00219	90001	35		77	0
83726	24351	0		65	0

Build a model that can predict this "target" variable given the values of all the other variables

Data where the target values are known is called "training data"

Prediction

Patient ID	Zipcode	Age	….	Test Score	Diagnosis
18261	92697	55		83	1
42356	92697	19		-99	1
00219	90001	35		77	0
83726	24351	0		65	0
12837	92697	40		70	??
72623	92697	32		44	??

We can then use the model to make predictions when target values are unknown

This is referred to as "test data"

Predictive Modeling Notation

Model: $f(\underline{x}; \underline{\theta})$ is our model, a scalar function of $\underline{x} = [x_1, x_2, \ldots, x_d]$

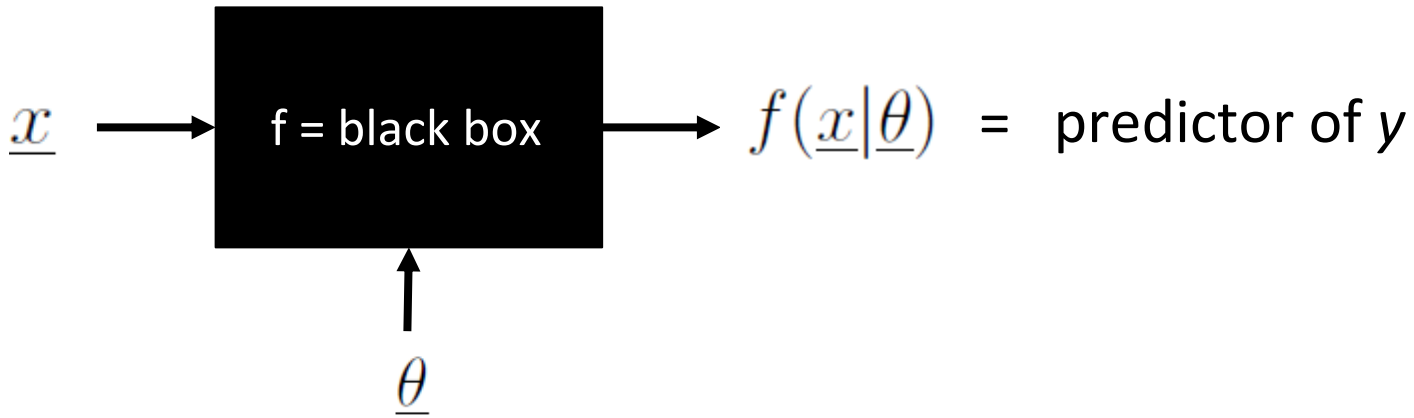
Parameters: $\underline{\theta}$ is a vector of unknown parameters that we want to estimate from data

Target or Output: y is what we want to predict, using f, given inputs $\underline{x}$

Regression: y's are real-valued

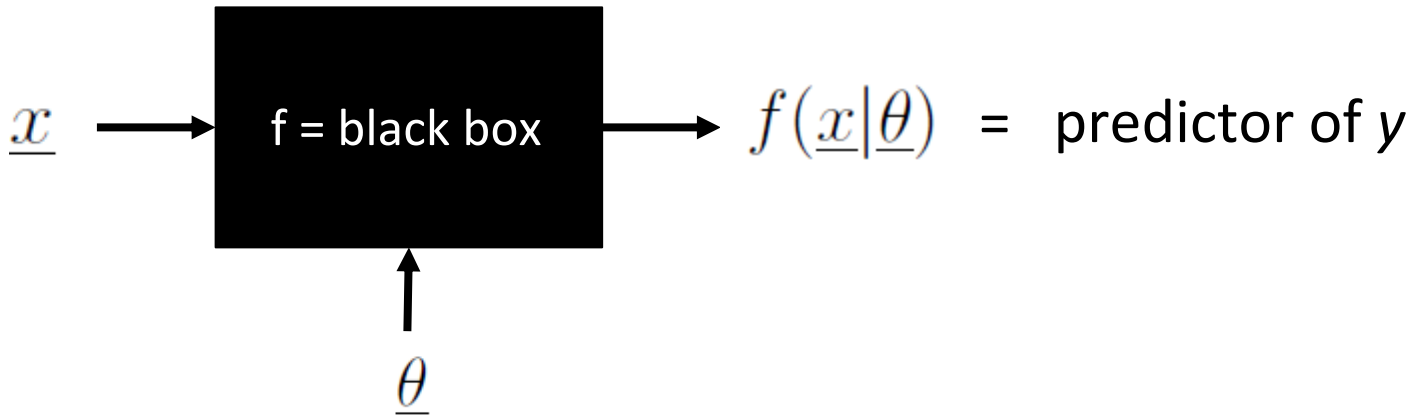
Classification: y's are categorical, e.g., $y \in \{0, 1\}$

Predictive Modeling



Goal is to use the training data to learn to predict unknown y's given new x's

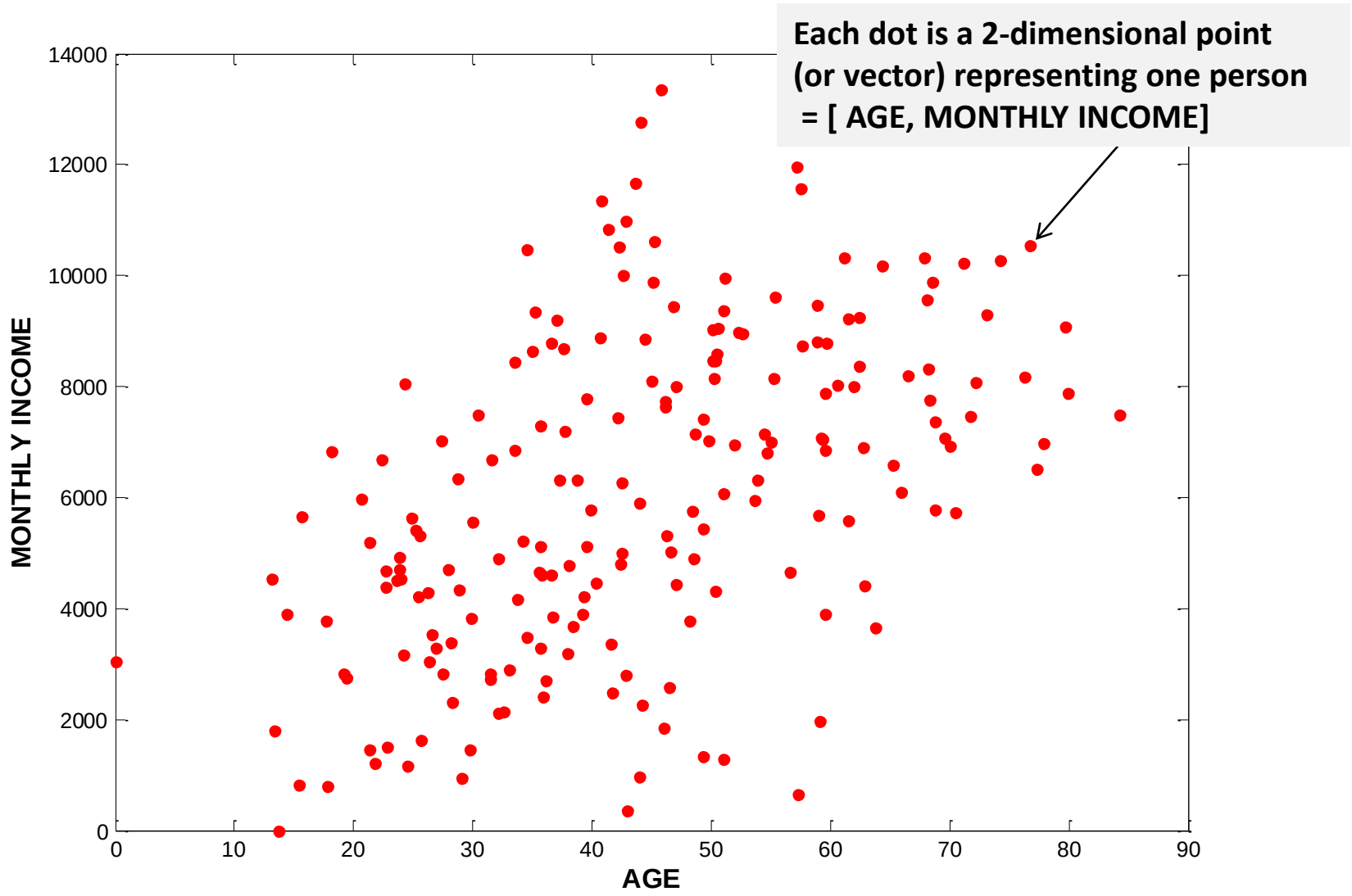
Predictive Modeling

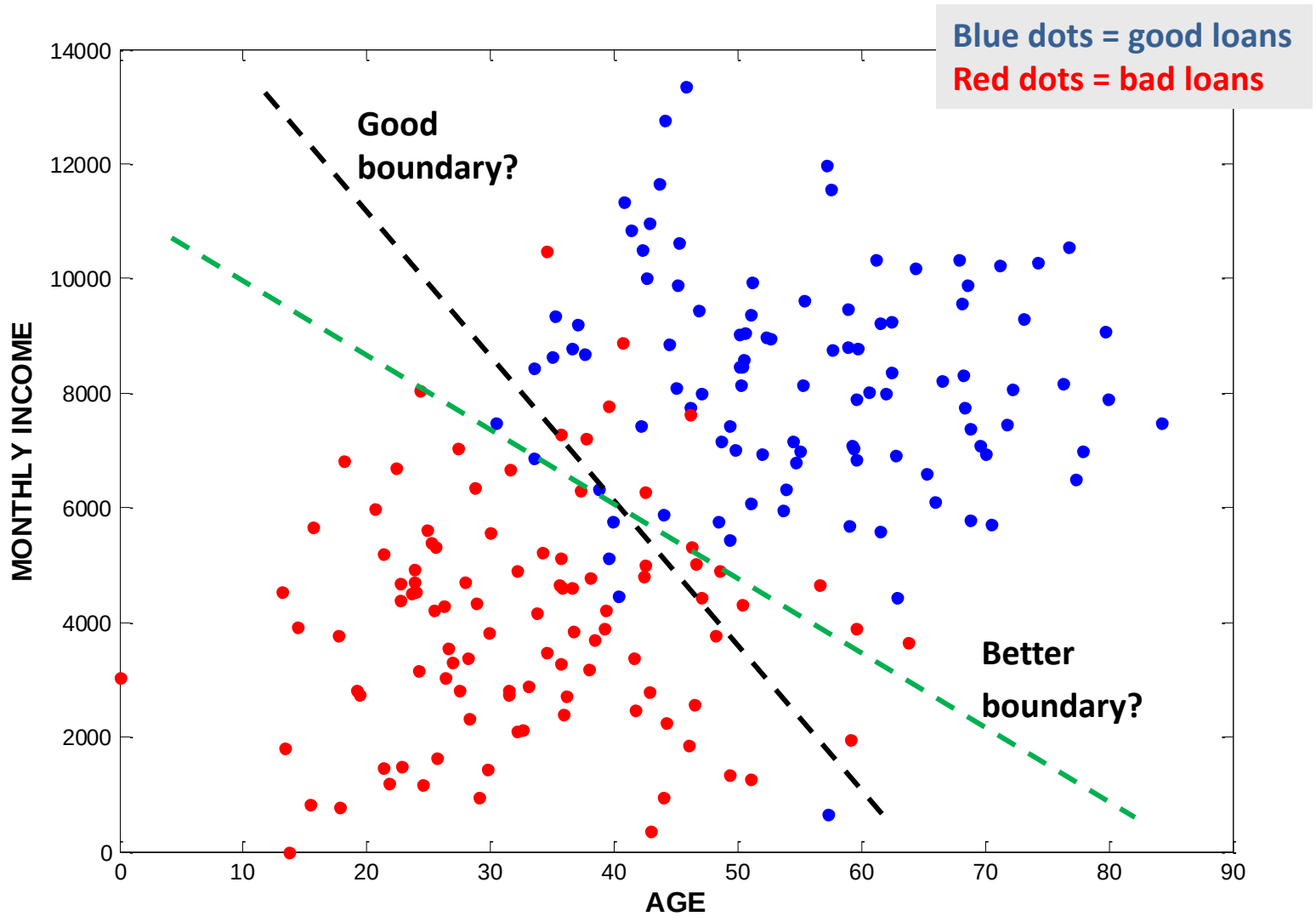


Goal is to use the training data to learn to predict unknown y's given new x's

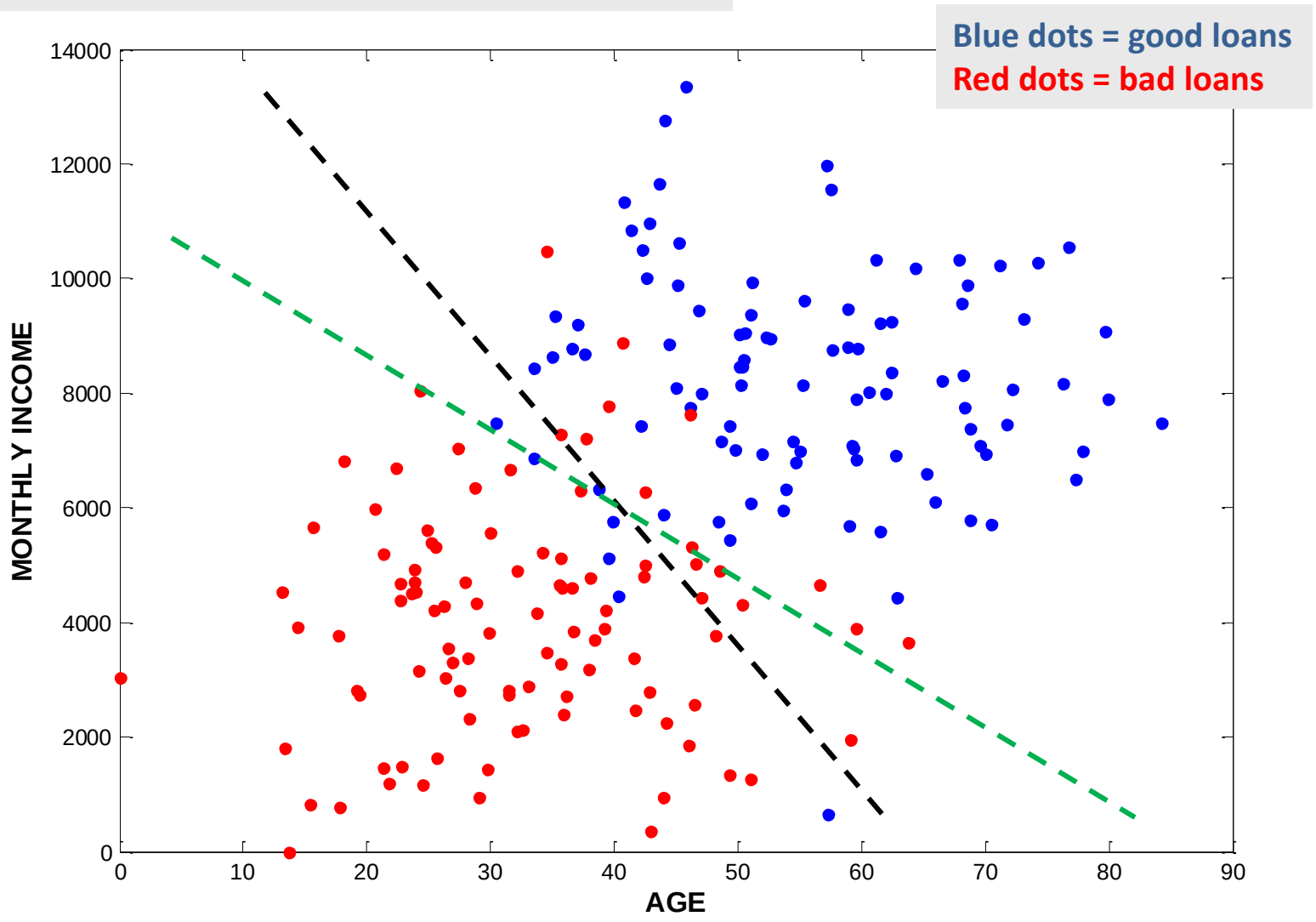
Many possible choices for *f*

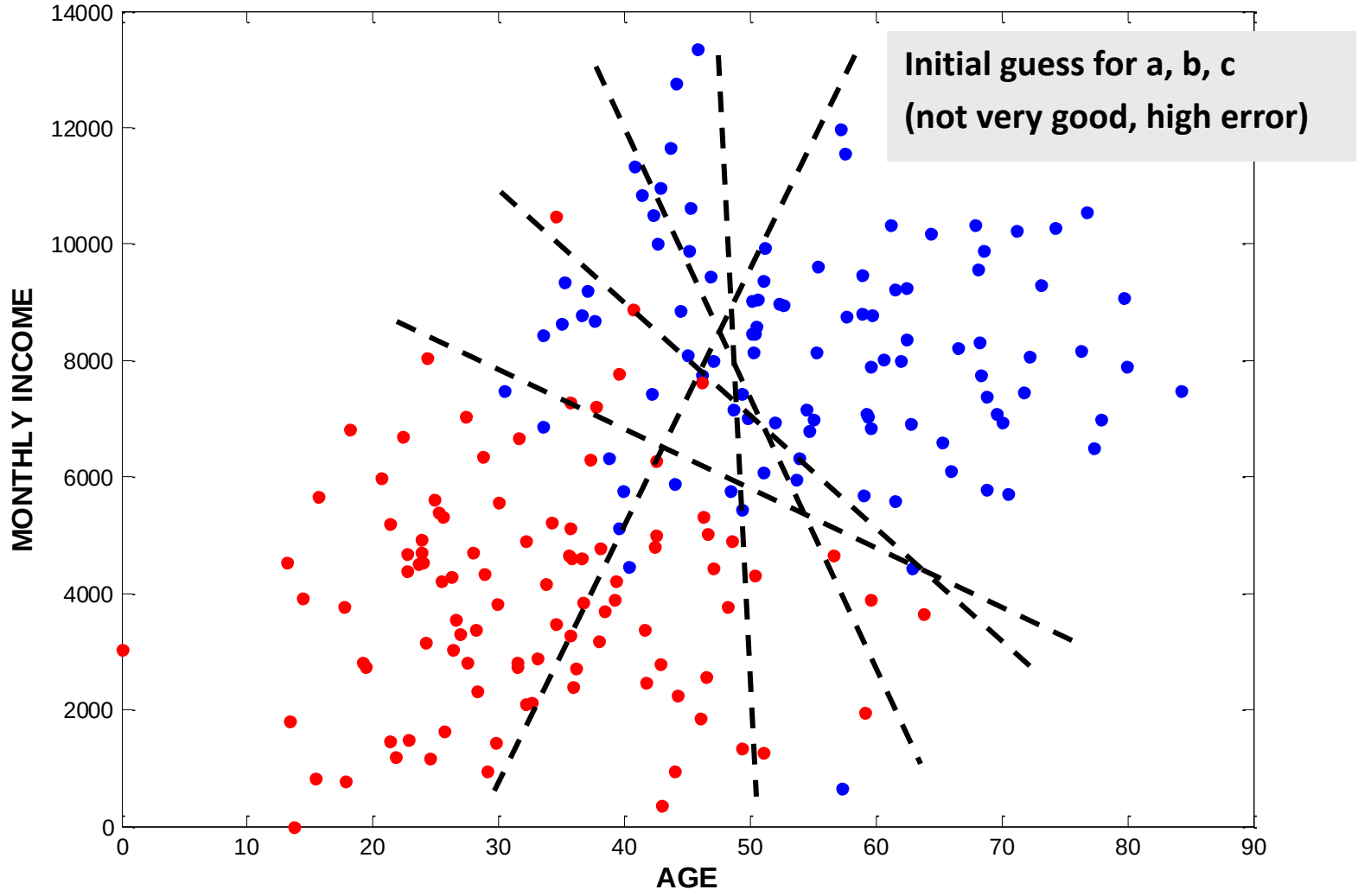
Important point: it is a function that maps *x* to *y*

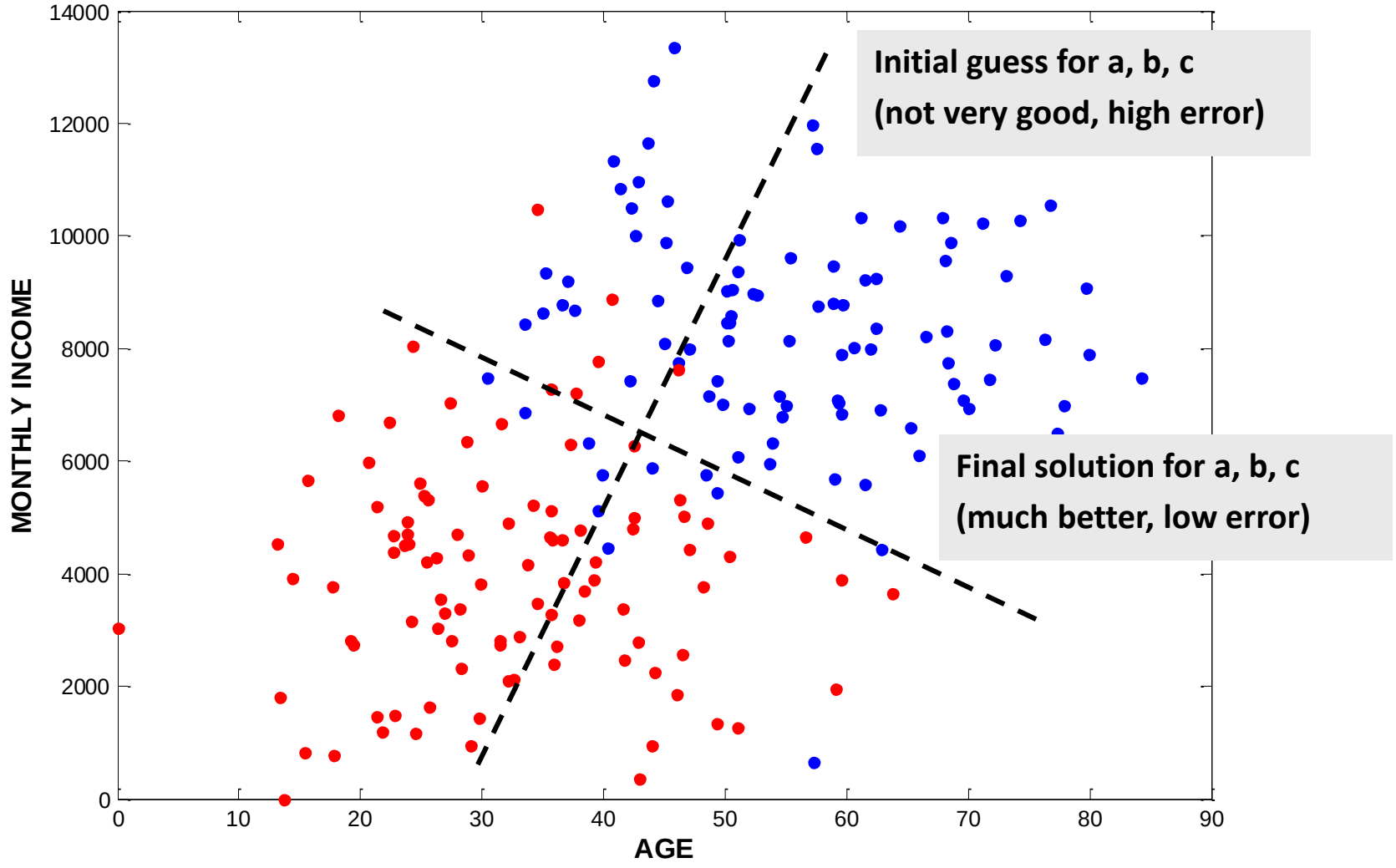

Each dot is a 2-dimensional point
(or vector) representing one person
= [AGE, MONTHLY INCOME]
MONTHL Y INCOME
AGE
0
2000
4000
6000
8000
10000
12000
14000
0
10
20
30
40
50
60
70
80
90

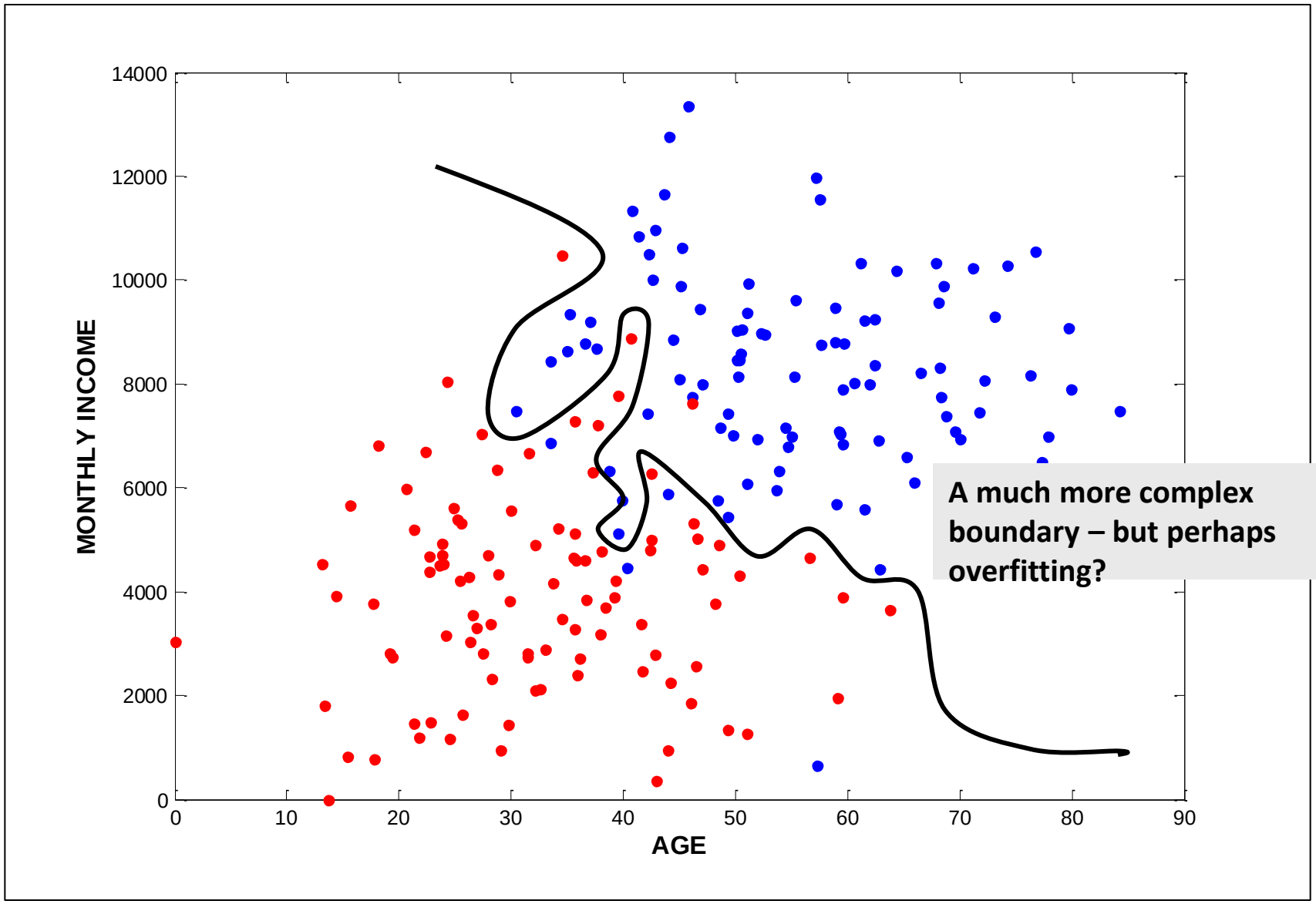

Blue dots = good loans
Red dots = bad loans
Good boundary?
Better boundary?
MONTHLY INCOME
AGE
14000
12000
10000
8000
6000
4000
2000
0
0
10
20
30
40
50
60
70
80
90

Model for decision boundary $f(x) = ax_1 + bx_2 + c$
Unknown parameters: a, b, c




Initial guess for a, b, c
(not very good, high error)
MONTHLY INCOME
AGE
0
2000
4000
6000
8000
10000
12000
14000
0
10
20
30
40
50
60
70
80
90


Initial guess for a, b, c
(not very good, high error)
Final solution for a, b, c
(much better, low error)
MONTHLY INCOME
AGE
0
2000
4000
6000
8000
10000
12000
14000
0
10
20
30
40
50
60
70
80
90

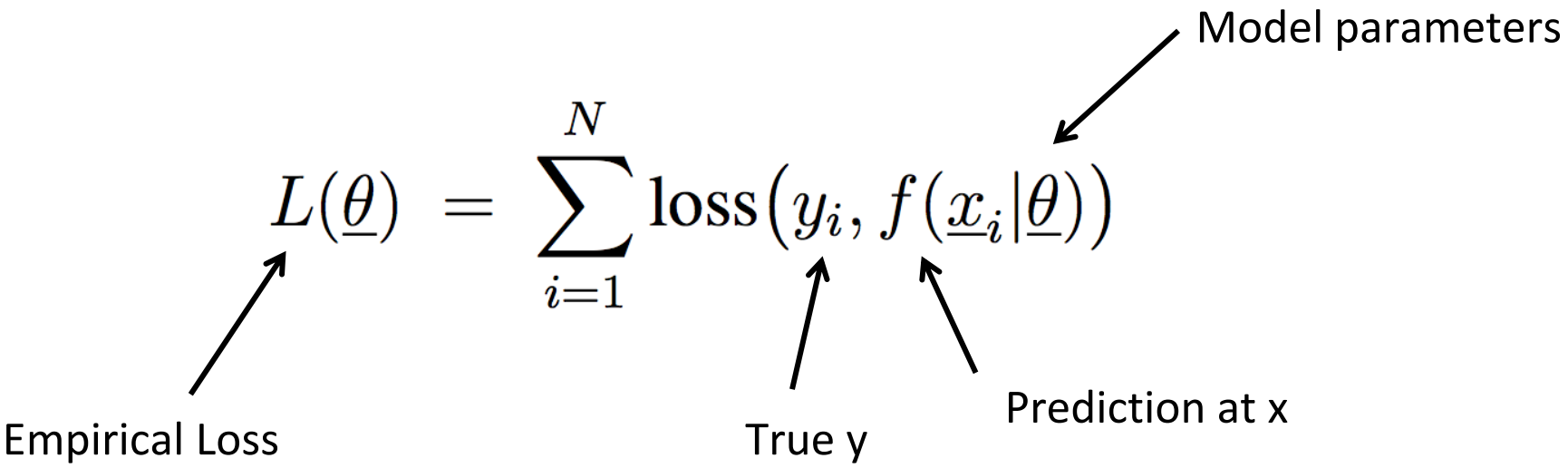

MONTHLY INCOME
AGE
0
2000
4000
6000
8000
10000
12000
14000
0
10
20
30
40
50
60
70
80
90
A much more complex boundary – but perhaps overfitting?

Empirical Learning

Training Data = $\{\underline{x}_i, y_i\}$

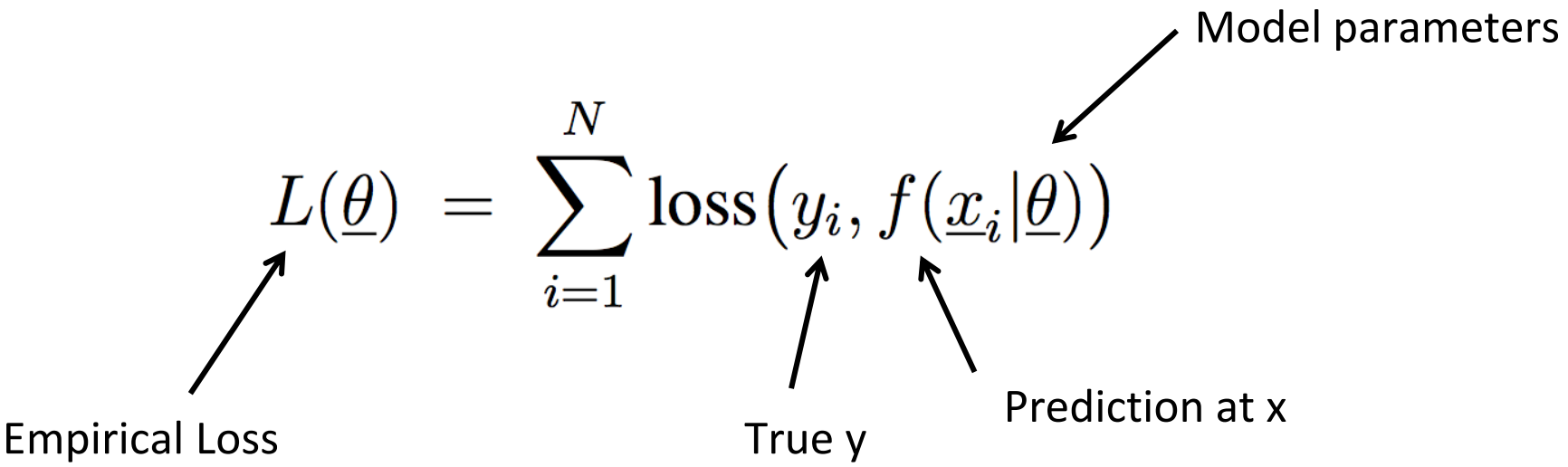
Empirical Learning

Training Data = $\{\underline{x}_i, y_i\}$



Empirical Learning

Training Data = $\{\underline{x}_i, y_i\}$



Empirical learning:
Find the parameters that minimize the empirical loss on the training data

Key point: "Learning" = Optimization

Key Components of any Predictive Modeling Algorithm

1. Prediction Model f:

   What functional form should we choose for f?

2. Loss function

   How do we compare f's predictions to y?

3. Optimization

   Given f and a loss function, how can we learn f's parameters

Every predictive learning algorithm is an instantiation of these 3 components

Key Components of any Predictive Modeling Algorithm

1. Prediction Model f:

   What functional form should we choose for f?

2. Loss function

   How do we compare f's predictions to y?

3. Optimization

   Given f and a loss function, how can we learn f's parameters

Examples of Prediction Models

Linear Regression

$$
\begin{aligned}
f(\underline{x}|\theta) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d \\
&= \sum_{j=0}^{d} \theta_j x_j = \underline{\theta}^T \underline{x}
\end{aligned}
$$

Examples of Prediction Models

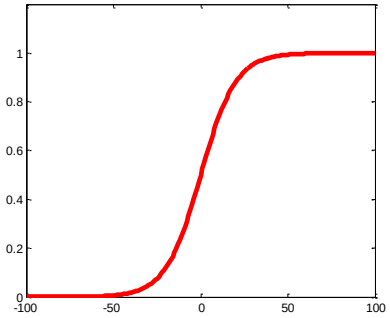
Linear Regression

$$f(\underline{x}|\theta) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d$$

$$= \sum_{j=0}^{d} \theta_j x_j = \underline{\theta}^T \underline{x}$$

Logistic Regression

$$f(\underline{x}|\theta) = \frac{1}{1 + e^{-z}}, \quad z = \underline{\theta}^T \underline{x}$$



Examples of Prediction Models

Logistic Regression

$$f(\underline{x}|\theta) = \frac{1}{1+e^{-z}}, \quad z = \underline{\theta}^T \underline{x}$$

Neural Network

$$f(\underline{x}|\theta) = g\left(\sum_{k=1}^{H} \theta_k h_k(\underline{x})\right) \quad \text{where } h_k = \frac{1}{1+e^{-z_k}}, \quad z_k = \sum_{j=0}^{d} \theta_{kj} x_j$$

Key Components of any Predictive Modeling Algorithm

1. Prediction Model f:

   What functional form should we choose for f?

2. Loss function

   How do we compare f's predictions to y?

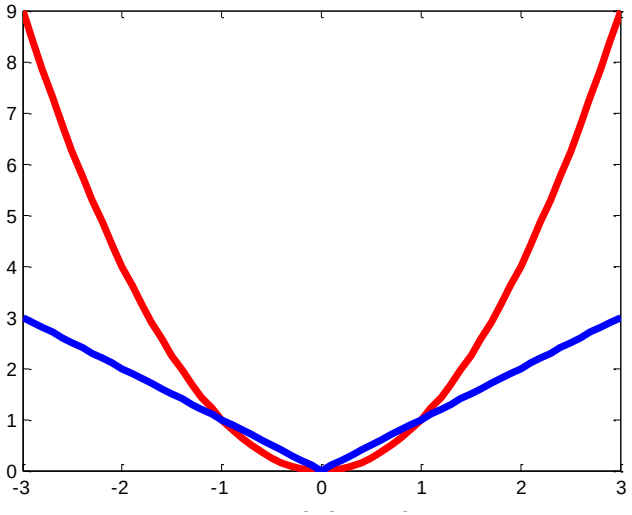
3. Optimization

   Given f and a loss function, how can we learn f's parameters

Examples of Loss Functions

Squared Loss

$$L(\underline{\theta}) = \sum_{i=1}^{N} \left(y_i - f(\underline{x}_i|\underline{\theta}) \right)^2$$



Absolute Loss

$$L(\underline{\theta}) = \sum_{i=1}^{N} \left| y_i - f(\underline{x}_i|\underline{\theta}) \right|$$

The Log Loss Function

$$L(\underline{\theta}) = \sum_{i=1}^{N} \log f(y_i | \underline{x}_i, \underline{\theta})$$

Here f is defined as the probability for y given the model

We want to maximize the sum of log probabilities

The Log Loss Function

$$L(\underline{\theta}) = \sum_{i=1}^{N} \log f(y_i | \underline{x}_i, \underline{\theta})$$

Here f is defined as the probability for y given the model

We want to maximize the sum of log probabilities

Same as maximizing the product of probabilities of individual data points

A key idea in statistical modeling: "maximum likelihood" (Fisher, 1928)

Empirical loss functions can often be cast as "log likelihoods"

Key Components of any Predictive Modeling Algorithm

1. Prediction Model f:

   What functional form should we choose for f?

2. Loss function

   How do we compare f's predictions to y?

3. Optimization

   Given f and a loss function, how can we learn f's parameters

Optimization: The Heart of Machine Learning

- We want to find θ that minimizes $L(\theta)$
- Note that θ may be a very high-dimensional vector…..this amounts to finding the extremum of a function in very high-dimensions
- Typically little is known about the shape of $L(\theta)$ in advance since it depends on the data to hand
- Local iterative search is widely used
  - Start at some random guess for θ
  - Move downhill on the surface of $L(\theta)$ using a locally-computed heuristic
  - Can use local information about the shape of the surface

Example: Language Modeling

Learning models for P(word $_t$ | word $_{t-1}$,.....)

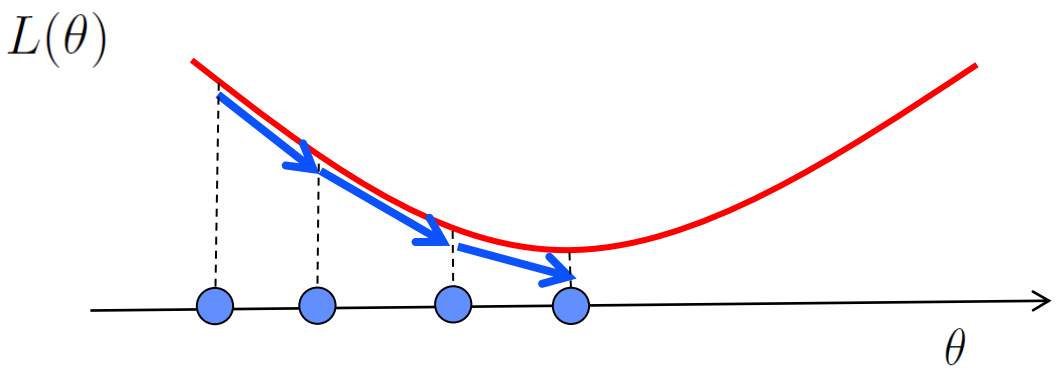
From Chelba et al, 2013 (Google)

Model	Num. Params [billions]	Training Time [hours]	[CPUs]	Perplexity
Interpolated KN 5-gram, 1.1B n-grams (KN)	1.76	3	100	67.6
Katz 5-gram, 1.1B n-grams	1.74	2	100	79.9
Stupid Backoff 5-gram (SBO)	1.13	0.4	200	87.9
Interpolated KN 5-gram, 15M n-grams	0.03	3	100	243.2
Katz 5-gram, 15M n-grams	0.03	2	100	127.5
Binary MaxEnt 5-gram (n-gram features)	1.13	1	5000	115.4
Binary MaxEnt 5-gram (n-gram + skip-1 features)	1.8	1.25	5000	107.1
Hierarchical Softmax MaxEnt 4-gram (HME)	6	3	1	101.3
Recurrent NN-256 + MaxEnt 9-gram	20	60	24	58.3
Recurrent NN-512 + MaxEnt 9-gram	20	120	24	54.5
Recurrent NN-1024 + MaxEnt 9-gram	20	240	24	51.3

Table 1: Results on the 1B Word Benchmark test set with various types of language models.

Lower perplexity scores indicate better predictions

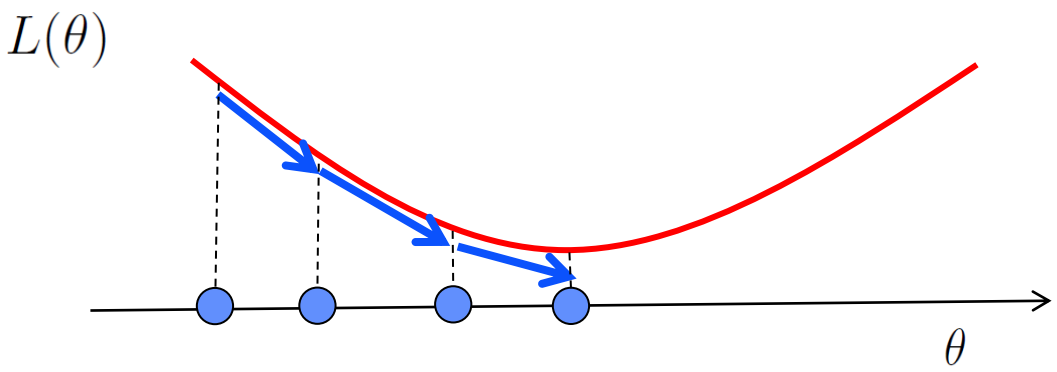
Optimization



Easy (convex)

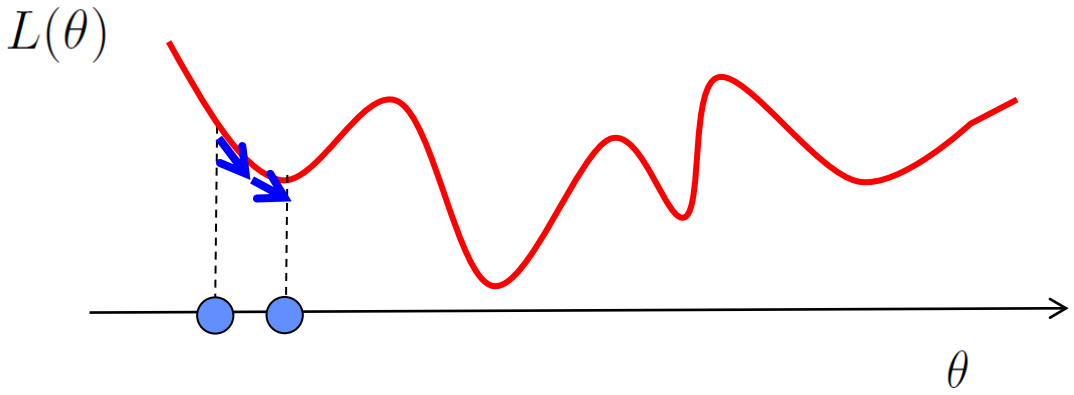
Example: Logistic regression

Optimization



Easy (convex)

Example: Logistic regression



Hard (non-convex)

Example: Neural network

Learning with Gradient Descent

New location in parameter space

Current location in parameter space

$$\theta^{\text{new}} = \theta^{\text{old}} - \lambda \nabla L(\theta^{\text{old}})$$

Learning with Gradient Descent

New location in parameter space

Current location in parameter space

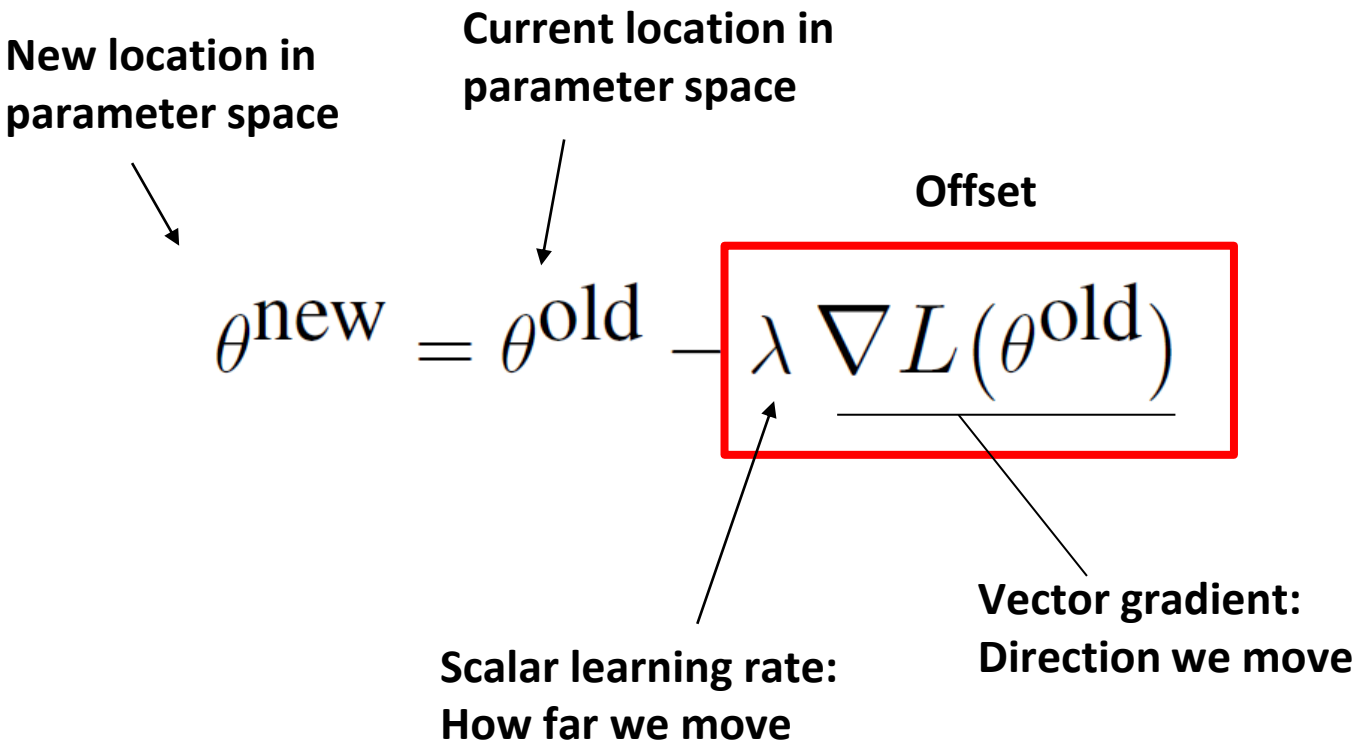
Offset

$$\theta^{\text{new}} = \theta^{\text{old}} - \lambda \nabla L(\theta^{\text{old}})$$

Scalar learning rate: How far we move

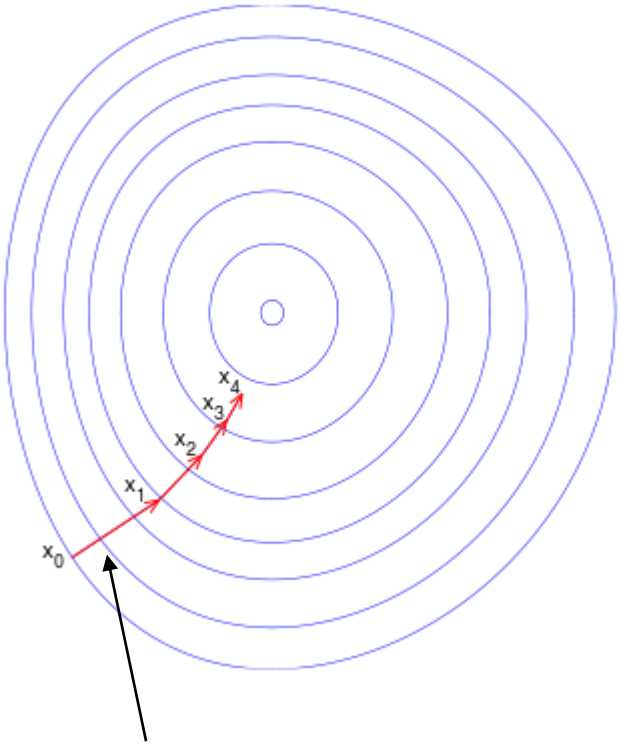
Vector gradient: Direction we move

Learning with Gradient Descent

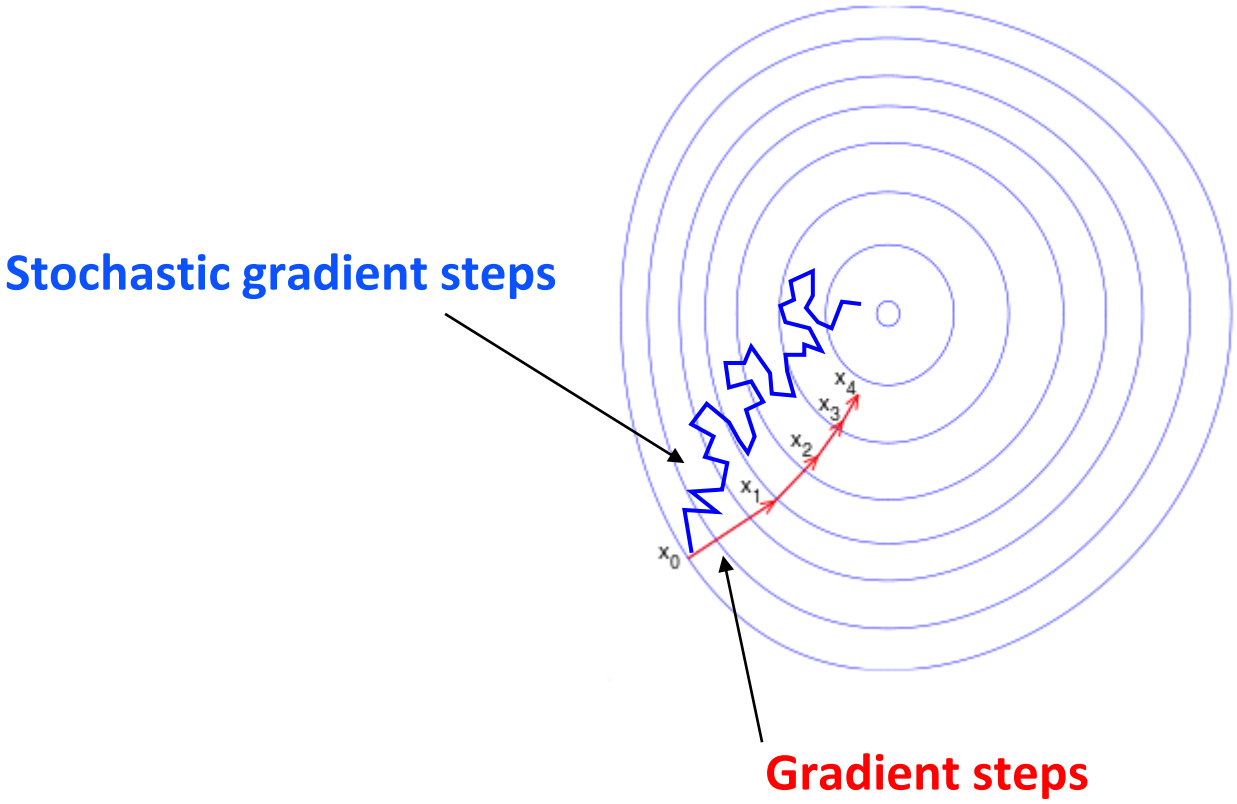


Theory: if learning rate is small enough we will converge to a (local) minimum

Gradient Descent in 2d Parameter Space

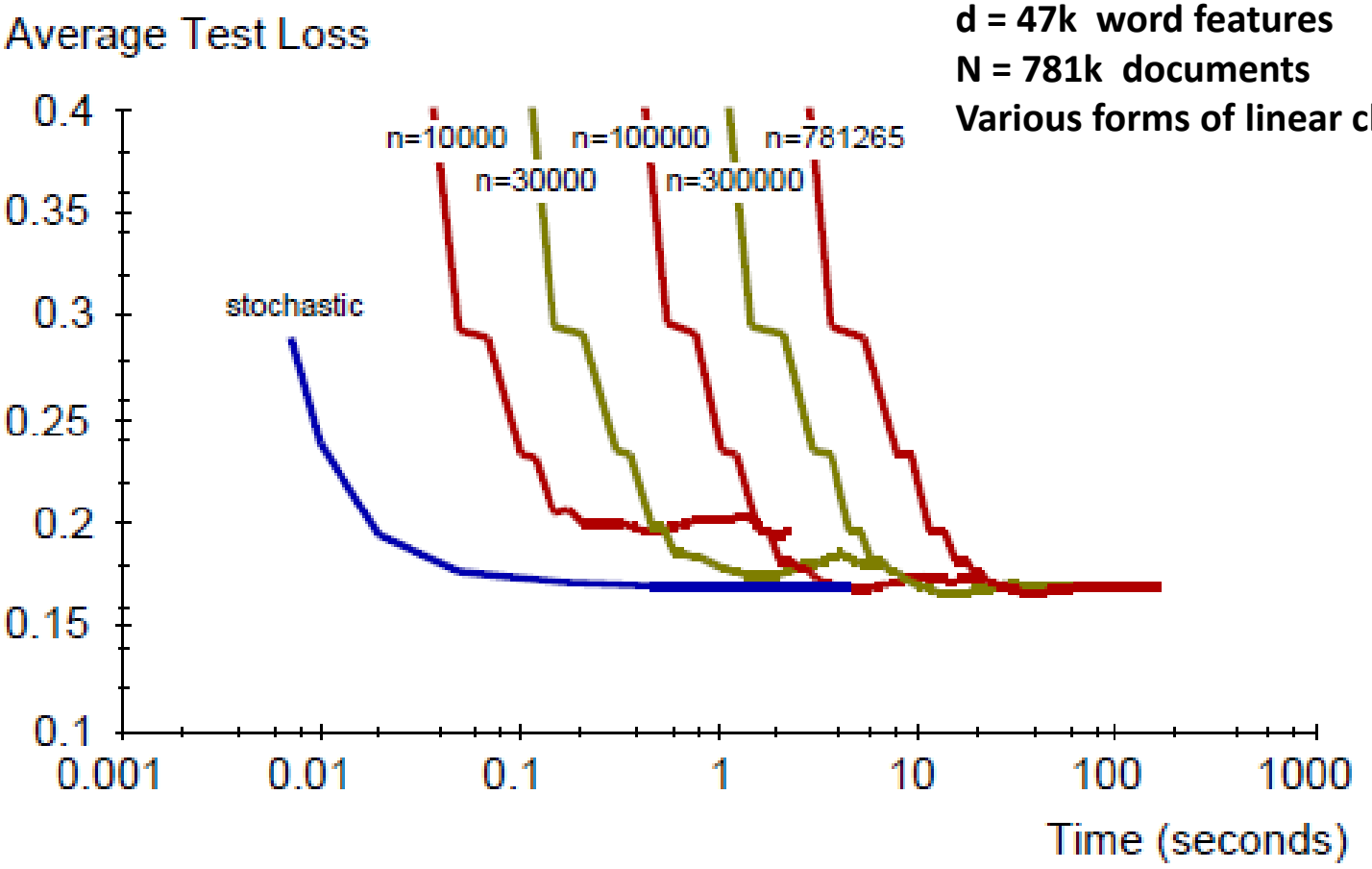


Gradient Descent in 2d Parameter Space



Stochastic Gradient Optimization

Binary Text Classification Problem
d = 47k word features
N = 781k documents
Various forms of linear classifiers



From Leon Bottou, Stochastic Gradient Learning, MLSS Summer School 2011, Purdue University, http://learning.stat.purdue.edu/mlss/_media/mlss/bottou.pdf

Success Stories

- When is machine learning most successful?
  - Large amounts of labeled data to train on
  - Prediction is more important than interpretation
  - Future data is similar to training data, i.e., data characteristics are not changing
  - Lack of theory, i.e., no strong first principles models available

- Examples
  - Speech recognition
  - Auto-completion for text
  - Face recognition
  - Document classification
  - Information extraction
  - ……and many more applications

?

?

The New York Times

?

TURMOIL IN UKRAINE

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEE MYERS
59 minutes ago

President Vladimir V. Putin's first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

Uriel Sinai for The New York Times

Ukrainian riot police officers stood guard at an anti-Russian rally in Donetsk on Tuesday.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Crimea's Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSZENHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

RELATED COVERAGE

- **Kerry Takes Offer of Aid to Ukraine** 33 minutes ago
- **Cyberattacks Rise as Crisis Spills to Internet** 8:47 PM ET
- VIDEO: **Confrontation in Crimea**

An Obama Budget Big on Ideals, but With Small Chances

By JACKIE CALMES 9:02 PM ET

President Obama sent

Some Who Fled Cuba Are Returning to Help

By DAMIEN CAVE 8:55 PM ET

Some members of the first Cuban families to leave after Fidel Castro took over are coming back, reuniting with the island and partnering with Cubans in direct new ways

The Opinion Pages

OP-ED CONTRIBUTOR

Has Privacy Become a Luxury Good?

By JULIA ANGWIN

It takes a lot of money and time to avoid hackers and data miners.

- **Editorial: Frustration With Afghanistan**
- **Brooks: Putin Can't Stop**
- **Cohen: Russia's Crimean Crime**

DRAFT

My Character to Kill

By ALEX BERENSON

I'm not sure I can say goodbye to a man who has defined my creative life for so long — and who will pay the mortgage for at least one more contract.

- **Op-Docs: 'Chinese, on the Inside'**

MARKETS » At 10:03 PM ET

JAPAN Nikkei	HangSeng	CHINA Shanghai
14,942.78	22,690.46	2,059.39
+221.30	+32.83	−12.09
+1.50%	+0.14%	−0.58%

Data delayed at least 15 minutes

Get Quotes | My Portfolios »

?

Each ? represents an "ad slot"

In a fraction of a second, predictive models predict which ads you are most likely to click on (from 1000's of ads)

U.S. INTERNATIONAL 中文网

The New York Times

Tuesday, March 4, 2014 | Today's Paper | Personalize Your Weather

TURMOIL IN UKRAINE

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEE MYERS
59 minutes ago

President Vladimir V. Putin's first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

Uriel Sinai for The New York Times

Ukrainian riot police officers stood guard at an anti-Russian rally in Donetsk on Tuesday.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Crimea's Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSZENHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

RELATED COVERAGE

- **Kerry Takes Offer of Aid to Ukraine** 33 minutes ago
- **Cyberattacks Rise as Crisis Spills to Internet** 8:47 PM ET
- VIDEO: **Confrontation in Crimea**

An Obama Budget Big on Ideals, but With Small Chances

By JACKIE CALMES 9:02 PM ET

President Obama sent

Some Who Fled Cuba Are Returning to Help

By DAMIEN CAVE 8:55 PM ET

Some members of the first Cuban families to leave after Fidel Castro took over are coming back, reuniting with the island and partnering with Cubans in direct new ways

The Opinion Pages

OP-ED CONTRIBUTOR

Has Privacy Become a Luxury Good?

By JULIA ANGWIN

It takes a lot of money and time to avoid hackers and data miners.

- **Editorial: Frustration With Afghanistan**
- **Brooks: Putin Can't Stop**
- **Cohen: Russia's Crimean Crime**

DRAFT

My Character to Kill

By ALEX BERENSON

I'm not sure I can say goodbye to a man who has defined my creative life for so long — and who will pay the mortgage for at least one more contract.

- **Op-Docs: 'Chinese, on the Inside'**

MARKETS »

At 10:03 PM ET

JAPAN Nikkei	HangSeng	CHINA Shanghai
14,942.78	22,690.46	2,059.39
+221.30	+32.83	−12.09
+1.50%	+0.14%	−0.58%

Data delayed at least 15 minutes

Get Quotes | My Portfolios »

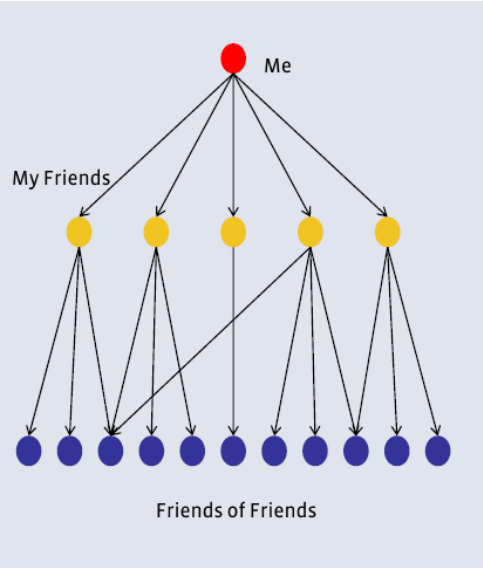
The ads that are most likely to lead to a click are selected and displayed

Application: Learning to Suggest Friends on Facebook

Problem: automatically suggest "friend" links

Can restrict to "friends of friends"

- Still leaves 40,000 possibilities on average



From:
L. Backstrom, invited keynote talk at ESWC 2011
Online video and slides at http://videolectures.net/eswc2011_backstrom_facebook/

L. Backstrom and J. Leskovec
Supervised Random Walks: Predicting and Recommending Links in Social Networks
ACM Conference on Web Search and Data Mining (WSDM), 2011

Application: Learning to Suggest Friends on Facebook

Solution: learn a prediction model

Target: whether user clicks or not on the recommendation

Features: mutual friends, age, geography, etc

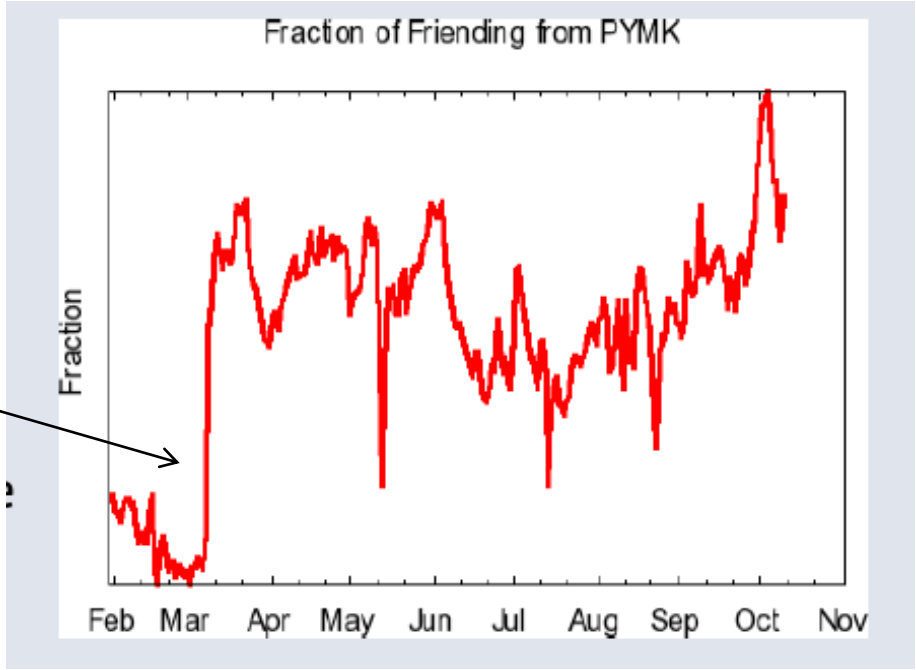
Models: logistic regression + other models

Significant engineering:
feature computation in real-time, plus real-time feedback

From Lars Backstrom, ESWC 2011

Application: Learning to Suggest Friends on Facebook

Significant improvement in click-through rate when system went live

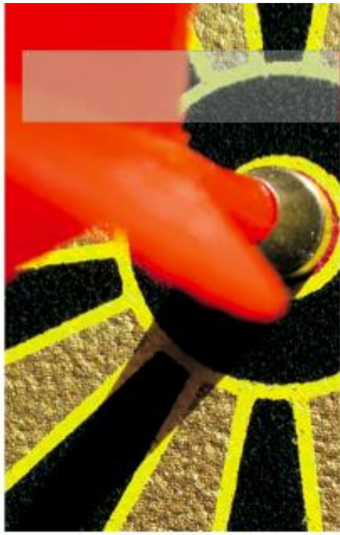


From:
L. Backstrom, invited keynote talk at ESWC 2011
Online video and slides at http://videolectures.net/eswc2011_backstrom_facebook/

L. Backstrom and J. Leskovec
Supervised Random Walks: Predicting and Recommending Links in Social Networks
ACM Conference on Web Search and Data Mining (WSDM), 2011

From Lars Backstrom, ESWC 2011

IEEE Intelligent Systems, 2009



EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas

such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary par-

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Neural Networks and Deep Learning

Neural Network Models

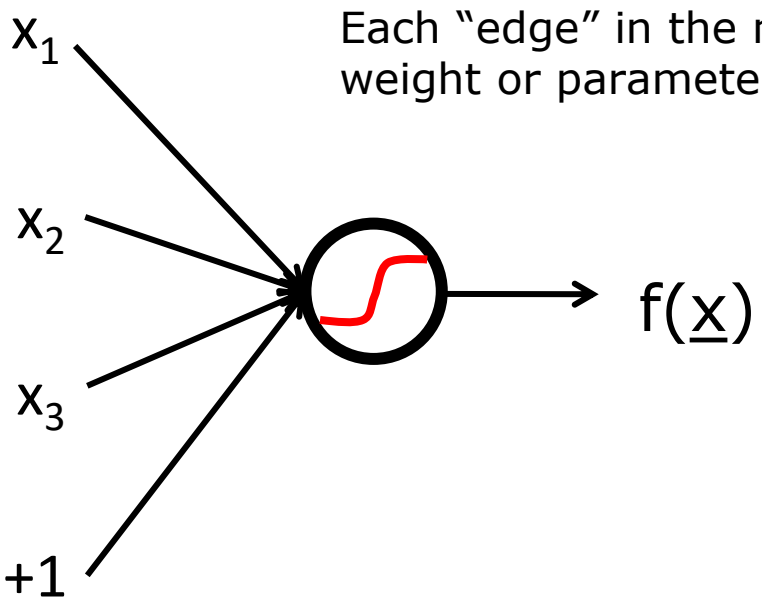
- **"Layers" of simple processing units with weights**
  - (superficial) similarity to real neural/brain models
  - Better to think of them as powerful flexible functions
  - We can learn the weights for these functions from data
  - Key aspect: each "layer" extracts useful features for later layers

- **3 phases of research**
  - Phase 1: early work in 1960's on very simple network models
  - Phase 2: late 80's, early 90's, a resurgence of interest
  - Phase 3: 2010 to now, huge interest in "deep neural networks"

- **Applications**
  - State-of-the-art for classifying objects in images, classifying words in speech
  - Significant investment by companies like Facebook, Google, and others
  - Often require very large labeled data sets (millions)
  - Significant engineering and know-how required

Neural Networks

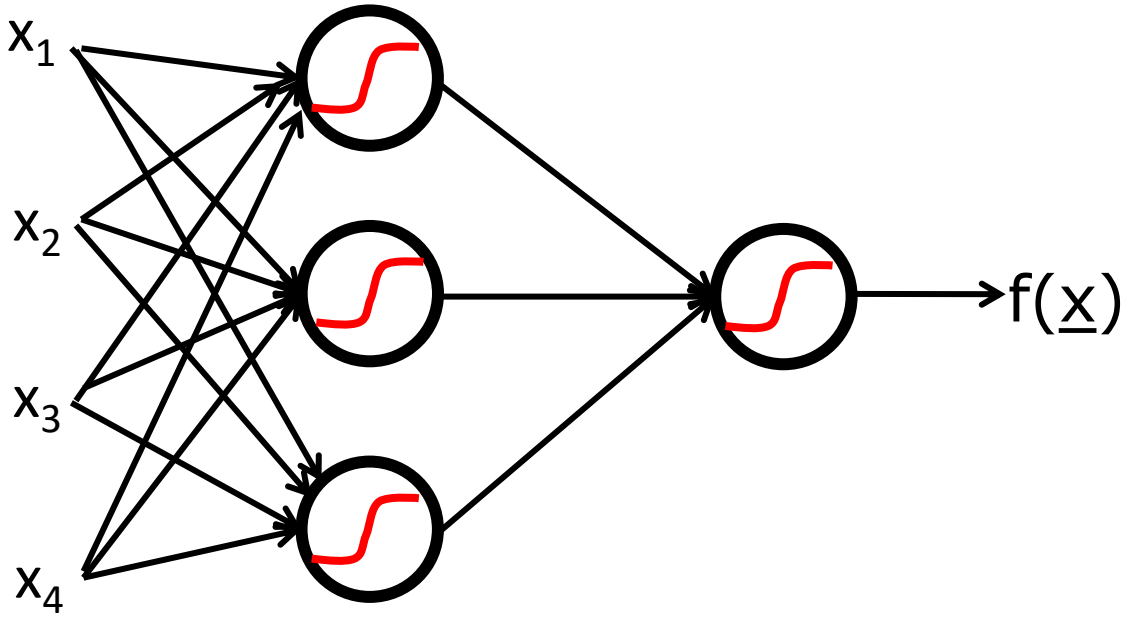
- Logistic regression can be viewed as a simple neural network



Each "edge" in the network has an associated weight or parameter, α_j

$$f(\underline{x}) = \hat{P}(Y = 1|\underline{x}) = \frac{1}{1 + e^{(-\sum_{j=1}^{d} \alpha_j x_j)}}$$

A Neural Network with 1 Hidden Layer

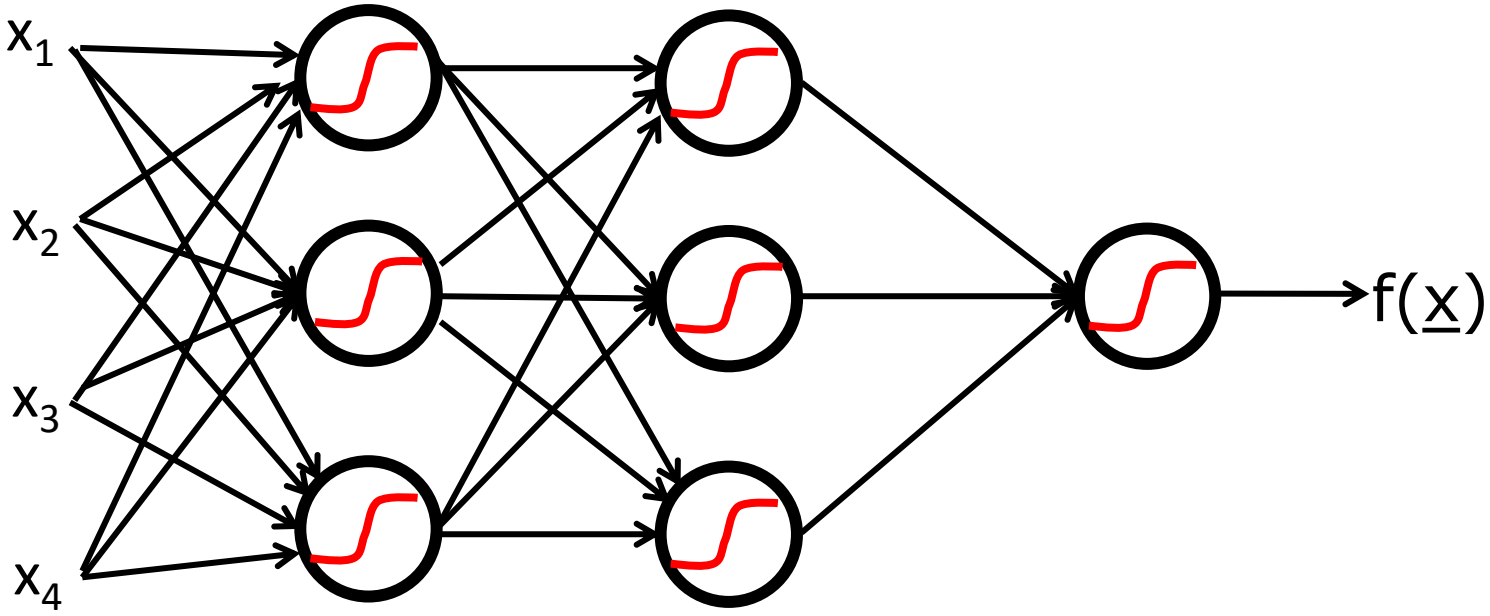


Can recursively create more complex prediction models

Many more weights now….requires more data to estimate

Deep Learning: Models with 2 or More Hidden Layers

We can build on this idea to create "deep models" with many hidden layers



The model is now a very flexible highly non-linear function

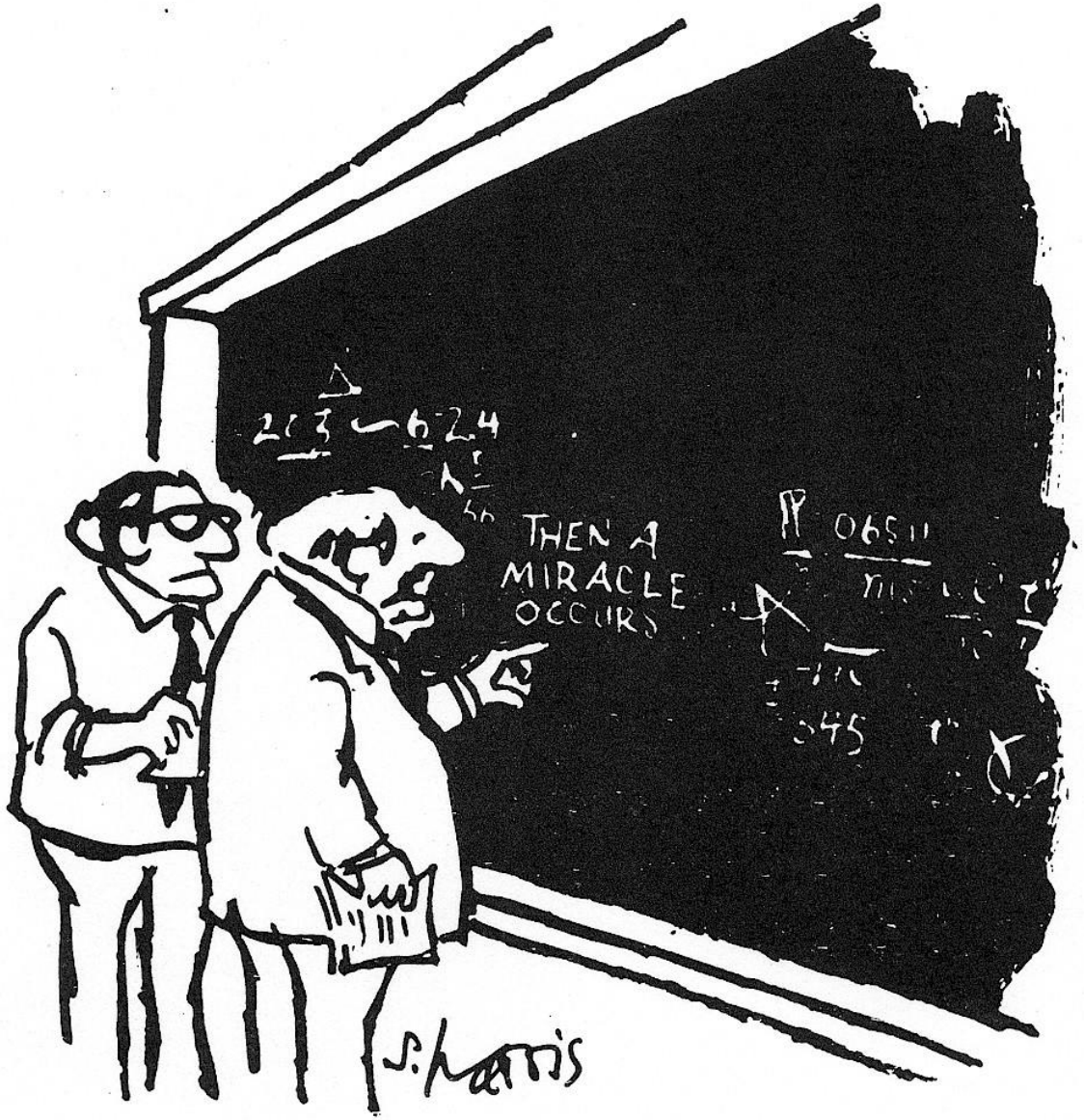
Training Neural Networks

- **There are quite a few "design parameters" that need to be selected or tuned when working with neural networks**

- **Network model**
  - Number of hidden layers and number of units per layer
  - The equation for the hidden unit (logistic is common, but others also used)

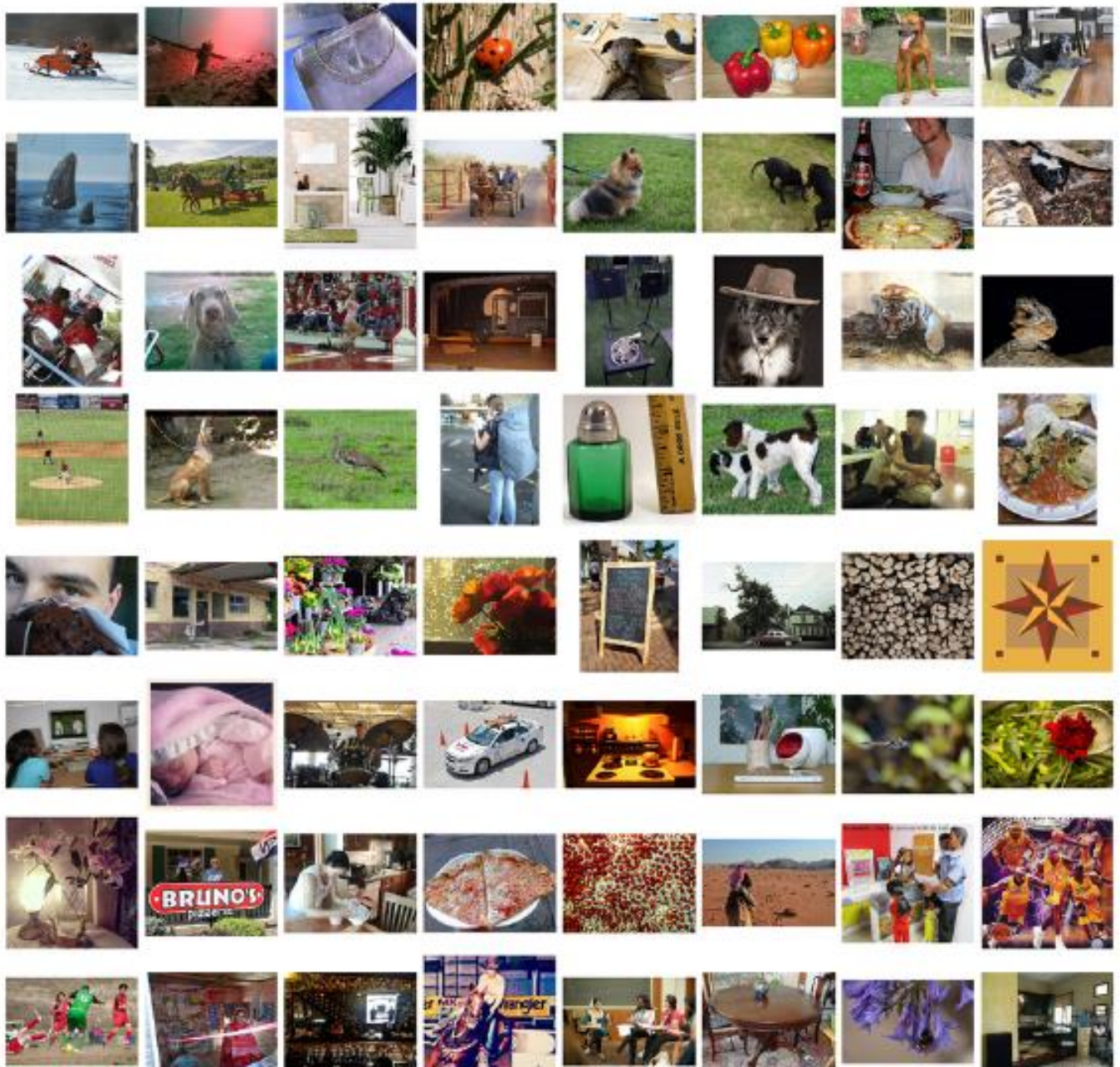
- **Learning algorithm ("gradient descent")**
  - Initialization of the weights
  - Scaling of the input variables
  - Learning rate: how far the algorithms moves the weights at each iteration
  - …..plus many more heuristics and ideas

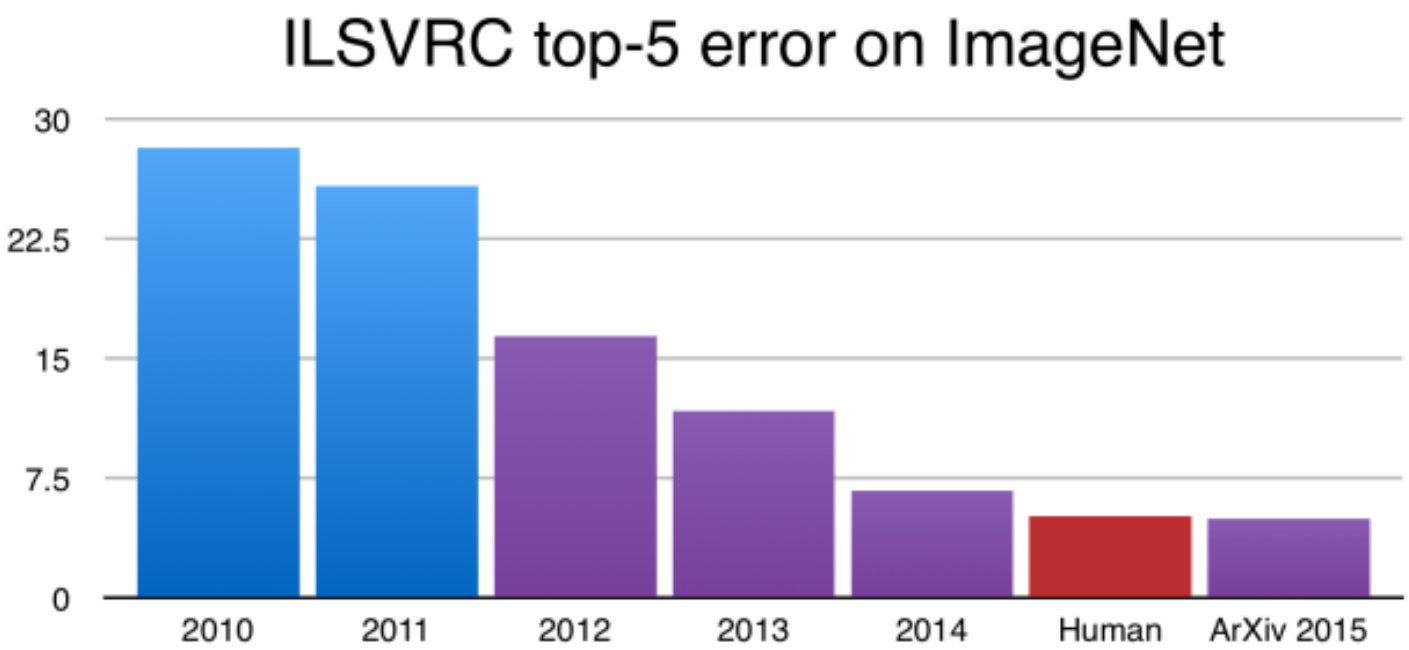
- **Bottom line: these are complex models to work with, requiring a fair bit of "know-how", and might be more complex than needed for many prediction problems**



"I think you should be more explicit here in step two."

Russakovsky et al, ImageNet Large Scale Visual Recognition Challenge, 2015




ILSVRC top-5 error on ImageNet
30
22.5
15
7.5
0
2010
2011
2012
2013
2014
Human
ArXiv 2015

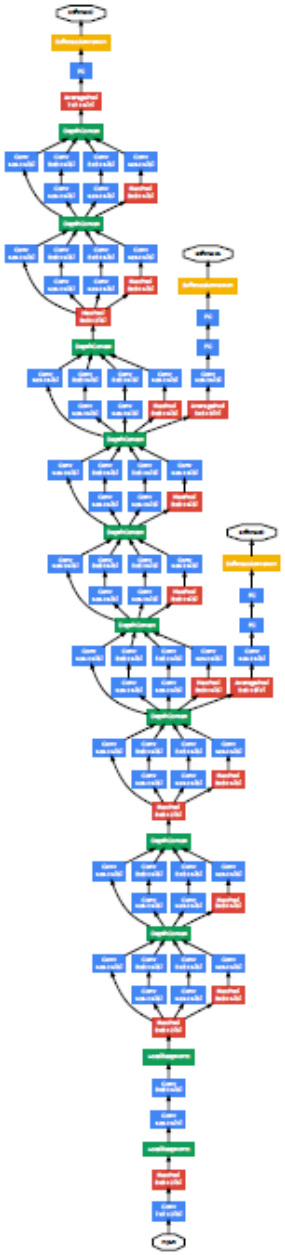


Figure 3: GoogLeNet network with all the bells and whistles

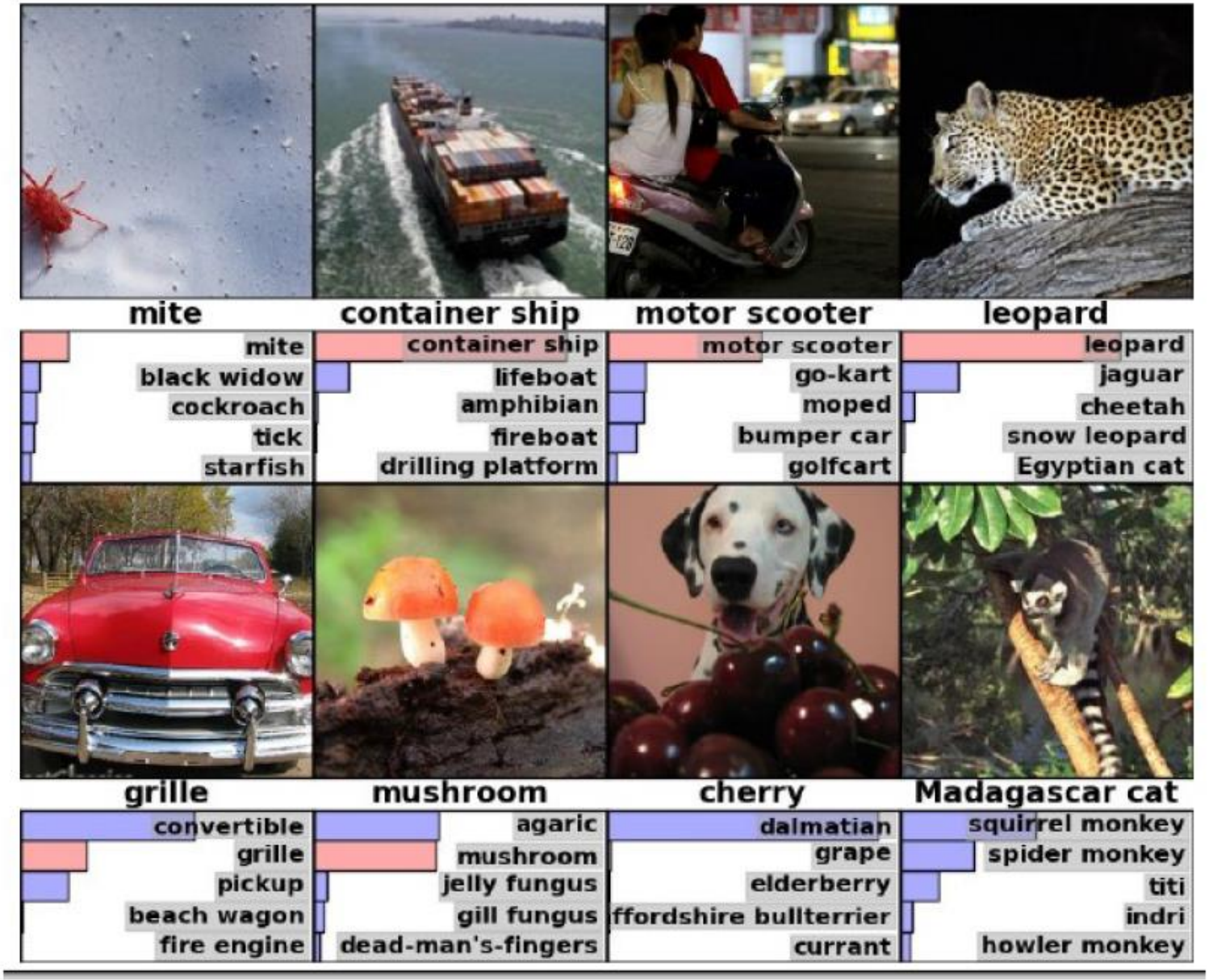
Deep Network architecture for GoogLeNet network, 27 layers

Don't try this at home….

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

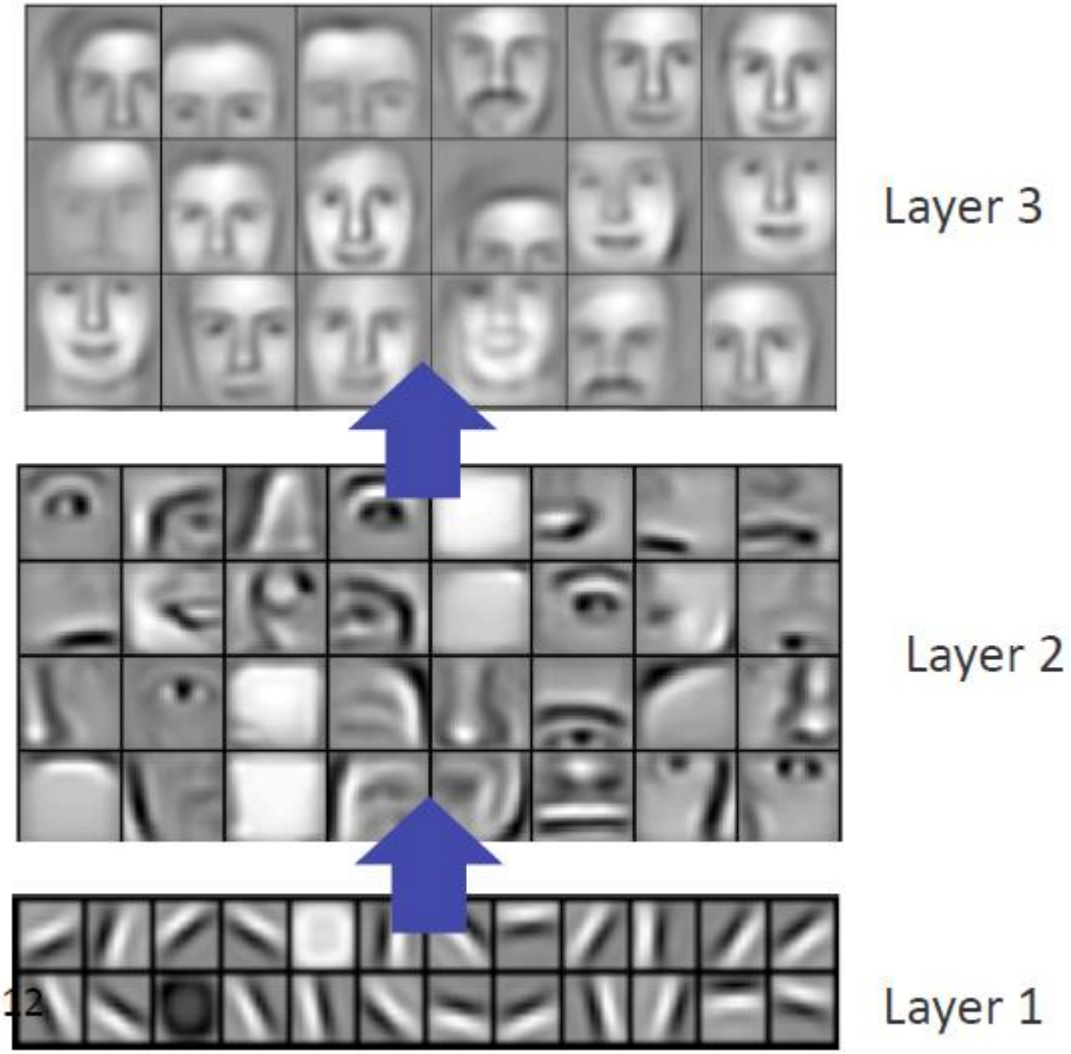
Table 1: GoogLeNet incarnation of the Inception architecture

Figure from Krizhevsky, Sutskever, Hinton, 2012



Visualizing what the Hidden Units are Learning

Figure from Lee et al., ICML 2009



THE DARK SIDE OF BIG DATA….

ALCHEMY IN THE BEHAVIORAL SCIENCES*

BY HILLEL J. EINHORN

Access to powerful new computers has encouraged routine use of highly complex analytic techniques, often in the absence of any theory, hypotheses, or model to guide the researcher's expectations of results. The author examines the potential of such techniques for generating spurious results, and urges that in exploratory work the outcome be subjected to a more rigorous criterion than the usual tests of statistical significance.

Hillel Einhorn is Assistant Professor of Behavioral Science, Graduate School of Business, University of Chicago.

WITH THE LARGE-SCALE use of electronic computers, powerful new methods for data analysis have become quite prevalent. Although this development may be viewed with considerable enthusiasm by some,[1] others may view the gains to be derived from increased ability to handle large amounts of data with increasingly sophicated tools with more than a certain degree of skepticism. Such skepticism is based on the observation that as methods and techniques get more complicated, the role of theory in research is being dangerously ignored in favor of purely empirical work that proceeds without so much as a hypothesis. Like Pirandello's characters in search of an author, many of today's researchers seem to have an assortment of techniques in search of a substantive problem.

The general question of proceeding inductively or deductively in science is not easily answered. As Armstrong has put it,

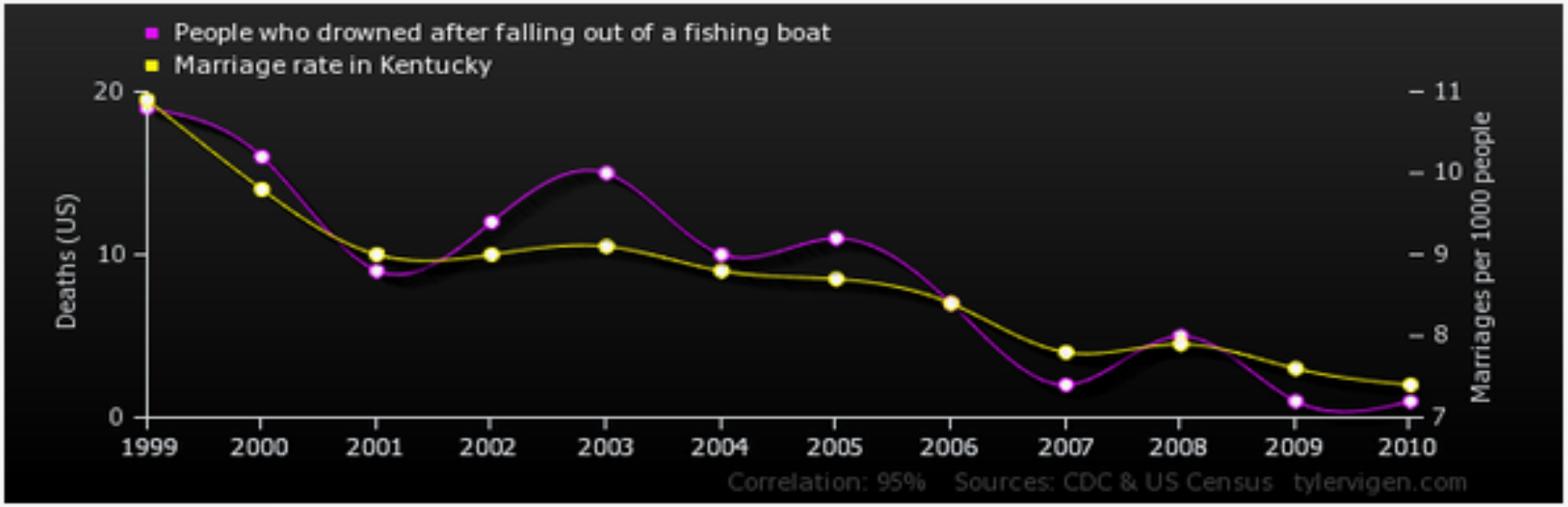
Public Opinion Quarterly, 1972

"The role of theory in research is being dangerously ignored in favor of purely empirical work that proceeds without so much as a hypothesis."

People who drowned after falling out of a fishing boat

correlates with

Marriage rate in Kentucky



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
People who drowned after falling out of a fishing boat *Deaths (US) (CDC)*	19	16	9	12	15	10	11	7	2	5	1	1
Marriage rate in Kentucky *Marriages per 1000 people (US Census)*	10.9	9.8	9	9	9.1	8.8	8.7	8.4	7.8	7.9	7.6	7.4

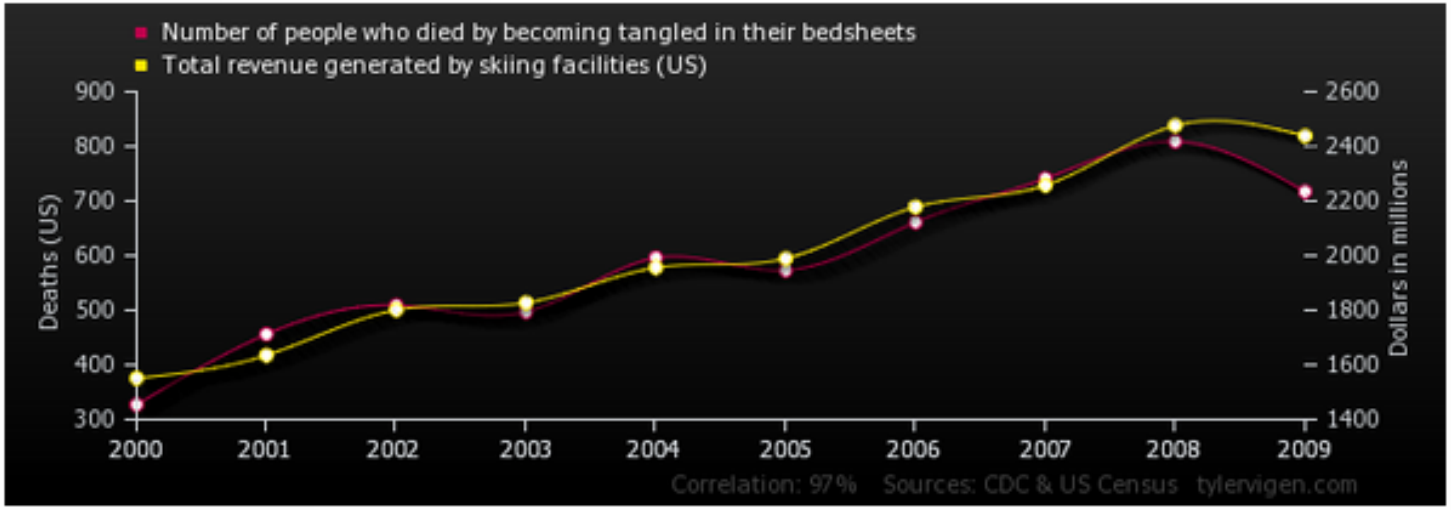
Correlation: 0.952407

Graphics from http://www.tylervigen.com/

Number of people who died by becoming tangled in their bedsheets

correlates with

Total revenue generated by skiing facilities (US)

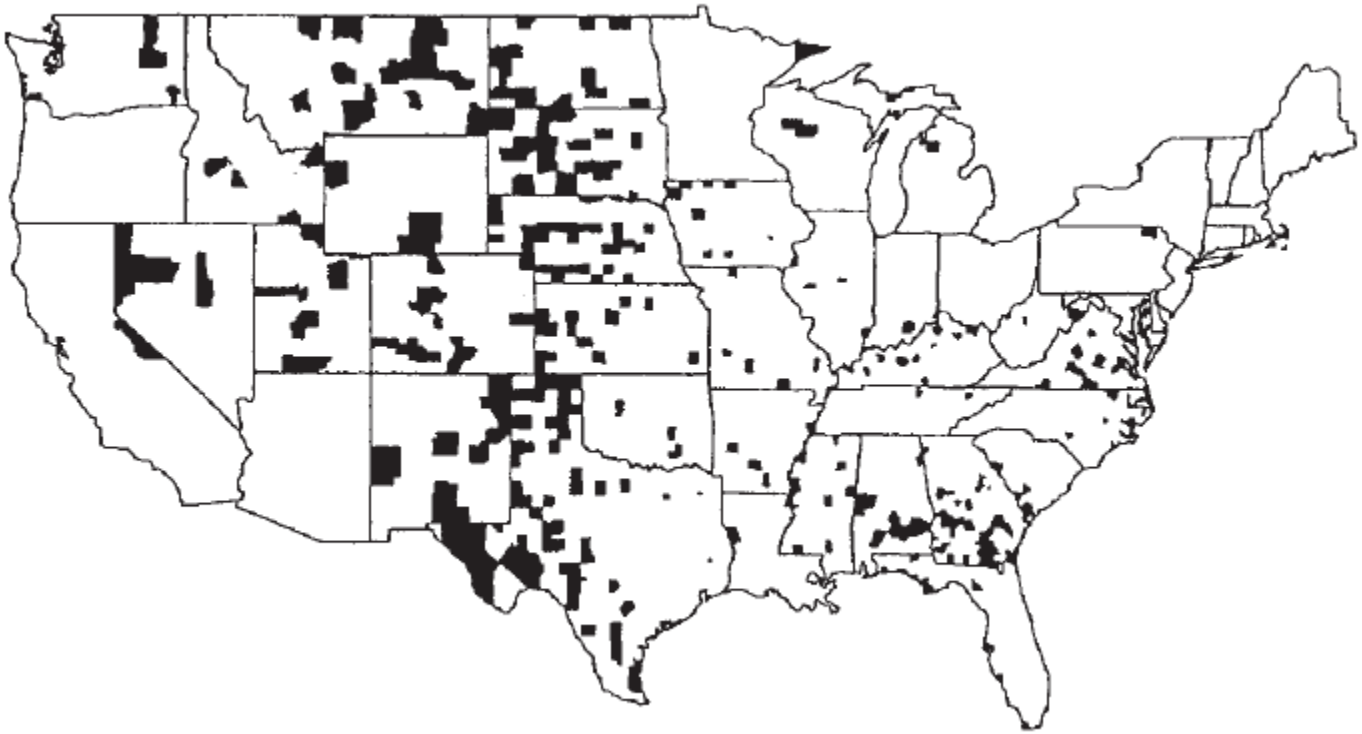


	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717
Total revenue generated by skiing facilities (US) *Dollars in millions (US Census)*	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Correlation: 0.969724

Graphics from http://www.tylervigen.com/

Counties with Lowest Kidney Cancer Death Rates in the US



From A. Gelman and D. Nolan
Oxford University Press, 2002

How Much Climate Data Do We Actually Have?

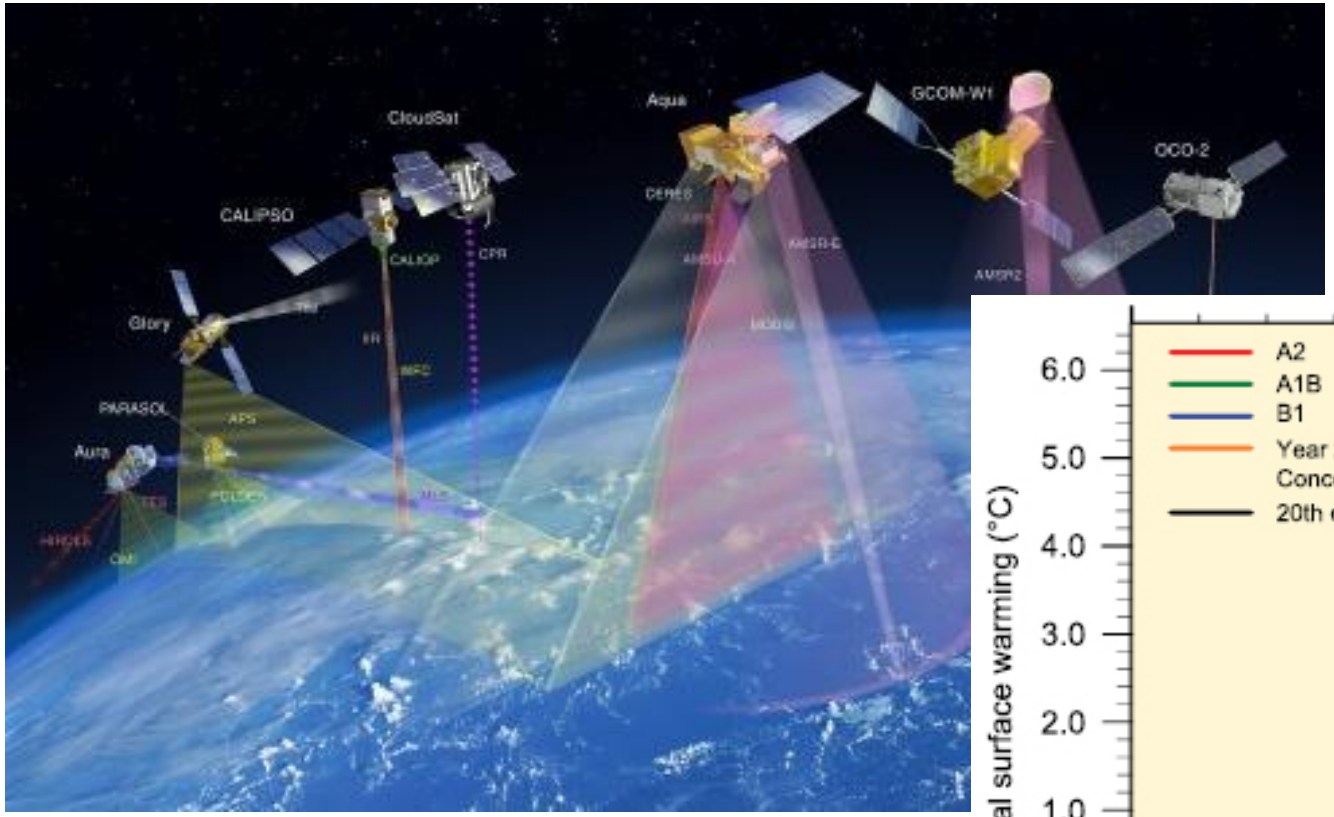


Image from http://cimss.ssec.wisc.edu/

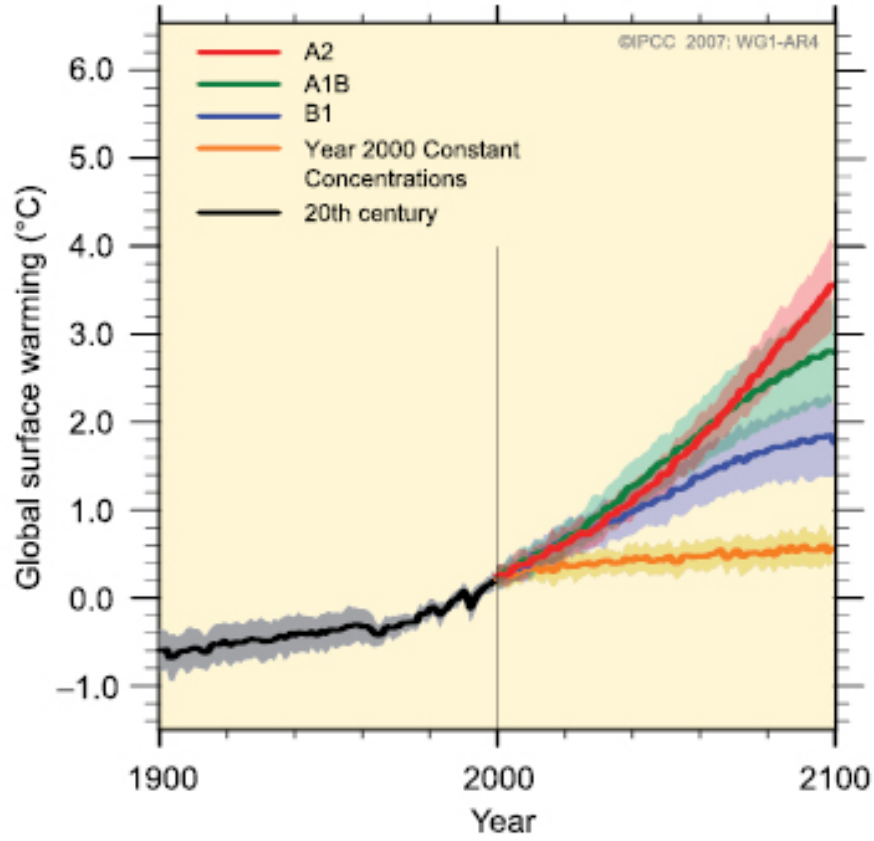


Image from ipcc.ch

Fooling Automated Essay Grading

From New Statesman and New York Times, April 2012

Les Perelman, MIT, experimented with different essays to test the Educational Testing Service (ETS)'s automated eRater program

All of his essays received a perfect score

SAT prompt:

"*The rising cost of a college education is the fault of students who demand that colleges offer students luxuries unheard of by earlier generations of college students -- single dorm rooms, private bathrooms, gourmet meals, etc.*"

Discuss the extent to which you agree or disagree with this opinion. Support your views with specific reasons and examples from your own experience, observations, or reading.

Portions of a Perfect-Scoring Essay

Teaching assistants are paid an excessive amount of money. The average teaching assistant makes six times as much money as college presidents. In addition, they often receive a plethora of extra benefits such as private jets, vacations in the south seas, a staring roles in motion pictures.

Portions of a Perfect-Scoring Essay

In Heart of Darkness, Mr. Kurtz is a teaching assistant because of his connections, and he ruins all the universities that employ him. Finally, teaching assistants are able to exercise mind control over the rest of the university community. The last reason to write this way is the most important. Once you have it down, you can use it for practically anything. Does God exist? Well, you can say yes and give three reasons, or no and give three different reasons. It doesn't really matter.

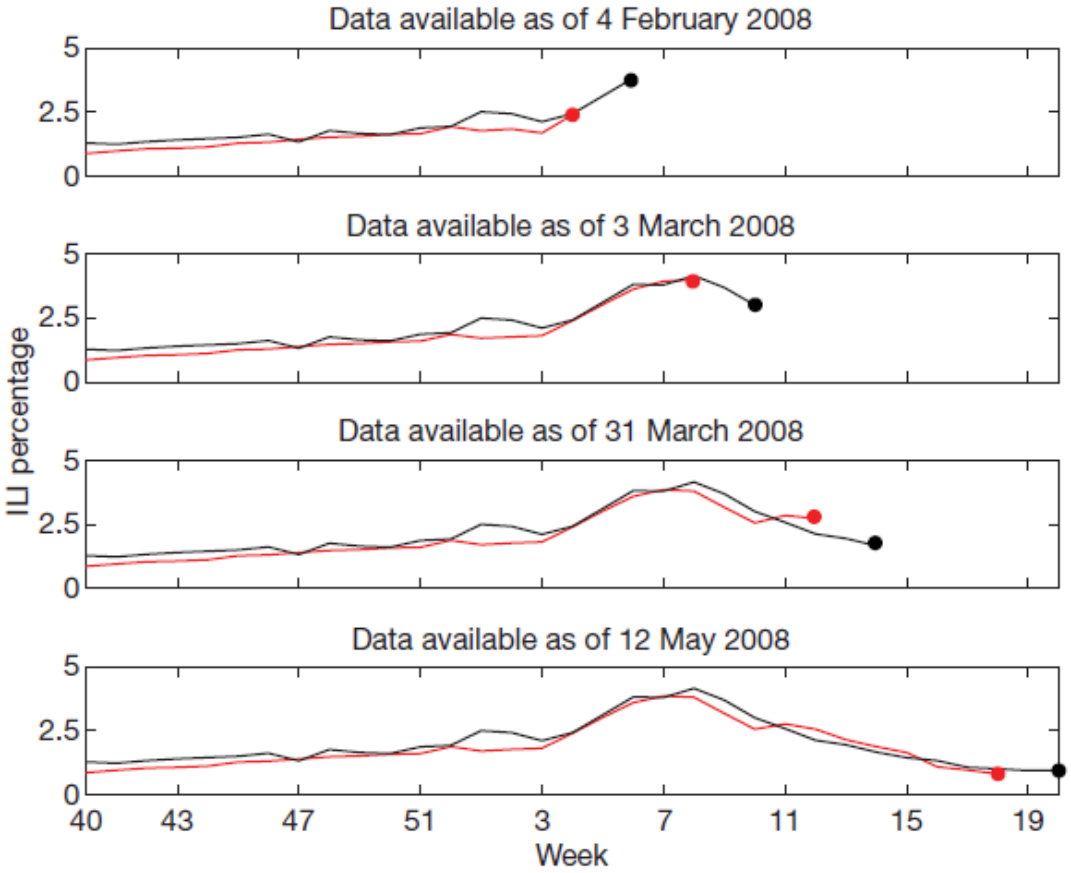
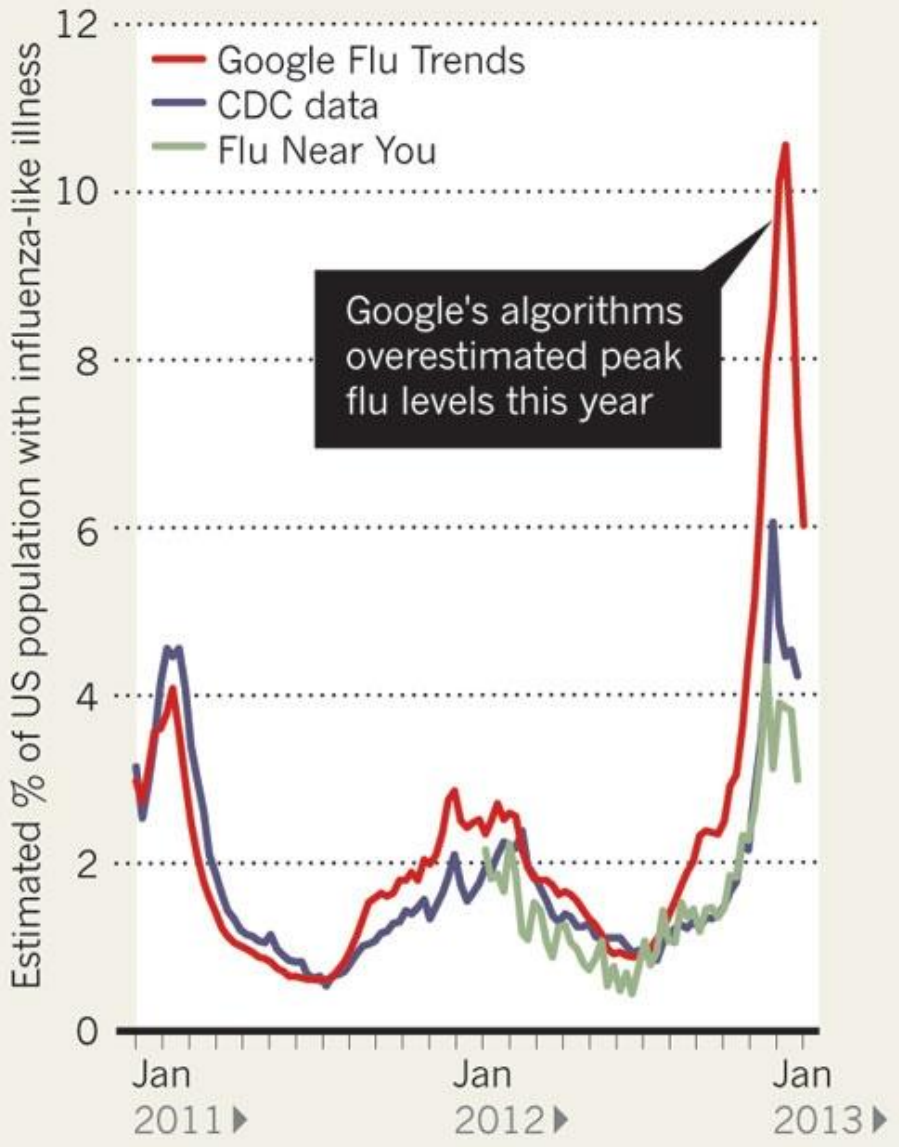


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3 March our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.

Google Flu predictions based on search queries were accurate 2 weeks ahead of CDC predictions

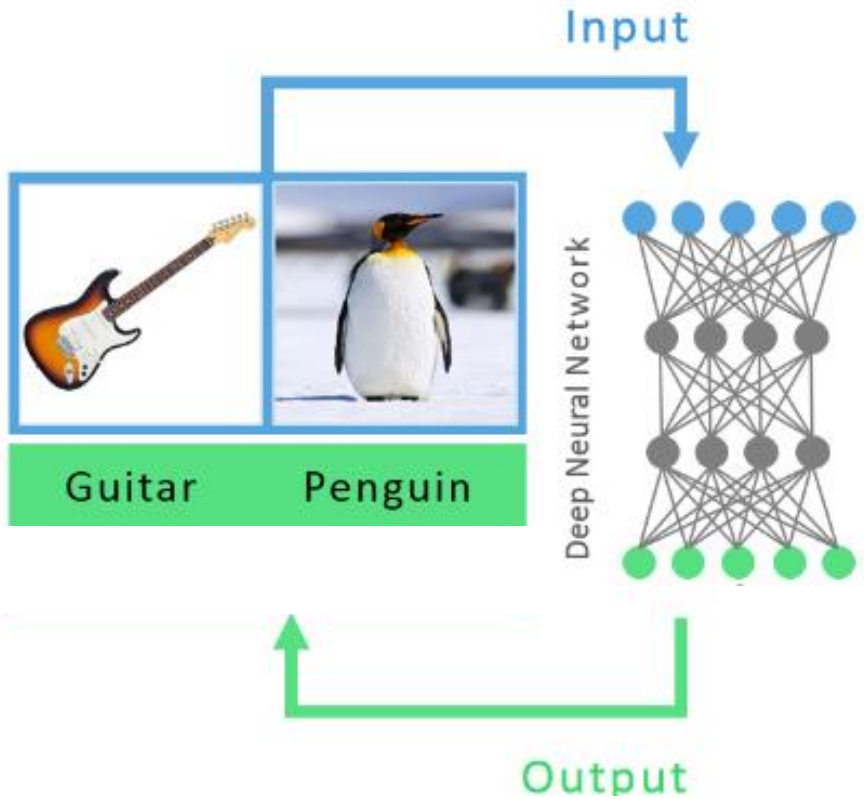
FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



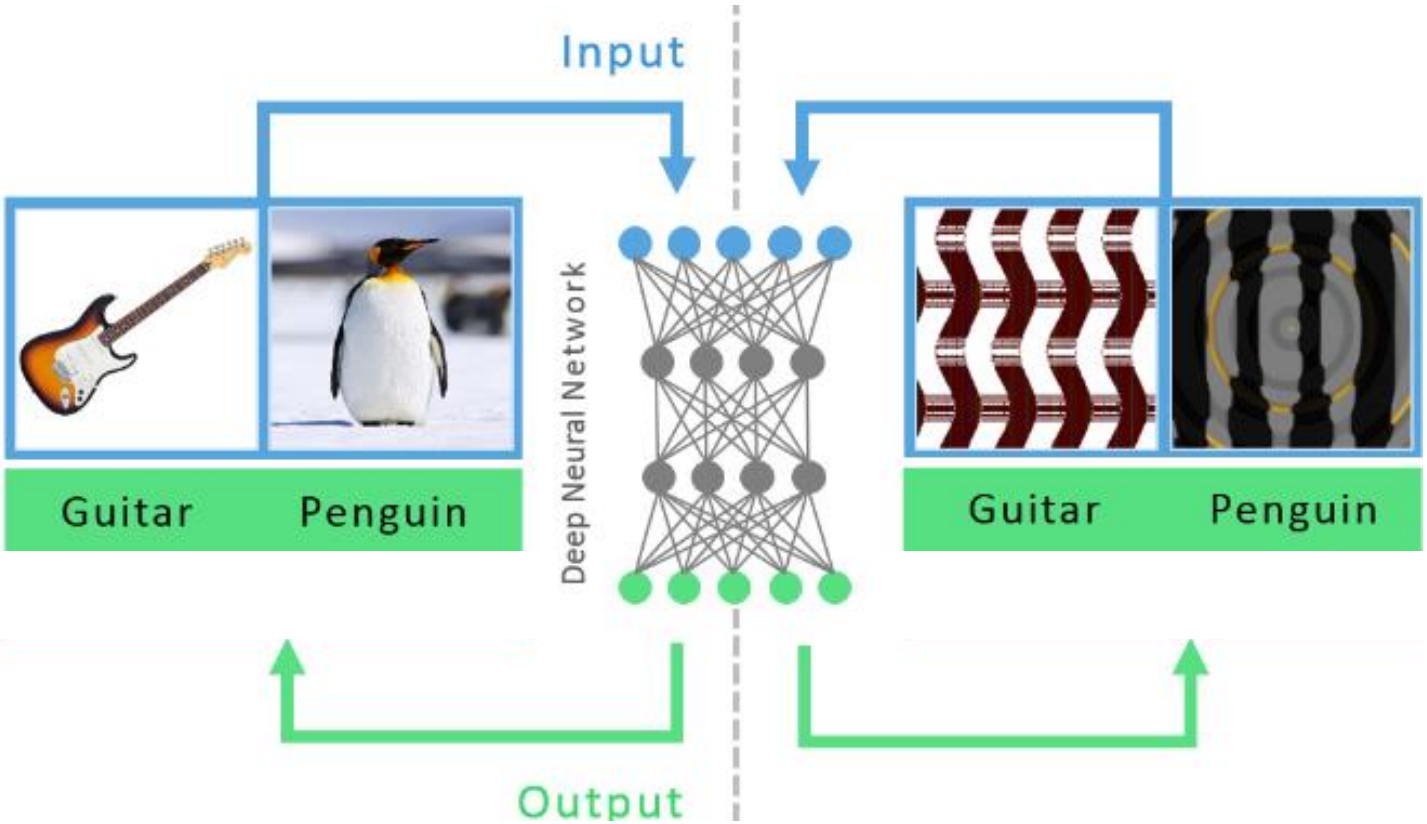
A Deep Neural Network for Image Recognition

From Nguyen, Yosinski, Clune, ArXiv preprint, 2014



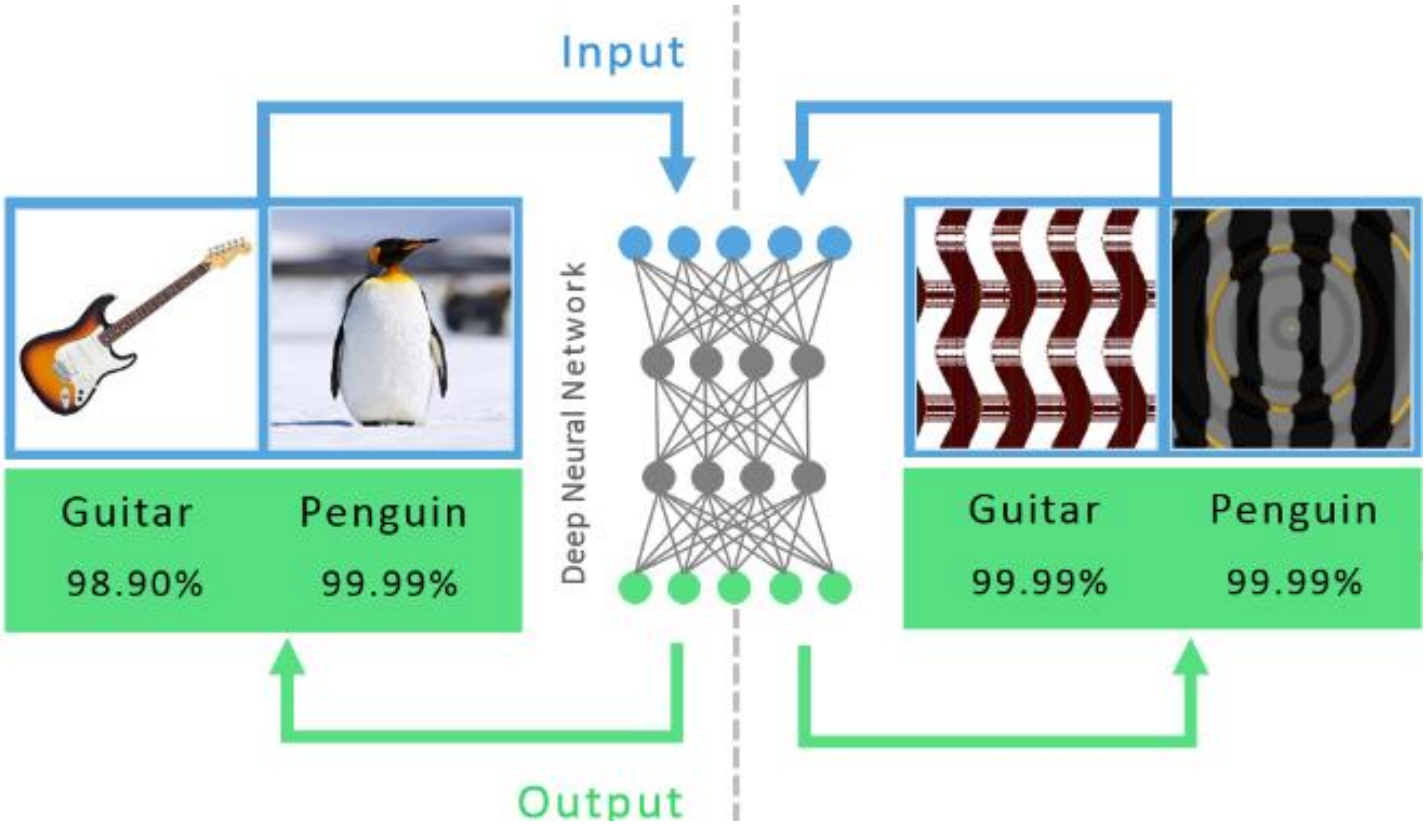
Poor Extrapolation

From Nguyen, Yosinski, Clune, ArXiv preprint, 2014



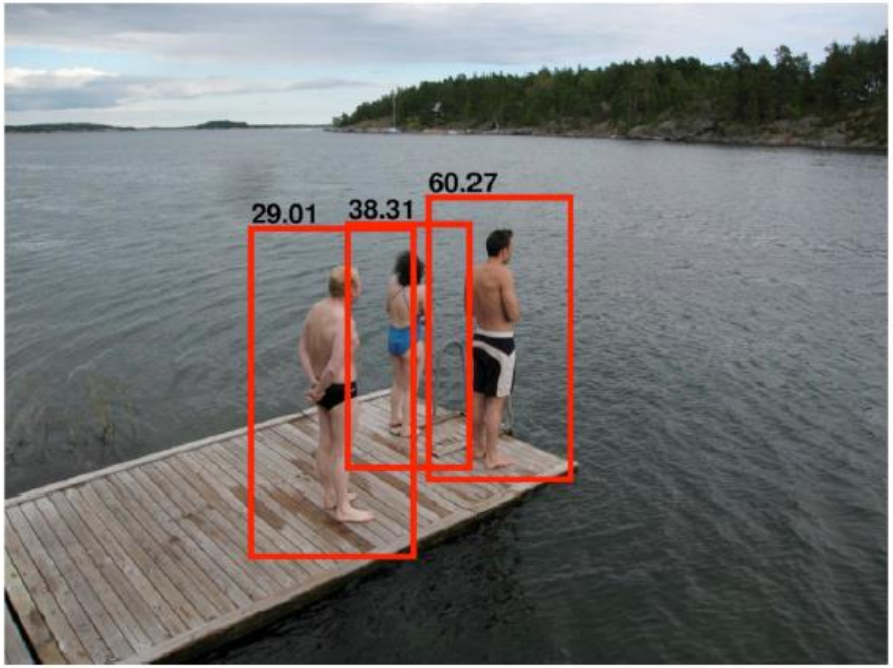
Lack of Calibration

From Nguyen, Yosinski, Clune, ArXiv preprint, 2014



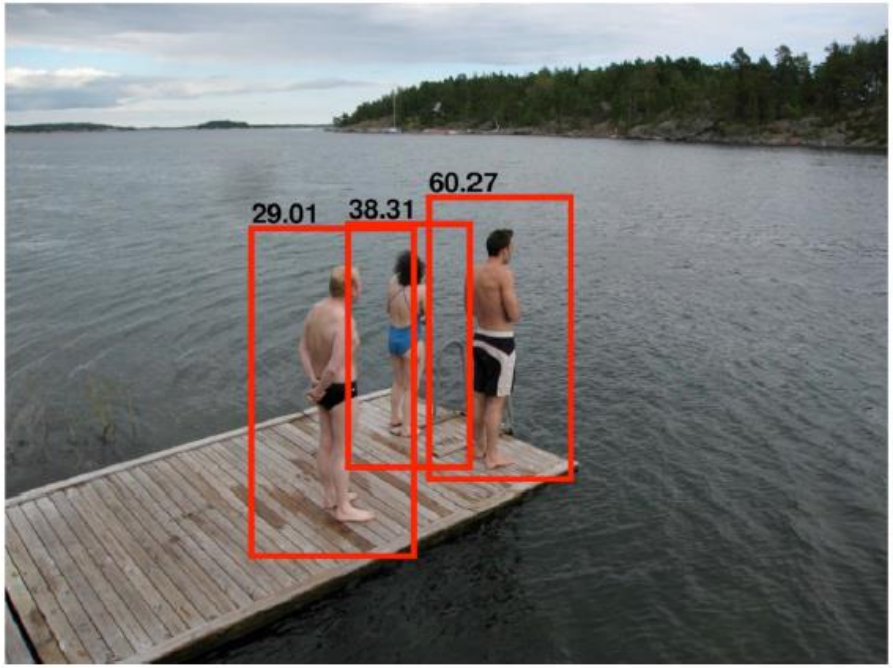
Computer Vision

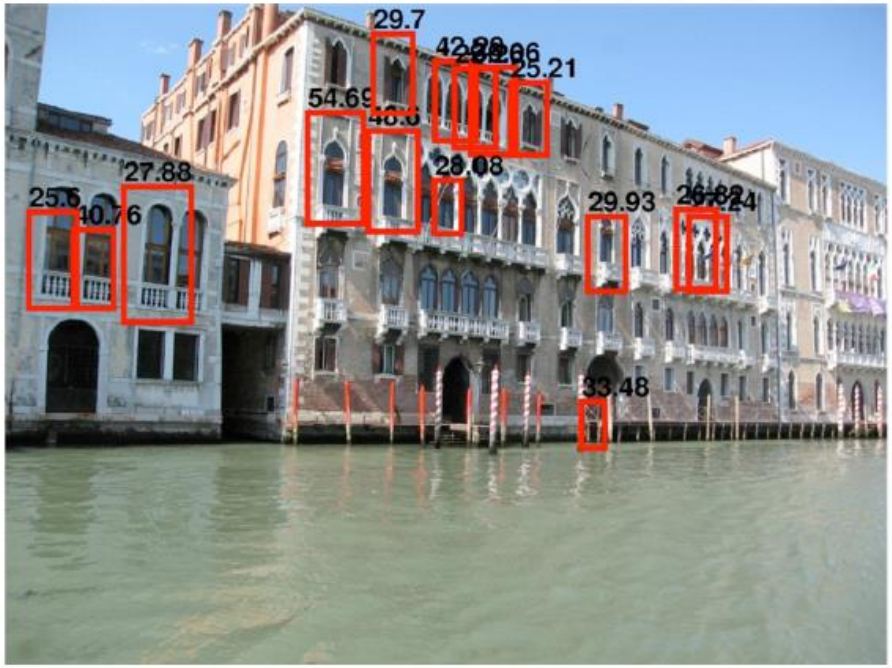
Example from Welinder, Welling, and Perona, CVPR 2013

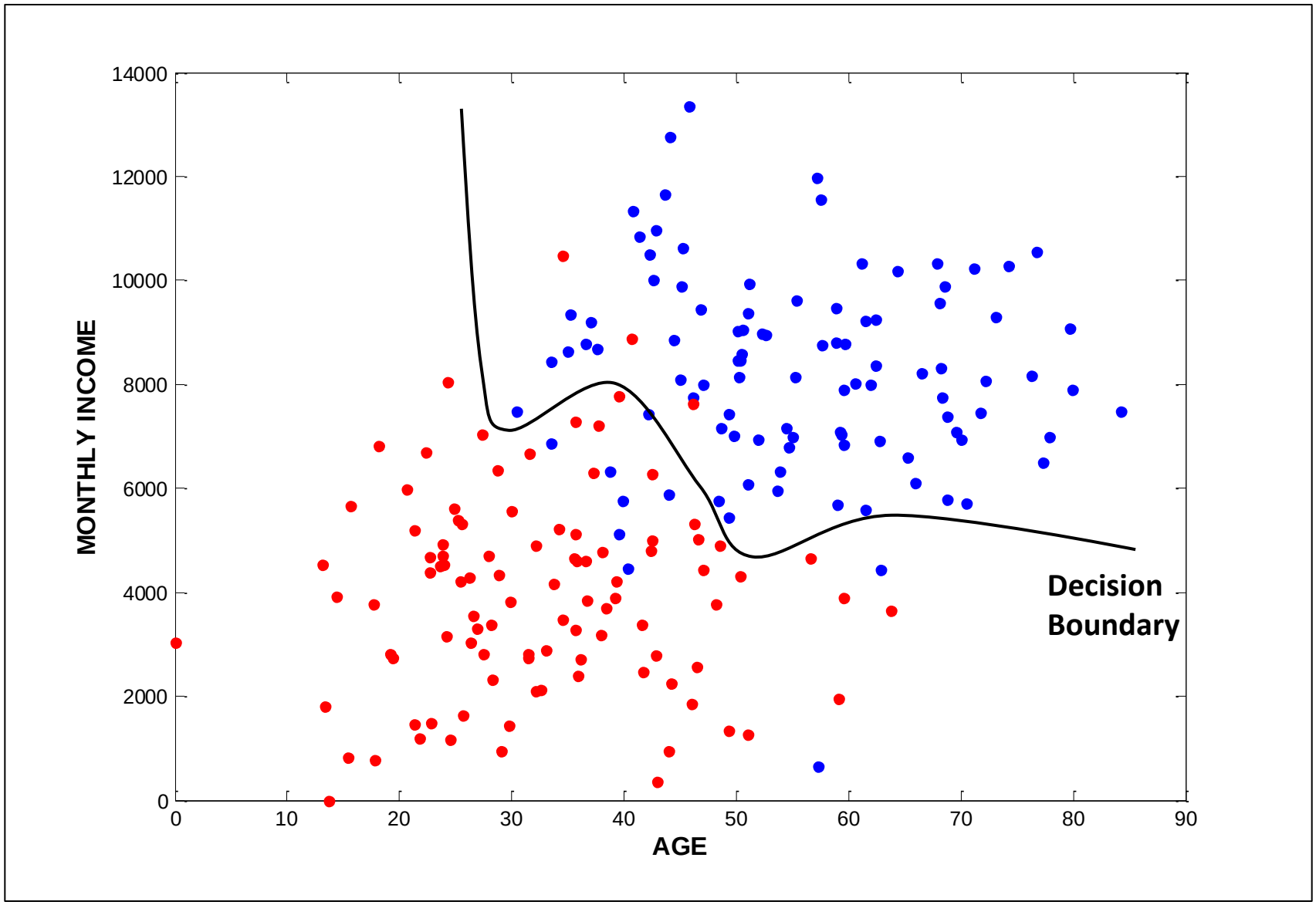


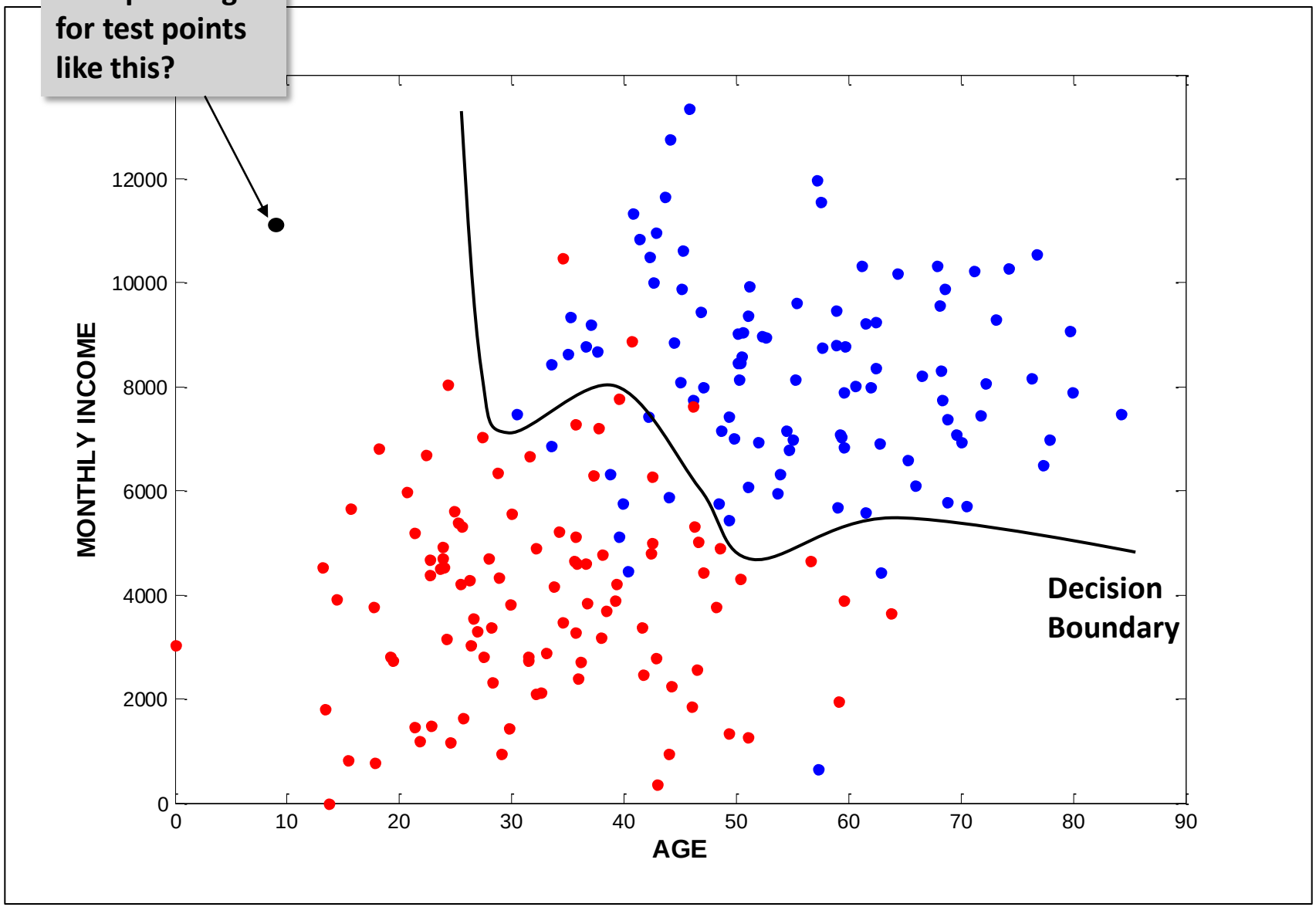
Computer Vision

Example from Welinder, Welling, and Perona, CVPR 2013






MONTHLY INCOME
AGE
Decision
Boundary
0
2000
4000
6000
8000
10000
12000
14000
0
10
20
30
40
50
60
70
80
90


Poor at extrapolating for test points like this?
MONTHLY INCOME
AGE
0
2000
4000
6000
8000
10000
12000
0
10
20
30
40
50
60
70
80
90
Decision Boundary

Interpolation and Extrapolation

- Interpolation
  - Machine learning and statistical approaches are very good when the training and test data come from (approximately) the same distributions
    - e.g., any research study where we randomly partition data into train and test
    - e.g., train on camera images of human faces, test on another 1 million

- Extrapolation
  - This is much harder: machine learning and statistics alone cannot work here
  - Examples:
    - Predicting the behavior of loan applicants 3 years from now
    - Predicting the effects of global temperature changes on storm frequency
    - Predicting how a particular patient will respond to chemotherapy
  - Typically notions of causality and first principles knowledge are required….in addition to empirical data

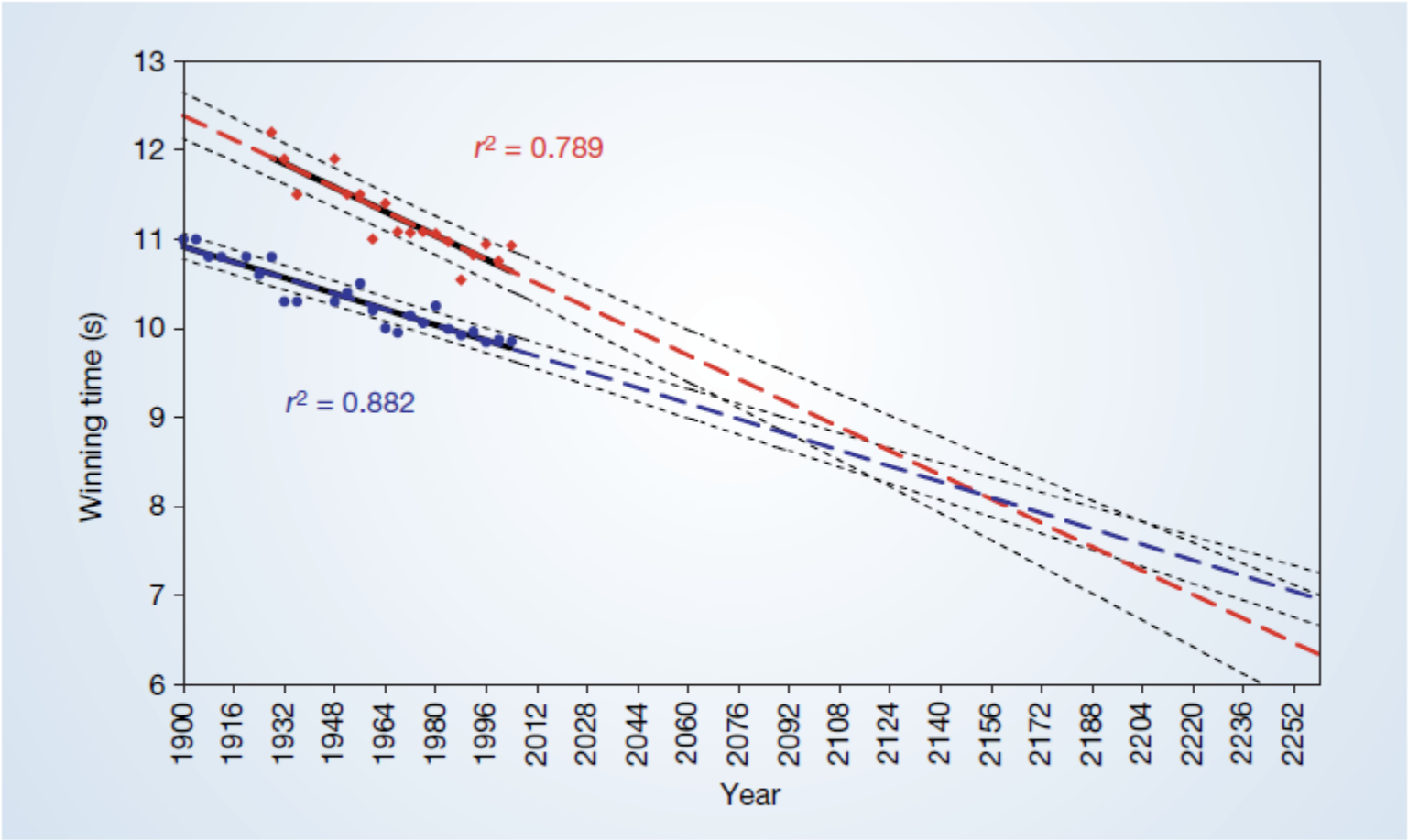


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

From Tatem et al., Nature 2004.
(see also response letters at http://faculty.washington.edu/kenrice/natureletter.pdf)

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

CONCLUDING COMMENTS

What are Machine Learning's Strengths?

Problems that are relatively theory-free or model-free

- Image classification, text analysis, speech recognition, bioinformatics

Labeled data is relatively plentiful and cheap

- spam email is a good example

Large number of observations and many variables

- Significant leverage of algorithmic and optimization methods

Non-conventional data

- Text, images, social media
- Requires algorithmic pipelines for preprocessing, filtering, linking

When prediction accuracy is more important than model interpretation

Practical Advice

- Start with simple models
  - Logistic regression, boosted decision trees, are useful starting points
  - Deep learning (for example) may be much more complex than you need

- The accuracy of a model depends critically on the input variables
  - No model or optimization algorithm can find predictive signal if it is not there
  - E.g., trying to predict if a person has cancer given only demographic variables

- The 80/20 rule
  - A significant amount of time will be spent on problem definition, data gathering, data cleaning, model evaluation…..rather than on the algorithm itself

- Be objective and skeptical
  - Evaluate models with realistic test data (e.g., A/B tests)
  - If the results are too good to be true….be suspicious!

Big Data: Hope or Hype?

Useful in prediction problems with lots of data and little theory

Problems in medicine, climate, education can be much more challenging

Statistical thinking is key, linear algebra and optimization also important

Common sense applies

- Data does not exist in a vacuum – interpretation is important

Considerable hype in this area……but there is also reason for hope

The UCI Data Science Initiative

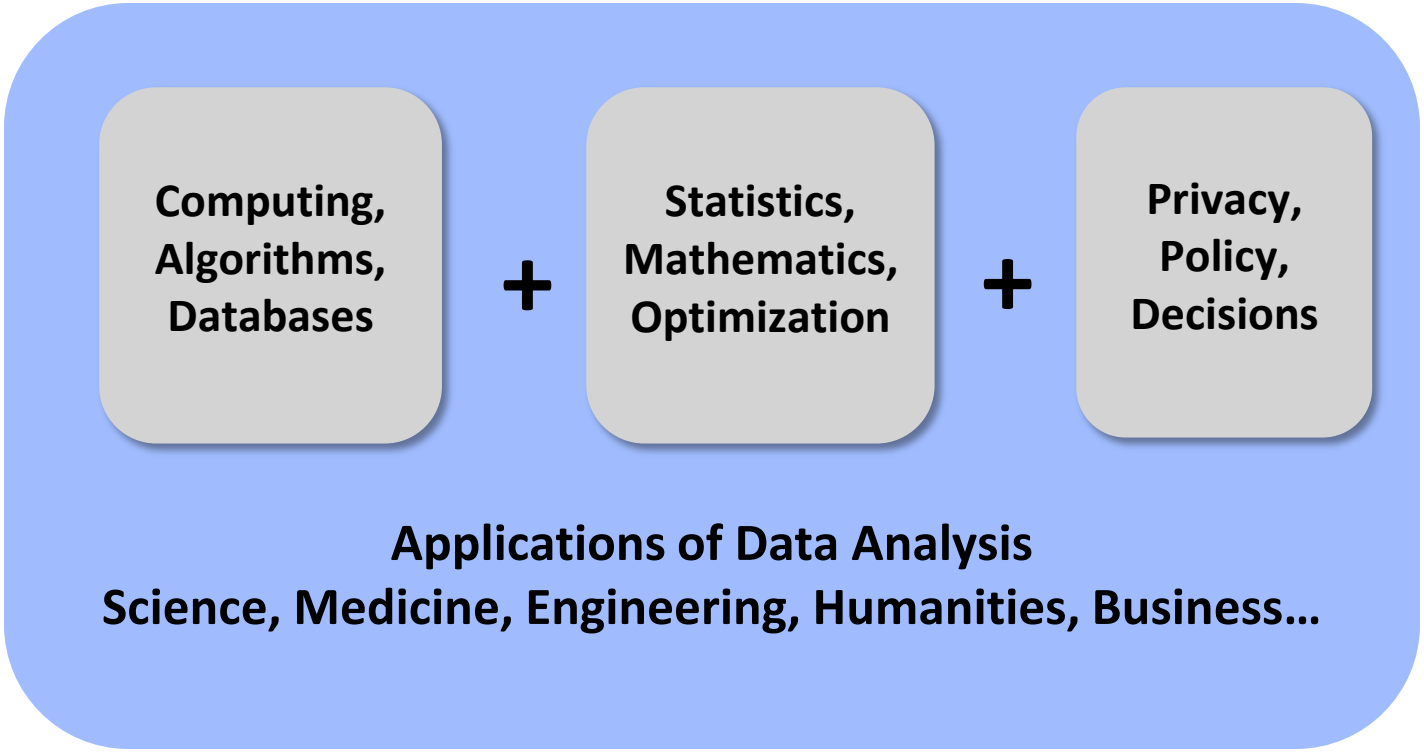
UCI Data Science Initiative

Kickoff Meeting

Friday, October 24, 2014

1:30-2:00 PM	Introduction
	What is the U Padhra
2:00-3:20 PM	Data Science
2:00	Bayesian Sta Hal St
2:20	Deep Learni Pierre
2:40	Platforms fo Michae
3:00	Me and My D Geof E
3:20	Questions an
3:30-3:45 PM	Break
3:45-4:45 PM	Data Science

The Science of Data Analysis



UCIrvine
University of California, Irvine

Faculty Advisory Board

Anima Anandkumar

Engineering

Jessica Utts

Pierre Baldi

Geof Bowker

Michael Carey

Information and Computer Sciences

Peter Krapp

Humanities

Andrew Noymer

Public Health

Jim Randerson

Physical Sciences

Suzanne Sandmeyer

Medicine

Vijay Gurbaxani

Business

Mark Warschauer

Education

Kevin Thornton

Biological Sciences

George Tita

Social Ecology

Mark Steyvers

Tom Boellstoff

Social Sciences

Activities

- Research Symposia
  - Algorithms for analyzing social network data (March 2015)
  - Text mining and education data (May 2015)
  - Digital humanities (Feb 2016)
  - Machine Learning (May 2016)
- Graduate Student Education
  - 1 and 2-day short courses on "hands-on" data science topics
    - Introduction to R, Predictive modeling in Python, etc
  - Over 400 graduate students have participated in 15 course offerings in the past 12 months
  - Summer fellowship program for PhD students on interdisciplinary data science research

Going Forward

- New Data Science Major (undergraduate)
  - Started in Fall 2015, first graduating class in 2018
  - Combination of courses from computer science and statistics
    - Statistical thinking plus computational skills
    - Emphasis on machine learning, statistical modeling, databases, algorithms, computational statistics, etc

- Short Courses for Industry
  - Currently under discussion
  - 2 to 3 day short courses on data science topics, e.g., predictive modeling
  - Taught by UCI faculty experts

- Longer term?
  - Permanent Data Science Institute?
  - Professional MS degree in predictive modeling/data science

The UCI Data Science Initiative

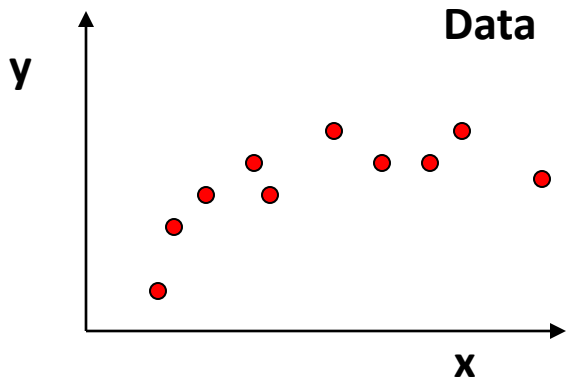
More information at http://datascience.uci.edu

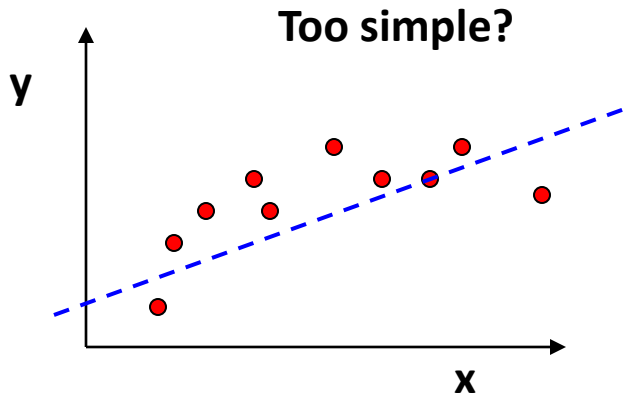
Sign up for mailing list (from the Web site) for announcements about symposia, etc

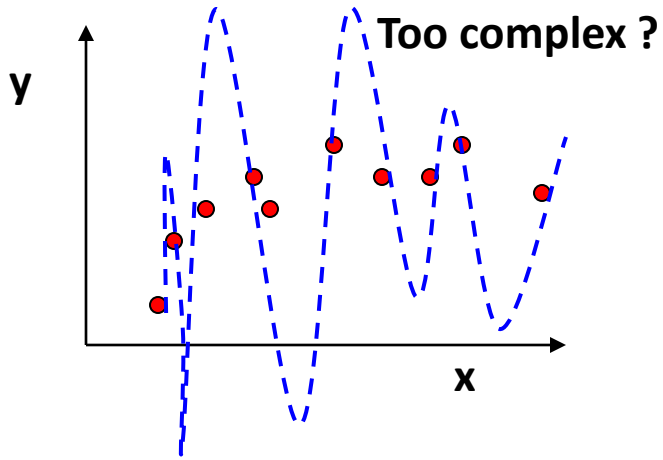
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

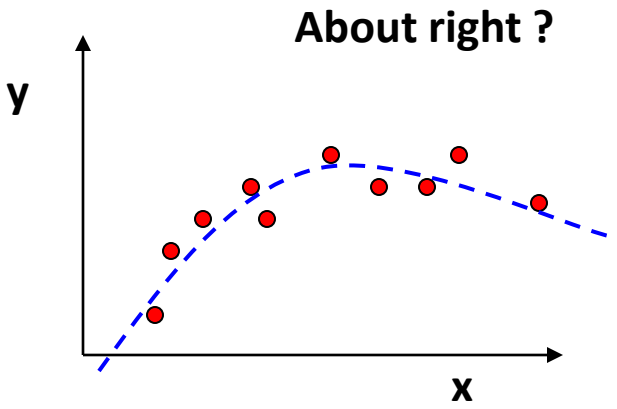
Backup Slides

Model Complexity

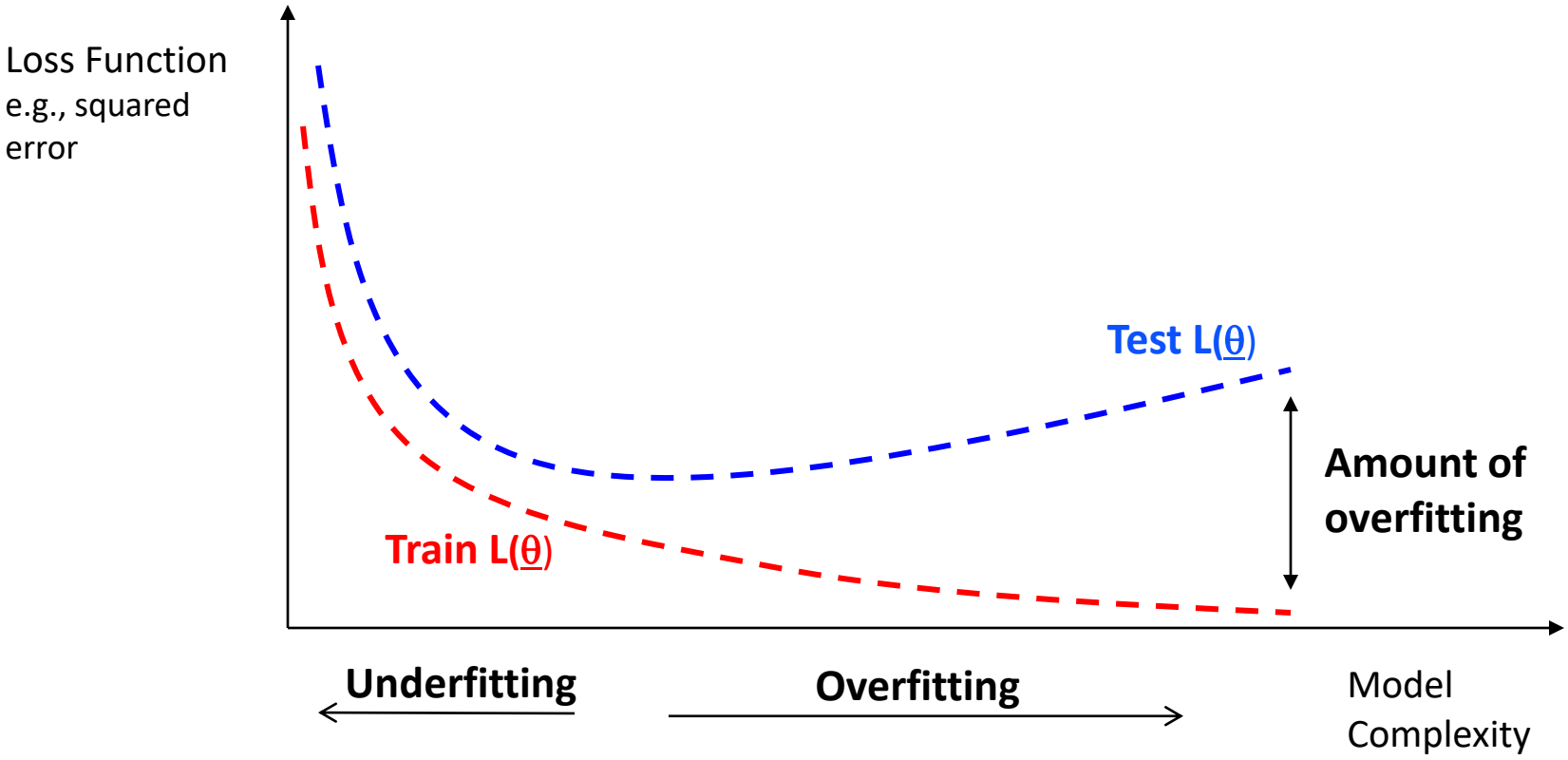








Model Complexity and Future Performance



Simple Example of a Document-Word Matrix

	water	farming	cattle	agriculture	land rights	use	stock	portfolio	bonds	return	loan	interest rate	profit
doc1	1	1			2		1						
doc2			1	4	1								
doc3	2	1	1		1	1	1						
doc4		1	2		1			2		1			1
doc5	1	1	1	1	3	1	1			1		3	1
doc6								2	1		1	1	
doc7				1			1		1	1	1	3	
doc8					1		1	1	1				
doc9										1	1		
doc10							1		1		1		1