# What is Data Science?

Peter Diao, SAMSI

Field of Dreams 2017

November 4, 2017

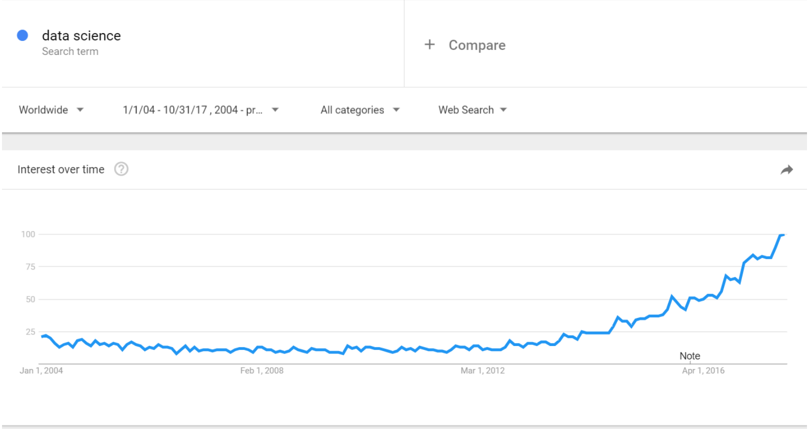
# Two Ways to Define a Field

1. "A mathematician, like a painter or a poet, is a maker of patterns. If his patterns are more permanent than theirs, it is because they are made with ideas." - Hardy, English Mathematician, 1877 - 1947

# Two Ways to Define a Field

1. "A mathematician, like a painter or a poet, is a maker of patterns. If his patterns are more permanent than theirs, it is because they are made with ideas." - Hardy, English Mathematician, 1877 - 1947
2. Mathematics is what mathematicians happen to be studying.

# Data Science as a term is getting very popular



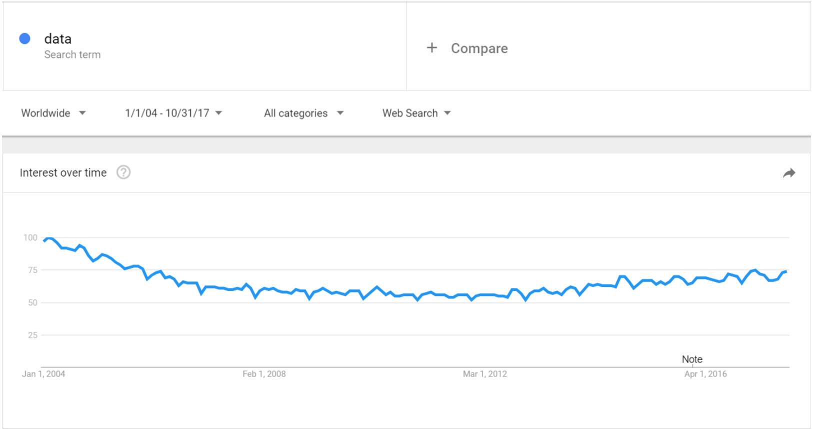
● data science
Search term

+ Compare

Worldwide ▾ 1/1/04 - 10/31/17 , 2004 - pr... ▾ All categories ▾ Web Search ▾

Interest over time

# Data Science Outpaces Data

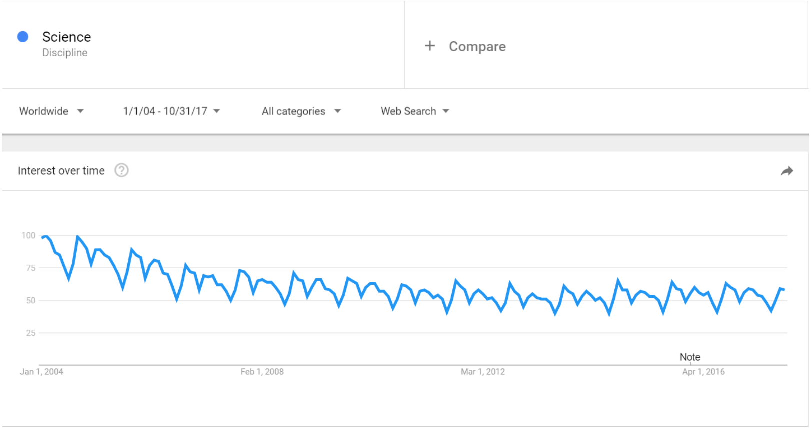


● data
Search term

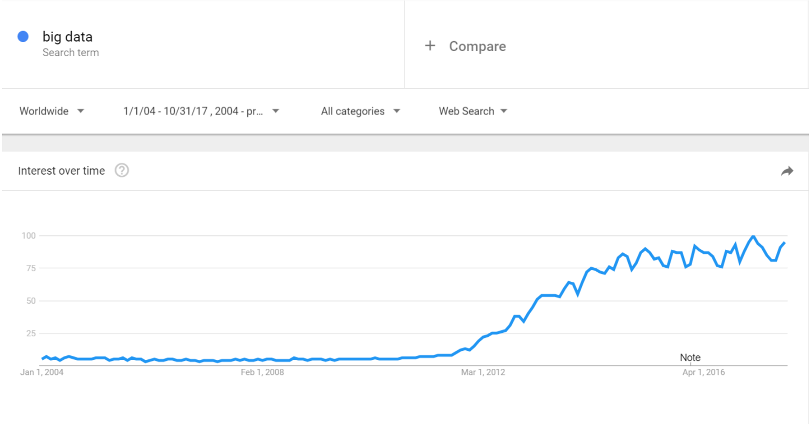
Worldwide ▾ 1/1/04 - 10/31/17 ▾ All categories ▾ Web Search ▾

Interest over time

# Science is not looking too good



# Driven by Desire to Capitalize on Growth in Data Sets

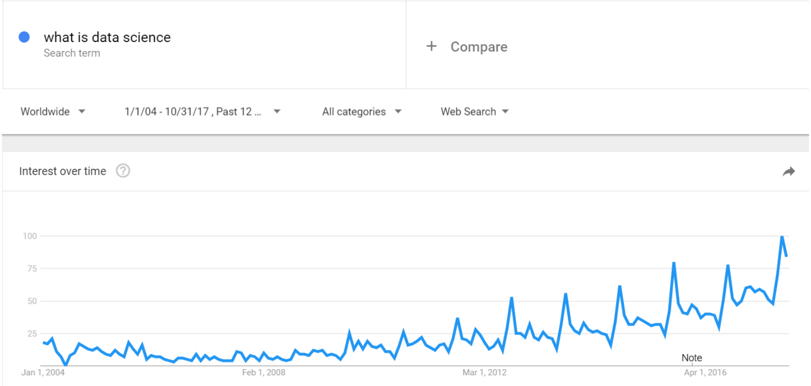


● big data
Search term

Worldwide ▾ 1/1/04 - 10/31/17 , 2004 - pr... ▾ All categories ▾ Web Search ▾

Interest over time

# Data Science as a term is getting very popular



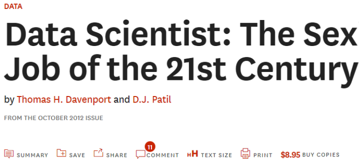
● what is data science
Search term

＋ Compare

Worldwide ▾ 1/1/04 - 10/31/17 , Past 12 ... ▾ All categories ▾ Web Search ▾

Interest over time

# Data Scientist as Job

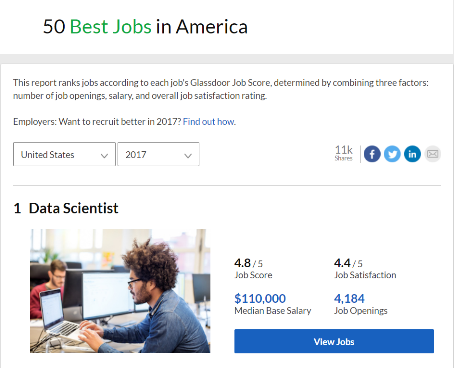


DATA

## Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE



50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? Find out how.

United States
2017

1 Data Scientist

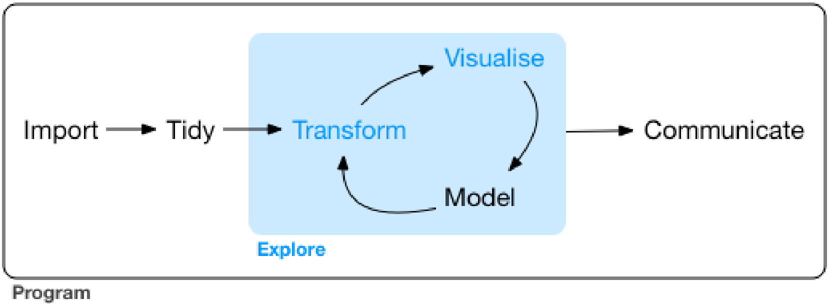
4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

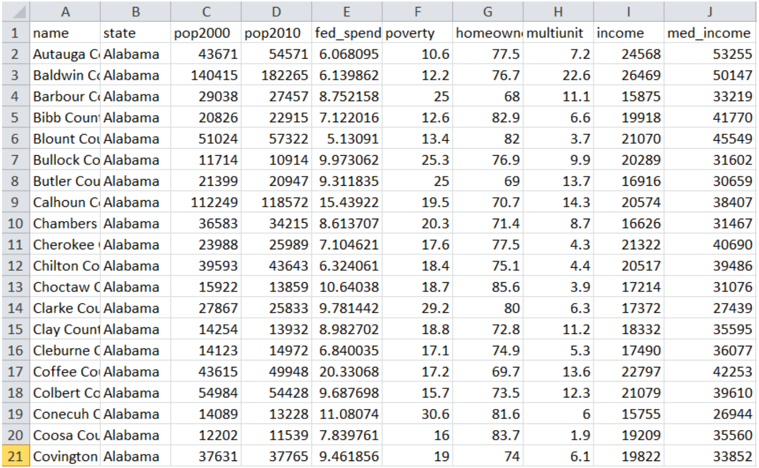
4,184
Job Openings

# What do they do?



- Image taken from “R for Data Science” by Grolemund and Wickham (free introduction to practical data science skills!)
- Your undergraduate days are a perfect time to acquire such practical skills. Could be helpful for employment and also very handy for analysis of scientific data.

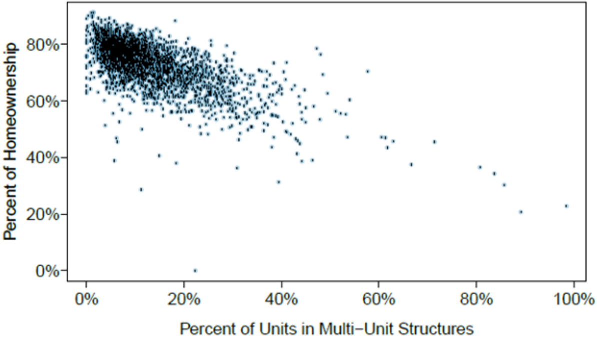
# 80% of the time spent Importing and Tidying Data



| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | name | state | pop2000 | pop2010 | fed_spend | poverty | homeown | multiunit | income | med_income |
| 2 | Autauga C | Alabama | 43671 | 54571 | 6.068095 | 10.6 | 77.5 | 7.2 | 24568 | 53255 |
| 3 | Baldwin C | Alabama | 140415 | 182265 | 6.139862 | 12.2 | 76.7 | 22.6 | 26469 | 50147 |
| 4 | Barbour C | Alabama | 29038 | 27457 | 8.752158 | 25 | 68 | 11.1 | 15875 | 33219 |
| 5 | Bibb Coun | Alabama | 20826 | 22915 | 7.122016 | 12.6 | 82.9 | 6.6 | 19918 | 41770 |
| 6 | Blount Cou | Alabama | 51024 | 57322 | 5.13091 | 13.4 | 82 | 3.7 | 21070 | 45549 |
| 7 | Bullock Co | Alabama | 11714 | 10914 | 9.973062 | 25.3 | 76.9 | 9.9 | 20289 | 31602 |
| 8 | Butler Cou | Alabama | 21399 | 20947 | 9.311835 | 25 | 69 | 13.7 | 16916 | 30659 |
| 9 | Calhoun C | Alabama | 112249 | 118572 | 15.43922 | 19.5 | 70.7 | 14.3 | 20574 | 38407 |
| 10 | Chambers | Alabama | 36583 | 34215 | 8.613707 | 20.3 | 71.4 | 8.7 | 16626 | 31467 |
| 11 | Cherokee ( | Alabama | 23988 | 25989 | 7.104621 | 17.6 | 77.5 | 4.3 | 21322 | 40690 |
| 12 | Chilton Co | Alabama | 39593 | 43643 | 6.324061 | 18.4 | 75.1 | 4.4 | 20517 | 39486 |
| 13 | Choctaw C | Alabama | 15922 | 13859 | 10.64038 | 18.7 | 85.6 | 3.9 | 17214 | 31076 |
| 14 | Clarke Cou | Alabama | 27867 | 25833 | 9.781442 | 29.2 | 80 | 6.3 | 17372 | 27439 |
| 15 | Clay Count | Alabama | 14254 | 13932 | 8.982702 | 18.8 | 72.8 | 11.2 | 18332 | 35595 |
| 16 | Cleburne C | Alabama | 14123 | 14972 | 6.840035 | 17.1 | 74.9 | 5.3 | 17490 | 36077 |
| 17 | Coffee Co | Alabama | 43615 | 49948 | 20.33068 | 17.2 | 69.7 | 13.6 | 22797 | 42253 |
| 18 | Colbert Co | Alabama | 54984 | 54428 | 9.687698 | 15.7 | 73.5 | 12.3 | 21079 | 39610 |
| 19 | Conecuh C | Alabama | 14089 | 13228 | 11.08074 | 30.6 | 81.6 | 6 | 15755 | 26944 |
| 20 | Coosa Cou | Alabama | 12202 | 11539 | 7.839761 | 16 | 83.7 | 1.9 | 19209 | 35560 |
| 21 | Covington | Alabama | 37631 | 37765 | 9.461856 | 19 | 74 | 6.1 | 19822 | 33852 |

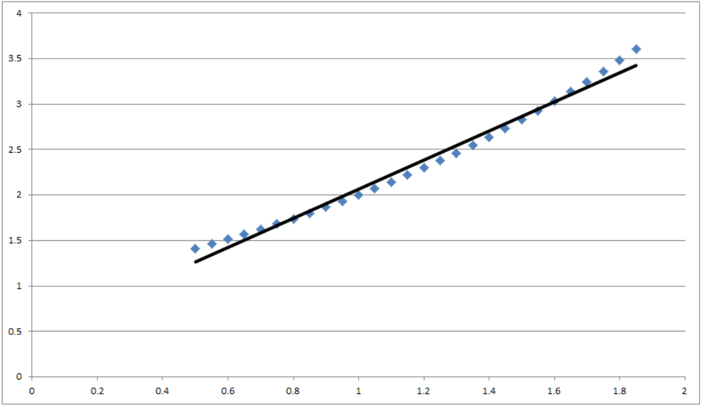
- From "OpenIntro Statistics" by Diez, Barr, Cetinkaya-Rundel.
- Columns: *variables* or *features*; Rows: *cases* or *examples*

# Visualizing



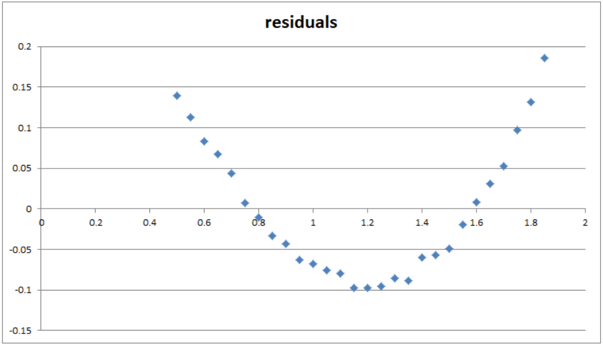
- From “OpenIntro Statistics” by Diez, Barr, Cetinkaya-Rundel.
- Scatterplots still the best for visualizing relationships.

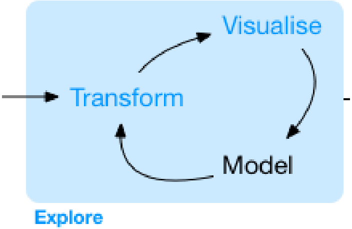
# Model: Mathematical Relationships



The most famous is simple linear regression, in which we try to find the line $y = b_0 + b_1 x$ that minimizes the sum of the squared errors for the data we are trying to fit.

# A Log Transformation was needed here



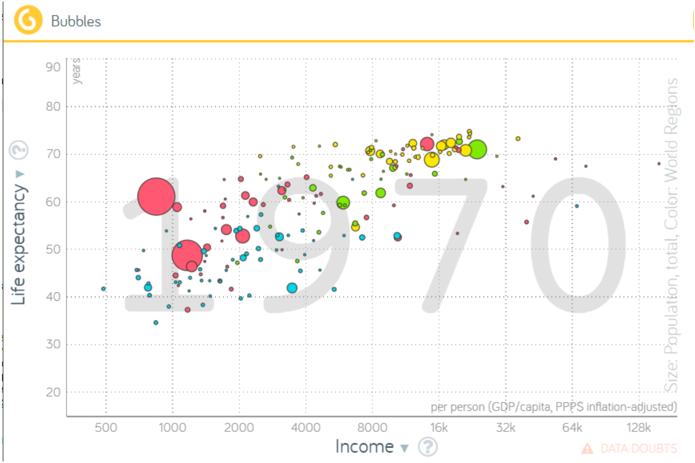

Visualise

Transform

Model

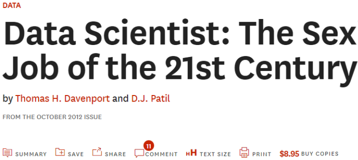
Explore

# Communicate



Take a look at this famous visualization of Gapminder. What transformation did he use on the x-axis and how does it change the story?

# So Far

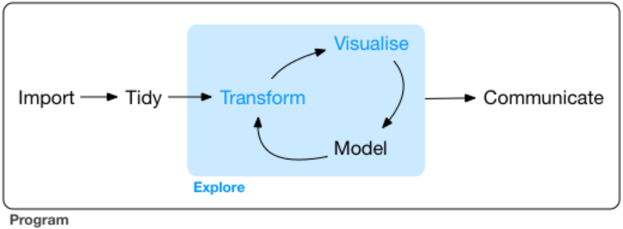


DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE



Employers looking for: **coding** skills, **math** skills, **hacking** together solutions skills

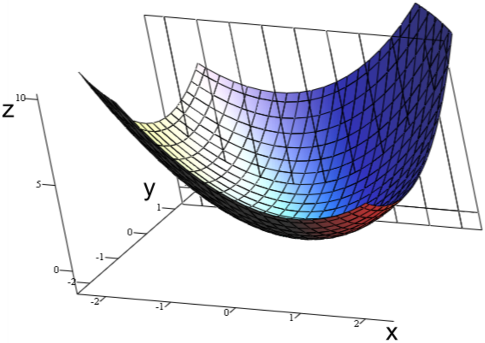
# What is Data Science?

**Using data to solve a problem.**

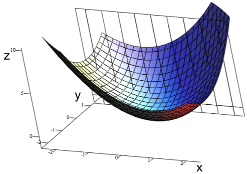
1. Using website traffic data to design a better website.
2. Using data on social network users to suggest contacts.
3. Using mobile phone data to track the formation of urban slums in developing countries.
4. Using text mining and sentiment analysis to see how the public feels about a stock in order to trade stocks.
5. Using a database of high level go play in order to make a machine capable of beating the world's best go players.
6. Using facial recognition software to identify individuals in order to pay for things.
7. Using ratings for previously seen movies to make suggestions for movies a person may like.
8. Using voice data to compile a national artificial intelligence to identify individuals by their voice.
9. Using brain activitity patterns to identify interesting components of the brain that function together.

# Simple Linear Regression

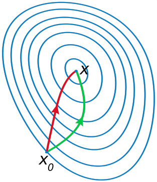
- Given finite data set: $(x_i, y_i)_{i=1}^n$.
- Find $b_0$ and $b_1$ so that $L(b_0, b_1) := \sum_{i=1}^n (y_i - b_1 x_i - b_0)^2$ is minimized.
- Notice that $L$ is a convex function. Therefore it has a unique minimum.



# Optimization as main tool!



Using the gradient, which is a generalization of the derivative to multiple dimensions, we can find a way to descend on the surface step by step. **Take Multivariable Calculus!**



Since our loss function $L(b_0, b_1)$ is convex, we will eventually reach the line of best fit. **Take Convex Optimization!**

# Stereotypical Prediction

1. The variable you want predicted $Y$ (say the price of Tesla stock tomorrow).
2. The features used to predict $X_1, X_2, \ldots, X_k$ (say the weather, the stock prices of a 100 different related stocks on the previous day, etc.)
3. The form of the prediction function and the parameters defining them $F_\theta : X_1 \times X_2 \cdots \times X_n \to Y$ (this varies for every kind of prediction strategy).
4. Large quantities of training data.
5. A loss function based on the data $L(\theta)$, which we are trying to minimize in order to find the best $F_\theta$.
6. An optimization algorithm for minimizing $L(\theta)$.
7. Validating the function on test data.

# Everything is a Long Vector

How to teach a robot to be able to recognize images as either a cat or a non-cat? This sounds like a biology problem. How can we formulate this as a mathematics problem?

# Everything is a Long Vector

How to teach a robot to be able to recognize images as either a cat or a non-cat? This sounds like a biology problem. How can we formulate this as a mathematics problem?

- $\mathbb{R}^{3\times 1000\times 1000}$ is a space of 1000 by 1000 rgb images

# Everything is a Long Vector

How to teach a robot to be able to recognize images as either a cat or a non-cat? This sounds like a biology problem. How can we formulate this as a mathematics problem?

- $\mathbb{R}^{3\times1000\times1000}$ is a space of 1000 by 1000 rgb images
- $C \subset \mathbb{R}^{3\times1000\times1000}$ is the cat subset.

# Everything is a Long Vector

How to teach a robot to be able to recognize images as either a cat or a non-cat? This sounds like a biology problem. How can we formulate this as a mathematics problem?

- $\mathbb{R}^{3\times 1000\times 1000}$ is a space of 1000 by 1000 rgb images
- $C \subset \mathbb{R}^{3\times 1000\times 1000}$ is the cat subset.
- Try to learn the classifier function $f_C : \mathbb{R}^{3000000} \to \{1, -1\}$ so that $f_C(x) = 1 \iff x \in C$.

# Everything is a Long Vector

How to teach a robot to be able to recognize images as either a cat or a non-cat? This sounds like a biology problem. How can we formulate this as a mathematics problem?

- $\mathbb{R}^{3\times 1000\times 1000}$ is a space of 1000 by 1000 rgb images
- $C \subset \mathbb{R}^{3\times 1000\times 1000}$ is the cat subset.
- Try to learn the classifier function $f_C : \mathbb{R}^{3000000} \to \{1, -1\}$ so that $f_C(x) = 1 \iff x \in C$.
- Let us play in a playground: `playground.tensorflow.org/`

# Everything is a Long Vector

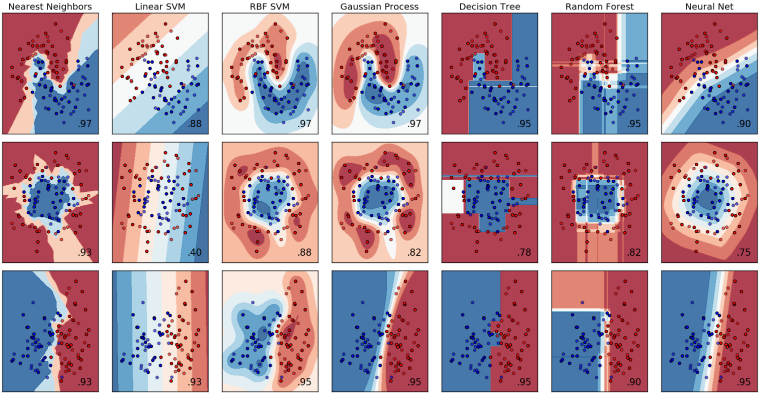
How to teach a robot to be able to recognize images as either a cat or a non-cat? This sounds like a biology problem. How can we formulate this as a mathematics problem?

- $\mathbb{R}^{3\times1000\times1000}$ is a space of 1000 by 1000 rgb images
- $C \subset \mathbb{R}^{3\times1000\times1000}$ is the cat subset.
- Try to learn the classifier function $f_C : \mathbb{R}^{3000000} \to \{1, -1\}$ so that $f_C(x) = 1 \iff x \in C$.
- Let us play in a playground: `playground.tensorflow.org/`

**Take Linear Algebra!**

# Many Different Kinds of Classifiers Out There

Helpful examples at
`http://scikit-learn.org/stable/index.html`
**Learn scikit-learn package of Python!**

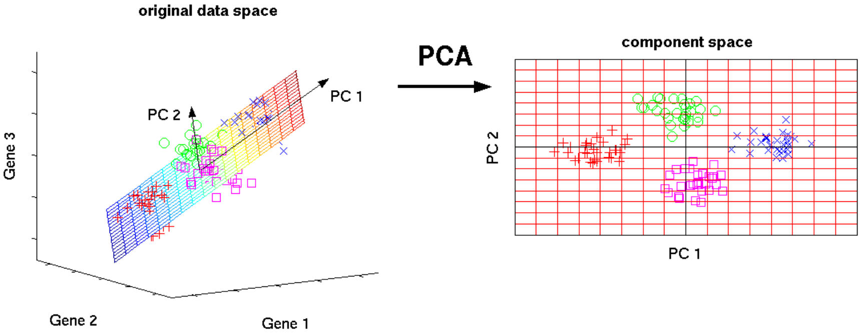


# Amazing Idea: Learning the Predictors $X_1, \ldots, X_k$

Say we want to classify $32 \times 32$ faces. That means 1024 features or dimensions. Hard problem! Curse of dimensionality.

# Amazing Idea: Learning the Predictors $X_1, \ldots, X_k$

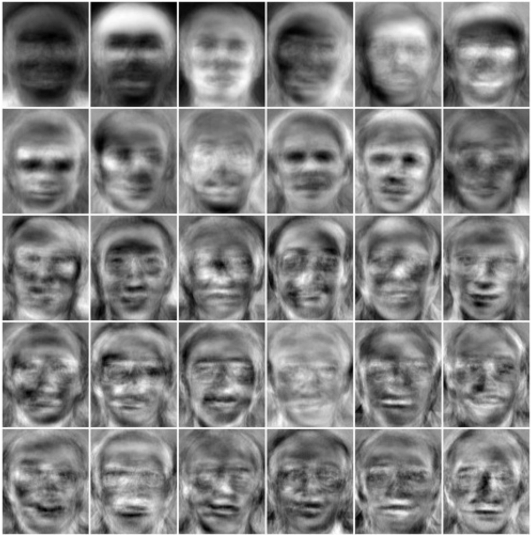
“Dimension Reduction” or “Representation Learning” **Take Linear Algebra!**



Mattias Scholz PhD Thesis 2006

# Amazing Idea: Learning the Predictors $X_1, \ldots, X_k$

$k$ Eigenfaces



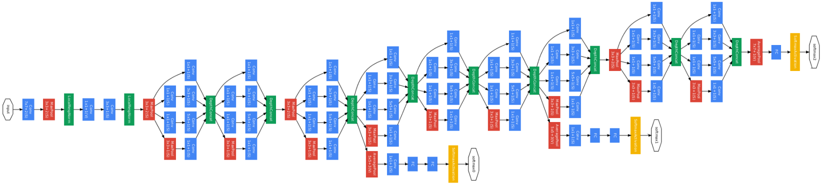
# Representation Learning + Prediction

Now we can classify faces:

- Raw images to Eigenface basis coordinates to Prediction
- $\mathbb{R}^{32\times 32} \to X_1 \times \ldots X_k \to Y$
- We learn the feature representation $F : \mathbb{R}^{32\times 32} \to X_1 \times \ldots X_k$ first.
- Then we learn classifier $X_1 \times \ldots X_k \to Y$.

# Several Layers of Feature Representations

Deep Learning



- From Szegedy et al. 2015.
- We don't really understand why it works, it is very hard to analyze non-convex heuristic optimization.

# Power of Representation Learning

- Vision: ImageNet classification with deep convolutional neural networks (2012), A. Krizhevsky et al.
- Language: Efficient estimation of word representations in vector space (2013), T. Mikolov et al
- Decision Making: Mastering the game of Go with deep neural networks and tree search (2016), D. Silver et al.
- The Representation can be reused for different tasks: CNN features off-the-Shelf: An astounding baseline for recognition (2014), A. Razavian et al.
- Unsupervised: Unsupervised representation learning with deep convolutional generative adversarial networks (2015), A. Radford et al.
- Art of Optimization: Training very deep networks (2015), R. Srivastava et al.

# Obligatory Slide on “Big Data"

How many images do you think we have?

# Obligatory Slide on “Big Data"

How many images do you think we have?

- 7 billion people, 3 billion people with smartphones, 1 picture a day = approximately 1 trillion pictures a year

# Obligatory Slide on “Big Data"

How many images do you think we have?

- 7 billion people, 3 billion people with smartphones, 1 picture a day = approximately 1 trillion pictures a year
- Some claim that more data was generated in the last 2 years than the rest of the history of mankind.

# Obligatory Slide on “Big Data"

How many images do you think we have?

- 7 billion people, 3 billion people with smartphones, 1 picture a day = approximately 1 trillion pictures a year
- Some claim that more data was generated in the last 2 years than the rest of the history of mankind.
- In comparison: there are around 3 billion seconds in a 100 year lifetime.

# Obligatory Slide on “Big Data"

How many images do you think we have?

- 7 billion people, 3 billion people with smartphones, 1 picture a day = approximately 1 trillion pictures a year
- Some claim that more data was generated in the last 2 years than the rest of the history of mankind.
- In comparison: there are around 3 billion seconds in a 100 year lifetime.
- Such deep representations can only be learned with such large data sets and massive computers (industry is outpacing academia).

# Obligatory Slide on “Big Data"

How many images do you think we have?

- 7 billion people, 3 billion people with smartphones, 1 picture a day = approximately 1 trillion pictures a year
- Some claim that more data was generated in the last 2 years than the rest of the history of mankind.
- In comparison: there are around 3 billion seconds in a 100 year lifetime.
- Such deep representations can only be learned with such large data sets and massive computers (industry is outpacing academia).
- If error = bias + variance, then we want a large and flexible class of functions so that bias is small since large enough data can control variance.

# Big Data and Mathematics

- Major technological advance of the last half century is information technology.

# Big Data and Mathematics

- Major technological advance of the last half century is information technology.
- The result is “Big Data.”

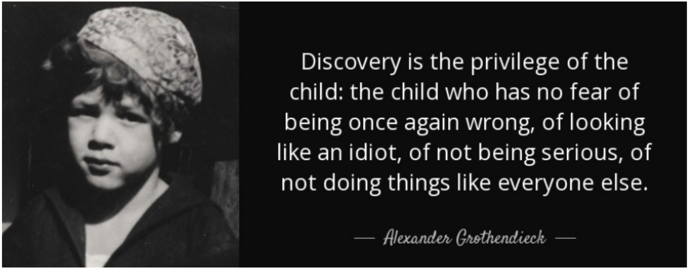
# Big Data and Mathematics

- Major technological advance of the last half century is information technology.
- The result is “Big Data.”
- Today, big data provides an opportunity to create AI; understand life and the mind; lay new foundations for computational sciences.

# Big Data and Mathematics

- Major technological advance of the last half century is information technology.
- The result is “Big Data.”
- Today, big data provides an opportunity to create AI; understand life and the mind; lay new foundations for computational sciences.
- For mathematicians, it is a chance to make discoveries on the order of the formulation of probability theory or calculus.

# Have fun!



Discovery is the privilege of the child: the child who has no fear of being once again wrong, of looking like an idiot, of not being serious, of not doing things like everyone else.

— Alexander Grothendieck —