



The Big Data Phenomenon

Socis promotores:



Contents

- Definitions and a bit of motivation:
 - Business Context
 - Big Data
 - Data Science
 - Artificial Intelligence: Machine Learning
- The “Data Science toolbox”
- The “Big Data” toolbox
- Market and industry
- Artificial Intelligence in the industry

01

Introduction

The value of Data

Making decisions based on data is nothing new. Now it is much easier, simply.

 **Sir William Davenant**
@SirWilliamD ⚙️ Segueix

The world before computers - staff sorting 4M used tickets from #London Underground to analyse line use in 1939.

← Respon → Retuitar ★ Marca com a preferit Pocket ... Més



RETUITES 105 PREFERITS 49

8:50 - 8 ag, 2014 Marca contingut

 **Old Pics Archive**
@oldpicsarchive ⚙️ Segueix

Computing Division at the Department of the Treasury, mid 1920s

← Respon → Retuitar ★ Marca com a preferit Pocket ... Més



RETUITES 264 PREFERITS 152

21:49 - 20 set, 2014

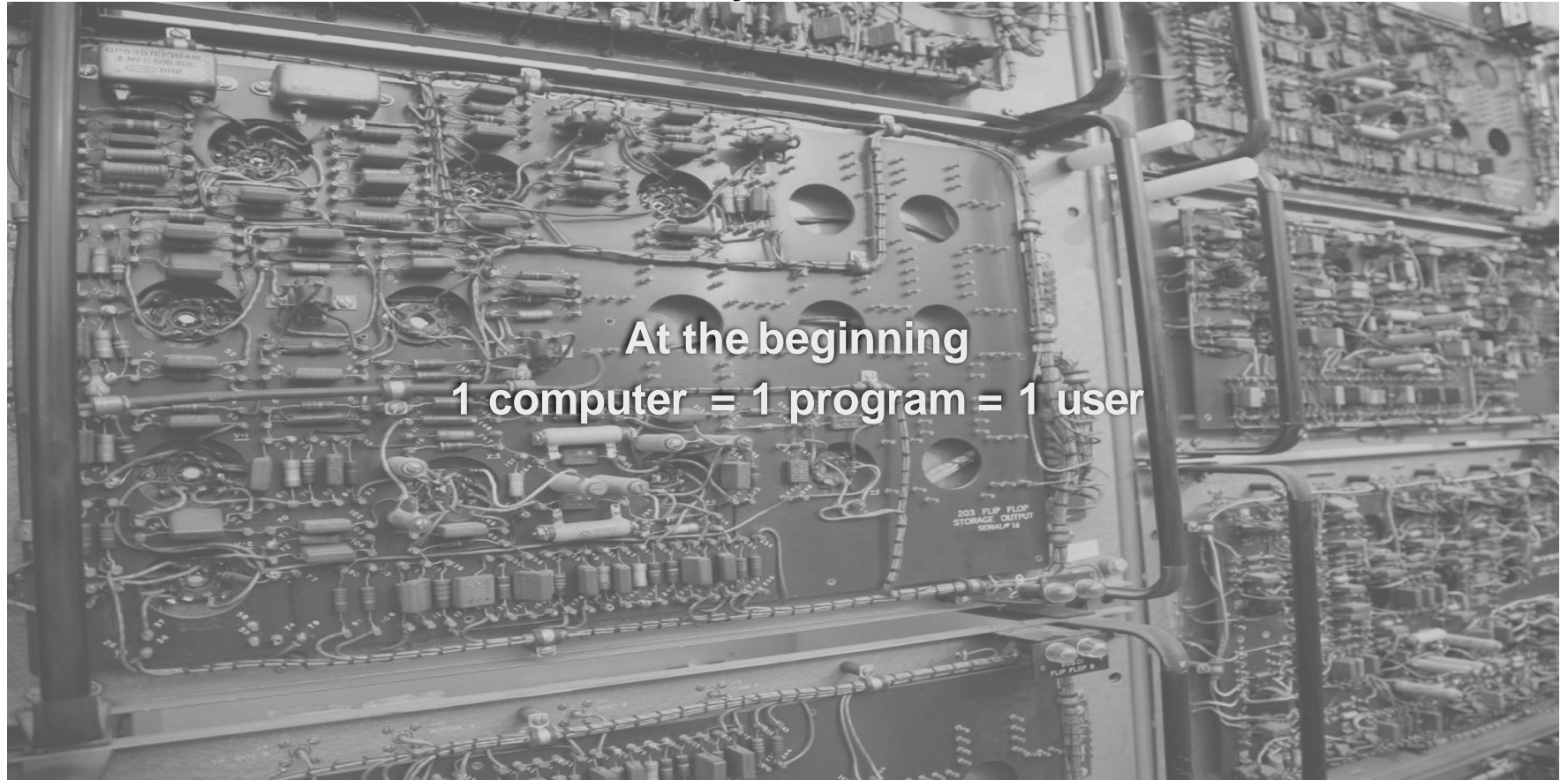
8:50 - 8 ag, 2014 Marca contingut

102 18

51:48 - 50 set, 2014

504 125

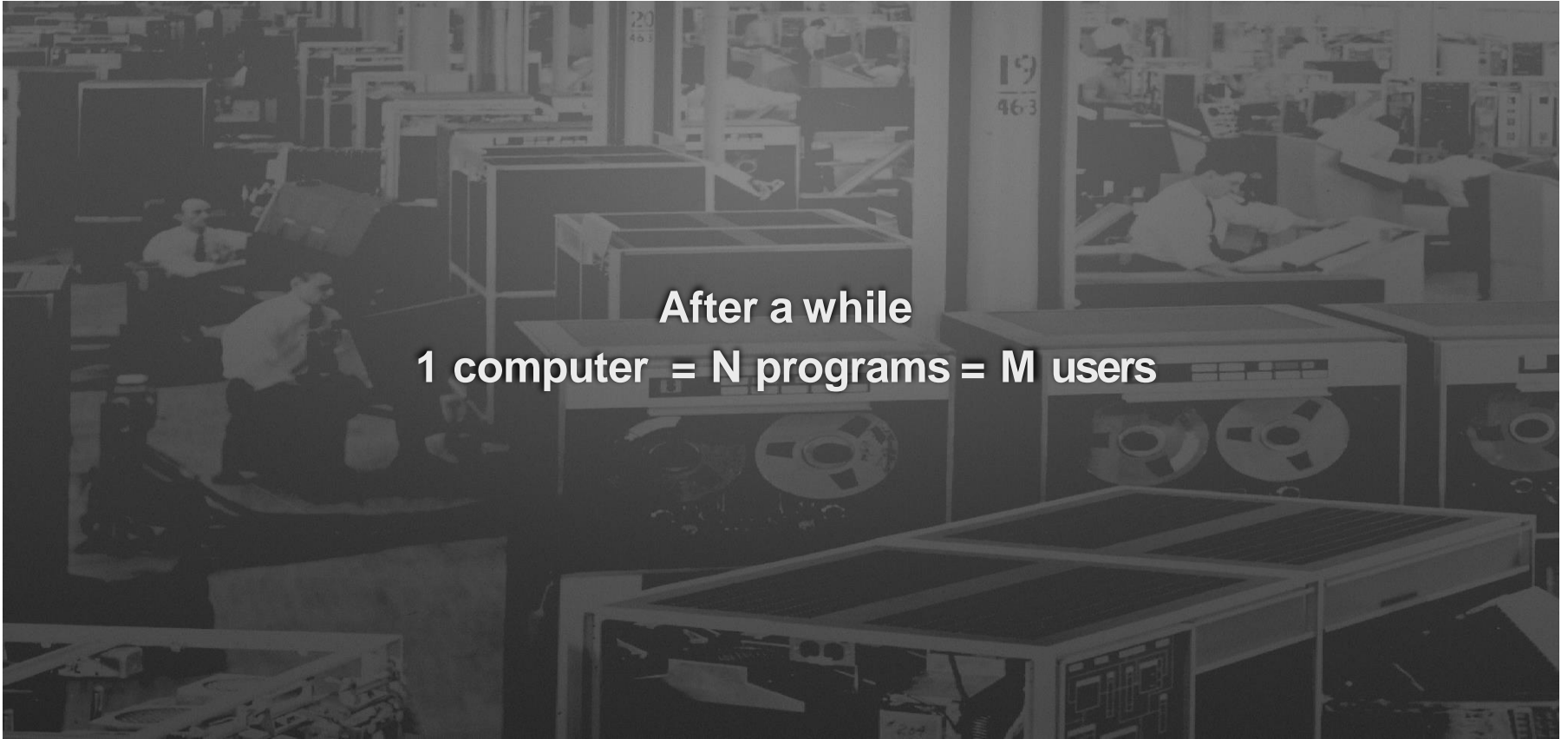
Why now?



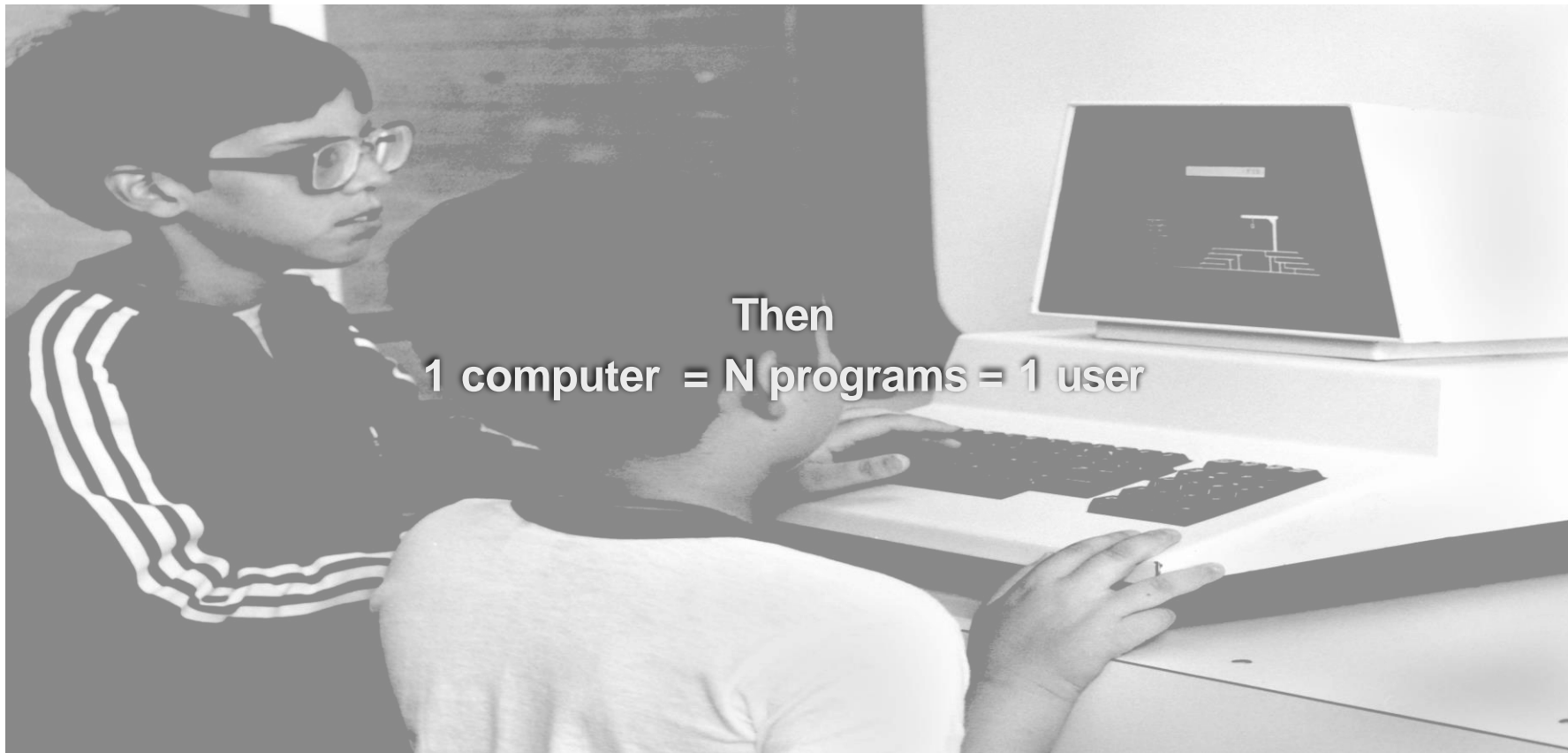
At the beginning
1 computer = 1 program = 1 user

Why now?

After a while
 $1 \text{ computer} = N \text{ programs} = M \text{ users}$



Why now?



Then

1 computer = N programs = 1 user

Why now?

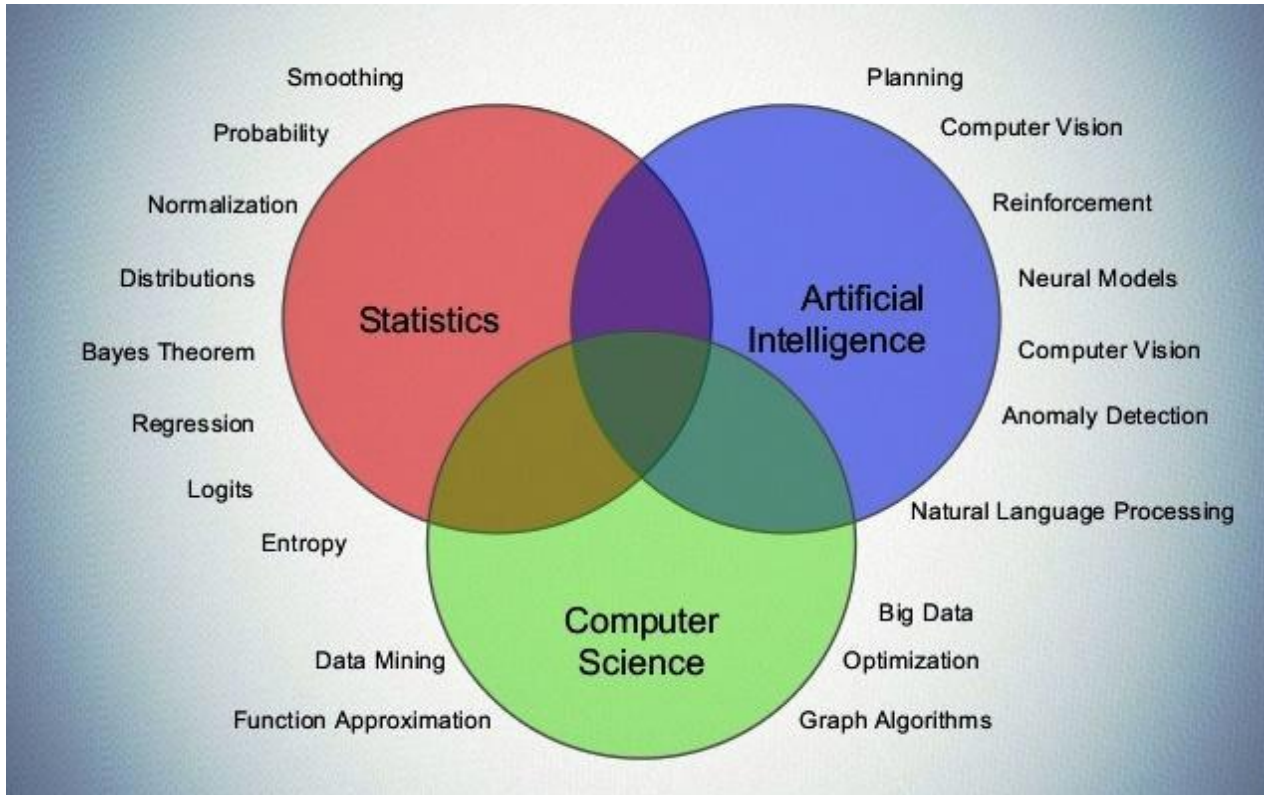
A few years ago we reach the present situation.

From a user perspective:

M computers = N programs = 1 user



A bit of terminology





Data Science
Big Data

Big Data

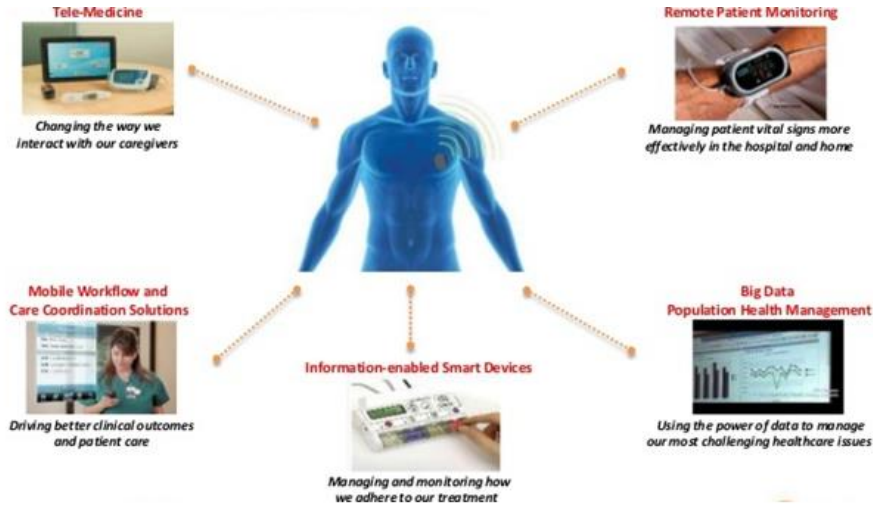
What is Big Data?

- For some people, they have big data when its size $> 65536 \times 256$.
- In general we have big data when its size does not allow its storage and analysis in a big computer.



Big Data

Wal-Mart handles over one million customer transaction per hour, the information is stored on a database sized in excess of 2.5 Petabytes ($2,0 \times 10^{16}$ bits).



By 2016 it is likely that a typical hospital will create 665 terabytes (5.32×10^{15} bits) of data a year.



Big Data



With a personal computer:

- You can find an element in a 1 MB file in less than a second.
- You can find an element in a 1 GB file in less than a minute.
- You can find an element in a 1 TB file in less than sixteen hours.
- You can find an element in a 1 PB file in less than two years.
- You can find an element in a 1 EB file in less than two thousand years.

Big Data

Big data is more than size.

It is commonly characterized with four V



Big Data

The cloud is key to deal with the three V, but the main phenomenon behind Big Data is **datification**.

Key enabler

The three V are a consequence of it.



Big Data

We are rendering into data many aspects
of the world that have never been
quantified before:

business networks

books I'm reading

location

physical activity

consumed food

purchases

physiological signals

straight thoughts

friendship

gaze

driving behavior

1 THE RAPID GROWTH OF GLOBAL DATA



The production of data is expanding at an astonishing pace. Experts now point to a 4300% increase in annual data generation by 2020. Drivers include the switch from analog to digital technologies and the rapid increase in data generation by individuals and corporations alike.

■ Size of Total Data
■ Enterprise Created Data
■ Enterprise Managed Data

2020: MORE THAN 1/3 OF THE DATA PRODUCED WILL LIVE IN OR PASS THROUGH THE CLOUD.

2012: CUSTOMERS WILL START STORING 1 EB OF INFORMATION.



WHAT IS A ZETTABYTE?

1,000,000,000,000 gigabytes
1,000,000,000,000 terabytes
1,000,000,000,000 petabytes
1,000,000,000,000 exabytes
1,000,000,000,000 zettabyte



1 terabyte holds the equivalent of roughly 210 single-sided DVDs.



It took roughly 1 petabyte of local storage to render the 3D CGI effects in Avatar.



In 2007, the estimated information content of all human knowledge was 295 exabytes.

DATA PRODUCTION WILL BE 44 TIMES GREATER IN 2020 THAN IT WAS IN 2009

More than 70% of the digital universe is generated by individuals. But enterprises have responsibility for the storage, protection and management of 80% of it.*

90% of world data was generated between 2012 and 2015

Big Data

Information comes from:

- Corporate Data Bases (structured information). Unstructured information in documents, Wikipedia, textbooks, journals, blogs, tweets, etc.
- Images in the web, public cameras, phones, TV, YouTube, etc.
- Public APIs: smart cities, government, search engines, etc.
- Sensor Data: GPS, accelerometer, physico- chemical sensors, sociometric sensors, super-colliders, telescopes, etc.

Big Data Characteristics

Volume	Variety	Velocity
<ul style="list-style-type: none">• Records• Pictures• Videos• Terabyte	<ul style="list-style-type: none">• Structured• Semi-structured• Unstructured	<ul style="list-style-type: none">• Batch• Stream• Realtime Processing

Data Science

Technology is the collection of tools, including machinery, modifications, arrangements and procedures used by humans.

Big Data is a key **technology** to process massive amounts of data (f.e. to count items).

Methodology is the systematic, theoretical analysis of the methods applied to a field of study.

Data Science is a **methodology** to define what we want to do with data, how do we evaluate our actions, what decisions can be grounded on data, how do we combine evidences from several sources, etc.

What are the limits of data science?

- Data science is a tool to inform, not to explain.
- Data science cannot substitute intuition or creativity.

If I had asked people what they wanted,
they would have said faster horses.
Henry Ford.

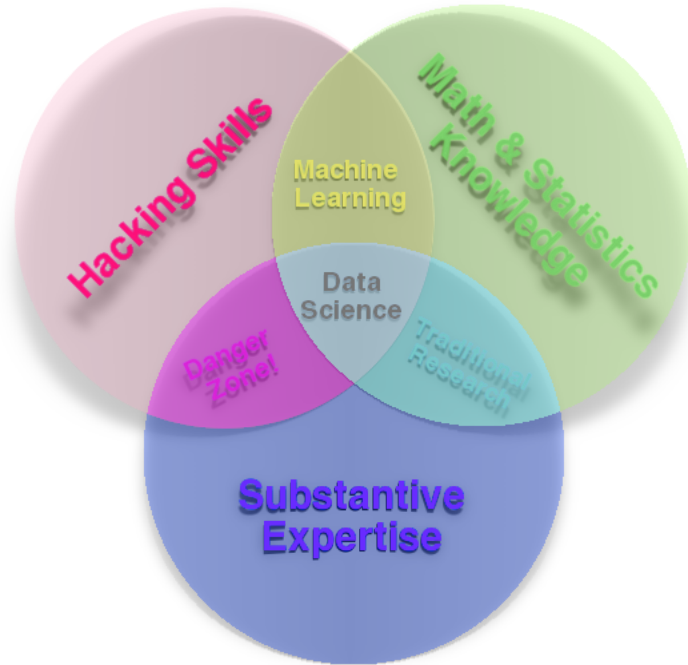
Data Science



**THE SEXIEST JOB OF
THE 21TH CENTURY.**

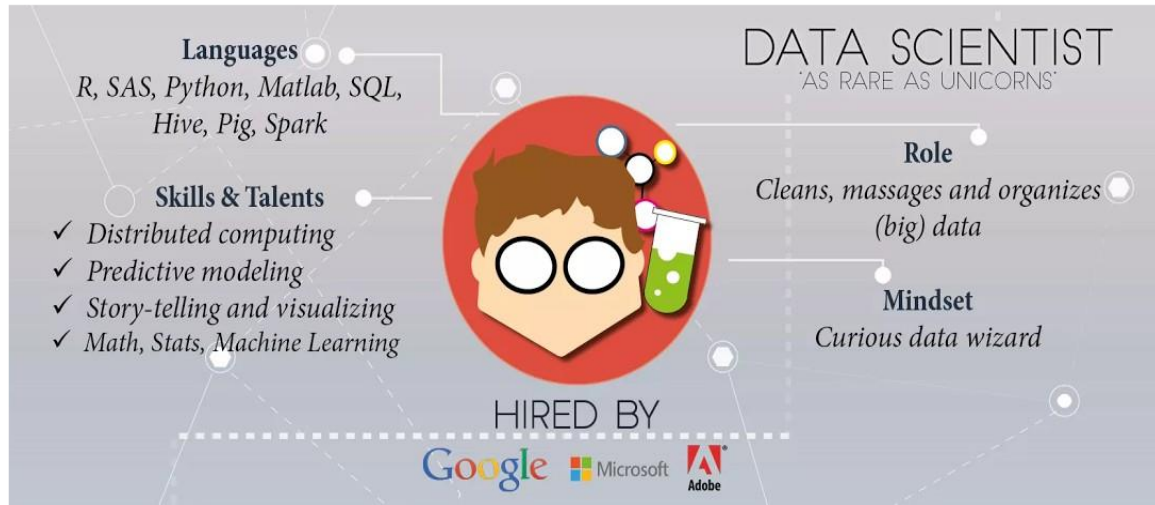
HARVARD BUSINESS REVIEW,
OCT. 2012

Data Science



Drew Conway's Data Science Venn Diagram

Big Data Roles



Big Data Roles



DATA ANALYST "DATA DETECTIVE"

Role

Collects, processes and performs statistical data analyses

Mindset

Intuitive data junkie with high "figure-it-out" quotient



Languages

R, Python, HTML, Javascript, C/C++, SQL

Skills & Talents

- ✓ *Spreadsheet tools (e.g. Excel)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Communication & visualization*
- ✓ *Math, Stats, Machine Learning*

HIRED BY



Big Data Roles

DATA ENGINEER

SOFTWARE ENGINEERS BY TRADE

Role

Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset

All-purpose everyman



HIRED BY



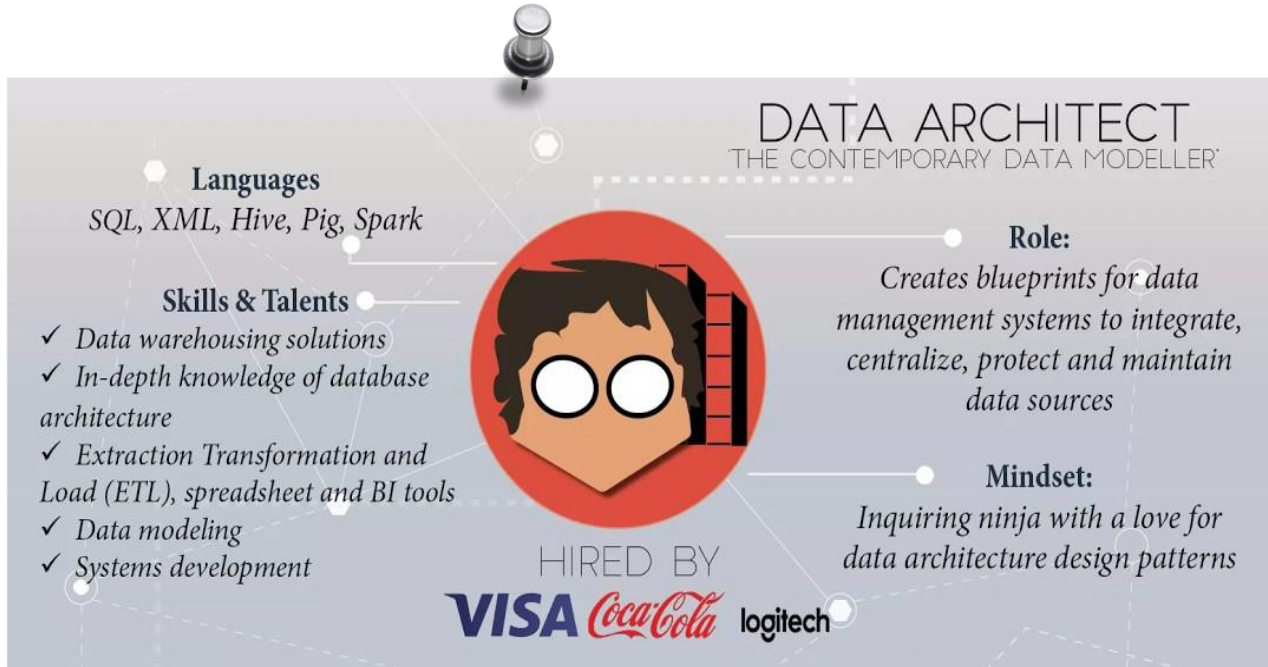
Languages

SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents

- ✓ *Database systems (SQL & NO SQL based)*
- ✓ *Data modeling & ETL tools*
- ✓ *Data APIs*
- ✓ *Data warehousing solutions*

Big Data Roles



02

**The “Data Science”
Toolbox**

Data Science

Maths and Statistics

Descriptive
statistics

Linear Algebra

Numerical
analysis

Optimization

Bayesian
probability
models

Data Science

Programming skills

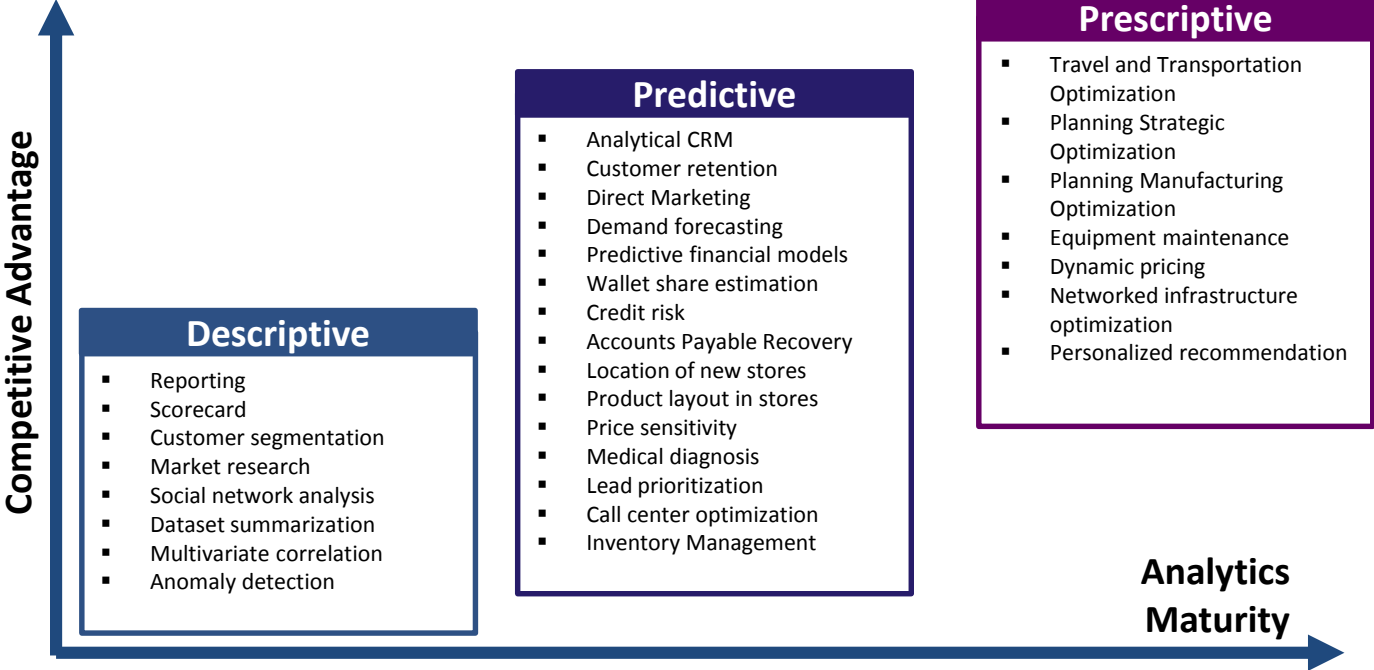
- **Algorithm prototyping**
- **Programming languages for prototyping**
 - ✓ **R**
 - ✓ **Python**
 - ✓ **Matlab**
 - ✓ **Julia**
 - ✓ **Java**
 - ✓ **Scala**
- **Big Data Tools: Hadoop, Spark, Amazon WS, Kafka, etc.**

Data Science

Techniques

- Classification and class probability
- Regression
- Similarity matching
- Clustering
- Co-occurrence grouping
- Profiling
- Data reduction
- Casual modeling
- A/B testing

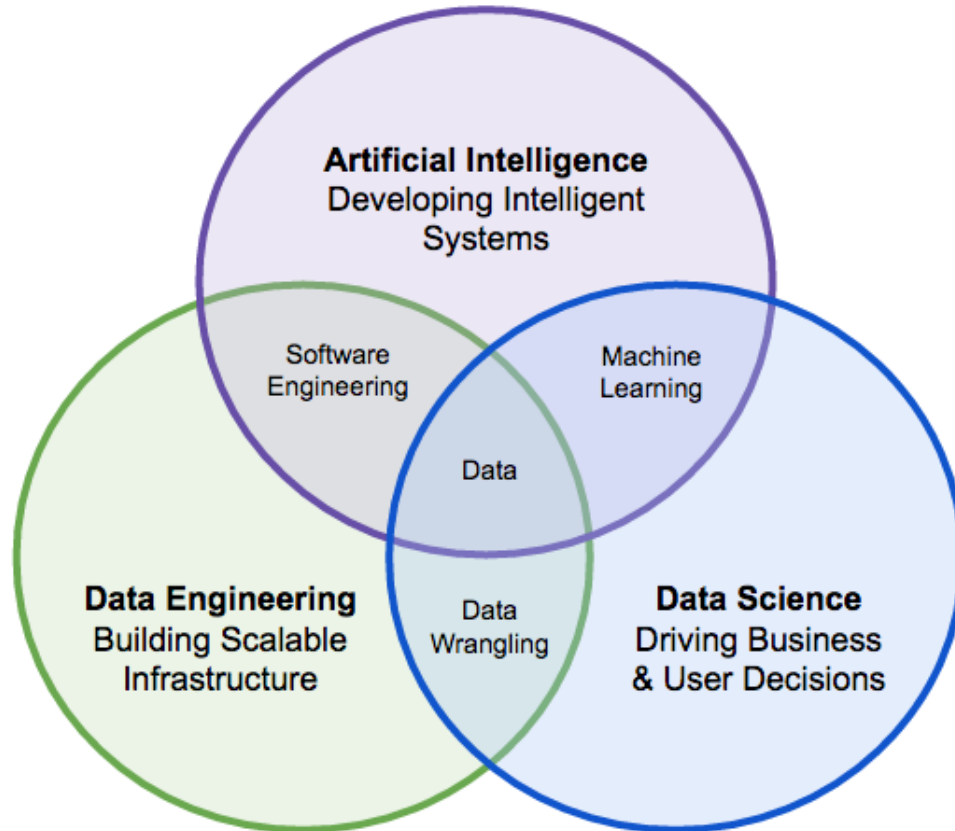
Data Analytics Capabilities



02

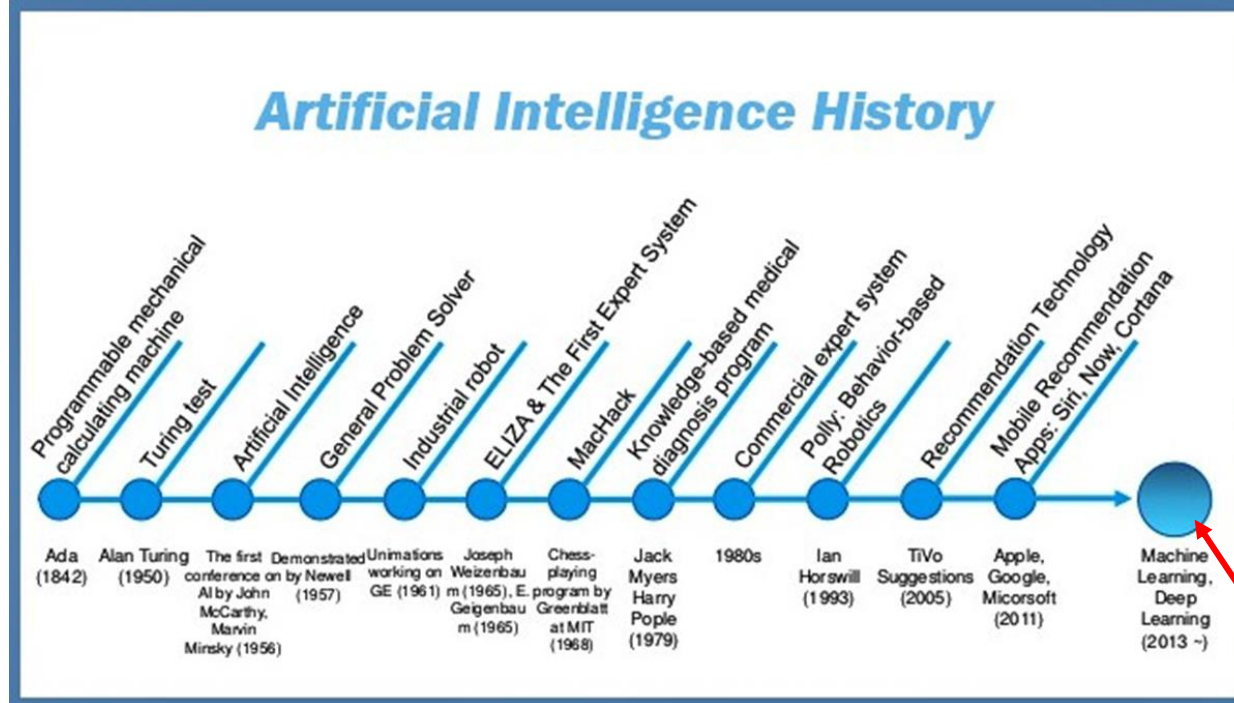
**Artificial Intelligence
and Big Data**

How is Artificial Intelligence related to Data Science and Big Data?



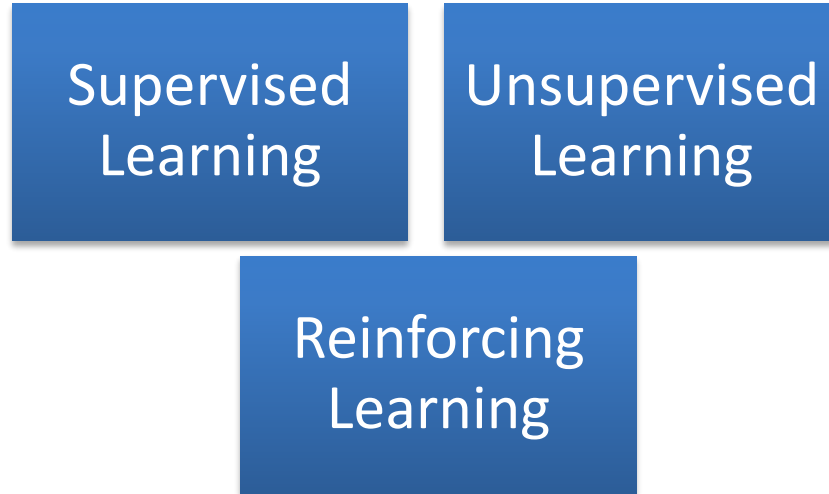
Artificial Intelligence is nothing new

Artificial Intelligence is nothing new...

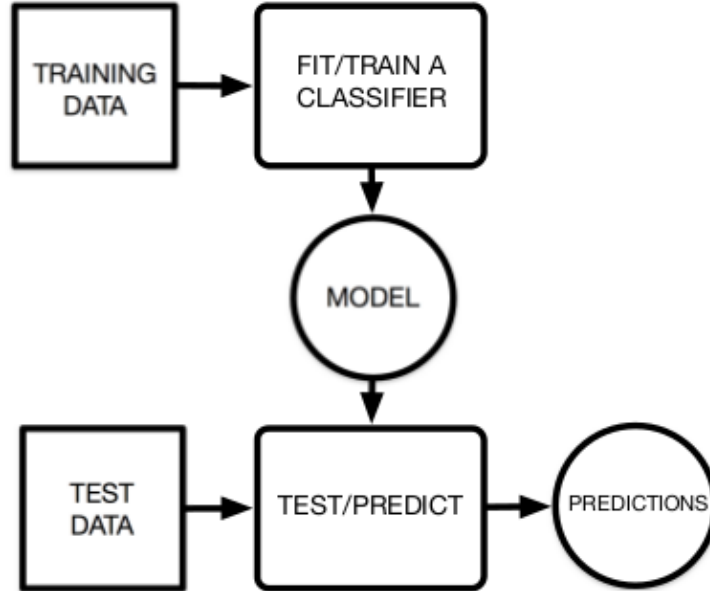


NOW!

Machine learning



Machine learning workflow





Historical metaphors of the brain:
Hydraulic (blood cooler, spirits),
Mechanical (clock, steam machine),...



- In 1943, neurophysiologist **Warren McCulloch** and mathematician **Walter Pitts** wrote a paper on how neurons might work. In order to describe how neurons in the brain might work, they modeled a simple neural network using **electrical circuits**.



In 1949, Donald **Hebb** wrote *The Organization of Behavior*, a work which pointed out the fact that **neural pathways are strengthened each time they are used**, a concept fundamentally essential to the ways in which humans **learn**.

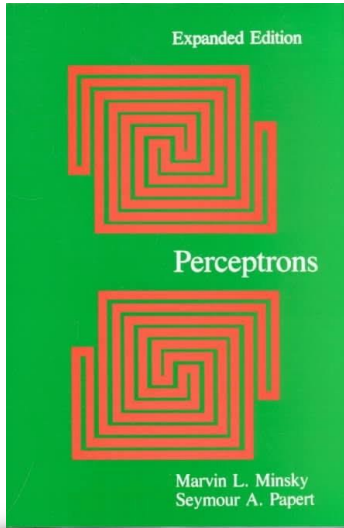
$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$



In 1957 **Frank Rosenblatt** attempted to build a kind of mechanical brain called the **Perceptron**, which was billed as “*a machine which senses, recognizes, remembers, and responds like the human mind*”.



A critical book written in 1969 by **Marvin Minsky** and his collaborator **Seymour Papert** showed that Rosenblatt's original system was **painfully limited**, literally blind to some simple logical functions like "*exclusive-or*".



It is claimed that pessimistic predictions made by the authors were responsible for an erroneous change in the direction of research in AI, concentrating efforts on so-called "symbolic" systems, and contributing to the so-called AI winter. This decision, supposedly, proved to be unfortunate in the 1980s, when new discoveries showed that the prognostics in the book were wrong.

Source: Wikipedia

70's: First neural network winter





In 1982, interest in the field was renewed. **John Hopfield** of Caltech presented a paper to the National Academy of Sciences. His approach was to create more useful machines by using **bidirectional lines**. Previously, the connections between neurons was only one way.



In 1986, the problem was how to **extend the Widrow-Hoff rule to multiple layers**. Three independent groups of researchers, which included **David E. Rumelhart**, **Geoffrey E. Hinton** and **Ronald J. Williams**, came up with similar ideas which are now called **back-propagation** networks because it distributes pattern recognition errors throughout the network.



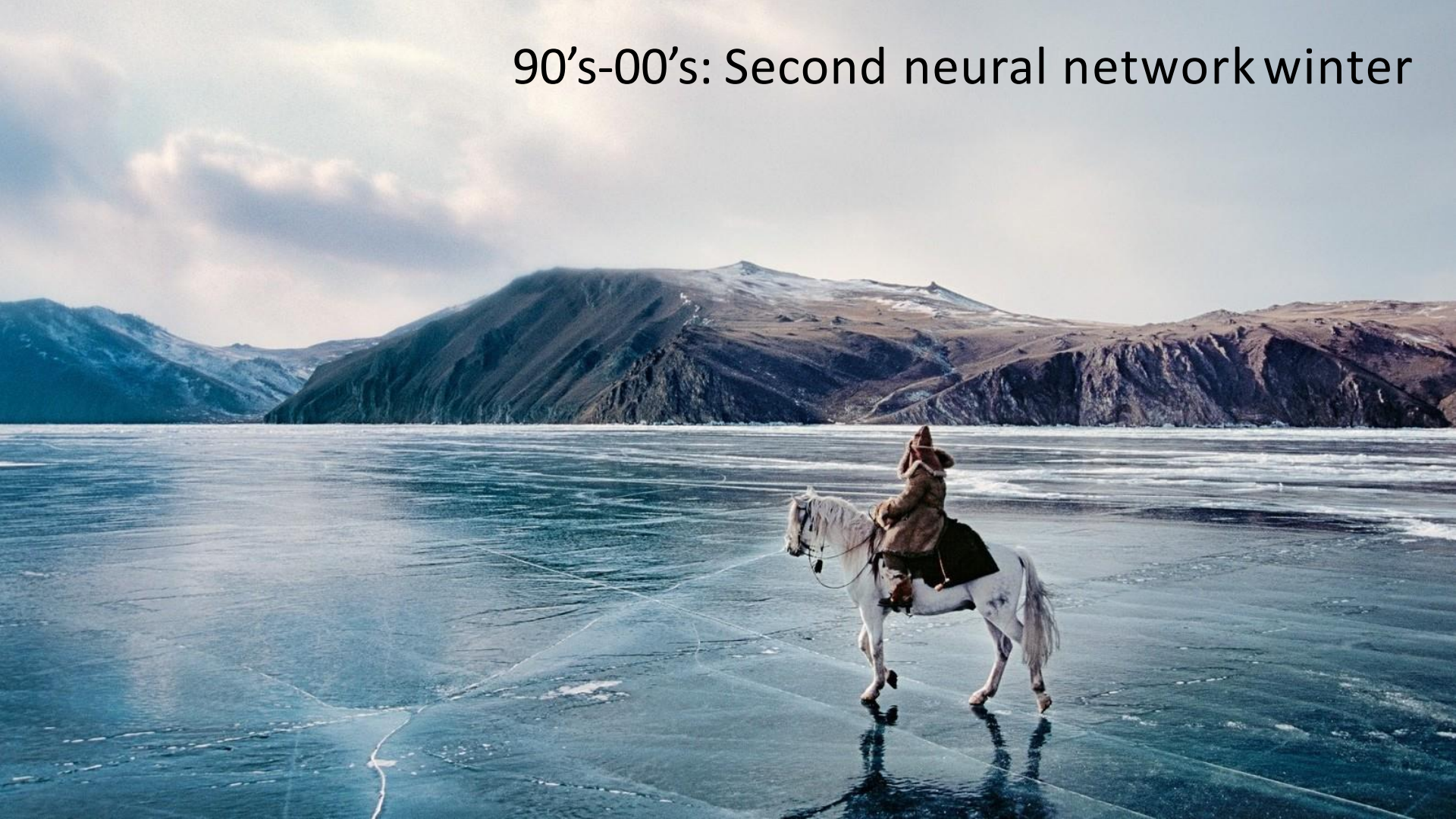
From 1986 to mid 90's new developments arised:
convolutional neural networks (**Y.LeCun**), unsupervised
learning (**Y.Bengio**), RBM(**G.Hinton**), recurrent networks
(**J.Schmidhuber**), etc.



But, by this point **new machine learning methods**
had begun to also emerge, and people were again
beginning to be skeptical of neural nets since they
seemed so intuition-based and since computers
were still barely able to meet their computational
needs.



90's-00's: Second neural network winter

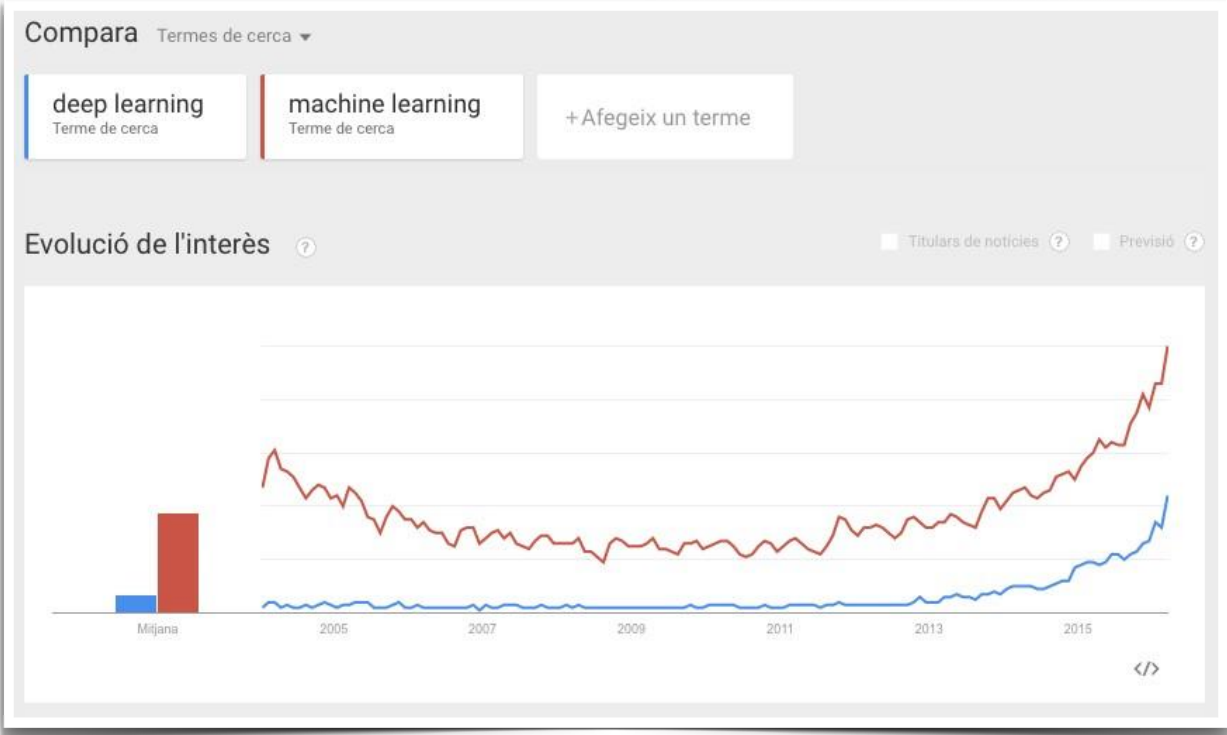




With the ascent of Support Vector Machines and the **failure of backpropagation**, the early 2000s were a dark time for neural net research.

- Then, what every researcher must dream of actually happened: G.Hinton, S.Osindero, and Y.W.Teh published a paper in 2006 that was seen as a breakthrough, a breakthrough significant enough to rekindle interest in neural nets: *A fast learning algorithm for **deep** belief nets.*
- After that, following Moore's law, computers got dozens of times faster (GPUs) since the slow days of the 90s, making learning with large datasets and many layers much more tractable.

Neural Networks Reborn

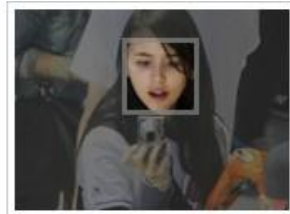


Google Trends

- NN and DL currently provide the best solutions to many problems in image recognition, speech recognition, and natural language processing.



Face recognition.



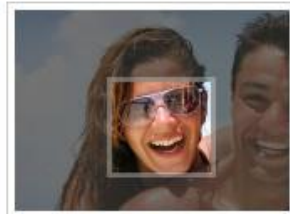
Who is this?



Who is this?



Who is this?



Who is this?



Who is this?



Who is this?

DeepFace (Facebook): Accuracy of 97.35%

New applications: navigation and mapping.

The image shows a screenshot of the Dyson website's product page for the Robot Dyson 360 Eye. At the top left is the Dyson logo. To its right is a navigation menu with links for 'Tienda', 'Aspiradoras', 'Ventiladores y Calefactores', 'Airblade™', 'Mi cuenta', and 'Soporte'. A globe icon is located in the top right corner. Below the navigation is the product name 'Robot Dyson 360 Eye™' and a yellow button that says 'Sea el primero en disfrutarlo'. The main heading reads 'El nuevo robot aspirador de Dyson'. On the left, there is a circular video thumbnail with a play button icon and the text 'Vea a James Dyson presentando el nuevo Dyson 360 Eye™ en Tokio'. The central focus is a large, detailed image of the robot vacuum, which is a cylindrical, silver-colored device with a blue base and a transparent top section showing internal components.

dyson

Tienda Aspiradoras Ventiladores y Calefactores Airblade™ Mi cuenta Soporte

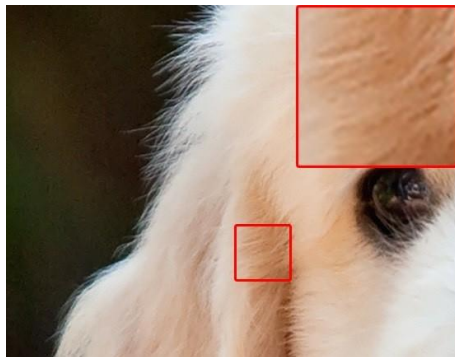
Robot Dyson 360 Eye™

Sea el primero en disfrutarlo

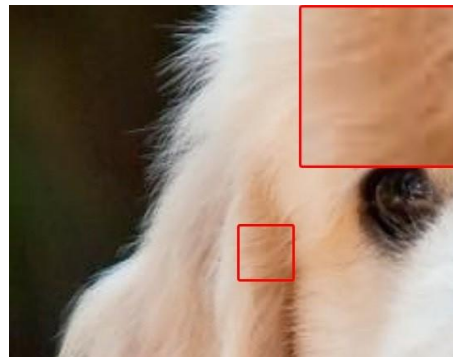
El nuevo robot aspirador de Dyson

Vea a James Dyson presentando el nuevo Dyson 360 Eye™ en Tokio

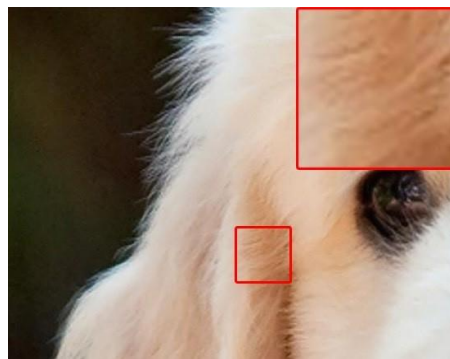
New applications: Image Upscaling (Flipboard)



Original

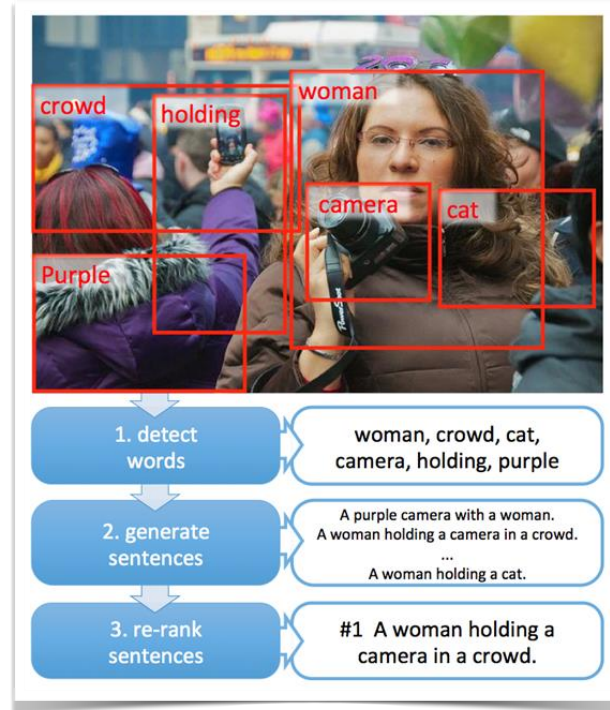


Bicubic



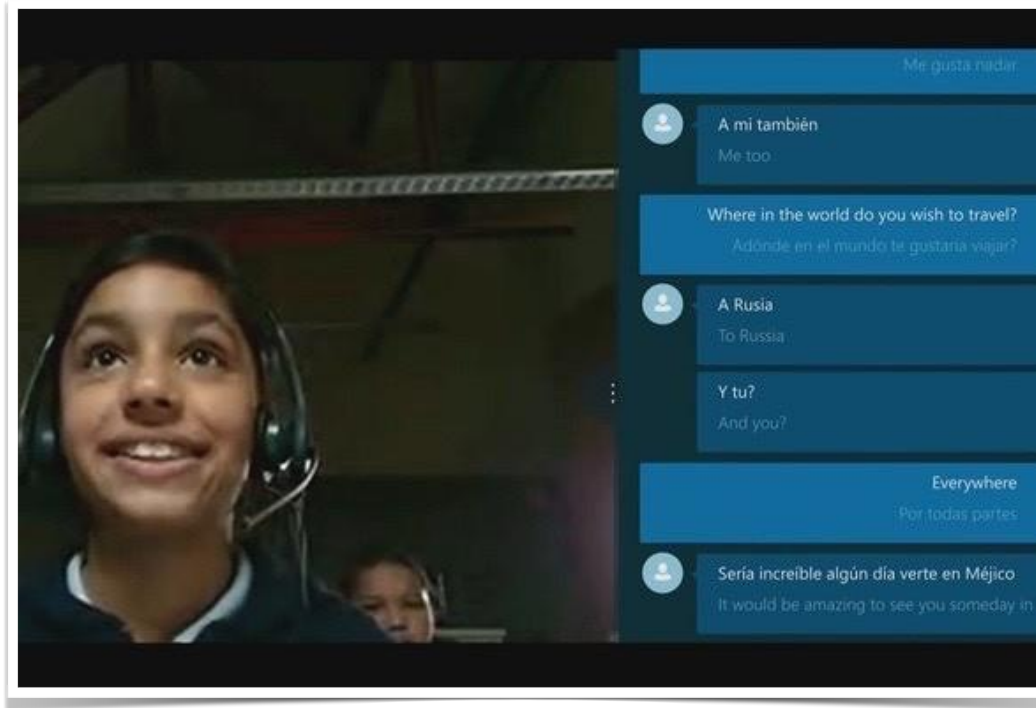
Model

New applications: Automatic Image Captioning

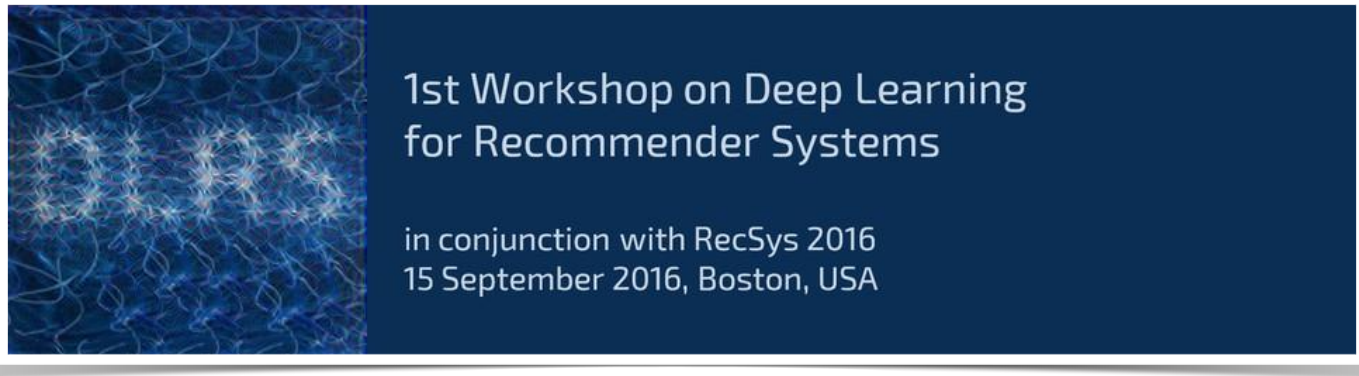


<http://blogs.technet.com/b/machinelearning/archive/2014/11/18/rapid-progress-in-automatic-image-captioning.aspx>

Speech translation



Recommenders



1st Workshop on Deep Learning
for Recommender Systems

in conjunction with RecSys 2016
15 September 2016, Boston, USA

Music Generation

The screenshot shows the SoundCloud profile for 'deepjazz'. The profile banner features a black and white photograph of a jazz band. The profile name is 'deepjazz' with a bio that reads 'I'm an AI built to make Jazz' and 'Princeton, United States'. The profile picture is a circular logo with the letters 'dj'. Below the banner, there are navigation tabs for 'All', 'Tracks', 'Playlists', and 'Reposts', along with 'Follow' and 'Share' buttons. The main content area displays a track titled 'deepjazz On Metheny' with a waveform and a duration of 0:53. Below the track, there is a list of three tracks generated by the AI, each with a play button and a view count. The right sidebar shows the user's statistics: 104 followers, 1 following, and 6 tracks. A bio paragraph explains that 'deepjazz' is an AI built by Ji-Sung Kim and provides links to its source code on GitHub and its website, deepjazz.io. There is also a '1 following' indicator and a 'View all' link.

SOUNDCLOUD Charts Search for artists, bands, tracks, podcasts Sign in or Create account Upload

deepjazz
I'm an AI built to make Jazz
Princeton, United States

dj

All Tracks Playlists Reposts Follow Share

dj 6 tracks

deepjazz 14 days
deepjazz On Metheny #Electronic

1 0:53

dj 1 deepjazz On Metheny ... 1 Epoch	▶ 6,142
dj 2 deepjazz On Metheny ... 16 Epochs	▶ 3,452
dj 3 deepjazz On Metheny ... 32 Epochs	▶ 1,908

Followers 104 Following 1 Tracks 6

Hi! I'm deepjazz, an AI built by Ji-Sung Kim. You can check out my source code on GitHub or visit my website, deepjazz.io

[my source code \(GitHub!\)](#)
[deepjazz.io](#)

1 following View all

Go



nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

At last — a computer program that
can beat a champion Go player **PAGE 484**

ALL SYSTEMS GO

CONSERVATION

**SONGBIRDS
À LA CARTE**
*Illegal harvest of millions
of Mediterranean birds*
PAGE 616

RESEARCH ETHICS

**SAFEGUARD
TRANSPARENCY**
*Don't let openness backfire
on individuals*
PAGE 619

POPULAR SCIENCE

**WHEN GENES
GOT 'SELFISH'**
*Darwin's 'calling
card' forty years on*
PAGE 412

NATURE.COM/NATURE

28 January 2016 £10
Vol. 529, No. 7587



Start Ups



vicarious home about news investors join contact

our mission:
build the next generation of A.I. algorithms

learn more join our team

NEWS

- Vicarious announces \$40M Series B Mar 25, 2014
- Vicarious passes first Turing Test - CAPTCHA Oct 28, 2013
- Vicarious announces \$15M Series A Aug 21, 2012
- Vicarious presents at Peter Thiel's Stanford Class Jun 5, 2012



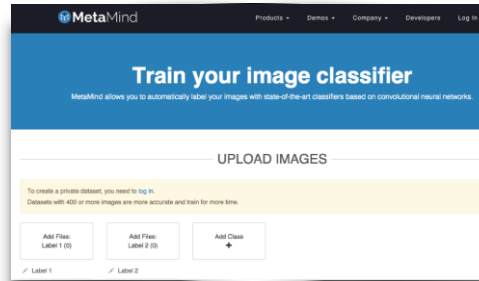
Dato Create Intelligence PRODUCTS USES LEARN EVENTS COMPANY BLOG DOWNLOAD

MACHINE LEARNING

that scales with your business

DOWNLOAD GRAPH LAB CREATE

- Ultra-Fast Data Analytics**
Wrangle terabytes of data at interactive speeds, even on your laptop. With DFrame, the most scalable data structure, optimized for machine learning, large-scale data transformations and feature engineering are easy now.
- Best-in-Class Predictive Modeling**
Build your models with the best applied predictive technologies and machine learning algorithms, including Python scikit-learn and GraphLab Create™. Model selection, parameter search, metrics and visualization are easy now.
- Production-Ready Data Science**
Make your predictive models available as RESTful services, on AWS or on-premises. Integrate with Hadoop and Spark for distributed execution. Deploying distributed learning and predictive services are easy now.



MetaMind Products Demo Company Developers Log In

Train your image classifier

MetaMind allows you to automatically label your images with state-of-the-art classifiers based on convolutional neural networks.

UPLOAD IMAGES

To create a private dataset, you need to log in.
Datasets with 400 or more images are more accurate and train for more time.

Add File: Label 1 (0) Add File: Label 2 (0) Add Class +

Label 1 Label 2



clarifai ABOUT TECHNOLOGY API NEWS BLOGS JOBS CONTACT

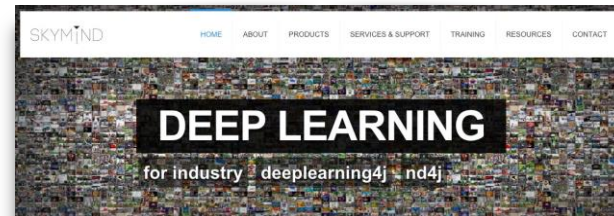
clarifai

Bring the future into focus with our world class visual recognition system

Make sense of your data with our deep learning system

TRY IT NOW API SIGN UP

- DAVE HASTON, GIGAMON, 05.06.14 Image recognition: Consumer products will drive enterprise breakthroughs
- ROBERT McLELLAN, WIRE, 07.03.14 Machines Finally Match Monkeys in Key Image Recognition Test
- TOM SHONTE, MIT TECH REVIEW, 04.22.13 A Startup's Neural Network Can Understand Video



SKYMIN D HOME ABOUT PRODUCTS SERVICES & SUPPORT TRAINING RESOURCES CONTACT

DEEP LEARNING

for industry deeplearning4j nd4j

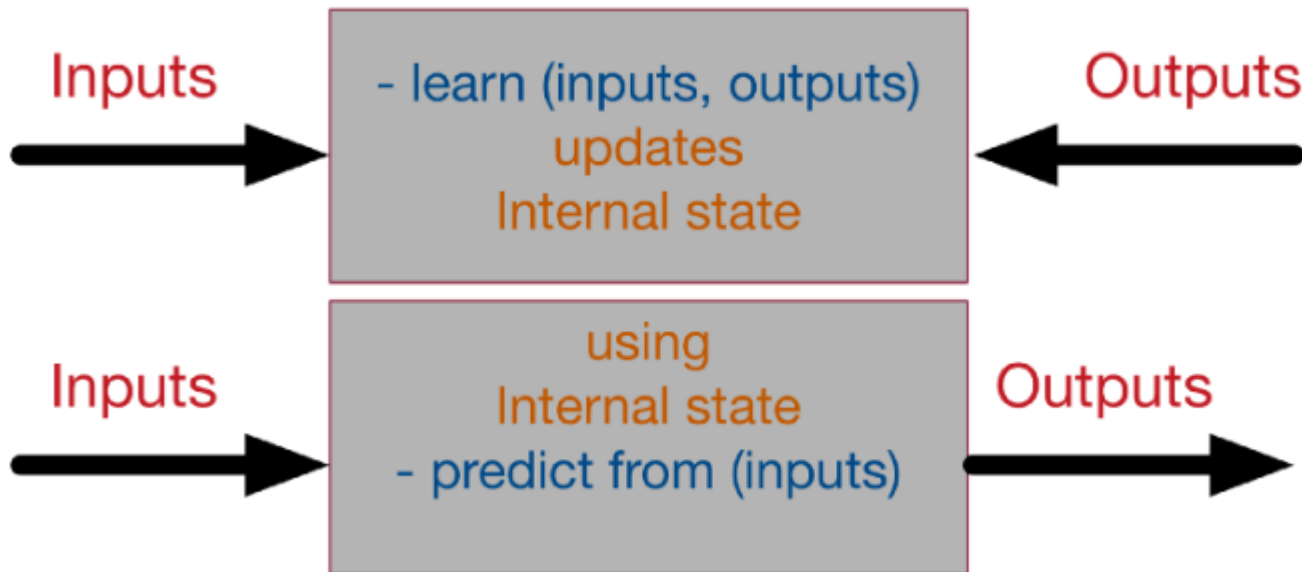
03

Mathematics behind Neural Networks

58

Neural networks and back-propagation

A supervised neural network, at the highest and simplest abstract representation, can be presented as a black box with 2 methods learn and predict



Neural Net Model

Parameters of a linear model

$$f(x) = o(w^T \cdot x + b)$$

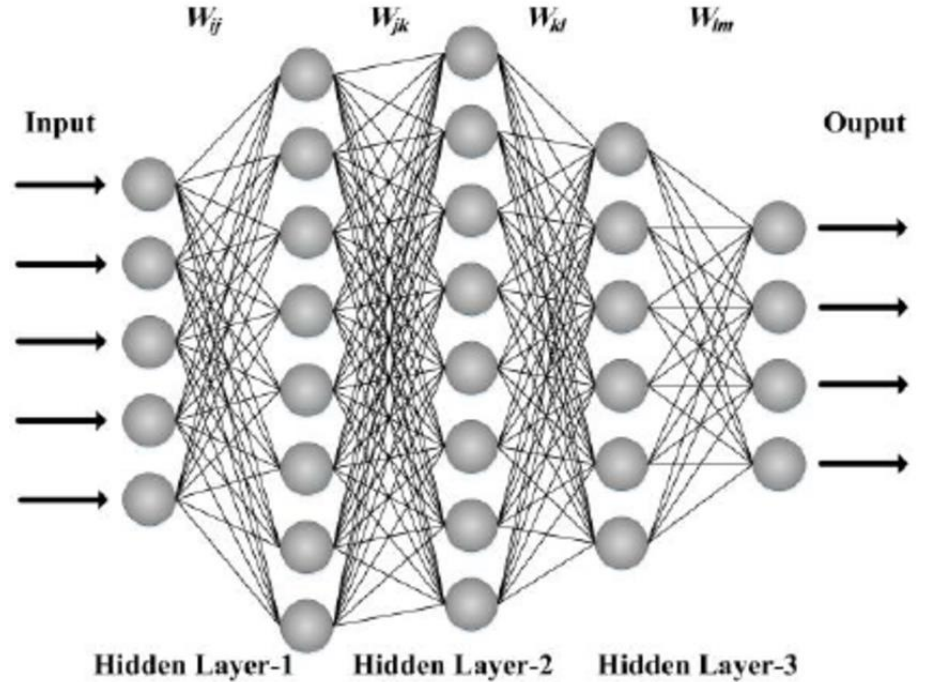
Diagram showing the equation $f(x) = o(w^T \cdot x + b)$ with arrows indicating the flow of information: Data (blue arrow) points to x ; Weights (red arrow) points to w ; Bias (red arrow) points to b ; and Dot Product (red arrow) points to $w^T \cdot x$.

Sigmoid
Function

$$o(x) = \frac{1}{1+e^{-x}}$$

ReLU
Function

$$o(x) = \max(0, x)$$



Neural Net: Mathematical steps

- **Model initialization:** Giving an initial value to the weights. Random initialization of the model is a common practice.
- **Forward propagate:** Evaluation of the initialized model.
- **Loss Function:** Function that compares the result of the evaluated model with the desired outputs.

As a whole, the process can be reduced to find the minimum of the Loss Function.

- **Differentiation: Gradient descent**
- **Back-propagation**
- **Weights Update**

Neural Net: Mathematical steps

$$Y = Wx$$

Input	Desired output
0	0
1	2
2	4
3	6
4	8

Input	Actual output of model 1 (y= 3.x)
0	0
1	3
2	6
3	9
4	12

Input	actual	Desired	Absolute Error	Square Error
0	0	0	0	0
1	3	2	1	1
2	6	4	2	4
3	9	6	3	9
4	12	8	4	16
Total:	-	-	10	30

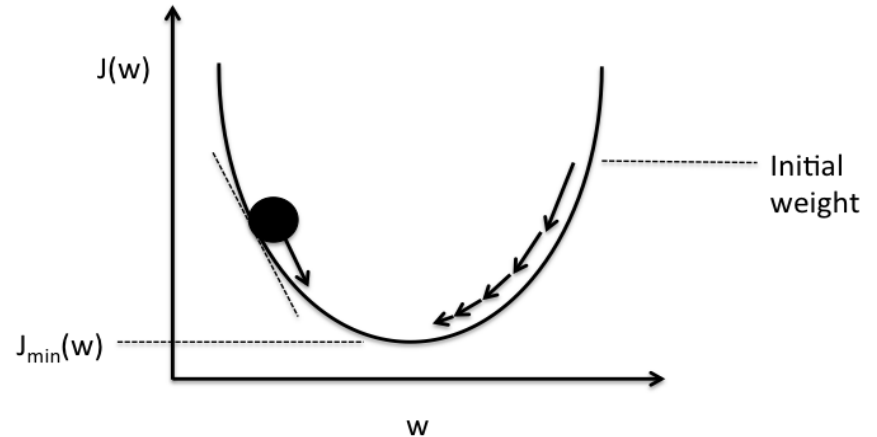
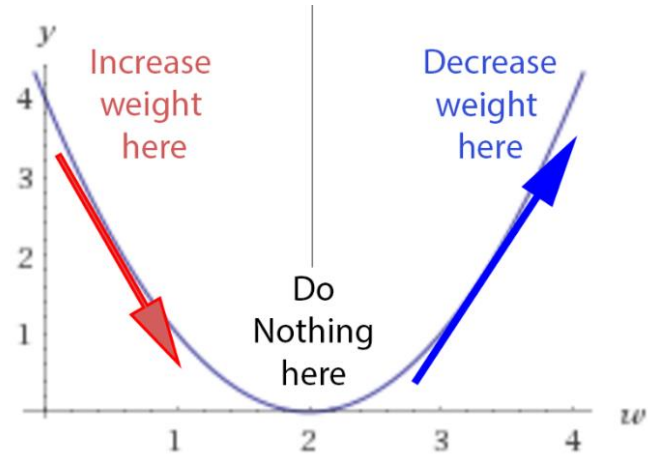
Loss
Function

Neural Net: Mathematical steps

Gradient Descent

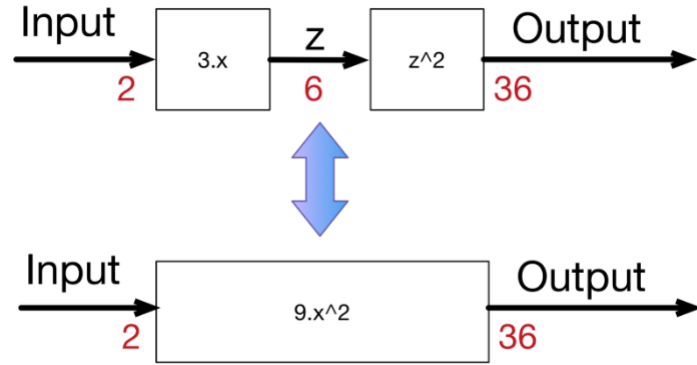
If we initialize randomly the network, we are putting any random point on this curve (let's say $w=3$). The learning process is actually saying this:

- Let's check the derivative.
- If it is positive, meaning the error increases if we increase the weights, then we should decrease the weight.
- If it's negative, meaning the error decreases if we increase the weights, then we should increase the weight.
- If it's 0, we do nothing, we reach our stable point.



Neural Net: Mathematical steps

Back-propagation



In most cases composing the functions is very hard. Plus for every composition one has to calculate the dedicated derivative of the composition (which is not at all scalable and very error prone). In order to solve the problem, luckily for us, derivative is decomposable, thus can be back-propagated. We have the starting point of errors, which is the loss function, and we know how to derivate it, and if we know how to derivate each function from the composition, we can propagate back the error from the end to the start. Let's consider the simple linear example: where we multiply the input 3 times to get a hidden layer, then we multiply the hidden (middle layer) 2 times to get the output.

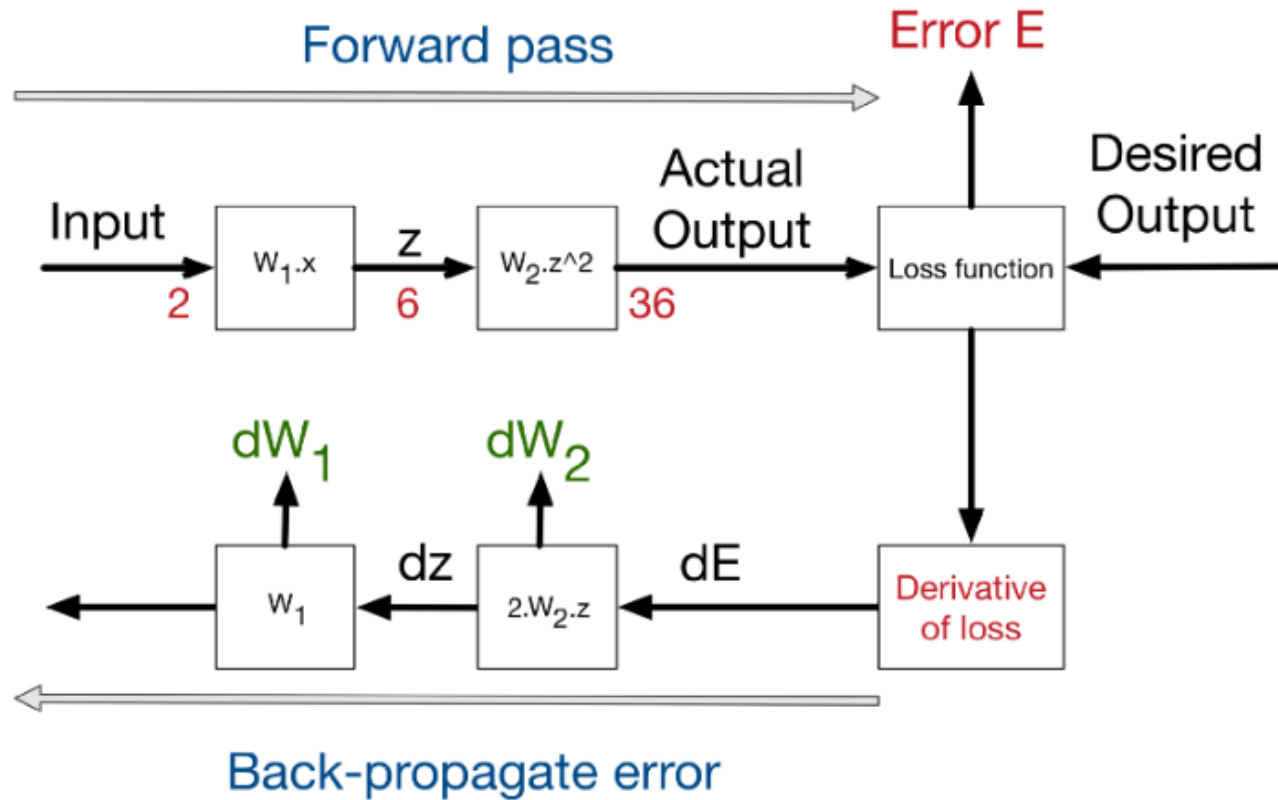
A 0.001 delta change on the input, will be translated to a 0.003 delta change after the first layer, then to 0.006 delta change on the output.

which is the case if we compose both functions into one:

input $\rightarrow 6 \cdot x \rightarrow$ output.

Similarly an error on the output of 0.006, can be **backpropagated** to an error of 0.003 in the middle hidden stage, then to 0.001 on the input.

Neural Net: Workflow diagram



04

Neural Networks for Business Problems

The Effectiveness of Personalized Product Recommendations

MarketingSherpa Study, 1.5 billion shopping sessions, 2015:

The recommendations used a variety of different common phrases on a product page, home page, shopping cart, category page or site wide. The actual product recommendations were dynamic and personalized based on visitor data, behavior, and history.

On the whole, **11.5% of the revenue** (whether from more volume or higher value of products) generated in the shopping sessions was attributable to **purchases from the product recommendations.**

- The companies that used the most common **“visitors who viewed this product also viewed”** on the product page had the highest success, **with a remarkable 68% of all revenue of those companies coming from the product recommendations.**
- The phrasing **“you might also like, ”** correlating to **16% of that group’s revenue to the recommendations.**
- The popular phrasing **“customers also bought”** on the cart page generated only **8% of revenues from recommendation sales.**





HMV, a British entertainment retailing company (music Retailer) realized that sending the same campaign message to all its customers is not appropriate anymore, as people start treating emails as spam and do not open them. The company uses **a recommendation system**, which analyses customer click streams and which products fits the customer's preferences. HMV sends out personalized recommendations, which increased the emails opening by over 70% on mobile phones, and PC mails by 50 %.



In 2013 **Item-to-Item collaborative filter**:
35% of all sales are estimated to be generated by the recommendation engine. In May 2016, Amazon opened up **DSSTNE** as open source software so that the promise of deep learning can extend beyond speech and object recognition to other areas such as search and recommendations



After a long refinement process, Netflix finally released its first “global” recommendation engine in December, 2016. Netflix will invest 1 billion of the total 5 billion of its budget in recommendation and personalization. Why?? Netflix estimates that only 20% of its subscriber video choices come from search, with the other 80% coming from recommendations

Examples

Customer Loyalty

Churn prediction is the task of identifying whether users are likely to stop using a service, product, or website.

- Churn Prediction model based on Machine Learning:

Decision Trees, SVM, Logistic Regression

Ensembles (Random Forests)

Boosting

Final method overall performance:

64 % accuracy on users who did churn
74% accuracy on users who did not churn

Optimize efforts: it is not worth trying to retain the 4.4% lower. We should focus only on 82.5% higher.



Look beyond just overall accuracy

Probability Range Bins	% of users in bin that actually churn
0% – 10%	4.4%
10% – 20%	14.1%
20% – 30%	25.8%
30% – 40%	35.5%
40% – 50%	44.9%
50% – 60%	55%
60% – 70%	65.5%
70% – 80%	76.8%
80% – 90%	82.5%
90% – 100%	NaN

Recent tests in churn prediction using Deep Learning show an overall accuracy higher than 78%.

Examples Optimal Shop location

In 2007 and 2008, Starbucks' CEO [Howard Schultz](#) was forced to come out of retirement to close hundreds of stores, and rethink the company's strategic growth plan.

"This time around, Starbucks took a more disciplined, data-driven approach to store openings and used mapping software to easily analyze massive amounts of data about planned store openings. The software analyzed location-based data and demographics to determine the best place to open Starbucks stores without hurting sales at other Starbucks locations.

"The software is also helping to determine where the next 1,500-plus stores should be placed not only to help the company expand, but drive revenue for new store developments."

Data used:

- *Mobile data*
- *Demographic and income data (CENSUS)*
- *Geoinformation (OpenStreet Maps and Google)*

