Introduzione ai Big Data e Intelligenza Artificiale

Dipartimento di Matematica e Informatica - 19 Feb 2020

Prof. Giovanni Giuffrida
“It is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change.”

-Charles Darwin, 1809
Introduction to Big Data
Figure 1
Revenue from big data and business analytics worldwide from 2015 to 2022 (in billion U.S. dollars)
Roots

- **Datification** is the revolution behind Big Data
- People have always tried to quantify the world around them
Datification is the revolution behind Big Data
People have always tried to quantify the world around them

Always-connected digital technologies made this a reality
Big Data releases

- **1.0**: Big Data become available and get collected
Big Data releases

- **1.0**: Big Data become available and get collected
- **2.0**: Technology to process Big Data develops
- **3.0**: Getting value out of Big Data (The most difficult!)
Big Data releases

- **1.0**: Big Data become available and get collected
- **2.0**: Technology to process Big Data develops
- **3.0**: Getting value out of Big Data (The most difficult!)
All around data and technology today

The top 6 most capitalized companies in the world

<table>
<thead>
<tr>
<th>Company</th>
<th>USbn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exxon Mobil</td>
<td>453</td>
</tr>
<tr>
<td>PetroChina China</td>
<td>424</td>
</tr>
<tr>
<td>General Electric</td>
<td>369</td>
</tr>
<tr>
<td>Gazprom Russia</td>
<td>300</td>
</tr>
<tr>
<td>China Mobile</td>
<td>298</td>
</tr>
<tr>
<td>Bank of China</td>
<td>278</td>
</tr>
</tbody>
</table>

2008
All around data and technology today

The top 6 most capitalized companies in the world

<table>
<thead>
<tr>
<th>Company</th>
<th>USbn</th>
<th>Company</th>
<th>USbn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exxon Mobil</td>
<td>453</td>
<td>Apple</td>
<td>927</td>
</tr>
<tr>
<td>PetroChina China</td>
<td>424</td>
<td>Amazon</td>
<td>778</td>
</tr>
<tr>
<td>General Electric</td>
<td>369</td>
<td>Google</td>
<td>767</td>
</tr>
<tr>
<td>Gazprom Russia</td>
<td>300</td>
<td>Microsoft</td>
<td>750</td>
</tr>
<tr>
<td>China Mobile</td>
<td>298</td>
<td>Facebook</td>
<td>542</td>
</tr>
<tr>
<td>Bank of China</td>
<td>278</td>
<td>Alibaba</td>
<td>500</td>
</tr>
</tbody>
</table>

2008

2018
When it became trendy?

Big Data vs Business Intelligence

Figure 2
A complicate landscape
Big Data according to Oxford Dictionary

**big data** n. Computing (also with capital initials): *data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.*
Anything that does not fit in Excel
How big is big?

2019: This is what happens in an Internet Minute

- 1 Million logs into Google
- 3.8 Million searches on Google
- 1 Million texts sent
- 4.5 Million videos viewed
- 694,444 queries
- 5996,956 spent online
- 2.1 Million snaps created
- 41,669,820 messages sent
- 4.8 Million GIFs served
- 180 smart speakers shipped
- 41 million music streaming subscriptions
- 1.4 Million swipes
- 87,500 people tweeting
- 347,222 scrolling Instagram
- 188 million emails sent
- 390,030 apps downloaded
- 550 million ads viewed
- 1.1 million views

Created by: @LoriLewis @OfficiallyChadd
How big is big?

Every day we create

2,500,000,000,000,000

(2.5 quintillion) bytes of data

90% of the world’s data today has been created in the last 2 years alone.

This would fill 10 million Blu-ray discs, the height of which stacked, would measure the height of 4 Eiffel Towers on top of one another.
How big is big? Let’s try to measure it

- Many attempts to measure the world information of all types
- Prof. Martin Hilbert from USC:
  - In 2000: Only 25% of the world information was digital
  - In 2007: Only 7% of the world information was analog
  - In 2013: 1200 Exabyte (1B gigabytes) of overall data, only 2% analog
  - If it were books: Cover entire US surface 52 layers of book
  - If it were CD-ROMs: 5 separate piles to the moon
"Big data is high-Volume, high-Velocity and high-Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."
"Big data is high-Volume, high-Velocity and high-Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- **Value** and **Veracity** added later
"Big data is high-Volume, high-Velocity and high-Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- **Value** and **Veracity** added later

- **Volatility** and **Validity** added later
"Big data is high-Volume, high-Velocity and high-Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- **Value** and **Veracity** added later
- **Volatility** and **Validity** added later
- **Virality** added even later
Frankly... No formal definition yet!

- These numbers outstrip our machines and our imagination
- Technology to process all this is behind
- "Big" depends on the context for many
The FOUR V’s of Big Data

Volume
SCALE OF DATA

40 ZETTABYTES

Volume of data will be created by 2020, an increase of 300 times from 2010

6 MILLION PEOPLE

30% of adults in the US will have at least 100 TERABYTES of data stored in 2015

Variety
DIFFERENT FORMS OF DATA

8 MILLION WEARABLE, WIRELESS HEALTH MONITORS

28 MILLION IMAGES (photos, x-rays, etc).

102 MILLION TWEETS: one sent per day by 12% of U.S. million income online users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures: 1 TB OF TRADE INFORMATION during each trading session

Medically, data from over 100 SENSORS that monitor heart rate, temperature, etc.

4.4 MILLION ATT.

By 2015, it is projected there will be 18.3 MILLION NETWORK CONNECTIONS. 500 million connections with 5G connections per user.

Veracity
UNCERTAINTY OF DATA

11% of U.S. business leaders don’t trust the information they use to make decisions

50 BILLION BOTTLES

In 2015, there were 1 billion people who drink at least 20 bottles of water a day

Sources: McKinsey Global Institute, Twitter, US Census, IDC, Gartner, FORRESTER, IBM
Volume: Data at rest

- About: Amount of data
- Unit: bytes
- Information about the general population, education, health, medicine, travel, geographic locations, shopping, financial transactions, jobs, scientific experiments, emails, sensors, texts, photos, videos, activity on social networks, etc.

- How much is "Big Data"?
Velocity: Data in motion

- About: Moving data
- Unit: Bytes per second
- Two possible interpretations
  - Data Generation Rate
  - Data Processing Rate
- Every minute (2018):
  - 187M emails sent
  - 3.7M searches on Google
  - 38M WhatsApp messages
  - 973K logins on Facebook
  - ...

Variety: Data in many forms

- About: Form of data
- Three basic types of data
  - Structured = Data in a fixed field within a record (spreadsheets, Relational Database)
  - Semi-Structured = XML, JSON, CSV
  - Unstructured = Data stored without any model, or that does not have any organisation
- Any of those types can be big
- Only 20% of data today is "structured"
Veracity: Data in doubt

- Uncertainty due to many factors
- Incompleteness
- Inconsistency
- Ambiguity
- Model approximation
- Technical constraints
- Often overlooked
- But... it could as important as the other Vs
Value: Actionable insights

People do not need data, they need insights!!

- Hidden in the data
- Value is a concentrated data-juice
- Gaining correct but irrelevant or un-actionable information is a (big) waste of time
Introduction to Artificial Intelligence
In 1956 the term **Artificial Intelligence** was coined by John McCarthy.
In 1956 the term **Artificial Intelligence** was coined by John McCarthy.

“The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.”
History repeats itself

- 90’s: Artificial Intelligence, Expert systems
- 00’s: Data/Text mining, Knowledge discovery, Machine learning
- Today: Big Data, DMP, Machine Learning, Deep learning

Same goal: Extracting meaningful info from data
History repeats itself

- 90’s: Artificial Intelligence, Expert systems
- 00’s: Data/Text mining, Knowledge discovery, Machine learning
- Today: Big Data, DMP, Machine Learning, Deep learning

Same goal: Extracting meaningful info from data

So... what's the difference today??
Three main reasons

1. **Data size**: Today’s data at companies disposal is really big. Traditional methods break.
Three main reasons

1. **Data size**: Today’s data at companies disposal is really big. Traditional methods break.

2. **Technology**: Tech infrastructure to handle big data is now available (e.g., Hadoop, NoSql, CouchBase, MongoDB, Dynamo, Hbase, RedShift, etc.)
Three main reasons

1. **Data size**: Today’s data at companies disposal is really big. Traditional methods break.

2. **Technology**: Tech infrastructure to handle big data is now available (e.g., Hadoop, NoSql, CouchBase, MongoDB, Dynamo, Hbase, RedShift, etc.)

3. **Data culture**: Companies now appreciate value of data to run their business
Now industry is really serious about it

- In the past, mostly limited to Universities and research centers
- Companies now aggressively investing big $$$ in Big Data
- Shifting budgets from offline to online
- They want more control on their investments: transparency and control
The evolution
The AI evolution

Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.
Fully automated learning

Machine Learning

Input → Feature extraction → Classification → Output (Car, Not Car)

Deep Learning

Input → Feature extraction + Classification → Output (Car, Not Car)
Self learning example

Learning to walk
A new vision of Artificial Intelligence today

- Before: Intelligence was *hardcoded* into machines
- Today: Machines learn by *observing* Big Data
A new vision of Artificial Intelligence today

- Before: Intelligence was **hardcoded** into machines
- Today: Machines learn by **observing** Big Data

Big Data $\Rightarrow$ A.I.
The old vs the new school

- In the past, many attempts to make machines "Intelligent": Expert systems, Artificial Intelligence, etc.
- Today, Big data/Artificial Intelligence is about deriving math models (insights) from huge data bases
- Being able to observe and learn models leads to *intelligent behavior*
  - IBM Watson
  - AlphaGo
The old vs the new school

- CYC vs Watson
- Two (very) different approaches
- CYC was “embedding” knowledge
- Watson is able to “learn” from huge amount of data
Cyc

From Wikipedia, the free encyclopedia.

For other uses, see Cyc (disambiguation).

Cyc (ˈsæk/) is the world's longest-lived artificial intelligence project,[citation needed] attempting to assemble a comprehensive ontology and knowledge base that spans the basic concepts and "rules of thumb" about how the world works (think common sense knowledge but focusing more on things that rarely get written down or said, in contrast with facts one might find somewhere on the internet or retrieve via a search engine or Wikipedia), with the goal of enabling AI applications to perform human-like reasoning and be less "brittle" when confronted with novel situations that were not preconceived.

Douglas Lenat began the project in July 1984 at MCC, where he was Principal Scientist 1984–1994, and then, since January 1995, has been under active development by the Cycorp company, where he is the CEO.
Oscar Wilde said of this title place "The warder is despair"

At the beginning of "A Tale of Two Cities", these 2 kings sit on the thrones of England & France

Around 1912, while recovering in a sanatorium, this former seaman decided to become a playwright

The accompanying text to this book was published separately as "Ornithological Biography" in the 1830s

In May 1973 Sports Illustrated ran one of his short stories under the title "A Day of Wine and Roses"

This author & biochemist who died in 1992 has at least one book in all 10 main Dewey Decimal categories

The Prague tombstone of this German-language writer who died in 1924 is inscribed in Hebrew

D.H. Lawrence called him "an adventurer into the vaults and... horrible underground passages of the human soul"

In 1935 she sent a telegram to a Macmillan editor: "Please send manuscript back I've changed my mind"
IBM’s Watson -- the language-fluent computer that beat the best human champions at a game of the US TV show *Jeopardy!* -- is being turned into a tool for medical diagnosis. Its ability to absorb and analyse vast quantities of data is, IBM claims, better than that of human doctors, and its deployment through the cloud could also reduce healthcare costs.
Two years ago, IBM announced that Watson had “learned” the same amount of knowledge as the average second-year medical student. For the last year, IBM, Sloan-Kettering and Wellpoint have been working to teach Watson how to understand and accumulate complicated peer-reviewed medical knowledge relating to oncology. That’s just lung, prostate and breast cancers to begin with, but with others to come in the next few years). Watson’s ingestion of more than 600,000 pieces of medical evidence, more than two million pages from medical journals and the further ability to search through up to 1.5 million patient records for further information gives it a breadth of knowledge no human doctor can match.
According to Sloan-Kettering, only around 20 percent of the knowledge that human doctors use when diagnosing patients and deciding on treatments relies on trial-based evidence. It would take at least 160 hours of reading a week just to keep up with new medical knowledge as it’s published, let alone consider its relevance or apply it practically. Watson’s ability to absorb this information faster than any human should, in theory, fix a flaw in the current healthcare model. Wellpoint’s Samuel Nessbaum has claimed that, in tests, Watson’s successful diagnosis rate for lung cancer is 90 percent, compared to 50 percent for human doctors.
Tom Mitchell says:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E"
or in a simpler way:

"The field of Machine Learning is concerned with the question of how to construct computer programs that automatically improve with experience."
A formal definition:

"Deep Learning is a particular kind of Machine Learning that achieves great power and flexibility by learning to represent the world as nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones."
Learning from data

Why deep learning

Performance vs. Amount of data

Deep learning vs. Older learning algorithms

How do data science techniques scale with amount of data?
In 1997
• 3000 years old game
• Simple board
• Before 2016 it was considered to be impossible to model
• Many (many) more combinations compared to chess
• It was said:
  • "the most elegant game that humans have ever invented";
  • "simple rules that give rise to endless complexity";
  • "more possible Go positions than there are atoms in the universe"
In 2016
It gets better

- In 2018 AlphaGo-Zero
- A new version based on Deep Learning techniques

Previous versions of AlphaGo initially trained on thousands of human amateur and professional games to learn how to play Go. AlphaGo Zero skips this step and learns to play simply by playing games against itself, starting from completely random play. In doing so, it quickly surpassed human level of play and defeated the previously published champion-defeating version of AlphaGo by 100 games to 0.
At the beginning of the training, AlphaGo Zero has no prior knowledge of the game and only the basic rules as an input.
After 3 days

3 days

AlphaGo Zero surpasses the abilities of AlphaGo Lee, the version that beat world champion Lee Sedol in 4 out of 5 games in 2016.
After 21 days

21 days

AlphaGo Zero reaches the level of AlphaGo Master, the version that defeated 60 top professionals online and world champion Ke Jie in 3 out of 3 games in 2017.
After 40 days

**40 days**

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.
The Turing test
The Turing test
Big Data is surely helping computers towards passing the Turing test!
Learning Models
Type of learning

Mostly, two different learning paradigms:

- Supervised
- Unsupervised
Supervised learning

- Data are *labelled*
- Labels are the targets (or output, or class), basically what we want to *learn*
- So, for each observation we have:
  - Input values
  - Label

The machine learning algorithm learns such associations over time
Supervised learning - example

We want to teach a small kid how to distinguish a *bike* from a *car*

He has not ever seen those before

Input = a set of labelled images
Supervised learning - example

Let's proceed as follows:

1. Let’s show the images of the bikes
2. We tell him those are "bikes"
3. We do not teach him about any specific characteristic

So we let the kid analyse those images to understand what makes those objects a “bike”
Supervised learning - example

We do the same with the cars

Figure 3
We let him “think and learn”
Eventually, we show him a picture and ask him to identify it.

**Notice: It’s a new picture, he has not seen it before.**
Supervised learning - example

Eventually, we show him a picture and ask him to identify it

**Notice: It’s a new picture, he has not seen it before**
Unsupervised learning

- Here the algorithm learns without any label
- The input to the algorithm is just a set of observations

In general, this is a more challenging class of problems
Unsupervised learning - Example

- Let’s repeat the previous example with no *supervision*
- This time we show the kid the images at once, *bikes and cars together*
- *We don’t tell him anything about the two type of objects!*
Unsupervised learning - Example

The kid has to learn by himself the two categories and what makes those different from each other

Figure 4
Then, like before, we show him a new unseen image
The kid in his learning may use

- more than two categories
- a very different set of categories of what we expect

For instance, he may decide to put together objects based on color, size, or number of wheels (that he sees!)
Unsupervised learning - Considerations

- The results is greatly dependent upon the quality of the input images
- As usual, the more data in input the more accurate the learning, at least, until a certain point
Reinforcement learning - The basic principle

- Learning is based on a *gain function*
- Each time the *machine* reaches a positive state it gains something
- The objective is to *maximize gain*
- Used by DeepMind to develop AlphaGo
We do it *naturally* in life, maybe without knowing it

- Basic task for "data miners"
- Statisticians have been doing it for many years
- It takes many different forms
- Today, *all managers should have at least a basic understanding*
The art of modeling: Classification
Say, we collect the following data:

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>65k</td>
<td>Y</td>
</tr>
<tr>
<td>68</td>
<td>83k</td>
<td>Y</td>
</tr>
<tr>
<td>43</td>
<td>61k</td>
<td>N</td>
</tr>
<tr>
<td>30</td>
<td>25k</td>
<td>Y</td>
</tr>
<tr>
<td>51</td>
<td>82k</td>
<td>N</td>
</tr>
<tr>
<td>78</td>
<td>67k</td>
<td>Y</td>
</tr>
</tbody>
</table>
Modeling: A simple “supervised” example

- Say, we collect the following data:

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>65k</td>
<td>Y</td>
</tr>
<tr>
<td>68</td>
<td>83k</td>
<td>Y</td>
</tr>
<tr>
<td>43</td>
<td>61k</td>
<td>N</td>
</tr>
<tr>
<td>30</td>
<td>25k</td>
<td>Y</td>
</tr>
<tr>
<td>51</td>
<td>82k</td>
<td>N</td>
</tr>
<tr>
<td>78</td>
<td>67k</td>
<td>Y</td>
</tr>
</tbody>
</table>

Goal: **Build a model to predict the dependent variable** $Out$
Modeling: A simple “supervised” example

- Say, we collect the following data:

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>65k</td>
<td>Y</td>
</tr>
<tr>
<td>68</td>
<td>83k</td>
<td>Y</td>
</tr>
<tr>
<td>43</td>
<td>61k</td>
<td>N</td>
</tr>
<tr>
<td>30</td>
<td>25k</td>
<td>Y</td>
</tr>
<tr>
<td>51</td>
<td>82k</td>
<td>N</td>
</tr>
<tr>
<td>78</td>
<td>67k</td>
<td>Y</td>
</tr>
</tbody>
</table>

Goal: **Build a model to predict the dependent variable** *Out*

- *Out* is a categorical variable - *Age* and *Income* are the independent variables
Let’s model

- We project the dependent variable *Out*, as + and ○, on a two-dimensional space
Let's model

Do we notice anything?
Let’s model
Let’s model
Let’s model

- We can model it through a **Classification tree**
Let’s model

- We can model it through a Classification tree

- The algorithm **finds** the optimal splits - It maximizes **prediction confidence**
Let’s apply the model now

- Let’s now **generalize** the model
- Say, we receive new data and we want to predict the *Out* variable
- For instance, a new person 25 years old and with an income of 70k
We can now predict
Classification tree considerations

- **C4.5** was the original idea
- Old technique very well established
- Works for *categorical* dependent variable
- It works with both *continuos* and *categorical* independent variables
- Fast to execute
- Easy to find free code on the web
The art of modeling: Clustering
Clustered • Unsupervised learning model
• Widely used in business/marketing applications
• Gives a structure to unsorted data points
• In simpler words: Aggregate similar items
k-means Clustering algorithm

- **Step 0**: Initialize K random centroids (*just pick randomly K data points*)
- **Step 1** For every data point:
  - assign it to the closest centroid (*any distance metric works*);
- **Step 2** For every centroid
  - move the centroid to the *average* among all its points;
- Repeat **Step 1** and **Step 2** until all centroids do not change anymore;

The algorithm "converged"
k-means example... Initial step

Figure 5
k-means example... Iterations

Figure 6
k-means, some considerations

- Easy to apply
- Various algorithms available
- You can find free, ready to use, code on the web
- Not suitable for categorical variables
- Need to normalize variable for scale uniformity
- Otherwise, distance calculation may be affected
- It scales on Big Data, that is, it can be parallelized
- How to pick initial K?
The art of modeling: Associations
Associations

What can we infer just by observing?
**Market-basket analysis**: Understanding meaningful *patterns* by analysing baskets

A *basket* is a generic set of items

- **Pattern**: A set of items
- **Frequent pattern**: A pattern that appears *frequently*
- We infer rules such as: \( A, B, ..., C \implies X \)
- *Beer and diapers on friday evening*?!
- It may depend on the context: "Have kids?", "Travelling for work?", etc.
Metrics: Support and Confidence

- We are only interested in rules with *high support*
- Statistically meaningful
- \( s(X \implies Y) = \text{"\# of transactions containing both } X \text{ and } Y\)
Metrics: Support and Confidence

- We are only interested in rules with **high support**
- Statistically meaningful
- \( \text{support } s(X \implies Y) = \# \text{ of transactions containing both } X \text{ and } Y \)

- We are only interested in rules with **high confidence**
- \( \text{Confidence } c(X \implies Y) = \text{“conditional probability that a transaction containing } X \text{ also contains } Y \)
We are only interested in rules with *high support*

Statistically meaningful

\[ \text{support } s(X \implies Y) = \text{“# of transactions containing both } X \text{ and } Y” \]

We are only interested in rules with *high confidence*

\[ \text{Confidence } c(X \implies Y) = \text{“conditional probability that a transaction containing } X \text{ also contains } Y” \]

Be careful: The rule \( \text{milk, butter } \implies \text{ bread } \) may derive from the fact that many baskets contain \( \text{bread} \)

Basically, we need to select rules such as:

\[ c(\text{milk, butter } \implies \text{ bread}) >> c(\text{bread}) \]
The Apriori algorithm

- Find all $X \implies Y$
- Supports in the given example:
  - $s(\text{Birra}) = 3$
  - $s(\text{Noccioline}) = 3$
  - $s(\text{Pannolini}) = 4$
  - $s(\text{Uova}) = 3$
  - $s(\text{Birra, Pannolini}) = 3$

- $\text{Birra} \implies \text{Pannolini}(3, 100\%)$
- $\text{Pannolini} \implies \text{Birra}(3, 75\%)$
- $\text{Latte} \implies \text{Uova}(?, ?)$
- $\text{Noccioline} \implies \text{Latte}(?, ?)$
Apriori, execution example
Association rules applications

Association rules are used in many different contexts:

- Combined promotions
- Optimized shelf allocation
- Text documents based on shared concepts
- Targeted promotions of movies/books/articles/etc
- CTR banner optimization
- Mechanical fault prevention
- Medical diagnosis
- etc.
Big Data Infrastructure
Serial computation

- Traditional computation is “serial”
- One instruction after the other one
- Overall speed depends on the CPU speed

With Big Data, serial computation poses serious limits
The problem is broken into **parts**
- Each part is executed on a different CPU
- Needs proper coordination
Parallel computing prime
Parallel computing prime
Parallel computing prime
Parallel computing prime
Parallel computing considerations

- In theory, if we break a problem into N parts it should take 1/N of the time to execute it in parallel
- But this is not the case because:
  - Uneven workload distribution
  - Execution dependency
  - Communication time between parts
  - Coordination time
Infrastructure for parallel computing
- Introduced in 2005 by Doug Cutting and Mike Cafarella at Yahoo!
- Name and logo from a toy of Doug’s son
- Started as an open source software project
Hadoop, what is it?

- A *software layer* to coordinate multiple PC execution
- Provide distributed storage
- Control distributed processing
- Fault tolerant
- A robust platform for storing and analysing Big Data
- Easily scalable to thousands of nodes
- At the base of all tech giants tech infrastructure today
A processing model to process Big Data in a parallel fashion over a large number of computers

- Algorithm to parallelize execution
- Two phases: Map and Reduce
- Based on a “key-value” concept
MapReduce prime

A **processing model** to process Big Data in a parallel fashion over a large number of computers

- Algorithm to parallelize execution
- Two phases: Map and Reduce
- Based on a “key-value” concept

Many problems can be mapped into a MapReduce model
MapReduce word count example

Fig. Workflow of MapReducing
“It does look similar—but this one is powered by Hadoop”
A not so far-fetched scenario:

“One of your loyal customers posts on Facebook that she’s going shopping at one of your stores today. You know that she just purchased a pair of pants online last week, and that her abandoned online shopping cart has a few cute tops in it to go with the pants. She goes to the store, the retail assistant is able to identify who she is and brings out the tops she abandoned online to try on with her new pants. But since your customer isn’t wearing her new pants, the retail assistant knows which size pants to go grab. Then while shopping, your customer gets a 25% off coupon delivered to her smartphone-good for today only.”
Many types of data

- The big challenge is to “use” all available data

<table>
<thead>
<tr>
<th>POS DATA</th>
<th>CRM</th>
<th>FINANCIAL DATA</th>
<th>LOYALTY CARD DATA</th>
<th>TROUBLE TICKETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMAIL</td>
<td>PDF FILES</td>
<td>SPREADSHEETS</td>
<td>WORD PROCESSING DOCUMENTS</td>
<td>RFID TAGS</td>
</tr>
<tr>
<td>GPS</td>
<td>WEB LOG DATA</td>
<td>PHOTOS</td>
<td>SATELLITE IMAGES</td>
<td>SOCIAL MEDIA DATA</td>
</tr>
<tr>
<td>BLOGS</td>
<td>FORUMS</td>
<td>CLICK-STREAM DATA</td>
<td>VIDEOS</td>
<td>XML DATA</td>
</tr>
<tr>
<td>MOBILE DATA</td>
<td>WEBSITE CONTENT</td>
<td>RSS FEEDS</td>
<td>AUDIO FILES</td>
<td>CALL CENTER TRANSCRIPTS</td>
</tr>
</tbody>
</table>
Data warehouse, data lakes

- Data are not preprocessed and imported apriori
- Data are kept in their native format
- In general, more flexibility for future unplanned applications
Datalake overall architecture
<table>
<thead>
<tr>
<th>DATA WAREHOUSE</th>
<th>vs.</th>
<th>DATA LAKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>structured, processed</td>
<td>DATA</td>
<td>structured / semi-structured / unstructured, raw</td>
</tr>
<tr>
<td>schema-on-write</td>
<td>PROCESSING</td>
<td>schema-on-read</td>
</tr>
<tr>
<td>expensive for large data</td>
<td>STORAGE</td>
<td>designed for low-cost storage</td>
</tr>
<tr>
<td>volumes</td>
<td>AGILITY</td>
<td>highly agile, configure and reconfigure as needed</td>
</tr>
<tr>
<td>less agile, fixed configuration</td>
<td>SECURITY</td>
<td>maturing</td>
</tr>
<tr>
<td>mature</td>
<td>USERS</td>
<td>data scientists et. al.</td>
</tr>
<tr>
<td>business professionals</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The new Big Data Science
The new Science of Data
Big Data Science is a special mix

- **Science**: You can learn it
- **Crafts**: You can learn it
- **Creativity**: in part natural, in part it may came through experience
- **Common Sense**: You need to have/develop it
Big Data Science is multidisciplinary
Data Science is becoming ubiquitous

- Analytical thinking will be *pervasive* in companies
- Managers should have a *basic understanding of fundamental principles*
- Managers will lead data-analytics teams and data-driven projects
- Company’s culture should support data-driven processes
- All aspects: from production to marketing to sales
The ideal data scientist profile

- Mostly a domain expert (70%)
- A strong background on data (30%)
- Ideally, *all managers should be like this*
- See A.I. like "Excel"
McKinsey: "There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."
The “Data Scientist” must-have skills

- Business understanding
- Basic statistics
- Basic statistical programming: R and Python, SQL
- Machine learning concepts
- Data Visualization: Many tools available
- Thinking like a “Data Scientist”: see data everywhere
- Thinking like a problem solver
- Exceptional oral and written communication abilities
- Customer oriented approach
Data Science Pros & Cons

**Advantages**

A. It’s in Demand
B. Abundance of Positions
C. Highly Paid Career
D. Highly Prestigious
E. Versatile

**Disadvantages**

A. It is a Blurry Term
B. Mastering Data Science is near to impossible
C. Large amount of domain knowledge required
D. Arbitrary Data May Yield Unexpected Results
E. Problem of Data Privacy
Thanks