



Cloud Computing and Big Data Fundamentals

H2020 COST ACTION TD1403 “BIG SKY EARTH”
2016 TRAINING SCHOOL

Stefano Cavuoti

INAF – Capodimonte Astronomical Observatory – Napoli

Cloud Computing

Oversimplification: Cloud computing is computing based on the internet.

Where in the past, people would run applications or programs from software downloaded on a physical computer or server in their building, cloud computing allows people access to the same kinds of applications through the internet.



Cloud Computing

When you update your Facebook status, you're using cloud computing. Checking your bank balance on your phone? You're in the cloud again.

In short, cloud is fast becoming the new normal. By the end of 2016 it was estimated that 90% of UK businesses will be using at least one cloud service.

Why are so many businesses moving to the cloud? It's because cloud computing increases efficiency, helps improve cash flow and offers several benefits.



Opportunities and Challenges

The use of the cloud provides a number of opportunities:

- It enables services to be used without any understanding of their infrastructure.
- Cloud computing works using economies of scale:
 - It potentially lowers the outlay expense for start up companies, as they would no longer need to buy their own software or servers.
 - Cost would be by on-demand pricing.
 - Vendors and Service providers claim costs by establishing an ongoing revenue stream.
- Data and services are stored remotely but accessible from “anywhere”.

Opportunities and Challenges

In parallel there has been backlash against cloud computing:

- Use of cloud computing means dependence on others and that could possibly limit flexibility and innovation:
 - The others are likely become the bigger Internet companies like Google and IBM, who may monopolise the market.
 - Some argue that this use of supercomputers is a return to the time of mainframe computing that the PC was a reaction against.
- Security could prove to be a big issue:
 - It is still unclear how safe out-sourced data is and when using these services ownership of data is not always clear.
- There are also issues relating to policy and access:
 - If your data is stored abroad whose policy do you adhere to?
 - What happens if the remote server goes down?
 - How will you then access files?
 - There have been cases of users being locked out of accounts and losing access to data.

Advantages of Cloud Computing

- Lower computer costs:
 - You do not need a high-powered and high-priced computer to run cloud computing's web-based applications.
 - Since applications run in the cloud, not on the desktop PC, your desktop PC does not need the processing power or hard disk space demanded by traditional desktop software.
 - When you are using web-based applications, your PC can be less expensive, with a smaller hard disk, less memory, more efficient processor...
 - In fact, your PC in this scenario does not even need a CD or DVD drive, as no software programs have to be loaded and no document files need to be saved.
- Reduced software costs:
 - Instead of purchasing expensive software applications, you can get most of what you need for free!
 - most cloud computing applications today, such as the Google Docs suite.
 - better than paying for similar commercial software
 - which alone may be justification for switching to cloud applications.

Advantages of Cloud Computing

- Instant software updates:
 - Another advantage to cloud computing is that you are no longer faced with choosing between obsolete software and high upgrade costs.
 - When the application is web-based, updates happen automatically
 - available the next time you log into the cloud.
 - When you access a web-based application, you get the latest version
 - without needing to pay for or to download an upgrade.
- Improved performance:
 - With few large programs hogging your computer's memory, you will see better performance from your PC.
 - Computers in a cloud computing system boot and run faster because they have fewer programs and processes loaded into memory...

Advantages of Cloud Computing

- Unlimited storage capacity:
 - Cloud computing offers virtually limitless storage.
 - Your computer's current 1 Tbyte hard drive is small compared to the hundreds of Pbytes available in the cloud.
- Increased data reliability:
 - Unlike desktop computing, in which if a hard disk crashes will destroy all your valuable data, a computer crashing in the cloud should not affect the storage of your data.
 - if your personal computer crashes, all your data is still out there in the cloud, still accessible
 - In a world where few individual desktop PC users back up their data on a regular basis, cloud computing is a data-safe computing platform!

Advantages of Cloud Computing

- Universal document access:
 - That is not a problem with cloud computing, because you do not take your documents with you.
 - Instead, they stay in the cloud, and you can access them whenever you have a computer and an Internet connection
 - Documents are instantly available from wherever you are
- Latest version availability:
 - When you edit a document at home, that edited version is what you see when you access the document at work.
 - The cloud always hosts the latest version of your documents
 - as long as you are connected, you are not in danger of having an outdated version
- Improved document format compatibility.
 - You do not have to worry about the documents you create on your machine being compatible with other users' applications or OSes
 - There are potentially no format incompatibilities when everyone is sharing documents and applications into a cloud.

Advantages of Cloud Computing

- Easier group collaboration:
 - Sharing documents leads directly to better collaboration.
 - Many users do this as it is an important advantage of cloud computing
 - multiple users can collaborate easily on documents and projects
- Device independence.
 - You are no longer tethered to a single computer or network.
 - Changes to computers, applications and documents follow you through the cloud.
 - Move to a portable device, and your applications and documents are still available.

Disadvantages of Cloud Computing

- Requires a constant Internet connection:
 - Cloud computing is impossible if you cannot connect to the Internet.
 - Since you use the Internet to connect to both your applications and documents, if you do not have an Internet connection you cannot access anything, even your own documents.
 - A dead Internet connection means no work and in areas where Internet connections are few or inherently unreliable, this could be a deal-breaker.

Disadvantages of Cloud Computing

- Does not work well with low-speed connections:
 - Similarly, a low-speed Internet connection, such as that found with dial-up services, makes cloud computing painful at best and often impossible.
 - Web-based applications require a lot of bandwidth to download, as do large documents.
- Features might be limited:
 - This situation is bound to change, but today many web-based applications simply are not as full-featured as their desktop-based applications.
 - For example, you can do a lot more with Microsoft PowerPoint than with Google Presentation's web-based offering

Disadvantages of Cloud Computing

- Can be slow:
 - Even with a fast connection, web-based applications can sometimes be slower than accessing a similar software program on your desktop PC.
 - Everything about the program, from the interface to the current document, has to be sent back and forth from your computer to the computers in the cloud.
 - If the cloud servers happen to be backed up at that moment, or if the Internet is having a slow day, you would not get the instantaneous access you might expect from desktop applications.

Concerns about Cloud Computing

- Stored data might not be secure:
 - With cloud computing, all your data are stored on the cloud.
 - The questions is How secure is the cloud?
 - Can unauthorised users gain access to your confidential data?
- Stored data can be lost:
 - Theoretically, data stored in the cloud are safe, replicated across multiple machines.
 - But on the off chance that your data goes missing, you have no physical or local backup.
 - Put simply, relying on the cloud puts you at risk if the cloud lets you down.

Open Issues of Cloud Computing

- HPC Systems:
 - Not clear that you can run compute-intensive HPC applications that use MPI/OpenMP!
 - Scheduling is important with this type of application
 - as you want all the VM to be co-located to minimize communication latency!
- General Concerns:
 - Each cloud systems uses different protocols and different APIs
 - may not be possible to run applications between cloud based systems
 - Amazon has created its own DB system (not SQL 92), and workflow system (many popular workflow systems out there)
 - so your normal applications will have to be adapted to execute on these platforms.

Example of Cloud Computing

Cloud Market Types	Types of Offerings	Examples
Software-as-a-Service	<ul style="list-style-type: none"> • Rich Internet application web sites • Application as Web Sites • Collaboration and email • Office Productivity • Client apps that connect to services in the cloud 	<ul style="list-style-type: none"> • Flickr • Myspace.com • Cisco WebEx office • Gmail • IBM Bluehouse
App-components-as-a-Service	<ul style="list-style-type: none"> • APIs for specific service access for integration • Web-based software service than can combine to create new services, as in a mashup 	<ul style="list-style-type: none"> • Amazon Flexible Payments Service and DevPay • Salesforce.com's AppExchange • Yahoo! Maps API • Google Calendar API • zembly
Platform-as-a-Service	<ul style="list-style-type: none"> • Development-platform-as-a-service • Database • Message Queue • App Servicer • Blob or object data stores 	<ul style="list-style-type: none"> • Google App Engine and BigTable • Microsoft SQL Server Data Services • Engine Yard • Salesforce.com's Force.com
Infrastructure-as-a-Service	<ul style="list-style-type: none"> • Virtual servers • Logical disks • VLAN networks • Systems Management 	<ul style="list-style-type: none"> • Akamai • Amazon EC2 • CohesiveFT • Mosso (from Rackspace) • Joyent Accelerators • Nirvanix Storage Delivery Network
Physical Infrastructure	<ul style="list-style-type: none"> • Managed Hosting • Collocation • Internet Service Provider • Unmanaged hosting 	<ul style="list-style-type: none"> • GoDaddy.com • Rackspace • Savvis

Big Data

CONSULTANTS SAY
THREE QUINTILLION
BYTES OF DATA ARE
CREATED EVERY DAY.

dilbertcartoonist@gmail.com

IT COMES FROM
EVERYWHERE. IT
KNOWS ALL.

ACCORDING TO THE
BOOK OF WIKIPEDIA,
ITS NAME IS "BIG
DATA."

BIG
DATA

BIG DATA LIVES
IN THE CLOUD. IT
KNOWS WHAT WE
DO.

© 2002 Scott Adams, Inc. All rights reserved.

IN THE PAST, OUR
COMPANY DID MANY
EVIL THINGS.

BUT IF WE ACCEPT
BIG DATA IN OUR
SERVERS, WE WILL
BE SAVED FROM
BANKRUPTCY.

LET US PAY.

IS IT TOO
LATE TO
SIDE WITH
EVIL?

SHHHH!
IT HEARS
YOU.

www.dilbert.com

1-37-12

**Big Data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it...**

Dan Ariely



Are big data changing everything?

You take the **Blue Pill**,
The story ends. You wake up in your bed and believe whatever you want to believe.
You take the **Red Pill**,
You stay in Wonderland and I show You how deep the rabbit hole goes




I'm only offering You the **TRUTH**... Nothing more.

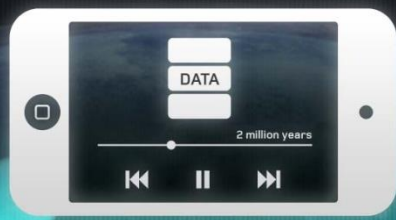
As an example: astronomy has become a data rich science, but do we grasp the depth of the problem?



As an example: astronomy has become a data rich science, but do we grasp the depth of the problem?



**SKA – first light planned 2020 –
will produce at least 1.5 PB/day
(100 PB/day are foreseeable)
Great! But it is just a number...
What does 1.5 PB mean???**



Did you know?

The data collected by the SKA in a single day would take nearly two million years to playback on an ipod.



Did you know?

The SKA will generate enough raw data to fill 15 million 64GB iPods every day!



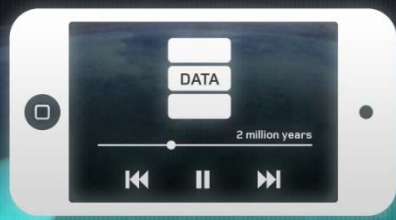
SKA WILL ALSO FILL ABOUT 1.000.000.000 AMAZON KINDLE PER DAY

The largest library in the world is the **Library of Congress**, Washington, D.C., USA with **ONLY 30.000.000 books...**

US Census Bureau (December 2010) estimates for 2020 is 7.7 billion of person...

So to SEE each day the amount of SKA data, each person in the world should read about 100.000 books per day..

ARE YOU READY FOR THIS???



Did you know?

The data collected by the SKA in a single day would take nearly two million years to playback on an ipod.



Did you know?

The SKA will generate enough raw data to fill 15 million 64GB iPods every day!



SKA WILL ALSO FILL ABOUT 1.000.000.000 AMAZON KINDLE PER DAY

The largest library in the world is the **Library of Congress**, Washington, D.C., USA with **ONLY 30.000.000 books...**

US Census Bureau (December 2010) estimates for 2020 is 7.7 billion of person...

So to SEE each day the amount of SKA data, each person in the world should read about 100.000 books per day..

ARE YOU READY FOR THIS???

AND THIS IS JUST ONE SURVEY!!!



I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I've watched c-beams glitter in the dark near the Tannhäuser Gate. All those ... *moments* will be lost in time, like tears...in rain. Time to die...

**ROY EFFECT:
(Blade Runner)
MOST DATA WILL
NEVER BE SEEN BY
HUMANS!!!**

What is big data?

Big Data is any thing which is crash Excel.

Small Data is when is fit in RAM. Big Data is when is crash because is not fit in RAM.



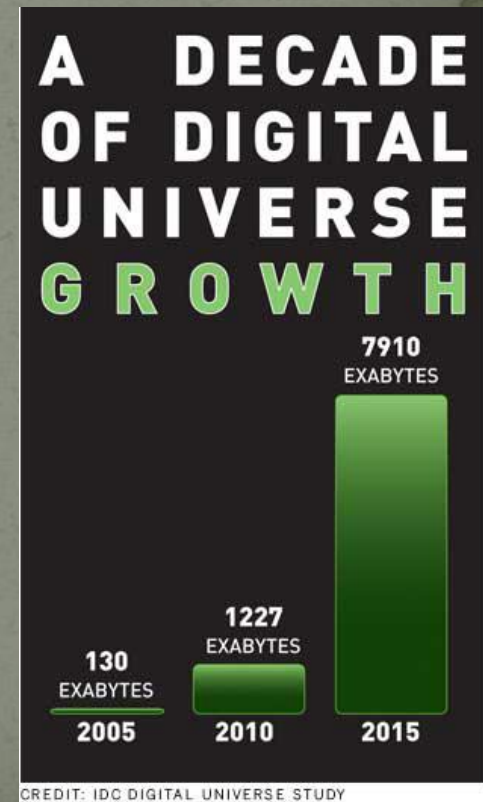
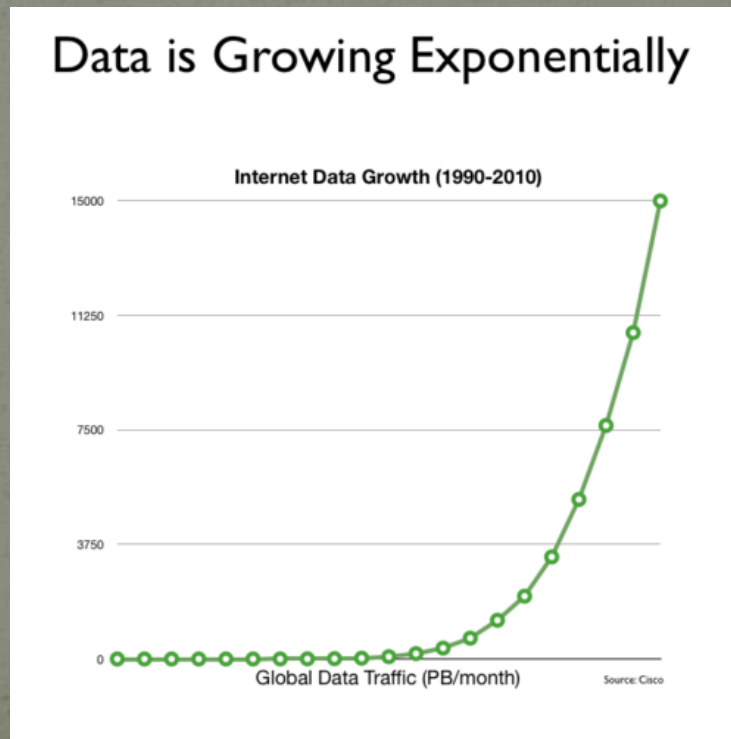
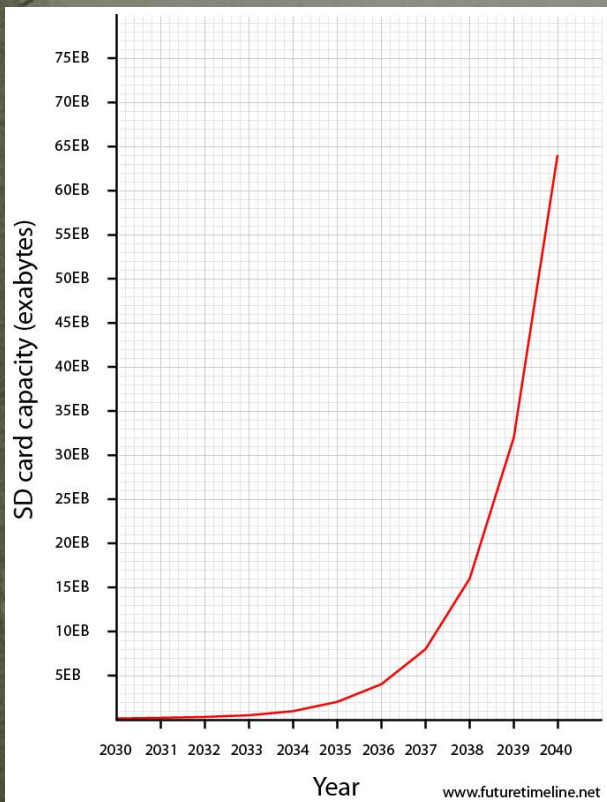
Or, in other words, Big Data is data in volumes too great to process by traditional methods.

Characteristics of Big Data

- Volume
 - data volumes are becoming unmanageable
- Variety
 - data complexity is growing
 - more types of data captured than previously
- Velocity
 - some data is arriving so rapidly that it must either be processed instantly, or lost
 - this is a whole subfield called “stream processing”

Characteristics of Big Data

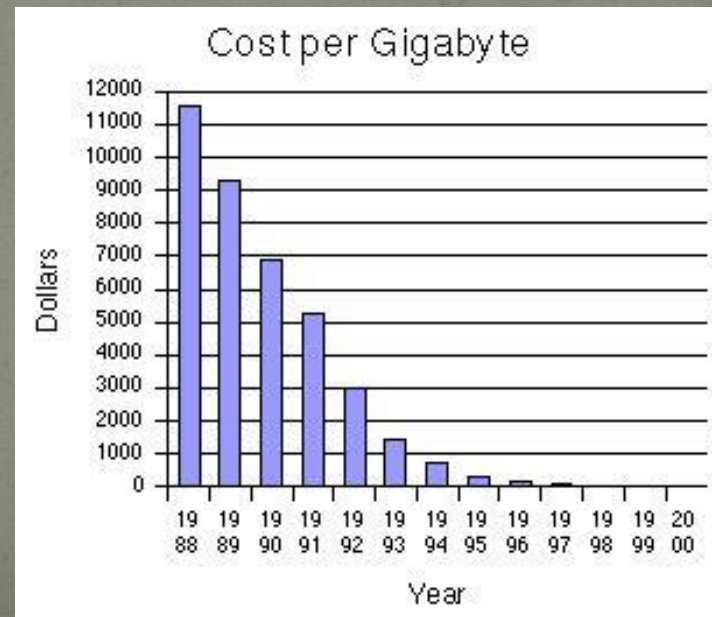
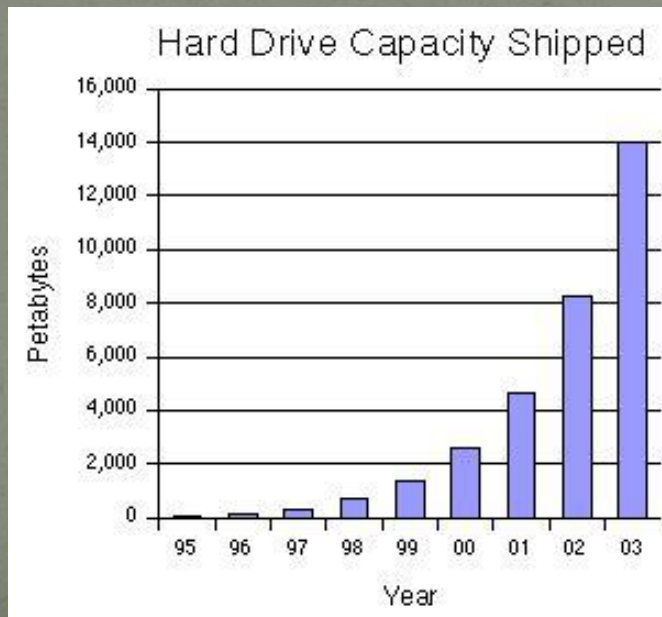
- Big quantities of data are acquired everywhere.
- It is now a big issue in all aspects of life: science, business, healthcare, gov, social networks, national security, media, etc.



Characteristics of Big Data

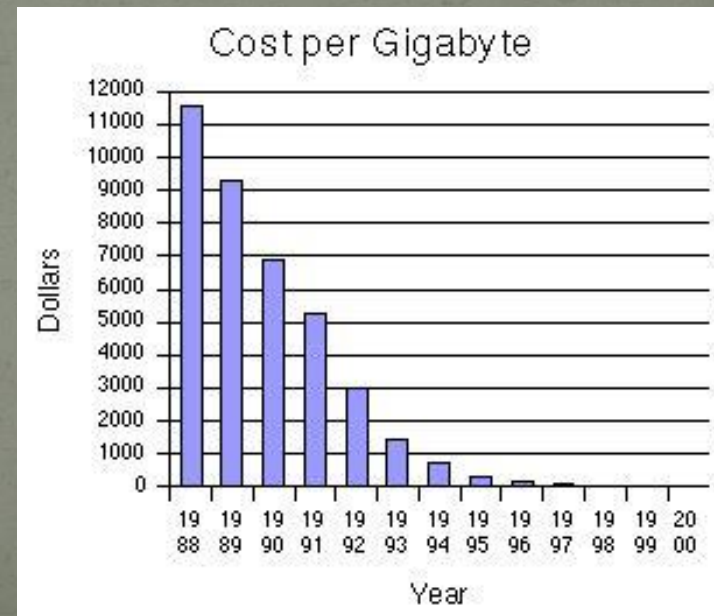
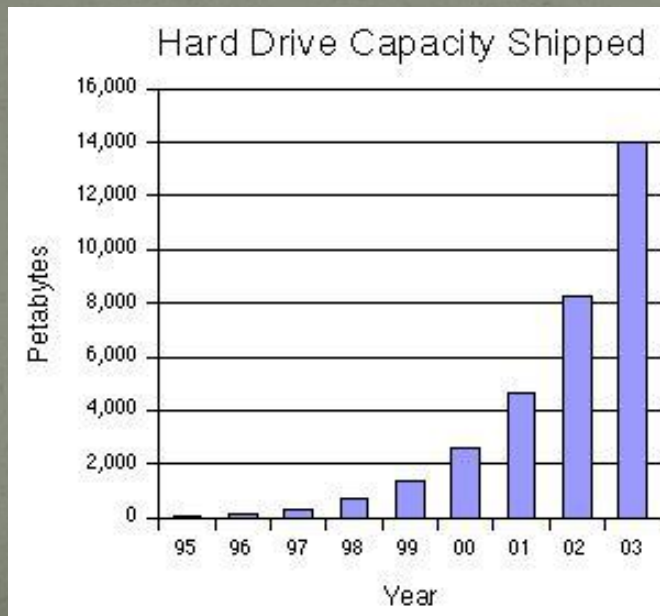
How much data are there in the world?

- From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data.
- In 2011 the same amount was created every two days.
- In 2013, the same amount is created every 10 minutes.



Characteristics of Big Data

- Computing power doubles every 18 months (Moore's Law) ...
 - **100x improvement in 10 years**
- The amount of data doubles every year (or faster!) ...
 - **1000x in 10 years, and 1,000,000x in 20 yrs.**
- I/O bandwidth increases ~10% / year
 - **<3x improvement in 10 years.**



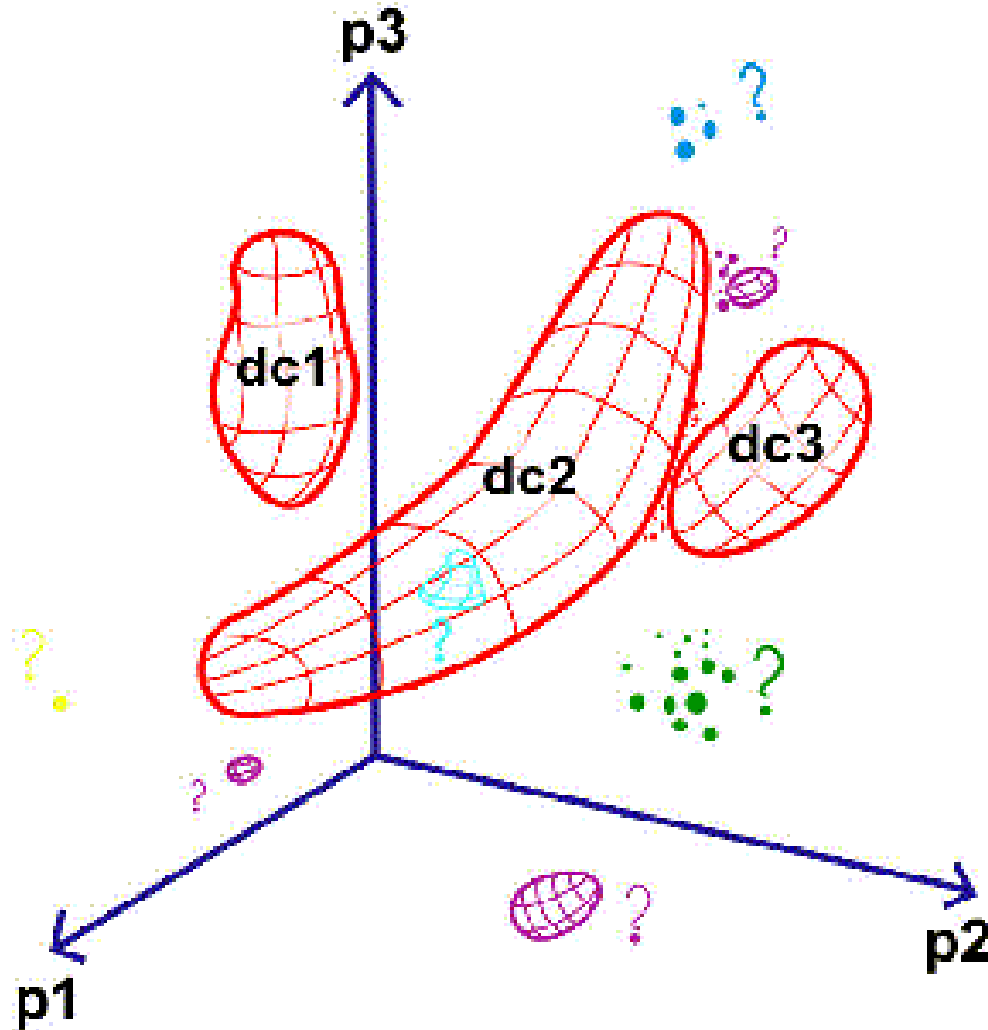
How to deal with Big data

- Do not move data over the network: **bring the computation to data!**
 - The Beowulf paradigm: Datawulf clusters, smart disks ...
 - The Grid paradigm (done right): move only the questions and answers, and the flow control
- You *will* learn to use databases!
- *Everybody* needs efficient db techniques, DM (searches, trends & correlations, anomaly/outlier detection, clustering/classification, summarization, visualization, etc.)

Types of algorithms

- Clustering
- Association learning
- Parameter estimation
- Recommendation engines
- Classification
- Similarity matching
- Neural networks
- Bayesian networks
- Genetic algorithms
- Etc.

Data Mapping and a Search for Outliers



- **Clustering** – examine the data and find the data clusters (clouds), without considering what the items are = **Characterization !**
- **Classification** – for each new data item, try to place it within a known class (i.e., a known category or cluster) = **Classify !**
- **Outlier Detection** – identify those data items that don't fit into the known classes or clusters = **Surprise !**

Two kinds of learning

- Supervised
 - we have training data with correct answers
 - use training data to prepare the algorithm
 - then apply it to data without a known answer
- Unsupervised
 - no training data
 - throw data into the algorithm, hope it makes some kind of sense out of the data

Some types of task

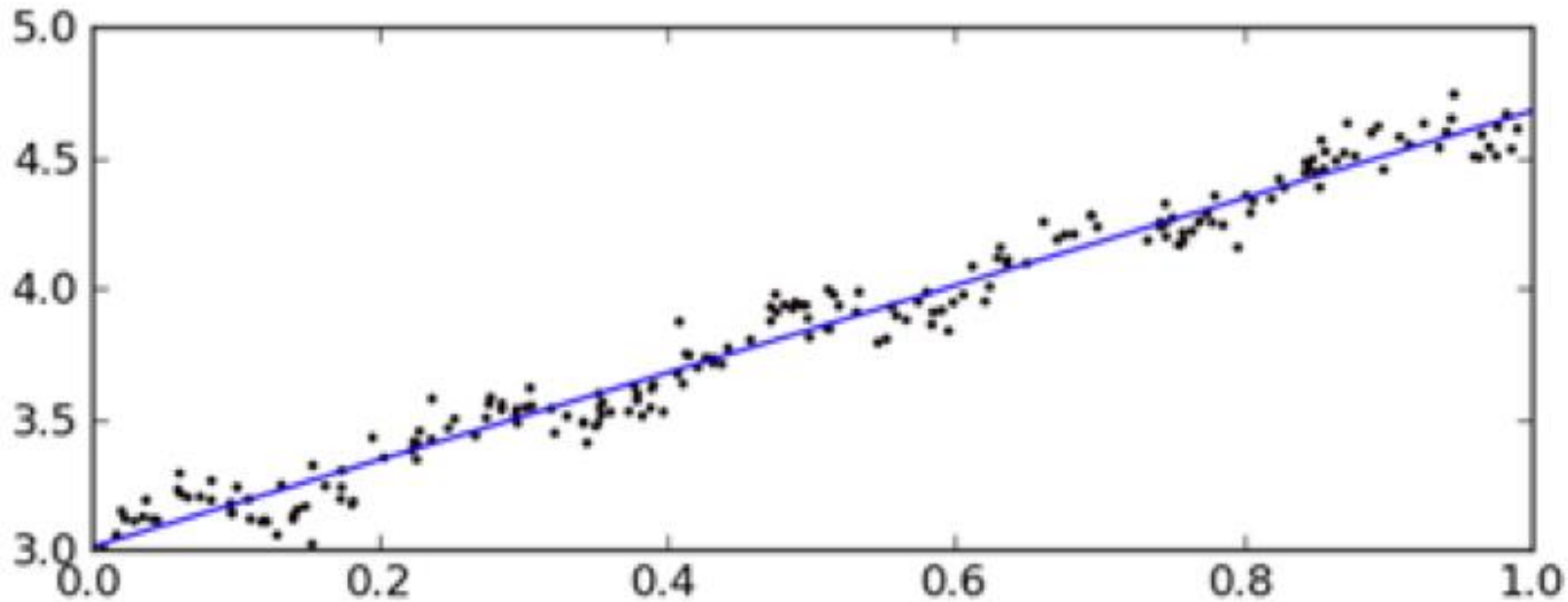
- Reduction
 - predicting a variable from data
- Classification
 - assigning records to predefined groups
- Clustering
 - splitting records into groups based on similarity
- Association learning
 - seeing what often appears together with what

Issues

- Data is usually noisy in some way
 - imprecise input values
 - hidden/latent input values
- Inductive bias
 - basically, the shape of the algorithm we choose:
 - may not fit the data at all
 - may induce underfitting or overfitting
- Machine learning without inductive bias is not possible

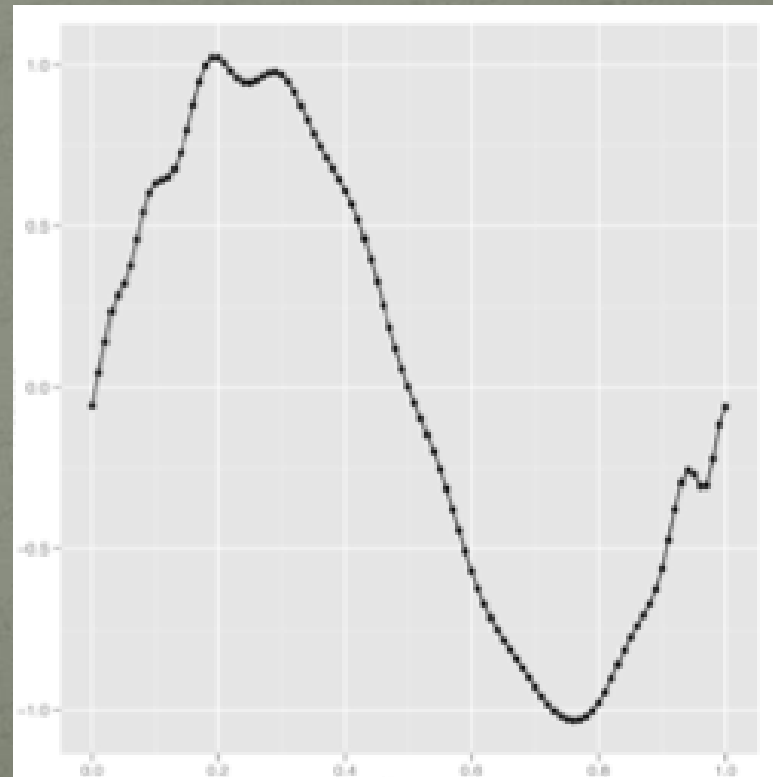
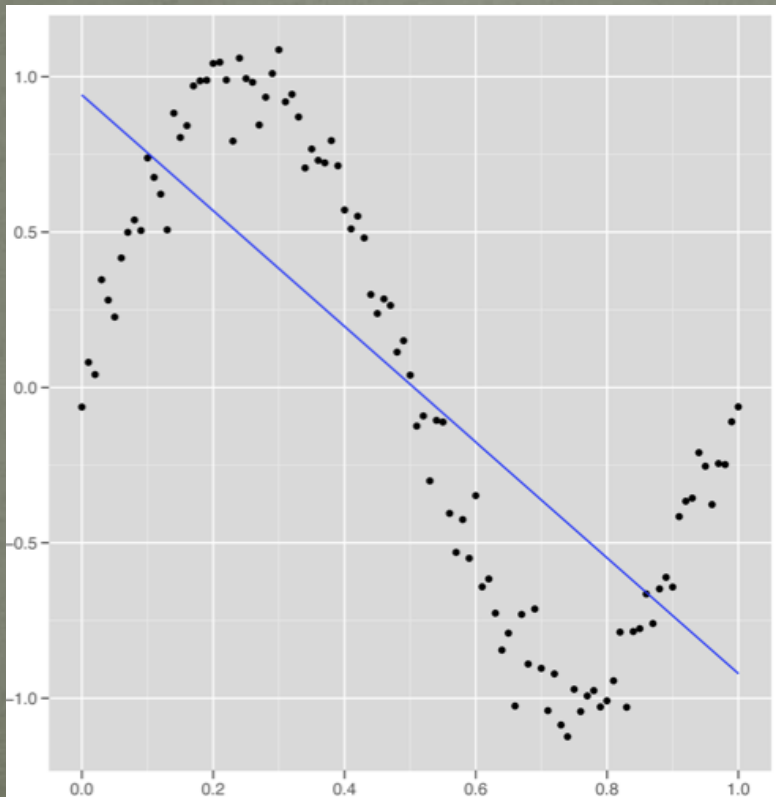
Underfitting

- Using an algorithm that cannot capture the full complexity of the data



Overfitting

- Tuning the algorithm so carefully it starts matching the noise in the training data



Good News: Big Data is Sexy

The image shows a screenshot of a web browser displaying the Harvard Business Review website. The browser's address bar shows the URL <http://hbr.org/2>. The page features the Harvard Business Review logo and a search bar. A navigation menu includes links for THE MAGAZINE, BLOGS, AUDIO & VIDEO, BOOKS, WEBINARS, and COURSES. A banner for 'Guest | limited access' is visible, along with a promotional message: 'Register today and save 20%* off your first order! Details'. The main content area displays the article title 'Data Scientist: The Sexiest Job of the 21st Century' by Thomas H. Davenport and D.J. Patil. A 'Buy Reprint »' link is present above the title. At the bottom, there are social media sharing icons for email, Twitter, LinkedIn, Facebook, Google+, and a printer icon, along with a 'Comments (39)' link.

Harvard Business Review

THE MAGAZINE | BLOGS | AUDIO & VIDEO | BOOKS | WEBINARS | COURSES

Guest | limited access Register today and save 20%* off your first order! [Details](#)

THE MAGAZINE
October 2012

[Buy Reprint »](#)

Data Scientist: The Sexiest Job of the 21st Century

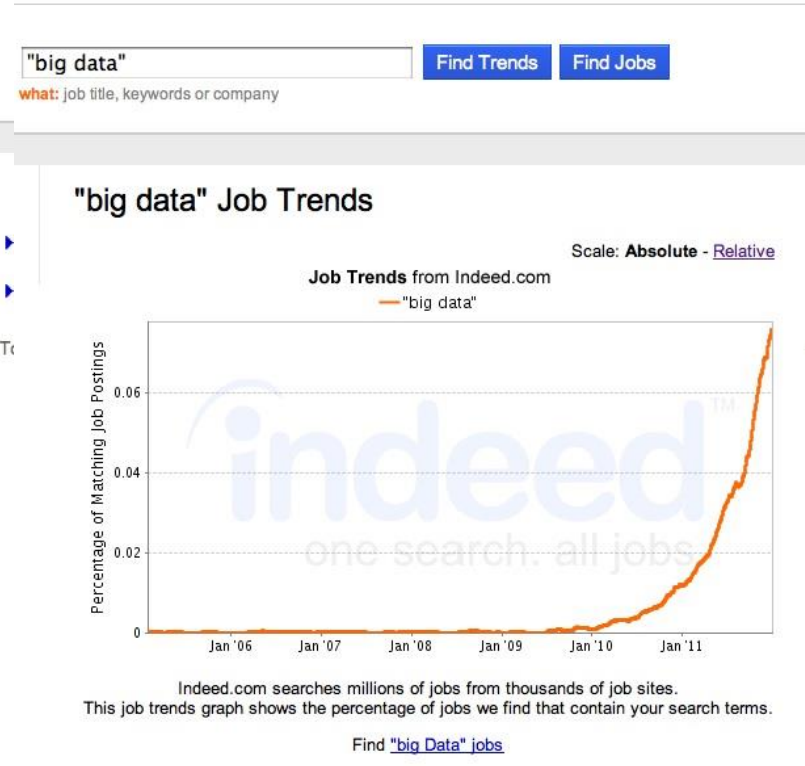
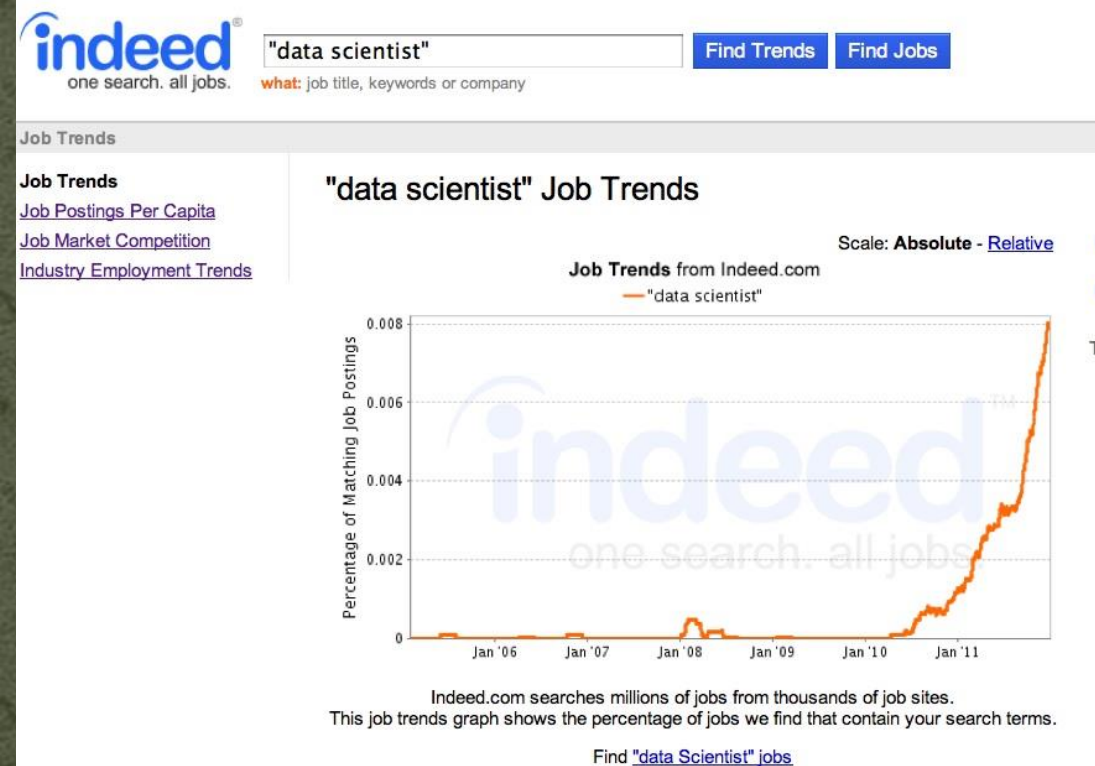
by Thomas H. Davenport and D.J. Patil

Comments (39)

[Email](#) [Twitter](#) [LinkedIn](#) [Facebook](#) [Google+](#) [Print](#)

Characteristics of Big Data

- Job opportunities are sky-rocketing
- Extremely high demand for Data Science skills
- Demand will continue to increase
- Old: “100 applicants per job”. Future: “100 jobs per applicant”



Characteristics of Big Data

- Job opportunities are sky-rocketing
- Extremely high demand for Data Science skills
- Demand will continue to increase
- Old: “100 applicants per job”. Future: “100 jobs per applicant”



Characteristics of Big Data

- Job opportunities are sky-rocketing
- Extremely high demand for Data Science skills
- Demand will continue to increase

Google

big data skills|

big data skills

big data skills shortage

big data skills gap



Job Trends

Job Tr
Job Pos
Job Ma
Industr

- Hardly anyone knows this stuff
- It's a big field, with lots and lots of theory
- And it's all maths, so it's tricky to learn



Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows the percentage of jobs we find that contain your search terms.

Find ["data Scientist" jobs](#)



Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows the percentage of jobs we find that contain your search terms.

Find ["big Data" jobs](#)

Data Scientist Definition



Josh Wills
@josh_wills

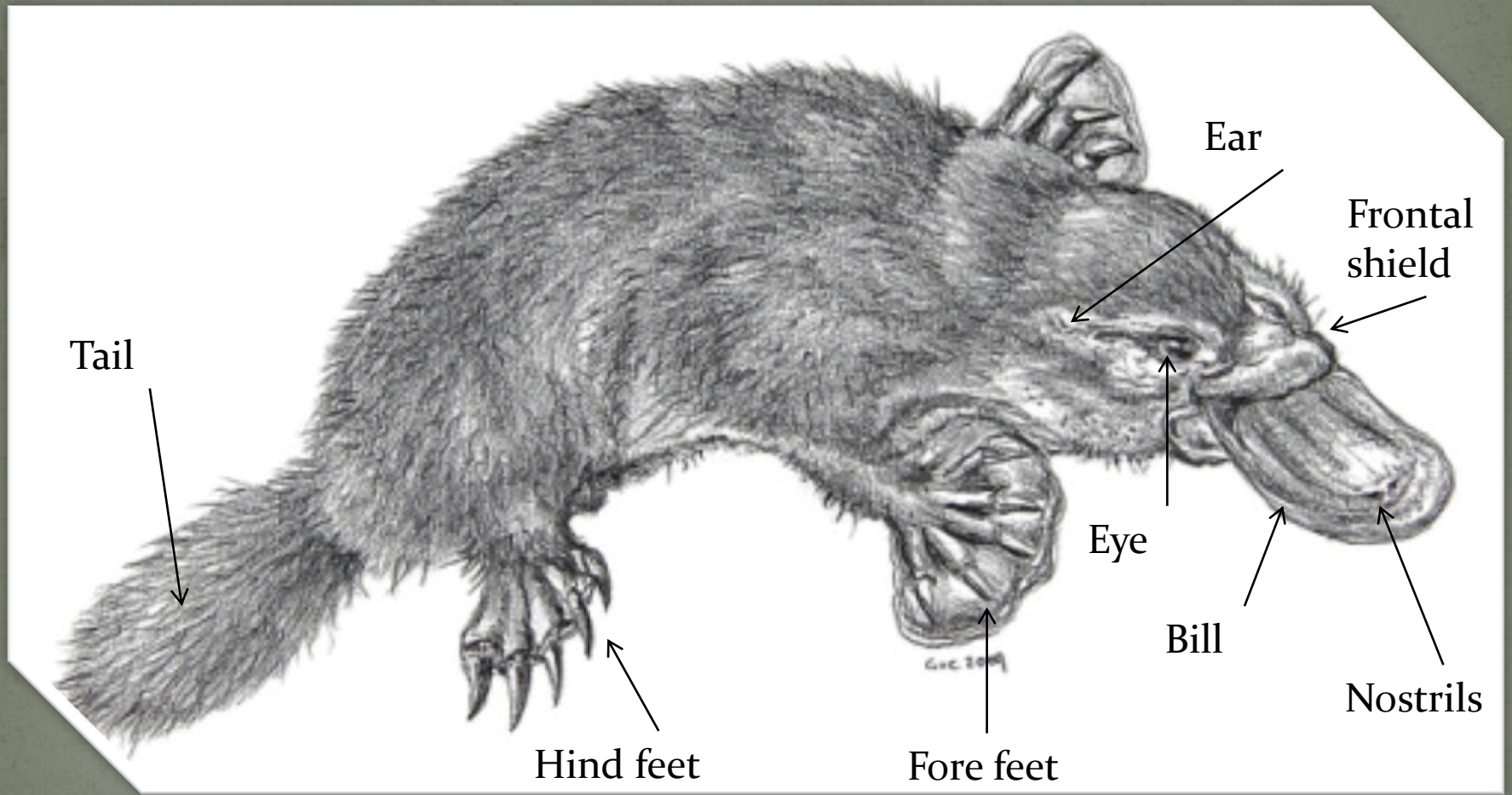
 Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

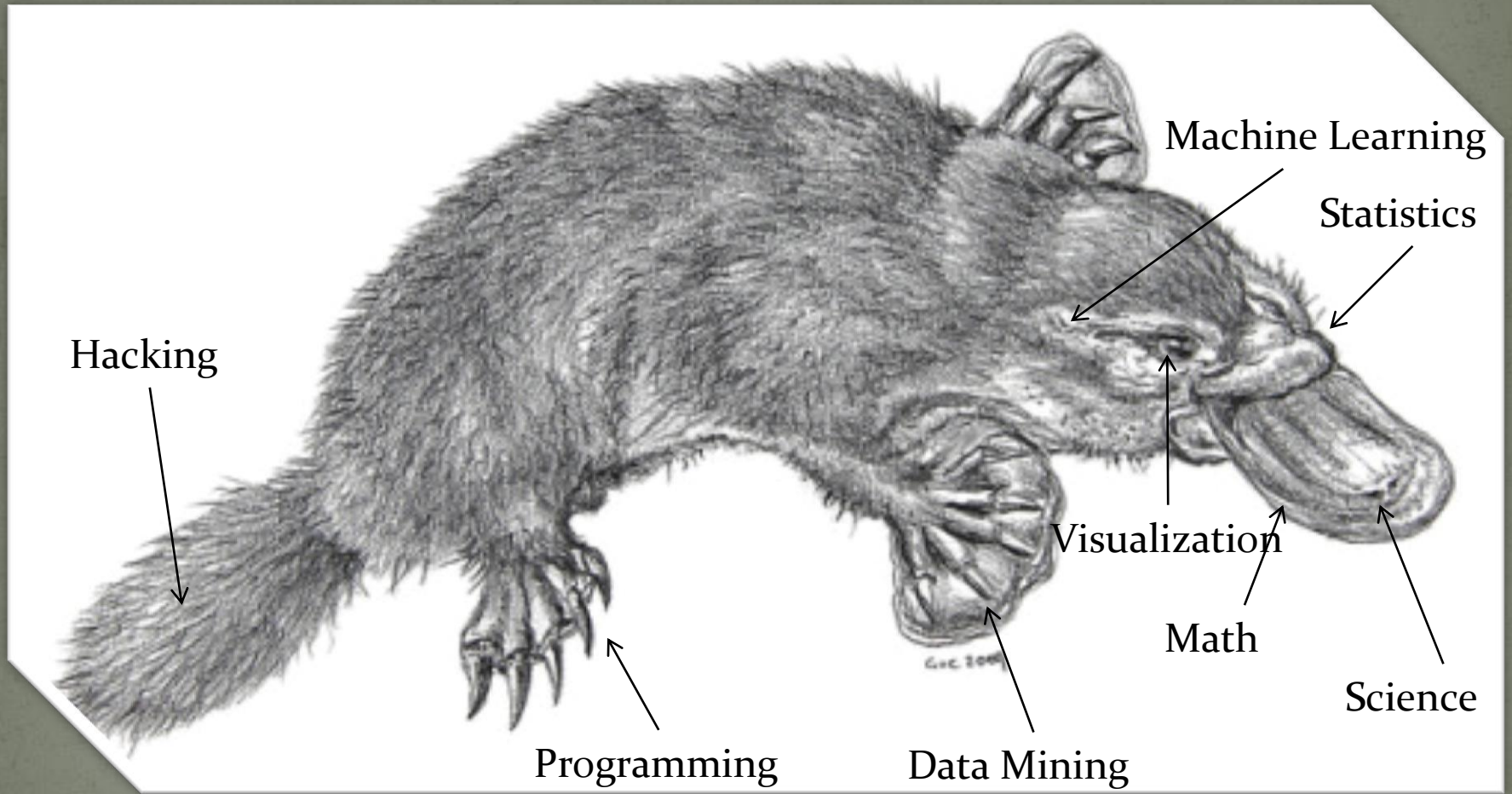
 Reply  Retweet  Favorite  More

9:55 AM - 3 May 12

Duck Billed Platypus



Data Scientist Platypus



Reality

Your boss says something vague

You think very hard on how to move the needle

Where's the data?

What's in this dataset?

What's all the f#\$!* crap in the data?

Clean the data

Run some off-the-shelf data mining algorithm

...

Productionize, act on the insight

Rinse, repeat

Conclusions, in the middle of the white Rabbit Hole...

Well, in conclusion... we have not yet (we'll never do) concluded, in reality: we just started...

- Big data are already here for some use cases and just around the corner for many other cases

THE CORE IS:

For the **Red Pill** consumers: **YES**

Big Data are changing our way to do science and/or business.
A new era of analysis has started.

For the **Blue Pill** consumers:

Don't worry, tomorrow you forget everything, you just have a little déjà vu...

**N-N-N-NO TIME, NO TIME, NO TIME!
HELLO, GOOD BYE,
I AM LATE, I AM LATE....**

JUST TIME FOR A FEW QUESTIONS!

Big Bang

Radiation era

~300,000 years:
"Dark ages" begin

~400 million years: Stars
and nascent galaxies form

~1 billion years: Dark ages end

Galaxies evolve

~9.2 billion years: Sun, Earth, and solar system have formed

~13.7 billion years: Present

