BIG DATA AND ANALYTICS

Daniel Gillblad, dgi@sics.se
Swedish Institute of Computer Science

OUR RESEARCH: SURVIVING THE DATA FLOOD

SOME QUESTIONS AROUND BIG DATA ANALYTICS

1. What is Big Data?

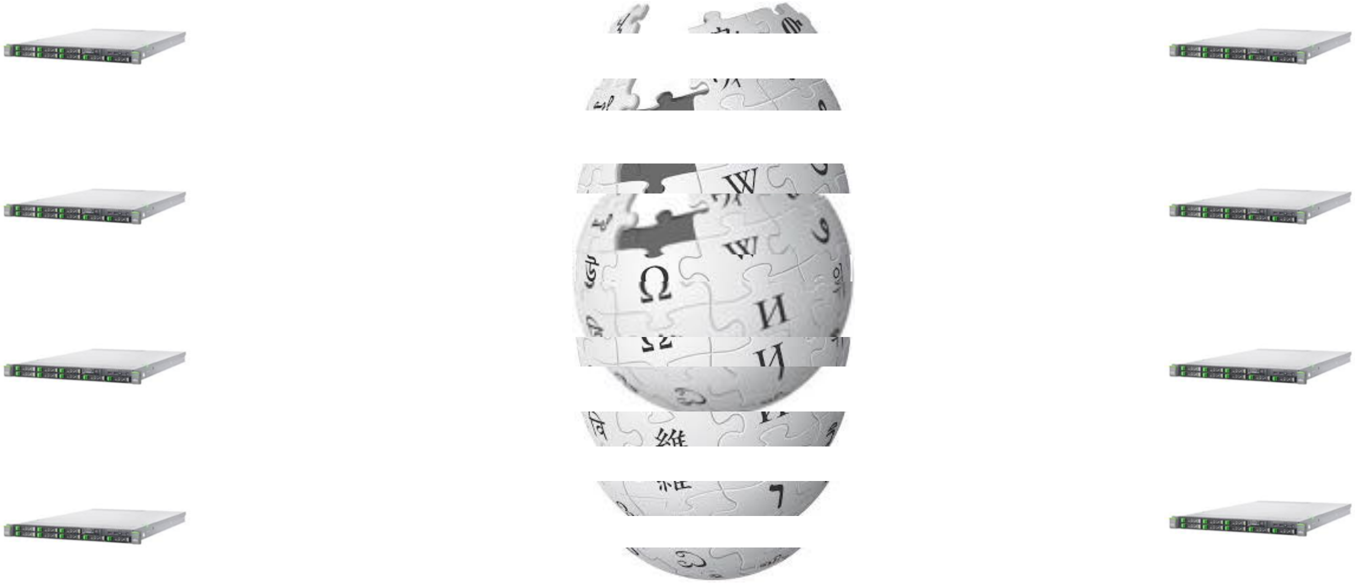
2. What is different this time?

3. Why is it important?

4. What does the future look like?

BIG DATA – 3V, AND MORE?

BIG DATA IS PARALLELIZATION



Read on 10 000 machines:
10 000 times faster

SWEDISH ICT SICS

BIG DATA IS SCALABILITY

ON COMMODITY HARDWARE

LOTS OF COMMODITY HARDWARE...
...BIG DATA IS FAULT TOLERANCE

BIG DATA IS PLATFORMS AND MIDDLEWARE

hadoop

Spark

Lightning-Fast Cluster Computing

amazon web services

Windows Azure

Storm

Distributed and fault-tolerant realtime computation

GraphLab

H_2O

DATASIFT

GNIP

FIRST, THERE WAS MAP REDUCE

- Distribute data over commodity hardware (HDFS etc.) in data center
- Map and distribute computation to computers storing the data
- Choose fix, batch-oriented form of calculation: Map -> Reduce

MAP REDUCE, EXAMPLE

- Word count: “read -> map -> reduce -> write”

```
...
public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
     private final static IntWritable one = new IntWritable(1);
     private Text word = new Text();

     public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
       String line = value.toString();
       StringTokenizer tokenizer = new StringTokenizer(line);
       while (tokenizer.hasMoreTokens()) {
         word.set(tokenizer.nextToken());
         output.collect(word, one);
       }
     }
}

public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
     public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter reporter)
     throws IOException {
       int sum = 0;
       while (values.hasNext()) {
         sum += values.next().get();
       }
       output.collect(key, new IntWritable(sum));
     }
}
...
```

NEXT-GEN MAP REDUCE

Spark

Lightning-Fast Cluster Computing

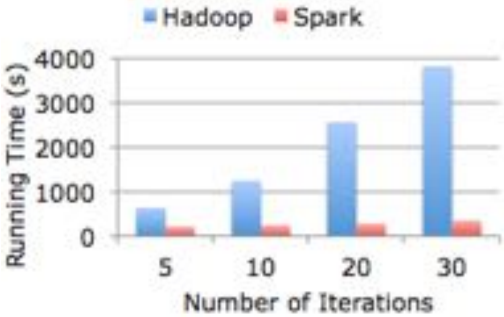
- Handle real-time, interactive and iterative tasks: Necessary for machine learning etc.
- Resilient distributed data sets – in-memory caching etc.

1. Transformations: Map, filter, join…
2. Actions: Count, collect, save…

- Faster, easier, more powerful

```
file = spark.textFile("hdfs://...")

file.flatMap(line => line.split(" "))
    .map(word => (word, 1))
    .reduceByKey(_ + _)
```



SWEDISH ICT SICS

SPARK, REAL-WORLD EXAMPLE

- Incomprehensible example:

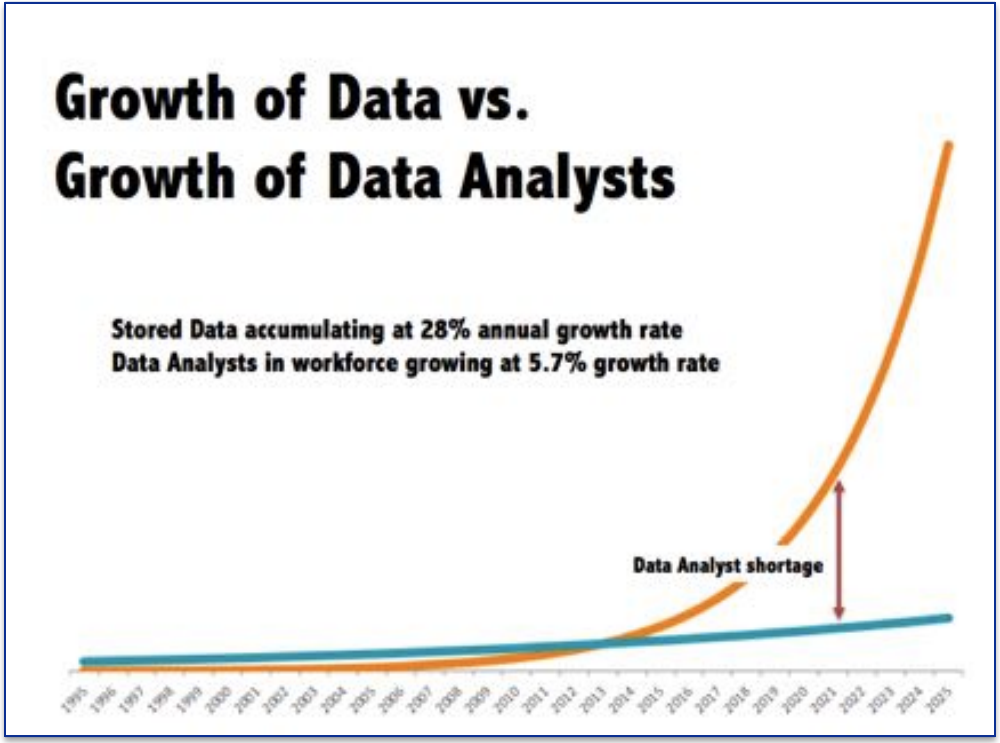
```
logs.map(log => (log.data, log)).groupByKey()
        .filter(_._2.size >= minimumValue)
        .map(xLogsPair => new Aggregate(Pair(xLogsPair._1,
         logsExpanded(xLogsPair._2))))
```

- Defines both *what/how* and hints on *how to parallelize*

IS THAT IT?

- Still not that easy to write – special kind of developers
- Fantastic productivity in the right hands
- Developer support limited
- Parallelization is difficult
- Pretty difficult to write advanced analytics

WHO USES THIS?

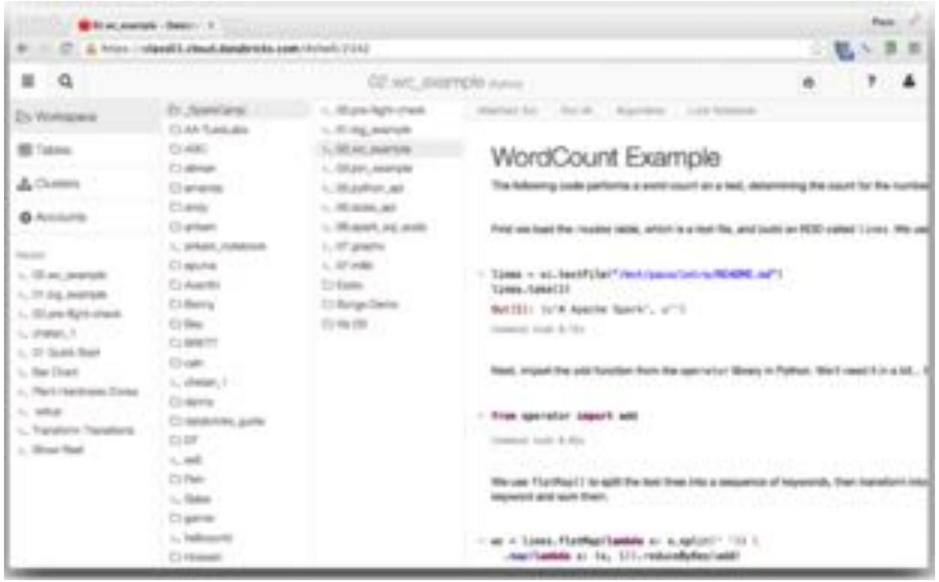


http://www.delphianalytics.net

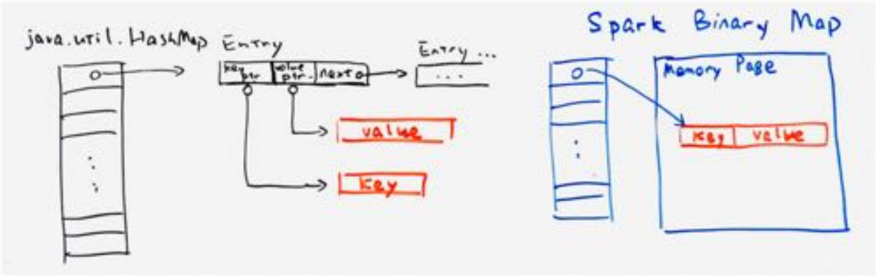
SCALABILITY IS OFTEN NO LONGER THE MAIN ISSUE – ANALYTICS, EFFICIENCY AND PRODUCTIVITY ARE

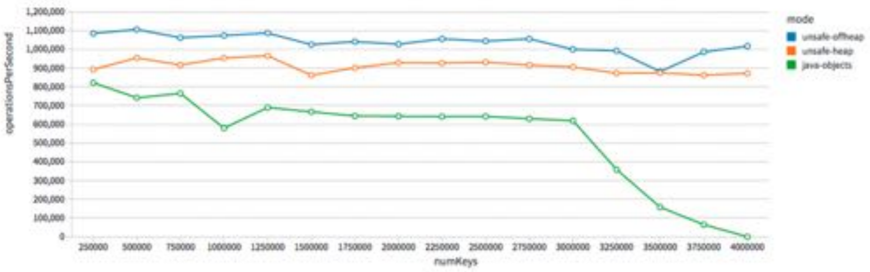
ACCESSIBLE BIG DATA TOOLS?

- Higher-level abstractions (RDDs -> Dataframes etc.)
- R and Python interfaces
- Libraries:
  - Spark MLib, GraphX
  - FlinkML, Gelly
  - …
- Higher level languages
- Imperative Big Data processing
- …



RAW PERFORMANCE MATTERS





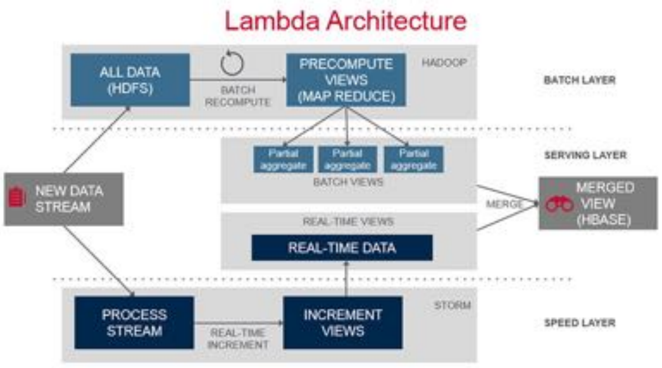
https://databricks.com/blog/2015/04/28/project-tungsten-bringing-spark-closer-to-bare-metal.html

http://bid2.berkeley.edu/bid-data-project/

DATA AT REST...

… TO DATA IN MOTION

FROM LAMBDA ARCHITECTURE…



https://www.mapr.com/sites/default/files/otherpageimages/lambda-architecture-2-800.jpg

…TO UNIFIED ARCHITECTURE

WHAT IS DIFFERENT THIS TIME?

- Haven't we seen this before? AI, Data Mining etc.?
- Things are a bit different this time around:
  1. We have the computational power and middleware (cloud, Big Data middleware)
  2. We have the data (Human generated, IoT)
  3. We have the algorithms

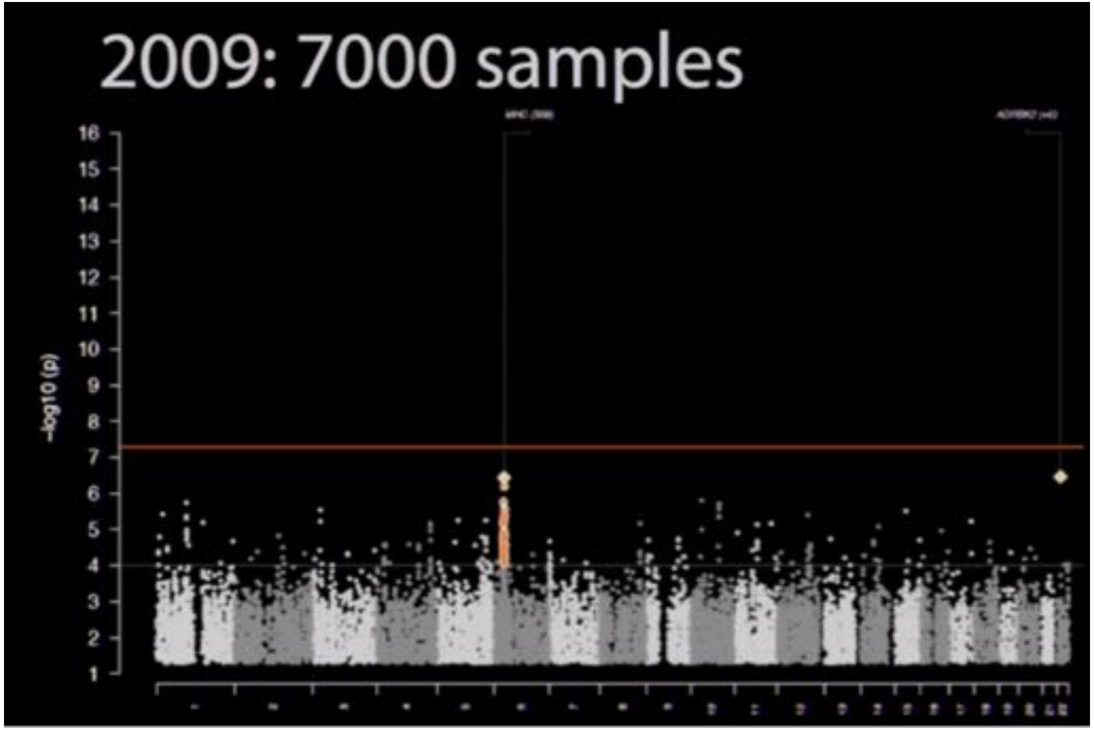
THE UNREASONABLE EFFECTIVENESS OF DATA

- In a wide array of academic fields, the ability to effectively process data is superseding other more classical modes of research.

*"More data trumps better algorithms"**

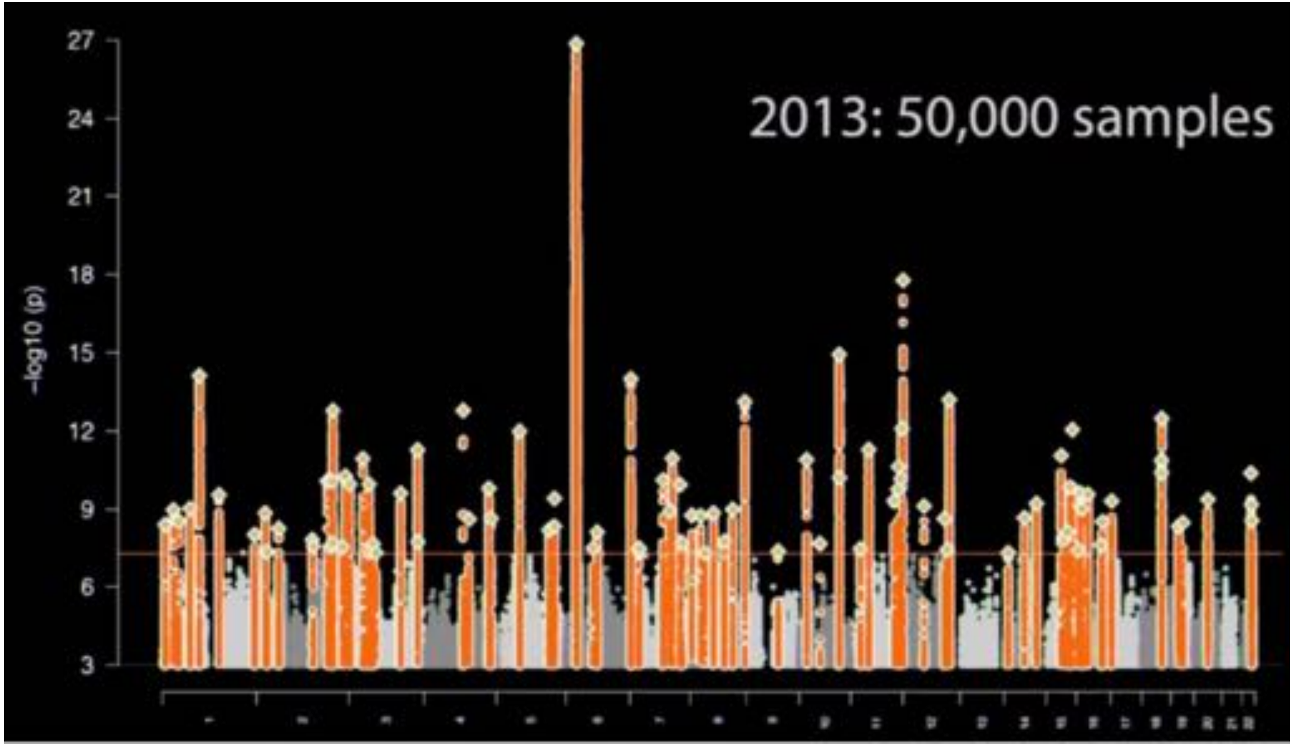
*"The Unreasonable Effectiveness of Data" [Halevey et al 09]

EXAMPLE: POPULATION SCALE GENOMICS

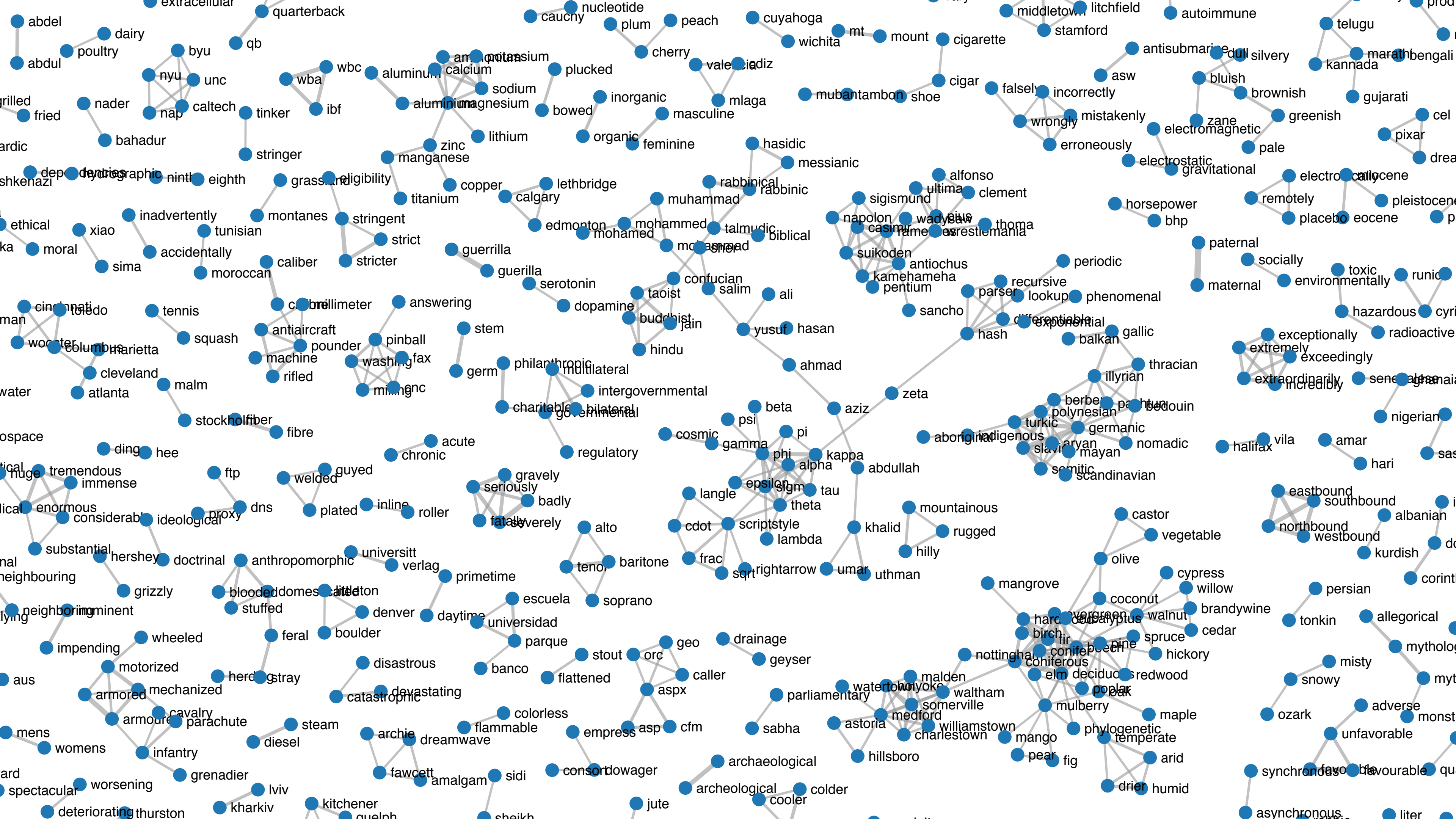


[Image source: Patterson, Fighting the Big C with the Big D, 2014]

EXAMPLE: POPULATION SCALE GENOMICS



[Image source: Patterson, Fighting the Big C with the Big D, 2014]



abdel
abdul
dairy
poultry
byu
nyu
unc
nap
caltech
fried
nader
bahadur
quarterback
qb
tinker
stringer
wba
wbc
ibf
aluminum
calcium
potassium
sodium
magnesium
lithium
zinc
manganese
copper
titanium
caucny
nucleotide
plum
peach
cherry
plucked
bowed
inorganic
organic
masculine
feminine
cuyahoga
wichita
valencia
cadiz
mlaga
mt
mount
cigarette
cigar
shoe
middletown
litchfield
stamford
autoimmune
antisubmarine
asw
falsely
incorrectly
wrongly
mistakenly
erroneously
dull
silvery
bluish
brownish
greenish
pale
zane
electromagnetic
electrostatic
gravitational
telugu
kannada
marathi
bengali
gujarati
pixar
eighth
eligibility
montanes
stringent
strict
stricter
calgary
lethbridge
edmonton
hasidic
messianic
rabbinical
rabbinic
muhammad
mohammed
mohamed
talmudic
biblical
alfonso
clement
sigismund
napoleon
casimir
suikoden
antiochus
kamehameha
pentium
sancho
thoma
horsepower
bhp
remotely
placebo
eocene
pleistocene
ethical
moral
xiao
sima
inadvertently
accidentally
tunisian
moroccan
caliber
guerrilla
guerilla
serotonin
dopamine
taoist
confucian
buddhist
jain
hindu
salim
ali
yusuf
hasan
ahmad
paternal
maternal
socially
environmentally
toxic
hazardous
radioactive
runic
periodic
recursive
parser
lookup
phenomenal
differentiable
exponential
hash
gallic
balkan
thracian
illyrian
exceptionally
extremely
exceedingly
extraordinarily
incredibly
nigerian
cincinnati
toledo
worcester
columbus
marietta
cleveland
atlanta
tennis
squash
malm
stockholm
fiber
fibre
millimeter
antiaircraft
pounder
machine
rifled
answering
pinball
washing
fax
mining
cnc
stem
germ
philanthropic
multilateral
intergovernmental
charitable
bilateral
governmental
regulatory
zeta
aziz
beta
psi
cosmic
gamma
pi
phi
kappa
alpha
epsilon
sigma
tau
theta
langle
cdot
scriptstyle
lambda
frac
sqrt
rightarrow
abdullah
khalid
umar
uthman
berber
polynesian
pashtun
bedouin
turkic
germanic
aboriginal
indigenous
aryan
mayan
nomadic
semitic
scandinavian
halifax
vila
amar
hari
ding
hee
tremendous
huge
immense
enormous
considerable
substantial
acute
chronic
ftp
dns
proxy
ideological
welded
guyed
plated
inline
roller
gravely
seriously
badly
fatally
severely
mountainous
rugged
hilly
castor
vegetable
olive
eastbound
southbound
northbound
westbound
albanian
kurdish
hershey
doctrinal
grizzly
anthropomorphic
universitt
verlag
primetime
alto
baritone
tenor
soprano
blooded
domesticated
stuffed
littleton
denver
boulder
feral
daytime
escuela
universidad
parque
banco
neighbouring
imminent
impending
wheeled
motorized
armored
mechanized
armoured
cavalry
parachute
infantry
grenadier
mens
womens
aus
herd
stray
disastrous
catastrophic
devastating
steam
diesel
stout
orc
geo
caller
aspx
asp
cfm
flattened
drainage
geyser
parliamentary
sabha
mangrove
cypress
willow
coconut
brandywine
walnut
cedar
eucalyptus
evergreen
birch
fir
conifer
beech
pine
spruce
hickory
coniferous
elm
deciduous
redwood
poplar
oak
mulberry
maple
phylogenetic
temperate
mango
pear
fig
arid
drier
humid
nottingham
watertown
holyoke
malden
waltham
somerville
medford
astoria
williamstown
charlestown
hillsboro
persian
tonkin
allegorical
misty
snowy
ozark
adverse
unfavorable
favourable
synchronous
asynchronous
colorless
flammable
archie
dreamwave
fawcett
amalgam
sidi
empress
consort
dowager
archaeological
archeological
colder
cooler
jute
worsening
spectacular
deteriorating
thurston
lviv
kharkiv
kitchener
guelph
sheikh
liter

BIG DATA – A BIG MISTAKE?

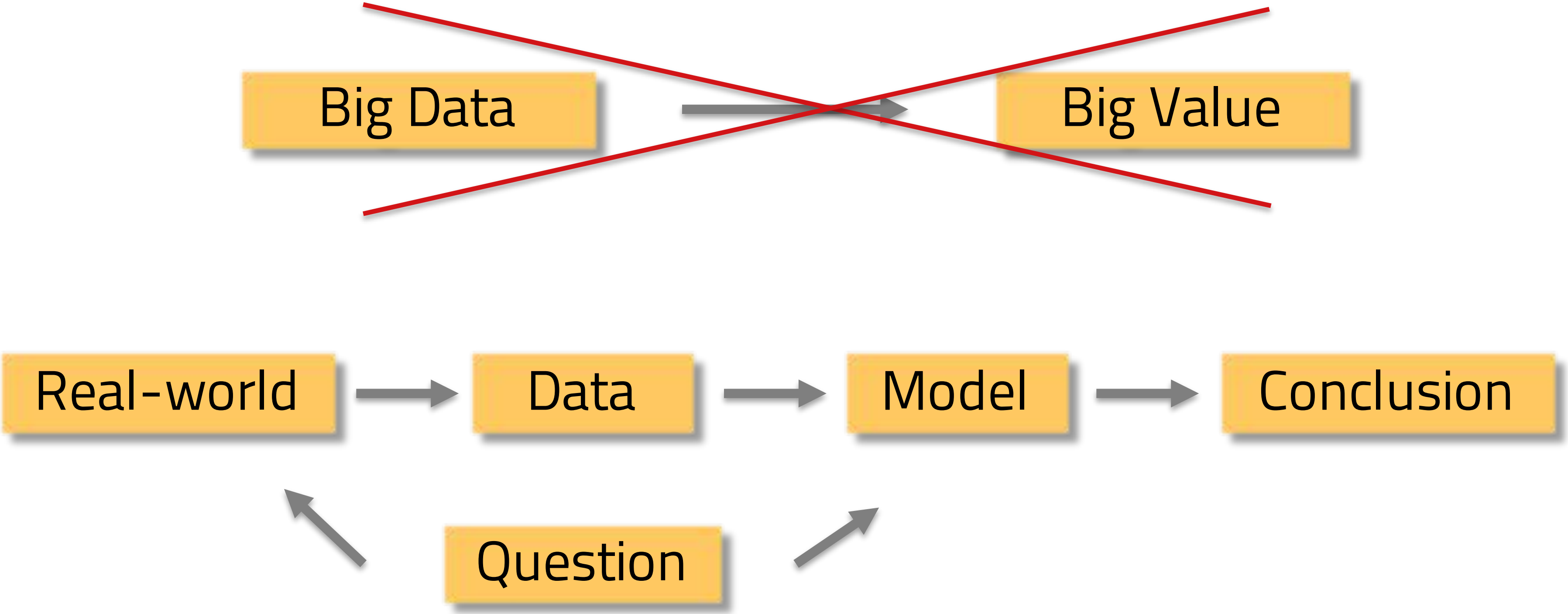
- "Four oversimplified articles of faith:"
  1. Data analysis produces uncannily accurate results
  2. Every single data point can be captured, making old statistical sampling techniques obsolete
  3. It is passé to fret about what causes what, because statistical correlation tells us what we need to know
  4. "The End of Theory": With enough data, the numbers speak for themselves
- Practitioners do not necessarily believe this, and neither should you

From *Big Data: are we making a big mistake?*, Tim Harford, Financial Times March 28 2014

BIG DATA TO BIG VALUE?

Big Data → Big Value

BIG DATA TO BIG VALUE?



BDA: WORKFLOW NOT FUNDAMENTALLY DIFFERENT

Acquisition

Cleaning

Representation

Case-based | Statistical | Logical | Kernel-based

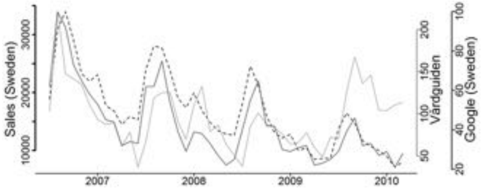
Validation

Deployment

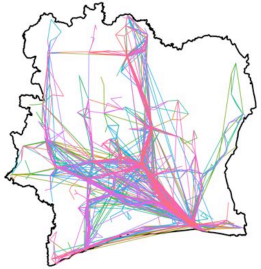
EVERYTHING IS ABOUT MACHINE LEARNING ANYWAY…

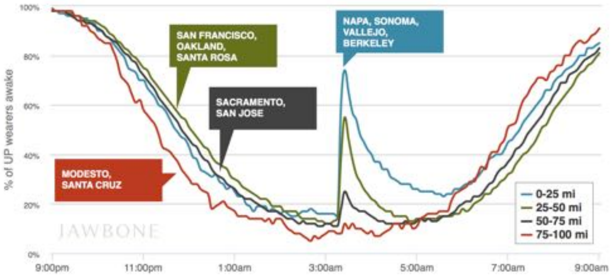
A Few Useful Things to Know About Machine Learning
Pedro Domingos
CACM 55:10
http://dl.acm.org/citation.cfm?id=2347755

LET'S START WITH THE OBVIOUS



Lindh, J *et al.*: *PLoS ONE* 7(11), 2012.

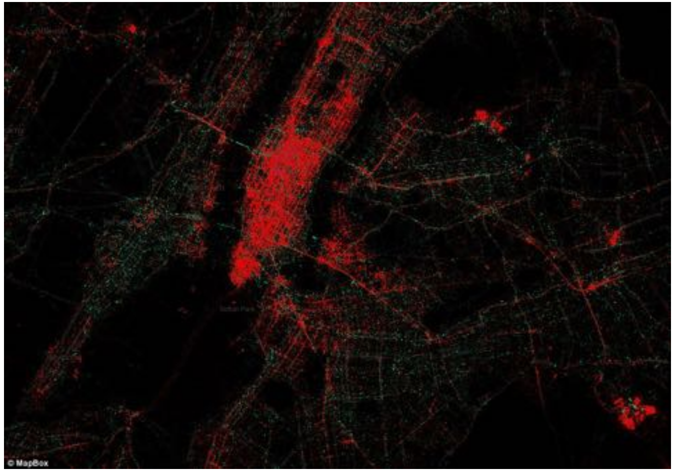




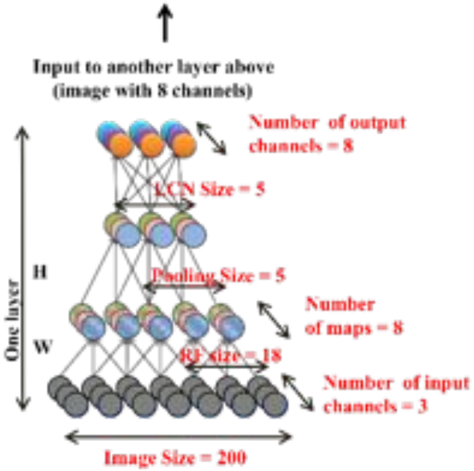
It is a key component of digital services

It will help us understand the world better

...BUT IT IS REALLY, REALLY DIFFICULT



Sample bias, data veracity



Model complexity

...+ user and analyst bias...

SCALABLE, PRIVACY-PRESERVING MOBILITY ANALYTICS FROM NETWORK DATA

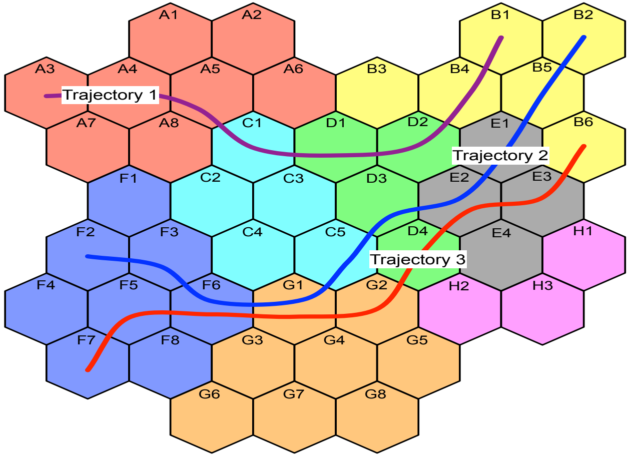
Urban planning

Traffic management

Crisis management

Consumer applications

Network management



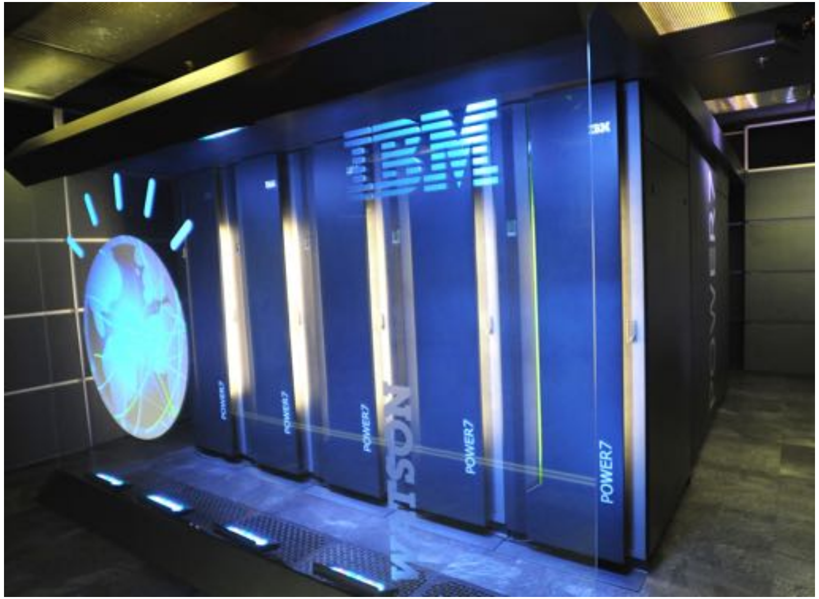
A3, A4, A5, ***C1***, ***D1***, D2, ***B4***, B1
F2, F3, F6, ***G1***, ***C5***, ***D3***, ***E2***, E1, ***B5***, B2
F7, F5, F6, ***G1***, G2, ***D4***, E2, E3, ***B6***

AN UNSOLVABLE PROBLEM?

- Moving datasets can be difficult
  - Large and sensitive (business value, integrity issues, etc.)
  - Can allow for re-identification in anonymized data
- Combining several datasets outside their collection entity is key to many (public) applications, *but…*

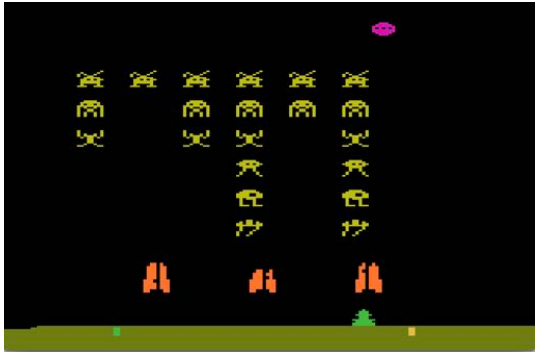
Centralized datasets:	*Technically feasible*	*Politically hard*
Federated datasets:	*Technically difficult*	*Politically solvable*

BIG DATA: KEY COMPONENT IN AUTOMATION

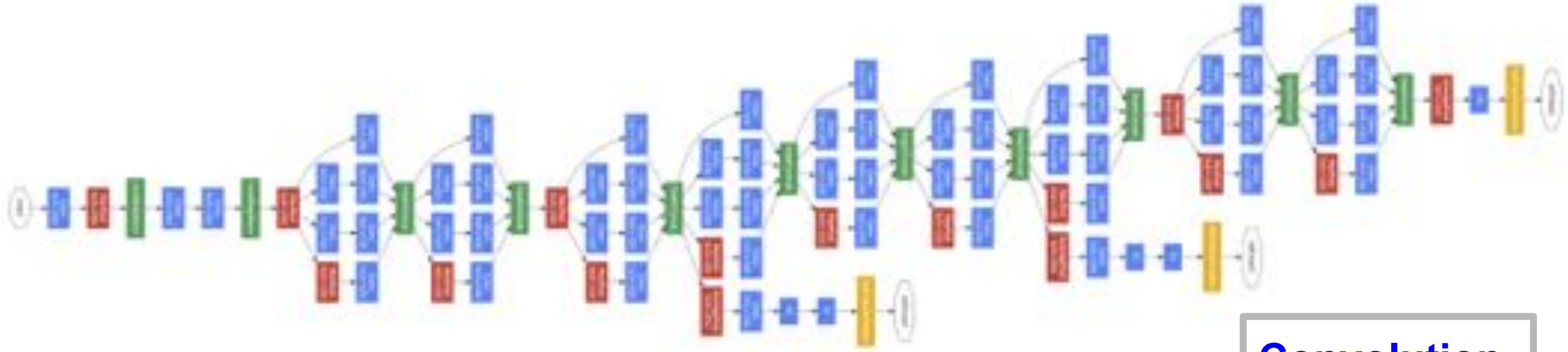


SWEDISH ICT SICS

MASSIVE LEARNING SYSTEMS

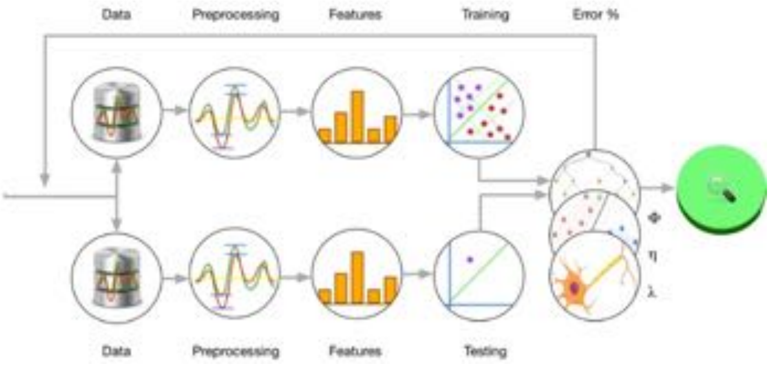


LARGER AND LARGER REPRESENTATIONS



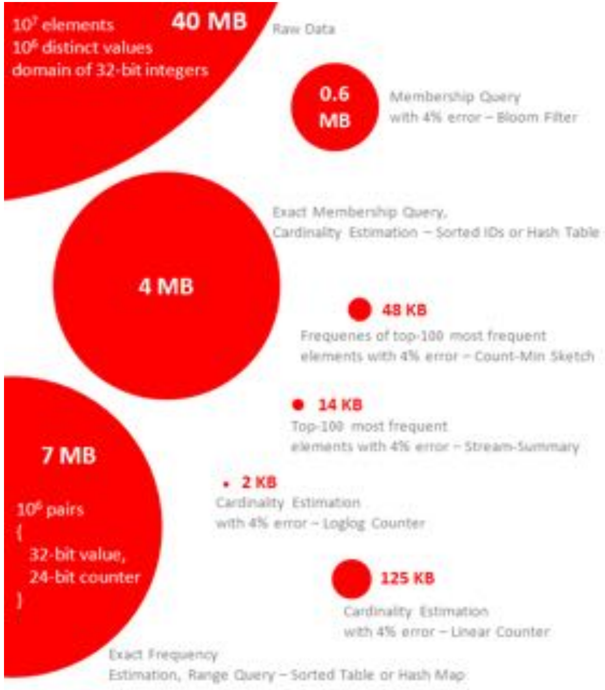
Convolution
Pooling
Softmax
Other

UNIFICATION OF MACHINE LEARNING?



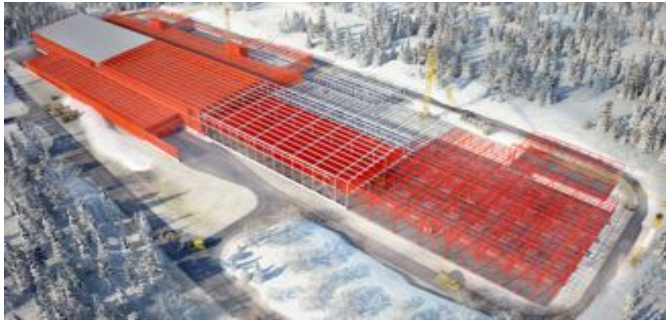
Alternating Direction Method of Multipliers
S. Boyd, N. Parikh, et al.
http://stanford.edu/~boyd/papers/admm_distr_stats.html

MOVING TO PROBABILISTIC APPROXIMATIONS



https://highlyscalable.wordpress.com/2012/05/01/probabilistic-structures-web-analytics-data-mining/

ANALYTICS EVERYWHERE



SWEDISH ICT SICS

ANALYTICS AS A SERVICE

Gavagai MONITOR

Gavagai API

Gavagai PRO SERVICES

WHERE ARE WE GOING?

① Advanced analytics is moving towards large-scale Machine Learning

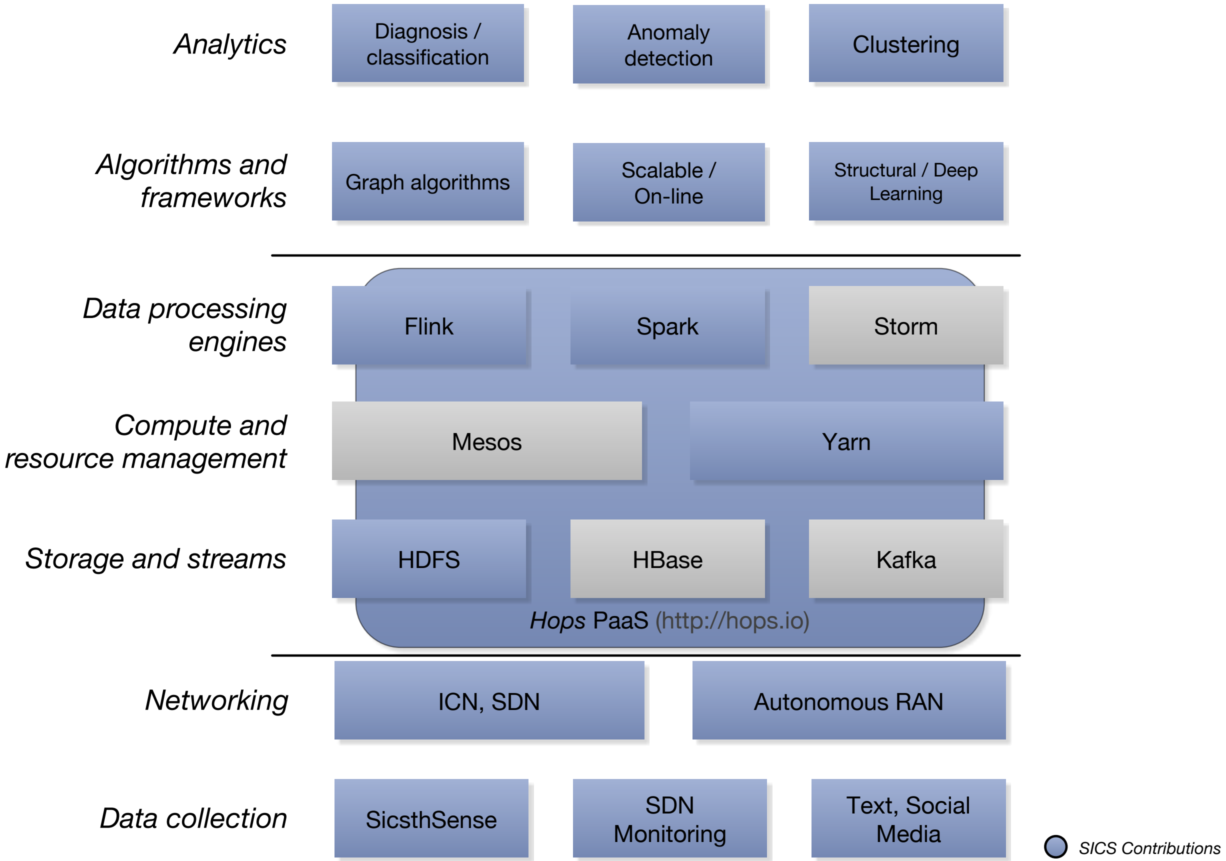
② Computation and storage platforms need to adapt and develop

⇩

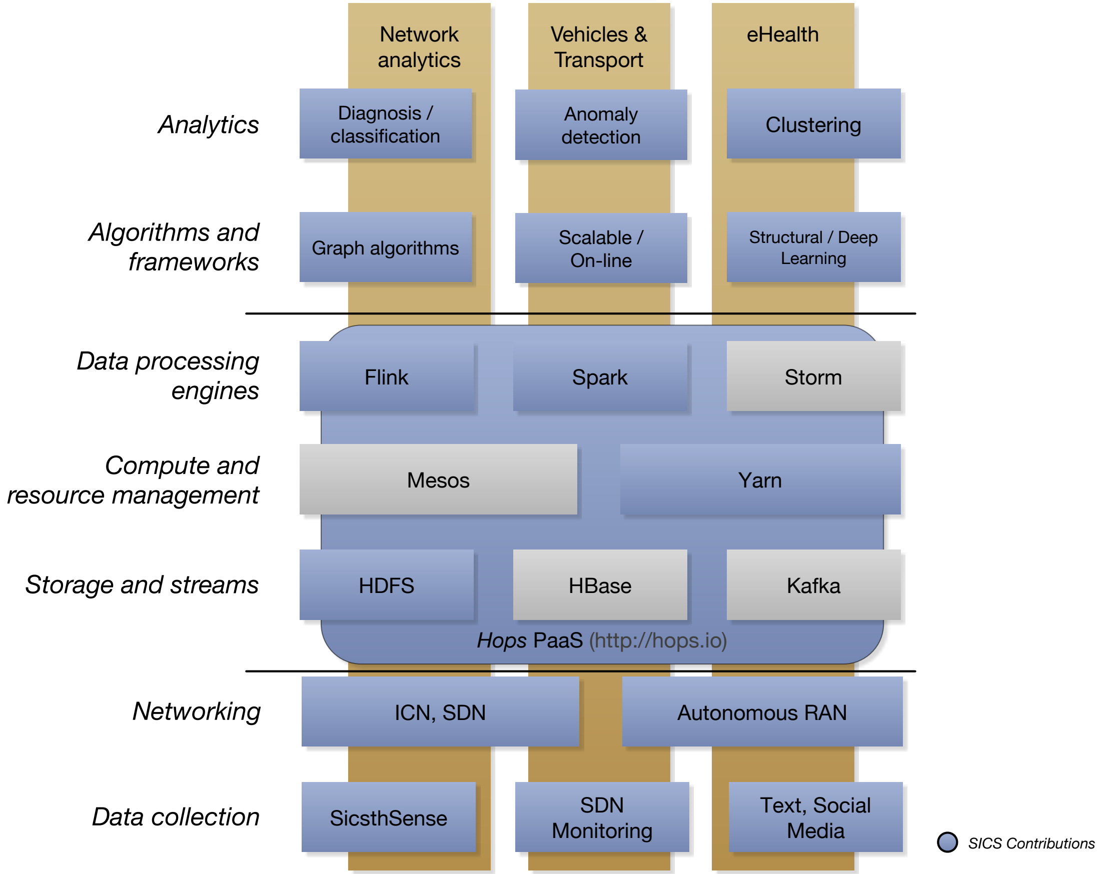
LearningMachines@SICS

Analytics and system development on real data and use cases

THE DATA DRIVEN SYSTEMS STACK, SICS



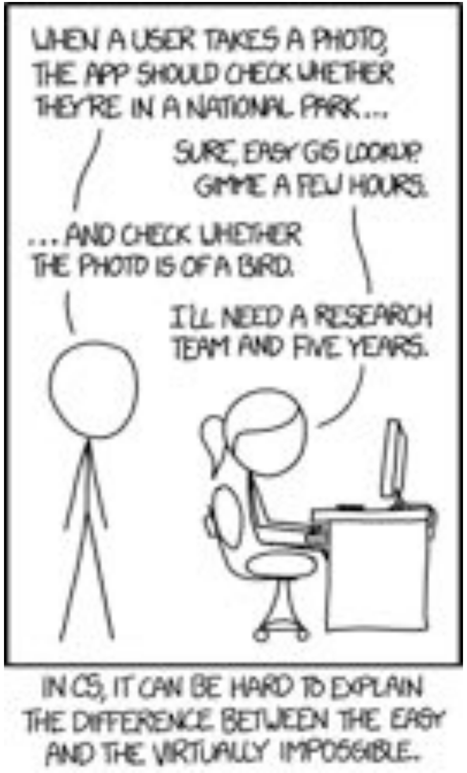
DATA DRIVEN SYSTEMS, APPLICATION EXAMPLES



EXPERIENCE FROM APPLICATIONS

- Big Data can be *made* small – consider the complete application
- Big Data can become small very quickly
- Beware of sample bias
- Distribute models, not data
- Distributed solutions, Statistical Machine Learning, and Bayesian statistics can help

DISCUSSING APPLICATIONS: IT IS NOT OBVIOUS WHAT IS DIFFICULT

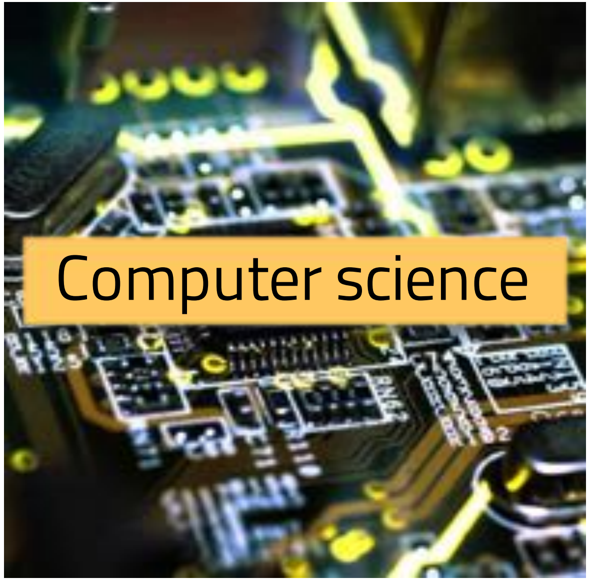


SWEDISH ICT SICS

WHAT ABOUT INTEGRITY?

- Not all Big Data data is integrity sensitive!
  - Media, measurements, science, ...
- How data is (or potentially is) used is everything
- Surveillance and data driven services are very different
- It all comes down to trust
  - Transparency is key
  - Laws and regulations? How do we manage sales, sharing?

THE END OF COMPUTER SCIENCE



…BUT JUST REMEMBER…

We're not there yet – things are moving fast!

WWW.SICS.SE