CERN

Scientific computing in the era of big data …

Alberto Pace, alberto.pace@cern.ch

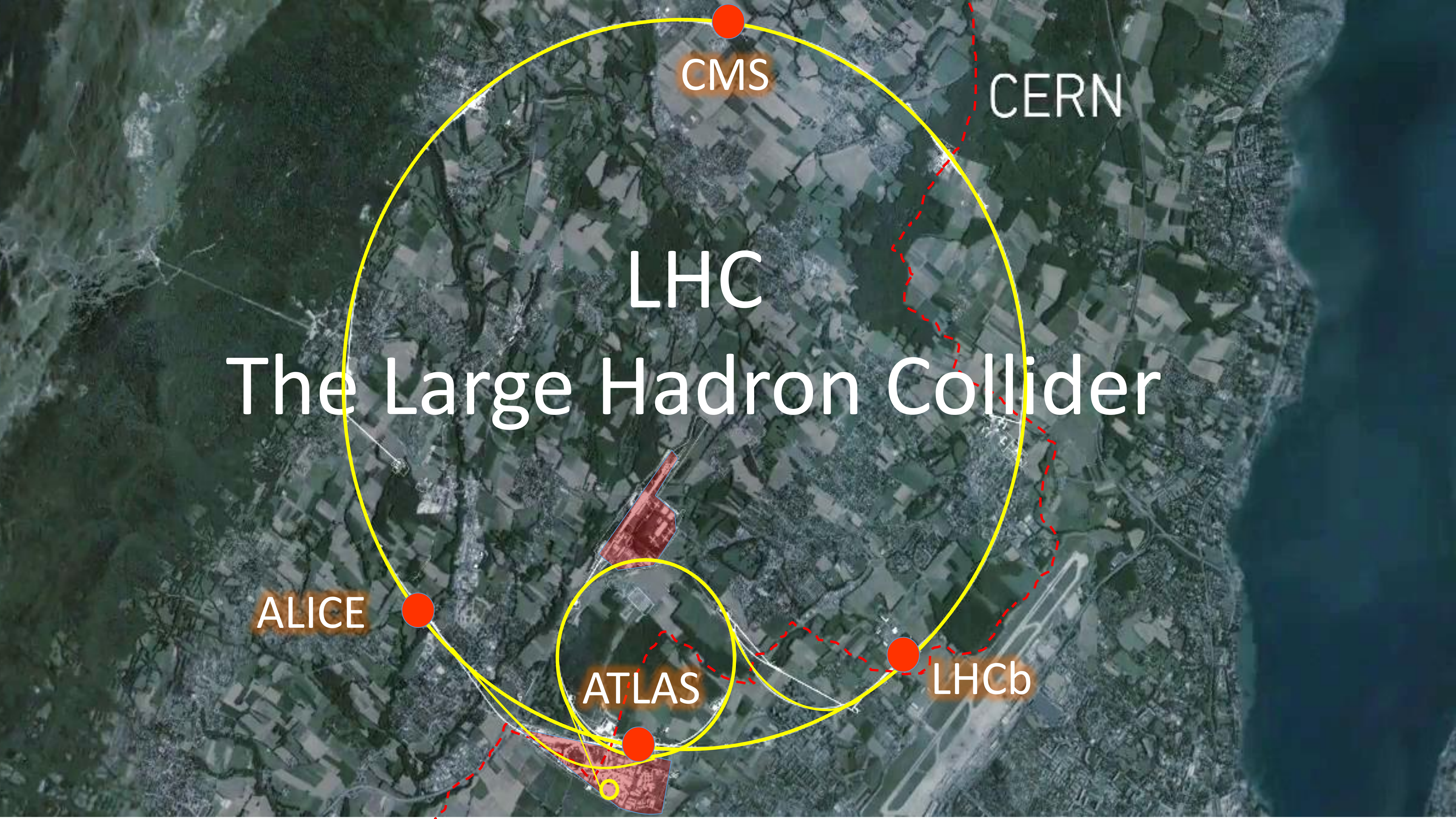
Head, Storage group

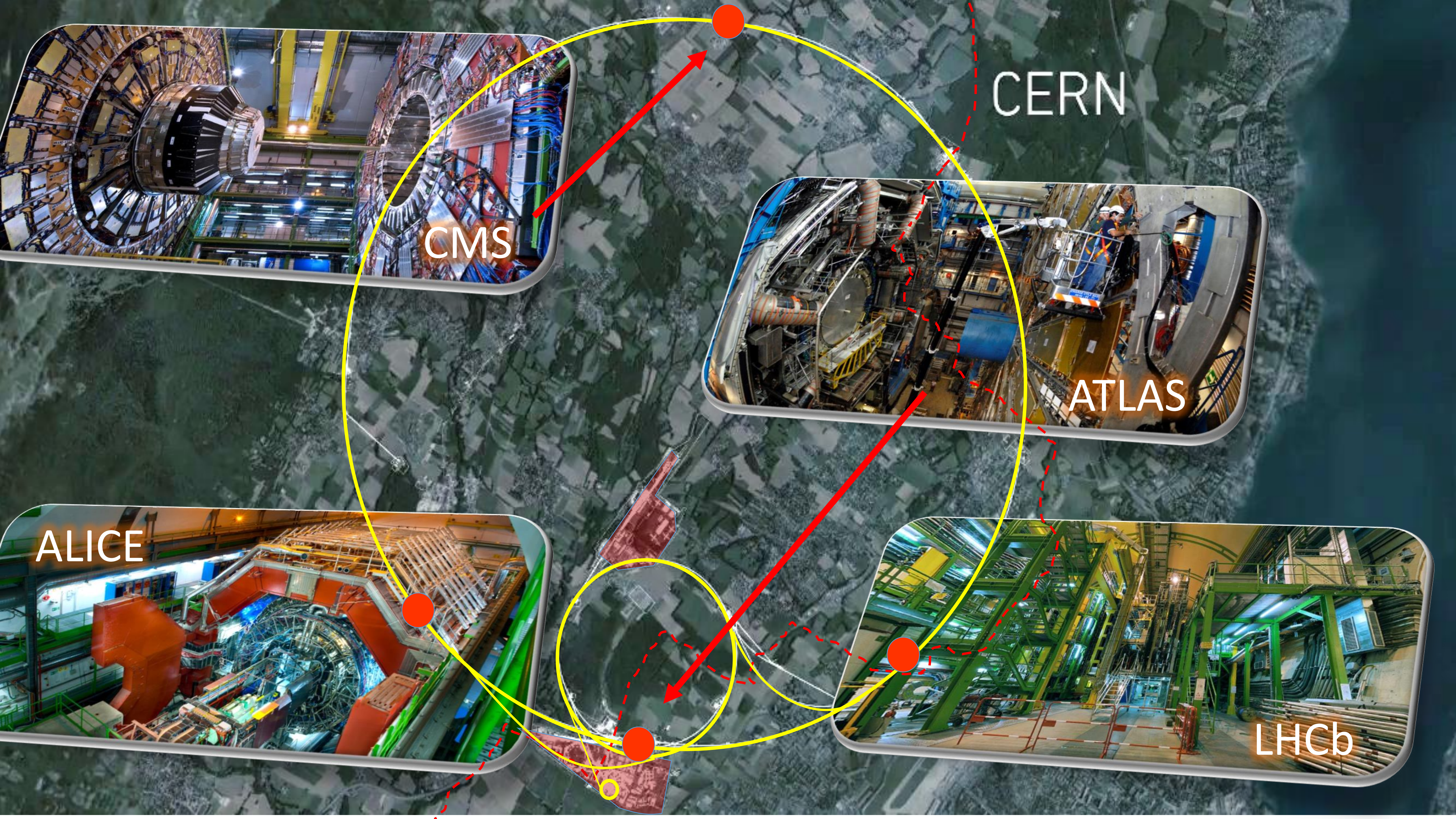
CERN, Geneva, Switzerland

Intégrité et partage de la Science : les données de la Recherche, 13 Décembre 2018, Université de Grenoble, France

Setting the scene …

- CERN has a 27 km machine
  - One of the most complex machine ever built

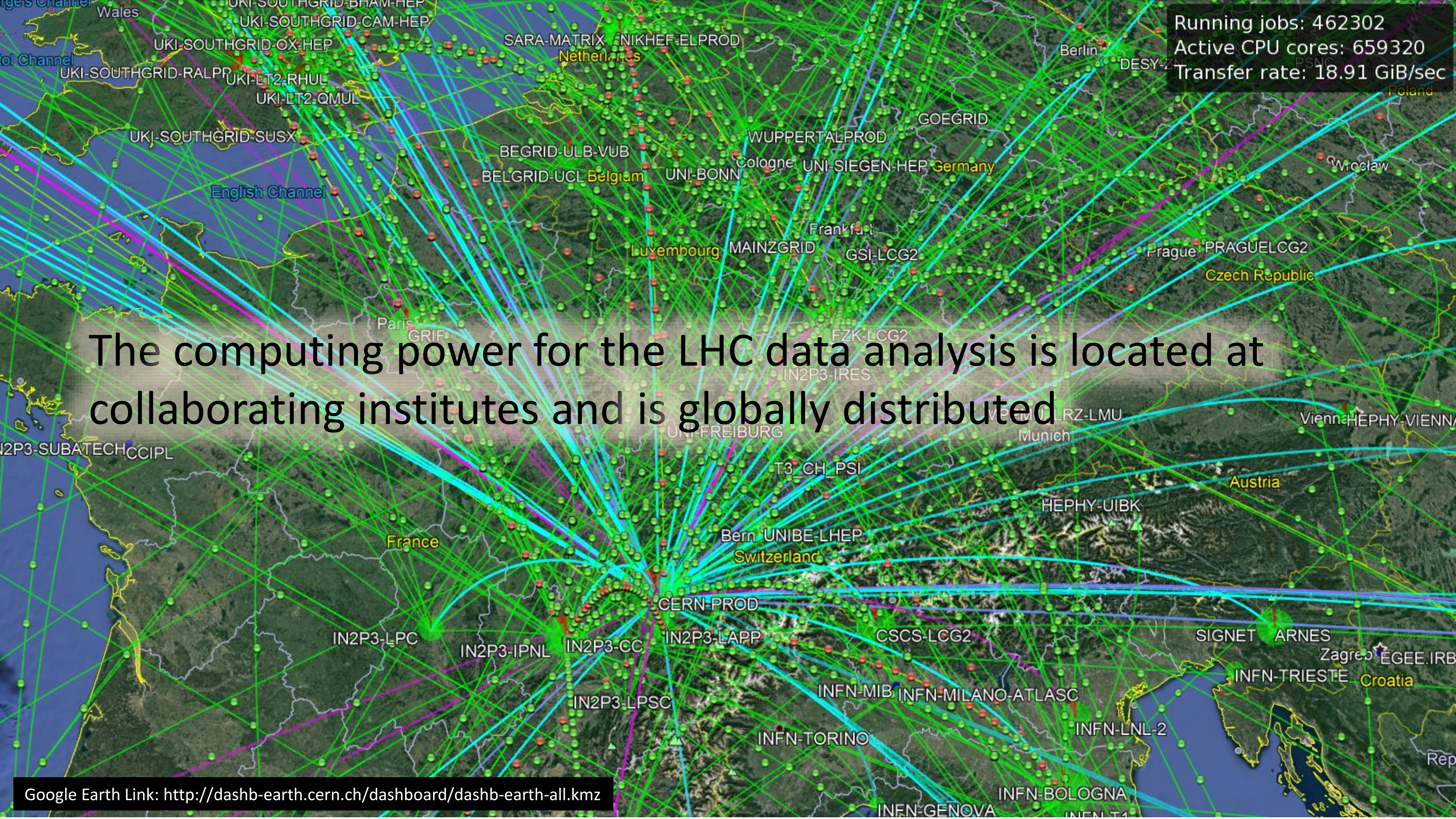

CMS
CERN
LHC
The Large Hadron Collider
ALICE
ATLAS
LHCb


CERN
CMS
ATLAS
ALICE
LHCb

Setting the scene …

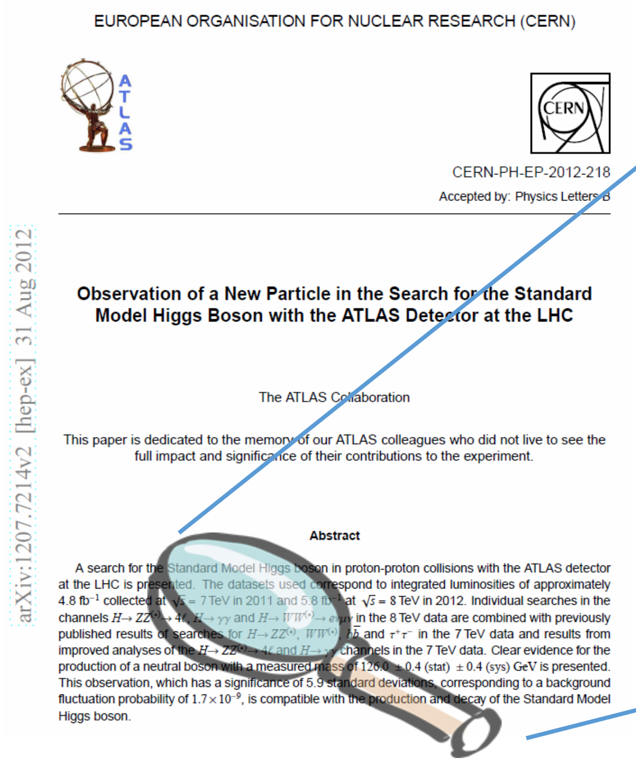
- CERN has a 27 km machine
  - One of the most complex machine ever built
- There are 4 large scientific experiments
  - These experiments produced 80 million GB in 2018 which had to be stored, need to be preserved and kept easily accessible
  - More than 200 data centres around the planet are required to analyse LHC data

The computing power for the LHC data analysis is located at collaborating institutes and is globally distributed



The (r)evolution of scientific research at CERN

- Why more than 600'000 computing cores 24h/day for years ?
  - Because we heavily use statistics, to prove that the theoretical model (the existence of the Higgs boson) is compatible with the experimental measurements… with a certain probability !



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP-2012-218
Accepted by: Physics Letters B

arXiv:1207.7214v2 [hep-ex] 31 Aug 2012

Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\mathrm{stat}) \pm 0.4\,(\mathrm{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\mathrm{stat}) \pm 0.4\,(\mathrm{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

P (Model | Data)

What is the probability of the theory, given the observed data ?

Unstructured data analysis uses Bayesian theory

- Everything is derived from the probability formula P() of two dependent events to both happen
  - P(A and B)
  - P(A) x P(B|A) = P(B) x P(A|B)

P(A|B): conditional probability
The probability of A knowing B has occurred

- When you suppose a model and you observe some data: P(Data | Model)
  - What is the probability of our observation given a certain model?
- and of course, concept of Likelihood : P(Model | Data)
  - What is the probability of our theory, given the observation ?
    - Will we have inflation ? What will be the interest rate next year ? Will it be sunny tomorrow ?
- Bayes' Theorem (simplified)
  - P(Data) x P (Model | Data) = P(Model) x P(Data | Model)

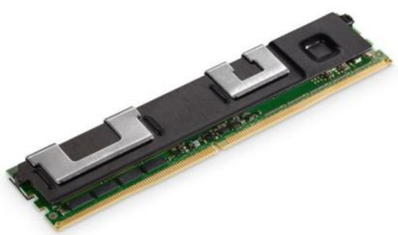
This you measure	This is your unknown. It tells you if the data are compatible with your model	This is your subjective prior belief	This you calculate (or maximise)

What has recently changed ?

- Statistics and Mathematics are sciences, they hardly change.
- But computers can do a lot more:
  - The explosion of data processing possibilities
    - CPU performance (10^3 increase) and number of CPUs available (10^3 increase)
  - New storage possibilities
    - From few GB to many PB (10^6 increase) – Big Data
  - The possibility of collect / transfer / store these data in a distributed environment
    - From few Mbit/s to Gbit/s (10^4 increase)

Has something changed ?

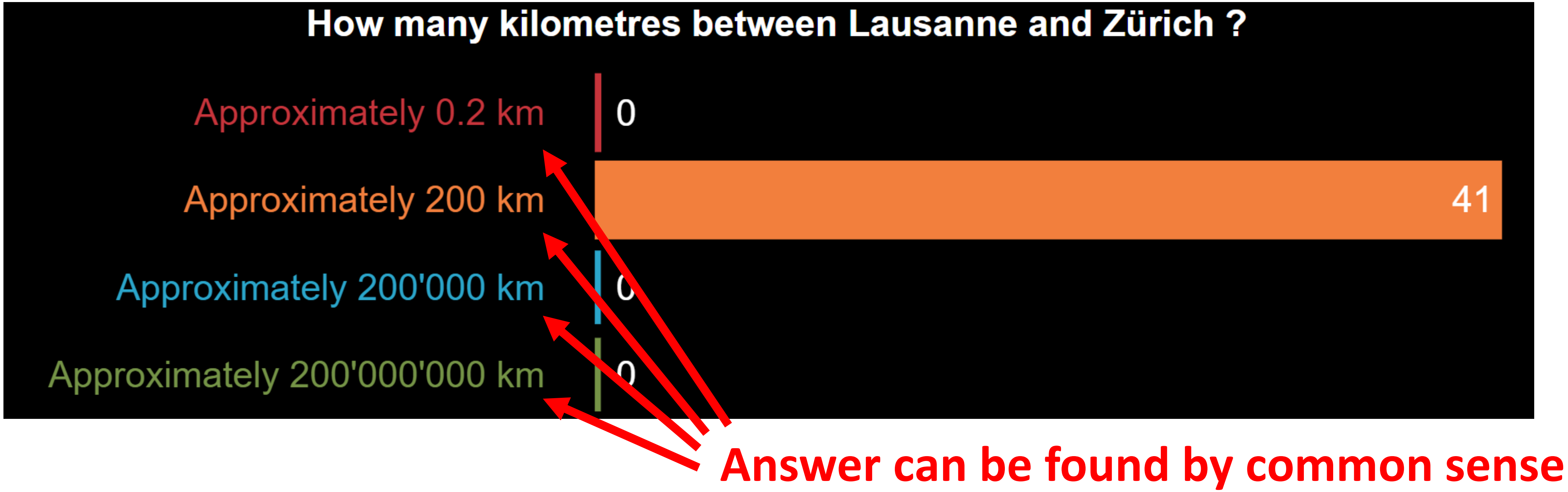
- Progress in computing technology has been exponential from its inception
  - (for fixed investment amount)
- In many areas, return of investment (ROI) in computing has been the highest compared to other investment areas
  - This attracts more investments
  - Consequence: more than exponential growth !
- And it is not over yet !
  - we are only at the very beginning …



Just announced: 512 GB non volatile capcacity in a single DIMM module. Faster than RAM

The general population is not aware of these changes

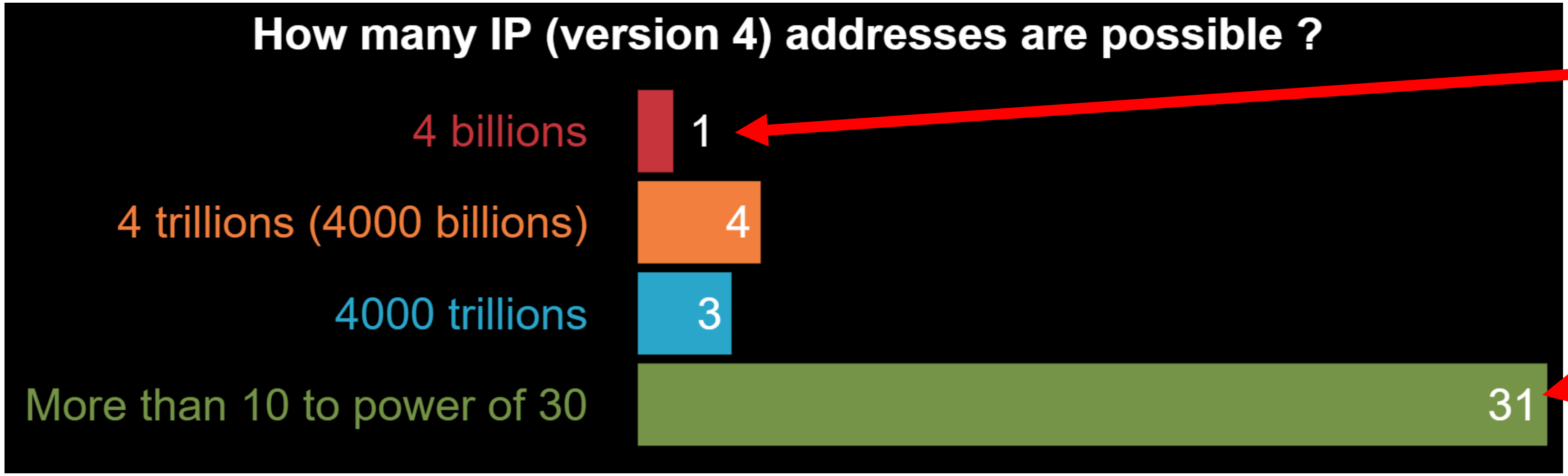
- An example of a survey made in February 2018 with Master of Finance students from University of Lausanne:



but … what about questions on computing ?

most popular answer
has 3 orders of magnitude error

less than 1/5
Correct answers !



only 1
Correct answers !

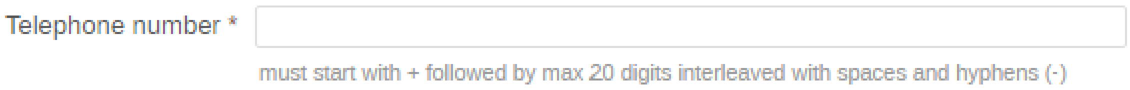
20 orders of
magnitude error

More computing power … for what ?

- This has enabled the use of methods and techniques that few years ago where "computationally impossible"
  - Current research at CERN would not have been possible without these computing evolutions
  - From few targeted calculations to a systematic wide-scale approach
- In practice:
  - The increasing power of computing allows statistic correlation on (un)structured, unverified or unreliable data

The (traditional) structured approach …

- Data is collected in databases:



- The databases is supposed to contain the "truth"
  - Integrity, constraints and input validation exists at all levels
- Information is processed by deterministic software
  - predictions are considered true with a probability of 1

The unstructured approach …

- Data is collected unstructured, without constraints:

… please call me at 0754113800 so we can discuss …

- Is it a Phone number ? French or Swiss mobile ? Is it a German fixed line from Friedrichshafen ?
- Can we correlate with the knowledge that the internet connection was from Lausanne, Switzerland ?
- The analytic software can be designed to guess the most probable outcome
  - This may not be as useful as you think
- Or used to identify "less probable" but alternative "truths"
  - Unexploited niche markets, uncovered needs, services missing, alternative scenarios …

Multiple techniques and technologies available

- Several techniques (and buzzwords)
  - Data mining
  - Machine learning
  - Artificial intelligence
- Several computing technologies
  - Databases (Sql / NoSql)
  - Structured / Unstructured data
  - Map - Reduce architectures
  - Hadoop + its ecosystem
    - Hive, HBase, Spark, Impala, Flume, …

There is a need for both approaches
When possible, structured approaches are preferable

real time statistics: http://monit-grafana-open.cern.ch/d/000000884/it-overview

Dealing with unstructured information

- Typically text based (documents, email, messages)
  - Contains dates, numbers, and assertions.
  - Also images, audio files, and videos - which can be analysed
- Every information has a probability of being incorrect
  - Contains irregularities, ambiguities, and errors.
- Possible to make accurate predictions despite poor data quality
  - provided you have a lot of data (the "Big Data" buzzword)

Practical prediction example …

- You have a coin, you flip it 10 times.
  - You obtain 6 Heads and 4 Tails: H-H-T-H-H-T-T-H-T-H

- What is the probability of having Head on the next coin flip ?

a) less than 0.5 — Tail is slightly more likely
b) 0.5 — ? Head and tail have equal probability
c) 0.6 — Head is significantly more likely
d) between 0.5 and 0.6 — Head is slightly more likely

Multiple approaches

- You think that after many throws you will have equal number of heads and tails, so … you believe the coin will more likely flip on tail to compensate the many heads that happened in previous throws. (Wrong reasoning)
  - As you believe that heads are less probable than tails, your probability is below 0.5
- You use the prior probability (Theoretic approach)
  - You have a model of the coin that has two faces. You suppose the coin is well balanced and each face has equal probability to be drawn. You ignore the fact that you have observed more heads than tails.
    - The probability to have Head on the next coin flip is 0.5
- You rely only on your limited experience (Frequentist probability inference)
  - You have no model and calculate the probability of the next draw using you observation which showed 6 heads of 10 draws.
    - Despite you have a limited sample, your best guess for the probability to have Head on the next coin flip is 6/10 = 0.6
- You combine both (Bayesian approach)
  - You refine the prior probability with the event you have observed
    - So the next draw will be Head with a 0.52 probability
  - … or … you reject / accept (with a given probability) a model based on your measured events: The 6 head events for 10 draws are compatible with the 0.5 probability of the model
    - So the next draw will be Head with a 0.5 probability

https://en.wikipedia.org/wiki/Bayesian_inference https://en.wikipedia.org/wiki/Frequentist_inference

Bayesian example

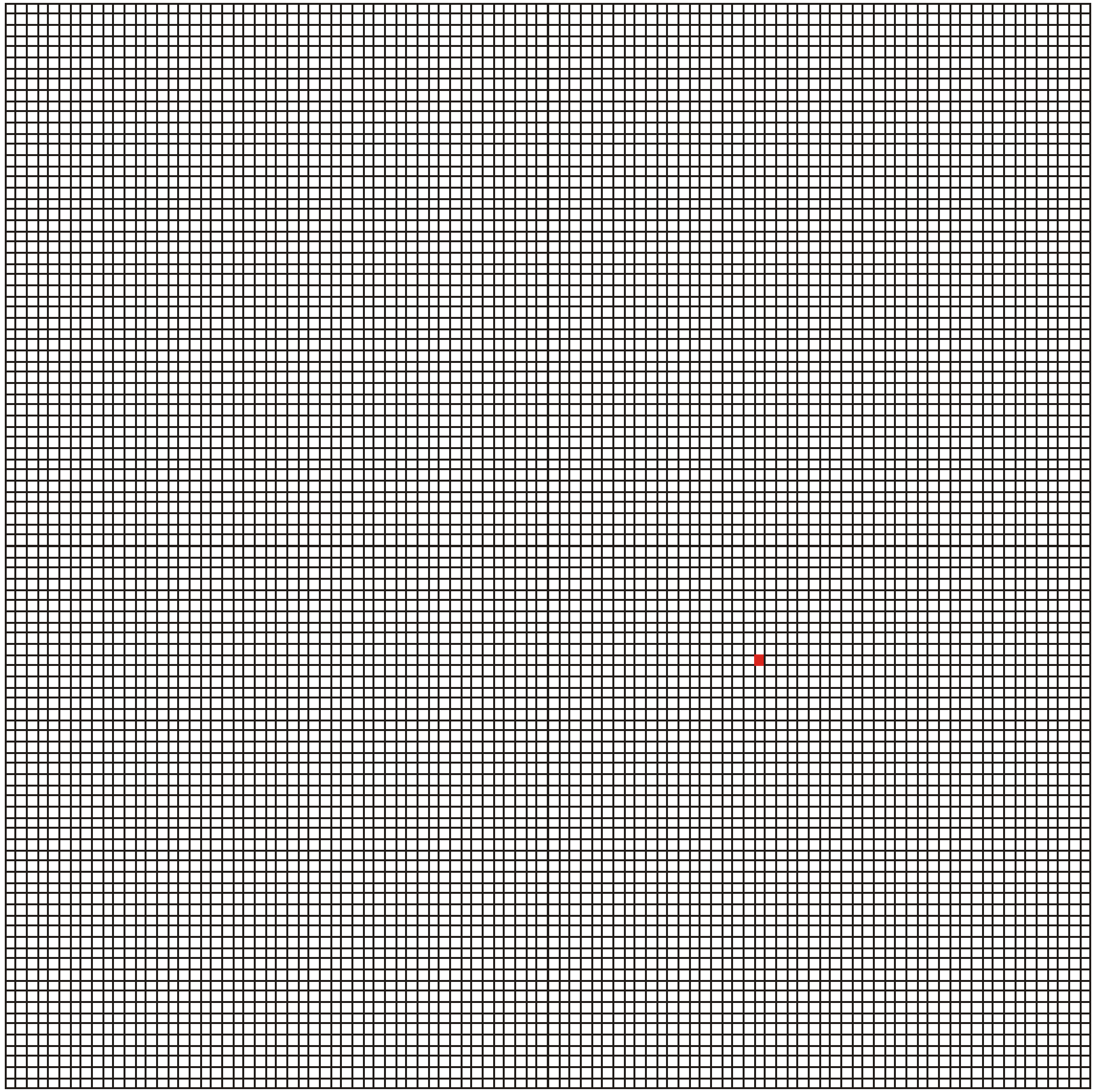
- Mongolian swamp fever (MSF) is a deadly disease that hits 1 person every 10'000
- Luckily there is a reliable test:
  - If you have MSF, the test will report positive at 99.99% probability
  - if you do have MSF, the test will report positive at 1 % probability
- You take the test and it reports positive.
- What is the probability that you have MSF ?

P(MSF | TestPositive) = P(TestPositive | MSF) x P(MSF) / P(TestPositive) = 0.9999 x 0.0001 / 0.009999 = 0.0099 (less than 1%)

P(TestPositive) = P(TestPositive | MSF) x P(MSF) + P(TestPositive | NoMSF) x P(NoMSF) =
= 0.9999 x 0.0001 + 0.01 x 0.9999 = 0.009999

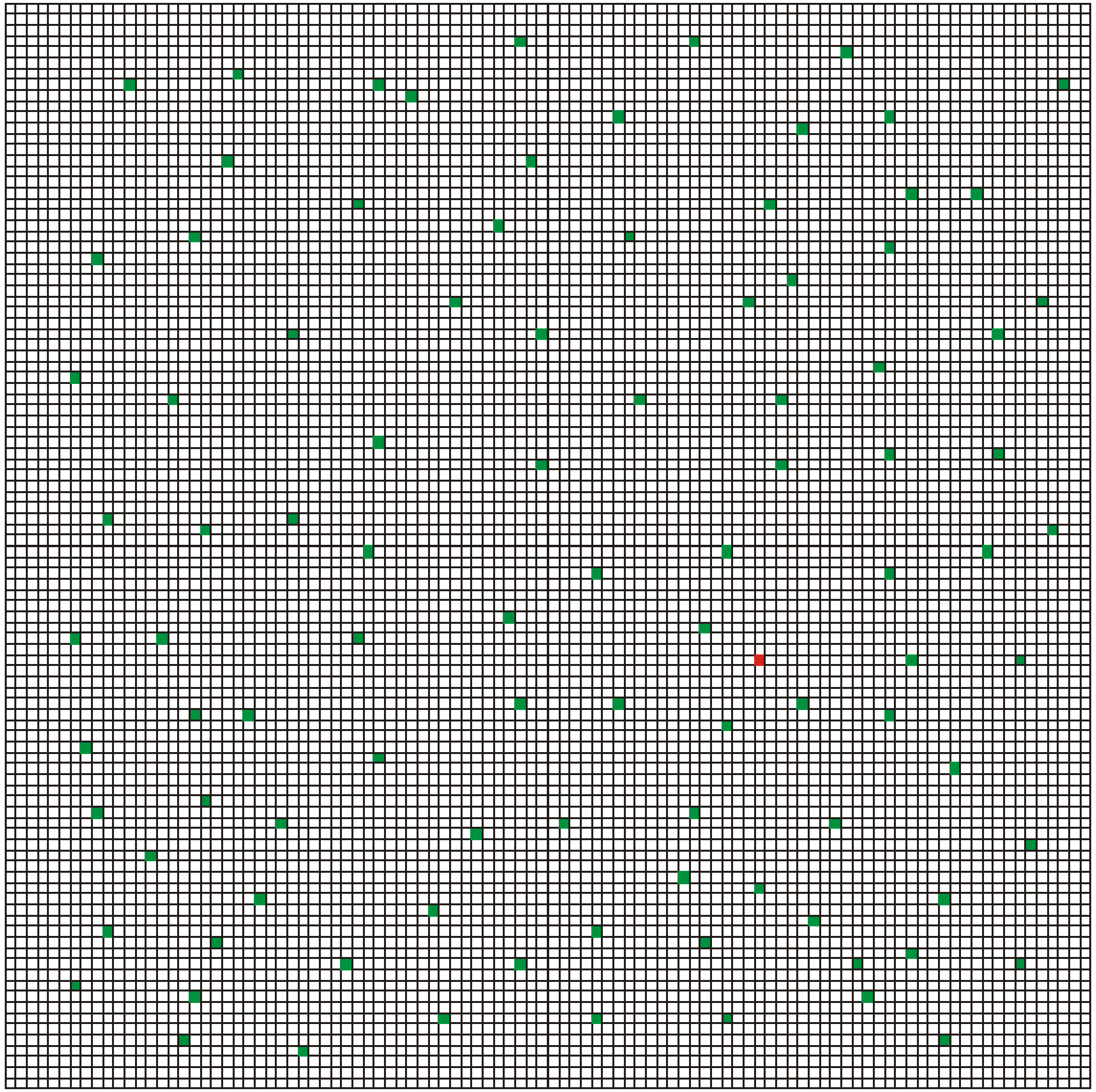
Total population (10'000)

- Rare disease
  - 1 person sick



Total population (10'000)

- Rare disease
  - 1 person sick
- 1 in 100 false positive
  - 101 positive test for 1 person sick
- 1 % probability to be sick if the test is positive



What if you do the test again ?

- You do the test again and … bad luck. You are again positive to MSF.
- Does this change the probability that you have MSF ?
- What is the probability that you have MSF ?

P(MSF | TestPositive) = P(TestPositive | MSF) x **P(MSF)** / P(TestPositive) = 0.9999 x **0.0099** / 0.01989 = 0.497 (~ 50%)

P(TestPositive) = P(TestPositive | MSF) x **P(MSF)** + P(TestPositive | NoMSF) x P(NoMSF) =
= 0.9999 x **0.0099** + 0.01 x 0.9999 = 0.01989

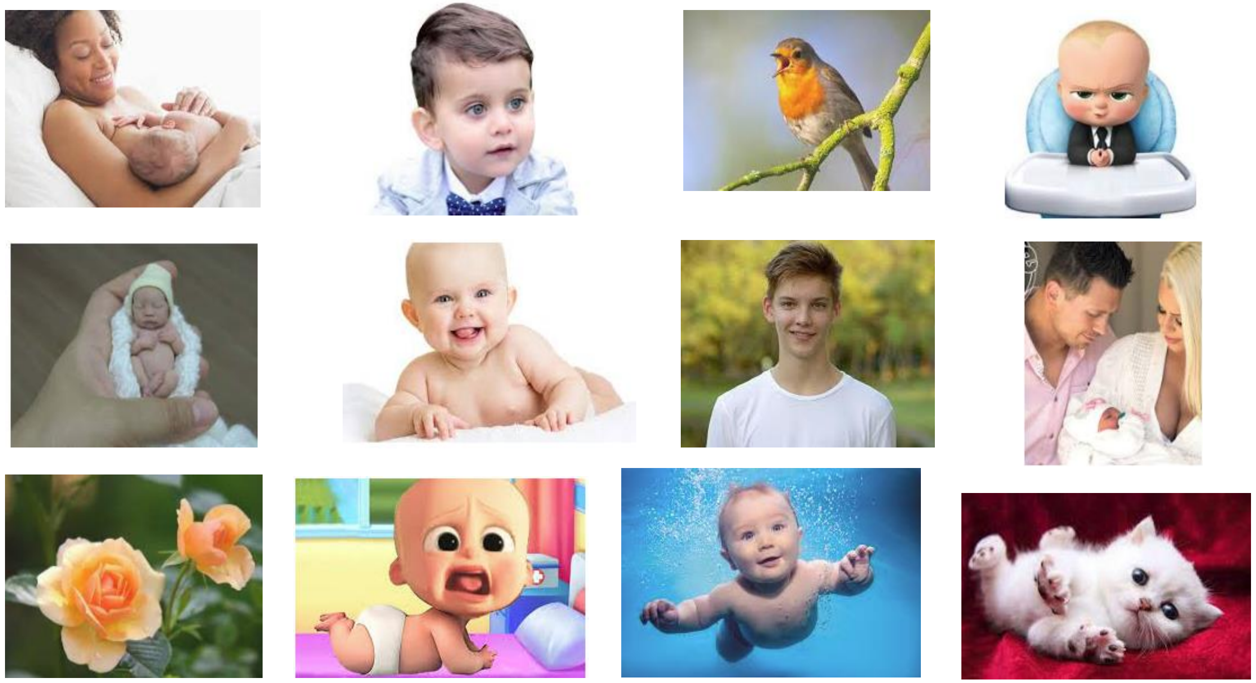
101 people where positive to the test and run the test again. 1 person is sick

amont the 100 non sick persons, you have 1% failure rate. you have two positive tests. One true, one false positive (50%)

- The more observations you make, the more you reduce the uncertainty.
  - What about doing a third test ?

Another Example - image recognition

- How many of these photos represent a baby ?
  - Can you give a definitive answer ?
  - Can you estimate a (subjective) probability for each image ?
  - Can you transform millions of subjective answers into an objective one ?
  - Can you develop an algorithm that can analyse a new photo from the millions of subjective answers analysed from a larger image database ?



Another example: accurate predictions from poor data quality

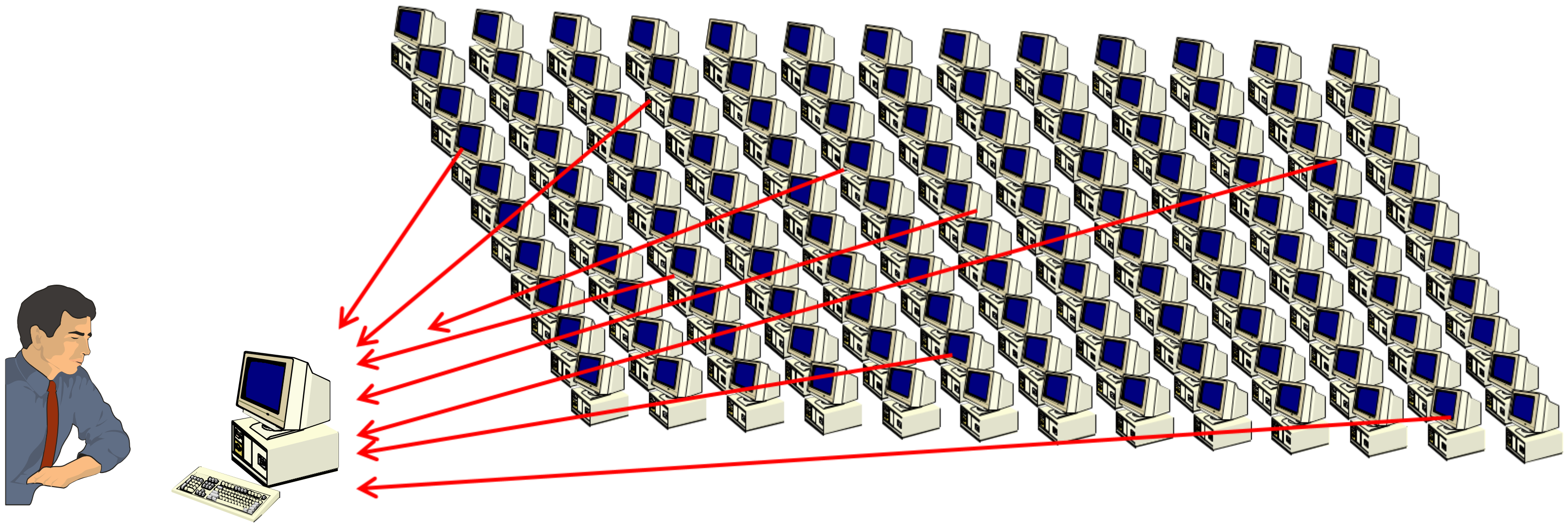
- If you are using a mobile phone or a portable computer ...
  - Some companies knows your location - within one meter and millisecond precision
  - The same companies may be able to read your email, access your contacts
  - The same companies may know all web pages you have been reading, at what time, from which location. Probably for the last 10 years or more.
- Good news: under GDPR, you have ownership of your data
- But … this information can be correlated with other data
  - To know who is in this room, to guess what we are talking about …
  - To know who we meet, political opinions, industrial plans, …
  - The information inferred using analytic techniques doesn't belongs to you anymore

try it yourself: http://google.com/takeout

Another change is happening

- Some "notary" roles of transactions can be guaranteed by algorithms
  - An application distributed across thousands of computers can ensure, verify and guarantee by itself its own integrity
    - No need of third party intermediary, no commissions
    - No possibility to manipulate the book of writings
    - No possibility to change the "rule of the game" (the contracts) in an ongoing process

A first example: Bittorrent file download



Data reassembled directly on the client from untrusted computers.

The software does all necessary cryptographic verifications which guarantee data integrity.

A second example - blockchains

- A distributed database validated by large number of computers
  - Everyone can read, everyone can write new data (append)
  - Everyone can validate, validation is rewarded
    - There is a large number of computers validating the truth.
    - Cheating is "probabilistically impossible"

Why this technology has potential ?

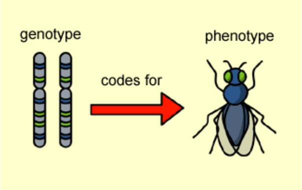
- Every activity requiring a certification authority can profit from a distributed database (the blockchain)
  - Can store information about contracts, or arbitrary complex transaction
- New business models appear
  - The distributed database ensures the notary role of contract certification
  - Strategic to focus on issues related to contracts enforcements and resolving disputes

The Bitcoin blockchain

- The blockchain database contains the list of Bitcoin transactions that describe changes of ownership
  - Does not require permissions, it is resilient and it has a cost
  - it seems immune to security attacks, censorship, race conditions, tampering of the past
  - produces a single version of the 'truth'
- Validating the blockchain is rewarded using the same Bitcoins, therefore subjective. So, why do Bitcoins have some value ?
  - It is portable, it is scarce, new coin creation is rate-limited, predictable and not infinite.
  - It can be used as a currency
- If the value is >0, and it is subjective - it can be traded !

Where will we see most spectacular changes ?

- High Energy Physics has been profiting, because the community is historically organized, but all other sciences can expect similar benefits:
  - Biology, Medicine, Climate and Weather forecast, …



- Finance and market analytics
  - Insurances, loans, derivatives, forex, … anywhere there is a contract or a risk
- Marketing, targeted advertisements, lobbying, identifying markets
  - from data collected worldwide

Conclusion

- Everything you learned in hard science during traditional studies is still valid:
  - Mathematics, Statistics, Physics laws still applies. Nothing changes.
- But … computers and networks can do more today than a few years ago
  - Statistics and analytics solutions that where computationally unfeasible in the past become possible today.
- Cannot fight progress
  - Many of these approaches can bring significant improvements to everyone's life
    - plenty of new business opportunities, ethical consequences must be understood and handled
  - Education is of the utmost importance

CERN Data management for the LHC experiments

Alberto Pace, alberto.pace@cern.ch

Head, storage group

CERN, IT department, Geneva, Switzerland

Roles Storage Services

- Three main roles
  - Storage (store the data)
  - Distribution (ensure that data is accessible)
  - Preservation (ensure that data is not lost)

Size in PB + performance

Availability

Reliability

"Why" data management ?

- Data Management solves the following problems
  - Data reliability
  - Access control
  - Data distribution
  - Data archives, history, long term preservation
  - In general:
    - Empower the implementation of a workflow for data processing

CERN Computing Infrastructure

CPUs

Network

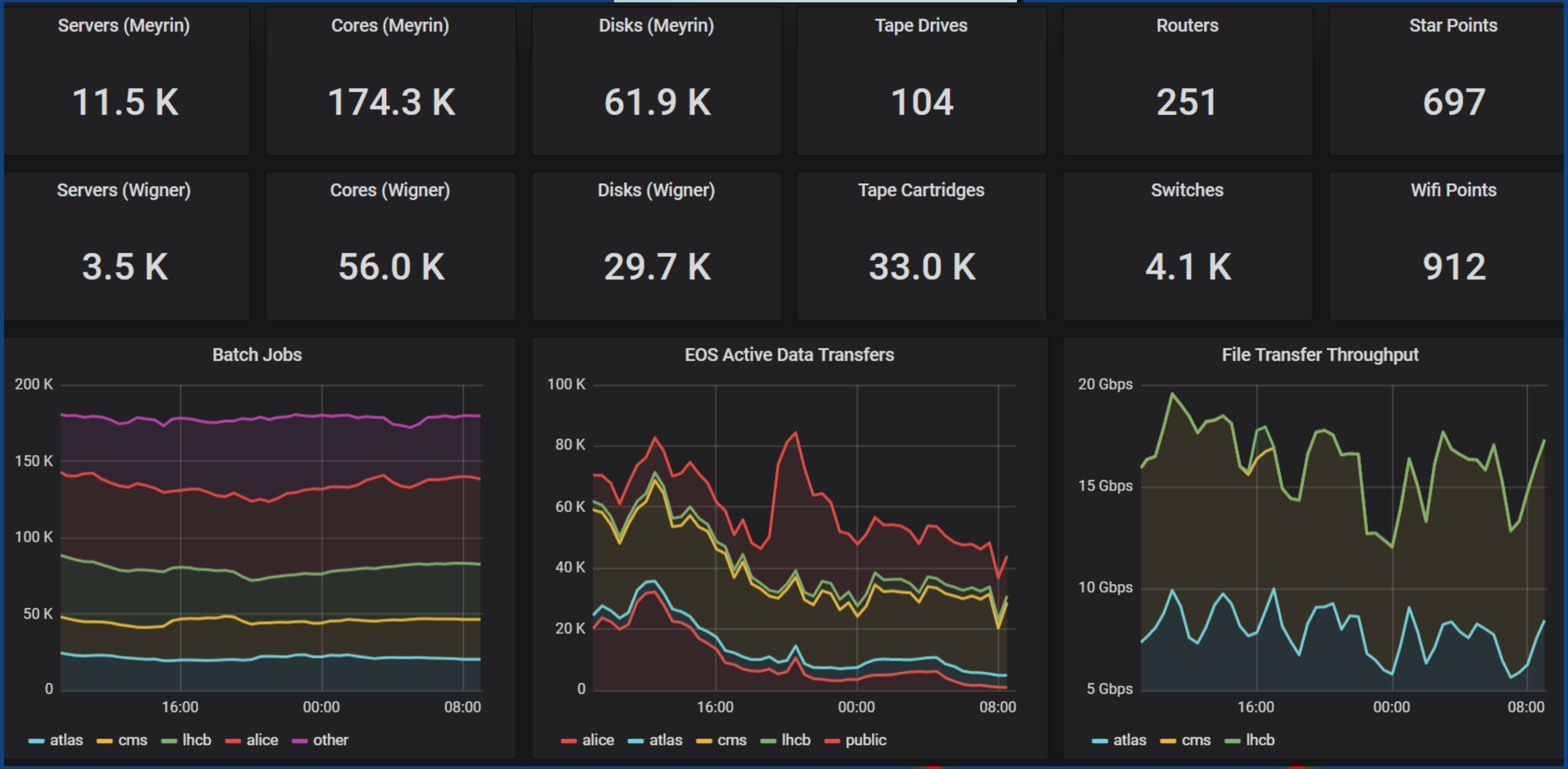
Databases

Storage

Infrastructure

CERN Computing Infrastructure

Tue Nov 27th, 2018 at 11:00

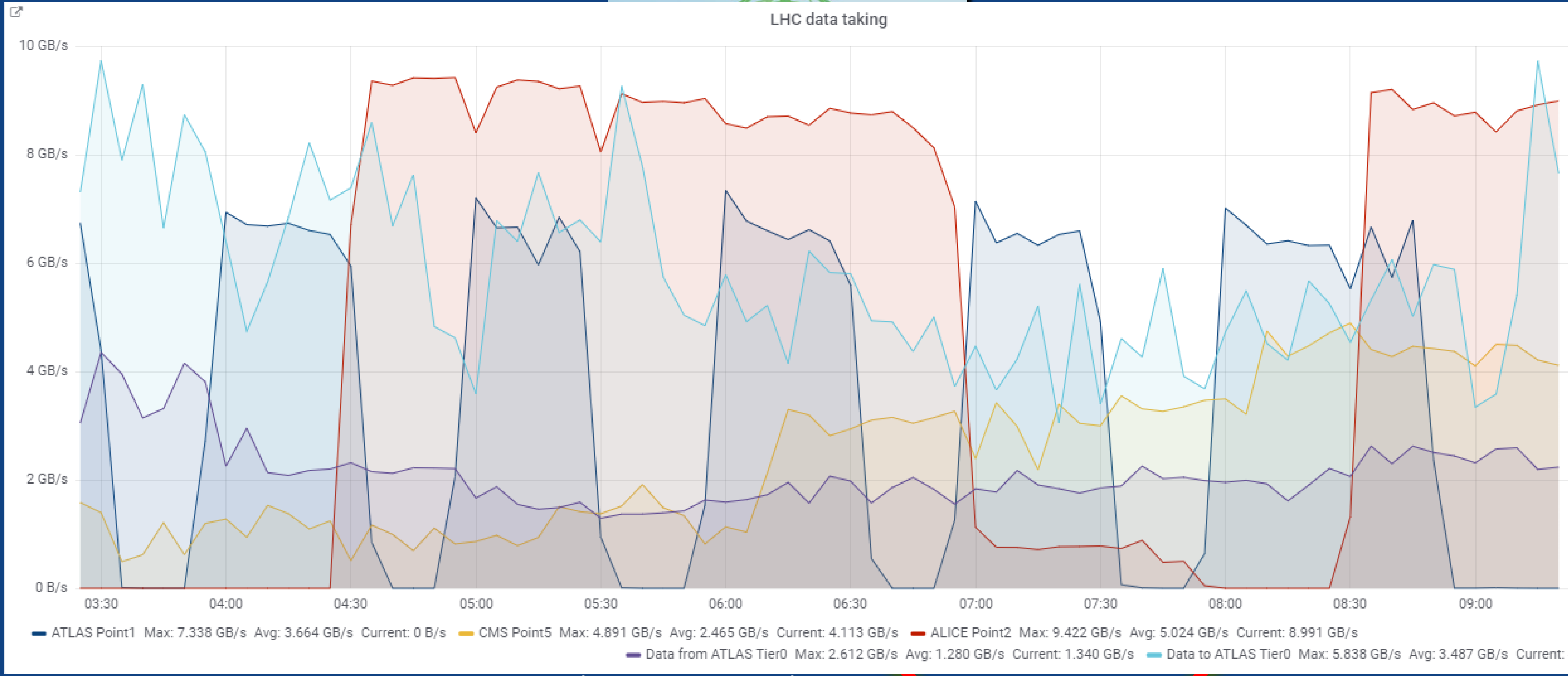


CPUs Network Databases Storage Infrastructure

http://monit-grafana-open.cern.ch/d/000000884/it-overview?orgId=16

CERN Computing Infrastructure

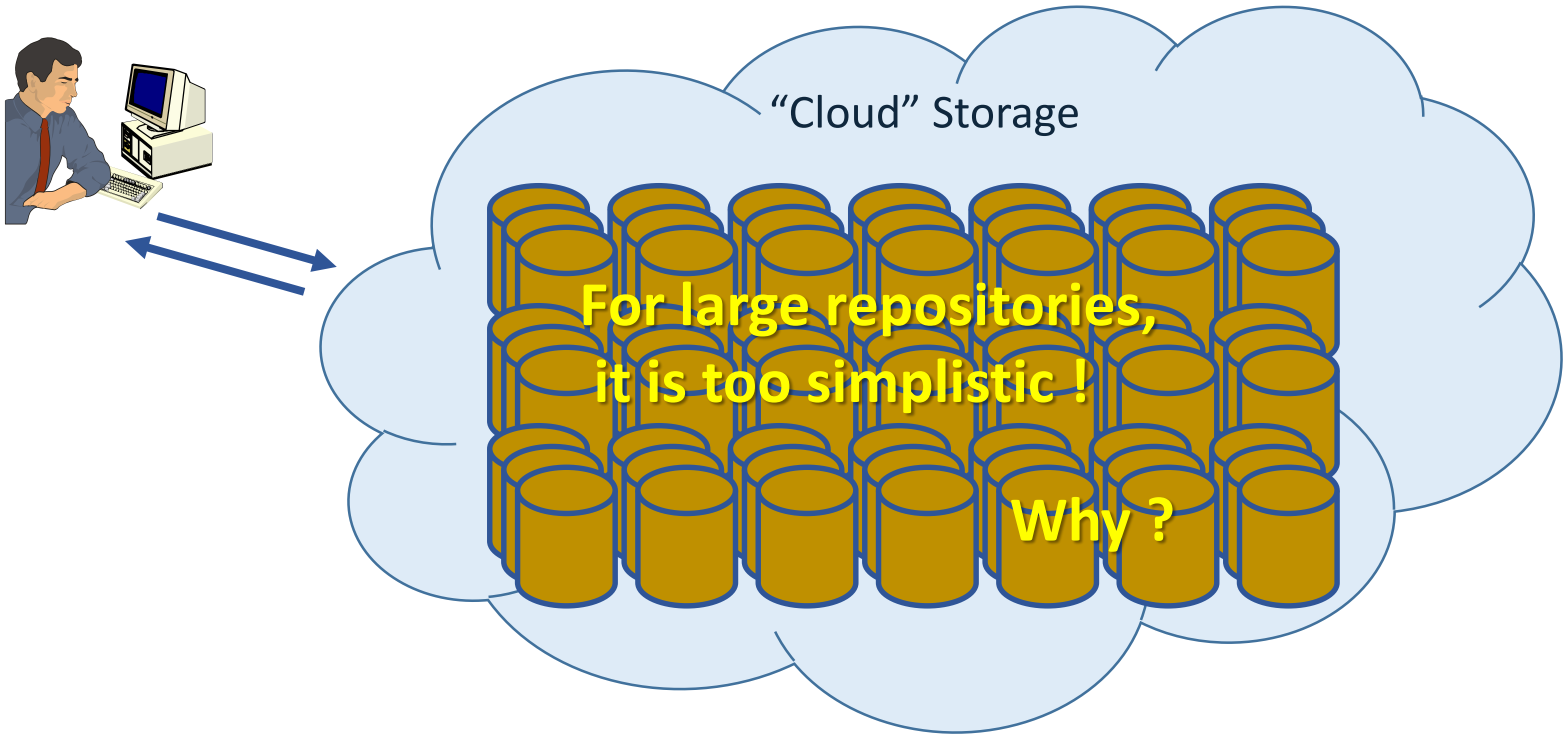
Tue Nov 27th, 2018 at 11:00



CPUs Network Databases Storage Infrastructure

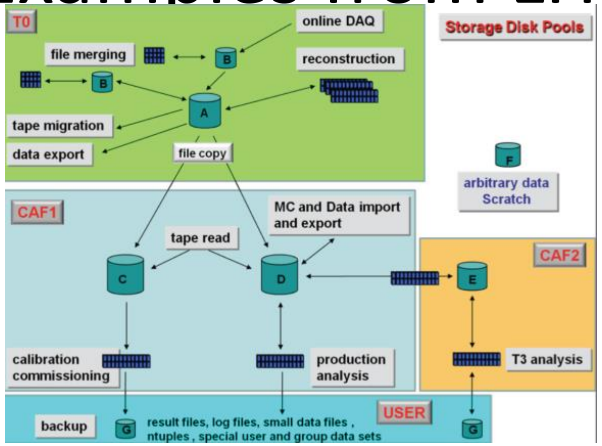
Can we make it simple ?

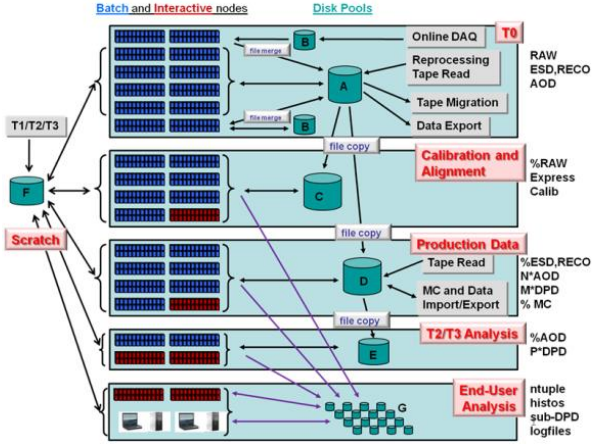
- A simple storage model: all data into the same container
  - Uniform, simple, easy to manage, no need to move data
  - Can provide sufficient level of performance and reliability

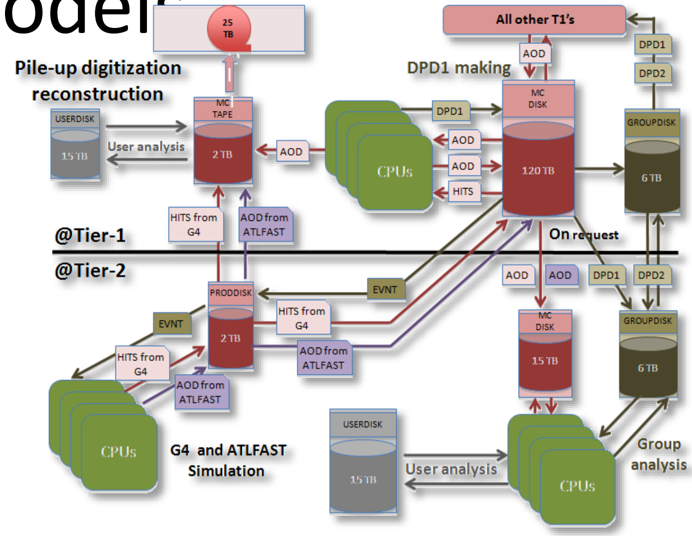


So, … what is data management ?

- Examples from LHC experiment data models







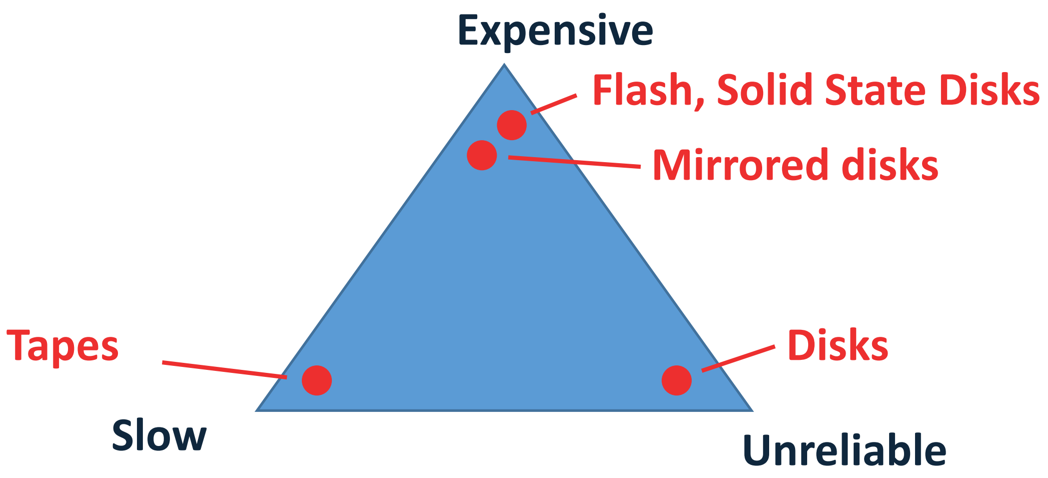
- Two building blocks to empower data processing
  - Data pools with different quality of services
  - Tools for data transfer between pools

Why multiple pools and quality ?

- Derived data used for analysis and accessed by thousands of nodes
  - Need high performance, Low cost, minimal reliability (derived data can be recalculated)
- Raw data that need to be analyzed
  - Need high performance, High reliability, can be expensive (small sizes)
- Raw data that has been analyzed and archived
  - Must be low cost (huge volumes), High reliability (must be preserved), performance not necessary

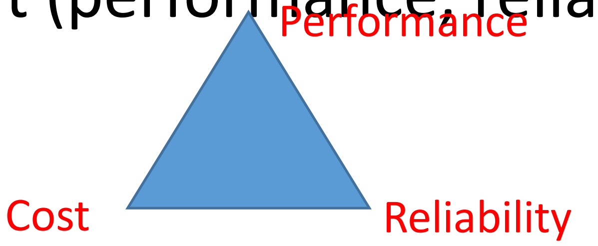
Data pools

- Different quality of services
  - Three parameters: (Performance, Reliability, Cost)
  - You can have two but not three

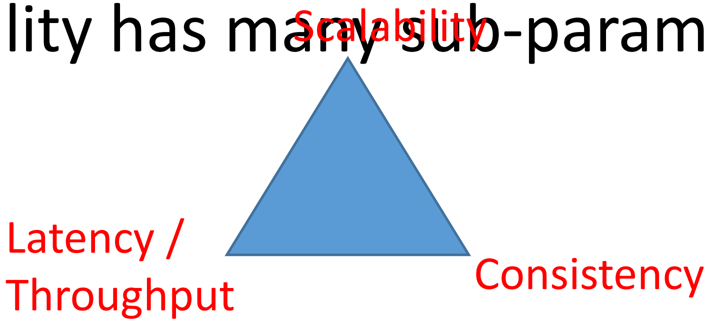


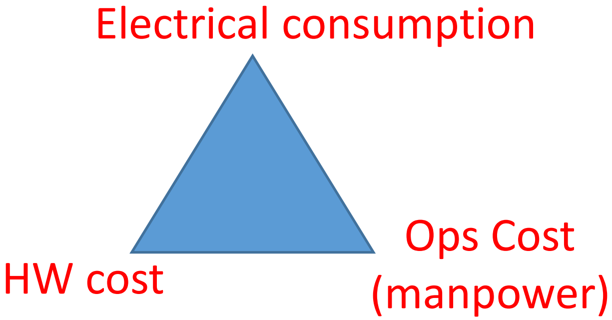
But the balance is not as simple

- Many ways to split (performance, reliability, cost)



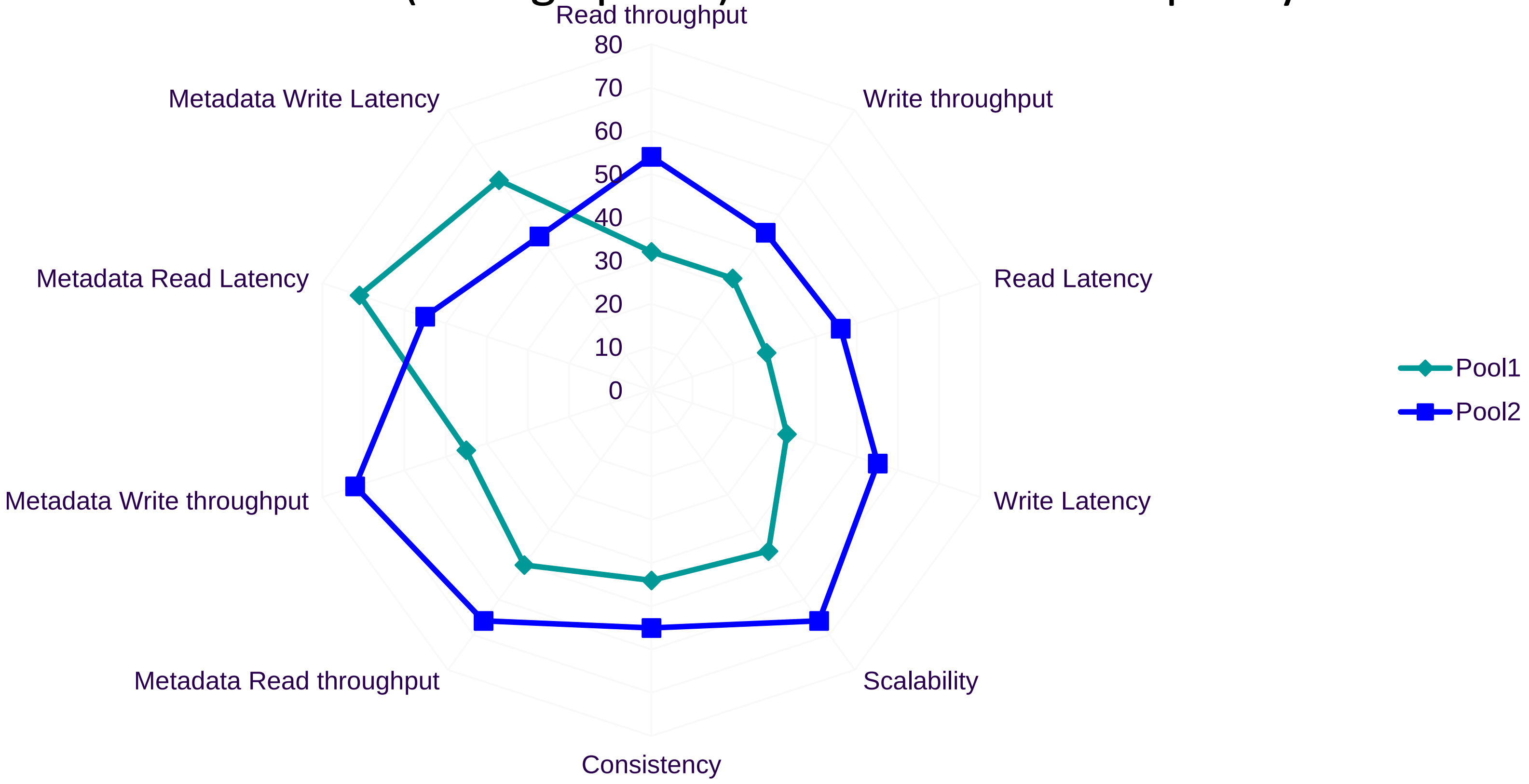
- Performance has many sub-parameters
- Cost has many sub-parameters
- Reliability has many sub-parameters





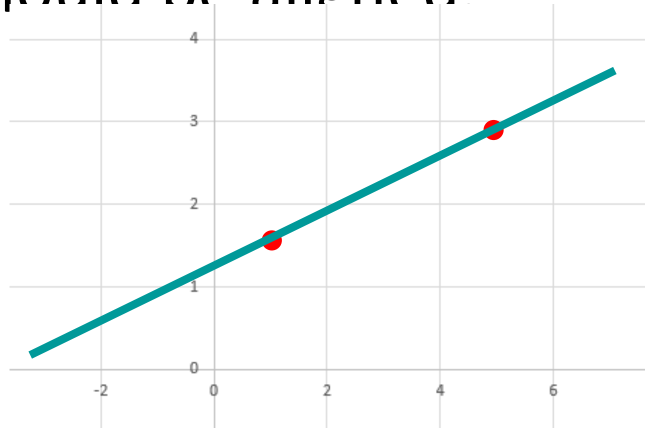
And reality is complicated

- Key requirements: Simple, Scalable, Consistent, Reliable, Available, Manageable, Flexible, Performing, Cheap, Secure.
- Aiming for “à la carte” services (storage pools) with on-demand “quality of service”

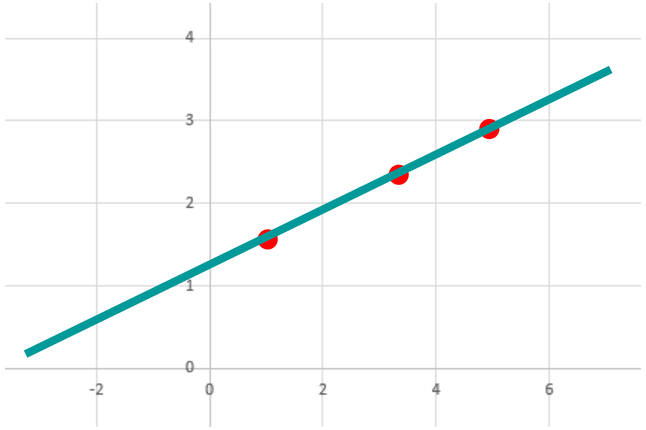


Understanding error correction

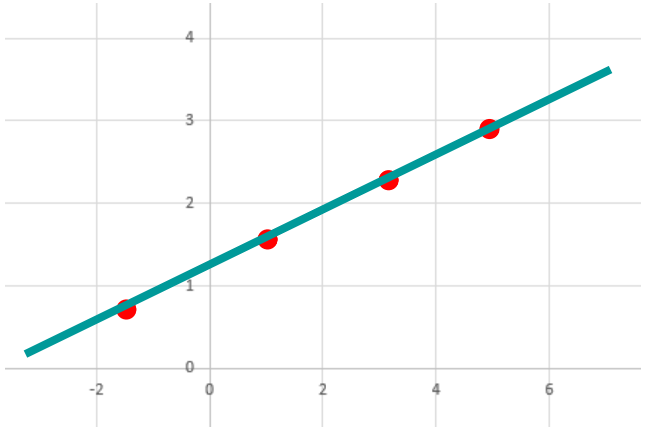
- A line is defined by 2 numbers: a, b
  - (a, b) is the information
  - $y = ax + b$
- Instead of transmitting a and b, transmit some points on the line at known abscissa. 2 points define a line. If I transmit more points, these should be aligned.



2 points



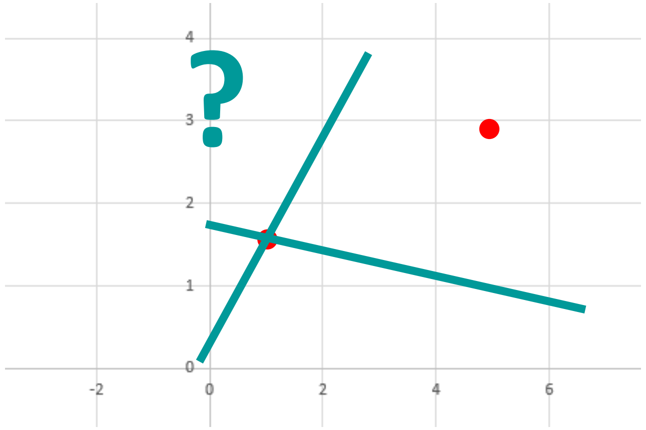
3 points



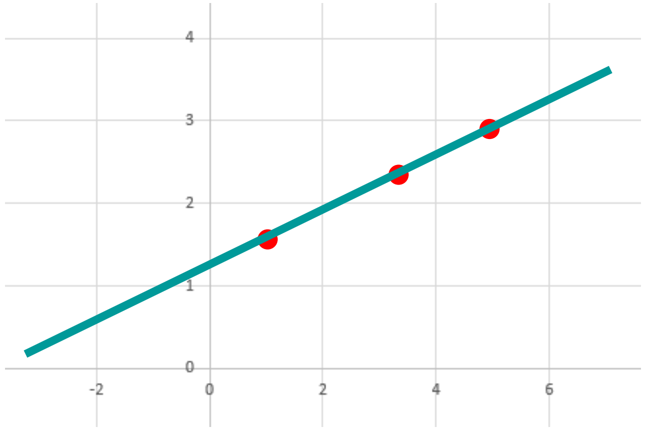
4 points

If we lose some information …

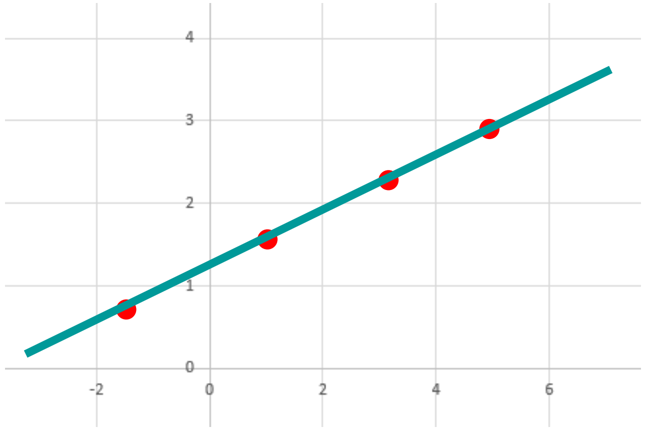
- If we transmit more than 2 points, we can lose any point, provided the tot`al number of point left is >= 2



1 point instead of 2
information lost



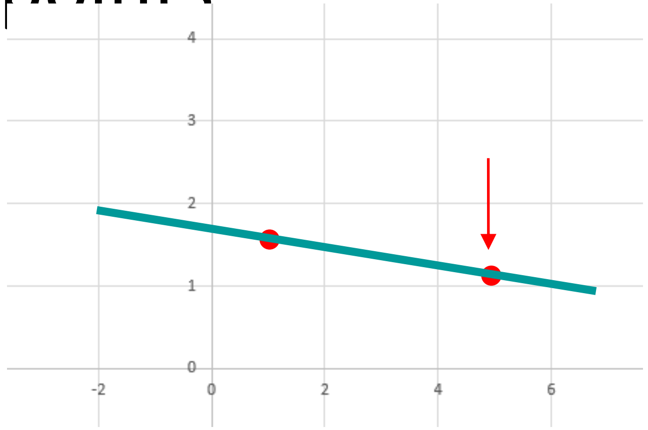
2 points instead of 3



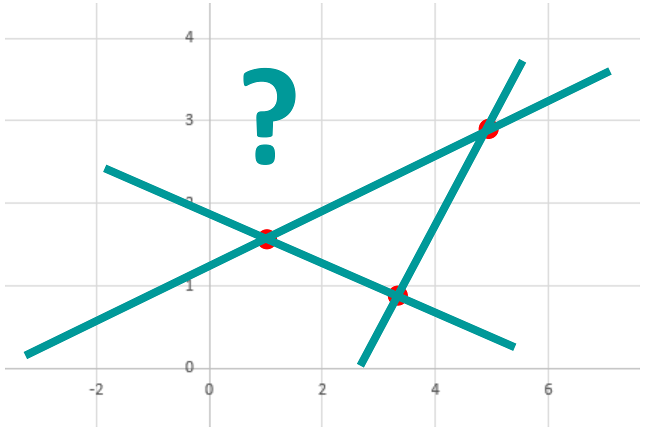
2 or 3 points instead of 4

If we have an error …

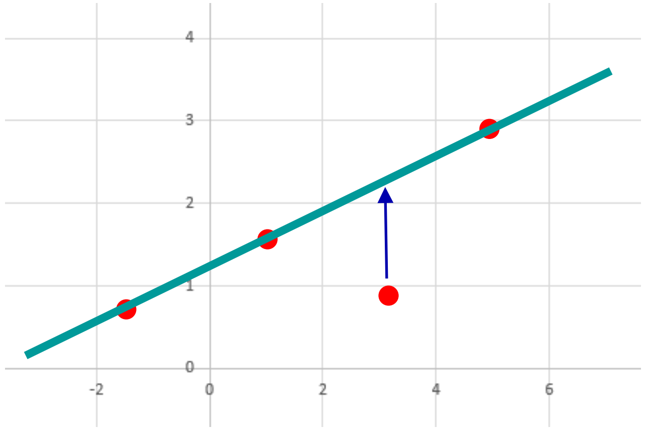
- If there is an error, I can detect it if I have transmitted more than 2 points, and correct it if have transmitted more than 3 points



Information lost
(and you do not notice)



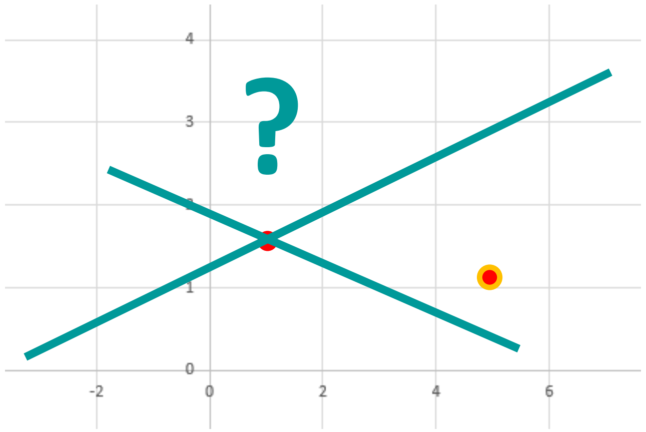
Error detection
Information is lost
(and you notice)



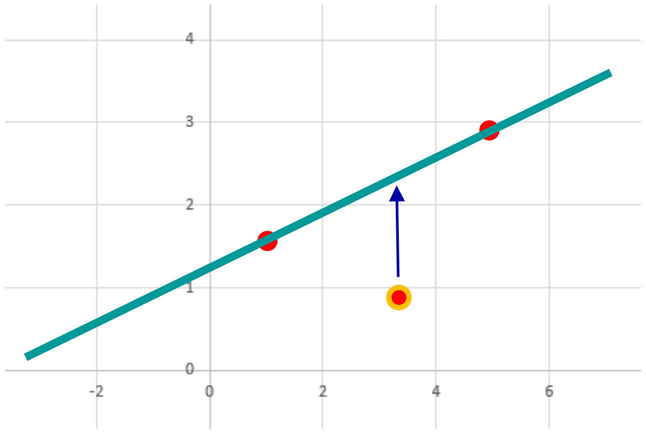
Error correction
Information is recovered

If you have checksumming on data …

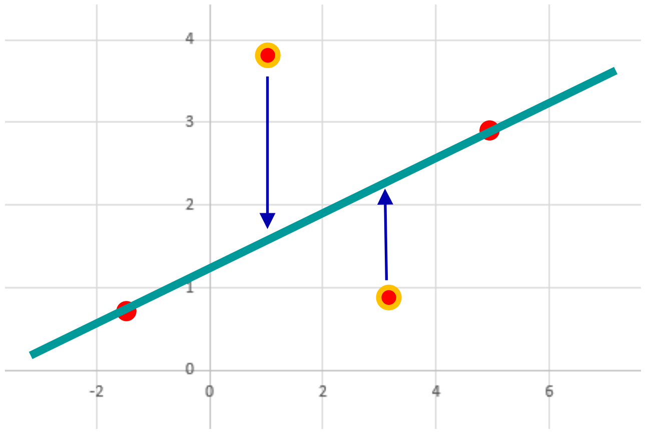
- You can detect errors by verifying the consistency of the data with the respective checksums. So you can detect errors independently.
- … and use all redundancy for error correction



Information lost
(and you notice)



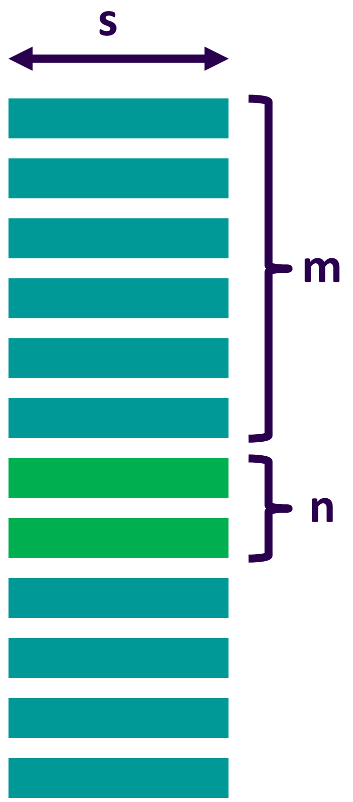
Error correction
Information is recovered



2 Error corrections possible
Information is recovered

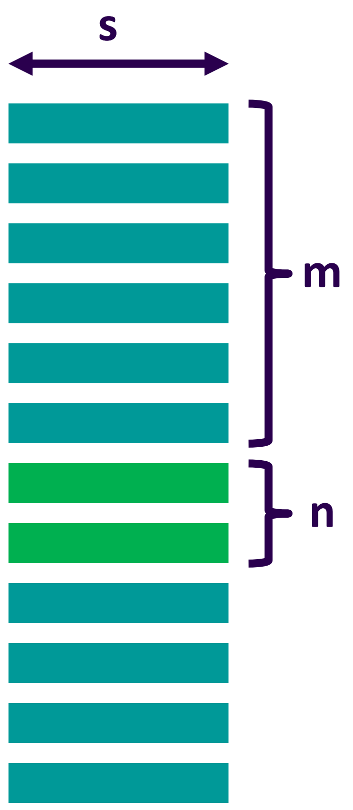
Arbitrary reliability

- For increased flexibility, we could use files ... but files do not have constant size
- File “chunks” (or “blocks”) is the solution
  - Split files in chunks of size “s”
  - Group them in sets of “m” chunks
  - For each group of “m” chunks, generate “n” additional chunks so that
    - For any set of “m” chunks chosen among the “m+n” you can reconstruct the missing “n” chunks
  - Scatter the “m+n” chunks on independent storage



Arbitrary reliability with the “chunk” based solution

- The reliability is independent form the size “s” which is arbitrary.
  - Note: both large and small “s” impact performance
- Whatever the reliability of the hardware is, the system is immune to the loss of “n” simultaneous failures from pools of “m+n” storage chunks
  - Both “m” and “n” are arbitrary. Therefore arbitrary reliability can be achieved
- The fraction of raw storage space loss is n / (n + m)
- Note that space loss can also be reduced arbitrarily by increasing m
  - At the cost of increasing the amount of data loss if this would ever happen



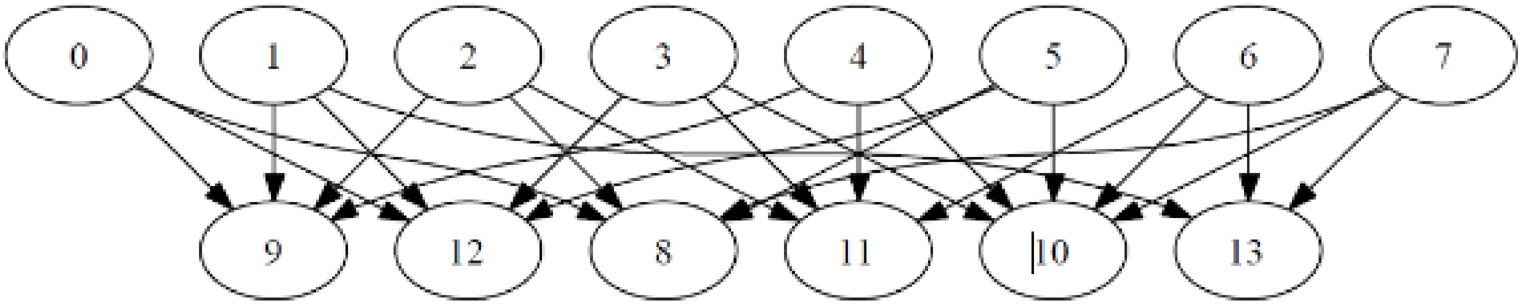
Analogy with the gambling world

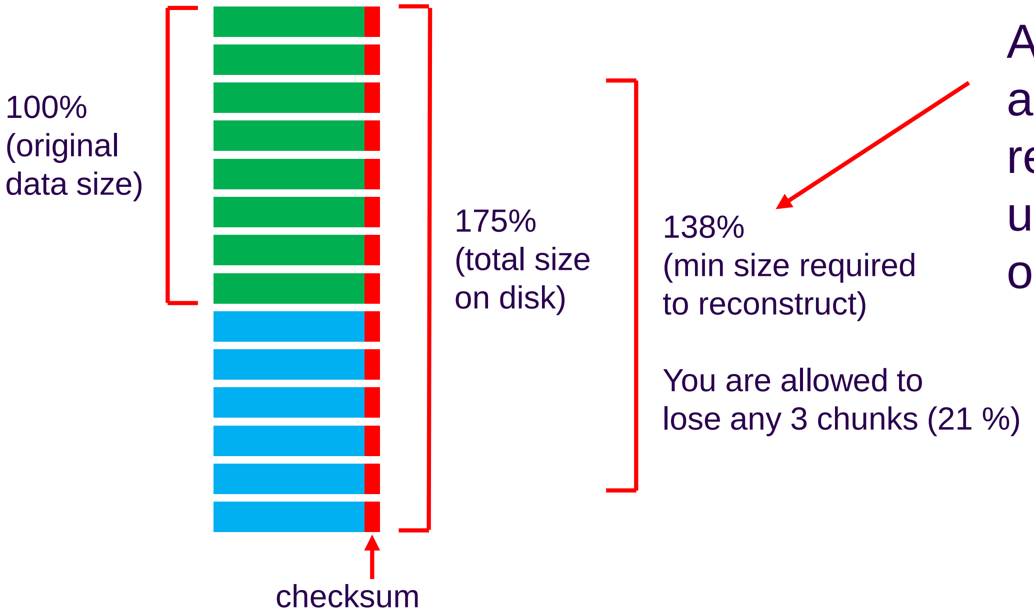
- We just demonstrated that you can achieve “arbitrary reliability” at the cost of an “arbitrary low” amount of disk space. This is possible because you increase the amount of data you accept loosing when this rare event happens.
- In the gambling world there are several playing schemes that allows you to win an arbitrary amount of money with an arbitrary probability.
- Example: you can easily win 100 Euros at > 99 % probability ...
  - By playing up to 7 times on the “Red” of a French Roulette and doubling the bet until you win.
  - The probability of not having a “Red” for 7 times is (19/37)7 = 0.0094)
  - You just need to take the risk of loosing 12’700 euros with a 0.94 % probability

Amount		Win		Lost	
Bet	Cumulated	Probability	Amount	Probability	Amount
100	100	48.65%	100	51.35%	100
200	300	73.63%	100	26.37%	300
400	700	86.46%	100	13.54%	700
800	1500	93.05%	100	6.95%	1500
1600	3100	96.43%	100	3.57%	3100
3200	6300	98.17%	100	1.83%	6300
6400	12700	99.06%	100	0.94%	12700

Error correction example: 8+6 LDPC

0 .. 7: original data
8 .. 13: data xor-ed following the arrows in the graph

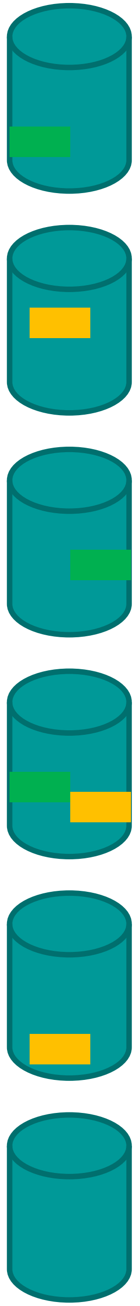




Any 11 of the 14 chunks are enough to reconstruct the data using only XOR operations (very fast)

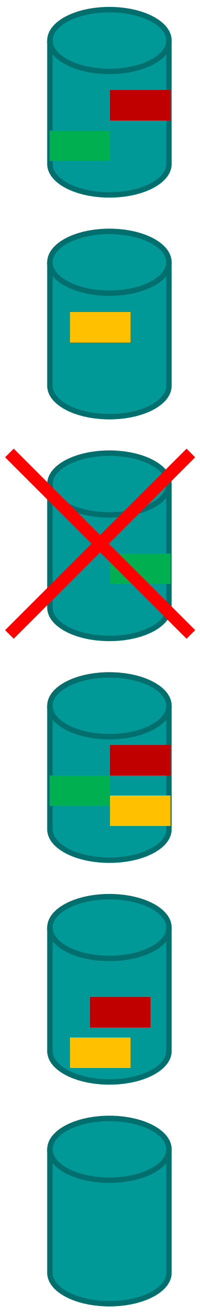
Example: High Availability with replication

- We have “sets” of T independent storage
  - This example has T=6
- The storage pool is configured to replicate files R times, with R < T
  - This example: R=3 every file is written 3 times on 3 independent storage out of the 6 available
  - When a client read a file, any copy can be used
  - Load can be spread across the multiple servers to ensure high throughput (better than mirrored disks, and much better than Raid 5 or Raid 6)



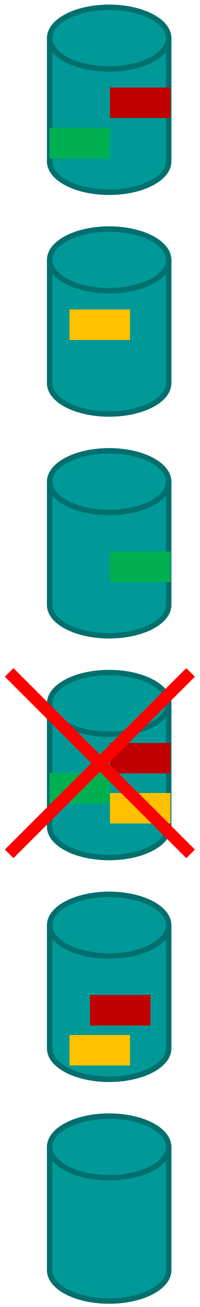
Example scenario: hardware failure

- The loss of a storage component is detected. The storage component is disabled automatically
- File Read requests can continue if R>1 (at least 1 replica), at reduced throughput
  - The example has R=3
- File Creation / Write requests can continue
  - New files will be written to the remaining T – 1 = 6 – 1 = 5 storage components
- File Delete request can continue
- File Write / Update requests can continue
  - Either by just modifying the remaining replicas or by creating on the fly the missing replica on another storage component
- Service operation continues despite hardware failure. (remember: independent storage)



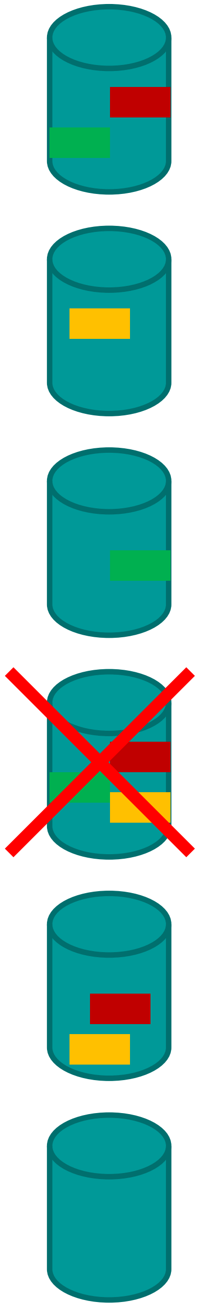
Example scenario: failure response

- The disabled faulty storage is not used anymore
- There is “Spare Storage” that can be used to replace faulty storage
  - manually or automatically
- The lost replicas are regenerated from the existing replicas
  - Manually or automatically



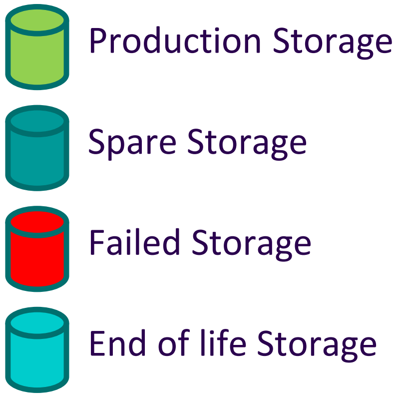
Example scenario: draining a server

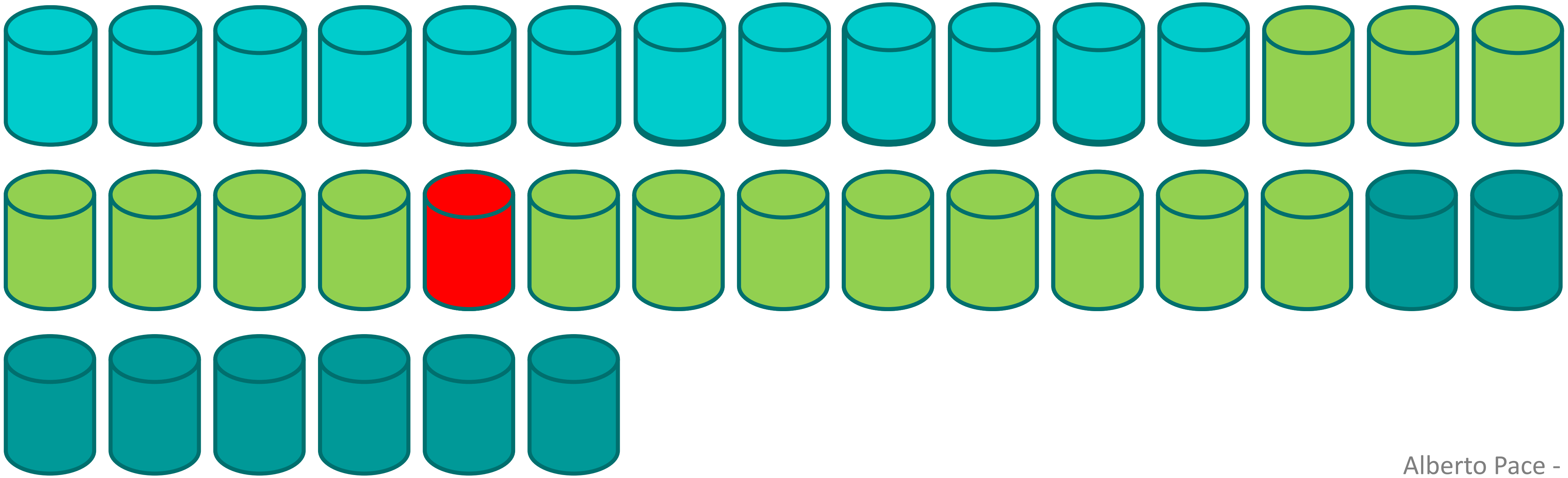
- To drain a server, just power it off
- Will be seen as faulty and disabled (it will not used anymore)
- The available “Spare Storage” will be used to replace faulty storage
  - manually or automatically
- The lost replicas are regenerated from the existing replicas
  - Manually or automatically



Service operation eased …

- Production cluster, 15 Server with 9 spare
- Server Failure (servers)
- New HW delivery (6 servers)Out of warranty (6 servers)
- End of life





Summary

- Data Management solves the following problems
  - Data reliability
  - Access control
  - Data distribution
  - Data archives, history, long term preservation
  - In general:
    - Empower the implementation of a workflow for data processing

www.cern.ch