Quantitative data analysis #2: practical examples

Domenico Giordano, Andrea Valassi (CERN IT-SDC)
With contributions from and many thanks to Hassen Riahi

White Area Lecture, 3rd June 2015

(Follow-up to the previous White Area Lecture on 18th February 2015)

REMINDER!
(WA#1 Feb 2015)

Outline (WA #1)

- Measurements and errors
  - Probability and distributions, mean and standard deviation
  - *Introduction to tools and first demo*
- Populations and samples
  - What is statistics and what do we do we do with it?
  - The Law of Large Numbers and the Central Limit Theorem
  - *Second demo*
- Designing experiments
- Presenting results
  - Error bars: which ones?
  - Displaying distributions: histograms or smoothing?
- Conclusions and references

REMINDER!
(WA#1 Feb 2015)

Conclusions (WA #1)

- Statistics has implications at many levels and in many fields
  - Daily needs, global economy, HEP, formal mathematics and more
  - Different fields may have different buzzwords for similar concepts
- We reminded a few basic concepts
- And we suggested a few tools and practices

REMINDER! (WA#1 Feb 2015)

Take-away messages? (WA #1)

- Do use errors and error bars!
  - When quoting measurements and errors, check your significant figures!
  - Different types of error bars for different needs! Say which ones you are using!
    - Descriptive, width of distributions – standard deviations σ, box plots…
    - Inferential, population mean estimate uncertainty – standard errors $\sigma/\sqrt{n}$, CIs…
    - *[Why do we use $\sigma/\sqrt{n}$? Because of the Central Limit Theorem!]*
    - *[Ask yourself: are you describing a sample or inferring population properties?]*

- Beware of long tails and of outliers!
  - More generally: we all love Gaussians but reality is often different!
    - *[Why do we love Gaussians? Because maths becomes so much easier with them!]*
    - *[Why do we feel ok to abuse Gaussians? Because of the Central Limit Theorem!]*

- Before analyzing data, design your experiment!
  - Aim for reproducibility, reduce external factors – and it is an iterative process

- Make your plots understandable and consistent with one another
  - Label your axes and use similar ranges and styles across different plots
  - Be aware of binning effects (do you really prefer KDEs to histograms?)

Outline (WA #2)

- Follow-up about IPython, tools, repositories...
- Data analysis in practice
  - Data analysis is an iterative process!
  - Data samples
  - Data granularity
- Practical examples as meta-analyses *(with demos)*
  - Hassen's case study: FTS transfer monitoring and optimization
  - Domenico's case study: analysis of Ganglia metrics
- Conclusions

Python analysis tools on AFS at CERN

- Goal: one-click script for anyone to setup the full environment
  - ***DONE!*** *(details on next slide) – already used successfully by LucaMe*

- A consistent set of Python packages is now installed on AFS
  - Many thanks to Patricia Mendez Lorenzo from PH-SFT
  - SLC6 and CC7, part of same software stack as ROOT, CORAL, COOL
    - *Better out-of-the-box build integration between Python tools and ROOT*

- This is being continuously maintained and improved
  - Missing some packages (e.g. pandas), will be added in a next iteration
  - For this iteration I added these packages on another public AFS area
  - You can also easily complement this setup using python easy-install

Useful tools on gitlab.cern.ch

http://letsgo.gorizia.it/ristorazione/ricette/gubana

NB: No pythons were harmed to make this pie!

- Committed all tools to gitlab as package ipypies (yummy!)
  - one-click setup script
  - one-click startup script (but you may prefer your own configuration)
  - the notebooks used for these two White Area lectures

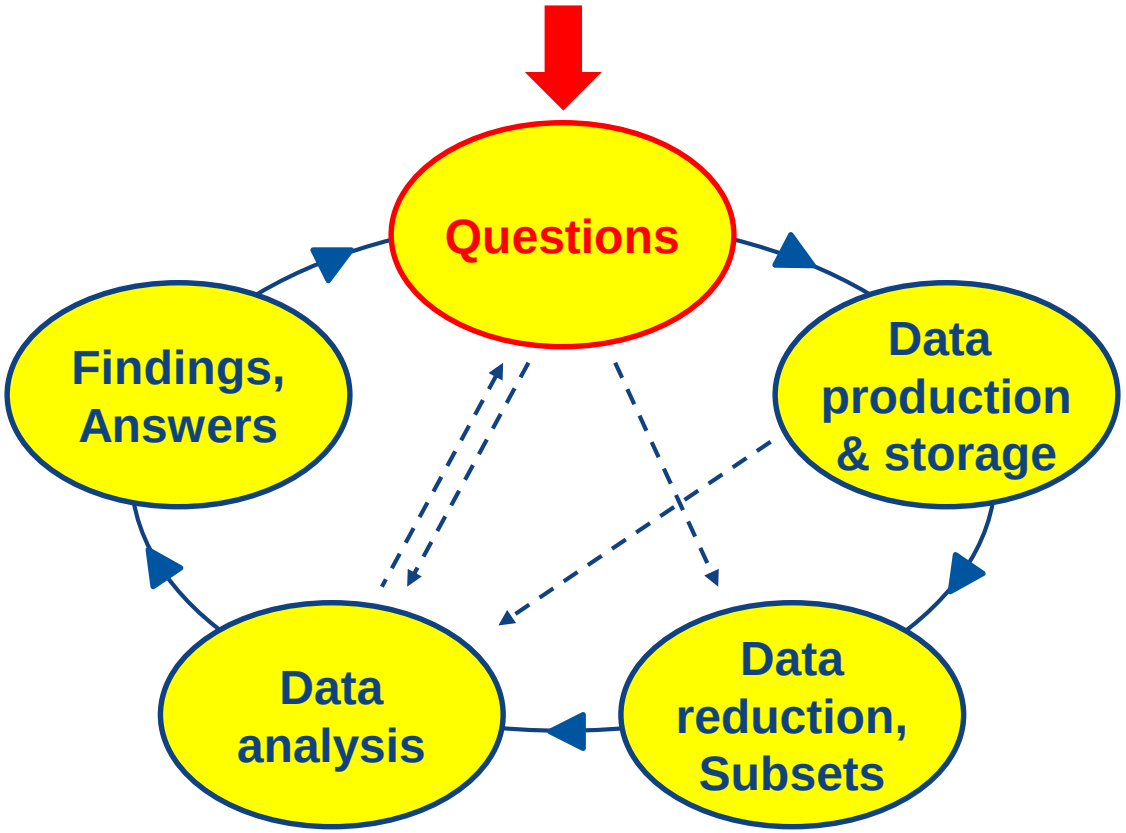
- View notebooks from any *public* URL on nbviewer.ipython.org
  - added direct links in the README.md of the relevant directories
    - example: http://nbviewer.ipython.org/urls/gitlab.cern.ch/avalassi/ipypies/raw/master/NOTEBOOKS/WhiteArea2015/Lecture1_Feb2015/WA_AV/Hello_World.ipynb
  - GitHub provides better integration with nbviewer than GitLab
    - directory navigation within nbviewer, links to nbviewer within GitHub
    - discussed this with IT-PES (need own nbviewer for private notebooks)

Other news related to IPython

- Major changes in IPython v3: two separate components
  - Jupyter is now the language-agnostic part, including notebooks
  - IPython is the language-specific kernel (non-Python kernels also exist)
  - This is the version included in the ipypies setup

- The ROOT team are also interested in IPython and notebooks
  - As new GUI, as new parallel processing engine (ROOT-as-a-service)…
    - Investigating ROOT as a new non-Python kernel within Jupyter
    - See Pere Mato's talk at the recent LHCb Computing Workshop
  - We had a chat with them last week and plan to follow up

Data analysis is an ITERATIVE process!



It starts with questions!

(You would not even store data if you did not think that it could eventually be useful to address some questions → data model)

There is often a "default" loop, but may also take sub-loops

The more you analyse the data, the more you have new questions!

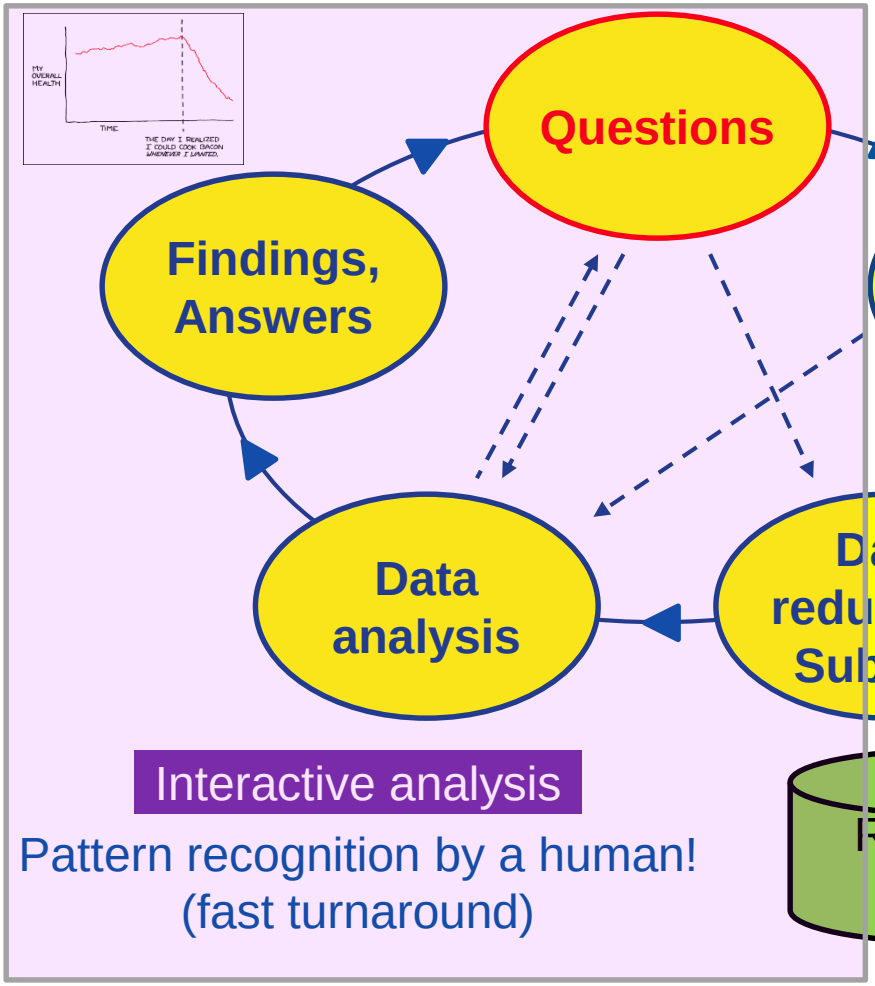
(And the more you will need to rethink your data model and your data storage and processing strategies…!)

Data granularity and aggregation

- Storing all data generated by an experiment is often impossible
  - HEP experiments use “triggers” to select the relevant data to keep
    - Some triggers may even be “downscaled” to randomly select a data fraction
    - And even after triggering they may store only pre-reduced data
  - For the IT “monitoring” data we are most concerned with, some raw data are kept, some are thrown away and then only sums/averages are kept

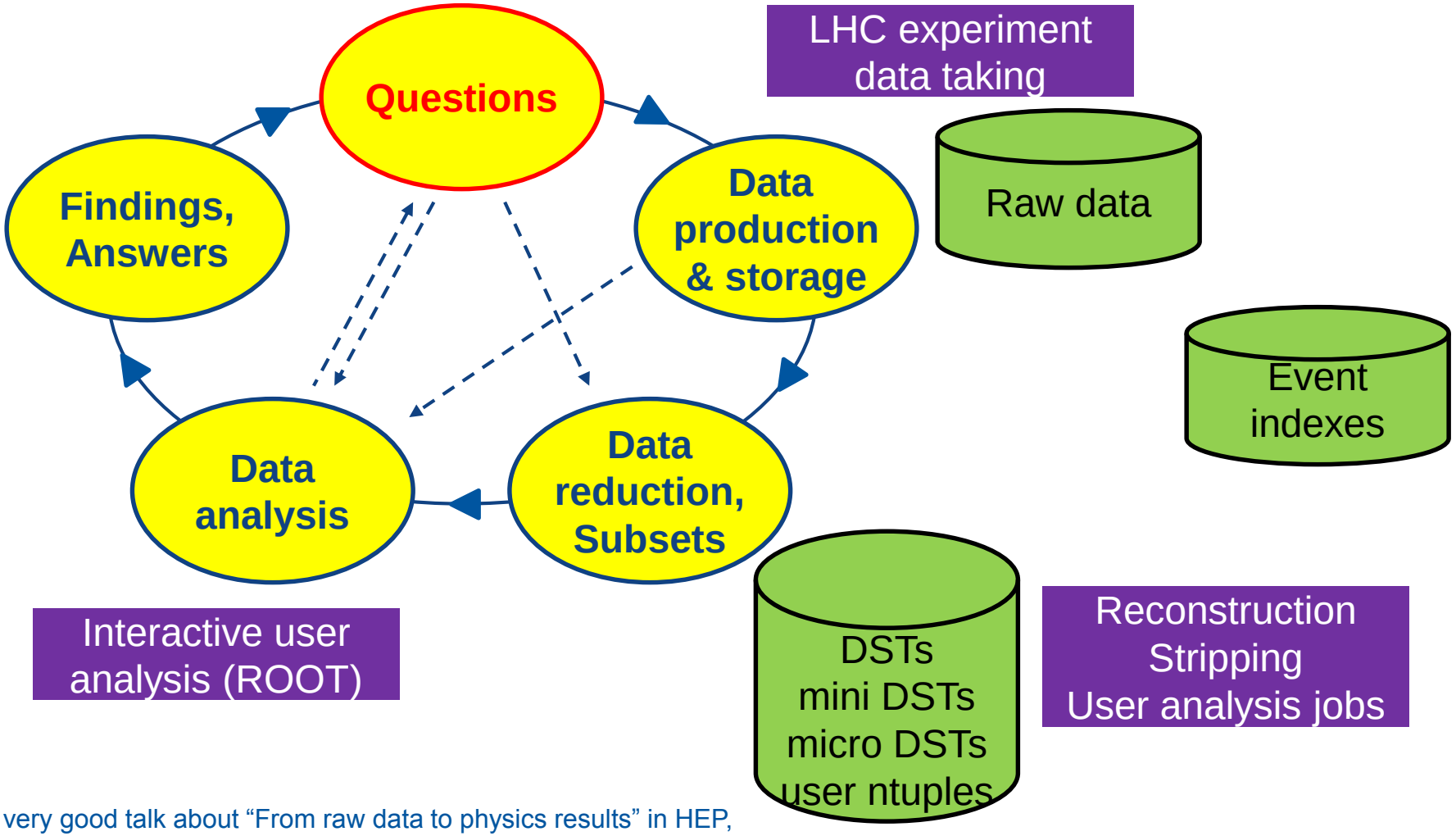
- Aggregate data contains less information than individual data
  - By only storing sums and averages, you are likely to lose information about the differences between the categories of data in your sample
  - Kind of obvious, but related to very fundamental concepts in statistics (sufficient statistics in parameter estimation, Fisher information…)

Data samples and data processing



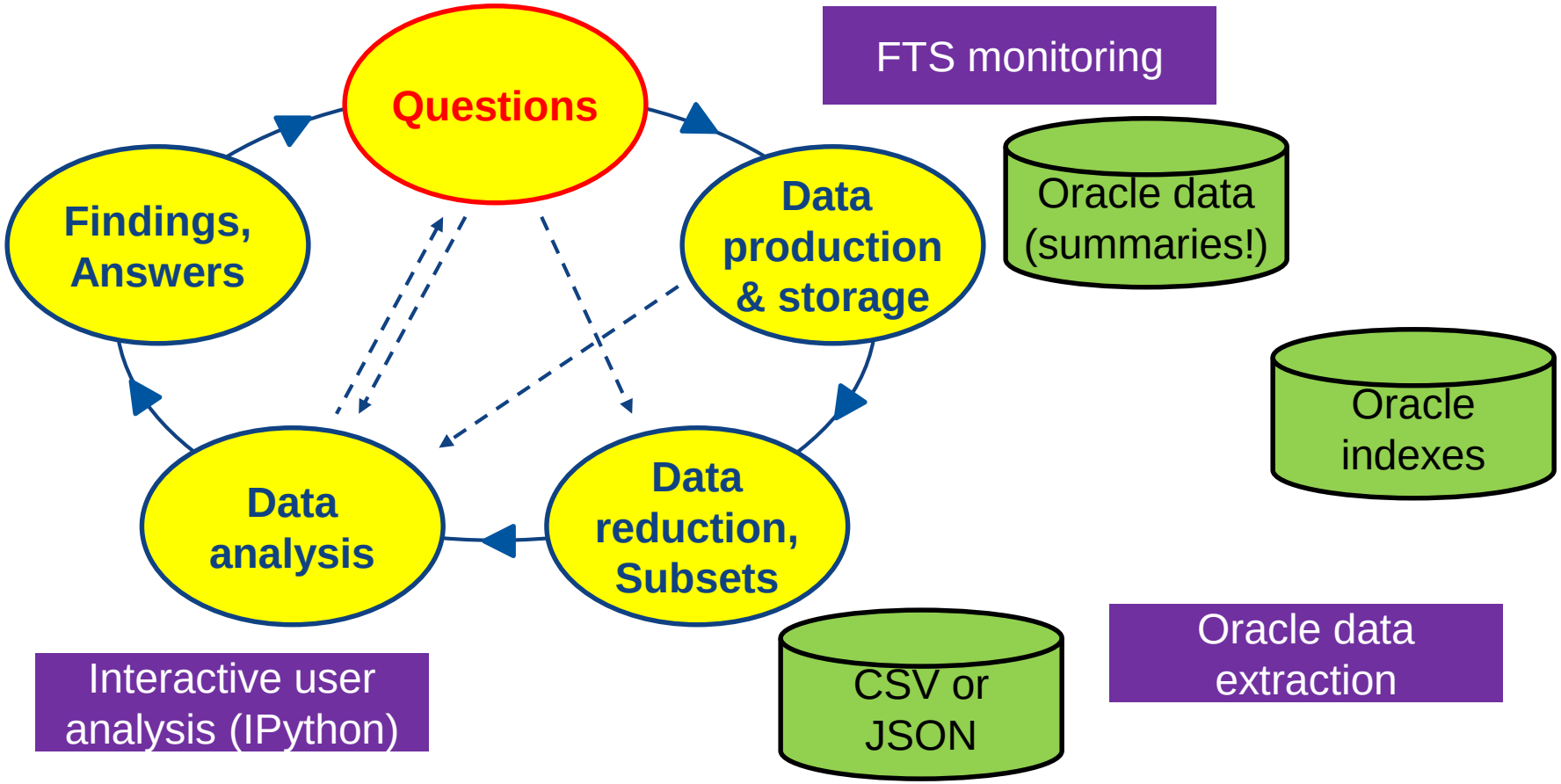
- These WAs focus on interactive analysis (e.g. IPython)
  - but we will also cover the other phases as they are very relevant too!

HEP data samples and data processing



For a very good talk about "From raw data to physics results" in HEP, see G. Dissertori's CERN Summer Student Lecture 2010

FTS transfer analysis data samples



This is the analysis model for Hassen's case study that will be presented later

Case studies (as “meta-analyses”)

- We now present some case studies as practical examples

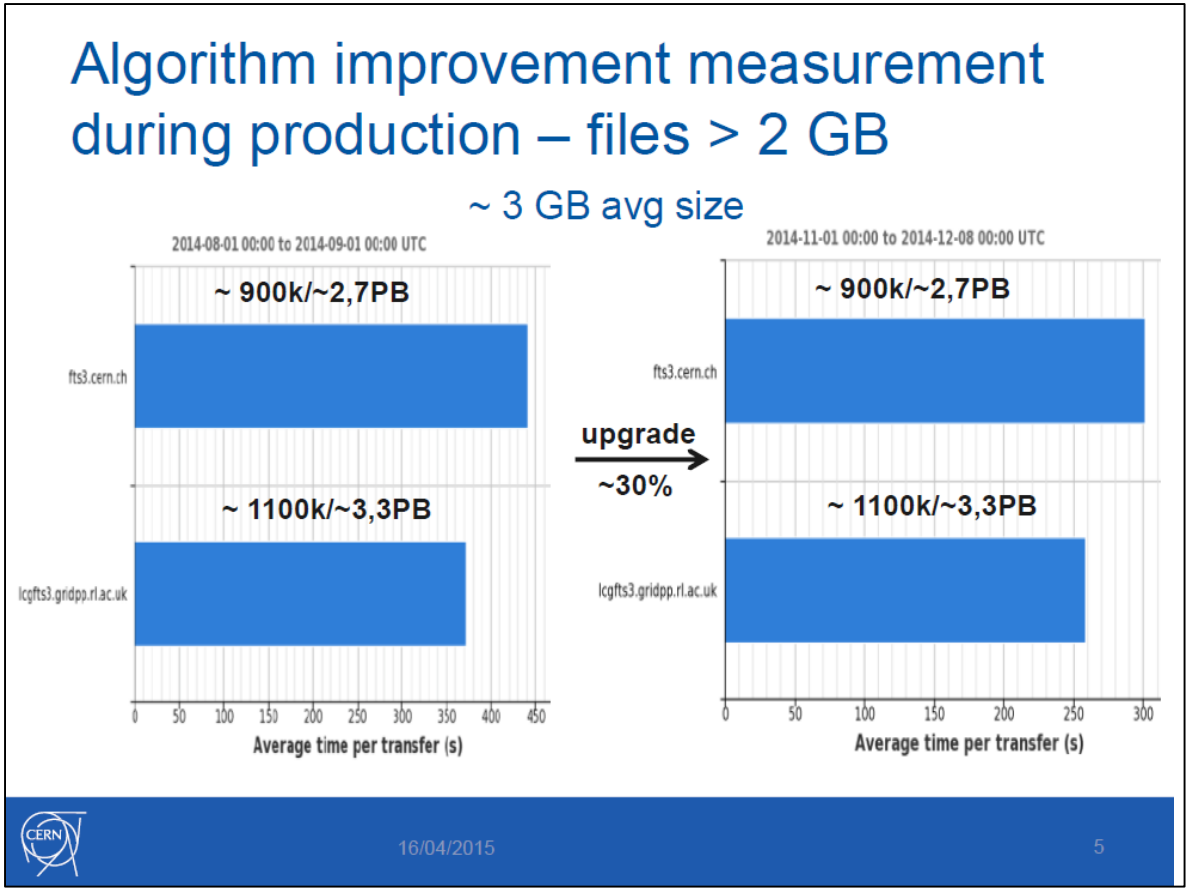
- ***NB: DISCLAIMER! The point here is only to describe the analysis process, not to present any actual results!!***

- This is why I used the term “meta-analysis”, an analysis of the analysis
  - (even if this term is actually used in statistics with a slightly different specific meaning...)

http://xkcd.com

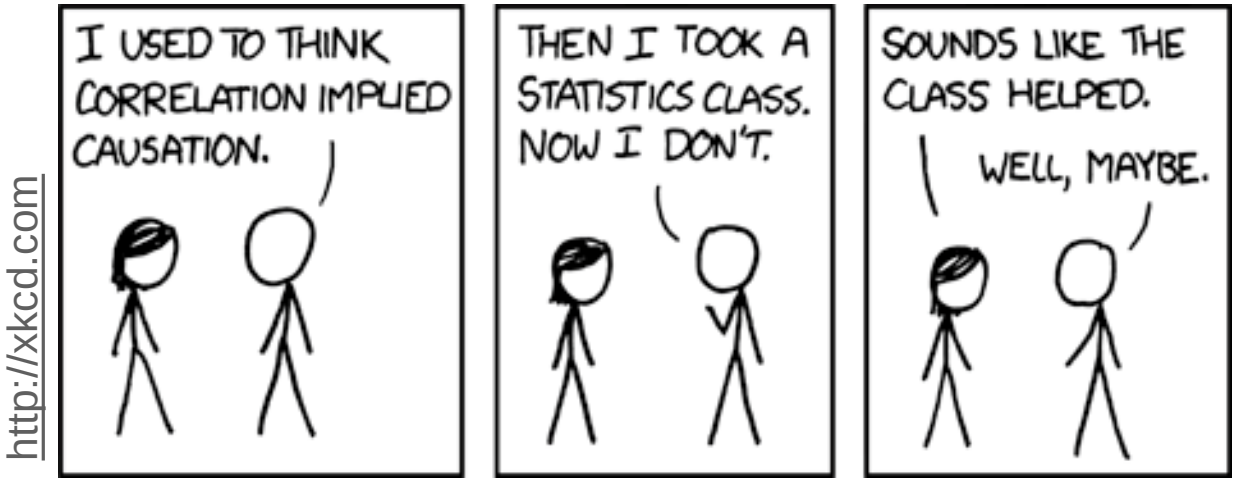
Hassen's case study – FTS transfers

- FTS entered production in August 2014
- Changed an algorithm during September
  - Question: did this lead to an improvement?
- Hassen's presentation:
  - compared transfers for files >2 GB over ~1 month, Aug vs Nov
  - average transfer time decreased by ~30%
  - Question: does this prove that the new algorithm is better?



Correlation does not imply causation

- In the FTS transfer time case study:
  - did anything else change from Aug to Nov, apart from the algorithm?
  - are there other variables more relevant than the algorithm choice?
  - are (average) transfer times the most relevant metric?
- Remember: by aggregating data you may lose information
  - look at distributions, not only at averages
  - look at multiple variables (multi-D distributions), not at a single metric



http://xkcd.com

Overview of the analysis

- Extracted a data subset from Oracle for interactive analysis
  - first in ~json from detailed tables (1 row per transfer)
    - understood this is only available for the last month, gave up
    - lesson on ~json: keep one row per line and make sure you can read back!
  - then in csv from summary tables (1 row per 10 minutes per channel)
    - only aggregate info available (e.g. average file size in 10 minutes > 2GB)
    - good enough to identify some interesting patterns, but granularity could be improved if necessary (file categories by size? downscaled full detailed?)
- Analysis using pandas DataFrame's (~ntuples)
  - transfer time vs file size – better use throughput?
  - transfer time or throughput vs channel categories

“Demo”

... or rather, scroll through the notebooks in nbviewer...

(notebook1 – read from Oracle and create csv)

(notebook2 – load the csv into pandas and analyse data)

Summary of the FTS case study (1)

CERN endpoint:

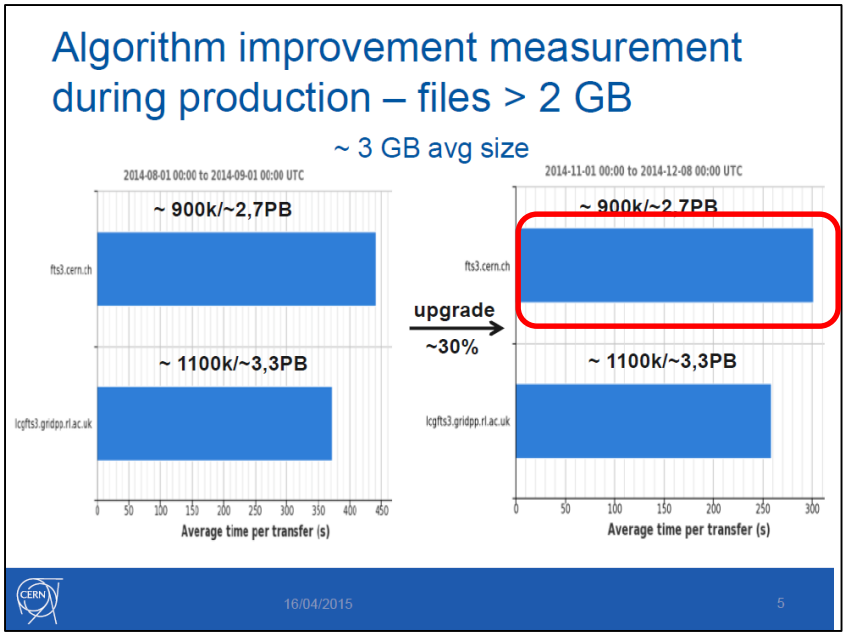
	thrmbs_aug2014	thrmbs_nov2014	avgdttr_aug2014	avgdttr_nov2014
chnctg				
0	31.409692	68.361672	114.674330	43.133902
1	20.069616	20.236231	158.240580	144.101148
2	19.060817	14.270635	179.711897	208.700323
3	5.801345	8.697614	522.881514	340.562432
4	4.350692	6.824802	651.070185	429.125395

RAL endpoint:

	thrmbs_aug2014	thrmbs_nov2014	avgdttr_aug2014	avgdttr_nov2014
chnctg				
0	8.658036	13.813119	357.502049	214.865096
1	11.110499	17.137084	275.451613	180.808996
2	17.596013	16.126469	189.611936	174.578528
3	8.924890	11.244989	338.823639	276.301510
4	6.760997	8.754934	429.360283	341.644535

```
if src == dst: return 0, "LOCALSITE  "
elif srcdom == dstdom: return 1, "LOCALDOMAIN"
elif srct1 and dstt1: return 2, "TIER1TIER1 "
elif srct1 or dstt1: return 3, "TIER1OTHER "
else: return 4, "OTHEROTHER "
```

Average transfer time (all channels) ~ 300s CERN Nov



- Different channel categories have very different behaviours!
  - seems to be an improvement in all categories, but still too soon to tell
  - the fraction/weight of each category in the overall average is important!

Summary of the FTS case study (2)

- Many things could be studied
  - 1-D distribution showing contributions of different categories
  - why CERN and RAL endpoints are so different?
  - better and finer-grained categorization of channels
  - box plots (y axis) grouped by channel categories (x axis)
  - relevance of #streams (the actual thing that changed in the algorithm)

- The point was to discuss a method, not the results...

Summary (before Domenico's part)

- Data analysis is an iterative process
  - You start with questions, don't know what you'll find, apart from more questions – you need to review your process at each step
  - Use small data sets for fast interactive data analysis!
- Aggregating data you may lose some relevant information
  - Look at individual data (if available, else review raw storage policies)
  - *It is often difficult to draw conclusions from a single number*
- Correlation does not imply causation
  - Look at what else changed – look at multidimensional distributions

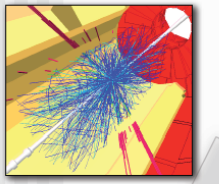
Caravaggio – I bari (1594)

Backup slides

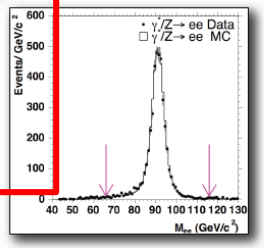
HEP data analysis chain in one slide

Data Analysis Chain

ETH Institute for Particle Physics



- Have to collect data from many channels on many sub-detectors (millions)
- Decide to read out everything or throw event away (Trigger)
- Build the event (put info together)
- Store the data
- Analyze them
  - reconstruction, user analysis algorithms, data volume reduction
- do the same with a simulation
  - correct data for detector effects
- Compare data and theory

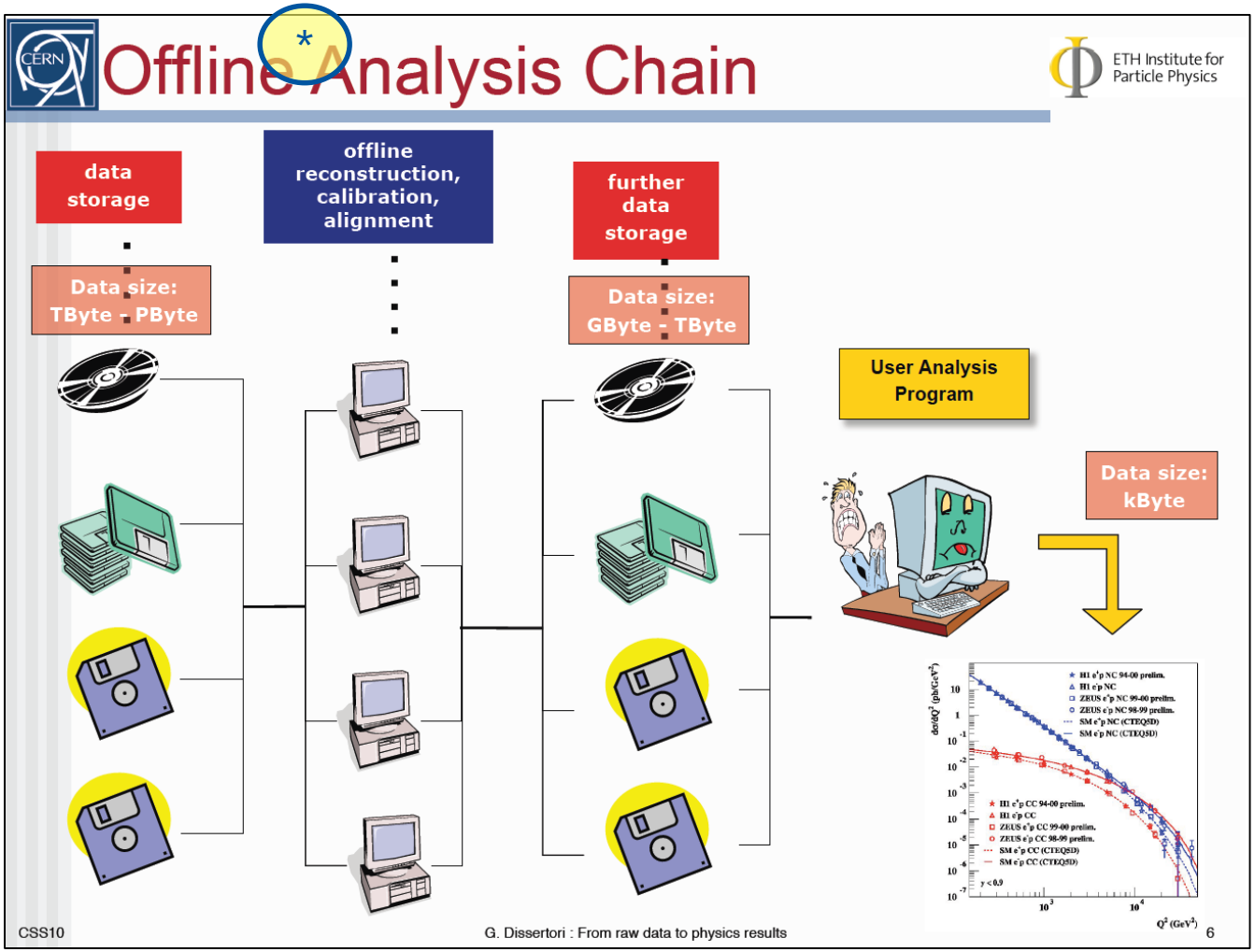


CSS10 G. Dissertori : From raw data to physics results 4

G. Dissertori, CERN Summer Student Lectures 2010

A brilliant talk that I highly recommend!

Raw data → Intermediate data → Plots

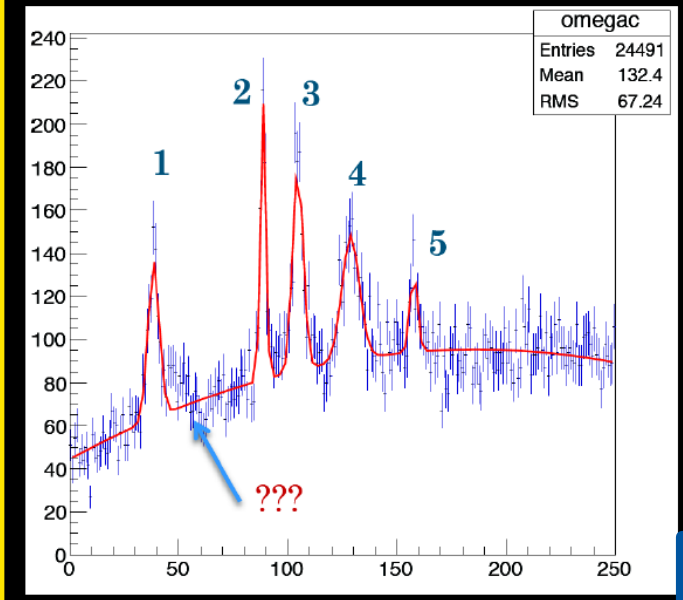


*Distinction between online and offline is getting more blurred in HEP these days!

G. Dissertori, CERN Summer Student Lectures 2010

An HEP example from LHCb

More flexibility



$\Omega_c^0 K^+$ mass plot. [Marco Pappagallo]

Very exciting $\Omega_c^0 K^+$ spectrum with five peaks seen.

➜ Publish asap

- But stripping line does not contain the wrong sign $\Omega_c^0 K^-$ candidates.
- Could wait for a restripping
- Instead, decided to add 2015 data in the publication

We don't know in advance what we'll find and thus it's hard to plan everything.

That's why it's called research

Raw data vs reduced data

Research is an iterative process!!!