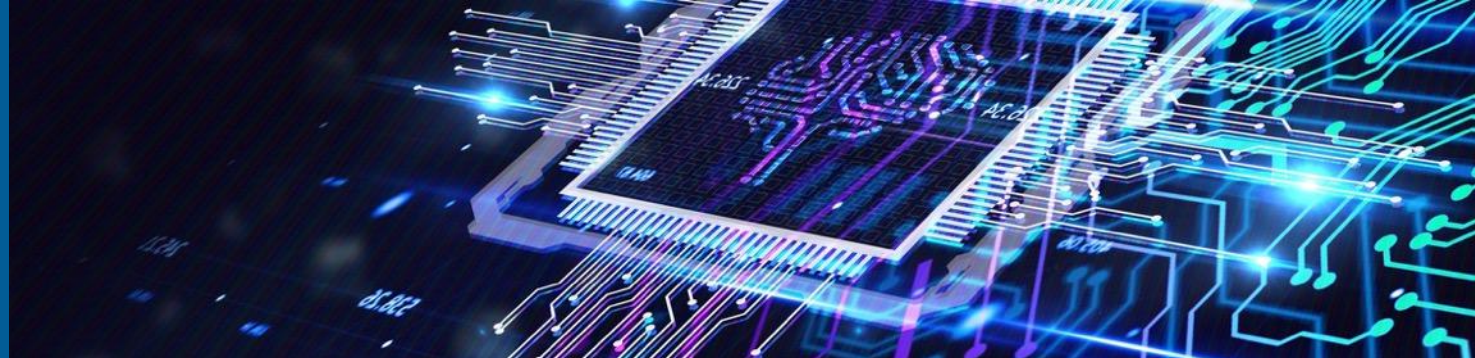




CSC

ICT Solutions for
Brilliant Minds



Hiukkasen toistettavuudesta: case CERN

Katja Mankinen

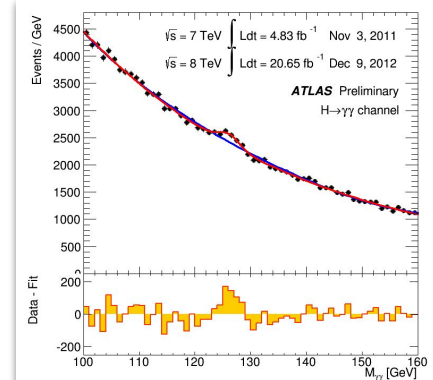
Toistettavuus-webinaari 23.4.2021



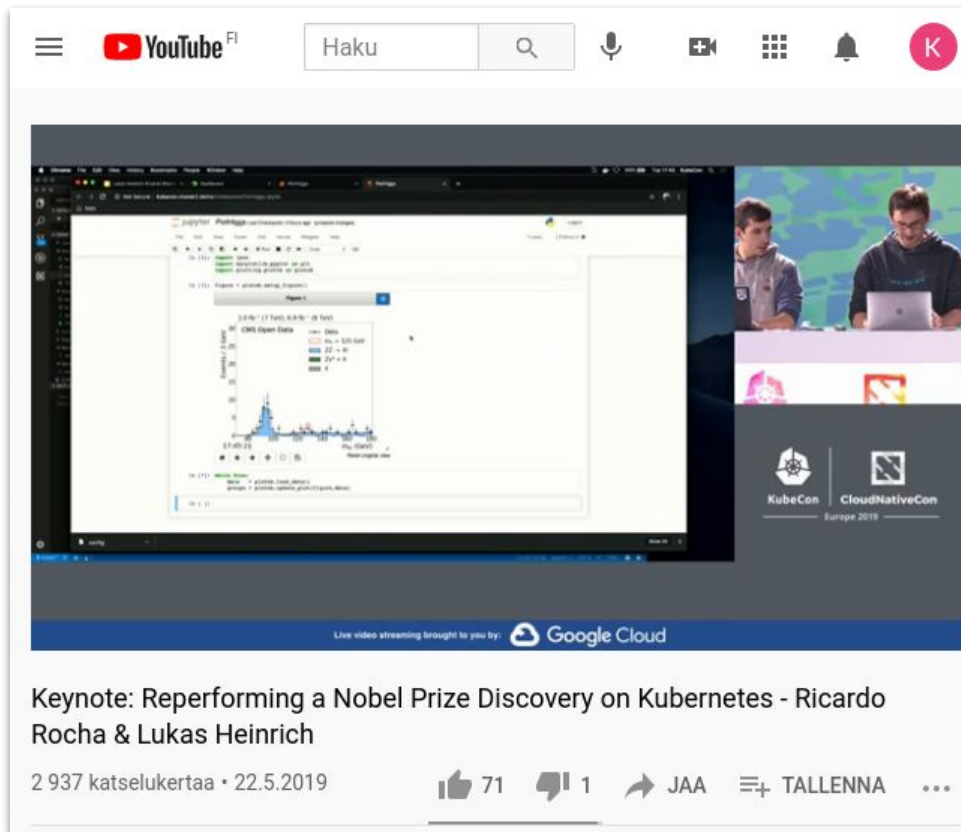
Higgsin bosoni 4.7.2012



Voidaanko analyysi
toistaa lähes 10 vuoden
jälkeen?



Kyllä!



The image shows a YouTube video player interface. The video content displays a presentation slide with a histogram titled "CMS Open Data". The histogram shows a distribution of values, with a peak around 0.145. The slide also includes a table of statistics:

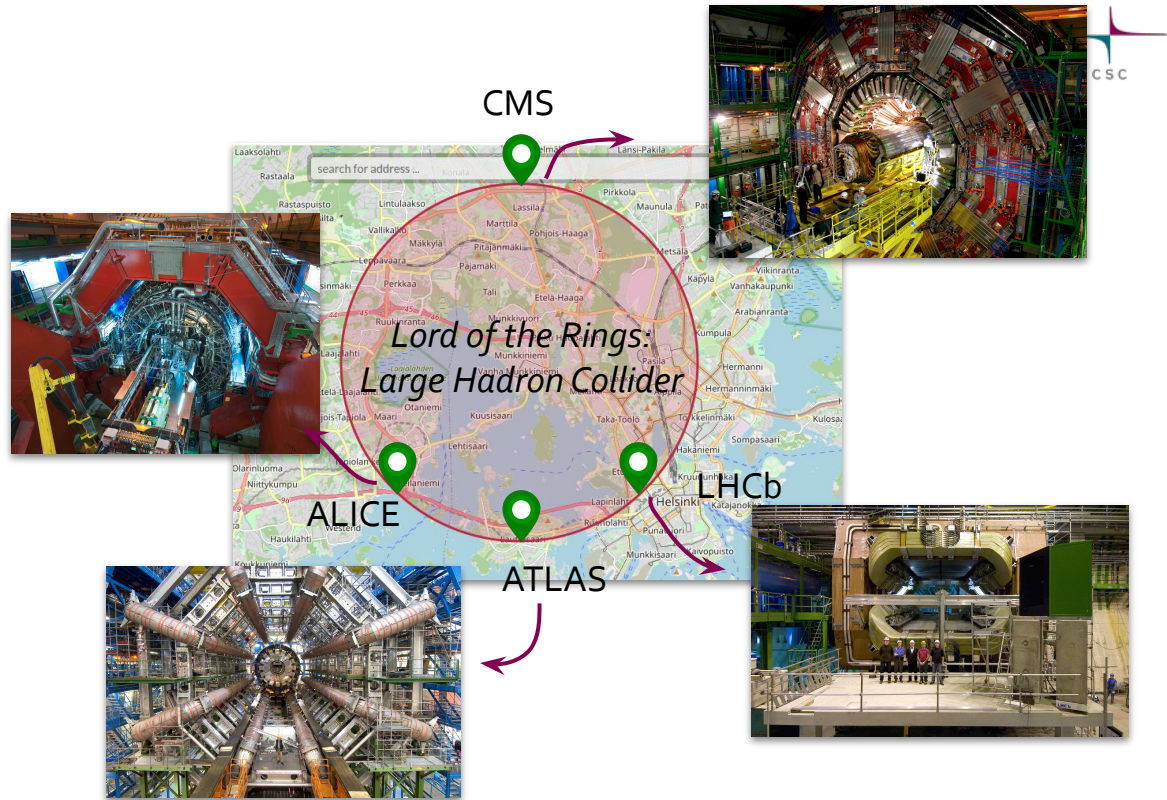
Stat	Value
Mean	0.145
Std	0.005
Min	0.140
Max	0.150

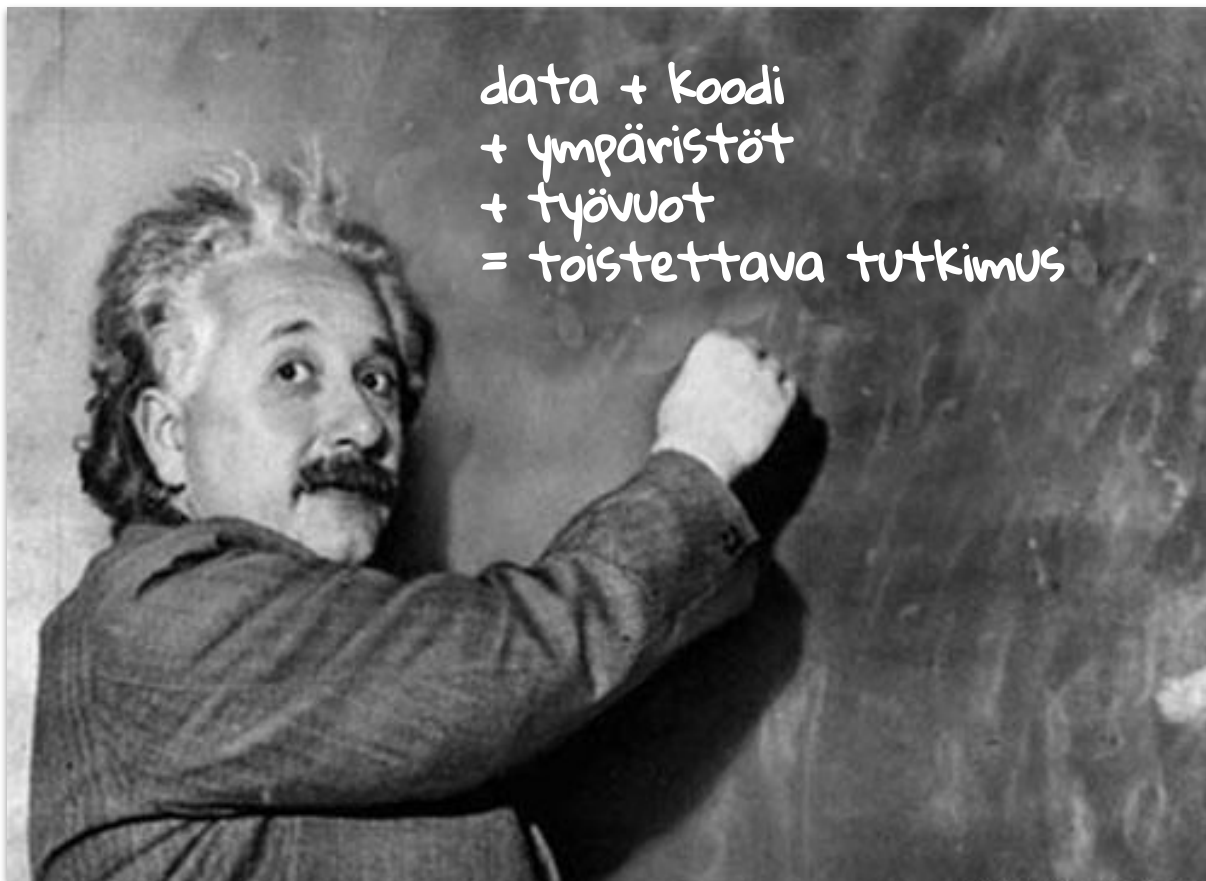
. The video player interface includes a search bar with the text "Haku", a microphone icon, a video camera icon, a grid icon, a bell icon, and a profile icon with the letter "K". Below the video player, the title "Keynote: Reperforming a Nobel Prize Discovery on Kubernetes - Ricardo Rocha & Lukas Heinrich" is displayed, along with the view count "2 937 katselukertaa" and the date "22.5.2019". The video player also shows engagement icons for likes (71), comments (1), and shares (JAA), as well as a list of related videos (TALLENNA).

<https://www.youtube.com/watch?v=CTfp2woVEkA>

CERN

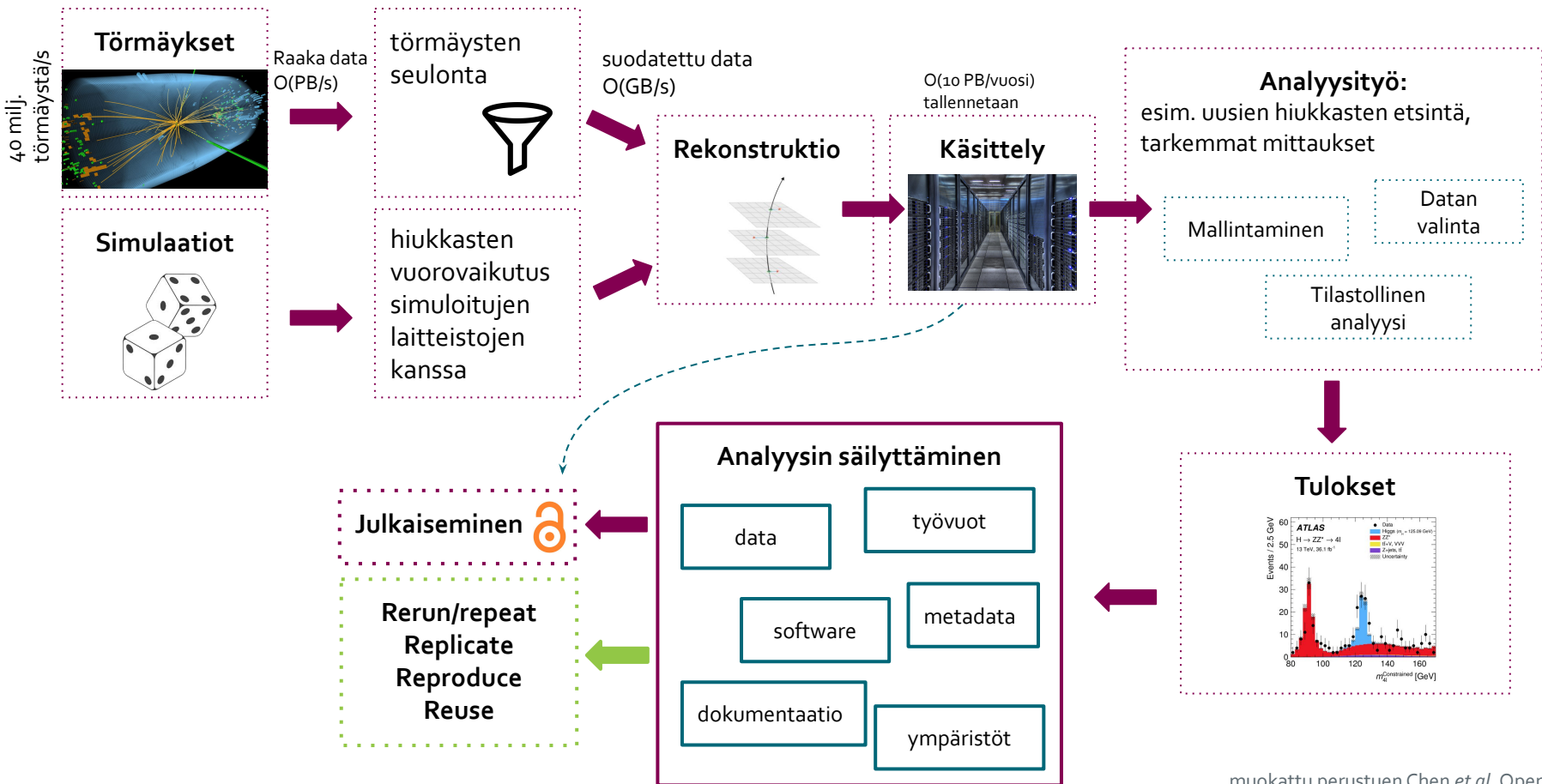
- Miten maailmankaikkeus toimii?
- Perustettu 1954:
"The Organization shall have no concern with work for military requirements and the **results of its experimental and theoretical work shall be published or otherwise made generally available.**"
- ~12 000 vierailevaa tutkijaa
- ~90 PB dataa/vuosi Suuresta Hadronitörmäytimestä (Large Hadron Collider)





data + koodi
+ ympäristöt
+ työvuot
= toistettava tutkimus

Tyypillinen LHC-analyysi



muokattu perustuen Chen *et al.* Open is not enough (2019) ja Diego Rodríguezin kuvaan

hajautetut tiedontallennusjärjestelmät

Monta palasta!



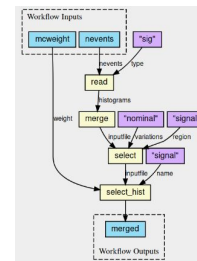
kaggle



HEPData



COMMON
WORKFLOW
LANGUAGE



Singularity



kubernetes

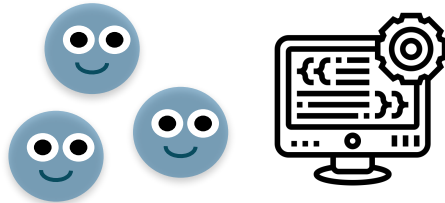


HTCondor
High Throughput Computing



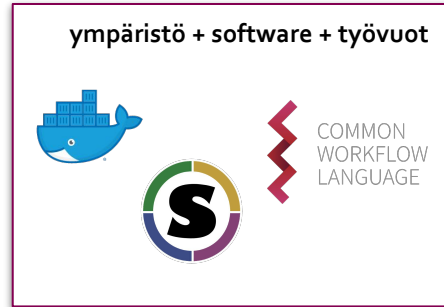
Käyttötapaus: *Rerun, repeat*

Toistetaan tutkimus tismalleen samoilla asetuksilla, tai tutkitaan miten pieni muutos vaikuttaa tulokseen



Samat tutkijat, sama tai eri toteutus

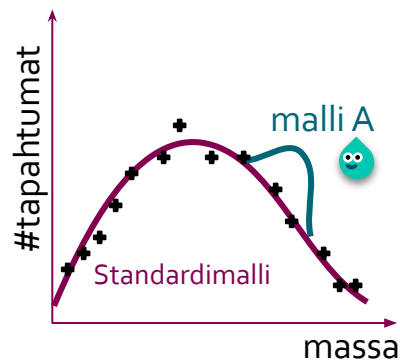
- *eri energia-alue*
- *hieman eri data*
- *hieman eri malli*



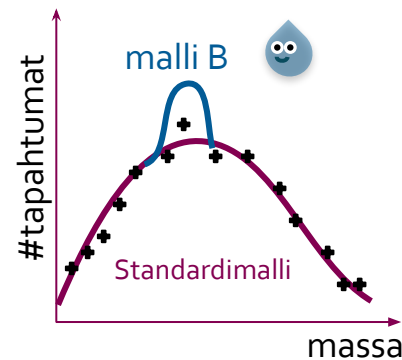
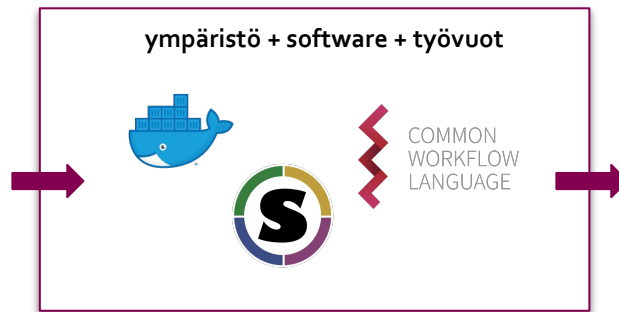
Eri tutkijat, sama toteutus

Käyttötapaus: *Reuse analysis*

“Hei, mulle tuli tosi hyvä idea, joka vois selittää viimeaikaiset havainnot...”



Alkuperäinen analyysi



Alkuperäinen analyysi tulkittu uudelleen toisella mallilla

Käyttötapaus: *Reproduce*

Higsin löytäminen, CMS avoin data

ympäristö + software + työvuot



COMMON
WORKFLOW
LANGUAGE

reana-demo-cms-h4l/reana.yaml at master · reanahub/reana-demo-cms-h4l · Mozilla Firefox

reana-demo-cms-h4l/reana.yaml

Go to file

dpelipcean reana: transformed into a full reana... Latest commit 238f0e5 on Sep 13, 2019 History

2 contributors TS

13 lines (13 sloc) 216 Bytes

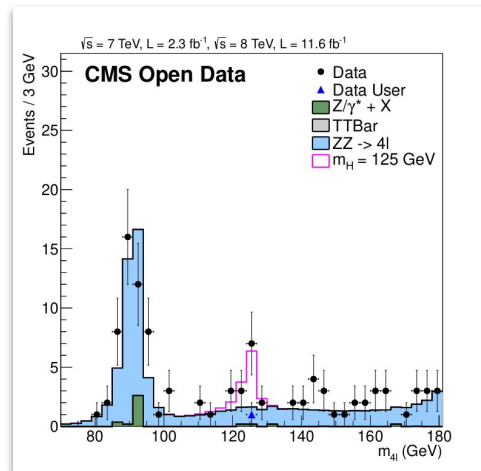
```

1 version: 0.6.0
2 inputs:
3   parameters:
4     input: workflow/input.yaml
5   directories:
6     - data
7     - code
8 outputs:
9   files:
10    - outputs/mass4l_combine_user1v13.pdf
11 workflow:
12   type: cwl
13   file: workflow/workflow.cwl
  
```

koodi, data, parametrit

kuvaajat, taulukot...

komennot, ajoympäristö, konttikuvat



Esimerkkejä:

<https://github.com/reanahub/reana-demo-cms-h4l>

<http://opendata.cern.ch/record/5500>

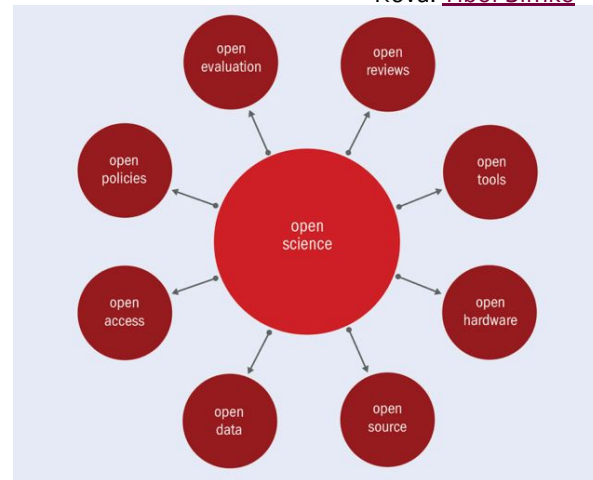
49

Kohti avointa tiedettä!

- Hiukkasfysiikan yhteisö edistää aktiivisesti toistettavaa tutkimusta ja avointa tiedettä
- Avoin data ei yksinään riitä:
tarvitaan myös koodi, ympäristö ja työvuot
 - **data:** sama data, uusi analyysi
 - **koodi:** sama koodi, uusi data
 - **analyysi:** toista aiempi analyysi tai kokeile uutta mallia
- Työkaluja datan ja analyysin säilyttämiseen ja toistamiseen on jo käytössä ja kehitteillä
- Haasteina mm. uudet teknologiat, lisääntyvä datan määrä
- ... mutta myös itse toteutus:
 - työkalujen täytyy sopia tutkijoiden päivittäiseen työhön
 - työkalujen käyttö heti alusta asti helpottaa

Toistettava tutkimus ei ole pelkkää sci-fiä!

Kuva: [Tibor Šimko](#)





Katja Mankinen

PhD, former exotic Higgs hunter at the ATLAS experiment at CERN & Lund University

Data scientist at Research Information Management group
katja.mankinen@csc.fi



facebook.com/CSCfi



twitter.com/CSCfi



linkedin.com/company/csc---it-center-for-science



github.com/CSCfi



Term	Purpose	Description adapted to the HEP context	Example
Rerun Repeat ^a	Defend and make robust	Same implementation, same analysts	These actions are integral parts of HEP analyses. Using datasets recorded in slightly different conditions (for example, different energies or magnet orientation) to verify that the results are consistent across the data subsets.
		Variations in implementation, same analysts	Ensuring robustness by using the same analysis model (experiment simulation) to learn about the process or to verify the results (for example, checking the impact of slight model differences to the final result).
Replicate ^b	Certify	Same dataset, same implementation, independent analysts	Replicate an analysis Analysis certification includes a review using prepared written documentation, and in some scenarios even reproducing the entire (or some parts of the) analysis. Another common scenario of reproducing an analysis occurs when personnel changes.
Reproduce	Compare	Variations in implementation, independent analysts	Reproduce a measurement After the Higgs boson discovery was published by the CMS experiment, a group of independent researchers reproduced the measurement using CMS open data and obtained similar results (Fig. 3).
Reuse	Transfer, new (different) purpose	Original (or subset of) data, different implementation, independent analysts	Reusing analysis data Using data for a new independent discovery, such as the use of CMS data via the CERN Open Data portal by Thaler (MIT) and colleagues ^{31,32} .
	Certify or transfer to a new (different) purpose	Same implementation, same or independent analysts	Reusing analysis code Using analysis elements for a new independent study. For example, redoing an analysis by reusing the code on a more recent dataset to improve the sensitivity (statistics) of a measurement (for example, ‘evidence for’ may become ‘observation’).
	Transfer	Same implementation, same or independent analysts	Reusing complete analyses Testing different analysis models using existing analyses’ resources. This application can be adapted in some (more inclusive) HEP analyses, and it is captured by the RECAST project.

Lähteitä

- Diego Rodríguez (2021). [Open and Reproducible Research Services in LHC Particle Physics](#)
- Fokianos Pamfilos et al. (2020) [CERN Analysis Preservation and Reuse Framework: FAIR research data services for LHC experiment](#)
- Ioannis Tsanaktsidis et al. (2016). [CERN data stories: CERN Analysis Preservation, CERN Open Data and HEPData.](#)
- Sünje Dallmeier-Tiessen and Tibor Šimko (2019). [Open science: A vision for collaborative, reproducible and reusable research](#)
- Tibor Šimko (2017). [Towards Reproducible Research Data Analyses in LHC Particle Physics.](#)
- Tibor Šimko et al. (2019) [REANA: A System for Reusable Research Data Analyses](#)
- Xiaoli Chen et al. (2019) [Open is not enough.](#) Nature Phys 15, 113–119, 2019.