Data analysis of large measurement sets

Ian Marsh, Ph.D, 2020-01-19

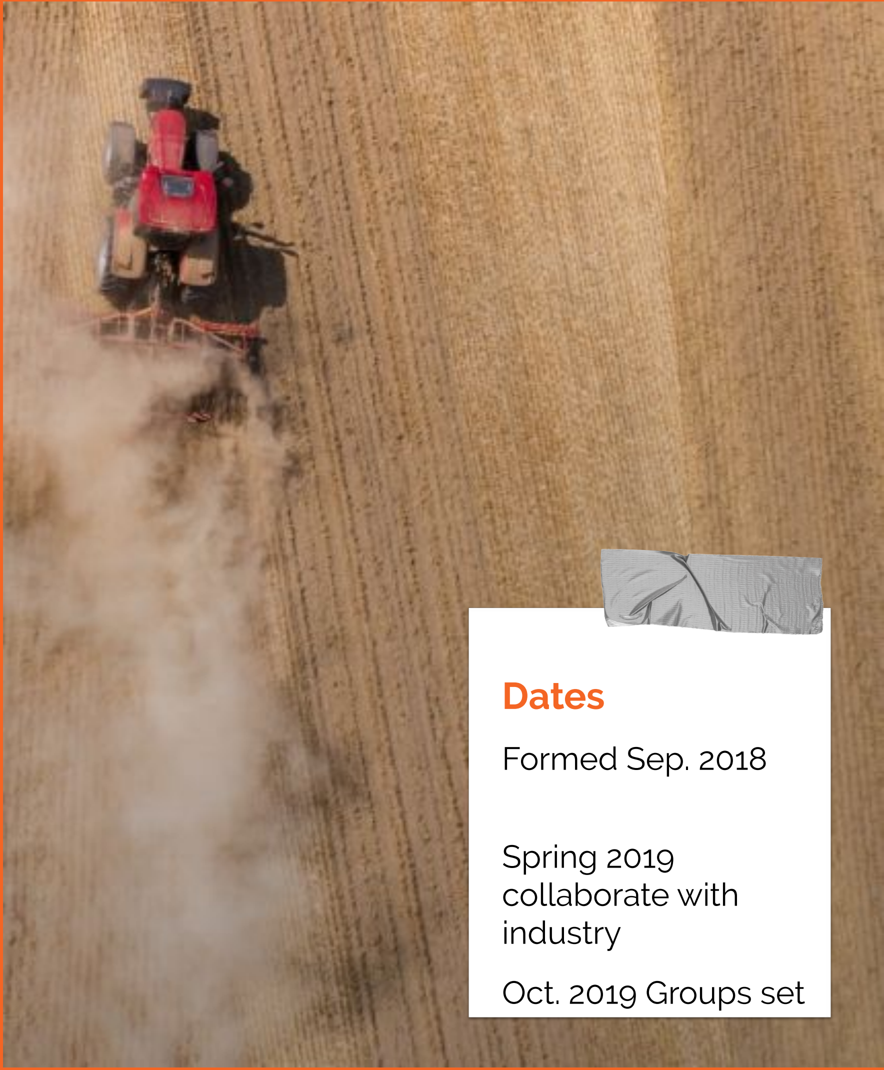
New national institute of Sweden.

2500 employees, 30% PhDs

Swedish industries: Ericsson, Volvo, Scania, AstraZenica, Spotify, H&M

RI.SE covers most areas

https://www.ri.se/en



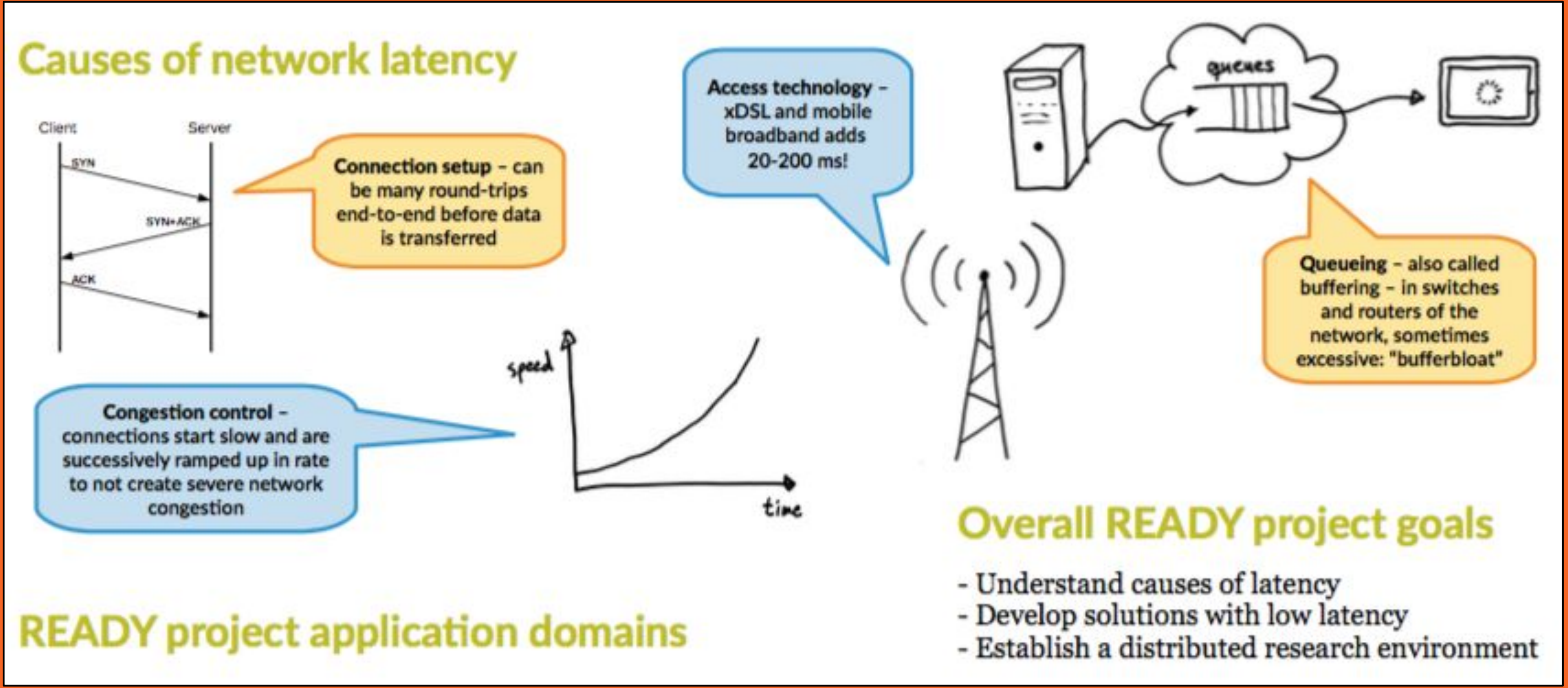
Dates

Formed Sep. 2018

Spring 2019 collaborate with industry

Oct. 2019 Groups set

Network delay



Causes of network latency

Connection setup – can be many round-trips end-to-end before data is transferred

Access technology – xDSL and mobile broadband adds 20-200 ms!

Queueing – also called buffering – in switches and routers of the network, sometimes excessive: "bufferbloat"

Congestion control – connections start slow and are successively ramped up in rate to not create severe network congestion

Overall READY project goals

- Understand causes of latency
- Develop solutions with low latency
- Establish a distributed research environment

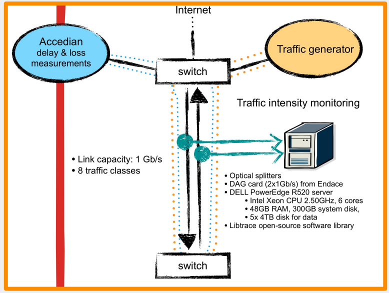
READY project application domains

Project https://ready-sidus.se/

Measurements & datasets

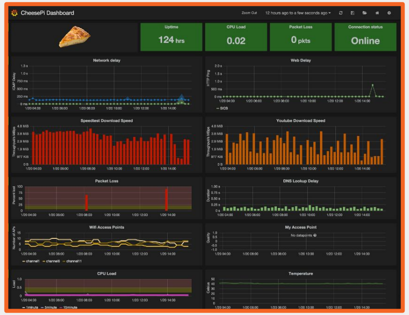
Four example projects:

1. Telia Swedish incumbant measurements
   - ○ Access delay in home & mobile networks
2. CheesePi
   - ○ Home-monitoring, more below
3. EU FIRE+ : MONROE Mobile broadband measurements
   - ■ 4 operators, 4 countries
   - ■ Ours : busses
4. Orange labs
   - ○ Cellular http accesses for 3 months (5TB)
     - ■ Apps: Caching, mobility patterns, ...

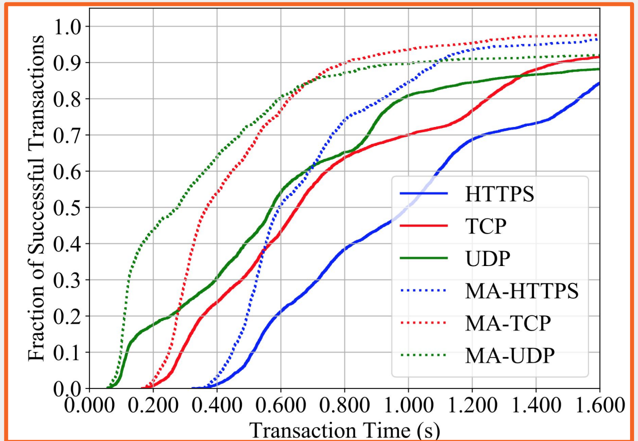

Internet
Accedian delay & loss measurements
switch
Traffic generator
Traffic intensity monitoring
switch

- Link capacity: 1 Gb/s
- 8 traffic classes

- Optical splitters
- DAG card (2x1Gb/s) from Endace
- DELL PowerEdge R520 server
  - Intel Xeon CPU 2.50GHz, 6 cores
  - 48GB RAM, 300GB system disk,
  - 5x 4TB disk for data
- Libtrace open-source software library



Uptime: 124 hrs
CPU Load: 0.02
Packet Loss: 0 pkts
Connection status: Online



Analysis

Different approaches over the years. We have progressed from low-level to a “more” human-friendly presentation.

- ➔ **Protocol**
  Round-trip times, timers, queue lengths
- ➔ **Statistical**
  Summary statistics
- ➔ **Visualisation**
  Plots, dashboards, video
- ➔ **Big data**
  Handling large amounts of data (see URL)

Hops.io RI.SE’s & Europe’s only big data platform)

Lessons learned: provider dialog, capture small data (only), include implicit environment, experimental design, ...

Issues

1. Getting complete data (URL below)
2. Timing always problematic
3. Machine learning needs minimal # of features

https://www.dropbox.com/s/ohq8i5xnn0incmg/FULLTEXT01.pdf

Each measurement campaign needs to learn from previous experiences and often needs 4-5 attempts (sometimes more) to get your campaign correct

http://tma.ifip.org/2018/special-session-cfp/

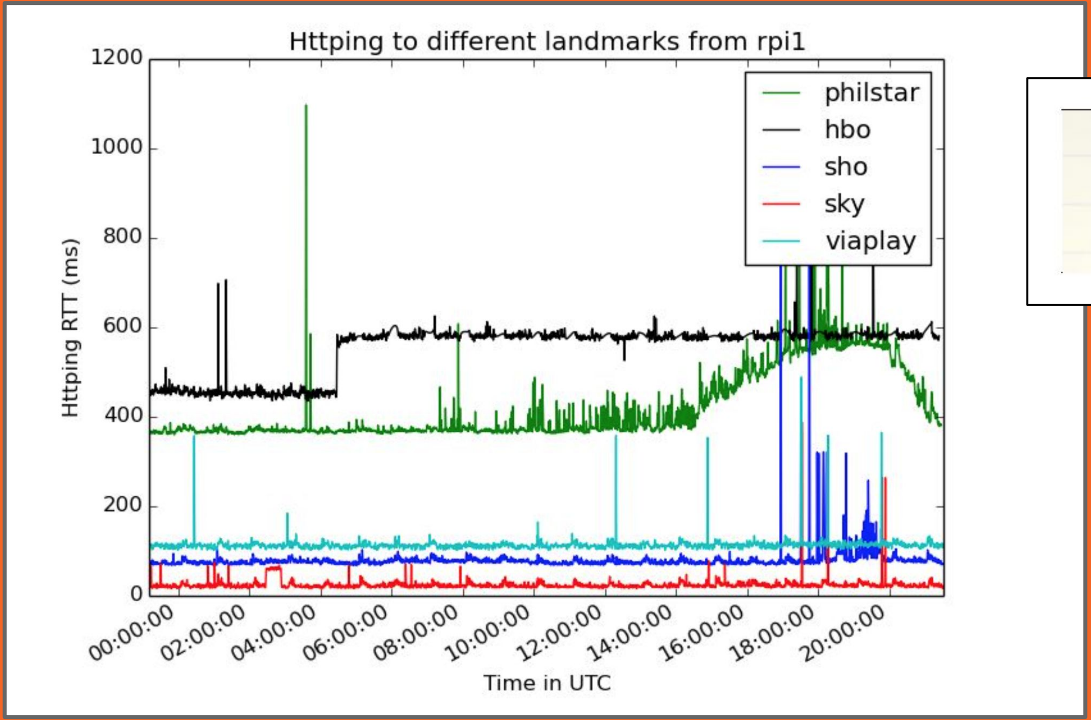
Tips

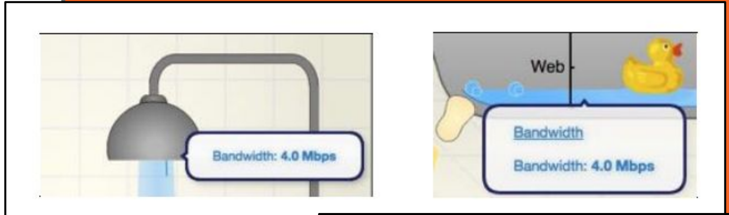
Inferring behaviour from measurements is hard

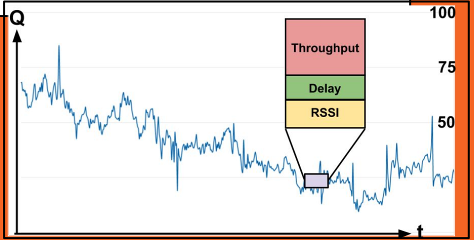
General failure of QoE

Active measurements need to be conducted fairly, to ensure no blame game

Home measurements





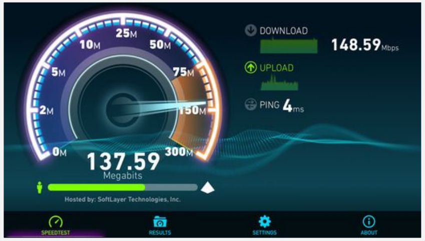


Our project cheesepi.sics.se

Help! I'm home (alone)

Mass-scale measurements:
RIPE Atlas, Geant, ARK, ...

NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community
Connected
Disconnected
Abandoned



Numerous 1-off measurements done by users: Speedtest

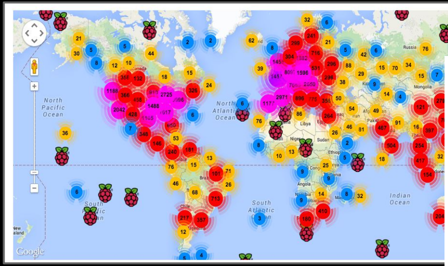
But what if I want to find why *my* Internet is dodgy? and **now**!

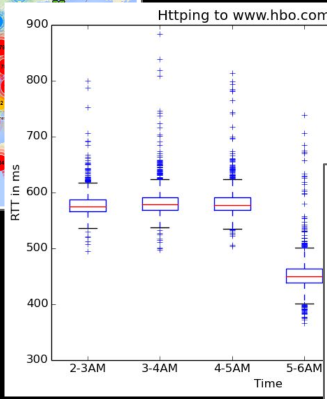
Is this video stream going to frustrate me?

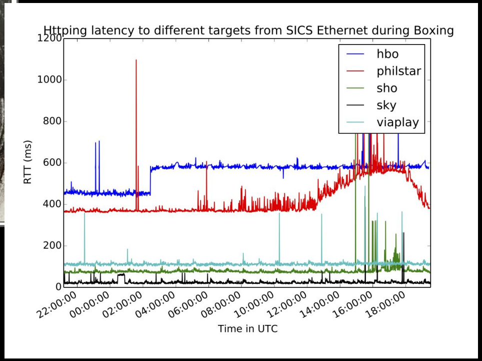
Can we as a community do something?

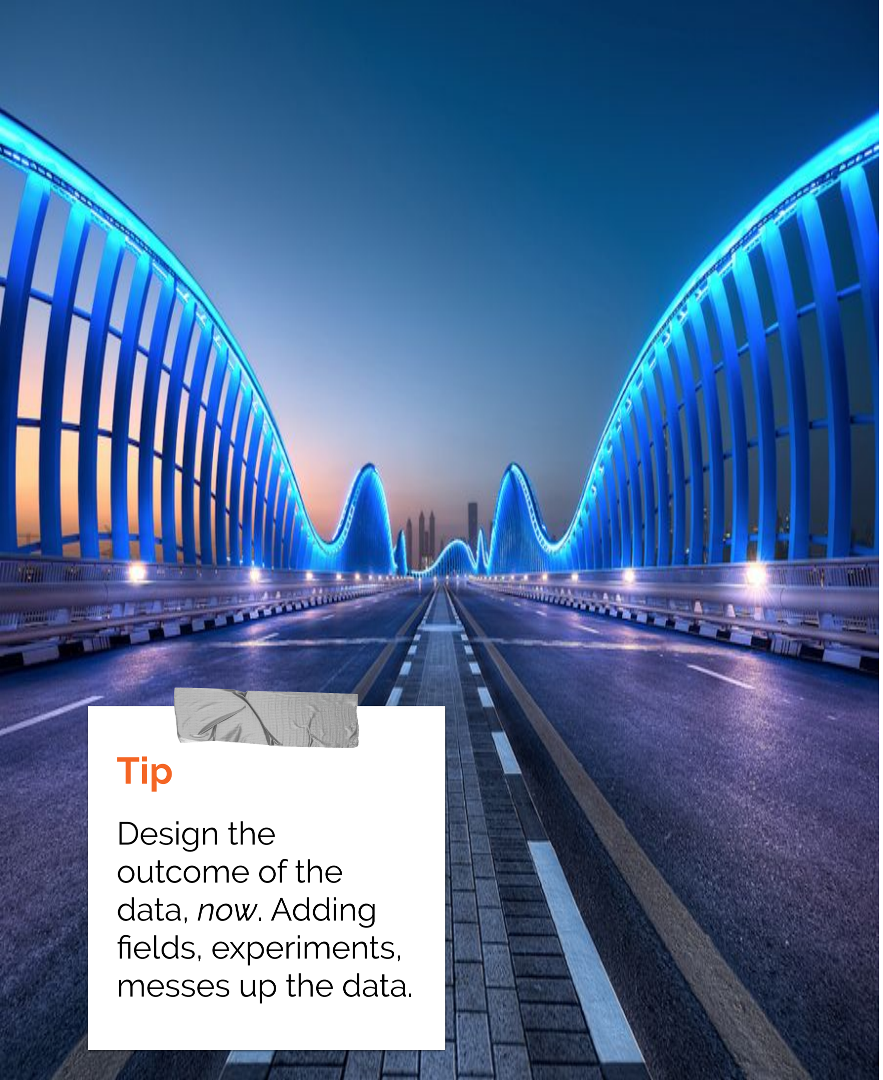
By measuring *colloboratively*, better network inference can be drawn
Where the problems occur, which can’t be reached from the outside

However measurements **must** be fair, unbiased, neutral, …







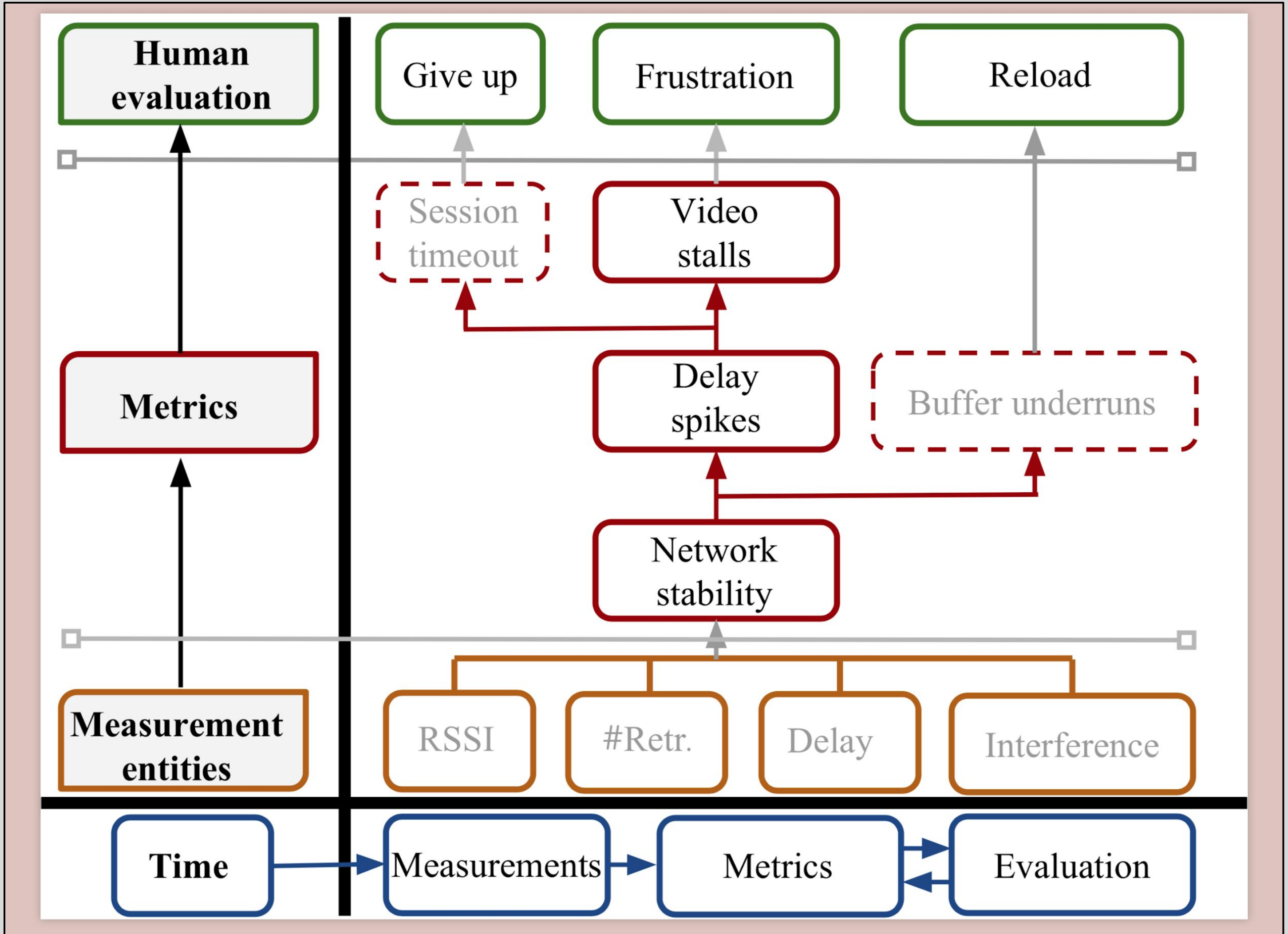


Tip

Design the outcome of the data, *now*. Adding fields, experiments, messes up the data.

Future.

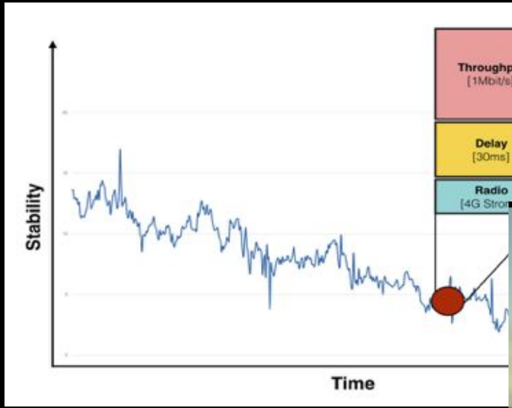
Measurements + Analysis = Pipeline

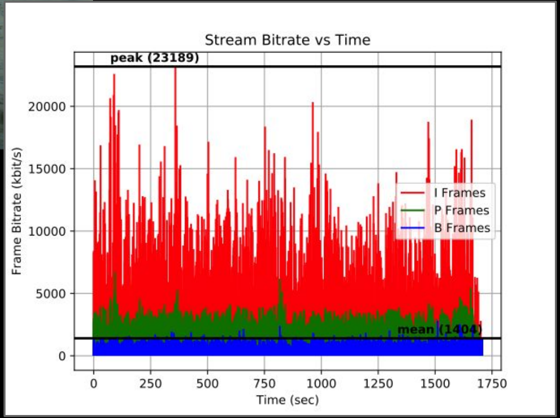


Metrics

1. Build blocks, not systems
2. Reuse each
3. And compose for each application

Video buffering ~ still broken





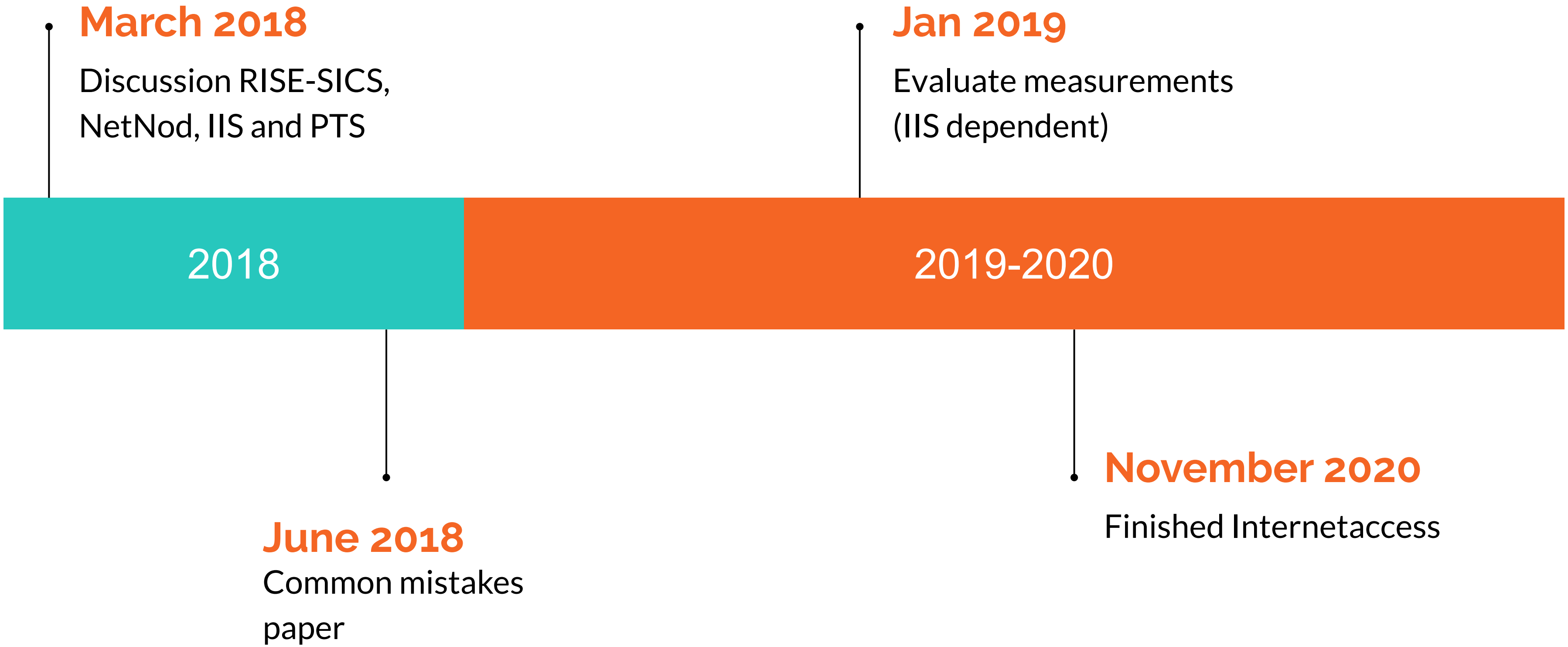
Internet access project

The project consists of three phases:

1. Provide a clear definition of what constitutes Internet ccess
2. Develop a measurement tool that customers and operators can use to measure Internet access.
3. Introduce a self-certification that may be used by operators and public procurement entities.

https://www.netnod.se/internetaccess

Milestones



2018

2019-2020

March 2018

Discussion RISE-SICS, NetNod, IIS and PTS

June 2018

Common mistakes paper

Jan 2019

Evaluate measurements (IIS dependent)

November 2020

Finished Internetaccess

European colloboration

“Common mistakes in measurements”

A (white) paper, will be summarised at one meeting.

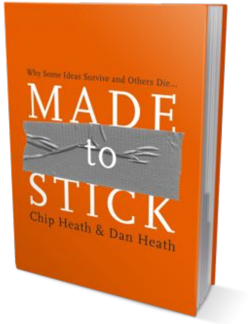
Way ahead

1. Summarise experiences over 10 campaigns
2. In a report
3. Data analysis included ☺

Dimension reduction

1. Reduce explicit and implicit artifacts
2. Makes visualisation more effectibe
3. ML performs better
4. Data exploration more intuitive

Theme stolen from Google docs.



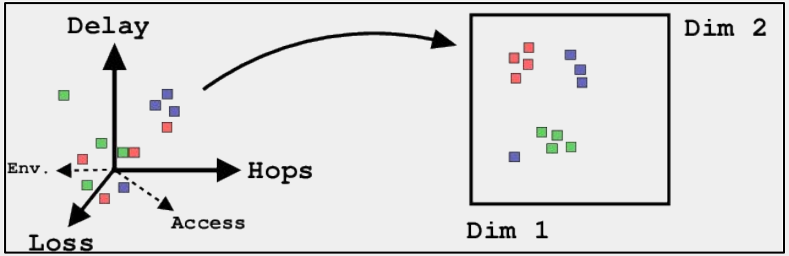
Dimensionality reduction

1. Modern network measurements may contain hard to discern facets.

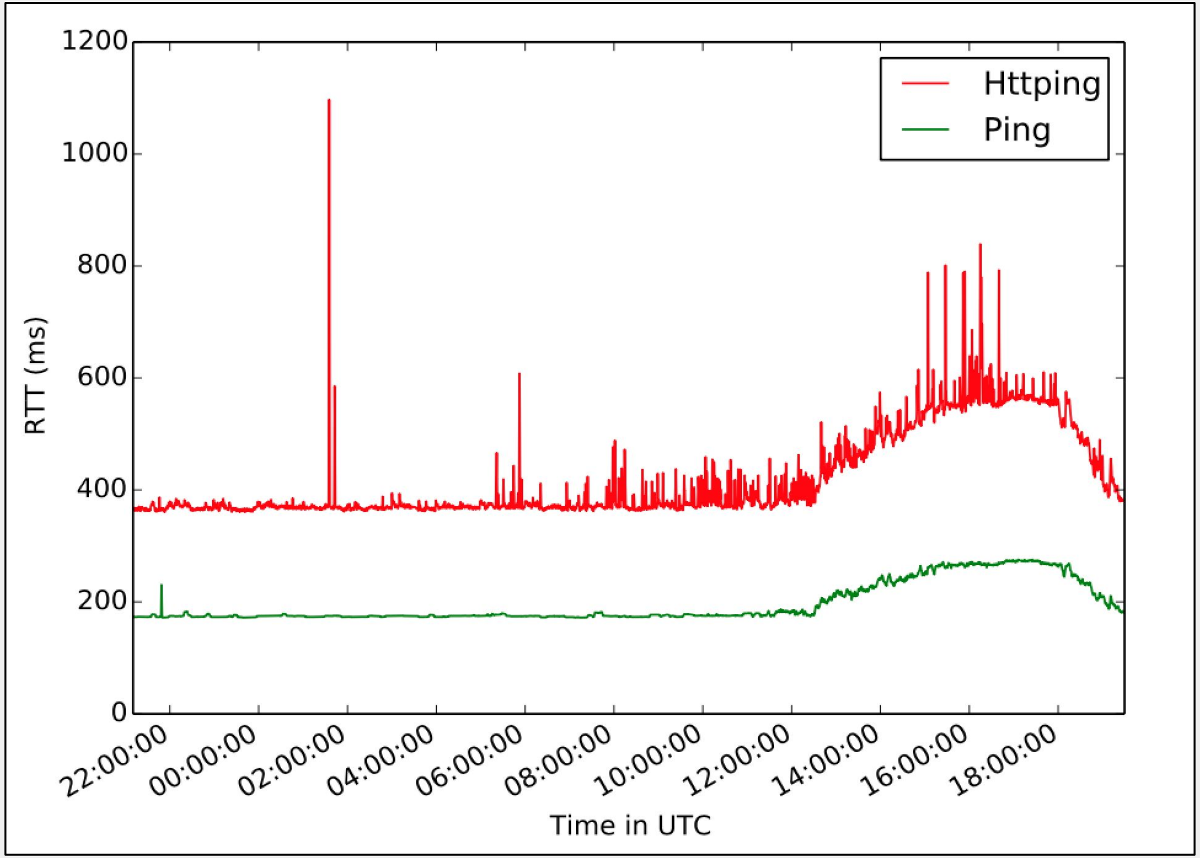
2. Metrics may be derived from basic, measurable entities, for example, quality might be a function of the delay, loss, environment, access technology…

3. System delays may be coupled, network load and server response times may be correlated.

4. Endpoints, moving users or link types affect feedback timers



Network and server delays



Example - coupled delays

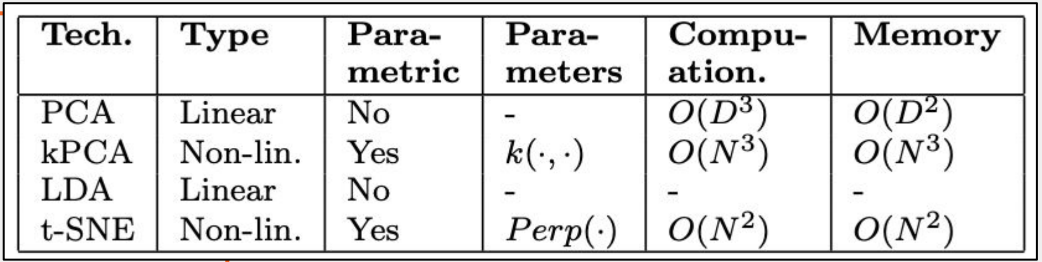
Delays arising from a network and a server interact. A loaded network will result in longer response times from a server, and a busy server will produce longer latency for the network.

From an external measurement perspective these delays might be indistinguishable and change in contribution over time. Systems theory, stability and coupled systems have a rich engineering history.

Clearly the average delay is due to the network, however the variance in the delay is due to the server response. To quantify the contributions of the network and server, therefore visualisations become important.

Techniques

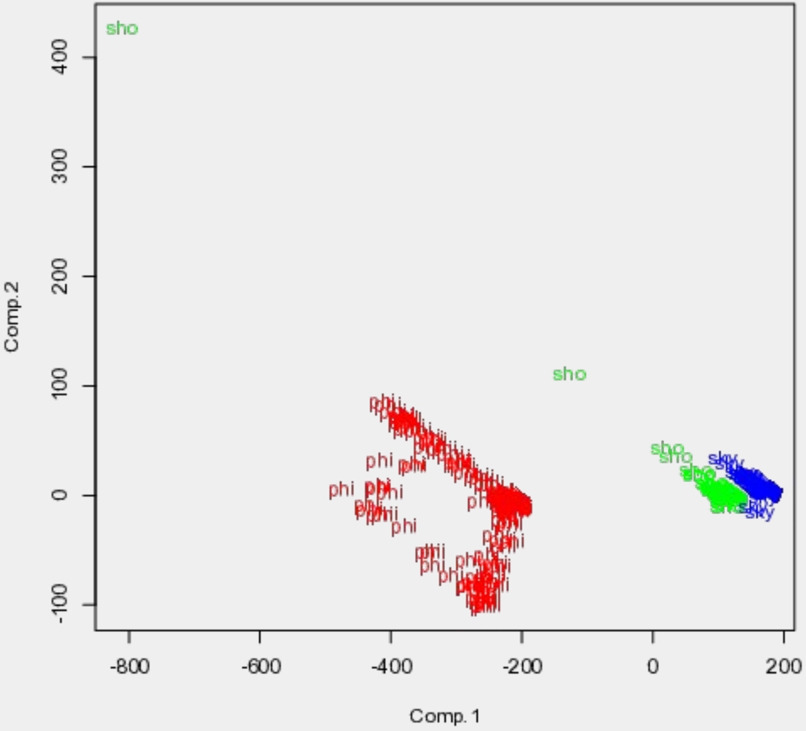
1. PCA
2. kPCA
3. LDA
4. t-SNE



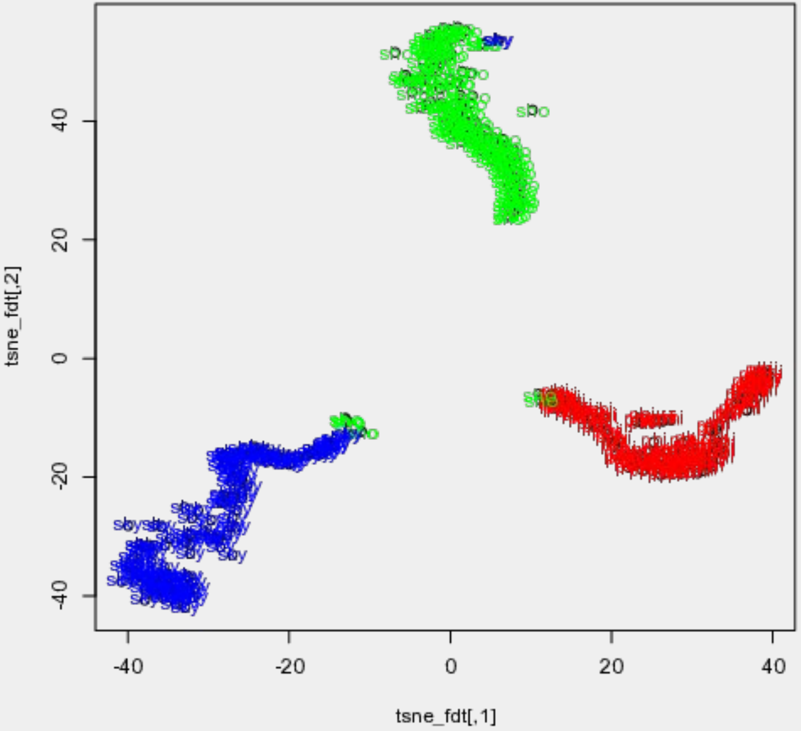
Tech.	Type	Parametric	Parameters	Computation.	Memory
PCA	Linear	No	-	$O(D^3)$	$O(D^2)$
kPCA	Non-lin.	Yes	$k(\cdot,\cdot)$	$O(N^3)$	$O(N^3)$
LDA	Linear	No	-	-	-
t-SNE	Non-lin.	Yes	$Perp(\cdot)$	$O(N^2)$	$O(N^2)$

We chose PCA as it is ubiquitous in dimensionality reduction, it is computationally efficient, linear and parameterless. It also has many variants, kernel, probablistic, discriminet. Kernel PCA performs analysis in the high dimensional space, using a kernel function, to find the principal components, see Table. LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. t-SNE is attractive due to its separation and visualization properties with recent improvements in its performance, examples next

Visualisation (coupled network and server delays)

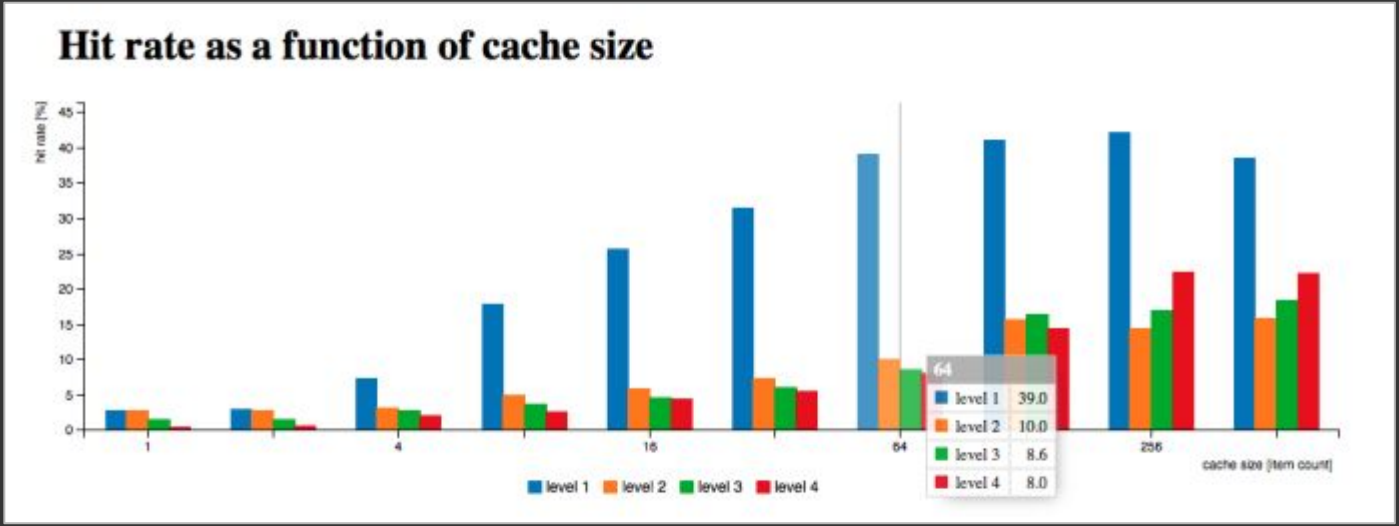


PCA



t-SNE

Caching analysis
experimental design, ...

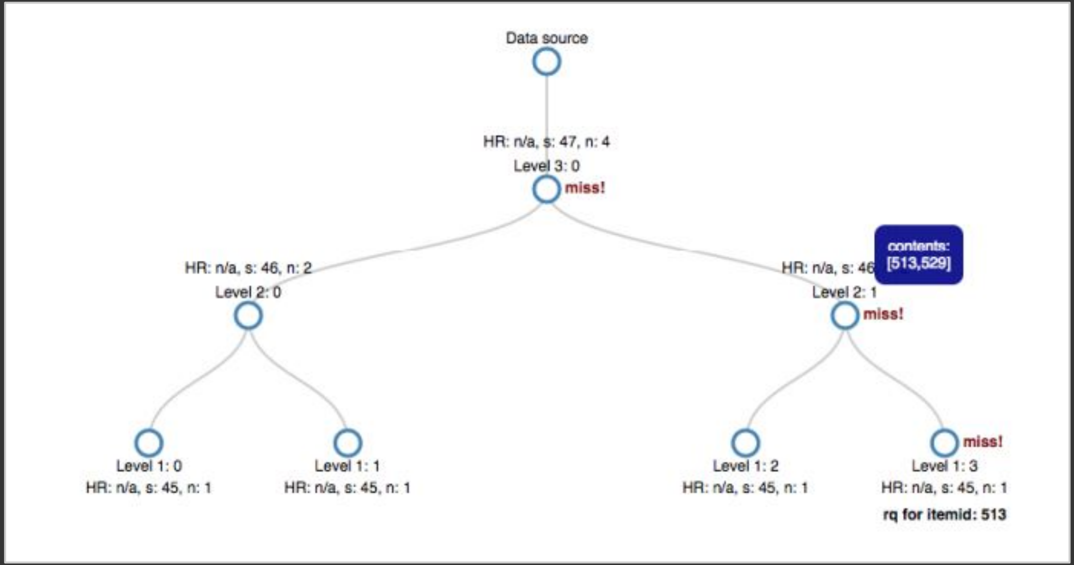


Issues

1. Complex
2. Dynamic
3. Policy-driven
4. Expensive (storage)
5. Poor quality (underdimensioned)

http://www.diva-portal.org/smash/get/diva2:861210/FULLTEXT01.pdf

Caching analysis
Visualisation



Solution

1. Inspection
2. Insight
3. Hard to do otherwise

Related topics

1. Data Readiness (TRL for data)
2. Sharing data incentives
3. Transparent AI
4. Data analysis landscape

Available from https://ianmarsh.org