Experimental Design and Analysis

IN5060: Quantitative Performance Analysis

UiO : University of Oslo

How to

- Design a proper set of experiments for measurement.
- Estimate the contribution of each alternative to the performance.
- Check if the alternatives are significantly different.
- Isolate the measurement errors.
- Develop a model that best describes the data obtained.
- Check if the model is adequate.

Example

- Personal workstation design
  1. Processor: AMD Ryzen 3,5,9 or Intel Core i5, i7, i9.
  2. RAM size: 8G, 16GB, or 32GB bytes
  3. Storage: 512GB, 1TB, 2TB
  4. Workload: Gaming, managerial, or scientific.

Terminology

- **Response Variable**: Outcome. E.g., throughput, response time
- **Factors (predictors)**: Variables that affect the response variable. E.g., CPU type, RAM size, etc…
- **Levels (treatment)**: The values that a factor can assume, E.g., RAM size has three levels: 8G, 16GB, or 32GB bytes
- **Primary Factors**: The factors whose effects need to be quantified. E.g., CPU type, RAM size only, etc…
- **Secondary Factors:** Factors whose impact need not be quantified. E.g., the workloads.
- **Replication:** Repetition of all or some experiments.

Experimental Design

The number of experiments, the factor level and number of **replications** for each experiment.

- Simple Designs: Vary one factor at a time

$$1 + \sum_{i=1}^{k} (n_i - 1)$$

  - Not statistically efficient.
  - Wrong conclusions if the factors have interaction.

- Full Factorial Design: All combinations

$$\prod_{i=1}^{k} n_i$$

  - Can find the effect of all factors.
  - Too much time and money.

How to choose the factors?

- Domain knowledge is needed to define the primary/secondary factors
- System limitations: controllable parameters, number of repetitions, etc…

- Assume you want to measure the performance of a oprational 4G network, what factors can you control?

Statistical Significance

Dunn's test is used to identify significant difference between the means of two or more distributions for a given confidence interval

- ***Null Hypothesis:*** There is no significant difference between the two distributions
- ***Alternative Hypothesis:*** There is significant difference between the two distributions.
- ***Calculating the p-value****:* The p-value associated with each pairwise comparison is a measure of the statistical significance of the difference between the groups being compared. It tells you whether the observed difference between the groups is likely to be due to random chance or if it is statistically significant.

Statistical Significance

Interpretating the Dunn’s test

- If the p-value is less than a specified significance level (α) (usually 0.05, 0.01 or 0.001), you can declare the difference to be statistically significant and reject the test's null hypothesis.
- If the p-value is greater than the significance level, you fail to reject the null hypothesis, indicating that there is insufficient evidence to conclude a significant difference between the groups.

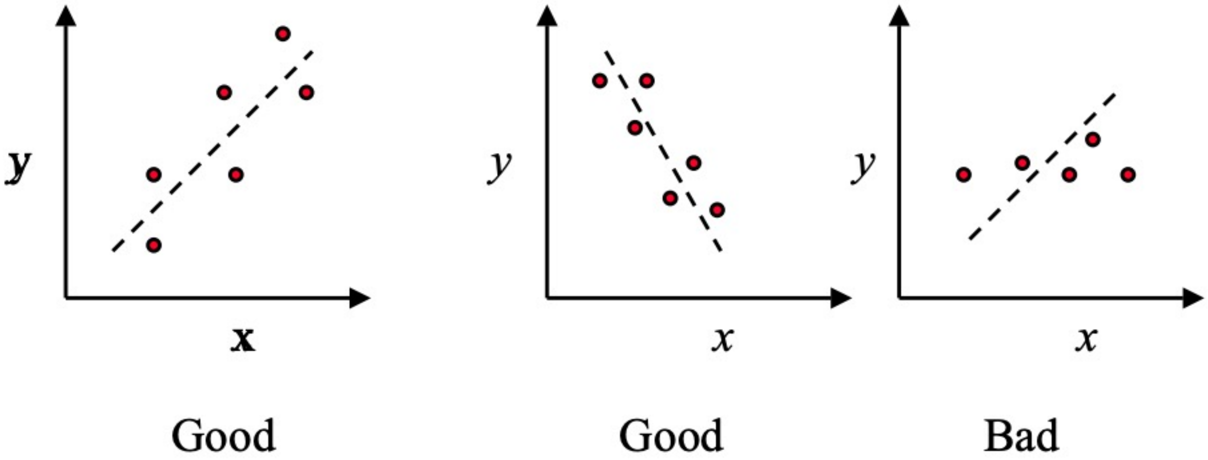
- The lower the p-value, the more significant the difference is!

Modelling: Simple Linear Regression

- Regression Model: Predict a response for a given set of predictor variables.
- Response Variable: Estimated variable
- Predictor Variables: Variables used to predict the response. predictors or factors
- Linear Regression Models: Response is a linear function of predictors.
- Simple Linear Regression Models: Only one predictor

What is a good model?

- Check visiually first how your data looks



A simple good model

Choose the line that minimizes the sum of squares of the errors

$$\hat{y} = b_0 + b_1 x$$

where, $\hat{y}$ is the predicted response when the predictor variable is x.

- The parameter b_0 and b_1 are fixed regression parameters to be determined from the data.
- Given n observation pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the estimated response for the ith observation is:

$$\hat{y}_i = b_0 + b_1 x_i$$

Minimizing the error

- The error is:

$$e_i = y_i - \hat{y}_i$$

- The best linear model minimizes the sum of squared errors (SSE):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Common Mistakes in Experimentation

- The variation due to experimental error is ignored.
- Important parameters are not controlled.
- Effects of different factors are not isolated
- Simple one-factor-at-a-time designs are used
- Interactions are ignored
- Too many experiments are conducted

- Remember the noise in real world measurements! We need STATISTICS ☺

Quantitative Performance Analysis

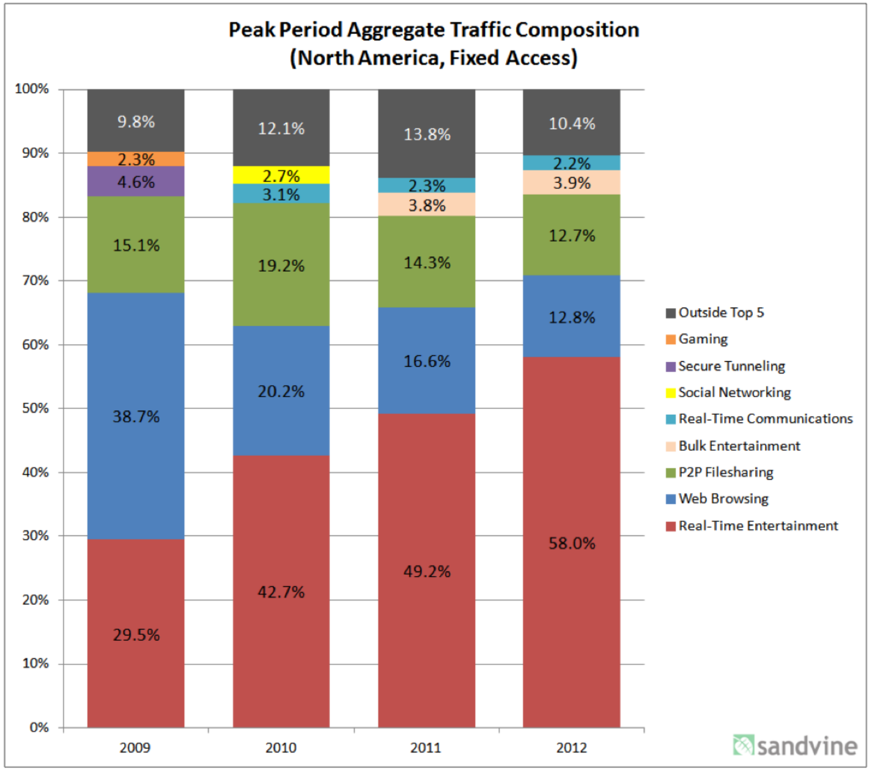
Real-world examples

Measurement example

Development of traffic shares over time

A graph using percentages to express the share of application types on the Internet

- no absolute values, only percentages
- color as well as order allows easy recognition of types, as well as appearance of new types

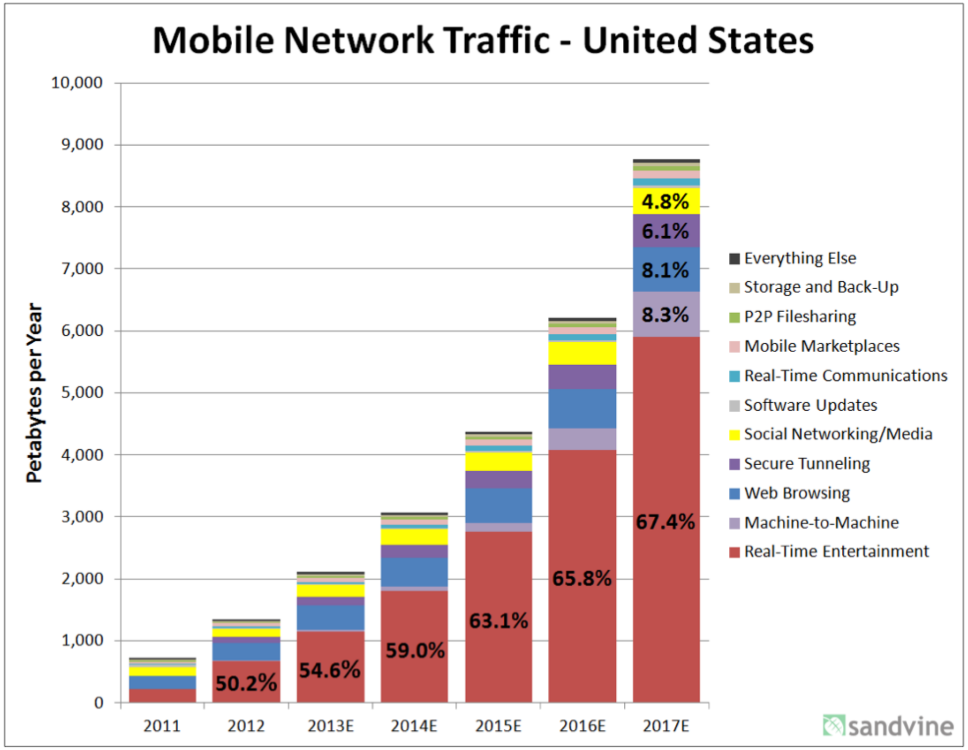


Measurement example

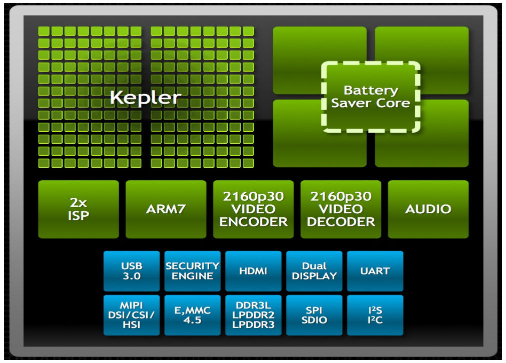
Development of absolute mobile traffic over time

A graph using absolute value to communicate the rapid growth of mobile traffic

- percentages provided as text in graph
- color as well as order allows easy recognition of types, as well as appearance of new types
- note ”E” for estimates

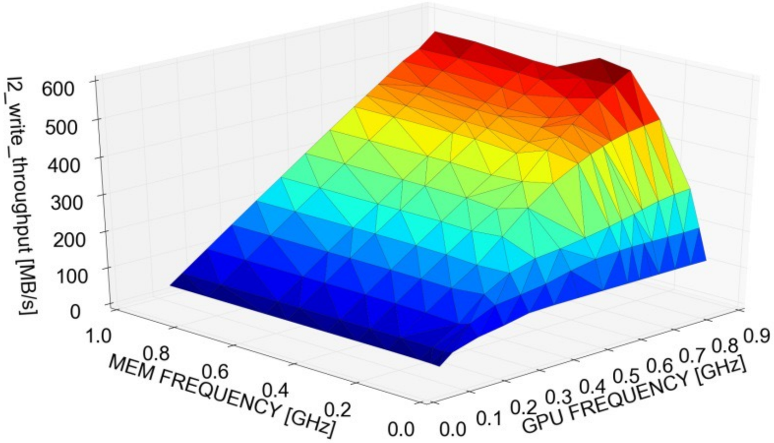


Measurement example



NVidia Tegra K1

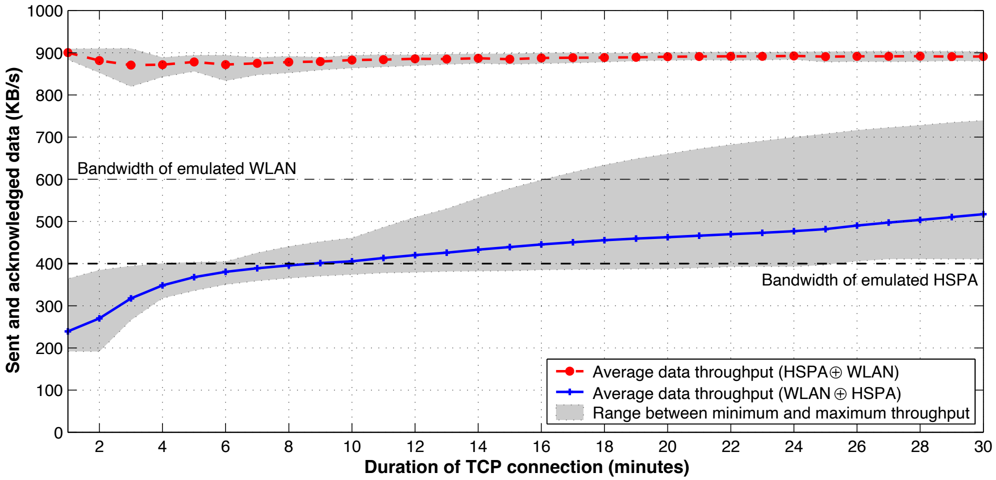
impact of frequency on throughput



- note: *4 dimensions* in the presentation
- additional dimension can be used to add information or to add expressiveness to one or more of the dimensions

Measurement example

Linux TCP's ability to recover from out-of-order delivery of packets



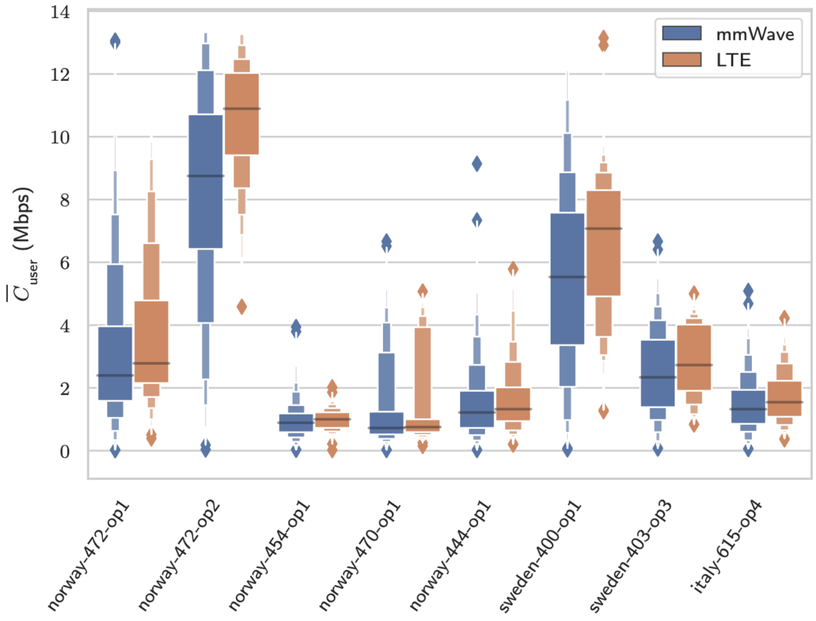
- 2D graph studying values for the average bandwidth of very long-lived TCP flows whose packets are alternately sent over 2 very different paths

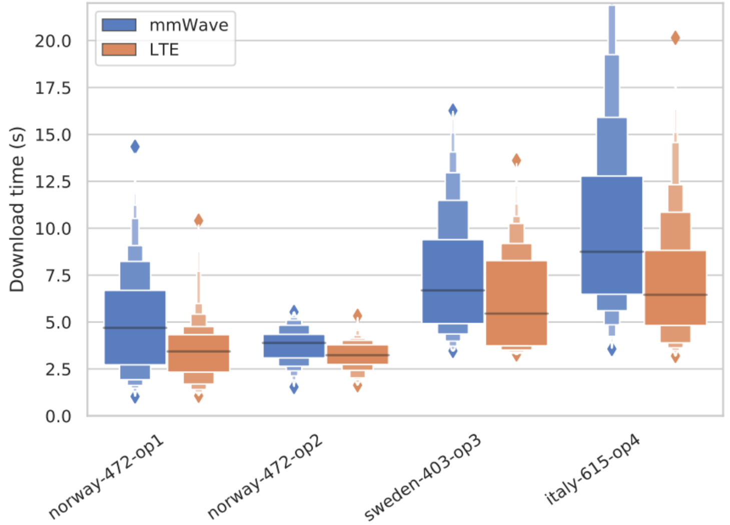
- details of short-term TCP behaviour are completely hidden
- smoothness achieved by averaging
- *shaded areas illustrate uncertainty* (range from min to max average throughput)

Measurement example

Investigation of the current operator SPEPs (used in 4G LTE networks) under mmWave-like dynamics using MONROE testbed

- Different colors highlight the different settings
- Boxplots display the first two letter values (the median and quartiles); letter-value plots display further letter values therefore more suitable for large datasets.

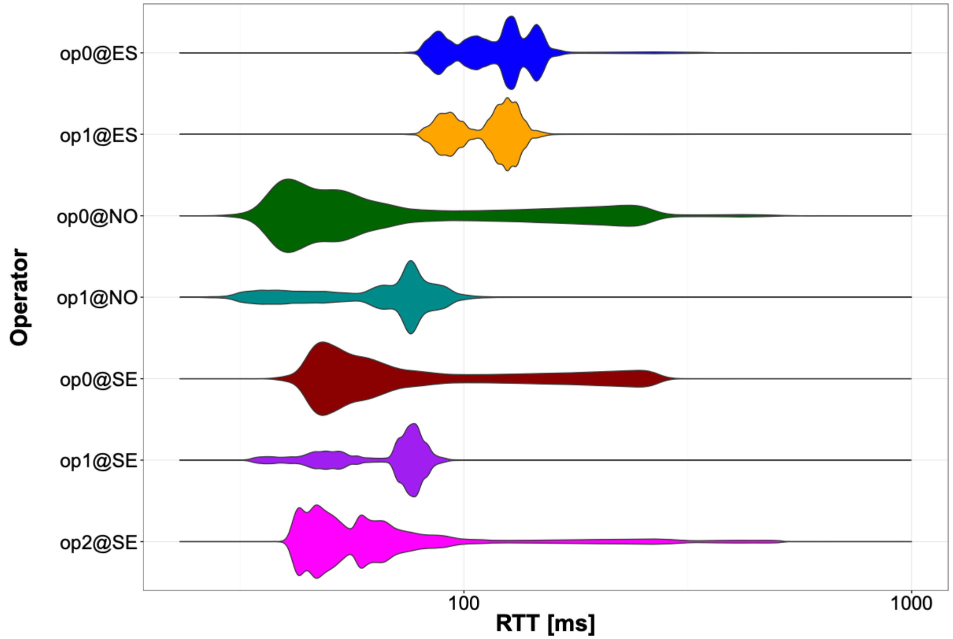




Measurement example

Investigation of latency for 4G LTE networks using MONROE testbed

- Different colors highlight the different settings
- Violin plots are similar to box plots, except that they also show the probability density of the data at different values.

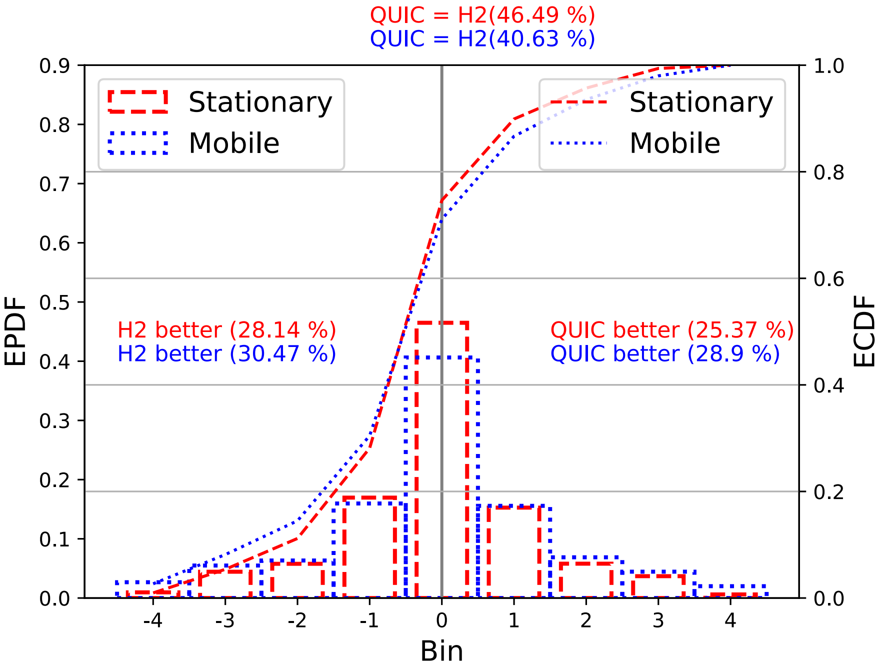


Measurement example

Performance comparison of HTTP2 and QUIC over 4G networks under stationary and mobile scenarios

- Color together with pattern is used to diffentiate different settings

- Empirical PDF is presented together with empirical CDF
- The bins are used to illustrate the relative difference levels

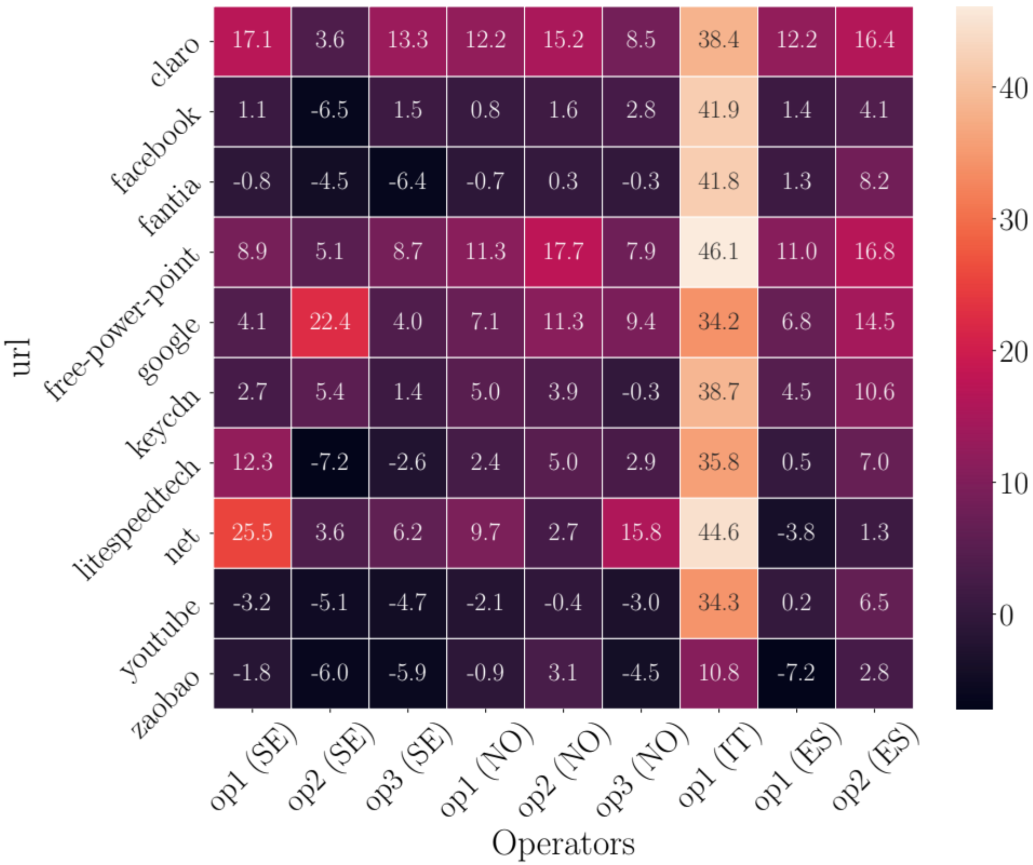


Measurement example

The impact of COVID-19 on the web performance over mobile networks

- 2D matrix form together with heatmap
- The heatmap together with values in it provide more information

- highly aggregated data
- concept of certainty (e.g. confidence intervals) gets lost

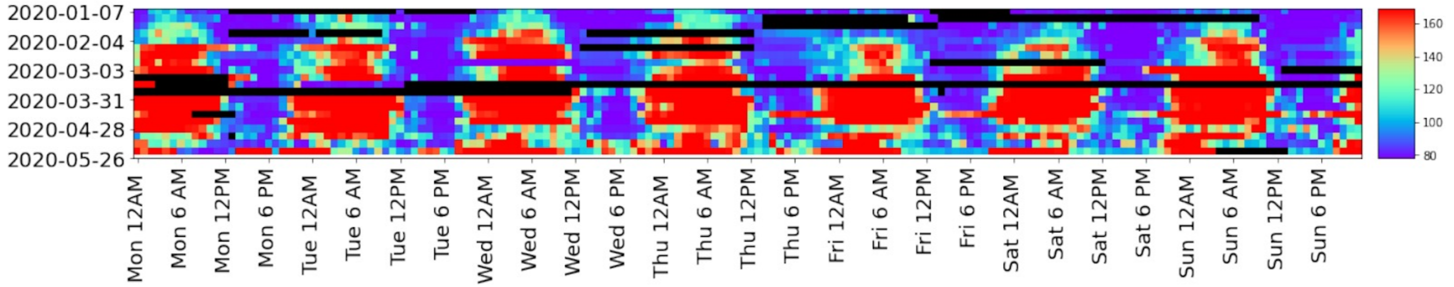


url	op1 (SE)	op2 (SE)	op3 (SE)	op1 (NO)	op2 (NO)	op3 (NO)	op1 (IT)	op1 (ES)	op2 (ES)
claro	17.1	3.6	13.3	12.2	15.2	8.5	38.4	12.2	16.4
facebook	1.1	-6.5	1.5	0.8	1.6	2.8	41.9	1.4	4.1
fantia	-0.8	-4.5	-6.4	-0.7	0.3	-0.3	41.8	1.3	8.2
free-power-point	8.9	5.1	8.7	11.3	17.7	7.9	46.1	11.0	16.8
google	4.1	22.4	4.0	7.1	11.3	9.4	34.2	6.8	14.5
keycdn	2.7	5.4	1.4	5.0	3.9	-0.3	38.7	4.5	10.6
litespeedtech	12.3	-7.2	-2.6	2.4	5.0	2.9	35.8	0.5	7.0
net	25.5	3.6	6.2	9.7	2.7	15.8	44.6	-3.8	1.3
youtube	-3.2	-5.1	-4.7	-2.1	-0.4	-3.0	34.3	0.2	6.5
zaobao	-1.8	-6.0	-5.9	-0.9	3.1	-4.5	10.8	-7.2	2.8

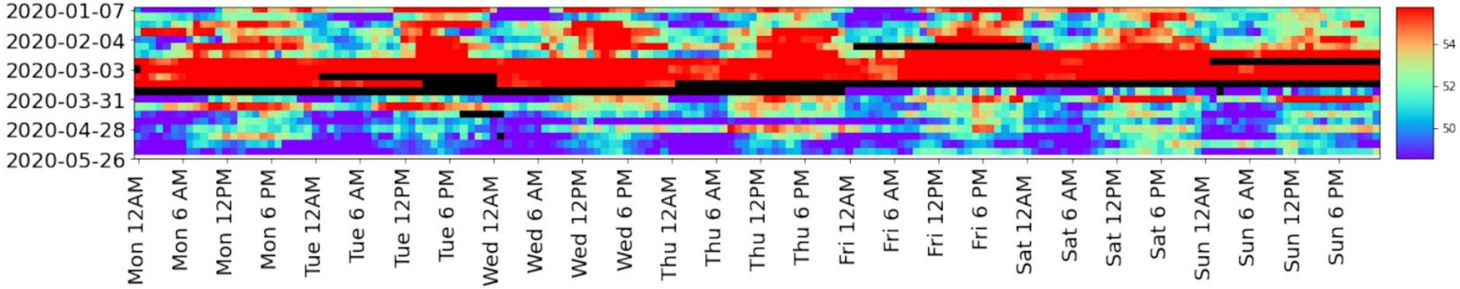
Figure 1: Change in PLT(%) from February to March.

Measurement example

The impact of COVID-19 on the performance of mobile networks



(a) RTT [ms] for op1 (IT).



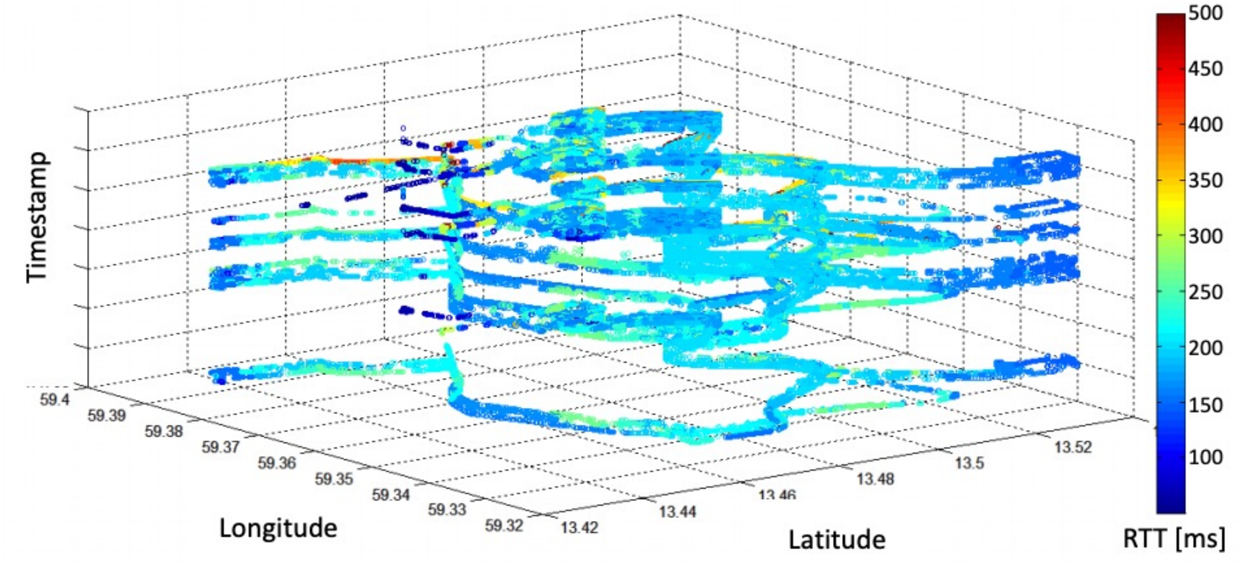
(c) RTT [ms] for op1 (SE).

- 2D information (date vs time of the dat) together with heatmap

- Diurnal pattern comparison
- Visual patterns are visible, but details are hidden

Measurement example

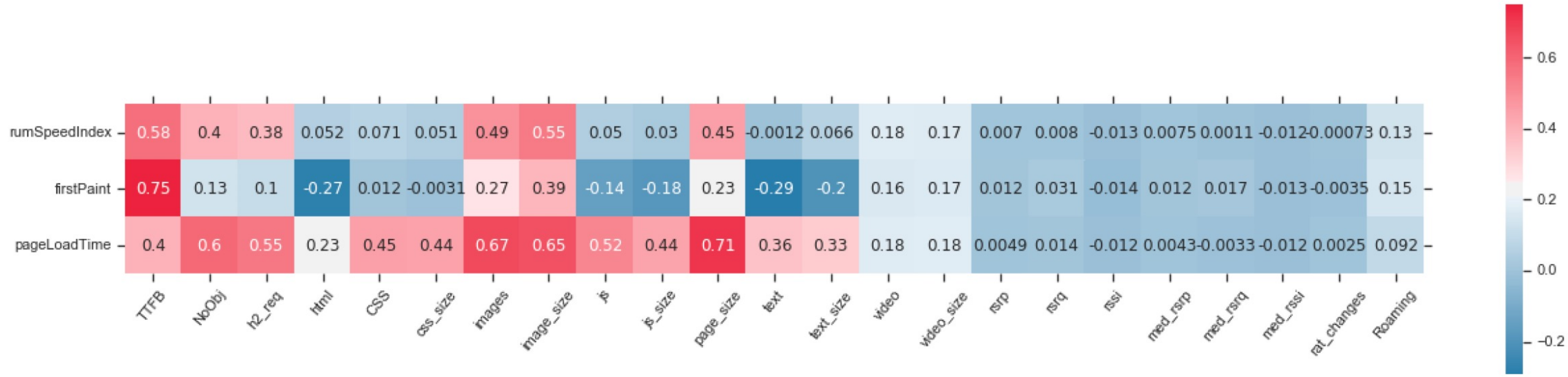
Investigation of latency over multiple trips under mobility for mobile networks



- 3D information together with heatmap
- Multiple laps are shown using the Y-axis offset based on relative timestamps to visually show the different trips
- Potentially have the map to visualize the location as well

Measurement example (with some ML flavour)

The impact of different browsing features on the web quality of experience



- Spearman’s correlation matrix
- Heatmap has negative and positive values