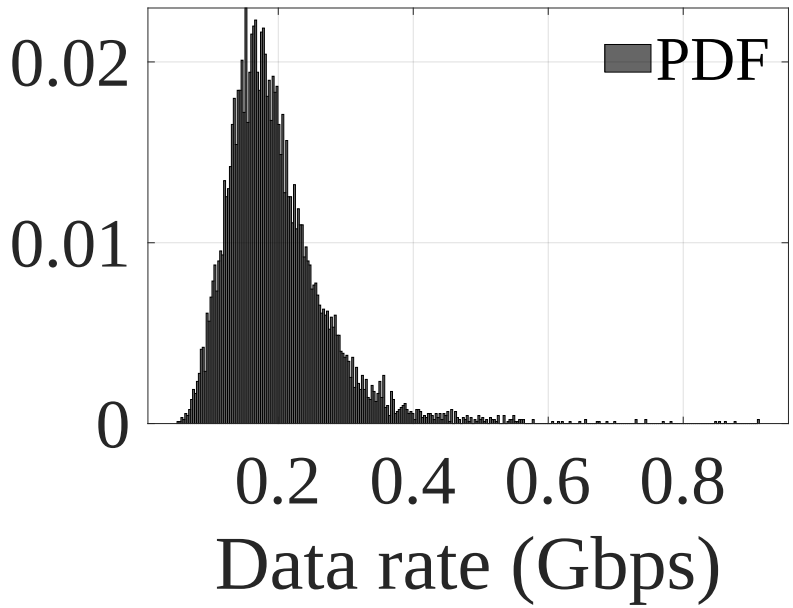
Internet Traffic Volumes Are Not Gaussian - They Are Log-Normal: An 18-Year Longitudinal Study With Implications for Modelling and Prediction



Presenter: Richard G. Clegg,
Authors: Mohammed Alasmar, Richard Clegg, Nickolay Zakhleniuk, George Parisis
(Prepared using LaTeX and beamer.)

Aims/hypothesis

Aim of this research

What is the best 'simple' statistical model of 'background traffic' on an Internet link that aggregates traffic from many users? Core question: 'How much traffic is there per second and how does it vary'?

Aims/hypothesis

Aim of this research

What is the best 'simple' statistical model of 'background traffic' on an Internet link that aggregates traffic from many users? Core question: 'How much traffic is there per second and how does it vary'?

Initial hypothesis

Measure traffic past a point on a wire for some 'small' unit of time (second or less?). Literature says this is normally distributed (or some claim a power law) – truth is between the two.

Aims/hypothesis

Aim of this research

What is the best 'simple' statistical model of 'background traffic' on an Internet link that aggregates traffic from many users? Core question: 'How much traffic is there per second and how does it vary'?

Initial hypothesis

Measure traffic past a point on a wire for some 'small' unit of time (second or less?). Literature says this is normally distributed (or some claim a power law) – truth is between the two.

Aim of this talk

How do you responsibly write a data analysis paper and prove a statistical result?

Frame the question properly

A precise question

Take some time series of Internet data from 'reasonably' sized links. Split it into time periods of length τ. Let X_i be the number of bytes in time period $t \in [\tau i, \tau(i+1))$. What is the distribution of X_i?
Note clopen set $t \in [t_1, t_2)$ meaning $t_1 \leq t < t_2$. You see this a lot in creating time series.

Data problems

How do we get a 'representative' set of data? Data from a reasonable number of different sources that are still 'typical' in some way.

Analysis problems

What length of data should we look at? What size of τ should we pick?

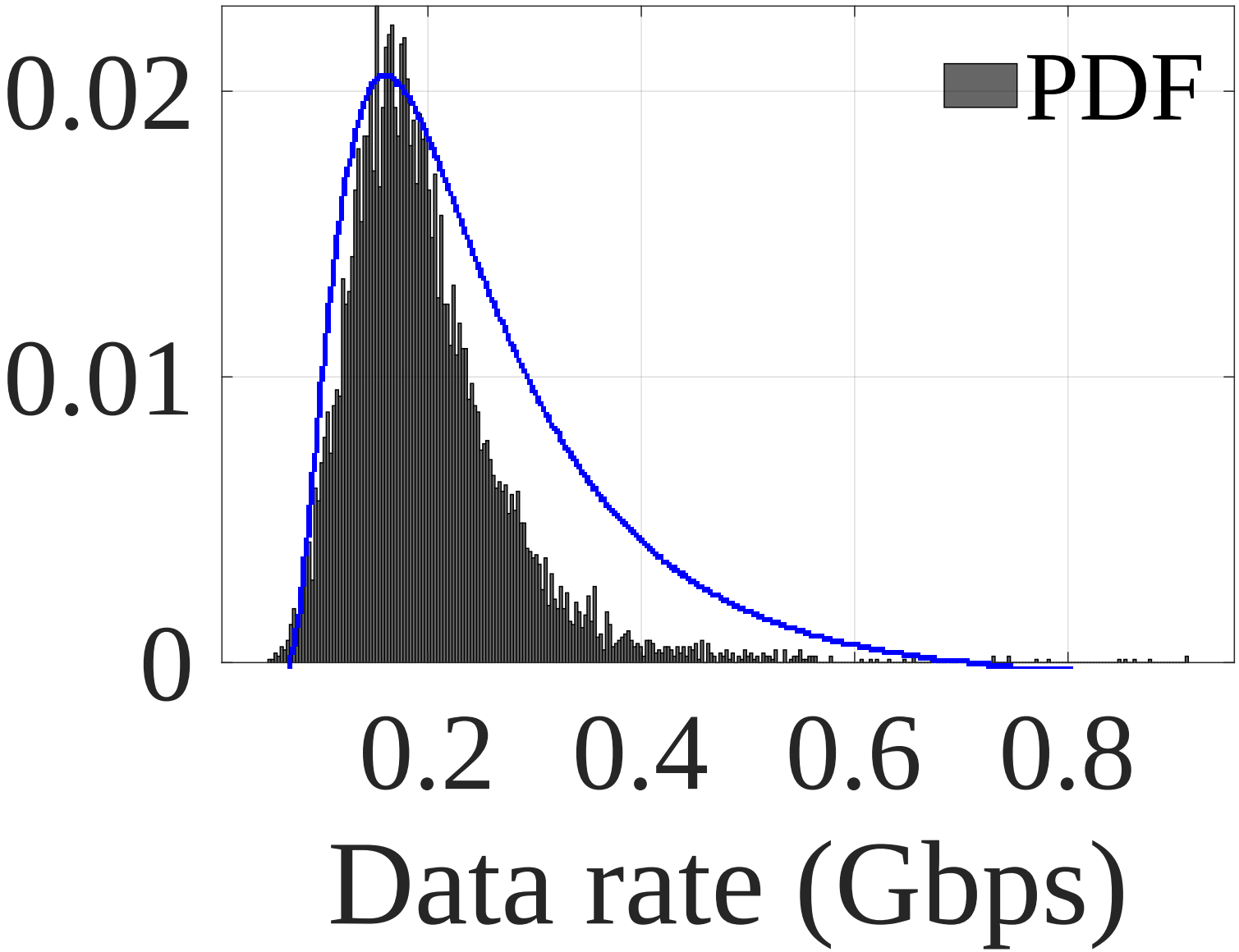
Data used

Data summary

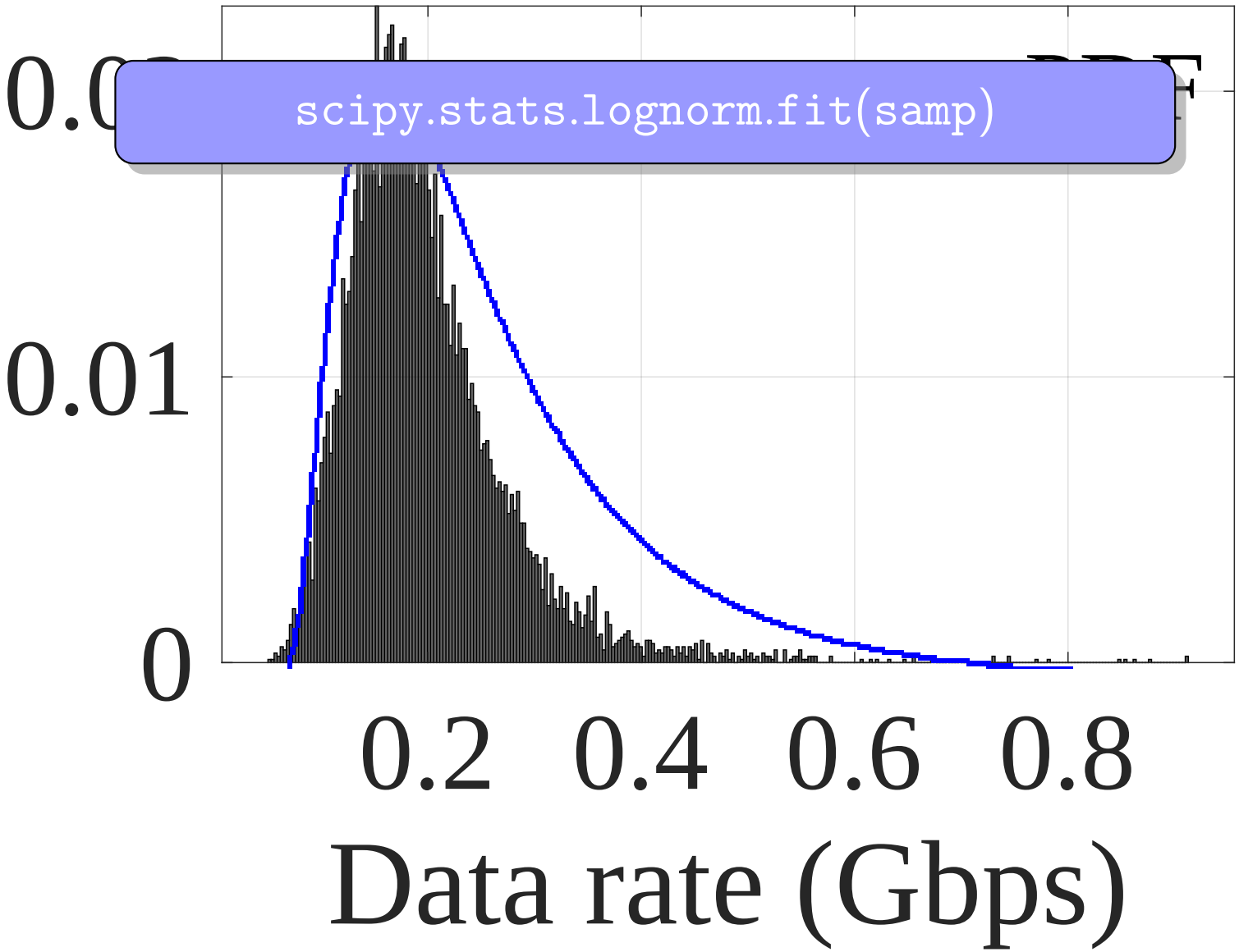
Total traces 232, spanning 18 years from various public data sets. Some residential, some academic, some commercial. It is very hard to avoid the referee comment 'your result is atypical because' in this type of study.

Trace	**Type**	**Number**	**year**
CAIDA	Tier 1 ISP (Chicago)	27	2013–2016
MAWI	Academic backbone (Japan)	110	2014–2020
Twente	Various inc residential	40	2003–2007
Waikato	Academic (New Zealand)	30	2011
Auckland	Academic (New Zealand)	25	2009

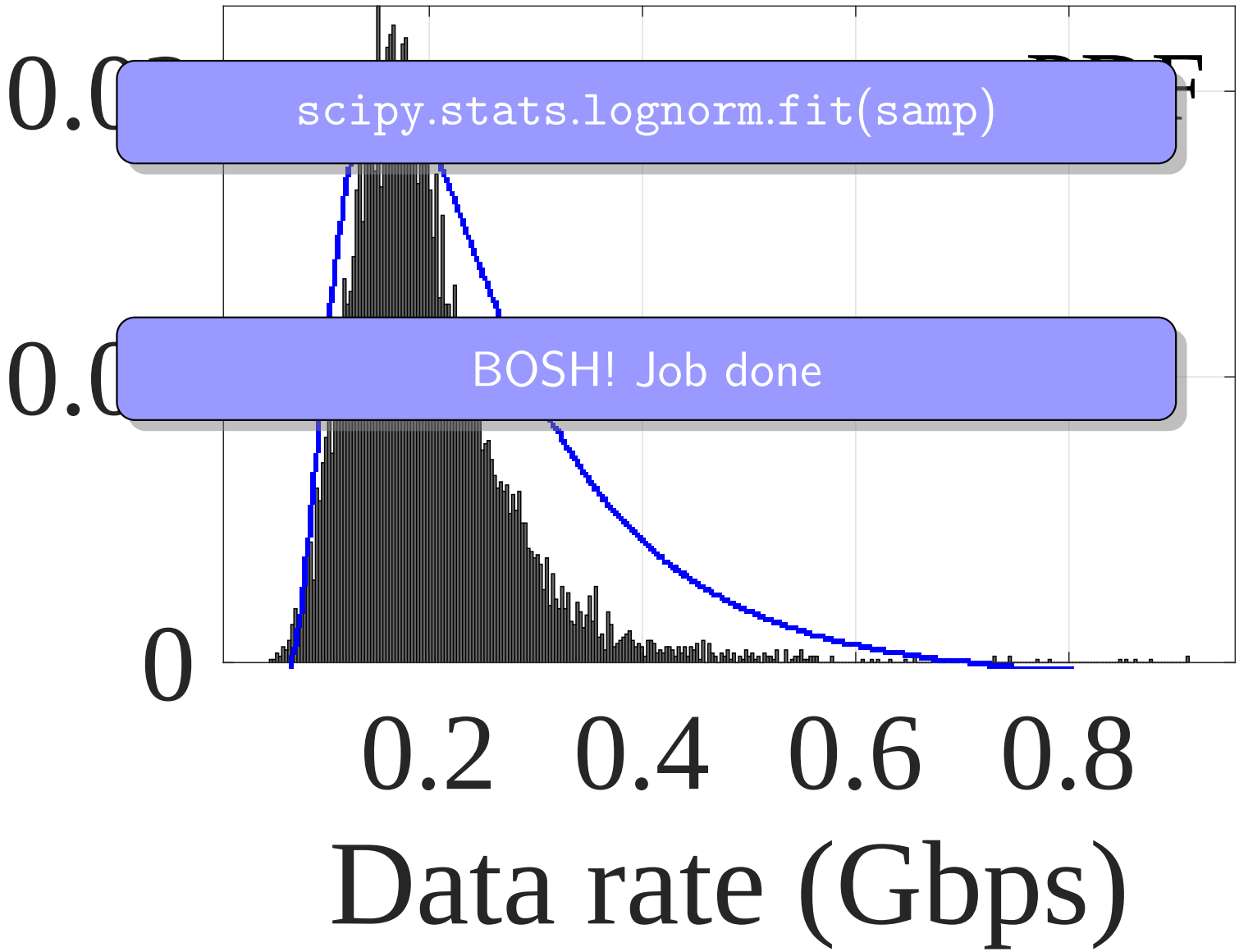
How to do it badly



How to do it badly



How to do it badly



How to do it badly

Length of data/stationarity

Stationarity, definition

Process X_t wide sense stationary iff:

$\mathrm{E}[X_1] = \mathrm{E}[X_2]$ mean does not vary in time.

$\mathrm{Cov}_X(t_1, t_2) = \mathrm{Cov}_X(0, t_2 - t_1)$ covariance depends on interval

$\mathrm{E}\left[(X_t)^2\right]$ second moment bounded.

Simplest way to think about it: Process looks kind of the same at whatever time of day you view it.

Length of data/stationarity

Stationarity, definition

Process X_t wide sense stationary iff:
$\mathrm{E}[X_1] = \mathrm{E}[X_2]$ mean does not vary in time.
$\mathrm{Cov}_X(t_1, t_2) = \mathrm{Cov}_X(0, t_2 - t_1)$ covariance depends on interval
$\mathrm{E}\left[(X_t)^2\right]$ second moment bounded.
Simplest way to think about it: Process looks kind of the same at whatever time of day you view it.

Why do we care?

Is distribution meaningful? Imagine traffic is tiny 25% of time, huge the rest. Fitted distribution is terrible at any time.

Length of data/stationarity

Stationarity, definition

Process X_t wide sense stationary iff:
$\mathrm{E}[X_1] = \mathrm{E}[X_2]$ mean does not vary in time.
$\mathrm{Cov}_X(t_1, t_2) = \mathrm{Cov}_X(0, t_2 - t_1)$ covariance depends on interval
$\mathrm{E}\left[(X_t)^2\right]$ second moment bounded.
Simplest way to think about it: Process looks kind of the same at whatever time of day you view it.

Why do we care?

Is distribution meaningful? Imagine traffic is tiny 25% of time, huge the rest. Fitted distribution is terrible at any time.

Is Internet traffic stationary?

God no! Traffic volumes are small at 5am for your local population but large when people are, say, streaming video or doing zoom calls. But maybe, if I take a small enough time period it is stationary.

OK you convinced me – how do I test stationarity?

Phillips–Perron test, Augmented Dickey Fuller test

Really easy to use robust tests. However you're testing that the data is non-stationary.
Failing to prove data is non-stationary is not the same as proving it is stationary.

Kwiatkowski–Phillips–Schmidt–Shin test

Tests the data is stationary. However, not robust to trends in the data.
If test is correct we get a p-value asserting the data is stationary but test is easily fooled and p-values are a pain to explain.

In short

The tools are easy to use but interpreting what they are saying takes experience. Call someone who has studied stats. Some of our data was proved to be both stationary and not-stationary (because p-values)!

OK you convinced me – how do I test stationarity?

Phillips–Perron test, Augmented Dickey Fuller test

Really easy to use robust tests. However you're testing that the data is non-stationary.
Failing to prove data is non-stationary is not the same as proving it is stationary.

Kwia

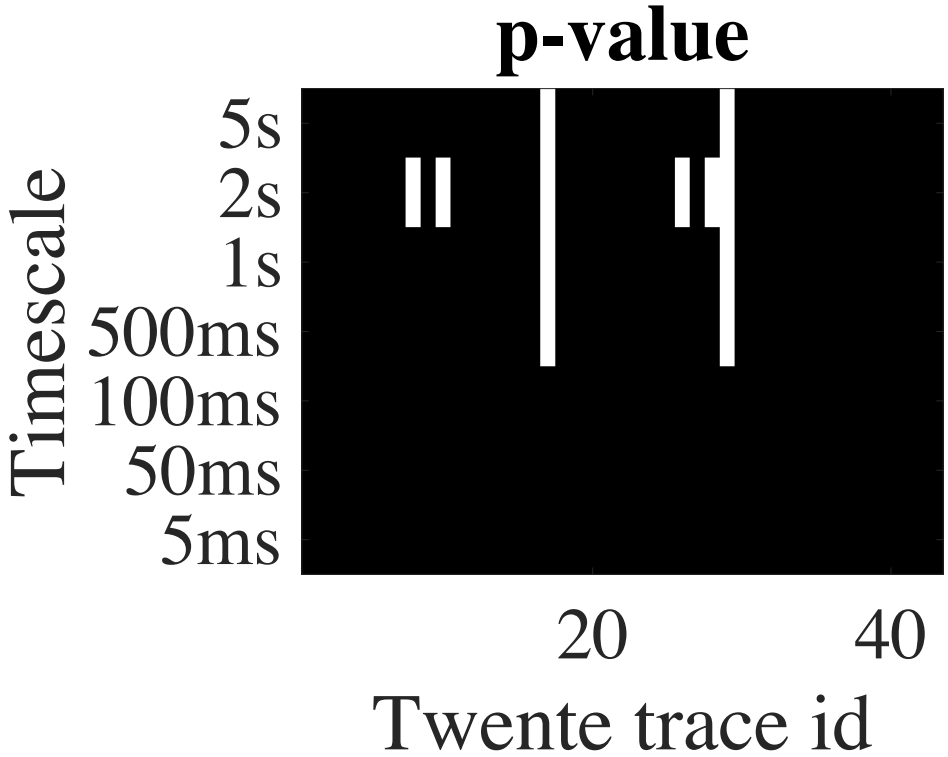
Statistics why you so mean?

Tests the data is stationary. However, not robust to trends in the data.
If test is correct we get a p-value asserting the data is stationary but test is easily fooled and p-values are a pain to explain.

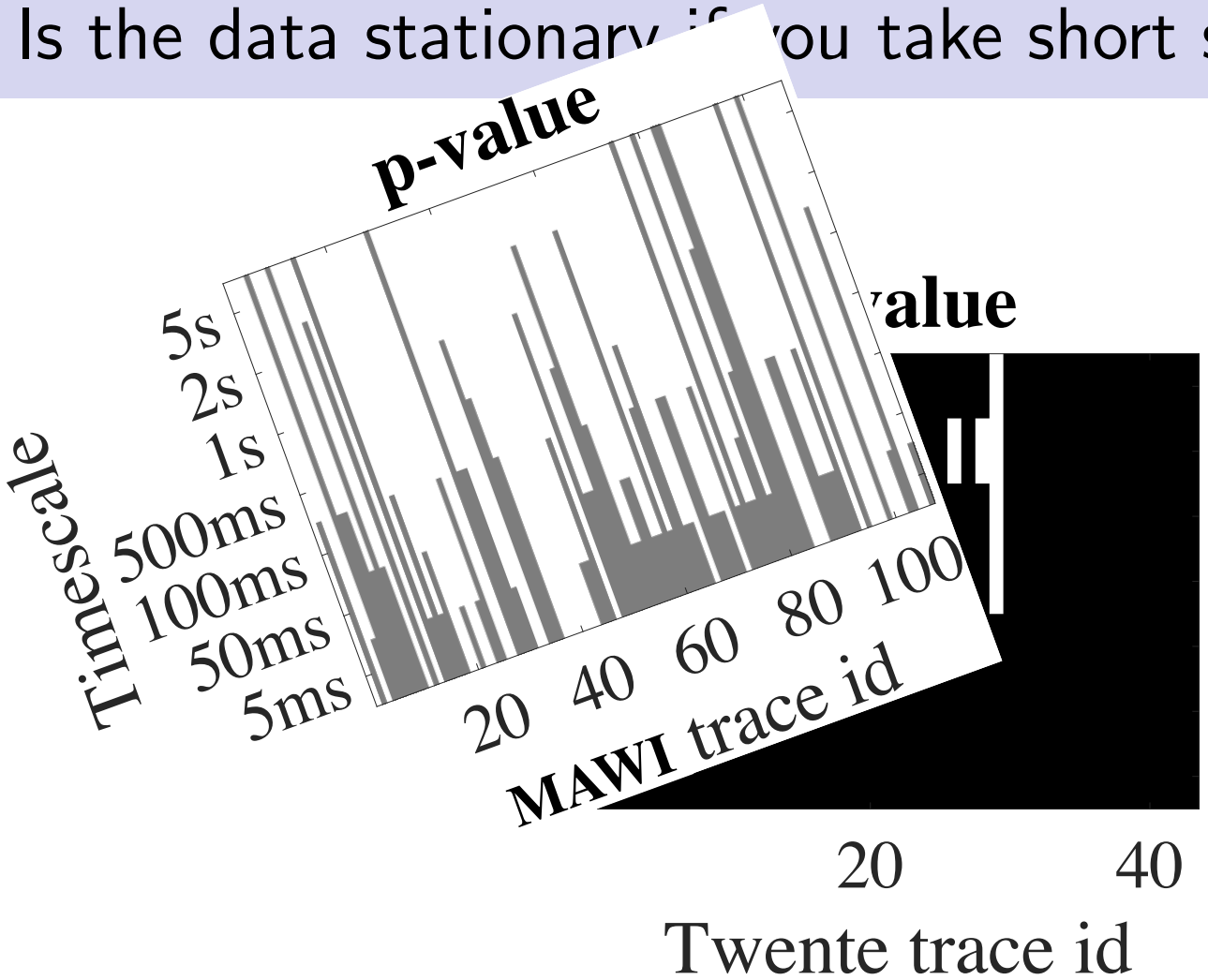
In short

The tools are easy to use but interpreting what they are saying takes experience. Call someone who has studied stats. Some of our data was proved to be both stationary and not-stationary (because p-values)!

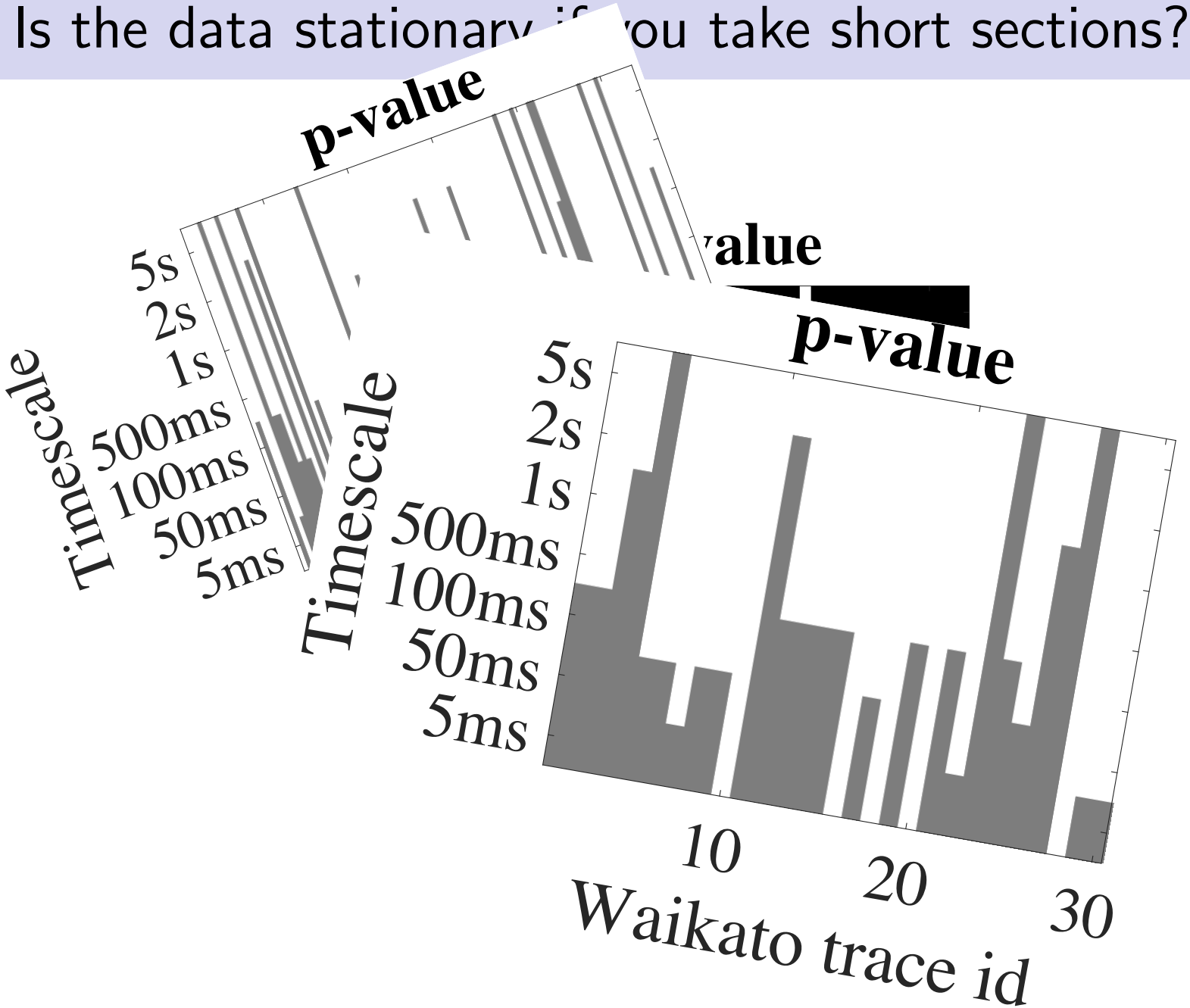
Is the data stationary if you take short sections?



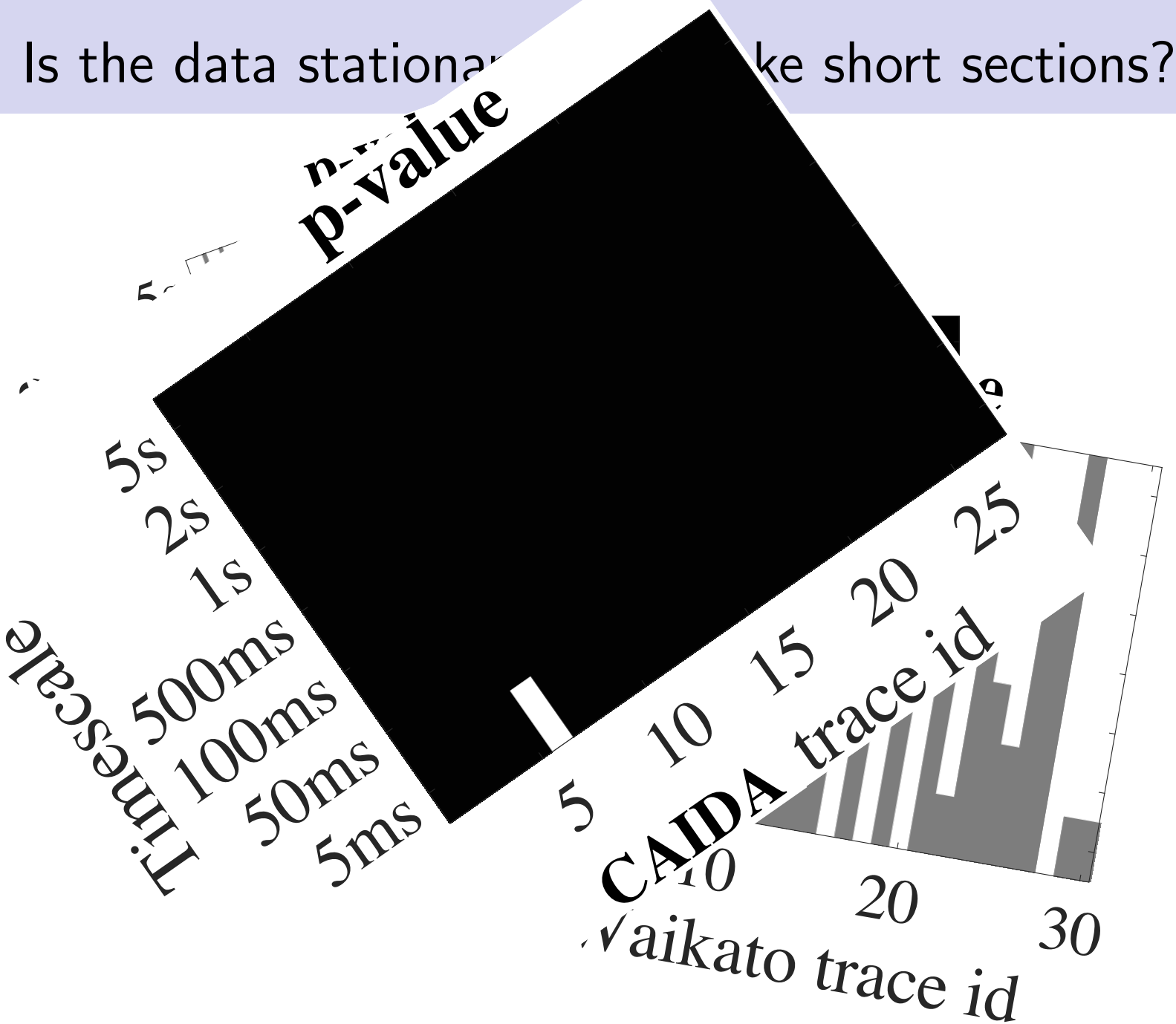
Is the data stationary if you take short sections?



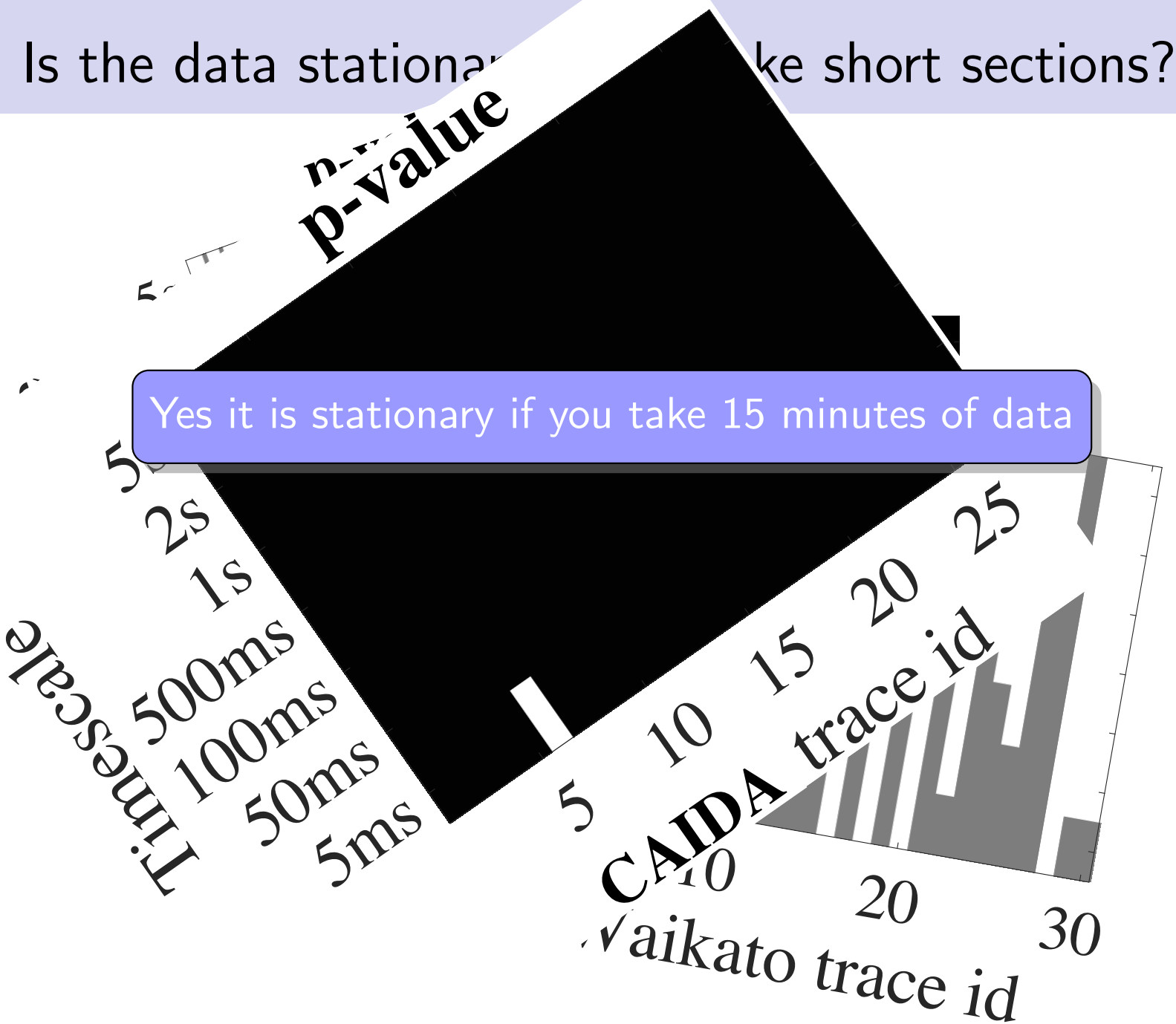
Is the data stationary if you take short sections?



Is the data stationar… ke short sections?

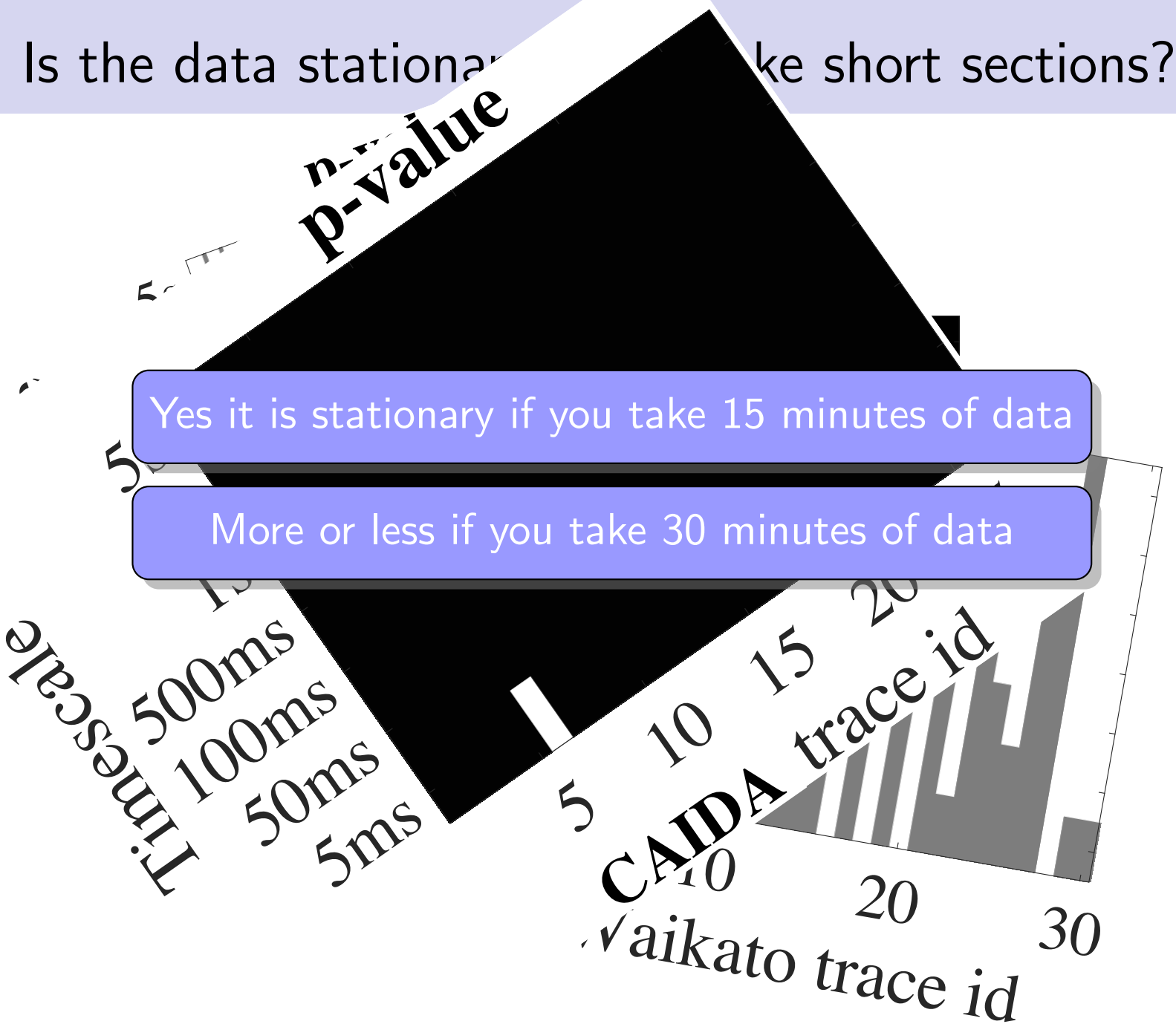


Is the data stationa[illegible]ke short sections?



Yes it is stationary if you take 15 minutes of data

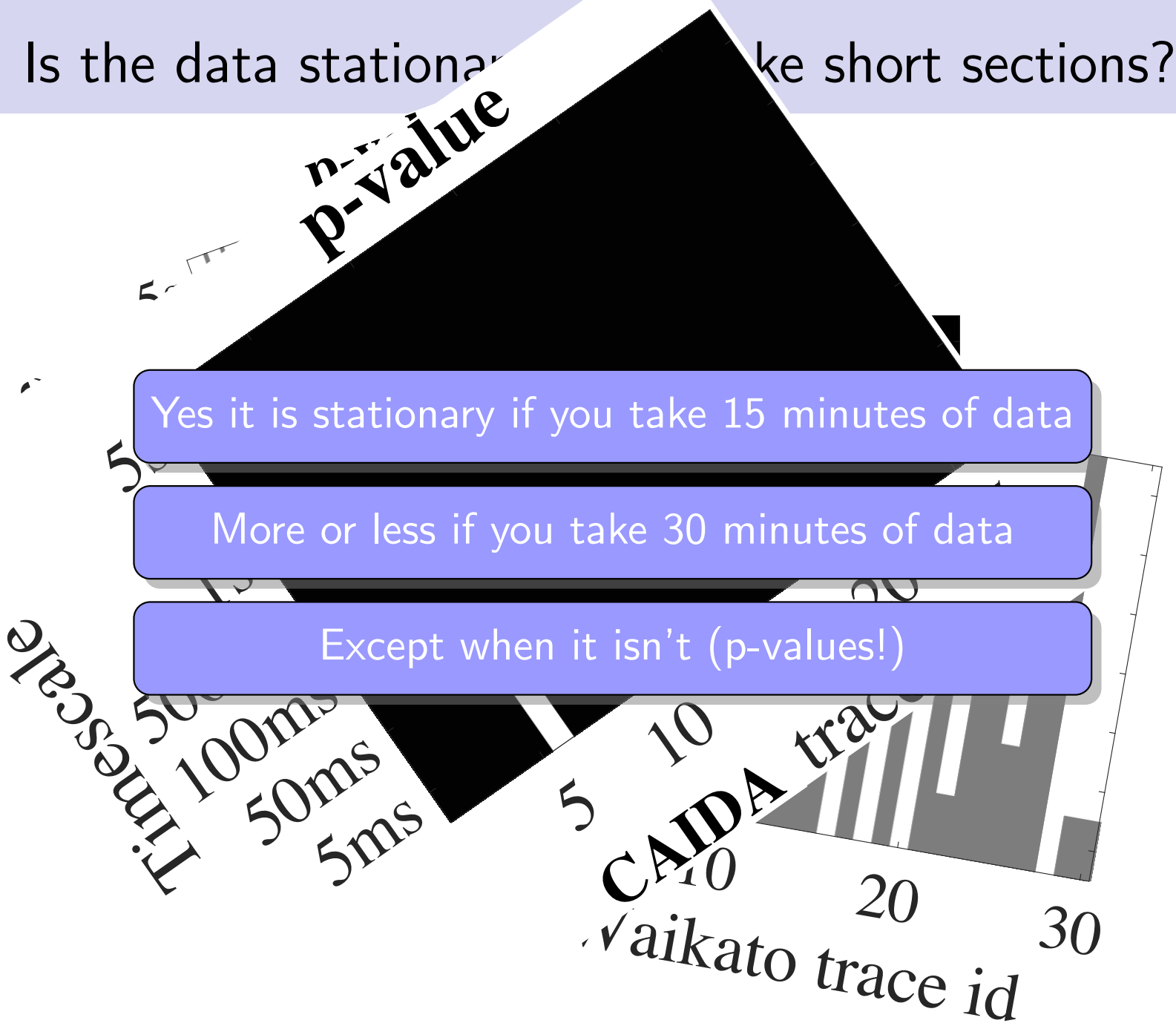
Is the data stationa ke short sections?



Yes it is stationary if you take 15 minutes of data

More or less if you take 30 minutes of data

Is the data stationa ke short sections?

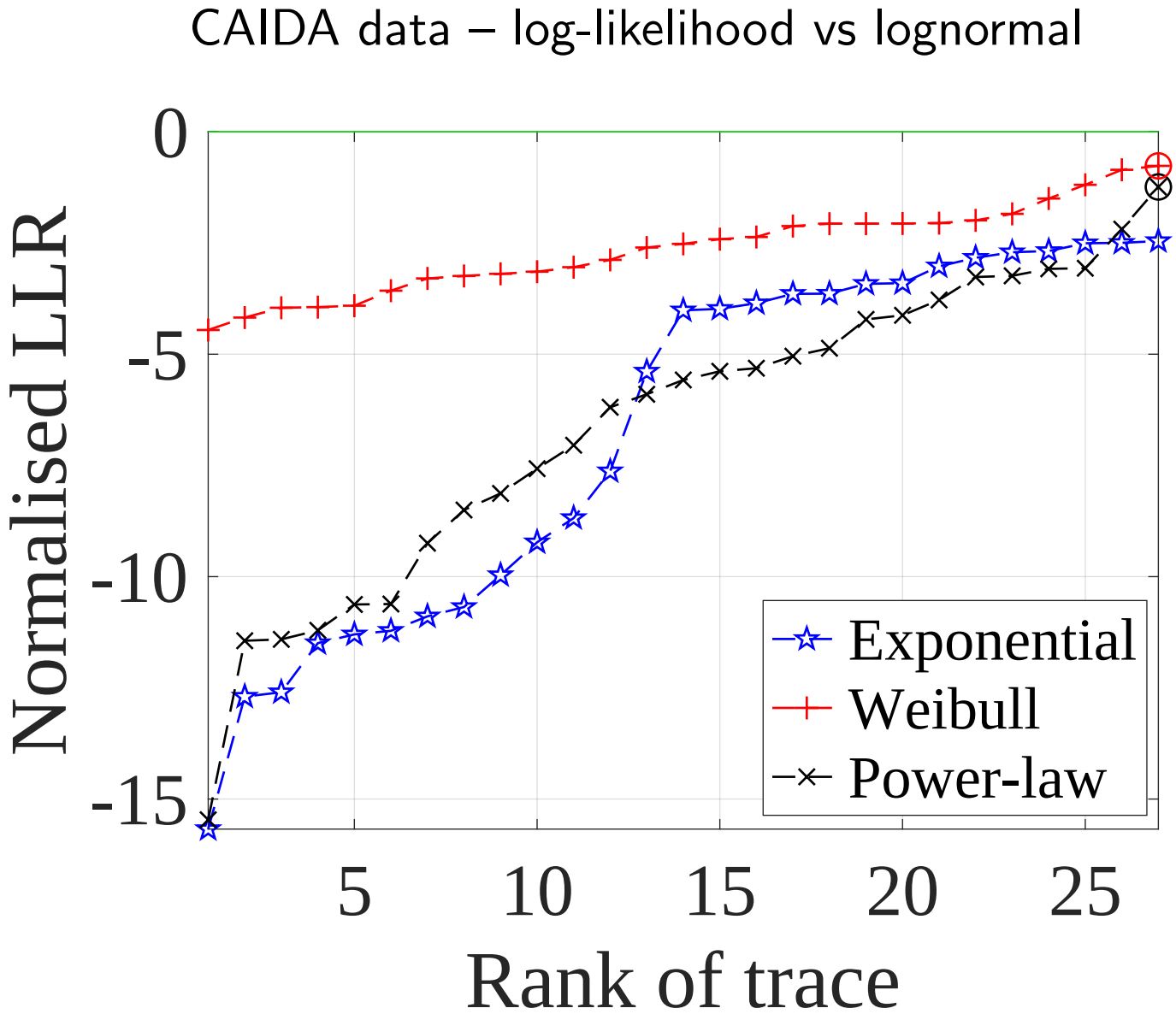


Yes it is stationary if you take 15 minutes of data

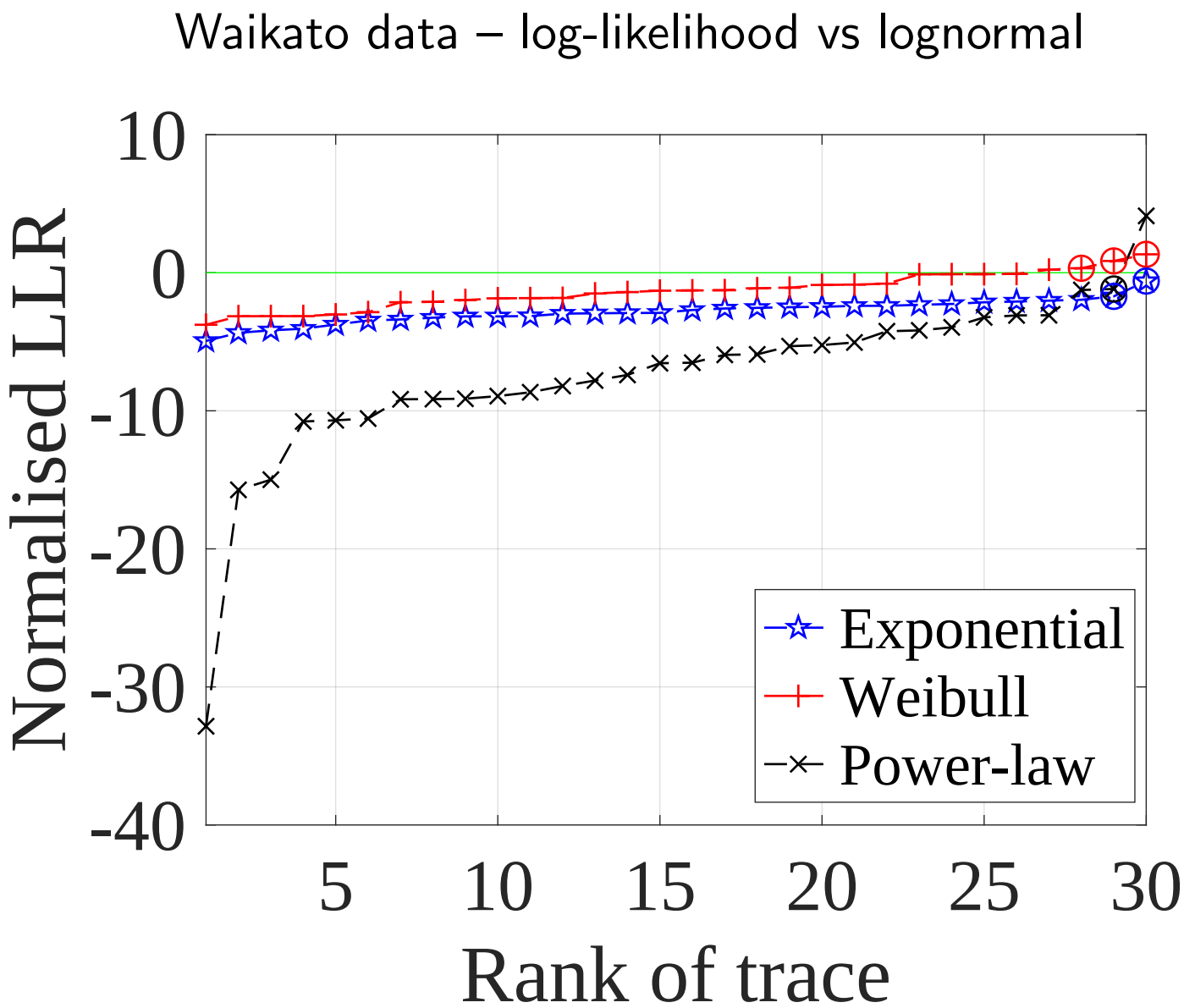
More or less if you take 30 minutes of data

Except when it isn't (p-values!)

So after all that, is lognormal a good fit? (1)



So after all that, is lognormal a good fit? (2)



Conclusions on the lognormal distribution

- The testing procedure used is robust for data with a suspected power law (Clauset et al 2009)– this is the methodology for testing this.
- By far the majority of traces lognormal was a best fit. Some anomalous traces existed (eg low traffic for half the trace, perhaps some outage existed).
- Use cases – we gave proof-of-concept tests to demonstrate that lognormal distribution gives improvement for estimating link sizes and for 95th percentile billing estimates.
- We did not give a complete model since the harder part remains: how does traffic vary across a day, week and year?
- Tremendous amount of work (huge amount of data analysed) by Mohammad Alasmar. It was not a trivial question to answer.

Conclusions of this research

Conclusion of this research

In fifteen minute long traces of internet traffic from typical aggregated sources the time series of bytes per unit time fits a log normal distribution when broken into time units between 5ms and 5secs.

Conclusions of this research

Conclusion of this research

In fifteen minute long traces of internet traffic from typical aggregated sources the time series of bytes per unit time fits a log normal distribution when broken into time units between 5ms and 5secs.

Use of this finding

If you want a reasonable 'background' traffic level for your research and you have an approximate mean for various times of day you could use this for your model.

Conclusions of this research

Conclusion of this research

In fifteen minute long traces of internet traffic from typical aggregated sources the time series of bytes per unit time fits a log normal distribution when broken into time units between 5ms and 5secs.

Use of this finding

If you want a reasonable 'background' traffic level for your research and you have an approximate mean for various times of day you could use this for your model.

Conclusion of this talk

Doing statistics responsibly can be relatively hard even if you're experienced and your question is quite simple. It took us several goes to ask the right questions of this data. If your work needs reasonable quality statistical support find a reasonable quality statistician (and a lot of data).