

Are popular websites' load times different from those of less popular ones?

It's more complicated than you think.....

Advanced Statistics Final Project

We browse the web every day.

When we visit a website, we wait before it fully loads.

- This is called the load time, T_{load} : it measures the time it takes for your browser to download all the resources on the webpage and do book-keeping work. Note that this does not include the rendering time, the time it takes for your graphics to render the webpage.

Question: Given two categories of websites, are their load times different?

Website Popularity Measure

We use the ALEXA ranking, which provides a daily ranking of the top 1m most visited websites in the world.

Data url: <http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>,
changes daily, random day \triangleq random sample

Two-sample t -test for means: Sample n_1 top-ranked ALEXA 1m websites and n_2 bottom-ranked ALEXA 1m websites. Compare the distributions.

Dilemma: ALEXA accounts for both resource-provider websites (a website that serves e.g. images to others) and content-provider websites (a website that serves actual content); we don't want to include resource provider websites' hompages (usually a landing page) in our sample. Solution next page

Loading Time Measure: T_{load}

We use the Chrome web browser in the headless mode, driven by a script, to visit each website. ← Simulation of user-controlled browsing experience

$$T_{load} = T_{resources} + T_{page} + T_{storage} \leftarrow \text{excludes only page rendering time (device-dependent)}$$

Which websites to sample?

A webpage is valid iff:

- Either HTTP mode or HTTPS mode returns status code 200 (success)
- It may be accessible either via `http://` or `https://`, or both;
- For `https`, the encryption certificate must be valid.
- Take

$$T_{\text{load}} = \begin{cases} \min\{T_{\text{https}}, T_{\text{http}}\} & \text{if both} \\ T_{\text{https}} & \text{if only https} \\ T_{\text{http}} & \text{if only http} \end{cases}$$

Remarks

This weeds out a large number of "resource provider" websites, e.g.
`googleapis.com` returns a `404 Not Found` status code.

For network turbulence, We take the avg in the load times distribution X :
 $T_{load} = \bar{X}$ for $n = 3$

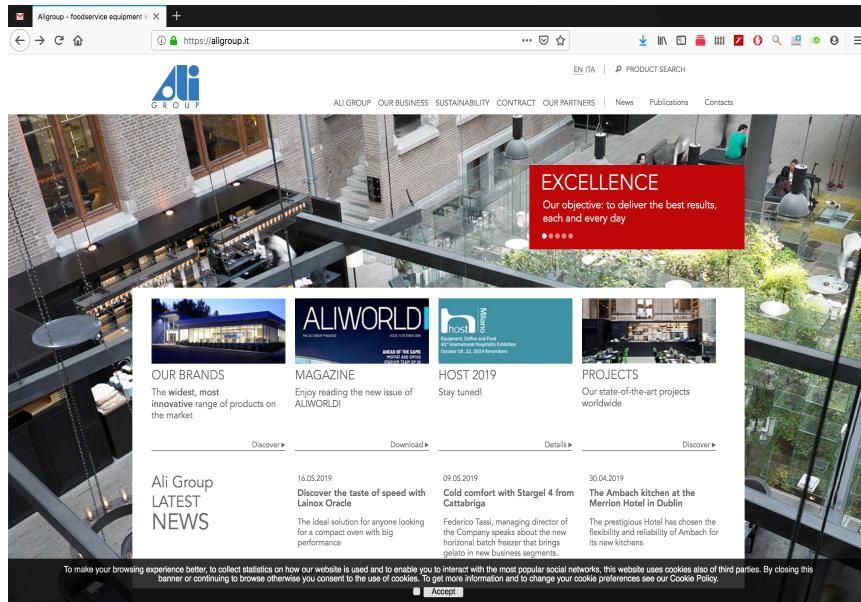
Our sampling distribution comes from the population distribution of all valid websites in the ALEXA top 1m ranking.

Examples of top-ranked websites from all over the world

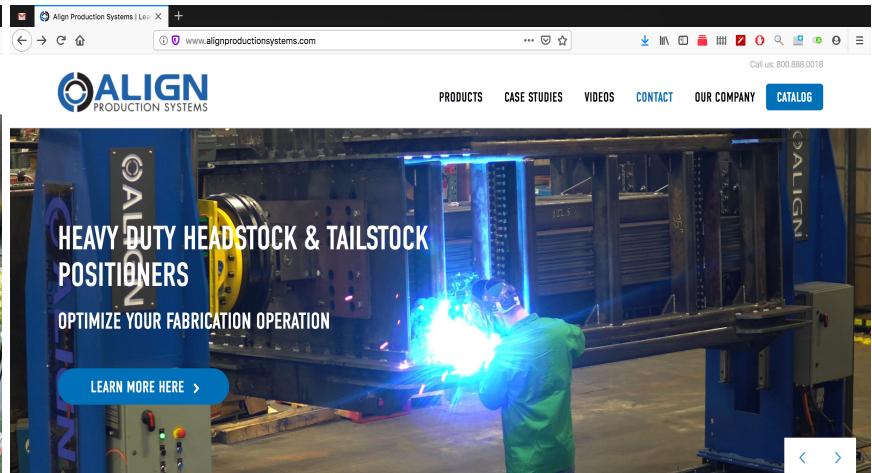
The image displays a grid of screenshots from several well-known international websites:

- Hao123 (中国):** A Chinese search engine and portal featuring a weather forecast (Beijing, 29-30°C), news links, and a news feed.
- Baidu (中国):** A Chinese search engine with a search bar, news links, and a news feed.
- Amazon.co.jp (日本):** The Japanese version of the e-commerce giant, showing promotional banners for Prime Video, Time Sales, and Points Up Chances, along with a "We deliver the world. Directly, 65+ countries." message.
- Yandex (Russia):** A Russian search engine with a weather forecast (+24°), news links, and a news feed.
- Mail.ru (Russia):** The Russian version of the webmail service, showing a search bar, news links, and a news feed.

Examples of bottom-ranked websites from all over the world



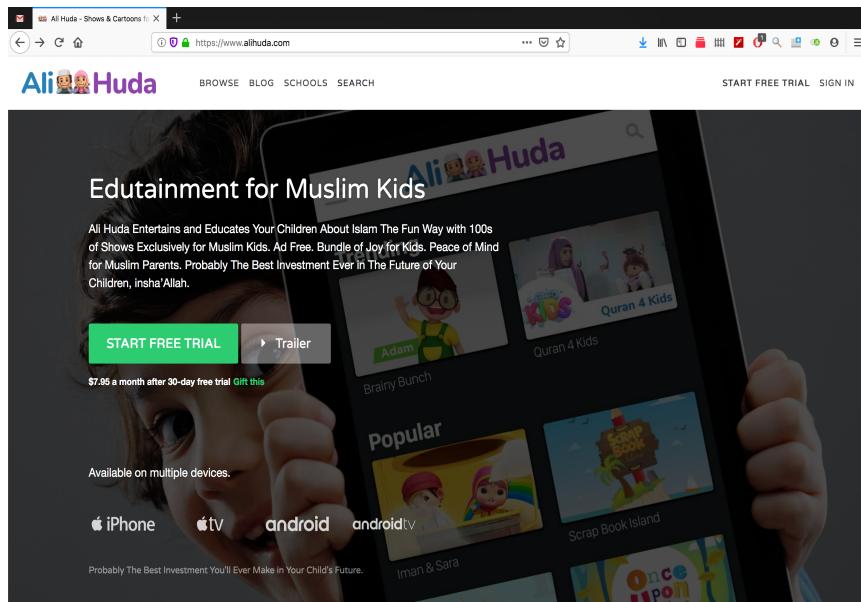
The screenshot shows the homepage of Ali Group. The header features the Ali Group logo and navigation links for EN, ITA, PRODUCT SEARCH, ALI GROUP, OUR BUSINESS, SUSTAINABILITY, CONTRACT, OUR PARTNERS, News, Publications, and Contacts. A red call-to-action box in the center says "EXCELLENCE Our objective: to deliver the best results, each and every day" with five stars. Below it, there are sections for "OUR BRANDS", "MAGAZINE", "HOST 2019", and "PROJECTS". A large image of a modern building with a glass facade is the background. At the bottom, there's a cookie consent banner and a "Discover" button.



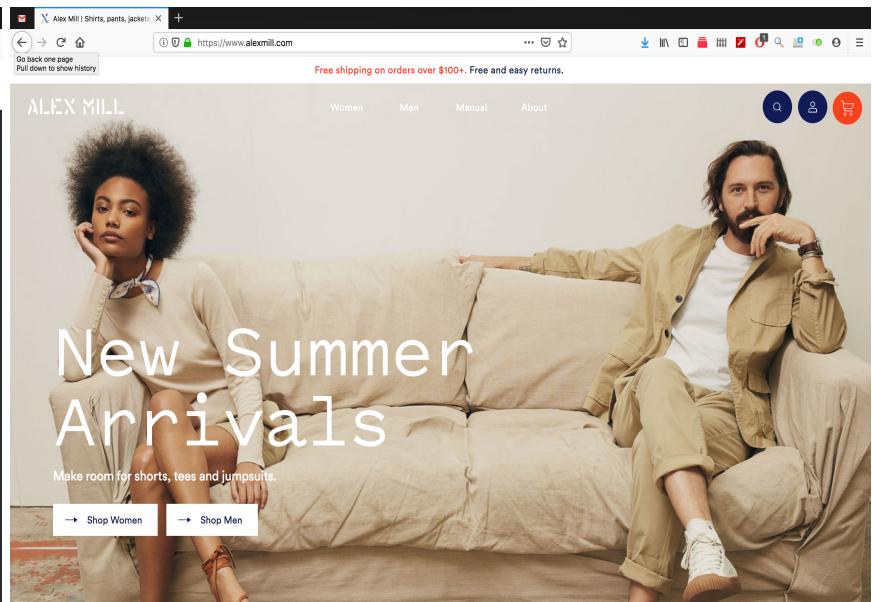
The screenshot shows the homepage of Align Production Systems. The header includes the Align logo, navigation links for PRODUCTS, CASE STUDIES, VIDEOS, CONTACT, OUR COMPANY, and CATALOG, and a phone number: Call us: 800.888.0018. The main content features a large image of a worker welding a large metal structure labeled "ALIGN". Text on the left reads "HEAVY DUTY HEADSTOCK & TAILSTOCK POSITIONERS" and "OPTIMIZE YOUR FABRICATION OPERATION". A blue "LEARN MORE HERE" button is present.

ABOUT ALIGN PRODUCTION SYSTEMS

OUR EMPHASIS: TO HELP OUR CLIENTS OVERCOME PRODUCTION OBSTACLES AND MAXIMIZE



The screenshot shows the homepage of Ali Huda. The header has the Ali Huda logo and links for BROWSE, BLOG, SCHOOLS, and SEARCH. It features a large image of a child holding a smartphone displaying the Ali Huda app. Text on the page includes "Edutainment for Muslim Kids", "Ali Huda Entertains and Educates Your Children About Islam The Fun Way with 100s of Shows Exclusively for Muslim Kids. Ad Free. Bundle of Joy for Kids. Peace of Mind for Muslim Parents. Probably The Best Investment Ever in The Future of Your Children, insha'Allah.", and "START FREE TRIAL". There are also sections for "Available on multiple devices." and download links for iPhone, iPad, and Android.



The screenshot shows the homepage of Alex Mill. The header features the Alex Mill logo and links for Women, Men, Manual, and About. It includes a search bar and social media icons. The main image shows a woman and a man sitting on a couch. Text on the page reads "New Summer Arrivals" and "Make room for shorts, tees and jumpsuits.". Buttons for "Shop Women" and "Shop Men" are at the bottom.

Potential problems

- Landing pages might still be included in the sample
- Measurement is inherently US-centric; we're taking latency measurements from US east coast, so websites in other parts of the world (e.g. Russia, China, India) will behave poorly for T_{load} .

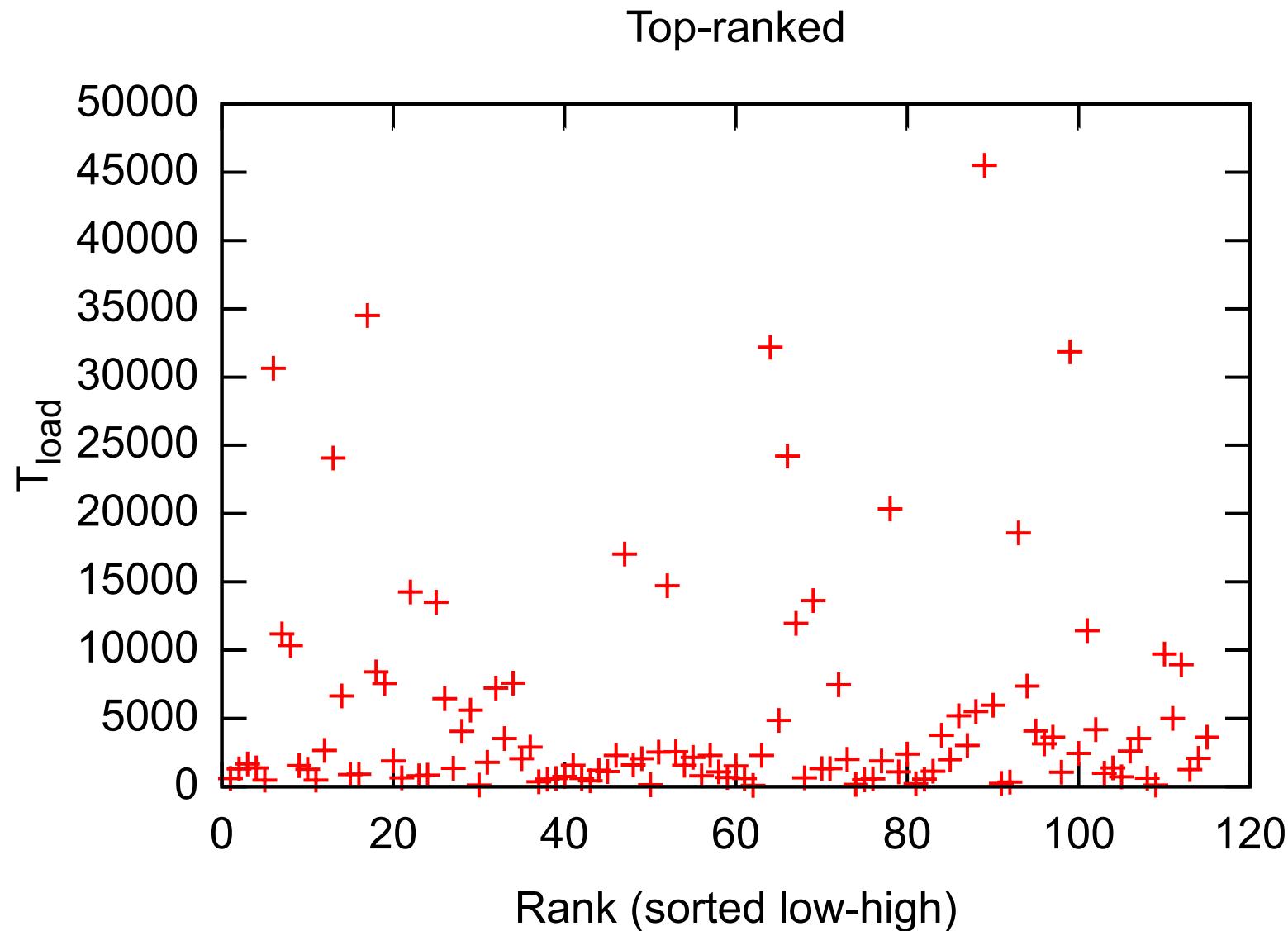
Test: 2-sample t -test for means, two tails

Conditions ($n_1 = 115$, $n_2 = 138$)

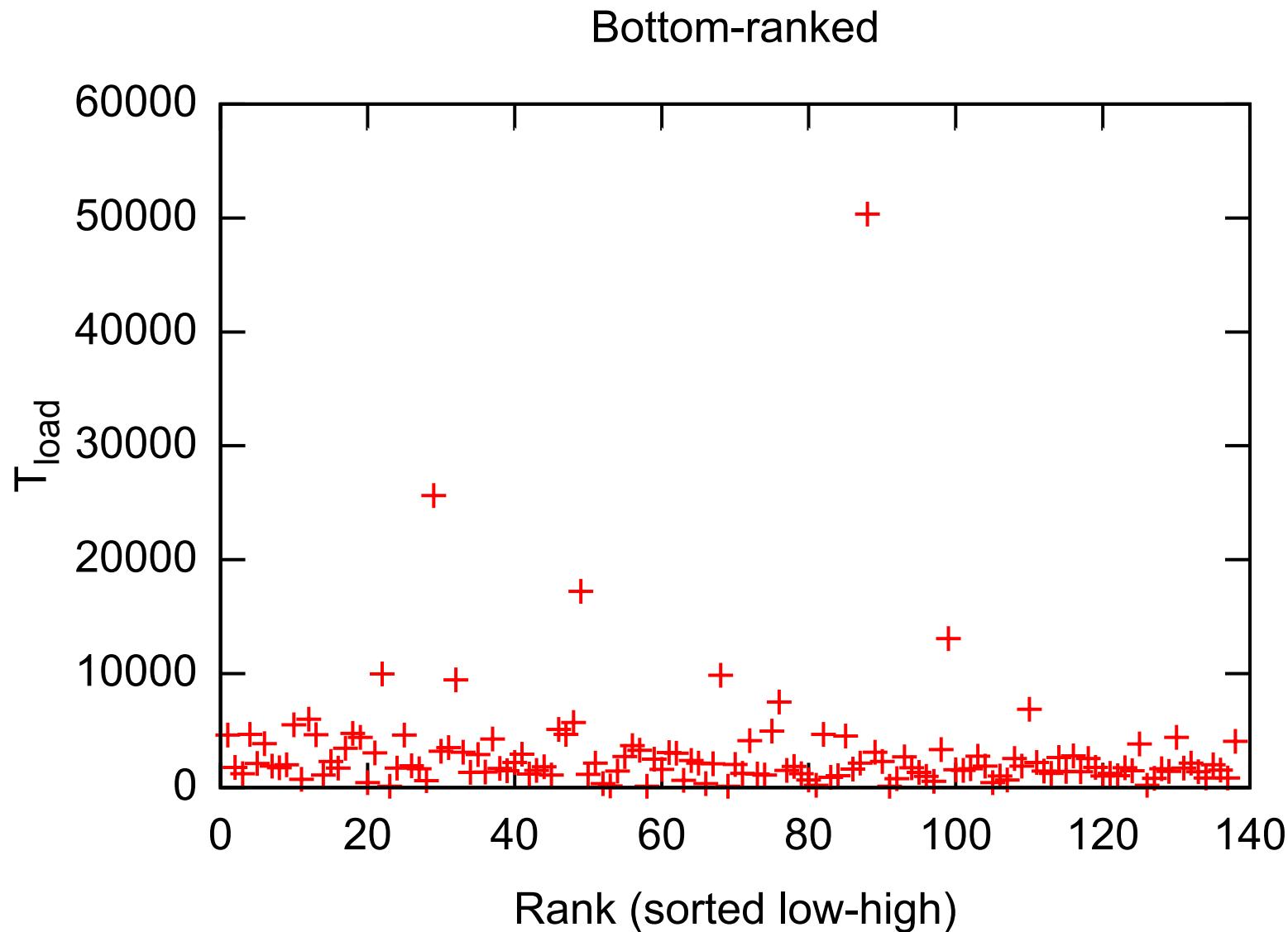
- Random: ✓, the ALEXA rank is taken for a random day
- Independent: ✓, our population size N is the total number of active websites, which is very very large... But $n_1, n_2 < 200$ is small.
- Normal: ✓, $n_1, n_2 > 30 \rightarrow$ CLT holds.

We do the 2-sample t -test for two distributions: the distribution of T_{load} for top-ranked websites and the distribution of T_{load} for bottom-ranked websites.

T_{load} distribution for Top-ranked 115 sites



T_{load} distribution for bottom-ranked 138 sites



Hypotheses

H_0 : That the mean of avg page load time for sites ranked 1-115 = the mean of avg page load time for sites ranked bottom 1-138

H_a : That the mean of avg page load time for sites ranked 1-115 \neq the mean of avg page load time for sites ranked bottom 1-138.

t-test results ($\alpha = 0.05$)

Let T_1, T_2 be our distributions. T_1 is the top-ranked website distribution, T_2 is the bottom-ranked website distribution.

$$S_{T_1} = 8106.177131633802$$

$$S_{T_2} = 5085.949930330646$$

$$\overline{T_1} = 5329.4502753620545$$

$$\overline{T_2} = 3064.815464975765$$

$$t^* = \frac{\overline{T_1} - \overline{T_2}}{\sqrt{S_{T_1}^2/n_1 + S_{T_2}^2/n_2}} = 2.5997102453932923$$

$$p(t > t^* | \text{df} = \min\{n_1, n_2\} - 1 = 114) = 0.00528 < \alpha = 0.05.$$

Conclusion

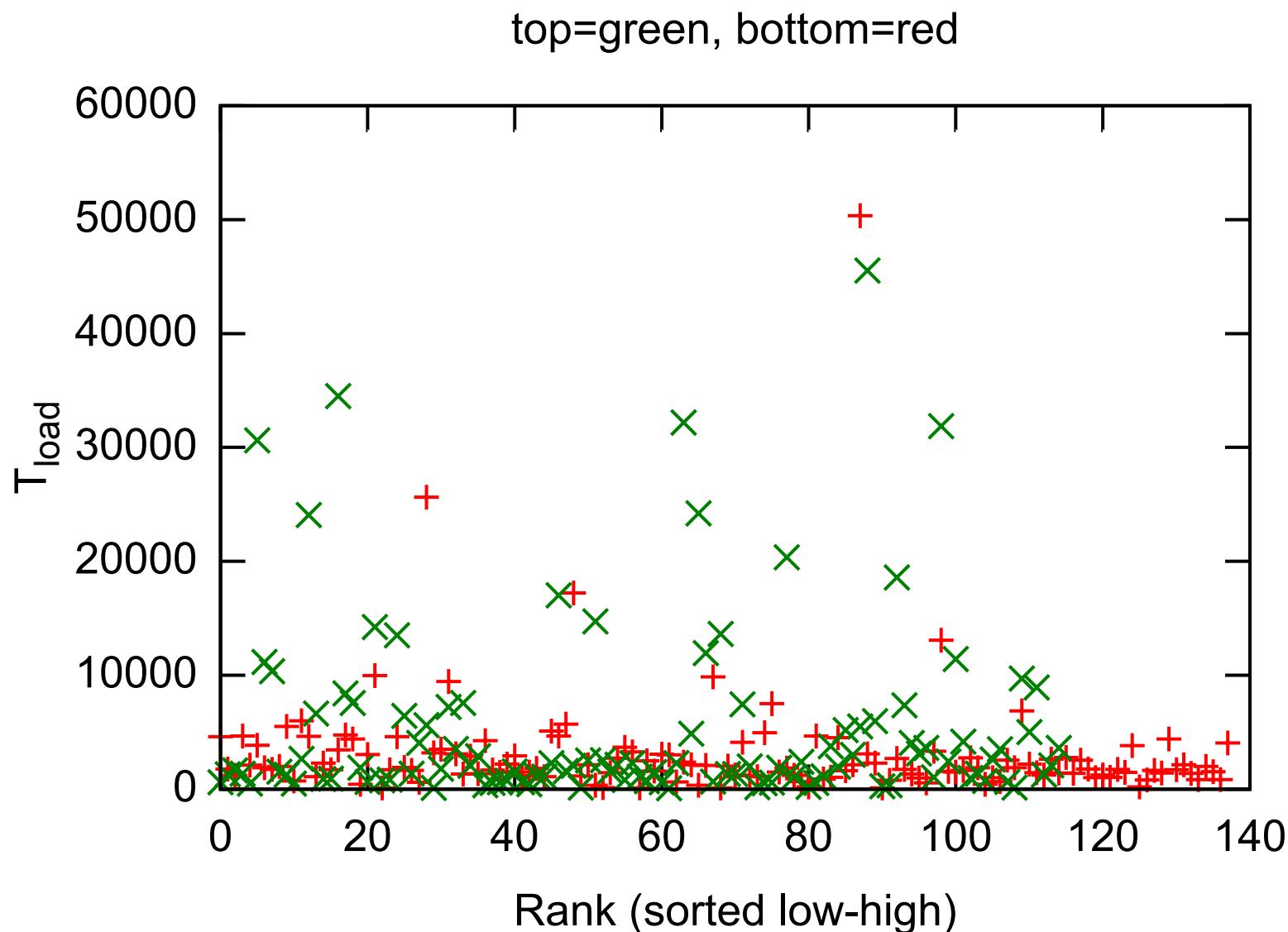
We have to reject H_0 . There is suff evidence to conclude that bottom-ranked sites load faster, on average.

- $p = 0.005282249701030417$ with $\alpha = 0.05!$

Why do low-ranked sites load faster? Small sites are frequently located on large platforms (e.g. `wordpress.com`, `tumblr.com`); these large platforms have the infrastructure and the sites might have less multimedia content.)

- Also, large sites might serve a more local audience (e.g. [163.com](#) or [mail.ru](#))

Combined scatterplot, T in millisecs



Future work

- Sample more websites
- We took the mean of three load times. We should take more?
- Stddev is really, really high. Is there a way to further minimize stddev?
- Take geolocated measurements.
- Do more work on identifying landing pages and content farms
 - they may harm the integrity of sample data collected