

R USERS ALL AROUND THE WORLD

SOME DATA COLLECTED AND PRESENTED BY GERGELY DAROCZI



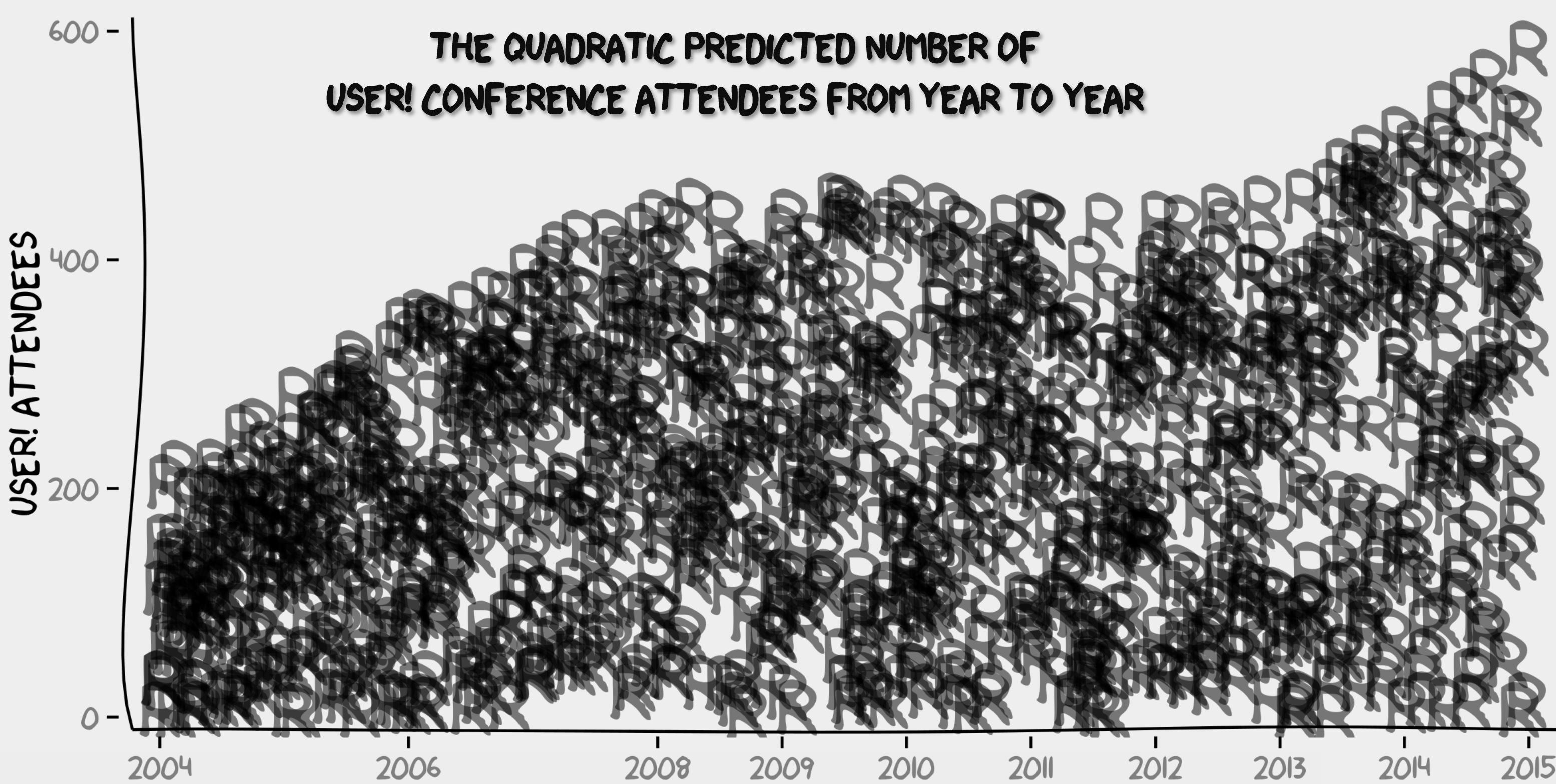
OVERVIEW

This poster was inspired by a 2013 blog post on „*Where is the R activity?*” and a few follow-up articles, where the authors tried to determine the **number of R users all around the world**.

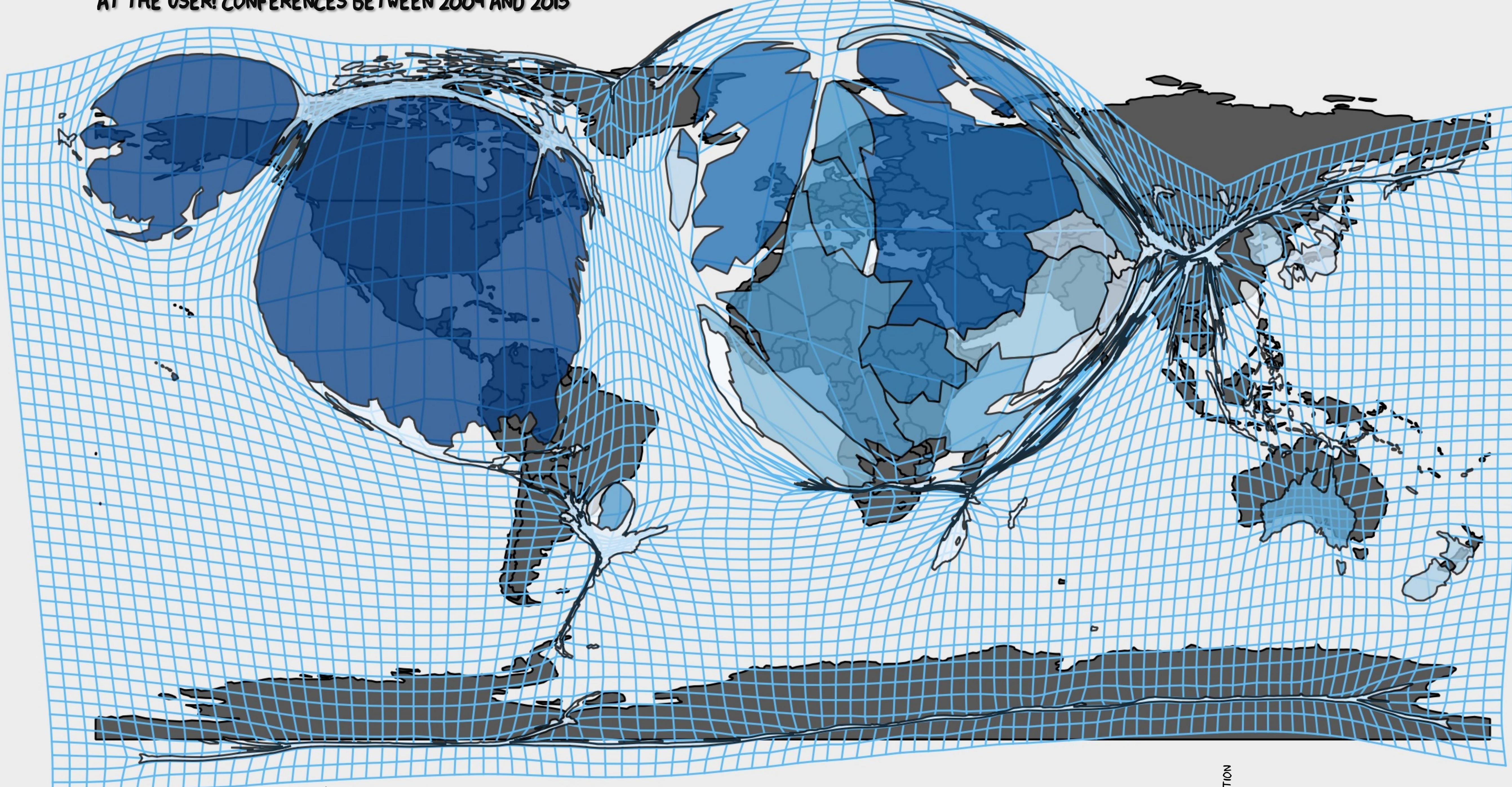
The first version of my related experiment included various data sources on the number of **R Foundation members**, lists of attendees of **useR! conferences**, the number and size of **R User Groups** registered on meetup.com, **R package downloads** from the RStudio cloud CRAN servers, R-related **search queries via Google** and the geographical distribution of **GitHub users** with at least one R repository. The results and a combined R score were published on an interactive D3.js application, and a paper-based poster at the useR! 2014 conference in Los Angeles – of which this is a revised and updated version.

Now I focus on the number of attendees at all previous useR! conferences along with the fresh dataset of **500+ registrants** here (aka *useR! 2015*). First, I visualized and fitted simple statistical models on the number of participants in time and space. After that, the actual research in this paper was collecting and cleaning the names of all previous useR! attendees from different data sources, which list can be used as the base of the **capture-recapture** method to **estimate** the size of the unknown population with **loglinear models**. How many guys and girls were using R in the past 10 years? Is it 2 or 3 millions? Even more?

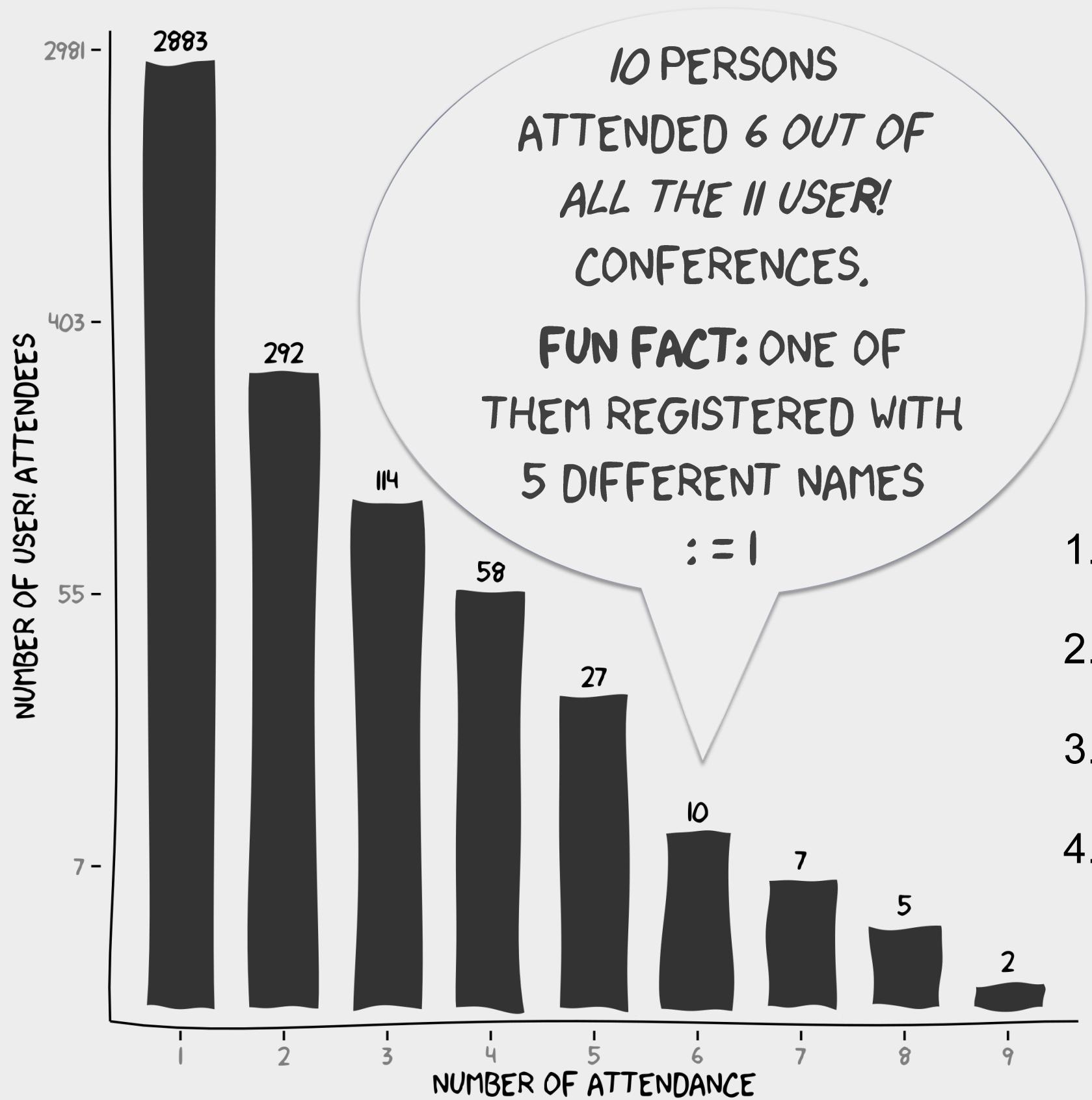
Although this poster was not able to answer this question after all, it still features some related statistical methods along with a few **XKCD** webcomic-styled plots of the results. **I hope you enjoy the show!**



OVERALL NUMBER OF ATTENDEES AT THE USER! CONFERENCES BETWEEN 2004 AND 2015



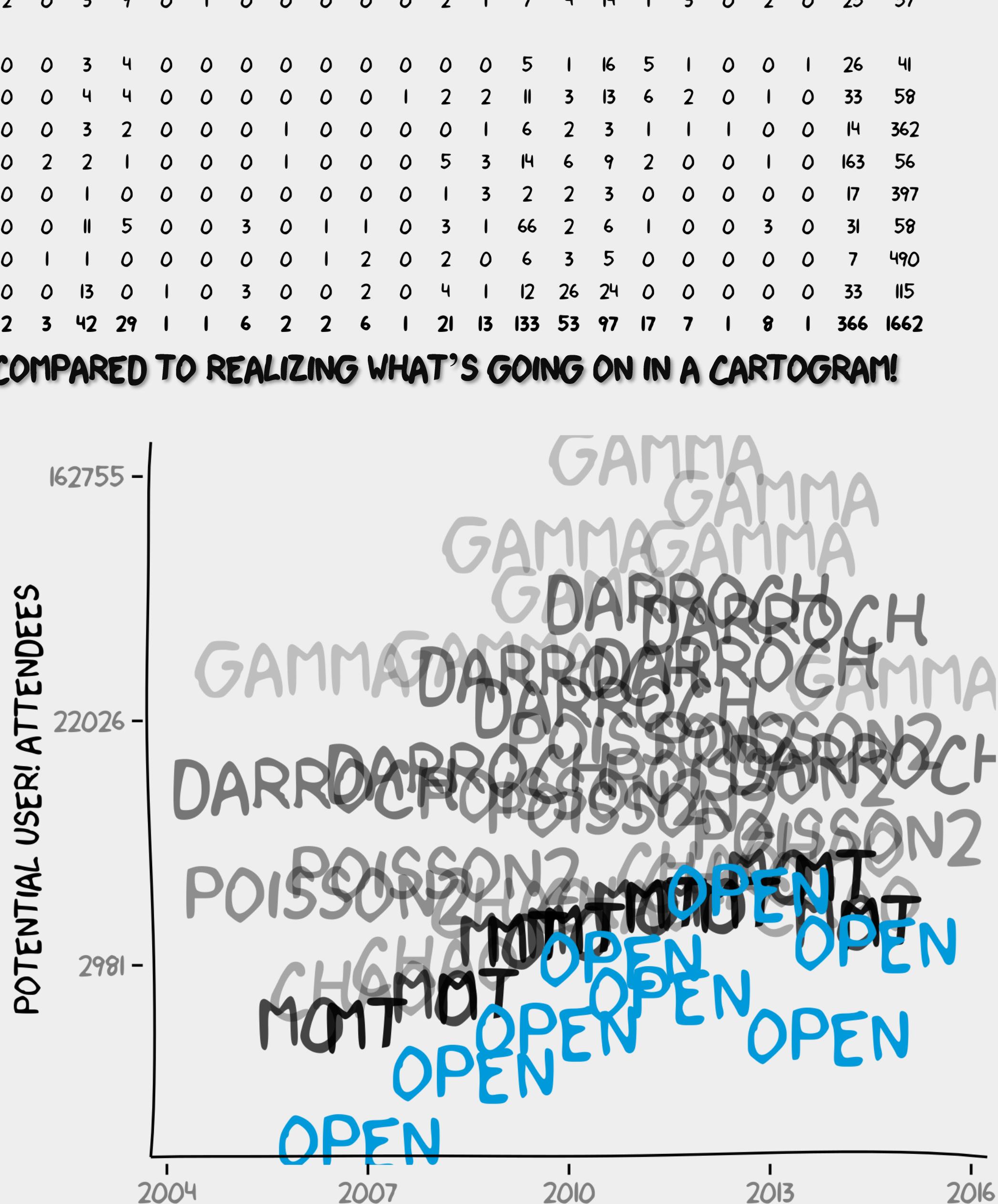
NOTE TO DATA VÍZ GUYS: MOST STATISTICIANS CAN READ, ANALYSE AND INTERPRET HUGE TABLES EASIER AND QUICKER COMPARED TO REALIZING WHAT'S GOING ON IN A CARTOGRAM!



UX RESEARCH METHODS RESULTS

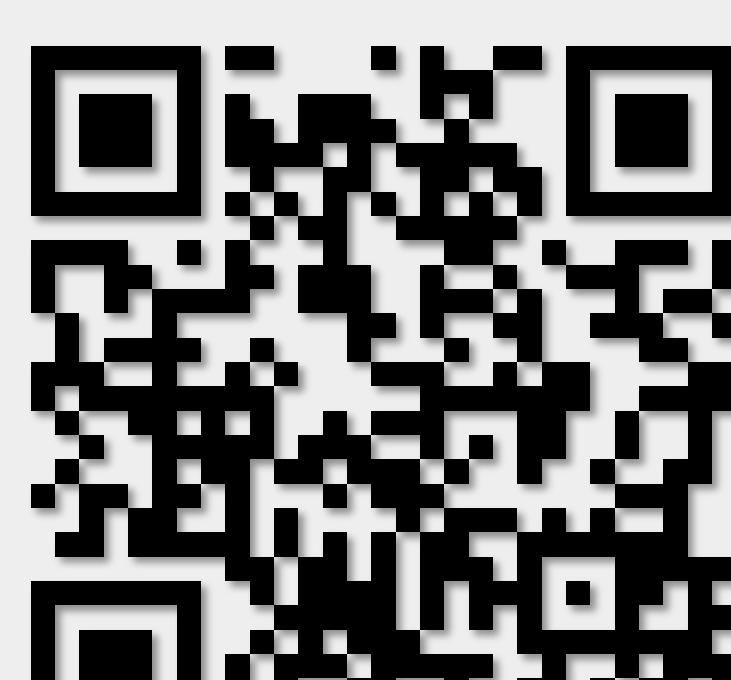
1. The name of all **useR! 2004-2015** attendees was compiled in a list, where some (left figure) **recurring participants** were identified after manual **data cleaning**.
 2. These results can be used as the "number of units captured i time" in a "mark and recapture" research method for **estimating the unknown population size**.
 3. Building open-population models by estimating the **observation probability**, the **survival rate** and the **number of new arrivals** based on historical data.
 4. Building 6x8 **closed population** models based on 3-years long time intervals (right figure): the *lower bound* estimate of Chao showed an increase from 2,500 to 7,000 even with and without a time effect, while Daroch's model estimates tend to spread between 15,000 and 60,000. The Gamma and Poisson models of Rivest and Baillargeon (2012) returns some restrained (5,000-20,000) and some much higher Gamma results (30,000-200,000).

Σ: Can an ecologist please help me interpret the results?



SOME KNOWN LIMITATIONS

- UseR! attendees do not represent R users.
 - Further data cleaning is required.
 - Missing data for 2007. Any help appreciated!
 - Chinese R Conference is not included :o
 - Closed population model is not appropriate.
 - Improved model selection is desired.



REFERENCES

- James Cheshire (2013). *Where is the R Activity?* Blog post at <http://spatial.ly>
 - Gergely Daroczi (2013). *The attendants of useR! 2013 around the world.* Blog post at <http://blog.rapporter.net>
 - Gergely Daroczi (2014). *R activity around the world.* Web applicatoin published at <http://blog.rapporter.net>
 - Gergely Daroczi (2014). *R users all around the world.* Poster presented at useR! 2014. <http://blog.rapporter.net>
 - Dominique Andrieu, Christian Kaiser and André Ourednik: *ScapeToad.* v1.1. <http://scapetoad.choros.ch>
 - Gastner, M.T. and Newman, M. (2004). *Diffusion-based method for producing density equalizing maps.* PNAS 101(20): 7499-7504.
 - Sophie Baillargeon, Louis-Paul Rivest (2014). *Rcapture: Loglinear Models for Capture-Recapture Experiments.* R pkg v1.4-2. CRAN.
 - Emilio Torres Manzanera (2014). *xkcd: Plotting ggplot2 graphics in a XKCD style.* R pkg v0.0.4. CRAN.