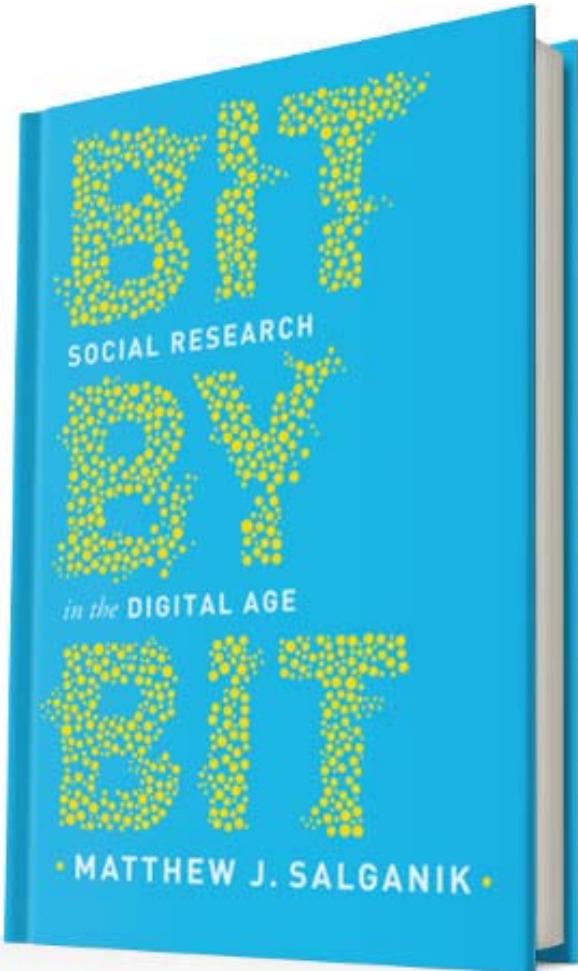


# Two sides of the same coin

## The use of Big Data in social research through theory & practice

Ságvári Bence, MTA TK

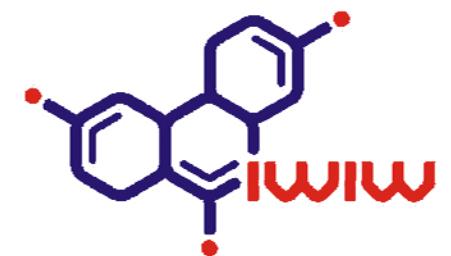
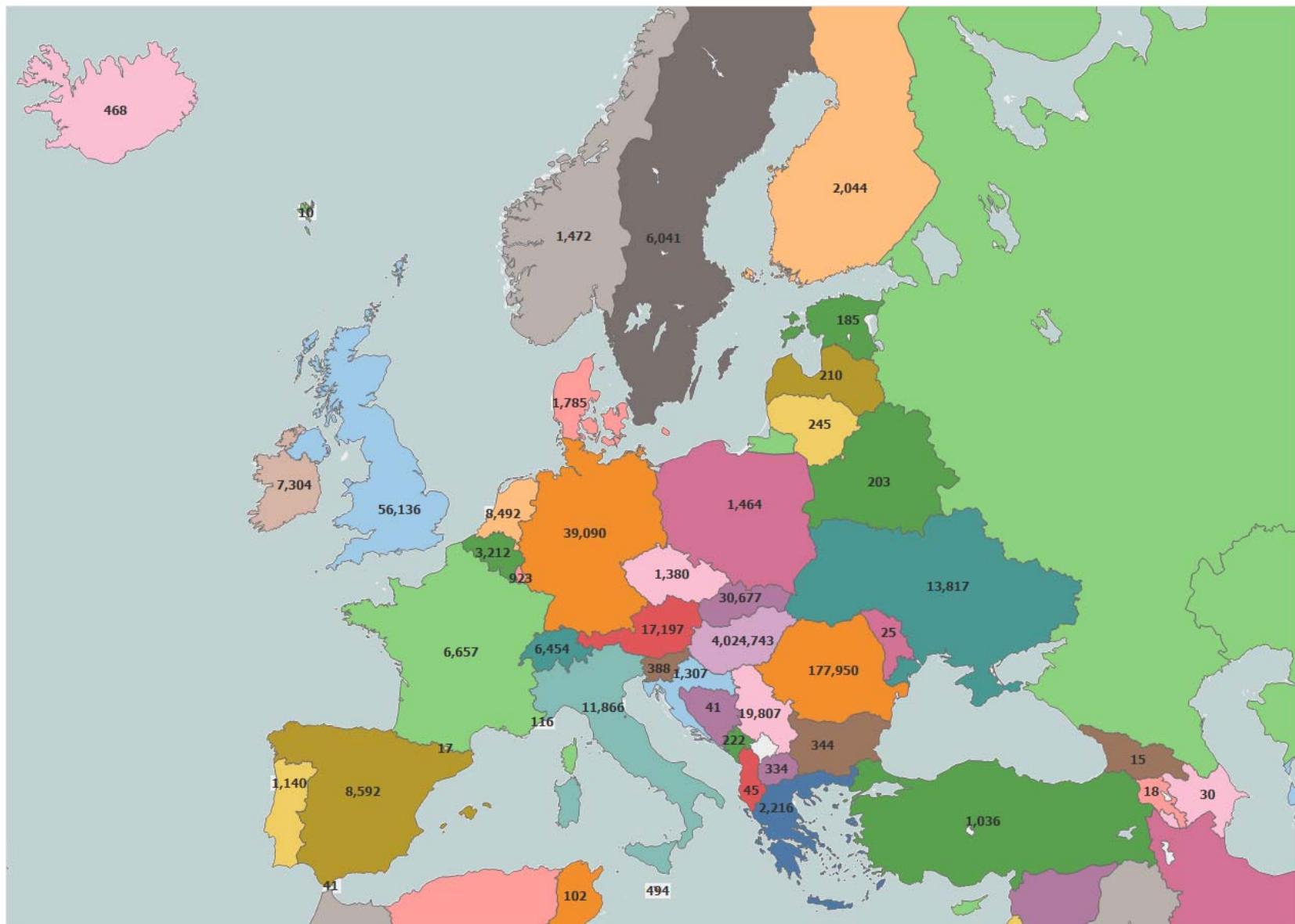


# Ten common characteristics of big data

Matthew J. Salganik (2017) Bit by Bit: Social Research in the Digital Age. Princeton University Press

# 1. BIG

“Large datasets are a means to an end;  
they are not an end in themselves.”



## **2. ALWAYS-ON**

**"Always-on big data enables the study of unexpected events and real-time measurement."**

# **3. NONREACTIVE**

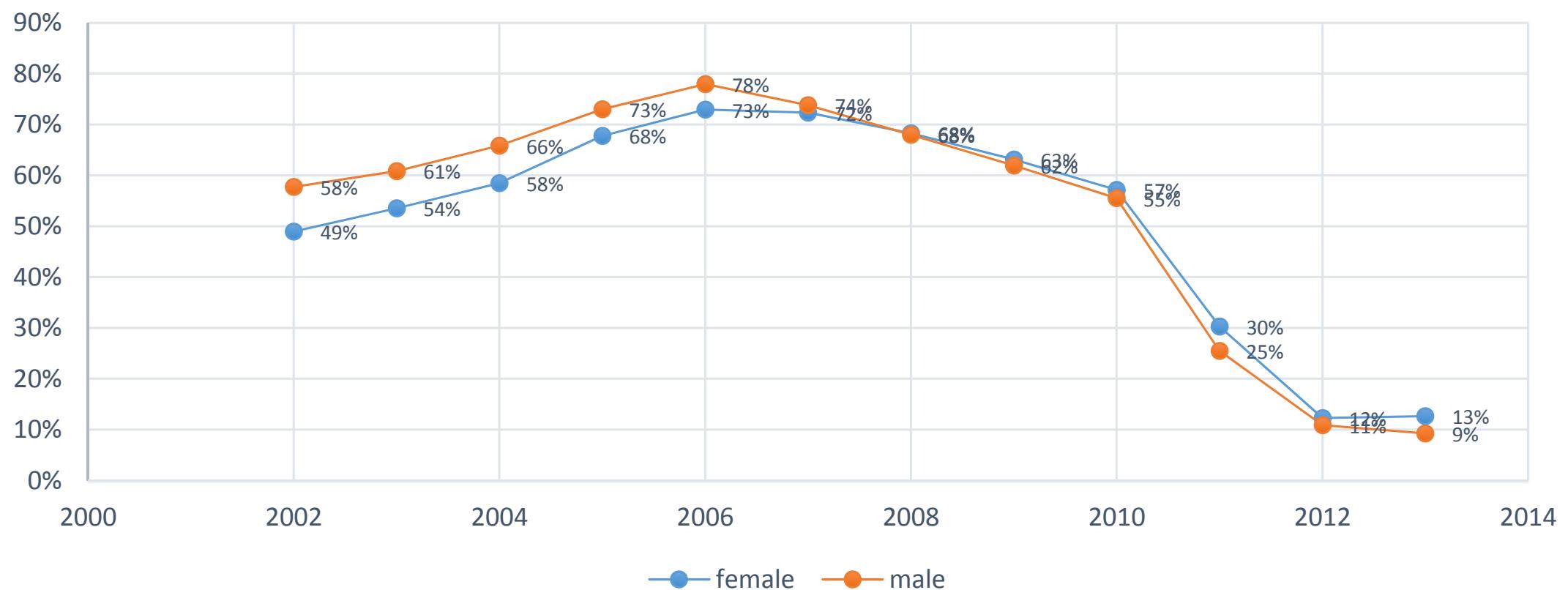
**"Measurement in big data sources is much less likely to change behavior."**

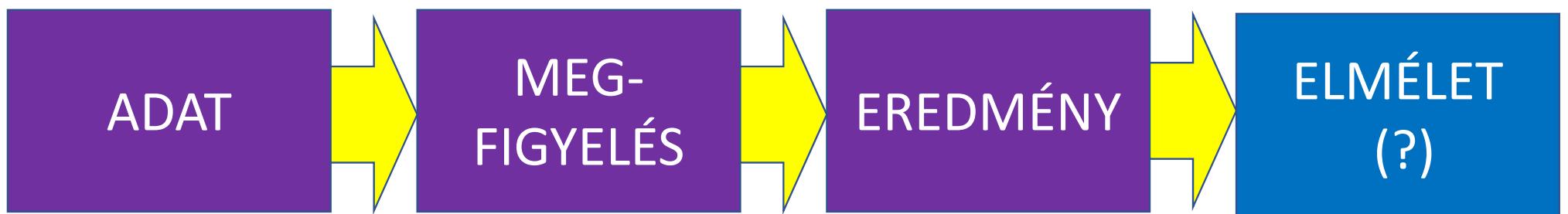
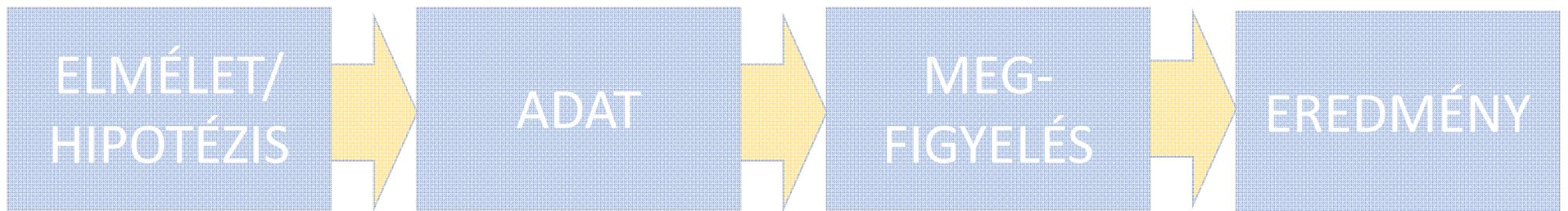
# **4.** INCOMPLETE

“No matter how big your big data, it probably doesn’t have the information you want.”

# Massive share of missing information on age

## Users with valid age by gender





# 5. INACCESSIBLE

“Data held by companies and governments  
are difficult for researchers to access.”

Az Üzleti titok megismerője **teljes körű felelősséggel tartozik Üzleti titokkal kapcsolatos kötelezettségei bármilyen megsértésért, ezekkel kapcsolatos kárért**, ha a titoksértés vagy kár az Üzleti titok megismerőjének felerőható, vagy egyébként ő okozta, beleértve az alkalmazottai, Üzleti Partnerei és Jogosított megismerői általi károkozást, vagy titoksértést is. Az Üzleti titok megismerője a tudományos kutatásban részt vevő alkalmazottak és partnerek felelősségét szerződésben rögzíti.

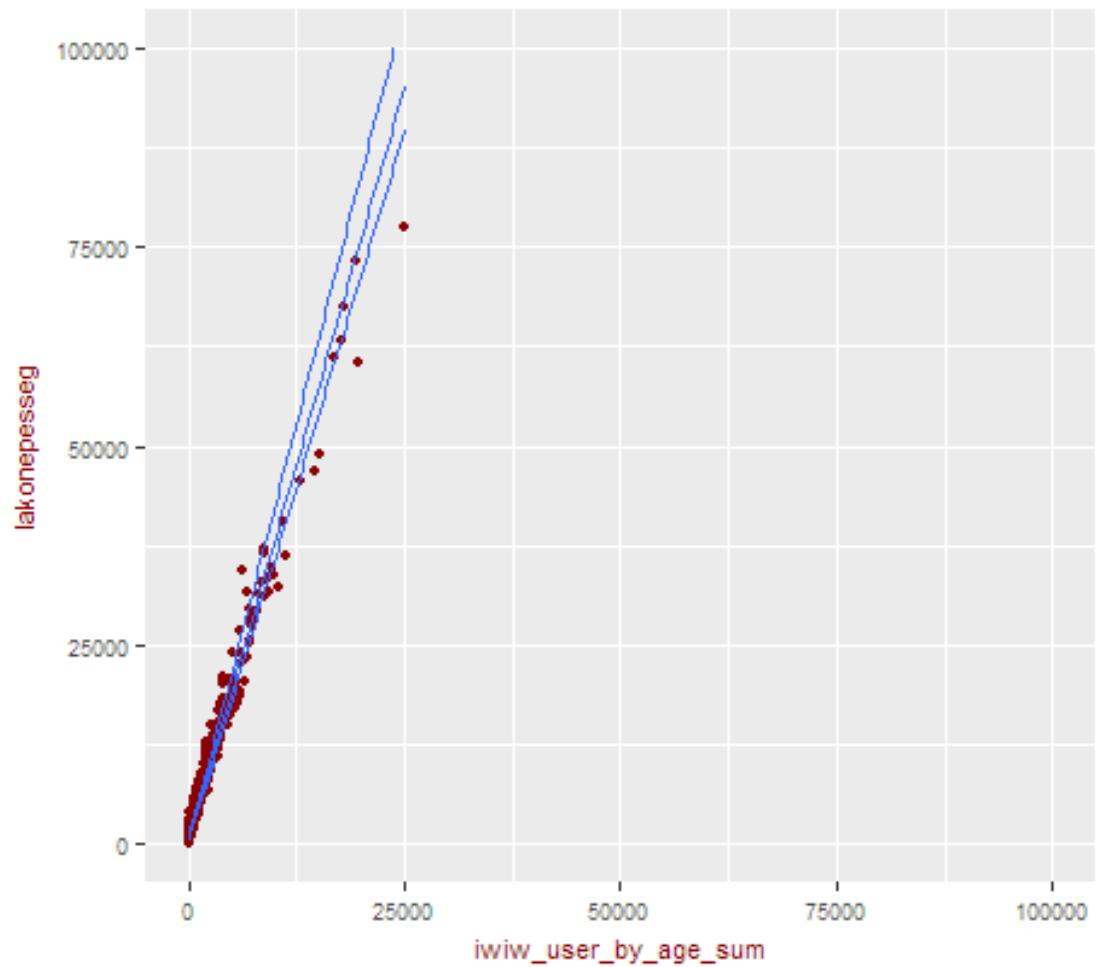
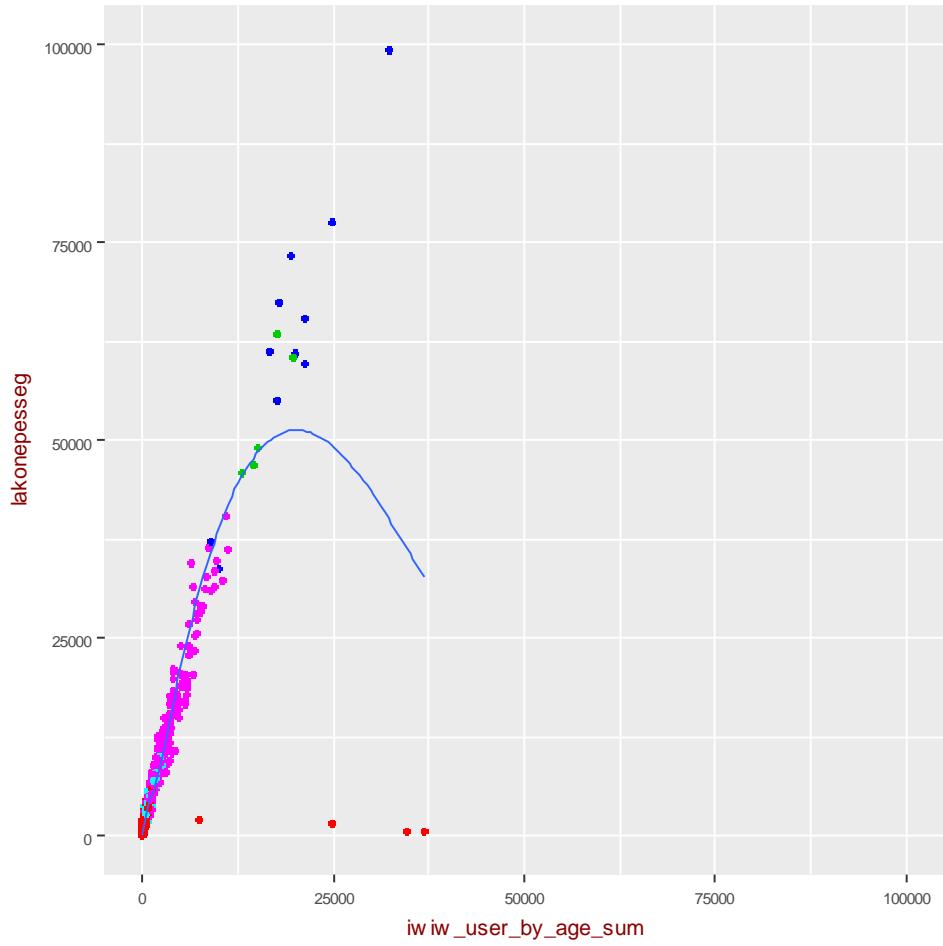
A tudományos kutatás eredményeit a Partner közölheti olyan módon, hogy a tudományos és ismeretterjesztő publikációkba üzleti titok nem kerülhet. **A publikációk közlésének menetét a Felek a kizárolagos felhasználási szerződésben fektetik le.**

A jelen Megállapodás hatálya alatt az Üzleti titok jogosultja írásbeli kérésére az Üzleti titok megismerője a fizikailag birtokában lévő Üzleti **titkokat köteles az Üzleti titok jogosultjának visszaszolgáltatni vagy megsemmisíteni** (az Üzleti titok jogosultja utasítása szerint), annyiban amennyiben az ésszerűen végrehajtható.

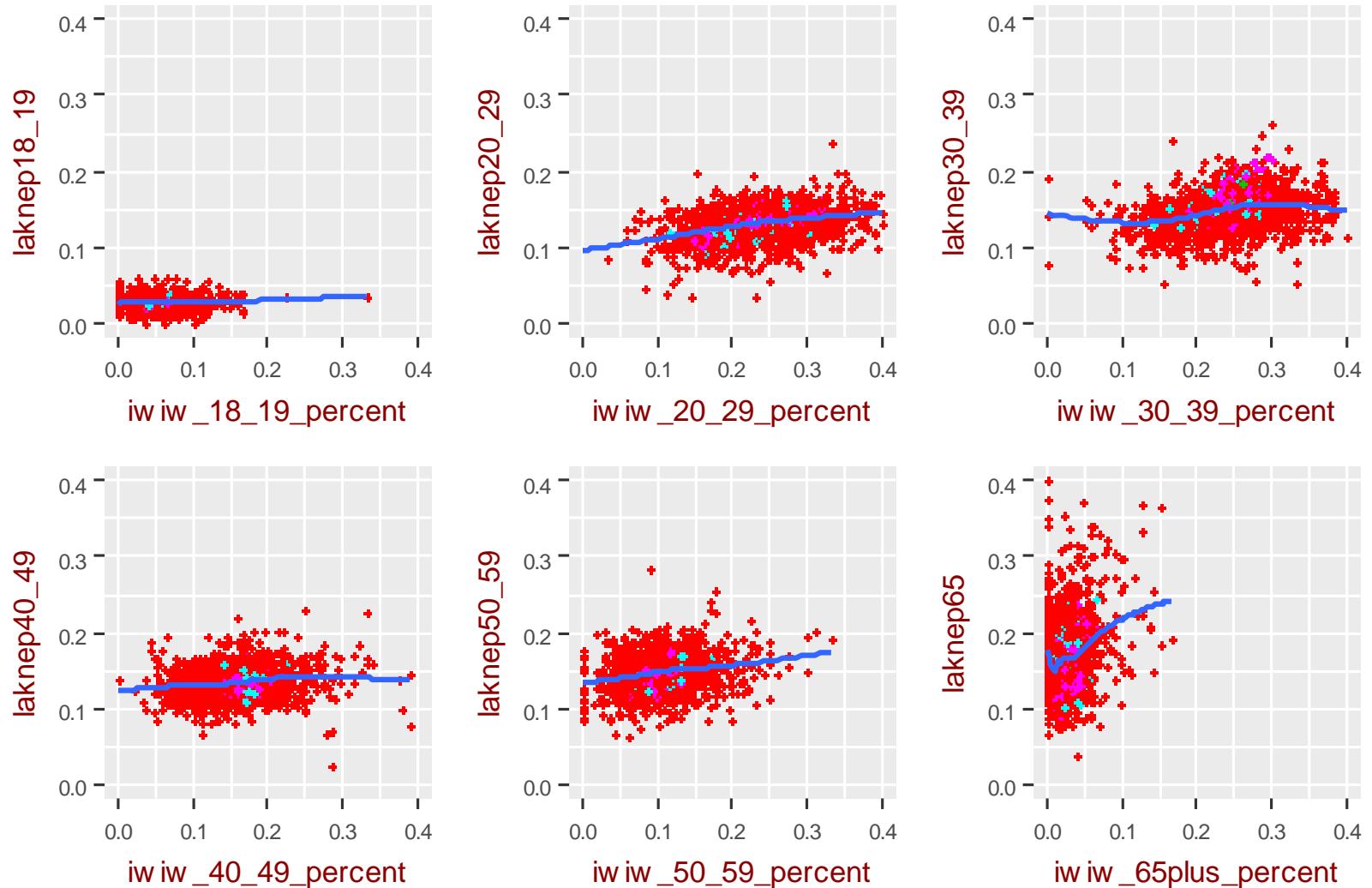
# **6.** NONREPRESENTATIVE

"Non-representative data are bad for out-of-sample generalizations, but can be quite useful for within-sample comparisons."

# Real vs. iWiW population in Hungarian settlements



# Age distribution of real vs. iWiW population in Hungarian settlements



# 7. DRIFTING

“Population drift, usage drift, and system drift make it hard to use big data sources to study long-term trends.”

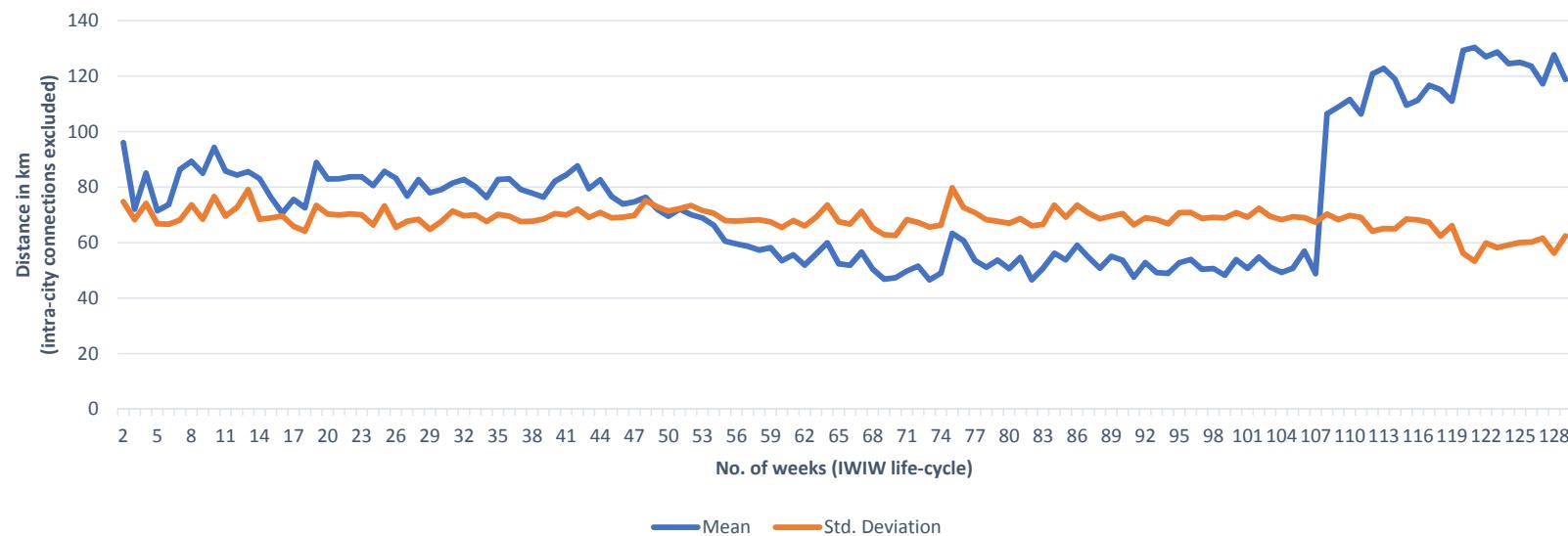
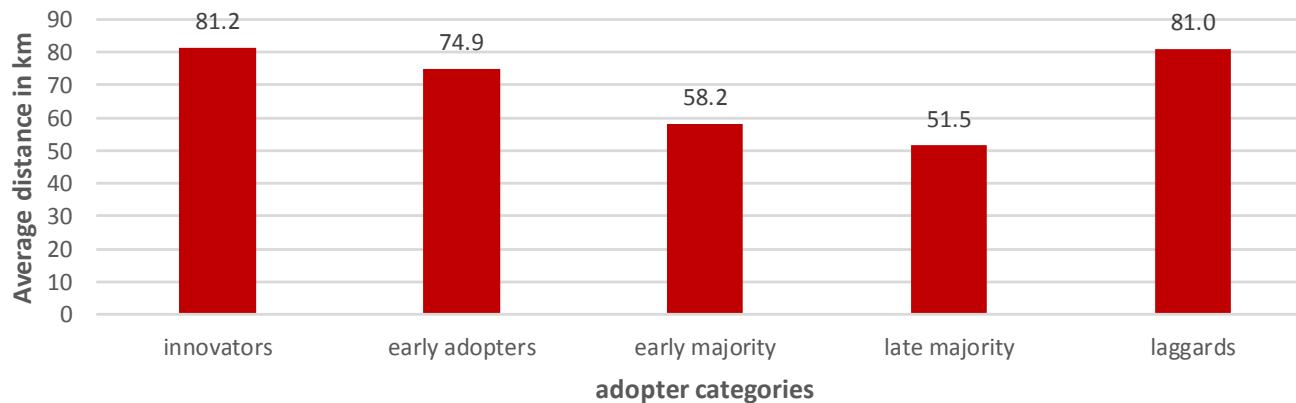
# **8. ALGORITHMICALLY COUNFOUNDED**

**"Behavior in big data systems is not natural; it is driven by the engineering goals of the system."**

# 9. DIRTY

“Big data sources can be loaded with junk and spam.”

Average distance of invitor-invited connections across adopter categories  
 (intra-city connections excluded)



# **10. SENSITIVE**

“Some of the information that companies and governments have is sensitive.”

# Thank you...

A screenshot of the RStudio IDE interface. The main workspace shows R code for reading data from URLs and performing operations like `head` and `gather`. The environment pane shows variables like `url`, `names`, `queries`, `search\_term`, and `search\_url`. The user library pane lists several packages such as `acepack`, `aod`, `AlgDesign`, `assertthat`, `backports`, `base64enc`, and `BH`.

```
1 # Load required packages
2 library(tidyverse)
3 library(data.table)
4 library(ggplot2)
5
6
7 # Set working directory
8 setwd("C:/Users/Bence")
9 getwd()
10
11 # Load required data tables
12 edges <- fread(file = "edges_out_all.csv")
13 distance <- fread(file = "city_distances.csv")
14 city <- fread(file = "city_1000_top_1000.csv")
15
16 colnames(edges) <- c("id1", "id2", "city_id", "distance")
17 head(edges)
```

R is a collaborative project with many contributors.  
Type `"contributors()"` for more information and  
`"citation()"` on how to cite R or R packages in publications.

Type `"demo()"` for some demos, `"help()"` for on-line help, or  
`"help.start()"` for an HTML browser interface to help.  
Type `"q()"` to quit R.

(workspace loaded from ~/RData)

User Library

- acepack ACE and ACO for Selecting Multiple Regression Transformations 1.4.1
- aod Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences 1.7.0
- AlgDesign Algorithmic Experimental Design 1.1-5.3
- assertthat Easy Pre-and Post Assertions 0.2.0
- backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.1
- base64enc Tools for Base64 encoding 0.1.3
- BH Boost C++ Header Files 1.62.0