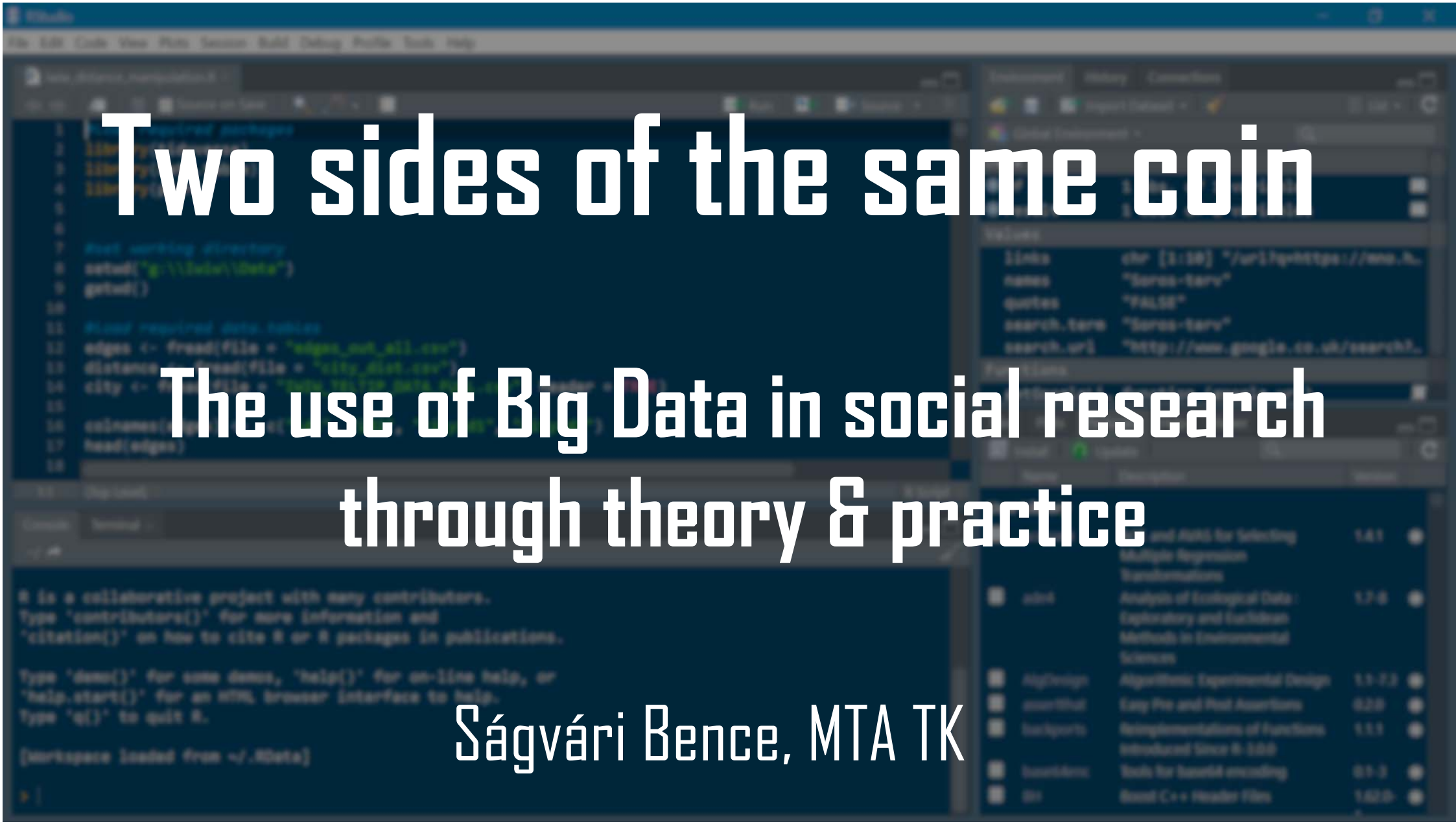
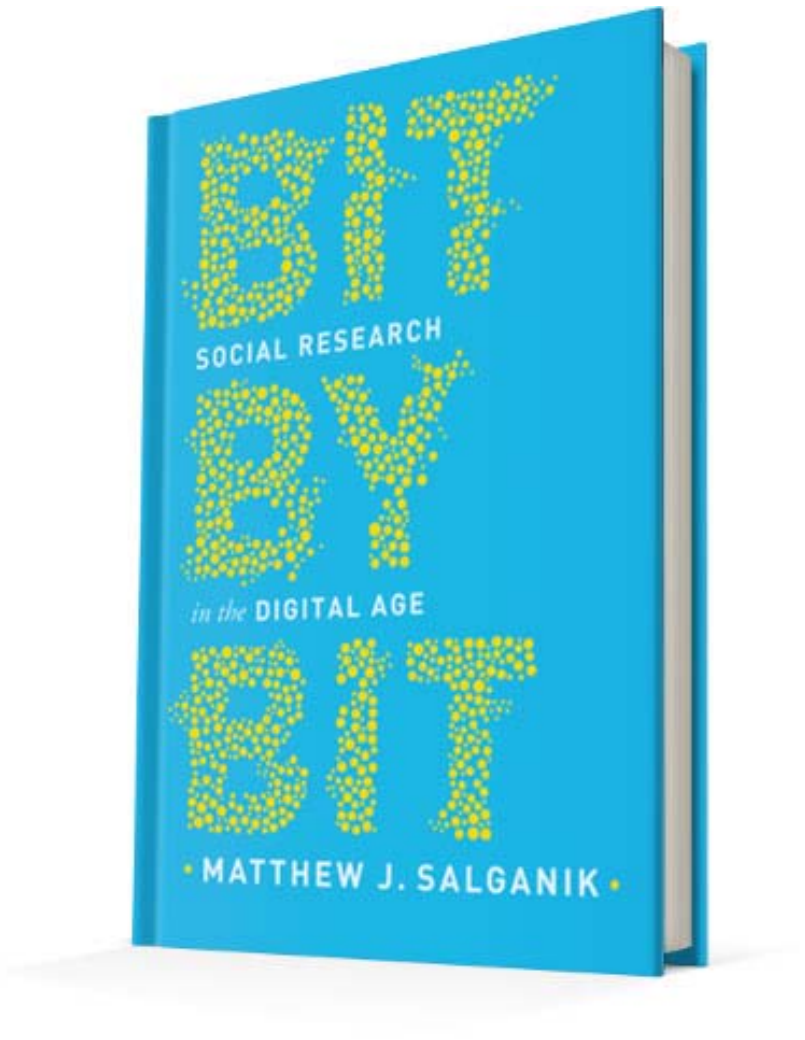


Two sides of the same coin

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

Ságvári Bence, MTA TK

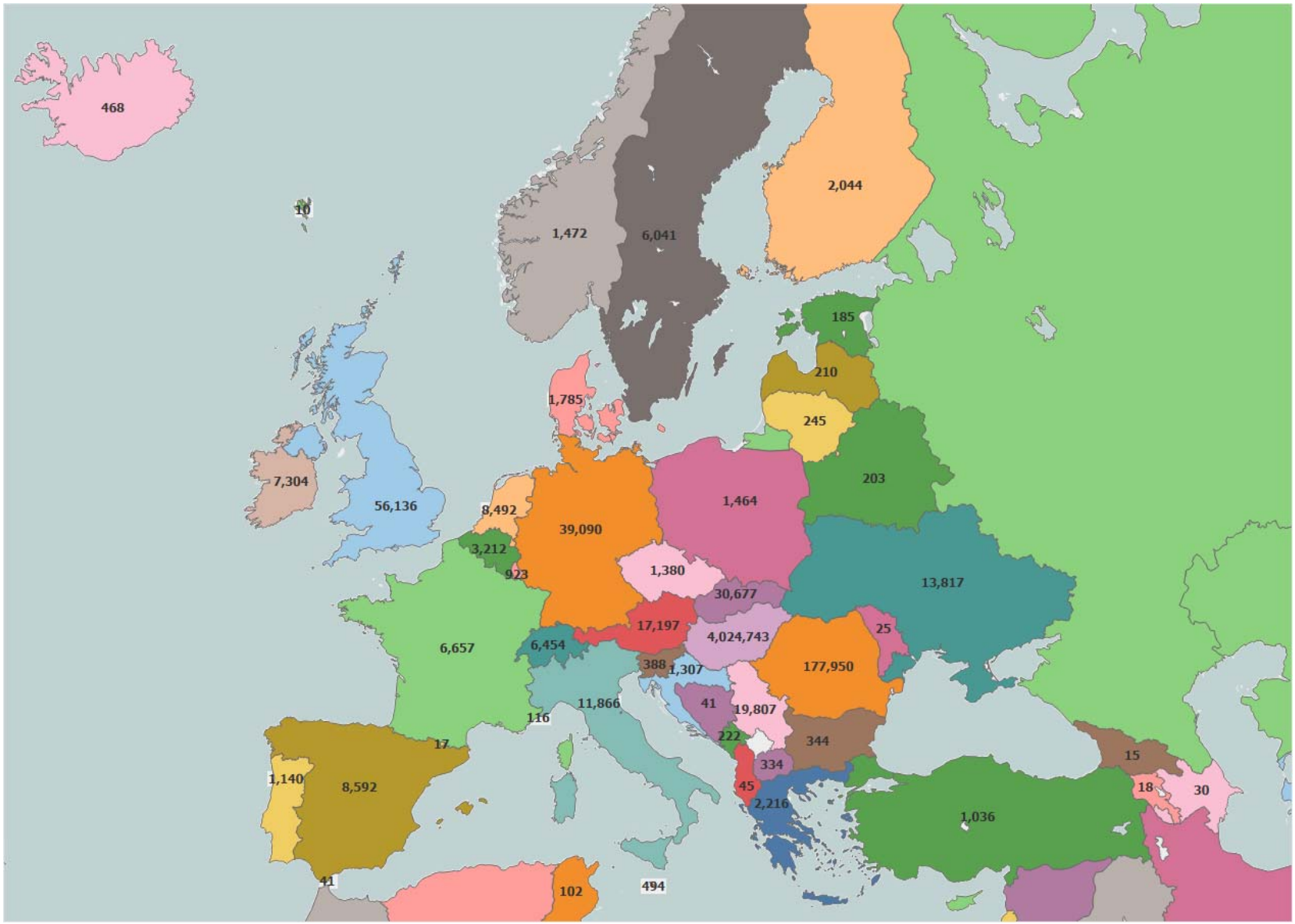


Ten common characteristics of big data

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."


468
10
1,472
6,041
2,044
185
210
1,785
245
203
7,304
56,136
8,492
1,464
39,090
3,212
923
1,380
13,817
30,677
17,197
4,024,743
25
6,657
6,454
388
1,307
177,950
11,866
41
19,807
116
222
344
17
15
1,140
8,592
334
45
18
30
2,216
1,036
41
102
494

2. ALWAYS-ON

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

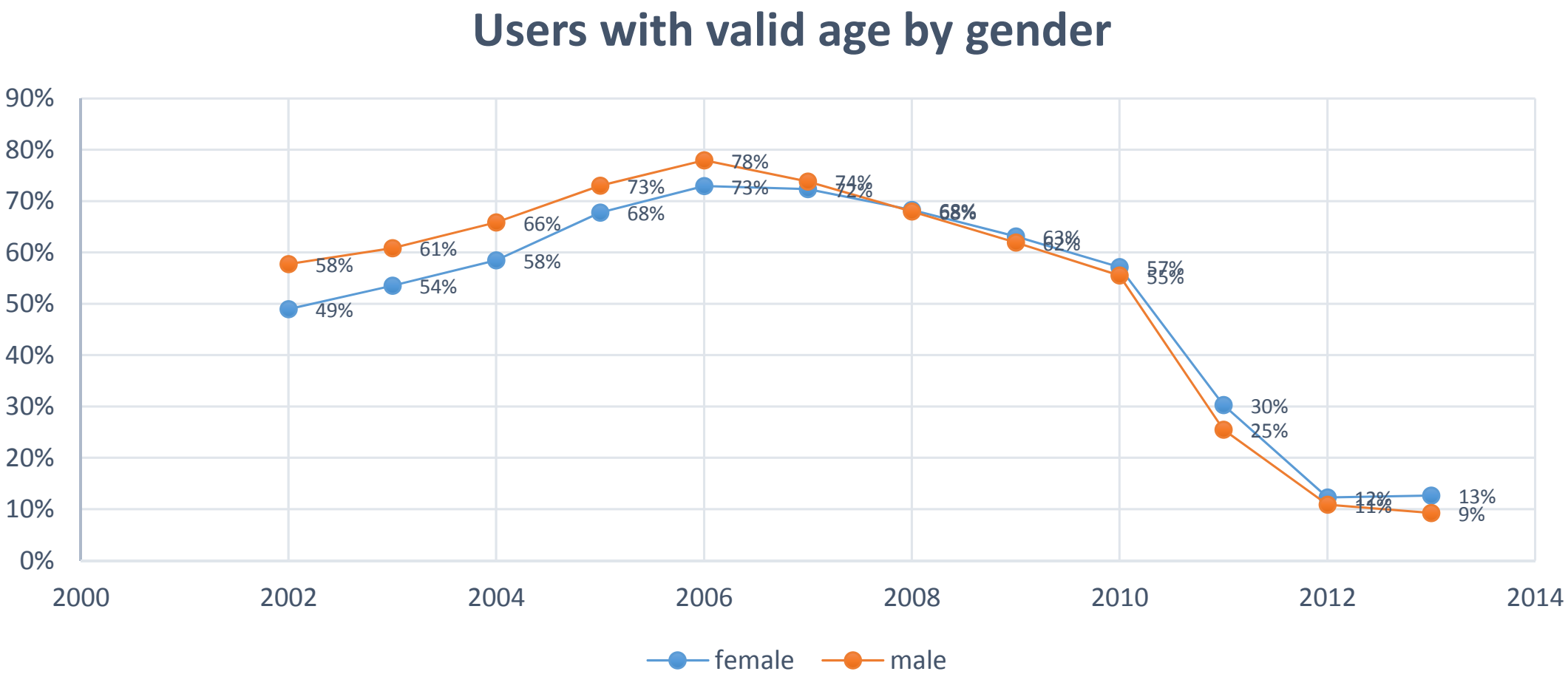
3. NONREACTIVE

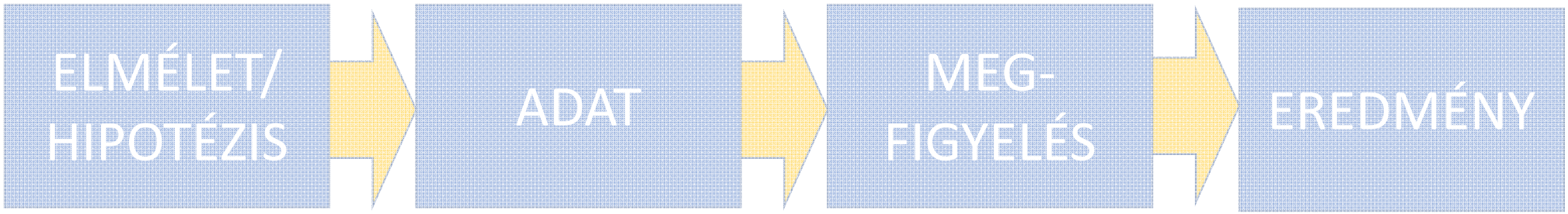
"Measurement in big data sources is much less likely to change behavior."

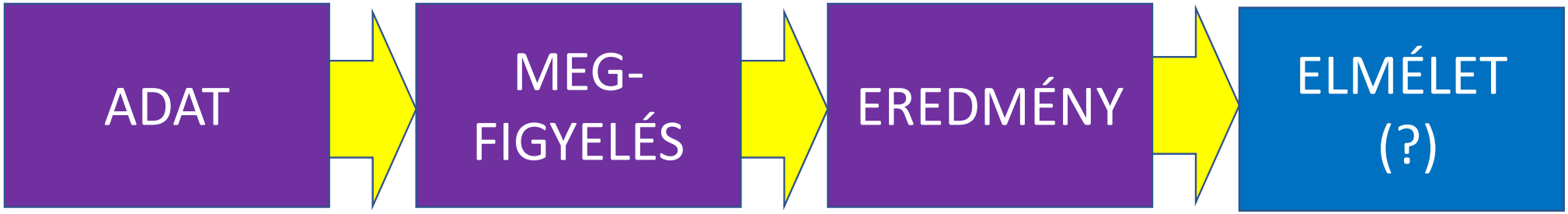
4. INCOMPLETE

"No matter how big your big data, it probably doesn't have the information you want."

Massive share of missing information on age




ELMÉLET/
HIPOTÉZIS
ADAT
MEG-
FIGYELÉS
EREDMÉNY


ADAT
MEG-
FIGYELÉS
EREDMÉNY
ELMÉLET
(?)

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 megsértéséért, ezekkel kapcsolatos kárért**, ha a titoksértés vagy kár az Üzleti titok megismerőjének felróható, 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.

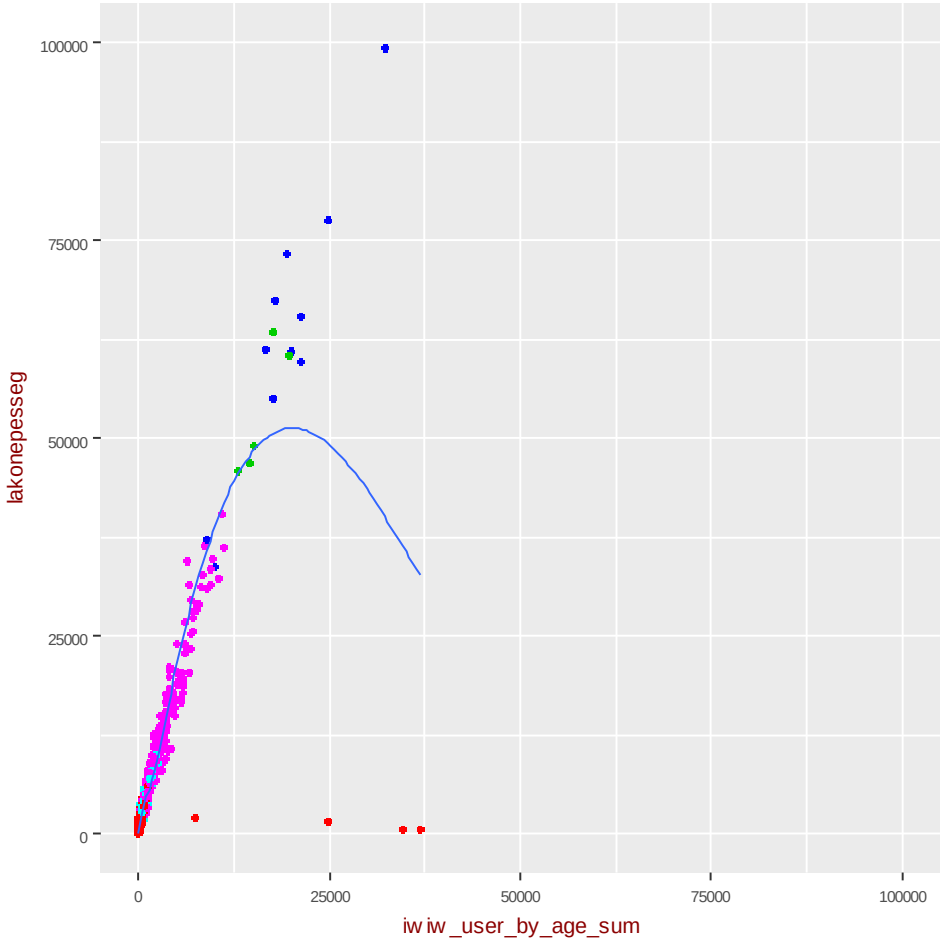
A tudományos kutatás eredményeit a Partner közölheti olyan módon, hogy a tudományos és ismeretterjesztő publikációkba üzleti titok nem kerülhet. **A publikációk közlésének menetét a Felek a kizárólagos felhasználási szerződésben fektetik le.**

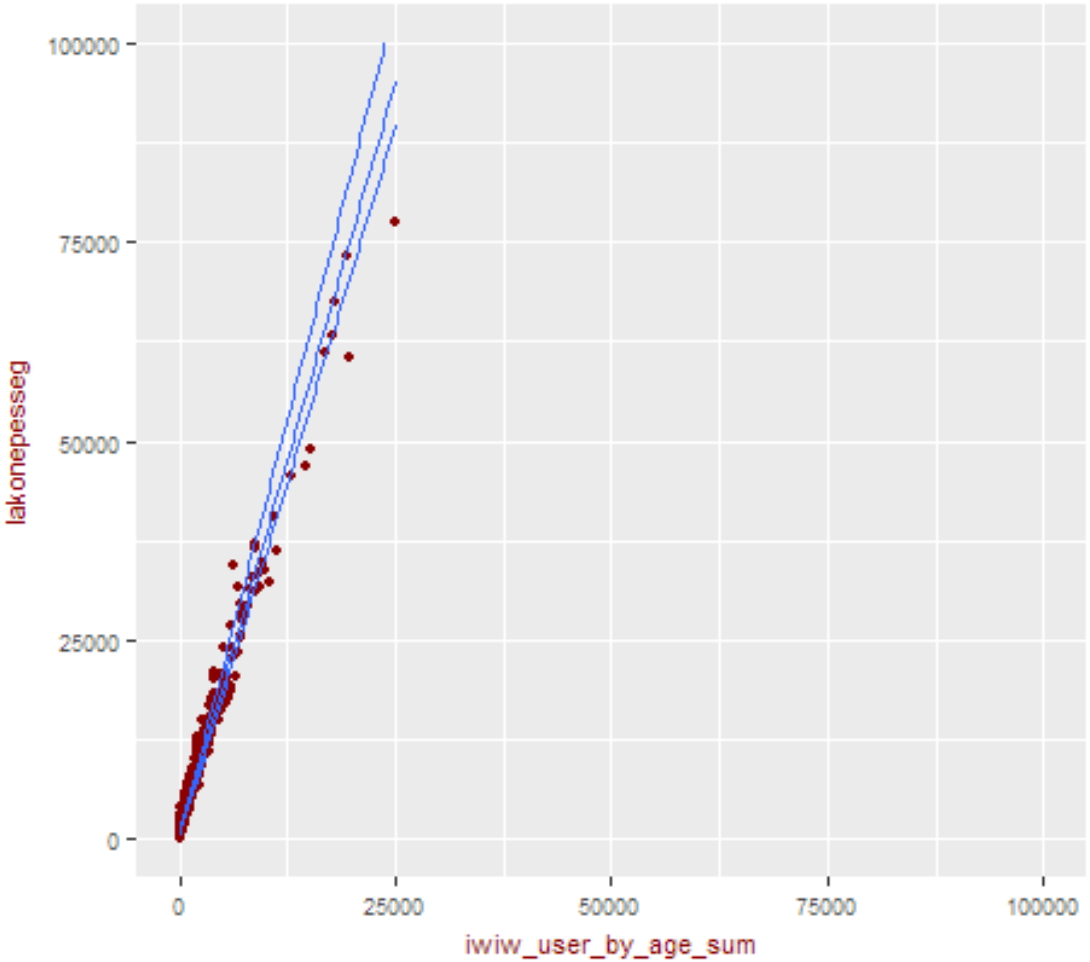
A jelen Megállapodás hatálya alatt az Üzleti titok jogosultja írásbeli kérésére az Üzleti titok megismerője a fizikailag birtokában lévő Üzleti **titkokat köteles az Üzleti titok jogosultjának visszaszolgáltatni vagy megsemmisíteni** (az Üzleti titok jogosultja utasítása szerint), annyiban amennyiben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

6. NONREPRESENTATIVE

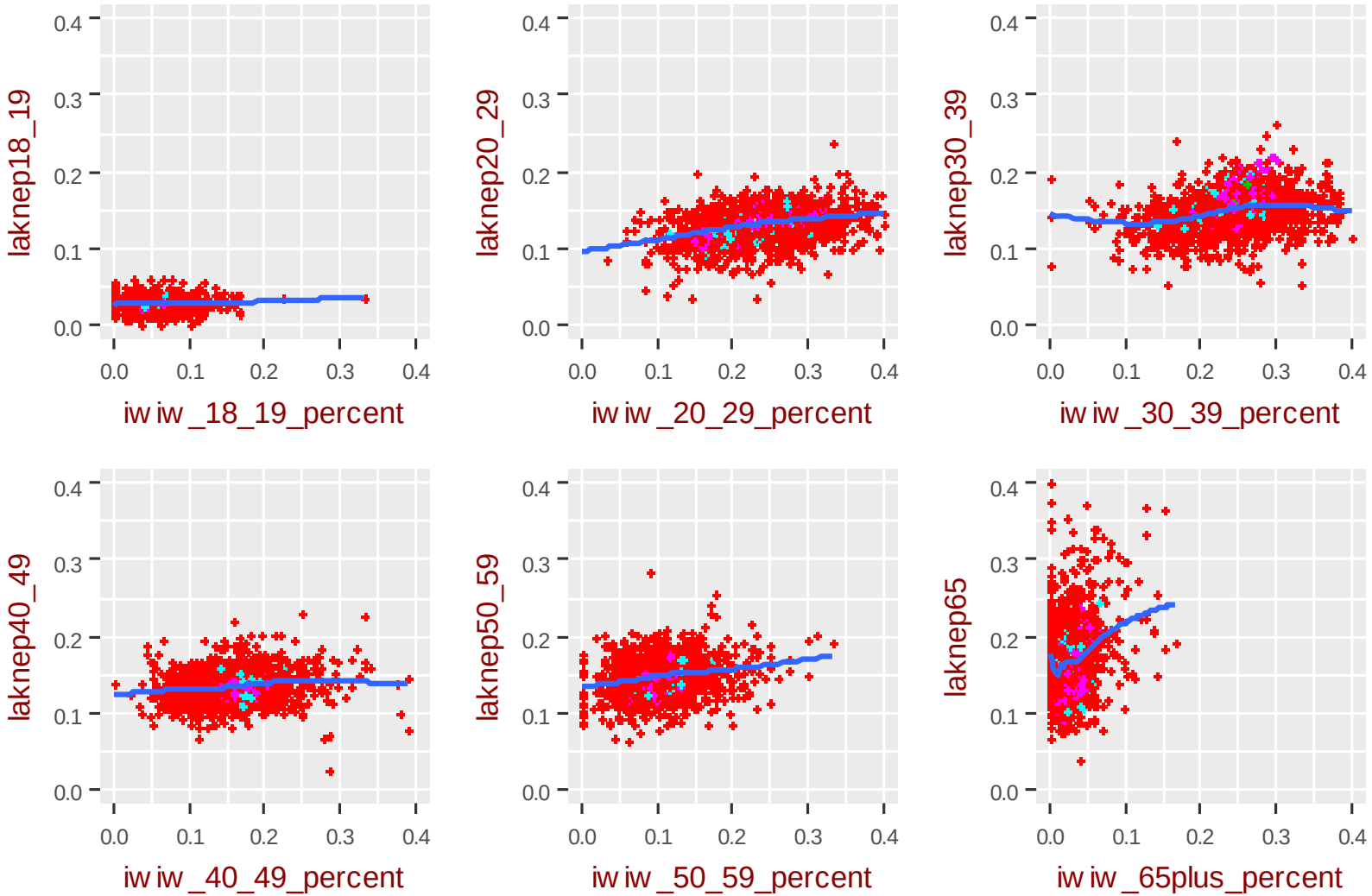
"Non-representative data are bad for out-of-sample generalizations, but can be quite useful for within-sample comparisons."

Real vs. iWiW population in Hungarian settlements





Age distribution of real vs. iWiW population in Hungarian settlements



7. DRIFTING

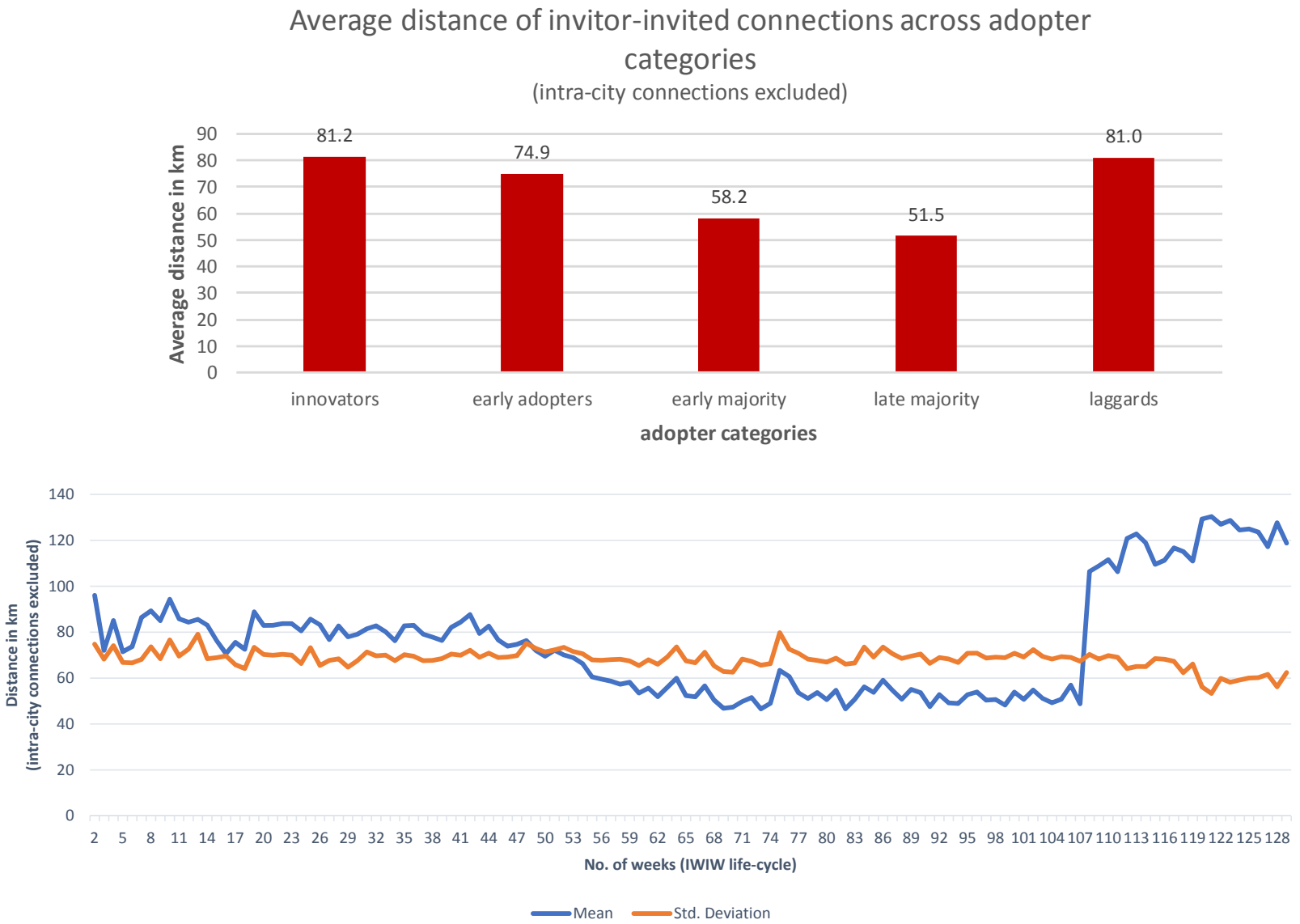
"Population drift, usage drift, and system drift make it hard to use big data sources to study long-term trends."

8. ALGORITHMICALLY COUNFOUNDED

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

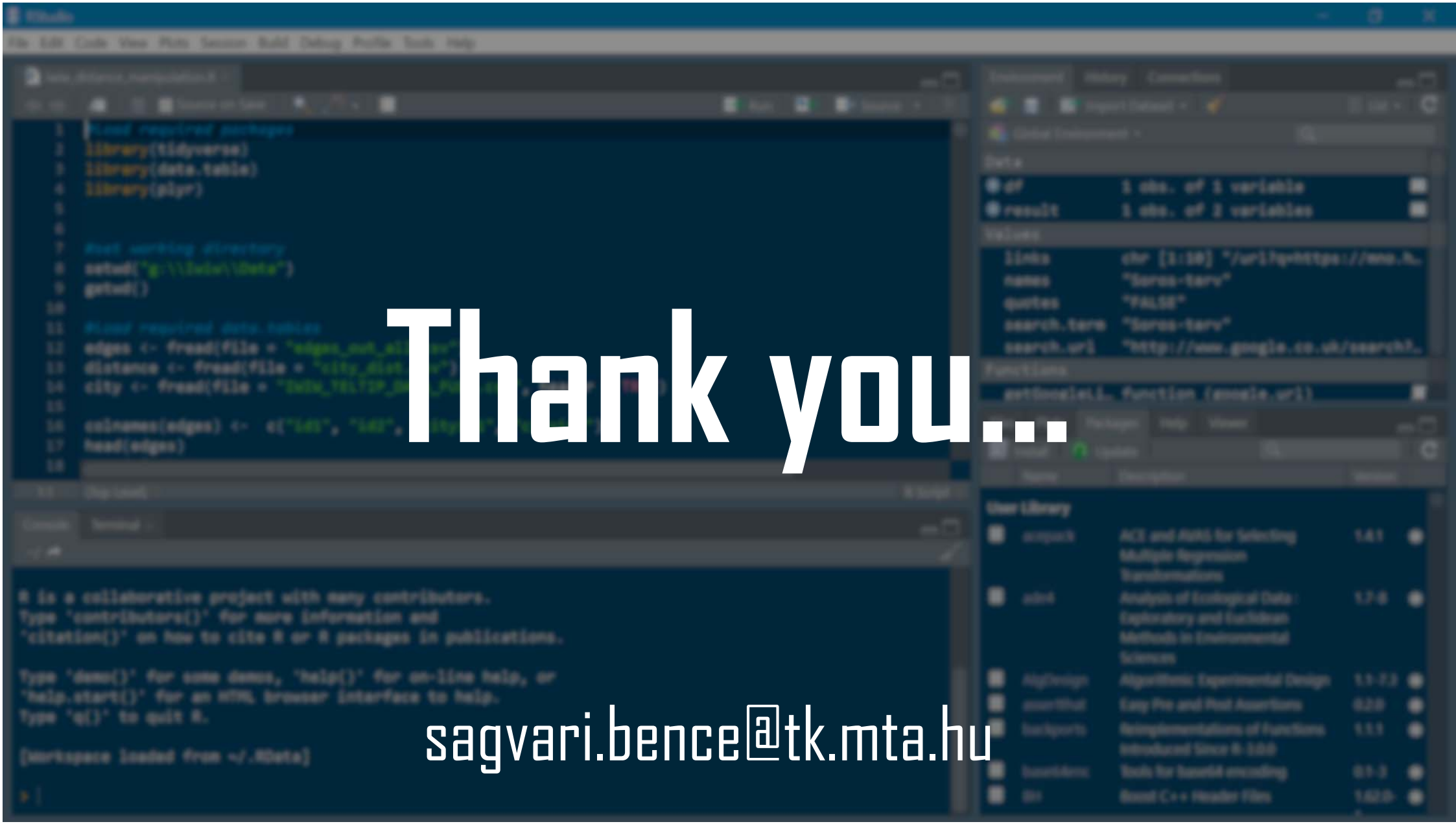
9. DIRTY

"Big data sources can be loaded with junk and spam."


Average distance of invitor-invited connections across adopter categories
(intra-city connections excluded)
Average distance in km
90
80
70
60
50
40
30
20
10
0
81.2
74.9
58.2
51.5
81.0
innovators
early adopters
early majority
late majority
laggards
adopter categories
Distance in km
(intra-city connections excluded)
140
120
100
80
60
40
20
0
2 5 8 11 14 17 20 23 26 29 32 35 38 41 44 47 50 53 56 59 62 65 68 71 74 77 80 83 86 89 92 95 98 101 104 107 110 113 116 119 122 125 128
No. of weeks (IWIW life-cycle)
Mean
Std. Deviation

10. SENSITIVE

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


Thank you...
sagvari.bence@tk.mta.hu