UNIVERSITÄT
BIELEFELD

Fakultät für
Wirtschaftswissenschaften

Our contact points with big data in empirical methods

BGTS Doctoral Day 2022

Lennart Oelschläger, Carlina Feldmann & Jonas Bauer

Bielefeld, October 7, 2022

1. The empirical methods group

Empirical methods is a research and teaching center

- Data Science (Head Prof. Fuchs)
- Statistics and data analysis (Head Prof. Langrock)
- Econometrics (Head Prof. Bauer)

UNIVERSITÄT BIELEFELD
Fakultät für Wirtschaftswissenschaften

1. The empirical methods group

Empirical methods is a research and teaching center

- Data Science (Head Prof. Fuchs)
- Statistics and data analysis (Head Prof. Langrock)
- Econometrics (Head Prof. Bauer)

The three V's (Laney, 2001)

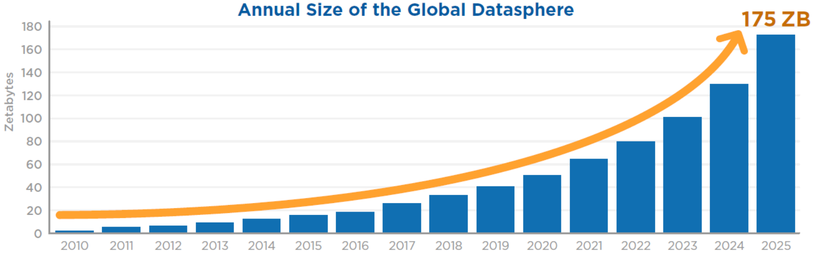
- Volume
  - *"the volume criterion is met if the dataset is such that we cannot collect, store, and analyze it using traditional computing and statistical methods"*
  - Moore's Law: number of transistors on microchips doubles every two years
- Variety
  - structured (spreadsheets, databases) and unstructured data (photos, tweets)
- Velocity
  - continuous streams from the Web, smartphones, sensors, Teslas

The three V's (Laney, 2001)

- Volume
  - *"the volume criterion is met if the dataset is such that we cannot collect, store, and analyze it using traditional computing and statistical methods"*
  - Moore's Law: number of transistors on microchips doubles every two years
- Variety
  - structured (spreadsheets, databases) and unstructured data (photos, tweets)
- Velocity
  - continuous streams from the Web, smartphones, sensors, Teslas

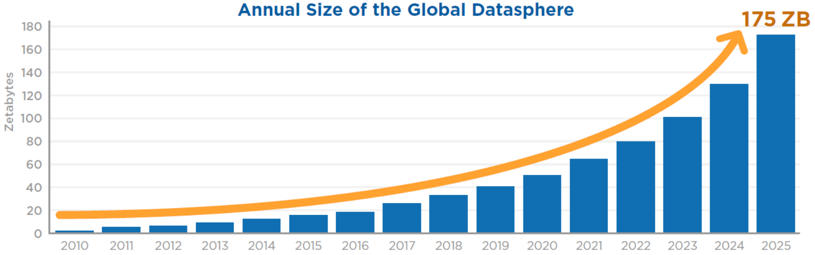
4. Value
   - pattern hidden in and knowledge gained from big data
5. Veracity
   - trust in and quality of data sources

How big is big?



Source: The digitization of the world from edge to core (Rydning et al., 2018, IDC)

How big is big?

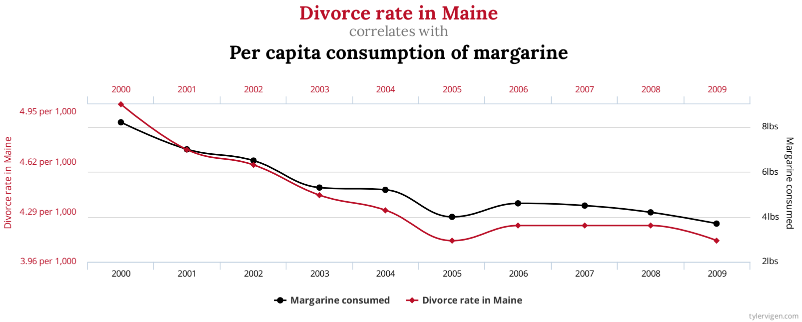


Source: The digitization of the world from edge to core (Rydning et al., 2018, IDC)

1 ZB $= 10^{12}$ GB $= 10^{21}$ bytes

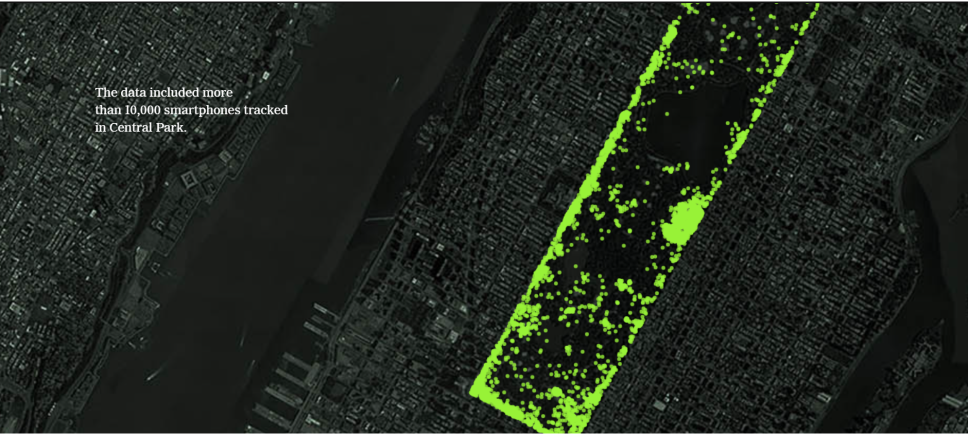
If one attempted to download 80 ZB with 50 Mbit/s, it would take about 400 million years.

With enough data, the numbers speak for themselves?



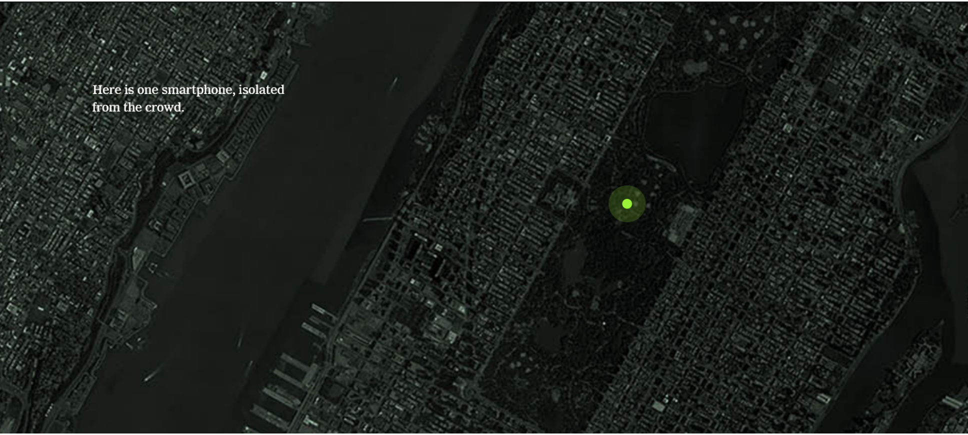
Calude et al. (2016) shows (using ergodic theory, Ramsey theory and algorithmic information theory) that *"very large databases have to contain arbitrary (spurious) correlations. These correlations appear only due to the size, not the nature, of the data."*

Privacy in times of big data



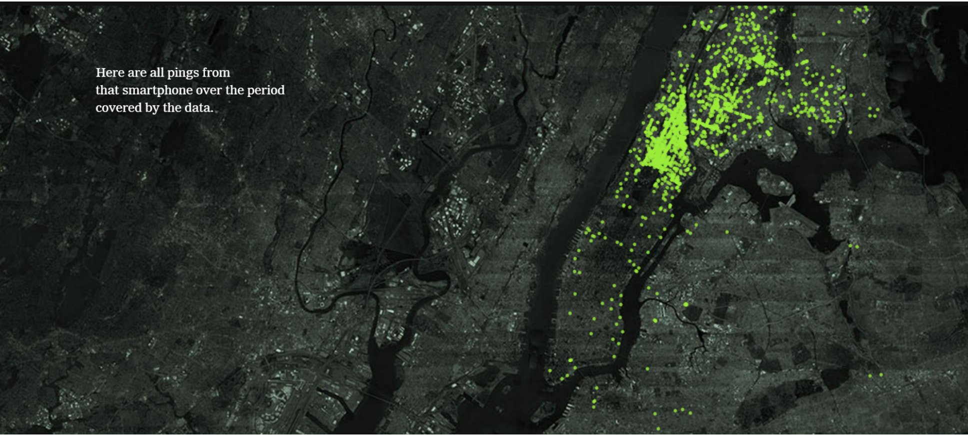
Source: "Twelve Million Phones, One Dataset, Zero Privacy" (The New York Times, 2019)

Privacy in times of big data



Source: "Twelve Million Phones, One Dataset, Zero Privacy" (The New York Times, 2019)

Privacy in times of big data

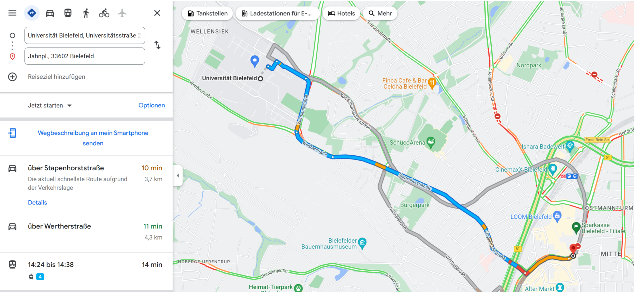


Source: "Twelve Million Phones, One Dataset, Zero Privacy" (The New York Times, 2019)

Privacy in times of big data

Source: "Twelve Million Phones, One Dataset, Zero Privacy" (The New York Times, 2019)

The value of route choice data



- Google aggregates smartphone data to suggest routes and predict traffic (somehow).
- Can we personalize the ranking of travel options with individual characteristics (daily schedule, green-life propensity, traffic jam aversion)?

The MOP data

- German mobility panel from Karlsruhe Institute of Technology
- from 1994 to 2013:
  - 8722 households, 15864 individuals, 230769 daily mobility diaries
  - numerous sociodemographic data (age, sex, employment, education, ...)
- goal: explain, predict, and simulate mobility behavior

The MOP data

- German mobility panel from Karlsruhe Institute of Technology
- from 1994 to 2013:
  - 8722 households, 15864 individuals, 230769 daily mobility diaries
  - numerous sociodemographic data (age, sex, employment, education, ...)
- goal: explain, predict, and simulate mobility behavior

Stadtbahn- und Busverkehr in Bielefeld soll ausgebaut werden

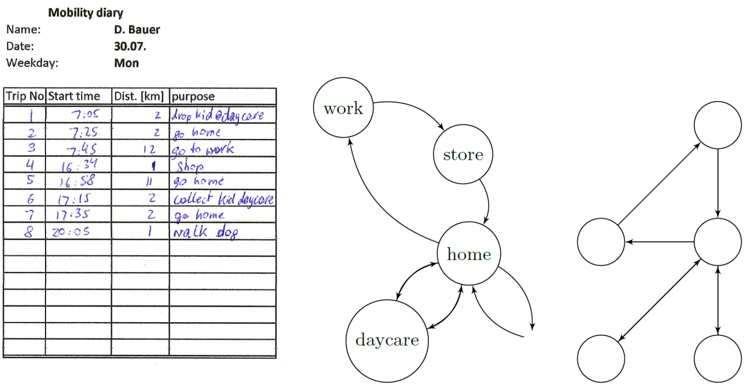
am Freitag, 10.12.2021 | Lokalnachrichten

Im Bielefelder Rat haben SPD, Grünen und Linken den neuen Nahverkehrsplan beschlossen, gegen die Stimmen der Opposition. Der Stadtbahn- und Busverkehr soll deutlich ausgebaut werden. Ein dreistelliger Millionenbetrag wird in den kommenden zehn Jahren investiert. Die CDU scheiterte mit ihrem Antrag auf ein ganzheitlicheres Verkehrskonzept.

alle Lokalnachrichten

Source: Radio Bielefeld (2021)

Motifs



Mobility diary

Name: **D. Bauer**
Date: **30.07.**
Weekday: **Mon**

Trip No	Start time	Dist. [km]	purpose
1	7:05	2	drop kid @daycare
2	7:25	2	go home
3	7:45	12	go to work
4	16:34	1	shop
5	16:58	11	go home
6	17:15	2	collect kid daycare
7	17:35	2	go home
8	20:05	1	walk dog

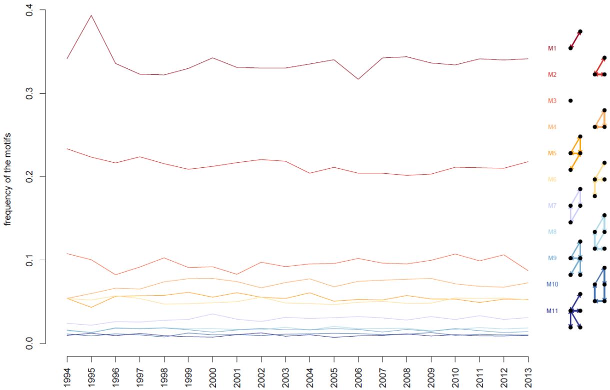
(A) Example page from a stylized mobility diary

(B) Resulting mobility graph

(C) Resulting motif

Source: Büscher et al. (2019)

Temporal stability



Source: Büscher et al. (2019)

Discrete choice models

Probit model of decider n's utility at time t for alternative j

$$
\begin{aligned}
U_{ntj} &= X_{nt}\beta_n + \epsilon_{ntj} && \text{(latent utilities)} \\
\beta_n &\sim N(0, \Omega) && \text{(mixing distribution)} \\
\epsilon_{nt:} &\sim N(0, \Sigma) && \text{(residual)} \\
y_{nt} &= \arg\max U_{ntj} && \text{(observed choice)}
\end{aligned}
$$

Discrete choice models

Probit model of decider n's utility at time t for alternative j

$$\begin{aligned} U_{ntj} &= X_{nt}\beta_n + \epsilon_{ntj} && \text{(latent utilities)} \\ \beta_n &\sim N(0,\Omega) && \text{(mixing distribution)} \\ \epsilon_{nt:} &\sim N(0,\Sigma) && \text{(residual)} \\ y_{nt} &= \arg\max U_{ntj} && \text{(observed choice)} \end{aligned}$$

Choice probability

$$\Pr(y_{n:}) = \int 1(y_{n:} = \arg\max U_{n:j})\phi(\epsilon_{n:})\ d\epsilon_{n:}$$

(dimension "time points $\times$ (alternatives -1)")

Discrete choice models

Probit model of decider n's utility at time t for alternative j

$$\begin{aligned} U_{ntj} &= X_{nt}\beta_n + \epsilon_{ntj} && \text{(latent utilities)} \\ \beta_n &\sim N(0,\Omega) && \text{(mixing distribution)} \\ \epsilon_{nt:} &\sim N(0,\Sigma) && \text{(residual)} \\ y_{nt} &= \arg\max U_{ntj} && \text{(observed choice)} \end{aligned}$$

Choice probability

$$\Pr(y_{n:}) = \int 1(y_{n:} = \arg\max U_{n:j})\phi(\epsilon_{n:})\ d\epsilon_{n:}$$

(dimension "time points $\times$ (alternatives -1)")

Composite marginal likelihood (CML) and CDF approximation

$$L = \prod_n \Pr(y_{n:}) \approx \prod_n \prod_{(t_1,t_2)} \Pr(y_{nt_1}, y_{nt_2}) \approx \prod_n \prod_{(t_1,t_2)} \tilde{\Pr}(y_{nt_1}, y_{nt_2})$$

Current research

DFG Project of Dietmar Bauer: *Using the Composite Likelihood Methods for Estimation of Probit models*

Manuel Batram	Approximation of Gaussian CDF and their statistical implications (asymptotic bias, relative efficiency depending on the CML, model selection procedures)
Sebastian Büscher	Weighting of pairs in CML
myself	Initialization of likelihood optimization and Bayesian alternative

4. Big movement data

- high-throughput tracking systems: temporal resolution, tracking duration & concurrency, cost-effectiveness
- Nyquist–Shannon sampling theorem: to characterize a signal of the duration $2\delta t$, an observation frequency of δt is needed

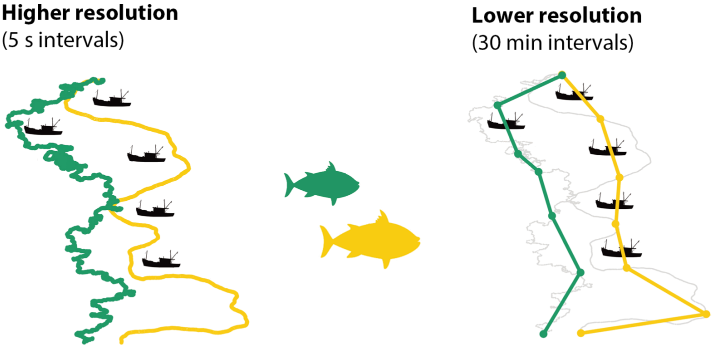
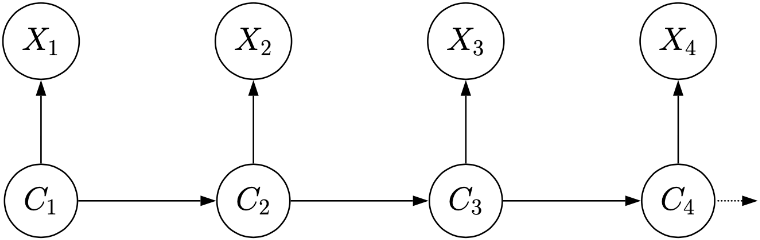


Figure: Nathan et al. 2022

4. Big movement data

- costs of big-data
  - data management and processing
  - challenging statistical analyses, in time series especially because of autocorrelation
- hidden Markov models: assume underlying Markov chain (of behavioural states) determining distributions of observations

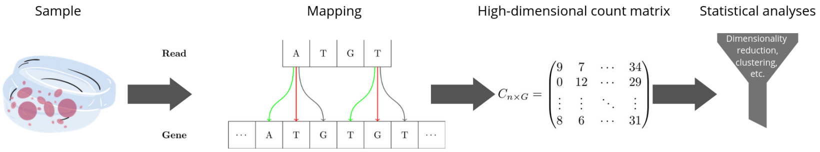


5. Big data in biological data analyses

Goal: distinguish & classify diseased vs. healthy individuals/samples

- many diseases origin in the wrong concentration of expressed genes (GELs)
- technological advancements enabled study of single cells (heuristically)
- Homo Sapiens has 20000 genes, 250000 transcripts and a lot more exons[1]

$\Rightarrow$ big data with $p >> n$



[1] Ensembl primary assembly

Challenges:

- a priori unknown number of classes due to cell heterogeneity
- $p(\theta|C)$ is important $\Rightarrow$ Markov chain Monte Carlo
- curse of dimensionality since $\theta = (\theta_1, \ldots, \theta_K, \pi)^T \in [0,1]^{K \cdot G + K}$
- MCMC costs proportional to parameter dimension (e. g. gradient calculation)
- sampling produces $N \times K \cdot G + K$ dimensional matrix

Challenges:

- a priori unknown number of classes due to cell heterogeneity
- $p(\theta|C)$ is important $\Rightarrow$ Markov chain Monte Carlo
- curse of dimensionality since $\theta = (\theta_1, \ldots, \theta_K, \pi)^T \in [0,1]^{K \cdot G + K}$
- MCMC costs proportional to parameter dimension (e. g. gradient calculation)
- sampling produces $N \times K \cdot G + K$ dimensional matrix

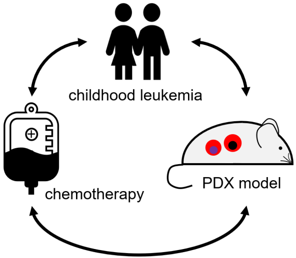
"Solutions":

- blocking, i.e. change only a subset of the parameters per iteration
- penalized gene-specific step lengths ϵ, e. g. via $||\epsilon||_1$
- boosting, e. g. by treating class-specific distributions as "weak learner"

6. The opposite of big data

The model:

$$dx_1(t) = (r_1 - r_2)x_1(t)dt, \quad x_1(0) = x_1,$$
$$dx_2(t) = (r_3 - r_4)x_2(t)dt, \quad x_2(0) = x_2$$

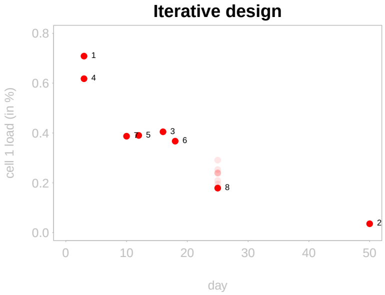


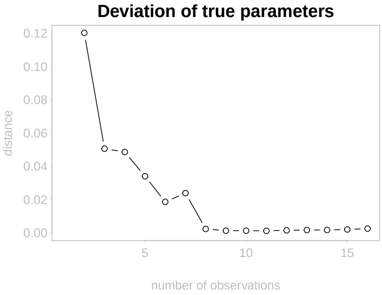
Research questions: How to choose measurement points s.t.

- knowledge is maximal &
- resources spent are minimal?

Iterative point selection:

- Which next point brings me closest to the DGP?
- repeat *n* times
- no guarantee for global optimum

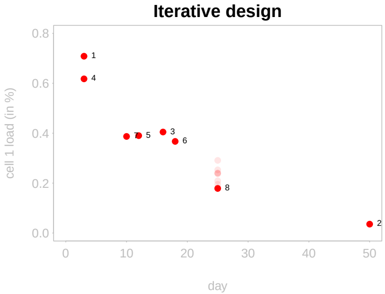


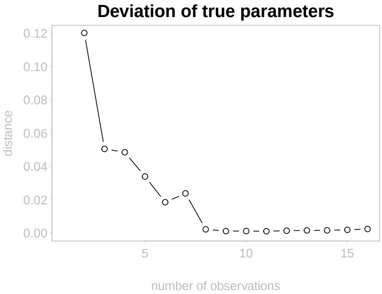


6. The opposite of big data

Iterative point selection:

- Which next point brings me closest to the DGP?
- repeat *n* times
- no guarantee for global optimum





Best subset selection:

- Which subset of time points yields minimal costs?
- costly calculations

Thanks for your attention!

Your thoughts on big data?