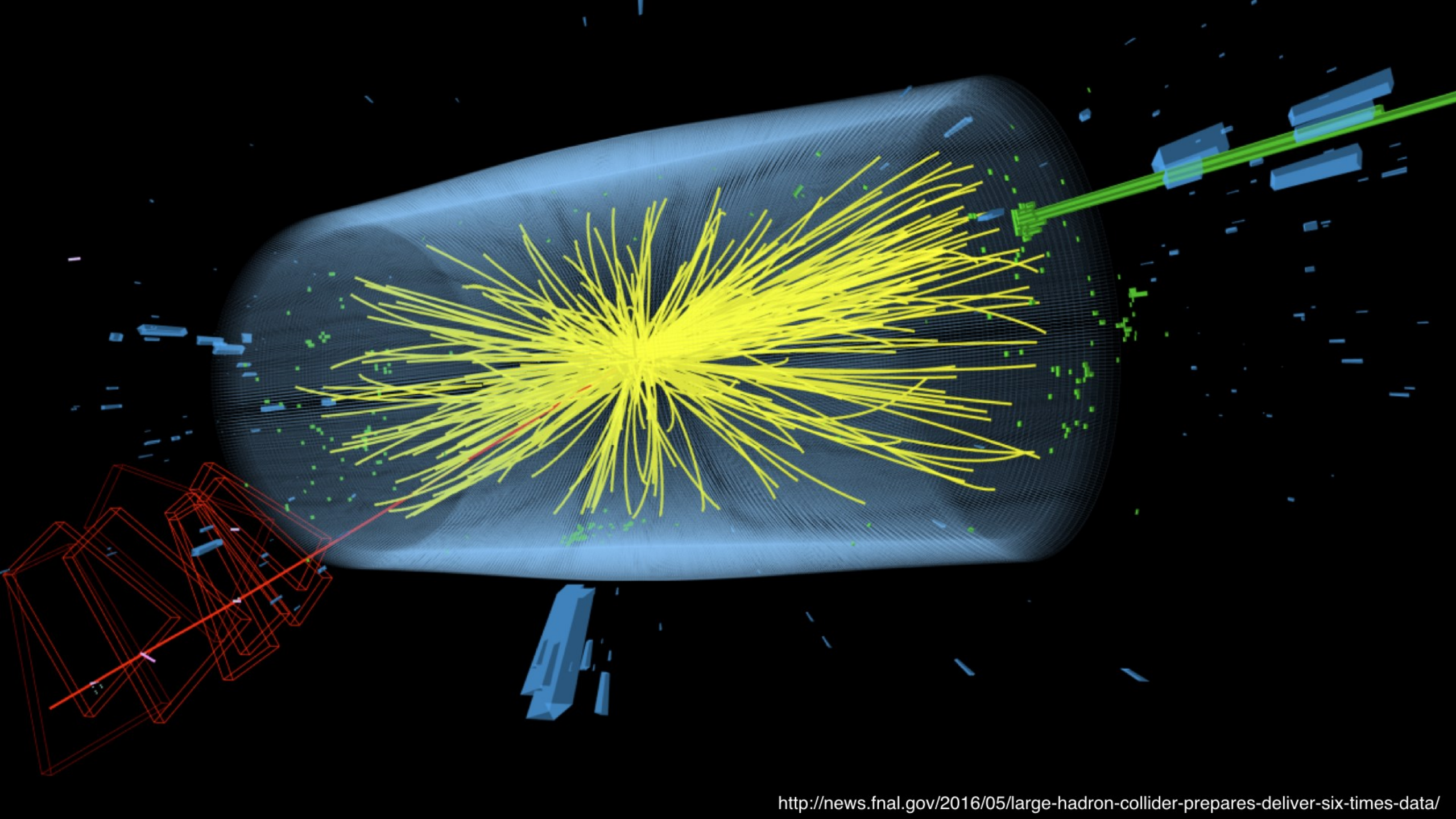
Social Data Science

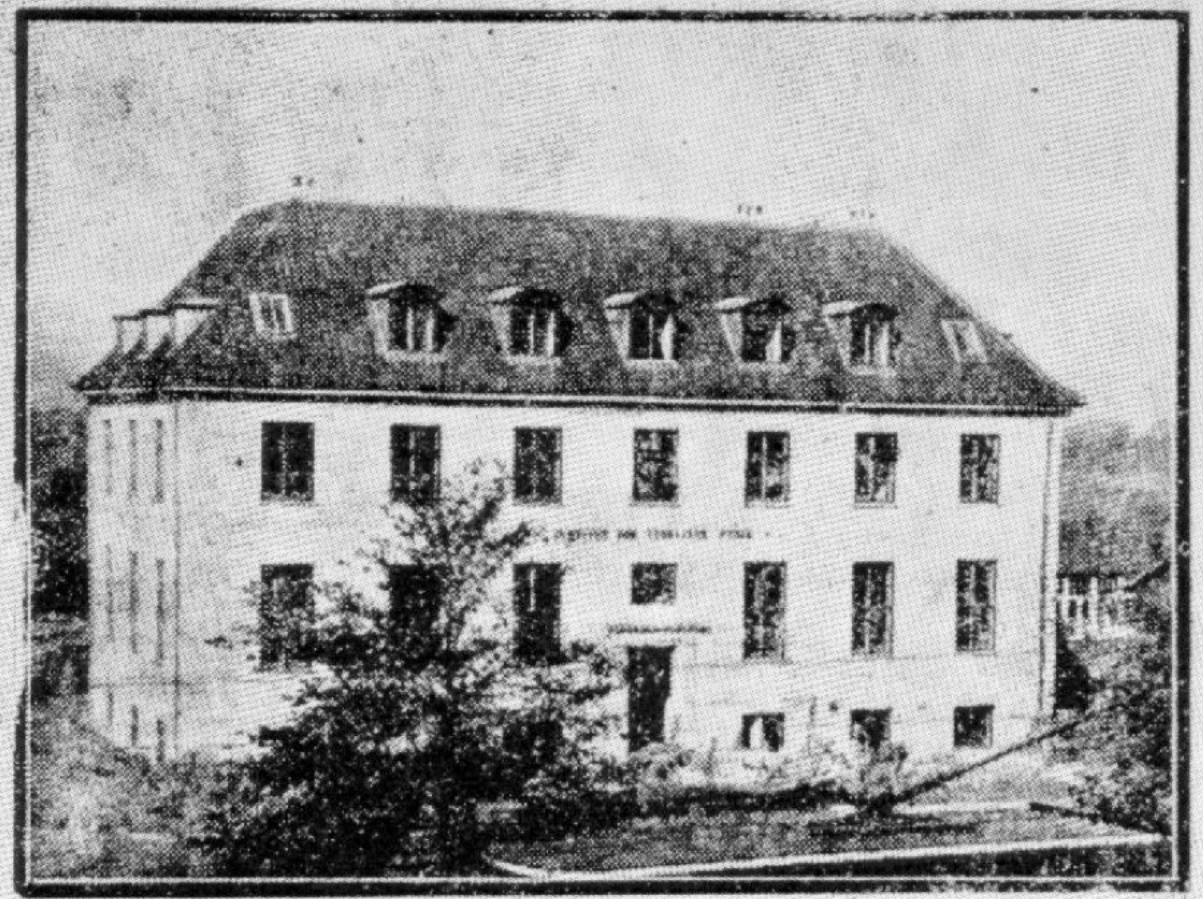
Sune Lehmann
- Associate Professor, DTU Compute. Technical University of Denmark.
- Adjunct Associate Professor, Niels Bohr Institute. University of Copenhagen
- @suneman



http://news.fnal.gov/2016/05/large-hadron-collider-prepares-deliver-six-times-data/

But social data is different.

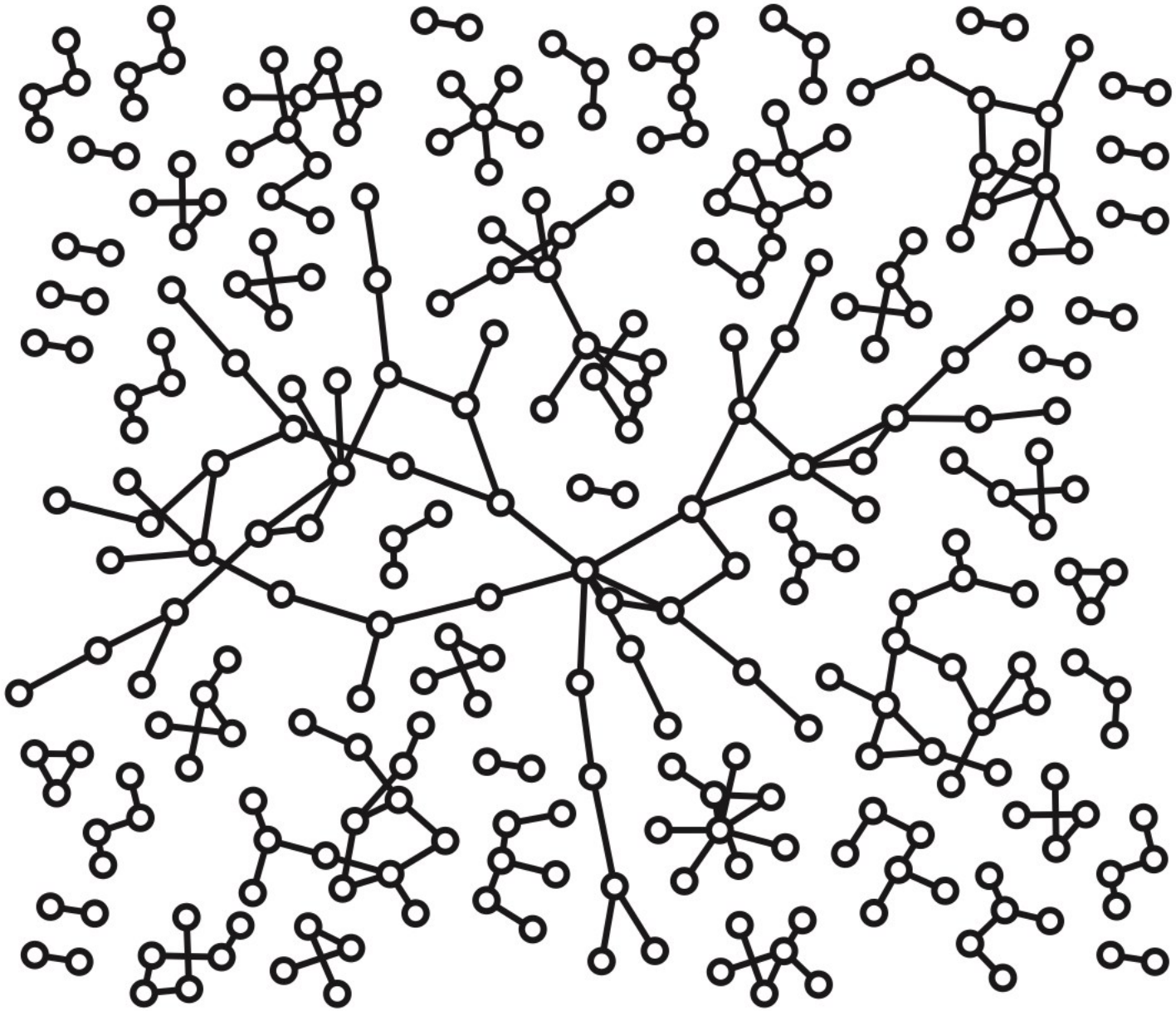
En ny Læreanstalt.

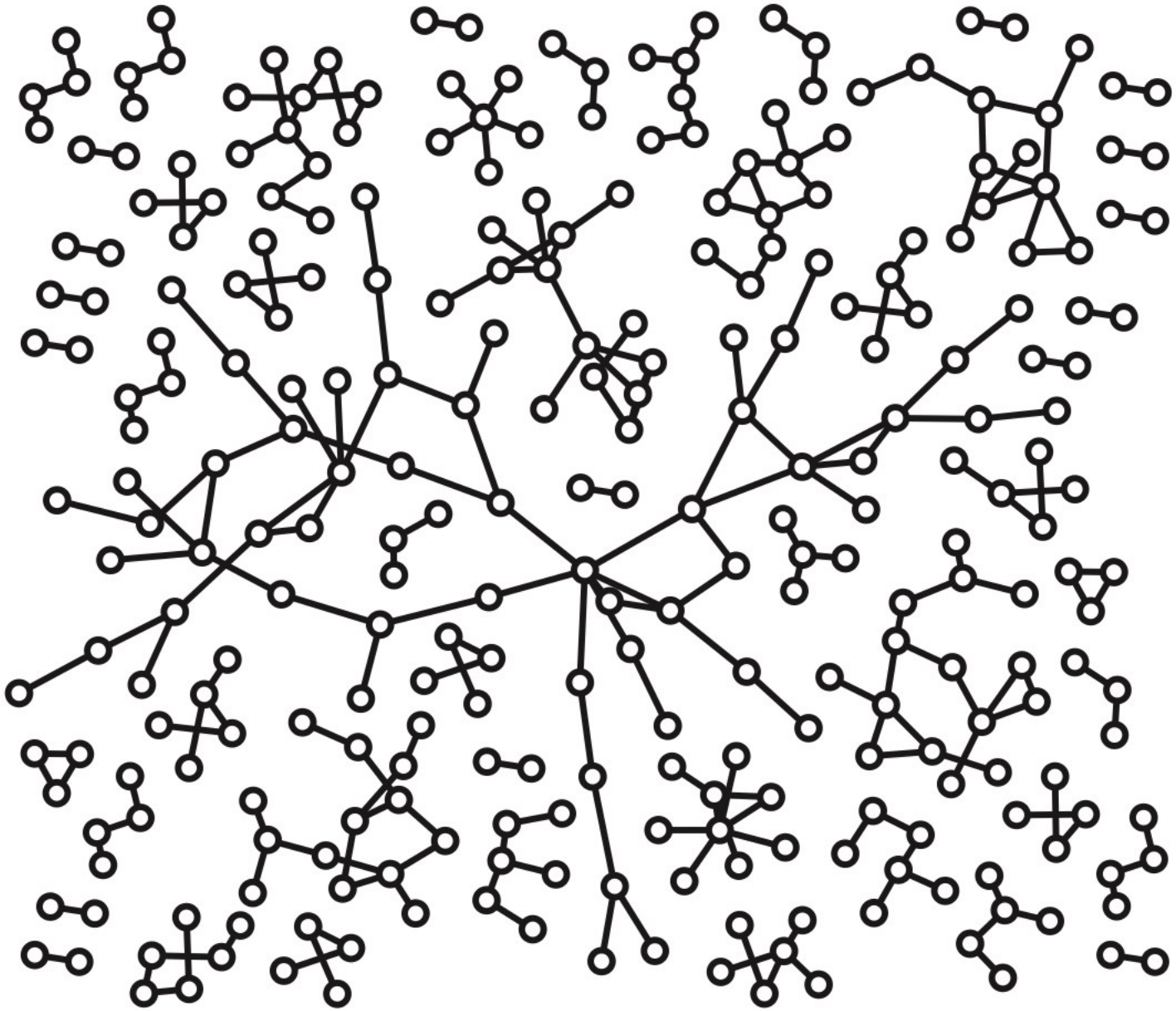


Vi bringer her et Billede af det nye Institut for teoretisk Fysik paa Blegdamsvej, der nu staar omtrent færdigt til at tages i Brug.

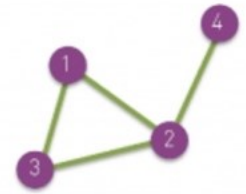
Fra B.T. (1920).

http://www.nbi.dk/NBI-HISTORY/whouse.jpg





a. Undirected

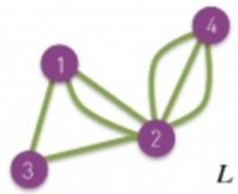


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{2L}{N}$$

c. Multigraph
(undirected)

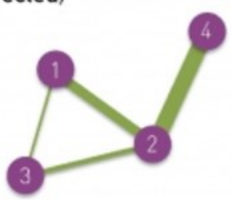


$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{2L}{N}$$

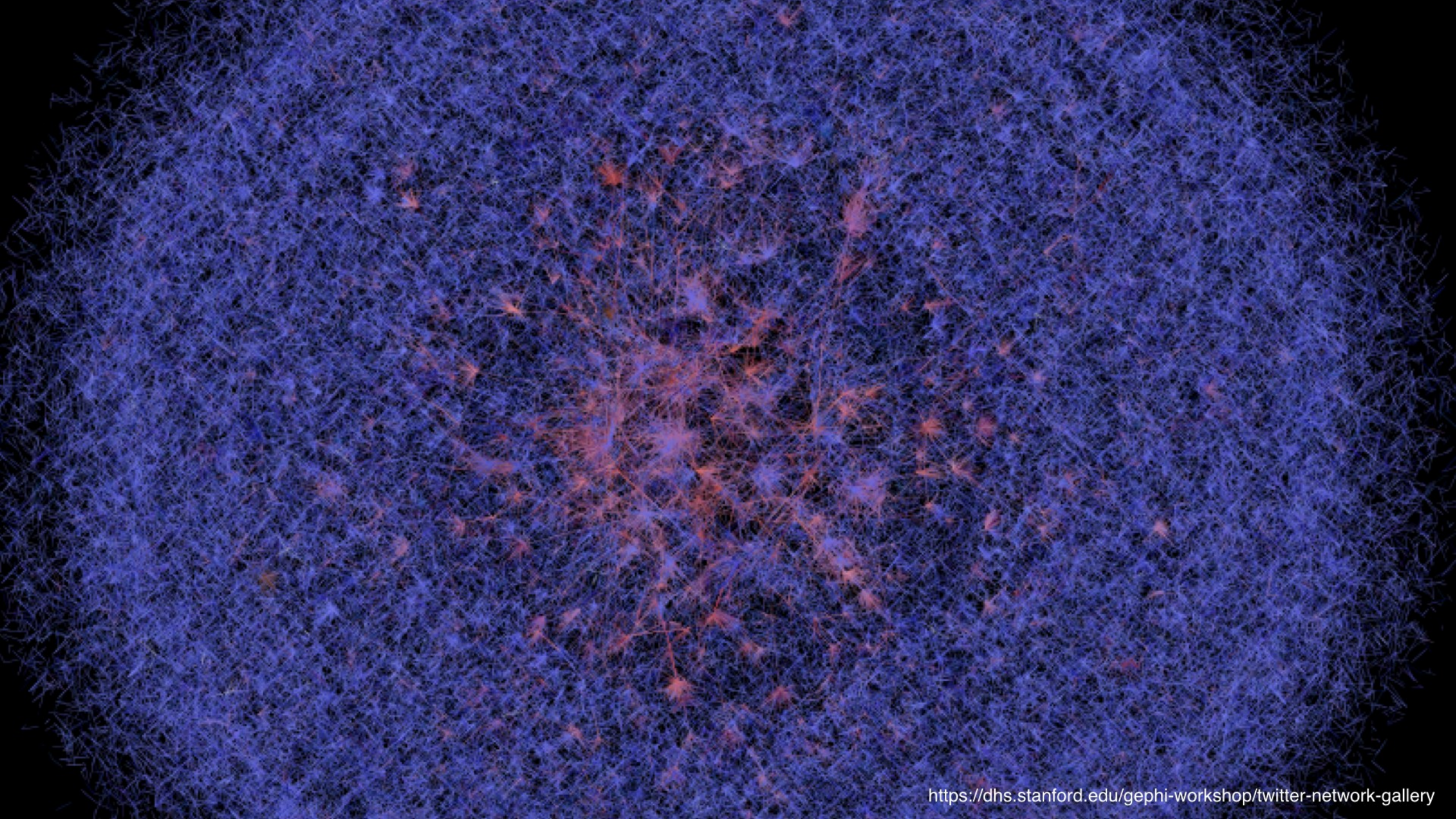
e. Weighted
(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$< k >= \frac{2L}{N}$$



https://dhs.stanford.edu/gephi-workshop/twitter-network-gallery

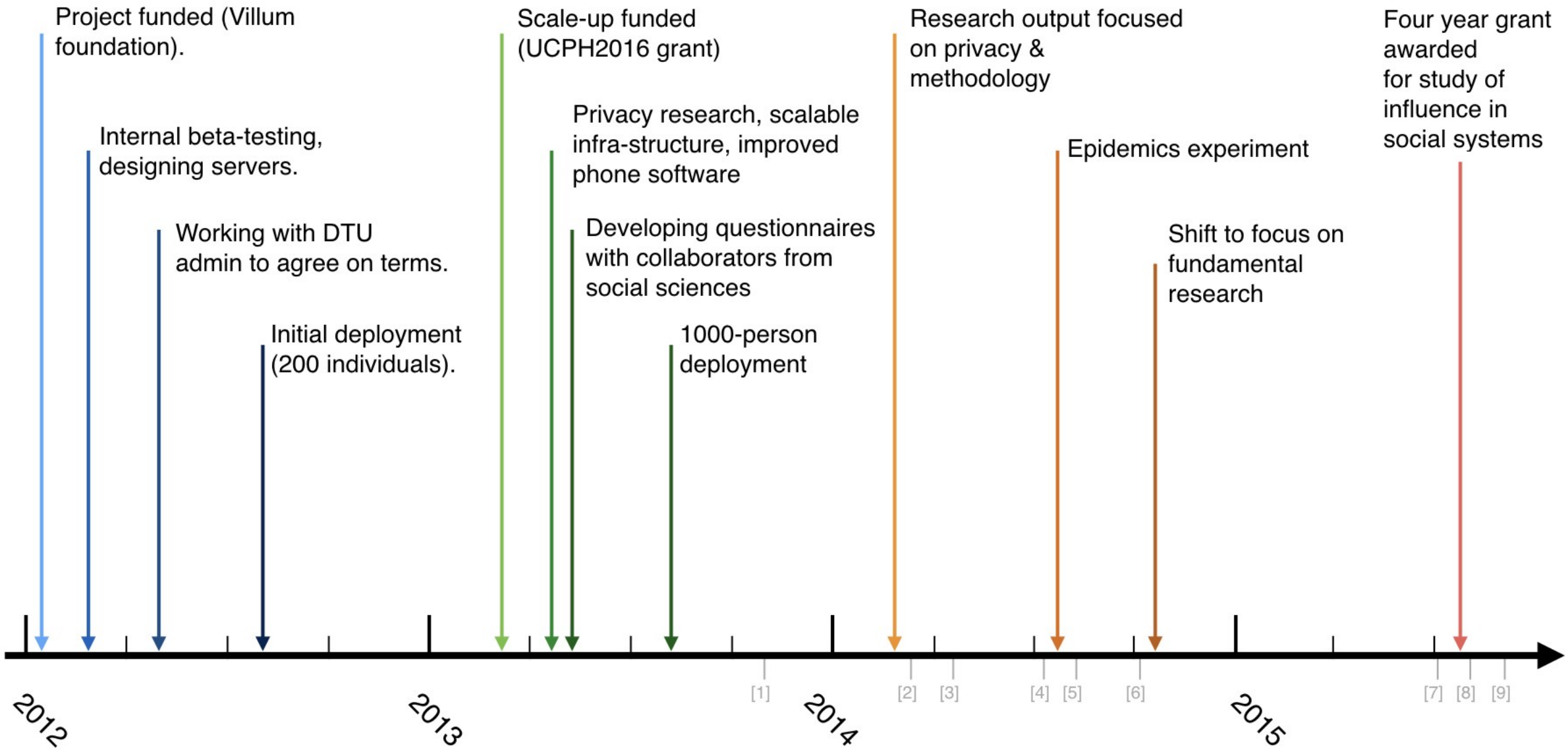

SAMSUNG
Galaxy S8


Pizza Pizza Pizza

Data types

- Physical proximity (Bluetooth)
- Telecommunication network (phone calls, SMS)
- Facebook friendships & interactions
- Geolocation (GPS & WiFi)
- A few more nerdy things
- Also did extensive battery of paneled personality tests (Big 5, narcissim, etc, etc)

[data collection for approximately 2 years]


Project funded (Villum foundation).
Internal beta-testing, designing servers.
Working with DTU admin to agree on terms.
Initial deployment (200 individuals).
Scale-up funded (UCPH2016 grant)
Privacy research, scalable infra-structure, improved phone software
Developing questionnaires with collaborators from social sciences
1000-person deployment
Research output focused on privacy & methodology
Epidemics experiment
Shift to focus on fundamental research
Four year grant awarded for study of influence in social systems
[1]
[2]
[3]
[4]
[5]
[6]
[7]
[8]
[9]
2012
2013
2014
2015

September 6, 2016 | vol. 113 | no. 36 | pp. 9947–10220

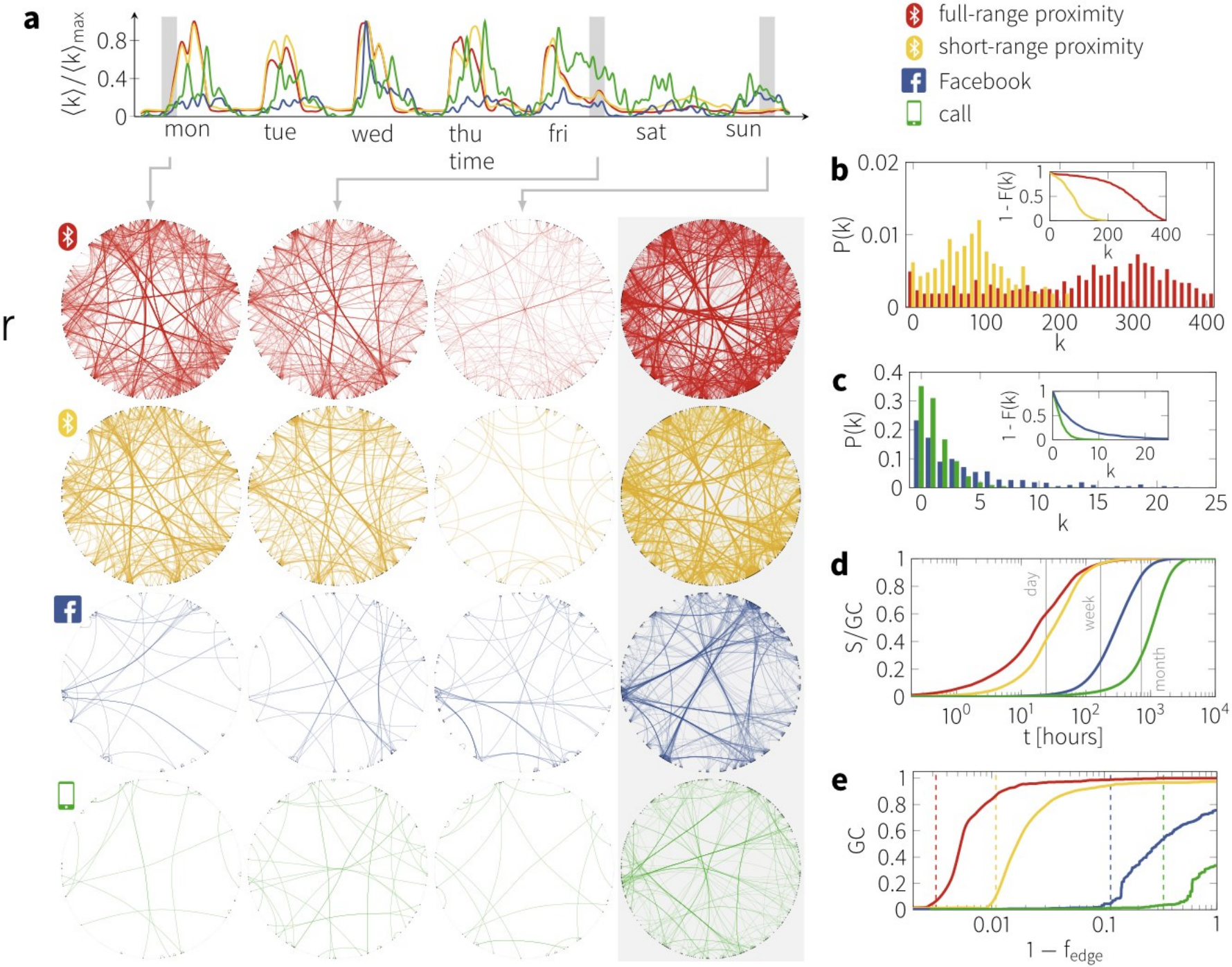
PNAS

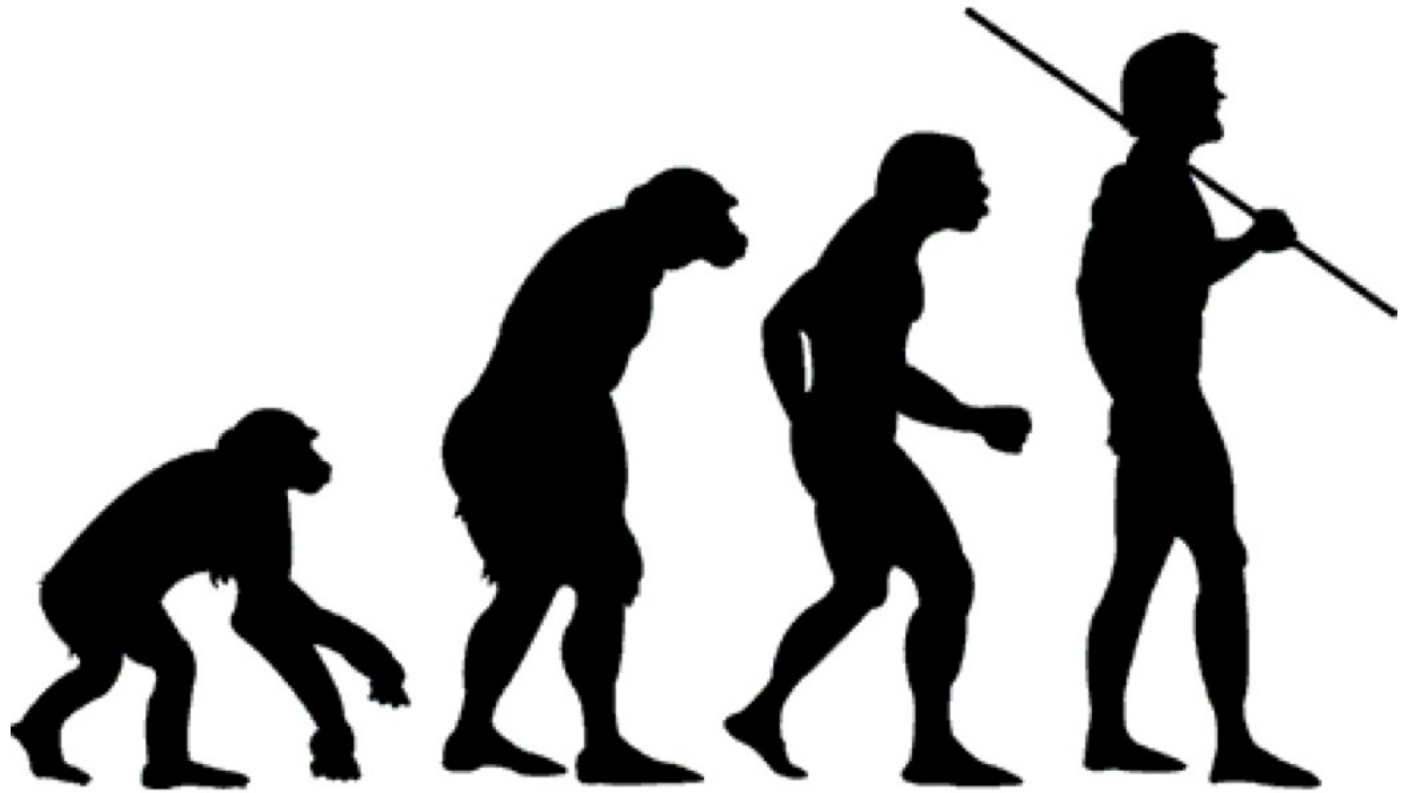
Proceedings of the National Academy of Sciences of the United States of America www.pnas.org

Principles of human social interactions

Some (of many) topics

- Epidemics
- Mobility
- Sleep
- Attributes and behavior
- Privacy
- And stuff





Privacy in Sensor-Driven Human Data Collection:

A Guide for Practitioners

Working Paper

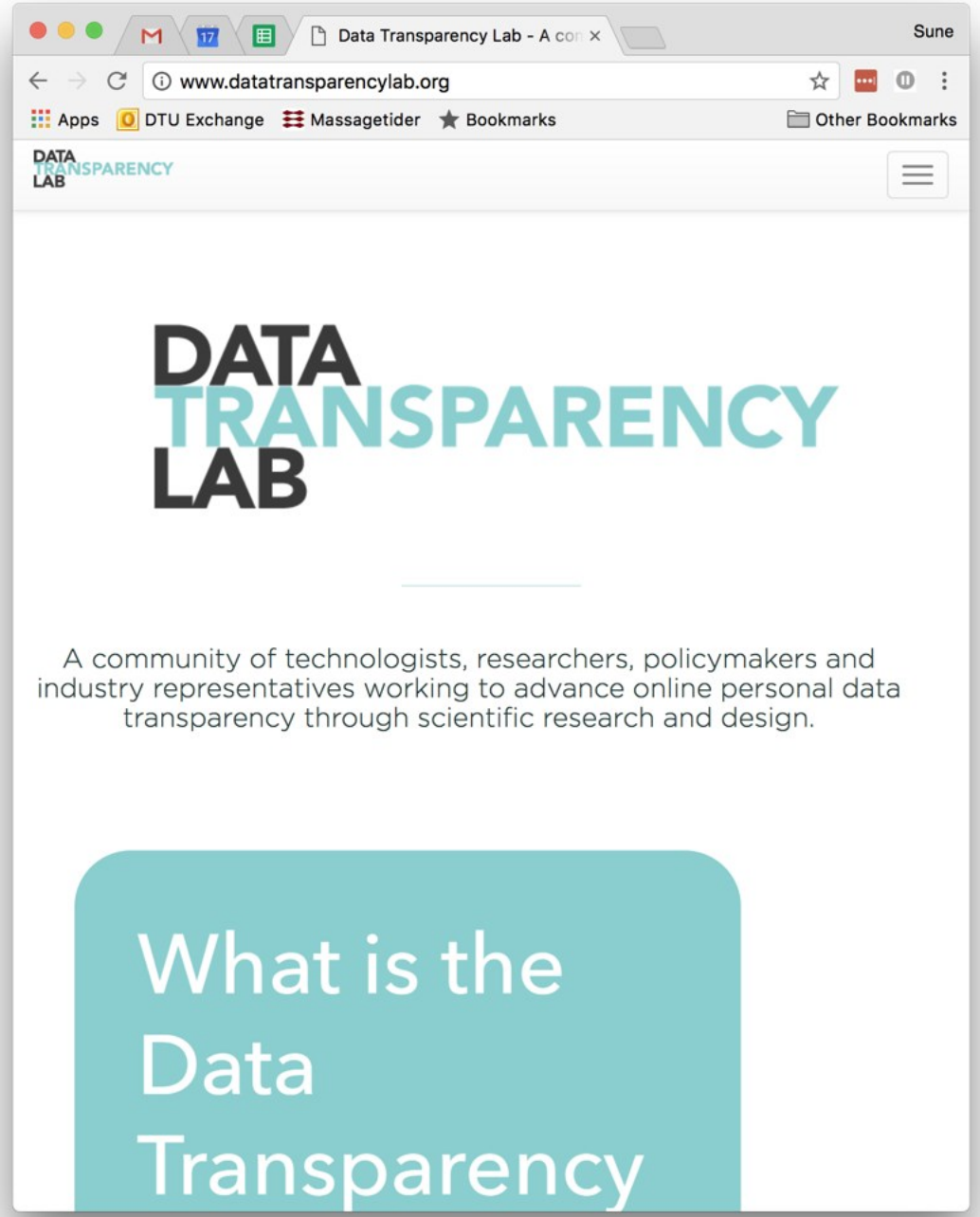
Arkadiusz Stopczynski[1,2,*], Riccardo Pietri[1], Alex 'Sandy' Pentland[2], David Lazer[3], Sune Lehmann[1,4]

1 Technical University of Denmark
2 MIT Media Lab
3 Northeastern University
4 The Niels Bohr Institute

arks@dtu.dk, riccardo.pietri@gmail.com, sandy@media.mit.edu, d.lazer@neu.edu, sljo@dtu.dk

Contents

'bæredygtig' datascience



BRIEF BIO

Privacy Part I: Why everyone is complaining, but no one is taking action.

[This is part II of a series, you can find the overview here]

We all have a sense that privacy is important. A sense that our ability to freely express "who we are" is slowly eroded by large corporations and governments collecting data on our actions for purposes not clear to us (and maybe not to them either). But on the other hand, no one is doing anything about this. Why is that?

I think that there are two central reasons for this.

The first reason is that humans are not very good at handling situations where cause and effect is separated by a lots of time and space. (I was made aware of this point by an excellent column in the Guardian by the author Cory Doctorow, who I will be stealing from in the following). There are lots of examples of this: No one would smoke if you developed cancer immediately upon the first drag of a cigarette. The possibility of cancer is so far away in time that it feels like the consequences happen to another person. You would be less likely to binge-eat if the food

The good: Big data

(Or more precisely: What is there to be optimistic about?)

Big data algorithms allow for

- Personalization
- Prediction
- Efficiency

The good: Data Science Impact

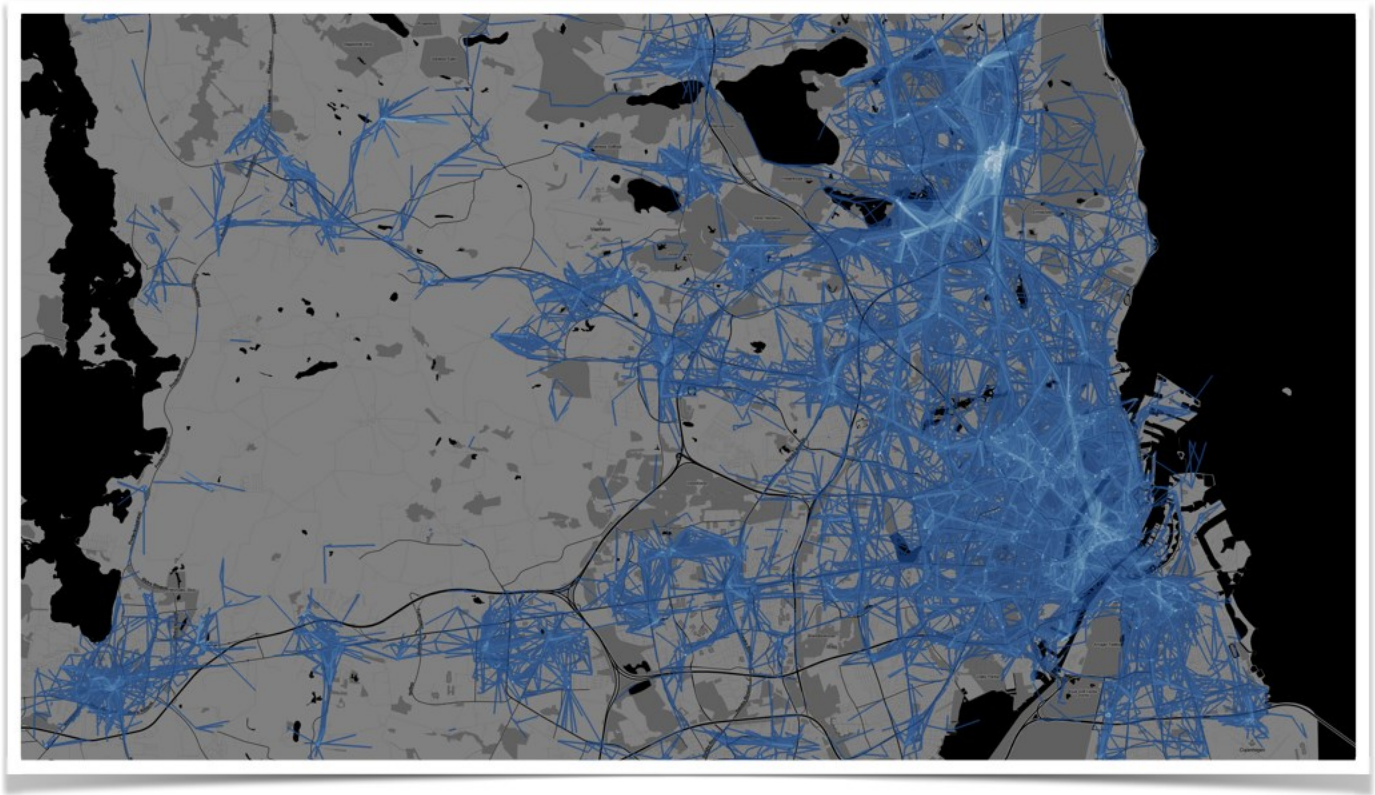
(Or more precisely: What is there to be optimistic about?)

Big data has impacts across

- *Health*
- Businesses
- Government

My work

https://hiveminer.com/Tags/dendrites/Interesting



The bad

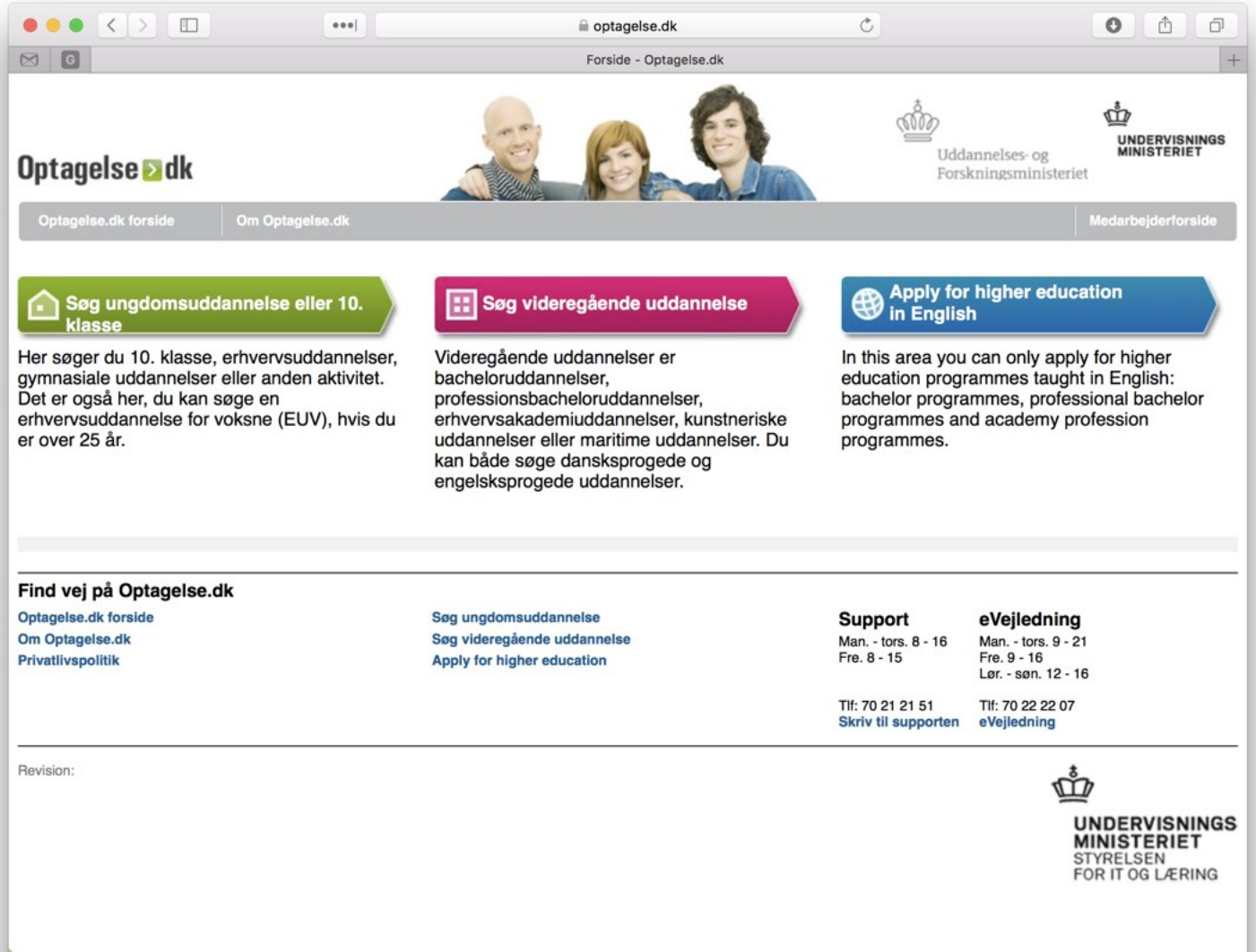
(Or more precisely: What should we be careful about?)

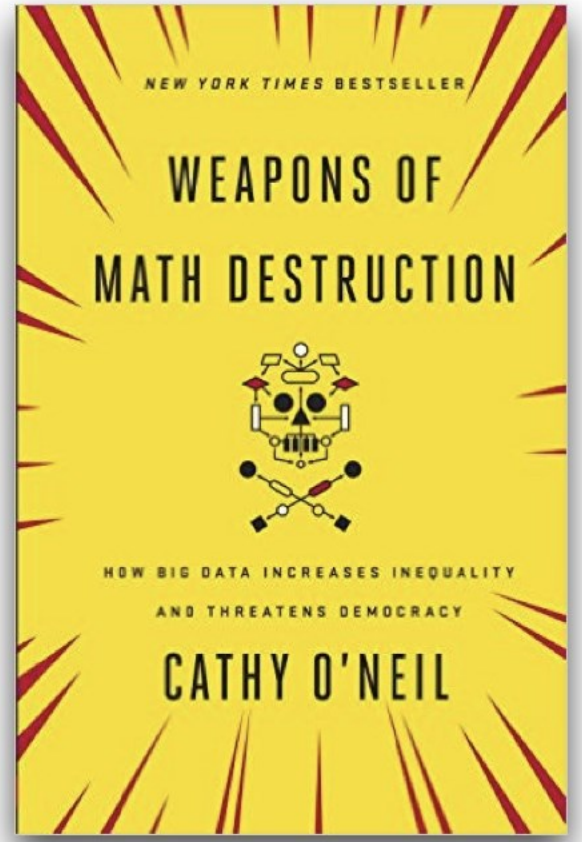
I see three key areas

- Algorithms.
- Nudging.
- Personal freedom.

The bad: Algorithms

As an illustration, let's consider a hypothetical case





The bad: Algorithms

Are we optimizing the right thing?

The bad: Algorithms

Are we optimizing the right thing?

What happens to people in the “wrong bucket”?

The bad: Algorithms

Are we optimizing the right thing?

What happens to people in the “wrong bucket”?

Existing data codifies our biases

The bad: Algorithms

Are we optimizing the right thing?

What happens to people in the “wrong bucket”?

Existing data codifies our biases

Algorithms often lack feedback

The bad: Algorithms

Are we optimizing the right thing?

What happens to people in the “wrong bucket”?

Existing data codifies our biases

Algorithms often lack feedback

Data quality

The bad: Algorithms

Are we optimizing the right thing?

What happens to people in the “wrong bucket”?

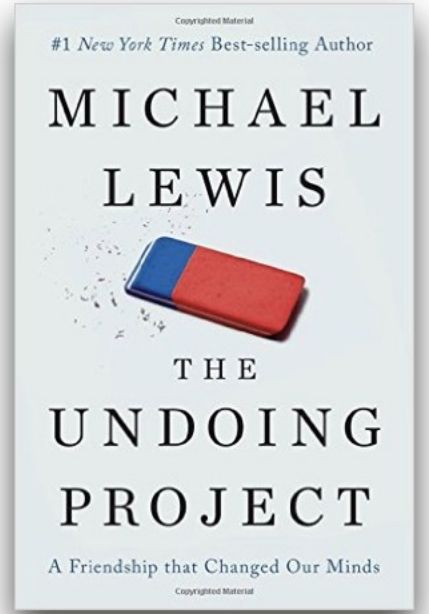
Existing data codifies our biases

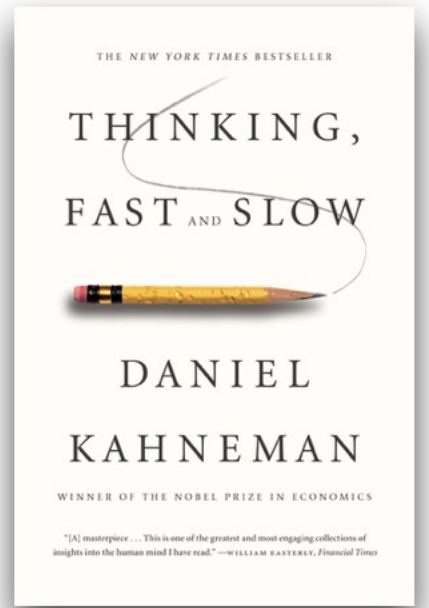
Algorithms often lack feedback

Data quality

How do you argue with an algorithm?

The bad: Nudging



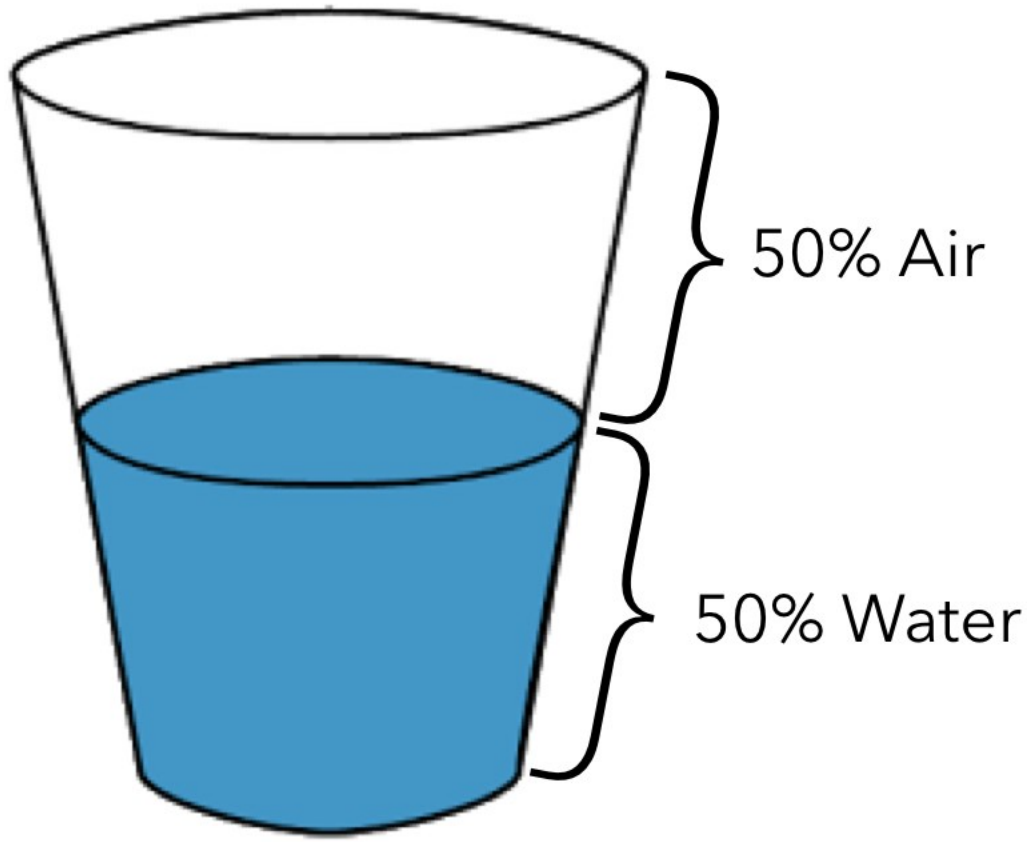


The bad: Nudging

The bad: Personal freedom

"I lay all this while, as the reader may believe, in great uneasiness."—*Page* 8.

‘Sustainable’ Big Data



Technically, **the glass is 100% full!**