Causality Detection in Health Econometrics Using Big Data

Dr. Sayan Chakrabarty

Research Fellow (Economics)

University of Southern Queensland

Arpita Chakraborty

PhD Fellow

School of ICT, Griffith University

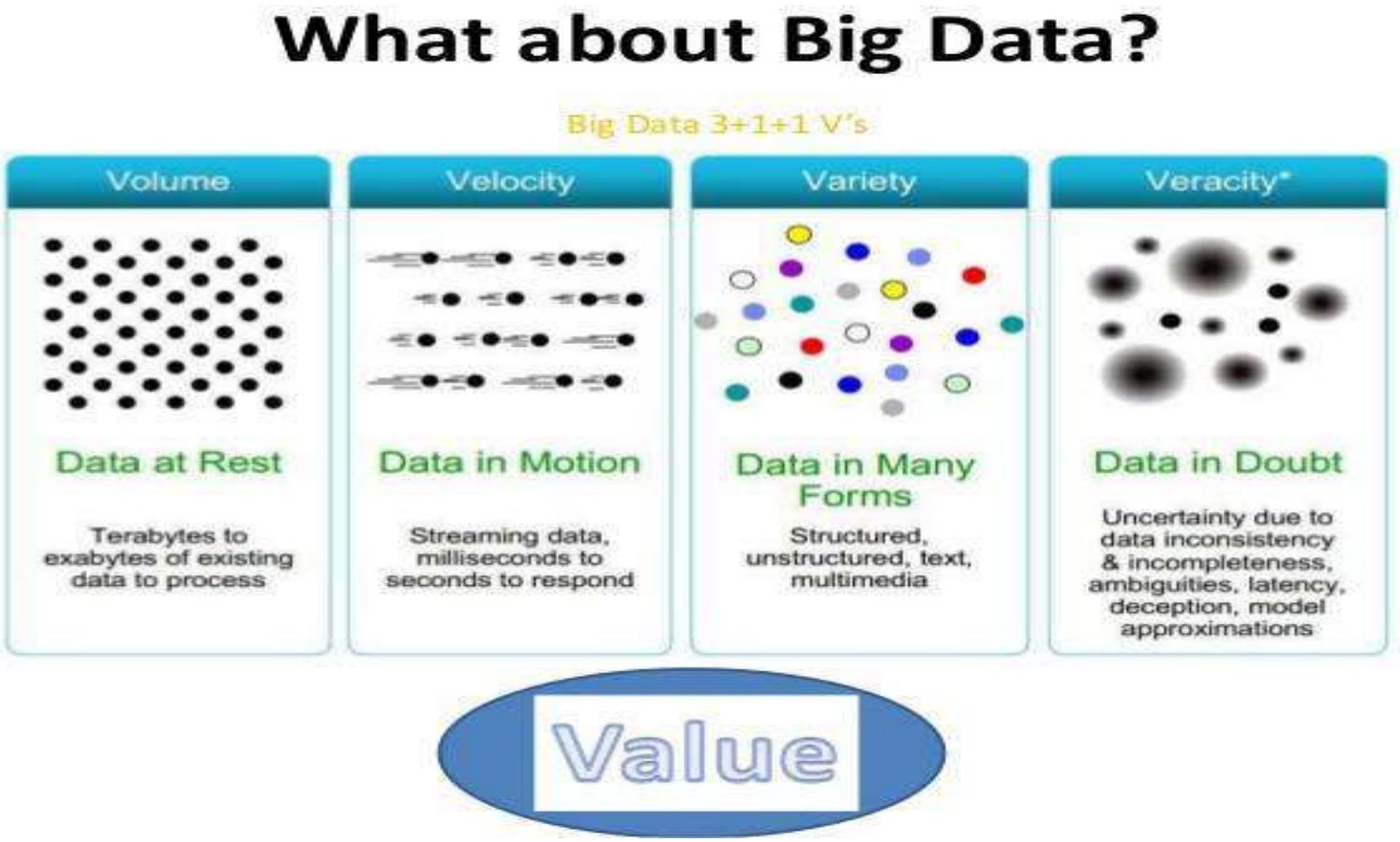
Content

- Digital foot print
- Big data
- Sources of big data
- Conventional econometrics and big data
- Machine learning
- Big data variety
- Data extraction
- Data visualization

Digital foot print

https://www.teacherspayteachers.com/Product/What-is-my-digital-footprint-poster-3005468

Big data



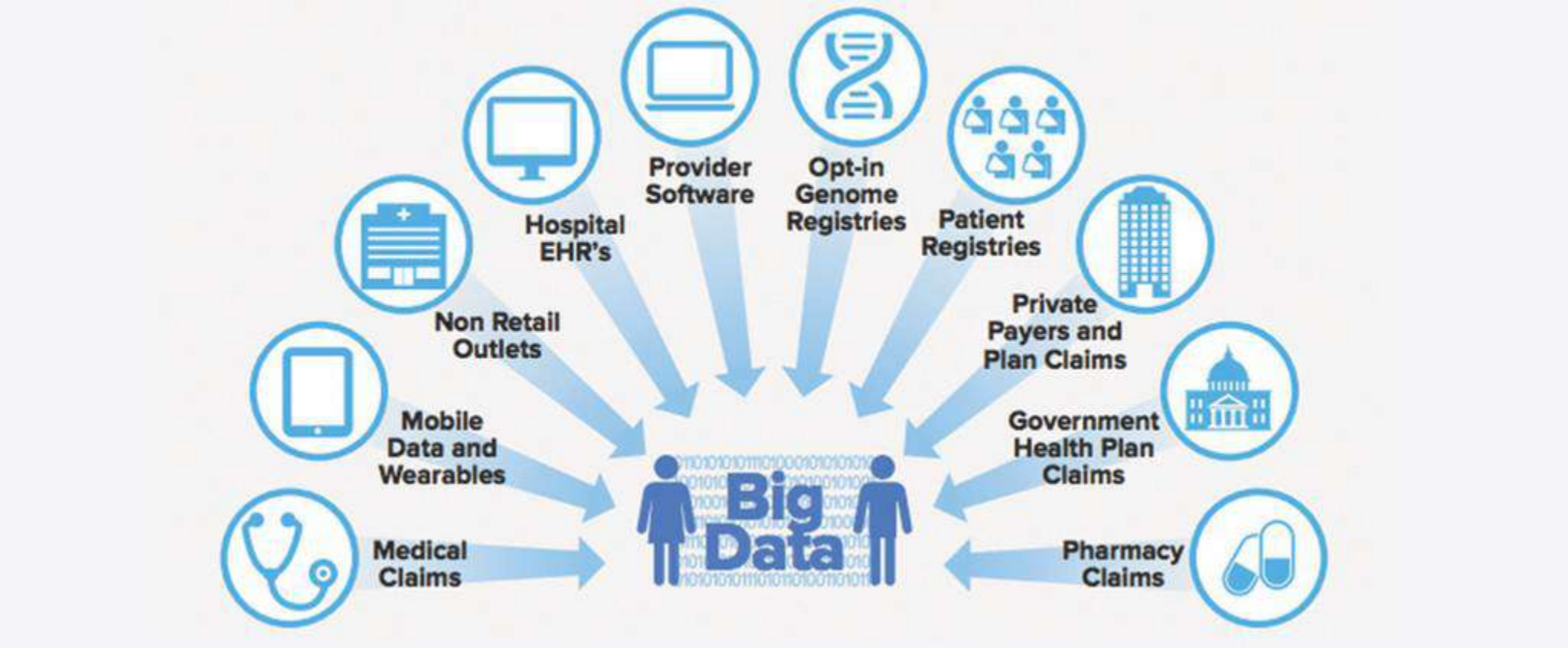
https://www.google.com.au/search?q=what+is+big+data&source=lnms&tbm=isch&sa=X&ved=0ahUKEwimrrP575XcAhXFHpQKHZqqBCcQ_AUICygC&biw=1920&bih=901#imgrc=LdhqJ29DA0cq-M
:

Sources of big data

Sources	Some examples
Administrative data	• Electronic medical records • Insurance records • Tax records
Commercial transactions	• Bank transactions (inter-bank as well as personal) • Credit card transactions • Supermarket purchases • Online purchases
Sensors and tracking devices	• Road and traffic sensors • Climate sensors • Equipment and infrastructure sensors • Mobile phones • Satellite/GPS devices
Online activities/social media	• Online search activities • Online page views • Blogs and posts and other authored and unauthored online content and social media activities • Audio/images/videos

Source: ITU, adapted from UNSC (2013).

Sources of big data in health



https://tcf.org/content/report/strengthening-protection-patient-medical-data/?session=1

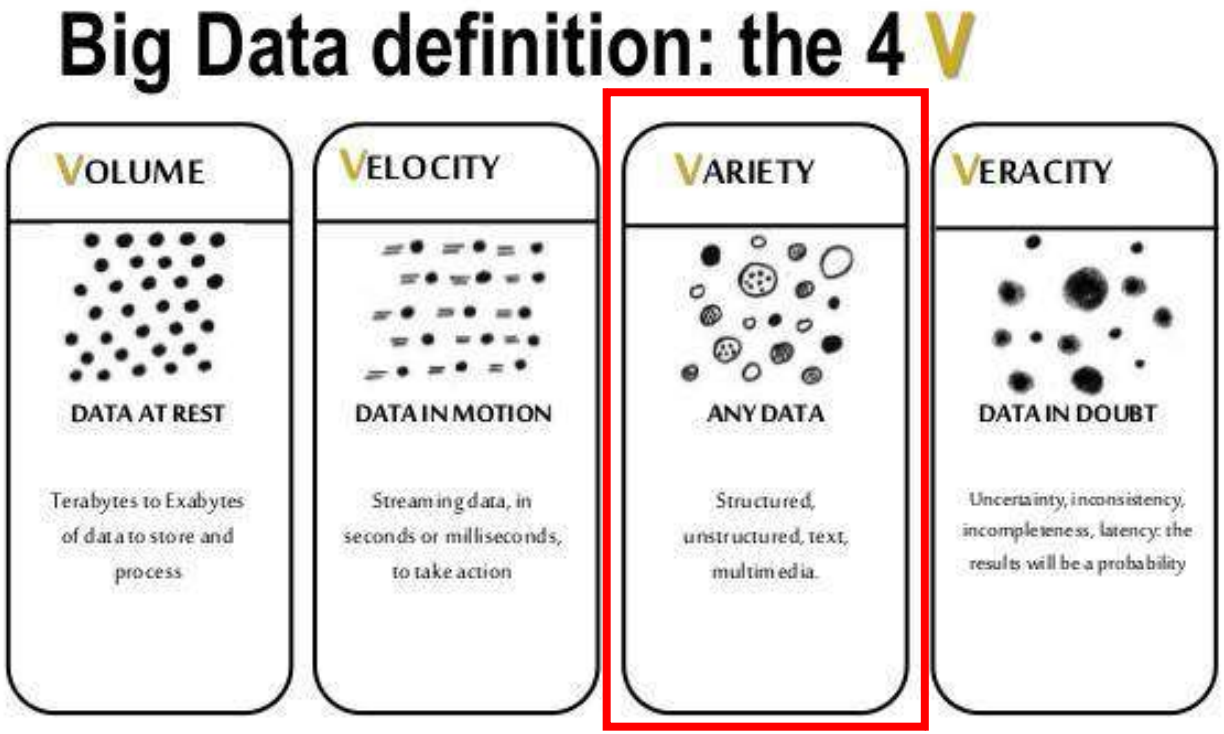
Conventional Econometrics and Big Data

- Conventional econometric techniques (regression) often work well, but there are issues unique to big datasets that may require different tools (Varian, 2014).
- Why?
  - Size of the data and p value
  - More potential predictors (variable selection)
  - Non linear relationship

Machine Learning

- Divide the data into training, testing and validation.
- Machine learning finds particular function (s) that provide a good prediction of y as a function of x.
- Out of sample performance (overfitting problem) minimizing sum of square residuals.
- Historically, machine learning deals cross-section data.
- The data may be “fat,” (lots of predictors relative to the number of observations) or “tall” (lots of observations relative to the number of predictors).

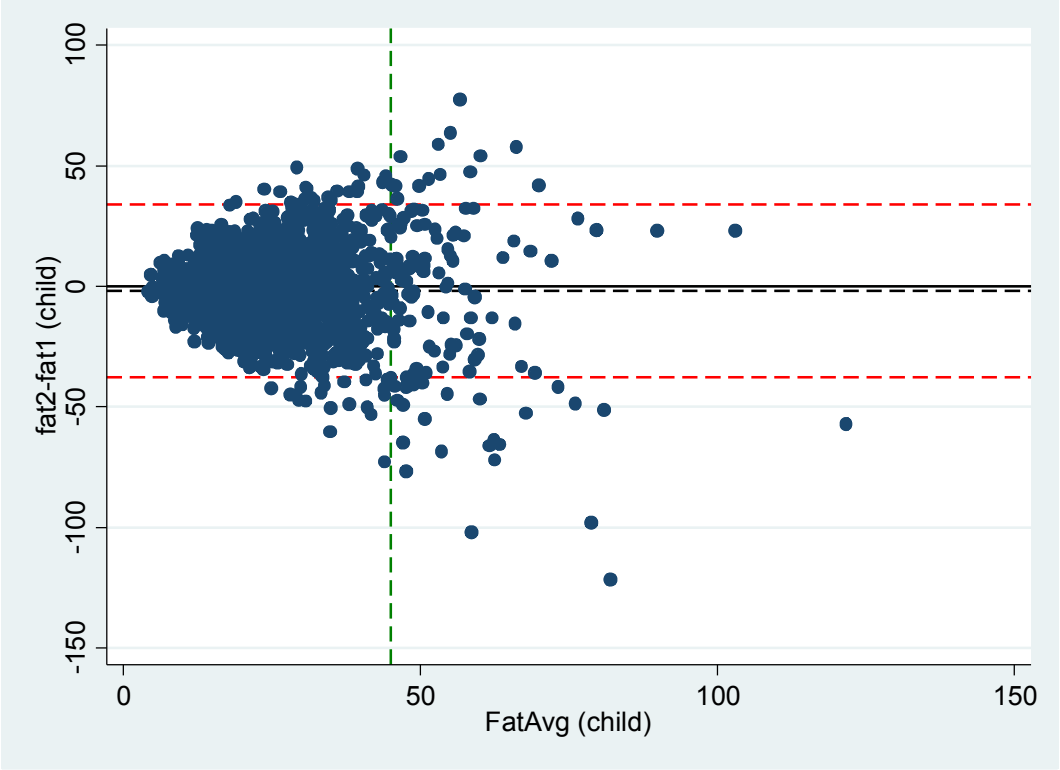
Big Data Variety



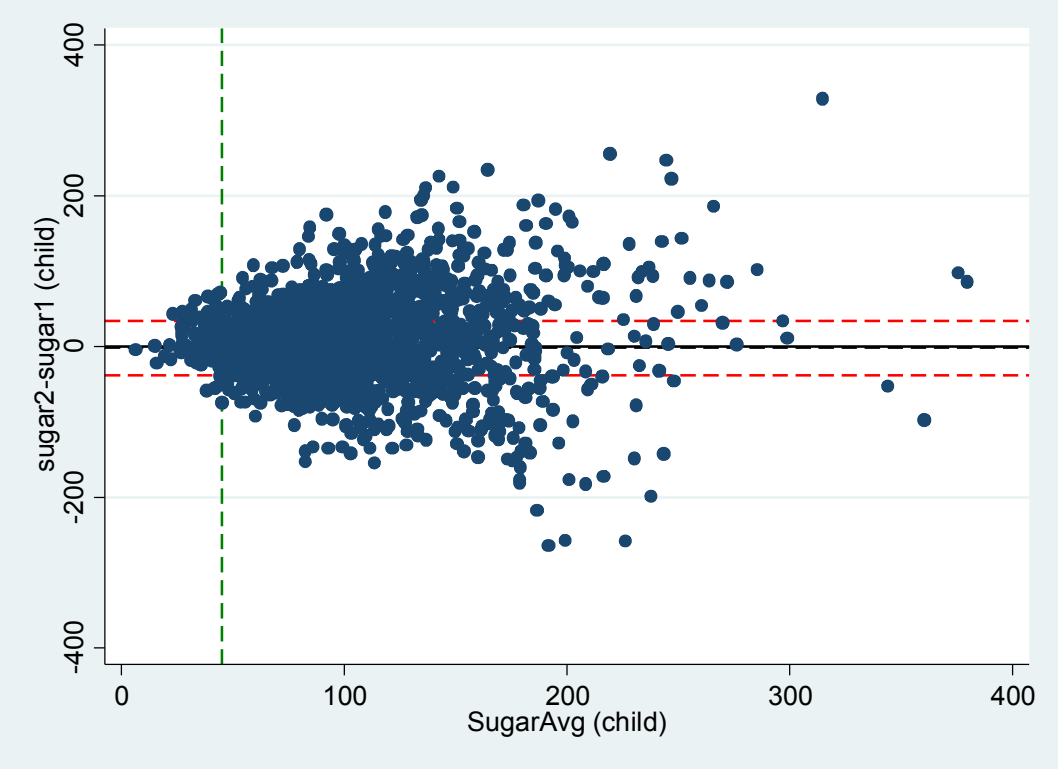
Structured Data

- Automated Multiple-Pass Method (AMPM), developed by the Agricultural Research Service of the United States Department of Agriculture (USDA)
- ABS follow AMPM approach for Australian National Nutrition and Physical Activity Survey.
- Face to face and telephone interview

Bland and Altman plot for saturated fat intake for the two methods

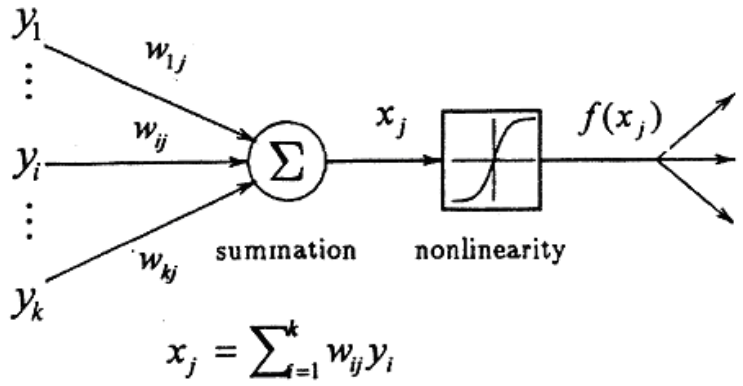


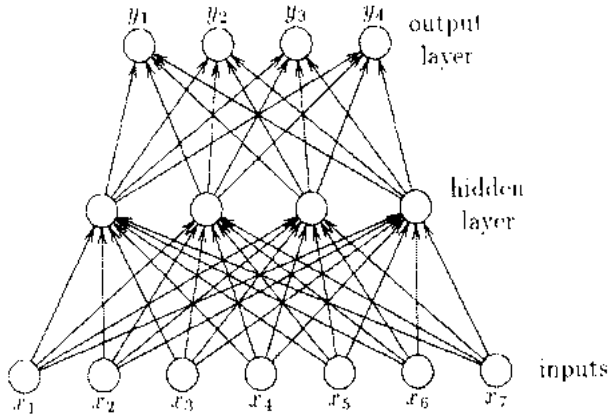
Bland and Altman plot for sugar intake for the two methods



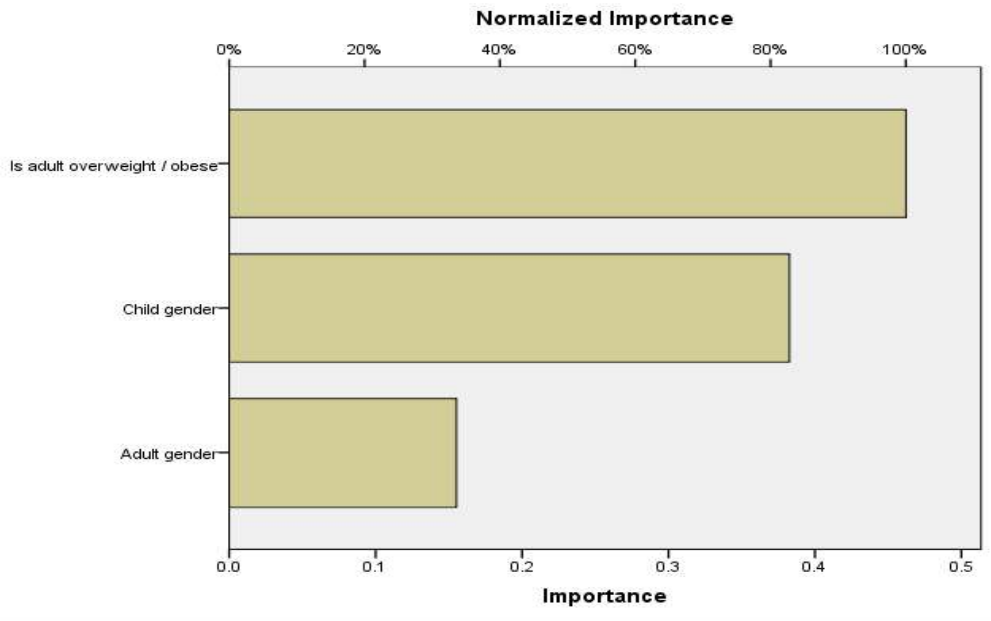
Supervised Learning: Neural Networks

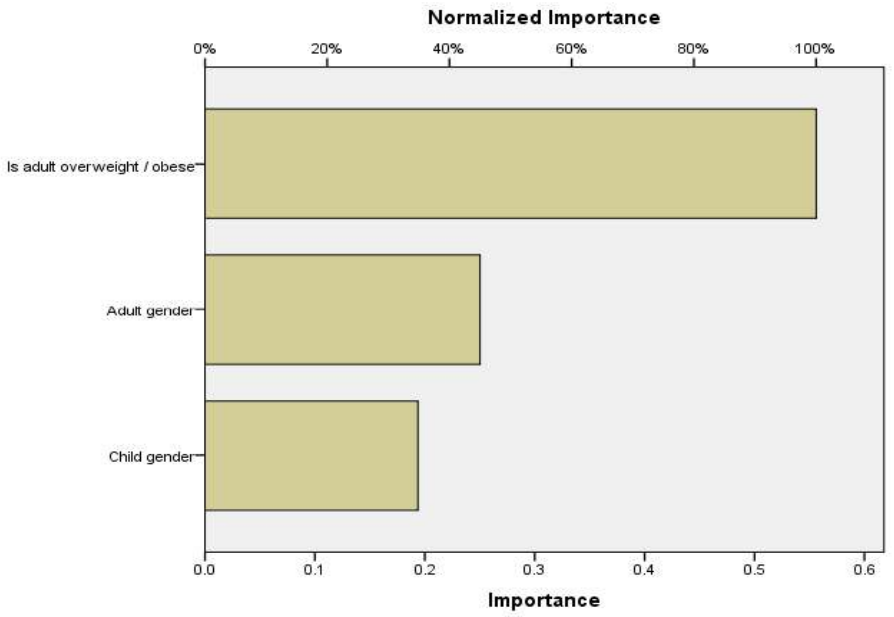
- Most work in machine learning has involved cross-section data (Varian, 2014, 2016)
- The MLP Neural networks are composed of layers of elementary units, called neurones, linked to one another by weighted connections.



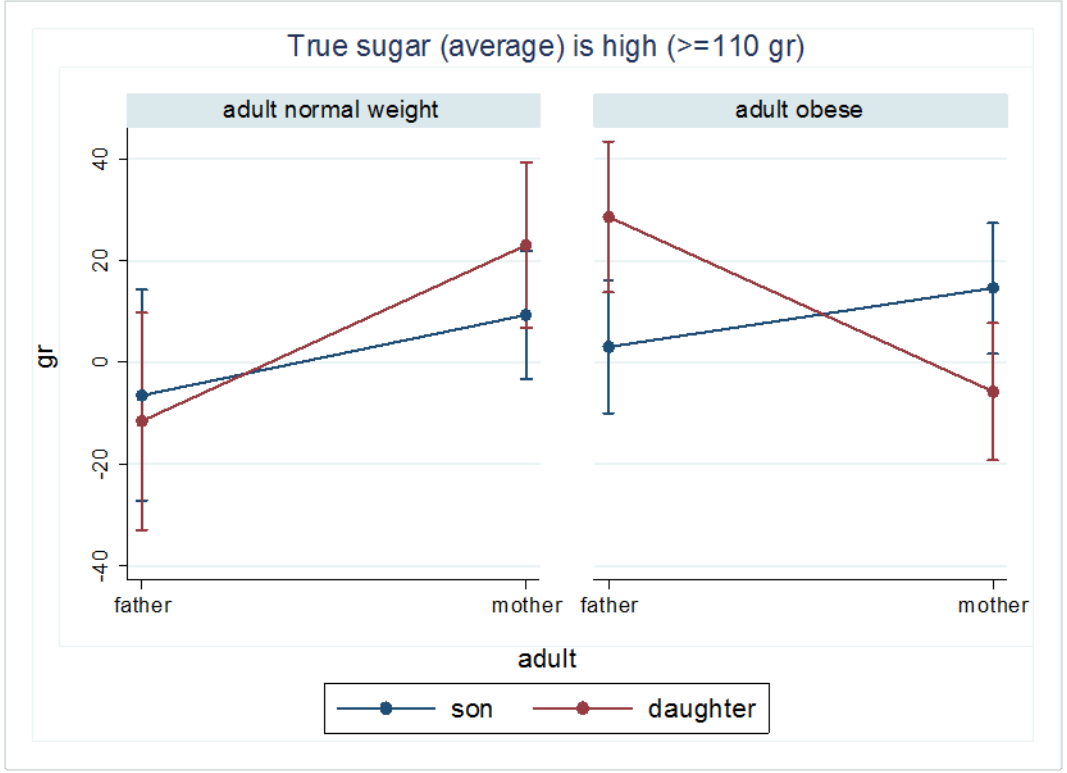


Effect size of the sources of disagreement of two measurements (fat and sugar intake)

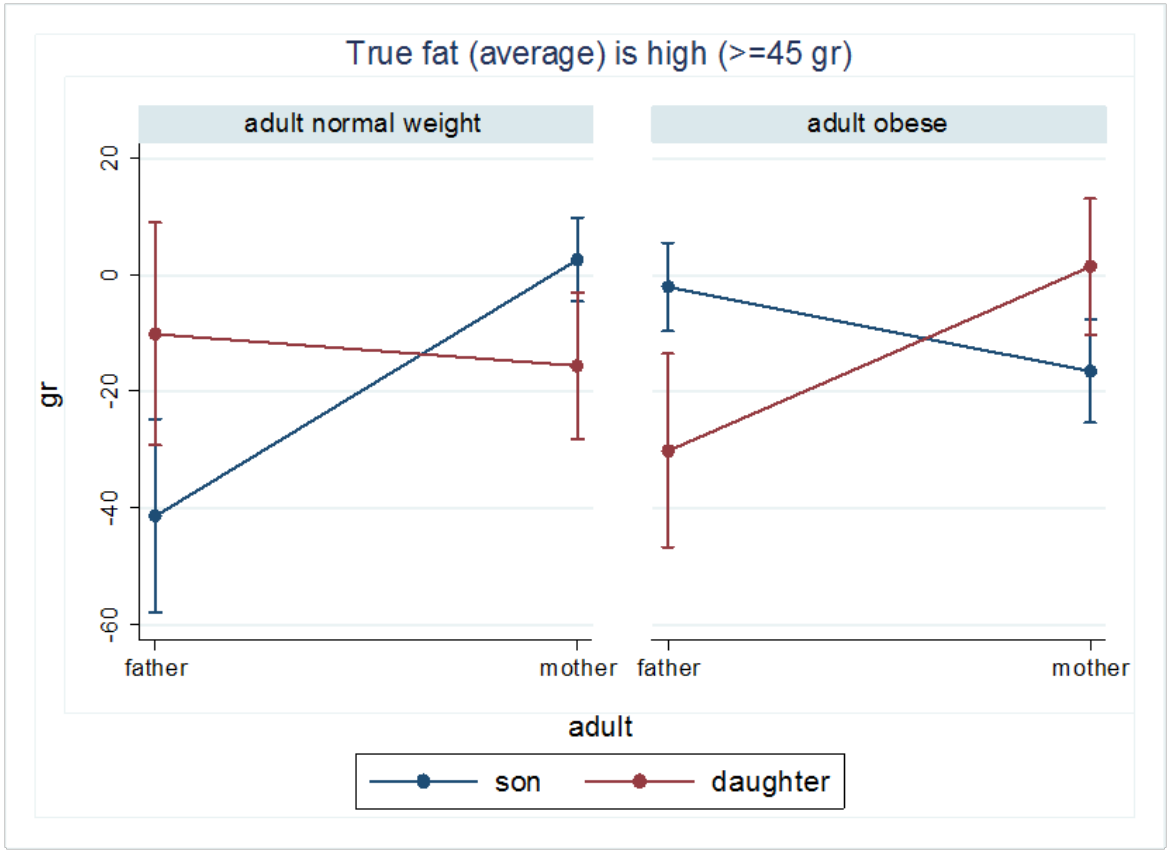




Source of disagreement of two measurements for over reporting sugar value



Source of disagreement of two measurements for over reporting fat value



Unstructured data

- When big data is observational, generated from uncontrolled experiments/environment and often non-random.
- Often less expensive to collect.
- No statistical sampling (example: electronic health records (EHRs)).

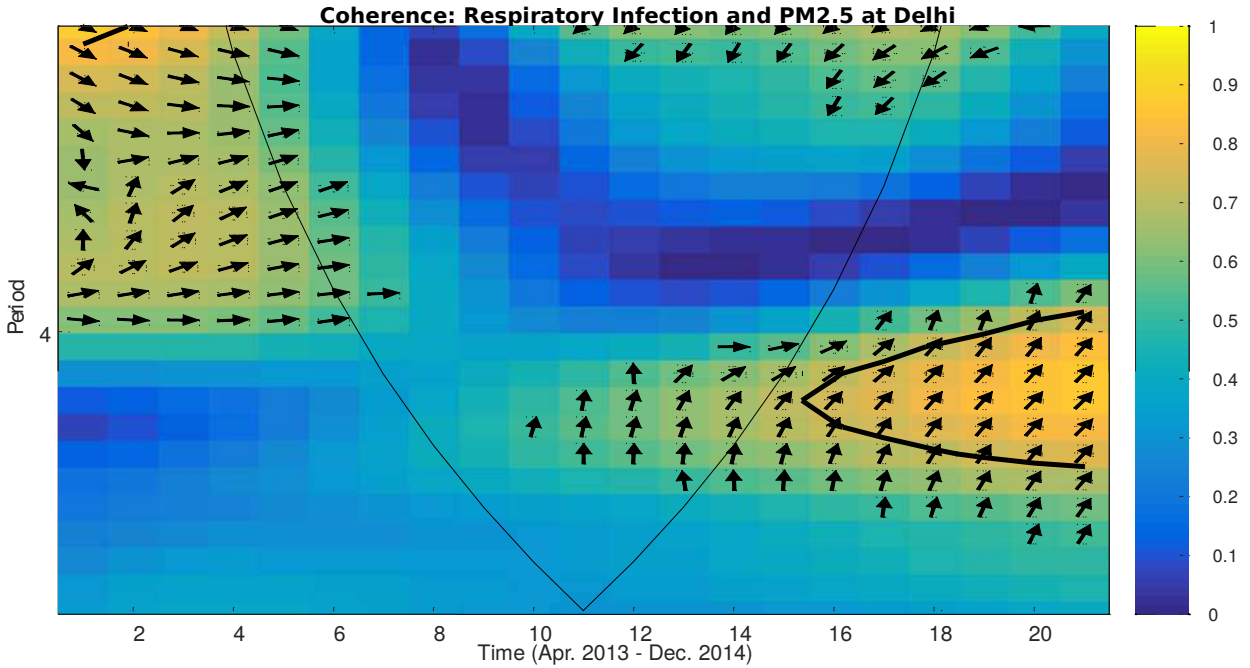
Unstructured data: unsupervised and supervised learning (cross section)

- Unsupervised learning can be used for clustering, grouping, autonomous post stratification.

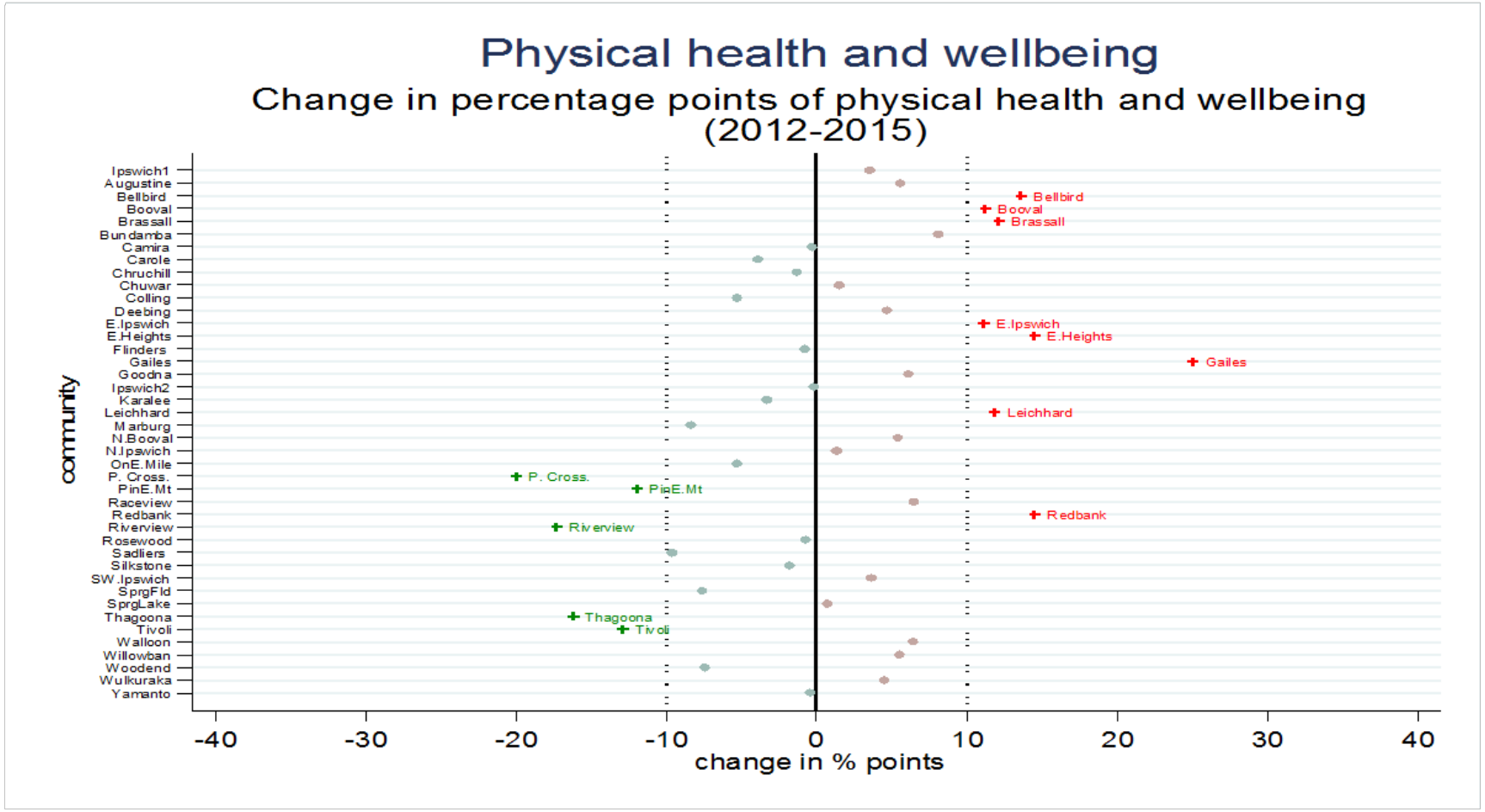
- Supervised learning could be used for testing theories by controlling the learning process.

Data extraction with wavelet analysis (big data time series)

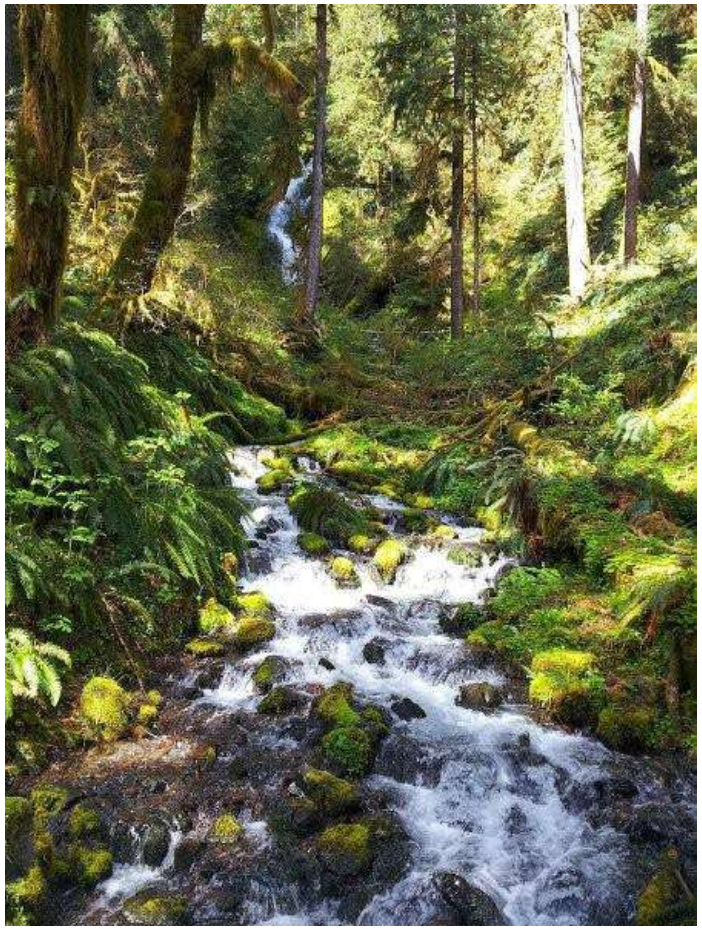
Coherence

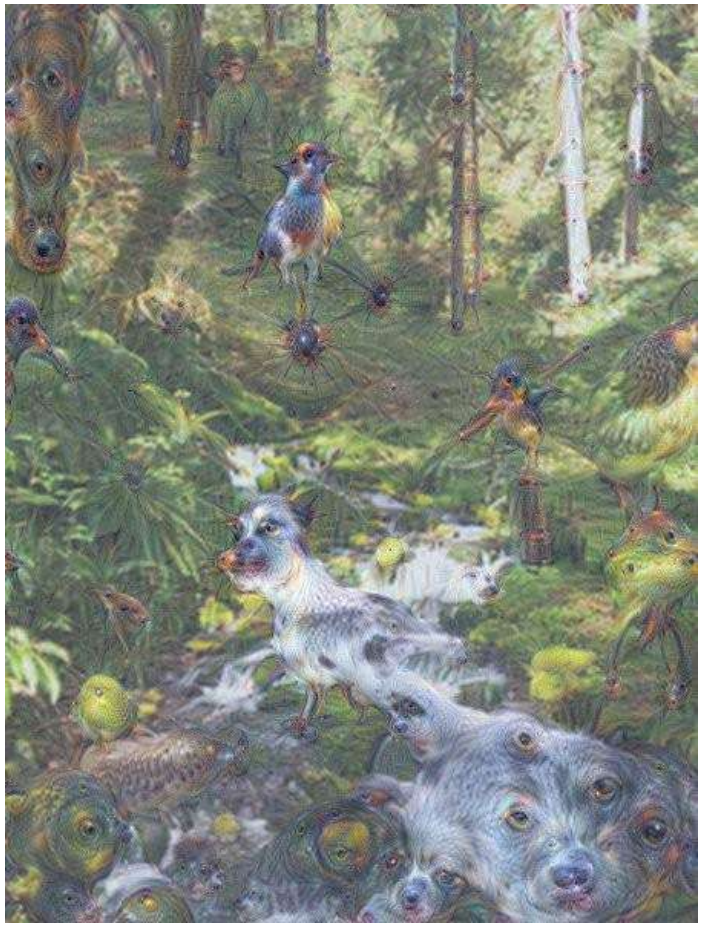


Data visualization



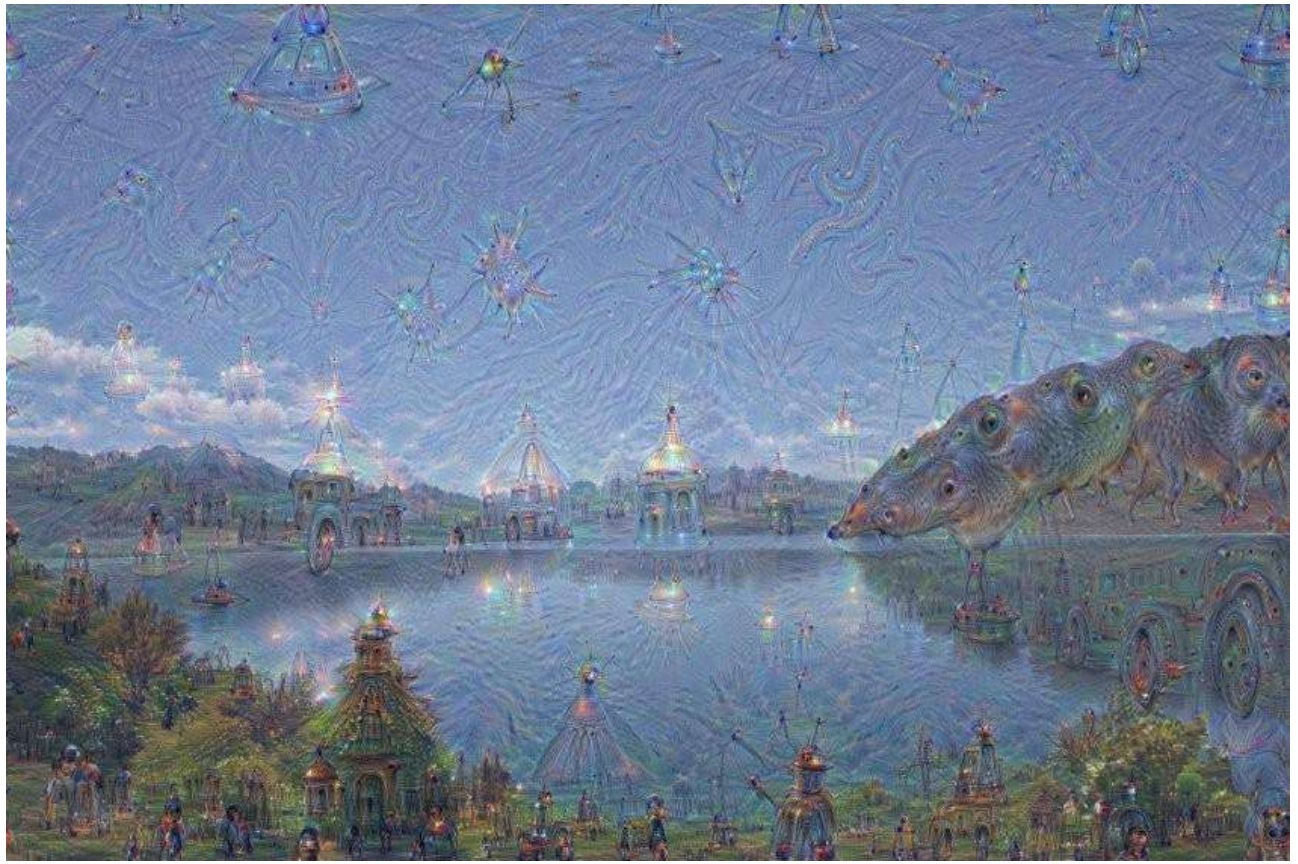
Neuronal Networks





Lüchters, G. 2017

Neuronal Networks



Lüchters, G. 2017

Neuronal Networks

Lüchters, G. 2017

Reference

- ITU World Telecommunication/ICT Indicators database, 17th edition, 2014, available at: http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx.
- Varian HR. 2014. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28 (2).
- Varian HR. 2016. Causal inference in economics and marketing. Proc Natl Acad Sci USAdoi:10.1073/pnas.1510479113.
- Wasserstein L. R., & Lazar, A.N. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician,* 70, 129-133.