En statistikers syn på Big Data
Analysmetoder för stora datamängder

Timo Koski, matematisk statistik, KTH

Ett halvdagsseminarium om Big Data måndag den 30e november 2015

November 29, 2015

Gemensamt arbete med:

KTH Working Group on Big Data: Erik Aurell koordinator, Gunnar Karlsson, Timo Koski, Mikael Skoglund och Ozan Öktem. Två anslag (2013, 2014) från KTH ICT Platform.

1. White Paper on Big Data (2013)
2. KTH Roadmap on Big Data (2014)

C.R. Rao: Statistics: reflections on the past and visions for the future. Communications in Statistics , Theory and Methods, 30, pp. 2235–2257, 2001

The 21st century will be characterized by complex multidisciplinary problems accompanied by massive data sets, and information technology. Both research and teaching of statistics will have to prepare for interaction with intelligent artificial systems and complex networks of information processing. There will be more and more statistical problems where the data is not 'flat' but consists of , e.g., images, relations between concepts, and/or entries or documents in data warehouses.

Läsvärda och tänkvärda skrifter:

- Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council: *Frontiers in Massive Data Analysis*, The National Academies Press, 2013.
- David Bollier and Charles M Firestone: *The promise and peril of big data.* Aspen Institute, Communications and Society Program Washington, DC, USA, 2010.

What is Big?

Let us recall the storage units:

- kilobyte (kB 10^3 bytes)
- megabyte (MB, 10^6 bytes)
- gigabyte (GB, 10^9 bytes)
- terabyte (TB, 10^{12} bytes)
- petabytes (PB, 10^{15} bytes)
- exabyte (EB, 10^{18} bytes)
- zettabyte (ZB, 10^{21} bytes) zetta= en triljard
- yottabytes (YB, 10^{24} bytes).

Now exactly what is a Zeta Byte ?

Mega Byte
Giga Byte
Tera Byte
Peta Byte
Exa Byte
Zeta Byte
Yotta Byte

XpressAnalytics

www.xpress-analytics.com

Ph: 8775734486

XpressAnalytics

Big?

PBs of data are reached by today's largest commercial actors and single research projects.
Smartphone traffic will create over 17 EBs of data by the year 2020, marking an eightfold increase on current levels, according to Ericsson. An EB is a thousand million gigabytes.

Big?

Big data is often said to refer broadly to data sets so large, or ≈ PB scale, or complex that traditional database and data processing methods are inadequate.

Are the traditional statistical methods inadequate for big data, too ?

Big?

According to well-known estimates the world's annual production of digital data reached ZBs in 2011, and could reach YBs in the next ten years. To put this in perspective, one YB would weigh about two hundred million tons if stored in current top-of-the-line hard disks.
Wikipedia:

http://www.invitrogen.com/site/us/en/home/References/Ambion-Tech-Support/rna-tools-and-calculators/dna-and-rna-molecular-weights-and-conversions.html

Big data in physics

- Large Hadron Collider
  - CERN particle physics collider
  - 1 million PB of raw data, 107× faster than it can be moved to disk
  - 15 PB of data stored per year
- Large Synoptic Survey Telescope
  - Fully operational by 2022
  - Image the entire visible sky (2D+time imagery of 10^{10} galaxies & 10^{9} stars) every few nights for at least a 10 year period
  - 10^{15} pixels, 30 TB data per night

Scandinavia and Sweden in particular has large and very valuable public data resources. Much of this data is obviously sensitive, and its use is carefully regulated.

Nationella Kvalitetsregister

The public in a modern state directly or indirectly also owns Big Data. Maybe as (or more) valuable as that of commercial actors. Examples: mobility and traffic patterns, tax receipts, health statistics, etc. etc.

Due to advances in computer hardware– faster CPUs, cheaper memory, and in new technologies such as Hadoop, MapReduce, and text analytics for processing big data, it seems now feasible to collect, analyze, and mine massive amounts of structured and unstructured data. But it could be asked whether **'big data is driven more by storage capabilities than superior ways to ascertain new knowledge'** .

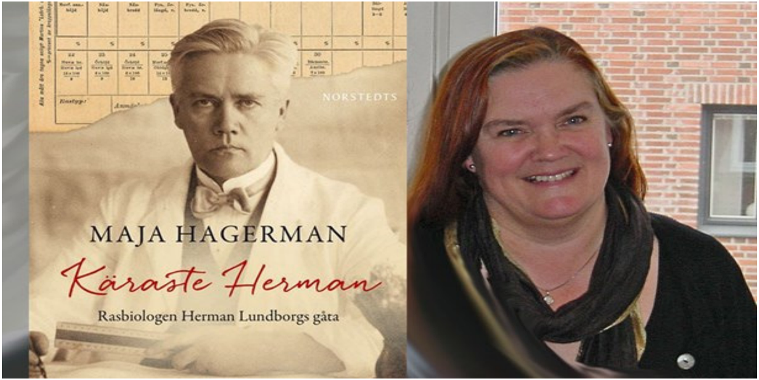
Typical challenges

- Data-intensive discovery – No first-principles for modelling that provide simplistic descriptions
- Heterogeneity $< --$ **No single ”true” source of data**
- Reliability – Large uncertainties (missing data and/or highly noisy data), insufficient expertise
  - 89% of research is not reproducible
  - 27% (±9%) of cancer cell lines are misidentified
  - 85% of research efforts are wasted due to inadequate production and reporting practices

Big data hubris

- Big data are a substitute for, rather than a supplement to, traditional data collection and analysis. The quantity of data does not mean that one can ignore foundational issues of measurement, validity and reliability, and dependencies among data.
- There is no longer place for scientific theories of any kind: all truths are to be found by data associations.

– ”samlade kranier, mätte, fyllde hålkort med oändliga data från de många tiotusentals individer han kartlade”.



C.R. Rao, 2001

- It is therefore safe to assume that the coming (-written in 2001) era of massive data bases will require new solutions and theories ensuring a sound statistical basis.
- One has already for the last two decades been able to see a change of statistical practice and research in probability in the direction of methods of simulation (McMC, simulated annealing, particle filtering and re-sampling).
- The emergence of computer intensive practices of data analysis will probably require a complete re-thinking of the theories of model based hypothesis testing and estimation theory.

Nytt datacenter ska göra Luleå till huvudstad för big data

1. Judgements
2. Bayesian classification
3. High Dimension
4. Manifold learning
5. Sparse processing

Analysis of data and modelling

Draper, David and Hodges, James S and Mallows, Colin L and Pregibon, Daryl: Exchangeability and data analysis, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 9–37, 1993.

> . . . *defining relevant populations of devices and potential measurements on them, and with whether the available data can be regarded as a realization of some random sampling mechanism applied to the relevant populations. Such judgements must precede the application of standard probability-based methodology. Few would disagree that judgements of this type should be based on data; the question is precisely how data and contextual information are used to make them* . . . *the role of formal statistical methodology is often minimal, with the difficulty lying in the decision about what is relevant.*

Scalability

The computational complexity of algorithms. Most state-of-the-art nonparametric learning algorithms have a complexity of $O(N^2)$ or $O(N^3)$, where N is the number of training examples. This seriously restricts the use of massive data sets.

Bayes classifier

The *Bayes classification strategy* is to minimize the expected loss computed according to the posterior $P(H = h \mid z^n)$. Or,

$$\mathrm{Bayes}_{\mathcal{S}}(\mathbf{x}) = \mathrm{argmin}_{y \in \mathcal{Y}} E_{H|z^n} \left[l(H(\mathbf{x}), y) \right]$$

Note that this strategy does not correspond to any fixed $h \in \mathcal{H}$(=hypothesis space), and is therefore computationally demanding. With large data sets the appetite for hypotheses tends to get even larger.

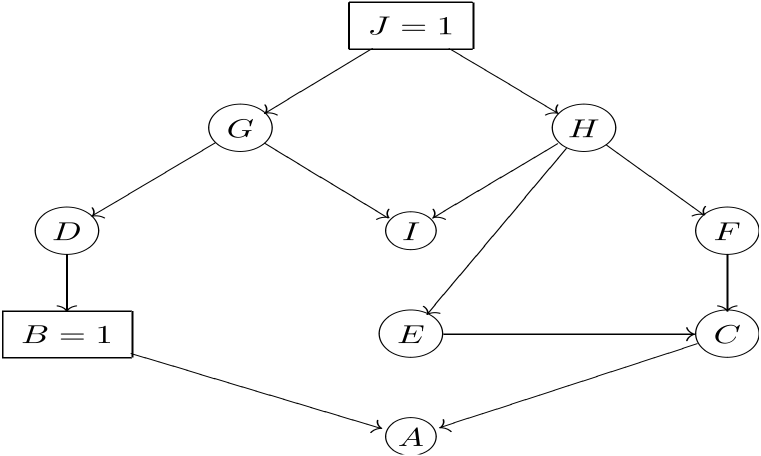
Scalability: Naive Bayes Classifiers

R.v.'s $X_1, \ldots, X_d$ are conditionally independent given a classification variable $C \in \{c_1, \ldots, c_K\}$, i.e.,

$$P(X_1, \ldots, X_n \mid C = c_l) = \prod_{i=1}^{d} P(X_i \mid C = c_l)$$

Suppose each X_i has v values, and that there is a data set of N of data $X_1, \ldots, X_d$. Then time complexity of estimation (''training''; examining every feature of every sample) time of the naive classifier is $O(Nd)$ (C. Elkan, 1997) independently of v. Naive Bayes Classifier scales: they are well suited for very large data sets.

Bayesian Network Structure Learning



Big Data Bayesian Network Structure Learning

- Slice your data and send slices to local processors
- Each local processor evaluates the quality of the slices by a BDe score standardized
- Each local processor uses ensemble learning to find the network structure Max-Min-Hill Climbing and its best data slice.
- These are combined by Bayesian model combination to give the optimal network structure

Tang, Yan and Wang, Yu and Cooper, Kendra ML and Li, Ling: Towards big data Bayesian network learning-an ensemble learning based approach, *2014 IEEE International Congress on Big Data*, pp. 355−357, 2014

Key Words in the above

- **Ensemble learning** : Uses multiple predictive models (each developed using statistics and/or machine learning) to obtain better predictive performance than could be obtained from any of the constituent models.
- slicing, local learning, combining, distributed computing

Big Data: High Dimension

- There is recent research showing that in high dimensional space, the concept of proximity, distance or nearest neighbour may not even be qualitatively meaningful.
- The reliance on time-honoured statistical concepts like sufficiency and coefficient of correlation can turn out to be obsolete, when dealing with massive data sets. Montanari, Andrea: *Computational implications of reducing data to sufficient statistics*, arXiv preprint arXiv:1409.3821.

Analysis of data: Manifold learning

- Data of interest lie on an embedded (non-linear) d-dimensional manifold within the higher n-dimensional space (dimensionality reduction)
- A large class of methods (ISOMAP, LLE, Hessian, Laplacian, and kNN Diffusion) building a k-nearest neighbor graph, estimate local properties of the manifold though neighborhoods, and construct a global embedding that preserves these properties.
- Directly deals with data where points are separated non-linearly. No need to recast the problem to a linear one.
- Handles multi-scale problems.
- Access to a large variety of fast algorithms.
- Infers the geometry by capturing local geometric information (curvature and corners).

Analysis of data: Sparsity

For data analyzed in batch two simple ideas have here recently shown to be very powerful: (i) learning models with many more than n parameters and (ii) learning such models approximately, as maximum likelihood is computationally unfeasible. The first idea calls for regularization, as otherwise the problem is ill-defined, which can be given a Bayesian interpretation as learning with weakly informative priors, an a priori assumption that the world, in the large, is somehow simple. The second idea is surveyed in the monograph by M. Jordan and M. Wainwright "Graphical Models, Exponential Families, and Variational Inference" (2008).

Pitfalls of Massive Data

- The standard of scientific research is that the measurements are to be replicable and comparable across cases and over time. And we need to ascertain whether measurement errors are systematic or not.
  In today's organizations the engineers are incessantly changing the algorithms to improve the service. Platforms such as Twitter and Facebook are always being re-engineered. Whether studies conducted even a year ago on data collected from these platforms can be replicated in later or earlier periods is an open question.
  M.I. Jordan, one of the main authors of the Report *Frontiers in Massive Data Analysis*, in an interview in The IEEE Spectrum on 3 October 2014
  `http://spectrum.ieee.org/robotics/artificial-inte`
  `-on-the-delusions-of-big-data-and-other-huge-engi`

Pitfalls of Massive Data

- It is needed to add error bars to the inferred analysis. This point of view is missing in much of the current machine learning literature.
- Due to huge amount of data analysed, it is very easy to find spurious dependencies in projects with Big Data. It is growing faster than the statistical strength of the data, then many of the inferences are likely to be false.

Pitfalls of Massive Data

- Hal Varian from Google is quoted in stating that *the point of big data is . . . to be able to pick a random sample and to analyze it*. One gets a result from the random sample as good as looking at everything. But the difficulty is to make sure that it is really a random sample.
- There may be a tendency to overlook the old and elementary pieces of statistical wisdom calling attention to issues like selection bias, endogeneity/exogeneity and confounding. There is in addition the phenomenon of *dark data*.

"90 % av all Big Data är Dark Data"

dark data: *By this one means that most data is is created, used and thrown away without any person being aware of its existence.*

Pitfalls of Massive Data

- There must be an awareness of *overfitting* a small number of cases to a huge number of models:
  Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343 (6176) (March 14): 1203–1205.

It will take decades, at least according to M.I. Jordan (interview, loc.cit), to get a an engineering approach that contains understanding of where results came from and/or why models are not working and also of necessary exploratory tools for visualizing data and models.

In the international as well as in the national perspective, the developments outlined above pose a challenge to every statistics department and unit, as there are many categories of researchers outside the cadre of professional statisticians willing and able to become active in large complex data sets. C.f., The Swedish Big Data Analytics Network (2013): *The Big Data Analytics. The Research and Innovation Agenda for Sweden*, SICS *https://www.sics.se/projects/big-data-analytics*

A Paraphrase of C.R. Rao (2001)

Statistical methodology has for long been and is being employed to find useful and usable information in data. During the recent years computer scientists have harnessed the power of computer technology to find useful and usable patterns in massive data sets.

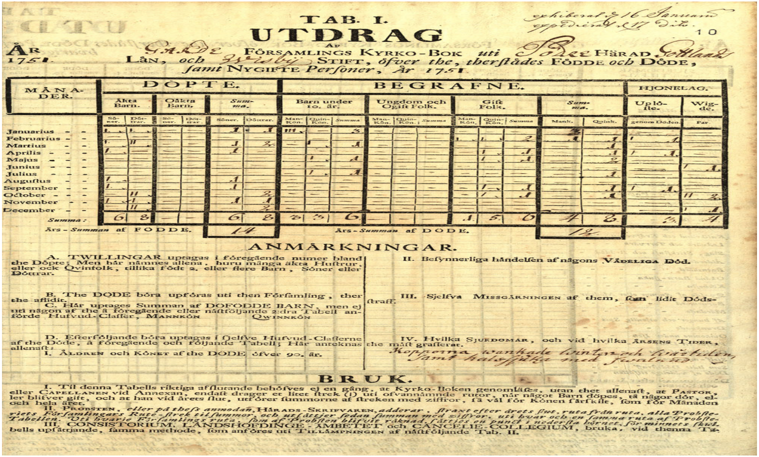
On the other hand, statistics as a discipline has grown by solving practical problems in other subjects and by developing suitable methods to a given situation. Later these methods have consolidated to unified mathematical theories.

En röst från konsultfirman McKinsey

McKinsey Global Institute *Big data: The next frontier for innovation, competition, and productivity, May 2011*

p. 10: A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics . . .
The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings.

Pionjärer i Big Data: Tabellverket



exhiberat d. 16 Januarii
expedierat d. 17 dito.

10

TAB. I.
UTDRAG

ÅR 1751. AF Garde FÖRSAMLINGS KYRKO-BOK uti Södre HÄRAD Gottlands LÅN, och Wisby STIFT, öfver the, therſtädes FÖDDE och DÖDE, ſamt NYGIFTE Perſoner, År 1751.

MÅNADER.	DÖPTE.						BEGRAFNE.											HJONELAG.	
	Äkta Barn.		Oäkta Barn.		Summa.		Barn under 10. år.			Ungdom och Ogift Folk.			Gift Folk.			Summa.		Uplöſte.	Wigde.
	Söner.	Döttrar.	Söner.	Döttrar.	Söner.	Döttrar.	Mankön.	Quinkön.	Summa	Mankön.	Quinkön.	Summa	Mankön.	Quinkön.	Summa	Mank.	Quink.	genom Döden.	Par.
Januarius	1	1	—	—	1	1	111	—	3	—	—	—	—	—	—	3	—	—	—
Februarius	—	—	—	—	—	—	—	—	—	—	—	—	1	1	2	1	1	1 1	—
Martius	1	11	—	—	1	2	—	1	1	—	—	—	—	—	—	—	1	—	—
Aprilis	1	—	—	—	1	—	—	—	—	—	—	—	—	1	1	—	1	1 1	—
Majus	—	—	—	—	—	—	—	1	1	—	—	—	—	1	1	—	2	—	—
Junius	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1 1
Julius	—	—	—	—	—	—	—	1	1	—	—	—	—	—	—	—	1	—	—
Auguſtus	1	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
September	1	—	—	—	1	—	—	—	—	—	—	—	—	1	1	—	1	1 1	—
October	—	11	—	—	—	2	—	—	—	—	—	—	—	1	1	—	1	—	11 2
November	1	1	—	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—
December	—	11	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	1 1
Summa:	6	8	—	—	6	8	3	3	6	—	—	—	1	5	6	4	8	3	4

Års-Summan af FÖDDE. 14

Års-Summan af DÖDE. 12

ANMÄRKNINGAR.

A. TWILLINGAR uptagas i föregående numer bland the Döpte; Men här nämnes allena, huru många äkta Huſtrur, eller ock Qvinfolk, tillika födt 2. eller flere Barn, Söner eller Döttrar.

B. The DÖDE böra upföras uti then Förſamling, ther the aflidit.

C. Här uptages Summan af DÖFÖDDE BARN, men ej uti någon af the å föregående eller näſtföljande 2:dra Tabell anförde Hufvud-Claſſer, Mannkön Qwinnkön

D. Efterföljande böra uptagas i ſjelfve Hufvud-Claſſerne af the Döde, å föregående och följande Tabell; Här anteknas allenaſt:

I. Åldren och Könet af the DÖDE öfver 90. år.

II. Beſynnerliga händelſen af någons Vådeliga Död.

III. Sjelfva Missgärningen af them, ſom lidit Döds-ſtraff.

IV. Hvilka Sjukdomar, och vid hvilka Årsens Tider, the mäſt graſſerat.

Kopporna wankade winter och wårtiden, ſamt halsſjuka med ſwalnad

BRUK.

I. Til denna Tabells riktiga afſlutande behöfves ej en gång, at Kyrko-Boken genomläſes, utan thet allenaſt, at Pastor, eller Capellanen vid Annexan, endaſt drager et litet ſtrek (I) uti ofvannämnde rutor, när något Barn döpes, tå någor dör, eller blifver gift, och at han vid årets ſlut, utförer ſummorne af ſtreken med ziffror, ſå väl för Könen ſärſkilt, ſom för Månaden och hela året.

II. Probſten, *eller på theſs anmodan,* Härads-Skrifvaren, *adderar, ſtraxt efter årets ſlut, ruta från ruta, alla Probſteriets Förſamlingars Rute-ſtrek til ſummor, och utſättjer ſedan ſumman med ziffror, uti hvar och en ſamma ruta af Probſte-Tabellen. Uti hvarje Förſamlings ruta, ſom af Probſten blifvit räknad, ſättjes en punct i nederſta hörnet, för minnets ſkull.*

III. CONSISTORIUM, LANDSHÖFDINGE-ÄMBETET och CANCELIE-COLLEGIUM, bruka, vid thenna Tabells upſättjande, ſamma methode, ſom anföres uti Tillämpningen af näſtföljande Tab. II.