



Statistics for BIG data

Dennis Lin
Department of Statistics
The Pennsylvania State University

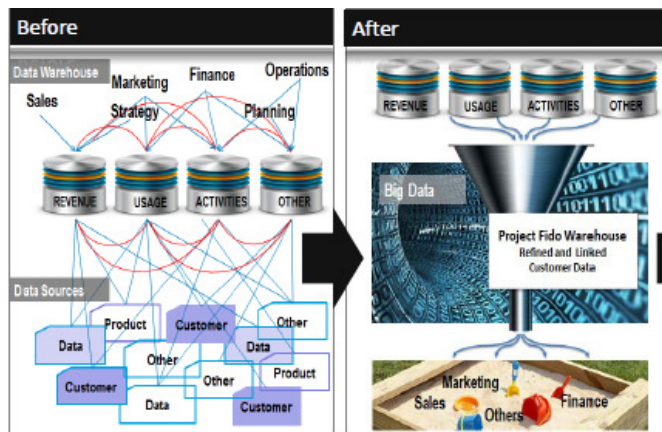


Statistics for Big Data: Are Statisticians Ready?

John Jordan and Dennis K.J. Lin
(ICSA-Bulletine 2014)



Before and After



What is BIGdata?

Who is working on BIGdata?

*What kinds of Statistics are
needed?*



*What gets measured,
get managed.*

If you cannot measure it,
You cannot improve it!



Wikipedia—Big Data

A collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization.



It looks at large data sets and turns raw data into “actionable intelligence” and insights to make better, more valuable marketing decisions.



Why Now?

- ✦ Computer power
- ✦ The price of digital storage
- ✦ Data warehouses are being widely implemented
- ✦ Astronomy, Genetic, etc
- ✦ Amazon, Google, Facebook, Yahoo, Youtube.
- ✦ Censors/RFID



DeVeaux



Big Data

- ✦ 2.5 EB bytes of data is created every day.
2,500,000,000,000,000 bytes
- ✦ More than 30 million sensors are being used.
- ✦ More than 4 billion people were using mobile phones in 2010.
- ✦ 90% of the total data was created in last two years.

10



Who is working on BIGdata?

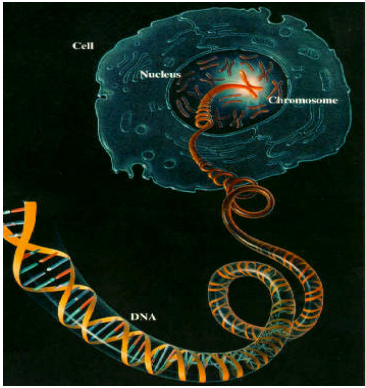
- ✦ Computer Science
- ✦ Information Technology
- ✦ Applied Mathematics
- ✦ Computational Math/Science
- ✦ Operation Research
- ✦ Industrial Engineering
- ✦ Informatics Science/Data Scientist
- ✦ Electronic Engineering
- ✦ Political Science
- ✦ *Anyone who needs funding \$\$\$*




BIG Data

This is a golden age for Statistics,
But not necessary for statisticians.

Genetic Study



Cell
Nucleus
Chromosome
DNA



```

AGAGTTCTGCTCG
AGGTTATGCGCG
CGTTCGGGAATCC
CGTTAGGAAATCT
TCTTTGACGACTC
TCTTAGAGGACTC
          
```

Astronomy









Solar System
Planets
Dwarf Planets


Geology






G.I.S & Geo-Statistics

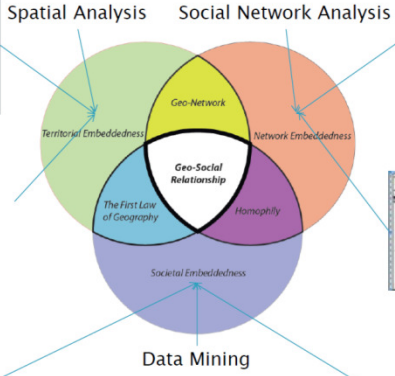
Geo-Social Analysis



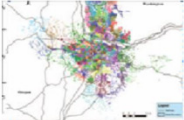
Spatial Analysis



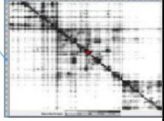
Social Network Analysis



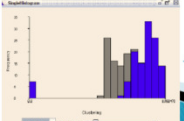
Geo-Social Relationship
The First Law of Geography
Territorial Embeddedness
Network Embeddedness
Homophily
Societal Embeddedness
Data Mining




(Guo 2007)



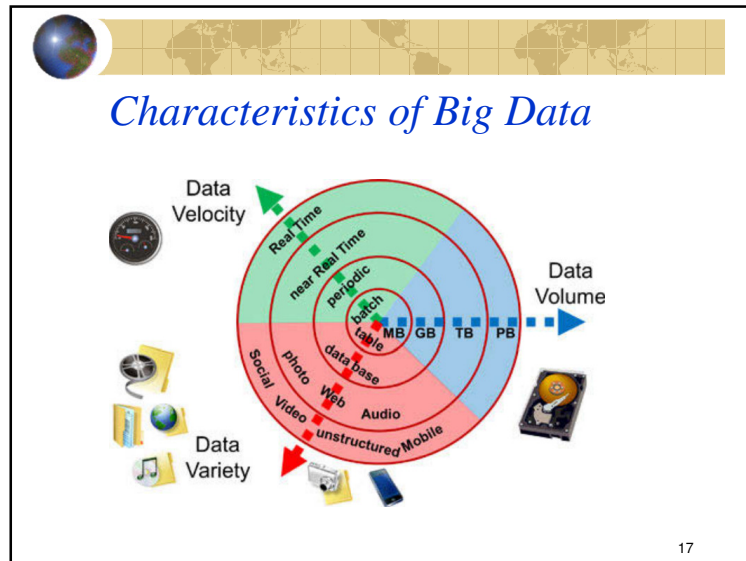
(Guo 2007)





Clusters = 5.57082
Relativity = 269.2
Outdegree = 178
Indegree = 172218.0
Weight = 984357.8
Magnitude = 182.9

<http://www.geovista.psu.edu/GeoSocialApp/>



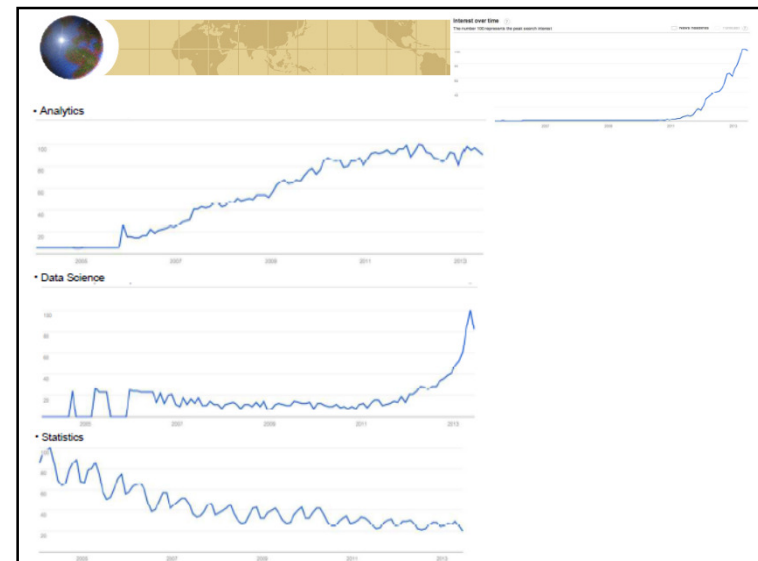
Emphasis on Automatic

- Computer users are domain experts, not computer professionals
 - Too much data
 - Too much technology
 - Not enough useful information
- When is Automatic Analysis Useful?

DeVeaux

BIGdata:

The word “Statistics” was nowhere to be seen!





*BIG Data:
What should be the fourth V?*

Value?



Keep in mind

- ✦ Big data is not necessarily complete, accurate, or true!
- ✦ Value is in the eye of beholder, not the person crunching the numbers!
- ✦ Bigger does not always implies better!
 - ▣ If there is "initial" bias...



*BIGdata typically may
not have much Value at all:
DRIP effect*

Data Rich Information Poor



Challenges

- ✦ 87.5% global data hasn't been really developed and used.
- ✦ Large databases only used in data entry and query— *statistics without deep analysis on big data.*
- ✦ Structured databases are used to store data, resulting in distortion and missing of large amounts on unstructured data.



Randomness

Opportunity for Statisticians

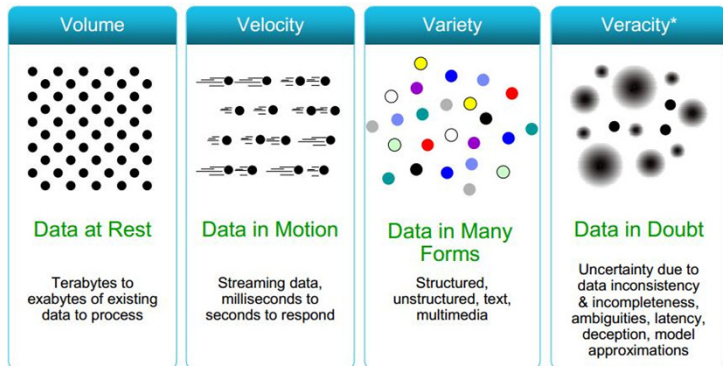


Big Data: 3+1=4 V's

- ✦ Volume
 - ▣ Data at rest
- ✦ Velocity
 - ▣ Data in Motion
- ✦ Variety
 - ▣ Data in many forms
- ✦ Veracity
 - ▣ Data in Doubt



Big Data: The 4V's



Big data in real world

- ✦ Skills
- ✦ Politics
- ✦ Techniques
- ✦ Cognition
- ✦ Privacy



What kind of Stat is needed: General

- ✦ Prediction vs Estimation (Inference)
 - ▣ Only Statistician talk about inference
- ✦ Common Belief
 - ▣ Big data is better than Small data
 - ▣ New Methodologies are so powerful and work well...
 - ▣ The death of p-value and the death of science
- ✦ Fundamental Theory
 - ▣ All theories with iid one population assumption—CLT, LLN, etc



What Statistics: Wish-List

- ✦ Seek out high-impact problems
- ✦ Provide structure for poorly defined problems
- ✦ Develop new theories for new (reality) BIG data world.



Big Data: Basic concerns

- ✦ What to collect?
 - ▣ Bias?
- ✦ How to collect?
- ✦ How to store?
- ✦ How to use?
- ✦ What to use?

- ✦ Analysis and potential risks?



What Stat: Technical Details

- ✦ Statistics of (many descriptive) statistics
 - ▣ how could we summarize thousands of correlations?
 - ▣ How about thousands of p-values? ANOVA's? Regression models? Histograms?
- ✦ Classification and Clustering
- ✦ Low-dimension behavior
- ✦ Feature Extraction (for extremes) and pattern recognition (for norm)
- ✦ New type/structure of data
 - ▣ How to build up a regression (say) model when both input and output variables are network?




Journal of the Royal Statistical Society
Series B

*A
good
Example*


Volume 76 Part 2 2014

Association pattern discovery via theme dictionary models K. Deng, Z. Geng and J. S. Liu	319
Two-sample test of high dimensional means under dependence T. T. Cai, W. Liu and Y. Xia	349
The joint graphical lasso for inverse covariance estimation across multiple classes P. Danaher, P. Wang and D. M. Witten	373
Preadjusted non-parametric estimation of a conditional distribution function N. Veraverbeke, I. Gijbels and M. Omelka	399
Space-time modelling of extreme events R. Huser and A. C. Davison	439
Regularized matrix regression H. Zhou and L. Li	463




Some Remarks

- ✦ "Data have complete meaning in themselves, no theory is required." cf. "Data has no meaning in themselves" (BH², 1978)
- ✦ Using a "Scientific" approach, as oppose to an "Algorithm" approach, to Big data—including data collection, data analysis, mode selection, and feature interpretation.



Susan Hockfield (former MIT President)

- ✦ "Around the dawn of the 20th century, physicists discovered the basic building blocks of the universe; a "parts list", if you will. Engineers said "we can build something from this list," and produced the electronics revolution, and subsequently the computer revolution.
- ✦ More recently, biologists have discovered and mapped the basic "parts list" of life – the human genome. Engineers have said "we can build something from this list," and are producing a revolution in personalized medicine.
- ✦ Who is Building Something Meaningful from the Statistical Science Parts List of Tools?"



*What's
Statistics
All About?*

Classical View



科学探索

◆ Deduction (演绎)

- ❖ Physics (物理)
- ❖ Mathematics (数学)

◆ Induction (归纳)

- ❖ Biology (生物)
- ❖ Chemistry (化学)

◆ Statistics(统计分析) 比较偏重于观察与试验的累积 (也就是“归纳”类型的研究)。



万事万物有其**共同性**,
这使得统计变成**可能**;

万事万物有其**不同性**,
这使得统计变成**必须**。

Things are the same, this makes Statistics possible.
Things are different, this makes Statistics necessary.



鏡子與窗子

成功者
失敗者



*A
Data Explosion
is Coming!*

Are You Ready?

RFID Study Group



BIGdata: Statistics is essential!

Speak Loud and
Remind all your friends and enemies!



Computer Science Root

- ✦ Database
- ✦ *SVM: Support Vector Machine*
- ✦ *ANN: Artificial Neural Network*
- ✦ MBA: Market Basket Analysis (Association Rule)
- ✦ Genetic Algorithms
- ✦ OLAP: On-Line Analytic Processing
- ✦ Link Analysis
- ✦ High Dimensional Plots
- ✦ KDD Process
- ✦ Machine Learning
- ✦ Text Mining



Statistical Root

- ✦ *(Linear & Nonlinear) Regression*
- ✦ Logistic Regression
- ✦ Classification
- ✦ Clustering
- ✦ Density Estimation, Bumps and Ridges
- ✦ Time Series: Trend & Spectral Estimation
- ✦ Dimension Reduction Techniques
- ✦ High Dimensional Plots
- ✦ Classical Multivariate Methods



Statistical & Computer Science Mixture

- ✦ Data Visualization
- ✦ Dimension Reduction
 - ✦ Multi-dimensional Scaling
 - ✦ Local Linear Embedding
 - ✦ Principal Component & Principal Manifolds
 - ✦ Local Tangent Space Alignment
- ✦ Independent Component Analysis
- ✦ K-means & kd-Tree
- ✦ Support Vector Machine



Which Professions???

Statistics

- ✦ All kind of errors
 - ✦ Type-I error
 - ✦ Type-II error
 - ✦ Pure error
- ✦ Lack of fit
- ✦ Loss function
- ✦ Failure Rate
- ✦ Hazard Model
- ✦ Penalty
- ✦ False Discovery Rate
- ✦ Risk ...

Computer Science

- ✦ Faithfulness
- ✦ Greedy Search
- ✦ Smart Algorithm
- ✦ Intelligent Procedure
- ✦ Golden Standard
- ✦ Knowledge Discovery
- ✦ ...



Old Business Model



New Business Model



STILL QUESTION?



Send \$500 to

- ✦ Dennis Lin
University Distinguished Professor
*317 Thomas Building
Department of Statistics
Penn State University*
- ✦ +1 814 865-0377 (phone)
- ✦ +1 814 863-7114 (fax)
- ✦ DennisLin@psu.edu



(Customer Satisfaction or your money back!)