



Ghost Data (幽灵数据)

Dennis Lin
University Distinguished Professor
Department of Statistics
The Pennsylvania State University



Statistics: Two Directions

Probability-driven
Data-driven




Statistics based upon Probability

- ✦ Probability Theory
- ✦ Likelihood Function
- ✦ Maximal Likelihood Principle
 - ❖ Point Estimate
 - ❖ Hypothesis Testing
 - ❖ Inference
- ✦ The role of the data—
 - ❖ if you have data, great!
 - ❖ If not, that's fine too!



Likelihood Principle

- ✦ Full Likelihood
- ✦ Partial Likelihood
- ✦ Empirical Likelihood
- ✦ Pseudo Likelihood
- ✦ Quasi Likelihood
- ✦ Penalized Likelihood
- ✦ Posterior Likelihood
- ✦ Composite Likelihood
- ✦ Profile Likelihood



Statistics based upon data


EDA: Exploratory Data Analysis (Tukey)

- ✦ Learning from Data
- ✦ What can the data tell you?






- ✦ Show me your problems?
- ✦ Show me your data?
- ✦ Tell me how were the data collected?
- ✦ What can the data tell you (& how)?




Where did Data come from?



The Evolution of Science


- **I. Observational Science**
 - ☞ Scientist gathers data by direct observation
 - ☞ Scientist analyzes data
- **II. Analytical Science**
 - ☞ Scientist builds analytical model
 - ☞ Makes predictions.
- **III. Computational Science**
 - ☞ Simulate analytical model
 - ☞ Validate model and makes predictions
- **IV. Data Exploration Science**
 - ☞ **Data-driven science**
Data captured by instruments or data generated by simulation
 - ☞ Processed by software
 - ☞ Placed in a database / files
 - ☞ Scientist(s) analyze(s) database / files
 - ☞ Access crucial




Fourth Paradigm — Jim Gray's (2009)

The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science

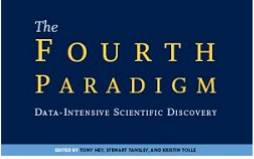


Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

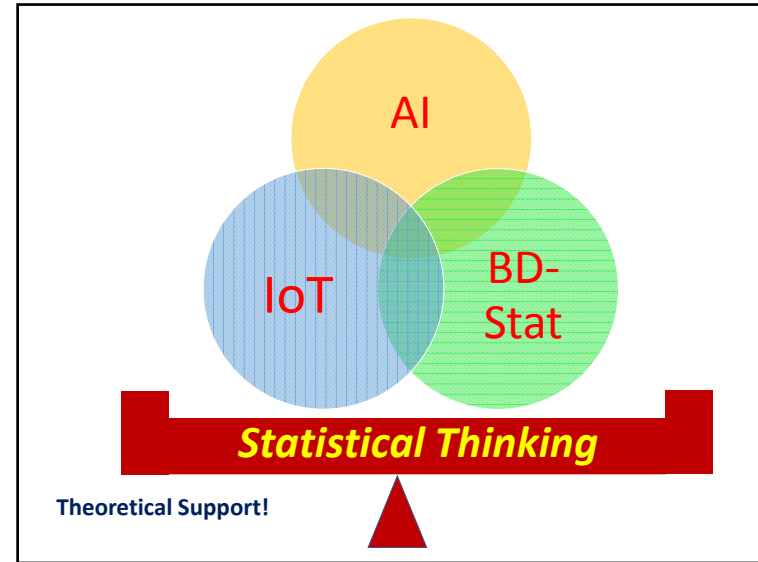
In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

- Download
 - [Full text](#)
 - [By chap](#)
- Related Re
 - [Microsoft](#)
 - [eScience](#)






The fifth Paradigm?




Journal of Data Science (since 2002)

Journal of Data Science
an international journal devoted to applications of statistical methods at large

Editorial Board

Chief Editor:

- Michael T. S. Lee, College of Management, Fu-Jen Catholic University, New Taipei City

Executive Managing Editors:

- Ming-Chih Chen, College of Management, Fu-Jen Catholic University, New Taipei City
- Han-Sheng Wang, Department of Business Statistics & Econometrics, Peking University, Beijing
- Bo Zhang, School of Statistics, Renmin University, Beijing

Associate Editors:

- Zhong-Bai, Department of Mathematics, Northeast Normal University, Changchun
- Mingxiao Bao, Research Institute of State Sports Bureau, Beijing
- Carmon Bateman, Department of Mathematics Education, University of Granada, Granada
- Michael V. Chan, Biology Department, California State University at San Bernardino, California

- Dennis K. J. Liu, Department of Statistics, Penn State University, University Park
- Shuangge Steven Ma, School of Public Health, Yale University, Connecticut
- Bing Han, Department of Statistics and Finance, University of Science and Technology of Hefei
- Xizhi Wu, Data Mining Center, Renmin University, Beijing

M.T. Chao

JDS
Journal of Data Science
Volume 12 Number 1 January 2014

Two Factor Stochastic Modeling with Generalized Hyperbolic Distribution
Seyed Saad Alnabi and Patrick Golleker.....1

Does Satisfaction or Anxiety Drive Consumer Demand?
Gordon G. Borchert.....19

A New Variable Selection Approach Inspired by Supermodular Design: Case a Large Dimensional Dataset
Christos Pappas, Kyriakos Ouzas, Christos Katsourakis and Kallipe Nifliou.....33

The Weibull of Family of Probability Distributions
Manohar Bhargava, Rodrigo B. Silva and Gerson M. Cordeiro.....33

J-Measures Estimation for Mixtures of Weibull Distributions
Elke Utrichter and Marc Engelke.....49

The Chen of α 's Extreme Value Distribution
Sanku Sanjib and Sumita Singh.....87

A New Class of Survival Regression Models with Cox Fraction
Edwin M. M. Ortega, Chedy D. C. Ortega, Elizabeth M. Holliman, Vicente G. Cancho and Gerson M. Cordeiro.....107

On the Estimation of Probability Model for the Number of Female Child Birth among Females
Pratik Sanjib, Sanku Sanjib and Sumita Singh.....137

Estimation for the Parameter of Poisson-Exponential Distribution under Bayesian Paradigm
Sanjay Kumar Singh, Anshu Singh and Mitesh Kumar.....157

A Comparison of Statistical Tests for Identifying Modality in Body Mass Distributions
Ling Xu, Edward J. Dubicki, Timothy Hansen and Carlo Rovato.....175

Does Globalization Mitigate Income Inequality?
Gordon G. Borchert.....197

Contents of Volume 12.....1

Copyright 2014 Journal of Data Science
ISSN 1548-7657 (print) 1548-7665 (online)
Hempel, New York, Beijing

Ghost Data

DennisLin@psu.edu



 **Ghost Data:**

*from the art of luck
to the
management of ghosts*



The Sixth Sense





***"I can see things
that you cannot see!"***




 **Design & Analysis of
Ghost Data**





Sixth Senses



*I can analyze
the data
that you cannot see!*



The Sixth Sense



*A theory of the sixth sense
—the frequencies spectrum!*







***Can we "tune" our frequency to read
those out-of-range "ghost" messages?***


 **Public Health—bacteria**






- ✦ Bacteria is everywhere...
- ✦ We cannot see *bacteria*, until the invention of *microscope*.
- ✦ Can we develop some new tools/methodologies to improve our “seeings?”

 **Some Statistical Telescopes**

- ✦ Sensors 
- ✦ RFID 
- ✦ Machine Data 
- ✦ more

 **Ghost Data—**
*As natural as the real data,
ghost data is everywhere
—it is just data that you cannot see.*

Need to learn
*how to handle it,
how to model with it, and
how to put it to work.*

 **Ghost Data—examples** 
*It is as natural as real data.
It's just data that isn't there.*
(John Sall)

- ✦ Virtual data (虚拟数据)
 - ▣ *Virtual Reality—it isn't there until you look at it;*
- ✦ Missing data (缺失数据)
 - ▣ *there is a slot to hold a value, but the slot is empty;*
- ✦ Pretend data (做作数据)
 - ▣ *data that is made up;*
- ✦ Simulation data (模拟数据)
 - ▣ *data to answer "what if."*
- ✦ Highly Sparse Data (高度稀疏数据)
 - ▣ *whose absence implies a zero*



Sherlock Holmes

A poster for the movie 'Sherlock Holmes' featuring Robert Downey Jr. as Sherlock Holmes and Jude Law as Dr. Watson. The title 'SHERLOCK HOLMES' is at the top, and 'A GAME OF SHADOWS' is at the bottom.

Sherlock Holmes

—The dog that did not bark at night

- ✦ Gregory (Scotland Yard detective): *"Is there any other point to which you would wish to draw my attention?"*
- ✦ Holmes: *"To the curious incident of the dog in the night-time."*
- ✦ Gregory: *"The dog did nothing in the night-time."*
- ✦ Holmes: *"That was the curious incident."*



A background image of a world map in a light tan color.



*Absence of evidence is **NOT** evidence of absence*

(in fact, it can be evidence of something).

A silhouette of Sherlock Holmes wearing his iconic deerstalker hat and smoking a pipe, holding a magnifying glass.

Sherlock Holmes

✦ "The best clue is not always what you see, but what you didn't see."

—Sherlock Holmes

(The dog didn't bark in the night)

- ✦ The missing values can be the most predictive of the predictors.
- ✦ Need to learn to work well with data that isn't there—*Ghost Data!*



Absence of Expected Facts

Expected—model based
Absence—data based



DMNAR— Data Missing Not At Random



Don Rubin (1972, *Biometrika*)
Special Issue of *Statistica Sinica* (2018)

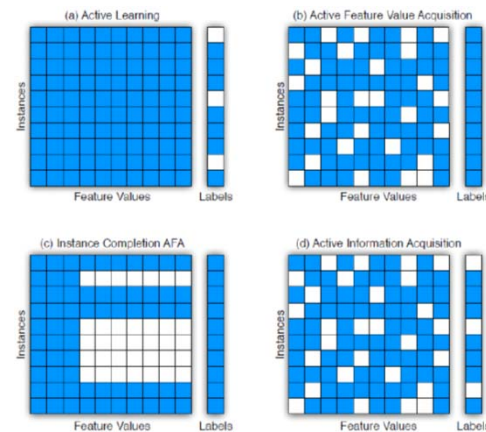


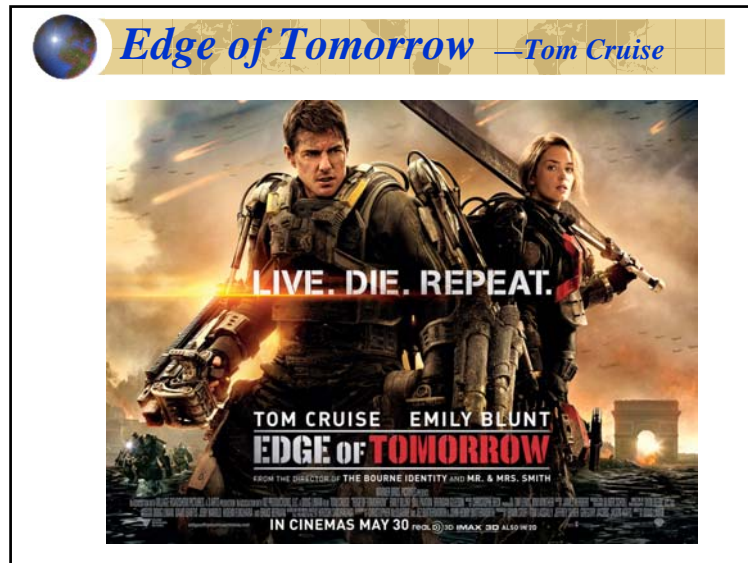
When a student receives “0” in her/his exam...this could be

- ✦ Did come & hand in Exam, and indeed get a “0”
 - ✦ Not a missing
- ✦ Did not come to Exam
 - ✦ Missing in Action
- ✦ Did come & hand in Exam, but get lost in process
 - ✦ Missing at Random
- ✦ Did come but did not hand in Exam
 - ✦ Missing Not at Random
- ✦ etc



Pattern of Missing—Medical Example





Spear and Shield


- ✦ "My **spear** can penetrate anything (including shield)!"
- ✦ "My **shield** can block anything (including spear)!"



矛盾 (Mao-Spear/Dun-Shield)
They are contradiction, but they can learn from each other!
—this is the basic idea of AlphaGo-Zero




How to speed up the learning process?

How to improve the success probability of the active learning?




Ghost Movie—Three Movies

- ✦ The Sixth Sense
 - ❑ *I can see things that you cannot see*
- ✦ Sherlock Holmes
 - ❑ *absence of expected facts*
- ✦ Edge of Tomorrow
 - ❑ *how to speed up your learning*




Statistical Microscope

- ✦ You have to decide
*What to see, and
How to see?*
- ✦ The power of DoE
(Design of Experiment)




Body and Soul—

*you can see **body**,
you cannot see the **soul!***

***X** is what you “observed” (**body**), while
X* is the “truth” (**soul-ghost**)!*

- ✦ $X = X^*$ lucky you!
- ✦ **X** is **X*** “plus” some kind of noises
 - ❑ $X = X^* + \Delta$ or $X = X^* \cdot \delta$, or any other form
- ✦ **X** is a low-dimension projection of **X***
- ✦ **X** is a “transformation” of **X***
 - ❑ Confidentiality/privacy Data
- ✦ **X** is empty (while **X*** is not)

Walking Dog...

This is X* (arrow pointing to paw prints)

This is X —what you've observed (text below dog)

Example—X and X*

- What you see this
- What the camera sees this

This is the "data" that will be analyzed!

194	213	201	212	199	213	211	195	178	158	192	209
180	189	193	221	209	203	191	157	147	115	129	163
114	125	140	109	175	165	152	140	170	106	78	80
87	103	115	154	148	142	149	153	175	101	57	57
102	112	105	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	119	107	87	69	67
68	71	62	98	89	92	98	95	89	88	76	67
41	55	68	99	63	45	60	82	58	76	75	65
20	43	69	75	55	41	51	73	53	70	63	44
63	63	57	62	75	71	73	74	52	68	59	27
72	59	53	65	84	92	84	74	57	72	63	42
67	61	55	65	75	78	76	73	59	75	69	30

Flow vs. Snapshot

What we observe—flow

What we analyze—snapshot

The missing slides in between can be view as Ghost data —no matter how high its frequency is.

Big Data: The 4V's

Volume	Velocity	Variety	Veracity
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations




Variety: New Types of Data

- ✦ Text Data
- ✦ Audio Data
- ✦ Image Data
- ✦ Video Data
- ✦ Social Media Data
- ✦ Network Data
- ✦ Finger Print Data
- ✦ Citation Data (ranking/rating)
- ✦ etc...

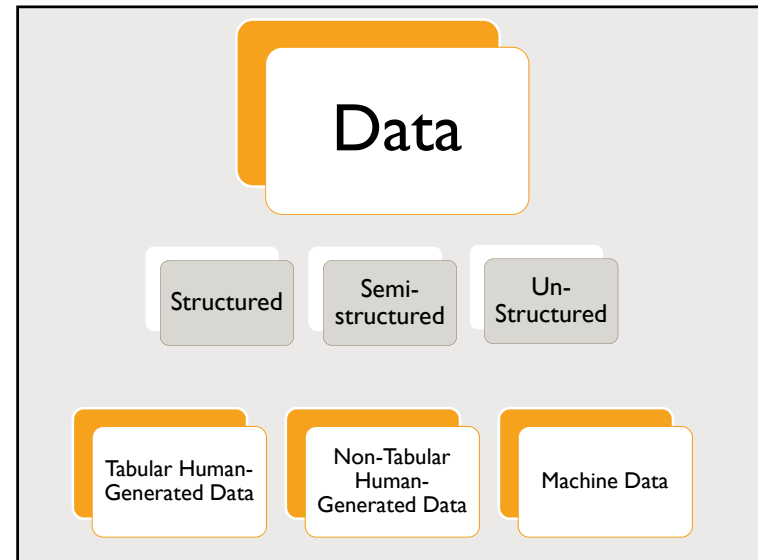


Machine Data

Machine (Generated/Collected) Data

- ✦ **Machine (Generated) Data**
 - ❑ Computer Simulation/Experiment
 - ❑ Pretend Data
 - ❑ AI data
- ✦ **Machine (Collected) Data**
 - ❑ Sensors (Security Camcorder etc)
 - ❑ Traffic
 - ❑ Stock ...



Why Don't Traditional models fit?

<p>Data Structure</p> <ul style="list-style-type: none"> ✦ Traditional data is structured ✦ Machine data is a mix of data structure types 	<p>Data Complexity</p> <ul style="list-style-type: none"> ✦ Loses the ability to present a simple and clear picture of nature's mechanism ✦ Too complex to fit to a pre-existing model
--	---

Camera: new and old





Data: new and old
Statistics: new and old

Observational Data
vs.
Inferred or Derived Values?

The Others — Nicole Kidman





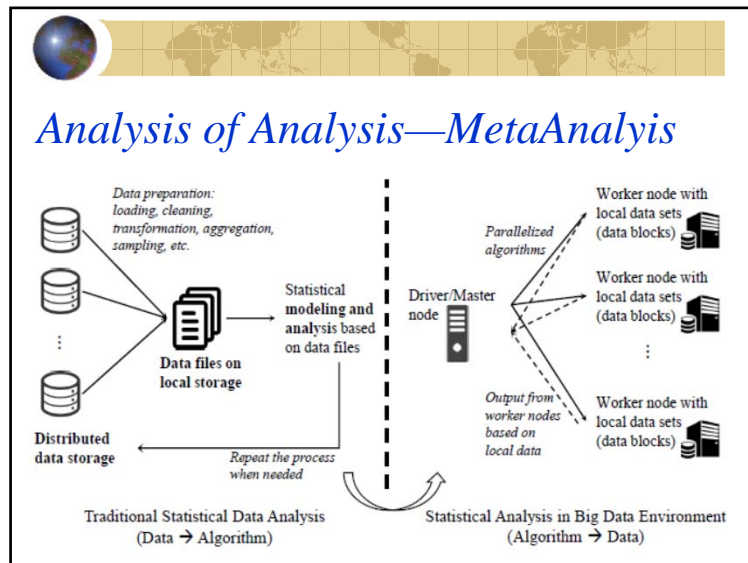
 *When I encounter "ghost data," should I assume if I am a **human** or, if I too am a **ghost** ???*

Is ghost data problematic to AI's, or only to human statisticians?



 *Data of Data*

Network of Network




 *Analysis of Analyzes*

Meta-analysis etc



Ghost Data—examples

*It is as natural as real data.
It's just data that isn't there.*

(John Sall)

- ✦ Virtual data (虚拟数据)
 - ❖ *Virtual Reality—it isn't there until you look at it;*
- ✦ Missing data (缺失数据)
 - ❖ *there is a slot to hold a value, but the slot is empty;*
- ✦ Pretend data (做作数据)
 - ❖ *data that is made up;*
- ✦ Simulation data (模拟数据)
 - ❖ *data to answer "what if."*
- ✦ Highly Sparse Data (高度稀疏数据)
 - ❖ *whose absence implies a zero*



Are they *i.i.d. Normal ???*

If so, I have a lots to offer...

Central Limit Theorem, Asymptotical Normality,
Law of Large Number, etc.

If not...well...



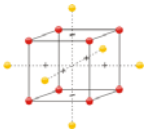
Ghost Data—some extensions

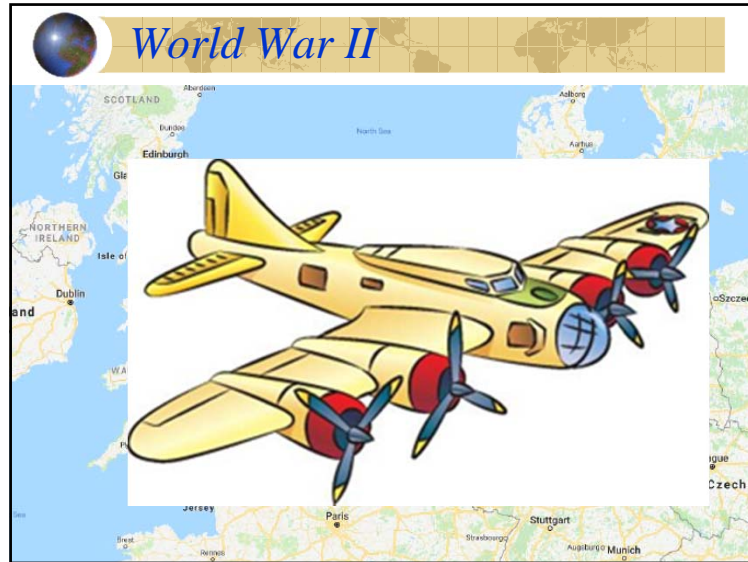
- ✦ Optimization via Simulation
- ✦ Partition Model
- ✦ Two-Stage Least Squares Estimate
- ✦ Hidden Markov Chain
- ✦ Adversarial Learning
- ✦ Topological data
- ✦ Feature engineering
- ✦ Unknown tuning parameters



Ghost Data for Industrial Statistics

- ✦ Design of Experiment
 - ❖ *Decide what to see and how to see?*
- ✦ Reliability/Survival
 - ❖ *What we see is "dead" (failures), but what we are interested in is "alive".*
- ✦ Control Charts
 - ❖ *How to monitoring Ghost data?*
(*Ghost Hunting—TV series*)





Ghost Data in Economics

Ghost with one hand
Ghost with two hands —
on one hand...on the other...

Invisible Hand

Adam Smith's : "Invisible Hand"

"The sole purpose of all production is to provide the best possible goods to the consumer at the lowest possible price. Society should assist producers of goods and services only to the extent that assisting them benefits the consumer...he intends his own gain; and he is in this, as in many other causes, led by an invisible hand to promote an end which was no part of his intention...By pursuing his own interest, he frequently promotes that of society."

Price
Quantity
Equilibrium
Q*

Government does not get involved
Needs of society automatically met
Profit-seeking producers will make more
Competition keeps quality high
Competition keeps prices low
Competition & self-interest act as an invisible hand that regulates the free market

A slide titled "Ghost Data in Economics" with a supply and demand graph. The graph shows a downward-sloping Demand curve and an upward-sloping Supply curve intersecting at an Equilibrium point (Q*, P*). To the right is an illustration of a hand with text describing the "Invisible Hand" concept.

One-Handed Economist

*Economist—on one hand, since..., it is A;
one the other hand, if ..., it could be A^c.*

Give me a one-handed economist!
All my economists say, On the one hand on the other.

— Harry S. Truman —

AZ QUOTES

Statistician—with probability 95%, it is A.
Mathematician—it is A.

A slide titled "One-Handed Economist" featuring a portrait of Harry S. Truman and a quote. The quote is: "Give me a one-handed economist! All my economists say, On the one hand on the other." Below the quote is the signature "— Harry S. Truman —" and the source "AZ QUOTES". At the bottom, it says "Statistician—with probability 95%, it is A. Mathematician—it is A."

Ghost Money **bitcoin**

Bitcoin

US\$1.00 = 1,000 (2006)

1.00 = US\$20K+ (December, 2017)

What Next?

A slide titled "Ghost Money" featuring a Bitcoin logo and the word "bitcoin". It shows the price of Bitcoin in 2006 (US\$1.00 = 1,000) and in December 2017 (1.00 = US\$20K+). The slide ends with the question "What Next?"



Ghost Data—some examples

*It is as natural as real data.
It's just data that isn't there.* (John Sall)

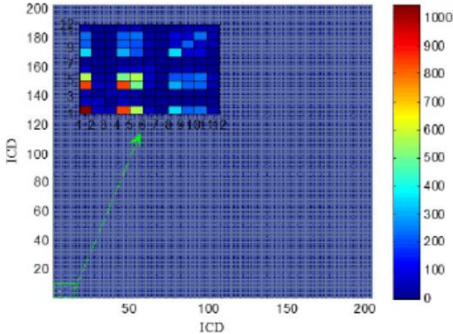
- ✦ Virtual data (虚拟数据)
 - ❏ it isn't there until you look at it;
- ✦ Missing data (缺失数据)
 - ❏ there is a slot to hold a value, but the slot is empty;
- ✦ Pretend data (做作数据)
 - ❏ data that is made up;
- ✦ Simulation data (模拟数据)
 - ❏ data to answer "what if."
- ✦ Highly Sparse Data (高度稀疏数据)
 - ❏ whose absence implies a zero

Virtual data —
it isn't there until you look at it.

- ✦ **Movie Plus Game**
 - ❏ Have to decide what to see —design of experiment
 - ❏ Design and Analysis of Virtual Reality Data
- ✦ Detect whether there is a bug—software reliability (see Kevin Quinlan, t-covering)
- ✦ Compact Urban Cell

Highly Sparse Data
—whose absence implies a zero



Haiyen Yu

Missing data —
there is a slot to hold a value, but the slot is empty

- ✦ Informative Missing
- ✦ Informative Present

W_i	$Y_i(1)$	$Y_i(0)$
1	✓	??
1	✓	??
...
1	✓	??
0	??	✓
0	??	✓
...
0	??	✓



Notations

- ✦ Y: incomplete variable
- ✦ R: response indicator
 - ▣ R = 1 if Y is observed; R=0 otherwise
- ✦ X: fully observed covariate
- ✦ Y_{obs} and Y_{mis} : the observed and missing parts of Y



Missing (*not*) At Random

- ✦ Imputation under MAR (Missing at Random)

$$P(Y|X, R = 0) = P(Y|X, R = 1)$$
- ✦ Imputation under MNAR (Missing Not at Random)

$$P(Y|X, R = 0) \neq P(Y|X, R = 1)$$



Government data— *public data under protection (confidentiality)*

- ✦ Lin in Advisory Board for
 - ▣ Bureau of Statistics (Taiwan) &
 - ▣ State Bureau of Statistics (China)
 since 1998.
- ✦ There is Statistics Canada, but there is no Bureau of Statistics in USA—try NISS!



Pretend data — *data that is made up*

Under Construction!



Simulation data —data to answer “what if.”

Design of Computer Experiment
(**Stat 597A** in 18S by Lin)



Design of Experiments for Big Data Era

Joint Statistical Meetings 2017
Round Table Lunch Discussion#21
(Lead by Lin)



DoE for Big Data

Conventional Statistical Methods

- ✦ Random
- ✦ i.i.d.
- ✦ Normal
- ✦ etc ...

**Broken
Gap**

Big Data Machine (Generate) Data

- ✦ “Just in case”
 - ▣ sensor
 - ▣ RFID
 - ▣ Camcorder
 - ▣ etc ...



Two folds

- ✦ *Is the conventional statistical methods ideal for model/analysis ideal for machine data?*
- ✦ **NO!**
- ✦ *If the conventional statistical modeling/analysis method is to be used, what should we do (design) about the big data?*
- ✦ **Data Aggregation/Fusion** (for example).

Different Culture

✦ Large Data Set

Statistics: $n > 30$

Comp Sci: $n > 30\text{GB}$ (or even 30TB)

✦ Dimension Reduction

Statistics: from $d=10$ into $d=2$

Comp Sci: from $d=10^{10}$ to $d=10^2$.

✦ Ultra High-Dimension

Statistics: $d=50$

Comp Sci: $d=10^{12} = 1,000,000,000,000$

✦ Fast Computing

Statistics: from $O(n^{1/4})$ to $O(n^{1/5})$

Comp Sci: from 20 minutes to 6 seconds (4G→5G).

Real & Ghost

Live Happily Ever After...



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!)