Promoting Statistics for the Youth

August 28, 2013

Byeong U. Park

Department of Statistics

Seoul National University

What is Statistics?

- A study of the
  - *collection,*
  - *analysis,*
  - *interpretation*

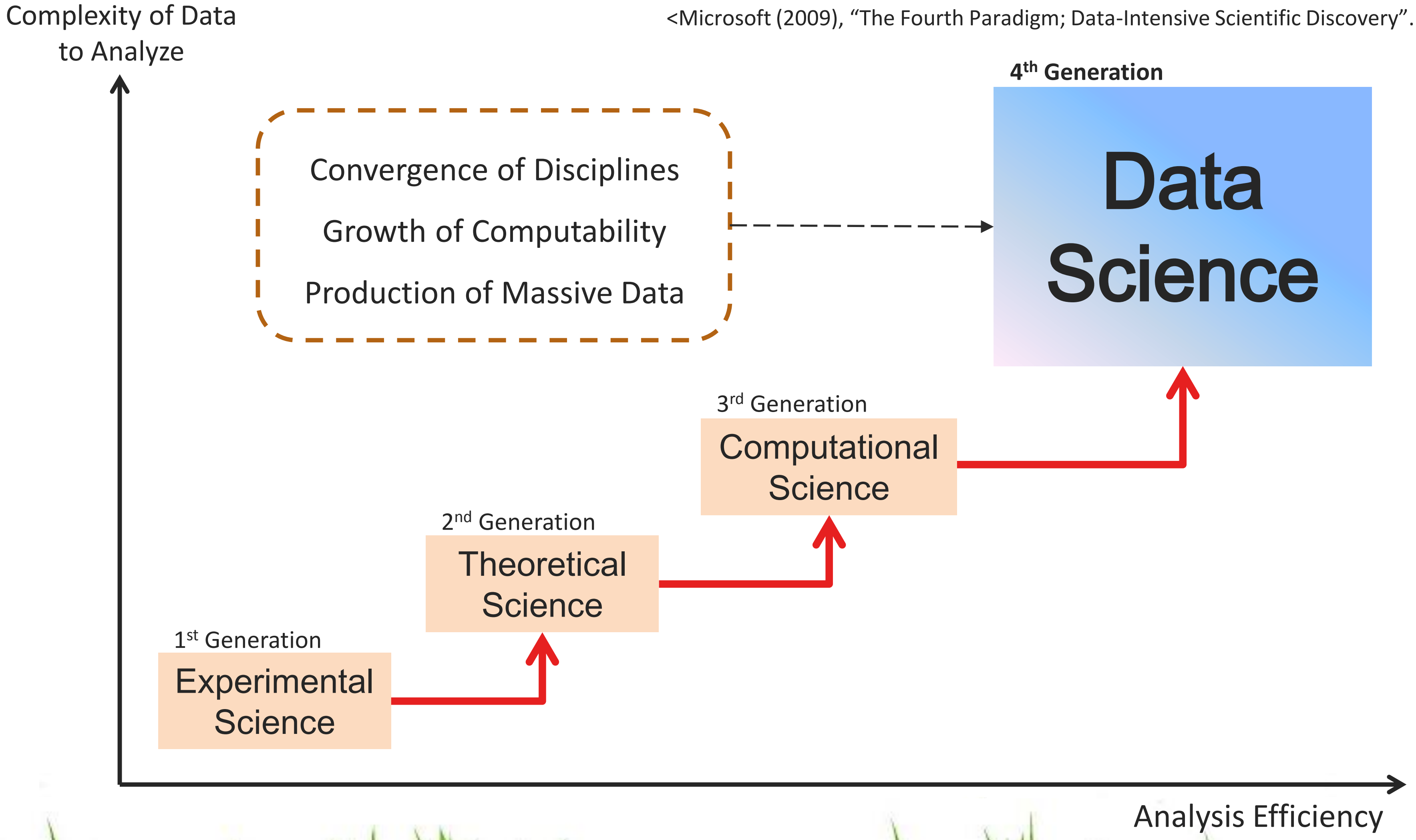
  of **Data**.
- Linked to and applied to various real problems in industry as well as scientific problems in other disciplines, founded on mathematical thinking.

What is Statistics?

100

Science Paradigm for 21st Century

<Microsoft (2009), “The Fourth Paradigm; Data-Intensive Scientific Discovery”. >



Complexity of Data to Analyze

Convergence of Disciplines

Growth of Computability

Production of Massive Data

4th Generation

Data Science

3rd Generation

Computational Science

2nd Generation

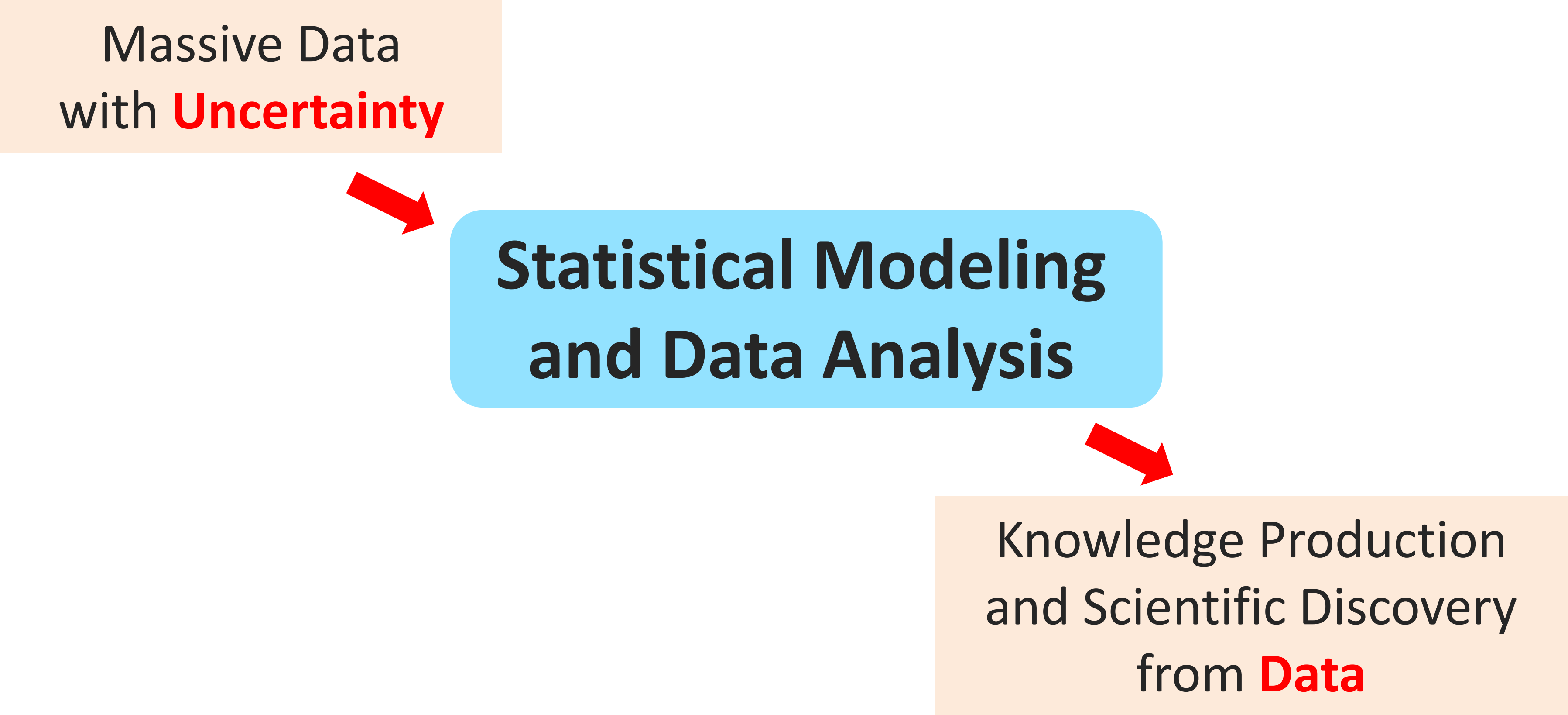
Theoretical Science

1st Generation

Experimental Science

Analysis Efficiency

Role of Statistics in Data Science



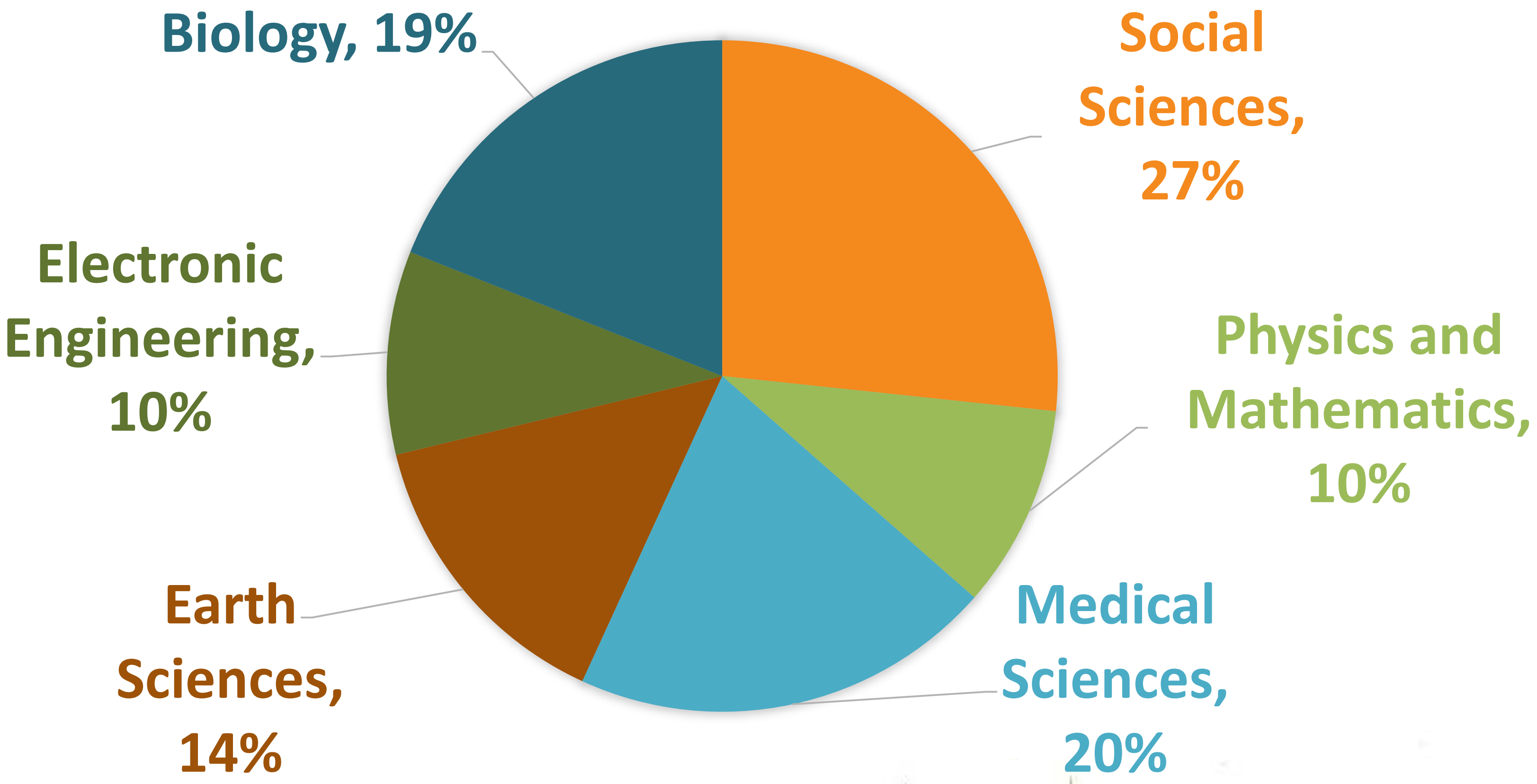
Massive Data
with **Uncertainty**

Statistical Modeling and Data Analysis

Knowledge Production
and Scientific Discovery
from **Data**

Far-reaching Influence of Statistics

- Citation distribution of Bootstrap paper (Efron, 1979)



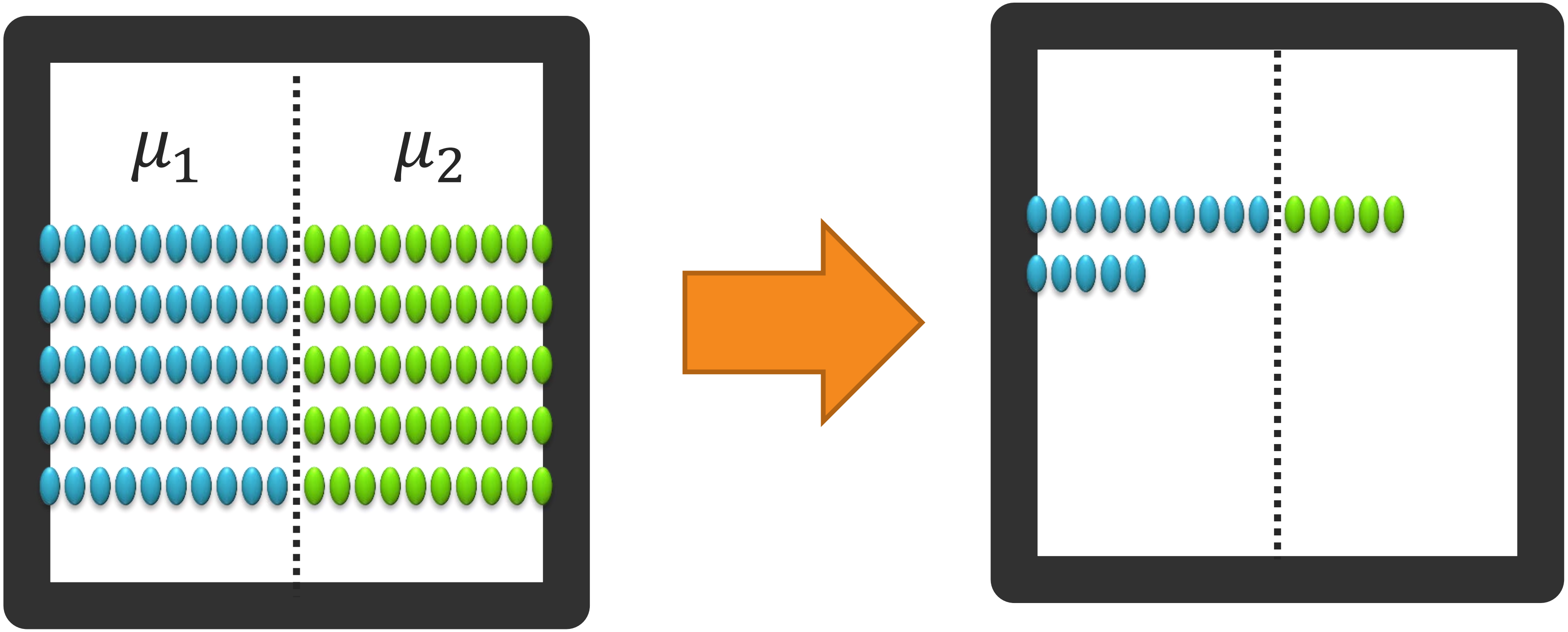
Far-reaching Influence of Statistics

Statistical Topic	Author	Journal Title	Citation*
EM Algorithm	Dempster et al. (1977)	Journal of the Royal Statistical Society Series B	33,645
False Discovery Rate	Benjamini and Hochberg (1995)	Journal of the Royal Statistical Society Series B	17,964
Co-integration**	Engle and Granger (1987)	Econometrica	20,249

* 2013.06.06

** Nobel Prize in Economic Sciences (2003)

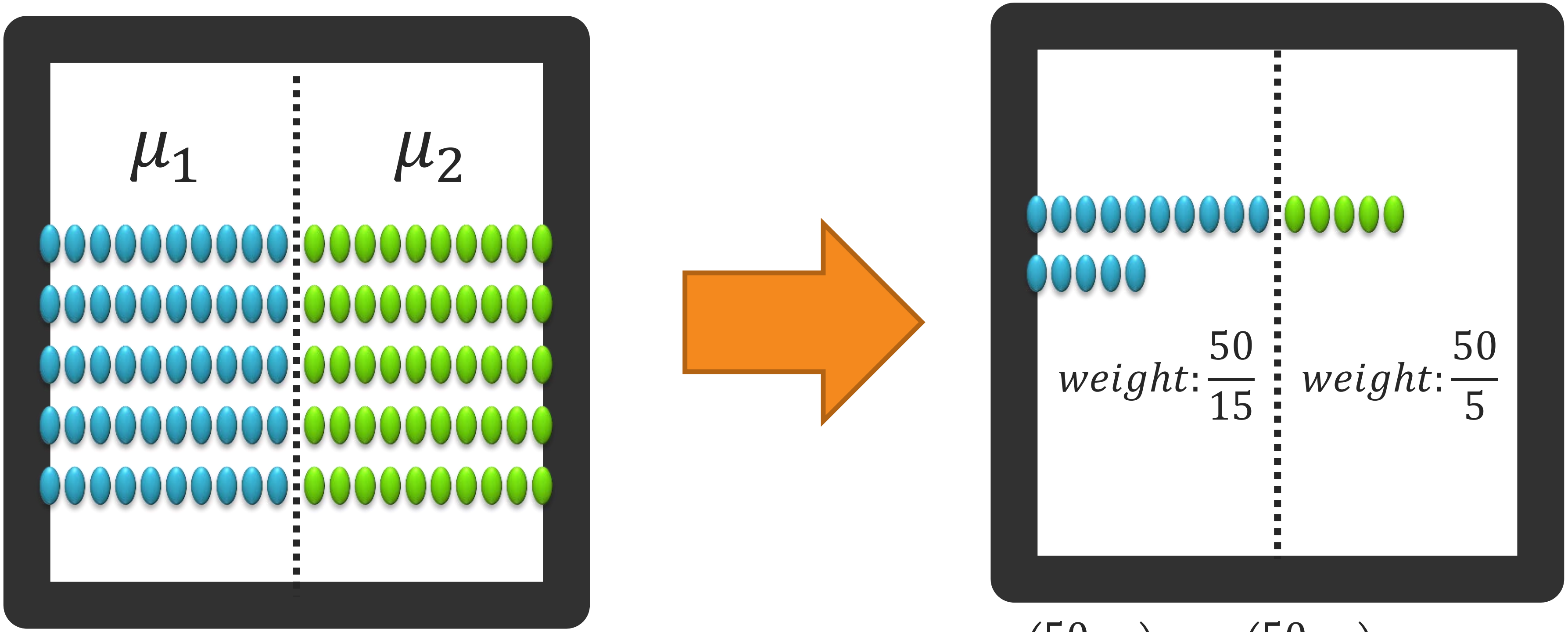
Ball Sampling



$$Average\colon \frac{\mu_1 + \mu_2}{2}$$

$$Average\colon \frac{15\mu_1 + 5\mu_2}{20} \left(\neq \frac{\mu_1 + \mu_2}{2} \right)$$

Ball Sampling



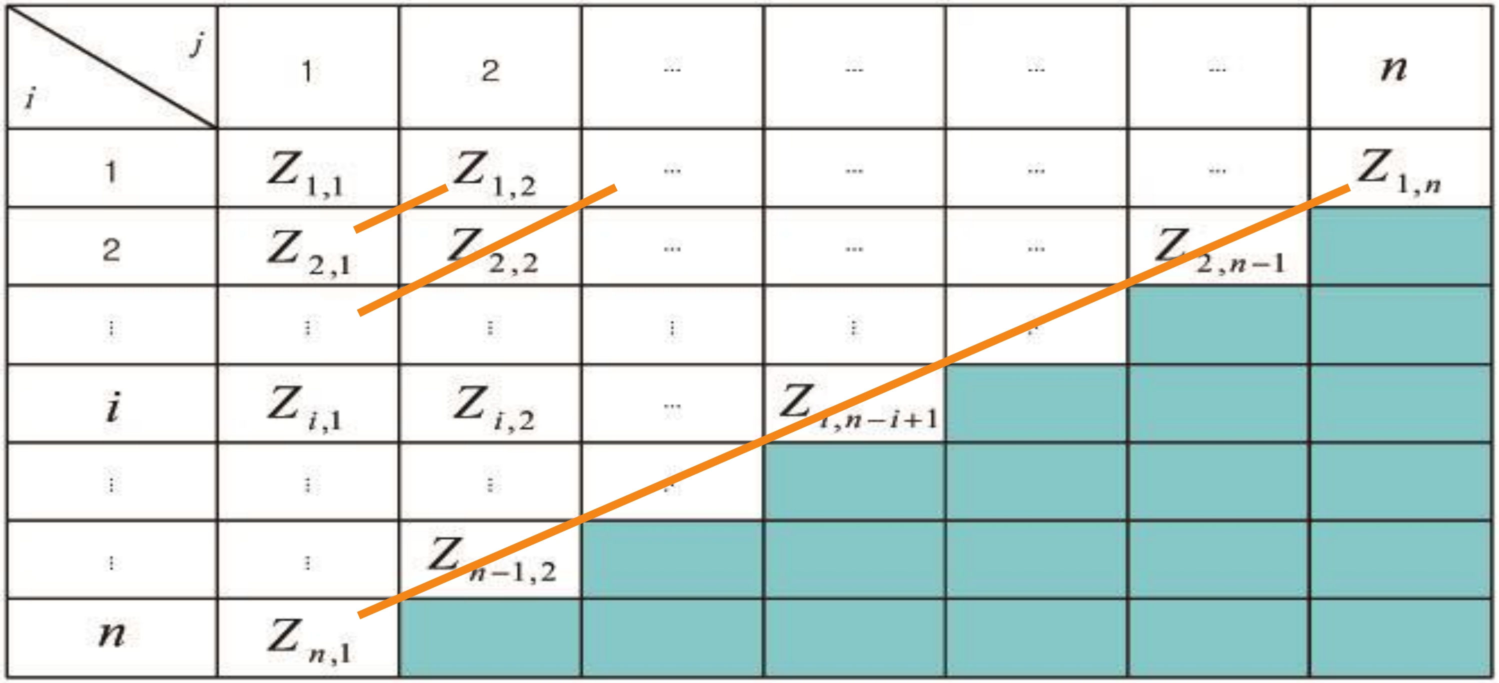
$$Average: \frac{\mu_1 + \mu_2}{2}$$

$$Weighted\ Average: \frac{15\left(\frac{50}{15}\mu_1\right) + 5\left(\frac{50}{5}\mu_2\right)}{15 \times \frac{50}{15} + 5 \times \frac{50}{5}} \left(= \frac{\mu_1 + \mu_2}{2}\right)$$

Credit Loan Survey

- Investigated the relation between overdue status of credit loan and employment status to establish a strategy for allowing loan or not.
- Found that the group of people in their employment has a much higher overdue rate than the group of unemployed.
- **Why?**

Chain-Ladder Data



j i	1	2	…	…	…	…	n
1	$Z_{1,1}$	$Z_{1,2}$	…	…	…	…	$Z_{1,n}$
2	$Z_{2,1}$	$Z_{2,2}$	…	…	…	$Z_{2,n-1}$	
⋮	⋮	⋮	⋮	⋮	⋮		
i	$Z_{i,1}$	$Z_{i,2}$	…	$Z_{i,n-i+1}$			
⋮	⋮	⋮	⋮				
⋮	⋮	$Z_{n-1,2}$					
n	$Z_{n,1}$						

(e.g. 1) Future liability for insurance
i = Accident year, j= development year

(e.g. 2) Divorce rate
i = Marriage year, j=Divorce year

Body Fat Data

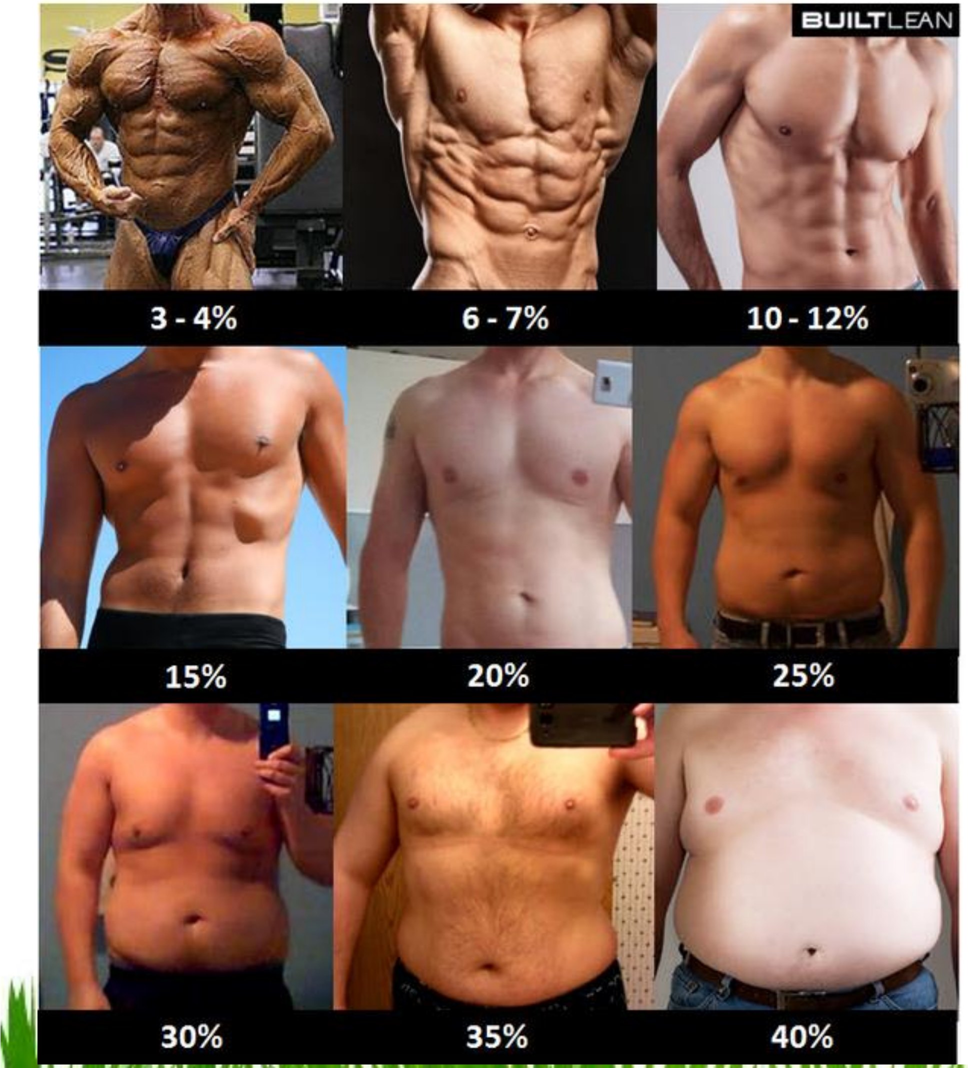
- Body Fat (%) = $\frac{\text{Total Weight of Fat}}{\text{Weight}} \times 100$
- Typical Body Fat Amounts

Description	Men	Women
Exceptionally lean	6 ~ 10 %	10 ~ 15 %
Very lean	11 ~ 14 %	16 ~ 20 %
Lean*	15 ~ 18 %	21 ~ 25 %
Moderate	19 ~ 24 %	26 ~ 29 %
Obese	≥25 %	≥30 %

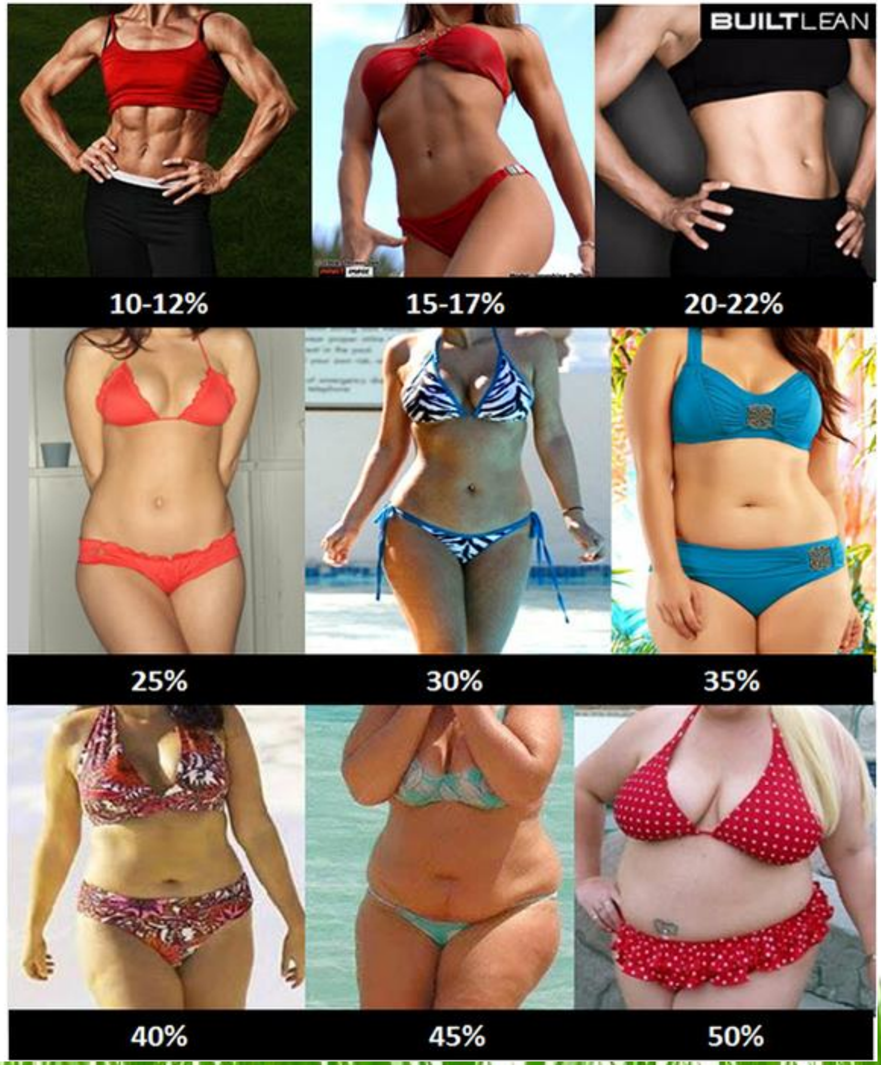
* Thin but looks strong and healthy

Body Fat Data

Men



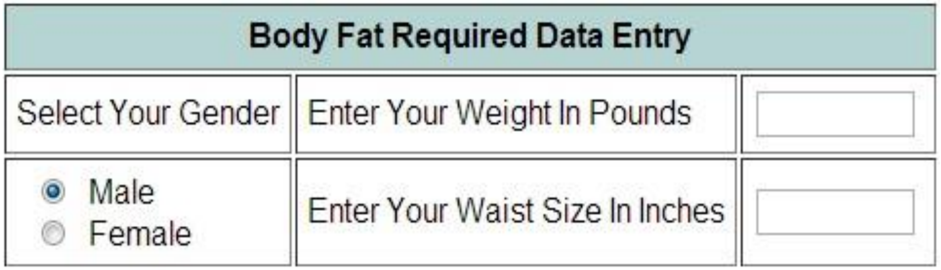
Women



Body Fat Data

- Body fat percentage is a good health indicator.
- How to measure body fat percentage?
- MRI, CT: Accurate, but **expensive**

Cheap but less accurate equipment



Body Fat Required Data Entry		
Select Your Gender	Enter Your Weight In Pounds	
● Male ○ Female	Enter Your Waist Size In Inches	

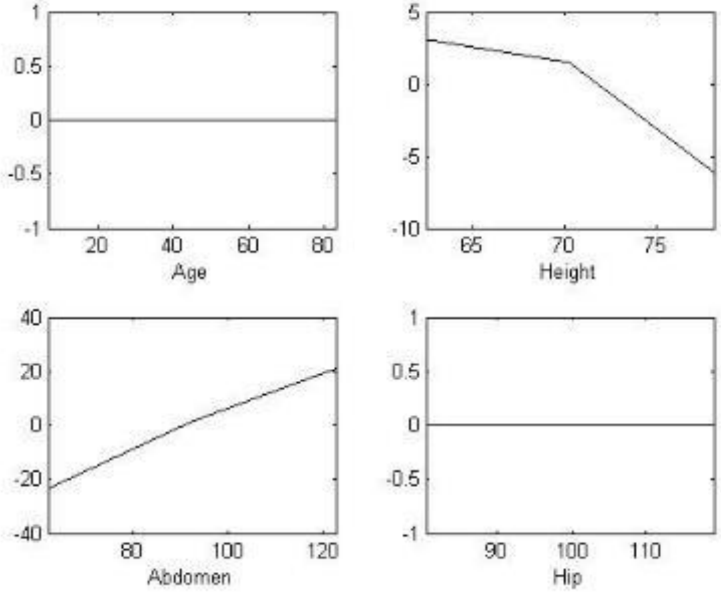
Free but fairly accurate calculator (statistical modeling and prediction)

http://www.csgnetwork.com/bodyfatcalc.html
http://extremebodyfit.com/fat-loss/how-to-calculate-your-body-fat-percentage

Body Fat Data

- Nonparametric function estimation
  - Quantile additive model
  - Response(Y) = Body Fat percentage
  - Covariate(**X**) = (Age, Height, Abdomen, Hip)
  - $(Body\ Fat) = f_1(Age) + f_2(Height) + f_3(Abdomen) + f_4(Hip)$
    $+(measurement\ error)$

quantile(τ)=0.5
(median estimation)



Era of Big Data

- **Production of Massive Data**
  - Over 95% of human-made data have been generated just during last 2 years.
  - 1.8 zettabytes in 2011 ($2x10^{11}$ HD movies that can be watched for 47 million years)
  - Expect 50-fold increase by 2020.
- **Type of Data**
  - Conventional: number
  - New: images, sounds, texts, web logs (Facebook, Twitter, …)

Simpson’s Paradox

- **Berkeley Gender Bias**
  - Lawsuit for gender bias against women for admission to graduate schools in the fall 1973
  - Men were more likely than women to be admitted.

Gender	Total Applicants	Admitted Percentage
Men	8,442	**44 %**
Women	4,321	**35 %**

https://en.wikipedia.org/wiki/Simpson's_paradox

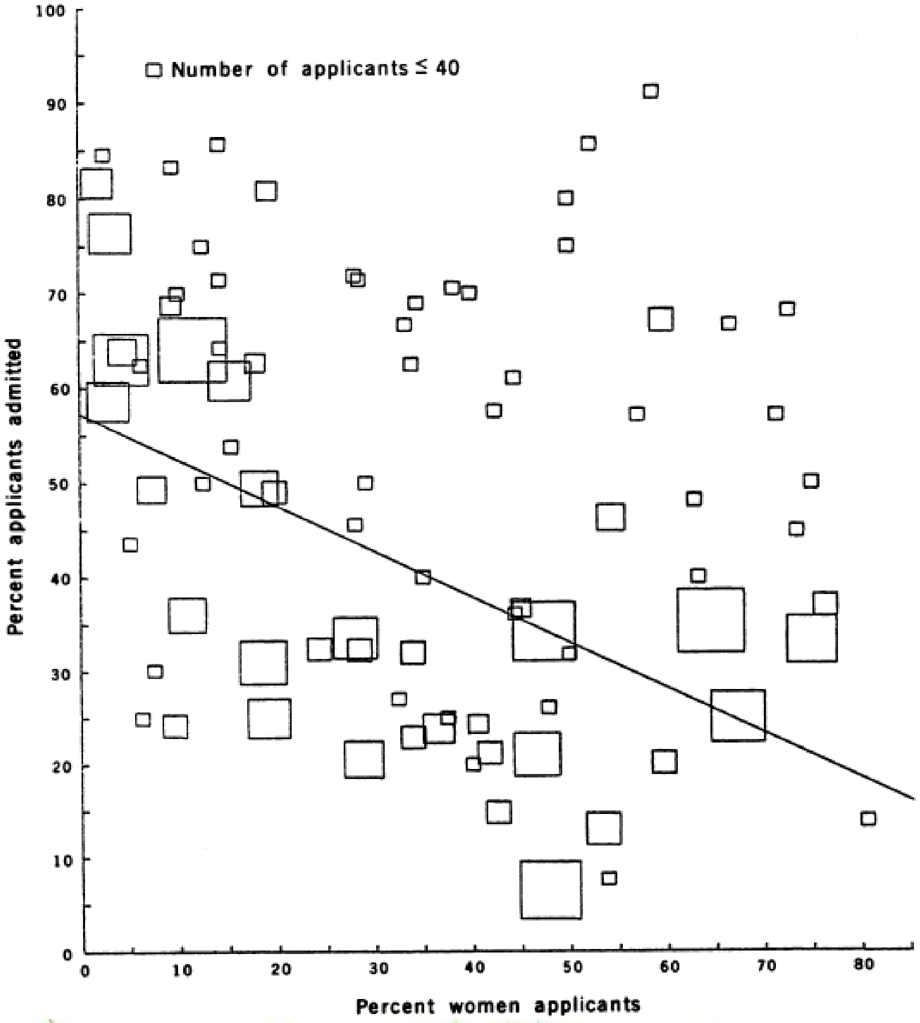
Simpson’s Paradox

– But, disaggregation of the data showed:

Department	Men		Women	
	Applicants	Admitted (%)	Applicants	Admitted (%)
A	825	**62 %**	108	**82 %**
B	560	**63 %**	25	**68 %**
C	325	**37 %**	593	**34 %**
D	417	**33 %**	375	**35 %**
E	191	**28 %**	393	**24 %**
F	272	**6 %**	341	**7 %**

https://en.wikipedia.org/wiki/Simpson's_paradox

Simpson’s Paradox



- **Data Analysis Paper in *Science***
  - Bickel, Hammel and O’Connell (1975), “Sex Bias in Graduate Admissions: Data from Berkeley”, *Science*, Vol. 187, p. 498-404.
- Size of box indicates relative number of applicants to the department.

PageRank

- Since web pages are extremely diverse, the role of search engines are very important.

- Earlier search engines showed web pages related to the query only based on contents in the page.

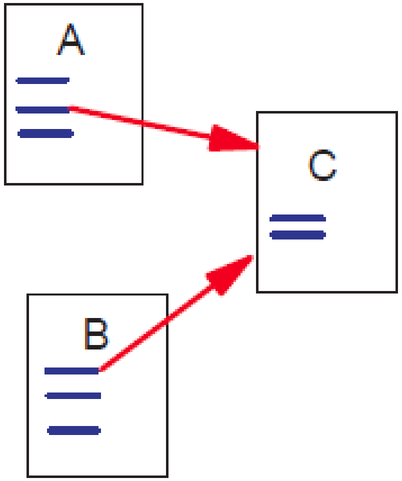
  (e.g.) There are too many web pages containing the word “university”.

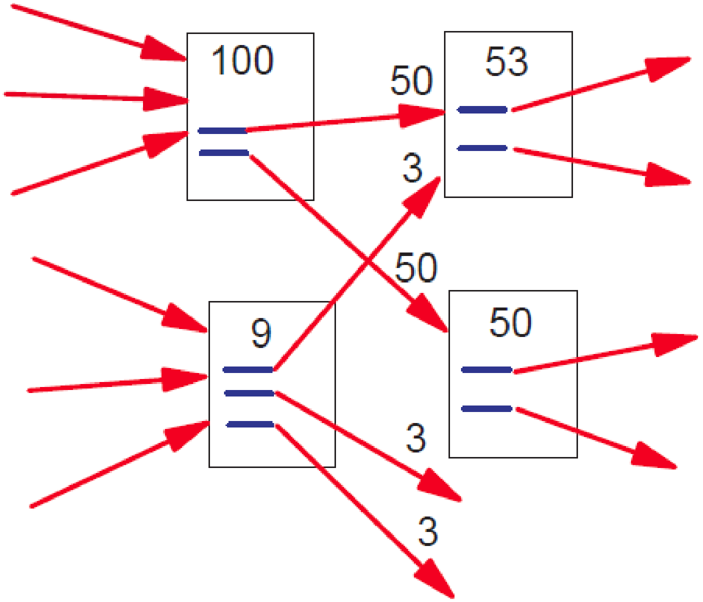
  → Is it reasonable to rank pages just in the order of appearance counts of the word “university”?

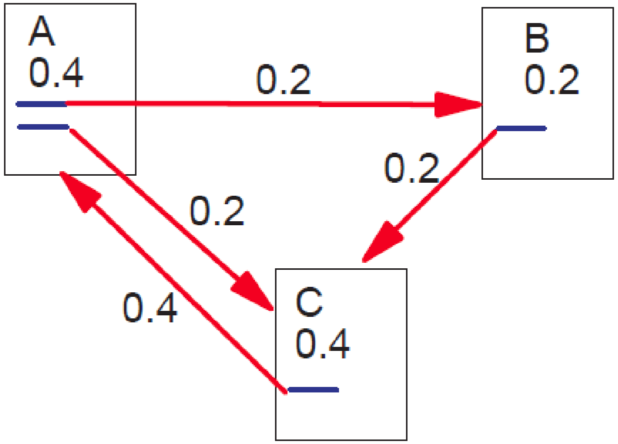
- We need to consider which page is of the most importance.

PageRank

- Measure of Importance
  - Page, et al. (1999) “The PageRank Citation Ranking: Bringing Order to the Web”, *Technical Report*. Stanford InfoLab.
  - Importance based on the webpages linked to it

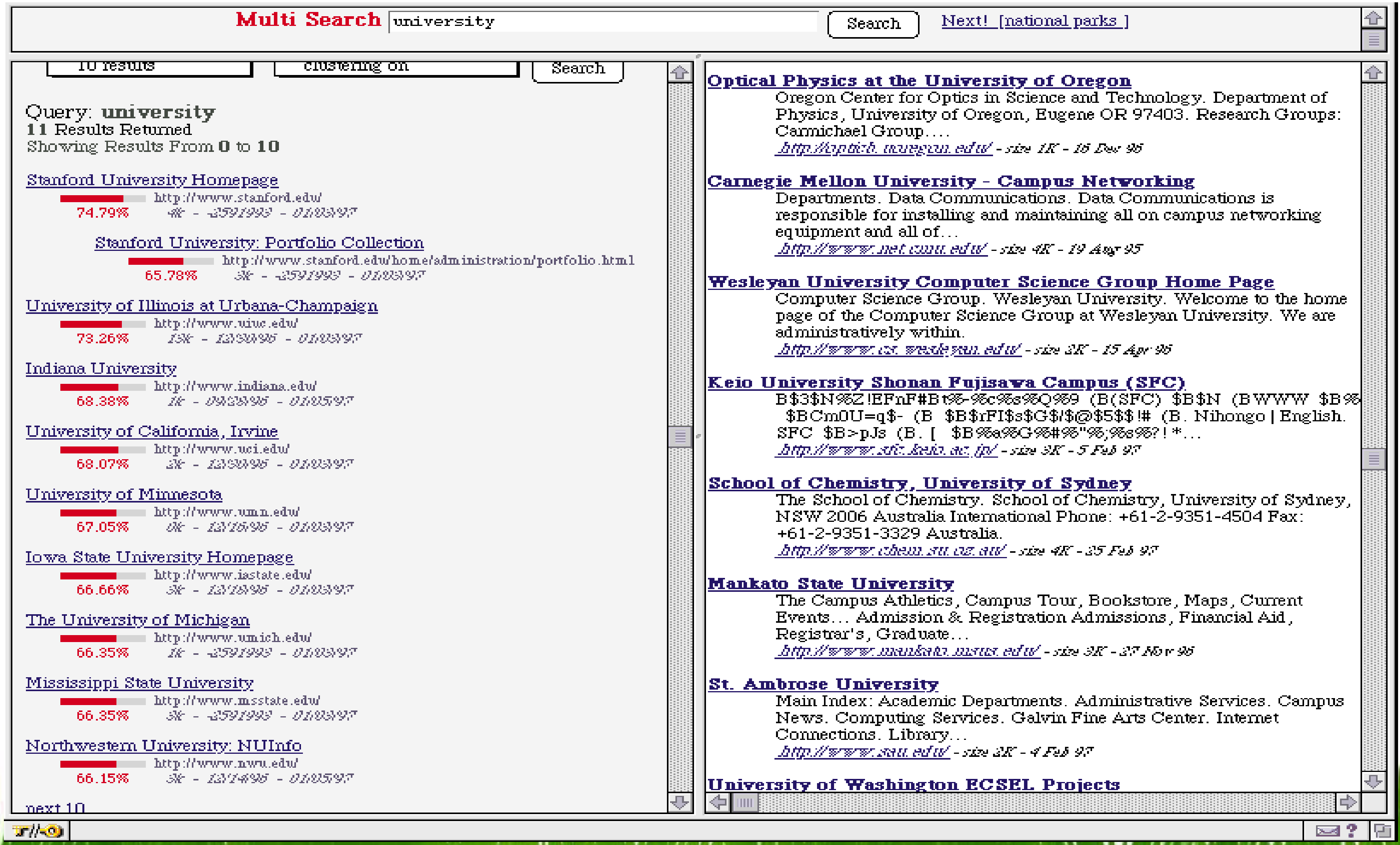






PageRank

- Comparison of query for “university”



Query: **university**
11 Results Returned
Showing Results From **0** to **10**

Stanford University Homepage
74.79% http://www.stanford.edu/
4k - [illegible] - 01/03/97

Stanford University: Portfolio Collection
65.78% http://www.stanford.edu/home/administration/portfolio.html
3k - [illegible] - 01/03/97

University of Illinois at Urbana-Champaign
73.26% http://www.uiuc.edu/
13k - 12/30/96 - 01/03/97

Indiana University
68.38% http://www.indiana.edu/
1k - 09/28/96 - 01/05/97

University of California, Irvine
68.07% http://www.uci.edu/
3k - 12/30/96 - 01/03/97

University of Minnesota
67.05% http://www.umn.edu/
6k - 12/16/96 - 01/03/97

Iowa State University Homepage
66.66% http://www.iastate.edu/
3k - 12/18/96 - 01/03/97

The University of Michigan
66.35% http://www.umich.edu/
1k - [illegible] - 01/03/97

Mississippi State University
66.35% http://www.msstate.edu/
3k - [illegible] - 01/03/97

Northwestern University: NUInfo
66.15% http://www.nwu.edu/
3k - 12/14/96 - 01/05/97

next 10

Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
http://optics.uoregon.edu/ - size 1K - 15 Dec 96

Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
http://www.net.cmu.edu/ - size 4K - 19 Aug 95

Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
http://www.cs.wesleyan.edu/ - size 3K - 15 Apr 96

Keio University Shonan Fujisawa Campus (SFC)
B$3$N%Z!EFnF#Bt%-%c%s%Q%9 (B(SFC) BN (BWWW $B%
$BCm0U=q$- (B BrFIsG$/$@$5$$!# (B. Nihongo | English.
SFC $B>pJs (B. [$B%a%G%#%"%;%s%?!*...
http://www.sfc.keio.ac.jp/ - size 3K - 5 Feb 97

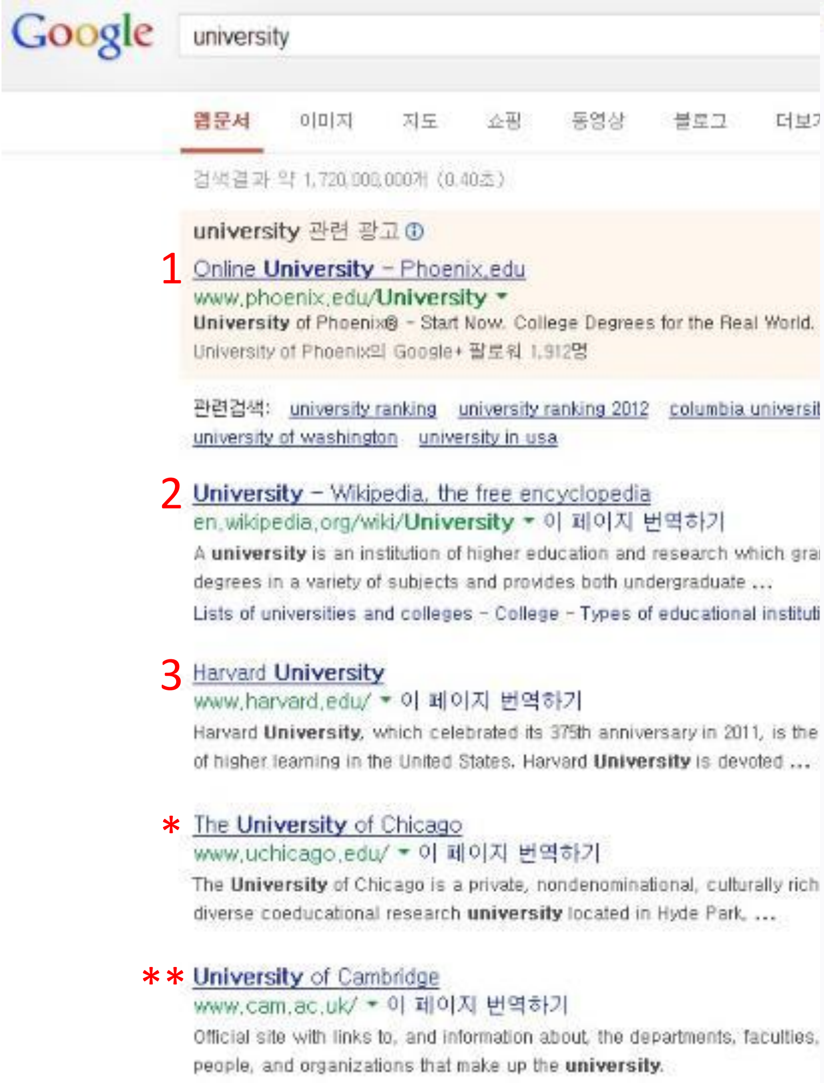
School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
http://www.chem.su.oz.au/ - size 4K - 25 Feb 97

Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
http://www.mankato.msus.edu/ - size 3K - 27 Nov 96

St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
http://www.sau.edu/ - size 3K - 4 Feb 97

University of Washington ECSEL Projects

PageRank



검색결과 약 1,720,000,000개 (0.40초)

university 관련 광고 ⓘ

1 Online University – Phoenix.edu
www.phoenix.edu/University
University of Phoenix® - Start Now. College Degrees for the Real World.
University of Phoenix의 Google+ 팔로워 1,912명

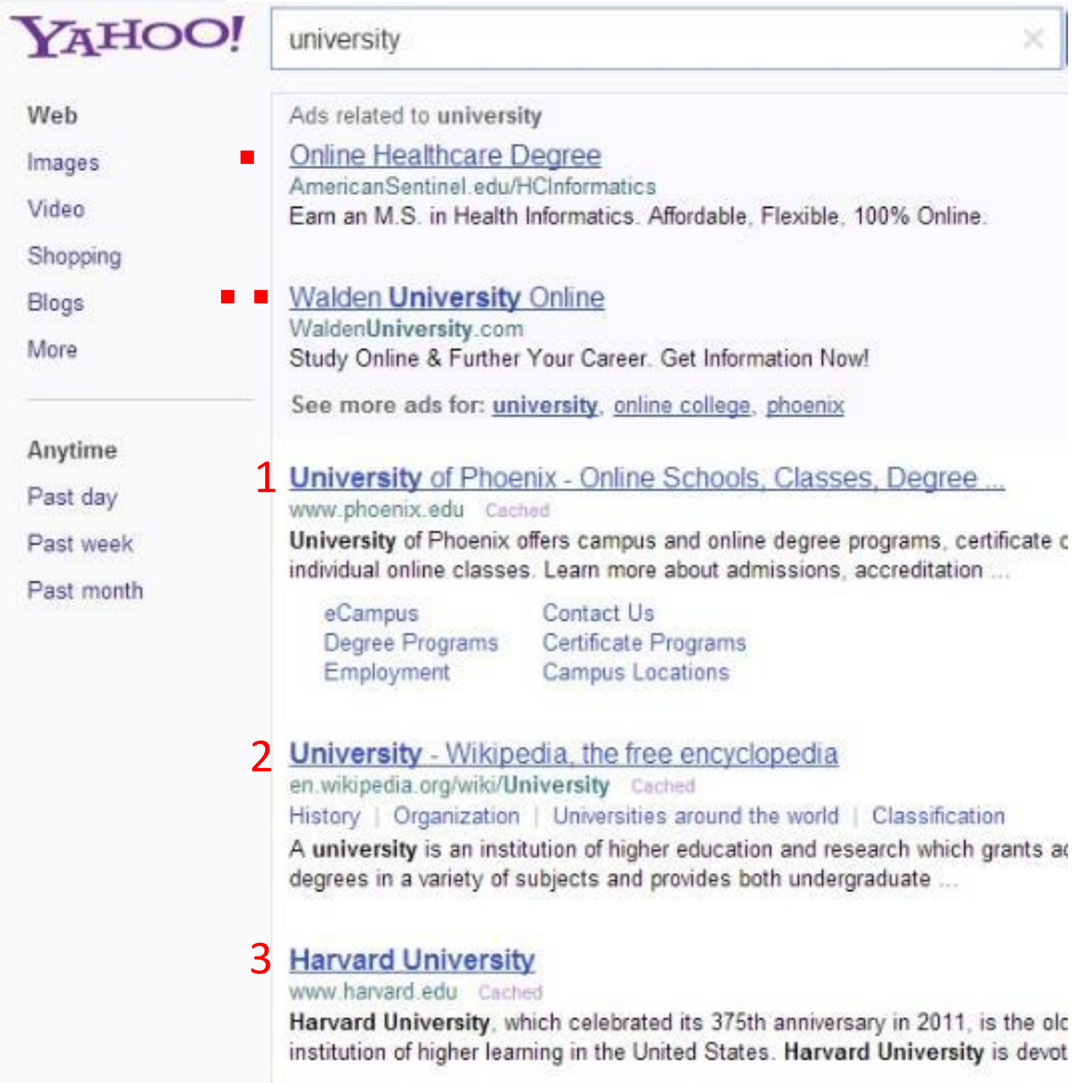
관련검색: university ranking university ranking 2012 columbia universit
university of washington university in usa

2 University – Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/University ▾ 이 페이지 번역하기
A university is an institution of higher education and research which gra
degrees in a variety of subjects and provides both undergraduate ...
Lists of universities and colleges - College - Types of educational instituti

3 Harvard University
www.harvard.edu/ ▾ 이 페이지 번역하기
Harvard University, which celebrated its 375th anniversary in 2011, is the
of higher learning in the United States. Harvard University is devoted ...

* The University of Chicago
www.uchicago.edu/ ▾ 이 페이지 번역하기
The University of Chicago is a private, nondenominational, culturally rich
diverse coeducational research university located in Hyde Park, ...

** University of Cambridge
www.cam.ac.uk/ ▾ 이 페이지 번역하기
Official site with links to, and information about, the departments, faculties,
people, and organizations that make up the university.



Ads related to university

■ Online Healthcare Degree
AmericanSentinel.edu/HCInformatics
Earn an M.S. in Health Informatics. Affordable, Flexible, 100% Online.

■ ■ Walden University Online
WaldenUniversity.com
Study Online & Further Your Career. Get Information Now!

See more ads for: university, online college, phoenix

1 University of Phoenix - Online Schools, Classes, Degree ...
www.phoenix.edu Cached
University of Phoenix offers campus and online degree programs, certificate c
individual online classes. Learn more about admissions, accreditation ...

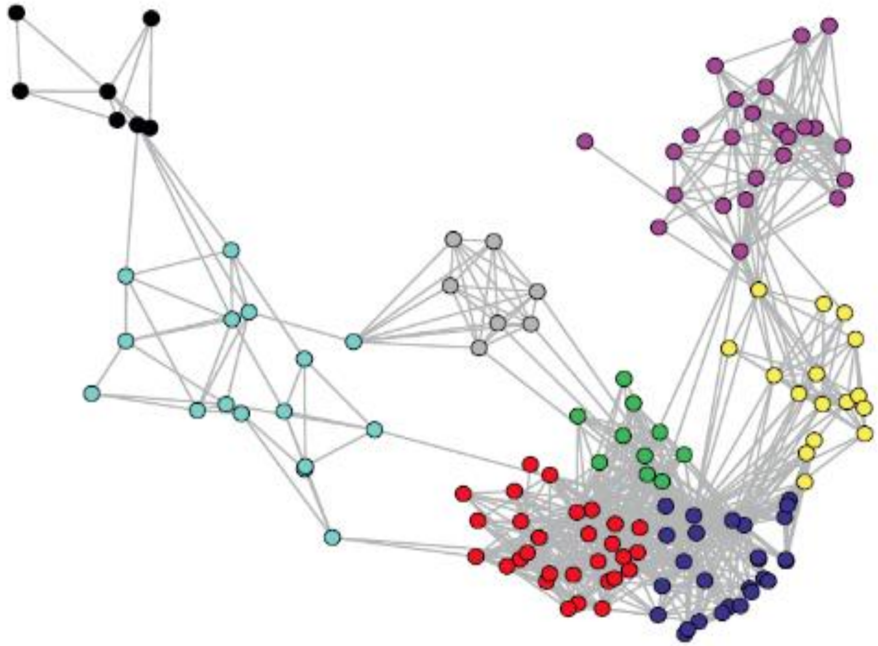
eCampus
Degree Programs
Employment
Contact Us
Certificate Programs
Campus Locations

2 University - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/University Cached
History | Organization | Universities around the world | Classification
A university is an institution of higher education and research which grants ac
degrees in a variety of subjects and provides both undergraduate ...

3 Harvard University
www.harvard.edu Cached
Harvard University, which celebrated its 375th anniversary in 2011, is the olc
institution of higher learning in the United States. Harvard University is devot

Cluster Analysis on Facebook

- Salter-Townshend (2012), “Analysing My Facebook Friends”, *Significance*, p. 40-42.
- A statistical analysis of the link pattern (grey lines) reveals information on the Facebook friends (circles).
- The algorithm picks up 8 groups as different colors.
- In fact Facebook uses closeness measures to make their friend suggestions.



Cf) Yellow(family), Blue(girl friends), Red(dormitory friends), Green(ski club members), …

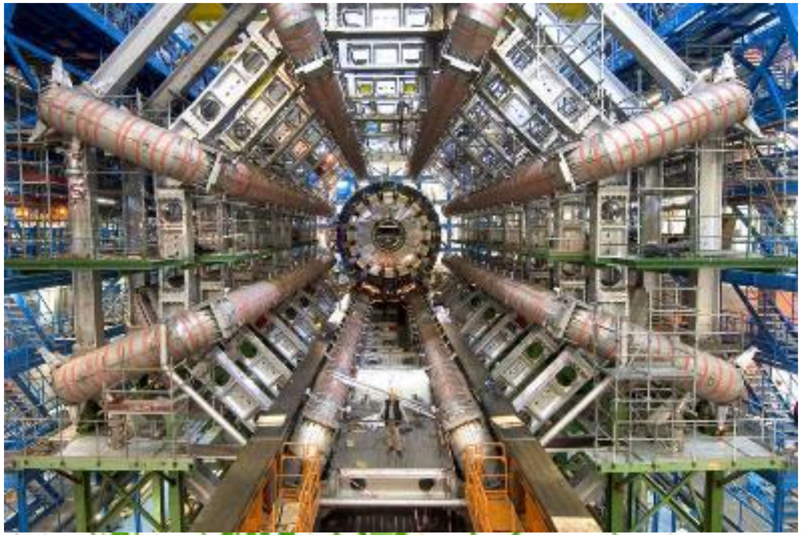
Netflix Prize

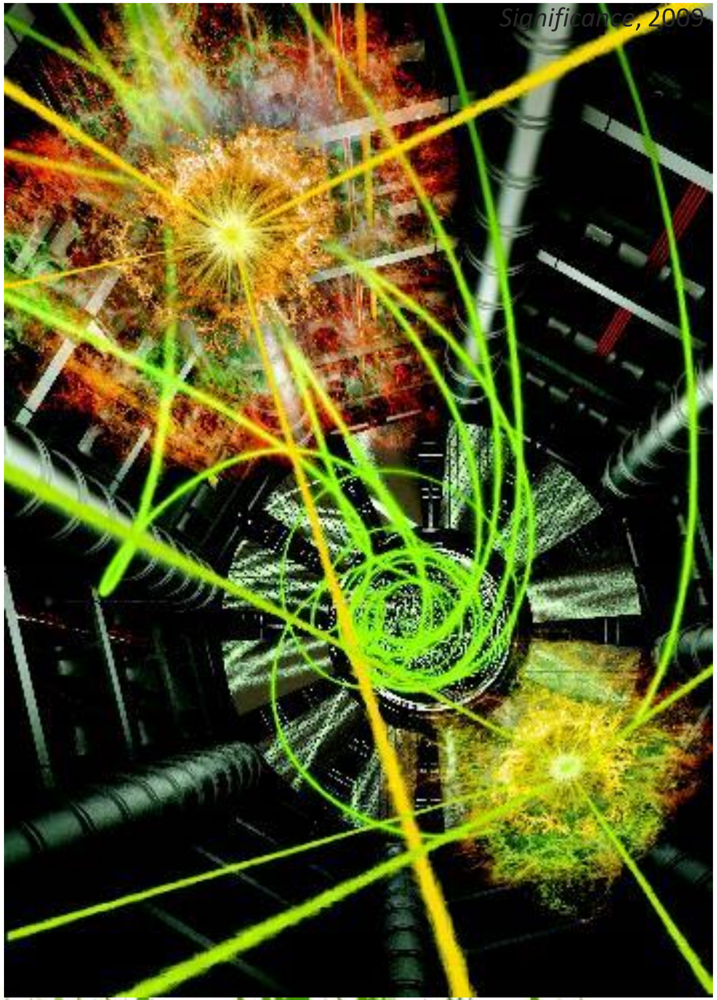
- **Netflix**
  - Online DVD rental service company
  - (data) 100,480,507 ratings that 480,189 users gave to 17,770 movies
- **Objective**
  - Three-year (2007-2009) contest for movie recommendation system
  - Improving prediction algorithm for user ratings based on the **data**
  - At least 10% better performance than *Cinematch,* Netflix's algorithm
- **Competition Result**
  - $ 100,000,000 prize open competition
  - Winning to *BellKor* team, AT&T **statistician** group
  - The winning algorithm is also applicable to other marketing area.

Higgs Boson Data

- Large Hadron Collider (LHC)

"Testing nature to the limit: the Large Hadron Collider",
Significance, 2009





Higgs Boson Data

- **Standard Model in Physics**
  - A theory concerning the electromagnetic, weak and strong nuclear interactions
  - There exists 25 adjustable parameters.
- **Higgs Boson**
  - An unconfirmed piece of the theory
  - Exist theoretically but no evidence
  - Strong *belief* for most physicists

Higgs Boson Data

- **Key Goal of LHC**
  - Establish whether Higgs boson actually exists or not.
  - Measure, if so, properties of Higgs boson.
- **LHC Data**
  - LHC produce close to a billion events per seconds. (≈ 10^{15} bytes)
  - (1 DVD ≈ 5 GB) 200,000 DVDs are needed for a second.
  - But, only a tiny fraction of those are of potential interest.

Statistics in Other Disciplines

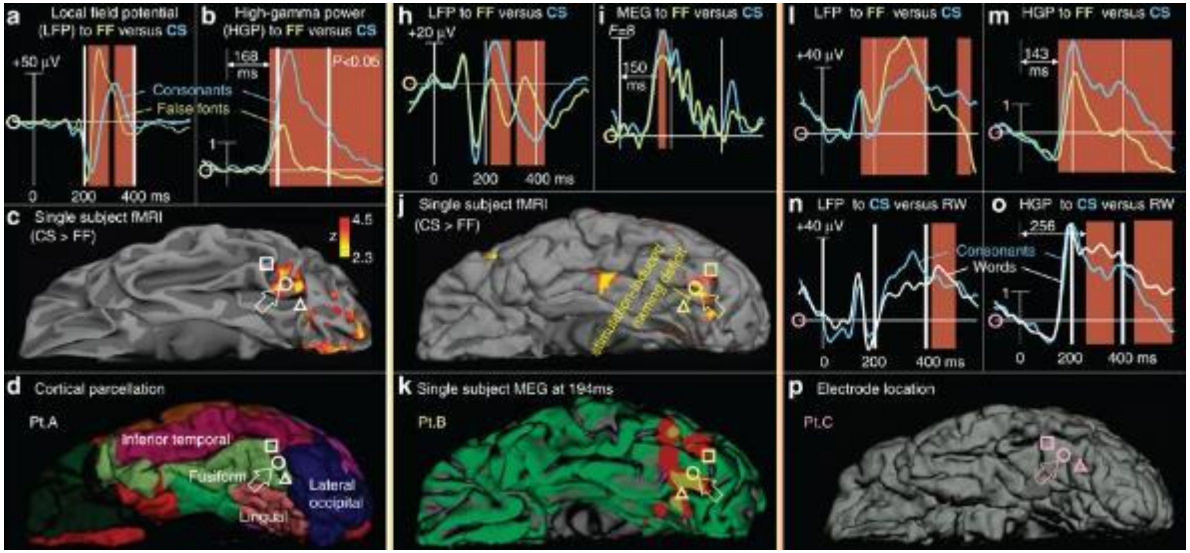
- **Bioinformatics**
  - **Bio**logy + **Informatics**
  - Genome sequence analysis and its management
  - Biologically novel discovery from genomic data

- **Medical Science**
  - Finding optimal therapy based on massive medical records combining health insurance data

Statistics in Other Disciplines

- **Brain Science**
  - Complex understanding on brain function
  - Statistical analysis of signals on brain with high-dimensional images

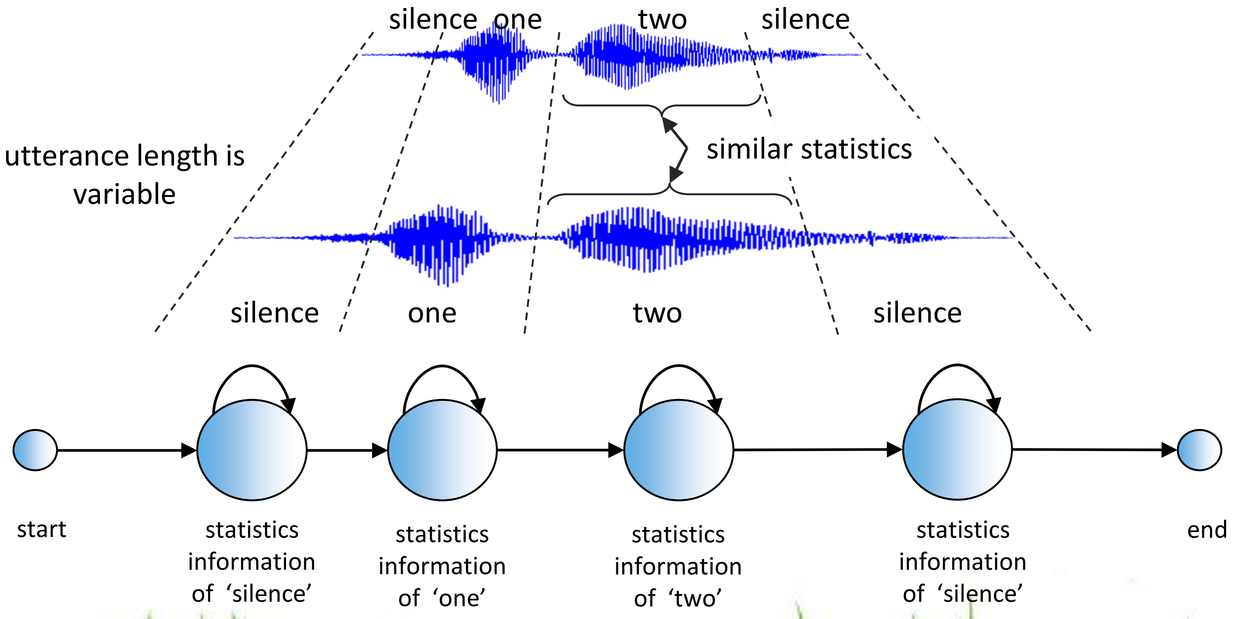
  (e.g.) Functional Magnetic Resonance Image (fMRI) data



Thesen et al. (2012), “Sequential then interactive processing of letters and words in the left fusiform gyrus, *Nature Communications*, Vol. 3.

Statistics in Other Disciplines

- **Signal Processing**
  - Voice/picture recognition, Image denoising algorithm, …
  - Conventional: Physical pattern analysis
  - New: Statistical/Machine learning


silence
one
two
silence
utterance length is variable
similar statistics
silence
one
two
silence
start
statistics information of 'silence'
statistics information of 'one'
statistics information of 'two'
statistics information of 'silence'
end

Statistics in Other Disciplines

- **Earth Science**
  - Meteorology, Climatology, Oceanography, Seismology, …
  - Probabilistic/Statistical models perform more efficient than computationally heavy physical-dynamic models

  (e.g.) Ensemble method

  - 3rd generation → 4th generation science

- **Astronomy**
  - Spatial-temporal distribution analysis for cosmic evolution
  - Measurement error or observation truncation on magnitude of stars

Statistics in Other Disciplines

And MANY others…!

Prospects

- ***For Today's Graduate, Just One Word: Statistics***
  - New York Times, August 5, 2009
  - "People think of field Archaeology as Indiana Jones, but much of what you really do is **data analysis**." – Carrie Grimes, Google statistician
  - "We are rapidly entering a world where everything is monitored and measured but the big problem is going to be **the ability of human to use, analyze and make sense of the data**" – Erik Brynjolfsson, director of MIT center for digital business

- ***Statistics - Dream Job of the Next Decade***
  - "People can make the data tell a story, and everybody has data, but the problem is **how to utilize the data** more effectively." – Hal Varian, Google chief economist

Thank You!

bupark@stats.snu.ac.kr