Web and Data Analysis

Minkoo Seo

June 2015

About

- R user since 2011
- Wrote this book →
- Software Engineer at Google Korea
- http://mkseo.pe.kr/

- These views are mine and mine alone and do not reflect the views of my employer.

Web and Data Analysis

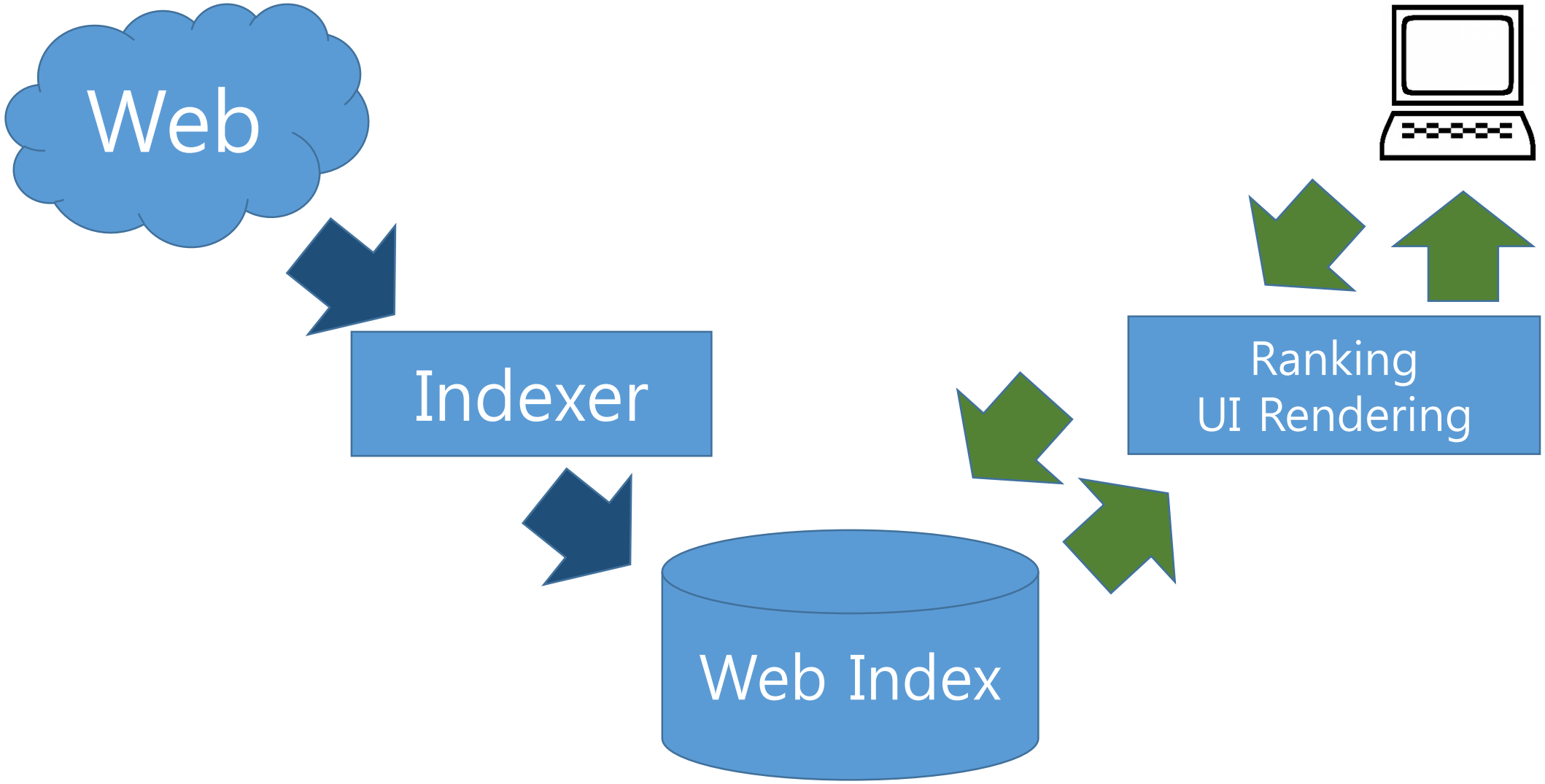
- Big data for web
  - Infrastructure
  - Way of thinking

- Analyzing web site performance
  - A/B Testing
  - Considerations for controlled experiments

Big data for Web

Infrastructure and as a way of thinking

Big data on the web: Search Engine



Inverted Index

Doc 0: It is what it is
Doc 1: What is a banana

Word	Location
a	D1
banana	D1
is	D0, D1
it	D0
what	D0, D1

http://en.wikipedia.org/wiki/Inverted_index

How do we find documents for a query?

Query: [what is a banana]

Word	Location
a	**D1**
banana	**D1**
is	D0, **D1**
it	D0
what	D0, **D1**

Answer: D1

http://en.wikipedia.org/wiki/Inverted_index

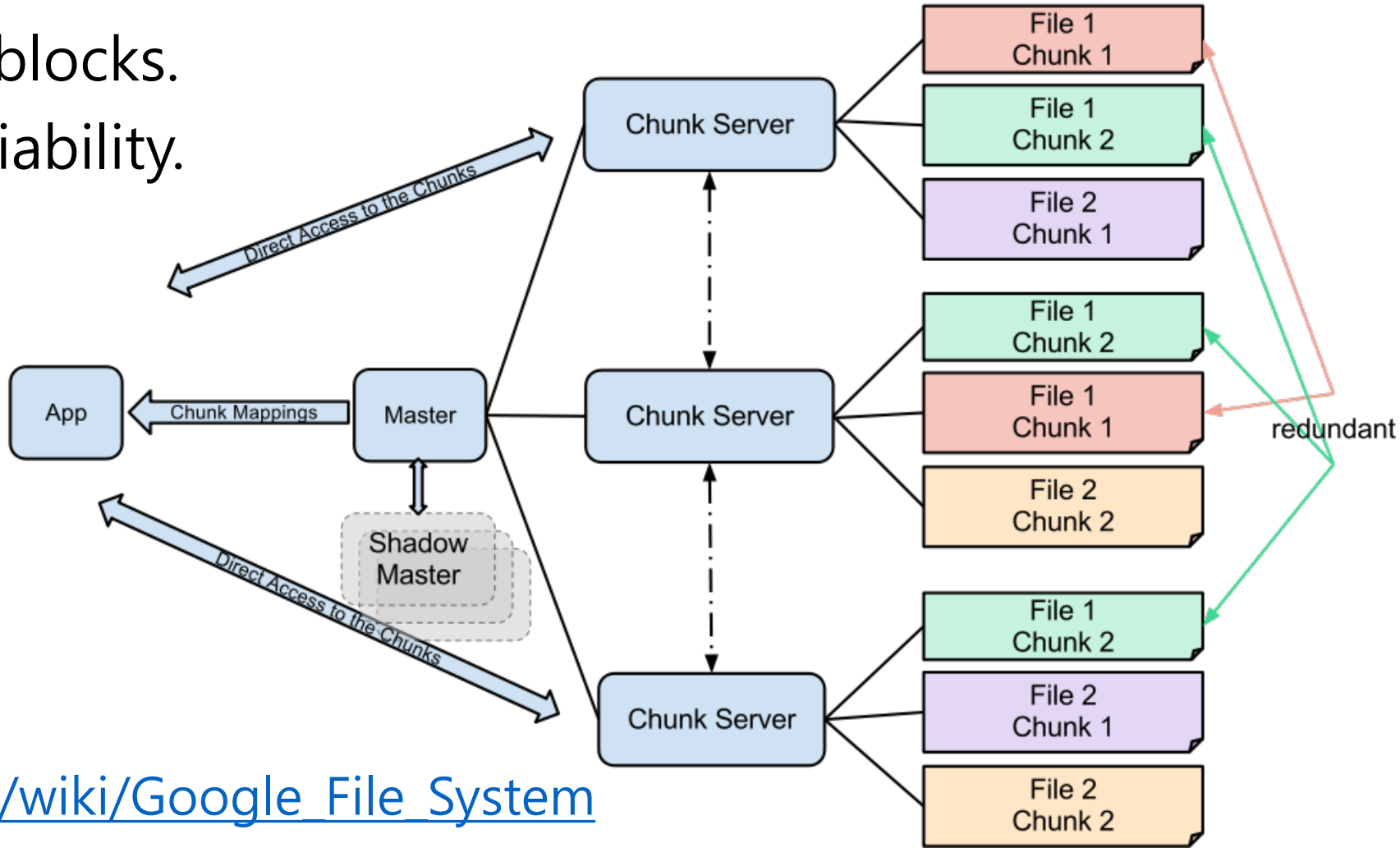
How big is the web?

- +4.67 billion pages on Sunday, 14 June, 2015.
  - http://www.worldwidewebsize.com/

- Need a storage system to store web index.

GFS, HDFS

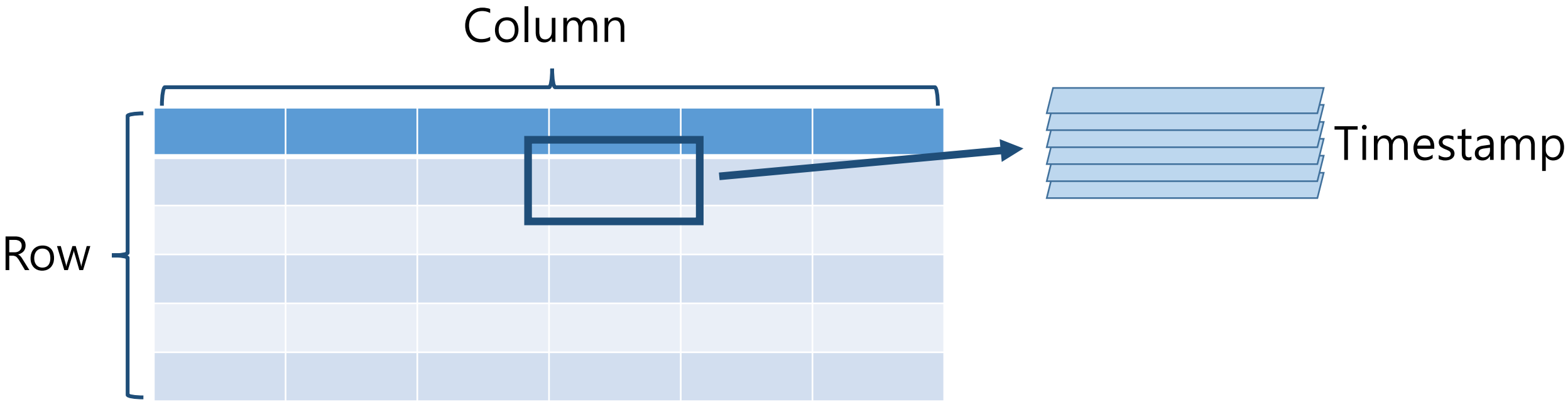
- Master knows data location.
- Data is stored as blocks.
- Replication for reliability.



http://en.wikipedia.org/wiki/Google_File_System

BigTable, HBase

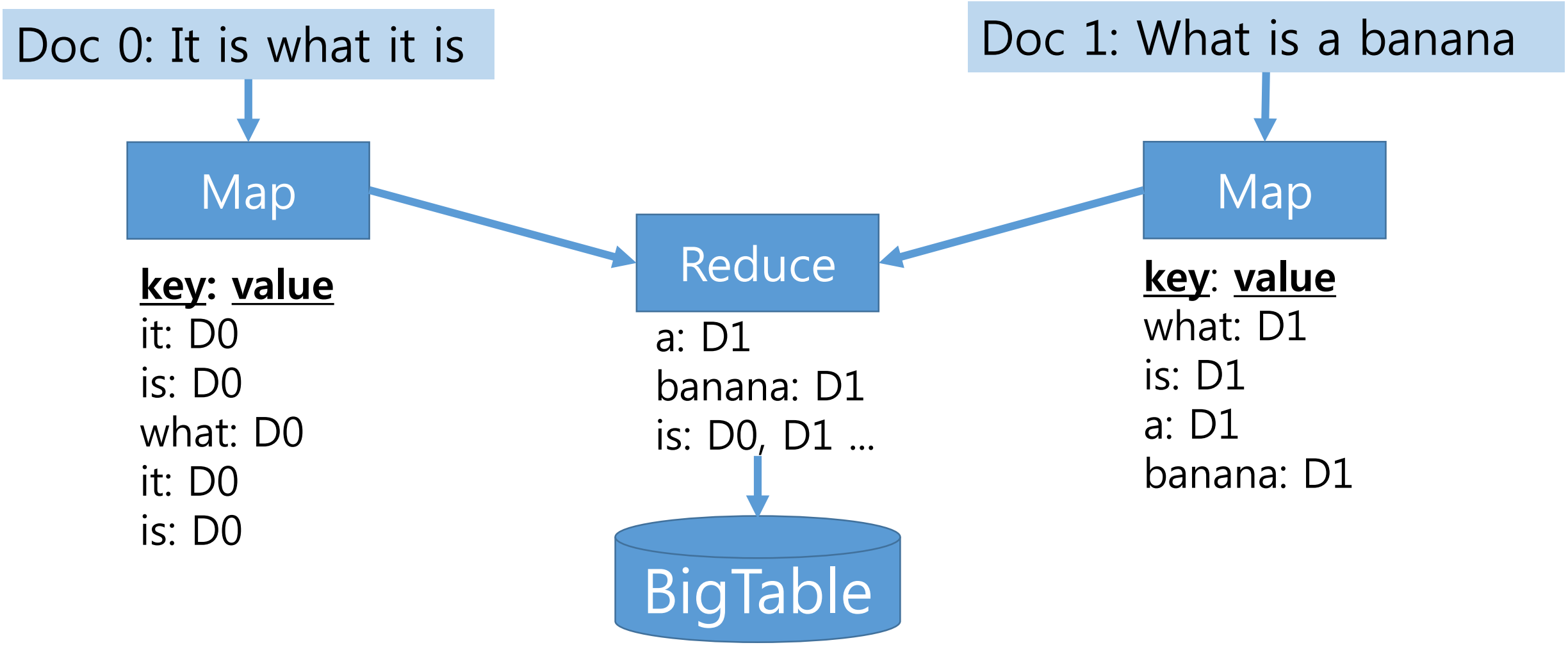
- GFS is great, but that's too primitive.
- BigTable: A table with row key, column key and timestamp.



https://en.wikipedia.org/wiki/BigTable

MapReduce

- Building an index for the web.



Big Data

- So far
  - Big data processing in web scale data.

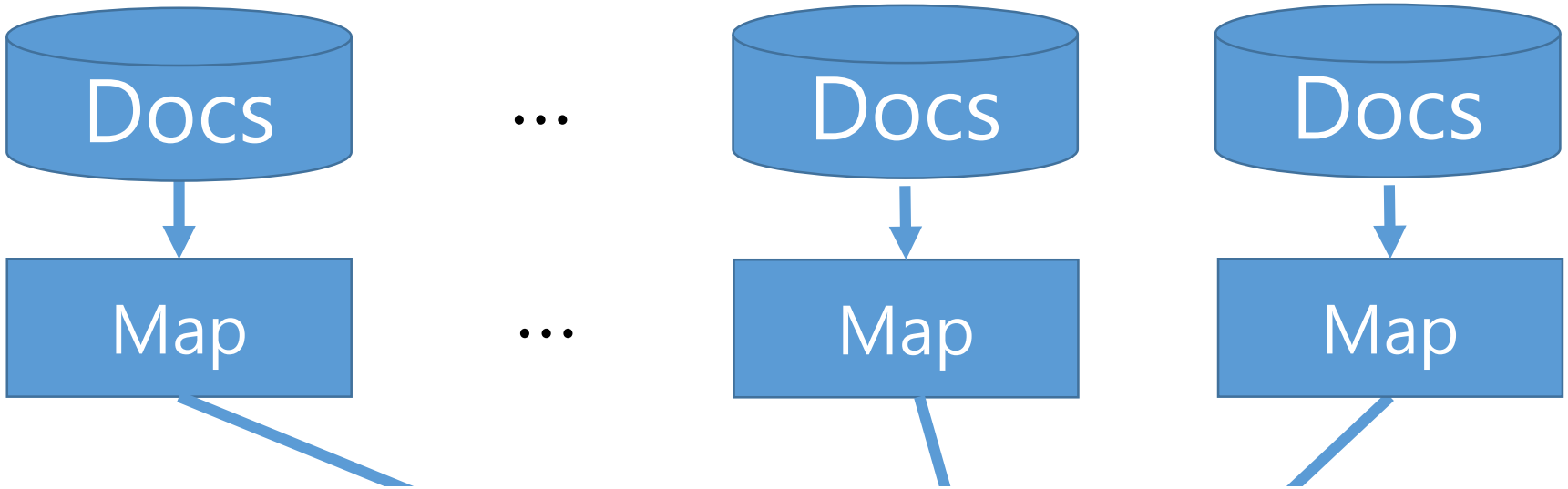
- Next
  - Data size changes the basic.
  - Simple algorithm + Big data.

sample(x, k)

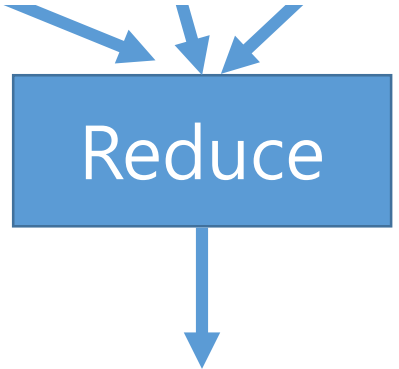
- After sampling, big data is an well known statistical problem.

- But, sampling itself is a difficult problem.

sample(x, k) (cont)



- For each record, get a random number between [0, 1].
- Keep K records with the smallest random number.

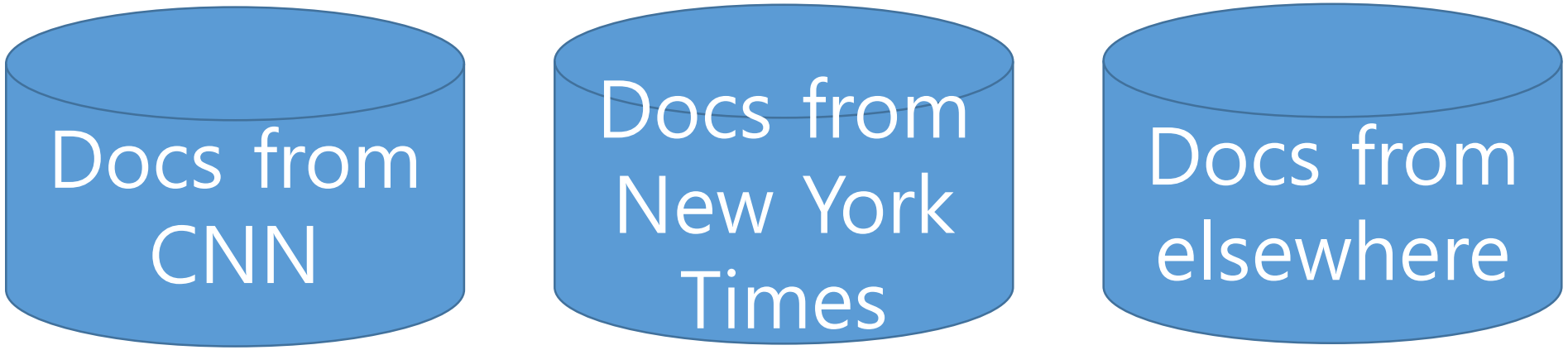


Keep K records with the smallest random number.

http://had00b.blogspot.kr/2013/07/random-subset-in-mapreduce.html

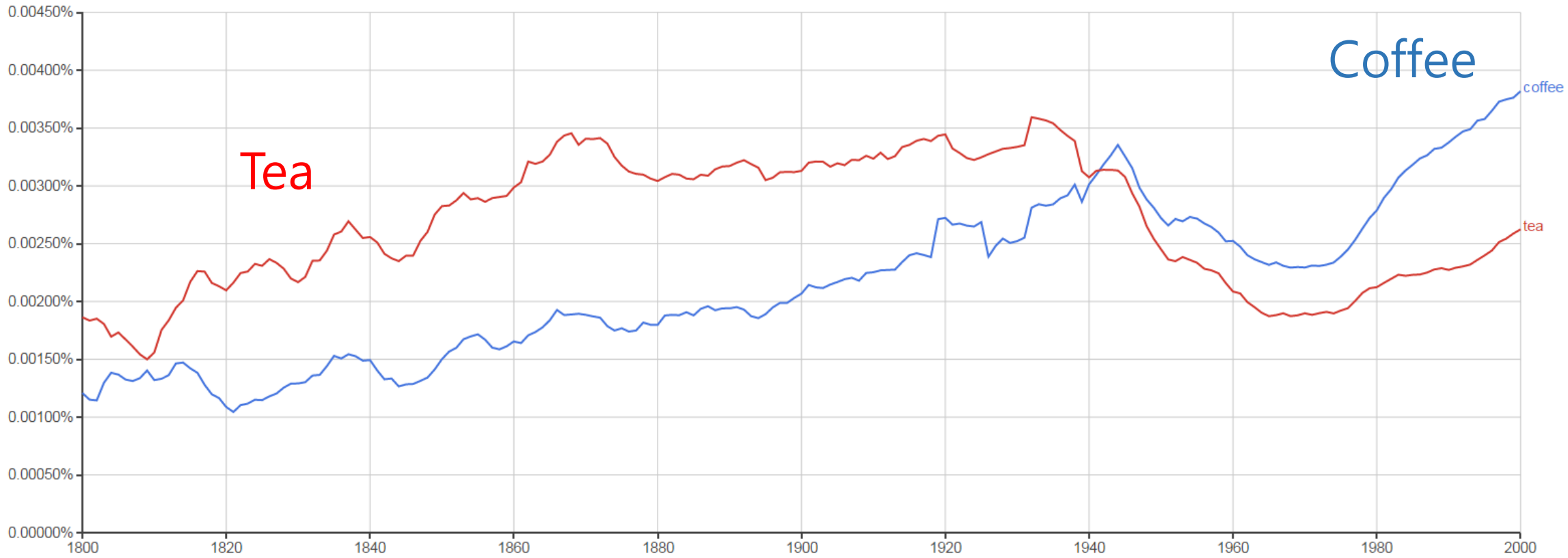
sample(x, k) (cont)

- Need a function to randomly distribute documents.



Simple algorithm + Big Data

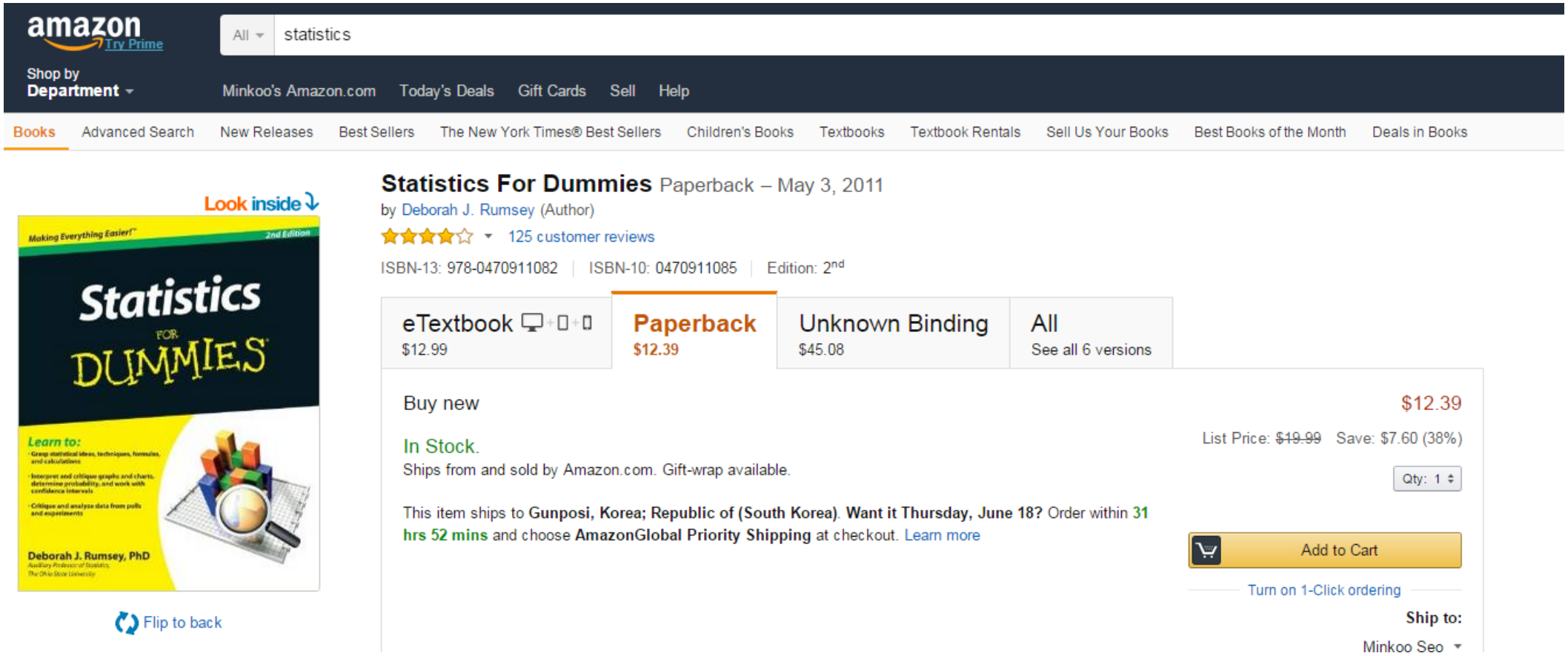
- Counting N-gram (i.e., N consecutive words)



http://books.google.com/ngrams

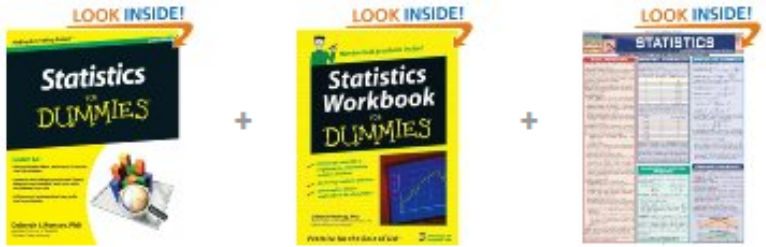
Simple algorithm + Big Data (cont)

- Collaborative Filtering: collecting preference from many users.



http://en.wikipedia.org/wiki/Collaborative_filtering

Frequently Bought Together



Price for all three: $31.45

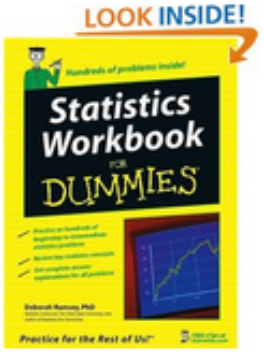
Add all three to Cart

Add all three to Wish List

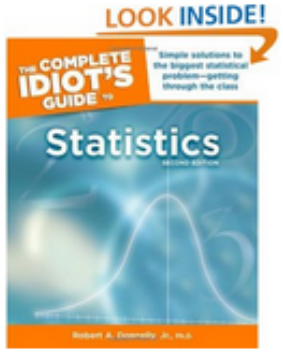
Show availability and shipping details

- **This item:** Statistics For Dummies by Deborah J. Rumsey Paperback $12.39
- Statistics Workbook For Dummies by Deborah J. Rumsey Paperback $13.83
- Statistics Laminate Reference Chart: Parameters, Variables, Intervals, Proportions (Quickstudy: Academic ... by Inc. BarCharts Pamphlet $5.23

Customers Who Bought This Item Also Bought

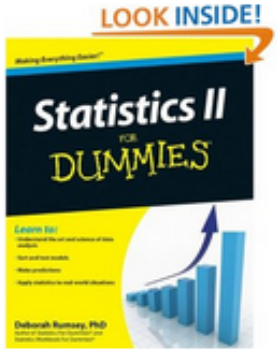


Statistics Workbook...
› Deborah J. Rumsey
47
Paperback
$13.83 Prime
Get it by **Tuesday**

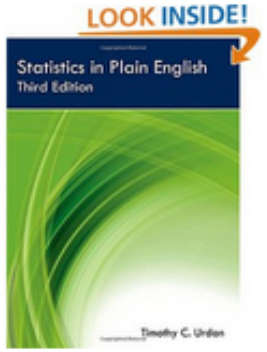


The Complete Idiot's...
Robert A. Donnelly Jr....
93
Paperback
$15.65 Prime
Get it by **Tuesday**

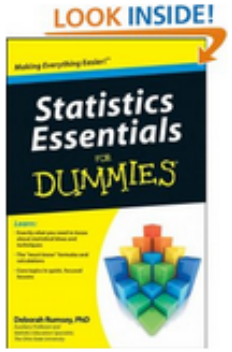
Statistics Laminate...
Inc. BarCharts
112
Pamphlet
$5.23 Prime
Get it by **Tuesday**



Statistics II for...
› Deborah J. Rumsey
32
Paperback
$15.92 Prime
Get it by **Tuesday**



Statistics in Plain...
› Timothy C. Urdan
70
Paperback
$35.71 Prime
Get it by **Tuesday**



Statistics Essentials
› Deborah J. Rumsey
15
Paperback
$9.18 Prime
Get it by **Tuesday**

Simple Algorithm + Big Data (cont)

- Query logs and web documents are valuable resources.
  - Query recommendation: from query sequence [qi09]
  - Learning new entities: rows of tables on the web [que13]

Museum

Museum

Name	Country
Musee de Louvre	France
Metropolitan Museum of Art	USA
National Museum of Korea	Korea

???

- For R users
  - Getting Data
  - Parsing Data
  - Building models
  - **Environment to run analysis w/o written permission, planning, etc.**

Analyzing web site performance

A/B Testing and Consideration

Measuring User Experience

- HEART [ker10]
  - Happiness: satisfaction, likelihood to recommend, visual appeal
  - Engagement: frequency, intensity, depth of interaction
  - Adoption and Retention: new users, revisit
  - Task success: time to complete, percent of completed

- Pageviews, # of unique visitors, # of visits, clicks, earning, etc.

A/B Testing Example

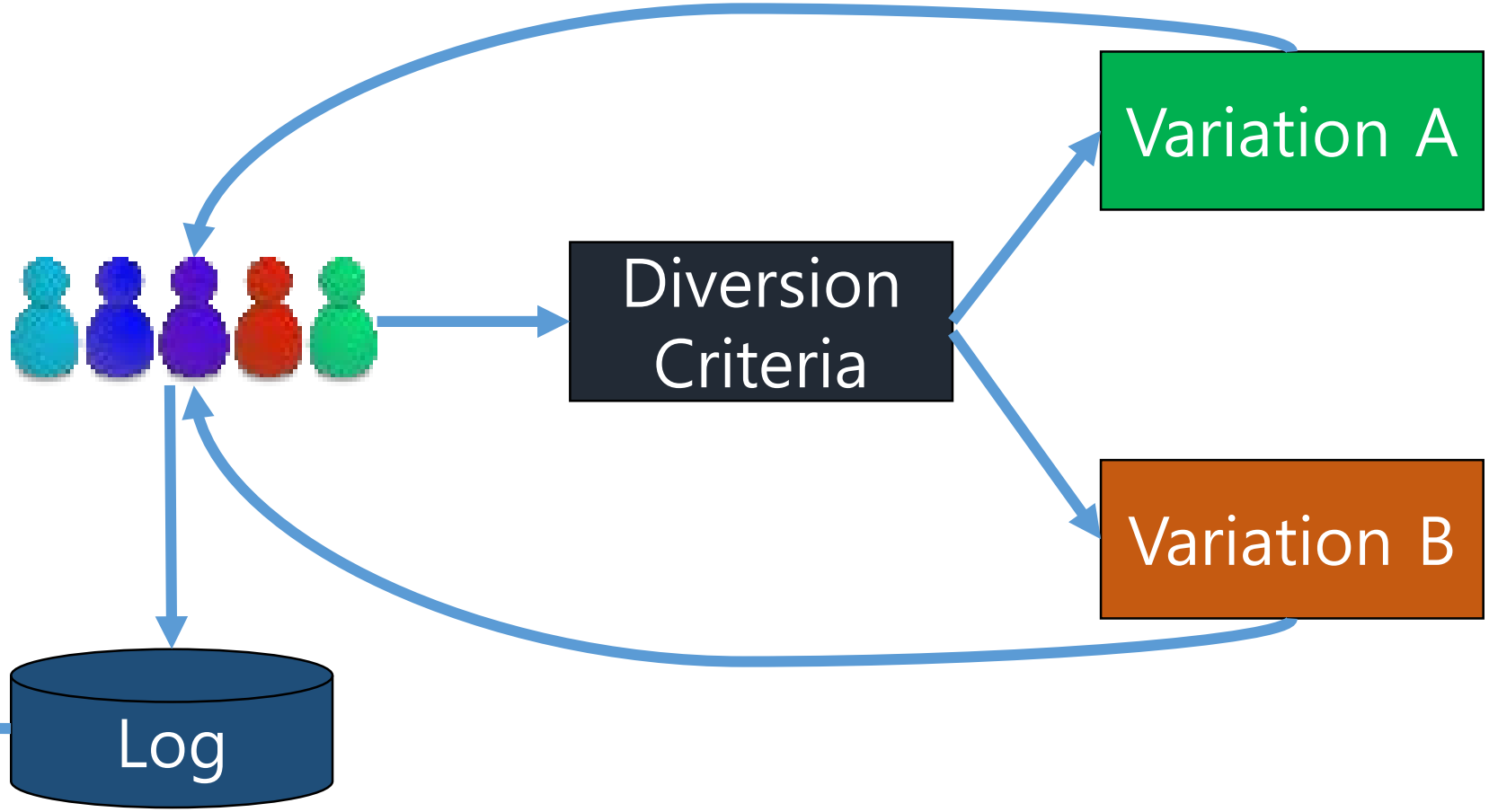
Media
- IMAGE
- VIDEO

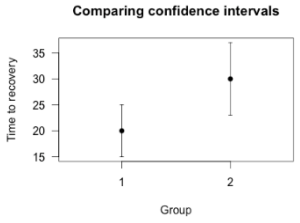
Button:
- JOIN US NOW
- LEARN MORE
- SIGN UP NOW

Result:
- 8.26% -> 11.6%
- 2,880,000 more signup.

https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

A/B Testing


Variation A
Diversion Criteria
Variation B
Log


Comparing confidence intervals
Time to recovery
35
30
25
20
15
1
2
Group

% of total experiment visits : 100.00%

Explorer

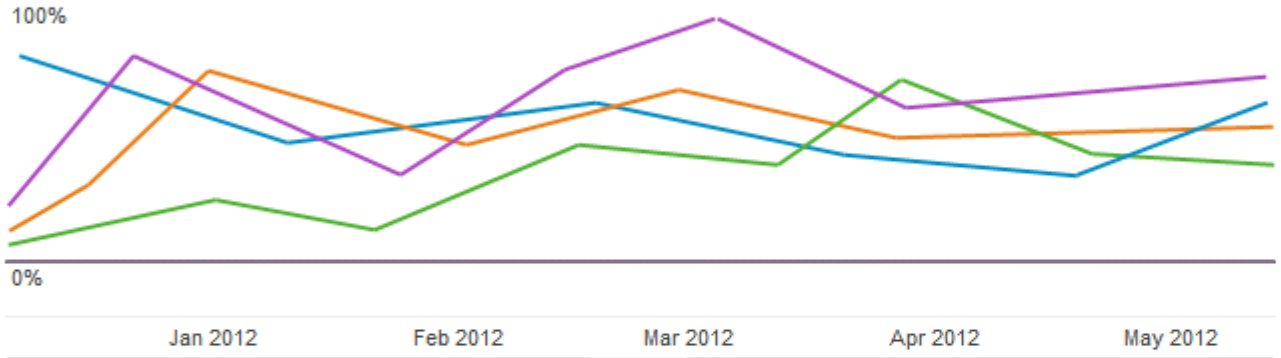
Conversions | Site Usage | Goal Set 1 | Goal Set 2 | Goal Set 3 | Goal Set 4 | Ecommerce

Experiment running - no winner yet

Conversion Rate vs. Select a metric

Day | Week | Month

Original | Variation 1 | Variation 2 | Variation 3



331 visits

161 days of data

100% visitors included

Status:

No winner yet - Experiment still running

Primary Dimension: Variation

Plot Rows

Variation	Experiment Visits	Conversions	Conversion Rate	Compare to Original	Probability of Outperforming Original
Original	64	34	52.31%	0%	0.00%
Variation 2	70	39	55.71%	7%	66.11%
Variation 3	96	45	46.88%	-10%	25.14%
Variation 1	101	3	2.97%	-94%	0.00%

https://support.google.com/analytics/answer/1745152?hl=en&ref_topic=1745207

Running Experiments

- Experiment is not cheap.
  - Each variation should be high quality as experiment == real product.
  - Consider different approach. e.g., user survey.

- Design
  - Hypothesis
    Information browsing -> Pageview increase
    Better answer -> Pageview decrease
  - Logging
  - Sample Size

Running Experiments (cont)

- Before starting an experiment, test it.
  - Bing had a bug that resulted in poor search quality. Poor quality increased # of clicks and revenue [ron12].
  - Test logging [cor09].

- Before and after experiment [dia10].
  - Pre-Period or A/A testing:
    Comparable traffic in control and experiment.
  - Post-period:
    Learned effect from experiment.

Running Experiments (cont)

- Understanding the significance [mar14]
  - Stopping experiments early as soon as one sees significant.
  - Significant but the difference is too small.

- Bots [cor09], e.g., crawler.

- Simpson's Paradox [cor09]
  - Sampling is not uniform, and some browsers are sampled at higher rate.
  - Most of browsers performed worse in treatment, but overall treatment looks better.

Simpson's Paradox

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

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%27s_paradox

Presenting the Result

- Report
  - Experiment design.
  - Confidence Interval, Visualization.
  - Custom log analysis: # of clicks on this after a click on that.

- Experiment Council [dia10]
  - Experiment set up: diversion criteria, triggering, analysis, sizing and duration.
  - Interpreting data: validity of result, completeness of metrics, discussion if the result is positive or negative.

- For R users
  - Understanding infra, analyzing logs, deriving metrics, figuring out confidence interval, and presenting the results.

Summary

- Big data for web
  - Infrastructure
  - Way of thinking

- Analyzing web site performance
  - A/B Testing
  - Considerations for controlled experiments

Reference

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/

- http://www.worldwidewebsize.com/

- http://googleblog.blogspot.kr/2008/07/we-knew-web-was-big.html

- How to pick random (small) data samples using Map/Reduce? http://stackoverflow.com/questions/2514061/how-to-pick-random-small-data-samples-using-map-reduce

- Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber, Bigtable: A Distributed Storage System for Structured Data, OSDI, 2006.
- http://en.wikipedia.org/wiki/Collaborative_filtering
- http://had00b.blogspot.kr/2013/07/random-subset-in-mapreduce.html
- http://books.google.com/ngrams

- [qi09] Qi He, et. al, “Web Query Recommendation via Sequential Query Prediction”, ICDE, 2009.

- [que13] Quercini, Gianluca, and Chantal Reynaud. "Entity discovery and annotation in tables." *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 2013.

- [ker10] Kerry Rodden, Hilary Hutchinson, Xin Fu, Measuring the User Experience on a Large Scale: User-Centered Metrics for Web Applications, CHI, 2010.

- [dia10] Diane Tang, Ashish Agarwal, Deirdre O'Brien, Mike Meyer, “Overlapping Experiment Infrastructure: More, Better, Faster Experimentation”, Conference on Knowledge Discovery and Data Mining, ACM, 2010.

- [ron12] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, Ya Xu, “Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained”, KDD 2012.

- [mar14] Martin Goodson, “Most Winning A/B Test Results are Illusory”, Qubit, 2014. http://www.qubit.com/research/most-winning-ab-test-results-are-illusory

- [cor09] T. Corrk, Brian Frasca, R. Kohavi, R. Longbotham, “Seven Pitfalls when Running Controlled Experiments on the Web”, KDD, 2009.