XONA PARTNERS

Data Science as a Service: The Present and the Future

James G. Shanahan Ph.D.
Xona Partners

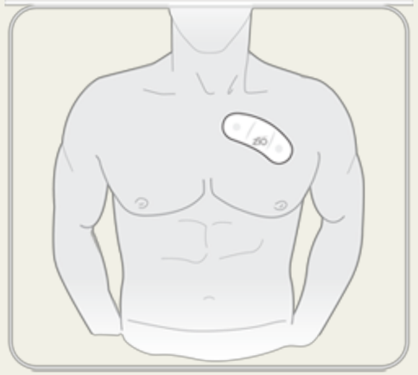
February 2013

Talk Outline

- Big Data Case Studies
- Data Science
- Data Science as a Service (DSaaS)
- Data Science Bubble
- Conclusions

Sensors + Services → Big Data

- **Personal devices (with GPS' and accelerometers)**
  - Earphones; Nike+ (measures and records the distance and pace of a walk or run); asthma inhaler with built-in GPS tracking; Zio Patch that helps doctors detect cardiac problems before they become fatal



- **Personal/social services**
  - Mint, Twitter, diets, health, exercise, FaceBook

- **(These data streams create a huge privacy problem)**

3rdi Art Project

- **A New York University arts, Professor Bilal**
- **A surgically-implanted camera (12/15/2010)**

  3rdi Project, has already generated international media attention and anticipation. On Dec. 15 images from the "third eye" in the back of Bilal's head -- a surgically-impanted camera -- will be unveiled in Doha, Qatar as part of the Told/Untold/Retold exhibition that inaugurates the new Arab Museum of Modern Art near Education City, Doha's intellectual hub.
- **Transmits one image per minute to a website (www.3rdi.me), displayed a Doha gallery**

  with the inaugural images to be displayed in a custom-designed room in the Doha gallery. Bilal's piece will be part of the museum's new permanent collection, 20 years in the making, including more than 6,000 works by Arab artists from North Africa to the Gulf, from the 1920s to the present day.

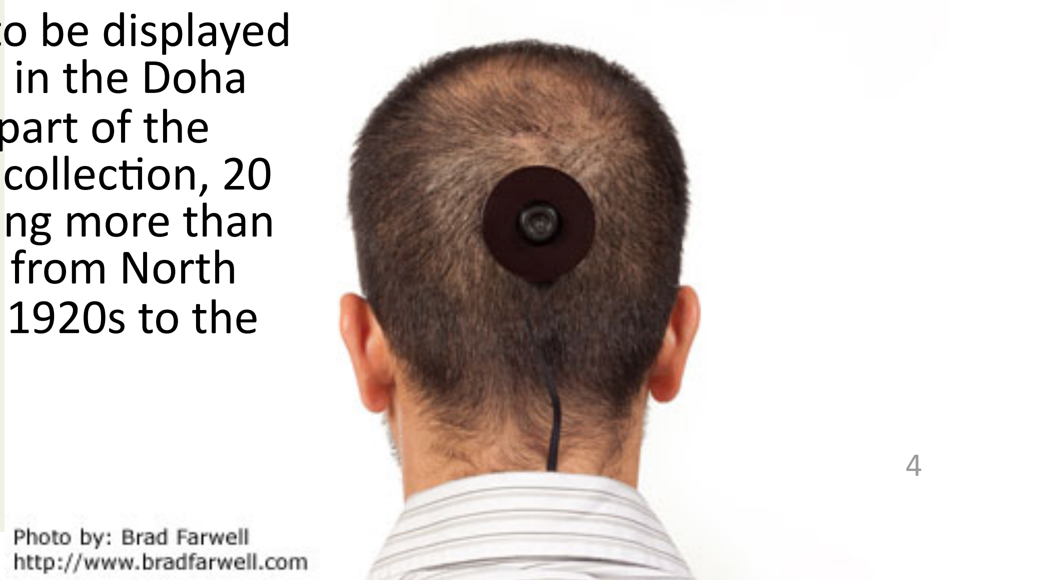
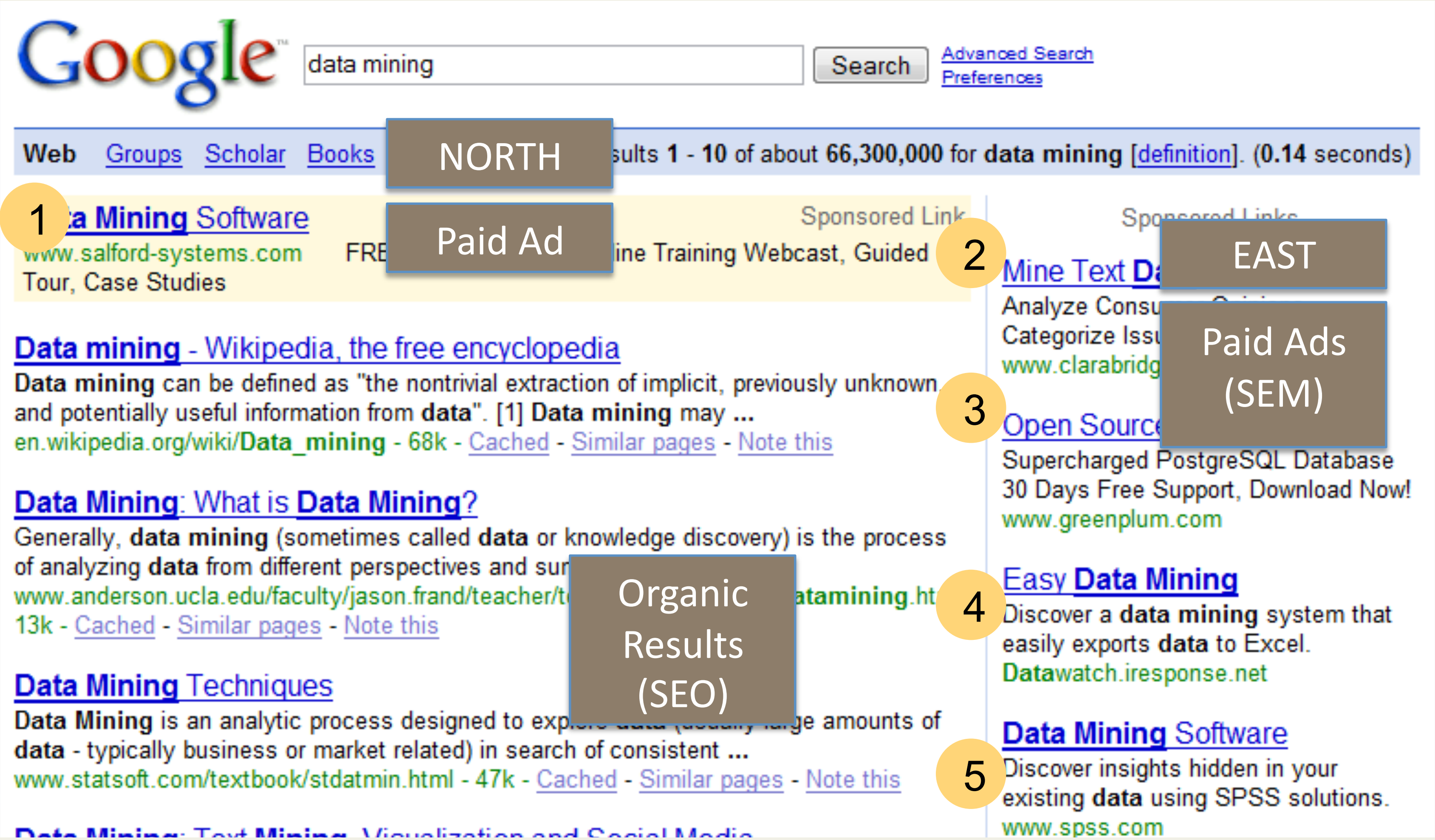


Photo by: Brad Farwell
http://www.bradfarwell.com

Case Studies

- **Travel: Improved ETAs**
  - Plane to gate ETA
  - Traffic
- **Healthcare:**
  - Prostrate cancer
  - Germ Tracker
- **Websearch**
- **Online Advertising:**
  - Personalized promotions
  - RTB for display advertising
- **Politics**

Web Search & Sponsored Search



data mining

Search

Advanced Search
Preferences

Web Groups Scholar Books sults 1 - 10 of about 66,300,000 for data mining [definition]. (0.14 seconds)

ta Mining Software
www.salford-systems.com FRE ine Training Webcast, Guided Tour, Case Studies
Sponsored Link

Data mining - Wikipedia, the free encyclopedia
Data mining can be defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data". [1] Data mining may ...
en.wikipedia.org/wiki/Data_mining - 68k - Cached - Similar pages - Note this

Data Mining: What is Data Mining?
Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and sur
www.anderson.ucla.edu/faculty/jason.frand/teacher/t atamining.ht
13k - Cached - Similar pages - Note this

Data Mining Techniques
Data Mining is an analytic process designed to exp ge amounts of data - typically business or market related) in search of consistent ...
www.statsoft.com/textbook/stdatmin.html - 47k - Cached - Similar pages - Note this

Data Mining: Text Mining, Visualization and Social Media

Sponsored Links

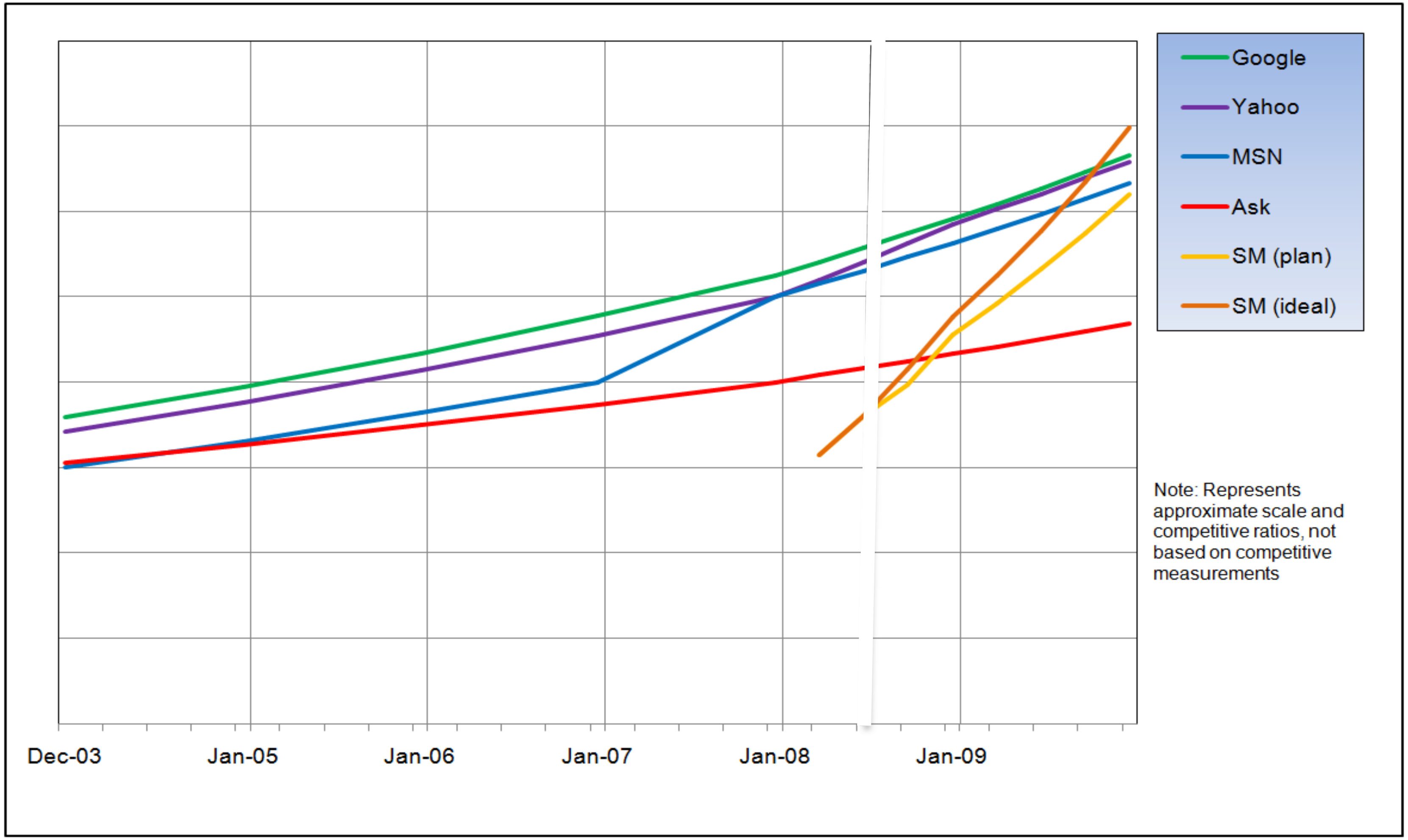
Mine Text Da
Analyze Consu
Categorize Issu
www.clarabridg

Open Sourc
Supercharged PostgreSQL Database
30 Days Free Support, Download Now!
www.greenplum.com

Easy Data Mining
Discover a data mining system that easily exports data to Excel.
Datawatch.iresponse.net

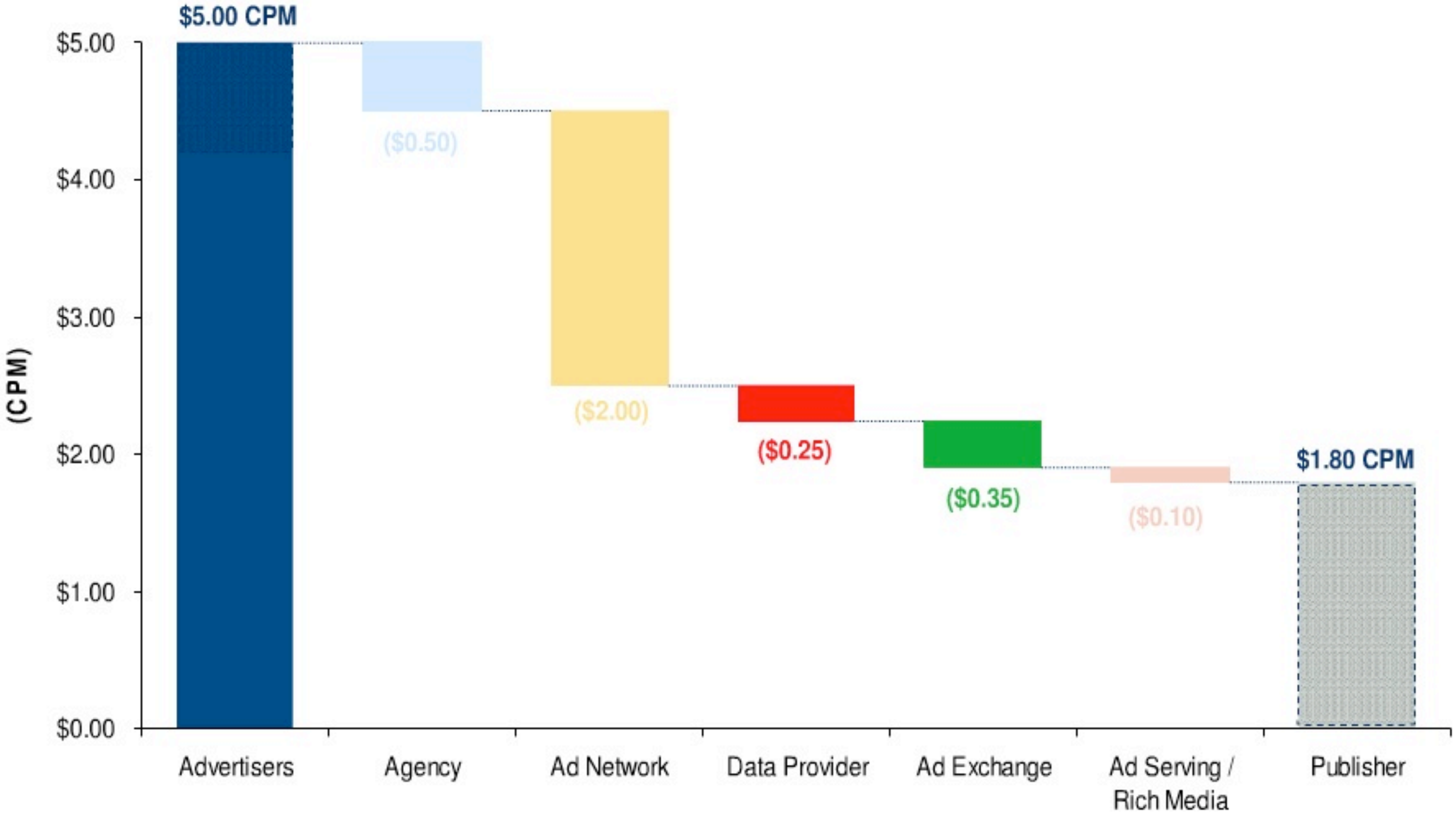
Data Mining Software
Discover insights hidden in your existing data using SPSS solutions.
www.spss.com

Ranking: Improve in a measured way



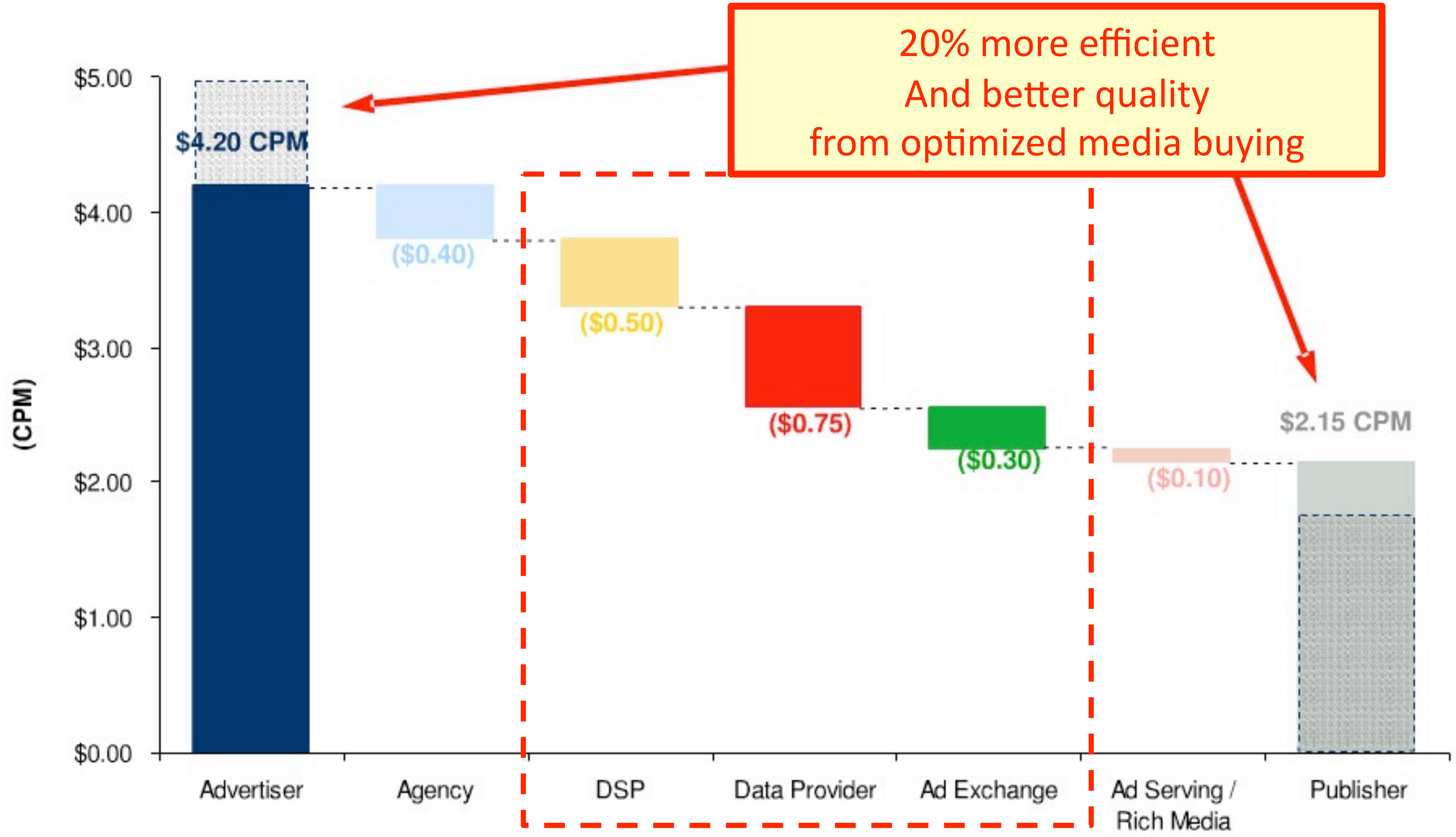
Note: Represents approximate scale and competitive ratios, not based on competitive measurements

Carving Up the Stack - Network World



GCA SAVVIAN [Kawaja, 2010]

Carving Up the Stack - DSP / Exchange World



20% more efficient
And better quality
from optimized media buying

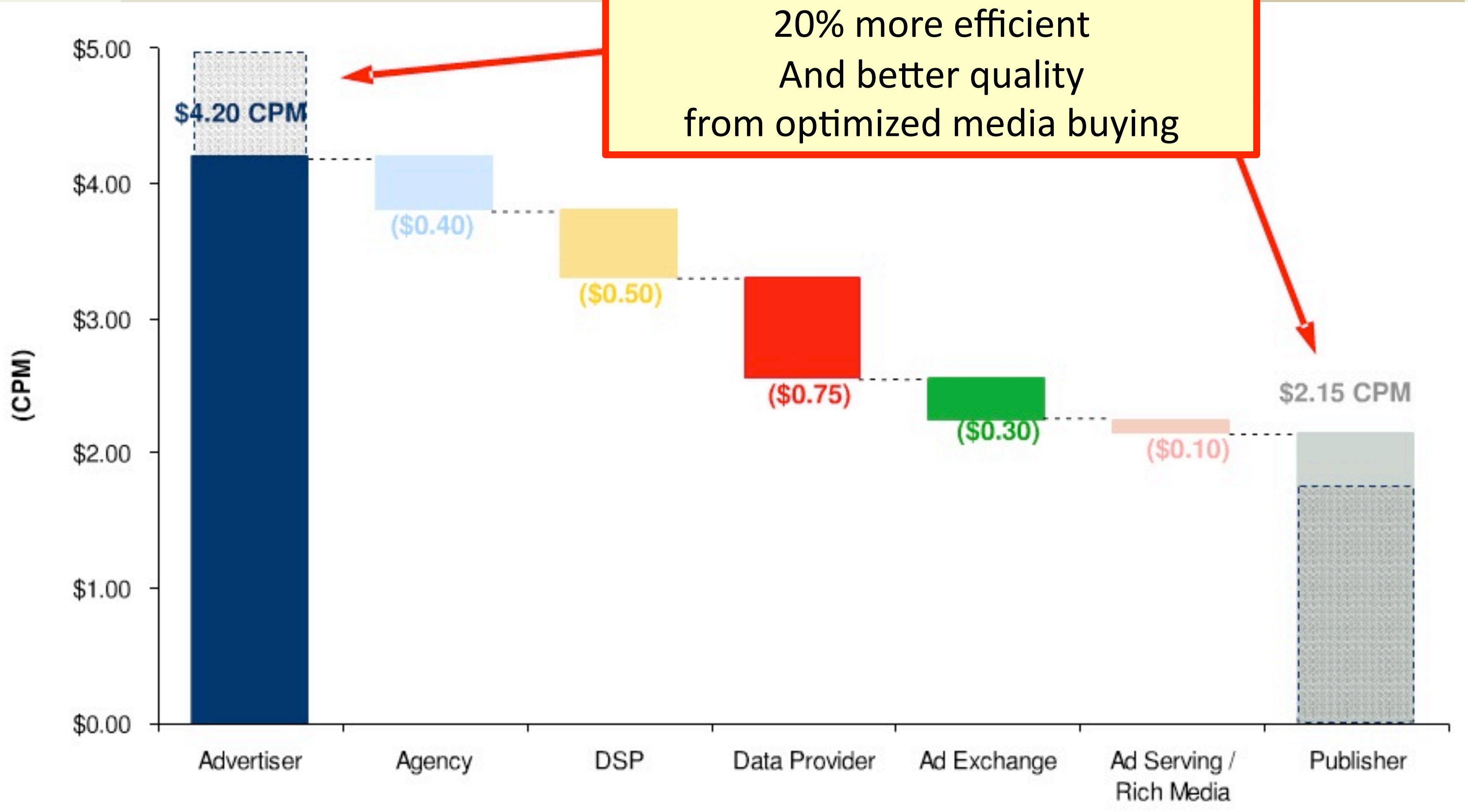
GCA SAVVIAN [Kawaja, 2010]

3rd Generation

Empowering advertisers thru DSPs and RTB to buy audiences

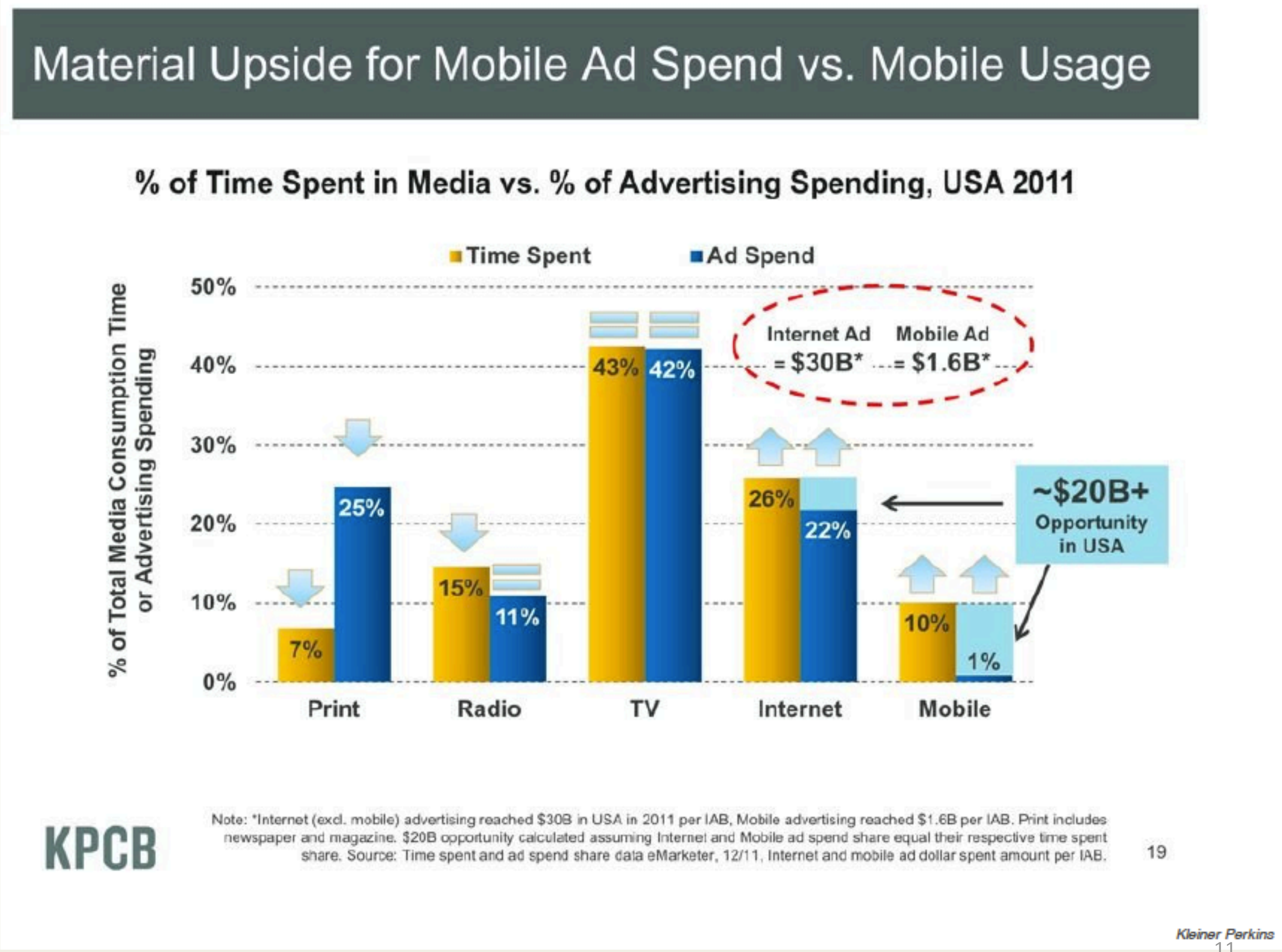
DEMAND

SUPPLY



20% more efficient
And better quality
from optimized media buying

Empowering advertisers thru DSPs connect the data dots and buy audiences



Material Upside for Mobile Ad Spend vs. Mobile Usage

% of Time Spent in Media vs. % of Advertising Spending, USA 2011

Note: *Internet (excl. mobile) advertising reached $30B in USA in 2011 per IAB, Mobile advertising reached $1.6B per IAB. Print includes newspaper and magazine. $20B opportunity calculated assuming Internet and Mobile ad spend share equal their respective time spent share. Source: Time spent and ad spend share data eMarketer, 12/11, Internet and mobile ad dollar spent amount per IAB.

KPCB

19

Kleiner Perkins

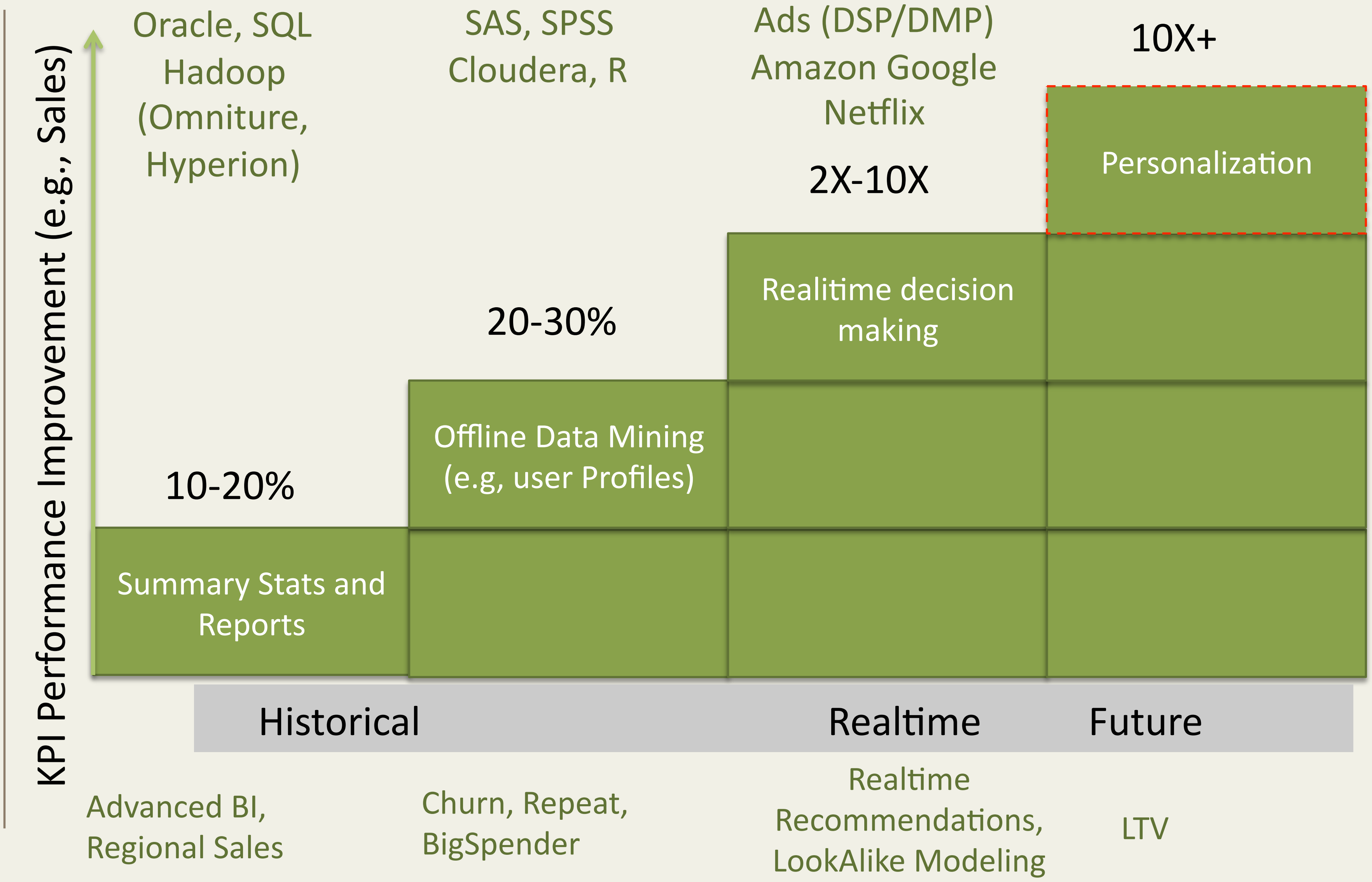
Obama's ‘Data Science’ Victory

- **Raising a billion dollars**
- **Customized fundraising invitations**
  - When the Obama campaign emailed supporters to join a $40,000-a-ticket dinner in June at the New York home of actress Sarah Jessica Parker, journalists at ProPublica noticed something odd. They uncovered seven versions of the email solicitation for the fundraiser, some mentioning a second fundraiser that night, a concert by Mariah Carey, others that Ms. Parker is a mother, and still others that Vogue editor Anna Wintour would be at the dinner. [WSJ, 11/18/2012]
- **Predicting Turnout and "persuadability" of voters.**
  - Multivariate tests identified issues and positions that could move undecided voters, ProPublica said: "The persuasion scores allowed the campaign to focus its outreach efforts—and their volunteer calls—on voters who might actually change their minds as the result. It also guided them in what policy messages individual voters should hear."

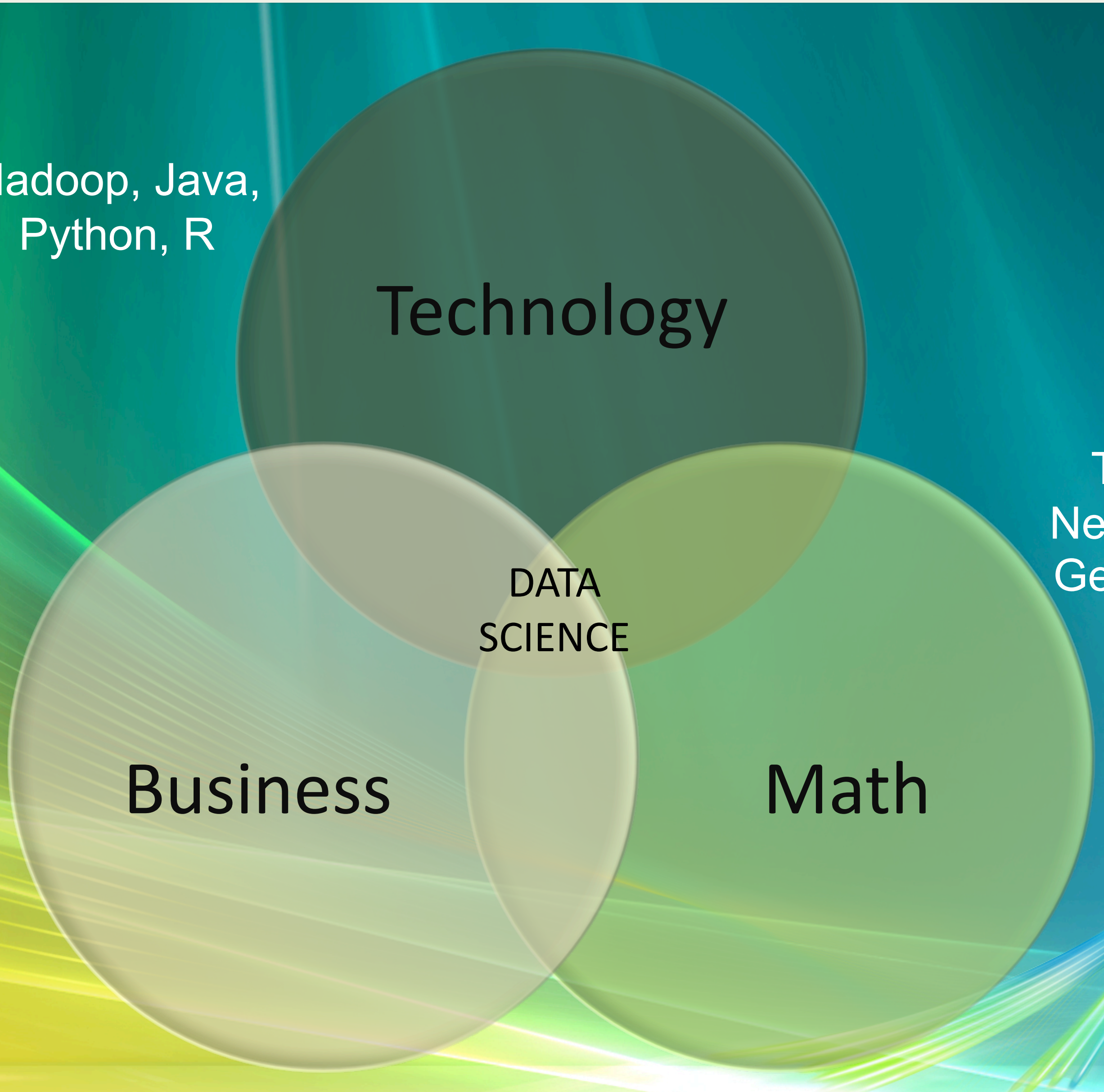
80 pieces of information on each person

- The Obama campaign has used cookies to track its supporters online since the 2008 election.
- It spent the past 18 months creating a new, unified database, factoring in some 80 pieces of information about each person, from age, race and sex to voting history.
- The Romney campaign says it tried to match the Obama campaign's collection and analysis of data but had to start from scratch and had just seven months after the primaries.

80 pieces of information on each person


KPI Performance Improvement (e.g., Sales)
Oracle, SQL
Hadoop
(Omniture,
Hyperion)
SAS, SPSS
Cloudera, R
Ads (DSP/DMP)
Amazon Google
Netflix
10X+
Personalization
2X-10X
Realitime decision making
20-30%
Offline Data Mining (e.g, user Profiles)
10-20%
Summary Stats and Reports
Historical
Realtime
Future
Advanced BI, Regional Sales
Churn, Repeat, BigSpender
Realtime Recommendations, LookAlike Modeling
LTV

Data Science

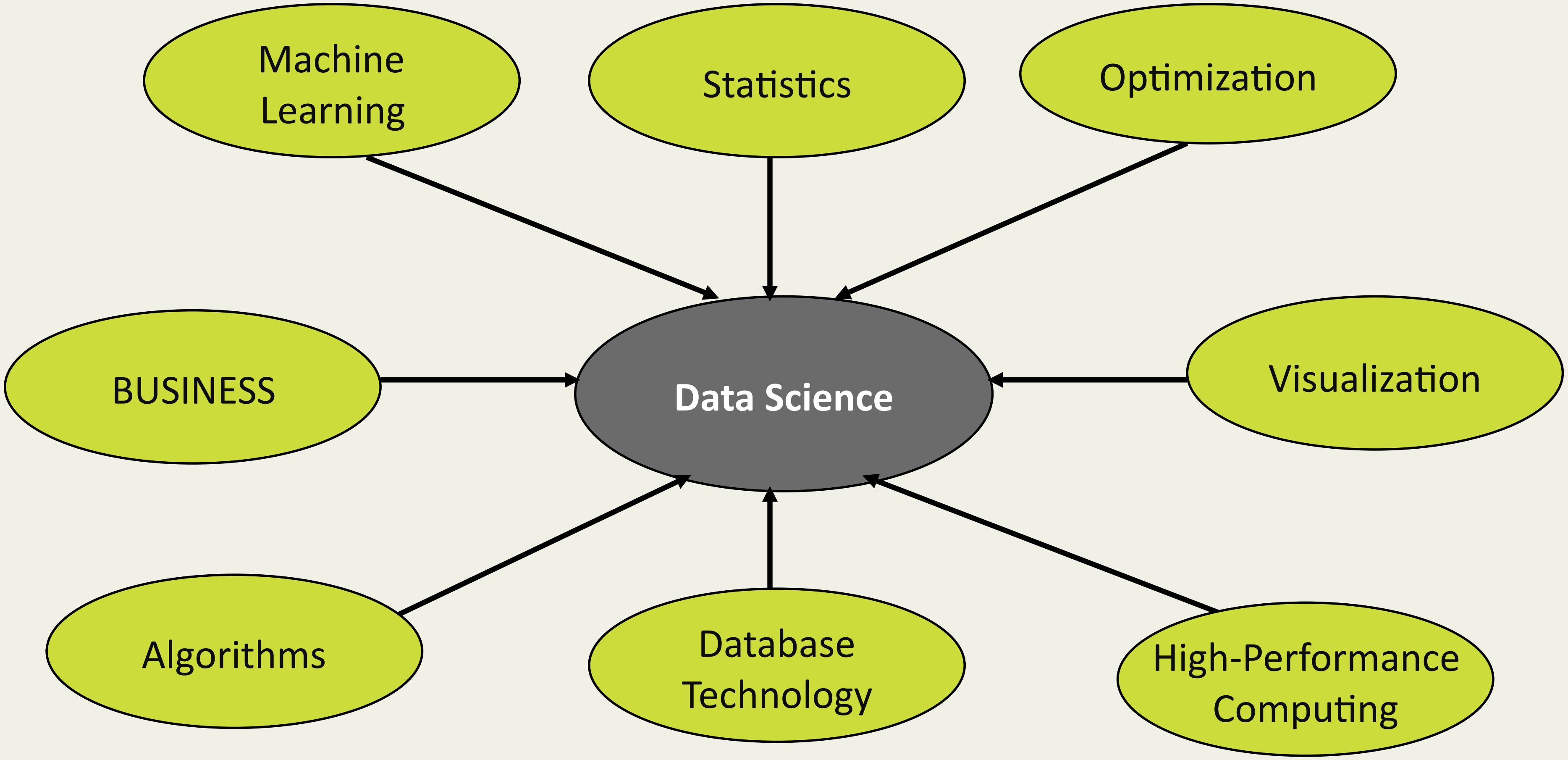


Hadoop, Java, Python, R

Statistics, Optimization Theory, Social Network Analytics, Geo-Informational Science

Digital Advertising & Marketing, Econometrics, Web Search

Data Science: Confluence of Multiple Disciplines



Data Science

○ **Data science incorporates varying elements and builds on techniques and theories from many fields, including**

- math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing
- with the goal of extracting meaning from data and creating data products.

○ **Data products**

- Reports, scoring systems, classifiers, clustering systems, forecasting systems, optimization systems
- Batch or realtime
- Cloud-based or on premise…
- As a service or in-house

Talk Outline

- Big Data Case Studies
- Data Science
- Data Science as a Service (DSaaS)
- Data Science Bubble
- Conclusions

W3i.com

- **W3i is a leader in monetization and user acquisition services for mobile and desktop apps**.
  - For monetization W3i maximizes app revenue through promoting high-yielding offers.
  - For user acquisition, W3i delivers a high-volume, high-quality audience at performance-based pricing.
  - W3i's technology has been optimized on more than 700 million app installs, that's 500 installs per minute.
  - HQ at Saint Cloud, Minnesota
- **Generating Terrabytes of data per month**

W3i.com

- ○ **W3i is a leader in monetization and user acquisition services for mobile and desktop apps**.
  - For monetization W3i maximizes app revenue through promoting high-yielding offers.
  - For user acquisition, W3i delivers a high-volume, high-quality audience at performance-based pricing.
  - W3i's technology has been optimized on more than 700 million app installs, that's 500 installs per minute.

  - HQ at Saint Cloud, Minnesota

- ○ **Generating Terrabytes of … month**

Hiring data scientists

Embracing Data Science

- Consultants
- Data Science as a Service (DSaaS)
- Build in-house team

Embracing Data Science

- Consultants
  - Most consulting firms have yet to assemble data science teams (e.g., Accenture, Deloitte IBM are all in the early stages of leading big data projects for their clients)
  - Data scientists prefer to build as opposed to give advice
- Data Science as a Service (DSaaS)
- Build in-house team

Embracing Data Science

- Consultants
- Data Science as a Service (DSaaS)
- Build in-house team

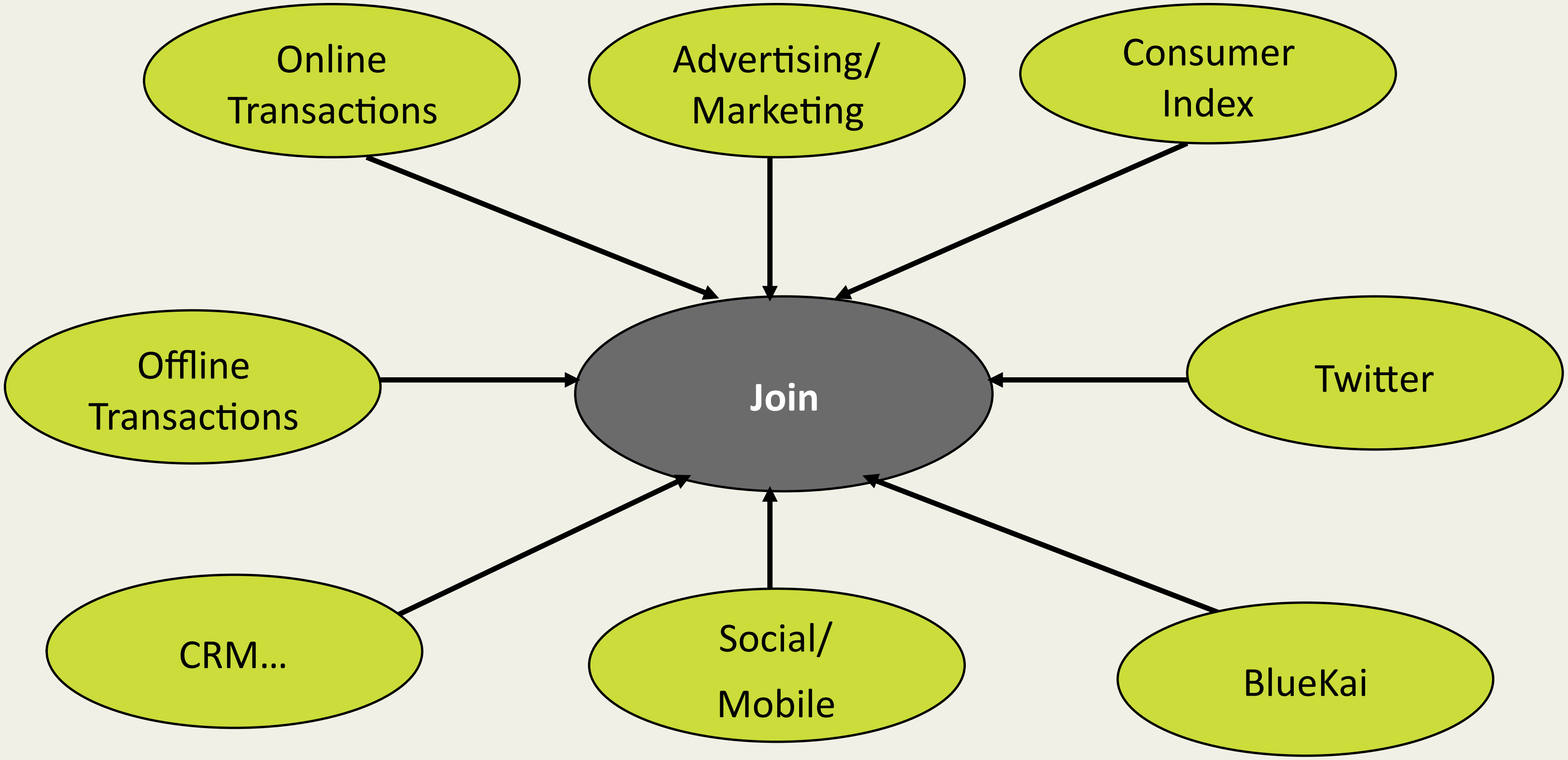
Software as a service (SaaS)

- **Software as a service is a software delivery model in which software and associated data are centrally hosted on the cloud.**
- **SaaS is typically accessed by users using a thin client via a web browser.**
- **Applications include accounting, collaboration, customer relationship management (CRM), management information systems (MIS), enterprise resource planning (ERP), invoicing, human resource management (HRM), content management (CM) and service desk management.**
- **Advantages:**
  - Reduce IT support costs by outsourcing hardware and software maintenance and support to the SaaS provider.

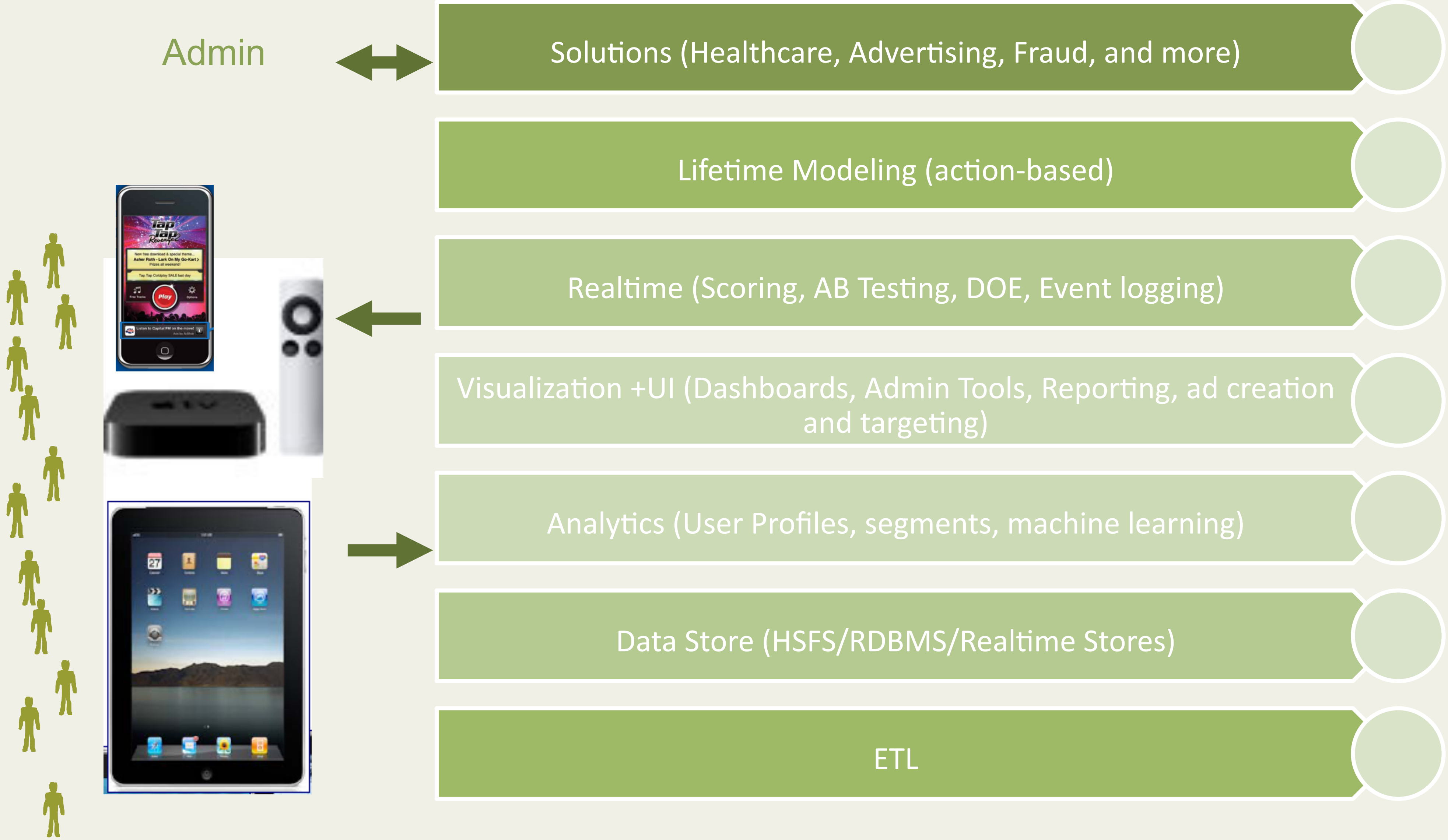
Data Science as a Service (DSaaS)

- **DSaaS unlocks value and insight from large-scale transaction streams.**
- **Using a cloud-based solution, data scientists deliver powerful dashboards to business users and accelerate their own discovery efforts.**
    - Work with data at the speed of thought
    - Make realtime predictions
    - Realtime ingestion of data? Logfile transfer based?
- **DSaaS is typically accessed by business users using a thin client via a web browser**
    - Cloud-based or on premise

ETL: Joining of disparate data sources



DSaaS Stack



Admin

Solutions (Healthcare, Advertising, Fraud, and more)

Lifetime Modeling (action-based)

Realtime (Scoring, AB Testing, DOE, Event logging)

Visualization +UI (Dashboards, Admin Tools, Reporting, ad creation and targeting)

Analytics (User Profiles, segments, machine learning)

Data Store (HSFS/RDBMS/Realtime Stores)

ETL

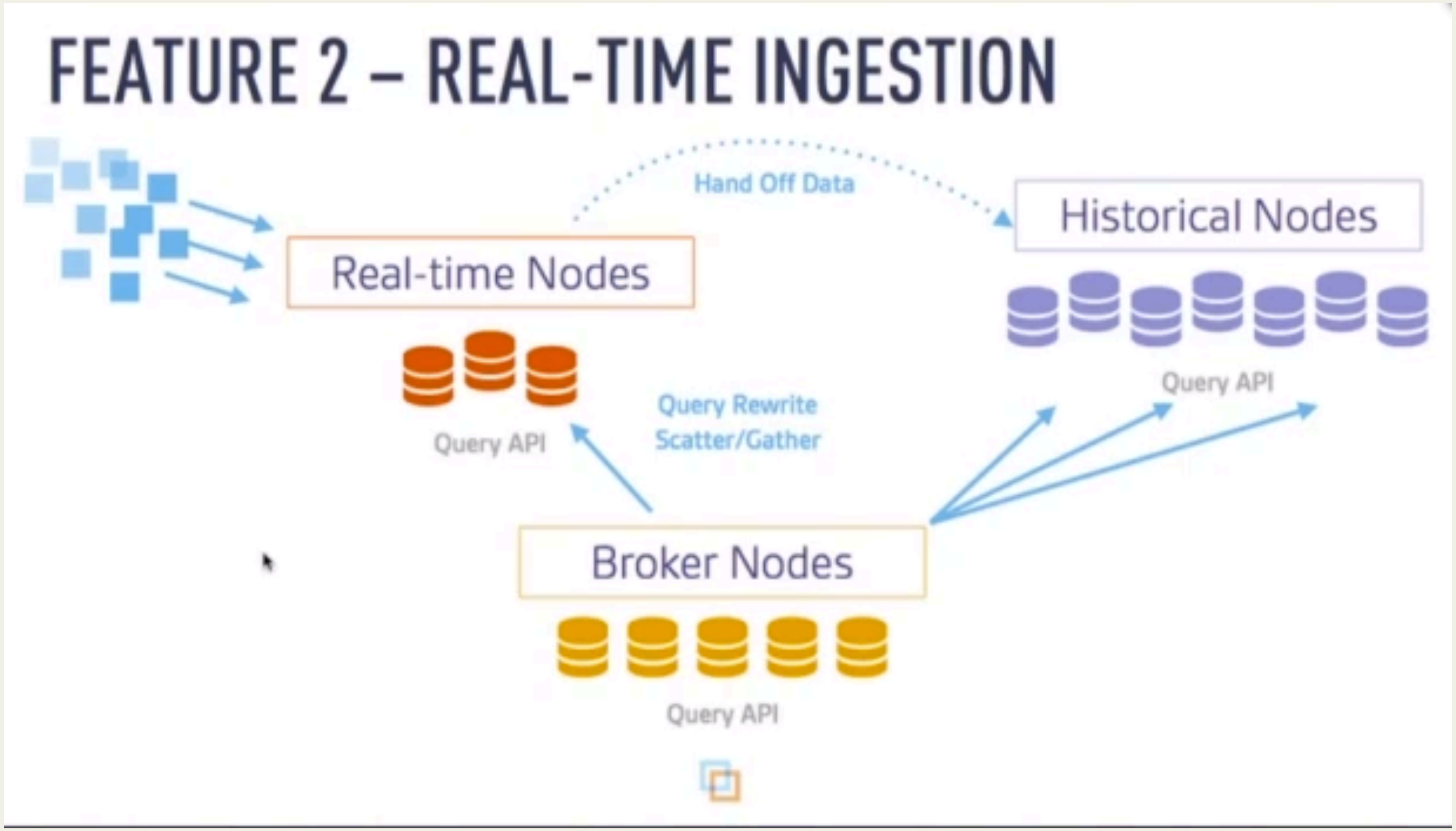
Data Science as a Service (DSaaS)

- **HDFS/HIVE/HBASE, RDBMS, Data warehouses**
- **Realtime analytics on high velocity, voluminous data, and varied data**
  - Sub-second response times

Data Science as a Service (DSaaS)

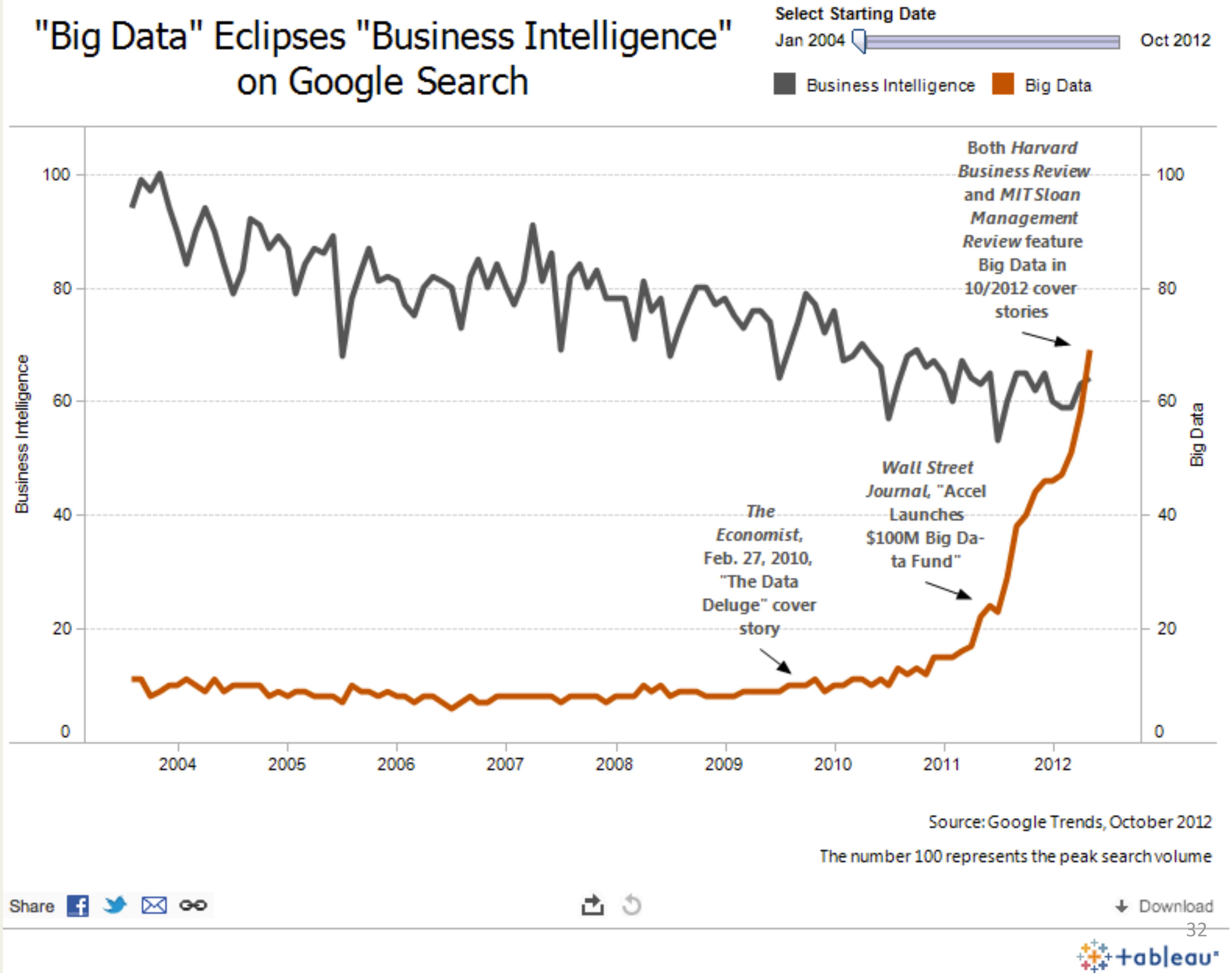
- **Age the data**
- **Summarize the data**
  - Discretize the data
- **Project the data**
- **Time series analysis**
- **Feature engineering**

Realtime Analytics Data Stores



FEATURE 2 – REAL-TIME INGESTION

- Druid System From MetaMarkets;
- Similar to Cloudera’s Impala and Google’s Dremel



"Big Data" Eclipses "Business Intelligence" on Google Search

Select Starting Date

Jan 2004 — Oct 2012

■ Business Intelligence ■ Big Data

Both *Harvard Business Review* and *MIT Sloan Management Review* feature Big Data in 10/2012 cover stories

Wall Street Journal, "Accel Launches $100M Big Da-ta Fund"

The Economist, Feb. 27, 2010, "The Data Deluge" cover story

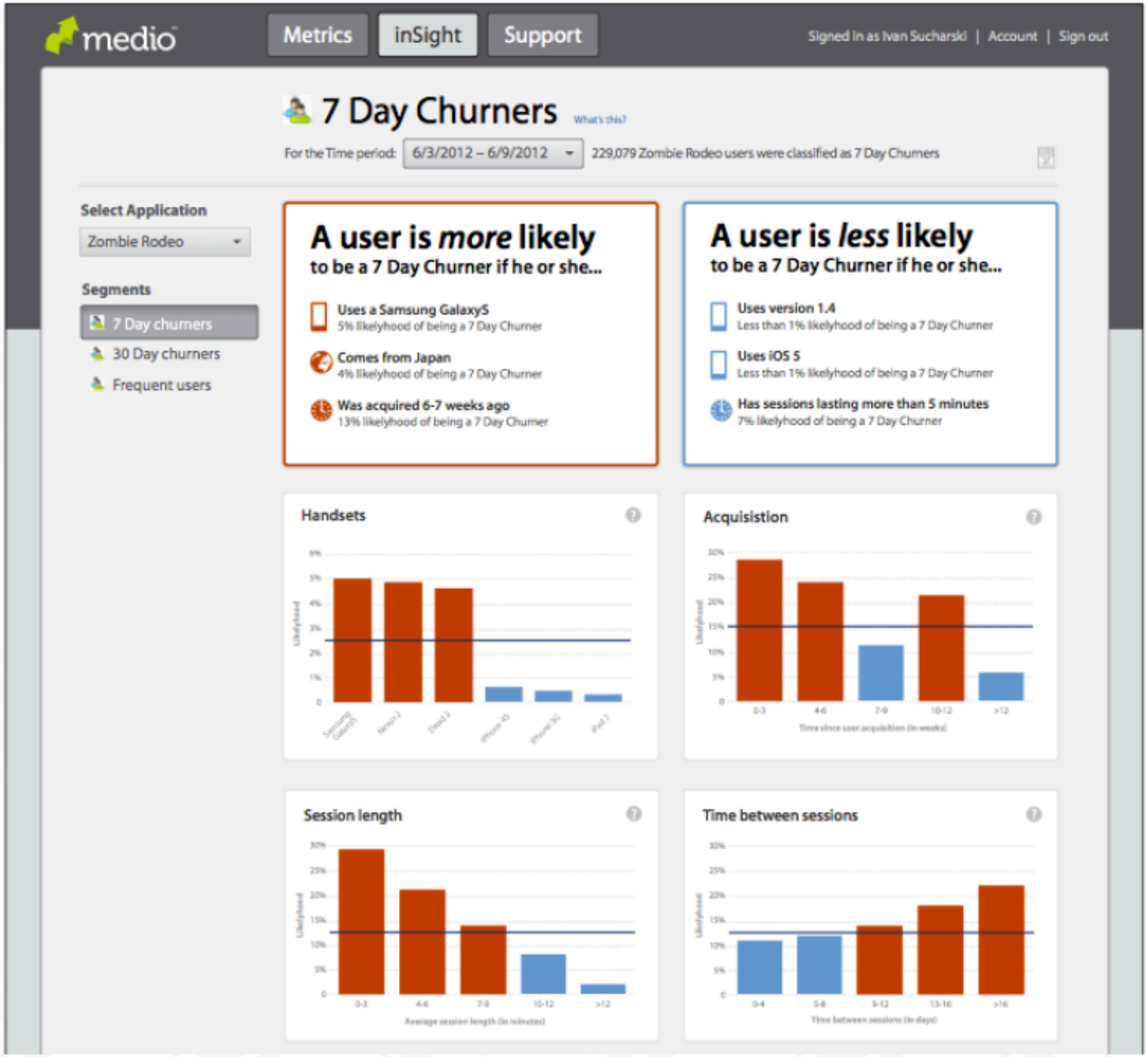
Source: Google Trends, October 2012

The number 100 represents the peak search volume

Data Science as a Service (DSaaS)

- ○ **Classification (logistic regression, decision trees, support vector machines)**
  - Churner or not
- ○ **Clustering (K-means, EM, LDA)**
  - Group customers into segments
- ○ **Prediction and Ranking (linear regression, boosted decision trees)**
  - Click through rate
  - Number of days in hospital
- ○ **Optimization (portfolio optimization)**

Segment Consumer Population



7 Day Churners

For the Time period: 6/3/2012 – 6/9/2012 ▾ 229,079 Zombie Rodeo users were classified as 7 Day Churners

Select Application

Zombie Rodeo ▾

Segments

- 7 Day churners
- 30 Day churners
- Frequent users

A user is *more* likely
to be a 7 Day Churner if he or she...

- Uses a Samsung GalaxyS
  5% likelyhood of being a 7 Day Churner
- Comes from Japan
  4% likelyhood of being a 7 Day Churner
- Was acquired 6-7 weeks ago
  13% likelyhood of being a 7 Day Churner

A user is *less* likely
to be a 7 Day Churner if he or she...

- Uses version 1.4
  Less than 1% likelyhood of being a 7 Day Churner
- Uses iOS 5
  Less than 1% likelyhood of being a 7 Day Churner
- Has sessions lasting more than 5 minutes
  7% likelyhood of being a 7 Day Churner

Look-alike modeling

- **Observe individuals who just transacted; find more who look-like these**
  - Positive and negative example
  - Learn a model from data based on consumer RFM and demographic features
- **Data driven e.g., from QuantCast [KDD 2010]**
  - Millions of partner sites
  - 10 Billion weblog records (ad tag firing events from publisher); 250 Billion per month
  - 1 Billion users globally
  - 15 terabytes per day of new data
- **Forecasting (see CIKM 2010 papers and posters)**
  - How many of uniques? What-if I increase by bid price?
- **Model drift**

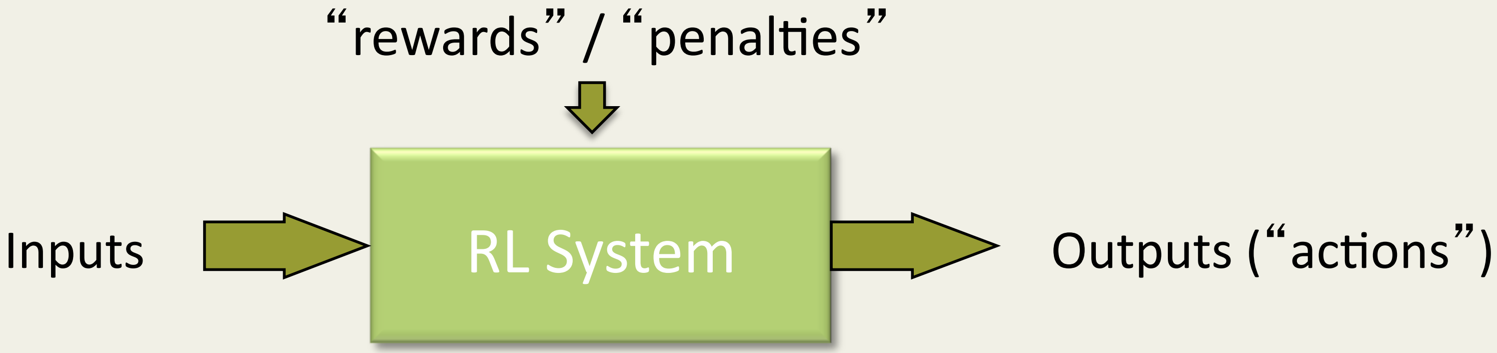
Econometrics: Causality versus Correlation

- **Econometrics is the study of the applications of statistical methods to the analysis of economic phenomena.**

- **What distinguishes an econometrician from a statistician is the former's preoccupation with problems caused by violations of statistician's standard assumptions**

- **Econometrics is not about prediction. It is about understanding the relationships and causality.**

Reinforcement Learning



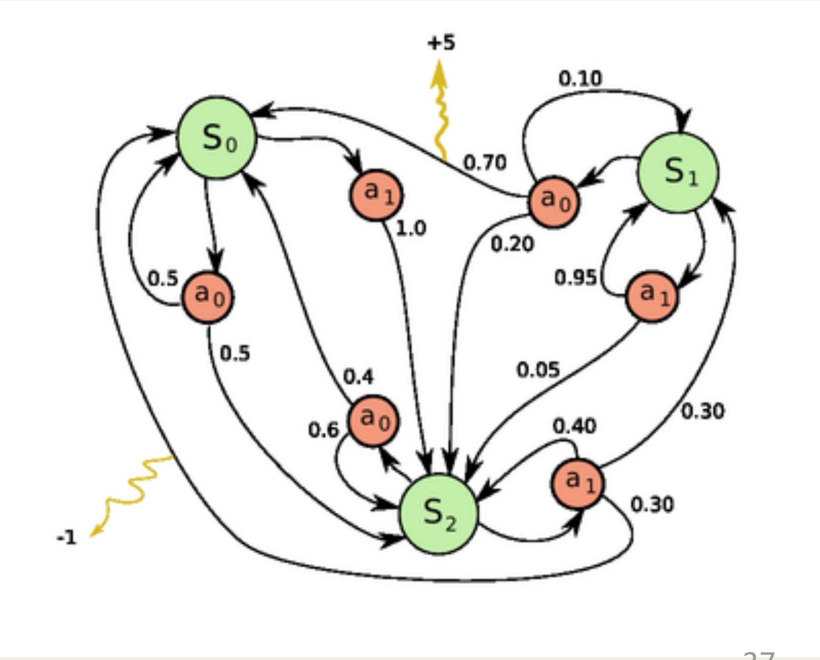
“rewards” / “penalties”

Inputs

RL System

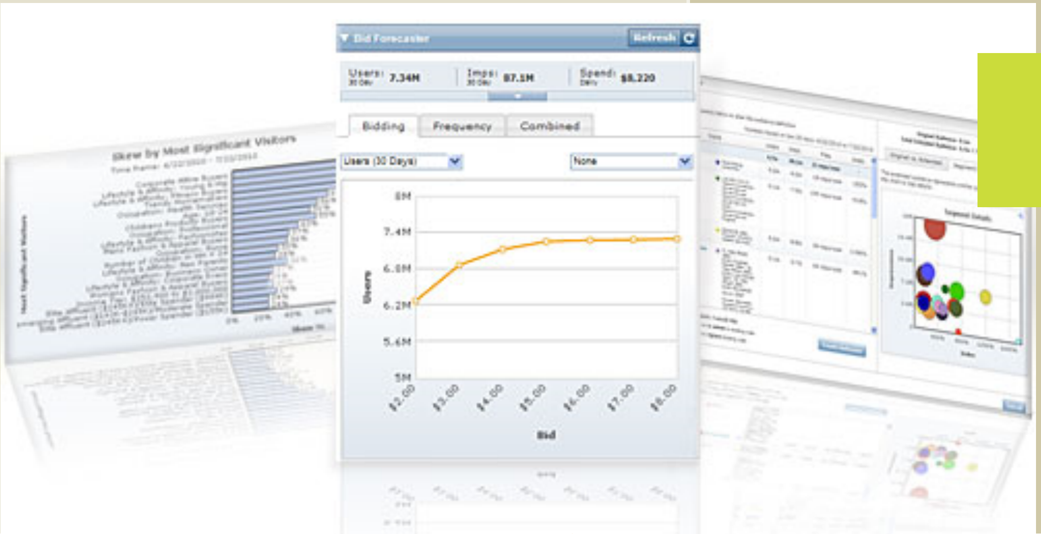
Outputs (“actions”)

Objective: get as much reward as possible

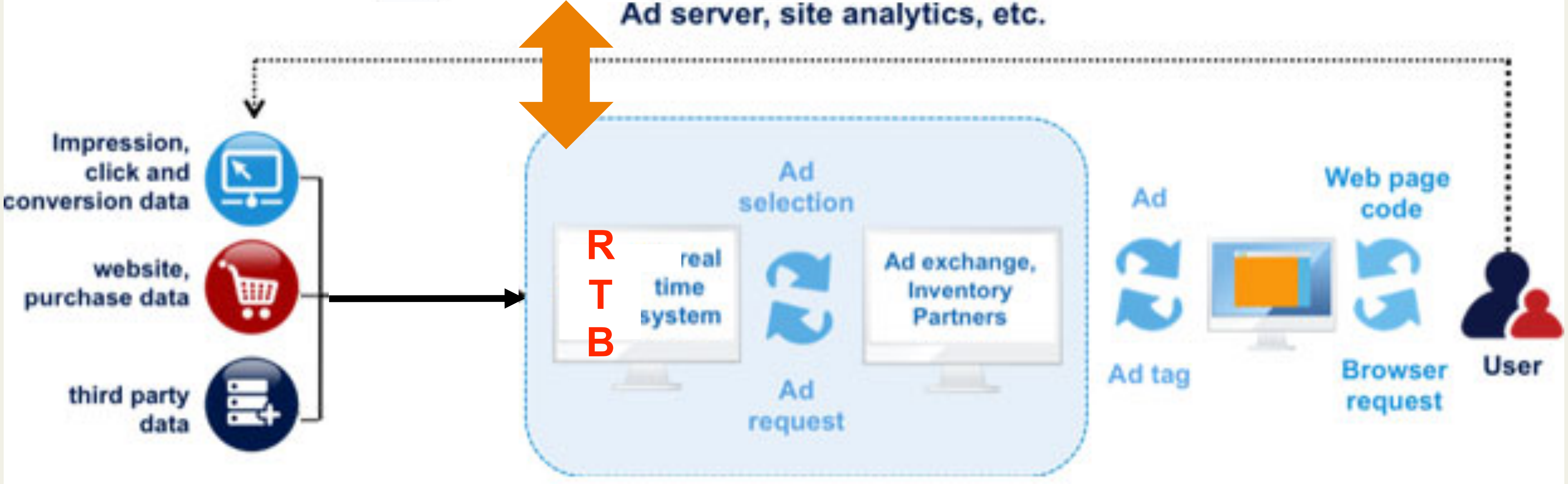


Demand side platform <-> Exchanges

Advertiser



Speak the language of marketing folk



Ad server, site analytics, etc.

Impression, click and conversion data

website, purchase data

third party data

Ad selection

RTB

real time system

Ad exchange, Inventory Partners

Ad request

Ad

Ad tag

Web page code

Browser request

User

Behavior | Advertiser/DSP | Exchange | Publisher | Consumer

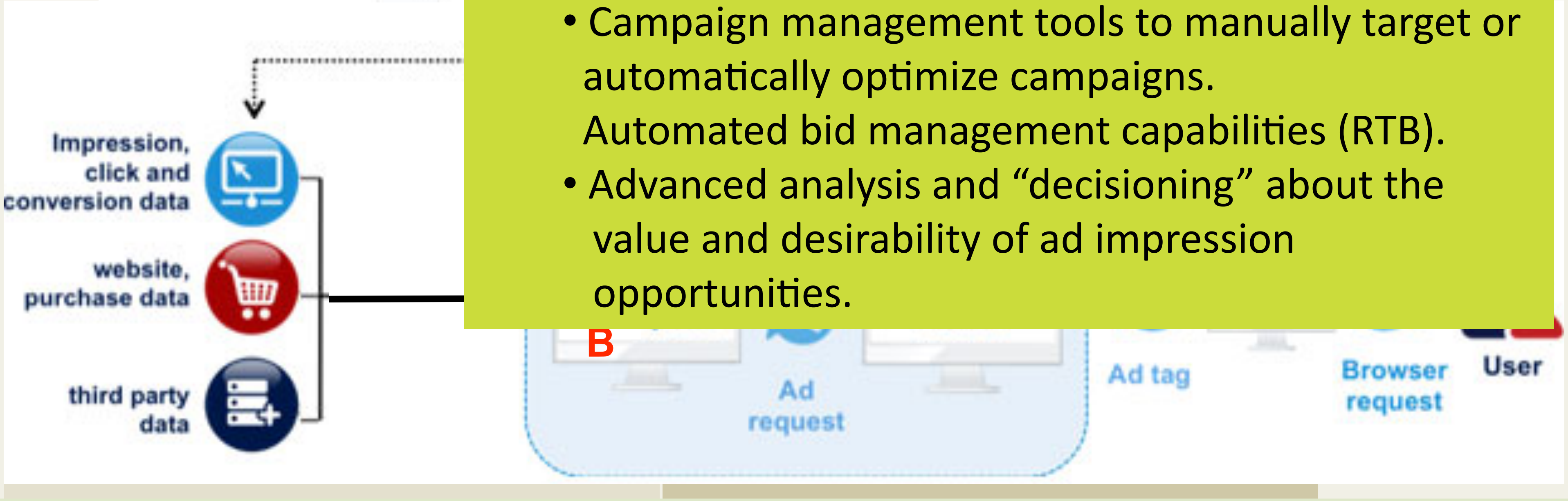
DEMAND | Exchange | SUPPLY

Demand side platform <-> Exchanges

Advertiser

Speak the language
of marketing folk

- A DSP is a demand side trading desk; is a SaaS
  Connects to multiple ad exchanges and other media suppliers
  - Campaign management tools to manually target or automatically optimize campaigns.
    Automated bid management capabilities (RTB).
  - Advanced analysis and "decisioning" about the value and desirability of ad impression opportunities.



Behavior | Advertiser/DSP | Exchange | Publisher | Consumer

DEMAND | Exchange | SUPPLY

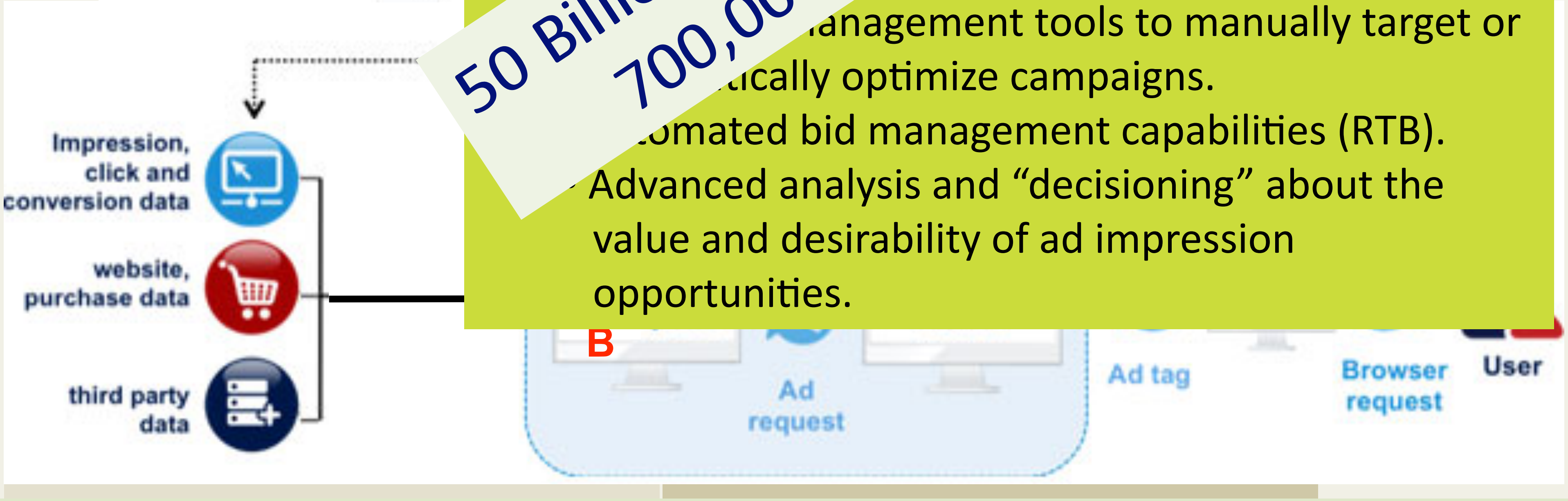
Demand side platform <-> Exchanges

Advertiser

Speak the language
of marketing folk

- A DSP is a d… g desk; is a SaaS
  Con… d exchanges and other media
- …anagement tools to manually target or …tically optimize campaigns.
- …tomated bid management capabilities (RTB).
- Advanced analysis and "decisioning" about the value and desirability of ad impression opportunities.

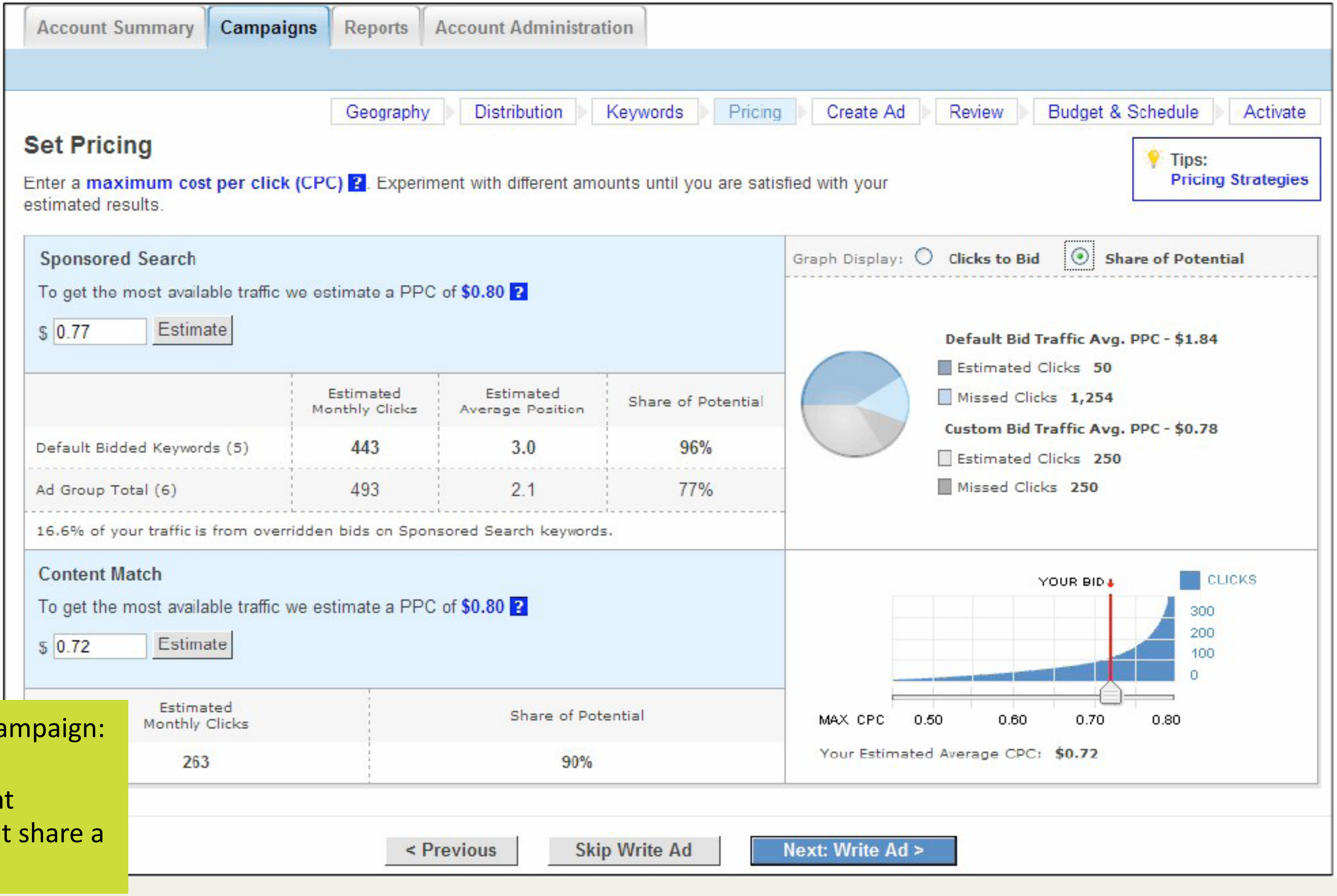
50 Billion Impressions
700,000 QPS



Behavior | Advertiser/DSP | Exchange | Publisher | Consumer

DEMAND | Exchange | SUPPLY

Set bid based on targeting criteria



Set Pricing

Enter a **maximum cost per click (CPC)** ? . Experiment with different amounts until you are satisfied with your estimated results.

Tips: Pricing Strategies

Sponsored Search

To get the most available traffic we estimate a PPC of **$0.80** ?

$ 0.77 Estimate

	Estimated Monthly Clicks	Estimated Average Position	Share of Potential
Default Bidded Keywords (5)	443	3.0	96%
Ad Group Total (6)	493	2.1	77%

16.6% of your traffic is from overridden bids on Sponsored Search keywords.

Graph Display: ○ Clicks to Bid ● Share of Potential

Default Bid Traffic Avg. PPC - $1.84

- ▨ Estimated Clicks **50**
- □ Missed Clicks **1,254**

Custom Bid Traffic Avg. PPC - $0.78

- □ Estimated Clicks **250**
- ▨ Missed Clicks **250**

Content Match

To get the most available traffic we estimate a PPC of **$0.80** ?

$ 0.72 Estimate

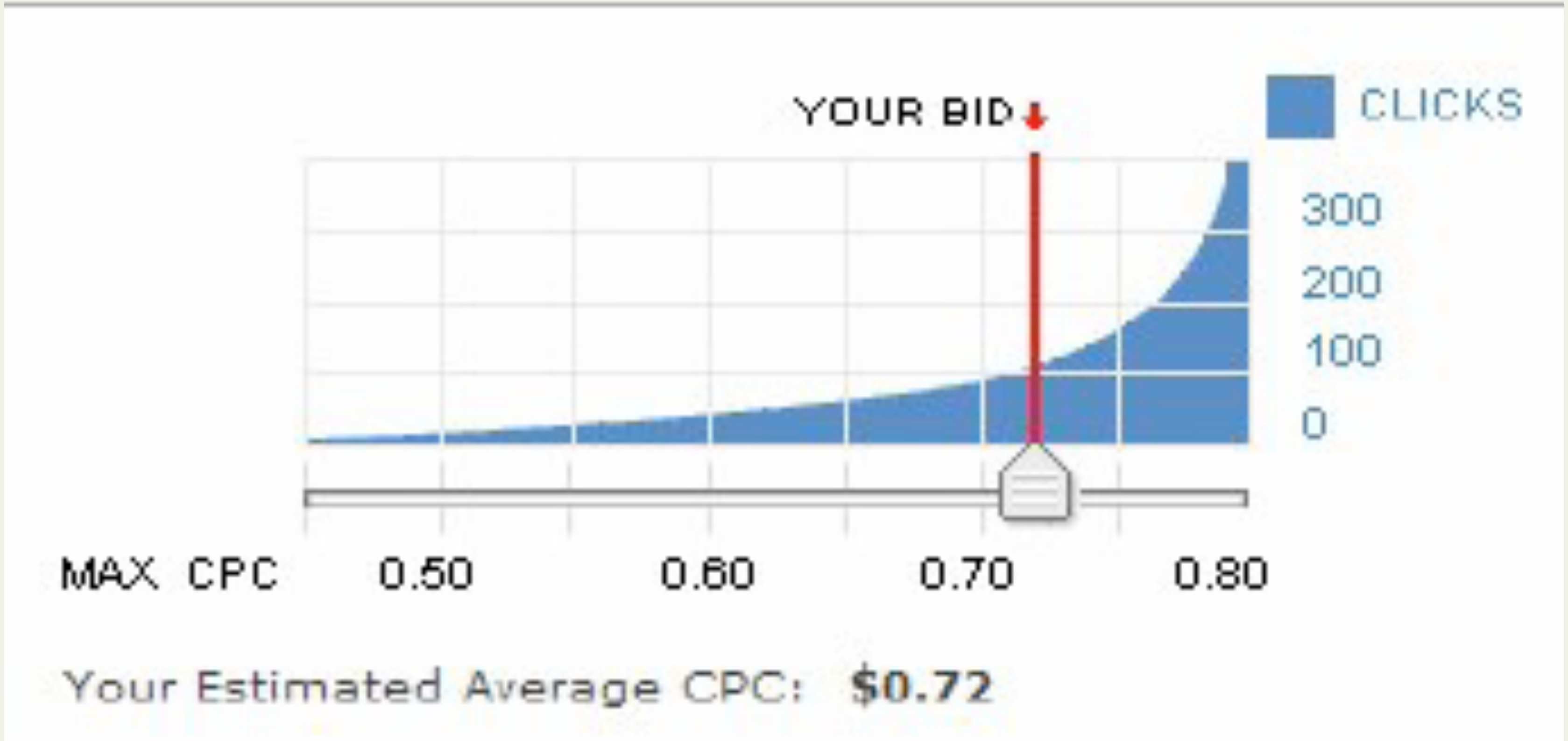
Estimated Monthly Clicks	Share of Potential
263	90%

Your Estimated Average CPC: **$0.72**

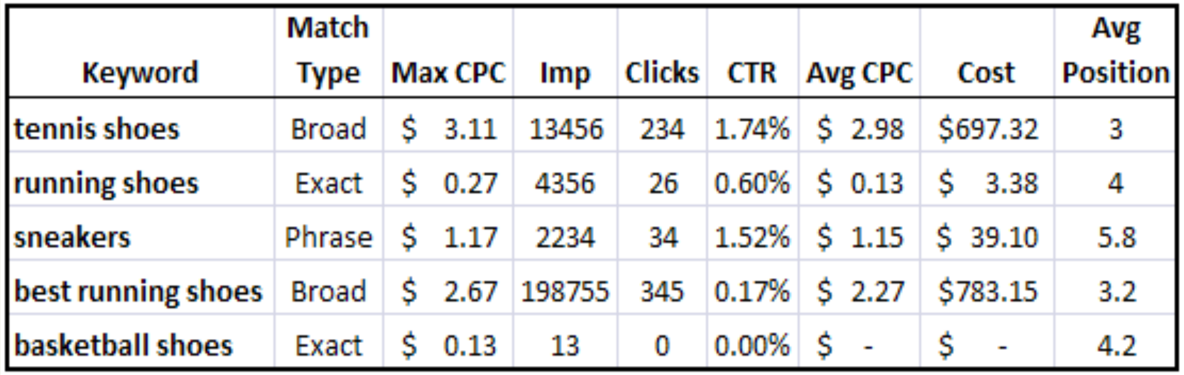
< Previous　Skip Write Ad　Next: Write Ad >

Advertising campaign: a series of advertisement messages that share a single idea

Forecasting

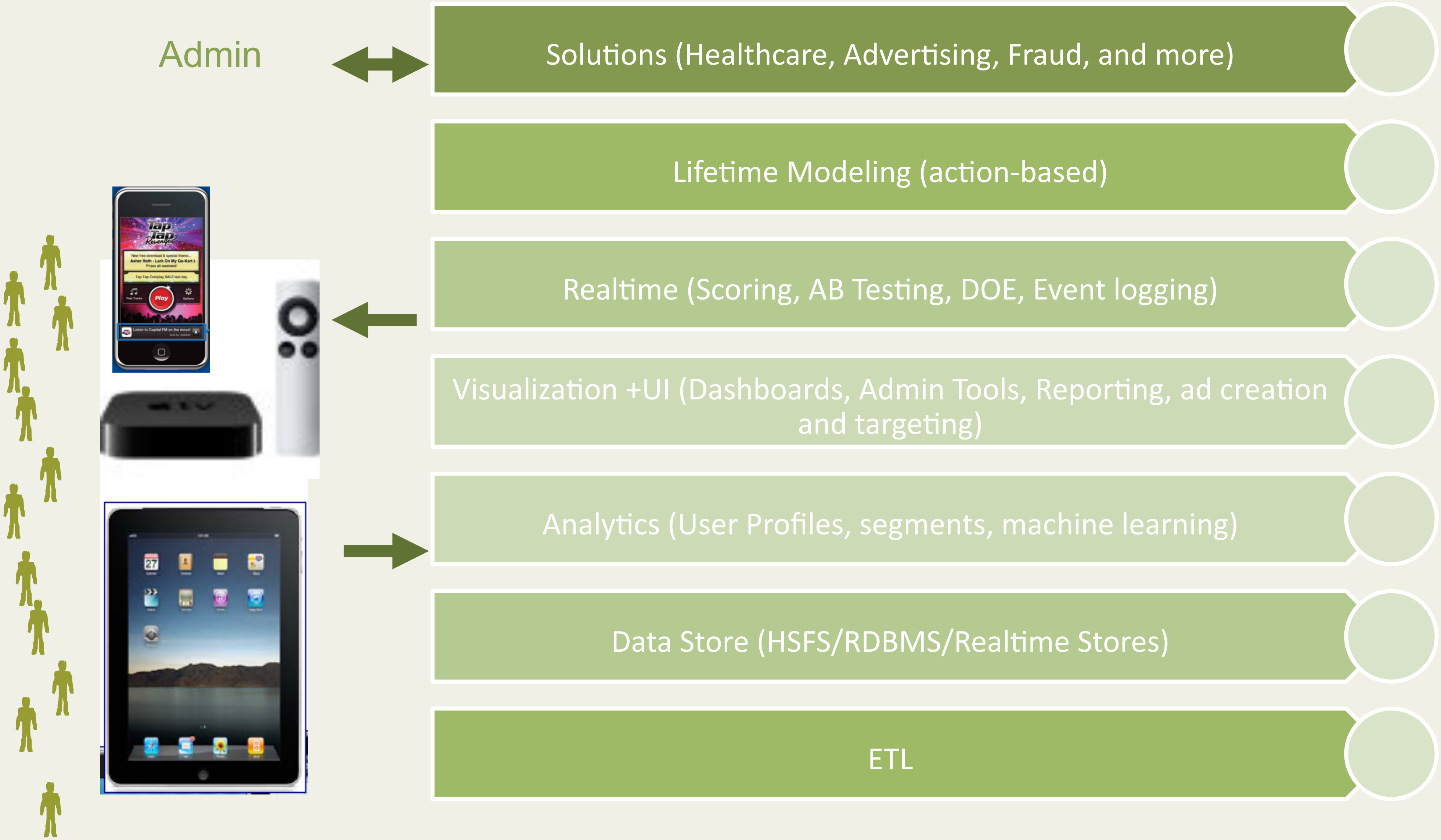


Your Estimated Average CPC: $0.72



Keyword	Match Type	Max CPC	Imp	Clicks	CTR	Avg CPC	Cost	Avg Position
tennis shoes	Broad	$ 3.11	13456	234	1.74%	$ 2.98	$697.32	3
running shoes	Exact	$ 0.27	4356	26	0.60%	$ 0.13	$ 3.38	4
sneakers	Phrase	$ 1.17	2234	34	1.52%	$ 1.15	$ 39.10	5.8
best running shoes	Broad	$ 2.67	198755	345	0.17%	$ 2.27	$783.15	3.2
basketball shoes	Exact	$ 0.13	13	0	0.00%	$ -	$ -	4.2

DSaaS Stack



Admin

Solutions (Healthcare, Advertising, Fraud, and more)

Lifetime Modeling (action-based)

Realtime (Scoring, AB Testing, DOE, Event logging)

Visualization +UI (Dashboards, Admin Tools, Reporting, ad creation and targeting)

Analytics (User Profiles, segments, machine learning)

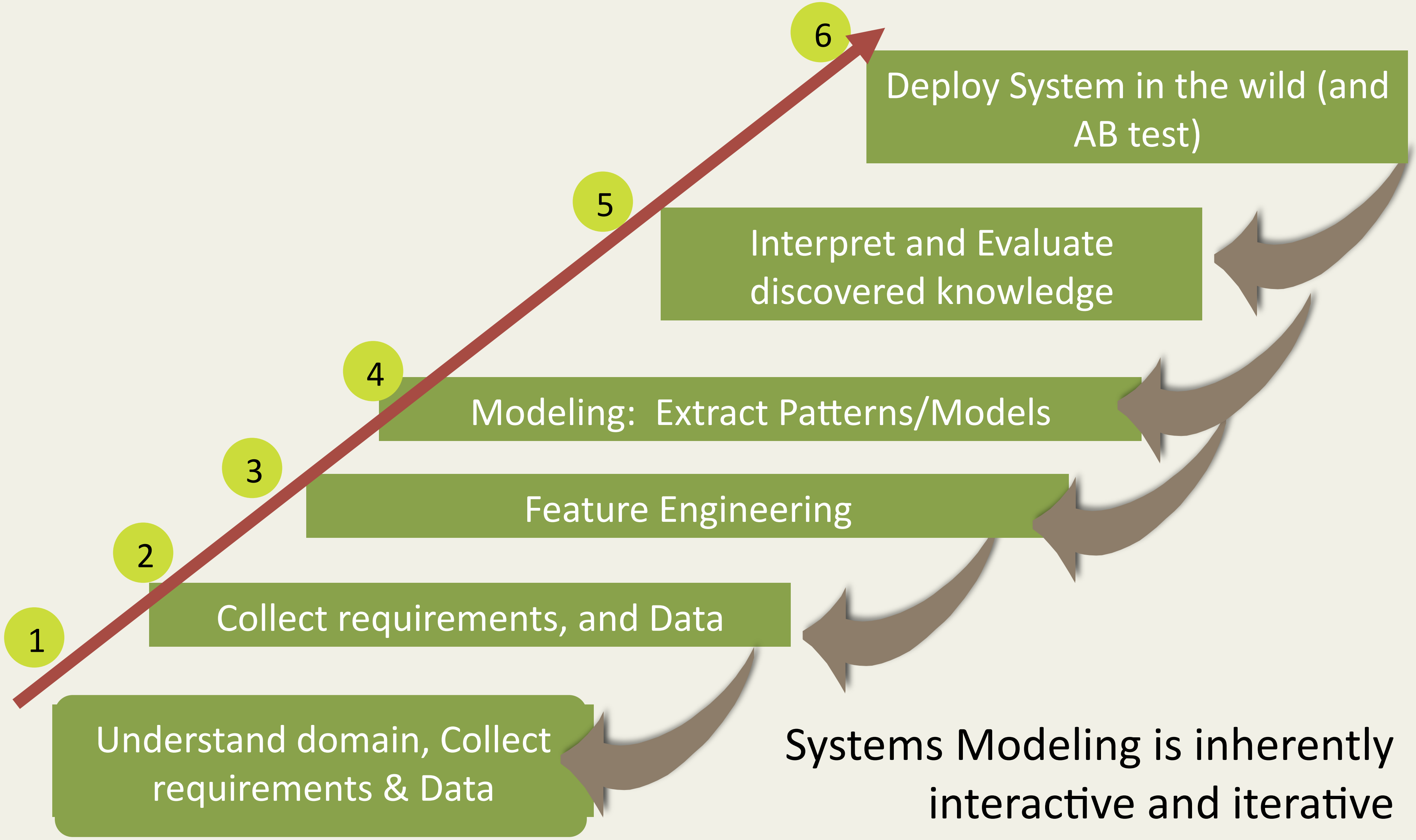
Data Store (HSFS/RDBMS/Realtime Stores)

ETL

DSaaS Companies

- **MetaMarkets.com**
- **Medio**
- **InferSystems**
- **BigML:**
  - Much like Prior Knowledge, BigML is a startup that combines data with machine learning to help give normal people access to the smarts to help them answer questions with their data.
  - It hopes to let people do machine learning in four easy steps: set up a data source; create a dataset; create a model; and generate predictions.
- **IBM Smarter Analytics**

6 Steps to Data Modeling in Practice


6
Deploy System in the wild (and AB test)
5
Interpret and Evaluate discovered knowledge
4
Modeling: Extract Patterns/Models
3
Feature Engineering
2
Collect requirements, and Data
1
Understand domain, Collect requirements & Data

Systems Modeling is inherently interactive and iterative

Key Core Skills

- **Metrics (accuracy, precision, recall, AUC, ROC, MSE)**
- **Linear Algebra, Matrices**
- **Optimization, Gradient Descent**
- **Machine learning (linear regression, decision trees, Kmeans)**
- **Bayesian Statistics (Hierarchical models)**
- **Markov (decision) Processes**
- **Hadoop/SQL, R, Matlab, Python, Java**

Types of Machine Learning

- **Supervised Learning**

  Generates a function that maps inputs to desired outputs. For example, in a classification problem, the learner approximates a function mapping a vector into classes by looking at input-output examples of the function.

- **Unsupervised Learning**

  Models a set of inputs: like clustering

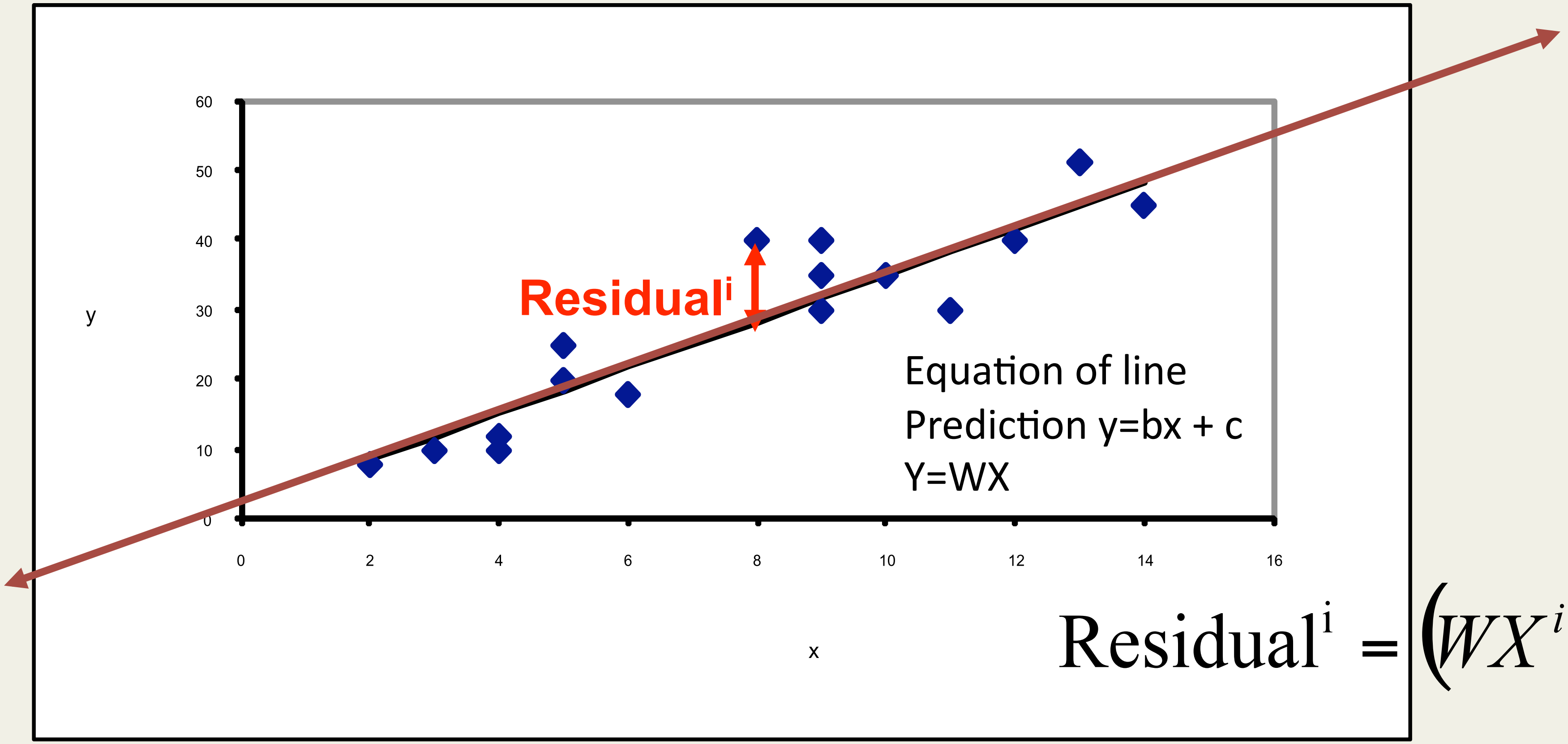
- **Semi-supervised Learning**

  Combines both labeled and unlabeled examples to generate an appropriate function or classifier.

- **Reinforcement Learning**

  Learns how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback in the form of rewards that guides the learning algorithm.

Linear Regression: Minimize Residuals

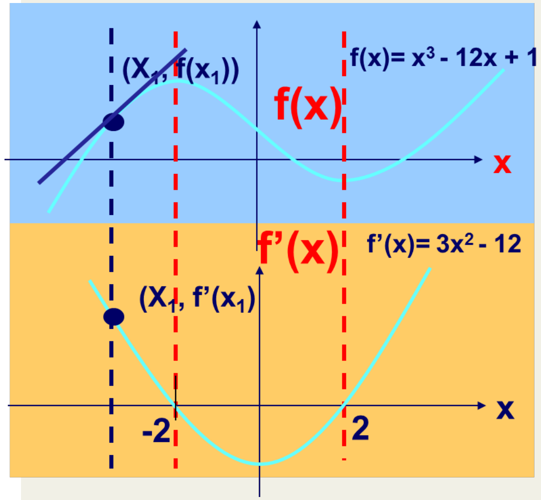


Equation of line
Prediction y=bx + c
Y=WX

$$\mathrm{Residual}^{i} = \left(WX^{i} - y^{i}\right)$$

$$J(W) = \tfrac{1}{2}\sum_{i=1}^{m}\left(WX^{i} - y^{i}\right)^{2} = \tfrac{1}{2}\sum_{i=1}^{m}\left(\mathrm{Re}\,sidual^{i}\right)^{2}$$

Gradient Descent (a simpler root finder)



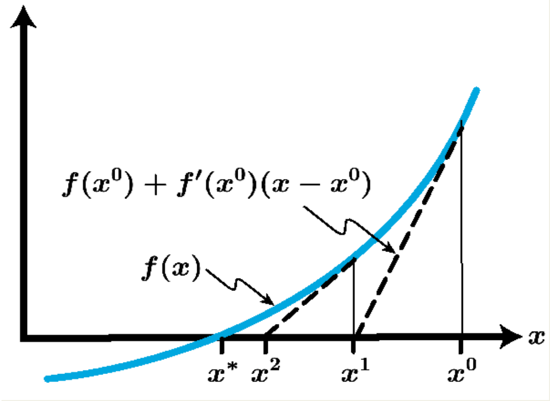
$$x^{i+1} = x^i - \frac{f'(x^i)}{f''(x^i)} \qquad \textit{Iteration function}$$

Newton-Raphson in 1-Dimension

Calculating f''(x), the Hessian H, and inverting it is complex so simpler algorithms have been developed such gradient descent

$$x^{i+1} = x^i - a^i f'(x^i)$$

Gradient Descent



How large should I step in the positive gradient direction (gradient ascent)

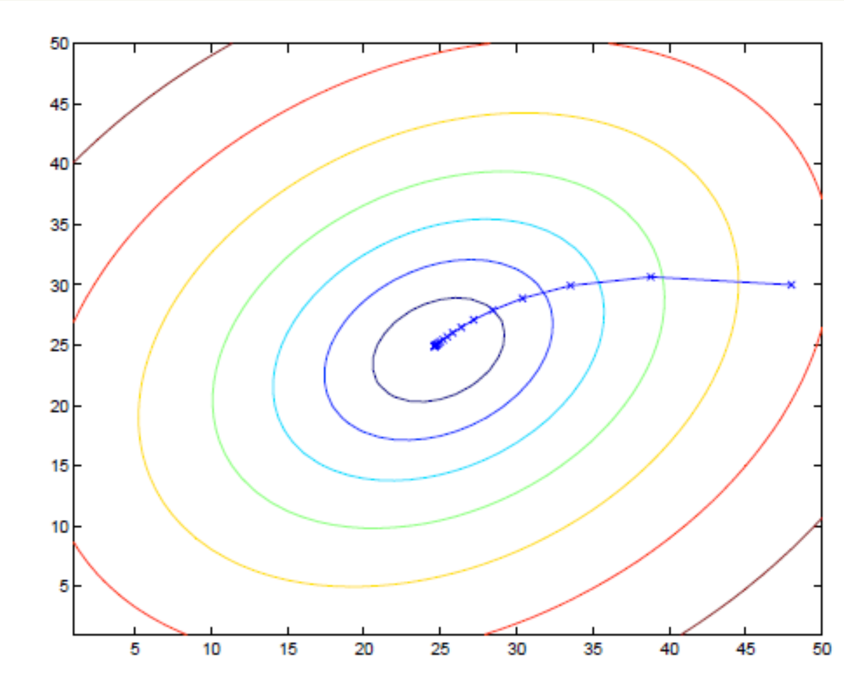
- or in the negative gradient direction (gradient descent)

Gradient Descent: surf downhill

- **Goal: Choose W so as to minimize J(W)**
- **Algorithm**
  - Start with some random guess for W
  - Repeat
    - Use gradient to travel downhill
      - Update each weight w_i
  - Until convergence (to global minimum)

$$J(W) = \tfrac{1}{2}\sum_{i=1}^{m}\left(W^i X^i - y^i\right)^2$$

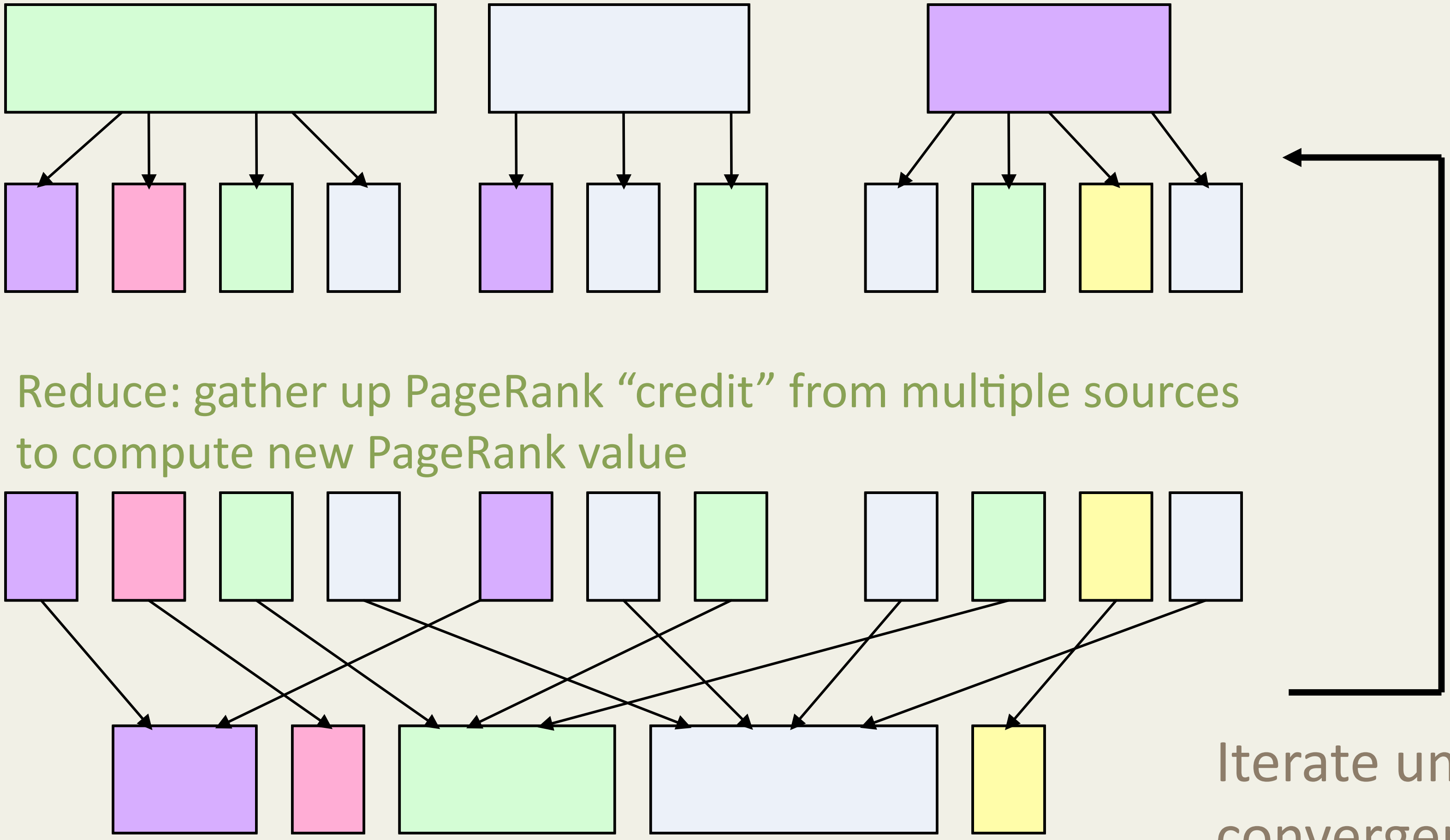
Contour Map of J(W)



```
Let  W = (0,0)
Repeat
    $W_{t+1} = W_t - \alpha * \nabla J(W_t)$
until convergence (i.e., no big changes in W or error)
```

PageRank in MapReduce

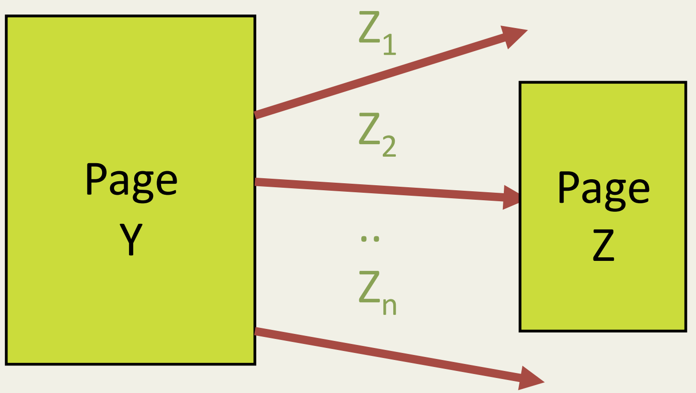
Map: distribute PageRank “credit” to link targets



Reduce: gather up PageRank “credit” from multiple sources to compute new PageRank value

Iterate until convergence

PageRank in MapReduce

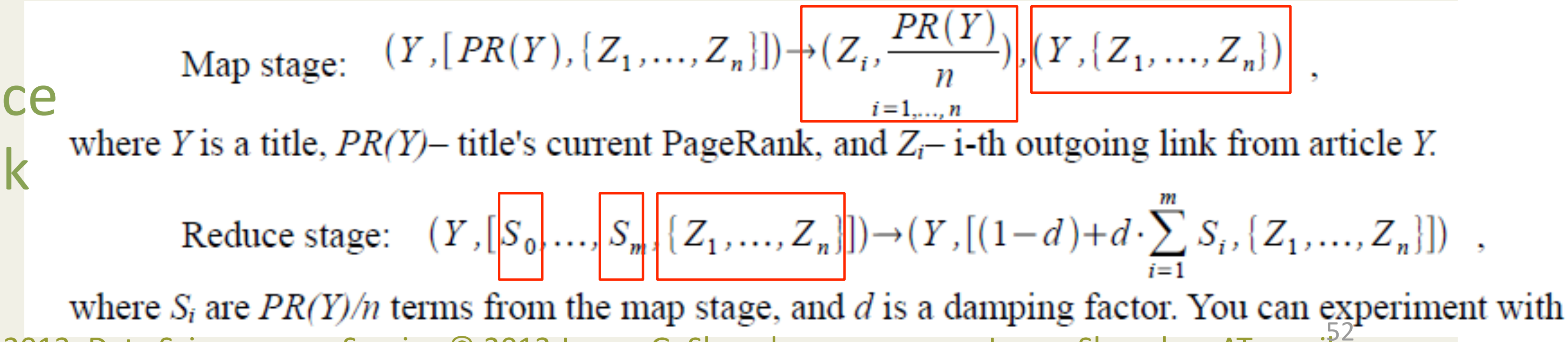


Assume m pages
Leads to m key-value pairs in the
Key=pageID, Value=PageRank(page), OutLinks(Page)
Y, PageRank(Y), Outlinks(Page)
Repeat
  Map: For each page: Spread PageRank of current page to outlinks
  Reduce: accumulate PageRank from inlinks
Until Convergence

Input

Key	Value has two parts	
PageID(i.e., Y)	PageRank(Y)	Outlinks(Y) [Z_1, Z_2, …Z_n]
Y_1	PR(Y_1)	[Z_{11}, Z_{12}, …Z_{1n}]
……	….	……
Y_m	PR(Y_m)	[Z_{m1}, Z_{m2}, …Z_{mn}]

MapReduce PageRank



Map stage: $(Y,[PR(Y),\{Z_1,\dots,Z_n\}])\rightarrow\boxed{(Z_i,\frac{PR(Y)}{n})_{i=1,\dots,n}},\boxed{(Y,\{Z_1,\dots,Z_n\})}$,

where Y is a title, $PR(Y)$– title's current PageRank, and Z_i– i-th outgoing link from article Y.

Reduce stage: $(Y,[\boxed{S_0},\dots,\boxed{S_m},\boxed{\{Z_1,\dots,Z_n\}}])\rightarrow(Y,[(1-d)+d\cdot\sum_{i=1}^{m}S_i,\{Z_1,\dots,Z_n\}])$,

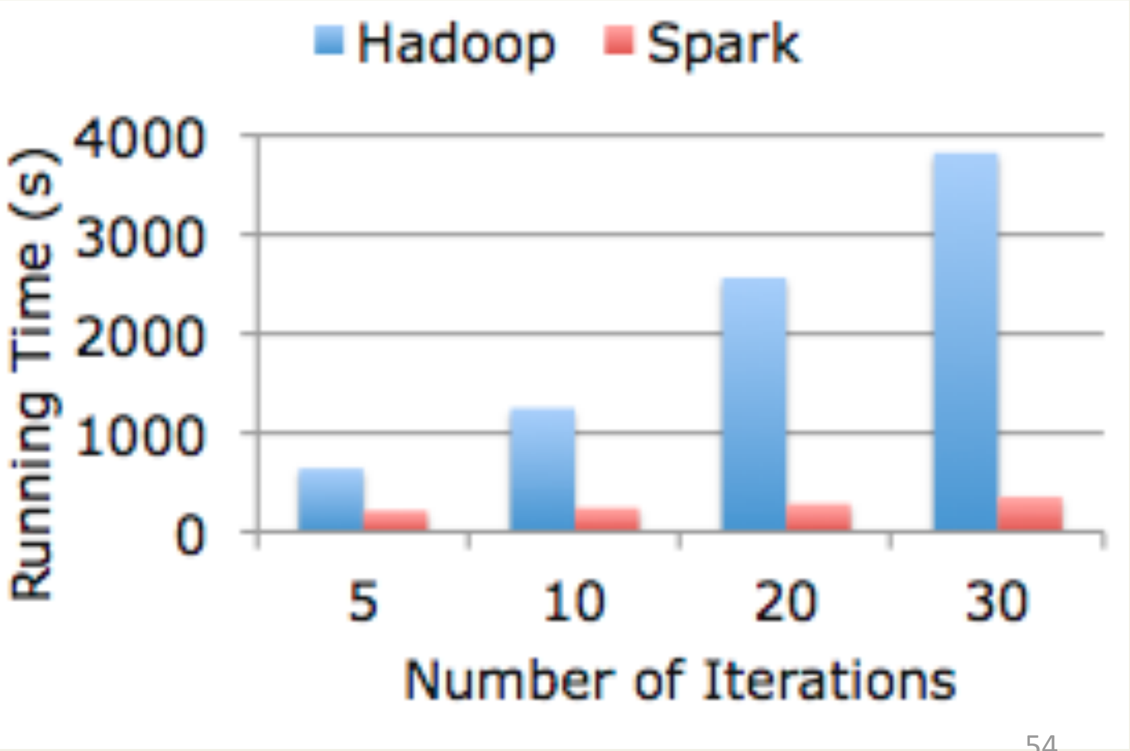
where S_i are $PR(Y)/n$ terms from the map stage, and d is a damping factor. You can experiment with

Opensource Predictive Platforms

- **Mahout**
  - Collaborative Filtering, Clustering (K-Means, Fuzzy K-Means clustering), Latent Dirichlet Allocation, Singular value decomposition, Parallel Frequent Pattern mining, Naive Bayes classifier, Random forest decision tree based classifier, datasets
  - A vibrant community
- **Spark (enables faster ML) (UC Berkeley)**
  - Spark is an open source cluster computing framework that can outperform Hadoop by 30x through a combination of in-memory computation and a richer execution engine.
- **SHARK (UC Berkeley)**
  - Speeded up SQL analysis on Hadoop
- **R**
  - Is an open source programming language and software environment for statistical computing and graphics.

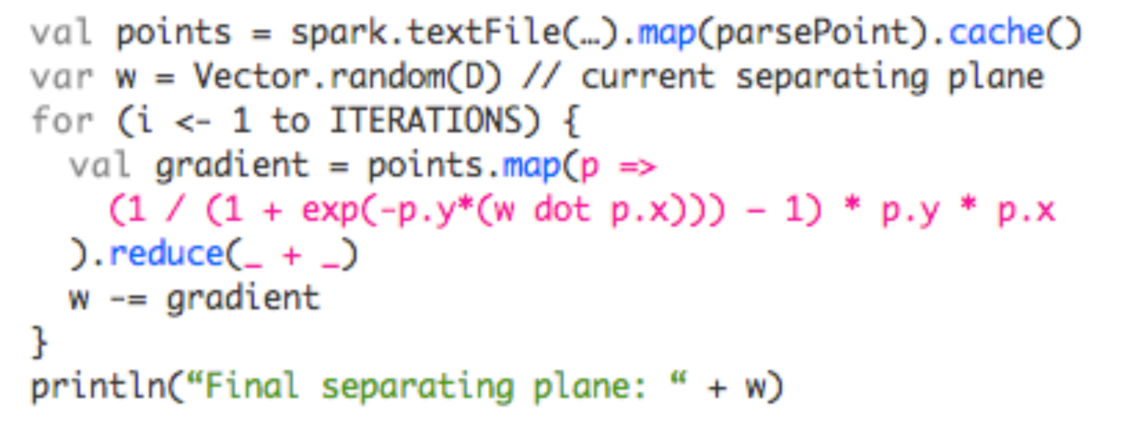
Logistic Regression in Spark rocks!

This is an iterative machine learning algorithm that seeks to find the best hyperplane that separates two sets of points in a multi-dimensional feature space. It can be used to classify messages into spam vs non-spam, for example. Because the algorithm applies the same MapReduce operation repeatedly to the same dataset, it benefits greatly from caching the input data in RAM across iterations.



Logistic Regression in Spark rocks!

This is an iterative machine learning algorithm that seeks to find the best hyperplane that separates two sets of points in a multi-dimensional feature space. It can be used to classify messages into spam vs non-spam, for example. Because the algorithm applies the same MapReduce operation repeatedly to the same dataset, it benefits greatly from caching the input data in RAM across iterations.

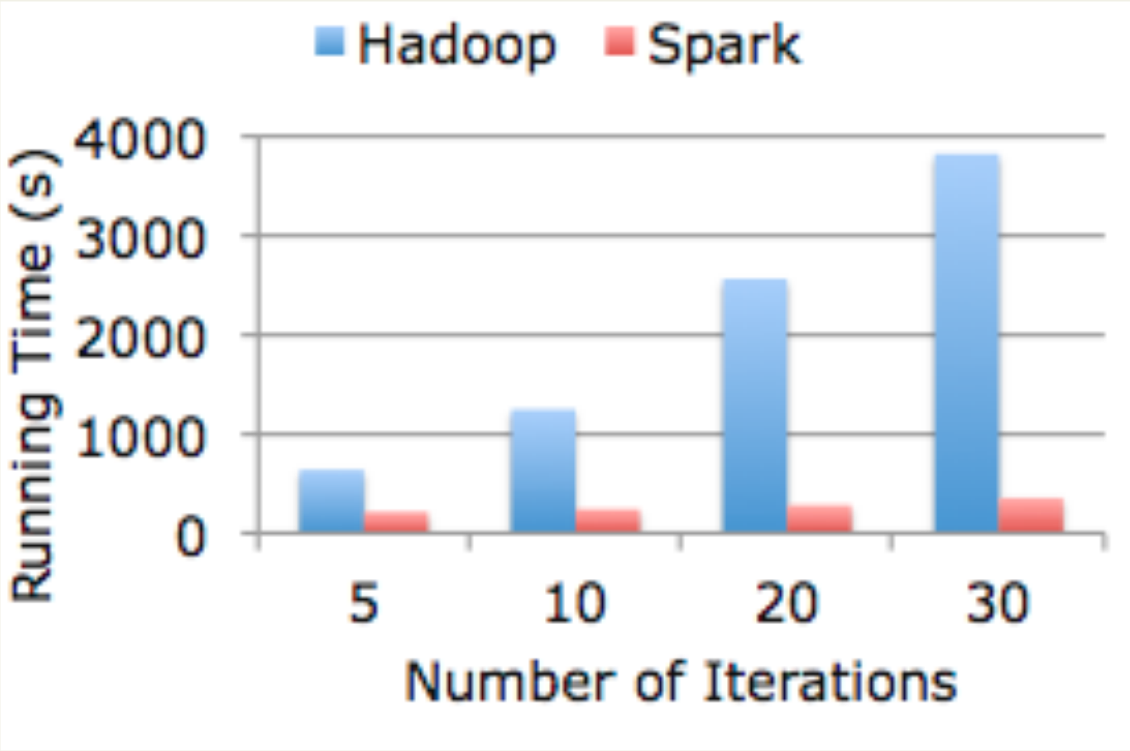


```
val points = spark.textFile(…).map(parsePoint).cache()
var w = Vector.random(D) // current separating plane
for (i <- 1 to ITERATIONS) {
  val gradient = points.map(p =>
    (1 / (1 + exp(-p.y*(w dot p.x))) - 1) * p.y * p.x
  ).reduce(_ + _)
  w -= gradient
}
println("Final separating plane: " + w)
```

Note that w gets shipped automatically to the cluster with every map call.

Logistic Regression in Spark rocks!

The graph below compares the performance of this Spark program against a Hadoop implementation on 30 GB of data on an 80-core cluster, showing the benefit of in-memory caching:



X O N Δ partners

More Data versus Rocket Science

Some simple math using a mountain of data can get you 80% of the way!

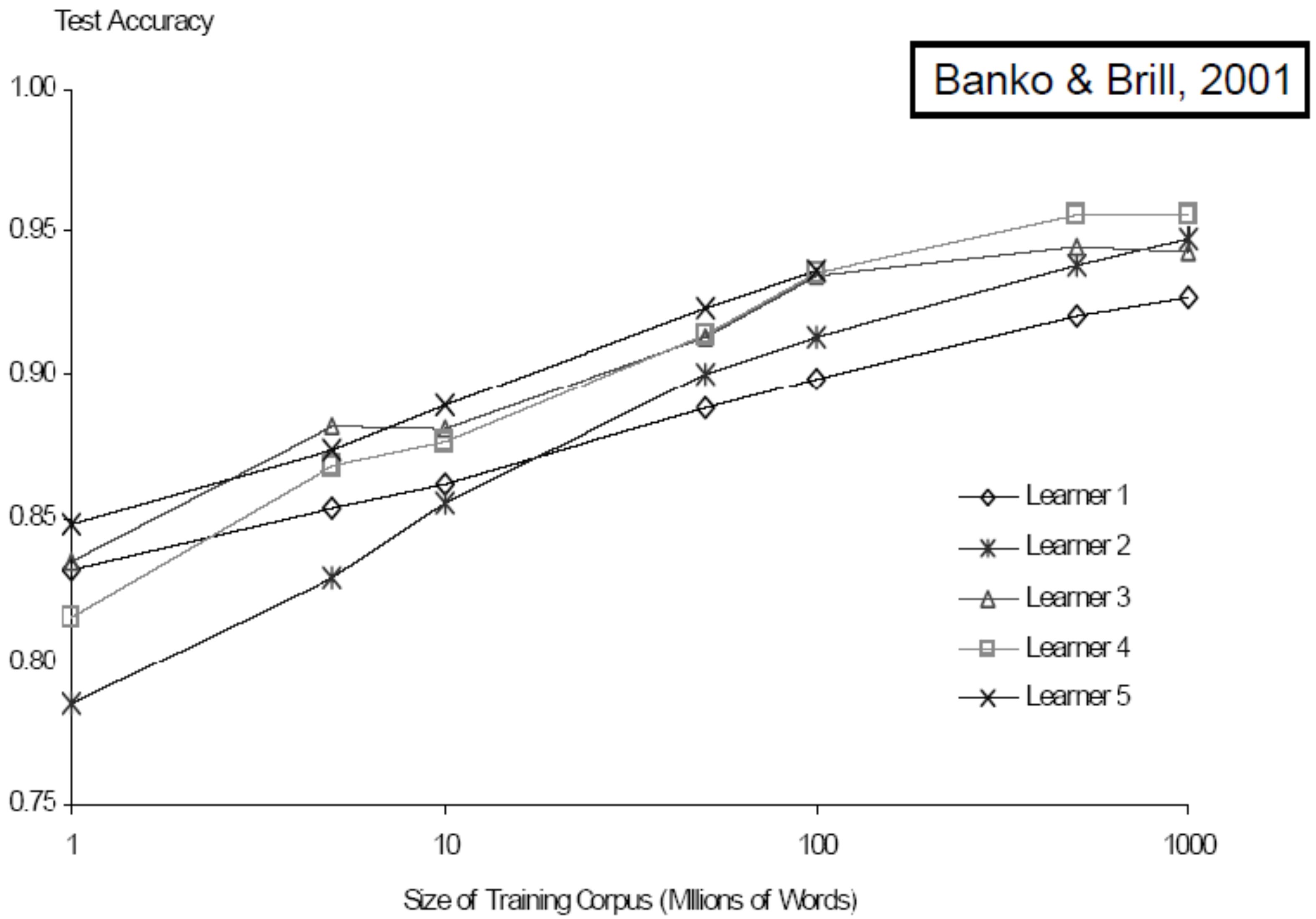


Figure 2. Learning Curves for Confusable Disambiguation

HERITAGE PROVIDER NETWORK
HEALTH PRIZE

- A competition, powered by Kaggle
- De-identified dataset containing medical records of 100,000 Americans
- $3 million prize

http://www.heritagehealthprize.com

NETFLIX

NetFlix Prize

2006 – 2009

$1 million prize

50,000 registrations

HERITAGE PROVIDER NETWORK
HEALTH PRIZE

2011

$3 million prize

Projected 100,000 registrations

UCSC TIM 209 (Machine Learning, Data Mining)
Had my students participate; one student got to rank 52!

Conclusions

○ **Data is disruptive**

○ **Data is an enabler**

○ **Data science is a discipline that organizations will need to embrace**

○ **DsaaS is still embryonic as is the field itself**

○ **Grow your own DS stack (lots of opensouce systems out there)**

○ **Core principles and skills matter not the hype**

○ **Closing thought**

> *"We don't have better algorithms. We just have more data"*
> [Peter Norvig, Google]