Big Data in B2C

Jean-Paul Schmetz
Chief Scientist
Hubert Burda Media

Hubert Burda Media

What is Big Data?

- mochi-wa mochi-ya – if you want mochi, go to the mochi maker

What is Big Data?

- mochi-wa mochi-ya – if you want mochi, go to the mochi maker
- Stanford is amongst the best big data people maker in the world (brought us the people that created things like Yahoo, Google, Netflix, Instagram, Pulse, …)

What is Big Data?

- mochi-wa mochi-ya – if you want mochi, go to the mochi maker
- Stanford is amongst the best big data people maker in the world (brought us the people that created things like Yahoo, Google, Netflix, Instagram, Pulse, …)
- They have a “big data” program that they call “Mining Massive Data Sets” – the name is a bit older, it started in the 90s

Mining Massive Data Sets

- It’s not about data mining as we normally see

Mining Massive Data Sets

- It’s not about data mining as we normally see
- It’s not about creating visualization of data and making nice presentations that lead to decision

Mining Massive Data Sets

- It’s not about data mining as we normally see
- It’s not about creating visualization of data and making nice presentations that lead to decision
- It’s about making products specifically turning data into software

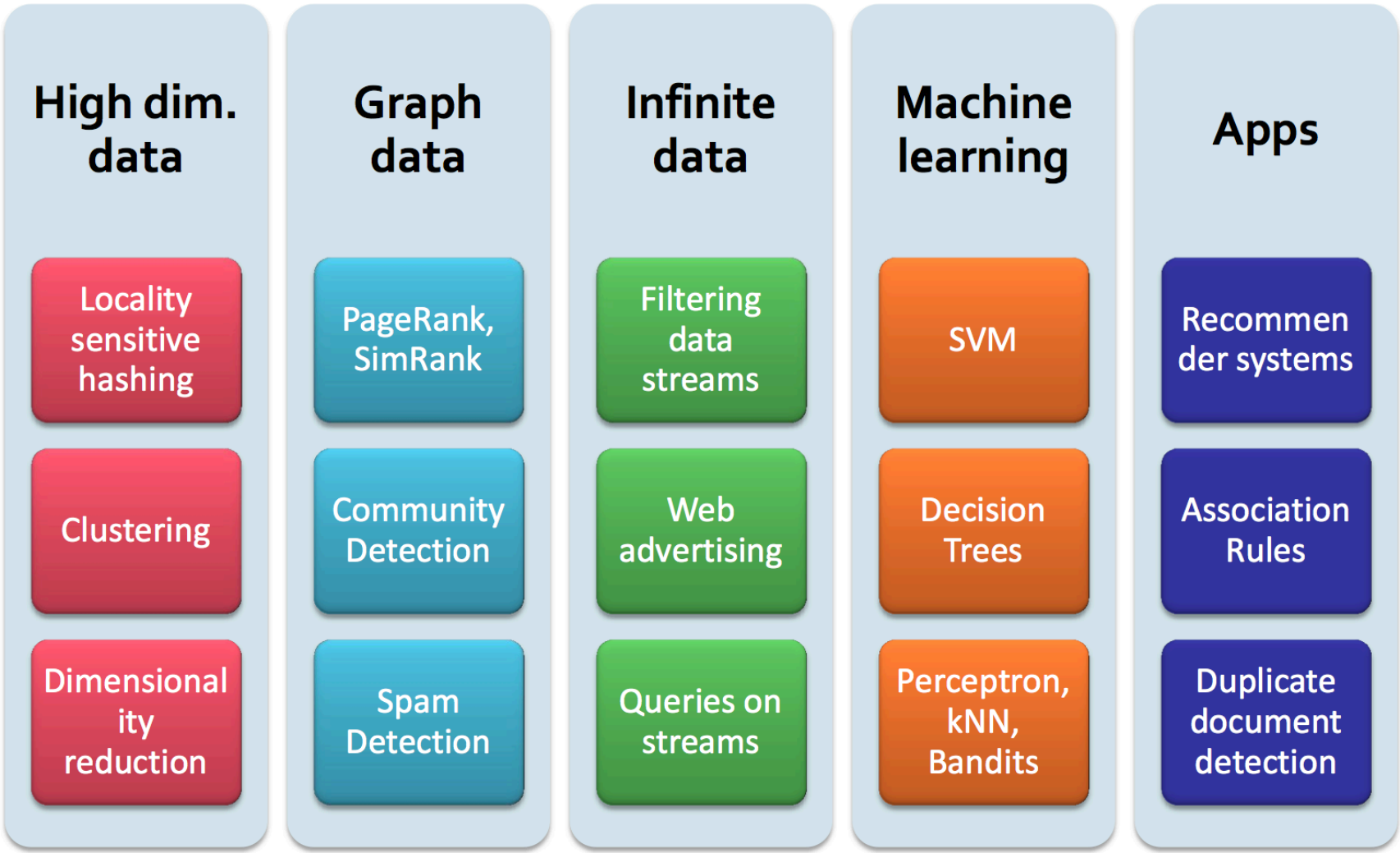
Mining Massive Data Sets

- It's not about data mining as we normally see
- It's not about creating visualization of data and making nice presentations that lead to decision
- It's about making products specifically turning data into software
- It is taught by people who brought you
  - Google
  - the algorithm that recommends friends in Facebook

Mining Massive Data Sets

- Mining Massive Data Sets:
  - The mechanics of big data
  - An overview of all the methods to handle massive size

How It All Fits Together



Mining Massive Data Sets

- Mining Massive Data Sets:
  - The mechanics of big data
  - An overview of all the methods to handle massive size
- Machine Learning:
  - Turning data into software

Mining Massive Data Sets

- Mining Massive Data Sets:
  - The mechanics of big data
  - An overview of all the methods to handle massive size
- Machine Learning:
  - Turning data into software
- Search:
  - Interpreting users' needs/intent and finding results

Mining Massive Data Sets

- Mining Massive Data Sets:
  - The mechanics of big data
  - An overview of all the methods to handle massive size
- Machine Learning:
  - Turning data into software
- Search:
  - Interpreting users' needs/intent and finding results
- Information Network Analysis:
  - Dealing with data in Networks

Why does it matter to us?

- It does matter because:
  - It enables to better understand users and provides them better content

Why does it matter to us?

- It does matter because:
  - It enables to better understand users and provides them better content
  - It enables more effective advertising (one of the biggest methods – adwords - is now bigger than the whole print industry in the US)

Why does it matter to us?

- It does matter because:
  - It enables to better understand users and provides them better content
  - It enables more effective advertising (one of the biggest methods – adwords - is now bigger than the whole print industry in the US)
  - Their applications capture the time of people (e.g. the content of facebook is highly addictive to 100s million of people)

Magazines and Big Data

- Magazine making does not and should not really change with big data

Magazines and Big Data

- Magazine making does not and should not really change with big data
- However “big data” can throw serious curve balls at magazines making some formats simply ineffective in the future (some examples to follow)

Magazines and Big Data

- Magazine making does not and should not really change with big data
- However “big data” can throw serious curve balls at magazines making some formats simply ineffective in the future (some examples to follow)
- At Burda, all of digital is affected and is “big data’ through and through (i.e. we are successful where we are good at it and lose where we are not)

Holiday vs. HolidayCheck

- We had a magazine called Holiday (1990s)

Holiday vs. HolidayCheck

- We had a magazine called Holiday (1990s)
- Then we had a website for it (1996-2003)

Holiday vs. HolidayCheck

- We had a magazine called Holiday (1990s)
- Then we had a website for it (1996-2003)
- Then we killed the magazine (late 1990s)

Holiday vs. HolidayCheck

- We had a magazine called Holiday (1990s)
- Then we had a website for it (1996-2003)
- Then we killed the magazine (late 1990s)
- Then we killed the website

Holiday vs. HolidayCheck

- We had a magazine called Holiday (1990s)
- Then we had a website for it (1996-2003)
- Then we killed the magazine (late 1990s)
- Then we killed the website
- Then we started to make serious money and delight users
  - Because we looked at the problem from the users perspective (quality and price)
  - Because we did not shy away from size (millions of reviews)

Abebooks/LibraryThing

- This one did not have a magazine behind it

Abebooks/LibraryThing

- This one did not have a magazine behind it
- But it became a “cultural good” in Canada

Abebooks/LibraryThing

- This one did not have a magazine behind it
- But it became a “cultural good” in Canada
- Big data user delight
  - We had 100s million of books “in store”
  - We were making millions of customers happy
  - Library thing is the social network for book lovers

Abebooks/LibraryThing

- This one did not have a magazine behind it
- But it became a “cultural good” in Canada
- Big data user delight
  - We had 100s million of books “in store”
  - We were making millions of customers happy
  - Library thing is the social network for book lovers
- Uniquely about big data: we know where the books are and we know who reads what book

XING/LinkedIn vs. Business Press

- LinkedIn has pretty much become the number 1 referrer in the “serious” content category

XING/LinkedIn vs. Business Press

- LinkedIn has pretty much become the number 1 referrer in the “serious” content category
- LinkedIn and XING will also become the number 1 creator in these categories

XING/LinkedIn vs. Business Press

- LinkedIn has pretty much become the number 1 referrer in the “serious” content category
- LinkedIn and XING will also become the number 1 creator in these categories
- They are both able to determine the strength of relationship/interest between millions of their members

Netflix etc.. vs. TV guides

- Movie recommendation is one of the example of big data in action

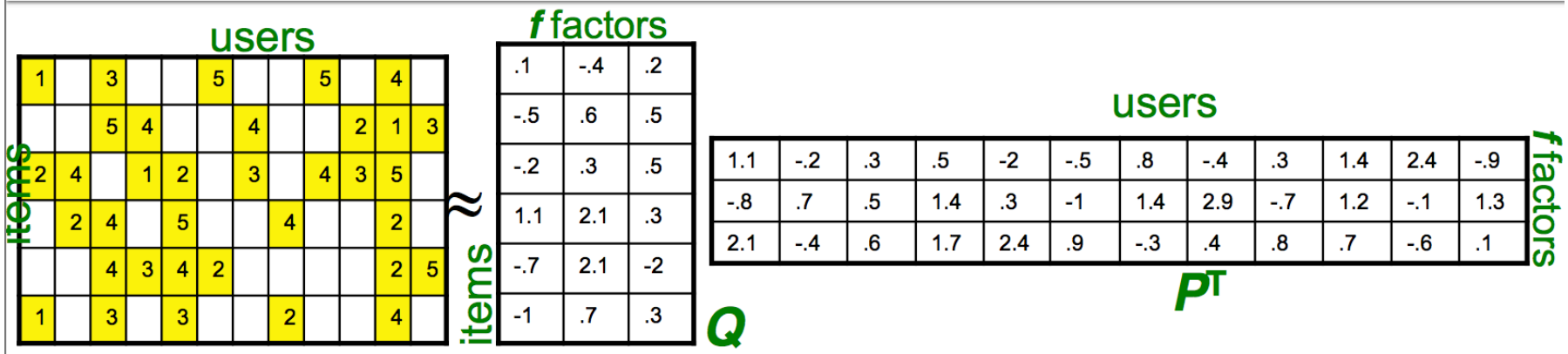
Netflix etc.. vs. TV guides

- Movie recommendation is one of the example of big data in action
- Netflix created the benchmark with their Netflix challenge where 2700 teams competed to make better movie recommendations (by over 10%) and win $1 million dollar

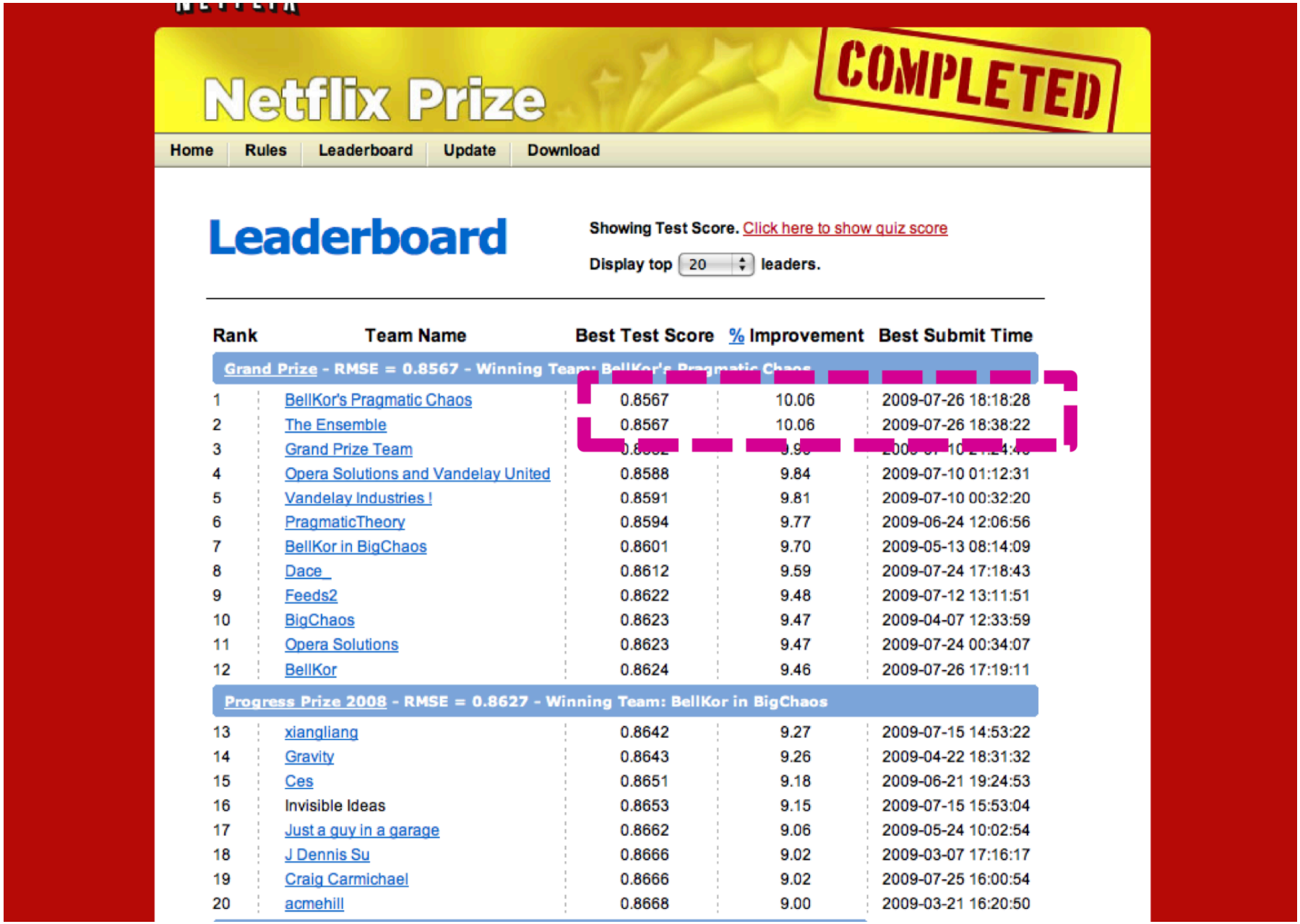
Netflix etc.. vs. TV guides

- Movie recommendation is one of the example of big data in action
- Netflix created the benchmark with their Netflix challenge where 2700 teams competed to make better movie recommendations (by over 10%) and win $1 million dollar
- After years of work, it came down to 20 minutes with a score similar to 4 decimals after the comma

Latent Factor Models



- **SVD isn't defined when entries are missing!**
- **Use specialized methods to find *P*, *Q***
  - $\min_{P,Q} \sum_{(i,x)\in R} (r_{xi} - q_i \cdot p_x^T)^2$
  - $\hat{r}_{xi} = q_i \cdot p_x^T$
  - **Note:**
    - We don't require cols of ***P, Q*** to be orthogonal/unit length
    - ***P, Q*** map users/movies to a latent space
    - The most popular model among Netflix contestants


Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Download
Leaderboard
Showing Test Score. Click here to show quiz score
Display top 20 leaders.
Rank
Team Name
Best Test Score
% Improvement
Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos
1 BellKor's Pragmatic Chaos 0.8567 10.06 2009-07-26 18:18:28
2 The Ensemble 0.8567 10.06 2009-07-26 18:38:22
3 Grand Prize Team
4 Opera Solutions and Vandelay United 0.8588 9.84 2009-07-10 01:12:31
5 Vandelay Industries ! 0.8591 9.81 2009-07-10 00:32:20
6 PragmaticTheory 0.8594 9.77 2009-06-24 12:06:56
7 BellKor in BigChaos 0.8601 9.70 2009-05-13 08:14:09
8 Dace 0.8612 9.59 2009-07-24 17:18:43
9 Feeds2 0.8622 9.48 2009-07-12 13:11:51
10 BigChaos 0.8623 9.47 2009-04-07 12:33:59
11 Opera Solutions 0.8623 9.47 2009-07-24 00:34:07
12 BellKor 0.8624 9.46 2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos
13 xiangliang 0.8642 9.27 2009-07-15 14:53:22
14 Gravity 0.8643 9.26 2009-04-22 18:31:32
15 Ces 0.8651 9.18 2009-06-21 19:24:53
16 Invisible Ideas 0.8653 9.15 2009-07-15 15:53:04
17 Just a guy in a garage 0.8662 9.06 2009-05-24 10:02:54
18 J Dennis Su 0.8666 9.02 2009-03-07 17:16:17
19 Craig Carmichael 0.8666 9.02 2009-07-25 16:00:54
20 acmehill 0.8668 9.00 2009-03-21 16:20:50

Pinterest/Facebook/Instagram vs. Women Magazines

- Bottomline is: Facebook manages to create 1 billion newsfeed every minute that captures the attention of many 100s of million of people for an average length of 7 hours and 46 minutes per month. Beats us by a mile

Pinterest/Facebook/Instagram vs. Women Magazines

- Bottomline is: Facebook manages to create 1 billion newsfeed every minute that captures the attention of many 100s of million of people for an average length of 7 hours and 46 minutes per month.
- Because of their beacons everywhere (like buttons) these services (also Twitter) managed to know you before you join them.

E-commerce vs. publishing

- In some of our magazines (CHIP e.g.) we start to consider Amazon to be the #1 competitor

E-commerce vs. publishing

- In some of our magazines (CHIP e.g.) we start to consider Amazon to be the #1 competitor
- We do see an increase of “publishing”-like activity in many e-commerce firms.

E-commerce vs. publishing

- In some of our magazines (CHIP e.g.) we start to consider Amazon to be the #1 competitor
- We do see an increase of “publishing”-like activity in many e-commerce firms.
- The reason is deeper than naïvity allows:
  - They do pride themselves in objectivity – most of them see themselves as working on behalf of the customer
  - They do know more: if you have 100s of people who have bought something and (maybe) reviewed it or returned it – you know

Search vs. everything

- Search is where “big data” is most used. All the techniques are present at every step:
  - Intent classification
  - Spell correction, query expansion
  - Indexing and core search
  - Ranking
  - SERP
  - Advertising

Search vs. everything

- Search is where “big data” is most used. All the techniques are present at every step:
  - Intent classification
  - Spell correction, query expansion
  - Indexing and core search
  - Ranking
  - SERP
  - Advertising
- If there is one company in the world who does “big data”, it is Google.

Advertising and Big Data

- Adwords and Adsense are now generating more revenue than the whole of print in the US

Advertising and Big Data

- Adwords and Adsense are now generating more revenue than the whole of print in the US
- The core algorithm behind them is based on a few big data technique.
  - It ranks by CTR * bid (= revenue to Google)
  - Sounds simple but CTR is not known and needs to be discovered
  - Exploration vs. Exploitation is a core big data skill

Conclusions

- mochi-wa mochi-ya – make sure you have some people who went to a big data "maker"
- Big data is really about products not presentations
- While big data will not change magazines it will throw serious curve balls at them
- Your digital strategy is and must be centered around big data/search