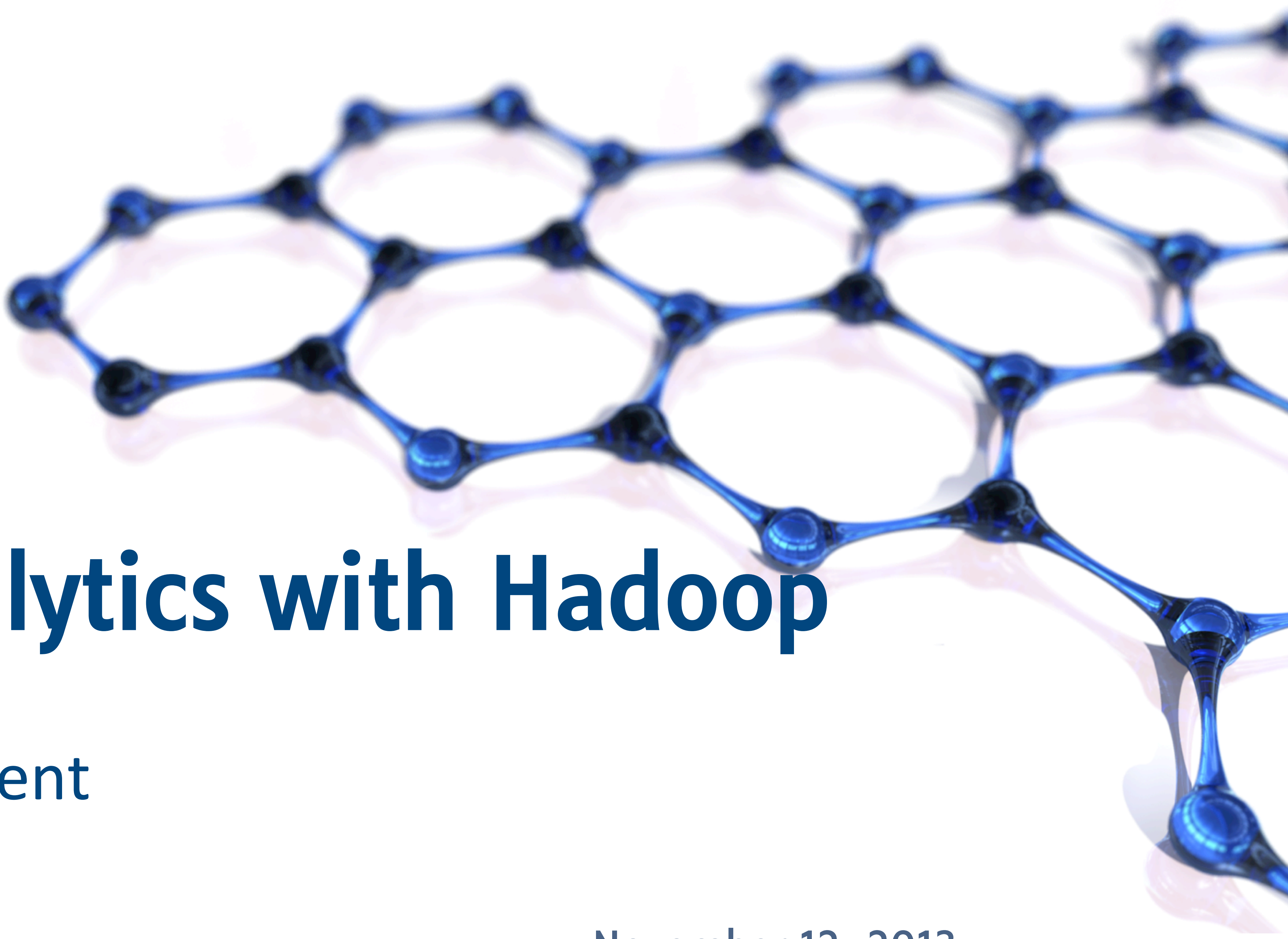


Predictive Analytics with Hadoop

Tomer Shiran
VP Product Management
MapR Technologies

November 12, 2013

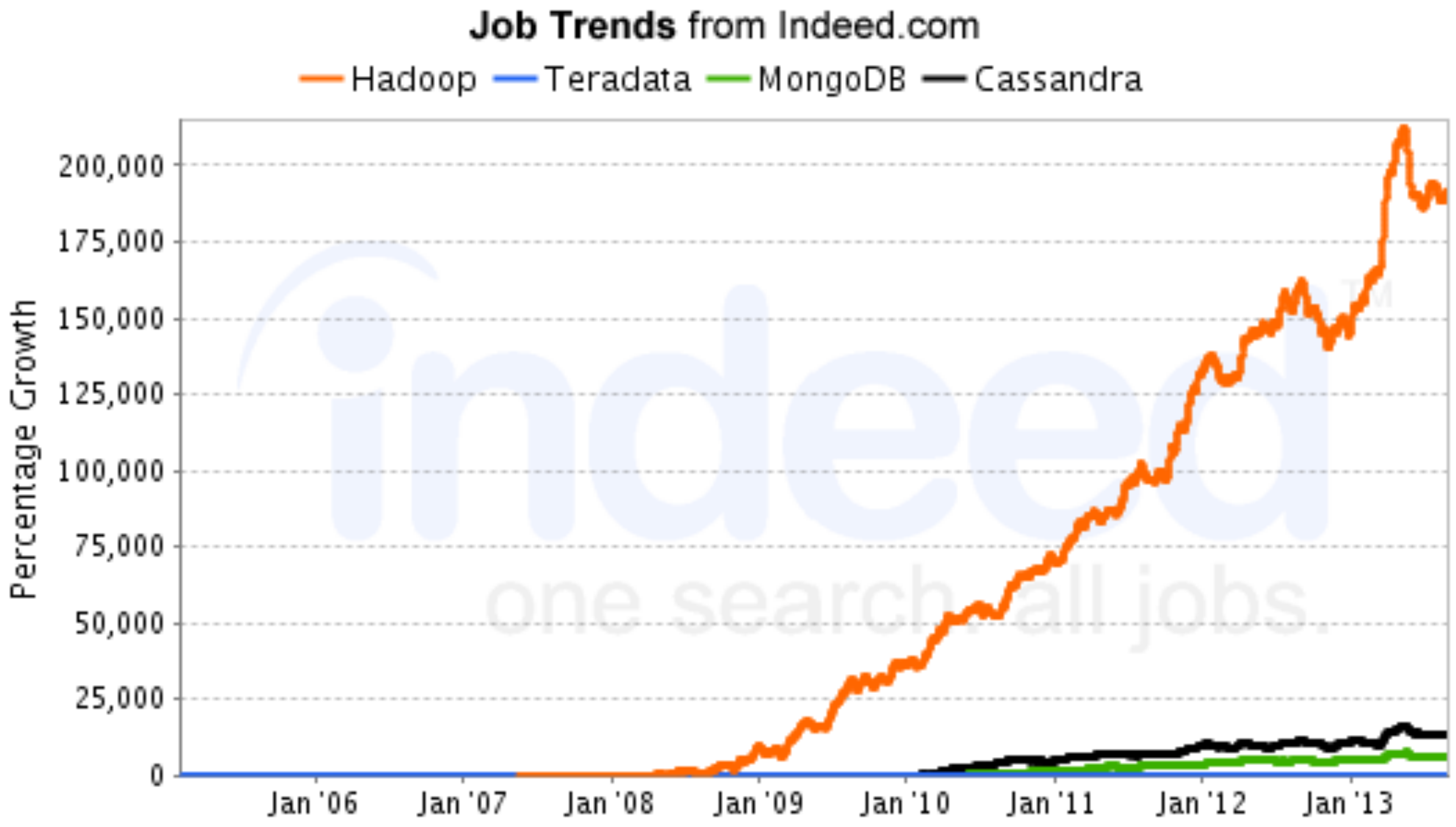
Me, Us

- Tomer Shiran
  - VP Product Management, MapR Technologies
  - tshiran@maprtech.com

- MapR
  - Enterprise-grade Hadoop distribution
  - Apache Hadoop + infrastructure and management innovation
  - > 500 paying customers
  - EMEA offices in UK, Germany, Sweden and France (HQ in London)

- Twitter: #mapr

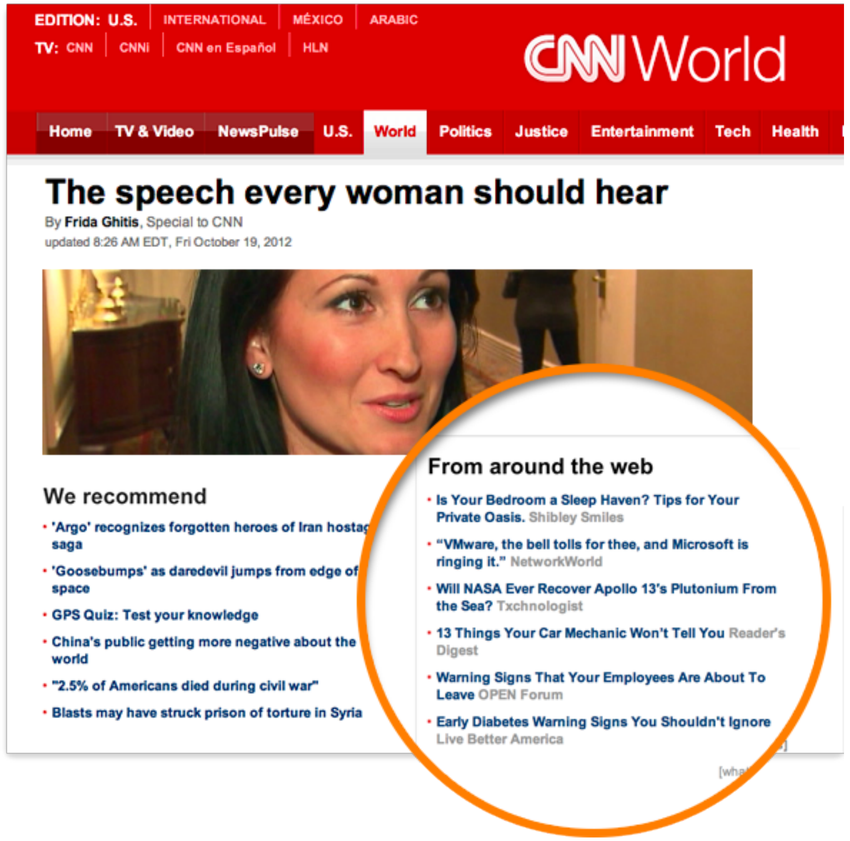
Hadoop Job Growth



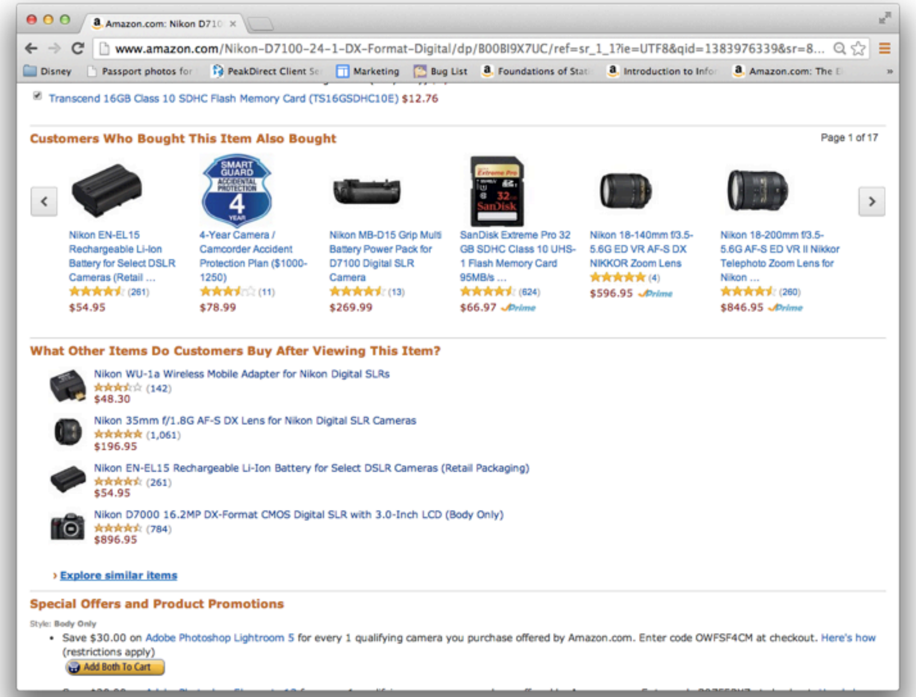
Agenda

- Examples
- Data-driven solutions
- Obtaining <u>big</u> training data
- Recommendation with Mahout and Solr
- Operational considerations

Recommendation is Everywhere



Media and Advertising



e-commerce

- Recommend sales opportunities to partners
- $40M revenue in year 1
- 1.5B records per day
- Using MapR

Enterprise Sales

Classification is Everywhere

Return Path

- 600+ variables considered for every IP address
- Billions of data points
- Using MapR

IP address blacklisting

ZIONS BANK®

- Identify anomalous patterns indicating fraud, theft and criminal activity
- Stop phishing attempts
- Using MapR

Fraud Detection

Fortune 100 Telco

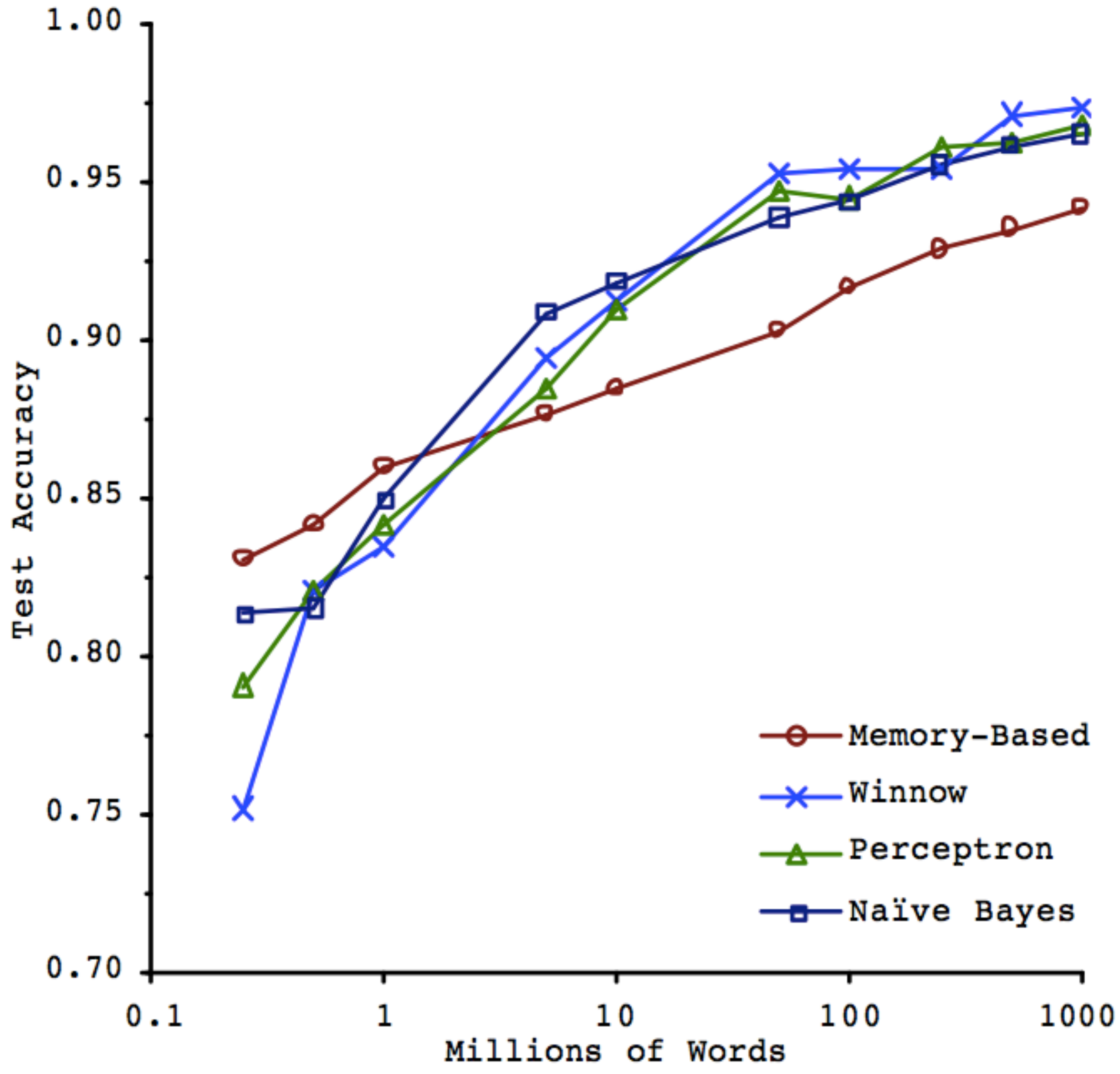
- Customer 360 application
- Each customer is scored and categorized based on all their activity
- Data from hundreds of streams and databases
- Using MapR

Customer 360
Scoring & Categorization

Data-Driven Solutions

- Physics is simple: $f = ma$; $E=mc^2$
- Human behavior is much more complex
  - Which ad will they click?
  - Is a behavior fraudulent? Why?
- Don't look for complex models that try to discover general rules
  - The size of the dataset is the most important factor
  - Simple models (n-gram, linear classifiers) with Big Data
- A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2):8-12, 2009.

The Algorithms Are Less Important

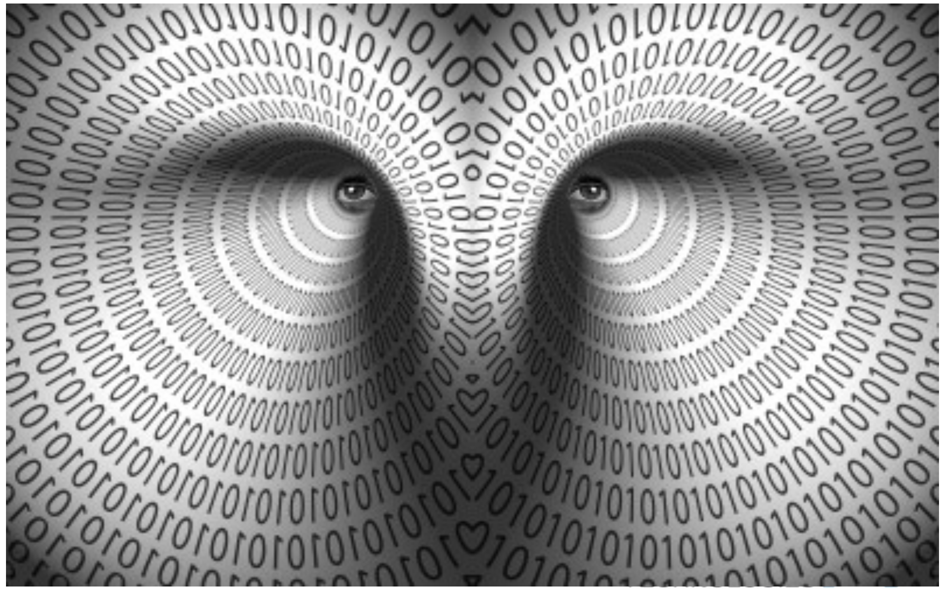


Focus on the Data

- Most algorithms come down to counting and simple math

- Invest your time where you can make a difference
  - Getting more data can improve results by 2x
    - eg, add beacons everywhere to instrument user behavior
  - Tweaking an ML algorithm will yield a fraction of 1%

- Data wrangling
  - Feature engineering
  - Moving data around
  - …



Obtaining Big Training Data

- Can’t really rely on experts to label the data
  - Doesn’t scale (not enough experts out there)
  - Too expensive

- So how do you get the training data?
  - Crowdsourcing
  - Implicit feedback
    - “Obvious” features
    - User engagement

MAPR TECHNOLOGIES

Using Crowdsourcing for Annotation

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. EMNLP, 2008.

Task	Labels	Cost (USD)	Time (hrs)	Labels per USD	Labels per hr
Affect	7000	$2.00	5.93	3500	1180.4
WSim	300	$0.20	0.174	1500	1724.1
RTE	8000	$8.00	89.3	1000	89.59
Event	4620	$13.86	39.9	333.3	115.85
WSD	1770	$1.76	8.59	1005.7	206.1
Total	21690	25.82	143.9	840.0	150.7

Table 3: Summary of costs for non-expert labels

Emotion	1-Expert	10-NE	k	k-NE
Anger	0.459	0.675	2	0.536
Disgust	0.583	0.746	2	0.627
Fear	0.711	0.689	–	–
Joy	0.596	0.632	7	0.600
Sadness	0.645	0.776	2	0.656
Surprise	0.464	0.496	9	0.481
Valence	0.759	0.844	5	0.803
Avg. Emo.	0.576	0.669	4	0.589
Avg. All	0.603	0.694	4	0.613

Table 2: Average expert and averaged correlation over 10 non-experts on test-set. k is the minimum number of non-experts needed to beat an average expert.

Quantity: $2 for 7000 annotations (leveraging Amazon Mechanical Turk and a "flat world")

Quality: 4 non-experts = 1 expert

Using "Obvious" Features for Annotation

Siah @siah 15 May

95 percent of my money comes from my R and **Hadoop** skills. Only 5 percent from the PhD that I spent 4 years of my life on :) #rstats

Expand

:)

aw @_a_w_ 25 Nov

Started uploading some old #hadoop presentations to @slideshare . Looks like it doesn't like Keynote speaker notes though. :(

Expand

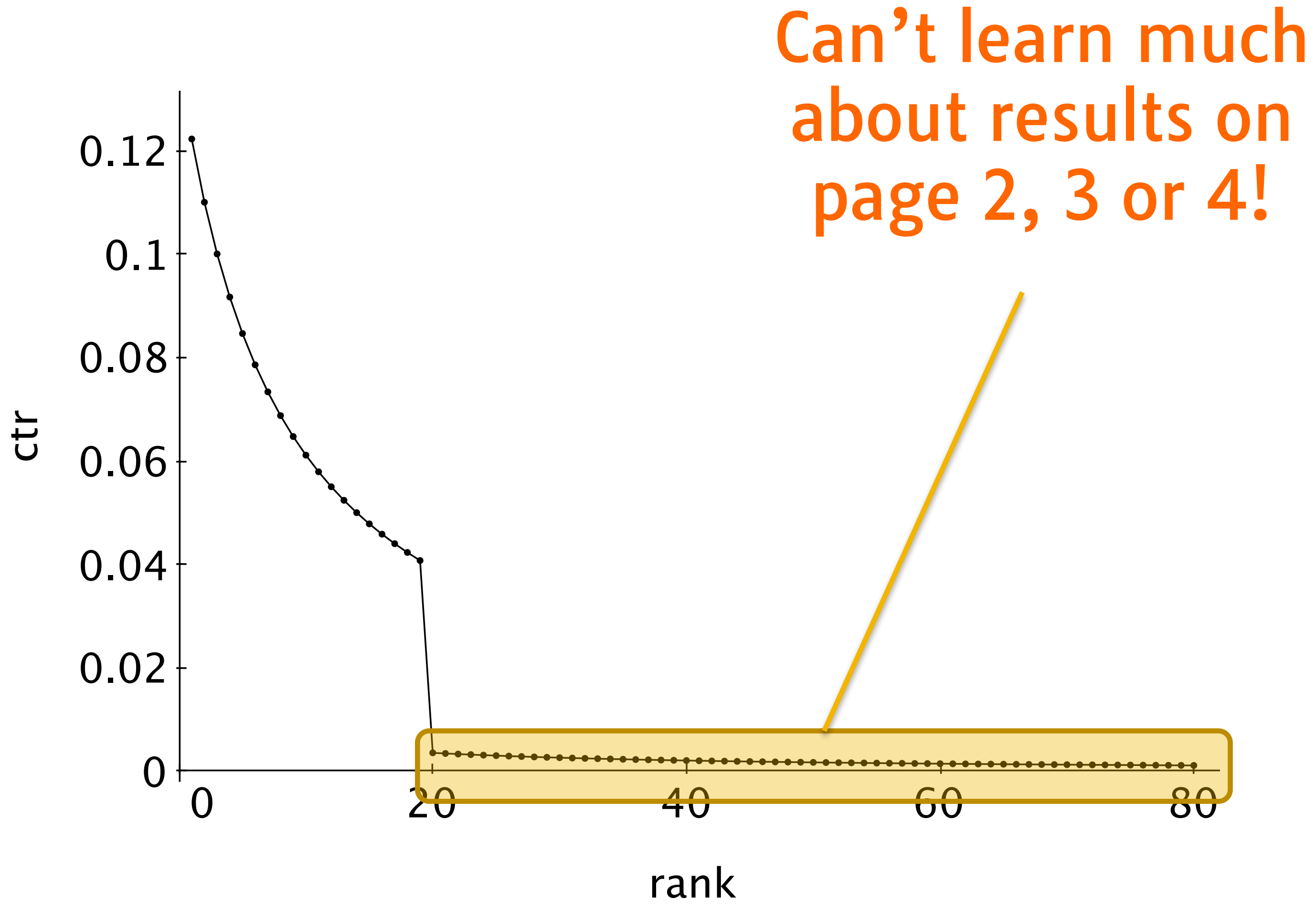
:(

MAPR TECHNOLOGIES

Leveraging Implicit Feedback

- Users behavior provides valuable training data
- Google adjusts search rankings based on engagement
  - Did the user click on the result?
  - Did the user come back to the search page within seconds?
- Most recommendation algorithms are based solely on user activity
  - What products did they view/buy?
  - What ads did they click on?
- T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. ACM TOIS, 25(2):1{27, 2007.

Increasing Exploration



Exploration of the second page

We need to find ways to increase exploration to broaden our learning

Result Dithering

- Dithering is used to re-order recommendation results
  - Re-ordering is done randomly

- Dithering is *guaranteed* to make off-line performance worse

- Dithering also has a near perfect record of making actual performance much better
  - “Made more difference than any other change”

Simple Dithering Algorithm

- Generate synthetic score from log rank plus Gaussian

$$s = \log r + N(0, \varepsilon)$$

- Pick noise scale to provide desired level of mixing

$$\Delta r \propto r \exp \varepsilon$$

- Typically:

$$\varepsilon \in [0.4, 0.8]$$

- Oh… use floor(t/T) as seed so results don't change too often

Example: ε = 0.5

1	2	6	5	3	4	13	16
1	2	3	8	5	7	6	34
1	4	3	2	6	7	11	10
1	2	4	3	15	7	13	19
1	6	2	3	4	16	9	5
1	2	3	5	24	7	17	13
1	2	3	4	6	12	5	14
2	1	3	5	7	6	4	17
4	1	2	7	3	9	8	5
2	1	5	3	4	7	13	6
3	1	5	4	2	7	8	6
2	1	3	4	7	12	17	16

- Each line represents a recommendation of 8 items
- The non-dithered recommendation would be 1, 2, …, 8

MAPR TECHNOLOGIES

Example: ε = log 2 = 0.69

1	2	8	3	9	15	7	6
1	8	14	15	3	2	22	10
1	3	8	2	10	5	7	4
1	2	10	7	3	8	6	14
1	5	33	15	2	9	11	29
1	2	7	3	5	4	19	6
1	3	5	23	9	7	4	2
2	4	11	8	3	1	44	9
2	3	1	4	6	7	8	33
3	4	1	2	10	11	15	14
11	1	2	4	5	7	3	14
1	8	7	3	22	11	2	33

- Each line represents a recommendation of 8 items
- The non-dithered recommendation would be 1, 2, …, 8

Recommendations with Mahout and Solr

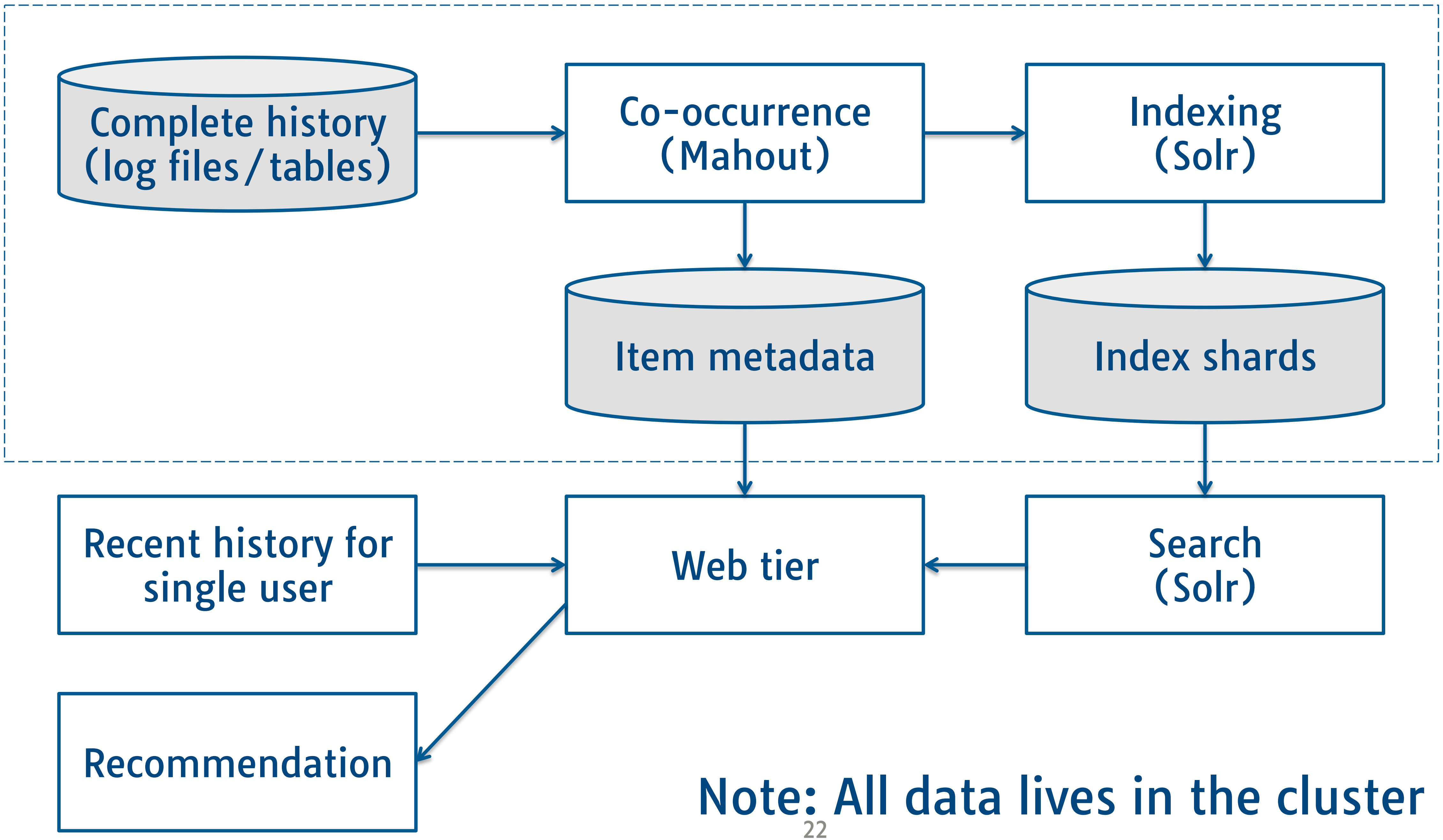
What is Recommendation?

The behavior of a crowd helps us understand what individuals will do...

Batch and Real-Time

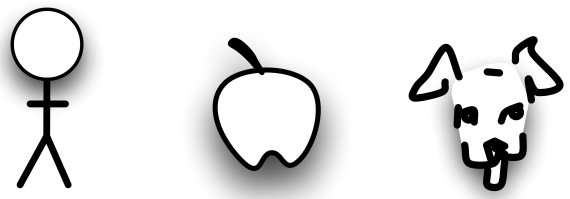
- We can learn about the relationship between items every X hours
  - These relationships don't change often
  - People who buy Nikon D7100 cameras also buy a Nikon EN-EL15 battery
- What to recommend to Bob has to be determined in real-time
  - Bob may be a new user with no history
  - Bob is shopping for a camera right now, but he was buying a baby bottle an hour ago
- How do we do that?
  - Mahout for the heavy number crunching
  - Solr/Elasticsearch for the real-time recommendations

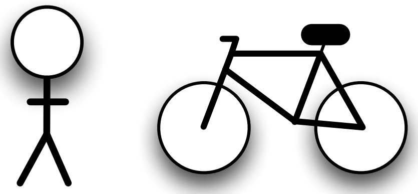
Real-Time Recommender Architecture



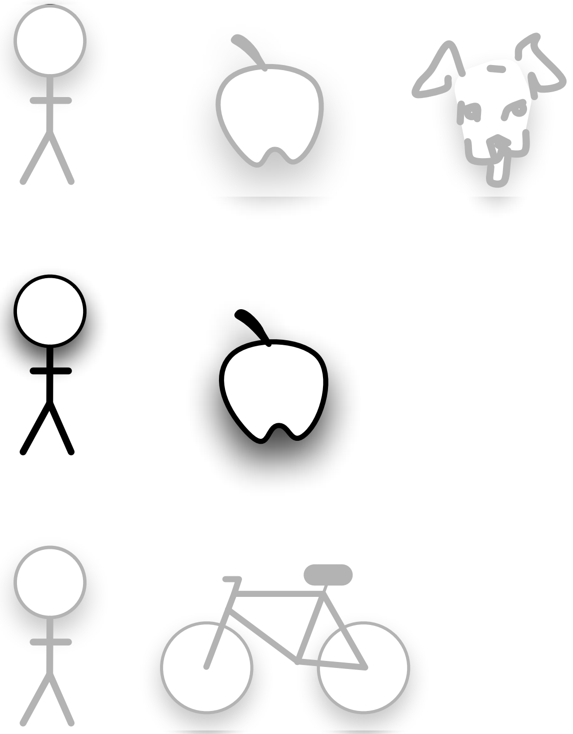
Note: All data lives in the cluster

Recommendations

Alice  Alice got an apple and a puppy

Charles  Charles got a bicycle

Recommendations

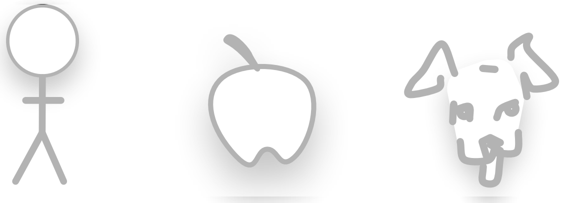


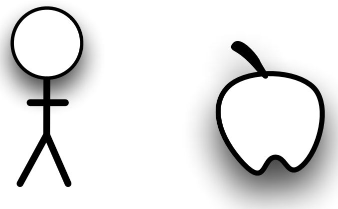
Alice — Alice got an apple and a puppy

Bob — Bob got an apple

Charles — Charles got a bicycle

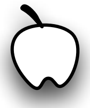
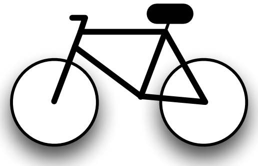
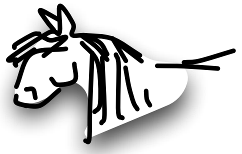
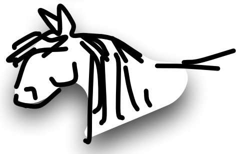
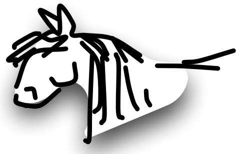
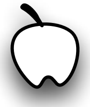
Recommendations

Alice 

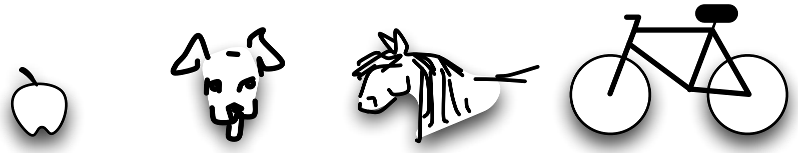
Bob  ? What else would Bob like?

Charles 

Log Files

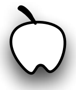
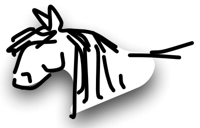
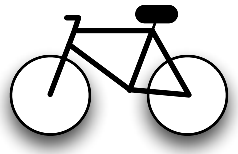
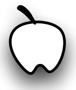
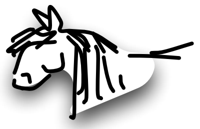
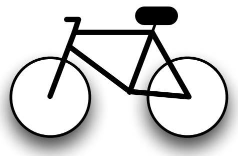
Alice	
Charles	
Charles	
Alice	
Alice	
Bob	
Bob	

History Matrix



Alice	✔	✔	✔	
Bob	✔		✔	
Charles			✔	✔

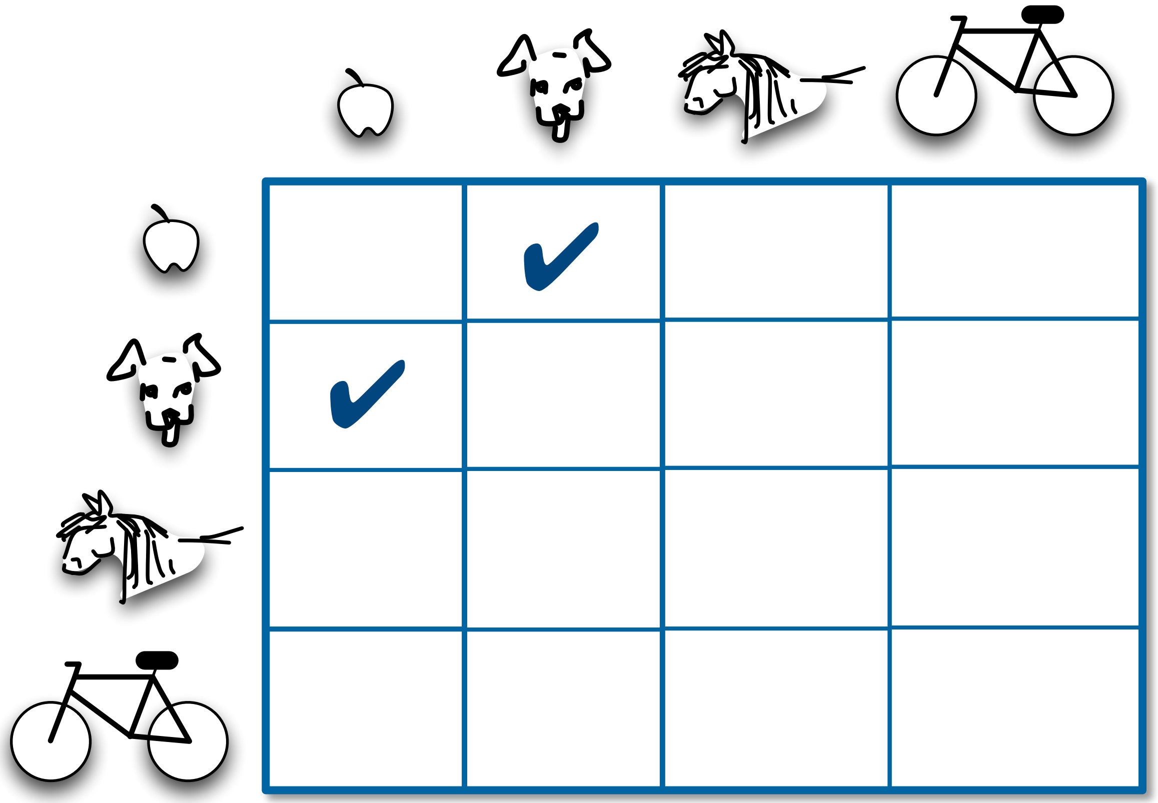
Co-Occurrence Matrix: Items by Items

				
		1	2	0
	1		1	0
	2	1		1
	0	0	1	

Q: How do you tell which co-occurrences are useful?

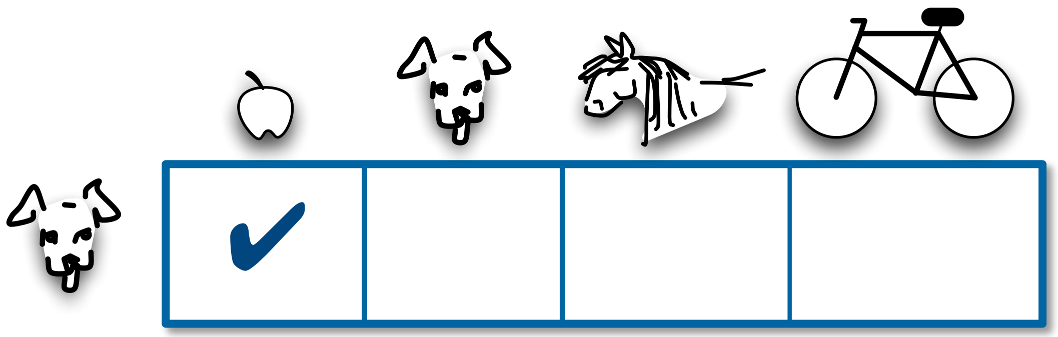
A: Let Mahout do the math...

Indicator Matrix: Anomalous Co-Occurrences



Result: The marked row will be added to the indicator field in the item document…

Indicator Matrix



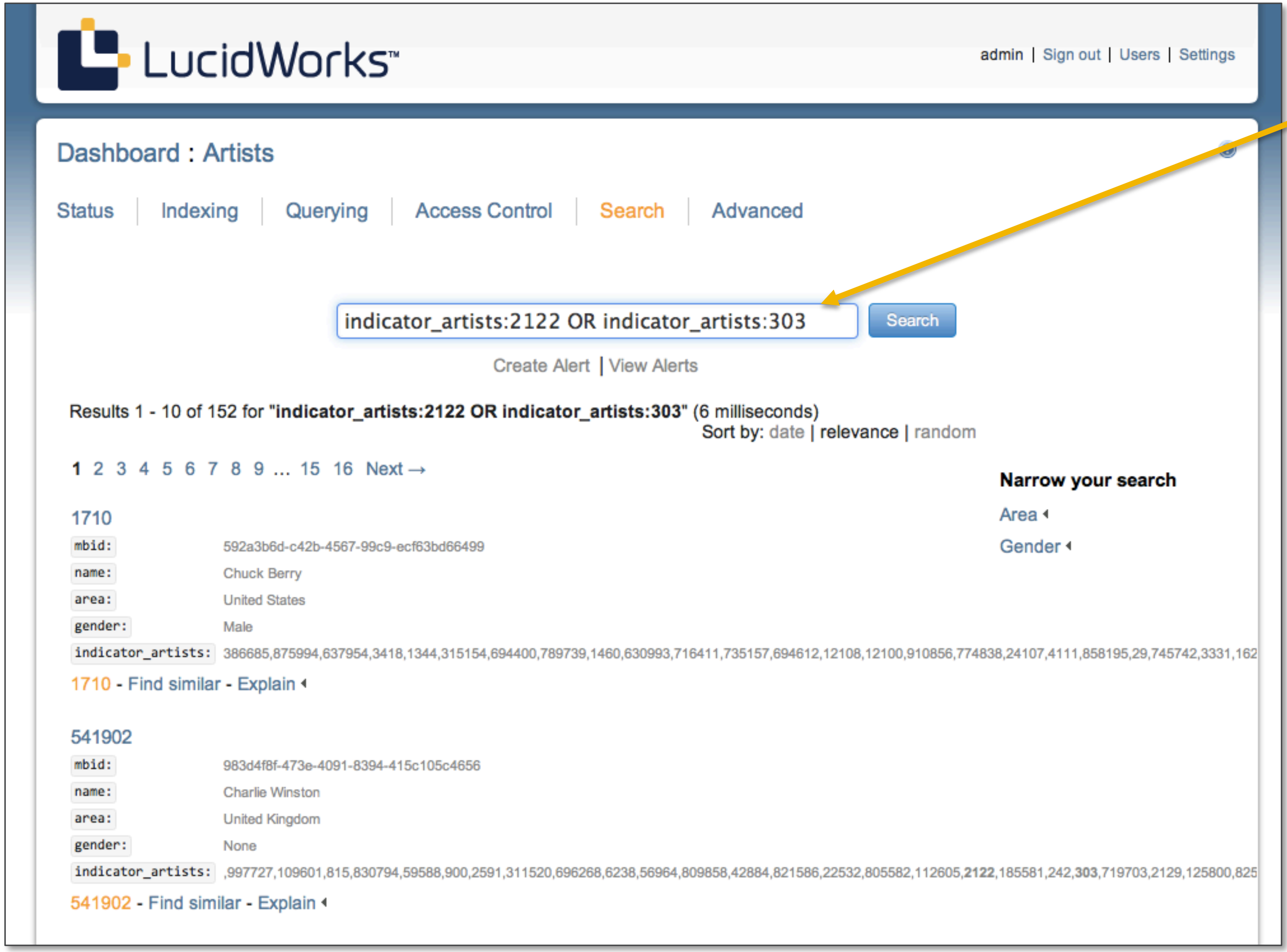
That one row from indicator matrix becomes the indicator field in the Solr document used to deploy the recommendation engine.

```
id: t4
title: puppy
desc: The sweetest little puppy ever.
keywords: puppy, dog, pet

indicators: (t1)
```

Note: Data for the indicator field is added directly to meta-data for a document in Solr index. You don't need to create a separate index for the indicators.

Internals of the Recommender Engine

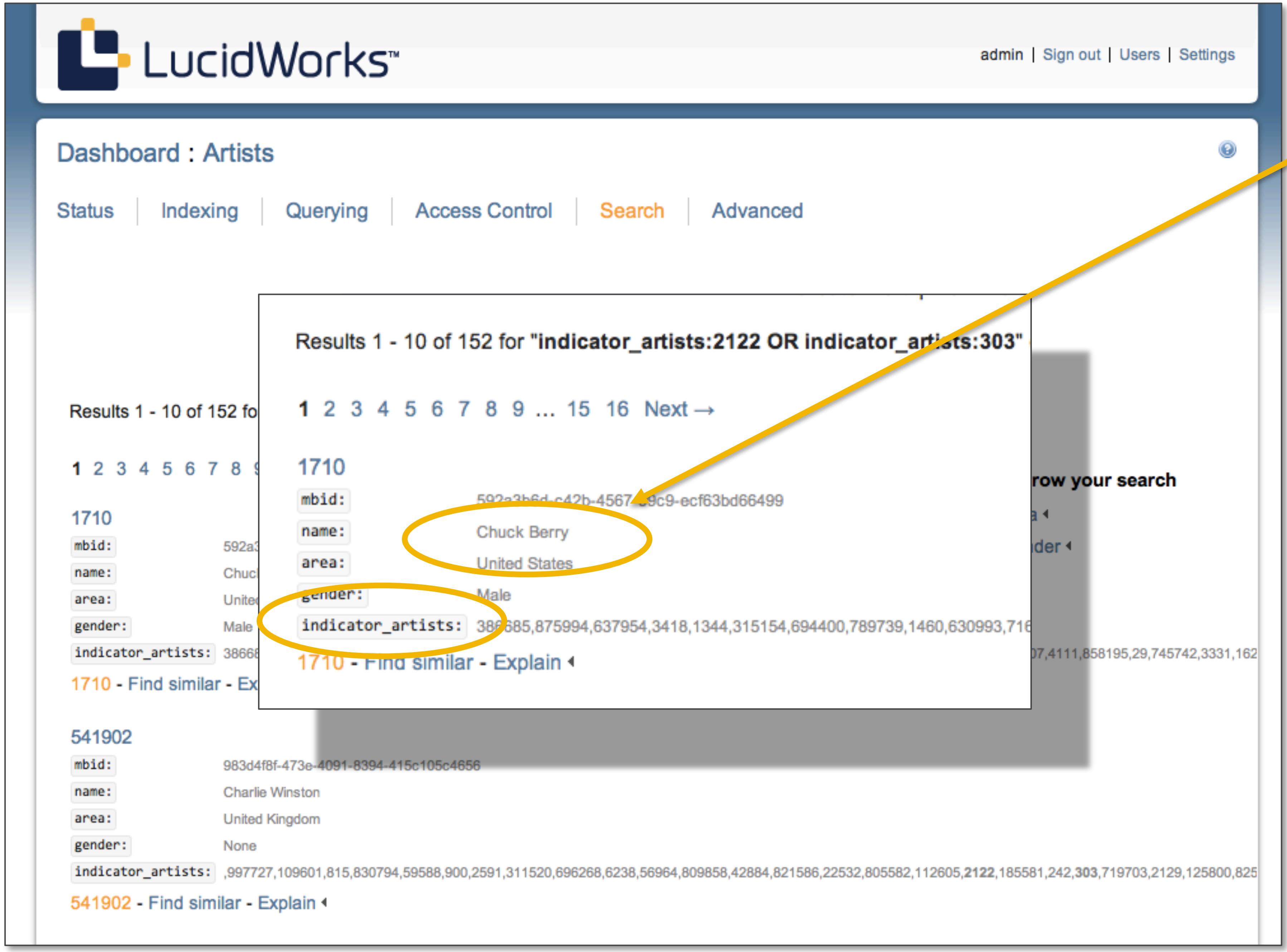


Q: What should we recommend if new user listened to 2122:Fats Domino & 303:Beatles?

A: Search the index with “indicator_artists:2122 OR indicator_artists: 303”

MAPR™ TECHNOLOGIES

Internals of the Recommender Engine



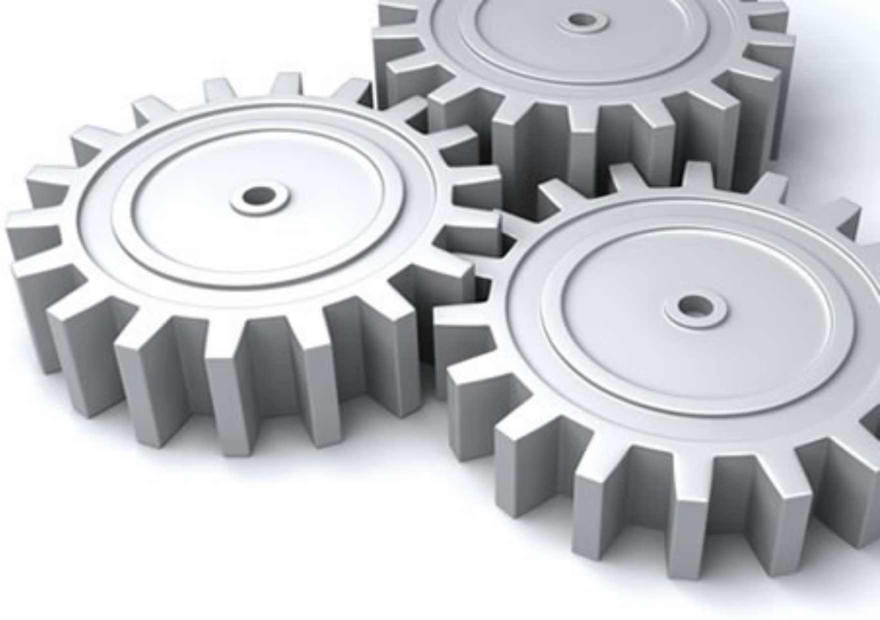
1710:Check Berry is the top recommendation

MAPR TECHNOLOGIES

Toolbox for Predictive Analytics with Hadoop

- Mahout
  - Use it for Recommendations, Clustering, Math
- Vowpal Wabbit
  - Use it for Classification (but it's harder to use)
- SkyTree
  - Commercial (not open source) implementation of common algorithms

Operational Considerations



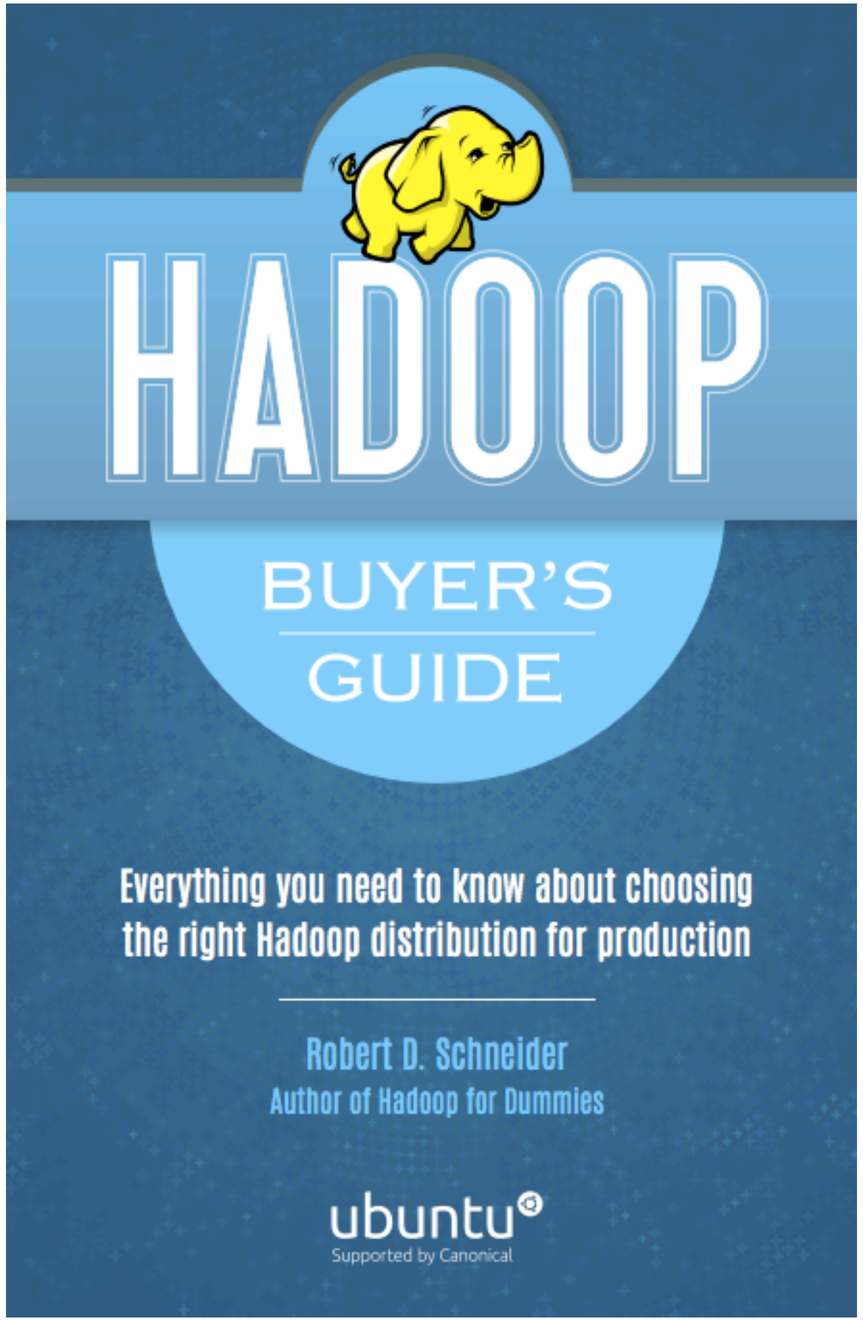
- Snapshot your raw data before training
  - Reproducible process
  - MapR snapshots
  - (Beware of HDFS so-called “snapshots” – not consistent…)

- Leverage NFS for real-time data ingestion
  - Train the model on today’s data
  - Learning schedule independent from ingestion schedule

- Look for HA, data protection, disaster recovery
  - Predictive analytics increases revenue or reduces cost
  - Quickly becomes a must-have, not a nice-to-have

MAPR TECHNOLOGIES

Starting a Project or Moving to Production?



- Read the Hadoop Buyer's Guide
  - Free hard copies at the MapR booth
  - Or read it online: www.HadoopBuyersGuide.com
- Come to our office hour (right now, 2:45pm)
  - Will be joined by Ted Dunning, Mahout Committer and author of Mahout In Action

- Contact MapR
  - Data Science team that can help you with Hadoop and predictive analytics

MAPR TECHNOLOGIES

Thank You