Welcome to the O'Reilly Strata Online Conference

Big Data and the Ethics and Challenges of Living in a Connected Society

September 27, 2013
9:00 AM -11:00 AM (PT)

MAPR means business

40 million
incremental revenue
in the first year
with MapR

FORTUNE 100 COMPANY

90 billion
real-time auctions
per day
conducted with MapR

rubicon PROJECT

1.7 trillion
events processed
per month
using MapR

comScore

come
visit
booth
102

Agenda

9:00 – 9:10 Intro by Alistair Croll

9:10– 9:30 What Makes Us Human? A Tale of Advertising Fraud – Claudia Perlich (Dstillery)

9:30 – 9:50 Machine Learning Applications: Recommendation Engines Using Multiple Behavior Sources – Ted Dunning (MapR)

9:50 – 10:10 Real-time Recommendations for Retail: Architecture, Algorithms, and Design - Juliet Hougland and Jonathan Natkins (WibiData)

10:10 – 10:30 Leveling Up With Hadoop: How Blizzard's Business Intelligence Supports The Worlds of Warcraft, Starcraft, and Diablo – Amanda Gerdes (Blizzard Entertainment)

10:30 – 10:50 Interactive Visualization of "Big" Data – Jeffrey Heer (Tifacta)

10:50 – 11:00 Closing by Alistair Croll

Intro

9:00 – 9:10 am PT

Alistair Croll
Solve for Interesting
Strata Program Chair

What Makes Us Human? A Tale of Advertising Fraud

9:10 – 9:30

Claudia Perlich

Chief Scientist, Dstillery


dstillery
What makes us human?
A tale of advertising fraud.

Alan Turing

(1912-1954)

"A computer would deserve to be called intelligent if it could deceive a human into believing that it was human."

Dark Side of Display Advertising

Click fraud botnet defrauds ...

advetising spider

PCWorld AUSTRALIA

ORECK Magnesium RS LIGHTEST FULL POWERED U JUST 7.7 POUNDS

All Categories | Shop & Compare | Business Centre | Jobs | Guides

Back to Uni Survival Guide

News ▸ Click fraud botnet defrauds advertisers up to $6 m...

Click fraud botnet defrauds advertisers up to $6 million

The 'Chameleon' botnet generates high traffic on low-quality websites

Jeremy Kirk (IDG News Service) — 20 March, 2013 00:55

Share +1 0 Tweet 0 Be the first to comment 0

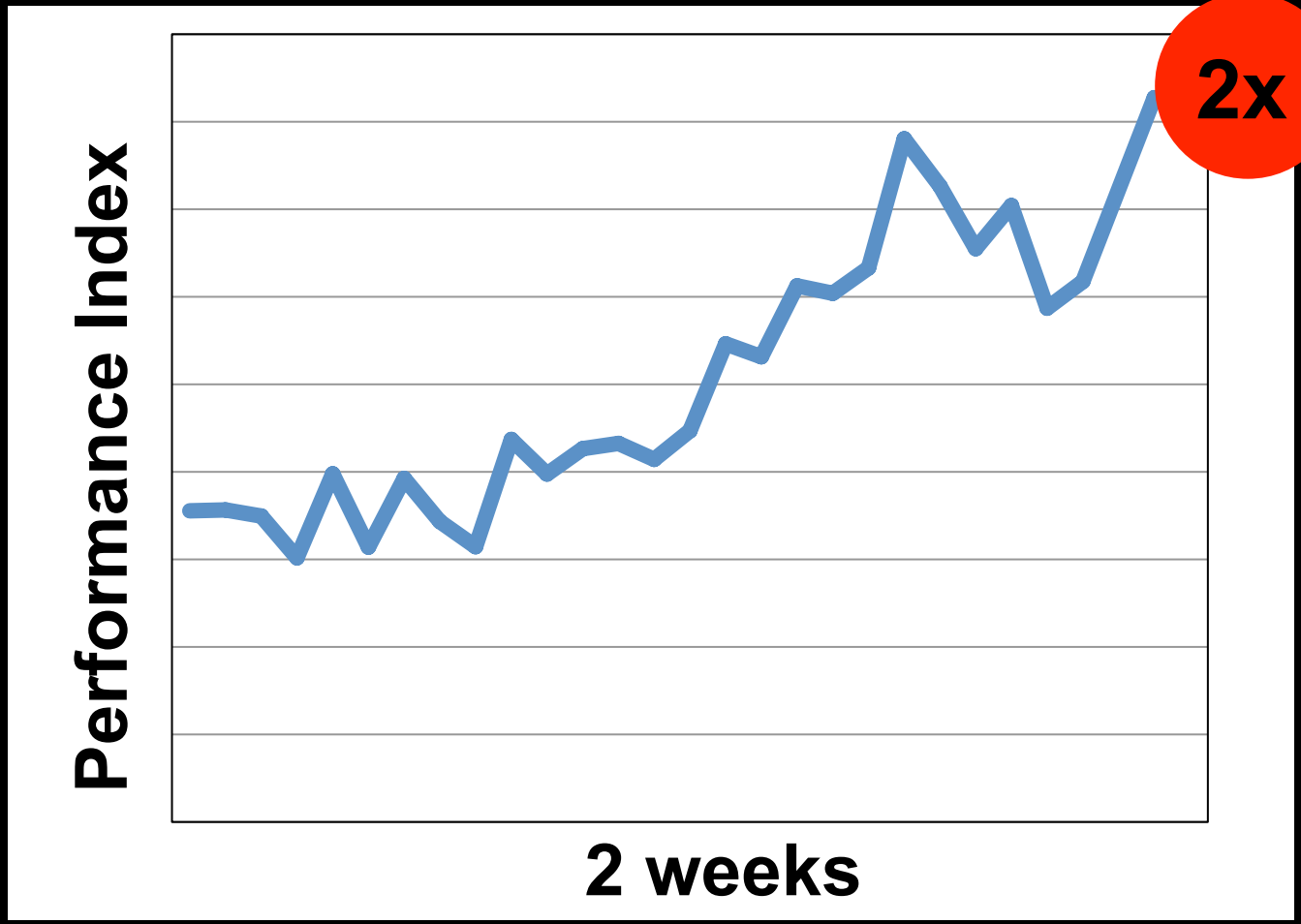
An advertising analytics company said it has discovered a botnet that generates upwards of US$6 million per month by generating bogus clicks on display advertisements.

Spider.io, based in the U.K., wrote that the botnet code, called Chameleon, has infected about 120,000 residential computers in the U.S. and perpetrates click fraud on 202 websites that collectively deliver 14 billion ad impressions. Chameleon is responsible for 9 billion of those impressions, Spider.io said.

Click fraud cheats Web advertisers by making them pay for clicks on ads that are not legitimate, depriving them of customers and revenue. Spider.io said advertisers pay an average of $0.69 per one thousand impressions.

dstillery

Unreasonable Performance Increase Spring 12



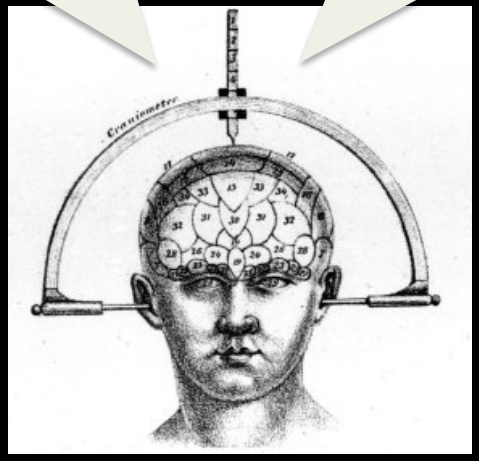
dstillery

Dstillery Targeting: Agnostic Data

A consumer's online activity → **gets recorded like this:**

The Non-Branded Web

The Branded Web



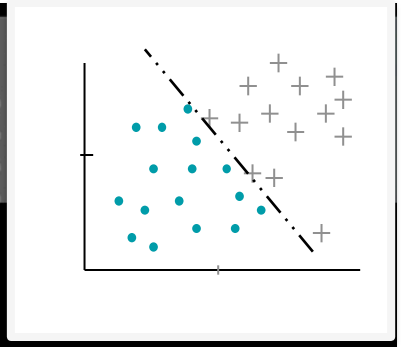
Purchases
Encoded
date1 3012L20
date 2 4199L30
…
date n 3075L50

Browsing History
Hashed URL's:
date1 abkcc
date2 kkllo
date3 88iok
date4 7uiol
…

computers do not 'understand' who you are

dstillery

Model of browsing histories



Using Logistic Regression, we estimate statistical correlations between 10s of millions of web URLs and 1000s of actions.

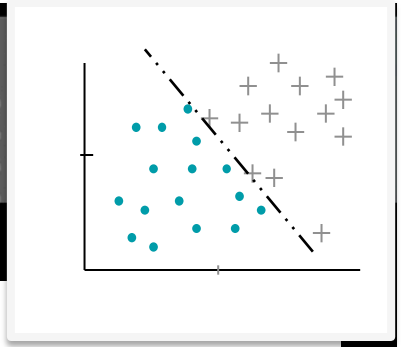
Likelihood to Convert given Visit

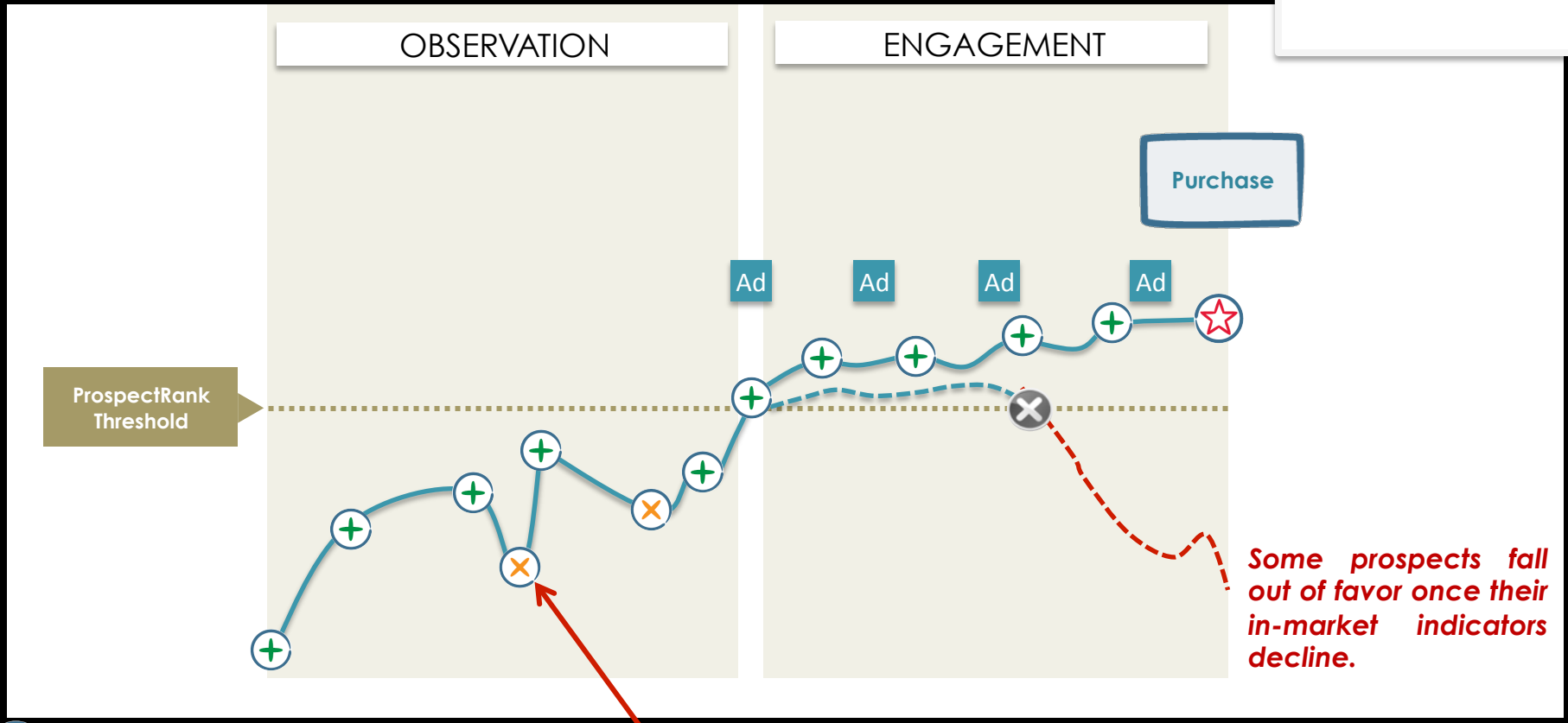
non-branded websites

$$p(\text{buy}|\text{urls}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

dstillery

Real-time Scoring of a User





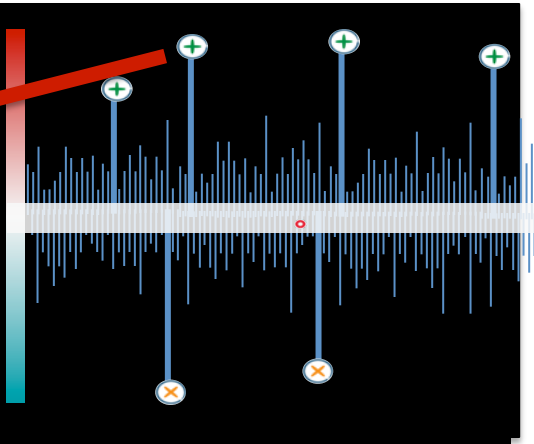
site visit with positive correlation

site visit with negative correlation

p(buy|urls) $\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

dstillery

Oddly predictive websites?



Sheraton HOTELS & RESORTS

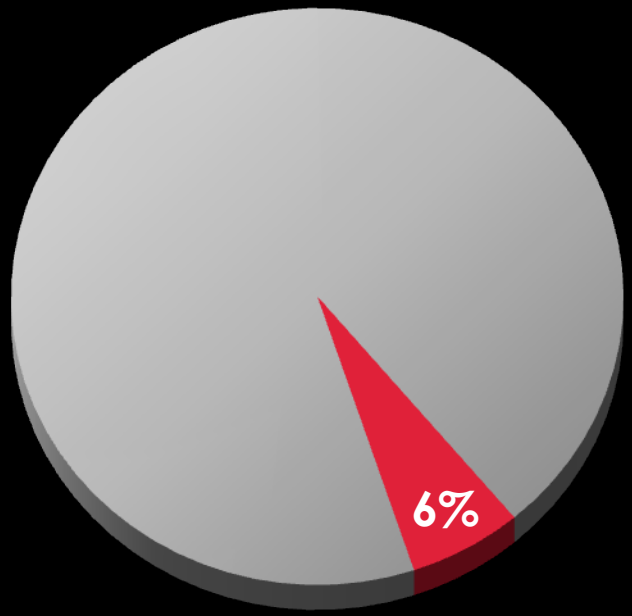
Microsoft Office

dstillery

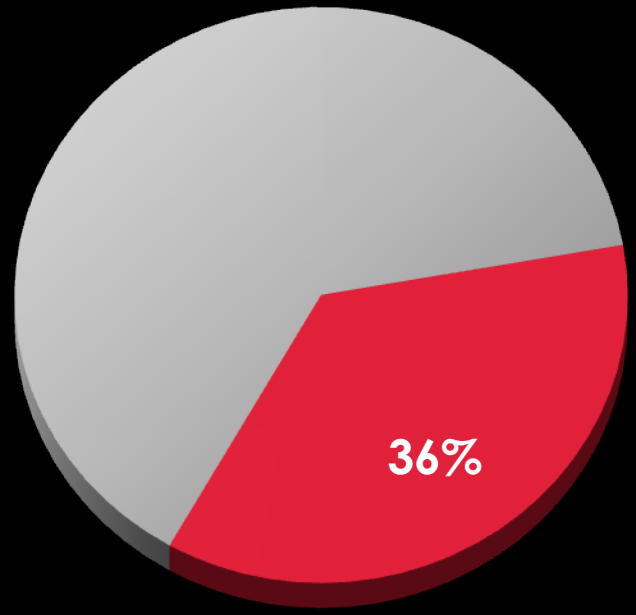
Appearances are deceiving …

dstillery

36% traffic is Non-Intentional



2011



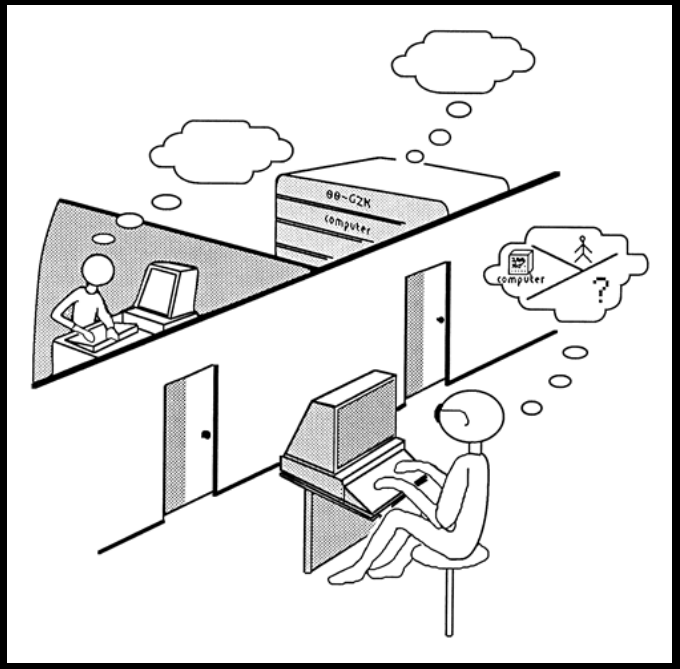
2012

comScore

dstillery

Telling the difference between an algorithm and a human

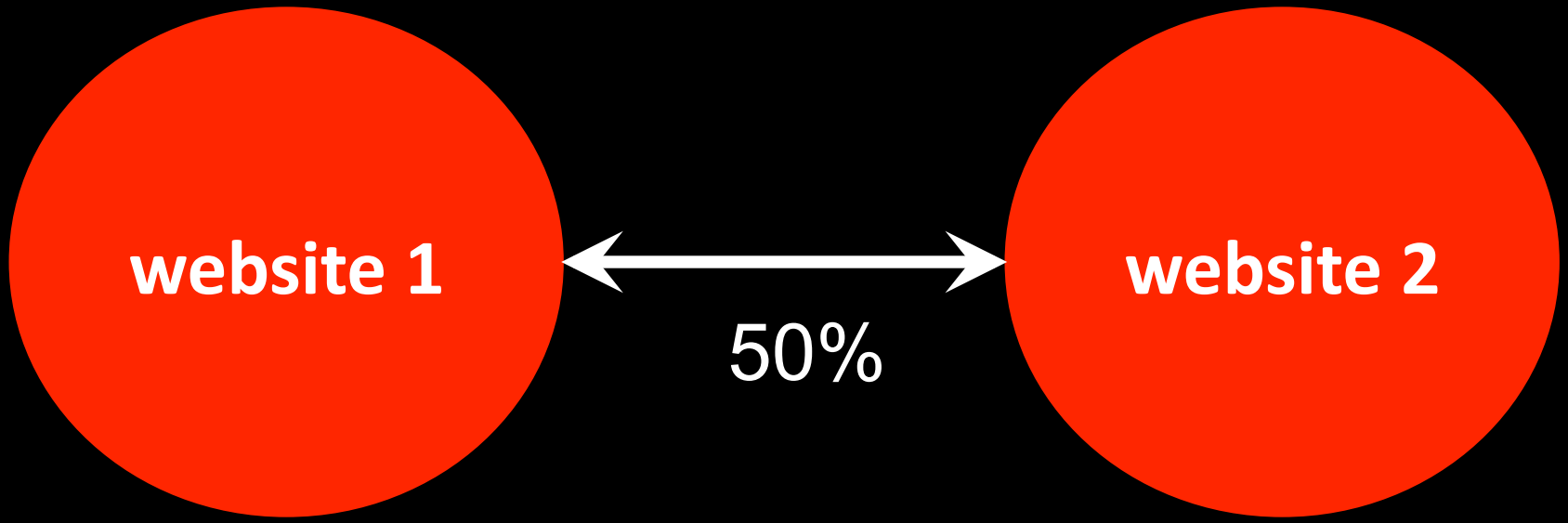
Turing test



KAPTCHA

dstillery

Traffic patterns are ‘non - human’



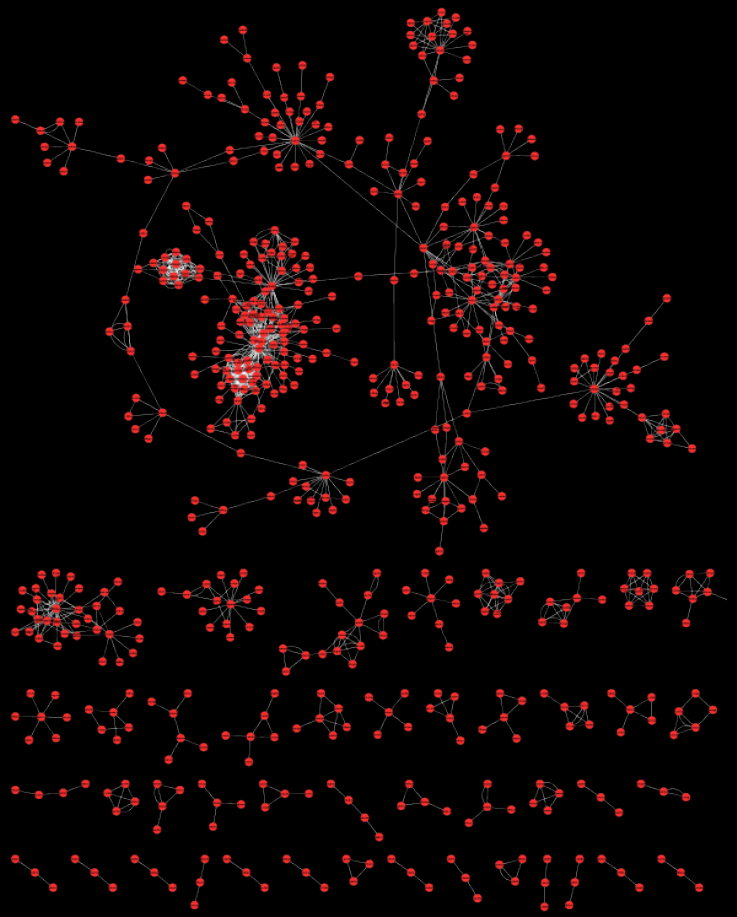
Data from Bid Requests in Ad-Exchanges

dstillery

WWW 2010

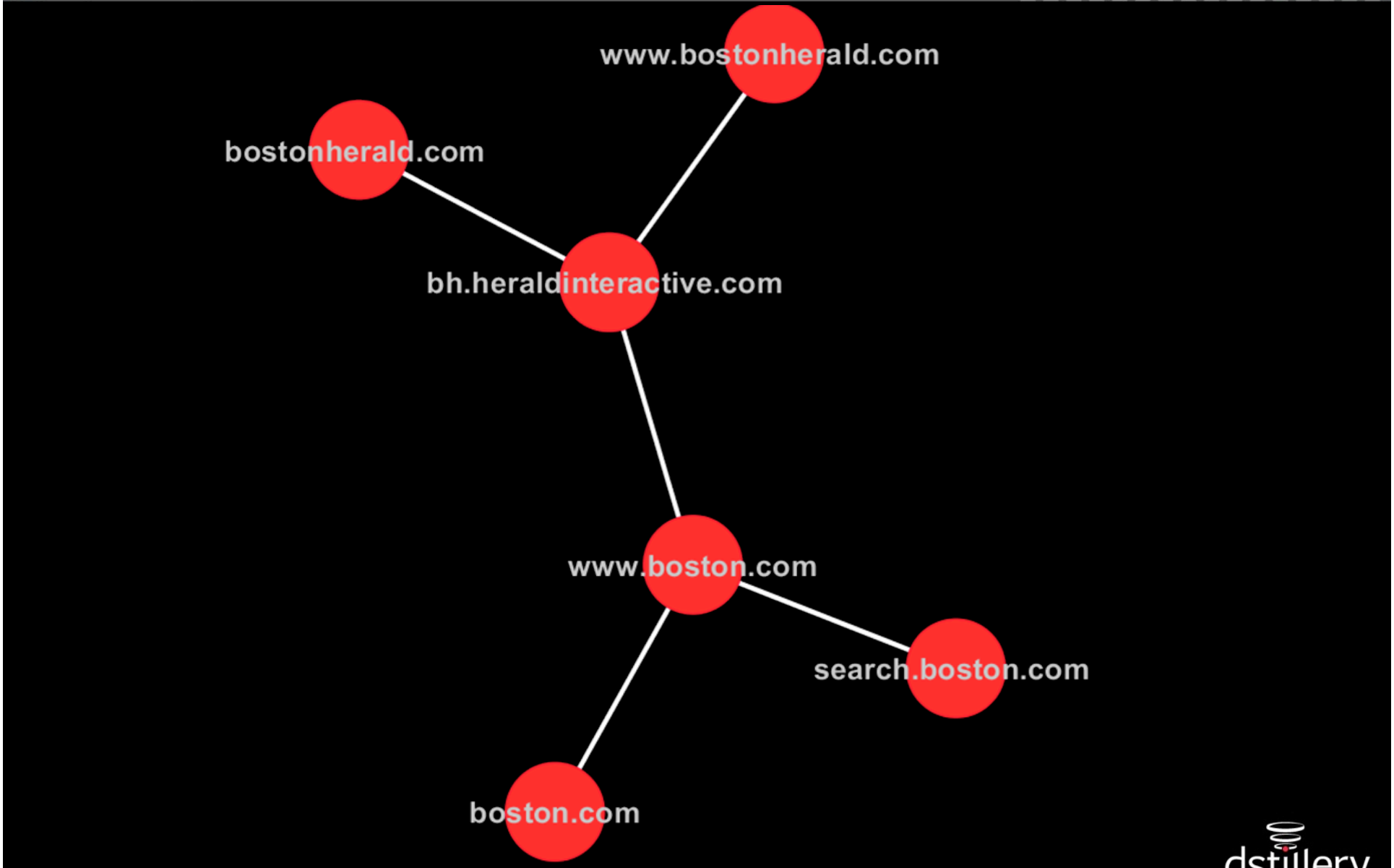
Node:
hostname

Edge:
50% co-visitation



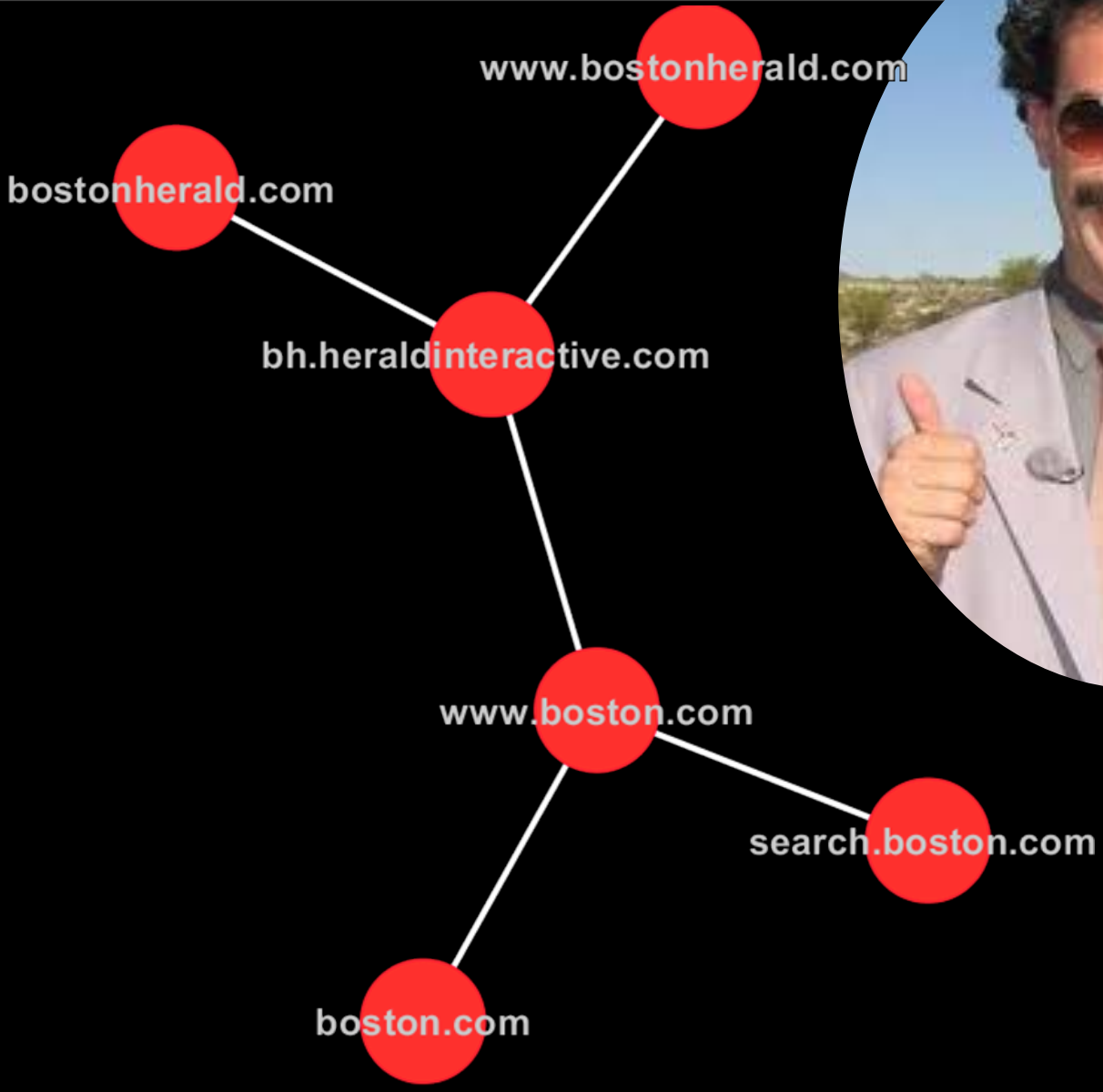
dstillery

Boston Herald



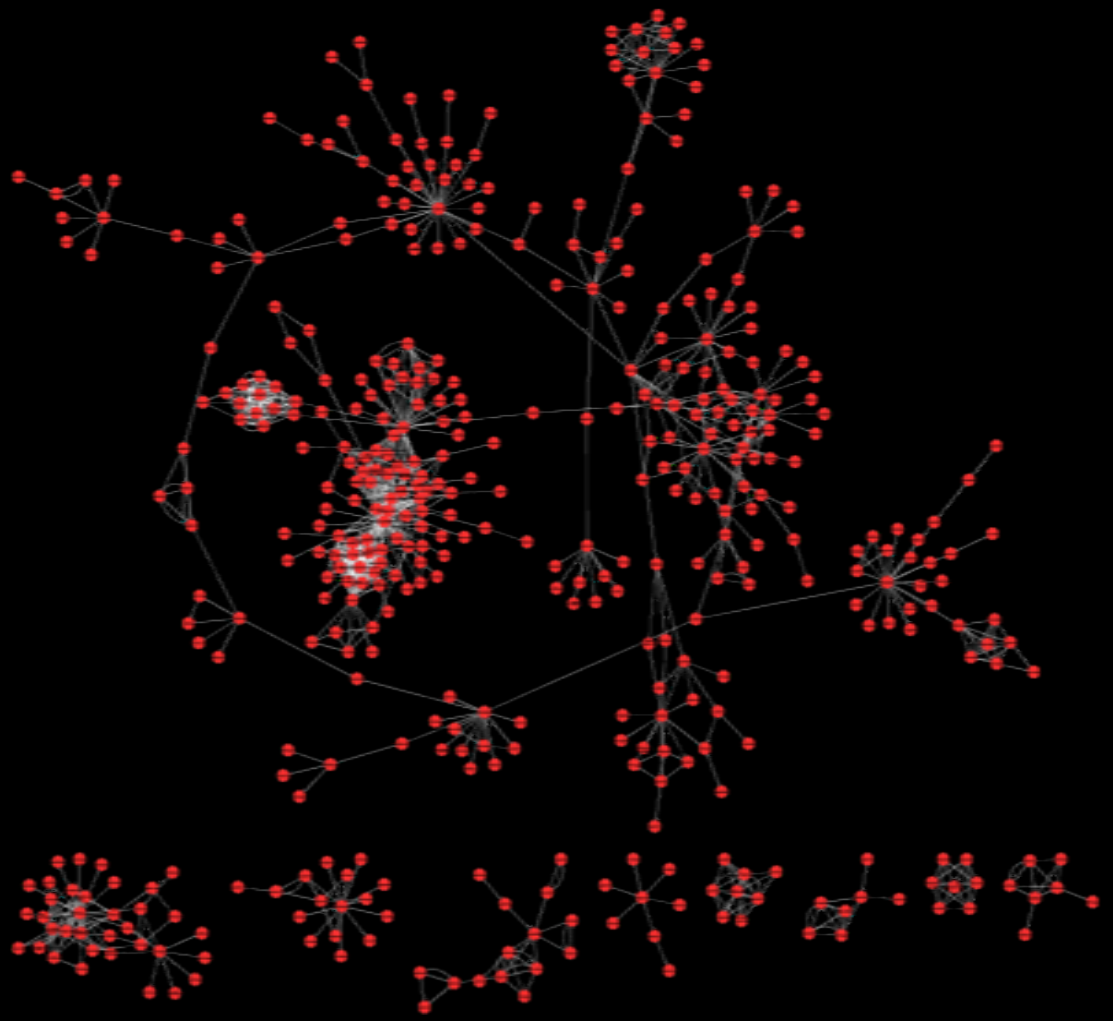
dstillery

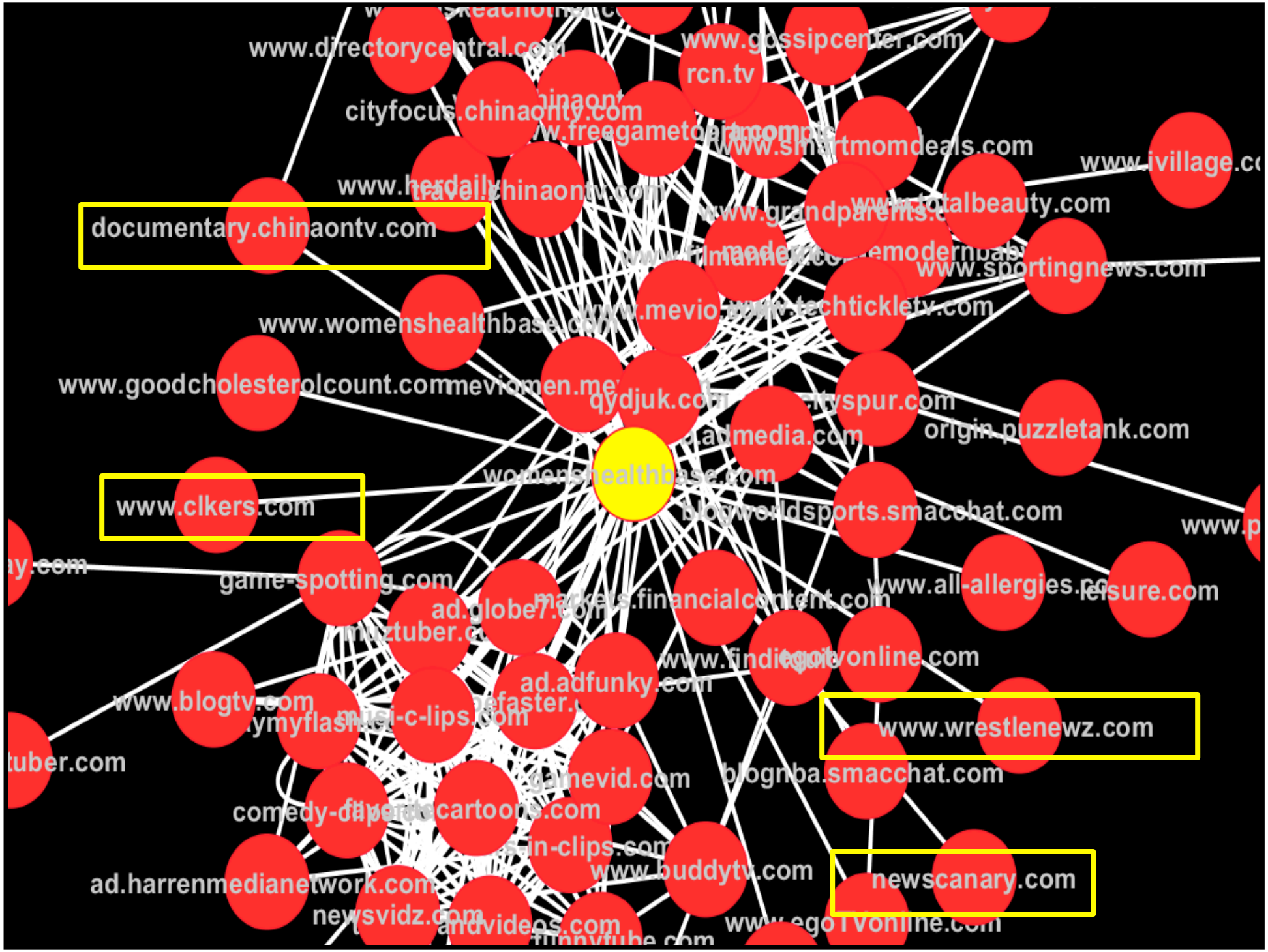
Boston Herald

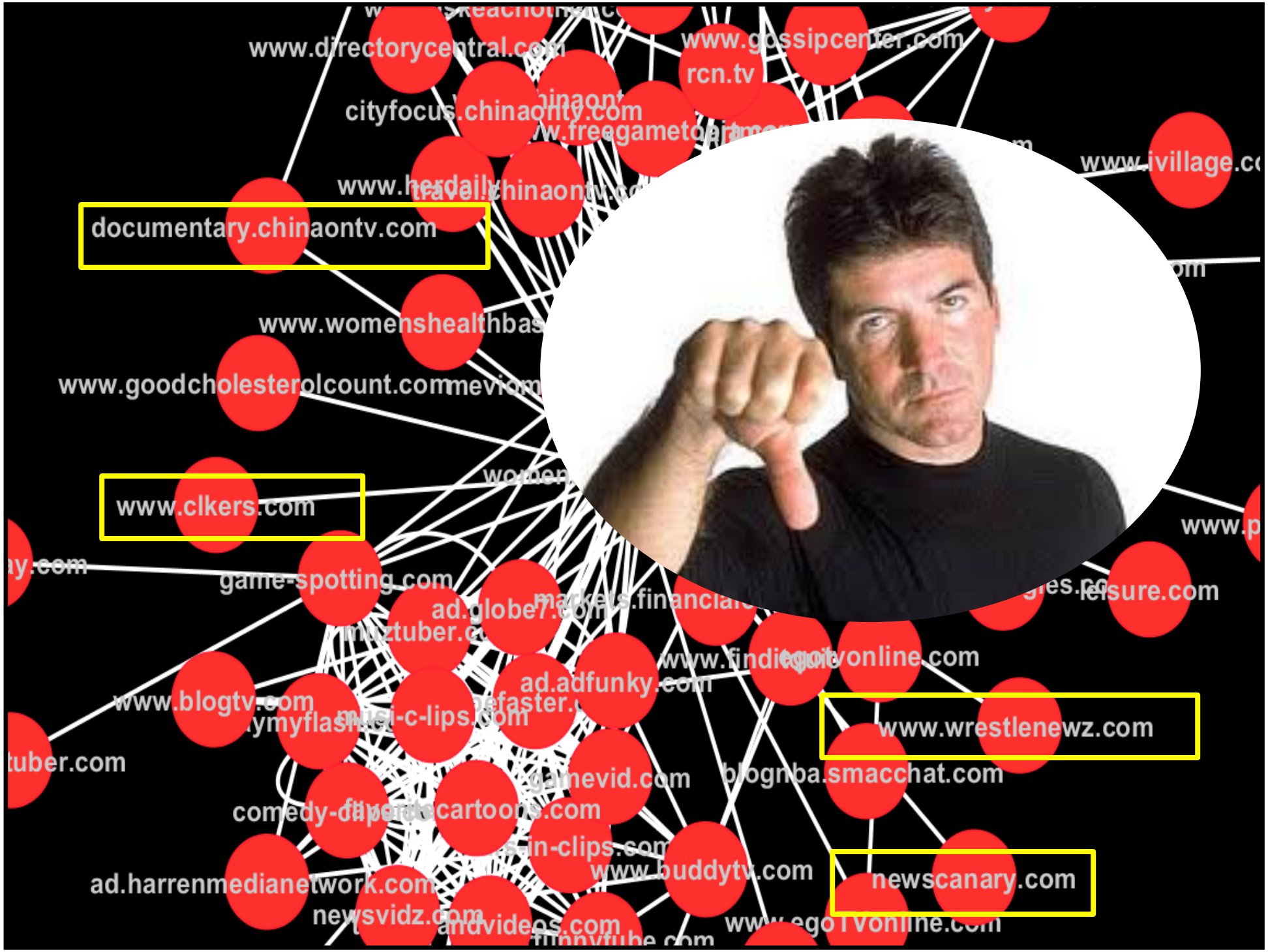


dstillery

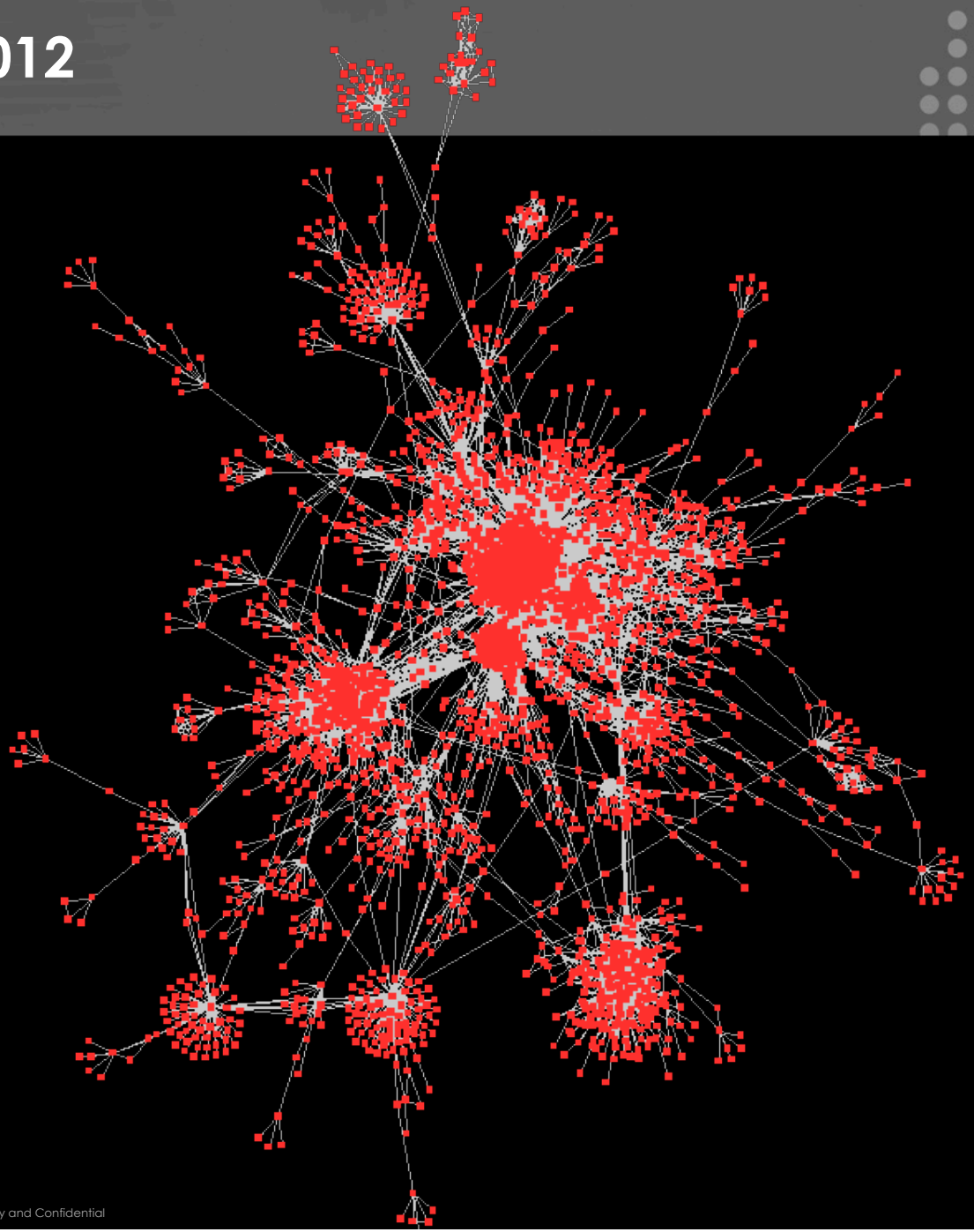
womenshealthbase?




www.directorycentral.com
www.gossipcenter.com
rcn.tv
cityfocus.chinaontv.com
www.smartmomdeals.com
www.ivillage.co
www.herdaily
travel.chinaontv.com
documentary.chinaontv.com
www.totalbeauty.com
www.sportingnews.com
www.womenshealthbase.com
www.goodcholesterolcount.com
qydjuk.com
origin.puzzletank.com
www.clkers.com
womenshealthbase.com
blogworldsports.smacchat.com
game-spotting.com
www.all-allergies.com
leisure.com
ad.globe7.com
www.findingtvonline.com
ad.adfunky.com
www.blogtv.com
www.wrestlenewz.com
blognba.smacchat.com
comedy-dailycartoons.com
www.buddytv.com
newscanary.com
ad.harrenmedianetwork.com
newsvidz.com

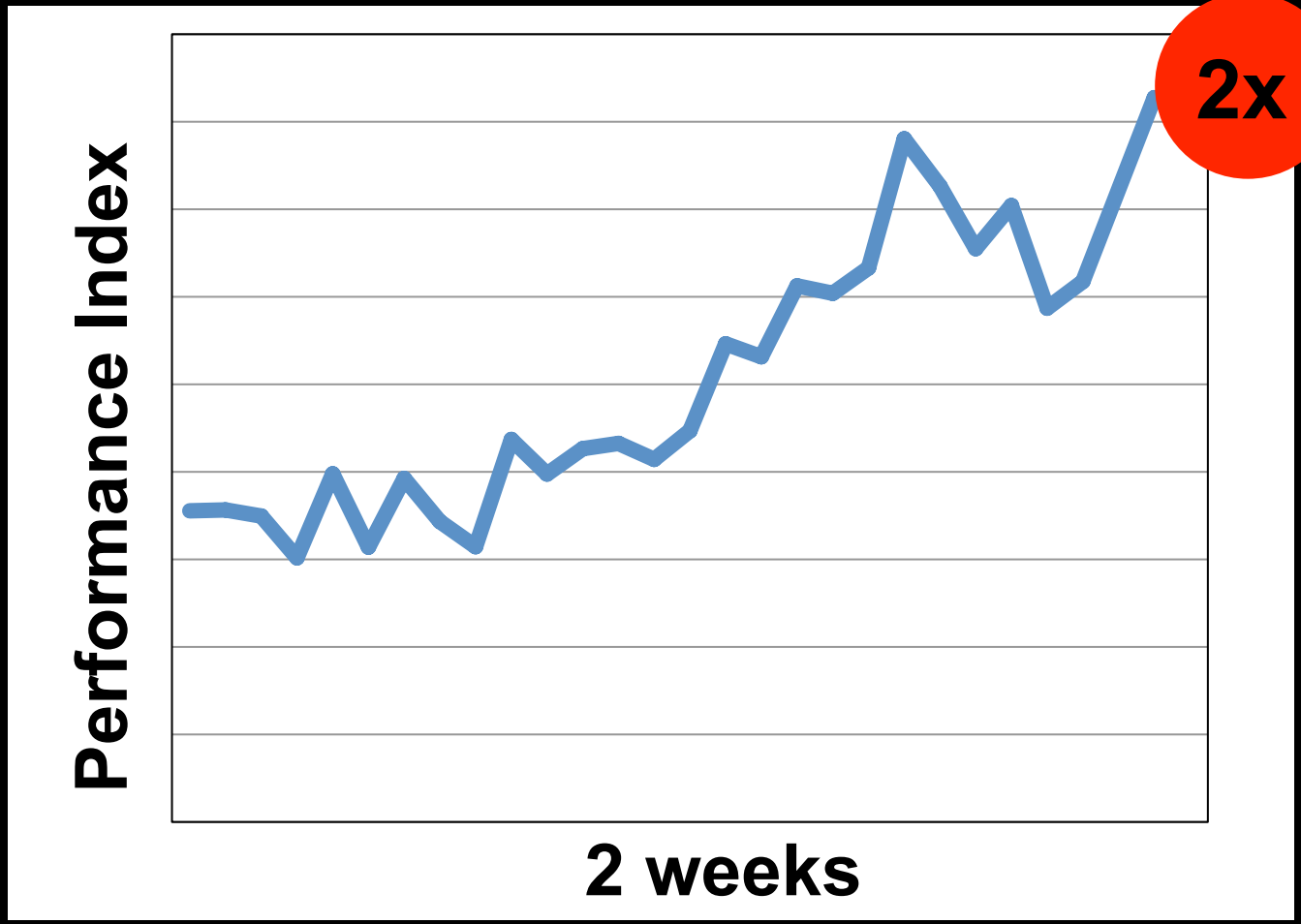

www.directorycentral.com
www.gossipcenter.com
rcn.tv
cityfocus.chinaontv.com
www.herdaily
documentary.chinaontv.com
www.ivillage.co
www.womenshealthbas
www.goodcholesterolcount.com
www.clkers.com
game-spotting.com
ad.globe7.com
leisure.com
www.findingegovonline.com
ad.adfunky.com
www.blogtv.com
www.wrestlenewz.com
gamevid.com
blogmba.smacchat.com
comedy-dailycartoons.com
www.buddytv.com
newscanary.com
ad.harrenmedianetwork.com
newsvidz.com

WWW 2012



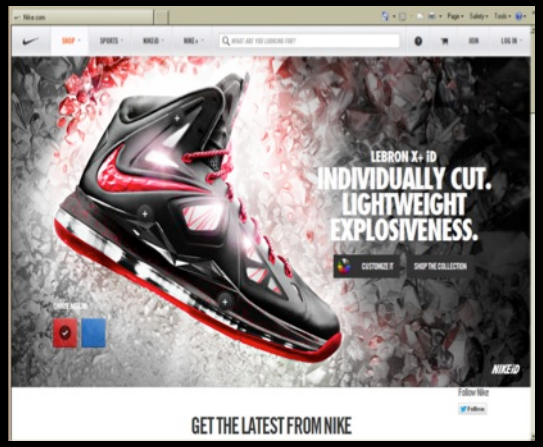
dstillery

Unreasonable Performance Increase Spring 12



dstillery

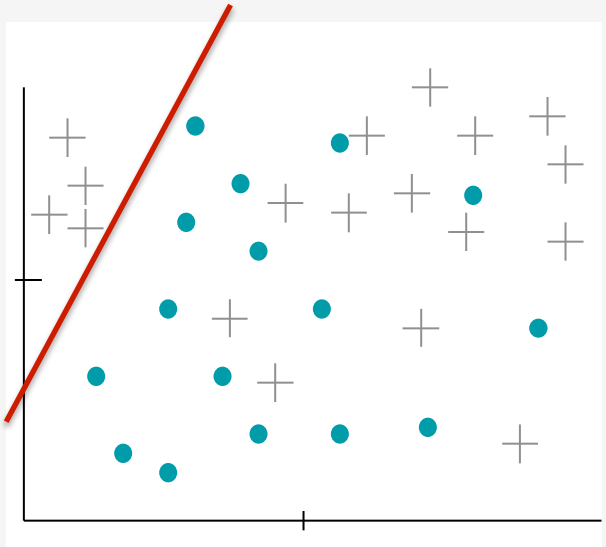
Now it is coming also to brands



- 'Cookie Stuffing' increases the value of the ad for all the retargeters
- Messing up your Web analytics …
- Messes up my models because a botnet is easier to predict than a human

dstillery

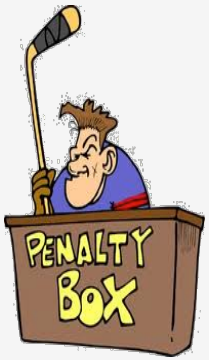
Fraud pollutes my models



- Don't show ads on those sites
- Don't show ads to a high jacked browser

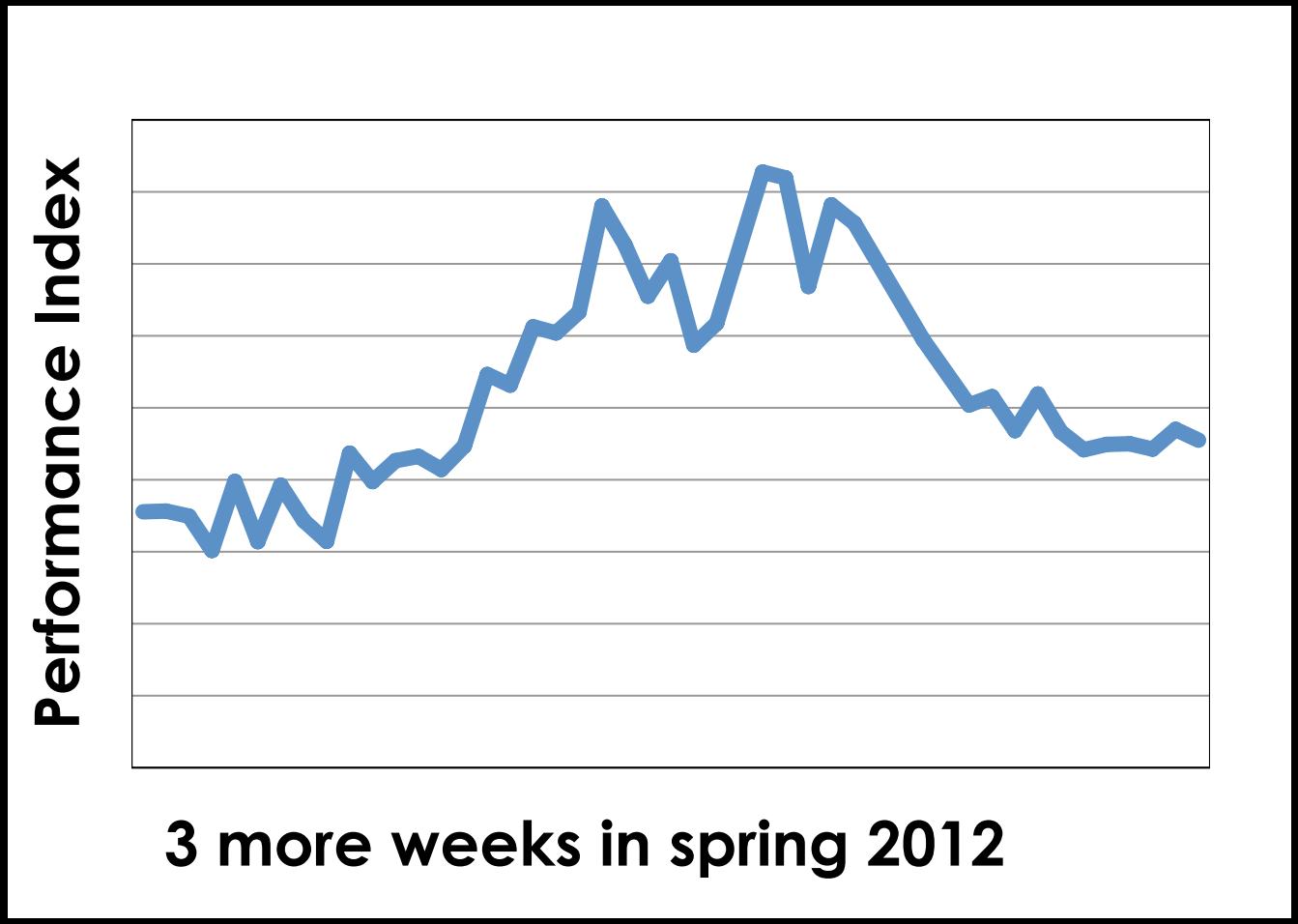
- Need to remove the visits to the fraud sites
- Need to remove the fraudulent site visits

When we see a browser on caught up in fraudulent activity: send him to the penalty box where we ignore all his actions



dstillery

Using the penalty box: all back to normal



3 more weeks in spring 2012

dstillery

Fun Facts: Trying to look human?

dstillery

Thanks!
To the Data science Team

More technical details:
http://m6d.com/who-we-are/datascience/

dstillery

Machine Learning Applications: Recommendation Engines Using Multiple Behavior Sources

9:30 – 9:50

Ted Dunning

Chief Application Architect, MapR

MAPR™ TECHNOLOGIES

Multi-input Recommendations

Topic For Today

- What is recommendation?
- What makes it different?
- What is multi-model recommendation?
- How can I build it using common household items?

Oh … Also This

- Detailed break-down of a live machine learning system running with Mahout on MapR
- With code examples

I may have to summarize

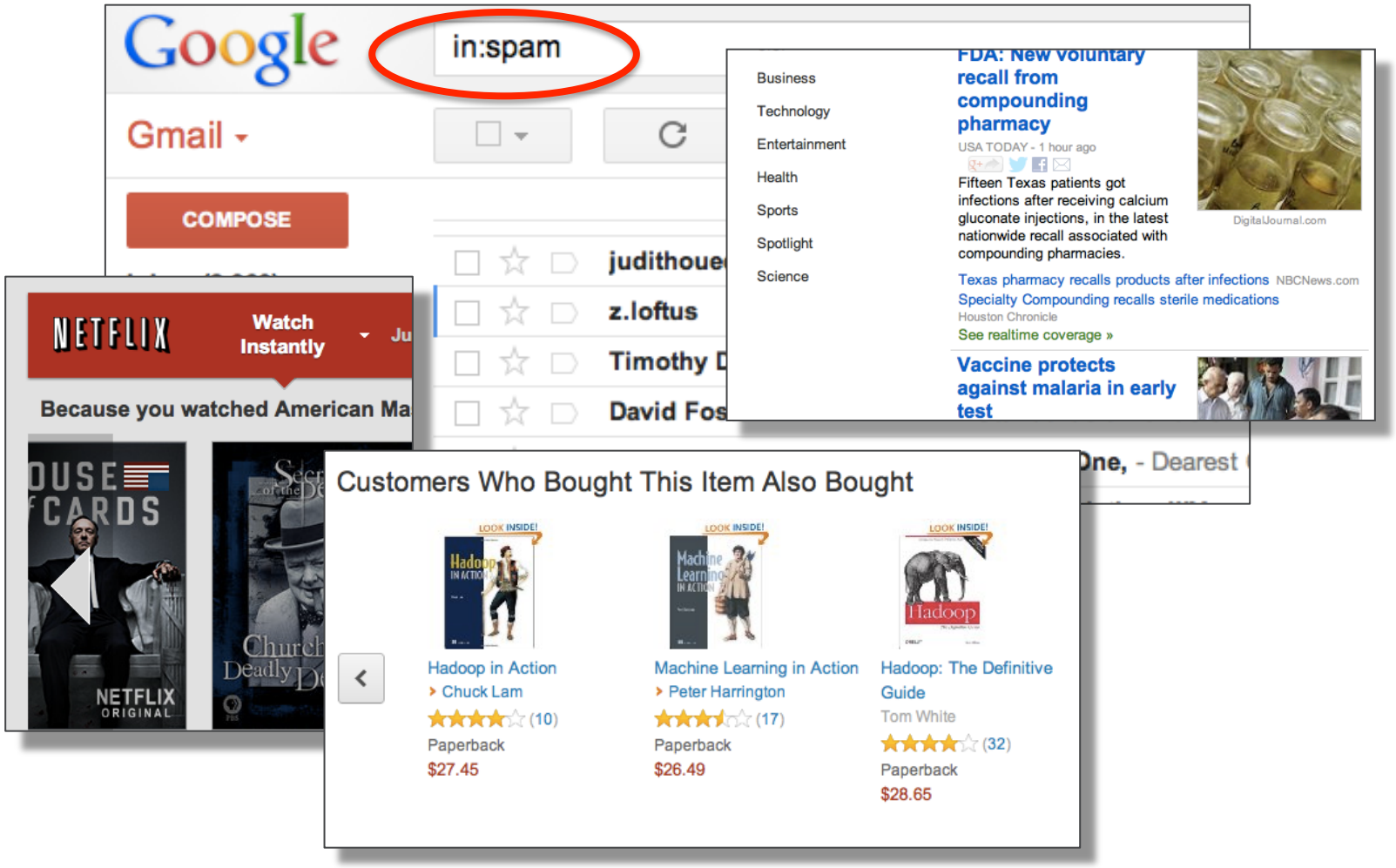
I may have to
summarize
just a bit

Part 1:
5 minutes of background

Part 2:
5 minutes: I want a pony

Part 1:
5 minutes of background

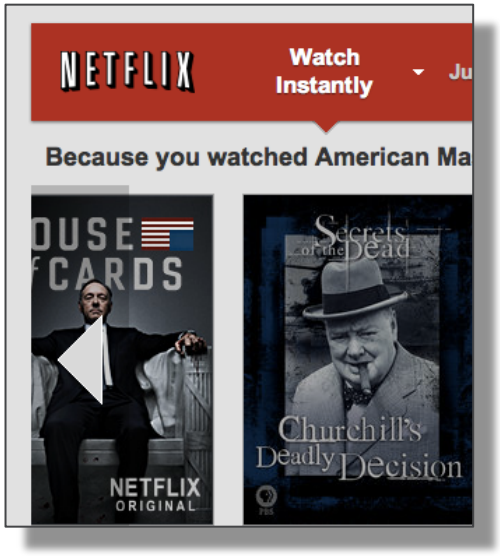
What Does Machine Learning Look Like?

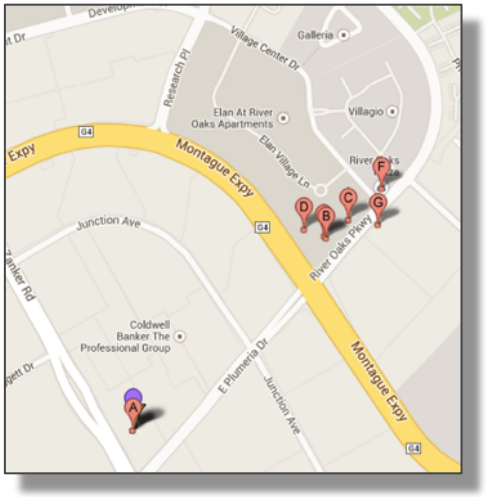


MAPR TECHNOLOGIES

Recommendations as Machine Learning

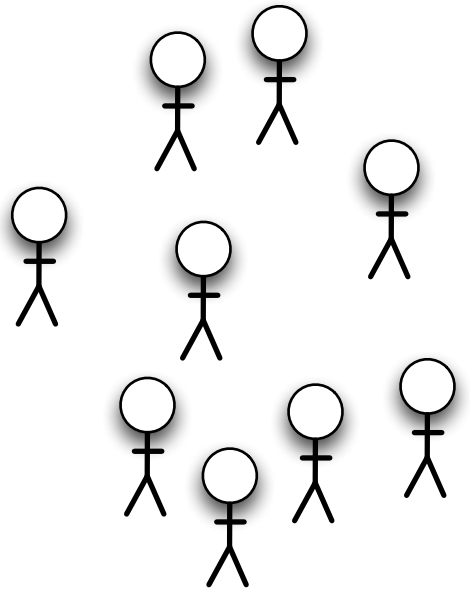
- **Recommendation**:
  - Involves observation of interactions between people taking action (users) and items for input data to the recommender model
  - Goal is to suggest additional appropriate or desirable interactions
  - **Applications include:** movie, music or map-based restaurant choices; suggesting sale items for e-stores or via cash-register receipts





Part 2:
How recommenders work
(I still want a pony)

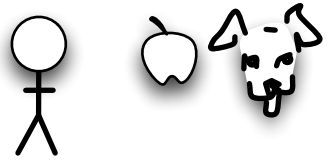
Recommendations



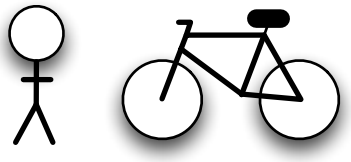
Recap:

Behavior of a crowd helps us understand what individuals will do

Recommendations

Alice 

Alice got an apple and a puppy

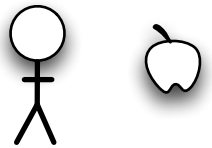
Charles 

Charles got a bicycle

Recommendations

Alice 

Alice got an apple and a puppy

Bob 

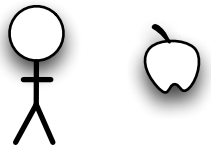
Bob got an apple

Charles 

Charles got a bicycle

Recommendations

Alice 

Bob  ?

What else would Bob like?

Charles 

Recommendations

Alice 

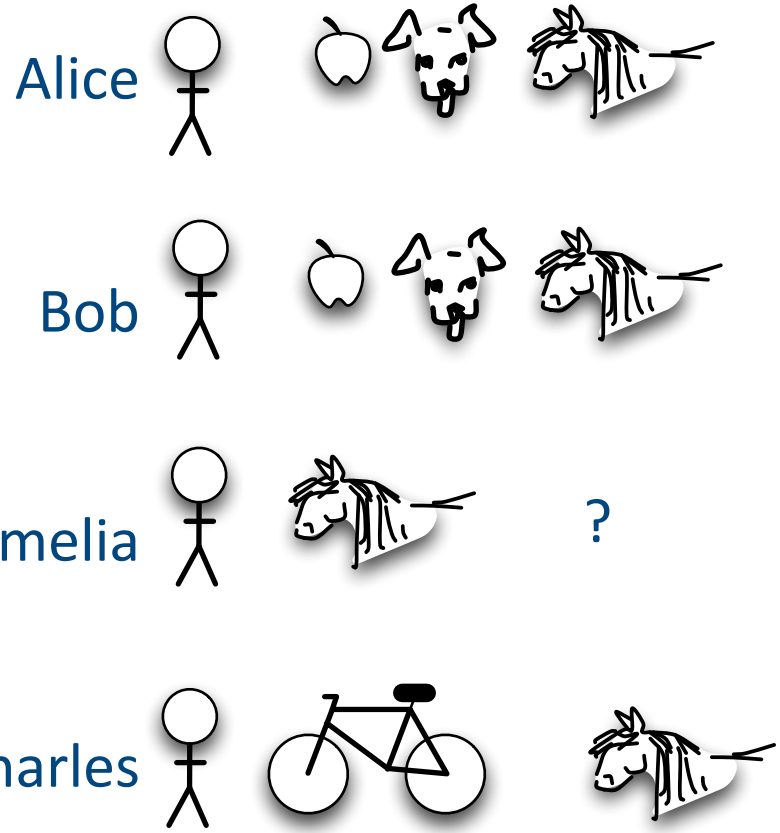
Bob 

A puppy, of course!

Charles 

You get the idea of how recommenders work…
(By the way, like me, Bob also wants a pony)

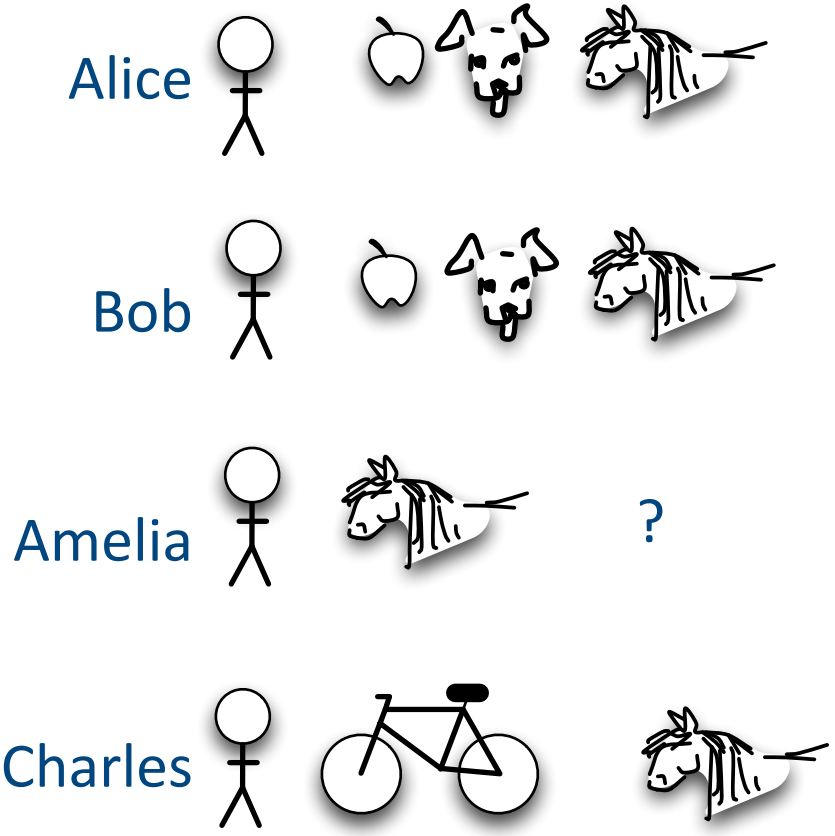
Recommendations



What if everybody gets a pony?

What else would you recommend for Amelia?

Recommendations



If everybody gets a pony, it's not a very good indicator of what to else predict...

Problems with Raw Co-occurrence

- **Very popular items co-occur with everything (or why it's not very helpful to know that everybody wants a pony…)**
  - Examples: Welcome document; Elevator music
- **Very widespread occurrence is not interesting as a way to generate indicators**
  - Unless you want to offer an item that is constantly desired, such as razor blades (or ponies)
- What we want is ***anomalous*** co-occurrence
  - This is the source of interesting indicators of preference on which to base recommendation

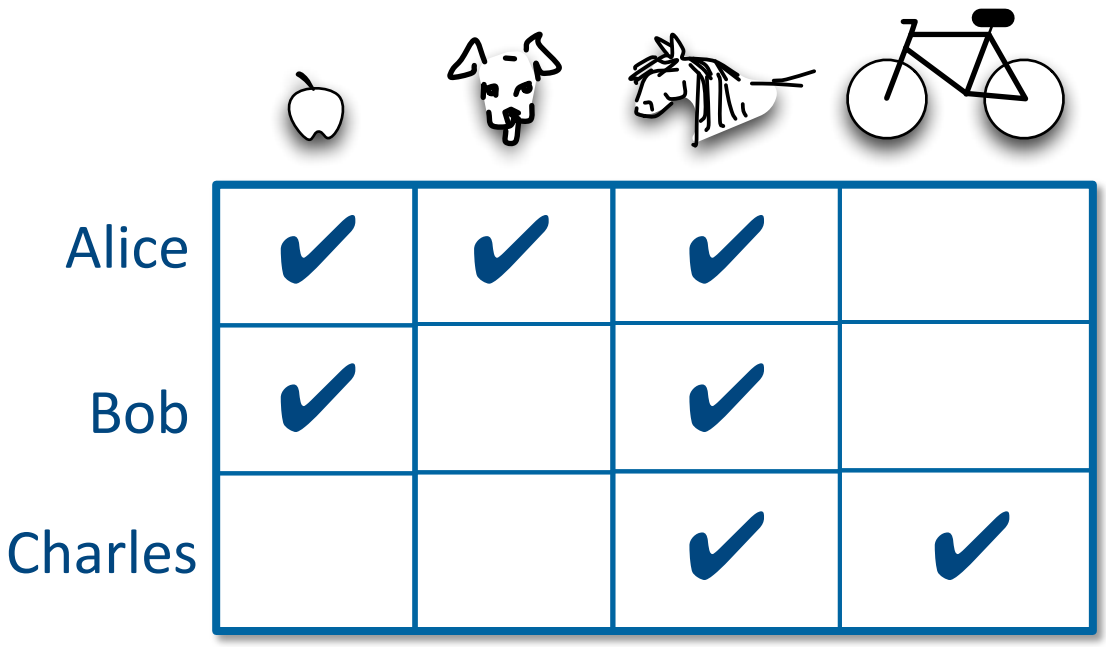
Get Useful Indicators from Behaviors

1. Use log files to build **history matrix** of users x items
   - Remember: this history of interactions will be sparse compared to all potential combinations
2. Transform to a **co-occurrence matrix of** items x items
3. Look for *useful* co-occurrence by looking for *anomalous* co-occurrences to make an **indicator matrix**
   - **Log Likelihood Ratio (LLR**) can be helpful to judge which co-occurrences can with confidence be used as indicators of preference
   - `RowSimilarityJob` in Apache Mahout uses LLR

Log Files

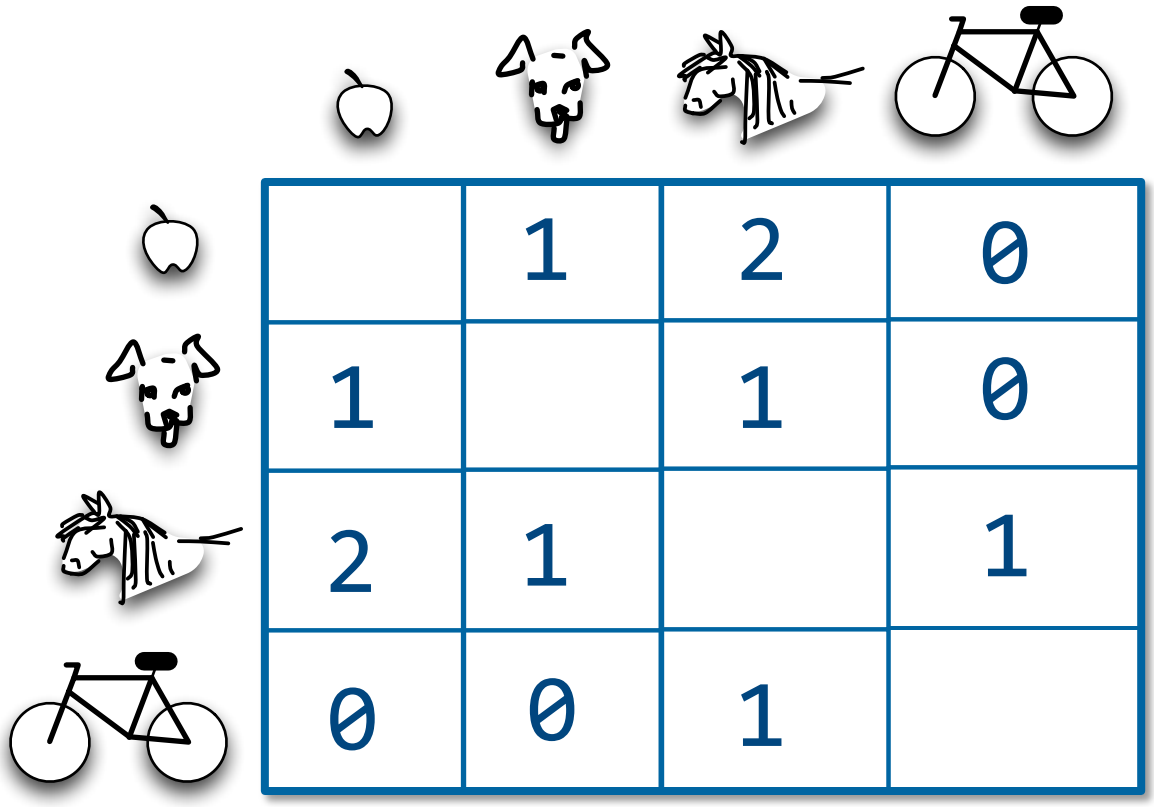
Alice	
Charles	
Charles	
Alice	
Alice	
Bob	
Bob	

History Matrix: Users by Items



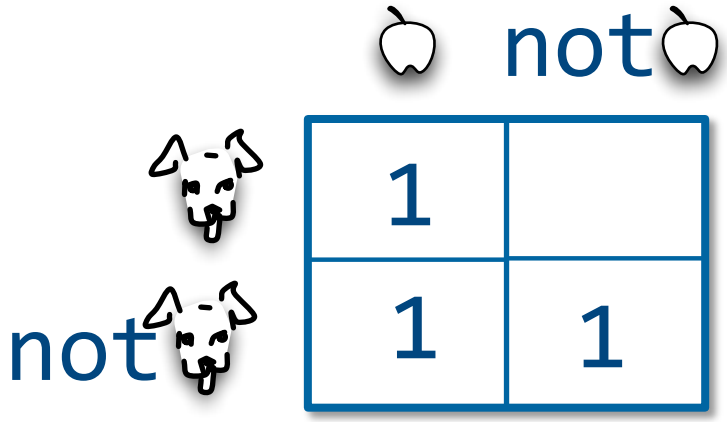
Co-occurrence Matrix: Items by Items

How do you tell which co-occurrences are useful?.



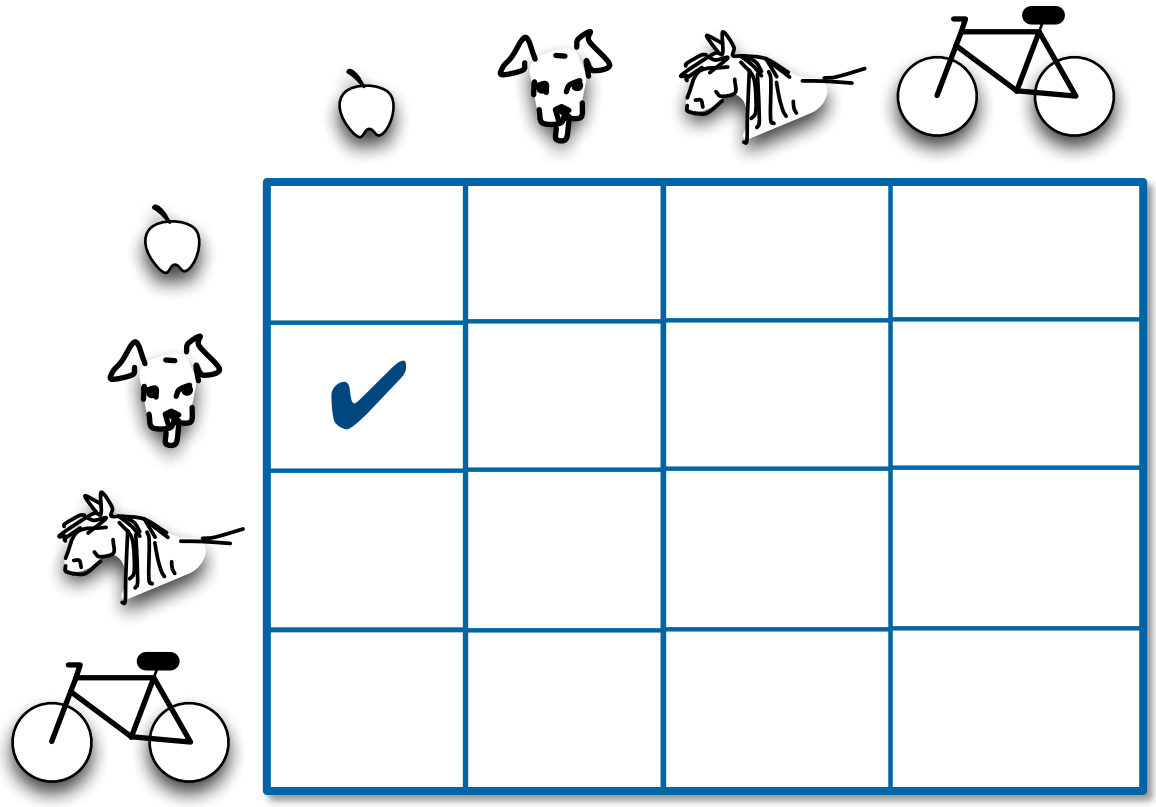
	1	2	0
1		1	0
2	1		1
0	0	1	

Co-occurrence Binary Matrix



Indicator Matrix: Anomalous Co-Occurrence

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

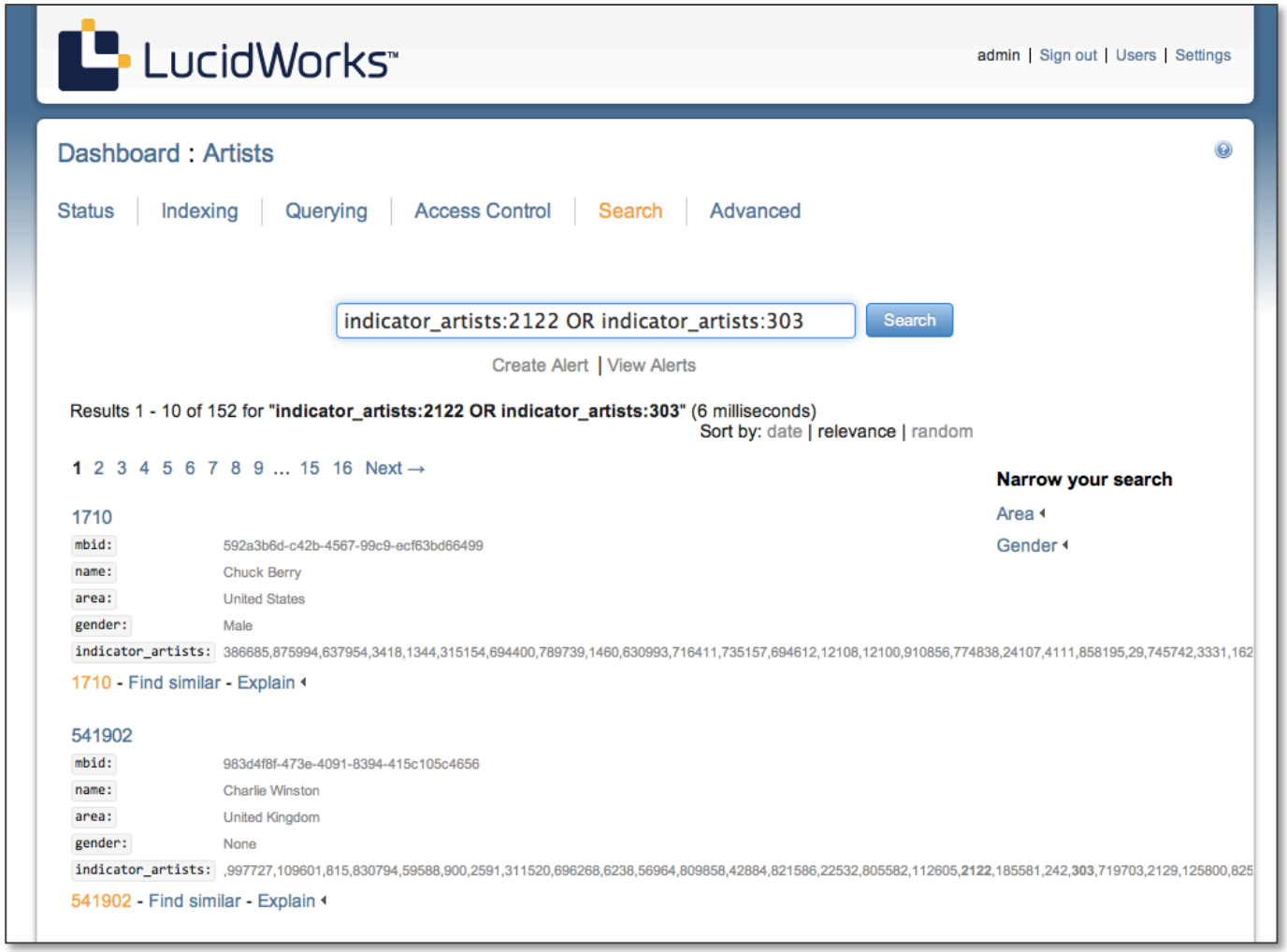


Indicator Matrix

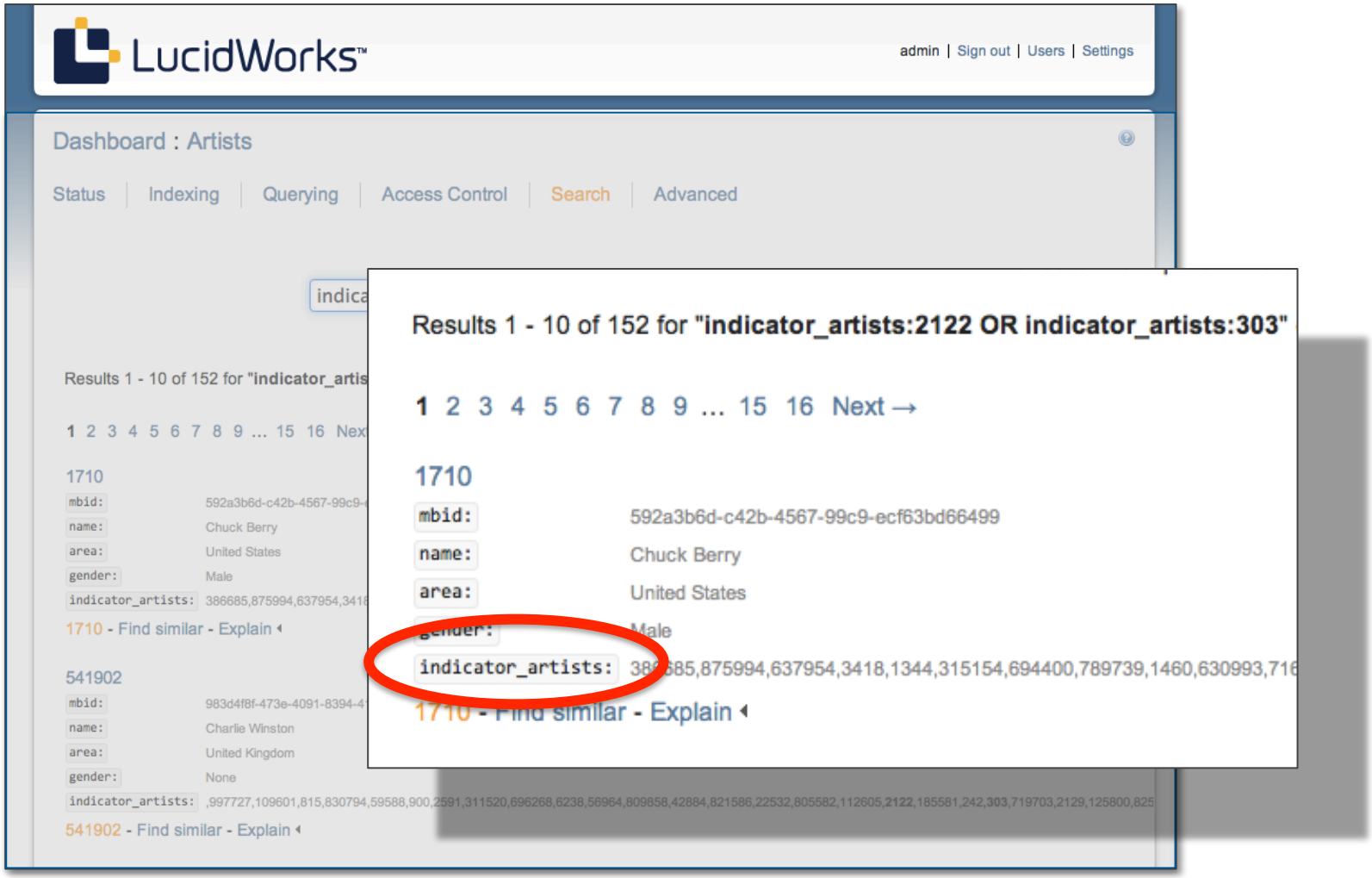
That one row from indicator matrix becomes the indicator field in the Solr document used to deploy the recommendation engine.

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



Internals of the Recommender Engine



A Quick Simplification

- Users who do *h* (a vector of things a user has done)

$$\mathbf{Ah}$$

A translates things into users

- Also do *r*

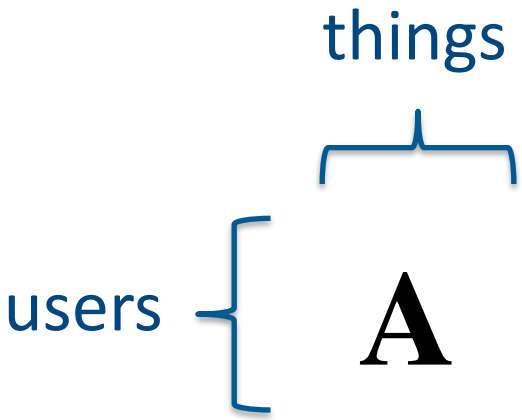
$$\mathbf{A}^T\left(\mathbf{Ah}\right)$$

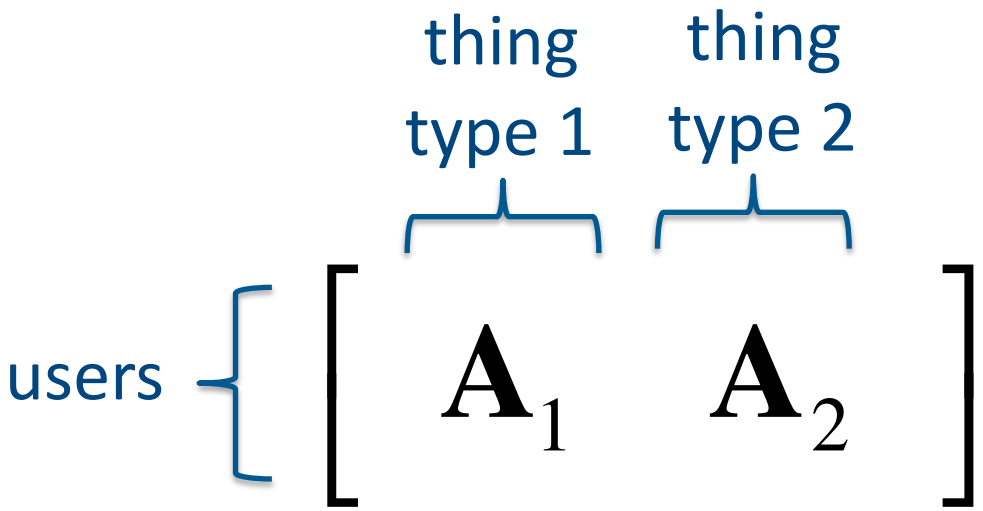
User-centric recommendations
(transpose translates back to things)

$$\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{h}$$

Item-centric recommendations
(change the order of operations)

Nice. But we can do better?





Now again, without the scary math

For example

- Users enter queries (A)
  - (actor = user, item=query)
- Users view videos (B)
  - (actor = user, item=video)
- A^TA gives query recommendation
  - "did you mean to ask for"
- B^TB gives video recommendation
  - "you might like these videos"

The punch-line

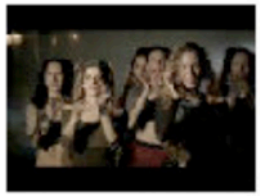
- B^TA recommends videos in response to a query
  - (isn't that a search engine?)
  - (not quite, it doesn't look at content or meta-data)

Real-life example

- Query: “Paco de Lucia”
- Conventional meta-data search results:
  - “hombres del paco” times 400
  - not much else
- Recommendation based search:
  - Flamenco guitar and dancers
  - Spanish and classical guitar
  - Van Halen doing a classical/flamenco riff

Real-life example

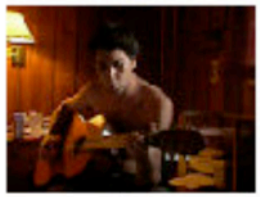
CONCIERTO CIUDAD DE LAS IDEAS PARTE FINAL
Music
58 views



Siudy / Buleria
Music
722 views

Vicente Amigo 2ª parte Ciudad de las Ideas
Music
124 views

Van Halen's Eruption
Music
4400 views



Freestyle Flamenco
Music
653 views

Search-based Recommendations

- Sample docume
  - Merchant Id
  - Field for text description
  - Phone
  - Address
  - Location

  - Indicator merchant id's
  - Indicator industry (SIC) id's
  - Indicator offers
  - Indicator text
  - Local top40

Original data and meta-data

- Sample query
  - Current location
  - Recent merchant descriptions
  - Recent merchant id's
  - Recent SIC codes
  - Recent accepted offers
  - Local top40

Recommendation query

Derived from cooccurrence and cross-occurrence analysis

Me, Us

- Ted Dunning, Chief Application Architect, MapR
  Committer PMC member, Mahout, Zookeeper, Drill
  Bought the beer at the first HUG

- MapR
  Distributes more open source components for Hadoop
  Adds major technology for performance, HA, industry standard API's

- Info
  Hash tag - #mapr
  See also - @ApacheMahout @ApacheDrill
  @ted_dunning and @mapR