Who’s Ready for BigData!!

Istanbul - SoLoMo Presentaion - Fall 2012

Your Data is Not Big

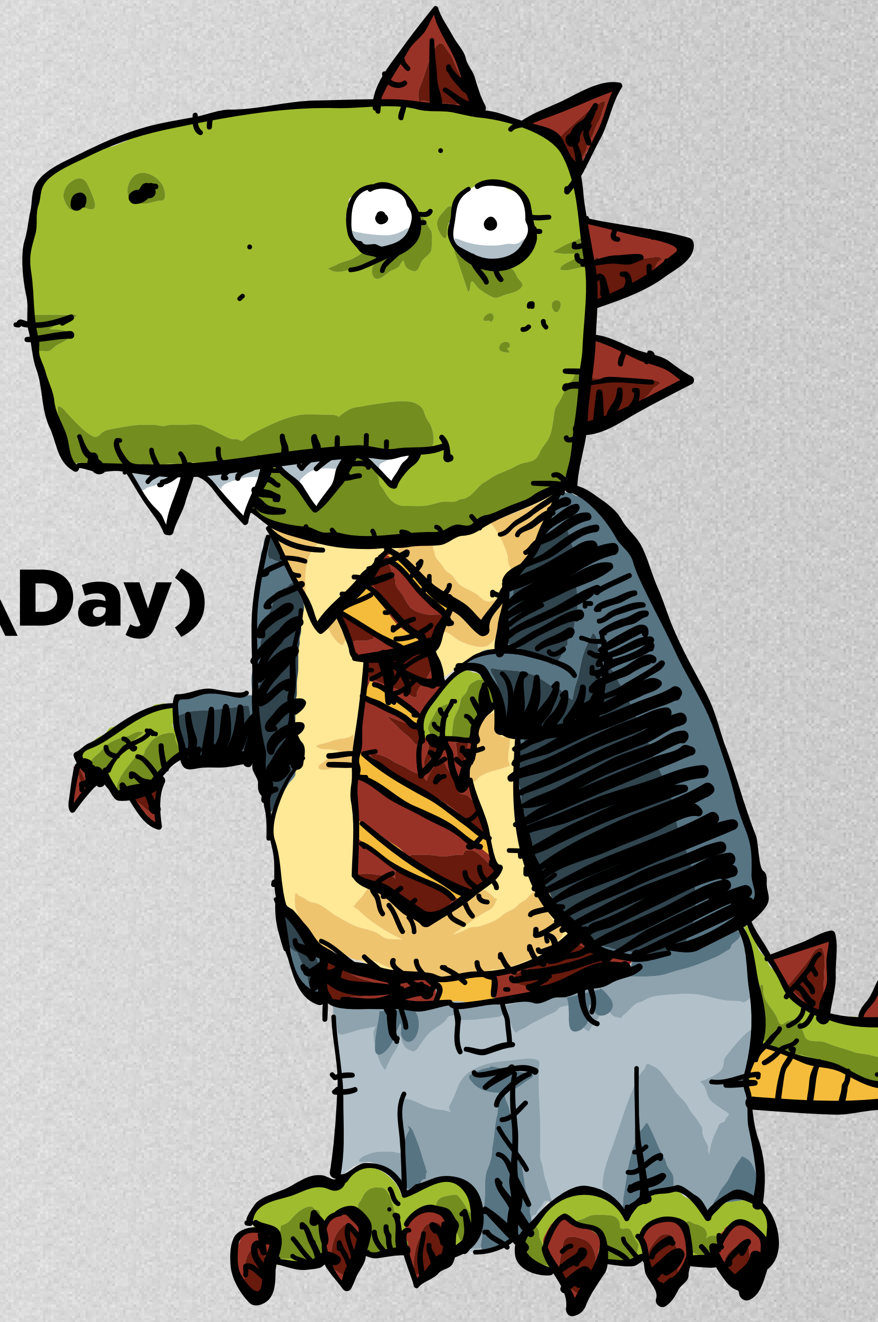
And Why You Don’t Want It To Be...

A hairsplitting dialogue on the biggest technological sea changeto NOT hit your company since Y2K

Who Am I?

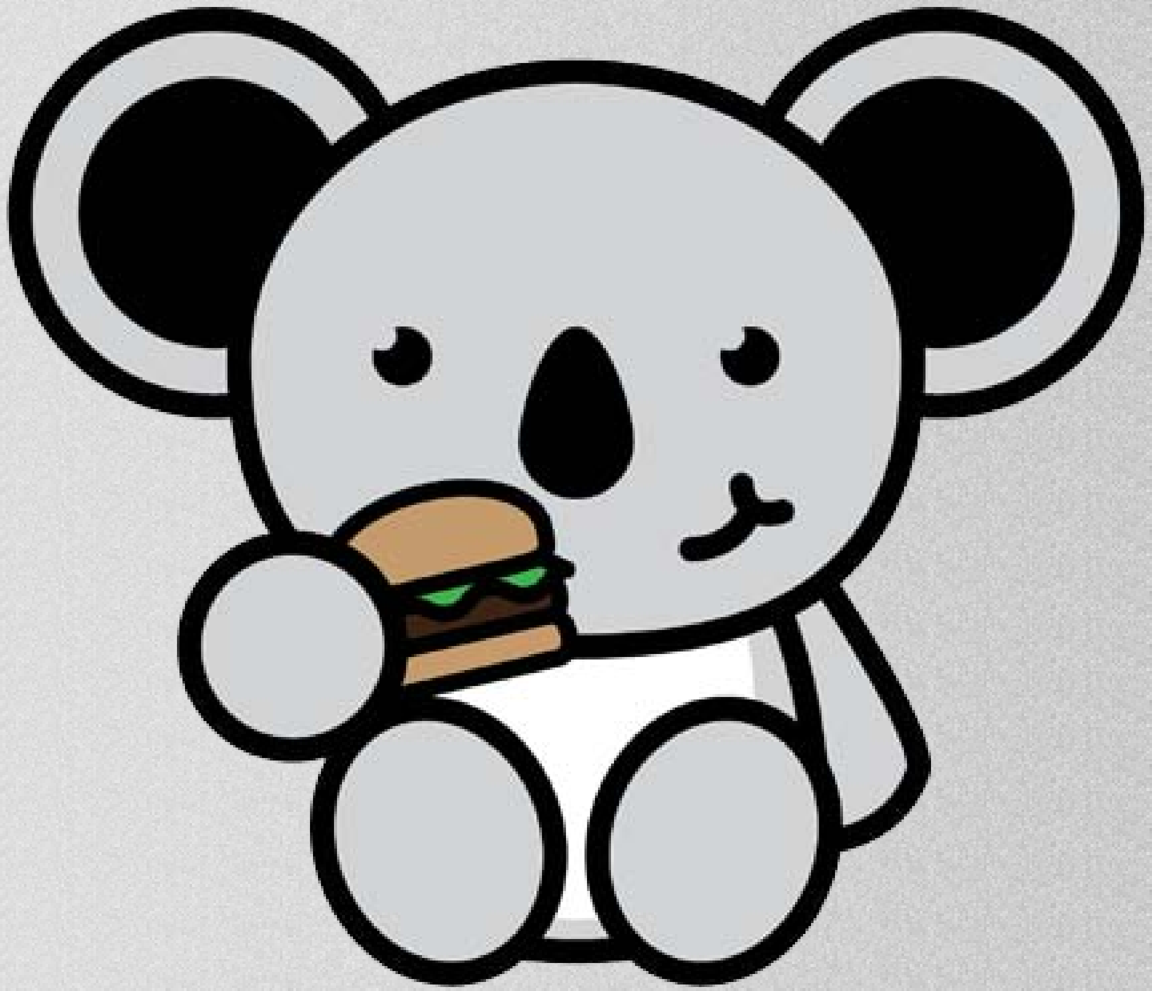
Tim Shea - Follow me @sheanineseven

- Data Scientist and Software Developer
- Ad Agency Guy (Razorfish, Universal, TBWA\Chiat\Day)
- Founder and CTO of WhatsGood.com
- Contrarian/Old Man/Dinosaur
- Pretty Good Guy



What is What’sGood?

Klout/PageRank for Restaurant Menu Item’s



Allow people on-the-go:

- Find food
- Follow tastes
- Collect rewards

We use BigData and NLP to determine

- Vegan
- Gluten Free
- Calories
- Popularity

Who Are You?

Any Programmers or DBA’s?

Anyone wrestling with BigData problem today?

Anyone *never* heard of BigData before?

What is BigData?

The 3 V’s:

1. **Volume**

2. **Velocity**

3. **Variety**

Why is Big Bad?

"Umm...Small Data, please?" -- Everyone

The Bad

Bottom line:

Big is Unweildy

Big is Unstructured

Big is Constantly Changing

What if your "stuff" was literally larger than any single array of disks, RAM, CPU available in the world today?

Big Use Cases

CERN - 200 PB DB, 200MB/sec deemed unusable

Hubble Space Telescope - Produces ~120GB per week!

Human Genome Project - 3.2GB per human

HFT & Algo Trading - 1000’sTB, Billions of “ticks”/day

Twitter - 400MM Tweets/Day

Why is Big Good?

Think Stats: Large data sets eliminate anomolies.

Mining Insights: This customer is pregnant!

Discovering Patterns: “Customers also bought…”

The Fine Print:

Even entire Twitter Firehose or last 30 years of stock market tick data still cannot predict the future.

One is a representation of reality.

The other is only a tiny snapshot of reality.

Examples

Nate Silver - The 2008/2012 Presidential Election

Alpha Genius - Semantic Analysis of Social Media to Trade

Drew Conway - Shades of TIME

Why am I only hearing about this now?

2012

A “New” “Problem”

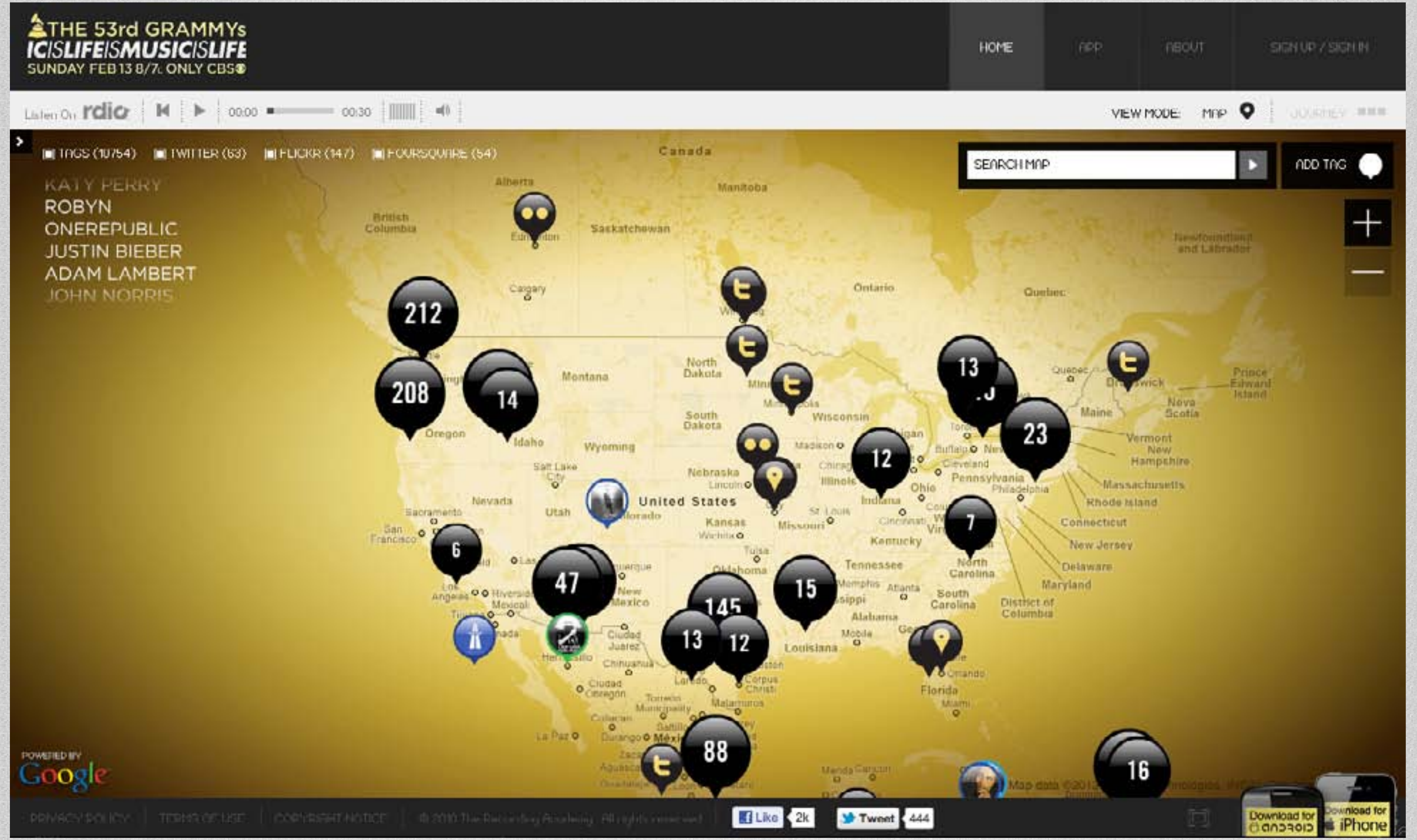
2002

Wildly successful company => Database grows very large, maybe very quickly, and potentially is seen by lots of people.

2012

Day One => Companies start with Millions/Billions/Trillions of records, sometimes growing at massive speeds, with tremendous traffic.

The GRAMMY's


THE 53rd GRAMMYs
ICISLIFEISMUSICISLIFE
SUNDAY FEB 13 8/7c ONLY CBS
HOME
APP
ABOUT
SIGN UP / SIGN IN
Listen On rdio
00:00
00:30
VIEW MODE: MAP
JOURNEY
TAGS (10754)
TWITTER (63)
FLICKR (147)
FOURSQUARE (54)
SEARCH MAP
ADD TAG
KATY PERRY
ROBYN
ONEREPUBLIC
JUSTIN BIEBER
ADAM LAMBERT
JOHN NORRIS
212
208
14
13
23
12
7
6
47
145
15
13
12
88
16
POWERED BY Google
PRIVACY POLICY
TERMS OF USE
COPYRIGHT NOTICE
Like 2k
Tweet 444
Download for ANDROID
Download for iPhone

Back in the Day....

A little SQL

A little Analytics or Stats

A little Business Sense

You could be a very powerful “BI Programmer”

But Today…

A background in AI/ML/Linguisitic Programming

A “full-stack” understanding of your DB vendor

The entire suite of tools from your ecosystem

A very solid grasp of Statistics

2012: The Data Scientist

Lives at the intersection of:

Computer Science, Stats, Business Development

Classifieds <3 BigData?

BigData-By-Design?

Think “Big”

Think Algorithmically

Think like some someone is posied to eat your lunch

Think!

What does my business look like if I had the power of:

- Big-ness
- AI
- Insights
- Recommendations
- Personalization

At my fingertips

Really Quick Dive....

Why Does
So + Mo + Lo = Big

Social: The social graph is terribly difficult to model. "Virality" is terribly difficult to scale.

Mobile: Low power, "always on" devices serving context to users on the go. All mobile "exhaust" as an important data point to capture.

Local: Hundreds of Millons/Billions of POS. Constantly changing. Think Apple Maps.

Local: The 4th Dimension

Volume+Velocity+Variety+ .... Harmonization??

Steps for keeping good data:

1. Download 100MM Locations 50GB
2. SSIS Package to Import
3. Deal with Loss
4. Deal with Obsolete Data
5. Allow UGC Submissions
6. Refresh Data from data source
7. Merge Original/UGC/Refresh
8. Switch Data Providers

.....

9. Leap off Bridge....

Again with the What'sGood

350mm records

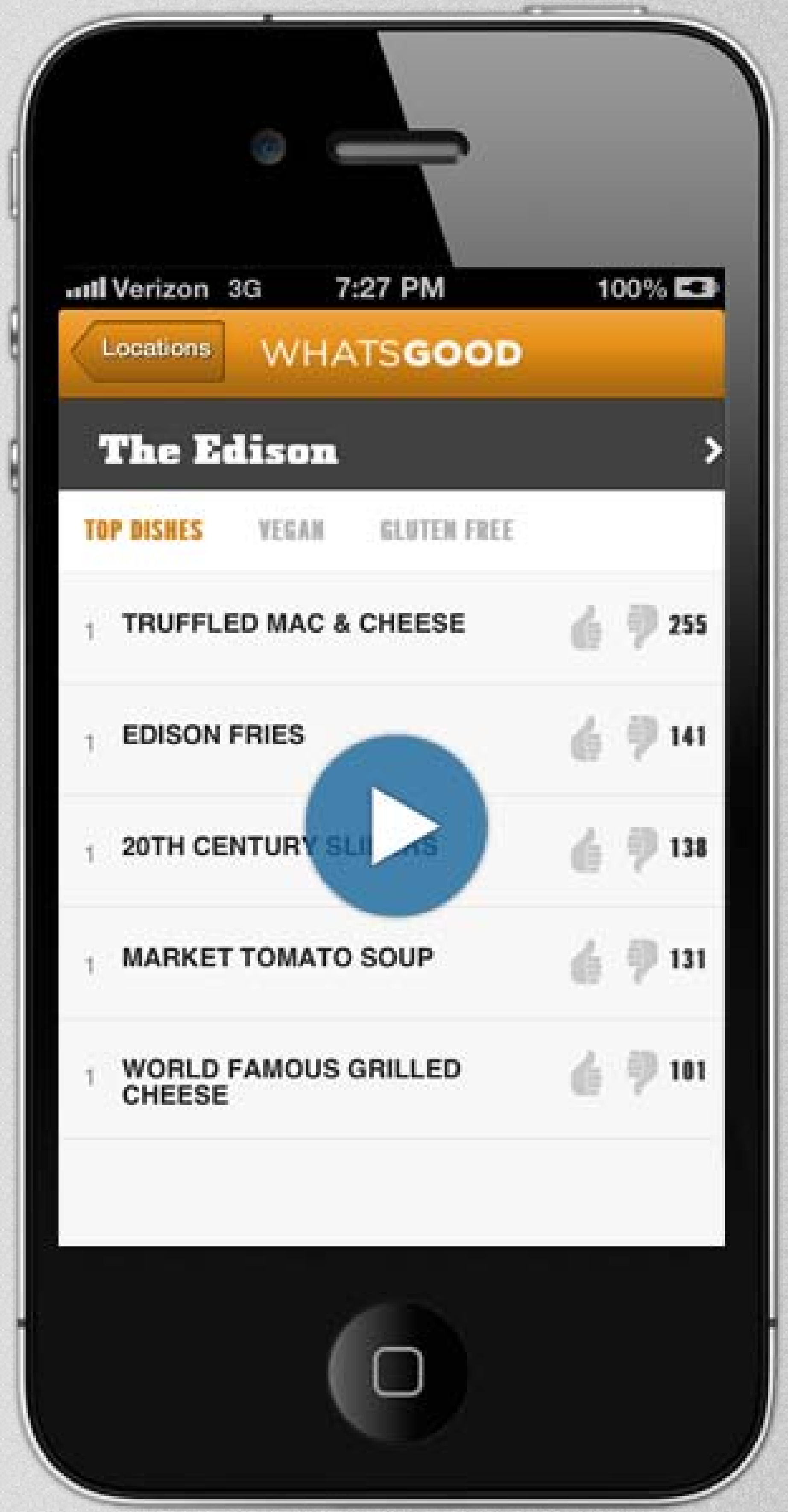
7 NLP classifiers

~20 votes per record

350*7 + 350*20 = 9.4B records

Avg user session 5 searches

47B records traversed



1	TRUFFLED MAC & CHEESE	255
1	EDISON FRIES	141
1	20TH CENTURY SLI[illegible]RS	138
1	MARKET TOMATO SOUP	131
1	WORLD FAMOUS GRILLED CHEESE	101

Haversine Search

“Get All Italian Restaurants in My Vicinity”:

```
select top(@_num_records) *,
3956 * 2 *
ASIN(
SQRT(
POWER(SIN((@orig_lat - abs(latitude)) * pi()/180 / 2), 2) +
COS(@orig_lat * pi()/180) *
COS(abs(latitude) * pi()/180) *
POWER(SIN((@orig_lon - longitude) * pi()/180 / 2), 2)
)
) as distance
from LocationTable
having distance < @distance
where name like ‘%italian%’
order by distance
```

What’s Irritating

We can’t leverage:

Hadoop: We had to build our own parallelisation platform.

NoSQL: We can’t do aggregates. Tools are immature.

Cloud: Shared CPU, “Noisey Neighbors”, can’t scale “Up”.

The Backlash…

Medium Data?!?

So...

What if I have a Medium Data problem?

No worries:

You can still leverage the BigData ecosystem.

NoSQL

MongoDB, CouchDB, RavenDB, etc

Reality is: Not ready for Prime Time!

So, how do I query the database?
It's not a database. It's a key-value store!
Ok, it's not a database. How do I query it?
You write a distributed map reduce function in Erlang!
Did you just tell me to go fuck myself?
I believe I did, Bob.

Hadoop

Getting value from a new product, shouldnt be like getting your PhD.

“My CEO went to a BigData conference and all I got was this lousy Distributed Key/Value framework for parallelising MapReduce jobs over muti-node commodity hardware in the cloud.”

The Cloud

Vertical is the new horizontal & Expensive is new inexpensive

Is there a Panacea?

C'mon SON!

Play Nice!

Sid Anand - LinkedIn

“Many of the NoSQL vendors view the ‘battle of NoSQL’ to be akin to the RDBMS battle of the 80s:

A winner-take-all battle.

In the NoSQL world, it is by no means a winner-take-all battle.

Distributed Systems are about compromises.”

Some Tools

Public BigData Sets

http://aws.amazon.com/datasets

Enron Email Database (1.2MM Emails)

The Cannabis Sativa Genome - “Chemdawg”

Marvel Universe Social Graph (tab delimited file)

Daily Global Weather Measurements 1929 - 2009

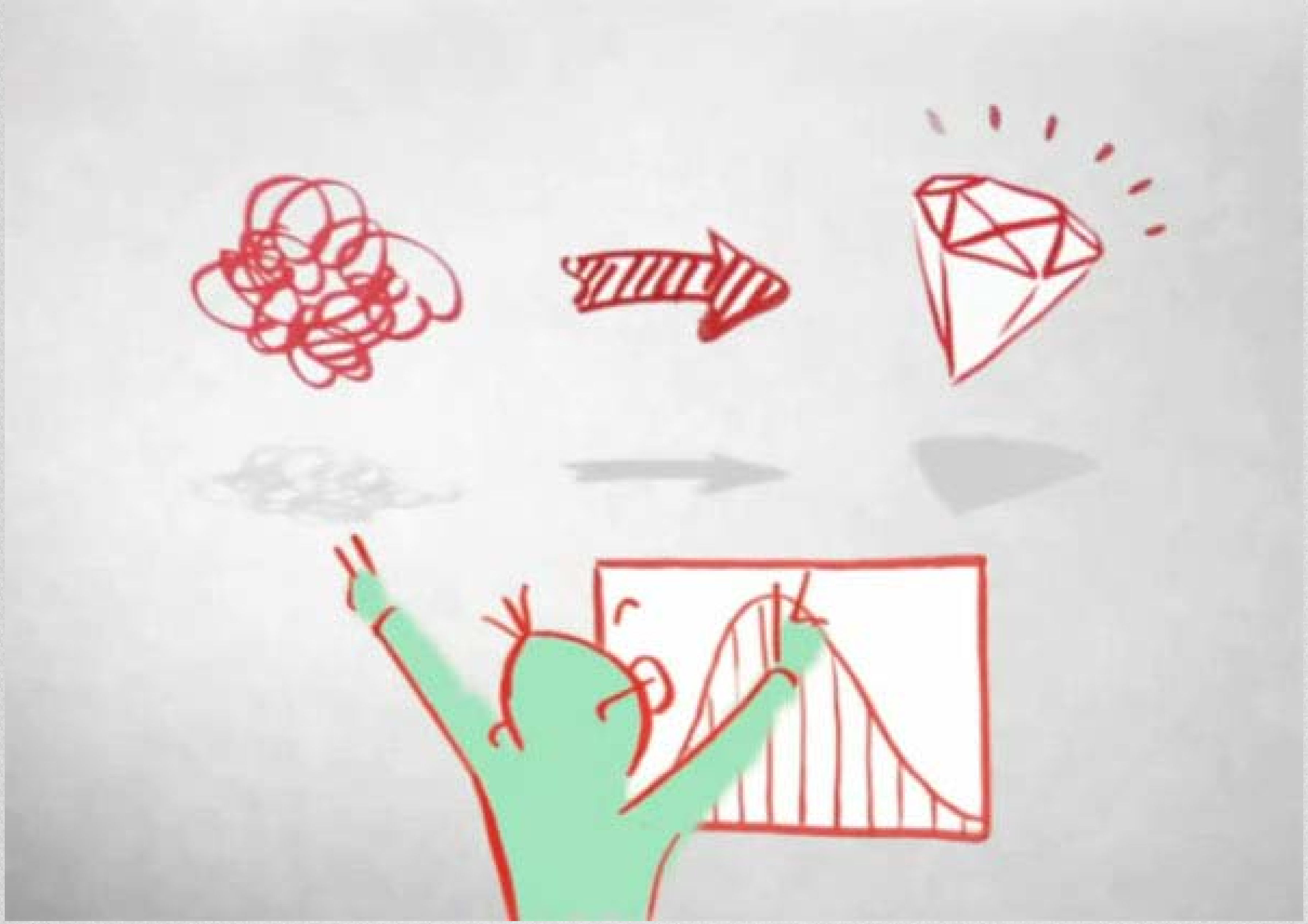
Entire English Wikipedia Extraction

Python’s NLTK

- Pre-Parsed, 60 Corpora Database
    - Wall Street Journal
    - Shakespeare’s Works
    - The Book of Genesis
    - Every State of the Union Address
- Really cool Python SDK
- Training your own Linguistic Applications

Google Refine

Cleaning, searching, & sorting big data sets quickly



Cloudera Hadoop

Download & Practice with a psuedo-cluster on your MacBook Air

Some Final Thoughts…

WHAT IF THE MAYAN CALENDAR ENDS IN 5105
AND WE'VE JUST BEEN HOLDING IT UPSIDE DOWN

Don’t Go Crazy

1. **Work with Small Data sets where you can.**

2. **No magic bullets. All compromise.**

3. **ER Modeling - Don’t Stop Believing.**

4. **Right tool, right job.**

I hope that helps!

Tim Shea

@sheanineseven

tim@whatsgood.com