Big Data

SU 5050

LECTURE 3

JESSICA L. MCCARTY, PH.D.

What is Big Data?

Data sets that grow so large that they become awkward to work with using on-hand database management tools? (Wikipedia)

What is Big Data?

- distributed file systems
- NoSQL databases
- grid computing, cloud computing
- MapReduce and other new paradigms
- large-scale machine learning

MapReduce = programming model and associated implementation for processing & generating large data sets with parallel, distributed algorithm on a cluster.

What is Big Data?

from “Big Data and the Web: Algorithms for Data Intensive Scalable Computing” PhD Thesis, Gianmarco De Francisci Morales

What is Big Data?

- buzzword?
- bubble?
- gold rush?
- revolution?
- funding fad?
    - DARPA XDATA project, March 2012

Who's profiting?

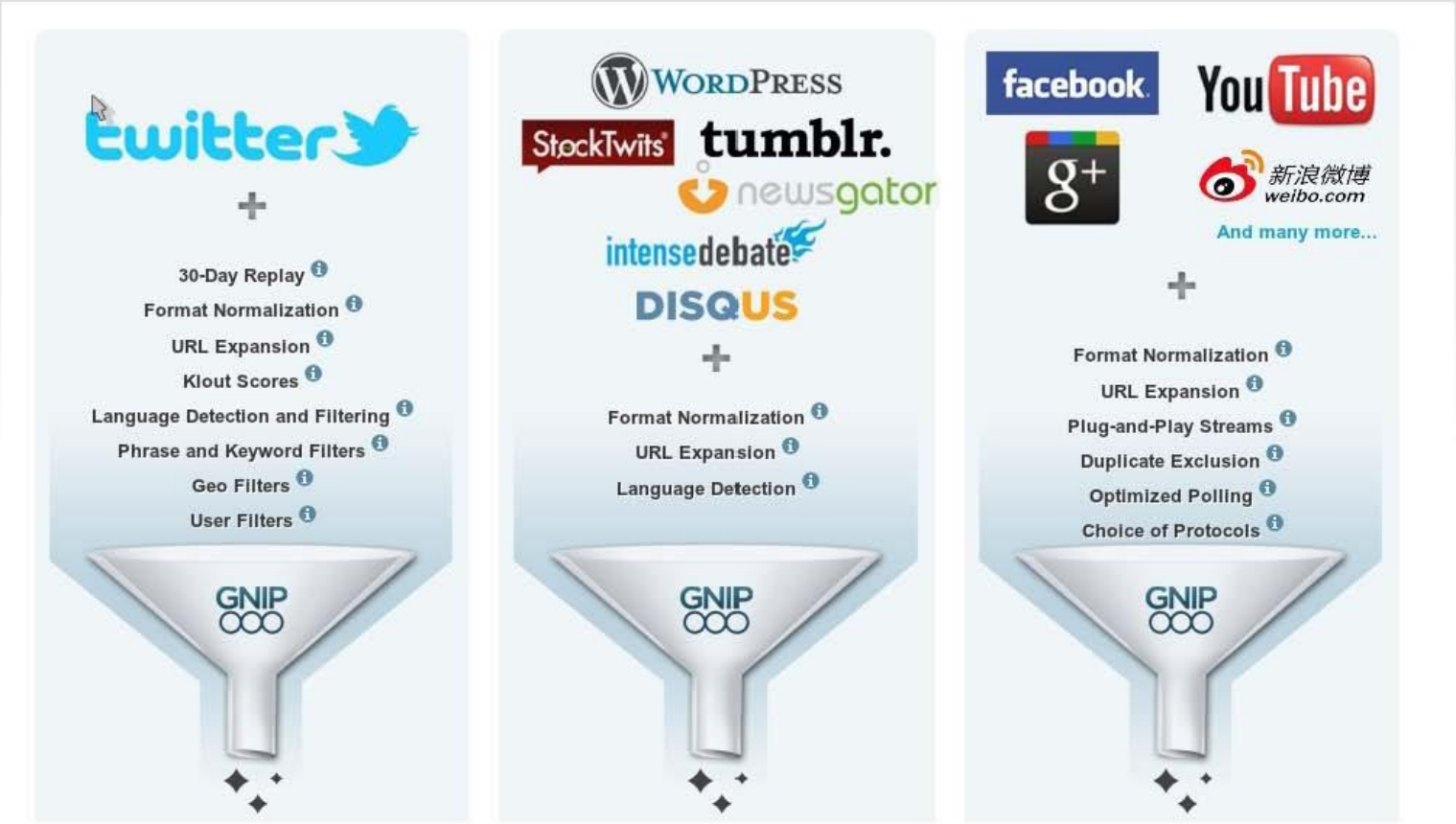
RapLeaf


```
{
  "age":"21-24",
  "gender":"Male",
  "interests":{
    "Blogging":true,
    "High-End Brand Buyer":true,
    "Sports":true,
  },
  "education":"Completed Graduate School",
  "occupation":"Professional",
  "children":"No",
  "household_income":"75k-100k",
  "marital_status":"Single",
  "home_owner_status":"Rent"
}
```


Now **intelligence.towerdata.com**; help marketers target consumers via mail, email, social media

- 10 TB of user data
- all computation done using local Hadoop cluster

Gnip



- "grand central station" for social web streams
- aggregates several TB of new social data **per day**
- all data stored on Amazon S3

Acquired by Twitter in April 2014; **blog.gnip.com**

Climate Corporation

- 14 TB of historical weather data
- all computation done using Amazon EC2
- 30 technical staff, including 12 PhDs
- 10,000 sales agents

climate.com;
Direct competitor in Ann Arbor - farmlogs.com/

FICO

- 50+ years of experience doing credit ratings
- transitioning to predictive analytics

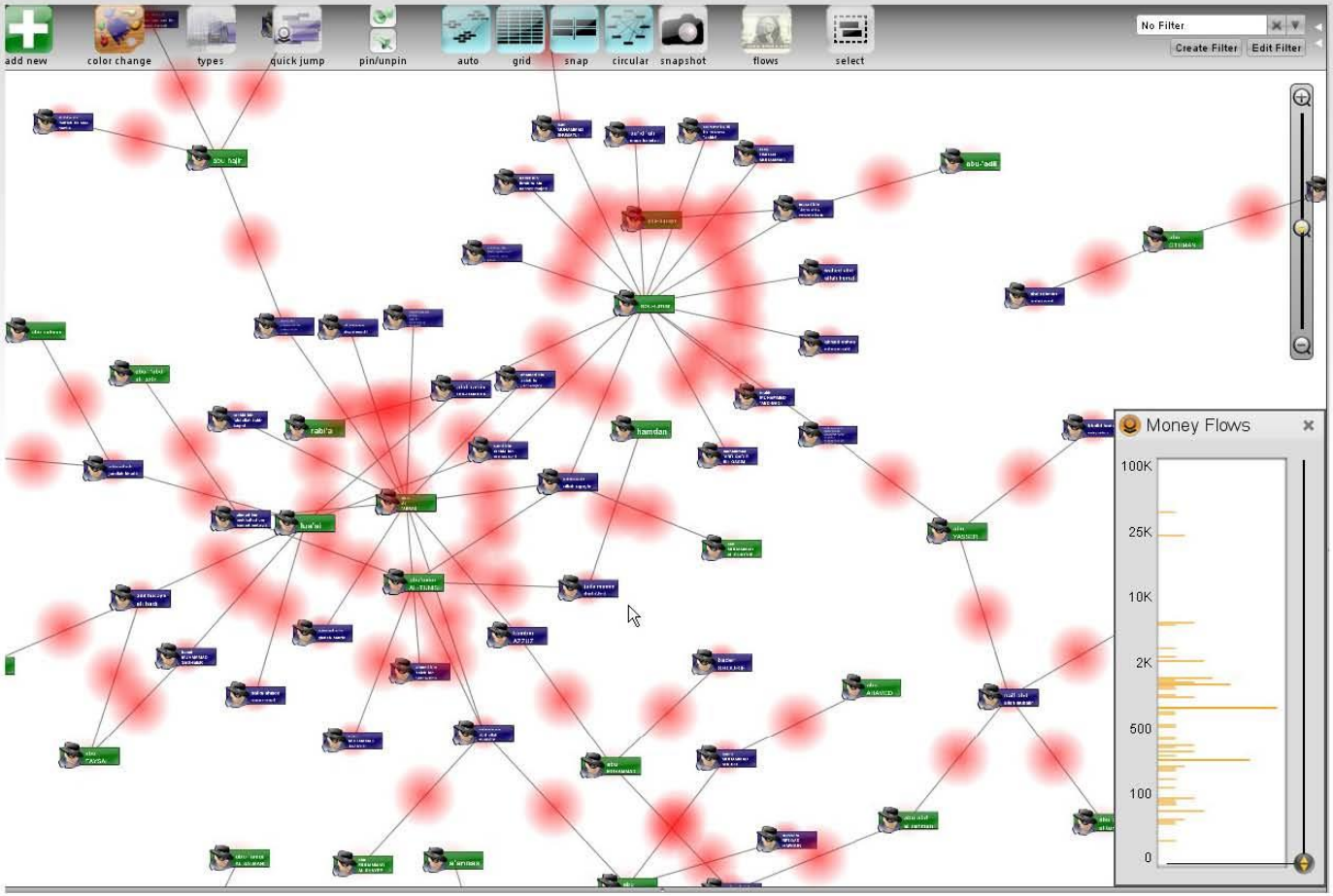
fico.com;
New direction is financial habits of the Millennials

Cloudera

- employs many of the original Hadoop developers
- now facing many competitors

cloudera.com;
First to offer big data and analytics training

Palantir Technologies



- data analysis, exploration tools for government, finance
- 0-$2 billion in 5 years

palantir.com ; DC, NYC, Palo Alto

BIG DATA

Companies Move On From Big Data Technology Hadoop

By QUENTIN HARDY JUNE 15, 2015 12:30 PM 7 Comments

Email

Share

Tweet

Save

More

There is increasing evidence that Hadoop — one of the most important technologies of the past several years for big data analysis —is not keeping up with the world that created it.

On Monday, IBM, which has championed Hadoop and put it at the center of its big data strategy, announced it is working on a faster data-processing engine, called Spark.

IBM is working with a company called Databricks, which also announced the general release of its product on Monday, at a conference in San Francisco on Spark. The product was previously used by just a few customers.

Additionally, a senior executive at Cloudera, probably the largest Hadoop company, said Cloudera is prepared to see key parts of Hadoop diminish in importance, and was increasingly distributing Spark.

In the broadest terms, these shifts make sense. Hadoop is an open-source architecture that drew heavily on work published by Google in 2004, in particular two different papers on managing, processing and generating very large data sets -- the kind of stuff a company

http://bits.blogs.nytimes.com/2015/06/15/companies-are-moving-on-from-big-data-technology-hadoop/

The modern data scientist

- Engineer
  - collect & scrub disparate data sources
  - manage a large computing cluster
- Mathematician
  - machine learning
  - statistics
- Artist
  - visualise data beautifully
  - tell a convincing story

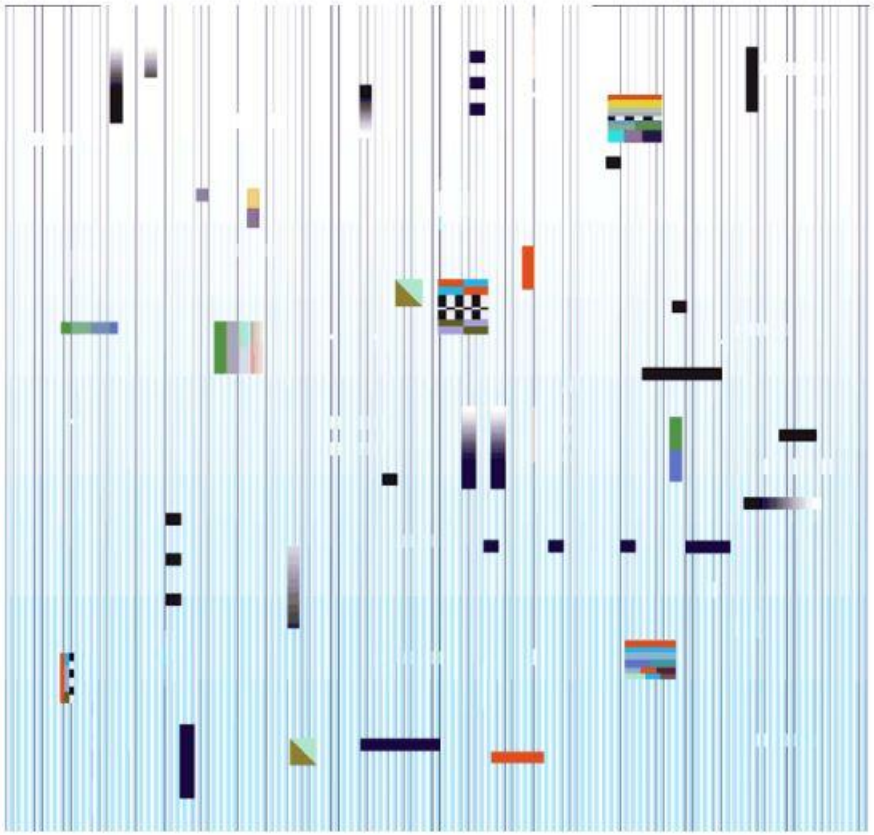
Learning to See Data

By BENEDICT CAREY MARCH 27, 2015

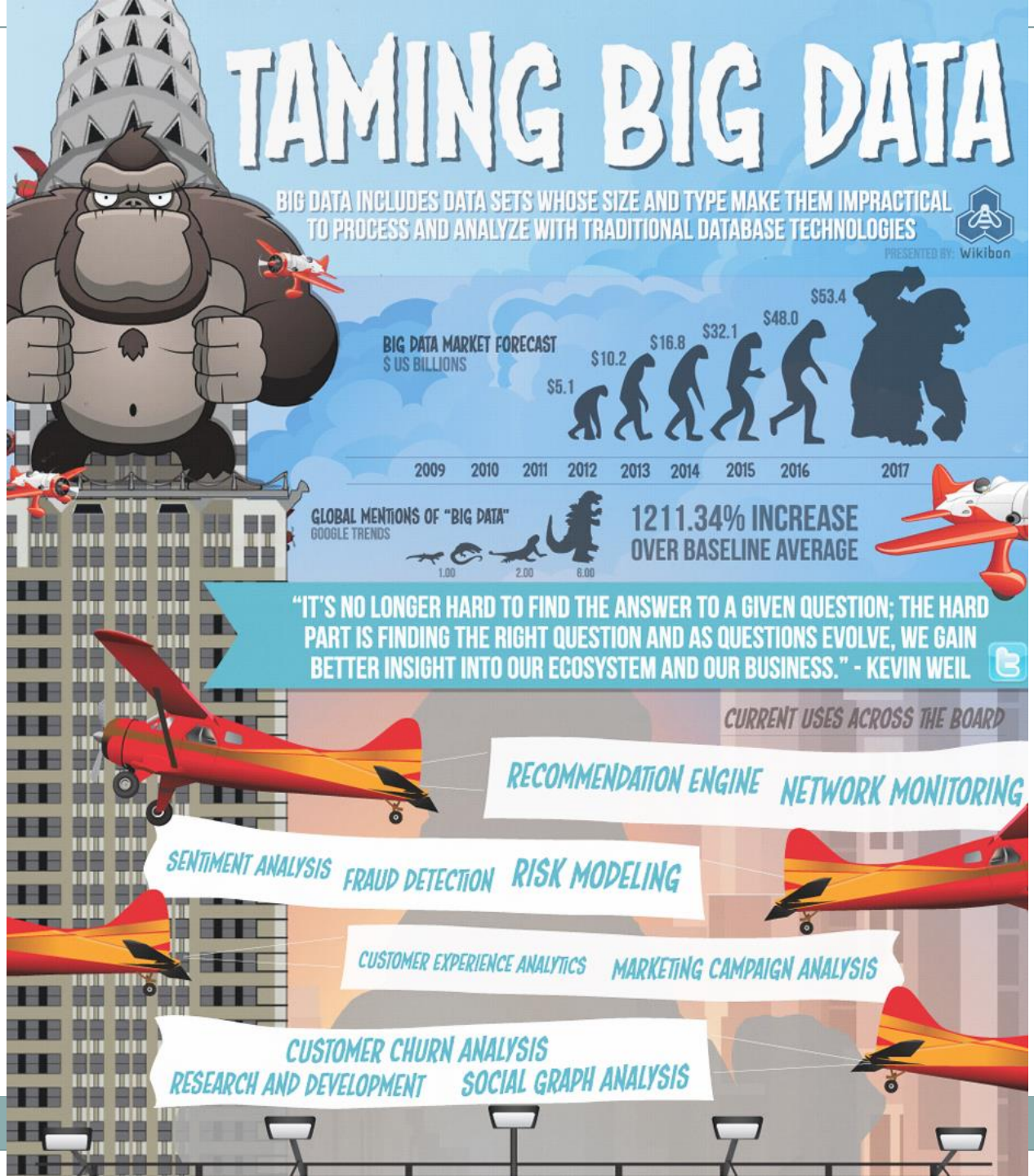
FOR the past year or so genetic scientists at the Albert Einstein College of Medicine in New York have been collaborating with a specialist from another universe: Daniel Kohn, a Brooklyn-based painter and conceptual artist.

Mr. Kohn has no training in computers or genetics, and he's not there to conduct art therapy classes. His role is to help the scientists with a signature 21st-century problem: Big Data overload.

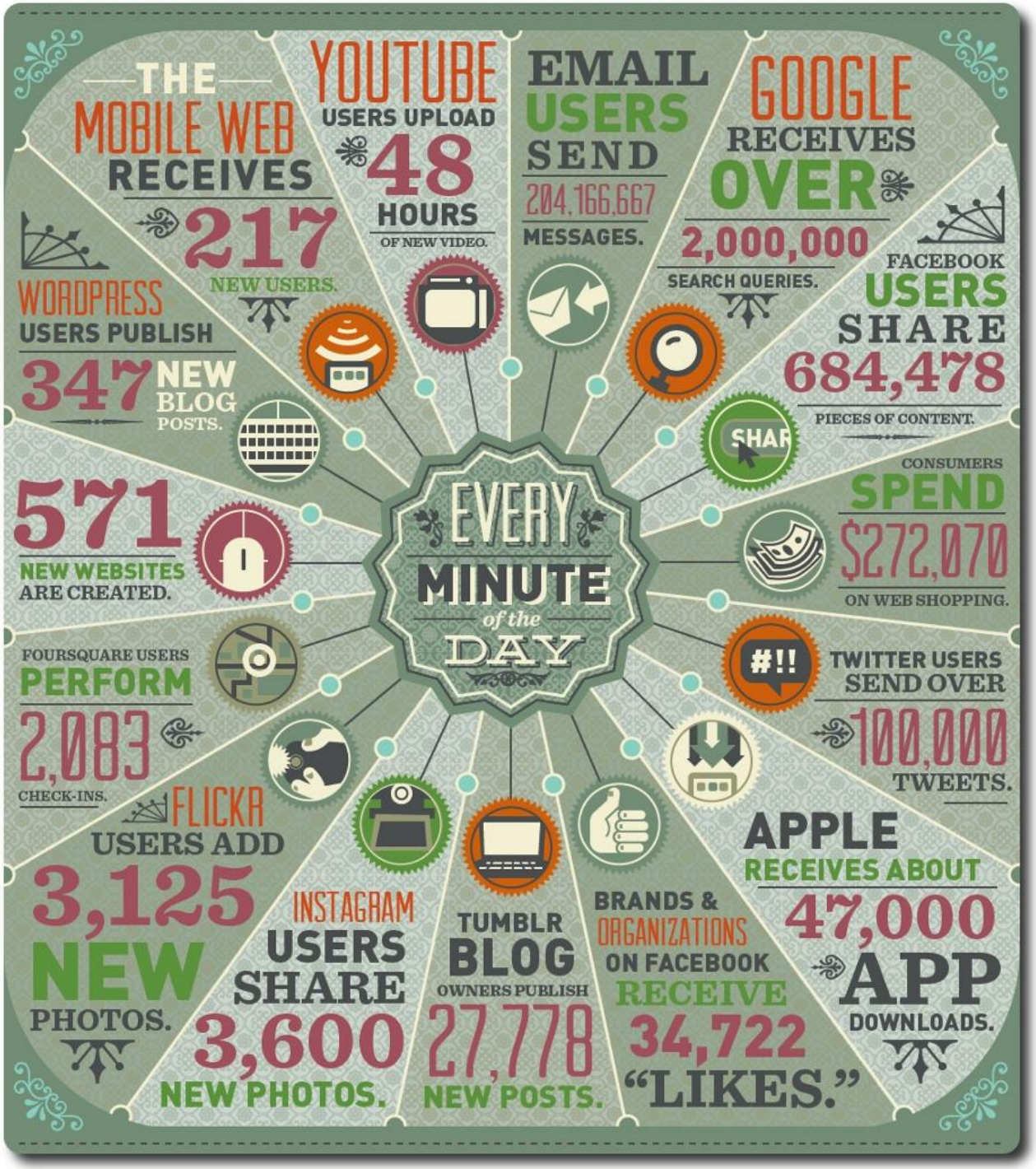
Advanced computing produces waves of abstract digital data that in many cases defy interpretation; there's no way to discern a meaningful pattern in any intuitive way. To extract some order from this chaos, analysts need to continually reimagine the ways in which they represent their data — which is where Mr. Kohn comes in. He spent 10 years working with scientists and knows how to pose useful questions. He might ask, for instance, What if the data were turned sideways? Or upside down? Or what if you could click on a point on the plotted data and see another dimension?



Hvass & Hannibal


TAMING BIG DATA
BIG DATA INCLUDES DATA SETS WHOSE SIZE AND TYPE MAKE THEM IMPRACTICAL TO PROCESS AND ANALYZE WITH TRADITIONAL DATABASE TECHNOLOGIES
PRESENTED BY: Wikibon
BIG DATA MARKET FORECAST
$ US BILLIONS
$5.1
$10.2
$16.8
$32.1
$48.0
$53.4
2009 2010 2011 2012 2013 2014 2015 2016 2017
GLOBAL MENTIONS OF "BIG DATA"
GOOGLE TRENDS
1.00
2.00
6.00
1211.34% INCREASE OVER BASELINE AVERAGE
"IT'S NO LONGER HARD TO FIND THE ANSWER TO A GIVEN QUESTION; THE HARD PART IS FINDING THE RIGHT QUESTION AND AS QUESTIONS EVOLVE, WE GAIN BETTER INSIGHT INTO OUR ECOSYSTEM AND OUR BUSINESS." - KEVIN WEIL
CURRENT USES ACROSS THE BOARD
RECOMMENDATION ENGINE
NETWORK MONITORING
SENTIMENT ANALYSIS
FRAUD DETECTION
RISK MODELING
CUSTOMER EXPERIENCE ANALYTICS
MARKETING CAMPAIGN ANALYSIS
CUSTOMER CHURN ANALYSIS
RESEARCH AND DEVELOPMENT
SOCIAL GRAPH ANALYSIS


IN 60 SECONDS_
1 NEW DEFINITION IS ADDED ON urban dictionary
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WordPress DOWNLOADS
125+ PLUGIN DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
695,000+ facebook STATUS UPDATES
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO!
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
THE LARGEST
New Craigslist Ads
GO-Globe.com
Google
Google Search


THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
EVERY MINUTE of the DAY
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.

The Power of Healthcare Data

The Body as a Source of Big Data

Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.

Access to electronic patient data beyond the desktop

3D MRI 150MB

MAMMOGRAMS 120MB

X-RAY 30MB

3D CT SCAN 1GB

0.5MB is generated

There are currently 425K telehealth providers in the U.S.[7]

It is estimated that by 2015, the average hospital will generate

665TB of data[1]

Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations[4]

36.6M

Total admissions in U.S. registered hospitals, according to the American Hospital Association[5]

Medical image archives are increasing by

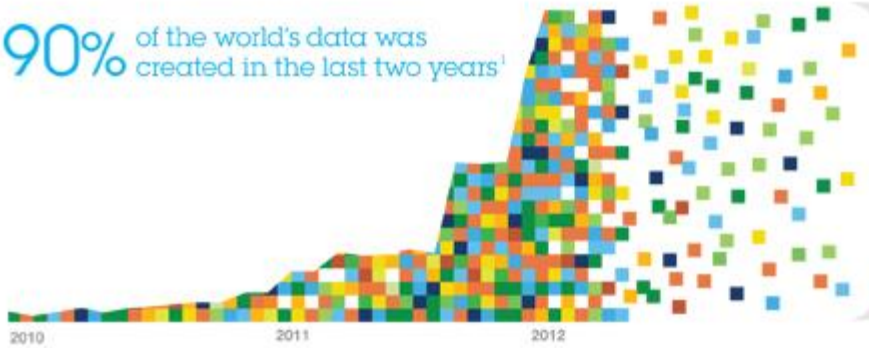
20-40% annually[6]

Today,

80% of data is unstructured, such as images, video and email.[3]

Why Big Data?

90% of the world's data was created in the last two years[1]



80% of the world's data today is unstructured

1 in 2 business leaders dont have access to data they need[2]

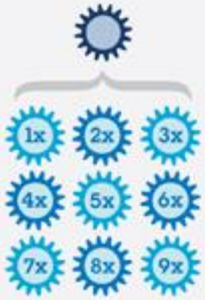
83% of CIOs cited BI and analytics as part of their visionary plan[3]

54% of companies use analytics for competitive advantage[3]

IBM System x Reference Architectures

Delivering business insights with speed and flexibility

Enhance performance



Achieve up to 9x the performance of open-source Hadoop with enhanced scheduler[8]

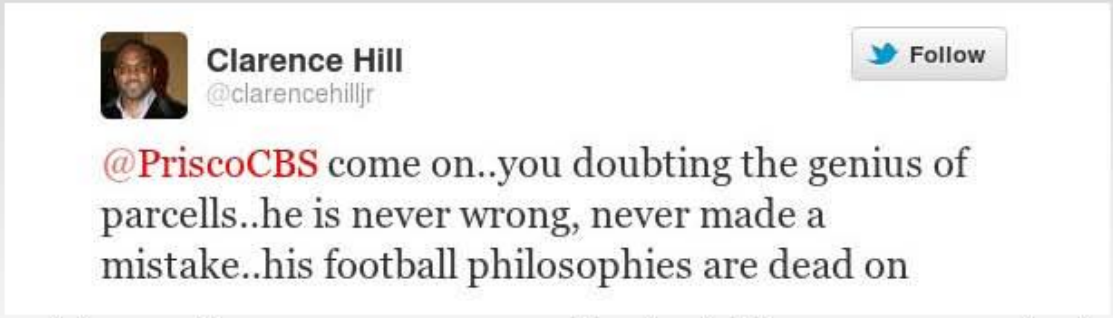
Simplify deployment

Utilize IBM System x hardware to leverage IBM software and services in standardized blueprints created from IBM enterprise experience and expertise.

Improve platform flexibility

Interpret vast amounts of heterogeneous data in real time[4]

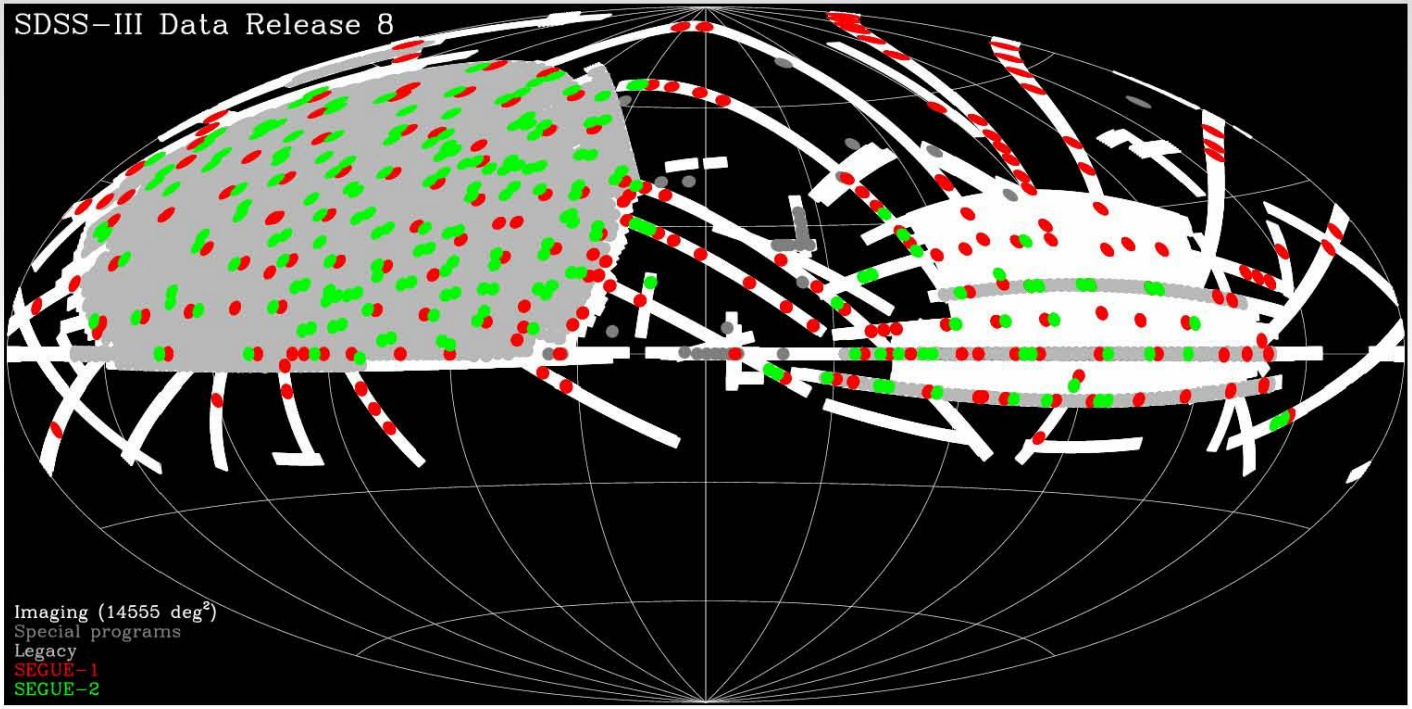
Twitter “Firehose”




```
{ "in_reply_to_status_id_str":null, "retweet_count":0, "favorited":false, "text":"New iPad vs iPad 2: which should
you choose? http://t.co/EpygqtlE RedsnOw 0.9.10b6 i4Siri Proxy Server For Spire GEVEY Ultra S WP7
_83", "in_reply_to_user_id_str":null, "in_reply_to_status_id":null, "created_at":"Mon Mar 19 22:22:32 +0000
2012", "geo":null, "in_reply_to_user_id":null, "truncated":false, "source":"<a href=\"http://google.com\"
rel=\"nofollow\">Tech Discovery</a>", "id_str":"181868286009548802", "entities":{ "hashtags":[], "urls":
[ { "indices":
[45,65], "expanded_url":"http://ow.ly/9A4kC", "url":"http://t.co/EpygqtlE", "display_url":"ow.ly/9A4
kC" } ], "user_mentions":
[] }, "contributors":null, "in_reply_to_screen_name":null, "place":null, "retweeted":false, "possibly_sensitive_edi
table":true, "possibly_sensitive":false, "coordinates":null," user":
{ "profile_text_color":"333333", "profile_image_url_https":"https://si0.twimg.com/profile_images/1817878884/ironman
_normal.png", "screen_name":"ricegyeat", "default_profile_image":false, "profile_background_image_url":"http://a
0.twimg.com/images/themes/theme1/bg.png", "favourites_count":0, "created_at":"Thu Feb 09 18:26:24 +0000
2012", "profile_link_color":"0084B4", "verified":false, "friends_count":0, "url":null, "description":"",
"profile_background_color":"C0DEED", "id_str":"487764951", "lang":"en", "profile_background_tile":false, "l
isted_count":0, "contributors_enabled":false, "geo_enabled":false, "profile_sidebar_fill_color":"DDEEF6", "lo
cation":"", "time_zone":null, "protected":false, "default_profile":true, "following":null, "notifications"
:null, "profile_sidebar_border_color":"C0DEED", "name":"rice
gyeat", "is_translator":false, "show_all_inline_media":false, "follow_request_sent":null, "statuses_count":82
53, "followers_count":109, "profile_image_url":"http://a0.twimg.com/profile_images/1817878884/ironman_normal.png",
"id":487764951, "profile_use_background_image":true, "profile_background_image_url_https":"https://si0.twimg.com
/images/themes/theme1/bg.png", "utc_offset":null }, "id":181868286009548802}
```


350 M tweets/day x 2-3Kb/tweet ≈ 1 TB/day

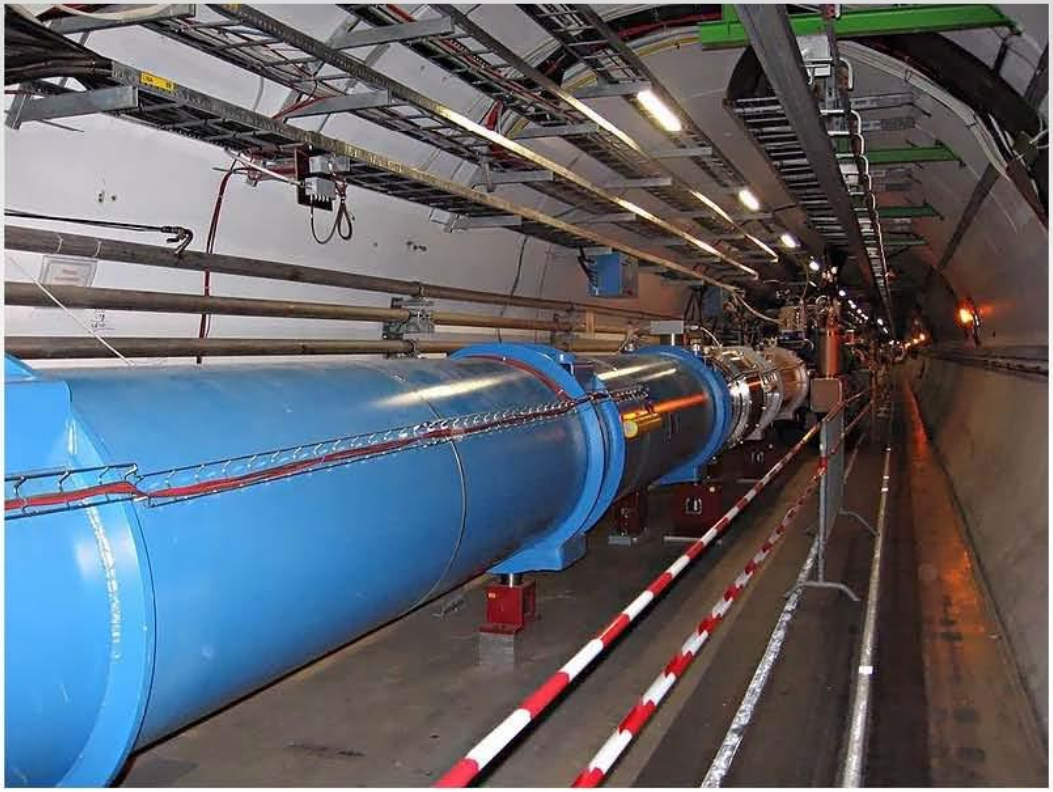
Sloan Digital Sky Survey



- 35% of the sky mapped
- 500 million objects classified
- 50 TB of data available

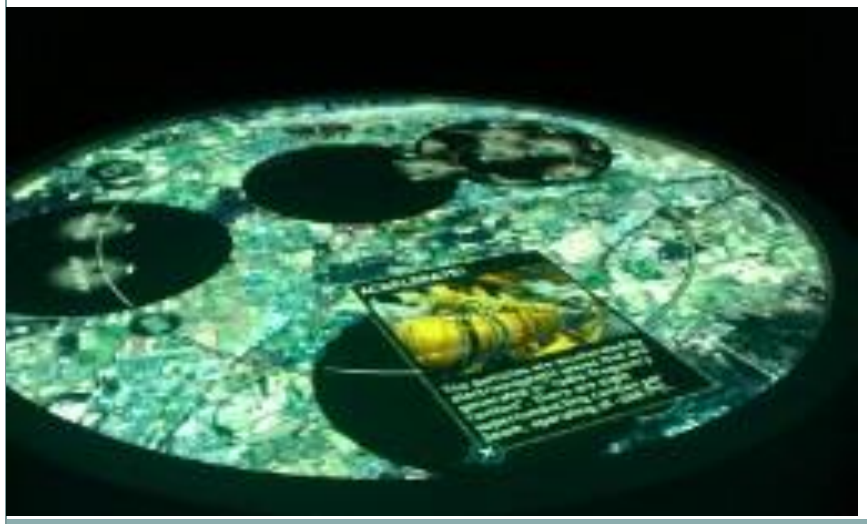
sdss.org;
Anyone can get access, on Data Release 10

Large Hadron Collider



- 15 PB of data generated annually
- mostly stored in Oracle databases (SQL)

My Trip to CERN in March 2013



Human genome

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

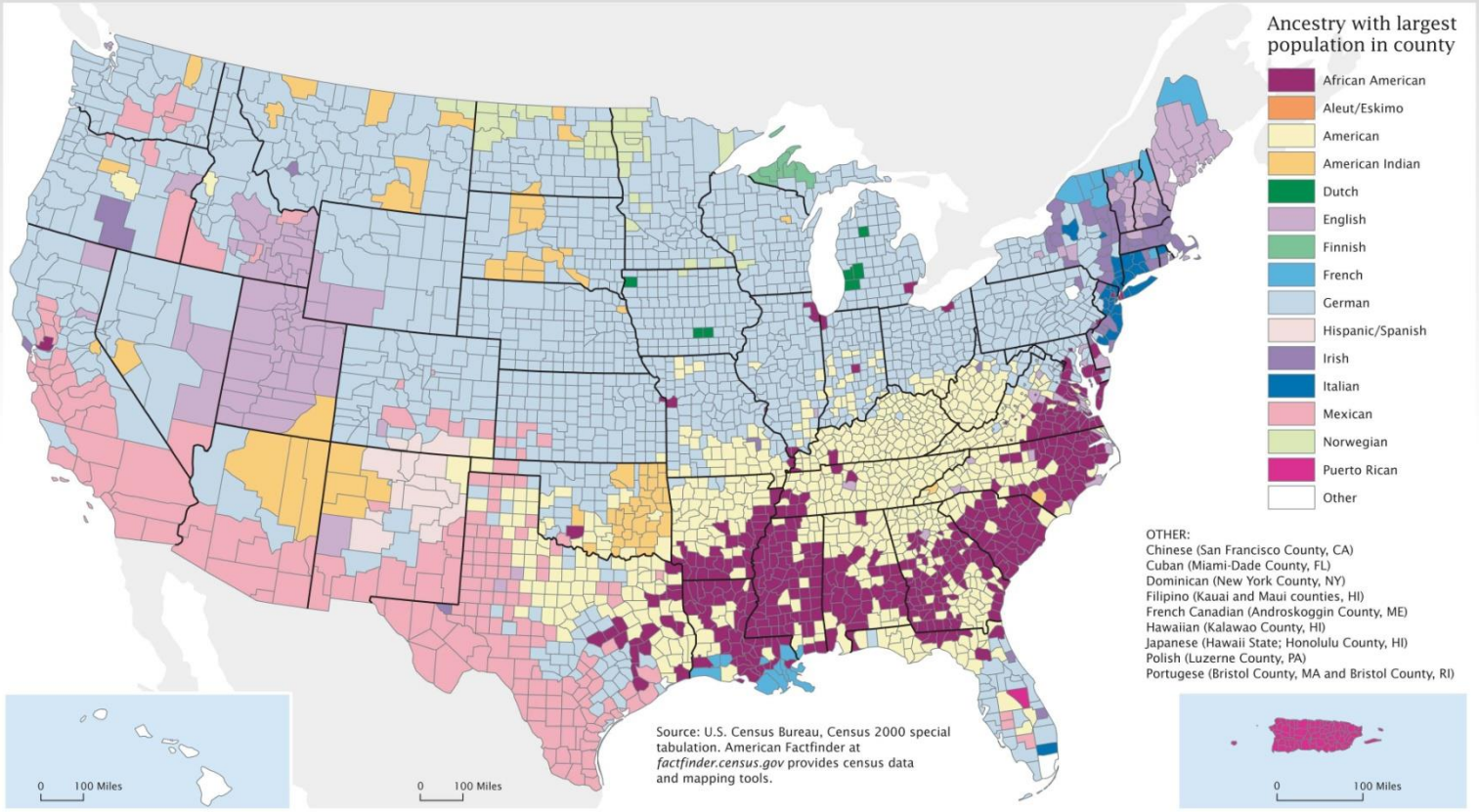
Human
GRCh37

Mouse
NCBIM37

Zebrafish
Zv9

- humans and 50 other species
- "only" 250 GB

US census data (2000)



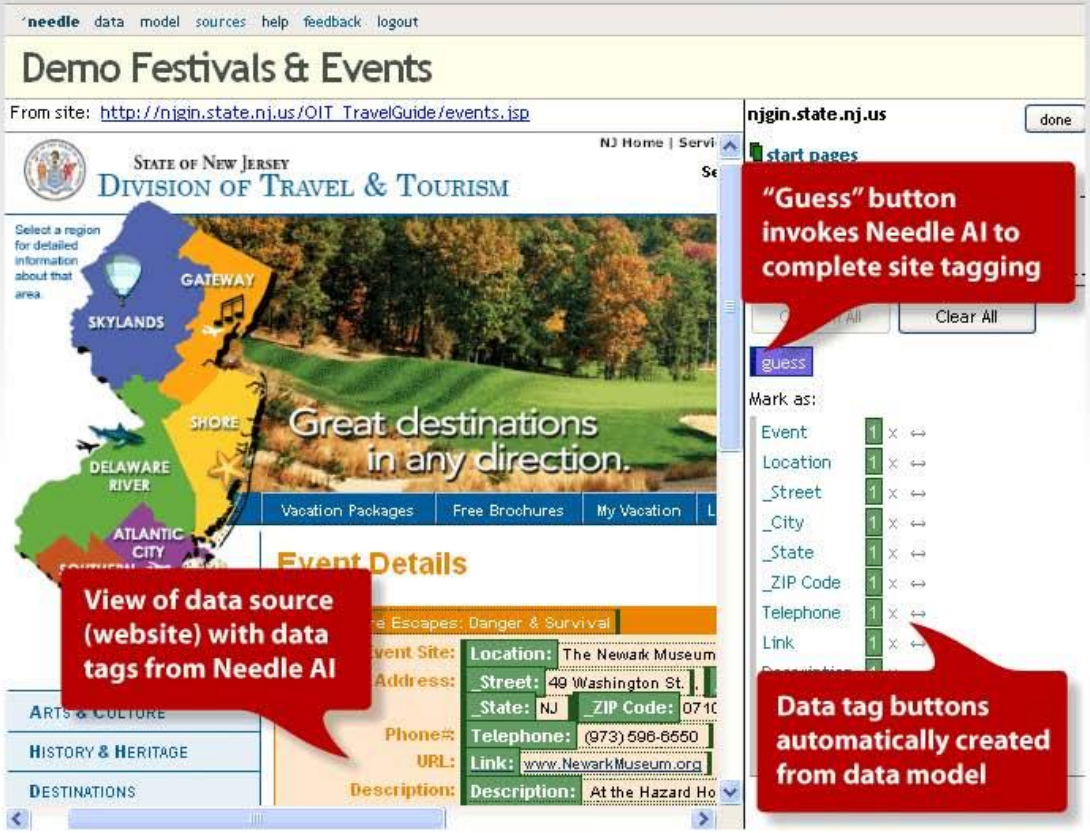
- detailed demographic data for small localities
- “only” 200 GB

Roll your own big data

- crawling as a service
- web sites, social profiles, product listings, etc.
- free accounts offer crawls of up to 10k URLs

80legs.com; *Datafiniti* search engine (businesses, people, products)

Roll your own big data



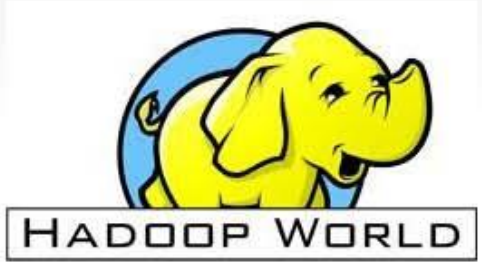
- Needlebase: graphical tagging of website structure

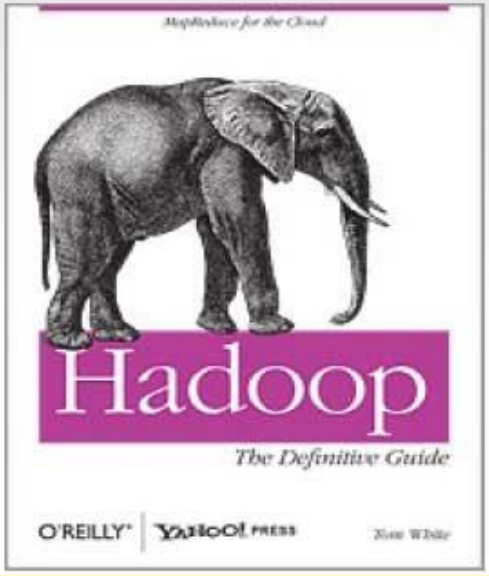
Newer GUI web scraping competition: **www.reporterslab.org/needlebase-dead/**

Roll your own big data

- Boilerpipes: remove “clutter” from web pages
  - Metadata, JavaScript, etc.
- Google Refine: clean up human-entered data
  - fix common typos, spacing, etc.
- NLPToolkit: simplify natural language
  - stem words, replace synonyms
- Lucene: index terms for text search
- Amazon MTurk: human analysis

The Hadoop Industry

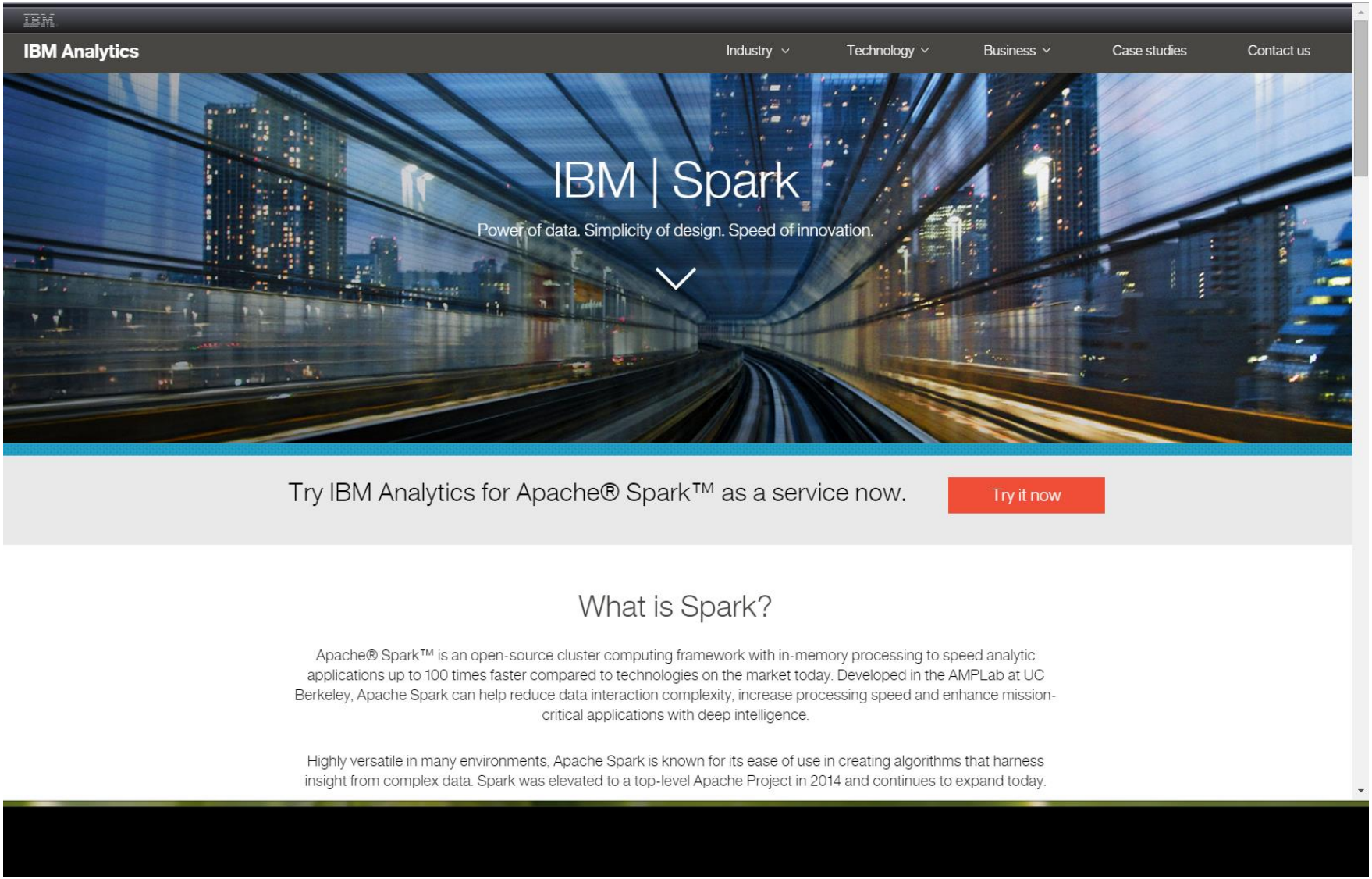




Large scale processing on data clusters: **hadoop.apache.org**

siliconangle.com/blog/2014/09/08/what-you-missed-in-big-data-hadoop-bandwagon-gets-even-more-crowded/

www.zdnet.com/teradata-acquires-hadoop-consulting-firm-think-big-analytics-7000033220/


IBM
IBM Analytics
Industry
Technology
Business
Case studies
Contact us
IBM | Spark
Power of data. Simplicity of design. Speed of innovation.
Try IBM Analytics for Apache® Spark™ as a service now.
Try it now
What is Spark?
Apache® Spark™ is an open-source cluster computing framework with in-memory processing to speed analytic applications up to 100 times faster compared to technologies on the market today. Developed in the AMPLab at UC Berkeley, Apache Spark can help reduce data interaction complexity, increase processing speed and enhance mission-critical applications with deep intelligence.
Highly versatile in many environments, Apache Spark is known for its ease of use in creating algorithms that harness insight from complex data. Spark was elevated to a top-level Apache Project in 2014 and continues to expand today.

Amazon web services

- launched 2006
- largest, most popular cloud computing platform
  - others: Rackspace, Azure, Google App Engine

Elastic Compute Cloud (EC2)

- rent “Elastic compute units” by the hour
  - one ECU = one 1 GHz processor machine
- can choose Linux, FreeBSD, Solaris, Windows
- virtual private servers running on Xen
- pricing: US$0.02-2.50 per hour
  - varies by machine capacity
  - spot pricing-varies by demand

aws.amazon.com/ec2/

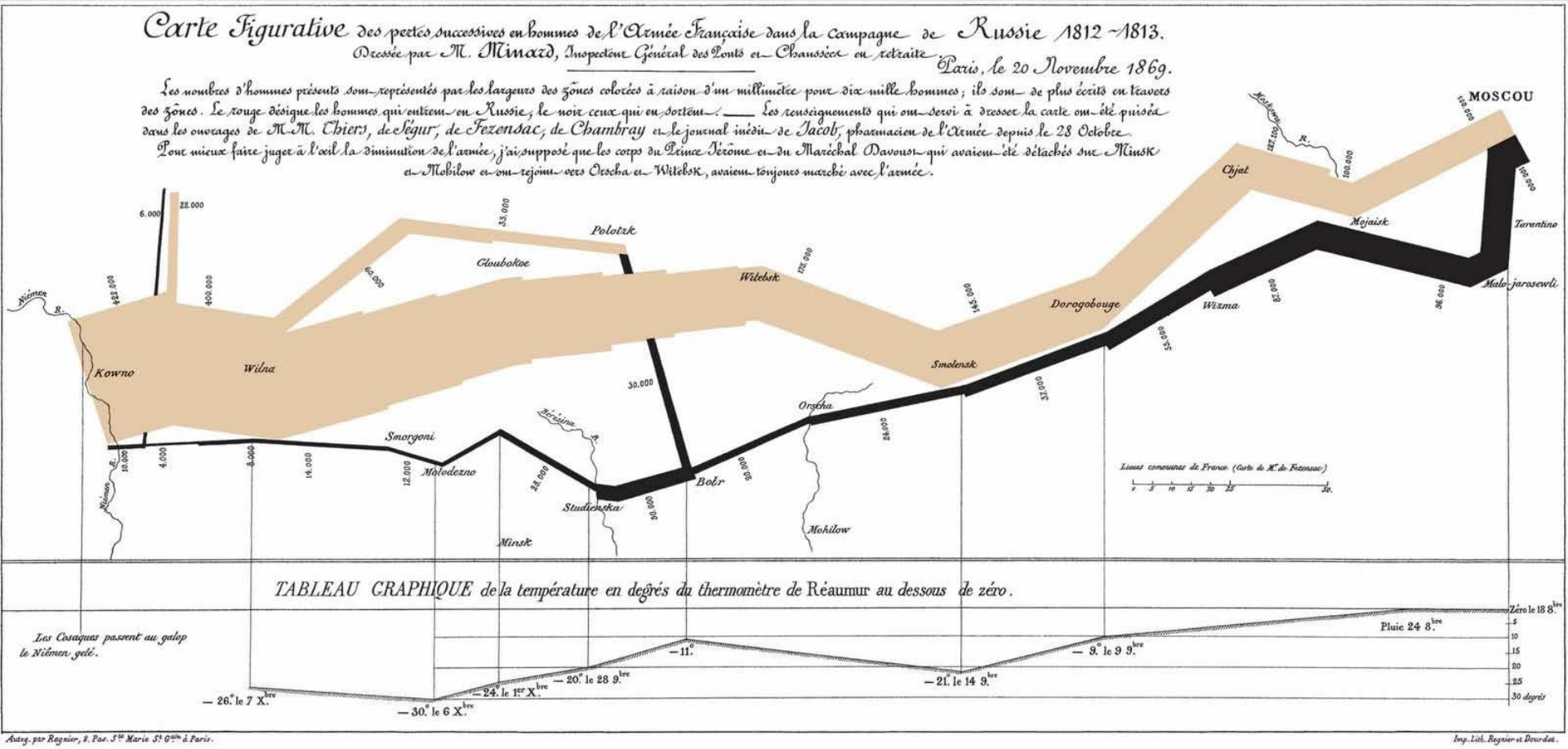
Other AWS elements

- elastic MapReduce
  - run Hadoop on EC2 machines with S3 storage
  - free data transfer
- relational Database Service
  - SQL database
- many features for running a data-driven website
  - content delivery networks, caches, etc.

Comparison

	Google	Hadoop	Amazon WS
storage	GoogleFS	HDFS	S3
caching	memcacheg		ElastiCache
locking	chubby	Zookeeper	
key-value	LevelDB		
column-oriented	BigTable	Cassandra	DynamoDB
document-oriented		CouchDB	SimpleDB

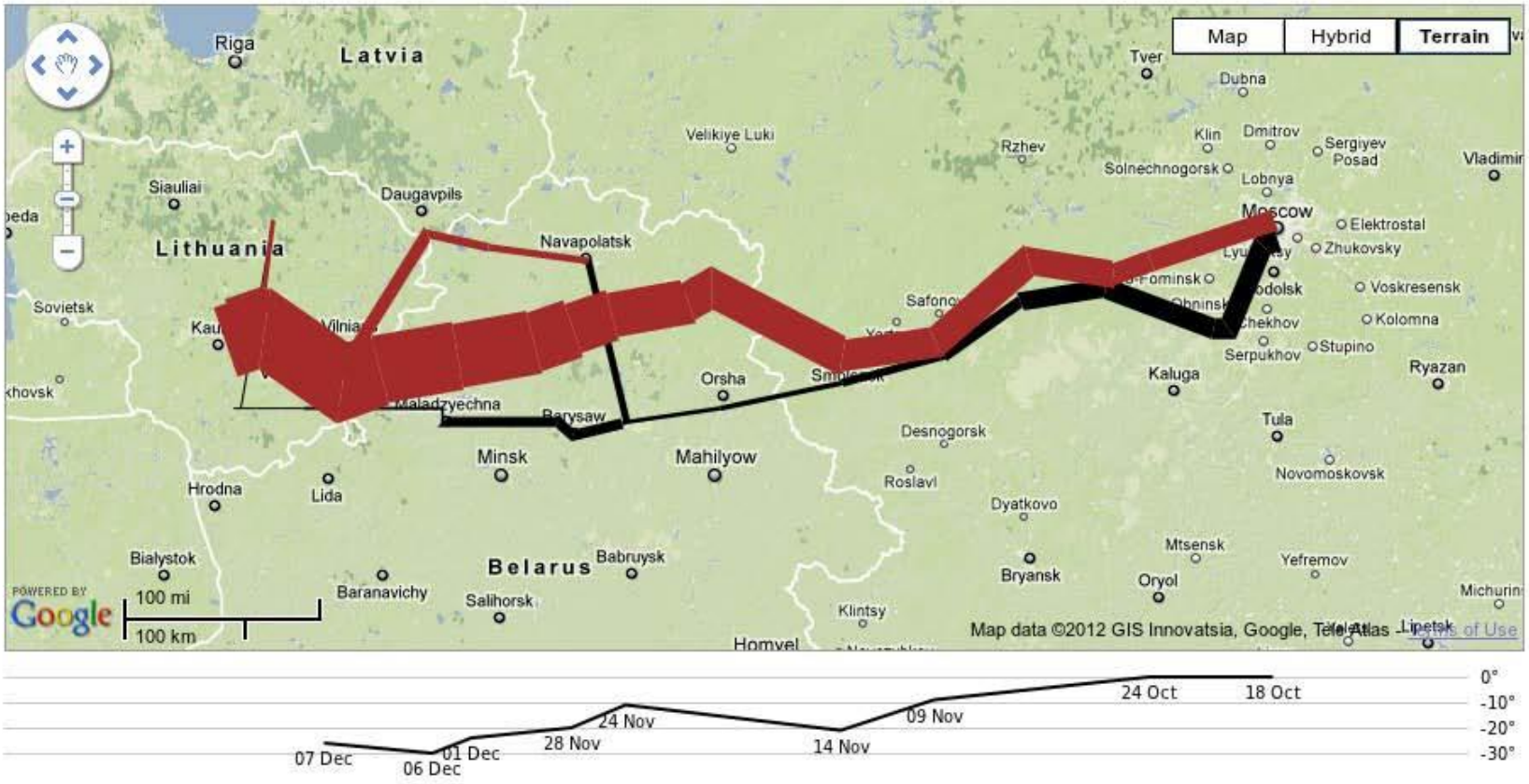
Visualisation



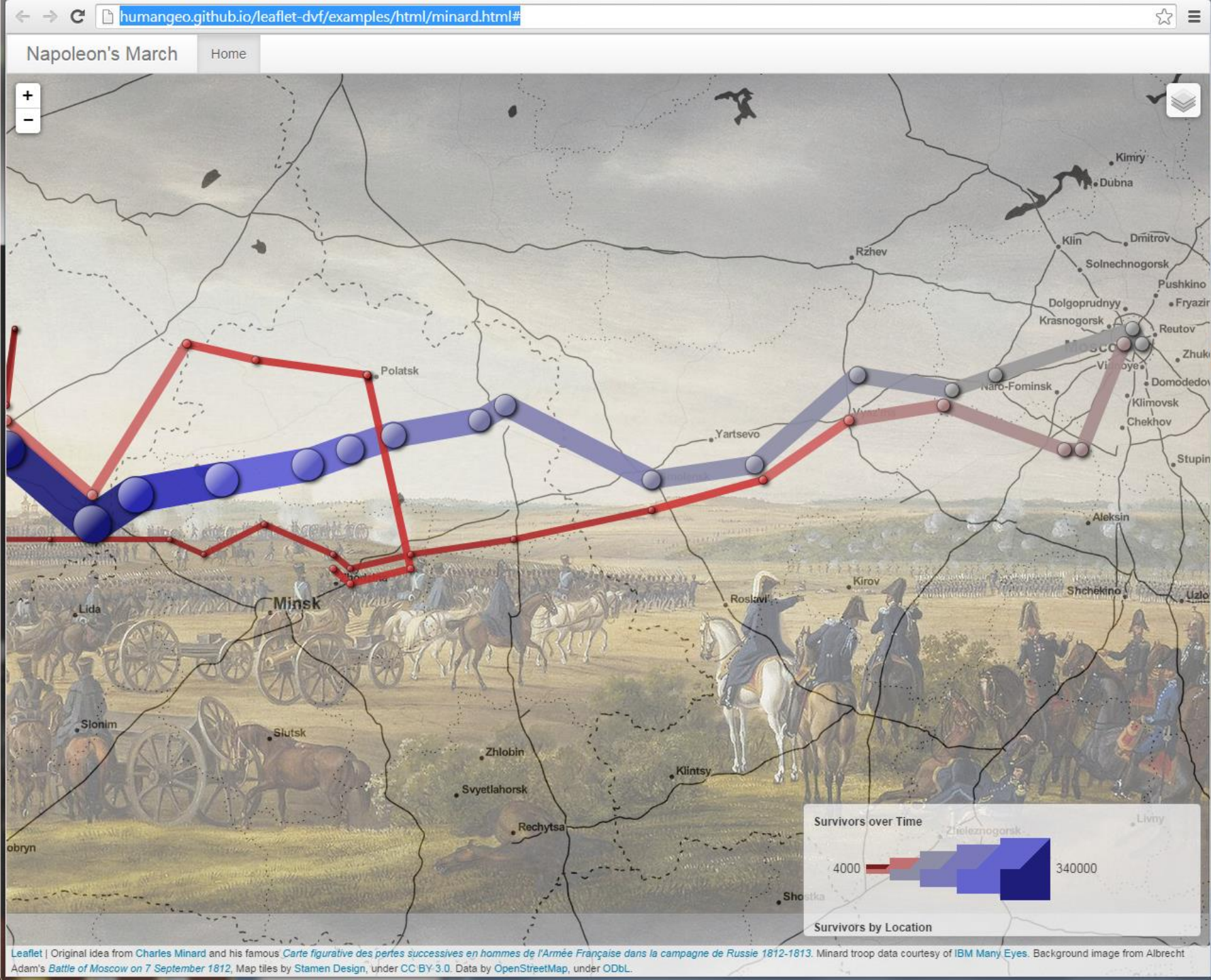
Charles Minard 1869

Visualisation

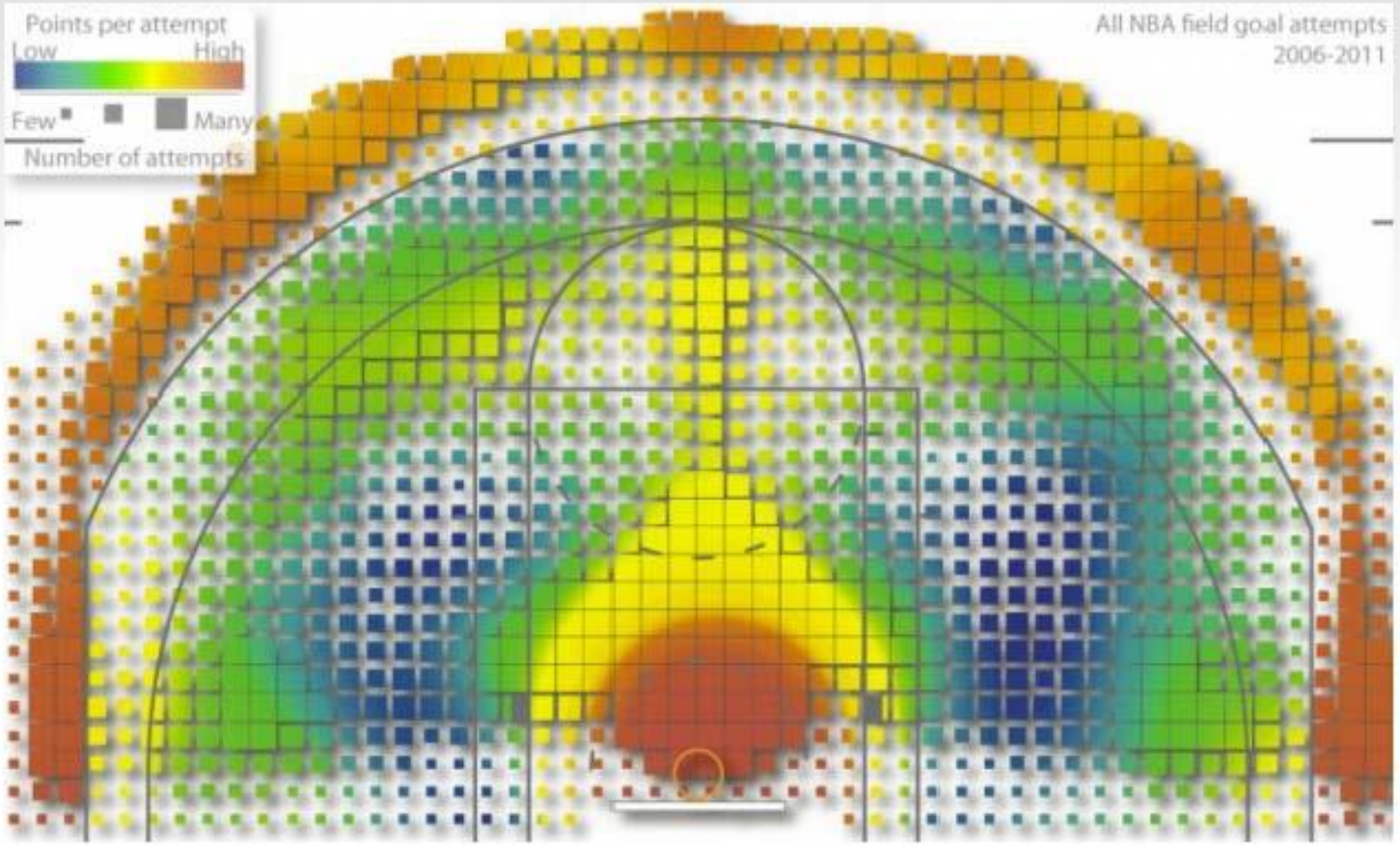
Minard's Napoleon



re-creation with Protovis, Google Maps

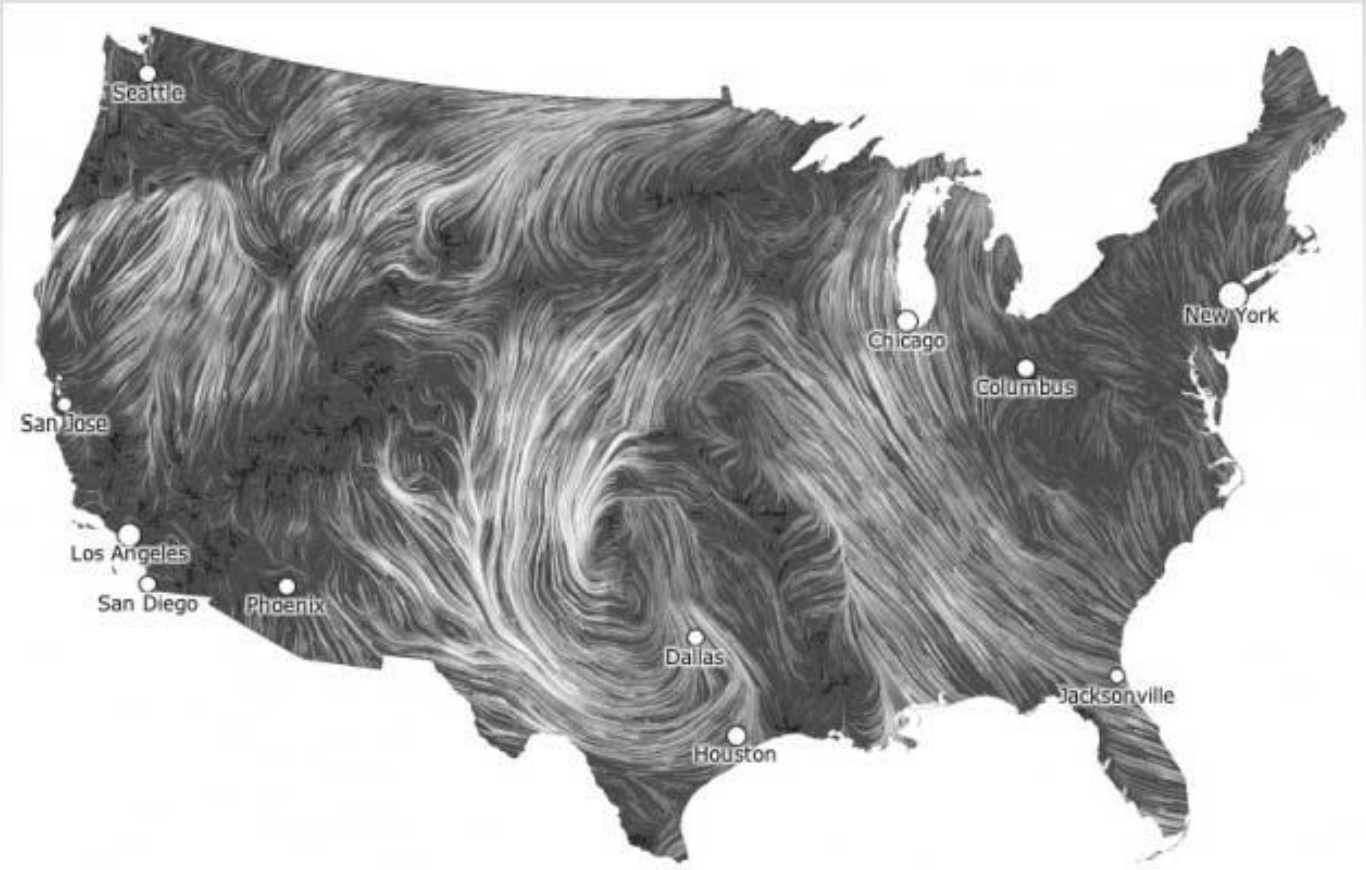

humangeo.github.io/leaflet-dvf/examples/html/minard.html#
Napoleon's March
Home
Kimry
Dubna
Klin
Dmitrov
Rzhev
Solnechnogorsk
Pushkino
Dolgoprudnyy
Krasnogorsk
Reutov
Polatsk
Naro-Fominsk
Domodedov
Klimovsk
Chekhov
Yartsevo
Stupin
Aleksin
Kirov
Roslavl'
Shchekino
Lida
Minsk
Slonim
Slutsk
Zhlobin
Klintsy
Svyetlahorsk
Rechytsa
Survivors over Time
4000
340000
Survivors by Location
Leaflet | Original idea from Charles Minard and his famous Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813. Minard troop data courtesy of IBM Many Eyes. Background image from Albrecht Adam's Battle of Moscow on 7 September 1812, Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

Visualisation



Kirk Goldberry

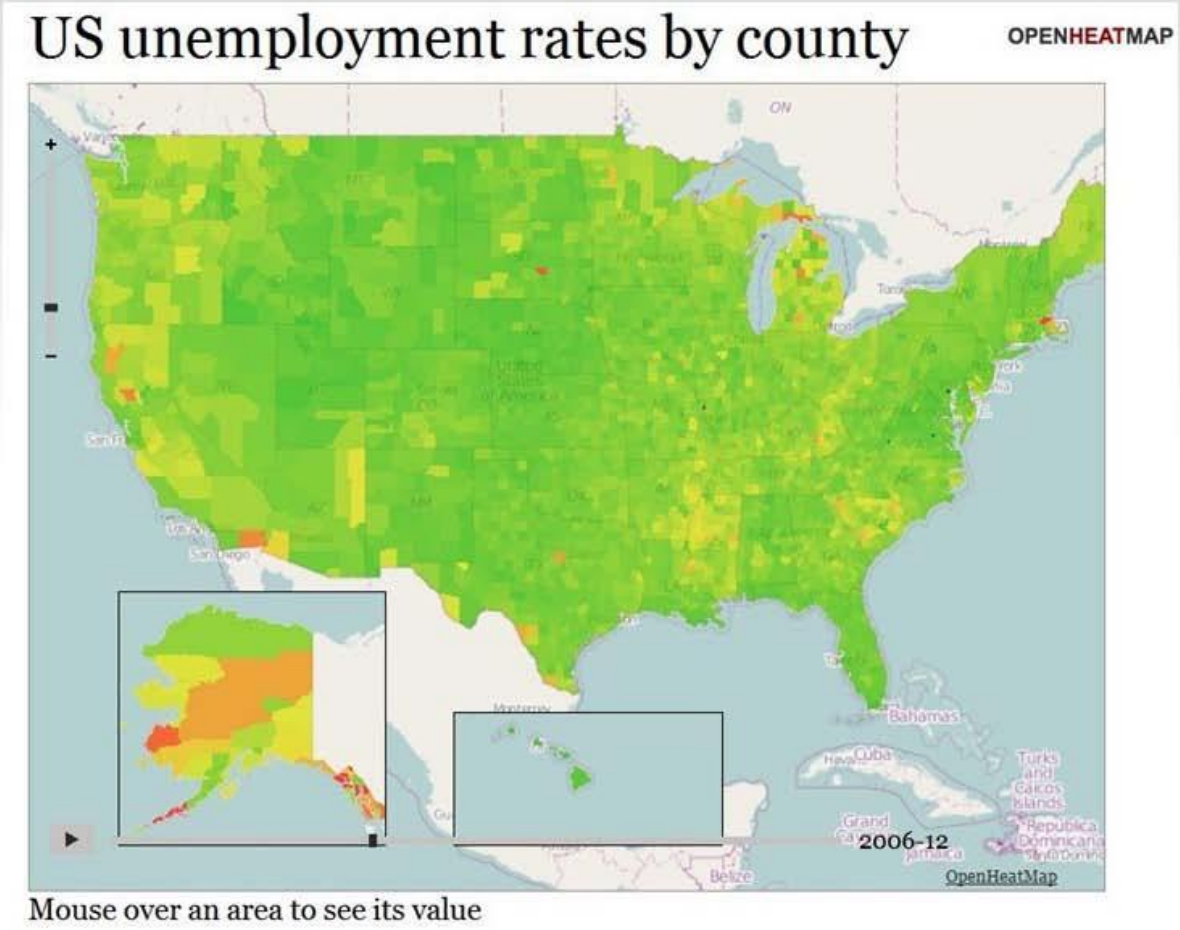
Visualisation



http://hint.fm/wind/

Martin Wattenburg
Fernanda Viegas

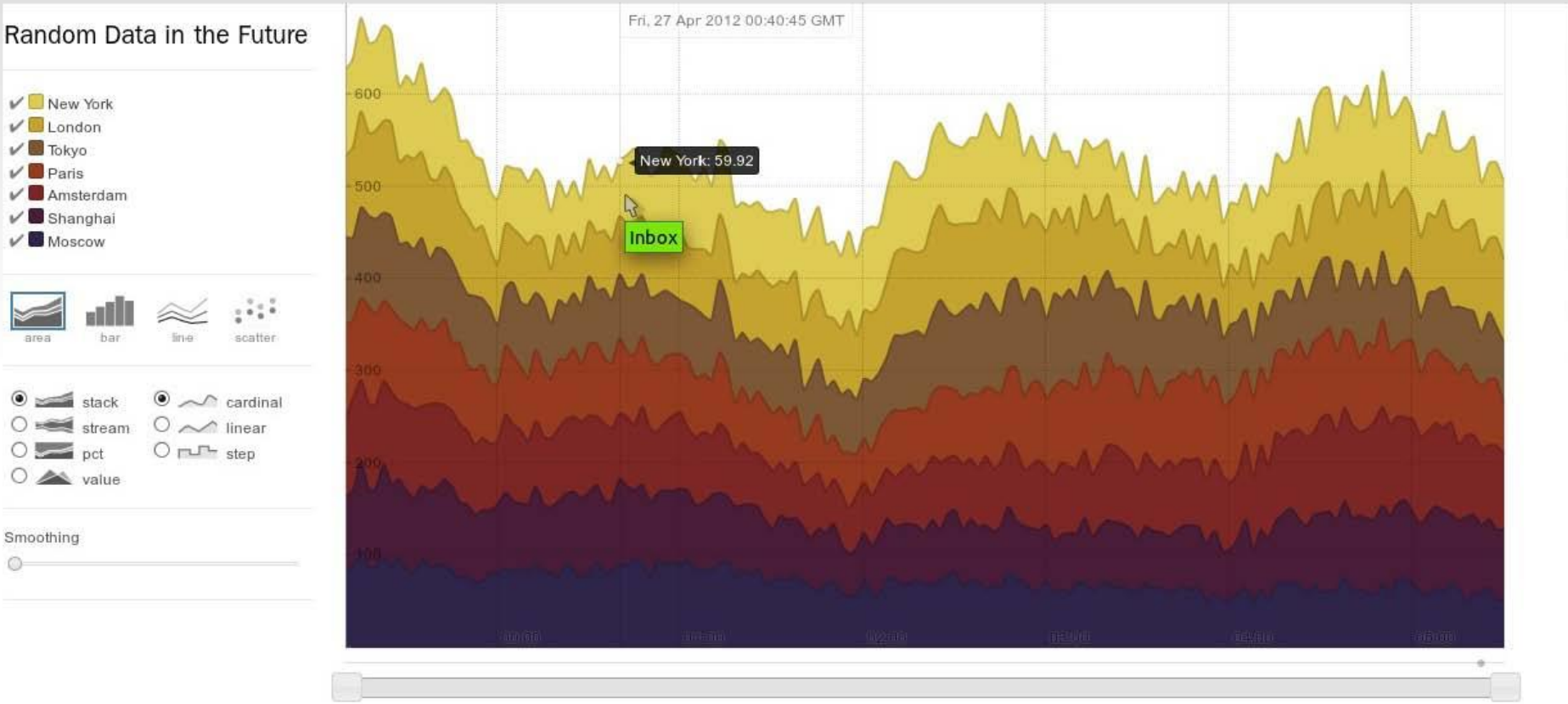
Geographic visualisation tools



OpenHeatMap

See also: FusionTables, GoogleMaps API

Interactive chart tools



Rickshaw

See also: Tableau, Highcharts JS, ExtJS, Raphaël, flot, dojox.charting

Data Engagement

- Opening up data is NOT new;

- "Introducing the ***human element*** to open data initiatives involves recognizing the key role that existing data stewards can play in making their datasets ***easier to understand***, ***interpret and work with***, and supporting the emergence of new or expanded roles focusing on ***open data engagement***.

Five Stars of Data Engagement

- Be demand driven
- Put data in context
- Support conversation around data
- Build capacity, skills, and networks
- Collaborate on data as a common resource

- *http://www.opendataimpacts.net/engagement/*

★	make your stuff available on the web (whatever format)
★★	make it available as structured data (e.g. excel instead of image scan of a table)
★★★	non-proprietary format (e.g. csv instead of excel)
★★★★	use URLs to identify things, so that people can point at your stuff
★★★★★	link your data to other people's data to provide context

Demand and Engagement

Central and local governments around the world are increasingly 'opening' a range of data, for free, including as part of continuing global efforts to strengthen 'open government.' While this has resulted in excitement from development practitioners, government sponsors, and technologists, much of the public has been left behind. As a result, the level of informed public debate across regions on data-driven issues – from budgets to service delivery to the practical effectiveness of donor aid – in 'opened' sectors is low.

So, now that this data has been 'opened', how can it capture the attention and imaginations of the full spectrum of users? How can we focus on the other side - the demand side - of the open data phenomenon? How can we grow communities of data users, and encourage data 'ownership' by the media, civic hackers, community groups, NGOs, labor unions, professional associations, universities, and more?

To help developing countries address the demand-side of open data, the World Bank offers a 'menu' of services to promote and support 'Open Data Literacy', the goal of which is to catalyze, engage, and inspire strategic multi-stakeholder groups to see the value and potential of open data, and what it means for local, national, and regional development in a practical, hands-on way.

STEP 1: Early Engagement

Catalyzing Supply/Demand Side Actors to Support Opened Data; Surfacing Data

- **'Open Data Sensitization' Roundtable** – an overview of open data processes crystallized from World Bank and partner experiences. Examples include sessions at the first and second International Open Government Data Conferences, Open Government Conference in Moscow, and at the Regional Conference on Open Data for Latin America and the Caribbean.
- **The 'Business Case for Open Data' Roundtable** – a high-level roundtable to encourage use, analysis, and ownership of Open Data, and to catalyze leadership on both supply/demand sides. This information resonate with particular audiences and twinned with support opportunities from the Bank and close partners, including the 'Data Literacy Bootcamp' detailed below. Examples include Media Leaders Roundtables on 'Open Data and the Future of News' in Ghana, Kenya, Nepal, Tanzania, and South Africa.
- **Data Liberation Scrape-a-thon** – an 'early engagement' 2-day program which convenes international coders (and sources local civic coders) to scrape as much existing (unstructured) government data off current websites, structuring it (CSV format); and using it populate a nascent or temporary open data platform. The goal is to start to gain traction on early open data process by capitalizing on what's already available, and leveraging it to build momentum toward further opening by individual ministries. Examples include support

Open Government Data Toolkit

Open Data Essentials
Technology Options
Demand and Engagement
STEP 1: Early Engagement
STEP 2: Capacity Development
STEP 3: Use + Ownership + Reuse
STEP 4: Further Development
Additional Resources
Supply and Quality of Data
Readiness Assessment Tool