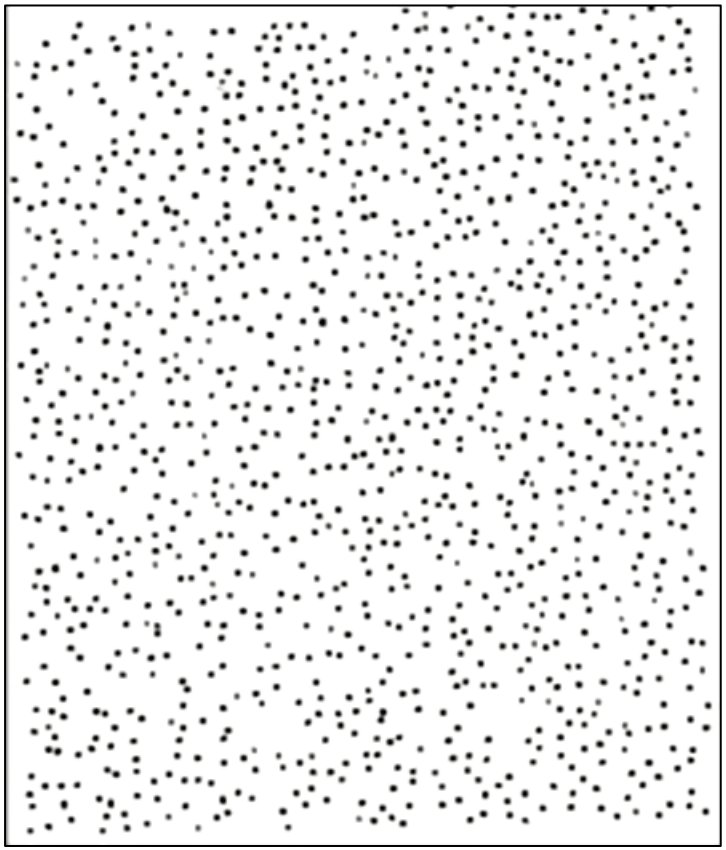
Big Data

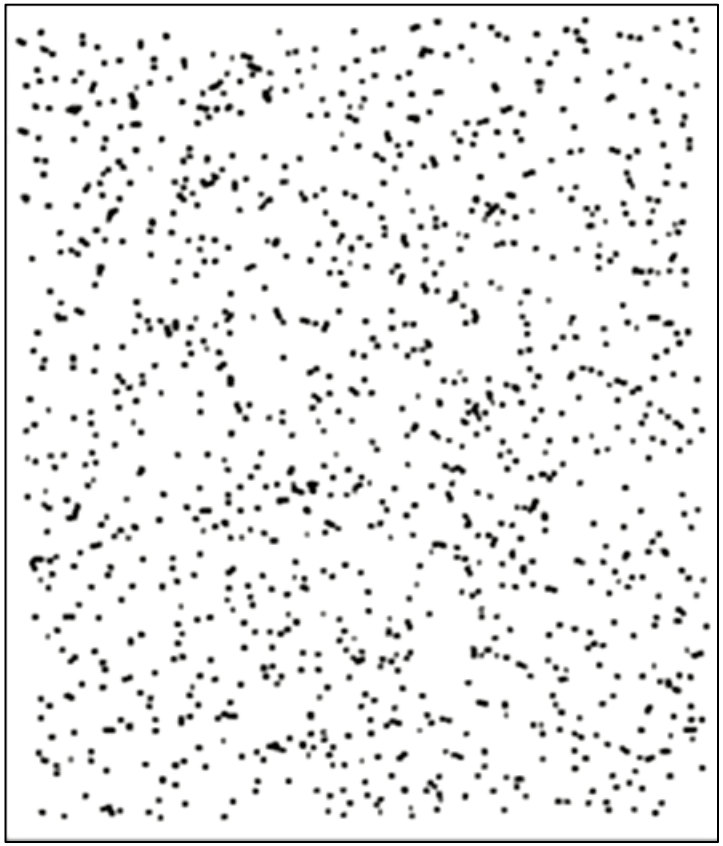
Overview, Analysis, and Visualization

Ng Yen Kaow

The basics of analytics (1/3)

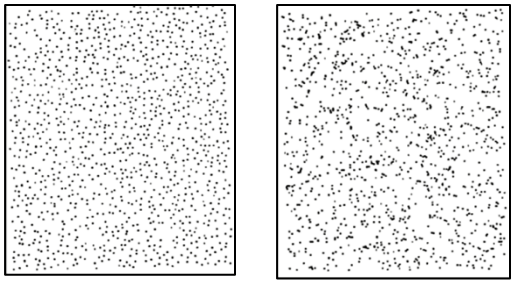
- Discover patterns, or regularities in data
  - Sometimes they not be meaningful to human
  - Example





The basics of analytics (2/3)

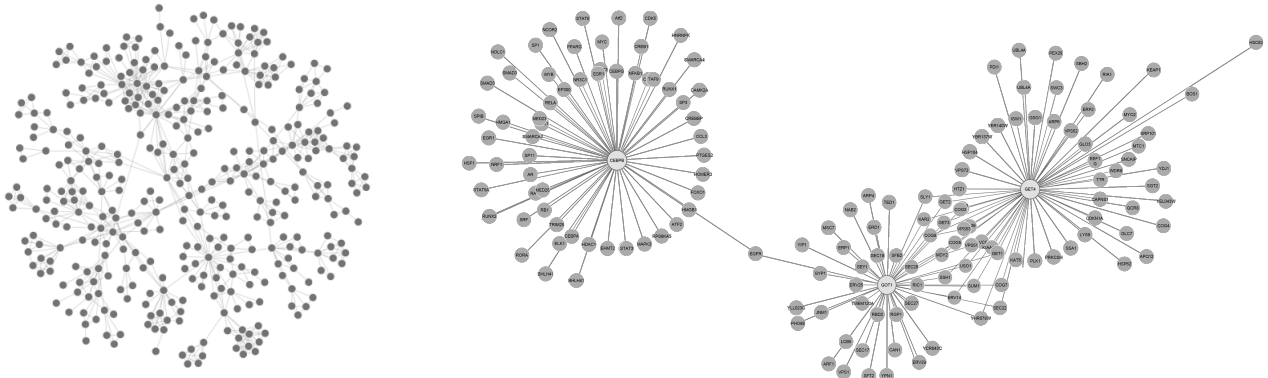
- Discover patterns, or regularities in data
  - Sometimes they not be meaningful to human
  - Example



  - Example

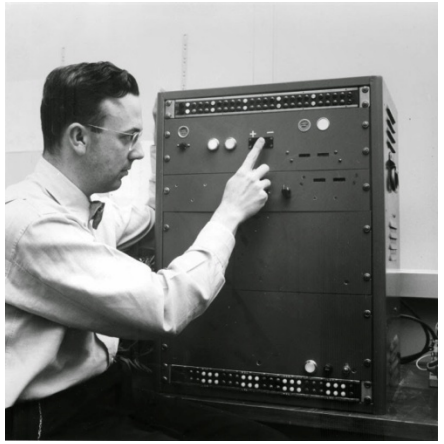
11000011100001111000011111000…

  - Example



The basics of analytics (3/3)

- Consider the game of "**matching pennies**"
  - Two players each conceal a coin –head or tail– and reveal them simultaneously
  - Player 1 wins if the coins matches, otherwise player 2 wins
- A sequence of plays is usually not random – we have a tendency to play in specific manners
- In the 1950s, Hagelbarger built a machine that plays the game
  - **One difficulty was in getting people to play so that the machine can collect data…**



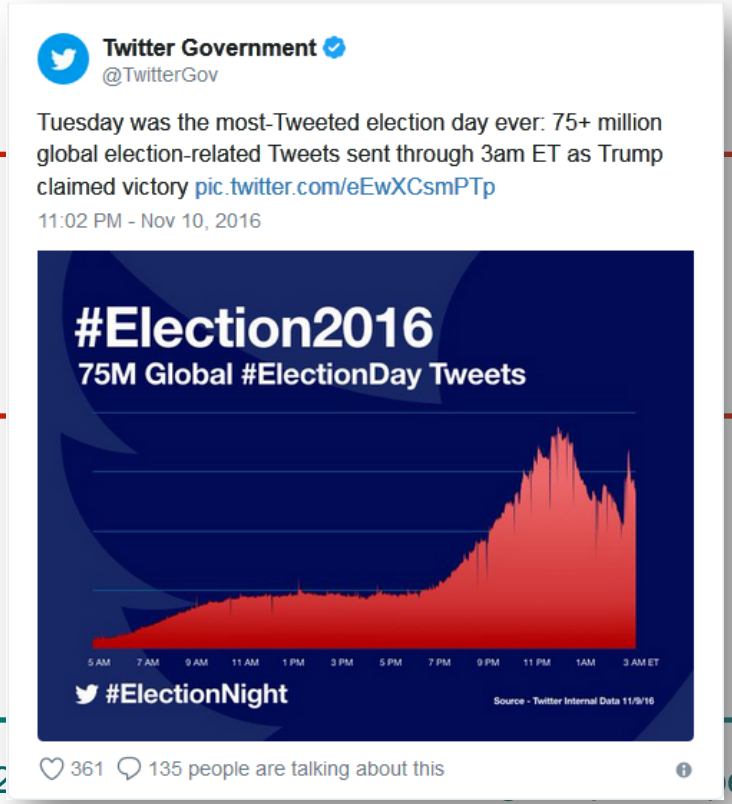
Limits to analytics in the past

- We know that
  - More data ⇒ more accurate results
  - More data ⇒ more kinds of analyses

- However, in the past, we were restricted by
  - Lack of data collection mechanism
  - Insufficient data processing power

Present state of data collection

- Data Sources
  - **Internet Content**
    - Social Media
  - **Logs**
    - Sales transactions
    - Medical records
  - **Closed-circuit cameras**
  - **Sensors**
  - **Scientific data**



Present processing power (1/6)

- Commercial-class rack server
  - Example: Dell M1000e w/ M830 blades
    - 1 M1000e rack
    - 8 blades in rack
    - 64TB (8x4x2TB) HD
    - 6TB (8x48x16GB) RAM

LET'S BUILD THE ULTIMATE SERVER		
1 M1000e rack		$ 2,000
8 blades in rack	8x$5000=	$ 40,000
480TB (8x4x15TB) HD	8x4x$10000=	$ 320,000
24TB (8x48x64GB) RAM	8x48x$700=	$ 268,800
		$ 630,800

Present processing power (2/6)

- Powerful, affordable GPUs
  - $800~$1000

  - >$7,000

Present processing power (3/6)

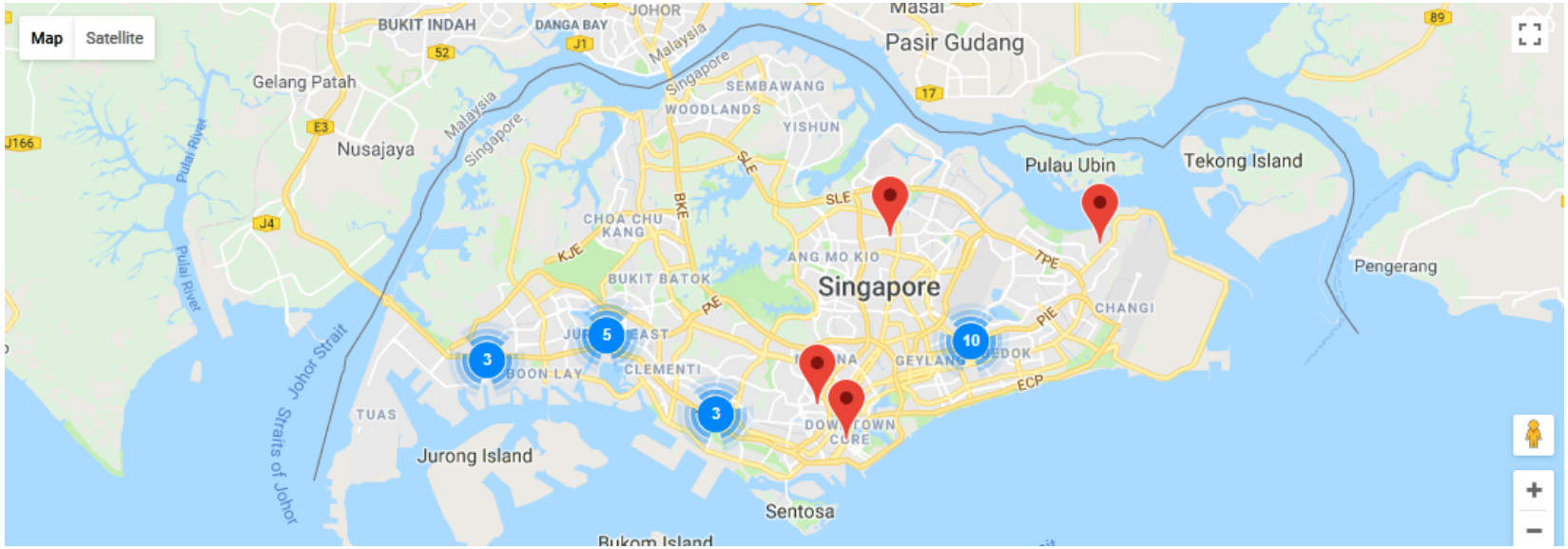
- Supercomputers
  - Example: Sunway TaihuLight
    - 10,649,600 cores
    - 1,310,720 GB RAM

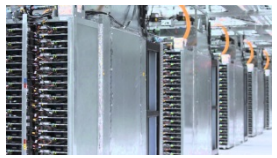
Year	Supercomputer	Peak speed (Rmax)	Location
2016	Sunway TaihuLight	93.01 PFLOPS	Wuxi, China
2013	NUDT Tianhe-2	33.86 PFLOPS	Guangzhou, China
2016 (2012)	Piz Daint	19.59 PFLOPS	Switzerland
2012	Cray Titan	17.59 PFLOPS	Oak Ridge, USA
2012	IBM Sequoia	17.17 PFLOPS	Livermore, USA
2016	Cori	14.01 PFLOPS	NERSC, USA
2016	Oakforest-PACS	13.55 PFLOPS	Tokyo, Japan

Present processing power (4/6)

- Data centers

https://www.datacenters.com/locations/Singapore





Google's data center in Douglas County, Georgia

Present processing power (5/6)

- Sharing of processing power

  ⇒ Cloud service

  - Example: Amazon Elastic Compute Cloud (EC2), Microsoft Azure, Google Cloud Platform
  - Merits and demerits

    Pros:

    - Many hardware configurations to choose from
    - No need for setup
    - Pay only what you use

    Cons:

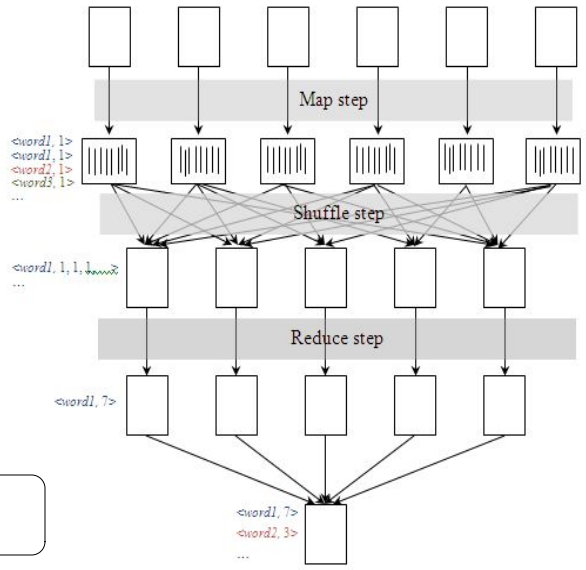
    - Security and privacy concerns
    - Waste time in data upload

Present processing power (6/6)

- Distributed technology
  - Parallelized document search operations
    - MapReduce

      e.g. Google search
  - Content distribution over multiple servers
    - *e.g.* Youtube
      http://www.youtube.com/watch?v=**nbp3Ra3Yp74**
  - *etc.*, *etc.*

Determines the server



The MapReduce framework

- The market is flooded with supporting products
  - Hadoop / Spark
  - Cassandra, MongoDB, Riak, CouchDB, Redis, Hbase, Couchbase, Neo4j, Hypertable, ElasticSearch, Accumulo, VoltDB, Scalaris, Bigtable, Dynamo, ***etc., etc., etc. …***

Data sources

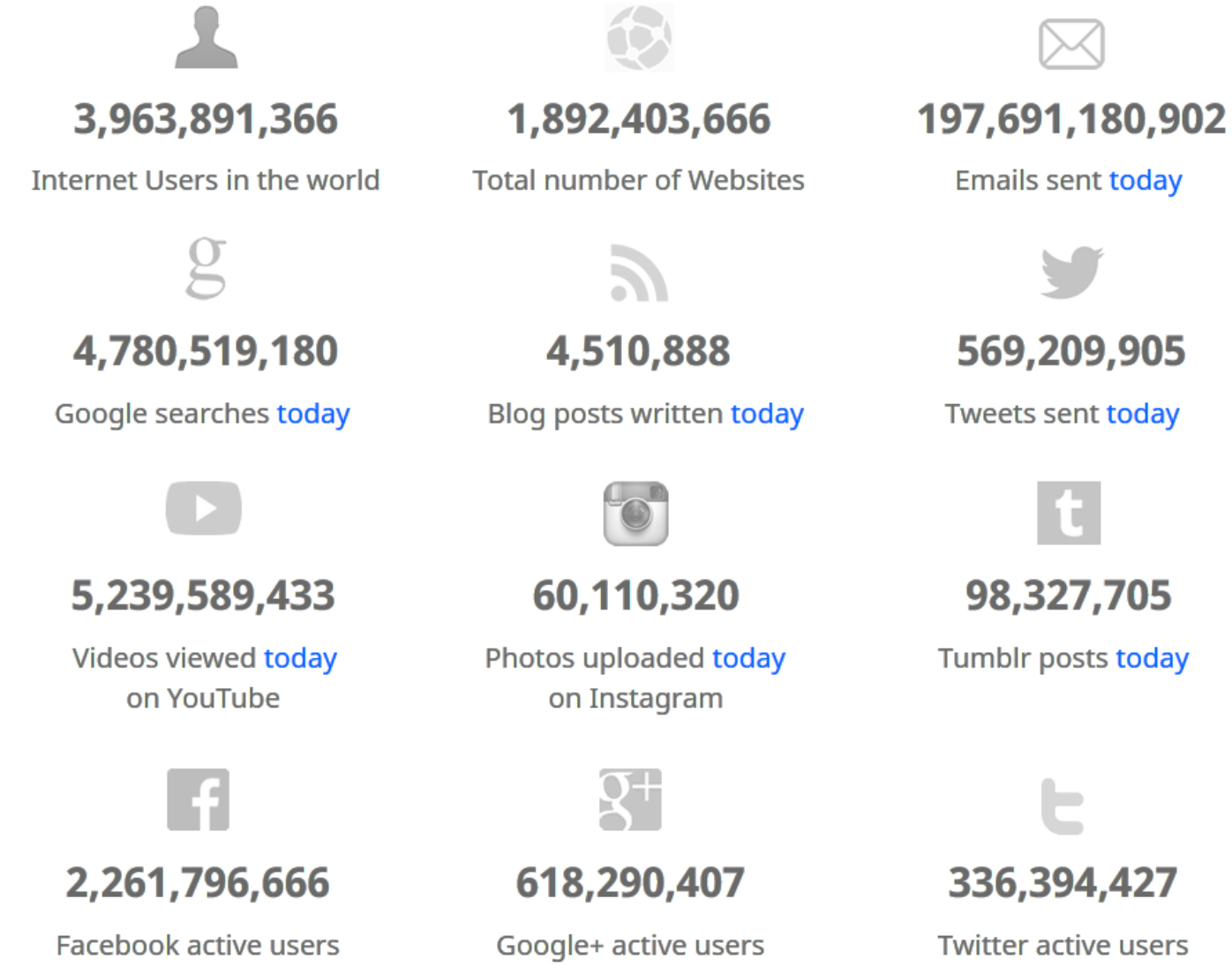
- Internet
- Social networks
- Public knowledge bases
- Log files, sales transactionss
- Data from devices: phones, camera, sensors
- Scientific data
  - NASA
  - Biological data

Internet sources

- World Wide Web
  - **At least 4.5 billion pages**
    - According to http://www.worldwidewebsize.com (8 July, 2018),
    - Does not include pages that are only accessible only upon authentication, or dynamic pages generated from database only upon access
  - Primarily HTML and XML
  - Increasingly more image and video data
  - **Rich natural language data**
    - CMU's "Read the Web" project (NELL)

Social networks (1/3)

- http://www.internetlivestats.com/



Social networks (2/3)

- Mining social networks ("gone viral") has replaced reporting

This Sydney train trip video has gone viral across the world, but why?

8 July. https://www.sbs.com.au/news/this-sydney-train-trip-video-has-gone-viral-across-the-world-but-why

This picture of a dad 'breastfeeding' his newborn has gone viral

3 July. http://www.wbaltv.com/article/this-picture-of-a-dad-breastfeeding-his-newborn-has-gone-viral/22035181

This Mom's Hack For Soothing A Sunburn Has Gone Viral

5 July. https://www.buzzfeed.com/laurenstrapagiel/sunburn-hack-shaving-foam?utm_term=.nhj706pP9#.kid28oeY0

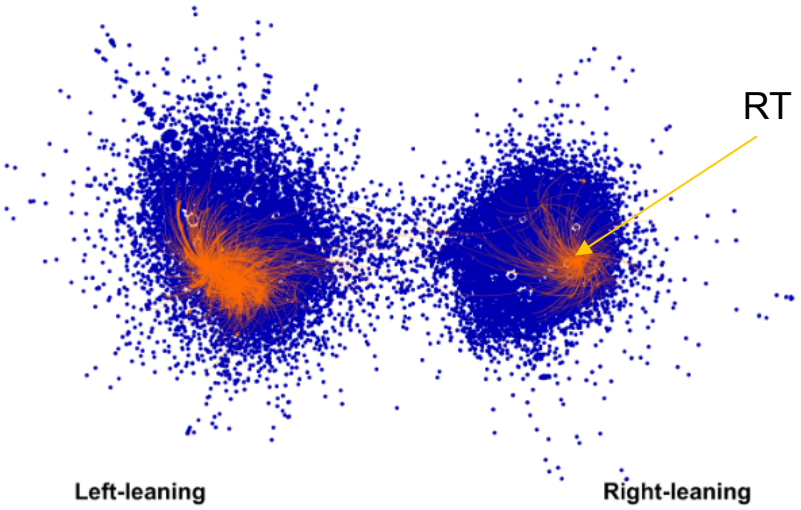
Cleve farmer Damien Elson has gone viral with his sheep art

https://www.adelaidenow.com.au/news/south-australia/cleve-farmer-damien-elson-has-gone-viral-with-his-sheep-art/news-story/1321faf6d3a60d4f50b6f115f57ceb9f

Social networks (3/3)

- Social networks have been weaponized
  - Activism
    - Small group of activists generate large amount of data
  - Sabotage
    - Activities on social media during US election 2016
    - Accounts traced to Russia-based Internet Research Agency (IRA)

RT network
Constructed from
tweets on BLM
of IRA origin



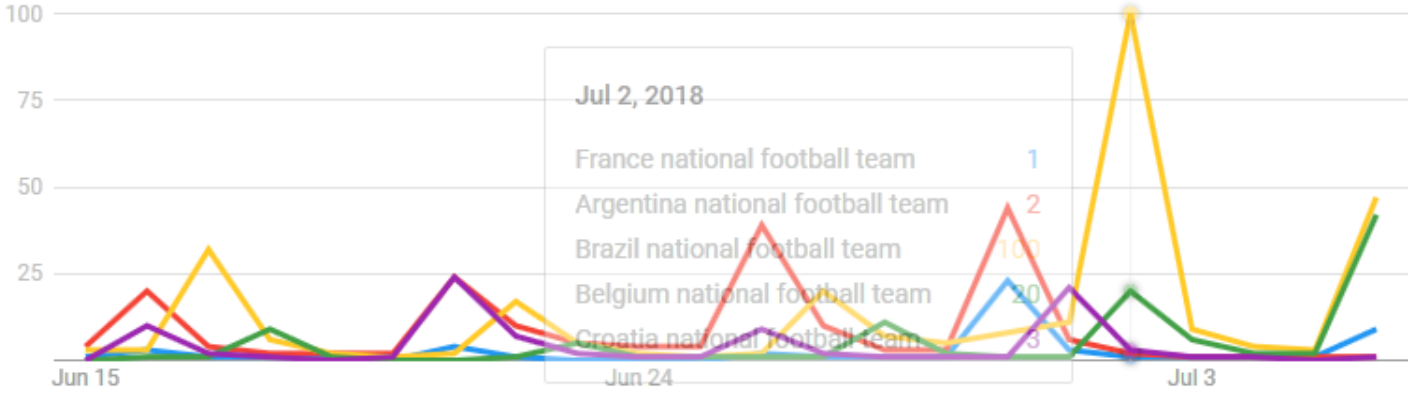
L. G. Stewart, A. Arif, K. Starbird
"Examining Trolls and Polarization with a Retweet Network"
MIS2: Misinformation and Misbehavior Mining on the Web 2018

Public data

- Massive amount of public domain data has been curated
- GSOD data from WMO (114,420,316 rows) and NOAA (lots, 1929~)
  - Temperature, precipitation, dew point, pressure, visibility, etc
- GitHub project information (2,541,639 rows)
  - Project URL, language, location, etc
- Shakespeare (164,656 rows – got so many words in English meh?)
  - word, word count, from which work, date of work
- 3-grams (68,051,509 rows)
- Wikipedia edits (313,797,035 rows)
  - Page title, revision_id, contributor_id
- Persons names (collected processing social security number) (5,552,452 rows)
  - Name, gender, state, number (note: no surname)

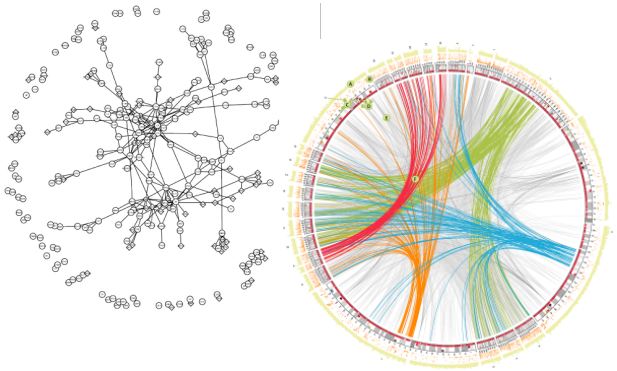
Data ⇒ analyses opportunities

- Simple analyses give interesting results
  - Counting-based analyses
    - How do we know something is popular? Count internet or social media appearances!



How can we find out our probable friends in **Facebook**? Count the number of mutual friends

- More sophisticated analysis
  - **Network mining**



Points to ponder

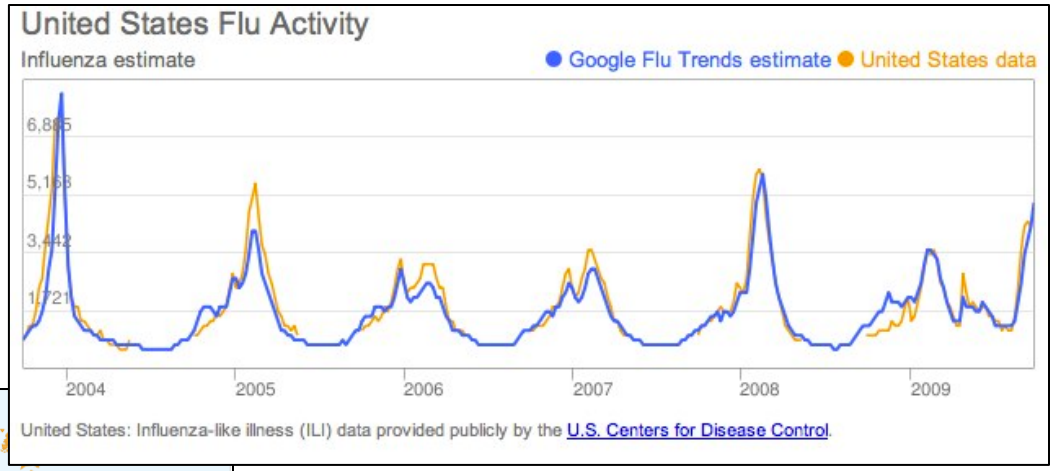
- Why couldn't we do "Big Data" in the past?
- Why is it possible today?

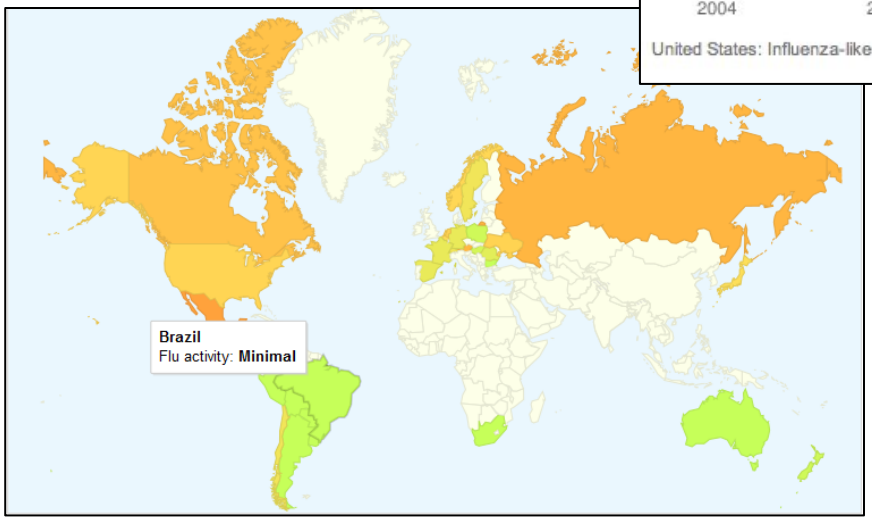
- Next point: Why do I want this "Big Data"?

Correlation analysis

- Correlation analysis reveals possible relationship between events

Allows **prediction** of events





Allows more **strategic planning**

(Internet) activity analysis

- Various activities can be tracked for our benefits
  - Sites such as alexa tracks internet traffic
  - Google keeps record of their searches
  - Facebook keeps track of the number of shares and likes
  - Twitter tracks the number of retweets
  - Quora analyzes upvotes

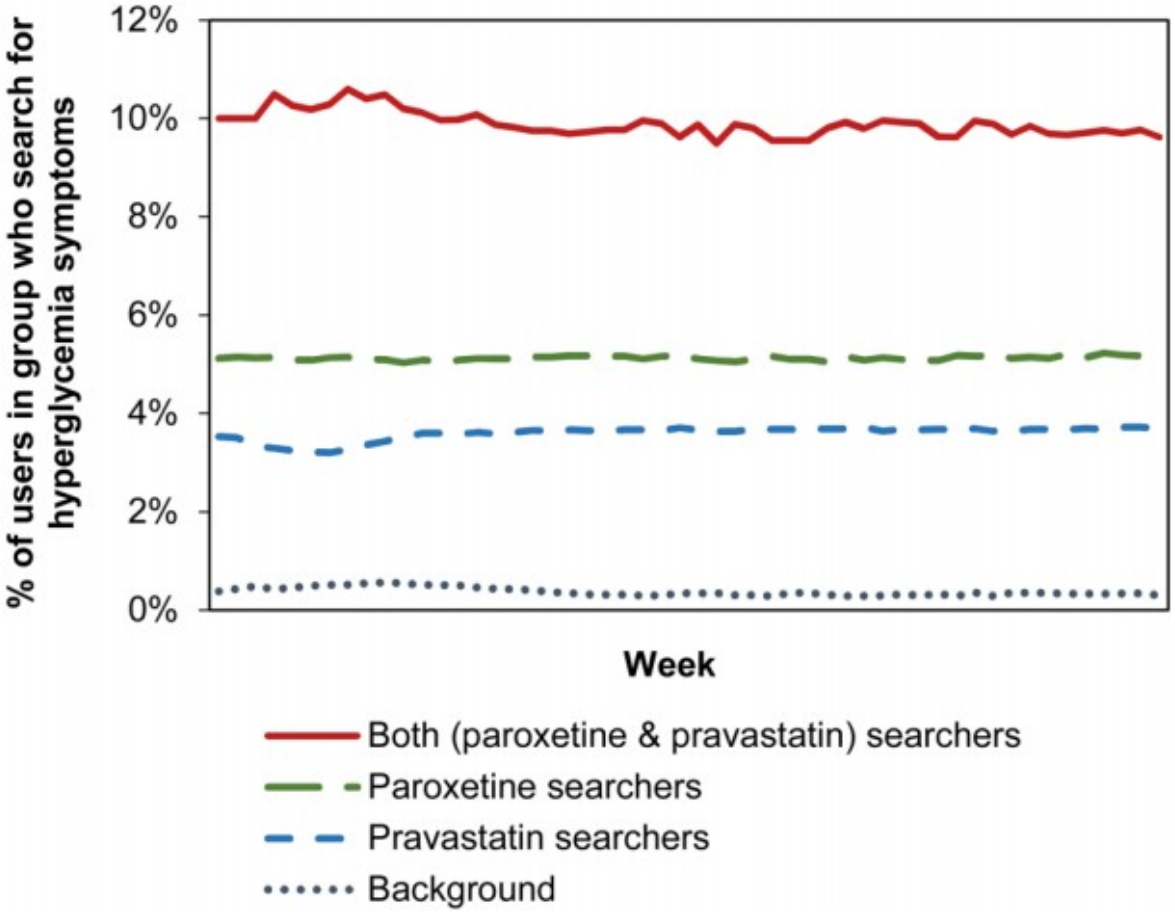
(Internet) activity analysis *(cont.)*

- Example: Discover side-effects of drugs from Internet searches
  - It is suspected that the simultaneous use of the drugs paroxetine and pravastatin causes hyperglycemia – a condition in which an excessive amount of glucose circulates in the blood plasma
  - This association is examined using web log data

R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, E. Horvitz
"Web-scale pharmacovigilance: listening to signals from the crowd"
J Am Med Inform Assoc, 20(3):404-8, 2013

(Internet) activity analysis *(cont.)*

- Example: Discover side-effects of drugs from Internet searches



(Internet) activity analysis *(cont.)*

- Example: The Tiobe index tracks internet activity to gauge the popularity of programming languages

Jul 2018	Jul 2017	Change	Programming Language	Ratings	Change
1	1		Java	16.139%	+2.37%
2	2		C	14.662%	+7.34%
3	3		C++	7.615%	+2.04%
4	4		Python	6.361%	+2.82%
5	7	^	Visual Basic .NET	4.247%	+1.20%
6	5	v	C#	3.795%	+0.28%
7	6	v	PHP	2.832%	-0.26%
8	8		JavaScript	2.831%	+0.22%
9	-	^^	SQL	2.334%	+2.33%
10	18	^^	Objective-C	1.453%	-0.44%
11	12	^	Swift	1.412%	-0.84%

Text analysis

- Sentiments analysis
  - Some words carry positive or negative meanings

# POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by `to') having
a	00002098	0	0.75	unable#1	(usually followed by `to') not
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward
a	00002730	0	0	acroscopic#1	facing or on the side toward the
a	00002843	0	0	basiscopic#1	facing or on the side toward the
a	00002956	0	0	abducting#1 abducent#1	especially of muscles;
a	00003131	0	0	adductive#1 adducting#1 adducent#1	especially
a	00003356	0	0	nascent#1	being born or beginning; "the nascent
a	00003553	0	0	emerging#2 emergent#2	coming into existence;
a	00003700	0.25	0	dissilient#1	bursting open with force, as
a	00003829	0.25	0	parturient#2	giving birth; "a parturient
a	00003939	0	0	dying#1	in or associated with the process of
a	00004171	0	0	moribund#2	being on the point of death;
a	00004296	0	0	last#5	occurring at the time of death; "his
a	00004413	0	0	abridged#1	(used of texts) shortened by
a	00004615	0	0	shortened#4 cut#3	with parts removed; "the
a	00004723	0	0	half-length#2	abridged to half its original
a	00004817	0	0	potted#3	(British informal) summarized or
a	00004980	0	0	unabridged#1	(used of texts) not shortened;
a	00005107	0.5	0	uncut#7 full-length#2	complete; "the full-
a	00005205	0.5	0	absolute#1	perfect or complete or pure;

First few lines from the SentiWordNet

Text analysis *(cont.)*

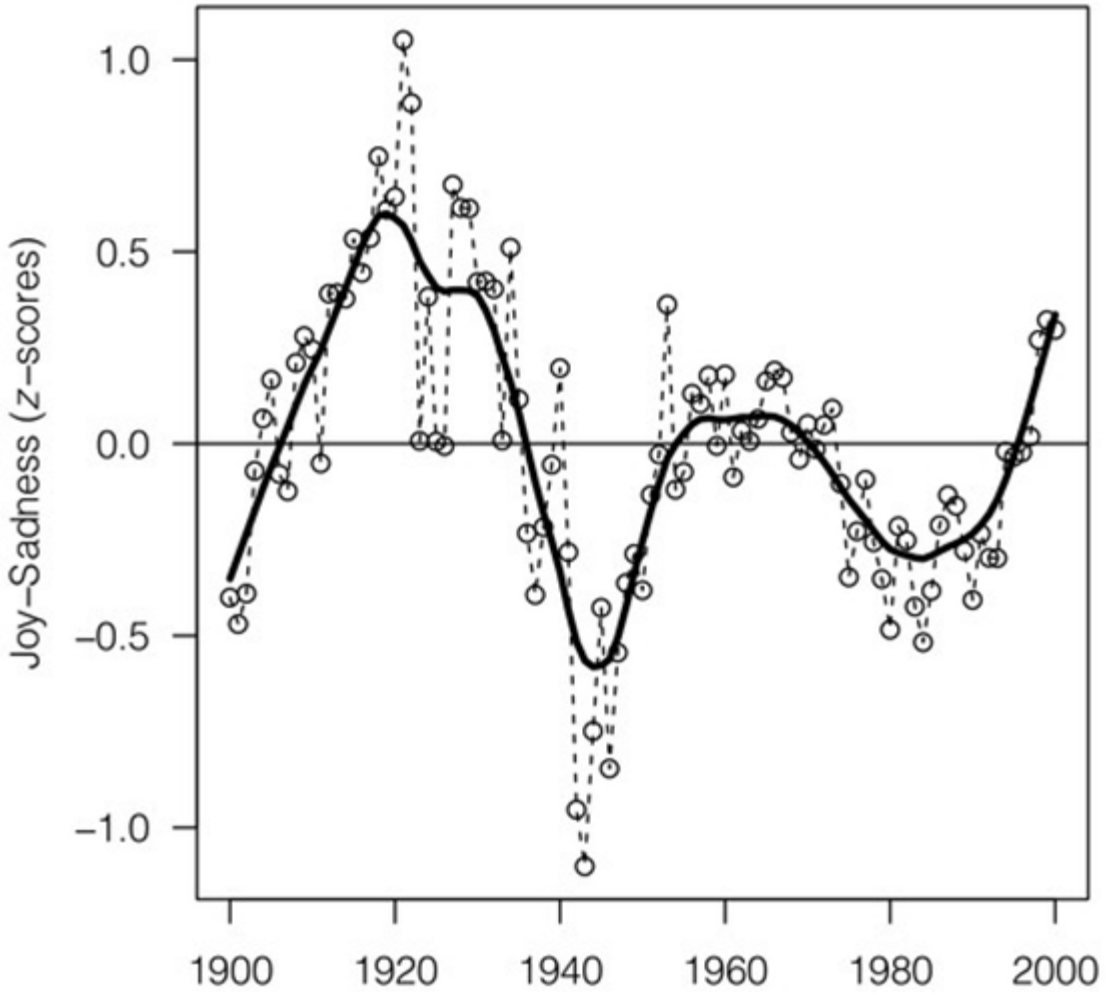
- Example: Sentiments of literature correlates accurately with world events during their year of publication

A. Acerbi, V. Lampos, P. Garnett, R. A. Bentley
"The Expression of Emotions in 20th Century Books"
PLoS ONE 8(3), 2013

  - Procedure:
    - Convert all the digitized books in the 20th century into n-grams and label each 1-gram (word) with a mood score
    - Collect all the literature for each year and count the occurrences of each mood word for that year
    - This gives us a mood count for the literature of each year
    - Plot the (mood) counts

Text analysis *(cont.)*

- Example: Sentiments of literature correlates…



Text analysis *(cont.)*

- Already a large body of work in sentiments analysis

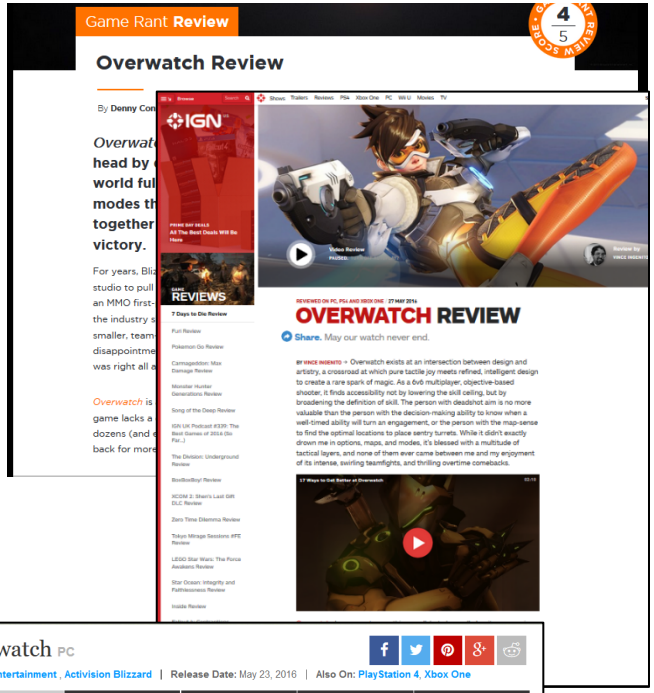
…

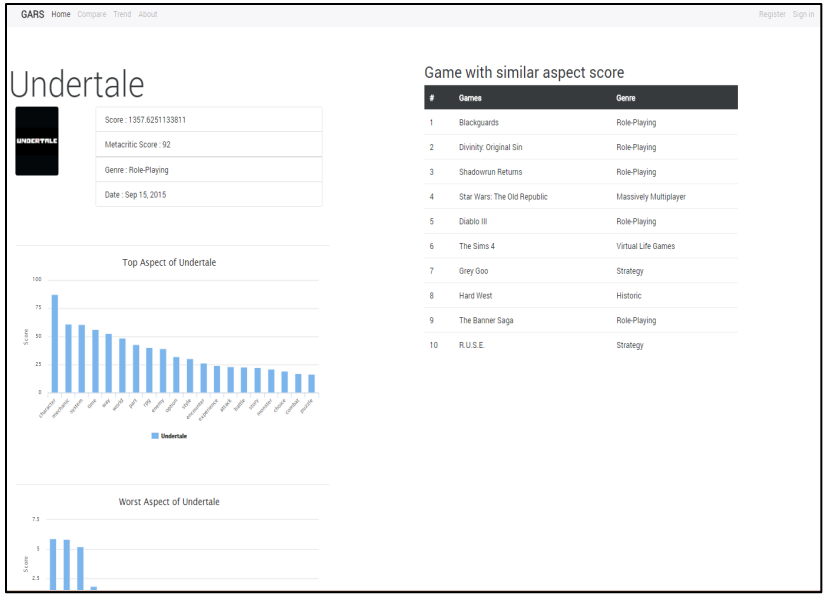
2. Michel J-P, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) ***Quantitative analysis of culture using millions of digitized books***. Science 331: 176–182. doi: 10.1126/science.1199644. Find this article online
3. Lieberman E, Michel J-P, Jackson J, Tang T, Nowak MA (2007) ***Quantifying the evolutionary dynamics of language***. Nature 449: 713– 716. doi: 10.1038/nature06137. Find this article online
4. Pagel M, Atkinson QD, Meade A (2007) ***Frequency of word-use predicts rates of lexical evolution throughout Indo-European history***. Nature 449: 717–720. doi: 10.1038/nature06176. Find this article online…
5. DeWall CN, Pond RS Jr, Campbell WK, Twenge JM (2011) ***Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions Over Time in Popular U.S. Song Lyrics***. Psychology of Aesthetics, Creativity and the Arts 5: 200–207. doi: 10.1037/a0023195

…

Text analysis *(cont.)*

- Aspect ranking
  - Automatically obtain the relevant aspects of a product from its review articles
  - Input: review articles of product
  - Output: important aspects of the product and how the product scores in these aspects

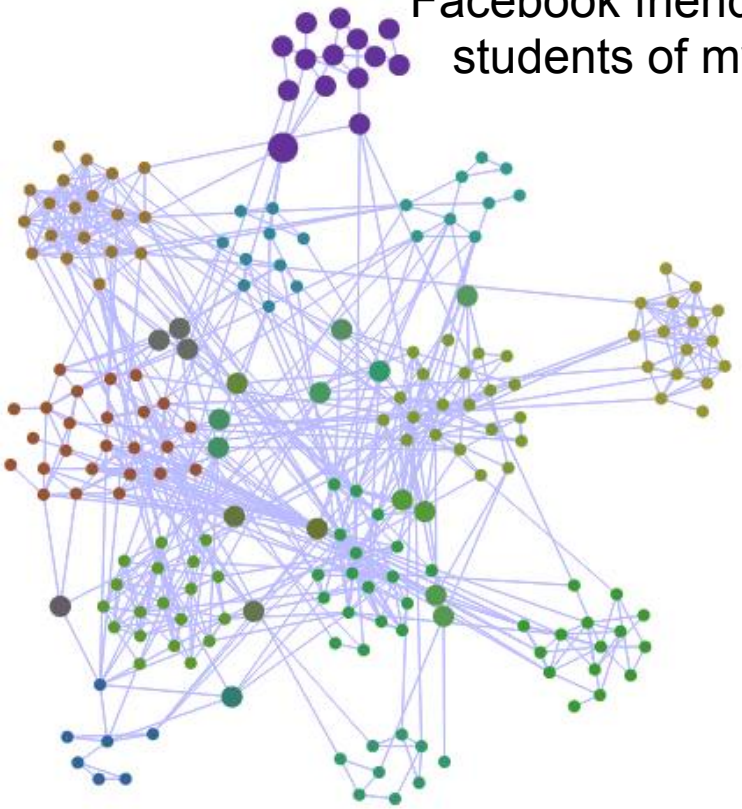




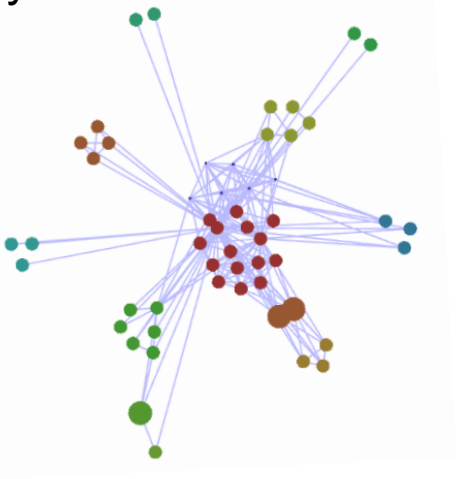
Graph / Structure mining

- Identifying and extracting useful information from semi-structured data
  - Community detection

Facebook friends network of the students of my Statistics class in May 2014

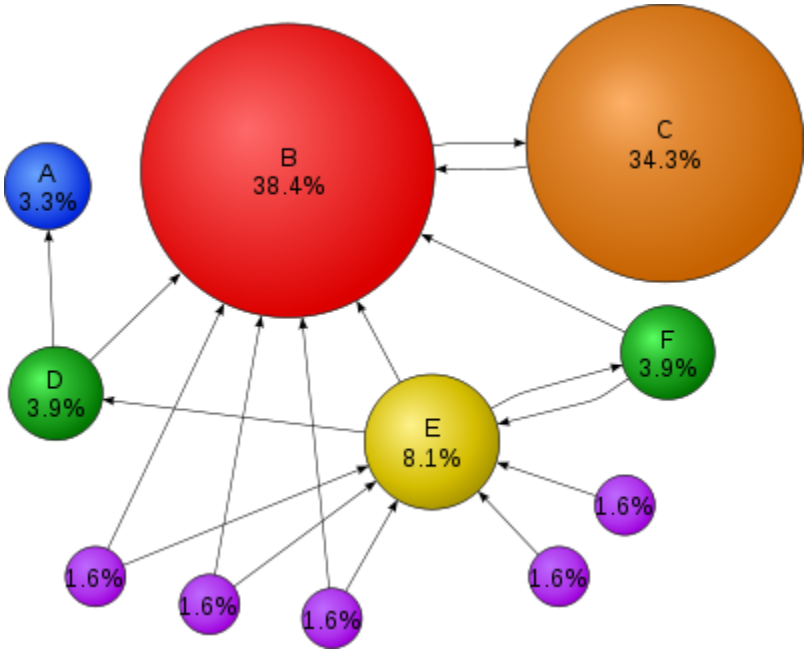


My Facebook network



Graph / Structure mining *(cont.)*

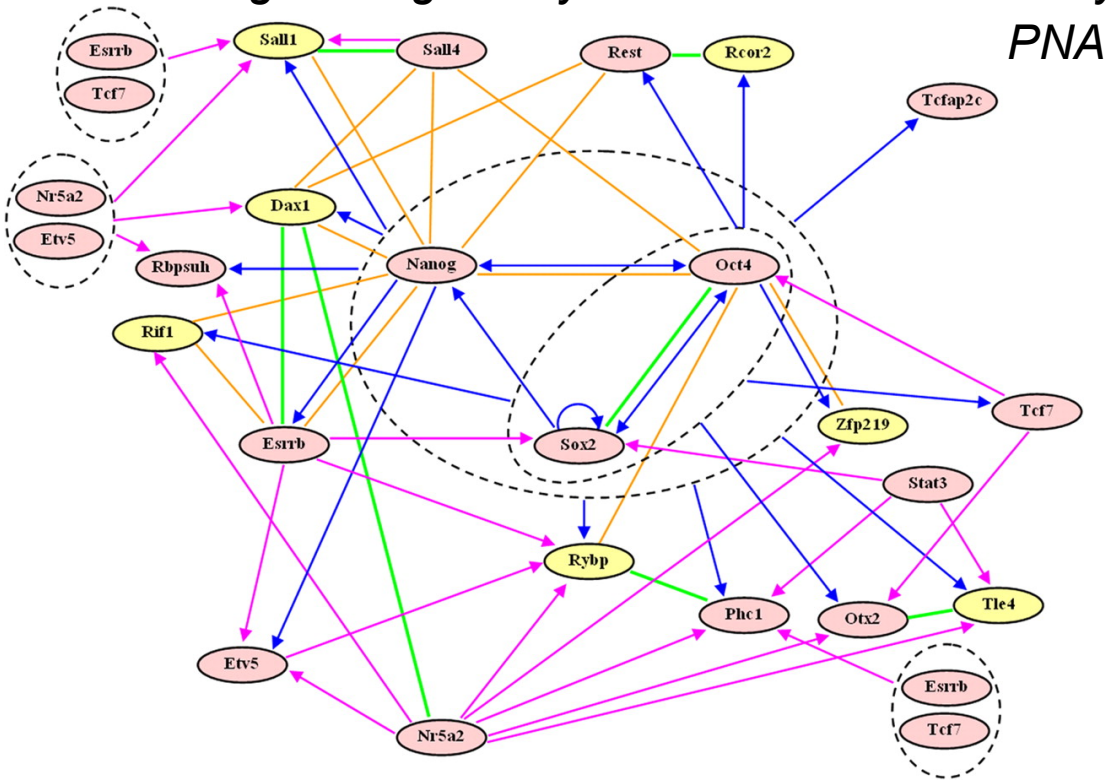
- Example: Google's PageRank
  - The mechanism behind Google search
  - Creates a graph from the links in pages
  - Determines the importance of each page from the number of times that a page is referenced, weighted by the importance of the page that reference it



Graph / Structure mining *(cont.)*

- Example: Analyze the interactions between genes to identify gene regulation
  - Gene regulatory network

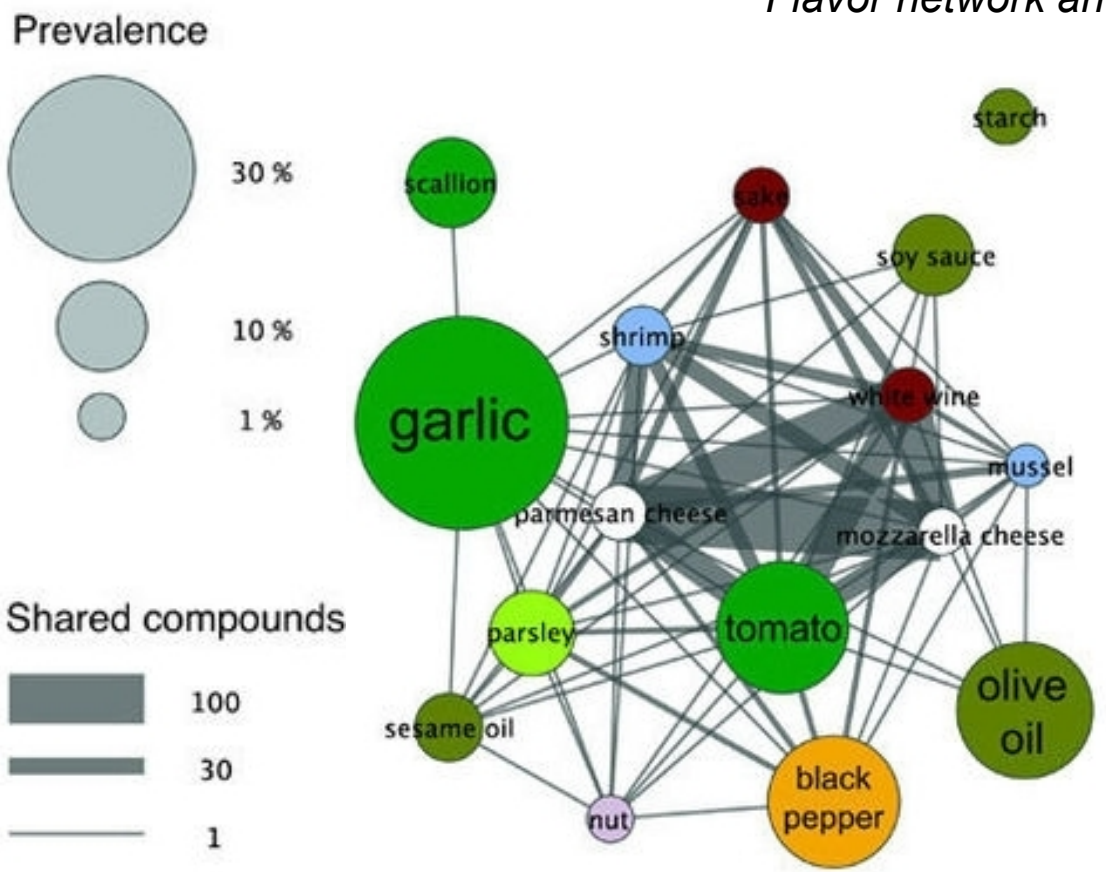
Q. Zhou, H. Chipperfield, D. A. Melton, W. H. Wong
"A gene regulatory network in mouse embryonic stem cells"
PNAS 104(42), 2007



Graph / Structure mining *(cont.)*

- Example: Analyze the co-occurrence graph of ingredients in recipes to analyze food-pairing principles

Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow and A-L Barabási
"Flavor network and the principles of food pairing"
Scientific Reports 1(196), 2011



"ingredients sharing flavor compounds are more likely to taste well together", e.g.

- white chocolate and caviar
- chocolate and blue cheese

Connect ingredients which share at least one flavor compound

Graph / Structure mining *(cont.)*

- More will be covered in this course

Points to ponder

- Is there any limit / trend / pattern in these works?
- What else can be done with data?

- Next point: The next frontier

New development in AI (1/3)

Domingos' Five "Tribes" of ML

Tribe	Strength	Technology
SYMBOLIST	Structure inference	Production rule system Inverse deduction
CONNECTIONIST	Estimating parameters	Backpropagation Deep Learning
BAYESIAN	Weighing evidence	HMM Graphical models
EVOLUTIONARY	Structure learning	Genetic algorithm Evolutionary algorithm
ANALOGIZER	Mapping to novelty	kNN SVM

http://homes.cs.washington.edu/~pedrod/smlr.pptx

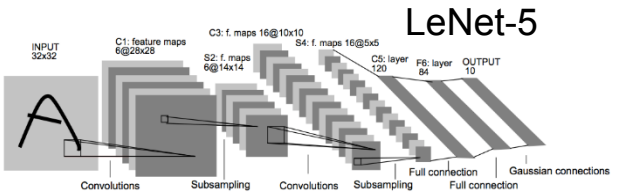
Also read the very funny https://medium.com/intuitionmachine/the-many-tribes-problem-of-artificial-intelligence-ai-1300faba5b60

New development in AI (2/3)

1960~	Naïve Bayes	BAYESIAN
	Decision trees	SYMBOLIST
	Nearest-neighbor	ANALOGIZER
	Perceptron	CONNECTIONIST
1970~	Spectral methods (PCA, MDS)	
	Hidden Markov Model	BAYESIAN
1980~	**Neural networks**	CONNECTIONIST
	Bayesian networks	BAYESIAN
1990~	SVM (kernel trick)	ANALOGIZER
	Ensemble learning (random forest)	TREE HUGGER (Kaggle's favorite)
	Boosting (AdaBoost, XGBoost)	
	Probabilistic Graphical Models (CRF)	BAYESIAN

New development in AI (3/3)

- 1989 AT&T handwritten zip code recognition system
  - Unlike neocognitron which trained layer by layer, LeCun *et al.* trained entirely using backpropagation
  - 3 hidden layers
- By 1998, most of the major components (except ReLU) of convolutional neural networks have been established
  - Convolutional layer
  - Pooling (or subsampling) layer
  - Backpropagation



Relatedly, Hochreiter and Schmidhuber introduced LSTM in 1997

- Main difficulty was in training
- In 2006, Hinton rebranded NN “deep learning” (DL)

The rise of big data sets

1999 MNIST

- Modified National Institute of Standards and Technology database
- Released in 1999 by LeCun *et al.*
- Handwritten digits
- 60,000 training images + 10,000 testing images

2009 CIFAR

- Funded by the Canadian Institute For Advanced Research
- Compiled in 2009 by Krizhevsky *at al.* (U of T) from the unlabeled “80 Million Tiny Images” database
- CIFAR-10: 10 classes of 6,000 images each
- CIFAR-100: 100 classes of 600 images each

ImageNet

- Released in 2009 by Li’s group
- ~15 million annotated images of over 22,000 categories

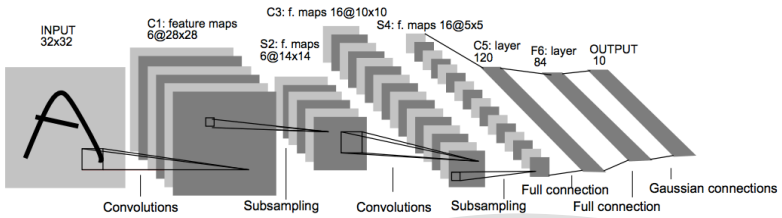
Massive computing power

- GPUs have been studied for matrix multiplication since the GeForce 3 (2001)
  - Pixar's Junior Lamp and Doom 3
  - Pre-Cuda (Cuda 1.0 released June 2007)

- In 2009, Raina *et al.* trained a deep belief network using a GTX 280
  - 70 times speed gain over CPU
- In 2010, a max pooling CNN system (MPCNN) trained with GPU (2 GTX 480 + 2 GTX 580) by Ciresan (Schmidhuber's group) started to win competitions
  - Registered new record in MNIST recognition rate
  - ICDAR Chinese handwriting 2011
    - 60 times speed gain over CPU reported
    - Schmidhuber's group won the ICDAR handwriting competition 2009 for French, Farsi, Arabic with CPU-trained LSTM
  - IJCNN 2011 traffic sign recognition competition

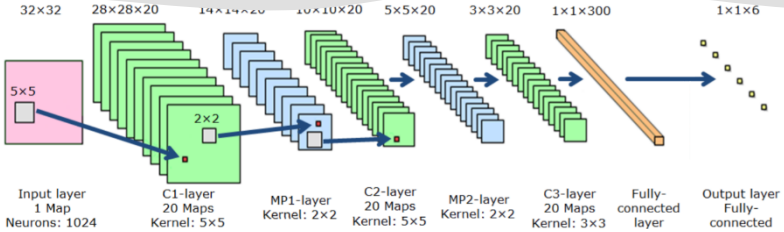
Monstrous CNN

1998 LeNet-5



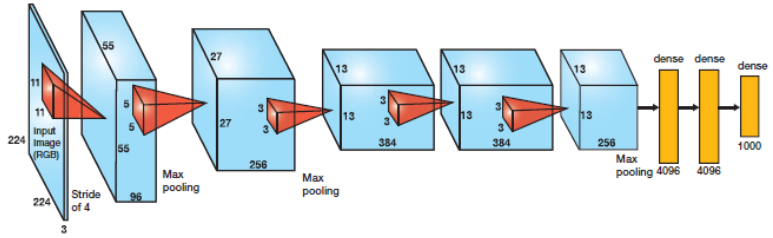
7 layers

2011 MPCNN



4~7 layers

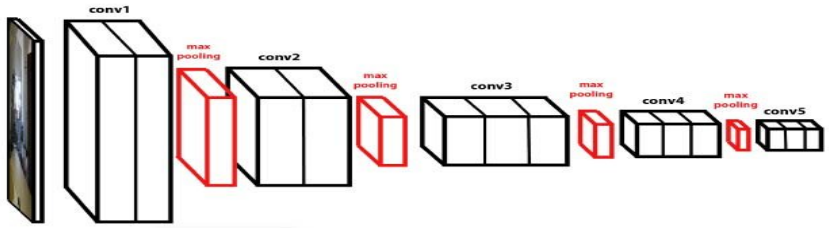
2012 AlexNet
(ReLU)



8 layers

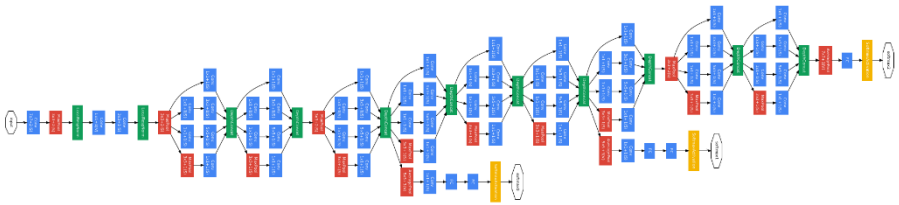
ZFNet similar to AlexNet but more elegant

2014 VGG
(Simple, just deep)



19 layers

2014 GoogLeNet
(Inception)



22 layers

Guess the number of layers in ResNet152

Conquests of CNN (and DL)

2012 AlexNet (U of T) wins ImageNet (ILSVRC) competition

LSTM wins ISBI challenge on brain segmentation

MPCNN wins ICPR challenge on mitosis detection

2013 ZFNet (NYU) wins ImageNet competition

MPCNN defended mitosis detection in MICCAI challenge

2014 GoogLeNet and VGGNet wins ImageNet competition

2015 ResNet (Kaiming He@MSRA) wins ImageNet competition

AlphaGo beats human (MC tree search + deep learning)

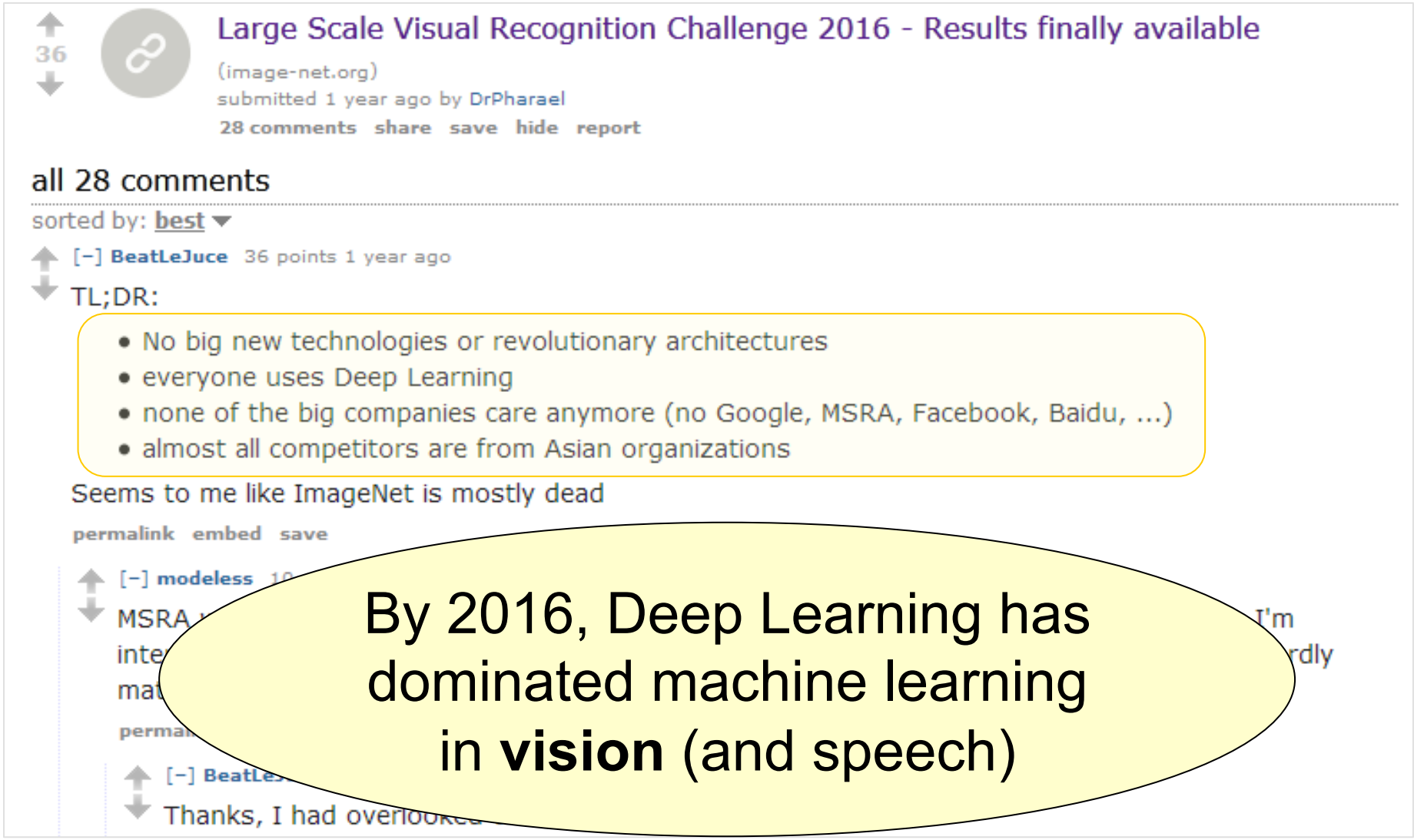
PReLU beats human at ImageNet classification

2016 Camelyon beats human at metastatic cancer detection

2017 Libratus beats human at Poker

2018 SLQA+ (iDST) and r-net+ (MSRA) beat human at Stanford Question Answering Dataset (SQuAD)

You didn't mention ImageNet 2016?



https://www.reddit.com/r/MachineLearning/comments/54jiyy/large_scale_visual_recognition_challenge_2016/

What we can readily do with CNN

- Recognize faces
- Identify gesture/action
- Identify objects
- Classify scenery
- Analyze visual similarity between objects

Points to ponder

- What hit me? (Too much data and too much computing power, as usual)
- Now what are we going to do with all the image data?
- Next point: What you need to get started

Standard data formats

- CSV / TSV / Ctl+A-delimited
- HTML / XML / JSON / YAML
- W3C's Resource Description Framework (RDF)
  - **RSS** (newsfeed), *etc.*
- Serialized data structures / objects
  - External Data Representation (XDR)
  - Python's *pickle*
  - Java's *Serializable*, *etc.*
- HDF5 / netCDF
  - Machine independent data format
- The Web Ontology Language (OWL)
- Various image file formats: JPG / PNG / PGM / TIFF / GIF

Obtaining data

- Typically obtained from the internet automatically using UNIX scripts
  - Commonly used languages
    - Shells (bash, tcsh, etc.)
    - Perl
    - Python
    - Ruby
- Services often provide protocols for data access
  - e.g. Facebook, Twitter
- Standard parsers for standard data formats
  - Apache Tika: extracts metadata and text from >100 different file types
  - e.g. PPT, XLS, PDF, XML, ODF, EPUB, ZIP, MP4, FLV, …

Obtaining data *(cont.)*

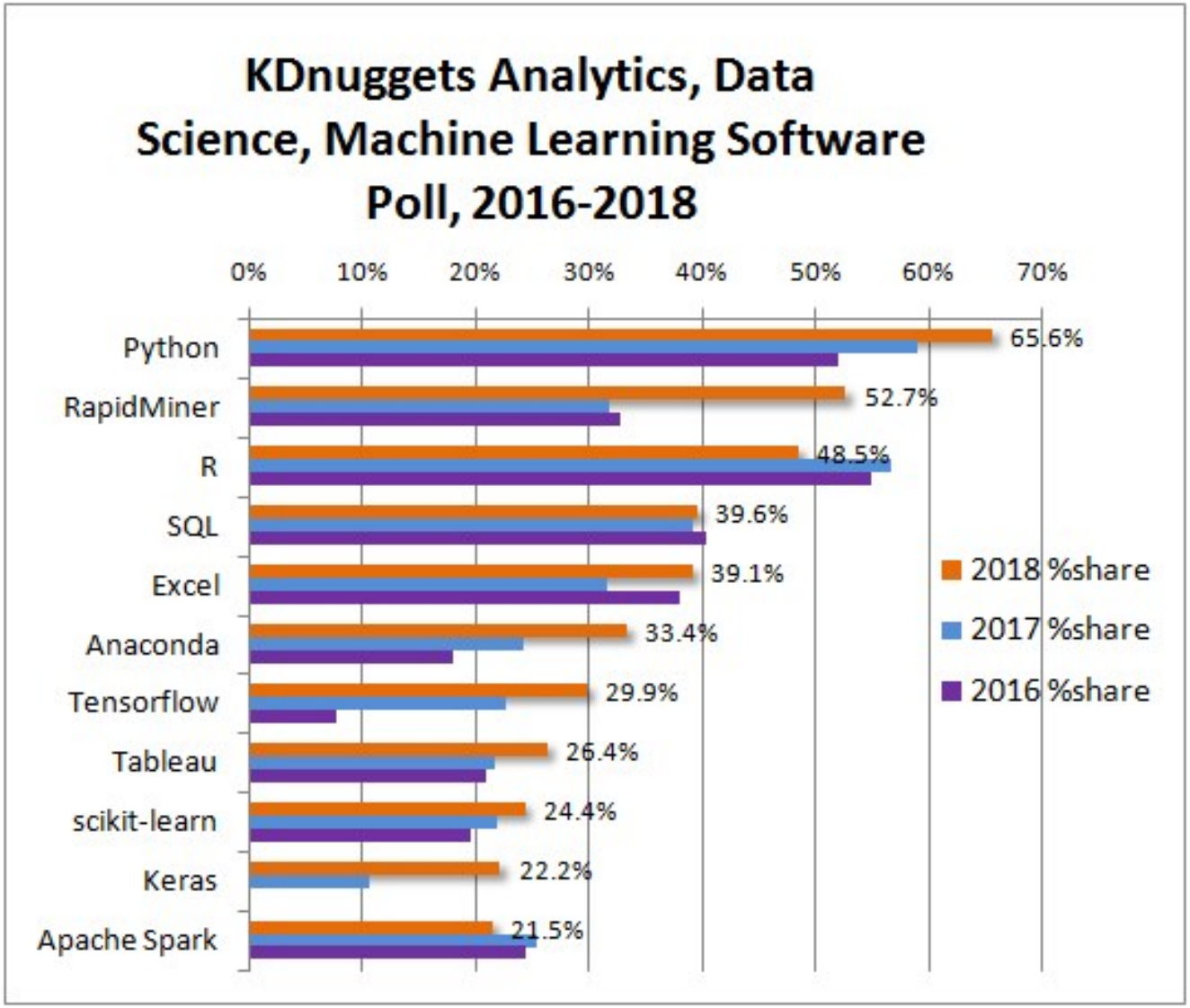
- Web-scraping
  - complicated due to the need for user-interaction in retrieving online documents, especially from DHTML-rich websites
    - e.g. Some pages load only when you scroll to the end of the page
  - Many tools allow us to mimic user-interaction as if accessed through a web-browser
  - **Scripting environment**
    - Selenium Webdriver (Java, Python, C#, Groovy, Perl, PHP, Ruby)
    - PhantomJS (Javascript)
    - Capybara (Ruby)
    - Zombie.js (Javascript)
    - WebHarvest (Java)
    - *etc*
  - **Learn from user-interaction**
    - Selenium IDE
    - Commercial ($$$): Mozenda, Import.io
    - *etc*

Obtaining data *(cont.)*

- Online image repositories

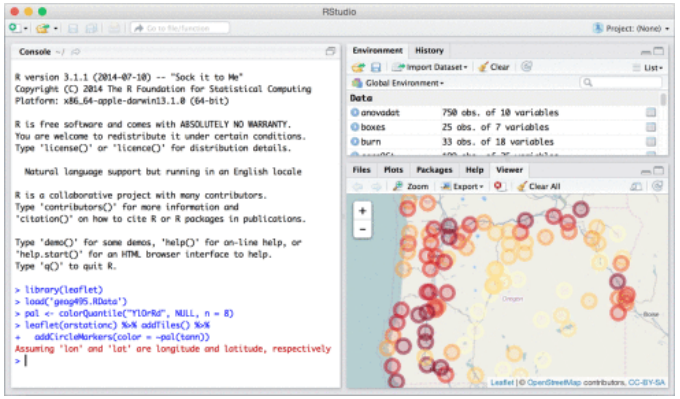
- Stalk your friends

Analyzing data



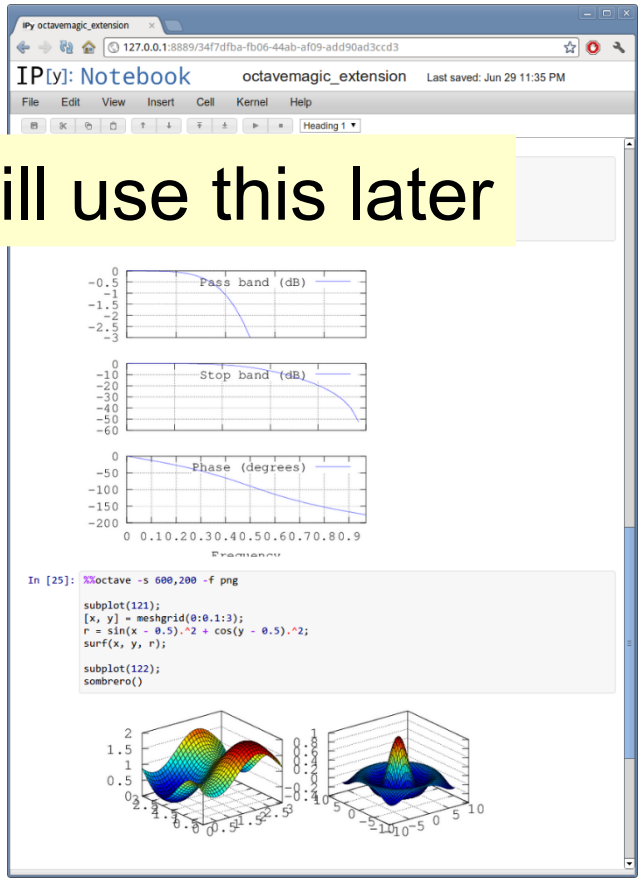
https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

Analyzing data *(cont.)*



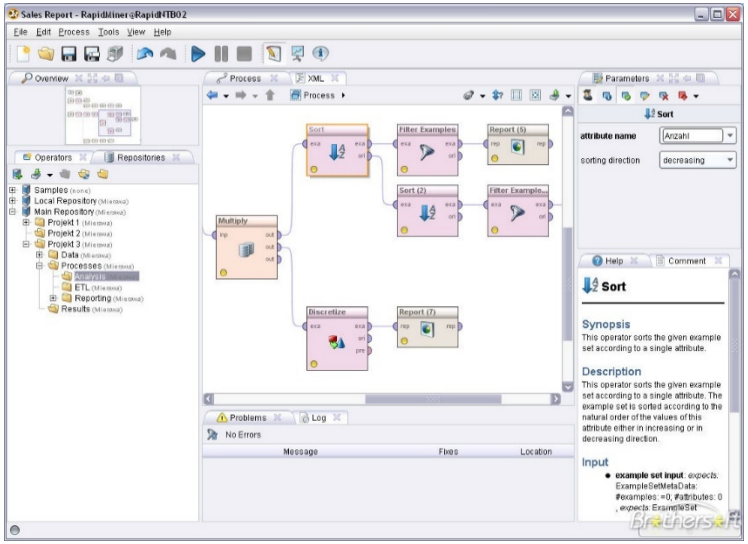
R

IPython/Jupyter



We will use this later

RapidMiner



DL Frameworks

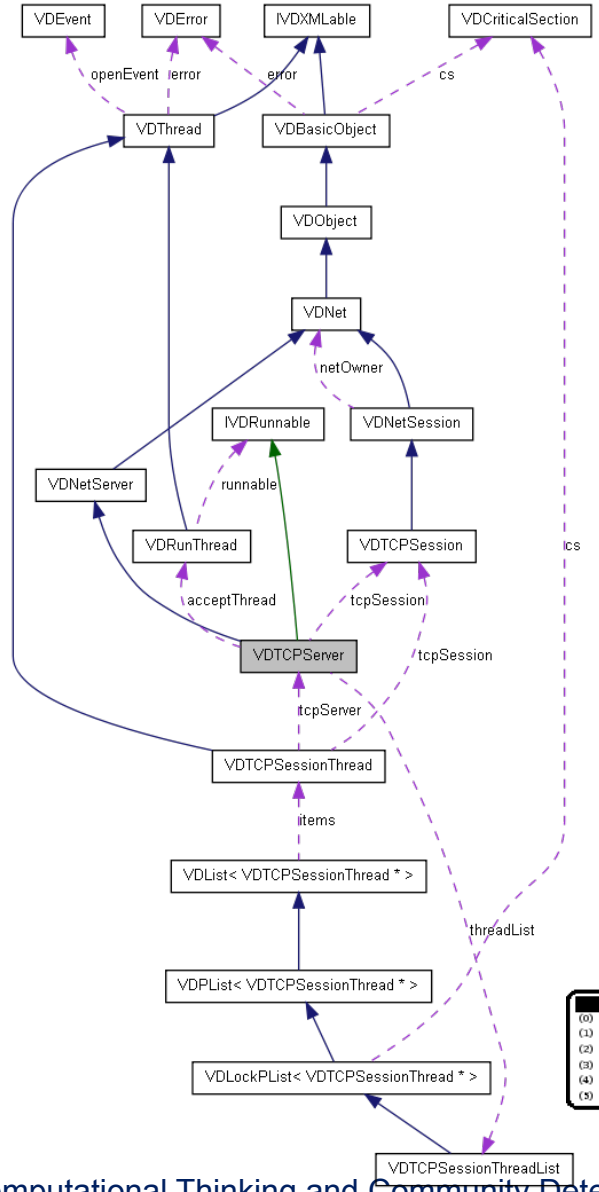
- Library functions to DL components
- Backpropagation gradient computation
- GPU-support
- Popular frameworks

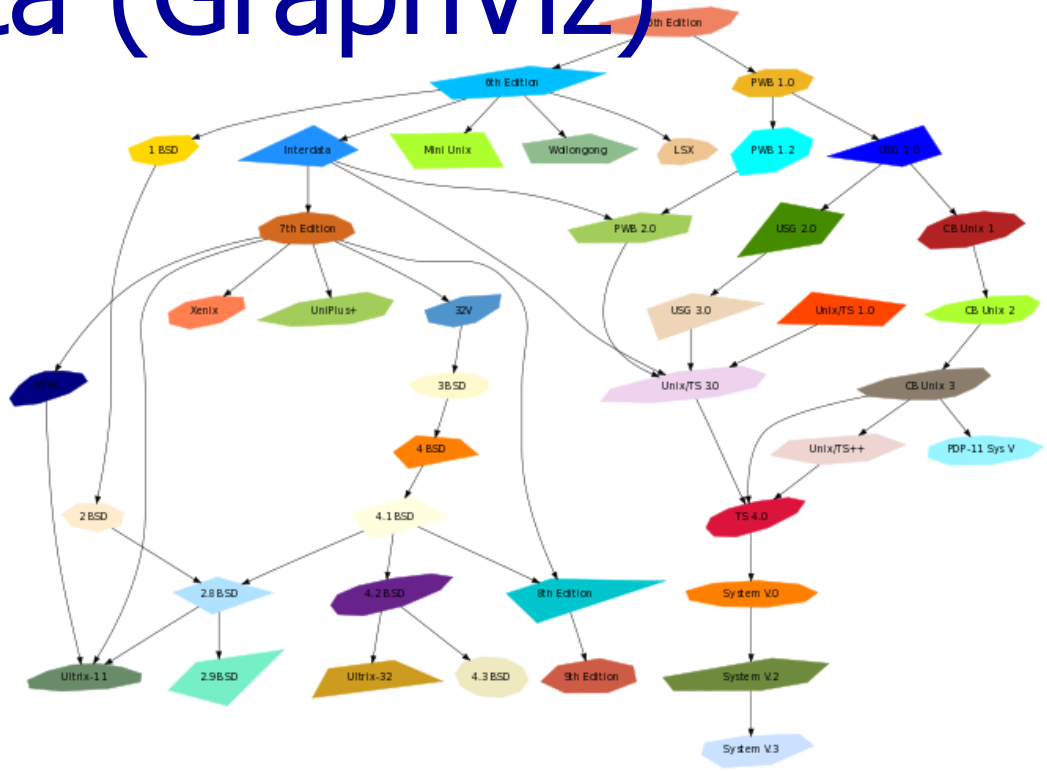
- Theano family
  - Theano (U of Montreal)
  - TensorFlow (Google)
- Torch family
  - Torch (NYU)
  - PyTorch (Facebook)
- Caffe family
  - Caffe (UC Berkeley)
  - Caffe2 (Facebook)
- CNTK (Microsoft)
- MXNet (Amazon)
- Wrapper
  - keras (works with Theano, TensorFlow, CNTK)

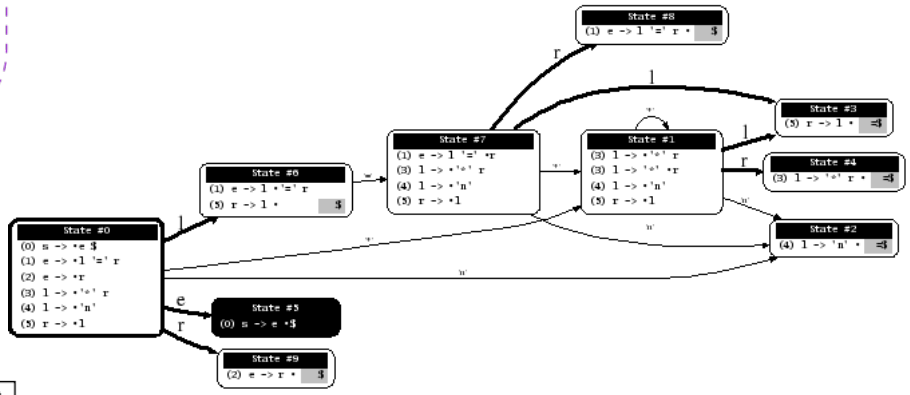
We will use this later

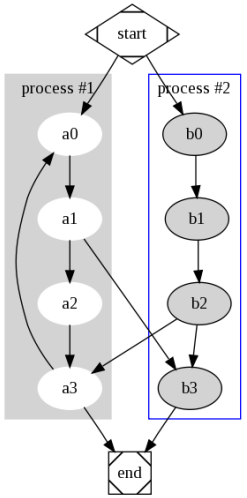
Visualizing data (Gephi)

Visualizing data (GraphViz)









Visualizing data (Tableau)

Tale of 100

Author: **Christian Chabot** published on Wall Street Journal

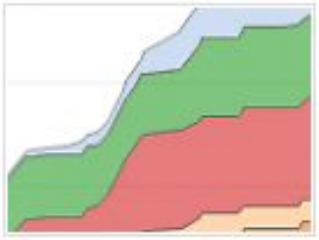
How fast do successful tech companies grow? The Wall Street Journal posted this visualization that compares the performance of 100 fast growing software companies.

Standalone version not free

College Football Recruiting

Author: **Scott Wasserman** published on Tableau Software

This year's national championship might have just been decided, but the struggle for years of future domination has already started. The recruiting wars are in full swing. Use this viz to see how your team stacks up.



Tech IPOs Finish the Year Strong

Author: **Daniel Hom** published on IPO Dashboards

Tech IPOs have been under scrutiny this year but ended up leading the way in total capital raised during 2011.

Forest Fire Hot Spots

Author: **Ben Jones** published on DataRemixed

Today's viz by Ben Jones is all about US forest fires; size, cause, where and when. Filter down to a specific year to see specific details and trends.

Points to ponder

- Is it really that hard to get the tools?
- How about the data? Which is more difficult to get?

- Next point: With great power come great responsibility

Biases in data

- Learned models (especially neural networks) are approximate representations of their input data
  - Garbage in, garbage out
  - Biases in data will be reflected in the learned models
- Biases may lead to discrimination
  - Commonly accepted protected groups
    - Color / Race
    - Gender
    - Age
    - Religion
  - Less commonly accepted protected groups
    - Sexual orientation
    - Immigrants

Consequences of biases

Privacy and cookies | Jobs | Dating | Offers | Shop | Puzzles | Investor

The Telegraph

Home Video News World Sport Business Money Comment Culture

Apple | iPhone | Technology News | Technology Companies | Technology Reviews

HOME » TECHNOLOGY » GOOGLE

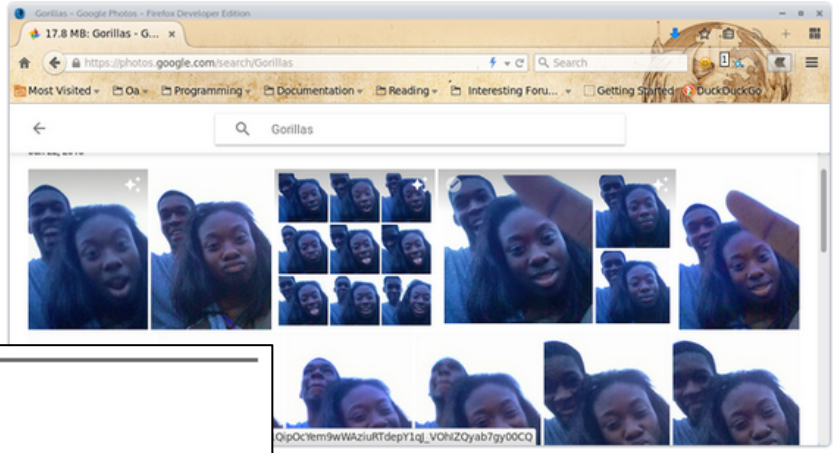
Google Photos labels black people as 'gorillas'

Google has removed the 'gorilla' tag from its new Photos app, aft
found to be misidentifying images of black people

jalciné
@jackyalcine

Follow

And it's only photos I have with her it's doing this with (results truncated b/c personal):



115

By Sophie Curtis
11:20AM BST 01 Jul 2015

Follow 11K followers

Google has removed the 'gorilla' tag from its new Photos app, after a user noticed it had filed a number of photos of him and his black friend in an automatically generated album named 'gorillas'.

The affected user, computer programmer Jacky Alciné, took to Twitter to post proof of the Google Photos error, along with the question: "What kind of sample image data you collected that would result in this son?"

Consequences of biases

Newsweek

ROBOTS WITH ARTIFICIAL INTELLIGENCE BECOME RACIST AND SEXIST—SCIENTISTS THINK THEY'VE FOUND A WAY TO CHANGE THEIR MINDS

BY **ANTHONY CUTHBERTSON** ON 10/26/17 AT 7:34 AM

In 2016, Microsoft released a "playful" chatbot named Tay onto Twitter designed to show off the tech giant's burgeoning artificial intelligence research. Within 24 hours, it had become one of the internet's ugliest experiments.

By learning from its interactions with other Twitter users, Tay quickly went from tweeting about how "humans are super cool," to claiming "Hitler was right I hate the jews."

Consequences of biases

- AI products
  - Search engines, chatbots, assistants
  - Lack of human values poses obstacles to public trust
  - May hurt acceptance of general AI

- Reporting / Journalism
  - Deplete public trust
- Decision-making
  - Affect prediction accuracy
  - Even worse, law offences…

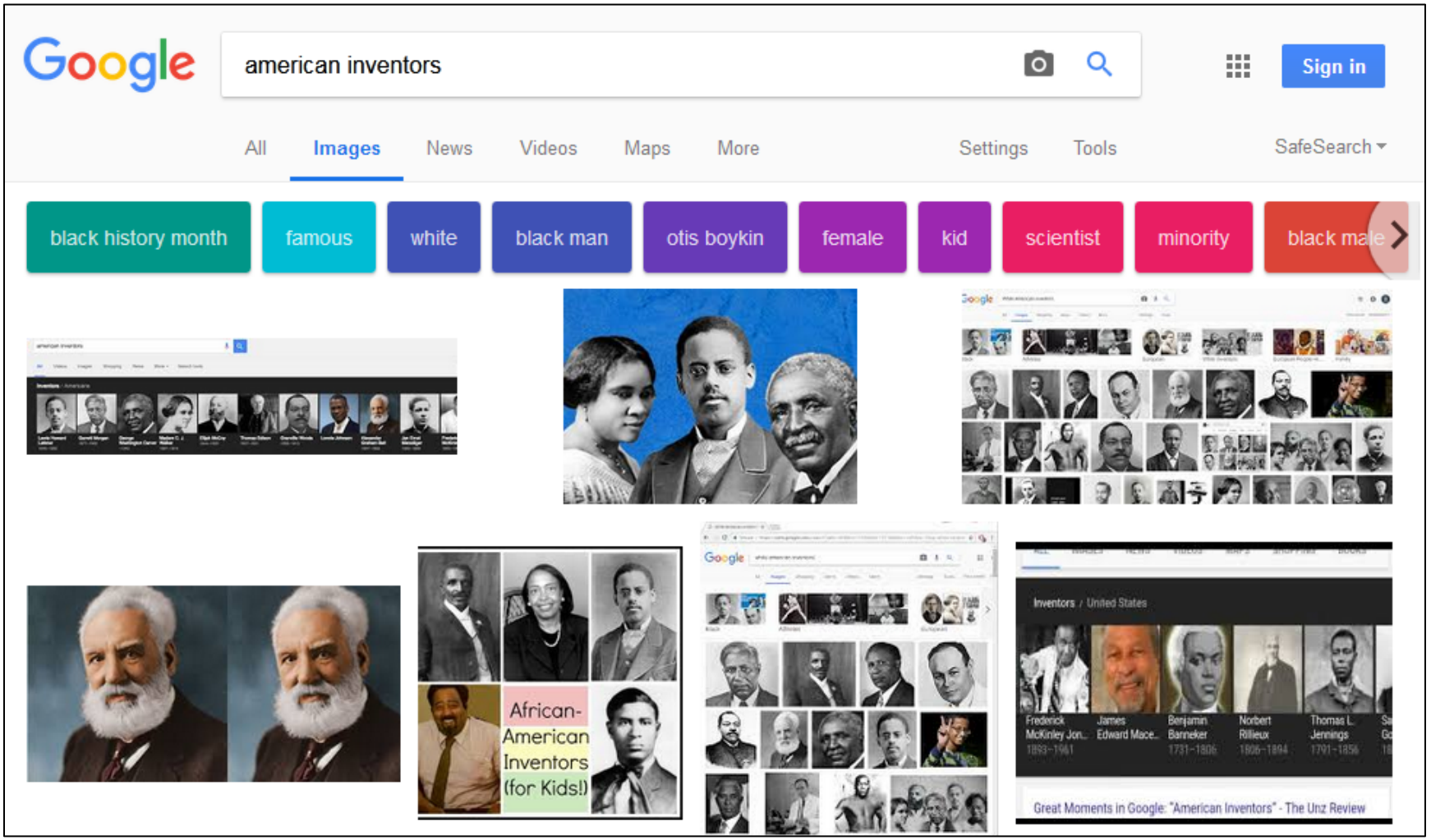
Law offences

- Laws against discrimination of protected classes
  - Hiring (US, UK, Canada, EU, *etc.*)
  - Loan applications (US, UK, Canada, most of EU, *etc.*)
  - School enrollment (most countries)
  - Housing (US)
    - Under US Fair Housing Act, housing providers cannot discriminate against a tenant based on his or her sex, except: (1) a home in which the landlord lives and rents out only one room; (2) where the renters will be sharing a living space with the landlord; (3) single-sex dormitories at colleges or other educational institutions.
  - *etc.*

Debiasing

- Many approaches to mitigate bias
  - Independence
    - Leads to statistical parity
    - Restrictive
  - Separatedness
    - Relaxed condition
    - Conditional independence
- Objections to corrections
  - Sub-optimal utility
  - Self-fulfilling prophecy
  - Long-term goal questionable

Dangers of debiasing



Possibility: Perhaps the search engine confused “African-American” with “American”?

The extent of statistical decision making

- Predictions are often meaningful only when they are to be repeated over many times, and the results aggregated over all the trials
- Use of such predictions in other situations needs caution

A cab was involved in a hit and run accident. Two kinds of cabs operate in the city, Yellow and Green. You know that:

- 91% of the cabs in the city are Green
- 9% of the cabs in the city are Yellow

A witness says the cab was Yellow. When tested, the witness was able to identify the cabs correctly 90% of the time.

$$P(\text{Yellow} \mid \text{witness}) = P(w|Y)P(Y) \,/\, [\, P(w|Y)P(Y) + P(w|!Y)P(!Y)\,]$$
$$= .9 * .09 \,/\, (\, .9 * .09 + .1 * .91\,)$$
$$= .471$$
$$P(\text{Green} \mid \text{witness}) = 1 - .471 = .529$$

Points to ponder

- Be responsible

- Next point: Big Data trends in the industry

Farecast.com (2003)

- Most likely the first ever data product that caught media's attention
- Founded in 1999
- Launched in 2003
- Uses Big Data to predict air ticket prices

- Collected over 175 billion airfare observations as of 2007
- Acquired by Microsoft in April, 2008
- Officially taken offline in April, 2014

Use kayak.com instead!

Data-driven appliances control (2012)

- Analyzes optimal air-con setup

ComputerWeekly.com

TechTarget

News | IT Management | Industry Sectors | Technology Topics | Blogs | Multimedia | Vendor Content | Jobs | Premium Content | Awards | SEARCH

Home > Topics > Information management > Big data analytics > Tesco uses big data to cut cooling costs by up to €20m

ANALYSIS

Tesco uses big data to cut cooling costs by up to €20m

Bill Goodwin
Wednesday 22 May 2013
11:00

Share 74 | g+1 2 | Tweet 95

Supermarket chain Tesco aims to save over €20m a year by using sophisticated business intelligence technology to ensure its refrigerators operate at the right temperature.

The move will help the retailer cut its refrigeration energy costs by up to 20% across 3000 stores in the UK and Ireland.

An experimental research project between Tesco in Ireland and IBM Dublin's research laboratories has shown the retailer can save a significant part of its total energy costs by optimising the performance of its in-store refrigerators.

Latest News

- £15bn gap in VAT and excise might be filled by big data analytics
- Quiksilver expands e-commerce and sees 65% increase in revenue
- The Interview to screen despite threats of furt attacks
- Top 10 cyber security stories of 2014
- Top 10 IT outsourcing stories of 2014

MORE NEWS

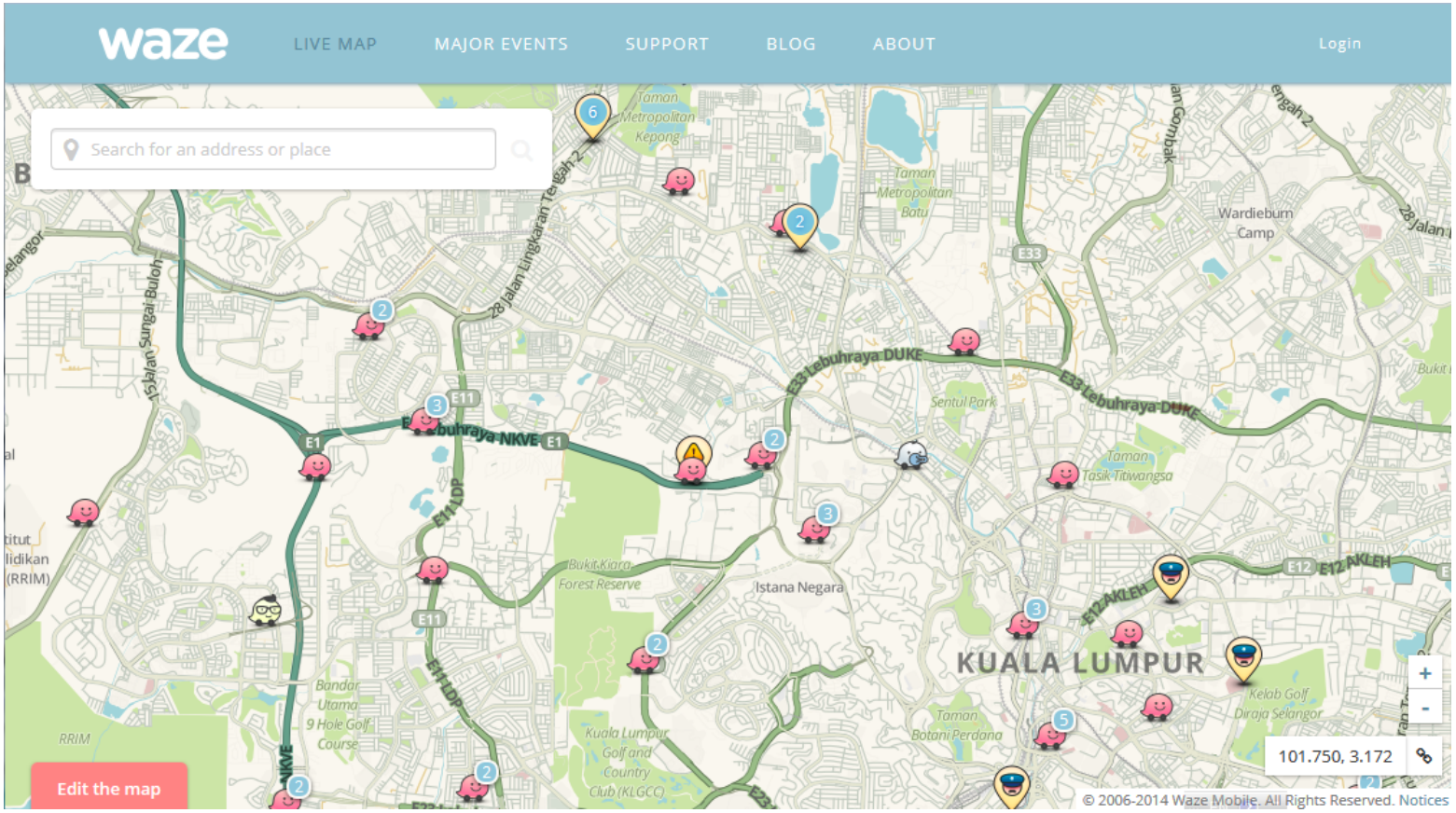
Free Download

Hot Topics

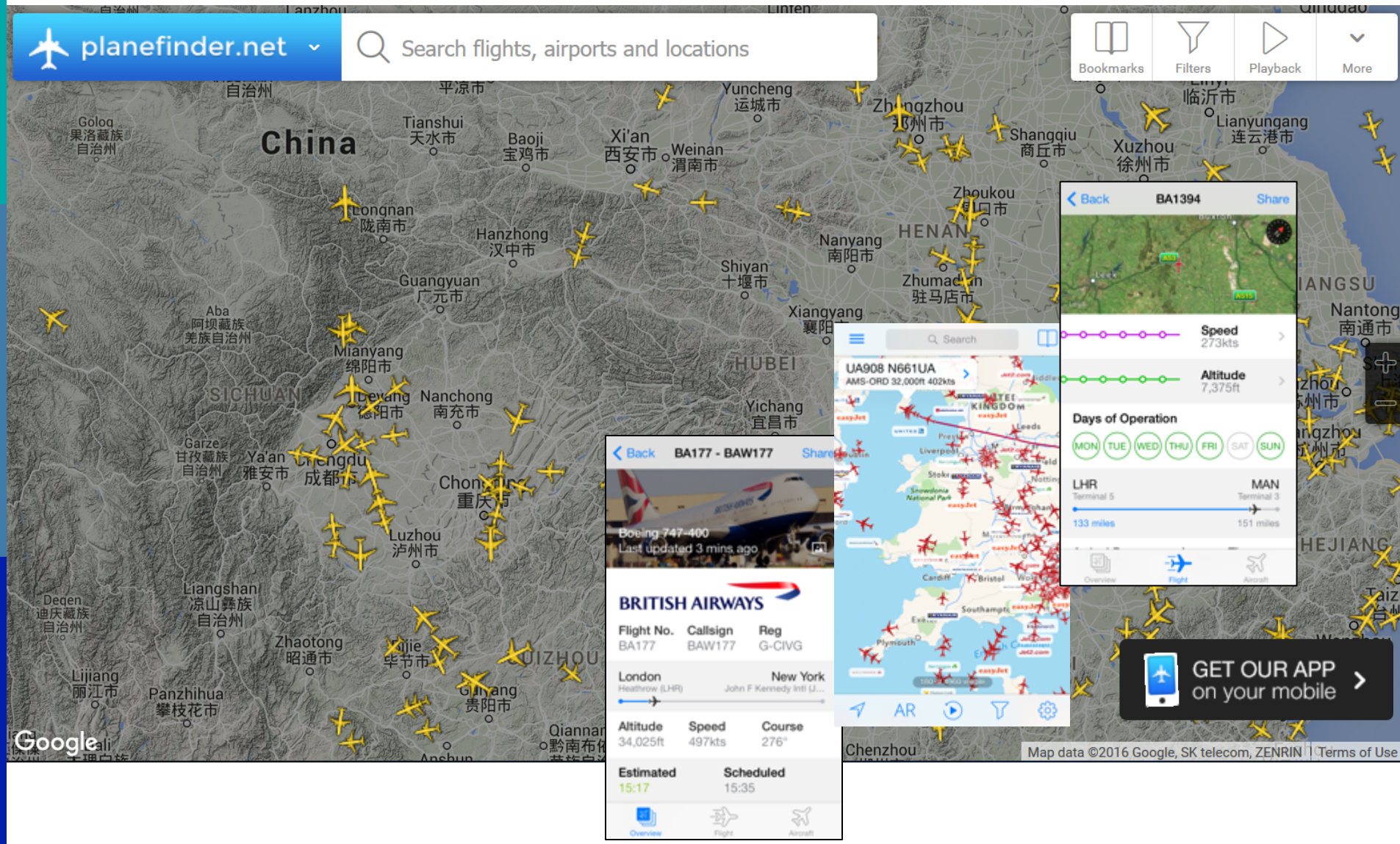
- Hitachi Innovation Forum: Tokyo

Waze (2012)

- Revolutionary symbiotic, collaborative data collection (like Wiki)
- Smartphone app collects car movement data from the installed base (crowdsourcing) for traffic report
- Apple, Facebook, and Google all tried to buy Waze – sold to Google

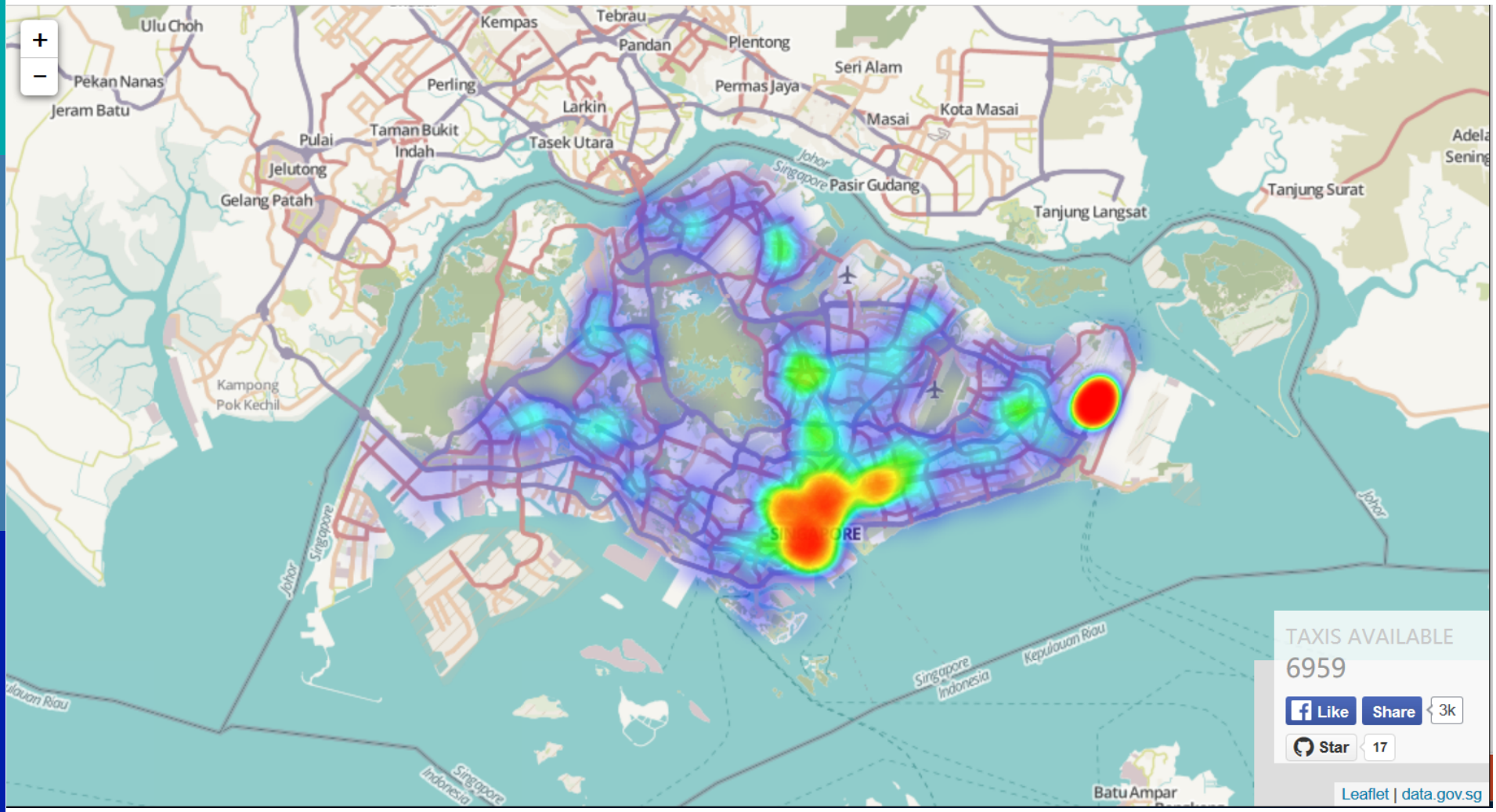


Tracking commercial aviation flights

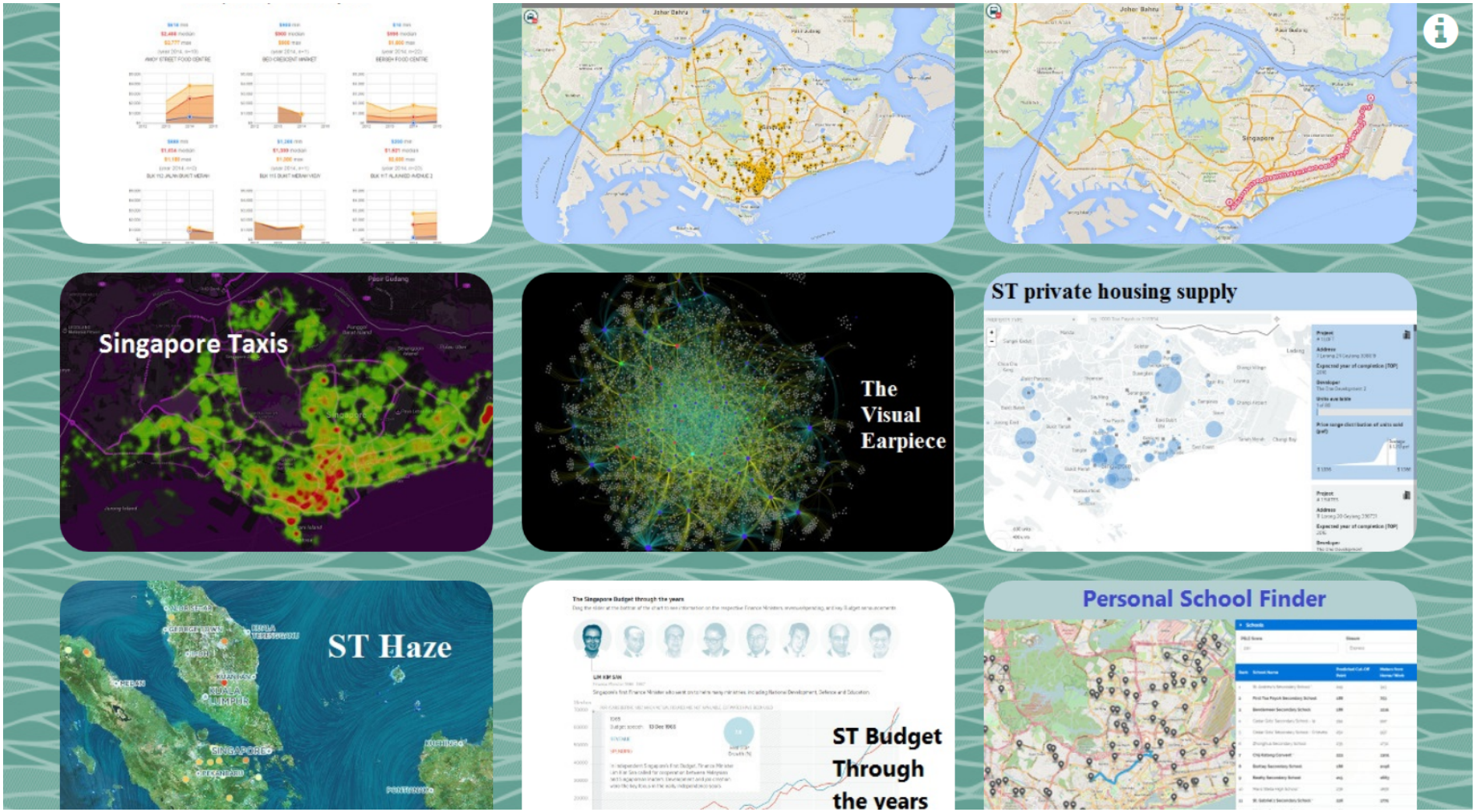


Tracking taxis

- http://www.comp.nus.edu.sg/~josepht/taxi.html



More @ www.viz.sg



"Unconventional" AI (2014)

- Fast-food drive-thru menus

Companies are also using big data to improve or alter their business models. According to case studies released by Gartner, a technology research group and leader in big data consulting, a major fast-food retailer "is training cameras on drive-through lanes to determine what to display on its digital menu board. When the lines are longer, the menu features products that can be served up quickly; when the lines are shorter, the menu features higher-margin items that take longer to prepare." Precise calculations of waste means better and better ideas of how to increase efficiency at every level. Operations, manufacturing, supply chain management — the applications for big data across industry are as wide as the human imagination. Gartner also cites a thing called dark data — information collected by an organization for one purpose, long ago, and then reanalyzed using new tools for insight still buried within it.

http://www.institutionalinvestor.com/blogarticle/3333390/blog/everything-you-need-to-know-about-big-data-to-keep-your-job.html#.VJxxNv9BbA

“Unconventional” AI (2017)

- G-Assist

More recent advances

- Move from big businesses to everyday personal decision
  - Which smartphone to buy
  - Which movie to watch
  - Which language to learn
- Natural human-computer interactions
  - Chatbots
  - Computer-generated speech/video

Points to ponder

- Data ⇒ analyses opportunities
- Everybody is doing something that seems impossible yesterday
- Take action
  - **Know the tools**
  - **Look for the data you need**