# 大數據分析方法
# Introduction of Big Data Analytics

曾意儒 助理教授

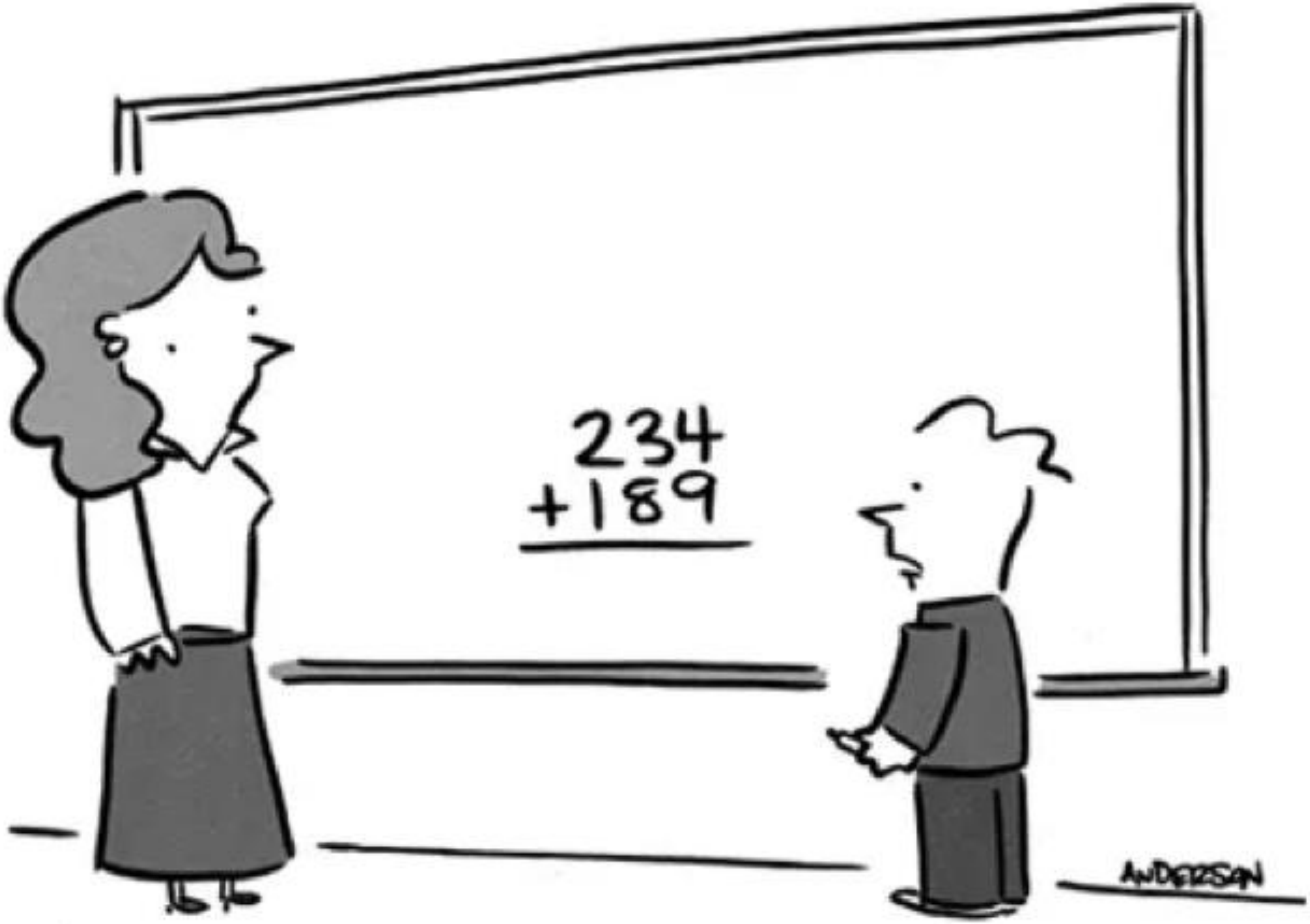
長庚大學 資訊管理學系

"After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snockered. Who's with me?!"

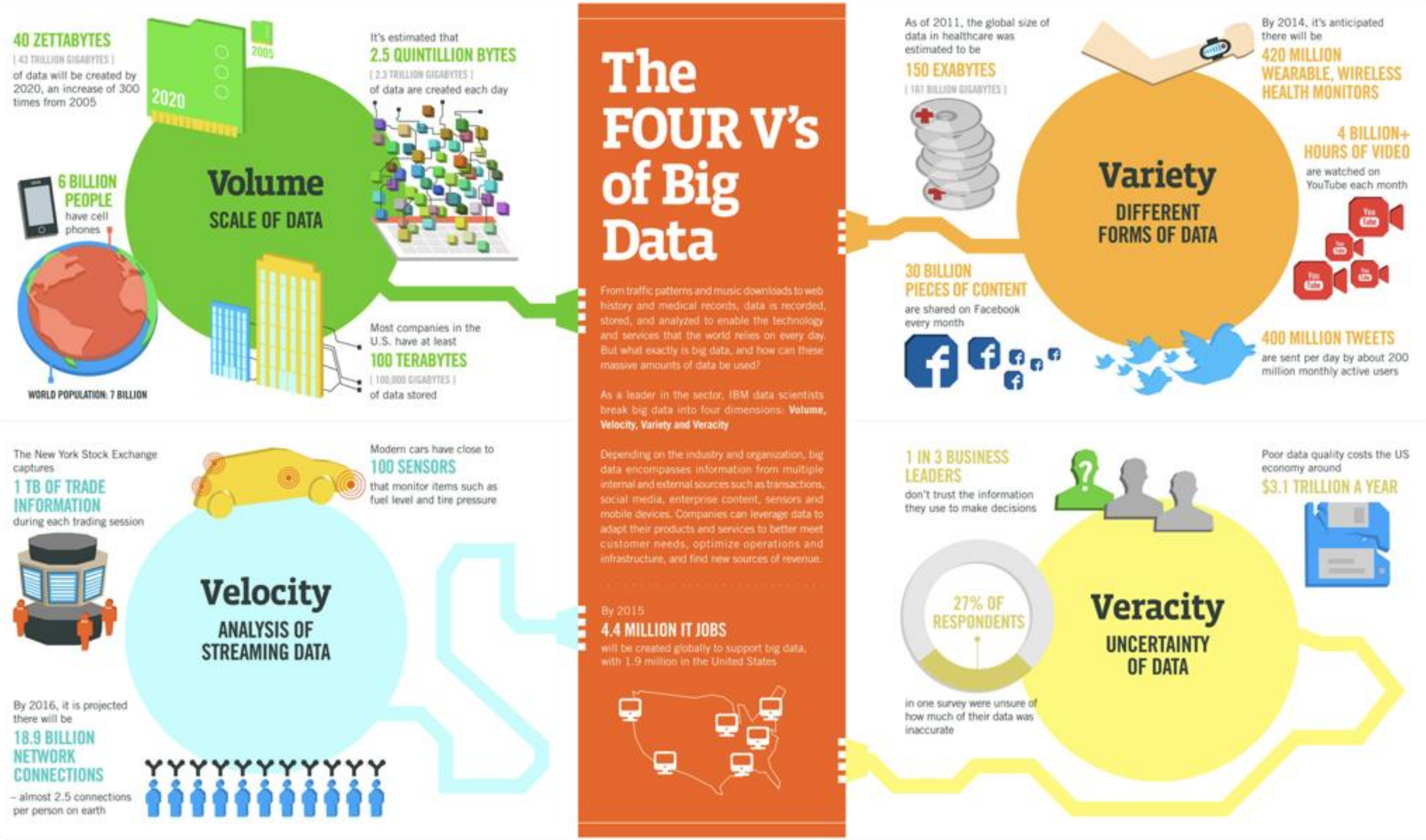
# Outlines

- What is Big Data?
- What is Big Data Analytics?
- Why We Need Big Data Analytics?
- What is Data Science?

# What is Big Data?



"Does this count as big data?"



# The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

**4.4 MILLION IT JOBS**

will be created globally to support big data, with 1.9 million in the United States

## Volume

SCALE OF DATA

**40 ZETTABYTES**

[ 43 TRILLION GIGABYTES ]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

**2.5 QUINTILLION BYTES**

[ 2.3 TRILLION GIGABYTES ]

of data are created each day

**6 BILLION PEOPLE**

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

**100 TERABYTES**

[ 100,000 GIGABYTES ]

of data stored

## Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

**1 TB OF TRADE INFORMATION**

during each trading session

Modern cars have close to

**100 SENSORS**

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

**18.9 BILLION NETWORK CONNECTIONS**

– almost 2.5 connections per person on earth

## Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**

[ 161 BILLION GIGABYTES ]

By 2014, it's anticipated there will be

**420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO**

are watched on YouTube each month

**30 BILLION PIECES OF CONTENT**

are shared on Facebook every month

**400 MILLION TWEETS**

are sent per day by about 200 million monthly active users

## Veracity

UNCERTAINTY OF DATA

**1 IN 3 BUSINESS LEADERS**

don't trust the information they use to make decisions

Poor data quality costs the US economy around

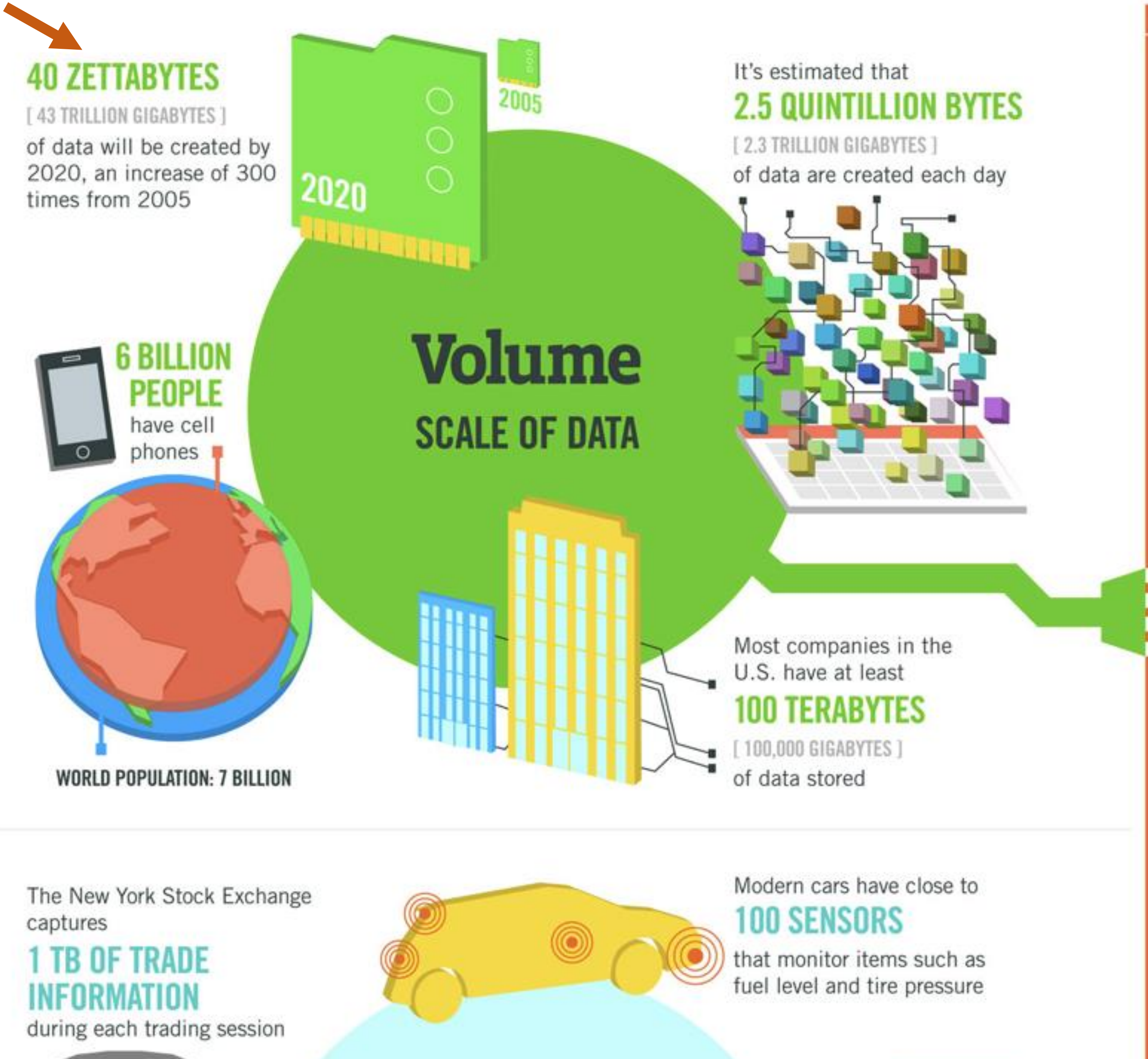
**$3.1 TRILLION A YEAR**

**27% OF RESPONDENTS**

in one survey were unsure of how much of their data was inaccurate

**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM



**40 ZETTABYTES**

[ 43 TRILLION GIGABYTES ]

of data will be created by 2020, an increase of 300 times from 2005

**6 BILLION PEOPLE**

have cell phones

WORLD POPULATION: 7 BILLION

## Volume

### SCALE OF DATA

It's estimated that

**2.5 QUINTILLION BYTES**

[ 2.3 TRILLION GIGABYTES ]

of data are created each day

Most companies in the U.S. have at least

**100 TERABYTES**

[ 100,000 GIGABYTES ]

of data stored

The New York Stock Exchange captures

**1 TB OF TRADE INFORMATION**

during each trading session

Modern cars have close to

**100 SENSORS**

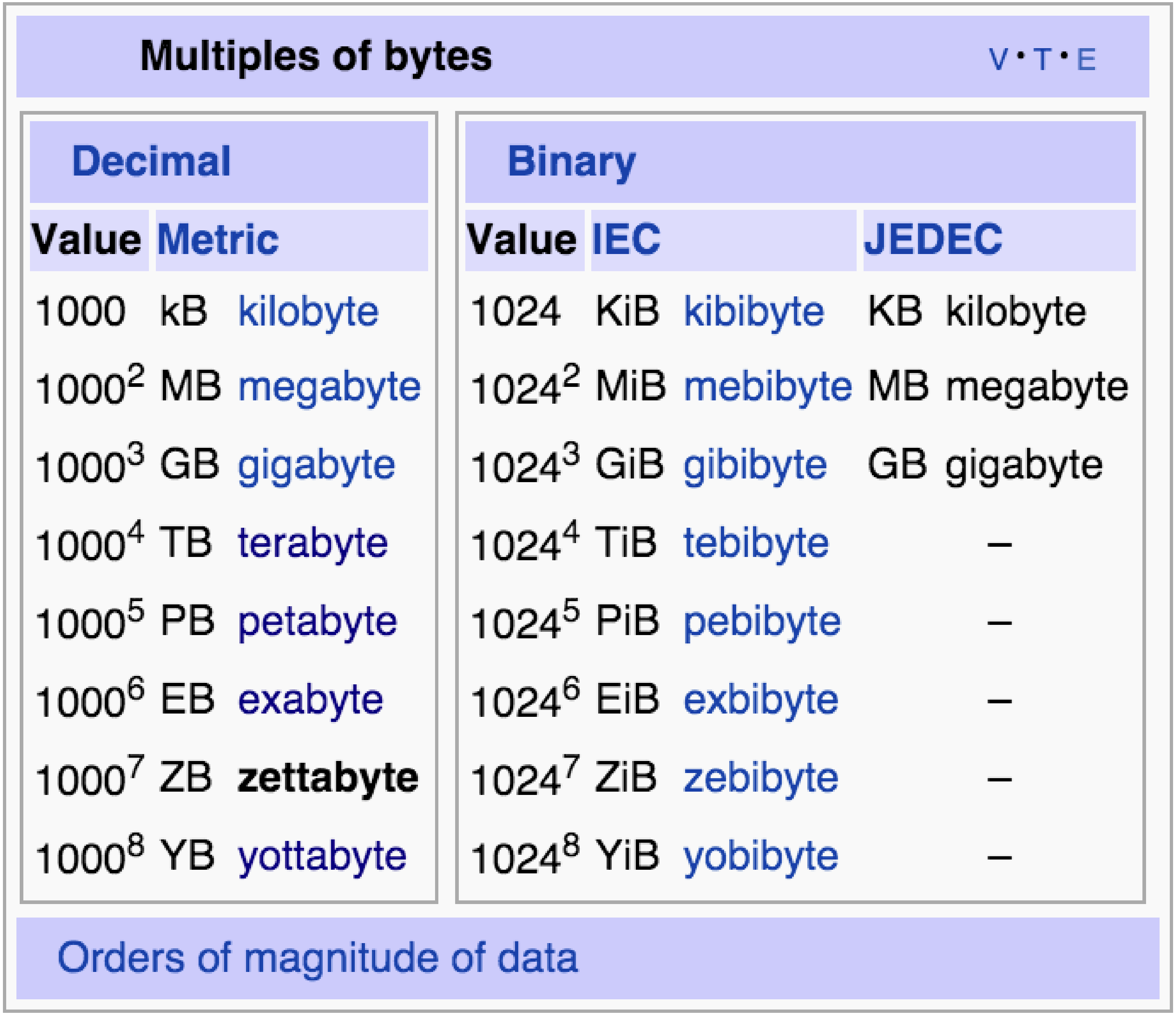
that monitor items such as fuel level and tire pressure

# The FOUR of Big Data

From traffic patterns and music d
history and medical records, d
stored, and analyzed to enable
and services that the world reli
But what exactly is big data, and
massive amounts of data be use

As a leader in the sector, IBM
break big data into four dime
**Velocity, Variety and Veracity**

Depending on the industry and o
data encompasses information
internal and external sources such
social media, enterprise conte
mobile devices. Companies can
adapt their products and service



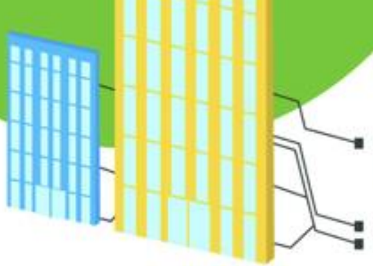
## Multiples of bytes

V · T · E

|  | Decimal |  |  | Binary |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | Value | Metric |  | Value | IEC |  | JEDEC |  |
| $10^3$ | 1000 | kB | kilobyte | 1024 | KiB | kibibyte | KB | kilobyte |
| $10^6$ | $1000^2$ | MB | megabyte | $1024^2$ | MiB | mebibyte | MB | megabyte |
| $10^9$ | $1000^3$ | GB | gigabyte | $1024^3$ | GiB | gibibyte | GB | gigabyte |
| $10^{12}$ | $1000^4$ | TB | terabyte | $1024^4$ | TiB | tebibyte | – |  |
| $10^{15}$ | $1000^5$ | PB | petabyte | $1024^5$ | PiB | pebibyte | – |  |
| $10^{18}$ | $1000^6$ | EB | exabyte | $1024^6$ | EiB | exbibyte | – |  |
| $10^{21}$ | $1000^7$ | ZB | **zettabyte** | $1024^7$ | ZiB | zebibyte | – |  |
| $10^{24}$ | $1000^8$ | YB | yottabyte | $1024^8$ | YiB | yobibyte | – |  |

Orders of magnitude of data

WORLD POPULATION: 7 BILLION



Most companies in the U.S. have at least

**100 TERABYTES**

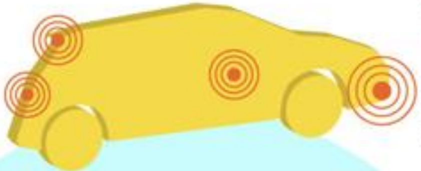
[ 100,000 GIGABYTES ]

of data stored

The New York Stock Exchange captures

**1 TB OF TRADE INFORMATION**

during each trading session



Modern cars have close to

**100 SENSORS**

that monitor items such as fuel level and tire pressure

# Velocity

## ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

**18.9 BILLION NETWORK CONNECTIONS**

– almost 2.5 connections per person on earth

From traffic patterns and mu
history and medical record
stored, and analyzed to ena
and services that the world
But what exactly is big data
massive amounts of data be

As a leader in the sector,
break big data into four d
**Velocity, Variety and Veracity**

Depending on the industry a
data encompasses informa
internal and external sources
social media, enterprise co
mobile devices. Companies
adapt their products and ser
customer needs, optimiz
infrastructure, and find new

By 2015

**4.4 MILLION IT JOBS**

will be created globally to s
with 1.9 million in the Unit

**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

# UR V's
# ig
# a

and music downloads to web
records, data is recorded,
d to enable the technology
e world relies on every day.
ig data, and how can these
data be used?

sector, IBM data scientists
four dimensions: **Volume,**
**Veracity**

dustry and organization, big
information from multiple
sources such as transactions,
rise content, sensors and

## Variety

**DIFFERENT FORMS OF DATA**

As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**

[ 161 BILLION GIGABYTES ]

By 2014, it's anticipated there will be

**420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO**

are watched on YouTube each month

**30 BILLION PIECES OF CONTENT**

are shared on Facebook every month

**400 MILLION TWEETS**

are sent per day by about 200 million monthly active users

**1 IN 3 BUSINESS LEADERS**

don't trust the information

Poor data quality costs the US economy around

**$3.1 TRILLION A YEAR**

d analyzed to enable the technology es that the world relies on every day. xactly is big data, and how can these nounts of data be used?

r in the sector, IBM data scientists data into four dimensions: **Volume, riety and Veracity**

on the industry and organization, big mpasses information from multiple d external sources such as transactions, ia, enterprise content, sensors and ices. Companies can leverage data to products and services to better meet needs, optimize operations and ire, and find new sources of revenue.

## LION IT JOBS

ated globally to support big data, illion in the United States

are shared on Facebook every month

## 400 MILLION TWEETS

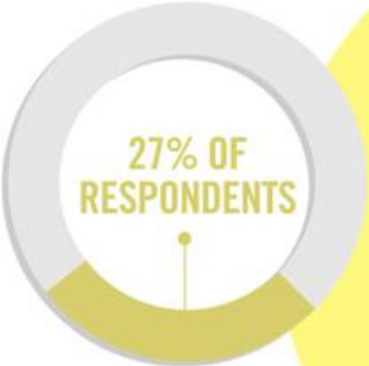
are sent per day by about 200 million monthly active users

## 1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

## $3.1 TRILLION A YEAR



in one survey were unsure of how much of their data was inaccurate

# Veracity

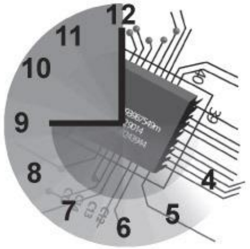
## UNCERTAINTY OF DATA

# Why Big Data is Popular Now

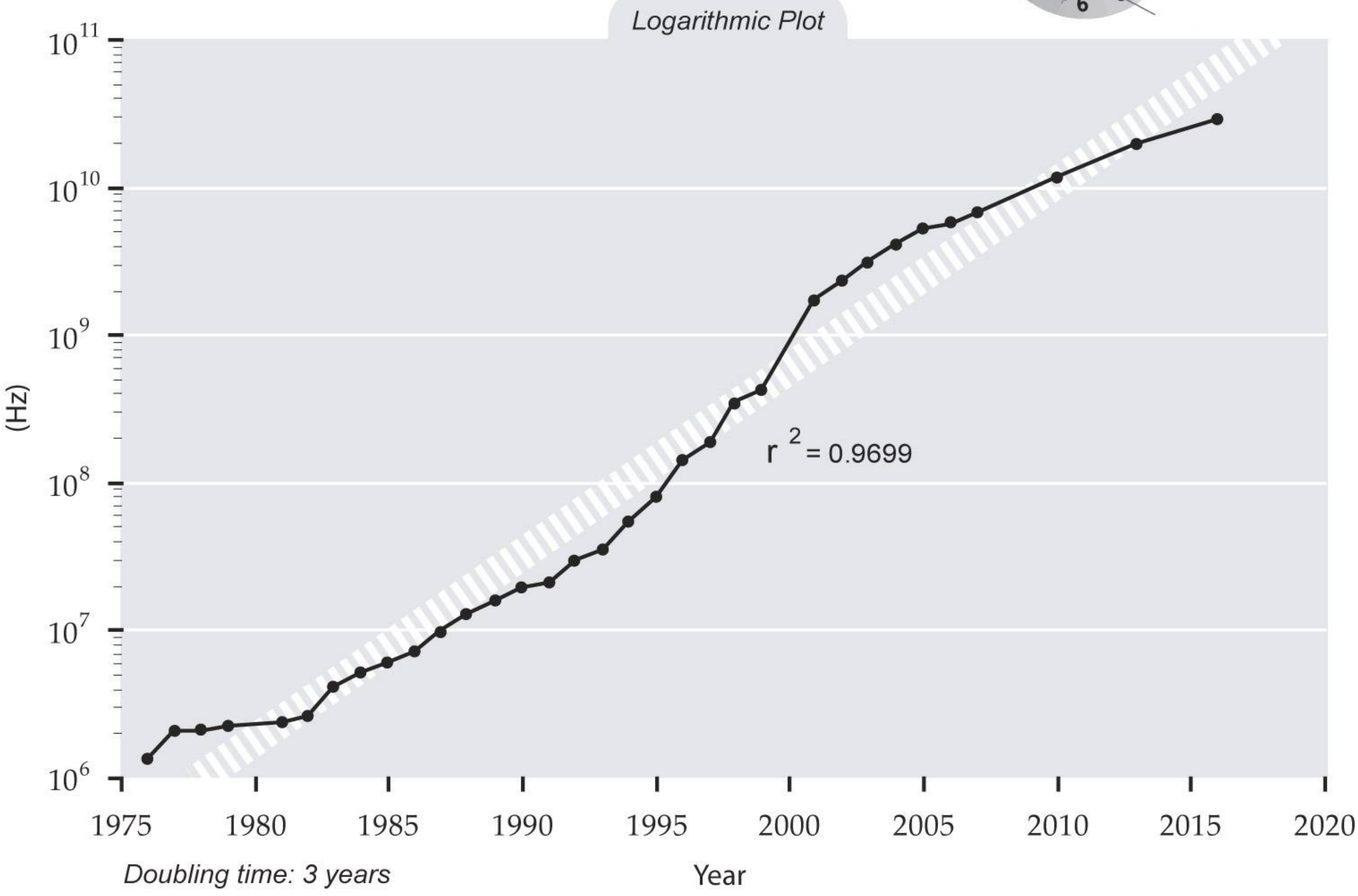
1. Technological progress
2. Development of infrastructure
3. Accessibility of data

# Technological Progress in Big Data

- Computing power
- Price drop of the hardware
- Appearance of cloud storage and reduction of prices for storage devices

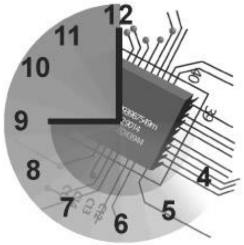


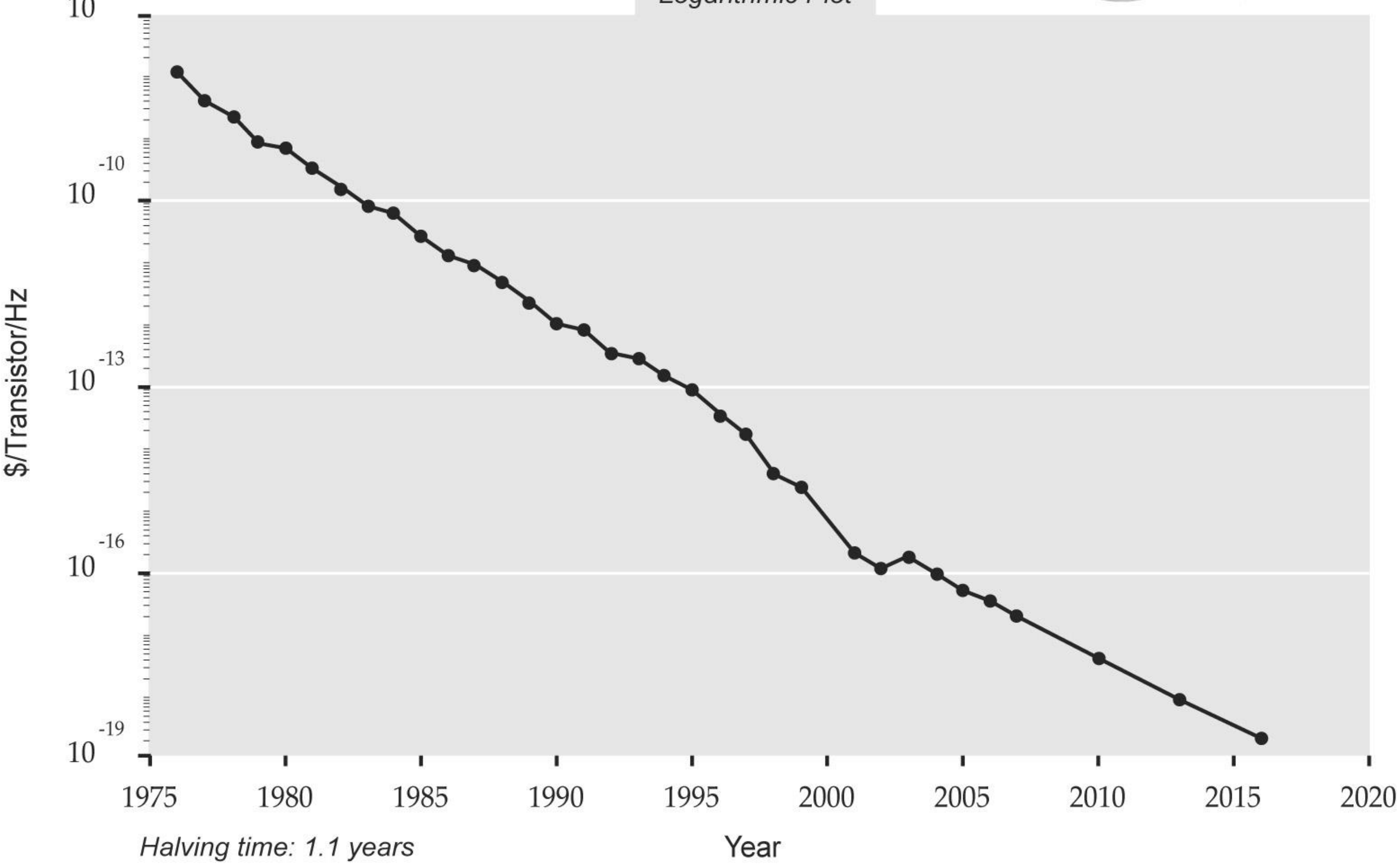
# Microprocessor Clock Speed



http://tvtropes.org/pmwiki/pmwiki.php/UsefulNotes/ClockSpeed

# Microprocessor Cost Per Transistor Cycle





http://www.singularity.com/charts/page62.html

https://onedrive.live.com/about/zh-tw/plans/
https://www.dropbox.com/pro/buy

https://www.google.com/intl/zh-tw/drive/pricing/
https://www.box.com/pricing/individual

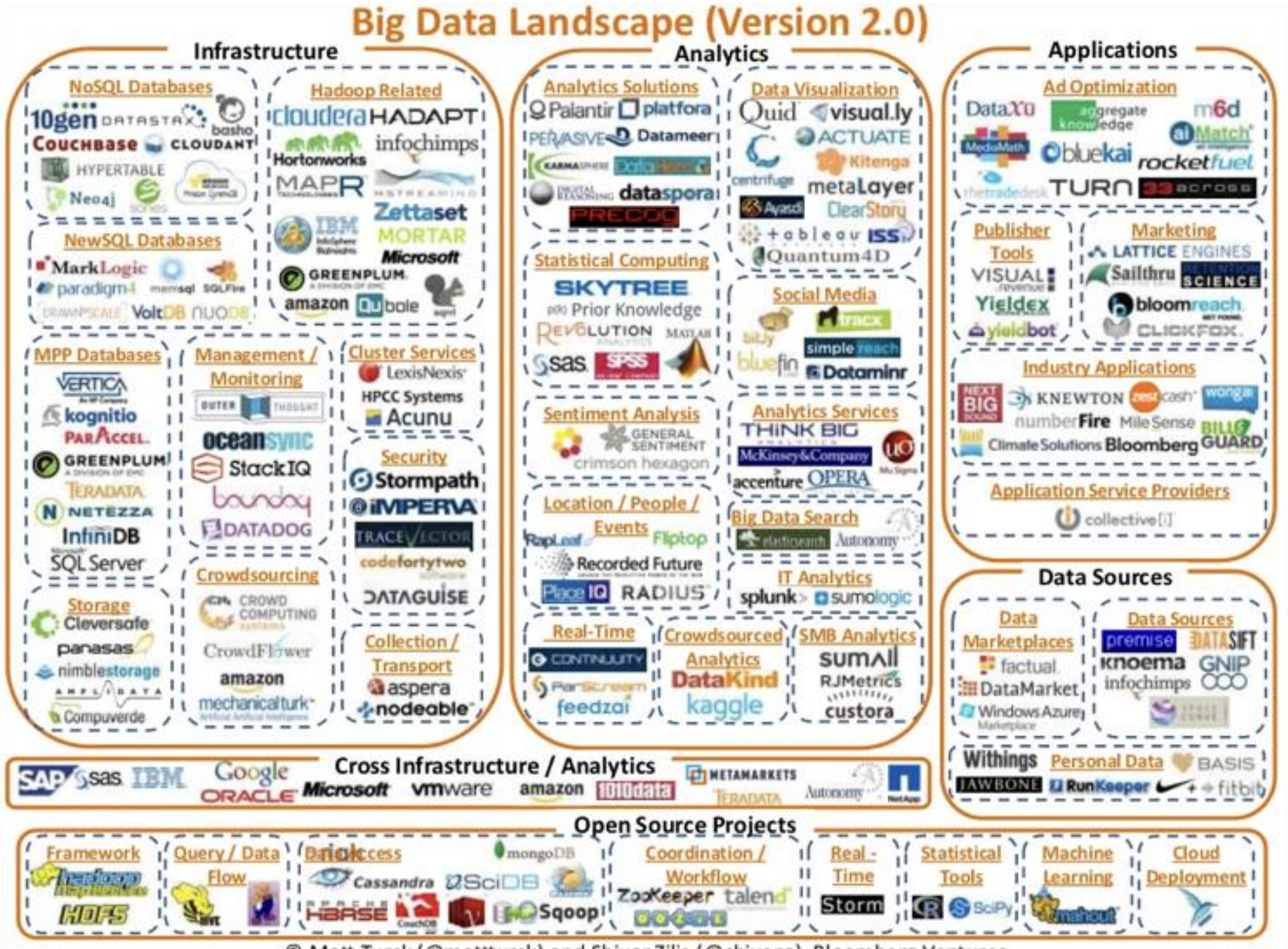
| | | | |
|---|---|---|---|
| 5 GB free | 2 GB free | 15 GB free | 10 GB free |
| 50 GB NT$60/m | 1 TB ~NT$330/m | 100 GB NT$65/m | 100 GB ~NT$35/m |
| 1 TB NT$2190/y<br>(+Office 365) | | 1 TB NT$330/m | 1 TB NT$330/m |

2017/02/11

# Cloud Storage

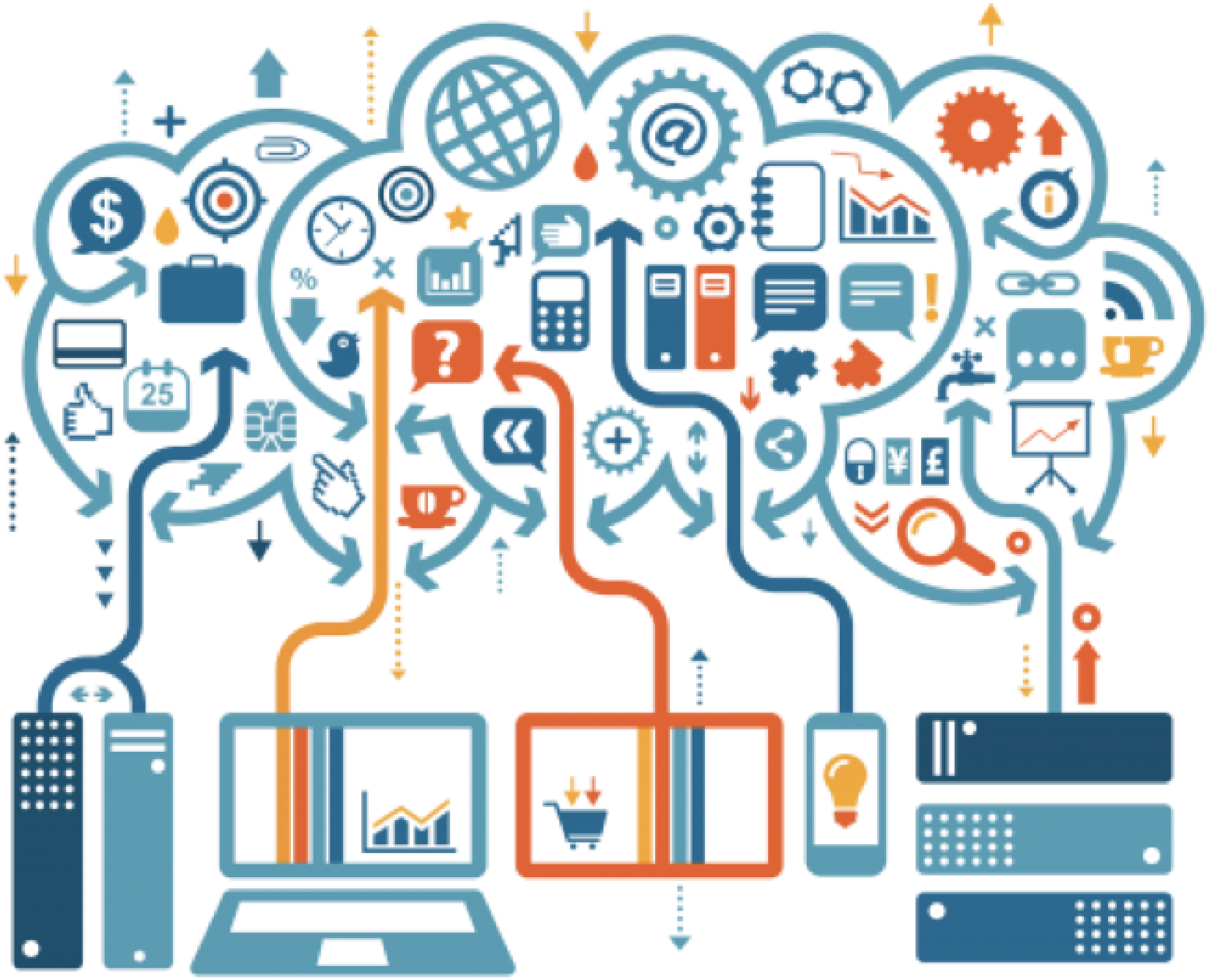
- AWS (Amazon Web Services) - Simple Storage Service (S3)
- Microsoft Azure - Blob Storage
- Google Cloud Platform - Cloud Storage
- ….

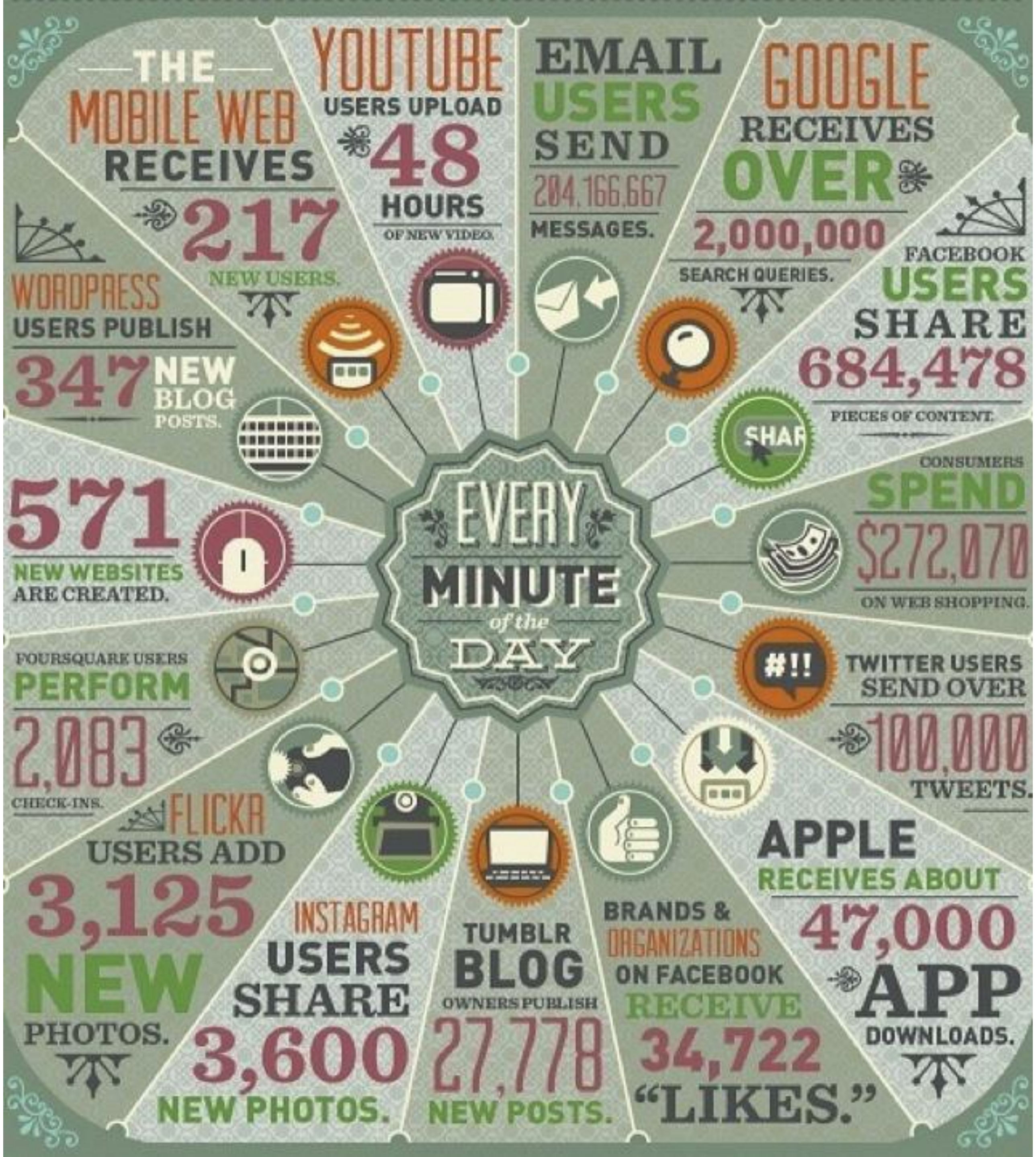
# Development of Infrastructure in Big Data



© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

# Accessibility of Data





# EVERY MINUTE of the DAY

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.

CONSUMERS SPEND $272,070 ON WEB SHOPPING.

TWITTER USERS SEND OVER 100,000 TWEETS.

APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.

BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."

TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.

INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.

FLICKR USERS ADD 3,125 NEW PHOTOS.

FOURSQUARE USERS PERFORM 2,083 CHECK-INS.

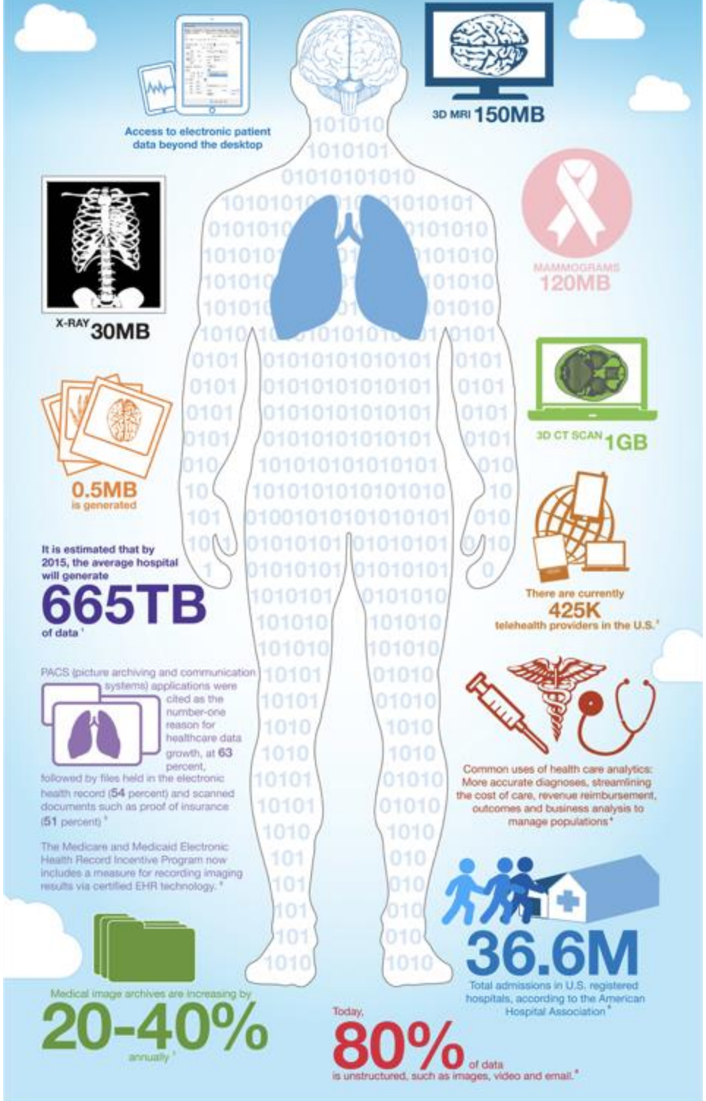
571 NEW WEBSITES ARE CREATED.

WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.

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

X-RAY 30MB

MAMMOGRAMS 120MB

3D CT SCAN 1GB

0.5MB is generated

It is estimated that by 2015, the average hospital will generate **665TB** of data [1]

There are currently **425K** telehealth providers in the U.S. [2]

PACS (picture archiving and communication systems) applications were cited as the number-one reason for healthcare data growth, at **63** percent, followed by files held in the electronic health record (**54** percent) and scanned documents such as proof of insurance (**51** percent) [3]

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

The Medicare and Medicaid Electronic Health Record Incentive Program now includes a measure for recording imaging results via certified EHR technology. [5]

**36.6M** Total admissions in U.S. registered hospitals, according to the American Hospital Association [6]

Medical image archives are increasing by **20-40%** annually [7]

Today, **80%** of data is unstructured, such as images, video and email. [8]

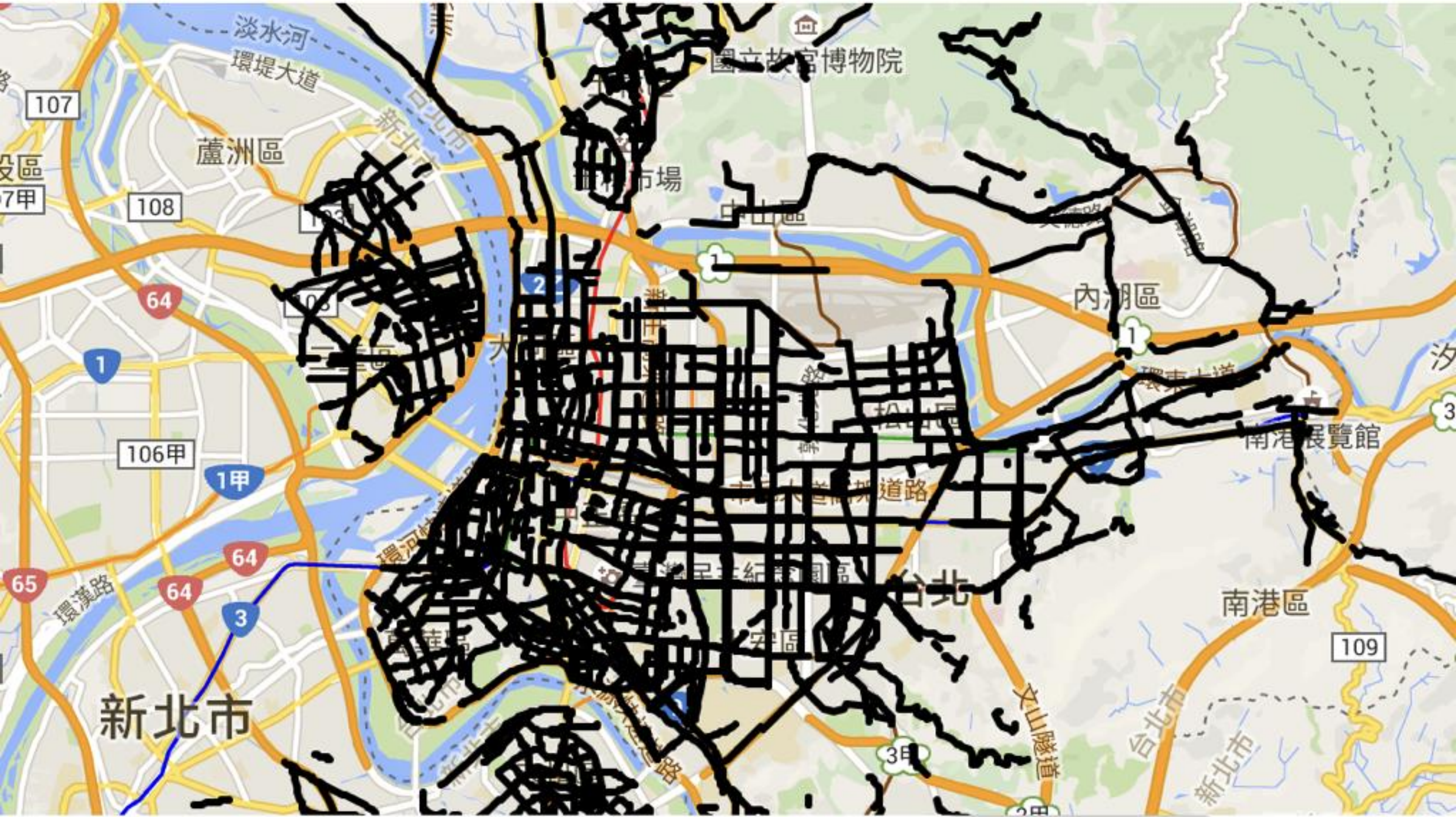
Sources:

[illegible]

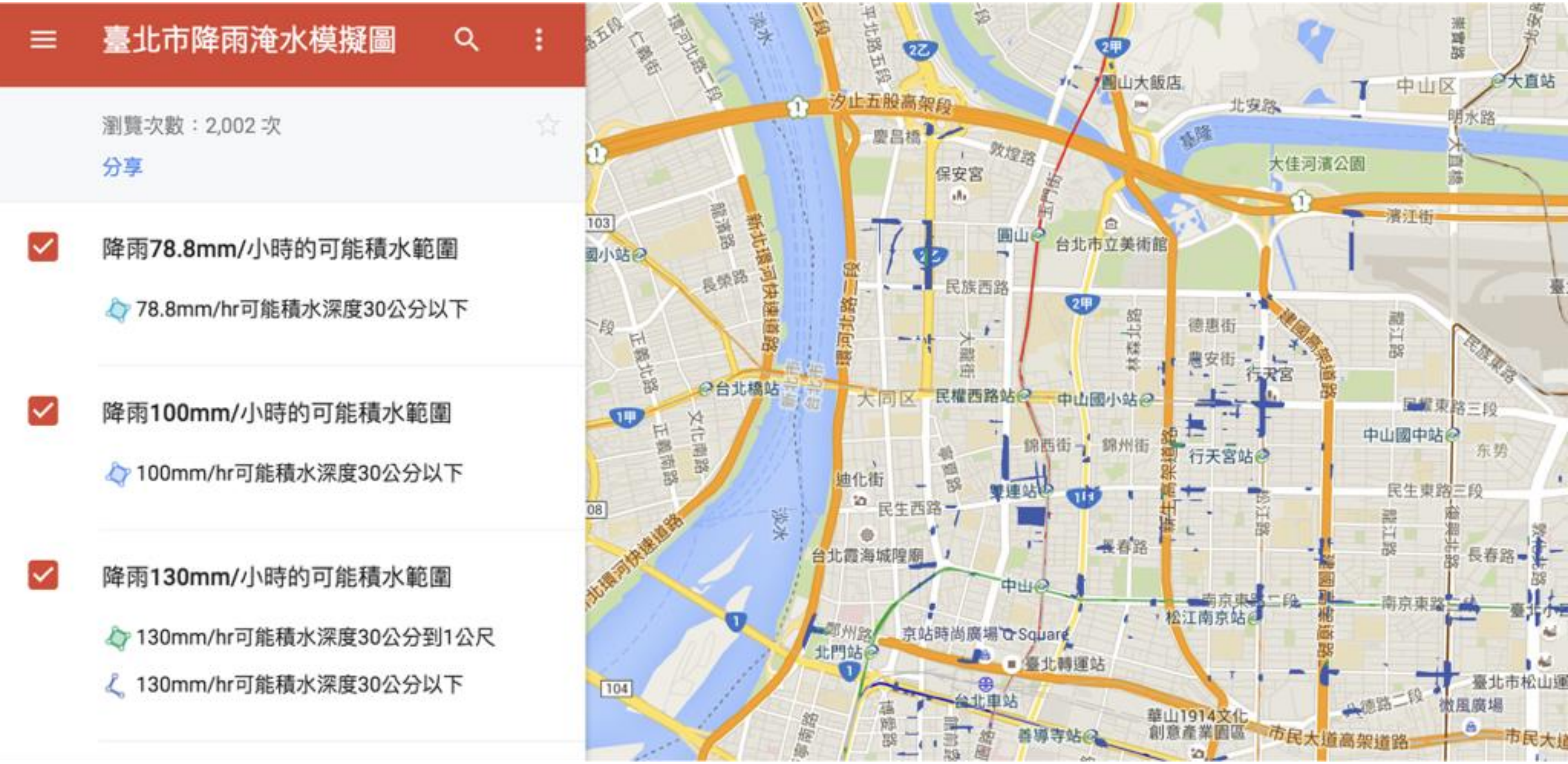
NetApp

# 台北市鉛管地圖

http://kiang.github.io/lead_pipes/



# 台北市淹水地圖



臺北市降雨淹水模擬圖

瀏覽次數：2,002 次

分享

☑ 降雨78.8mm/小時的可能積水範圍

78.8mm/hr可能積水深度30公分以下

☑ 降雨100mm/小時的可能積水範圍

100mm/hr可能積水深度30公分以下

☑ 降雨130mm/小時的可能積水範圍

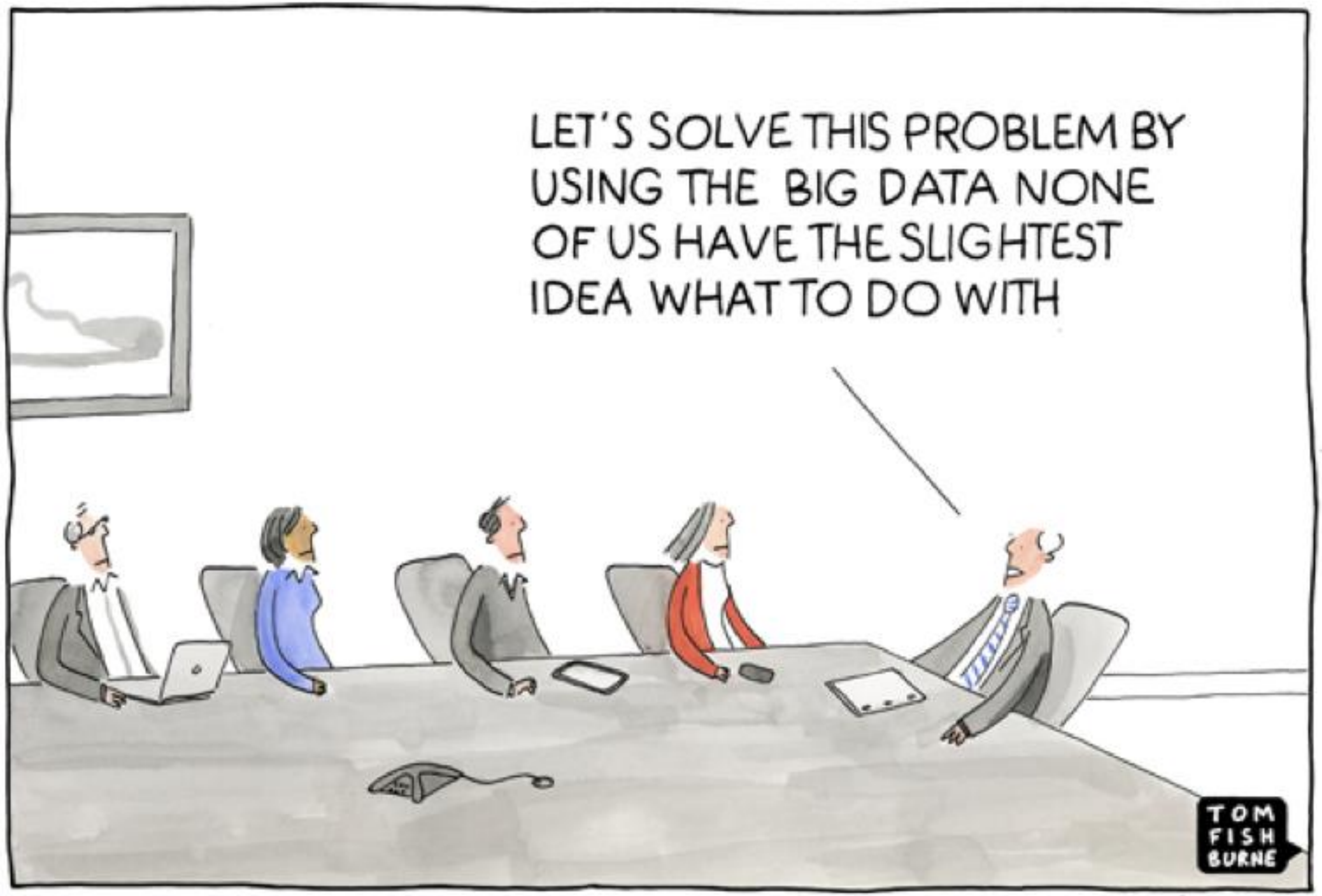
130mm/hr可能積水深度30公分到1公尺

130mm/hr可能積水深度30公分以下

https://www.google.com/maps/d/viewer?mid=zCRWCdi-t4dk.kwfkt9RpU_8o

# Questions?

# What is Big Data Analytics?


LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE

© marketoonist.com

# Let’s Start From ‘Small’ Data Analytics….

- What is data analytics?
- Data analytics (DA) is the process of **examining data sets** in order to **draw conclusions about the information** they contain, increasingly with the aid of specialized systems and software.

# What is Big Data Analytics?
# From IBM

- Big data analytics is the use of **advanced analytic techniques** against (*very large, diverse data sets that include different types such as structured/unstructured and streaming/batch, and different sizes from terabytes to zettabytes*).

- Big data analytics is the use of **advanced analytic techniques** against ***big data***.

http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html

# Advanced Analytic Techniques?

- Text analytics
- Machine learning
- Predictive analytics
- Data mining
- Statistics
- Natural language processing
- …etc

- To analyze such a large volume of data, big data analytics is typically performed using **specialized software tools and applications**

http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html

# Before Applying Advanced Analytic Techniques

- Big data analytics is the process of **collecting**, **organizing** and **analyzing** large sets of data to **discover patterns and other useful information**.
  - Vangie Beal

- Before **analyzing** large sets:
  - Collect the data
  - Organize the data

# Pipelines for Data Analysis

1. Asking a Question
2. Data Collection
3. Data Import
4. Data Pre-processing (cleaning)
5. Exploratory Data Analysis
6. Data Visualization
7. Data Modeling
8. Data Communication (Report)

Repeat n times

# Pipelines for Data Analysis in R (& Hadoop)

1. ~~Asking a Question~~
2. ~~Data Collection~~
3. Data Import
4. Data Pre-processing (cleaning)
5. Exploratory Data Analysis
6. Data Visualization
7. Data Modeling
8. Data Communication (Report)

Repeat n times

# Questions?

# Why We Need Big Data Analytics?

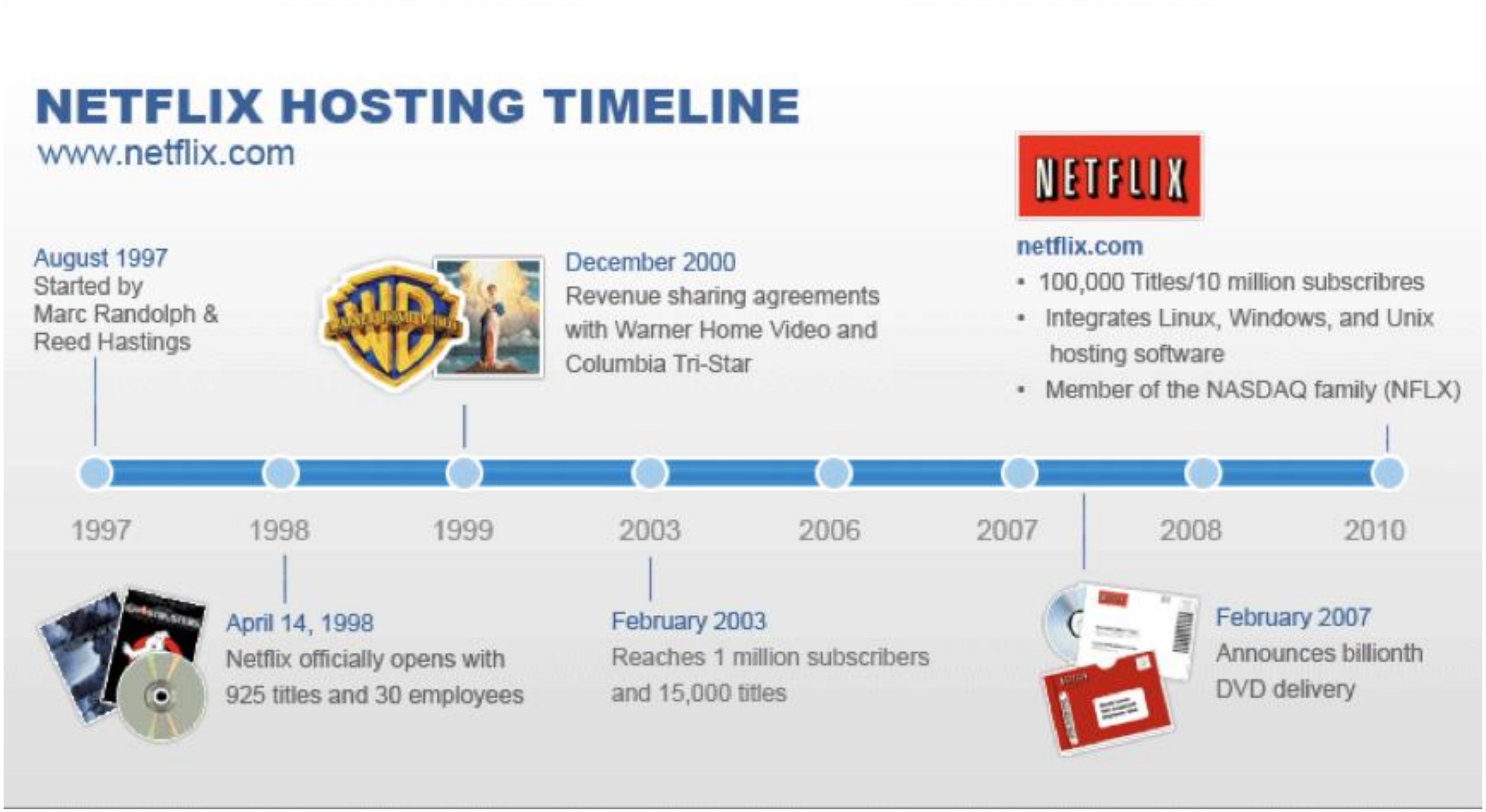
# Why We Need Big Data Analytics? From IBM

- Analyzing big data allows analysts, researchers, and business users to **make better and faster decisions** using **data that was previously inaccessible or unusable.**

http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html

# Netflix

Netflix is said to account for one third of peak-time internet traffic in the US



## NETFLIX HOSTING TIMELINE

www.netflix.com

August 1997
Started by
Marc Randolph &
Reed Hastings

December 2000
Revenue sharing agreements
with Warner Home Video and
Columbia Tri-Star

NETFLIX

netflix.com

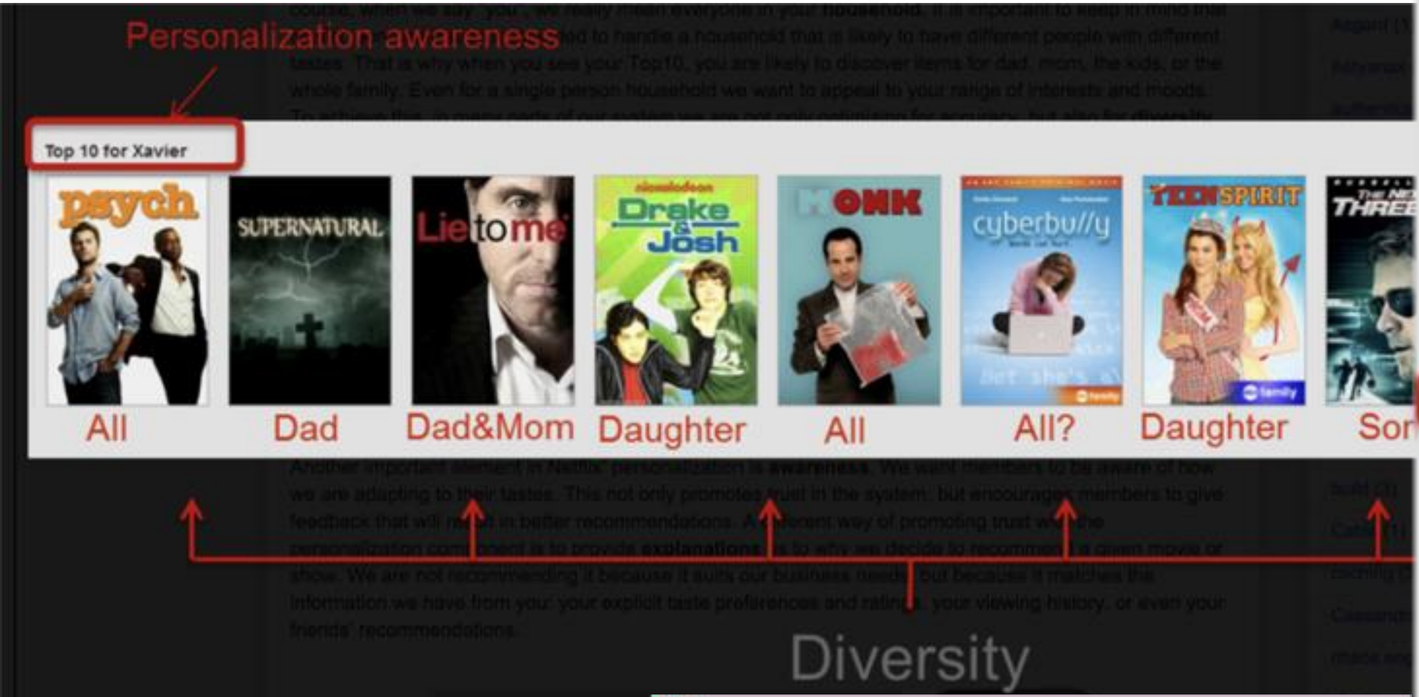
- 100,000 Titles/10 million subscribres
- Integrates Linux, Windows, and Unix hosting software
- Member of the NASDAQ family (NFLX)

April 14, 1998
Netflix officially opens with
925 titles and 30 employees

February 2003
Reaches 1 million subscribers
and 15,000 titles

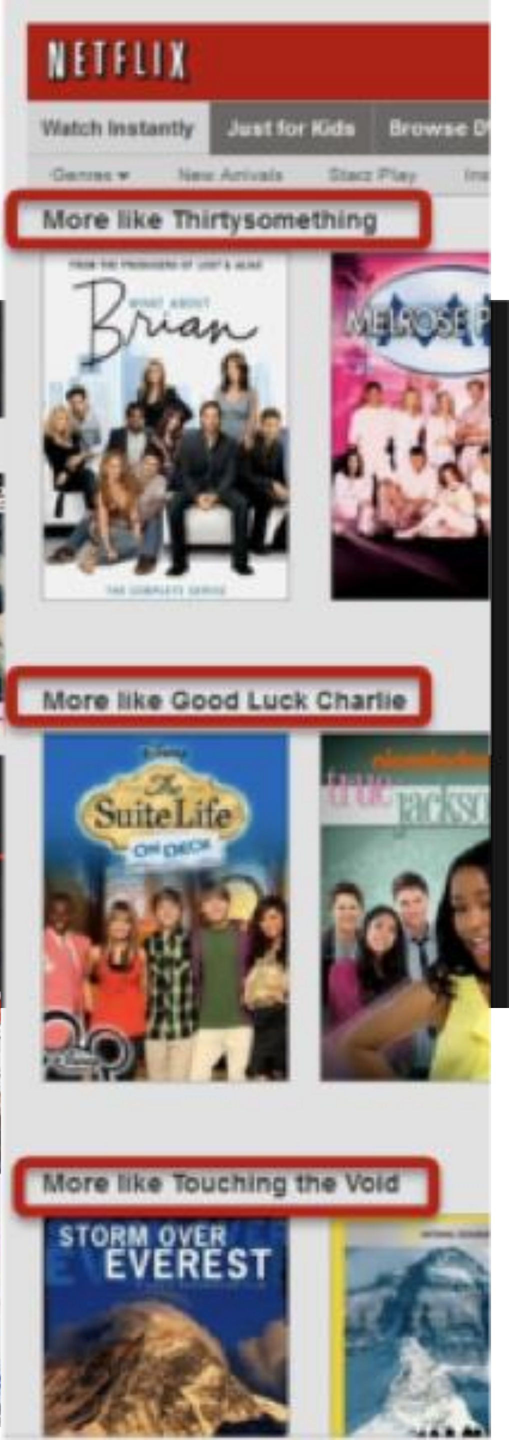
February 2007
Announces billionth
DVD delivery

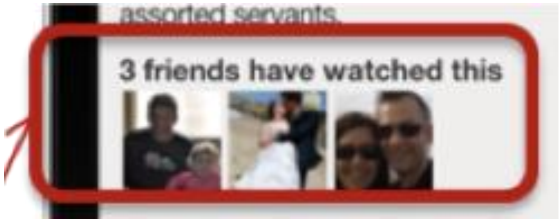
# Netflix Recommendations

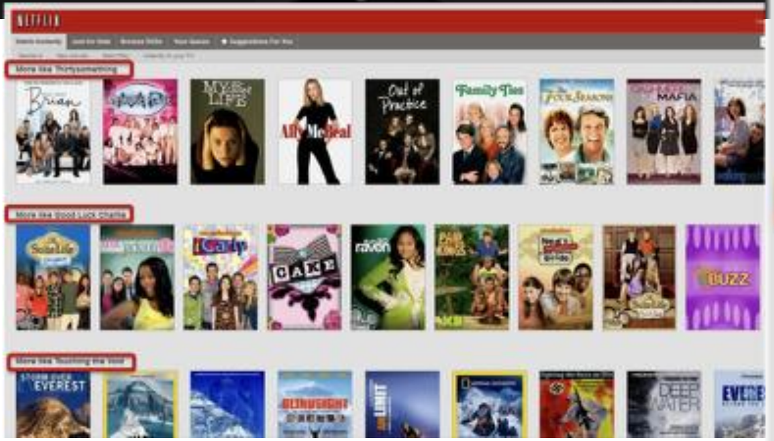


course, when we say "you", we really mean everyone in your household. It is important to keep in mind that [illegible] need to handle a household that is likely to have different people with different tastes. That is why when you see your Top10, you are likely to discover items for dad, mom, the kids, or the whole family. Even for a single person household we want to appeal to your range of interests and moods. To achieve this, in many parts of our system we are not only optimizing for accuracy, but also for diversity.

Another important element in Netflix' personalization is awareness. We want members to be aware of how we are adapting to their tastes. This not only promotes trust in the system, but encourages members to give feedback that will result in better recommendations. A different way of promoting trust with the personalization component is to provide explanations as to why we decide to recommend a given movie or show. We are not recommending it because it suits our business needs, but because it matches the information we have from you: your explicit taste preferences and ratings, your viewing history, or even your friends' recommendations.

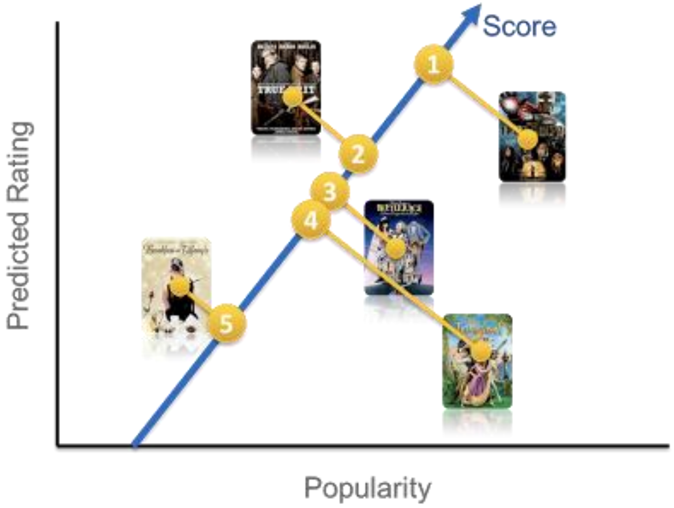

NETFLIX
Watch Instantly
Just for Kids
Browse D
Genres
New Arrivals
Starz Play
More like Thirtysomething
More like Good Luck Charlie
More like Touching the Void

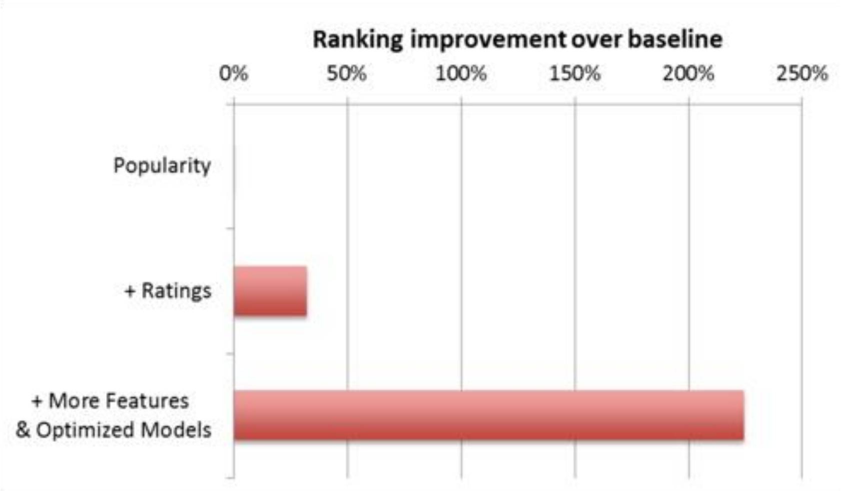


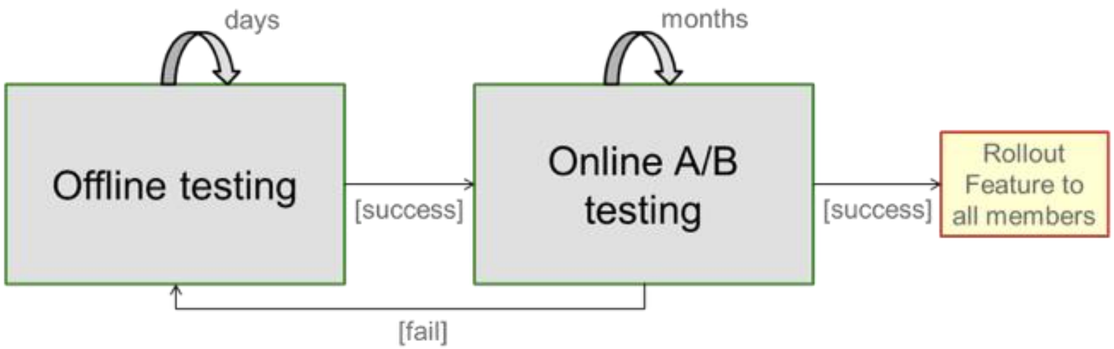

NETFLIX
More like Thirtysomething
More like Good Luck Charlie
More like Touching the Void

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

# Netflix Recommendations







days

Offline testing

[success]

months

Online A/B testing

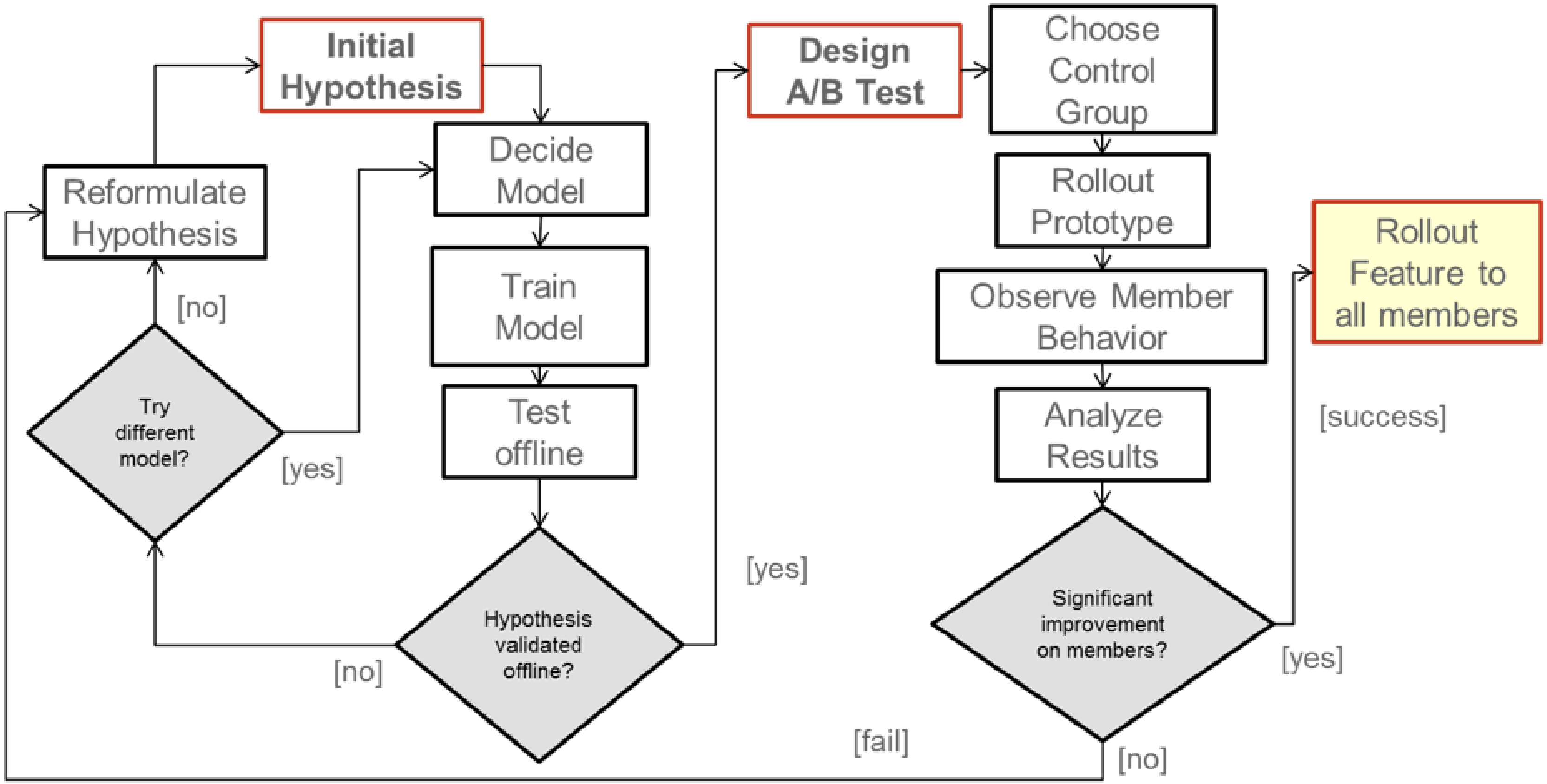
[success]

Rollout Feature to all members

[fail]

http://techblog.netflix.com/2012/06/netflix-recommendations-beyond-5-stars.html

# Netflix Recommendations

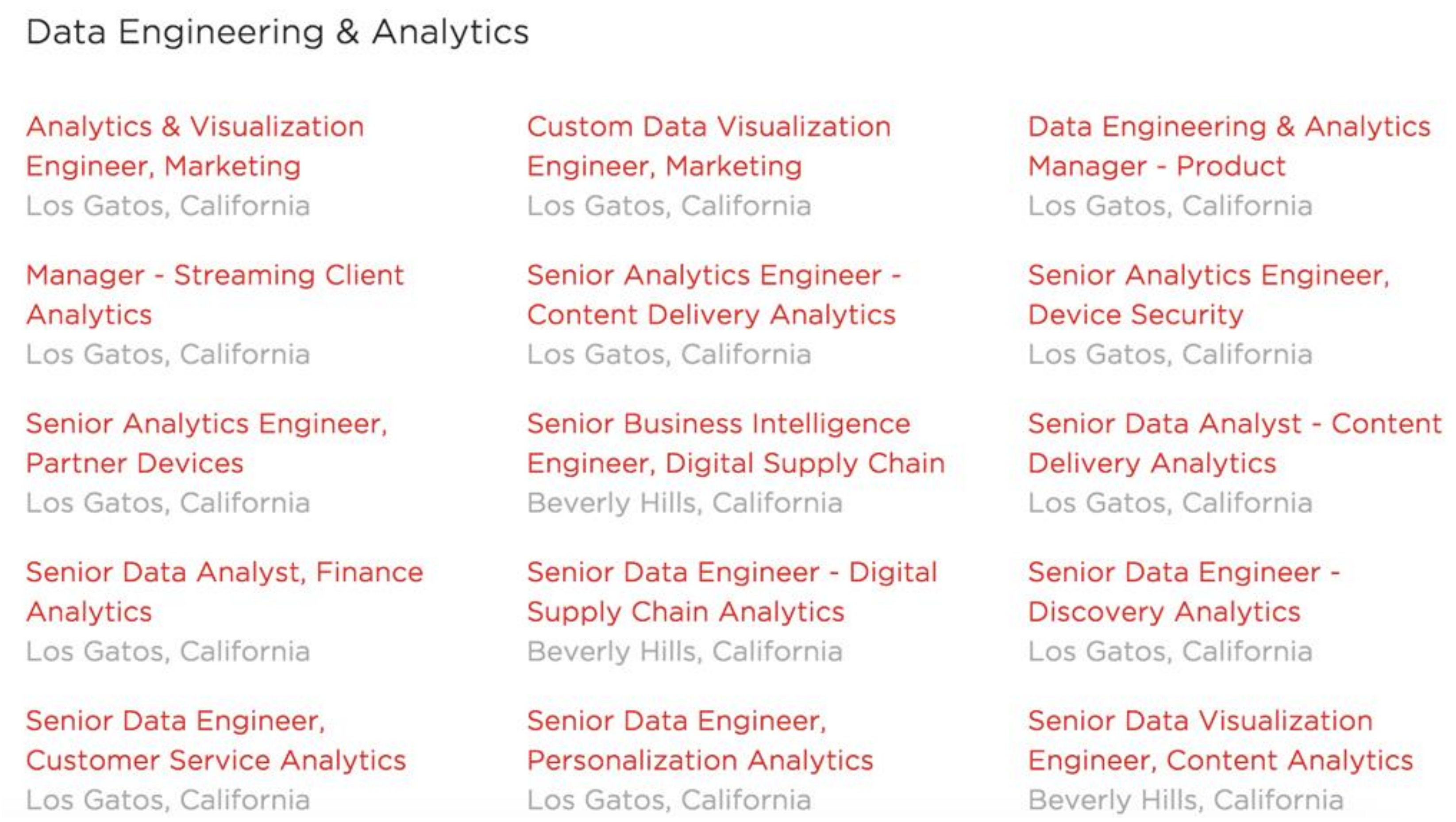


http://techblog.netflix.com/2012/06/netflix-recommendations-beyond-5-stars.html

# Data used in Netflix Recommendations

- When you pause, rewind, or fast forward
- What day you watch content (Netflix has found people watch **TV shows during the weekday** and **movies during the weekend**.)
- The date you watch
- What time you watch content
- Where you watch (zip code)
- What device you use to watch (Do you like to use your tablet for TV shows and your Roku for movies? Do people access the Just for Kids feature more on their iPads, etc.?)
- When you pause and leave content (and if you ever come back)
- The ratings given (about 4 million per day)
- Searches (about 3 million per day)
- Browsing and scrolling behavior
- Netflix also looks at data within movies.

# https://jobs.netflix.com/jobs



## Data Engineering & Analytics

Analytics & Visualization Engineer, Marketing
Los Gatos, California

Custom Data Visualization Engineer, Marketing
Los Gatos, California

Data Engineering & Analytics Manager - Product
Los Gatos, California

Manager - Streaming Client Analytics
Los Gatos, California

Senior Analytics Engineer - Content Delivery Analytics
Los Gatos, California

Senior Analytics Engineer, Device Security
Los Gatos, California

Senior Analytics Engineer, Partner Devices
Los Gatos, California

Senior Business Intelligence Engineer, Digital Supply Chain
Beverly Hills, California

Senior Data Analyst - Content Delivery Analytics
Los Gatos, California

Senior Data Analyst, Finance Analytics
Los Gatos, California

Senior Data Engineer - Digital Supply Chain Analytics
Beverly Hills, California

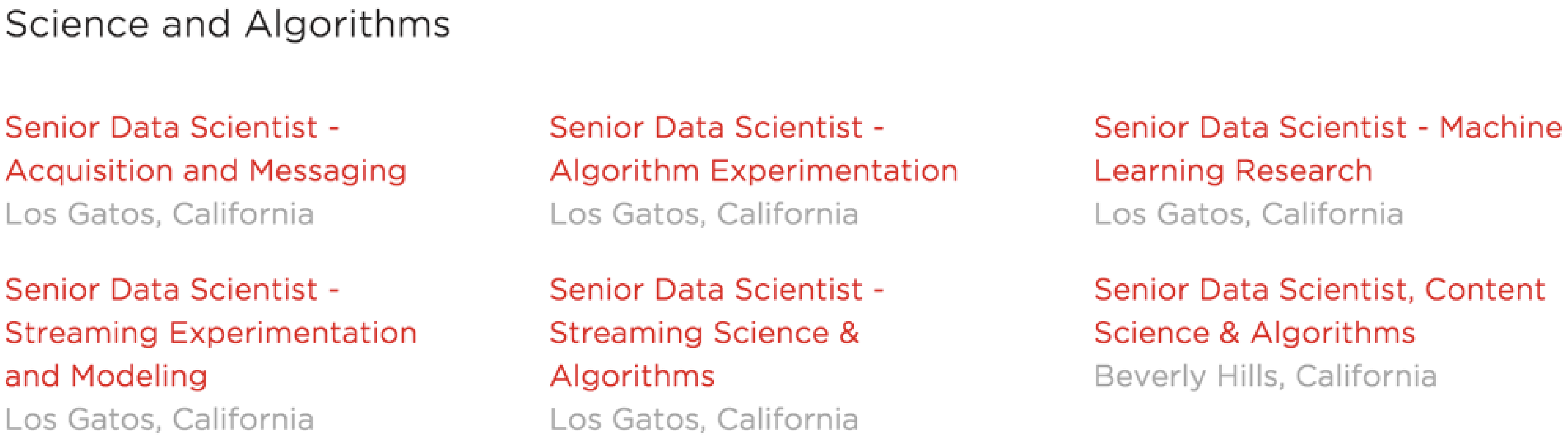
Senior Data Engineer - Discovery Analytics
Los Gatos, California

Senior Data Engineer, Customer Service Analytics
Los Gatos, California

Senior Data Engineer, Personalization Analytics
Los Gatos, California

Senior Data Visualization Engineer, Content Analytics
Beverly Hills, California

# https://jobs.netflix.com/jobs



Science and Algorithms

Senior Data Scientist - Acquisition and Messaging
Los Gatos, California

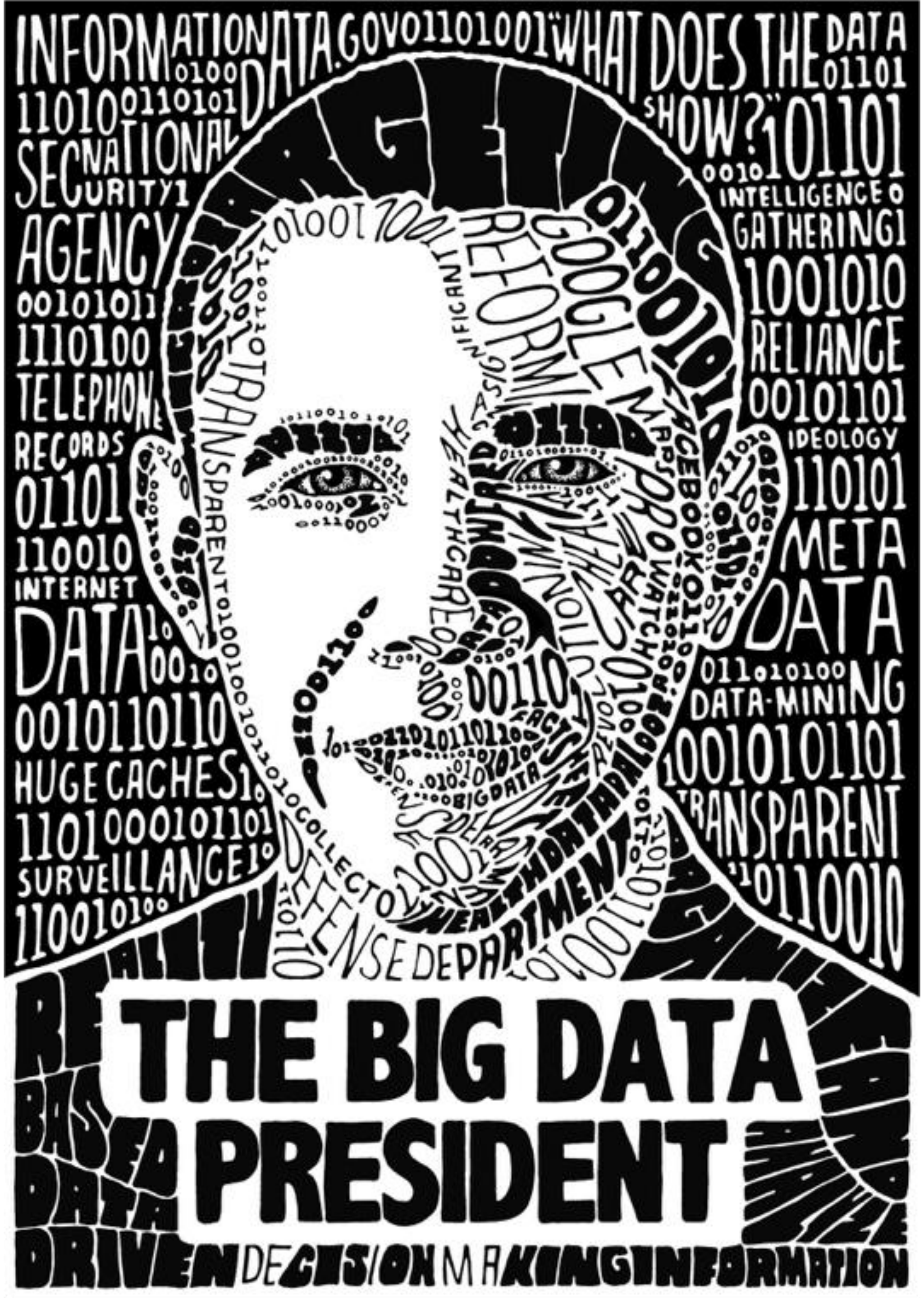
Senior Data Scientist - Algorithm Experimentation
Los Gatos, California

Senior Data Scientist - Machine Learning Research
Los Gatos, California

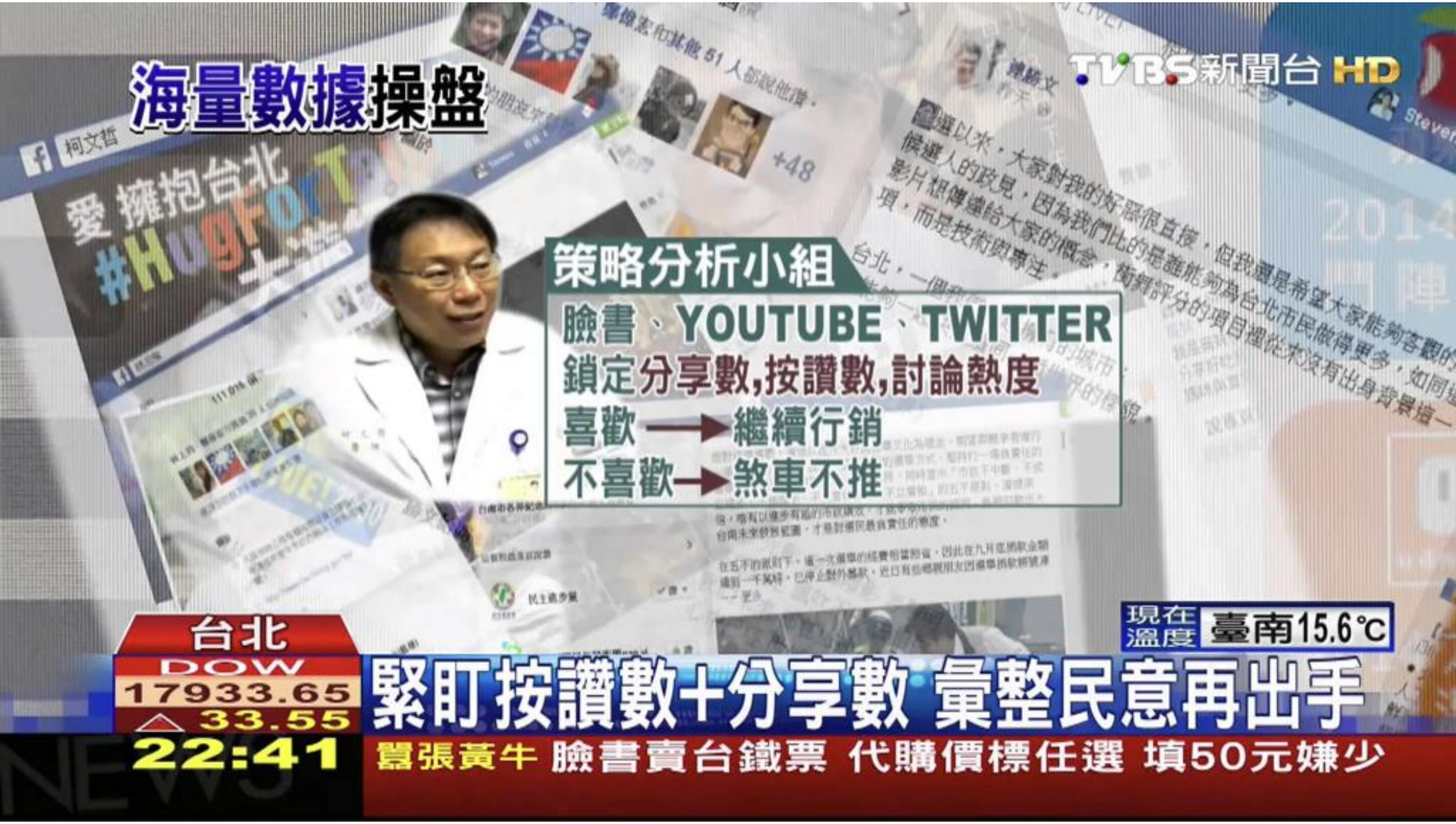
Senior Data Scientist - Streaming Experimentation and Modeling
Los Gatos, California

Senior Data Scientist - Streaming Science & Algorithms
Los Gatos, California

Senior Data Scientist, Content Science & Algorithms
Beverly Hills, California



Sarah A. King for The Washington Post



海量數據操盤

TVBS新聞台 HD

策略分析小組

臉書、YOUTUBE、TWITTER

鎖定分享數,按讚數,討論熱度

喜歡→繼續行銷

不喜歡→煞車不推

現在溫度 臺南15.6℃

台北

DOW 17933.65 △33.55

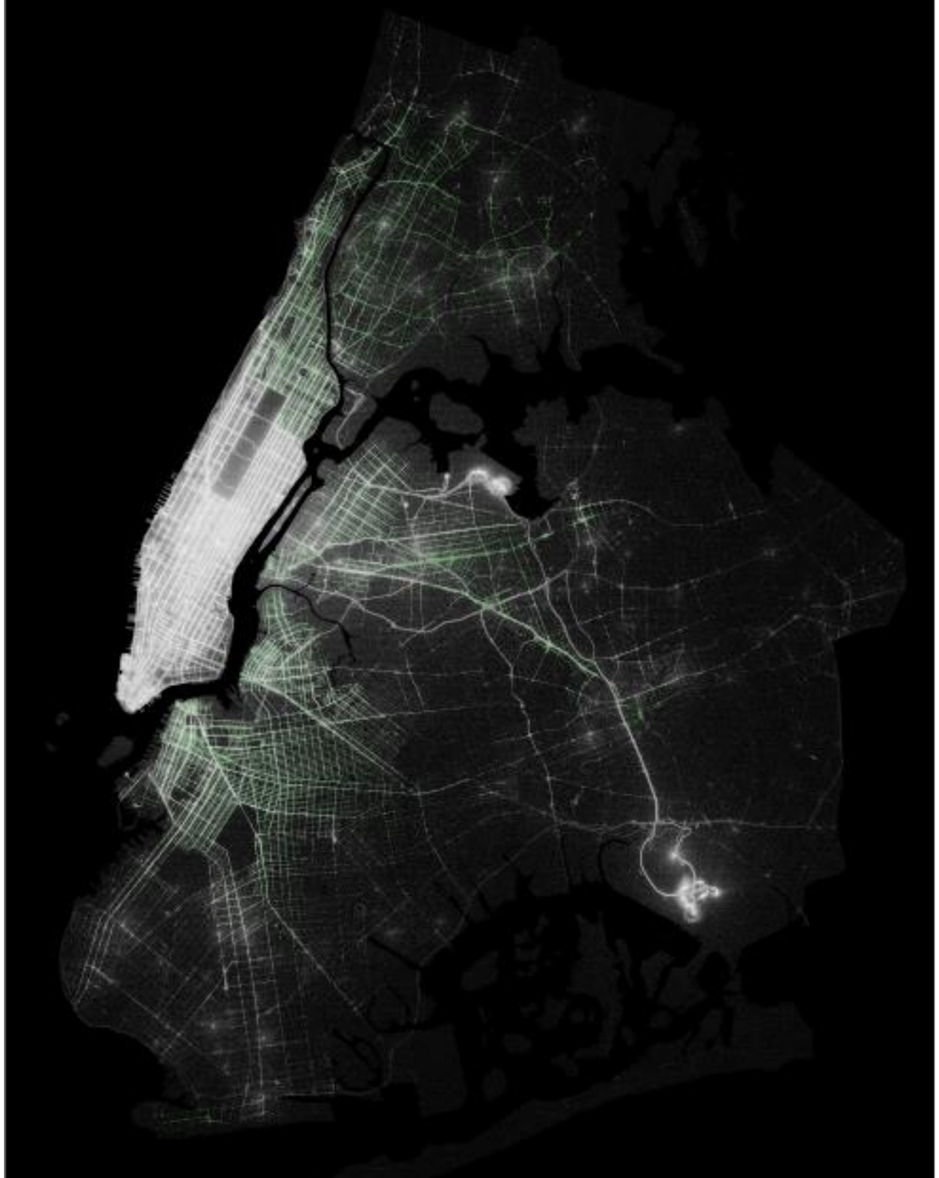
緊盯按讚數+分享數 彙整民意再出手

22:41 囂張黃牛 臉書賣台鐵票 代購價標任選 填50元嫌少

MG149, Keywords, Young people…etc.

## New York City Taxi Pickups

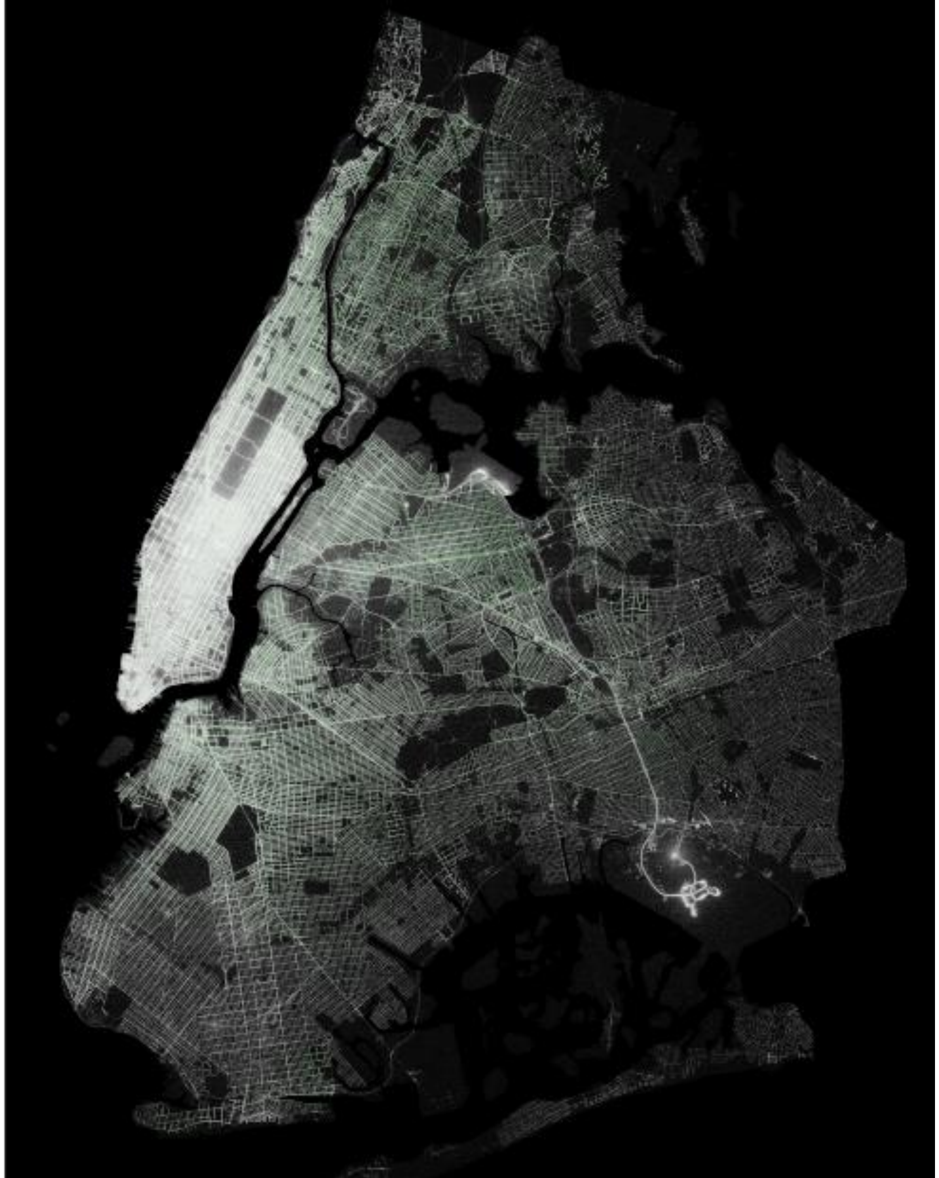
2009–2015



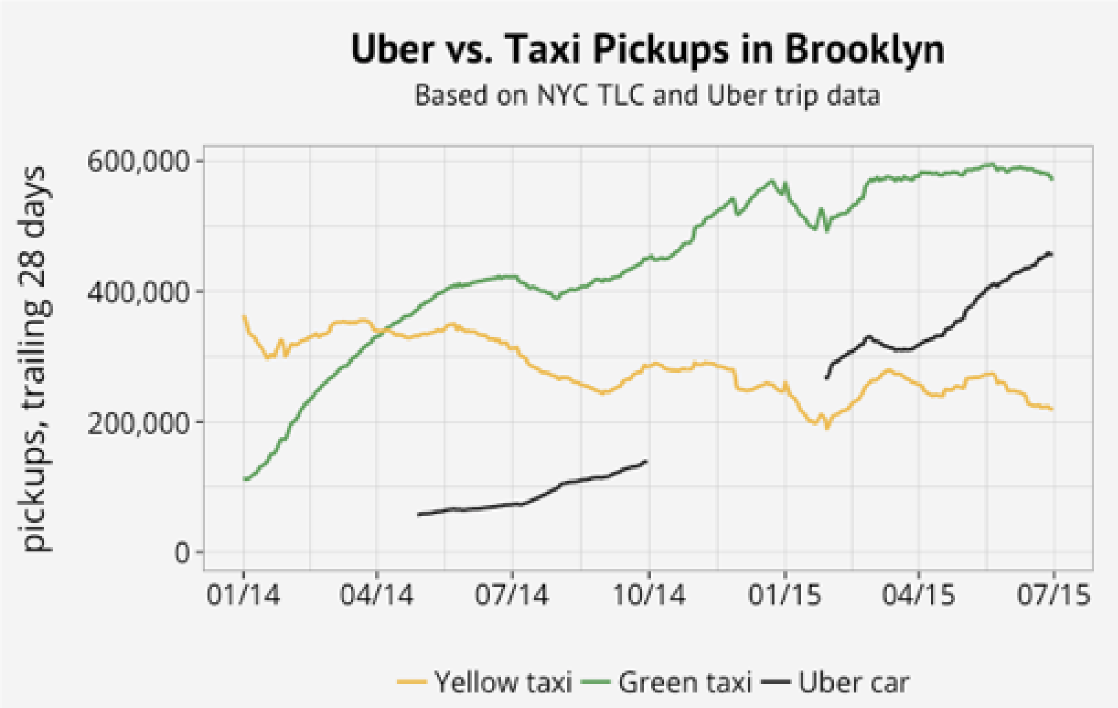
toddwschneider.com

## New York City Taxi Drop Offs

2009–2015

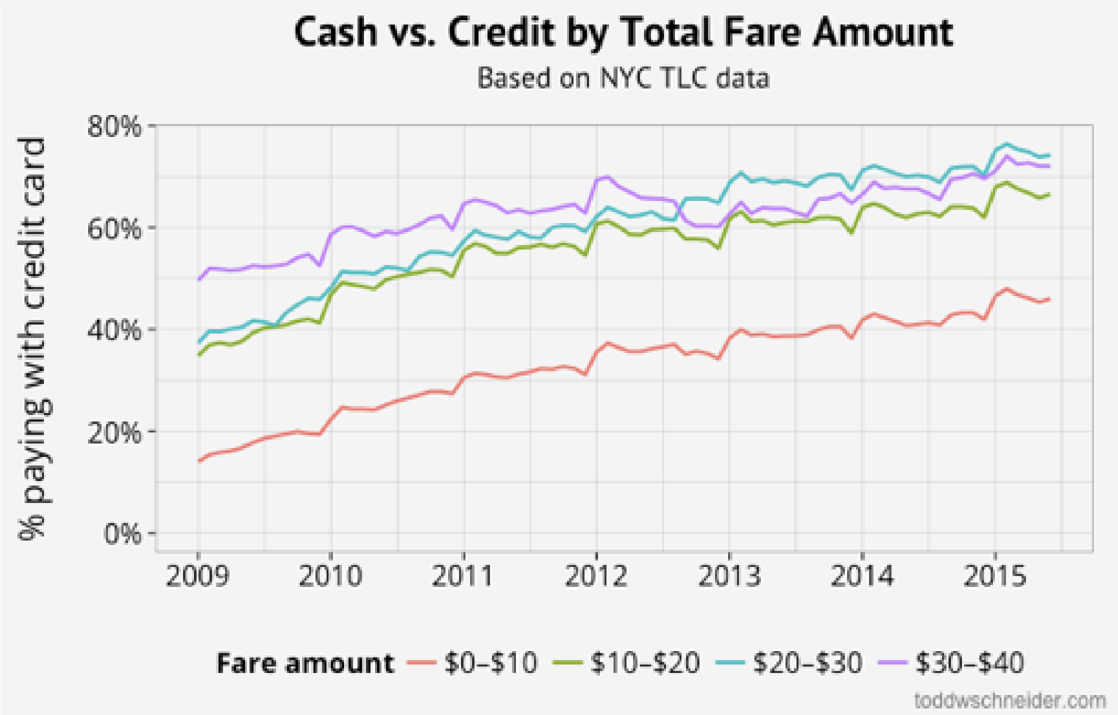


toddwschneider.com



Uber vs. Taxi Pickups in Brooklyn

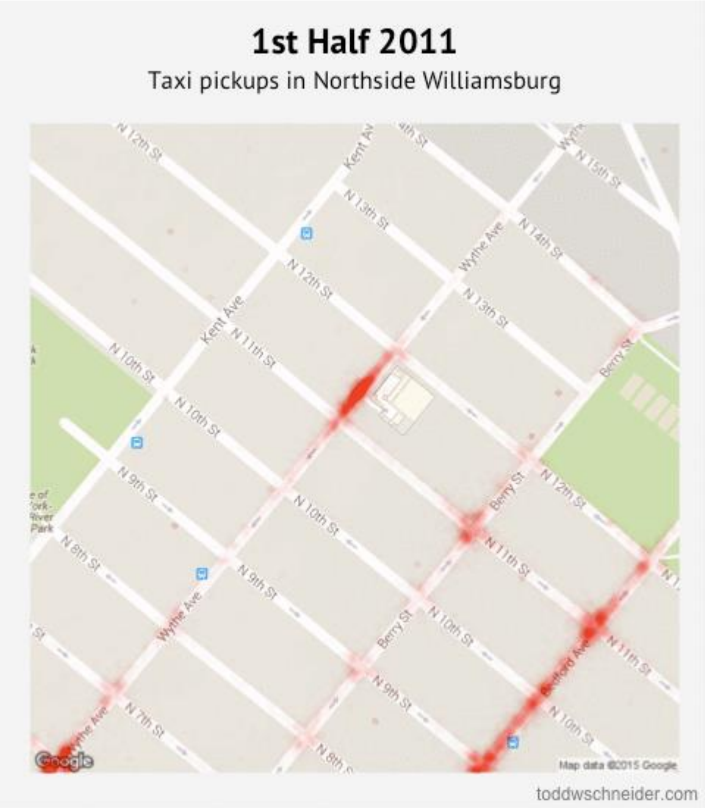
Based on NYC TLC and Uber trip data



Cash vs. Credit by Total Fare Amount

Based on NYC TLC data

toddwschneider.com



1st Half 2011

Taxi pickups in Northside Williamsburg

toddwschneider.com

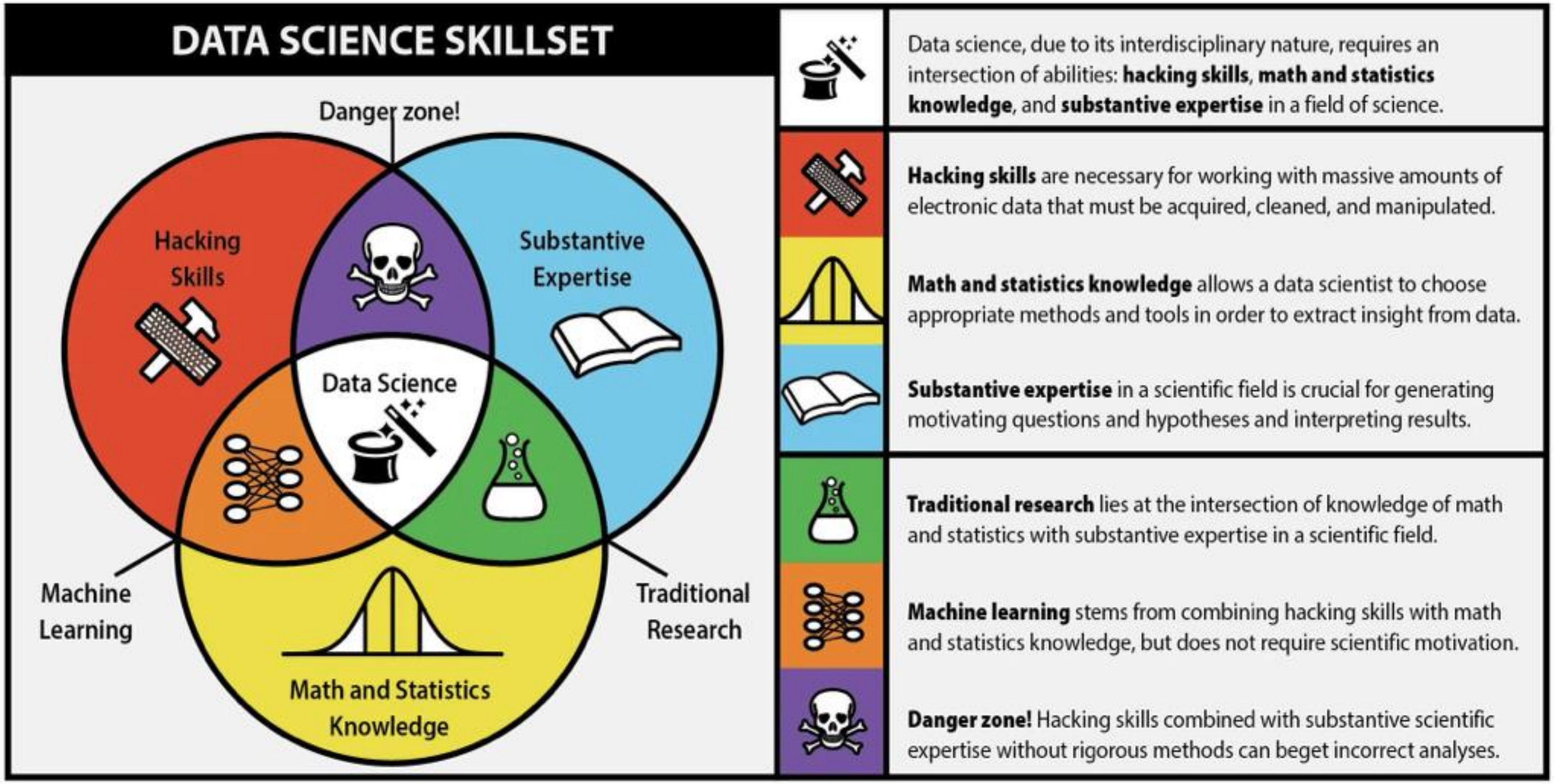
## The Boston Celtics are seeking a Basketball Analytics Database Programmer

- This full time position will report to the CTO and the Assistant General Manager / Team Counsel.
- This position will work with the information technology group and basketball operations in the development of basketball analytics infrastructure and applications.

# Questions?

# What is Data Science?

# 資料科學 Data Science



## DATA SCIENCE SKILLSET

Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.

**Hacking skills** are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

**Math and statistics knowledge** allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

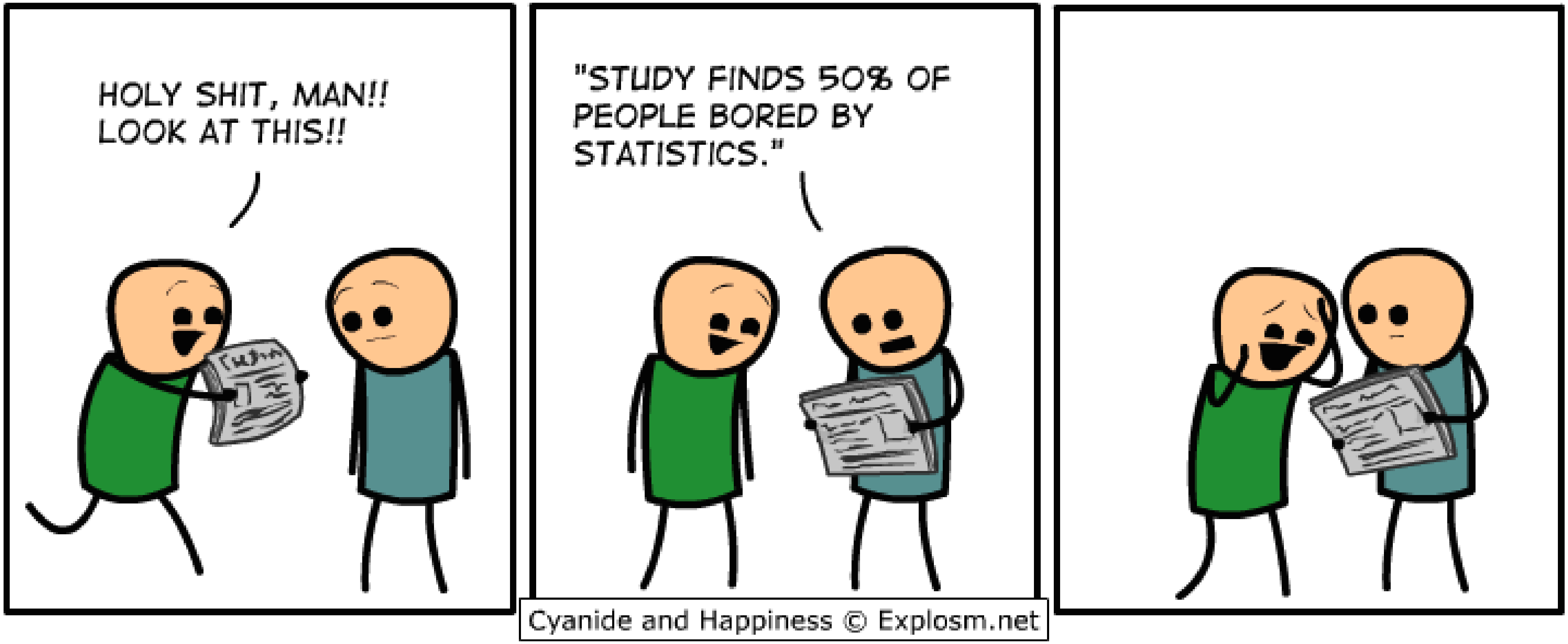
**Substantive expertise** in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

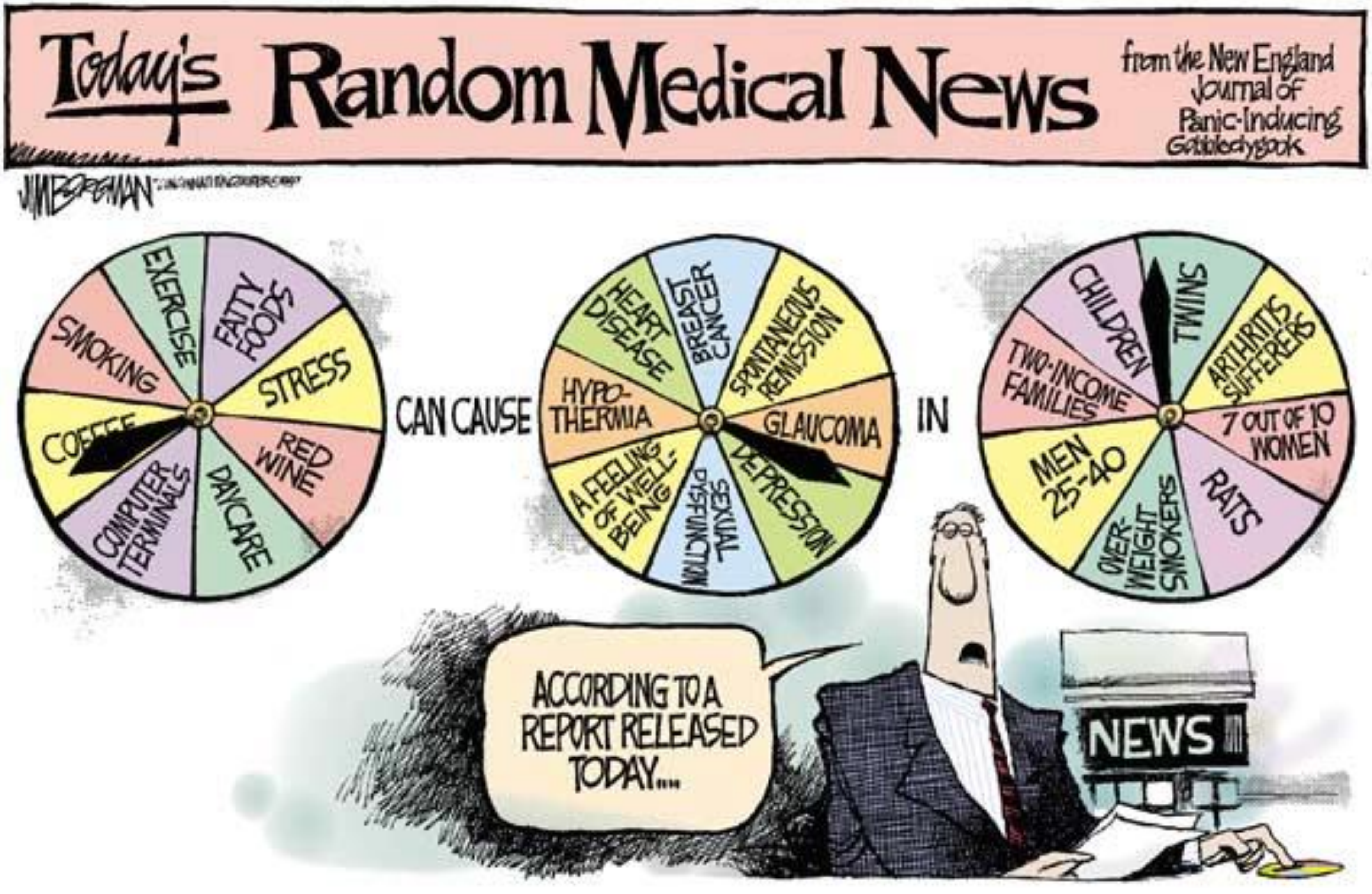
**Traditional research** lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

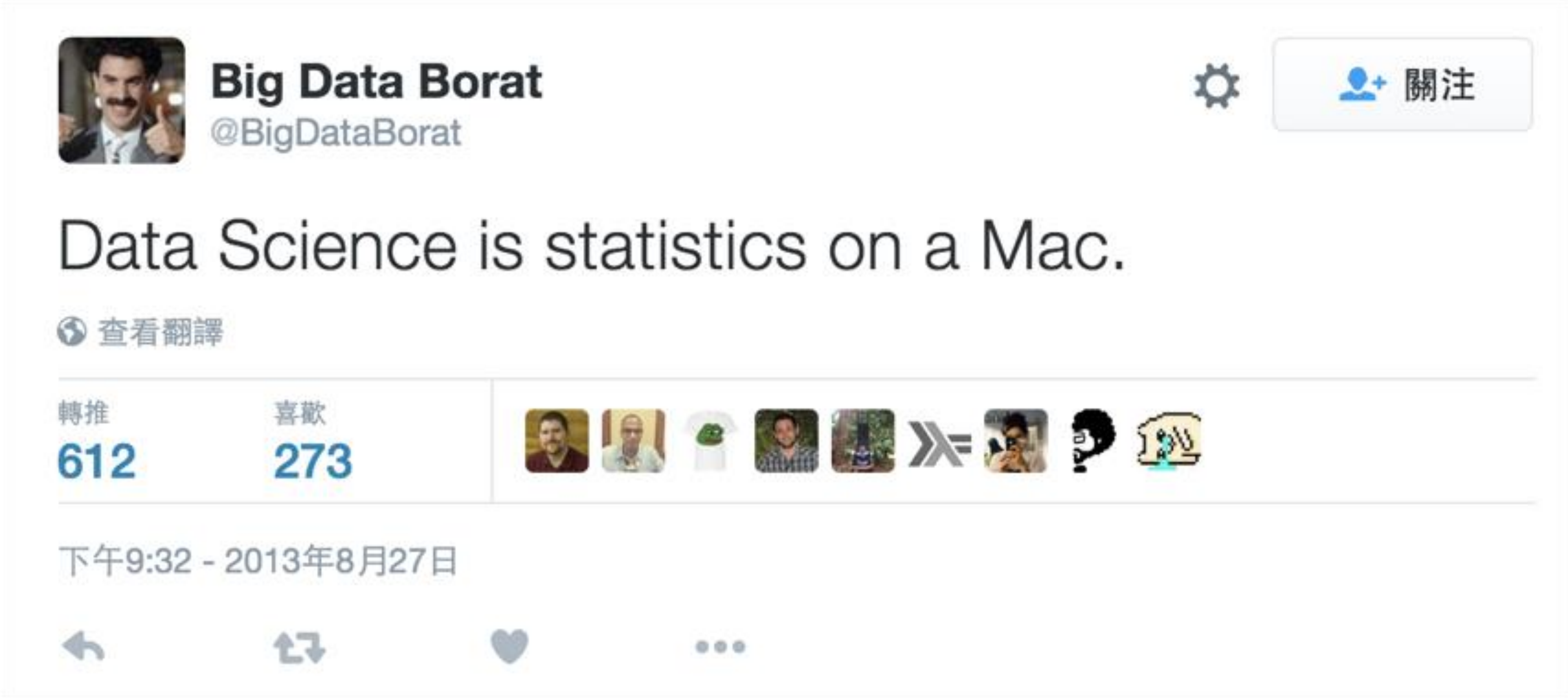
**Machine learning** stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

**Danger zone!** Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

Drew Conway's Data Science Venn Diagram


HOLY SHIT, MAN!! LOOK AT THIS!!
"STUDY FINDS 50% OF PEOPLE BORED BY STATISTICS."
Cyanide and Happiness © Explosm.net


Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
SMOKING
EXERCISE
FATTY FOODS
STRESS
RED WINE
DAYCARE
COMPUTER TERMINALS
COFFEE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
GLAUCOMA
DEPRESSION
SEXUAL DYSFUNCTION
A FEELING OF WELL-BEING
HYPO-THERMIA
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
7 OUT OF 10 WOMEN
RATS
OVER-WEIGHT SMOKERS
MEN 25-40
TWO-INCOME FAMILIES
ACCORDING TO A REPORT RELEASED TODAY...
NEWS



Big Data Borat
@BigDataBorat

關注

Data Science is statistics on a Mac.

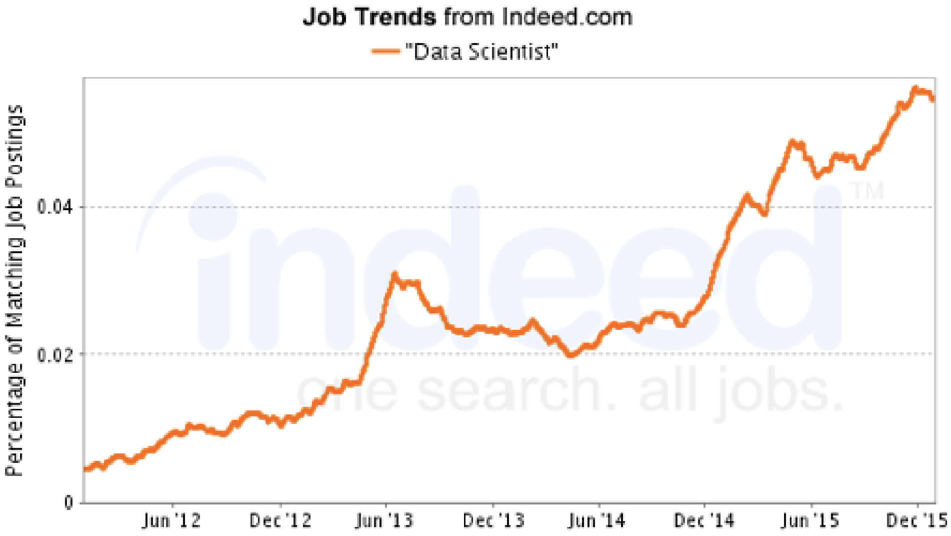
查看翻譯

| 轉推 | 喜歡 |
|---|---|
| 612 | 273 |

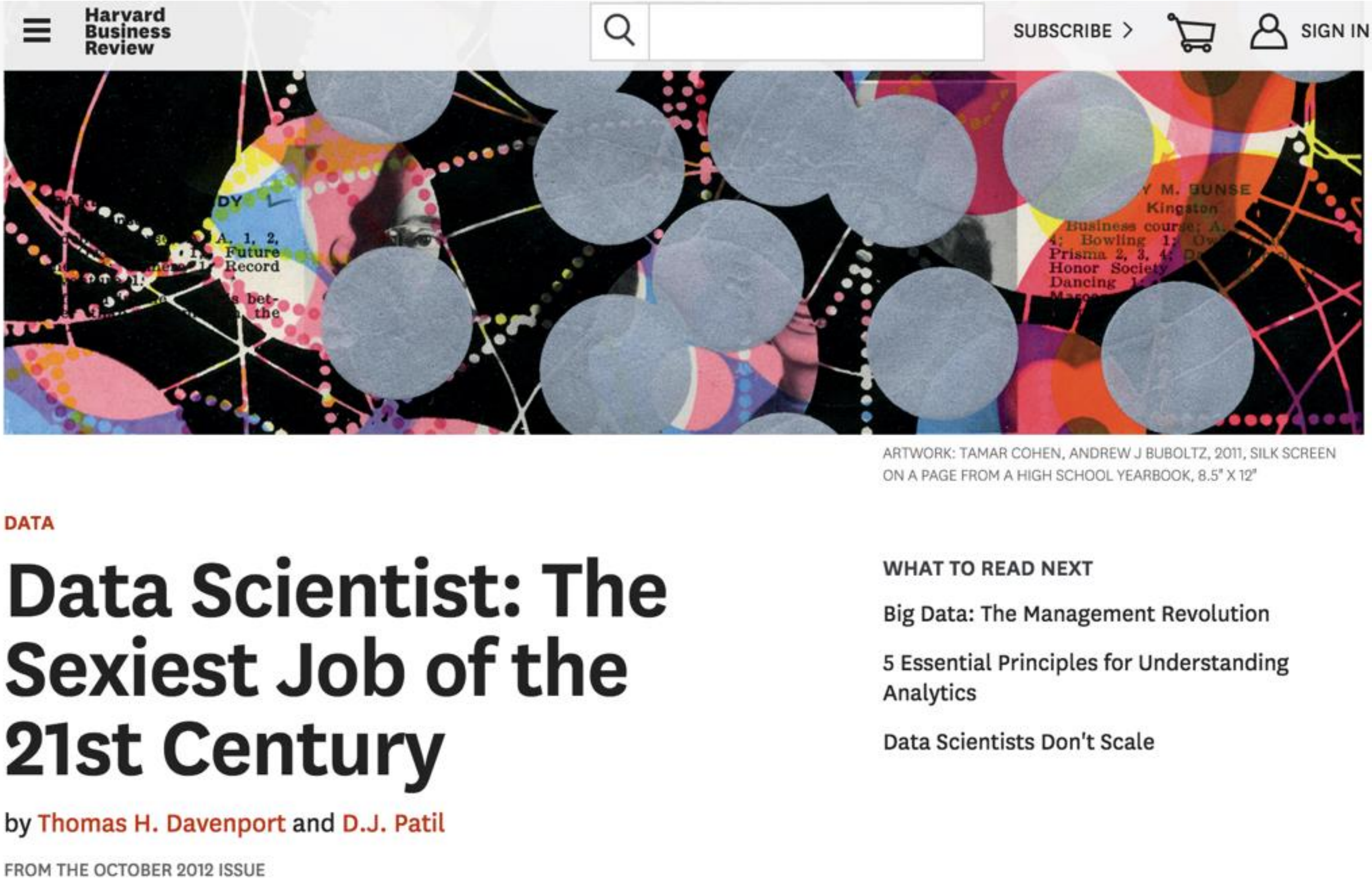
下午9:32 - 2013年8月27日

# 資料科學家 **Data Scientists**

- The ability to take data- to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it
  — that’s going to be a hugely important skill, NYT



http://www.indeed.com/publicanalytics/jobtrends?q=%22Data+Scientist%22&l=



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

# Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT

Big Data: The Management Revolution

5 Essential Principles for Understanding Analytics

Data Scientists Don't Scale

# Data Scientist

HELPING YOU REACH
VELOCITY

Data Science allows front offices to better predict what and when consumers are likely to buy.
The ability to write algorithms that find relationships in datasets is usable to provide actionable insight.

## The Challenge

-Data Mining
-Analysis
-Communication

## Industry Niche Titles

- Financial Institutions/ Decision Scienist
- Retailers/Omni Channel Expert
- Marketing Agencies/Consumer Behaviour Analyst
- Ecommerce/Analytics Expert

## Urgent Need

Data Scienticts - those with the technical savvy and analytical chops to derive meaning from all the information- are in high demand

## Skills by the Numbers

The skills and talents that make a fantastic Data Scientist

| | |
|---|---|
| Complex Formulas | 40% |
| Consumer Psychology | 25% |
| Business Acumen | 25% |
| Programming Languages | 10% |

## Did you Know?

Google's Eric Schmidt claims that every two days now we create as much information as we did from the dawn of civilization up until 2003

# What Do Data Scientists Do?

- Define the question
- Define the ideal data set
- Determine what data you can access

- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling

- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code
- Distribute results to other people

# Become A Data Scientist

- DBA: deal with unstructured data
- Statistician: data that does not fit in memory
- Software engineer: learn statistical modeling + communicate results
- Business analyst: learn algorithm + trade of scale

# Questions?