

# Introduction to Data Science (Khoa học dữ liệu)

**Khoa** Than

Hanoi University of Science and Technology  
[khoattq@soict.hust.edu.vn](mailto:khoattq@soict.hust.edu.vn)

IT4142E, SOICT, HUST, 2019

# Contents

---

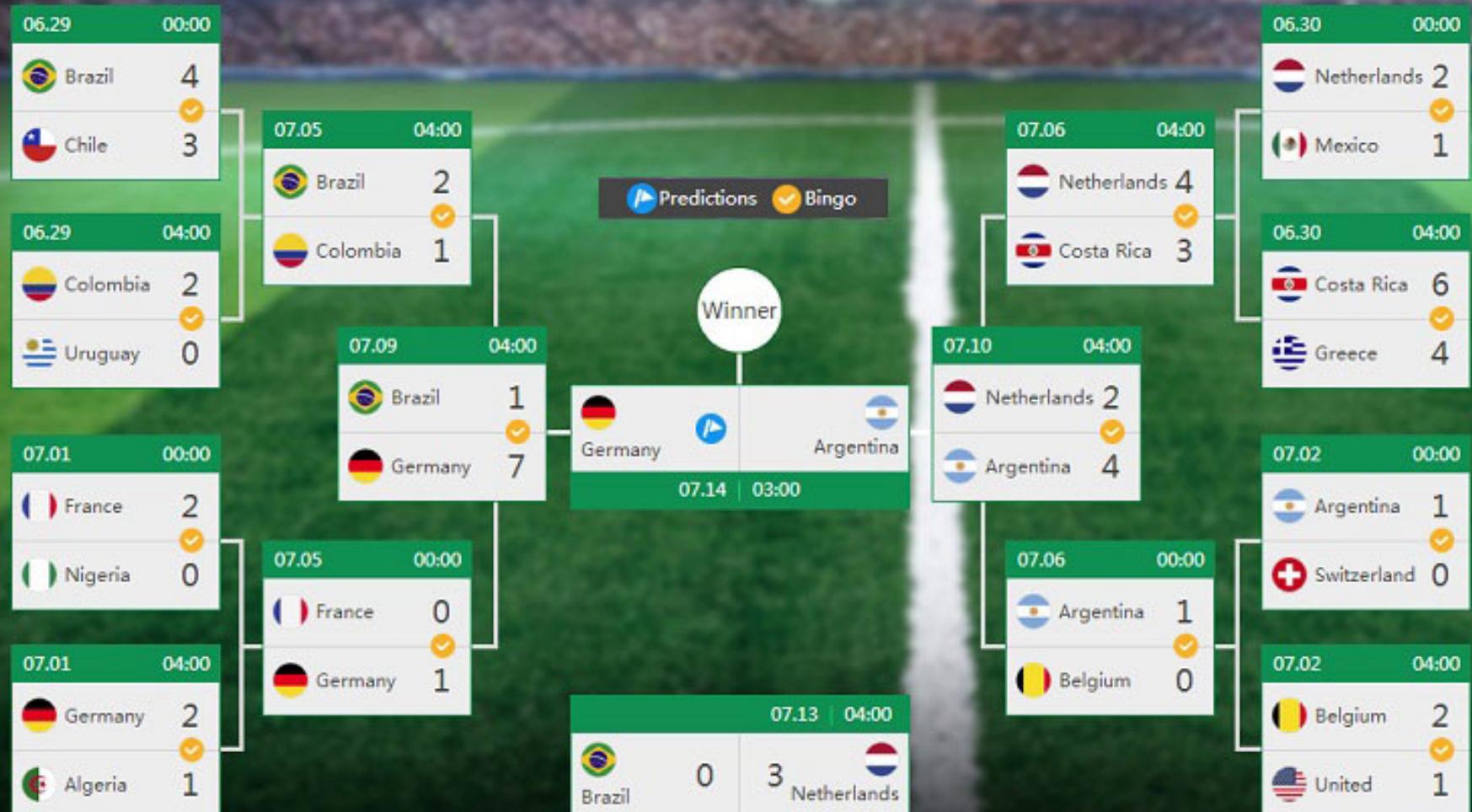
- **Introduction to Data Science**
- Data crawling and processing
- Data cleaning and integration
- Exploratory data analysis
- Machine Learning
- Big data analysis
- Visualization
- Text analysis
- Image and video analysis
- Graph analysis
- Recommender system

## Some questions

- Ung thư ngày càng nhiều, nguyên nhân do đâu?
  - Bộ trưởng Y tế: "Ung thư chết nhiều không phải do thực phẩm bẩn"
  - ??
- Ảnh hưởng của báo chí
  - "Báo Đất Việt" - "Thực phẩm bẩn là nguyên nhân hàng đầu gây ung thư" - BC.com"
  - ??
- Thị yếu xem truyền hình hiện nay như thế nào?  
Kênh nào hấp dẫn khán giả nhất?
  - ??

**Let the data speak**

# Some examples: FIFA prediction (2014)



# Amazon's recommendation secret



"The company reported a **29% sales increase** to \$12.83 billion during its second fiscal quarter, up from \$9.9 billion during the same time last year."

– Fortune, July 30, 2012

## Lower Priced Items to Consider



LG 34UM68-P 34-Inch 21:9...

★★★★★ 164

\$389.89 ✓Prime



LG 27UD68-P 27-Inch...

★★★★★ 54

\$439.00 ✓Prime

Is this feature helpful?



## Customers Who Bought This Item Also Bought



Cable Matters Thunderbolt  
2 Cable in White 6.6 Feet /  
2m

★★★★★ 10



Cable Matters Thunderbolt  
2 Cable in Black 6.6 Feet /  
2m

★★★★★ 38

\$38.99 ✓Prime



Cable Mat...

★★★★★

\$31.99 ✓P...

## Some examples: IBM's Watson



IBM's Watson Supercomputer Destroys Humans in Jeopardy (2011)

# IBM's massive bet on Data Analytics

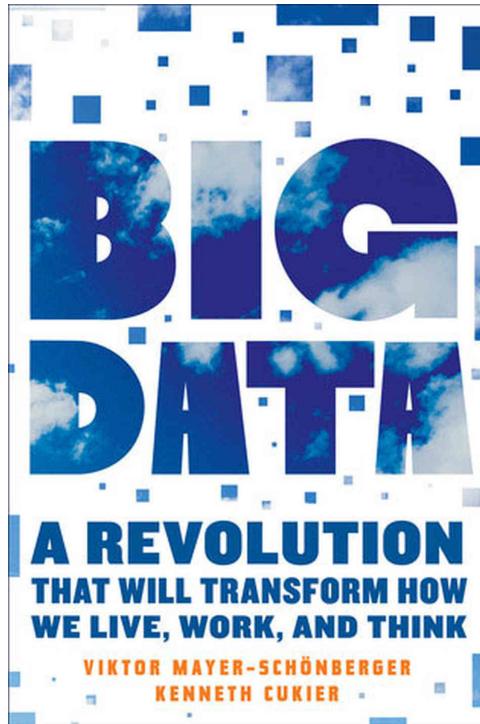
The illustration features a central glowing blue sphere with green and yellow orbits, symbolizing data and connectivity. A large circuit board forms the background, representing technology. Several panels and callouts describe the product's capabilities:

- Left Panel:** "DISCOVER NEW INSIGHTS LOCKED AWAY IN MILLIONS OF PAGES. IN SECONDS." Includes icons of a DNA helix, laboratory glassware, and lungs.
- Middle Left:** "IF YOU'RE LIKE MANY ORGANIZATIONS, INSIGHT IS HARD TO COME BY." Includes a DNA helix, a book, and lungs.
- Middle Left Text:** "THE PROCESS IS MANUAL AND SLOW. IT'S FRAGMENTED AND PIECE MEAL. IT'S LIMITED TO A NARROW POINT OF VIEW IN A WORD. IT'S OUTDATED."
- Middle Left Callout:** "IMAGINE INSTEAD INSIGHT FROM MILLIONS OF PAGES. IN SECONDS. WITH TIMELY UPDATES."
- Center:** "INTRODUCING IBM WATSON DISCOVERY ADVISOR."
- Bottom Center:** "YOU ASK THE QUESTION. WATSON LOOKS FOR PATTERNS, THEN PROVIDES RELEVANT RESPONSES BACKED BY EVIDENCE. ALL OF WHICH CAN HELP YOU UNCOVER WHAT'S LIKELY NEVER BEEN DISCOVERED."
- Bottom Left:** "IN THE PAST, TOOLS PROVIDED A LIST OF DOCUMENTS TO GO THROUGH MANUALLY. NOW, WATSON CAN SYNTHESIZE MILLIONS OF PAGES FOR INSIGHT WHERE IT'S NEEDED MOST."
- Right Side:** "WHEN INDUSTRIES NEED ANSWERS, WATSON DELIVERS INSIGHTS YOU CAN ACT ON. FASTER." Includes a doctor and patient icon.
- Top Right:** "ACCELERATING BREAKTHROUGHS IN CLINICAL TRIALS WITH THE GOAL OF IMPROVING CARE." Includes a doctor and patient icon.
- Bottom Right:** "ASSISTING WITH NEW DRUG DISCOVERY IN PHARMA TO HELP COMBAT DISEASE AND EASE PAIN." Includes illustrations of prescription bottles and pills.
- Bottom Right Callout:** "EXPANDING RESEARCH POSSIBILITIES IN EDUCATION WITH UNIVERSITIES SUCH AS NORTH CAROLINA STATE."
- Bottom Right Text:** "WATSON DISCOVERY ADVISOR. WHAT MIGHT YOU DISCOVER NEXT?"
- Bottom Right Logo:** IBM logo with the text "IBMWATSON.COM".

# 21<sup>st</sup> century

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

by Thomas H. Davenport and D.J. Patil



# Where the data? Social networks

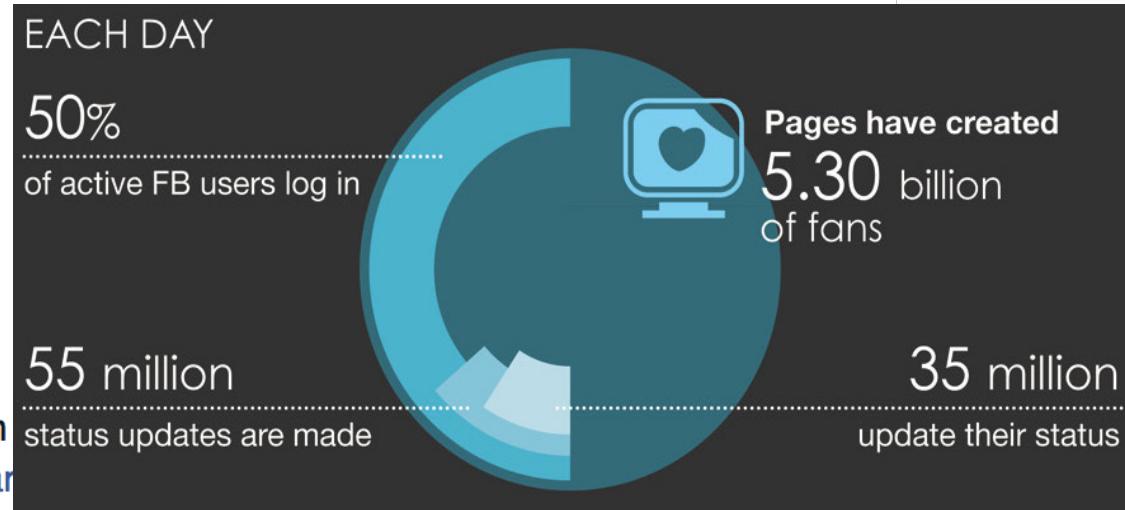
# facebook®



**Taylor Swift** đã thêm 4 ảnh mới.

4 Tháng 4 lúc 19:52 · ●

What an unbelievable run we've had with these memories & all of you. #iHeartAwar...



7,174 Tweets sent in 1 second

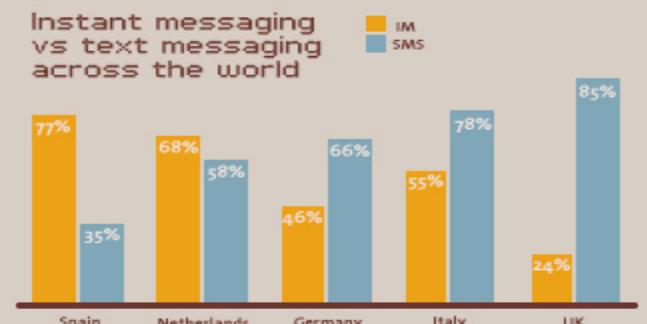
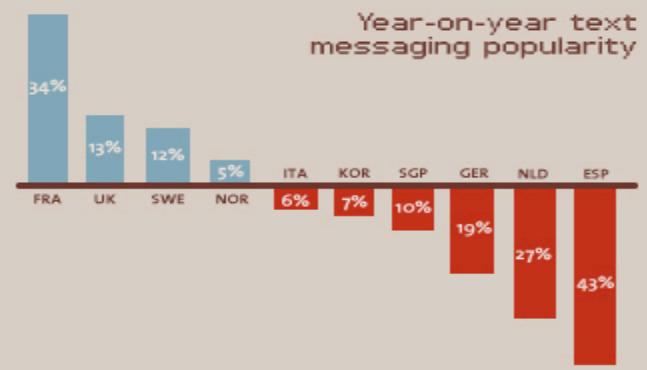
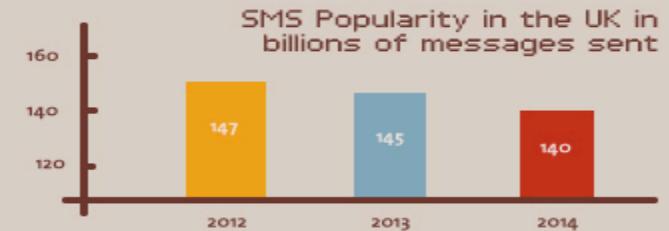
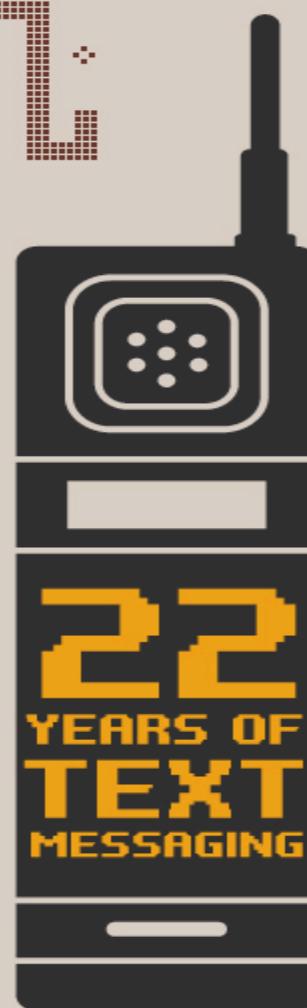
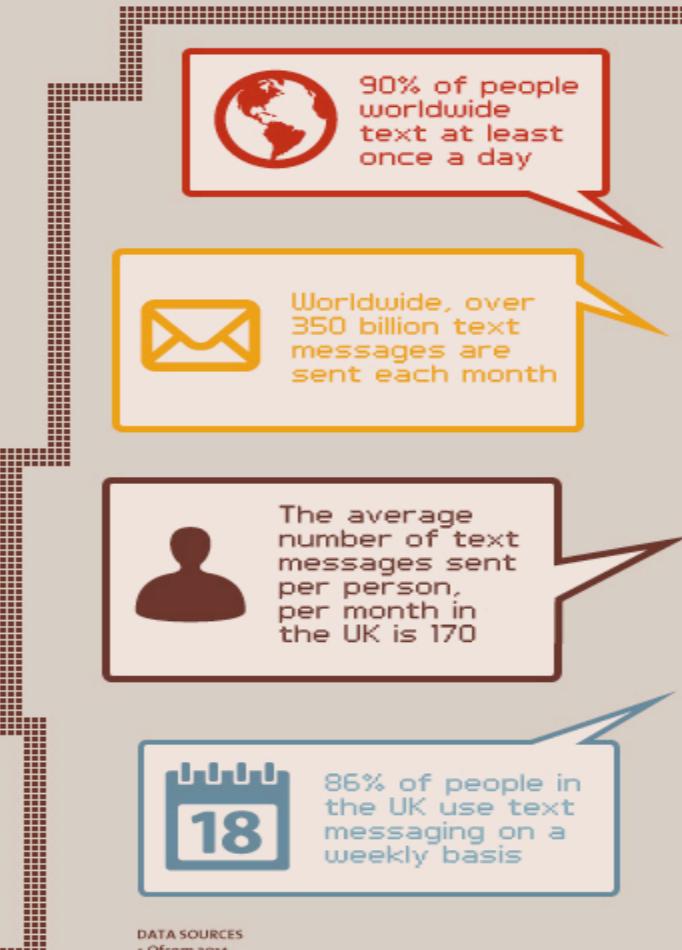


862,696 Tweets since opening this page  
0:02:00 seconds ago

**Basit Alvi** @bpk69 · 6m  
Swiss banker whistleblower: CIA behind **Panama Papers** [cnb.cx/1WpVjgK](http://cnb.cx/1WpVjgK)

**Violamagic** @TrautCarol · 6m  
Why The **Panama Papers** Scandal Is About Cheating School Children  
[educationopportunitynetwork.org/why-the-panama...](http://educationopportunitynetwork.org/why-the-panama...)

# Where the data? Mobile messages



# Where the data? Internet



**3,669,276,617**

Internet Users in the world



**1,215,275,272**

Total number of Websites



**156,807,290,512**

Emails sent **today**



**3,542,616,621**

Google searches **today**



**3,316,277**

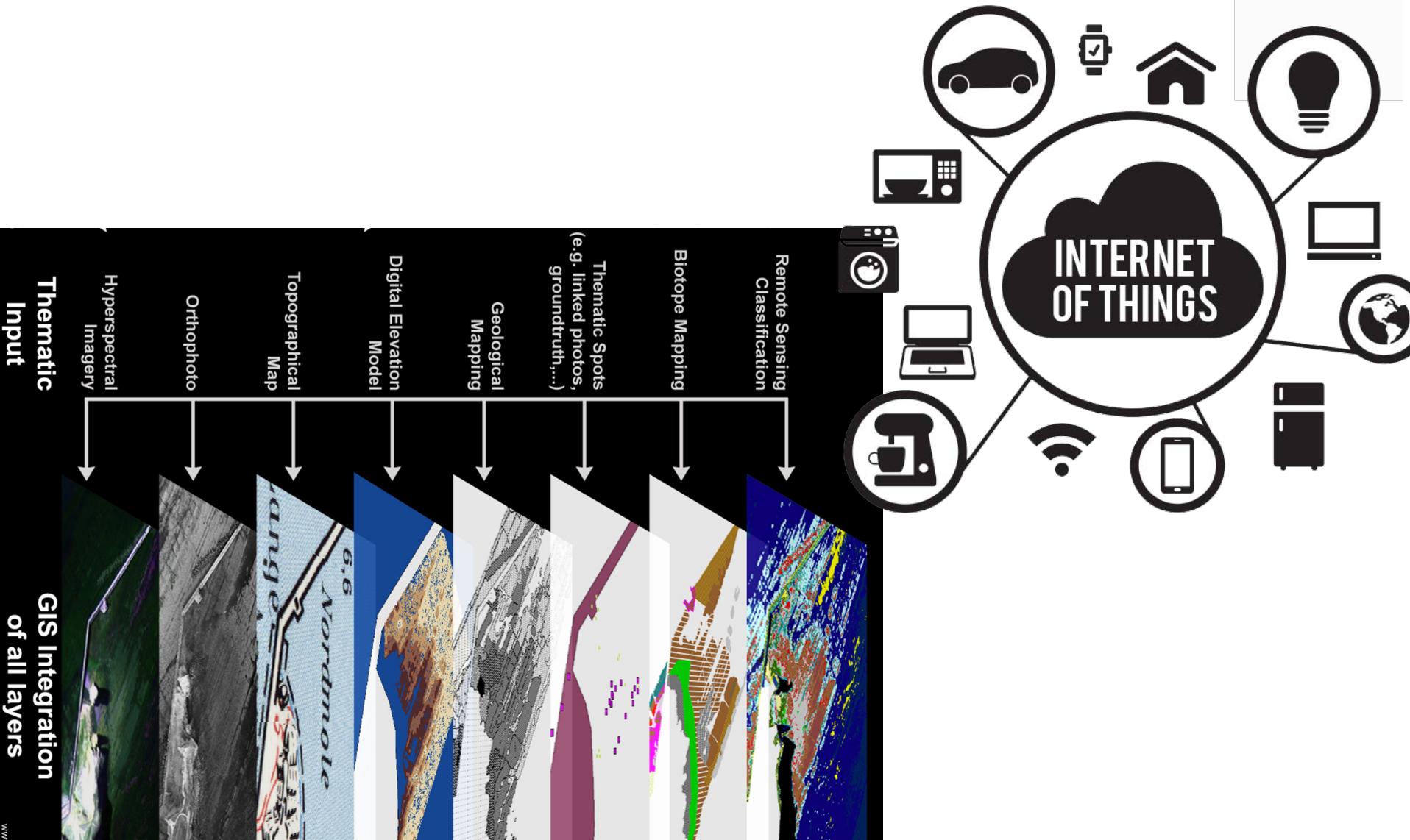
Blog posts written **today**



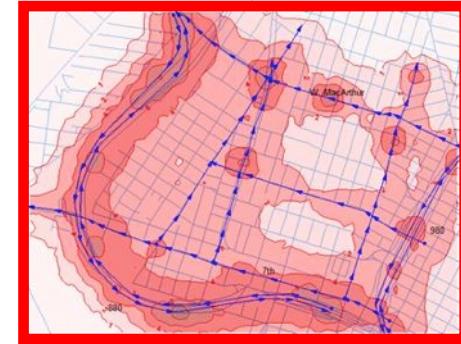
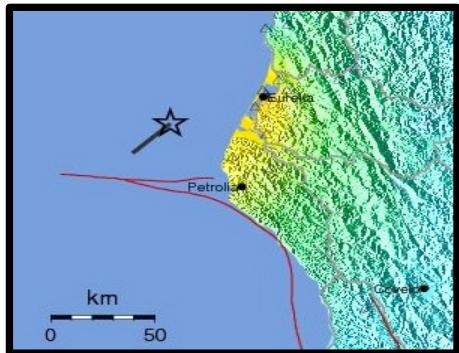
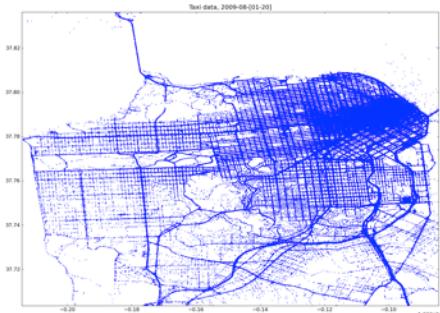
**444,323,737**

Tweets sent **today**

# Where the data? And more



# What can we do with the data?



Crowdsourcing + physical modeling + sensing + data assimilation

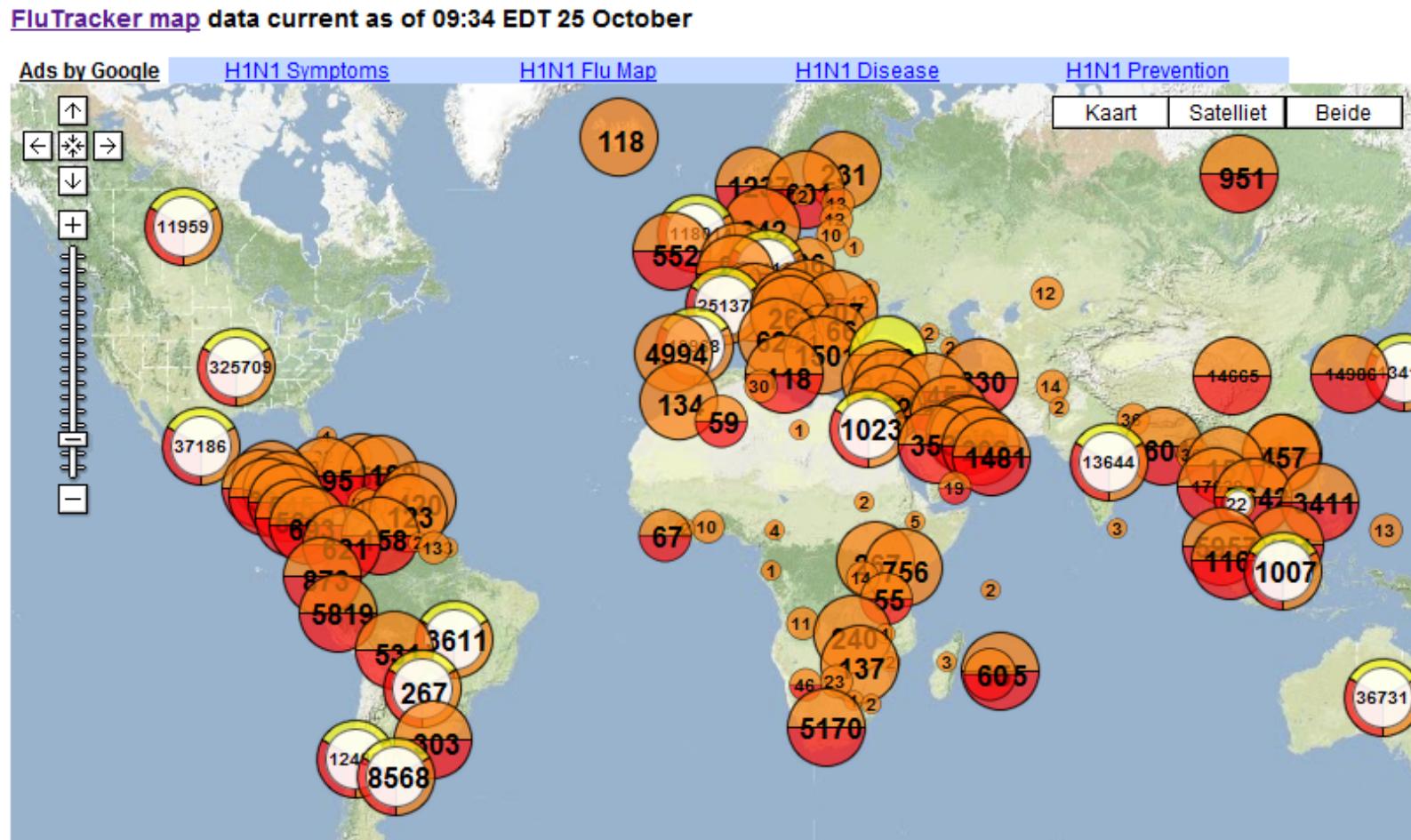
to produce:



(Alex Bayen, UC Berkeley)

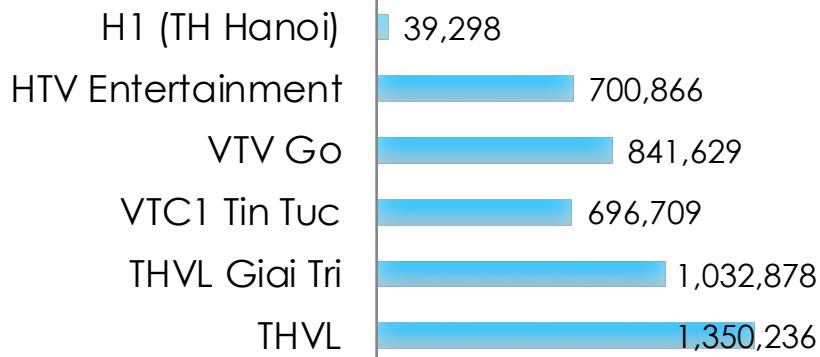
# What can do? Prediction

- **Google Flu Trends:** detecting outbreaks *two weeks ahead* of CDC data

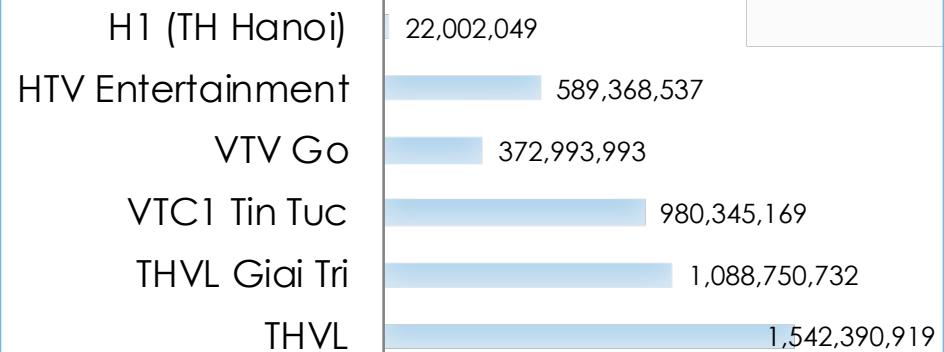


# What can do? Exploration

## Subscribers in Youtube



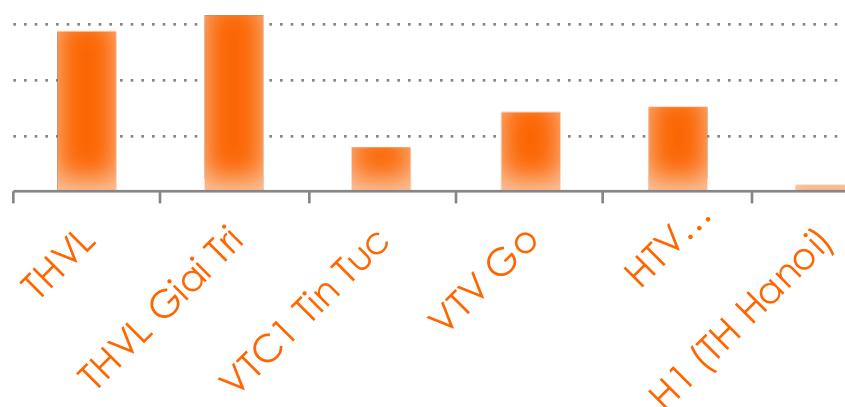
## Views in Youtube



Effective TV channels?

(July 4, 2017)

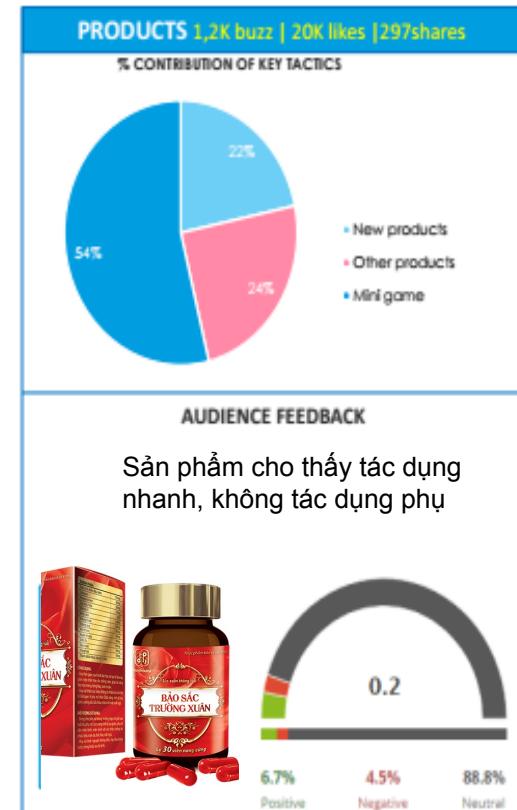
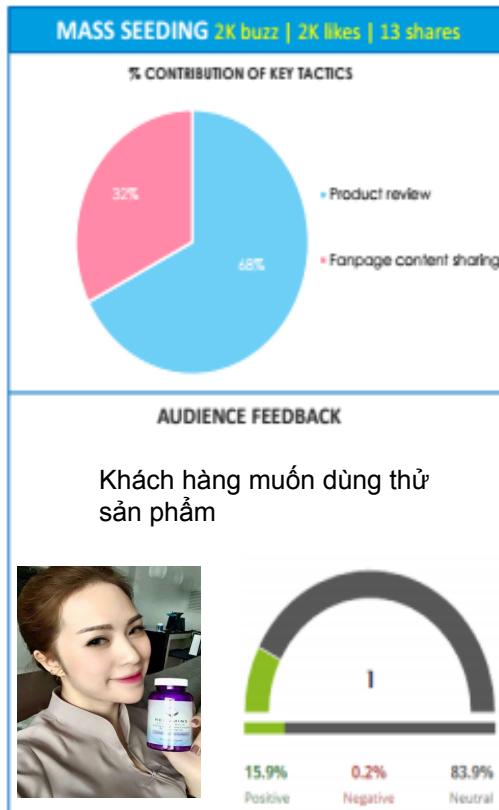
## Attractiveness



# What can do? Planning

- Make plan and test it (effectiveness, changer, ...)

## DEEP DIVE INTO EACH ACTIVITY



# Data Science

# Data Analysis has been around

1935: "The Design of Experiments"

R.A. Fisher



1939: "Quality Control"

W.E.  
Demming

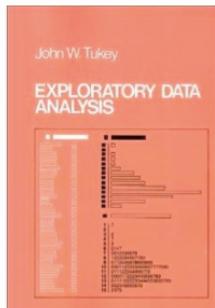


1958: "A Business Intelligence System"

Peter Luhn



1977: "Exploratory Data Analysis"



Howard  
Dresner



1997: "Machine Learning"



2010: "The Data Deluge"

1996: Google



2007: "The Fourth Paradigm"



(John Canny, UC Berkeley)



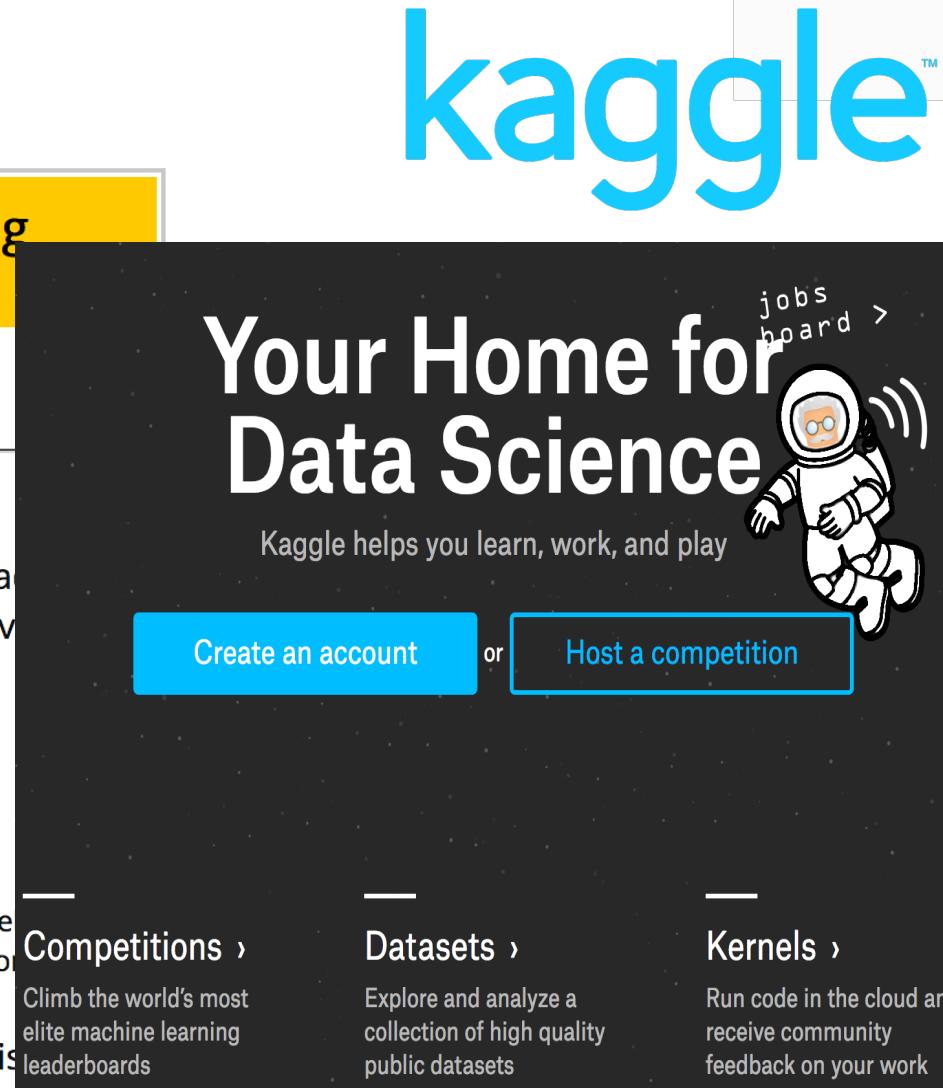
# Online platforms



The image shows the KDnuggets logo, which consists of the letters 'KD' in a large, bold, yellow font on a dark background, followed by the word 'nuggets' in a smaller, regular yellow font. Below the logo is a yellow banner with the text 'Analytics, Data Science, Data Mining Competitions' in black.

## Notable Recent Competitions

- [GE NFL \\$10 Million Head Health Challenge](#), for more accurate diagnoses of mild brain injury and prognosis for recovery following acute and/or repetitive injuries.
- [GE Hospital Quest on Kaggle](#).  
Your challenge: Contribute to the design of the ultimate patient experience. Prize Pool: \$100,000
- [GE Flight Quest on Kaggle](#).  
Your Challenge: Develop a usable and scalable algorithm that defines real-time flight profile to the pilot, helping them make flights more efficient and reliably on time. Prize Pool: \$250,000
- [Heritage Health Data Analysis Prize \(\\$3M\)](#), can administered health care data be used to accurately predict which patients



The image shows the homepage of Kaggle. The top half features a large blue 'kaggle' logo. Below it, the text 'Your Home for Data Science' is displayed in white. To the right, there is a cartoon illustration of an astronaut in a spacesuit. Above the astronaut, the words 'jobs board >' are written in a small font. Below the main title, the text 'Kaggle helps you learn, work, and play' is shown. At the bottom, there are two blue buttons: 'Create an account' and 'Host a competition'. On the left side, there are three sections with arrows: 'Competitions >', 'Datasets >', and 'Kernels >'. Each section has a brief description below it. The 'Competitions' section says 'Climb the world's most elite machine learning leaderboards'. The 'Datasets' section says 'Explore and analyze a collection of high quality public datasets'. The 'Kernels' section says 'Run code in the cloud and receive community feedback on your work'.

# What is it?

**Khoa học dữ liệu** là  
ngành *học (tìm ra tri thức)*  
*từ dữ liệu.*

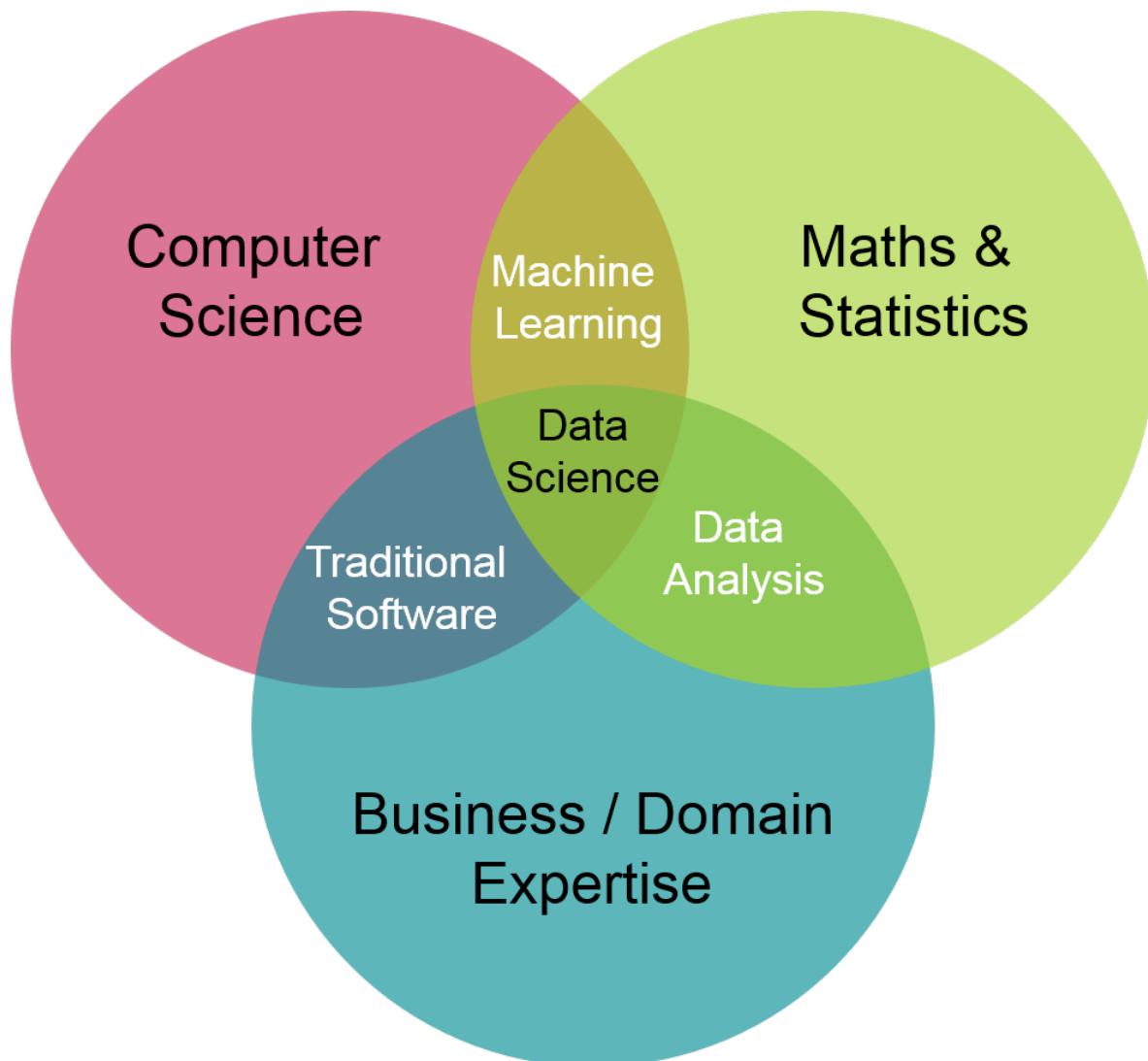


**Data science** is  
the science of *learning from data.*

(David Donoho, Stanford University)

# What is it?

---



# Why is Data Science important?

“Khả năng biết lấy dữ liệu, để **hiểu**, để **xử lý**, để **tạo ra giá trị** từ chúng, để **trao đổi** những giá trị đó với người khác, là một trong những kỹ năng cực kỳ quan trọng trong những thập kỷ tới. Những kỹ năng đó cần thiết ngay cả đối với các học sinh phổ thông. **Lý do là vì chúng ta thực sự đang có dữ liệu ở khắp nơi và miễn phí.**”

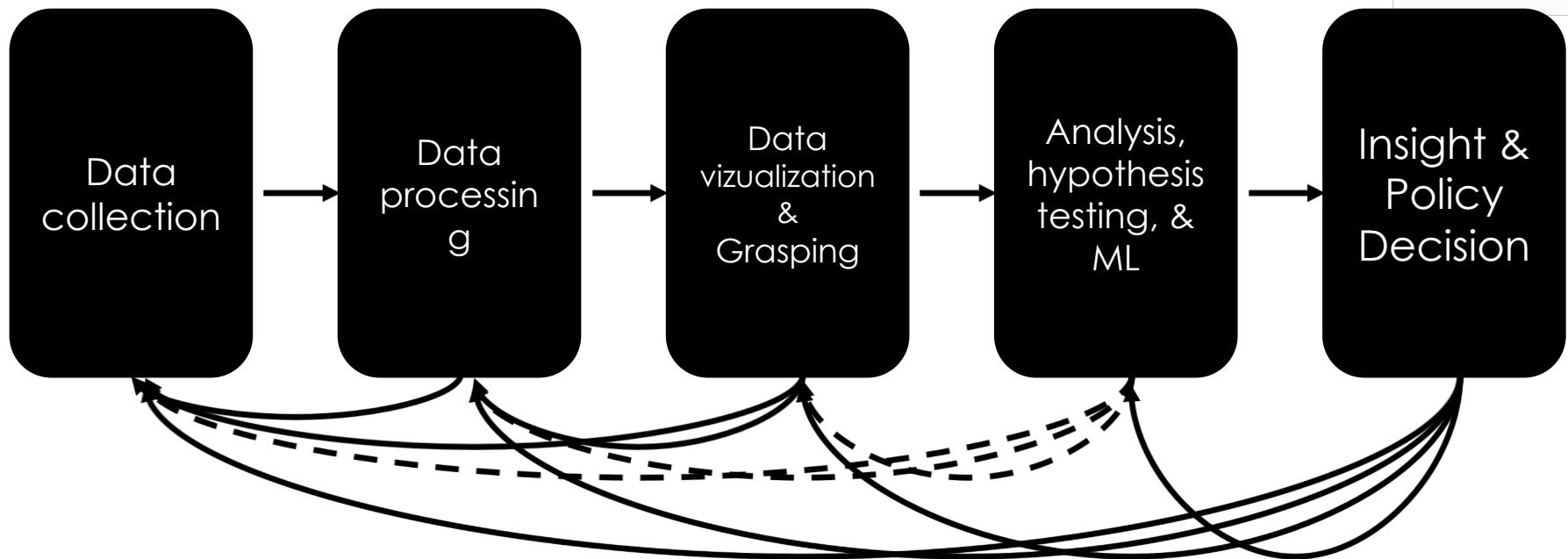
- Hal Varian, Nhà kinh tế Trưởng của Google



“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's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. **Because now we really do have essentially free and ubiquitous data.**”

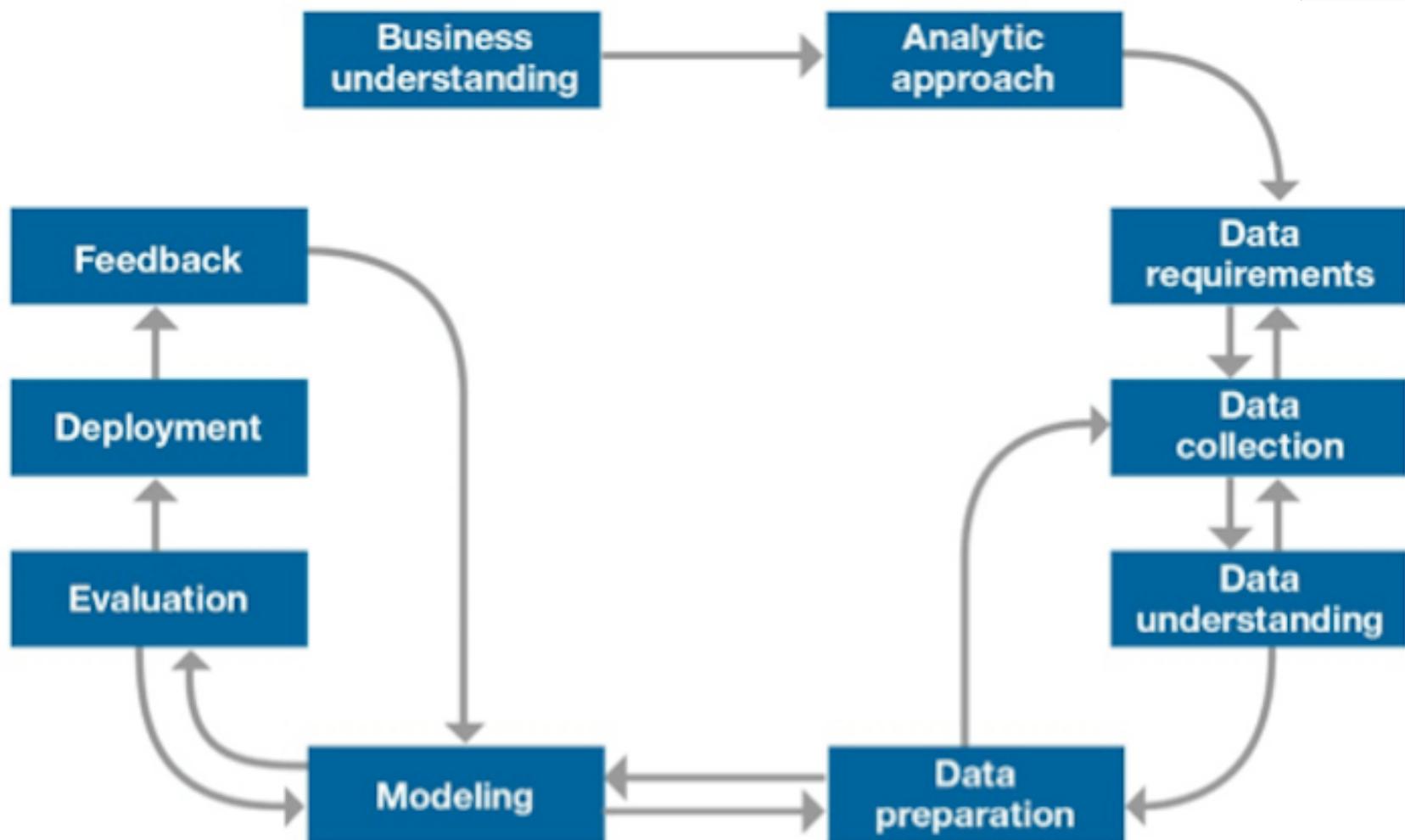
- Hal Varian, Google's Chief Economist

# DS methodology: insight-driven



(John Dickerson, University of Maryland)

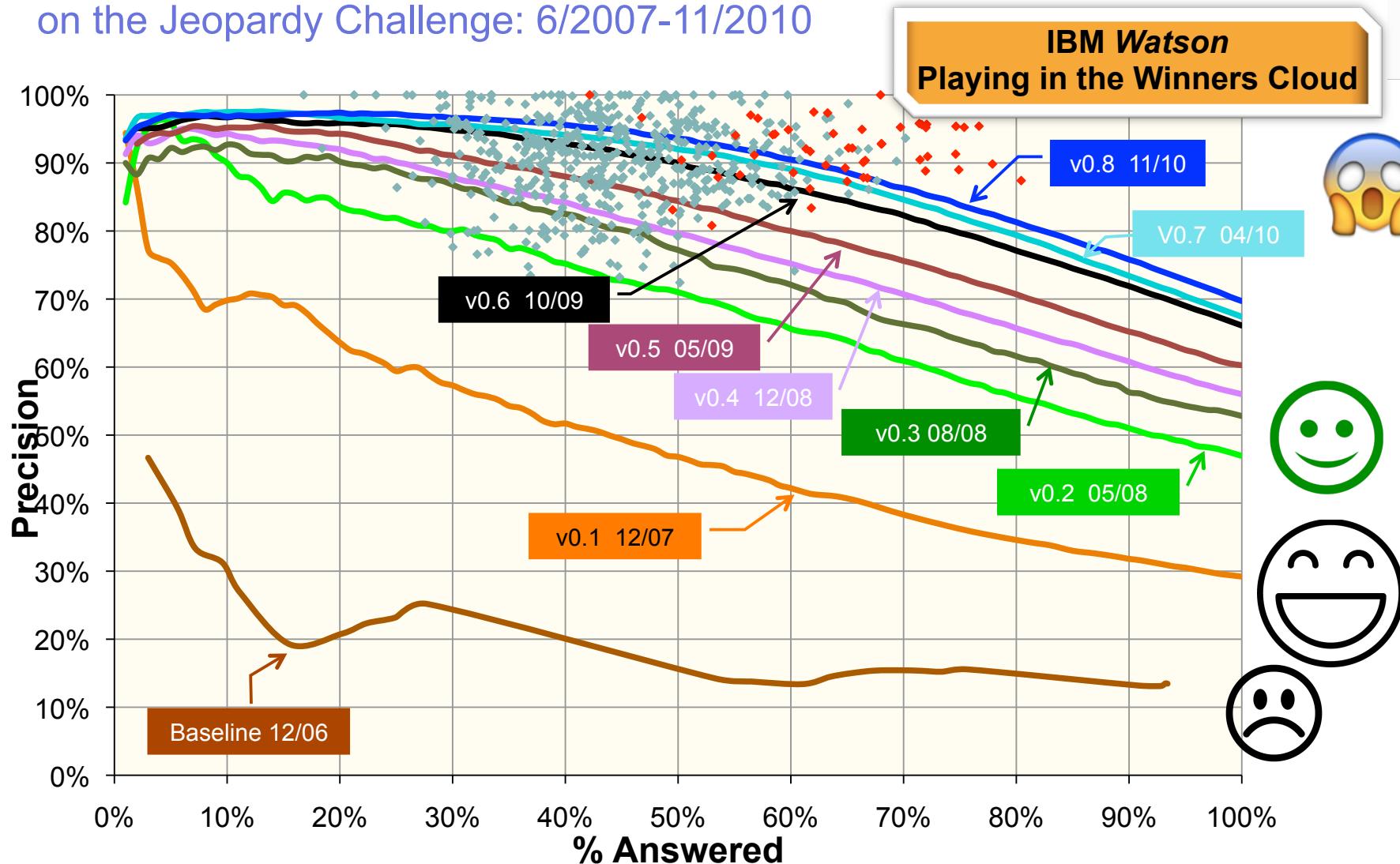
# DS methodology: product-driven



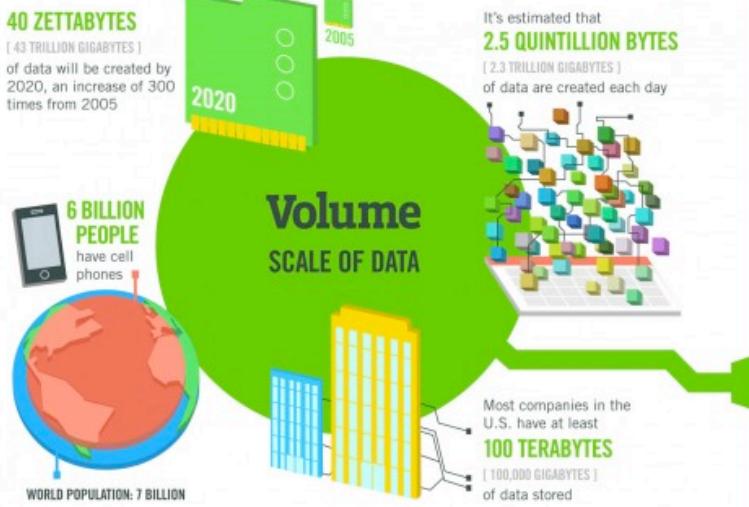
(<http://www.theta.co.nz/>)

# Product development: experience

DeepQA: Incremental Progress in Answering Precision  
on the Jeopardy Challenge: 6/2007-11/2010



# Challenges



By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth.

## 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.

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

**150 EXABYTES** [ 161 BILLION GIGABYTES ]

**30 BILLION PIECES OF CONTENT**

are shared on Facebook every month

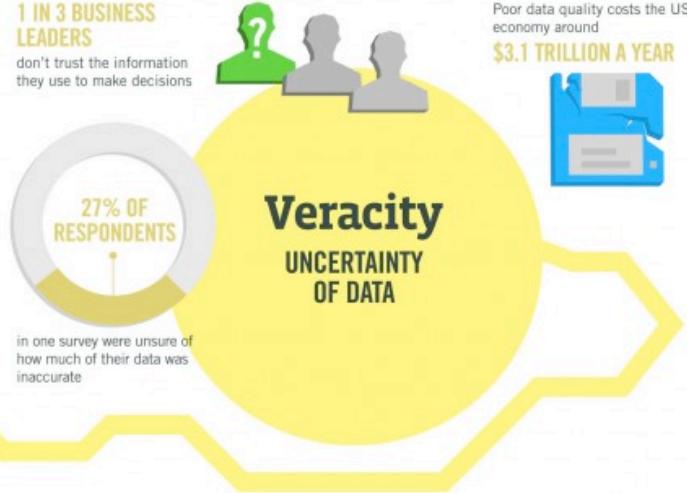


**400 MILLION TWEETS**

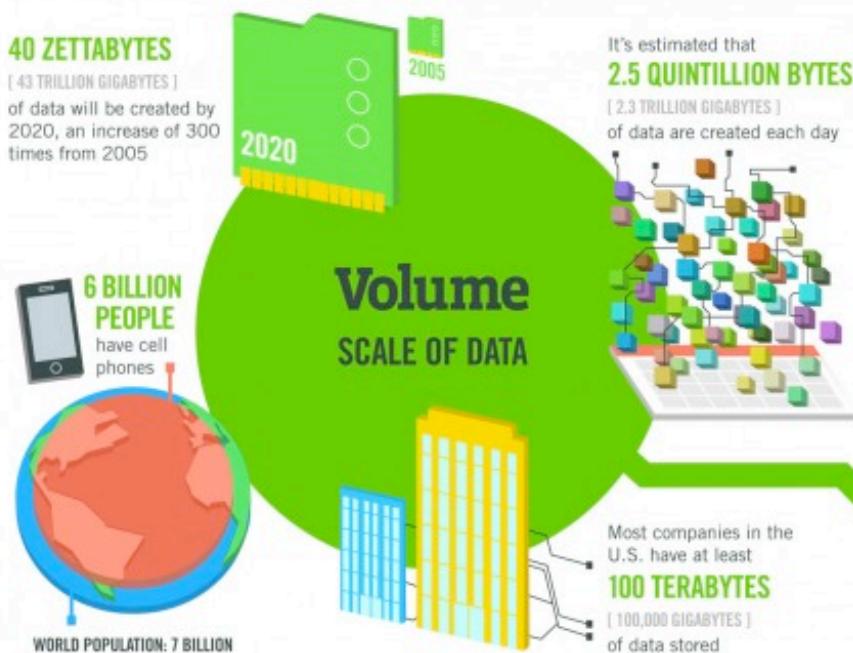
are sent per day by about 200 million monthly active users

**1 IN 3 BUSINESS LEADERS**

don't trust the information they use to make decisions

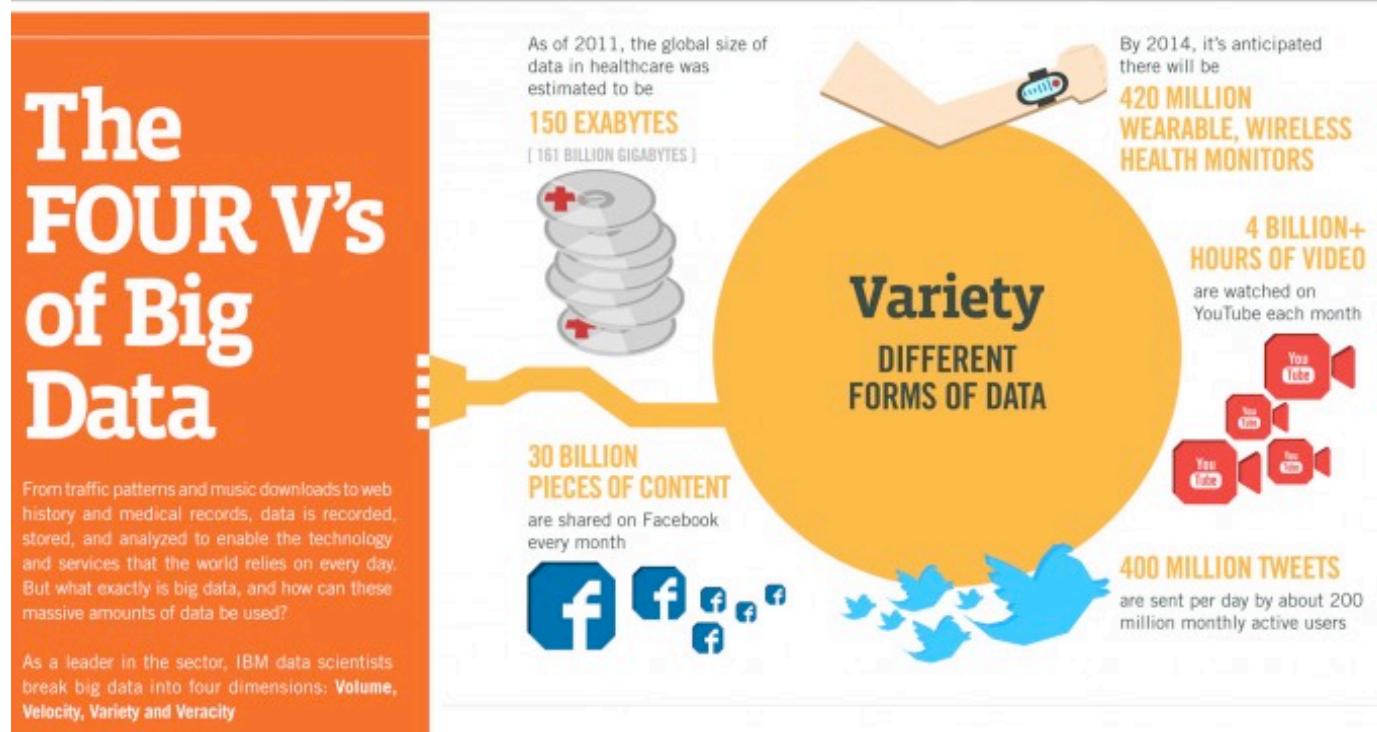


# Challenges: volume



Số lượng dữ liệu  
thu thập được  
quá nhiều

# Challenges: variety



Dữ liệu  
quá đa dạng

# Challenges: velocity

Dữ liệu  
đến liên tục  
và nhanh

The New York Stock Exchange captures  
1 TB OF TRADE INFORMATION  
during each trading session

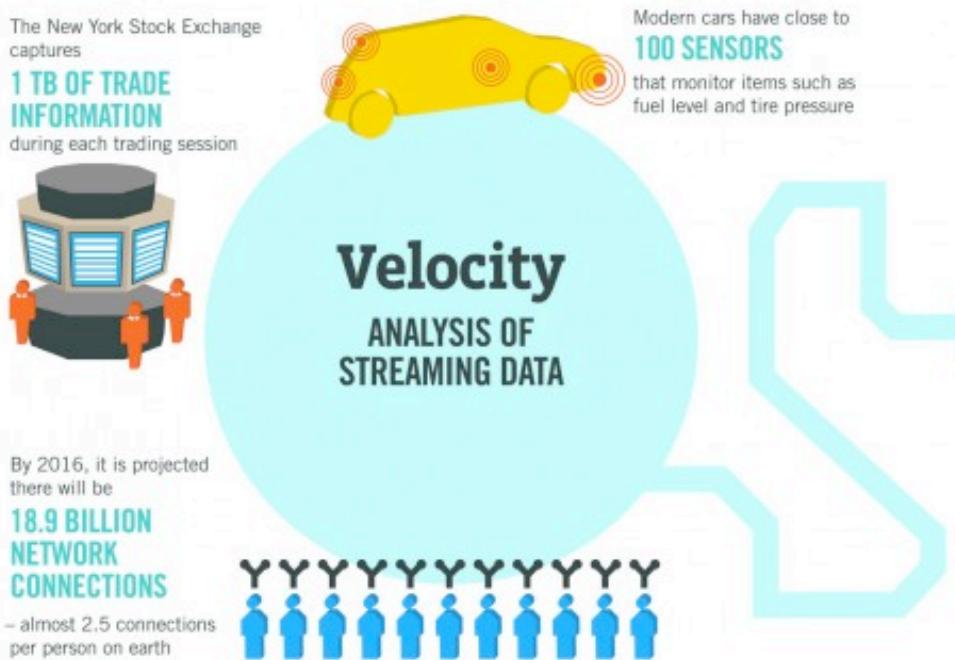


By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

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

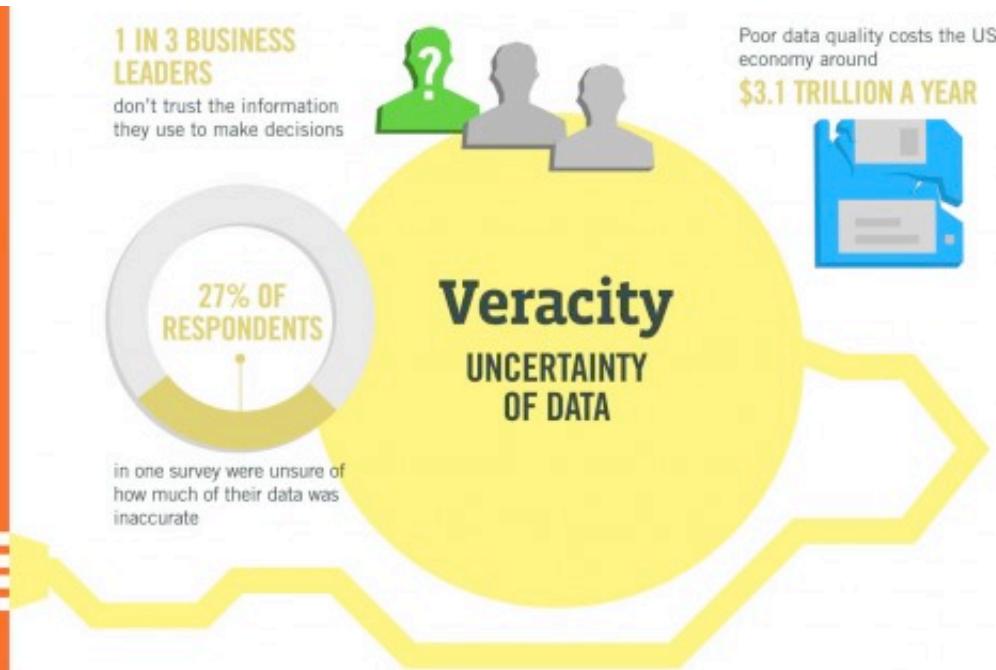


Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs,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

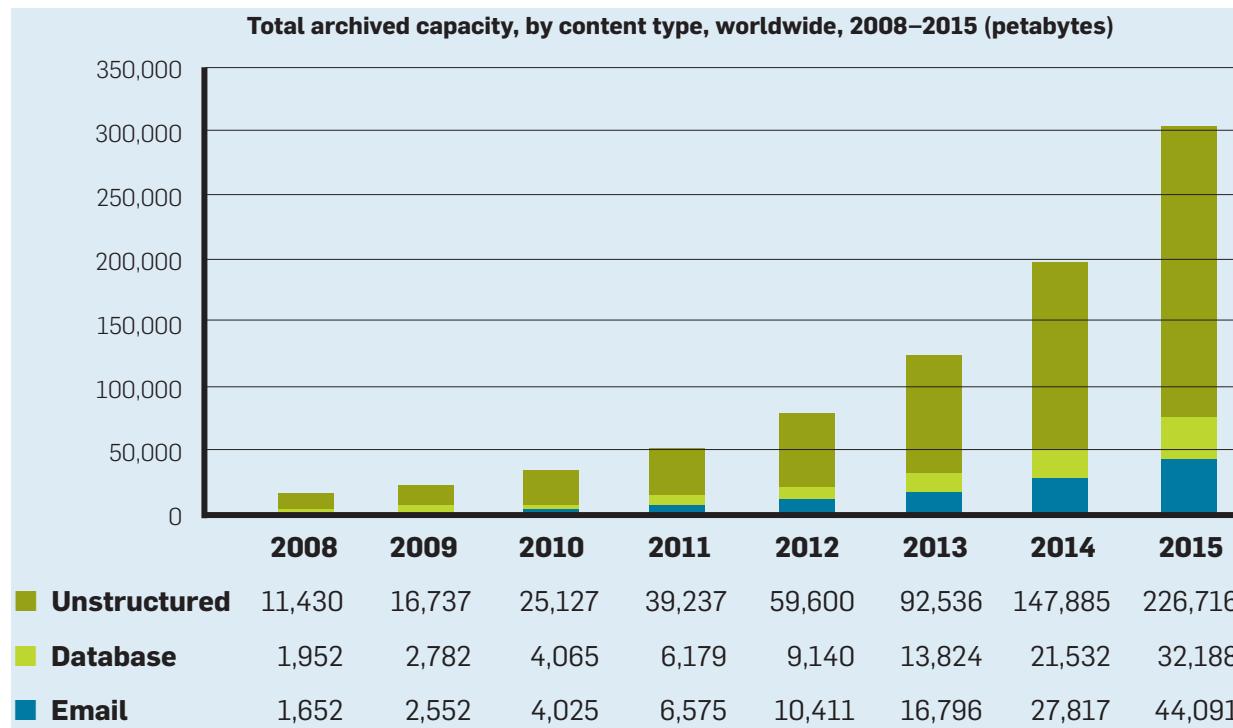


# Challenges: veracity



# Challenges: free structure

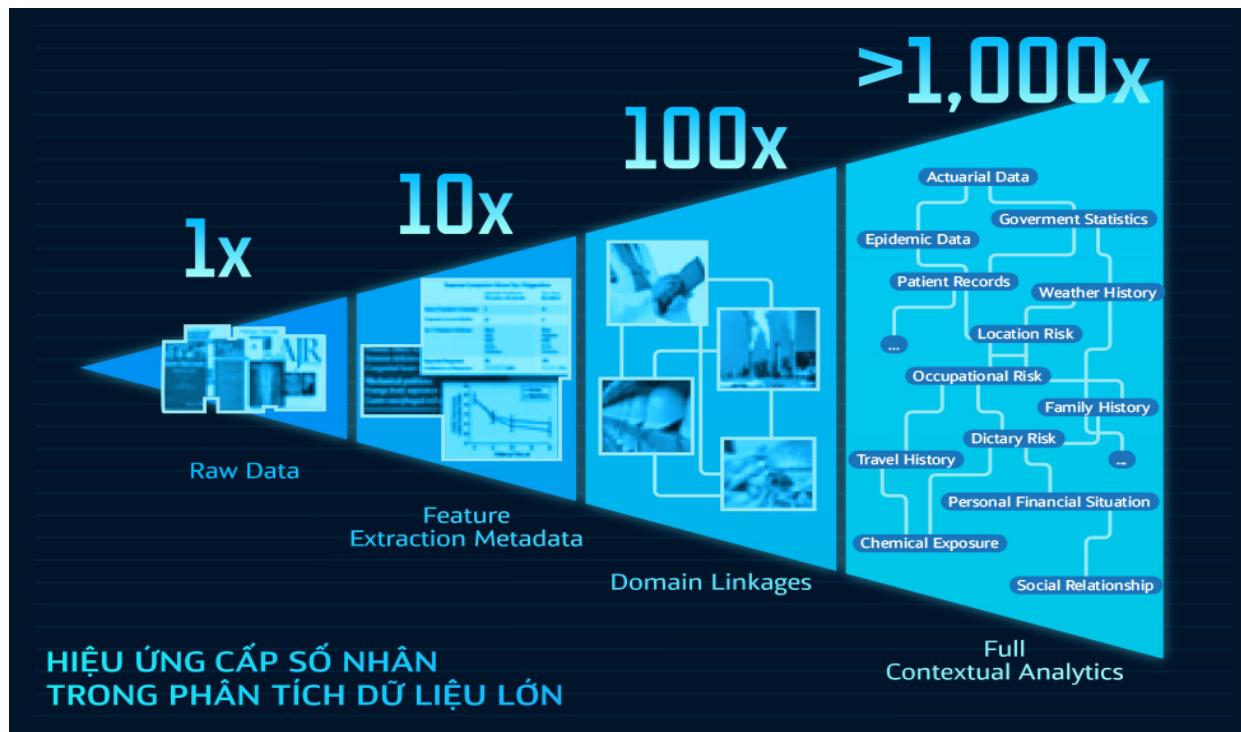
- Unstructured data increases extremely fast
  - Texts, images, tags, links, likes, emotions, ...



(Vasant Dhar, CACM, 2013)

# Challenges: hidden interaction

- The interactions or correlations hidden in data might be really huge
  - (Những mối tương tác ẩn chứa bên trong dữ liệu có thể rất lớn)



(<http://genk.vn/ibm-va-cuoc-cach-mang-tri-tue-nhan-tao-mang-ten-watson-20160830152953753.chn>)

# Challenges: high dimensionality

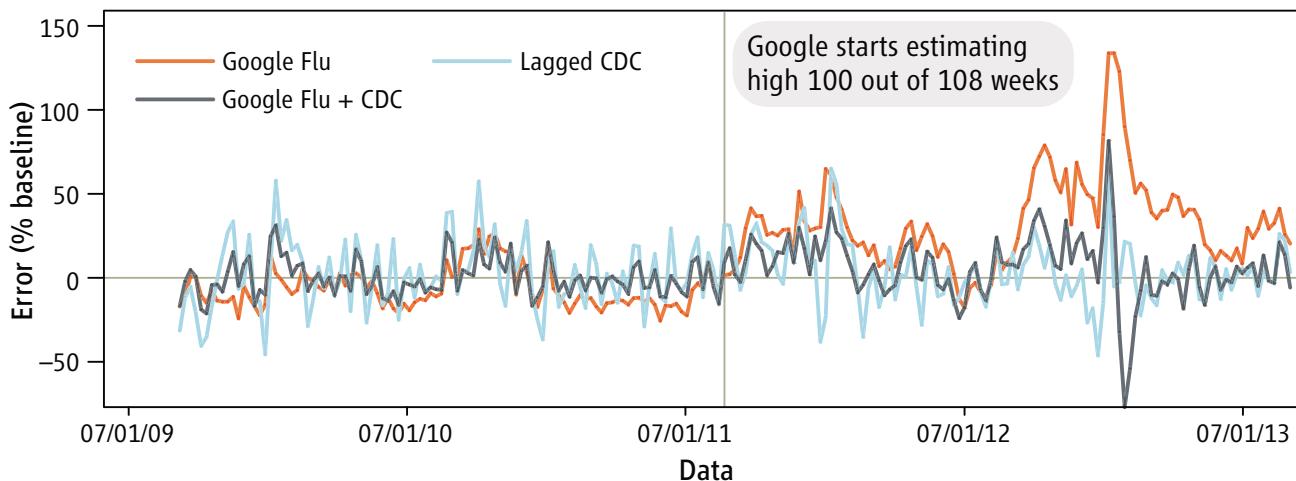
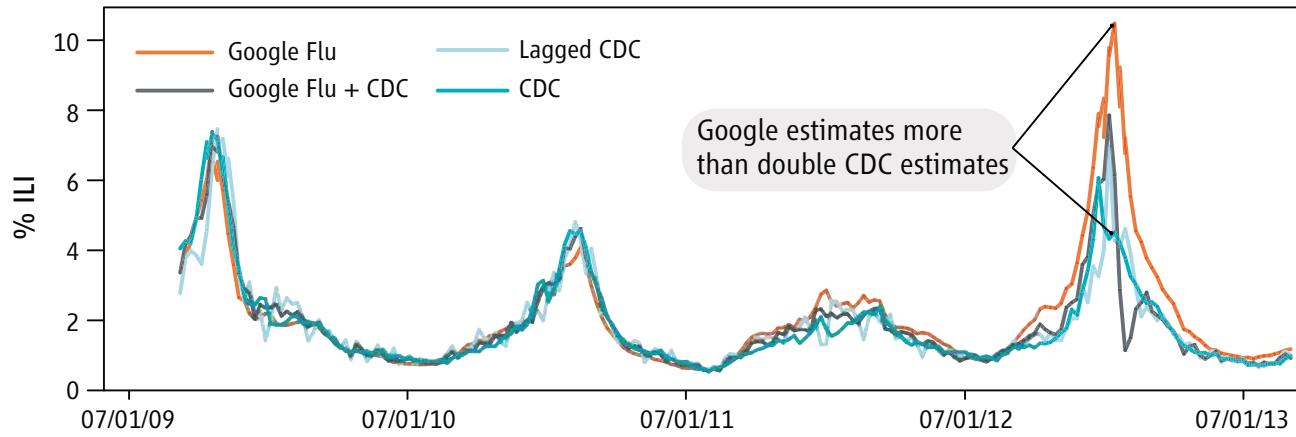
- Real problems often have extremely **high dimensions** (số chiều của dữ liệu quá lớn)
  - Bicycle runs: 2 dimensions (a road)
  - We live: 4 dimensions
  - But an image 1024x1024: **~1 million** dimensions
  - Text collections: **million** dimensions
  - Recommenders' system: **billion** dimensions (items/products)
- The **Curse of dimensionality**



Dữ liệu dù thu thập được  
lớn đến đâu thì cũng là  
**quá nhỏ** so với không  
gian của chúng

# Issues from big data

## ■ Google Flu Trends: Traps in Big Data Analysis



"Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data." [David Lazer et al., 2013]

# Issues from big data

- Ethical issue (đạo đức)
  - breach of privacy, collection of data without informed consent
- Security and privacy (bảo mật và thông tin cá nhân)
  - the ease of stealing, including identity theft, the stealing of national security information
- Issue of exploitation (trục lợi bất hợp pháp)
  - commercial mining of information; targeting for commercial gain
- Issue of Power and politics (quyền lực và chính trị)
  - the use of data to perpetuate particular views, ideologies
- Issue of Truth (sự thật)
  - the perpetuation of falsehoods; propaganda
- Issue of social justice (công bằng trong xã hội)
  - Information is overwhelmingly skewed towards certain groups and leaves others out of the 'digital revolution'

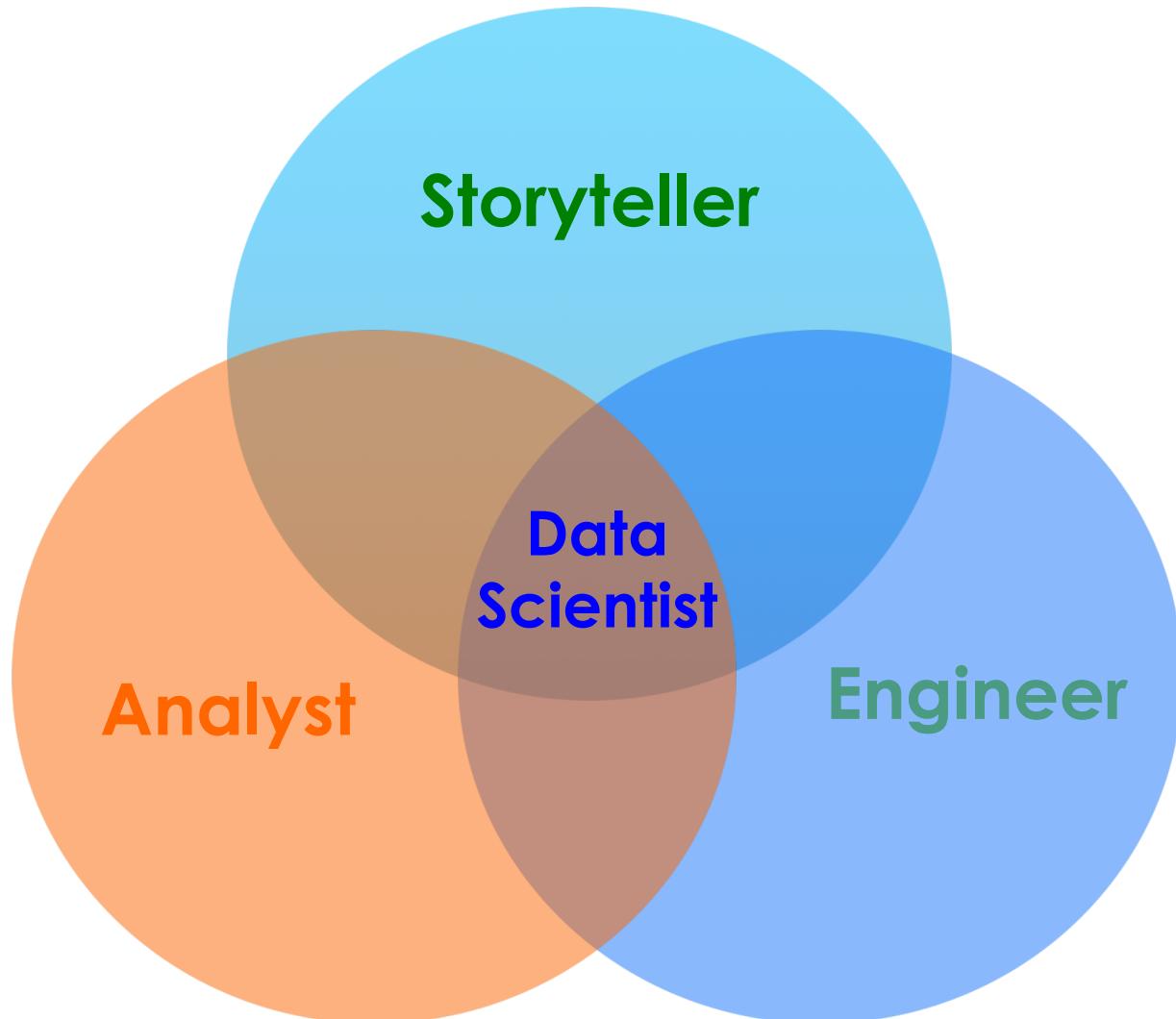
# Skillset



(source: <http://datasciencedojo.com/>)

# Superman

---



## Further reading

---

- “Job Comparison – Data Scientist vs Data Engineer vs Statistician”  
<https://www.analyticsvidhya.com/blog/2015/10/job-comparison-data-scientist-data-engineer-statistician/>
- Big Data Landscape 3.0  
<http://mattturck.com/big-data-landscape-2016-v18-final/>
- Ten Lessons Learned from Building (real-life impactful) Machine Learning Systems  
<http://technocalifornia.blogspot.com/2014/12/ten-lessons-learned-from-building-real.html>

# References

---

- John Dickerson. *Lectures on Introduction to Data Science*. University of Maryland, 2017.
- Longbing Cao. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 43, 2017.
- Longbing Cao. Data science: nature and pitfalls. *IEEE Intelligent Systems*, 31(5), 66-75, 2016.
- David Donoho. "50 years of Data Science." In *Princeton NJ, Tukey Centennial Workshop*. 2015.
- L. Duan, Y. Xiong. Big data analytics and business analytics. *Journal of Management Analytics*, vol 2 (2), pp 1-21, 2015.
- X. Wu, X. Zhu, G. Wu, W. Ding. Data mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering*, vol 26 (1), pp 97-107, 2014.
- Rafael Irizarry & Verena Kaynig-Fittau. *Lectures on Data Science*. Harvard Univ., 2014.
- John Canny. *Lectures on Introduction to Data Science*. University of California, Berkeley, 2014.
- Vasant Dhar. Data Science and Prediction. *Communication of the ACM*, vol 56 (12), pp 64-73, 2013.
- Michael Perrone. *What is Watson – an overview*. 2011.