Introduction to Data Science

(Khoa học dữ liệu)

Khoat Than

Hanoi University of Science and Technology

khoattq@soict.hust.edu.vn

IT4142E, SOICT, HUST, 2019

Contents

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 hưở
  - “Báo Đả 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)

Accuracy ~93%.

(http://yourstory.com/2014/07/germany-argentina-fifa-world-cup-2014/)

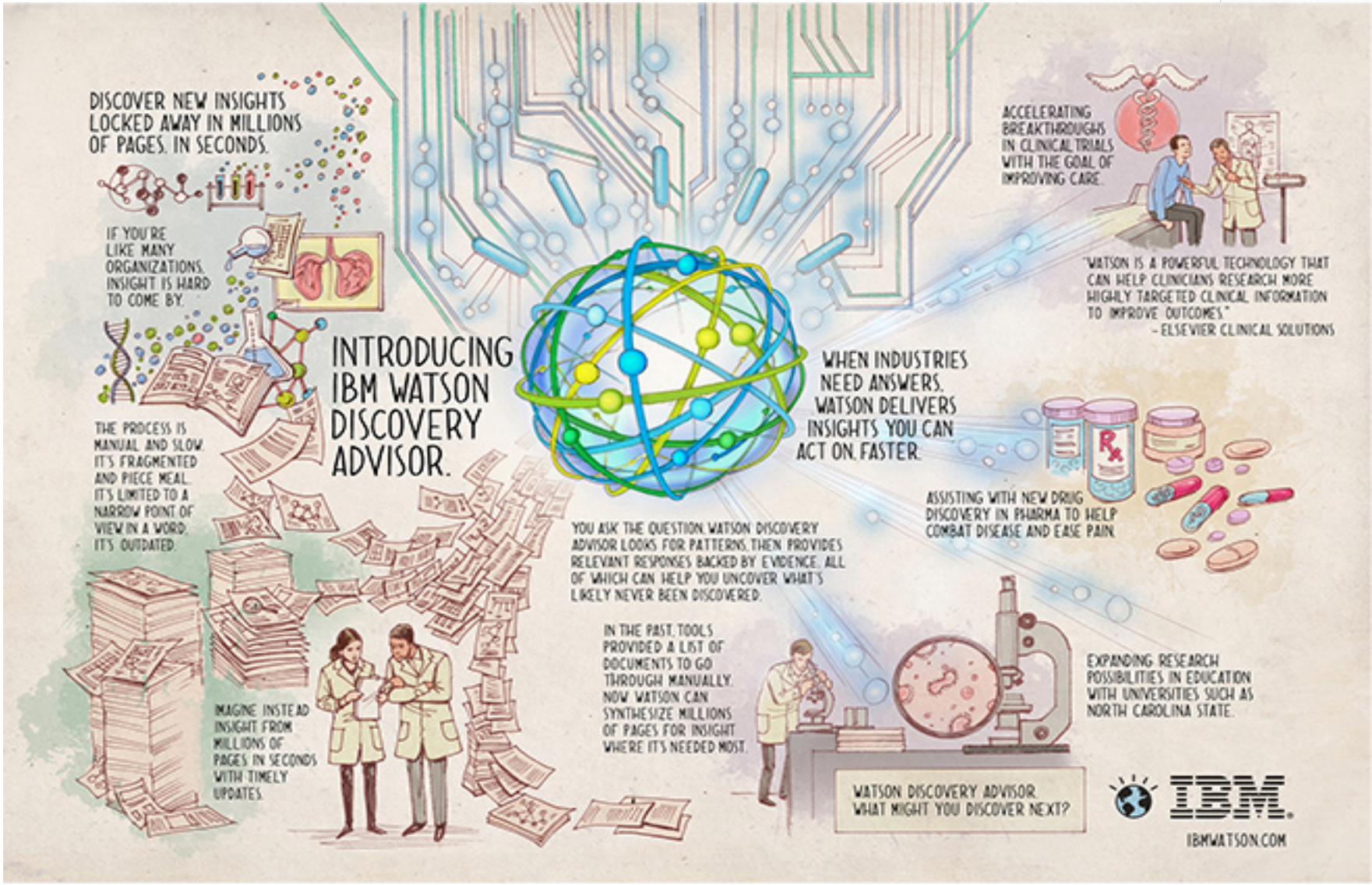
Amazon's recommendation secret

"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

Some examples: IBM's Watson

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

IBM's massive bet on Data Analytics



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

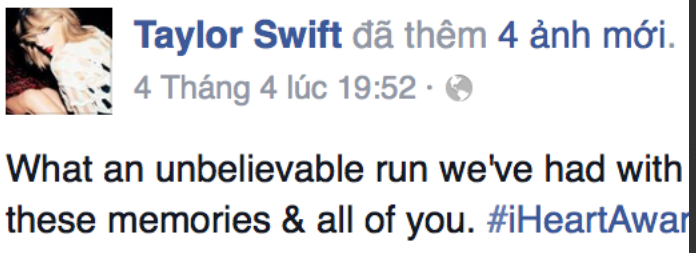
21st century

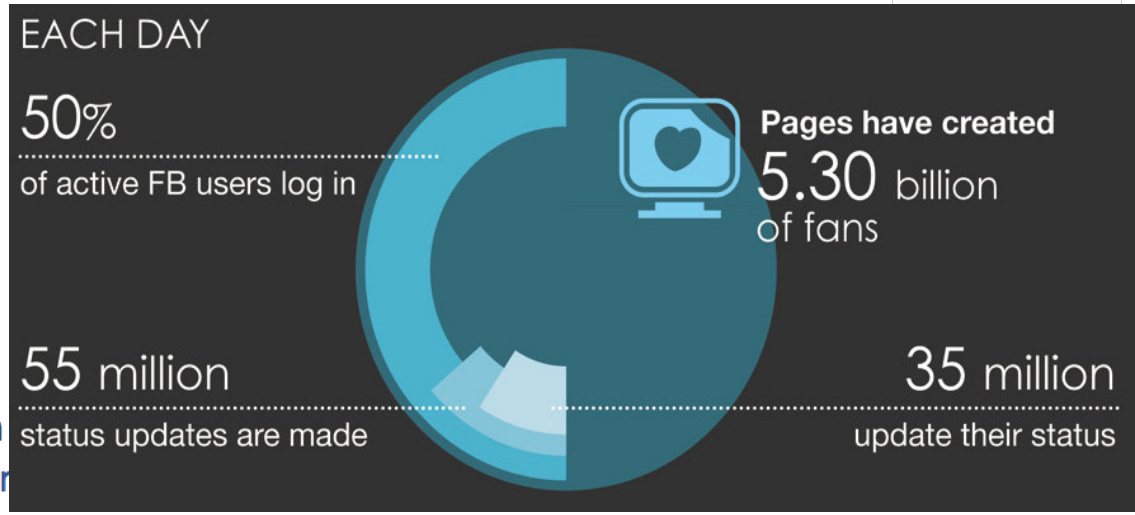
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

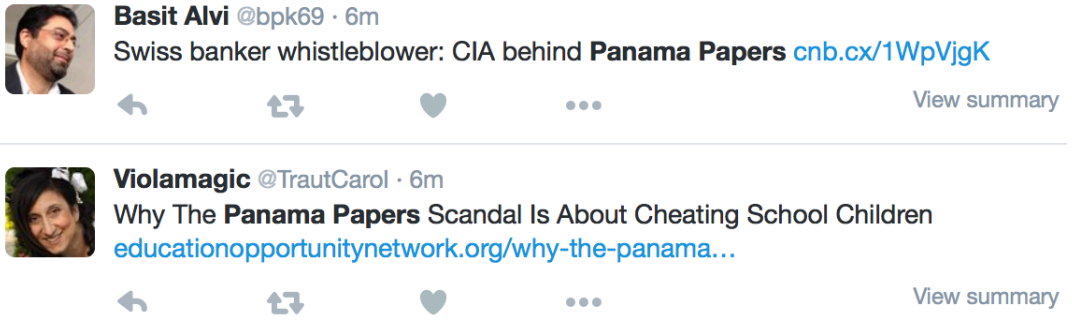
Where the data? Social networks

facebook





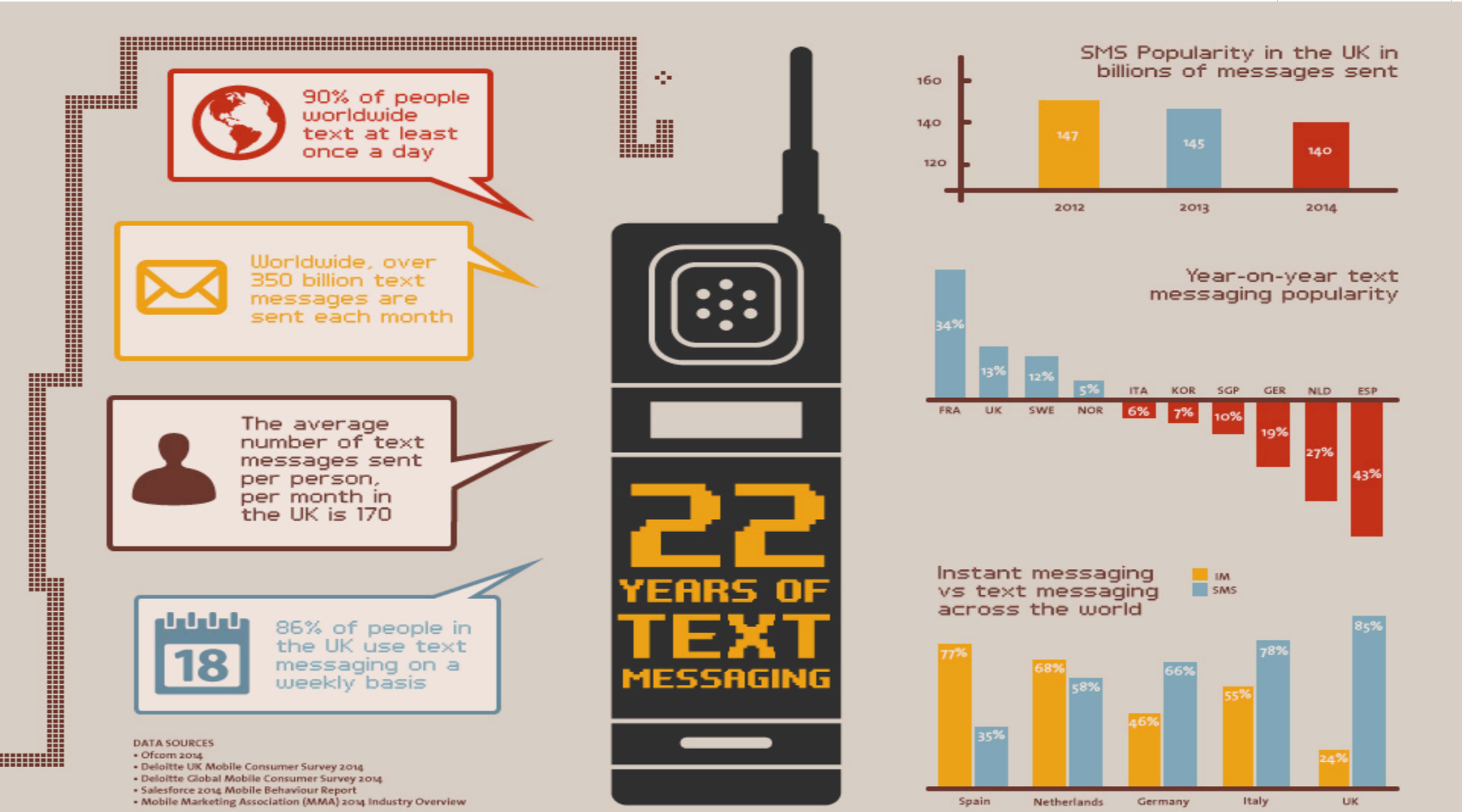
twitter



7,174 Tweets sent in 1 second

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

Where the data? Mobile messages



(http://www.openuniversity.edu/news/news/2014-text-messaging-usage-statistics)

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

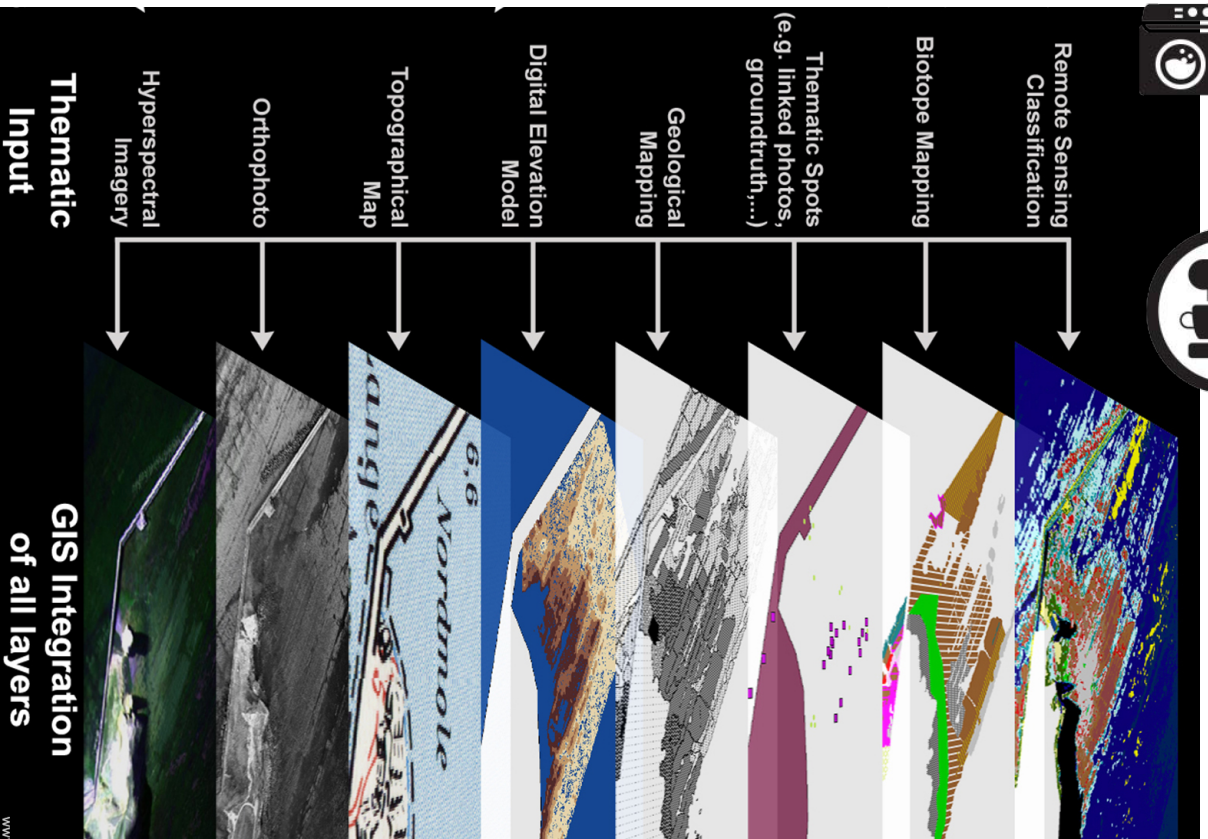
Blog posts written today

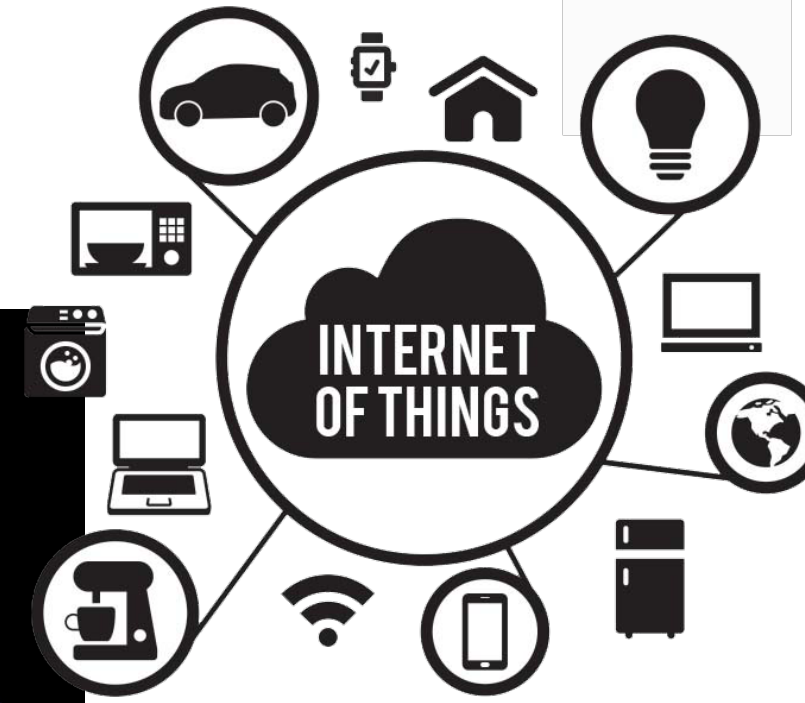
444,323,737

Tweets sent today

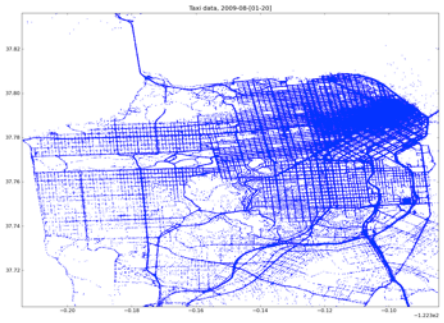
(http://www.internetlivestats.com)

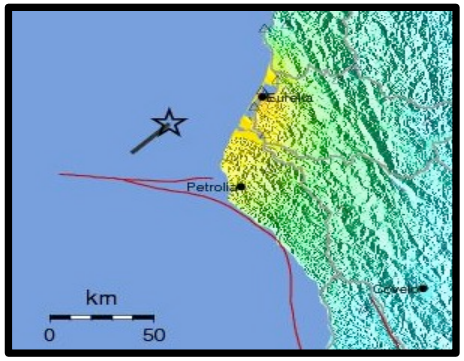
Where the data? And more

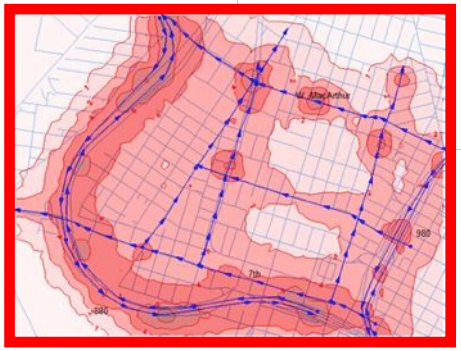




What can we do with the data?

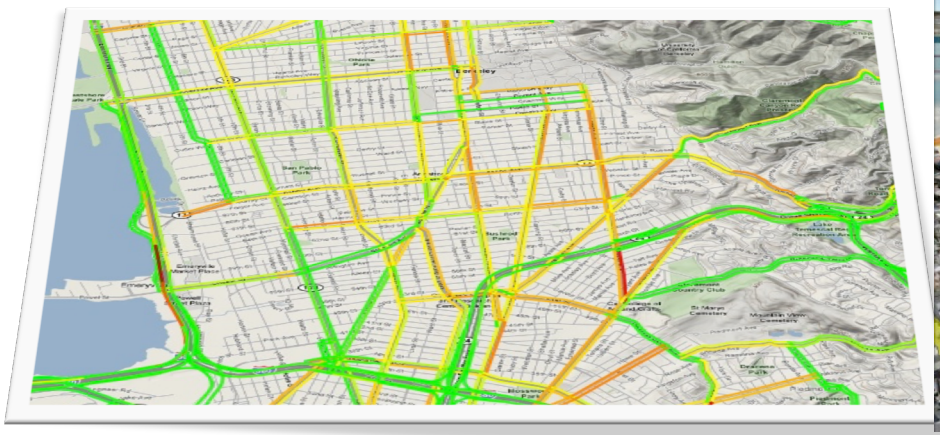






Crowdsourcing + physical modeling + sensing + data assimilation

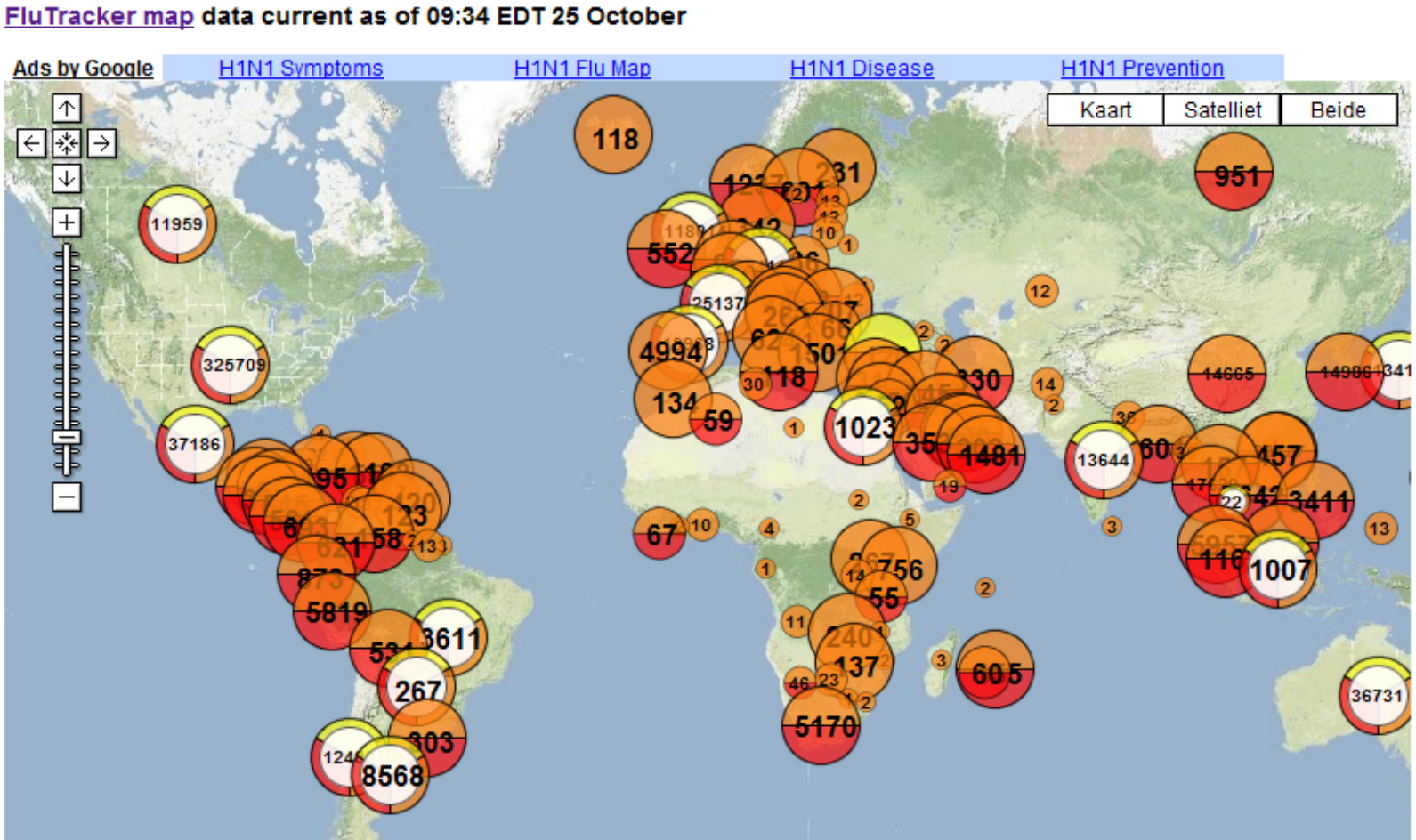
to produce:



(Alex Bayen, UC Berkeley)

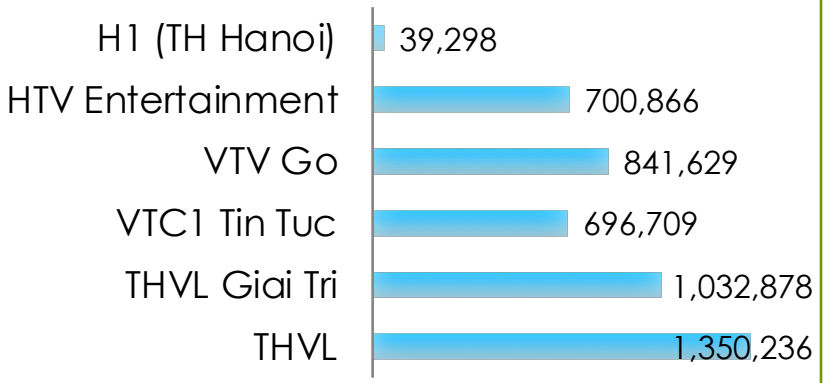
What can do? Prediction

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

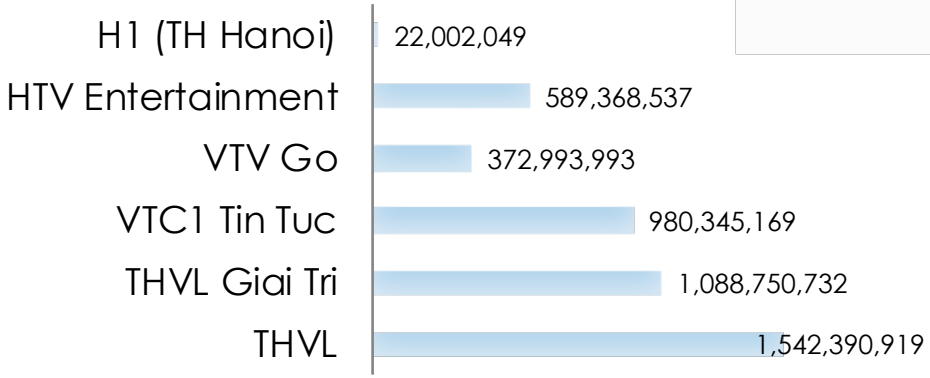


What can do? Exploration

Subscribers in Youtube

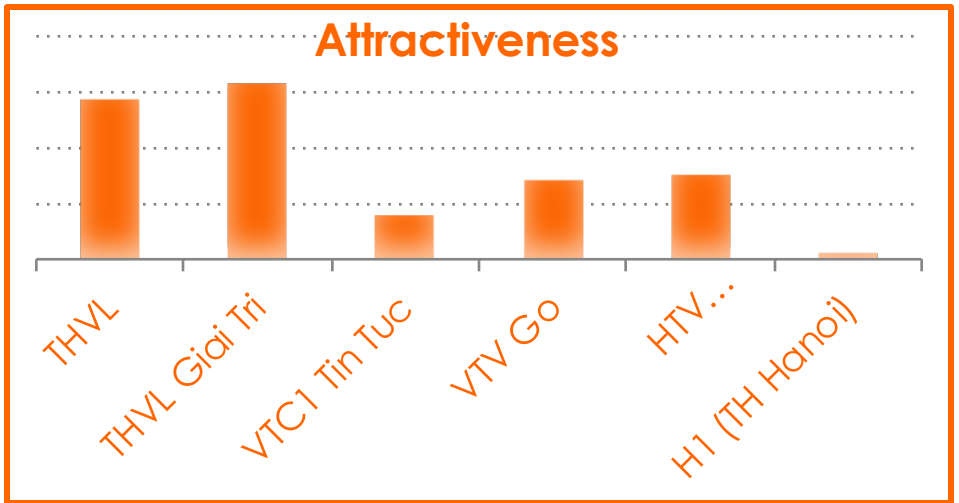


Views in Youtube



Effective TV channels?

(July 4, 2017)



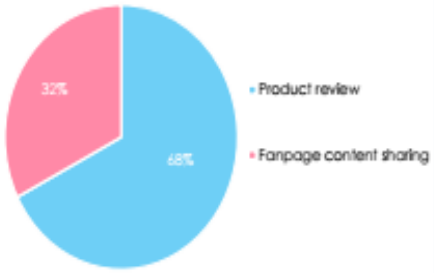
What can do? Planning

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

DEEP DIVE INTO EACH ACTIVITY

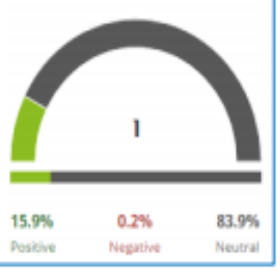
MASS SEEDING 2K buzz | 2K likes | 13 shares

% CONTRIBUTION OF KEY TACTICS



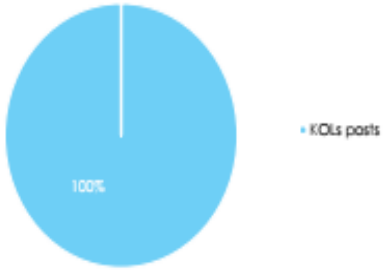
AUDIENCE FEEDBACK

Khách hàng muốn dùng thử sản phẩm



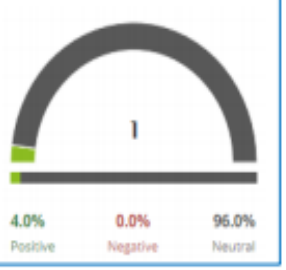
GIVEAWAY 610 buzz | 4K likes | 105 shares

% CONTRIBUTION OF KEY TACTICS



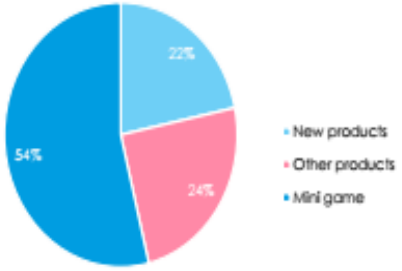
AUDIENCE FEEDBACK

Người tham gia tag bạn bè mình cùng tham gia



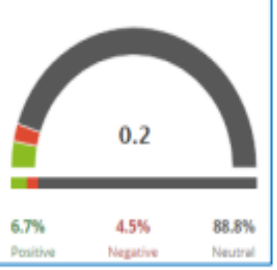
PRODUCTS 1,2K buzz | 20K likes |297shares

% CONTRIBUTION OF KEY TACTICS



AUDIENCE FEEDBACK

Sản phẩm cho thấy tác dụng nhanh, không tác dụng phụ



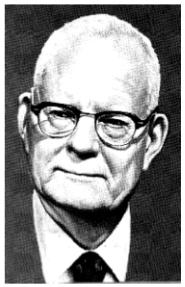
Data Science

Data Analysis has been around

1935: "The Design of Experiments"

R.A. Fisher 

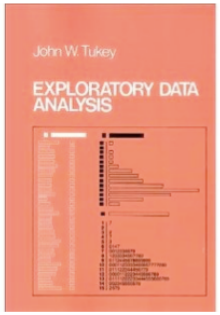
1939: "Quality Control"

W.E. Demming 

1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"



1989: "Business Intelligence"

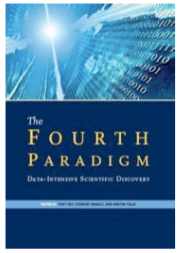
Howard Dresner 

1997: "Machine Learning"

2010:"The Data Deluge"

2009:"The Unreasonable Effectiveness of Data"

2007:"The Fourth Paradigm"



1996: Google

Google™

(John Canny, UC Berkeley)

Online platforms

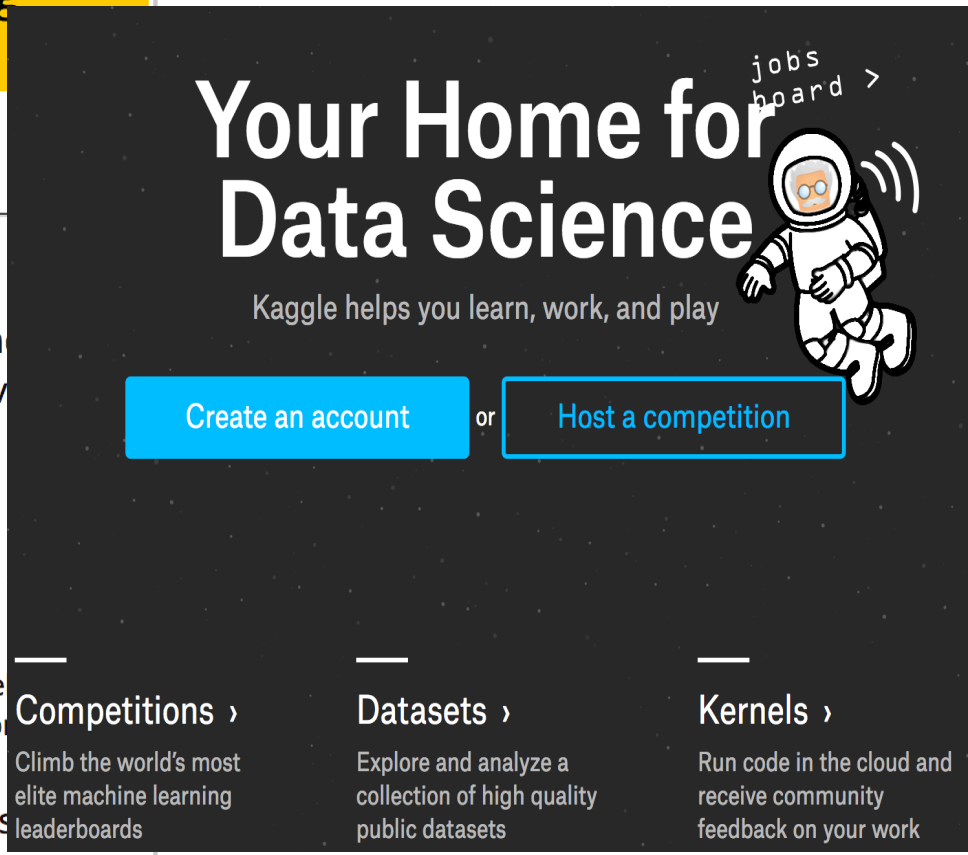
KDnuggets

Analytics, Data Science, Data Mining Competitions

Notable Recent Competitions

- GE NFL $10 Million Head Health Challenge, for more a
  diagnoses of mild brain injury and prognosis for recov
  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 de
  real-time flight profile to the pilot, helping them make flights mo
  efficient and reliably on time. Prize Pool: $250,000
- Heritage Health Data Analysis Prize ($3M), can adminis

kaggle™



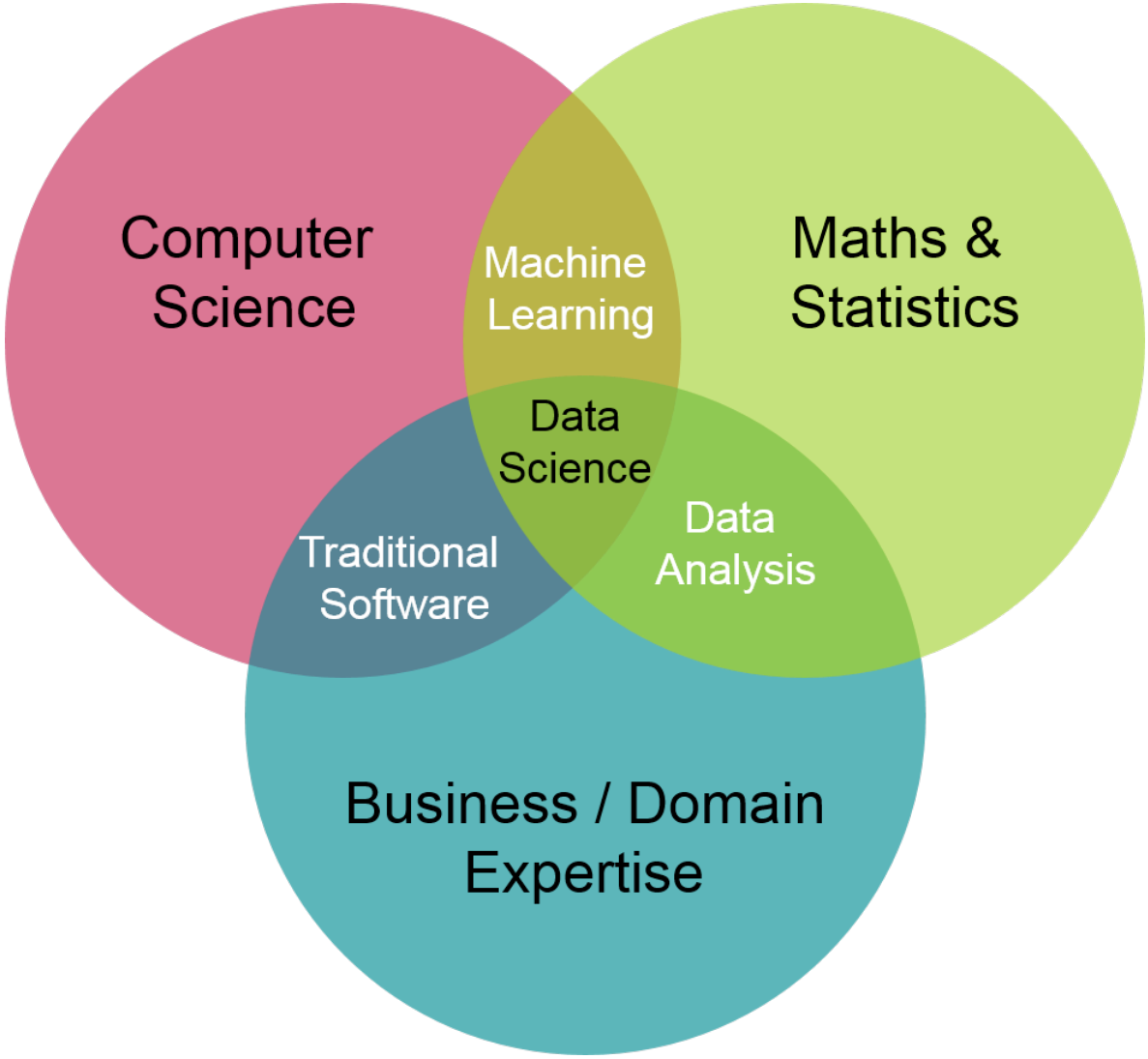
What is it?

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?


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise

Why is Data Science important?

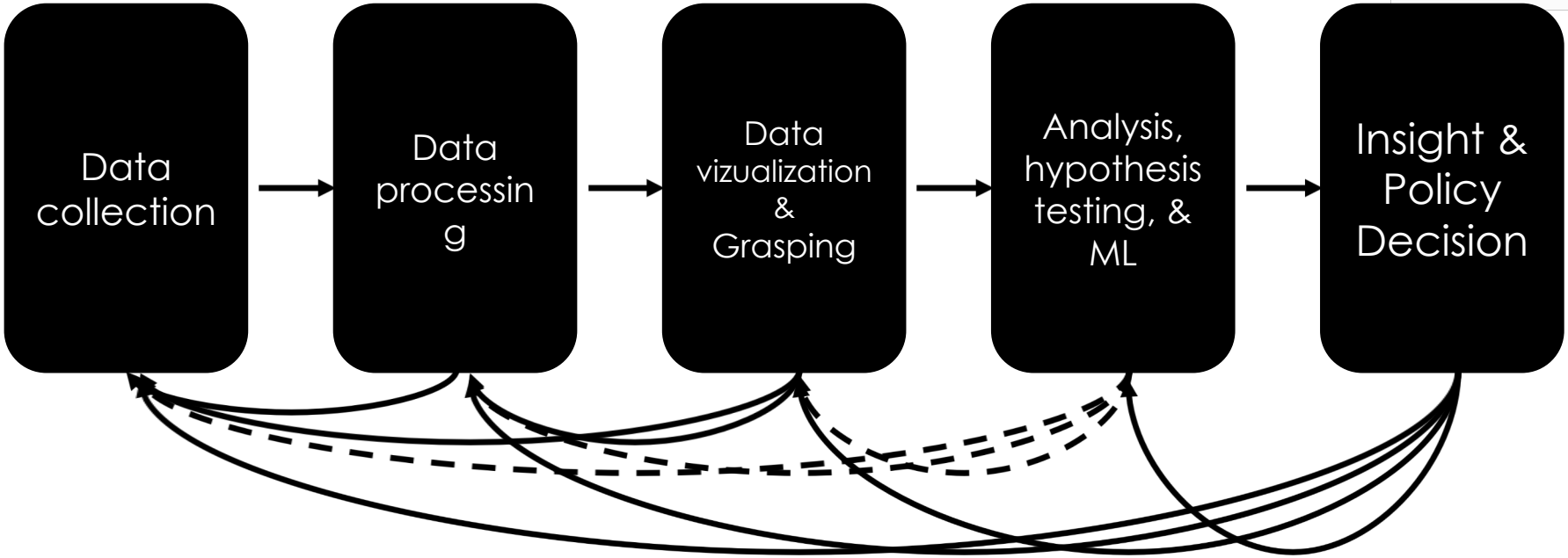
"Khả năng biết lấy dữ liệu, để **hiểu**, để **xử lý**, để **tạo ra giá trị** từ chúng, để **trao đổi** những giá trị đó với người khác, là một trong những kỹ năng cực kỳ quan trọng trong những thập kỷ tới. Những kỹ năng đó cần thiết ngay cả đối với các học sinh phổ thông. **Lý do là vì chúng ta thực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

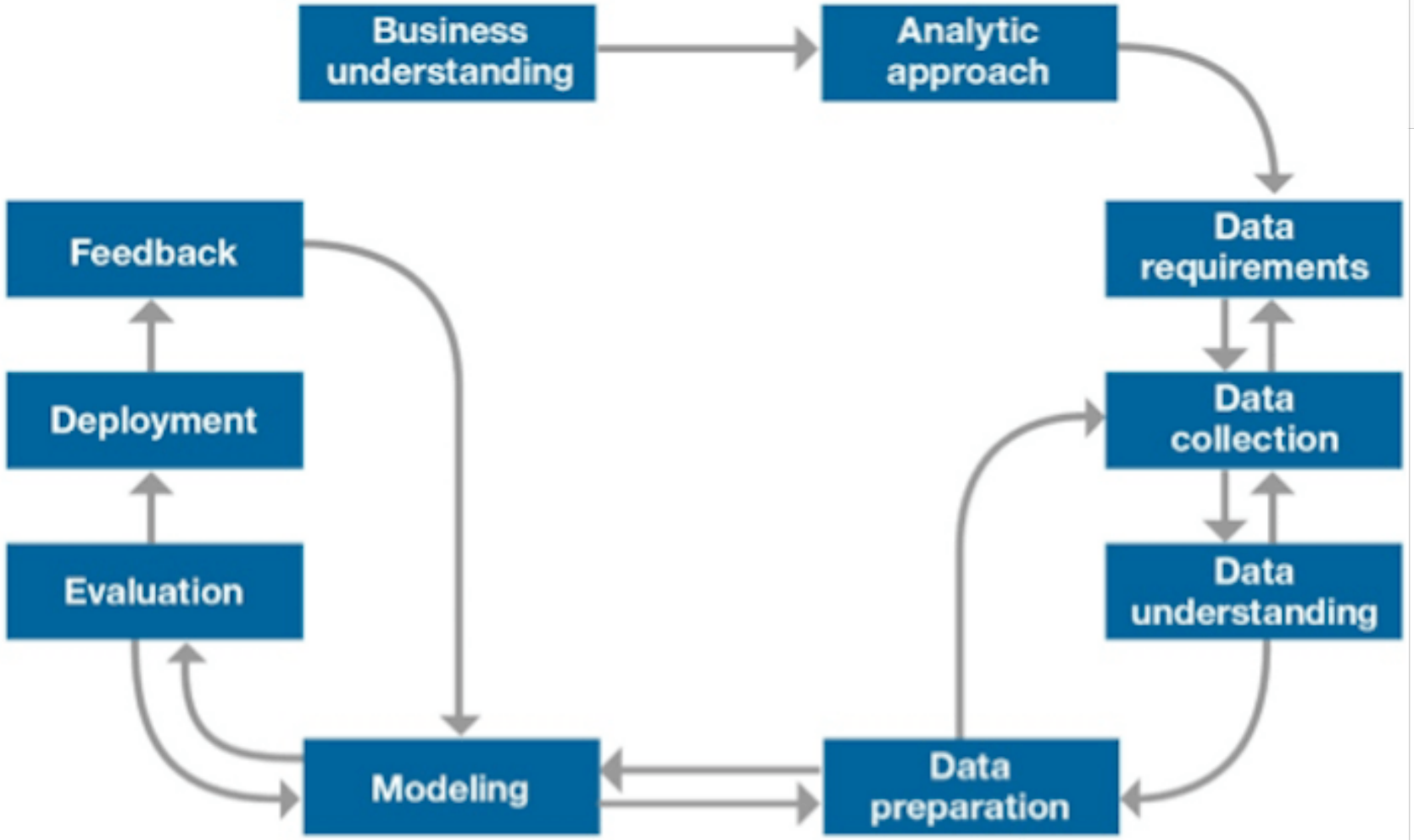
- Hal Varian, Google's Chief Economist

DS methodology: insight-driven



(John Dickerson, University of Maryland)

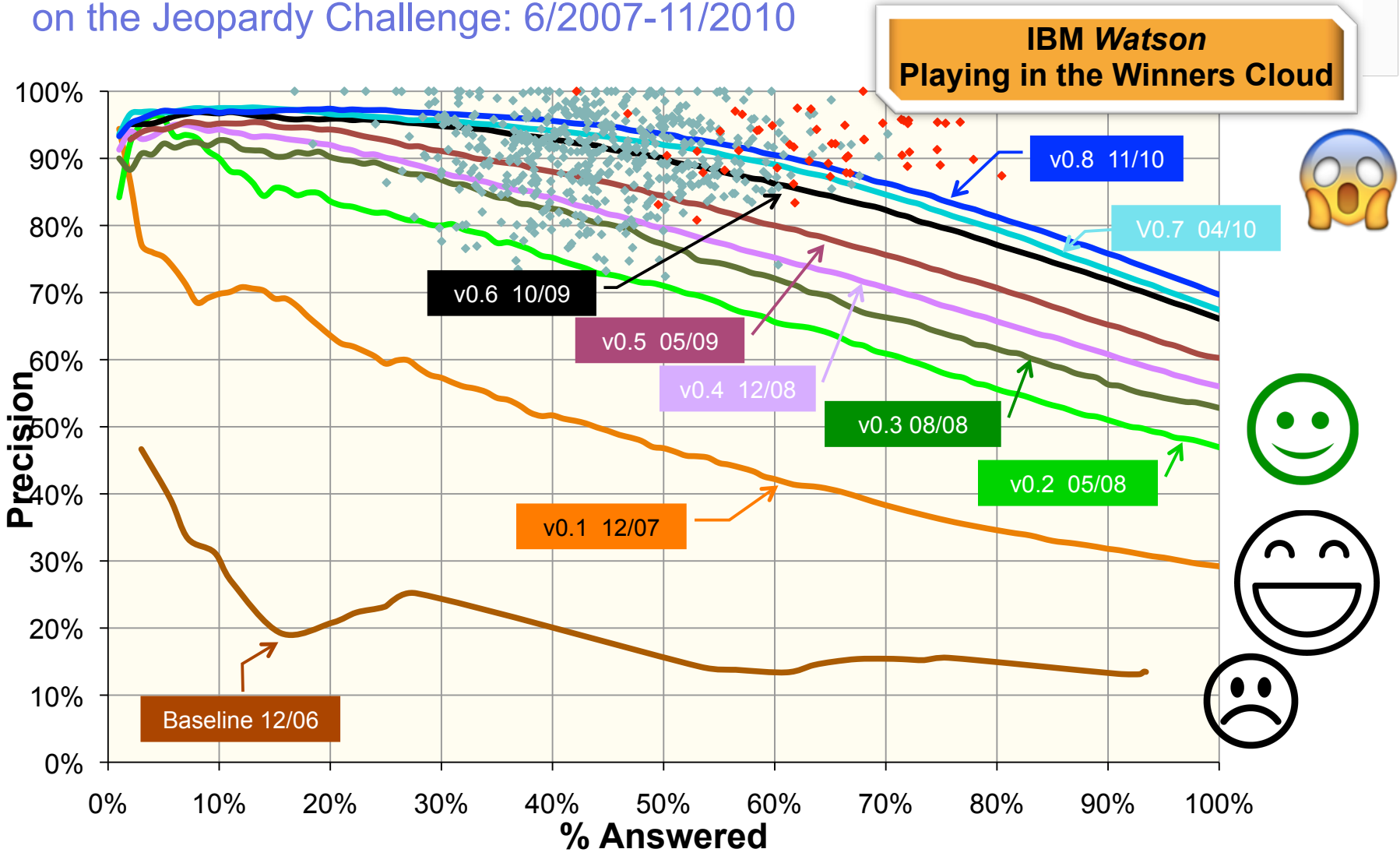
DS methodology: product-driven



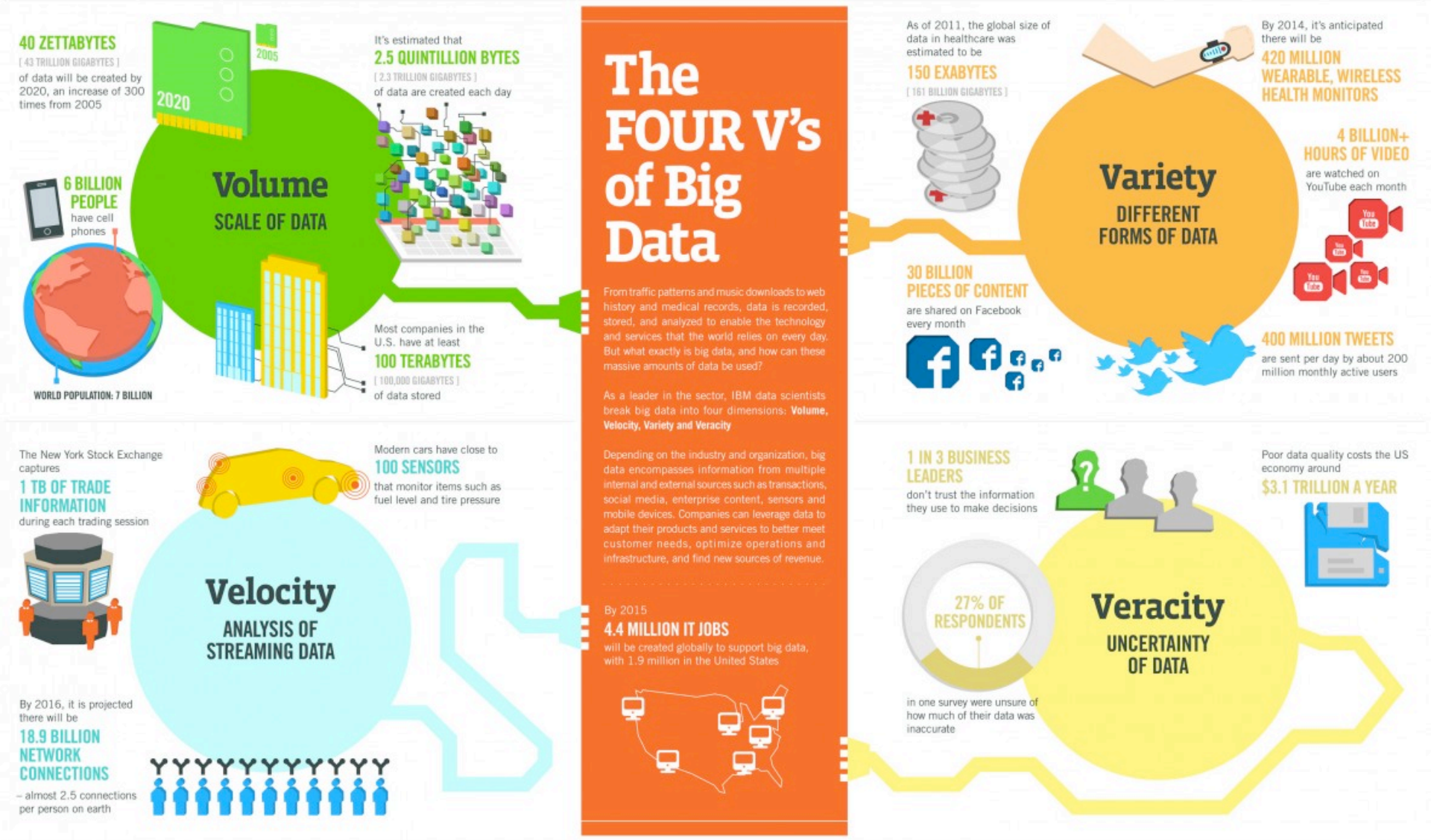
(http://www.theta.co.nz/)

Product development: experience

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



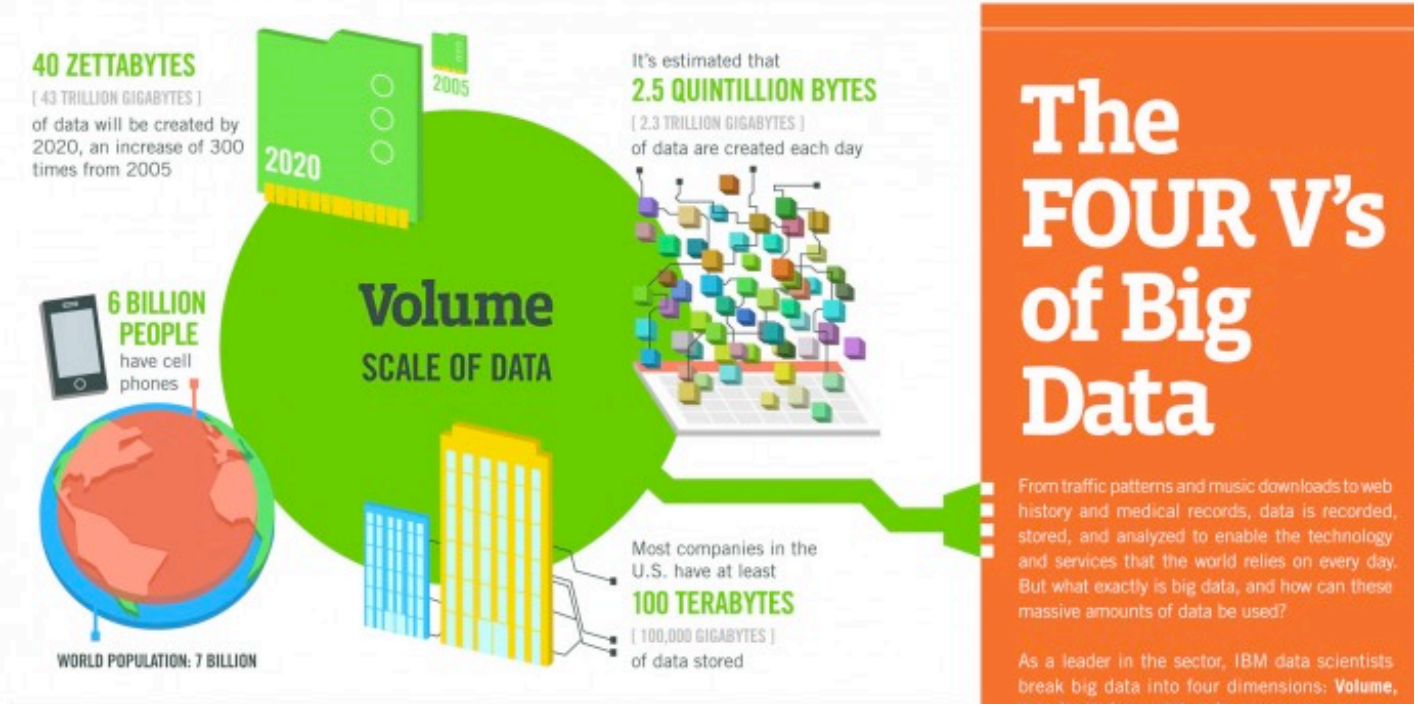
Challenges



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

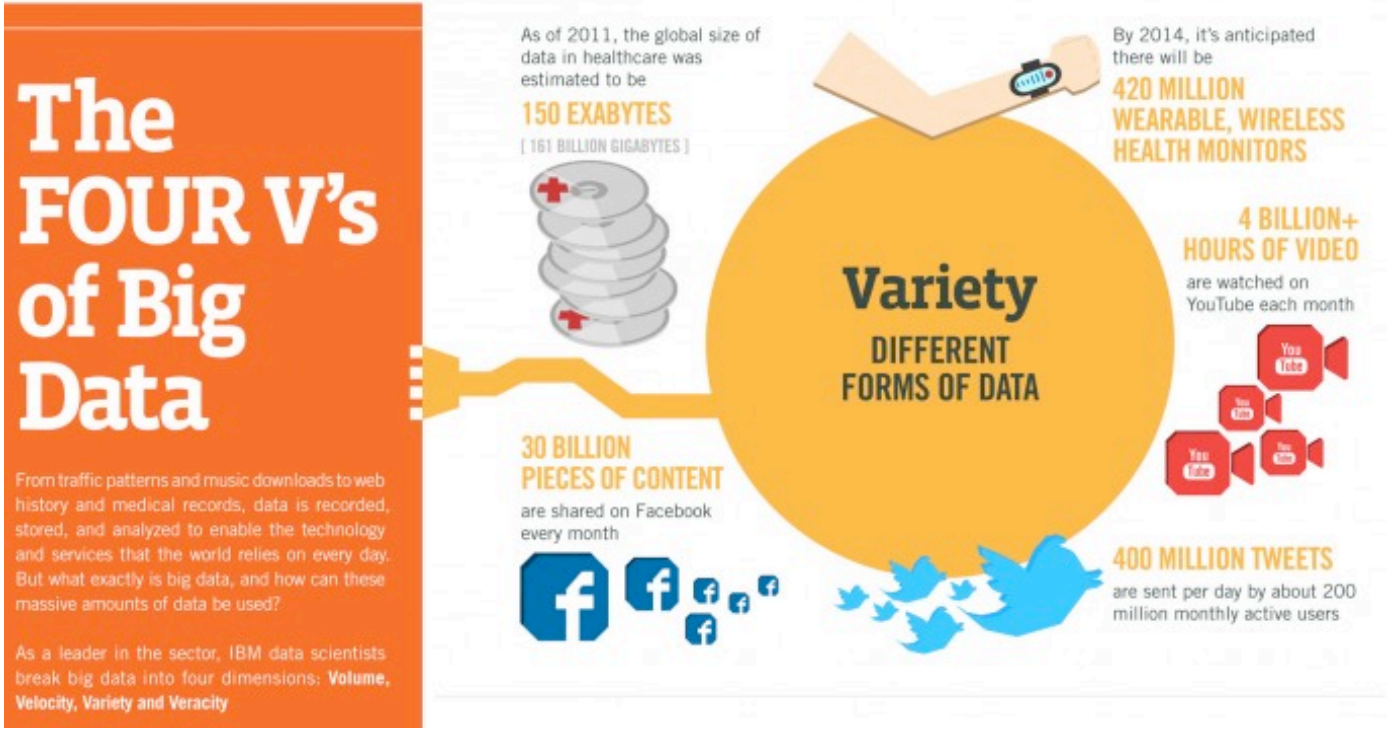
IBM

Challenges: volumn



Số lượng dữ liệu thu thập được **quá nhiều**

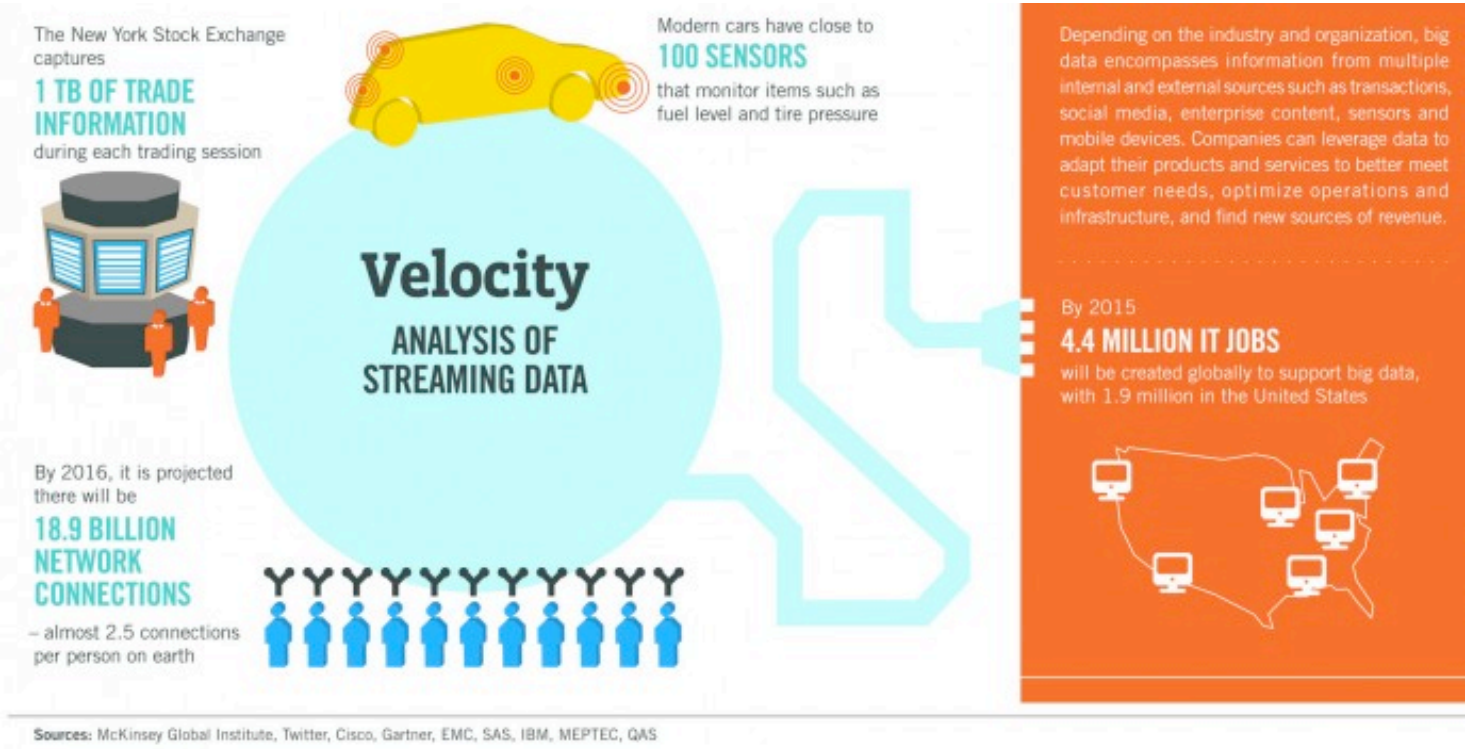
Challenges: variety



Dữ liệu
quá **đa dạng**

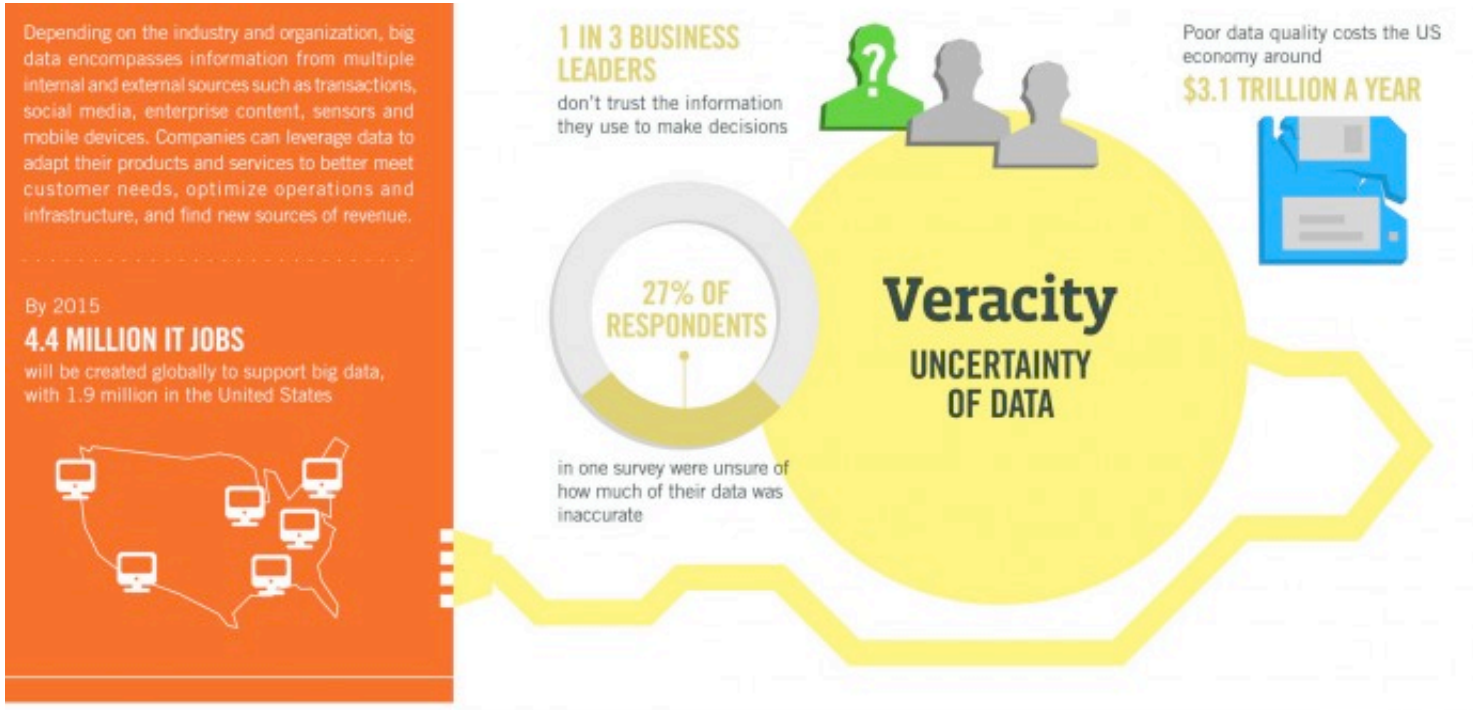
Challenges: velocity

Dữ liệu đến **liên tục** và **nhanh**



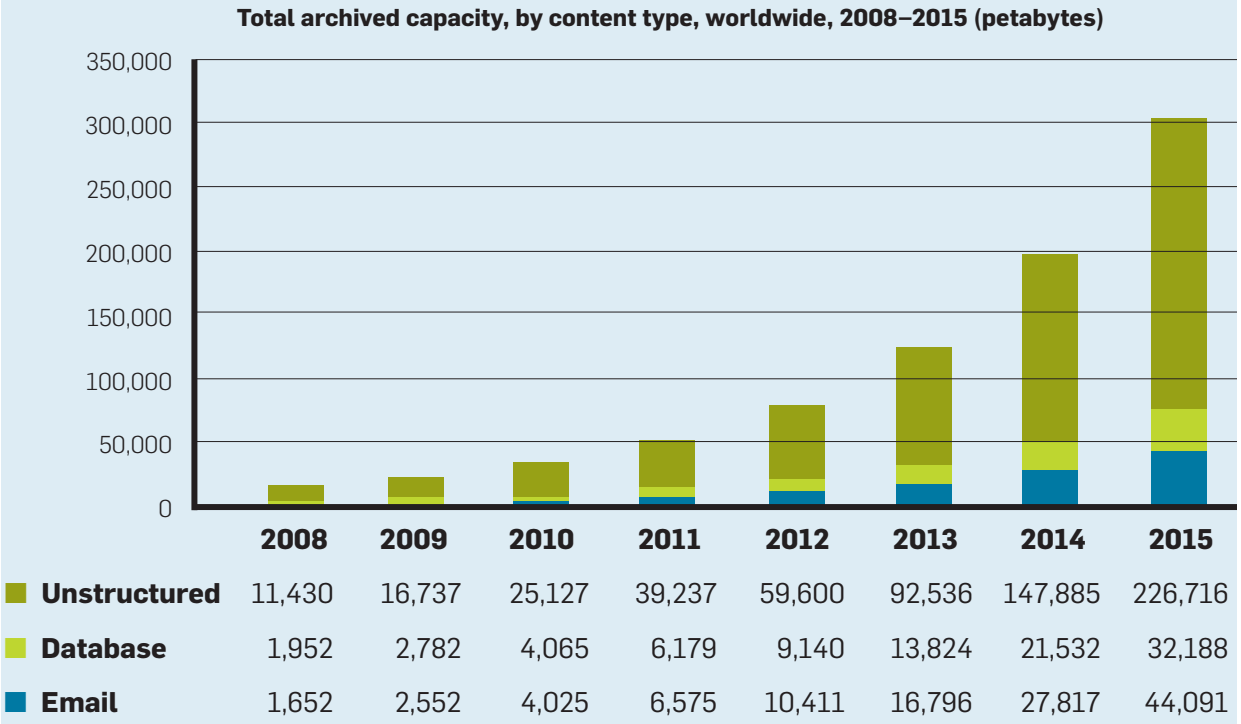
Challenges: veracity

Dữ liệu
hỏng, nhiễu, lỗi
quá nhiều



Challenges: free structure

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

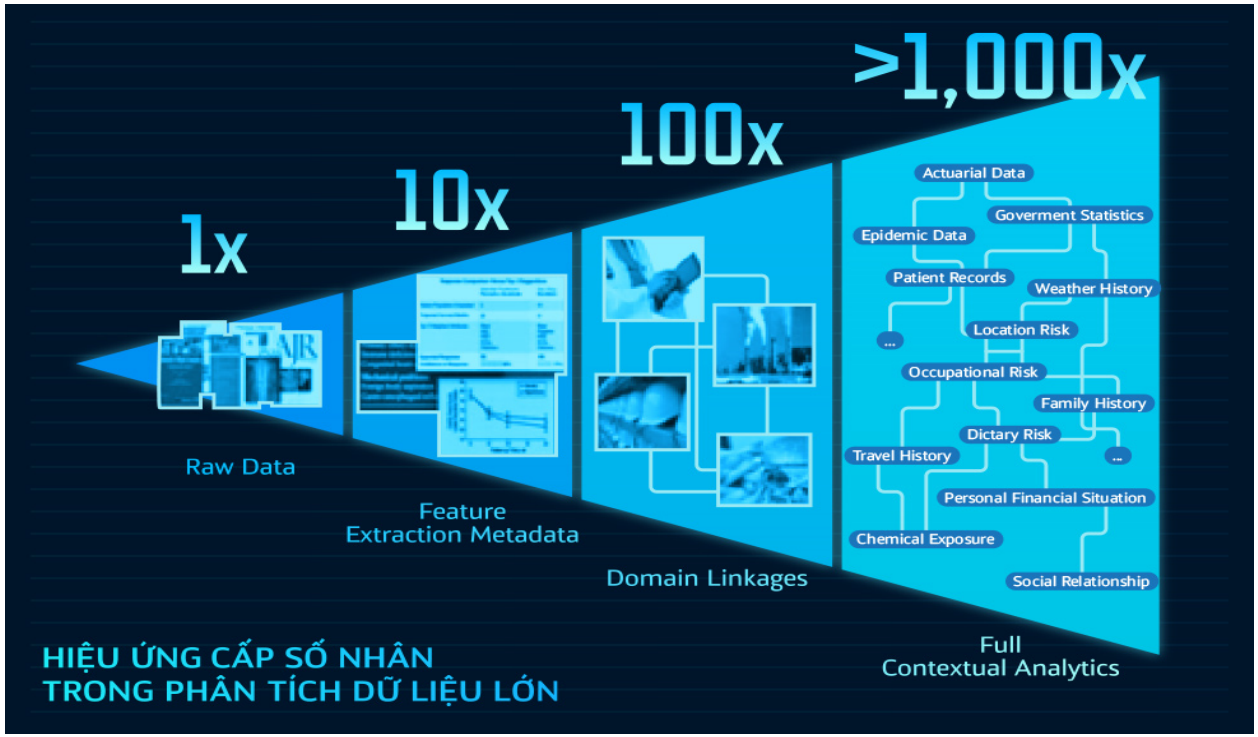


	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

(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)



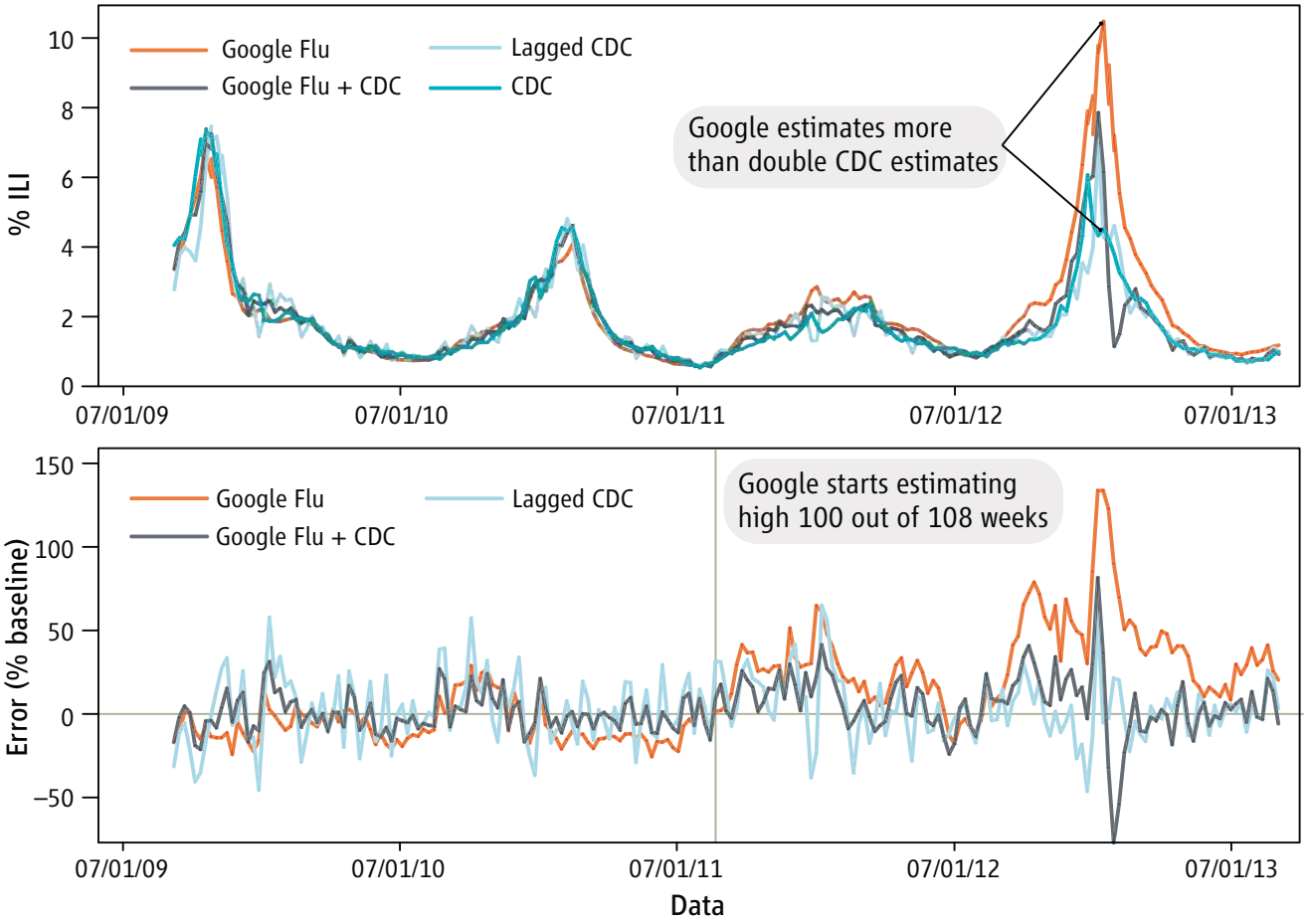
(http://genk.vn/ibm-va-cuoc-cach-mang-tri-tue-nhan-tao-mang-ten-watson-20160830152953753.chn)

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

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

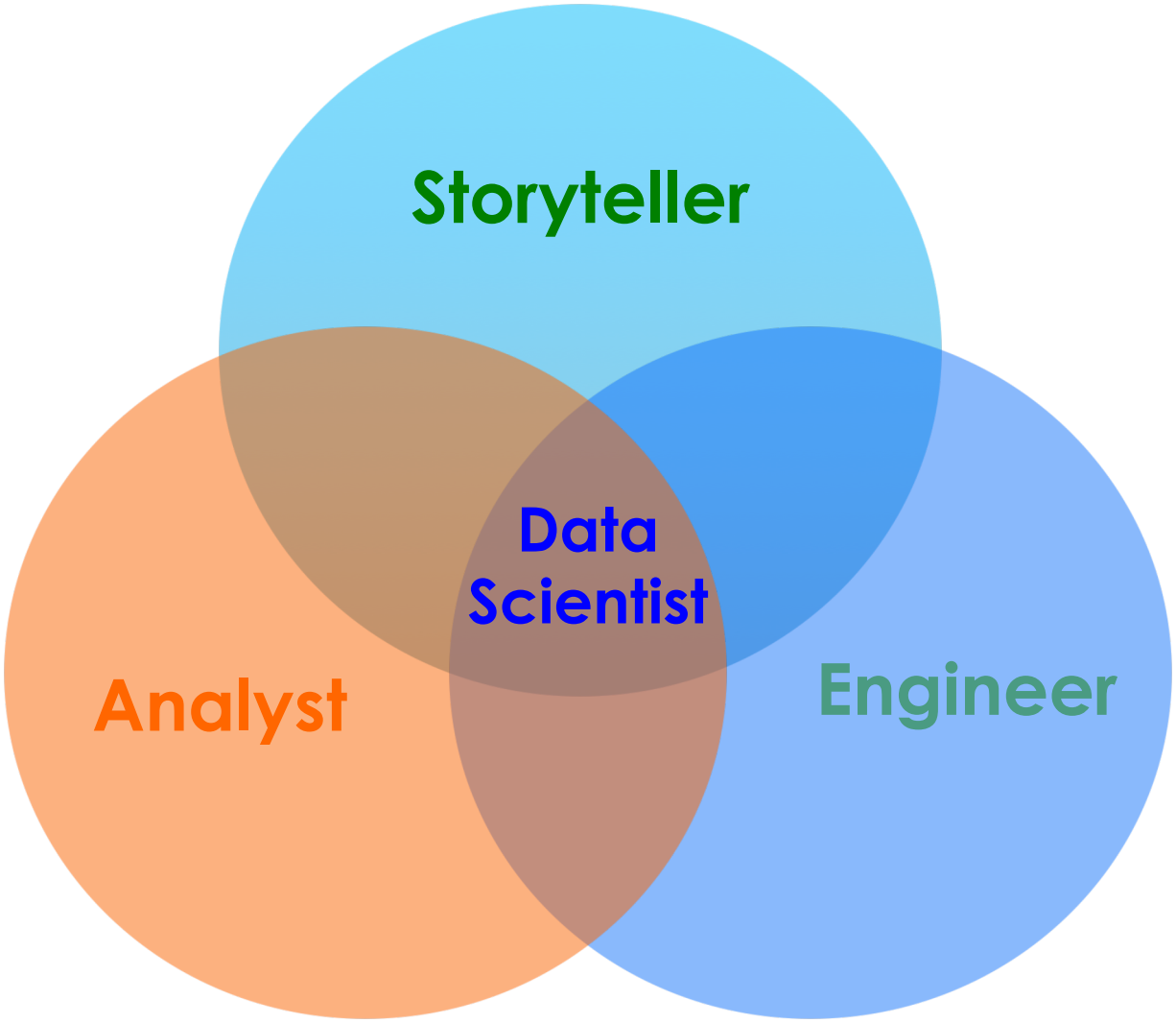
- Ethical issue (đạo đức)
  - breach of privacy, collection of data without informed consent
- Security and privacy (bảo mật và thông tin cá nhân)
  - the ease of stealing, including identity theft, the stealing of national security information
- Issue of exploitation (trục lợi bất hợp pháp)
  - commercial 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'

(Dinh Phung, Monash University)

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.