Outline


science
value
Hadoop
petabytes
computers
huge
visualization
unstructured
database
Facebook
images
private
Statistics
Association
memory
storage
Google
online
Prediction
videos
Velocity
fast
access
Variety
emails
Veracity
Petabytes
challenges
big
data
Volume
modeling
transactions
zettabytes
machines
sensors
search
Twitter
security
Classification
analyze
photos
open
datamining
analytics
complex
terabytes
sensor
exabytes
amount
mining
vast
Science
enterprise
dimensions
privacy

What is big data?

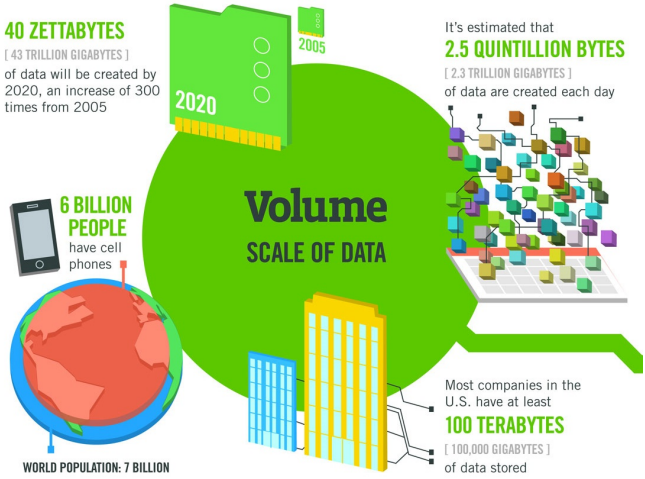
the 3 V's of big data

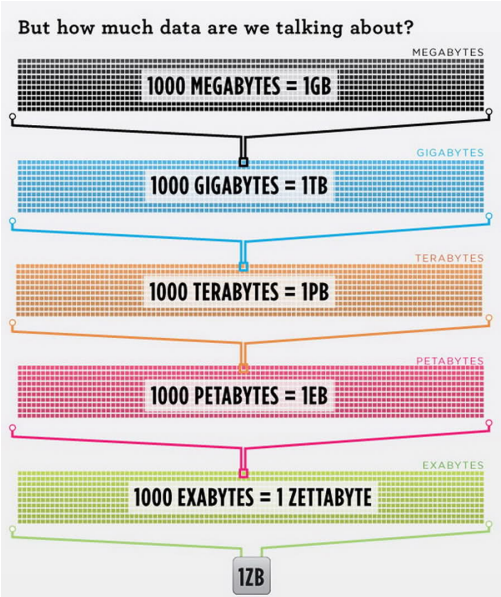
Volume, Velocity, Variety

What is big data?

The 3 V's of big data

- Volume
- Velocity
- Variety



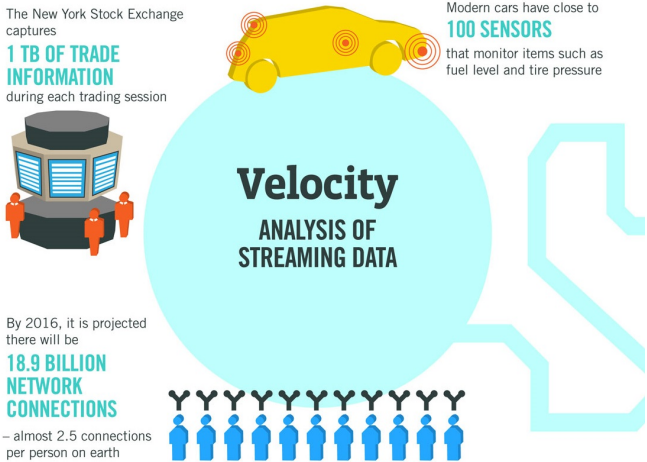

But how much data are we talking about?
MEGABYTES
1000 MEGABYTES = 1GB
GIGABYTES
1000 GIGABYTES = 1TB
TERABYTES
1000 TERABYTES = 1PB
PETABYTES
1000 PETABYTES = 1EB
EXABYTES
1000 EXABYTES = 1 ZETTABYTE
1ZB


EXABYTE
(1,152,921,504,606,846,976 BYTES; 2 60)
approx. 1,000,000,000,000,000,000 or 10 18
5 EXABYTES: ALL WORDS EVER SPOKEN BY HUMAN BEINGS
ZETTABYTE
(1,180,591,620,717,411,303,424 BYTES; 2 70)
approx. 1,000,000,000,000,000,000,000 or 10 21

What is big data?

The 3 V's of big data

- Volume
- Velocity
- Variety



Batch vs. Real-Time Processing

Batch Processing

Occurs at a scheduled time when a sufficient amount of data has been accumulated and computational load will not impact other processes.

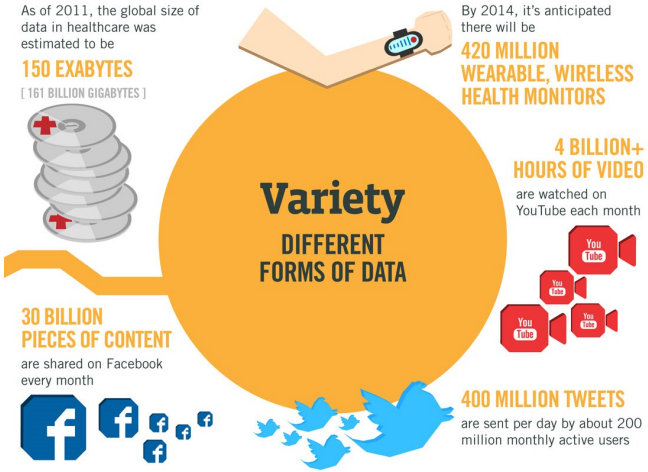
Real-time Processing

Occurs when the data is generated.

What is big data?

The 3 V's of big data

- Volume
- Velocity
- Variety



Structured data

highly organized information

easily “fits” into a relational database in rows and columns

searchable and well-defined

Unstructured data

may have an internal structure, but does not fit a relational data model

includes textual and multimedia content

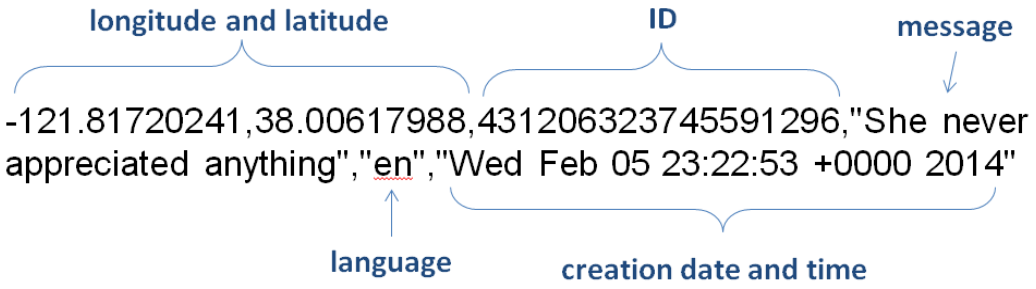
e.g.

- email, instant messages, tweets, chat
- business reports and presentations
- web pages, blogs, wikis, photos, videos
- satellite imagery

Twitter feed excerpt

-121.81720241,38.00617988,431206323745591296,”She never appreciated anything”,”en”,”Wed Feb 05 23:22:53 +0000 2014”
-80.39720584,25.68434391,431206327940304896,”I dont really care if I dont please you”,”en”,”Wed Feb 05 23:22:54 +0000 2014”
-78.82004388,42.74178633,431206382470443008,”maybe you little brat”,”en”,”Wed Feb 05 23:23:07 +0000 2014”
-120.06433452,36.72913789,431206390833500160,”lly lt”,”en”,”Wed Feb 05 23:23:09 +0000 2014”
-75.37382296,40.62207079,431206411805413376,”All the pictures on my phone consist of ugly pictures of me and really attractive famous people”,”en”,”Wed Feb 05 23:23:14 +0000 2014”

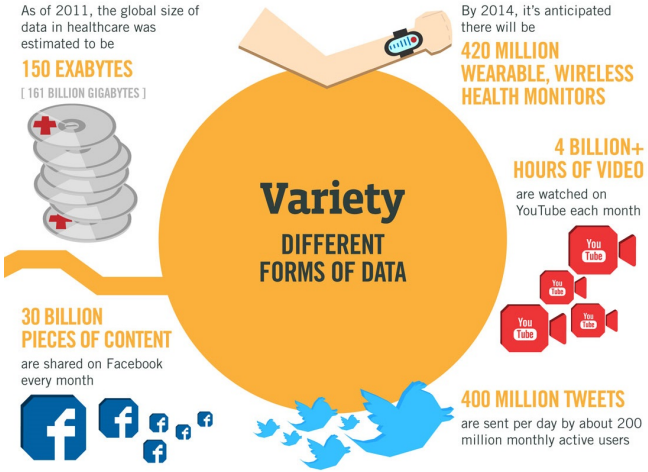
Twitter feed excerpt



What is big data?

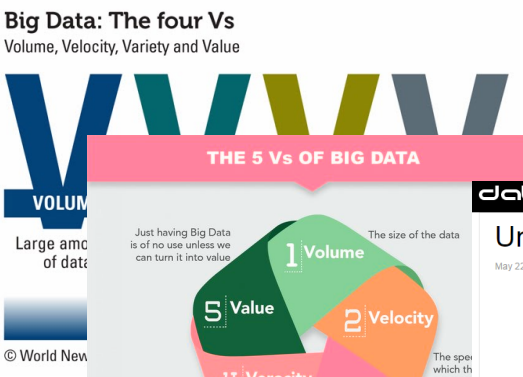
The 3 V's of big data

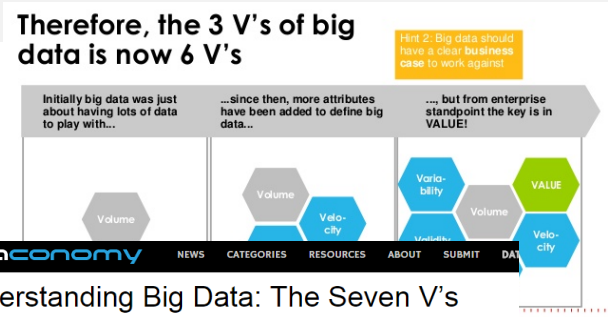
- Volume
- Velocity
- Variety



What is big data?

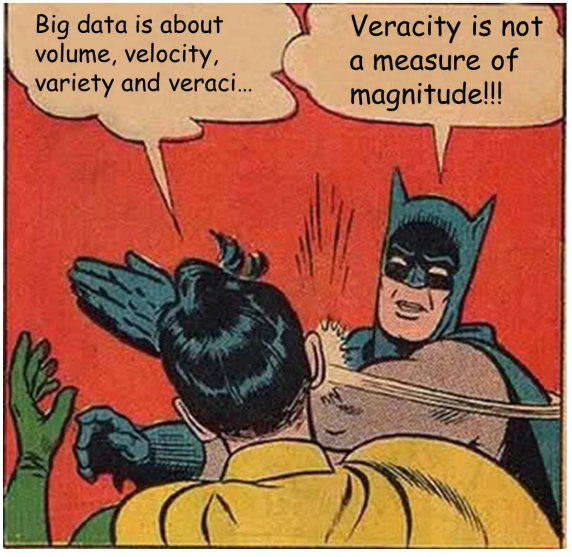
even more V's relating to (big) data...





Understanding Big Data: The Seven V's

May 22, 2014 Written by: Eileen McNulty 4 Comments


Big data is about volume, velocity, variety and veraci…
Veracity is not a measure of magnitude!!!

What is big data?

even more V's relating to (big) data...

- Veracity
- Variability
- Visualization
- Value

IBM found that:

- 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

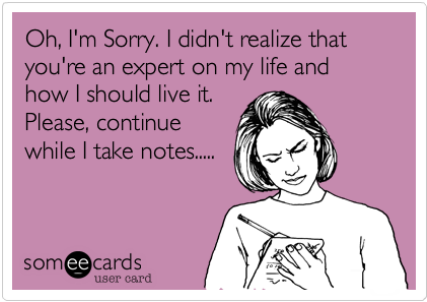
What is big data?

even more V's relating to (big) data...

- Veracity
- Variability
- Visualization
- Value

→ the *meaning* of the data is changing

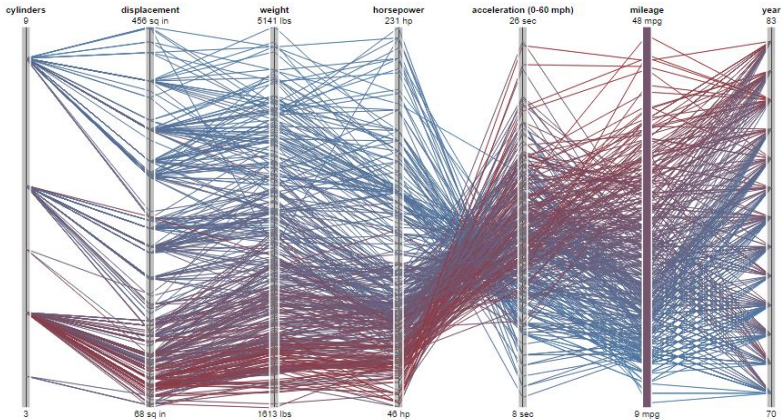
e.g., sarcasm in natural language processing...



What is big data?

even more V's relating to (big) data...

- Veracity
- Variability
- Visualization
- Value



What is big data?

even more V's relating to (big) data...

- Veracity
- Variability
- Visualization
- Value

"**Big data describes datasets that are so large, complex, or rapidly changing that they push the very limits of our analytical capability.** It's a subjective term: What seems "big" today may seem modest in a few years when our analytic capacity has improved.

While big data can be about anything, the most important kinds of big data – and perhaps the only ones worth the effort – are those that can have a big impact through what they tell us about society, public health, the economy, scientific research, or any number of other large-scale subjects."

– Joel Gurin, author of *Open Data Now*

Data vs. Knowledge

Data

- Most enterprises use information systems
  - Production and distribution planning
  - Stock and supply management
  - Customer and personnel management
- Often coupled with a database system (e.g. databases of customers, suppliers, parts, etc.).
- Every possible individual piece of information can be retrieved.

Data

However: **Data alone are not enough.**

- “not see the forest for the trees”.
- patterns and structures go undetected
- Often such patterns can be exploited (e.g. for profit, efficiency, save lives...)

It comes down to the issue of ***data versus knowledge***.

Data

Examples of Data

- “Columbus discovered America in 1492.”
- “Mr. Jones owns a Ford Fusion.”

Characteristics of Data

- refer to single instances
- describe individual properties
- often available in huge amounts (databases, archives)
- relatively easy to obtain
- do not allow us to make predictions

Knowledge

Examples of Knowledge

- "All masses attract each other."

Characteristic of Knowledge

- refers to *classes* of instances
  (*sets* of objects, persons, points in time, etc.)
- describes general patterns, structure, laws, principles, etc.
- is usually difficult to find or to obtain
- allows us to make predictions

Criteria to Assess Knowledge

Not all statements are equally important, equally substantial, equally useful. **Knowledge must be assessed**

Assessment Criteria

- Correctness (probability, success in tests)
- Generality (range of validity, conditions of validity)
- Usefulness (relevance, predictive power)
- Comprehensibility (simplicity, clarity, parsimony)
- Novelty (previously unknown, unexpected)

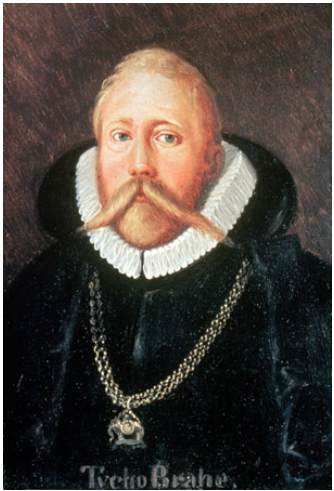
Who was Tycho Brahe? (1546–1601)

- Danish nobleman and astronomer
- 1582: built two observatories
- Determined positions of sun, moon and planets (accuracy: one angle minute, without a telescope!)
- Recorded the motions of the celestial bodies for several years.

Who was Tycho Brahe? (1546–1601)

Brahe's Problem

- Could not summarize the data in a consistent scheme.
- The planetary system he developed (the Tychonic system) did not stand the test of time.



Who was Johannes Kepler? (1571–1630)

- Assistant of Tycho Brahe
- He started from the data that Tycho Brahe had collected.
- Worked all his life to find laws that govern motion of planets

Who was Johannes Kepler? (1571–1630)

Kepler's Three Laws

1. Law of Orbits
2. Law of Areas
3. Law of Periods

are still taught in Physics classes nearly 400 years later!

hyperphysics.phy-astr.gsu.edu/hbase/kepler.html

How to find Knowledge?

We are drowning in information, but starving for knowledge

— John Naisbett

- Today huge amounts of data are available
- Manual methods of analysis have long ceased to be feasible.
- Simple aids (e.g. displaying data in charts) are too limited.

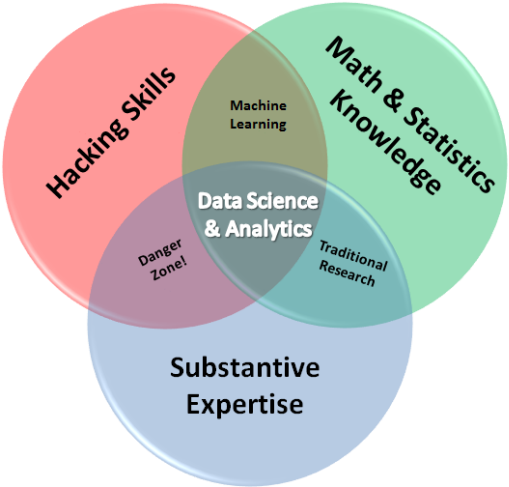
We need $\longrightarrow$

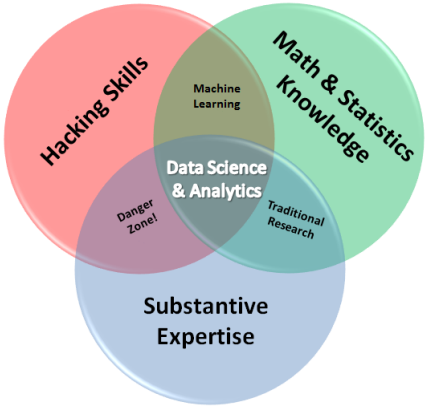
...serious computing power,

...statistical skills, hacking skills

...human intuition, creativity

What is Data Science and Analytics?





- Homeworks $\Rightarrow$ R, data, stats
- Projects $\Rightarrow$ R, data, stats *and* subject area, communication

"Make sure a candidate can find a story in a data set and provide a coherent narrative about a key data insight... "

— Harvard Business Review