SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA

BIG DATA

Marco Brambilla

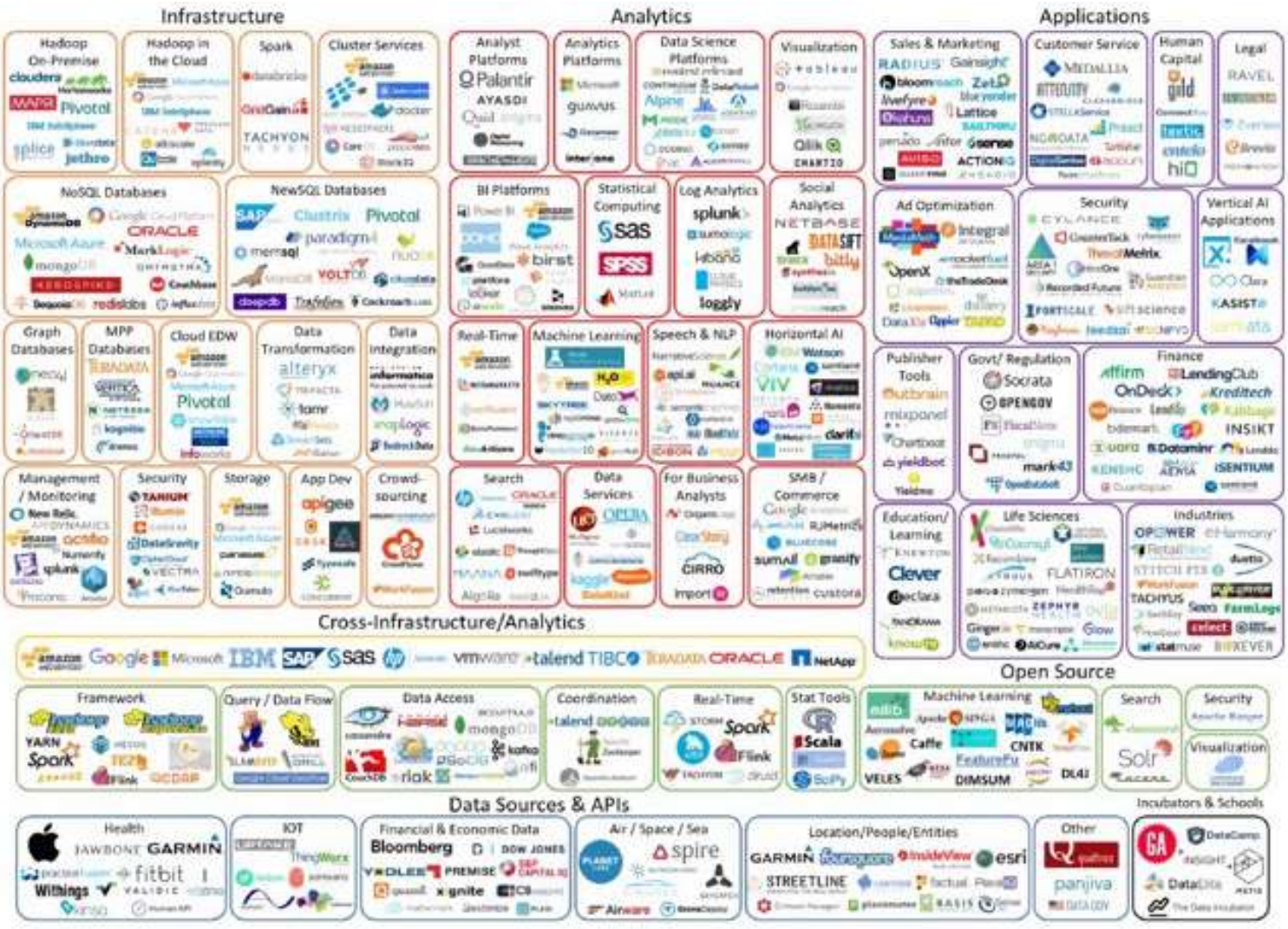
marco.brambilla@polimi.it

@marcobrambi

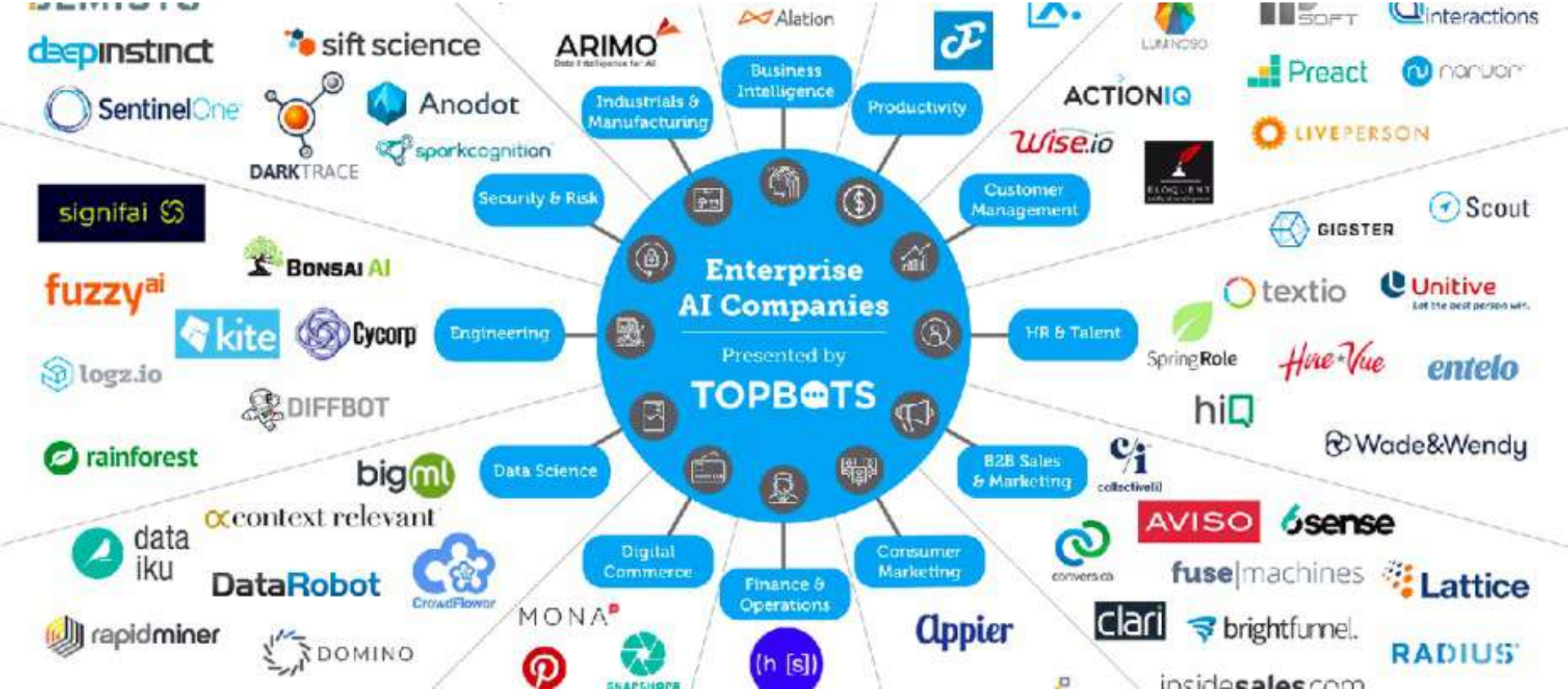
Big Data

It’s big!
A value model
Why now?
What is it?
Paradigm shifts enabled
Tools
Market Landscape
Trends

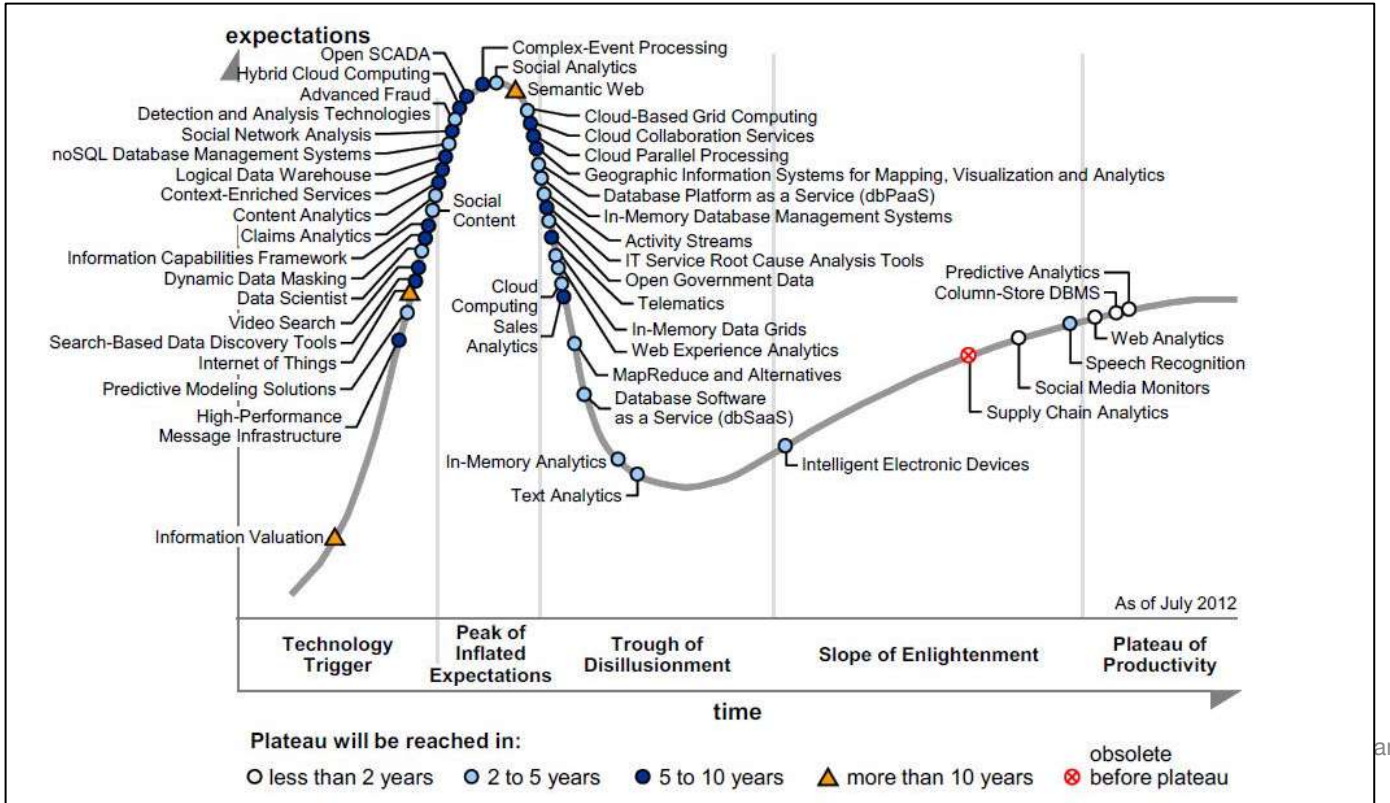
The Big Data Mkt.



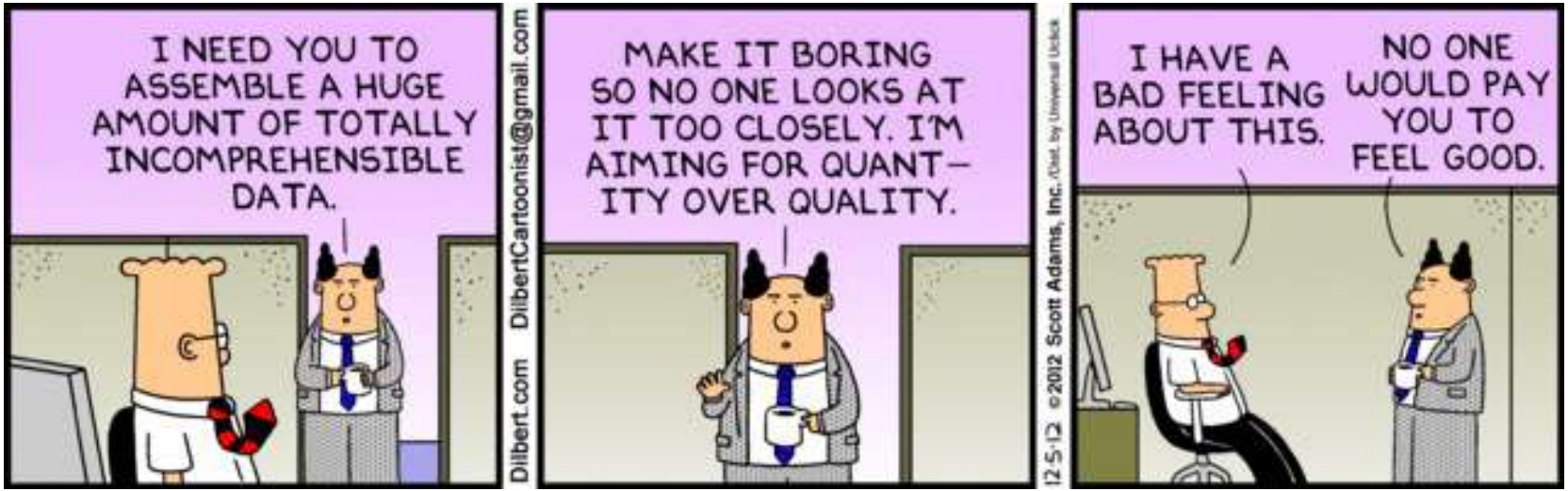
The AI Market



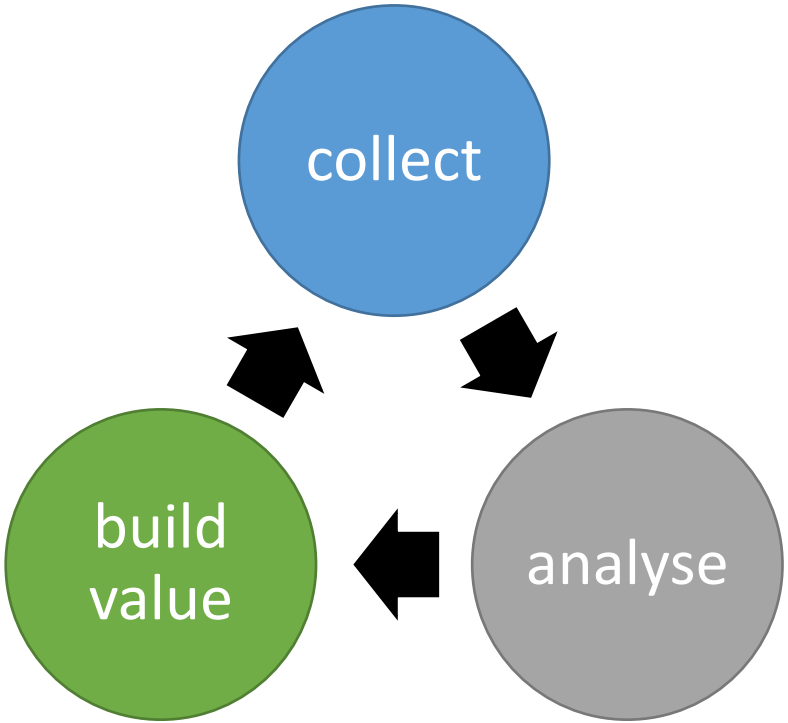
The big picture: the Hype Cycle



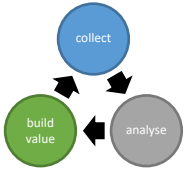
The Big Data Deluge



The Data-driven Virtuous Cycle



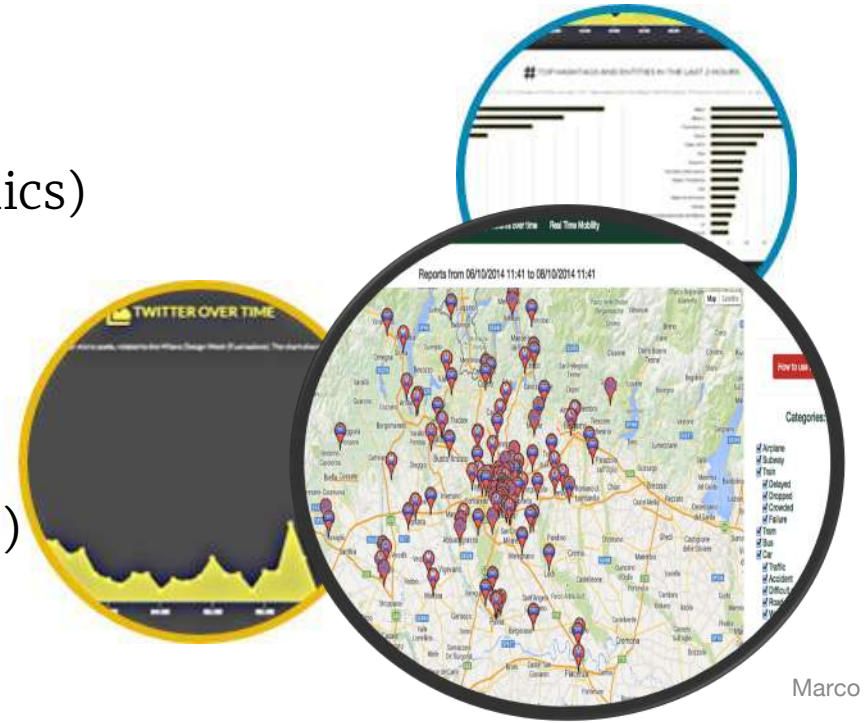
Method – Collect



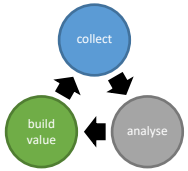
Data sources:

- WiFi logs
- Telco (incl. demographics)
- Transport
- IoT
- Social Media

  (s. networks, profiling, ...)



Method – Analyse



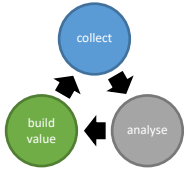
Stakeholders

- Digital Interaction on existing platforms (web, apps, social networks)
- Customers / Buyers / Citizens
- Economic Actor Association
- Citizen Groups
- Public Events

Analysis

- Description
- Prediction

Method – Build Value



Increase economic transactions

Brand awareness

Make cities a better place

Provide cultural value

Collect and build knowledge

Motivational Use Case

Netflix in numbers

Netflix streaming service accounts for
1/3 of peak time internet traffic in the US.

Last year, Netflix announced that it signed on **50 million** accounts worldwide.

Understanding viewing habits

Netflix specialists capture important information from a variety of different analytics streams:

- Personalization analytics
- Messaging analytics
- Content delivery analytics
- Device analytics
- ...

Back in 2006 …

Marco Brambilla.

With four data points …

Netflix was a DVD delivery in need for data analysis but with only **4 data points**:

- Customer ID
- Movie ID
- Rating
- The date the movie was watched

Nowadays, with more data points …

As streaming video became the primary focus → more insights on the customers

Time of day something was watched
User gender and age (based on individual logins)
Time spent selecting movies
How often a movie or program was paused
…

Netflix predicts the "perfect situation"

Netflix can now **analyze** how these factors impacted **viewers' enjoyment**

The happier the customers, the longer they stay

Models to **predict the "perfect situation"** for the customer experience

… creates "specific" categories

Tagging

Preferences

76,897 ways to describe types of movies

“alt-genres”: e.g.,“Comedies Featuring a Strong Female Lead”

… bids on original programming

When the series **House of Cards** began shopping around for a home, **Netflix** aggressively jumped on it, outbidding major cable networks with a massive **two-season order**.

Why the enthusiasm?

By analyzing its data, Netflix saw that a large majority of its viewers enjoy programs :

directed by David Fincher (who directed Se7en, Fight Club and The Social Network)

starring Kevin Spacey.

House of Cards was exactly that!

Data Promotes Shows

Netflix took a **data-driven approach to promote** House of Cards.

For example, the company **modelled the show's cover image on the colours and styles for successful, similarly tagged programs**, to help draw new viewers in

Big Data. Why now?

Why now? 1

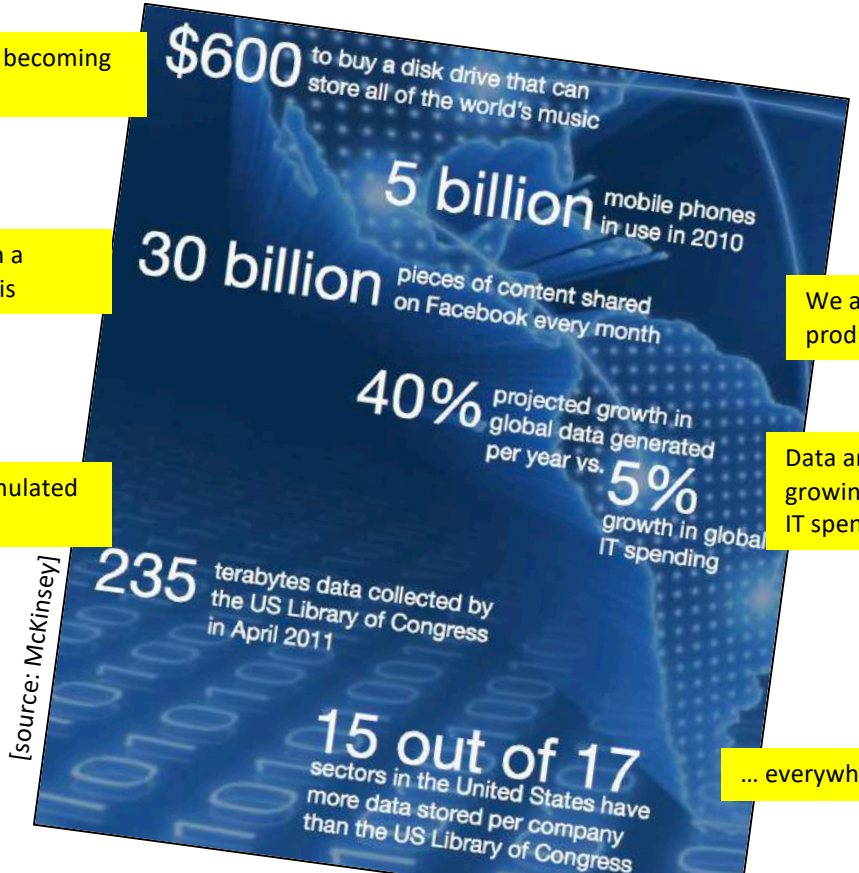

710 COMPUTERS SOLD
232 COMPUTERS GOT INFECTED BY MALWARE
555 OF THEM WITH intel
2,6 MILLION CD'S
1,820 TB OF DATA CREATED
450 Windows 7 CDs SOLD
12 WEBSITES GOT HACKED
416 ATTEMPTS
redbox
1,400 DISCS ARE RENTED ON ONLINE MOVIE RENTAL SERVICE
180+ BY MOBILE
ebay
950+ PURCHASES ON ebay
$10,000 BY MOBILE
PayPal
$219,000 OF TOTAL PAYMENTS
IN 60 SECONDS
What Happens in an Internet Minute?
639,800 GB of global IP data transferred
20 New victims of identity theft
204 million Emails sent
47,000 App downloads
$83,000 In sales
61,141 Hours of music
20 million Photo views
3,000 Photo uploads
320+ New Twitter accounts
100,000 New tweets
135 Botnet infections
1,300 New mobile users
100+ New LinkedIn accounts
6 New Wikipedia articles published
277,000 Logins
6 million Facebook views
2+ million Search queries
30 Hours of video uploaded
1.3 million Video views
And Future Growth is Staggering
intel

Why now? 2

Hard-drive are becoming cheaper

We do it on a regular basis

Data gets cumulated …



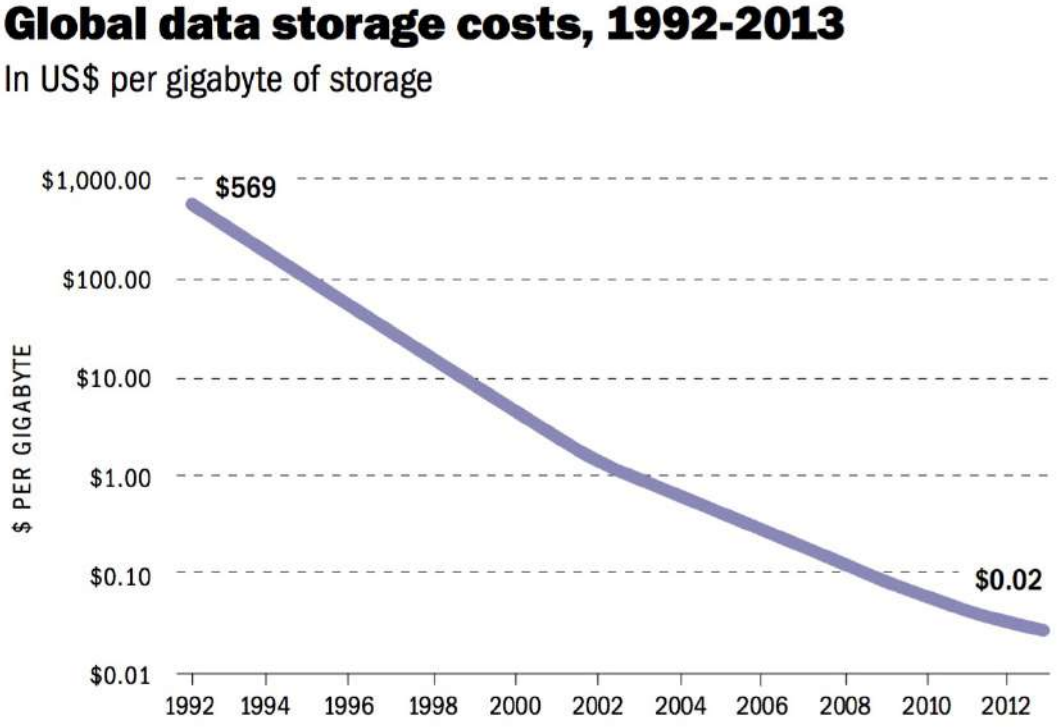
[source: McKinsey]

We are all data producers

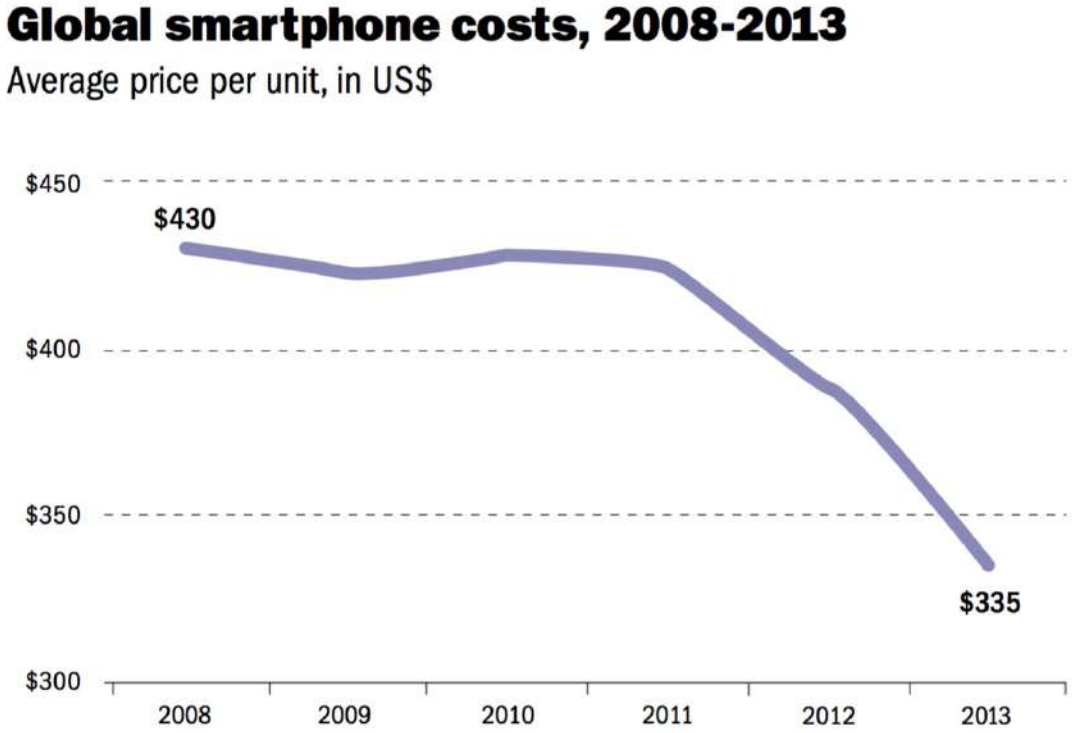
Data are growing faster than IT spending

… everywhere

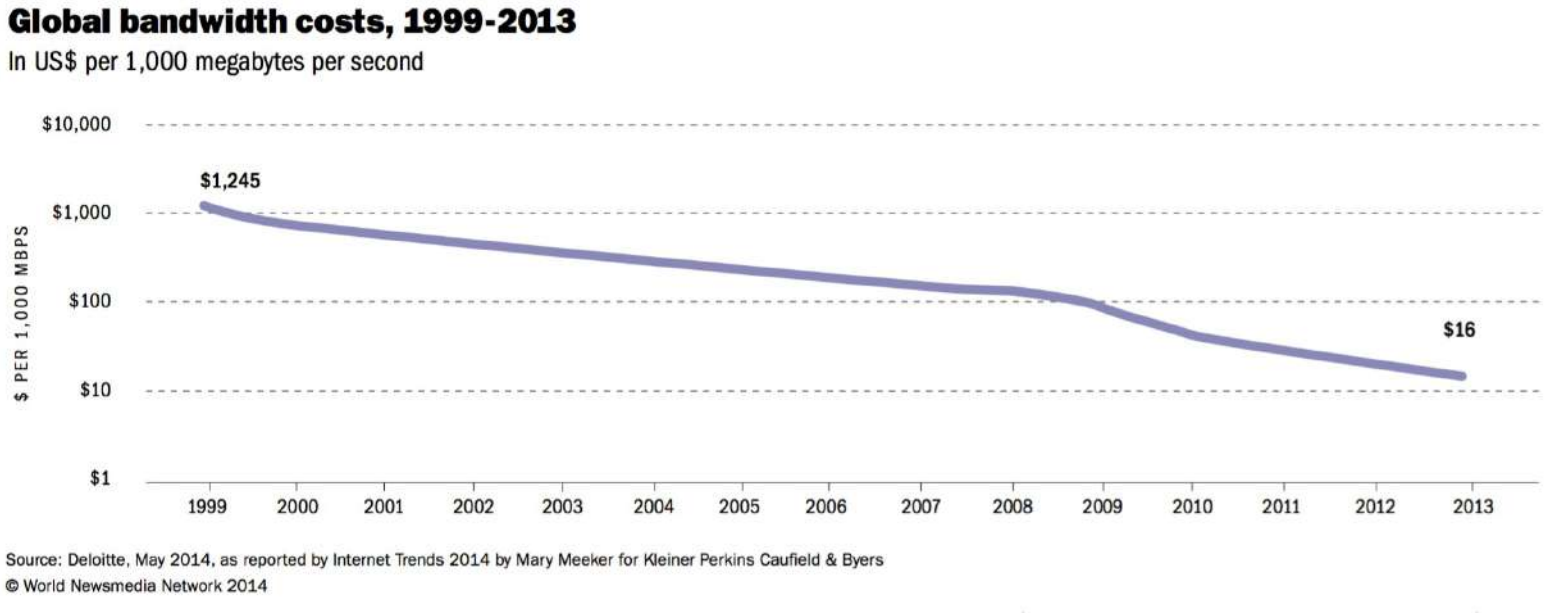
Why now? 3


Global data storage costs, 1992-2013
In US$ per gigabyte of storage
$1,000.00
$100.00
$10.00
$1.00
$0.10
$0.01
$ PER GIGABYTE
$569
$0.02
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Source: Deloitte, May 2014, as reported by Internet Trends 2014 by Mary Meeker for Kleiner Perkins Caufield & Byers
© World Newsmedia Network 2014

Why now? 4


Global smartphone costs, 2008-2013
Average price per unit, in US$
$450
$400
$350
$300
$430
$335
2008
2009
2010
2011
2012
2013
Source: Deloitte, May 2014, as reported by Internet Trends 2014 by Mary Meeker for Kleiner Perkins Caufield & Byers

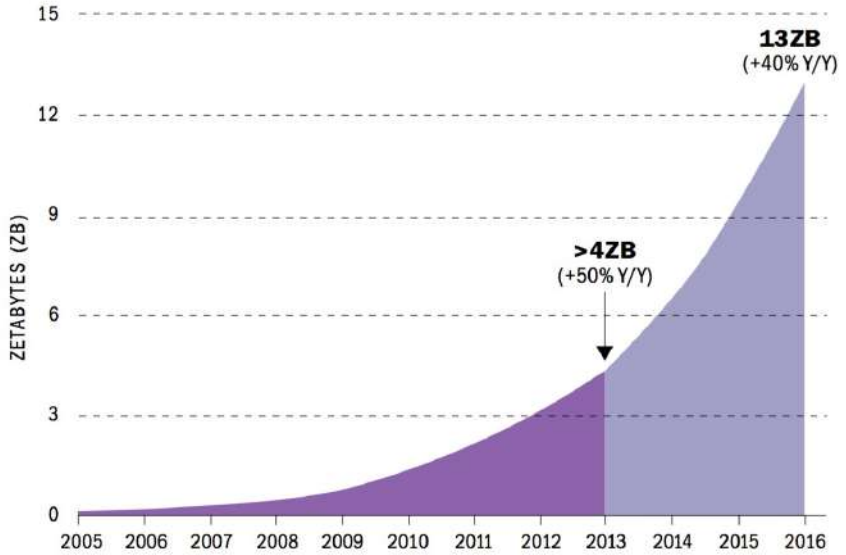
Why now? 5


Global bandwidth costs, 1999-2013
In US$ per 1,000 megabytes per second
$10,000
$1,000
$100
$10
$1
$ PER 1,000 MBPS
$1,245
$16
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Source: Deloitte, May 2014, as reported by Internet Trends 2014 by Mary Meeker for Kleiner Perkins Caufield & Byers
© World Newsmedia Network 2014

Why now? 6

Digital content universe generated by consumers

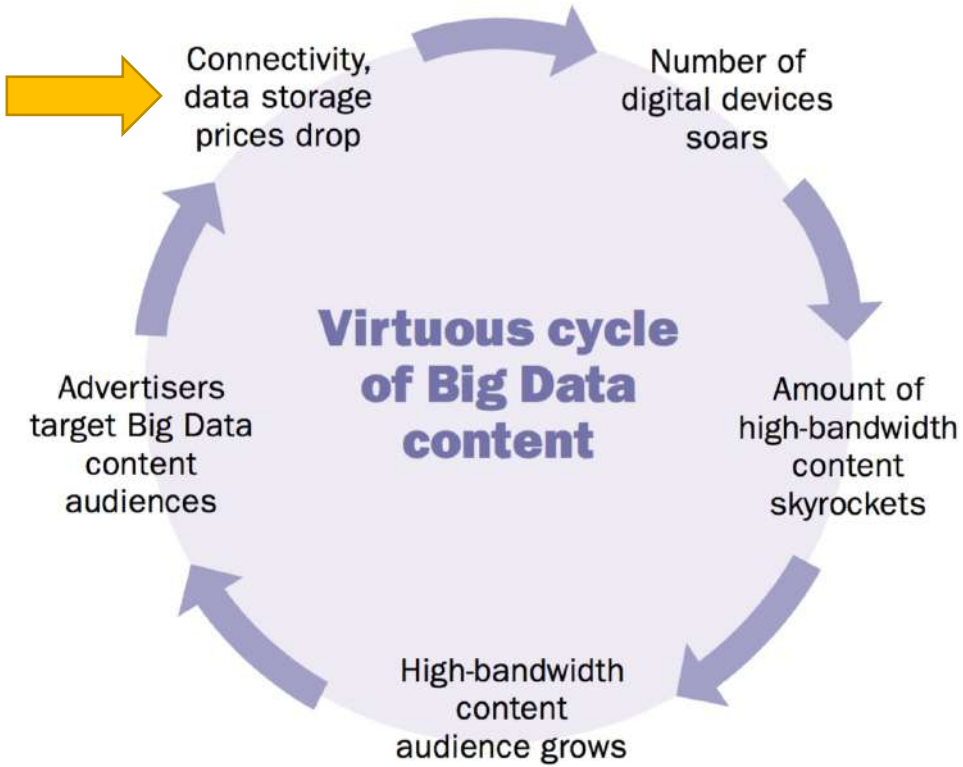
Video and photo generation, consumption and sharing and social media usage made up the bulk of online content in 2013



Note: 1 petabyte = 1 million gigabytes. 1 zetabyte = 1 million petabytes

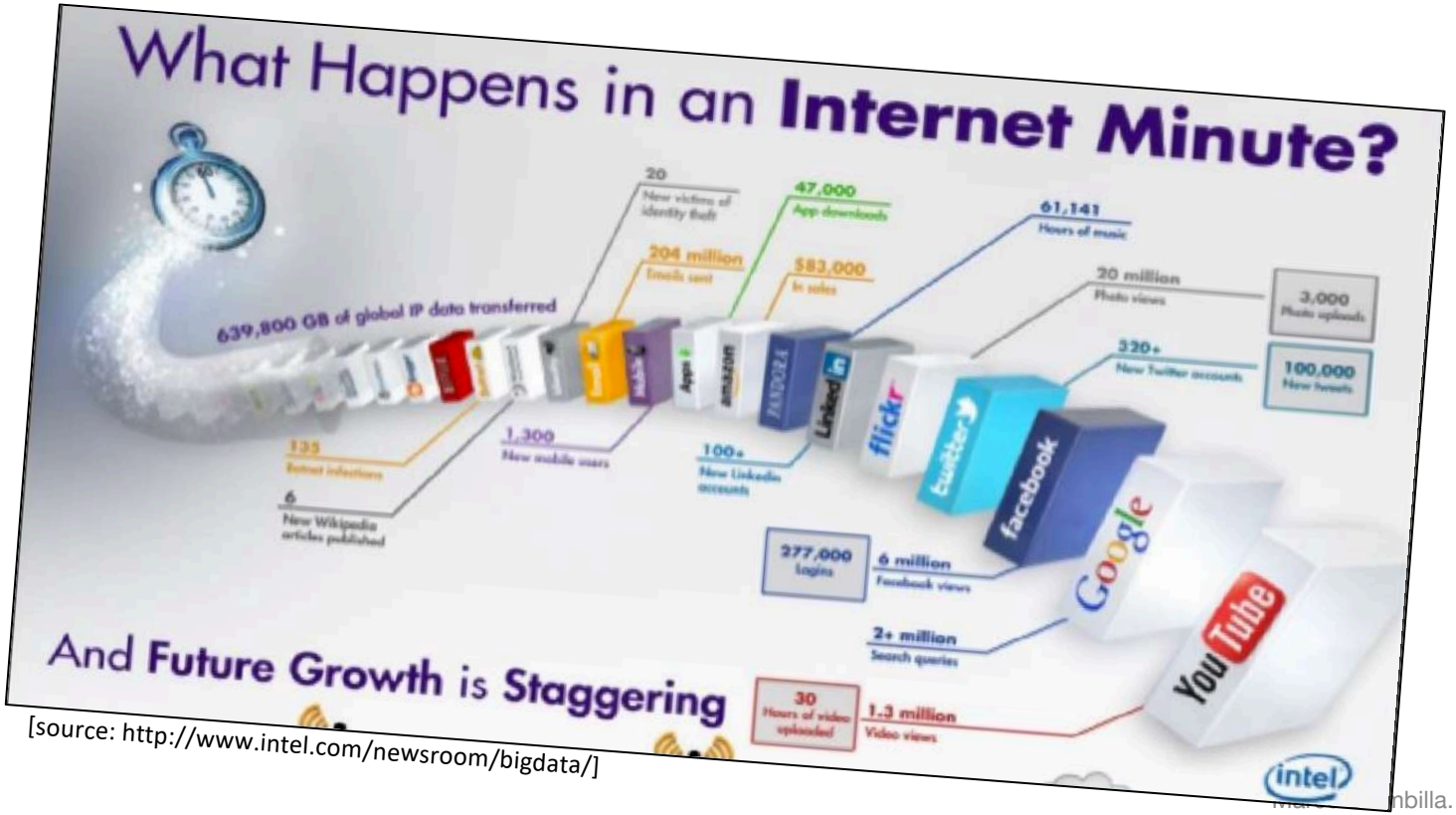
Source: Deloitte, May 2014, as reported by Internet Trends 2014 by Mary Meeker for Kleiner Perkins Caufield & Byers

Why now? 7



Source: World Newsmedia Network 2014

Why now? 8

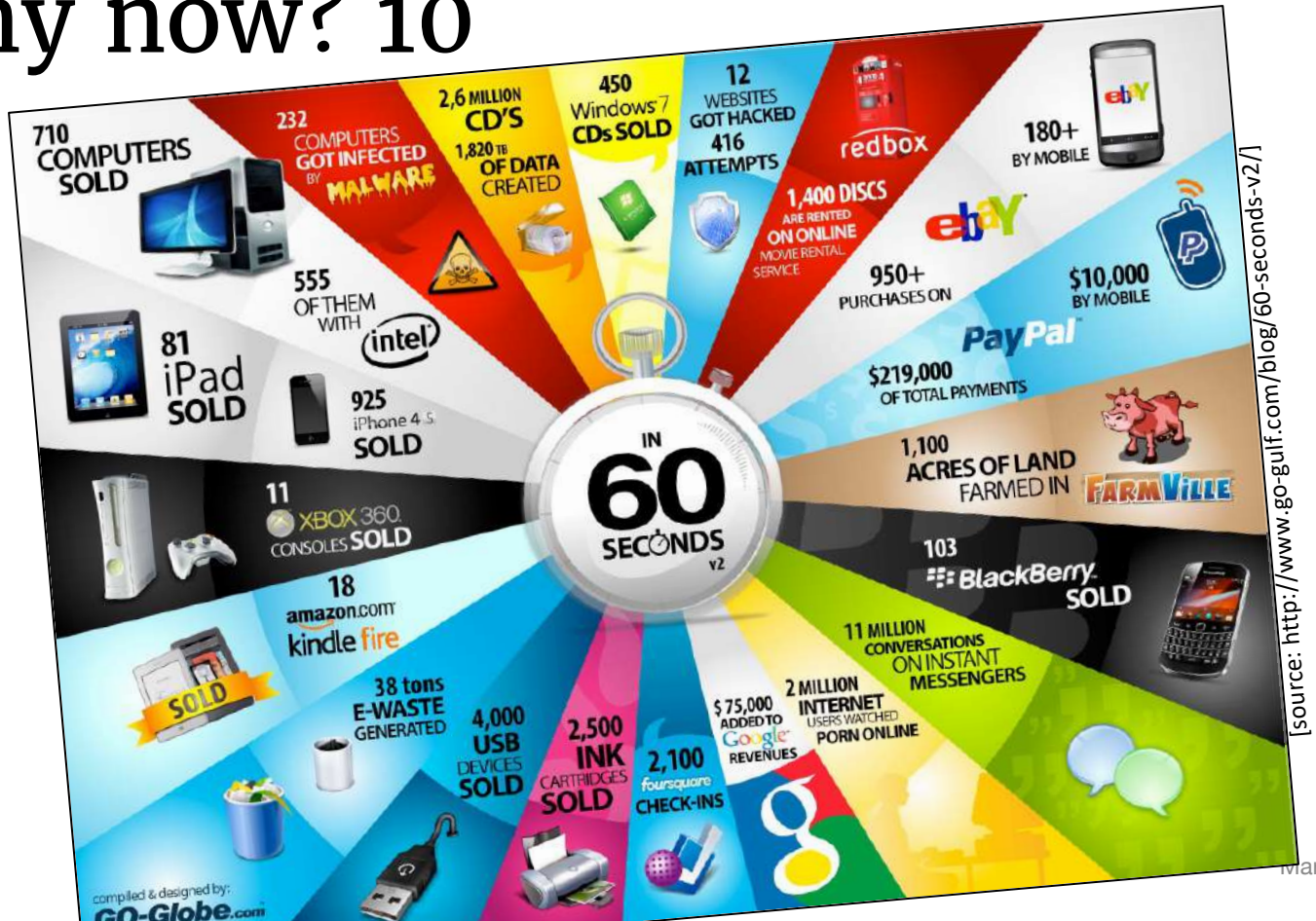


[source: http://www.intel.com/newsroom/bigdata/]

Why now? 9

[source: http://www.go-gulf.com/blog/60-seconds/]

Why now? 10



[source: http://www.go-gulf.com/blog/60-seconds-v2/]

Its value 1

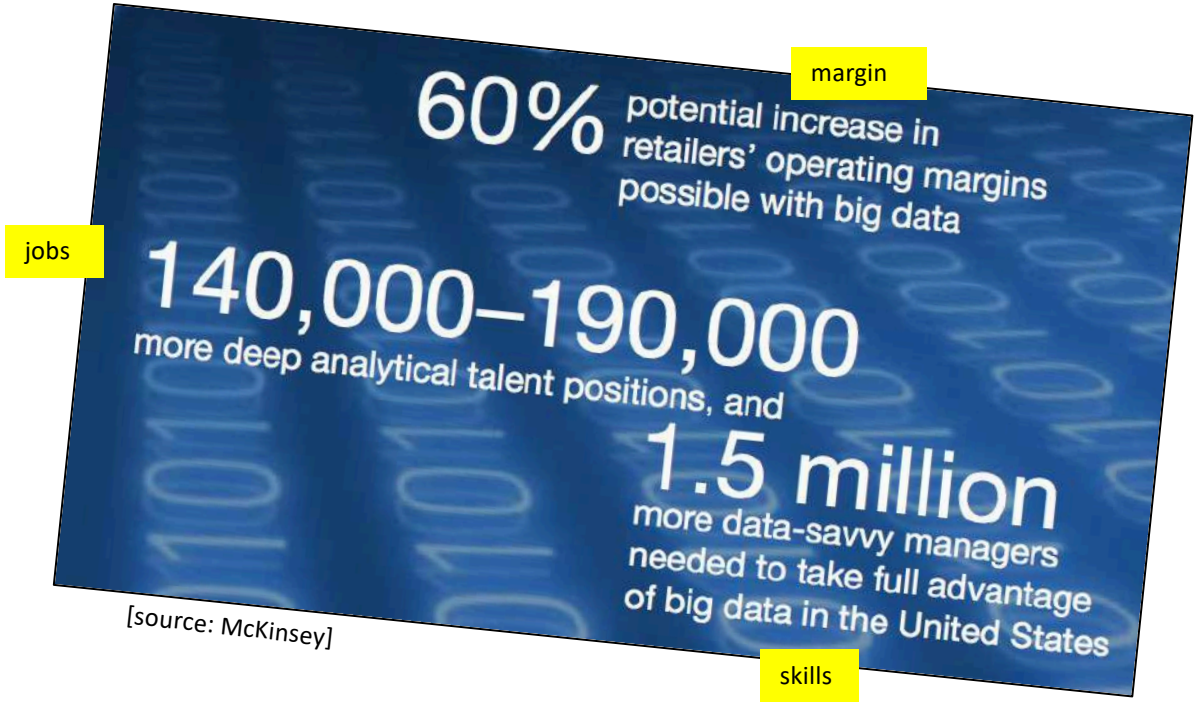
healthcare

Public Administrations

[source: McKinsey]

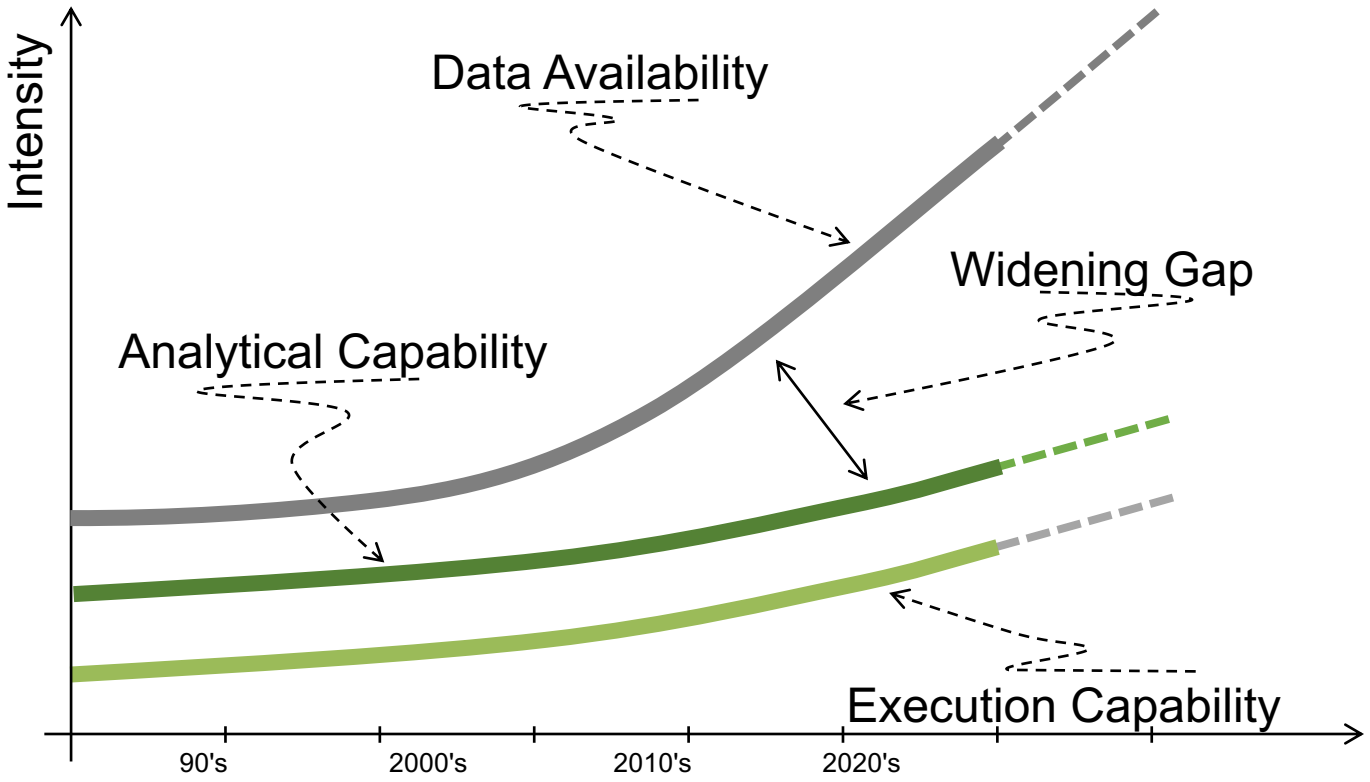
Personal location

Its value 2



[source: McKinsey]

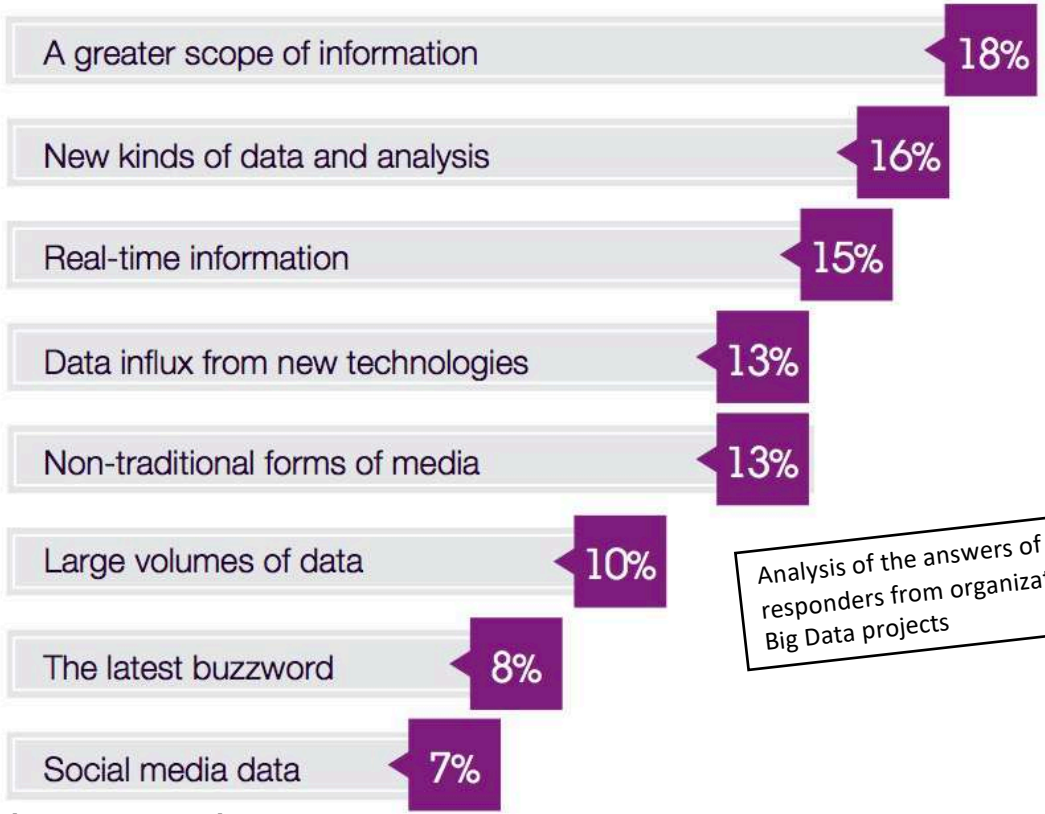
Data vs. Analytical vs. Execution Capability



Marco Brambilla.

Big Data. What?

… so what's Big Data?



[source: IBM, 2012]

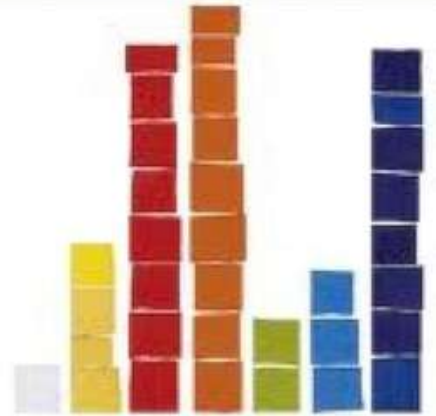
Analysis of the answers of 1144 responders from organizations that run Big Data projects

What's Big Data?

Volume

Data at scale

Terabytes to hexabyte of data cumulated on cheaper and cheaper storages

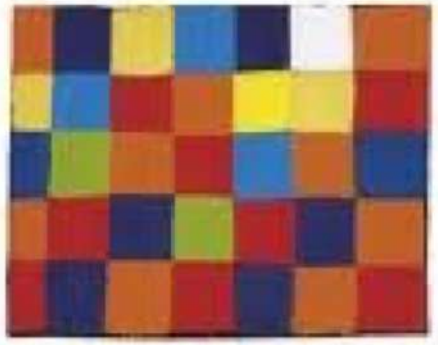


What's Big Data?

Variety

Data in many form

Structured, unstructured, text, images, videos, multimedia

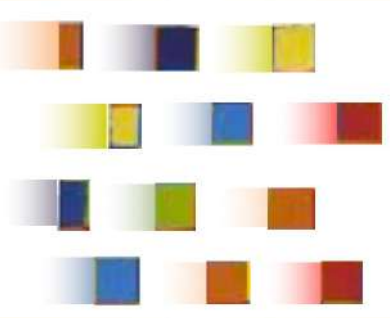


What's Big Data?

Velocity

Data in motion

Analysis of streaming data to enable decision within fractions of a second

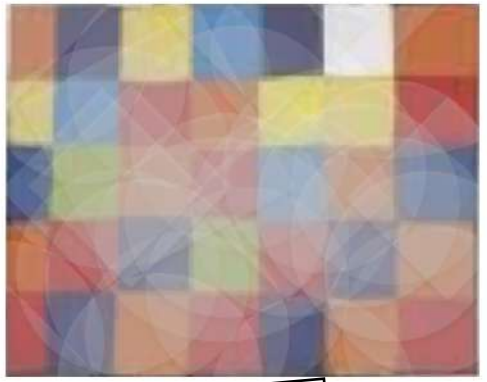


What's Big Data?

Veracity

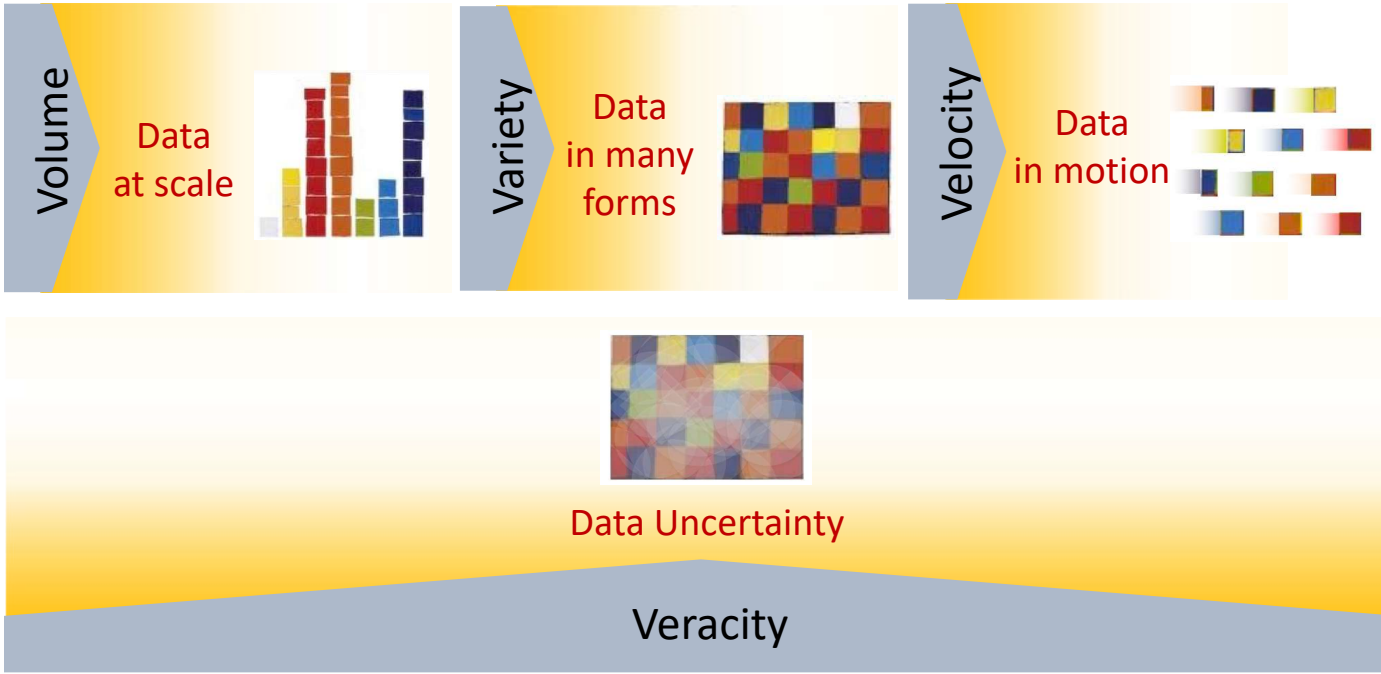
Data Uncertainty

Managing the reliability and predictability of inherently imprecise data type



the one certainty about uncertainty is that it is not likely to go away

What's Big Data? (cont.)



More Vs

Value

Volatility

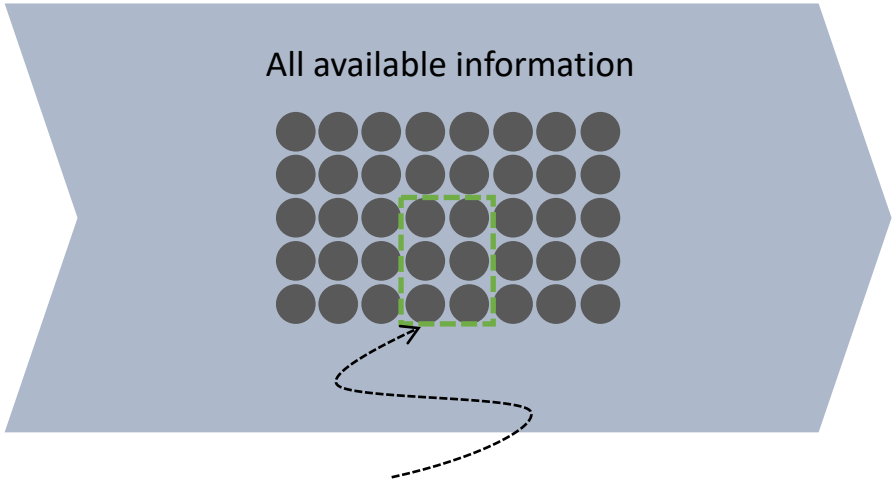
Validity

...

Solving problems in new ways

Leverage more of the data being captured

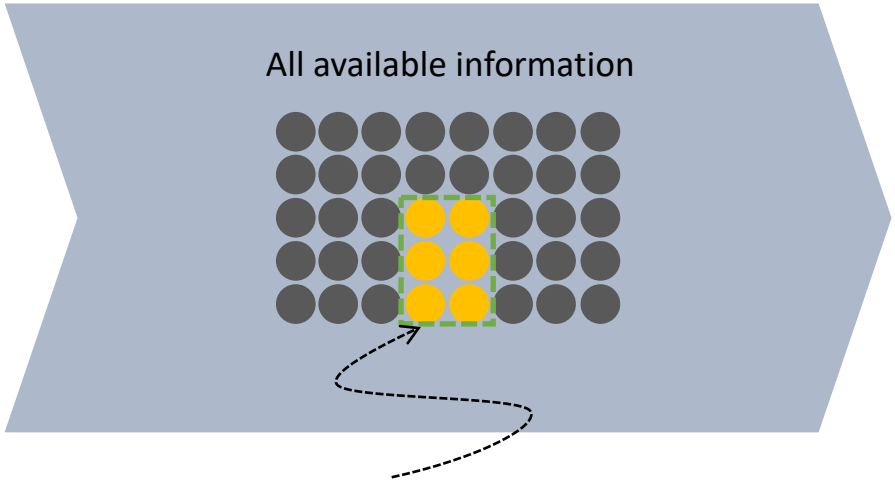
TRADITIONAL APPROACH



Solving problems in new ways (cont.)

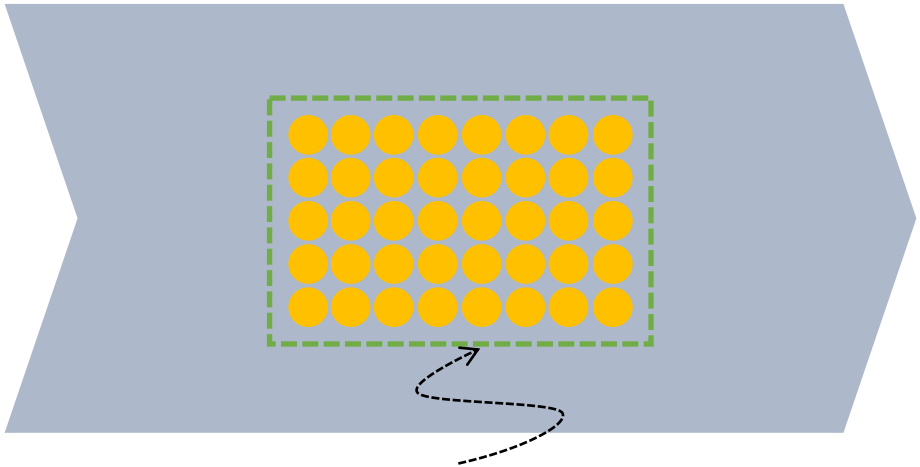
Leverage more of the data being captured

TRADITIONAL APPROACH



Analyze small subset of information

INNOVATIVE APPROACH



Analyze **all** available information

Solving problems in new ways (cont.)

Leverage more of the data being captured

TRADITIONAL APPROACH

Analyze small subset of information

UNION PACIFIC

INNOVATIVE APPROACH

Analyze **all** available information

Statistical foundations!

Central Limit Theorem

The Central Limit Theorem (CLT) states that the sample mean of a sufficiently large number of i.i.d. random variables is approximately normally distributed. The larger the sample, the better the approximation.

Change the parameters α and β to change the distribution from which to sample.

$\alpha = 0.21$ ○

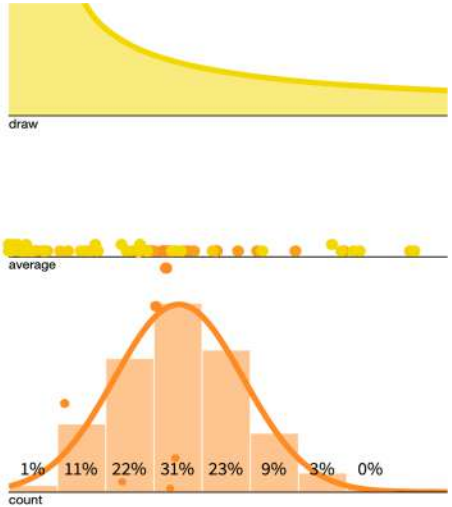
$\beta = 0.98$ ○

Choose the sample size and how many sample means should be computed (draw number), then press "Sample." Check the box to display the true distribution of the sample mean.

Sample size = 15 ○

Draws = 50 ○

☑ Theoretical

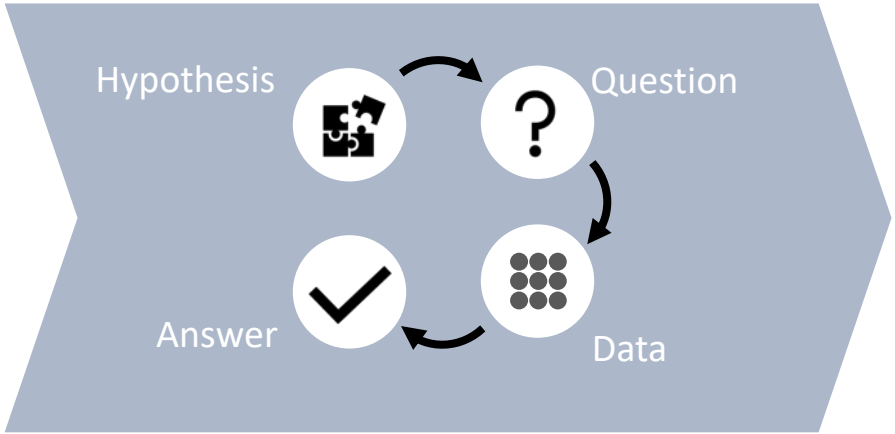


[source: https://seeing-theory.brown.edu/probability-distributions/index.html#section3]

Solving problems in new ways (cont.)

Data-driven exploration looking for correlation

TRADITIONAL APPROACH

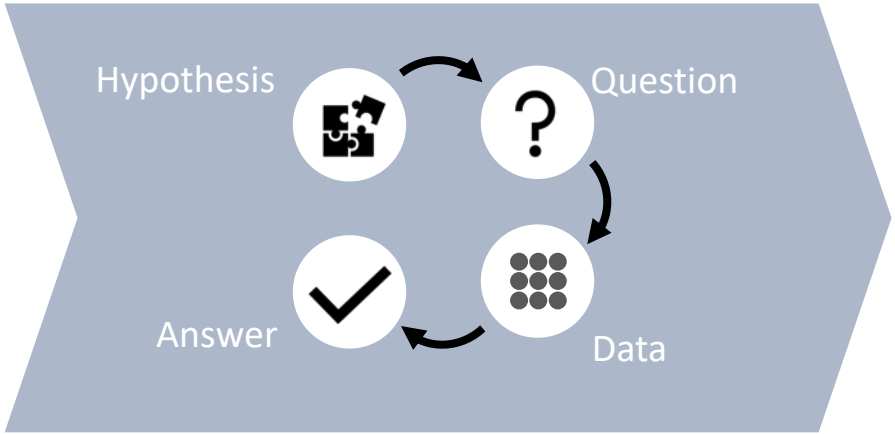


Start with hypothesis and test against selected data

Solving problems in new ways (cont.)

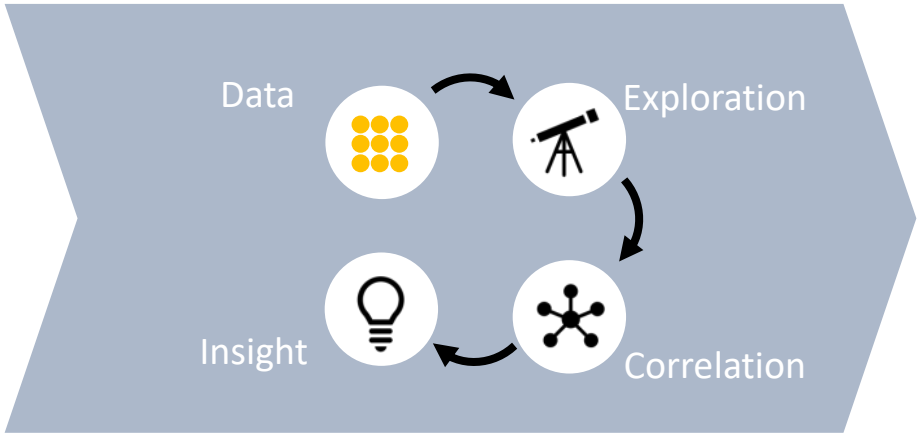
Data-driven exploration looking for correlation

TRADITIONAL APPROACH



Start with hypothesis and test against selected data

INNOVATIVE APPROACH

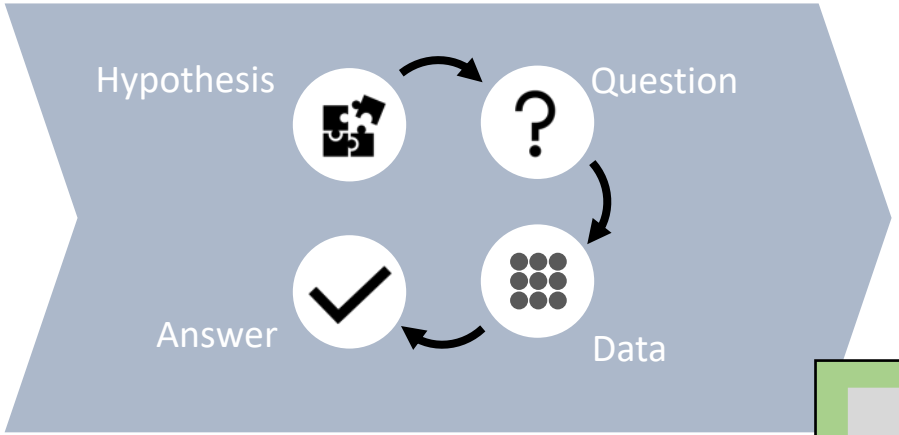


Explore **all** data and identify correlations

Solving problems in new ways (cont.)

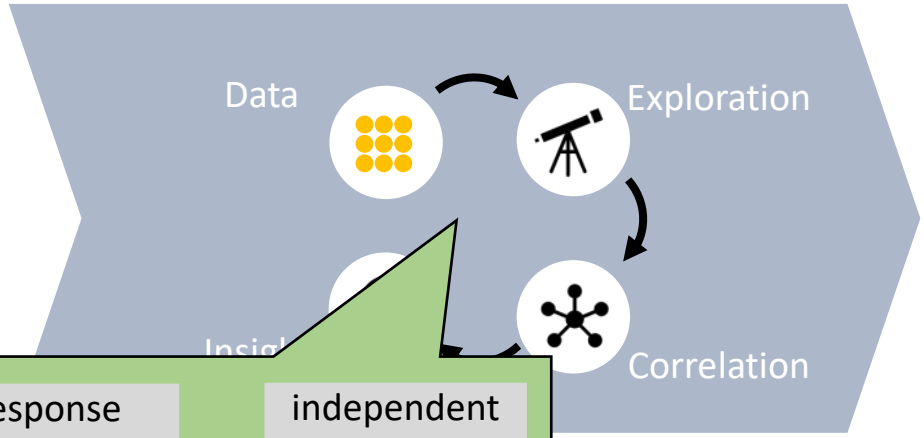
Data-driven exploration looking for correlation

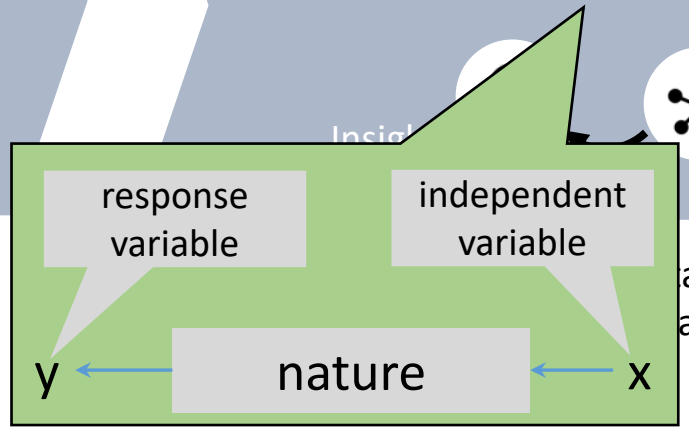
TRADITIONAL APPROACH



Start with hypothesis and test against selected data

INNOVATIVE APPROACH





a and
ations

Marco Brambilla.

Your butcher …

Marco Brambilla.

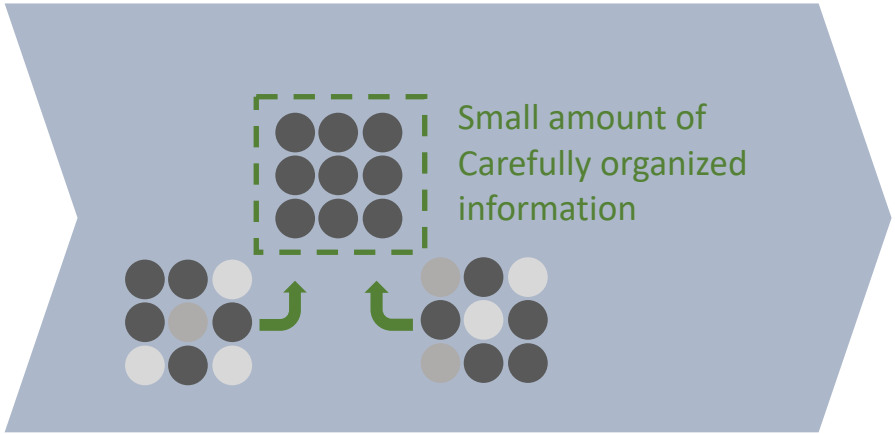
… at scale!

Marco Brambilla.

Solving problems in new ways (cont.)

Reduce effort required to leverage data

TRADITIONAL APPROACH

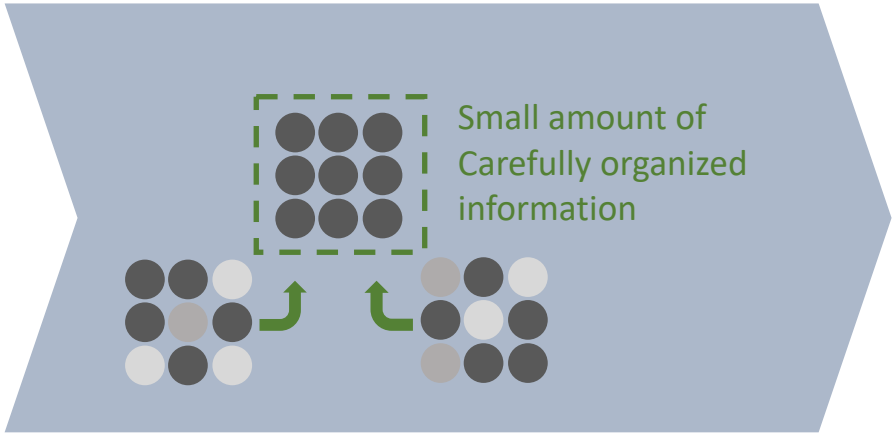


Carefully cleanse information
Before any analysis

Solving problems in new ways (cont.)

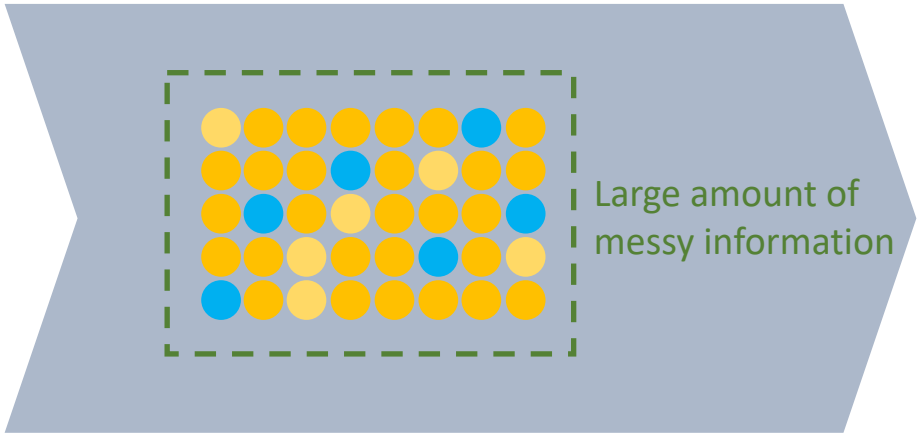
Reduce effort required to leverage data

TRADITIONAL APPROACH



Carefully cleanse information
Before any analysis

INNOVATIVE APPROACH



Analyze information as is,
Cleanse as needed

Solving problems in new ways (cont.)

Reduce effort required to leverage data

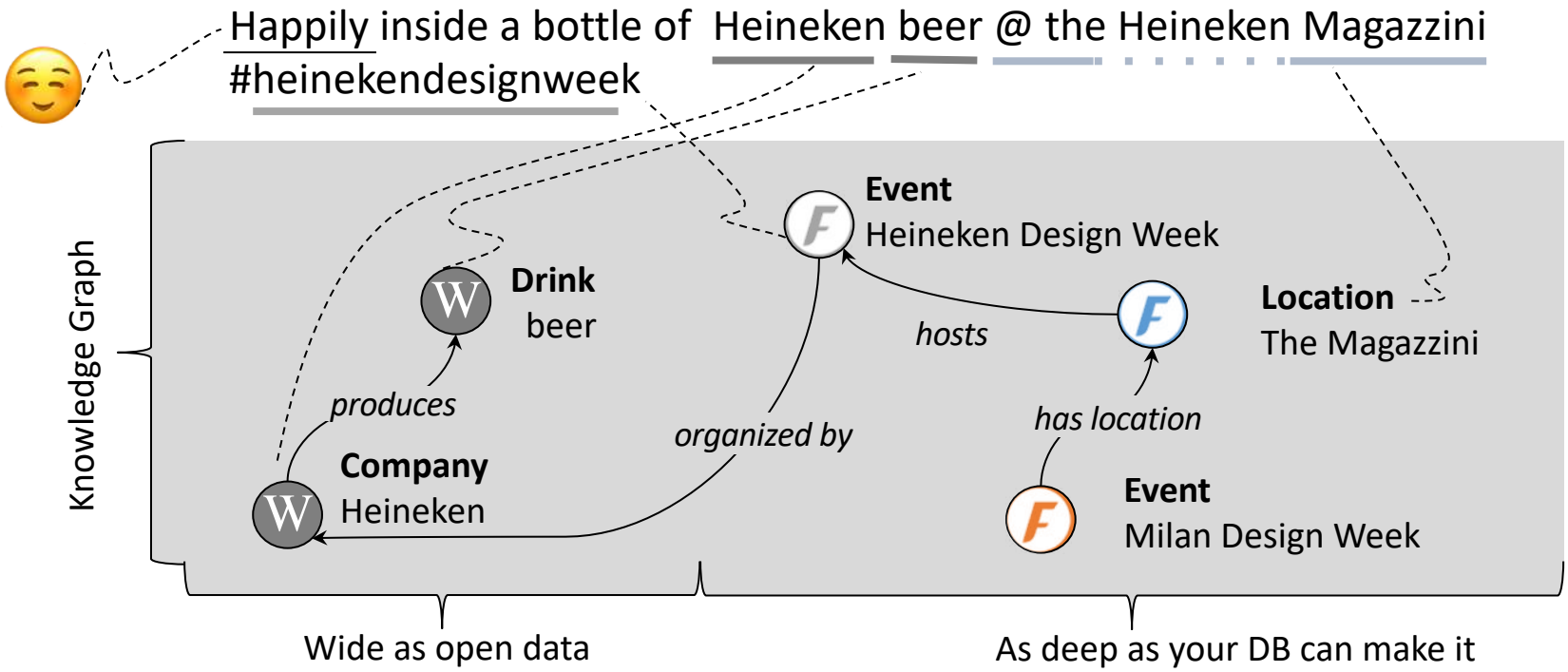
TRADITIONAL APPROACH

Carefully cleanse information
Before any analysis

INNOVATIVE APPROACH

Analyze information as is,
Cleanse as needed

Reduce effort required to leverage data



The case of IBM Watson

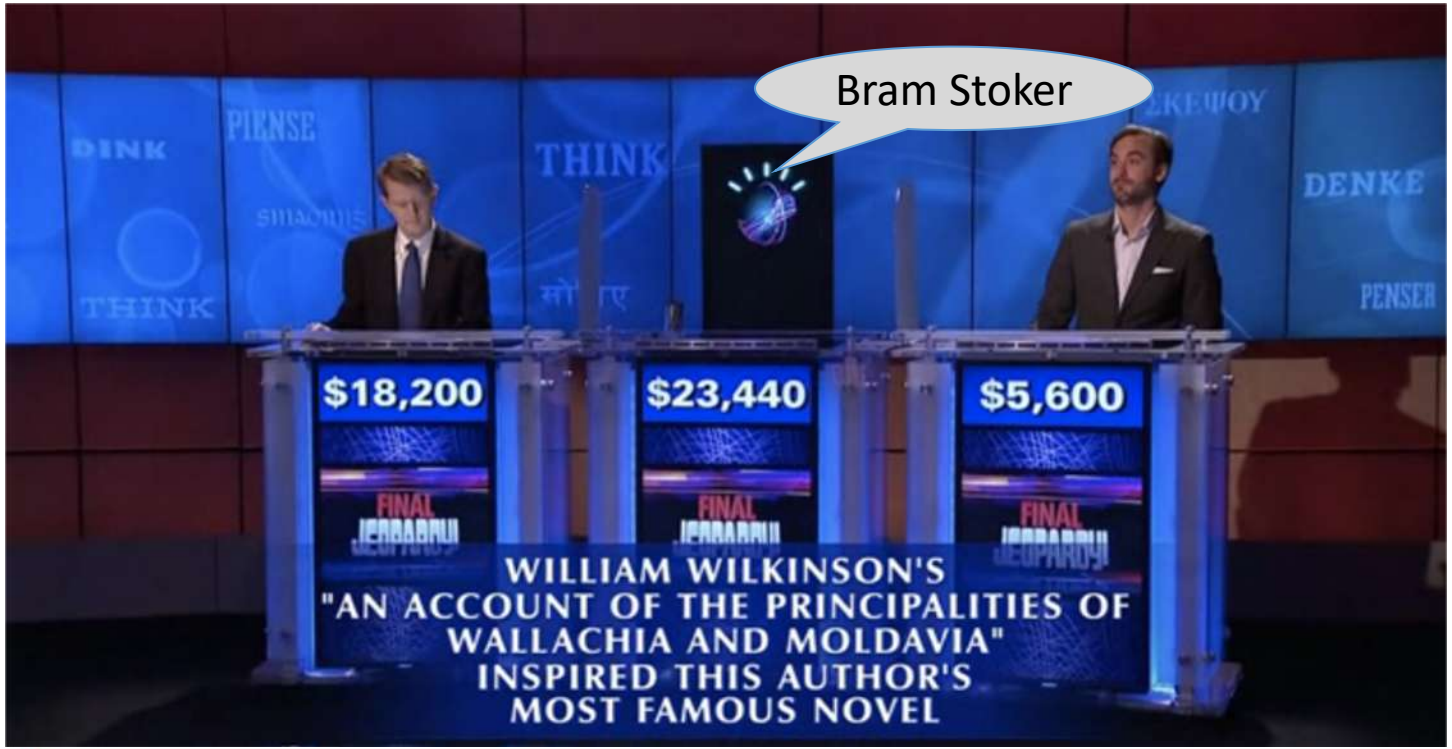
IBM Watson

The case of IBM Watson (cont.)


```
?x rdf:type dbo:Writer .
  ?x dbo:notableWork ?w .
  ?w dbp:name ?name .
<http://dbpedia.org/resource/William_Wilkinson_(diplomat)> rdfs:comment ?t .
FILTER regex(?t, concat(".*",?name), "i") }
```

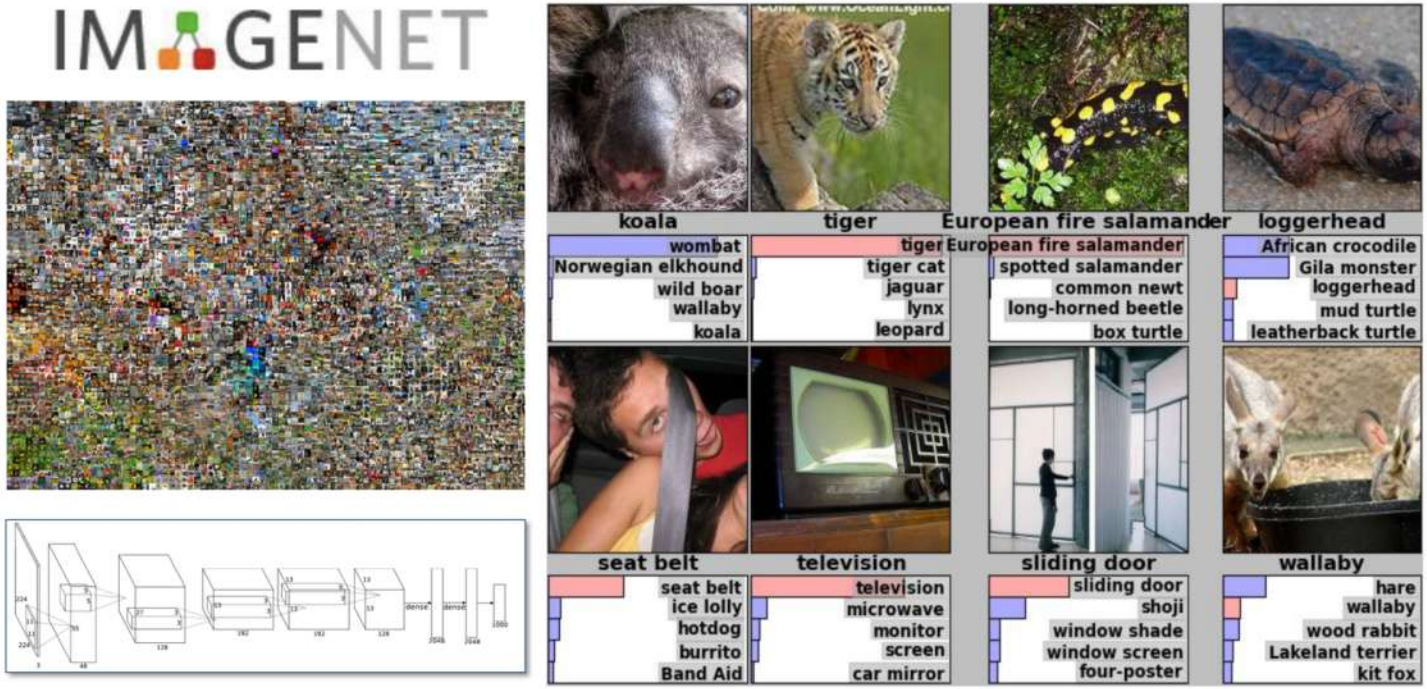
IBM Watson

The case of IBM Watson (cont.)

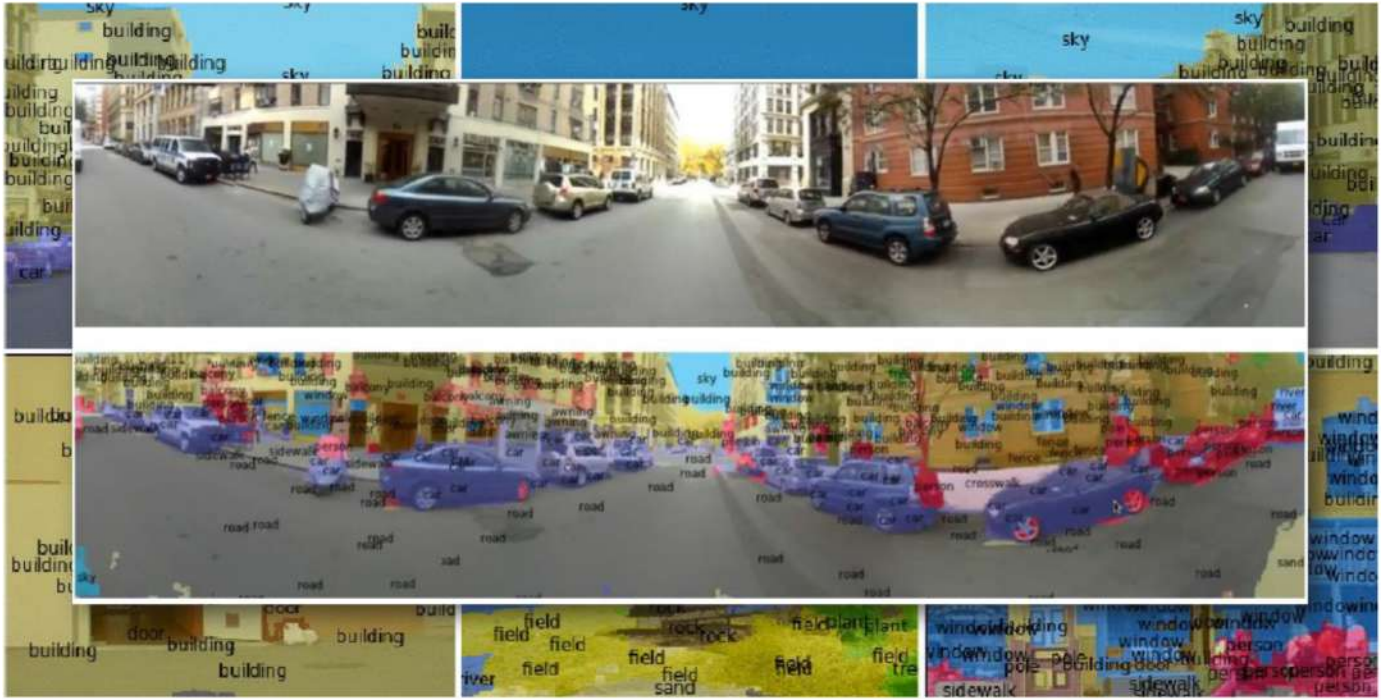


IBM Watson

Reduce effort required to leverage data



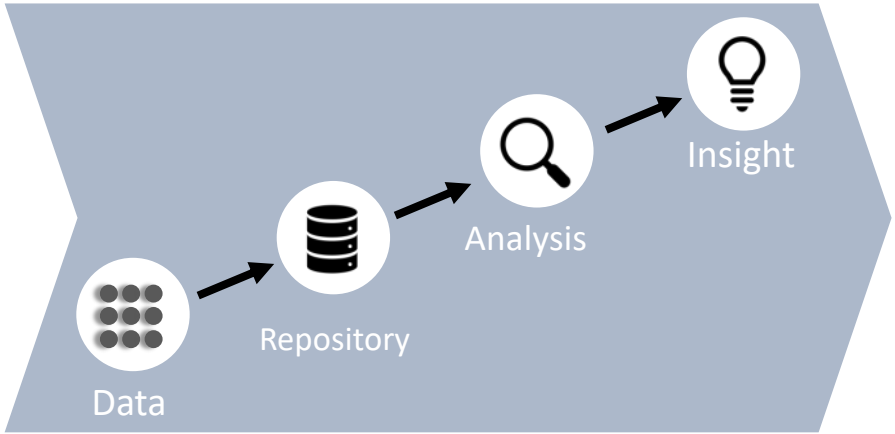
Reduce effort required to leverage data



Solving problems in new ways (cont.)

Leverage data as it is captured

TRADITIONAL APPROACH

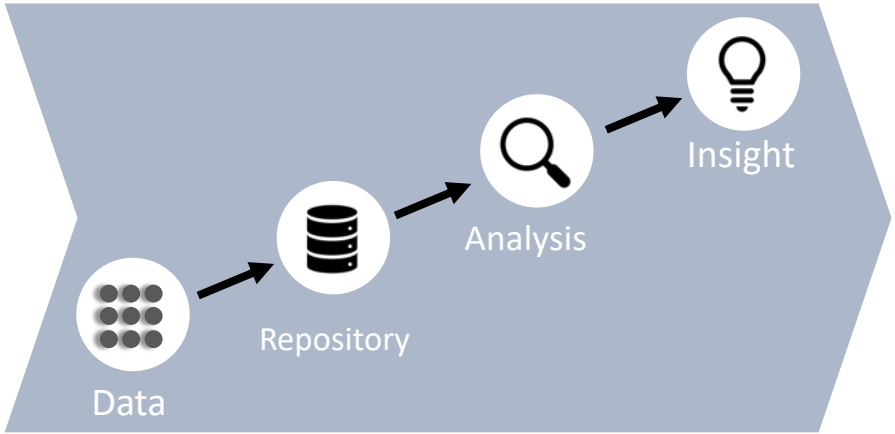


Analyze data **after** it's been processed
And landed in a warehouse or mart

Solving problems in new ways (cont.)

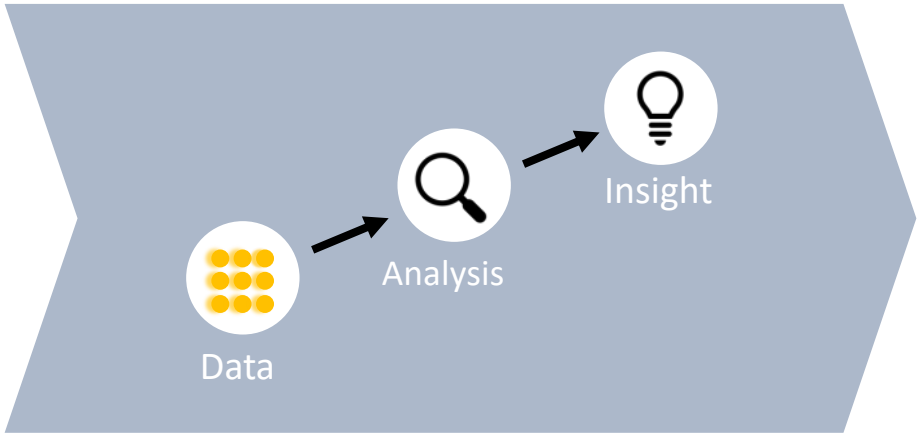
Leverage data as it is captured

TRADITIONAL APPROACH



Analyze data **after** it's been processed
And landed in a warehouse or mart

INNOVATIVE APPROACH

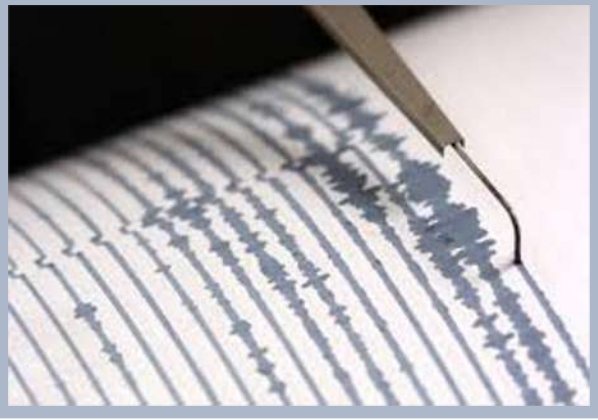


Analyze data **in motion** as it's
Generated, in real-time

Solving problems in new ways (cont.)

Leverage data as it is captured

TRADITIONAL APPROACH



Analyze data **after** it's been processed
And landed in a warehouse or mart

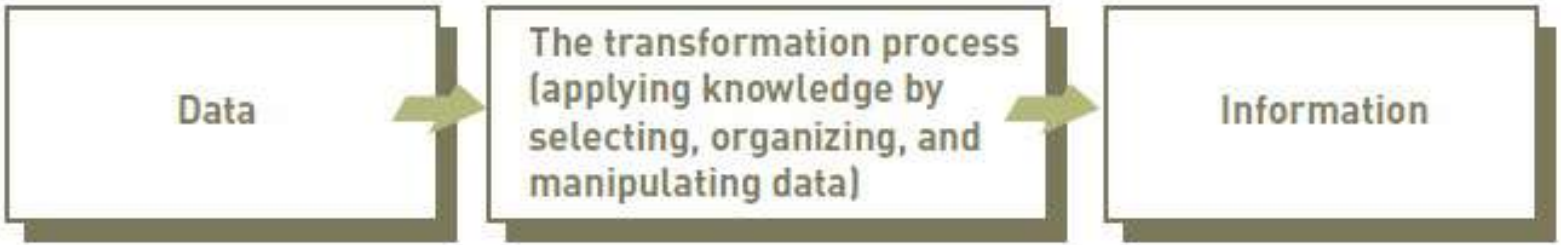
INNOVATIVE APPROACH

Analyze data **in motion** as it's
Generated, in real-time

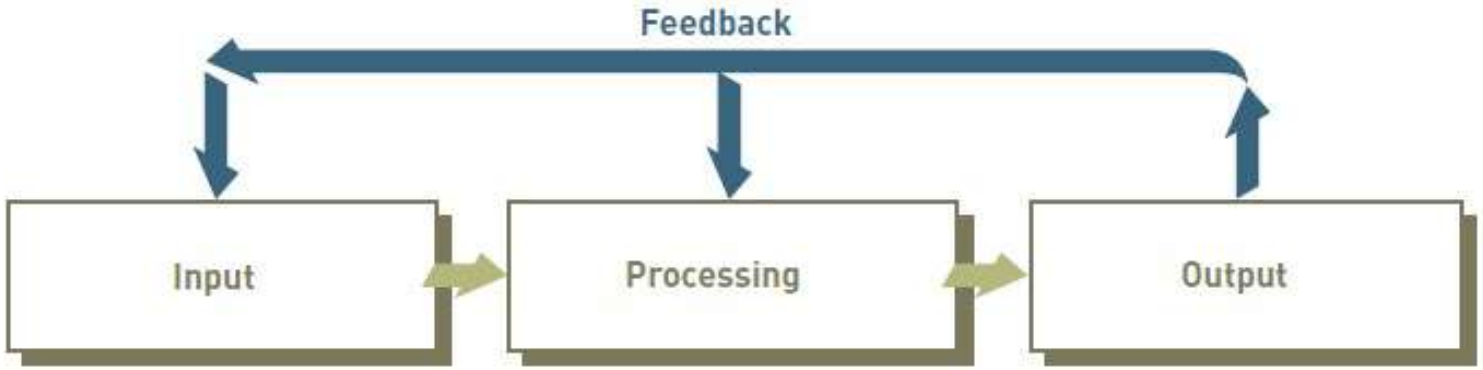
The case of Google's Self-Driving Car

[Src. https://www.youtube.com/watch?v=YXylqtEQ0tk]

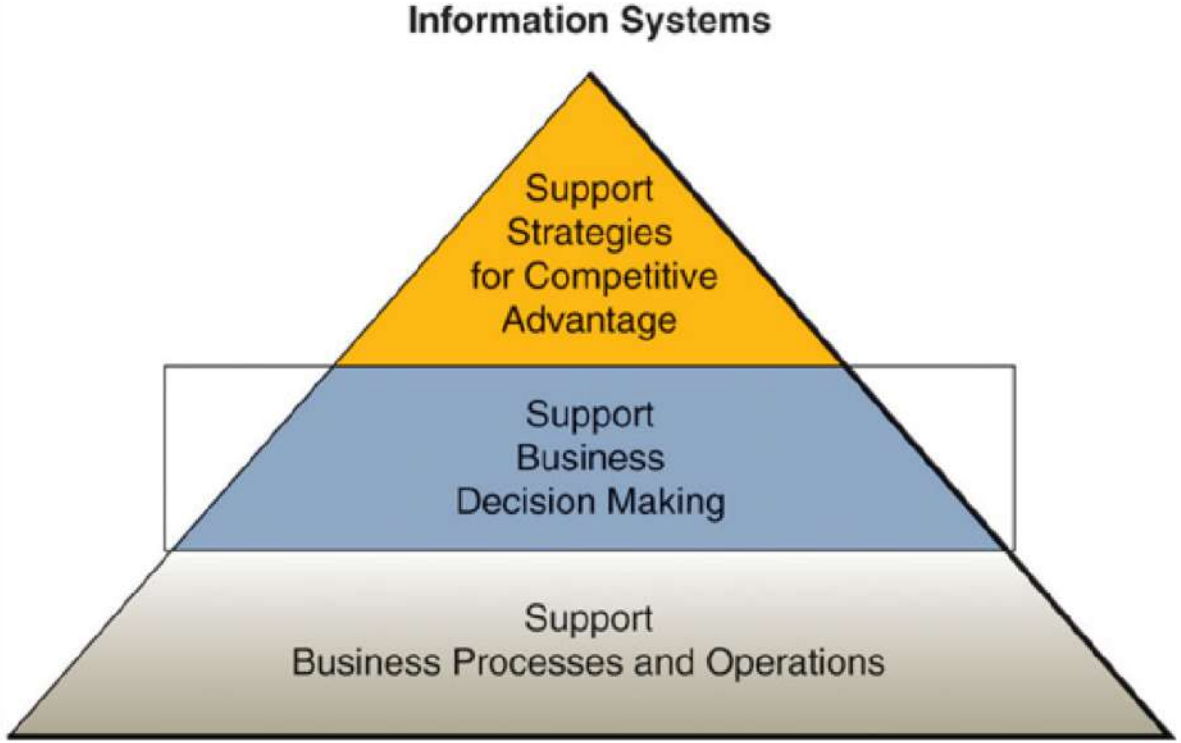
Data, Information, and Knowledge



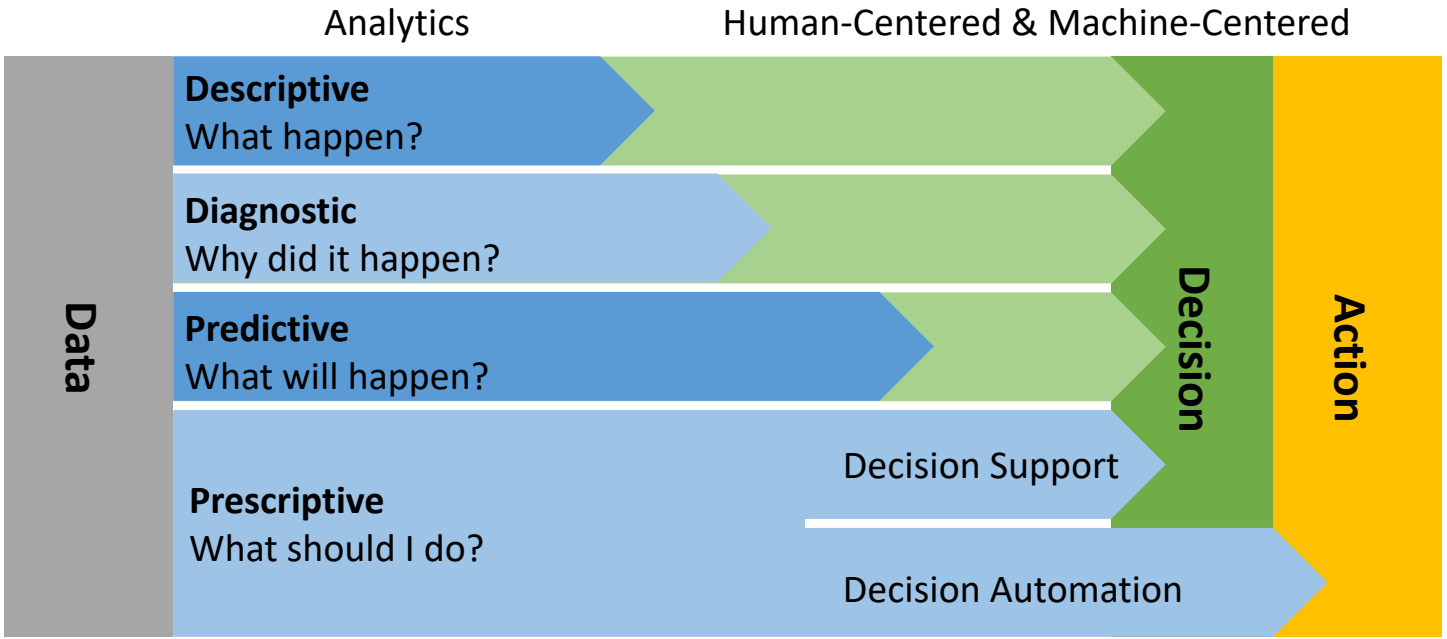
Feedback Value



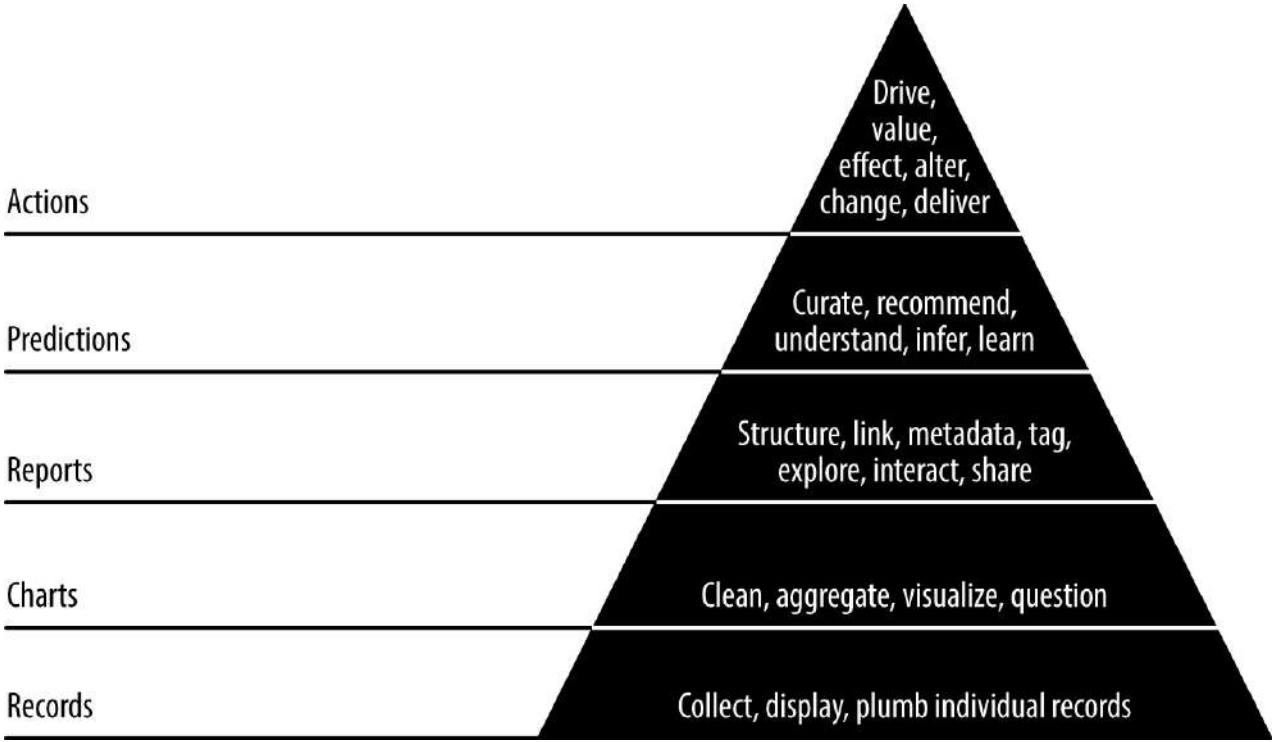
IS Pyramid


Information Systems
Support
Strategies
for Competitive
Advantage
Support
Business
Decision Making
Support
Business Processes and Operations

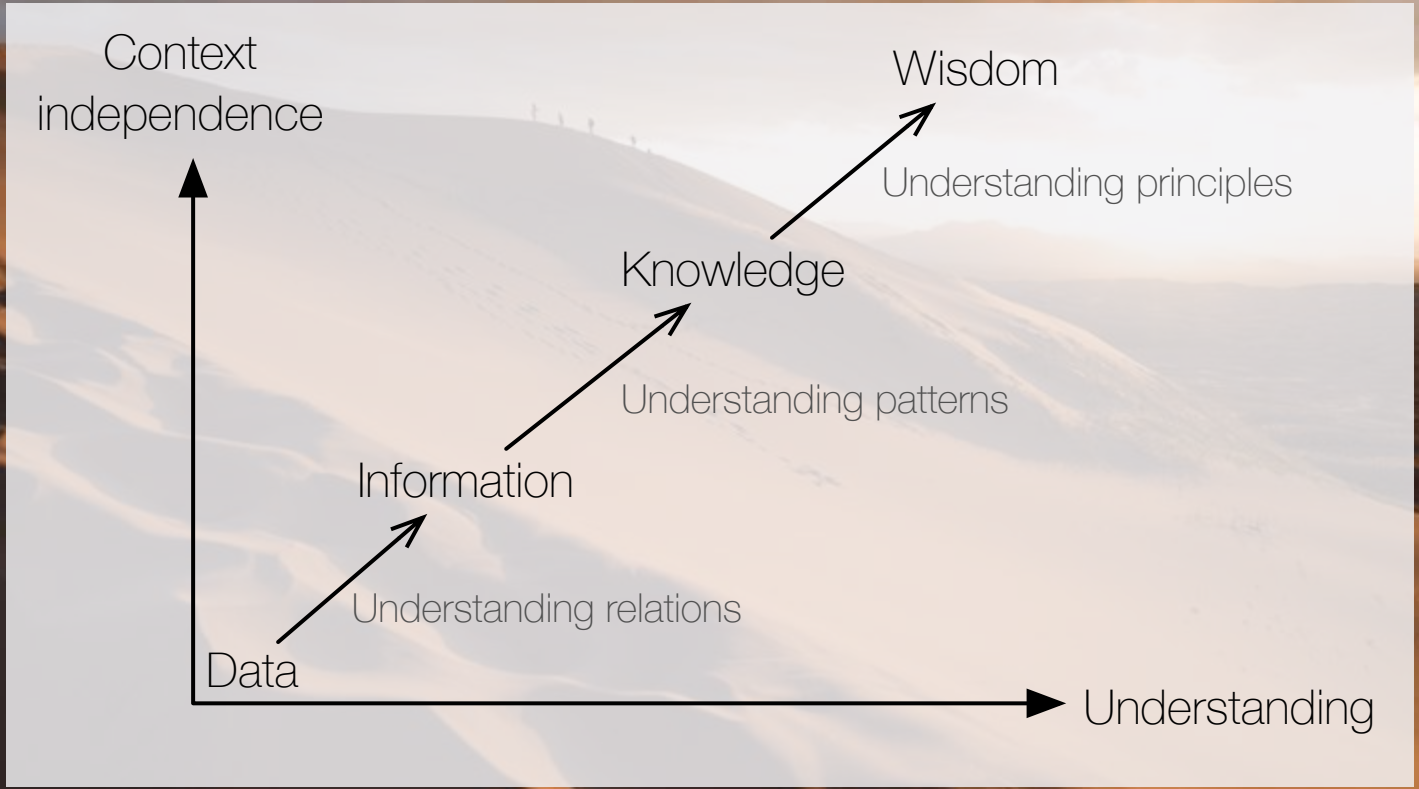
Data-driven decision (by Gartner)



Data-driven Value

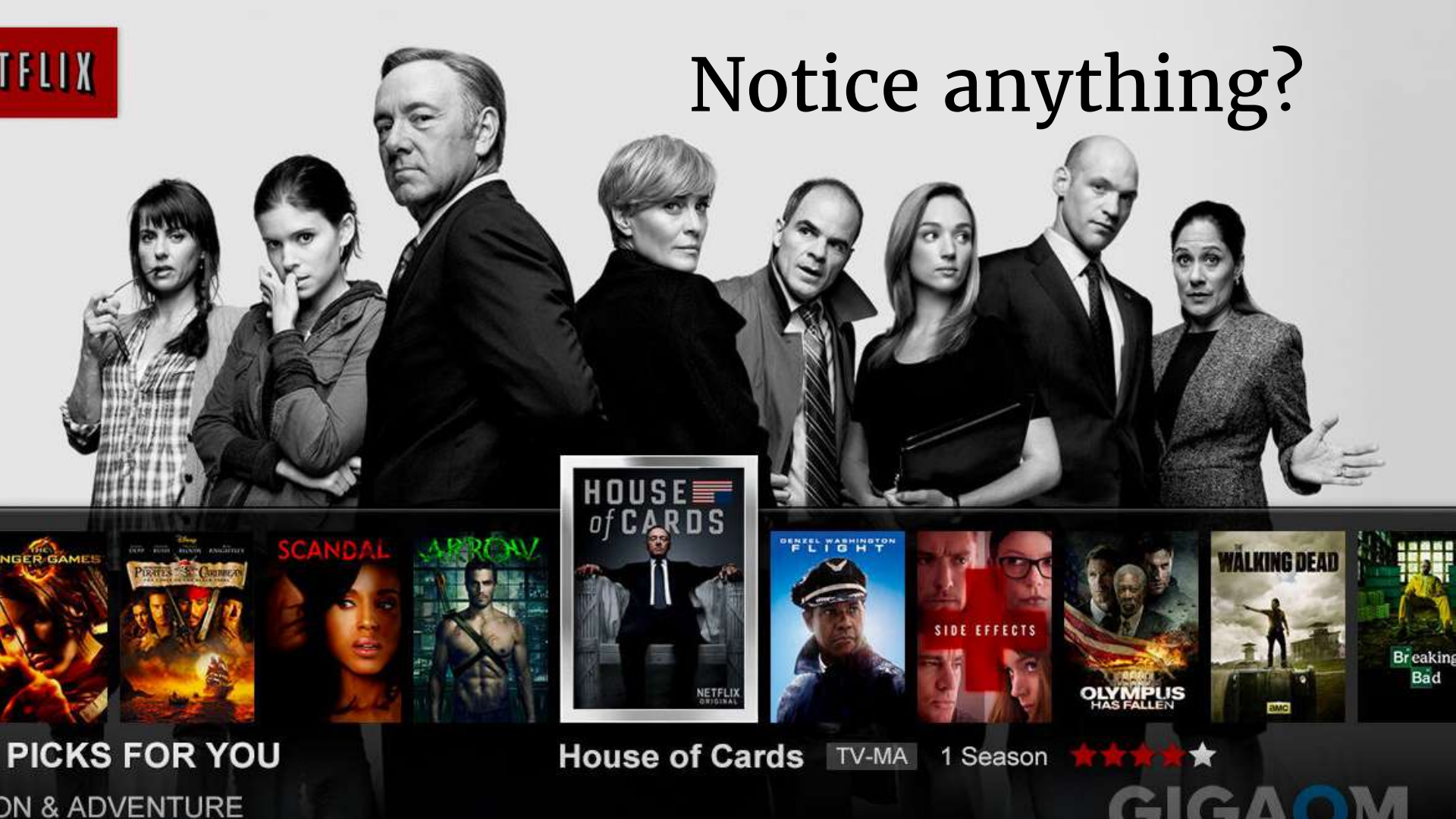


The Answer to the Great Question...
Of Life, the Universe and Everything



Big Data = Big Problems?

Notice anything?



Problems

Data access

Noise and trustworthiness

Algorithmic (and many other types of) bias

Cost of maintenance / problem solving