Intro to Big Data

ITU ASP COE TRAINING ON

“Developing the ICT ecosystem to harness IoTs”

Marco Zennaro, PhD

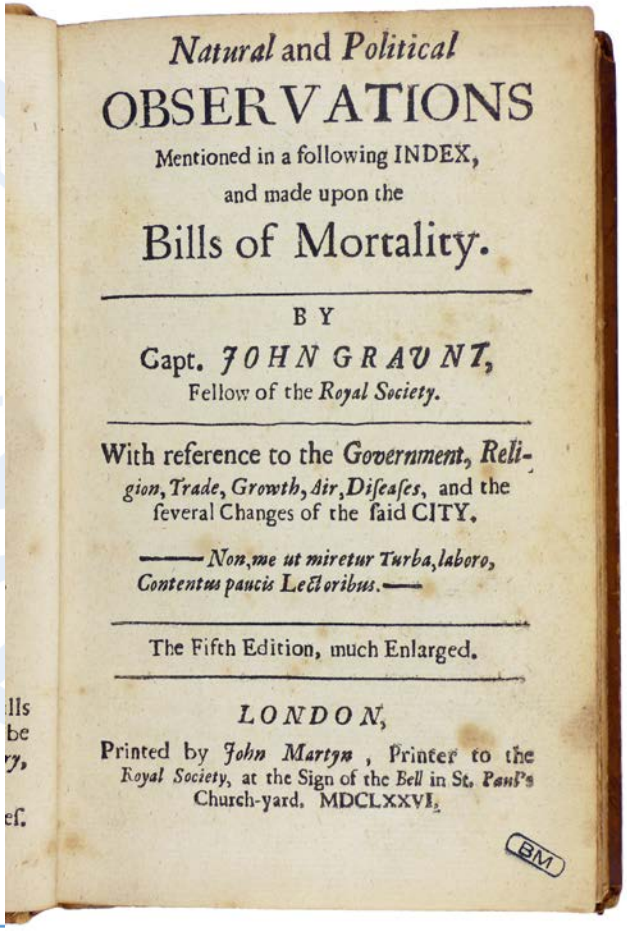
13-15 December 2016
Bangkok, Thailand

History of Big Data

- **300 BC – 48 AD** The Library of Alexandria is the world's largest data storage center – until it is destroyed by the Romans.
- 180,000 books, for a total of 20 Gb

History of Big Data

- **1663** John Graunt conducts the first recorded statistical-analysis experiments in an attempt to curb the spread of the bubonic plague in Europe.



Natural and *Political*
OBSERVATIONS
Mentioned in a following INDEX,
and made upon the
Bills of Mortality.

BY
Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Air*, *Diſeaſes*, and the ſeveral Changes of the ſaid CITY.

——*Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

LONDON,
Printed by *John Martyn*, Printer to the
Royal Society, at the Sign of the *Bell* in St. *Paul's*
Church-yard. MDCLXXVI.

BM

History of Big Data

- **1881** Herman Hollerith creates the Hollerith Tabulating Machine which uses punch cards to vastly reduce the workload of the US Census. He is one of the founders of IBM.
- **1926** Nikola Tesla predicts that in the future, a man will be able to access and analyze vast amounts of data using a device small enough to fit in his pocket.

History of Big Data

- **1965** The US Government plans the world's first data center to store 742 million tax returns and 175 million sets of fingerprints on magnetic tape.
- **1989** Early use of term Big Data in magazine article by fiction author Erik Larson – commenting on advertisers' use of data to target customers.

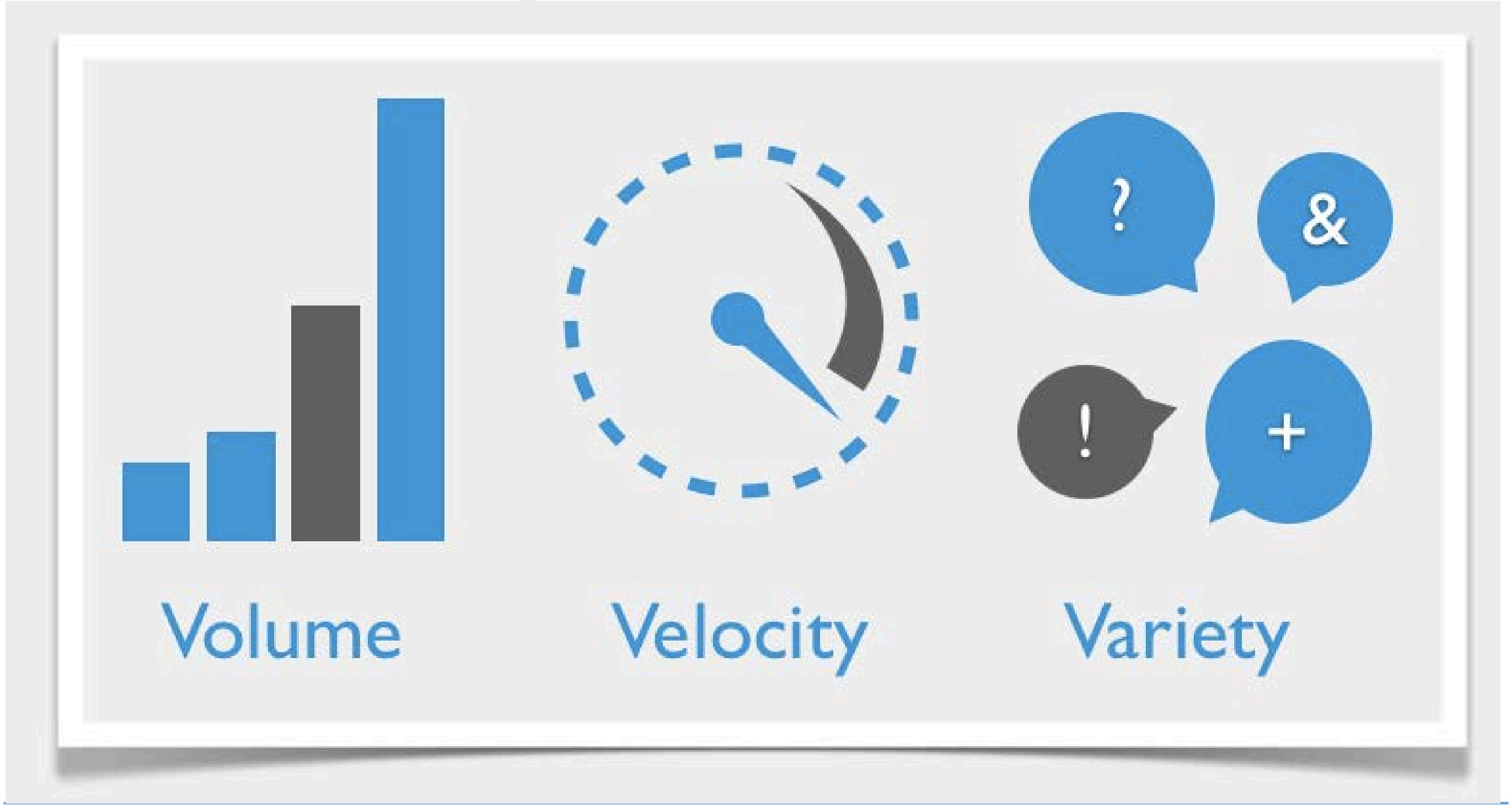
History of Big Data

- **2010** Eric Schmidt, executive chairman of Google, tells a conference that as much data is now being created every two days, as was created from the beginning of human civilization to the year 2003.

Definition

“Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.” – Wikipedia

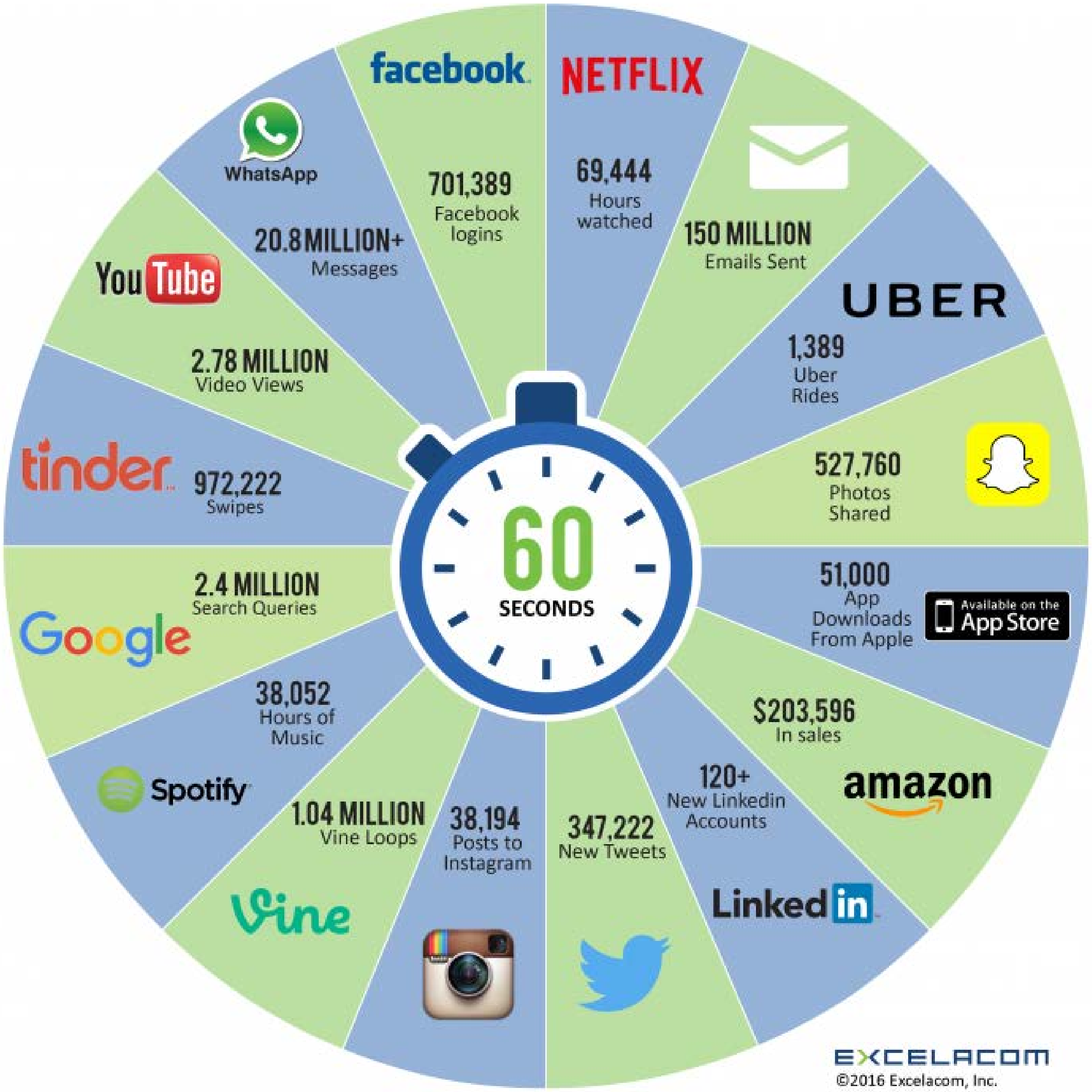
Three Characteristics of Big Data



Volume

- The sheer size of data in terms of storage and access.
- For example: unstructured data from social media in form of posts, video, audio with relational data such as comments, discussions, likes, etc.

2016 What happens in an INTERNET MINUTE?



facebook
701,389 Facebook logins

NETFLIX
69,444 Hours watched

150 MILLION Emails Sent

UBER
1,389 Uber Rides

527,760 Photos Shared

51,000 App Downloads From Apple
Available on the App Store

$203,596 In sales
amazon

120+ New Linkedin Accounts
LinkedIn

347,222 New Tweets

38,194 Posts to Instagram

1.04 MILLION Vine Loops
Vine

38,052 Hours of Music
Spotify

2.4 MILLION Search Queries
Google

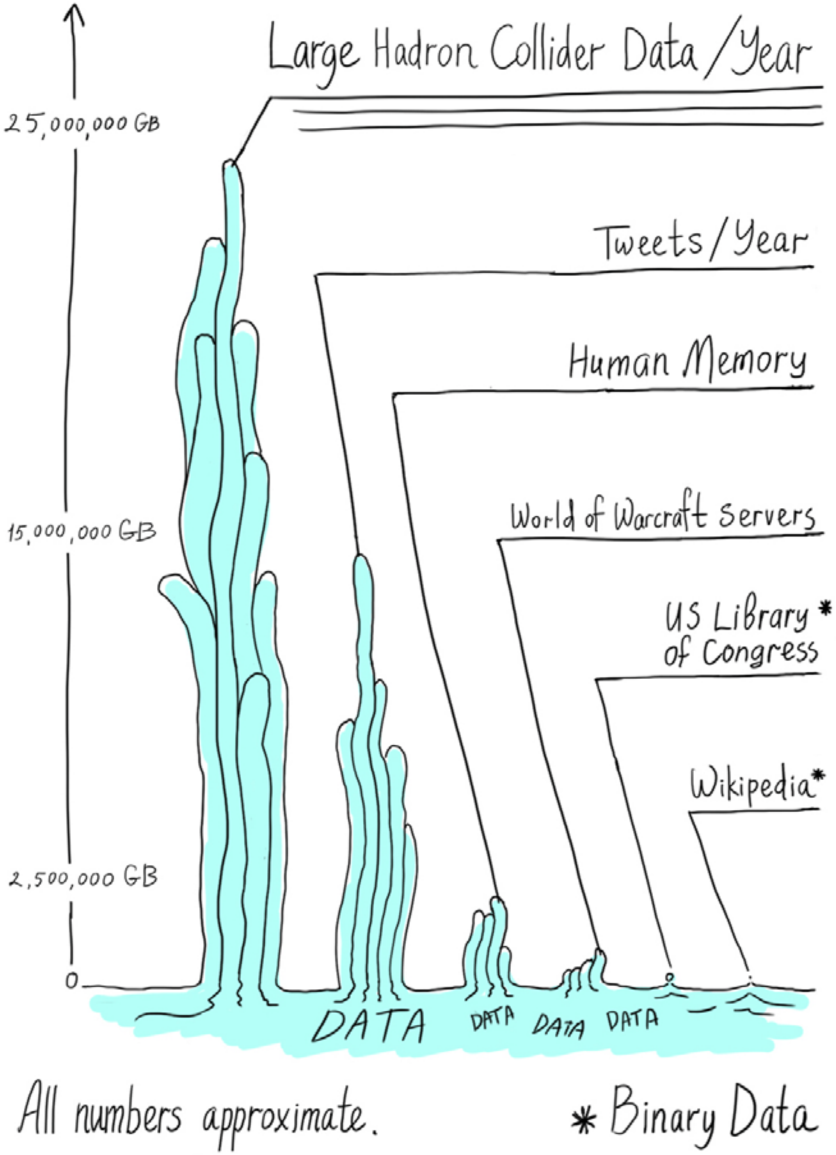
972,222 Swipes
tinder

2.78 MILLION Video Views
YouTube

20.8 MILLION+ Messages
WhatsApp

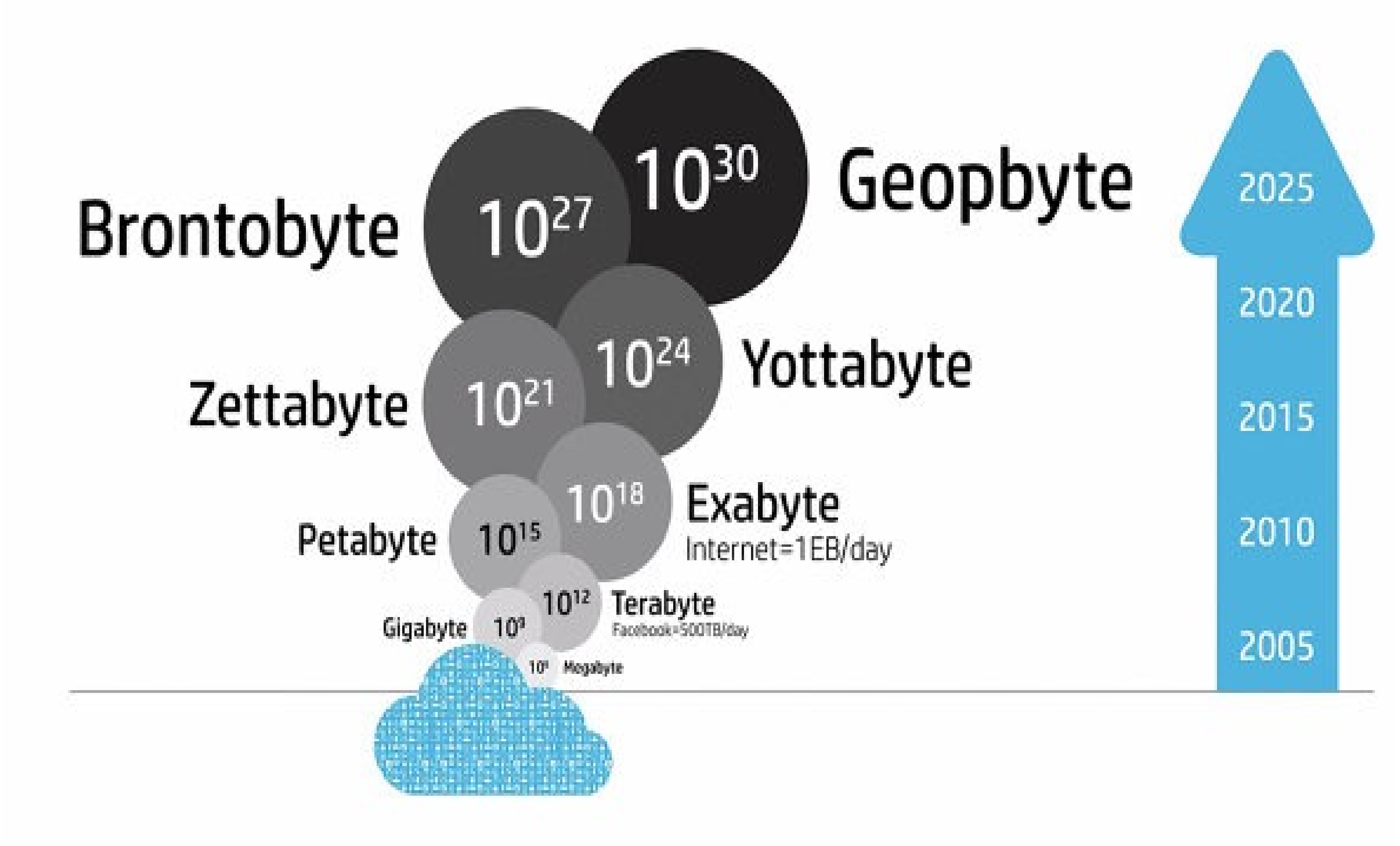
60 SECONDS

EXCELACOM
©2016 Excelacom, Inc.



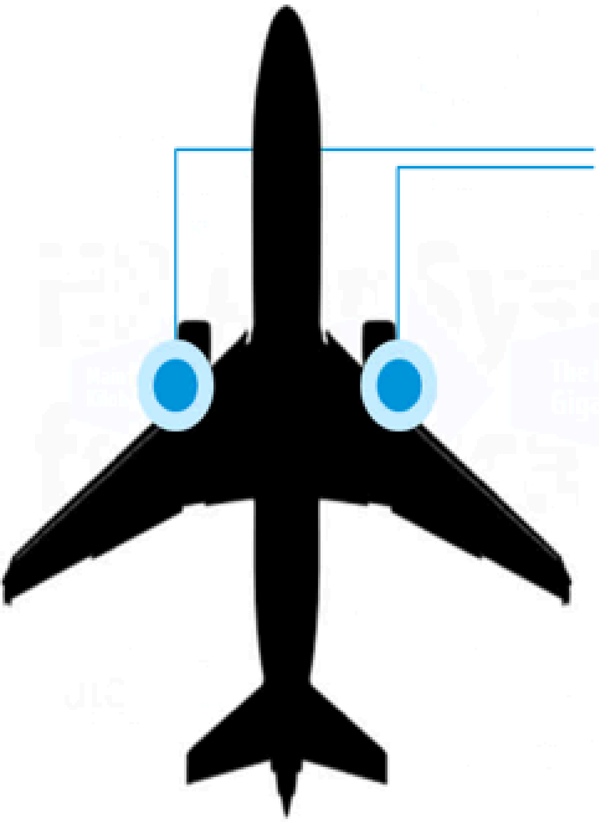
- The Large Hadron Collider (LHC) will generate 60 terabytes of data per day, 25 petabytes annually
- Wallmart generates 2.5 petabytes per hour

Volume



Brontobyte
10^{27}
10^{30}
Geopbyte
Zettabyte
10^{21}
10^{24}
Yottabyte
Petabyte
10^{15}
10^{18}
Exabyte
Internet=1EB/day
Gigabyte
10^{9}
10^{12}
Terabyte
Facebook=500TB/day
10^{6}
Megabyte
2025
2020
2015
2010
2005

Sensor data from a cross-country flight



20 TB × 2 × 6 × 28,537 × 365

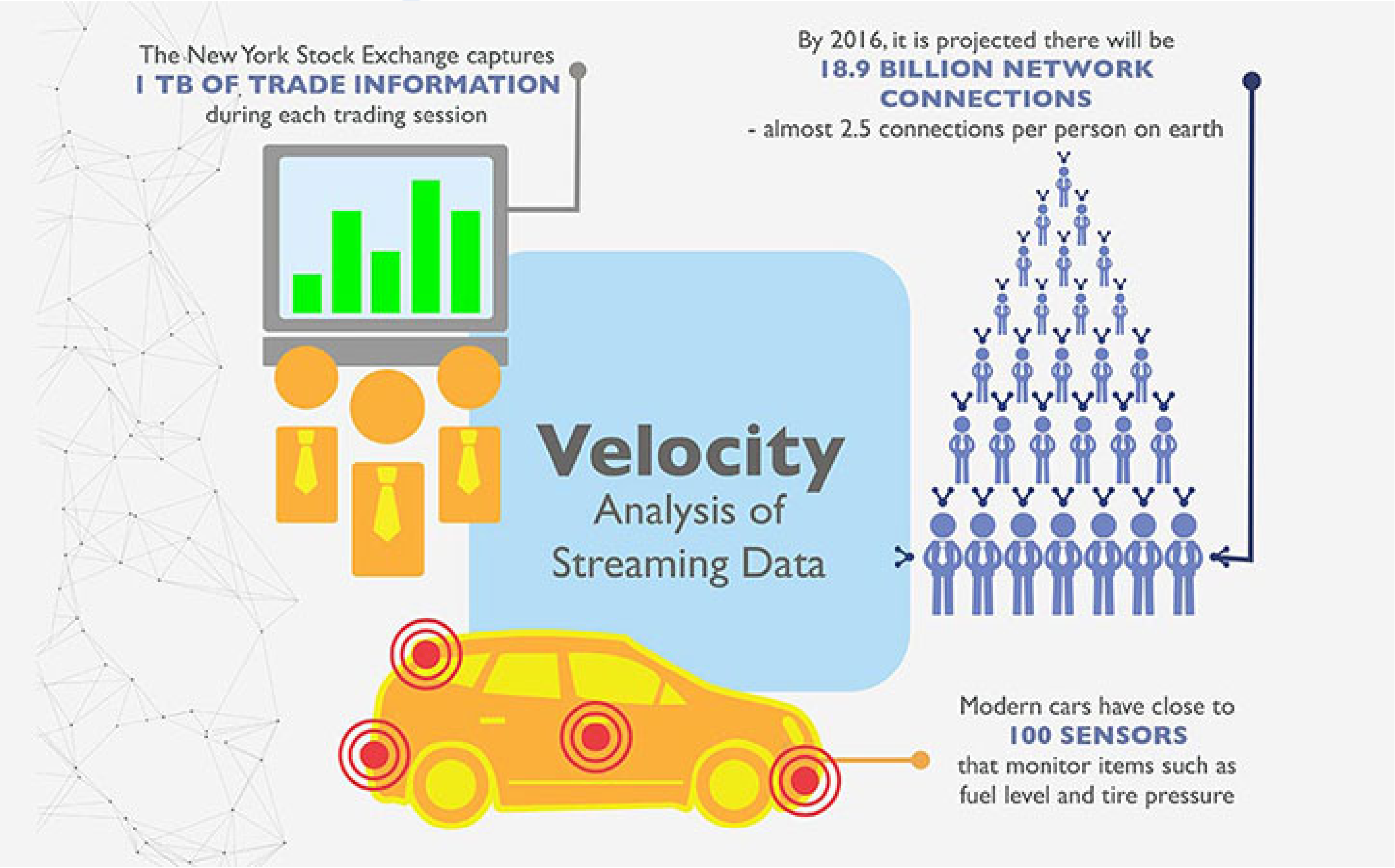
20 TB	2	6	28,537	365
20 terabytes of information per engine every hour	twin-engine Boeing 737	six-hour, cross-country flight from New York to Los Angeles	# of commercial flights in the sky in the United States on any given day.	days in a year

= 2,499,841,200 TB

Velocity

- The speed of incoming data and the time it takes to process it.
- With the advent of IoT, streaming data is driving the need to process and analyze data in near-real time.

Velocity



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

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Variety

- The type of files and format of data as well as sources.
- Data can be **structured,** such as a traditional database (pre-formatted data collected over time), or **unstructured** (unrelated data from unstructured sources such as social media, email, etc)

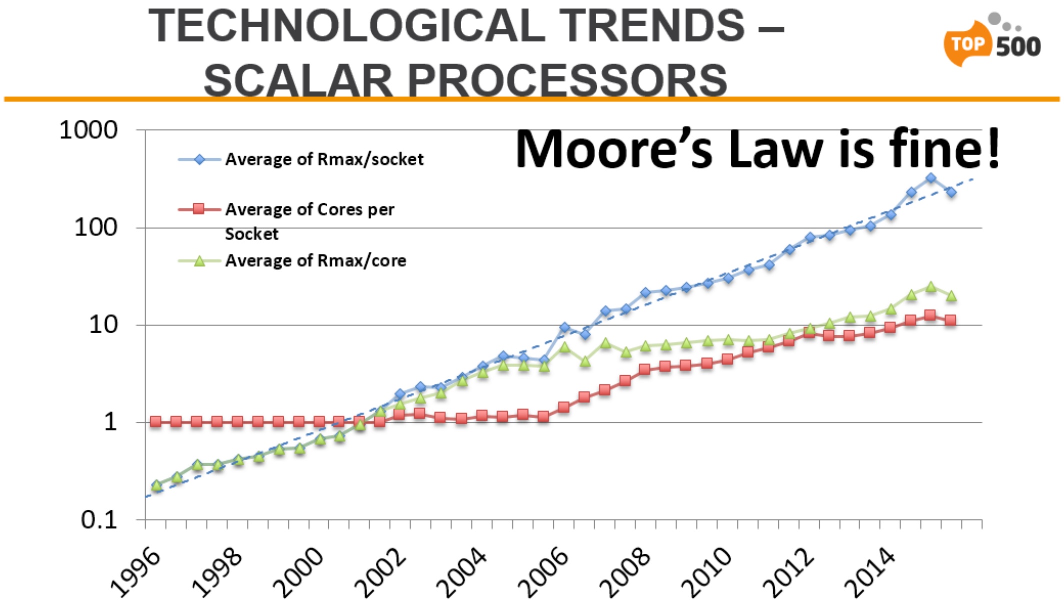
Data Definition Framework

Data Format

Data Source	Structured	Unstructured
Internal	**Human-Generated** • Survey ratings • Aptitude testing **Machine-Generated** • Web metrics from Web logs • Product purchase from sales Records • Process control measures	**Human-Generated** • Emails, letters, text messages • Audio transcripts • Customer comments • Voicemails • Corporate video/communications • Pictures, illustrations • Employee reviews
External	**Human-Generated** • Number of Retweets, Facebook likes, Google Plus +1s • Ratings on Yelp • Patient ratings **Machine-Generated** • GPS for tweets • Time of tweet/updates/postings	**Human-Generated** • Content of social media updates • Comments in online forums • Comments on Yelp • Video reviews • Pinterest images • Surveillance video

Big Data – Supporting Trends

- **Moore's Law**: an observation that the number of transistors on integrated circuits doubles every two years.

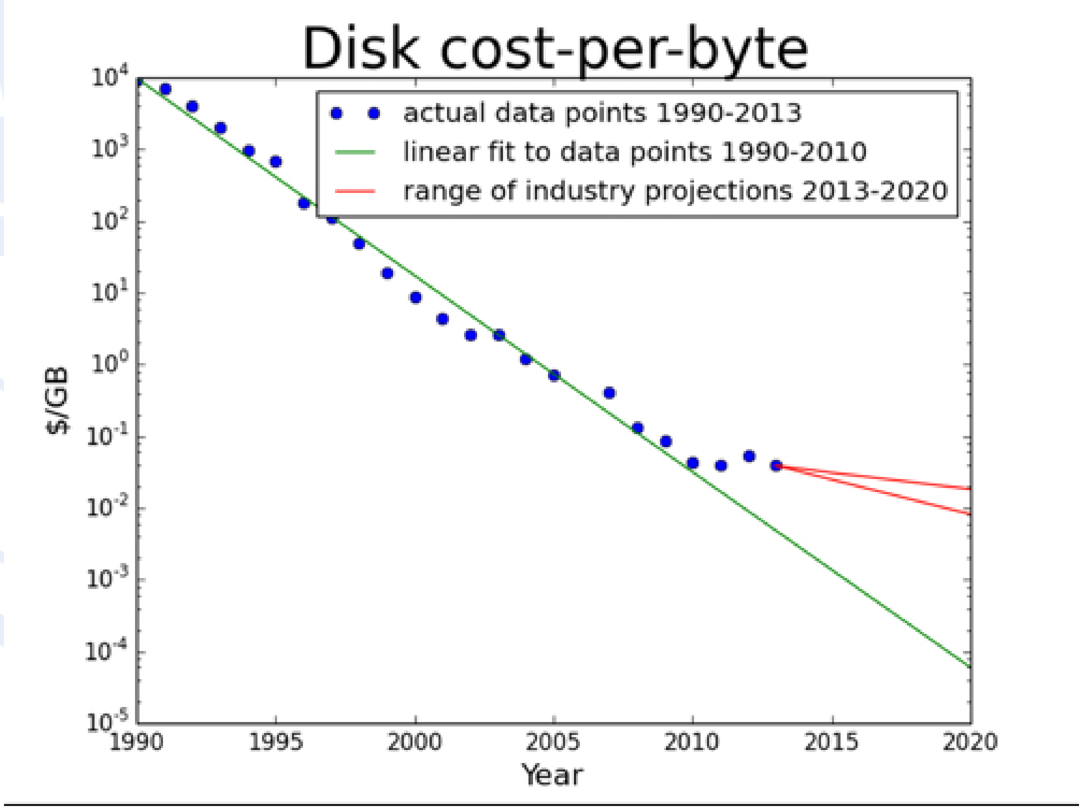


TECHNOLOGICAL TRENDS – SCALAR PROCESSORS

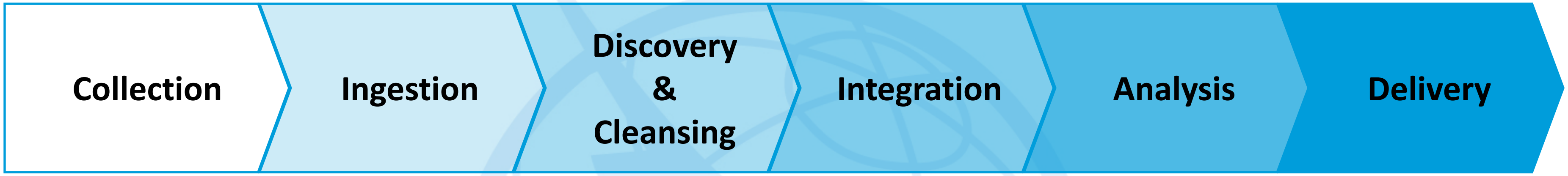
Moore’s Law is fine!

Big Data – Supporting Trends

- **Kryder's Law:** the density of storage is increasing and the cost decreasing at a rate faster than Moore's Law



Big Data Value Chain



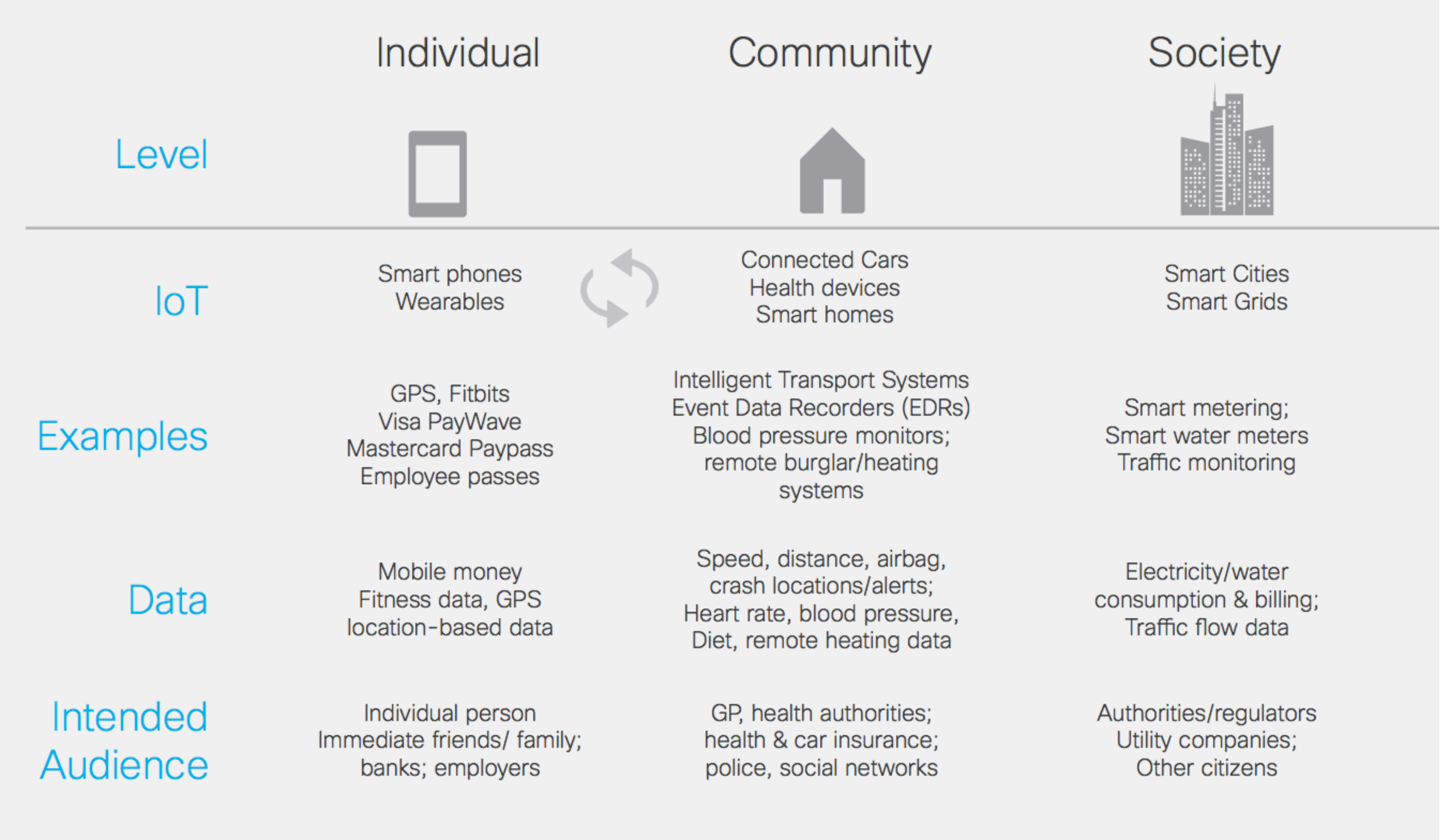
- **Collection** – Structured, unstructured and semi-structured data from multiple sources
- **Ingestion** – loading vast amounts of data onto a single data store
- **Discovery & Cleansing** – understanding format and content; clean up and formatting
- **Integration** – linking, entity extraction, entity resolution, indexing and data fusion
- **Analysis** – Intelligence, statistics, predictive and text analytics, machine learning
- **Delivery** – querying, visualization, real time delivery on enterprise-class availability

Source O’Reilly Strata 2012

Big Data – Tools

- **Hadoop** is often used at the server level to organise the cluster along with a NoSQL database for data storage.
- **NoSQL** are databases that use looser consistency models than relational databases. Performance gains via simplification using key value stores.

Examples of data generated by IoT



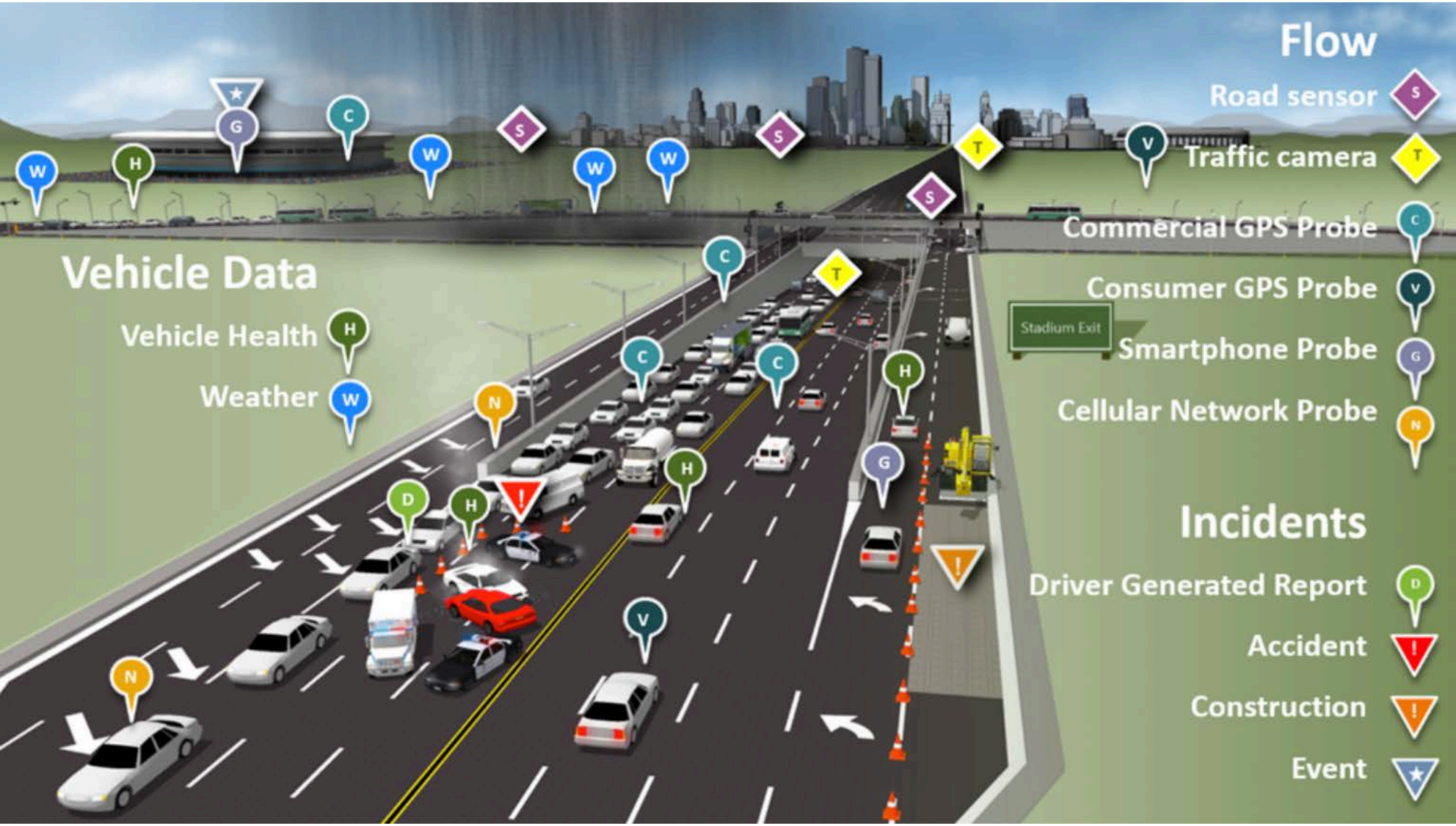
Level	Individual	Community	Society
IoT	Smart phones Wearables	Connected Cars Health devices Smart homes	Smart Cities Smart Grids
Examples	GPS, Fitbits Visa PayWave Mastercard Paypass Employee passes	Intelligent Transport Systems Event Data Recorders (EDRs) Blood pressure monitors; remote burglar/heating systems	Smart metering; Smart water meters Traffic monitoring
Data	Mobile money Fitness data, GPS location-based data	Speed, distance, airbag, crash locations/alerts; Heart rate, blood pressure, Diet, remote heating data	Electricity/water consumption & billing; Traffic flow data

Intended Audience	Individual person
	Immediate friends/ family;
	banks; employers
		GP, health authorities;
		health & car insurance;
		police, social networks
			Authorities/regulators
			Utility companies;
			Other citizens

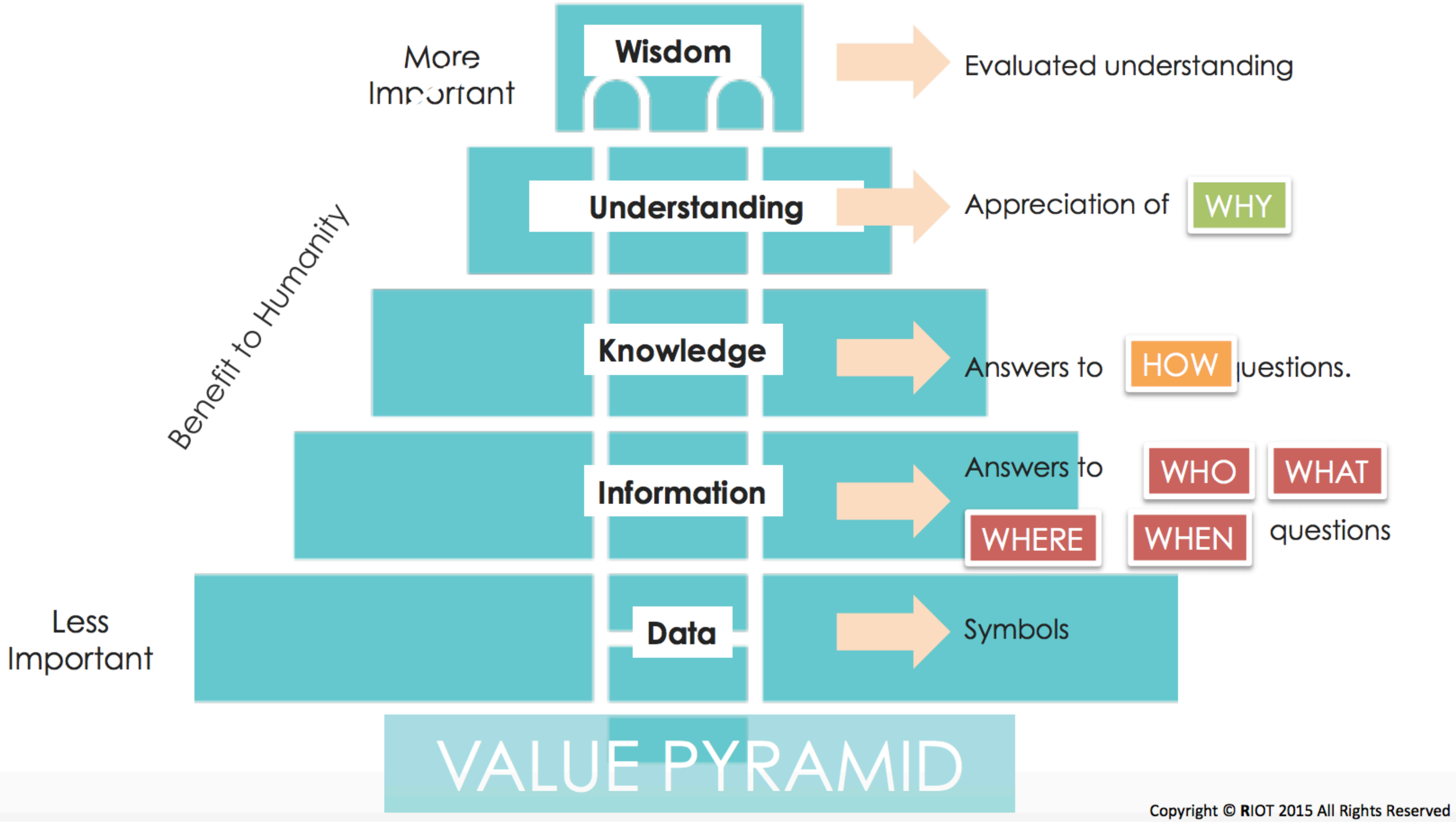
Source: Adapted from the ITU Draft GSR Discussion Paper 2015, “Regulation and the Internet of Things”,

Usecase: Traffic

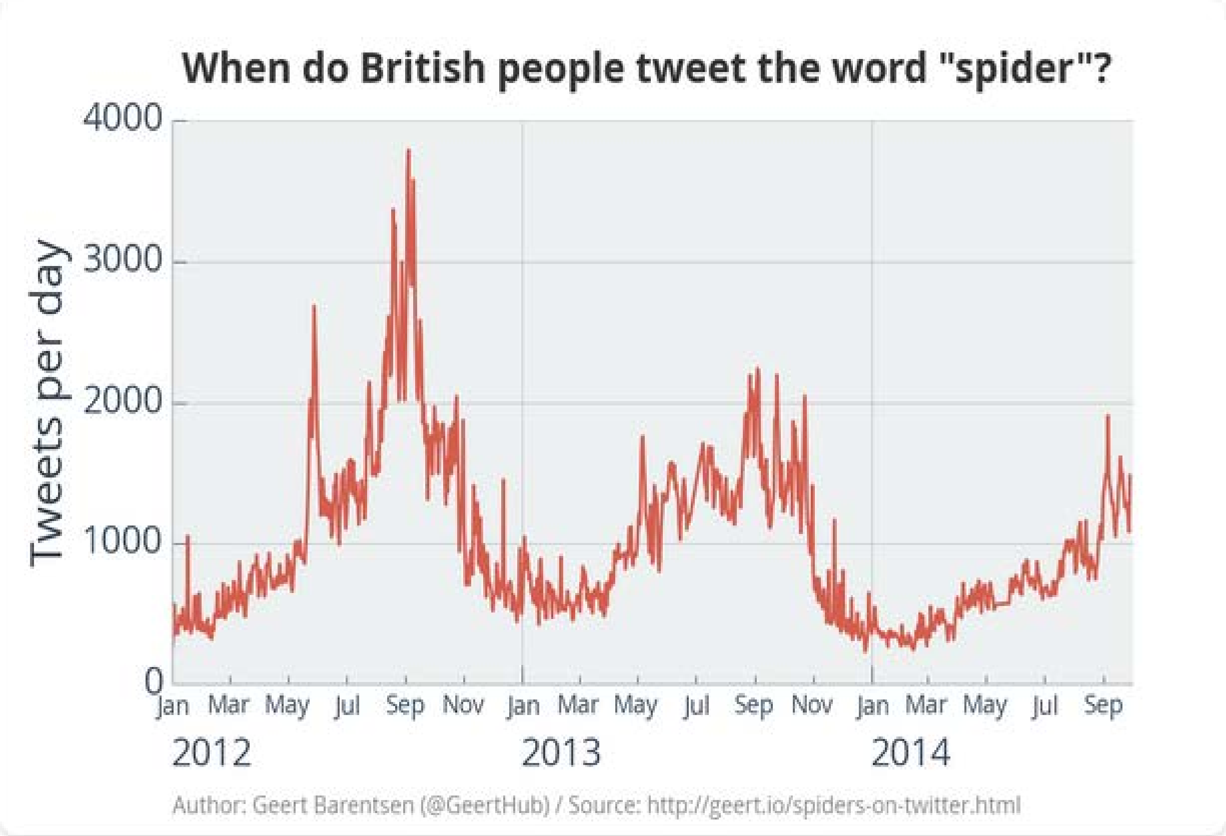
- **Collect** traffic data and transportation data from sensors
- **Build** a model of traffic patterns
- **Predict** the traffic and congestions
- **Act**: divert traffic, adjust troll, adjust traffic lights



Source: http://inrix.com/



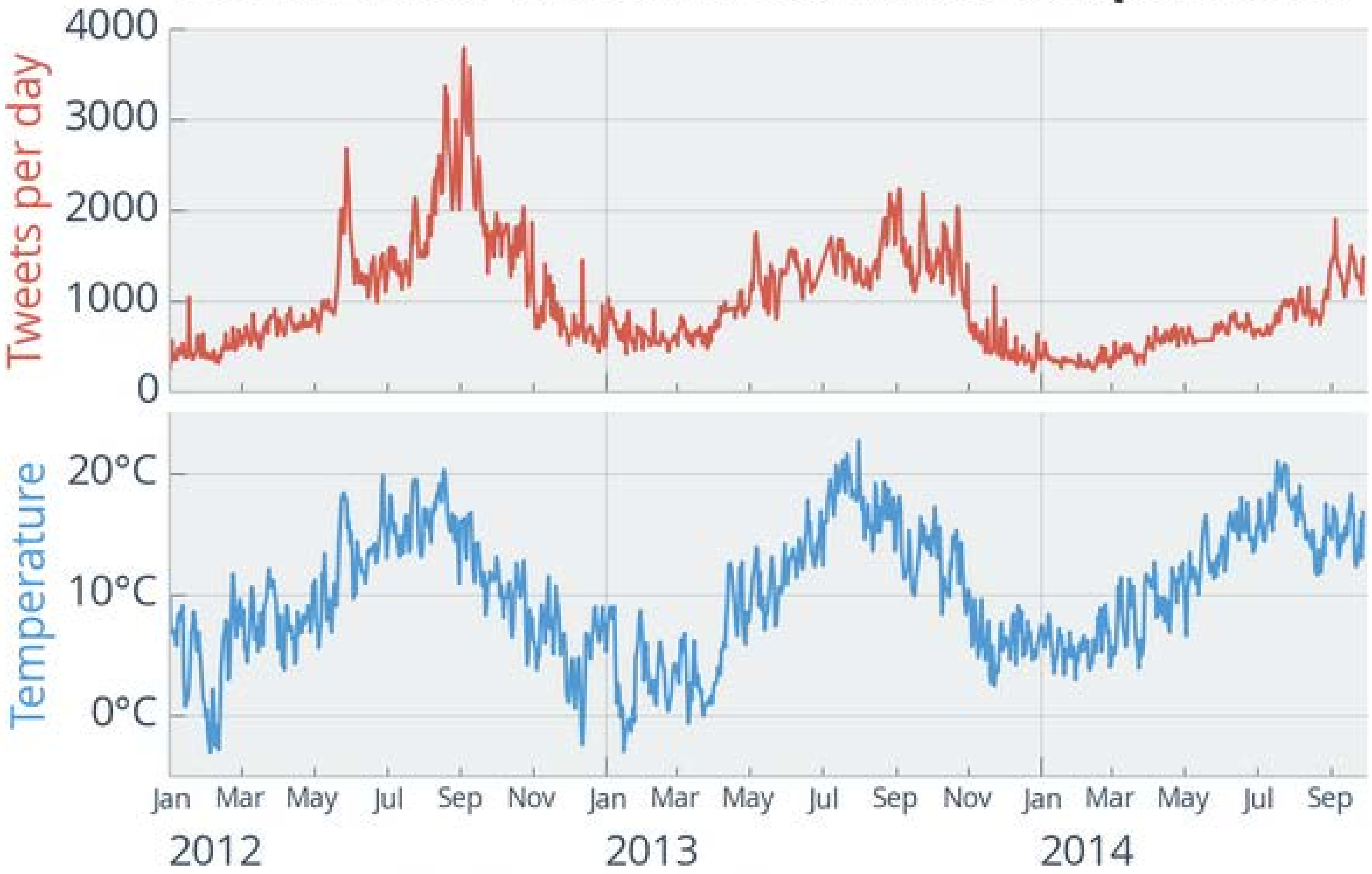
Source: REDtone IOT



When do British people tweet the word "spider"?

Author: Geert Barentsen (@GeertHub) / Source: http://geert.io/spiders-on-twitter.html

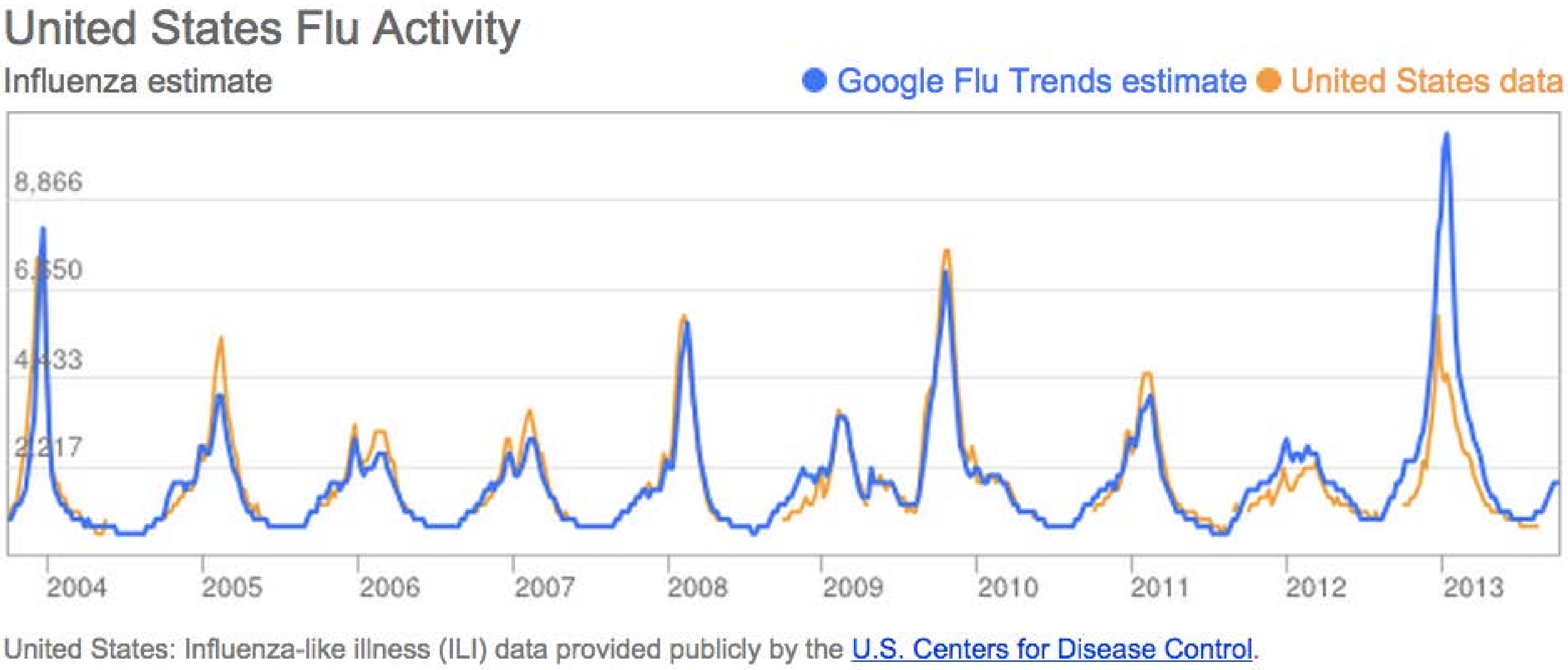
Number of tweets per day in Britain that contained the word "spider". Retweets, replies, and tweets about Spider-Man have been excluded.



The number of tweets about spiders in Britain is correlated with the local mean temperature:

Author: Geert Barentsen (@GeertHub) / Source: http://geert.io/spiders-on-twitter.html

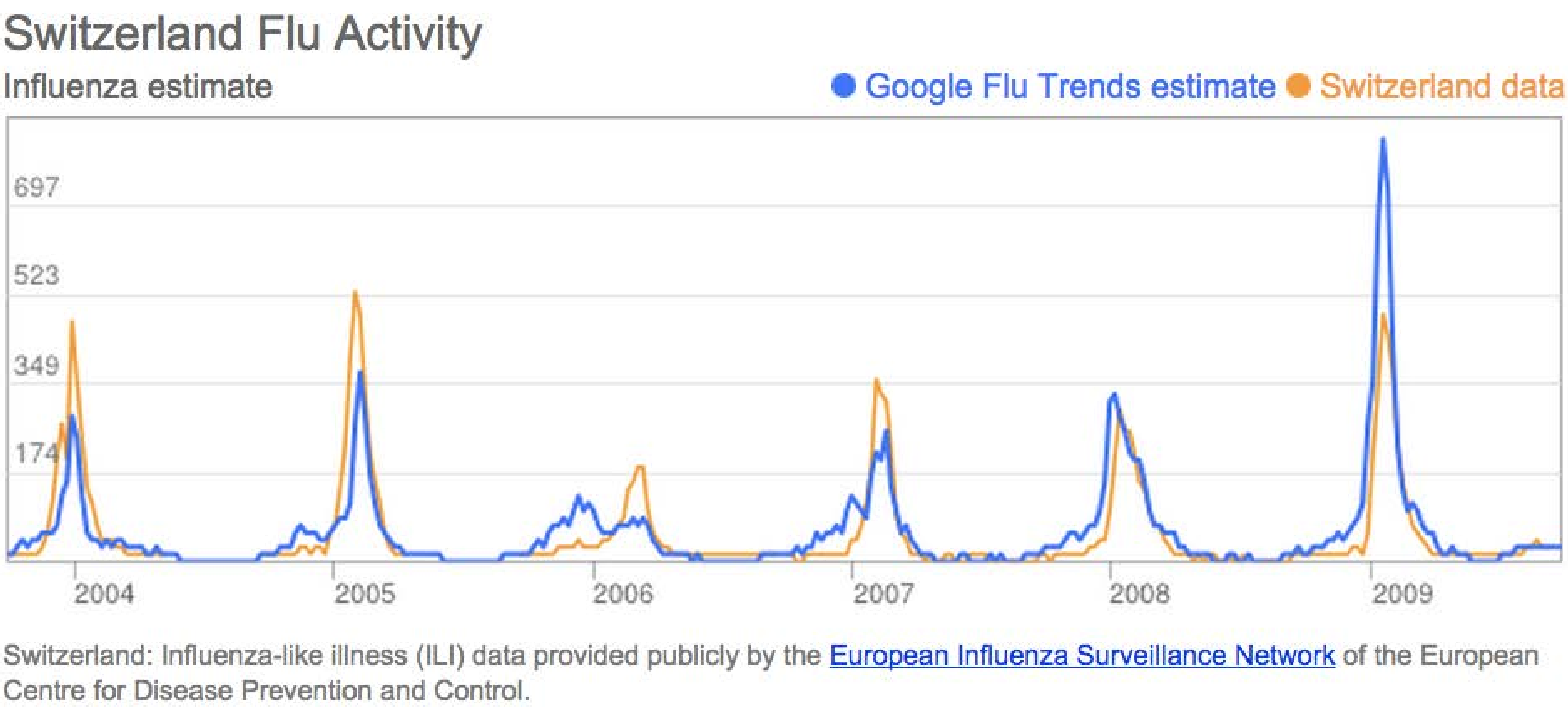
Number of British spider tweets per day (top), shown against the mean temperature in Central England obtained from the UK Met Office (bottom).



United States Flu Activity

Influenza estimate

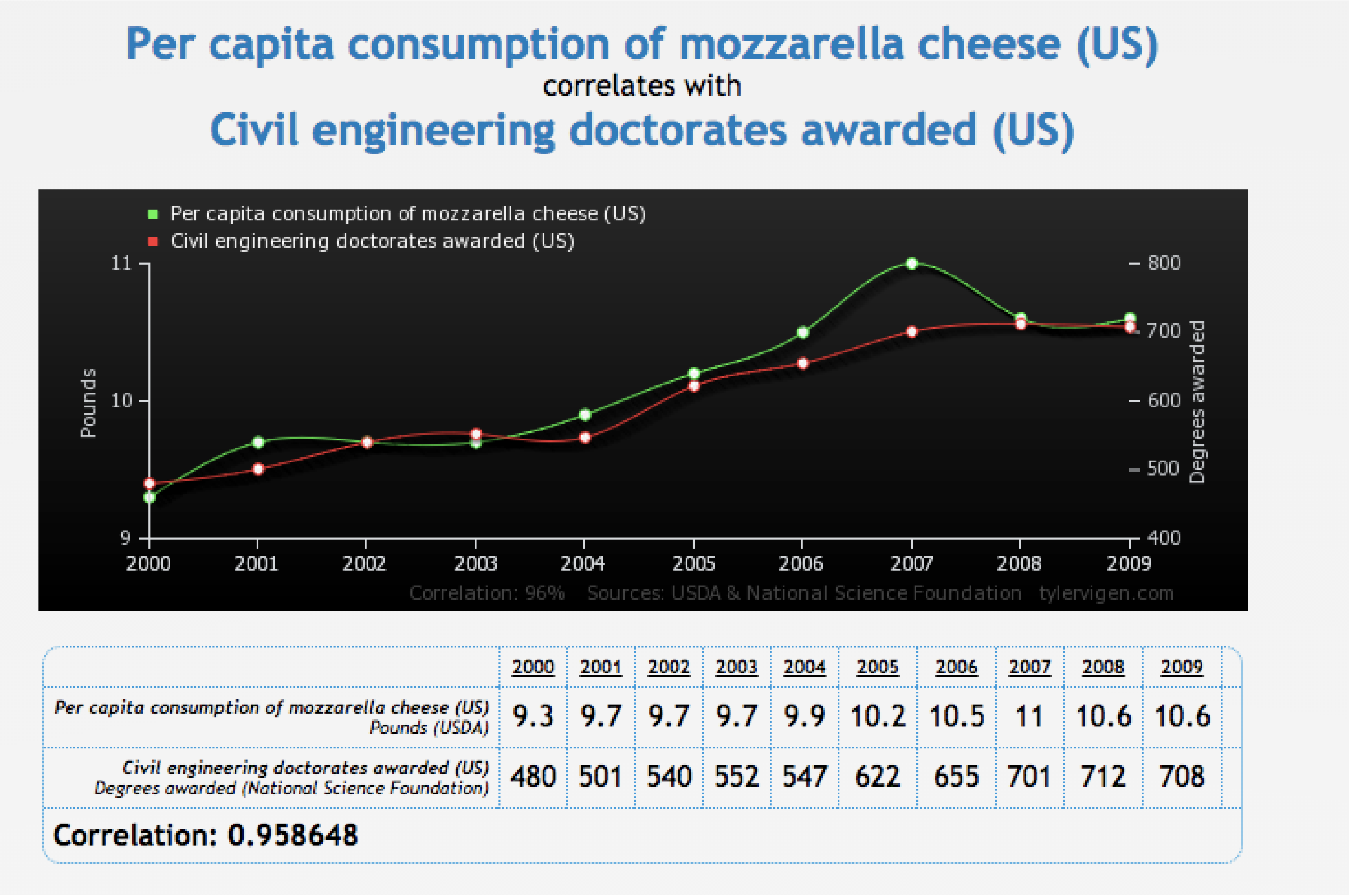
United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.



Switzerland Flu Activity

Influenza estimate

Switzerland: Influenza-like illness (ILI) data provided publicly by the European Influenza Surveillance Network of the European Centre for Disease Prevention and Control.



Per capita consumption of mozzarella cheese (US)

correlates with

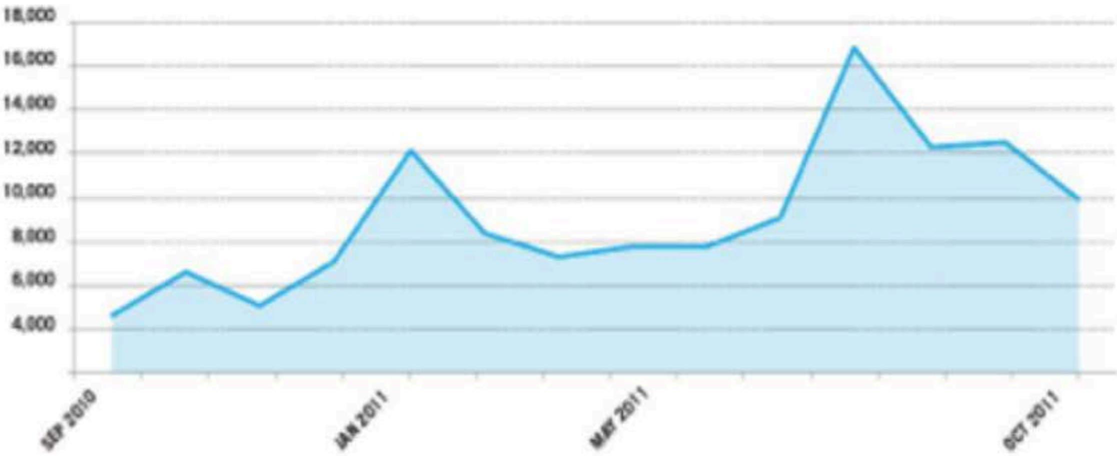
Civil engineering doctorates awarded (US)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of mozzarella cheese (US) *Pounds (USDA)*	9.3	9.7	9.7	9.7	9.9	10.2	10.5	11	10.6	10.6
Civil engineering doctorates awarded (US) *Degrees awarded (National Science Foundation)*	480	501	540	552	547	622	655	701	712	708

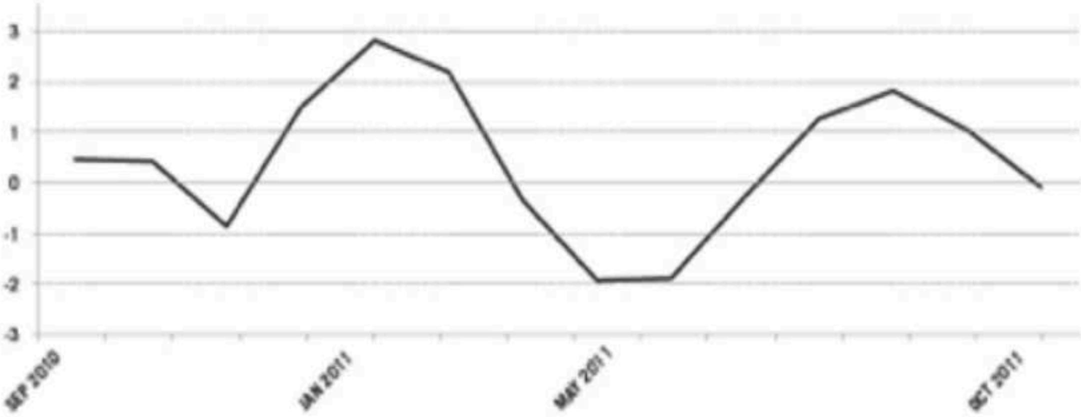
Correlation: 0.958648

Source: http://tylervigen.com/

Tweets about the price of rice
(per month)

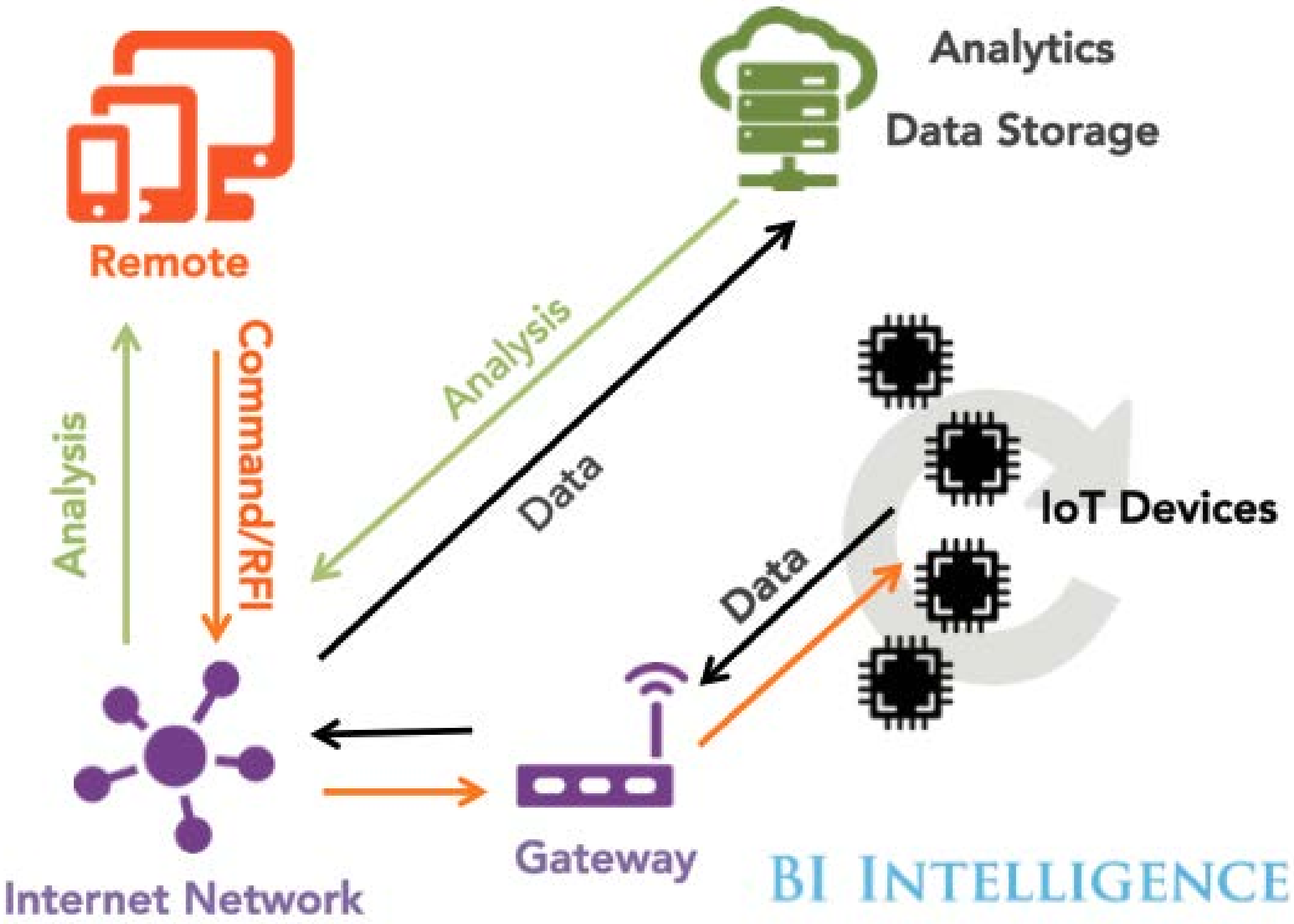


Food Price Inflation



Source:UN Global Pulse

The Internet of Things Ecosystem



Demo

http://discover-iot.eu-gb.mybluemix.net/#/play/device/smartphone

Thank You