

Introduction to Big Data

Second International School on Open Science Cloud –
Perugia – Italy

Daniele Cesini INFN-CNAF

+ Outline

- This presentation is about:
 - Big Data definition
 - Big Data examples
 - Big Data computing infrastructures

- This presentation is **NOT** about
 - Big Data analytics technologies and strategies
 - No map-reduce
 - No Hadoop
 - No spark
 - Ethics and privacy issues about Big Data

+ Let's start with an example....

3



Bill Gates ✓
@BillGates

Follow

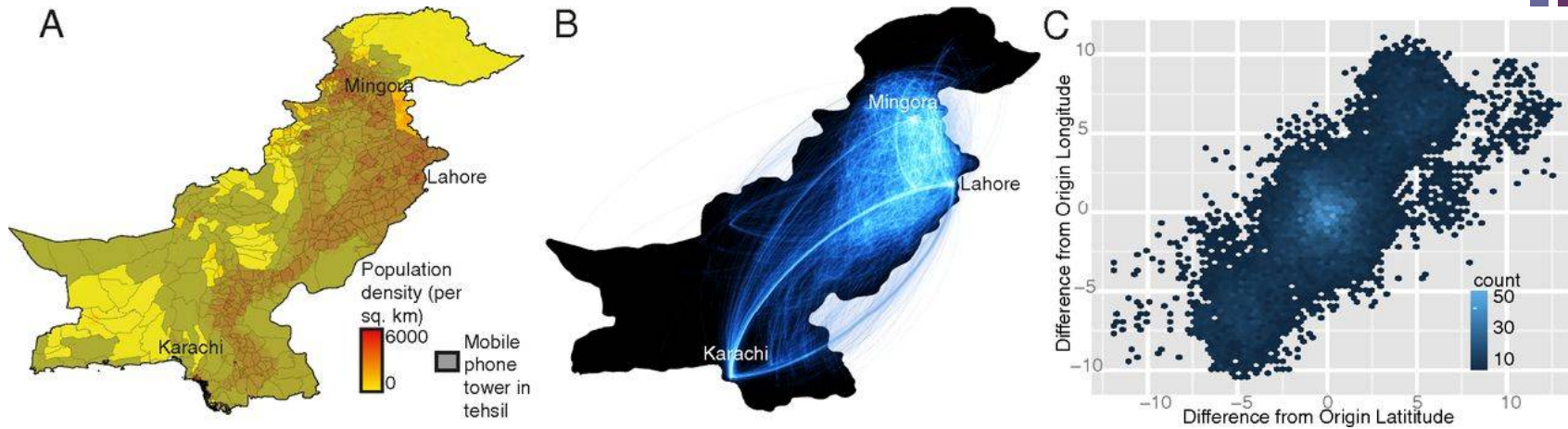
Cellphone records could help doctors predict which places might be hit by dengue: [b-gat.es/1Obnehb](https://twitter.com/billgates/status/1028888888888888888)



Paper available at: <http://www.pnas.org/content/112/38/11887>



Human mobility dynamics in Pakistan



- (A) Population density and mobile phone tower coverage from the mobile phone operator in Pakistan tehsil.
- (B) The top routes of travel between pairs of tehsils in Pakistan. A line is drawn if at least 20,000 trips occurred between the origin and destination between June and December 2013. The top routes occur between Karachi and cities in northern Punjab province, particularly Lahore tehsil.
- (C) Relative direction and volume of travel.

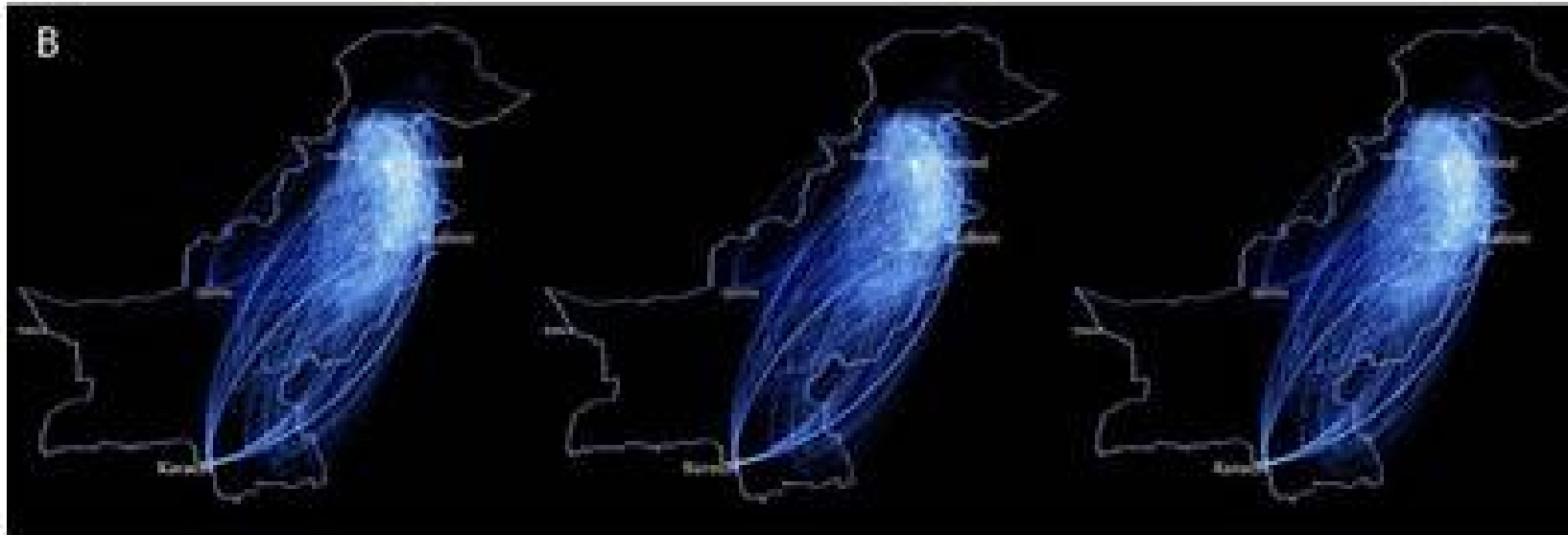


Amy Wesolowski et al. PNAS 2015;112:38:11887-11892

+

Human mobility dynamics in Pakistan: the cell phone data

5

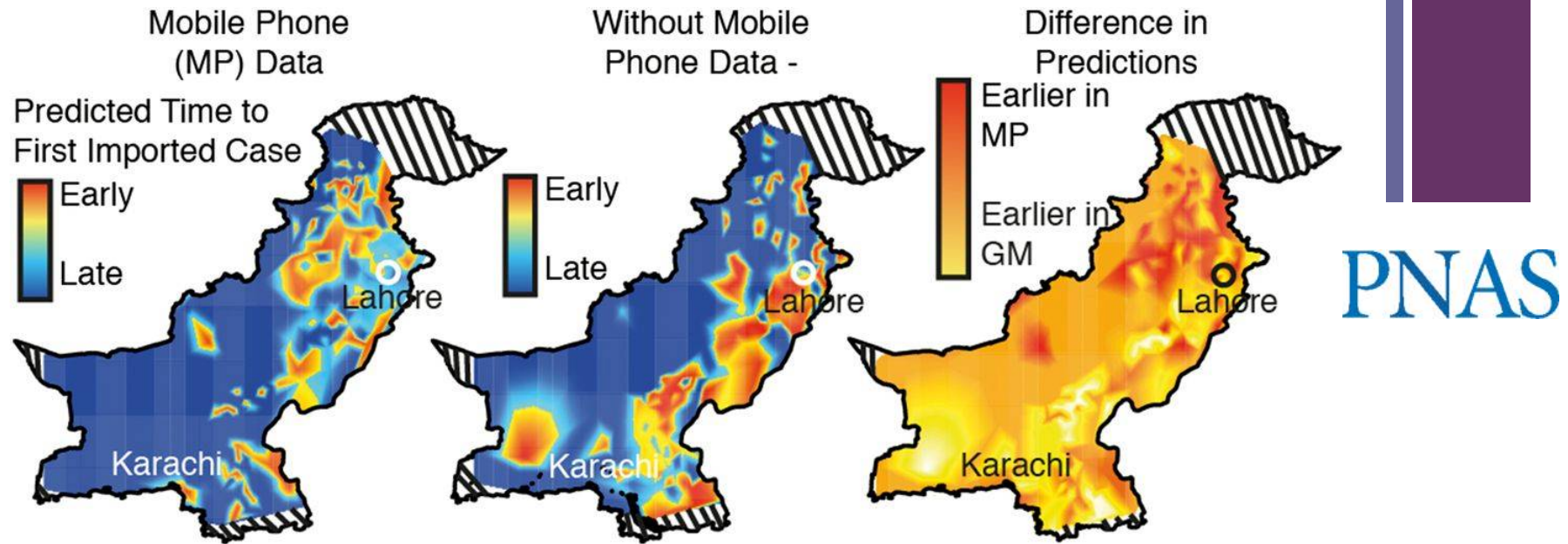


- The most frequently traveled routes based on the mobile phone data. The figure compare the most traveled routes in different period of the year, before, during and after Ramadan
- 40 MILLIONS OF CELL PHONES WERE USED

PNAS

Amy Wesolowski et al. PNAS 2015;112:38:11887-11892

+ Dengue diffusion model



- Dengue diffusion model prediction using and not using the cell phone data
- “Mobile phone data provide dynamic population mobility estimates that can be combined with infectious disease surveillance data and seasonally varying environmental data to map these changing patterns of vulnerability in a country where dengue outbreaks are emerging and irregular in many regions”

Amy Wesolowski et al. PNAS 2015;112:38:11887-11892

+ Another example...



Official Blog

Insights from Googlers into our products, technology, and the Google culture

Stuck in traffic?

February 28, 2007

Posted by David Wang, Software Engineer

There's nothing worse than getting stuck in traffic when you have some place to go, so I'm happy to tell you about a new feature on [Google Maps](#) that can help. For more than 30 major U.S. cities, you can now see up-to-date traffic conditions to help you plan your schedule and route. If you're in [San Francisco](#), [New York](#), [Chicago](#), [Dallas](#), or any of the other cities we now include, just click on the traffic button to show current traffic speeds directly on the map. If your route shows red, you're looking at a stop-and-go commute; yellow, you could be a little late for dinner; green, you've got smooth sailing.

We can't make traffic go away, but we hope Google Maps traffic info helps you avoid it whenever possible.



7 min

18/09/2018

+ Google Traffic

- Google Traffic is a feature on Google Maps that displays traffic conditions in **real time** on major roads and highway
- Works by analyzing the GPS-determined locations transmitted to Google by a **large number** of **mobile phone users**
- By calculating the speed of users along a length of road, Google is able to **generate a live traffic map**
- Google has stated that the speed and location information it collects to calculate traffic conditions is **anonymous**.
- Excludes **anomalies such as postal vehicles** which make frequent stops



Traffic congestion on a snowy Wednesday night in Washington in 2016

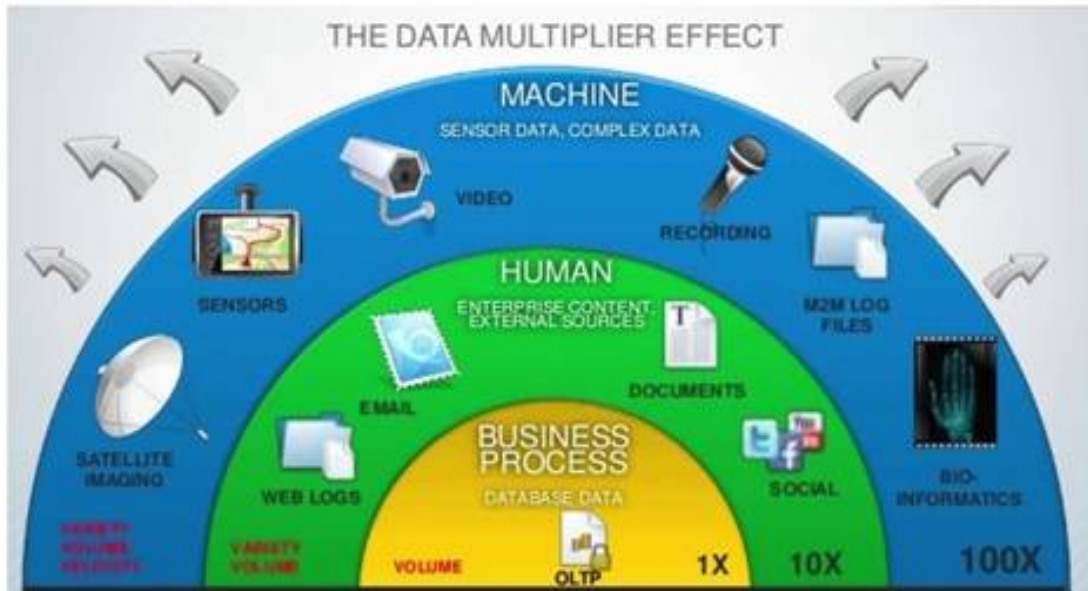
© Source: Wikipedia

+ What do they have in common?

9

- They process a lot of data from a lot of sources
- Data are coming without a pre-defined schedule
- They need a technology that is able to process quickly that amount of data in order to take actions
- They are based on just 5 numbers.....
 - Time and geographic position
 - An ID of the source
-and can extract information that greatly impact on our lives

+ Data explosion



Annual size of global data



+ Data explosion - one minute on the Internet

2017 *This Is What Happens In An Internet Minute*



+ It's all about Vs – part 1

■ Volume

- Large amounts of data
- Typically sizes ranges from several terabytes to zettabyte



+ It's all about Vs – part 2

■ Velocity

- Data from transactions with high refresh rate
- Data streams coming at great speed
- Short time to react
 - from batch processing to real time streaming



+ It's all about Vs - part3

■ Variety:

- Data come from different data sources
- Data can come in various formats
 - **Structured** data as database table
 - **Semi-structured** data such as XML data
 - **Unstructured** data
 - Text
 - Images
 - Video streams
 - Audio recording
 - There is a shift from sole structured data to increasingly more unstructured data



+ The Gartner definition

15

- Most widely used definition in the industry by Gartner (2012) “Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.”
- **It should by now be clear that the “big” in Big Data is not just about volume**
 - ...but certainly involves having a lot of Data
- You are not only getting a lot of data
 - It is also coming at you fast
 - It is coming at you in complex format
 - It is coming at you from a variety of sources

+ Another definition...

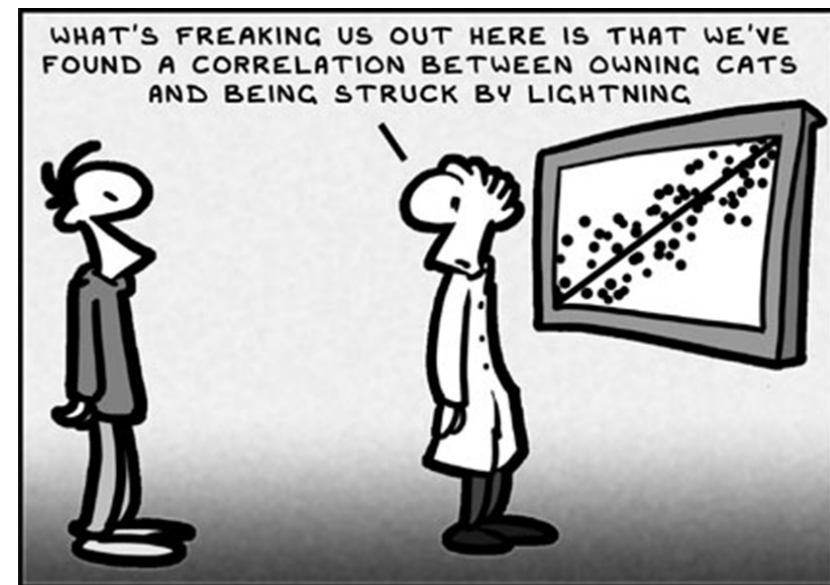
O' Reilly Strata group states that

“Big Data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your data database architectures”

+ It's all about Vs - part4

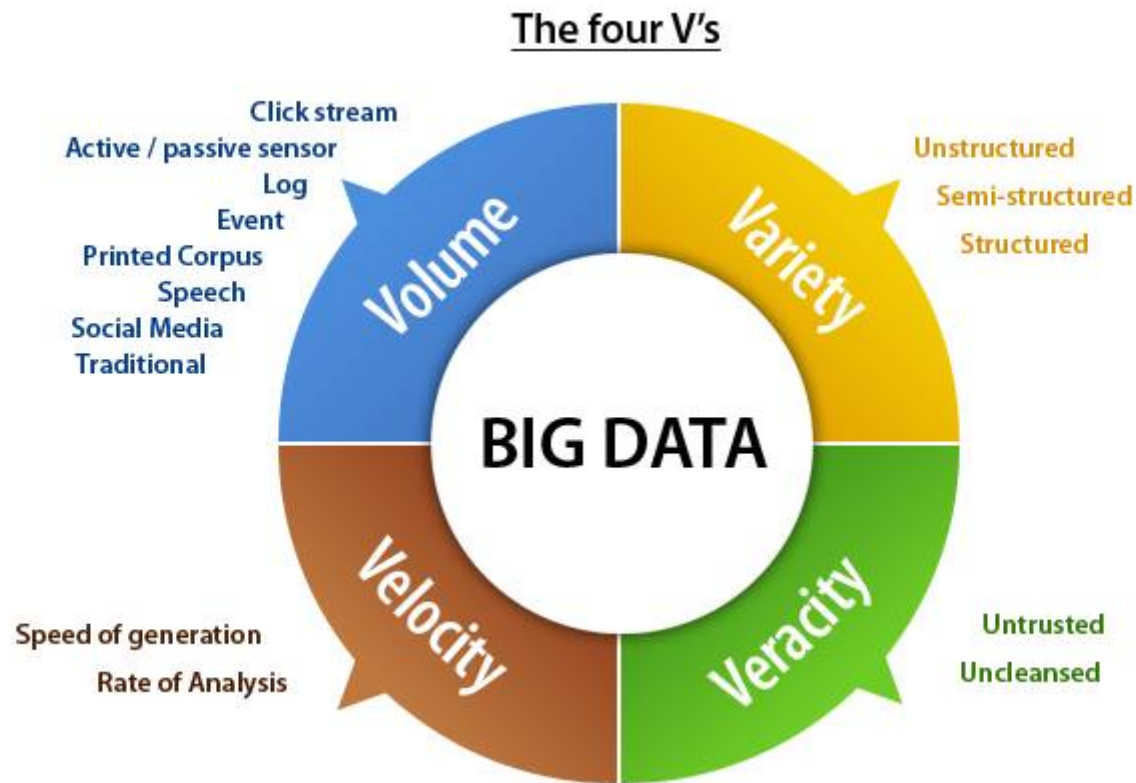
■ Veracity:

- The degree to which data is **accurate, precise and trusted**
- Problem spaces, data sets and operational environments is often uncertain, imprecise and difficult to trust



+ The four Vs

18



- There are many people that are champions in adding more Vs...
 - ...let's stay with the first four
 - See i.e. <https://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077>

+ It is a relative concept

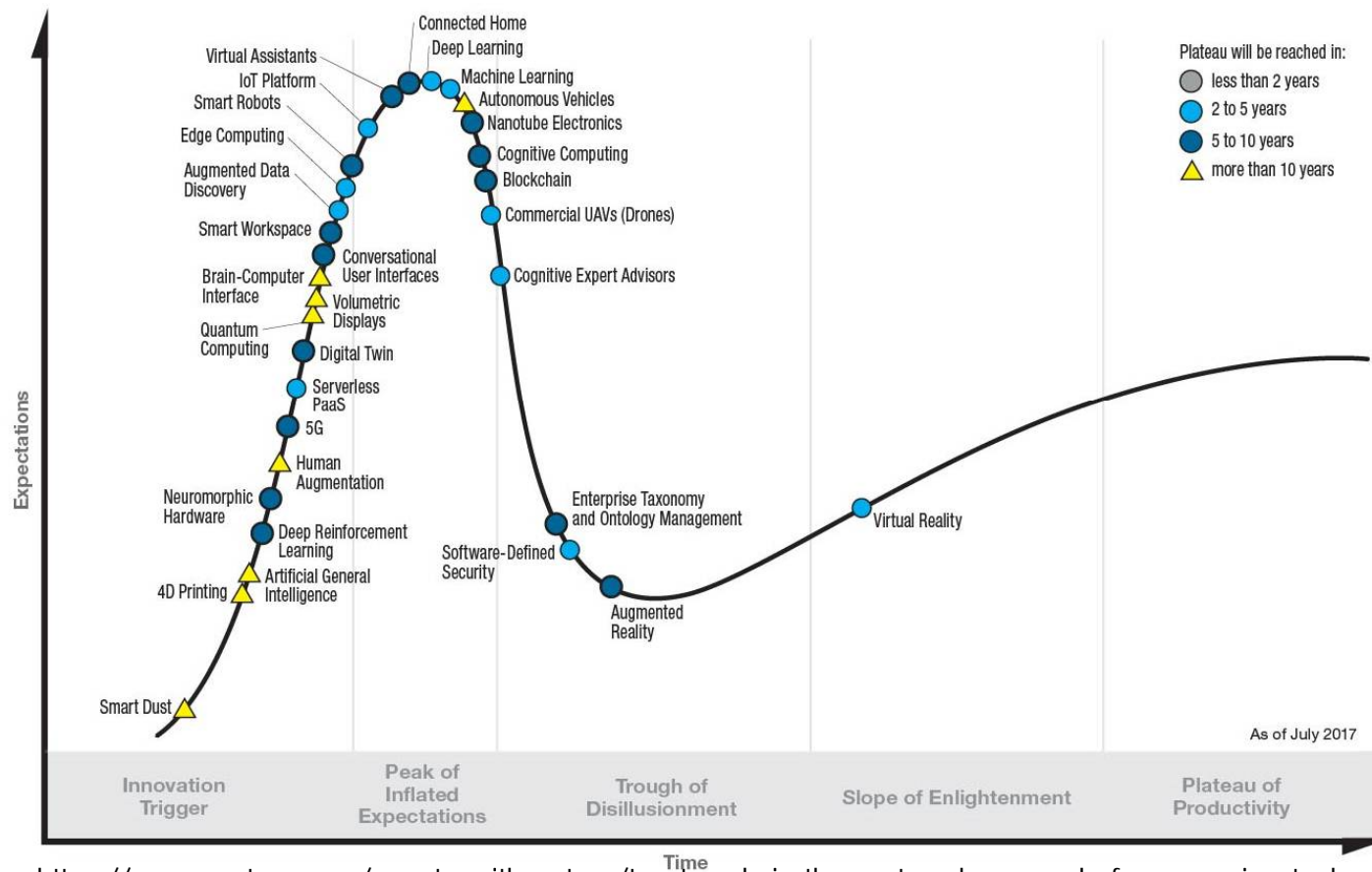
19

- The Big Data concept is not absolute, it is also varying in time
 - What constitutes Big Data today may not be tomorrow's Big Data
- From anyone's given perspective, if you are facing significant challenges around data's volume, velocity and variety, it is your big data challenge
 - But big challenges may also provide big opportunities

+ Big Data hype – where is it?

- One of the most “hyped” terms today

Gartner **Hype Cycle** for Emerging Technologies, 2017

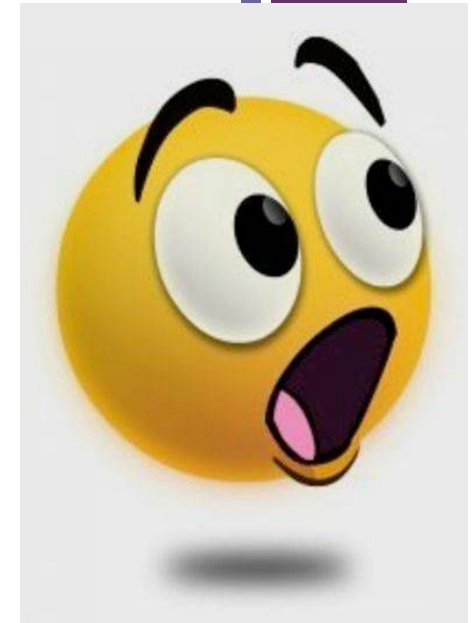


© Source: <https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/>

+ ...there isn't...

21

- Betsy Burton (Gartner Analyst) says:
“We’ve retired the big data hype cycle (in 2015) [...] big data has quickly moved over the Peak of Inflated Expectations [...] and has become prevalent in our lives across many hype cycles. [...] “I would not consider big data to be an emerging technology”



© Source: <https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/>

+ Data sources examples - 1

22

■ Web data

- Customer level web behavior data
 - page views, searches, reading reviews, purchasing
 - next best offer
 - churn modelling
 - targeted advertisement

■ Text data

- email, news, web pages, documents uploaded
- one of the biggest and most widely applicable types of Big Data.
- The focus is typically on extracting key facts from the text and then use the facts as inputs to other analytic process

+ Data sources examples - 2

23

■ Time and location data

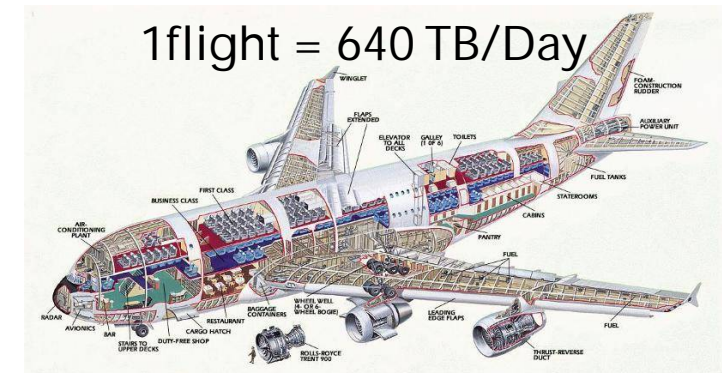
- GPS and mobile phone as well as Wi-Fi connection makes time and location information a growing source of data.
 - **Individual** level
 - Companies realize the power of knowing when their customers are at which location.
 - Advertisements and tailored information
 - At an **aggregated** level
 - Shaping and dimensioning infrastructure
 - Placing vending spots
 - **One of the most privacy-sensitive types of Big Data**

+ Data sources examples - 3

24

■ Smart grids and sensor data

- Cars, Oil pipes, Windmill turbines,
 - Are now collected at extremely high frequency
 - Help to improve performance of engines and machinery
 - Enables diagnosis of problems more easily and faster



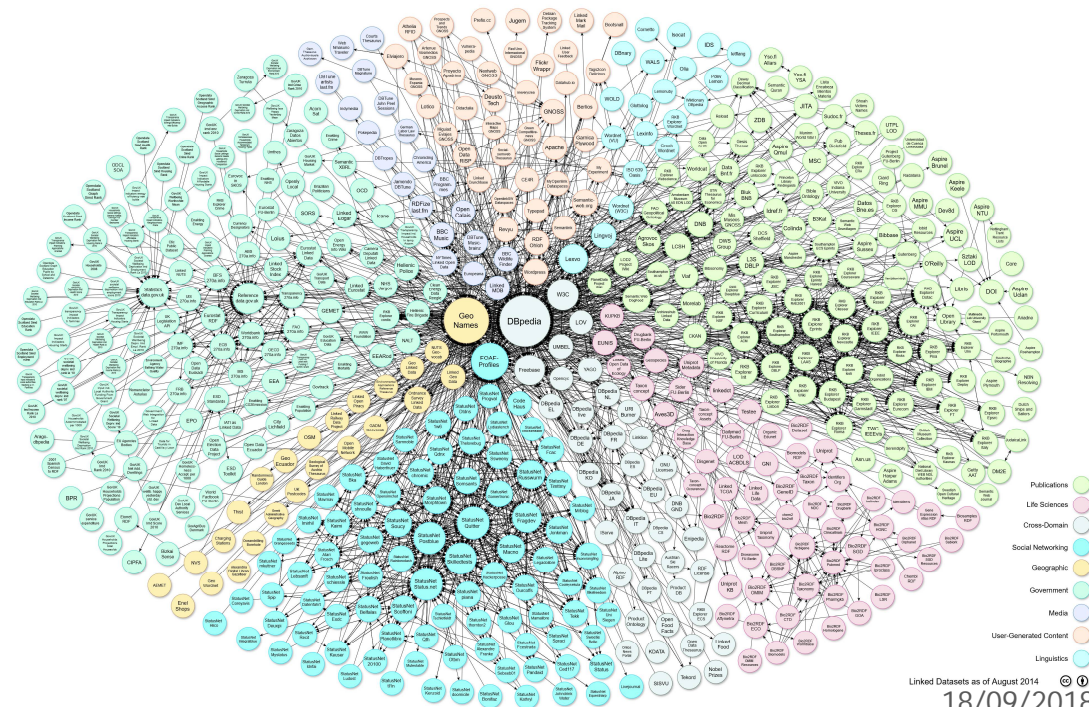
■ Social network data

- Link analysis to uncover the network of a given user
 - Targeted advertisements by considering not only interests the customers have personally stated, but also knowing what it is that their circle of friends or colleagues has an interest in.



+ The value is in the combination

- The power is not just in what that particular source of data can tell you uniquely by itself
- The value is in what it can tell you in combination with other data



+ How is big data different from traditional data sources?

- Big Data can be an entirely new source of data
 - Browsing behavior in online shopping
- The speed of data feeds can transform old data into new data
 - enables a very different, more in-depth level of analytics that such data is really appears as a new data source
- Structure of data
 - Most traditional data sources are in the structured forms
 - Every piece of information included is known ahead of time, comes in a specified format and occurs in a specified order
 - database tables
 - spreadsheets
 - Big Data are increasingly more semi-structured and unstructured
 - You have little or no control over its format.
 - Text data, video data and audio recordings

+ Semi-structured data

- Semi-structured data are data that may be irregular or incomplete and have a structure that may change rapidly or unpredictably
 - Have some structure, but does not conform to a fixed schema
 - Web logs are good example of semi-structured data.
 - The log text generated by a click n a website right now can be longer or shorter than the log text generated by a click from a different page a minute later.
 - Semi-structured data does have an underlying logic

+ Is Big Data better than traditional data?

- The power of Big Data is in the analysis and the actions you take as the result of the analysis.
- Big Data or “small” data does not in and by itself possession any value
- Big Data are valuable only when you manage to get some insight out of the data



+ Analytics types - 1

29

■ Descriptive analytics

- answers the questions about “what happened in the past?”
 - What was the sales revenue in the first quarter of the year?
 - Which is our most profitable product/region/customer?
 - How many of the won customers can be attributed to the promotional campaign

■ Predictive analytics

- aim to say something about “what might happen next?”
- Typically is harder!!
 - What will the number of complaints to our call center next month?
 - which type of customer are most likely to churn (e.g. cancel her subscription)?
 - What is the next best offer for this customer?

+ Analytics types - 2

30

■ Prescriptive analytics

- Tries to answer, “how do I deal with this” ?
- Analytics gets operational
 - We know that this person has a high chance to churn, we can offer her a value package
 - We know the viewing history of this customer on our news site, we can recommend articles that we think she would like to read next
 - From analyzing various sensor data we know that part of my datacenter or my windmill is about to break, a replacement part is automatic ordered → Predictive maintenance

+ Is this new?

31

- All three type of analytics existed before Big Data
- Big Data improve the ability for precise forward-looking insight
 - to predict what might happen next
- Big Data improve the ability for **fast and actionable** insight
 - The impact is embedded in the process
 - i.e. recommender systems automatically generate personalized recommendations
 - i.e. Amazon recommendation for you are different from those recommended to me (we have different purchasing and viewing history) right after a purchasing transaction in the hope to increase sales there and then.

+ Applications - Segmentation

32

■ Segmentation

■ Banks

- Every day millions of people apply for new credit cards, loans, and mortgages
 - Big Data analytics on all the data available (to the bank) about you to **calculate the credit score**

■ Insurance companies

- wearing devices to measure biometrics, such as fitness activity, sleep patterns, calorie intake to **predict health outcomes**

+ Applications - Churn prediction

■ Churn prediction

- I.e. in the telecom sector customers switch from one company to another is called churn.
 - Flag customer that at the risk of churning
 - i.e. using web data about cancellation pages navigation
 - i.e using publically available social media data to improve their churn model
 - Find ways to retain them



+ Applications - Recommender systems

■ Recommender systems and targeted marketing

- Amazon.com (“customer who bought this item also bought...”)
- Music recommendations on Spotify
- Movie recommendations on Netflix
- News recommendations on almost all news portals

- Recommendation can be based on general trends
 - most read news for today...

- Can be personalized
 - “recommended to you because you have watched...” on Netflix,

- Recommender system can affect business significantly
 - Netflix reported that 2/3 of the movies watched are recommended,
 - Google News stated that recommendations generate 38% more click-through
 - Amazon claimed that 35% sales come from recommendations.



+ Applications - Sentiment Analysis

35

■ Sentiment analysis

- Looks at the general direction of opinions across a large number of people to provide information on what the market is saying, thinking, and feeling about an organization.
- It often uses data from **social** media sites
 - what is the buzz around a company or product?
 - Are people saying good or bad things about an organization and the services it offers?
- It can also be used at an **individual** level
 - can use pattern recognition to detect a caller's mood at the start of a call.
 - routed to a specialist for careful treatment



+ Other Applications

36

■ Operational Analytics

- Embedding analytics within business processes and automate decisions
 - Without human intervention
 - i.e. airlines automatically reroute customers when a flight is delayed

■ Social good, research and personalized medicine

- Remote diagnosis
- Prevention
 - Just think to wearable and connected devices that monitor medical parameters



Devices and computing infrastructures for Big Data

+ Low-Power System on Chip (SoCs)

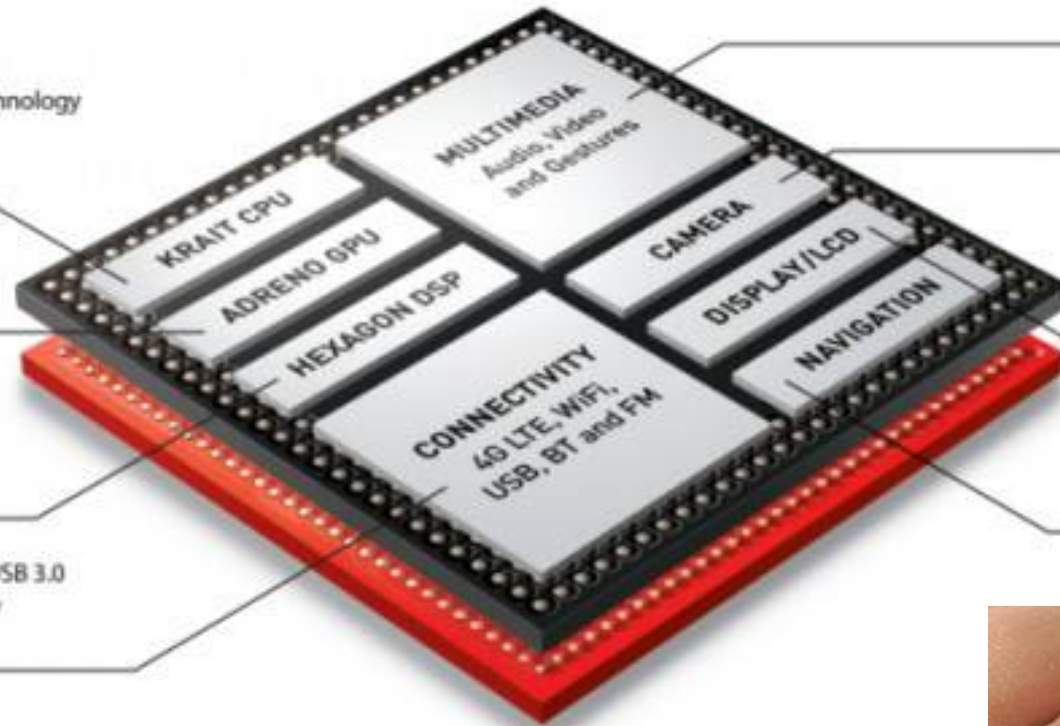
800 PROCESSOR

Krait 400 CPU
features 28HPm process technology
superior
2GHz+ performance

Adreno 330 for
advanced graphics

Hexagon QDSP6
for ultra low power
applications and custom
programmability

Integrated LTE⁺, 802.11ac⁺, USB 3.0
and BT 4.0 offers broad array
of high speed connectivity



Ultra HD Capture
and Playback
DTS-HD and Dolby
Digital Plus audio
Expanded Gestures

55MP with dual ISP

Support for up
to 2560x2048 display
Miracast 1080p
HD support

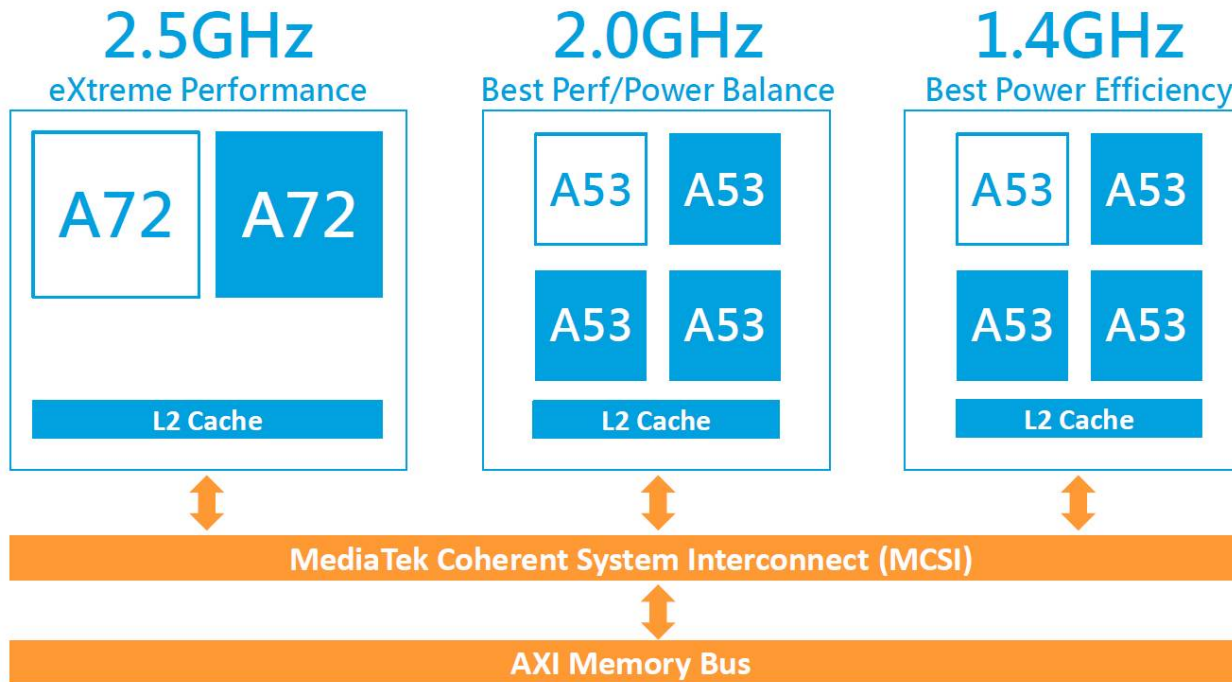
IZat GNSS with
support for three
GPS constellations



+ SoC Multicore Madness

big.Medium.LITTLE

Deca/10-Core CPU Architecture



+ NVIDIA JETSON TK1



- First **ARM+CUDA programmable SoC based** Linux development board

- 4 cores ARM A15 CPU

- 192 cores NVIDIA GPU
 → 300 GFLOPS (peak sp)

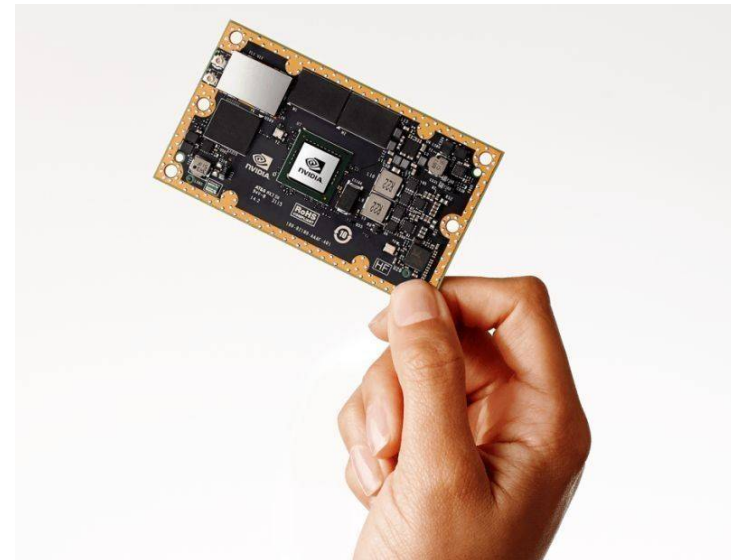
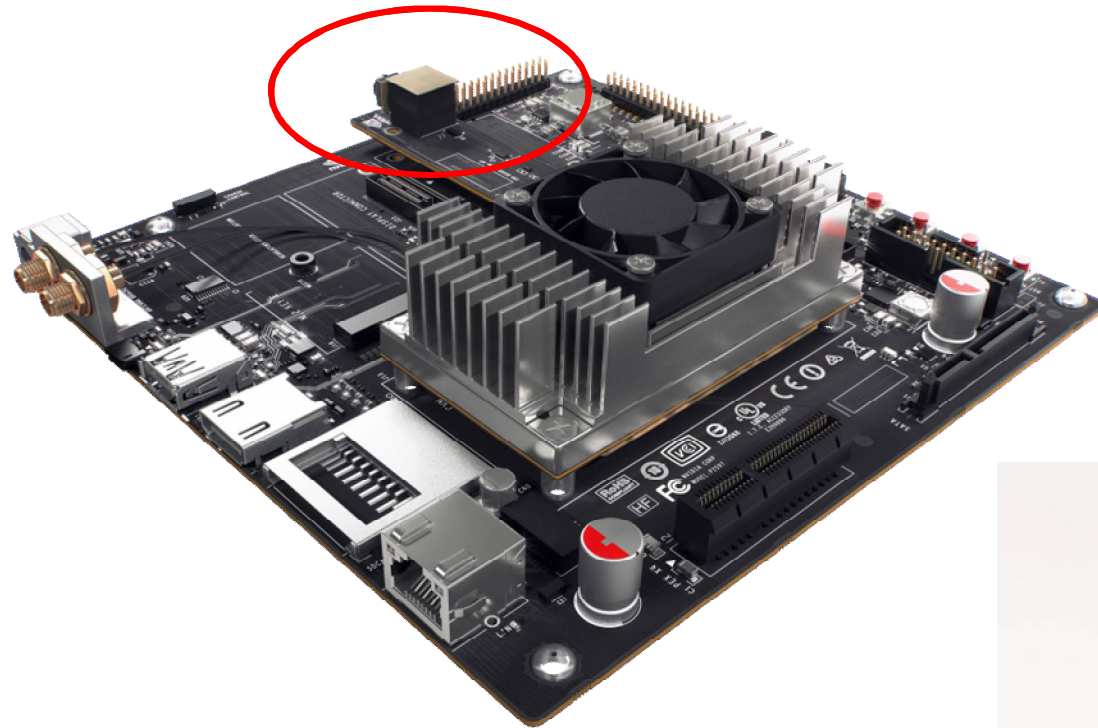
~ **21 GFLOPS/W (sp)**

- ... for less than 200 Euros

- 32bit

- 64bit version announced

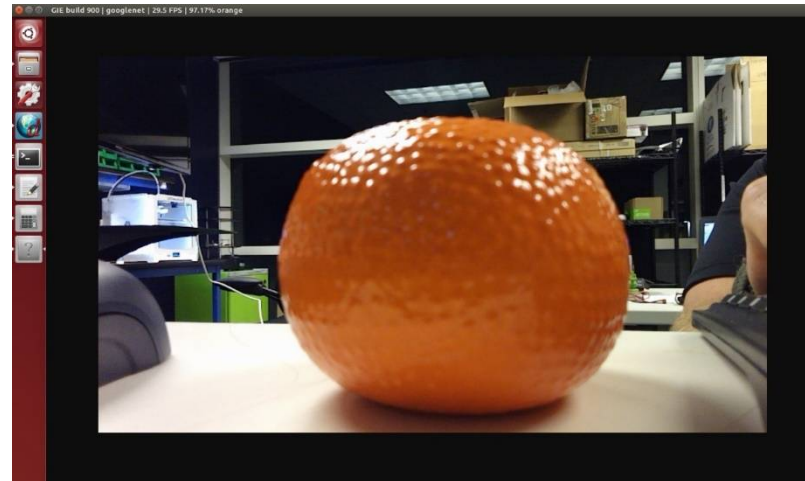
+ Jetson TX1



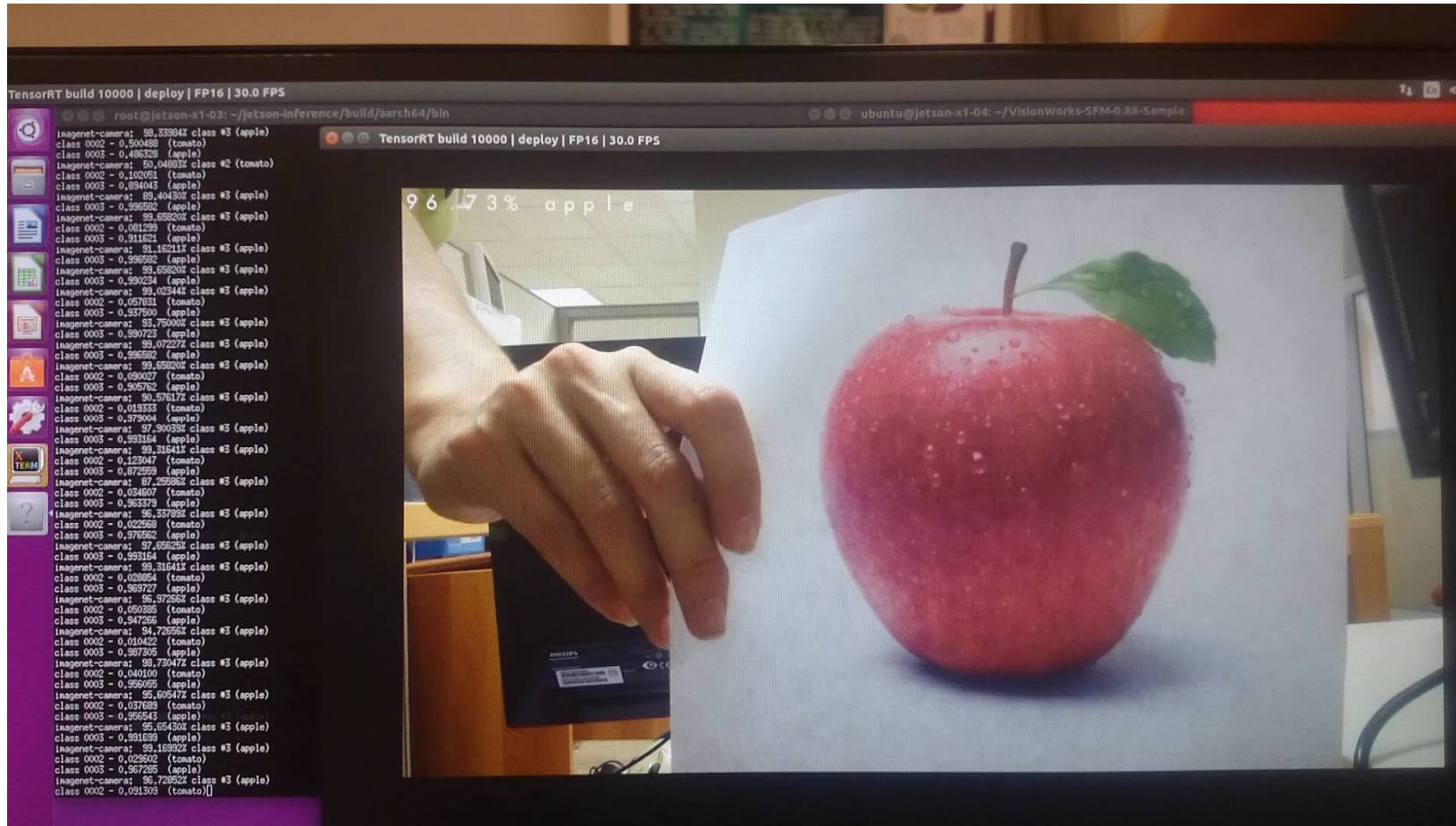
+ Image classification on the TX1

42

- Dataset creation using ImageNet
- Training using DIGITS + Caffe on 2 K40 or XeonD+K20
- Inference on Jetson X1 using NVidia TensorRT
- Real time classification



+ Image classification on the TX1



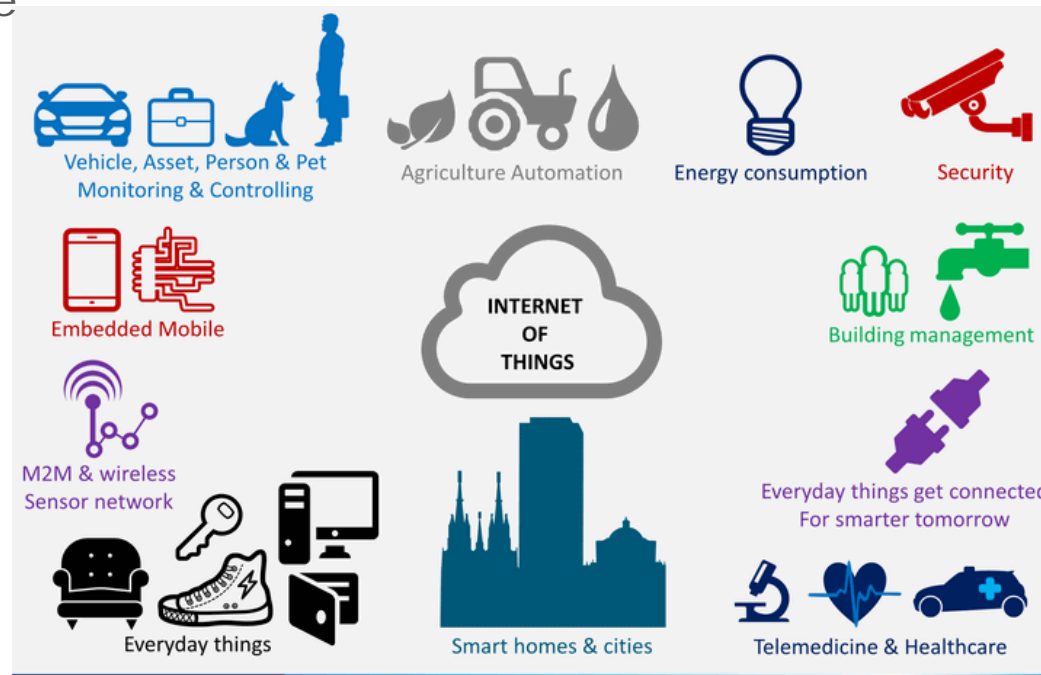


Internet of Things

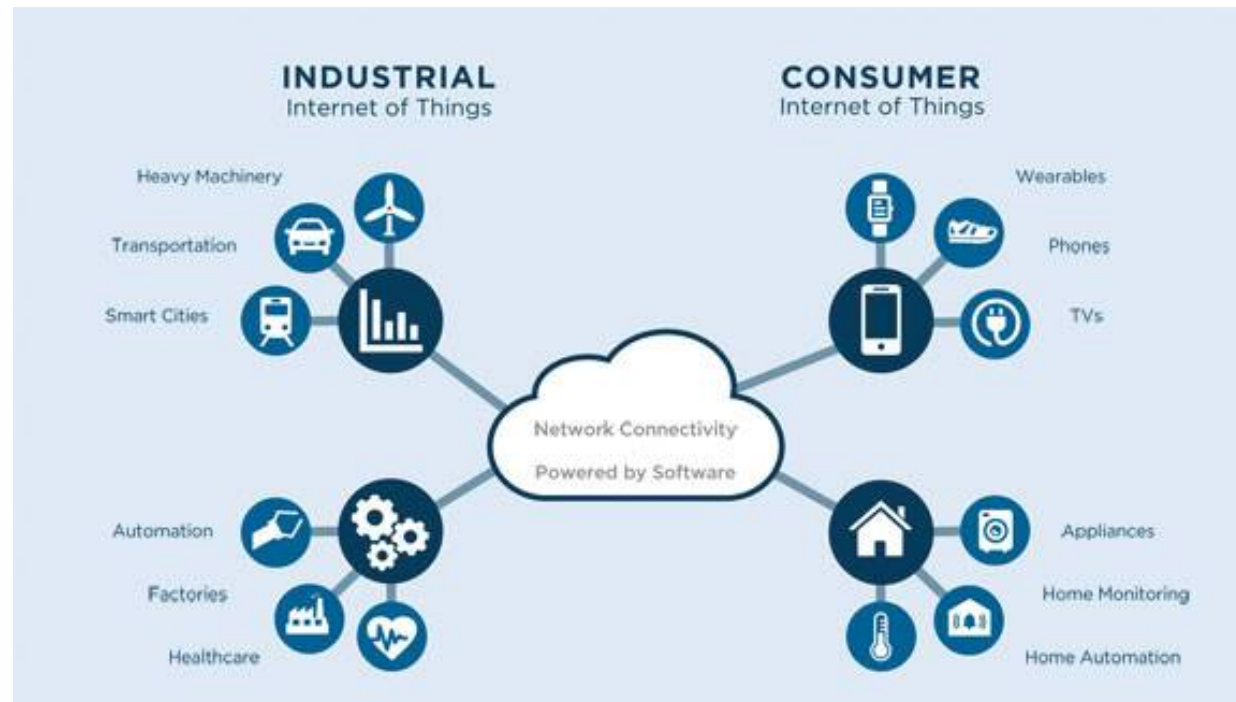
+ Internet Of Things (IoT)

- ❑ Internet of Things (IoT) is the **network of physical devices**
 - ❑ vehicles, home appliances and other items
- ❑ Embedded with electronics, software, sensors, actuators, and **connectivity** which enables these objects to connect and **exchange data**
- ❑ Each **“Thing”** is uniquely identifiable through its embedded computing system but is able to inter-operate within the existing Internet infrastructure

(Source: Wikipedia)



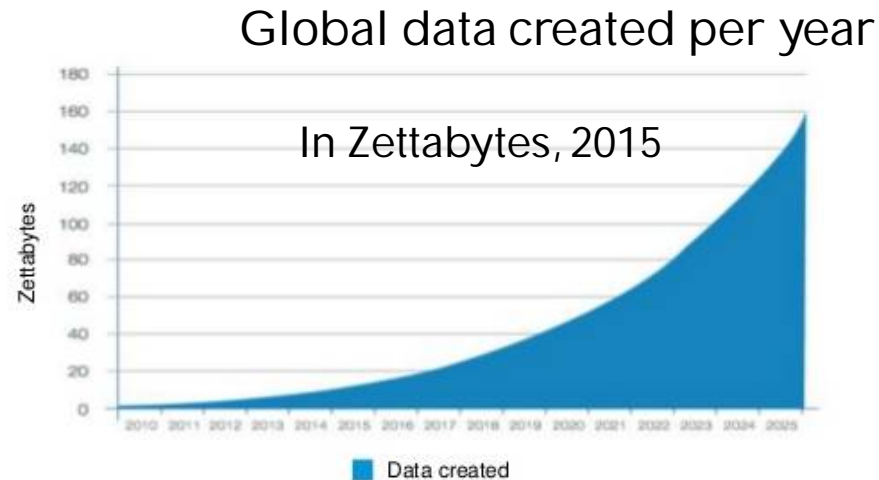
+ Internet Of Things (IoT)



- ❑ The IoT is a network of smart computers, devices, and objects that collect and share huge amounts of data
- ❑ The collected data is sent to a **central Cloud-based service** where it is aggregated with other data and then shared with end users in a helpful way
- ❑ The IoT increases automation in homes, schools, stores, industries

+ Fog/Edge Computing for IoT devices

47



Source: IDC Data Age 2025 study, sponsored by Seagate, April 2017




- Traditionally, much of this data has been managed and **stored through cloud computing**, a **centralized** network of computers and servers connected together over the Internet
- However, **access to data through the cloud can be slow** at times, because the data needs to be transported to the cloud **for processing, analysis and storage**.
- Low latency, high velocity, multiple location distribution -> the Cloud fails
- Moreover, there are privacy concerns


+ Fog/Edge Computing for IoT devices

- For example, autonomous cars, with hundreds of on-vehicle sensors, will generate **40TB of data for every eight hours of driving**.
- That's a lot of data. **It is unsafe, unnecessary, and impractical to send all that data to the cloud.**

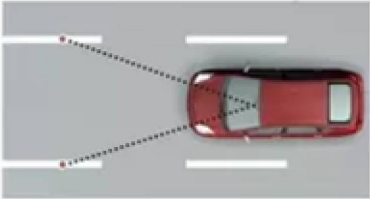
Radar
Used for adaptive cruise control. Reflected microwaves can identify location and speed — but not always type — of nearby vehicles.



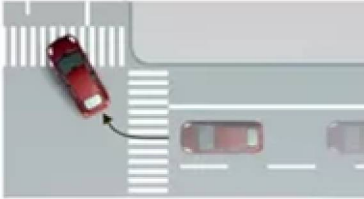
Ultrasound
Used for assisted parking. Reflected sound waves detect distance to nearby objects. Some cars use short-range radar instead.



Cameras
Used for lane-keeping and back-up assistance. Image-processing software can detect lane stripes, signs, stop lights, road signs and other objects.



Navigation Aids
Global positioning system unit determines car's position. Accelerometers and wheel sensors help with navigation when satellite signals are blocked.



Central Diagram: A top-down view of a car with various sensors labeled: RADAR, CAMERA, GPS UNIT, ACCELEROMETER, ULTRASOUND SENSOR, and WHEEL SENSOR.

LIDAR
Google's autonomous vehicle project uses a spinning range-finding unit, called lidar, on top of the car. It has 64 lasers and receivers.



The device creates a detailed map of the car's surroundings as it moves. Software adds information from other sensors and compares the map with existing maps, alerting the system to any differences.



+ Fog/Edge Computing for IoT devices

49

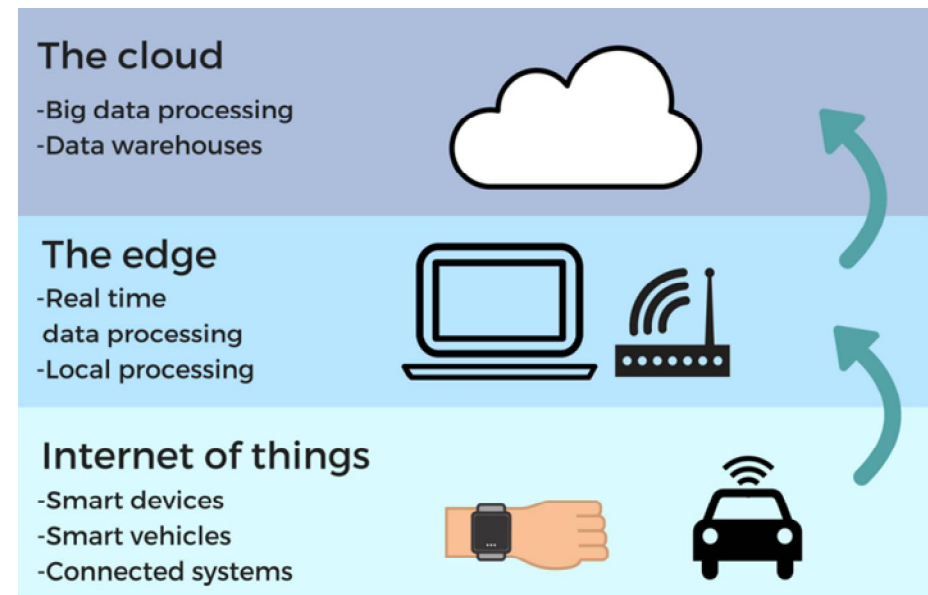
- Need for alternative, **adaptive, decentralized computational paradigms**
 - **complement the centralized Cloud Computing** model serving IoT networks
- In this alternative paradigms, **data is accessed locally**
 - reduces the amount of time it takes to access the data
 - real-time decisions can be made locally
- Main features
 - reduction of network traffic
 - low-latency requirement addressed
 - scalability
- Fog Computing and Edge Computing provide faster approaches that gain better situational awareness in a far more timely manner



+ Edge Computing: on-device processing

50

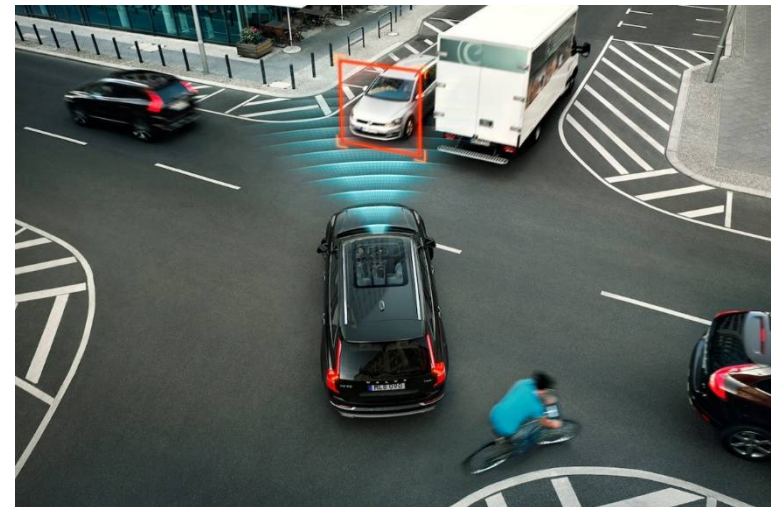
- ❑ “Edge” refers to the computing infrastructure that exists **close to the sources of data**, and typically away from the centralized computing available in the Cloud
- ❑ The role of Edge Computing to date has mostly been used to ingest, gather, store, filter, and send data to cloud systems
- ❑ A lot of data processing (compute and analytics) can be done on the devices themselves, at the edge of the network, near the source of the data
- ❑ Significant impact on **latency**
- ❑ The **volume of data** moved and the distance it travels are drastically **reduced**



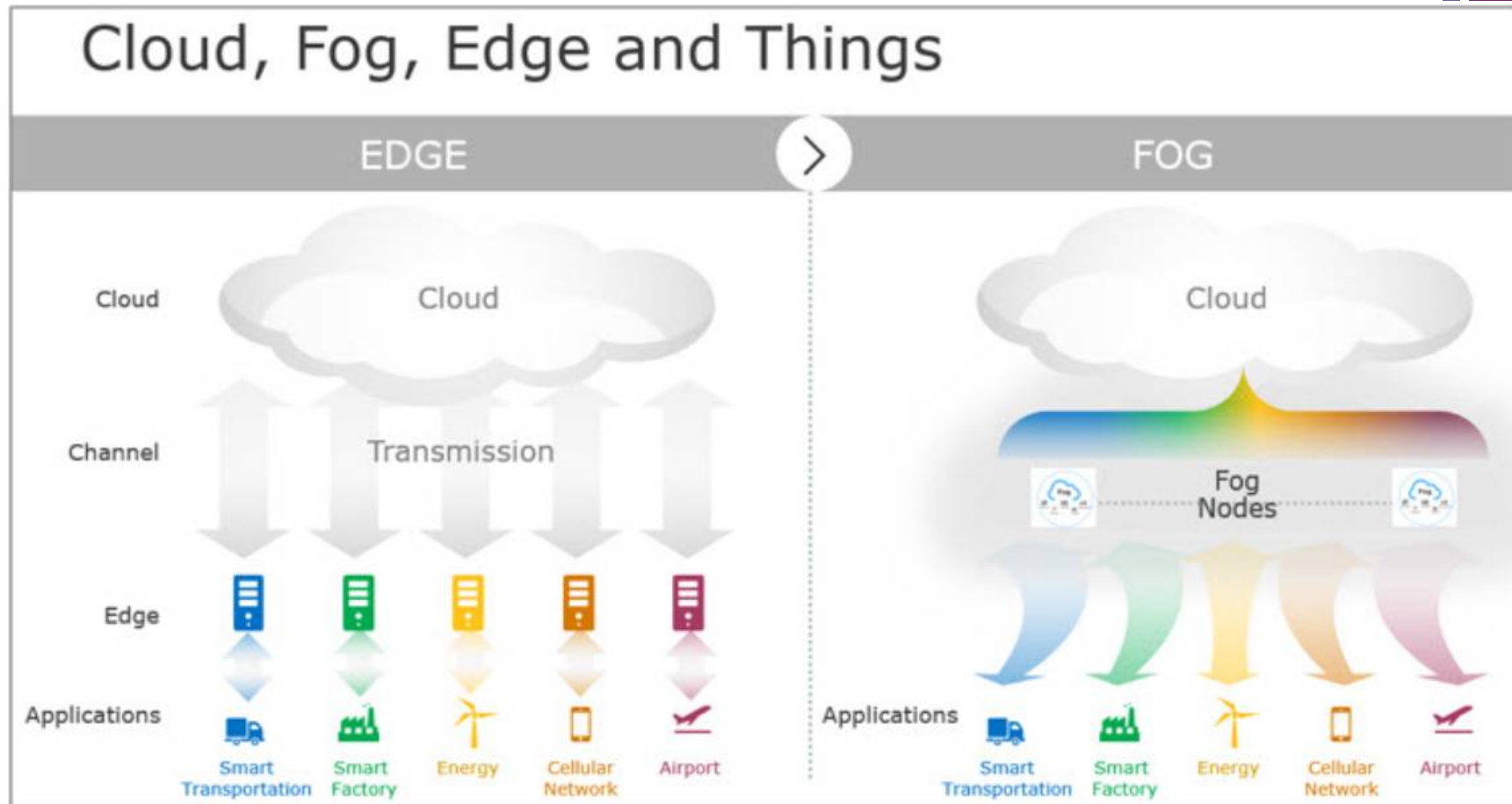
+ Edge Computing: on-device processing

51

- ❑ Edge Computing provides new possibilities in IoT applications, particularly for those relying on
 - ❑ machine learning for tasks such as object detection, face recognition, language processing, and obstacle avoidance
 - ❑ low/intermittent connectivity
 - ❑ remote locations
 - ❑ bandwidth and associated high cost of transferring data to the cloud
- ❑ Low latency, such as closed-loop interaction between machine insights and actuation (i.e. taking action on the machine)
- ❑ Immediacy of analysis
- ❑ Access to temporal data for real-time analytics
- ❑ Security and privacy, which can be improved with Edge Computing by keeping sensitive data within the device

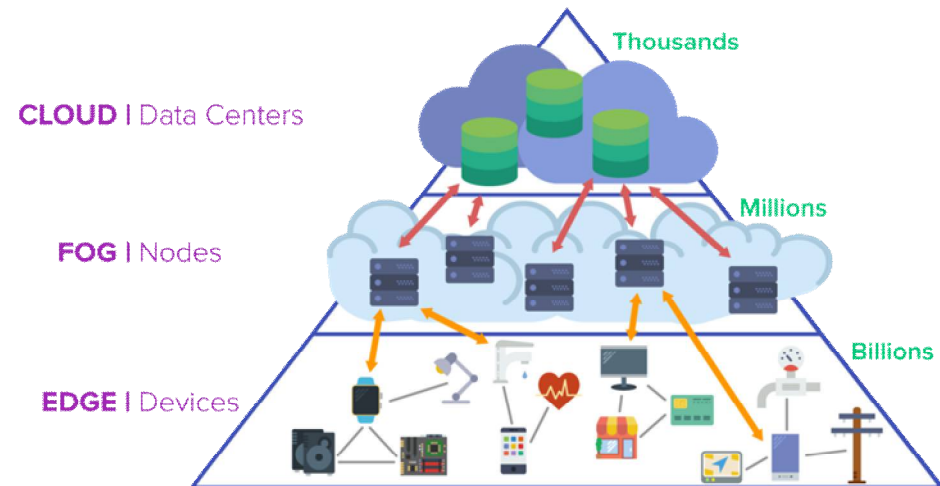


+ Edge/Fog Computing

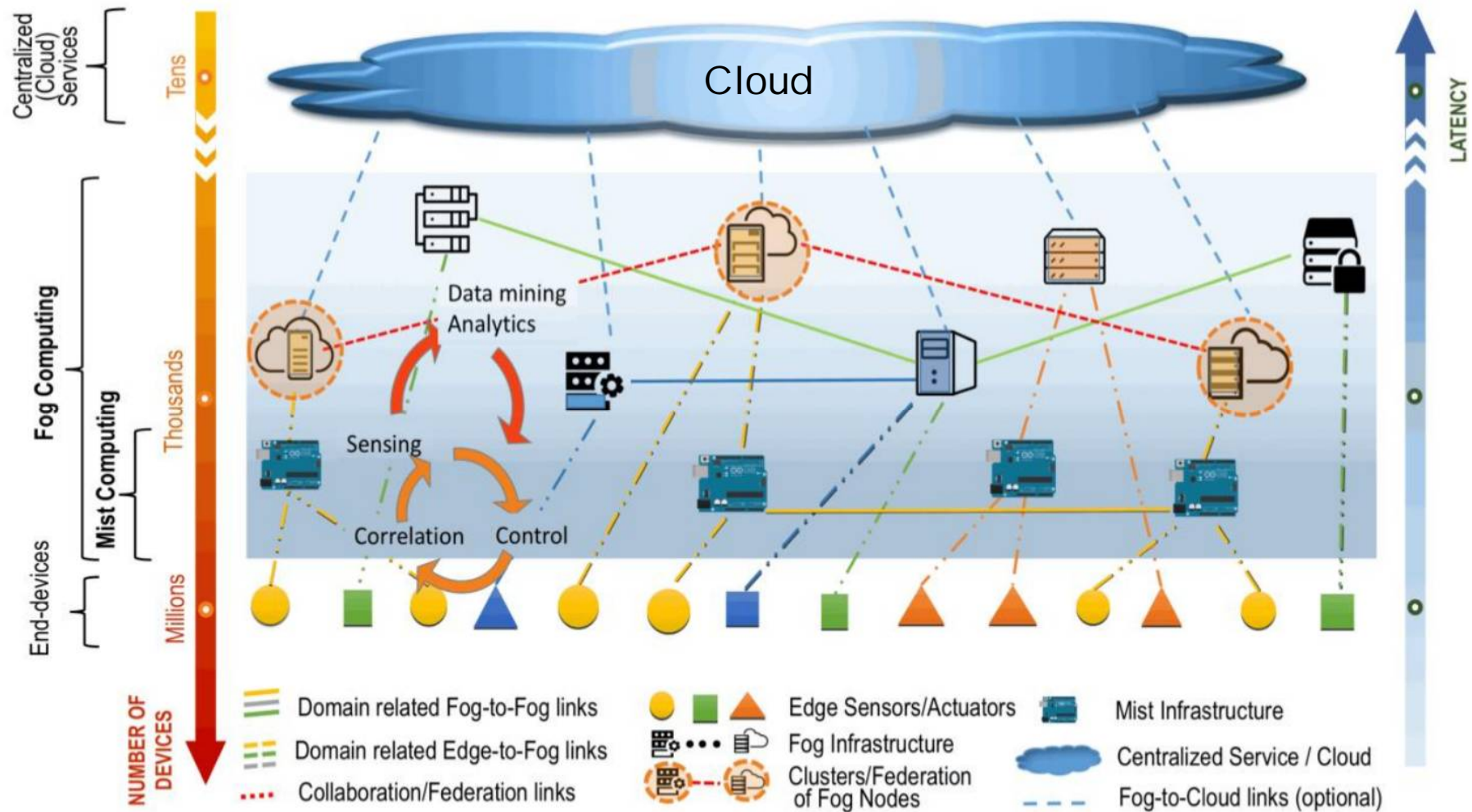


+ Fog Computing: bridging the continuum from cloud networks to things

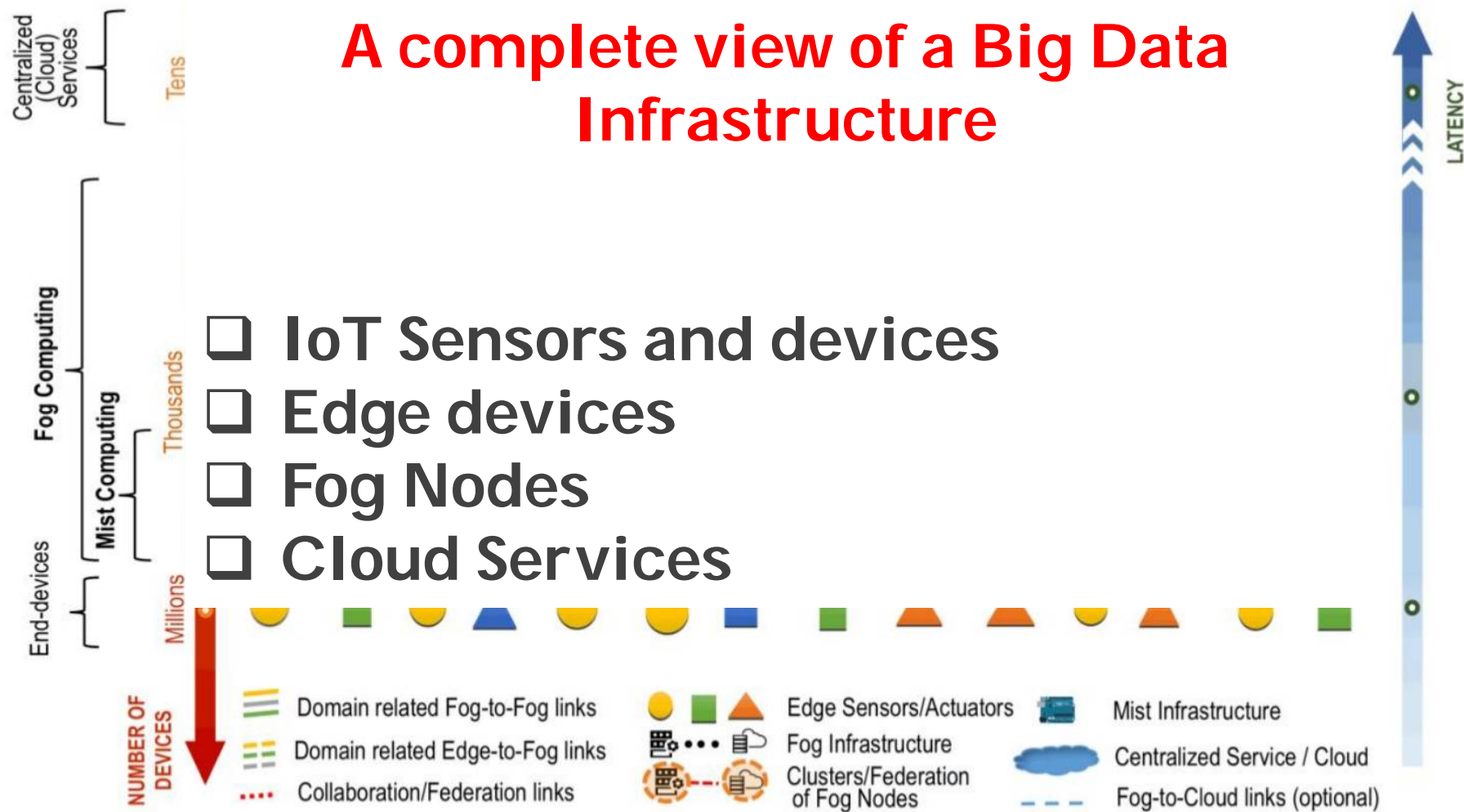
- A layered model for enabling ubiquitous access to a **shared continuum of scalable computing resources**
- Facilitates the deployment of distributed, latency-aware applications and services, and consists of **Fog Nodes** (physical or virtual), **residing between smart end-devices and centralized (cloud) services**.
- The Fog Nodes are context aware and support a common data management and communication system. They can be organized in **clusters**
- Fog Computing minimizes the request-response time from/to supported applications
- Provides, for the end-devices, local computing resources and, when needed, **network connectivity to centralized services**



+ Fog Computing: bridging the continuum from cloud networks to things



+ Fog Computing: bridging the continuum from cloud networks to things



+ Essential characteristics of Fog computing

56

- ❑ Contextual location awareness of the fog nodes, and low latency
- ❑ Can be geographically distributed
- ❑ Heterogeneity, in supporting collection and processing of data of different form factors through multiple types of network communication capabilities
- ❑ Interoperability and federation
- ❑ Real-time interactions rather than batch processing
- ❑ Scalability and agility of federated, fog-node clusters
- ❑ Adaptive by nature



+ Additional characteristics of Fog Computing

57

- Predominance of wireless access, well suitable for wireless IoT access networks
- Support for mobility techniques, such as the Locator/ID Separation Protocol (LISP)
- Edge may deliver cloud networked capabilities but without orchestration for connecting Edge nodes
- Fog orchestration and management is intended to be more universal, modern, and automated



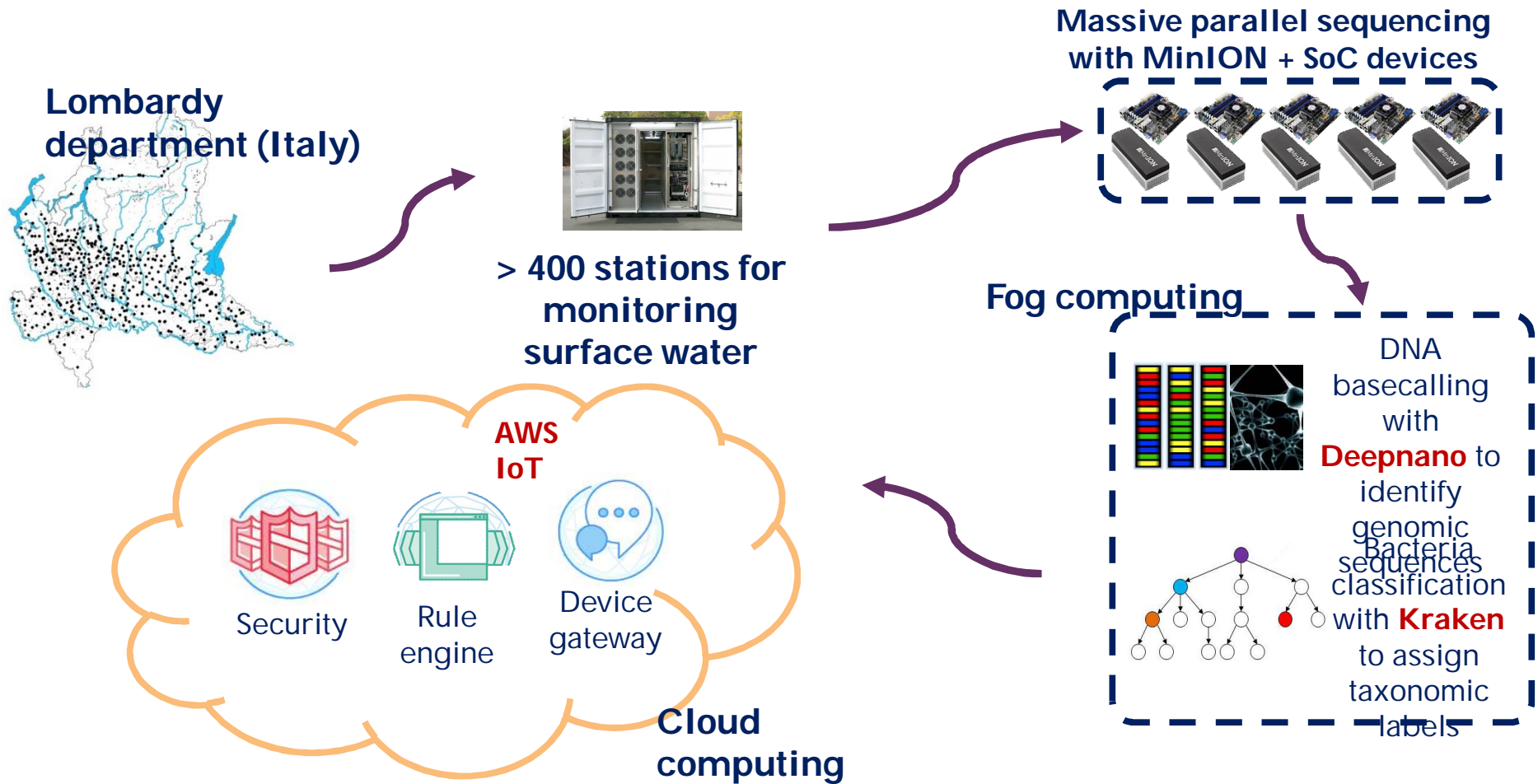
+ Fog node deployment models

58

- **Private Fog Nodes:** provisioned for exclusive use by a single organization comprising multiple consumers (e.g. business units)
- **Community Fog Node:** provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns
- **Public Fog Node:** provisioned for open use by the general public
- **Hybrid fog node:** a complex fog node that is a composition of two or more distinct fog nodes (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability



Fog Computing for Metagenomics



Low-Power Portable Devices for Metagenomics Analysis: Fog Computing Makes Bioinformatics Ready for the Internet of Things, Future Generation Computer Systems, 2017

+ References - 1

- <https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf>
- <http://www.pnas.org/content/112/38/11887>
- <https://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077>
- <https://ggwash.org/view/40557/how-snow-exacerbates-the-weaknesses-of-suburban-road-design>
- <https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017>
- <https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve>
- http://www.martinhilbert.net/wp-content/uploads/2015/01/BigData4Dev_Hilbert2014.pdf
- <https://www.youtube.com/watch?v=0hPb0S4zMSg>

+ References – 2

- https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf
- <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-325.pdf>
- <https://www.lanner-america.com/blog/5-examples-edge-computing-solutions-use-today/>
- http://www.apc.com/salestools/VAVR-A4M867/VAVR-A4M867_R0_EN.pdf?sdirect=true
- <https://www.networkworld.com/article/3234708/internet-of-things/why-edge-computing-is-critical-for-the-iot.html>

+ Break!



© Kai Godehusen /Getty Images