QCon

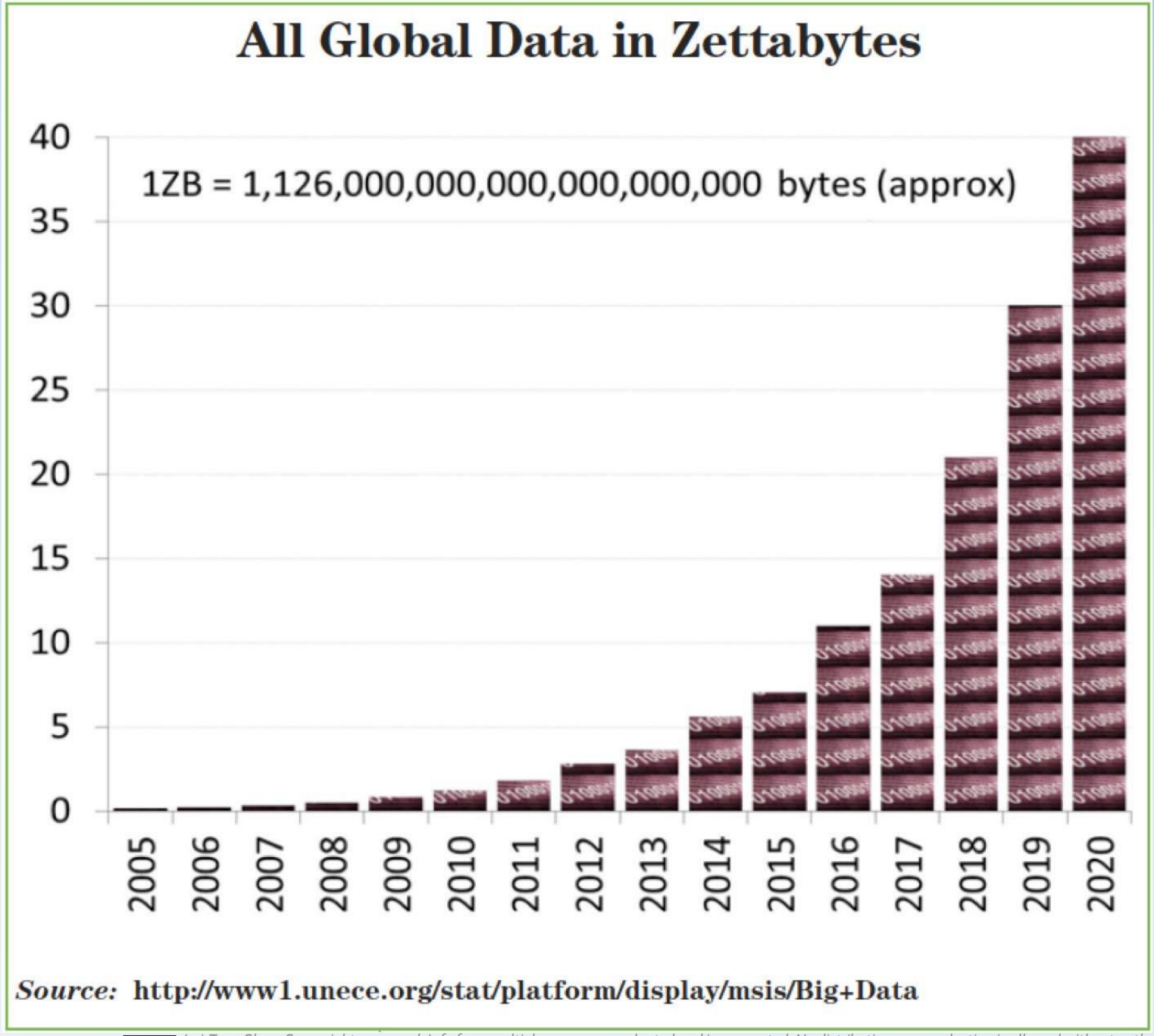
Uplifting Big Data

Tony Shan
mail@tonyshan.com

@TonyShan #UpliftBigData

October 20, 2016

Table of Contents



Ride the Big Data Wave

90%

of the world's existing data has been created in the last

2 YEARS

By 2020

35 Zettabytes worth of data will be created

*35 Zettabytes = 35,000,000,000,000,000,000,000 bytes

1/3 of that data will be stored in or will have passed through the cloud

10,000

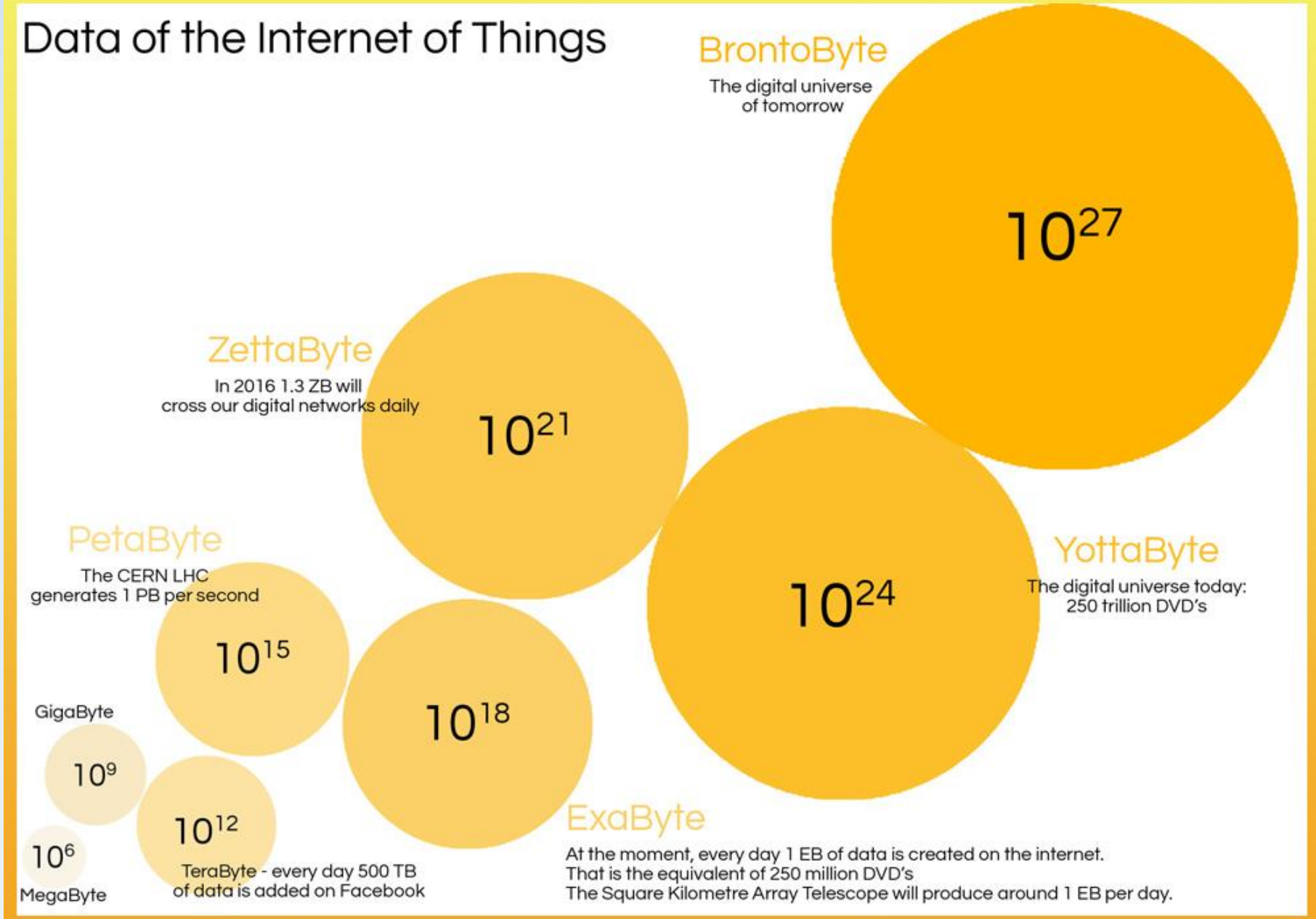
Libraries of Congress worth of data are created annually by U.S. companies

80,000 km

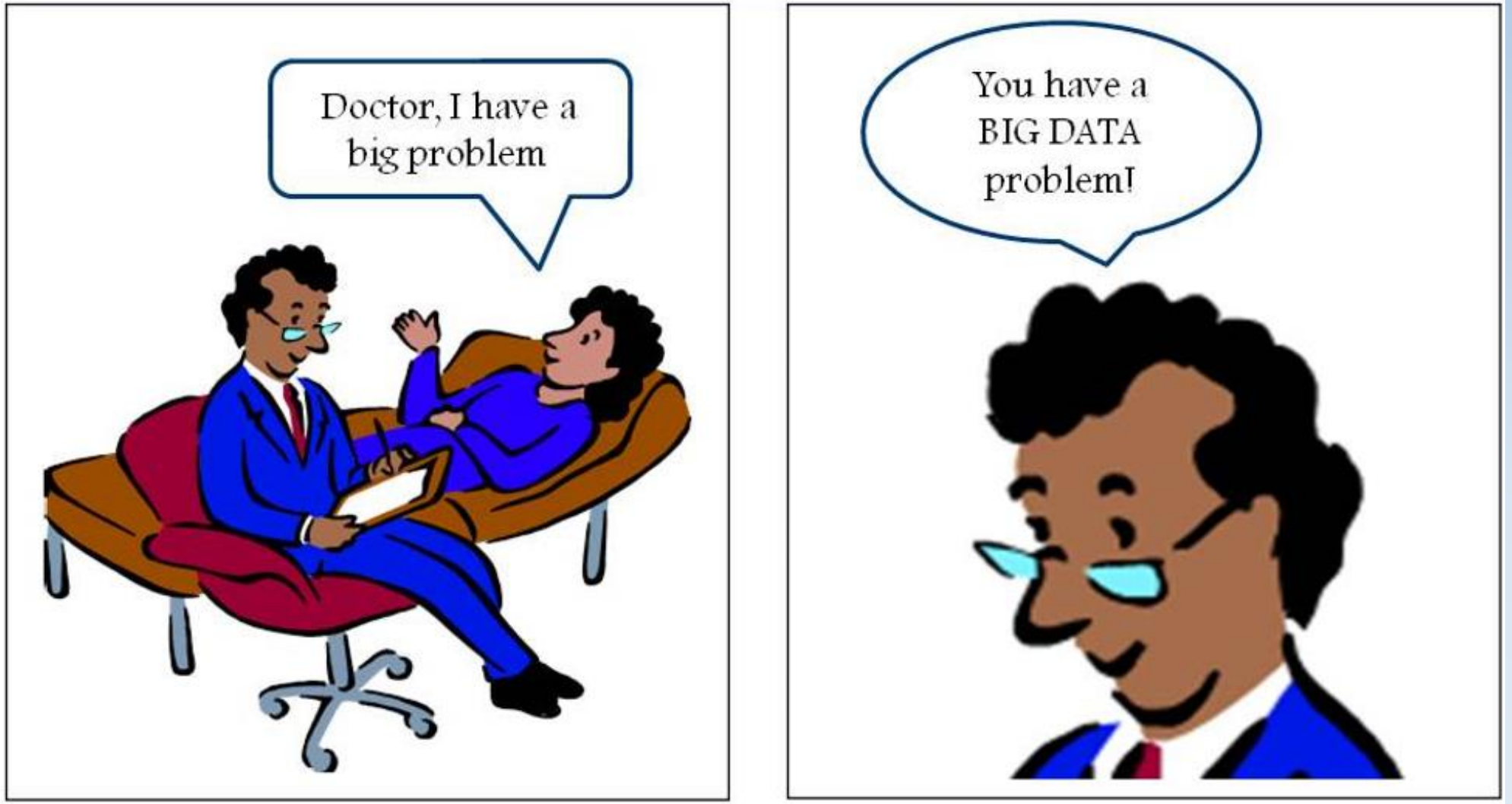
is the distance past the moon the current global storage capacity for digital information stored on a CD stack would stretch

Every 2 days as much information is created as from the dawn of civilization up until 2003

There are actually more pieces of digital information than stars in the universe



In Search of Solutions



Do You Have a Big Data Problem?

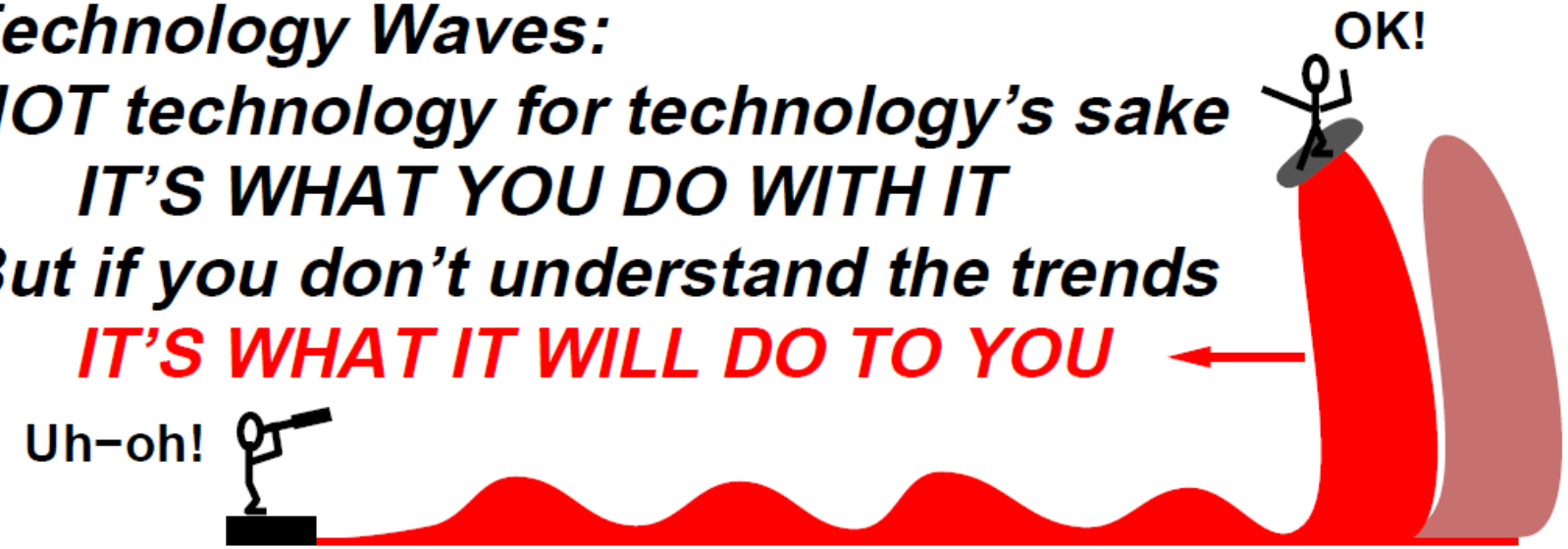
What is Big Data?

Who coined "Big Data" Term?

John Mashey was arguably the first to use the term Big Data in his public talks in **late 1990s**.

Big Data ...
and the Next Wave of InfraStress
John R. Mashey
Chief Scientist, SGI

Technology Waves:
NOT technology for technology's sake
IT'S WHAT YOU DO WITH IT
But if you don't understand the trends
IT'S WHAT IT WILL DO TO YOU



4/25/98 page 1

Father of "Big Data" Concept

Francis X. Diebold presented a paper, in which he states "Recently, much good science, whether physical, biological, or social, has been forced to confront—and has often benefited from—the 'Big Data' phenomenon. *Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology.*"

Diebold, F.X. (2003),
"'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting"
(Discussion of Reichlin and Watson papers), in M. Dewatripont, L.P. Hansen and S.Turnovsky (Eds.),
Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society.
Cambridge: Cambridge University Press, 115-122.

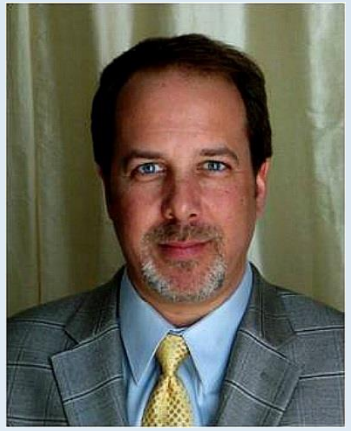
"Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting

Francis X. Diebold

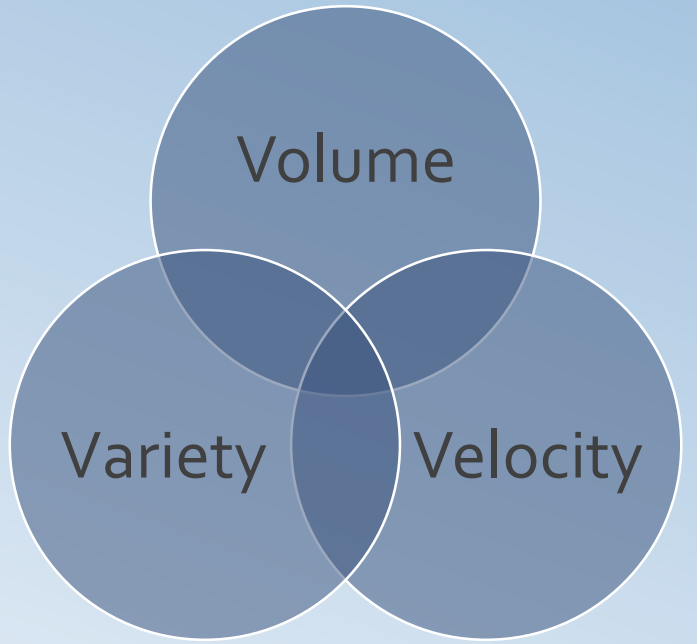
University of Pennsylvania
and NBER

First Version, July 2000
November 28, 2000

Big Data Definition



The definition of Big Data is typically traced back to the 3Vs model defined by Doug Laney in **2001**: Volume, Velocity, and Variety.



The fourth V was later added in different fashions, such as "Value" or "Veracity".

Valid?

1998
John Mashey coined "Big Data"

2000
Francis Diebold conceptualized "Big Data"

2001
Doug Laney defined "Big Data" as 3Vs

2016 "Big Data" explosively used

2016 "Huge Computing"

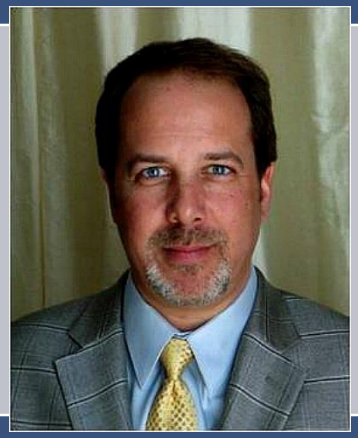
2030 "Huge Computing" booming!!!

1995 2005 2015 2025

2000 2010 2020 2030

Issues

The NIST report of "Big Data Interoperability Framework" lists the problem of multiple descriptions about Big Data in fragmented and incoherent efforts.

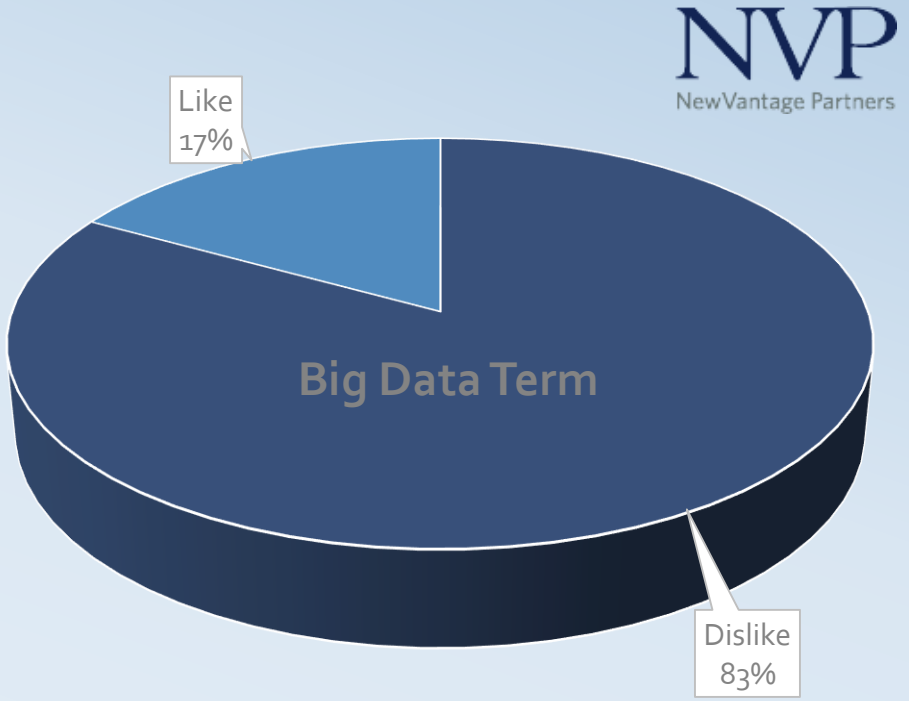


Even the original author of 3Vs admitted that he was simply writing about the burgeoning data in the data warehousing and business intelligence world.

Report by NewVantage Partner

83% of the executives surveyed thought that the term of Big Data was overstated, confusing, or misleading.

They liked the concept, but hated the phrase.



Expert's View

- There is Just One V in Big Data.

- Amar Nadig

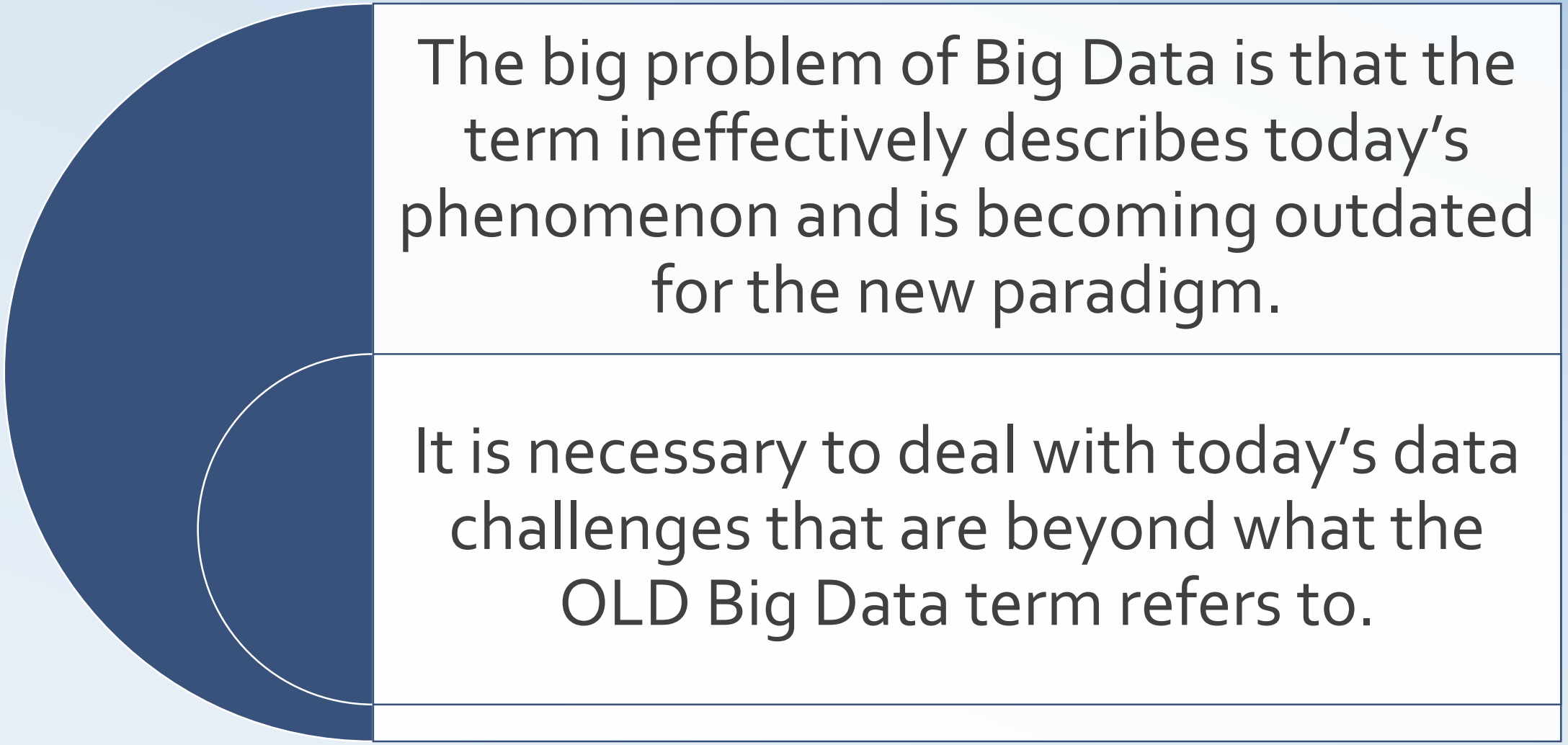
- I do think that big data can be better explained by adding a few more V's.

- Mark Van Rijmenam

- Nobody likes the Big Data term and almost everybody wishes for a better, more descriptive name for it.

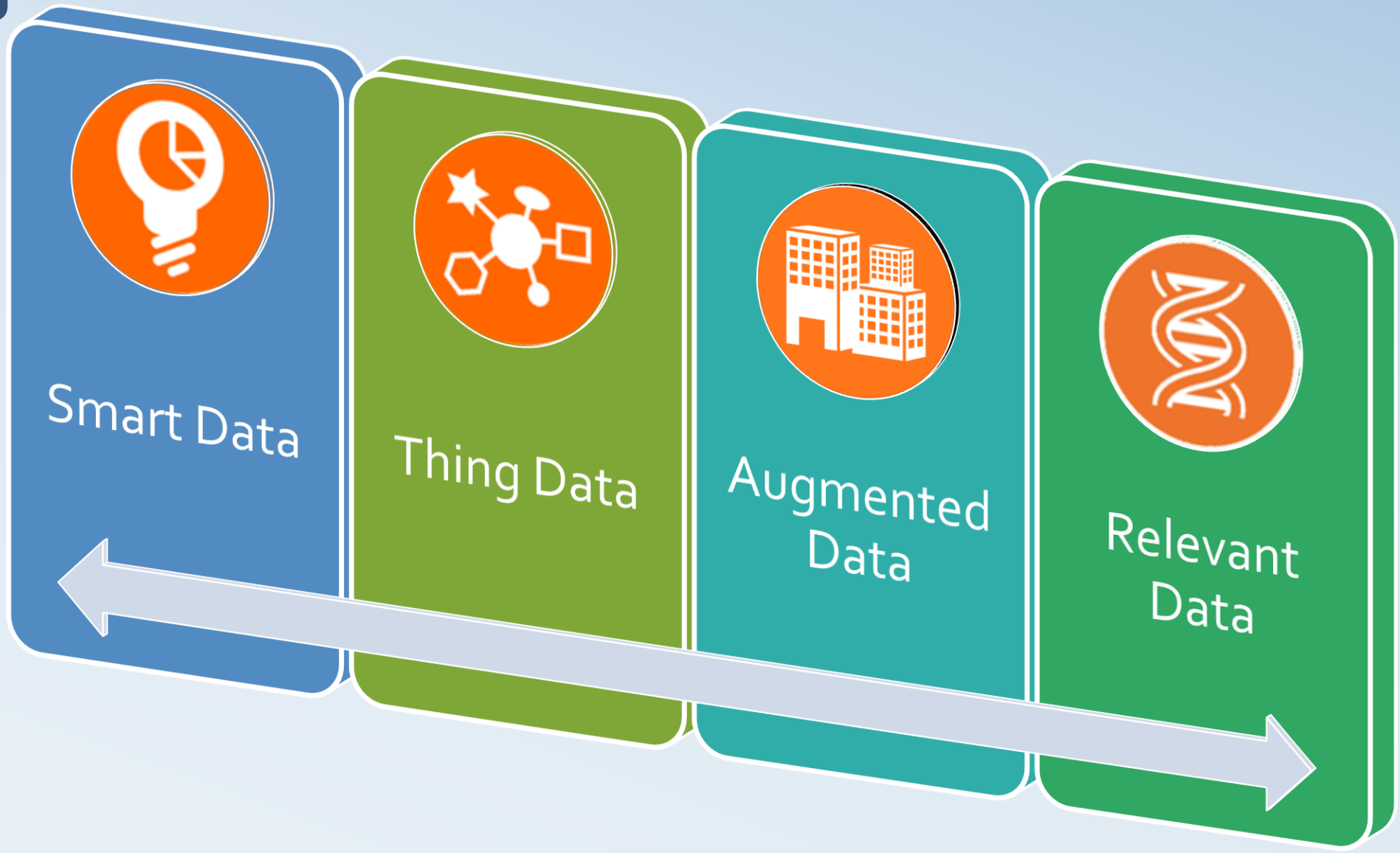
- Tom Davenport

Time to Rethink Big Data



Why Rethink?

Trending

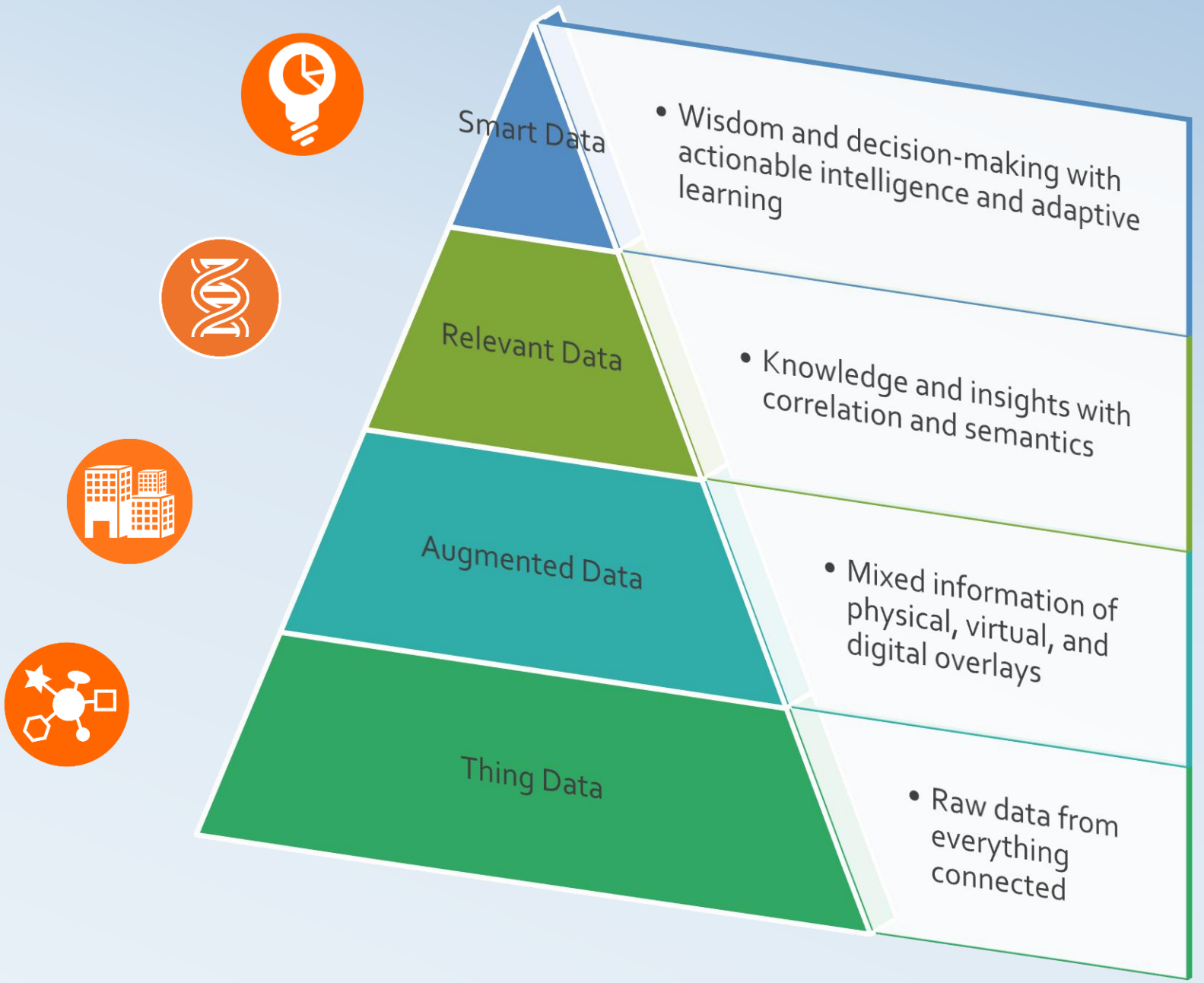


Anatomy

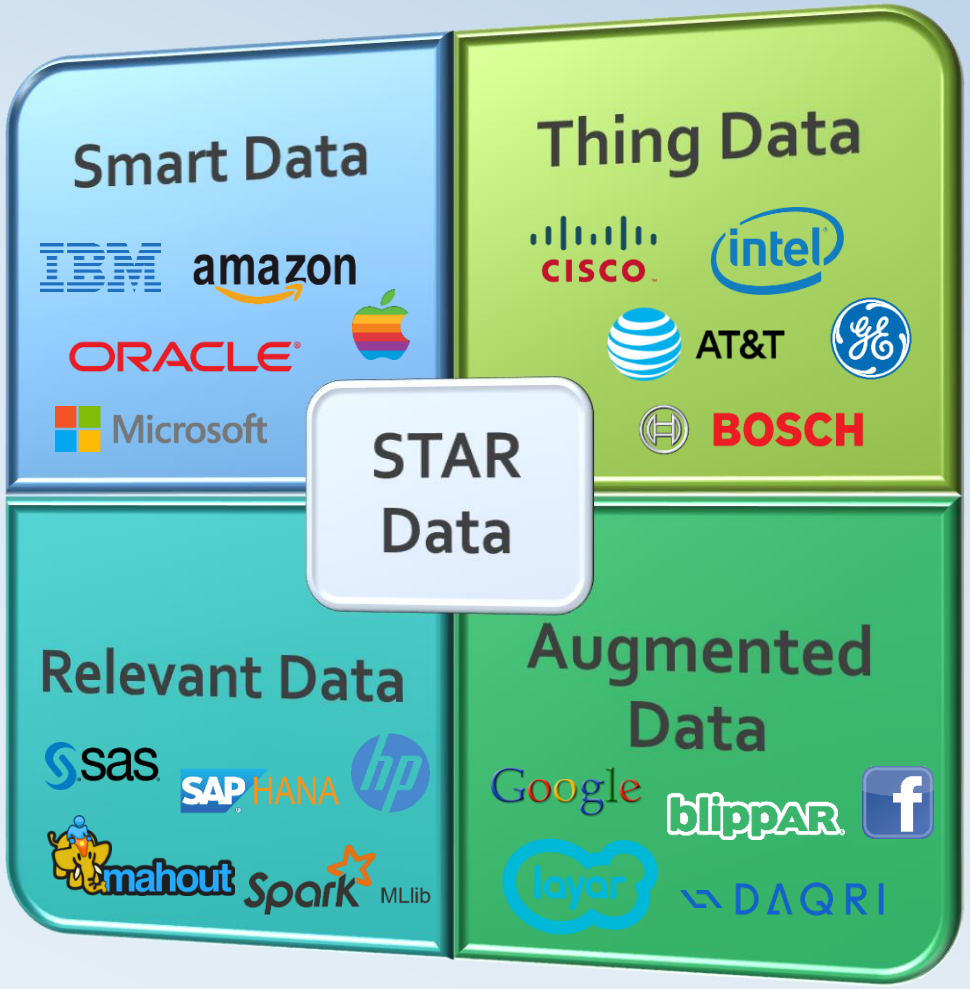
STAR Data

	Description	Application Areas
Smart Data 	Filter out the noise and hold the valuable data. The cognitive computing makes solutions intelligent and self-improving. Knowledgeable reasoning results in more sound decisions and better customer engagements.	Intelligent content personalization leverages all the data that are accumulated from B2C, B2B and C2C channels, to not only optimize the display of customized contents, but also heighten the user experience.
Thing Data 	Data from connected things like devices, mostly machine generated data. IoT Data is raw in nature, small in size, frequent in transmission, and large in volume.	A physical object like a sensor monitors a specific condition such as location, vibration, motion and temperature to collect data. IoT-enabled objects share information about their condition and the surrounding environment with others.
Augmented Data 	Audio, visual and interaction data for Augmented Reality (AR) and Virtual Reality (VR) in a physical 3D or cyber-physical converged environment. The vision, NLP, gesture and emotion recognition enable the surrounding real world become more interactive and digitally manipulatable.	AR enhances product previews by allowing a customer to view what's inside a product's packaging without opening it. It is used to let a doctor look inside a patient by combining one source of images such as an X-ray/ultrasound with another such as video of an endoscope to inspect a tumor.
Relevant Data 	Data relationship is critical to identify pertinence in the data set, which leads to profounder understanding of seemingly unrelated events and sequence. The focus is to contextualize each bit of data with its own specific semantics in predictive and preventive analytics.	Social analytics, as an example, empower businesses to distill the hidden values behind the reams of social data and activities, in order to glean actionable insights about the user's preference, behavior, habits and influence.

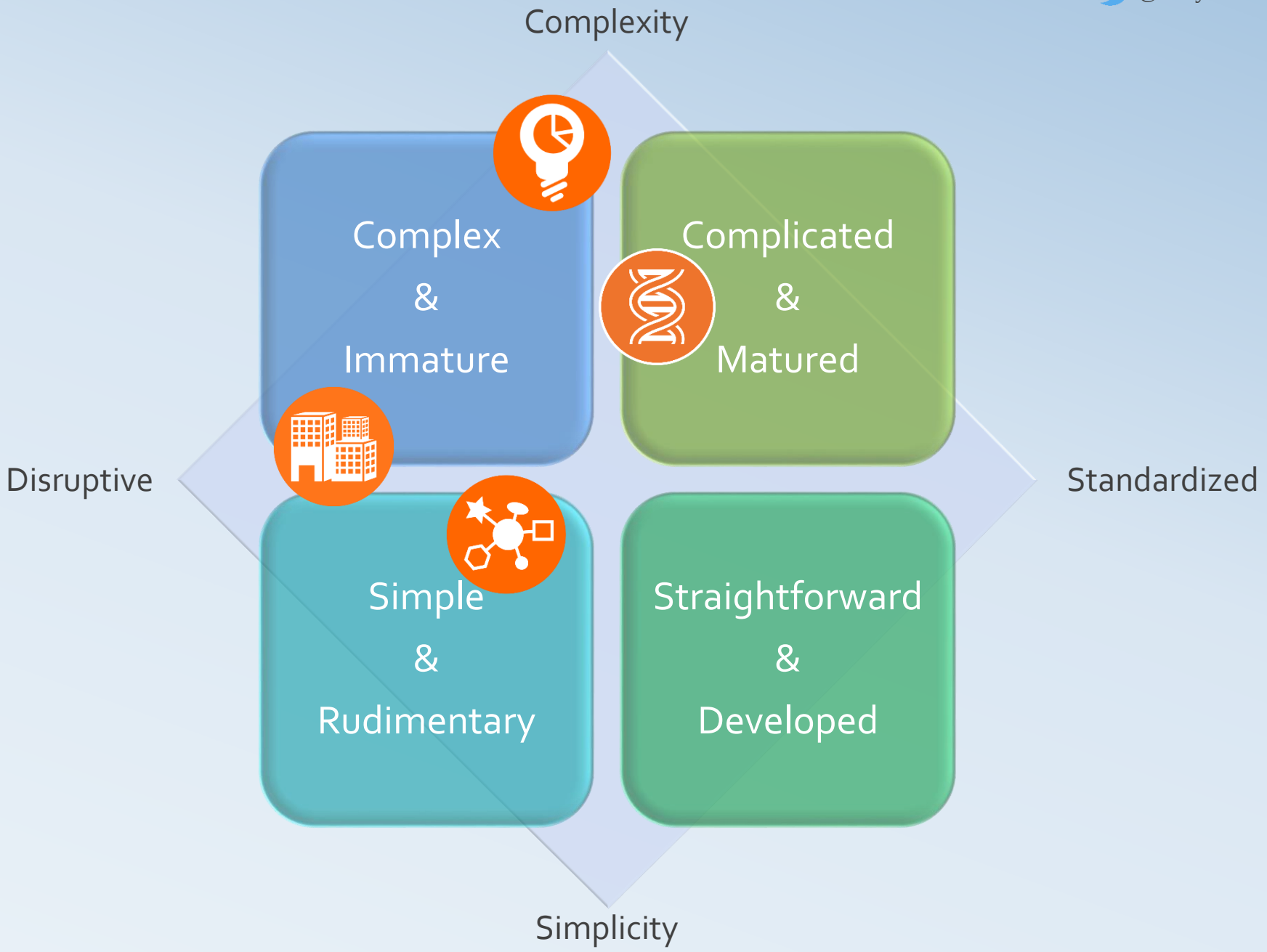
Why



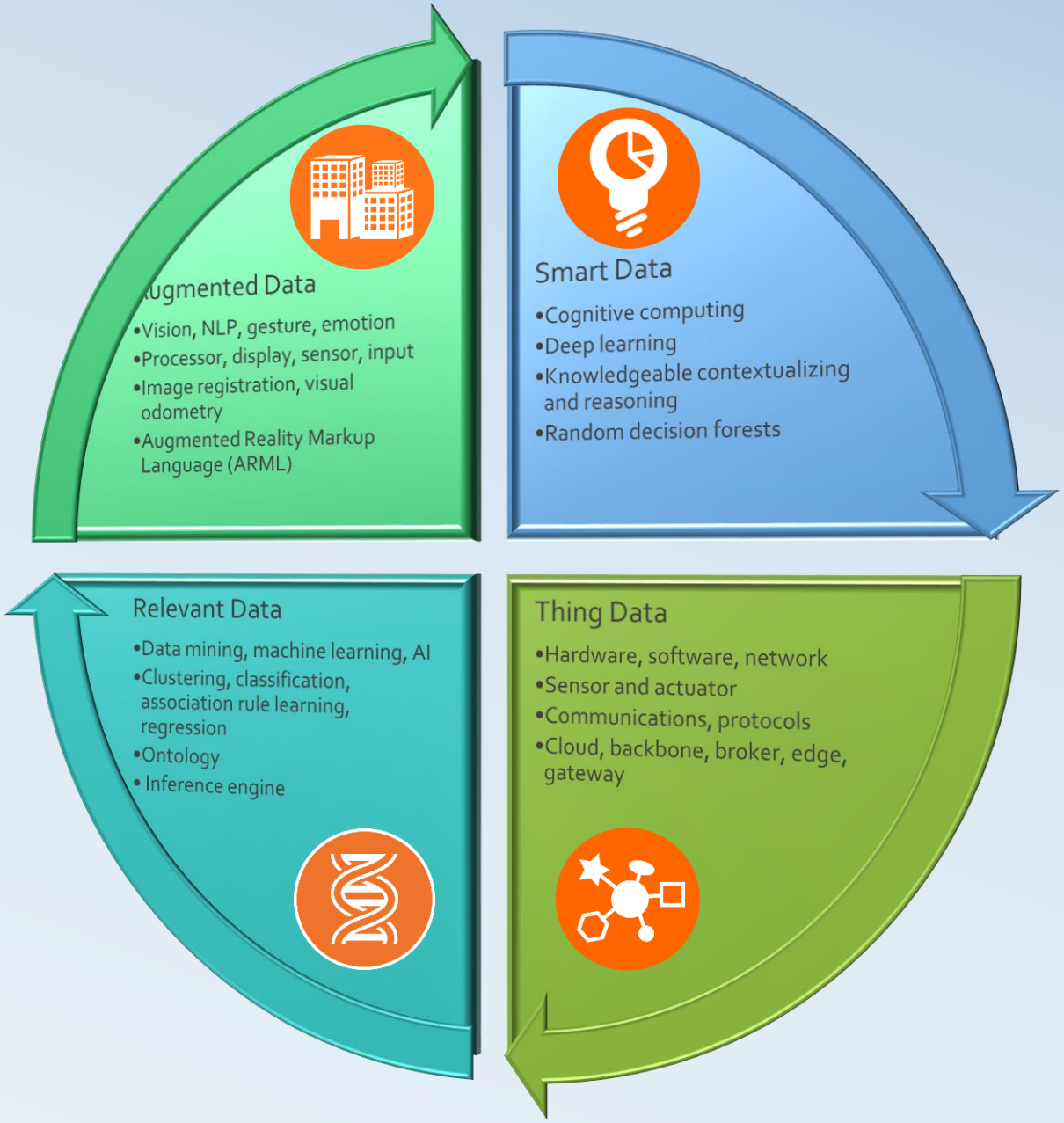
Who



When



How



IoT Protocols

Framework
- Alljoyn
- IoTivity
- Weave
- Homekit

Semantic
- JSON-LD
- Web Thing Model

Device Management
- TR-069
- OMA-DM

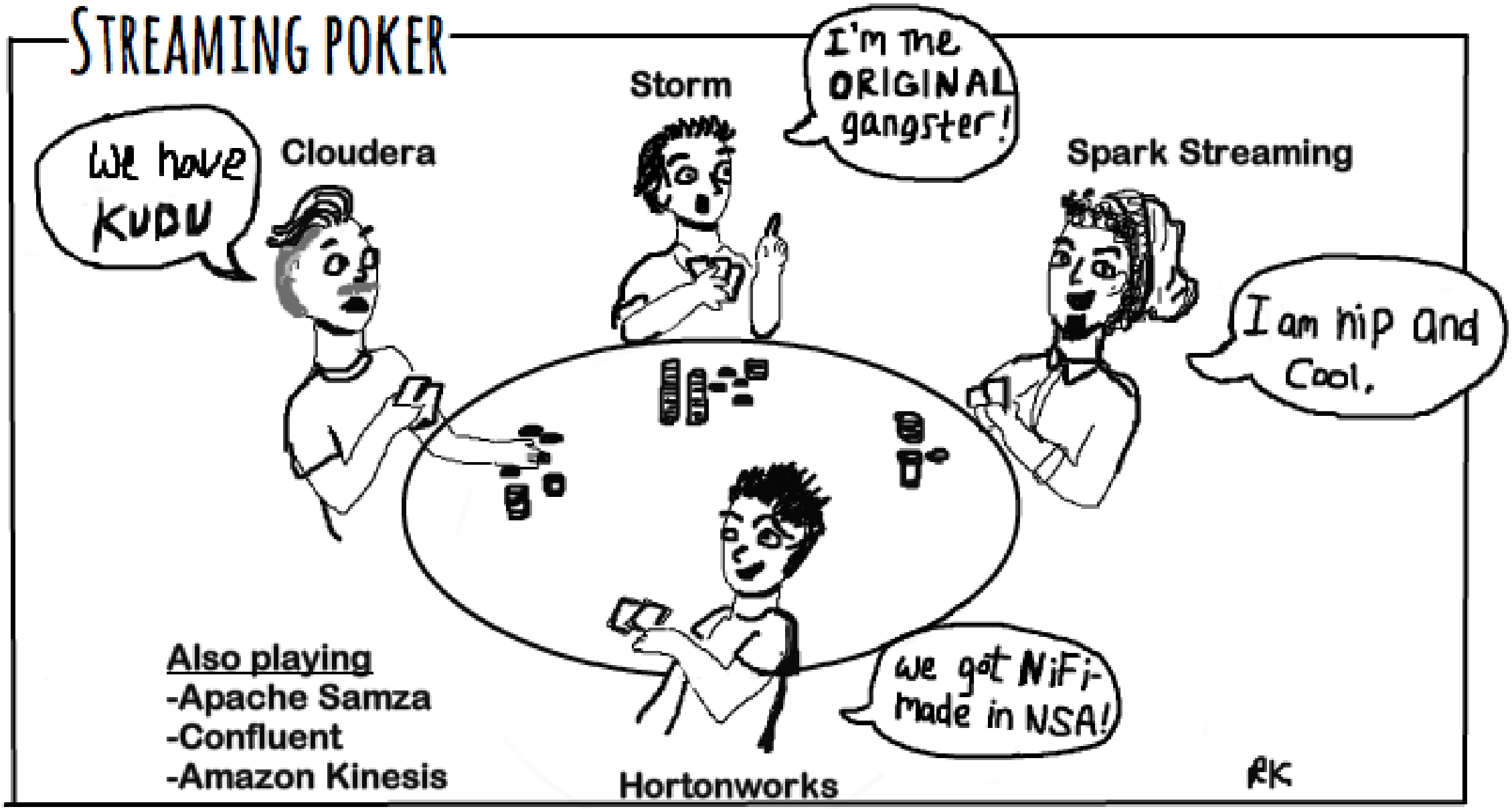
Data
- MQTT
- CoAP
- AMQP
- Websocket
- Node

Discovery
- Physical Web
- mDNS
- DNS-SD

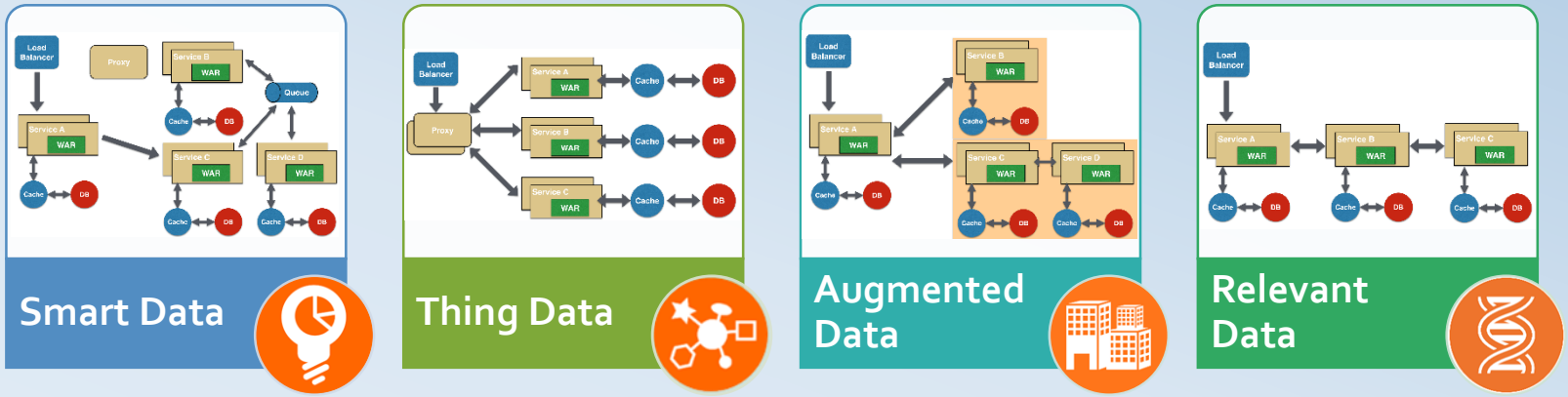
Transport
- Wifi
- Bluetooth
- LPWAN

Identification
- EPC
- uCode
- IPv6
- URIs

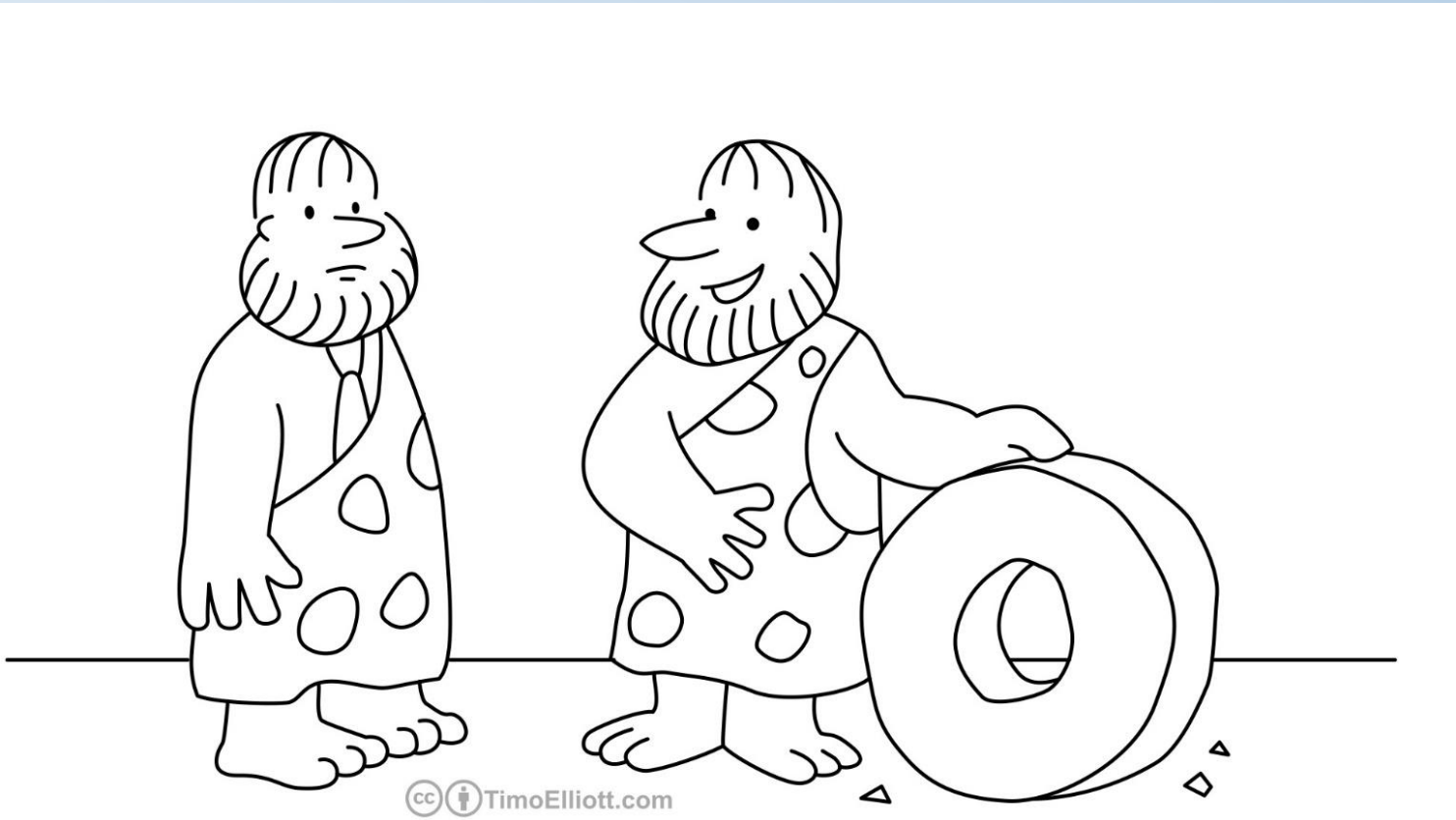
Infrastructure
- 6LowPAN
- IPv4/IPv6
- RPL



STAR Data Execution



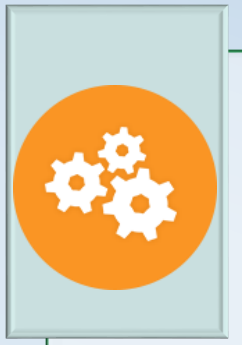
Do STAR Data Wisely



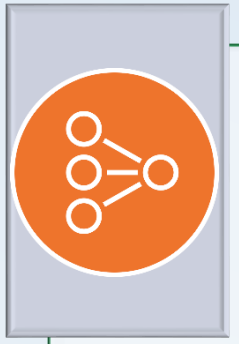
"It does look similar—but this one is powered by Hadoop"

Best Practices

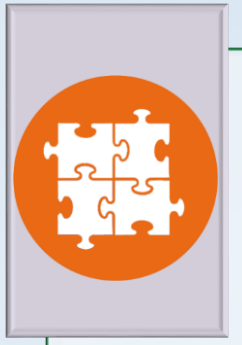
Convert



Condense

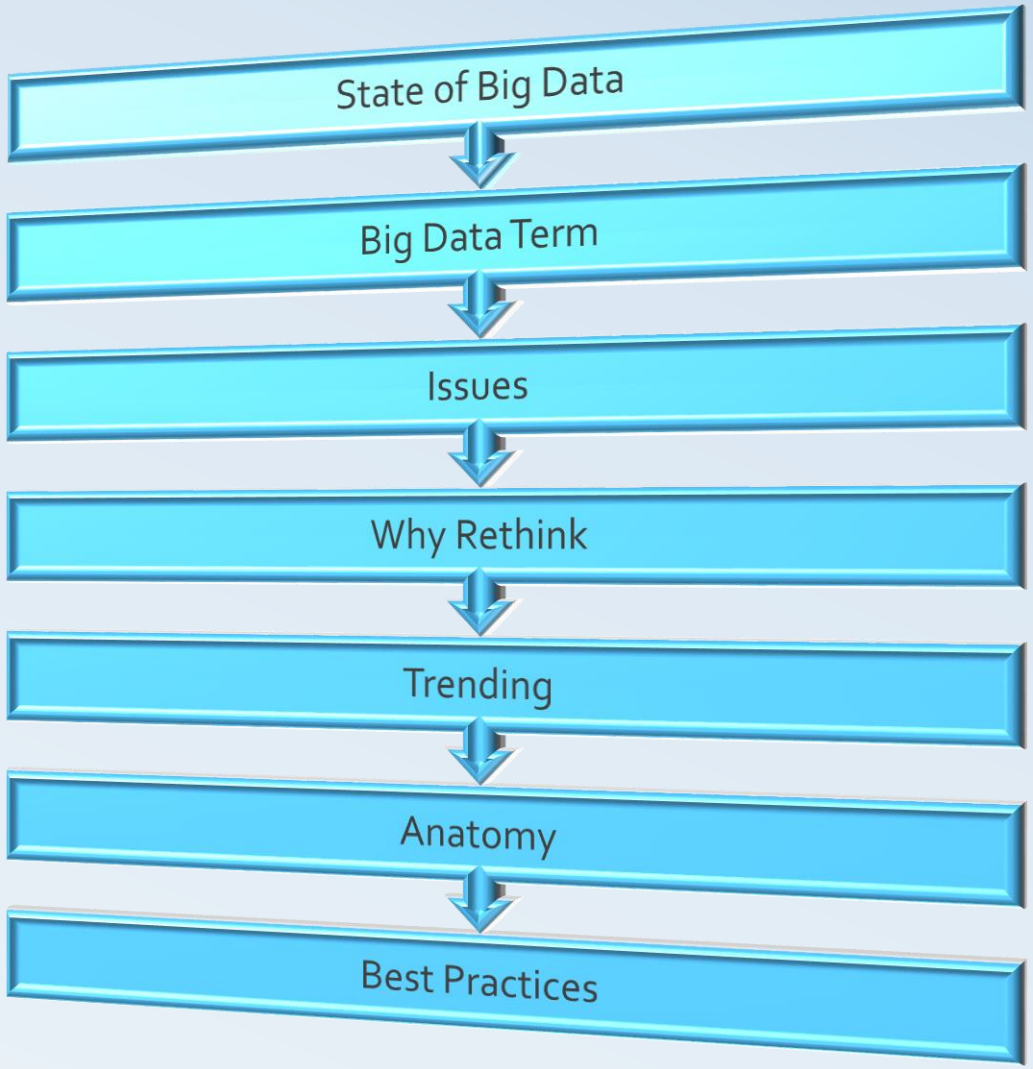


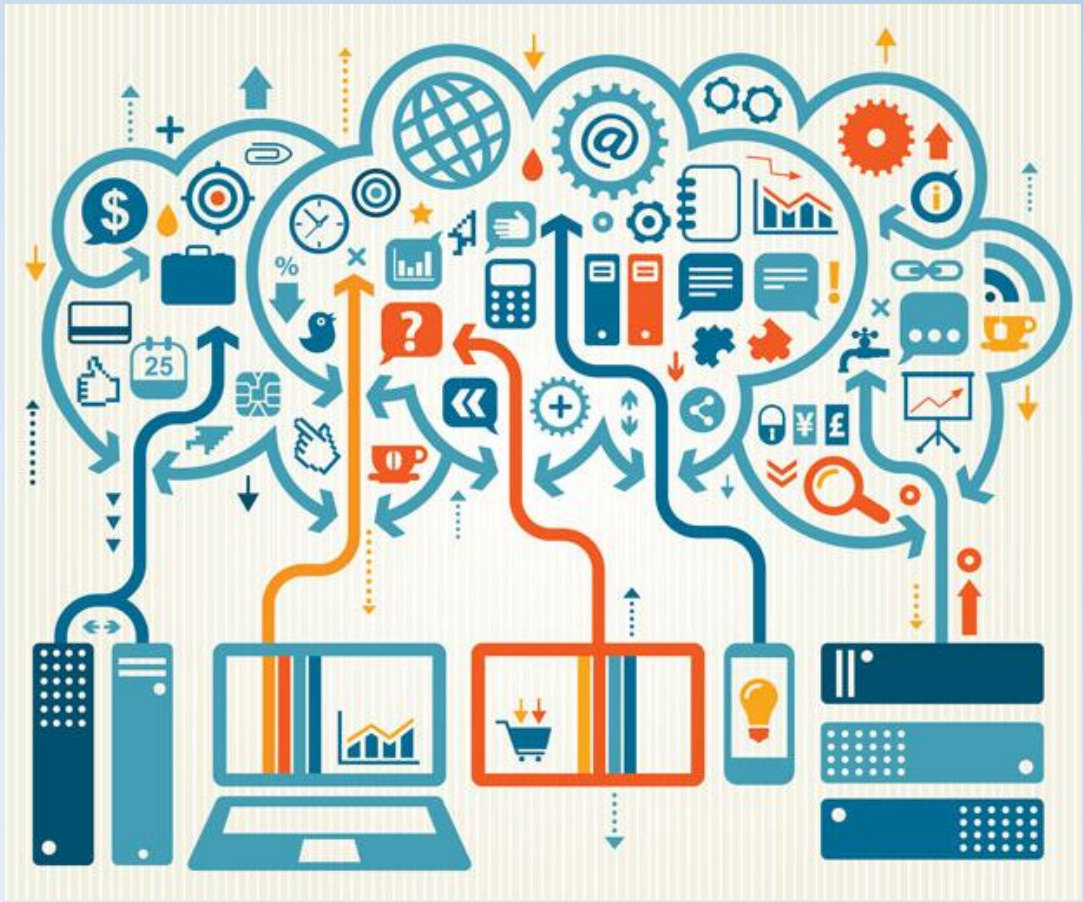
Converge



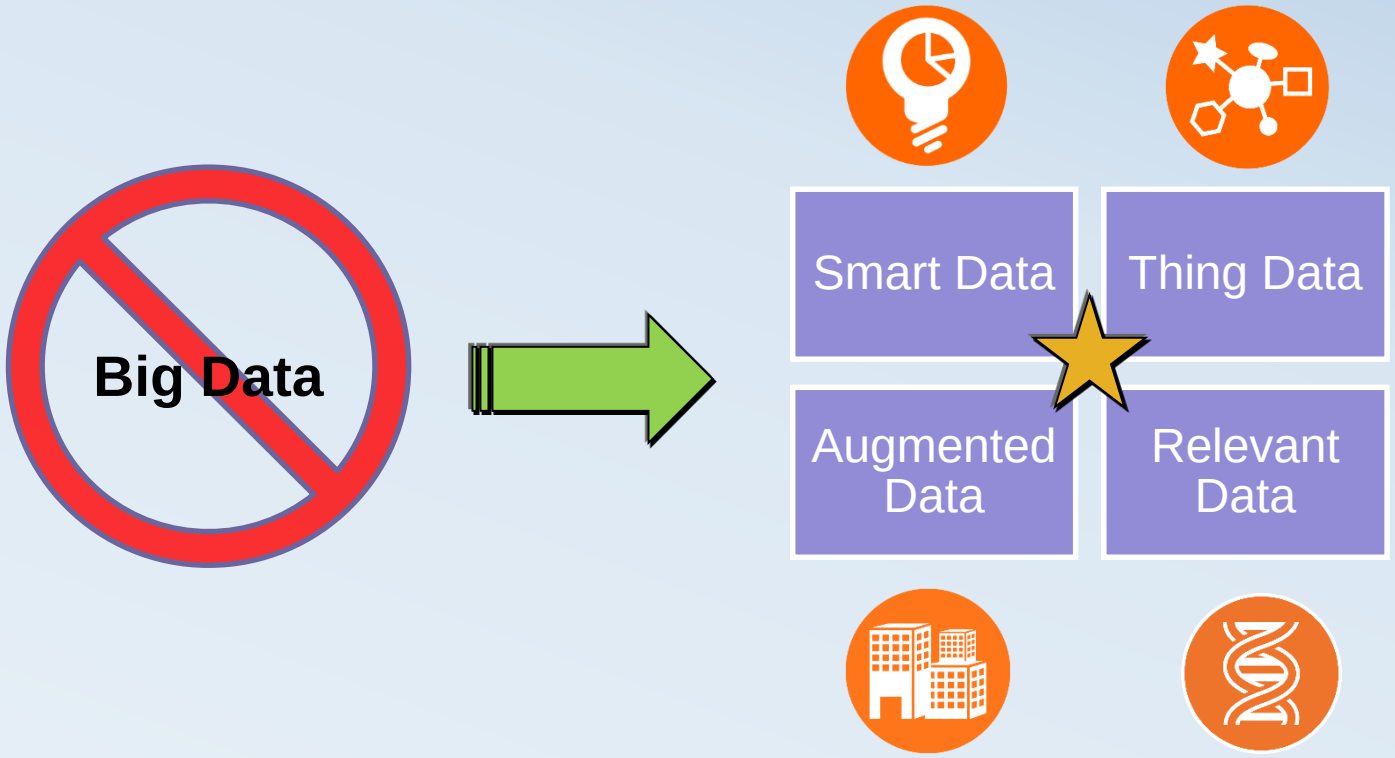
Convince

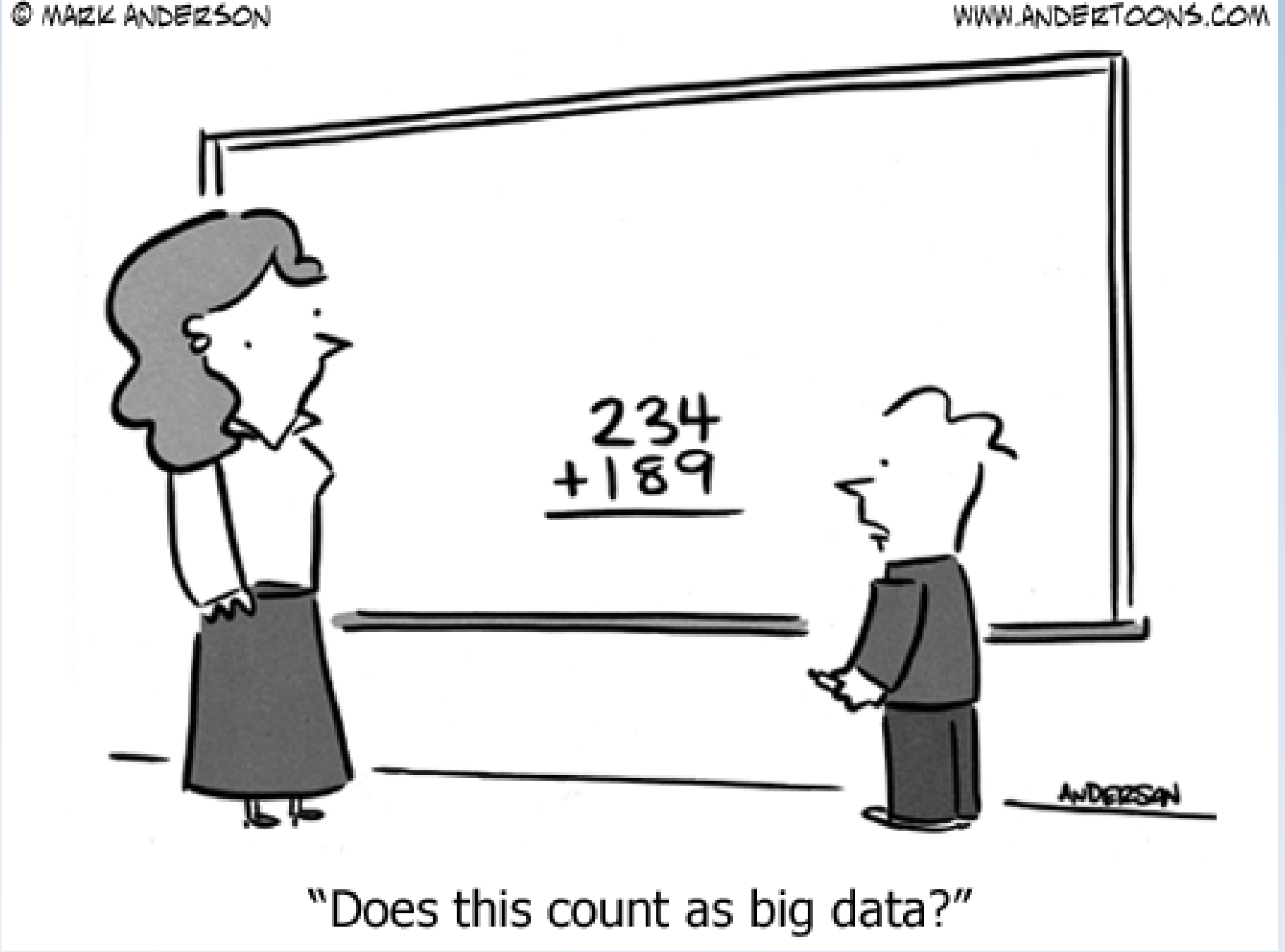
Summary





Key Takeaways





"Does this count as big data?"

Contact

mail@tonyshan.com