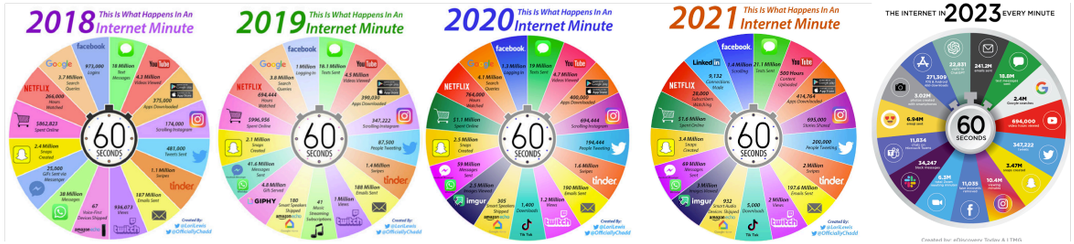
Data*

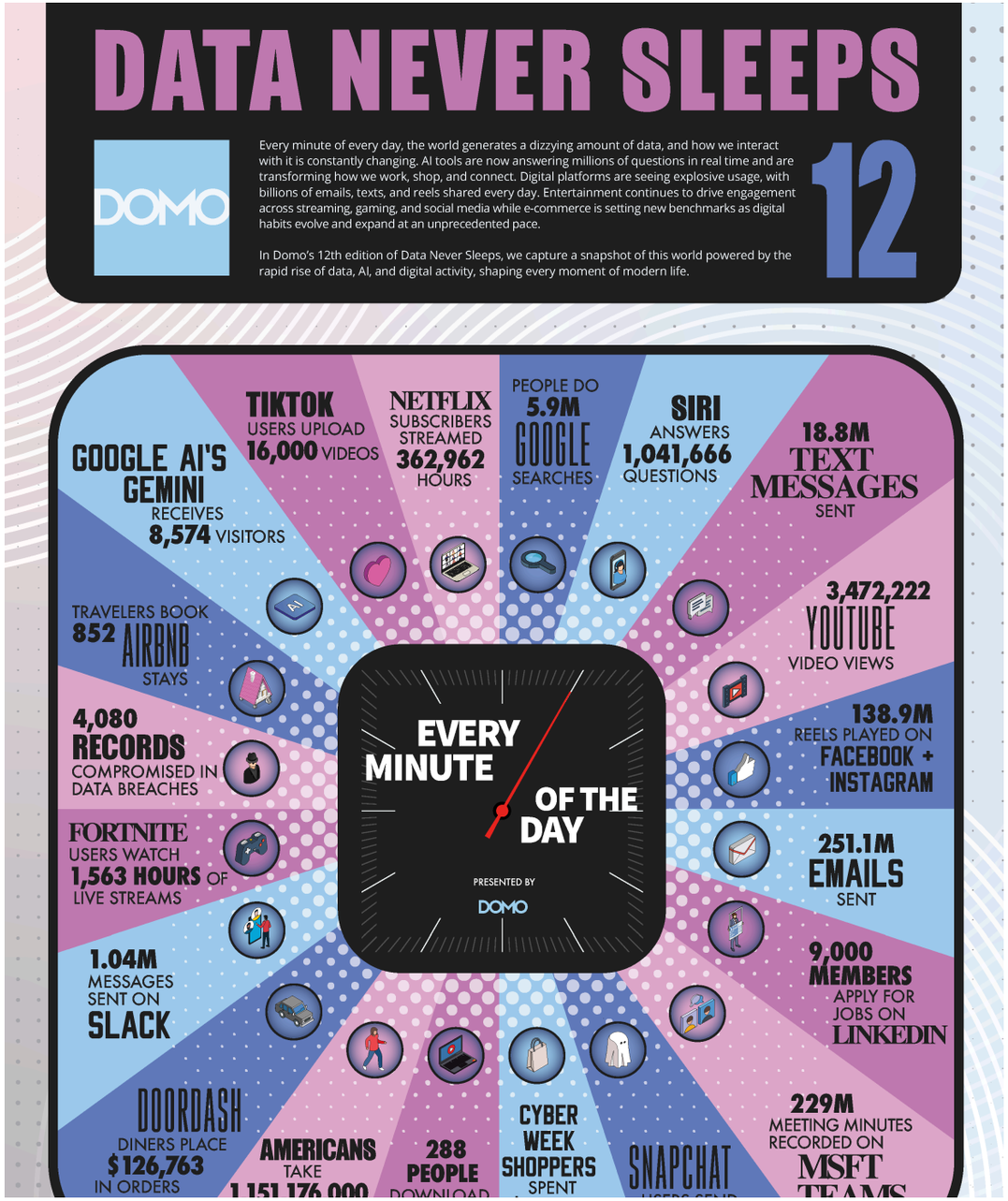
Salvo Nicotra

We live in a data World

Internet Minute



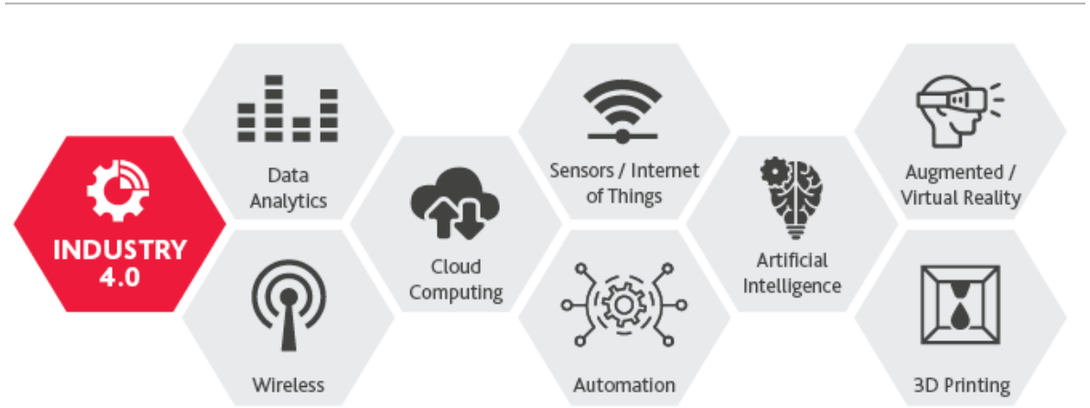
Data never sleeps



Data Revolution

From 4.0: The Information Age

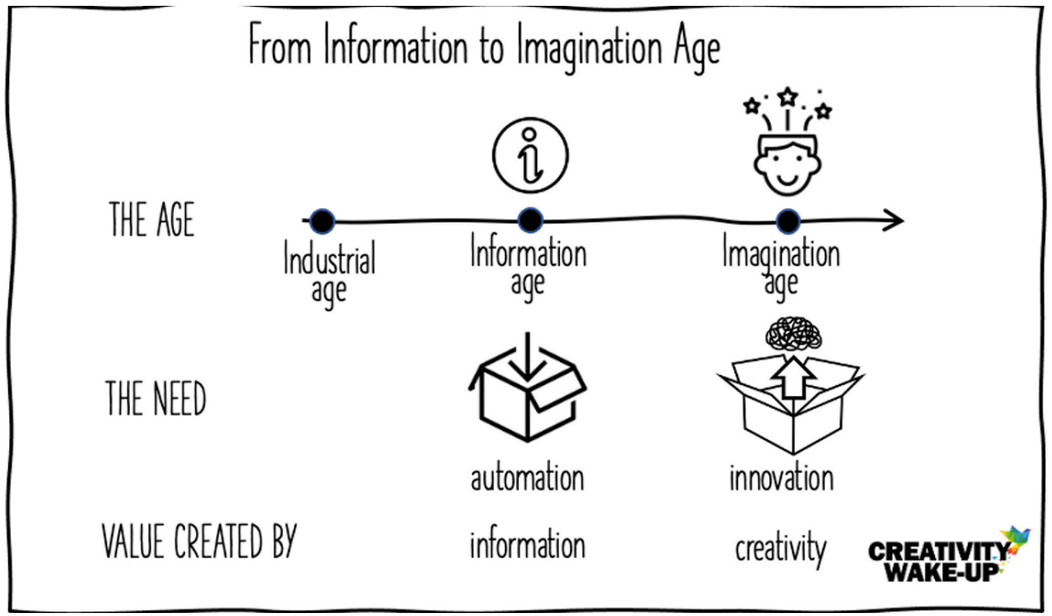
> In essence, the Fourth Industrial Revolution is the trend towards automation and data exchange in manufacturing technologies and processes which include cyber-physical systems (CPS), IoT, industrial internet of things,cloud computing,cognitive computing, and artificial intelligence.



Industry 4.0

https://en.wikipedia.org/wiki/Fourth_Industrial_Revolution

to the Imagination Age



NicsMeme

https://www.creativitywakeup.com/blog/imagination_age

towards Industry 5.0

> The Industry of the Future approach brings benefits for industry, for workers and for society. It empowers workers, as well as addresses the evolving skills and training needs of employees. It increases the competitiveness of industry and helps attract the best talents.

https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en

What is data ?

data

> The Latin word data is the plural of datum, “(thing) given”, neuter past participle of dare, “to give”. The first English use of the word “data” is from the 1640s. The word “data” was first used to mean “transmissible and storable computer information” in 1946. The expression “data processing” was first used in 1954.

Wikipedia

data

> **Information**, especially facts or numbers, collected to be examined and considered and used to help decision-making, or **information** in an electronic form that can be stored and used by a computer

Cambridge Dictionary

information

> facts about a situation, person, event, etc.:

Cambridge Dictionary

The Key Differences Between Data vs Information

https://bloomfire.com/blog/data-vs-information/

- Data is a collection of facts, while information puts those facts into context.
- While data is raw and unorganized, information is organized.
- While data is raw and unorganized, information is organized.
- Data points are individual and sometimes unrelated. Information maps out that data to provide a big-picture view of how it all fits together.
- Data, on its own, is meaningless. When it’s analyzed and interpreted, it becomes meaningful information.
- Data does not depend on information; however, information depends on data.
- Data typically comes in the form of graphs, numbers, figures, or statistics. Information is typically presented through words, language, thoughts, and ideas.
- Data isn’t sufficient for decision-making, but you can make decisions based on information.

Data -> Information

data is **information** stored that can be analyzed in a specific context to get the same or new information

Suppose we do have the “data” written on a paper, what does it mean ?

We need some context (or metadata) to understand what these data means, it could be:

- Alice has 10 years;
- Alice has 10$ in the wallet

Qualitative vs Qualitative

Quantitative data

- is numeric.
- Measures peoples, things and events.
- Can be integer, float. Data
- Operations: It’s possible to sum, compute max, min, avg

Qualitative data

- include descriptions, like name, surname, address or product name
- May be associated with numeric information such as price or age.
- It’s normally a text, can be a categorical value or a description
- Operations: It’s possible to count

What about time ?

Chat Gpt response:

> The answer to your question is not straightforward. Time can be both quantitative and qualitative depending on the context in which it is used In general, time is considered a quantitative data type because it can be measured and expressed numerically, such as in seconds, minutes, hours, days, weeks, months, or years. For example, the duration of a movie can be expressed in minutes, and the time it takes to complete a task can be expressed in hours. However, time can also be considered qualitative when it is used to describe a subjective experience or a phenomenon that cannot be measured precisely. For example, the phrase “time flies when you’re having fun” is a qualitative expression that describes the subjective experience of time passing quickly when one is enjoying oneself

ISO 11404:2007 Information Technlogy - General purpose datatypes

http://standards.iso.org/ittf/PubliclyAvailableStandards/c039479_ISO_IEC_11404_2007(E).zip

Three notions of datatype:

- the **conceptual**, or abstract, notion of a datatype, which characterizes the datatype by its nominal values and properties;
- the **structural** notion of a datatype, which characterizes the datatype as a conceptual organization of specific component datatypes with specific functionalities;
- the **implementation** notion of a datatype, which characterizes the datatype by defining the rules for representation of the datatype in a given environment.

Data Types

Numeric vs Non Numeric

- A datatype is said to be **numeric** if its values are conceptually quantities (in some mathematical number system).
- A datatype whose values do not have this property is said to be **non-numeric**.

Primitive vs Non primivite

- **primitive** datatypes
  - defined axiomatically without reference to other datatypes
  - Atomic: not divisible
  - Examples: boolean, integer, time, enumerate, char, varchar
- **generated** datatypes
  - specified, and partly defined, in terms of other datatypes.
  - aggregate: values organized for a specific function
  - Examples: array, bag, sequence, xml, json

Big Data

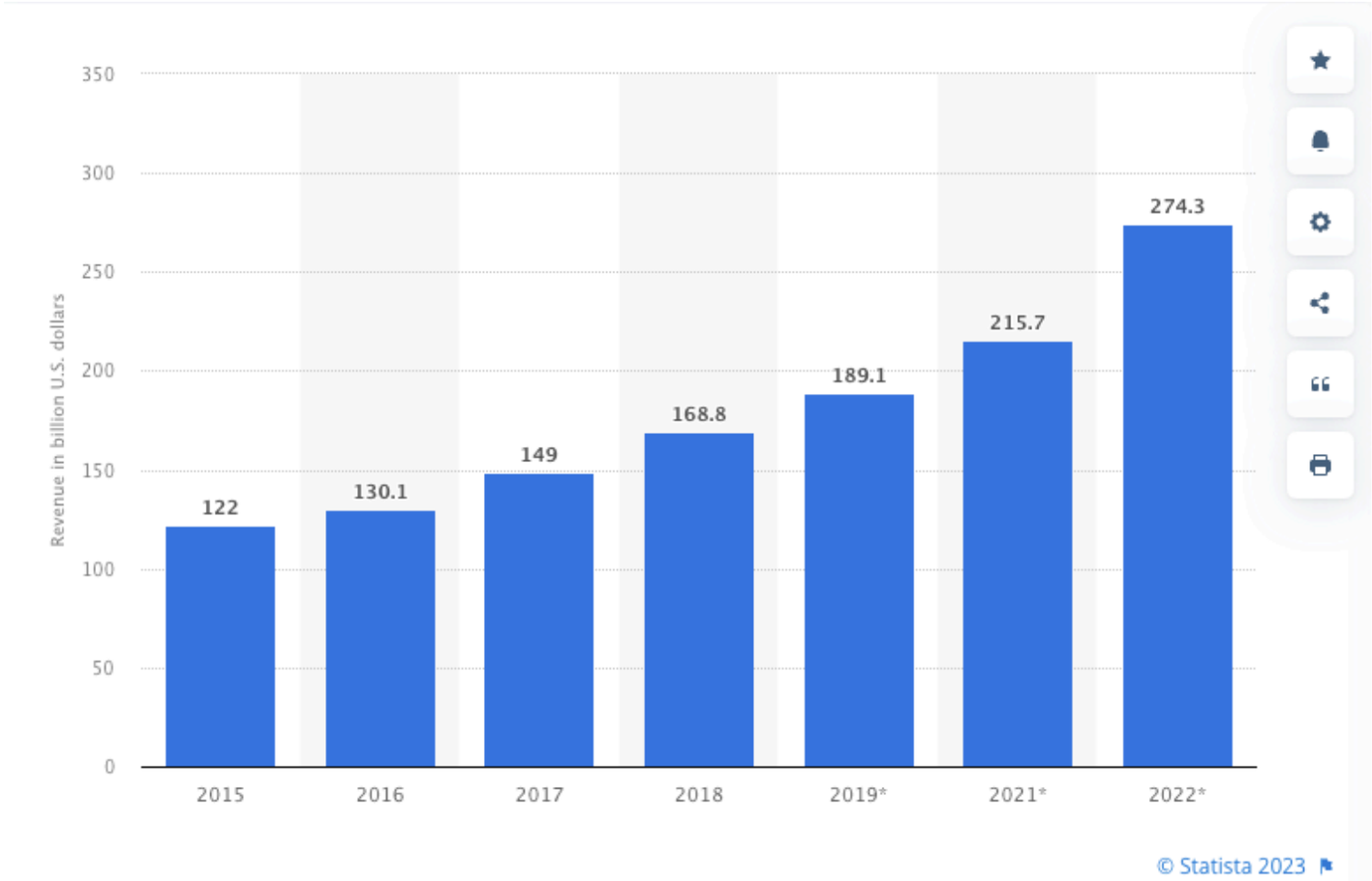
Warning

> Big data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.” (Dan Ariely, American Professor)

https://towardsdatascience.com/top-15-famous-data-science-quotes-f2e010b8d214

What is Big Data ?

A market perspective Revenue from big data and business analytics worldwide from 2015 to 2022 (Statista -

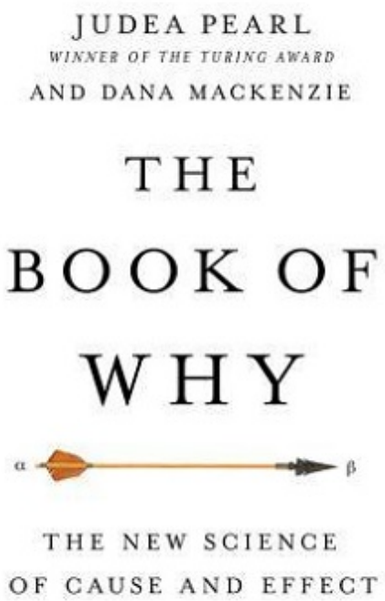


IDC)

https://www.statista.com/statistics/551501/worldwide-big-data-business-analytics-revenue/

Big data is better data

The Book of Why



> Data can tell you that the people who took a medicine recovered faster than those who did not take it, but they can't tell you why. Maybe those who took the medicine did so because they could afford it and would have recovered just as fast without it.

V for ?

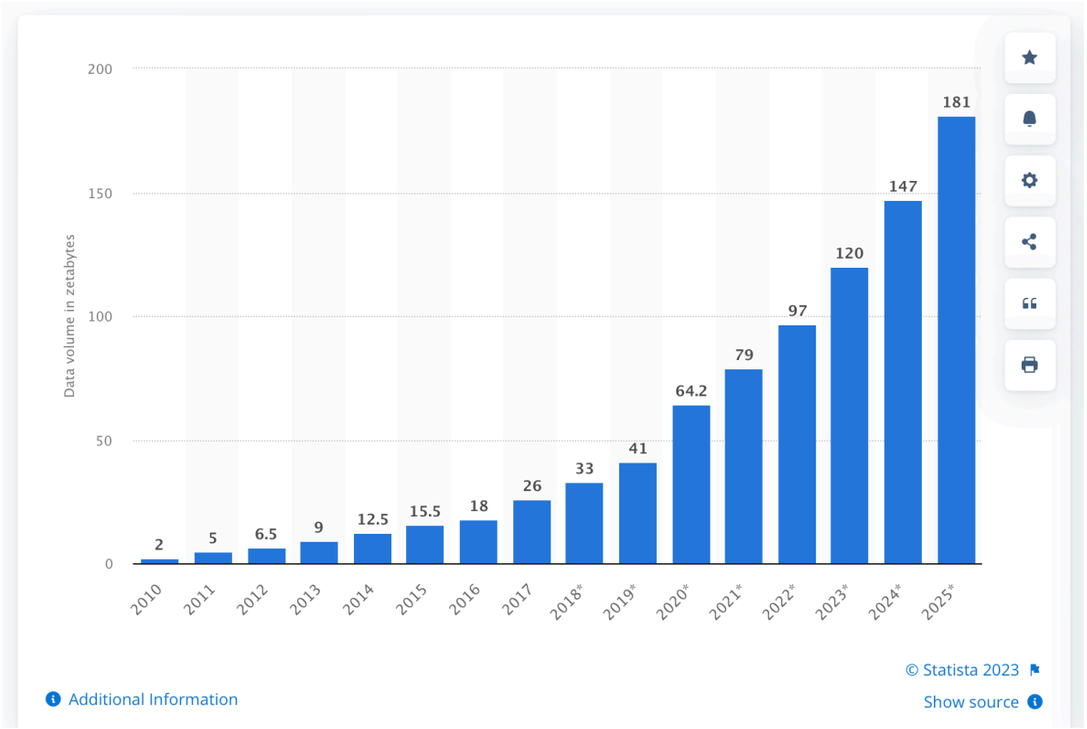
Volume

From day one, the **Zettabyte** has been a benchmark that our analysts have targeted as a major networking milestone.

When will global Internet traffic reach an annual run rate of one Zettabyte? Well, that day has finally come. According to our estimate, the world's collective Internet use will reach the Zettabyte threshold for this calendar year on September 9, 2016

https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-that

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes)



https://www.statista.com/statistics/871513/worldwid data-created/

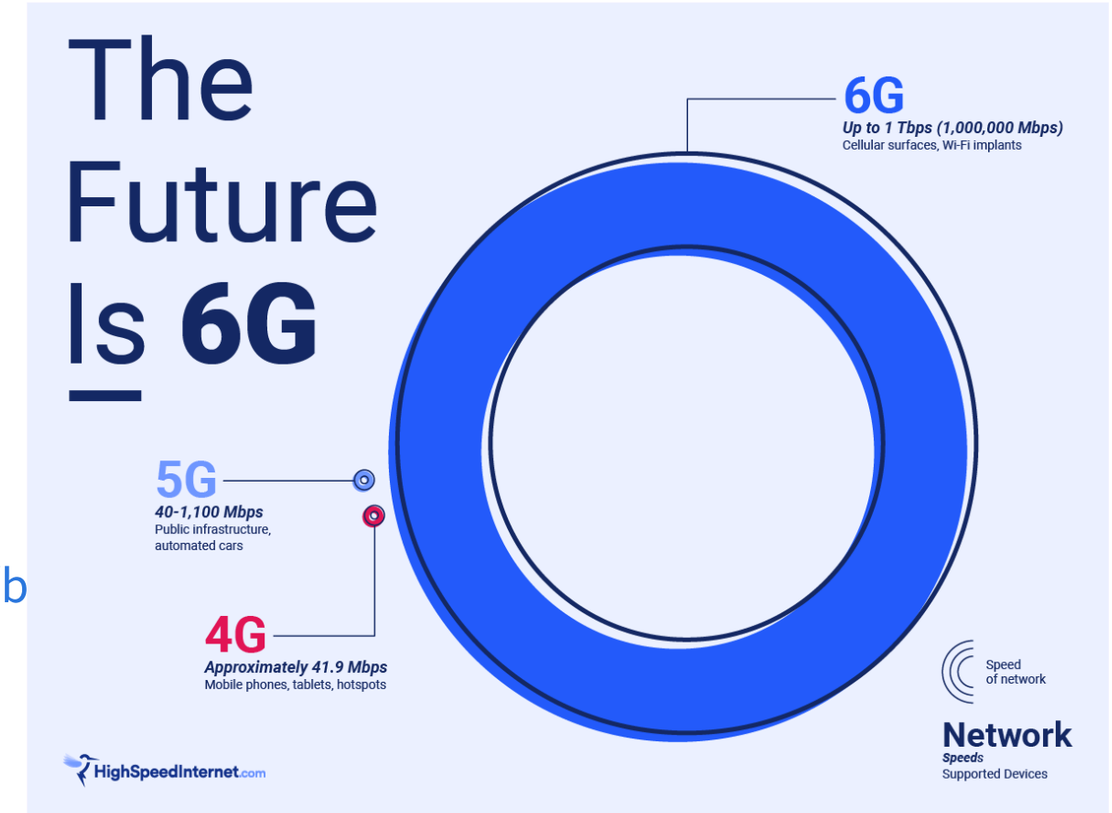
https://www.reddit.com/r/196/comments/uonbxf/rulezip/

Velocity

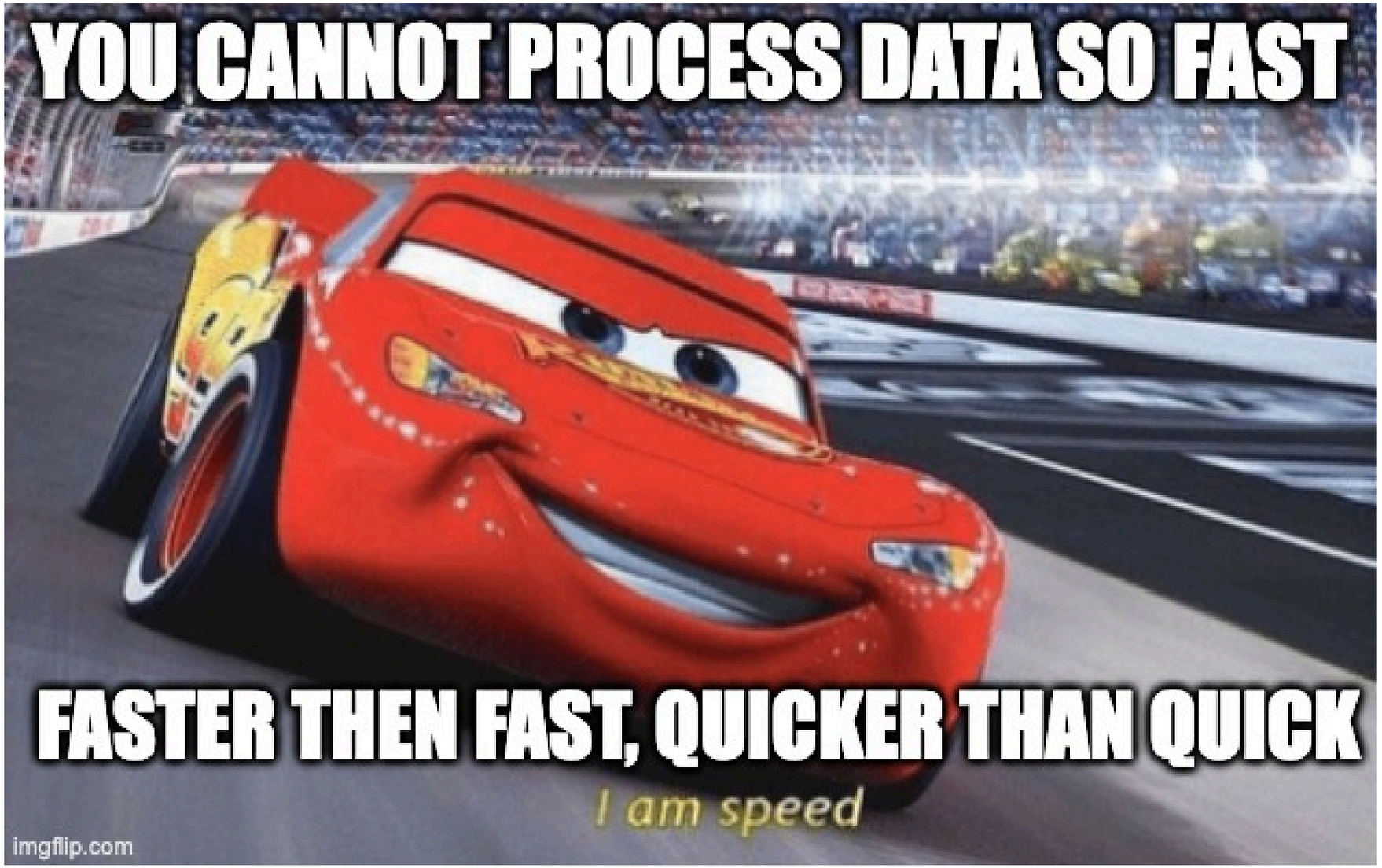
In contrast, data growth shows no sign of stopping. The ongoing explosion of data will result in approximately 40 trillion gigabytes (or 40 zettabytes) in 2020 alone. That's 1.7MB of data created every second for every person on earth.

https://www.ibm.com/blogs/journey-to-ai/2020/06/netezza-and-ibm-cloud-pak-a-knockout-combo-for-tough-data/

https://www.forbes.com/sites/brentdykes/2017/06/28/b data-forget-volume-and-variety-focus-on-velocity/?sh=e9e8b906f7d6



https://www.highspeedinternet.com/resources/6g-internet


YOU CANNOT PROCESS DATA SO FAST
FASTER THEN FAST, QUICKER THAN QUICK
I am speed
imgflip.com

NicsMeme

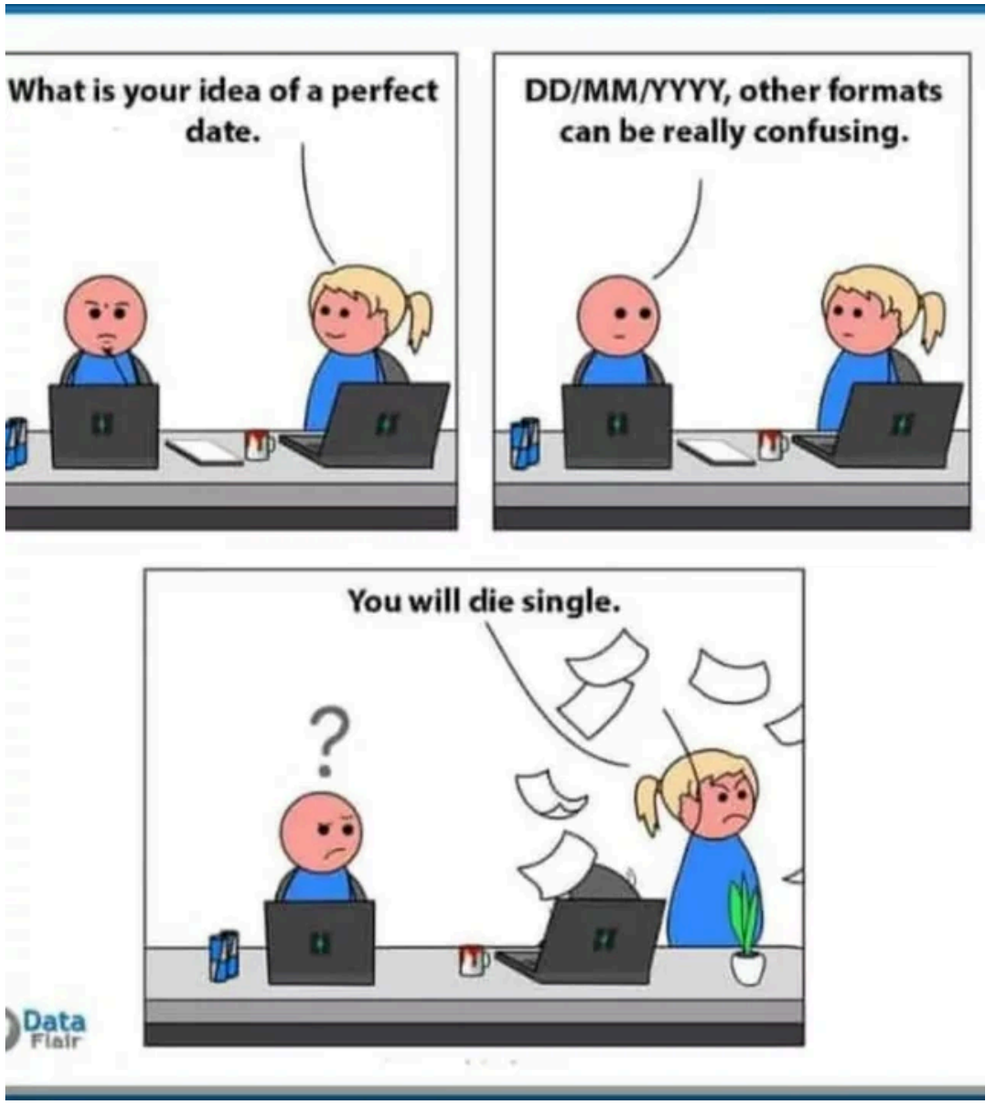
Variety

https://www.futurelearn.com/info/courses/applied-big-data-analytics/0/steps/52404

Variety refers to the complexity of data formats. Big data consists of different forms of data. For example, when a telecommunications company like Telstra records data on calls to its call centre, this data includes both:

- structured data, which conforms to a predefined data model (e.g., your customer ID, the timestamp of your call, your service type), and
- unstructured data (e.g., the recording of the call, notes that the call centre operator makes during the call, the problem history related to your call).

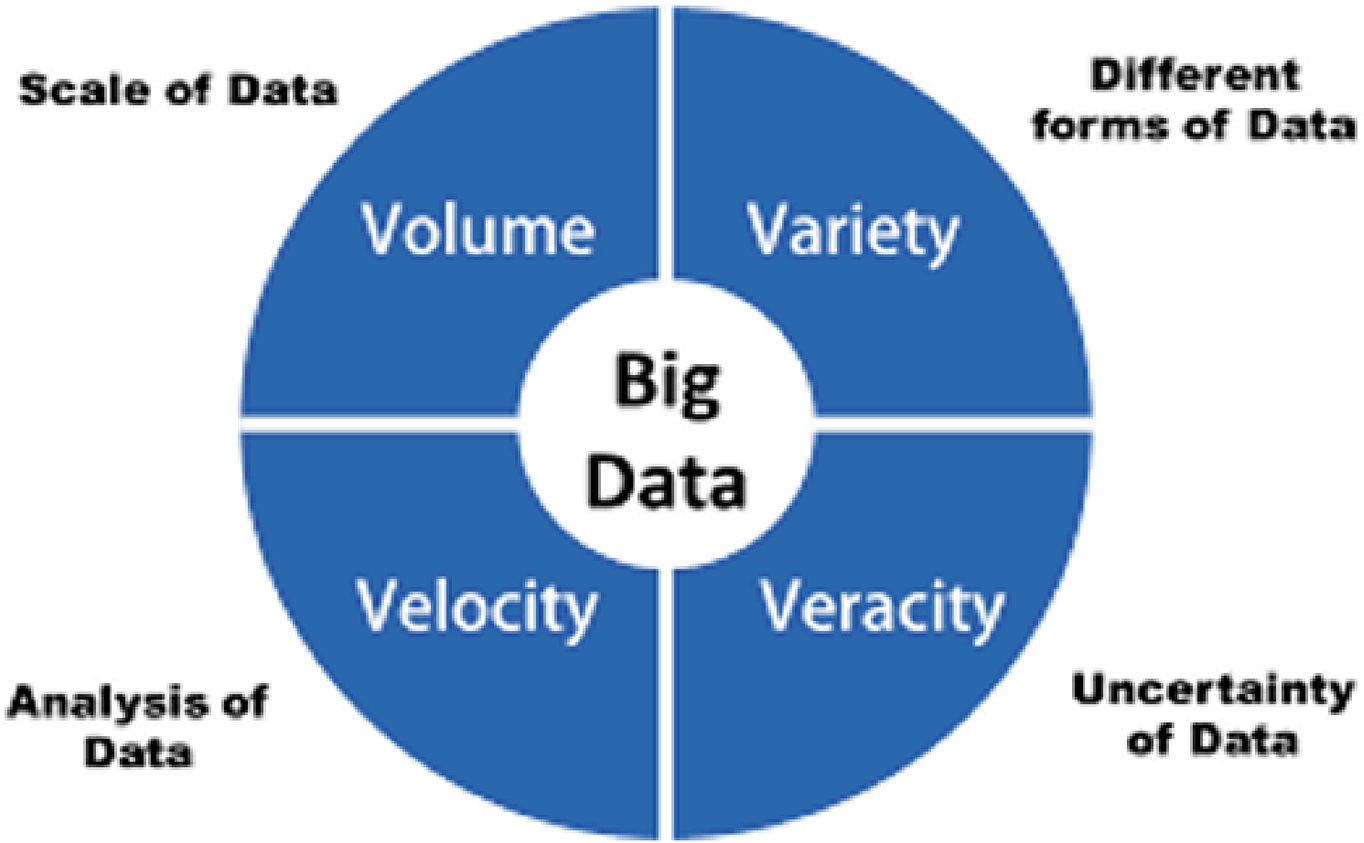
https://towardsdatascience.com/big-data-file-formats-explained-dfaabe9e8b33



https://www.reddit.com/r/ProgrammerHumor/comments/aehe8q/perfect_date/

4V

Oracle added Veracity



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

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

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

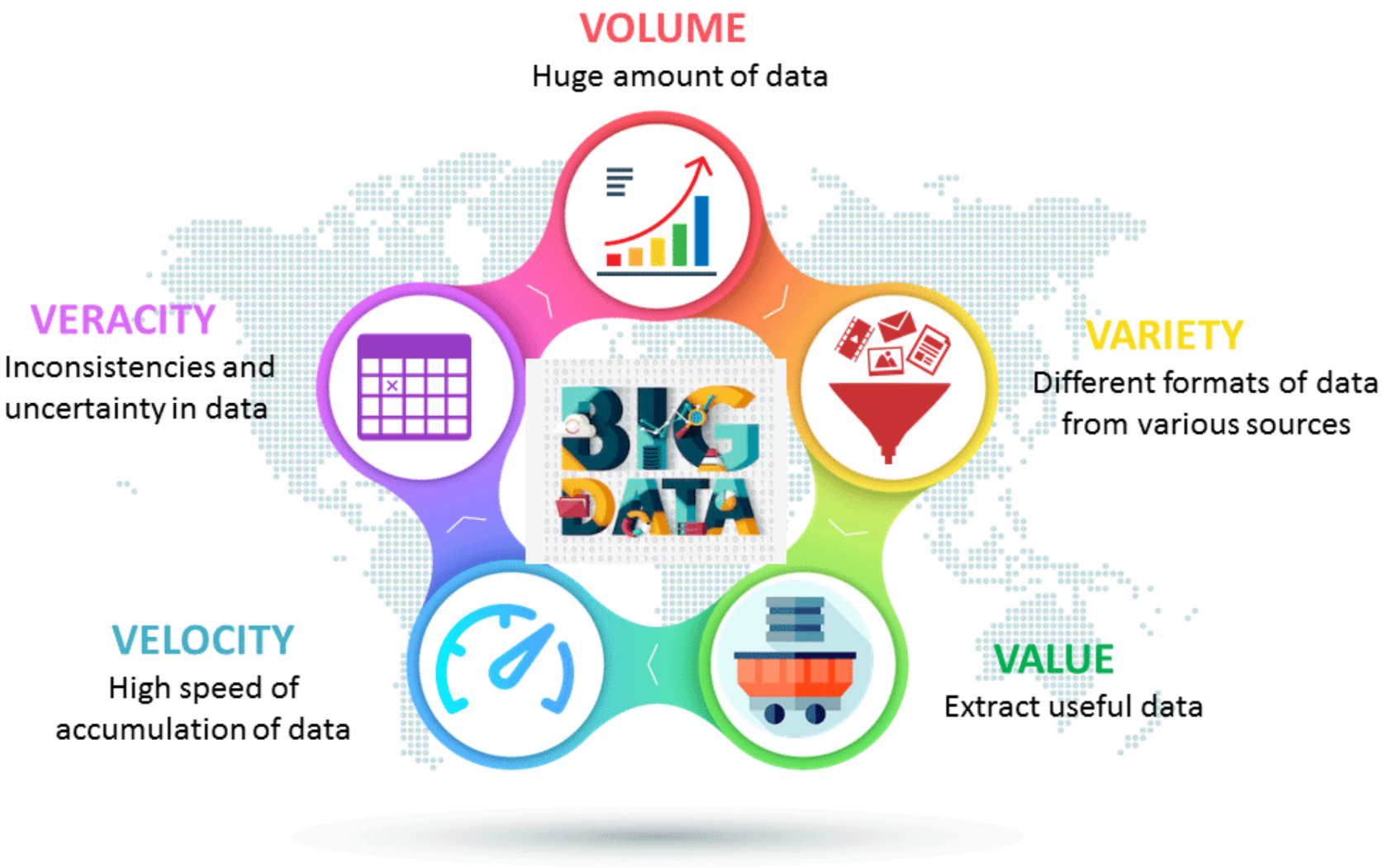
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

5V

Oguntimilehin added Value

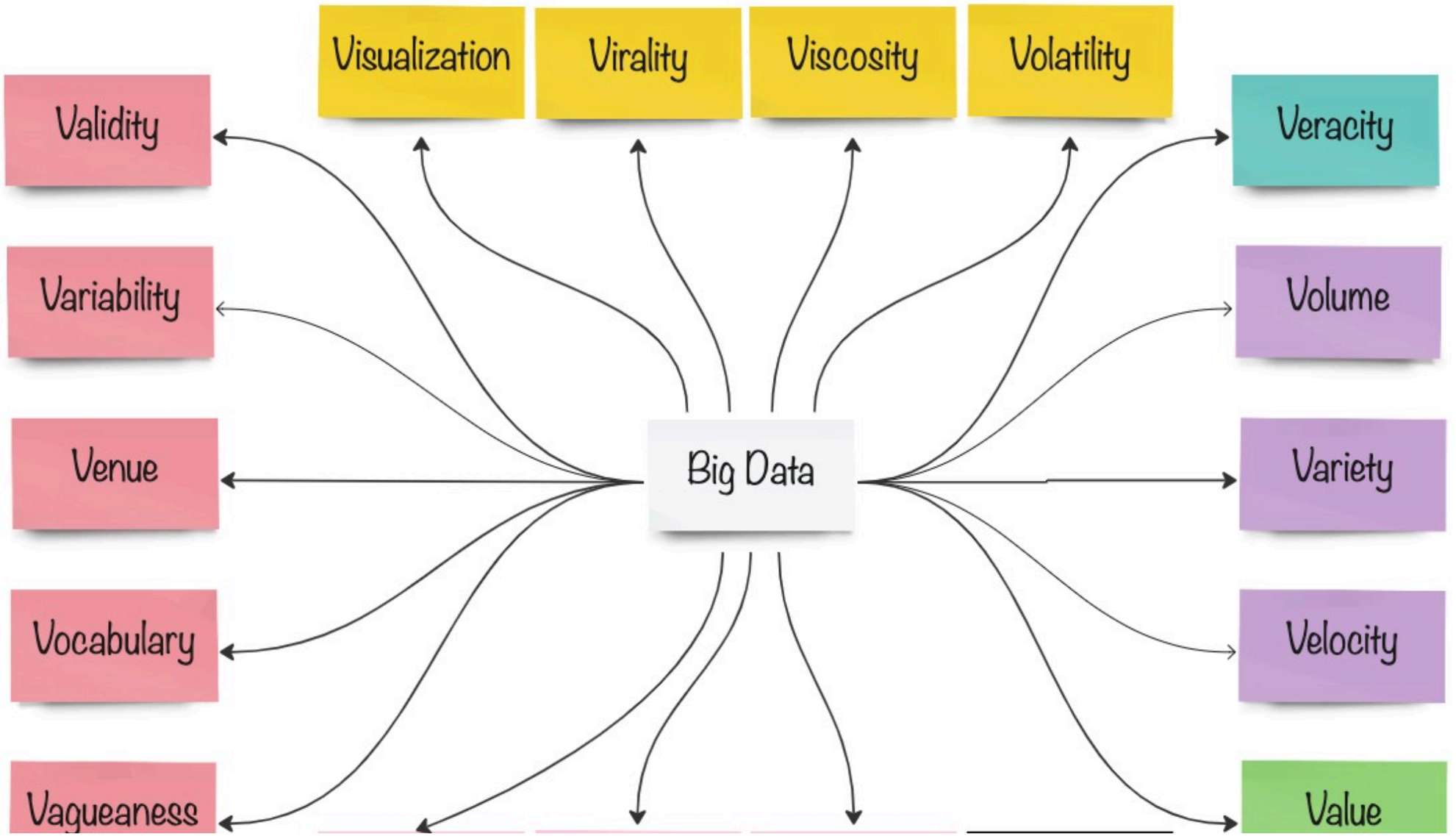


10V

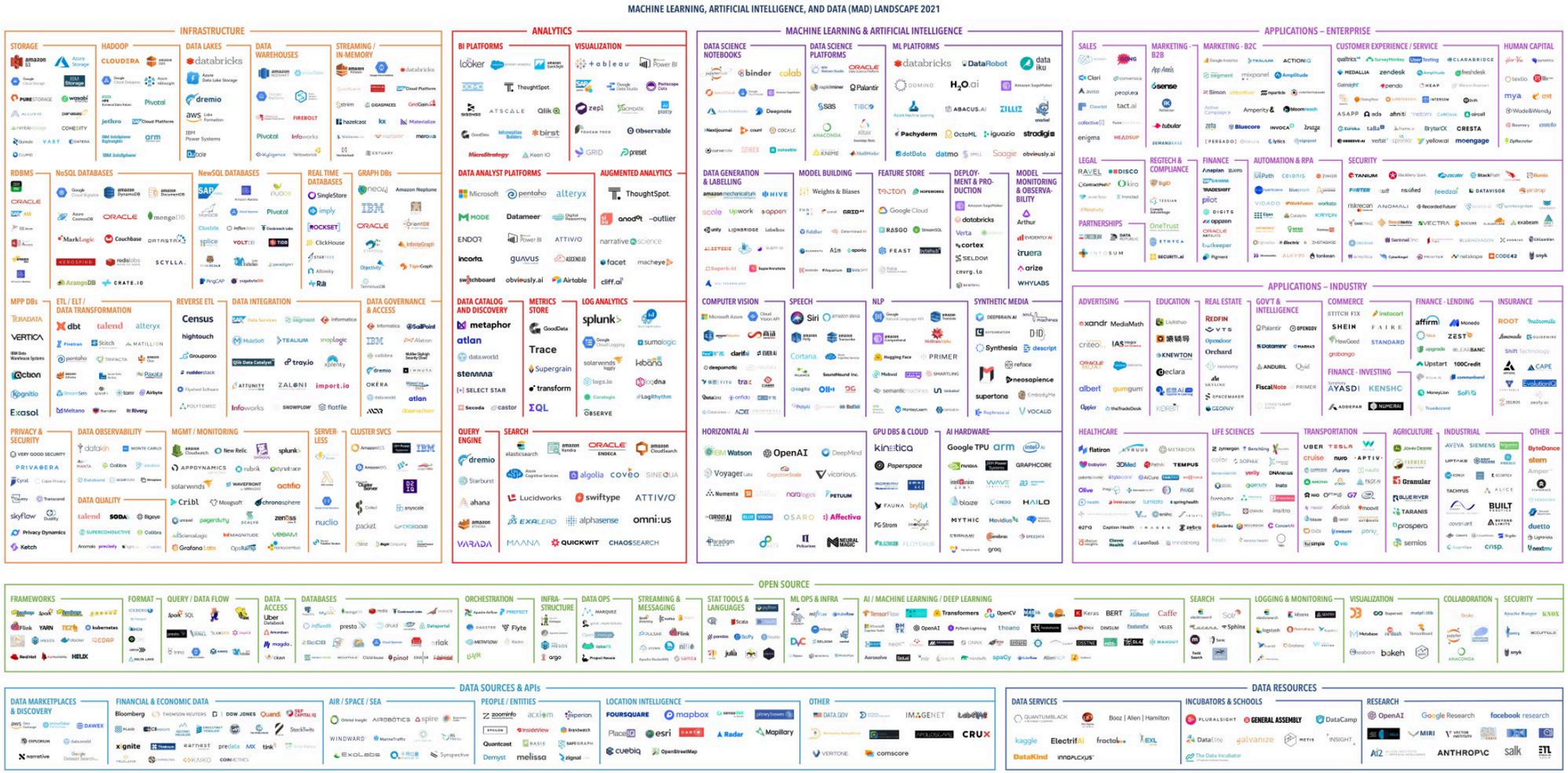
In 2014, Data Science Central, Kirk Born has defined big data in 10 V’s

17V

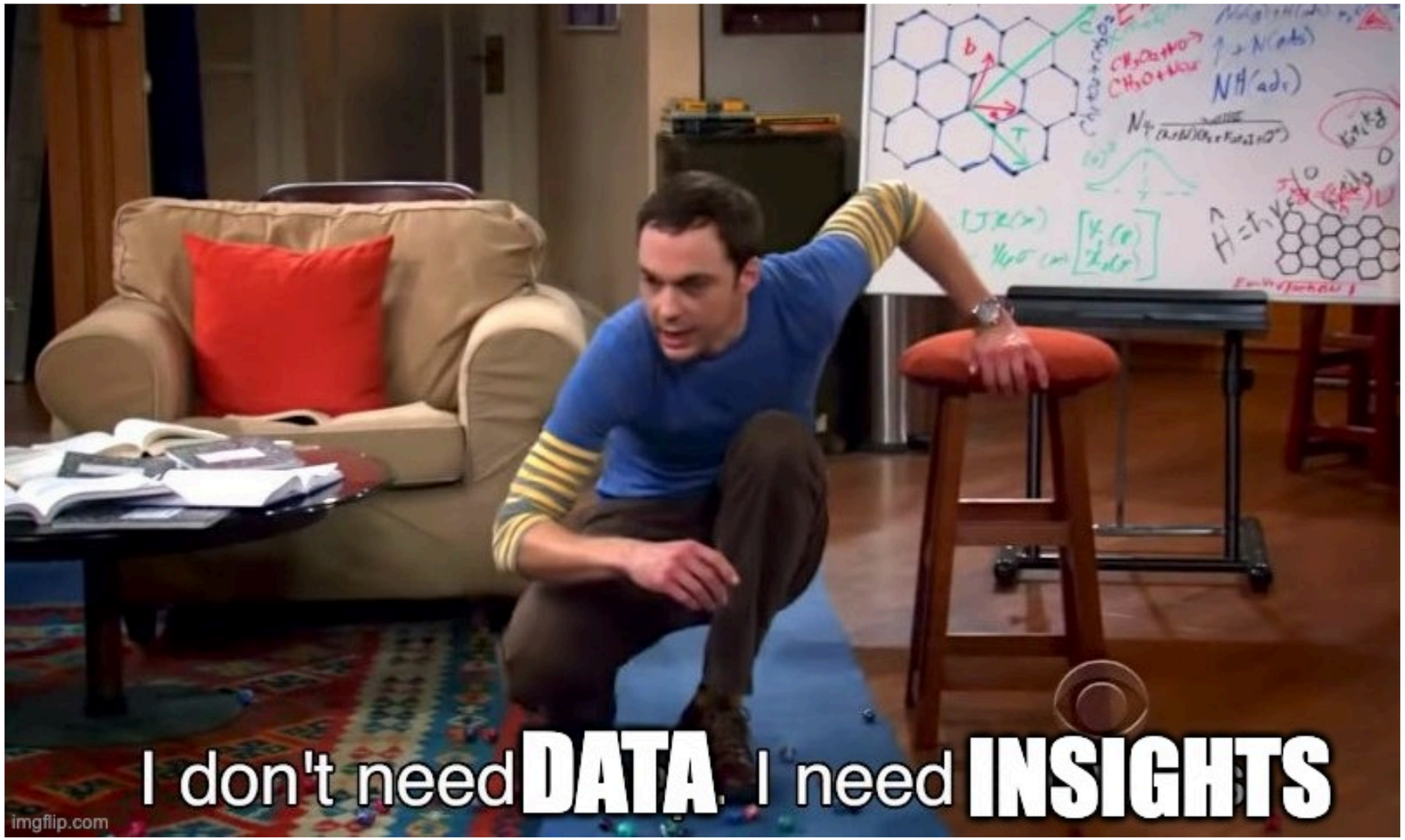
17 V AND 1 C


Visualization
Virality
Viscosity
Volatility
Validity
Veracity
Variability
Volume
Venue
Big Data
Variety
Vocabulary
Velocity
Vagueaness
Value

Big Data Landscape Big Data Landscape

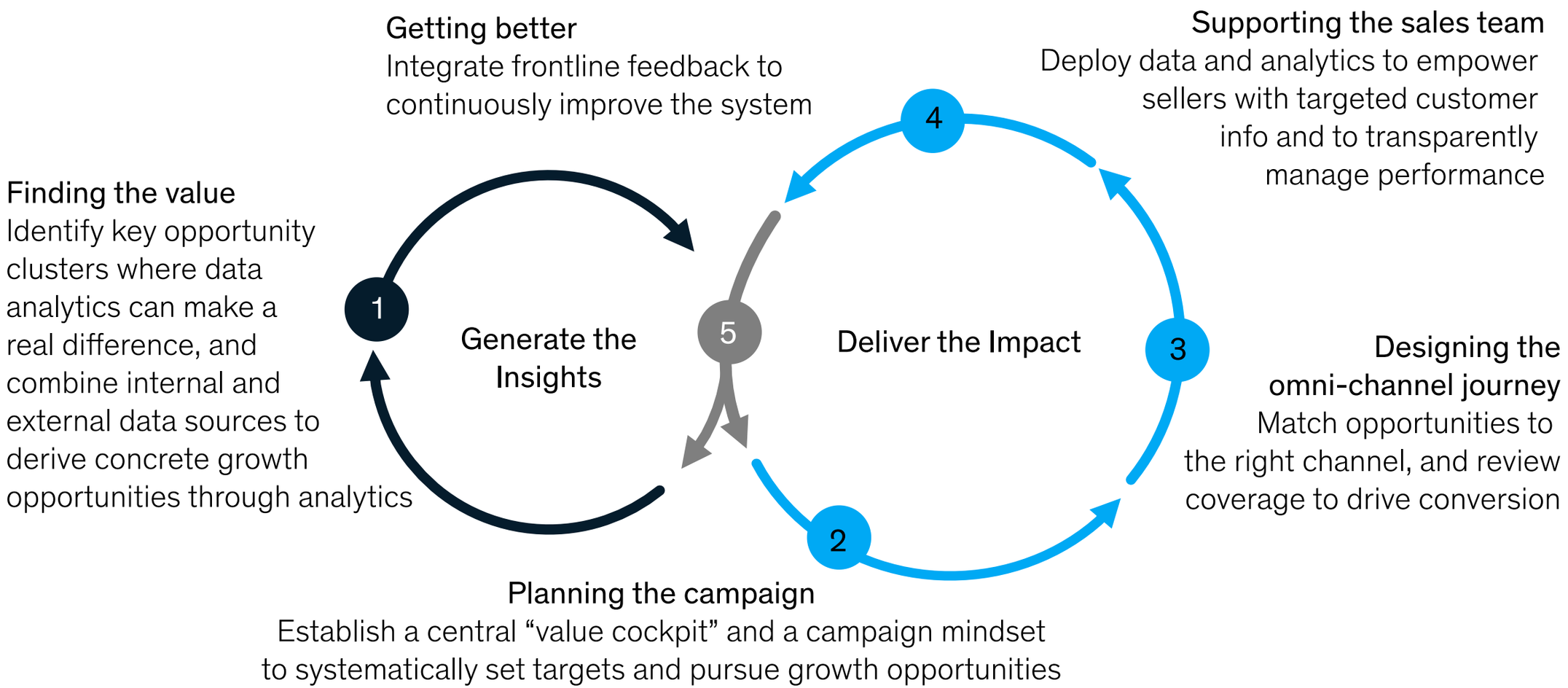


Big Data in real world…



https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/insights-to-impact-creating-and-sustaining-data-driven-commercial-growth

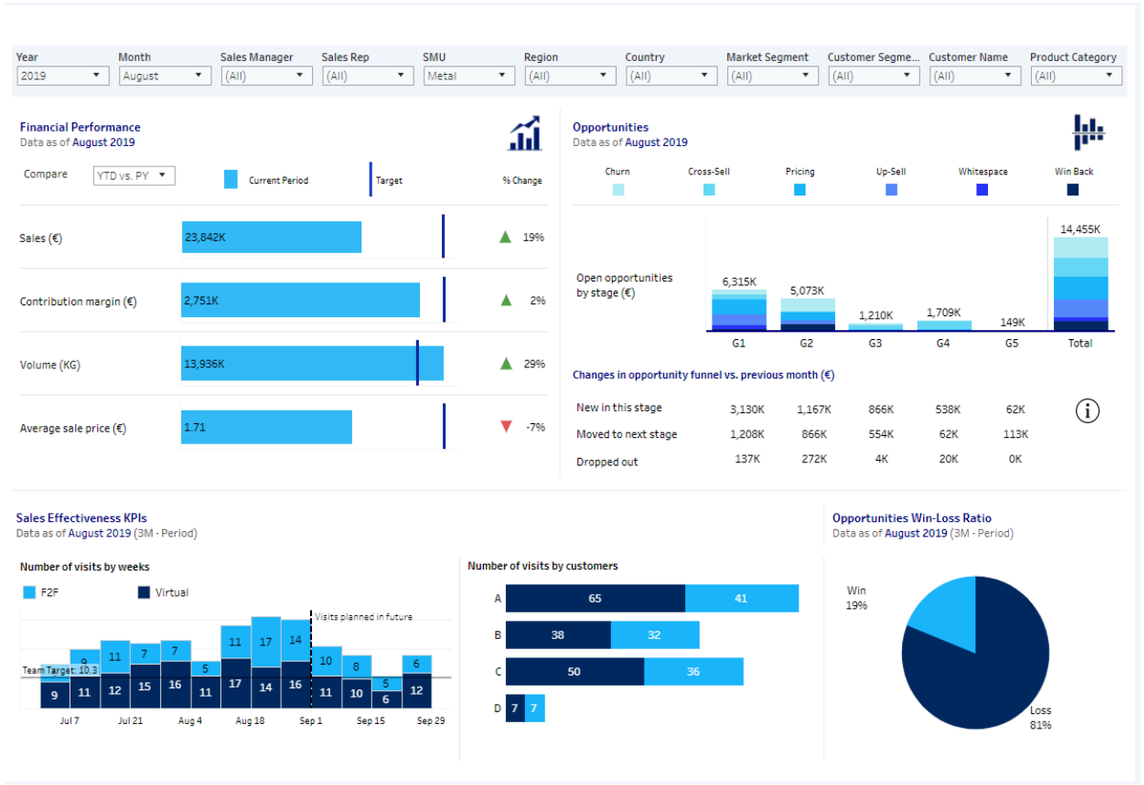
Insights to impact: Five levers enable data-driven sales growth by using insights to achieve impact.



A data-driven commercial performance dashboard serves as a single source of truth for everyone, from CEO to sales force.

1. Filters to slice and dice all key commercial indicators to understand performance from CEO to sales front line

2. **Backward-looking performance**
Measure selected performance indicators linked to budget and understand gap to target (less is more)



3. **Forward-looking actions**
Monitor progress in opportunity funnel on “customer x product” level to consistently prioritize and drive commercial actions

4. Underlying behavioral metrics (eg, number of visits, CRM usage) to understand and influence drivers

McKinsey & Company

Future ?