Funded by
the European Union

ASSISTANT

Evaldas Vaičiukynas (evaldas.vaiciukynas@ktu.lt)

https://www.assistant-erasmus.eu

Definition of Big Data

What is Data?

Attributes (Dimension; Features; Variables)

Objects (Samples, Individuals)

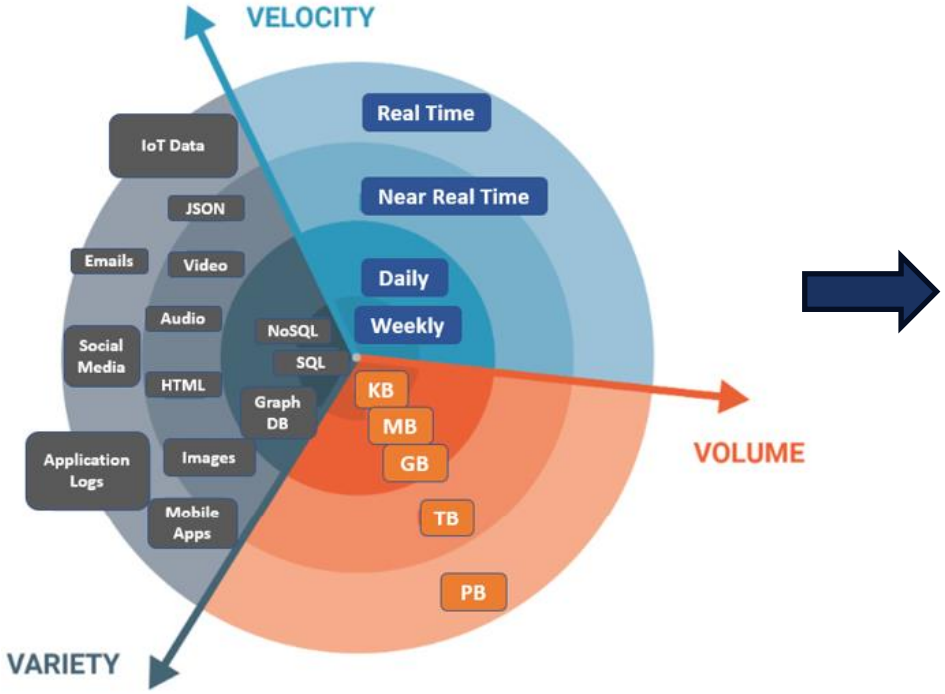
ID	Height	Weight	Age
Student 1	189	81	24
Student 2	210	90	26
Student 3	191	92	27
…	…	…	…
Student N	162	71	21

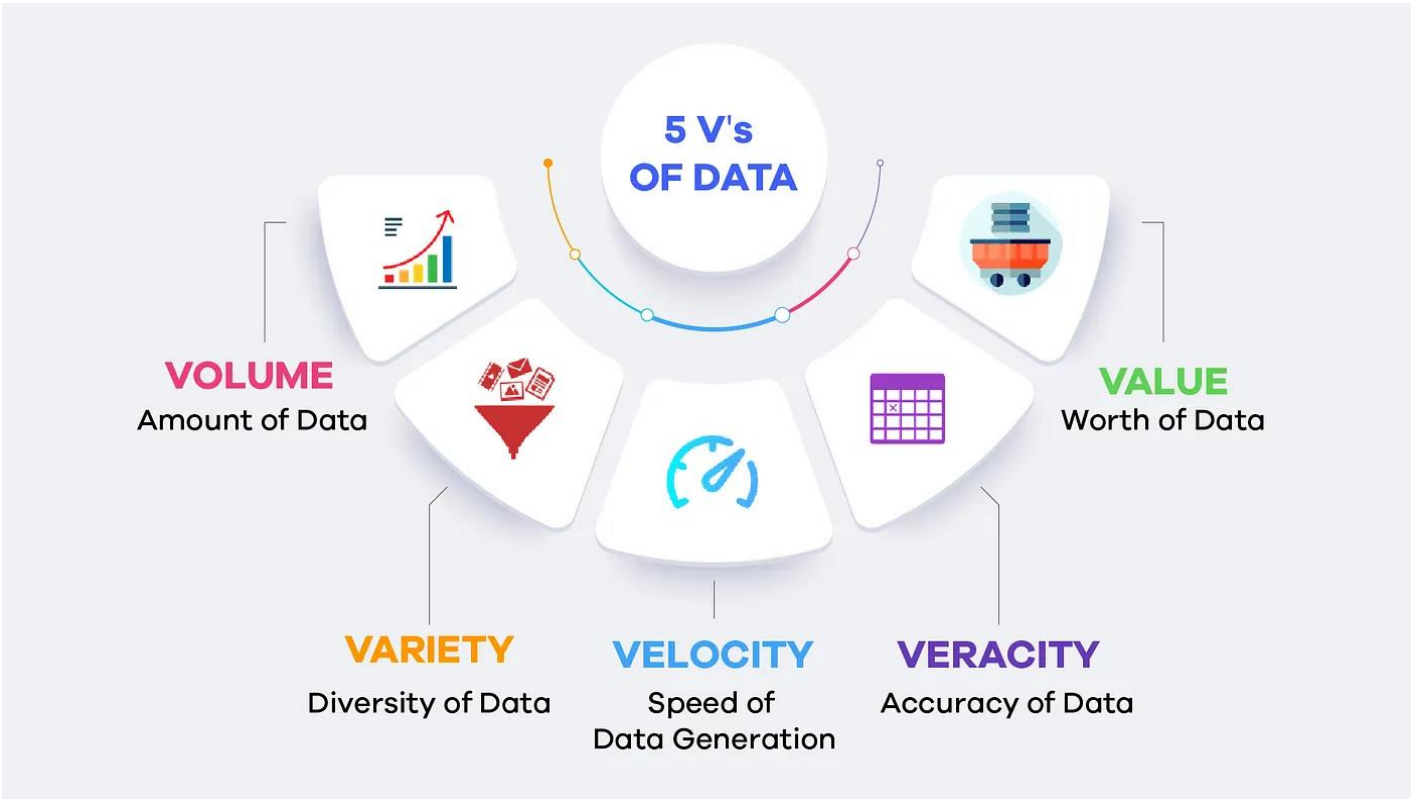
- Rows: examples, cases.
- Columns: variables, features.
- Big data = more rows (larger data sample size)
- Curse of dimensionality = more columns (problem for algorithms, when number of columns > rows).

What is Big Data?

- Big data is a great quantity of diverse information that arrives in increasing volumes and with ever-higher velocity.
- Big data can be structured (tables with numbers or text) or unstructured (different formats, images, etc).
- Big data is most often stored in computer databases and is analyzed using tools specifically designed to handle large, complex data sets.
- Fun definition: it's big, if Excel does not open it. ☺

From 3V's to 5 V's of Big Data:
Volume (size), Variety (types), Velocity (freq), Veracity (trust), Value ($?)



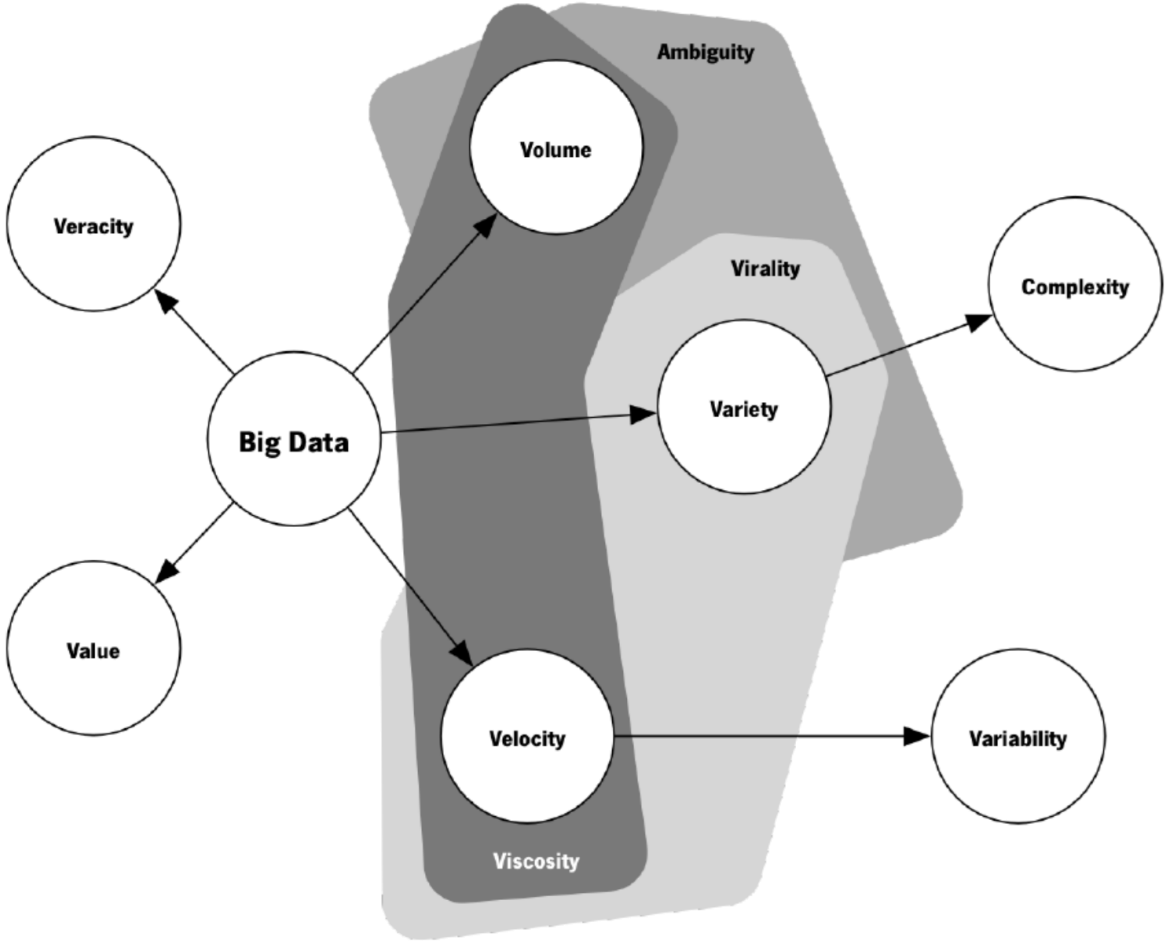


The V's of Big Data: Expansion?

9V's (Owais, 2016)	10 V's (Data Science Central)	17 V's (Panimalar, 2017)	42 V's (KDNuggets, 2017)		
Volume	Volume	Volume	Vagueness	Vault	Viral
Variety	Variety	Variety	Validity	Veer	Virtuosity
Velocity	Velocity	Velocity	Valor	Veil	Viscosity
Veracity	Veracity	Veracity	Value	Velocity	Visibility
Value	Value	Value	Vane	Venue	Visualization
Visualization	Variability	Variability	Vanilla	Veracity	Vivify
Variability	Validity	Validity	Vantage	Verdict	Vocabulary
Validity	Venue	Venue	Variability	Versed	Vogue
Volatility	Vocabulary	Vocabulary	Variety	Version Control	Voice
	Vagueness	Vagueness	Varifocal	Vet	Volatility
		Volatility	Varmint	Vexed	Volume
		Visualization	Varnish	Viability	Voodoo
		Viscosity	Vastness	Vibrant	Voyage
		Virality	Vaticination	Victual	Vulpine
		Verbosity			
		Voluntariness			
		Versality			

https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

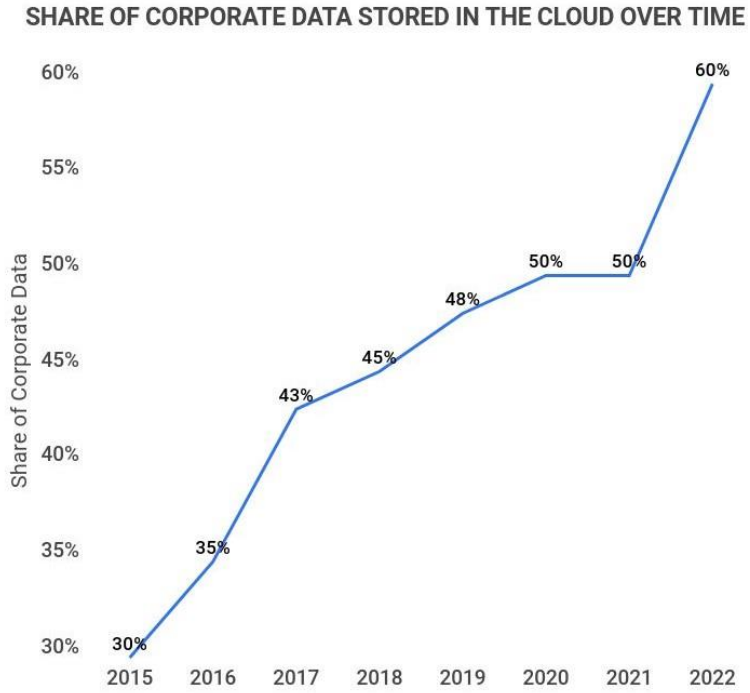
Big Data is an abstract concept



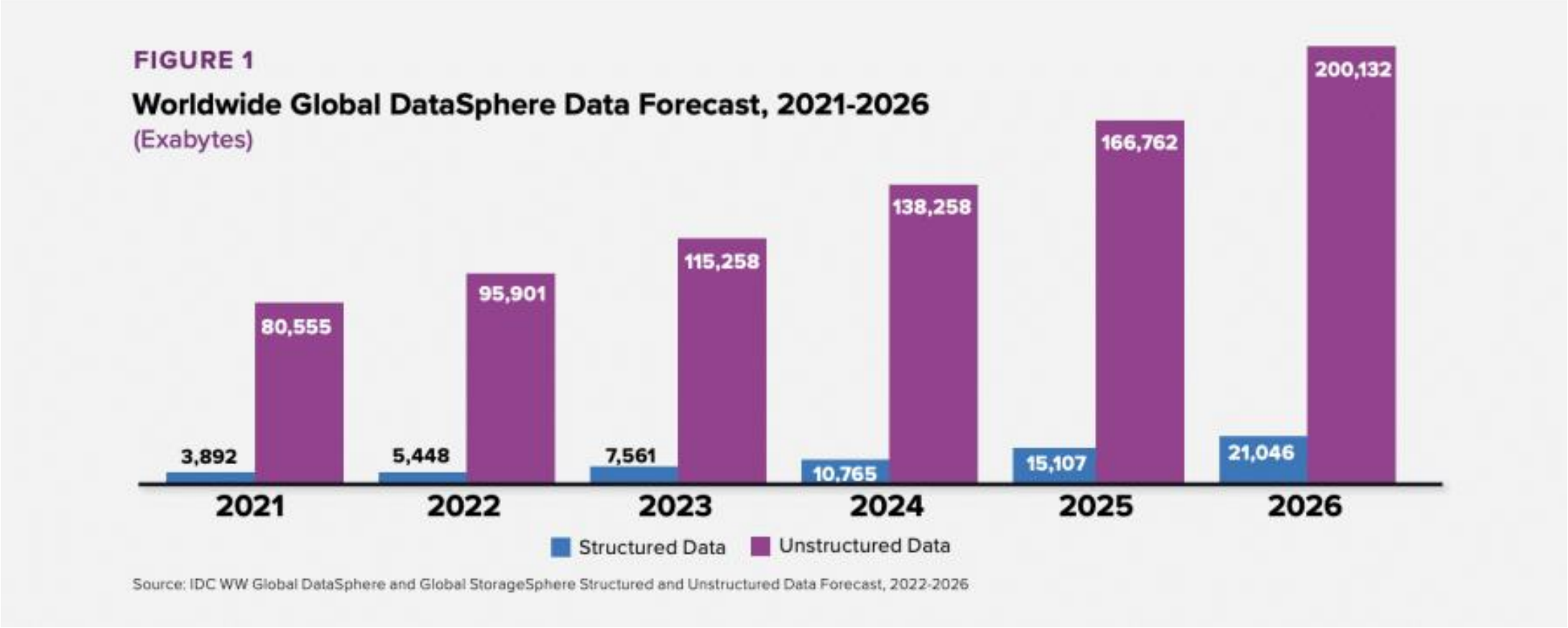
- Ambiguity = lack of metadata after combining.
- Virality = data propagation speed among peers.
- Viscosity = drag in flows of streaming data.
- Complexity = challenge of multiple sources.
- Variability = inconsistent data velocity.

Some statistics on the amounts of data (2023)

- 90% of the world's data has been created in the last two years.
- 60% of corporate data worldwide is stored in the cloud (2022).
- 1.7 MB of data is created every 1 s for every person on earth.
- Approx. 328.77 million terabytes of data are created each day.
- Video content covers over half (53.72%) of all global data traffic.



Forecast: 90% of all data generated annually will be unstructured in 2026



Data Sources & Data Format

Data Sources

Internet

Social Media/
Multimedia Data

Sensors

Web Services

Other Devices

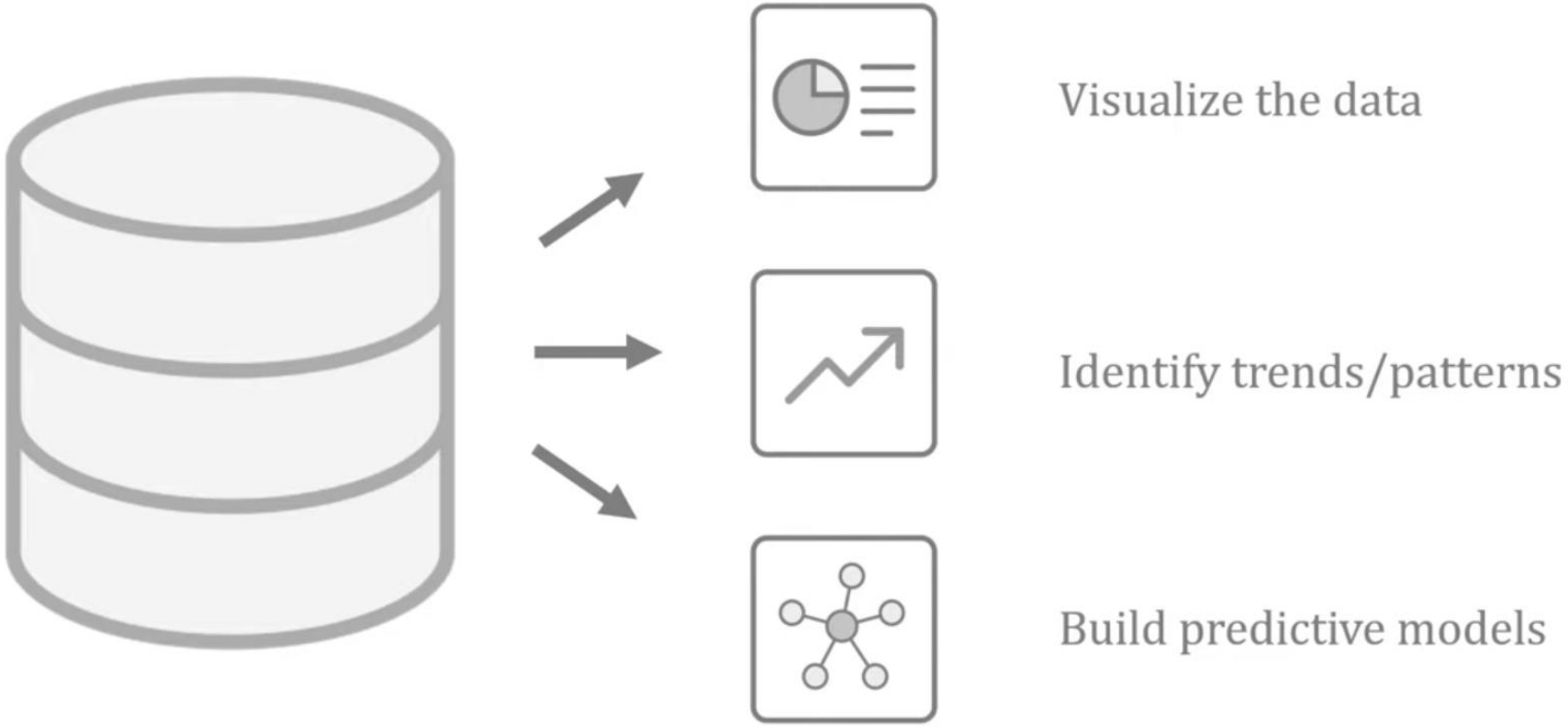
Data Format

Structured Data
- Enterprise Data
- CRM
- ERP

Semi Structured Data
- XML/JSON
- EDI
- Transaction

Unstructured Data
- Documents
- Videos
- Machine Sensors
- Social Media

Extracting the Value of Big Data



Thank you for attention

Funded by
the European Union