Ministry of Science and Education of the Republic of Azerbaijan
Institute of Information Technology

Opportunity of AzScienceNet in Big Data Analytics

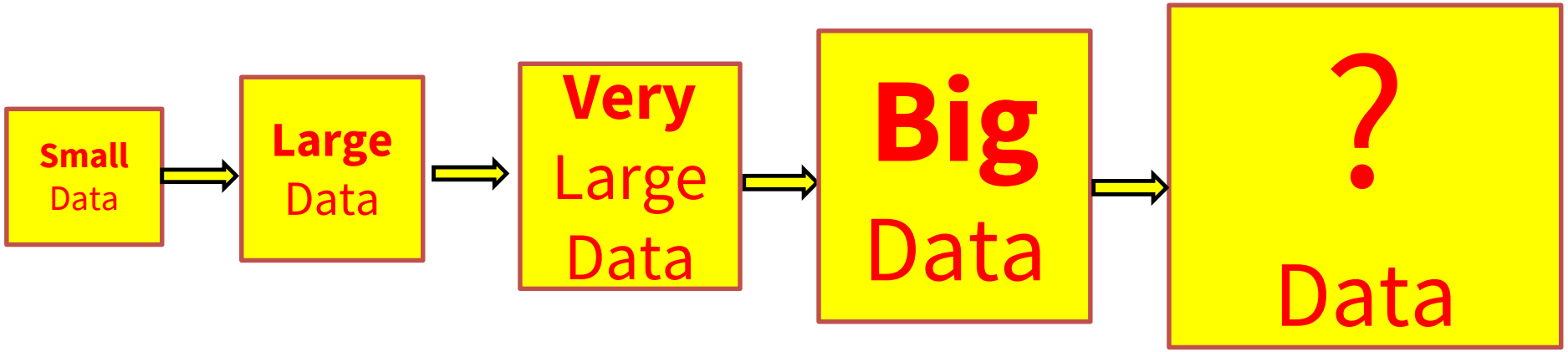
Prof.Dr. Ramiz Aliguliyev
Dr. Lyudmila Sukhostat

The 5th Eastern Partnership E-infrastructure Conference (EaPEC 2022)
Baku, 28-29 September 2022

What is Big Data?

Data, measured in Giga-, Tera-, Peta-, Exa- etc. bytes?

Is it just another IT buzzword…?



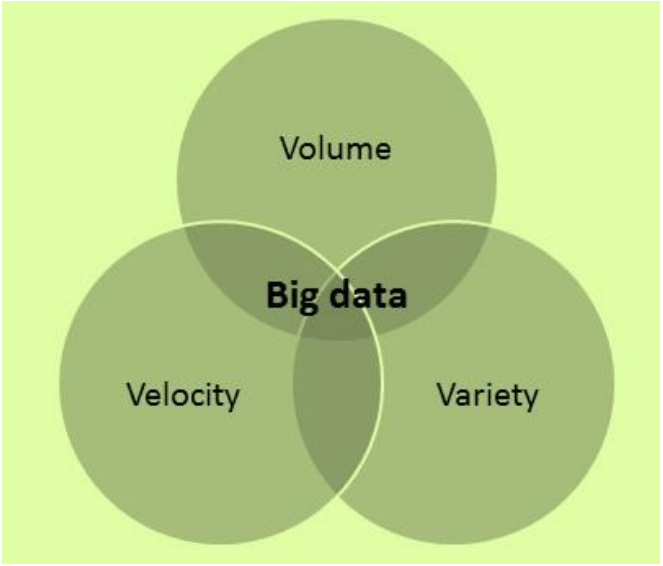
Who can answer the question: What is Big Data? Data that is small to you may be large (very large, big) to someone else or vice versa. Then the question arises: how to define Big Data? So the volume of data is relative; it is not absolute.

"The concept of **Big** is problematic to pinpoint, not least because a dataset that appears to be massive today will almost surely appear small in the near future" (MIT Technology Review, 2013).

What is Big Data?

For the definition of Big Data, there are different explanations:

The concept of Big Data dates back to 2001, when the challenges of increasing data were addressed with a **3V**s model by Doug Laney. He used volume, velocity and variety, known as **3V**s, to characterize the concept of Big Data. The term **volume** is the size of the data set, **velocity** indicates the speed of data in and out, and **variety** describes the range of data types and sources.



Oracle has added an additional dimension and defined Big Data in terms of **4 V**'s, i.e., Volume, Velocity, Variety and **Veracity** (how accurate or truthful a data set may be**).**

In 2012, Gartner gave a more detailed definition: **"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."**

Clifford A. Lynch, "**Big data: How do your data grow?**", *Nature*, vol. 455, no.7209, 2008.

What is Big Data?

"**Big Data** is a collection of very huge data sets with a great diversity of types so that it becomes difficult to process by using **state-of-the-art data processing techniques** or **traditional data processing platforms**."

(C.L. Philip Chen & Chun-Yang Zhang, «Data-intensive applications, challenges, techniques and technologies: a survey on big data», Information Sciences, 275, 314-347, 2014.)

Clifford A. Lynch, "Big data: How do your data grow?", Nature, vol. 455, no.7209, 2008.

What is Big Data?

1 zettabyte = 10^{21} byte

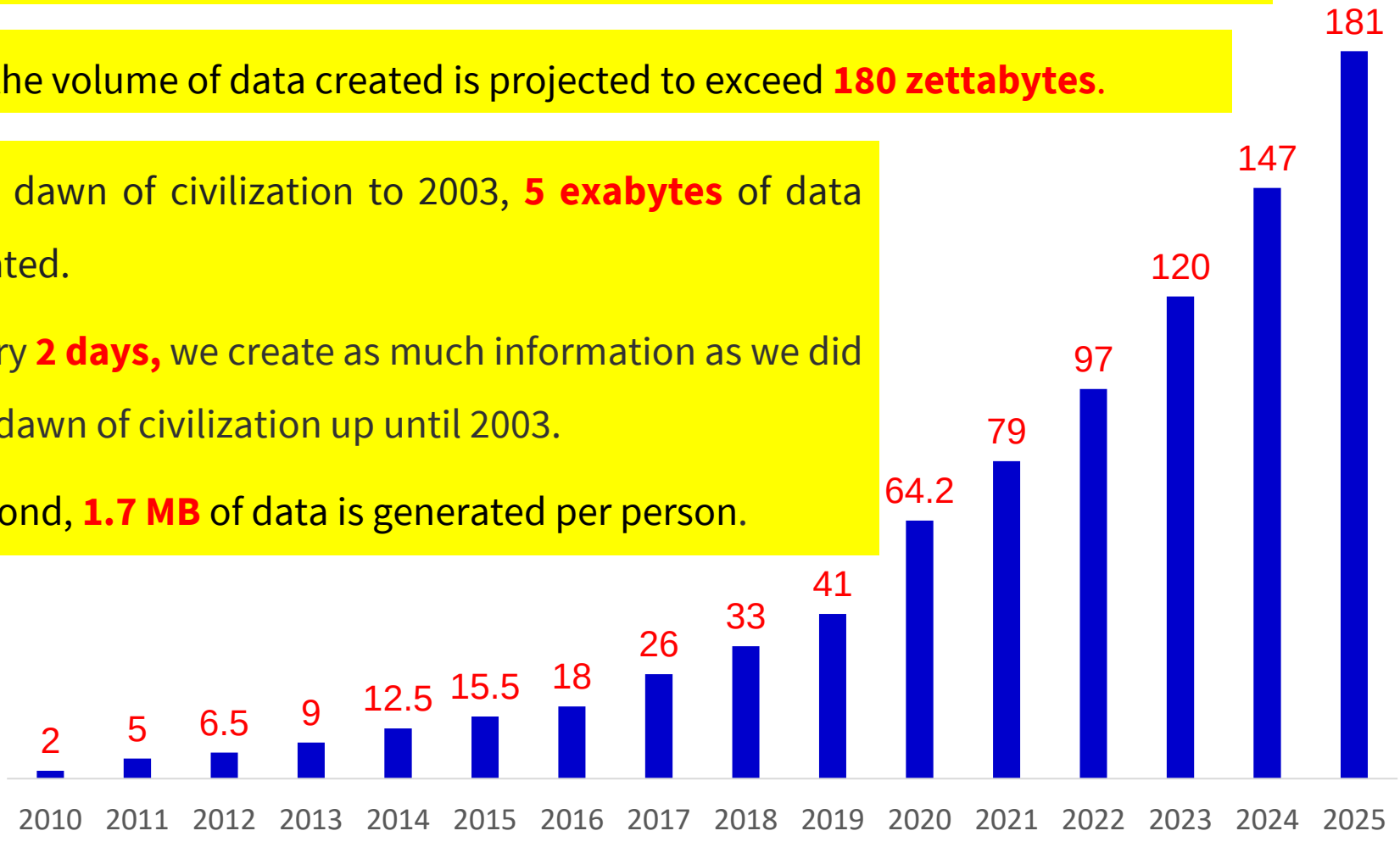
Volume of Data Created from 2010 to 2021, with forecasts from **2022** to **2025** (in zettabytes)

In 2025, the volume of data created is projected to exceed **180 zettabytes**.

From the dawn of civilization to 2003, **5 exabytes** of data were created.

Now, every **2 days,** we create as much information as we did from the dawn of civilization up until 2003.

Every second, **1.7 MB** of data is generated per person.



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

Why Big Data?

"**Big Data is the oil of new economy**" (World Economic Forum, Davos, 2011)

"Data is the new oil. Like oil, data is valuable, but if unrefined, it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity. So, data must be broken down, analyzed for it to have value."

Prof. Gary King (Harvard University) famously said that "Big Data is not about the Data." In other words, the data itself has no inherent value. It is like crude oil that needs to be refined before it can be put to good use. This refinement of "crude" data is achieved with analytics.

At the 2012 Annual Meeting in Davos, the World Economic Forum published a white paper entitled "**Big Data, Big Impact: New Possibilities for International Development**."

Big Data has been one of the current and future research frontiers. In 2012, Gartner listed the ''Top 10 Strategic Technology Trends For 2013'' and ''Top 10 Critical Tech Trends For The Next Five Years'', and **Big Data** is listed in both.

Why Big Data?

Global Big Data Initiatives

Most of the developed countries – US, Canada, G. Britain, Australia, France, Japan has a national agenda and initiatives on Big Data.

1. In 2012, **US** launched Big Data Research and Development initiative with an investment of more than US$ 200M.
2. In **2012**, Japan announced their Big Data Strategy named "The Integrated ICT strategy for 2020," which developed framework for open data.
3. In Jan 2013, the **UK** government announced a BIG DATA plan with initial funding of 189M pounds.
4. In Feb 2013, **French** government published the "Digital Roadmap," which invested 11.5M euros in the development of seven future projects, including BIG DATA.
5. In Aug 2013, **Australia** announced Australian National Big Data Strategy.

Why Big Data?

		Annual Revenue (2021)	Number of employees (2022)	Launch Date
1	**Uber** (the world's largest taxi company, owns no vehicles)	**$ 17.4 B**	**29,300**	**2010**
2	**Facebook** (the world's most popular media owner, creates no content)	**$ 117.92 B**	**60,600**	**2004**
3	**Alibaba** (the most valuable retailer, has no inventory)	**$109.48 B**	**254,941**	**1999**
4	**Airbnb** (the world's largest accommodation provider, owns no real estate)	**$ 5.9 B**	**5,597**	**2008**

Klaus Schwab, "4th Industrial Revolution," 2016.

What is Big Data Analytics?

Definition. Big data analytics is the process of collecting, examining, and analyzing large amounts of data to discover market trends, insights, and patterns that can help companies make better business decisions.

Why is big data analytics important?
Big data analytics is important because it helps companies leverage their data to identify opportunities for improvement and optimization.

Big data in the real world

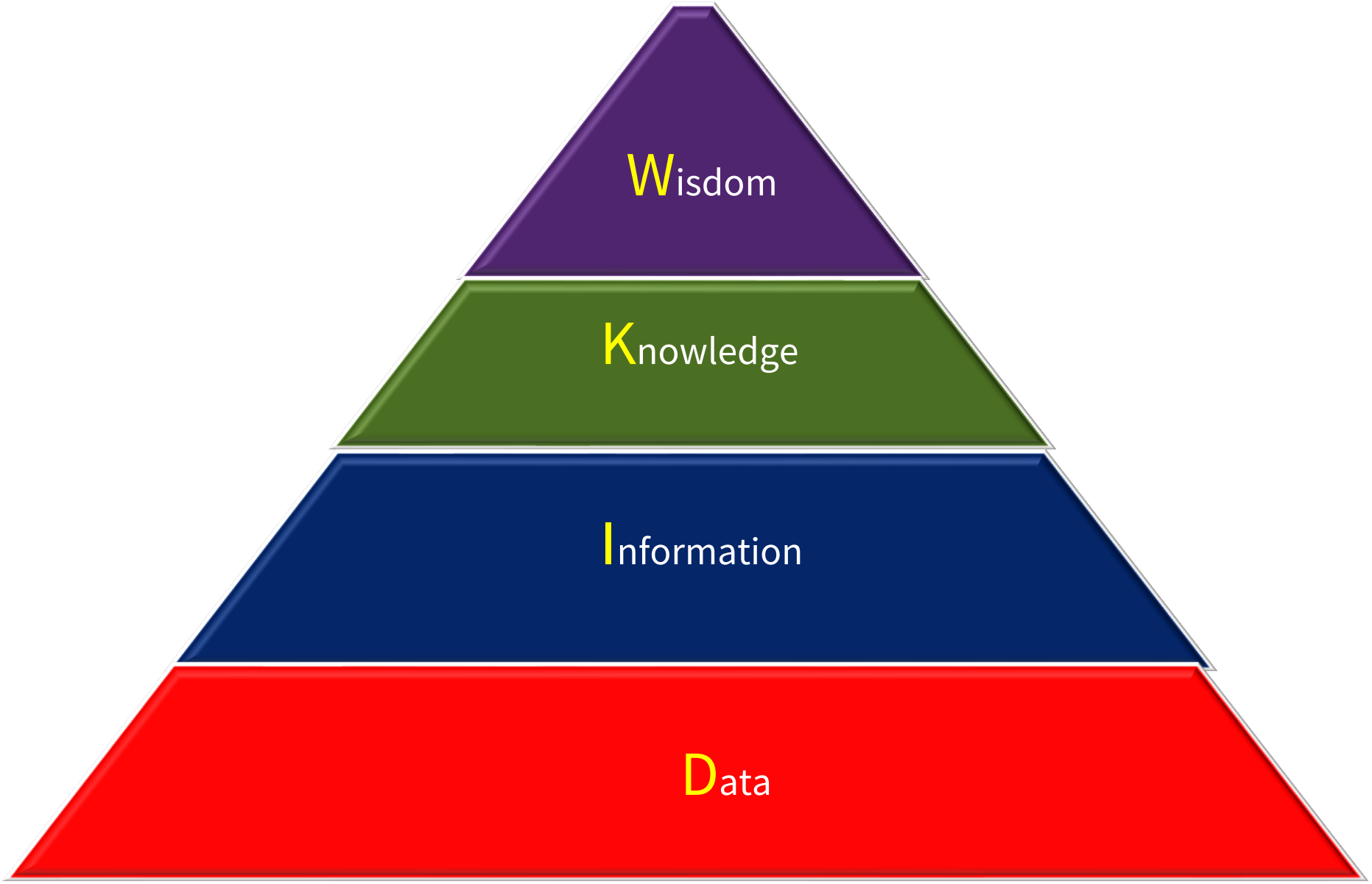
Big data analytics helps companies and governments make sense of data and make better, informed decisions.

Use big data to stay competitive

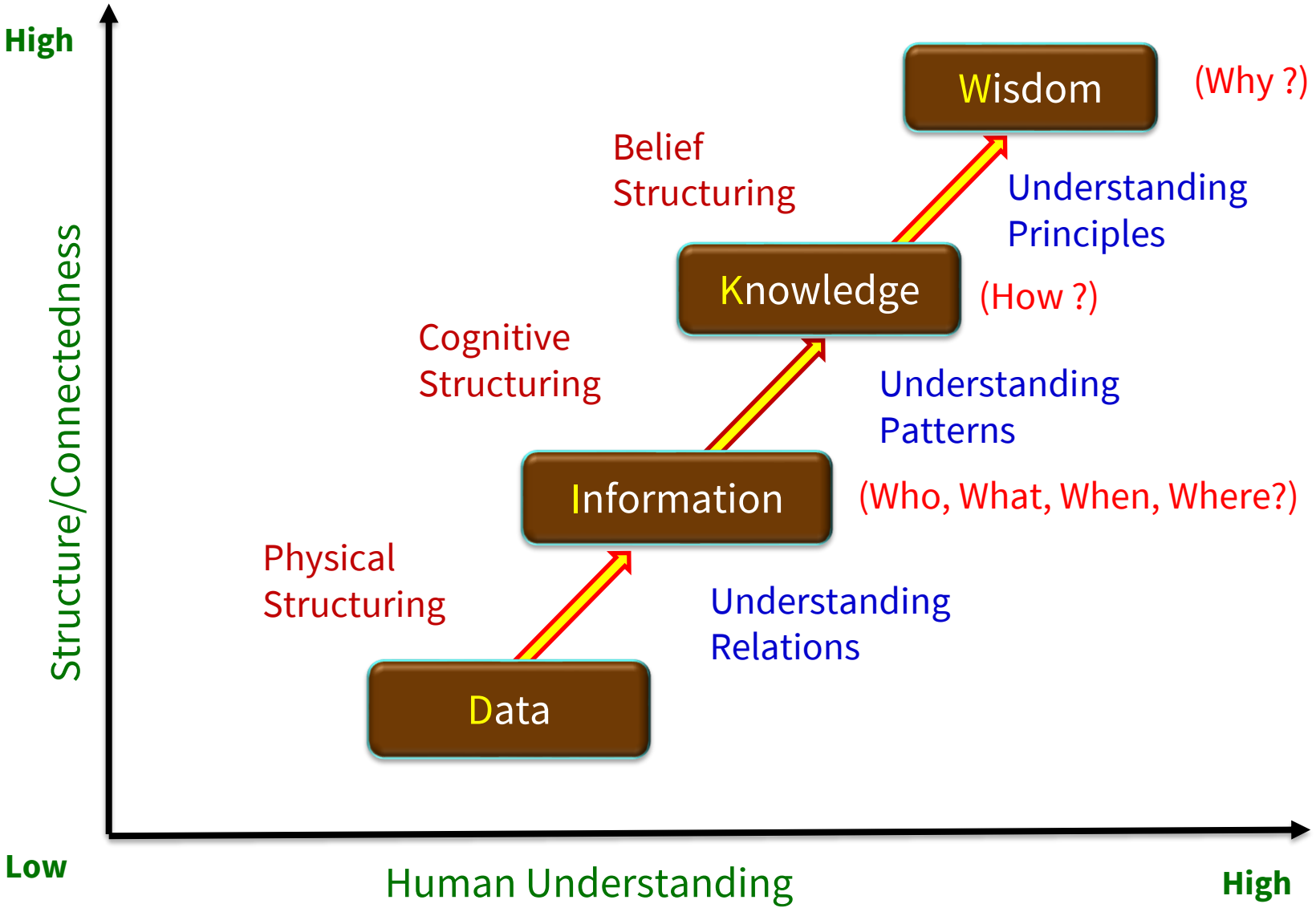
Almost eight in ten users (**79%**) believe that "companies that do not embrace big data will lose their competitive position and may even face extinction," according to an Accenture report. In their survey of Fortune **500** companies, Accenture found that **95%** of companies with revenues over **$10 billion** reported being "highly satisfied" or "satisfied" with their big data-driven business outcomes.

https://www.accenture.com/us-en/_acnmedia/accenture/conversion-assets/dotcom/documents/global/pdf/industries_14/accenture-big-data-pov.pdf

The DIKW Pyramid



The DIKW Pyramid



Opportunities of Big Data

There are quite a few advantages to incorporating big data analytics into a business or organization. These include:

- **Cost reduction**: Big data can reduce costs in storing all the business data in one place. Tracking analytics also helps companies find ways to work more efficiently to cut costs wherever possible.
- **Product development**: Developing and marketing new products, services, or brands is much easier when based on data collected from customers' needs and wants. Big data analytics also helps businesses understand product viability and keep up with trends.
- **Strategic business decisions**: The ability to constantly analyze data helps businesses make better and faster decisions, such as cost and supply chain optimization.
- **Customer experience**: Data-driven algorithms help marketing efforts (targeted ads, for example) and increase customer satisfaction by delivering an enhanced customer experience.
- **Risk management:** Businesses can identify risks by analyzing data patterns and developing solutions for managing those risks.

Opportunities of Big Data

- **Entertainment:** Providing a personalized recommendation of movies and music according to a customer's individual preferences has been transformative for the entertainment industry (for example, Spotify and Netflix).
- **Education:** Big data helps schools and educational technology companies develop new curriculums while improving existing plans based on needs and demands.
- **Health care:** Monitoring patients' medical histories helps doctors detect and prevent diseases.
- **Government:** Big data can be used to collect data from closed-circuit television and traffic cameras, satellites, body cameras and sensors, emails, calls, and more, to help manage the public sector.
- **Marketing:** Customer information and preferences can be used to create targeted advertising campaigns with a high return on investment.
- **Banking:** Data analytics can help track and monitor illegal money laundering.

Types of Big Data Analytics

4 types of Big Data Analytics support and inform different business decisions.

1. Descriptive analytics (helps in understanding **"What has happened?"**)

Descriptive analytics refers to data that can be easily read and interpreted. This data helps create reports and visualize information that can detail company profits and sales.

***Example*:** During the pandemic, a leading pharmaceuticals company conducted data analysis on its offices and research labs. Descriptive analytics helped them identify unutilized spaces and departments that were consolidated, saving the company millions of dollars.

2. Diagnostics analytics (helps in understanding **"Why happened?"**)

Big data technologies and tools allow users to mine and recover data that helps dissect an issue and prevent it from happening in the future.

***Example*:** A clothing company's sales have decreased even though customers continue to add items to their shopping carts. Diagnostics analytics helped to understand that the payment page was not working properly for a few weeks.

https://www.coursera.org/articles/big-data-analytics

Types of Big Data Analytics

3. Predictive analytics (helps in anticipating **"What could happen?**)

Predictive analytics looks at past and present data to make predictions. With artificial intelligence (AI), machine learning (ML), and data mining, users can analyze the data to predict market trends.

***Example*:** In the manufacturing sector, companies can use algorithms based on historical data to predict if or when a piece of equipment will malfunction or break down.

4. Prescriptive analytics (helps in responding **"What should we do?"**)

Prescriptive analytics provides a solution to a problem, relying on AI and ML to gather data and use it for risk management.

***Example*:** Within the energy sector, utility companies, gas producers, and pipeline owners identify factors that affect the price of oil and gas to hedge risks.

Tools used in Big Data Analytics

Harnessing all of that data requires tools. Thankfully, technology has advanced so that many intuitive software systems are available for data analysts to use.

- **Hadoop:** An open-source framework that stores and processes big data sets. Hadoop is able to handle and analyze structured and unstructured data.
- **Spark:** An open-source cluster computing framework for real-time processing and analyzing data.
- **Data integration software:** Programs that allow big data to be streamlined across different platforms, such as MongoDB, Apache, Hadoop, and Amazon EMR (Elastic MapReduce).
- **Stream analytics tools:** Systems that filter, aggregate, and analyze data that might be stored in different platforms and formats, such as Kafka.
- **Distributed storage:** Databases that can split data across multiple servers and can identify lost or corrupt data, such as Cassandra.
- **Predictive analytics hardware and software:** Systems that process large amounts of complex data, using machine learning to predict future outcomes, such as fraud detection, marketing, and risk assessments.
- **Data mining tools:** Programs that allow users to search within structured and unstructured big data.
- **NoSQL databases:** Non-relational data management systems ideal for dealing with raw and unstructured data.
- **Data warehouses:** Storage for large amounts of data collected from many different sources, typically using predefined schemas.

Big Data Challenges

Challenges of Big Data can be grouped into **3** main categories, based on the data life cycle:

DATA Challenges are related to the characteristics of the data itself (Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value)

PROCESS Challenges are related to a series of how techniques: how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to provide the results. (Data Acquisition & Warehousing, Data Mining and Cleansing, Data Aggregation & Integration, Analysis & Modeling, Data Interpretation).

MANAGEMENT Challenges cover, for example, the privacy, security, governance, ethical aspects and lack of skills in understanding and analyzing data (Privacy, Security, Data Governance, Data & Information Sharing, Cost/Operational Expenditures, Data Ownership)

Principles for Designing Big Data Systems

Big Data analytics in a highly distributed system cannot be achievable without the following principles:

Principle 1. Good architectures and frameworks.

Big Data cannot be solved effectively and approvingly if there are no good and proper architecture for the whole Big Data systems.

Principle 2. **Support a variety of analytical methods**.

Big Data applications often produce complex tasks that make it impossible to be resolved by using one or a few of disciplines and analytical methods.

Principle 3. No size fits all.

When it comes to Big Data analytics, there is no one size which can fit all solutions.

Principle 4. **Bring the analysis to data**.

Big Data set is extremely large; it is inadvisable and infeasible to collect and move data to only one or several centers for analysis. Data-driven analysis needs to bring the analysis tasks to data sites.

Principle 5. **Processing must be distributable for in-memory computation**.

Principle 6. **Data storage must be distributable for in-memory storage**.

Principle 7. **Coordination is needed between processing and data units**.

This principle guarantees the low latency of response which is particularly required in real-time analytics.

5 Trends in Computer Science Research in 2022

1. Artificial Intelligence and Robotics

2. Big Data Analytics

Back in 2012, the Harvard Business Review branded data science the most demanded job of the 21st century. There has been a surge in demand for experts in this field and doubled efforts on the part of brands and agencies to boost salaries and attract data science talents. **Big data analytics** is everywhere, from banking to healthcare, as companies increasingly attempt to use the enormous datasets they have to personalize and improve their services.

3. Computer-assisted education

4. Bioinformatics

5. Cyber security

QS: https://www.topuniversities.com/courses/computer-science-information-systems/5-trends-computer-science-research

Latest 10 Trends in Big Data Analytics

1. DaaS - Data as a Service

As with SaaS applications, DaaS uses cloud technology to give users and applications with on-demand access to information without depending on where the users or applications may be.

2. Responsible and Smarter AI

Responsible and Scalable AI will enable better learning algorithms with shorter time to market.

3. Predictive analytics

Predictive analysis in big data can predict what may occur in the future. This strategy is extremely efficient in correcting analyzed assembled data to predict customer response.

4. Quantum computing

The *quantum computer Sycamore (it has 53 qubits)* completed the complex computation in just 200 seconds, which would have taken approximately 10,000 years to finish on other powerful supercomputers. It is almost 1.5 trillion times faster!

All our hope is in quantum computing.

5. Edge computing

Edge computing brings computation to a network's edge and reduces the amount of long-distance connection that has to happen between a customer and a server, which is making it the latest trend in big data analytics.

https://www.xenonstack.com/blog/latest-trends-in-big-data-analytics

Latest 10 Trends in Big Data Analytics

6. Natural Language Processing

NLP lies inside AI and works to develop communication between computers and humans.

7. Hybrid Clouds

The hybrid cloud provides excellent flexibility and more data deployment options by moving the processes between private and public clouds.

8. Dark Data

Gartner defines dark data as the information assets organizations collect, process and store during regular business activities but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing). Similar to dark matter in physics, dark data often comprises most organizations' universe of information assets.

9. Data Fabric

Data fabric is an architecture and collection of data networks. Data fabric can use analytics to learn and actively recommend where data should be used and changed. It can reduce data management efforts by up to 70%.

10. XOps - DataOps + ModelOps + DevOps

The aim of XOps (data, ML, model, platform) is to achieve efficiencies and economies of scale.

Top 10 Programming Languages in Big Data Analytics

1. Python
2. Java
3. JavaScript
4. R
5. SQL
6. MATLAB
7. C/C++
8. Scala
9. Julia
10. SAS

https://www.analyticsinsight.net/10-best-data-science-programming-languages-for-data-aspirants-in-2021/

Our Contribution to Big Data Analytics

With the support of **GÉANT** and the **EaPConnect** project, the reconstruction of **AzScienceNet** played a significant role in making our research more effective

Algorithm 1: Weighted Consensus Clustering for Big Data

Algorithm 2: Parallel batch k-means for Big Data clustering

Algorithm 3: Weighted Clustering for Anomaly Detection in Big Data

Algorithm 4: Anomaly Detection in Big Data based on Clustering

Algorithm 5: Anomaly Detection based on Optimization

Algorithm 1: Weighted Consensus Clustering

Suppose that we have *r* basic partitions. The task is to find a weighted consensus π with a set of $\{w_1, w_2, \ldots, w_r\}$ weights assigned to each method.

To calculate the purity based utility function, we can use the co-association matrix.

π_i

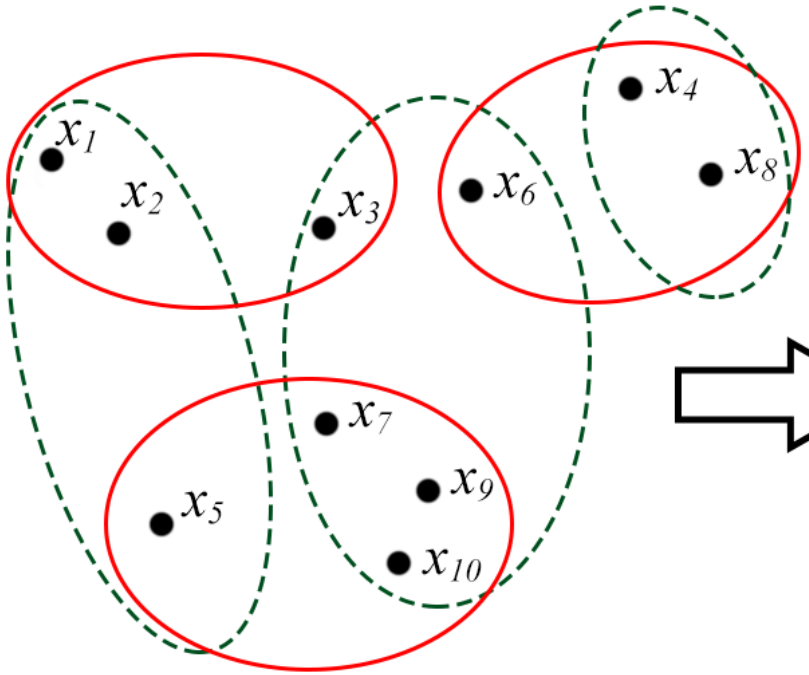
π	$C_1^{(i)}$	$C_2^{(i)}$	$\cdots$	$C_{K_i}^{(i)}$	$\sum$
C_1	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$\cdots$	$n_{1K_i}^{(i)}$	n_{1+}
C_2	$n_{21}^{(i)}$	$n_{22}^{(i)}$	$\cdots$	$n_{2K_i}^{(i)}$	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_K	$n_{K1}^{(i)}$	$n_{K2}^{(i)}$	$\cdots$	$n_{KK_i}^{(i)}$	n_{K+}
$\sum$	$n_{+1}^{(i)}$	$n_{+2}^{(i)}$	$\cdots$	$n_{+K_i}^{(i)}$	n

According to $m_{jk}^{(i)} = \left(\max n_{j1}^{(i)}, \max n_{j2}^{(i)}, \ldots, \max n_{jK_i}^{(i)}\right)$ and $m_k^{(i)} = n_{+k}^{(i)}$ the utility function can measure the similarity between the two partitions.

The purity based utility function has the following form

$$U(\pi, \pi_i) = \sum_{k=1}^{K_i} \frac{n_{+k}^{(i)}}{n} \max_j n_{jk}^{(i)} = \sum_{k=1}^{K_i} \frac{m_k}{n} m_{jk}^{(i)}$$

Algorithm 1: Weighted Consensus Clustering



	$C_1^{\pi_1}$	$C_2^{\pi_1}$	$C_3^{\pi_1}$	Σ
C_1^{π}	2	1	0	3
C_2^{π}	1	3	0	4
C_3^{π}	0	1	2	3
Σ	3	5	2	10

Illustration of a co-association matrix for three clusters

$n_{kj}^{(i)}$ is the number of points contained by both cluster $C_j^{(i)}$ and cluster C_k,

$n_{k+} = \sum_{j=1}^{K_i} n_{kj}^{(i)}$ is the number of points in C_k, $n_{+j}^{(i)} = \sum_{k=1}^{K} n_{kj}^{(i)}$ is the number of points in $C_j^{(i)}$

Algorithm 1: Weighted Consensus Clustering

The resultant rank of the clustering methods

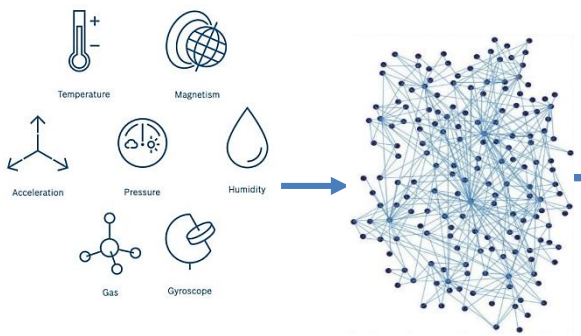
Method	The number of times the method is in the s^{th} rank s=							Resultant rank
	1	**2**	**3**	**4**	**5**	**6**	**7**	
PWCC+SqEucl	24	2	4	0	0	0	0	**28.5714**
PWCC+Cosine	15	4	7	1	2	1	0	25.1429
CLARANS	14	4	7	1	4	0	0	24.7143
PWCC+Eucl	8	8	9	1	1	3	0	23.1429
PWCC+Mink3	8	8	9	1	1	3	0	23.1429
PWCC+Mink4	8	8	9	1	1	3	0	23.1429
PWCC+Cheb	8	8	9	1	1	3	0	23.1429
OPTICS	8	3	7	7	3	2	0	21.4286
DBSCAN	2	9	7	7	3	1	1	20.4286
k-means	4	6	8	4	4	4	0	20.0000
SNNC	6	5	4	6	5	4	0	19.8571

$$rank(method) = \sum_{i=1}^{M} \frac{(M - r + 1) \cdot r_s}{M}$$

where ***M*** is he total number of methods, r_s is the number of times the method appears in the s^{th} rank. Here **s** = 7.

Algorithm 2: Parallel Batch k-means for Big Data Clustering

Flow diagram of the proposed algorithm



Sensors

Big Data

Batch 1

Batch 2

Batch *L*

K-means clustering

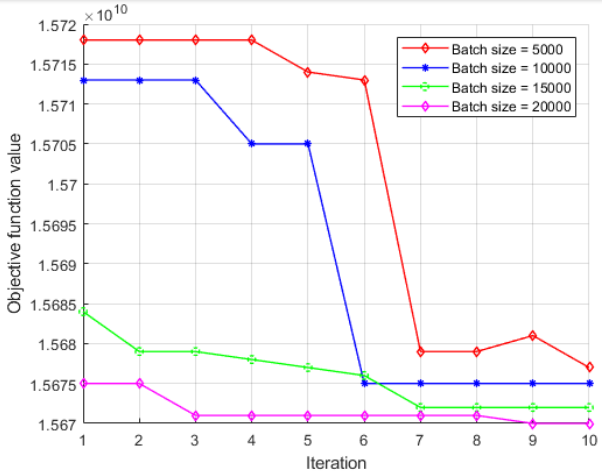
K-means clustering

K-means clustering

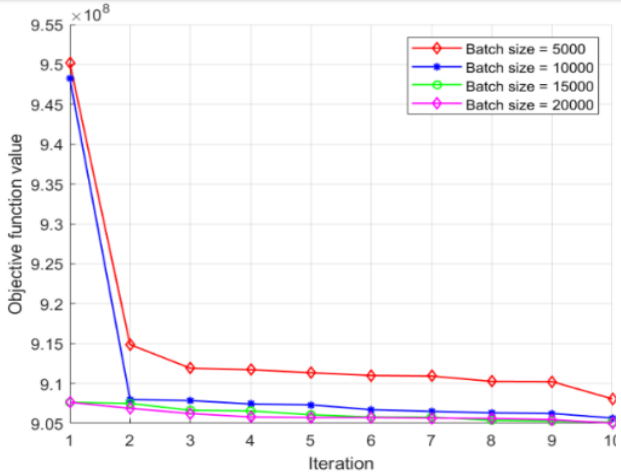
Set of centroids

K-means clustering

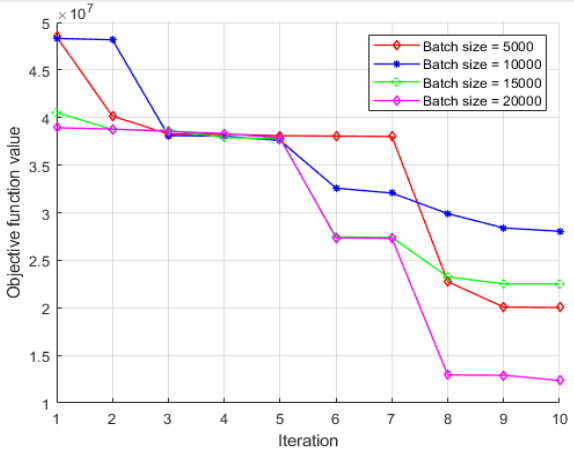
Algorithm 2: Parallel Batch k-means for Big Data Clustering



(a)

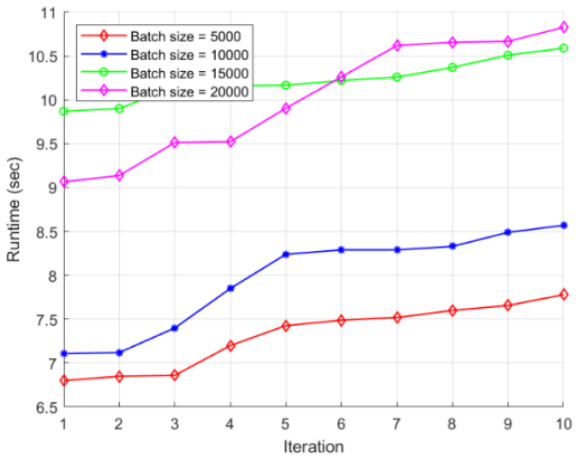


(b)

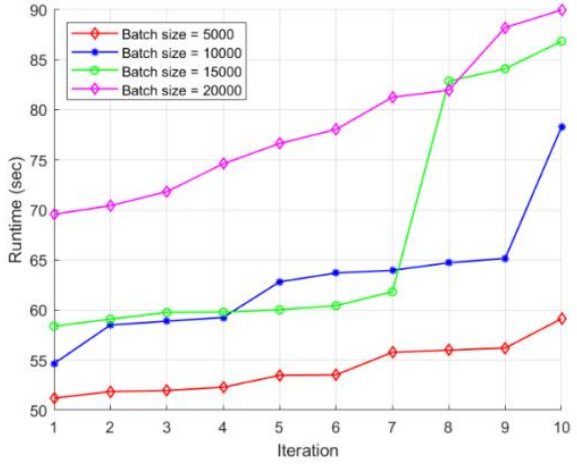


(c)

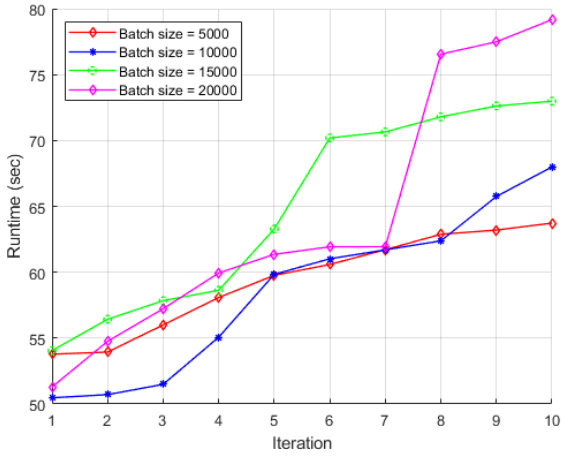
Evaluation of the objective function value with different number of iterations on Phone Accelerometer (a), YearPredictionMSD (b), US Census (1990) dataset (c)



(a)



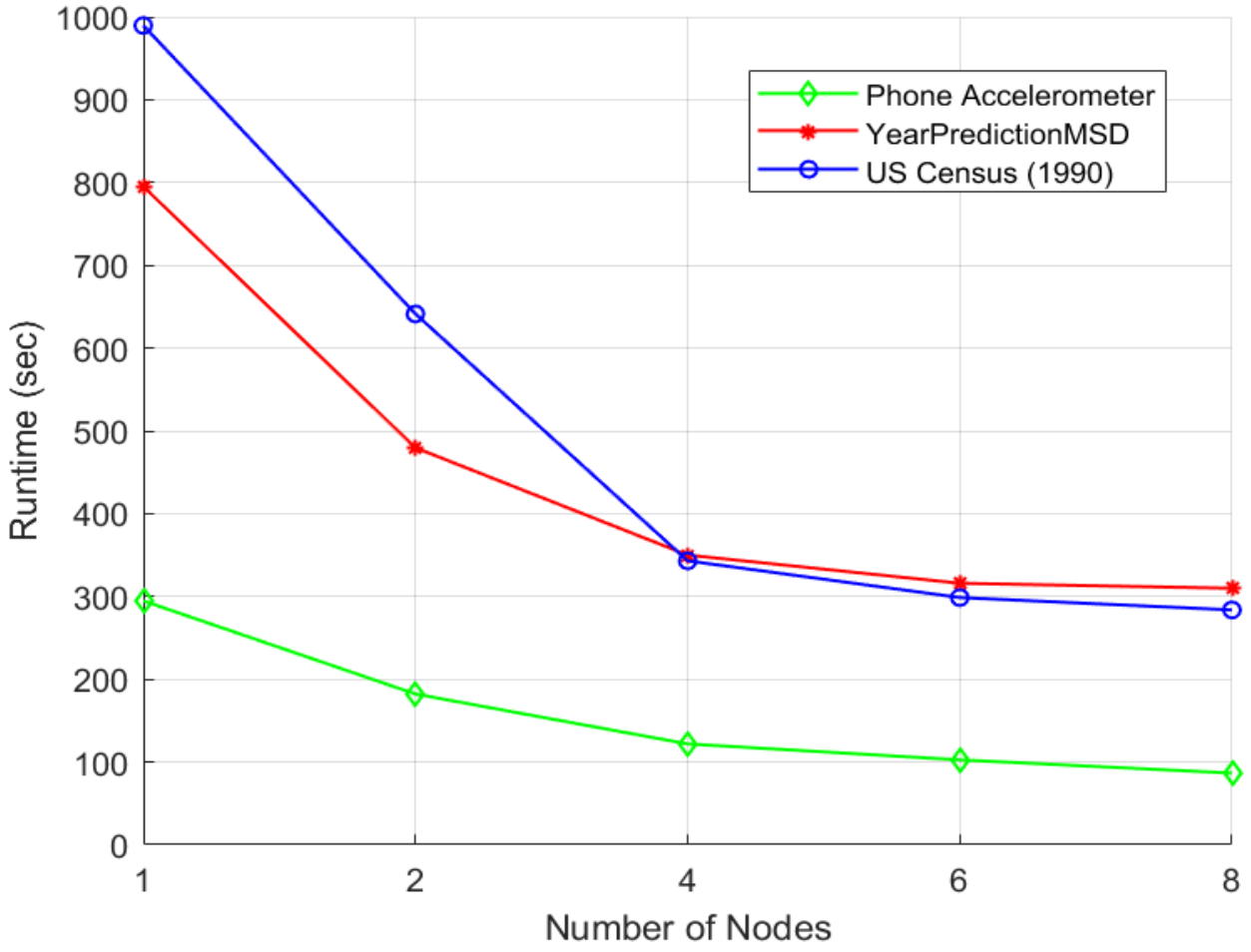
(b)



(c)

Evaluation of the runtime (sec) with different number of iterations on Phone Accelerometer (a), YearPredictionMSD (b), US Census (1990) dataset (c)

Algorithm 2: Parallel Batch k-means for Big Data Clustering



Runtime performance of the proposed approach for different number of nodes on the three datasets

Algorithm 3: Weighted Clustering for Anomaly Detection in Big Data

Input: $X = (x_1, x_2, ..., x_n)$

$w = (w_1, w_2, ..., w_n)$

k : number of clusters

Output: Vector of cluster indices $IDX = (idx_1, idx_2, ..., idx_n)$

Step 1. Find the center of all points in the dataset (*O*)

Step 2. Calculate the weights of all points x_i

$$w(x_i) = \|x_i - O\|, \quad O = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Step 3. $s = 0$

Step 4. Calculate the function value

$$f^{(s)} = \sum_{p=1}^{k} \sum_{x_i \in C_p} \left|C_p\right|_W * \left\| x_i - O_p{}^{(s)} \right\|^2$$

taking into account $\left|C_p\right|_W = \sum_{x_i \in C_p} w(x_i), \quad p = 1,2,...,k$

Step 5. $s = s + 1$

Step 6. Repeat steps 3-5 until the convergence condition is met:

$$\left| \frac{f^{(s+1)} - f^{(s)}}{f^{(s)}} \right| \le \varepsilon$$

Step 7. Return the values of IDX

End

Algorithm 3: Weighted Clustering for Anomaly Detection in Big Data

Performance evaluation of the proposed algorithm with the k-means algorithm

Dataset	Purity (%)	Mirkin (%)	F-measure (%)	Variation of information (%)	Partition coefficient (%)
NSL_KDD	3.78 (+)	0.44 (+)	0.66 (+)	13.03 (-)	10.55 (+)
Forest Cover Type	0	14.97 (+)	18.18 (+)	12.76 (+)	1.88 (+)

$$\frac{our\_algorithm - another\_algorithm}{another\_algorithm} \times 100\%$$

"**+**" means that the result outperforms, and "**-**" the opposite.

Algorithm 4: Anomaly Detection in Big Data based on Clustering

Input: $X = (x_1, x_2, ..., x_n)$

α : regularization parameter

k : number of clusters

Output: Vector of cluster indices $IDX = (idx_1, idx_2, ..., idx_n)$

Step 1. Find the center of all points in the dataset (*O*)

Step 2. $s = 0$

Step 3. Calculate the compactness

$$S_W = \sum_{p=1}^{k} \sum_{i=1}^{n} (x_i - O_p)(x_i - O_p)^T,$$

where $O = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad O_p = \frac{1}{n_p} \sum_{x_i \in C_p} x_i, \quad n_p = |C_p|, \quad p = 1,2,...,k$

Step 4. Calculate the separation of clusters

$$S_{BW} = \sum_{p=1}^{k-1} \sum_{q=p+1}^{k} (O_p - O_q)(O_p - O_q)^T$$

Step 5. Calculate the remoteness

$$S_B = \sum_{p=1}^{k} (O_p - O)(O - O_p)^T$$

Step 6. Calculate the value of the following function

$$F_3(x) = \frac{1}{S_W^{(s)}} \left(\alpha S_B^{(s)} + (1 - \alpha) S_{BW}^{(s)} \right) \to \max$$

Step 7. $s = s + 1$

Step 8. Repeat steps 3-7 until the convergence condition is met $\left| \frac{f^{(s+1)} - f^{(s)}}{f^{(s)}} \right| \le \varepsilon$

Step 9. Return the values of IDX

End

Algorithm 4: Anomaly Detection in Big Data based on Clustering

Performance of the proposed algorithm for α=0.1

Dataset	Purity	Mirkin	F-measure	Variation of information	Partition coefficient
NSL_KDD	0.5457	0.4958	0.6270	0.0736	0.3538
Forest Cover Type	**0.9647**	**0.0753**	**0.9764**	**0.0135**	**0.4757**

Performance of the proposed algorithm for α=0.3

Dataset	Purity	Mirkin	F-measure	Variation of information	Partition coefficient
NSL_KDD	0.5525	0.4945	0.6214	0.0747	0.3521
Forest Cover Type	**0.9647**	**0.0893**	**0.9652**	**0.0163**	**0.4756**

Performance of the proposed algorithm for α=0.5

Dataset	Purity	Mirkin	F-measure	Variation of information	Partition coefficient
NSL_KDD	0.5601	0.4928	0.6152	0.0759	0.3505
Forest Cover Type	**0.9647**	0.0933	**0.9621**	**0.0171**	**0.4759**

Algorithm 5: Anomaly Detection based on Optimization

Input: $r = \{r_1, r_2, ..., r_n\}$

λ : regularization parameter

$\beta^{(0)} = \{\beta_1^{(0)}, ..., \beta_n^{(0)}\}$: initial weights of data points

k : number of clusters

Output: $\beta^{(s)} = \{\beta_1^{(s)}, ..., \beta_n^{(s)}\}$

Step 1. Calculate the centers of clusters r^*

Step 2. $s = 0$

Step 3. for all $r_i \in R^n$ do

$$c_i = \sum_{i=1}^{n} \beta_i^{(s)} \left\| r^* - r_i \right\|^2$$

end for

Step 4. Calculate the function value

$$f^{(s)} = (1-\lambda)\sum_{i=1}^{n} c_i + \lambda \left\| \beta^{(s)} \right\|^2$$

where $\sum_{i=1}^{n} \beta_i = 1; \beta_i \geq 0 \ \forall i$, $0 \leq \lambda \leq 1$

Step 5. $s = s + 1$

Step 6. Repeat steps 3-5 until the convergence condition is met

Step 7. Return $\beta^{(s)}$

End

Algorithm 5: Anomaly Detection based on Optimization

Performance evaluation of the proposed approach on NSL-KDD dataset

λ	Purity (%)	Mirkin (%)	F-measure (%)	Variation of information (%)	Partition coefficient (%)
0.1	0.5330	**0.4982**	0.6371	0.1267	**0.3542**
0.2	0.5330	0.4985	0.6405	0.1261	0.3533
0.3	0.5330	**0.4982**	0.6371	0.1267	**0.3542**
0.4	0.5330	0.4997	**0.6931**	**0.1139**	0.3371
0.5	0.5330	0.4984	0.6394	0.1263	0.3536
0.6	0.5330	0.4996	0.6560	0.1233	0.3484
0.7	0.5330	0.4983	0.6383	0.1265	0.3539
0.8	0.5330	**0.4982**	0.6360	0.1280	0.3484
0.9	0.5330	0.4983	0.6348	0.1300	0.3377

Acknowledgments

It is well known that it is impossible to analyze the amount of Data created today with yesterday's resources. **AzScienceNet** has a special role in resolving these challenges. In recent years, with the support of **GÉANT** and the **EaPConnect** project, the reconstruction of **AzScienceNet** played a significant role in making our research more effective. Previously, storage problems were emerging. Even though it was possible to process data in parts, it would take days to resolve those issues. Thanks to **GÉANT** and **EaPConnect**, using **AzScienceNet's** opportunities, it was possible to resolve these problems in a reasonable time. With technology and methods, Big Data can be analyzed, and new knowledge can be gained. AzScienceNet is of great importance for Azerbaijani science and education. Scholars will continue to benefit from opportunities provided by **GÉANT** and **EaPConnect** through **AzScienceNet**.

Acknowledgments

The following projects were technically supported by **GÉANT**, **EaPConnect** and **AzScienceNet**:

1. **"Methods and algorithms for providing information security in Big Data environment and some of their applications"** (funded by the Science Development Foundation under the President of the Republic of Azerbaijan)
2. "**Development of methods and algorithms for increasing the efficiency of electronic government using Big Data analytics technology**" (funded by the Science Development Foundation under the President of the Republic of Azerbaijan)
3. "**Analysis of Big Data collected in the oil and gas sector**" (funded by the Science Foundation of SOCAR)
4. **"Investigation of Big Data Analytics technology for oil and gas industry"** (funded by the Science Foundation of SOCAR)

List of Selected Papers Published by us

1. R.M. Alguliyev, R.M. Aliguliyev, L.V. Sukhostat, “Parallel batch k-means for Big Data clustering” // Computers & Industrial Engineering, vol.152, pp.1-16, 2021. (WoS, **IF: 7.180**)
2. R.M. Alguliyev, R.M. Aliguliyev, L.V. Sukhostat, “Weighted consensus clustering and its application to Big Data” // Expert Systems with Applications, vol.150, pp.1-15, 2020. (WoS, **IF: 8.665**)
3. R.M. Alguliyev, R.M. Aliguliyev, L.V. Sukhostat, “Efficient algorithm for Big Data clustering on single machine” // CAAI Transactions on Intelligence Technology, vol.5, no.1, pp.9-14, 2020. (WoS, **IF: 7.985**)
4. R.M. Alguliyev, R.M. Aliguliyev, F.D. Abdullayeva, “Privacy-preserving deep learning algorithm for Big personal Data analysis” // Journal of Industrial Information Integration, vol.15, pp.1-14, 2019. (WoS, **IF: 10.615**)
5. R.M. Alguliyev, R.M. Aliguliyev, F.D. Abdullayeva, “Multidisciplinary study of the problems of Big Data technologies in the oil and gas industry” // International Journal of Oil, Gas and Coal Technology, vol.23, no.1, pp.92-105, 2020. (WoS, **IF: 0.820**)
6. R.M. Aliguliyev, R.K. Alekberov, S.F. Tahirzada, “An Architecture for Big IoT Data Analytics in the Oil and Gas Industry” // International Journal of Hyperconnectivity and the Internet of Things, vol.4, no.2, pp.25-37, 2020. (WoS: ESCI)
7. R.M. Alguliyev, R.M. Aliguliyev, F.J. Abdullayeva, “Hybridisation of classifiers for anomaly detection in Big Data” // International Journal of Big Data Intelligence, vol.6, no. 1, pp.11-19, 2019. (WoS: ESCI)
8. R.M. Alguliyev, R.M. Aliguliyev, Y.N. Imamverdiyev, L.V. Sukhostat, “Weighted clustering for anomaly detection in Big Data” // Statistics, Optimization and Information Computing, vol.6, no.2, pp.178-188, 2018. (Scopus)
9. R.M. Alguliyev, R.M. Aliguliyev, F.D. Abdullayev, “PSO+k-means algorithm for anomaly detection in Big Data” // Statistics, Optimization and Information Computing, vol.7, pp. 348-359, 2019. (Scopus)
10. R.M. Alguliyev, R.M. Aliguliyev, L.V. Sukhostat, “Anomaly detection in Big Data based on clustering” // Statistic Optimization and Information Computing, vol.5, no.4, pp.325-340, 2017. (Scopus)
11. R.M. Alguliyev, R.M. Aliguliyev, Y.N. Imamverdiyev, L.V. Sukhostat, “An anomaly detection based on optimization” // International Journal of Intelligent Systems and Applications, vol. 9, no. 12, pp. 87–96, 2017. (Scopus)

B I G

Thank You

For Your Attention!