
Formation Smart Data Science for Automated Analytics of Modeling of Scientific Experiments

Evgeniy Bryndin

Scientific Department, Research Center "Estestvoinformatika", Novosibirsk, Russia

Email address:

bryndin15@yandex.ru

To cite this article:

Evgeniy Bryndin. Formation Smart Data Science for Automated Analytics of Modeling of Scientific Experiments. *American Journal of Software Engineering and Applications*. Vol. 8, No. 2, 2019, pp. 36-43. doi: 10.11648/j.ajsea.20190802.11

Received: October 7, 2019; **Accepted:** October 25, 2019; **Published:** October 31, 2019

Abstract: Specialists of past generations have accumulated huge and valuable experience of analytics and forecasting in a variety of subject areas. Within the framework of the development of methods of processing scientific data, taking into account the accumulated experience of generations of specialists, it is possible to structure, specify, classify and rank them into flows of smart data for automation of research of various scientific problems by cognitive systems of artificial intelligence. Automated analytics is based on big data. A large amount of data is generated in real time by modeling a scientific experiment. When working with data, it must be processed as efficiently as possible to get the correct output. The main thing is to prepare the training sample correctly. If you select the training data sampling principle correctly, you can scale the task using a more complete set of data. It should be understood that rationing and data preparation is crucial for traditional machine learning. This process significantly affects the choice of the architecture of neural networks used, especially in so-called deep learning, when it is necessary to correctly determine the number of hidden layers in the neural network and the number of artificial neurons in them. One of the main advantages of multilayer neural networks is the simulation of the work of some complex mathematical dependence.

Keywords: Scientific Data, Topology of Data, Artificial Intelligence, Trance - Disciplinary Researches

1. Introduction

Data topologies classify, cluster, and manage data scenarios that embrace the competing priorities and needs of any experiment. A well-designed data topology considers the users, use, constraints, and flow of data, and is resilient to future needs and the adoption of new technologies.

"Single source of the truth" is a frequent practice in designing analytic data architectures to minimize redundancy, improve efficiency, and derive trust through shared data. In support of this practice, data silos are often considered an obstacle to productive use of data.

Many data silos are closed environments that can't readily participate with processes that are external to the silo. However, if you manage and access the data through integration or interoperability, the silo can serve as an asset.

Most organizations, and their individuals and teams, operate within areas of specialization. Those areas of specialization have corresponding specialized data needs.

Data silos that are deployed with managed interoperability

and integration can support specialized data needs and provide an optimal means to unravel the corporate system. Designing data flows as part of a planned data topology is the difference between a good data silo and a bad data silo.

To meet the needs of complex data organization, you can use different technologies to solve individual problems. The capability that is part of the database is separate from the capability in the transactional process used in the scientific experiment. Scientific research may require preservation of the same data using multiple technologies.

Today's information architecture may require multiple database technologies in the data center or their distribution across multiple data centers and external cloud environments.

Isolated data stores that are planned, managed, and compatible can provide an advantage by using data topology to provide a planned and organized environment and to maintain compatibility.

The practice of managing environments is aimed at perfect analytics. As analytics is embedded in an increasing number of processes and applications, analytics can go beyond the

environment. The data environment is often divided into areas such as raw, stepped, twisted, sandbox, and general analytics. However, these high-level zone definitions typically do not provide the means to implement or manage data scenarios, given such complexities as multi-cloud deployment, security, multi-data processing technologies, and the nature of the experiment.

Data zones represent a representation of them with general purpose, necessity and use (qualitative characteristics). Data zones do not prescribe technology such as hardware or software. Knowledge of the purpose, necessity and use of the data area leads to the development and adoption of appropriate technological solutions. When designing zones, it is necessary to take into account restrictions that are beyond direct control.

Matching a data zone can reduce the likelihood of an arbitrary storage decision and ensure that the zone design supports ongoing experiments. You can create zones to support persistent data, time data, or both.

Data topologies classify and manage real-world data scenarios in a holistic and meaningful manner, encompassing competing priorities and technologies of scientific experimentation. To create a properly designed data topology, consider users (or endpoints), usage, purpose, demand, constraints, and data flow.

2. The Data Topology Is Resilient to Future Needs

A well-designed data topology is resilient and resilient to future needs, new technologies, and continuous changes in data characteristics, including volume, diversity, speed, credibility, and perceived value.

Data topology design is an iterative process:

1. Consolidate users or endpoints into communities of interest to identify common needs.
2. Classification and clustering of data into zones with common qualitative characteristics, such as usage, purpose and need, which are not limited to specific technologies or quantitative characteristics.
3. Display and align communities of interest to data zones.
4. Add restrictions to further develop the zone map and align it with functional and non-functional requirements and capabilities.
5. Working backwards in the data stream to identify areas of synergy and reuse.
6. Define data flow between and within zones to support certain constraints.

Limitations are real factors that need to be taken into account when constructing a zone design for scientific experiments. The restricted zone map and corresponding data streams form the basis for implementing the data topology.

Location and security are key constraints to support distributed and multi-core deployments. It is always good to look at progress outside the data area and you need to consider this when designing your data topology. Many

functional and non-functional limitations need to be considered when designing zones and zone maps.

A data zone can be fed by more than one zone. In some cases, data is created in a zone as a result of an analytic process. The persistence of data might result from feedback or the creation of intermediate results. Aggregated and derived data can also be created and stored in a data zone and shared through a data flow. If the data in a data zone must be shared or flowed to another zone, make sure that the flow comes through a common staging data zone to maintain data lineage and provenance.

The flow of data can use multiple types of technology, from basic push-and-pull mechanisms to elaborate trigger-and-event engines. Data might travel over data streams, message queues, message buses, and batch processes.

If you apply analytics during the flight path from a source to a target, store any interim results that must be persisted in the raw data zone. Then, propagate the results to other data zones as necessary.

The successful implementation of data zones provides high-quality, ready-to-use data in a self-service model for all data users. This approach ensures access to data for those who need it, while respecting regulations.

Externally, the concept of data zones seems sound: store data from multiple sources - regardless of format - in one place, use innovation of open source technologies and big data, deploy on research infrastructure. Following these simple steps can provide a platform for self-service analytics and the development of data science. Building data zones is an iterative process that is tied to desired results, constantly measured against goals, and adapted to lessons learned to support new goals and priorities, leverage innovation in multidimensional work, automation, and management.

Artificial intelligence, data science and analytics are the same as your data. If you cannot trust or understand data from data zones, it is not ready for use. Data alone makes no sense. File formats such as JSON and XML, as well as schemas can help analyze data, but they do not provide insight into its meaning. Attributes with the same name can have different values, especially in different sources or instances.

The knowledge catalog is a critical component of any data zone engineering design. It provides a framework for understanding data, sharing knowledge, and reuse. Just as libraries use the catalog to find books and related materials, the knowledge catalog provides an inventory to find and understand information resources. The catalog brings together employees and policies to support the management and availability of this data to the business while meeting ever-changing regulatory requirements.

The Knowledge Catalog contains a lot of information to make information resources meaningful:

1. Semantic data value;
2. Data format;
3. Linearity (where it fits into the end-to-end flow) and data source;
4. Trust level or data quality;

5. Level of detail or aggregation;
6. Confidential, restricted, or regulated data;
7. Data policies;
8. Use of data and self-service ("shop-for-data");

You don't need to wait until data zones are built to start creating a knowledge catalog and inventory information resources. When you register new data in a catalog, you can capture the data source, data quality analysis, and classification rules using automation tools and tools. Similarly, when you use the catalog to "purchase data" the metrics of the data sources used, you can obtain confidence estimates within the management workflow.

Some data in the knowledge catalog is created and managed by employees, such as terminology, data owners and managers, and processes and rules. Together, the knowledge catalog provides the meaning of the data and provides the basis for trust and reuse in any data zone project.

Traditional approaches to big data projects begin with technology and involve some level of centralization and consolidation. The challenge with these approaches is that they fail to recognize the diversity of user, device, and API needs in specialized data in the broader ecosystem of transsectoral research partners.

A data topology is a user-oriented design technique that simplifies the organization, flow, and management of data to support goals and results. The data topology takes into account the reality of the need for specialized data in teams of researchers. With user-oriented data design and processing methods, the data topology provides a bridge between goals and technology and infrastructure to ensure and achieve these goals.

Data topologies begin with the individual needs of users, devices, and APIs. The data flow between zones drives the data topology, providing data to users, devices, and APIs to support the goals. The data stream may be unidirectional, bidirectional, and cyclic. By managing access and interaction between zones, isolated repositories of individual needs become the domain of researchers.

Shared data is an important set of zones in any topology. Shared data is the minimum set of atomic data required to level zones and facilitate trust, reuse, and compliance.

Data topologies do not prescribe, but define technological solutions. A well-designed data topology must be sustainable to meet future needs and introduce new technologies to support goals.

The data topology zone map and data flow structure define common characteristics for data management and integration to support the specialized needs of users, devices, and APIs.

Today's data zone architecture may require multiple database or storage technologies.

3. Leverage Innovation in Multidimensional Work, Automation, and Management

While successful data zone projects start small, they grow

and expand in support of goals, organizations and researchers, and the entire research ecosystem. With innovations in technical architectures and multi-cloud capabilities, you can simplify the deployment and management of data zones in data centers, cloud providers, local and private clouds.

Automation and containerization platforms can significantly accelerate the deployment of physical architectures. Any modern data array should consider architecture concepts, including infrastructure virtualization, separation of computing from storage, automated monitoring and maintenance to reduce overall costs and increase availability. Similarly, machine learning and automation must be part of the data management and data flow capabilities of any data zone design to complement traditionally manual processes:

1. Data discovery and registration of new data sources;
2. Classification and understanding of data;
3. Content-aware data extraction tools such as text, natural language, documents, and image;
4. Identification of confidential, private and regulated data;
5. Object resolution and compliance;
6. Data quality analysis;
7. Integrate and move data with rules to improve its quality;
8. Measurement of data quality;

Data zones are as successful as the people, policies, and processes that support them. Data placement, understanding, defining and managing security, protecting and maintaining data compliance, and delivering data using data zones requires an approach that is consistent with business needs and management authority.

With the right mix of staff, processes and technologies, you can create architecture of data and analytics that meets the objectives of the study [1-17].

4. Preparing Scientific Data for Artificial Intelligence

Preparing initial data for artificial intelligence is one of the most important and often the most time-consuming parts of data analysis. It is best practice to record any data preparation tasks in the data preparation report.

After performing preliminary tasks to transform the problem into an artificial intelligence solution and understanding the data needs to support the problem, you need to prepare the data. Data must be prepared in a format that can be used to design, measure, and train the machine learning model.

Most artificial intelligence and data models require data to be combined and de-normalized into one large analytical record before data mining, element selection, model development and optimization, and learning begin.

Data preparation includes the following tasks:

1. Select a sample subset of data.

Filter by strings targeted at specific researchers or technologies that help respond to data analysis. You should also filter attributes related to data analysis. Some scientific data models may require new data.

2. Merge datasets.

A shared key is required to join datasets. Aggregation of records and aggregation based on similarity grouping operations.

3. Getting new attributes

When merging datasets, it may be useful to derive new attributes.

4. Format and sort data for modeling

Sequential and temporal algorithms may require pre-sorting the data in a particular order. Categorical data fields may need to be converted from text categories to numeric ones.

There are two categorical data variables:

- 1. Denomination: categories are marked without any order of precedence, e.g. London, Paris and Berlin.
- 2. Sequence number: categories are marked where priority order exists, e.g. low, medium and high.

Although some mining algorithms can handle categorical data, many require it to be converted into a numerical representation. This transformation can be performed using a variety of approaches, often referred to as coding techniques.

In the label encoding approach, each unique category value is assigned a specified numeric value. This approach is easy to apply and suitable for ordinal categorical data, but can mislead nominal data. The problem with label coding is that some algorithms may view a value as a relative measure of magnitude:

Category Value Encoding

Low	1
Medium	2
High	3

Some algorithms can see a value 3 times that of 1.

Category Value Encoding

Electron	1
Ion	2
Nanoparticles	3
Photon	4

Another approach is one hot coding. This approach converts each category value to a new column and assigns a value of 1 or 0 (true or false). This approach overcomes the problem of label coding weight and can add many dimensions to the data.

Must be done before creating the algorithm itself - processing (preparation, clustering) and data normalization. This step is needed to prepare a sample for proper interpretation by the mathematical model of AI. For example, operations can be performed on a numerical sample, such as raising variables to a degree or multiplying by a constant, which will allow linear models to model nonlinear dependencies, to identify common patterns. It is sometimes necessary to perform a Fourier transform to correctly

interpret frequency characteristics in audio processing or to use a SIFT algorithm in solving an image mapping problem.

When we have a clear research goal, a true set of source data for the sample and the sample itself, we can start developing neural network models, programming and further training the neural network. The learning stage will include selecting the learning algorithm, applying the learning algorithm, visualizing it, and evaluating the quality of learning. Its efficient and correct execution on large data samples is not easy to ensure. It is necessary to correctly select the algorithm of neural network training, otherwise the artificial intelligence created can learn to misinterpret the incoming data flow. The finite behavior created by the artificial intelligence model is derived from the set of source data, the procedures for processing and normalizing them, and the learning algorithm used and the criterion for validating the learning result. Several approaches to learning allow the neural network to be properly trained so that the developed mathematical model receives the result of learning as reliably as possible. One of the defining solutions will be what percentage of the training sample will be used to train the neural network, and which will be presented to the neural network later, to obtain results and new simulation data of the scientific experiment. Modern deep learning approaches involve using up to 99% of the data for training, and 1% for testing.

A team of researchers should take on work on neural network learning, particularly the development and validation of machine learning algorithms. One part of the team will conduct training, and the other will test the algorithm, checking how accurately the model of the scientific experiment solves the obtained task. Once a team of researchers is confident in the correct solution to the problem, it can work to transform the results of the neural network into ideas, action elements, predictions, or simply use them as a result of data processing. Scheme of universal simulation of scientific experiments based on strong AI is shown in application.

Only by ensuring that the mathematical model really works correctly and is able to control complex modeling processes can the work be considered completed. Achieving this result requires serious skills. This is helped by backward links between the stages, which will make the scheme universal. By adhering to this scheme, one can learn to model and investigate a large number of scientific experiments.

5. Conclusion

When a sufficient amount of data is available that describes process, it is necessary to search for information explaining the correlation of the individual its parameters. Such information about correlations is sufficient to predict further development of the process under consideration by induction and derivation by cognitive methods. Trends in the development of modern digital technologies make it quite clear that such information is inevitable. Correlations must be compatible with the underlying patterns of the process under

investigation to choose the right path and get the right conclusions from the opportunities facing the researchers. One should not forget the fundamental role of the experiment for correlation. The study of big data in the context of the nature of cognitive creativity opens up new perspectives for researchers. All phenomena, objects, processes of Nature are connected with oscillatory processes (vibrations and their energy). Oscillatory processes are characterized by the frequency, intensity, the period, a form, distribution speed,

recurrence, a resonance or attenuation, interrelation and other parameters. This Smart Big Data can be structured, classified, provided to a logical format and on the basis of them to reveal regularities of wave processes and to visualize them and related natural entities with them. Engineers and programmers create cognitive robots of structuring and classification of Smart Big Data for identification of regularities by a logical format in the conditions of space.

6. Application

1. General

Use case name	Application of Strong Artificial Intelligence			
Application domain	Hi-Tech Labor Market			
Deployment Model	Human digital double			
Status	Results of research: Strong Artificial Intelligence			
Scope	Economic sectors and social services			
Objective (s)	Find accurate and universal application of strong artificial intelligence.			
Narrative	Short description (not more than 150 words)	Strong Artificial Intelligence - scientific applied direction on development and creation of technological and program cognitive complexes of the digital double of intelligence of the person capable to training, retraining, self-realization and self-improvement on the basis of criterion of preferences and to improvement of functional activity by the high-quality choice and development of creative innovative hi-tech professional and behavioural skills and competences. Technology that studies the development of chelatinous digital twins capable of acquiring, processing and applying human knowledge and skills, purchased through training, to solve problems, adapt to changing circumstances with or without human or external control in physical work, as well as in mental or cognitive work. Technology builds models by analyzing quantitative and qualitative data from different perspectives and measurements, classifying them, and summarizing potential relationships and impacts. The technology uses natural language processing and machine learning to interact more naturally and expand human experience and knowledge on a permanent basis during operation. The technology has robust mechanisms by which to ensure security in way that humans would understand. Technology showing smart behavior comparable to human across the range of cognitive abilities. The technology models the spectrum of human abilities by retraining. The technology relies on the infrastructure of interconnected actors, people, systems and information resources of high-tech industry and social sphere, as well as on services that process and respond to information from the physical and virtual world of social cognitive smart robots: guide, seller, teacher, nurse, volunteer, guard, administrator.		
	Complete description			
Stakeholders	Highly technological producer			
Stakeholders' assets, values	Reputation			
System's threats and vulnerabilities	Legal and ethical aspects of interaction with society.			
Key performance indicators (KPIs)	ID	Name	Description	Reference to mentioned use case objectives
	1	AI management of professional cooperation process	The technology of creative processes control can itself predict optimal terms of execution of certain stages on the basis of accumulated information about their labour intensity, selection of the route of staff load and competences of employees. Optimize processes during their execution - automatic delegation of tasks taking into account the load of employees and their competences. Strong artificial intelligence works with fewer mistakes and is safer.	Improve accuracy
	2	Productivity and quality AI	Strong artificial improves the quality of life of man and society in daily concerns, as well as productivity in high-tech industry and production.	Improve efficiency

AI features	Task (s) Method (s) Hardware Topology Terms and concepts used	Creative activity Deep learning Supercomputer with Strong Artificial Intelligence Distributed Modular Interconnect Topology Deep learning, "imagification", neural network, training, training data set
Standardization opportunities/ requirements	Strong artificial intelligence requires process standardization, as does every human activity.	
Challenges and issues	Qualitatively new type of thinking not available to humans.	
Societal concerns	Description SDGs to be achieved	Security and ethical and legal aspects Universal approach to big data processing with smart cognitive systems

2. Data

Data characteristics	
Description	Strong Artificial Intelligence Data
Source	Model and technology of Strong Artificial Intelligence
Type	Strong
Volume (size)	Hi-Tech Labor Market
Velocity (e.g. real time)	Supercomputing Velocity
Variety (multiple datasets)	streams of multiple datasets
Variability (rate of change)	Retraining
Quality	High

3. Process Scenario

Scenario conditions					
No.	Scenario name	Scenario description	Triggering event	Pre-condition	Post-condition
1	Training	Train a model (deep neural network) with training data set	Technological process raw data set is ready	Formatting of data	Management of safety
2	Evaluation	Expansion of the trained model	Development of technological thinking and behaviour	Cognitive thinking patterns and psychological behaviors	Meeting KPI requirements is condition of development
3	Execution	Model and Technology Tooling	Interaction	Activization of Model	Completion of interaction
4	Retraining	Retrain a model with training data set	Certain period of time has passed since the last training/ retraining	Additional data and knowledge	Combining Data and Knowledge

4. Training

Scenario name		Training			
Step No.	Event	Name of process/Activity	Primary actor	Description of process/activity	Requirement
1	Sample raw data set is ready	Specification and classification	Manufacturer	Transform sample raw data	Strong AI Software
2	Completion of Step 1	Creating Set of Experimental Data	Manufacturer	Development of a set of experimental data through job modelling	Software of modelling
3	Completion of Step 2	Model training	AI solution provider	Train a model (deep neural network) with experimental data set created by Step 2	

5. Evaluation

Scenario name		Evaluation			
Step No.	Event	Name of process/Activity	Primary actor	Description of process/activity	Requirement
1	Completion of training/retraining	Research	Manufacturer	Train model (deep neural network) with experimental data set created	
2	Completion of Step 1	Identification	AI solution provider	Based on data, detect execution using a deep neural network trained in a learning scenario	
3	Completion of Step 2	Evaluation	Manufacturer	Comparison of phase 2 results with human performance	
Input of evaluation					
Output of evaluation					

6. Execution

Scenario name		Execution			
Step No.	Event	Name of process/Activity	Primary actor	Description of process/activity	Requirement
1	Completion of comparison of modeling results with human performance	Research	Manufacturer	Development of a set of experimental data through job modelling	
2	Completion of Step 1	Identification	Manufacturer	Based on modified data train model (deep neural network) with experimental data set created	The trained model with deep neural network has to be handed over to the manufacturer
Input of Execution					
Output of Execution					

7. Retraining

Scenario name		Retraining			
Step No.	Event	Name of process/Activity	Primary actor	Description of process/activity	Requirement
1	Certain period of time has passed since the last training/retraining	Research	Manufacturer	Additional data and knowledge	
2	Completion of Step 1	Experimental data set creation	Manufacturer	Combining Data and Knowledge Based on modified data train model (deep neural network) with experimental data set created	
3	Completion of Step 2	Model training	AI solution provider	Comparison of phase 2 results with human performance	
Specification of retraining data		Retraining data set has to include recent data			

References

- [1] Glukhova O. E. Theoretical methods of a research of nanostructures. Messenger of SSU, Natural-science series, Release 9. 2012. Page 106-117.
- [2] Brazhe R. A. Mathematical modeling of nanostructures and their physical properties. USTUY. 2014. 99 pages.
- [3] Evgeniy Bryndin. BIG DATA MODELLING of TRANSFORMATION and BIFURCATION of NANOSTRUCTURES. Inter conference "Management of development of large-scale systems (MLSD'2018)". T. 2 - M.: IPM RAS, 2018. Page 340-343.
- [4] Demchenko Y., Laat C. De, Membrey P. Defining architecture components of the Big Data Ecosystem. Collaboration Technologies and Systems (CTS), 2014 International Conference. 2014. May. P. 104-112.
- [5] Evgeniy Bryndin. Cognitive Robots with Imitative Thinking for Digital Libraries, Banks, Universities and Smart Factories. International Journal of Management and Fuzzy Systems. V. 3, N. 5, 2017, pp 57-66.
- [6] Evgeniy Bryndin. Program Hierarchical Realization of Adaptation Behavior of the Cognitive Mobile Robot with Imitative Thinking. International Journal of Engineering Management. Volume 1, Issue 4. 2017, pp. 74-79.
- [7] Evgeniy Bryndin. Technological Thinking, Communication and Behavior of Androids. Communications. Vol. 6, No. 1, 2018. Pages: 13-19.
- [8] Evgeniy Bryndin. Communicative Associative Logic of Cognitive Professional Robot with Imitative Thinking. Journal Engineering Mathematics, V. 2, Issue 2. 2018. Pages: 79-85.
- [9] Victor Maier-Shenberger, Kenneth Kukyer. Big data. Revolution which will change how we live we work and we think. — M.: Mann, Ivanov and Ferber, 2014. —240 pages.
- [10] Evgeniy Bryndin. Modeling of Transformation of Nanostructures by Cognitive Systems on the Basis of Big Smart Data. International Journal of Artificial Intelligence and Mechatronics. Volume 7, Issue 4. 2019. P. 19-22.
- [11] Evgeniy Bryndin. Digital technologies of the industry 4.0. Chapter 10, C. 201-222, Book: Computer Science Advances: Research and Applications. USA: Nova Science Publisher. 2019. 252 pages.
- [12] Evgeniy Bryndin. System retraining to professional competences of cognitive robots on basis of communicative associative logic of technological thinking. International Robotics Automation Journal. 2019; 5 (3): 112-119.
- [13] Evgeniy Bryndin. Mobile Innovative Transformational Ecosystem of Management of Humane Technological Society. Integrative Journal of Conference Proceedings. Volume 1, Issue 3, 2019.
- [14] Evgeniy Bryndin. Human Digital Doubles with Technological Cognitive Thinking and Adaptive Behaviour. Software Engineering, Volume 7, Issue 1, 2019. P. 1-9.
- [15] Evgeniy Bryndin, Irina Bryndina. Technological Diagnostics of Human Condition According to Spectral Analysis of Biofield. *Advances in Bioscience and Bioengineering. Volume 7, Issue 3, 2019. Pages: 64-68.*
- [16] Evgeniy Bryndin. Mainstreaming technological development of industrial production based on artificial intelligence. *COJ Technical & Scientific Research, 2 (3). 2019. Pages: 1-5.*

- [17] *Evgeniy Bryndin*. Robots with Artificial Intelligence and Spectroscopic Sight in Hi-Tech Labor Market. *International Journal of Systems Science and Applied Mathematic*, V. 4, № 3, 2019. Pages: 31-37.