11TH ISC2019 INTERNATIONAL STATISTICS CONGRESS

Data Science and Big Data Analytics Education - Serbian perspective –

Dragan Vukmirović[1], Tijana Čomić, Zoran Radojičić[1]
[1]Belgrade University, Serbia
[2]GOPA mbH, Bad Homburg, Germany

dragan.vukmirovic@fon.bg.ac.rs

VOGUE HOTEL BODRUM Muğla

4-8 OCTOBER 2019

Data science is a multidisciplinary scientific field

- Data never exists in a vacuum.
- Like a biological organism, data has a life cycle, from birth through an active life to “immortality” or some form of expiration.
- Also like a living and intelligent organism, it survives in an environment that provides physical support, social context, and existential meaning.
- The data life cycle is critical to understanding the opportunities and challenges of making the most of digital data;

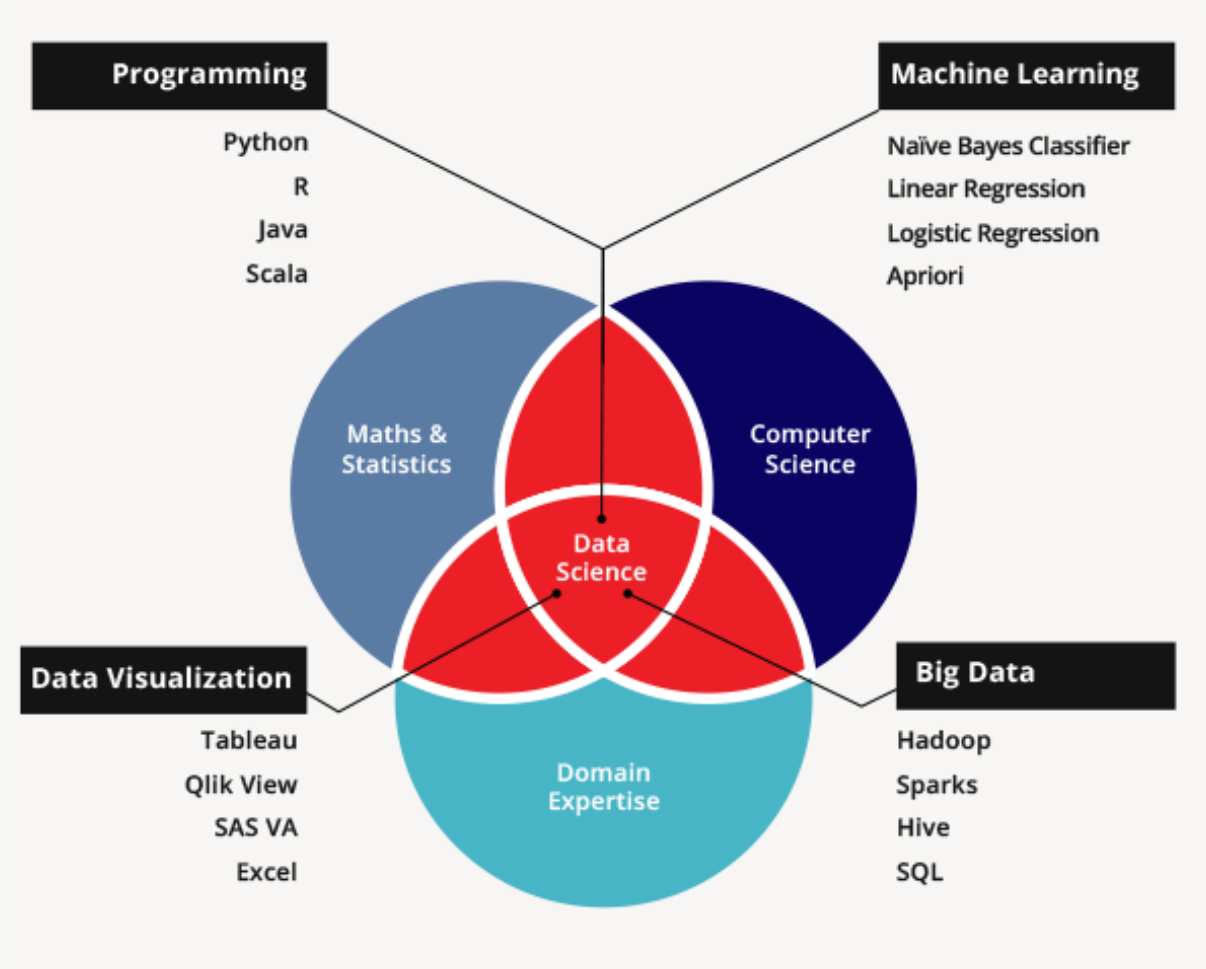
(https://www.nsf.gov/cise/ac-data-science-report/ CISEACDataScienceReport1.19.17.pdf).

Summary

- Data science is a multidisciplinary scientific **field** based on 3 pillars:

1. **Classics (computational) science**: Mathematics, Probability, Statistics, Operational research, Expert systems, Optimization, Decision Theory....
2. **Computer & Information science**, and
3. **Application domains** (including management).

Data Science



https://acadgild.com/blog/managers-data-science

Many universities have begun to offer data science educational programs at all levels of undergraduate, Master's, and doctoral study… but… **as a discipline, data science is only in its infancy**

- "From the current diversity of curricula and programs, data science is going through an important and healthy period of experimentation.

- It is important that we do not "standardize" data science too quickly, continuing to explore configurations of courses, areas, projects, faculty, and partnerships to gain critical experience in how to best educate new generations of data scientists"

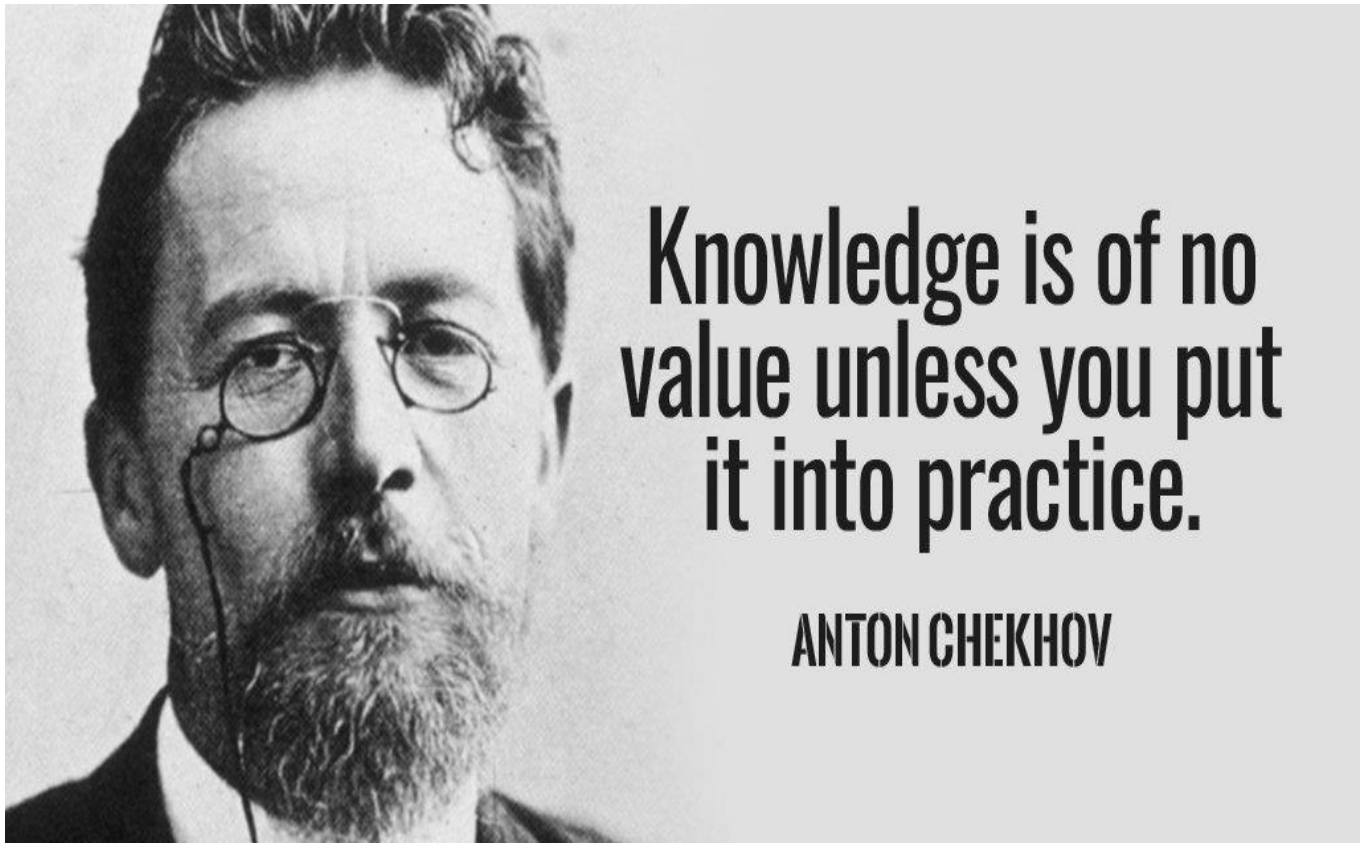
REALIZING THE POTENTIAL OF DATA SCIENCE, Final Report

from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group,

December 2016

There is a growing gap between commercial and academic research practice for data systems that needs to be addressed.

- … and it's NOT THIS GAP only ONE
- and the reasons that are much deeper...



as well as for other advanced technologies

- young-old

"Because all the peoples of the world are part of one electronically based, intercommunicating network, young people everywhere share a kind of experience that none of the elders ever had. . . . This break between generations is wholly new: it is planetary and universal".

- Cultural anthropologist Margaret Mead, 1970

Managers and executors

The research showed that students of applied mathematics programs are not interested in technical fields, but vice versa - those with predominant technical studies do not want to devote additional attention to mathematics.

Students from business faculties are the most aware of the importance of interdisciplinary and computing and mathematics in solving the complex problems of organizational systems, but in the absence of formal knowledge, they believe that someone else needs to solve these problems.

- Ivan Luković & Milan Šolaja (https://startit.rs/data-science-srbija-sta-nude-fakulteti-sta-traze-kompanije/)

The problem, in fact, is inherited from their professor

On the GitHub platform is currently registered more than 600 programs in Data Science, Big Data Analytics and related fields at over 200 universities around the World

(http://datascience.community/colleges)

- There has been a significant increase of **undergraduate programs** being conducted at research institutions and liberal arts colleges
- The vast majority of Data Science and Big Data Analytics Education are **master's degree** and certificate programs that are conducted in traditional way and/or online as distance learning.
- **PhD programs** in these areas are still relatively rare and their expansion may be expected in the near future

Data science education at the undergraduate level

Guiding Principles:

1. **Data Science as Science**
2. **Interdisciplinary Nature of Data Science**
3. **Data at the Core**
4. **Analytical (Computational and Statistical)Thinking**
5. **Mathematical Foundations**
6. **Flexibility**

- (Veaux et al.,(2017) *Curriculum guidelines for undergraduate programs in data science*

  Annual Review of Statistics and Its Application, 4(1), 15-30

Key Competencies for an undergraduate Data Science Major

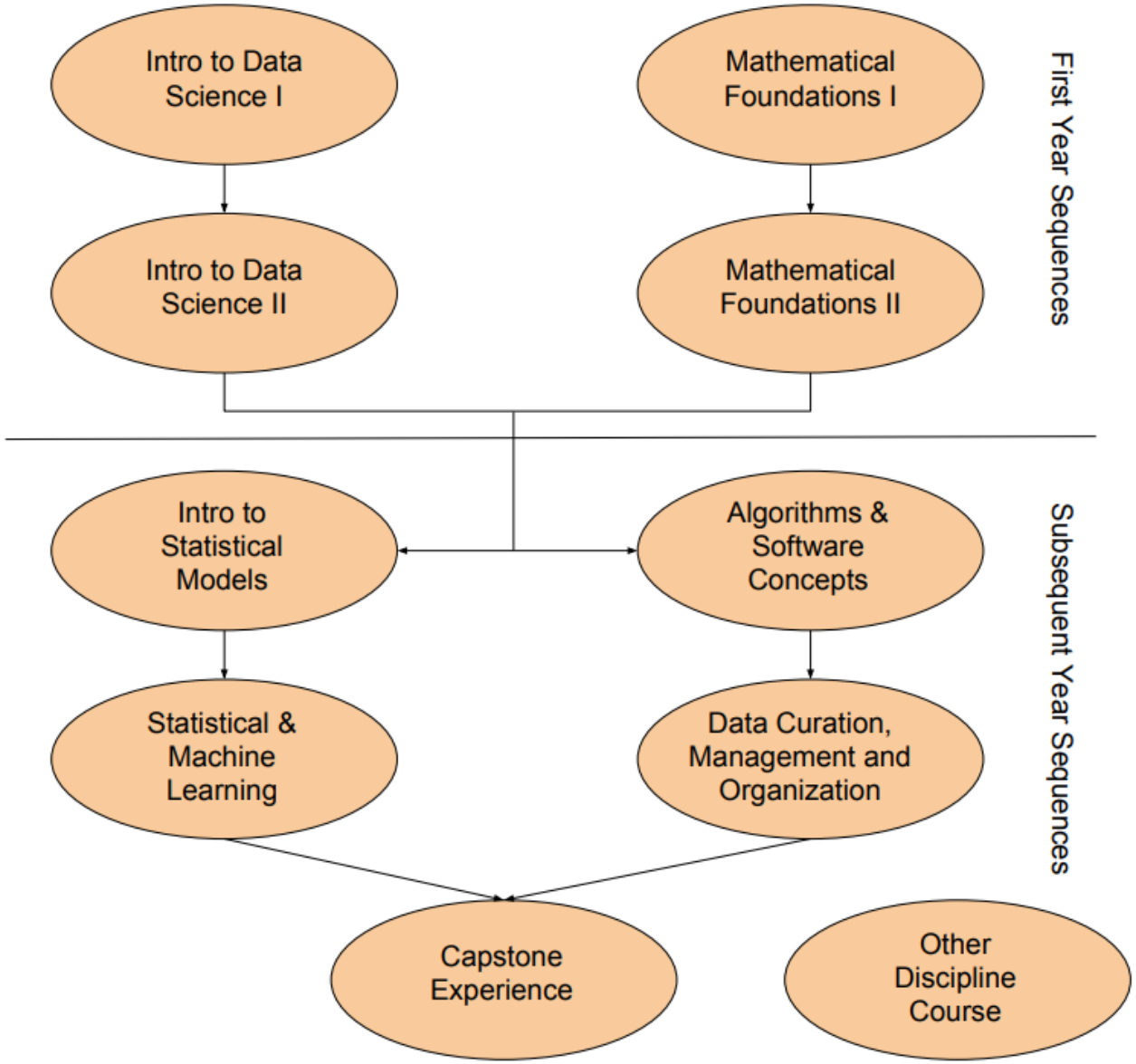
Computational and Statistical Thinking
Mathematical Foundations
Model Building and Assessment
Algorithms and Software Foundation
Data Curation
Knowledge Transference – Communication and Responsibility

Six Main Subject Areas of a Data Science Major

Data Description and Curation
Mathematical Foundations
Computational Thinking
Statistical Thinking
Data modeling
Communication, Reproducibility and Ethics

An Outline of the Data Science Major

1. Intro to Data Science
    - Intro to Data Science I
    - Intro to Data Science II
2. Mathematical Foundations
    - Mathematics for Data Science I
    - Mathematics for Data Science II
3. Computational Thinking
    - Algorithms and Software Concepts
    - Databases and Data Management
4. Statistical Thinking
    - Intro to Statistical Models
    - Statistical and Machine Learning
5. Course in an Outside Discipline
6. Capstone Course



A possible path through the major

SUMMARY POINTS

1. Data Science is a fast evolving discipline centered on the acquisition, curation and analysis of data.
2. Courses from the traditional disciplines of mathematics, statistics and computer science provide the basic infrastructure for the major at present.
3. A redesign of the curriculum, integrating the elements of mathematical foundations and computational and statistical thinking at all levels will provide a rich and effective series of courses to prepare graduates for a career in Data Science.

No single model of which department, school, or cross-unit collaboration within higher-education institutions should have the responsibility for Data Science & Big Data education and training

- Data science programs are being sited in departments and schools of:
  - Computer science,
  - Information science,
  - Statistics, and
  - Management.

- Many of the most successful, particularly at the undergraduate level, represent university wide coalitions frequently sponsored by interdisciplinary institutes, rather than by a particular department or school.

- **Computer scientists** often focus on platform and performance issues, including mining, organizing, modeling, and visualizing, as well as the mechanisms for eliciting meaning from the data through machine learning and other approaches.
- The physical processes of acquisition and instrument control are often the focus of **engineering**, or data as “dirty signals” or as control inputs for other equipment.
- **Statisticians** may focus on the mathematics of models for risk and inference.
- **Information scientists** and library scientists may focus on stewardship and preservation of data and the “back-end” of the pipeline, following acquisition, decisions, and action in the realm of publishing, archiving, and curation.

Master program Applied Mathematics – Data Science (120 ECTS)

- Two elective modules:
- Students of the **Data Analytics module** focus on the task of extracting knowledge from data, utilizing machine learning, optimization, and signal processing tools.
- Students of the **High Performance Computing module** focus on the computer engineering issues of storing, managing and manipulating large volumes of data;

- Their expertise will be on databases, high performance computing, and similar computer engineering aspects, observed through mathematical and computer science perspective.
- Students from both modules will be qualified to work in a very wide range of application domains, including business, finance, agriculture, medicine and industry.

Obligatory courses with the proposed scheduling:

		Code	MDS – obligatory courses	Hours	ECTS	Σ_{year}	Σ
First year	I	MDS01	Programming for Data Science	2+3	6	42	82
		MDS02	Stochastic Processes	2+3	6		
		MDS03	Numerical Linear Algebra 1	2+3	6		
		MDS04	Fundamentals of Numerical Optimization	2+3	6		
	II	MDS05	Graph Theory	2+3	6		
		MDS06	Pattern Recognition and Machine Learning	2+3	6		
		MDS07	Distributed Optimization with Applications	2+3	6		
Second year	III	MDS08	Network Science	2+2	5	40	
		MDS09	Large Scale Data Mining	2+2	5		
	IV	MDS10	Master thesis	-	30		

Elective courses :

	Code	MDS – elective courses *Module 1- Data Analytics*	Hours	ECTS
First year				
I	MDS11	Signals and Systems	2+3	6
	MDS12	Modelling Seminar	2+3	6
II	MDS13	Time series	2+3	6
	MDS14	Numerical Analysis	2+3	6
	MDS15	Numerical Linear Algebra 2	2+3	6
	MDS16	Introduction to Digital Signal Processing	2+3	6
	MDS17	Statistics Theory for Learning and Signal Processing	2+3	6
	MDS18	Information Theory and Networks	2+3	6
Second year				
III	MDS19	Communication and Storage Networks for Big Data	2+2	5
	MDS20	Introduction to Image Processing	2+2	5
	MDS21	Theory of Algorithms	2+2	5
	MDS22	Big Data in Medicine and Biology	2+2	5
	MDS23	Audio, Speech and Language Processing	2+2	5
	MDS24	Graphical Models and Probabilistic Inference	2+2	5
	MDS25	Research methodology	2+2	5
	MDS26	Operations research	2+2	5

	Code	MDS – elective courses *Module 2 – High Performance Computing*	Hours	ECTS
First year				
I	MDS11	Signals and Systems	2+3	6
	MDS12	Modelling Seminar	2+3	6
II	MDS13	Time series	2+3	6
	MDS14	Numerical Analysis	2+3	6
	MDS15	Numerical Linear Algebra 2	2+3	6
	MDS16	Introduction to Digital Signal Processing	2+3	6
	MDS17	Statistics Theory for Learning and Signal Processing	2+3	6
	MDS18	Information Theory and Networks	2+3	6
Second year				
III	MDS19	Communication and Storage Networks for Big Data	2+2	5
	MDS27	Databases	2+2	5
	MDS21	Theory of Algorithms	2+2	5
	MDS25	Research methodology	2+2	5
	MDS28	Software Engineering	2+2	5
	MDS29	High Performance Computing	2+2	5
	MDS30	Advanced Programming for	2+2	5
	MDS31	Data structures and Algorithms	2+2	5

Business analytics

Book of subjects

Curriculum of the study program Business analytics

- Management and organization
- Management
- Information Systems and Technologies
- Software Engineering and Computer Science
- Business analytics
- E-business and System management
- Management in public sector
- Business Administration

Study group **Operations research**

No.	Course	Semester		ECTS
		1	2	
1.	Business Analytics and Optimization	2+2		6
2.	Business Statistics	2+2		6
3.	Elective course G1	2+2		6
4.	Elective course G2	2+2		6
5.	Elective course G2	2+2		6
6.	Preparatory paper			8
7.	Internship			4
8.	Master thesis			18

Elective courses G1

- Games Theory and Business Strategies
- Risk Management
- Process Analysis and Petri Nets
- Combinatorial Optimization and Metaheuristics
- Measuring the Efficiency of Business Systems
- Advanced Planning and Scheduling
- Measuring the Preferences of Business Systems

Elective courses G2

- Games Theory and Business Strategies
- Biostatistics and Telemedicine
- Time Series Analysis and forecasting
- Data Analysis and R-software
- Measuring the Preferences of Business Systems
- Econometrics of Financial Markets
- Risk management
- Process Analysis and Petri Nets
- Combinatorial Optimization and Metaheuristics
- Measuring the Efficiency of Business Systems
- Multivariate Analysis – selected chapters
- Advanced Planning and Scheduling
- Statistics in Management – selected chapters
- Computational Statistics
- Data Mining
- Business Intelligence Systems
- Data Warehouse
- Advanced Planning in Marketing
- Simulation Models in Finance
- Supply Chain Management 2
- Numerical Methods in Finance
- Theory of Algorithms
- Mathematical Programming

Study group **Business Statistics**

No.	Course	Semester		ECTS
		1	2	
1.	Business Analytics and Optimization	2+2		6
2.	Business Statistics	2+2		6
3.	Elective course G1	2+2		6
4.	Elective course G2	2+2		6
5.	Elective course G2	2+2		6
6.	Preparatory paper			8
7.	Internship			4
8.	Master thesis			18

Elective courses G1

- Statistics in Management – selected chapters
- Biostatistics and Telemedicine
- Time Series Analysis and forecasting
- Data Analysis and R-software
- Econometrics of Financial Markets
- Multivariate Analysis – selected chapters
- Computational Statistics

tive courses G2

- Games Theory and Business Strategies
- Biostatistics and Telemedicine
- Time Series Analysis and forecasting
- Data Analysis and R-software
- Measuring the Preferences of Business Systems
- Econometrics of Financial Markets
- Risk management
- Process Analysis and Petri Nets
- Combinatorial Optimization and Metaheuristics
- Measuring the Efficiency of Business Systems
- Multivariate Analysis – selected chapters
- Advanced Planning and Scheduling
- Statistics in Management – selected chapters
- Computational Statistics
- Data Mining
- Business Intelligence Systems
- Data Warehouse
- Advanced Planning in Marketing
- Simulation Models in Finance
- Supply Chain Management 2
- Numerical Methods in Finance
- Theory of Algorithms
- Mathematical Programming

UNIVERSITY OF BELGRADE
FACULTY OF ORGANISATIONAL SCIENCES

Some basic assumptions, in order to develop discipline

Any innovative agenda in data science research and education will depend on a foundation of:

- Data Science Research and Education Infrastructure
- Access to sufficiently large and numerous datasets to illuminate and validate results
- Strategic frame
- Legislative

Recommendations

- National data science agenda
- Conglomerate
- **Private-Public-Academy partnership**

National data science agenda

- Strategy for Education Development in Serbia 2020
- Strategy and Action plan for the Scientific and Technological Development of the Republic of Serbia for 2016-2020 – “Research for Innovation”
- Open data initiative
- Open Science Platform (adopted 2018, July, 14th)
- OPEN ACCESS DATA SETS

- National Data Science Research Centre

Big Deal? Approach:

- To encourage a broader and more **holistic view of data** as integrating research opportunities across the classic sciences, engineering, and range of application domains.
- In parallel with development of data science in depth as a core component of computer science, data science should also evolve in breadth to address the needs of **domains outside computer science** -to applying data-driven strategies to individual domain research and cross domain research opportunities.
- "**Embodied intelligence**" scenarios that big data is enabling for the first time. Recent breakthroughs in a range of foundational artificial intelligence and "deep learning" technologies have made it possible to create sophisticated software artifacts that "act intelligently."
- **Bridge between commercial and academic research** practice for data systems at the edge of the state of the art

Conglomerate

- R&D organisations in Serbia within the public sector are: Serbian Academy of Sciences and Arts with its 10 scientific institutes; 28 other scientific institutes; a center of scientific excellence; 30 research institutes and 65 innovative organizations. There are seven public universities comprising 89 faculties and 10 private universities numbering 60 private faculties in Serbia, granted accreditation by the Ministry for school year 2012/2013.
- A strong data science curriculum requires faculty with appropriate expertise and engagement with the field. The pull of faculty with expertise in data science and related fields away from academia and toward industry creates a challenge for educational institutions in mounting such programs. It also presents a potential challenge to development of data science as a formal discipline
- Not only at the level of (state) universities, but all over the state. We are too small to waste resources in this area

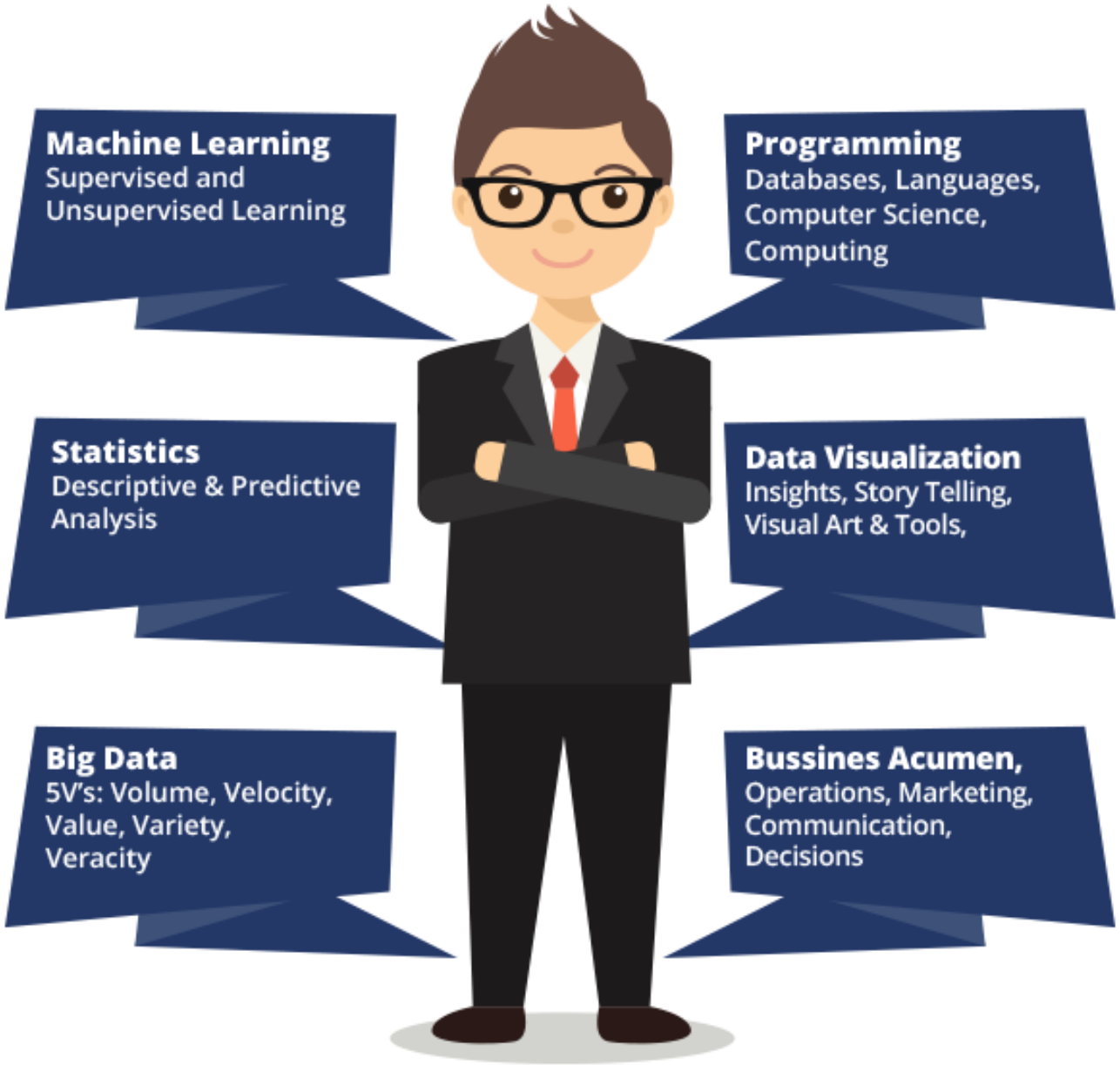
Private-Public-Academy partnership

- The “brain drain” from the research community into the private sector may be declining infrastructure-support environments, including the sparsity of large datasets and adequate infrastructure in academia that support data science research at scale.
- While this is likely good for the global economy in the near term, it is worrisome for the future of discovery-based open research, education, and training in the academic sector. In addition to the challenges of attracting sponsored research funding, another reason for “brain drain”.
- When the best infrastructure environment for cutting-edge research is consistently in the private sector, the opportunity for innovation in the public sector deteriorates – which leads to an even deeper gap, but the states can not force if the academic community is not ready to cope with the changes
- Government support for strategic and committed public-private (academic) partnerships that build adequate and representative at-scale infrastructure in the academic community for researchers can unlock innovation in academic research and ultimately support the private sector through development of a more sophisticated, educated, better-trained workforce.

Conclusion

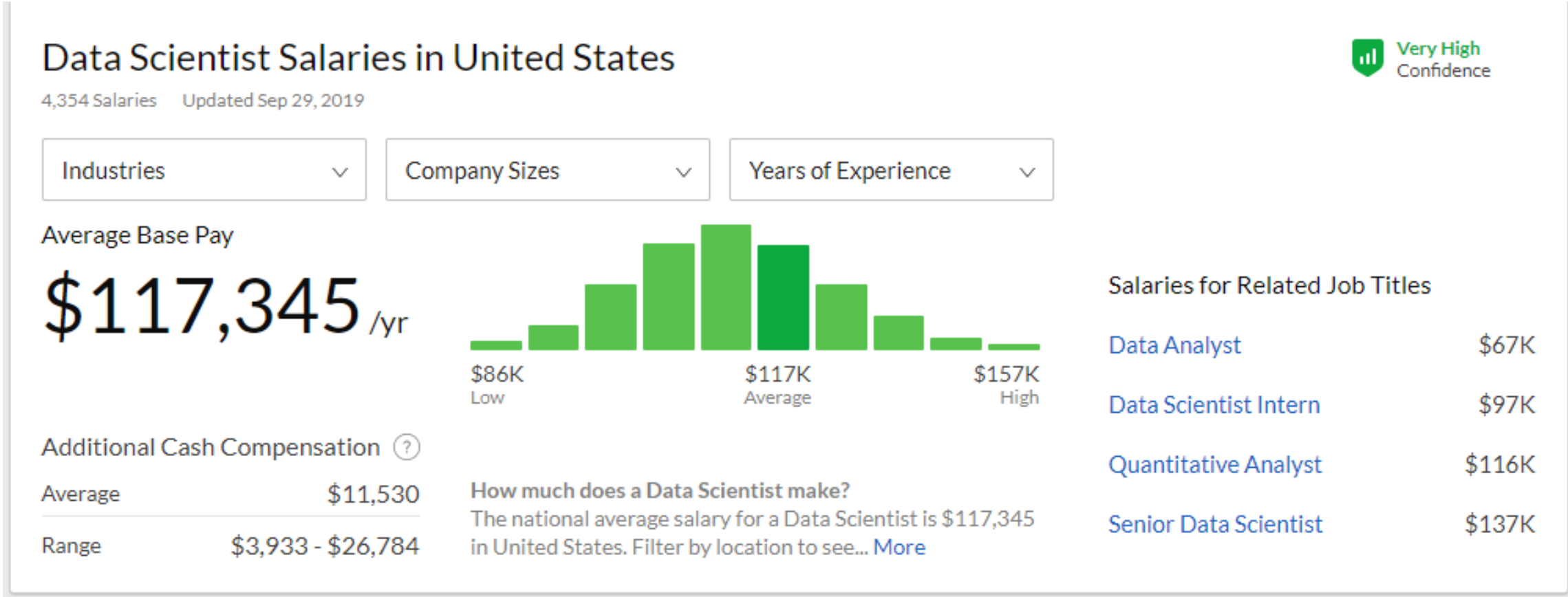
It seems that the prerequisites for the introduction of data science disciplines have been created…

Thank you



https://acadgild.com/blog/managers-data-science

…and goodbye



https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://www.techrepublic.com/article/why-data-scientist-is-the-most-promising-job-of-2019/