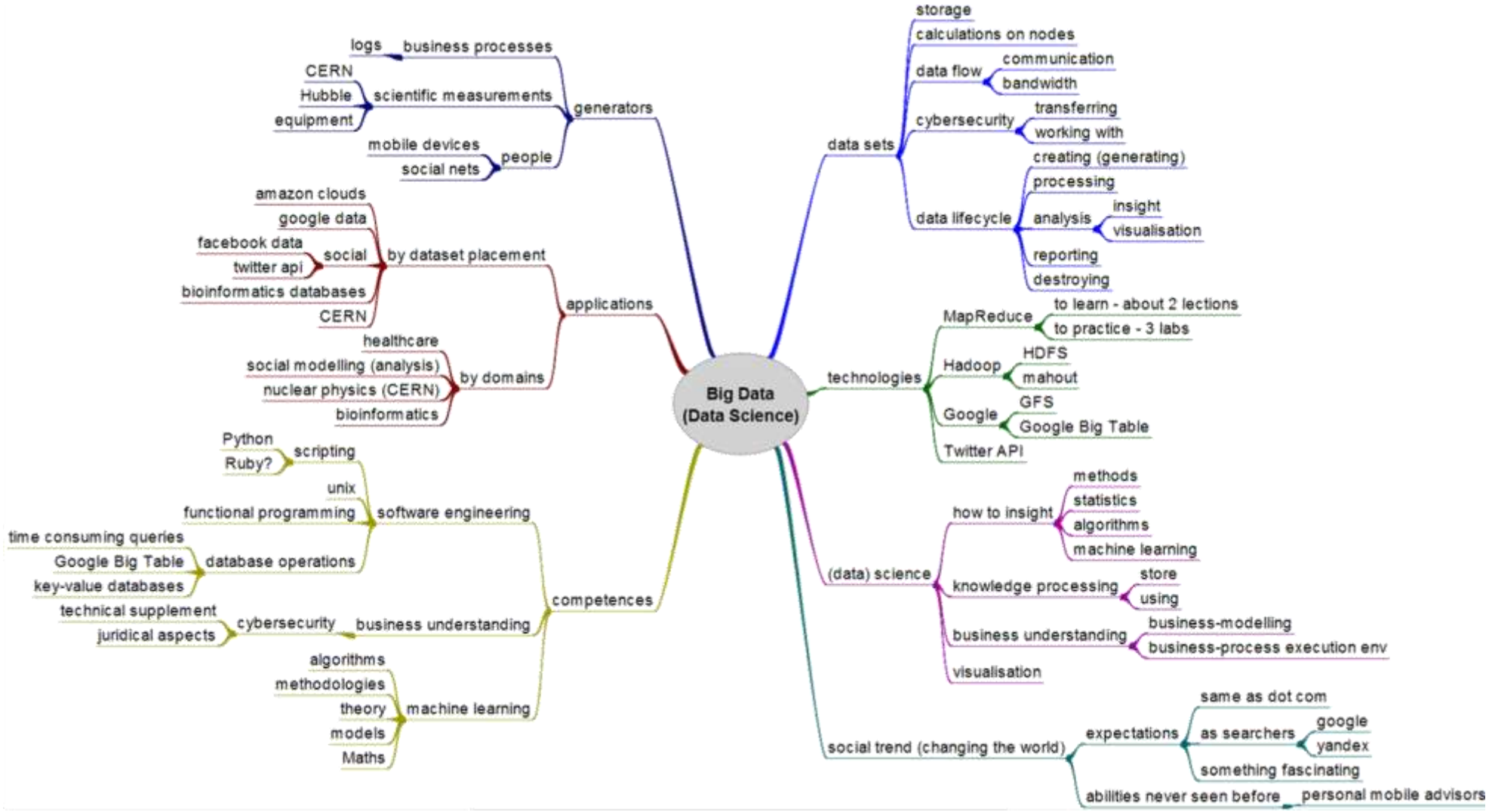
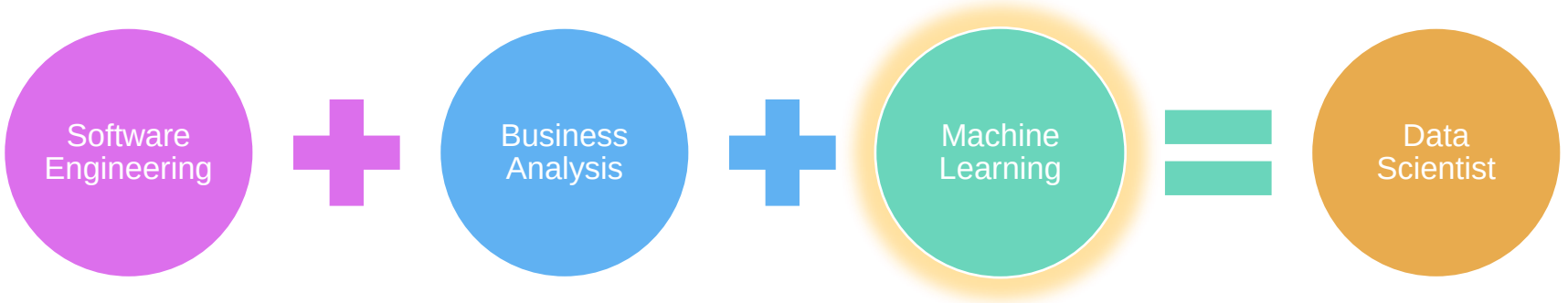
Научные вызовы больших данных

Руководитель проекта магистерской программы
«Аналитика больших массивов данных»
Старший преподаватель НГУ,
к.ф.-м.н. Павловский Евгений Николаевич
pavlovskiy@post.nsu.ru


Big Data
(Data Science)
data sets
storage
calculations on nodes
data flow
communication
bandwidth
cybersecurity
transferring
working with
data lifecycle
creating (generating)
processing
analysis
insight
visualisation
reporting
destroying
technologies
MapReduce
to learn - about 2 lections
to practice - 3 labs
Hadoop
HDFS
mahout
Google
GFS
Google Big Table
Twitter API
(data) science
how to insight
methods
statistics
algorithms
machine learning
knowledge processing
store
using
business understanding
business-modelling
business-process execution env
visualisation
social trend (changing the world)
expectations
same as dot com
as searchers
google
yandex
something fascinating
abilities never seen before
personal mobile advisors
applications
generators
logs
business processes
CERN
Hubble
equipment
scientific measurements
mobile devices
social nets
people
by dataset placement
amazon clouds
google data
facebook data
twitter api
social
bioinformatics databases
CERN
by domains
healthcare
social modelling (analysis)
nuclear physics (CERN)
bioinformatics
competences
software engineering
Python
Ruby?
scripting
unix
functional programming
database operations
time consuming queries
Google Big Table
key-value databases
business understanding
technical supplement
juridical aspects
cybersecurity
machine learning
algorithms
methodologies
theory
models
Maths

Факт

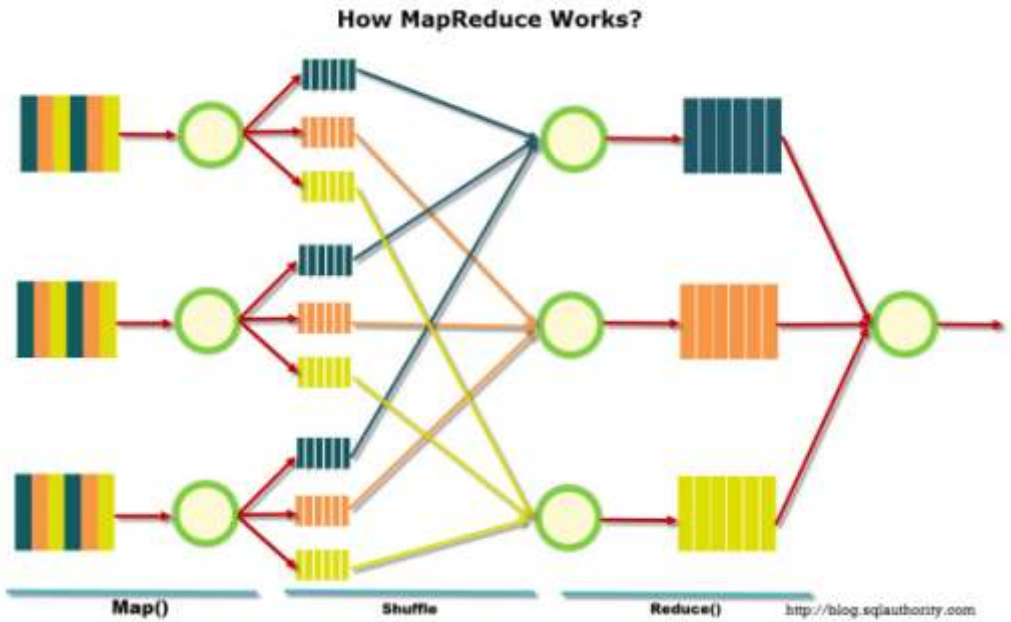
В ННЦ появился исследовательский, предпринимательский и преподавательский опыт для проведения исследований в области Data Science



Научные и технические проблемы

MapReduce

(Google, 2004)



Качество исходных данных

Проблема несоответствия цели сбора данных и цели их использования

Отсутствие должного внимания к качеству собираемых данных

Множество источников данных с неизвестной степенью истинности

Структурированность

Структурированная информация
- Таблицы СУБД с известными типами данных

Полуструктурированная информация
- XML-документы и XSD-схемы

Неструктурированная информация
- Текстовые документы
- Видеоконтент
- Аудиоконтент

Условность структурированности:
- Файл имеет структуру в рамках файловой системы
- Файл может не иметь структуры для исследователя (пока исследователь не узнает эту структуру)

Отказ от структурированности
- информационный поиск (индексы)

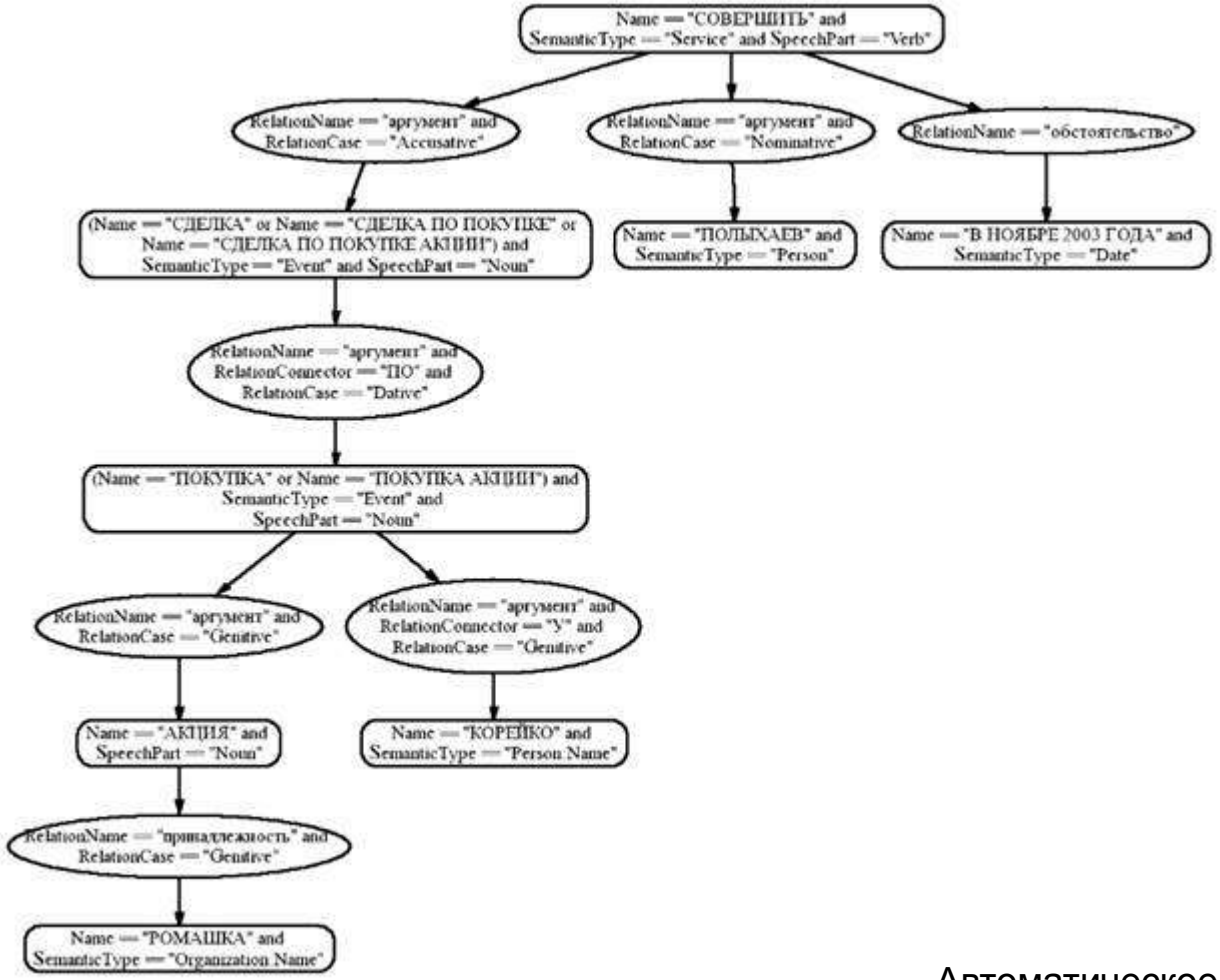
Данные – Информация – Знания

Данные – совокупность зафиксированных фактов

Информация – сведения, уменьшающие неопределённость

Знания – сведения, позволяющие действовать с прогнозируемым результатом

Мы располагаем данными, они хранятся в цифровом виде,
мы не знаем, что в них



Пример семантической сети, соответствующей предложению: В ноябре 2003 года Полыхаев совершил сделку по покупке акций ООО “Ромашка” у

Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей

А.Е. Ермаков

Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2007. – Москва, Наука, 2007

Тексты

Слова возникают не как буквы или слоги, а как обозначения смыслов сначала сенсорное представление, затем логическое.

Совокупность нейронов, связанных ассоциативно и иерархически.

Технологии направлены на возвращение к смыслам и решению попутно навязанных проблем:

1. Преобразование из двоичной информации в смысловую
2. Множественность смыслов
3. Наличие смысла в каком либо отношении

Новые постановки: (1) типы данных

Новые типы данных

- Представление текстов (нечёткие алгебраические системы*)
- Графы и графовые СУБД (соединение методов линейной алгебры и теории графов)

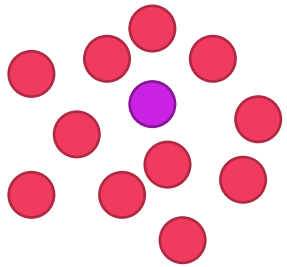
Проблема структурированности данных

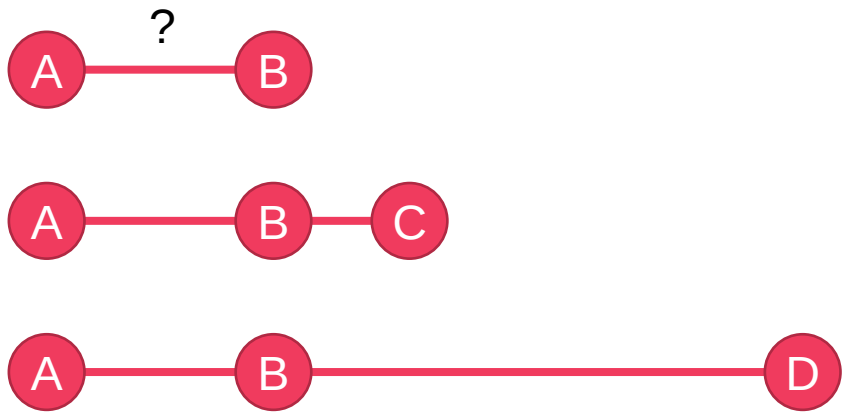
---

* Пальчунов Д.Е., Яхъяева Г.Э. Нечёткие алгебраические системы. Вестник НГУ. 2010.

Новые постановки: (2) измерительные шкалы

- Бинарные меры*
- Тернарные меры**





FRiS

---

* Пфанцагль И. Теория измерений. 1976.

** A quantitative measure of compactness and similarity in a competitive space // N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov and O. A. Kutnenko — Journal of Applied and Industrial Mathematics, 2011, Vol. 5, № 1, pp.144-154.

Jeffery I., Higgins D., Culhane A.: Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data, BMC Bioinformatics, 2006, 7:359. (http://www.biomedcentral.com/1471-2[9]5/7/359)

10 методов выбора * 4 типа реш. правил ……. 40 решений 9 задач

Задача	N0	m1/m2	max of 40	GRAD
ALL1	12625	95/33	**100.0**	**100.0**
ALL2	12625	24/101	**78.2**	**80.8**
ALL3	12625	65/35	**59.1**	**73.8**
ALL4	12625	26/67	**82.1**	**83.9**
Prostate	12625	50/53	**90.2**	**93.1**
Myeloma	12625	36/137	**82.9**	**81.4**
ALL/AML	7129	47/25	**95.9**	**100.0**
DLBCL	7129	58/19	**94.3**	**93.5**
Colon	2000	22/40	**88.6**	**89.5**
		average	**85.7**	**88.4**

Рейтинг методов выбора

Methods of feature selection	Rating
Fold change	47
Between group analysis	43
Analysis of variance (ANOVA)	43
Significance analysis of microarrays	42
Rank products	42
Welch t-statistic	39
Template matching	38
Area under the ROC curve	37
maxT	37
Empirical Bayes t-statistic	32
FRiS-GRAD	**12**

Загоруйко Н.Г.
Борисова И.А., Дюбанов В.В., Кутненко О.А., Леванов Д.А. 2013

Новые постановки: (3) выборка

Сокращение исходной выборки*

- Априорные методы (случайные выборки)
- Апостериорные методы (сокращение признакового пространства)

---

* National Research Council. 2013. *Frontiers in Massive Data Analysis.* Washington, D.C.: The National Academies Press.

Новые постановки: (4) индуктивно-дедуктивные системы

Индуктивные системы:

- Статистика
- Поиск неизвестных закономерностей
- Интеллектуальный анализ данных

Дедуктивные системы:

- Логический вывод
- Аксиоматизация
- Математическая логика

Соединение подходов даст новые возможности*

---

* Витяев Е.Е. Семантический вероятностный вывод. 2008.

Семантический вероятностный вывод

Правила вида: A & B & … & C -> D

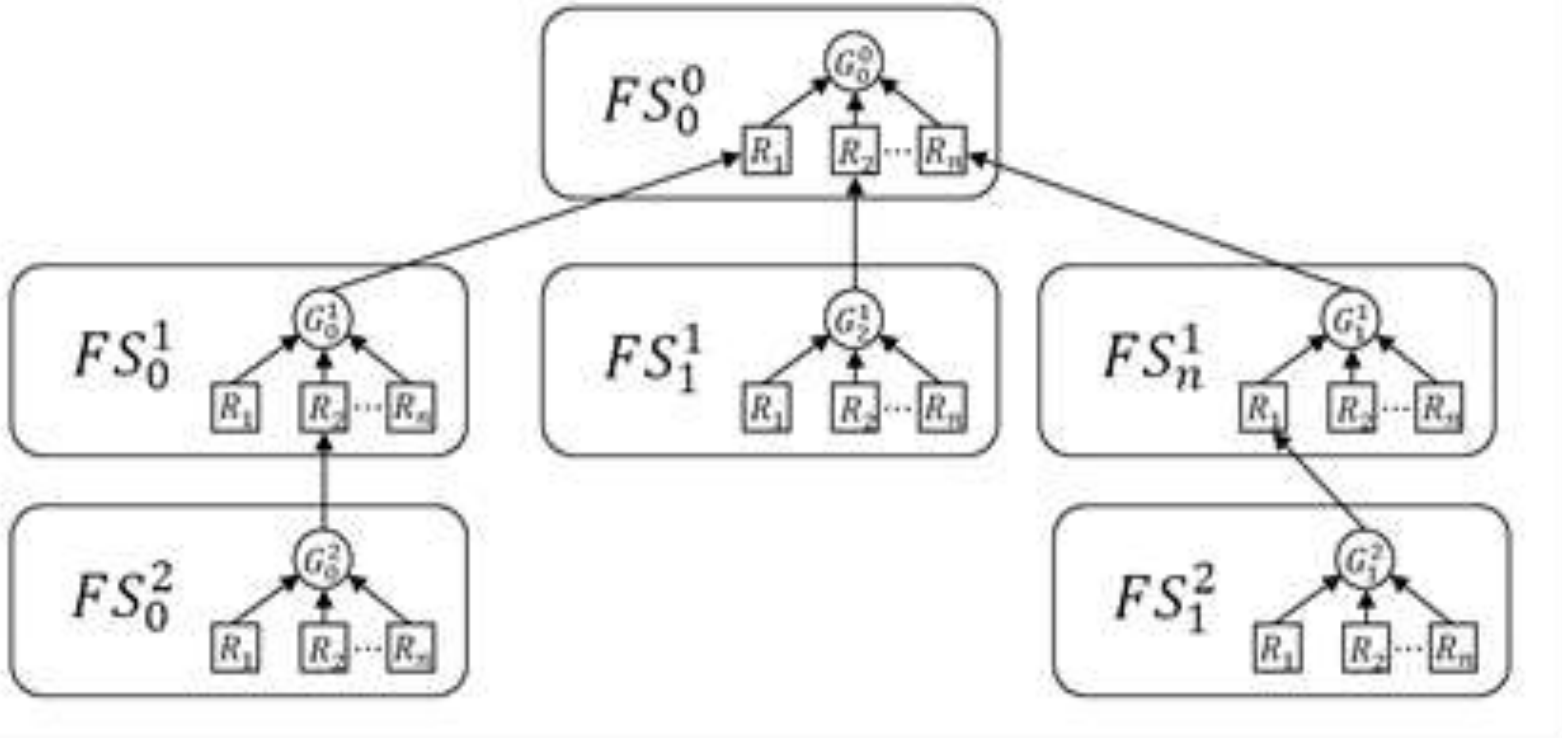
Статистика выполнения правила (условная вероятность).

На роль закономерностей претендуют максимально специфичные правила.

Логический вывод из набора правил.

Связь с физиологией:

1. теория функциональных систем Анохина.
2. Информационная теория эмоций Симонова.


FS_0^0
G_0^0
R_1 R_2 ⋯ R_n
FS_0^1
G_0^1
R_1 R_2 ⋯ R_n
FS_1^1
G_1^1
R_1 R_2 ⋯ R_n
FS_n^1
G_n^1
R_1 R_2 ⋯ R_n
FS_0^2
G_0^2
R_1 R_2 ⋯ R_n
FS_1^2
G_1^2
R_1 R_2 ⋯ R_n

Новые постановки: (5) субквадратичные алгоритмы

Большие данные требуют алгоритмов сложности не более O (N log N)

Кластеризация – одна из наиболее интересных задач

Модификация алгоритма кластеризации FRiS-Tax для работы с большими данными / Зырянов А.О. // Академический форум корпорации EMC: сборник тезисов докладов участников академической секции. 23-28 сентября 2013 г., Ялта, АРК, Украина. — Симферополь: изд-во "Ариал". — 2013. — С.17-18

Виды обучения

Онлайн-курсы	• самая широкая аудитория (школьники, разработчики, бакалавры) • средство привлечения из онлайн в офлайн
Магистратура	• вовлекаем в мобильность • Готовим для индустрии и для науки
Аспирантура	• укрепление научных школ
Дополнительное образование	• Повышение квалификации в области обработки больших данных

http://bit.ly/IntuitBDA

Первый в России онлайн-курс г Big Data Analytics

Загоруйко Николай Г[illegible]ич	Павловский Евгений Ни[illegible]ич	Борисова Ирина Ар[illegible]овна	Аникин Юрий Александрович	Зырянов Александр Олегович
				
д. т. н., академик МАИ, профессор, зав. лаб. анализа данных ИМ СО РАН	к.ф.-м.н., старший преподаватель кафедры общей информатики ФИТ НГУ	к.т.н., ассистент кафедры общей информатики ФИТ НГУ	к.т.н., преподаватель кафедры общей информатики ФИТ НГУ	Data-аналитик, ООО Экспасофт
Введение в когнитивный анализ данных	Введение в «большие данные» Области применения больших данных Основы языка R	Разработка алгоритмов на базе FRiS-функции	Обзор технологий хранения больших данных	Программирование на языке R Инструменты Data Mining

Учебный план магистратуры

	1st year. Known tech						2nd year. Innovate		
	1 Business			2 Science			3 Management		4 Thesis
	Business understanding	Ready solutions	Scripting	Access to data	Mining	Presenting	Deployment	Scaling	Final State Certification
BA	Business Analy BGoals, Commu		Business Analy Requirement	Business Analys		Presenting to SH		Marketing	
Engineering	Business Cases	Excel, Gree Deductor, V	SE: Programm (Python, R), Prot	Storage technol	Big Data Develop Environmets	Processing big data w cloud-based technolog	Virtualization an Consolidation	Couds	
Management	Project management							Product management	
Math	Knowledge presentation	Machine Lear	Operations Res		Machine Learni	Visualization methods and tools w practice	Fuzzy logic and rou computing		
Advancing	Theory of Constra		Enterpreneursh	Juridical issues	Decision making th		Enterpreneurship 2	Technology trer	
Elective business domain	Social networks analysis / Bioinformatics / Cognitive Data Mining / Instrumentation / Healthcare / Telecom								

Специализации

❤ Healthcare

medical experiment data processing; real-time patient data processing for alarming and prevention of risks, analytic modules for healthcare information systems.

Novosibirsk Research Institute of Circulation Pathology, Institute of fundamental medicine and physiology SB RAMS, Exploratory Systems (LLC.), Zdorovie Online

ⓘ Bioinformatics

gene pattern recognition, gene expression prediction

Institute of Cytology and Genetics, UniPro, Medical Genetics Technologies (LLC.), Novel Software Systems Company (LLC.), Development Group (LLC.)

Специализации

Instrumentation and scientific data processing

analyzing data from CERN, software for new electronic equipment

Budker Institute of Nuclear Physics, Uniscan (http://uniscan.biz)

Telecommunications

advertisement targeting

MTS, Eyeline Communications CIS

Social Networks

identifying social event preparation by social network activity, A/B testing, semantic analysis.

Alawar Entertainment, Futurolab LLC. (http://f-lab.pro)

Специализации

- **Cognitive Data Mining**

  development FRiS methodology of cognitive data mining for Big Data

  Sobolev Institute of Mathematics (science), Exploratory Systems, LLC. (http://xpss.ru research and development, business), Expasoft LLC. (http://expasoft.com, new algorithms)

Поддержка корпораций

EMC – Academic Alliance, курсы, продукты, продвижение, поездк

SAP – University Alliance, курсы, продукты, продвижение

Минкомсвязи РФ

ГК Ростехнологии

Индия

- Подписано соглашение с индийским университетом RGPV (250 тыс. студентов), Ноябрь 2013.
- Студенты приезжают на стажировку
- Есть один кандидат в на пост-док в Новосибирск

Публикации

1. Построение сжатого описания данных с использованием функции конкурентного сходства // Н. Г. Загоруйко, И. А. Борисова, О. А. Кутненко, В. В. Дюбанов — Сибирский журнал индустриальной математики Январь-март, 2013. Том XVI, № 1(53).
2. A construction of a compressed description of data using a function of rival similarity // N. G. Zagoruiko, I. A. Borisova, O. A. Kutnenko, V. V. Dyubanov — Journal of Applied and Industrial Mathematics, April 2013, Volume 7, Issue 2, pp 275-286.
3. Программная система, основанная на функции конкурентного сходства (проект FRiS-ОТЭКС) // Дюбанов В.В., Загоруйко Н.Г., Ижовкин И.Н., Леванов Д.А. — Интеллектуализация обработки информации: 9-я международная конференция. Республика Черногория, г. Будва, 16–22 сентября 2012 г.: Сборник докладов. – М.: Торус Пресс, 2012. – С. 21-24.
4. Использование алгоритма FRiS-GRAD для анализа активности генов при решении 9 медицинских задач // Загоруйко Н.Г., Борисова И.А., Дюбанов В.В., Кутненко О.А. — IV Международная конференция «Математическая биология и биоинформатика», Москва, 2012, с. 84-85.
5. Методы исследования операций и когнитивного анализа данных в решении задач лечебно-профилактических учреждений // Дюбанов В.В., Руднев А.С., Павловский Е.Н., Зозуля Ю.В., Самочернова А.С., Сандер Д.С. — Патология кровообращения и кардиохирургия. — 2011. — № 4. — С. 77-82.
6. A quantitative measure of compactness and similarity in a competitive space // N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov and O. A. Kutnenko — Journal of Applied and Industrial Mathematics, 2011, Vol. 5, № 1, pp.144-154.
7. Master’s Program “Big Data Analytics” In Novosibirsk State University // Pavlovskiy E.N. – International conference on Clouds, Big Data and Trust (ICCBDT- 2013), 13-15 November 2013, Bhopal, India.
8. Модификация алгоритма кластеризации FRiS Tax для работы с большими данными // Зырянов А.О. – Академический форум корпорации EMC 2013, 23-28 сентября 2013, Ялта, Украина. – С.15-16.

Евгений Павловский

Тел.: +79139117907
pavlovskiy@post.nsu.ru

ИТ-Центр, Академпарк