

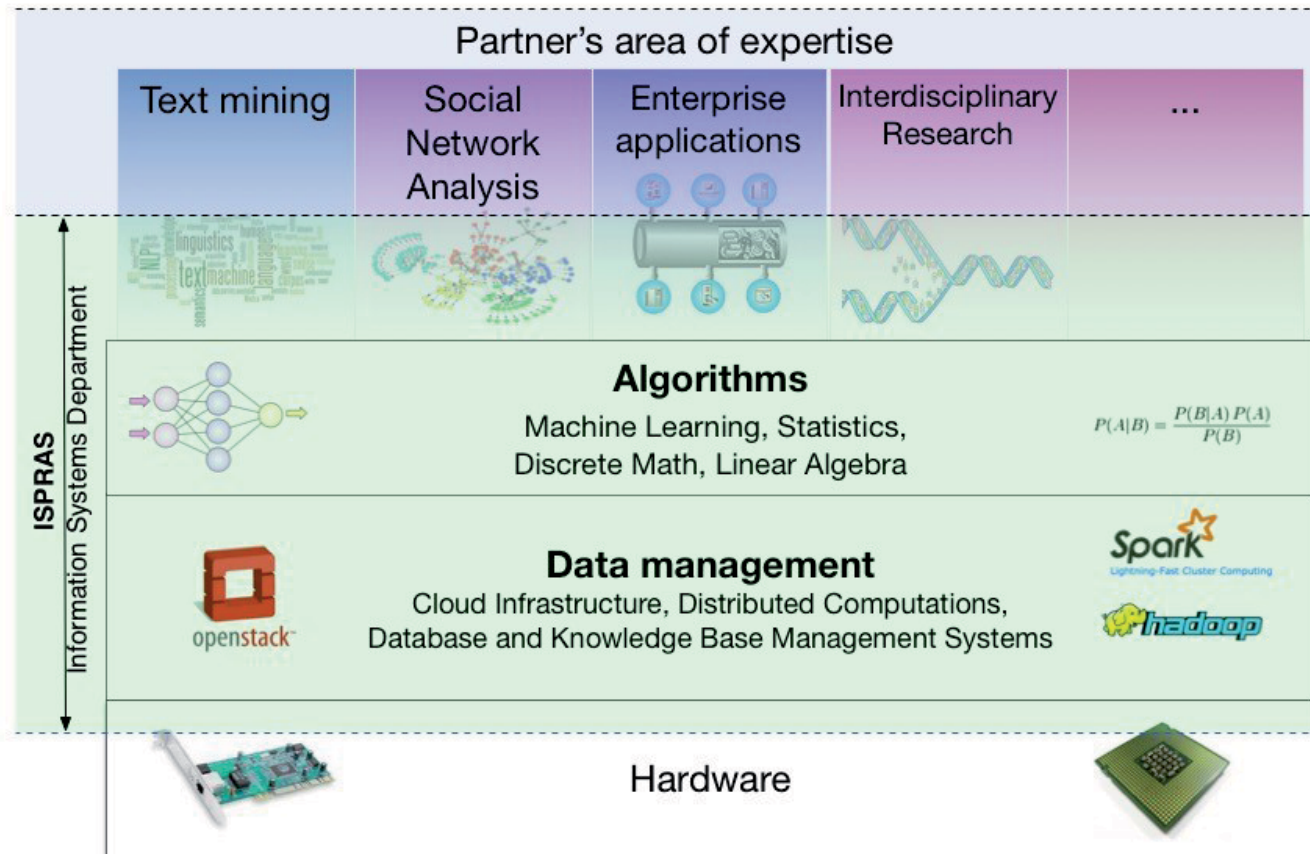
# Management of Data and Information Systems

Denis Turdakov

Head of the Department, ISP RAS



# Area of Expertise



1994

2008

2010

2012

2014

2016

2018

## Historical Background: Data Management

<b>Competence</b>	<b>Data Management</b> <ul style="list-style-type: none"><li>• DBMS development: <b>RDBMS, XML storages</b></li><li>• Query optimization: <b>SQL, XQuery, Xpath</b></li><li>• Data Integration</li></ul>
<b>Technologies</b>	ISPRAS: <ul style="list-style-type: none"><li>• <b>GNU SQL Server</b> (1994-1998)</li><li>• <b>Sedna</b> - Native XML DBMS (2004-2012)</li><li>• <b>BizQuery</b> - Data Integration Tool (1998-2003)</li></ul>



ISPI

1994

2008 2010 2012 2014 2016 2018

## Challenge: Unstructured Data Analysis

Partners searched for technologies that allows understand semantics of data and extract knowledge from unstructured data: Web sites, reports, etc.

<b>Partners</b>	 
<b>Competence</b>	<b>Data Analysis:</b> <ul style="list-style-type: none"><li>• Information Extraction, Recommender Systems</li><li>• Natural Language Processing</li><li>• Machine Learning, Probabilistic Graphical Models</li><li>• Information Retrieval, Semantic search, Exploratory Search</li></ul>
<b>Technologies</b>	ISPRAS: Texterra, BlogNoon Open Source: Apache Hadoop (first use in Russia for industrial research), Apache Lucene, Elastic Search

ISPI

1994

2008 2010 2012 2014 2016 2018

## Success Story: Texterra

<b>Challenge</b>	We need in fast tools for semantic analysis of unstructured data
<b>Solution</b>	<b>Texterra</b> -toolkit for text mining based on novel text processing methods that exploits semantics extracted from on-line sources
<b>Features</b>	<b>Fastest NLP toolkit</b> available (82Kb/sec full pipeline) <ul style="list-style-type: none"><li>• Multilingual support: <b>English, Russian</b>, <i>Korean (partially), Armenian (in progress)</i></li><li>• Fast knowledge base management engine</li><li>• Free Scalable API (<a href="https://api.ispras.ru/">https://api.ispras.ru/</a>)</li><li>• More than 20 tools, including: <i>Named Entity Recognition (F1: 0.77-0.87)</i>      <i>Sentiment Analysis (F1: 0.74-0.98)</i> <i>Term Detection (F1: 0.68-0.84)</i>                      <i>Word Sense Disambiguation (Acc: 0.76)</i></li></ul>
<b>Users</b>	HP (basic R&D), Samsung (Adding NERC, Korean Language), Internal projects
<b>Links</b>	<a href="https://texterra.ispras.ru/">https://texterra.ispras.ru/</a>


ISPI

1994

2008 2010 2012 2014 2016 2018

# Challenge: Complex Networks Analysis

Requirement from partner: Tools for analysis and modeling of large social networks (1B nodes, 100B edges).

<b>Partners</b>	
<b>Competence</b>	<p>Data acquisition: Large scale resistant Web crawling Data Analysis (with Applied Mathematics Department):</p> <ul style="list-style-type: none"> <li>• Large scale graph processing <b>1B</b> users network similar to real OSNs</li> <li>• Community detection</li> <li>• User identity resolution <b>89%</b> F1-measure in merging contact lists</li> <li>• Influence measurement <b>63%</b> precision in predicting retweets</li> </ul>
<b>Technologies</b>	<p>ISPRAS: Texterra Open Source: Apache Spark, GraphX</p> <ul style="list-style-type: none"> <li>• Random graph generation up to <b>1B</b> users, <b>+20%</b> precision in user recommendation in Hunch <b>1B</b> users network similar to real OSNs</li> <li>• Demographic Data Extraction: age, gender, religion, political views from Twitter messages in 9 languages <b>accuracy ~80%</b></li> </ul>

**ISPI**

1994

2008 2010 2012 2014 2016 2018

## Success Story: Apache Spark

<b>Challenge</b>	<ul style="list-style-type: none"><li>• We need tool to process networks with 1B vertices</li><li>• Usage of iterative algorithms</li><li>• Hadoop is too slow due to its architecture</li></ul>
<b>Solution</b>	Comprehensive overview of perspective technologies for scalable graph processing. As result we found <b>Berkeley Spark ver. 0.3</b> .
<b>Our experience</b>	<ul style="list-style-type: none"><li>• First in Russia who successfully used Spark for R&amp;D.</li><li>• <b>1B vertices and 100B edges in ~20 hours on 100 4xLarge VM in AWS</b></li><li>• Committed 3 patches</li><li>• New partitioning <b>&gt;50%</b> decrease in network traffic</li><li>• Integration Apache Spark and OpenStack (tool in Spark's main site)</li><li>• Developed Spark package extending <b>MLLib with PU-learning</b></li></ul>
<b>Links</b>	<a href="http://www.ispras.ru/en/technologies/apache_spark/">http://www.ispras.ru/en/technologies/apache_spark/</a>

ISPI




1994

2008 2010 2012 2014 2016 2018

## Challenge: Mobile Networks Analysis

Development of algorithms for large mobile networks analysis and modeling. Main specific: directed and weighted graphs.


<b>Partners</b>	 HUAWEI
<b>Competence</b>	Data Analysis: <ul style="list-style-type: none"><li>• Large scale graph processing</li><li>• Community detection</li><li>• Random graph generation</li><li>• <b>Deep Learning</b></li></ul>
<b>Results</b>	<ul style="list-style-type: none"><li>• Spark-based algorithms for overlapping community detection in directed weighted graphs</li><li>• Paper with results got <b>Best Student Paper Award</b> on ECML-PKDD 2017</li></ul>

ISPI

1994

2008 2010 2012 2014 2016 2018

# Success Story: TALISMAN

<b>Challenge</b>	Monitoring and Analysis of information flows in social media
<b>Partners</b>	
<b>Solution</b>	<p>Understanding (Russian companies) (slang, hashtags)</p> <ul style="list-style-type: none"> <li>• information flows analysis</li> <li>• Reconstruction of user's profile in the case of lack or distort data: F1 for gender <b>0.93</b>, age <b>0.73</b>, residence <b>0.83</b>, family status <b>0.76</b>, education <b>0.76</b></li> <li>• Influence bot detection <b>F1: 0.95</b></li> <li>• Object level sentiment analysis (integration with Texterra)</li> <li>• Scalable architecture (processing of News sites, VKontakte and LiveJournal): <b>~4 GB texts every day</b> on 10 VM (32G RAM, 4 cores) in ISPRAS cloud</li> </ul>


**ISPI**

1994

2008 2010 2012 2014 2016 2018

## Challenge: Resource Optimization

Big Data Open Lab: own small data center used for experiments and projects


<b>Partners</b>	
<b>Competence</b>	 Infrastructure Construction Computing Distributed Storage Engines Virtual Clustering, Big Data Distributed Systems
<b>Technologies</b>	<b>OpenStack</b> (with Compiler technologies department) <ul style="list-style-type: none"><li>• 9 reviews</li><li>• 3 discovered (and fixed with our help) bugs</li><li>• New functionality (in mainline and releases): 6 commits</li></ul> <b>Docker, Ansible</b>

ISPI

1994

2008 2010 2012 2014 2016 2018

## Success Story: GridGain / Ignite

<b>Challenge</b>	Find perspective technology for core-banking
<b>Partner</b>	 <b>SBERBANK</b>
<b>Solution</b>	<p>We developed a framework for comparison In-memory data grid systems (IMDG) based on containers (Docker) in order to reduce virtualization overhead.</p> <p>Support independent evaluation of different solutions: eXtreme Scale, <b>Hazelcast</b>, <b>GridGain</b>, <b>Infinispan</b>, Coherence, Tarantool, Terracotta, HANA, VoltDB, GemFire, XAP</p>
<b>Result</b>	<ul style="list-style-type: none"><li>• GridGain showed best results</li><li>• Sberbank became largest customer of GridGain company (2015)</li><li>• GridGain open sourced core of solution as <b>Apache Ignite</b> project</li></ul>

ISPI