Summary of results and experience in Data Science

Alexander P. Ryjov
Associated Professor, Chair of Mathematical Foundations of Intelligent Systems, Department of Mechanics and Mathematics, Lomonosov Moscow State University,
Professor and Head, Chair of Business Processes Management Systems, IT Business School, Russian Presidential Academy of National Economy and Public Administration.

1. Systems for evaluation and monitoring of complex processes. Information monitoring task includes evaluation of current status of some process and modeling of possible ways of its development based on all available information (structured, non-structured, weak-structured).

- system for monitoring and evaluation of state's nuclear activities (department of safeguards, IAEA)
- system for evaluation and monitoring of risks of cardiovascular disease (Center of Preventive Medicine of Ministry of Health of Russia)
- system for evaluation and monitoring of microelectronics design (Cadence Design Systems, Inc.)

More detailed presentation is available here:
http://intsys.msu.ru/en/staff/ryzhov/Systems%20for%20evaluation%20and%20monitoring%20of%20complex%20processes.pdf

2. Retail

2.1. Profiles

Input: receipts + database of discount program (personal data)
Question: who are our high profitable customers?
Solution: split (PROFIT) for categories (less than 8200; 8200 – 23300; 23300 – 60500; more then 60500). Build profile for "more than 60500"

	A	B	C	D	E	F	G
1	Если	То	Поддерж	Доля	Интерес	Длина	
4	if PROFIT is больше 60500 (max 351100)	then => nBUY is больше 100 (max 3800)	25%	87%	3.5	1	
6	if PROFIT is больше 60500 (max 351100)	then => AGE is до 30	25%	38%	3.04	1	
19	if PROFIT is больше 60500 (max 351100)	then => SUM is меньше 100	25%	68%	2.58	1	
23	if PROFIT is больше 60500 (max 351100)	then => nGOODS is [1 - 2]	25%	63%	2.32	1	
40	if PROFIT is больше 60500 (max 351100)	then => SEX is 2	25%	75%	1.6	1	
104	if PROFIT is больше 60500 (max 351100)	then => TIME is [7-13]	25%	45%	1.25	1	
125	if PROFIT is больше 60500 (max 351100)	then => TYPECARD is 10%	25%	100%	1.2	1	
223	if PROFIT is больше 60500 (max 351100)	then => DAY is конец месяца	25%	39%	1.12	1	
224	if PROFIT is больше 60500 (max 351100)	then => MONTH is апрель	25%	29%	1.12	1	
307	if PROFIT is больше 60500 (max 351100)	then => DAY is пн.-чт.	25%	61%	1.08	1	
308	if PROFIT is больше 60500 (max 351100)	then => MONTH is весна	25%	57%	1.08	1	
309	if PROFIT is больше 60500 (max 351100)	then => MONTH is февраль	25%	26%	1.08	1	
369	if PROFIT is больше 60500 (max 351100)	then => MONTH is март	25%	28%	1.05	1	
512	if PROFIT is больше 60500 (max 351100)	then => DAY is середина месяца	25%	34%	1.02	1	
864							
865							

2.2. Customers' behavior

Input: receipts for 1 year
Question: which goods are good for selling in particular time (of the year/ months/ weeks/ days)?
Solution: split goods for categories, time for periods
Results:

- customers with average check prefer to buy #10 at summer;
- these customers prefer to buy ## 5, 7, 9 in conjunction at winter

	A	B	C	D	E	F	G	H
2661	if SUM is [1000 - 40000] and MONTH is весна	then => 10	13%	2.11	47%	2	1	
2662	if SUM is [1000 - 40000] and MONTH is зима	then => 05 07 09	11%	2.11	63%	2	3	
2663	if MONTH is весна and nGOODS is [14- 105]	then => 13	13%	2.11	54%	2	1	
2664	if SEX is 1 and nGOODS is [14- 105]	then => 01 07	11%	2.1	27%	2	2	
2665	if SEX is 2 and SUM is [1000 - 40000]	then => 02 08	11%	2.1	35%	2	2	
2666	if SUM is [1000 - 40000] and MONTH is весна	then => 03 09	13%	2.1	31%	2	2	
2667	if SUM is [1000 - 40000] and TYPECARD is 10%	then => 10	23%	2.1	47%	2	1	
2668	if SUM is [1000 - 40000] and TYPECARD is 10%	then => 05 02 09	23%	2.1	47%	2	3	
2669	if SUM is [1000 - 40000] and TYPECARD is 10%	then => 05 07 09	23%	2.1	63%	2	3	
2670	if SEX is 2 and SUM is [1000 - 40000]	then => 07 08 09	11%	2.09	42%	2	3	
2671	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 05 03	12%	2.09	30%	2	2	
2672	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 02 08	12%	2.09	35%	2	2	
2673	if SUM is [1000 - 40000] and MONTH is весна	then => 05 07 09	13%	2.09	63%	2	3	
2674	if SUM is [1000 - 40000] and MONTH is зима	then => 05 02 09	11%	2.09	47%	2	3	
2675	if MONTH is зима and nGOODS is [14- 105]	then => 04	11%	2.09	60%	2	1	
2676	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 05 07 09	12%	2.08	63%	2	3	
2677	if SUM is [1000 - 40000] and MONTH is весна	then => 07 08	13%	2.08	48%	2	2	
2678	if SUM is [1000 - 40000] and MONTH is зима	then => 10	11%	2.08	47%	2	1	
2679	if DAY is пн.-чт. and nGOODS is [14- 105]	then => 10	13%	2.08	47%	2	1	
2680	if SEX is 2 and SUM is [1000 - 40000]	then => 04 07	11%	2.07	42%	2	2	
2681	if SUM is [1000 - 40000] and MONTH is весна	then => 02 08	13%	2.07	35%	2	2	
2682	if SUM is [1000 - 40000] and MONTH is весна	then => 05 02 09	13%	2.07	46%	2	3	
2683	if SEX is 2 and SUM is [1000 - 40000]	then => 05 02 09	11%	2.06	46%	2	3	
2684	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 10	12%	2.06	46%	2	1	
2685	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 03 09	12%	2.06	31%	2	2	
2686	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 02 04	12%	2.06	30%	2	2	
2687	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 05 04 09	12%	2.06	40%	2	3	
2688	if SEX is 2 and SUM is [1000 - 40000]	then => 05 03 09	11%	2.05	26%	2	3	
2689	if SEX is 2 and SUM is [1000 - 40000]	then => 10	11%	2.04	46%	2	1	
2690	if SUM is [1000 - 40000] and DAY is пн.-чт.	then => 06	12%	2.04	46%	2	1	
2691	if MONTH is зима and nGOODS is [14- 105]	then => 13	11%	2.04	52%	2	1	
2692	if MONTH is зима and nGOODS is [14- 105]	then => 10	11%	2.04	46%	2	1	

2.3. Retail/Cross-selling

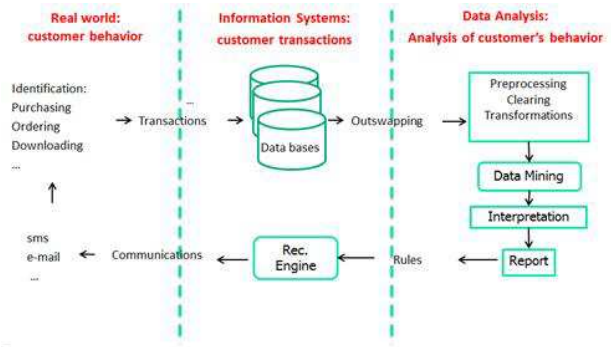
Input: receipts for 1 year
Question: which goods are most likely to be bought together?
Solution: split goods for categories.
Results:

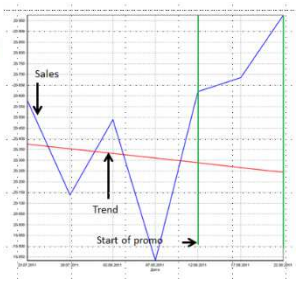
- goods #14 are bought together with #5, #4, #7, # 9
- if we add #6 or #13, we can increase the sales

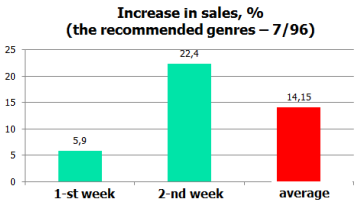
	A	B	C	D	E	F	G	H
1	Если	То	Поддержка	Интерес	Доля	Длина	Длина следствия	
2	if GOODS is [14]	then => 05 06 04 07 09	24%	3.64	26%	1	5	
3	if GOODS is [14]	then => 05 13 04 07 09	24%	3.62	26%	1	5	
4	if GOODS is [14]	then => 05 13 02 07 09	24%	3.59	28%	1	5	
5	if GOODS is [14]	then => 05 13 07 08 09	24%	3.59	27%	1	5	
6	if GOODS is [14]	then => 05 06 02 07 09	24%	3.58	28%	1	5	
7	if GOODS is [14]	then => 05 07 08 10 09	24%	3.57	26%	1	5	
8	if GOODS is [14]	then => 05 06 07 08 09	24%	3.56	28%	1	5	
9	if GOODS is [14]	then => 05 06 04 07	24%	3.51	27%	1	4	
10	if GOODS is [14]	then => 05 13 04 07	24%	3.49	28%	1	4	
11	if GOODS is [14]	then => 05 13 07 08	24%	3.49	28%	1	4	
12	if GOODS is [14]	then => 06 04 07 09	24%	3.49	28%	1	4	
13	if GOODS is [14]	then => 13 04 07 09	24%	3.49	28%	1	4	
14	if GOODS is [14]	then => 05 02 04 07 09	24%	3.48	31%	1	5	
15	if GOODS is [14]	then => 05 02 04 08	24%	3.47	25%	1	4	
16	if GOODS is [14]	then => 05 04 07 08 09	24%	3.47	33%	1	5	
17	if GOODS is [14]	then => 06 07 08 09	24%	3.46	30%	1	4	
18	if GOODS is [14]	then => 13 07 08 09	24%	3.46	29%	1	4	
19	if GOODS is [14]	then => 05 06 07 08	24%	3.45	29%	1	4	
20	if GOODS is [14]	then => 05 13 02 07	24%	3.45	30%	1	4	
21	if GOODS is [14]	then => 05 13 04 09	24%	3.45	29%	1	4	
22	if GOODS is [14]	then => 05 06 04 09	24%	3.44	29%	1	4	
23	if GOODS is [14]	then => 05 02 07 10	24%	3.44	27%	1	4	
24	if GOODS is [14]	then => 05 07 08 10	24%	3.44	28%	1	4	
25	if GOODS is [14]	then => 06 02 07 09	24%	3.43	29%	1	4	
26	if GOODS is [14]	then => 05 02 07 08 09	24%	3.43	35%	1	5	
27	if GOODS is [14]	then => 05 06 02 07	24%	3.42	29%	1	4	
28	if GOODS is [14]	then => 13 02 07 09	24%	3.42	30%	1	4	
29	if GOODS is [14]	then => 02 07 10 09	24%	3.42	27%	1	4	
30	if GOODS is [14]	then => 07 08 10 09	24%	3.42	28%	1	4	
31	if GOODS is [14]	then => 04 07 10	24%	3.39	26%	1	3	
32	if GOODS is [14]	then => 05 04 07 08	24%	3.39	34%	1	4	
33	if GOODS is [14]	then => 06 13 07	24%	3.38	25%	1	3	
34	if GOODS is [14]	then => 05 02 04 07	24%	3.38	33%	1	4	
35	if GOODS is [14]	then => 02 04 08 09	24%	3.38	26%	1	4	
36	if GOODS is [14]	then => 05 13 08 09	24%	3.36	31%	1	4	

3. Telecom

I used similar approaches for telecom (mobile content selling). Architecture for big data-based recommendation engine and results are below:



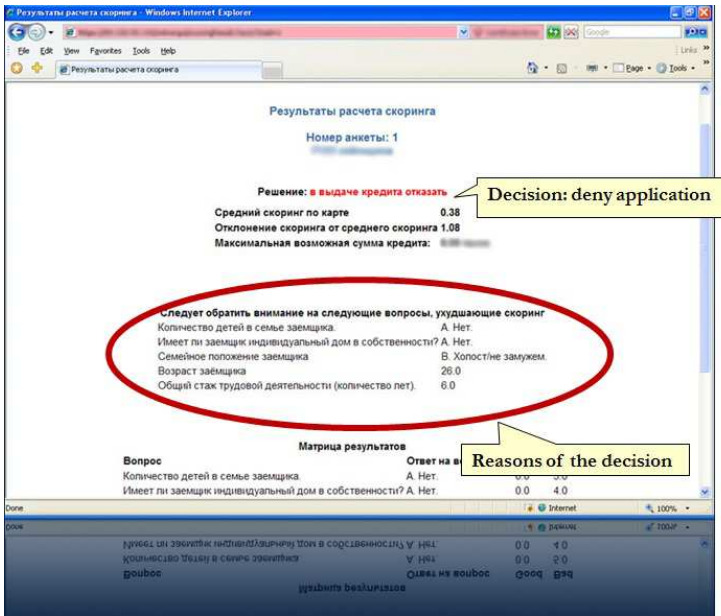




4. Banks: credit scoring system.

Business goal: objectively assess loan applicant's credit risks and decide whether to grant a loan or not.
Technology used in the product: data mining, associative rules induction.
Product main features:

- Automatic analysis of existing credit histories along with application forms of current borrowers. Identification of common



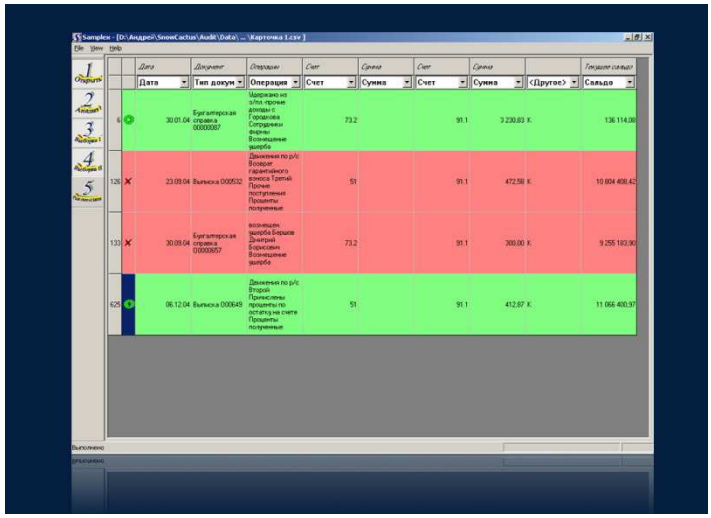
characteristics and building profiles of “good” and “bad” borrowers.

- Instant assessment of loan applicant’s credit risks and recommendation on granting a loan.
- Full integration with banking software.
- Quick integration: ready to use in 2-3 months; 6-9 months for full integration.
- Ease of use: operators do not need to have any science-specific knowledge.
- Reasoning of recommended decision: why should we deny application?

Results:

- Client using Score reports that bank’s share of bad loans is 2 times less than market average.

5. Finance: Suspicious transactions detection



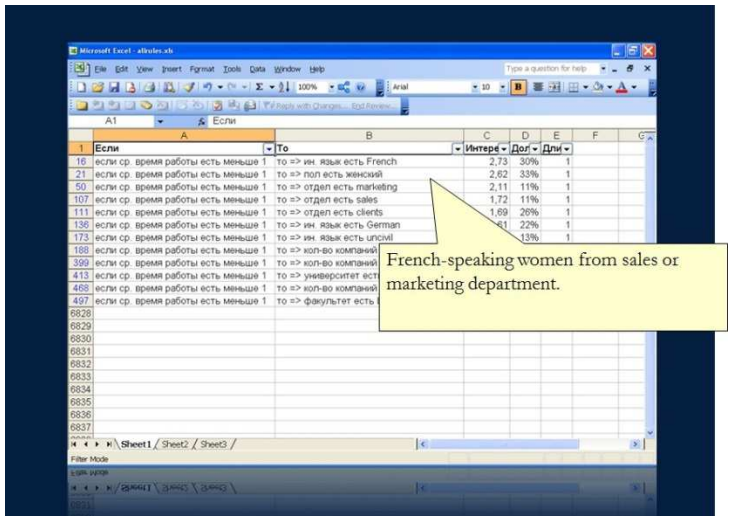
Target group: companies dealing with detection of suspicious or fraudulent transactions (auditors, banks, telecoms); companies with internal audit departments.
Business goal: to efficiently reveal suspicious transactions with greater accuracy and less time.
Technology used in the product: neural networks, cluster analysis.
Product main features:

- Automatic detection of non-typical (suspicious) transactions for further investigation.
- Automatic detection of transactions similar to fraudulent, specified by the user.

Results:

- Tests showed 7 times more accurate detection of suspicious transactions than currently widespread method.

6. HR: Evaluation of job applicants.



Target group: companies with 200+ employees or companies large turnover of certain types of employees (call centers, banks, shops).
Business goal: discover what makes best employees best; estimate job applicant’s potential loyalty.
Method: analysis of company employees’ resumes.
Technology we use: data mining, associative rules induction.
Tasks we solve:

- Structure and analyze company employees resumes, combine with performance data if available.
- Discover what is in common for top performing employees.
- Discover indicators of loyalty/unloyalty.

- Build profiles of different groups of employees (e.g. top-performers, most loyal employees, graduates of specific university, etc.)

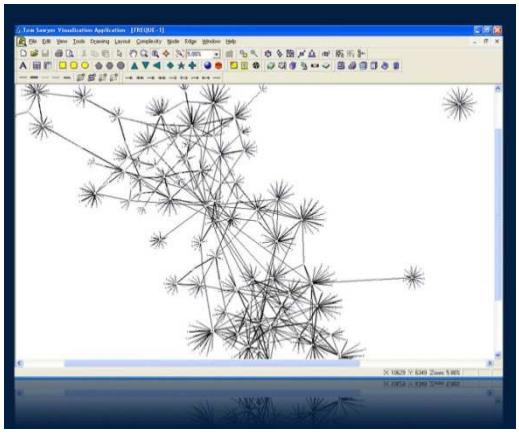
Result: interactive report with employees profiles and loyalty indicators.

7. HR: Improving employees' creativity and communications

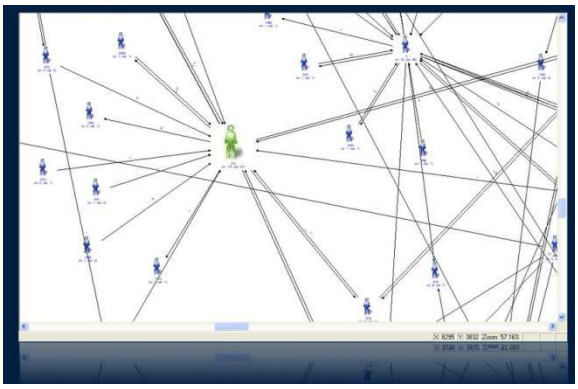
Target group: companies with 100+ employees.
Business goal: to boost employees' creativity, invention of new ideas and products.
Method: improving internal communication. The more people communicate the more creative they are.
Technology we use: social networks analysis.
Tasks we solve:

- Identification of key experts in the company. Who people go to for an advice?
- Identification of initiators. Who starts spreading new ideas and news?
- Identification of "bridges" between communities. Who connects departments?
- What-if analysis. What will happen if ... (key employee leaves, retires, gets sick; connection between teams breaks)?
- Recommendations: how to increase communications sustainability and knowledge sharing between departments?

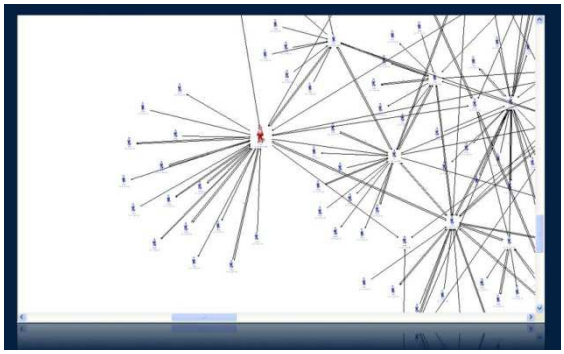
Employees communicate and share.
Here is how:



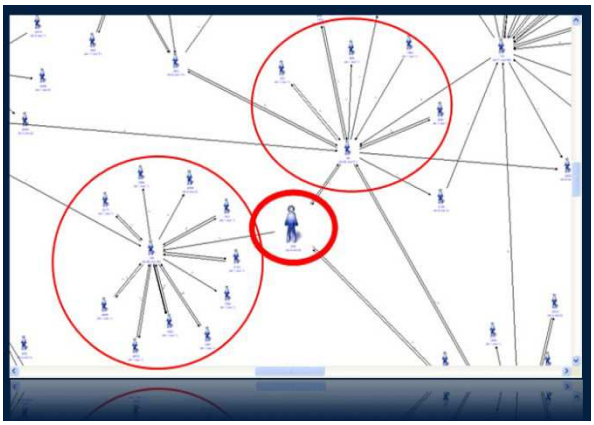
Who is an expert?
Who people consult with?



Who spreads the knowledge?



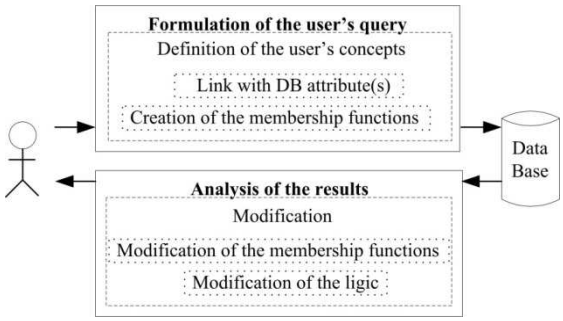
Who is the bridge between groups?



8. Large-scale databases: Adaptive semantic layer.

Adaptive semantic layer for large-scale databases allow to effectively handle a large amount of information. This effect is reached by providing an opportunity to search information on the basis of generalized concepts, or in other words, linguistic descriptions. These concepts are formulated by the user in natural language, and modelled by fuzzy sets, defined on the universe of the significances of the characteristics of the data base objects. After adjustment of user's concepts based on search results, we have "personalized semantics" for all terms which particular person uses for communications with data base or social networks (for example, "young person" will be different for teenager and for old person; "good restaurant" will be different for people with different income,
age, etc.

The structure of an adaptive semantic layer is shown here:



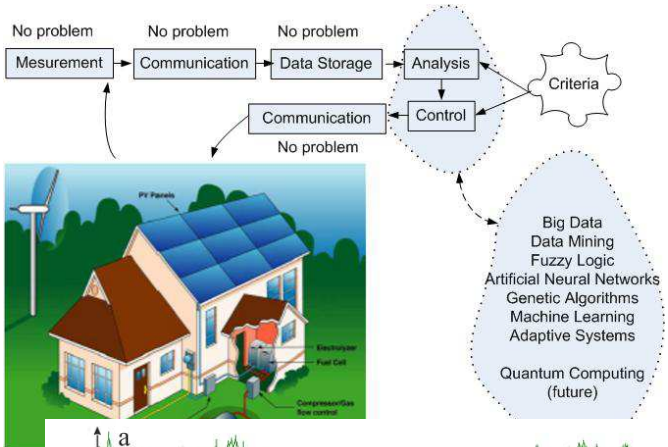
Based on theoretical results (section 1), we can develop *optimal* layer which allows:

- define user's concepts;
- search an information by these concepts;
- adjustment of user's concepts based on search results (GA-based tuning of membership functions and logic).

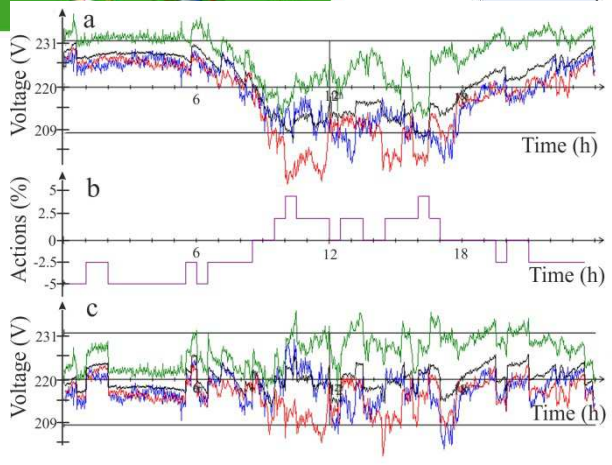
References:
Lyapin B. , Ryjov A. A Fuzzy Linguistic Interface for Data Bases in Nuclear Safety Problems. *Fuzzy Logic and Intelligent Technologies in Nuclear Science*. Proceedings of the 1st International FLINS Workshop, Mol, Belgium, September 14-16, 1994. Edited by Da Ruan, Pierre D'hondt, Paul Govaerts, Etienne E. Kerre, World Scientific. p. 212-215.
Alexander Ryjov. Personalization of Social Networks: Adaptive Semantic Layer Approach. In: *Social Networks: A Framework of Computational Intelligence*. Ed. by Witold Pedrycz and Shyi-Ming Chen. Springer Verlag, 2013 (will be published soon)

9. Energy: Smart Grid

For smart grid we can generate and use a huge amount of information from smart meters and other measurement devices. I have experience in usage customer's data for optimization consumption and energy quality. We use data mining for extracting patterns of customer's behavior from amount of data; fuzzy logic, artificial neural networks and genetic algorithms (soft computing approach) for development of monitoring and control systems; machine learning and adaptive systems approaches for optimization monitoring and control systems.



Mini-case: (Regional Energy Co.)
Problem definition: Local electric power substation has a retreating feeder and the meter installed on it. The switching station has a unit step voltage control under load with steps (0%, ± 2.5%, ± 5%, ±



7.5%, ± 10%). The task is to maintain the voltage deviation at the substation buses at a given level ± 5%. Data can be obtained from the meter: the current value of the phase voltages and currents. Quality measure: (total time period when the voltage deviation are out of level ± 5% without control)/(total time period when the voltage deviation are out of level ± 5% with control).
Results: up to 10 times quality increasing on real data (on the figure: real data (up), control, results (down)).