DataSys 2017, June 25 - 29, 2017 - Venice, Italy

--- Keynote Talk ---
Applications of Techno-social Systems in Economy and Governance

Alexander Troussov, Ph.D.

Mathematical Methods for Social Network Mining Laboratory,
The Russian Presidential Academy
of National Economy and Public Administration.

About AT

- 2014-Current : Director of the International Research Laboratory for Mathematical Methods for Social Network Mining at the Russian Presidential Academy of National Economy and Public Administration.
- 2000-2013: IBM Ireland Center for Advanced Studies Chief Scientist
  The Architect of IBM LanguageWare group
  (LanguageWare - the IBM suite for text analytics);
- Before joining IBM:
  - National Geophysical Data Center, Boulder, CO, USA - Visiting scientist
    - Fuzzy logic based search engine for search in large databases when exact parameters of search are hard to define
  - Observatoire de la Côte d'Azur, Nice, France – Visiting scientist
    - Numerical simulation in stochastic physics
  - Institute of Physics of the Earth (Russian Academy of Sciences) and the International Institute for Earthquake Prediction Theory and Mathematical Geophysics, Moscow, Russia – Lead Researcher
    - R&D in geophysics and geoinformatics
- Holds PhD in mathematics from Moscow State University

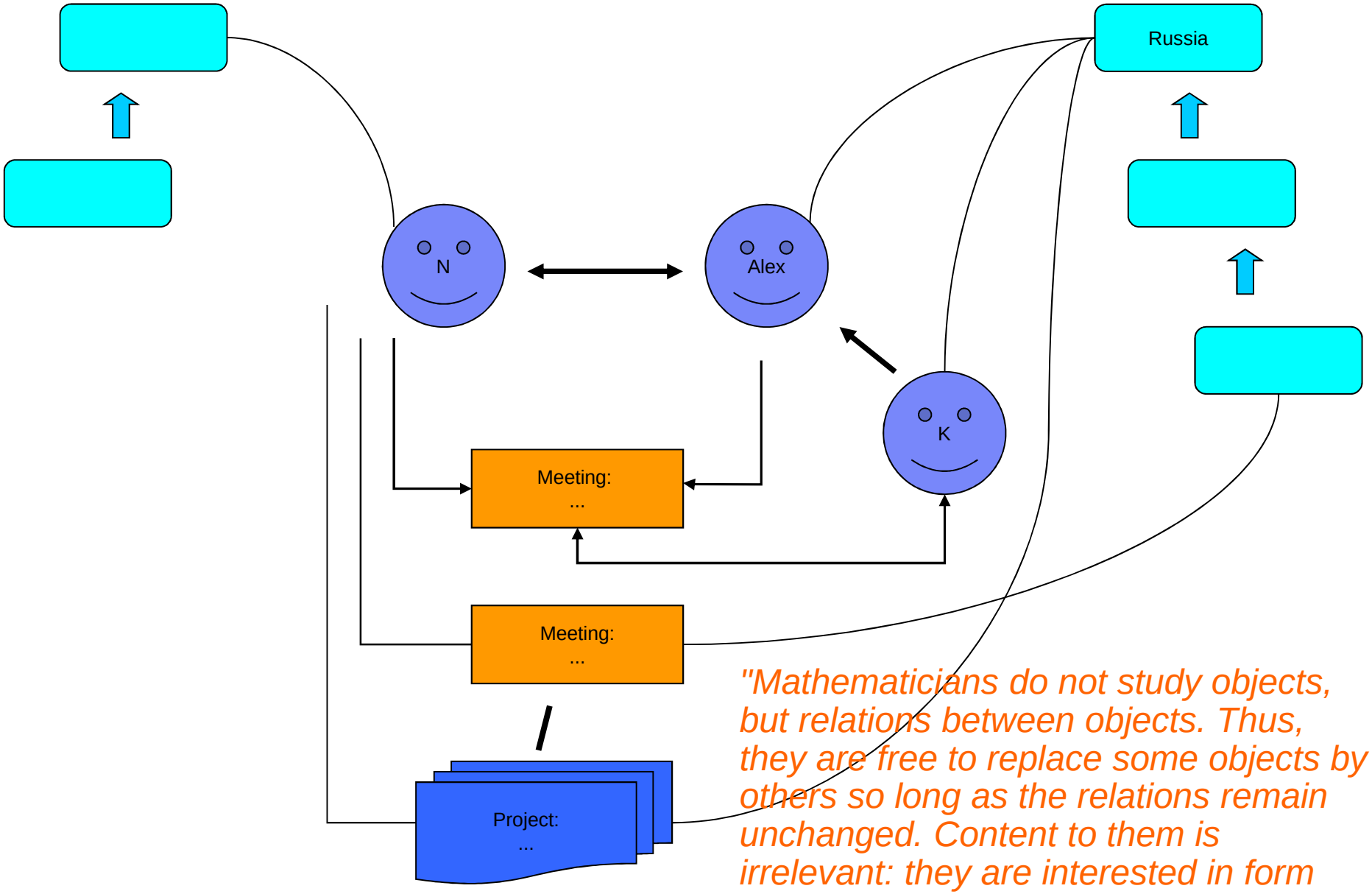
AGENDA

- Techno-Social systems – Facebook, Twitter etc,
- An illustrative example of modeling and mining
  And the focus of this talk is not on listing areas of applications, but on developing the most generic scalable technique to mining such systems
- Generalization of the definition – complex systems with free will agents and possibly a lot of other staff, including domain specific special codes, textual descriptions are usually important. Therefore for many tasks mining should be based on a good command of
  - Social network analysis, network analysis
  - Natural Language Engineering
- Should We Generalize? YES, we want to put many applications under one heading, provided that they have similar models. This will allow knowledge transfer between seemingly different domains.
- An example of a real problem with real data. Excellent solution has been achieved using the model successfully used in our Lab to study social networks. The model lends itself to applications of network analysis and natural language processing. This has been achieved by the introduction of the finite difference method on networks, and the use of L^p norm to implement fuzzy logic)
  - Good response from customers and scientists
- Another example - Multidimensional multiresolute clustering - showing
  - Another type of application
  - But the same type of mining techniques – finite difference method on network, used to emulate not heat propagation, but a process with different properties
- Concluding remarks on
  Data VS Models VS Information VS Knowledge VS Real Life
  Statistics VS Machine Learning VS Chaos

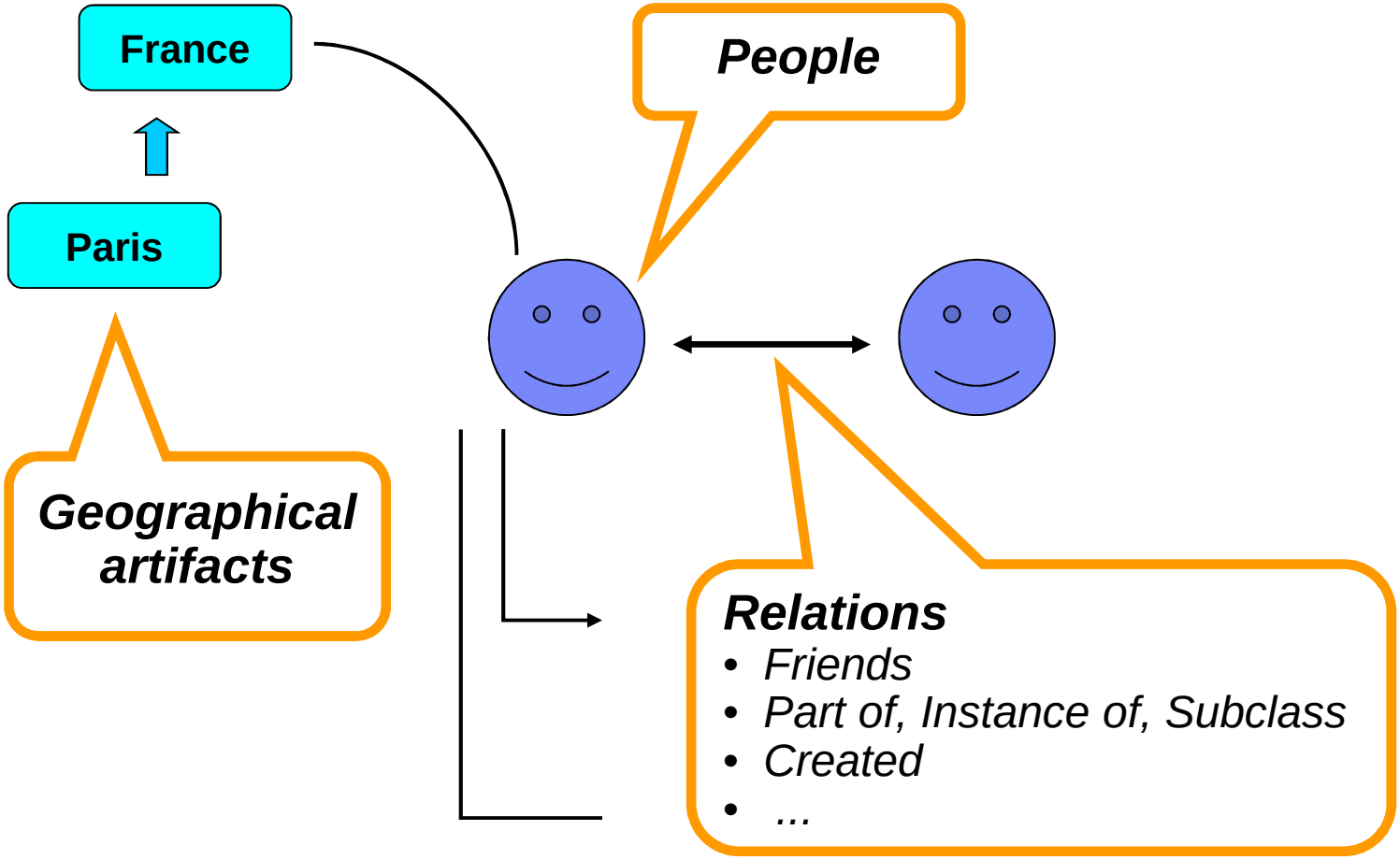
AN ILLUSTRATIVE EXAMPLE OF MODELING AND MINING

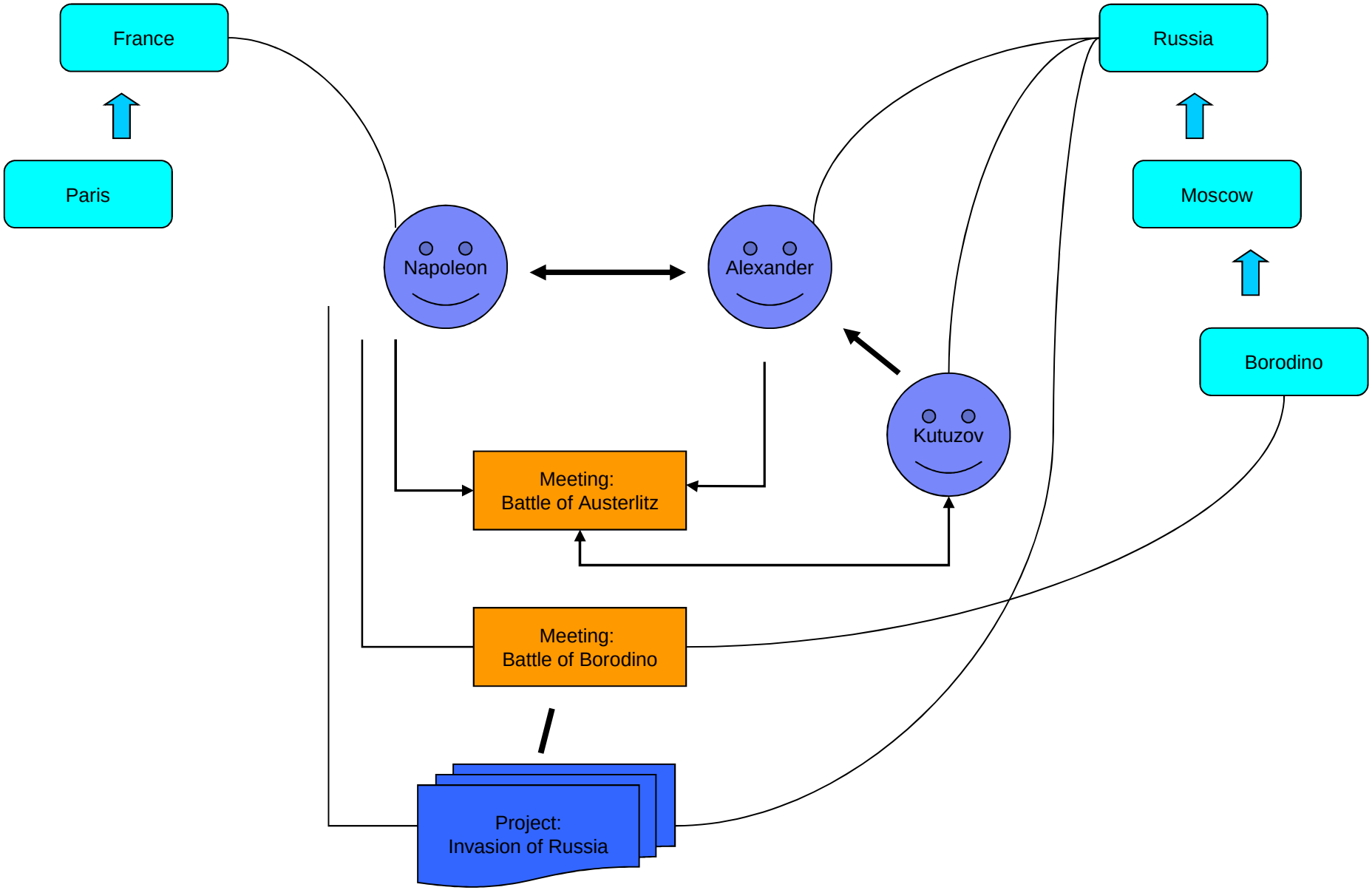
What does the following diagram represent?

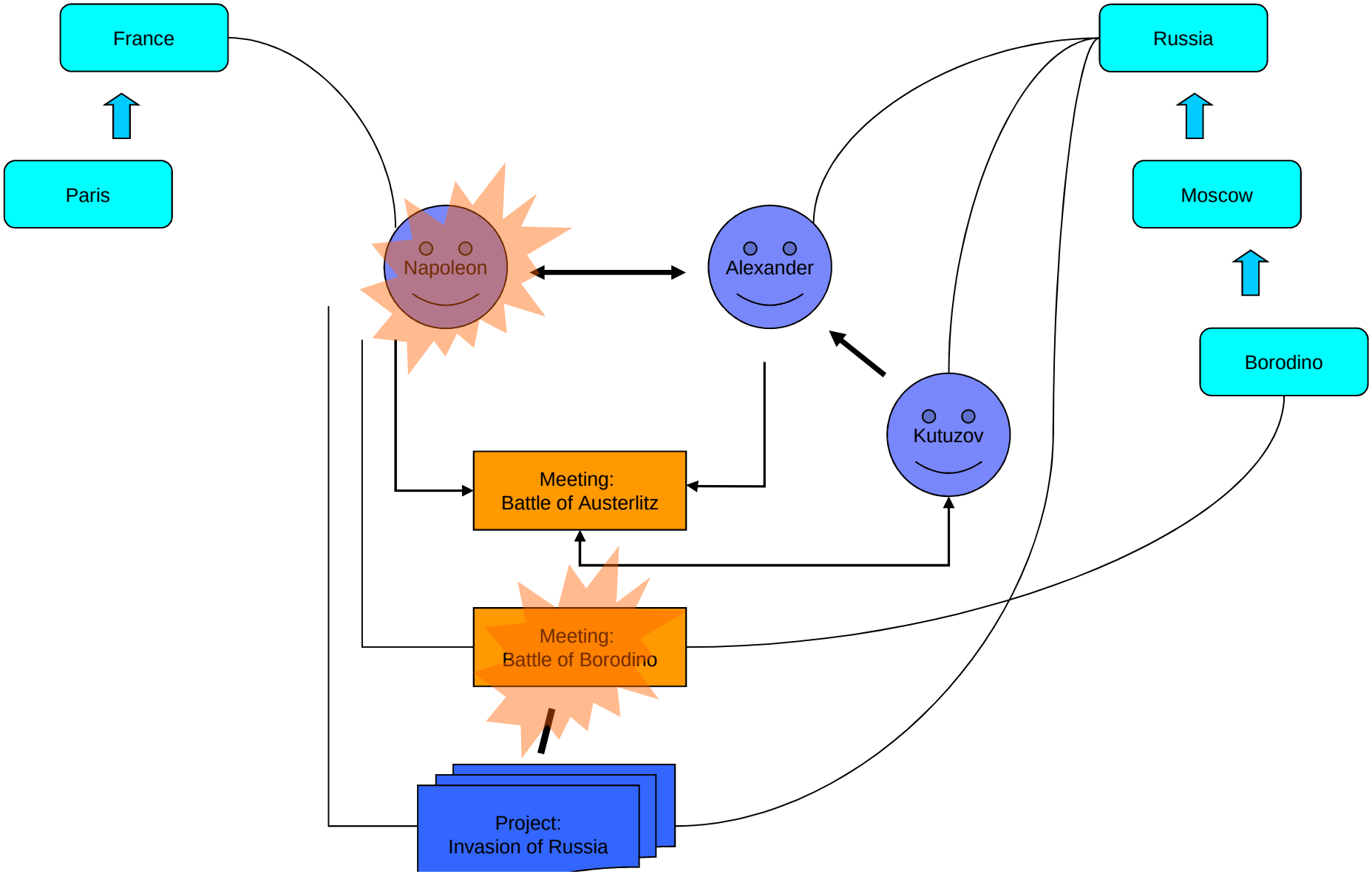
- Data from Facebook?
- An ontology?
- Collocation of terms in a text?
- Or something similar to Mind Maps, to visually organize information, to shows relationships among pieces of the whole?
- …



"Mathematicians do not study objects, but relations between objects. Thus, they are free to replace some objects by others so long as the relations remain unchanged. Content to them is irrelevant: they are interested in form only." - Henri Poincaré


France
Paris
Geographical artifacts
People
Relations
• Friends
• Part of, Instance of, Subclass
• Created
• ...


France
Paris
Napoleon
Alexander
Russia
Moscow
Borodino
Kutuzov
Meeting:
Battle of Austerlitz
Meeting:
Battle of Borodino
Project:
Invasion of Russia


France
Paris
Napoleon
Alexander
Russia
Moscow
Borodino
Kutuzov
Meeting:
Battle of Austerlitz
Meeting:
Battle of Borodino
Project:
Invasion of Russia

What does the previous diagram represent?

- Whatever it represents, it can be easily used,
  for instance, for recommendations like whom Napoleon should invite to this next business meeting “Battle of Borodino”, related to his new project “Invasion of Russia”, as is shown on the previous slide
- To compute recommendation
  - Put the initial activation at the point of interests
    - What - The meeting
    - For whom this recommendation – for Napoleon
  - Propagate this activation
  - Check Constraints on the recommendations – People
  - Select most activated nodes corresponding to People
  - Other highly activated nodes – use for explanations of the recommendation
- Computed recommendation implicitly takes into account most of the requirements for good recommendations for this invitation
  - Somebody who is an expert in Russia
  - Who is likely to accept the invitation
- All the information for such recommendations is encoded in the topology of the network data model

What does the following diagram represent?

- The models like this diagram could be created automatically out of data!
- The model really represents the knowledge about the data and can be used for inferencing
- A kind of fuzzy inferencing since the knowledge is “weak”

TECHNO-SOCIAL SYSTEMS

Techno-social systems

- DEFINITION
- Nowadays, most of the digital content is generated within public and enterprise techno-social systems like Facebook, Twitter, blogs, wiki systems, and other web-based collaboration and hosting tools, office suites, and project management tools. Enterprises use software tools for social collaboration and team collaboration, such as Microsoft SharePoint, IBM Lotus Notes and others. These applications have transformed the collaboration environment from a mere document collection into a highly interconnected social space, where documents are actively exchanged, filtered, organized, discussed and edited collaboratively.
- In techno-social systems infrastructures are composed of many layers (such as Internet communication protocols, markup languages, metadata models, knowledge representation languages which have spanned over two decades) and interoperate within a social and organizational context that drives their everyday use and development.

Techno-social systems – any data sets that have certain properties

- EXTENTION OF THE DEFINITION
- **1 Techno-social systems Collaboration is done using various technical layers**
- **2 Propritary data bases which simply REGISTER collaboration happening outside the system**
- Techno-social systems and various proprietary data bases are primary generators of Big Data about actors and various artifacts, and the relations between actors and the thing they create and do.
- Characteristics provide rich context Introduced and constitutes a new branch of Business Information Systems which requires some new approaches
- In these techno-social systems “everything is deeply intertwingled” using the term coined by the pioneer of information technologies Ted Nelson (Nelson, 1974): people are connected to other people and to “non-human agents” such as documents, datasets, analytic tools and concepts. These networks become increasingly multidimensional, providing rich context for understanding the role of particular nodes that represent both people and abstract concepts.
- Techno-social systems bear most of the general characteristics of Big Data; for instance, in these systems frequently it is easier to predict agent’s actions than to explain them. Mining of techno-social systems constitutes a new distinctive branch of Business Information Systems.

MINING TABULAR DATA

It has been a long, rainy day …

- **At a checkpoint between Russian Federation and an EU country a Russian customs officer is inspecting a track with commercial goods going to Russia.**

Customs Declaration – an example

An example of the itemised document to complete custom declaration for goods transported in one vehicle.

№	DI	E	C1	C2	C3	Goods-TNVEDCode	GoodsDescription	GW	InvoicedCost	CC	CR
7	11	0	967	218	204	9503003500	ИГРУШКИ	159.700	3985.64	EUR	44.0129
8	11	0	967	218	204	2208308200	ВИСКИ	295.950	943.20	EUR	44.0129
9	11	0	967	218	204	4820103000	БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ	15.140	128.64	EUR	44.0129

Another example

912	30	8486209009	СИСТЕМА КЛАСТЕР ПЛАЗМА ЛАБ 100 С КАМЕР	8326.000	4014349.56	USD
912	30	8465930000	ТОЧНОЕ ПОЛИРОВОЧНОЕ ОБОРУДОВАНИЕ	156.300	61285.00	USD
31	465	8708809909	СТУПИЦА ПЕРЕДНЕГО МОСТА	160.000	80.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА VOLVO FH 12	4400.000	1600.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА MERSEDES BENZ AC	3240.000	1200.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА IVECO STRALIS 5	4540.000	1640.00	USD
218	204	9503003500	ИГРУШКИ	159.700	3985.64	EUR
218	204	2208308200	ВИСКИ	295.950	943.20	EUR
218	204	9503004100	ИГРУШКИ НАБИВНЫЕ	7.310	207.00	EUR
218	204	4820103000	БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ	15.140	128.64	EUR
218	204	2402209000	СИГАРЕТЫ	2399.570	83950.80	EUR
218	204	2208701000	ЛИКЕР	2791.390	10220.34	EUR
218	204	3923210000	ПАКЕТЫ ПОЛИЭТИЛЕННОВЫЕ	877.500	5100.00	EUR
218	204	9608200000	РУЧКИ	1.860	74.16	EUR
218	204	2205101000	ВЕРМУТ	3663.860	9185.52	EUR
218	204	2208403100	РОМ	7004.870	25627.20	EUR
218	204	9503002100	КУКЛЫ	46.670	1737.95	EUR

- Gr

People make errors, some people commit fraud

- There are many “hard rules” in customs regulation, but many important decision are left at the discretion of the customs officer.
- AND
  - –People make errors,
  - –Some people commit fraud

- a computer system is able to help

- To a Russian customs officer
  - By raising red flags about potential discrepancies in declarations
  - By recommending actions (like assigning security convoy vehicle for a truck transporting goods to the Russian Federation)
- To a Dutch trader, Polish carrier, Swiss insurer by indicating potential problems with the customs declaration for goods to Russian Federation

MINING TABULAR DATA

- **Pilot project run in collaboration with the Federal Customs Service of Russia and Irish Office of the Revenue Commissioners to create proof-of-the-concept scalable technology platform for mining tabular data**
- **Particularly to create proof-of-the-concept prototype for mining custom declarations for cases not covered by hard rules of custom regulation**
  - Cold start: capable of finding patterns even in small amount of row data, test new records if they follow certain patterns
  - Ease of injection of external knowledge (for instance, relations between customs codes, models of misspellings)
  - Provide recommendation based on found patterns
  - Provide explanations of the rationale behind such recommendations

DATE MODELLING

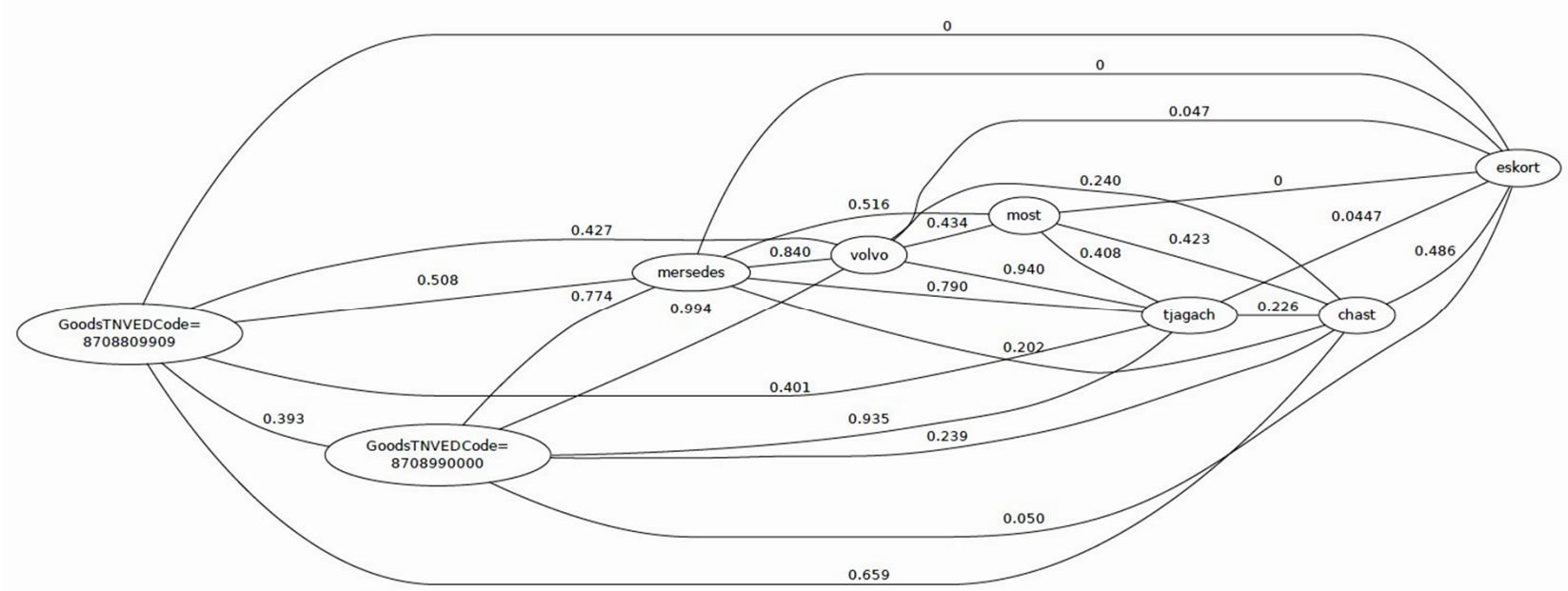
Tabular representation of customs declarations is not necessarily the best model

912	30	8486209009	СИСТЕМА КЛАСТЕР ПЛАЗМА ЛАБ 100 С КАМЕР	8326.000	4014349.56	USD
912	30	8465930000	ТОЧНОЕ ПОЛИРОВОЧНОЕ ОБОРУДОВАНИЕ	156.300	61285.00	USD
31	465	8708809909	СТУПИЦА ПЕРЕДНЕГО МОСТА	160.000	80.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА VOLVO FH 12	4400.000	1600.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА MERSEDES BENZ AC	3240.000	1200.00	USD
31	465	870899	ПЕРЕДНЯЯ ЧАСТЬ ТЯГАЧА IVECO STRALIS 5	4540.000	1640.00	USD
218	204	9503003500	ИГРУШКИ	159.700	3985.64	EUR
218	204	2208308200	ВИСКИ	295.950	943.20	EUR
218	204	9503004100	ИГРУШКИ НАБИВНЫЕ	7.310	207.00	EUR
218	204	4820103000	БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ	15.140	128.64	EUR
218	204	2402209000	СИГАРЕТЫ	2399.570	83950.80	EUR
218	204	2208701000	ЛИКЕР	2791.390	10220.34	EUR
218	204	3923210000	ПАКЕТЫ ПОЛИЭТИЛЕННОВЫЕ	877.500	5100.00	EUR
218	204	9608200000	РУЧКИ	1.860	74.16	EUR
218	204	2205101000	ВЕРМУТ	3663.860	9185.52	EUR
218	204	2208403100	РОМ	7004.870	25627.20	EUR
218	204	9503002100	КУКЛЫ	46.670	1737.95	EUR

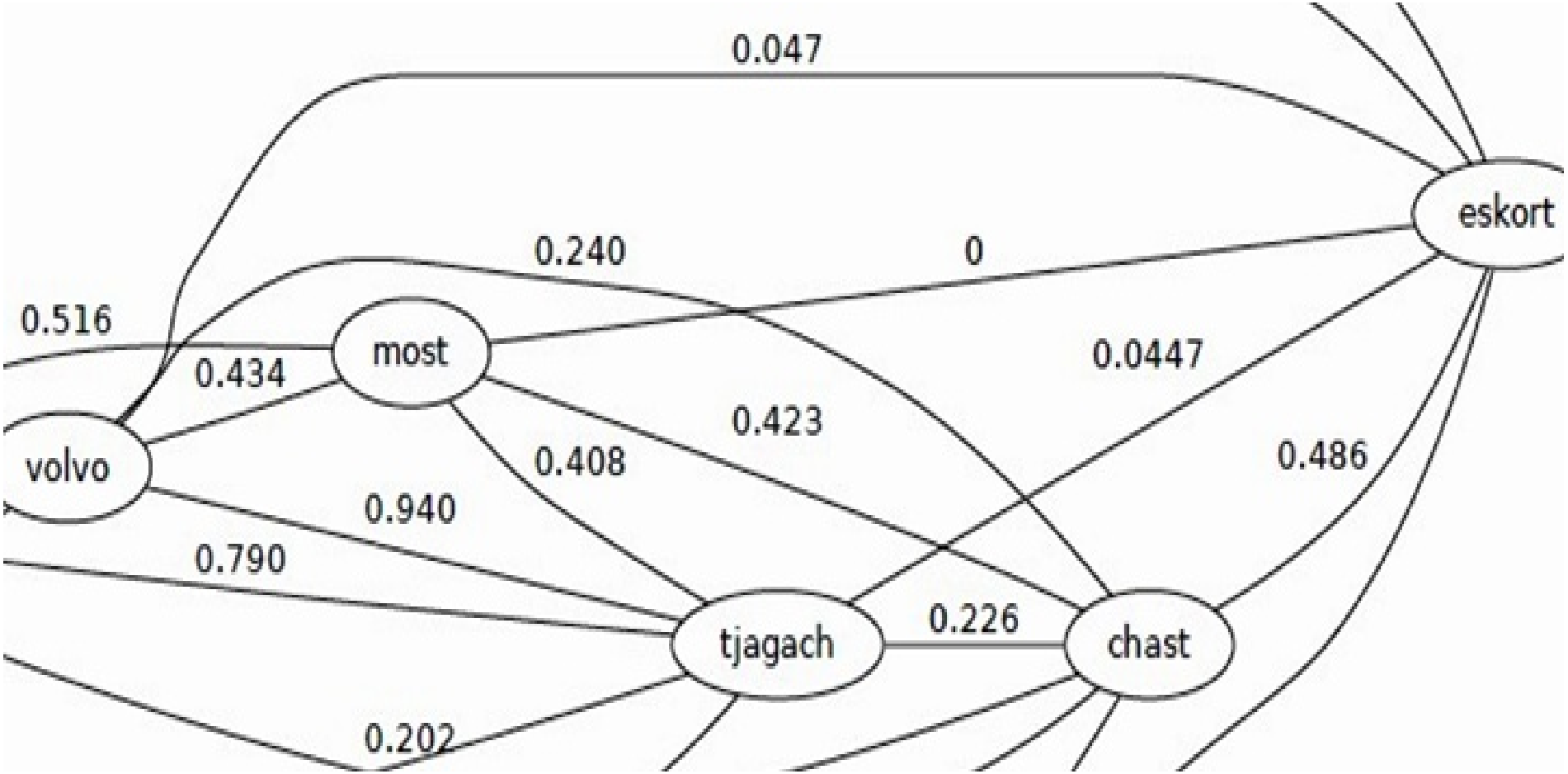
Modelling by multidimensional network

- *Multidimensional …*
  - *– having or involving or marked by several dimensions or aspects*
  - *–… A multi-dimensional database is structured by a combination of data from various sources*
- Multidimensional Network - nodes represent
  - –various abstract codes, for instance, assignment of the security escort, Customs Codes (Commodity codes, GoodsNTVEDCodes)
  - –numerical values measured in kg, seconds, USD etc,
  - –Words, natural language descriptions
  - –Actors (consignee, consignors, carriers)

Modelling by multidimensional network



- The fragment of the network which represents the data from custom declarations.
- Entities (Cells of the table shown before), after preprocessing are merged into one network, where nodes represent words, abstract codes (like assignment of the security escort)
- If two entities are met at least once in the same shipment document, the corresponding pair of nodes is connected by an arc. The weight of that arc represents how frequently the two entities are met in shipment documents (i.e the number of co-occurrences divided by the number of items)

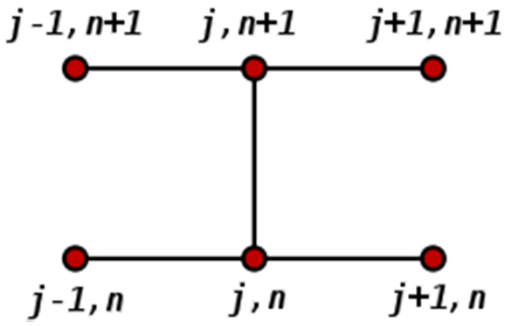


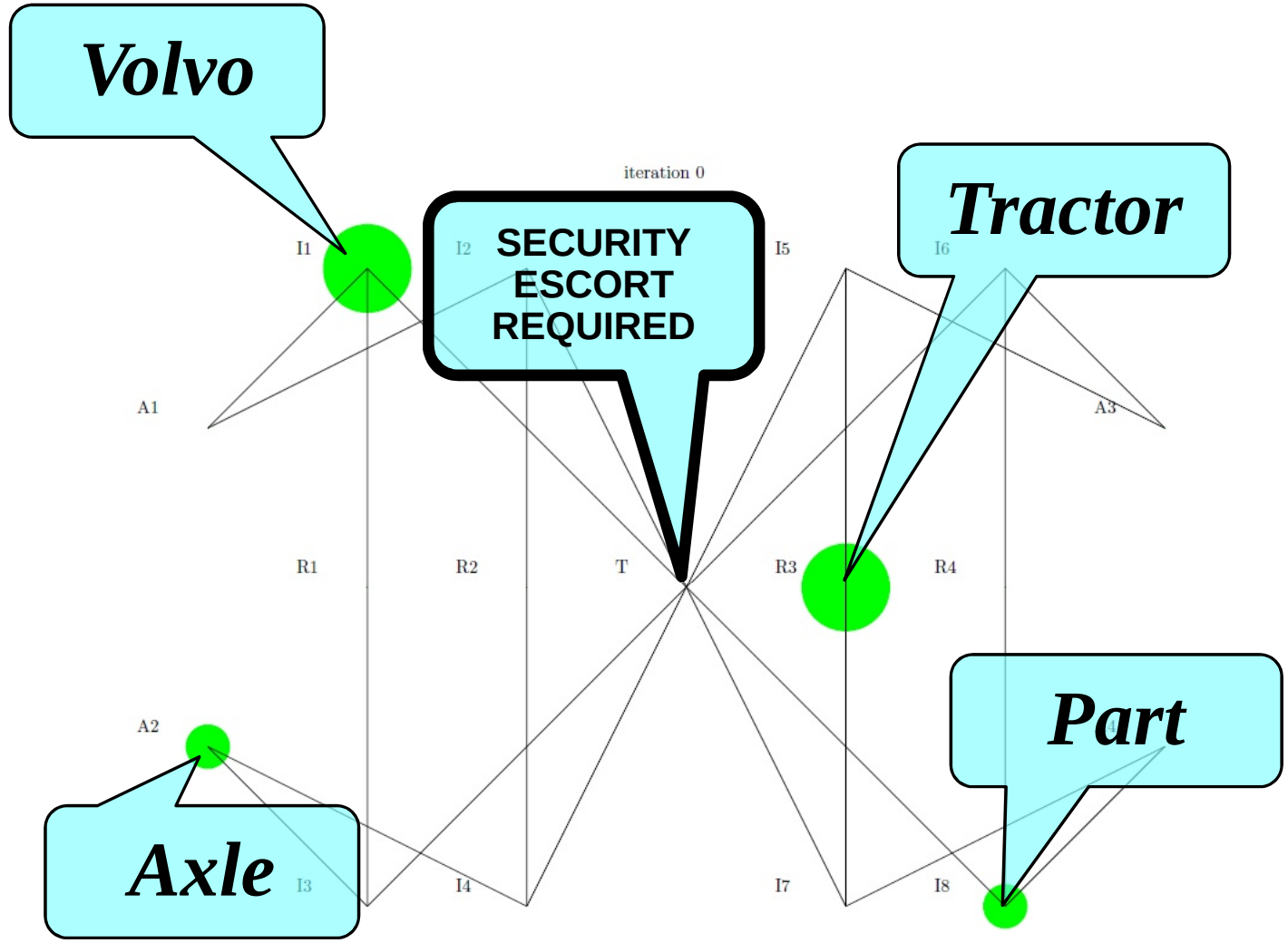
- The fragment of the network which represents the data from custom declarations. The are two types of nodes on this figure: words and security escort tag.
- This network shows, for example, that the word “tjagach” (“тягач” in Russian; rooter, tractor in English) was met in good descriptions which require armed escort in 0.0447% cases. The shipments which has in the description the word “mersedes” (“мерседес” in Russian; Mercedes-Benz international) never required armed escort, while “volvo” required escort in 0,047% of cases. Shipments with parts, details of something (“chast”, “часть” in Russian), frequently were escorted.

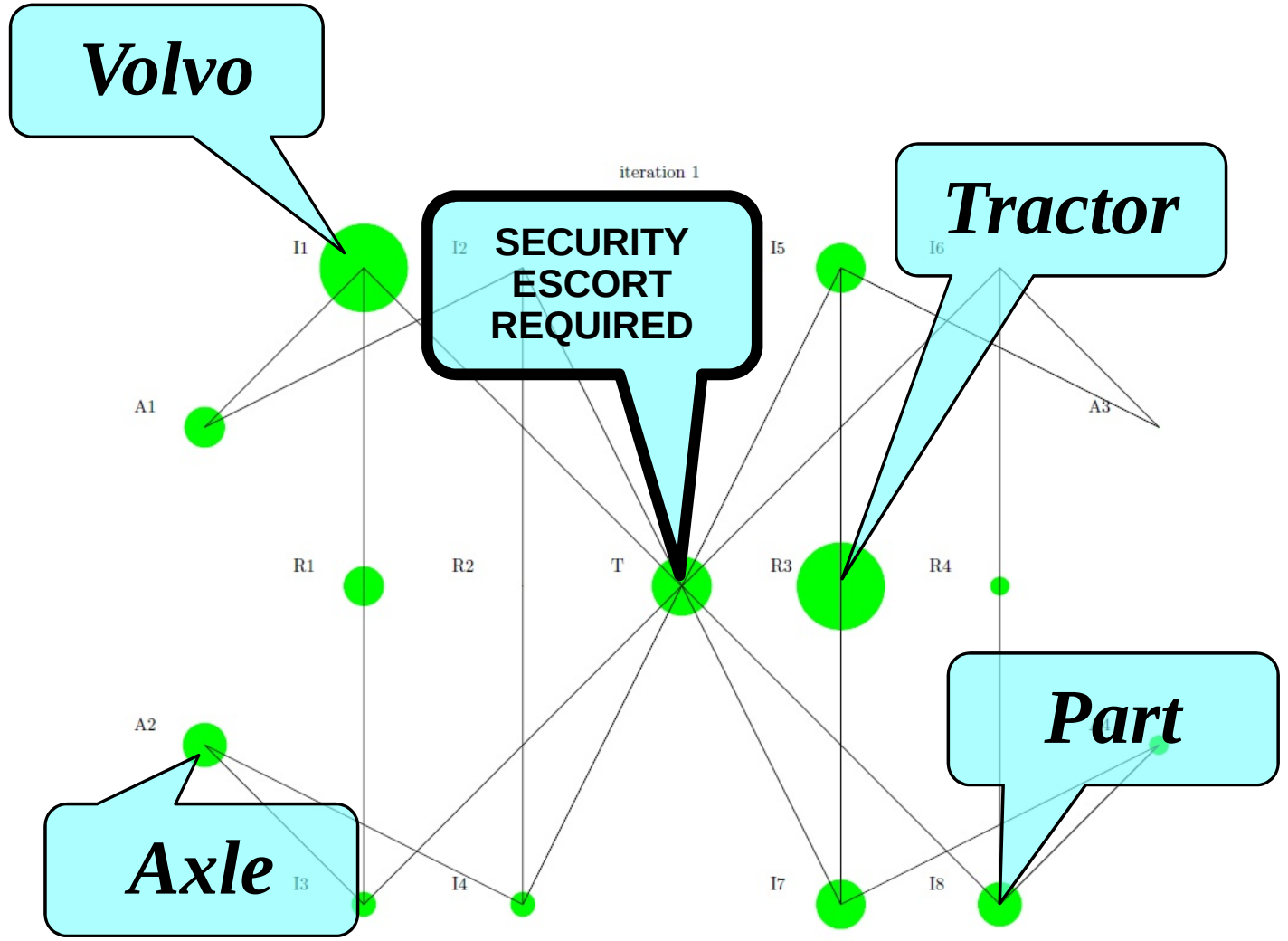
MINING

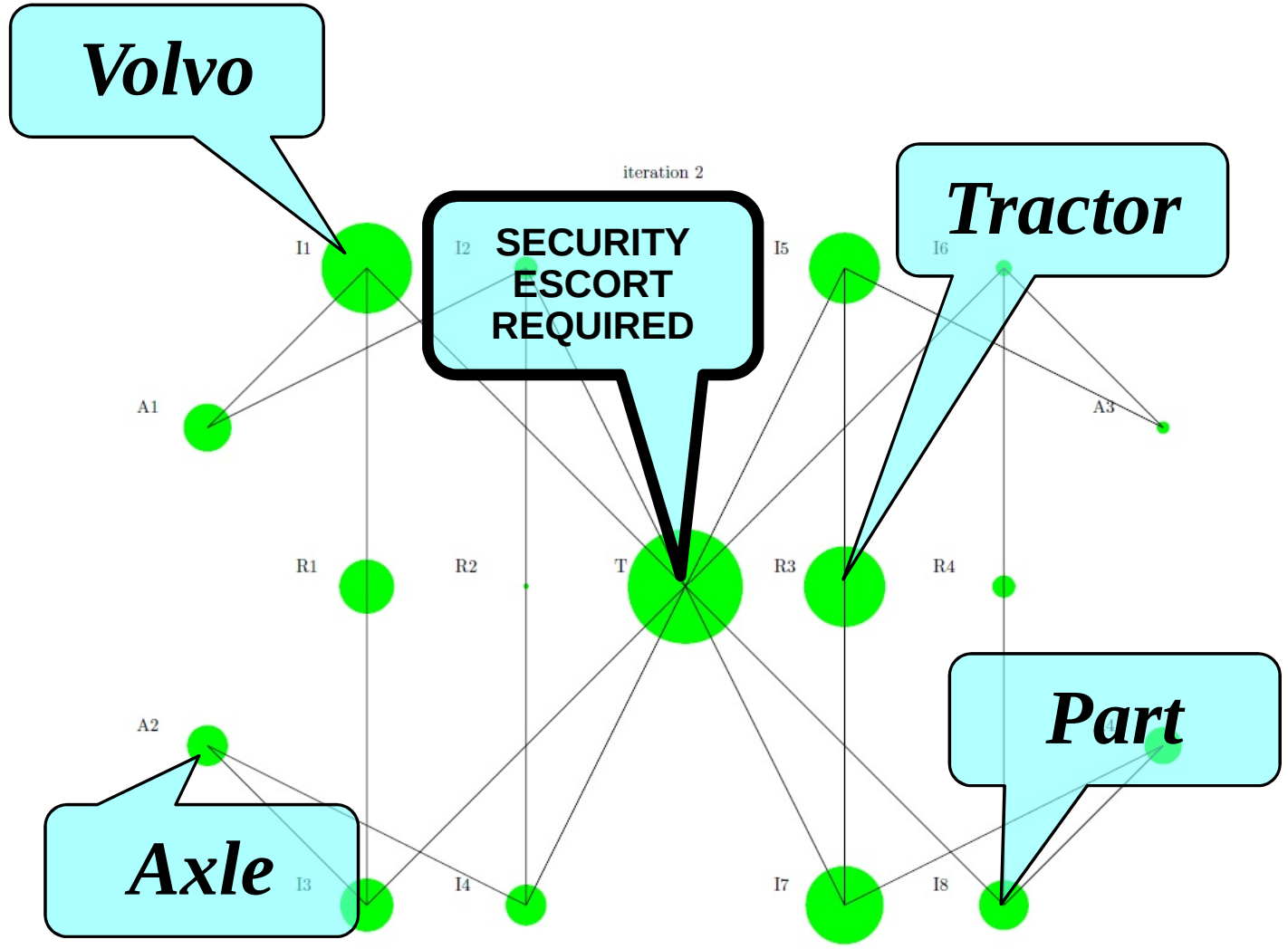
Finite difference method on networks

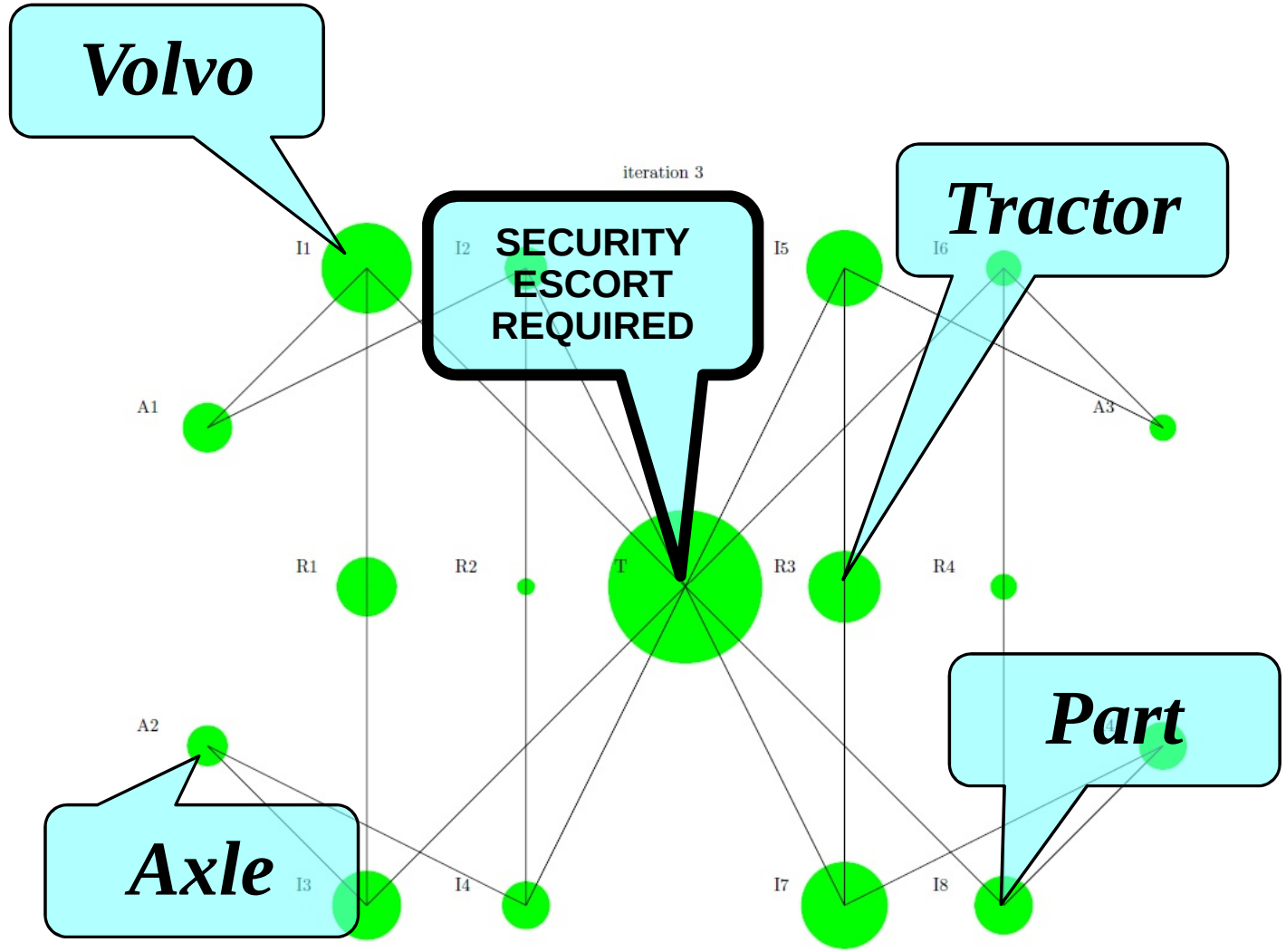
- "Plotting geometric arrangements and forces acting on small segments" evolved into
  - Finite difference method
    - In mathematics, finite-difference methods are numerical methods for approximating the solutions to differential equations using finite difference equations to approximate derivatives.
  - Stencil
    - In mathematics, especially the areas of numerical analysis concentrating on the numerical solution of partial differential equations, a stencil is a geometric arrangement of a nodal group that relate to the point of interest by using a numerical approximation routine. Stencils are the basis for many algorithms to numerically solve partial differential equations.

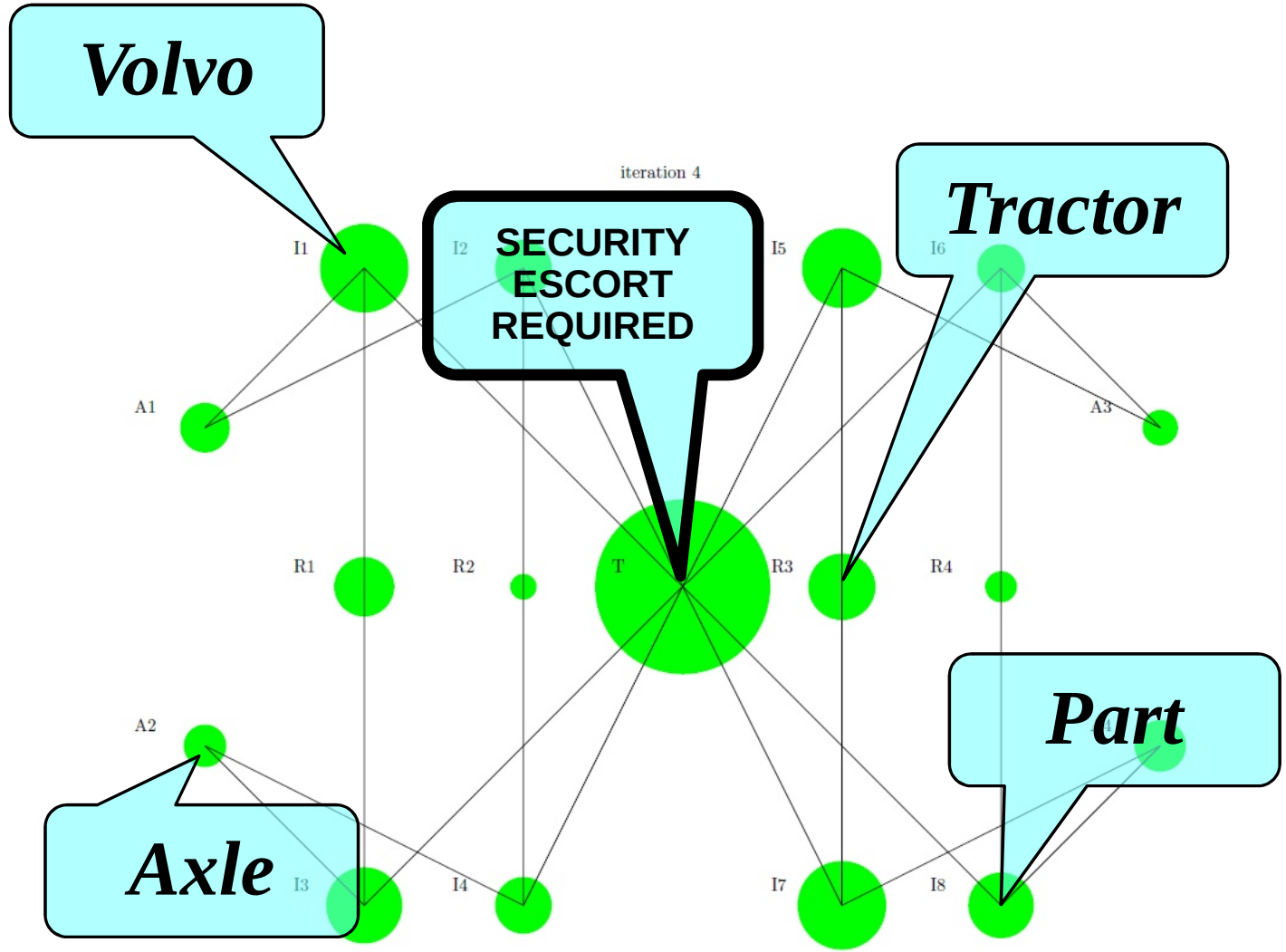


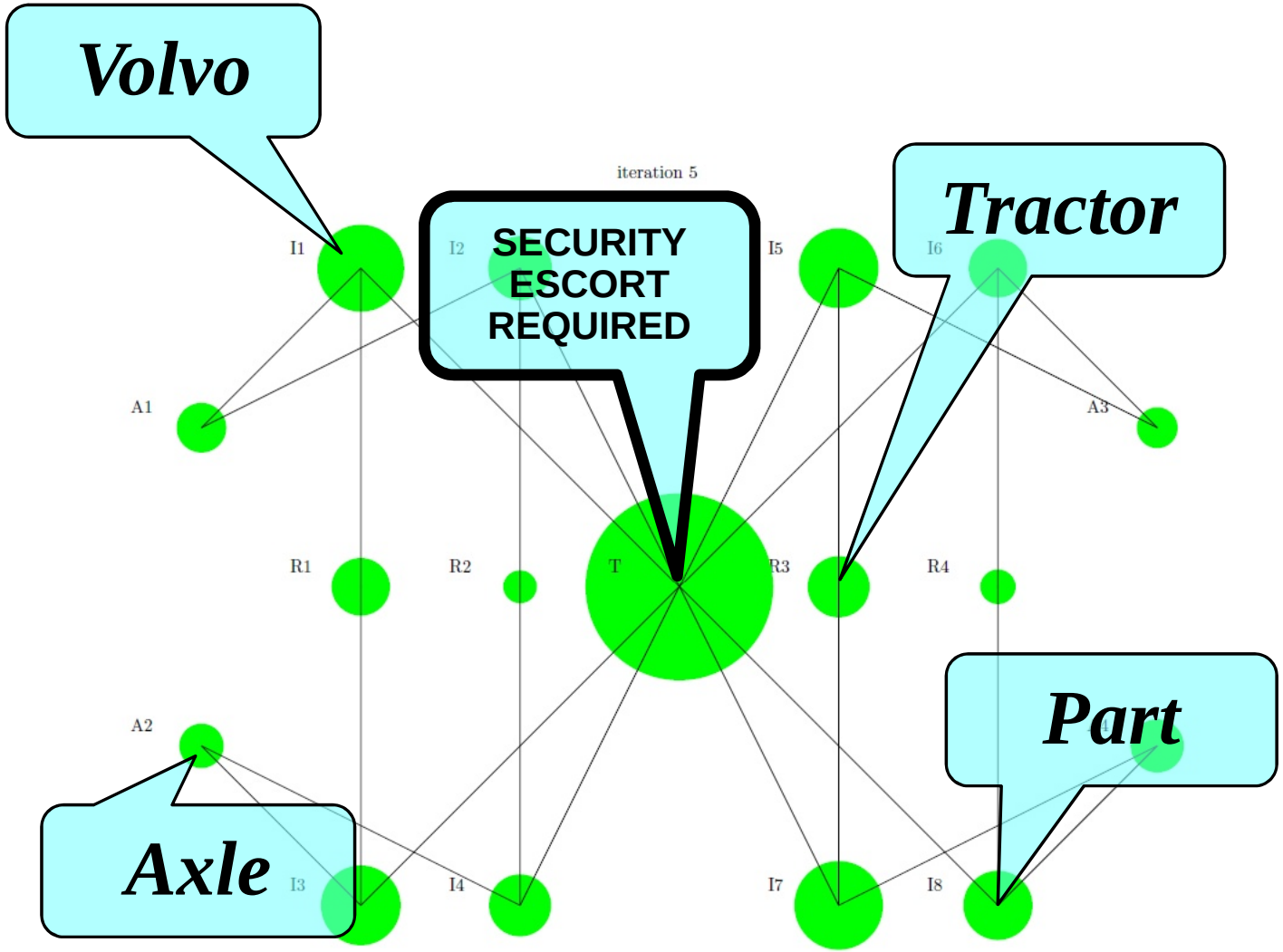

Volvo
iteration 0
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8

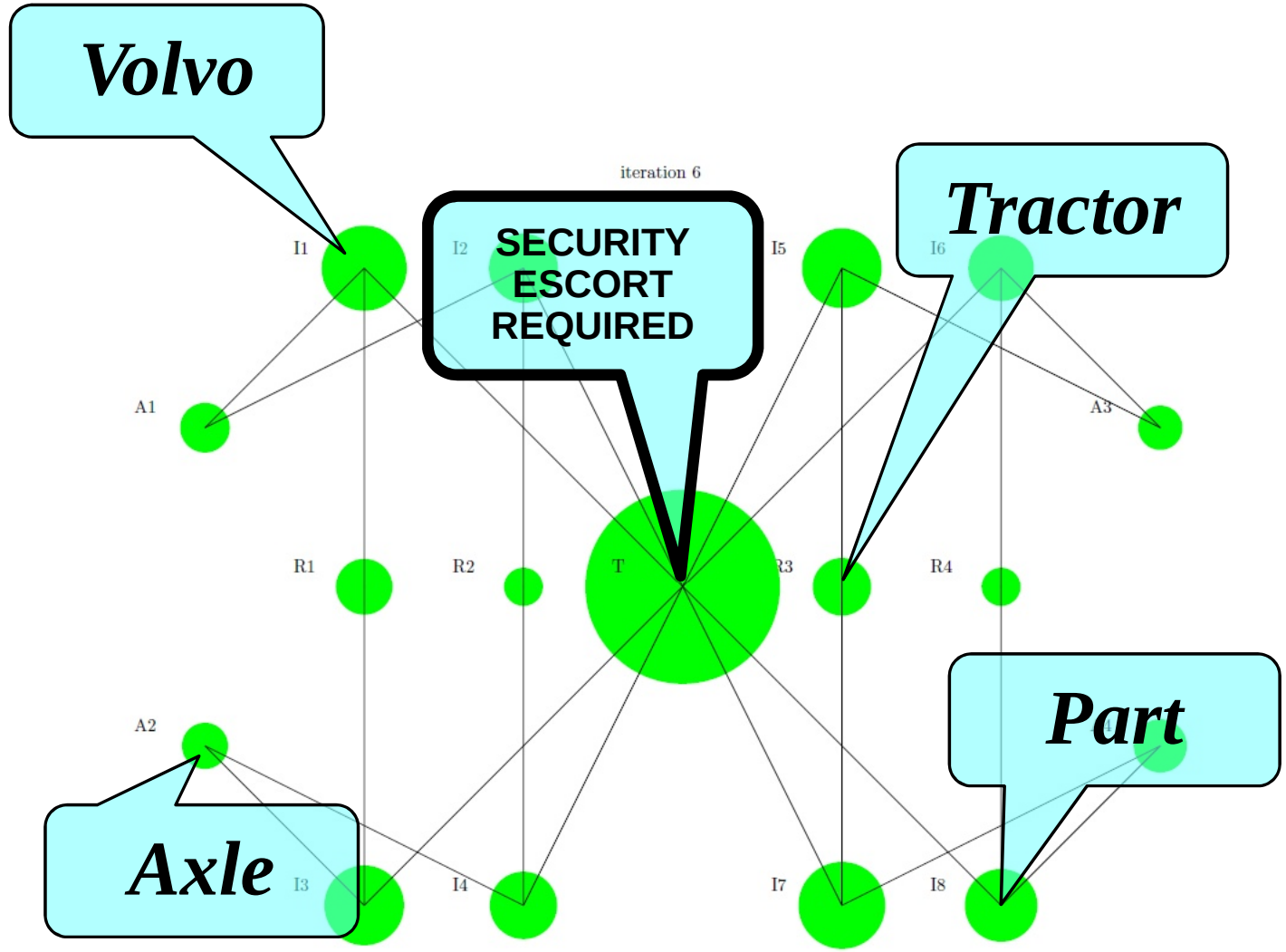

Volvo
iteration 1
SECURITY ESCORT REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8


Volvo
iteration 2
SECURITY ESCORT REQUIRED
Tractor
Part
Axle
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
I3
I4
I7
I8


Volvo
iteration 3
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8


Volvo
iteration 4
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8


Volvo
iteration 5
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8


Volvo
iteration 6
SECURITY
ESCORT
REQUIRED
Tractor
I1
I2
I5
I6
A1
A3
R1
R2
T
R3
R4
A2
Part
Axle
I3
I4
I7
I8

Scheme of the application

- Network – is a representation of the knowledge about the past
- Knowledge allows us make “inferencing” for new data:
  - –We project new data onto the network
    i.e. onto the knowledge about the past
  - –and investigate the consequences
- We can explain this mining in mathematical terms as the transformation of functions on the nodes of the network:

Scheme of the application

- Our approach – 100% Machine Learning
  – Data Black Box Input Output Validation
- However, it is a 100% knowledge based approach, where everything is almost evident.
  - Training – is substituted by automatic build of a suitable model out of data,
  - The work of the algorithm is the same transparent as the work of numerical simulation of heat propagation
  - If the results are not good for the application,
    for instance, some nodes become overheated,
    you look in the problem, and solve it
    by changing microforces, by removing excessive heat, or by simply changing *the physics* of the interactions

    Firstly, manually to see the implications of changes
    Later automatically, pretty much the same as in the artificial neural networks

Novel graph mining method

- We show that a wide range of graph-based algorithms popular in various application domains use the idea of propagation between nodes
- We discuss drawbacks and limitations of these algorithms belonging to the class of network flow algorithms (Borgatti 2005)
- To overcome these limitations, we are developing a much broader class of “physics-inspired” algorithms where the interaction between neighbor nodes could not always be interpreted as a flow producing a diffusion like process.
  - we show that iterative computational schemes similar to those used in network flow algorithms have been used for a long time in finite element analysis to solve physical problems and discuss a class of “physics-inspired” algorithms
- And show the road map to combine physics-inspired algorithms with “logic-inspired” algorithms on graph based on the extension of cellular automata procedure that determines the new state of a cell for the next generation
- We show the applications of these novel class of algorithms to core tasks of network mining – centrality measurements and clustering

PROPAGATION ALGORITHMS AND THEIR USE

- Formally, solution of many network data mining tasks boils down to the following problem: Given an initial function $F_0(v)$ on the network nodes, construct the function $F_{lim}(v)$ which provides the answer.
  - In different domains the function F_0 could be referred to as the initial conditions, the initial activation, semantic model of a text, etc.
  - In ontology based text processing, the initial function F_0 is the semantic model of a text w.r.t. to the knowledge: for instance, $F_0(v)=0$ if the concept v is not mentioned in the text, $F_0(v)=n$ if the concept v is mentioned n times.
    The function $F_{lim}(v)$ should show the foci of the text; for instance, $Argmax(F_{lim})$ is the most important focus of the text, while $F_{lim}(Argmax(F_{lim}))$ is the numerical value of the "relevancy"
  - In IR, the link analysis (such as Google's PageRank) ranks web pages based on the global topology of the network by computing $F_{lim}(v)$ using the iterative procedure where the initial condition is that all web pages are equally "important" ($F_0(v)\equiv 1$),

PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

- Computationally efficient and scalable algorithms usually compute the function F_{lim} (which provides the "answer") using iterations: on each iteration the value of $F_{n+1}(v)$ is computed depending on the values of the function F_n on the nodes connected to the node v
  - Very broad range of algorithms including Google's PageRank, spreading activation, computation of eigenvector centrality using the adjacency matrix.
- Most of the mathematical algorithms behind such iterative computations are propagation algorithms (the "network flow" algorithms):
  they are based on the idea that something is flowing between the nodes across the links, and the structural prominence of nodes could be explained and computed in terms of incoming, outcoming and passing through traffic
- Similar iterative computational schemes have been used for long time in finite element analysis to solve physical problems including propagation of heat, of mechanical tensions, oscillations, etc.
  - Although finite element analysis automata usually perform on rectangular (cubic, etc.) grids, the extension to arbitrary networks is feasible.

PROPAGATION ALGORITHMS AND THEIR USE (Cont.)

- However, the interaction between the material points in mechanics could not always be described as a flow, and such interactions could model more complex processes than diffusion
  - For instance, one dimensional heat transfer equations can be numerically simulated on a one-dimensional mesh by iterations. On each iteration recomputation is based on the formula below:

    $F_{new}(v) = (F(RightNeighbour(v)) + F(LeftNeighbour(v))) / 2$

    This linear equation confirms the perception of the heat transfer as a flow: on each iteration the heat – the value of the function F – flows from nodes to the neighbour nodes.

    In physics, a conservation law states that the amount of heat in an isolated physical system does not change as the system evolves, so “move mechanism” in heat propagation is a transfer. In network theory applications, network flow could be also done by “copy mechanism”, such as replication in spread of deceases.
  - At the same time, in physics, many processes can not be interpreted as a flow and can not be described by a function of one real variable. For instance, to simulate the behavior of an oscillating string one needs to operate with three values at each node - position, mass and velocity of the material point corresponding to the node. And none of these properties “flow” to the neighbors.

VALIDATION

- TWO USE CASES WHICH WERE IDENTIFIED BY EXPERTS IN THE FIELD AS THE MOST PRACTICALLY IMPORTANT
  - checking if the textual description of goods is consistent with the other parameters of the declaration,
  - Assigning or not assigning of a security escort vehicle for the truck transporting goods to the Russian Federation.
- Validation of the approach has been carried out on the data that were not used in building the models; in both scenarios the achieved accuracy was 100 percent in most important use cases. The accuracy measured in top two results, was 100 percent.
- The lowest reported accuracy was 90.5 percent, but the data used were insufficient for the completion of the task (our data model had no time stamps, while customs codes for fruits imported to Russian Federation, such as tomato, frequently depend on the season). However, such cases are easily solved by adding additional simple rules

Injection of Fuzzy Logic

- Our algorithm exploits generic computational scheme on graphs as it has been described in Troussov et al. Spreading Activation Methods. In Shawkat A., Xiang, Y. (eds). Dynamic and Advanced Data Mining for Progressing Technological Development, IGI Global, USA, 2009
- However, in this paper the fuzzy logic has been injected into this scheme; that is the affect that neighbour nodes produce on a node is considered as a logical operation AND.
  - The logical operation has a parameter which allows to regulate how well parameters of operation can compensate each other. To explain this phrase one can use the direct analogy between fuzzy logic operation AND and the operation used in arithmetic to compute means of several numbers. The case of full compensation corresponds to arithmetic mean, which is if one of the arguments will be increased by a certain value, and the other will be decreased by the same value, the results will be the same. However, in very many practical applications arithmetic mean is not an appropriate method for calculating an average, and other methods, including geometric mean, are used.

Fuzzy Logic: The experimental results strongly confirmed that:

- Graph mining using the generic spreading activation algorithm described in [Troussov et al., 2009] does not provide good results;
- Injection of fuzzy logic makes the spreading activation method, which is a wide class of algorithms, suitable for both cases;
- The two use cases considered in this paper are “the polar” use cases in terms of the essential properties of the algorithms providing accurate prediction.
  - In the task of security escort assigning, most reliable features should dominate the results trumping less reliable predictors; which in terms of fuzzy logic operations could be expressed as follows: the results of the operation AND should be highly skewed towards the value of the bigger operand.
    The logical aggregation skewed towards the value of the bigger operand spectacularly fails for the other use case, which is the prediction the nomenclature code of goods based on the words in the textual description. Authors demonstrated that for this use case, significant number of weak predictors easily overrules the judgment based on strong predictors.
- There is a one parametric family of logical AND operations, which covers both polar use cases discussed above. The value of the parameter which provides the best solution for a particular use case can be automatically learned from the data (and the formal description of the task modelled by subsets of nodes).

ANOTHER USE CASE OF OUR GRAPH MINING ALGORITHM
Multidimensional multiresolute clustering of massive socio-semantic networks

Multidimensional multiresolute clustering

- The same algorithm
  or to put it correctly, exactly the same iterative computational scheme of the finite difference method on networks
- The difference – initial conditions, and the physics of the interaction between neighbor nodes:
  - Napoleon: Whom invite to the meeting – Napoleon Meeting
    Physics – heat propagation provides ranking of nodes
    Everything else is outside the algorithm
    To provide recs
    explanations
  - Customs Declarations – Initial conditions – words
    Physics – propagation
    Measurements only heat in the node “Push The Red Button”
  - Clustering –
    Network a lot of dimensionalities
    Initial conditions – all nodes
    Physics - cellular automata, like Convey’s Game of Life

Multidimensional multiresolute clustering

- The algorithm has been validated on the task of clustering 46 million users of the largest European online social networking service VK, originally VKontakte

- Crawling
  - Crawling has been done through the official API of the VK social network. The limitations and low performance of this API has been largely overcome by tools, which are outside of the scope of this paper. The collected data include:
    - Profiles,
    - Lists of friends,
    - Text posts,
    - Social links between users.

Multidimensional multiresolute clustering

- Crawling has been performed in the period 1st of the August 2016 till 2nd of October 2016 by 25 high performance virtual servers. Technical tools used in multilink flexible structure Web crawler, data collection and analytic module include GreenPlum Database (GPDB) from Pivotal, JSONB, HTTP REST/JSON web-service, Nginx, Python3, DigitalOcean, Python scripts and libraries numpy, scipy, graph-tool, sklearn, xgboost.

- Data curation
  - On this stage we selected out of 320 millions of profiles only those, which look like profiles of real people from Russian Federation who are active on VK, and whose interests could be detected based on their posts, not on what they claim as their interests, since most of the announced interests are not reliable or could not be interpreted. Examples of such claimed interests include profile of a young lady “My interests are friends, brother and HIM!”, and the interest “It is good that I moved to St.Petersburg”. We also discarded profiles where names and family names could not belong and could not be transformed to names and surnames or real people.

Multidimensional multiresolute clustering

- Since many users registered using non-canonical or deliberately alternated forms of their name, we applied a recurrent neural network to check if names and surnames could be normalized to the form, which could belong to real people. This task has been performed based on the linguistic dictionary, the data base of names and surnames of 4 millions real people, the data base of one thousand pairs of transformation: string to real name.

- The performance of the clustering algorithm introduced in this paper is low degree polynomial; algorithm computed interpretable overlapping clustering in 4 days on two servers. The results are presented in several ways, including list of interests sorted by the number of users having this interests, the graph of connected interests, the overlapping clustering of users. All the operation used are “knowledge free”, parameters of modelling and algorithm include selection of several thresholds, most notably, thresholds for term frequencies.

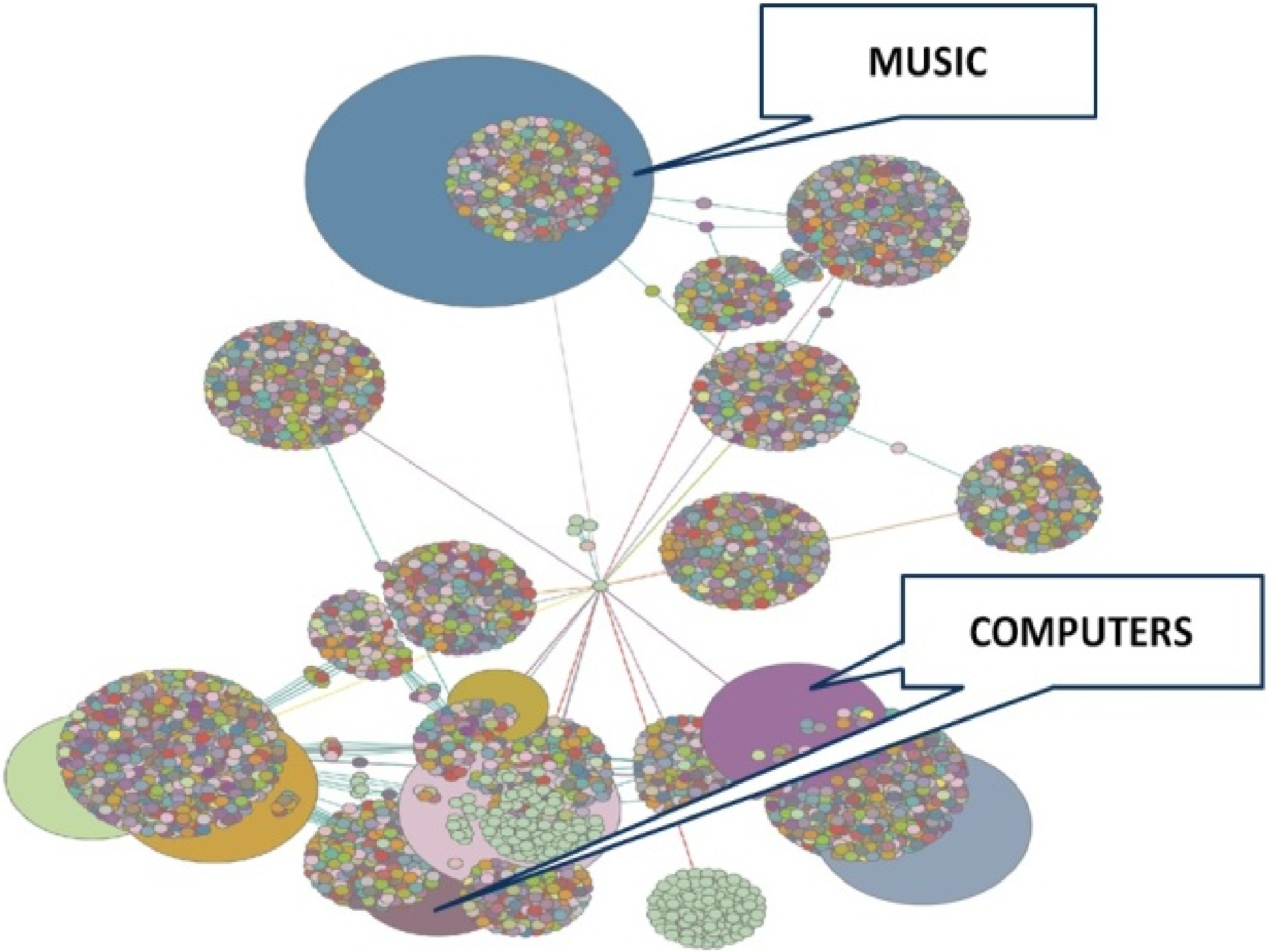
Multidimensional multiresolute clustering

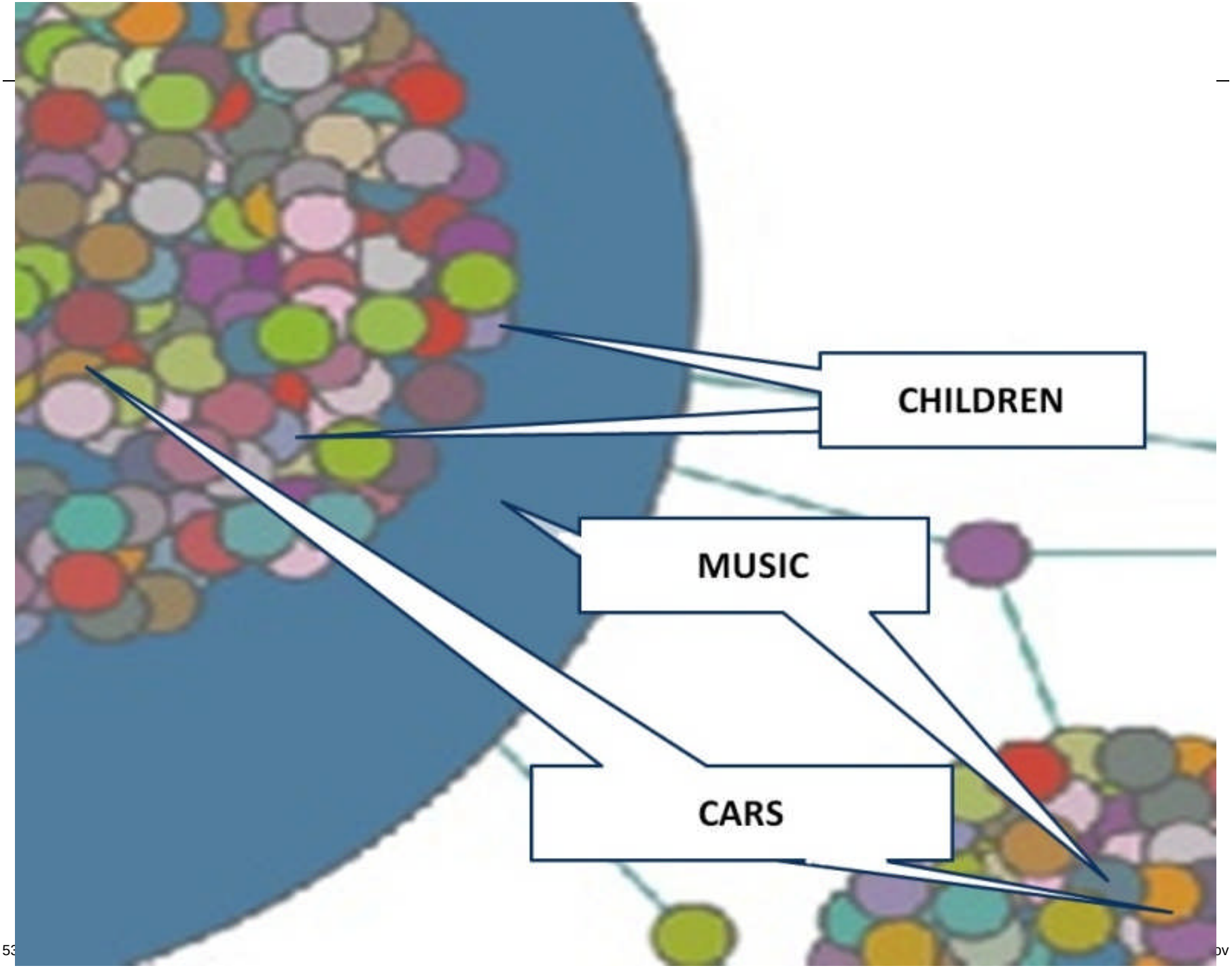
- Multidimensional –
  - we automatically detect the interests of users based on the results of textual analysis of their posts.
  - We automatically label each interest by a set of automatically detected keywords
  - This allows us
    - To give various dimensions, to contextualize social links. To overcome limitations of Friends of Friends notion

      Friend of a friend (FOAF) is a phrase used to refer to someone that one does not know well. The rise of social network services has led to increased use of this term. “Six degrees of separation” and the "small world" phenomenon are related terms.
    - *You are my soccer pal, but your mathematical friends are all boring*
- Cluster – a group of users
  - who all share a particular common interest (in addition to other interests)
  - Any two member of the cluster are connected by a chain of social links within the cluster

The most popular interests of Russian internet users

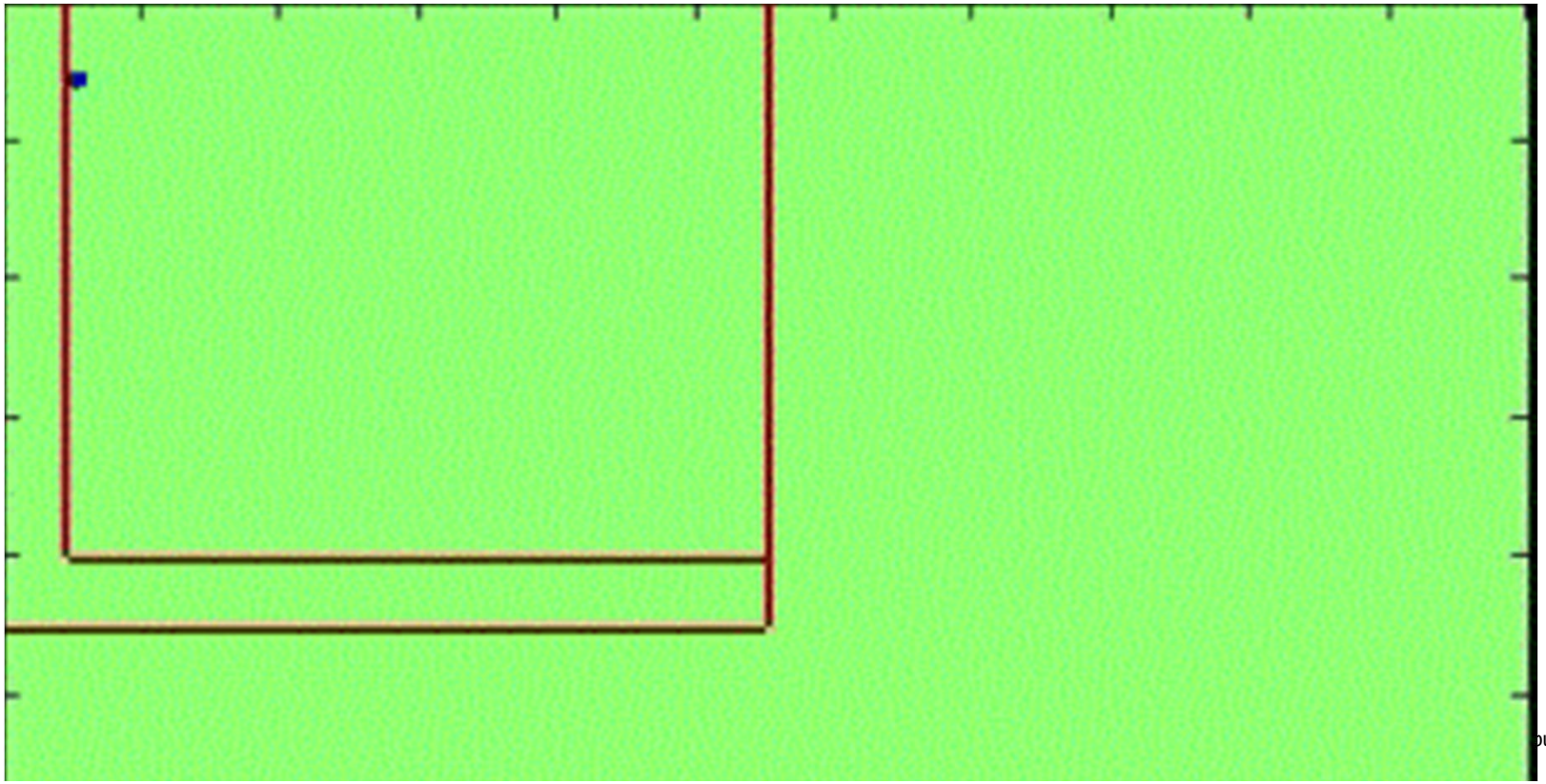
- the most populated groups
  are *music, computers, children*, etc.,
- Most of the user profiles falls simultaneously into several interest categories.
  - The top ten most populated intersections of interests in decreasing of popularity order are
    {recreation, outdoor activity, nature} {travel, children, internet}, {skiing, music, mountain skiing, foreign languages}, {vacations, people, outdoor activities}, {cars, snowboarding, business}, {computers, cars}, {soccer, vacations, cars}, {basketball, cars, business}, {cars, sport}, {psychology, arts}.
    In top 50 one can find the following combinations in decreasing order:
    {fishing, sport, music},{fishing, skiing, nature},{sex, internet, sport, business},{cars, girls},{dancing, theater, fitness, sport, work},{computers, games, computer games},{billiards, cats}, {photography, travels, summer, life, music, family},{photography, cinema, architecture, photo, music, design}
- Multiresolute – the algorithms produces in one go clusters on all levels (micro-, mezzo, macro-level), as well as relations between clusters.
  Topology of clusters is not dendroidal, but dendroidal-like visualisation is useful.


MUSIC
COMPUTERS


CHILDREN
MUSIC
CARS

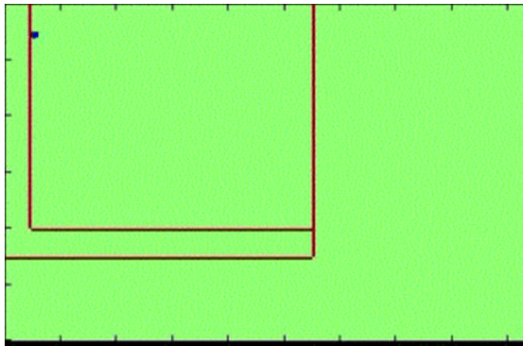
Multiscale Structure of the Media (continuous media) is frequently discovered by *Processes*

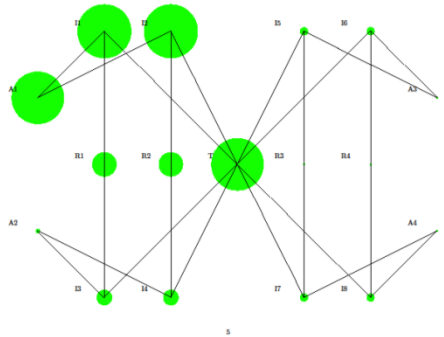
Processes on Networks:
How we study the Earth?
It is good to have master equation (?)
The story of Catenary



Processes in Networks

- How we study the structure of the Earth?
  - RADARS (on aircrafts, ships…) induces processes, The smaller the wave length – the better the resolution
  - THE INTERIOUR OF THE EARTH:
    Using opportunities (Earthquakes, Nuclear tests, …) or creating generating signals and measuring the input signal in one or more points.
- NETWORKS Similarly, the networks are frequently studied by network flow methods
  - introducing the processes where something is flowing from node to node across the edges
  - Why FLOWING? Probably, the interaction could be much more complex





Processes on Networks

- Our method – finite difference method on networks.
- This method also can be interpreted as an introduction of an artificial process to detect network structures.
  By doing so we actually tune the microforces and *physics* depending on the model and the application
- Mining of Customs Declarations demonstrate that different use cases require different structures, which can be achieved not by reworking the model, but by reworking *the physics*

Concluding remarks on various issues of data mining discussed at this conference

Data VS Models VS Information VS Knowledge VS Real Life … ….

- “The philosophers have only *interpreted* the world, in various ways. The point, however, is to *change* it.” — Karl Marx, Eleven Theses on Feuerbach.
- Data scientists *could* interpret data in various ways, The point, however, is to make a difference (in addition to publishing ☺). That is – to provide services to end-users, not only to other data scientists.

Data VS Models VS Information VS Knowledge VS Real Life … ….

- Let us consider cases in the healthcare or in mining data regarding International trade:
  - One can use statistics and provide the domain expert with valuable … statistics.
    80 percent of patients with these symptoms dye within a month
  - But statisticians usually start from data visualization (scatter plot). Application domain specialist frequently employ clustering – Clustering is good if it helps to interpret the data (so called external criteria)
  - The lecturer concurs with (Mirkin, 2016). *"The general opinion among specialists is that clustering is a tool to be applied at the very beginning of investigation into the nature of a phenomenon under consideration, to view the data structure and then decide upon applying better suited methodologies."*
  - In many applications, end-user needs just recommendation and the explanations he/she can understand
    "Should I push the red button"
    "Should I jump out of the window"
  - and the explanations he/she (who might be quite an expert in the domain) can understand
    Our state of the art artificial network advices you to jump out of the window
    NOT GOOD ENOUGH for any mentally healthy person

Data VS Models VS Information VS Knowledge VS Real Life … ….

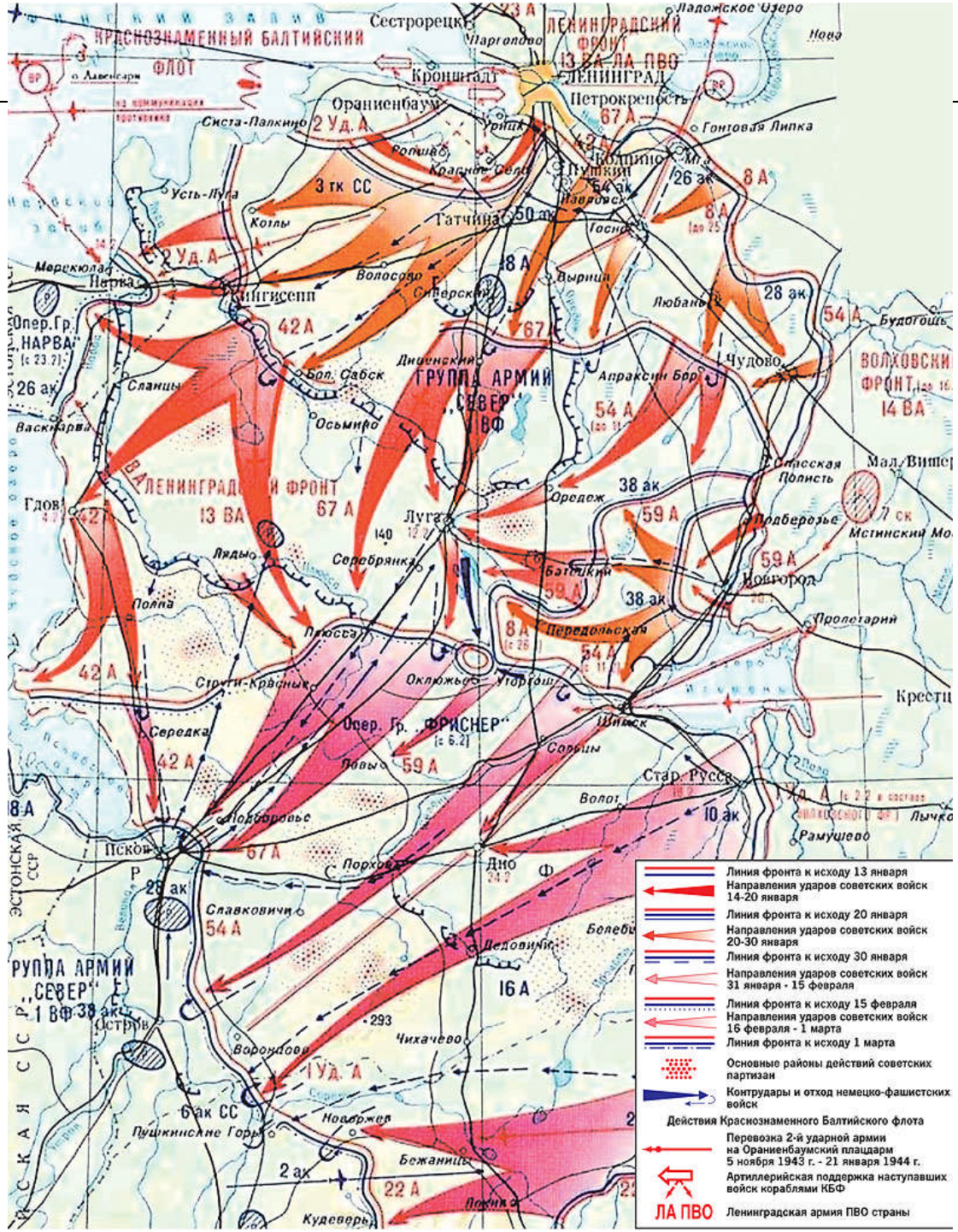
- ML, graph-based methods, rule based methods - are not statistics
- Forecast: ML will be used more and more
  and we will not be able always to understand how the method came to the results
- This is not a problem in processing sensorial information,
  adjective.
  1. of or relating to the senses or the power of sensation.
  2. of or relating to those processes and structures within an organism that receive stimuli from the environment and convey them to the brain.
- BUT …. ????

Data VS Models VS Information VS Knowledge VS Real Life … ….

- BUT …. ????
  But KNOWLEDGE WILL PREVAIL (Claus-Peter …)
  - Recommendations without explanations in most cases are not adequate
    Recommendation can be based on hybrid systems:
    We forecast the weather by solving Navier-Stocks
    But explained using military style maps with red and blue arrows showing the directions of troops movements and charges
  - Interpretational approaches will not die
    Strange case of Sir J. J. Thomson or how positivists missed the discovery of the electron

. . . . . .

- . . .
  . . .
- . . . .
  . . .



What is knowledge?

What is knowledge

- It is surely relevant to our ability to arrive to new conclusion, using network of related artifacts not necessarily following paths like Socrates is Human, Human but navigating multiple paths from several relevant nodes at the same time.
- BASED on the theory of Thomas Kuhn about the Structure of Scientific Revolution and several great thinkers
  - Hegel !
  - Laplace - "Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it — an intelligence sufficiently vast to submit these data to analysis — it would embrace in the same formula the movements of the greatest bod-ies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes." – Pinnacle of the view of the world as Clockwork -
    WAS GREAT ACHIEVEMENT now proved to be wrong
  - And other great mathematicians and thinkers.
  - "If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is." - John von Neumann
  - I. M. Gelfand - always discriminated between solving the mathematical problem and between understanding it.
  - Alexandre Groetendick - One thing Grothendieck said was that one should never try to prove anything that is not almost obvious. This does not mean that one should not be ambitious in choosing things to work on. Rather, "if you don't see that what you are working on is almost obvious, then you are not ready to work on that yet,"
- KNOWLEDGE – comes hand in hand with the conceptualization of new phenomena and artefacts, leading to the development of the language.

Thank you!

- Should you be interested
  - In collaboration with my Lab
  - To work with us on the curricula for teaching BIG DATA, big data analytics
- I'm working on an edited book (related to the topic of this talks)
  - To receive call for chapters
  - To write chapters and get Thomson Reuters, Scopus etc publications
- Please contact me at

troussov@gmail.com
Alexander Troussov

  - …