Human-in-the-loop and other human centric computing paradigms for modeling, simulation and decision support

Janusz Kacprzyk, *Fellow of IEEE, IET, IFSA, EurAI, SMIA*

Full Member, Polish Academy of Sciences
Member, Academia Europaea
Member, European Academy of Sciences and Arts
Foreign Member, Bulgarian Academy of Sciences
Foreign Member, Finnish Society for Sciences and Letters
Foreign Member, Spanish Royal Academy of Economic and Financial Sciences (RACEF)

Systems Research Institute, Polish Academy of Sciences Warsaw, Poland
Email: kacprzyk@ibspan.waw.pl

What is this talk basically about?

In general, a (plenary/keynote) talk should be concerned with a challenge or something that is important

This talk concerns one of big challenges facing IT/ICT

For instance, National Science Foundation says that:

- . . . there are the following . . . monumental research challenges, each requiring at least a decade of concentrated research, to make substantive progress:
  - . . .
  - Computers to be cognitive partners for humans,
  - Personalized lifelong learning environments,
  - Unfailingly reliable systems,
  - Making information technology less complex (to the humans!)
  - . . .

For our purposes:

> *Computers should be cognitive partners to the humans*

Unfortunately:

> *there is a huge (and growing!) gap between the human being and the "machine"(computer)*

because the power and capabilities of the computer systems (hardware and software) are growing, new computing paradigms are developed, etc. but the human cognitive, information processing, etc. capabilities remain practically the same, i.e. we are not better or smarter than our ancestors from, e.g., the ancient Greece or China

Therefore, the setting assumed here, which is in fact some "meta" problem:

- a growing complexity of social, technological, economic, etc. processes and systems which call for:
  - good (better?) decisions,
  - finding ways to an effective and efficient making (implementation) those good decisions,
- a growing discrepancy (gap) between the practically constant information/knowledge processing capabilities of the human beings and a growing capabilities (so far, mostly related to number crunching but maybe to "intelligent" capabilities, too) of the computers,
- a communication/articulation/cognitive gap between the computer and human being:
  - strings of 0/1s for the computer and
  - natural language for the human.

What we do (or want to do) in virtually all situations we deal with:

- We (want to) make decisions!
- These decisions should be "good", at least useful for somebody.
- These decisions are made by the humans and for the humans!

  Maybe by and for some inanimate agents who mimic (to some extent) the humans.
- We have to use information (data, knowledge, maybe wisdom) to make those decision.
- we have to use whatever tools and techniques may be effective and efficient (at least useful), in particular:
  - modeling,
  - "rational solutions" (optimization!),
  - simulation!, etc.

Very many aspects . . .

We will concentrate on just some aspects, notably:

- How to best solve the problem that all this is made by the humans and for the humans,
- How to make the best use of data, information, knowledge, . . . taking again into account the above fact: by the humans and for the humans.

Therefore:

- we will first advocate a more general philosophy, along the general idea of human-centric systems, and human-in-the loop
- we will then advocate the use of some less standard ways of making use of data, focusing on a wide use of natural language.

- Decision making is a “meta-problem”, omnipresent, in virtually all human activities,
- Decisions are made by humans, for (to suit) humans; may be mimicked by/in inanimate systems.

Decision making usually proceeds in a “multi-X” seting:

- multicriteria,
- multiperson (multiagent),
- multistage (dynamic).

Here:

- multiperson (multiagent), rather in a group decision making setting, preference, not utility function based,
- but our discussion is general.

What is new?

Now: modern, good, . . . decision making

Decision making process (DMP) (presumably introduced by Snyder in the early 1950s):

- Use of own and external knowledge,
- Involvement of various “actors”, aspects, etc.
- Use of explicit and tacit knowledge,
- Account for emotions, intuition, . . .
- Non-trivial rationality,
- Different paradigms when appropriate.

Virtually all elements are “human specific”, imprecisely specified (fuzzy logic should be a proper tool?)

Decision making process

Traditional decision making process:

the main stages are:

- Intelligence (information and data gathering),
- Design (selecting a model of a decision situation),
- Choice (of a best option),
- Implementation.

Modern decision making process (with creative, strategic, deliberative, etc. decision making) involves:

- Recognition,
- Deliberation and analysis,
- Gestation and enlightment (the „eureka!”, „aha” effects, very difficult to model, a nonlinear dynamics),
- Rationalization,
- Implementation.

All non-trivial decision making problems are complex: in addition to many variables, constraints, . . . :

- Self-organization: a change naturally occurs which leads to a better functioning of the system by making stronger parts and sub-processes that work well, and weaker parts and sub-processes that do not work well (natural selection!),
- "Non-linearity": all parts of the system affect many other parts throughout the system, and then affects them back, notably change, cause and effect are not due to a single one-way sequential line of events, but reflect interactive influence through feedback,
- Chaotic behavior: results inherently become less predictable getting farther from the original conditions,
- Emergent properties: completely unpredictable results can emerge from their original conditions.

Emergence:

- A direct expression of the vitality of complex “non-linear” dynamic systems,
- The most powerful manifestation of a remarkable self-organizing ability of complex dynamic systems,
- Relations to creativity and innovation,
- Relations to the so-called “aha” and “eureka” effects, etc.

But:

- “linear” decision making(simulation!) models are incompatible with “nonlinear” dynamics (chaos),
- valuation of decisions may loose its “objective” meaning (what is considered good now can be wrong pretty soon),
- many “non-scientific” human specific aspects like emotions, intuition, etc. can be decisive, etc.

So: A decision support paradigm, and a human-computer interaction, i.e. human-in-the-loop!

Modern decision making paradigms

- Heavily based on data, information and knowledge, but also on human specific characteristics (intuition, emotions, attitude, . . .),
- need number crunching, but also more "delicate" and sophisticated analyses,
- Heavily relying on computer systems, and capable of a synergistic human-computer interaction.

So:

- Decision support systems (DSSs)!
- Should be human centric/centered!

What are decision support systems?

Not clearly understood!

Usually:

- specific computerized information systems that support decision making activities,
- interactive computer based systems intended to help decision makers use data, documents, knowledge, models, etc. to identify and solve problems and make decisions,

Support, not **replace** the human being!

Because there are many approaches to decision-making and because of the wide range of domains in which decisions are made, the concept of decision support system (DSS) is very broad.

DSSs – characteristic features

Emphasis on:

- Ill/semi/un-structured questions and problems,
- Non-routine, one of a kind answers,
- A flexible combination of analytical models and data,
- Various kinds of data, e.g. numeric, textual, verbal,...
- Interactive interface (e.g. GUI),
- Iterative operation (WHAT – IF),
- Supporting various decision making styles,
- Supporting alternate decision making passes, etc.

Roots and history

The concept of decision support has evolved from two main areas of research:

- the theoretical studies of organizational decision making done at the Carnegie Institute of Technology (now Carnegie Mellon University) in Pittsburgh during the late 1950s and early 1960s, and
- the technical work on interactive computer systems, mainly at the Massachusetts Institute of Technology (MIT), Boston, in the 1960s, and development of IBM 360 and a wider use of distributed, time-sharing computing.

DSS became an area of active research of its own in the middle of the 1970s

DSSs – some milestones

Mid-1960s: development of IBM 360 and a wider use of distributed, time-sharing computing
Mid-1960s: MISs (management information systems) first to provide managers with structured, periodic reports,
Late 1960s-early 1970s: attempts to use analytical models, first attempts at interactive systems
Early 1980s: EISs (executive information systems) that use relational database, and use predefined screens, and are made by analysts for executives, knowledge-oriented DSSs (use of AI tools), group DSSs,
Early 1990s: Use of relational DBMS techniques, Shift from mainframe based to client-server based solutions, Object oriented technology for builing „reusable" systems.
Mid-1990s: Data warehouses and on line analytical processing (OLAP) tools, Web based and Web enabled systems, etc.

DSS is a multidisciplinary field including (but not only):

- database research,
- artificial intelligence,
- human-computer interaction,
- simulation methods,
- software engineering, item telecommunication, etc.

Basic types of DSSs

A traditional classification (cf. Dan Power's `www.dssresources.com`):

- Data driven,
- Communication driven and group DSSs,
- Document driven,
- Model driven,
- Knowledge driven,
- Web based and inter-organizational.

Basically, all non models driven ones:

- emphasize access to and manipulation of internal and external data, numerical or textual, even multimedia,
- facilitate collaboration between decision makers,

Only the model driven one explicitly uses formal (mathematical) models to derive solutions that can suggest the human decision makers a good (best?) course of action

The best: a synergistic combination

Is a model of a (decision making) problem considered necessary?

No! But maybe helpful. . .

A famous citation:

> *All models are wrong, some models are useful.*

Box, G.E.P., Robustness in the strategy of scientific model building, in Robustness in Statistics, R.L. Launer and G.N. Wilkinson, Editors. 1979, Academic Press: New York.

Our line of reasoning:

- for an effective and efficient human-computer interaction or human-in-the-loop we should:
  - try to bridge an inherent gap between the human being and the "machine" (computer) which, in our context, boils down to the following:
    - For the human being, natural language that is the only fully natural means of communication and articulation,
    - For the computer, "artificial" language of 0-1's is natural, and natural language is strange . . .
- Maybe, we should use natural language as much as possible? here, this philosophy!

Broadly perceived human centric computing/approaches

Our experience and practical implementations – a data-driven DSS!

Decision making is therefore a human centric/centered problem so that to deal with it some human consistent, human centric, . . . formulations, paradigms, frameworks, tools and techniques, etc. should be employed.

Human centric is meant in the traditional Dertouzos' (ca. 2000) sense:

> *no interface between the human being and the "machine" (computers, tools and techniques, . . .)*

Pedrycz (1996 – . . .): more technical!

Idea of human centric computing

Prof. Michael Dertouzos (1936-2001)
Director, Laboratory for Computer Science (LCS)
MIT, Boston, MA, USA

During his term at LCS: RSA encryption, the spreadsheet, the NuBus, the X Window System, and the Internet, defining the WWW Consortium (Tim Berners-Lee was there), support of the GNU Project, etc.

Influential books:
M. Dertouzos (2001) The Unfinished Revolution: Human-Centered Computers and What They Can Do for Us, Harper Collins.
M. Dertouzos (1997) What Will Be, Harper Collins.
M. Dertouzos, R.K. Lester, R.M. Solow (1986) Made in America, MIT Press.

Dertouzos (2001) introduced the concept of a human centric computing:

> *". . . I view human-centric computing as a total commitment to the human as the starting point . . . I start with the interface, and then I go down to all the applications. In the approach we have had for the last 40 years, there is a machine that has all this number crunching power, and then there is an interface that lets us talk to the machine . . . In the new approach, you're not talking to the interface, you're talking to the machine – it doesn't need an interface . . . "*

That is: "to make IT less complexąnd ą cognitive partner for humans"(NSF IT challenges!)

Human centric: in many ways and from many points of view

Many related ideas, for instance:

- Human centered computing: cf. A. Jasmine, D. Gatica-Perez, N. Sebe, Th. Huang: Human-centered computing: toward a human revolution. Computer (IEEE), May, 2007:

  A systems view integrating:
  - Computational tools,
  - Cognitive aspects,
  - Social aspects.

For instance:

HCC: Human-Centered Computing Consortium (University of California at Berkeley) Georgia Institute of Technology, Carnegie Mellon University, etc.

Many more, for instance:

- Human (based) computation (and interactive evolutionary computation) – the computer asks a person (group) to solve a problem, then collects, interprets and integrates the solutions obtained;
  so: the humans help the computer to solve a difficult problem (e.g. strategic planning), e.g.: University of Illinois at U.-Ch. (former David Goldberg's group),
- Humanistic intelligence (S. Mann): arises from the human in the feedback loop of a computations involving wearable „computers" (e.g. smartphones),
- related: social computing, social software, symbiotic intelligence, collaborative intelligence, etc. etc.
- ...

In general, all those approaches to human or human centric/centered/. . . computing try to attain a synergy and amplification between human abilities (e.g. intelligence) and computational power of computers!

Here:

- more into the “new” basically human-in-the-loop which is basically a paradigm (model) that requires human interaction increasing the efficiency of modeling and simulation, machine learning, decision making, problem solving (e.g. strategic planning), etc.

Human-in-the loop, e.g., MIT, UCBerkeley:

$$\left\{\begin{array}{l}\text{Human judgments:}\\ \\ \text{goals, constraints,}\\ \text{preferences, knowledge, etc.}\\ \\ \\ \\ \end{array}\right. \Longrightarrow \begin{array}{l}\text{Autonomous (AI?) system:}\\ \\ \text{data, algorithms,}\\ \text{expectations,}\\ \text{intentions,}\\ \text{models,}\\ \text{performance functions, etc.}\end{array}$$

A next step:

- "society-in-the-loop" (Rahwan, MIT), a scaled up version of the human-in-the-loop, basically:

$$\left\{\begin{array}{lcl} \text{Human values:} & & \text{Autonomous (AI?)system:} \\ & \Longrightarrow & \\ \text{Rights, ethics, law,} & & \text{data, algorithms,} \\ \text{equity, fairness,} & & \text{human/social rights, privacy,} \\ \text{social contract, etc.} & & \text{models, performance functions, etc.} \end{array}\right.$$

This all, to be used effectively and efficiently, is heavily interdisciplinary, using results of:

- Neuroscience,
- Psychology (cognitive, social),
- Economics, social choice, etc.
- Linguistics (natural language processing, computational linguistics),
- Cognitive science,
- Computer science and IT/ITC (human-centered computing, artificial intelligence, human-computer interfaces, etc.),
- Systems science, etc.

We have therefore some foundation, which is a very reasonable approach, for a general dealing with our problems:

> *try to be as much as possible consistent with how the humans reason, perceive, judge, act, etc. to increase your chances of success.*

Now, and this will be our main concern here:

> *try to make use of available data to obtain information and knowledge in a way that would be natural and easy for the humans, to again increase your chances of success – we will call this data mining here.*

What is basically data mining?

There are many definitions, for instance:

- "...the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data ..." (Gartner Group).
- "...the analysis of observational data sets to find unsuspected relationships and to summarize data in novel ways ..." (Hand et al.)
- "...an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization ..." (Cabana et al.)
- etc., etc.

Therefore, there is:

- a large amount of data (maybe even “big data”),
- an obvious need for some summarization (humans have limited cognitive and information processing capabilities!),
- a need for a convenient and intuitively appealing presentation of results (i.e. visualization).

However, notice that the visualization is only mentioned

A long tradition:

> *. . . one picture is worth thousands words . . .*

On the other hand:

- For the human being natural language is the only fully natural means of articulation and communication,
- visualization can distract attention (e.g. in military applications, car navigation GPS systems, in which there are images and voice commands),

Very many powerful tools and techniques have been developed in “natural language technology”, for instance in:

- Computational linguistics,
- NLP, NLG, NLU – natural language processing, natural language generation, natural language understanding,
- etc.

But:

- natural language is inherently imprecise („fuzzy”) and the above traditional natural language technology tools have problems with handling imprecision.

The data mining process, meant to end up in the derivation of a linguistic summary, has many aspects, hence its performance is to be evaluated with respect to various criteria, often:

- novelty,
- correctness,
- generality,
- usefulness, and
- comprehensibility

Comprehensibility:

- concerned with whether the data mining result obtained is comprehensible (clear, understandable, . . .) to the human user.

Difficult!

Traditional data mining tools and techniques are not comprehensible per se!

We will advocate the use of linguistic data summaries as being comprehensible "per se"!

Comprehensibility in data analysis, data mining, knowledge discovery, machine learning, etc. has been recognized for a long time.

Presumably, first explicitly pointed out by Michalski (1983), one of the founders of modern machine learning, who has formulated in 1982 the so called postulate of comprehensibility:

> *"... The results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single 'chunks' of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion ...".*

Michalski's idea of comprehensibility has been considerably extended, for instance, Craven and Shavlik (1995) have formulated as the main reasons for the importance of comprehensibility in machine learning:

- to be confident in the performance and usefulness of the algorithms, and hence to be willing to use them,
- the users have to understand how the result is obtained and what it really means,
- etc. etc.

A natural solution: natural language!

Linguistic data summaries!

The very purpose of linguistic data summaries:

> *to summarize the very meaning of a (usually huge) set of (numeric) data via a simple and short statement(s) in (quasi)natural language,*

exemplified by:

> *"most young and highly qualified employees are well paid"*

in the case of a personnel database.

Notice that they:

- summarize the very essence of what we are interested in,
- are meaningful for any size of a data set,
- are "naturally comprehensible" (short sentences in natural language).

Here: linguistic data(base) summaries in the sense of Yager (1982), and Kacprzyk and Yager (2001), notably:

- dealt with in terms of Zadeh's protoforms (Kacprzyk and Zadrożny, 2005),
- viewed from the perspectives of computational linguistics and natural language generation (Kacprzyk and Zadrożny, 2010 – . . .).

Basically, the linguistic summaries are:

- assumed to be linguistically quantified propositions,
- a reflection of the usuality modality in natural language.

Linguistic Data Summaries: An Approach based on Fuzzy Logic with Linguistic Quantifiers

Here: a seminal approach to linguistic data summarization by Yager (1982), and Kacprzyk and Yager (2001), and Kacprzyk, Yager and Zadrożny (2000)

We have:

- V, a quality (attribute) of interest, e.g. the salary in a database of workers,
- a set of objects (records) y_i that manifest quality V, e.g. the set of workers; hence $V(y_i)$ are values of quality V for objects y_i,
- $Y = \{V(y_1), \ldots, V(y_m)\}$ is a set of m data items (the "database" in question).

A linguistic summary of a data set consists of:

- a summarizer S, i.e., a fuzzy predicate describing a property, simple or compound, of the objects of interest to the user, and which is possibly exhibited by a reasonable quantity (cf. the Q below) of objects (e.g., "young", "young and well paid", etc.),
- a qualifier K, i.e., a fuzzy predicate describing a range of objects pertaining to the summarizers (e.g. "young", "young and well paid", etc.);

- . . .
- a quantity in agreement Q given as a fuzzy linguistic quantifier (e.g. *most*), which expresses how many objects from among those satisfying a qualifier K exhibit the property expressed by the summarizer S,
- truth degree T, exemplified by 0.7, meant as the truth degree of a linguistically quantified proposition $Q_{y \in Y}(K(y), S(y))$ as, e.g.,

$$T(\textit{most young employees are well-paid}) = 0.7$$

Notice that:

- the very essence of a linguistic summary coincides with the very essence of any data mining approach: indicates what usually occurs (or course, it can be reformulated to indicate what rarely occurs),
- the use of the linguistic quantifier "most" is natural (a reflection of the usuality: in most cases),
- but, "most" goes beyond the classic quantifiers "for all" and "for at least one" we all know,
- therefore, some unorthodox calculus of linguistically quantified propositions should be applied to find the truth of a linguistic summary.

Zadeh's (1983) fuzzy logic based calculus of linguistically quantified propositions is the easiest way to calculate the truth value of the propositions:

$$Q_{y \in Y}\, S(y) \tag{1}$$

(e.g., "Most elements of Y possess property S") or, more generally,

$$Q_{y \in Y}\, (K(y)\, , S(y)) \tag{2}$$

(e.g., "Most elements of Y with property K possess also property S").

In Zadeh's (1983) classic approach, by far the most widely used, first, a (relative) fuzzy linguistic quantifier is equated with a fuzzy set in $[0,1]$ as, e.g.:

$$\mu_{\text{“most”}}(x) = \begin{cases} 1 & \text{for } x > 0.8 \\ 2x - 0.6 & \text{for } 0.3 \leq x \leq 0.8 \\ 0 & \text{for } x < 0.3 \end{cases}$$

meant as:
if less than 30% of the objects considered possess some property, then it is sure that not most of them possess it, if more than 80% of them possess the property, then it is sure that most of them possess it, and for the cases in-between, this is true (sure) to an extent, from 0 to 1, the more the percentage the higher the truth.

And we obtain

$$\mathrm{truth}(Q\,S(y)) = \mu_Q\Big(\frac{\sum \mathrm{Count}(S)}{\sum \mathrm{Count}(Y)}\Big) = \mu_Q\Big(\frac{1}{m}\sum_{i=1}^{m}\mu_S(y_i)\Big) \quad (3)$$

$$\mathrm{truth}(Q\,(K(y)\,,S(y))) \;=\; \mu_Q\Big(\frac{\sum_{i=1}^{m}(\mu_S(y_i)\wedge\mu_k(y_i))}{\sum_{i=1}^{m}\mu_K(y_i)}\Big) \quad (4)$$

where $m = \mathrm{card}(Y)$, $\sum \mathrm{Count}(A) = \sum_{y_i\in Y}\mu_A(y_i)$, $\sum_{i=1}^{m}\mu_k(y_i) \neq 0$, and $\wedge$ is a t-norm.

Notice that:

- the basic validity criterion, i.e., the truth degree T, is the most important and widely employed, and is comprehensible, indeed,
- the other quality criteria (Kacprzyk and Yager 2000): measure of informativeness, focus, imprecision, covering, appropriateness, are not comprehensible; the a length of a summary is.

A natural question: are these formulas VERY complicated ("big data"!)? Maybe not ...

Mining Linguistic Data Summaries through Fuzzy Querying: A Protoform based Analysis

A natural question: how to mine the linguistic summaries?

As proposed by Kacprzyk and Zadrożny (2001), a linguistic summary can be generated by using a fuzzy querying add on to a database (e.g. Kacprzyk and Zadrożny's, 1989, or 1995 – . . . FQUERY for Access) that supports queries with fuzzy linguistic quantifiers,for instance:

> *find employees such that most of: "age is young, salary is low, sex is male, residence is close, . . . " are fulfilled*

Clearly, the fuzzy queries with linguistic quantifiers directly correspond to the linguistic summaries so that a linguistic summary may be derived as follows:

- the user formulates a set of linguistic summaries of interest (relevance) using the fuzzy querying add-on,
- the system retrieves records and calculates the validity of each summary in question,
- a most appropriate linguistic summary is chosen.

A key role of Zadeh's protoforms

To make this derivation process effective and efficient, some standardized forms of linguistic summaries would be desirable, and this is provided by Zadeh's protoform viewed as an abstract prototype of a linguistic summary given by, for instance:

$$QY's \text{ are}S \tag{5}$$

(e.g., "Most elements of Y possess property S") or, more generally,

$$QKY's \text{ are } S \tag{6}$$

(e.g., "Most elements of Y with property K possess also property S").

By relating the linguistic summaries to the protoforms we attain a high degree of comprehensibility operating within the same structure of the protoform (linguistic summary) and just instantiating or generalizing a particular element of the summary. The user stays therefore within his or her area of expertise as a proper type of protoform of a linguistic summary that is comprehensible in a particular domain is used.

Therefore, the more abstract forms of protoforms correspond to cases in which we assume less about the summaries to be mined:

- assume a totally abstract (top) protoform, or
- assume that all elements of a protoform are given, i.e., all attributes and all linguistic terms expressing their values are fixed,
- assume something in-between.

Then:

- In the former case data summarization by full search would be extremely time-consuming, but might produce interesting, unexpected patterns,
- in the latter case the user guesses in fact a good candidate summary but the evaluation is simple, related to ad hoc queries.

Remarks on Some Implementations

We have applied the linguistic data summaries to many problems, for instance:

- linguistic summarization of sales data and relations at a computer retailer (implemented and still in use!),
- linguistic summarization of corporate innovation data,
- linguistic summarization of Web server logs.

Basically:

- the first and second example concerns static data,
- the third is mainly concerned with static data but extends the analysis to dynamic data (summarization of time series).

Various protoforms are employed in those examples but, in general, they are highly comprehensible and can be well understood by domain experts.

Linguistic summarization of sales data at a computer retailer

Our main implementation:

A computer retailer:

- 15 employees (sales, service, . . .),
- individual and corporate customers,
- computers, printers, accessories, network elements, components, software, etc.,
- services (repairs, ...),
- competition (Vobis, national chains).

Owner: young, open-minded, cooperative

Interested, e.g., in:

- the staff size on Saturdays,
- whether to concentrate on advertisement to larger or smaller customers,
- commissions from suppliers, etc.

But: very busy (no time), needs a summary but a simple one

Examples of linguistic summaries obtained:

- relations between the commission and the type of goods sold:
    - About 1/2 of sales of network elements is with a high commission
    - About 1/2 of sales of computers is with a medium commission
    - Much sales of accessories is with a high commission
    - Much sales of components is with a low commission
    - About 1/2 of sales of software is with a low commission
    - About 1/2 of sales of computers is with a low commission
    - A few sales of components is without commission
    - A few sales of computers is with a high commission
    - Very few sales of printers is with a high commission

- relations between groups of products and times of sale
  - About 1/3 of sales of computers is by the end of year
  - About 1/2 of sales of accessories is in the fall
  - About 1/3 of sales of network elements is in the beginning of year
  - Very few sales of network elements is by the end of year
  - Very few sales of software is in the beginning of year
  - About 1/2 of sales in the beginning of year is of accessories
  - About 1/3 of sales in the summer is of accessories
  - About 1/3 of sales of peripherals is in the spring
  - About 1/3 of sales of software is by the end of year
  - About 1/3 of sales of network elements is in the spring
  - Very few sales of network elements is in the fall
  - A few sales of software is in the summer

- relations between the size of customer, regularity of customer, date of sale, time of sale, commission, group of product and day of sale
  - Much sales on Saturday is about noon with a low commission
  - Much sales on Saturday is about noon for bigger customers
  - Much sales on Saturday is about noon
  - Much sales on Saturday is about noon for regular customers
  - A few sales for regular customers is with a low commission
  - A few sales for small customers is with a low commission
  - A few sales for one-time customers is with a low commission
  - Much sales for small customers is for non-regular customers

However in the above approach:

- only the own database is employed, as if the company operated in a vacuum.

In reality, however, companies operate in an environment:

- socioeconomic (customers' age, income level, education, taxes, etc.)
- natural (climate, natural attractions, etc.), etc.

Much impact on company, sales, decision processes, etc.!

Much data freely (or cheaply) available on the Internet!

For instance on weather: temperature and precipitation.

These external data can enhance the summaries!

For instance:

- relations between groups of products, times of sale, temperature, rain, size of customers, etc.
  - Very few sales of software is in hot days in the summer to smaller customers
  - About 1/2 of sales of accessories is in rainy days in the end of year
  - About 1/3 of sales of computers is to smaller customers in rainy days
  - About 1/3 of sales of accessories is during vacations on rainy Saturdays
  - Very few sales of computers is in the summer to small customers on hot days

Very good experience!

Linguistic summaries of corporate innovativeness

Basically:

- Purpose: to develop a human consistent, linguistic summarization based tool for the analysis of data related to the innovativeness of Polish companies (cf. Baczko, Kacprzyk and Zadrożny, 2011, 2012).

Some examples:

- “Most companies with high net revenues from sales in 2004 had also high total assets in 2004”
- “Most companies with high R&D activities in 2006 had also at least some in 2005”
- “Most companies with many patents registered in 2006 AND high R&D activities in 2005 had also high R&D activities in 2006”

Thus, in general:

- companies being active in the R&D field in 2005 did not necessarily continue to do so in 2006,
- however, those with some patents in 2006 usually also had RTD related activities in 2006.

Notice that the very structure of the linguistic summaries, i.e. their underlying protoforms, have an extremely high degree of comprehensibility for the domain experts specializing in innovations, economics, etc. Moreover, to a large extent they are comprehensible even to an average reader.

Linguistic summaries of Web server logs

A Web server log file may be directly interpreted as a table of data with the columns corresponding to the fields (attributes) he requests. In this section we will discuss various linguistic summaries that may be derived using that type of data as proposed by Zadrożny and Kacprzyk (2011, 2012, 2016).

Very simple, for clarity!

Static summaries

Examples:

- *Almost all* requested files are *small*
- *Almost all* failures concern “ppt” files
- *Most* of the requests concerning *large* files happen in the *evening*

The first summary may indicate that the maintenance of the archive of the Powerpoint presentations should be carried out more carefully.
The second may suggest that the large reports that the company makes available at its Web server should be updated, if possible, in the afternoon rather than in the morning to provide useful and timely information.

Very useful!

What can be main next developments:

- linguistic summaries of time series introduced in Kacprzyk, Wilbik and Zadrożny's, 2006 – 2012; Wilbik and Kacprzyk, 2012; Zadrożny and Kacprzyk; 2016,
- use of tools and techniques of NLG (natural language generation) for the generation of linguistic summaries as proposed in Kacprzyk and Zadrożny (2009 – 2016) to use widely available open source and commercial NLG products.

Concluding Remarks

We have presented:

- the concept of a linguistic data summary as a human centric (consistent) tools for capturing the very essence of (large) sets of data,
- some tools and techniques to mine linguistic data summaries, even for large data sets,
- examples of our real applications.

In general: the visualization is fine but should be augmented with the verbalization when the human being are involved because for them natural langauge is the only fuzzy natural means of articulation and communication!

All that: an example of a human-centric approach!

Even: **human-in-the-loop!**