Preserving Vagueness: The Central Mission of Next Generation Digital Humanities

„I do not dare to decide what is the truth about this matter, given the high darkness of this story“
(Dimitrie Cantemir, Description of Moldavia, 1752)

Walther v.Hahn
Computer Science Department, University of Hamburg
Hamburg, Germany
vhahn@informatik.uni-hamburg.de

Charles-University Prague • Sept.11, 2015

Contents

- Digital Humanities (DH) – current status
- Theoretical background of Humanities
- Vagueness
- Dimitrie Cantemir‘s „Descriptio Moldaviae/Beschreybung der Moldau“ 1752
- Challenges for DH
- Examples and modelling
- Next steps

Main Research Goals of DH

The use of elaborated CS methods for humanities in order to

data level

- make data accessible by computer (editing, tagging, browsing, retrieval),

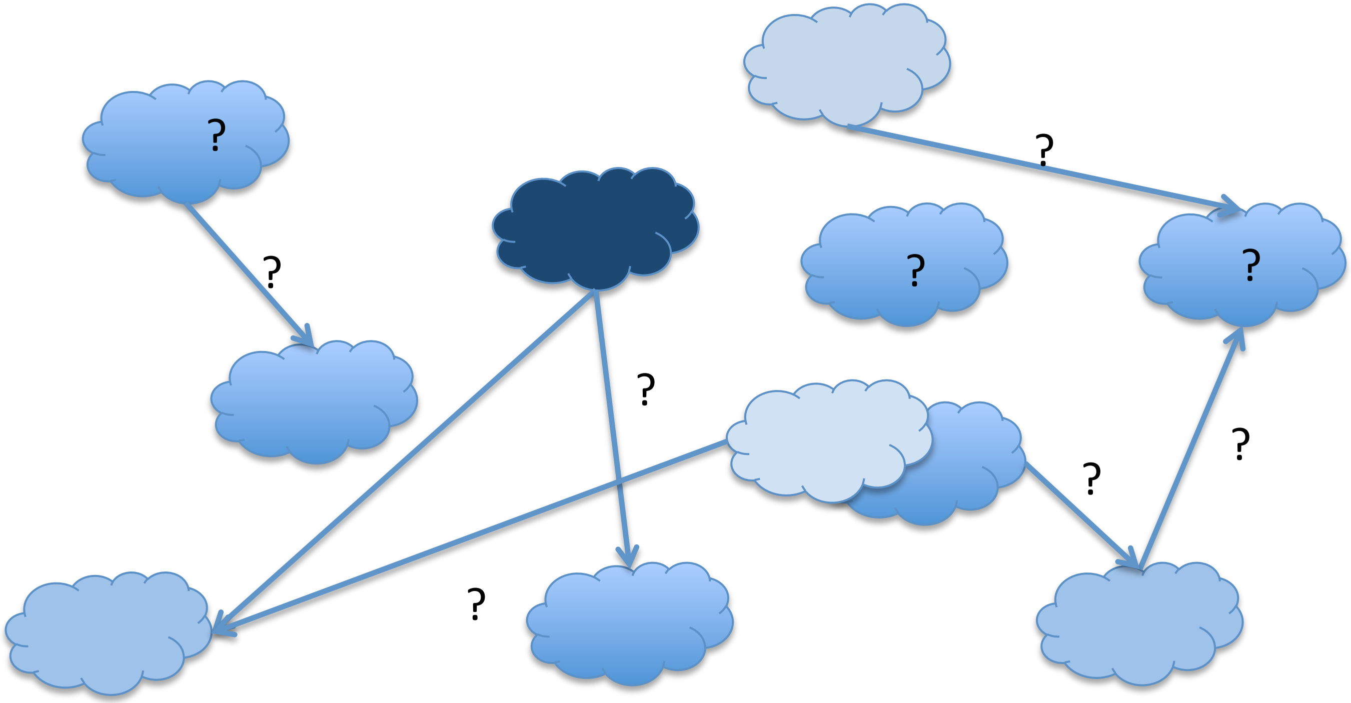
integration level

- link heterogeneous data, which differ, e.g., in
  - media (e.g. text, images, sounds, annotations),
  - time and eras, (e.g. time series, ages),
  - language and writing (e.g. multilinguality, scripts, transliterations),
  - areas (e.g. countries, locations, geo data),
  - [encoding].

interpretation level

- Investigate these data and achieve new interpretations
  in the humanities' research fields, <u>by applying specific humanities' methods.</u>

Humanities' Data: Vague, but relevant ...



DH‘s Favorite - Humanities‘ Nightmare:

Eliminating vagueness by omitting vague information

A	B	C	D	E
F	G	H	I	K
L	M	N	O	P

The Dagstuhl Seminar 2014: Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities.

“Further, it allows for analyzing much larger amounts of data in a quantitative and automated fashion – amounts of data that have never been analyzed before in the respective field of research. The question whether such steps ahead in terms of quantification lead also to steps ahead in terms of the quality of research has been at the core of the motivation of the seminar.“ …

“In particular, how can computer scientists convey the notion of uncertainties and processing errors to researchers in the humanities?”

“Which conditions influence the interpretability of the output generated by these algorithms from the point of view of researchers in the humanities? …

„It can be difficult for computer scientists to fully appreciate the concerns and research goals of their colleagues in the humanities. For humanities scholars, in turn, it is often hard to imagine what computer technology can and cannot provide, how to interpret automatically generated results, and how to judge the advantages of (even imperfect) automatic processing over manual analyse“

(Dagstuhl Seminar: Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities. Ed. by C. Biemann, G. R. Crane, Ch. D. Fellbaum, and A. Mehler.

One step ahead: Soft Computing

Hard Computing	Soft Computing
RIGID/CRISP/PRECISE	flexible /approximate
BI-VALUED	fuzzy-valued
TOTAL ORDER	partial order
ABSTRACT BASED	empirically (contextually) based
UNIQUE	hybrid/plural
NUMBERS	words

In: Seising, Rudolf and Veronica Sanz

Data quality forms a barrier to precision. However, the complexity of the human system of the First Global Age, 1400-1800,..., and the system's relation to geography constitutes a veritable fortress against numerical precision. Historians frequently resist any demands that they force their information into fixed categories required by a computational environment because so often, they feel that vital characteristics of some information will be lost when they do not fit precisely into a particular box. (J. B. Owens and Emery A. Coppola, Jr)

One step behind: DH

I believe that by conceptualizing our diverse cultural objects digitally and by developing analytical and representational tools and techniques for their investigation and manipulation that share the same fundamental principle, we have – possibly for the first time in human history – found a way to model the entire phenomenal spectrum from the concrete and material to the highly abstract by using one and the same 'language'. In other words, I believe that the digital is bound to become a sort of new lingua franca across the humanities, and perhaps even across all sciences.

In practice the Digital Humanities are methodologically defined by the principle of digital conceptualization of the objects and procedures of research. Who embarks upon Digital Humanities considers the objects of study implicitly as a complex of discrete measurable states, to apply, based upon this, computer based procedures: analytical, symbolizing or modelling. This mode of digital conceptualization of humanistic topics of research can in principle be used within all disciplines, as a digital lingua franca.

This fundamental methodological principle of digitality, so to speak, is discreteness.

(Jan Christoph Meister in: www.ssoar.info „DH is us or on the unbearable lightness of a shared methodology")

Science and Humanities

Wilhelm Dilthey (Einleitung in die Geisteswissenschaft, 1922):

Dilthey describes history as “a series of world views.” Man cannot understand himself through reflection or introspection, but only through what “history can tell him … never in objective concepts. Dilthey emphasizes the “intrinsic temporality of all understanding” i.e., that man’s understanding is dependent on past worldviews, interpretations, and a shared world.

Jürgen Habermas (Technik und Wissenschaft als Ideologie, Theorie des kommunikativen Handelns, 1968) distinguishes between purposive rational action and social action, the latter being the proper subject of humanities.

Jürgen Habermas’ concept and theory of communicative rationality distinguishes itself from the rationalist tradition, by locating rationality in structures of interpersonal linguistic communication rather than in the structure of the cosmos.

Consequences - 1 -

Humanities follow their own rationality (see Dilthey or Habermas), distinct from science and technology.

Todays humanities are often looking for higher public recognition and scientific acceptance from science and technology, Scholars from humanities are often skilled computer users and sometimes suffer from the weak precision of their topics and their intuitive techniques.

However: The integration of heterogeneous information and media (via concepts) distorts the data, because

- in most cases DH uses words instead of concepts and text sections instead of information bits,
- like texts, words are ambiguous or vague on several layers, esp. when being translated,
- simple annotations do not remedy the distortion,
- semantic string-tagging for humanistic interpretation often multiplies ambiguities (esp. in automatic tagging).

Consequences - 2 -

Ontologies suggest well-defined and systematic relations between well-defined concepts, but most research topics in humanities cannot be completely represented in a precise "technical" formalism, data often are in a "pre-research" status.

Grounding terms

- by using named entities does not change the situations, because NEs are not unambiguous by themselves (Istanbul or Constantinopol, Istria or Histria, Syrfia),
- even more titels (*Cesar*), names of empires (*Mesopotamia*), gods (*Astarte*), countries (*Walachia*), epochs (*Renaissance*) are vague,
- moreover, an elaborated time logic is needed, not even events with the same time stamp are necessarily synchronous, because they are discontinuously true („WWII was 1939 – 1945" „In AD 800 Charlemagne was crowned")

Annotations without reasoning (in OWL, e.g.) do not result in new knowledge, they only sum up what is written into the annotations.

Different knowledge classes (historical data, texts, images, beliefs, traditions, legends, rumors) behave differently when included into an inference chain.

What are Ontologies (according to Guarino)

- An ontology is a formal specification of a conceptualisation,

  not only

  - an annotation of words in a text, or
  - a set of „symbols“, but technically words, which are called „concepts“,
- The aim of an ontology is to derive, what is possible in a given domain,

  not only

  - to provide a vocabulary for annotation, or
  - to provide conceptual dependencies among words/concepts
- Main problems for DH:
  - how to abstract from words to language independent concepts?
  - how to write rules for processing the ontology?
  - what about alternative or concurrent ontologies?

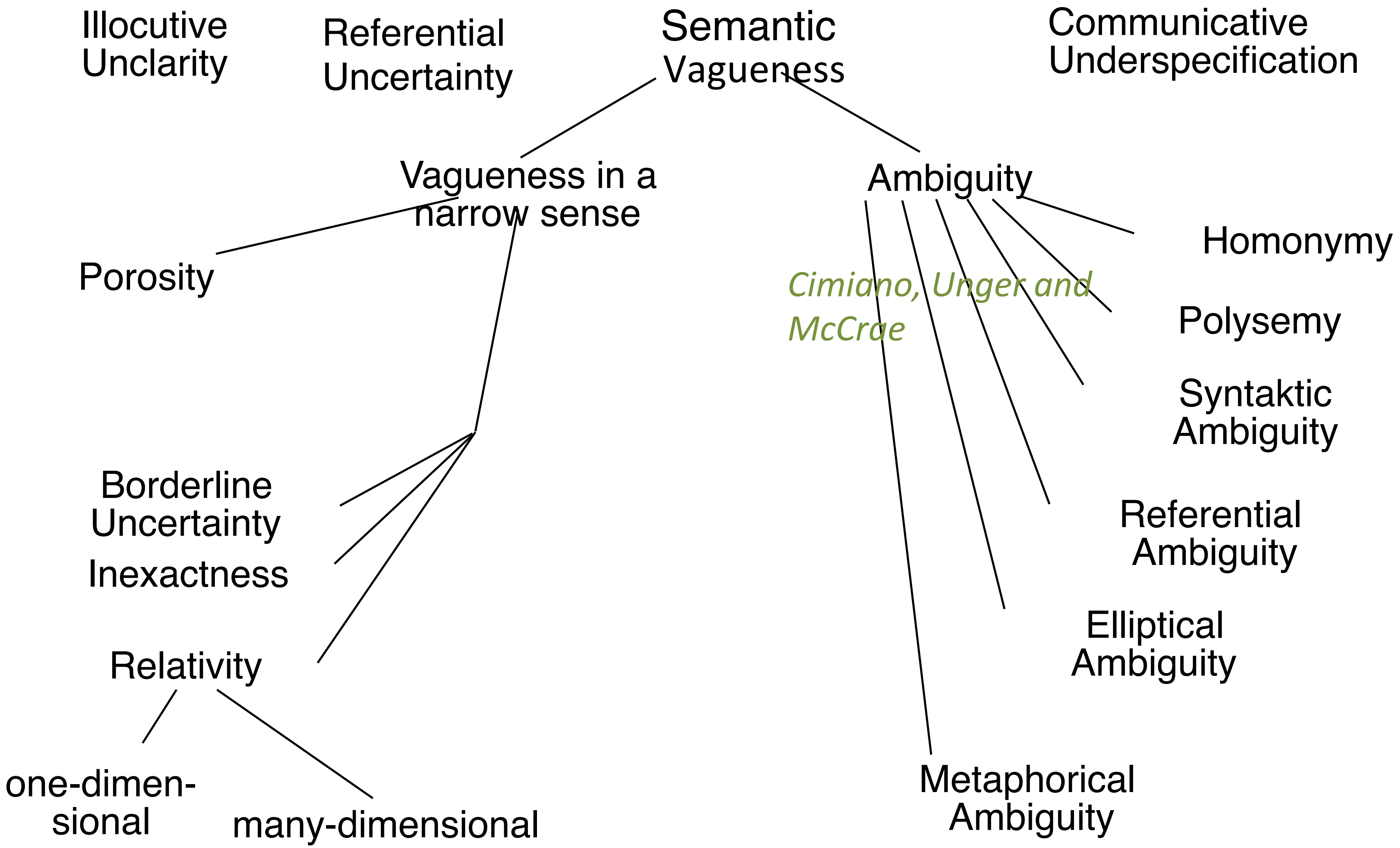
Chances for success of DH

- **Progress by**
  - better coverage (better statistics), whatever the statistics measure,
  - systematic processes (versus introspection),
  - using field specific annotations, reasoning and ontologies to obtain “new” knowledge,
  - Anyway international cooperation with computerized research infrastructure.
- **Preconditions:** Serious discussion within the (interdisciplinary) team about
  - field dependent cognitive interest (results do not emerge from the data)
  - Modelling (what is a historical event, a city, a date, etc?),
  - Formally reliable annotation with vagueness
    - anchors and tags
    - verification
  - Inferencing on an ontology and axioms in the domain,
  - Field dependent visualization for informal consistency estimation,
  - evaluation formalism and evaluation of results.
- **Result:** Acceptance of the humanistic character of the result: The result is still a social-historical interpretation of life experience.

Historical Remarks about Vagueness

- The formalisms for representing fuzziness in a rule-based way found their first technical application in Artificial Intelligence, esp. with Lotfi Zadeh‘s proposal of a fuzzy set theory, where the membership degree to a set is given in real numbers from 0 to 1. Later on, the notions of vagueness, uncertainty, credibility, and salience have been discussed and introduced in processing
- Vagueness in linguistics was first summarized for German by Manfred Pinkal:

M.Pinkal's Schema of Semantic Vagueness



Case Study: Dimitrie Cantemir‘s „History of Moldavia“ 1752

Example of a ‚Digital Humanities‘ Project

- Cantemirs Book was written in Latin, contains citations from
  - Romanian
  - Ancient Greek,
  - Turkish,
  - Tatar,
  - German,
  - Polish.
- quotes
  - various authors from Greek, Latin, German,
  - assumptions,
  - folk-belief,
  - myths,
  - etc.
- Translations from different times into
  - German, Romanian, Russian

What was known in central Europe about Moldavia in VIII. and XIX. century?
What was known about Turkish provinces at that time?

1. Challenge for DH: Mappings

Among

- historical text editions,
- translations,
- historical text and modern language,
- historical texts and historical maps,
- historical texts and modern maps (see „GeoNames forum“),
- texts and „official“ history,
- different source types,
- multilingual text elements.

2. Challenge for DH: Vagueness on several layers

- **Linguistic vagueness**
- **Fuzzy concepts**
  - “Before Stephan the Great, *all mountains around* Moldavia belonged to Transsilvania and the country was *narrow* on this side”…
- **Fuzzy maps or regions** Example: “Syrfia”,
- **Vague or concurrent ontologies:**
  - The Turkish and the Moldavian administration
- **Uncertain facts**
  - The origin of the hill “Chan Tepesi” or “Mogila Rabuy”
- **Naïve History** (derived from ‘naive physics’)
  - „The Roman Empire conquered Dacia“

Deep processing of historical texts requires handling of both challenges

Vagueness in historical texts: Lexical semantics

- obsolete words occur in texts,
- lexical semantics of known words changes over time,
- idioms change
- false friends over periods (Germ. *übel*, *wohl*),
- even technical terms change (Germ. *Verleger*),
- references might be wrong and might look like a translation error,

  See position of Constantinopel in Schedel's Nuremberg Chronicle
- Is *Istanbul* always the correct translation of *Konstantinopel*?

Vague mapping in historical text data: Orthography and Scripts

- Orthography over time:
  - For a long period there was no orthography, not even stable writing rules within a document:
    - Normal text sorting
    - writing error detection and
    - Retrieval

    will fail,
  - Arbitrary abbreviations: Normal expansion tools will fail ,
  - Change of scripts in Romanian or Turkish
  - Mixture of scripts: Latin, Black-letter Italics
  - Illegible sections: Transcription will fail.

Difficult mapping in historical texts: Multilinguality

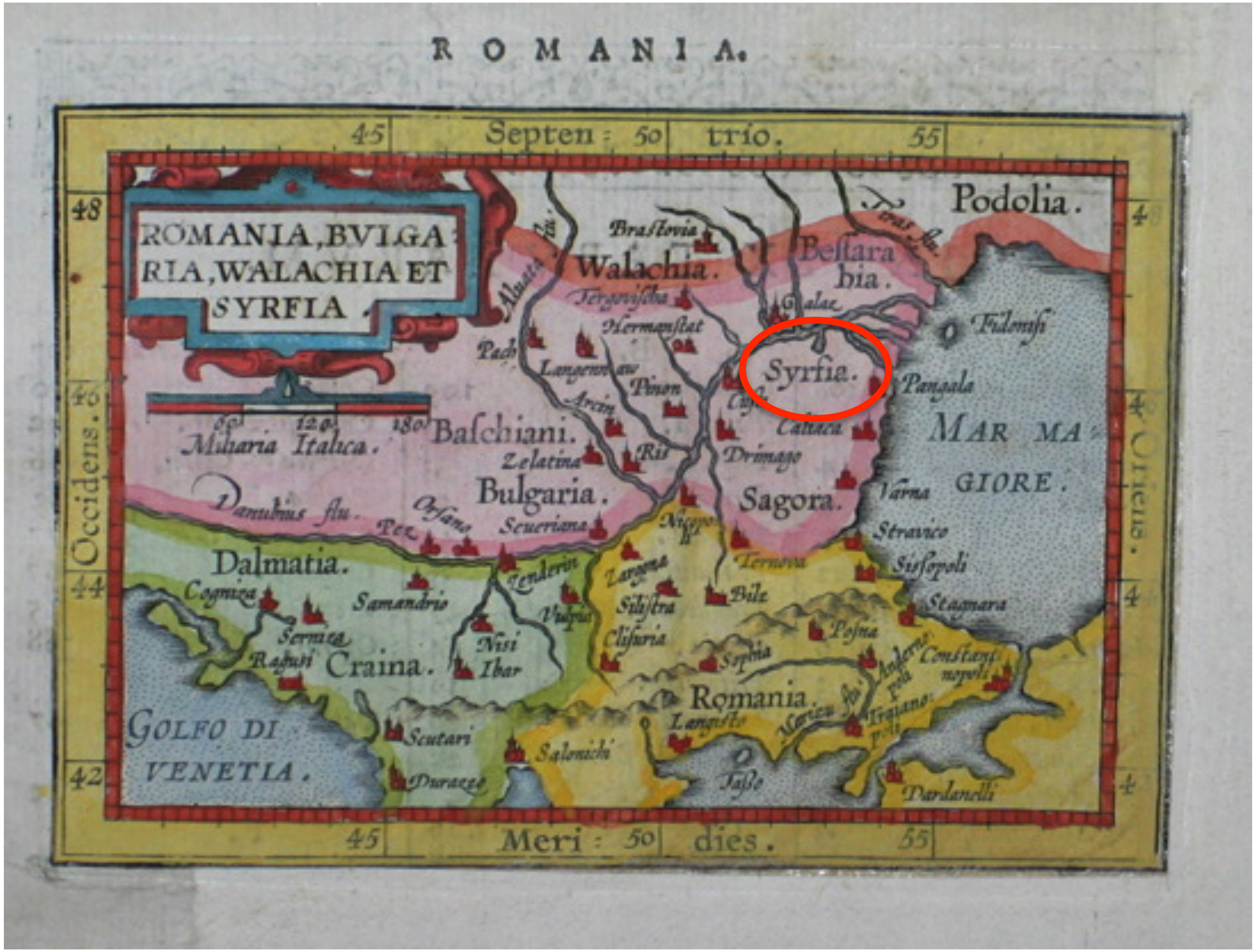
- Several languages with changing scripts are mixed in one document: Even digitization will produce errors,
- Historical languages are difficult for normal language tools,
- Translations are lexically not comparable over time.

Vague mapping in historical texts and maps: Named Entities

- Approximative “translations” of named entities,
- References might be wrong and might look like translation errors

  See position of Constantinopel in Schedel’s Nuremberg Chronicle
- Is *Istanbul* the correct translation of *Constantinopel*?
- What is *Marmora* in Cantemirs Historia …

Example: Ortelius‘ map 1570



Example for mapping maps: Syrfia

On the map of Ortelius and his successors there is a region called „Syrfia“ to the south of the Danube delta.

Wikipedia.rom says:

> ***Syrfia*** *is*
>
> *the abandoned name of a region in Eastern Europe, used on historical maps until 17th century, designating*
>
> *a part of Northern Dobrudja, coming from the Greek term Σύρφοι - Syrphoi, or*
>
> *The Cojani region from western Macedonia, today in Greece but in turkish times in the “Serfia sangiac” having the capital Σέρβια, Servia ;*
>
> *Sârbia, due to phonetic association.*

Modelling challenges -1-

Given an ontology (with the name space OntoHist) with classes and properties.

We create instances of regions:

```
<owl:NamedIndividual rdf:about="&OntoHist;SyrfiaNordDobrudja">
    <rdf:type rdf:resource="&OntoHist;Region"/>
<OntoHist:part_of rdf:resource=”&OntoHist;NorthDobrudja”>
<rdf:label xml:lang=”tr”>Sarf</rdf:label>
<rdf:label xml:lang=”gr”> Σύρφοι </rdf:label>
<rdf:label xml:lang=”ro”> cotitura Dunarii</rdf:label>
</owl>
```

Modelling challenges -2-

For Cojan region

```
<owl:NamedIndividual rdf:about="&OntoHist;CojanMacedoniaWest ">
<rdf:type rdf:resource="&OntoHist;Region"/>
<OntoHist:is_situated_now rdf:resource=”&OntoHist;Greece”>
<OntoHist:is_situated_old rdf:resource=”&OntoHist;SangiacSerfia”>
</owl:Namedindividual>
```

Where for Sangiac holds:

```
<owl:NamedIndividual rdf:about="&OntoHist;SangiacSerfia ">
<rdf:type rdf:resource="&OntoHist;HistoricalRegion"/>
<OntoHist:located_in rdf:resource=”&OntoHist;OttomanEmpire”>
<OntoHist:has_capital rdf:resource=”&OntoHist;Servia>
</owl>
```

and for Servia holds:

```
<owl:NamedIndividual rdf:about="&OntoHist;Servia ">
<rdf:type rdf:resource="&OntoHist;HistoricalCity"/>
<owl: rdf:resource="&OntoHist; Σέρβια"/>
</owl>
```

Modelling challenges -3-

The sameAs property does not really model the „phonetic confusion“ and we would need a fourth parameter for the fuzzy value

```
<owl:NamedIndividual rdf:about="&OntoHist;Sarbia">
<rdf:type rdf:resource="&OntoHist;Country"/>
<owl:sameAs rdf:resource="&OntoHist;Serbia"/>
</owl>
```

Summing up for Syrfia:

```
<owl:NamedIndividual rdf:about="&OntoHist;Syrfia">
<rdf:type rdf:resource="&OntoHist;Denomination"/>
<rdfs:label xml:lang="de">Bezeichnung</rdfs:label>
<OntoHist:used_on rdf:resource="&OntoHist;OldMapEuropa"/>
<OntoHist:used_when rdf:resource="&OntoHist;17Century"/>
<OntoHist:denominates    rdf:resource=”&OntoHist;SyrfiaNorthDobrudja”>
<OntoHist:denominates    rdf:resource=”&OntoHist;Sarbia”>
<OntoHist:denominates    rdf:resource=”&OntoHist;CojanRegion”>
</owl:NamedIndividual>
```

Modelling challenges -4-

Problems:

- We need an instance

```
<owl:NamedIndividual rdf:about="&cantemir;17thCentury">
       <rdf:type rdf:resource="&cantemir;TimeSpan">
 </owl:NamedIndividual>
```

  What is the granularity of time spans, how many of them do we need?

- Which Sârbia is a phonetic association to Syrfia?

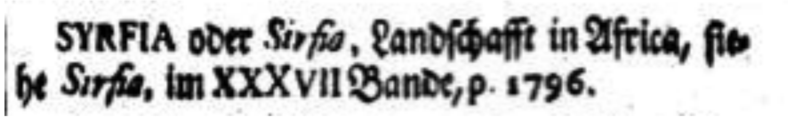
- How should the instance „Ottoman Empire“ be modelled?

A more general question is, how to represent fuzzyness in the RDF-triple formalism

```
<owl:NamedIndividual rdf:about="&OntoHist;Sarbia">
<rdf:type rdf:resource="&OntoHist;Country"/>
<owl:sameAs rdf:resource="&OntoHist;Serbia"/>
<rdf:comment>Probability 0,10</rdf:comment>
</owl>
```

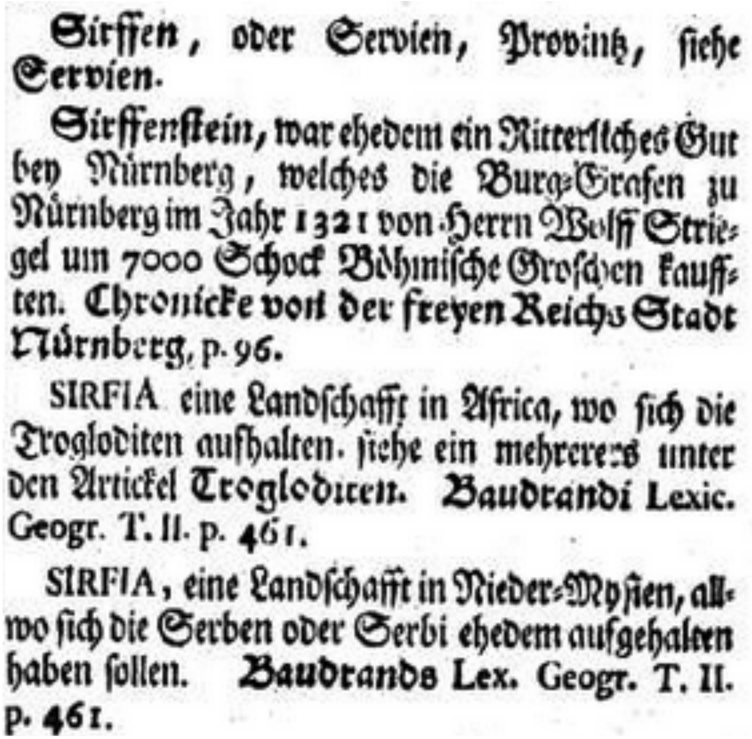
Here it is impossible to specify the scope of the fuzzyness

Comparing Older Lexical Source: Zedler, Universallexicon 1731 ff.



SYRFIA oder *Sirfia*, Landſchafft in Africa, ſiehe *Sirfia*, im XXXVII Bande, p. 1796.

↓



Sirffen, oder Servien, Provintz, ſiehe Servien.

Sirffenſtein, war ehedem ein Ritterliches Gut bey Nürnberg, welches die Burg-Grafen zu Nürnberg im Jahr 1321 von Herrn Wolff Striegel um 7000 Schock Böhmiſche Groſchen kaufften. Chronicke von der freyen Reichs Stadt Nürnberg, p. 96.

SIRFIA eine Landſchafft in Africa, wo ſich die Trogloditen aufhalten. ſiehe ein mehreres unter den Artickel Trogloditen. Baudrandi Lexic. Geogr. T. II. p. 461.

SIRFIA, eine Landſchafft in Nieder-Myſien, allwo ſich die Serben oder Serbi ehedem aufgehalten haben ſollen. Baudrands Lex. Geogr. T. II. p. 461.

SYRFIA oder Sirfia, Landschaft in Africa, siehe Sirfia, im XXXVII. Bande p.1796

↓

Sirffen, oder Servien, Provintz, siehe Servien.

Sirffenstein, war ehedem ein ritterliches Gut bey Nürnberg, welches ...

SIRFIA eine Landschaft in Africa, wo sich die Troglodíten aufhalten. siehe ein mehreres unter den Artickel Troglodíten. Baudrandi Lexic. Geogr. T.II. p. 461.

SIRFIA, eine Landschaft in Nieder Mysien, allwo sich die Serben oder Serbi ehedem aufgehalten haben sollen. Baudrands Lex.Geogr. T.II. p.461.

Consequences

- Does vagueness in DH require a heavy-weight fomalism?
- No, the fuzzyness on different layers is not comparable or computable in rules
- Possible solutions are
  - To indicate existing fuzzyness,
  - To give a ranking of the values/layers,
  - arrange corresponding visualization of [certainty|vagueness|credibility|plausibility].

Explicite Preciseness in Cantemirs „Descriptio Moldaviae...“

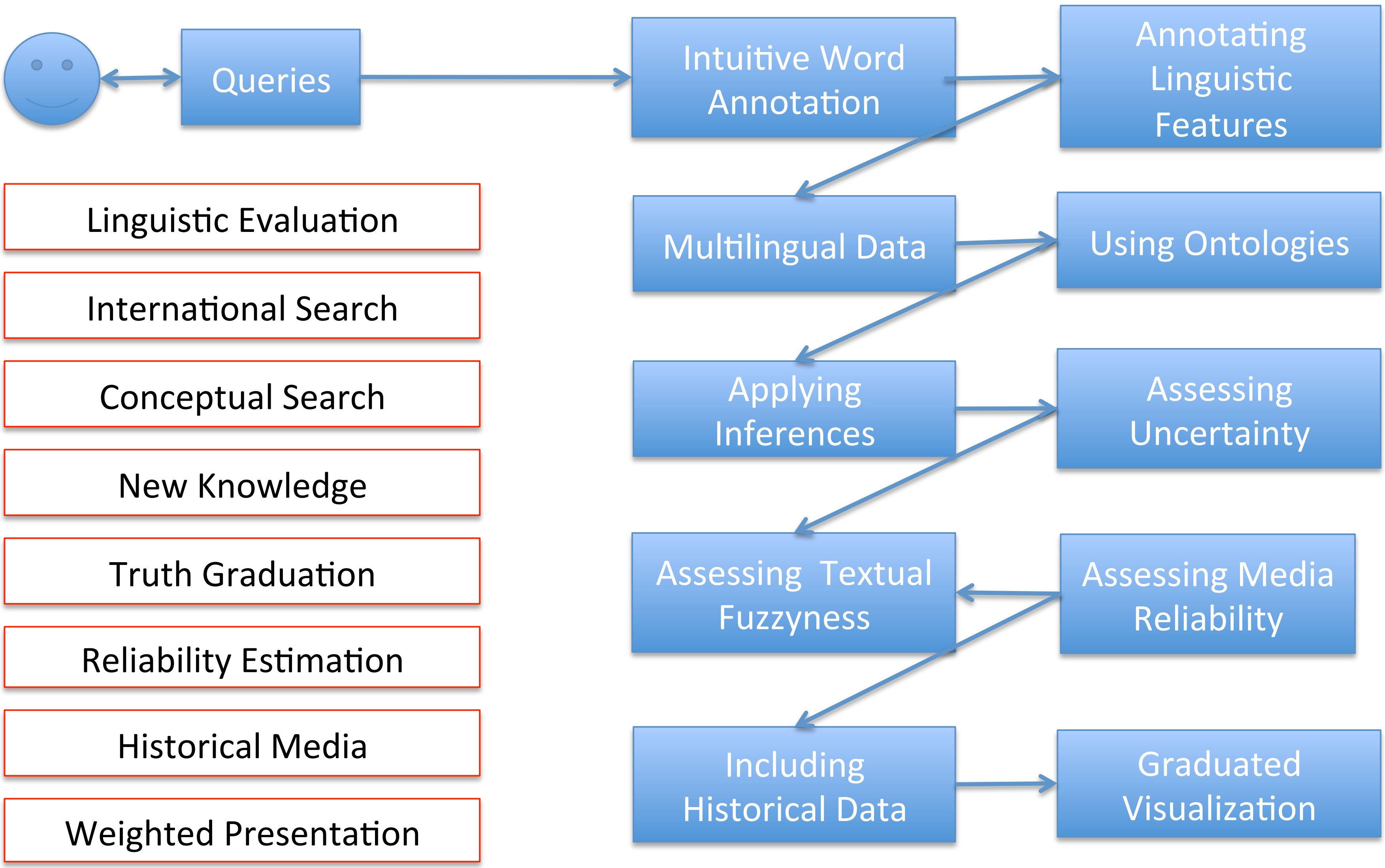
Da überdies alle Geschichtschreiber unserer Nation einstimmig bezeugen, daß die Moldauer, nachdem sie aus Maramor in ihr altes Vaterland zurückgekehrt, bevestigte Städte und Schlösser, die keine Einwohner halten, angetroffen haben , so laßt sich nicht ungereimt hieraus schlüssen, daß ihre Errichtung in weit entferntere Zeiten zurückzusetzen sey. Dieses bestätigt, ausser andern Umstanden, die Bauart der Mauern in den meisten Städten, welche gewiß keine andere als die Römische seyn kan, einiger wenigen ausgenommen, von welchen wir oben angezeigt, daß sie in neuern Zeiten wider die Einfalle der Tatarn errichtet zu seyn scheinen. Alles aber übertreffen die Zeugnisse bewährter römischer Geschichtschreiber, durch welche ausgemacht ist, daß der römische Kaiser Trajan grosse Colonien von römischen Bürgern nach Dacien geführt habe, und daß sein Nachfolger Hadrian, als er verschiedene im Oriente gelegene Gebiete an die Barbaren abgetreten, bloß aus Furcht vor derselben Verwüstung zurückgehalten worden sey, Dacien zu verlassen. Dazu kommt noch ein ewiges Denkmaal von dieser Sache, nehmlich Der Graben des Kaisers Trajan.

(Sounds like RST ...)

Explicite Vagueness in Cantemirs „Descriptio Moldaviae“

- „Daß unsere Landsleute, als sie wieder zurückkehrten, sich hier zu erst sollen niedergelassen haben, glauben sehr viele, aber nicht alle. Denn man zeigt nicht weit davon an dem östlichen Ufer des Sireth einen andern Ort, welcher heutiges Tages von den Einwohnern Smedorowa genennet wird, und giebt vor, daß daselbst die erste und größte Stadt sey angelegt worden. (p. 54)“
- „Ich las einmal in einer Handschrift von der Historie des Herodots, daß an dem Pruth, drey Tagreisen von der Donau, die so kriegerische Nation der Taiphalier gewohnt, und eine sehr grosse Stadt erbauet habe. (p. 58)“
- „Nicht weit von diesem Orte sieht man einen von Menschenhänden errichteten grossen Hügel, welcher bey den Tatarn Chan Tepesi d.i. des Chans Hügel, bey den Einwohnern Mogila Rabuy genennt wird. Von seinem Ursprunge giebt es verschiedene Meynungen. Einige Tatarn geben vor, daß ein gewisser Chan mit seiner ganzen Armee von den Moldauern daselbst zu Grunde gerichtet, und zum Andenken dieser Hügel aufgeworfen worden sey: andere erzählen, daß eine gewisse scythische Königin, Rabie genannt, an diesem Orte, als sie mit ihrer Armee gegen die in der Moldau wohnende Scythen auszog, erschlagen und von ihren Leuten allda begraben worden sey. Was an der Sache wahr oder falsch sey, unterstehe ich mich nicht bey einer so grossen Dunkelheit dieser Geschichte auszumachen.“(p. 60)

Bad Practice for „Avoiding Vagueness“



Queries

Intuitive Word Annotation

Annotating Linguistic Features

Multilingual Data

Using Ontologies

Applying Inferences

Assessing Uncertainty

Assessing Textual Fuzzyness

Assessing Media Reliability

Including Historical Data

Graduated Visualization

Linguistic Evaluation

International Search

Conceptual Search

New Knowledge

Truth Graduation

Reliability Estimation

Historical Media

Weighted Presentation

References

Dagstuhl Seminar: Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities. Edited by Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler

Dilthey, Wilhelm, Einleitung in die Geisteswissenschaft 1883

Cimiano, Ph., Unger, Chr. und McCrae, Ontology-Based Interpretation of Natural Language. San Rafael 2014

Guarino, Nicola, Some ontological principles for designing upper level lexical resources. In: Proceedings LREC 1998.

Habermas, Jürgen, Technik und Wissenschaft als Ideologie 1968

v.Hahn, Walther, Vagheit bei der Verwendung von Fachsprachen. In: Hoffmann / Kalverkämper / Wiegand: Fachsprachen. Band 1. Berlin 1998. S. 383 - 390.

Pinkal, Manfred, Logik und Lexikon: Die Semantik des Unbestimmten. Berlin 1985

Pinkal, Manfred, Semantische Vagheit: Phänomene und Theorien. In: Linguistische Berichte 70. 1980. 1-26. und 72. 1981. 1-26.

Seising, Rudolf and Veronica Sanz (Eds.) Soft computing in Humanities and Social Sciences. Berlin 2012

J. B. Owens and Emery A. Coppola, Jr, WHITE PAPER .Fuzzy Set Theory (or Fuzzy Logic) to Represent the Messy Data of Complex Human.

- .