Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Arbeitsstelle „Computerphilologie“

Representing and Processing Vague Data from Humanities

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

Walther v.Hahn, Cristina Vertan
Computer Science Department, University of Hamburg
Hamburg, Germany
{vhahn, vertan}@informatik.uni-hamburg.de

Universidad de Granada • May.11, 2017

Contents

- Challenges of representing and processing data from humanities
- Vagueness
- Levels of Vagueness

v. Hahn

- Example: Historical Works of Dimitrie Cantemir
- The project HerCoRe: Hermeneutic and Computer based Analysis of Reliability, Consistency and Vagueness in historical texts

Vertan

Humanities and Science I

- Whereas Science (at least above the molecular range) explains data of high reliability and predictive power (according to laws of nature),
- Humanities analyse historical data relative to the contemporary understanding of life, this means, that
  - there are no facts *like bank account numbers,*
  - there is no self-understanding use of the data *like withdrawing money,*
  - interpretation is always disputable*, in contrast to the law of gravitation.*
  - reasoning in humanities follows approximate and vague conclusion rules.
- Processing humanities' data has to leave open the final interpretation by the researcher observing the rules of "hermeneutics" (cf. Wilhelm Dilthey, Einleitung in die Geisteswissenschaften, 1883.

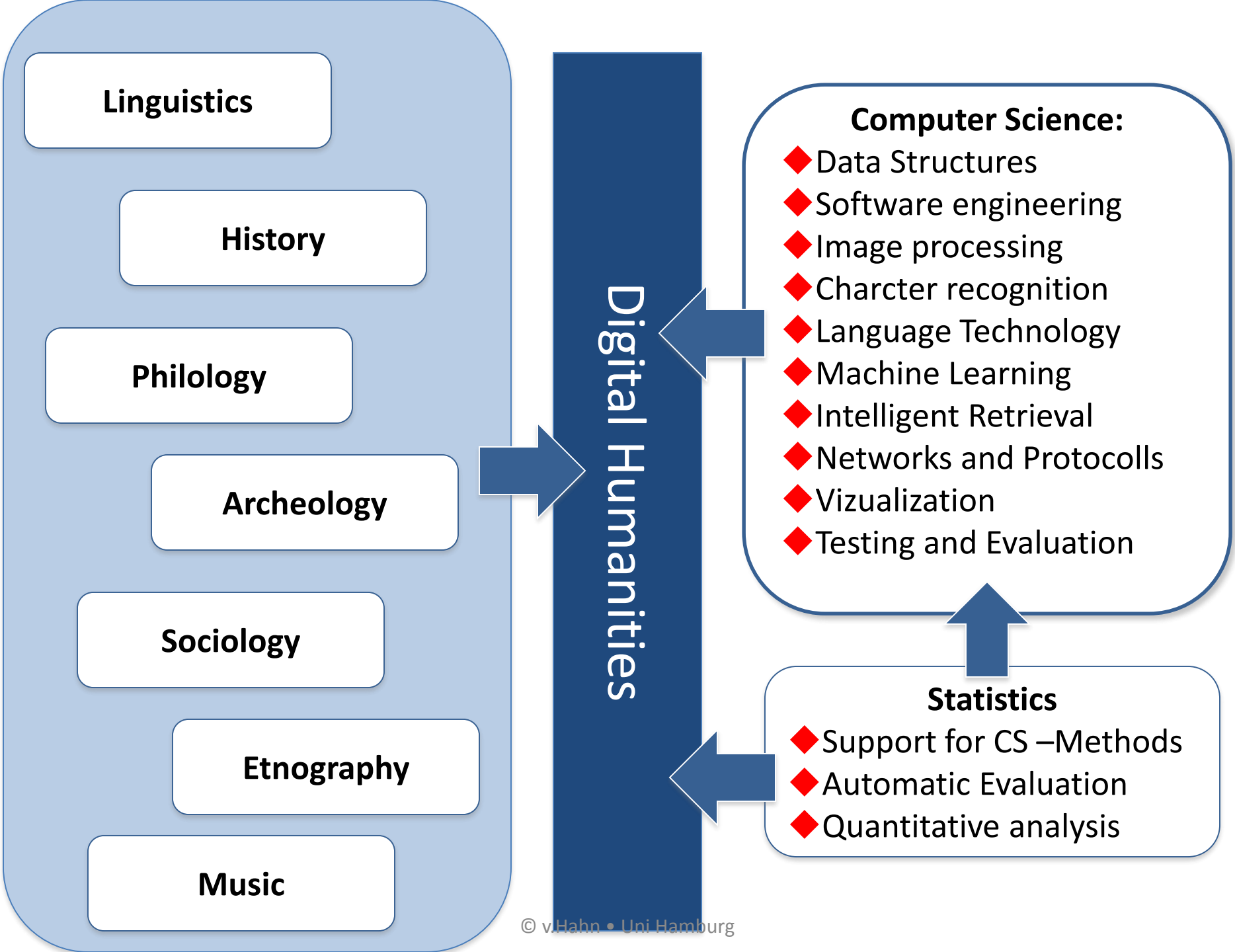
Science and Humanities II

Wilhelm 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 world views, 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.



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, by applying specific humanities' methods.

Workflow in Digital Humanities
- Desiderata -

- All stages of research involve researchers from computer science and humanities:

humanities researcher

- Provide specifications and require-ments list (they are end user),
- are acquainted with the state of the art in their field,
- define their research interests,
- explain the specifics of their data,
- report about the results of the CS methods,
- evaluate the user interface.

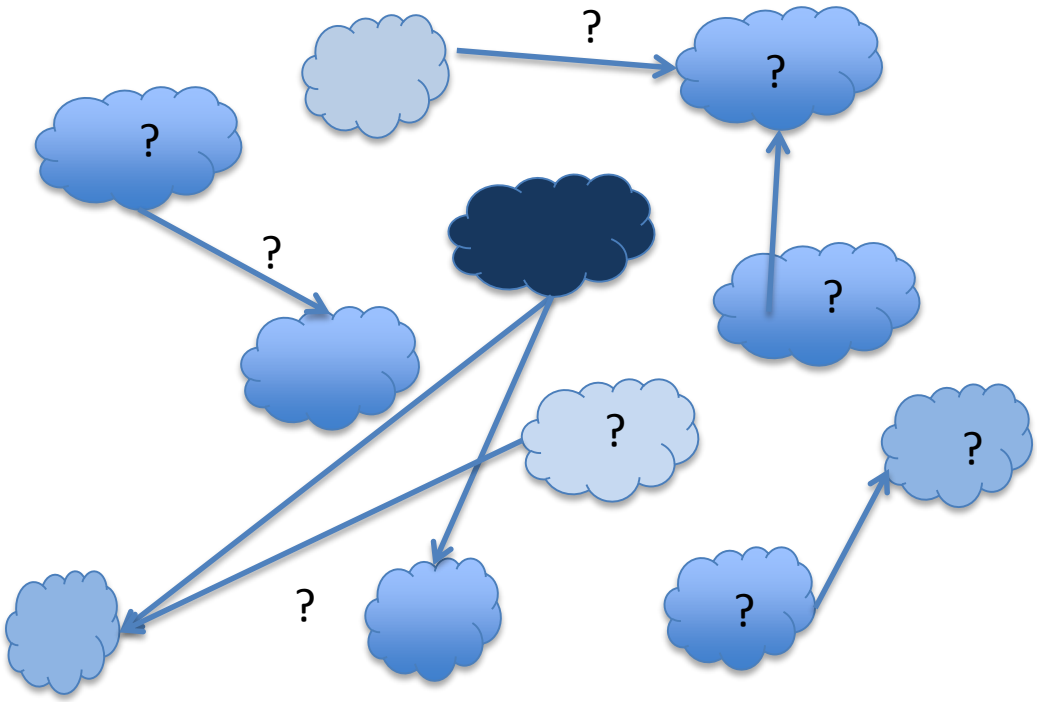
Computer Science Researcher

- are acquainted with the state of the art in CS and evaluate the usability of methods,
- implement new methods according to recent developments in CS,
- adapt existing systems according to user specifications,
- develop data models,
- evaluate performance.

DH means **research in Computer Science,** not only programming

DH means **research in humanities,** not only forced adaptation of data to existing systems or formalisms.

Humanities‘ Data: Vague, but relevant ...



DBMS‘ Favorite - Humanities‘ Nightmare:

Eliminating vagueness by omitting vague information

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

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

Table 1.1. Hard versus soft Computing.
In: Seising, Rudolf and Veronica Sanz (Eds.)
Soft Computing in Humanities and Social Sciences. Berlin 2012, p.25

Consequences - 1 -

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

Modern 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 “facts”, or “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 annotation techniques 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 situation, 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?

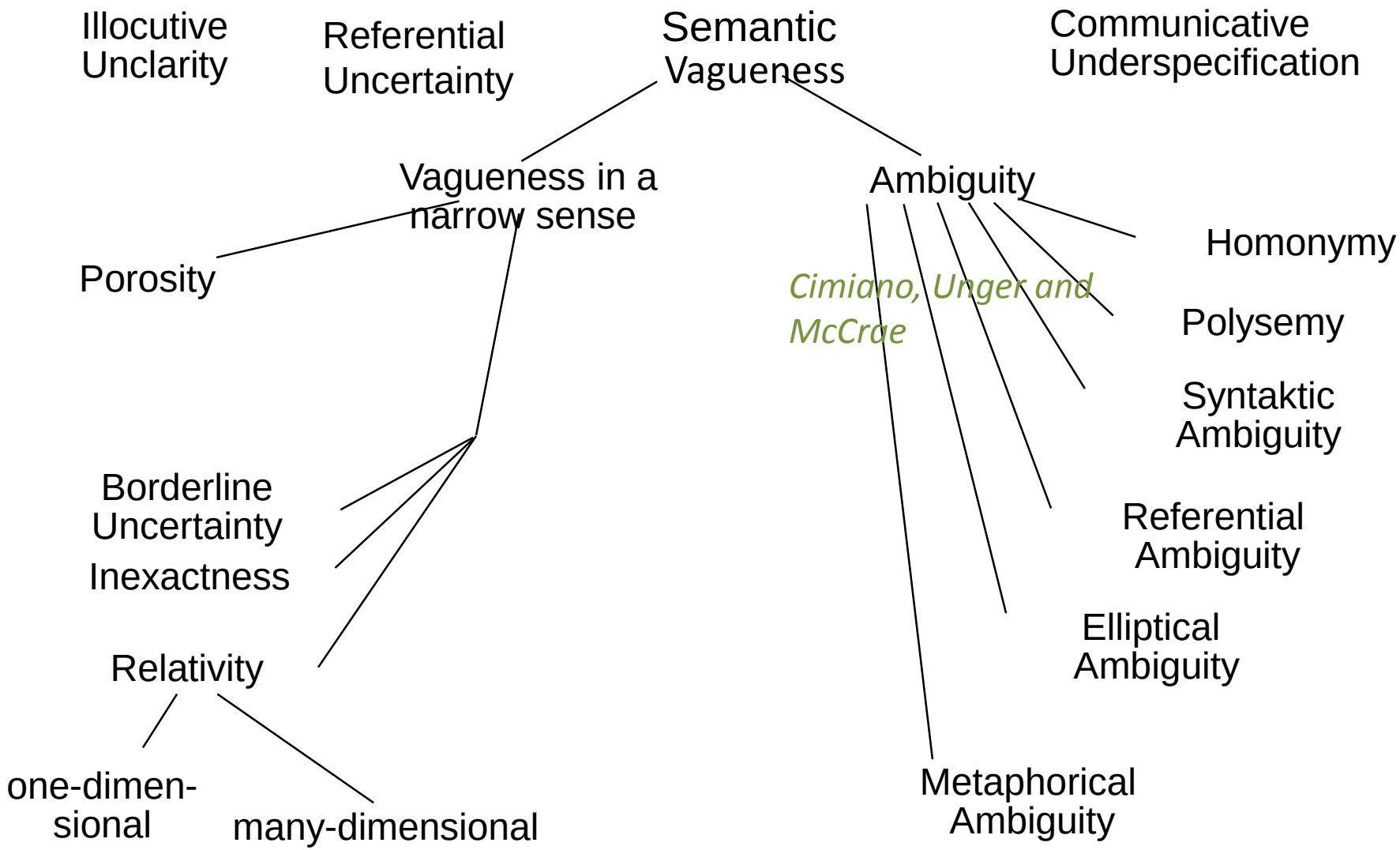
Chances for success of DH

- **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



Challenge for DH: Vagueness on several layers

- **Linguistic vagueness**
- **Fuzzy concepts**
  - "Before Stephan the Great, *all mountains around* Moldavia belonged to Transilvania 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"

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 Constantinopol in Schedel's Nuremberg Chronicle
- Is *Istanbul* always the correct translation of *Constantinopol*?

Vague mapping in historical text data: Orthography and Scripts

- Orthography over time:
  - For a long period there was no orthography in European languages, not even stable writing rules within a document. Standard tools for
    - 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/Translation will fail.
- Every Transliteration /transcription is already an interpretation and thus a candidate for vagueness.

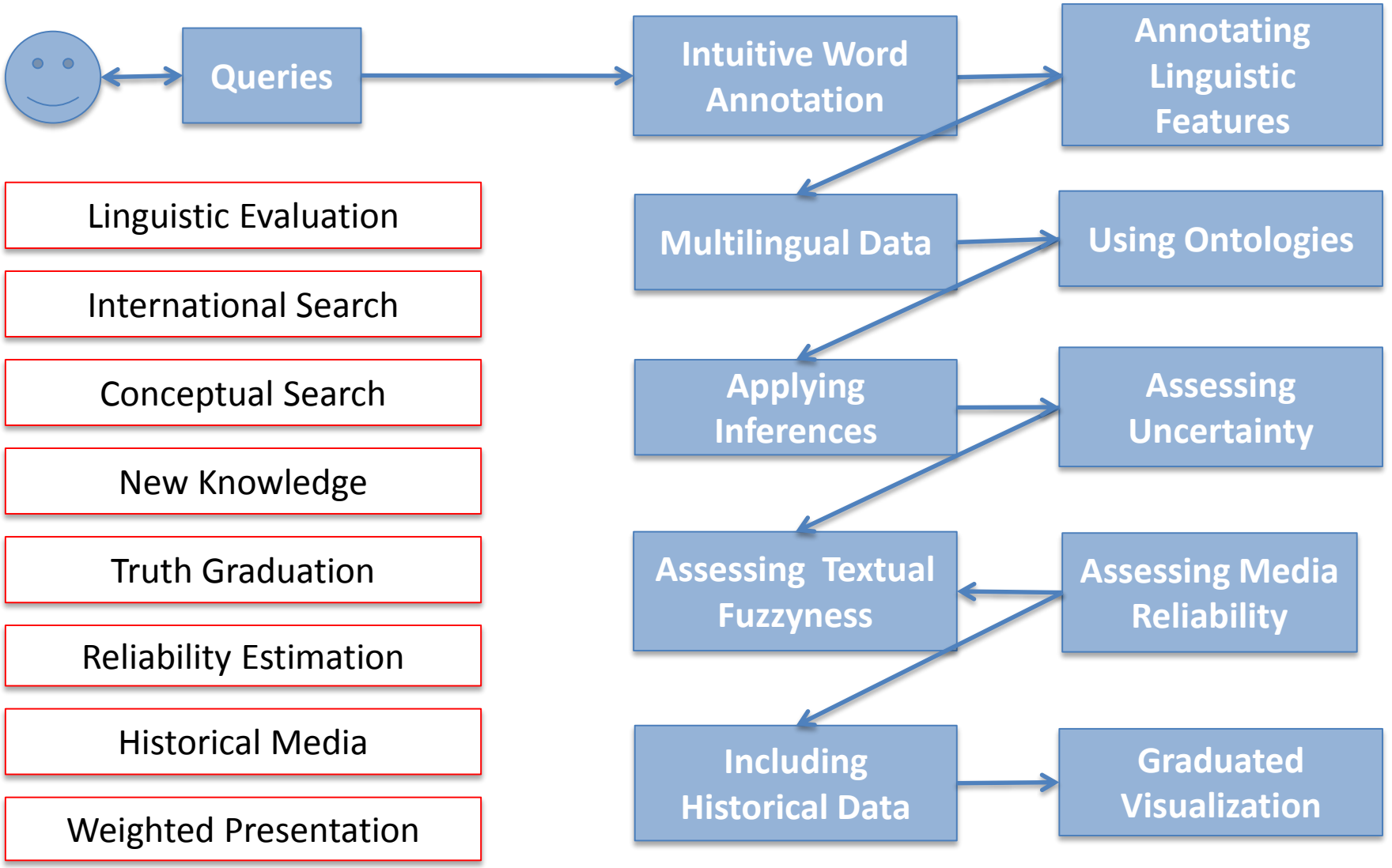
Vague mapping in historical texts and maps: Named Entities

- Approximate “translations” of named entities: “Milano” ➔ “Mailand”,
- 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 “*Marmor”* in D. Cantemirs “Descriptio Moldaviae”

Example: Ortelius‘ map 1570

Bad Practice for „Avoiding Vagueness“



Lots of work ahead!

- Thank you for your attention!