DataSys 2015
June 21 - 26, 2015 - Brussels, Belgium

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

bigM
BioInformatics Group Mittweida

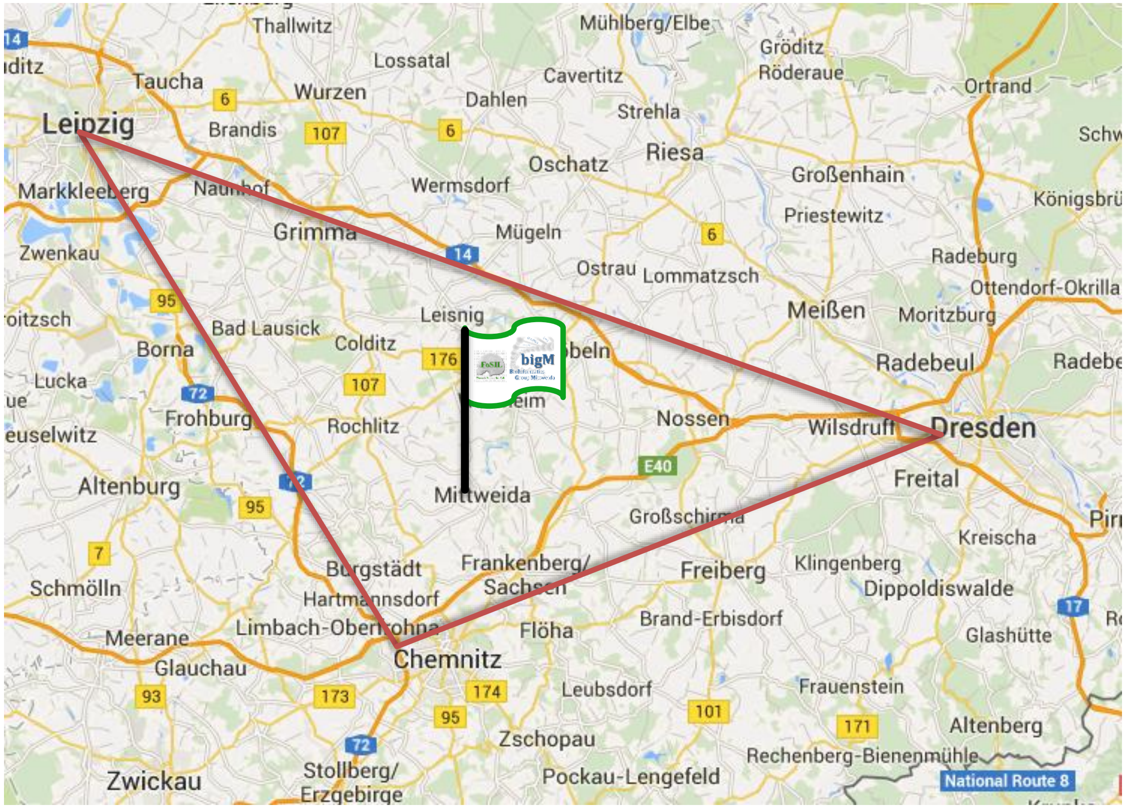
Ontologies - Useful tools in Life Sciences and Forensics

"How today's Life Science Technologies can shape the Crime Sciences of tomorrow"

Dirk Labudde
Mittweida

Mittweida

Schleswig Holstein
Hamburg
Mecklenburg Vorpommern
Bremen
Niedersachsen
Sachsen Anhalt
Brandenburg Berlin
Nordrhein-Westfalen
Hessen
Thüringen
Sachsen
Rheinland Pfalz
Saarland
Baden-Württemberg
Bayern



Leipzig
Dresden
Mittweida
Chemnitz
Grimma
Meißen
Freiberg
Zwickau
Altenburg
Borna
Rochlitz
Riesa
Oschatz
Wurzen
Radebeul
Freital

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Watson vs Watson

Dr. John H. Watson

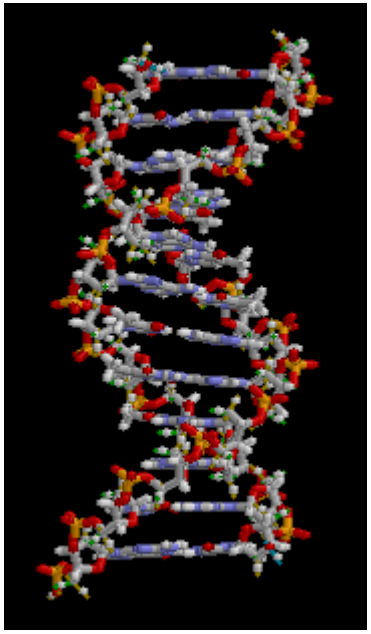
Investigator –
literary character

Dr. James Dewey Watson

Cofounder of the modern biology

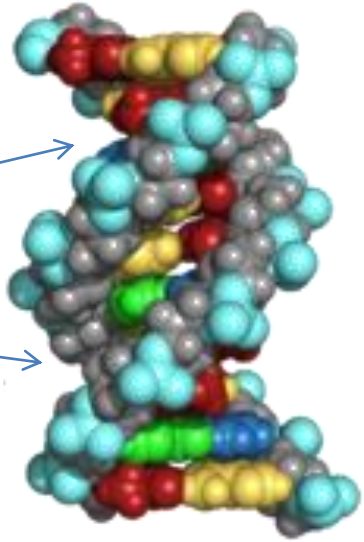
Modern Biology

1952 - King's College in London – **DNA** x-ray



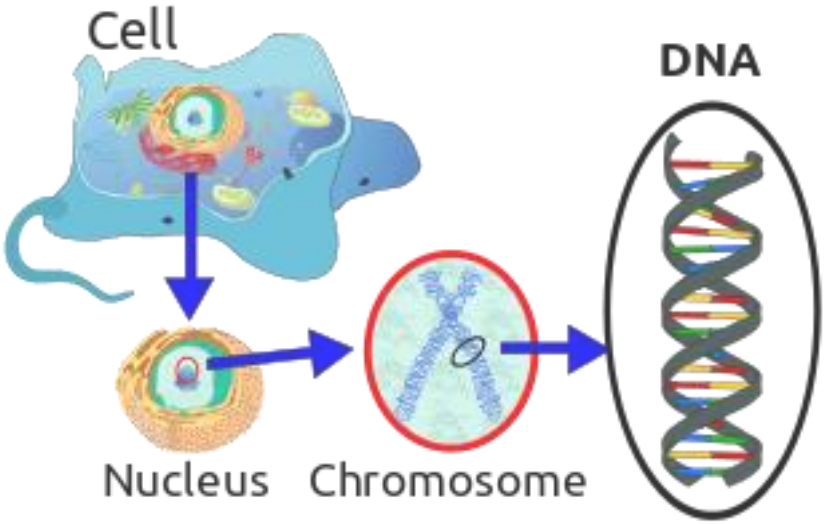
structure → Function

Major and minor grooves of DNA



The structure of part of a DNA double helix

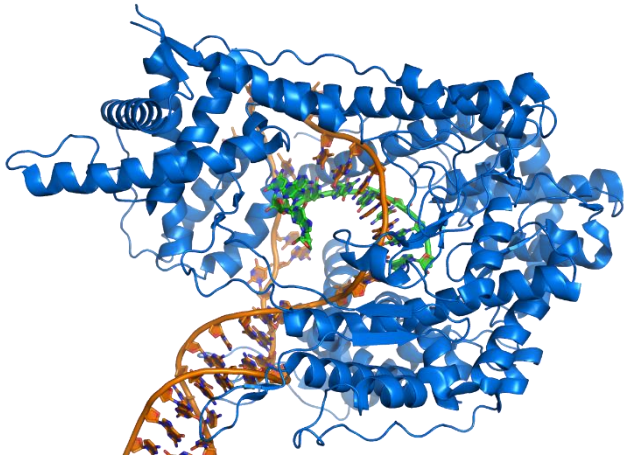
Modern Biology



DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome; the human **genome** has approximately **3 billion base pairs** of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence of pieces of DNA called **genes**. Transmission of genetic information in genes is achieved via complementary base pairing.

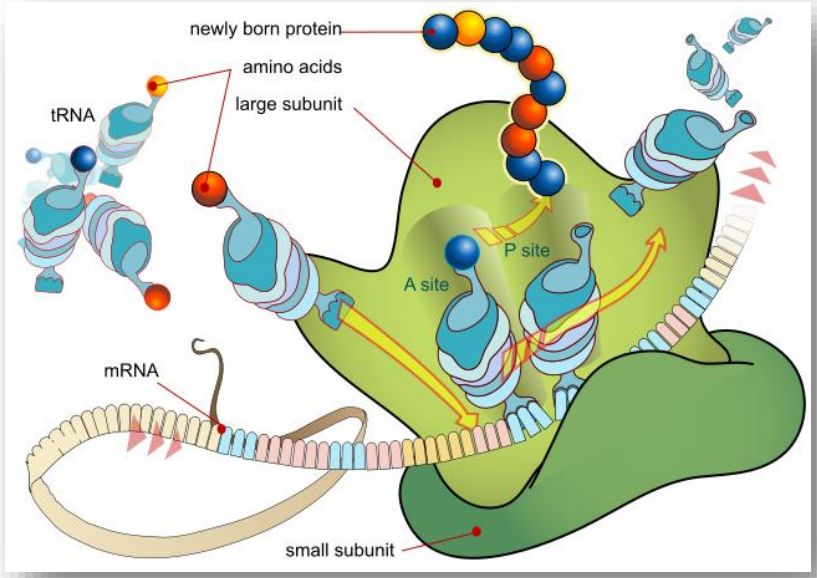
Modern Biology

Transcription and translation – biological information transfer



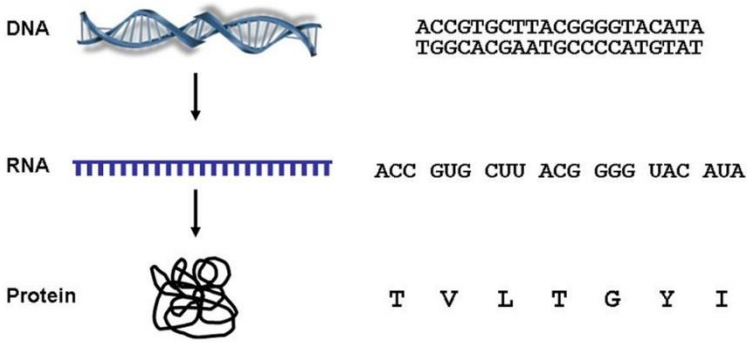
T7 RNA polymerase (blue) producing a mRNA (green) from a DNA template (orange).

In transcription, the codons of a gene are copied into messenger RNA by **RNA polymerase**.
Translation in Proteins



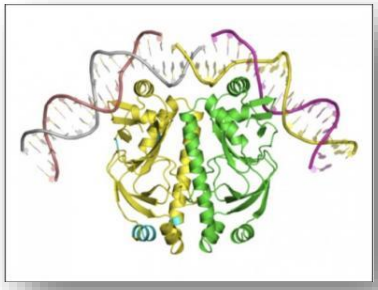
Modern Biology

Genes and **gen products**: Proteins – functional units of living cell



Sequence – structure - function

```
        10         20         30         40         50
MVLGKPQTDP TLEWFLSHCH IHKYPSKSTL IHQGEKAETL YYIVKGSVAV
        60         70         80         90        100
LIKDEEGKEM ILSYLNQGDF IGELGLFEEG QERSAWVRAK TACEVAEISY
       110        120        130        140        150
KKFRQLIQVN PDILMRLSAQ MARRLQVTSE KVGNLAFLDV TGRIAQTLLN
       160        170        180        190        200
LAKQPDAMTH PDGMQIKITR QEIGQIVGCS RETVGRILKM LEDQNLISAH
       210
GKTIVVYGTR
```

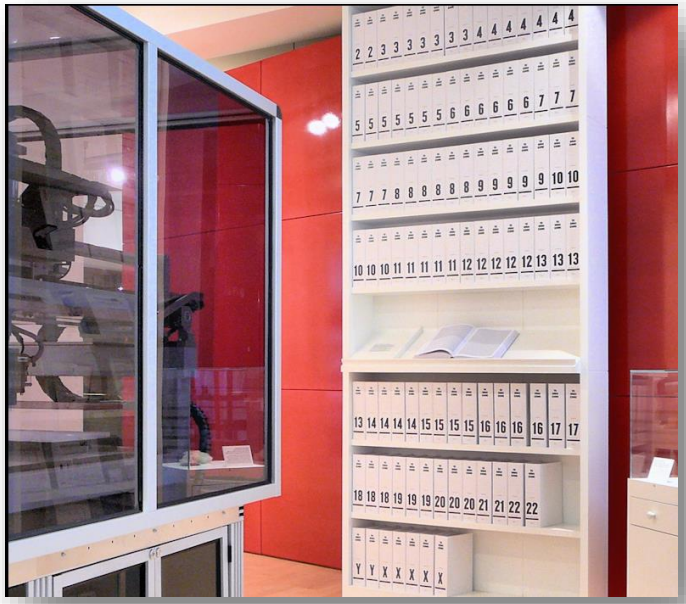


Human Genome Project 2001

Nature **409**, 860-921 (15 February 2001) | Accepted 9 January 2001

Initial sequencing and analysis of the human genome

The first printout of the human genome to be presented as a series of books



Modern Biology - bioinformatics

Start of the new bioinformatics in the omics era

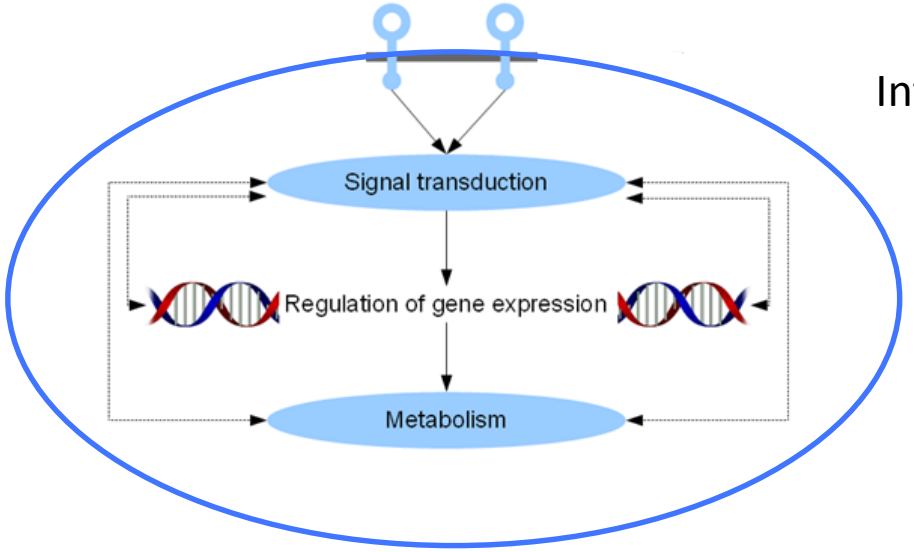
Data (sequences)
- Storage
- Integration and organization
- Analyses
- Standardization

high throughput methods

Simulation
Visualization
Modelling
Classifying

ontologies

genomics



Interatomics

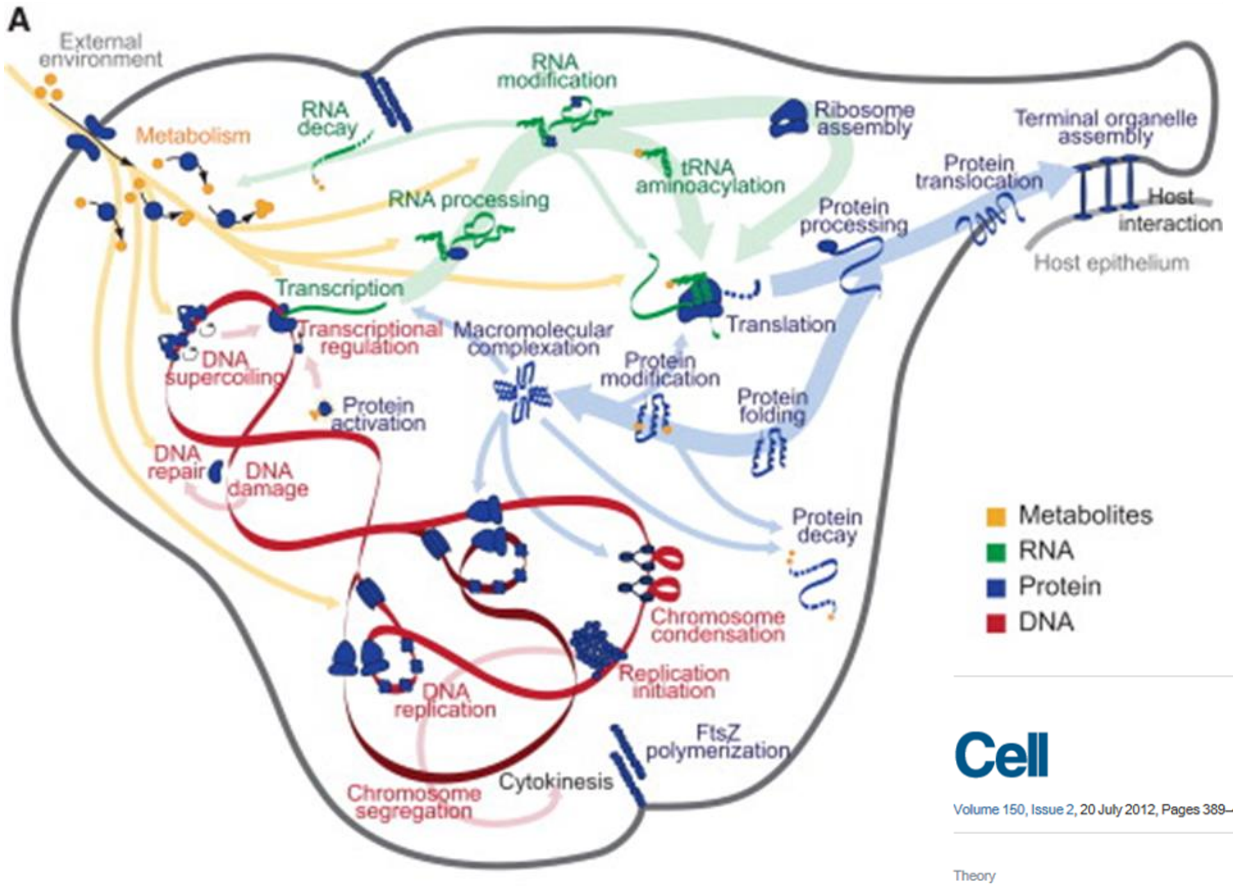
biblioomics

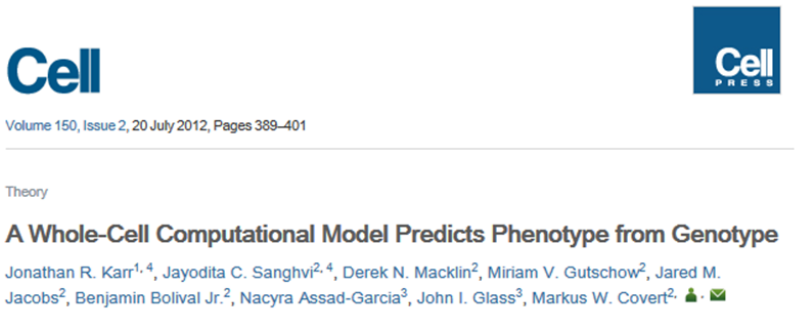
proteomics

metabolomics

transcriptomics

System biology – part of the bioinformatics



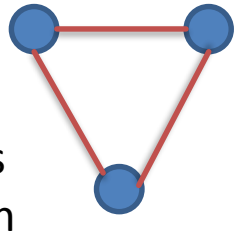


Gen ontology on different levels

System biology – part of the bioinformatics

Protein-Protein-Interaction network

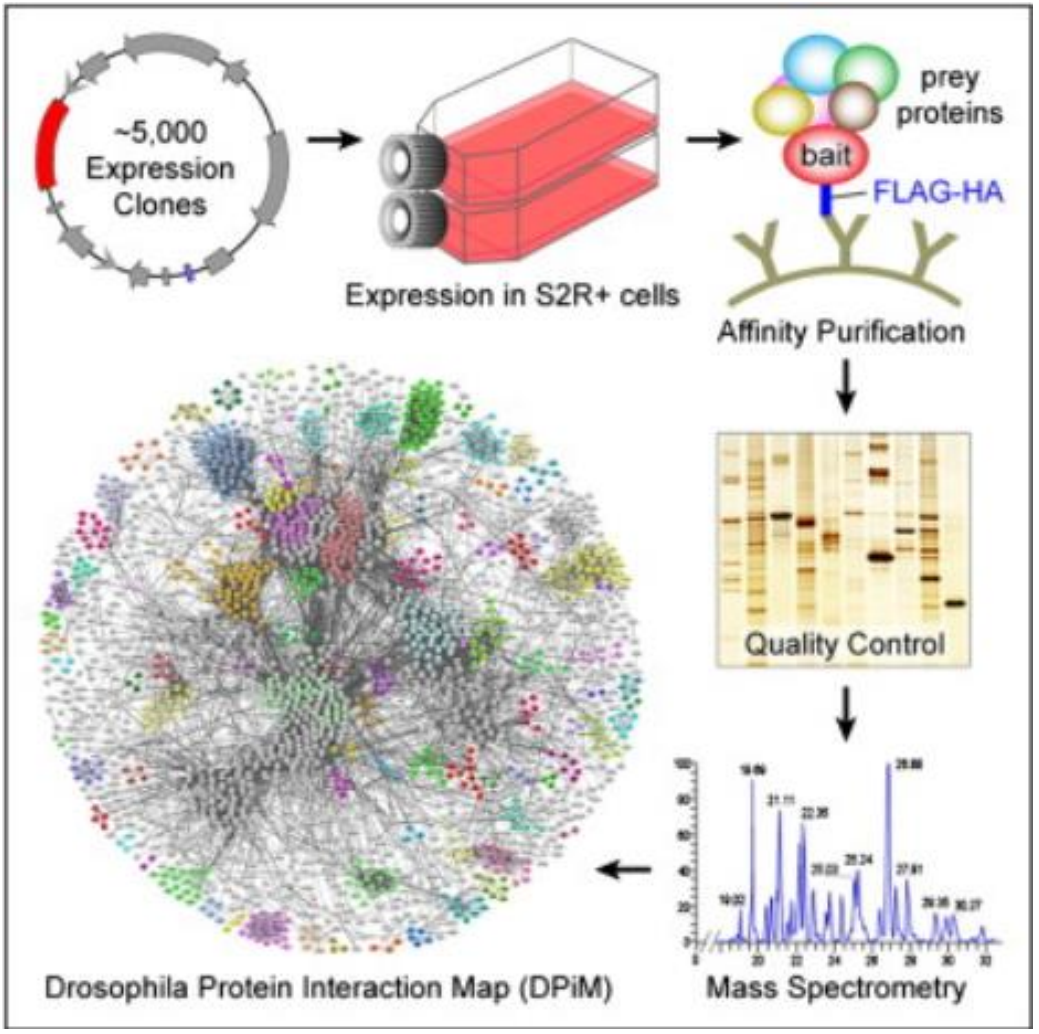
entities = proteins

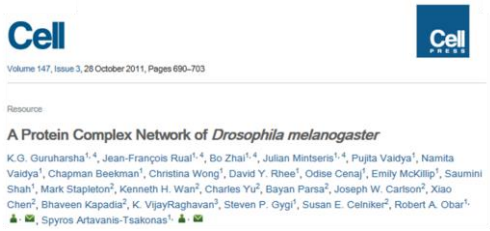


Kinetics reaction

experiments predicting

Interaction: Binding information transfer





Cell

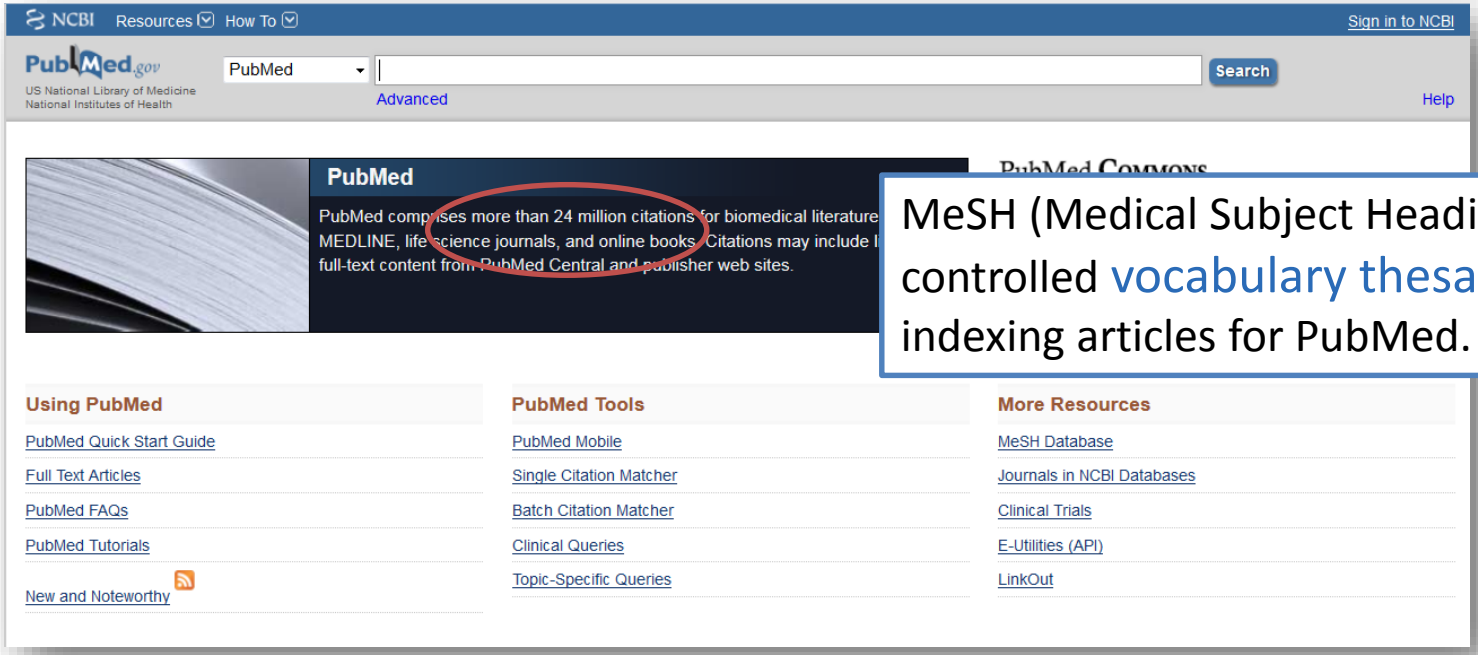
Volume 147, Issue 3, 28 October 2011, Pages 690–703

Resource

A Protein Complex Network of *Drosophila melanogaster*

K.G. Guruharsha[1,4], Jean-François Rual[1,4], Bo Zhai[1,4], Julian Mintseris[1,4], Pujita Vaidya[1], Namita Vaidya[1], Chapman Beekman[1], Christina Wong[1], David Y. Rhee[1], Odise Cenaj[1], Emily McKillip[1], Saumini Shah[1], Mark Stapleton[2], Kenneth H. Wan[2], Charles Yu[2], Bayan Parsa[2], Joseph W. Carlson[2], Xiao Chen[2], Bhaveen Kapadia[2], K. VijayRaghavan[3], Steven P. Gygi[1], Susan E. Celniker[2], Robert A. Obar[1,*], Spyros Artavanis-Tsakonas[1,*]

data in scientific journals - literature -omics

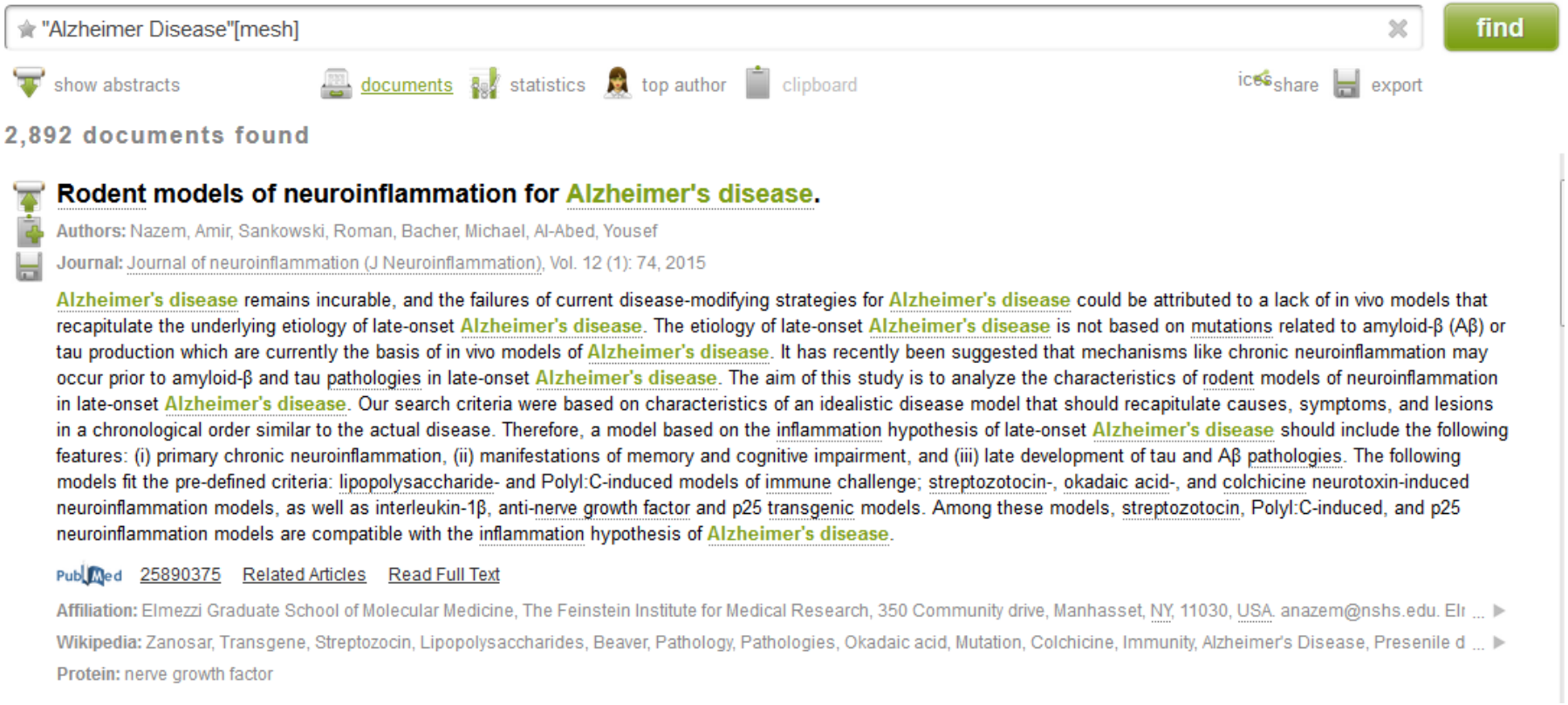


MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed.

Technics of TextMining

- Analyses of the abstract content
- Ontology and semantics

GOpubmed



background knowledge in the form of semantic networks
of concept categories (called ontologies or knowledge base)

Bioinformatics - summary

Data
- Storage
- Integration and organization
- Analyses
- Standardization

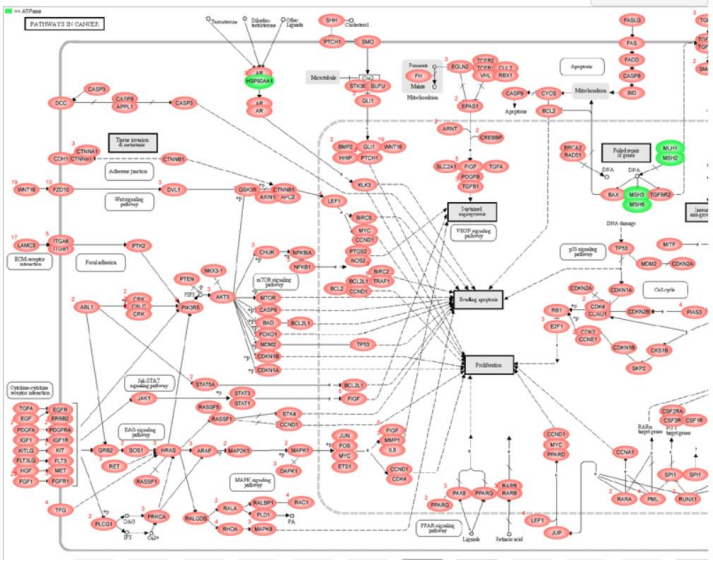
high throughput methods

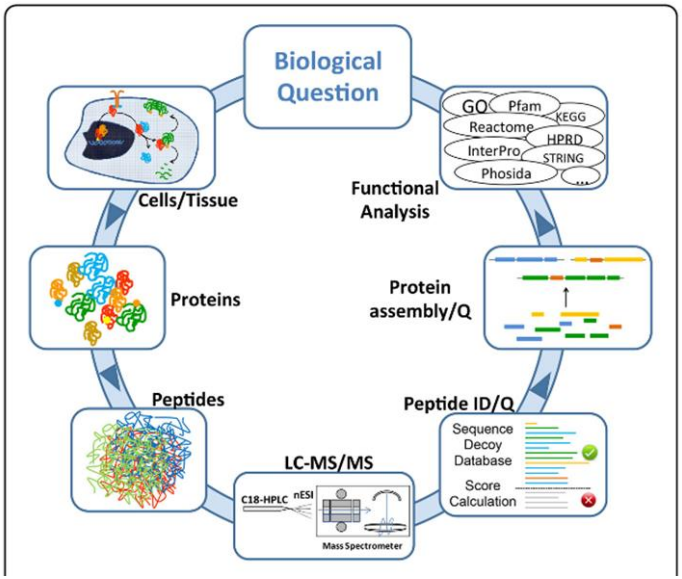
Big data
Ontologies
Databases and query languages

Simulation
Visualization
Modelling
Classifying

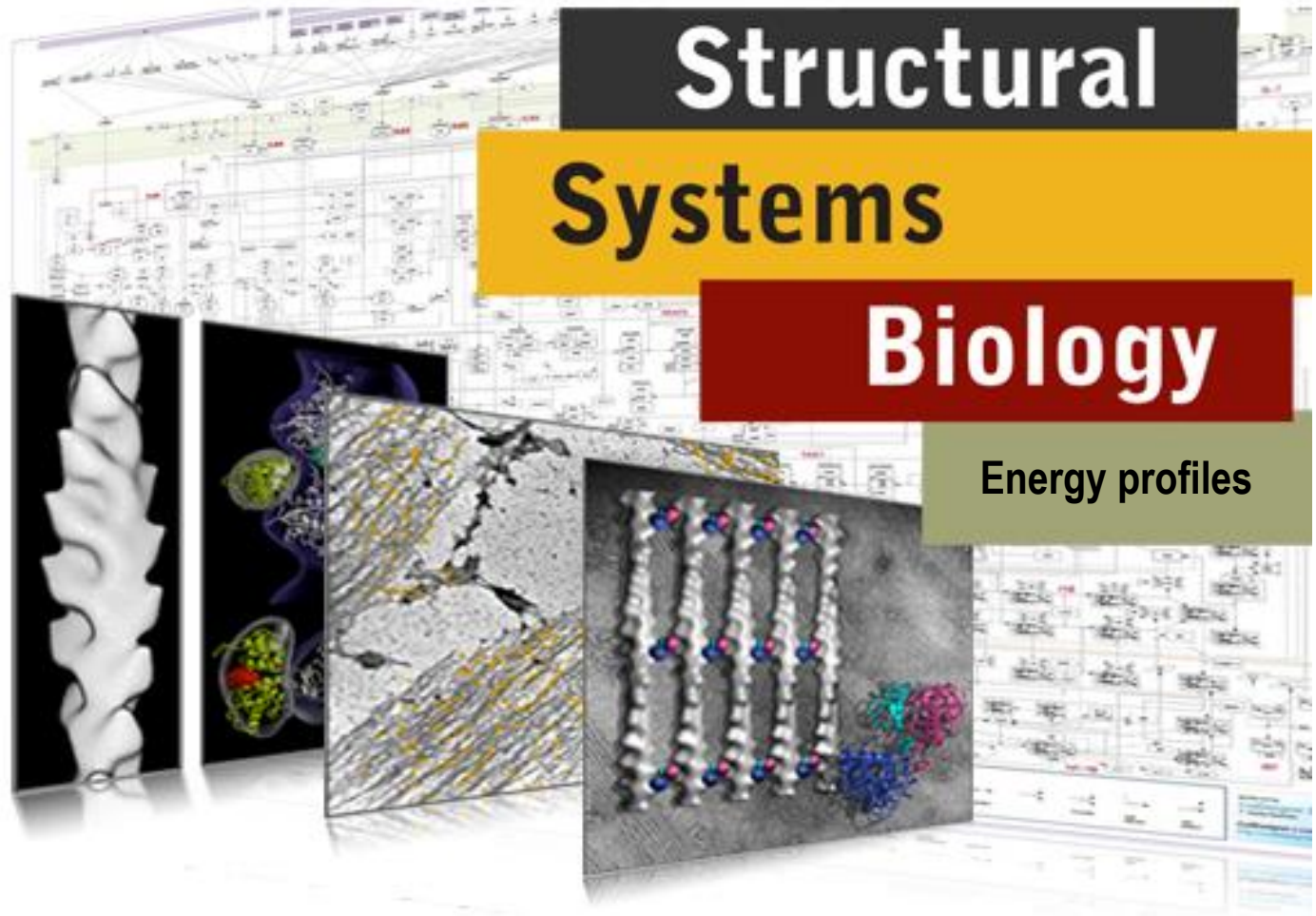
Evaluation/
validation

Statistics

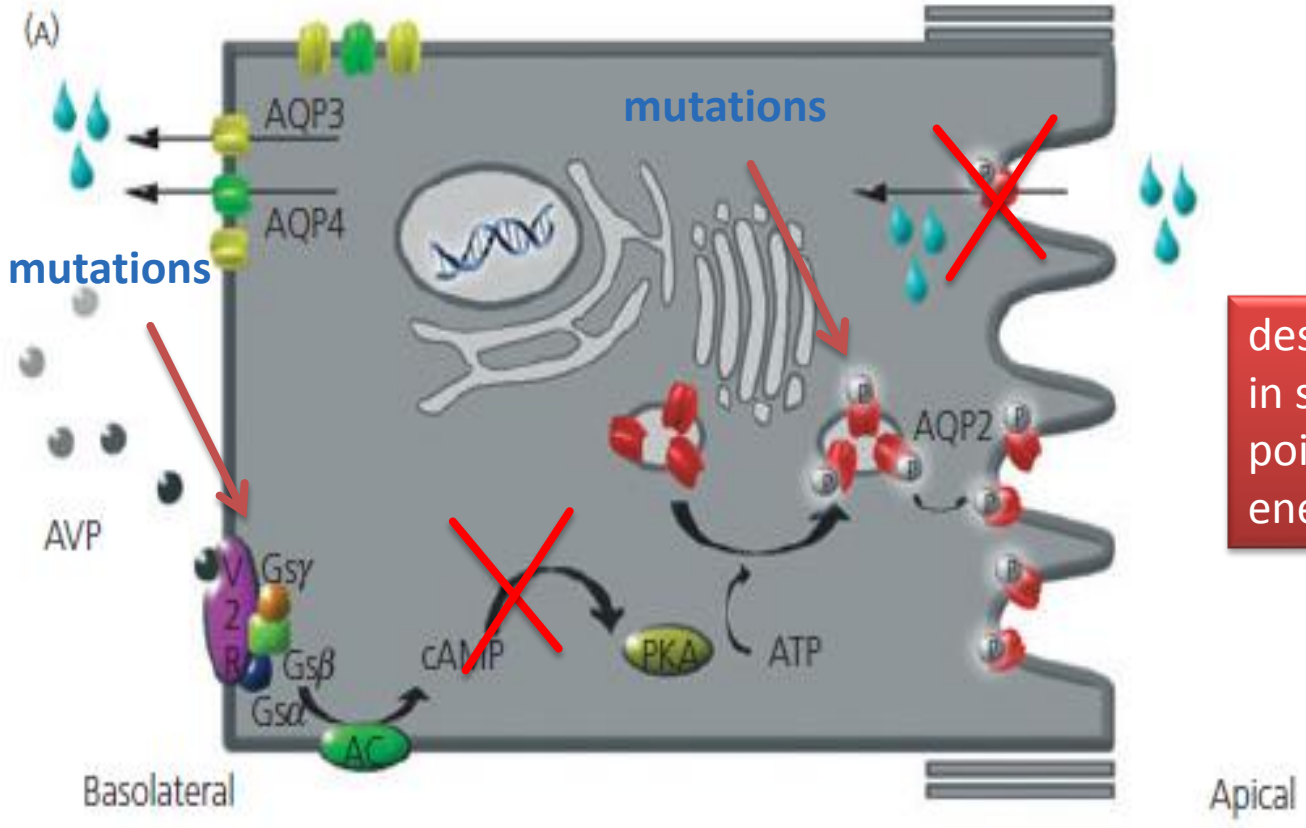




Energy profiles



Application - Diabetes insipidus in the context of systems biology



description of the changes in signaling pathway by point mutations based on energy profiles

Watson vs Watson

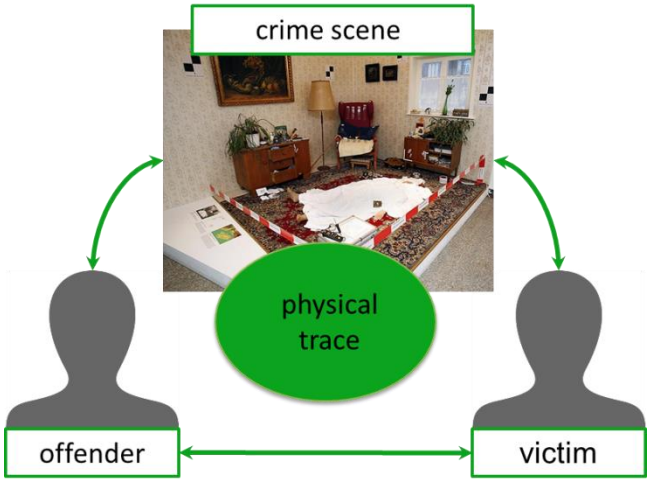
Dr. James Dewey Watson

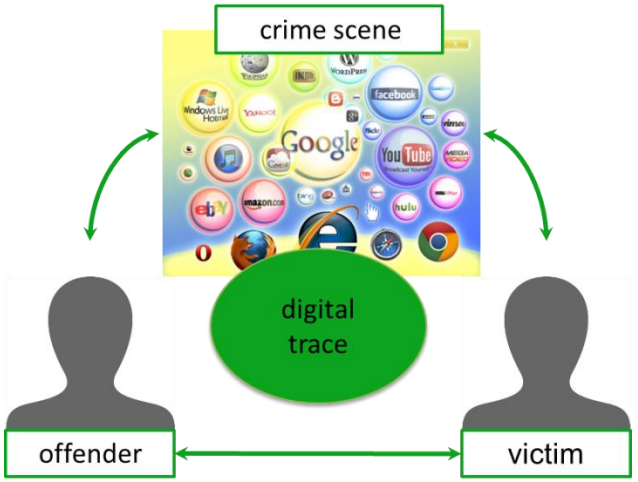
Cofounder of the modern biology

Dr. John H. *Watson*

Investigator –
literary character

Classical forensics and digital forensics



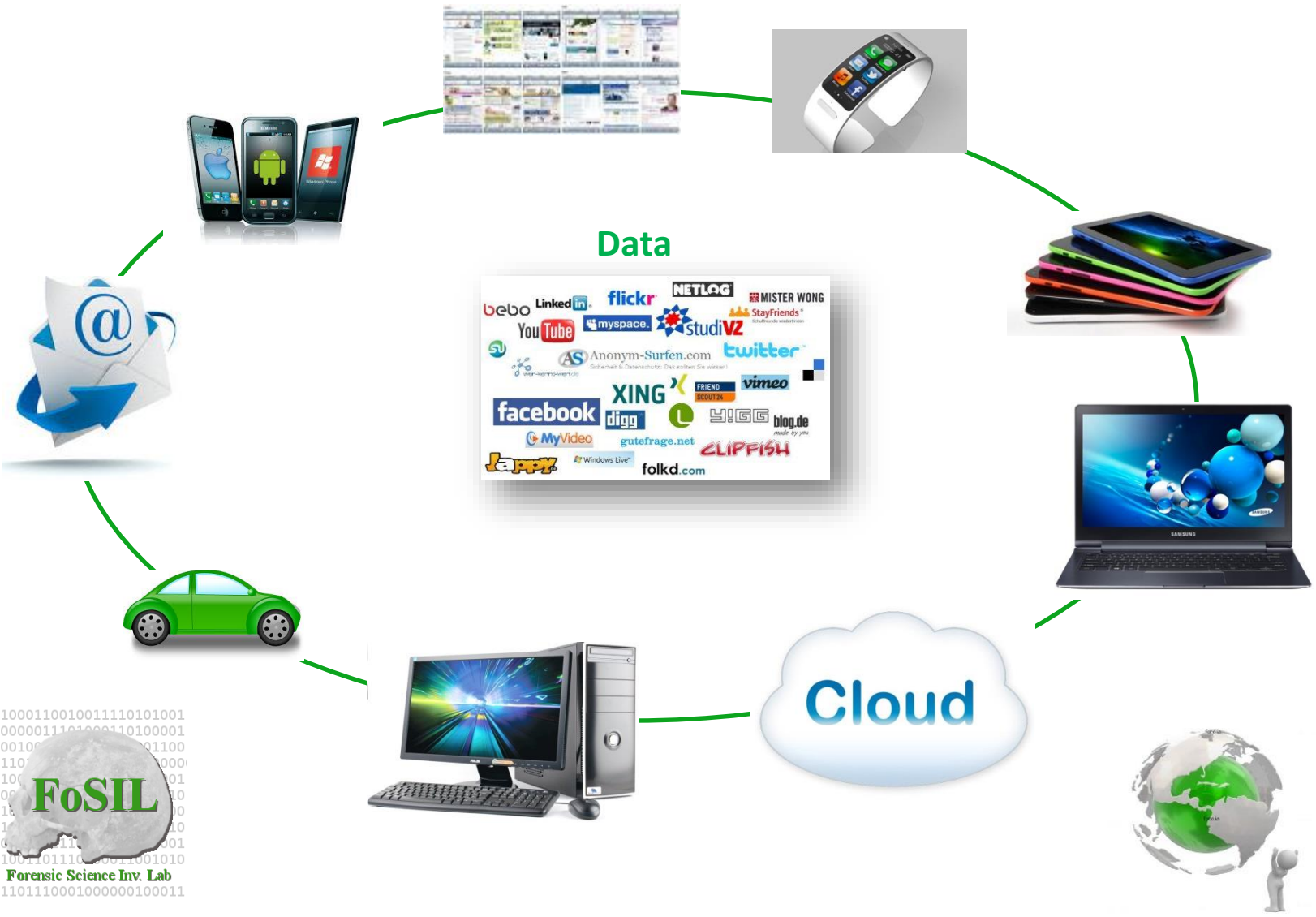


Locard's exchange principle

holds that the perpetrator of a crime will bring something into the crime scene and leave with something from it, and that both can be used as forensic evidence

Dr. Edmond Locard

Our daily Life



Classical forensics and digital forensics

- Analyses
- Validation
- **Evaluation**

circumstances of a crime

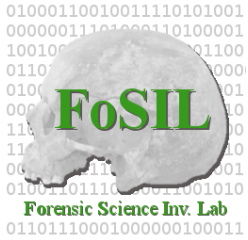
Physical or digital trace → evidence

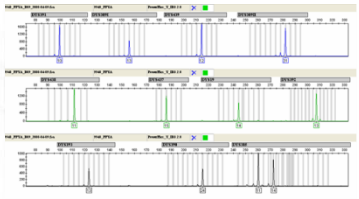
- fingerprint
- DNA traces
- traces of blood and pattern

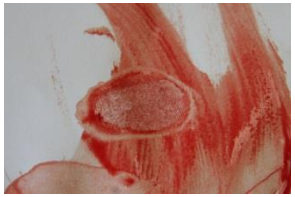
- Texts
- Short massages
- Chats
- …

comparison of traces:

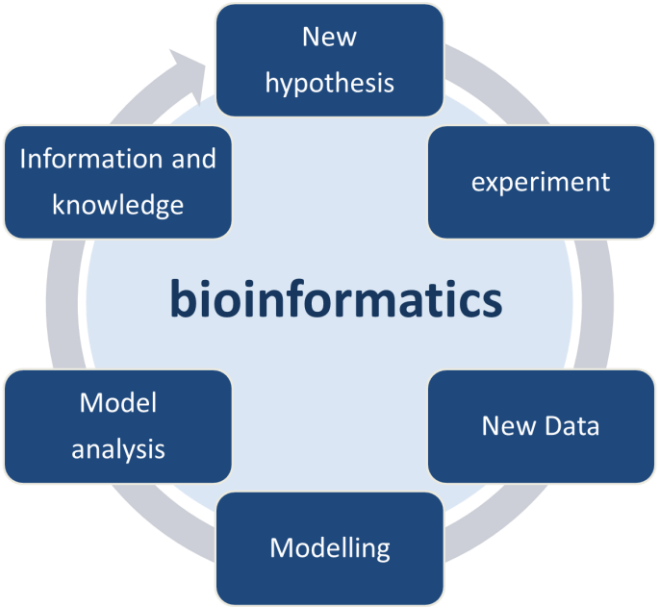
- Methods for the comparison
- Definition of similarity
- Evaluation of the results

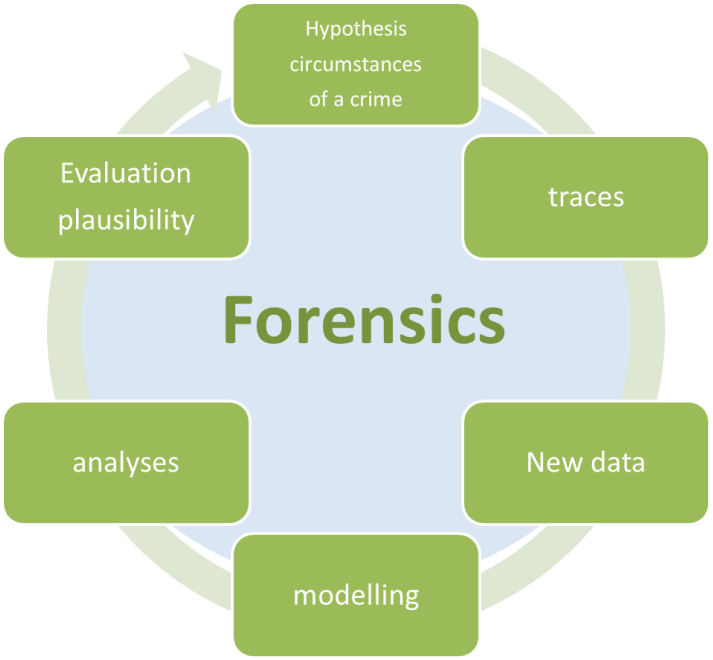






Classical forensics and digital forensics
-With respect to bioinformatics -





Classical forensics and digital forensics
-With respect to bioinformatics -

In cases of crime often computers and other data storage media will be seized or confiscated.

Case-relevant information needs to be separated and extracted to answer and prove criminalistic questions.

FoSIL
Forensic Science Inv. Lab

Classical forensics and digital forensics
-With respect to bioinformatics -

Pre-Process

- **text categorization**
- separation of case-relevant files
- text extraction/OCR
- modelling of a crime ontology

Main-Process

- syntactic annotation
- semantic annotation

Post-Process

- **detection of hidden semantics**

Pre-Process
- creating analysis corpus
- creating crime ontology

Main-Process
- basic text processing
- detecting secondary contexts
- instantiating crime ontology

Post-Process
- detecting hidden semantics

Classical forensics and digital forensics -With respect to bioinformatics -

Categorization of Forensic Texts

- bootstrapping ML combined with a set of rules
- rules determine the seed documents

investigator

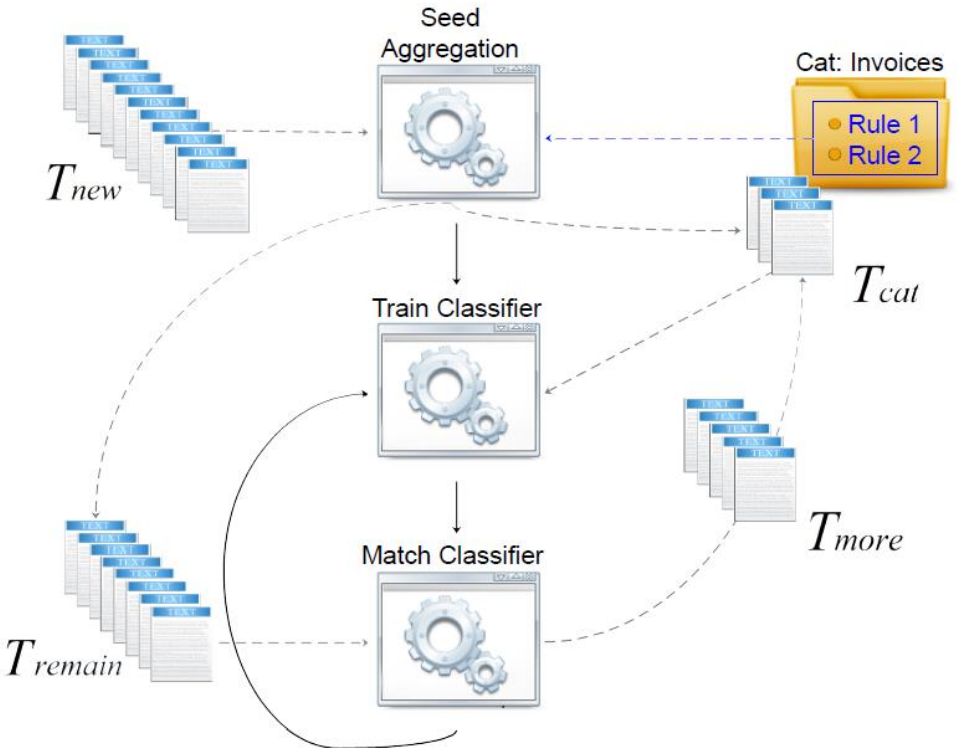
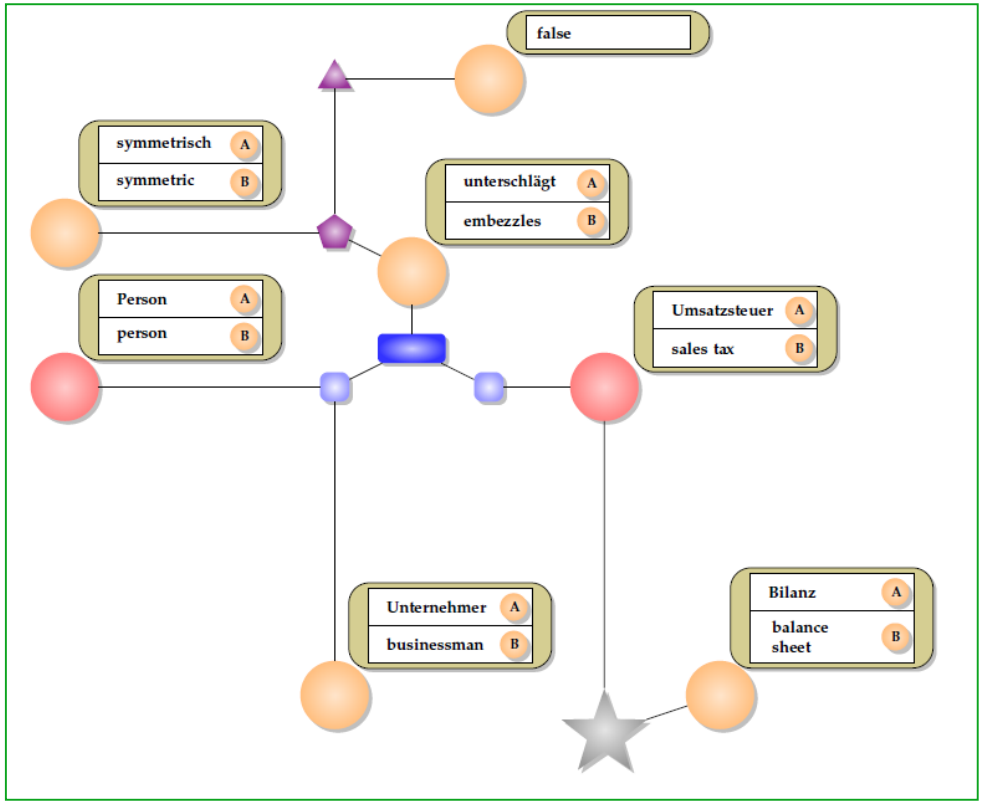


Figure 6: Bootstrapping algorithm for classifying forensic texts. From the texts T_{new} a set of seed documents for each category is acquired using the rules annotated in the taxonomy. This set T_{cat} is used to train one initial weak binary classifier per category. Subsequently, this classifier is used to classify the remaining texts T_{remain} and store the new labelled documents T_{more} to T_{cat}. Finally, the classifier is going to be improved iteratively using T_{cat} until no document is left or no further improvement is possible.

Ontologies und Semantics



Why do we need ontologies?

- they define terms and symbols referring to a syntax and an association network
- prior condition for raising questions
- can support the information extraction process in different ways
- can support the visualization of results

FoSIL
Forensic Science Inv. Lab

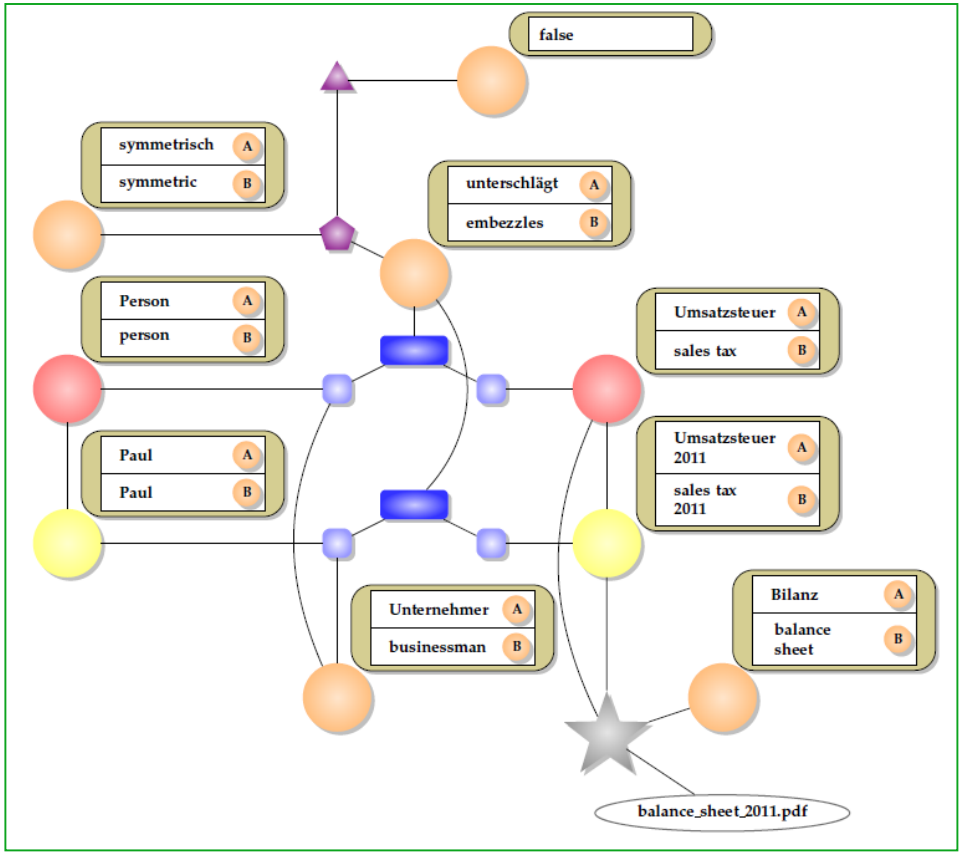
- oriented in the Topic Map ISO-standard
- readable for humans and processable for computers

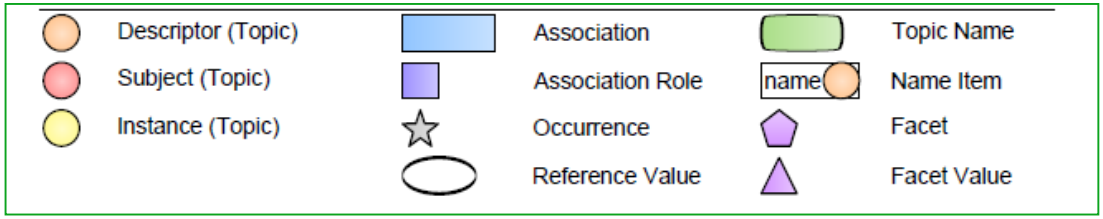
Ontologies und Semantics

Forensische Topic Map

instantiation

Question Answering System



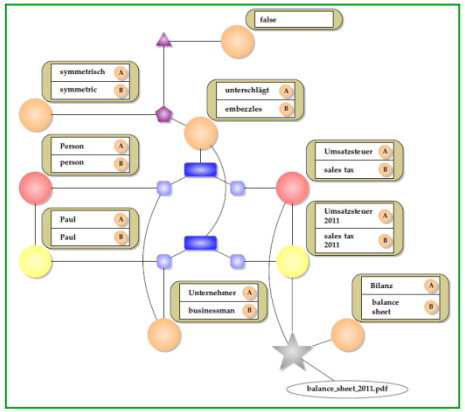


FoSIL
Forensic Science Inv. Lab

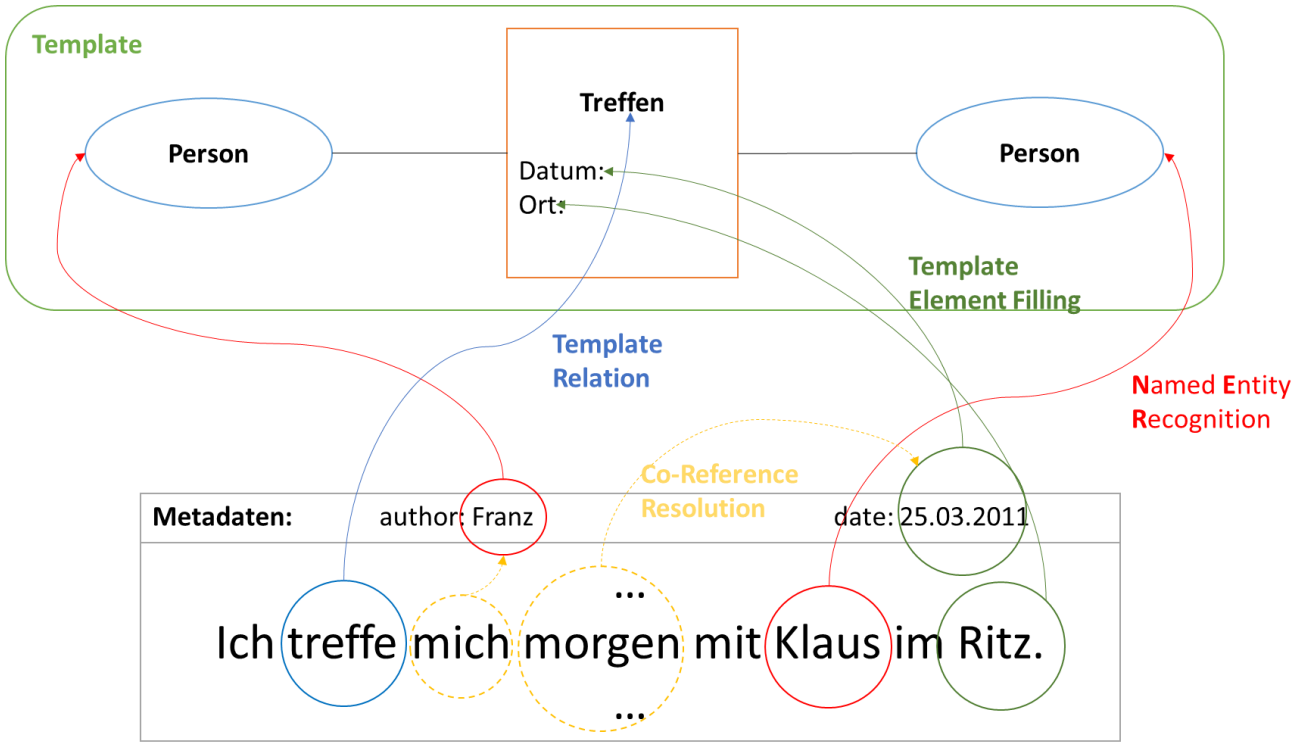
Texts – Information extraction

Forensische Topic Map

instantiation

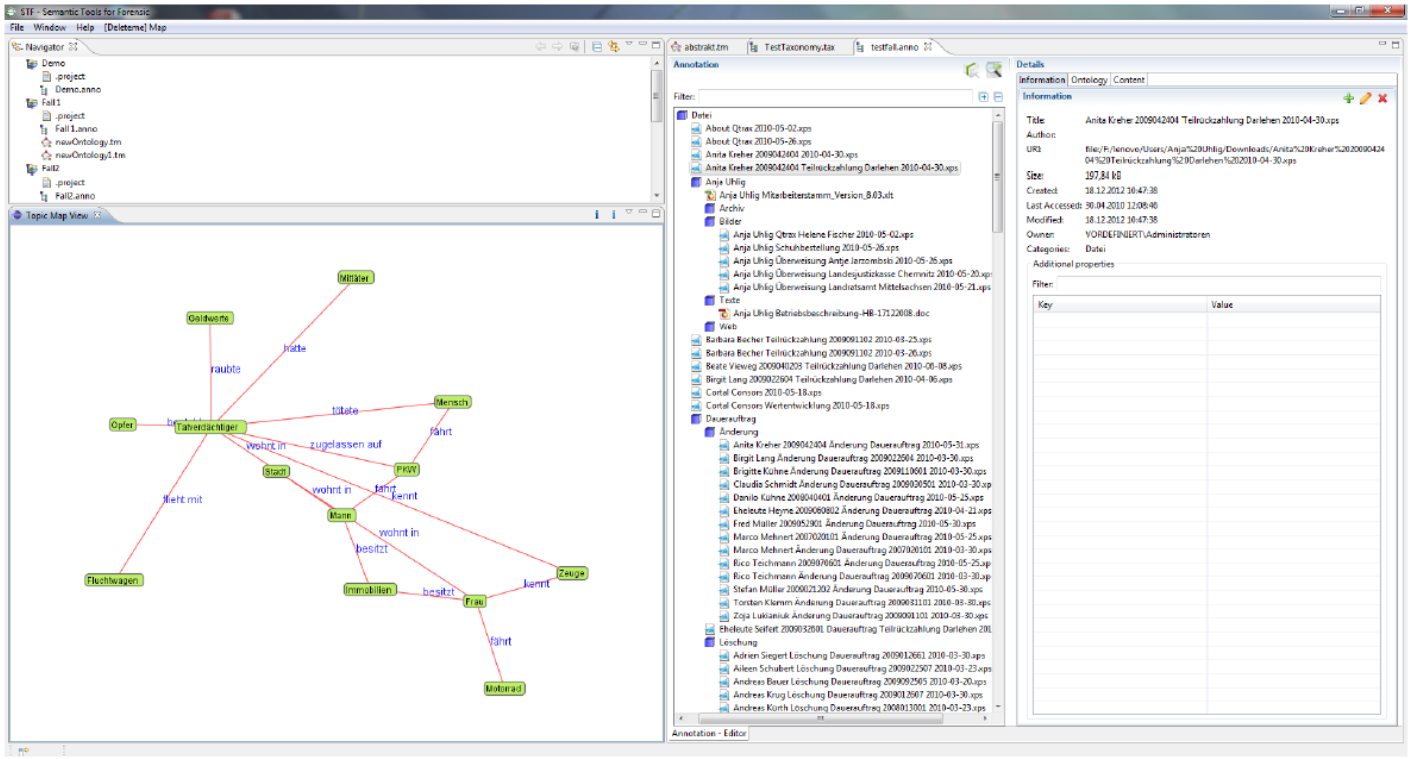


Information Extraction



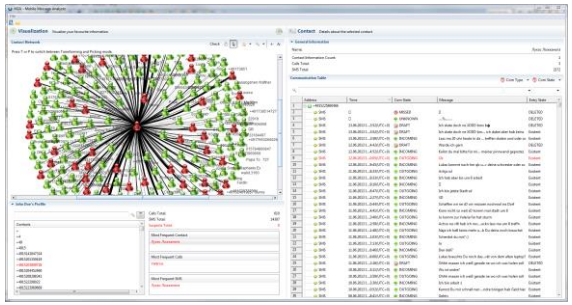
FoSIL
Forensic Science Inv. Lab

Ontologies und Semantics



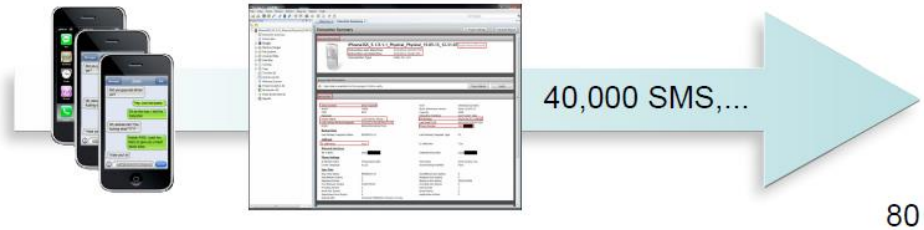
SEMANTIC TEXT ANALYZER

Ontologies und Semantics



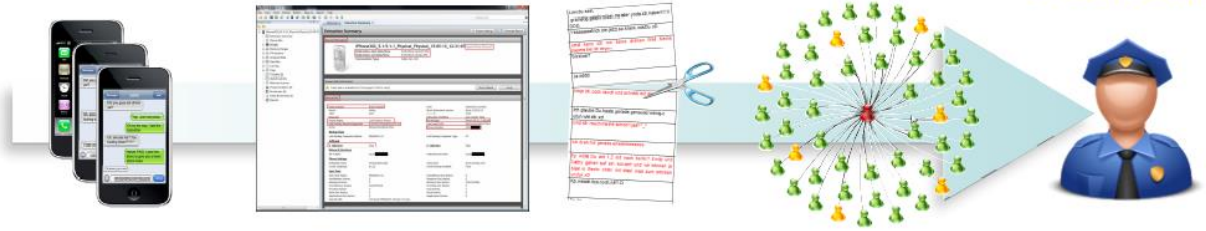
Mona

Mobile Message ANALYZER



8 weeks later

80 evidential SMS from 7suspects found manually



8 hours later

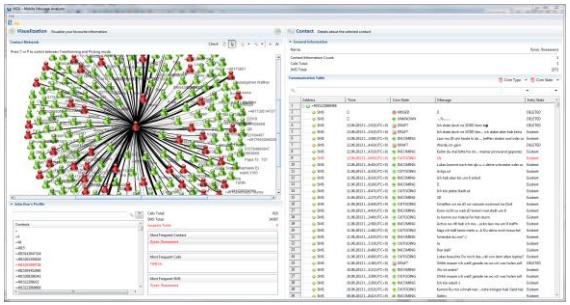
20 talks containing 500 SMS incl. 80 evidential SMS from 7suspects are extracted automatically

1-2 days manual effort

FoSIL Forensic Science Inv. Lab

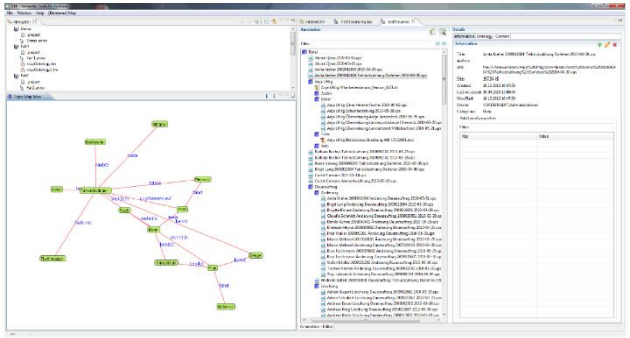
100% Precision und 67% Sensitivity

Ontologies und Semantics



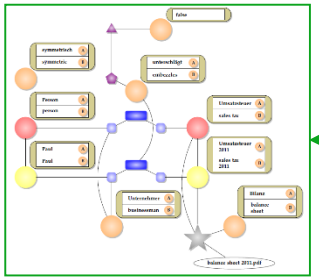
Mona

Mobile Message ANALYZER



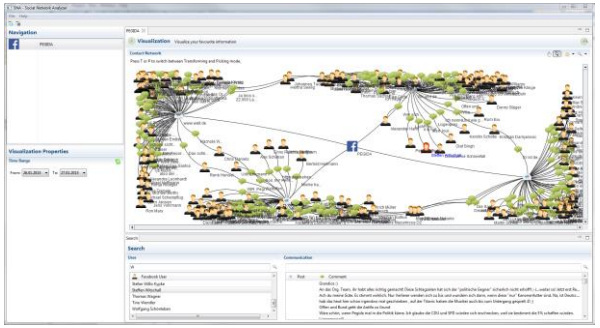
SemanTA

Semantic Text ANALYZER



AVATAR

A Victim Analysis Toolbox for Anatomic Reconstruction



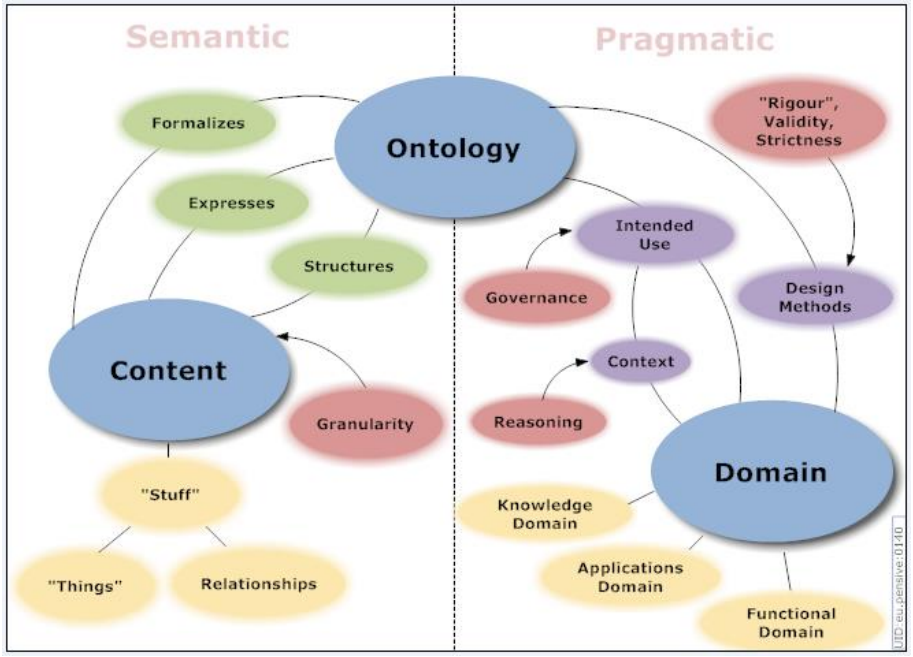
SoNA

Social Network ANALYZER

Ontology and Semantics

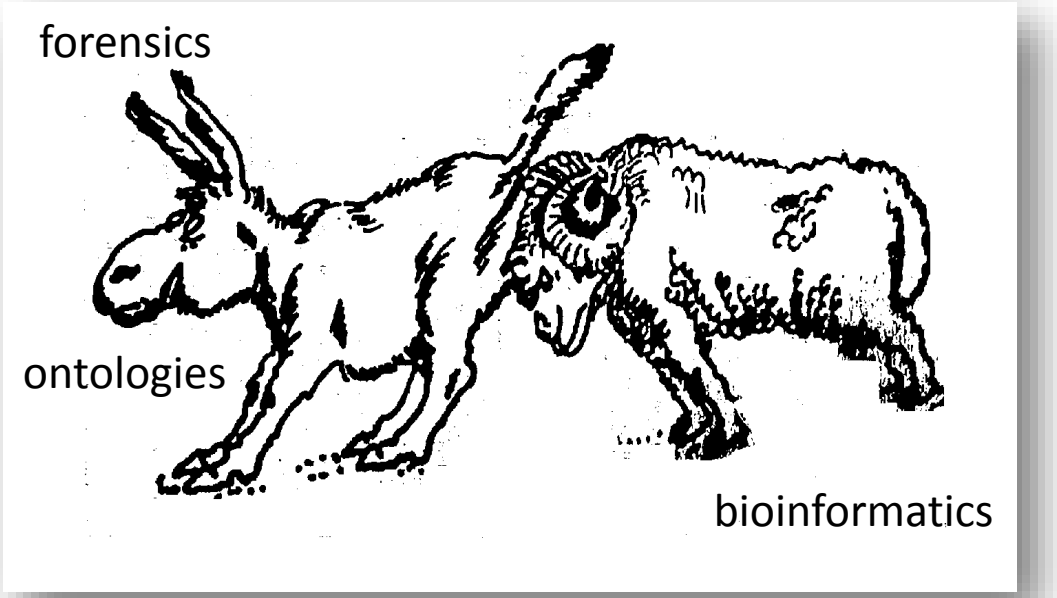
An ontology is a specification of a conceptualization.

Definition of syntax of terms and symbols in a network of associations



Summary

Bioinformatics and Forensics - How today's Life Science Technologies can shape the Crime Sciences of tomorrow



labudde@hs-mittweida.de www.bioforscher.de