

DataSys 2015

June 21 - 26, 2015 - Brussels, Belgium

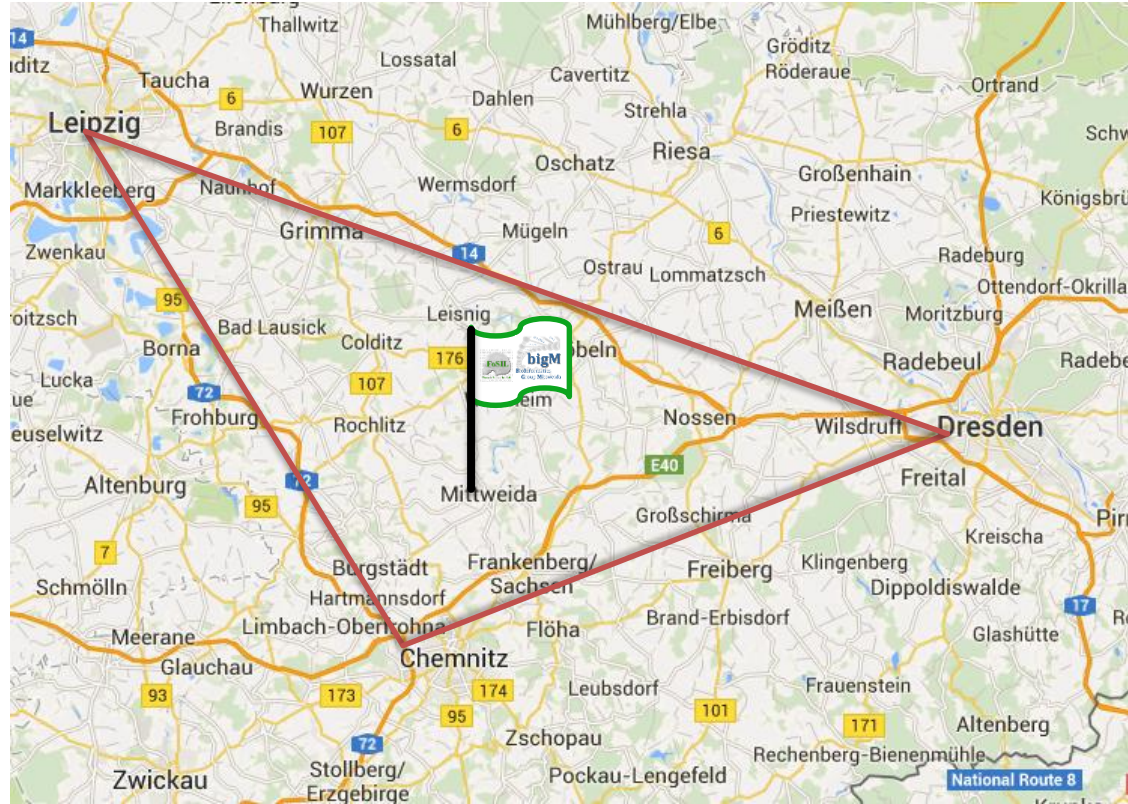
HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES



Ontologies - Useful tools in Life Sciences and Forensics

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

Dirk Labudde
Mittweida



Dr. John H. Watson



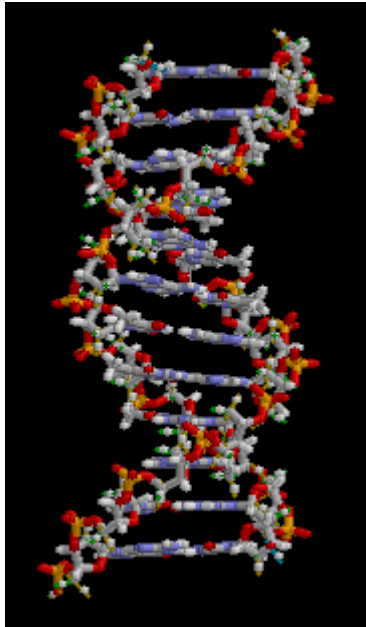
Investigator –
literary character

Dr. James Dewey Watson



Cofounder of the modern biology

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

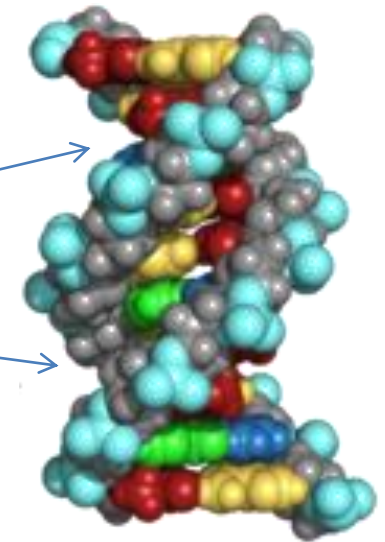


structure

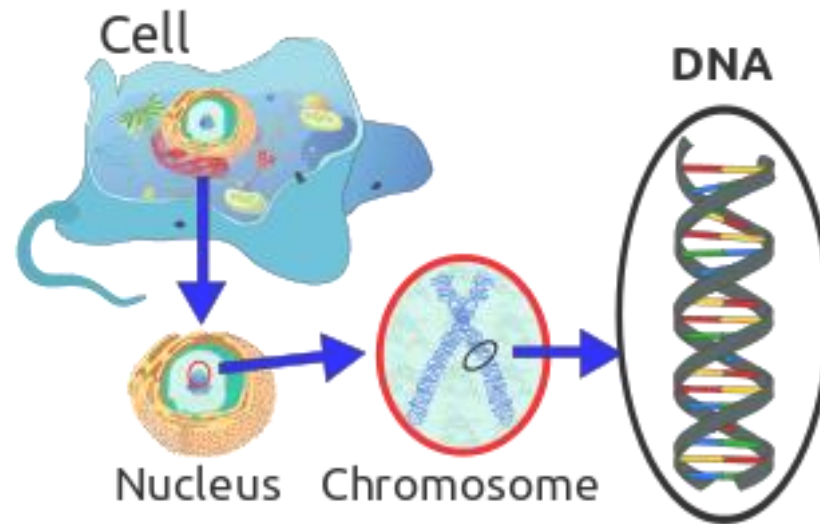


Function

Major
and
minor
grooves
of DNA

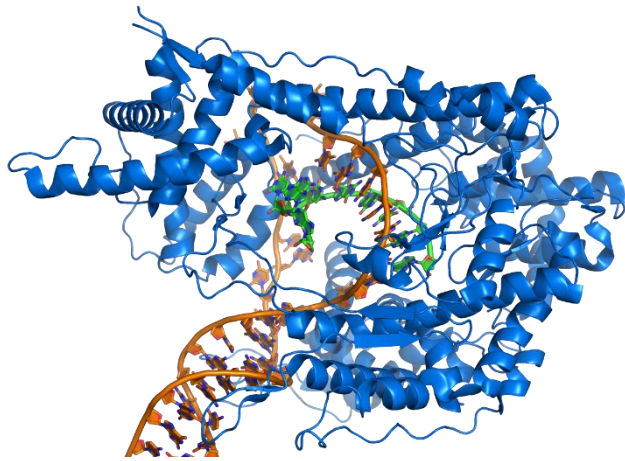


The structure of part of a
DNA double helix



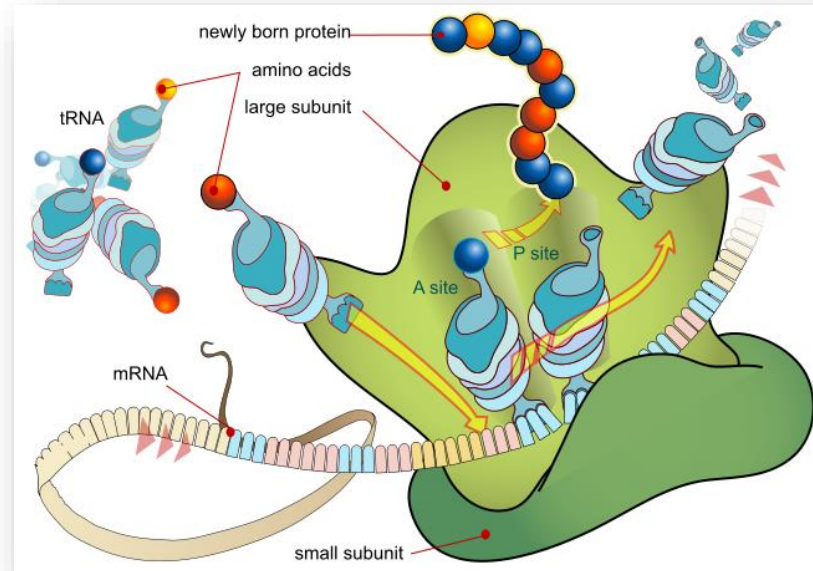
DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome; the human **genome** has approximately **3 billion base pairs** of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence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

Transcription and translation – biological information transfer

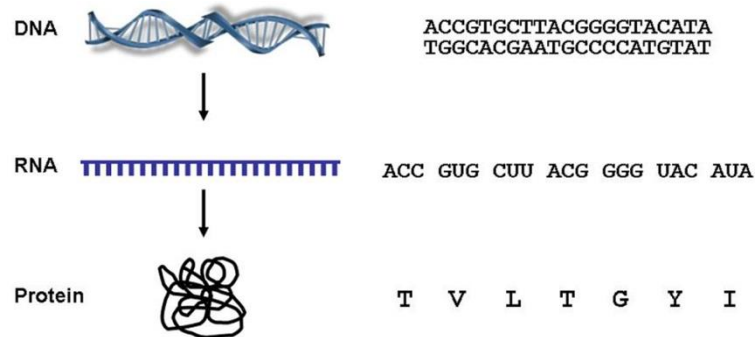


T7 RNA polymerase (blue)
producing a mRNA (green)
from a DNA template
(orange).

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



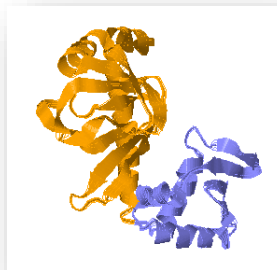
Genes and **gen products**: Proteins – functional units of living cell



Sequence – structure - function

```

10      20      30      40      50
MVLGKPTDF TLEWFLSHCH IHRYPSKSTL IHQGEKAETL YYIVKGSVAV
60      70      80      90     100
LIKDEEGKEM ILSYLNQGDF IGELGLFEEG QERSAWVRAK TACEVAEISY
110     120     130     140     150
KKFRQLIQVN PDILMRLSAQ MARRLQVISE KVGNLAPLDV TGRIAQTLN
160     170     180     190     200
LARQPDAMTH PDGMQIKITR QEIGQIVGCS RETVGRILEM LEDQNLISAH
210
GKTIIVVYGR
    
```



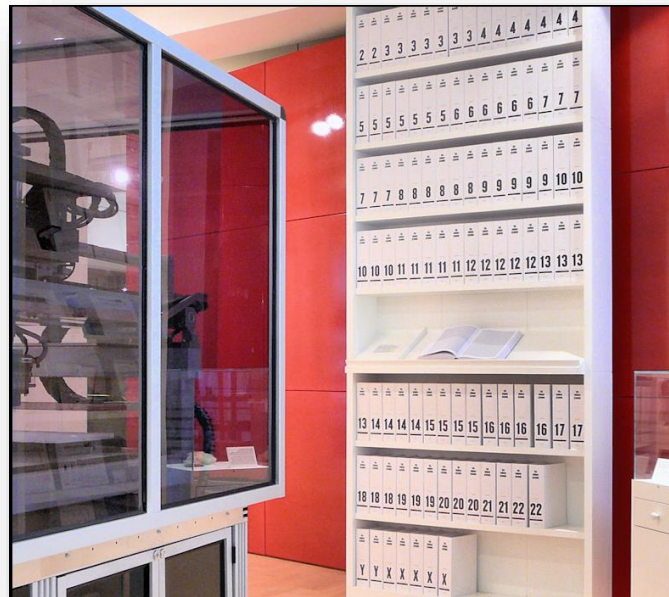
Human Genome Project 2001



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

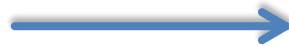


Start of the new bioinformatics in the omics era

Data (sequences)

- Storage
- Integration and organization
- Analyses
- Standardization

high throughput methods



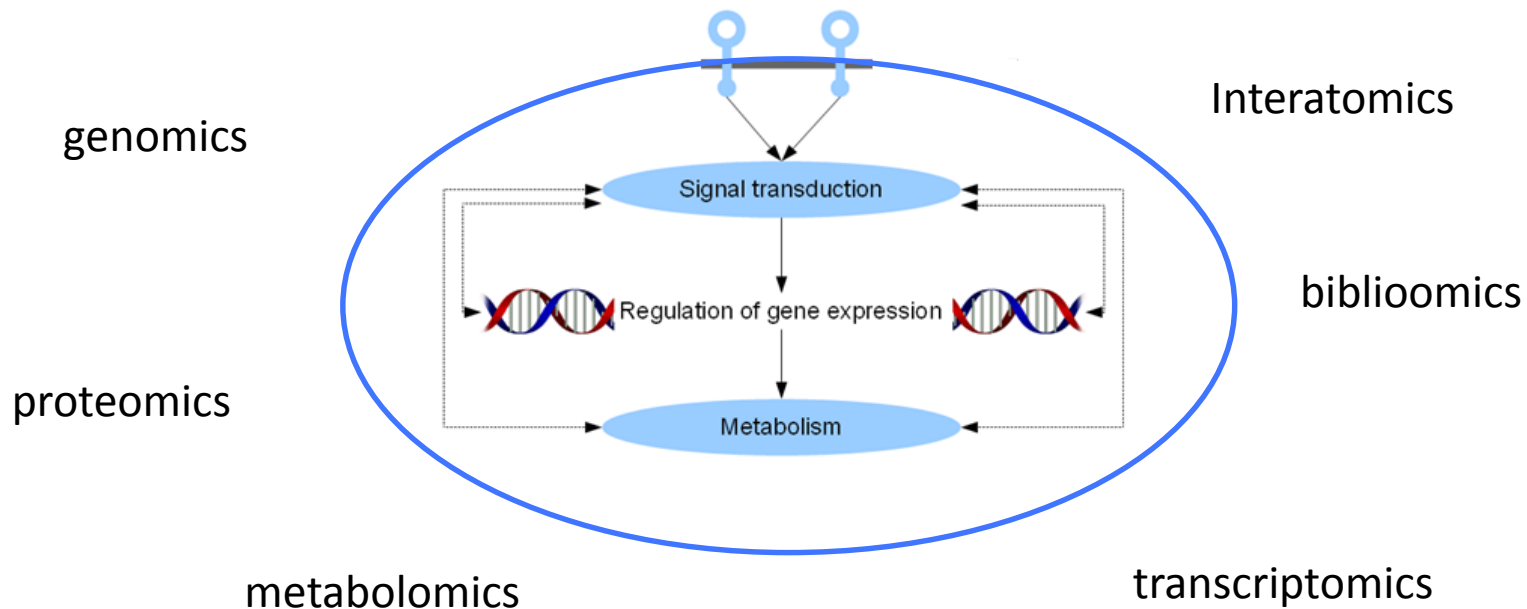
Simulation

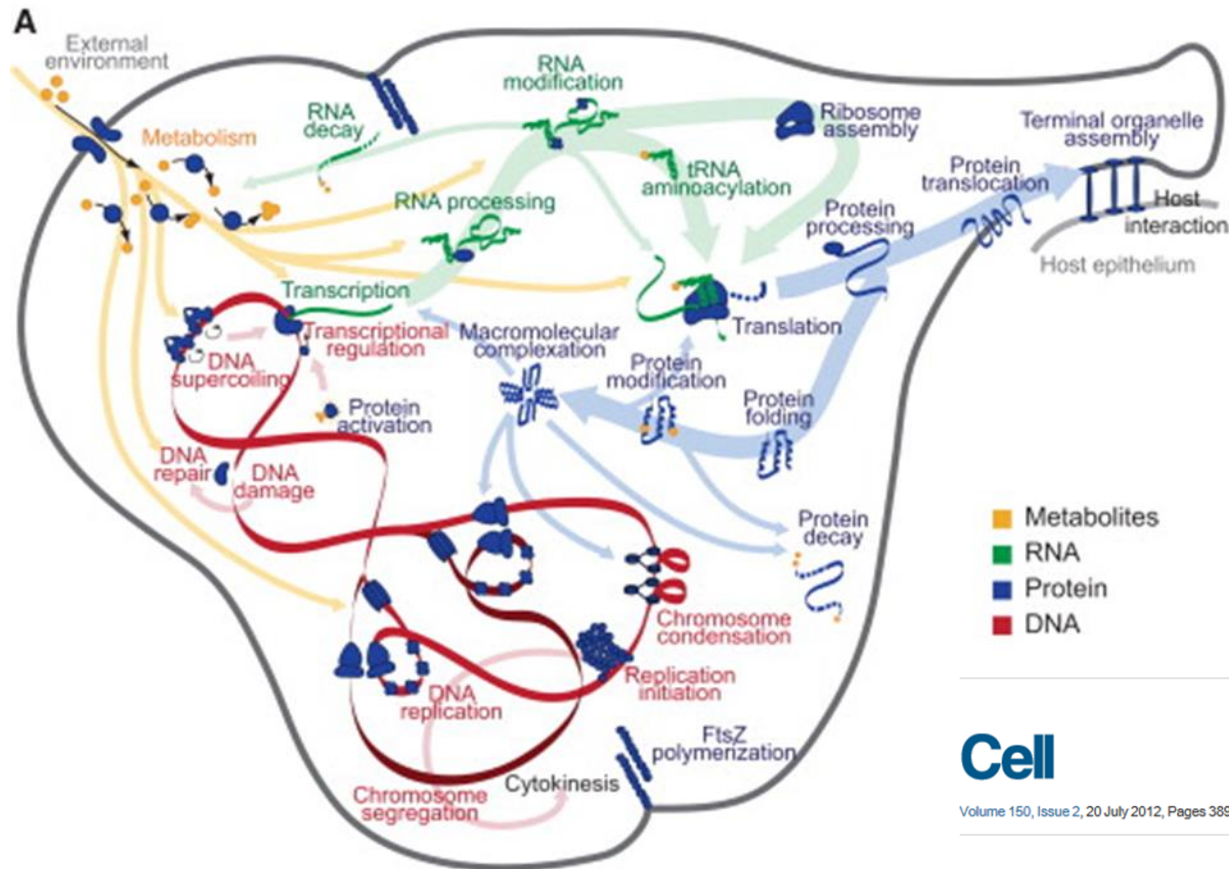
Visualization

Modelling

Classifying

} ontologies





Cell

Volume 150, Issue 2, 20 July 2012, Pages 389–401



Theory

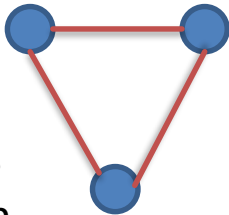
A Whole-Cell Computational Model Predicts Phenotype from Genotype

Jonathan R. Karr^{1,4}, Jayodita C. Sanghvi^{2,4}, Derek N. Macklin², Miriam V. Gutschow², Jared M. Jacobs², Benjamin Bolival Jr.², Nacyra Assad-Garcia³, John I. Glass³, Markus W. Covert²  

Gen ontology on different levels

Protein-Protein-Interaction network

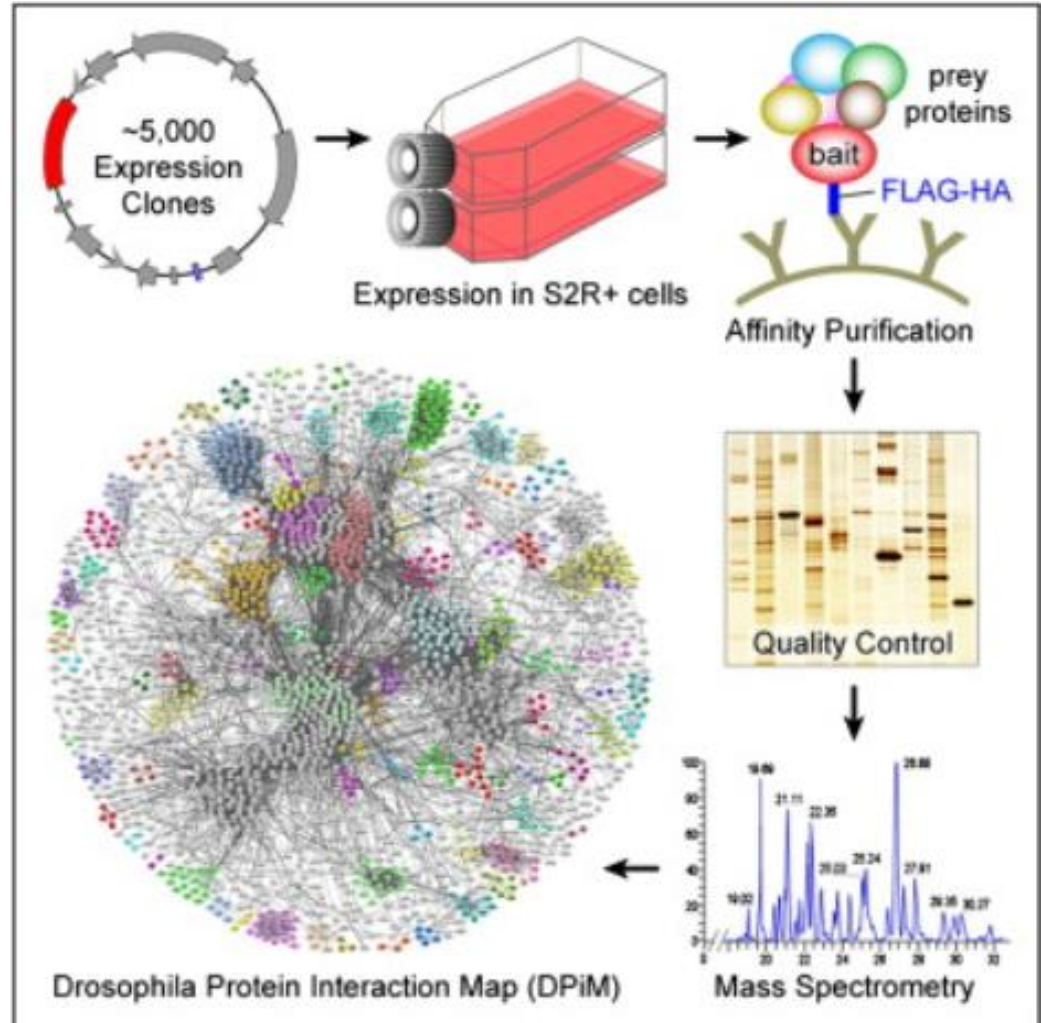
entities = proteins



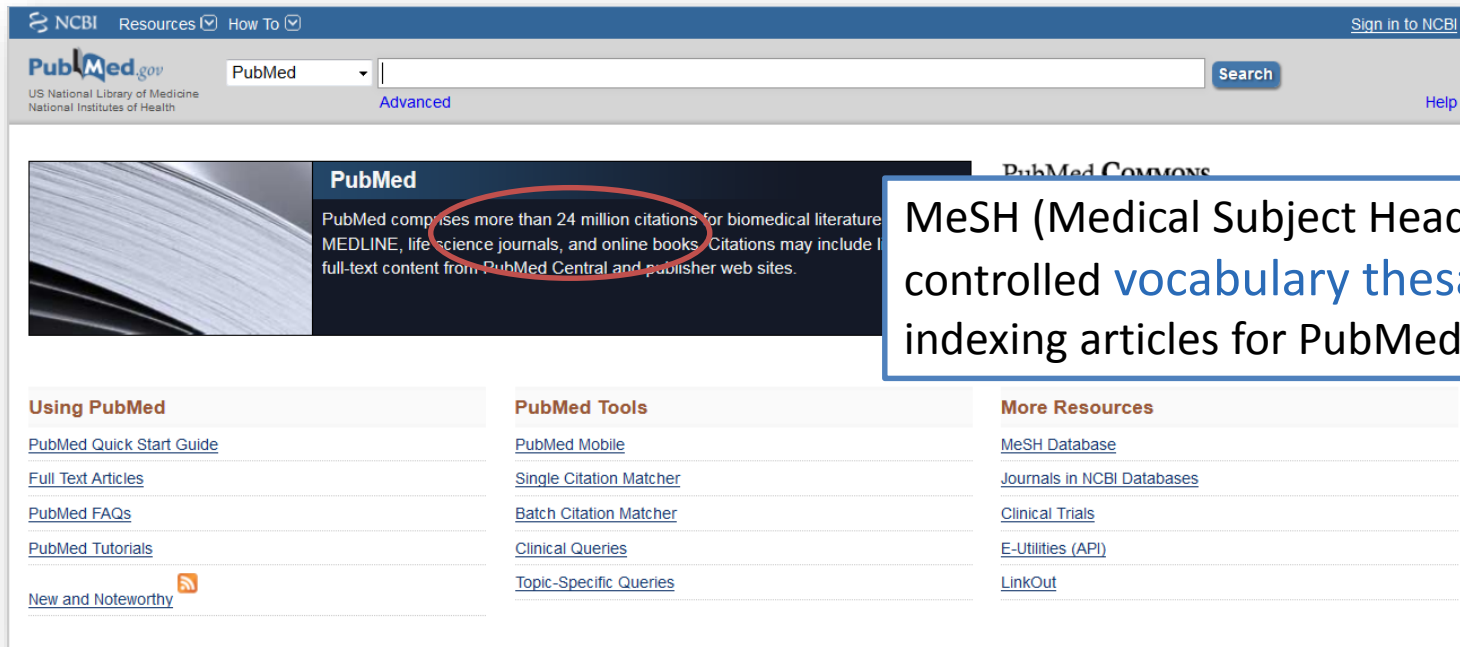
Kinetics
reaction

experiments
predicting

Interaction:
Binding
information transfer

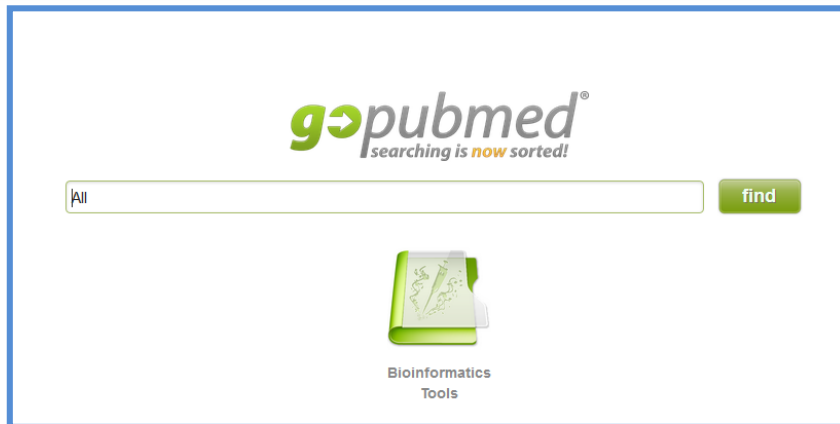


Resource
A Protein Complex Network of *Drosophila melanogaster*
K.G. Gurusantha^{1,4}, Jean-François Ruel^{1,4}, Bo Zhai^{1,4}, Julian Mintzer^{1,4}, Pujita Vaidya¹, Namita Vaidya¹, Chapman Beekman¹, Christina Wong¹, David Y. Rhee¹, Odise Censaj¹, Emily McKillop¹, Saumini Shah¹, Mark Stapleton², Kenneth H. Warr², Charles YU², Bayan Parsa², Joseph W. Carlson², Xiao Chen², Bhavleen Kappadia², K. VijayRaghavan², Steven P. Gygi¹, Susan E. Celis², Robert A. Obar¹, Spyros Artavanis-Tsakonas^{1,4}



The screenshot shows the PubMed website interface. At the top, there is a navigation bar with "NCBI Resources" and "How To" links, and a "Sign in to NCBI" button. Below this is the "PubMed.gov" header with a search bar containing "PubMed" and a "Search" button. The main content area features a "PubMed" section with a description: "PubMed comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include full-text content from PubMed Central and publisher web sites." A red circle highlights the text "PubMed comprises more than 24 million citations". To the right of this text, a blue-bordered box contains the text: "MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed." Below the main content are three columns of links: "Using PubMed" (including Quick Start Guide, Full Text Articles, FAQs, Tutorials, and New and Noteworthy), "PubMed Tools" (including Mobile, Citation Matcher, Batch Citation Matcher, Clinical Queries, and Topic-Specific Queries), and "More Resources" (including MeSH Database, Journals in NCBI Databases, Clinical Trials, E-Utilities (API), and LinkOut).

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










The screenshot shows the "goPubMed" search interface. It features the logo "goPubMed" with the tagline "searching is now sorted!". Below the logo is a search bar with the text "All" and a green "find" button. Underneath the search bar is a green folder icon labeled "Bioinformatics Tools".

Technics of TextMining

- Analyses of the abstract content
- Ontology and semantics

★ "Alzheimer Disease"[mesh] ✕ find

 show abstracts
  [documents](#)
  statistics
  top author
  clipboard
  share
  export

2,892 documents found

Rodent models of neuroinflammation for Alzheimer's disease.

 **Authors:** Nazem, Amir, Sankowski, Roman, Bacher, Michael, Al-Abed, Yousef

 **Journal:** Journal of neuroinflammation (J Neuroinflammation), Vol. 12 (1): 74, 2015

Alzheimer's disease remains incurable, and the failures of current disease-modifying strategies for Alzheimer's disease could be attributed to a lack of in vivo models that recapitulate the underlying etiology of late-onset Alzheimer's disease. The etiology of late-onset Alzheimer's disease is not based on mutations related to amyloid- β (A β) or tau production which are currently the basis of in vivo models of Alzheimer's disease. It has recently been suggested that mechanisms like chronic neuroinflammation may occur prior to amyloid- β and tau pathologies in late-onset Alzheimer's disease. The aim of this study is to analyze the characteristics of rodent models of neuroinflammation in late-onset Alzheimer's disease. Our search criteria were based on characteristics of an idealistic disease model that should recapitulate causes, symptoms, and lesions in a chronological order similar to the actual disease. Therefore, a model based on the inflammation hypothesis of late-onset Alzheimer's disease should include the following features: (i) primary chronic neuroinflammation, (ii) manifestations of memory and cognitive impairment, and (iii) late development of tau and A β pathologies. The following models fit the pre-defined criteria: lipopolysaccharide- and PolyI:C-induced models of immune challenge; streptozotocin-, okadaic acid-, and colchicine neurotoxin-induced neuroinflammation models, as well as interleukin-1 β , anti-nerve growth factor and p25 transgenic models. Among these models, streptozotocin, PolyI:C-induced, and p25 neuroinflammation models are compatible with the inflammation hypothesis of Alzheimer's disease.

PubMed [25890375](#) [Related Articles](#) [Read Full Text](#)

Affiliation: Elmezzi Graduate School of Molecular Medicine, The Feinstein Institute for Medical Research, 350 Community drive, Manhasset, NY, 11030, USA. anazem@nshs.edu. EIr ... ▶

Wikipedia: Zanosar, Transgene, Streptozocin, Lipopolysaccharides, Beaver, Pathology, Pathologies, Okadaic acid, Mutation, Colchicine, Immunity, Alzheimer's Disease, Presenile d ... ▶

Protein: nerve growth factor

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

Data

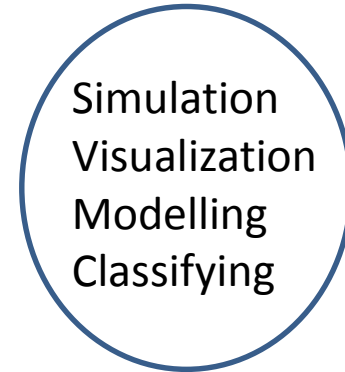
- Storage
- Integration and organization
- Analyses
- Standardization

high throughput methods



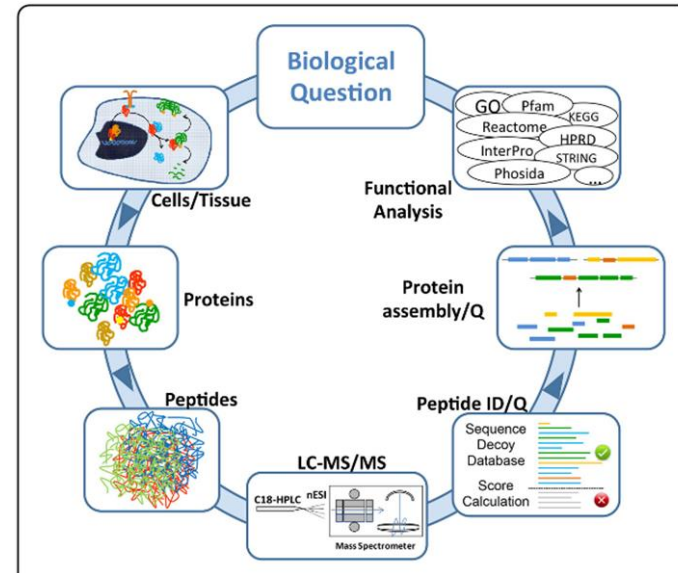
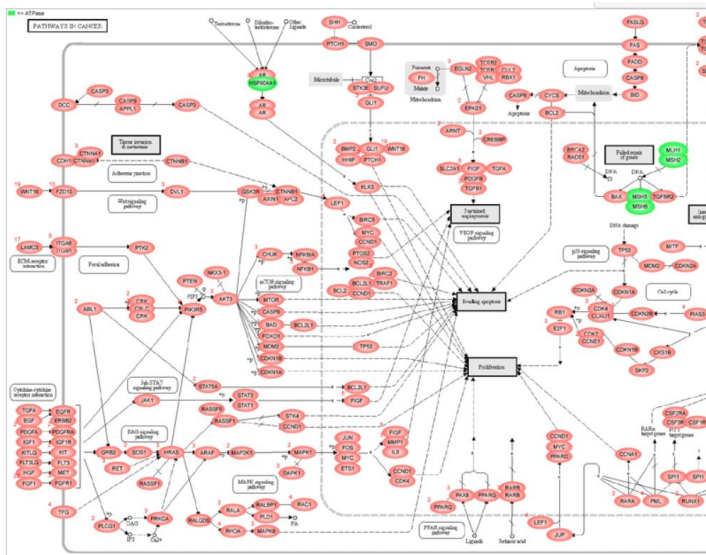
*Big data
Ontologies*

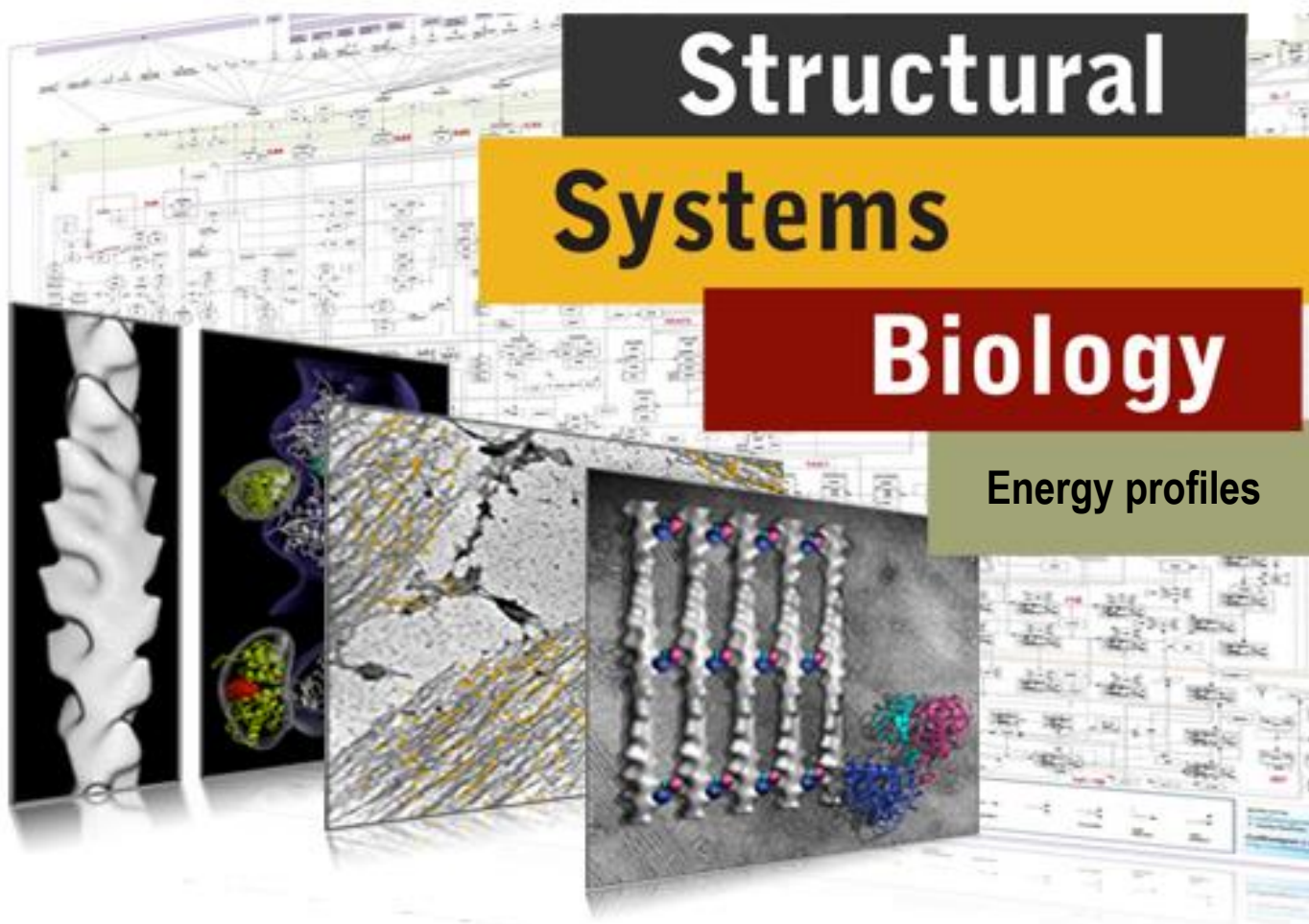
Databases and query languages



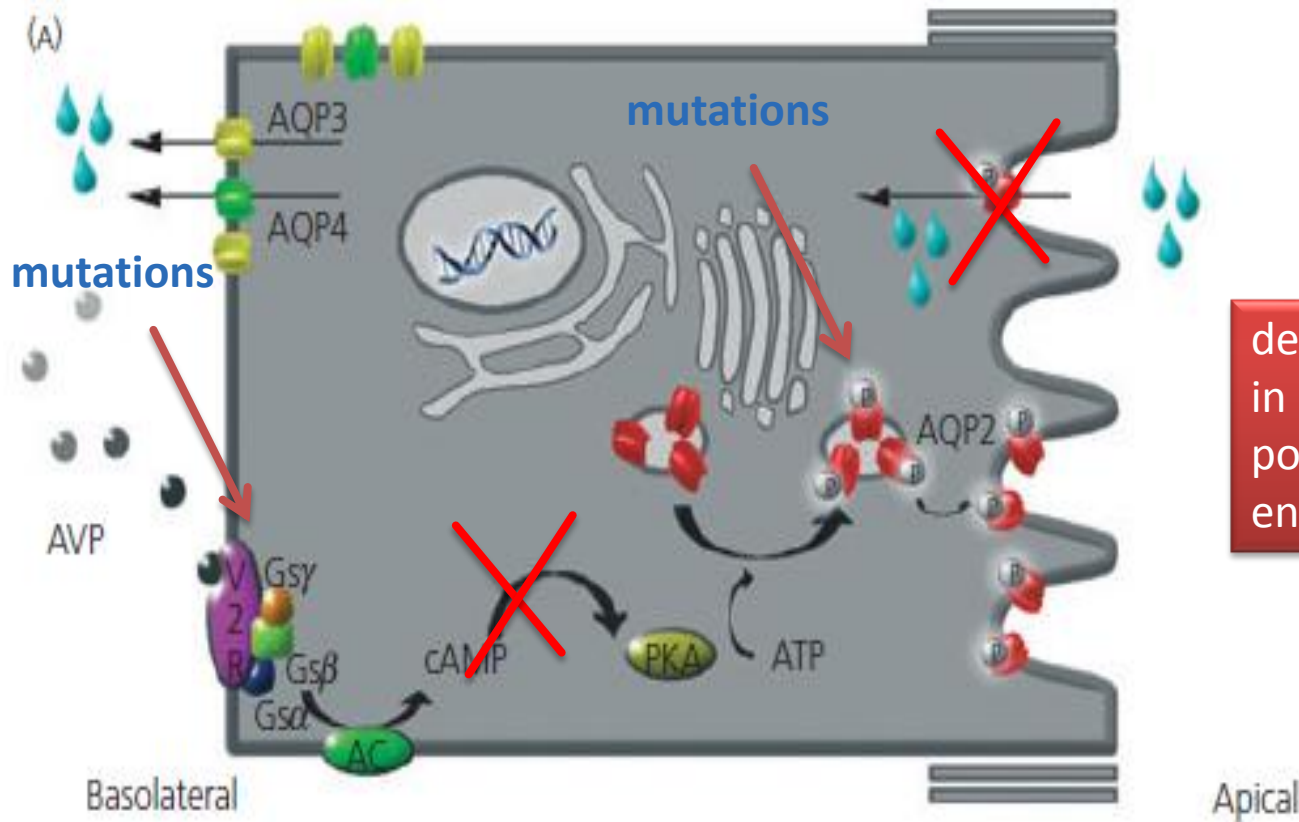
Evaluation/
validation

Statistics





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



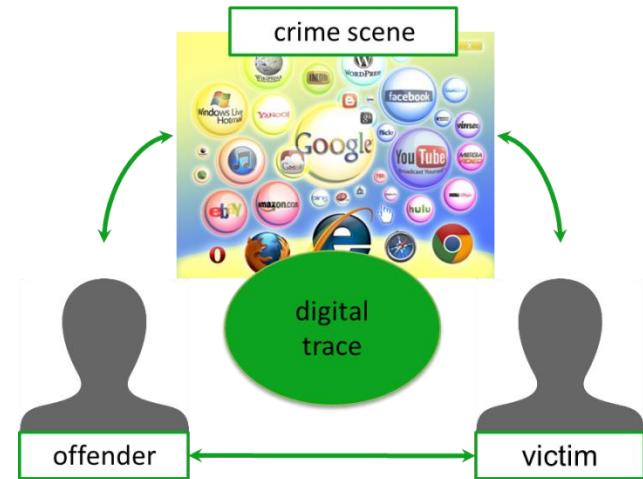
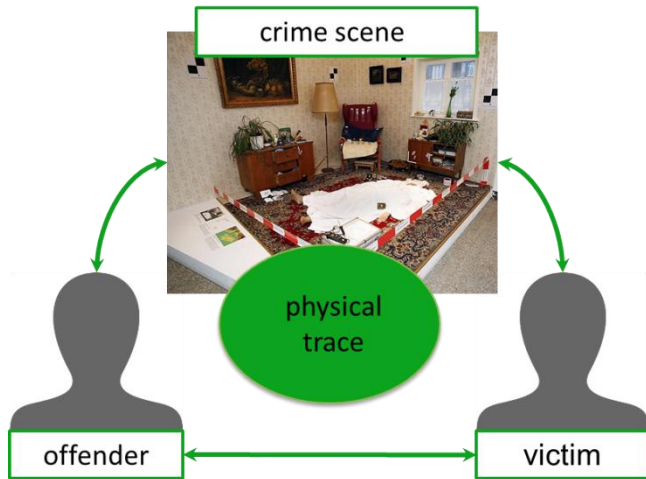
Cofounder of the modern biology

Dr. John H. *Watson*



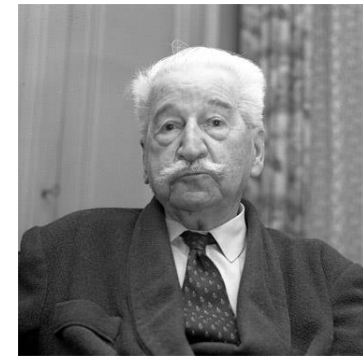
Investigator –
literary character

Classical forensics and digital forensics



Locard's exchange principle

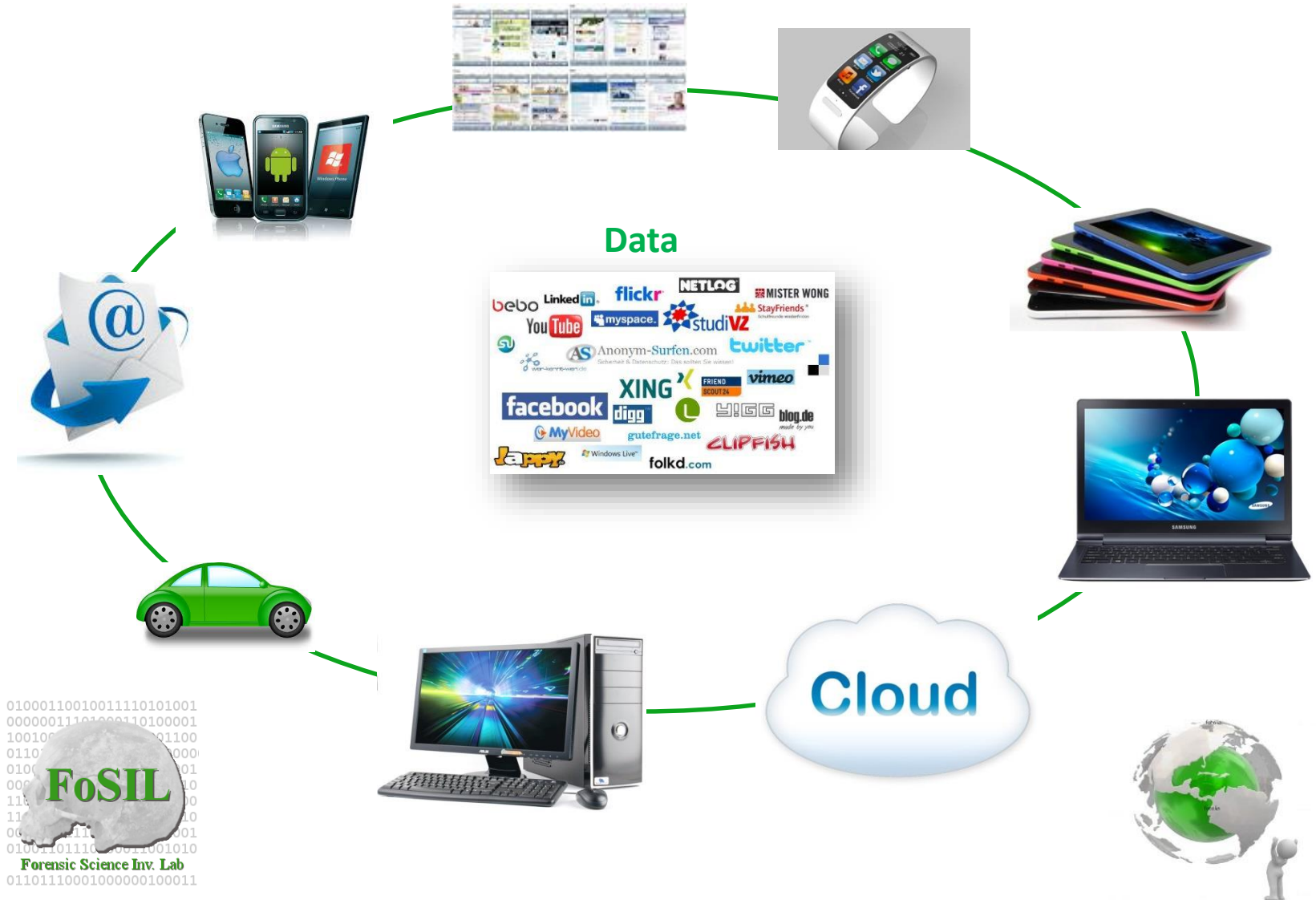
holds that the perpetrator of a crime will bring something into the crime scene and leave with something from it, and that both can be used as forensic evidence



Dr. Edmond Locard



Our daily Life



01000110010011110101001
00000011101000110100001
100100011100
011000000
0100010
000100
110000
110000
00010001
01001101100011001010
01001101100011001010
0110111000100000100011

Classical forensics and digital forensics

- Analyses
- Validation
- **Evaluation**

circumstances of a crime

Physical or digital trace



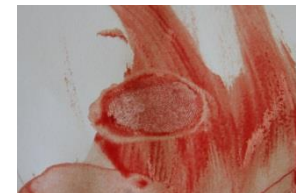
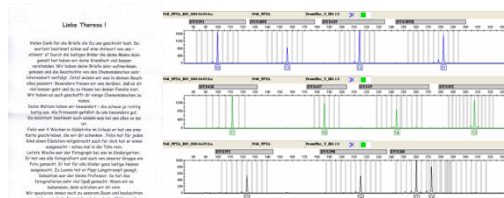
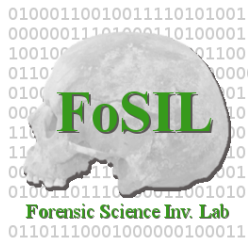
evidence



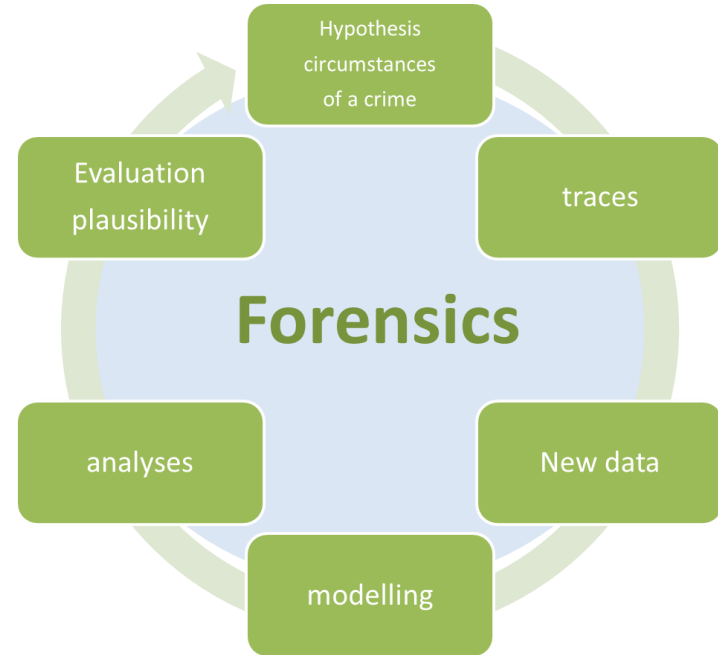
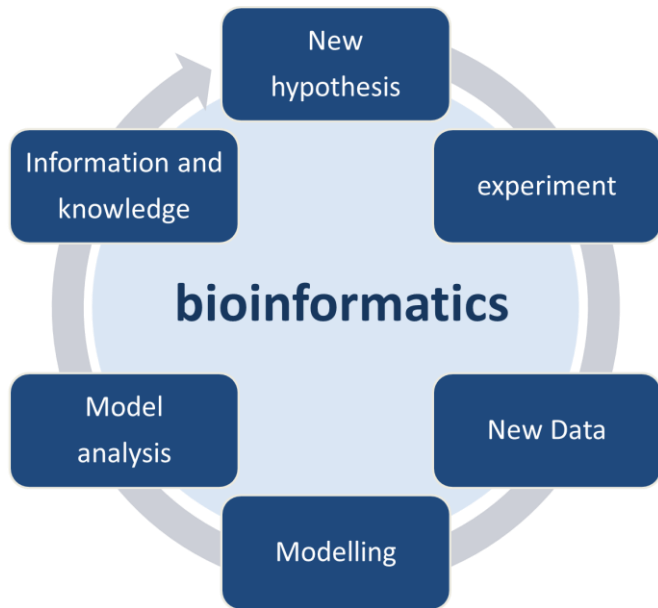
- fingerprint
- DNA traces
- traces of blood and pattern
- Texts
- Short messages
- Chats
- ...

comparison of traces:

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



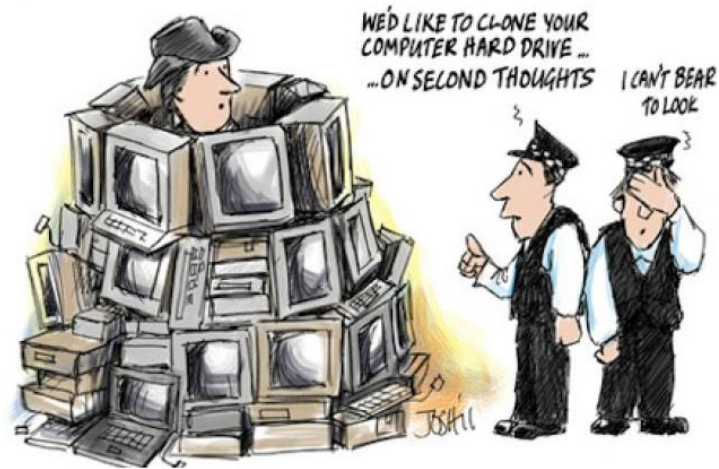
Classical forensics and digital forensics -With respect to bioinformatics -



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.

Classical forensics and digital forensics

-With respect to bioinformatics -

Pre-Process

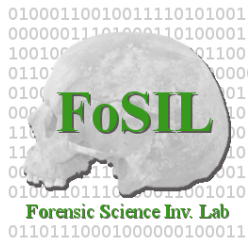
- ▶ text categorization
- ▶ separation of case-relevant files
- ▶ text extraction/OCR
- ▶ modelling of a crime ontology

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

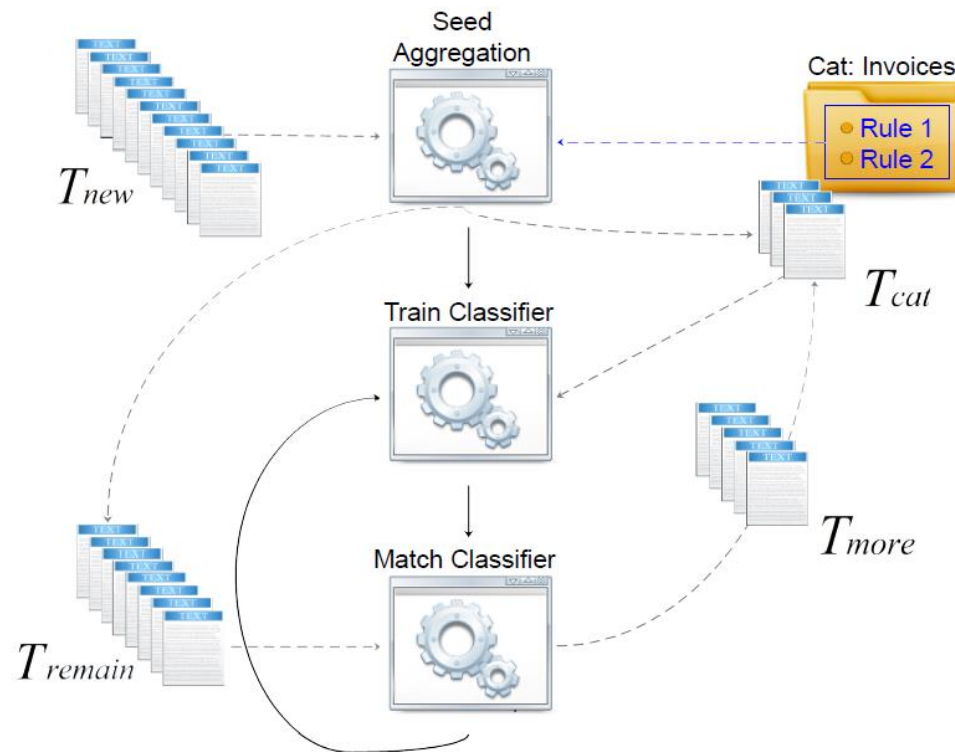
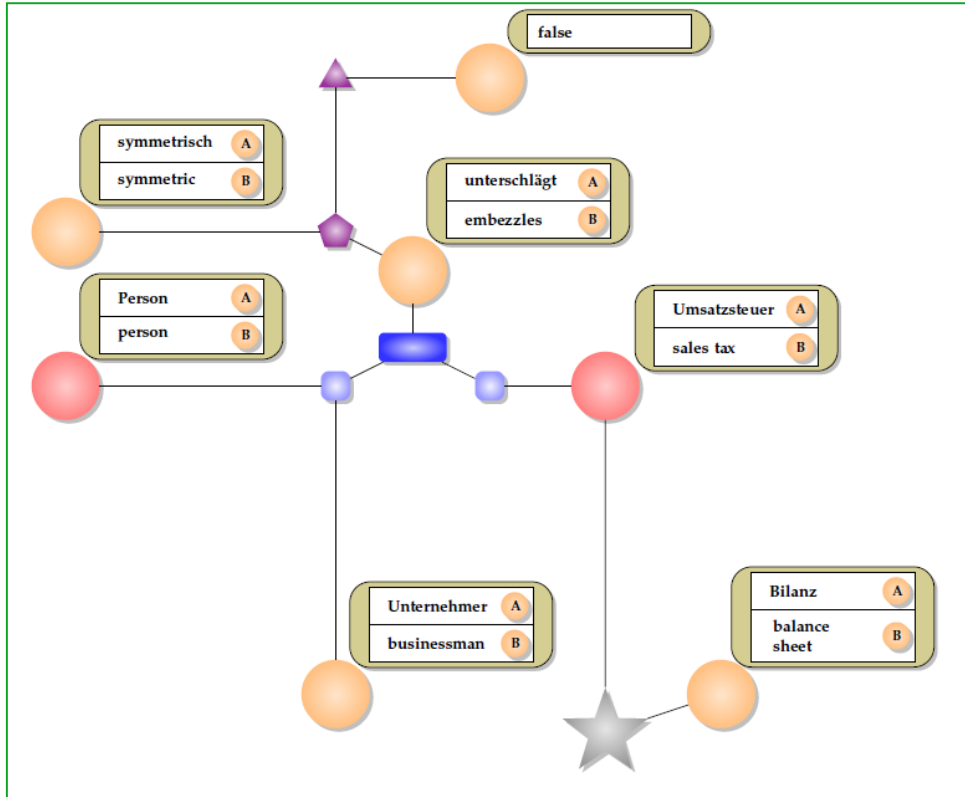


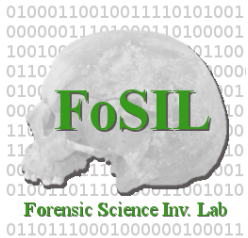
Figure 6: Bootstrapping algorithm for classifying forensic texts. From the texts T_{new} a set of seed documents for each category is acquired using the rules annotated in the taxonomy. This set T_{cat} is used to train one initial weak binary classifier per category. Subsequently, this classifier is 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



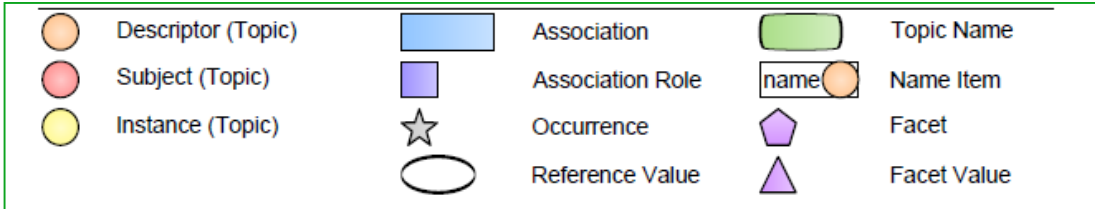
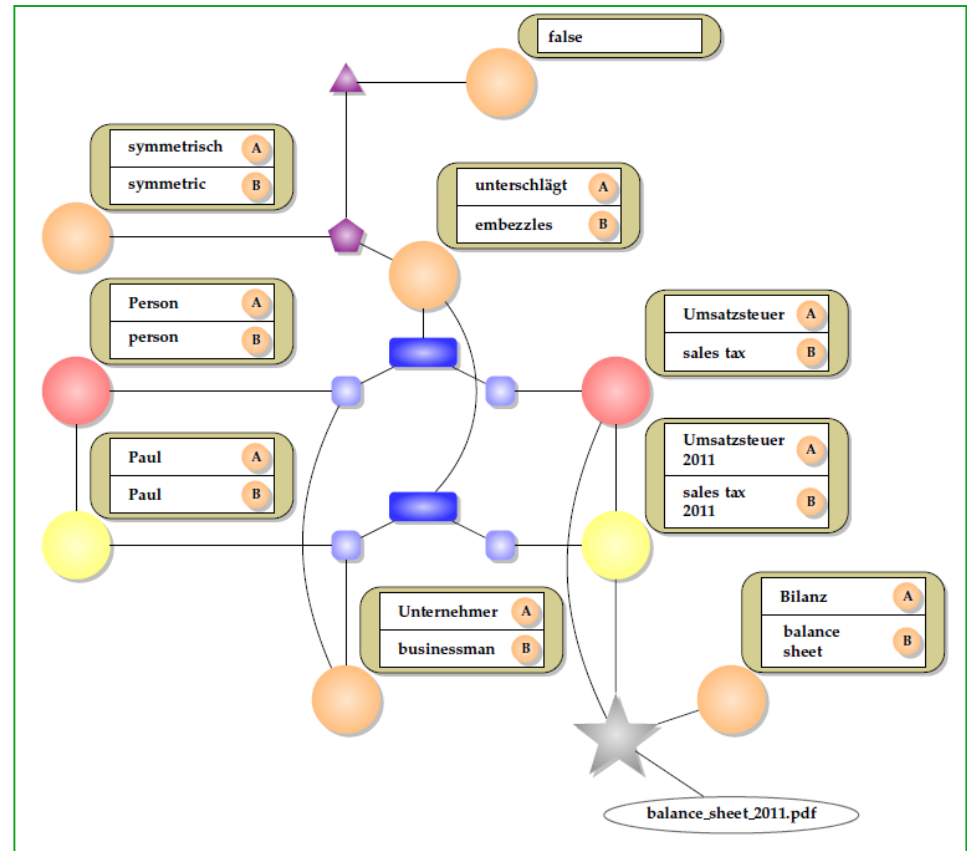
- oriented in the Topic Map ISO-standard
- readable for humans and processable for computers

Ontologies und Semantics

Forensische Topic Map
instantiation

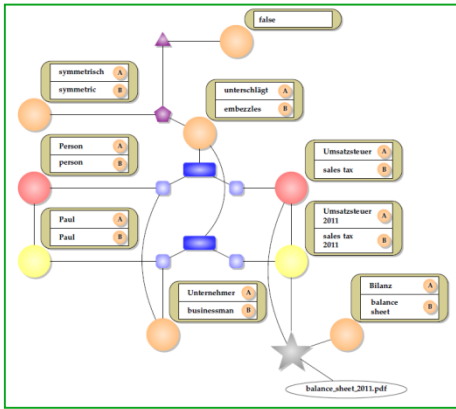


Question Answering System

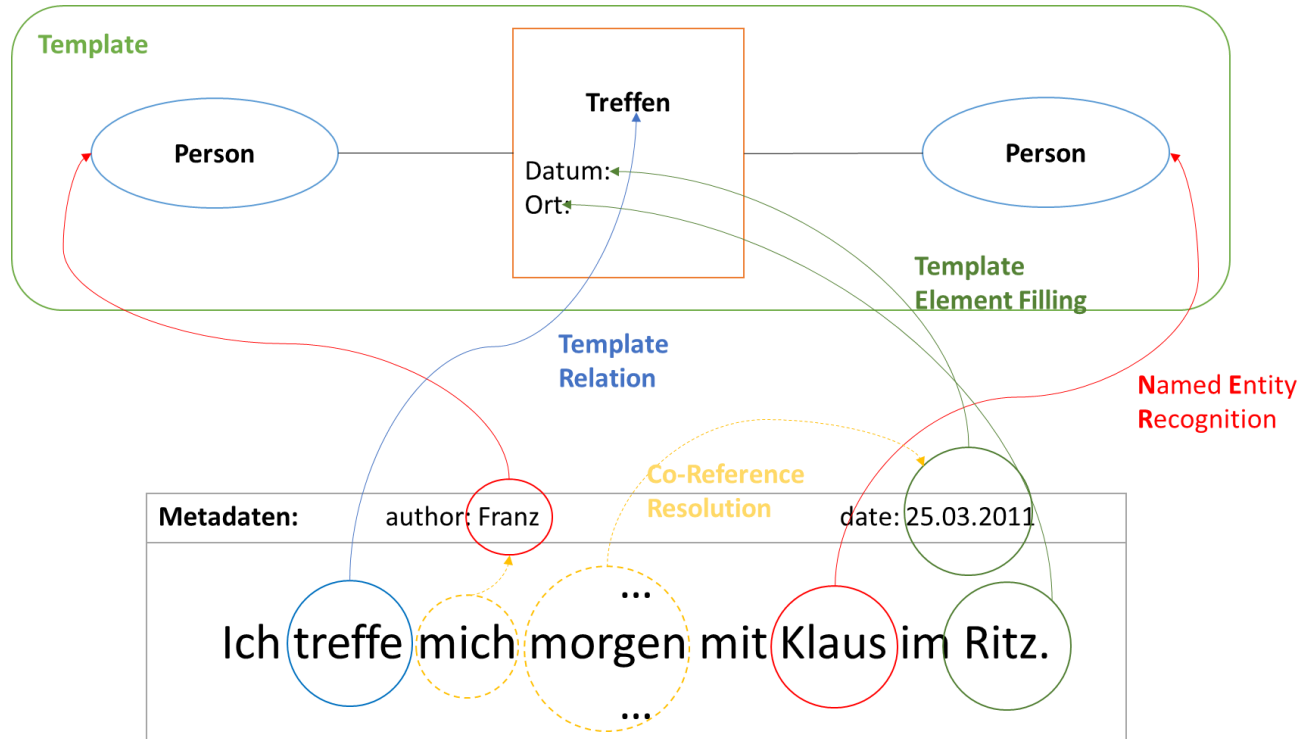


Texts – Information extraction

Forensische Topic Map instantiation



Information Extraction



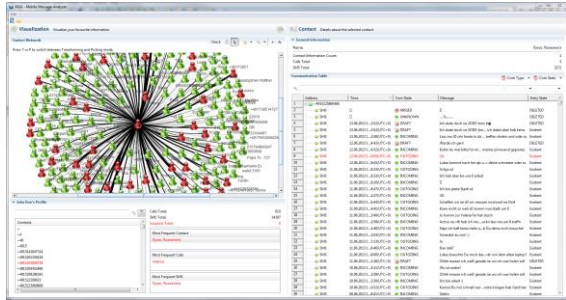
Ontologies und Semantics

The screenshot displays the 'Semantic Tools for Forensic' application. On the left, the 'Topic Map View' shows a network of concepts: 'Mensch' (Human) is connected to 'Fahrer' (Driver) and 'Fahrerlaubnis' (Driver's License). 'Fahrer' is connected to 'Fahrerzeug' (Vehicle) and 'Fahrerlaubnis'. 'Fahrerzeug' is connected to 'Fahrer' and 'Fahrerlaubnis'. Other concepts include 'Opfer' (Victim), 'Tatort' (Crime Scene), 'Wohnort' (Residence), 'Stadt' (City), 'Privat' (Private), 'Mann' (Man), 'Frau' (Woman), 'Merkmal' (Feature), 'besitzt' (Owns), 'kennt' (Knows), 'Zeuge' (Witness), 'Motorrad' (Motorcycle), 'föhrt' (Drives), 'wohnt in' (Lives in), 'fährt' (Drives), 'zugelassen auf' (Issued to), 'raubt' (Steals), 'führt' (Leads), and 'Gefährter' (Endangered). The right pane shows a list of files and a 'Details' panel for the file 'Anja Kreher 200904204 Teilrückzahlung Darlehen 2010-04-30.xls'. The details include: Title, Author, URN, Size (397.84 KB), Created (18.12.2012 16:47:38), Last Accessed (30.04.2010 11:08:46), Modified (18.12.2012 16:47:38), Owner (VORDEFINIERT\Administratoren), Categories (Datei), and Additional properties.

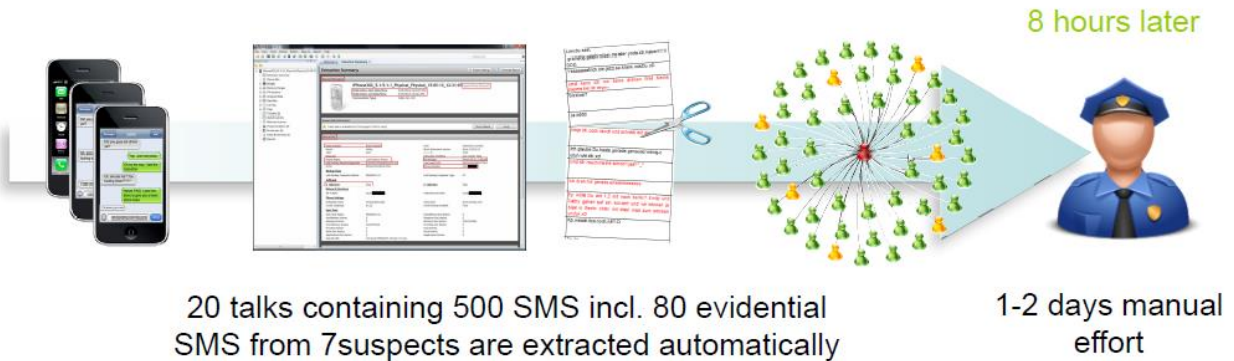
SEMANTIC TEXT ANALYZER



Ontologies und Semantics



Mona
Mobile Message ANALYZER



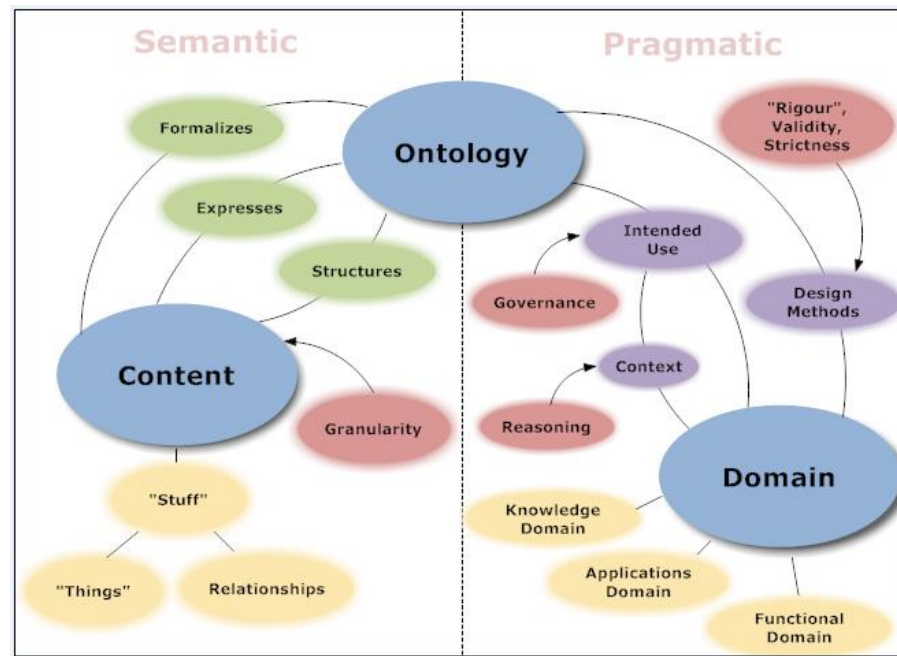
20 talks containing 500 SMS incl. 80 evidential SMS from 7suspects are extracted automatically

100% Precision und 67% Sensitivity

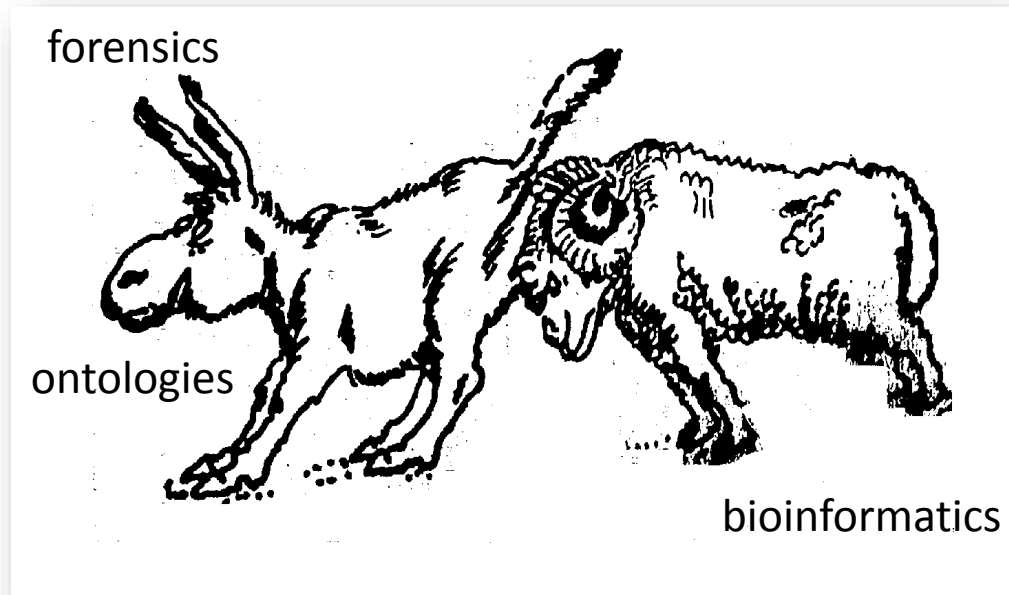


An ontology is a specification of a conceptualization.

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



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



labudde@hs-mittweida.de

www.bioforscher.de