Semantic resources for biomedical data integration: phenotypes

Presentation

Dietrich Rebholz-Schuhmann
University of Zurich

Faculdade de Medicina
da Universidade do Porto

Outline

Needs for phenotype data

A phenotype is . . . :

- Characterization of model organisms: to measure phenotype modifications in genetic experiments
- Judgement of medical patients: to assess the medical status of the patient
- Translational medicine: to judge a patient (or a drug) against the findings from biomedical research
- Biomedical data integration: to exchange data from the bench to the bedside, core research vs. lab results vs. anamnesis of a patient

What is a phenotype?

A phenotype is . . . :

- an observable trait concerning the representation of a body or organism (by measurement, by human judgment)
- induced by the genotype and the environment (e.g. increased size of muscle)
- giving clues to the understanding of the functionning of
  - the genotype or
  - the treatment,

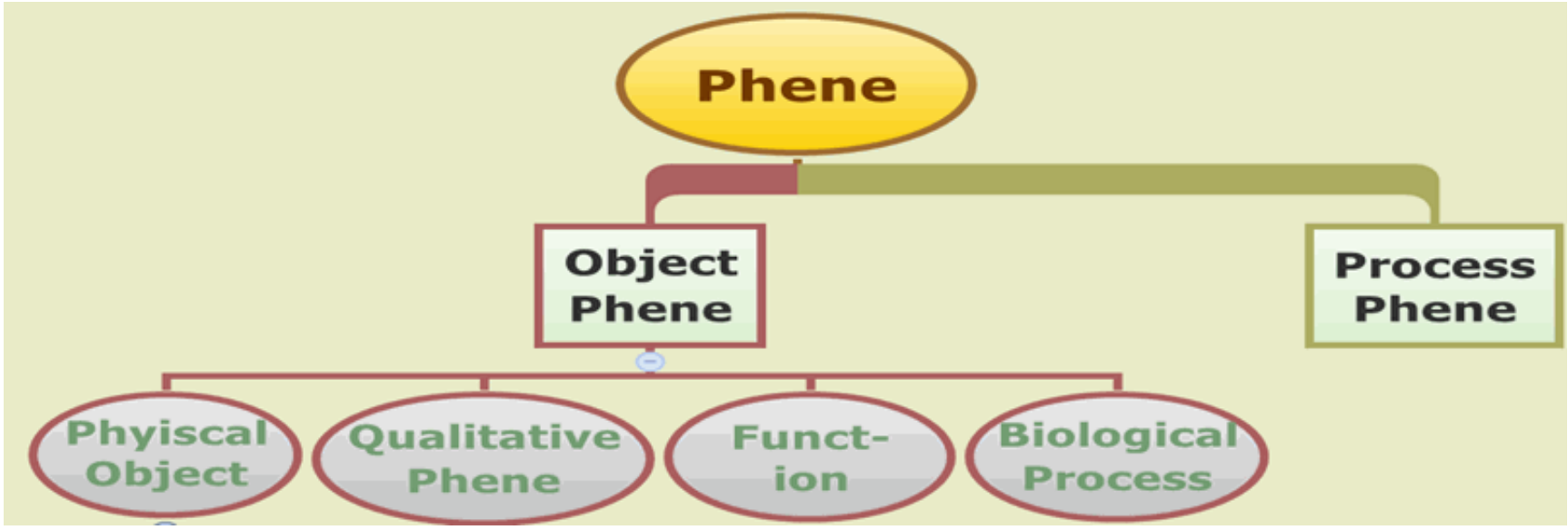
  if we resolve the dependencies
- may look quite different (human vs. mouse) although the genetic background is very similar (about 96 %).

Phenotype 'aging'?

Aging is . . . :

- (Wikipedia) . . . the accumulation of changes in a person over time
- Not a disease, but changes induced by diseases.
- Not a process, but changes could be induced by any processes.
- Changes are neither positive nor negative.
- It is a status as the result of changes.
- Is it a phenotype?
  The signs and symptoms of aging are phenotypes, i.e. changes to the expression of the functioning of the body.

Types of phenotypes



Hoehndorf, R., A. Oellrich und D. Rebholz-Schuhmann (2010). "Interoperability between phenotype and anatomy ontologies". In: Bioinformatics 26, S. 31123118.

Typical examples of phenotypes – in ontologies

- Increased / decreased size of an organ, cells, tissues, e.g. megalosplenie.
- Increase / decrease of functional physiological and metabolic processes, e.g. increase or decrease of the heart rate, the blood pressure, the blood glucose tolerance.
- Existing / Missing structural components, e.g. abnormal connections between vessels in the lungs or in the brain.
- Molecular abnormalities, e.g. missing enzymatic activities in genetic metabolic disorders.

Aging related phenotypes

- Reduced physiological functions:
  e.g. tissue perfusion, tissue regeneration, lower O2 perfusion of the brain
- Reduced metabolic functions:
  e.g. prediabetic conditions, increased blood sugar due to reduced insulin activity
- Changes to the body structure:
  e.g. lost elements (hair, teeth, kidney), broken bones

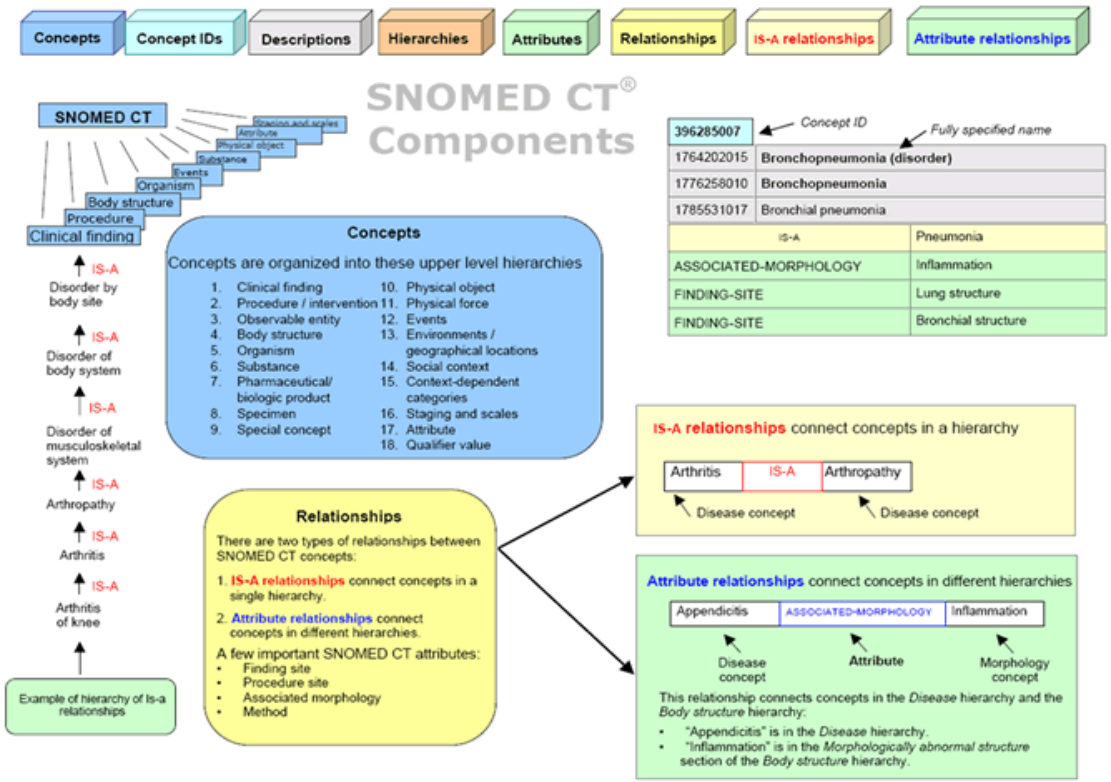
Sources of phenotypes

Description

Angelman syndrome is a neurodevelopmental disorder characterized by me
balance disorder, typical abnormal behaviors, and severe limitations in speec
caused by absence of a maternal contribution to the imprinted region on ch
Willi syndrome (PWS; 176270) is a clinically distinct disorder resulting from
15q11-q13 region. In addition, the chromosome 15q11-q13 duplicatio
overlapping clinical features.

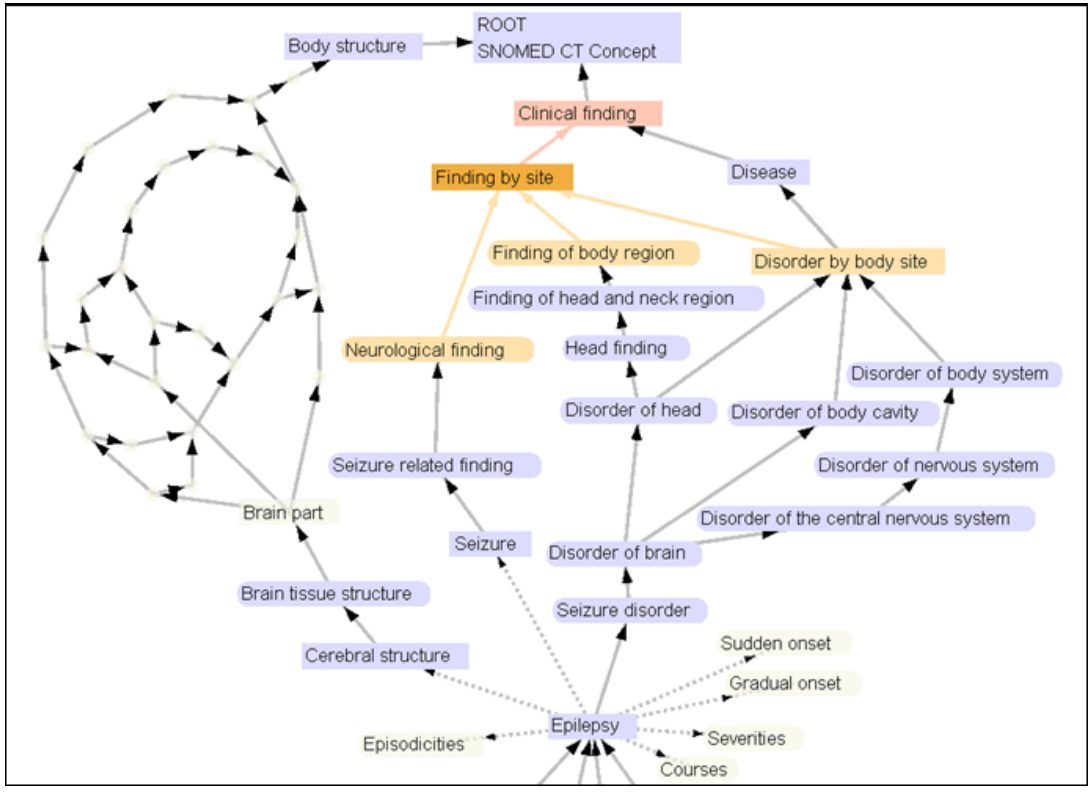
Phenotypes as free form text from the patient recored:
requires feature extraction,
morphological / syntactical normalisation,
mapping to semantic resources

Sources of phenotypes



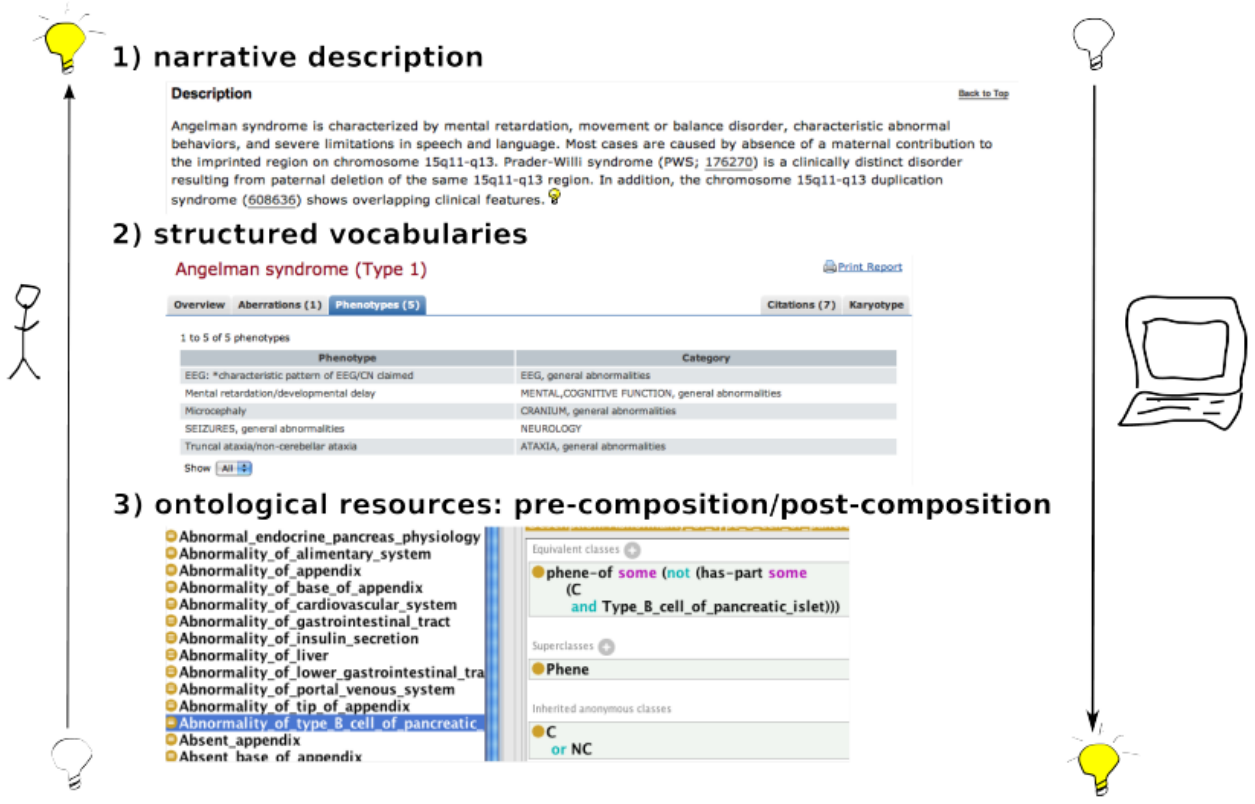
Phenotypes from public terminologies (Snomed):
mainly collection of terms, limited structure for semantic support.

Sources of phenotypes



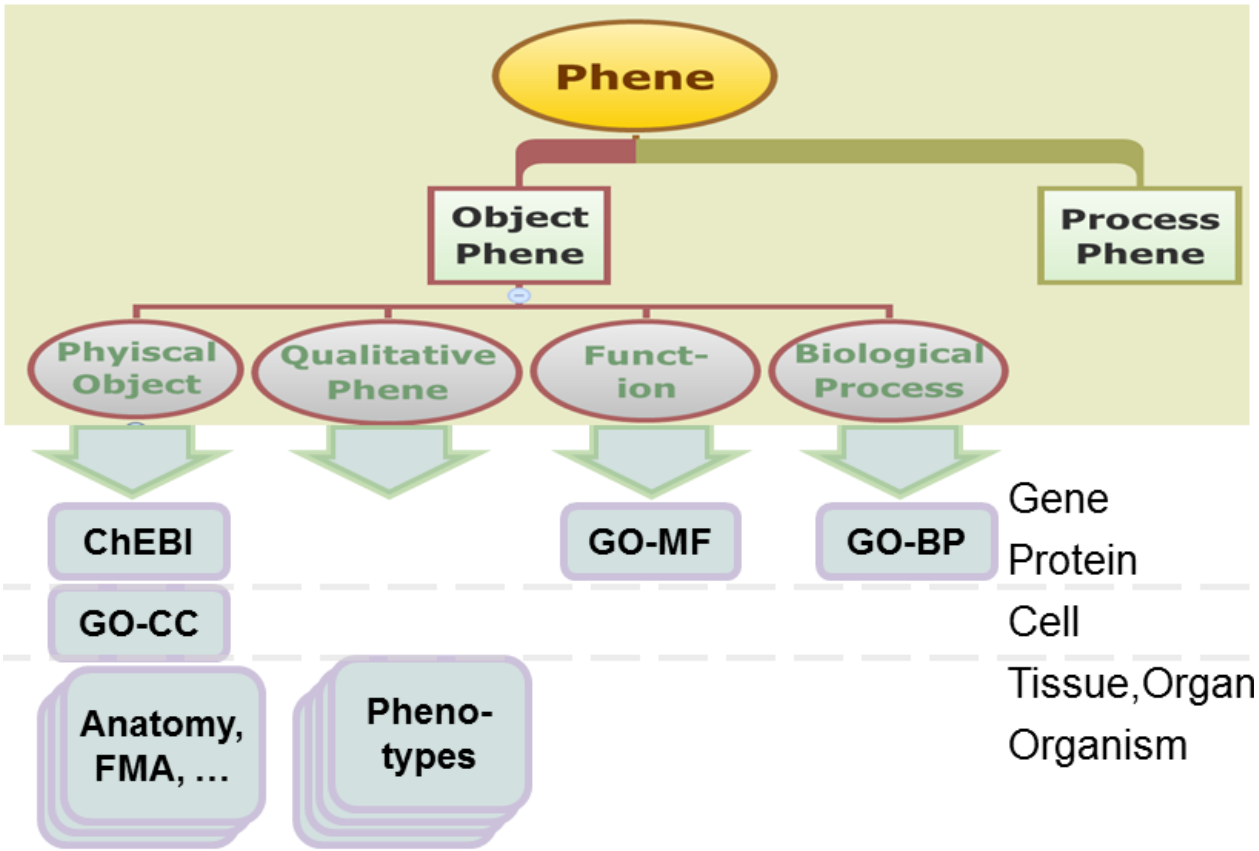
Phenotypes as a tree structure:
complex, well represented, difficult to build
Comprehensiveness impaired due to time-consuming development.

Phenotypes of different kinds



Phenotypes as (1) free-form text (patient recored), (2) predefined terminology, und (3) formal ontology.

Sources of phenotypes



Hoehndorf, R., A. Oellrich und D. Rebholz-Schuhmann (2010). "Interoperability between phenotype and anatomy ontologies". In: Bioinformatics 26, S. 31123118.

Example phenotypes

- HPO hearing loss
  MP hearing loss
- HPO Progressive childhood hearing loss
  MP abnormal hearing physiology – abnormal ear physiology – hearing/vestibular/ear phenotype
- HPO Optic nerve hypoplasia
  MP agnormal optic nerve morphology

Oellrich, A. u. a. (2012). "Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases". In: PLoS ONE 7.6, e38937.

Exploitation of phenotypes in research

- Comparison of model organisms with humans
  $\Rightarrow$ prediction/analysis of phenotypes induced by genes
- Influence of SNPs and genes on phenotypes
  $\Rightarrow$ prediction of adverse side effects of drugs
- Analysis of complex phenotypes
  $\Rightarrow$ modelling and analysis of genuine disease effects vs. treatment side effects
- Altogether: better interpretation of the patients condition

Analysis and normalisation of the patient record (similar to other texts)

- The patient record contains the information on the patient.
- Terminologies are available to normalize the record, but the level of normalisation is not sufficient.
- With text mining methods and statistical methods we can normalise the content.
- Using standardized interfaces adds to the degree of normalisation.
- We can compare and the patient record against other data resources, e.g. the scientific literature.

From the patient record to the patient profile

- Analysis of the patient record
- Generation of the profile: use terminology, statistical distribution, vector representation
- Normalisation of the profile using concept URIs instead of terms

URI = Uniform Resource Identifier

Example text: Sumo Wrestling [Source: Wikipedia]

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.

Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.

Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.

Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.

Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.

Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.

Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.

Statistical Analysis

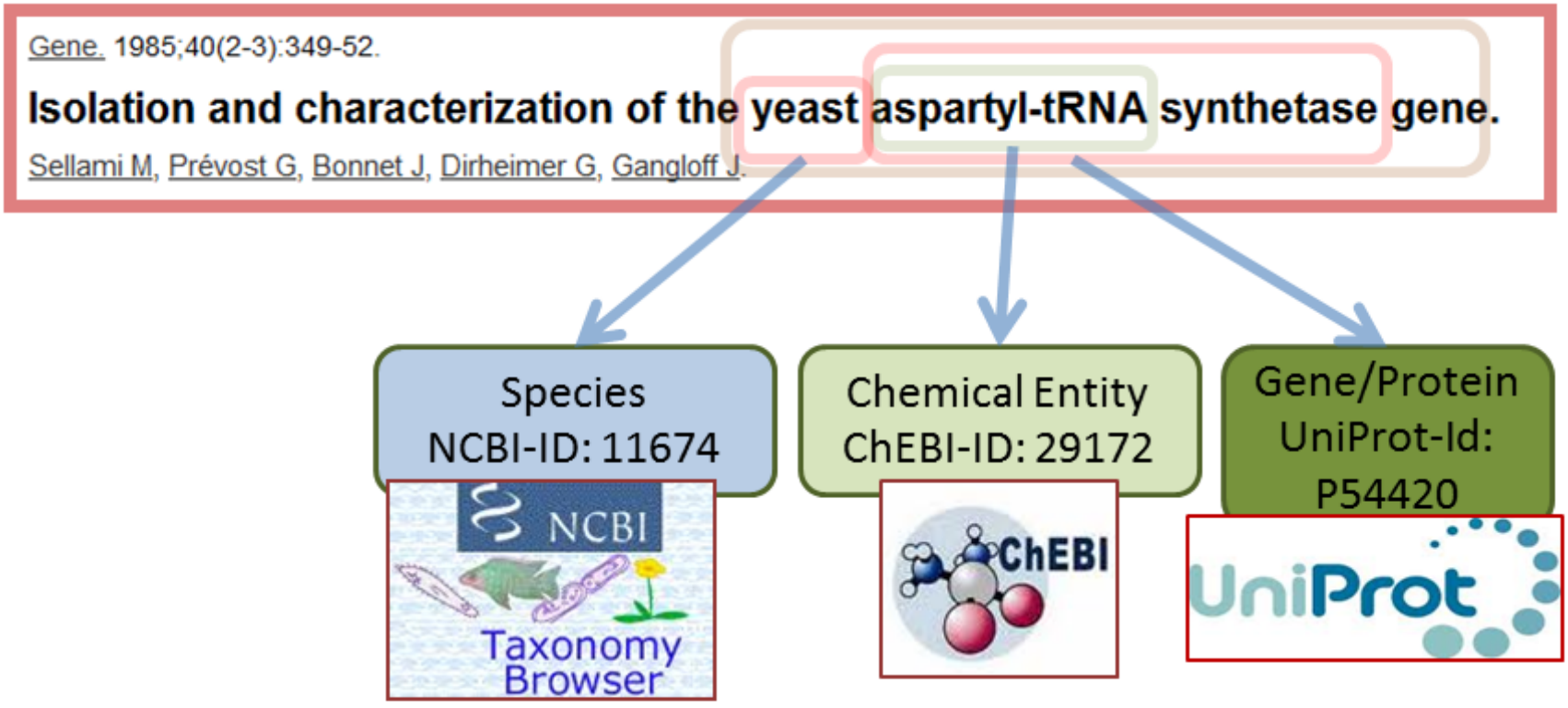
Number	Word
5	wrestler
3	force, ground
2	feet, first, opponent, ring, sumo, touch,
1	competitive, full-contact, sport, rikishi, attempts, ...

Example for Normalisation: SUMO

- SUMO is a martial art.
- SUMO (Suggested Upper Merged Ontology) is an ontology.
- SUMO (Small Ubiquitin-like Modifier) is a protein.
- SUMO (Surgery and Molecular Oncology) is a department of the Dundee University
- But also:
  - Software Update Monitor, Mozilla Support
    Sufficiently Uniform Memory Organization
  - Spacecraft for the Unmanned Modification of Orbits
  - Stanford University Mathematical Organization

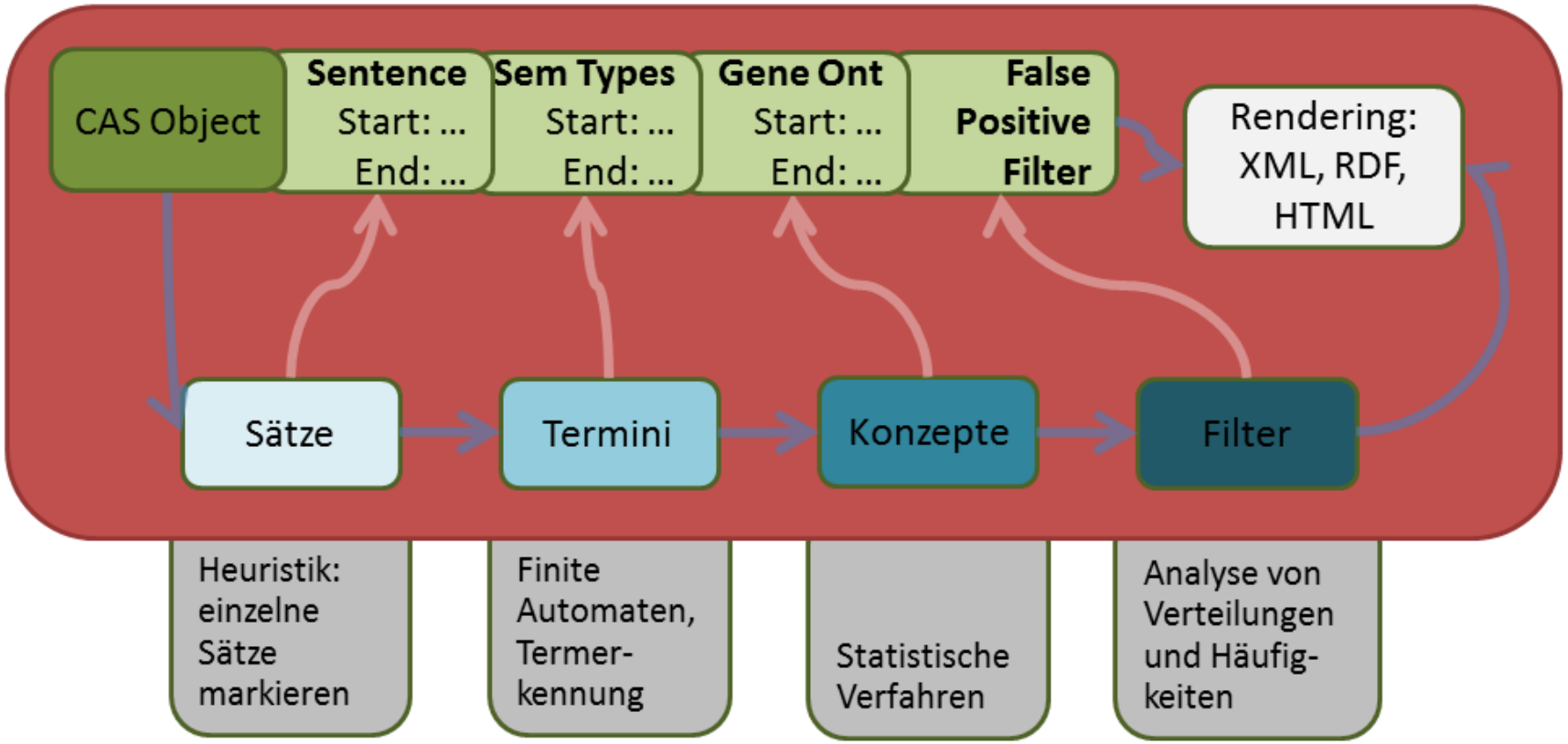
Example from the literature

A number of data resources are available to do the normalisation: : UMLS, ICD, Snomed-CT, UniProt, ChEBI, ...



Biomedizinische Datenbanken

Infrastructure for textual analysis



Semantic categorization of gene/protein naming conventions

Identified from the annotation guidelines for genes/proteins . . .
A gene/protein may represent

- a recognized macromolecule (e.g., "hemoglobulin", "prolactin")
- the function of the gene or protein (e.g., "methyl-transferase")
- (part of) a structure of a protein (e.g., "Cytochrome c oxidase subunit 2")
- (part of) a process (e.g.,"Mitochondrial fission process protein 1")
- an action on a target (e.g., "DNA gyrase inhibitor")
- a phenotype (e.g., "protein hunchback")
- a chemical or physical properties (e.g., "37.8 kD protein")
- a gene/proteine that is similar to another known gene/proteine (e.g., "Myc homolog protein")

Remaining problems in the semantic normalisation

- *mice* versus *MicE protein*
- *left breast cancer* is not a concept:
  - *breast cancer* and *left breast* do exist as concepts
  - Now, is *left breast cancer* different from *right breast cancer*
- *Retinoblastoma*: a disease and a gene, which initiates the cancer type.
- *Streptococcus pneumoniae* (is a species) but not *Streptococcus pneumonia* (is a disease)

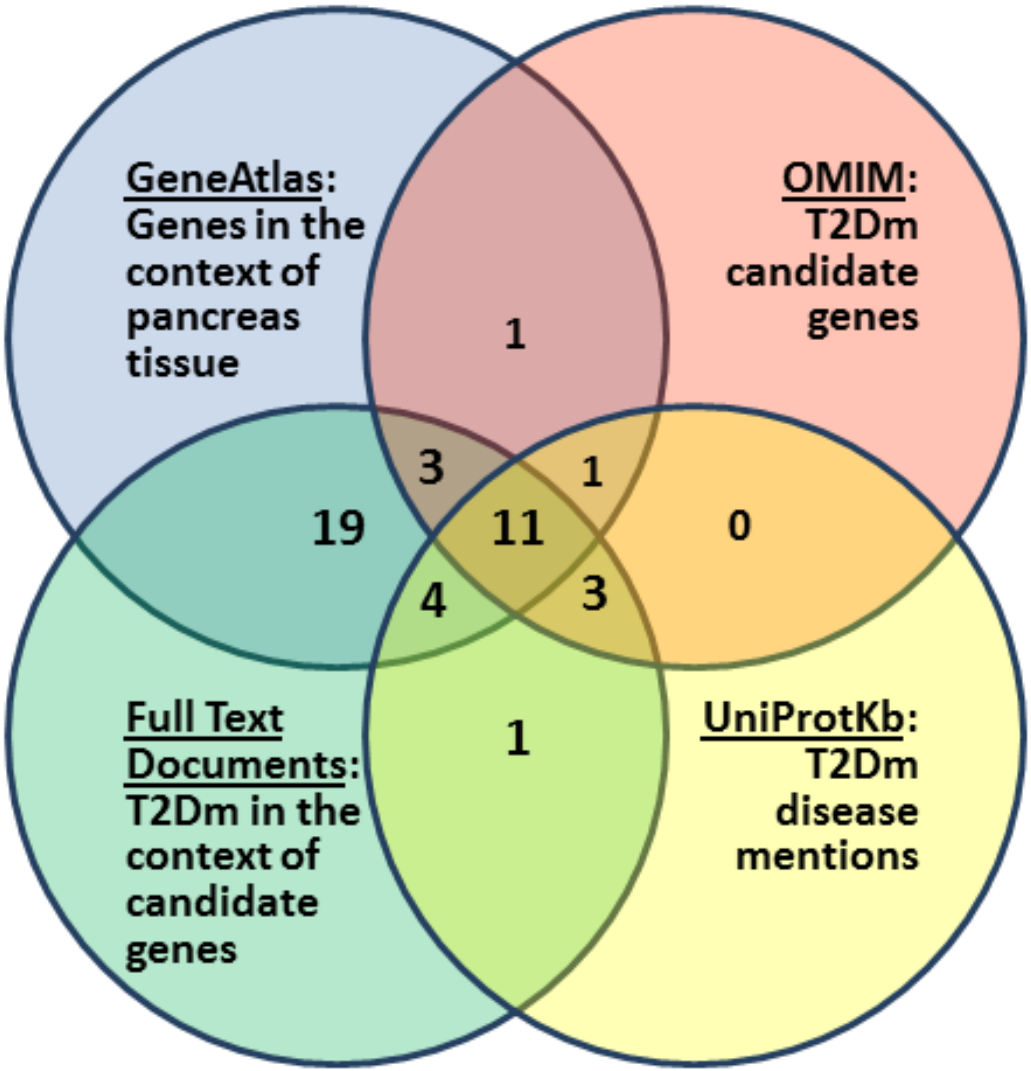
Use scenarios

- Cross evaluation of phenotypes
- Linking genes to diseases
  - Identification of candidate genes for Diabetes mellitus
  - Identification of candidate genes based on GO terms
  - Identification of candidate genes based on phenotype comparisons (mouse / human)
  - Identification of drugs for diseases (PharmGKB, PhenomeNet)
  - Identification of candidate genes for diseases, cross-species comparisons (PhenomeNet)
- Repurposing of drugs

SESL project: Genes linked to Diabetes mellitus type II (DmT2)

	Public & proprietary data	[%]	Public data only	[%]
ArrayExpress	182,840	0.5%	182,840	0.7%
EFO	49,026	0.1%	49,026	0.2%
UMLS, homebrew	6,906,735	18.8%	6,906,735	26.5%
Disease Ontology	1,863,664	5.1%	1,863,664	7.2%
Gene Ontology	495,595	1.3%	495,595	1.9%
UniProt filtered for Human	12,552,239	34.1%	12,552,239	48.2%
Triples on Meta-Data from FT Lit.	3,485,212	9.5%	1,949,293	7.5%
Triples with gene annot. From FT Lit.	2,373,584	6.5%	300,773	1.2%
Triples with disease annot. From FT Lit.	4,983,788	13.6%	662,824	2.5%
Triples with Go annot. From FT Lit.	3,870,834	10.5%	1,099,410	4.2%
Total number of triples	**36,763,517**		**26,062,399**	
Total number of public triples	**14,713,418**	**40.0%**	**4,012,300**	**15.4%**

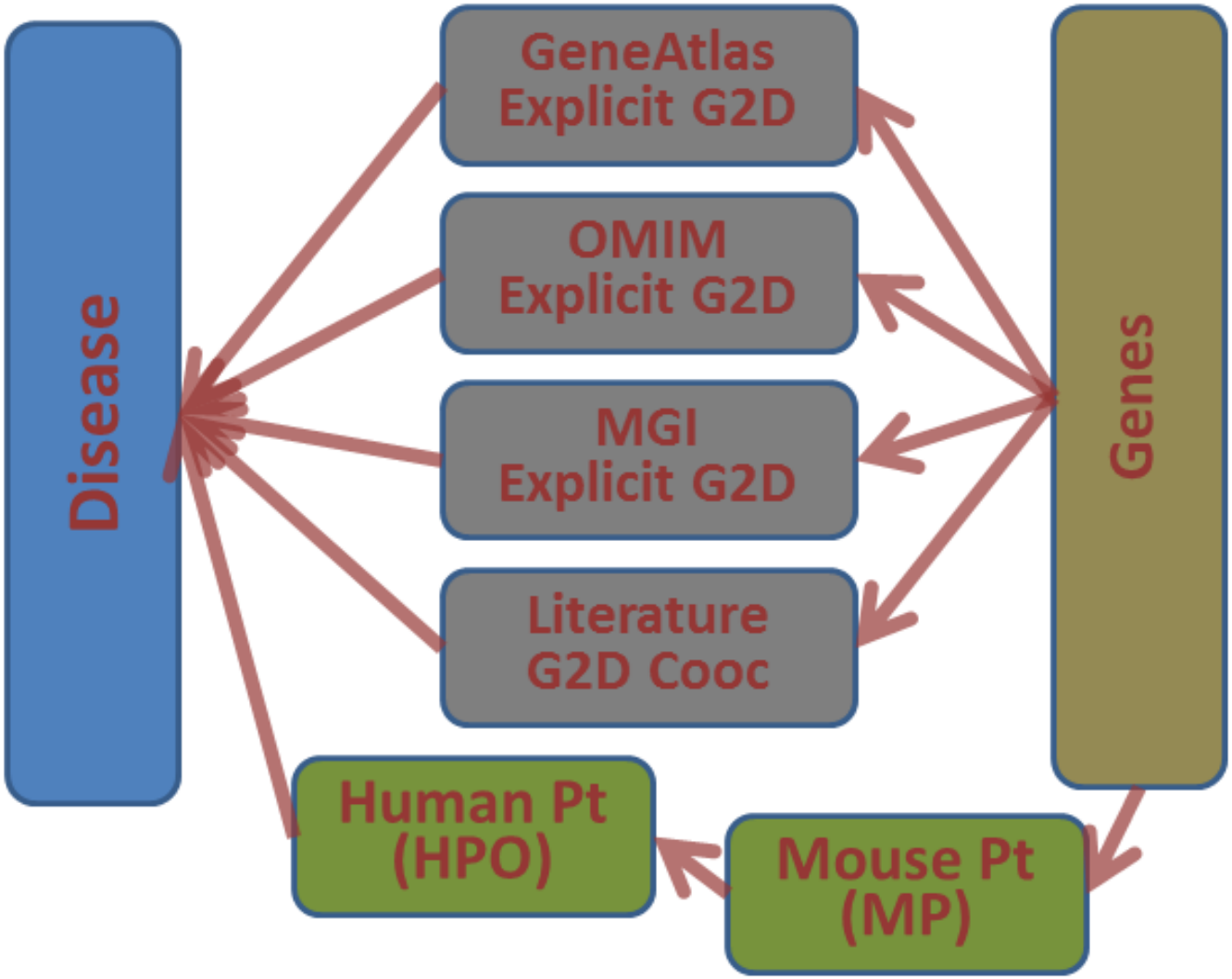
SESL project: Genes linked to Diabetes mellitus type II (DmT2)



SESL project: Genes linked to Diabetes mellitus type II (DmT2)

		OIMIM (+)	OIMIM (+)	OMIM (-)	OMIM (-)
		UniPro (+)	UniProt (-)	UniPro (+)	UniProt (-)
Review (+)	GXA (+)	ABCC8, CAPN10, HNF1A, HNF1B (TCF2), HNF4A, INSR, NeuroD1, PPARG, TCF7L2	WFS1	IRS1, PDX1	HHEX, JAZF1
Review (+)	GXA (-)	GCK, KCNJ11	IGF2BP2		
Review (-)	GXA (+)	MAPK8IP1, PAX4	LIPC, PTPN1	GBP28 (ADIPOQ), PPP1R3A	ACVR2A, ADCP2 (DPP4), ARCN1, FFAR1, GCG, GLP1R, IAPP, IDE, IL1B, MAP4K2, NEFA (NUCB2), NIF3 (CTDSP1), NOS3, PGC1A (PPARGC1A), PPARA, RBP4, UCP2
Review (-)	GXA (-)	SLC2A4	IL6, RETN	INS	

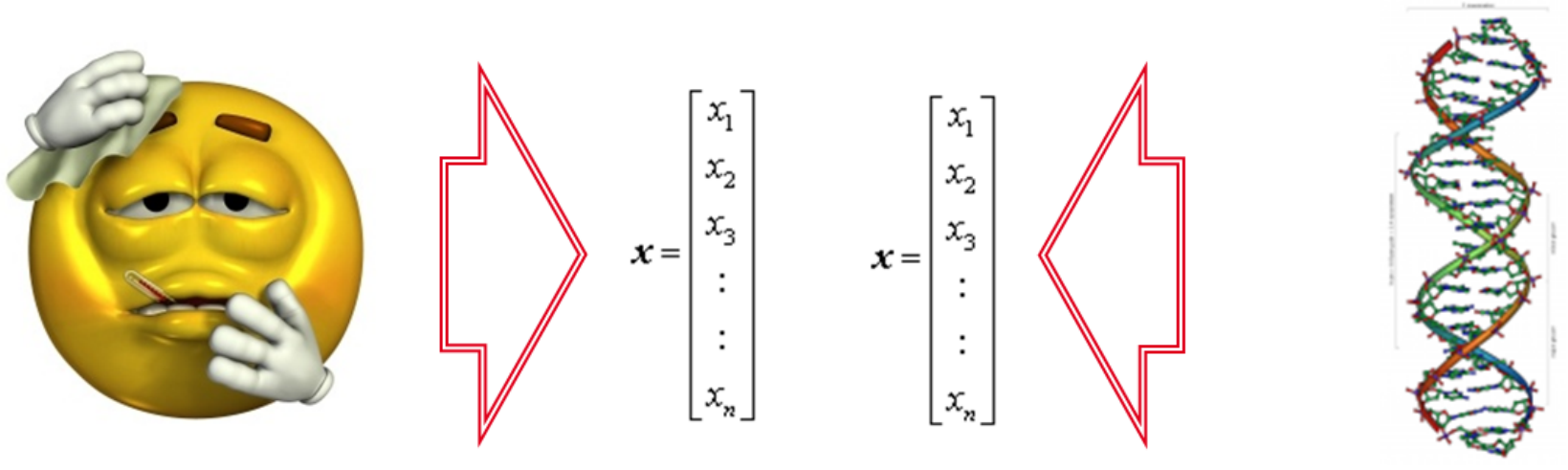
SESL project: Genes linked to Diabetes mellitus type II (DmT2)



Automatic analysis of the patient record for medical diagnostics

- The content is not well standardised.
- The data may be available in electronic form.
- We have efficient means to process the patient record and to draw conclusions.
- We can exploit multi-linguality and can identify hidden conncectsion and can draw conclusions (semi-)automatically.

Comparison of patient and disease profiles



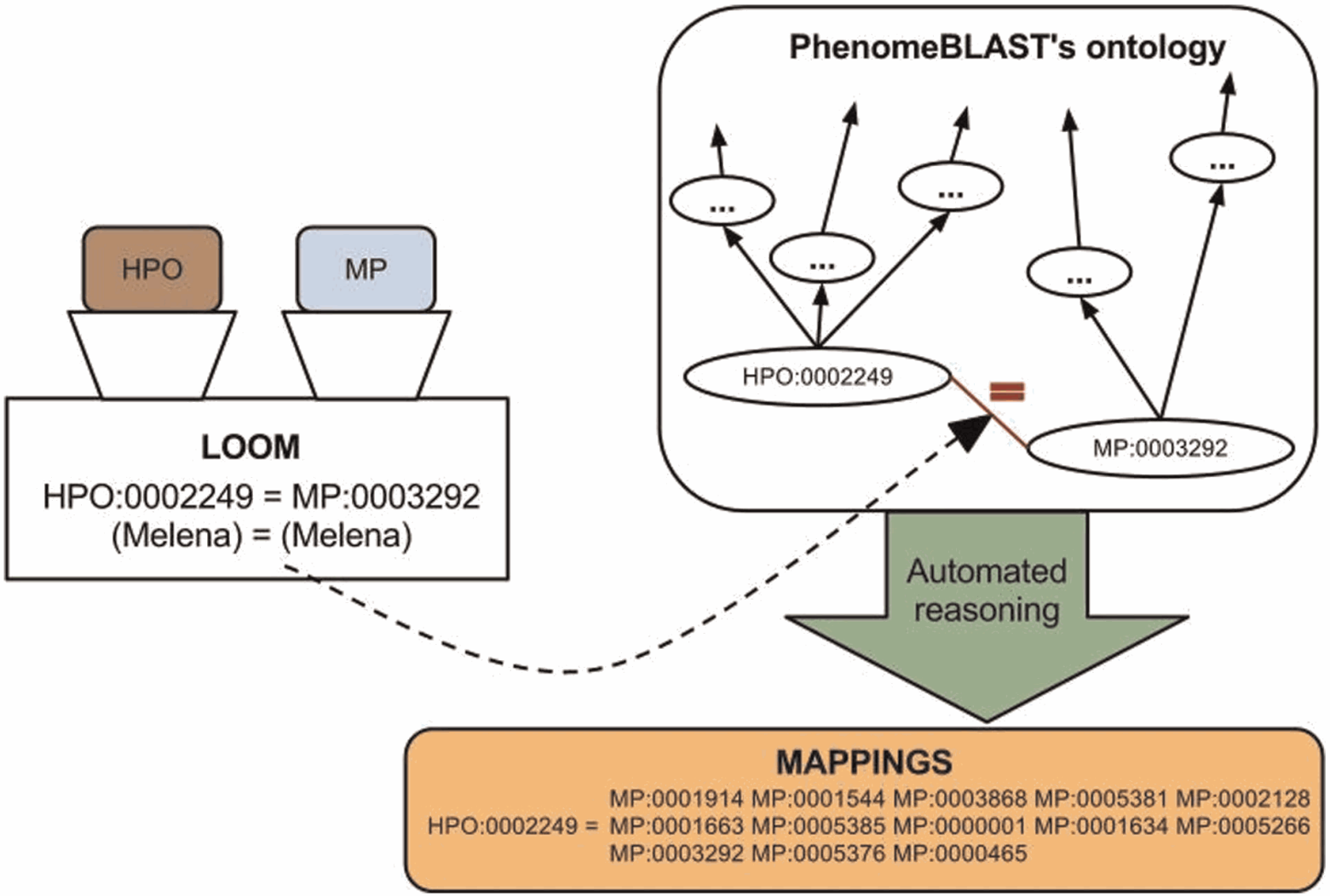
Disease profile (left): Generated from a database (OMIM).
Patient profile (right): Generated from the patient record.
Using similarity measures for the comparison (e.g. Cosinus)

Omim vs. literature

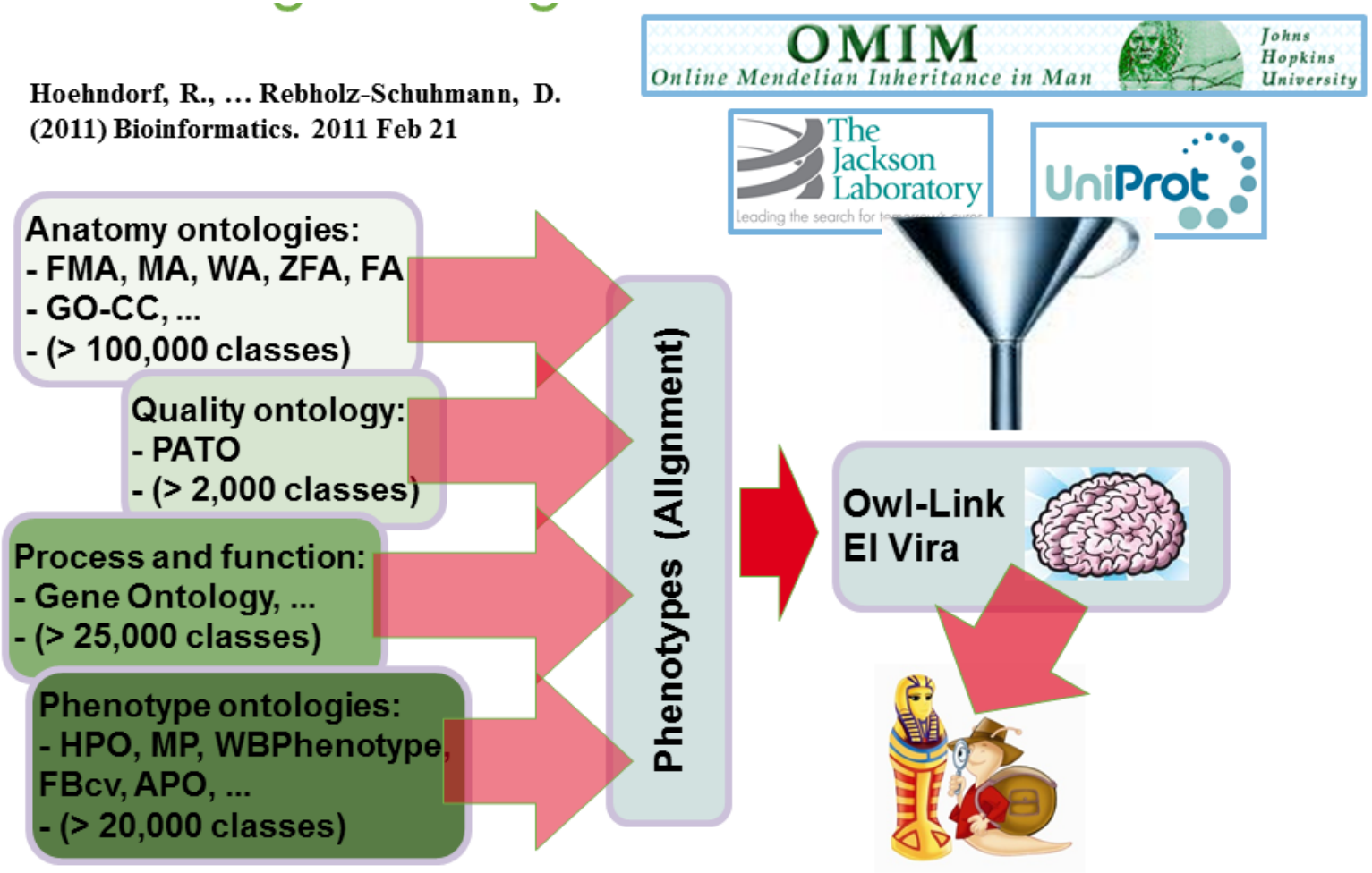
Gene	Mechanism / Function	Disease
SLC6A3	SLC6A3 mediates active reuptake of dopamine from the synapse, implicated in parkinsonism (Omim).	Parkinsonism
SCN5A	SCN5A is linked to long QT syndrome and is assumed to be involved in short QT syndrome too (Omim).	Short QT Syndrome
DAZL	DAZL expression has been shown in seminoma cells.	Germ cell tumor
ACVRL1	ACVRL1 increases the amyloid production in the brain leading to amyloid angiopathy.	Angiopathy

- Identification of 1,154 potential candidate genes
- About 63 % could be verified from the scientific literature.

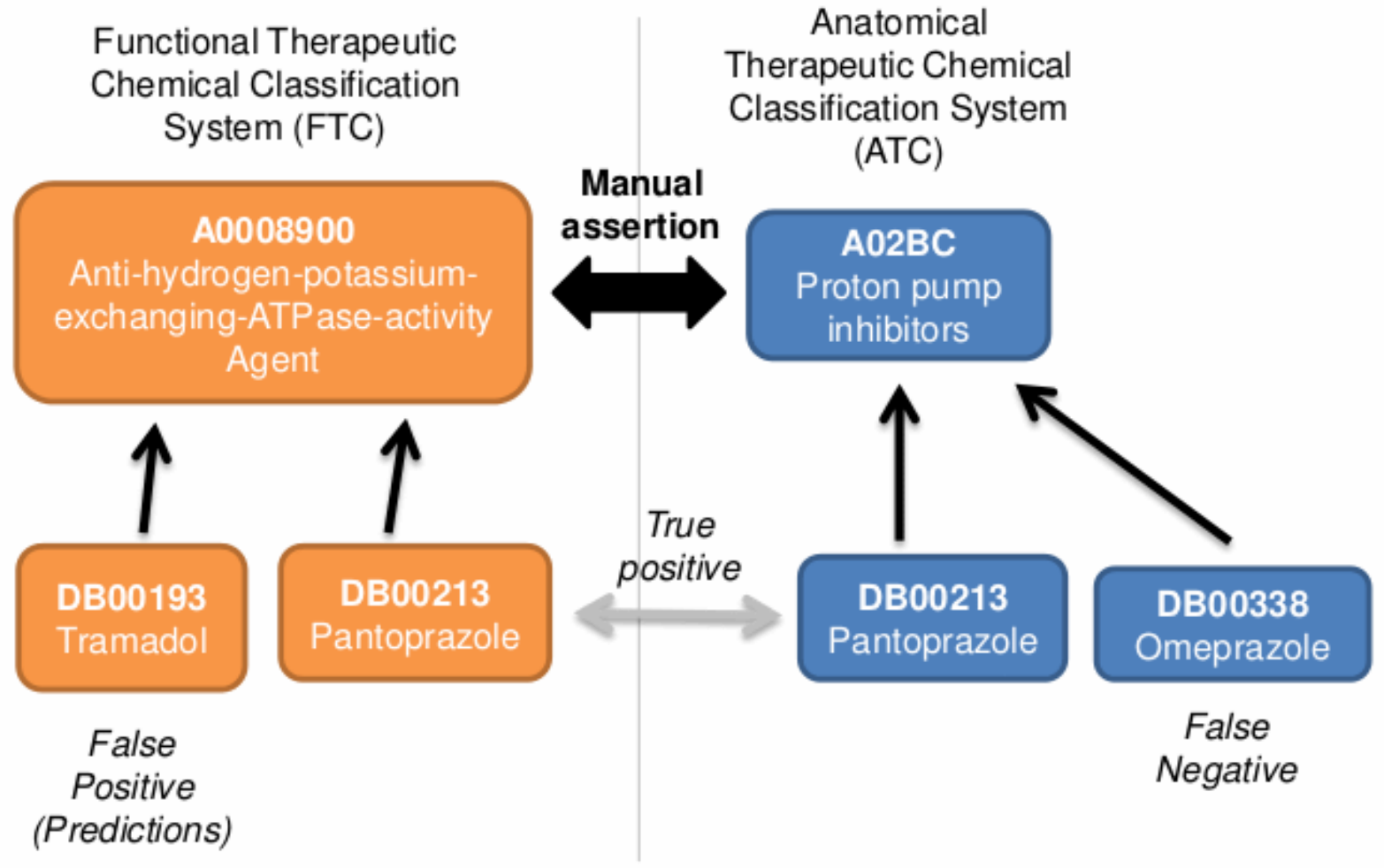
Reasoning over ontologies



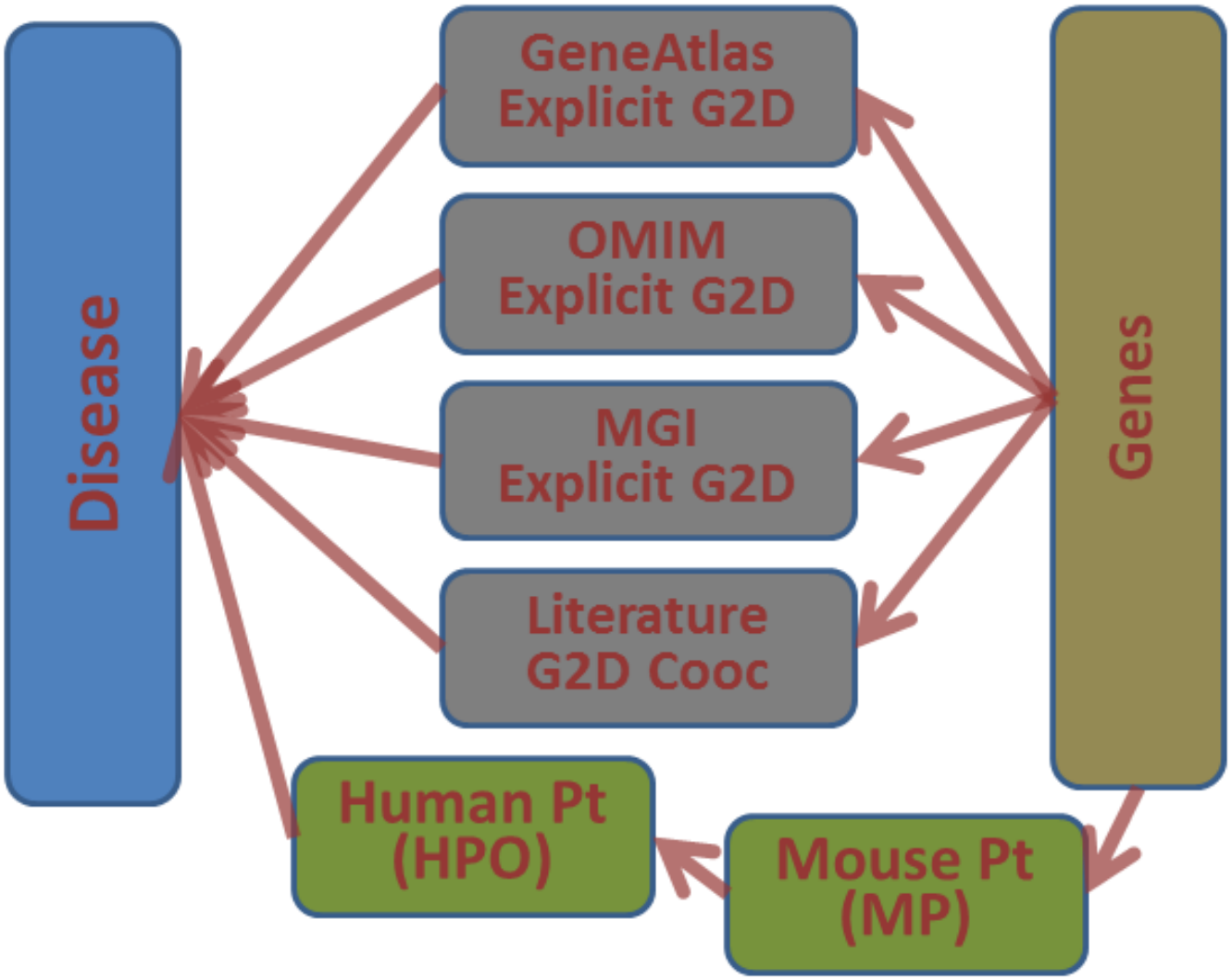
Reasoning over ontologies



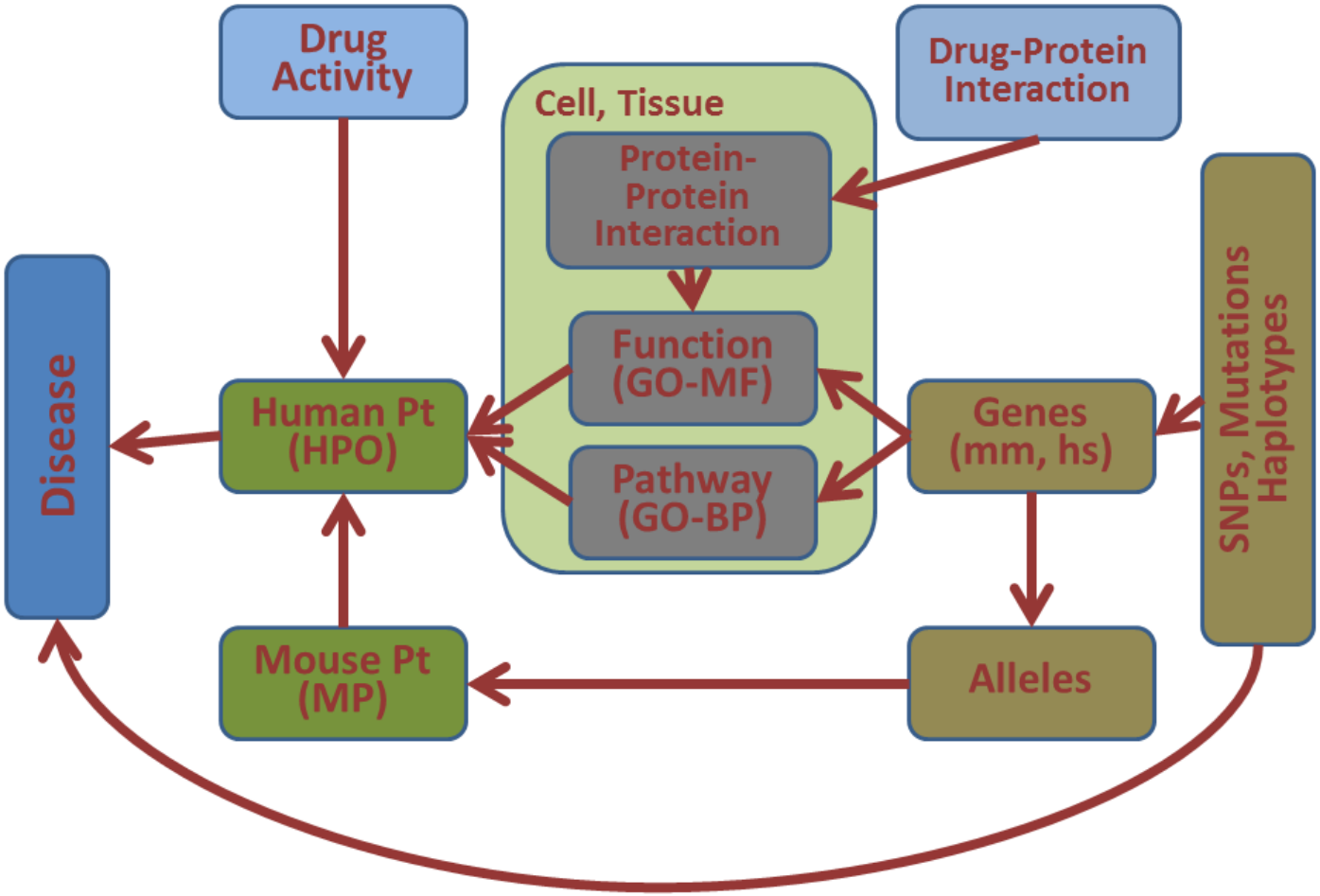
Repurposing of drugs



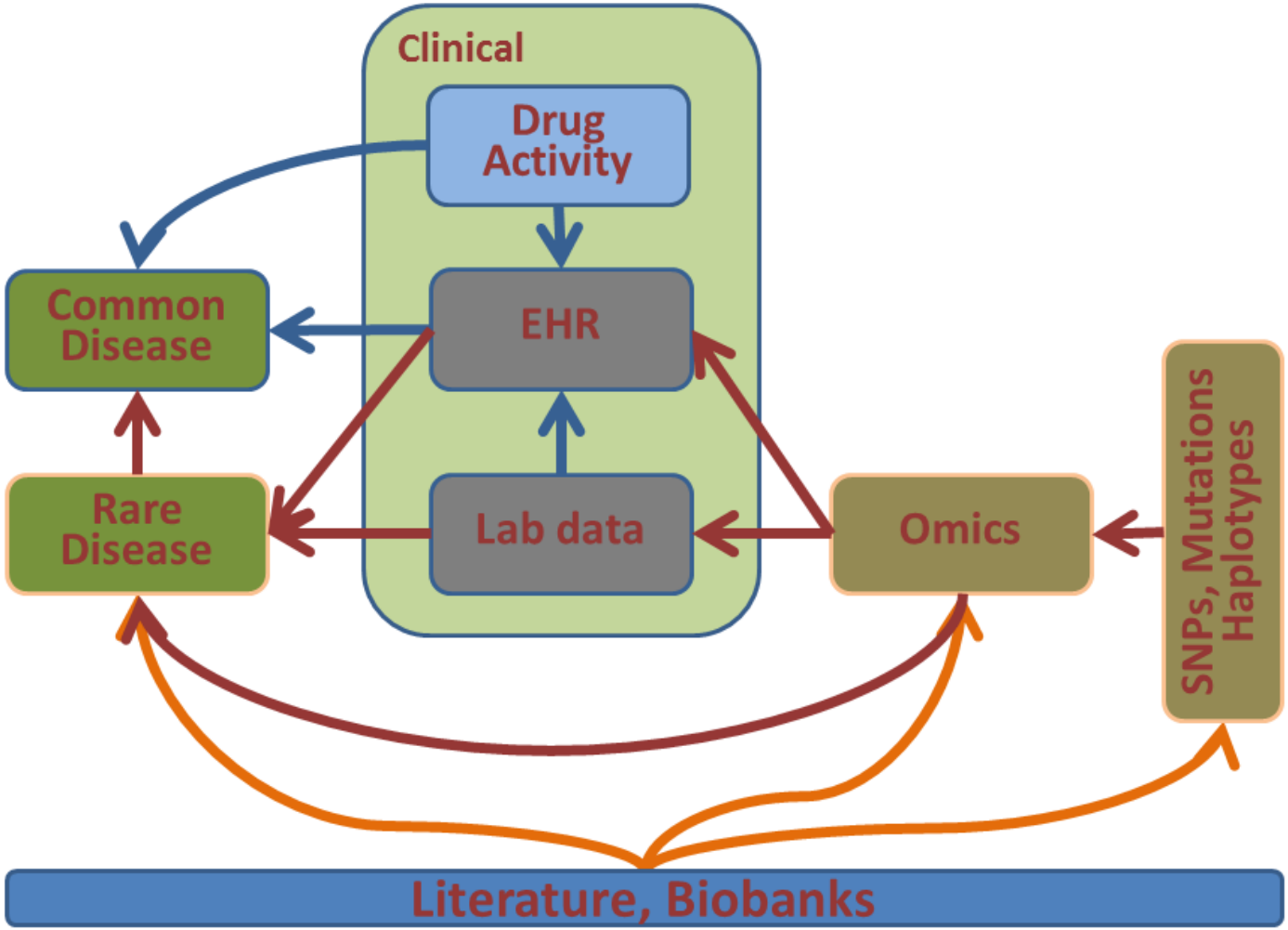
Model for genes, diseases



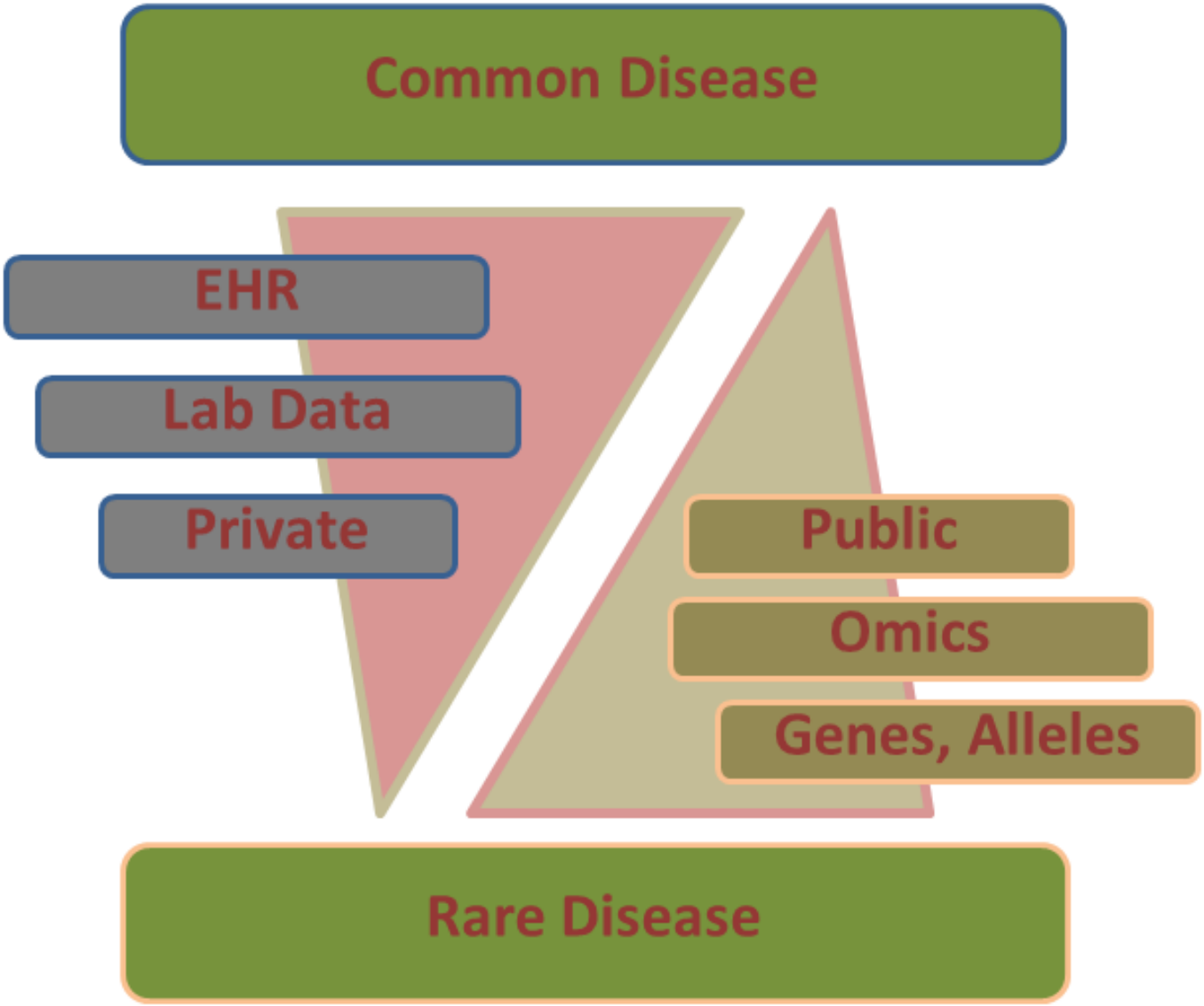
Model for genes, diseases



Model for genes, diseases



Model for genes, diseases



Scenario 1: Diagnostics support

Aim Propose relevant diagnostics decisions

Approach
- Comparison of the patient profile against the profiles from the diseases (see above)
- Identification of relevant parameters, e.g. dysfunctions of organs (liver, kidney)
- Consider genetic parameters

Gain
- Decision support to the medical doctor.
- Intgration of public data sources with patient information: easy access.

Szenario 2: Risks linked to drugs

Aim Identification of adverse side effects in the patient profile.

Approach

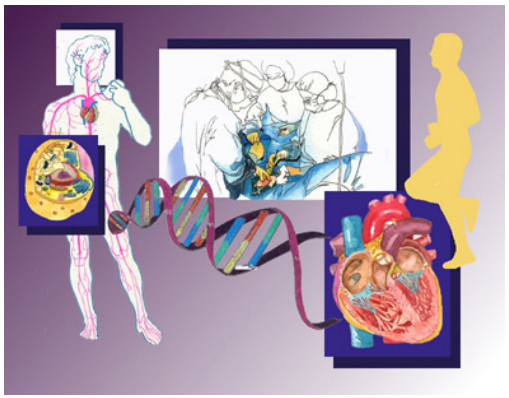
- Determine the drug profile: pos./neg. regulation
  of metabolism X, Y;
  of physiological parameters A, B;
  of phenotypes O, P.
- Comparison of the drug profile against the laboratory results and the patient record.
- Statistical methods for the identification of synergistic effects.
- Comparison between doctor's notes in the EHR against the predictions.

Gain Identification / reduction of unwanted adverse side effects of drugs.

Szenario 3: System medicine and Omim 2.0

Aim Integration of the patient data in system medicine / system biology (translational medicine).

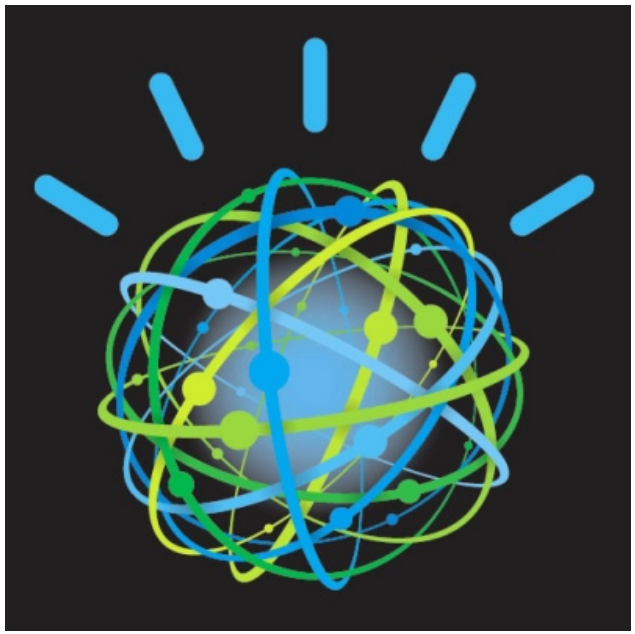
Approach



- Phenotype profile for all relevant diseases.
- Analysis and normalisation of public data resources: Omim, literature, databases in molecular biology, PharmGKB, ChEMBL.
- Identification of regulatory processes: positive regulation in process X increases parameter Y in phenotype Z
- Judgement on synergirstic / antagonistic effects: identification if risks, e.g. over-regulation.

Gain Comparison of the patient data against models from system biology, system medicin, personalized medicine, pharmacogenomics.

First Future: artificial intelligence

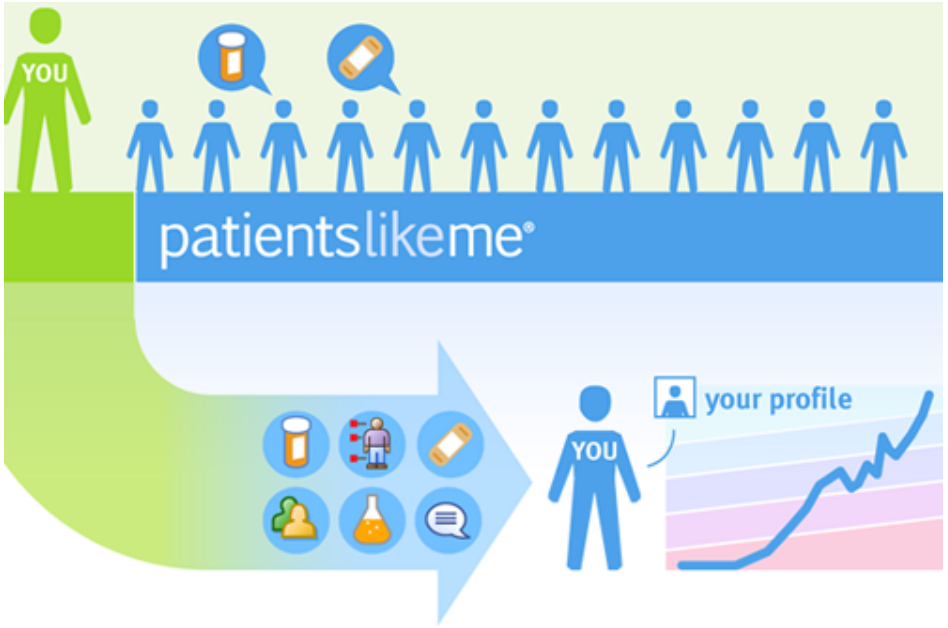


- IBM Watson: A cluster computer analyzes facts.
- Has been the chamption in "Jeopardy" (US).
- Could predict successfully the right diagnosis on existing patients.
- The biomedical knowledge domain provides lot of ontologies and electronic data bases.

Second future: Big data

- 23andme: Genomics analysis for everybody.
- Exom sequencing for about 5,000 USD, a complete SNP set for 99 USD.
- 10,000 people will be sequenced, maybe 100,000 soon.
- Peta-Bytes of data in biomedical data bases.
- Which domain knowledge is not not available yet, nor used yet?

Third future: social networks, Patients-like-me



- PatientsLikeMe: Exchange of patient-related data.
- Exchange of phenotype information.
- Identification of treatement.

Summary

- Semantic resources are still the key element for the normalisation of data resources.
- It is possible (easy) to process large data volumes in real time and on an everyday basis.
- The semantic integration requires textual data in the first place, but can be applied to numerical data as well.
- Profile comparisons, statistical methods and automatic reasoning are all equally important.
- The integration of the patient record with data from molecular biology and genetics data resources becomes increasingly important.

Fragen?