Data challenge in health care and life science

Bo Andersson, AstraZeneca R&D Lund

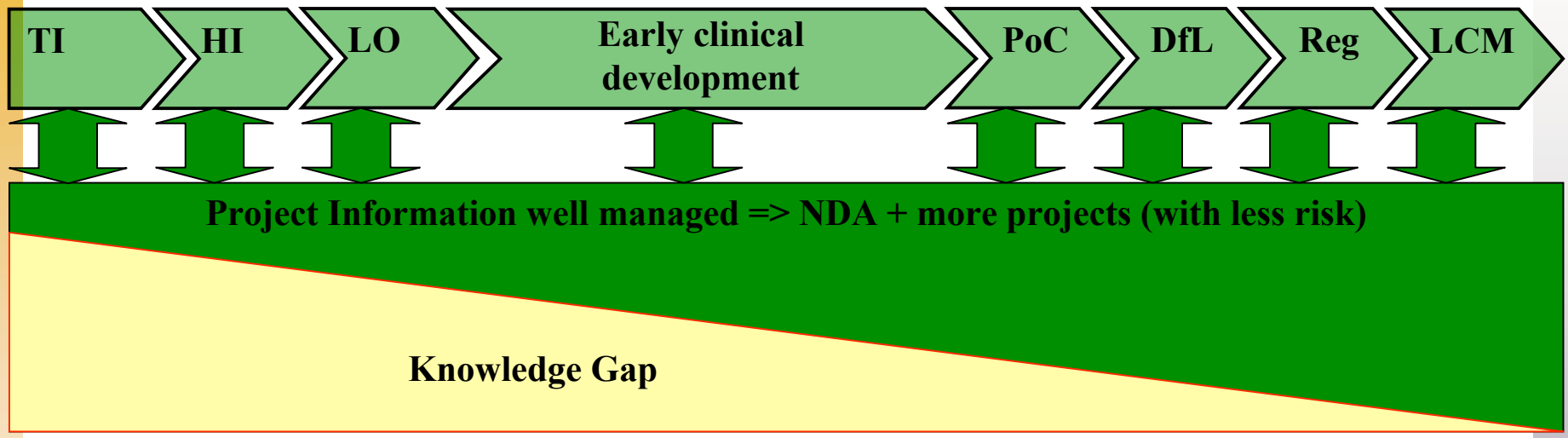
Semantic Web for Health Care and Life Sciences Interest Group

20 October 2008, F2F Meeting, Mandelieu, France

AstraZeneca

Outline

- **Data challenge**,
  - Drug development process
  - Complex requirements for new health care paradigm
  - Research scientists needs
- Activities in AZ with SW components
  - Clinical data repository
  - Clinical study information
  - Large Knowledge Collider (LarKC)
- Summary
  - Some thoughts for the future



AstraZeneca R&D is a
knowledge organization
in which teams create, use, search, combine, **interpret**, and manage information to develop drugs and services.

The BIG 3 concept

Smokers/
Noxious
gases

Screening Tool

Susceptible Individuals

Early diagnosis

Treatment for smokers and ex-smokers

Systemic disorder

Hospitalisation

MORTALITY

COPD

Lung cancer

CV

Mechanisms?

Influence guidelines

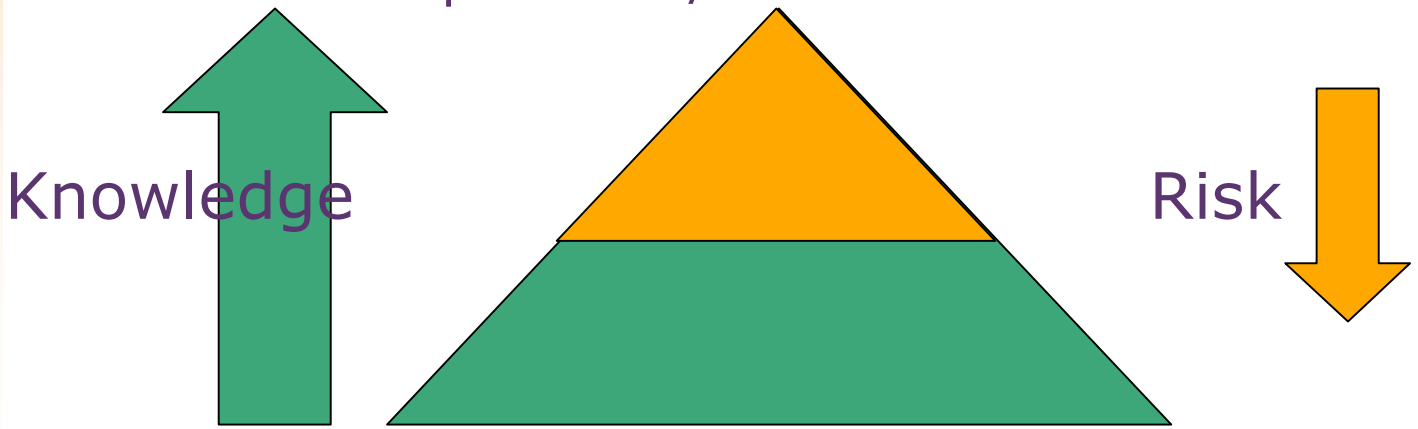
Maria Gerhardsson de Verdier, MD, PhD, AstraZeneca R&D Lund

Improve the capability to integrate and interpret heterogeneous data

- Build information management capability to support drug development:
  - Biological and environmental risk factors for developing a disease and prognosis for patients
  - Hypotheses for casual chains of diseases (early diagnosis)
  - Hypotheses about patient characteristics and other factors that can explain segmentation criteria

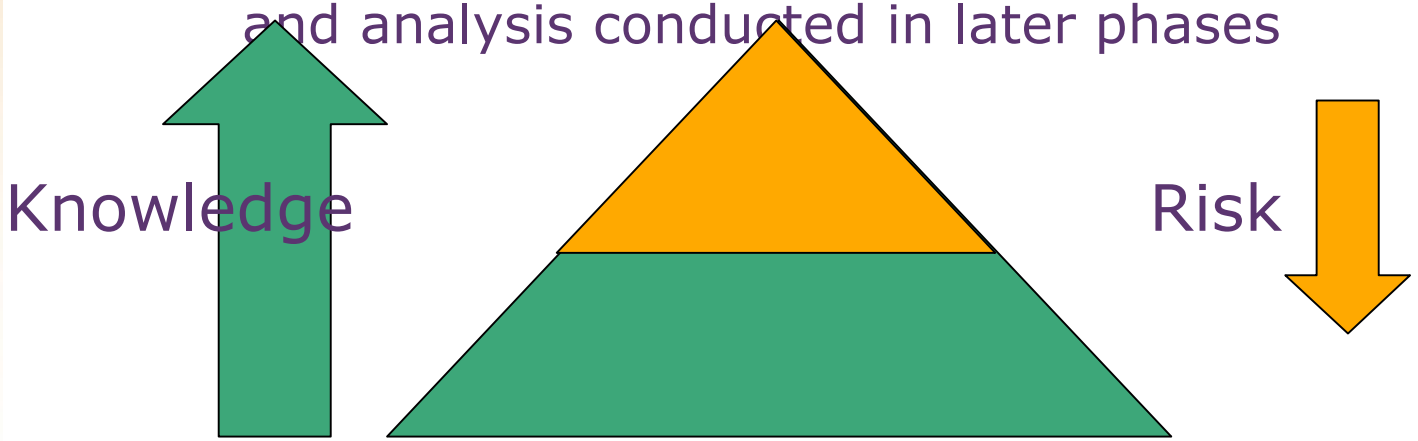
Project knowledge repository

- Build knowledge management capability to support early clinical project team:
  - Disease and patient segmentation
  - Risk factors for drug class and biological target
  - How does others do
  - Patient availability
  - Animal to human models
  - Known problems/failures



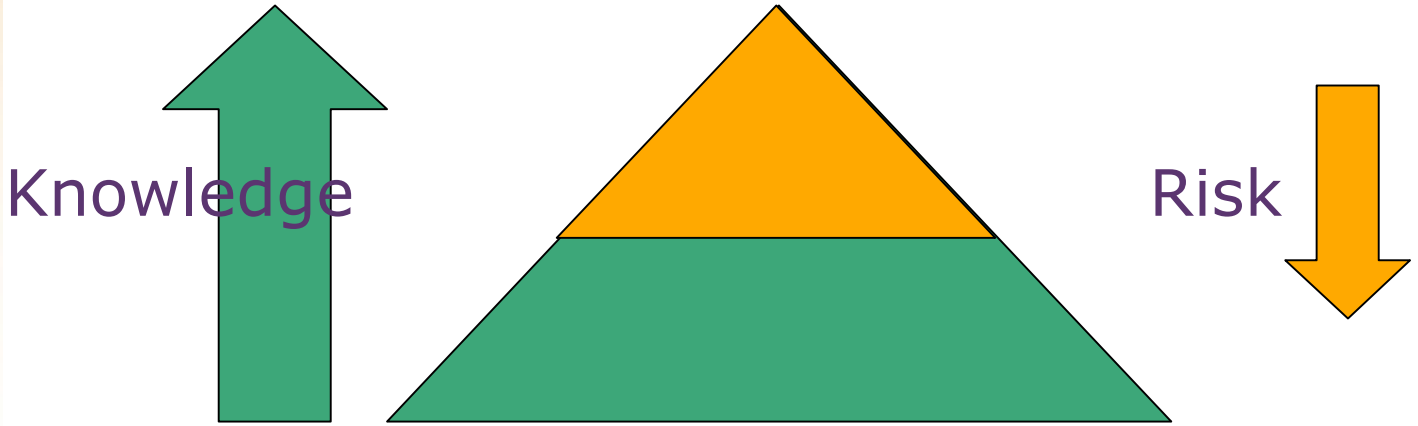
Identifying biomarkers and target mechanisms

- Data interpretation is a non-trivial process that requires overcoming:
  - Syntax differences in the generated format
  - Semantic differences in the format, e.g. used identifiers
  - Verify, validate and compare experimental results with other established data sets
  - Vast heterogeneity of the interpreted information
  - Efficient secondary usage of past experimental results and analysis conducted in later phases



Signal evaluation of adverse drug event reports

- During signal evaluation the safety expert will evaluate if there is a casual relationship between the drug and the adverse event (method RUCAM):
  - Time to onset of the reaction
  - Course of the reaction
  - **Risk factors for drug reaction**
  - **Concomitant drug(s)**
  - **Non-drug related causes of event**
  - **Previous information on the drug**
  - Response to readministration



Outline

- Data challenge,
  - Drug development process
  - Complex request for new health care paradigm
  - Scientists needs
- **Activities in AZ with SW components**
  - Clinical data repository
  - Clinical study information
  - Large Knowledge Collider (LarKC)
- Summary
  - Some ideas for the future

Consolidated clinical data repository

- The CRL and CCDS are designed based on the assumption that diversity in clinical data is part of "doing research".
  - Driver: Business value achieved by effective use of clinical data cross studies and over time
  - So, in CRL we will be able to specify the variances in what we observe on subjects in clinical studies, and the information about these observations.
  - CCDS will connect these specifications to the actual data. And thereby enable us to take informed decisions when we want to utilize data cross variances.
- Enforcement of standards to reduce diversity is a line organization decision.
  - Driver: Operational efficiency by rationalization of processes and tools for new studies.
  - So, CRL will make this task easier by making the preferred (standardized) variant of the specification available as first option when we will set up new studies and acquire new information.

Clinical Observation Concepts

To store the clinical observation within the CRL data model we need to define some terminology

What are we trying to measure?
*Systolic Blood pressure (**carrier of topic**)*

Could it be measured in a different way and would that affect the result? YES
- *Patient position (**qualifier**)*
- *Method/Tool/Equipment (**qualifier**)*
- *Location/Site -where you measure it (**qualifier**)*

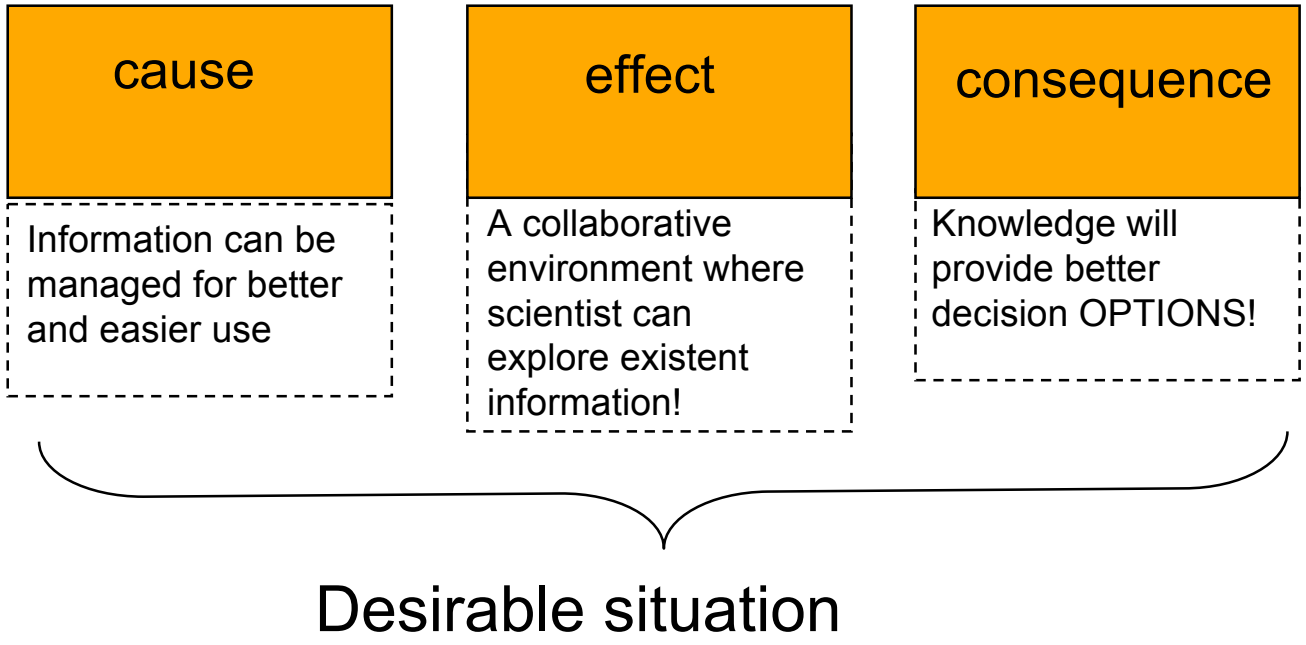
For the clinical trial is there anything I need to know? YES
- *When was it measured, date (**context**)*

Core part of JANUS have been normalized and implemented in CCDS

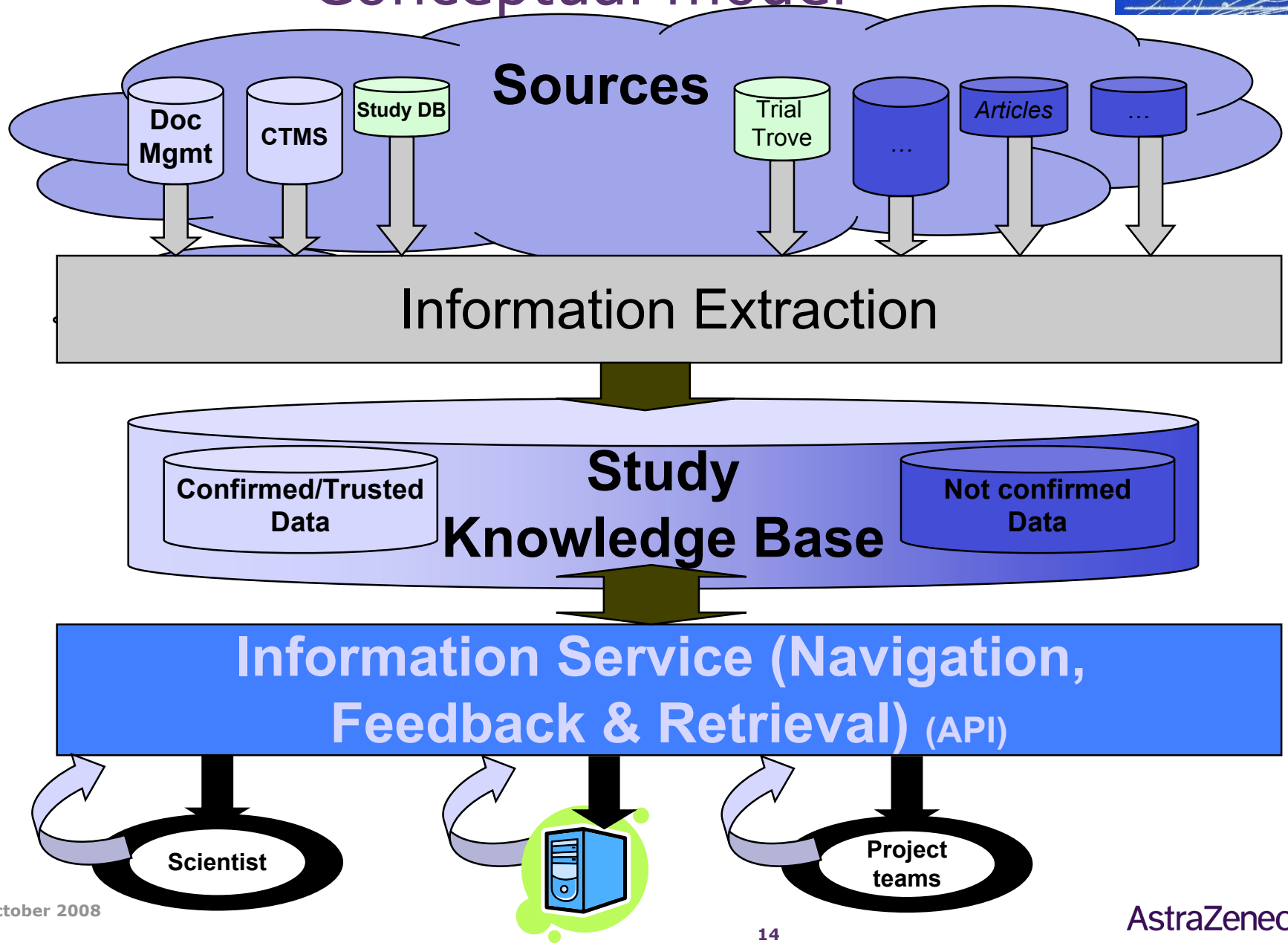
- Protocol – "what was supposed to happen"
- Trial structure (arms, visits)
  - Planned assessments
    - ✓ Like actual findings, but no result
  - Planned interventions
    - ✓ Like actual interventions
- Analysis plans and results
  - Analysis datasets (query rule)
  - Analytic plans
  - Analytic results
- Clinical Observations – "what happened"
  - Findings, Test types, Domains
  - Events
  - Interventions

Clinical study information

Opportunity

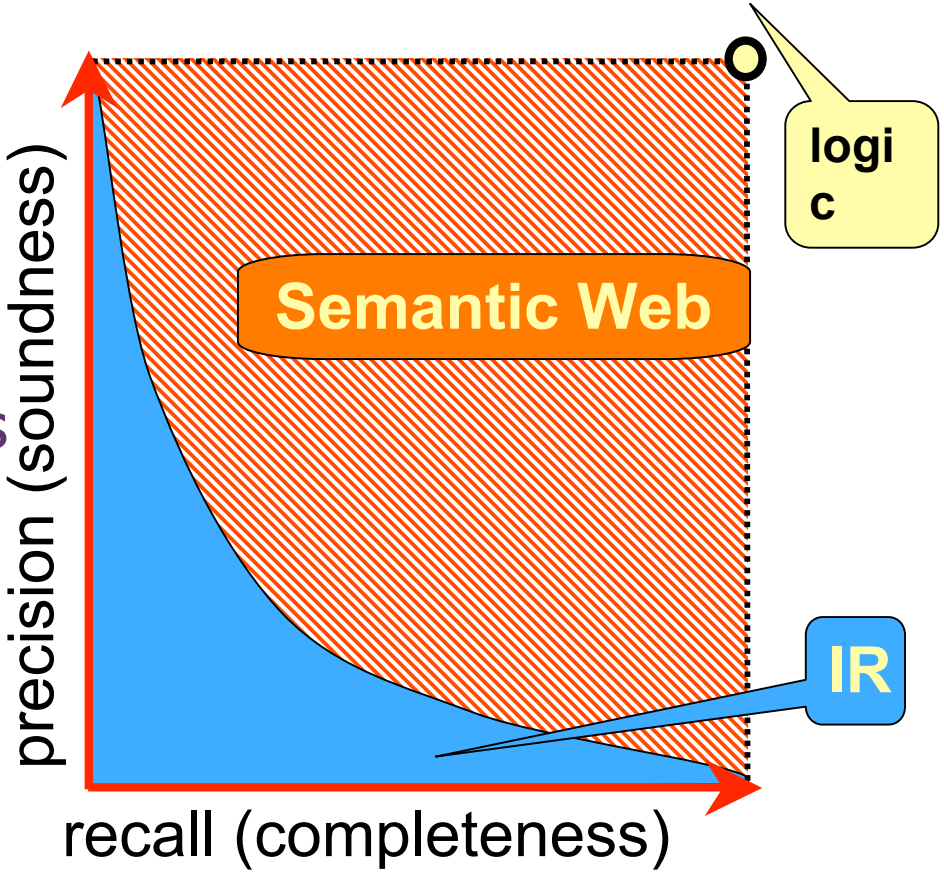


Clinical study information Conceptual model



LarKC in a Nutshell

- "Web Scale and Style Reasoning"
- Giving up 100% correctness:
  - trading quality for size
  - often completeness is not needed
  - sometimes even soundness is not needed



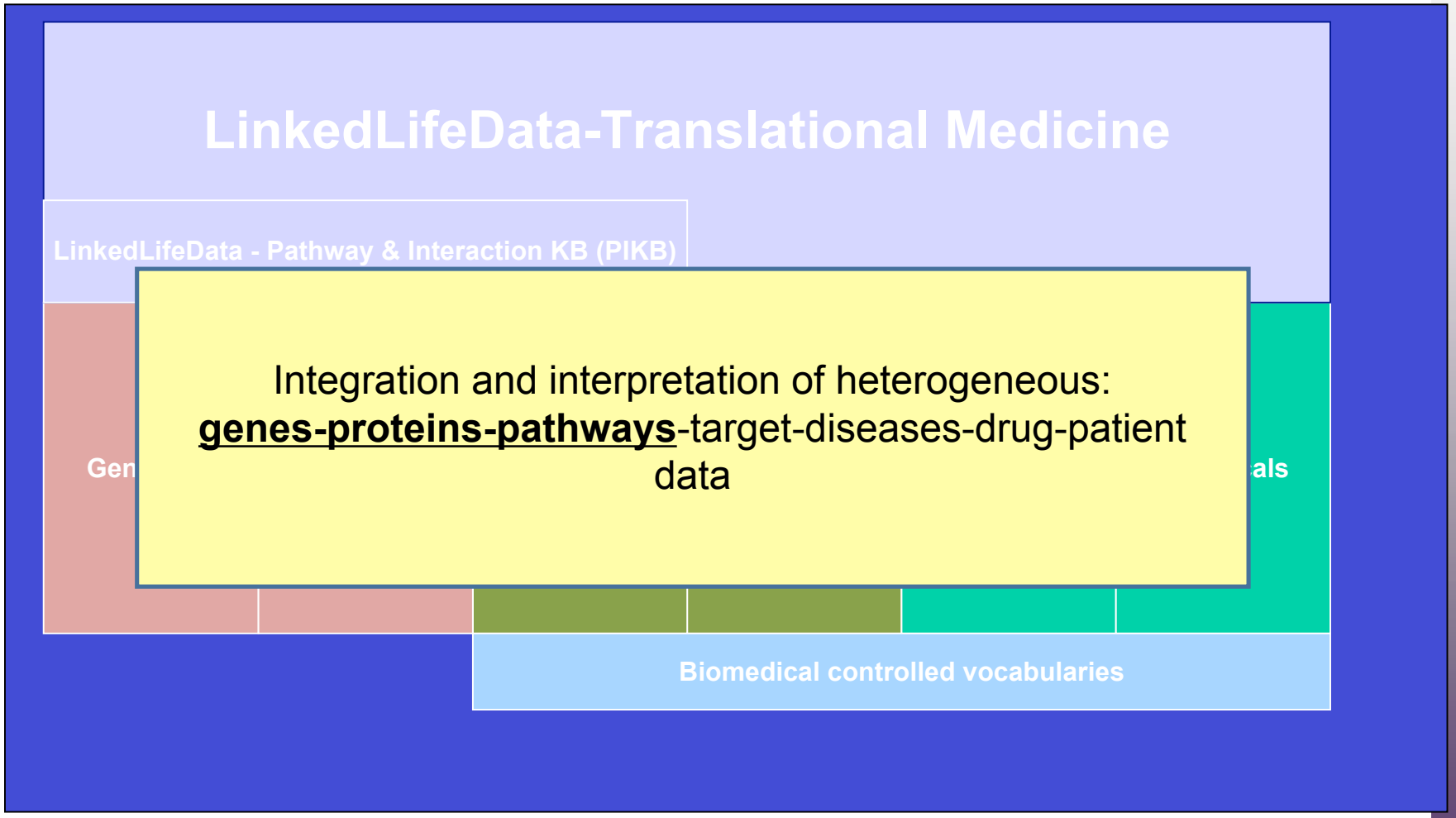
Main Innovations

- Enriching current logic-based Semantic Web reasoning
- Employing cognitively inspired approaches and techniques
- Achieve scalability trough giving up completeness
- Achieve scalability trough parallelization

LinkedLifeData

- Platform developed in context of LarKC
- Automates the process of:
  - Transformation of structured data sources to RDF
  - Load and reason on top of huge amounts of data
  - Provide web interface to access the data
- Currently running on top of BigOWLIM

Knowledge base for Early Clinical Drug Development



Pathway and Interaction Knowledge Base

- Dataset load in LinkedLifeData
- Integrates BioPAX and the related data sources
- First evaluation try!
- Take everything with pitch of salt!

Database	Dataset	Schema	Description
Uniprot	Curated entries	Original by the provider	Protein sequences and annotations
Entrez-Gene	Complete	Custom RDF schema	Genes and annotation
iProClass	Complete	Custom RDF schema	Protein cross-references
Gene Ontology	Complete	Schema by the provider	Gene and gene product annotation thesaurus
BioGRID	Complete	BioPAX 2.0 (custom generated)	Protein interactions extracted from the literature
NCI - Pathway Interaction Database	Complete	BioPAX 2.0 (original by the provider)	Human pathway interaction database
The Cancer Cell Map	Complete	BioPAX 2.0 (original by the provider)	Cancer pathways database
Reactome	Complete	BioPAX 2.0 (original by the provider)	Human pathways and interactions
BioCarta	Complete	BioPAX 2.0 (original by the provider)	Pathway database
KEGG	Complete	BioPAX 1.0 (original by the provider)	Molecular Interaction
BioCyc	Complete	BioPAX 1.0 (original by the provider)	Pathway database
NCBI Taxonomy	Complete	Custom RDF schema	Organisms

LinkedLifeData - PIKB

- Number of statements: **1,159,857,602**
- Number of explicit statements: **403,361,589**
- Number of entities: **128,948,564**
- Publicly available at:

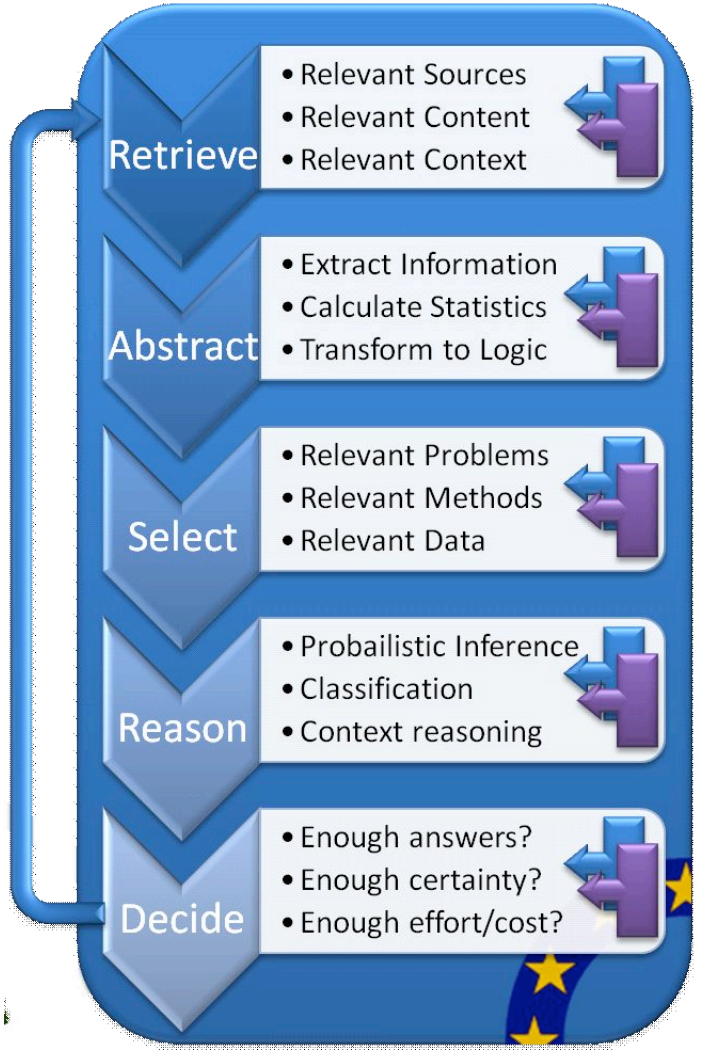
  http://www.linkedlifedata.com

Outline

- Data challenge,
  - Drug development process
  - Complex request for new health care paradigm
  - Scientists needs
- Activities in AZ with SW components
  - Clinical data repository
  - Clinical study information
  - Large Knowledge Collider (LarKC)
- **Summary**
  - Some ideas for the future

Summary

- Information integration and interpretation are huge challenges for scientists
- SW technology have showed potential
- Research scientist must be closely involved
- LarKC include many of the component we expect to need in the future



Some ideas for the future

- We need better solutions to describe information so that other humans and computers can use it, e.g. ontologies, identifiers, standards etc.
- We need personalized smooth tools to search, find, integrate and interpret information.
- We need computational support for ”annotation”, “reading” and writing
- We believe Semantic Web technologies will be an important part of the solution!

Read more about LarKC:

http://www.larkc.eu

http://www.linkedlifedata.com

Contributions from:

Maria Gerhardsson, AstraZeneca R&D Lund
Kerstin Forsberg, AstraZeneca R&D Mölndal
Vassil Momtchev, OntoText Bulgaria