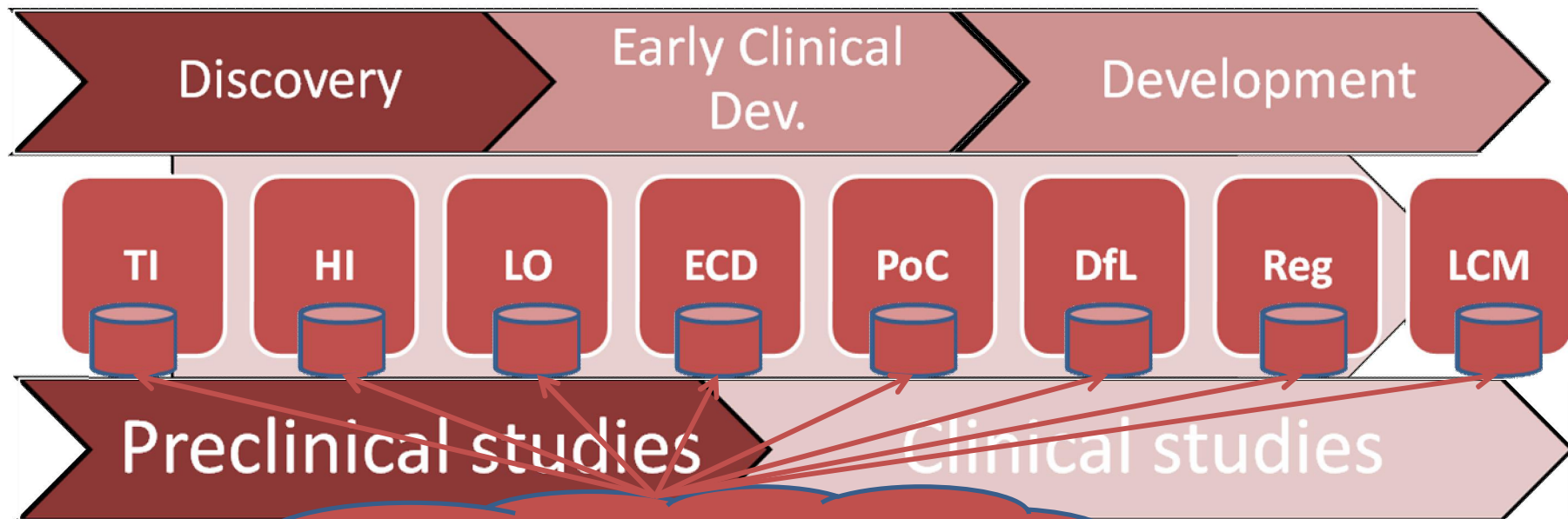


A very pragmatic view towards: Life Sciences and Health Care Vertical

SemData@Sofia

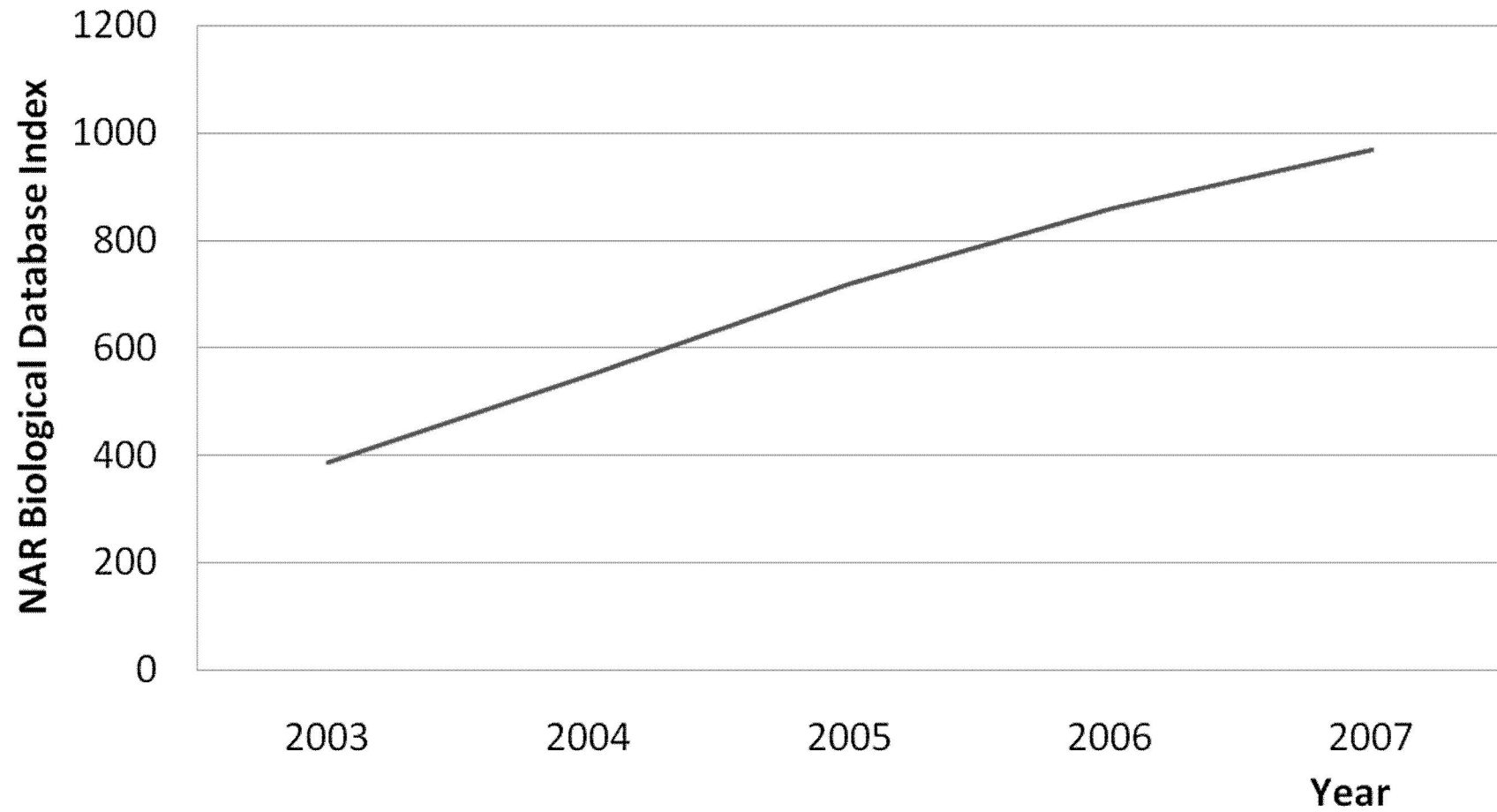
Vassil Momtchev

Knowledge Driven Process



- Target Identification
- Hit Identification
- Lead Optimisation
- Development for Launch
- Registration and Launch
- Life Cycle Management

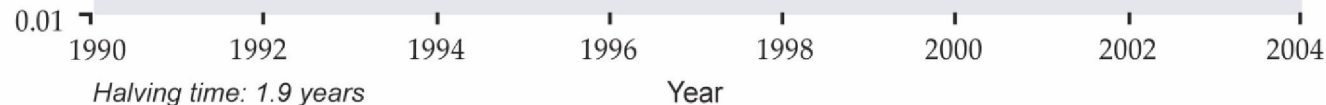
Public Data Sources



Big Data Silos

```
Terminal — ssh — 84x22
coil-blue ~ # df -H
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda3       56G   44G   8.6G  84% /
udev            3.7G  173k  3.7G   1% /dev
/dev/sda1       104M   15M   84M  15% /boot
/dev/sdb1       886G   21G  821G   3% /opt
/irstall        56G   44G   8.6G  84% /var/ftp/install
/tftpboot       56G   44G   8.6G  84% /var/ftp/tftpboot
/dev/ASDC_archive 1.1P  1.4T  1.1P   1% /ASDC_archive
/dev/SPG_ops    147T   52T   96T  36% /SPG_ops
/dev/homedir    6.0T   4.6G  6.0T   1% /homedir
/dev/scf0       90T   16T   75T  18% /SCF
coil-blue ~ #
```

Chris Dagdigan – Computing & Strage Trends, BioIT World '09



Life Science Models

- Biology is complex
 - > Complexity is bigger issue than the scale
- Applying expressive semantics is an even bigger issue
 - > Researchers have to understand the models
- Many exceptions
 - > Tomato: is it a fruit or vegetable?

Semantic Life Science Models

OWL	OBO
Generic or top-down	Specific or bottom-up
Describe any domain (in theory)	Focus on supporting existing users and applications
Background in AI	Background in genome annotations
Ontology (strict semantics)	Vocabularies (relaxed semantics)

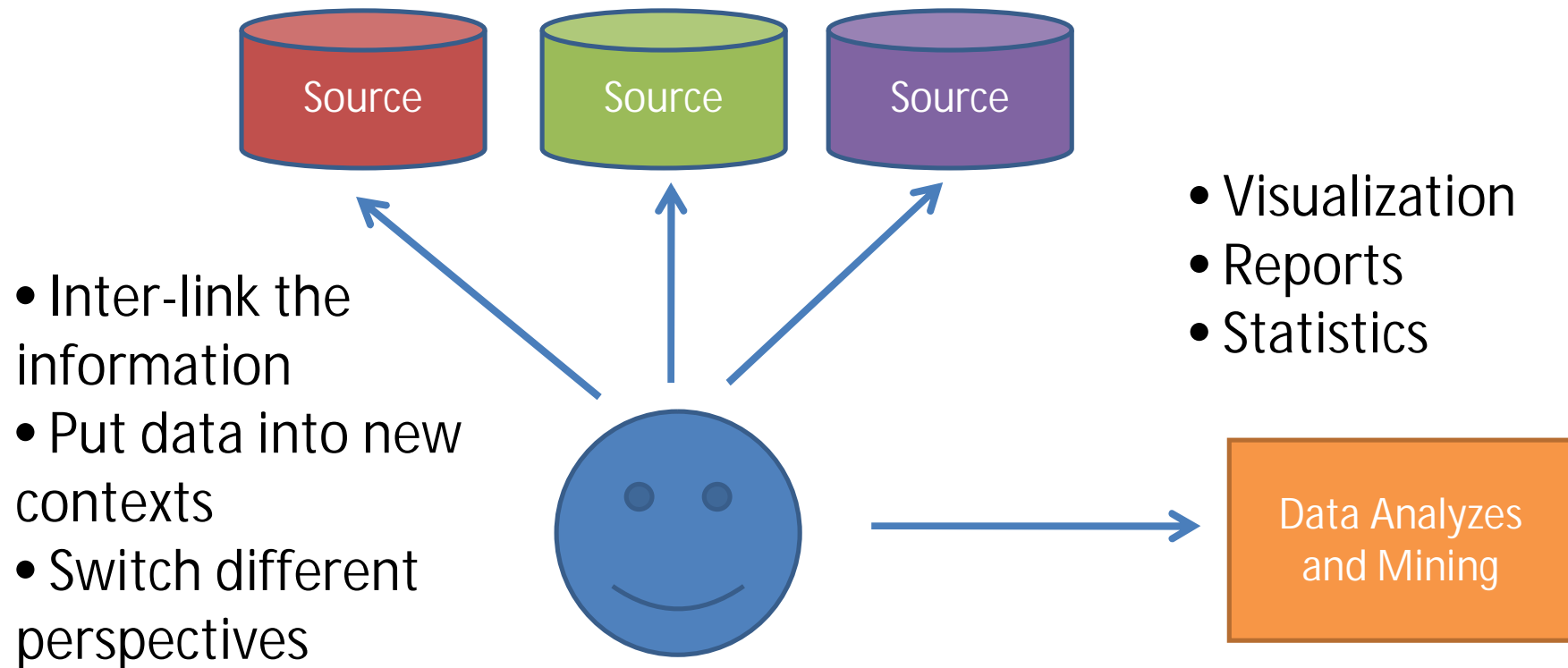
Semantic Expressivity Trade-offs

- Analyze the **Pain/Gain** reasoning curve
- Gold rule **80:20** (expressivity vs. consistency)
- **Linked data** is well accepted in the community
- **Simple and efficient** reasoning schema (SKOS)
- Custom light-weight **inference rules**

Nucleus part_of Cell Always

Cell has_part Nucleus Not Always

A Typical Use Case Scenario (highly-oversimplified)



Current Hot Problems

- Linked Life Data is a public RDF warehouse
 - More than 20 data sources and 5 billion statements
 - Post processing and instance alignment
 - Text mining of linked data
1. Data quality
 2. Licensing problems
 3. Tools for data exploring

http://en.wikipedia.org/wiki/AstraZeneca

Products

[\[edit\]](#)

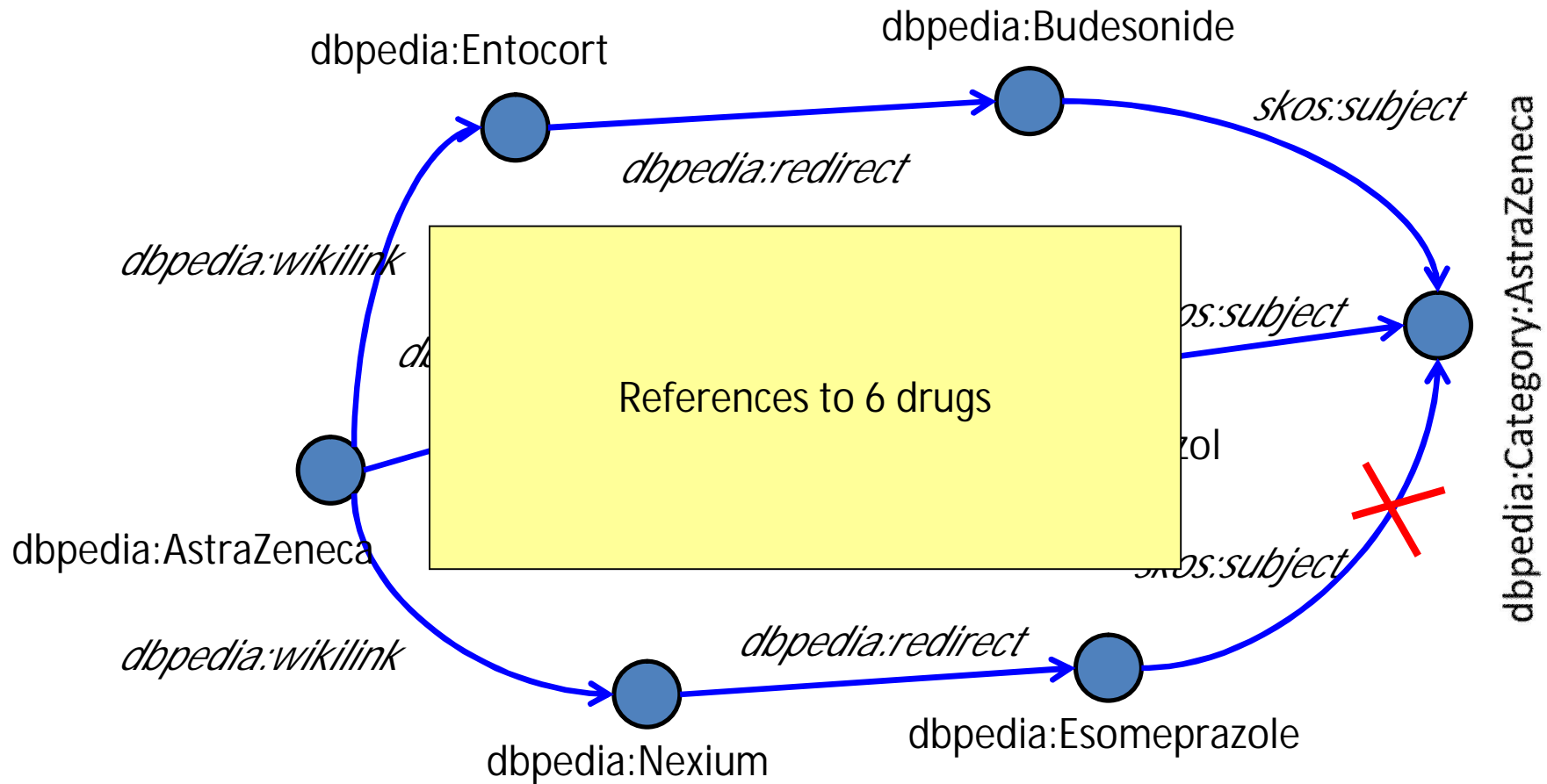
AstraZeneca specialises in prescription medicines to fight disease in several therapeutic areas. Year-on sales information can be found through AstraZeneca Annual Reports.^[16] The following is a list of key products found on the AstraZeneca website.^[17] Generic drug names are given in brackets following the brand name.

This list is incomplete; you can help by expanding it.

- **Gastrointestinal**
 - Entocort (budesonide)
 - Losec and Prilosec (esomeprazole)
 - Nexium (esomeprazole)
- **Cardiovascular**
 - Atacand (candesartan)
 - Crestor (rosuvastatin)
 - Exanta (ximelagatran, 2004 launch, not approved in the United States)
 - Imdur (isosorbide mononitrate)
 - Inderal (propranolol)
 - **Lexxel** (enalapril/felodipine ER; available only in the United States)
 - **Logimax** (felodipine/metoprolol ER)

References to 52 drugs
(the list is claimed to be incomplete)

http://dbpedia.org/resource/AstraZeneca



Data Quality

datasource:organization/AstraZeneca

datasource:organization/AstraZeneca_LP

datasource:organization/AstraZeneca_Pharmaceuticals%2C_LP

datasource:organization/AstraZeneca_Pharmaceuticals_LP

datasource:organization/AstraZeneca_Pharmaceuticals_LP

datasource:organization/AstraZeneca_Pharmaceuticals_LP

Data Licensing

- Data source licenses are loosely specified
- No simple way to solve the problem
 - Simpler licensing (highly unlikely)
 - “License agents” (more likely)
- Leave the problem to the end-user

Disclaimer: Part of the information in the Linked Life Data knowledge base is from copyrighted data sources.

Linked Life Data is a prototype demonstration service and its users are solely responsible for compliance with any copyright restrictions. [Report a copyright violation](#).

Linked Life Data is partly funded by the EU IST project [LarkC \(FP7-215535\)](#).

© 2009-2010 Ontotext AD. All rights reserved.

RDF Visualization and Navigation

- Generic RDF data is hard to explore and navigate
 - Coupling with a schema is easier
 - Nodes may have 000's of statements
- No good RDF model visualization or summarization tools
- Classical table presentation has limited applicability for graph models

Future Industry Trends

- Genomics Revolution
- Personalized Medicine
- New Pharmaceutical Business Models
- Electronic Health Records