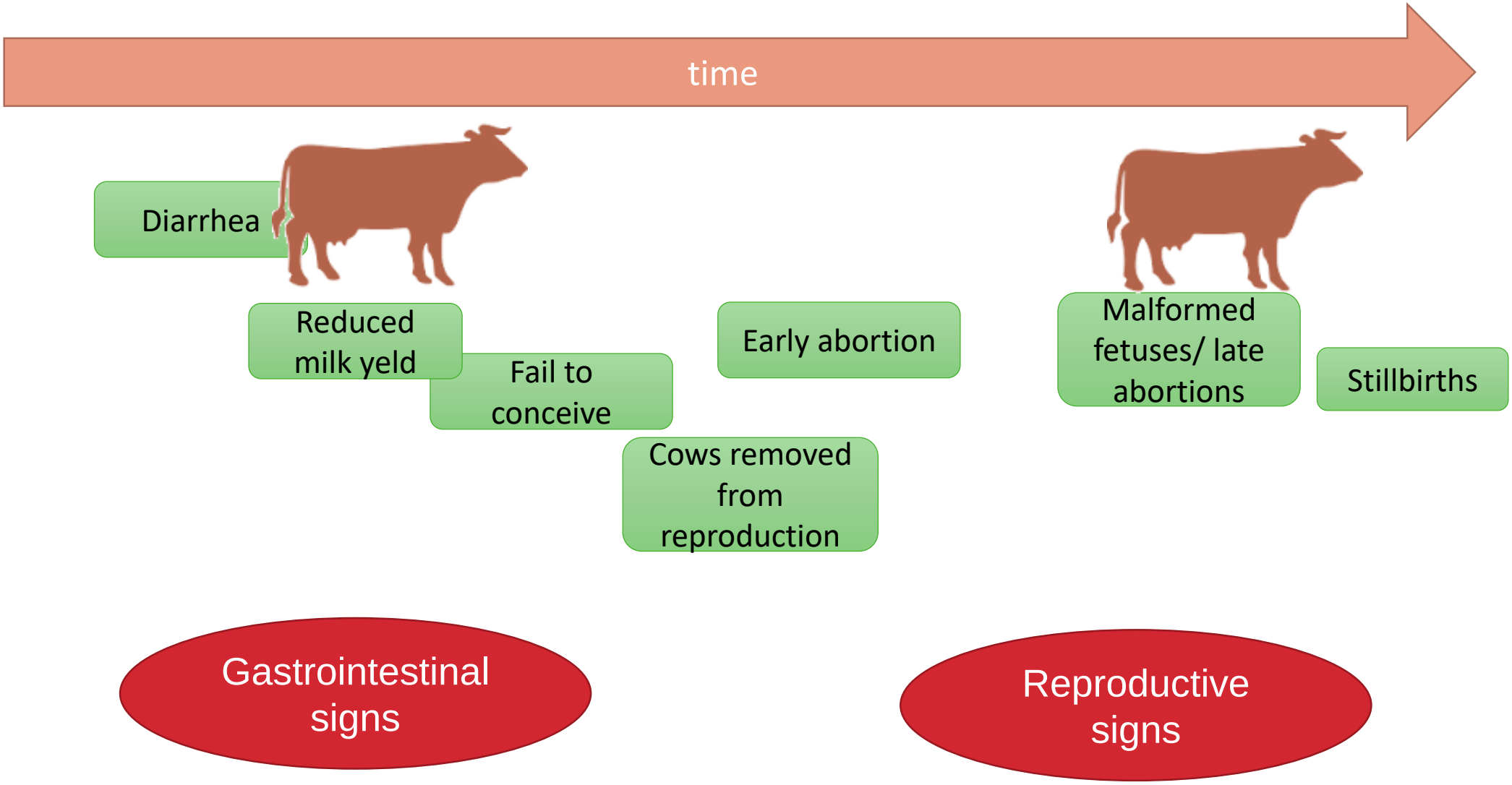
SEMANTIC MODELLING AND ONTOLOGIES IN EPIDEMIOLOGY AND SURVEILLANCE

Crawford Revie

Karl Hammar

Fernanda Dórea


How many diseases cases have we seen in which Schmallenberg could be a differential diagnosis?

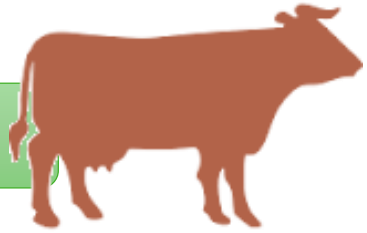


Consider the task of:

Extracting information about these cases from existing data sources

(These data sources were not collected for this purpose)

time



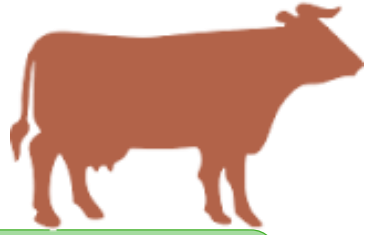
Diarrhea

Reduced milk yeld

Fail to conceive

Early abortion

Cows removed from reproduction



Malformed fetuses/ late abortions

Stillbirths

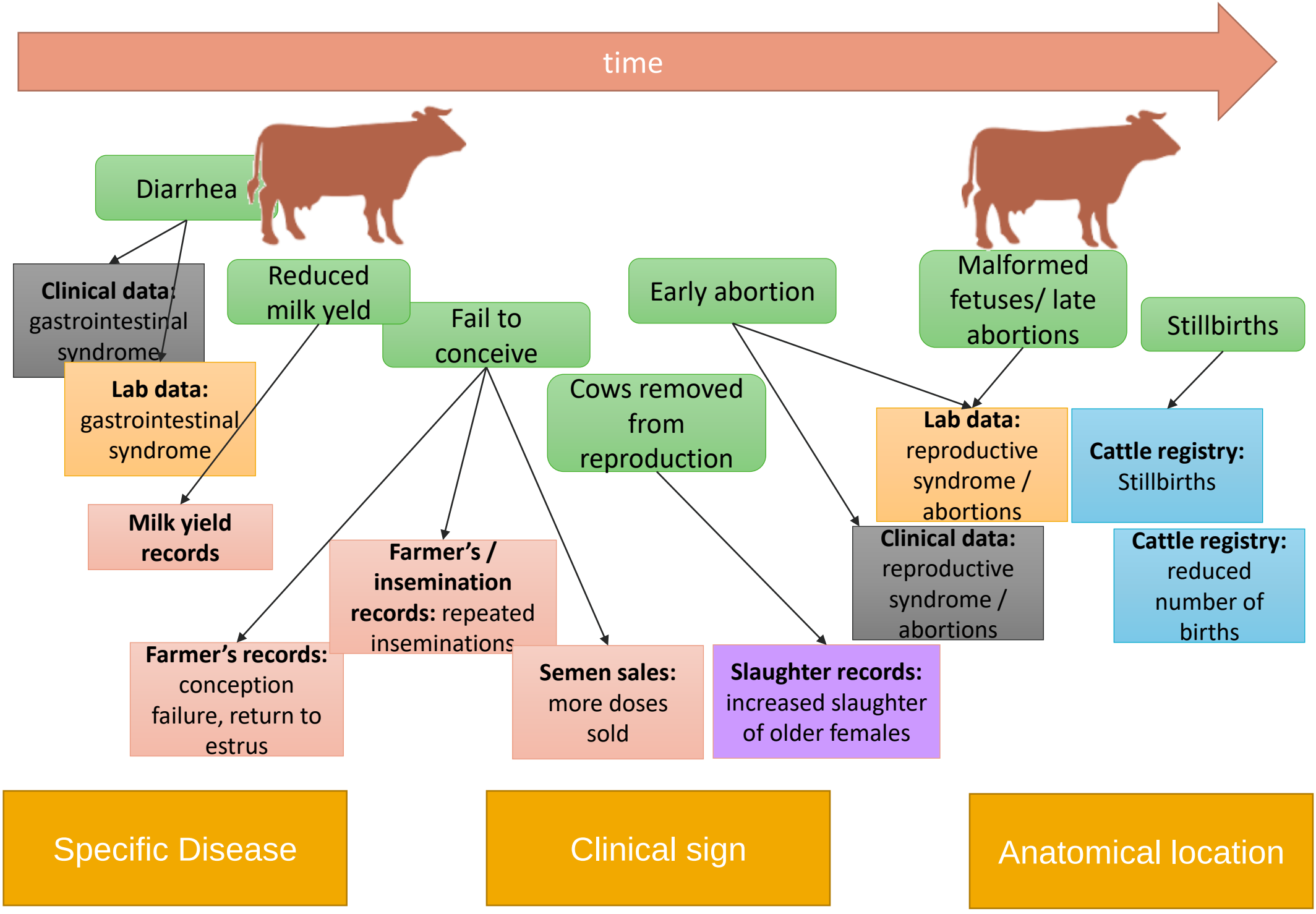
Gastrointestinal signs

Reproductive signs

Specific Disease

Clinical sign

Anatomical location


time
Diarrhea
Reduced milk yeld
Fail to conceive
Early abortion
Malformed fetuses/ late abortions
Stillbirths
Clinical data: gastrointestinal syndrome
Lab data: gastrointestinal syndrome
Milk yield records
Farmer's / insemination records: repeated inseminations
Farmer's records: conception failure, return to estrus
Semen sales: more doses sold
Cows removed from reproduction
Slaughter records: increased slaughter of older females
Lab data: reproductive syndrome / abortions
Clinical data: reproductive syndrome / abortions
Cattle registry: Stillbirths
Cattle registry: reduced number of births
Specific Disease
Clinical sign
Anatomical location

CHALLENGES

- One dataset
  - Secondary use
  - Historical compatibility
  - Prospective compatibility
    - Changes in the data
    - Changes in the question
- Interoperability
  - Datasets from the same type of activity
  - Datasets from a different source
- Reuse of work previously done (by you or others)

PURPOSE

- To illustrate the benefits of ***semantic*** technologies and ***ontology***-based tools that are available right now.
- To let you play around with some such tools yourselves.
- To describe some practical usage and deployment difficulties inherent to these technologies.
- To explore how these can be applied to syndromic, and particularly, animal health surveillance (**AHSO**)

MESSAGES TO TAKE AWAY

1. Data standardisation leads to static and largely ineffective knowledge sharing/use (and in most cases is impractical)
2. We live in an open world, and 'closed world' assumptions will inevitably lead to limitations in knowledge management
3. Snippets of knowledge/information can be useful and may (should) be captured as such
4. Medical data/knowledge is not uniquely complex

CLEVER THINGS 'GOOGLE' CAN DO?

- Ever better search results
- 'Conversational' search and "smart reply"

Transformation to a 'knowledge' company…

- 'Smarter' Gmail/message processing – e.g. direct to Calendar
  (Meetings, Flight Reservations, Hotel Bookings, Concert Tickets, etc.)
- Location-aware services

CLEVER THINGS 'GOOGLE' CAN DO #1

- **Ever better search results**

"Things not strings"

HOW MIGHT WE THINK ABOUT THINGS?

Object	Author	Title	Year published	Genre
Book	Jostein Gaarder	The Solitaire Mystery	1990	Fantasy, Philosophical fiction
Book	Marion Zimmer Bradley	The Mists of Avalon	1993	Fantasy
Book	V.C. Andrews	Flowers in the Attic	1979	Gothic fiction, Family saga

HOW MIGHT WE THINK ABOUT THINGS?

Object	Author	Title	Year published	Genre	Director
Book	Jostein Gaarder	The Solitaire Mystery	1990	Fantasy, Philosophical fiction	
Book	Marion Zimmer Bradley	The Mists of Avalon	1993	Fantasy	
Book	V.C. Andrews	Flowers in the Attic	1979	Gothic fiction, Family saga	
Film		Flowers in the Attic	2014	Drama	Deborah Chow

USING A SIMPLE DATA STANDARDISATION APPROACH TO THINK ABOUT THINGS

Object	Author	Title	Year published	Genre	Director
Book	Jostein Gaarder	The Solitaire Mystery	1990	Fantasy, Philosophical fiction	
Book	Marion Zimmer Bradley	The Mists of Avalon	1993	Fantasy	
Book	V.C. Andrews	Flowers in the Attic	1979	Gothic fiction, Family saga	
Film		Flowers in the Attic	2014	Drama	Deborah Chow
Person		Deborah Chow			

HOW DO 'SMART' APPLICATIONS THINK ABOUT THINGS?

- **Use Google to search for a book or film**

  e.g. "The Solitaire Mystery"

  or "Flowers in the Attic"

- **Use Google to search for a person**

  Vladimir Putin

  Stephen Hawking

HOW DO ‘SMART’ APPLICATIONS THINK ABOUT THINGS?

“Now! *That* should clear up a few things around here!”

STEPHEN HAWKING

Stephen Hawking

Theoretical physicist

Stephen William Hawking CH CBE FRS FRSA was an English theoretical physicist, cosmologist, author, and Director of Research at the Centre for Theoretical Cosmology within the University of Cambridge. Wikipedia

Born: January 8, 1942, Oxford, United Kingdom

Died: March 14, 2018, Cambridge, United Kingdom

Movies and TV shows: Into the Universe with Stephen Hawking, MORE

Spouse: Elaine Mason (m. 1995–2006), Jane Hawking (m. 1965–1995)

Quotes

View 7+ more

Intelligence is the ability to adapt to change.

The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge.

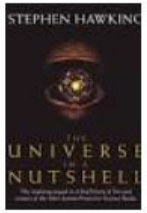
I have noticed even people who claim everything is predestined, and that we can do nothing to change it, look before they cross the road.

Books

View 25+ more

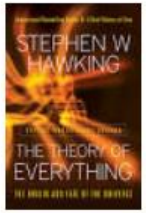
A Brief History of Time
1988

The Grand Design
2010



The Universe in a Nutshell
2001

George's Secret Key to the Uni...
2007



The Theory of Everything
2002

People also search for

View 15+ more

Jane Hawking
Former spouse

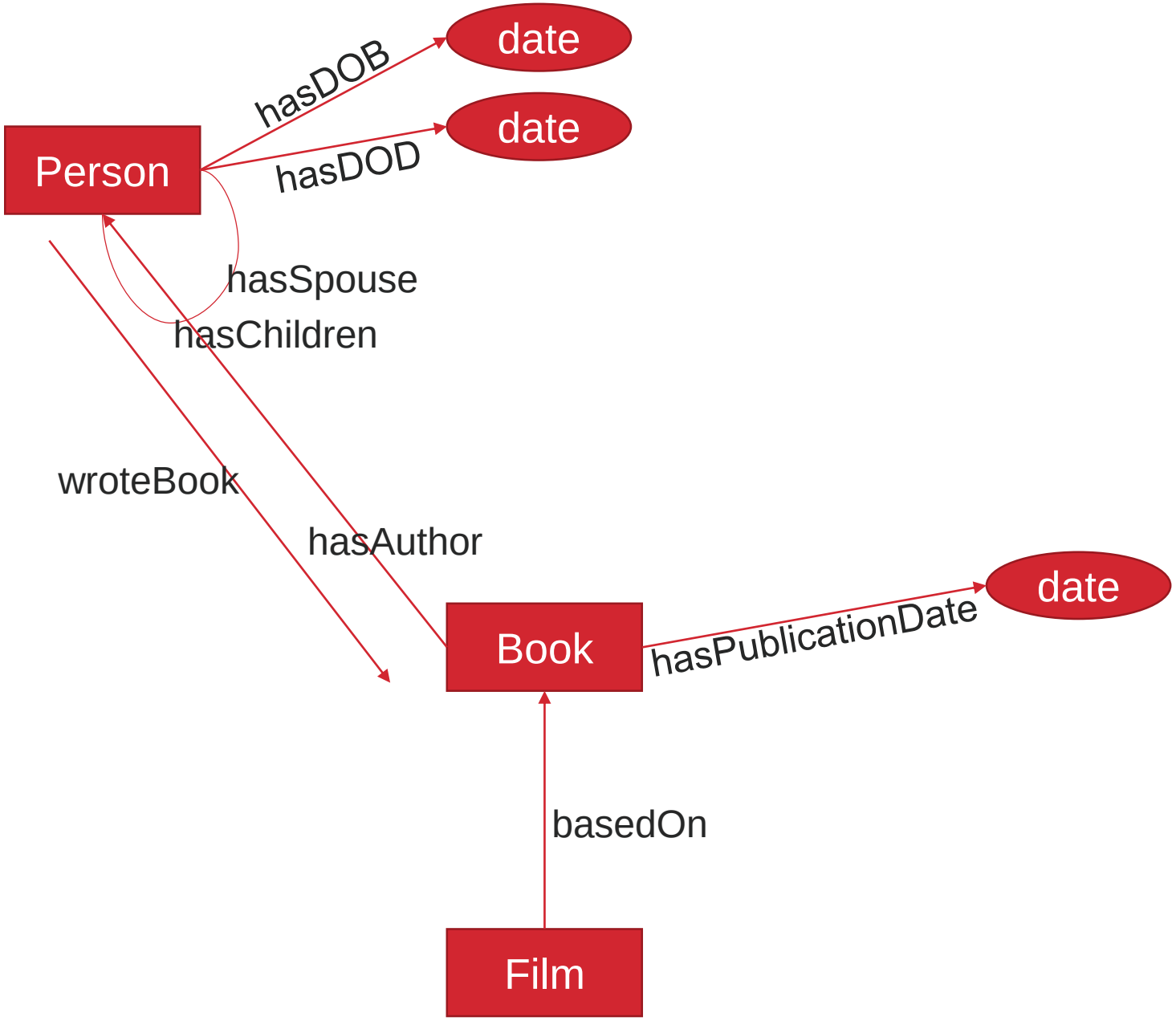
Lucy Hawking
Daughter

Elaine Mason
Former spouse

Albert Einstein

Eddie Redmayne

Feedback


date
hasDOB
date
Person
hasDOD
hasSpouse
hasChildren
wroteBook
hasAuthor
Book
hasPublicationDate
date
basedOn
Film

THINGS - PROPERTIES YOU CAN ASK ABOUT

how tall is the london shard

All Images News Videos Shopping More Settings Tools

About 745 000 results (0,88 seconds)

The Shard / Height

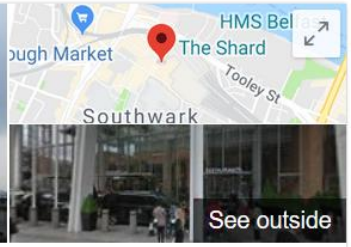
306 m, 310 m to tip

Sources include: CTBUH Feedback

People also ask

- Is the shard the tallest building in London?
- Is the shard the tallest building in the UK?
- How high is the 32nd floor of the Shard?
- How tall is a 10 story building?

Feedback



The Shard

Website Directions

4,5 ★★★★½ 8 039 Google reviews

Skyscraper in London, England

The Shard, also referred to as the Shard of Glass, Shard London Bridge and formerly London Bridge Tower, is a 95-storey skyscraper, designed by the Italian architect Renzo Piano, in Southwark, London, ... Wikipedia

Address: 32 London Bridge St, London SE1 9SG, UK

Height: 306 m, 310 m to tip CTBUH

Floors: 95

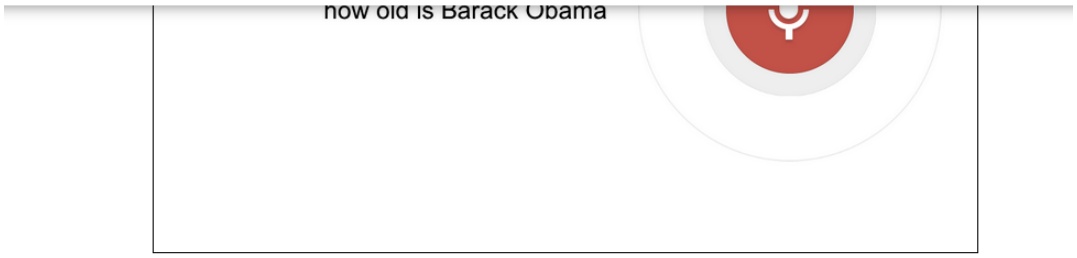
Opened: February 1, 2013

Hours: Open now · Add full hours

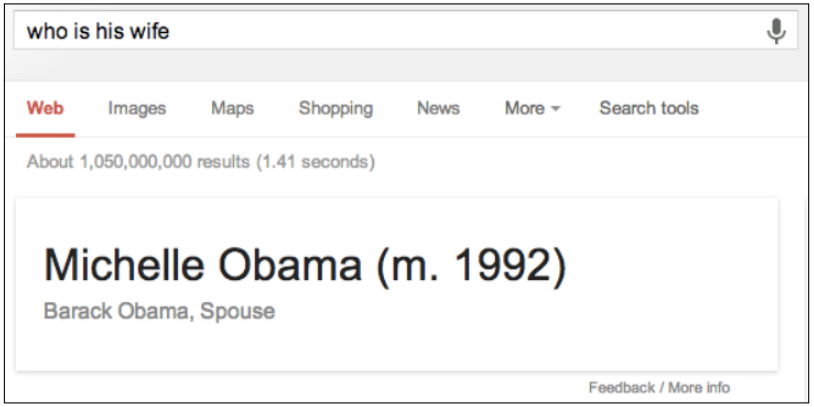
Did you know: The building is served by 36 lifts, some of which are

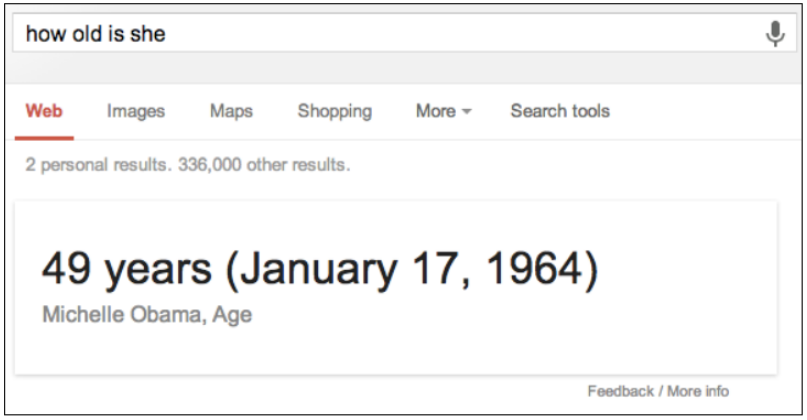
CLEVER THINGS ‘GOOGLE’ CAN DO #2

- **‘Conversational’ search (and “smart reply” in Gmail)**



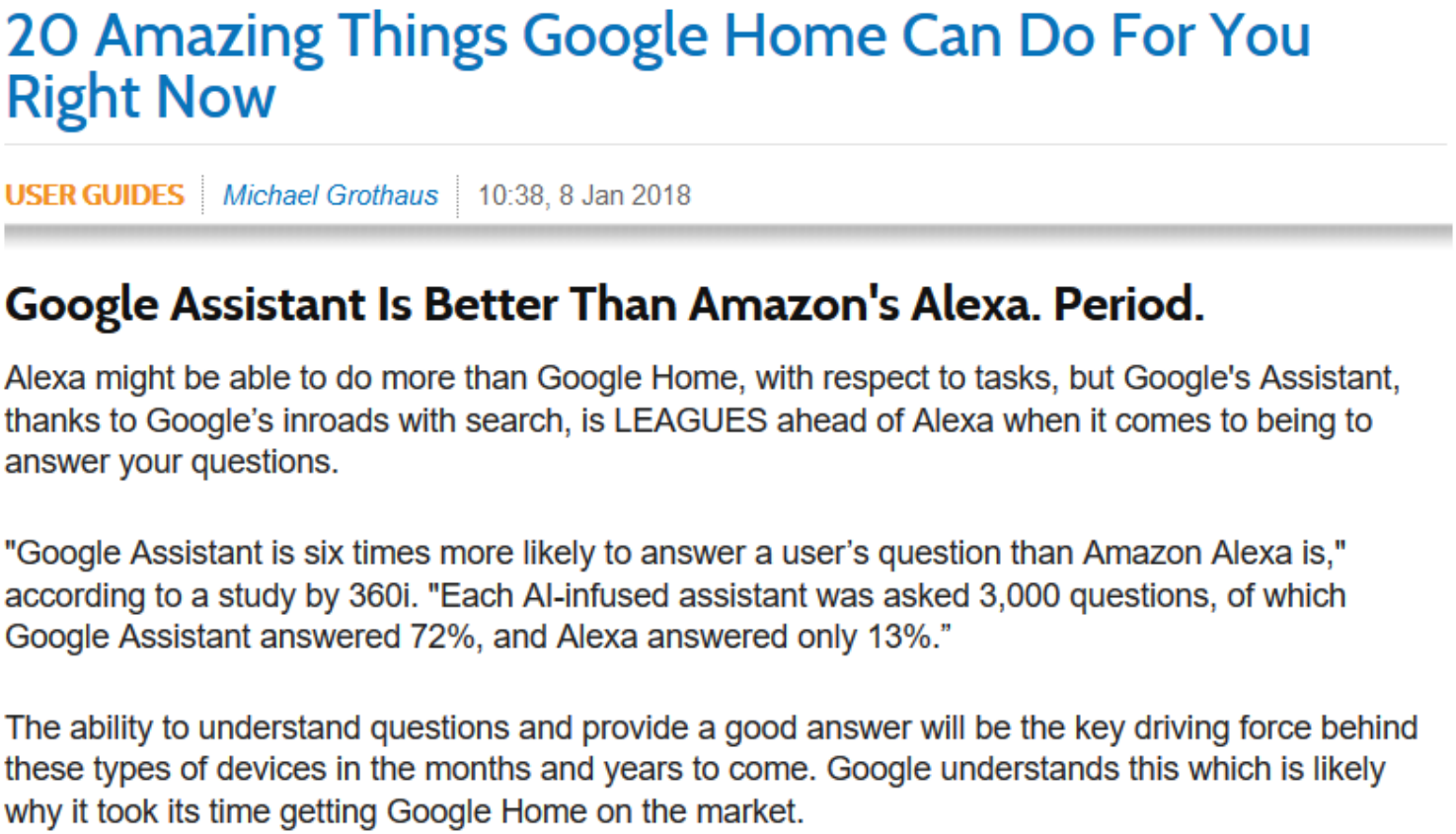
Then you’ll get your results:





MUCH 'CLEVERER' THAN ALEXA!

- **8 times as effective on general queries**



20 Amazing Things Google Home Can Do For You Right Now

USER GUIDES | *Michael Grothaus* | 10:38, 8 Jan 2018

Google Assistant Is Better Than Amazon's Alexa. Period.

Alexa might be able to do more than Google Home, with respect to tasks, but Google's Assistant, thanks to Google's inroads with search, is LEAGUES ahead of Alexa when it comes to being to answer your questions.

"Google Assistant is six times more likely to answer a user's question than Amazon Alexa is," according to a study by 360i. "Each AI-infused assistant was asked 3,000 questions, of which Google Assistant answered 72%, and Alexa answered only 13%."

The ability to understand questions and provide a good answer will be the key driving force behind these types of devices in the months and years to come. Google understands this which is likely why it took its time getting Google Home on the market.

GOOGLE - TRANSFORMED TO 'AI'/'ML' CO.

- **First two decades:**

  word-based statistics (relevance measures)

  link analysis (authority measures)

- **Knowledge-graph was first evidence of adding a 'semantic' layer with some searches based on formal knowledge models**

  bought *Freebase* in 2010 (1.9B triples)

"Strings => Things => Relationships"

SEMANTICS

Defn:

"the branch of linguistics and logic concerned with **meaning**."

"concerned with the **relationship** between linguistic symbols and their meaning or real-world objects the represent."

SEMANTICS AND 'TRIPLES'

- **Subject – Predicate – Object**

  "Bob is 35"
  "Bob is partner of Jack"
  "The sky" [S] "has the colour" [P] "blue" [O]

- "A semantic **triple** is the atomic data entity in the ***Resource Description Framework*** (*RDF*) data model" (*W3C*)

- Embed knowledge in a machine-readable format – each part of triple is individually addressable via a unique URI

CLEVER THINGS 'GOOGLE' CAN DO #3

- **Keep your calendar up to date and sort out all those annoying issues around time-zone changes, etc.**

  How was this done? (smart parsing / classification / etc.)

  How is this (mostly) done now?

  schema.org (our first 'proper' ontology)

CLEVER THINGS 'GOOGLE' CAN DO #3

- **Types of objects in schema.org**

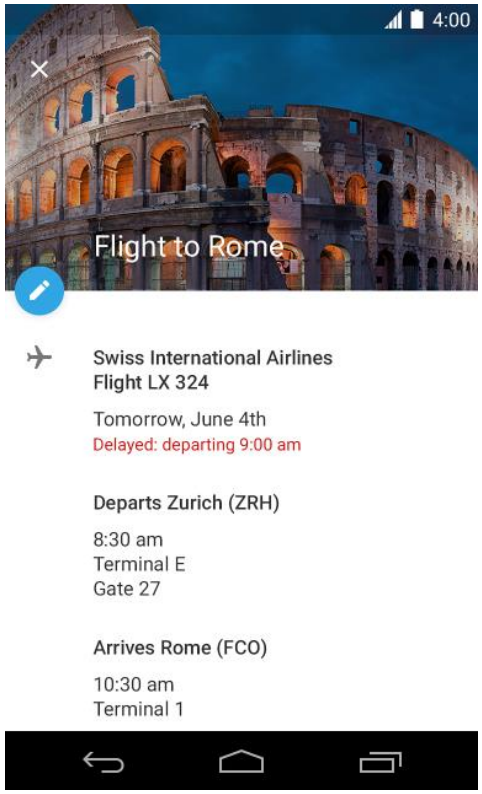
Here's a set of commonly used item types:

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
- Embedded non-text objects: AudioObject, ImageObject, VideoObject.
- Event.
- Organization.
- Person.
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer.
- Review, AggregateRating.

Getting Started - schema.org
schema.org/docs/gs.html

About this result Feedback

SCHEMA.ORG FOR FLIGHTS



SUPPORT

JSON-LD MICRODATA

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "FlightReservation",
  "reservationNumber": "RXJ34P",
  "reservationStatus": "http://schema.org/Confirmed",
  "underName": {
    "@type": "Person",
    "name": "Eva Green"
  },
  "reservationFor": {
    "@type": "Flight",
    "flightNumber": "110",
    "airline": {
      "@type": "Airline",
      "name": "United",
      "iataCode": "UA"
    },
    "departureAirport": {
      "@type": "Airport",
      "name": "San Francisco Airport",
      "iataCode": "SFO"
    },
    "departureTime": "2027-03-04T20:15:00-08:00",
    "arrivalAirport": {
      "@type": "Airport"
```

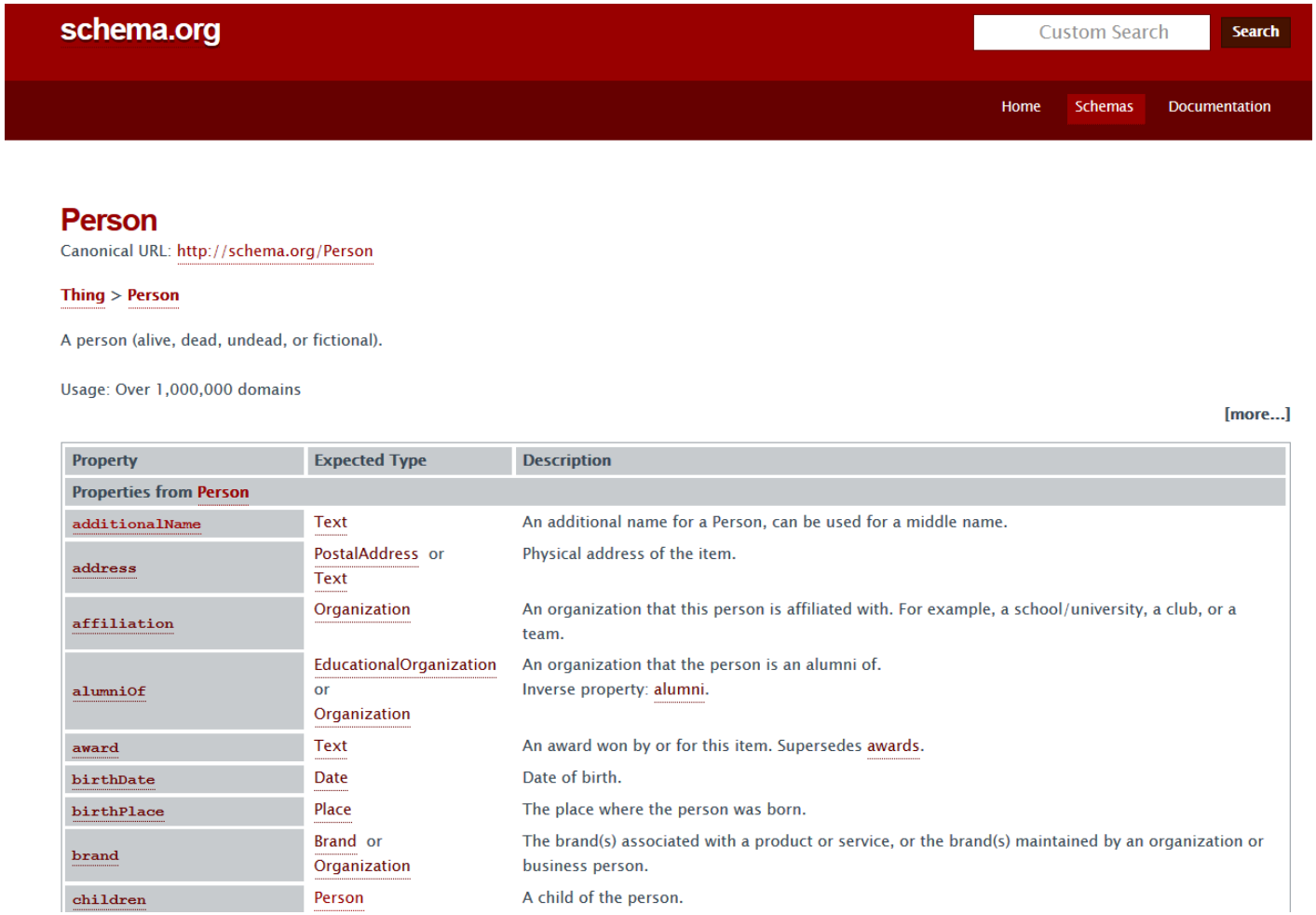
Person
Organization
Place
CreativeWork
Intangible
Action
MedicalEntity
Event

SCHEMA.ORG - FORMATS

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Person",
  "name": "John Doe",
  "jobTitle": "Graduate research assistar
  "affiliation": "University of Dreams",
  "additionalName": "Johnny",
  "url": "http://www.example.com",
  "address": {
    "@type": "PostalAddress",
    "streetAddress": "1234 Peach Drive",
    "addressLocality": "Wonderland",
    "addressRegion": "Georgia"
  }
}
</script>
```

```
<section itemscope itemtype="http://schema.org/Person">
  Hello, my name is
  <span itemprop="name">John Doe</span>,
  I am a
  <span itemprop="jobTitle">Graduate research assistant</span>
  at the
  <span itemprop="affiliation">University of Dreams</span>
  My friends call me
  <span itemprop="additionalName">Johnny</span>
  You can visit my homepage at
  <a href="http://www.example.com.com" itemprop="url">www.example
  <section itemprop="address" itemscope itemtype="http://schema.o
    I live at
    <span itemprop="streetAddress">1234 Peach Drive</span>
    <span itemprop="addressLocality">Warner Robins</span>
    <span itemprop="addressRegion">Georgia</span>.
  </section>
</section>
```

SCHEMA.ORG



schema.org

Custom Search | Search

Home | Schemas | Documentation

Person

Canonical URL: http://schema.org/Person

Thing > Person

A person (alive, dead, undead, or fictional).

Usage: Over 1,000,000 domains

[more...]

Property	Expected Type	Description
Properties from Person		
additionalName	Text	An additional name for a Person, can be used for a middle name.
address	PostalAddress or Text	Physical address of the item.
affiliation	Organization	An organization that this person is affiliated with. For example, a school/university, a club, or a team.
alumniOf	EducationalOrganization or Organization	An organization that the person is an alumni of. Inverse property: alumni.
award	Text	An award won by or for this item. Supersedes awards.
birthDate	Date	Date of birth.
birthPlace	Place	The place where the person was born.
brand	Brand or Organization	The brand(s) associated with a product or service, or the brand(s) maintained by an organization or business person.
children	Person	A child of the person.

A shared vocabulary (**ontology**) for representing *things* that typically occur on the web.

Founded and funded by Google, Microsoft, Yandex, and Yahoo. Developed through open community process.

Example:
http://schema.org/Person

Full list of types:
schema.org/docs/full.html

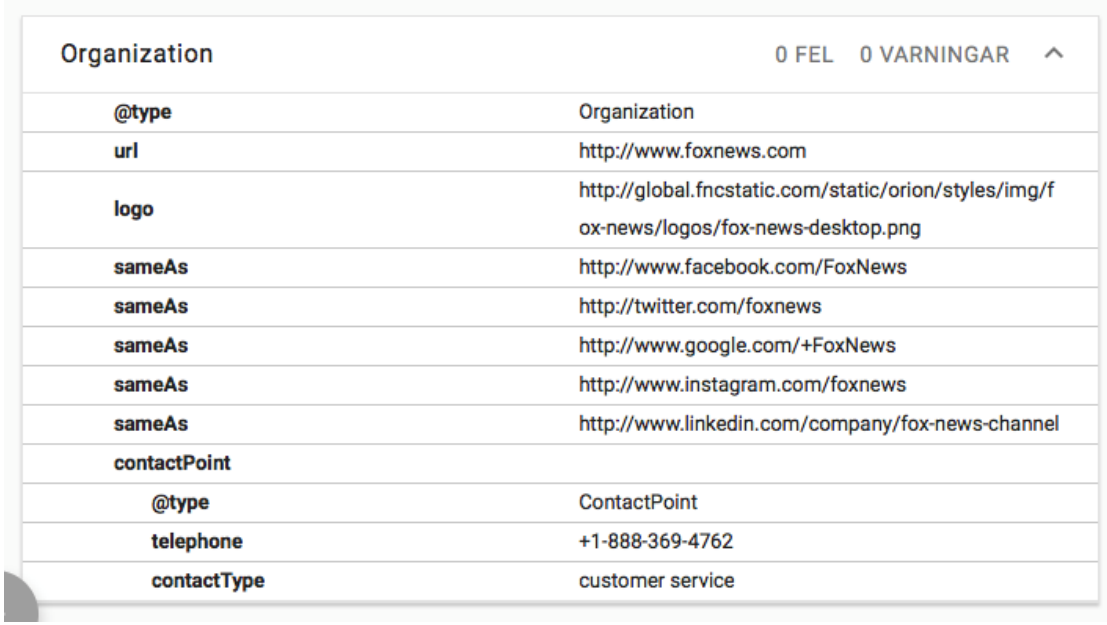
EMBEDDED STRUCTURED DATA

Google Structured Data Testing Tool:
https://goo.gl/cHMn5y

Yandex Structured data validator:
https://goo.gl/581Lrj

Addresses to test:

- http://xd-protege.com/~karl/
- http://www.foxnews.com
- https://www.theguardian.com/



Organization	0 FEL 0 VARNINGAR
@type	Organization
url	http://www.foxnews.com
logo	http://global.fncstatic.com/static/orion/styles/img/fox-news/logos/fox-news-desktop.png
sameAs	http://www.facebook.com/FoxNews
sameAs	http://twitter.com/foxnews
sameAs	http://www.google.com/+FoxNews
sameAs	http://www.instagram.com/foxnews
sameAs	http://www.linkedin.com/company/fox-news-channel
contactPoint	
@type	ContactPoint
telephone	+1-888-369-4762
contactType	customer service

CLEVER THINGS 'GOOGLE' CAN DO #4

- **Provide location-aware data / amazing map-linked functions**

  Geo ontologies:

  - WGS84 Geo
  - Linked Geo Data
  - OpenStreetView

- **Many open-source APIs**

  - RESTful API

THE LOD CLOUD DIAGRAM

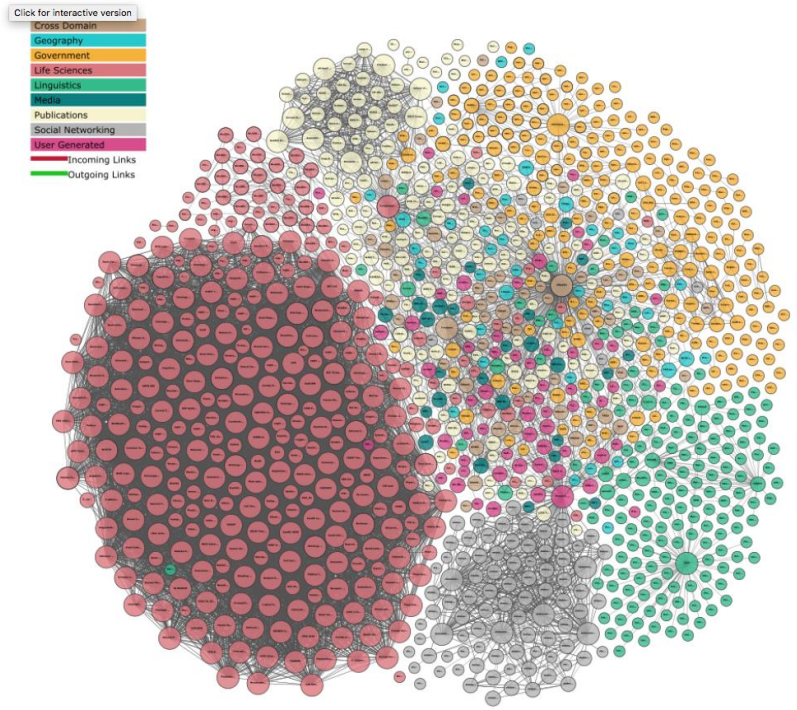
LOD (Linked Open Data) can be published as datasets, not just embedded in web pages.

http://lod-cloud.net indexes such published datasets. Go there and click on the picture to bring up an interactive SVG image.

Click-and-drag on the background to move the graph, scroll up or down to zoom in and out.

Click a single node to bring up that dataset's metadata.

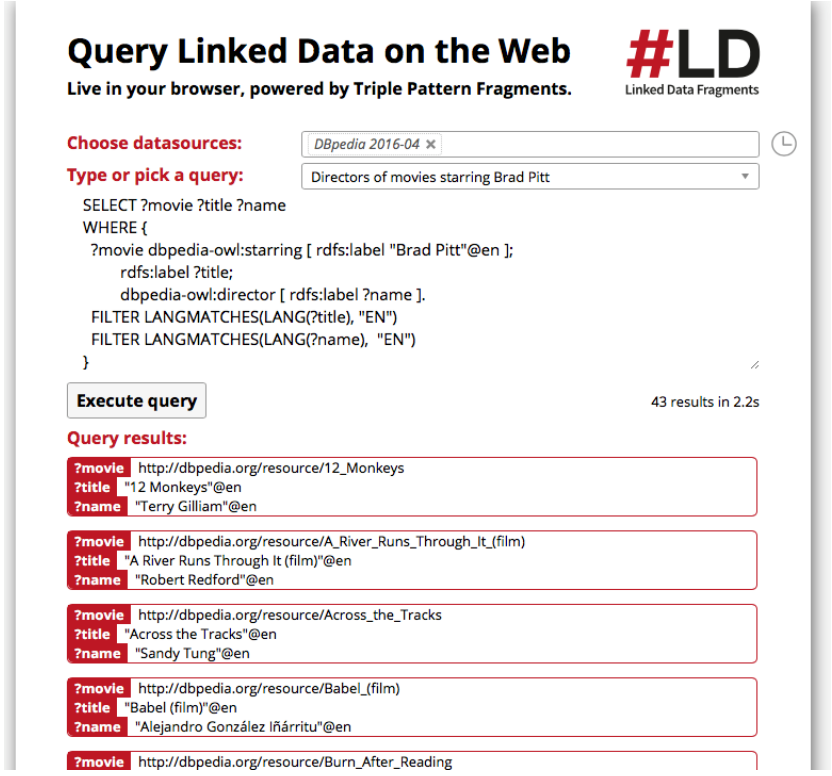
Everything red linked to Life Sciences



(If possible, use Google Chrome to avoid browser lag.)

Slide content by KARL HAMMAR

LOD CLIENT 1: LINKED DATA FRAGMENTS



http://client.linkeddatafragments.org/

Queries semantic data (possibly distributed over multiple endpoints) using the LDF protocol.

Example queries: See drop-down on client webpage.

Note in particular the query about San Francisco authors, which runs across two datasets *not designed with the intent to be interoperable with one another!*

Slide content by KARL HAMMAR

LOD CLIENT 2: WIKIDATA QUERY SERVICE

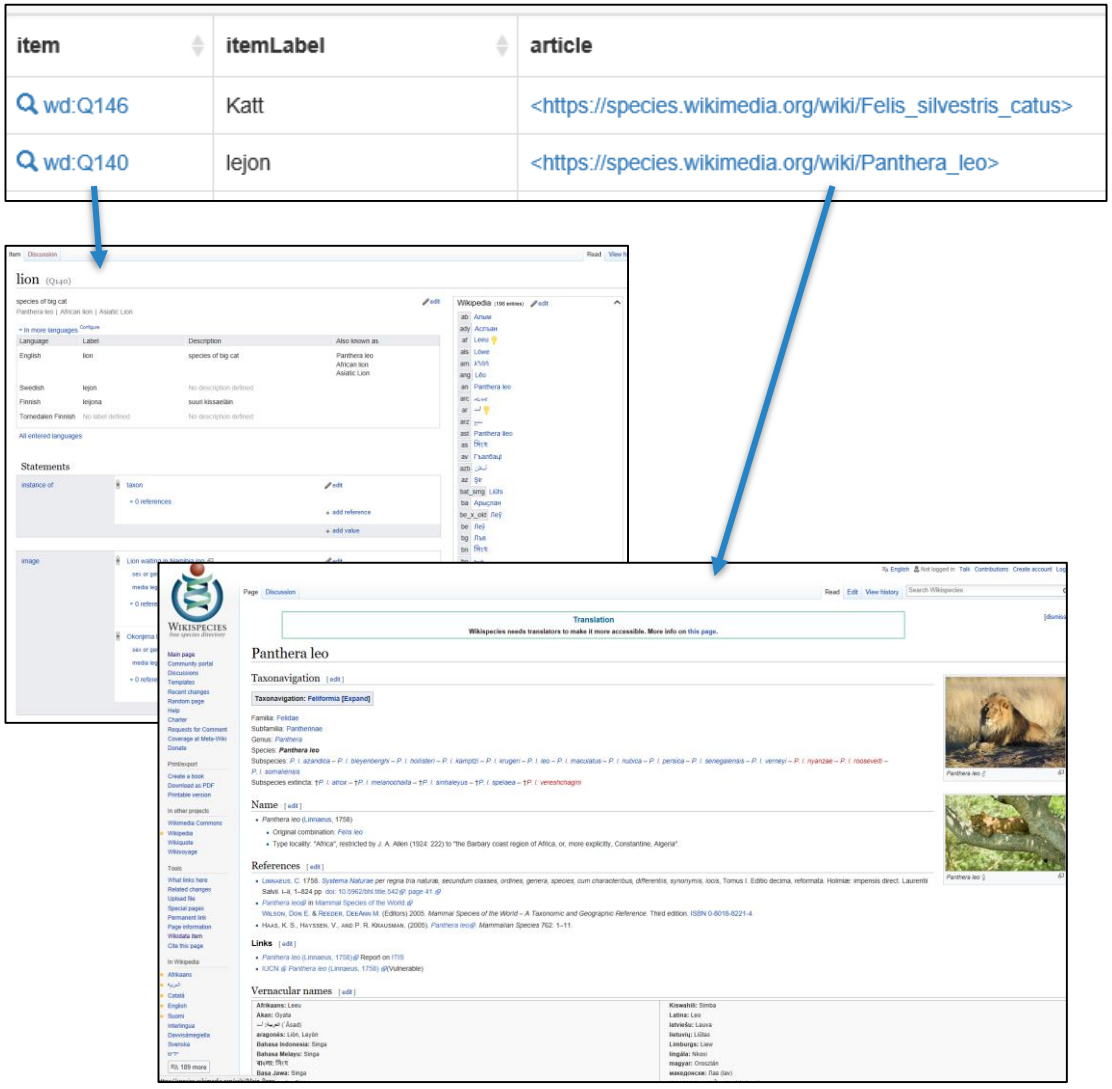
https://query.wikidata.org

Queries content in WikiData (the next-generation knowledge graph underpinnings of Wikipedia)

https://goo.gl/qhz88X - Query that links formal WikiData representations to corresponding WikiSpecies articles.

Example WikiData nodes:

- https://wikidata.org/wiki/Q140
- https://wikidata.org/wiki/Q7891
- https://wikidata.org/wiki/Q156050
- https://wikidata.org/wiki/Q192717



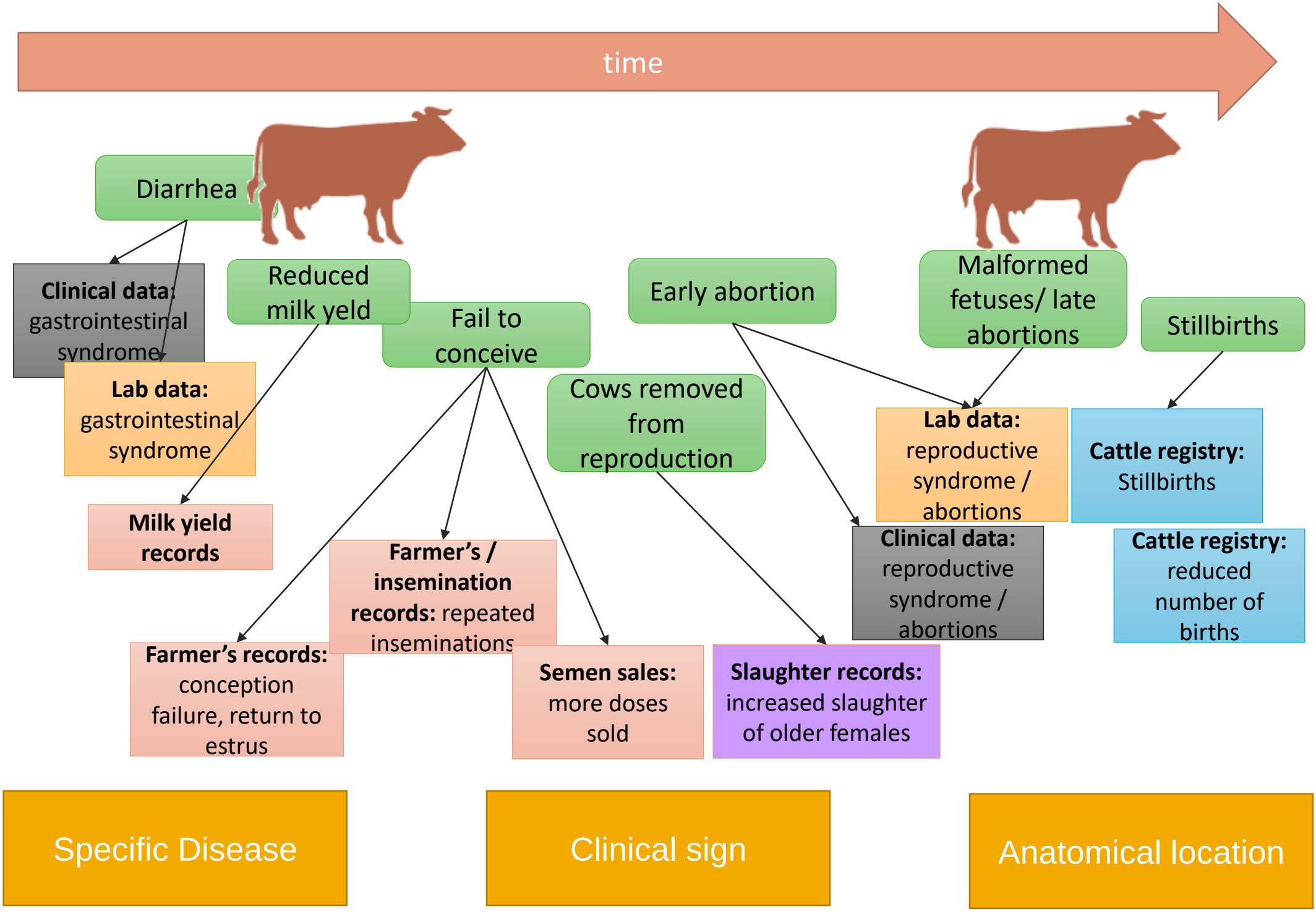
Slide content by KARL HAMMAR

MESSAGES TO TAKE AWAY

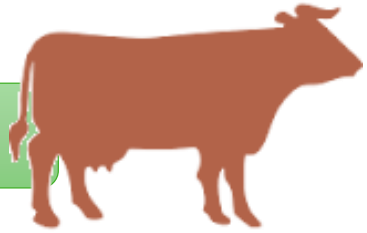
1. Data standardisation leads to static and largely ineffective knowledge sharing/use (and in most cases is impractical)
2. We live in an open world, and 'closed world' assumptions will inevitably lead to limitations in knowledge management
3. Snippets of knowledge/information can be useful and may (should) be captured as such
4. Medical data/knowledge is not uniquely complex

HOW CAN IT BE USEFUL TO US?

SVA


time
Diarrhea
Reduced milk yeld
Fail to conceive
Early abortion
Malformed fetuses/ late abortions
Stillbirths
Clinical data: gastrointestinal syndrome
Lab data: gastrointestinal syndrome
Milk yield records
Farmer's / insemination records: repeated inseminations
Farmer's records: conception failure, return to estrus
Semen sales: more doses sold
Cows removed from reproduction
Slaughter records: increased slaughter of older females
Lab data: reproductive syndrome / abortions
Clinical data: reproductive syndrome / abortions
Cattle registry: Stillbirths
Cattle registry: reduced number of births
Specific Disease
Clinical sign
Anatomical location

time



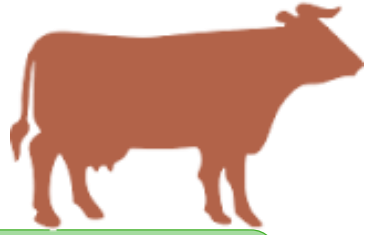
Diarrhea

Reduced milk yeld

Fail to conceive

Early abortion

Cows removed from reproduction



Malformed fetuses/ late abortions

Stillbirths

”Things not strings”

Specific Disease

Clinical sign

Anatomical location

HOW DO WE THINK ABOUT DATA?

Date	Herd	Animal	Test requested	Sample	Result
2018-01-01	2345	45198	Paramyxovirus PCR	Blood	Positive
2018-01-22	2345	Daisy	Salmonella serotyping	Feces	O4+ 5- 27-
2018-02-15	7531		Bacterial growth	Milk	Staphylococcus aureus growth
2018-03-01	2518	5458	Leisgmanua ELISA	Serum	183/

EXAMPLE: SYNDROMIC CLASSIFICATION

Microbial growth on trachea swab	Syndrome = Respiratory
Histological examination on sample from lung	Syndrome = Respiratory
Serological tests for IBR	Syndrome = Respiratory
PCR test for Porcine Influenza	Syndrome = Respiratory
Streptococcus serotyping on nasal swab	Syndrome = Respiratory

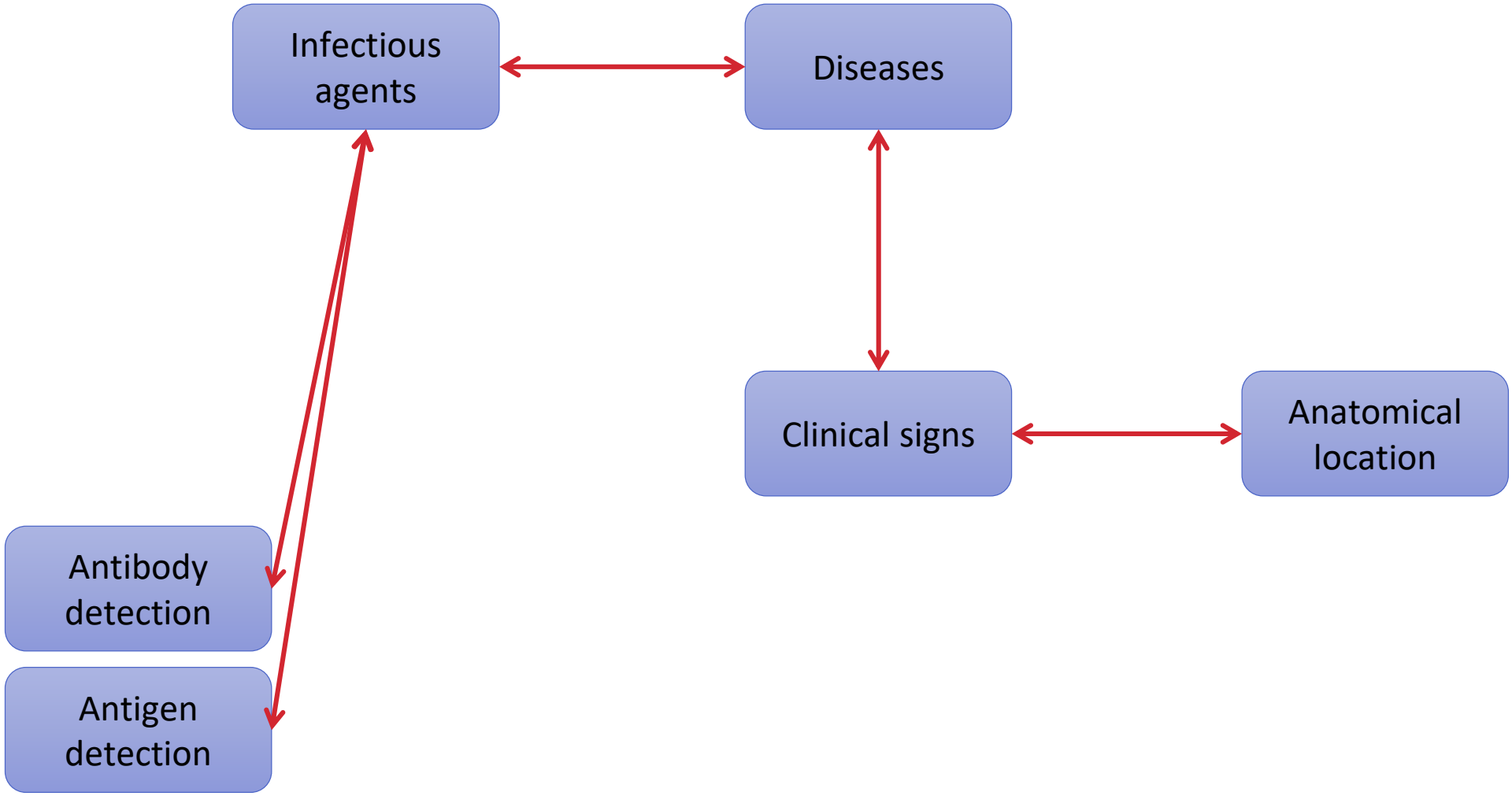
CHALLENGES

- One dataset
  - Secondary use
  - Historical compatibility
  - Prospective compatibility
    - Changes in the data
    - Changes in the question
- Interoperability
  - Datasets from the same type of activity
  - Datasets from a different source
- Reuse of work previously done (by you or others)
- Transparency

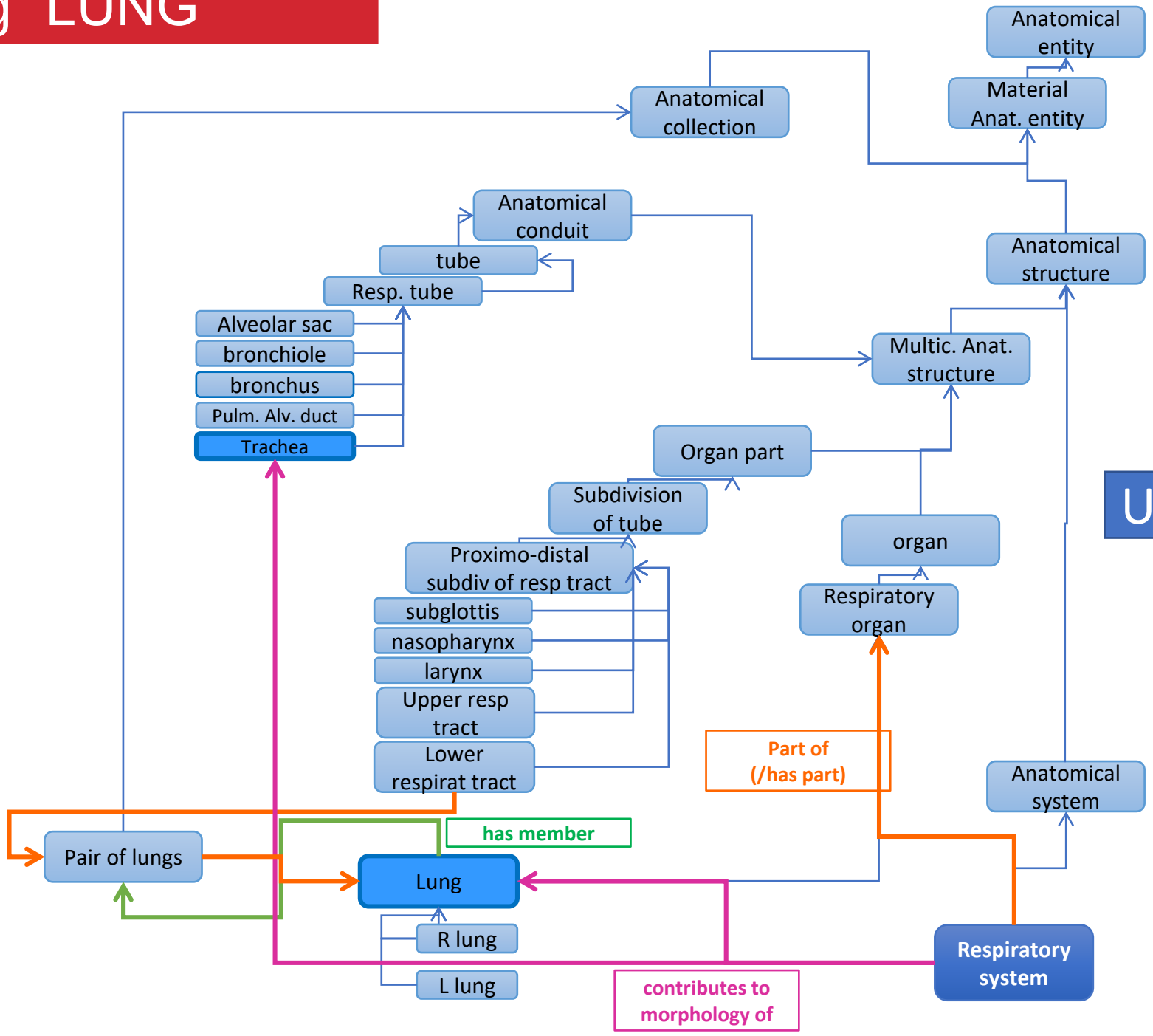
KNOWLEDGE MODEL?

Microbial growth on trachea swab	→	Syndrome = Respiratory
Histological examination on sample from lung	→	Syndrome = Respiratory
Serological tests for IBR	→	Syndrome = Respiratory
PCR test for Porcine Influenza	→	Syndrome = Respiratory
Streptococcus genotyping on nasal swab	→	Syndrome = Respiratory

Modeling Knowledge: Classes and Properties



The ”thing” LUNG

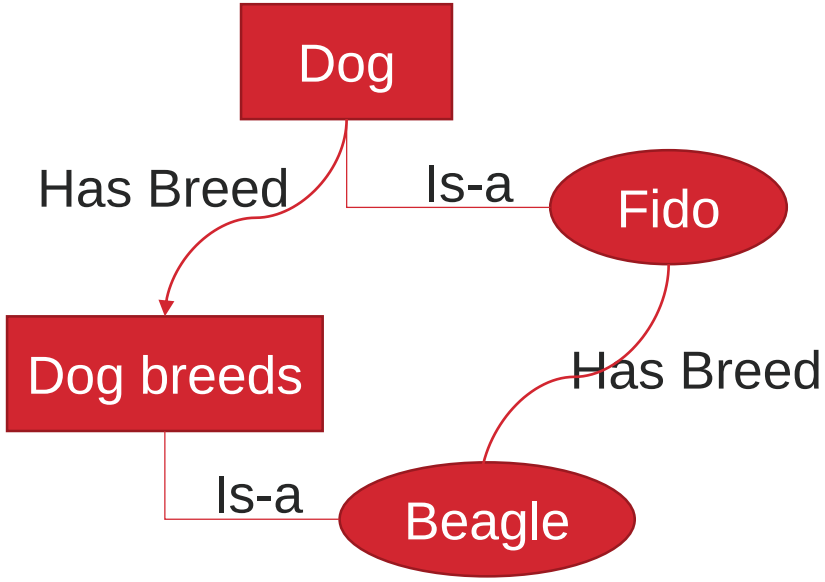


- https://bioportal.bioontology.org/
- Look for Uberon
- Visit the ”Classes” tab
- Look for ”lung” or play around

- Browse ”anatomical system”

CLASSES (CONCEPTS) AND PROPERTIES (RELATIONSHIPS)

- Object properties
  - Cows *are part of* Herds
  - Daisy *is a* cow
- Data properties
  - Daisy *has DOB* 2015-03-21
- Class subset only if ”is-a”
  - Fido *is a* dog
  - ~~Beagle *is a* dog~~
  - ~~Animal *is a* herd~~



WHY IS IT IMPORTANT?

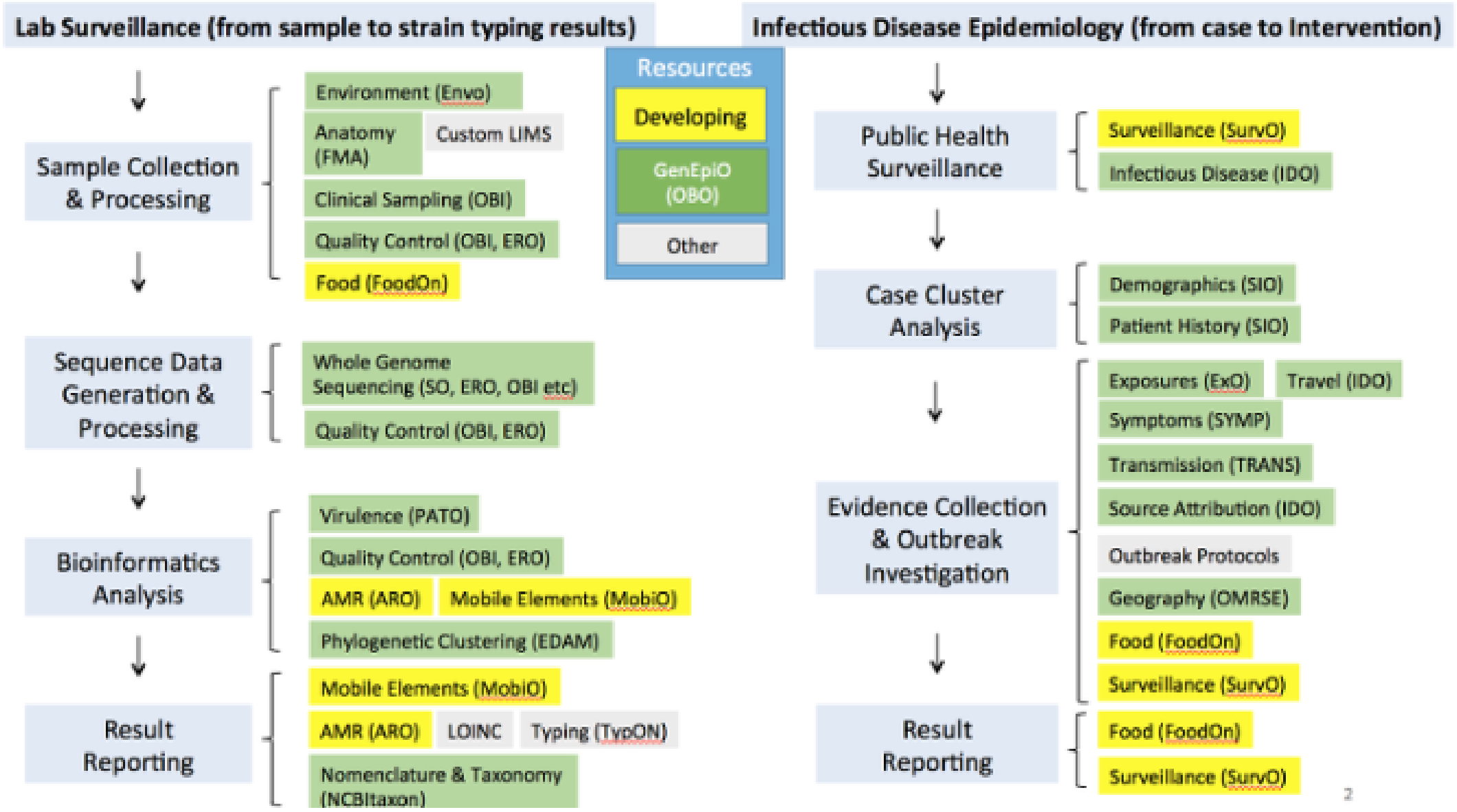
- Reasoning

- … besides everything else we have been talking about!

NOT TO DISPAIR WITH DETAILS...

- Remember the REUSE:

PRIMARY SURVEILLANCE DATA



SECONDARY SURVEILLANCE DATA

Health event

Recorded *Observations*

Geographical information

Observer (person)

Registry (database)

Target population

Observation context

Clinical observation

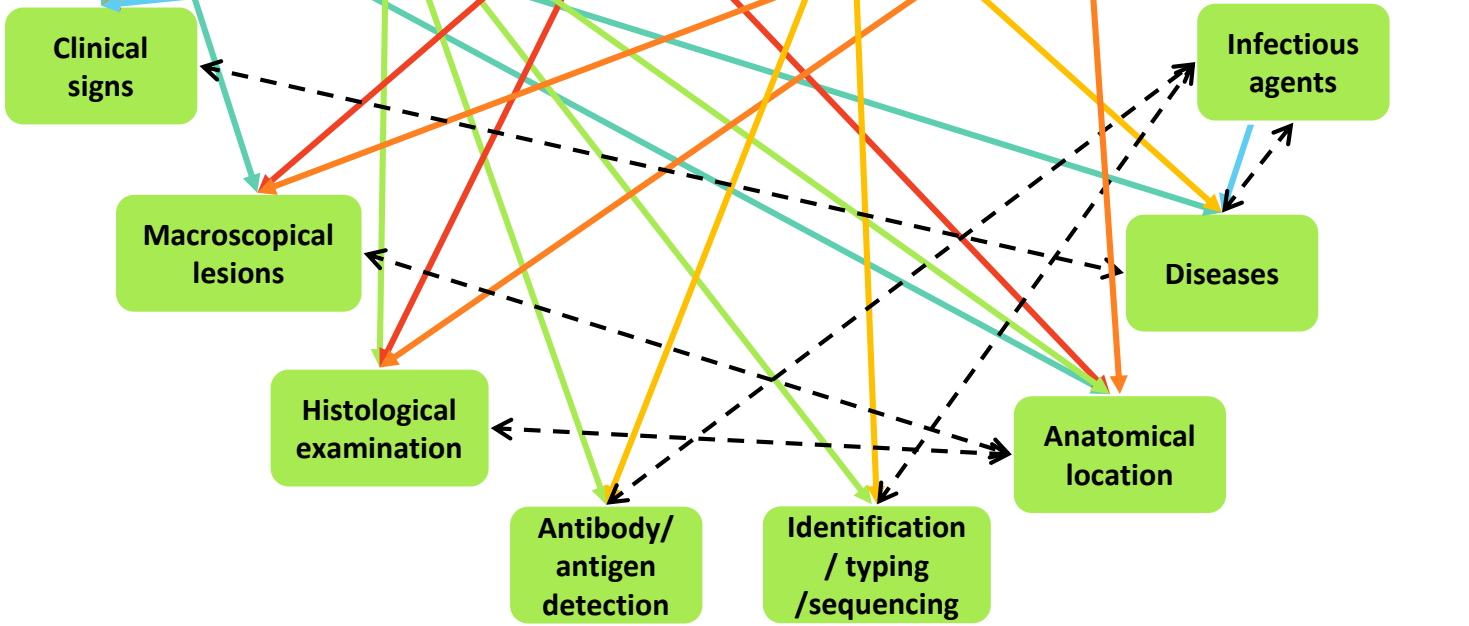
Laboratory tests (Diagnostic)

Slaughter

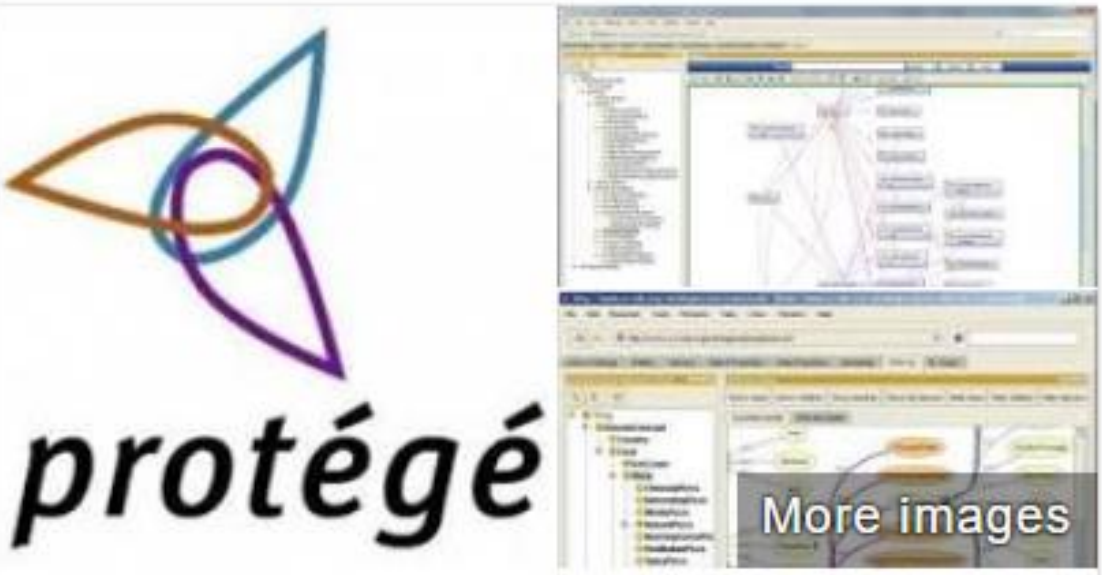
Surveillance (active and passive)

Necropsy

Mandatory life event reporting (birth, death, etc)



Health information



Protégé

Software

Protégé is a free, open source ontology editor and a knowledge management system. Protégé provides a graphic user interface to define ontologies. Wikipedia

Written in: Java

License: BSD 2-clause

Operating system: Linux, Mac OS X & Windows

Developer(s): Stanford Center for Biomedical Informatics Research

Stable release: 5.2.0 / 15 March 2017; 11 months ago

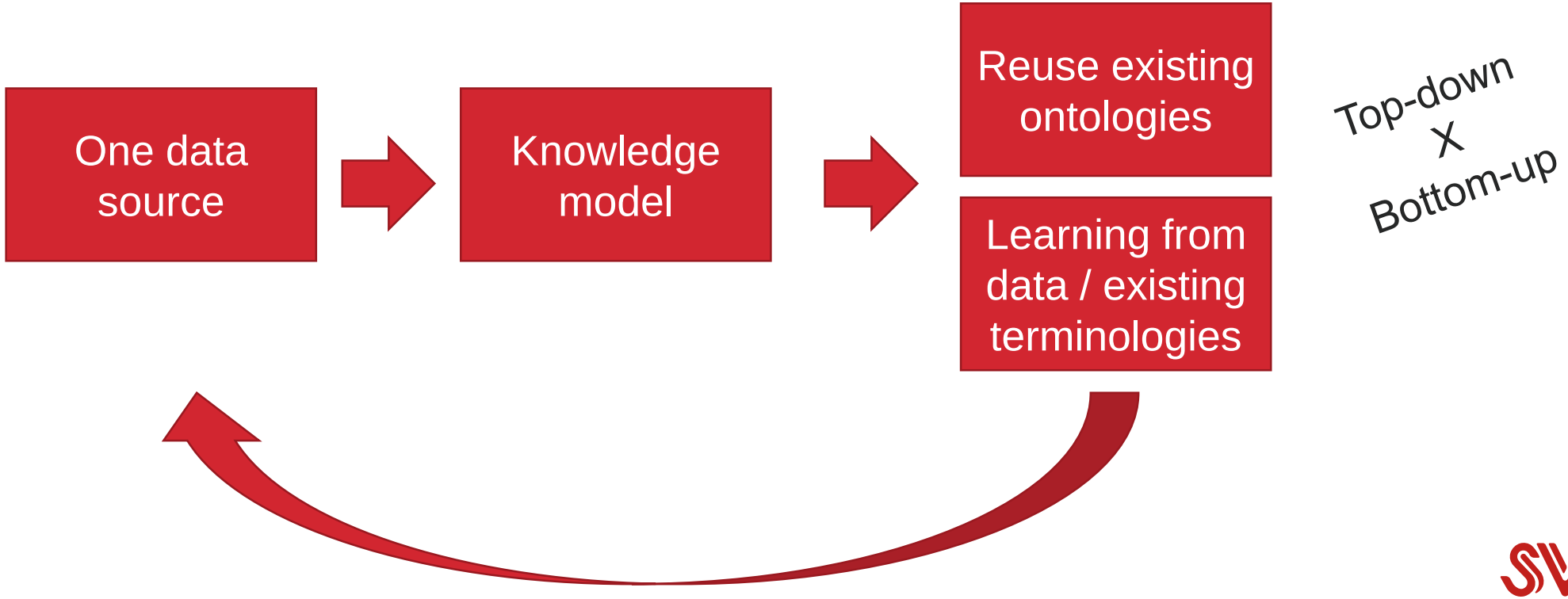
Initial release: 11 November 1999; 18 years ago

AHSO CONSTRUCTION

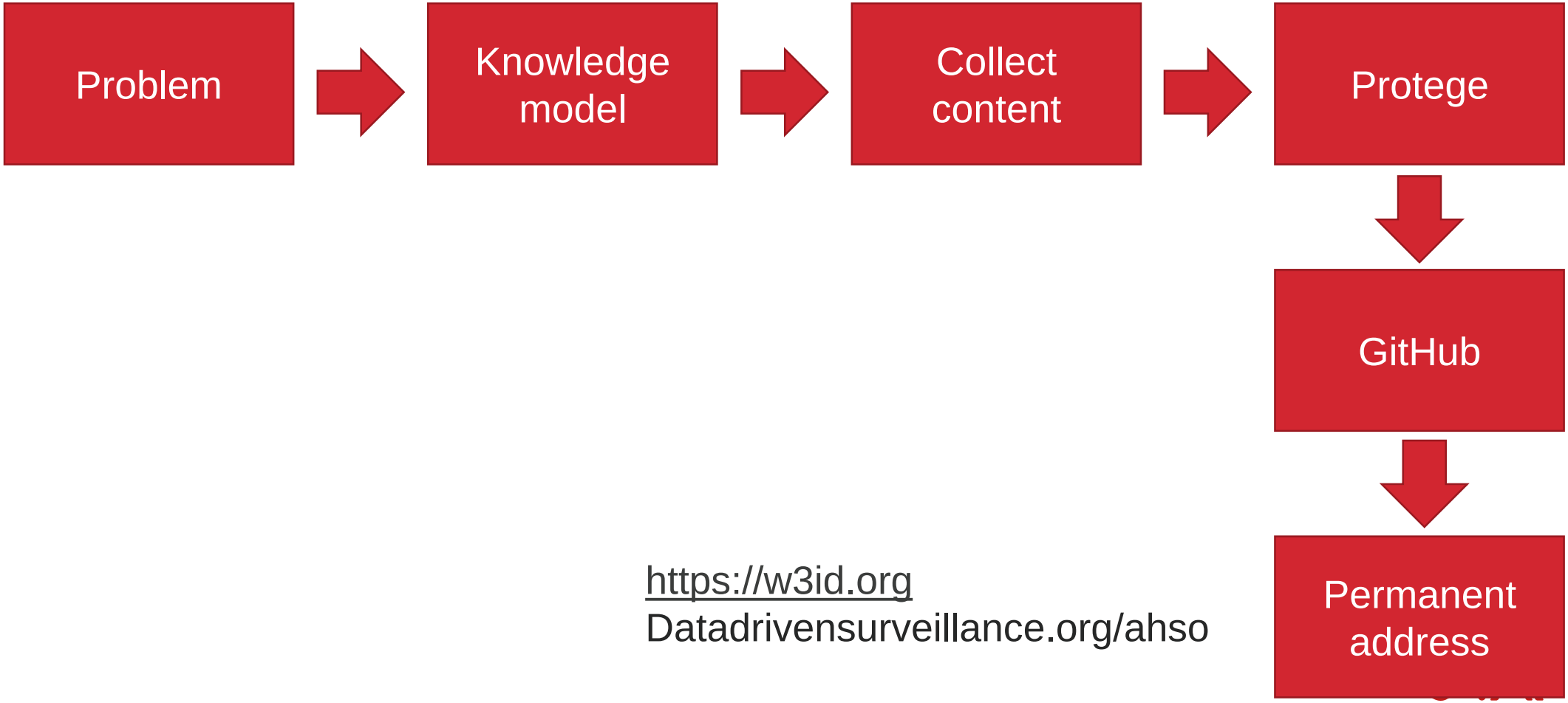
- Venom – terminology to description logics
- Pathology – 2 codes interoperability
- ORION – specific surveillance problems

AHSO CONSTRUCTION

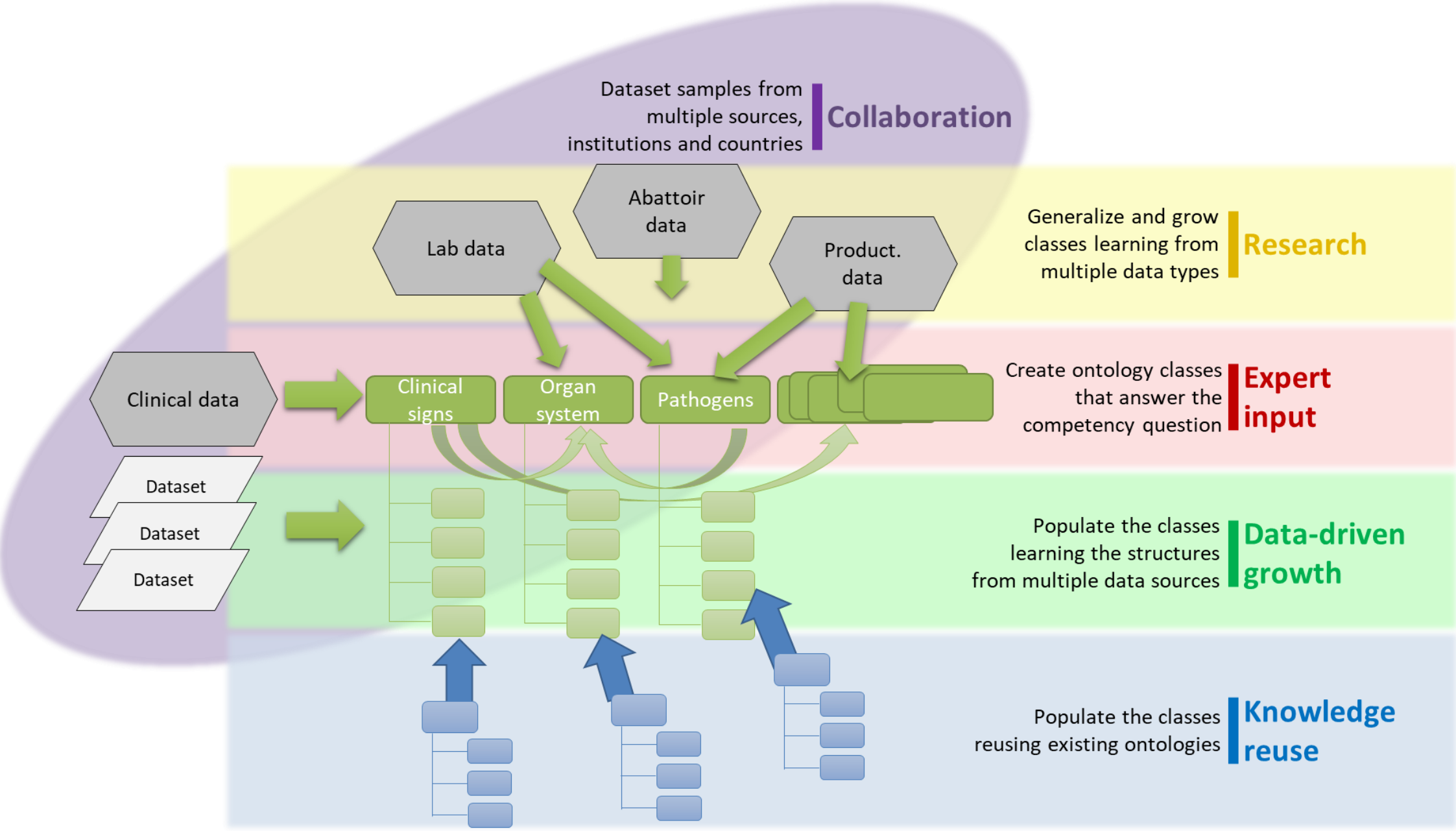
- Small functional modules



BUILDING IT



https://w3id.org
Datadrivensurveillance.org/ahso


Dataset samples from multiple sources, institutions and countries
Collaboration
Abattoir data
Lab data
Product. data
Generalize and grow classes learning from multiple data types
Research
Clinical data
Clinical signs
Organ system
Pathogens
Create ontology classes that answer the competency question
Expert input
Dataset
Dataset
Dataset
Populate the classes learning the structures from multiple data sources
Data-driven growth
Populate the classes reusing existing ontologies
Knowledge reuse


Datadrivensurveillance.org


SVA
Healthy animals, Safe people.

IS IT WORTH THE EFFORT?

IS IT WORTH THE EFFORT?

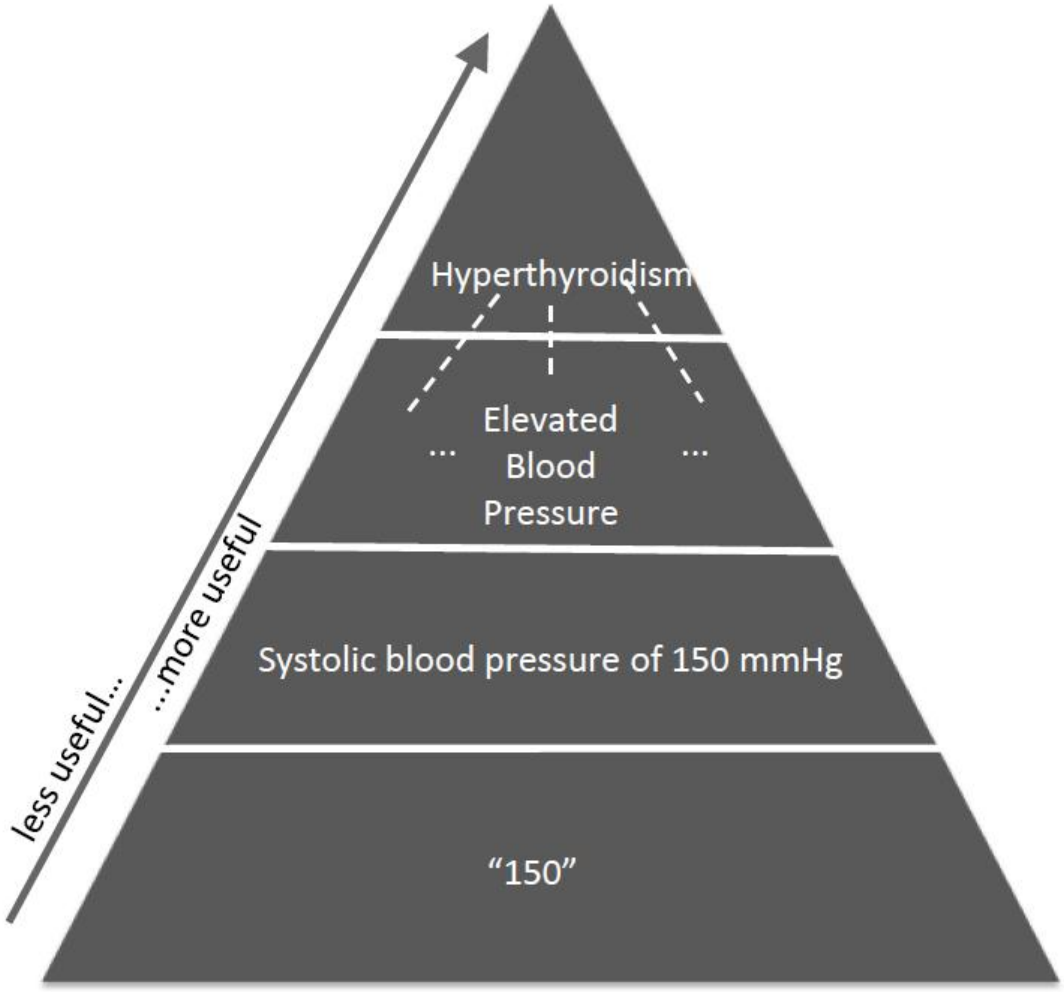
IS IT WORTH THE EFFORT?

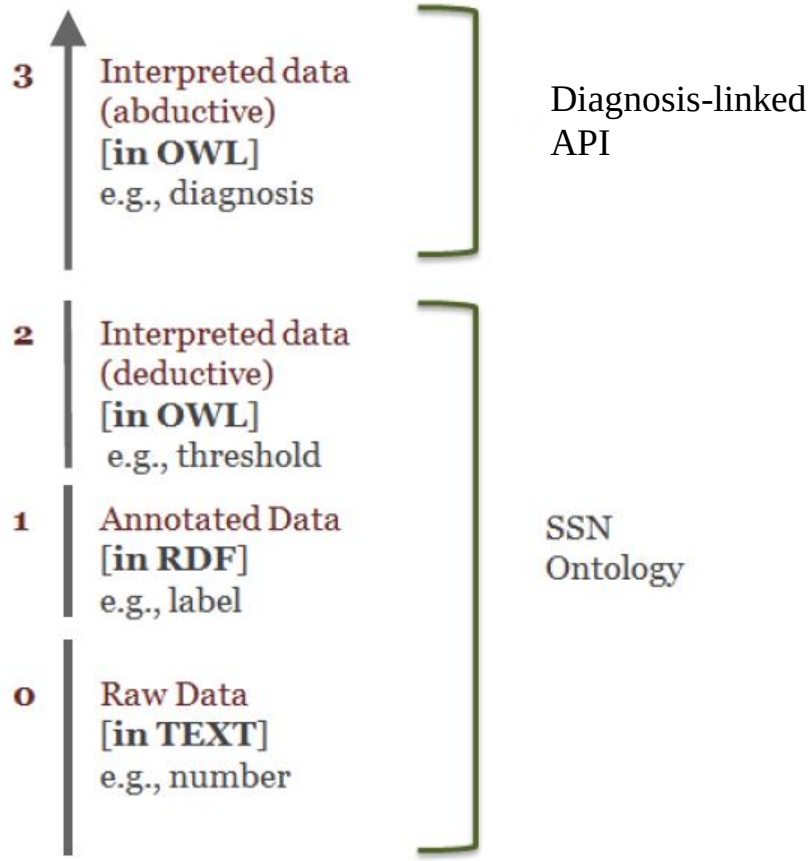
We need 'smart data' to make sense of *Big Data*

It provides value by harnessing the challenges posed by volume, velocity, variety, veracity…

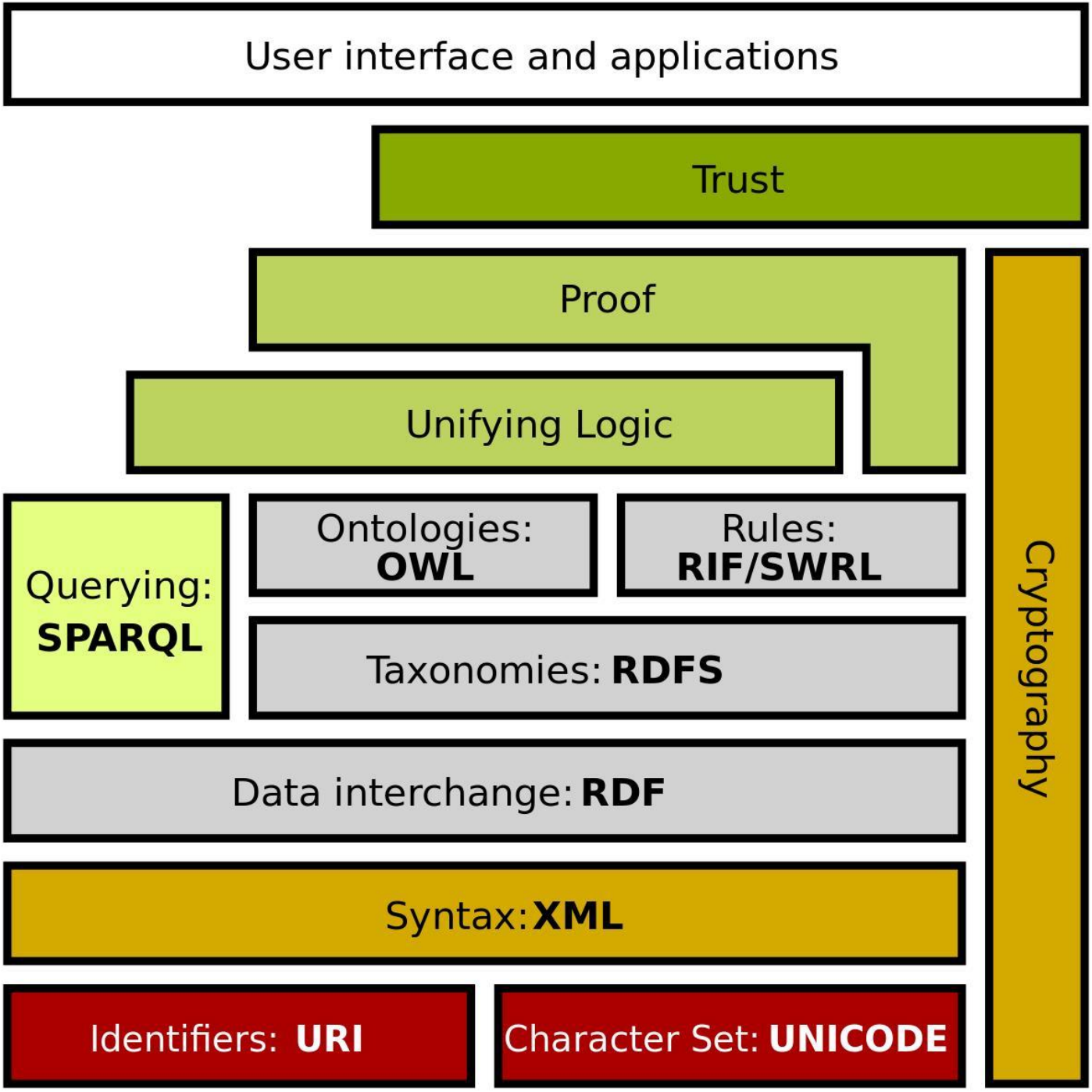
… to provide actionable information and improve decision making.

WHY DOES IT WORK?





WHY DOES IT WORK?



WHY DOES IT WORK?

Agreement on all levels of the stack:

- Shared network infrastructure (TCP/IP)
- Shared transmission protocol (HTTP)
- Shared node identifier structure (URIs)
- Shared data syntax for graphs (RDF)
- (Partially) shared semantics and schemas (RDFS, OWL ontologies, Schema.org)

=> Integration of data still non-trivial, but **significantly** easier.

No need for technical integration work. No need for protocol-level integration work. No need to align and mint new identifiers. No need to work around graph representation differences.

Just need to align concepts and relations within the domain of discourse

Slide content by KARL HAMMAR

SOME HEAVILY-USED ONTOLOGIES

Schema.org – Web content in general

Dublin Core - Metadata

FOAF – Friend-of-a-Friend, people and organisations

DBPedia – Contents of Wikipedia fact-boxes

WGS84 Geo – Latitude/longitude

Linked Geo Data

GoodRelations – E-Commerce

W3C ORG – Organisations

Semantic Sensor Network (SSN)

EXTENT OF ONTOLOGIES

Table 2. Description of the evaluation framework

	Data set	Triples	Size (MB)	Subjects	Predicates	Objects
Media	Jamendo	1,049,637	144	335,925	26	440,602
	LinkedMDB	6,147,996	850	694,400	222	2,052,959
	Dbtune	58,920,361	9566	12,401,228	394	14,264,221
	Flickr Event Media	49,107,168	6714	5,490,007	23	15,041,664
Publications	SWDF	101,321	16	10,476	132	34,609
	Faceted DBLP	60,139,734	9799	3,591,091	27	25,154,979
Knowledge	Wordnet 3.0	6,257,922	974	1,100,503	85	1,689,363
	Dbpedia 3-8	431,440,396	63,053	24,791,728	57,986	108,927,201
Government	2011 Australian Census	361,842	52	51,768	26	6901
	2000 US Census	149,182,415	21,796	23,904,658	429	23,996,813
Sensors	AEMET	3,547,154	726	394,289	23	793,664
	Ike	514,824,008	102,662	114,484,017	10	114,629,189
Geography	Linked Geo Data	274,668,813	39,423	51,916,995	18,272	121,749,861
Biology	Affymetrix	44,207,145	6526	1,421,763	105	13,240,270

Personalized Health and Objectives: one size does not fit all

Millions of people - > one treatment

Wearable and Sensor data

Near Future: Analyzing a Multifaceted Continuous Stream of Diverse Data

Personal level Signals

Population level Signals

Pollen.com

CO_2

Public level Signals

CDC

CENTERS FOR DISEASE CONTROL AND PREVENTION

"...**pediatric patients** report **variability** in **asthma symptoms** over time, even when asthma medications are taken."[1]