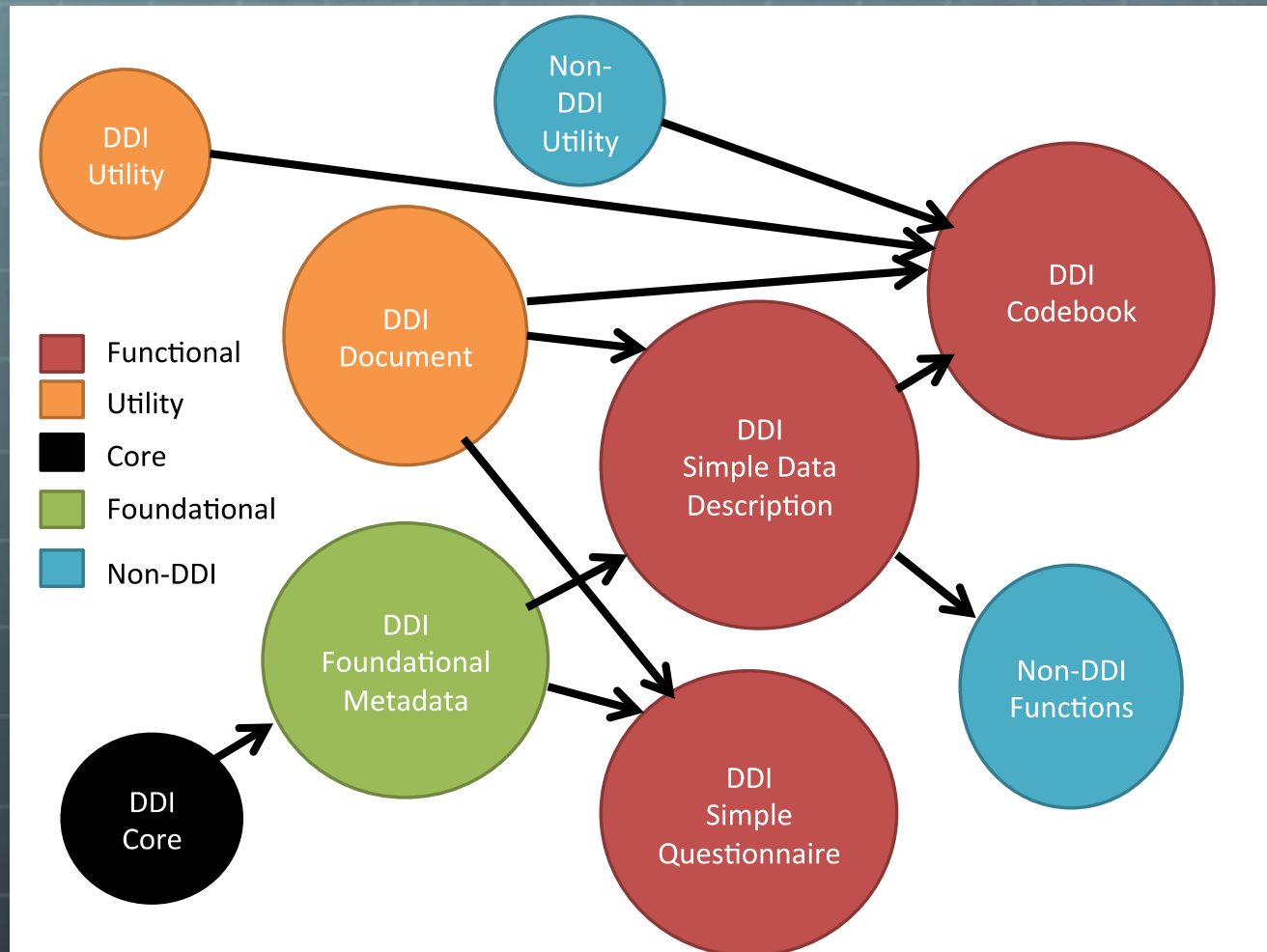


Moving Forward **DDI Support for Data Discovery**

Jay Greenfield, Ph.D.
EDDI 2013
Paris, France

Moving Forward DDI will become a model-based specification with **packages** and **views**



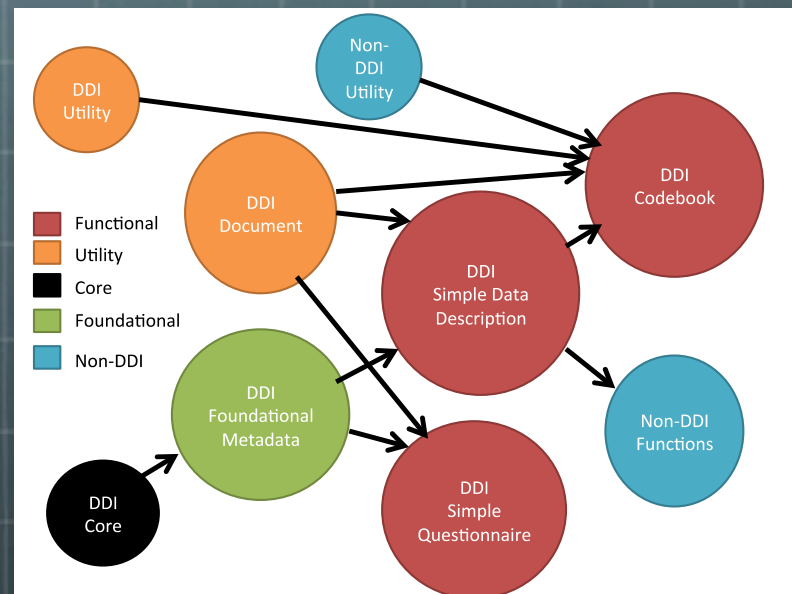
Moving Forward DDI will become a model-based specification with packages and views

- DDI **Foundational Metadata** will be built on top of **core classes and core relationships**.

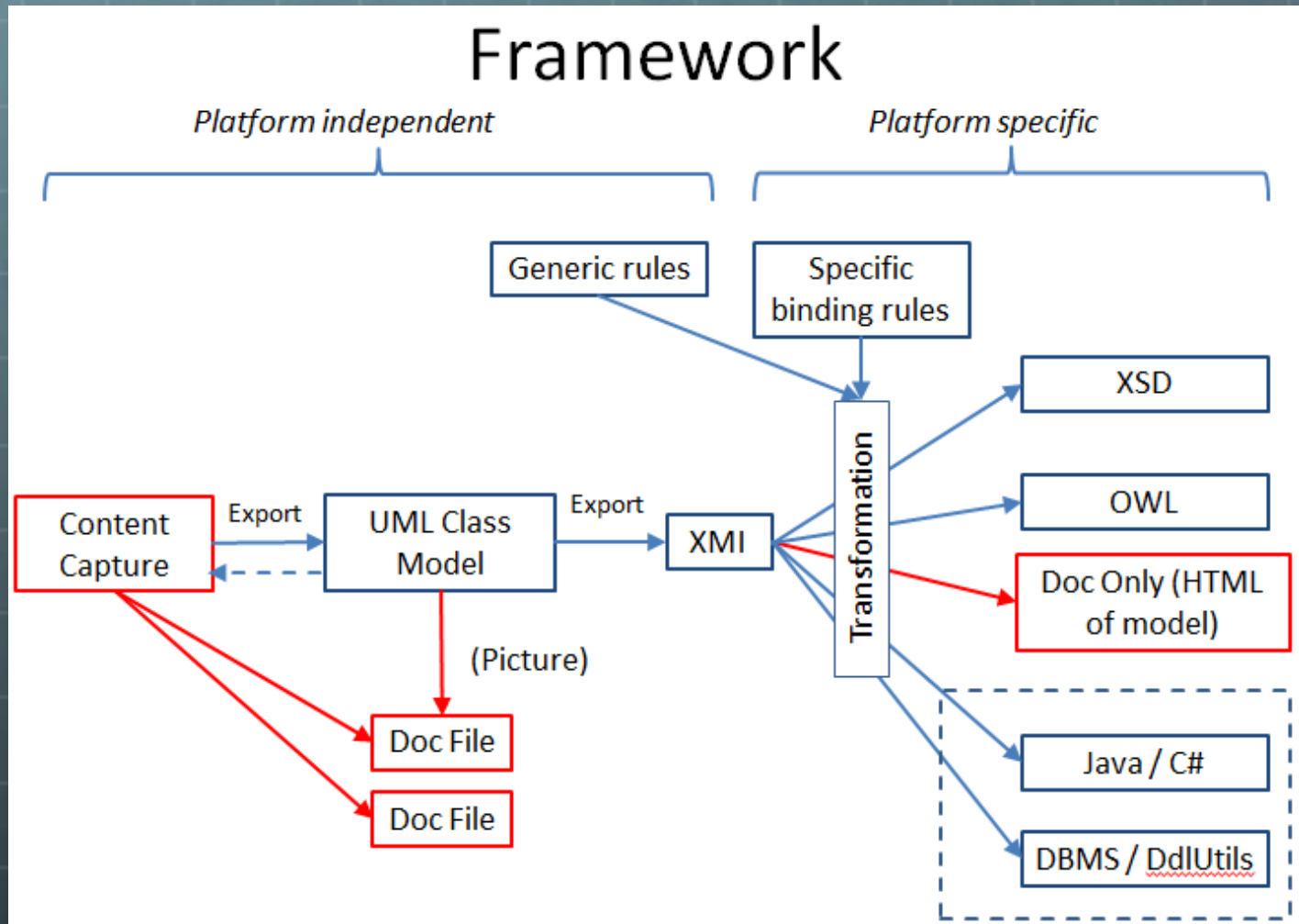
- Simple Codebook** will specialize the foundational metadata.

- Advanced Codebook** will specialize Simple Codebook.

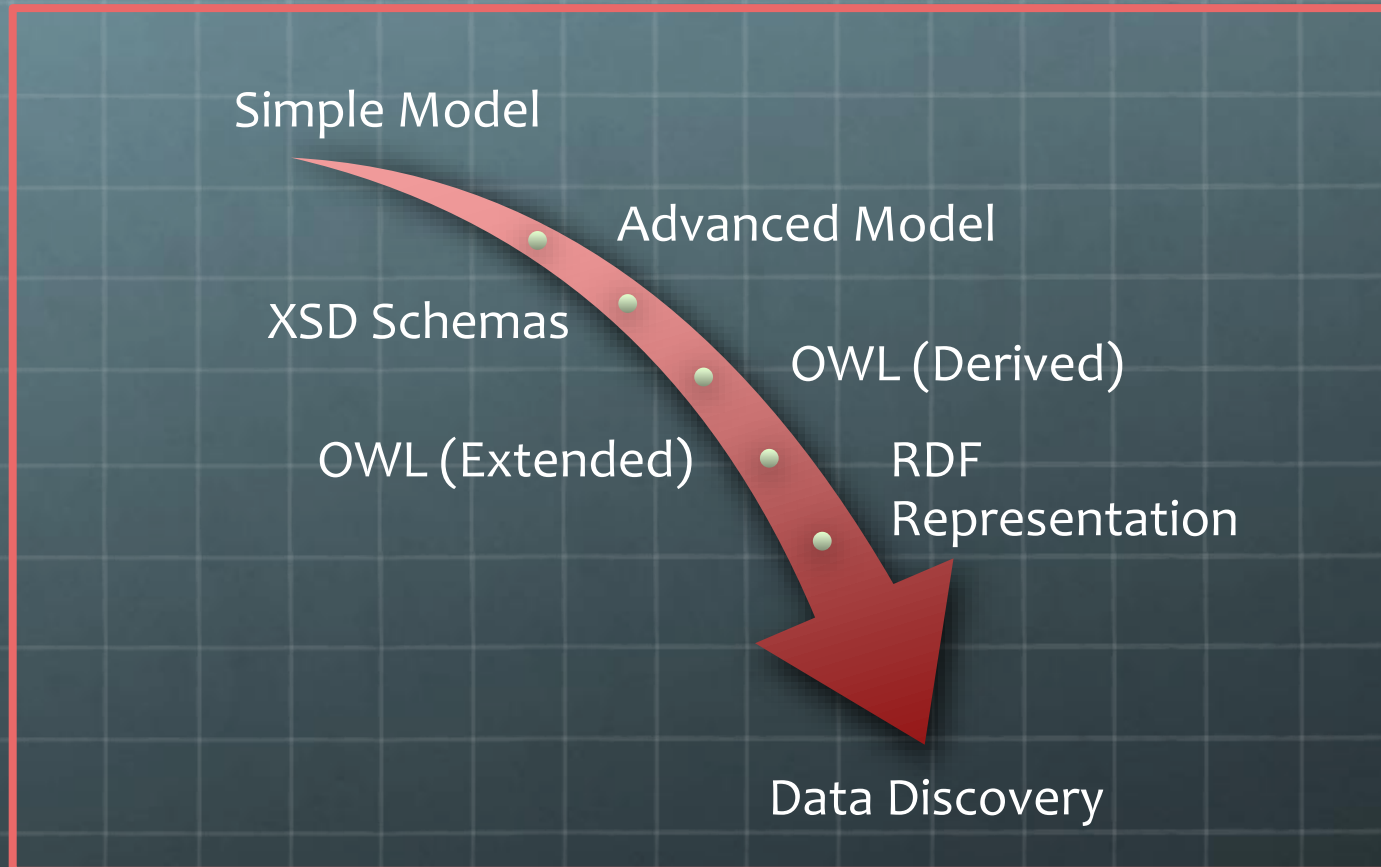
- Likewise there will be **simple and advanced studies**, be they longitudinal studies or meta-studies.



Driving the packages and views is a **framework**...

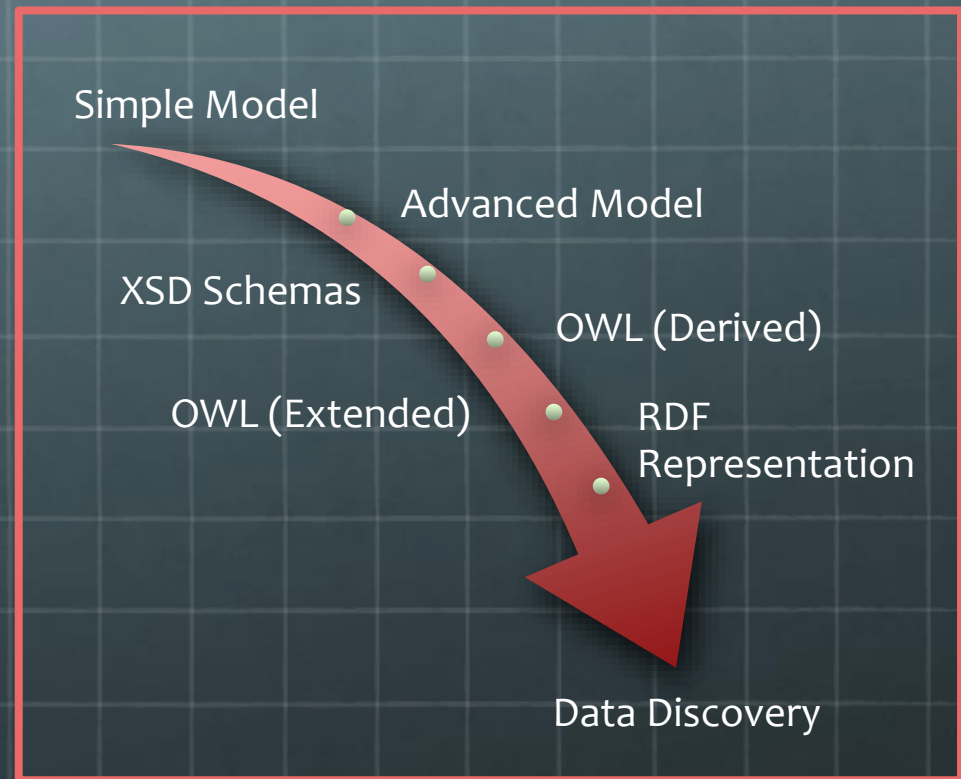


and a **Concept of Operations** (CONOPS)...



Data Discovery under this CONOPS is **semantic search** which “exposes” key elements from one or more packages in **views**...

- From these views first **XSD schemas** and then **OWL ontologies** are derived.
- As we shall see, there may be cause to **extend** OWL (Derived) after the fact.
- In Moving Forward codebooks and studies continue as **XML instances** of XSD schemas.
- OWL and **RDF instance-specific representations** of these XML codebook and study instances are generated.



Today we will propose three data discovery views (1 of 3)

View Domain	User Story
Data Elements and Common Data Elements (CDEs)	<ul style="list-style-type: none">• Measures sometimes change within surveys from one version or wave of a survey to the next.• Change happens.• Alternatively, sometimes different surveys use the same constructs.• As a data curator I want to aggregate data across participants who get different versions of the same survey or are enrolled in different surveys with similar constructs.• As a data curator I want to integrate data that repeats measures across waves in a longitudinal study.

DEs and CDEs

Today we will propose three data discovery views (2 of 3)

View Domain	User Story
DDI / RDF Discovery Vocabulary (Disco) and Beyond	<p>From Bosch et al (2012):</p> <ul style="list-style-type: none">• Researchers often want to know which studies exist for a specific country (e.g. France), time (e.g. 2005), and subject (e.g. election). Alternatively, sometimes different surveys use the same constructs.• Researchers want to be able to link data to similar concepts so they can better search other studies.• Researchers want to put data in context linking it to other research. <p>From Greenfield (2013):</p> <ul style="list-style-type: none">• As a data curator I want to support a research assets manager who needs to assess the current research value of a data set by linking it in real time to a citations database.


Disco


Today we will propose three data discovery views (3 of 3)

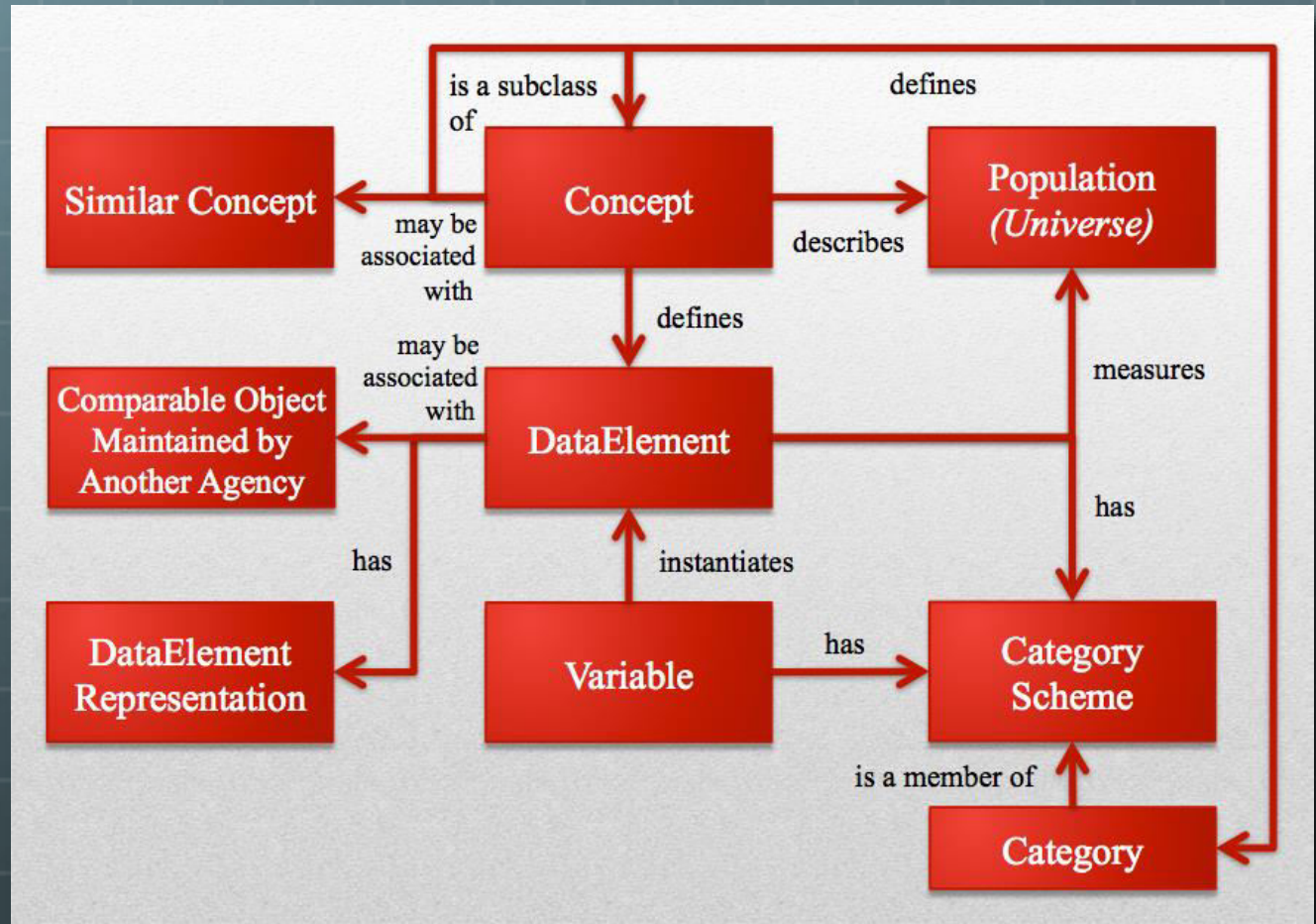
View Domain	User Story
Trajectories	<p>Trajectories refer to general models underlying path analysis in the social sciences and the representation of multifactorial traits in biomedical research.</p> <ul style="list-style-type: none">• As a data curator I want to support a data scientist who has developed a multifactorial path model of annual family income and wants to determine how well it fits a set of data.• As a data curator I want to support research on how a multifactorial model of BMI (Body Mass Index) might change over time as a result of a specific policy intervention such as the banning of trans fats by the American Food and Drug Administration (FDA).

Trajectories

Data Elements and Common Data Elements

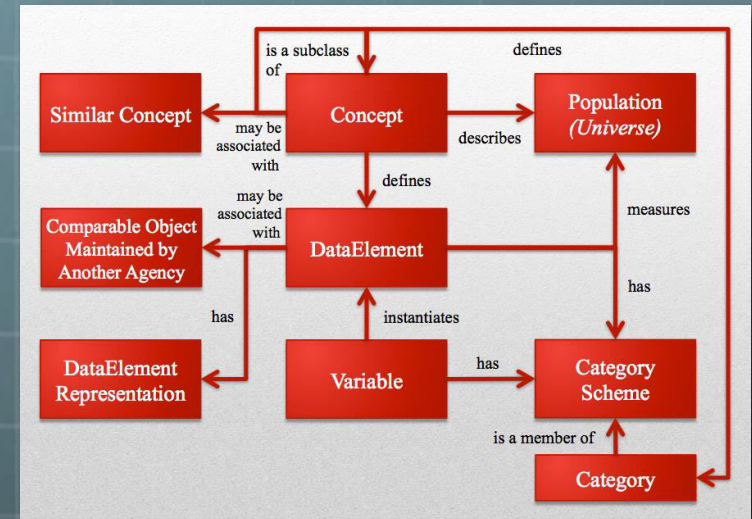
 **Data elements** were introduced in DDI 3.2.

 Before data elements there were just variables and their associated questions.



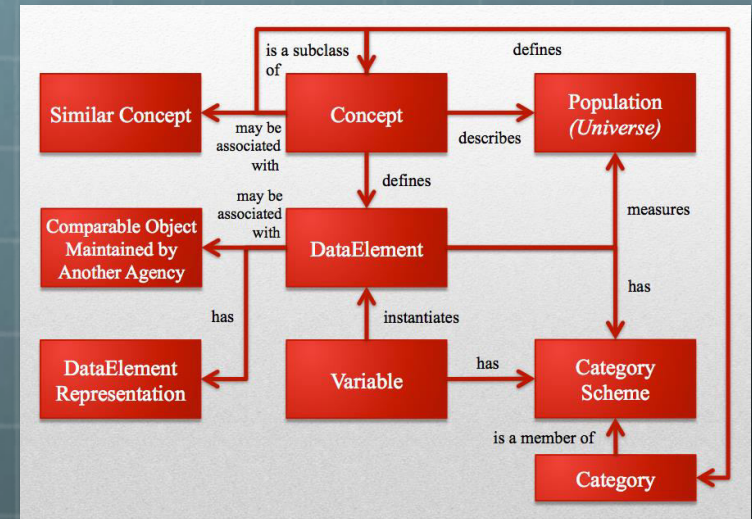
DEs and CDEs Moving Forward (1 of 4)

- Here data elements have **concepts** and there is a **one-to-many relationship between data elements and variables**.
- Also here, depending on whether a data element is associated with comparable objects maintained by other agencies, it may be a **common data element (CDE)**.
- Arguably CDEs were invented by NCI (US National Cancer Institute) at NIH (US National Institutes of Health).
- At the [NIH CDE Resource Portal](#) it is explained that CDEs were introduced to “improve the **comparability of data collected** in clinical research and/or with patient data in electronic health records”.



DEs and CDEs Moving Forward (2 of 4)

- 🌐 In terms of Moving Forward we might build, following the DDI CONOPS, a DDI **data normalization vocabulary**.
- 🌐 First we would derive a data normalization ontology from the DDI UML imagined here.
- 🌐 Perhaps, however, we will want to “tweak” this UML **specializing** it in a way that is not currently supported in DDI 3.2.
- 🌐 Specialization may be needed because across NIH **four subclasses of CDEs** have been identified...

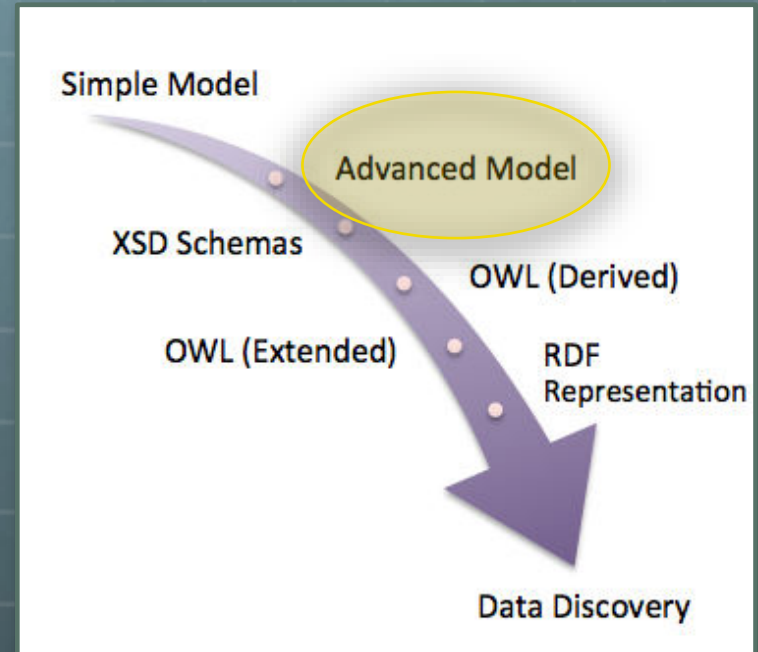


DEs and CDEs Moving Forward (3 of 4)

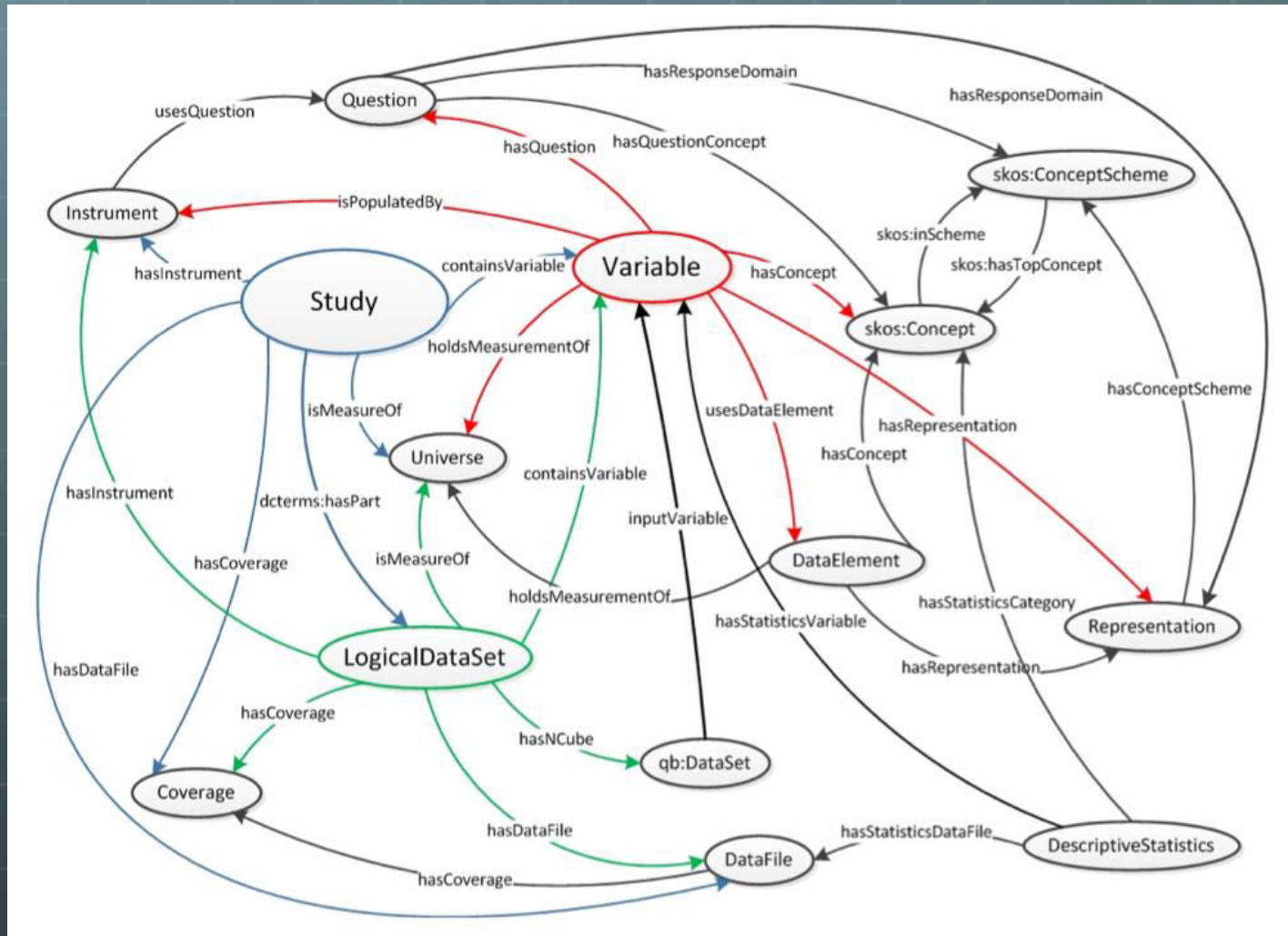
CDE Subclass	Description
Universal	CDEs that may be used in studies, regardless of the specific disease or condition of interest, e.g., demographic information of study subjects, medical history, certain patient-reported outcomes.
Domain Specific	CDEs that are designed and intended for use in studies of a particular topic, disease or condition, body system, or other classification , e.g., Parkinson's disease, Alzheimer's disease, diabetes, ophthalmology. Some domains are broadly applicable to a wide range of studies, while others are more useful in specific fields of clinical research.
Required	CDEs that are required or expected, as a matter of institutional policy (e.g., research funder or performer), to be collected for all subjects in studies of a particular type, e.g., NIH-funded studies of neurological disease, or NIH-funded studies of the genetics of eye disease.
Core	CDEs that are required or expected to be collected in particular classes of studies , e.g., any study of neurological disease or cancer, any genome-wide association study. Other, domain-specific common data elements may be suggested, expected, or required for collection, depending on the more specific focus of the study (e.g., Alzheimer's disease, ovarian cancer, genome-wide association study of diabetes).

DEs and CDEs Moving Forward (5 of 5)

- Specialization, however, creates issues...
- Any XML instance that a study populates in line with an XML schema derived from our “simple” UML data normalization model won’t identify CDEs in terms of their subclasses.
- Perhaps, however, we will want to “tweak” our simple UML data normalization model by “**advancing**” it with the addition of **CDE subclasses**.
- A **proliferation of advanced models** does not in principle lead to the “balkanization” of DDI but, because of the possibilities supported by the Moving Forward CONOPS, **more governance** may be required.



Disco (DDI-RDF Discovery Language) and Beyond – Part I: Architecture (1 of 4)

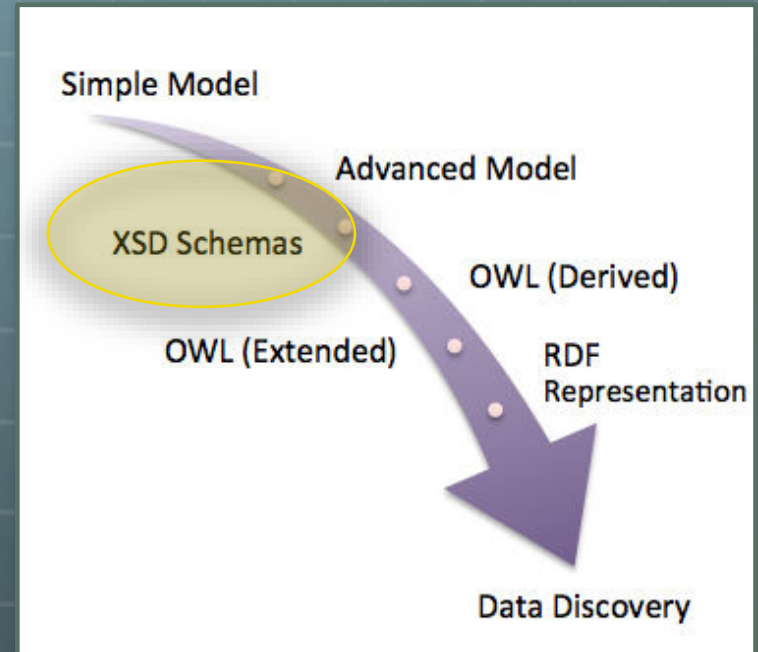


Here is a **directed graph** described by [Bosch et al \(2012\)](#) that is the basis for the Disco RDF discovery language

Disco and Beyond – Part I

(2 of 4)

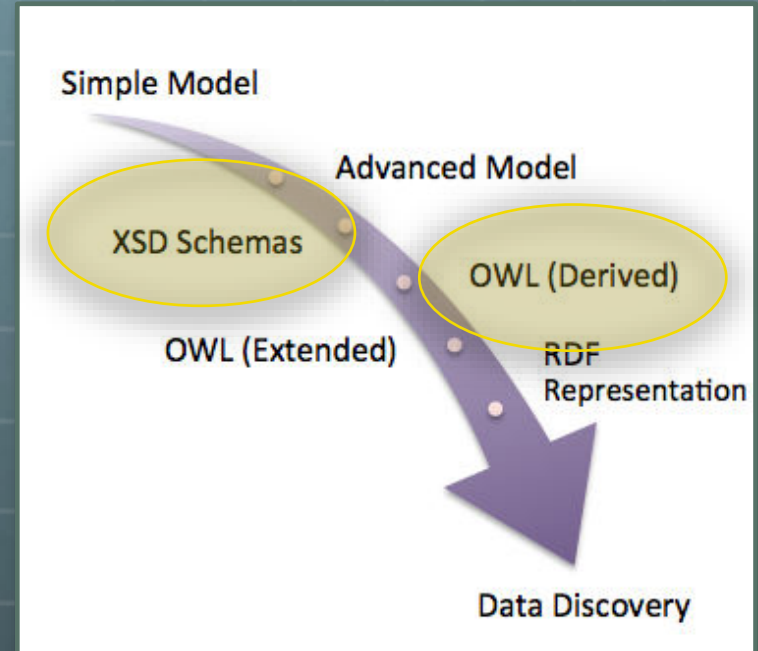
- Disco begins with an **XSD schema** based on this directed graph.
- [Bosch and Mathiak \(2011\)](#) and [Bosch and Mathiak \(2012\)](#) describe two approaches for **deriving OWL ontologies** from the Disco XSD.
- In terms of Moving Forward we note two things:
 - Disco doesn't begin with a view from either a simple or an advanced **UML model**.
 - The Disco directed graph / conceptual model includes a type of concept – **skos:Concept** – which is not part of DDI 3.2.



Disco and Beyond – Part I

(3 of 4)

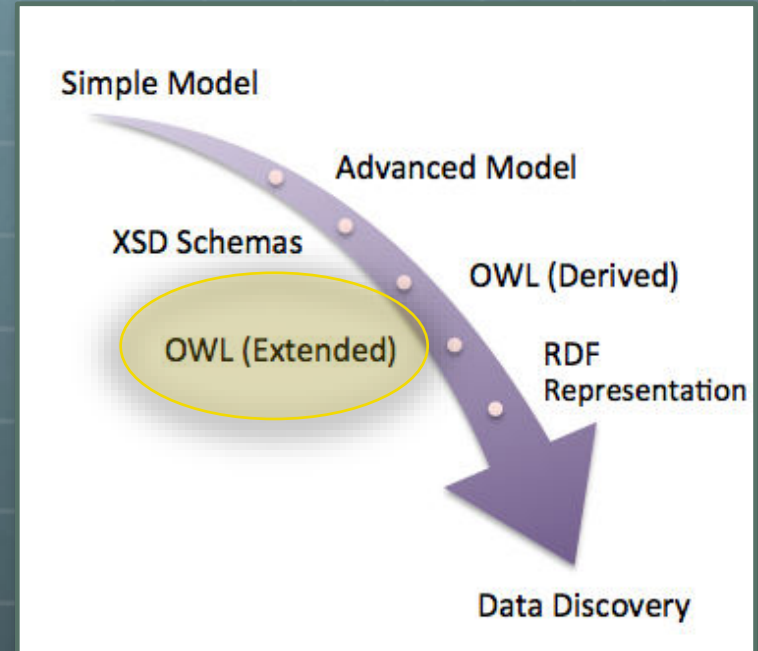
- Moving Forward, we will probably want to **derive the Disco XSD schema from UML XMI** although in principle it is possible to derive the Disco OWL ontology from UML XMI directly.
- A problem comes in because the **DDI specification doesn't support SKOS concepts**.
- SKOS organizes concepts around “broader”, “narrower” and “associative” relationships** and is well suited for describing statistical classifications and the standard thesaurus.
- Using SKOS, Disco can support **cohort identification** queries in which we search for populations that are more or less comparable on certain dimensions.



Disco and Beyond – Part I

(4 of 4)

- Since **SKOS integrates with OWL**, Moving Forward one approach for supporting SKOS concepts in DDI is to introduce them later on.
- “Later on” is after the Disco ontology is programmatically derived from DDI UML.
- In this approach OWL (Derived) can be **extended**.
- In their seminal work “Using OWL with SKOS” [Bechhofer and Miles \(2008\)](#) describe several models for **integrating a SKOS thesaurus with an OWL ontology**.



Disco and Beyond – Part II: Growing Disco to Support the Valuation of Study Variables (1 of 4)

- Imagine an instrument that uses your smartphone to take pictures that record **the food you eat**.
- Using the pictures as input, the instrument identifies food objects like the bread on your plate.
- The instrument records an annotation it places on the picture that marks the bread.
- It also records a measure of size based on the annotation.
- Finally it records the **nutritional value** of the piece of bread based on a lookup.



Disco and Beyond – Part II: Growing Disco to Support the Valuation of Study Variables (1 of 4)

Imagine
small
the

Using
instr
break

The
place
break

It all
the

Finally
the

Bread, egg

Serving size: 1 slice (5" x 3" x 1/2") (40g)

FOOD SUMMARY

Nutrition Facts

Serving Size 40 g

Amount Per Serving

Calories 113 Calories from Fat 22

% Daily Value*

Total Fat 2g 4%

Saturated Fat 1g 3%

Trans Fat

Cholesterol 20mg 7%

Sodium 197mg 8%

Total Carbohydrate 19g 6%

Dietary Fiber 1g 4%

Sugars 1g

Protein 4g

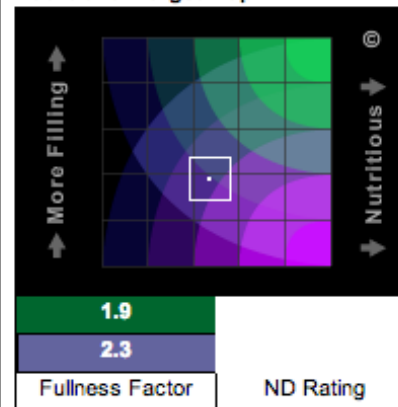
Vitamin A 2% • Vitamin C 0%

Calcium 4% • Iron 7%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

NutritionData.com

Nutritional Target Map



NutritionData's Opinion

? What is this?

Weight loss: ★★☆☆☆

Optimum health: ★★☆☆☆

Weight gain: ★★☆☆☆

The good: This food is a good source of Thiamin and Selenium.

Caloric Ratio Pyramid



Estimated Glycemic Load

11

0 250

Typical target total is 100/day or less

Inflammation Factor

-80

mildly inflammatory

- 0 +

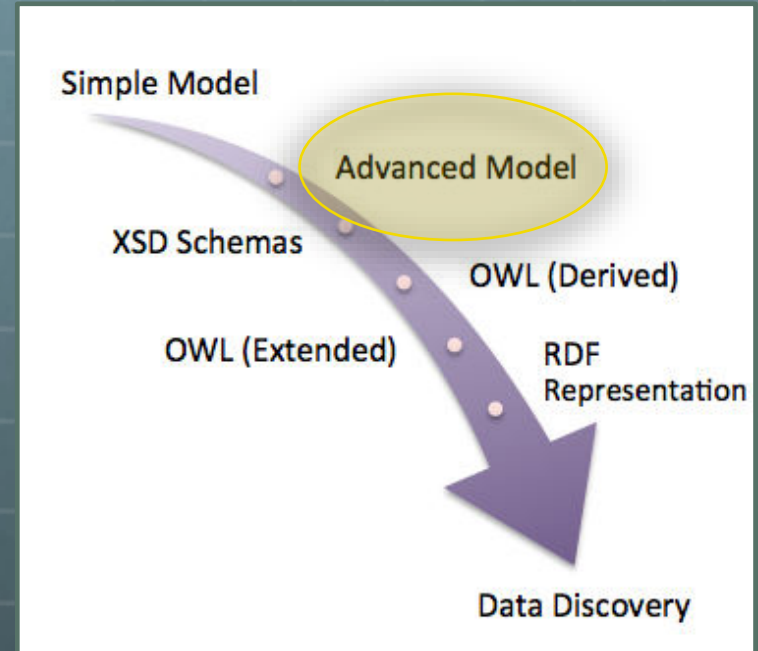
Typical target net is 50/day or higher



Disco and Beyond – Part II

(2 of 4)

- Eventually this information together with information about other foods on the plate finds its way into a dataset DDI describes with Variables and a LogicalDataSet.
- Let's imagine that Marie is its curator, and she just received a mission she chose to accept: determine the research value of this LogicalProduct.
- What Marie might do, under the circumstances, is **specialize** Disco...



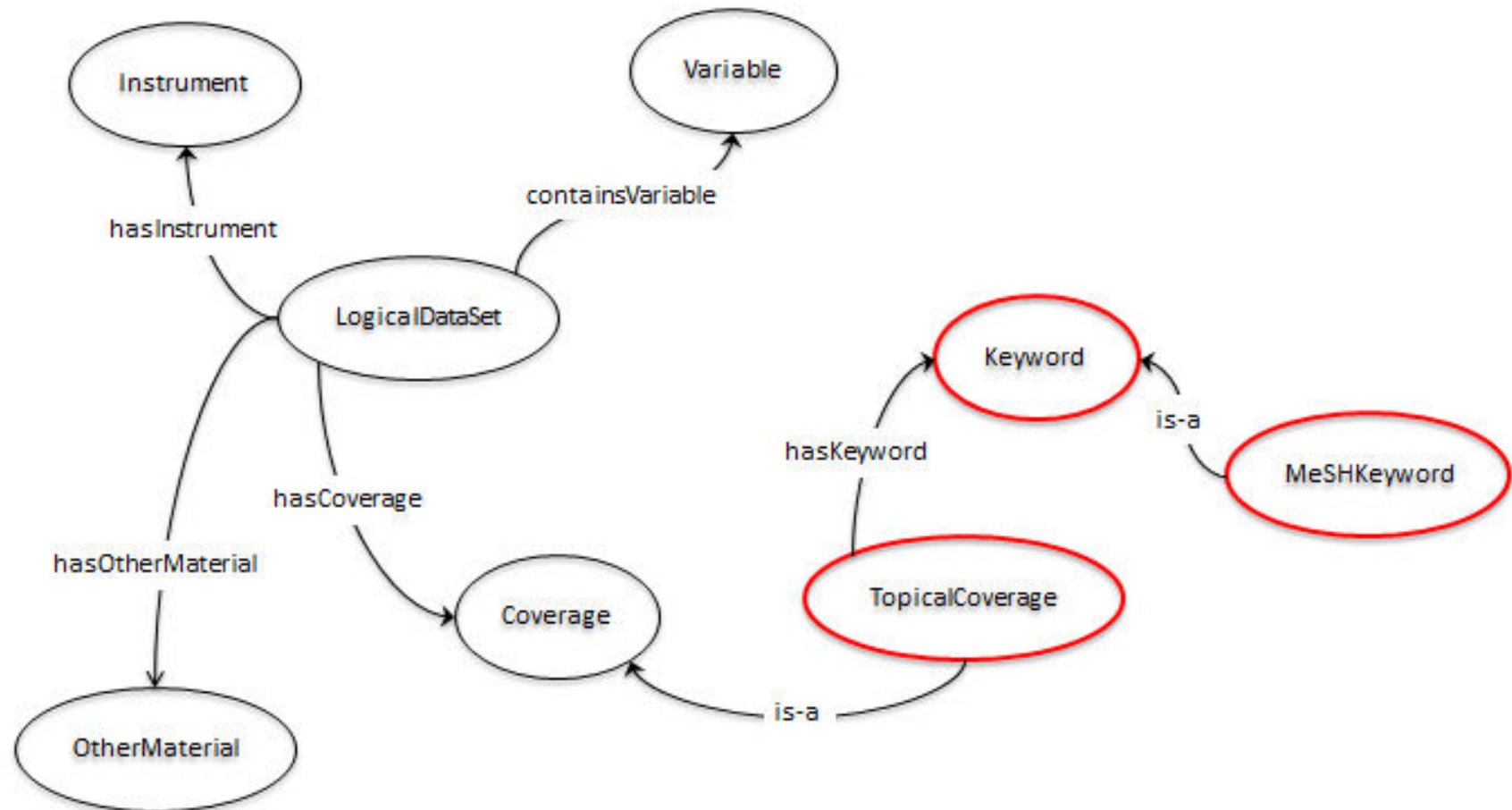
Disco and Beyond – Part II

(2 of 4)

Simple Model

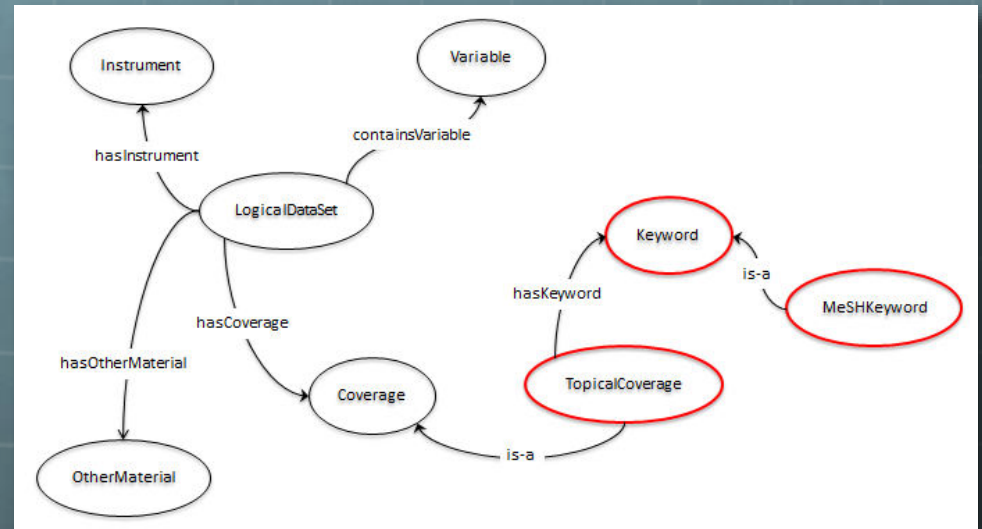
Advanced Model

VSD Schema

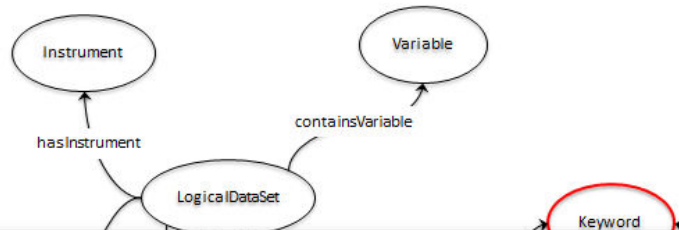



Disco and Beyond – Part II (3 of 4)

- Specializations are outlined in **red**.
- MeSH** is a taxonomy of Medical Subject Headers. **PubMed** consists of more than 23 million citations for biomedical literature. Those citations are indexed by MeSH.
- Finally, there is **Gene2MeSH**. Gene2MeSH “uses a statistical approach to reliably and automatically annotate genes with the concepts defined in MeSH, the National Library of Medicine’s controlled vocabulary for biology and medicine.”
- Back to Marie. She **tags** the nutritional LogicalDataSet with the MeSH term “Enzyme Activation” along with some other keywords.
- Next with the assistance of a software agent and a Gene2MeSH web service, she **retrieves** a list of at least 200 genes and the PubMed citations in which each is discussed.




Disco and Beyond – Part II (3 of 4)





Gene2MeSH – Gene Annotation with MeSH Terms



Limit Search by Organism:

☒ Substances only

history : [geneSymbol:m01b12.5 \(0\)](#) -> [descriptor:biocompatible materials \(5\)](#) -> [descriptor:metabolome \(1\)](#) -> [descriptor:biocompatible materials \(5\)](#) -> [descriptor:enzyme activation \(334\)](#)

334 genes found matching MeSH heading "enzyme activation"

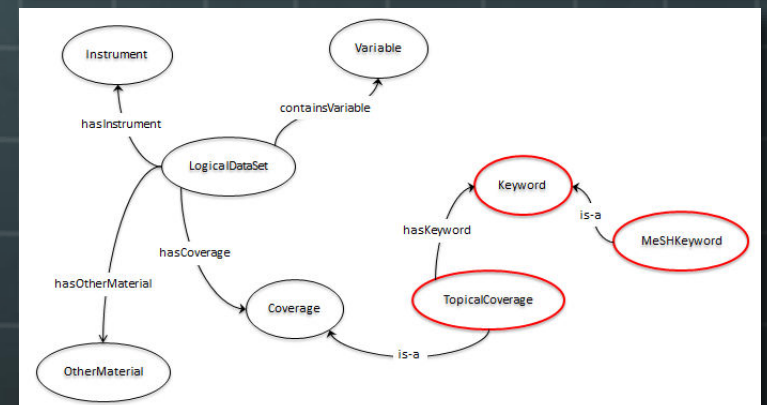
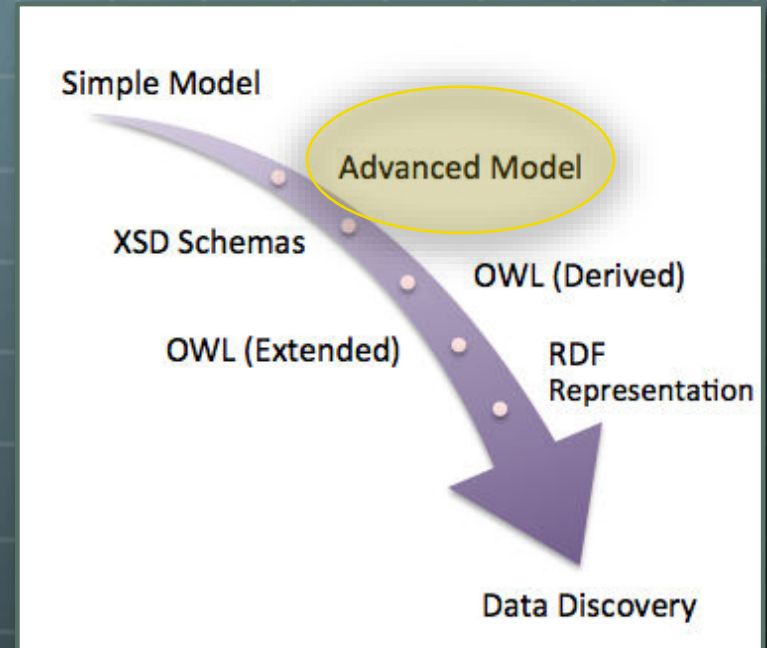
= lookup gene or MeSH heading at NCBI; = view interactions in MIM

Show All Columns ☐ | [download tab-delimited results](#)

Gene Symbol	MeSH Descriptor	TaxID	P-Value	MeSH Qualifier	Gene Description	PubMed Articles*
MAPK1	Enzyme Activation	9606	1.95e-285	-	mitogen-activated protein kinase 1	415
MAPK3	Enzyme Activation	9606	2.47e-219	-	mitogen-activated protein kinase 3	307
MAPK14	Enzyme Activation	9606	5.77e-181	-	mitogen-activated protein kinase 14	266
AKT1	Enzyme Activation	9606	3.10e-178	-	v-akt murine thymoma viral oncogene homolog 1	344
MAPK8	Enzyme Activation	9606	8.27e-177	-	mitogen-activated protein kinase 8	235
Mapk14	Enzyme Activation	10090	1.74e-148	-	mitogen-activated protein kinase 14	220
CASP3	Enzyme Activation	9606	2.54e-136	-	caspase 3, apoptosis-related cysteine peptidase	219
Mapk1	Enzyme Activation	10090	2.03e-134	-	mitogen-activated protein kinase 1	213
Mapk3	Enzyme Activation	10116	1.68e-130	-	mitogen activated protein kinase 3	183

Disco and Beyond – Part II (4 of 4)

- Next Marie and her software agent know to **filter the search** with a second keyword: *monosaccharide*.
- Now Marie and her agent have a count of ongoing genetic research into the digestion of bread.
- Marie stores the search URL that provides this account in **OtherMaterial**:
<http://gene2mesh.ncibi.org/index.php?term=descriptor%3Aenzyme+activation&taxid=ALL>
- Finally, **DDI** supports the real time valuation of study variables.



Hypothesis Discovery: What is it?

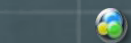
- Beyond common data elements, beyond cohort identification and beyond the DDI RDF vocabulary, what if we could **discover the research questions that studies are able to answer?**
- Note we don't say "the research questions that studies *intended* to answer".
- In a brave new world of research practiced by biology and the genomics revolution we divide samples into two parts:
 - In the first part we **train** programs to **explore** and learn the relationships in a data network.
 - In the second part we try to **confirm** what discovery has suggested to us.
- **Hypothesis generation**, however, depends on metadata that **annotates** the data elements that make up the **hypothesis discovery view**.

Hypothesis Discovery: What is it?

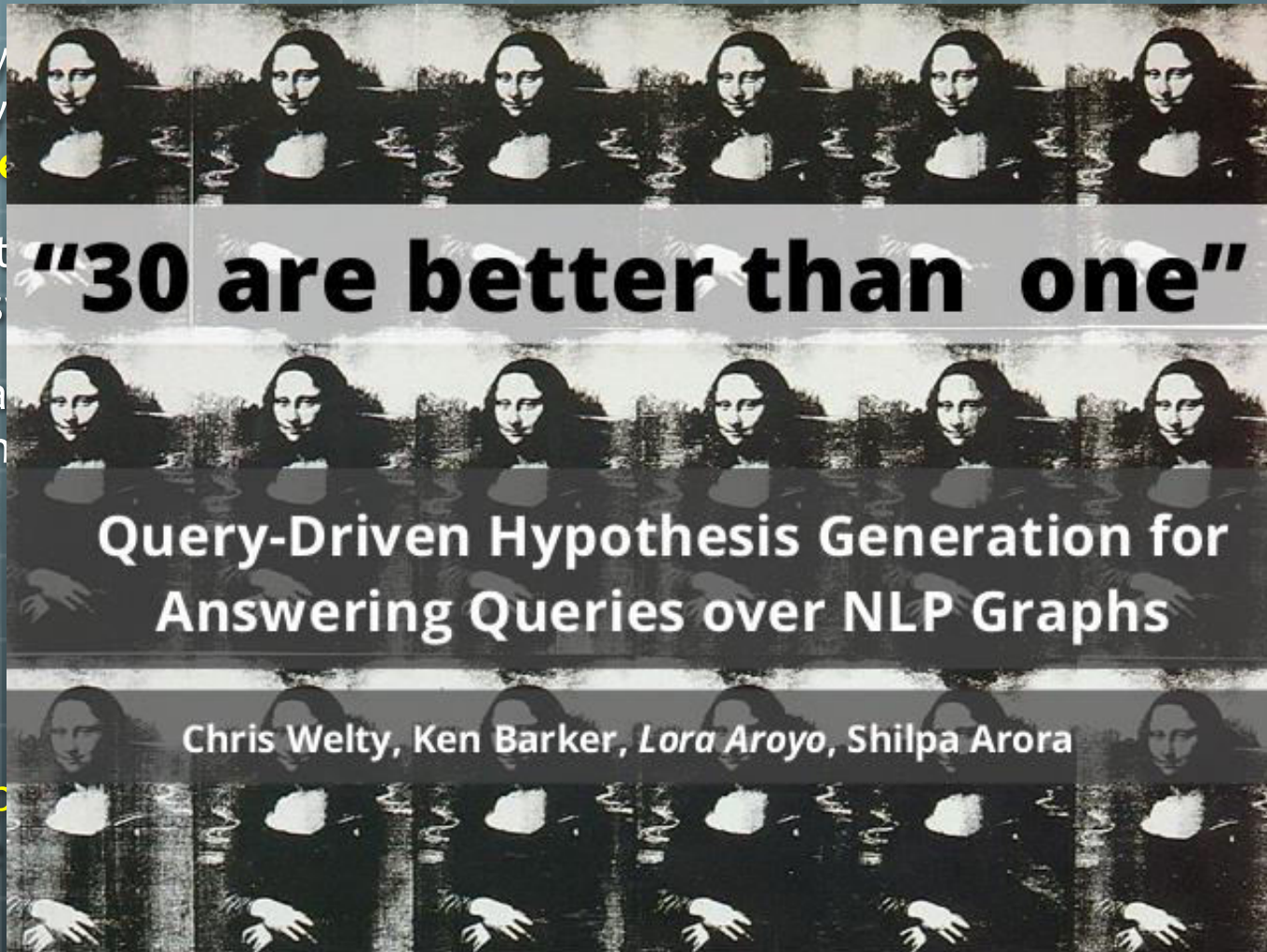
• Beyond
beyond
research

• Not
answers

• In a
general



• Hypothesis
the



and

to

suggested

notes

Hypothesis Discovery: What is it?



Crowd Truth

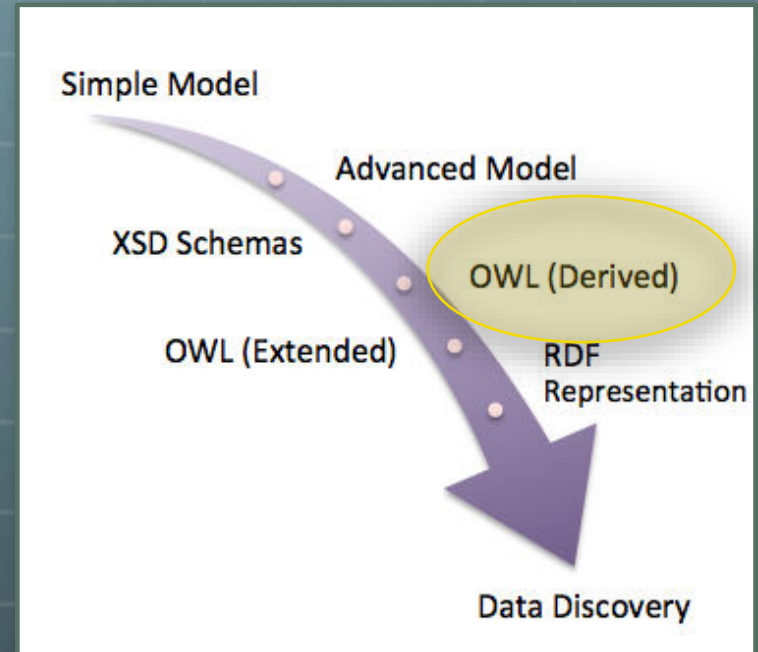
Harnessing Disagreement in Crowdsourcing

gathering gold standard annotations for relation extraction

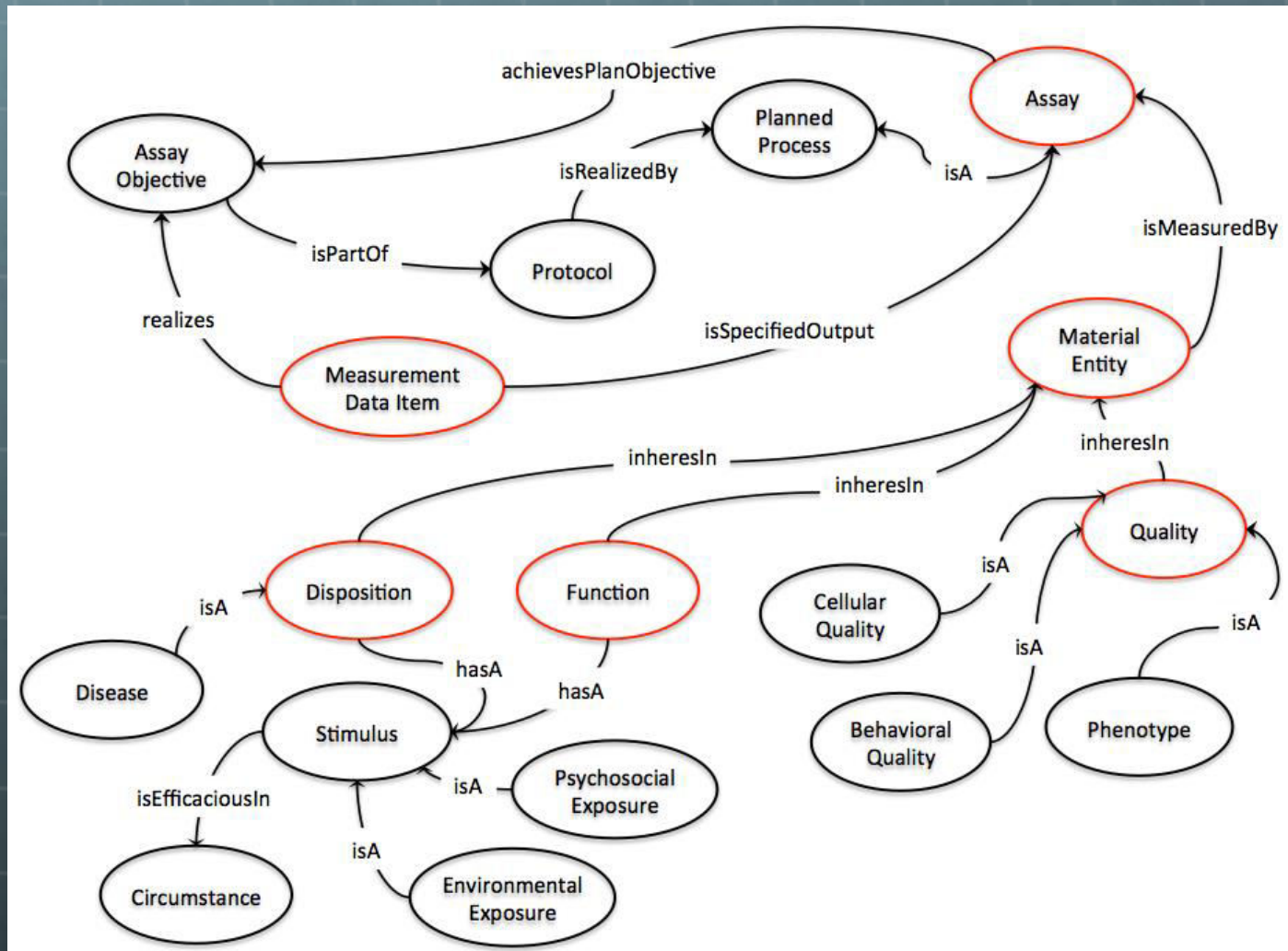
IBM Chris Wely Crowd Truth for Cognitive Computing Lora Aroyo **VU** UNIVERSITY AMSTERDAM

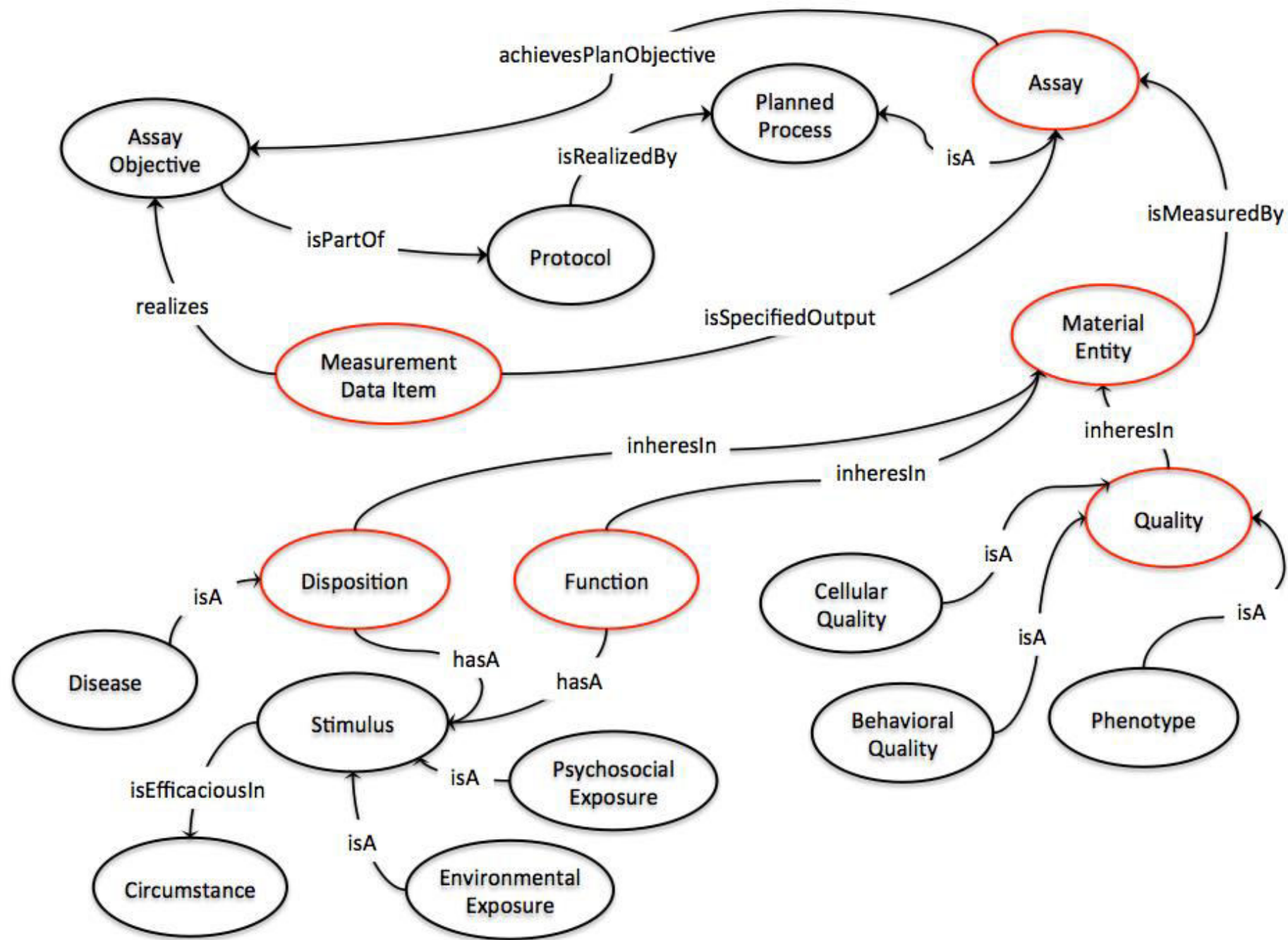
To support hypothesis generation we are proposing a *union*...

- 🌐 In this **union**, using the CONOPS, first DDI is projected into an OBO Foundry ontology called the Ontology of Biological Investigations (OBI).
- 🌐 In this projection OBI is specialized to include the very rich vocabulary of study objects and study processes DDI has developed.
- 🌐 And DDI gains a platform on which to describe **the succession of qualities and dispositions that humans and other entities undergo** as a result of a “treatment” in clinical studies or other “occurrences” such as environmental and psychosocial exposures in observational research.



After courting but before the union, OBI cuts the following figure...

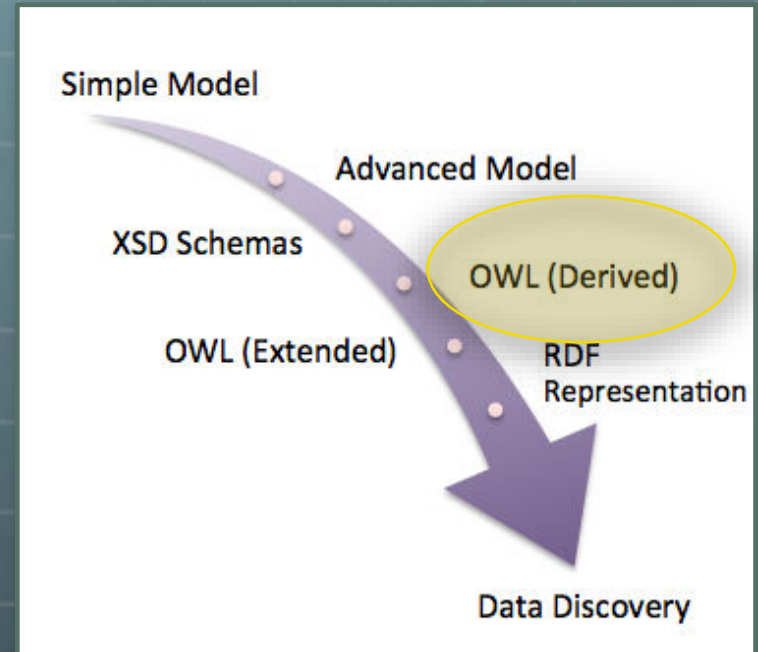




Upon the union OBI undergoes specialization...

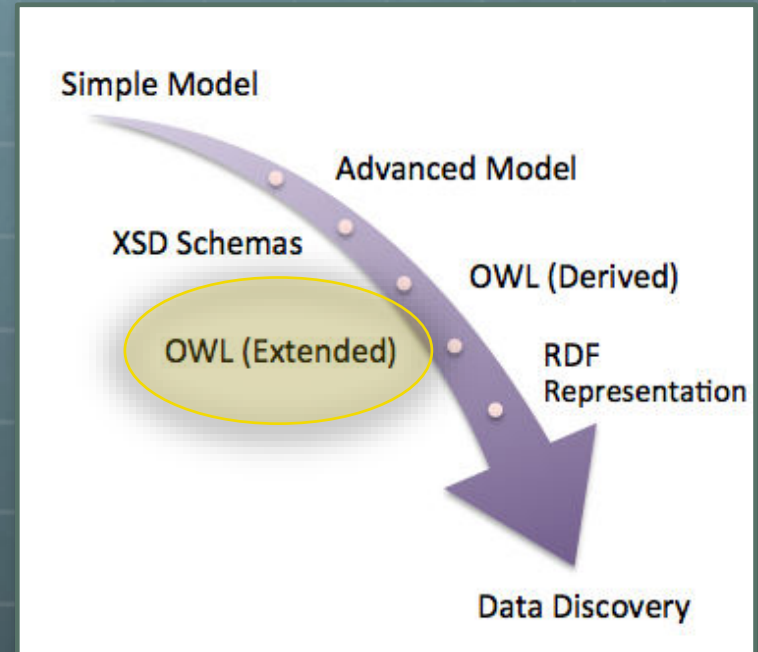
Using **DDI UML** and **certain rules**, in the future it will be possible to **derive** a DDI infused **OBI-based OWL** to describe, for example, a simple or an advanced survey:

- A **wave entity** might be added to the protocol.
- An **instrument entity** might be added to arrange assays.
- The **DataElement entity** might be introduced in order to account for successive versions of an assay and so forth.
- One set of specializations might grow OBI into a template suited for describing observational studies of behavior.
- Another set of specializations might enable this figure to describe clinical research.



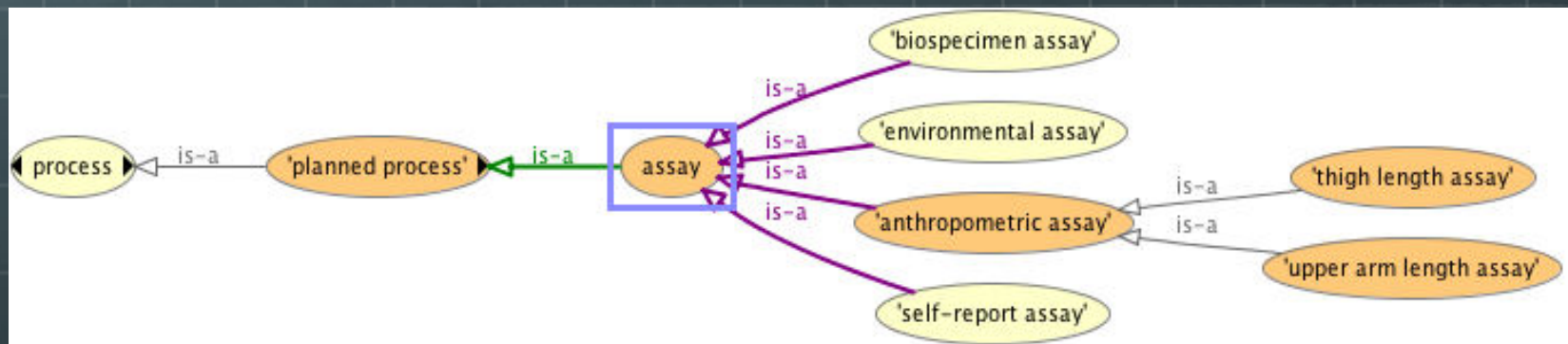
Next, using a DDI XML study instance, it will be possible...

- ...to complete this ontology for an actual study, filling in the specific measurements, qualities, dispositions and triggers that fall within its purview.
- Instantiating OWL Extended** using a DDI XML instance will require **leveraging study-specific DDI controlled vocabularies (CVS)**.
- Here is an example...



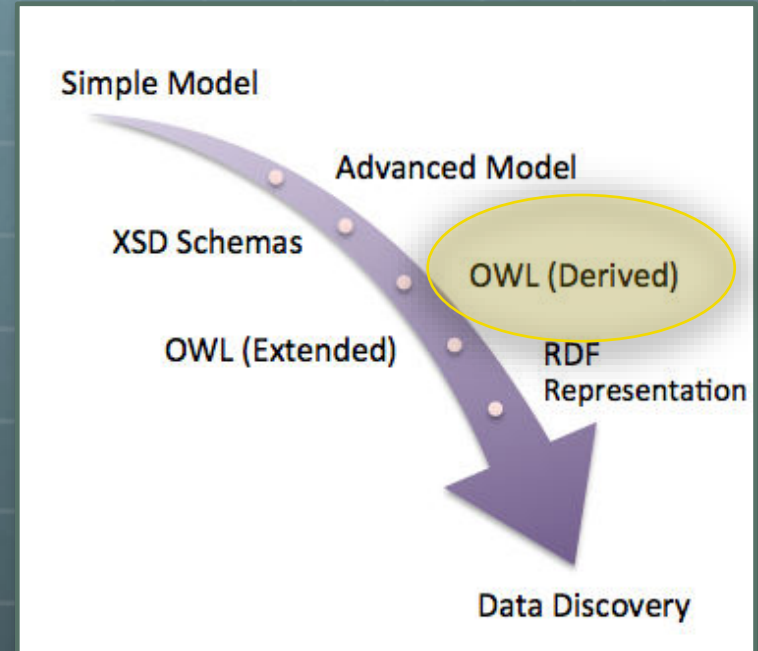
Here we subclass the OBI assay using a study-specific DDI controlled vocabulary...

- An “assay” is a planned process with the objective of producing information.
- An example of an assay might be a question an interviewer plans to ask, a blood draw a nurse plans to take or an anthropometric characteristic like the circumference of a child’s head that an interviewer plans to measure in line with the protocol.
- From this discussion we know that in a specific study the **assay entity has subclasses:**



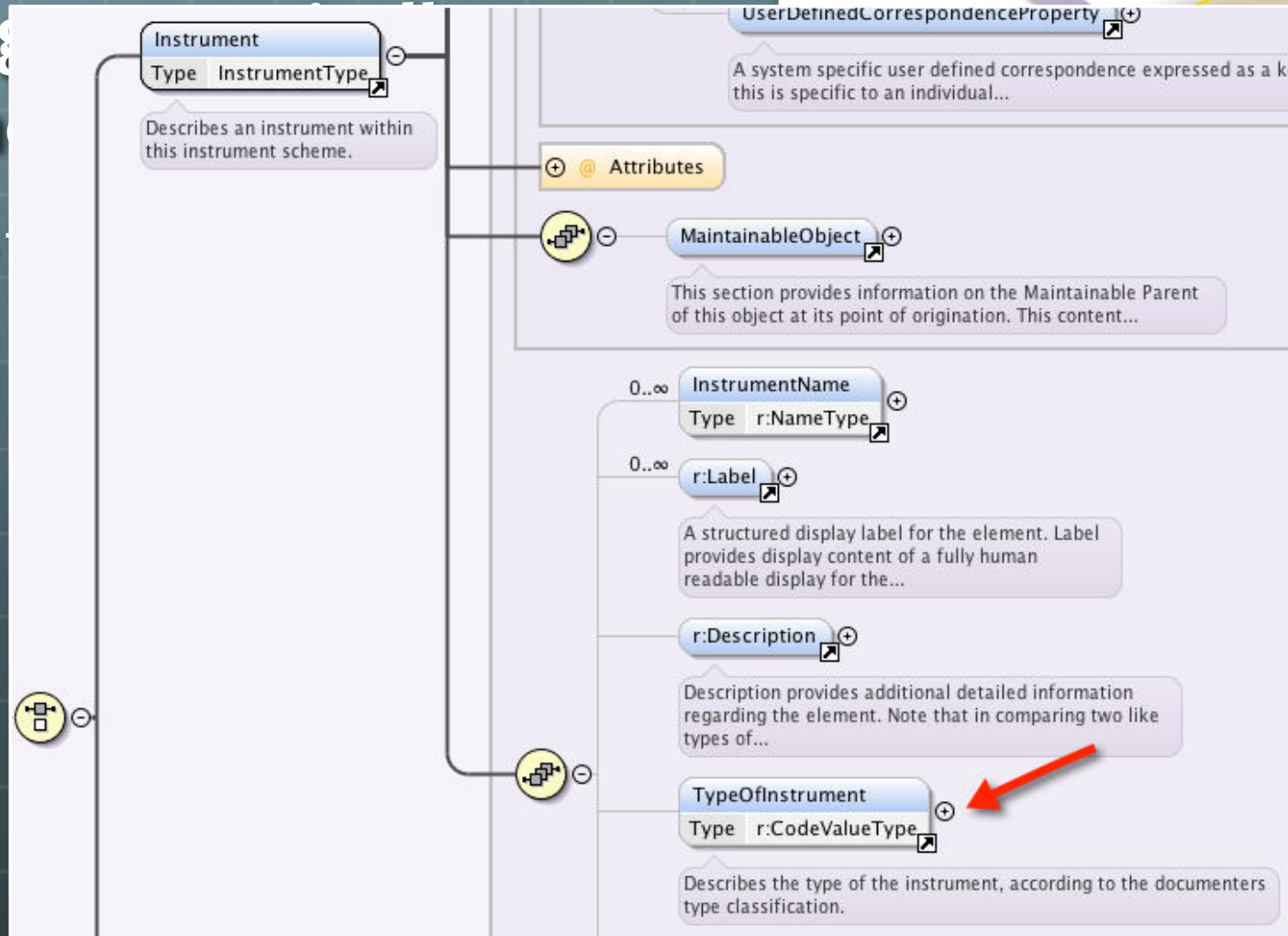
Appropriate subclasses for a specific study can be programmatically generated...

🌐 ... from CVs like this:



Appropriate subclasses for a specific study can be

prog
gen



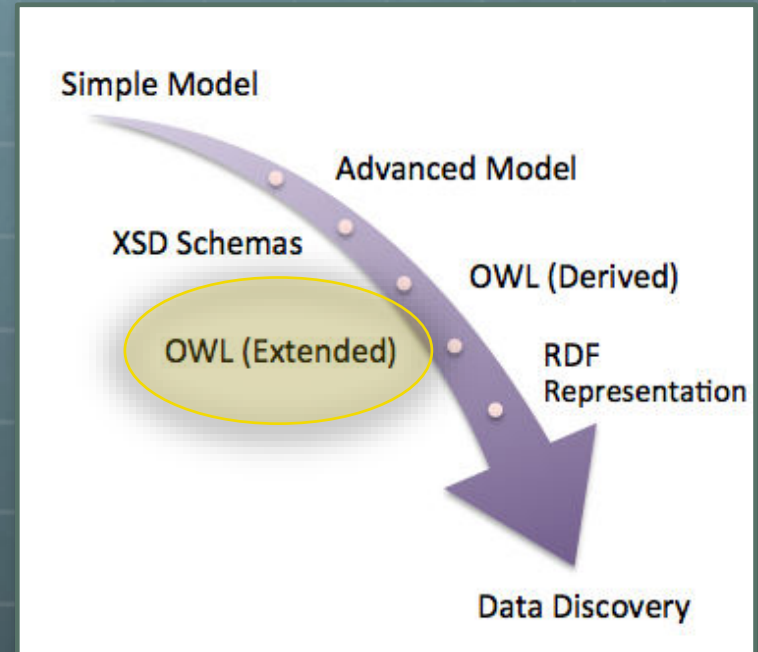
Simple Model

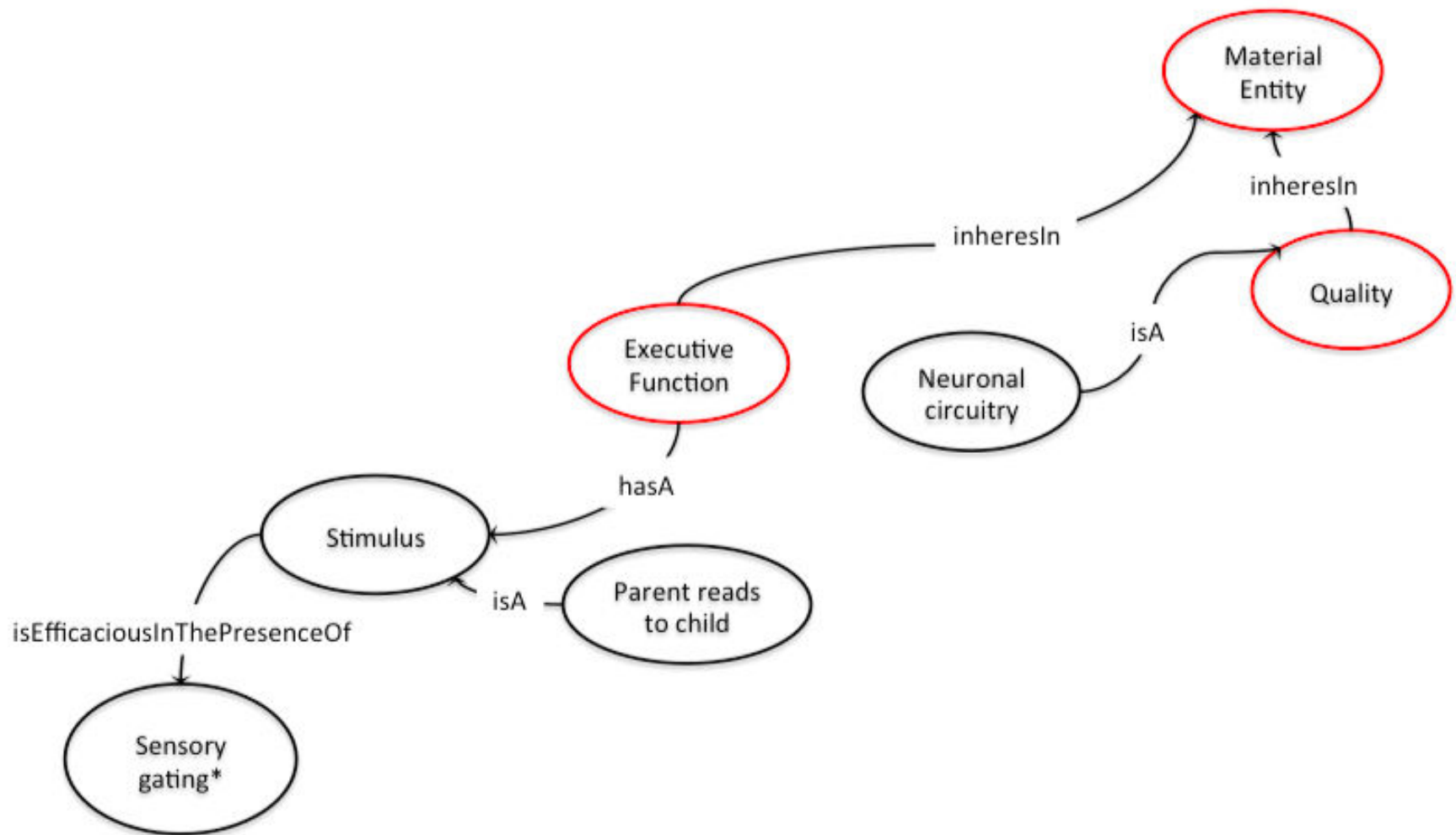
Advanced Model

derived)
presentation
discovery

However, in the end , hypothesis generation requires that we annotate DDI data elements...

- For instance, in OBI we can in principle make statements like this:
the executive function is stimulated by certain psychosocial exposures under certain circumstances as measured by an assessment, and humans who test well for executive function have certain neuronal qualities.
- For now this seems to require a curator to **tag** an instance of DDI data elements **with OBI concepts and relationships**.
- Eventually annotation will be machine-assisted but right now it is not...

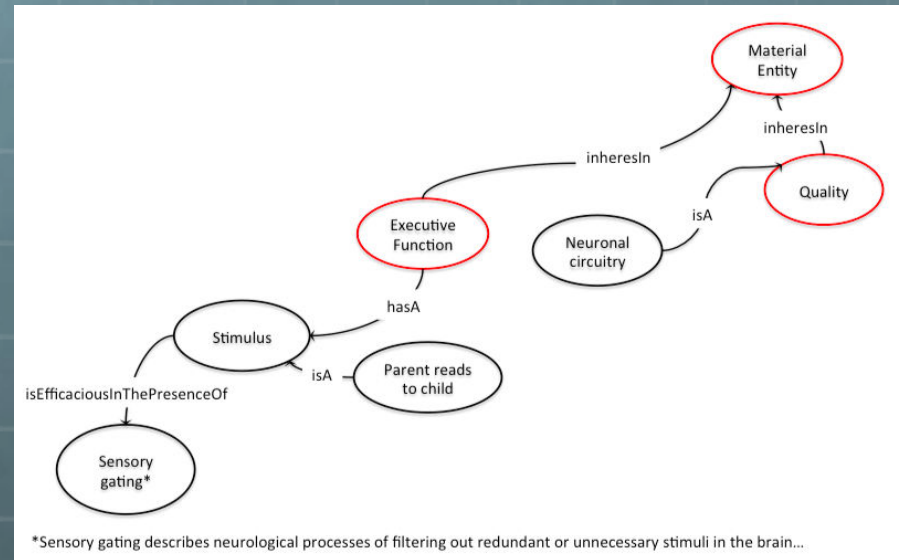




*Sensory gating describes neurological processes of filtering out redundant or unnecessary stimuli in the brain...

Towards a Hypothesis Vocabulary

- Basically OBI provides us with a biomedical research hypothesis template that we can instantiate with data elements
- We can also grow the template so it can express other hypothesis types
- OBI lends itself to many hypothesis views














Acknowledgements

Over the past several years the National Children's Study (NCS), its Director Dr. Steven Hirschfeld and the CIO David Songco have encouraged many of us to think broadly and creatively about terminology, **ontology** and study metadata. Many of the views described here originate in ongoing work that Booz Allen Hamilton and other NCS contractors are currently engaged in.

More specifically, this document draws from the work of two mentors:

-  Sophia Kuan is a colleague and an informaticist who over the years has led the development of a number of software systems that automate the annotation of data with metadata and semantics.
-  Mathias Brochhausen is a member of the faculty of the Biomedical Informatics Division of the College of Medicine at the University of Arkansas Medical Sciences and a contributor to the Basic Formal Ontology (BFO) Project. He has helped me grow my knowledge of the Open Biological and Biomedical Ontologies (OBO).

References

-  Bechhofer, Sean, and Alistair Miles. (2008). Using OWL and SKOS, [Online], Available: <http://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html>.
-  Bosch, Thomas, and Brigitte Mathiak. (2011). Generic Multilevel Approach Designing Domain Ontologies based on XML Schemas. Workshop Ontologies Come of Age in the Semantic Web, 2011, 1-12.
-  Bosch, Thomas, and Brigitte Mathiak. (2012). XSLT Transformation Generating OWL Ontologies Automatically Based on XML Schemas. The 6th International Conference for Internet Technology and Secured Transactions, 2012, 660-667.
-  Bosch, Thomas, Richard Cyganiak, Joachim Wackerow and Benjamin Zepilko. (2012). Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioral, and Economic Sciences. Proc. Int'l Conf. on Dublin Core and Metadata Applications, 2012.
-  Fisher, Justin. (2013). Dispositions, Conditionals and Auspicious Circumstances. Philosophical Studies 164 (2)443-464.

References (continued)

-  Greenfield, Jay. (2013). The Paradata Information Model. North American Data Documentation Conference (NADDI), 2013.
<http://kuscholarworks.ku.edu/dspace/handle/1808/11065>