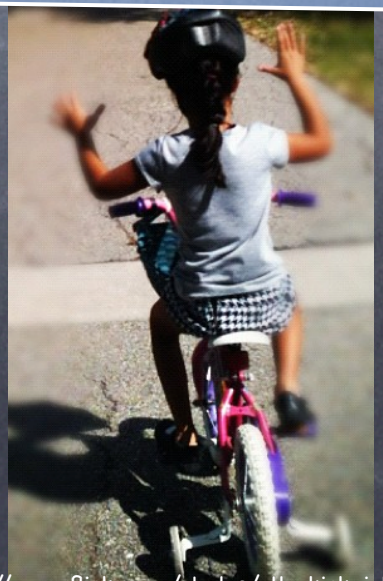
Evaluation in Semantic Web Research

Abraham Bernstein, University of Zurich, Switzerland
Natasha Noy, Stanford University, US

What's wrong with this picture?

How can we use Linked data to make movie recommend ations?

"Look, Ma, no hands!"
or
"We built an ontology editor"



http://www.flickr.com/photos/ultrakickgirl/8109854279

If we publish our papers in RDF, the world will be a better place

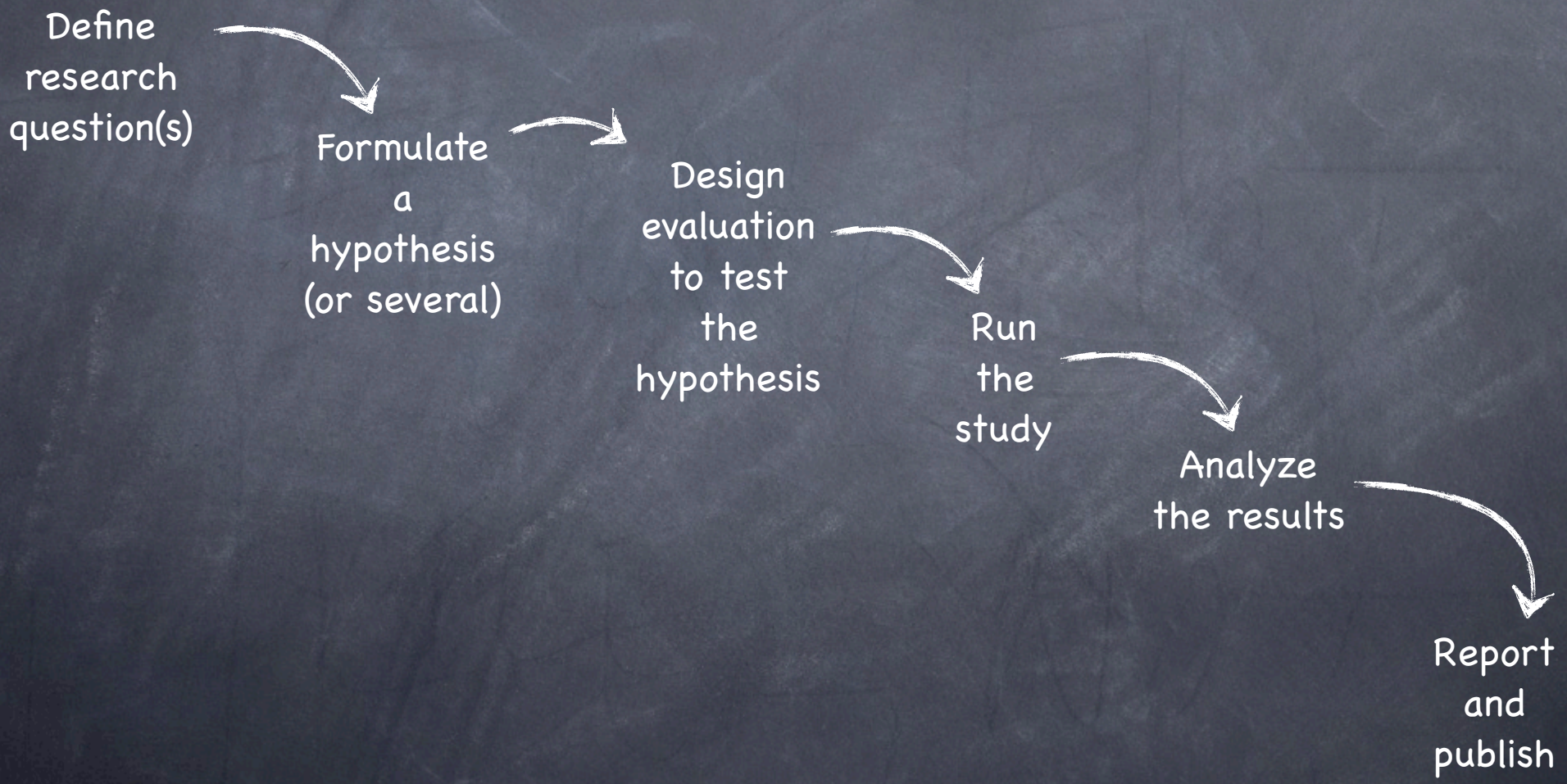
What is Scientific Research?

Theory

Observation

"The World"

Anatomy of a research project (or a PhD thesis)



What is a Research Question?

Research question

Determines where and what kind of research the writer will be doing

Identifies the specific objectives the study or the paper will address

NOTE:

there is a strong interaction between the RQ and the type of study (Qualitative, Quantitative, or Mixed)

What is a Research Question?

Qualitative Template:

_____ (How or what) is the _____ ("story for" for narrative research; "meaning of" the phenomenon for phenomenology; "theory that explains the process of" for grounded theory; "culture-sharing pattern" for ethnography; "issue" in the "case" for case study) of _____ (central phenomenon) for _____ (participants) at _____ (research site).

Quantitative Template:

Does _____ (name the theory) explain the relationship between _____ (independent variable) and _____ (dependent variable), controlling for the effects of _____ (control variable)?

Creswell's (2009)

RQs and their Validity

External Validity

Theory

Internal Validity

Cause and Effect

Construct

Observation

Construct Validity

Construct Validity

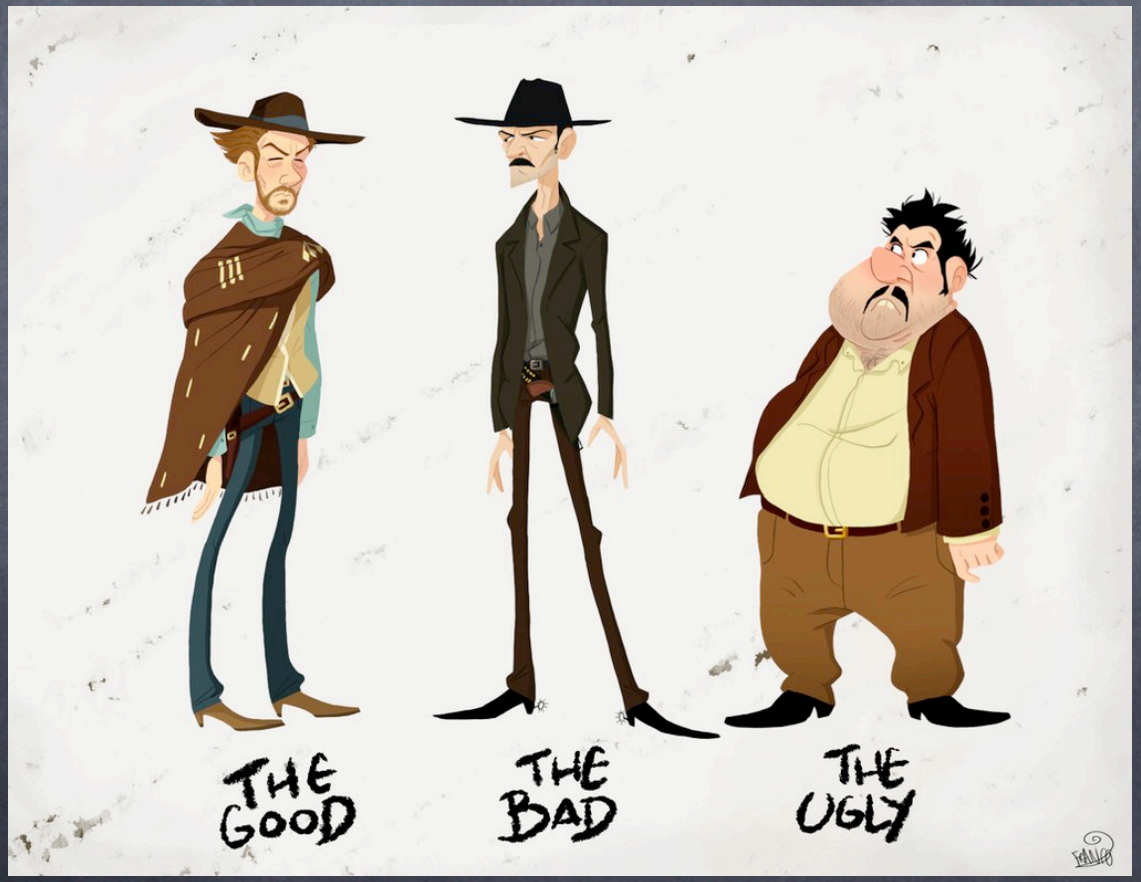
"The World"

Treatment

RQ and different methods...

- Feasibility study
- Case study (aka Demonstrator)
- Comparative study / Benchmark
- Observational Study [a.k.a. Ethnography]
- Experiment
- Literature survey (incl. Meta-Analysis)
- Formal Model
- Simulation

The Actual Semantic Web Research Projects

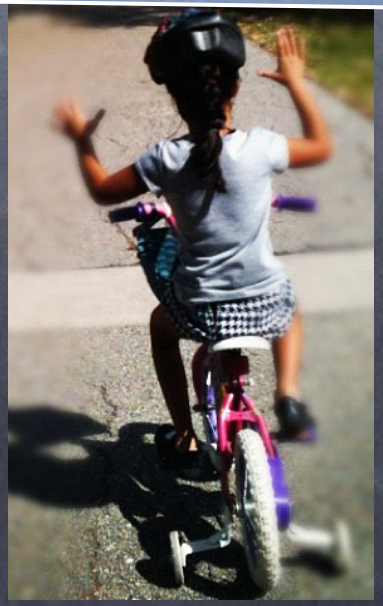


http://www.deviantart.com/art/The-Good-The-Bad-and-The-Ugly-320626352

What's wrong with this picture?

How can we use Linked data to solve this problem?

"Look, Ma, no hands!"
or
"We built a system"



We will put everything in RDF and the world will be a better place

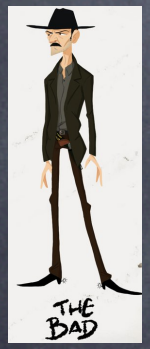
"My Semantic Web system is better than your Semantic Web system"

What does "better" mean and how do you measure it?

Improve performance of a system

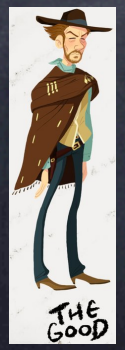
- Hypothesis: "My reasoner is very fast and efficient"

Nothing to measure here. You cannot falsify the hypothesis



- Hypothesis: "My reasoner is faster than a reasoner X on one specific ontology"

External validity:
how important is this one ontology?



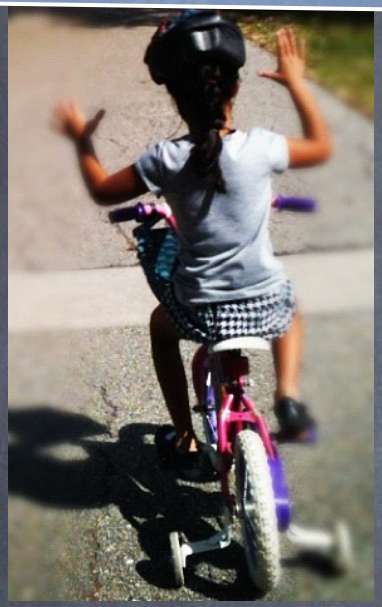
- Hypothesis: "Using X will improve the efficiency of reasoning on the class of languages Y, compared to the current state of the art"

Evaluation Methods (Construct validity)

- Run your reasoner on large ontologies in the class Y
- Run the best existing reasoner that is designed for the class of ontologies that you consider
  - Run experiments to understand why it is better
- Compare the performance of your reasoner to the existing one(s)
  - gold standard, published benchmark

- Make sure that your hypothesis is task-specific: X is better than Y for task Z (or in a context C)
- Maybe design a hierarchy of hypotheses (unfortunately, not very common in CS/AI/SemWeb)
- Make sure that your evaluation is designed to compare X and Y in the context C or for task Z
- When you report results and reach conclusions, do not over-generalize. The conclusions are valid only for these tasks/context.

"Look, Ma, no hands!"
or
"We built a system"

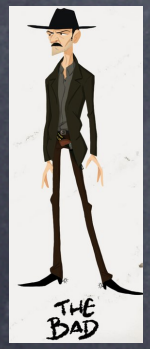


Where is a scientific problem in building an application?

Developing an application

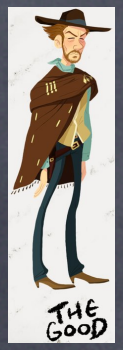
- "I have developed a tool. It works!"

Nothing to measure.
No way it can fail.

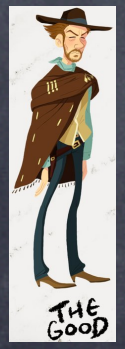


- Hypothesis: "My application works perfectly for displaying one ontology and if we ask users questions about this ontology, they can use the tool effectively"

External validity:
how important is this one ontology?



- Hypothesis: "Domain experts can use our system effectively to accomplish a task X (e.g., map between large ontologies)"



- Hypothesis: "Domain experts can use our system more effectively than another system Z to accomplish a task X (e.g., map between large ontologies)"

Evaluation Methods

- Experiment
  - Usability study
  - Successful completion of tasks
- Case study
- Comparative study or benchmark

- Just developing a system is not a research contribution in itself.
- Make sure that your hypothesis is task-specific: X is good for task Z, or in context C, or for users of type T
- Maybe design a hierarch
(unfortunate
- Make sure that your evaluation is designed to compare X and Y task Z, or in context C, or for users of type T
- When you report results and reach conclusions, do not over-generalize. The conclusions are valid only for these tasks/context/user types.

We have seen this before!

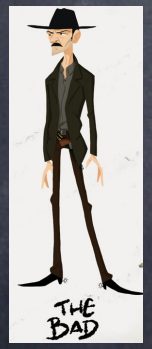
We will put everything in RDF and the world will be a better place

Solution in search of a problem: who should care if you succeed?

Convert unstructured data into RDF or Linked Data

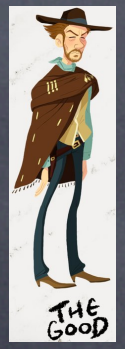
- "I will convert a corpus of abstracts into RDF"

So what? Who cares?

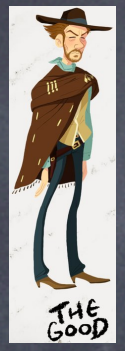


- Hypothesis: "My conversion process produces better linked data than conversion process X"

So what? Who cares?

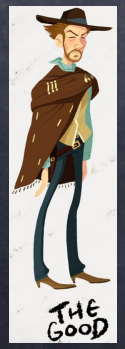


- Hypothesis: "Using extracted linked data will improve search performance on the corpus"



- Hypothesis: "Using extracted linked data will enable advanced querying that was not possible before"

Make sure there is somebody who actually wants to perform those queries on the corpus



- (Auxiliary hypothesis -- should not be the only one): "My method for extracting structured data has better accuracy/coverage/precision/recall/etc. than the state of the art."

Evaluation Methods

- Depending on the task, show that the task can be performed better if you use your structured data
- Compare the quality of the data to that produced by other algorithms

How can we use Linked data to solve this problem?

"Using our hammer for every nail"
or
does the use of semantic web technology actually improve anything?

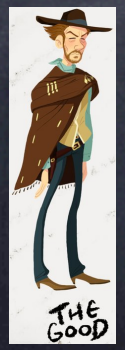
Novel Solution to an Old Problem

- "We will use Semantic Web technology to make movie recommendations"

Good for you, but who cares?

Key difference:

Our goal is to improve the efficiency of a task that someone cares about. Not to use SW technology per se



- Hypothesis: "We will improve the efficiency of social-network monitoring by using SW technology/improve the quality of recommendations"

Evaluation Methods

- Compare the accuracy of recommendations with and without the linked data component
- Compare the accuracy of your system to an existing non-Semantic Web system

- If the LD component is completely integral to your system and you cannot take it out, you will need to compare to another system
- You may need to compare to the state of the art to convince non-SemWebbies that your method has any value
- Make sure the metrics, the users, and the datasets are comparable

What (human) languages are ontologies published in?

http://www.flickr.com/photos/rachelfordjames/2833420148

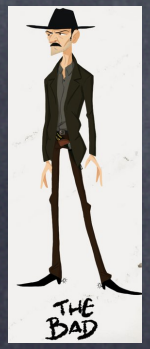
"All science is either physics or stamp collecting"
Lord Rutherford

Stamp collection for the sake of stamp collection

Improve performance of a system

- "We will create a set of features that ontologies have and will describe ontologies from our catalog according to these features"

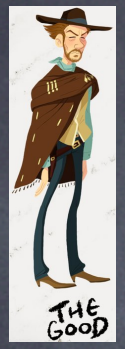
This is not a hypothesis; this is a research plan



- "We will create a set of features that ontologies have and will describe ontologies from representative ontology repositories according to these features."

Why would this information be useful?
What will drive you selection of features?
Imagination?

Improve performance of a system



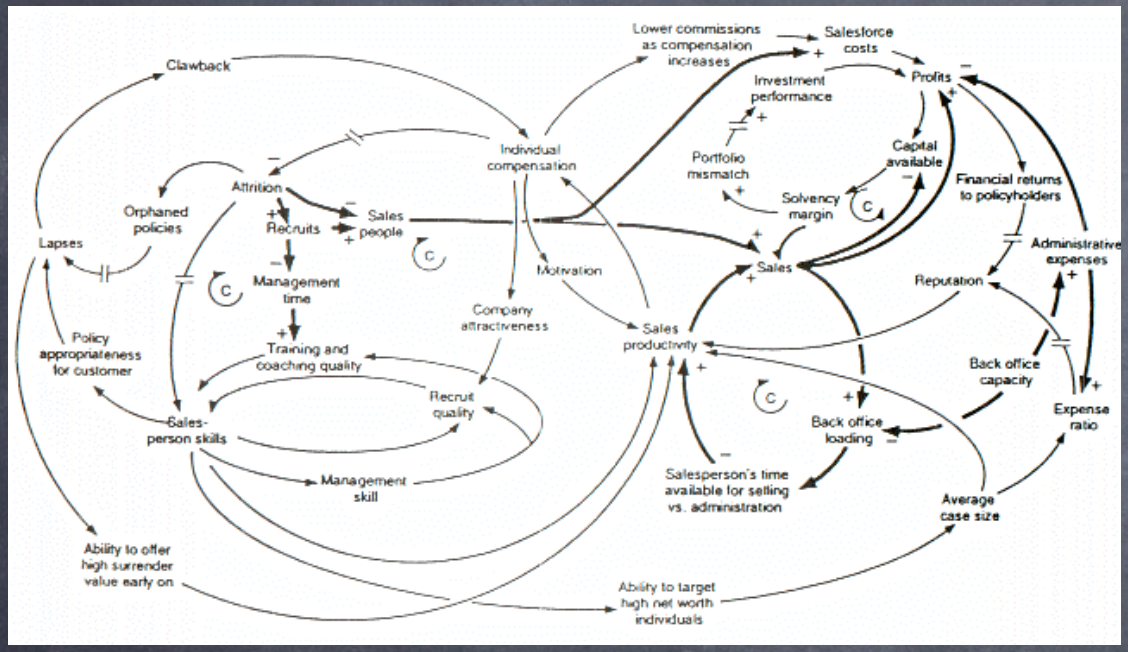
- Hypothesis: "Only a small number of OWL constructs are used in the publicly available ontologies."

Why should anyone care?

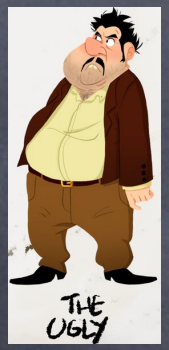
Example: if a particular set of features are almost never used, so it can be ok that your reasoner does not support it.

Evaluation Methods

- Collect a representative corpus
  - Representative is the operative word here
- Analyze the terms used and determine which ones are not used much

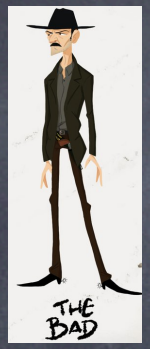


Develop a formal model/ workflow/etc for a task

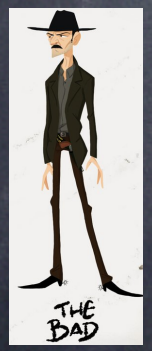


- "We will develop a workflow for creating linked data and annotating it with ontologies"

Nothing to measure here. You cannot falsify the hypothesis

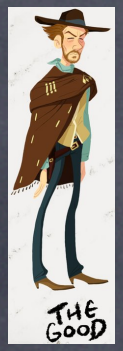


- "We will study a number of existing workflows and will create a more general one."

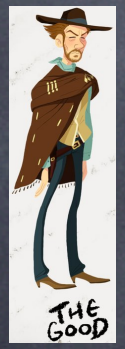


- Hypothesis: "It is possible to build a formal workflow for collaborative creation of linked data (similar: It is possible to develop a formal representation for X)"

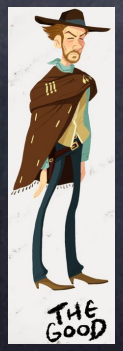
Not great: You cannot falsify this



- Hypothesis: "My workflow model is generic enough to represent a meaningful number of diverse published workflows".



- Auxiliary hypothesis: "My system provides sound and complete reasoning."



- Auxiliary hypothesis: "My formalism elements are symmetric, reflexive, transitive."

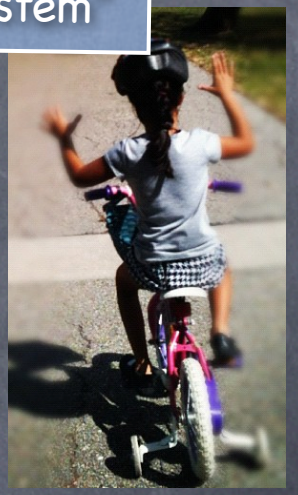
Evaluation Methods

- Prove a theorem!
- Find a representative set of workflows/ problems/etc and represent in your model

Re-cap: Types of problems

"My Semantic Web system is better than your Semantic Web system"

"Look, Ma, no hands!"
or
"We built a system"



How can we use Linked data to solve this problem?

We will put everything in RDF and the world will be a better place

Stamp collection for the sake of stamp collecting

What have we learned?

- Make sure
  - you have a good/appropriate research questions
  - you operationalized your research questions with (falsifiable) hypotheses
  - your evaluation plan is designed to test your hypothesis.
- "Who cares?" and "So what?"