Knowledge Graphs and Natural Language Processing (KG&NLP): a symbiosis

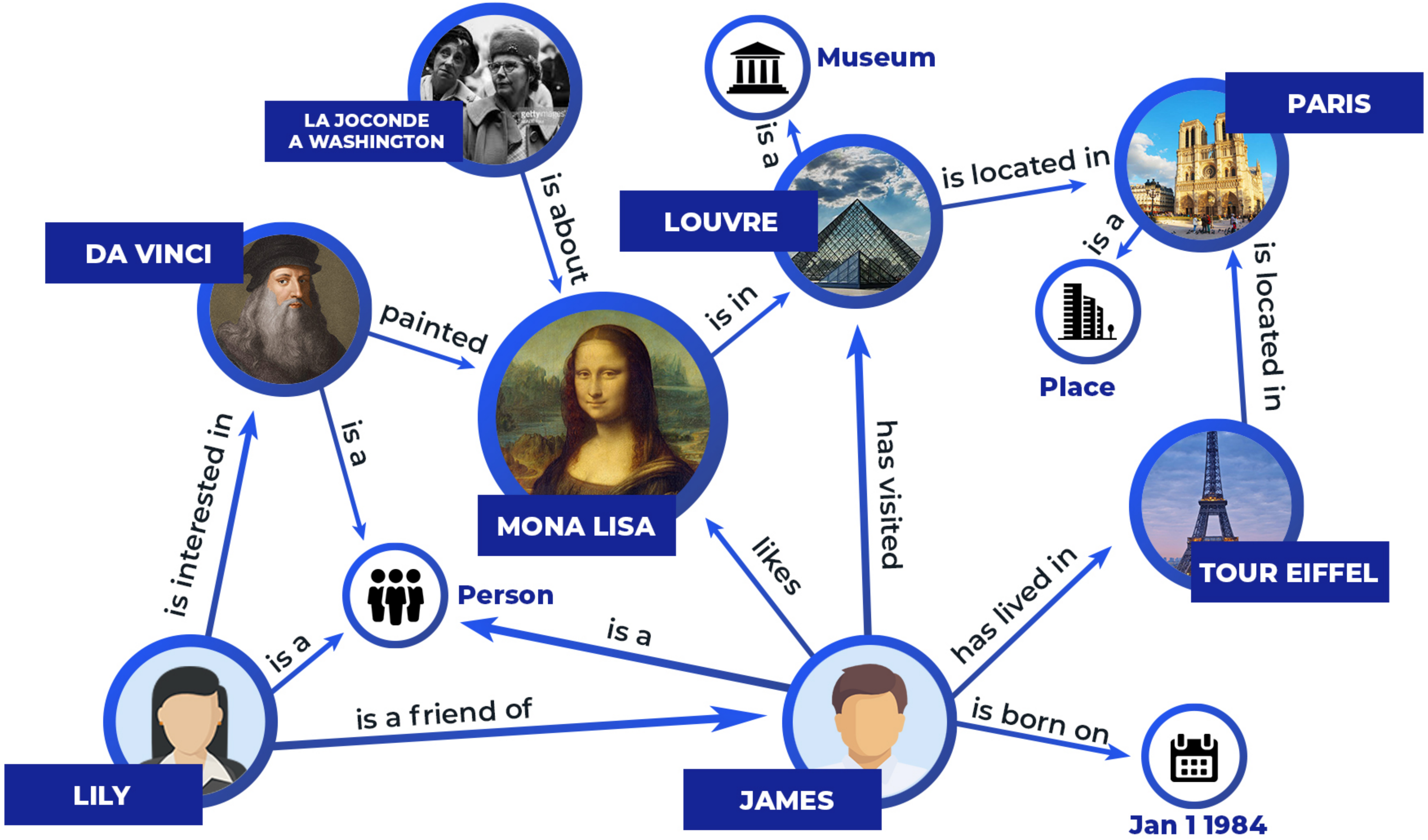
4th Workshop on NLP4AI

Prof. Philipp Cimiano
Semantic Computing Group
Bielefeld University

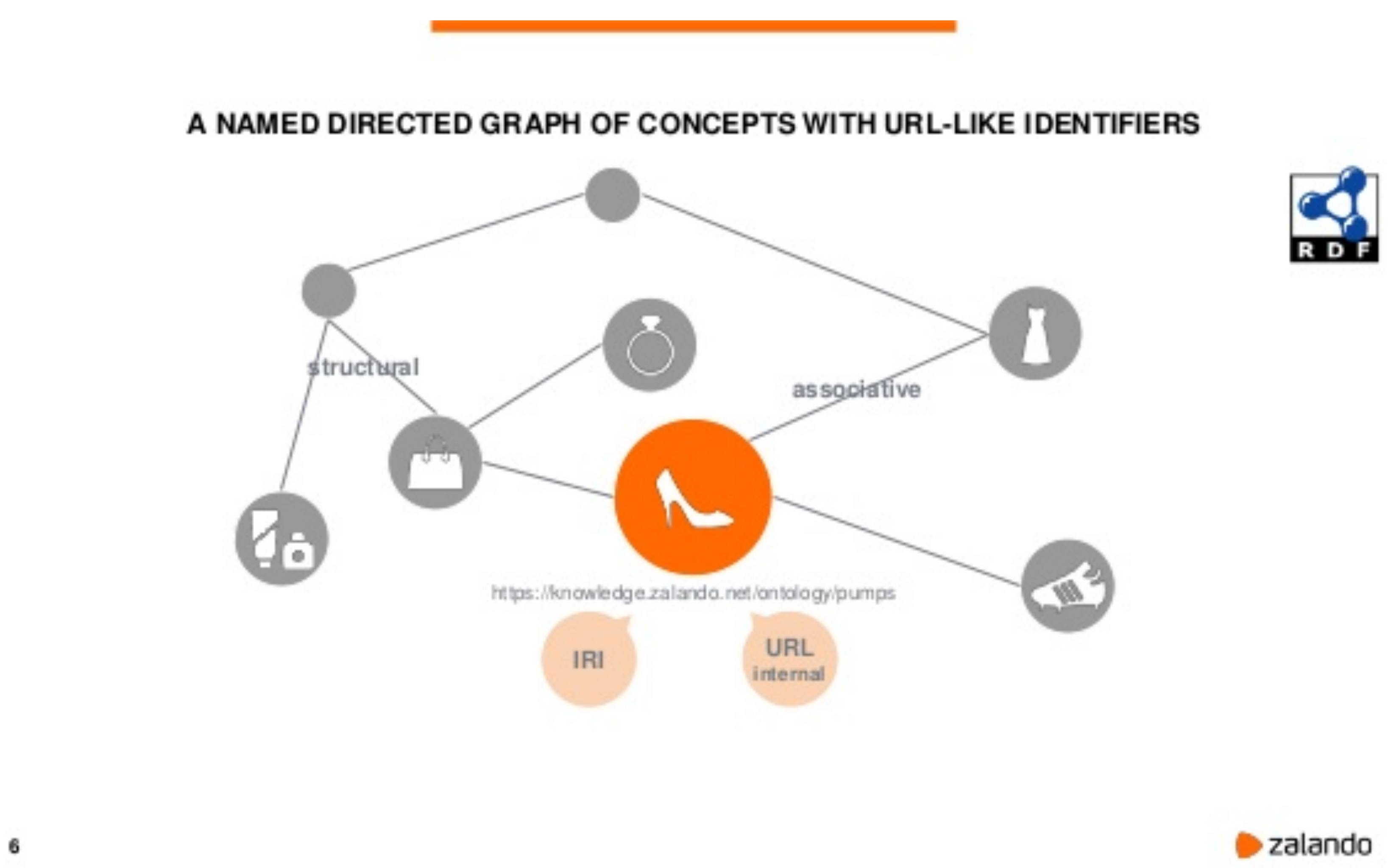
Agenda

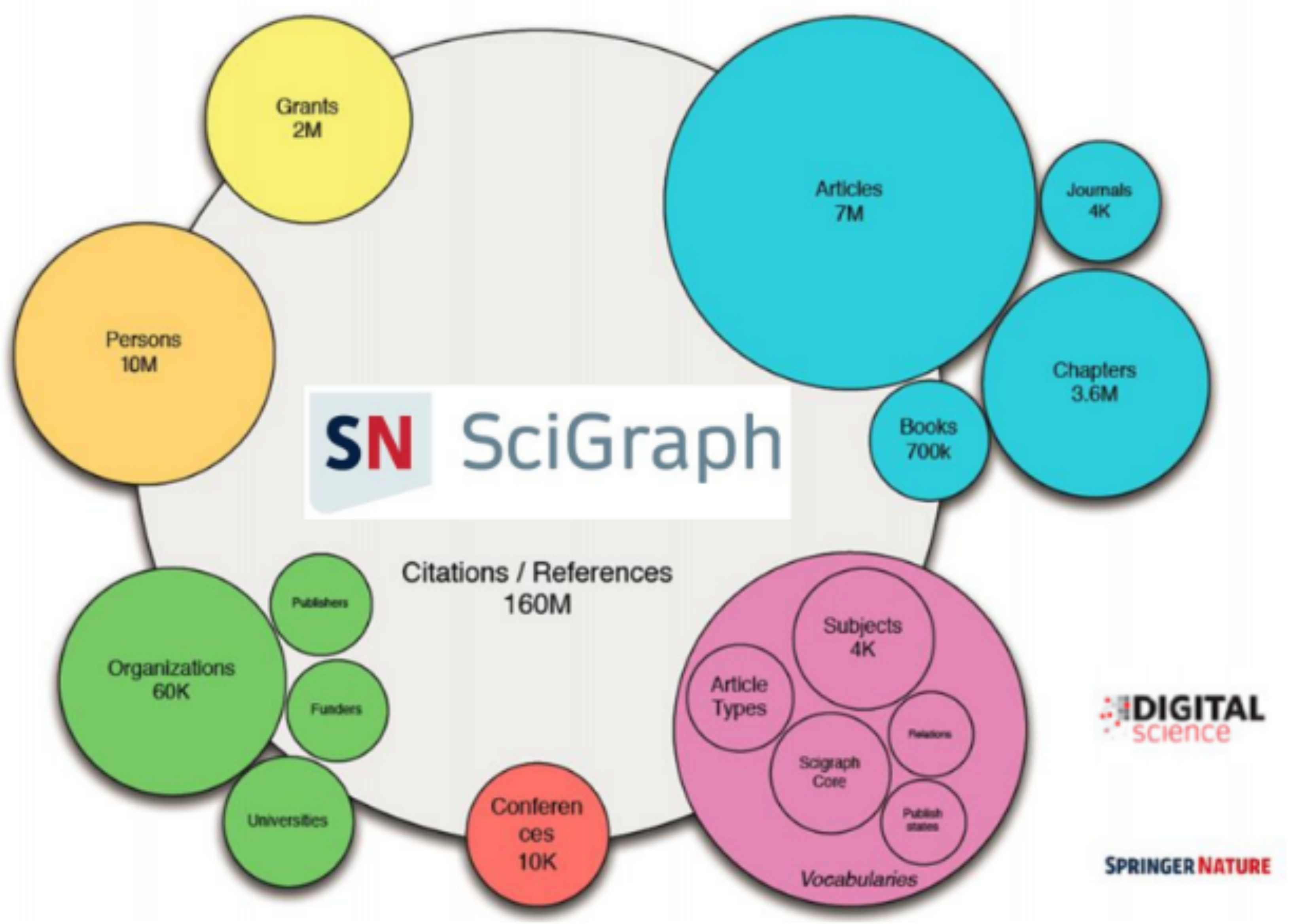
- Knowledge Graphs
- Knowledge Graphs for NLP
- Scientific Knowledge Graphs
- NLP for Knowledge Graphs
- Use Cases
- Future Work

What is a Knowledge Graph?



Example Knowledge Graphs

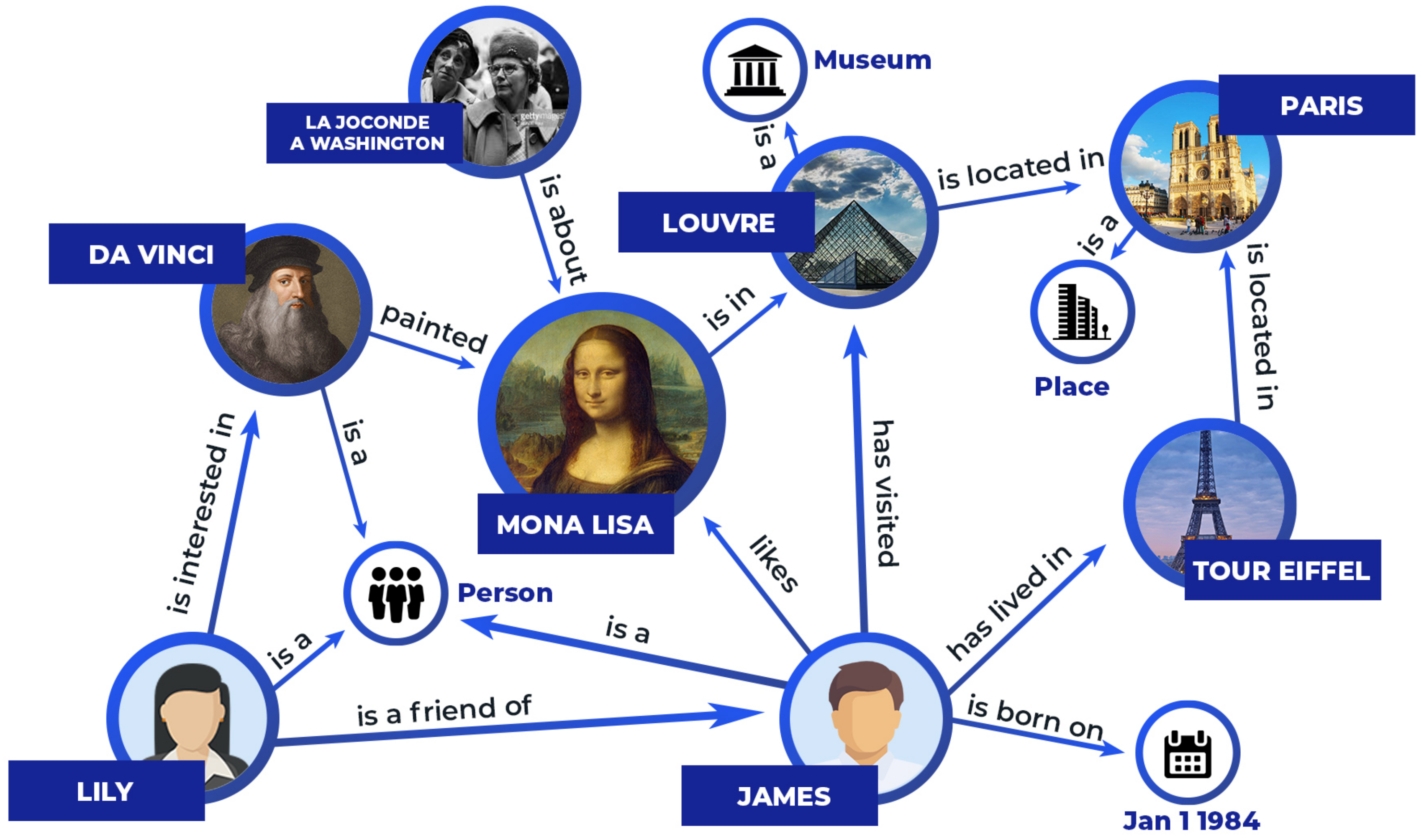




and many more…

Knowledge Graphs for NLP

- Knowledge graphs capture knowledge about
  - Entities
  - Relations
  - Facts
  - Broader Contexts
- NLU is a knowledge-intense task (although nowadays we are ignoring this!)
- KGs help to link a symbol in text to a `representation' of the real entity, that is a proxy of its denotation, introducing additional context



Knowledge Graphs for Information Extraction

Jack „Jacky“ Miller is an English soccer player born 15 December 1977 in Scunthorpe, now residing in Berlin. He began his soccer career as a youth player for F.C. Scunthorpe United.

Property	Value
Name	?
Year of Birth	?
Place of Birth	?
Soccer Club	?

Knowledge Graphs for Information Extraction

Jack "Jacky" Miller is an English soccer player born 15 December 1977 in Scunthorpe, now residing in Berlin. He began his soccer career as a youth player for F.C. Scunthorpe United.

Property	Value
Name	Jack „Jacky" Miller
Year of Birth	?
Place of Birth	?
Soccer Club	?

Knowledge Graphs for Information Extraction

Jack "Jacky" Miller is an English soccer player born 15 December 1977 in Scunthorpe, now residing in Berlin. He began his soccer career as a youth player for FC Scunthorpe United.

Property	Value
Name	Jack „Jacky" Miller
Year of Birth	15/12/1977
Place of Birth	?
Soccer Club	?

Knowledge Graphs for Information Extraction

Jack "Jacky" Miller is an English soccer player born 15 December 1977 in Scunthorpe, now residing in Berlin. He began his soccer career as a youth player for F.C. Scunthorpe United.

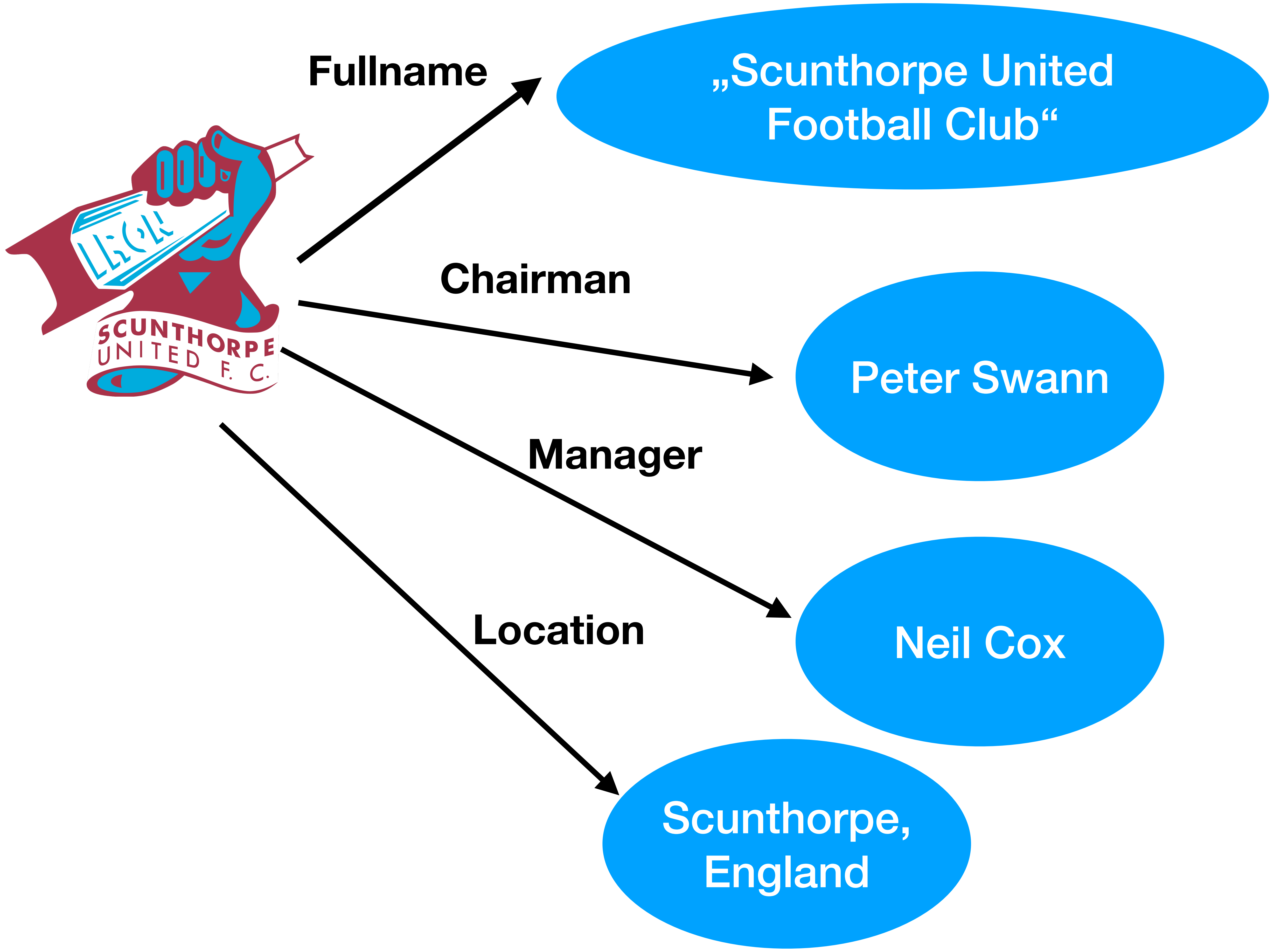
Property	Value
Name	Jack „Jacky" Miller
Year of Birth	15/12/1977
Place of Birth	?
Soccer Club	F.C. Scunthorpe United

Knowledge Graphs for Information Extraction

Jack "Jacky" Miller is an English soccer player born 15 December 1977 in Scunthorpe, now residing in Berlin. He began his soccer career as a youth player for F.C. Scunthorpe United.

Property	Value
Name	Jack „Jacky" Miller
Year of Birth	15/12/1977
Place of Birth	?
Soccer Club	F.C. Scunthorpe United.

Knowledge Graphs to the Rescue!



Structured Prediction with Conditional Random Field (CRFs)

Probabilistic model of the probabilities of target variables given the text x:

$$p(\boldsymbol{y}|\boldsymbol{x};\theta)$$

Maximum-a-posteriori inference to determine most likely slot filling:

$$\vec{y}\,' = \underset{\vec{y}}{\operatorname{argmax}}\, p(\vec{y}|\vec{x};\theta).$$

Property	Value
Name	Jack „Jacky“ Miller
Year of Birth	15/12/1977
Place of Birth	?
Soccer Club	F.C. Scunthorpe United.

CRFs with factor graphs

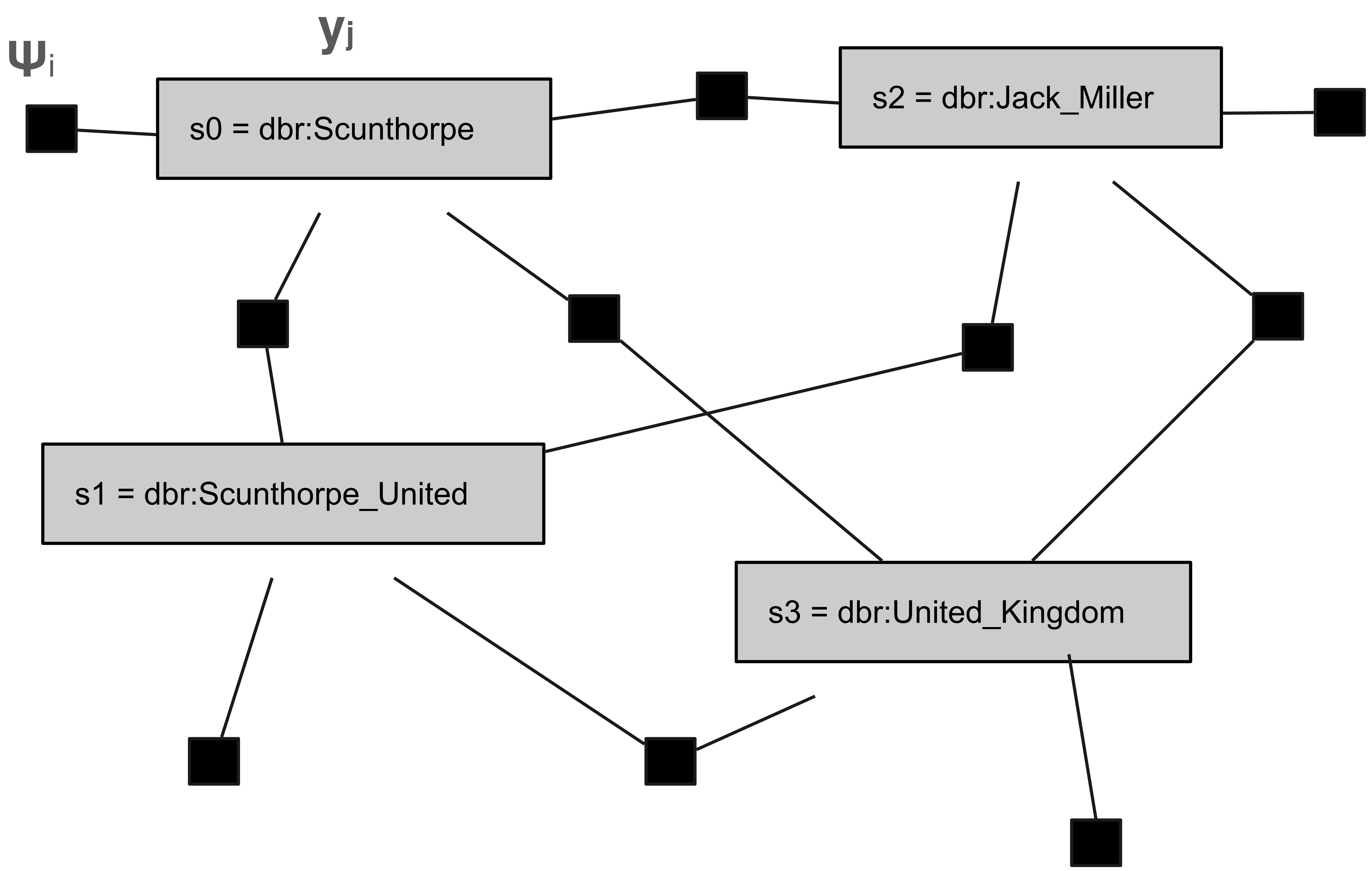
Formal Definition:

- A factor graph G consists of
  - random variables **x** and **y** (grey boxes)
  - factors **Ψ** (black boxes).

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \frac{1}{Z(\boldsymbol{x})} \prod_{\Psi_i \in \mathcal{G}} \Psi_i$$

- Each factor computes a **scalar score** based on a feature vector $f_i(\mathbf{x}_i, \mathbf{y}_i)$ and a set of parameters θ_i:

$$\Psi_i = e^{f_i(\boldsymbol{x}_i, \boldsymbol{y}_i) \cdot \theta_i}$$



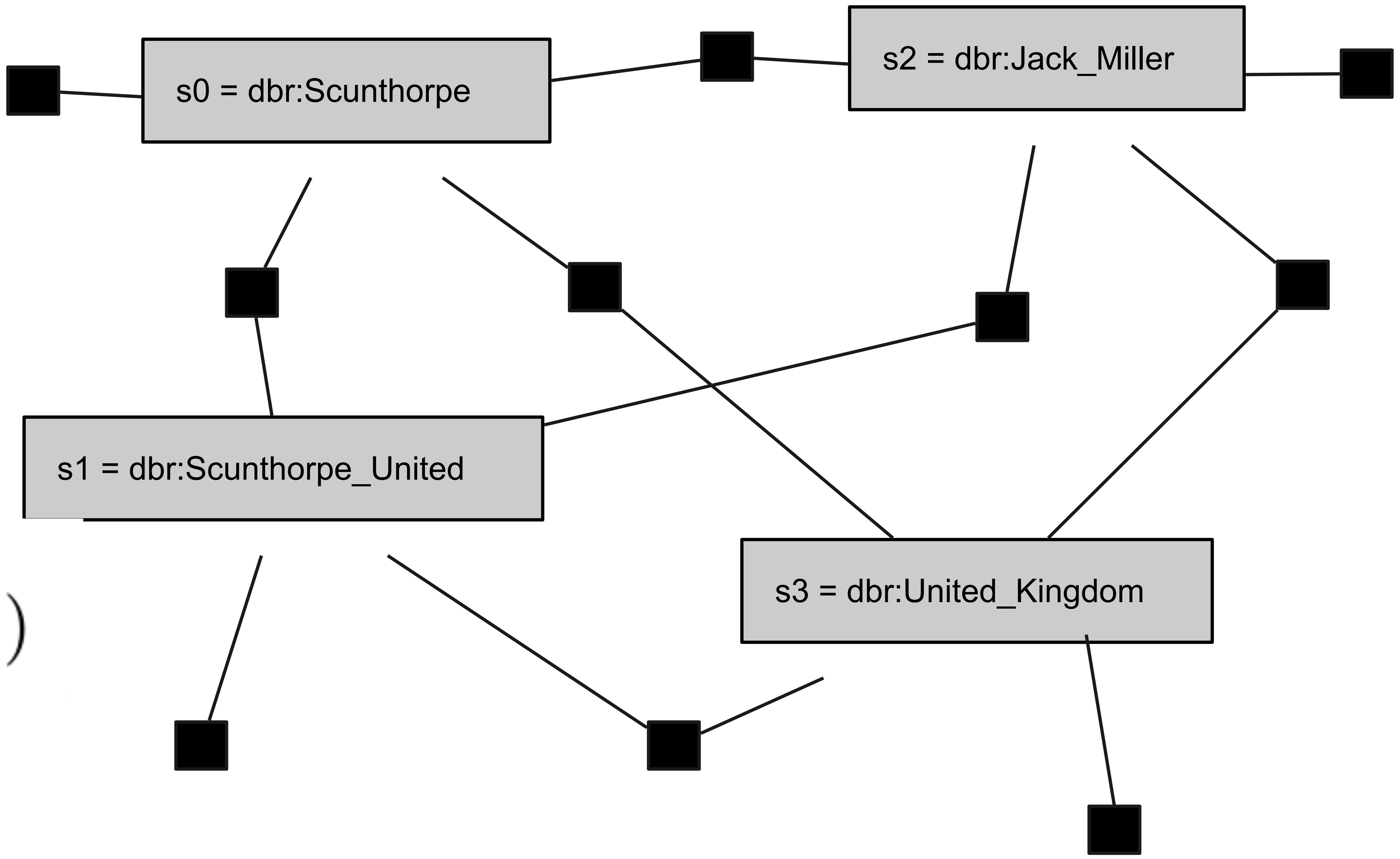
Incorporating Semantic Dependencies From Knowledge Graphs

Conditional Random Field:

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \frac{1}{Z(\boldsymbol{x})} \prod_{\Psi_i \in \mathcal{G}} e^{f_i(\boldsymbol{x}_i, \boldsymbol{y}_i) \cdot \theta_i}$$

Query as feature function:

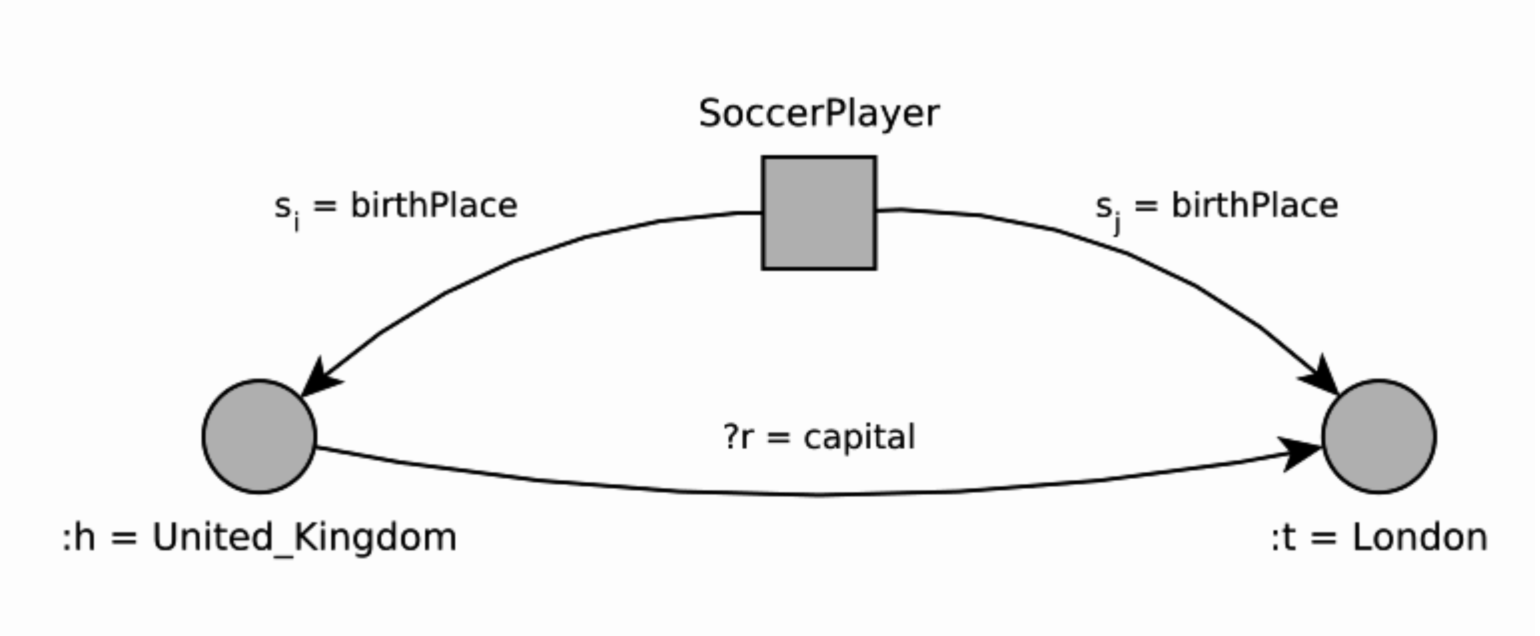
$$f_{s_i, s_j}(\hat{e}, e', p) = \begin{cases} 1 & \text{iff } p \in eval(Q(\hat{e}, e')) \\ 0 & \text{else} \end{cases}$$



Path Queries

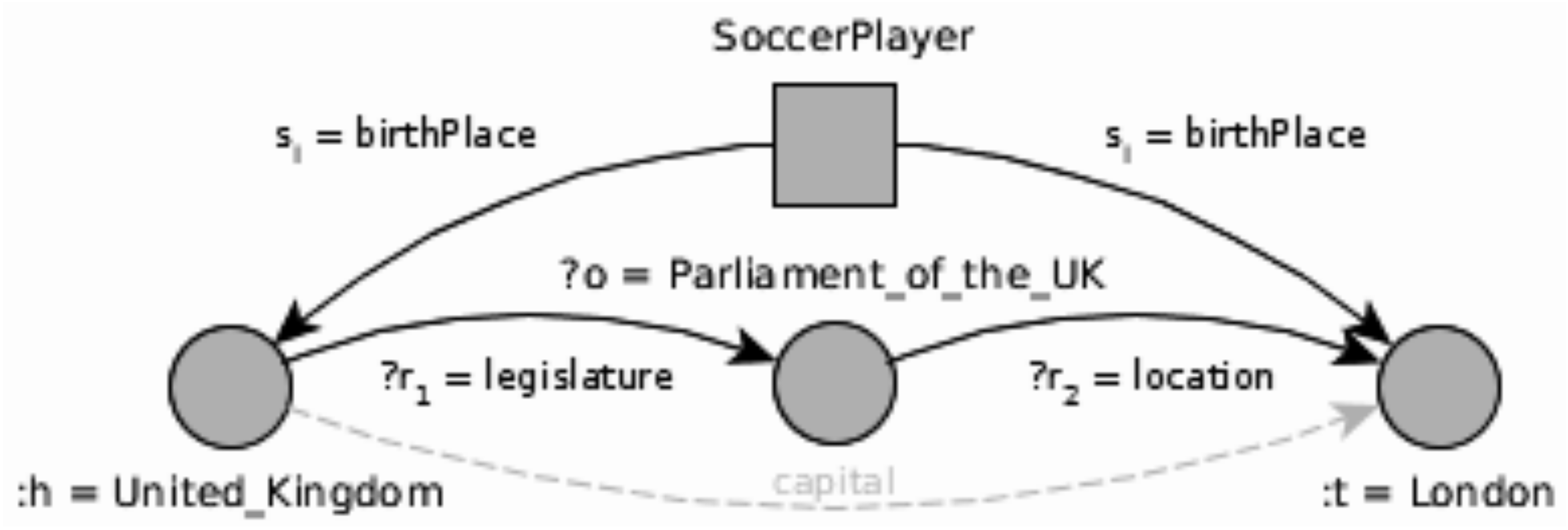
1. Undirected-1-hop-relation:

$$Q_1(e_1, e_2) = \{e_1 \ ?r \ e_2\}$$



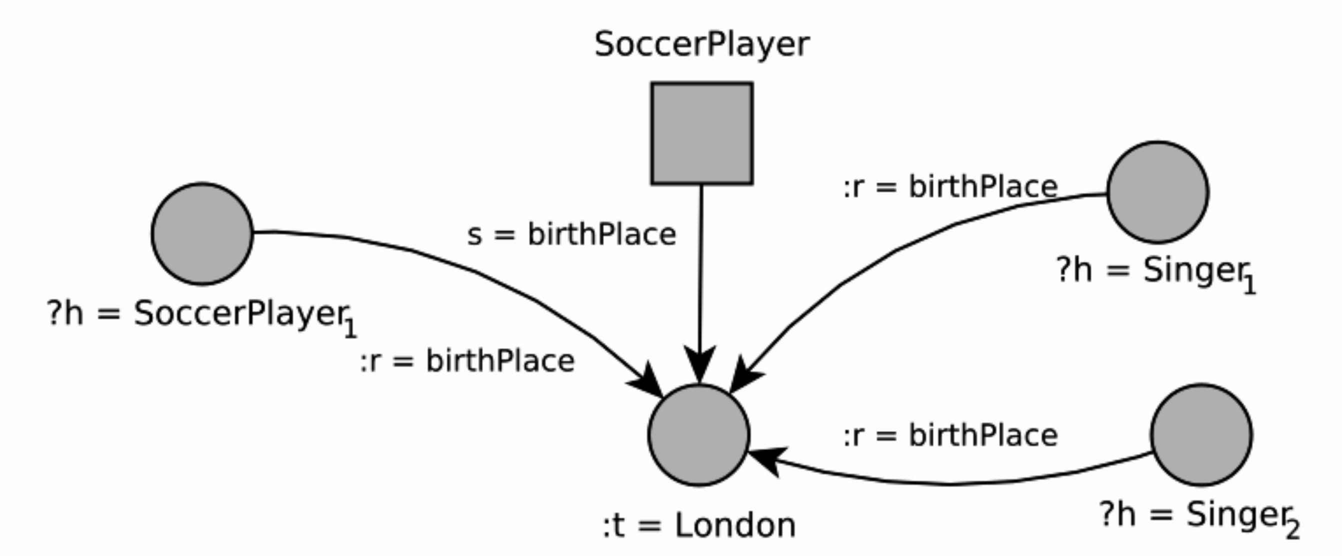
2. Undirected-2-hop-relation:

$$Q_2(e_1, e_2) = \exists o \ \{e_1 \ ?r_1 \ o \wedge o \ ?r_2 \ e_2\}$$



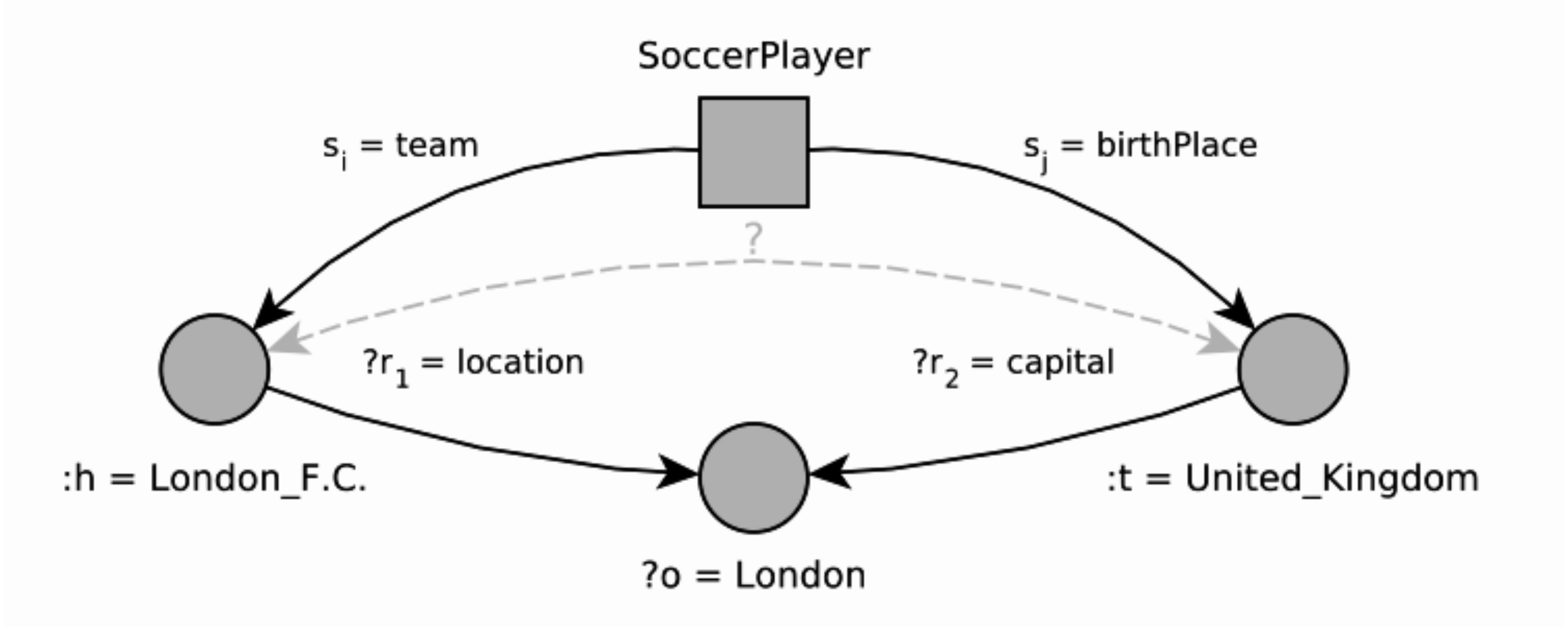
3. Tail Plausibility:

$$Q_3(r, e_1) = \exists h \ \{h \ r \ e_1\}$$



4. V-relation:

$$Q_4(e_1, e_2) = \exists o \ \{e_1 \ ?r_1 \ o \wedge e_2 \ ?r_2 \ o\}$$



Datasets

- dbr:SoccerPlayer (#2338)
- dbr:Film (#1000)
- dbr:Single (#1000)
- dbr:ArchitecturalStructure (#527)
- dbr:Dam (#491)

ARCHITECTURALSTRUCTURE			
resource	single	entity	530
architect	multi	entity	356
architecturalStyle	multi	entity	70
location	multi	entity	408
openingYear	single	literal	n/a
yearOfConstruction	multi	literal	n/a
DAM			
resource	single	entity	502
river	multi	entity	370
country	multi	entity	74
location	multi	entity	601
status	single	literal	n/a
openingYear	multi	literal	n/a
buildingStartYear	multi	literal	n/a

Dataset	Cardinality	Type	Values
SOCCERPLAYER			
resource	single	entity	4,905
birthPlace	multi	entity	2,314
birthYear	single	literal	n/a
deathYear	single	literal	n/a
position	single	entity	14
team	multi	entity	4,359
FILM			
resource	single	entity	5,405
producer	multi	entity	3,347
director	multi	entity	5,452
writer	multi	entity	5,452
starring	multi	entity	9,452
musicComposer	multi	entity	731
distributor	multi	entity	1,936
SINGLE			
resource	single	entity	17,881
writer	multi	entity	7,001
musicalBand	multi	entity	4,791
musicalArtist	multi	entity	4,791
album	multi	entity	10,090
producer	multi	entity	4,185
genre	multi	entity	548

Experiments on 5 datasets

- Experiments were run 3 times with a random 80/20 (train/test) split
- We report averaged harmonic macro F1
- Baseline: pick entity as property value based on the frequency in the text
- Intra-textual Model: CRF that relies on textual features only
- Intra-textual + SD Model: CRF with KG dependencies on top of textual features

	Baseline			Intra-textual Model			Intra-textual + SD Model		
Dataset	**prec.**	**rec.**	**F_1**	**prec.**	**rec.**	**F_1**	**prec.**	**rec.**	**F_1**
SOCCERPLAYER	0.35	0.31	0.33	0.82	0.77	0.79	**0.91**	**0.88**	**0.89**
FILM	0.09	0.09	0.09	*0.63*	0.65	*0.64*	*0.63*	**0.66**	*0.64*
SINGLE	0.11	0.12	0.11	0.62	0.64	0.61	**0.70**	**0.79**	**0.74**
ARCH_STRUCT*	0.25	0.28	0.26	0.76	0.68	0.71	**0.77**	**0.69**	**0.73**
DAM	0.39	0.42	0.40	*0.80*	0.66	0.72	0.80	**0.68**	**0.73**
Average	0.24	0.24	0.24	0.72	0.67	0.70	**0.76**	**0.74**	**0.75**

Ablation Experiments

Model	**precision**	**recall**	$\mathbf{F_1}$
Lin+BK	0.76	0.74	0.75
Lin+BK - Q_1	0.75	0.73	0.74
Lin+BK - Q_2	**0.74**	**0.70**	**0.72**
Lin+BK - Q_3	0.76	0.73	0.74
Lin+BK - Q_4	0.77	0.73	0.75

Interim Conclusions

1. Background knowledge can be successfully integrated into a CRF model via indicator functions.

2. The integration shows a positive impact in the slot filling task. Up to 10% F1 for individual datasets. (SoccerPlayer)

3. Ablation study shows that a wide variety of queries is important to retrieve information from the knowledge graph, two-hop dependencies being the most important.

Scientific Knowledge Graphs

- **Definition:** A scientific knowledge graph is a knowledge graph that describes the body of research in some discipline!
- What are Scientific Knowledge Graphs good for?
  - Knowledge Exploration
  - Knowledge Aggregation (e.g. Scientific Reviews, Meta-Analysis, …)
  - Inconsistency Detection

1st Workshop on Scientific Knowledge Graphs

February 25, 2020 | BY angelosalatino

Event Notification Type: Call for Papers
Abbreviated Title:
SKG2020

Tuesday, 25 August 2020
Country: France
Contact Email: skg2020@easychair.org
City: Lyon
Contact: Angelo Salatino
Andrea Mannocci
Francesco Osborne
Website: https://skg.kmi.open.ac.uk/SKG2020/
Submission Deadline: Saturday, 4 April 2020

web: https://skg.kmi.open.ac.uk, twitter: @skgworkshop

Held in conjunction with TPDL2020, 25th-28th August 2020, Lyon, France
----

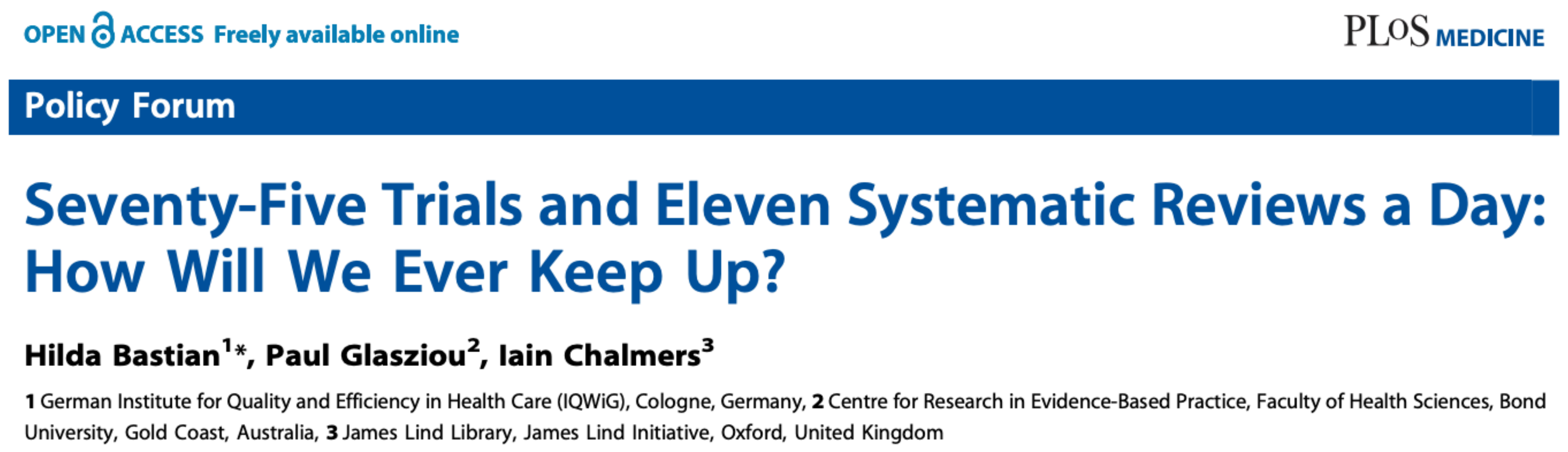
Knowledge aggregation in medicine

- Systematic reviews crucial for advance of knowledge in medicine, carried out since mid 70s

- Systematic reviews, meta-analysis, HTA require knowledge aggregation, currently done by hand (!)

- Outdated when they are published, although they are supposed to be updated, for most systematic reviews this is not the case

- 12 clinical trials per average in one systematic review => most clinical trials not „reviewed“


OPEN ACCESS Freely available online PLoS MEDICINE

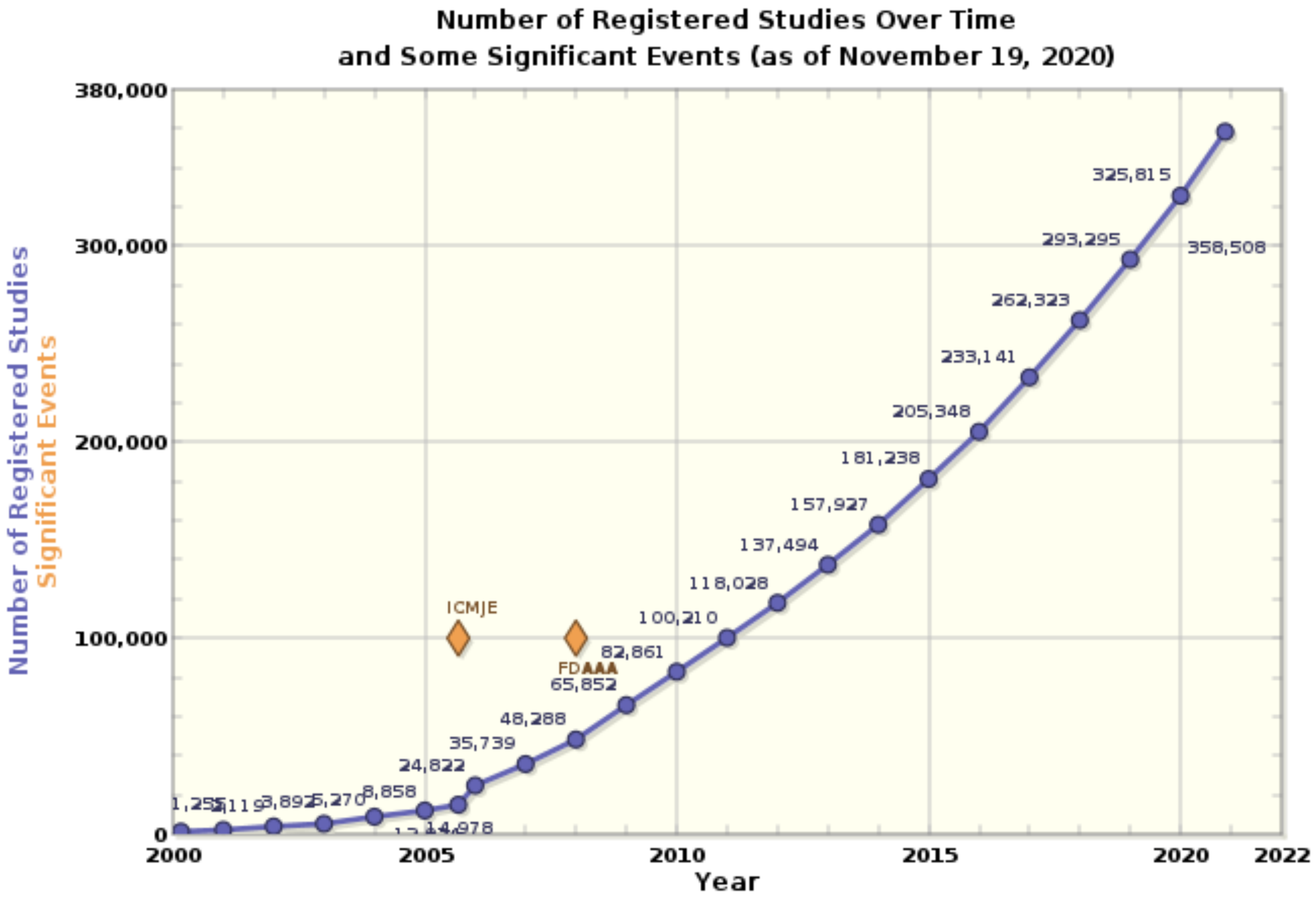
Policy Forum

Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up?

Hilda Bastian[1]*, Paul Glasziou[2], Iain Chalmers[3]

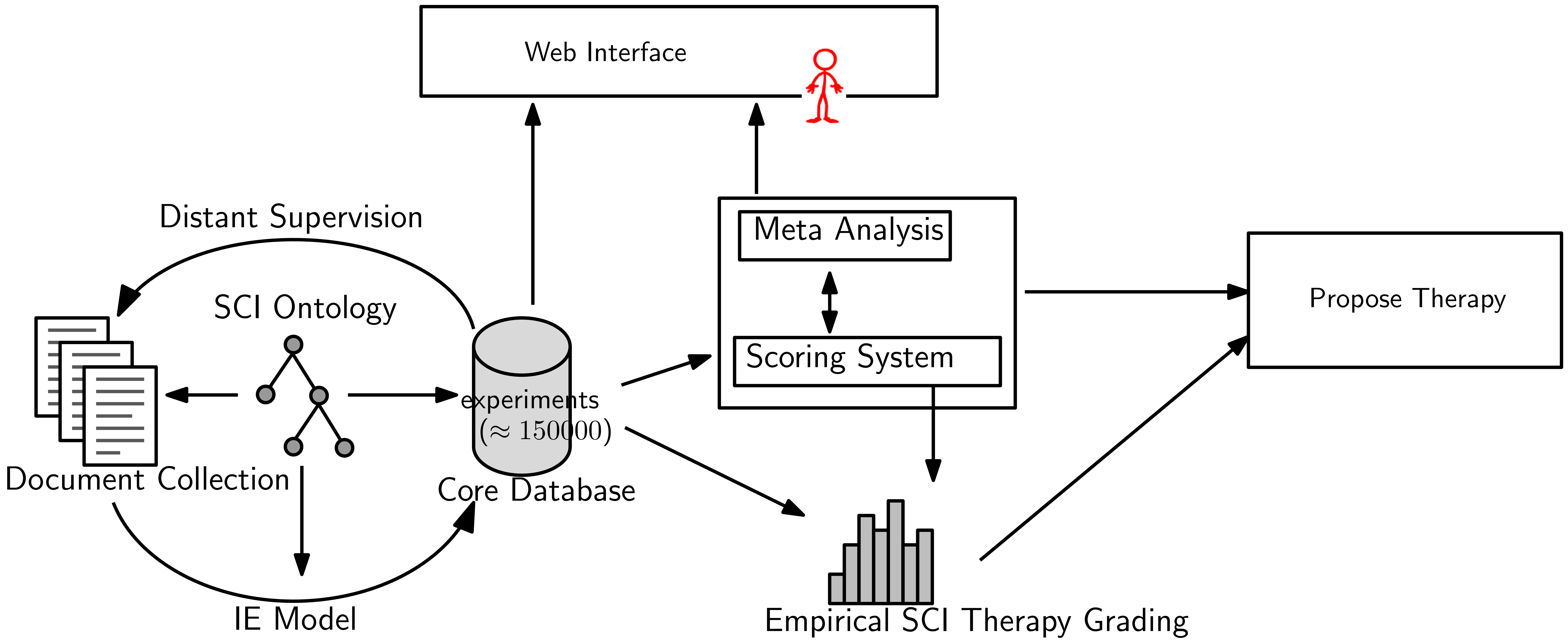
1 German Institute for Quality and Efficiency in Health Care (IQWiG), Cologne, Germany, 2 Centre for Research in Evidence-Based Practice, Faculty of Health Sciences, Bond University, Gold Coast, Australia, 3 James Lind Library, James Lind Initiative, Oxford, United Kingdom

Registered Clinical Trials

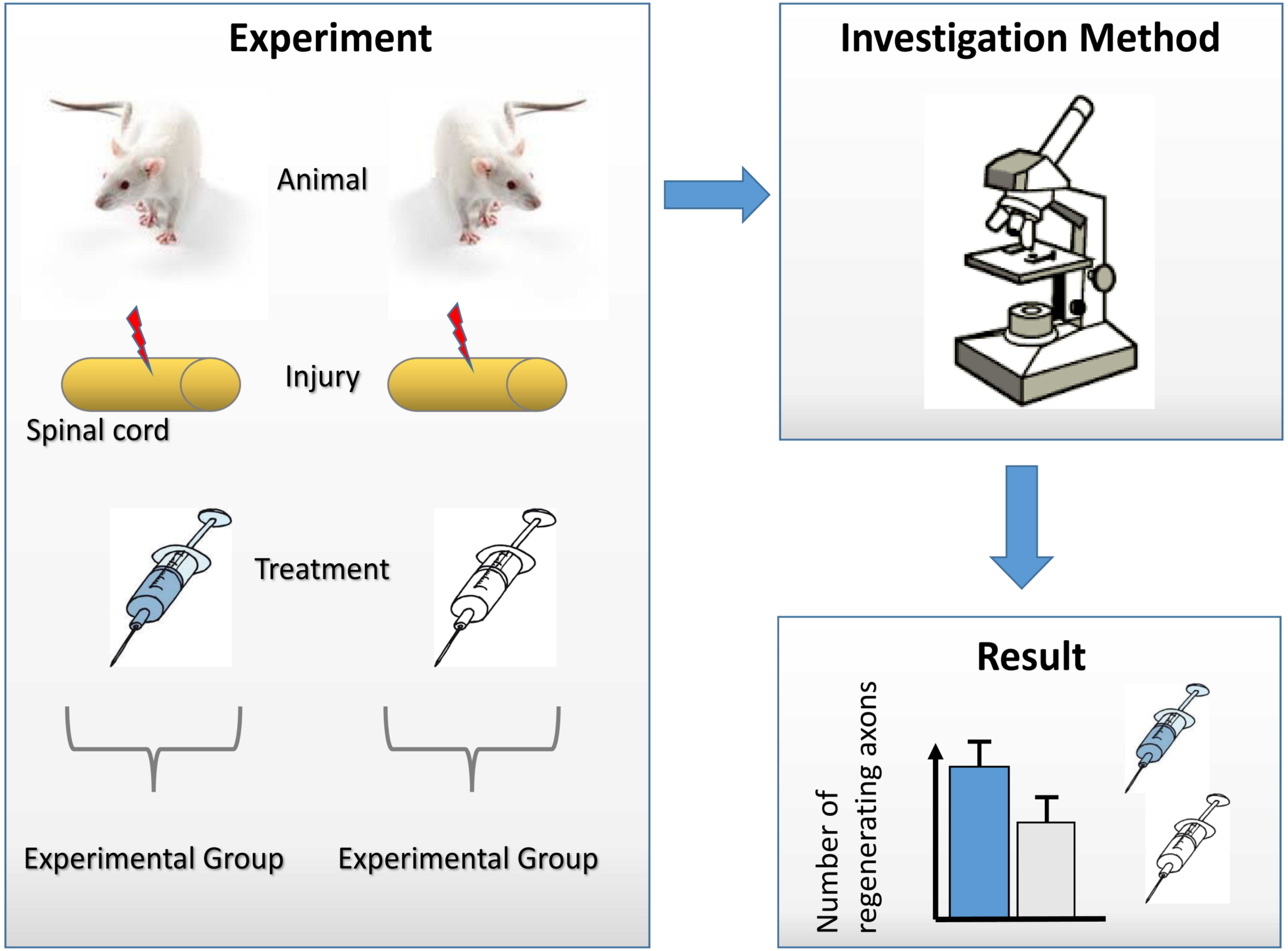


Source: https://ClinicalTrials.gov

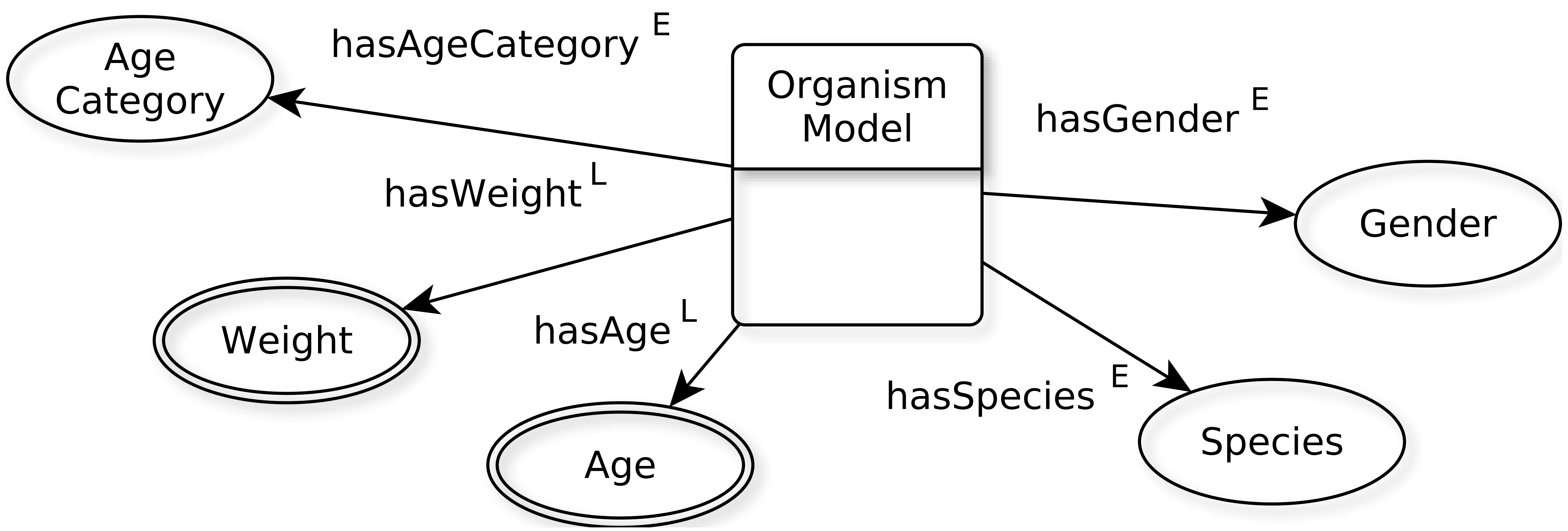
PSINK


Web Interface
Distant Supervision
SCI Ontology
Meta Analysis
Scoring System
Propose Therapy
experiments
(≈ 150000)
Document Collection
Core Database
IE Model
Empirical SCI Therapy Grading

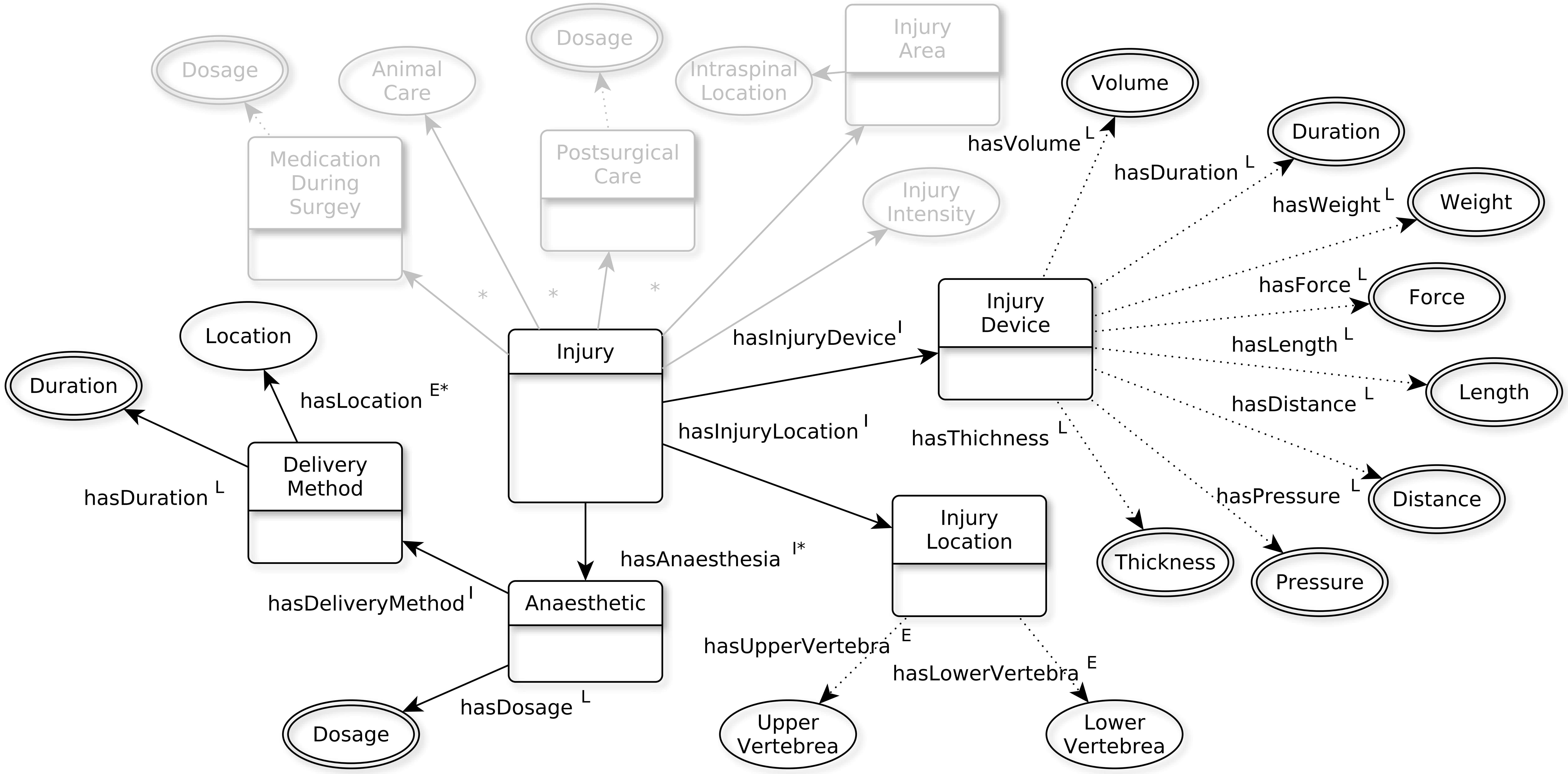
Extracting Outcomes from Pre-Clinical Literature



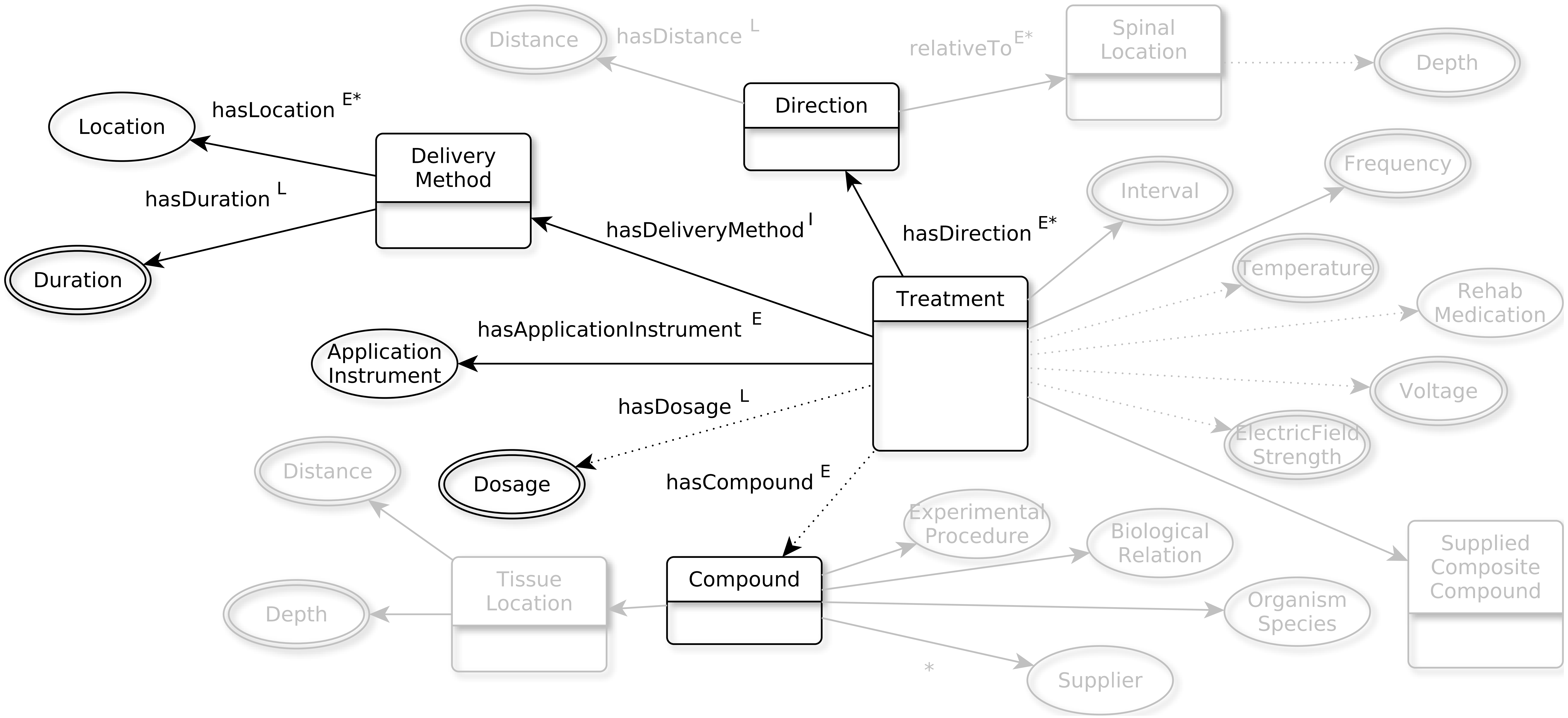
Spinal Cord Injury Ontology (SCIO)
Organism Model



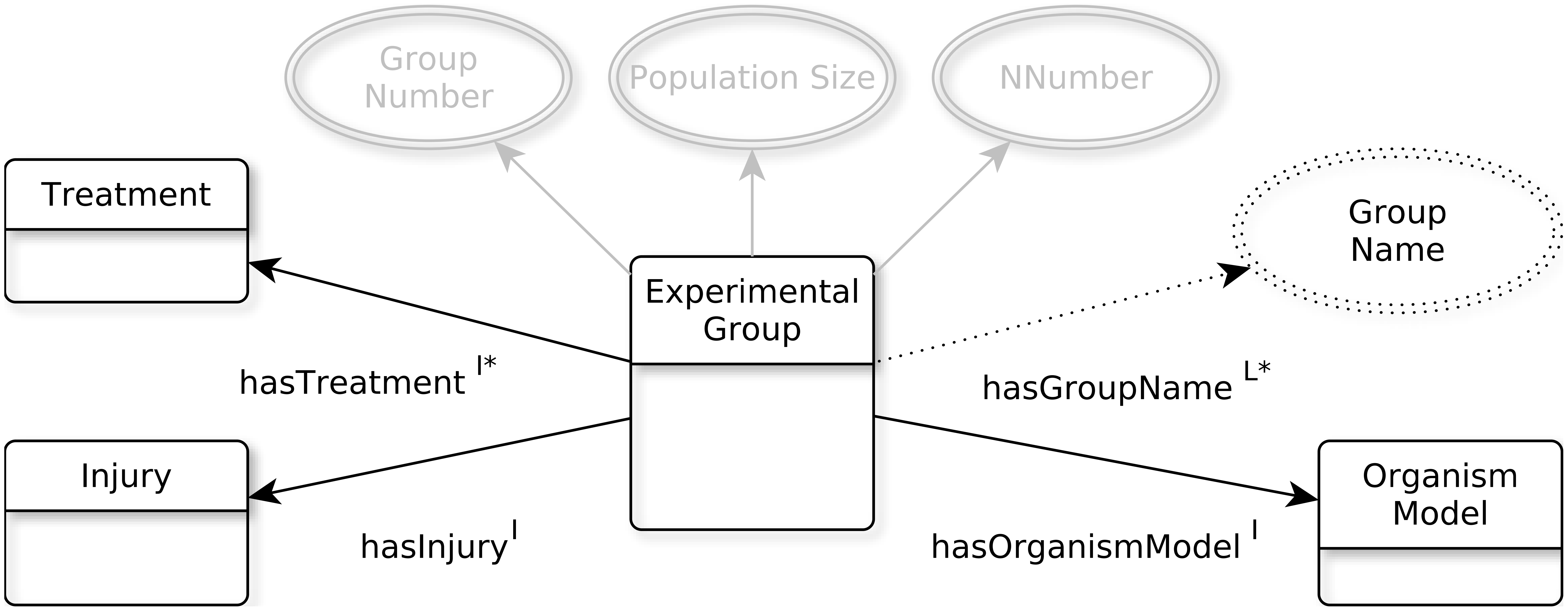
Spinal Cord Injury Ontology (SCIO)
Injury



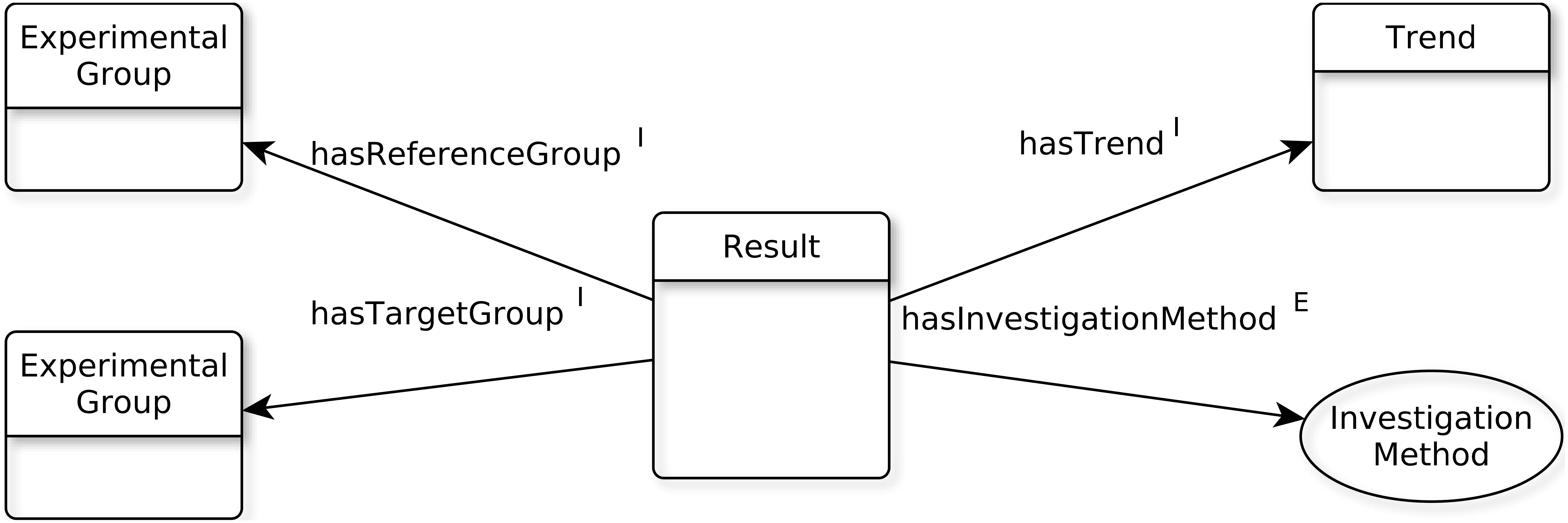
Spinal Cord Injury Ontology (SCIO)
Treatment



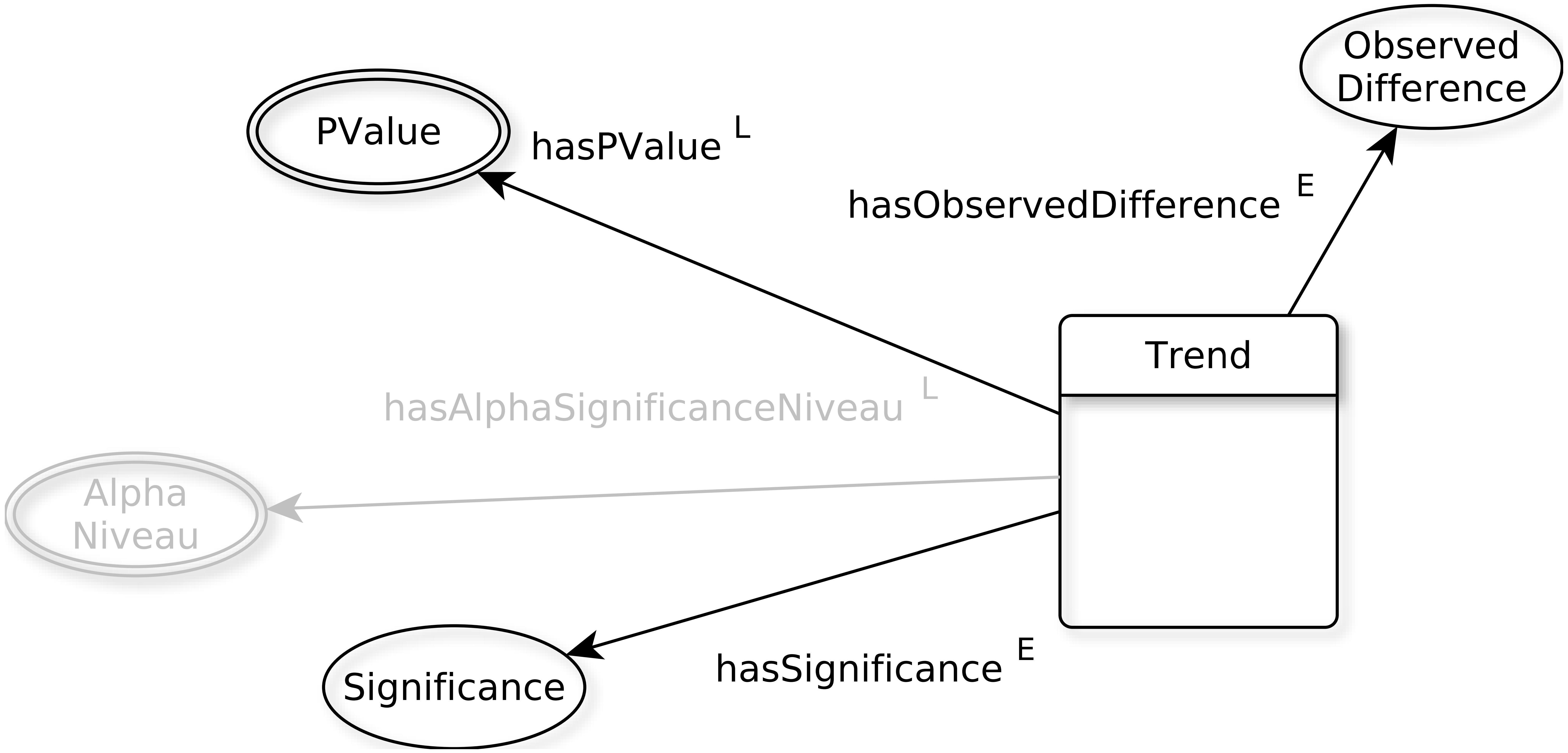
Spinal Cord Injury Ontology (SCIO)
Experimental Group



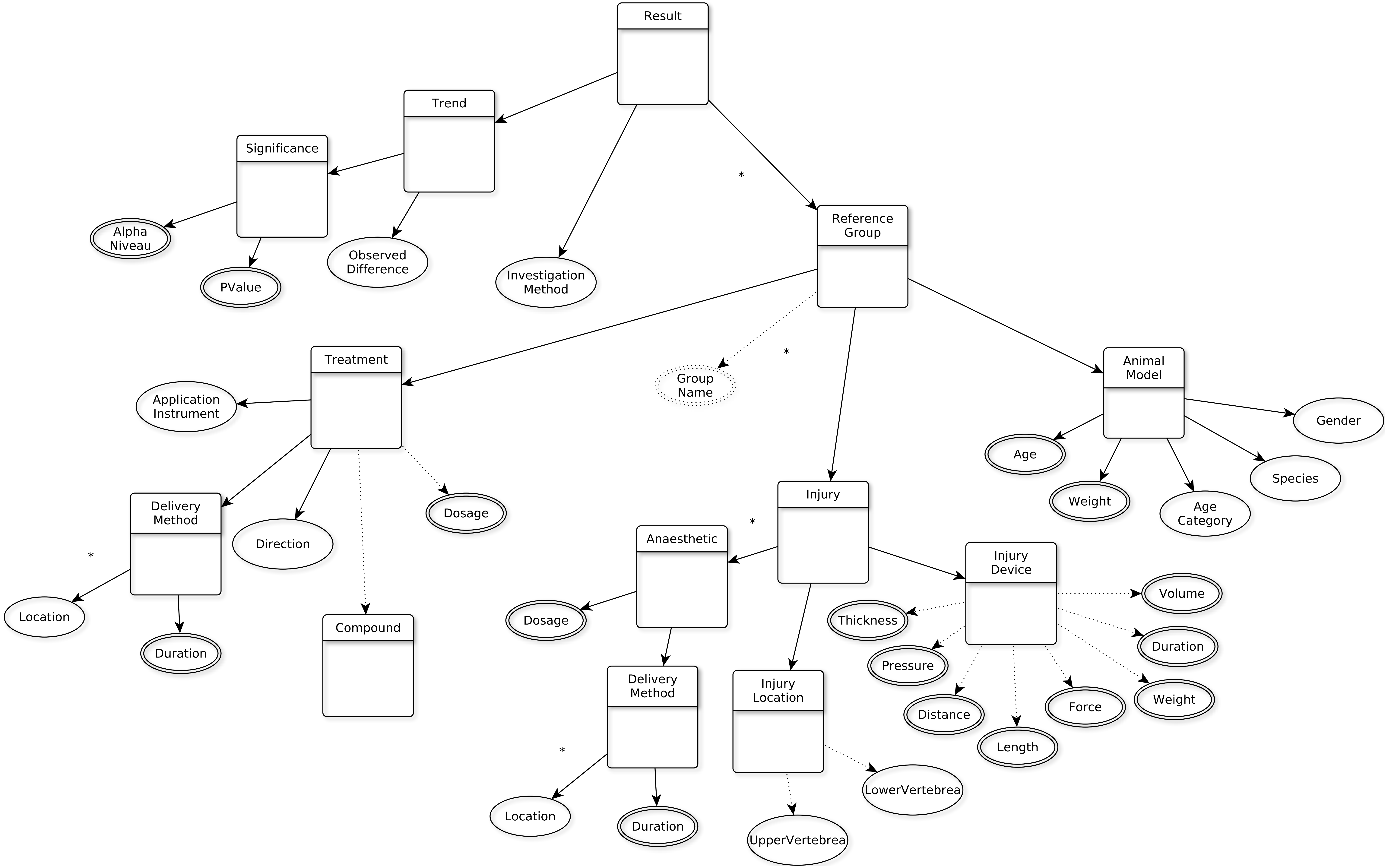
Spinal Cord Injury Ontology (SCIO)
Result



Spinal Cord Injury Ontology (SCIO)
Trend



Spinal Cord Injury Ontology



Problem Formulation

- Model-Complete Text Comprehension (MCTC)
  - Interpret text with respect to a given (ontological) model
  - Extract only those things that are relevant for a given ontology
  - Ignore everything else
- Important Subproblem: predict cardinality (inference!)

Model-complete text comprehension as structured (probabilistic) inference

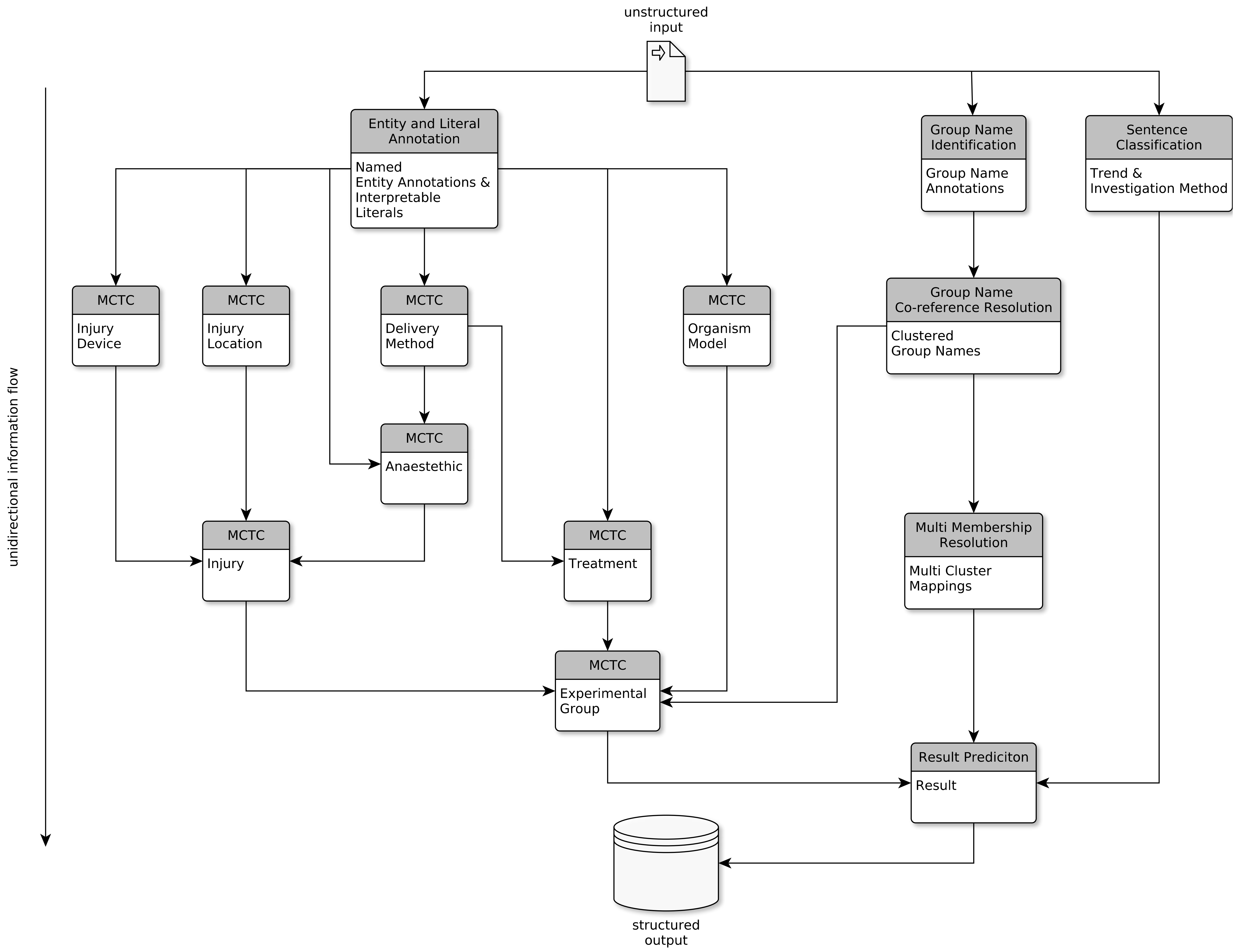
- Assume we have an ontology with two classes:
  - Class **A** with properties **hasProperty1, hasProperty2** and has **Property3**
  - Class **B** with properties **hasProperty4, hasProperty5**
- An instance of the model/ontology could be captured by the following vector:

  $(2,\{(a_1,p_1,p_2,p_3),(a_2,p_{1'},p_{2'},p_{3'})\},3,\{(b_1,p_4,p_5),(b_2,p_{4'},p_{5'}),(b_3,p_{4''},p_{5''})\})$

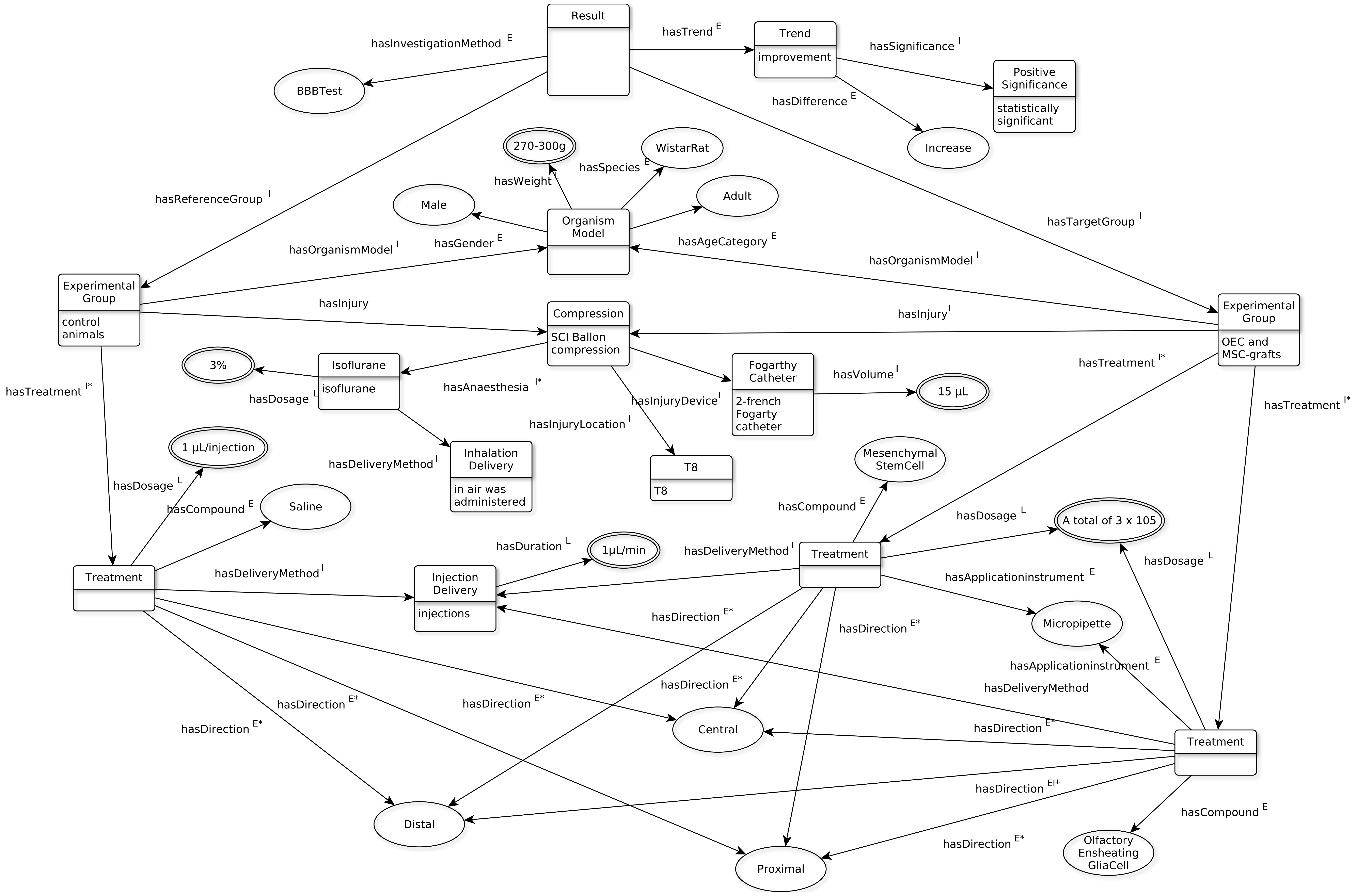
- Let ***Y*** be the set of all such vectors given a certain (ontological) vocabulary *O. We get the most likely instance of our model by maximum-a-posteriori inference:*

$$\vec{\mathbf{y}}\ ' = \operatorname*{argmax}_{\vec{y} \in \mathbf{Y}} p(\vec{y}|\vec{x}).$$

IE Architecture



Example target structure



Annotated corpus

Corpus	#doc.	#ind	#itp	#etp	#ltp	#tri	#ett	#ltt
OrganismModel	200	216	0	3	2	1.006	571	219
InjuryDevice	165	169	0	0	8	270	0	101
DeliveryMethod	161	268	0	1	1	736	296	29
InjuryLocation	198	221	0	0	2	355	134	0
Anaesthetic	160	269	1	0	1	986	311	264
Injury	199	220	3	0	0	2.308	1.150	363
Treatment	133	454	1	3	1	3.309	1.622	278
ExperimentalGroup	95	219	3	0	1	7.636	3.821	1.845
Trend	91	2.019	0	1	2	4.272	1.841	569
InvestigationMethod	91	843	0	0	0	843	0	0
Result	91	1.770	3	1	0	23.675	14.760	2.425
Overall (distinct)	200	6668	10	9	17	53.829	26.271	4.487

Organism Model

Adult male Wistar rats weighing **270-300 g** were used in our experiments.

ORGANISMMODEL^{C}	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
$hasWeight^{L}$	0.873	0.879	0.866	0.955	0.955	0.955	0.899	0.906	0.893
$hasAge^{L}$	0.681	0.712	0.654	0.957	0.957	0.957	0.662	0.684	0.640
$hasSpecies^{E}$	0.859	0.865	0.853	0.896	0.900	0.892	0.905	0.915	0.895
$hasAgeCategory^{E}$	0.884	0.889	0.880	0.997	1.000	0.993	0.882	0.886	0.877
$hasGender^{E}$	0.962	0.972	0.952	0.994	0.994	0.994	0.962	0.972	0.952
cardinality	0.990	1.000	0.980	0.995	1.000	0.995	0.990	1.000	0.980
overall	0.913	0.944	0.884	0.971	0.995	0.948	0.924	0.935	0.914

Type of Injury, Location and Device

SCI Balloon compression was used to create an SCI.
A **2-french Fogarty catheter** was inserted below **T8**, and
the balloon was inflated with **15 LVolume saline** for **5 min** at T8.

INJURY^C	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
$type^E$	0.596	0.610	0.583	0.992	0.995	0.990	0.595	0.610	0.580
$hasInjuryAnaesthesia^{I*}$	0.383	0.476	0.320	0.412	0.518	0.341	0.804	0.983	0.680
$hasInjuryDevice^I$	0.652	0.644	0.660	0.662	0.652	0.672	0.928	0.934	0.922
$hasInjuryLocation^I$	0.708	0.721	0.695	0.726	0.728	0.724	0.956	0.980	0.934
cardinality	0.977	1.000	0.955	0.997	1.000	0.995	0.977	1.000	0.955
overall	0.532	0.582	0.490	0.661	0.809	0.559	0.804	0.914	0.717

INJURYDEVICE^C	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
$type^E$	0.686	0.689	0.683	0.886	0.890	0.881	0.777	0.780	0.774
$hasWeight^L$	0.641	0.641	0.641	0.683	0.683	0.683	0.511	0.511	0.511
$hasForce^L$	0.704	0.704	0.704	0.778	0.778	0.778	0.852	0.852	0.852
$hasDuration^L$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$hasDistance^L$	0.625	0.625	0.625	0.735	0.735	0.735	0.658	0.658	0.658
$hasVolume^L$	0.000	0.000	0.000	0.500	0.500	0.500	0.500	0.500	0.500
cardinality	0.966	0.976	0.957	0.966	0.976	0.957	0.966	0.976	0.957
overall	0.704	0.702	0.705	0.866	0.890	0.843	0.777	0.765	0.789

INJURYLOCATION^C	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
$type^E$	0.689	0.710	0.669	0.912	0.915	0.908	0.800	0.820	0.781
$hasUpperVertebrae^E$	0.471	0.471	0.471	0.838	0.838	0.838	0.632	0.632	0.632
$hasLowerVertebrae^E$	0.206	0.206	0.206	0.794	0.794	0.794	0.412	0.412	0.412
cardinality	0.979	1.000	0.959	0.997	1.000	0.993	0.979	1.000	0.959
overall	0.620	0.640	0.601	0.908	0.915	0.901	0.741	0.758	0.725

Treatment, Delivery Method

A total of **3x105 OEG** and/or **MSC** was injected through a **glass pipette** at a concentration of 1x105 cells/L, into the **proximal, central and distal** parts of the lesioned spinal cord (each part received 1 L cell suspension), at a depth of 1 mm below the dorsal surface and a rate of 1 L/min using a **Nano-Injector** (Stoelt294 ing Co.); OEG/MSC-transplanted animals received **six injections** (3x105 OEG and 3 x105 MSC) instead of the three injections received by the other animals.

The **control group** received **three injections of saline** (1 L), also into the **proximal, central and distal** parts of the lesioned spinal cord.

TREATMENT^{C}	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
hasDeliveryMethod^{I}	0.282	0.321	0.251	0.333	0.384	0.294	0.715	0.757	0.677
hasDirection^{E*}	0.065	0.075	0.057	0.889	0.933	0.849	0.109	0.151	0.085
hasApp.Instrument^{E}	0.351	0.378	0.328	0.697	0.802	0.616	0.426	0.442	0.411
hasCompound^{E}	0.458	0.532	0.402	0.899	1.000	0.817	0.532	0.589	0.484
hasDosage^{L}	0.038	0.033	0.044	0.253	0.331	0.205	0.319	0.302	0.339
cardinality	0.817	0.860	0.778	1.000	1.000	1.000	0.819	0.738	0.920
overall	0.272	0.373	0.214	0.626	0.873	0.448	0.487	0.530	0.451

DELIVERYMETHOD^{C}	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
type^{E}	0.653	0.806	0.548	0.948	0.963	0.934	0.681	0.838	0.574
hasDuration^{L}	0.000	0.000	0.000	0.053	0.053	0.053	0.000	0.000	0.000
hasLocations^{E*}	0.346	0.438	0.286	0.623	0.683	0.573	0.417	0.525	0.346
cardinality	0.818	0.975	0.705	0.948	0.963	0.934	0.834	1.000	0.715
overall	0.485	0.622	0.398	0.782	0.848	0.725	0.536	0.681	0.441

Experimental Group

The lesioned animals were divided into **four groups.**The first group received both OEG and MSC (n = 21).

ExperimentalGroup^C	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
hasOrganismModel^I	0.605	0.614	0.597	0.627	0.634	0.620	0.748	0.748	0.748
hasInjuryModel^I	0.340	0.360	0.322	0.358	0.384	0.334	0.755	0.755	0.755
hasTreatment^{I*}	0.180	0.200	0.164	0.239	0.273	0.213	0.559	0.562	0.557
→ *mainTreatment*	0.164	0.184	0.148	0.210	0.246	0.183	0.531	0.533	0.529
→ *subTreatment*	0.026	0.031	0.022	0.041	0.040	0.041	0.622	0.626	0.619
cardinality	0.838	0.762	0.931	0.903	1.000	0.823	0.897	0.863	0.935
overall	0.377	0.422	0.341	0.450	0.538	0.387	0.720	0.729	0.711

Results, Investigation Method

The **control animals** achieved **BBB scores** of 7.08 +/- 0.24 at the end of the experiment (9 weeks after SCI, 8 weeks after transplantation) but never supported their body weight on their hind legs.
Animals with OEG and MSC co-grafts, even though they received six injections, showed a **statistically significant improvement** 6 weeks after SCI, with **BBB** of 9.18 +/-0.44.

Result^{C}	joint			entity			relation		
Macro	F_1	P	R	F_1	P	R	F_1	P	R
hasInvestigationMethod^{E}	0.229	0.274	0.196	0.714	0.884	0.598	0.711	0.714	0.708
hasTrend^{I}	0.186	0.392	0.122	0.716	0.902	0.593	0.638	0.665	0.613
hasTargetGroup^{I}	0.342	0.471	0.268	0.560	0.623	0.509	0.837	0.827	0.847
hasReferenceGroup^{I}	0.351	0.501	0.270	0.574	0.625	0.531	0.822	0.830	0.813
cardinality	0.628	0.779	0.526	1.000	1.000	1.000	0.860	0.861	0.860
overall	0.333	0.475	0.257	0.576	0.645	0.520	0.830	0.830	0.830

Prototype Applications

- SCIExplorer: http://psink.techfak.uni-bielefeld.de/SCIExplorer/
- Rationalizing Medical Evidence via Argumentation: http://scdemo.techfak.uni-bielefeld.de/ratio-argviz/
- Automatic Grading of Therapies: http://psink.techfak.uni-bielefeld.de/grading/kwon/

Summary

- Background knowledge in knowledge graphs can enhance / improve NLP
- Hypothesis: Scientific Knowledge Graphs will play a key role in reaching FAIR scientific knowledge dissemination
- NLP methods crucial to bootstrap the creation of scientific knowledge graphs
- Challenge: requires deep discipline-specific ontologies and NLP systems to extract fine-grained knowledge following these ontologie

Thanks for your attention!

- Dr. Olivia Sánchez-Graillet
- Hendrik ter Horst
- Frank Grimm
- Christian Grimm
- Dr. Roman Klinger
- Prof. Dr. Werner Müller
- Dr. Nicole Brazda







References

Hendrik ter Horst, Matthias Hartung, Philipp Cimiano, Nicole Brazda, Hans Werner Müller, Roman Klinger:
Learning soft domain constraints in a factor graph model for template-based information extraction. Data Knowl. Eng. 125: 101764 (2020)

Hendrik ter Horst, Philipp Cimiano:
Incorporating Semantic Dependencies Extracted from Knowledge Graphs into Joint Inference Template-Based Information Extraction. ECAI 2020: 2180-2187

Matthias Hartung, Hendrik ter Horst, Frank Grimm, Tim Diekmann, Roman Klinger, Philipp Cimiano:
SANTO: A Web-based Annotation Tool for Ontology-driven Slot Filling. ACL (4) 2018: 68-73

Hendrik ter Horst, Matthias Hartung, Roman Klinger, Nicole Brazda, Hans Werner Müller, Philipp Cimiano:
Assessing the Impact of Single and Pairwise Slot Constraints in a Factor Graph Model for Template-Based Information Extraction. NLDB 2018: 179-190

Hendrik ter Horst, Matthias Hartung, Philipp Cimiano:
Cold-Start Knowledge Base Population Using Ontology-Based Information Extraction with Conditional Random Fields. Reasoning Web 2018: 78-109

Alexander Borowi, Hendrik ter Horst, Matthias Hartung, Veronica Estrada, Nicole Brazda, Hans Werner Müller, Philipp Cimiano:
Ontology-driven Visual Exploration of Preclinical Research Data in the Spinal Cord Injury Domain. SEMANTICS Posters&Demos 2017

Nicole Brazda, Hendrik ter Horst, Matthias Hartung, Cord Wiljes, Veronica Estrada, Roman Klinger, Wolfgang Kuchinke, Hans Werner Müller, Philipp Cimiano:
SCIO: An Ontology to Support the Formalization of Pre-Clinical Spinal Cord Injury Experiments. JOWO 2017

Hendrik ter Horst, Matthias Hartung, Philipp Cimiano:
Joint Entity Recognition and Linking in Technical Domains Using Undirected Probabilistic Graphical Models. LDK 2017: 166-180