International Semantic Web Conference 2011

Conference Summary
by Dr. Dhaval Thakker

IMASH Group, University of Leeds
http://imash.leeds.ac.uk/

We acknowledge the fact that this presentation uses original slides from the authors of the papers and would like to thank them for making it publicly available.
Topics covered:
  - Keynotes:
    - Critical look @ 10 years to learn lessons: Frank
    - Critical friend & "outsider" to tell as it is: Gerhard
    - Completely Outsider to talk about things that might interest SW community: Sandy
  - Talks from:
    - Web of Data
    - Web scale knowledge extraction
    - Social Semantic Web
10 Years of Semantic Web research: Does it work in Theory?

Frank van Harmelen is a professor in Knowledge Representation & Reasoning in the AI department (Faculty of Science) at the Vrije Universiteit Amsterdam. He was one of the co-designers of the OWL Web Ontology Language. He is currently scientific director the LarKC project ([http://www.larkc.eu](http://www.larkc.eu)), aiming to develop the Large Knowledge Collider, a platform for very large scale semantic web reasoning.
At 10 years of age, there is little doubt that the Semantic Web is an engineering success, with substantial (and growing) take-up in business, government and media. However, as a scientific field, have we discovered any general principles?

Have we uncovered any universal patterns that give us insights into the structure of data, information and knowledge, patterns that are valid beyond the engineering of the Semantic Web in its current form?

**Did a decade of Semantic Web work help to discover any Computing Science laws?**
What have we built over the past 10 years

We can characterise what we have built over the past 10 years in 3 parts:

- We built a whole lot of vocabularies (including the languages to represent them, the tools to construct and deploy them, etc)
- We built a whole lot of URI's to name lots of things in the world, in fact, many billions of URI's
- We connected all of these in a very large network
It's been great engineering

But all of these have been mostly treated as one very large engineering exercise.

And it's obvious that as engineers we have succeeded.

- Governments (and not just US and UK anymore)
- BBC (worldcup football web site)
- Retail (GoodRelations),
- Search engines (schema.org)
- Oracle (DB product),
- Publishing industry (e.g. New York Times)
- Electricity de France (personalised energy saving plans for 350,000 customers a day) etc, just look at my [Good News Quiz slide deck at slideshare](https://www.slideshare.net/pacorowncost/19929829-good-news-quiz) for many more examples.
What if we go back and do all again?

- If we would build the Semantic Web again, surely some things would end up looking different, but are there things that would end up looking the same, simply because they have to be that way?
  - something that you would like to change: languages full of angle brackets.
  - but other features of what we've built what turn out in essentially the same way,
  - you would find the same pattern over and over again, every time we ran the experiment.
  - And that is because they are governed by fundamental laws that rule the structure and behaviour of information and knowledge.

- So, has HE discovered any of such laws, such stable patterns that we would rediscover by necessity every time we ran the experiment.
Factual knowledge is a graph

The dominant life-forms in our information space is the graph.

- The vast majority of our factual knowledge consists of simple relationships between things,

- represented as an ground instance of a binary predicate.

- And lots of these relations between things together form a giant graph.

Now this may sound obvious to us in this community, but stating that factual knowledge is a graph is not obvious at all.

For example, if you would ask this question to a DB person, they'd say: factual knowledge is a table. And a logician would say: knowledge is a set of sentences.
Terminological knowledge is a hierarchy

- this law has been rediscovered in knowledge representing and information modelling many times over.

- the details may differ, but the notion of simple hierarchies with property inheritance is widely recognised as the right way to represent terminological knowledge.

- And this observed repeated invention, makes this a much stronger law.

- So to say: this experiment has already been rerun many times in the history of computer science, and this has proven to be a stable finding.
Terminological knowledge is much smaller than the factual knowledge

In fact, it sharply contrasts with a long history of knowledge representation

Traditionally, KR has focussed on small and very intricate sets of axioms: a bunch of universally quantified complex sentences

But now it turns out that much of our knowledge comes in the form of very large but shallow sets of axioms.

Lots of the knowledge is in the ground facts, (not in the quantified formula's)
Terminological knowledge is much smaller than the factual knowledge.

To put this in a slogan:

- "It's the A-box, stupid"
- Knowledge is much more dominated by specific instances than by general rules.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Schema closure</th>
<th>Full closure</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUBM</td>
<td>8 sec</td>
<td>1h15min</td>
<td>562</td>
</tr>
<tr>
<td>Linked Life Data</td>
<td>332 sec</td>
<td>1h5min</td>
<td>11</td>
</tr>
<tr>
<td>FactForge</td>
<td>89 sec</td>
<td>2h45min</td>
<td>111</td>
</tr>
</tbody>
</table>

Apparently, the power of represented knowledge comes from representing a very small set of general rules that are true about the world in general.
Other rules:

Terminological knowledge is of low complexity
Heterogeneity is unavoidable

heterogeneity is solvable
socially,
culturally,
economically
(and a little bit algorithmically)
Publication should be distributed, computation should be centralised

The original dream of this community has sometimes been formulated as turning the Web into a database.

But unfortunately, observations from our 10 year experiment tell us rather the opposite:
● the Web is a good platform for data publication,
● but it's a pretty bad platform for data consumption.

And this is not just us finding this out.
● Google is combining our distributed publishing with their centralised processing,
● Facebook is combing our distributed publishing with their centralised processing,
● Wikipedia, etc.
● So, you might think that centralisation would become a bottle neck. wrong, distribution is the bottle neck,
Knowledge is layered

Contrary to the other laws, this law does not come so much yet from our own observations in this field. But other fields tell us that knowledge is like a set of Russian dolls with one doll nested inside the other.

We know that statements of knowledge need not only refer to the world, but that they may refer to other bits of knowledge, creating a multi-layered structure.

The examples are plenty: we may say that a fact in the world is true, and then we can say

- what the certainty of that statement is,
- or what the provenance of that statement is,
- or what our trust in that statement is
- or at what date that statement was made, etc.

Now curiously enough, there is lots and lots of demand in our community for this kind of layered representation, but our representation language serve this need very poorly. Re-ification can be seen as a failed experiment to obtain such layering, and now people are abusing named graphs because there is nothing better.
Finally... a call for arms

it's an invitation to journal editors and conference chairs to also consider papers that have the ridiculously ambitious aim to discuss one of these laws

and it's also a challenge to you:

Of course we won't really redo the last 10 years of our experiment, but when you do your research and write your papers, try to think about what are the repeatable patterns, these laws, and try to separate the incidental choices you make from the fundamental patterns you are uncovering.
Gerhard Weikum is a Research Director at the Max-Planck Institute for Informatics (MPII) in Saarbruecken, Germany, where he is leading the department on databases and information systems. He is also an adjunct professor in the Department of Computer Science of Saarland University in Saarbruecken, Germany, and he is a principal investigator of the Cluster of Excellence on Multimodal Computing and Interaction. Earlier he held positions at Saarland University in Saarbruecken, Germany, at ETH Zurich, Switzerland, at MCC in Austin, Texas, and he was a visiting senior researcher at Microsoft Research in Redmond, Washington. He received his diploma and doctoral degrees from the University of Darmstadt, Germany.

He claims that not from SW community but from database community and recently focused on "Web Mining".
The Web of Linked Data contains about 25 billion RDF triples and almost half a billion links across data sources; it is becoming a great asset for semantic applications. Linked Data comprises large general-purpose knowledge bases like DBpedia, Yago, and Freebase, as well as many reference collections in a wide variety of areas, spanning sciences, culture, entertainment, and more. Notwithstanding the great potential of Linked Data, this talk argues that there are significant limitations that need to be overcome for further progress. These limitations regard data scope and, especially, data quality. The talk discuss these issues and approaches to extending and enriching Linked Data, in order to improve its scope, quality, interpretability, cross-linking, and usefulness.

"What's wrong with the Linked Data Cloud and why we need few more (right) triples to fix it?"
LOD: Linked Data Cloud

LOD: Linked RDF Triples on the Web

- yago/wordnet: Artist109812338
- yago/wordnet:Actor109765278
- dbpedia.org/resource/Ennio_Morricone
- dbpedia.org/resource/Rome
- rdf.freebase.com/ns/en.rome
- data.nytimes.com/51688803696189142301
- geonames.org/3169070/roma
- imdb.com/title/tt0361748/
- rdf:type
- dbpprop: citizenOf
- owl:sameAs
- N 41° 54' 10" E 12° 29' 22"
Linked data - as it stands

- Size: 30 Billion triples
- Linkage: 500 Million links
- Dynamics: encyclopedic reference data

The problems: Dynamicity, Linkage and Ubiquity
Dynamicity

Where does the data come from for the LOD:

- Maintained, but mostly „static“ reference collections (e.g. geo)
- Periodic exports from curated databases (e.g. gov, bio, music)
- Periodic extraction from Web sources (e.g. encyclopedia, news)
- Tags in social streams and advertisements

Get closer to the data origin:

- RDF engines (Sparql APIs) for production DBs
- view-maintenance by pub-sub push (feeds) (subscribing to the change feeds and keeping up-to-date)
- Deep-Web crawl/query for surfacing of RDF data
- Nothing lasts forever: Even old and „static“ data often needs temporal scope (timepoint, timespan) for proper interpretation with reification, or use quads (quints, etc.)
Linkage: sameAs Links

LOD statistics: 30 Bio. triples, 500 Mio. links
330 Mio. links trivial (ID-based) within pub, within bio
10’s Mio. links near-trivial Dbpedia « Freebase « Yago « GeoNames
sameas.org: 17 Mio. bundles for 50 Mio. URIs
data.nytimes.com: 5000 people, 2000 locations
Way too few for a world with:
1 Mio. people, 10 Mio. locations, 10‘s Mio. species,
6 Mio. books, 2 Mio. movies, 10 Mio. songs, etc. etc.
Ubiquity

- To what extent the web of data (linked data) is connected with the Web of Contents?

- Entities in the LOD linked to "web documents" and having a search engines build around it (indexed and searching).

- Gave few examples of "Sindice".

- RDF data and Web contents need to be interconnected

- RDFa & microformats provide the mechanism

- How do we get the Web RDF-annotated (at large scale)?

- Largely automated, but allow humans in the loop
RDFa?

3. Ubiquity: Web of Data & Other Contents

May 2, 2011

Maestro Morricone will perform on the stage of the Smetana Hall to conduct the Czech National Symphony Orchestra and Choir. The concert will feature both Classical compositions and soundtracks such as the Ecstasy of Gold. In programme two concerts for July 14th and 15th.

<html ... May 2, 2011
<div typeof=event:music>

<span id="Maestro_Morricone">
Maestro Morricone
</span>
<a rel="sameAs" resource="dbpedia.../Ennio_Morricone ">
</a>
</span>
...
<br/>

<span property = "event:location" >
Smetana Hall </span>
...

<span property="rdf:type"
resource="yago:performance">
The concert </span> will feature
...

<span property="event:date"
content="14-07-2011"/>

July 1
</div>
Named-Entity Disambiguation: State-of-the-Art

Literature:
• Razvan Bunescu, Marius Pasca: EACL 2006
• Silviu Cucerzan: EMNLP 2007
• David Milne, Ian Witten: CIKM 2008
• S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti: KDD 2009
• G. Limaye, S. Sarawagi, S. Chakrabarti: VLDB 2010
• Paolo Ferragina, Ugo Scaella: CIKM 2010
• Mark Dredze et al.: COLING 2010
• Johannes Hoffart et al.: EMNLP 2011
etc. etc.

Online tools:
https://d5gate.ag5.mpi-sb.mpg.de/webaida/
http://tagme.di.unipi.it/
http://spotlight.dbpedia.org/demo/index.html
http://viewer.opencalais.com/
http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/
NED: Experimental Evaluation

Benchmark:
• Extended CoNLL 2003 dataset: 1400 newswire articles
• originally annotated with mention markup (NER), now with NED mappings to Yago and Freebase
• difficult texts:
  … Australia beats India … ® Australian_Cricket_Team
  … White House talks to Kreml … ® President_of_the_USA
  … EDS made a contract with … ® HP_Enterprise_Services

• J. Hoffart et al.: Robust Disambiguation of Named Entities in Text, EMNLP 2011
http://www.mpi-inf.mpg.de/yago-naga/aida/
Interesting Research Issues: NED

- More efficient graph algorithms (multicore, etc.)
- Allow mentions of unknown entities, mapped to null
- Leverage deep-parsing structures, leverage semantic types
- Short and difficult texts: tweets, headlines, etc.
  fictional texts: novels, song lyrics, etc.
  incoherent texts
- Disambiguation beyond entity names:
  coreferences: pronouns, paraphrases, etc.
  common nouns, verbal phrases (general WSD)
Why a Few Triples More?

Linked Data is great!
But still in its infancy
Need to add triples to capture further issues

Dynamics:
Where is the live data?
(Deep-Web) sources ® feeds, pub-sub, … ?  
® fresh & versioned triples

Linkage:
Where are the links in Linked Data?
LOD ® entity mapping ® user ® community

Ubiquity:
Where are the paths between the Web-of-Data and the Web?
RDFa ® entity disambiguation ® authoring
Take home message

- "we don’t need more triples, we need the right triples"
- "remove the half triples, clean up some and have to more address (above) to move higher in semantic chain"
- Linked data is best if it has many good links
- New & rich contents mostly in traditional Web
- Create sameAs links in (X)HTML contents, via RDFa
- Links for named entities give best mileage/effort
- Methods & tools greatly advanced & gradually maturing
- Keep human in the loop, embed NED in authoring tools
Extracting Binary relationships extraction and learning
- Very interesting and Challenging area
- Computationally expensive
- Big "Join" tasks
- Approach has been: get the unary relationship - entity detection (disambiguation) and move to this.
- Get the "unary" stuff to near human quality and move to "binary"

Is sameAs appropriate? may be "skos: broaderMatch " or "skos: closerMatch" is better?
- agree, limitations "sameAs" - and semantic shifts - is calculus same as mathematics? almost the same but not exactly same. In some cases it is straightforward (especially persons, locations) but when it comes to more philosophical - it does not always work.
Building A Nervous System for Society: The ‘New Deal on Data’ and how to make Health, Financial, Logistics, and Transportation Systems Work

Alex 'Sandy' Pentland

Alex 'Sandy' Pentland directs MIT’s Human Dynamics Laboratory and the MIT Media Lab Entrepreneurship Program, and advises the World Economic Forum, Nissan Motor Corporation, and a variety of start-up firms. Sandy is among the most-cited computational scientists in the world, and a pioneer in computational social science, organizational engineering, mobile computing, image understanding, and modern biometrics. His research has been featured in Nature, Science, the World Economic Forum, and Harvard Business Review, as well as being the focus of TV features including Nova and Scientific American Frontiers. His most recent book is ‘Honest Signals,’ published by MIT Press.

Alex ‘Sandy’ Pentland, was named among “The World’s 7 Most Powerful Data Scientists” by Tim O’Reilly in Forbes Magazine:
Sensor based applications

- There is "behavioral demographics"
- If you hang around certain people and like certain sort of restaurants and bars and shows then you are an iphone guy and
- if you like other types of things you are a blackburry guy.

- It is vital for advertising and marking people - the factor of 10 compared to demographics - if you just consider the demographics (age, gender, where you live) - you are missing the boat.

- He is a Computational social scientist - behavioral patterns that can affect health - if you are behaving in certain ways you are going to get certain disease and that information is worth a billion dollars. If you know that now, let alone money or you can save your life.
Exploiting this

Award someone for being more active, then instead given someone around you awards (that worked twice more effectively).

What he comes to say is: Quoting the EU commisioner for consumer affairs,

"Personal data is new oil of the internet and the new currency of the digital world"

That should scare you as well. You probably knew this going on but not this intimately and how it can manipulate people. It was also occurred to him - with the work he did - and he wanted to do something about.
Any solution?

Rethinking Personal Data:
Big companies, World Economic forums, MIT and Harvard law lab

There is a way to put the rabbit back in the hat?
No. But are there maybe ways to control.

Need all the parties to benefit: government, companies, you. "You can’t build effective transport, health, smart grid unless you have access to information"

New deal of data
Handshake agreement on what the future looks like so that everybody can nod their head and say I can live with that.
Who you can trust with data

Companies store it, regulators scold them for holding too much data, and companies also have liabilities.

Government? Not going to fly in many countries.

This is "Data" that is not in the wild already but is slowly getting out there. Things we can not do with internet, Google etc. we cant do much as we screwed up.

We are talking about mobile carrier, your bank and your doctor etc.

So who can be trusted? (and we are looking in the future).

"Its you. You can control and monetize the data"
Any time you sign up wireless carrier, doctors and they dump the data (may be in cloud) and there is trust authority to control who gets what?

Ecology of authorities working for you: Trust authorities (best practicencies and security of store), identity providers etc.

May be similar to the Amazon: Amazon runs some trust frameworks in the back-end to detect frauds.

All these information should be treated as assets (asset class). To give you control of your data so that it becomes asset that can be used for various applications. Also comes all the other things "auditing" etc. but it is cyber market so doesnt have to be complicated and can be transparent.
From Code to Contract

Code: Offer -- Reply Semantics

Personal Data Store  

Legal Contract 'Shadow'

Personal Data Store

Offers and replies also certified by trust network
No data exchanged, only actions
Where can semantics come into this?

- I have a data store and want to talk to you and send a little piece of codes and the other one replies and then you have binding contract that is machine interpretable and also NLP and binding in all the jurisdiction.

- To derive maximum value individuals need interoperable ontologies (market forces for semantic web - 80% of the digital economy)
- All the communication can be done using Semantics (p2p)
- can estimate risk from exchanges in advance.
Take home message

- The world is becoming full of sensor networks producing information about people, location and call data from cell phones, digital transactions, RFID and more.

- Information about people is power and money (all the basic infrastructure for our society: power grid, finance, advertising, media you name it, is at the end of the day driven by human demand and reactivity and this is the stuff you get from the personal data - thats why it is oil of digital economy)

- This information is creating a data-driven world, driving revolution in government, management and businesses. Make them asset, liquid etc. so that you have some control over it and structure around it. "It might not be the best thing in the world but better than where we started"
He finished his keynote by introducing ID Cubed, an institution to design, code, test and scale a new data ecosystem of innovative data-driven enterprises.

http://idcubed.org/

We are a loose assembly of entrepreneurs, technologists, attorneys, scientists, hackers, artists, data analysts, educators and activists who see open source software as a powerful vehicle for designing a new generation of transparent, networked enterprises and institutions.
Q & A

In addition to ontology - also "rule based technology" to do what you are trying to do.

Discussions they have is: How do you construct rule sets, who provides them? Can be group of people providers of rule sets.

Laws doesn't need to be "much" changed.

Assumption is that people are "capable" of managing data. There can be "parties" that can be doing that.
Summary of the summary

- Topics covered:
  - Keynotes:
    - Critical look @ 10 years to learn lessons: Frank
    - Critical friend & "outsider" to tell as it is: Gerhard
    - Completely Outsider to talk about things that might interest SW community: Sandy
  - Talks from:
    - Web of Data
    - Web scale knowledge extraction
    - Social Semantic Web
Web of Data
Connecting the Dots: A Multi-pivot Approach to Data Exploration

The focus of the paper is "Data Browsers". Purpose of such data browsers is to help users identify and query data effectively without being overwhelmed by large complex graphs of data.

The paper claims that: the semantic data browsers "can play the role of tackling the problem of making sense of rich and complex domains".

Challenge: "How to understand what's in an unknown dataset?" "How do I find and get the data I need"

Authors: Igor Popov, Wendy Hall, Nigel Shadbolt (Sothampton)
Pivoting

Authors lists the flaws in the existing solutions that they say are using "Pivoting (or set-oriented browsing)" - a many-to-many graph browsing technique that allows users to navigate the graph by starting from a set of instances followed by navigation through common links.
Multi-Pivot

The author build on the strengths of pivoting and present a multi-pivot approach (embodied in tool called Visor).

Visor allows users to explore from multiple points in the graph, helping users connect key points of interest in the graph on the conceptual level, visually occluding the remainder parts of the graph, thus helping create a road-map for navigation.
Multi-Pivot in Visor

multi-pivot approach

- No instance resources shown, navigation on the conceptual level (find a possible path first - instance data on demand)
- Start from multiple points in the graph (the system helps you find the connections)
- Allow to cherry pick data out of the generated sub-domain
Visor Tool
Evaluation

The "navigation" aspect (with respect to scaffolding) is interesting: while stating the state of the art they mention "navigation is typically supported in a single direction " and "immediate instance level exploration is regularly preferred without gaining familiarity with the domain or setting the exploration context first"

Authors carried out and reports on an user study to demonstrate the viability of their approach.

The evaluation carried out in this work involves: users were required to perform set of tasks using the tool and then observed by "think aloud" in terms of what they did and decisions they made.
The paper nicely articulates one of the major problems facing the uptake of linked data - large collections of descriptions becoming increasingly lengthy by the linkages and hard to comprehend for quick identification of entities, matching two entities manually.

The solution offered in the paper is a way of selecting a few "central data elements" (centrality-based) that are most useful in characterizing an entity. $k$ top-ranked features.

As the title of the paper suggests, they focus on the "Relatedness" and "Informativeness" to decide the centrality of data elements. The paper proposes a variant of the random surfer model.

Authors: Gong Cheng (Nanjing Uni), Thanh Tran (Karlsruhe)
Random Surfer Model : Page Rank

- Simulating a random surfer’s behavior who navigates from node to node
- Two types of action
  - Following a random edge (with a uniform probability distribution)
  - Jumping at random (with a uniform probability distribution)
- Ranking based on the stationary distribution of such a Markov chain
Random Surfer Model in RELIN

- This well-known model is used as the basis to support the idea of incorporating central elements into the summary.

- However, it is revised by a more specific notion of centrality, called RELIN, where the computation of central elements involves relatedness (or similarity) between elements as well as their informativeness, i.e. the amount of information carried that helps to identify the entity.

- Authors also propose an implementation of these notions of relatedness and informativeness that exploits the semantic information captured by the graph structure of the data as well as the labels of nodes and edges.
RELIN: RELatedness and INformativeness-based centrality

- Two kinds of action
  - Relational move --- more likely to a feature that carries related information about the theme currently under investigation
  - Informational jump --- more likely to a feature that provides a large amount of new information for clarifying the identity of the underlying entity
- Two non-uniform probability distributions
Evaluation

Intrinsic evaluation --- design

- Task
  - To manually construct ideal entity summaries as the gold standard
- Participants
  - 24 students majoring in computer science
- Test cases
  - 149 entity descriptions randomly selected from DBpedia 3.4
- Assignment
  - 4.43 participants per entity description
- Output
  - Top-5 features
  - Top-10 features
Evaluation

**Intrinsic evaluation --- results**

- Metric: overlap between summaries
- Agreement between participants about ideal summaries
  - 2.91 when $k=5$
  - 7.86 when $k=10$
- Quality of summaries computed under different approach settings

<table>
<thead>
<tr>
<th>Baselines</th>
<th>$k = 10$</th>
<th>$k = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OntoSum</strong></td>
<td>3.69</td>
<td>1.01</td>
</tr>
<tr>
<td><strong>RandomRank</strong></td>
<td>3.36</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RELIN, with $\lambda = 0.00$</td>
<td>3.58</td>
<td>1.61</td>
</tr>
<tr>
<td>RELIN, with $\lambda = 0.15$</td>
<td>3.84</td>
<td>1.73</td>
</tr>
<tr>
<td>RELIN, with $\lambda = 0.50$</td>
<td>4.40</td>
<td>1.99</td>
</tr>
<tr>
<td>RELIN, with $\lambda = 0.85$</td>
<td>4.88</td>
<td>2.29</td>
</tr>
<tr>
<td>RELIN, with $\lambda = 1.00$</td>
<td>4.86</td>
<td>2.40</td>
</tr>
</tbody>
</table>
Web scale knowledge extraction
Automatic Open-Domain Question Answering: Deep QA in Watson

A Long-Standing Challenge in Artificial Intelligence to emulate human expertise

Given
– Rich Natural Language Questions
– Over a Broad Domain of Knowledge

Deliver
– Precise Answers: Determine what is being asked & give precise response
– Accurate Confidences: Determine likelihood answer is correct
– Consumable Justifications: Explain why the answer is right
– Fast Response Time: Precision & Confidence in <3 seconds

Authors: Aditya Kalyanpur, James Fan, Chris Welthy,
Jeopardy

- **Jeopardy!** is an American quiz show
  - 1964 – Today
- **answer-and-question** format
  - contestants are presented with clues in the form of answers
  - must phrase their responses in question form.

- **Example**
  - Category: General Science
  - Clue: When hit by electrons, a phosphor gives off electromagnetic energy in this **form**
  - Answer: What is **light**?
DeepQA generates and scores many hypotheses using an extensible collection of Natural Language Processing, Machine Learning and Reasoning Algorithms. These gather and weigh evidence over both unstructured and structured content to determine the answer with the best confidence.
Leveraging Community-built Knowledge for **Type Coercion** in

Typing in Jeopardy!

- It's basically a big **kettle** with a close-fitting lid, used to cook pot roasts & stews
- Category: EUROPEAN NATIONALITIES
- Answer: **Dutch Oven**

- Unlucky things happen at Camp Crystal Lake in this 1980 **scarefest**
- Category: MOVIE CALENDAR
- Answer: **Friday the 13th**

- Wanted for general evil-ness; last seen at the Tower of Barad-Dur; it's a **giant eye**, folks. Kinda hard to miss
- Category: LITERARY CHARACTER APB
- Answer: **Sauron**
Closed Domain Type Checking

- Used in Traditional QA Systems
  - Based on “Type And Generate” Principle
  - Focus on a pre-determined set of interesting types
    - People, Places, Organizations, Dates
  - For these types, run Named Entity Recognizers (NER) over text corpus
    - People: {“Einstein”, “Sir I. Newton”..}
    - Places: {“Germany”, “UK”..}
    - Dates: {“1885”, “3rd April 1715”..}
  - At run-time, given a question, detect lexical answer type (LAT) and:
    - Generate candidates from pre-compiled list of LAT instances

Limitations
- Highly brittle – QA system breaks down if type not recognized
- Limited Coverage – need to enumerate all relevant types beforehand
- Dependent on quality of NERs used
Open Domain Type Coercion (TyCor)

- Approach taken in DeepQA
  - Based on "Generate-and-Type" Principle
    - Generate candidates without considering answer type (LAT)
    - Later check whether candidate can be coerced into LAT using a suite of Type-Coercion Algorithms
  - Use machine-learning to combine information from TyCors

- Advantages
  - More flexible as QA system does not break down if type is not detected or meaningful
  - Much wider type coverage possible using a variety of sources and analytics for TyCor
How TyCor Fits in DeepQA

IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE
How TyCor Fits in DeepQA

IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE.
TyCor Framework

- **Problem**: Compute type match for candidate w.r.t. LAT
  - Both candidate and LAT expressed as **Strings**
- **4 Steps:**
  1. EDM: Entity Disambiguation and Matching
  2. TR: Type Retrieval
  3. PDM: Predicate Disambiguation and Matching
  4. TA: Type Alignment

```
  "JFK"
  (Cand) -> EDM: Candidate → Instance ➔ Wikipedia:John_F_Kennedy_International (0.7)

  "facility"
  (LAT) -> PDM: LAT → Type ➔ Yago:Airport (1.0)

  TR: Instance → Type ➔ WN:Facility (0.9)

  TA: Compare LAT-type and Instance-type

Helps infer:
- "Ramadan" is a "month"
- "Interpreter" is a "job"
- "Castling" is a "maneuver"
- "Sauron" is an "eye"
```
Evaluating TyCors on Ground Truth

Benchmark creation:
- Annotated Top 10 Candidates for 1615 Jeopardy! Questions
  - Judgement: Does candidate match LAT – Y/N?
- Total <LAT, Candidate> Pairs for Testing: 25,991 (due to multiple LATs)
Evaluating TyCors on end-to-end QA

- Two Watson Configurations:
  1. **Watson-LITE**: Cand. Gen + Merging + Ranking (NO Answer Scoring)
  2. **Watson-FULL**: LITE + All Answer Scoring

- All gains over “No TyCor” are statistically significant
- Combining all 3 TyCors better than any one (Net gain: 5-6%)
Bootstrapping the Linked Data Web

Motivation:
- Most knowledge bases extracted from (semi)-structured data
- Only 15-20% of information in structured data
- Semantic Web <-> `Document Web
- How can we extract data from the document-oriented web?

Related Work:
- ReadTheWeb Project: N(ever) E(nding) L(anguage) L(earner)
- PROSPERA: Scalable Knowledge Harvesting with High Precision and High Recall

Authors: Daniel Gerber (Leipzig)
Approach:

The BOA approach
WHERE {
    ?x rdf:type dbpedia-owl:[Organisation|Person|Place] .
}

http://dbpedia.org/resource/Google
“Google”
http://dbpedia.org/ontology/subsidiary
http://dbpedia.org/resource/YouTube
“Youtube”

http://dbpedia.org/ontology/Company
http://dbpedia.org/ontology/Company
(1) Set of entities s and o connected through p
(2) Find all sentences which contain s and o
(3) Replace labels with variables (D?, R?)

BOA pattern:

dbpedia-owl:spouse
"D? with his wife R?"

BOA pattern mapping:

dbpedia-owl:spouse
"D? with his wife R?"

dbpedia-owl:spouse
"D? and her husband R?"

dbpedia-owl:spouse
"D? and his wife R?"
Social Semantic Web
Modeling and analysis of user behavior in online communities

The Utility of Online Communities

- Online communities yield value in terms of:
  - Idea generation
  - Customer support
  - Problem solving

- Managing and hosting communities can be:
  - Expensive
  - Time-consuming

- Large investments in communities, therefore they must:
  - flourish and remain active
  - remain... ‘healthy’
Gauging Health

- How can we gauge community health?
  - Post Count?
  - User Count?
  - Communication/Interaction?
  - Behaviour?

- Domination of one behaviour could lead to churn
  - Preece, 2000

- Behaviour in online community is influenced by the roles that users assume
  - Preece, 2001

- To provide health insights we need to monitor behaviour over time
  - Combined with basic health metrics (e.g. post count)
1. Monitor and capture member activities
2. Analyse emerging behaviour over time
3. Understand the correlation of behaviour with community evolution
4. Learn when to intervene to influence the community
Contributions

- Ontology to model behavioural roles and behaviour features
  - Capturing time stamped user attributes

- Method to infer user roles in online communities
  - Using semantic rules

- Analysis of community health through role composition
  - Identifying composition patterns for healthy communities
Modelling and Analysis of User Behaviour in Online Communities
Behaviour Features

- In-degree Ratio
  - Proportion of users that reply to user $i$
- Posts Replied Ratio
  - Proportion of posts by $i$ that yield a reply
- Thread Initiation Ratio
  - Proportion of threads started by $i$
- Bi-directional Threads Ratio
  - Proportion of threads where $i$ is involved in a reciprocal action
- Bi-directional Neighbours Ratio
  - Proportion of $i$'s neighbours with whom a reciprocal action has taken place
- Average Posts per Thread
  - Mean number of posts in the threads that $i$ has participated in
- Standard Deviation of Posts per Thread
  - Standard deviation of posts in the threads that $i$ has posted in
Community Roles

Elitist
Grunt
Joining Conversationalist
Popular Initiator
Popular Participant
Supporter
Taciturn
Ignored

### Table 1. Roles and the feature-to-level mappings

<table>
<thead>
<tr>
<th>Role</th>
<th>Feature</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elitist</td>
<td>In-Degree Ratio</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>Bi-directional Threads Ratio</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>Bi-directional Neighbours Ratio</td>
<td>low</td>
</tr>
<tr>
<td>Grunt</td>
<td>Bi-directional Threads Ratio</td>
<td>med</td>
</tr>
<tr>
<td></td>
<td>Bi-directional Neighbours Ratio</td>
<td>med</td>
</tr>
<tr>
<td></td>
<td>Average Posts per Thread</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>STD of Posts per Thread</td>
<td>low</td>
</tr>
<tr>
<td>Joining Conversationalist</td>
<td>Thread Initiation Ratio</td>
<td>low</td>
</tr>
<tr>
<td>Popular Initiator</td>
<td>In-Degree Ratio</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>Thread Initiation Ratio</td>
<td>high</td>
</tr>
<tr>
<td>Popular Participants</td>
<td>In-Degree Ratio</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>Thread Initiation Ratio</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>Average Posts per Thread</td>
<td>med</td>
</tr>
<tr>
<td></td>
<td>STD of Posts per Thread</td>
<td>med</td>
</tr>
<tr>
<td>Supporter</td>
<td>In-Degree Ratio</td>
<td>med</td>
</tr>
<tr>
<td></td>
<td>Bi-directional Threads Ratio</td>
<td>med</td>
</tr>
<tr>
<td></td>
<td>Bi-directional Neighbours Ratio</td>
<td>med</td>
</tr>
<tr>
<td>Taciturn</td>
<td>Bi-directional Threads Ratio</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>Bi-directional Neighbours Ratio</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>Average Posts per Thread</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>STD of Posts per Thread</td>
<td>low</td>
</tr>
<tr>
<td>Ignored</td>
<td>Posts Replied Ratio</td>
<td>low</td>
</tr>
</tbody>
</table>
Constructing Rules

Structural, social network, reciprocity, persistence, participation

Feature levels change with the dynamics of the community

Run rules over each user's features and derive the community role composition

Based on related work, we associate roles with a collection of feature-to-level mappings e.g. in-degree -> high, out-degree -> high

Modelling and Analysis of User Behaviour in Online Communities
Analysis of Community Health

- How is community role composition associated with activity?

- Dataset
  - Irish community message board: Boards.ie
  - All posts used from 2004 – 2006
  - Selected 3 forums for analysis
    - F246: Commuting and Transport
    - F388: Rugby
    - F411: Mobile Phones and PDAs

- Measured at 12-week increments:
  - Forum composition (% of roles)
    - E.g. 20% elitists, 10% grunts, etc
  - Number of posts
Modelling and Analysis of User Behaviour in Online Communities
Can we predict community health from role composition?

1. Predict either an increase or decrease in activity
   - Features: roles and percentages
   - Class label: increase/decrease
   - Performed 10-fold cross validation with J48 decision tree

2. Predict post count from role composition
   - Independent variables: roles and percentages
   - Dependent variable: post count
   - Induced linear regression model and assessed the model
### Table 3. Linear regression model induced from the forum composition of f388

<table>
<thead>
<tr>
<th>Role</th>
<th>Est' Coefficient</th>
<th>Standard Error</th>
<th>t-Value</th>
<th>P(x &gt; t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joining Conversationalist</td>
<td>69.20</td>
<td>43.82</td>
<td>1.579</td>
<td>0.1751</td>
</tr>
<tr>
<td>Popular Initiators</td>
<td>173.41</td>
<td>54.72</td>
<td>3.169</td>
<td>0.0248 **</td>
</tr>
<tr>
<td>Taciturns</td>
<td>-135.97</td>
<td>101.91</td>
<td>-1.334</td>
<td>0.2397</td>
</tr>
<tr>
<td>Supporters</td>
<td>-266.53</td>
<td>109.60</td>
<td>-2.432</td>
<td>0.0592 *</td>
</tr>
<tr>
<td>Elitists</td>
<td>-105.19</td>
<td>55.88</td>
<td>-1.882</td>
<td>0.1185</td>
</tr>
<tr>
<td>Popular Participants</td>
<td>372.44</td>
<td>103.24</td>
<td>3.608</td>
<td>0.0154 **</td>
</tr>
<tr>
<td>Ignored</td>
<td>-75.69</td>
<td>33.39</td>
<td>-2.267</td>
<td>0.0727 *</td>
</tr>
</tbody>
</table>

Summary: Res. St Err: 311.5, Adj R²: 0.8514, F₇,₅: 10.82, p-value: 0.0092

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1
Findings

1. Active communities contain more Elitists and Popular Participants

2. Unhealthy community contain more Tactiturns and Ignored users

3. Communities exhibit idiosyncratic compositions

4. A stable, mixed composition increases activity
Linked Data Across Universities: An Integrated Video Lectures Dataset

Miriam Fernandez, Mathieu d’Aquin, Enrico Motta
Goal

- Interlink educational information across universities through the use of LD principles and technologies
- Create a consolidated dataset of video material
• Use Linked Data
  1. Select and extract educational information from various sources
  2. Reuse well-known vocabularies to describe and structure the previously extracted data
  3. Integrate the educational information in a common categorization scheme
Vocabularies

- **Dublin Core**

- **FOAF**

- **The W3C ontology for media resources**
  - [http://www.w3.org/TR/mediaont-10/](http://www.w3.org/TR/mediaont-10/) (ma)

- **The Media Vocabulary**
  - [http://payswarm.com/vocabs/media](http://payswarm.com/vocabs/media) (media)

- **The Nice Tag Ontology**
Main design decisions:

1. The selected base URI is: http://linkeduniversities.org
2. VideoLectures objects are represented as media:Recording, authors are represented as foaf:Person
3. The video title is duplicated in the properties rdfs:label & dcterms:title

4. The set of tags and categories associated to a video is represented by the nt:isRelatedTo property
5. The assigned classification in the unified search space is represented by the dcterms:subject property
(1) Extract the information from the video lecture

(2) Generate an HTML document

(3) Provide the document to the textwise classification service

(4) Obtain the ODP document classification

Reference/Knowledge_Management (id=495), w=0.71
Results

- New LD educational dataset
  - More than **14,000** video lectures from **27** different institutions sharing the same search/browsing space
  - Quality
    - Coverage: **98%** of the videos were assigned at least one ODP category
    - Correctness and Specialization:
      - **89%** correct classifications
      - **51%** specialized classifications
      - Fleiss’ kappa statistic $k=0.71$
Discussion

• Lessons Learned
  - LD is simple, getting data and remodelling it is a high-cost process
  - Need to agree on a set of vocabularies
  - Need to agree on common searchable space
  - Need to establish qualitative criteria and quantitative evaluation measures
  - Educational LD is not about a killer app, but about multiple small things that are made easier!