ISWC 2013
Sydney, Australia

SemStats 2013

Towards a Vocabulary for Incorporating Predictive Models into the Linked Data Web

Evangelos Kalampokis, Areti Karamanou, Efthimios Tambouris, Konstantinos Tarabanis

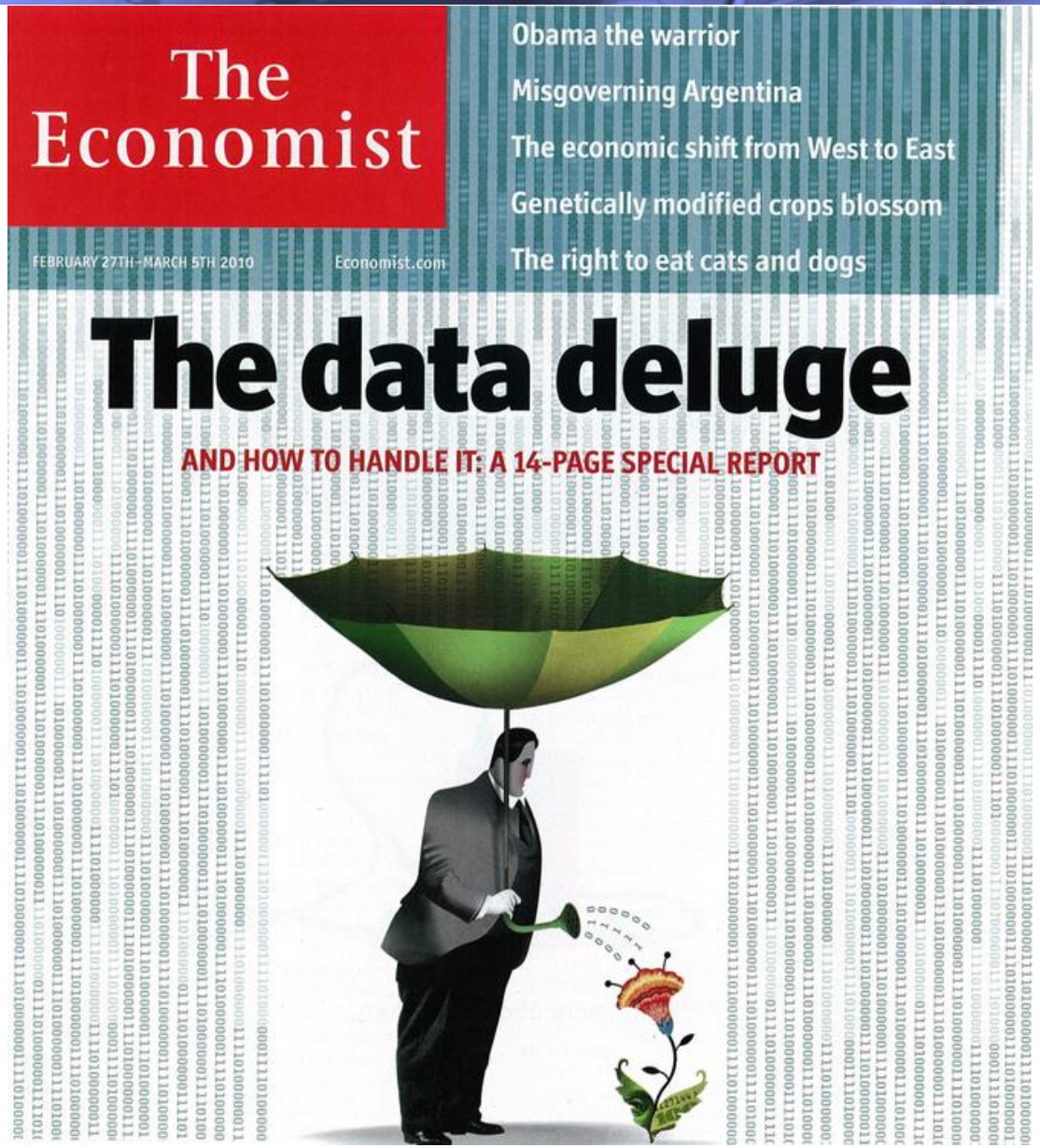
CERTH
THE CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

Information Systems Lab
University of Macedonia - Greece

http://islab.uom.gr

Objective

- *"To propose an RDF Schema vocabulary, named the **Linked Statistical Models (limo) vocabulary**, that will enable the incorporation of descriptions of **predictive models** into the **Linked Data Web** and establish links to other resources such as datasets, other models, academic articles and studies."*

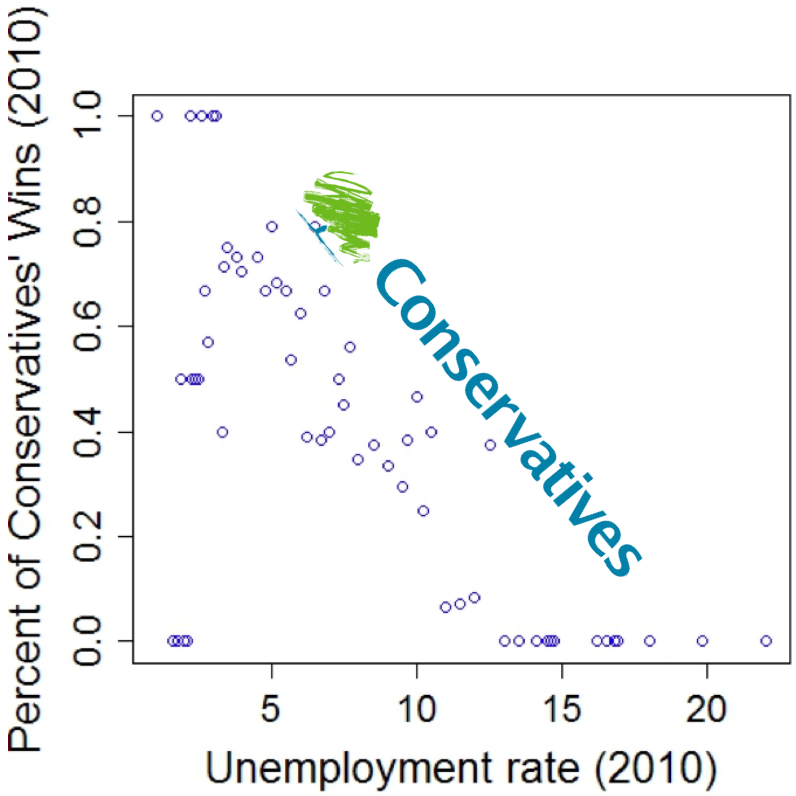


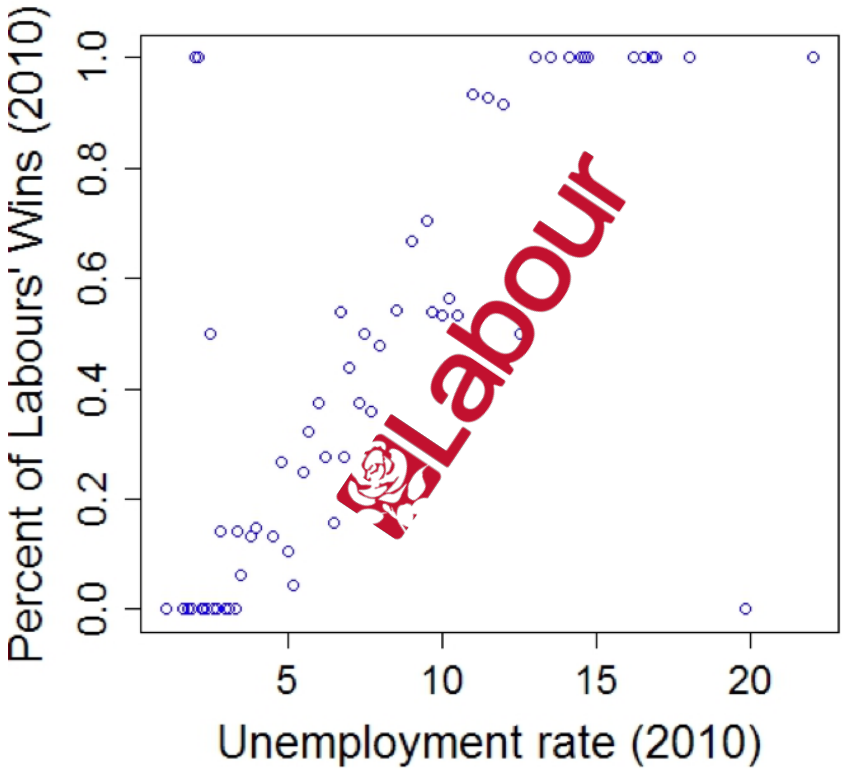
Data analytics

- **Easy access** to large amounts of data
- Combine data and perform **data analytics**
- Create **statistical or data mining models** for understanding and describing various problem areas and domains

http://www.flickr.com/photos/seandreilinger/2587606107

DATA.GOV.UK Beta
Opening up Government





E. Kalampokis, E. Tambouris and K. Tarabanis (2013) Linked Open Government Data Analytics, M.A. Wimmer, M. Janssen, and H.J. Scholl (Eds.): EGOV 2013, LNCS 8074, pp. 99-110. IFIP International Federation for Information Processing.


Home News World Sport Finance Comment Blogs
Film Music Art Books TV and Radio Theatre Comedy Culture Travel Life Women Fashion
Film Reviews Cinema Trailers Coming Soon Talking Movies Dance Opera Photography Film Life
Interviews DVDs
HOME » CULTURE » FILM » FILM NEWS
Can Google predict the next box office flop?
Google claims it can predict box office success with
searches for trailers made a month before a film
Taylor Kitsch,
BBC NEWS TECHNOLOGY
Home UK Africa Asia Europe Latin America Mid-East US & Canada Business Health Sci/Env
News Sport Weather Capital Culture
6 April 2011 Last updated at 09:44 GMT
Twitter predicts future of stocks
Twitter may not yet have found a
make money for itself but it
job of generating
research
Share
turn their screens to Twitter?
users.
Related Stories
Tweeting the American Dream
Twitter marks its fifth birthday
The Economist
World politics Business & finance Economics Science & technology Culture
Technology Quarterly: Q2 2011
Monitor
Can Twitter predict the future?
Internet forecasting: Businesses are mining online messages to unearth consumers' moods—and even make market predictions
Jun 2nd 2011 | From the print edition

Controversial results

- 11 models aiming at **predicting elections results** using Social Media (SM) related variables
- Only 3 of them included **sentiment** related variables
- Only 1 of them employed predictive analytics evaluation methods
- 6 supported SM predictive power while 5 challenged it

http://www.flickr.com/photos/cainandtoddbenson/717520970

E. Kalampokis, E. Tambouris and K. Tarabanis (2013) Understanding the Predictive Power of Social Media, Internet Research, Vol.23, No.5, pp. 544-559

Understanding the predictive power of SM

- 52 empirical studies that exploit Social Media for predictions
- The predictive power of a model is directly related to:
  - Selected predictors
  - Statistical or data mining method used
  - Evaluation method employed
  - Datasets selected
  - Approaches used to collect, filter and process data

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1066-2243.htm

INTR
23,5

544

Received 15 June 2012
Revised 4 February 2013
Accepted 11 February 2013

Understanding the predictive power of social media

Evangelos Kalampokis
Information Systems Laboratory, University of Macedonia, Thessaloniki, Greece and Information Technologies Institute, Centre for Research & Technology – Hellas, Thessaloniki, Greece
Efthimios Tambouris
Department of Technology and Management, University of Macedonia, Naousa, Greece and Information Technologies Institute, Centre for Research & Technology – Hellas, Thessaloniki, Greece, and
Konstantinos Tarabanis
Department of Business Administration, University of Macedonia, Thessaloniki, Greece and Information Technologies Institute, Centre for Research & Technology – Hellas, Thessaloniki, Greece

Abstract

Purpose – The purpose of this paper is to consolidate existing knowledge and provide a deeper understanding of the use of social media (SM) data for predictions in various areas, such as disease outbreaks, product sales, stock market volatility and elections outcome predictions.

Design/methodology/approach – The scientific literature was systematically reviewed to identify relevant empirical studies. These studies were analysed and synthesized in the form of a proposed conceptual framework, which was thereafter applied to further analyse this literature, hence gaining new insights into the field.

ed framework reveals that all relevant studies can be decomposed into a small erent approaches can be followed in each step. The application of the framework idings. For example, most studies support SM predictive power, however, more udies infer predictive power without employing predictive analytics. In addition, here is a clear need for more advanced sentiment analysis methods as well as search terms for collection and filtering of raw SM data. he proposed framework enables researchers to classify and evaluate existing tifically rigorous new studies and to identify the field's weaknesses, hence h directions.

orks, Data analysis, Open data, World Wide Web

paper

Internet Research

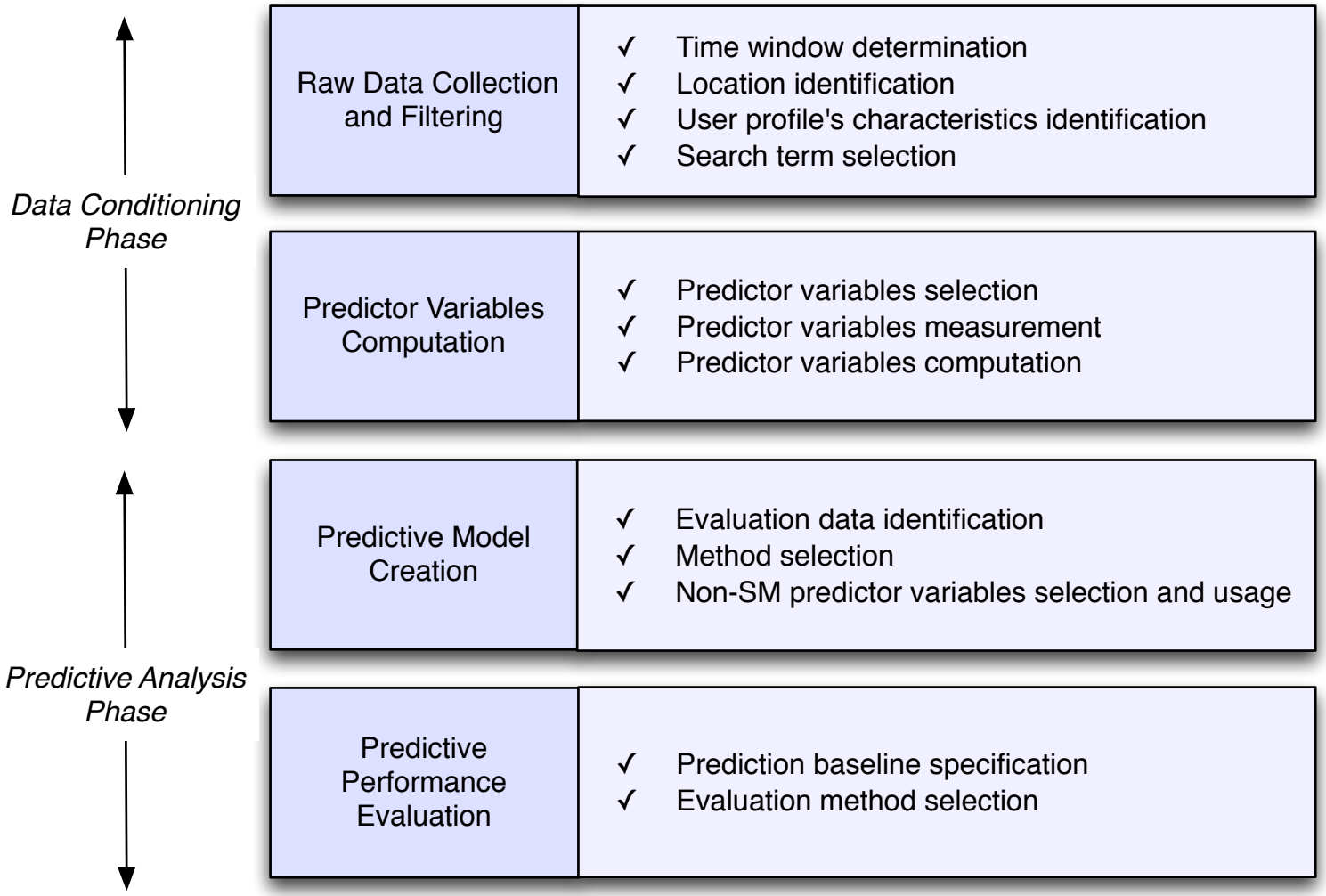
use of social media (SM) has dramatically increased with millions ssive amounts of data every day. As of September 2012, the online plication Facebook reached one billion monthly active users, while

to thank the anonymous reviewers for their valuable comments that have nt of manuscript's quality. They would also like to acknowledge that the s paper has been partially funded by the European Union through the pen Linked Data Platform for Semantically-Interconnecting Online, Social rporate Brand and Market Sector Reputation Analysis, FP7-SME-2011

E. Kalampokis, E. Tambouris and K. Tarabanis (2013) Understanding the Predictive Power of Social Media, Internet Research, Vol.23, No.5, pp. 544-559

- Why don’t we **reuse** all this information?

Social Media Data Analysis Process for Predictive Analytics

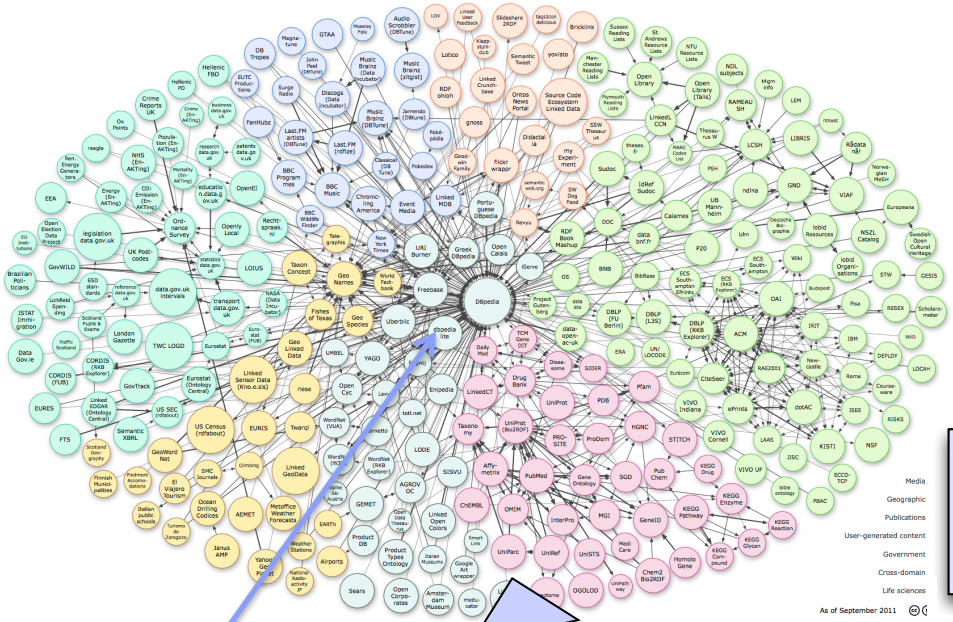


E. Kalampokis, E. Tambouris and K. Tarabanis (2013) Understanding the Predictive Power of Social Media, Internet Research, Vol.23, No.5, pp. 544-559

Reuse of Descriptions of Predictive Models

- Discover **variables** that a predictive relationship between them have been suggested by a model
- Discover **predictor variables** that are connected to the same **response**
- Discover **statistical or data mining methods** used in certain cases
- Discover **datasets** used or could be reused in existing or new models
- Discover **predictive models** that could be reused (e.g. for baseline predictions or with different data)

- This is where **Linked Data** comes in …



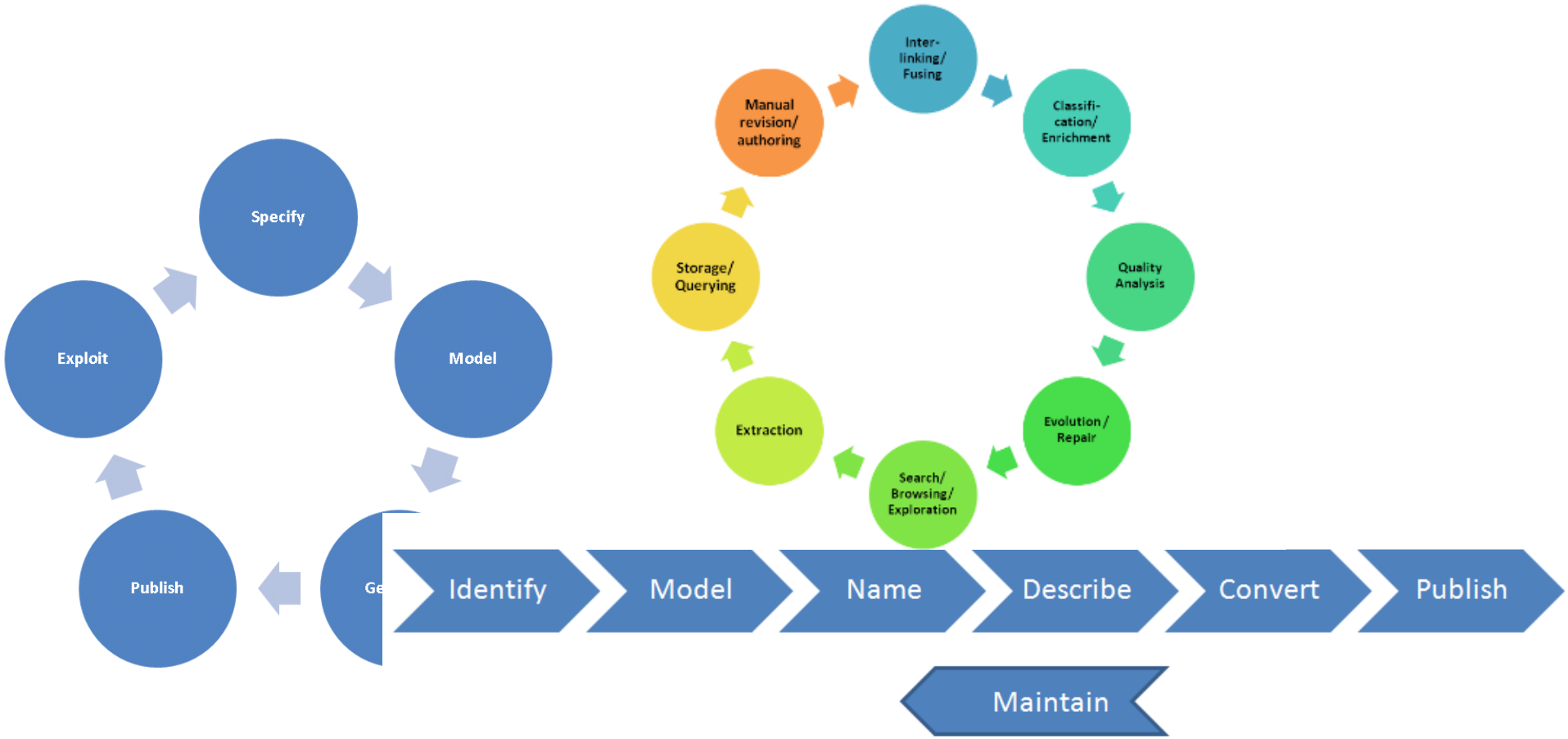
Predictive Models

http://www.flickr.com/photos/mbiskoping/6075387388

Raw Data Collection and Filtering	✓ Time window determination ✓ Location identification ✓ User profile's characteristics identification ✓ Search term selection
Predictor Variables Computation	✓ Predictor variables selection ✓ Predictor variables measurement ✓ Predictor variables computation
Predictive Model Creation	✓ Evaluation data identification ✓ Method selection ✓ Non-SM predictor variables selection and usage
Predictive Performance Evaluation	✓ Prediction baseline specification ✓ Evaluation method selection

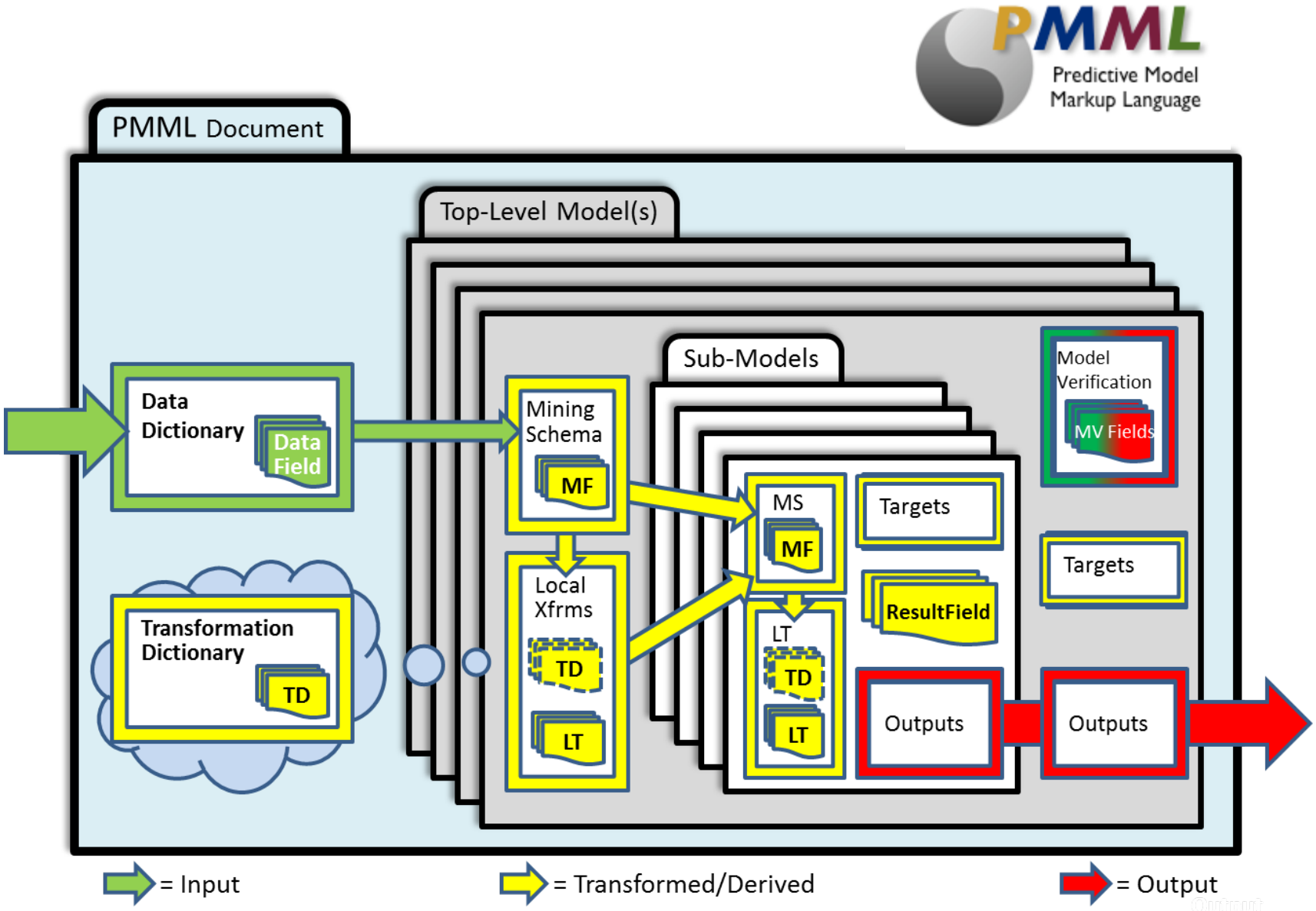
A vocabulary for describing predictive models as Linked Data

- A simple vocabulary that enables the creation of description of predictive models based on linked data principles

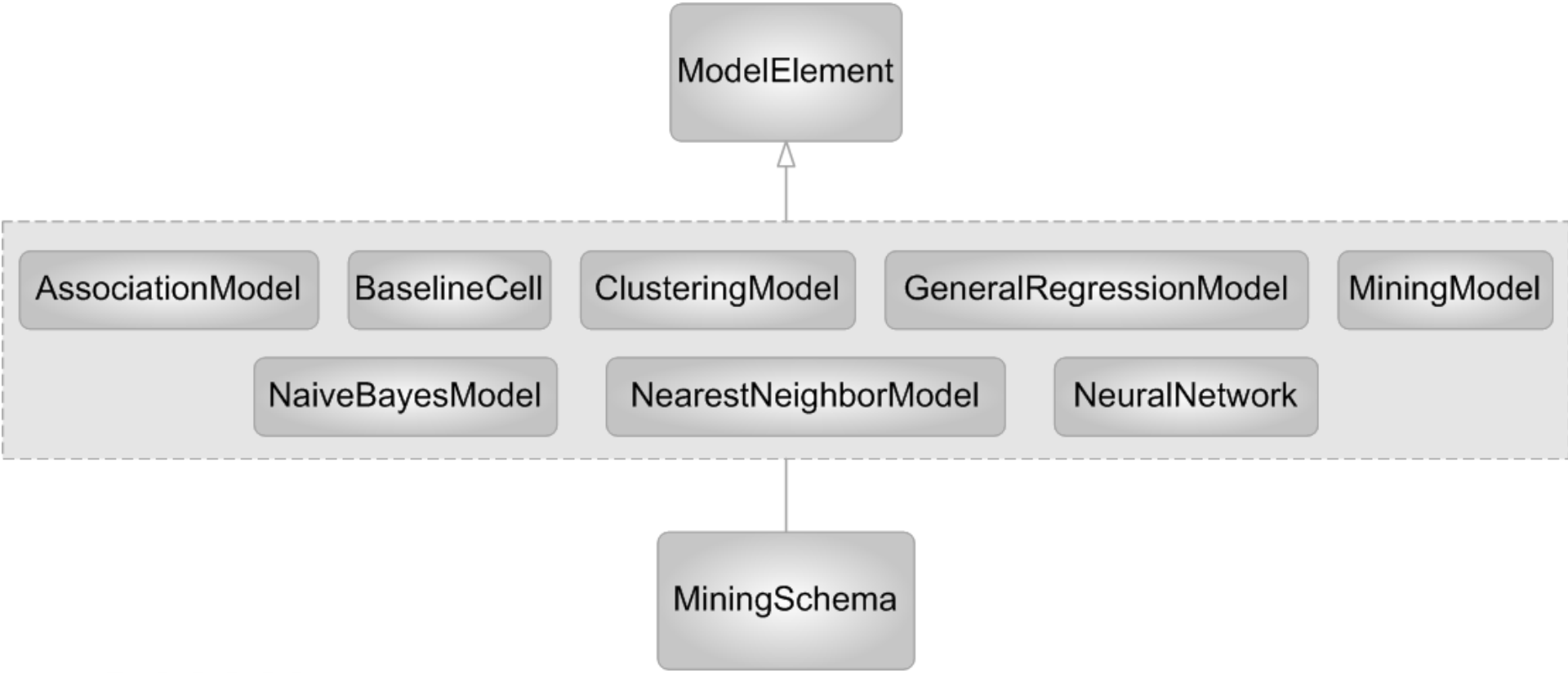


Relevant endeavors - PMML

- The **Predictive Model Markup Language (PMML)** is a standard for XML documents which express trained instances of analytic models
- **Main goal**: cross-platform interoperability
- PMML contains **over 700 elements**



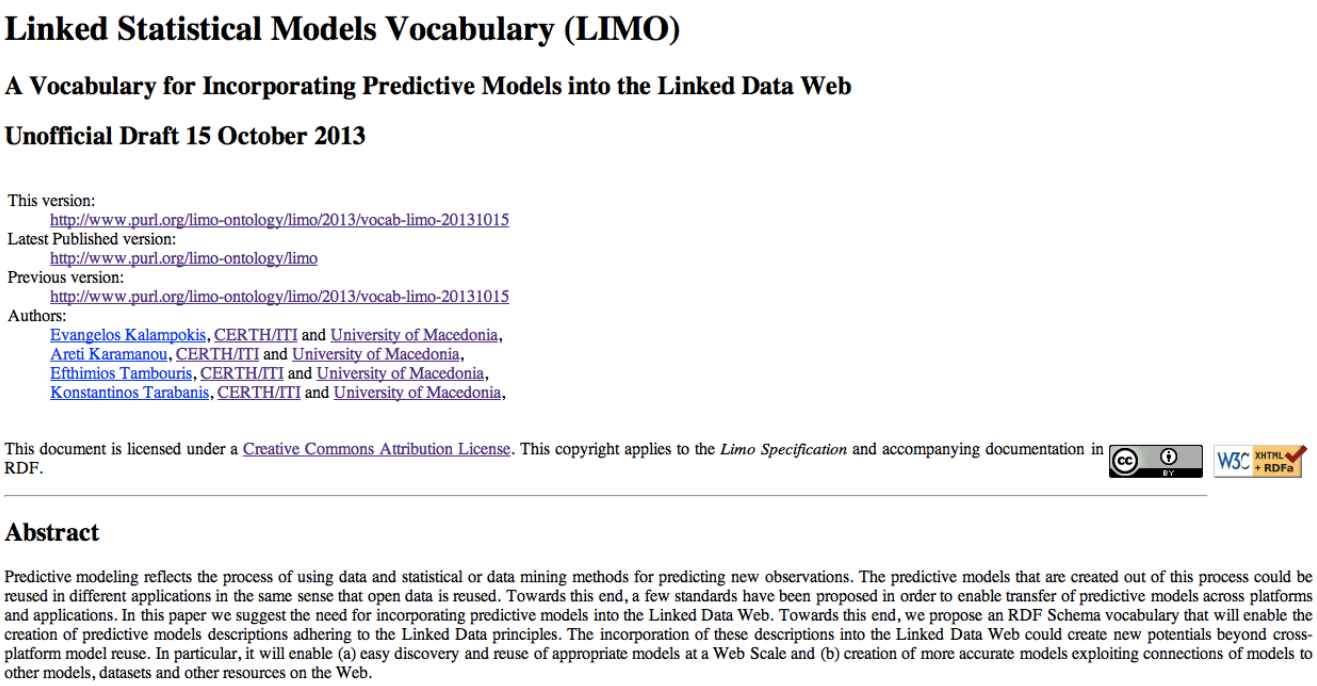
PMML's Model Element



PMML
Predictive Model
Markup Language

Linked Statistical Models Vocabulary (LIMO)

- LIMO will enable the creation of predictive models descriptions adhering to the Linked Data principles
- First unofficial draft in:
  - http://www.purl.org/limo-ontology/limo



Linked Statistical Models Vocabulary (LIMO)

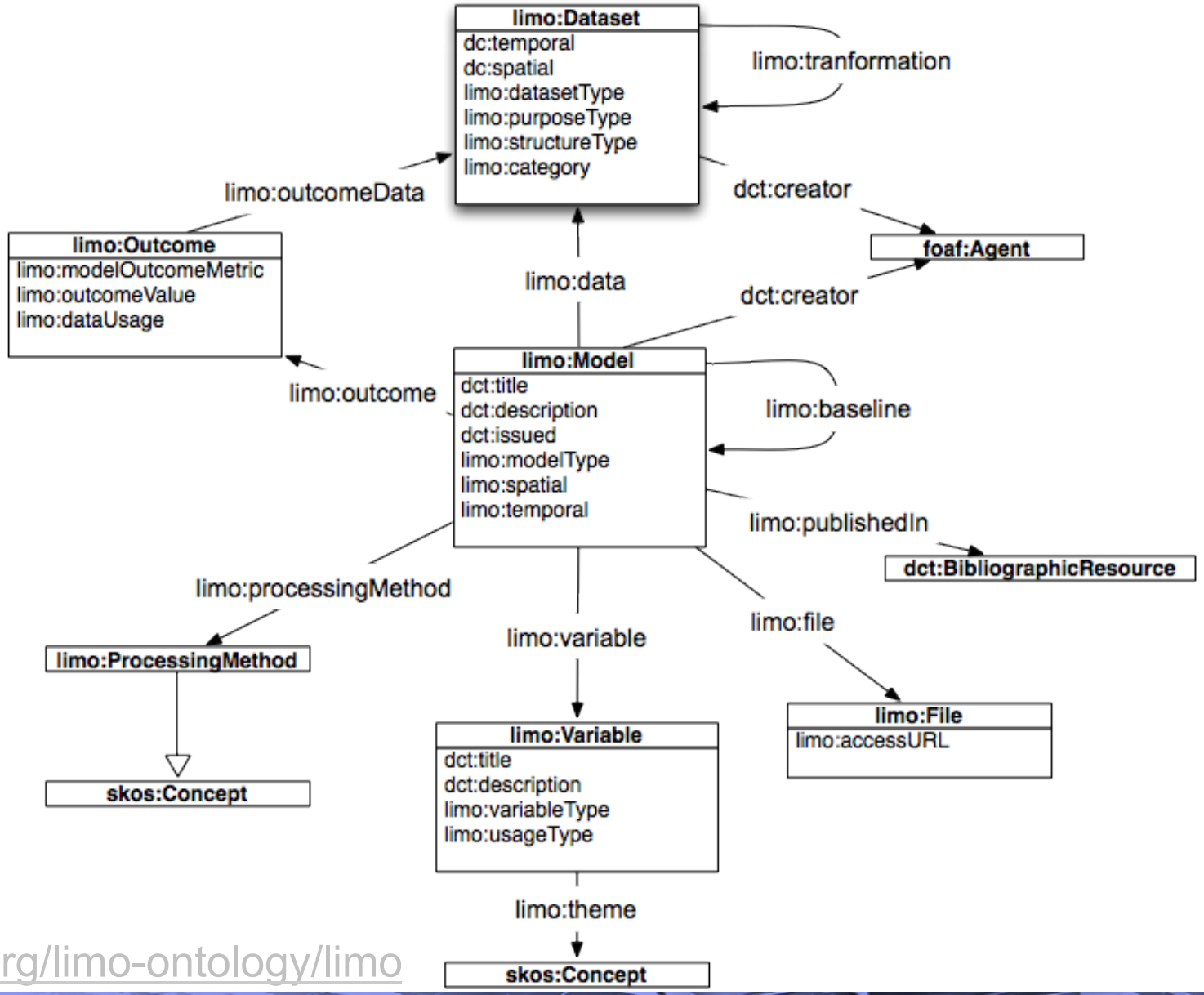
A Vocabulary for Incorporating Predictive Models into the Linked Data Web

Unofficial Draft 15 October 2013

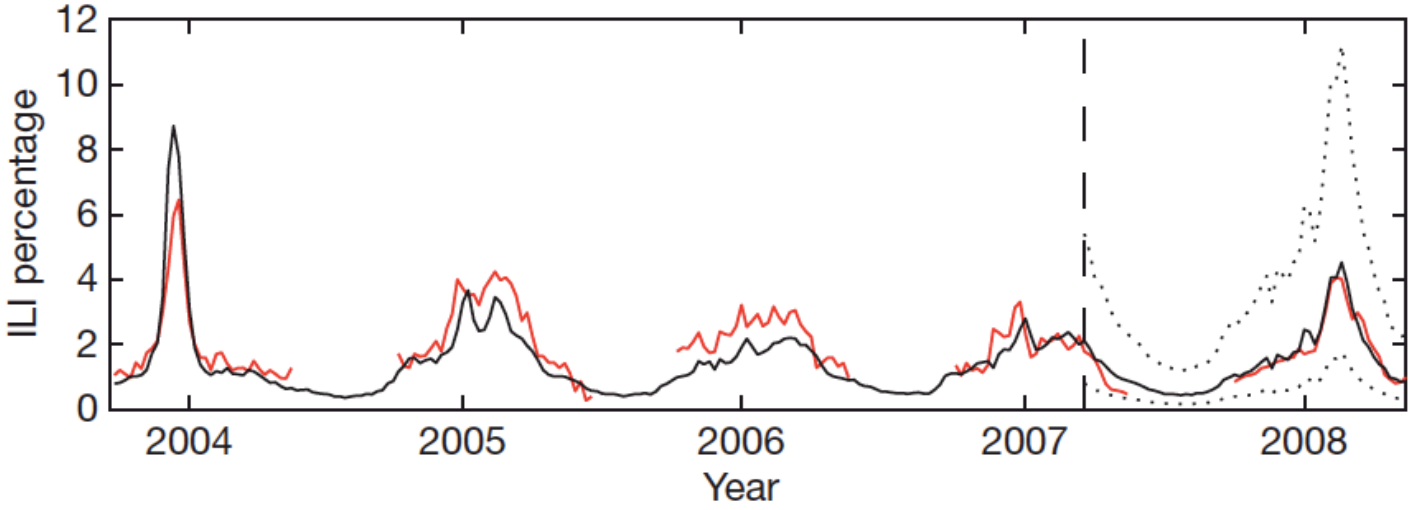
This version:
http://www.purl.org/limo-ontology/limo/2013/vocab-limo-20131015
Latest Published version:
http://www.purl.org/limo-ontology/limo
Previous version:
http://www.purl.org/limo-ontology/limo/2013/vocab-limo-20131015
Authors:
Evangelos Kalampokis, CERTH/ITI and University of Macedonia,
Areti Karamanou, CERTH/ITI and University of Macedonia,
Efthimios Tambouris, CERTH/ITI and University of Macedonia,
Konstantinos Tarabanis, CERTH/ITI and University of Macedonia,

Abstract

Predictive modeling reflects the process of using data and statistical or data mining methods for predicting new observations. The predictive models that are created out of this process could be reused in different applications in the same sense that open data is reused. Towards this end, a few standards have been proposed in order to enable transfer of predictive models across platforms and applications. In this paper we suggest the need for incorporating predictive models into the Linked Data Web. Towards this end, we propose an RDF Schema vocabulary that will enable the creation of predictive models descriptions adhering to the Linked Data principles. The incorporation of these descriptions into the Linked Data Web could create new potentials beyond cross-platform model reuse. In particular, it will enable (a) easy discovery and reuse of appropriate models at a Web Scale and (b) creation of more accurate models exploiting connections of models to other models, datasets and other resources on the Web.



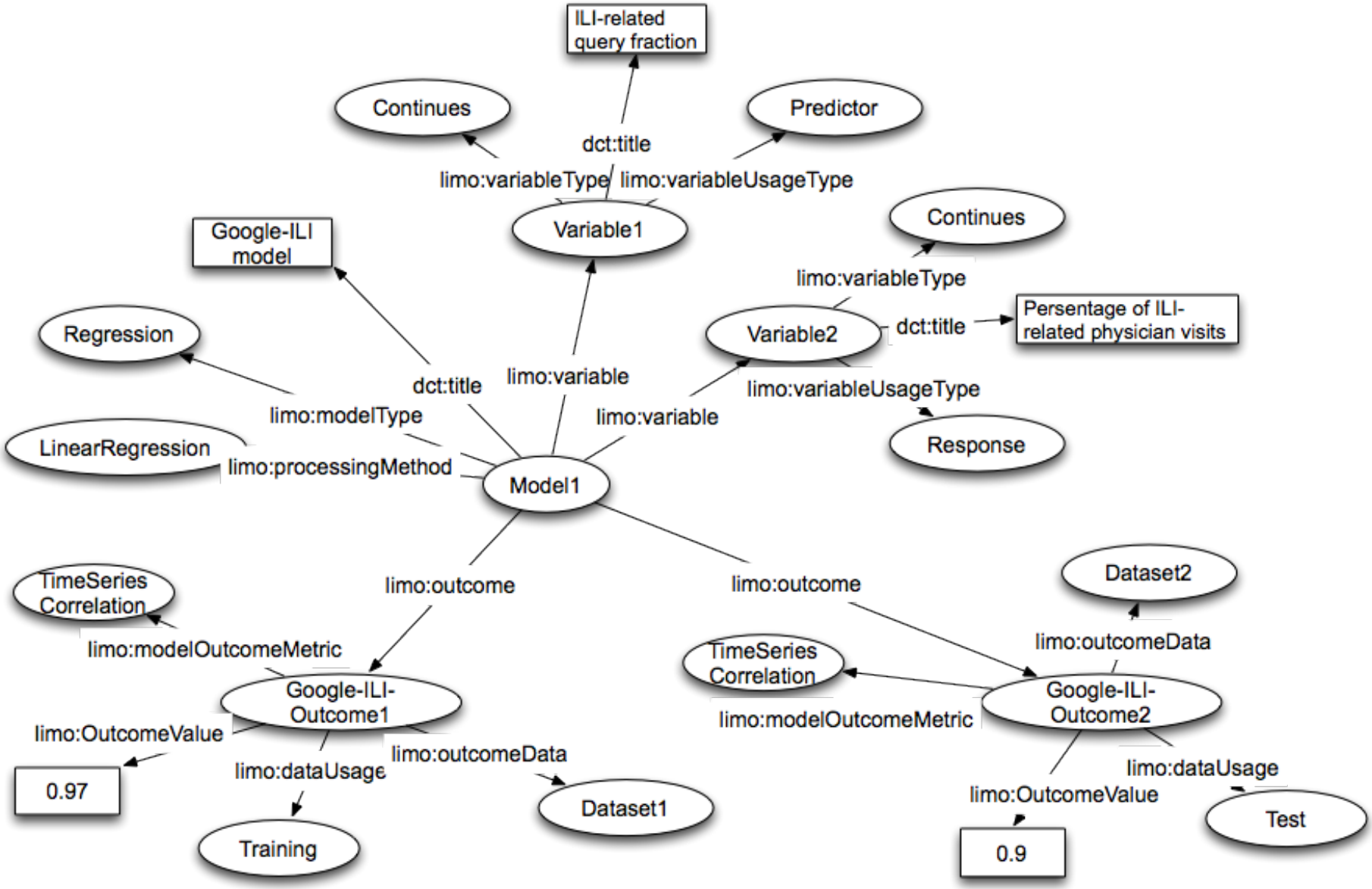
http://purl.org/limo-ontology/limo



Google

nature
International weekly journal of science

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L.: Detecting inuenza epidemics using search engine query data. Nature, 457(7232), 1012{4 (2009)



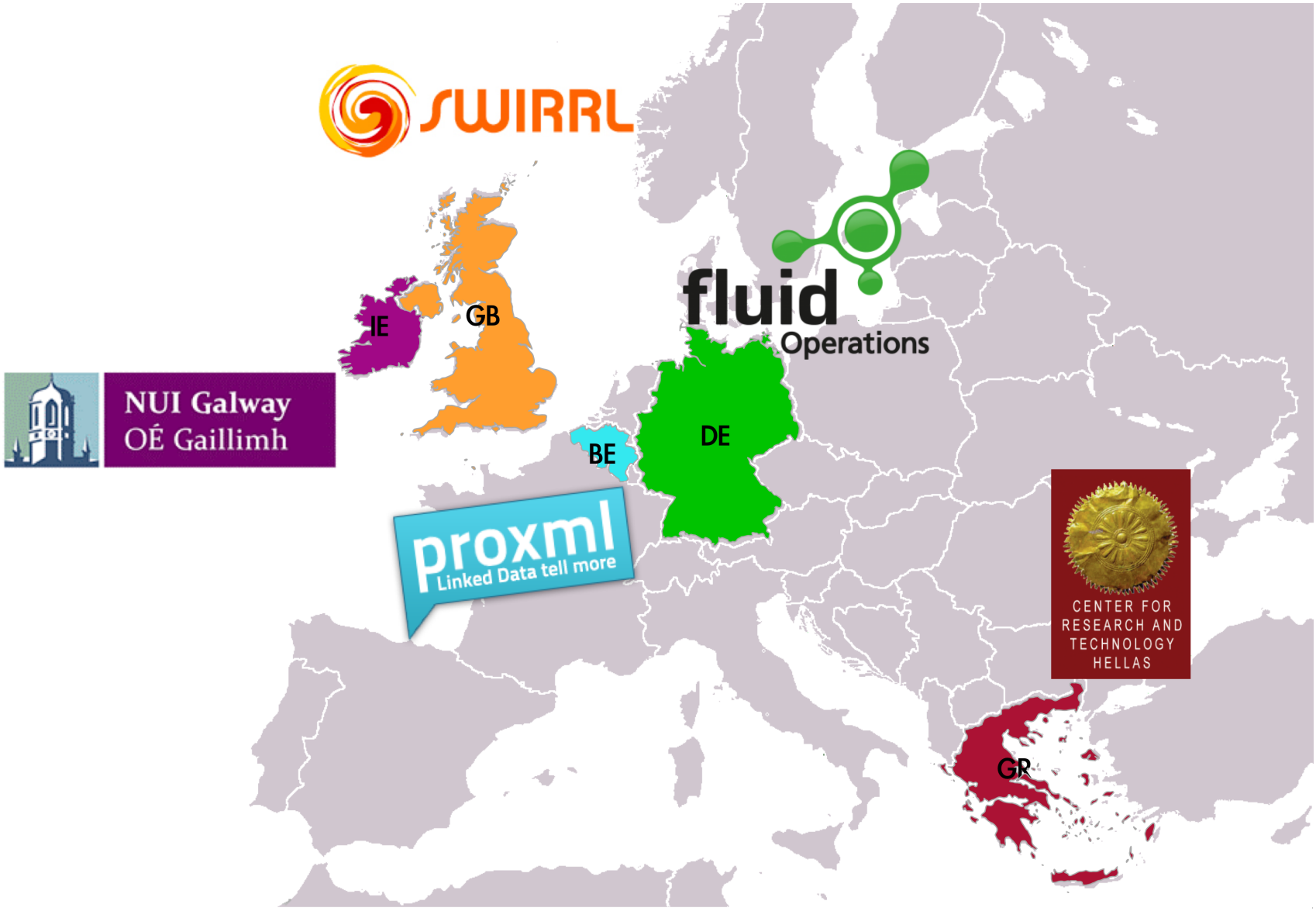
Future Work

- Finalize the model
- Create a dataset with predictive models described using LIMO
- Develop LIMO descriptions exporter

Future Work

- **OpenCube:** Publishing and Enriching Linked Open Statistical Data for the Development of Data Analytics and Enhanced Visualization Services
- *FP7-ICT-2013-SME-DCA No 611667*
- Start date: **1 November 2013**
- Duration: **24 months**

http://www.flickr.com/photos/stuckincustoms/524185543



Thank you for your attention!!

http://kalampok.is