Koen Smets

Personal Details

day of birth	7th January, 1984
place of birth	Wilrijk, Antwerpen (Belgium)
citizenship	Belgian
sex	Male
marital status	Cohabiting
drivers license	Category B

Education

2006–2012 **Ph.D. in Computer Science**, *Universiteit Antwerpen (UA)*.

dissertation Identifying and characterising anomalies in data

supervisor prof. dr. Bart Goethals (ADReM)

2004–2006 **M.Sc. in Mathematics (Computer Science)**, *Universiteit Antwerpen*, magna cum laude.

dissertation Study of kernel-based techniques for single-class classification and feature selection based on the optimisation of the kernel parameters

supervisor prof. dr. Brigitte Verdonk (ECT) and dr. Piet van Remortel (ISLab)

grade summa cum laude

2002–2004 **B.Sc. in Mathematics (Computer Science)**, *Universiteit Antwerpen*, magna cum laude.

1996–2002 **Sciences-Mathematics**, *Sint-Lievenscollege*, Antwerpen.

Research Grants

2013–2014 **RIOFI grant for Proof-of-Concept project**, *Universiteit Antwerpen*.

- Co-funding provided by two industrial partners: Forcea and UZA

2007–2012 **Ph.D. fellowship**, *Research Foundation - Flanders (FWO)*.

2006–2007 **Umbrella grant for FWO candidates**, *Universiteit Antwerpen*.

Professional Experience

Research & Development

current **Freelance Data Engineer/Scientist/. . . .**

- Architected platform to monitor energy consumption using TICK-stack for Bloomingfeld
- Implemented NLP-pipeline for IT Pieces by Flora Miranda
- Scraped Tax-on-Web for Moore Stephens and Ampe Accountancy

Senior Data Engineer, *Sentiance n.v./s.a.*, Antwerpen.

- Optimized and extended both batch layer (in Apache PySpark) and speed layer (in Apache Storm) of a λ-architecture processing data (in Thrift/Parquet format) from mobile devices and smarthome sensors to generate contextual user profiles (stored in PostgreSQL views)
- Implemented Kafka(-stream) based microservices, deployed as Docker containers
- All software and algorithms are developed in Java or Python, deployed using Ansible and monitored in an Elasticsearch/Logstash/Kibana stack.

2014-2015 **Software Developer**, *MDCPartners c.v.b.a*, Antwerpen.

- Setup Apache Nutch to crawl the internet for downloading webpages and documents to discover information about persons and organisations related to drugs and clinical trials
- Implemented a streaming document analysis pipeline on top of Twitter Storm to extract textual information integrating the GATE text-analysis platform and in-house developed algorithms
- Documents and discovered facts are interlinked and stored in a Titan graph database, searchable in Elasticsearch and demonstrated to end-users in a web-app displaying the network using sigma.js
- Performed ad hoc data explorations and statistical analyses using Kibana and IPython Notebook
- All software and algorithms are developed in Java, deployed using Ansible and monitored in an Elasticsearch/Logstash/Kibana stack.

jan-dec 2013 **Postdoctoral researcher**, *Advanced Database Research and Modelling (ADReM)*, UA.

- Developed an automated software platform to support and improve the quality of clinical coding, evaluated positively by pilot-users from 7 Flemish hospitals, and designed a valorisation plan
- Models in the back-end to identify anomalies in clinical coding and to automatically suggest codes are built using data mining and machine learning techniques, and are implemented in Python by extending the scikit-learn package
- The web-application for end-users is implemented using Python (Flask), HTML5 and Javascript (backbone.js and jQuery) and served through nginx and uWSGI
- All data resides in MongoDB, while back- and front-end are loosely coupled using a Redis key-value store and a RabbitMQ queuing system to provide horizontal and vertical scaling

oct-dec 2012 **Project leader data mining**, *biomedical informatics research group (biomina, i-ICT)*, UZA.

may-oct 2012 **Postdoctoral researcher**, *Advanced Database Research and Modelling (ADReM)*, UA.

- Involved in writing a research proposal and connecting research in ADReM with industry

2008–2012 **Ph.D. researcher**, *Advanced Database Research and Modelling (ADReM)*, UA.

- Developed Slim, an algorithm to mine high-quality patterns directly from data, written in C++ and optimised further using the open-source tools gprof and valgrind
- Implemented pattern-based algorithms in C++ not only to detect anomalies, but also to provide an explanation why an observation is regarded as unexpected
- Ported Krimp, an algorithm for mining compressing patterns in two phases, from Windows to GNU/Linux using CMake, g++ and gdb
- All experiments scripted in Python and their results visualised using Matplotlib

2006–2008 **Ph.D. researcher**, *Emerging Computational Techniques (ECT)*, UA.

- Initiated the data-driven identification of vandalism in Wikipedia by comparing the results of a Naive Bayes and compression-based classifier
- Reimplemented on top of Lucene the Explicit Semantic Analysis (ESA) method (Gabrilovich and Markovitch, 2007) to compute the semantic relatedness between two arbitrary texts
- Experimentally evaluated, in Matlab, the usefulness of several risk functionals for the selection and optimisation of hyperparameters for support vector regression

Internships

feb-apr 2012 **biomedical informatics research group (biomina, i-ICT)**, *UZA*, Antwerpen, Belgium.
Identified erroneous co-morbidities in the ICU database using Slim and participated in the PhysioNet challenge with a random forest classifier, in R, to predict the mortality risk of ICU patients

feb-aug 2007 **R&D**, *Dow Benelux*, Terneuzen, The Netherlands.
Implemented an algorithm, in Matlab, based on one-class support vector machines, to detect anomalies when monitoring several sensors during a production process in a chemical plant

Consulting

nov-dec 2010 **Data mining assistance**, *Noor Remmerie (CeProMa)*.
Applied tiling, or biclustering, to optimally group proteins by their migration profile and to correlate them to a specific complex

may 2010 **Statistical assistance**, *Marleen Eyckmans (ecotox, EB&T)*.
Explained ANOVA tests with Bonferonni correction to statistically analyse (using R) the differences between three groups of fish species

jan-jun 2008 **Machine learning assistance**, *Thanh Hai Dang (ISLab)*.
Provided insight in additional single-class techniques for predicting phosphorylation sites

Teaching

2008–2012 **Assistant**, *Artificial Intelligence (B.Sc.) and Database Security (M.Sc.)*.

2007–2013 **Co-supervisor**, *B.Sc. and M.Sc. dissertations*.

artificial intelligence	play Pac-Man with Lego Mindstorms robots
data mining	mining software repositories, interactive and visual pattern mining
natural language processing	extracting semantic knowledge form Wikipedia
operations research	transport optimisation, vehicle routing problems
security	intrusion detection in databases
software engineering	implementing web-application to register clinical codes

2006–2008 **Assistant**, *Artificial Intelligence (B.Sc.) and Capita Selecta Artificial Intelligence (M.Sc.)*.

Other

2009–2012 **Volunteer**, *Auxilia*, Antwerpen, teaching mathematics to disadvantaged across Antwerp.

aug 2004–2006 **Student worker**, *Digipolis*, Antwerpen.
Developed a project planner in Visual Basic, cleaned up digital archive using Perl scripts, upgraded network/telephone infrastructure at several municipal locations, actualised course material for introductory workshops about internet, computer and GSM for antwerpen.be-centrum (ABC), . . .

jul 2001–2003 **Student worker**, *Colruyt*, Antwerpen.

2000–2002 **Basketball youth coach**, *Olicsa*, Antwerpen.

Skills

computer **GNU/Linux sysadmin and free/open-source software user**, *18 years experience.*

- Using *Ansible* to deploy the clinical coding web-app and to configure ADReM data and crunch servers, and to orchestrate a handful of virtual machines in combination with *Vagrant*
- Monitoring all machines and services using *Icigna* and *Ganglia*

language **C++, Python and Java**, *16 years experience.*

Dutch, English and French, *resp. at 'mother tongue', 'fluent' and 'unrefreshed' level.*

communication **Technical writing and presenting**, *at several international conferences and workshops.*

personal **Studious, structured, versatile, independent**, *but aware that more is achievable in group*, hence, **team player**.

Leisure

2011– ... **Swimming**, *Nautica*, Antwerpen.

2010– ... **Developing**, virtual coach analysing my workout/health data and planning future workouts.

2008– ... **Running**, long distance.

1993–2010 **Basketball**, *Olicsa*, Antwerpen.

Primary Publications

- **K. Smets** and J. Vreeken. Slim: Directly mining descriptive patterns. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA*, pages 236–247, 2012.
- **K. Smets** and J. Vreeken. The odd one out: Identifying and characterising anomalies. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*, pages 804–815, 2011.
- **K. Smets**, B. Verdonk, and E. M. Jordaan. Discovering novelty in spatio/tem-poral data using one-class support vector machines. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks (IJCNN), Atlanta, GA*, pages 2956–2963, 2009.
- **K. Smets**, B. Goethals, and B. Verdonk. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI), Chicago, IL*, pages 43–48, 2008.
- **K. Smets**, B. Verdonk, and E. M. Jordaan. Evaluation of performance measures for SVR hyperparameter selection. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks (IJCNN), Orlando, FL*, pages 637–642, 2007.

Secundary & Tertiary Publications

- T. Vu, D. Valkenborg, **K. Smets**, K. Verwaest, R. Dommisse, F. Lemiere, A. Verschoren, B. Goethals, and K. Laukens. An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics*, 12(1):405, 2011.
- N. Remmerie, T. D. Vijlder, D. Valkenborg, K. Laukens, **K. Smets**, J. Vreeken, I. Mertens, S. C. Carpentier, B. Panis, G. D. Jaeger, R. Blust, E. Prinsen, and E. Witters. Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods. *Journal of Proteomics*, 74(8):1201 – 1217, 2011.
- M. Eyckmans, C. Tudorache, V.M. Darras, R. Blust and G. De Boeck. Hormonal and ion regulatory response in three freshwater fish species following waterborne copper exposure. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 152(3):270–278, 2010.
- T.H. Dang, K. Van Leemput, A. Verschoren and K. Laukens. Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics*, 24(24):2857–2864, 2008.