Insurance analytics

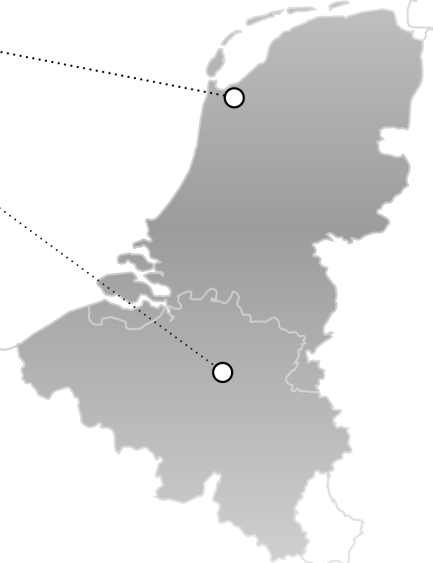
Knowing me, knowing you: actuarial science meets data science

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

May 26, 2019

UNIVERSITEIT VAN AMSTERDAM

AMSTERDAM SCHOOL OF ECONOMICS Economics



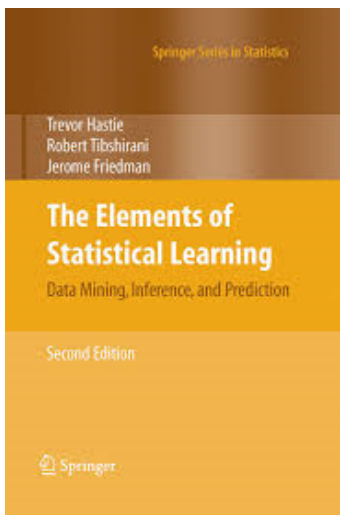
KU LEUVEN

LRISK

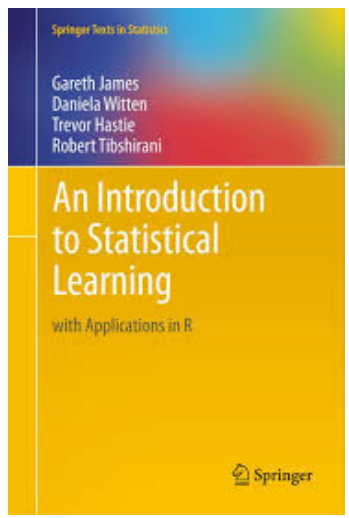
Analytics: what's in a name?

(Data) analytics or data science or data mining or predictive modeling or ...

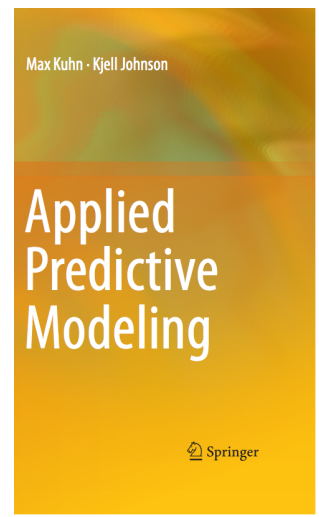
... refers to a vast set of tools for understanding data.



Hastie, Tibshirani & Friedman



James et al.

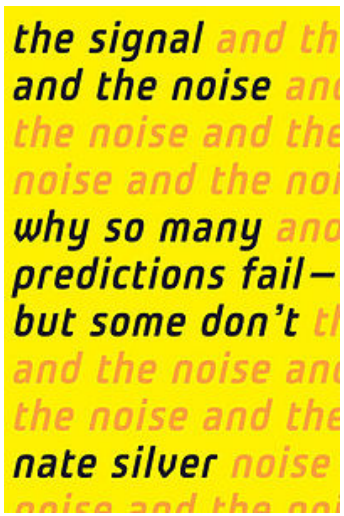


Kuhn & Johnson

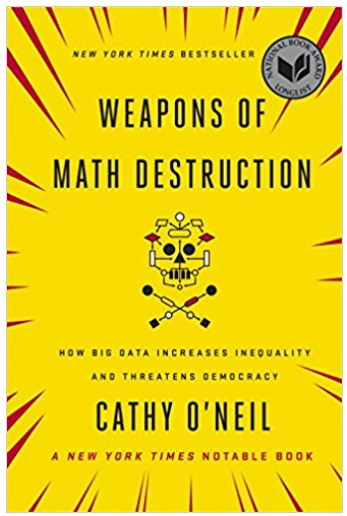
Analytics: what's in a name?

(Data) analytics or data science or data mining or predictive modeling or ...

... refers to a vast set of tools for understanding data.



Nate Silver



Cathy O'Neil

Hannah Fry

(Provocative) Statement Nr. 1

Doing data science - Straight talk from the frontline

What is the eyebrow-raising about big data and data science?

The hype is crazy.

Getting past the hype?

There might be some meat in the data science sandwich. Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.

Quote from *Doing data science - Straight talk from the frontline*, by Rachel Schutt and Cathy O'Neil, 2013.

What data scientists really do

- Technical tasks of a data scientist:
  - identifying models, including selecting/building appropriate features
  - training the models and testing their performance
  - interpreting the results and re-evaluating model selection
  - visualization of data and findings.
- Technical skills of a data scientist:
  - programming (e.g. R or Python), including standard packages for machine learning and visualization.
  - proficient knowledge of machine learning techniques and how they differ from each other.

Insurance analytics

- The actuary plays a central role in data analysis and predictive modeling:
  - insurance pricing and product development
  - reserving and accounting
  - risk management and Nat Cat modeling
  - marketing
  - claims handling.
- "All these actuarial fields go through massive, data driven changes."

  (quoting prof. Mario Wüthrich, ETH Zurich)

(Provocative) Statement Nr. 2

Aviva's CEO

> From a skills perspective, Wilson is aware of the need to reskill employees to navigate this digital era; for example, retraining actuaries to become data scientists. 'I'm desperate for that skill set but universities don't train people in it. I'm willing to pay more for a data scientist than an actuary,' he reveals.

Quote from *Reviving Aviva: Exclusive interview with Mark Wilson*, published on May 29, 2018.

☺ (New!) UvA course on Insurance analytics - een actuariële kijk op data science! ☺

Actuaries of the 5th kind

(Picture taken from Data Science Strategy, Working party of the Swiss Association of Actuaries, August 2018.)

Insurance analytics: 'the two cultures'

	Statistical Learning	Machine Learning
origin	statistics	computer science
$f(X)$	model	algorithm
emphasis	interpretability, precision and uncertainty	large scale applicability, prediction accuracy
jargon	parameters, estimation	weights, learning
CI	uncertainty of parameters	no notion of uncertainty
assumptions	explicit a priori assumption	no prior assumption, learn from the data

(Taken from Why a mathematician, statistician and machine learner solve the same problem differently.)

Actuarial learning: (some) challenges

- Past/Present

  Risk classification in competitive markets using (standard) regression models (~ GLMs) for frequency and severity.

- Ongoing

  From statistical learning to machine learning with shallow but also deep learning techniques.

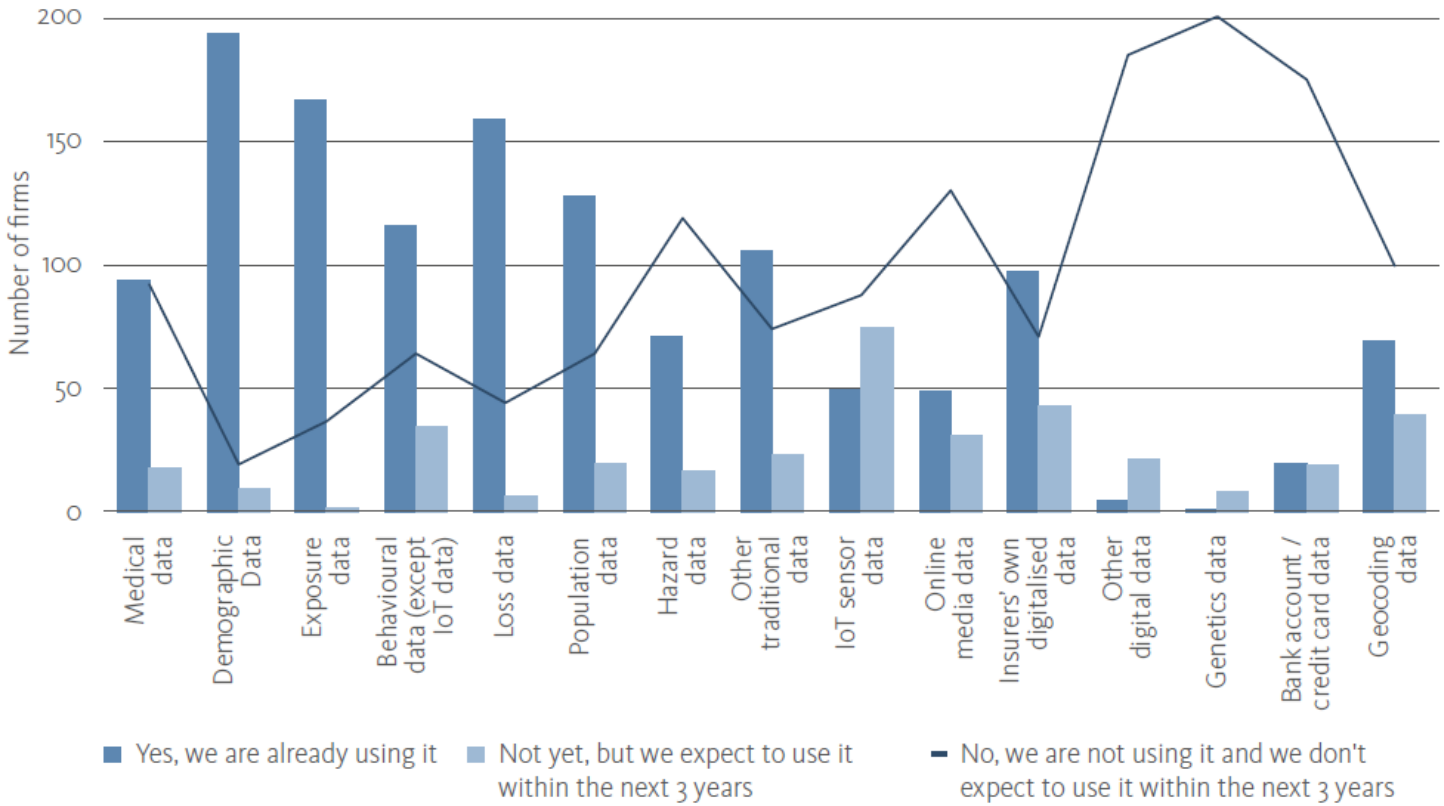
  New data sources (structured, but also unstructured).

- Challenges?

  - keep model explainable to clients, regulators, ICT
  - !!be aware of specific features of insurance data!!

Actuarial learning: (some) challenges

Figure 3 – Usage of different types of data



Source: EIOPA BDA thematic review

Actuarial learning: (some) challenges

Figure 7 – Usage of BDA tools such as AI and ML

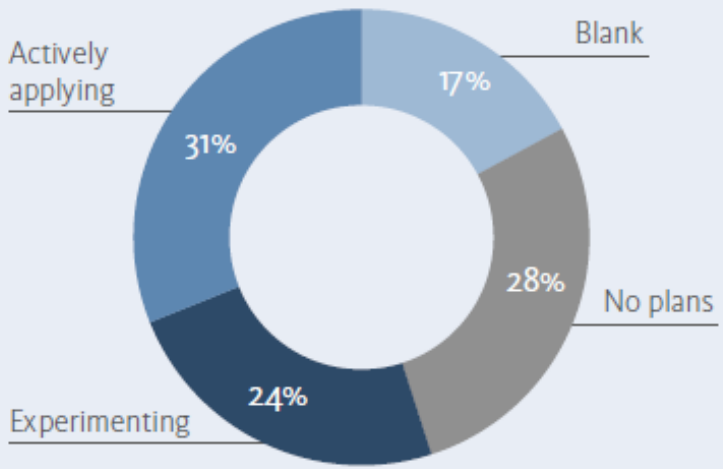
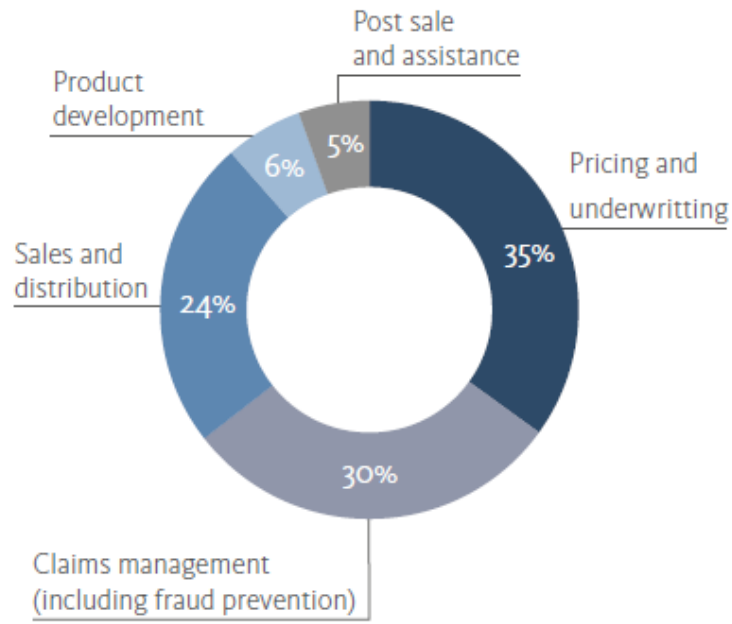
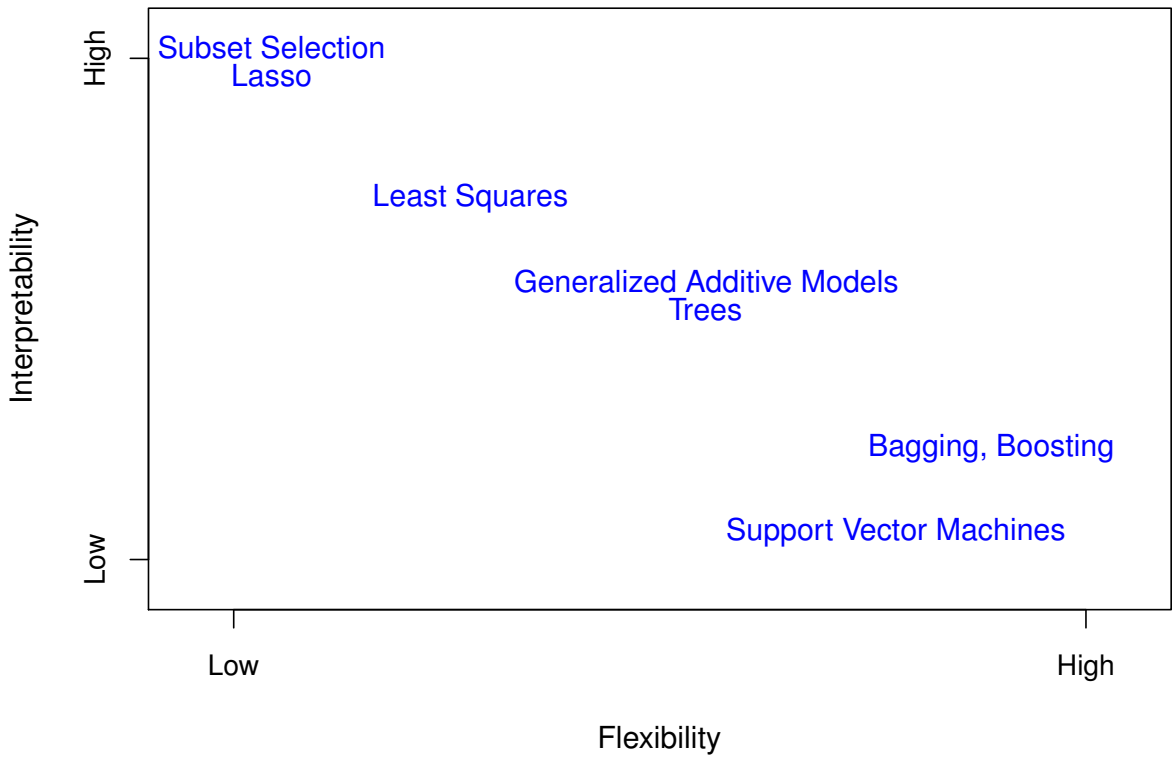


Figure 8 – Usage of BDA tools such as ML and AI across the value chain



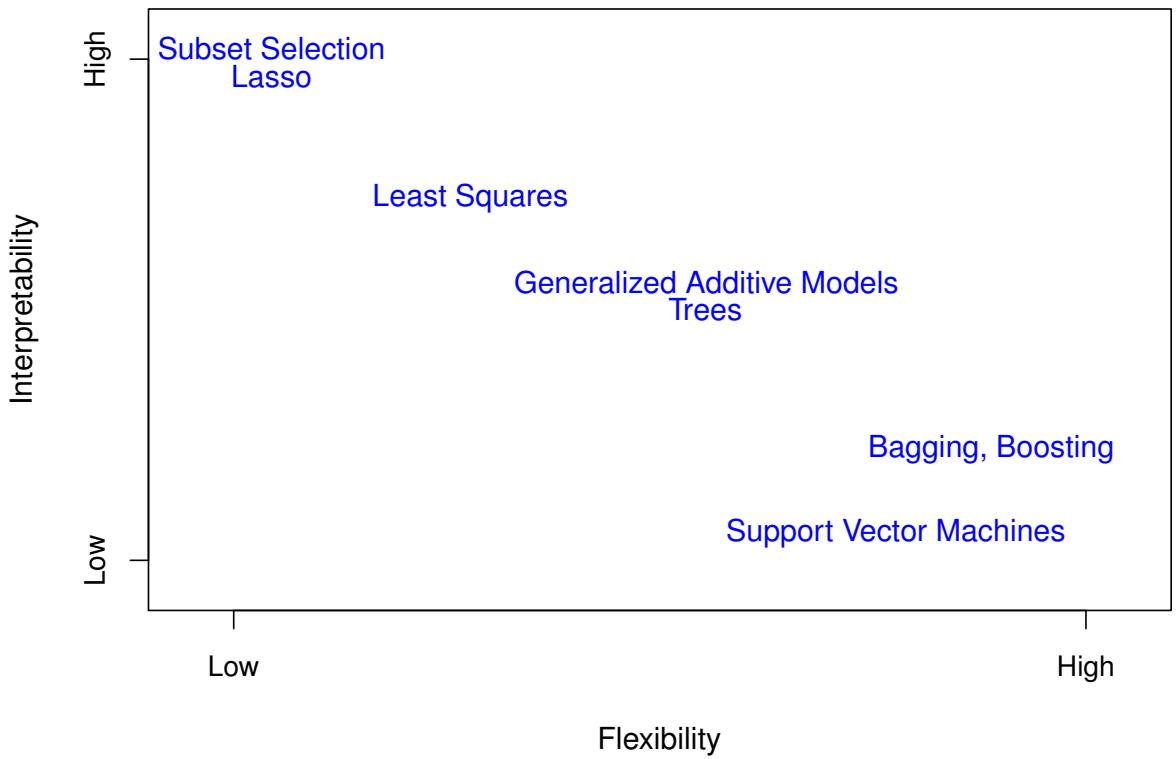
Source: EIOPA BDA thematic review

Actuarial learning: (some) challenges



Actuarial learning: (some) challenges

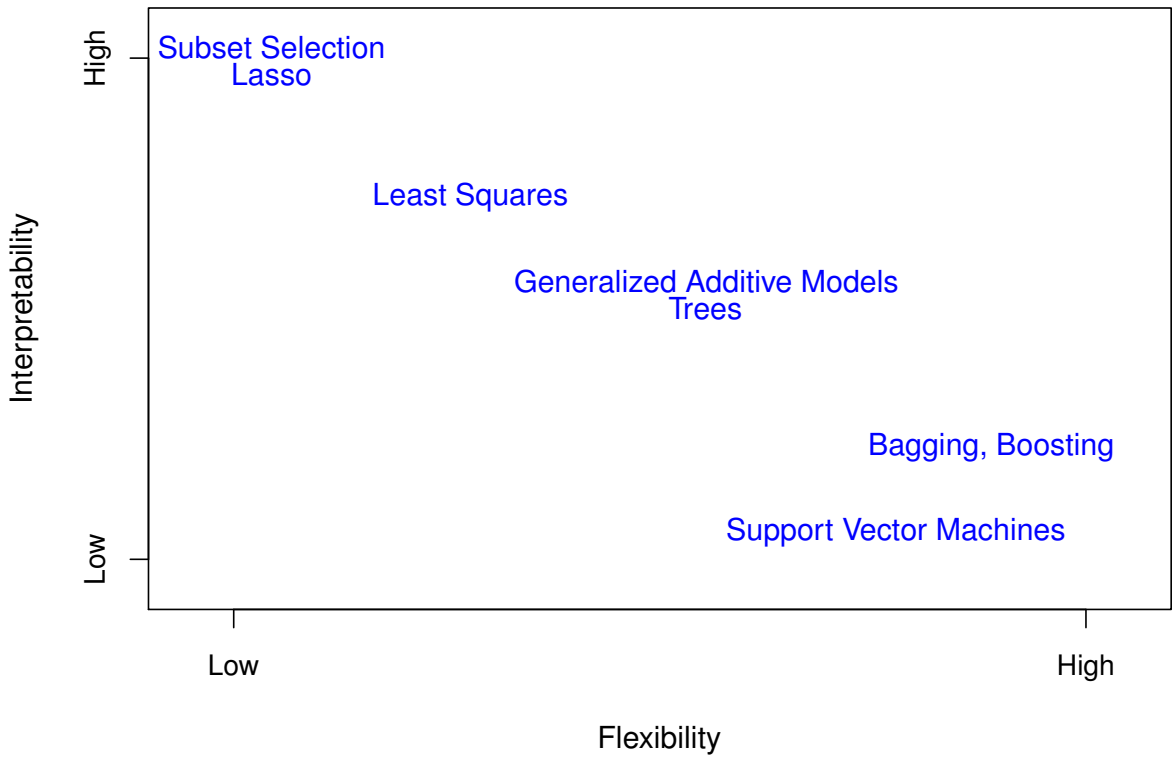
Trade off between flexibility and model interpretability!



(Picture taken from James et al., An introduction to statistical learning, 2017.)

Actuarial learning: (some) challenges

The 'black box' versus 'white box' discussion!



(Picture taken from James et al., An introduction to statistical learning, 2017.)

Use cases

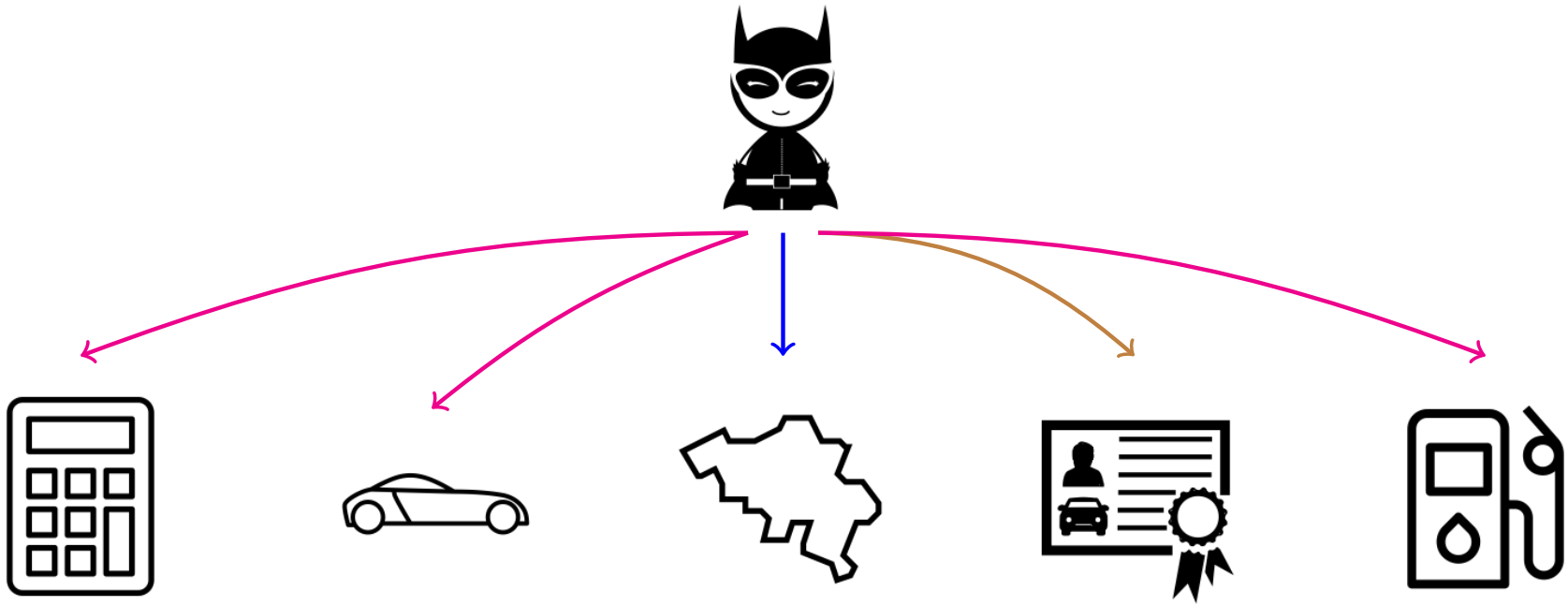
Figure 9 – BDA uses cases

Use Case	Output
Churn models	Use of ML churn models for the prediction of consumer's propensity to shop around at the renewal stage, which can be useful for pricing and underwriting (e.g. for price optimisation in combinaiton with a demand price-elasticity analysis) or for servicing the customer (e.g. "Next Best Action" approach)
Chatbot	Enable "human like" conversations with consumers by analysing customer unstructured data via text or voice with the use of natural language processing and other ML algorithms
Sentiment Analysis	Evaluate the sentiment in feedback provided by consumers to transform it into usable information to help improve customer satisfaction and engagement
Electronic document management	Robotic process automation (RPA) – Deep learning networks used for automatic classification of incoming documents of unstructured data (e.g. emails, claims statements), routing them to the correct department
Claims management	Optical character recognition (OCR) - Deep learning networks used to extract information from scanned documents such as images from damaged cars to estimate repair costs
Fraud prevention	Analysis of fraudulent claim patterns based on FNOL data provided by the consumer
Product development	Use of ML and graph database in predictive modeling for the identificaiton of disease development patterns
Pricing and underwriting	BDA tools used in motor and health insurance for processing large quantities of data from different sources, often on a real-time basis (e.g. quote manipulation), using a wide array of statistical techniques

Source: EIOPA BDA thematic review

Research on insurance pricing

Pricing through risk classification



Claim frequency and claim severity

as function of

nominal / numeric ~ ordinal / spatial

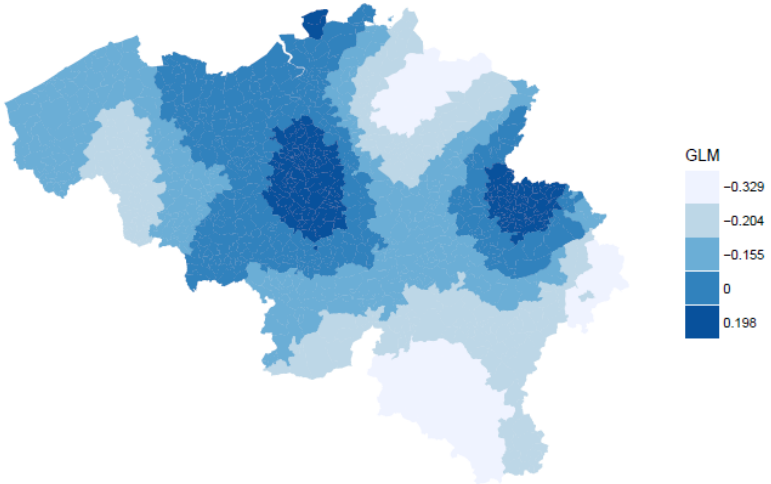
features

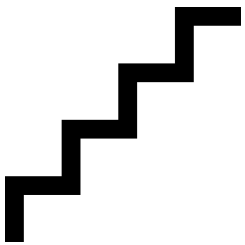
Pricing through risk classification

Research contributions

Pricing through risk classification

Research contributions





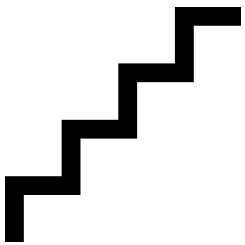
step-by-step

best subset selection

risk classes

Pricing through risk classification

Research contributions



step-by-step

best subset selection

risk classes

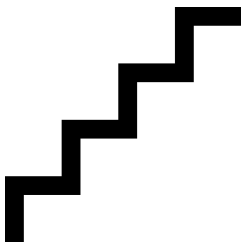
SMuRF

sparsity regularization

automatic feature selection

Pricing through risk classification

Research contributions



step-by-step

best subset selection

risk classes

SMuRF

sparsity regularization

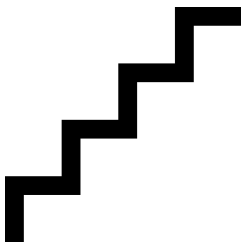
automatic feature selection

tree-based

CART, random forest gradient boosting

Pricing through risk classification

Research contributions



step-by-step

best subset selection

risk classes

SMuRF

sparsity regularization

automatic feature selection

Statistical Learning

tree-based

CART, random forest gradient boosting

Machine Learning

Pricing through risk classification

Research contributions

- Sparse regression with multi-type feature modelling by Devriendt, Antonio et al. (2018)
  - automatic feature selection and binning of risk factors
  - R package `smurf`
  - end product is a GLM!
- Boosting insights in insurance tariff plans with tree-based machine learning by Henckaerts, Côté, Antonio et al. (2019)
  - GLMs, GAMs, decision trees, random forests and gradient boosting machines
  - R packages extended to Poisson and gamma deviance
  - tuning strategy, interpretability tools and managerial insights.

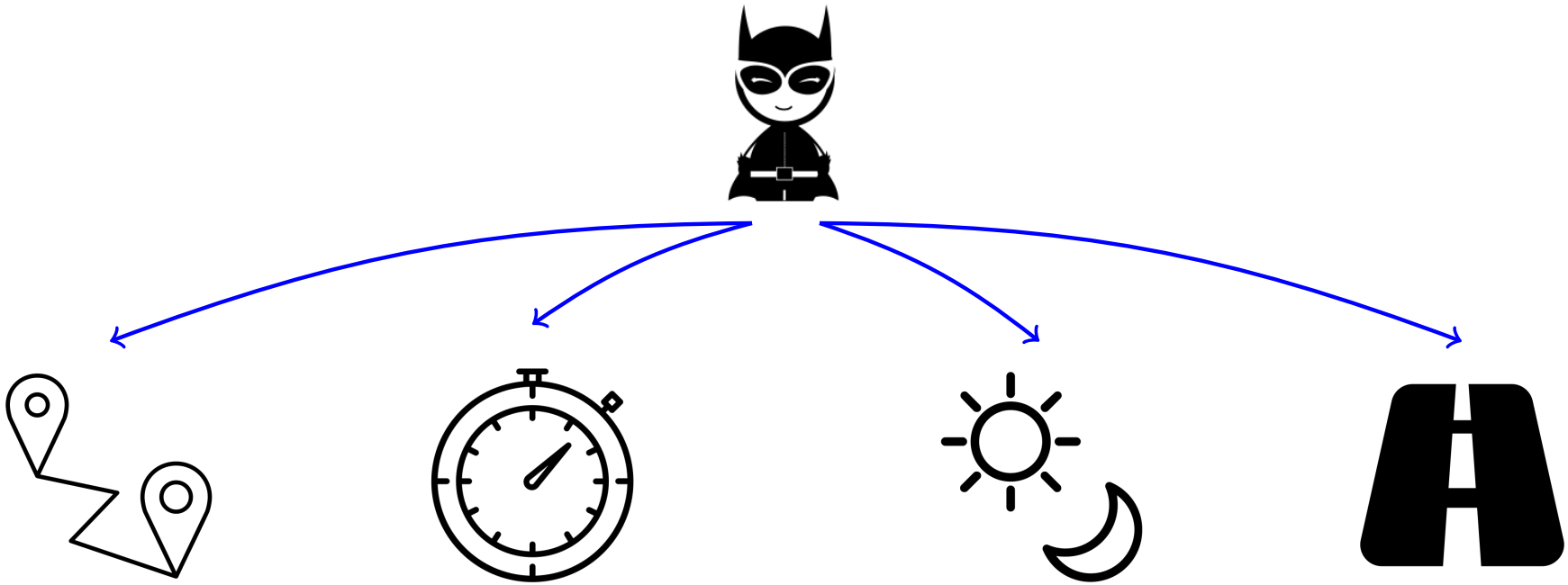
Telematics insurance

Verbelen, Antonio & Claeskens,

JRSS C, 2018

Telematics insurance

New features?



Claim frequency and claim severity

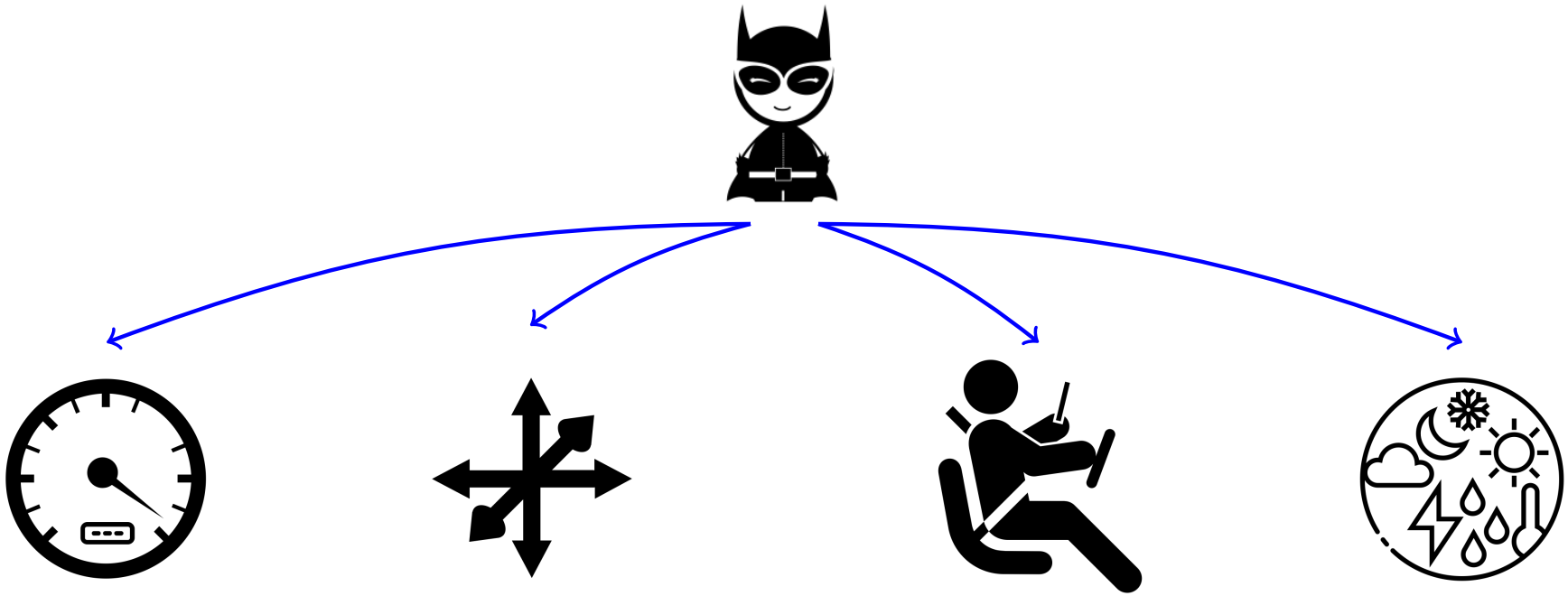
as function of

driving habits

e.g. mileage, travel time, time slot, road type

Telematics insurance

New features?



Claim frequency and claim severity

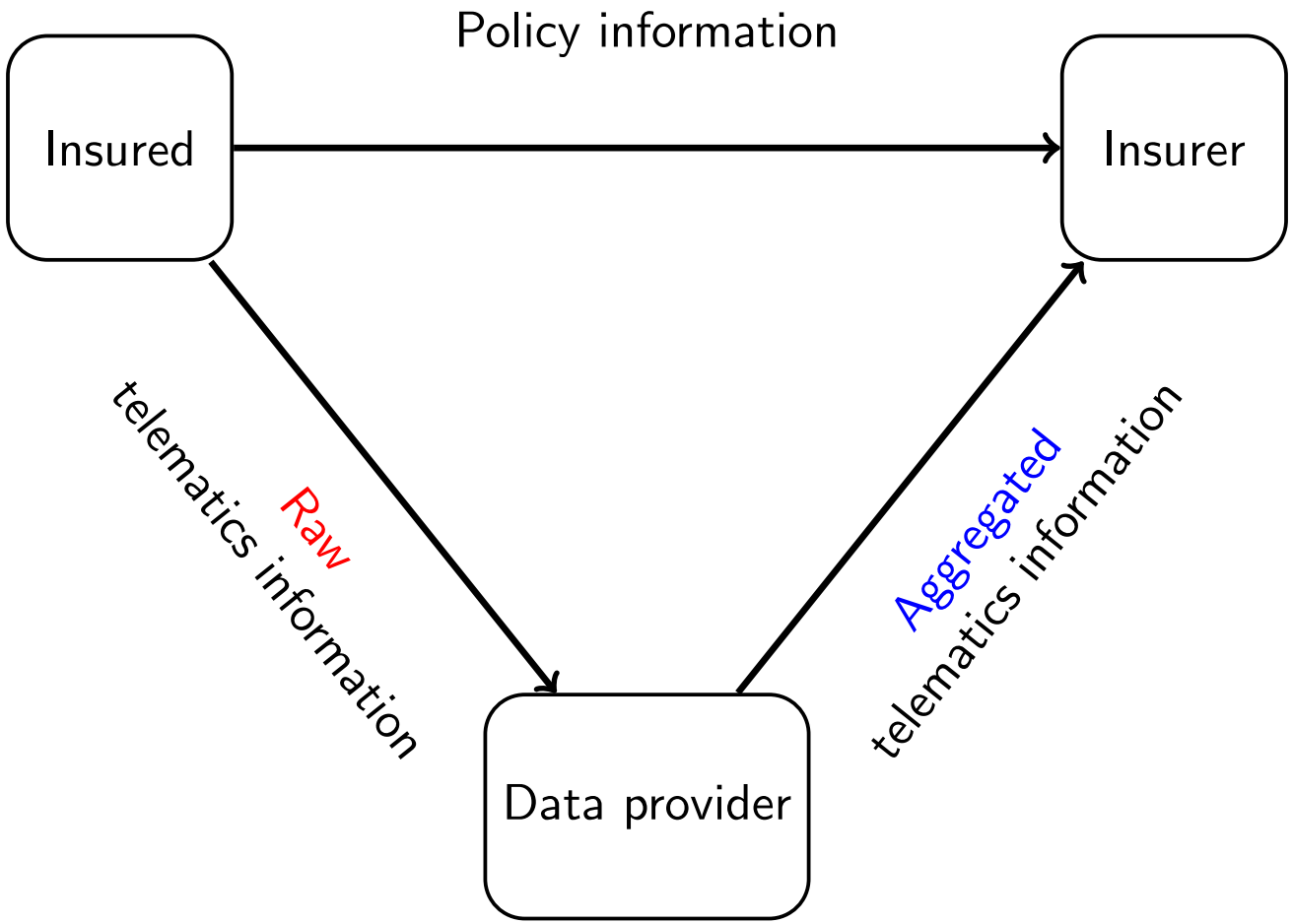
as function of

driving styles

e.g. speed, acceleration, attention, weather

Telematics insurance

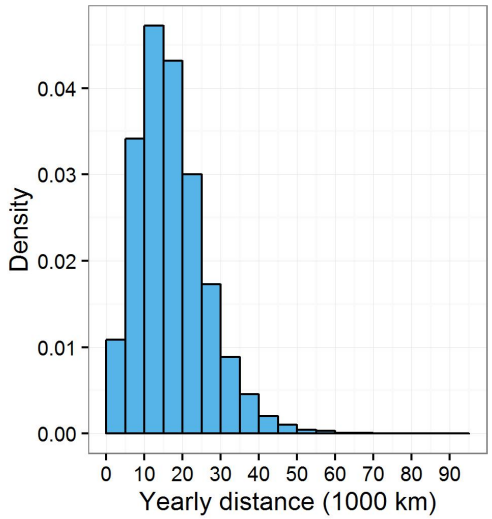
Data set from a Belgian insurer

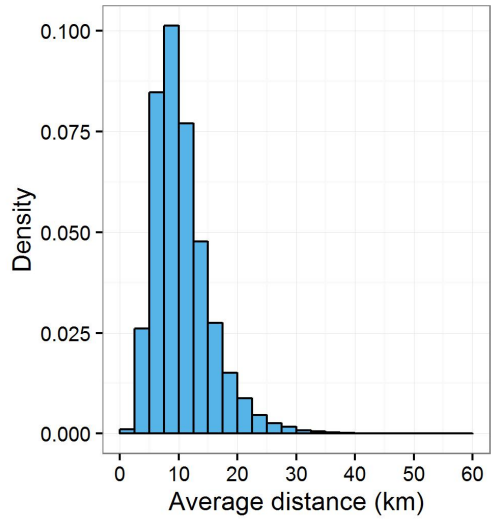


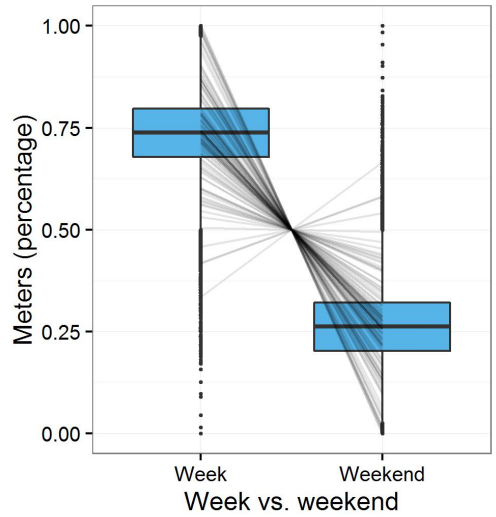
Mind data quality issues!

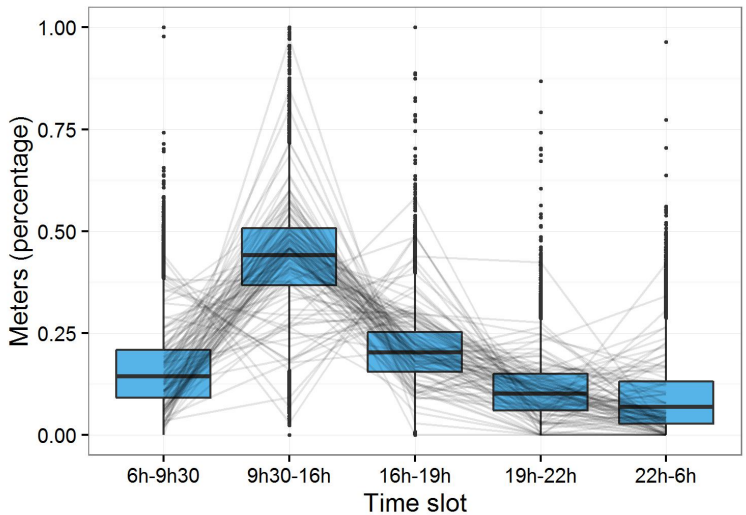
Telematics insurance

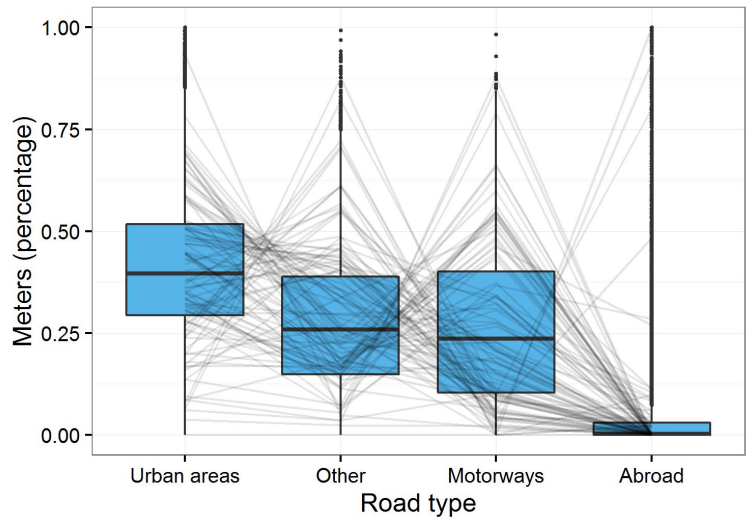
New features











Telematics insurance

New features

(Fictitious) Examples of driver profiles:

James B.

He drives 100 000 km during one year.

His road type composition is $(15\,000, 15\,000, 50\,000, 20\,000)$ or $(0.15, 0.15, 0.5, 0.2)$.

Eugène from Man Bijt Hond

He drives 1 000 km during one year.

His road type composition is $(500, 300, 200, 0)$ or $(0.5, 0.3, 0.2, 0)$.

Telematics insurance

Conclusions of our JRSS C paper

- Telematics information improves predictive power.

  (1) Hybrid model incorporating telematics through additional risk factors is optimal.

  (2) Classic approach performs worse.

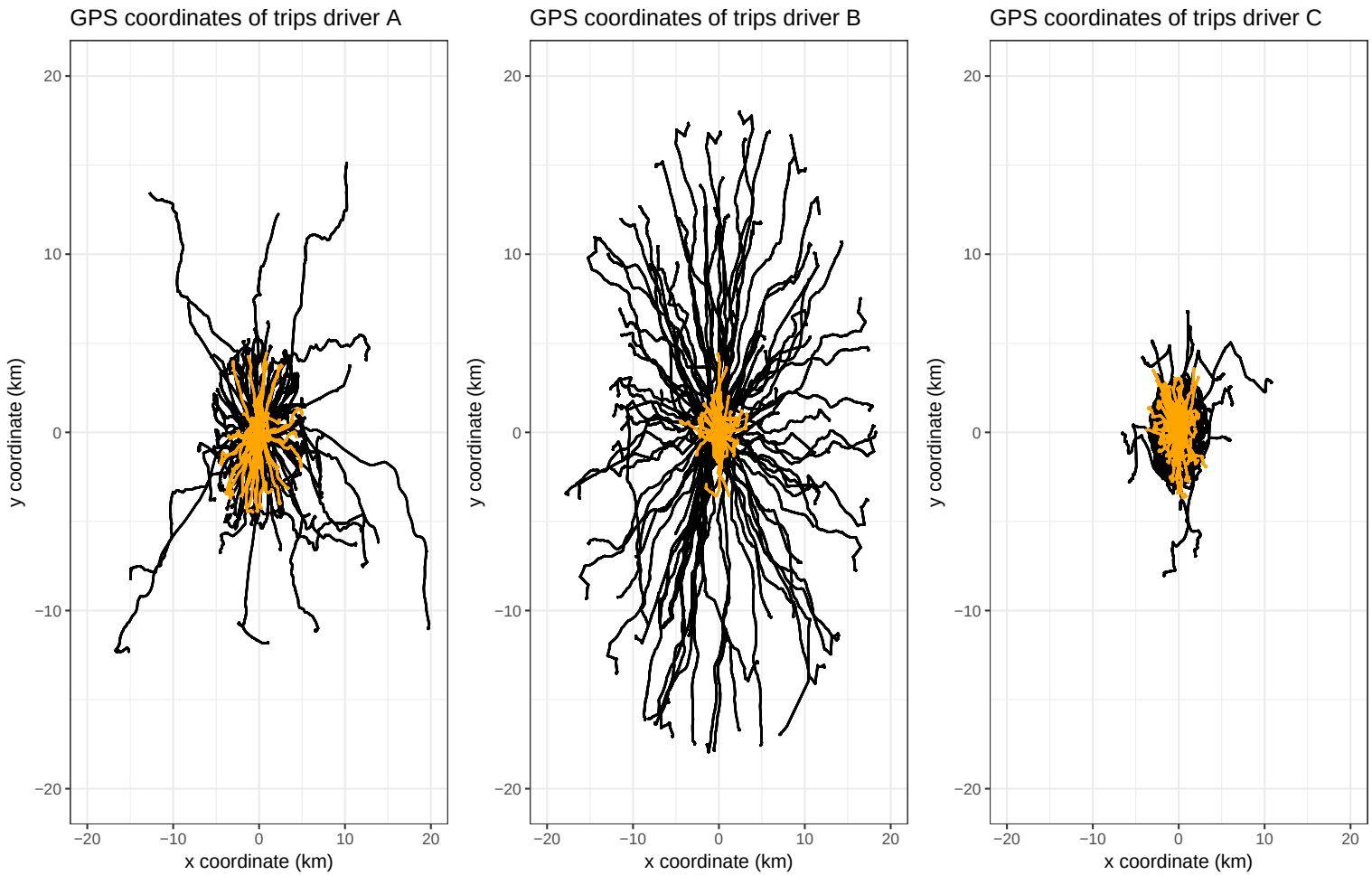
  (3) Gender plays no role anymore in models incorporating telematics information (cfr. Gender Directive).

  Spatial heterogeneity decreases.

  (4) Compositional driving habits have significant impact on riskiness.

Telematics insurance

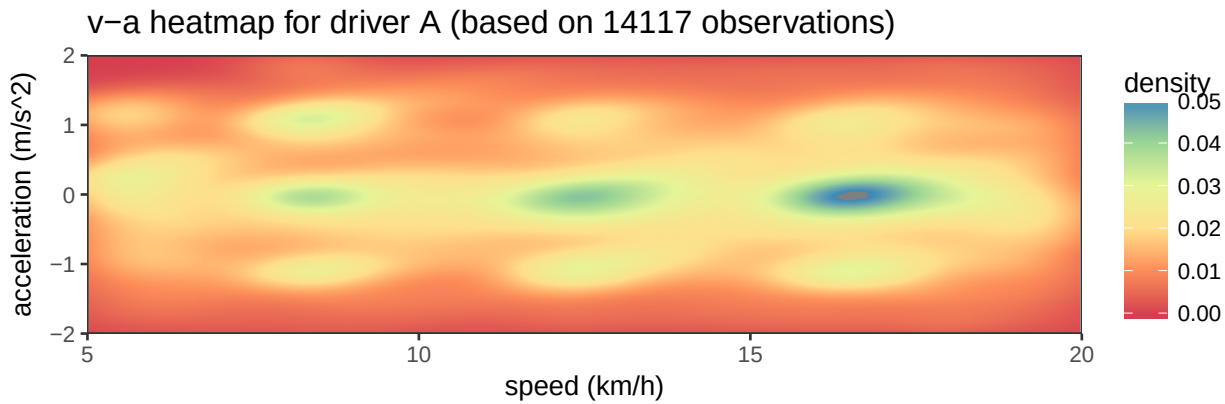
High-frequency GPS data

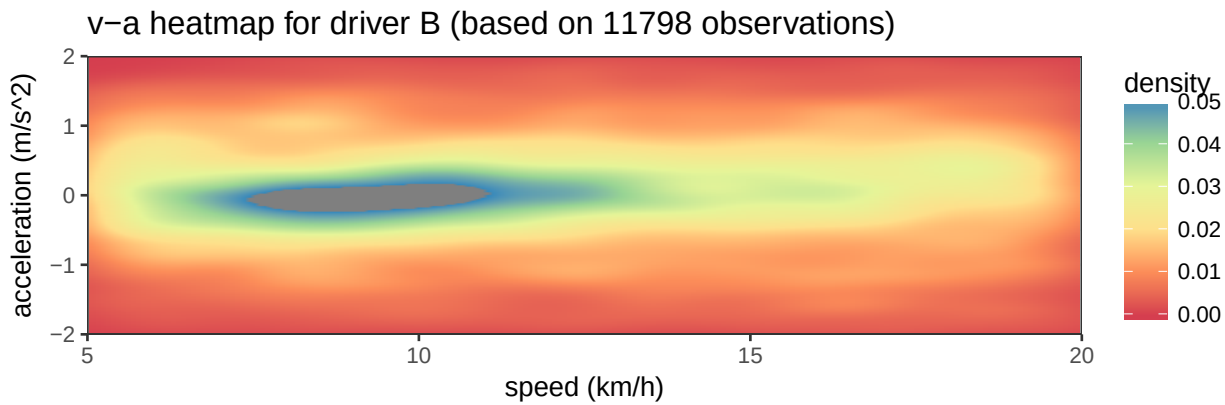


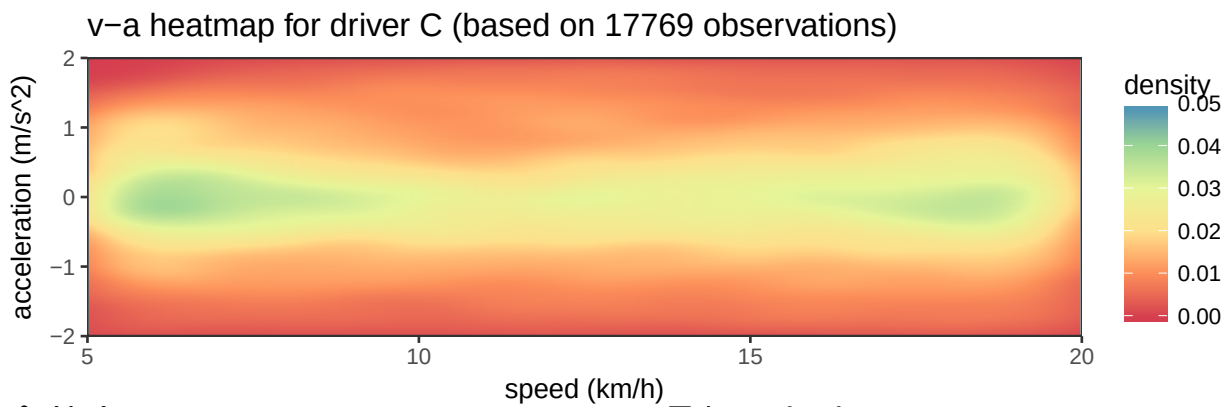
Reflect upon short distance versus long distance driver.

Telematics insurance

v-a heatmaps

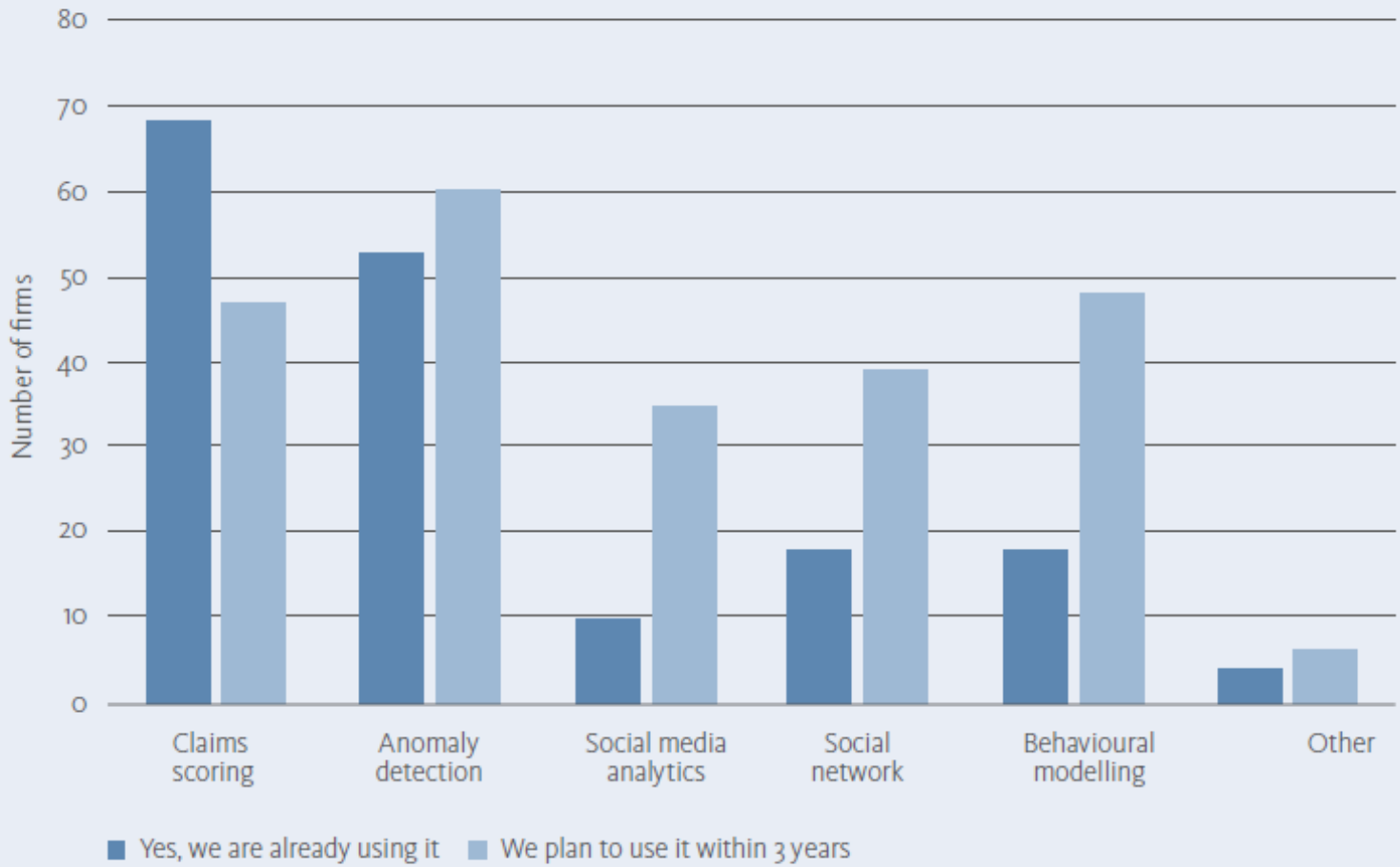






Fraud analytics

Figure 18 – Use of BDA in fraud detection



Source: EIOPA BDA thematic review, based on the classification of tools from Gartner†

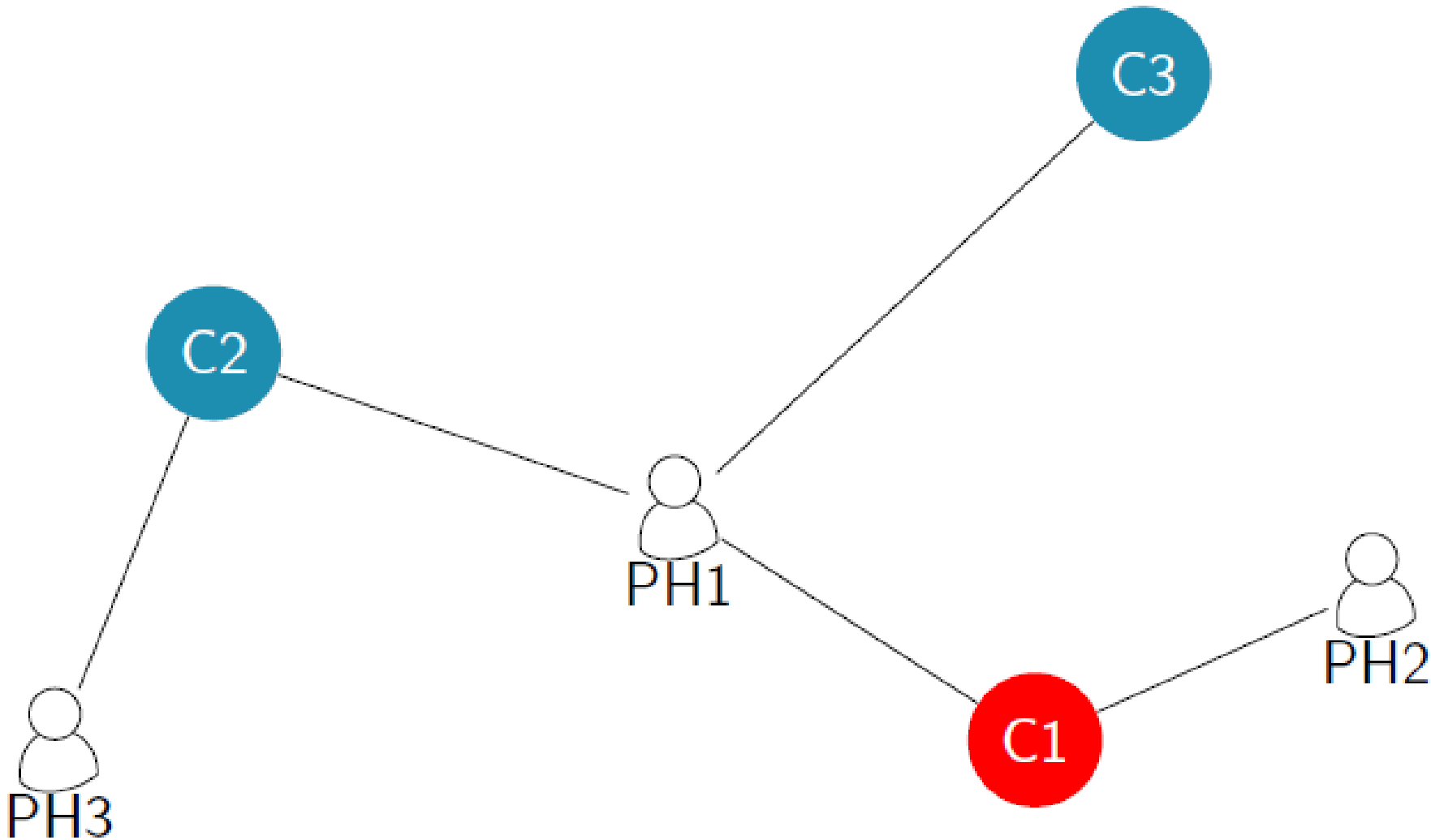
Fraud analytics

The goal

- Build a fraud detection model
  - use ‘classic‘ (or: local) features
  - use network data and extract useful features from network (new!)
  - apply supervised learning (or: unsupervised?, semi-supervised?).
- Flag suspicious claims for further investigation.

Fraud analytics

Network data



Fraud analytics

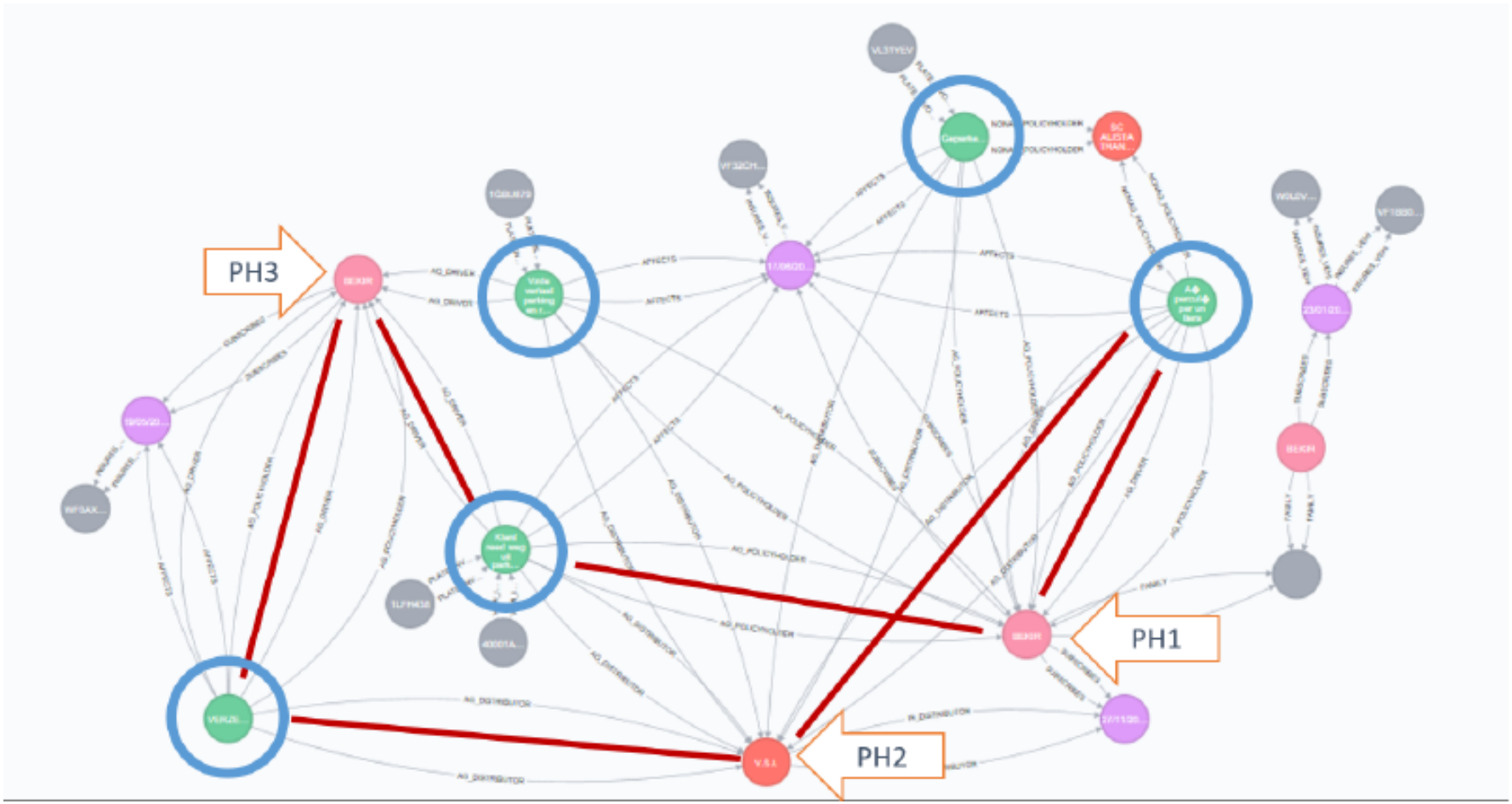
Network data

- Graph database
- Queried with Cypher
- Visualization

- Featurization
- Analytical model building

Fraud analytics

Network data - explore structure



Fraud analytics

Network data - calculate exposure scores

- Use the PageRank idea (Page, Brin et al., 1998)
  - assigns a *PageRank* (score, or measure of importance) to each webpage
  - exploit links between webpages
  - PageRank of webpage i is based on its linking webpages j, with weight

    proportional to the importance (or PageRank) of j

    inversely proportional to the number of pages j points to
  - random surfer with random jumps!
- Use personalized PageRank:

  biased towards (or personalized for) particular set of nodes (here: fraud)!

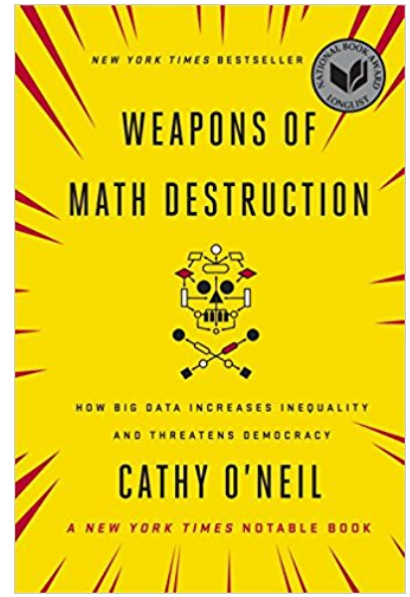
Fraud analytics

Project challenges

- Óskarsdóttir, Antonio et al. (ongoing):
  - large data set with over 4 million nodes
  - very few cases labelled (~ 0.2% of all claims filed can be investigated)
  - high class imbalance complicates supervised learning.

(Provocative) Statement Nr. 3

Cathy O'Neil



We urgently need an academic institute focused on algorithmic accountability.

Quote from *The ivory tower can't keep ignoring tech*, op-ed in The New York Times by Cathy O'Neil, published November 14, 2017.

Big data and insurance: society's perspective

Topics in this debate:

- solidarity, individualisation of insurance
- privacy and data protection
- competition
- innovation.

Read:

Big data and insurance. Implications for innovation, competition and privacy by the Geneva Association (2018).

Verzekeren, technische solidariteit en morele solidariteit from 2017.

Check-list for your project

A take home message from the Big Data debate at Verbond van Verzekeraars (April 1, 2019)

- Kan het?
- Mag het?
- Wil ik het?

These are not new, though become more prominent in the big data era!

(Provocative) Statement Nr. 4

The mindset of the actuary

> The narrative must be that actuaries are entering the data science world not entirely to compete but also to bring the element of the actuarial profession where we build integrity and transparency into any work that we do, and how documentation of that is possible.

Quote from What data science means for the future of the actuarial profession, British Actuarial Journal, June 2018.

Outlook

Common themes in my lab's research lines:

- open the black box (as much as possible) and document
- fill methodological gaps that arise when working with insurance data
- analyze real life data.

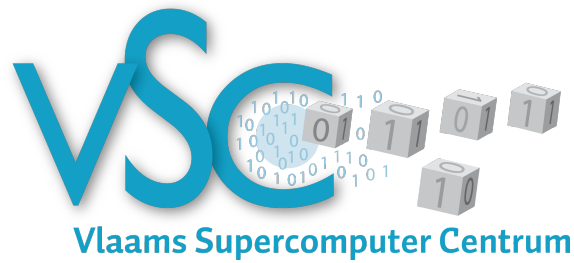
More information

For more information, please visit:

LRisk website, www.lrisk.be

https://katrienantonio.github.io

Thanks to



Research Foundation
Flanders
Opening new horizons

UNIVERSITY OF AMSTERDAM

View online

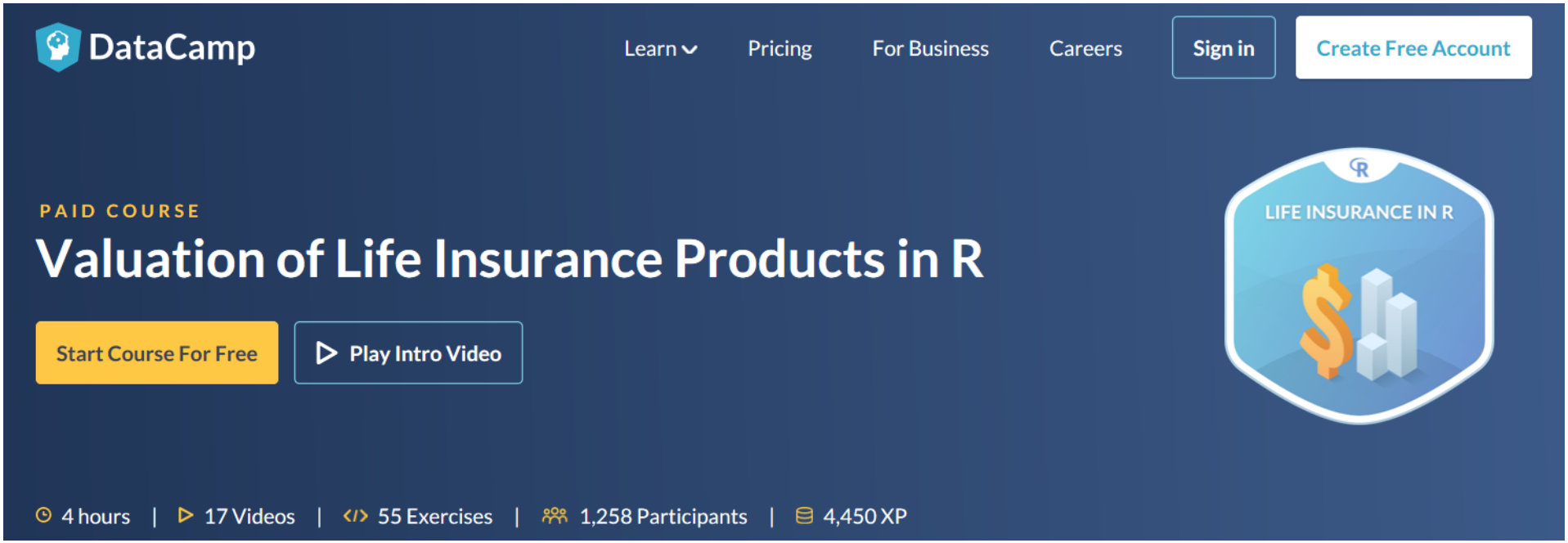
ABS | Postmaster Actuarial Practice Cycle

Insurance Analytics - Voorbij de Hype

Welke datavraagstukken kunnen actuarissen nú wel oplossen die vóór de Big Data innovatie onoplosbaar leken? | 2 PE

Beste Katrien,

Op 11 juni organiseren we vanuit de opleiding het eerste Alumni Event **Insurance Analytics - Voorbij de Hype** voor APC-alumni.



Online course with DataCamp on Valuation of Life Insurance Products in R

designed by Katrien Antonio & Roel Verbelen

http://www.datacamp.com/courses/2333