Reutlingen University

DBKDA 2017, Barcelona, 22.05.2017

Panel discussion on
DBKDA/WEB/GraphSM

Databases and Web Information in the Era of Big Data

Moderation: Fritz Laux, Reutlingen University, Germany

Panelists:

Hai Wang, Department of Finance, Information Systems, and Management Science, Sobey School of Business, Saint Mary's University, Canada

Alain Casali, Laboratoire d'Informatique Fondamentale de Marseille, France

Fritz Laux, Business Informatics, Reutlingen University, Germany

Reutlingen University

Initial Statements/Topics of Panelists

- *Hai Wang: „Knowledge Management and Business Analytics in the Era of Big Data"*
- *Alain Casali:*
  „ Selecting the right data!"
- *Fritz Laux: „Virtual data integration with transaction support"*

Reutlingen University

The Demand for Data Integration

- *Increasing number of heterogeneous sources need to be integrated to…*
  - gain added value (knowledge, insights) for decision support, predictive analysis, performance management, etc.
  - coordinate (complex) processes in (near) real-time with transaction support (e.g. traffic control, Industry 4.0, fight epidemic)
- *There is a need to*
  - not only analyze up-to-date mass data, but
  - write back to the source databases to control actors/systems
- *This calls for Big Data analysis in real-time with transactional support*

Reutlingen University

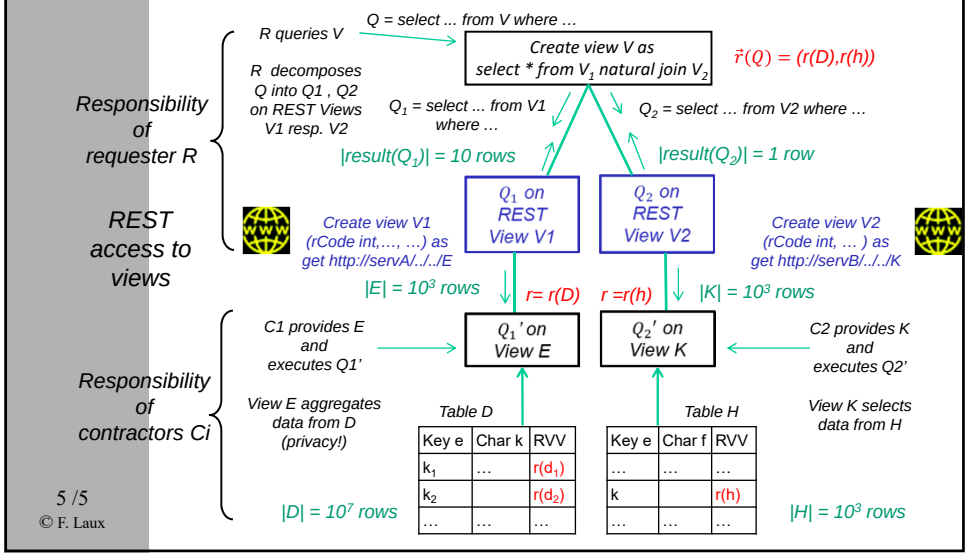
Solution (Idea)

- *A solution must handle*
  - Mass of data →Big Data integration
  - Freshness of data analysis → virtual data integration
  - Transactions → readCheck validation
- *Solution: Virtual data integration using REST programming over the Web with transaction support (PUT, POST)*
  → *Big Live Data*
- *"Big Live Data" = Big Data + Transaction support*
  - Volume, Velocity, and Variety of data impede classical ETL
  - Virtual integration is useful to
    - ⇨ Get latest data
    - ⇨ To reduce data load
    - ⇨ Distribute processing

Reutlingen University

Virtual Integration Example (Schema)

1. *Access latest data through Views*
2. *Reduce / distribute data transfer by caching / query decomposition*
3. *Transaction support through readCheck validator (optimistic CC)*



5 /5

Selecting the Right Data: a Time Series Point of View

Databases and Web Information in the Era of Big Data

Alain Casali

LIF / Aix Marseille Université - France

Wednesday, June 24

Classical Data Selection in RDBMS

A column is removed iff :

- it contains many null/default values;
- it is a duplication from another one;
- it has a small standard deviation;
- the number of distinct values $\notin [MinDV, MaxDV]$

More information: D. Pyle, Data Preparation for Data Mining. Morgan Kaufmann, 1999.

Buy something - 1 day vision

Michelin CrossClimate + 215/65 R17 103V XL Neumáticos de verano
€164 online ★★★★★ 11 product reviews

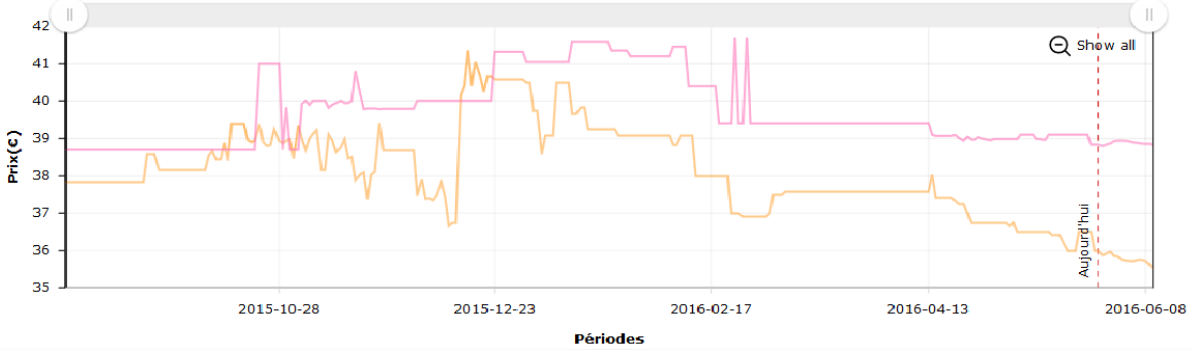
Online stores Your location: Barcelona

☐ Free shipping ☐ Refurbished / used

Sponsored

Sellers	Seller Rating	Details	Base Price	Total Price	
Epto - neumaticos.es	No rating	Free shipping	€163.99	€163.99	Shop
NeumaticosLider.es	★★★★☆ (1,327)	Free shipping	€196.63	€196.63	Shop
neumaticos-guru.es	No rating	Free shipping	€191.70	€191.70	Shop
neumaticos-online.es	★★★★½ (728)		€198.00 +€8.00 shipping	€206.00	Shop
neumaticos-outlet.es	No rating	Free shipping	€192.29	€192.29	Shop
eBay - giga-neumaticos + Show all 2	No rating	Free shipping	€199.10	€199.10	Shop
Tirendo.es	No rating	Free shipping	€196.23	€196.23	Shop
eBay - neumaticos-online	No rating	Free shipping	€203.90	€203.90	Shop
giga-neumaticos.es	No rating		€193.59 +€8.00 shipping	€201.59	Shop

Buy something - in near future



You can delay your buying if needed.

Forecasting Models

- Uni-variate models (AR, ARIMA, ...)
- Multi-variate models (VAR, VECM, ...)

Problems:

1. Sometimes uni-variate models give better forecsating results than multi-variate models;
2. Jiang et al.: "*there is little to no improvement in forecast accuracy when the number of predictors is expanded beyond 20–40 variables*"[1].

[1] http://business.monash.edu/econometrics-and-business-statistics/research/publications/ebs/wp02-17.pdf

Static Features Selection

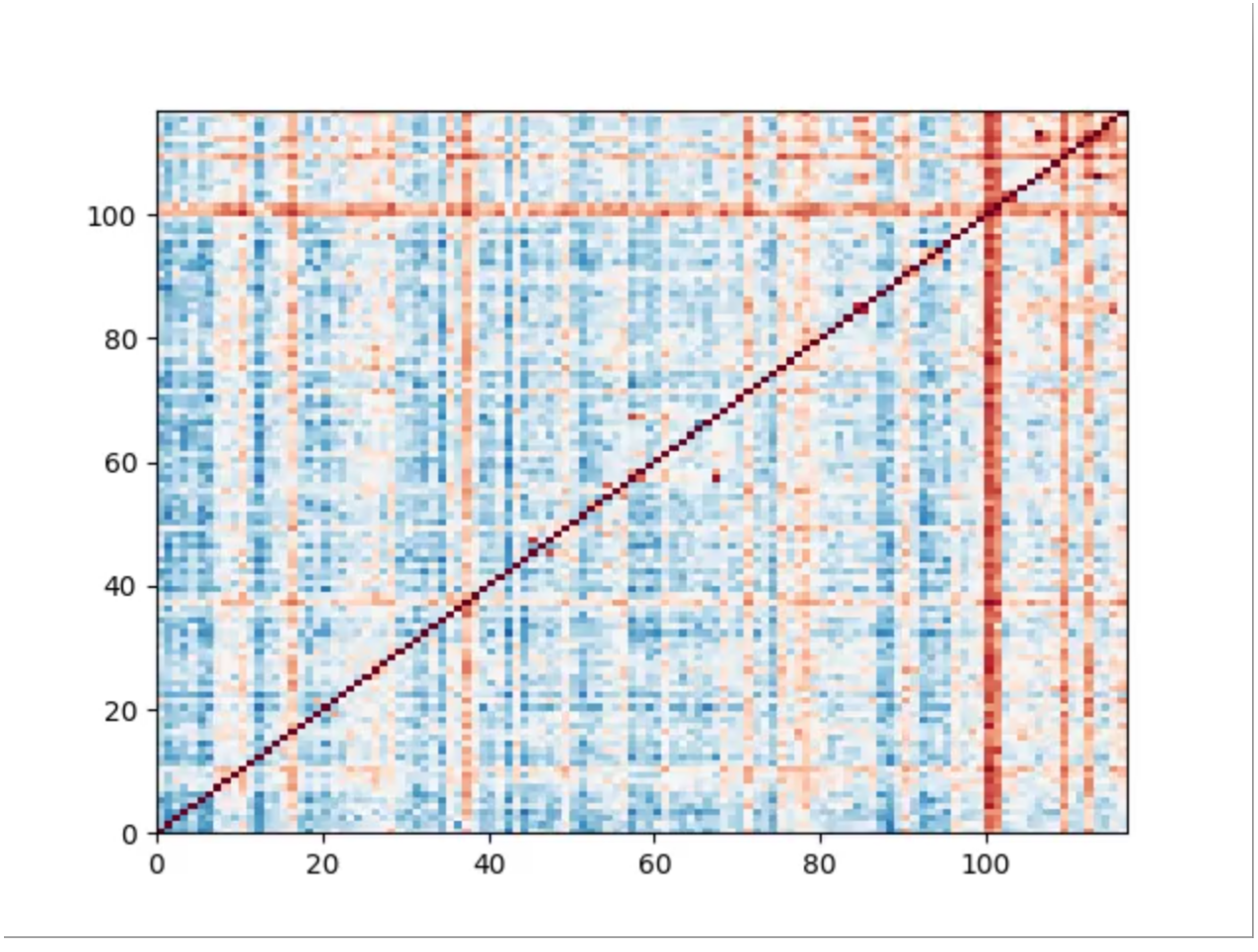
Goal: improve the forecasting results using multi-variate models.
Useful metrics:

- Granger Causality;
- Transfer Entropy;
- ...

Examples:

- Paper $\Leftarrow$ Paper Doe $\Leftarrow$ Wood
  good chain, forecasting results $\nearrow$;
- Paper $\Leftarrow$ lead
  bad chain, forecasting results $\searrow$;

Dynamic Features Selection



Some Solutions / Discussion

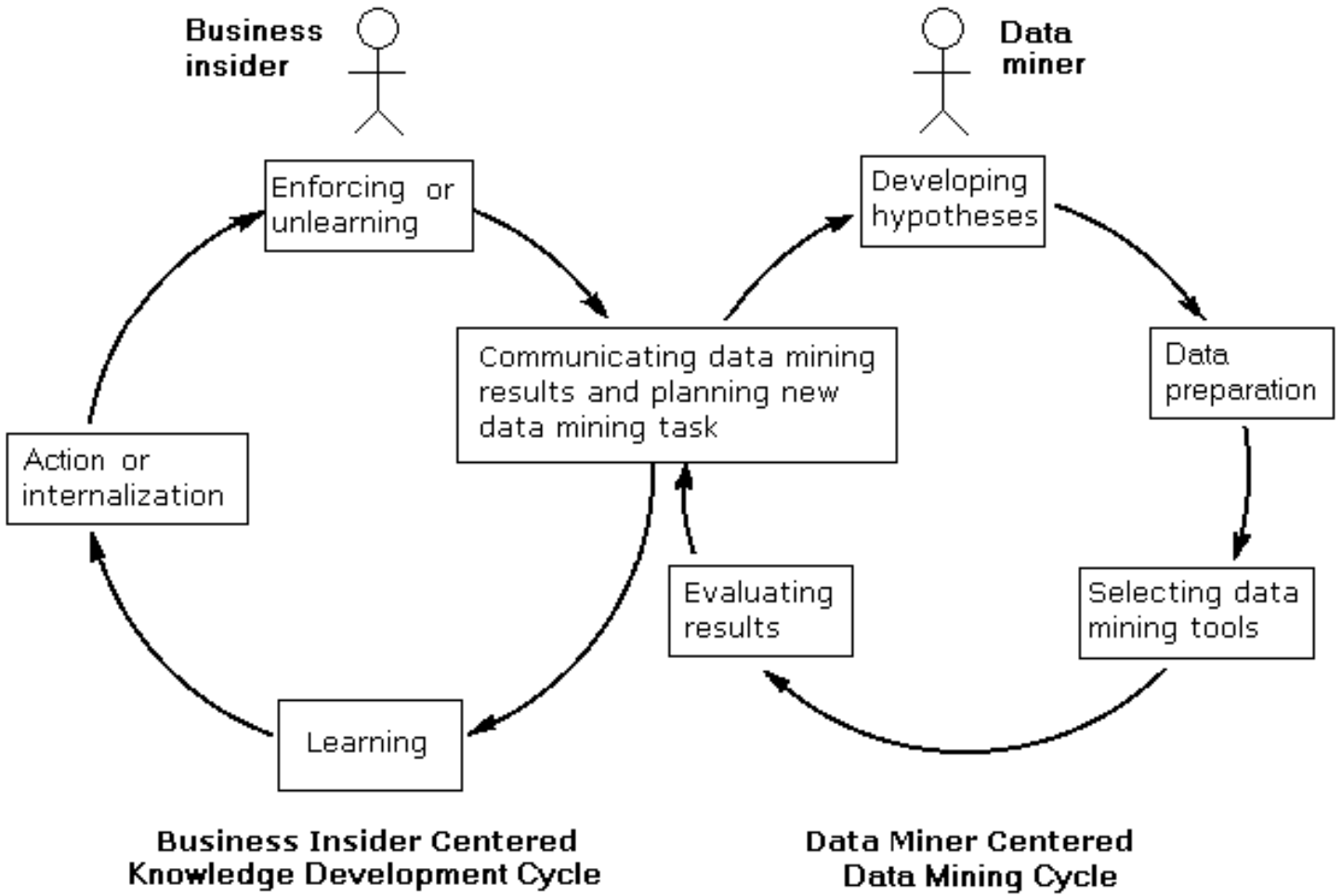
- Selecting the good data is very challenging especially in a big data context (a lot a noisy data);
- Even Amazon needs his little hands ("*turkers*") in MTurk (Amazon Mechanical Turk)[2].

[2] https://requester.mturk.com/tour/data_cleansing

Knowledge Management and Data Mining for Business Analytics

Hai Wang
Sobey School of Business
Saint Mary's University, Canada

Knowledge Management and Data Mining for Business Analytics



Reutlingen University

Outcome of Discussion

- *Analysis of Web Information adds (business) value if*
  - Data is combined/related (virtual data integration)
  - Right data is selected (most influential data, time dimension, Granger causality instead of correlation)
  - Analysis is interpreted in a business context (Knowledge creation)
  - Veracity (reliability, provenience) of data is granted
- *Issues*
  - Privacy is endangered
  - Veracity is hard to verify