Inductive models for automating data science

Luc De Raedt

KATHOLIEKE UNIVERSITEIT LEUVEN

ERC AdG - SYNTH

Synthesising Inductive Data Models

erc

Artificial Intelligence

- Can we build a machine that does X ?
  - X = game playing
  - X = proving a theorem
  - X = recognizing a cup
  - X = communicating in langauge
  - **X = science ?**

The Robot Scientist

- The robot scientists (Ross King et al. Nature 2004, Science 2009)

- Automates certain tasks in the life sciences

    - hypothesizing to explain observations

    - devising experiments to test these hypotheses

    - physically running the experiments using laboratory robotics

    - interpreting the results from the experiments

    - repeating the cycle as required

- Two versions : Adam (functional genomics) and Eve (drug screening)



Artificial Intelligence

- Can we build a machine that does X ?
  - X = game playing
  - X = proving a theorem
  - X = recognizing a cup
  - X = communicating in langauge
  - **X = DATA science ?**

Can we automate data science ?

(if we can, this would be pretty useful)

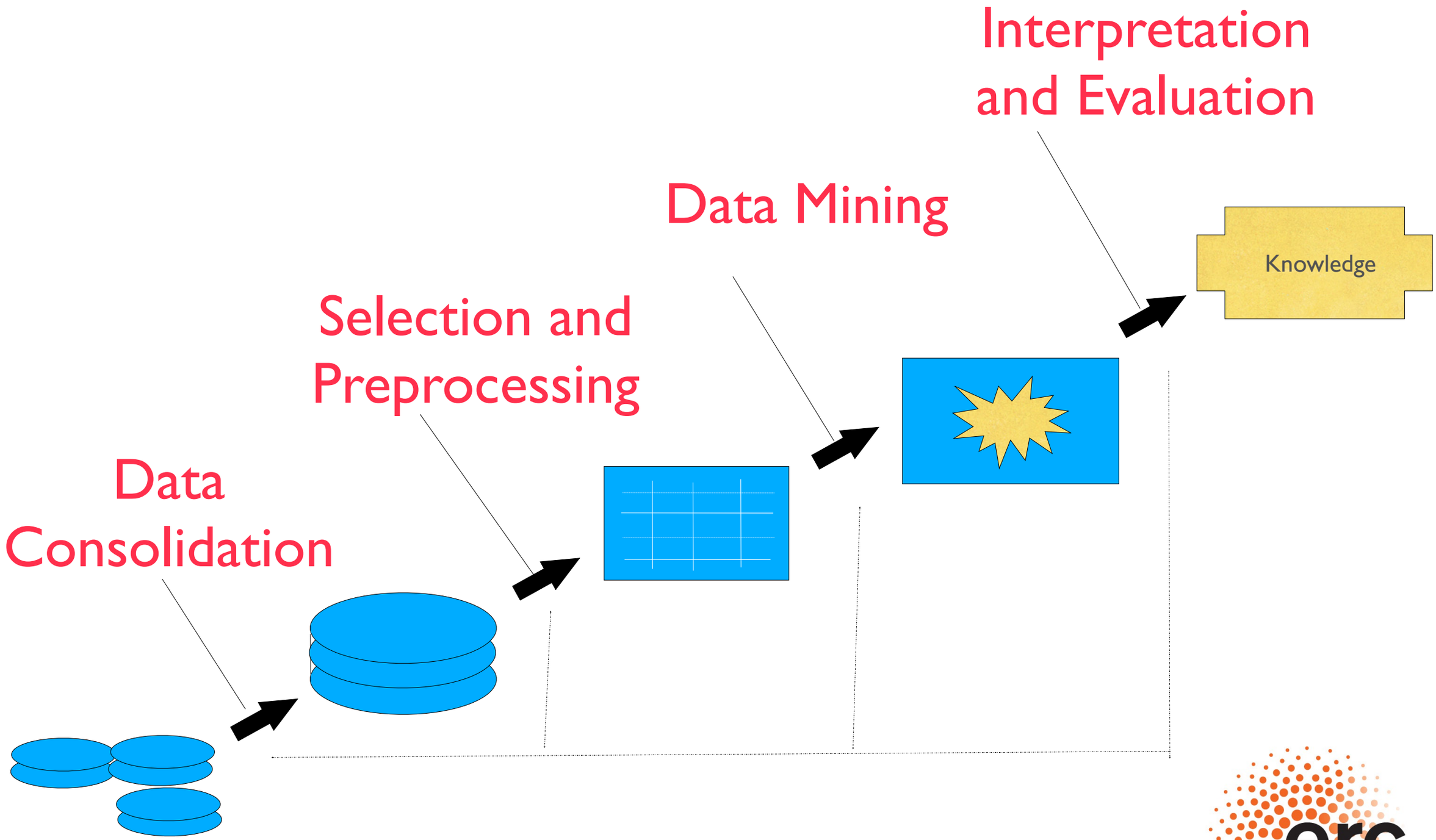
2017 Report of the Consumer Intelligence Series, Bot. Me' (PWC) automated data-analysists is second on the list of potentially most impactful AI techniques in business



Dagstuhl Seminar
Automating Data Science - AutoDS
1-5 October 2018



The KDD process

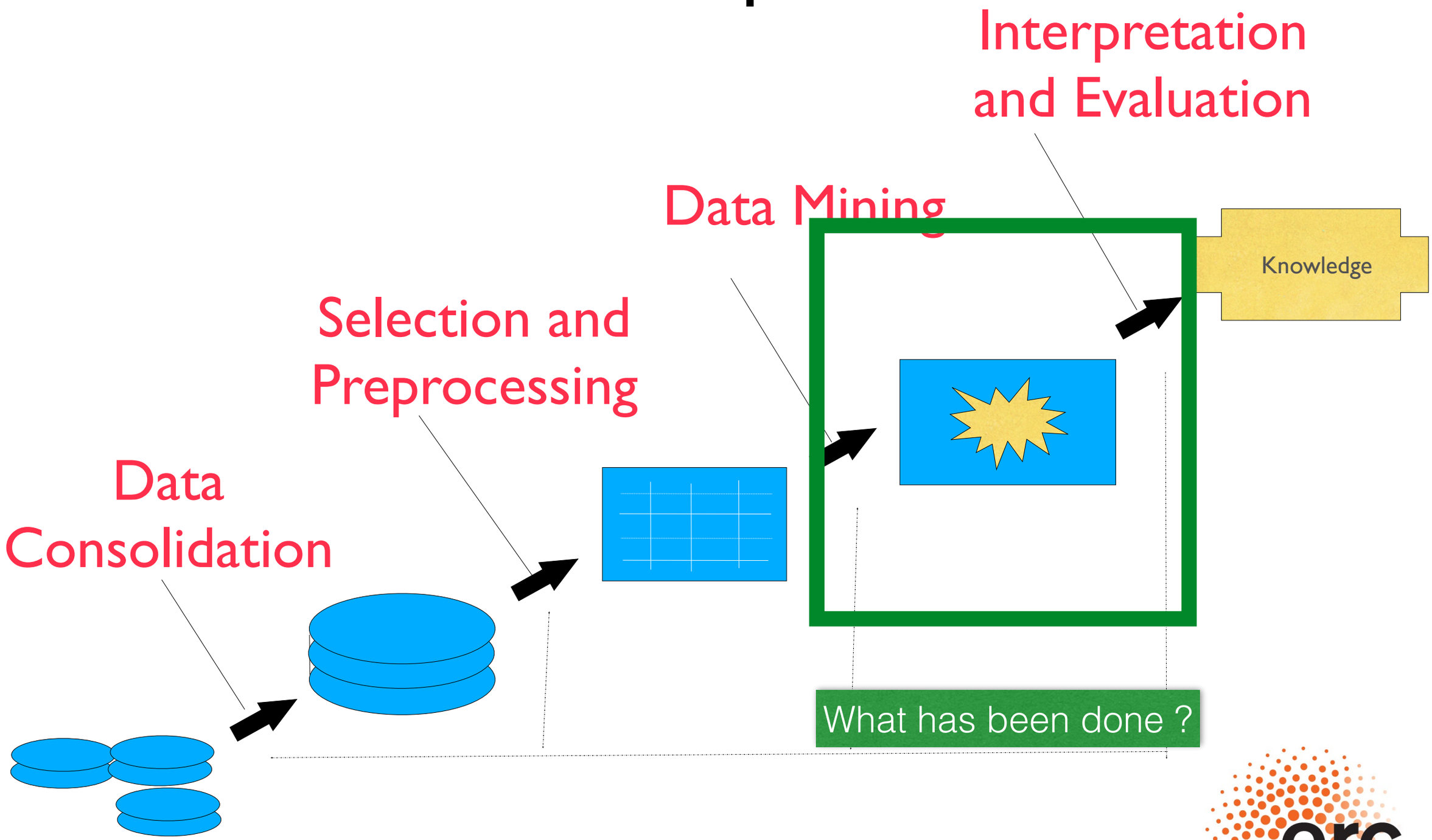


it is about automating the whole process

What has been done ?

1. AutoML - meta-learning and parameter configuration
2. model learning
3. data wrangling

The KDD process



Fayyad et al. 96

What has been done ?

1. **AUTO-ML - meta learning**

- **meta-learning** has been studied for at least 25 years
  - use decision trees or neural nets ?

- is quite popular and very effective (e.g. Auto WEKA, Auto SKlearn)

- automatic algorithm configuration / hyper parameter tuning (e.g. Hoos, Hutter et al.)

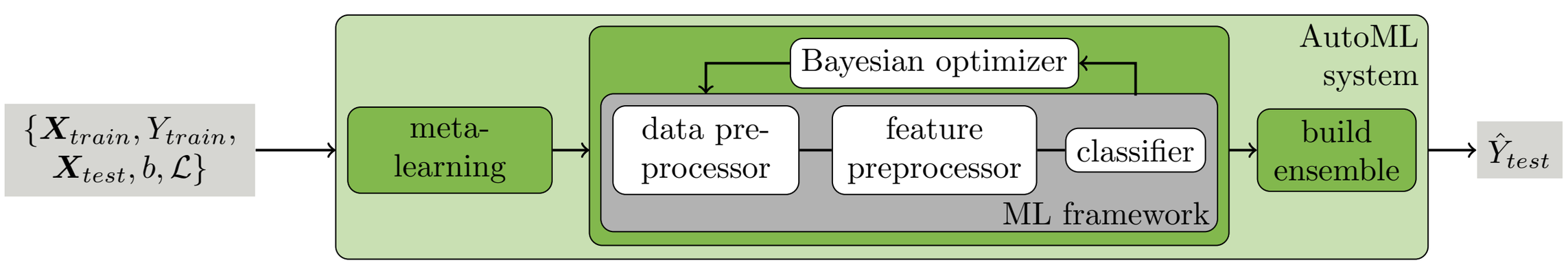
Algorithm selection and configuration

AutoML

Given a dataset

+ a task (e.g. predict X)

Select best algorithm and parameters



Examples: AutoWeka, AutoSkLearn

http://www.cs.ubc.ca/labs/beta/Projects/autoweka/
https://automl.github.io/auto-sklearn

task needs to be known

preprocessing has been accomplished

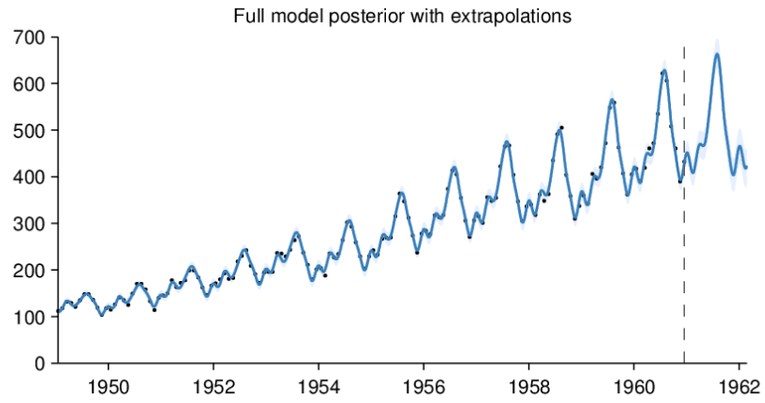
erc

What has been done ?

2. **model-learning**

https://www.automaticstatistician.com

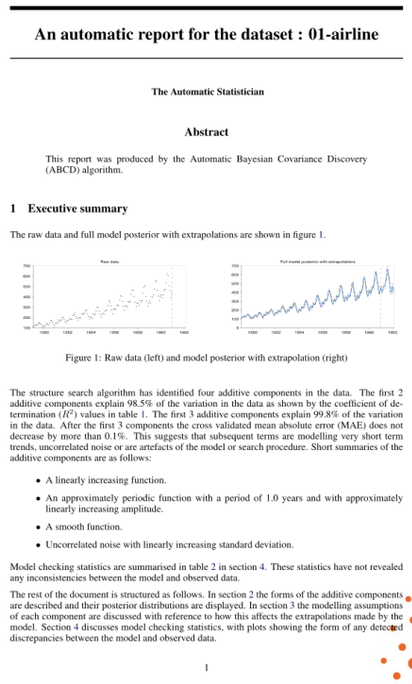
Automated statistician



Given 1) a dataset +

2) a class/space of statistical models

Generate a statistical report



An automatic report for the dataset : 01-airline

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.8% of the variation in the data. After the first 3 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with approximately linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.

Model checking statistics are summarised in table 2 in section 4. These statistics have not revealed any inconsistencies between the model and observed data.

The rest of the document is structured as follows. In section 2 the forms of the additive components are described and their posterior distributions are displayed. In section 3 the modelling assumptions of each component are discussed with reference to how this affects the extrapolations made by the model. Section 4 discusses model checking statistics, with plots showing the form of any detected discrepancies between the model and observed data.

1

preprocessing has been accomplished

only one type of analysis

report meant for statisticians

- Darpa's current projects on *Data-Driven Discovery of Models*

What has been done ?

Automated statistician

https://www.automaticstatistician.com

Given a dataset +
a class/space of statistical models

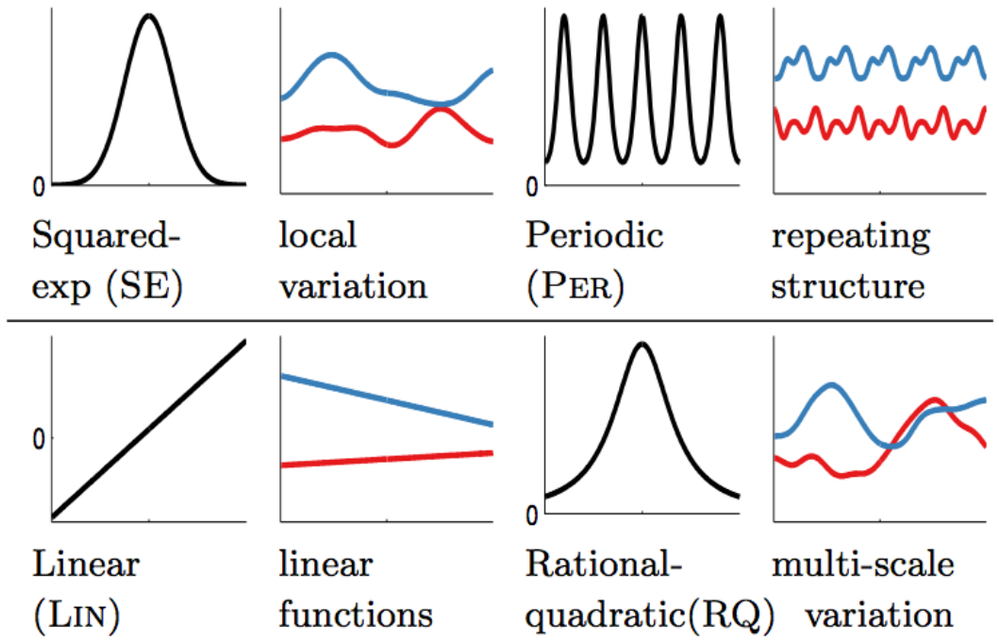


Figure 1. Left and third columns: base kernels $k(\cdot, 0)$. Second and fourth columns: draws from a GP with each repective kernel. The x-axis has the same range on all plots.

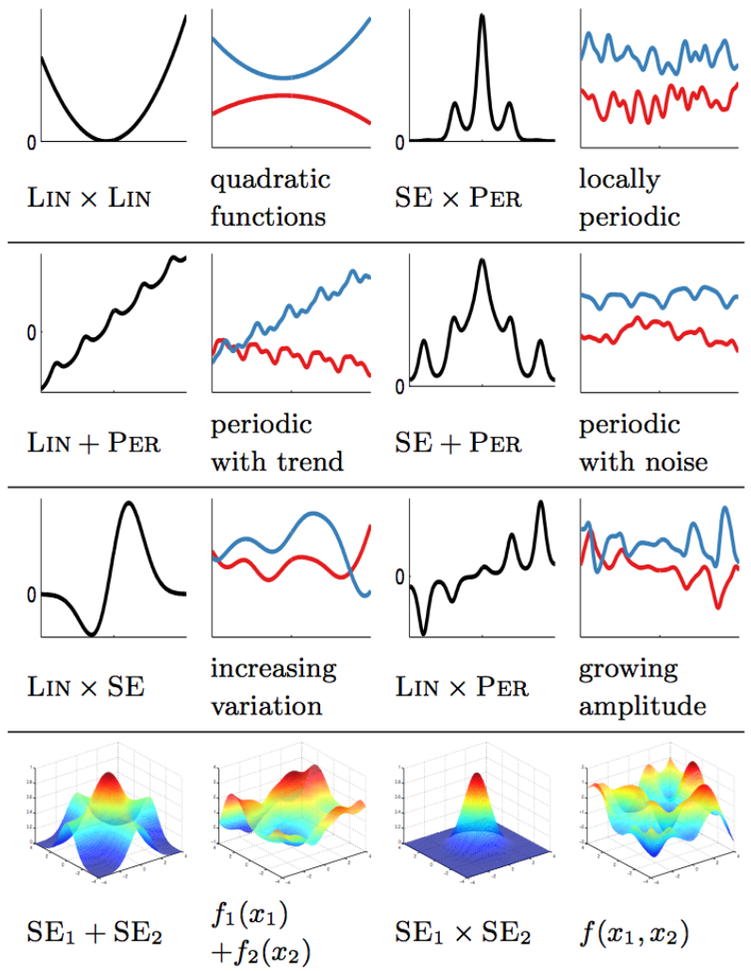


Figure 2. Examples of structures expressible by composite kernels. Left column and third columns: composite kernels $k(\cdot, 0)$. Plots have same meaning as in Figure 1.

[Duvenaud et al. ICML 13]

An automatic report for the dataset : 01-airline

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

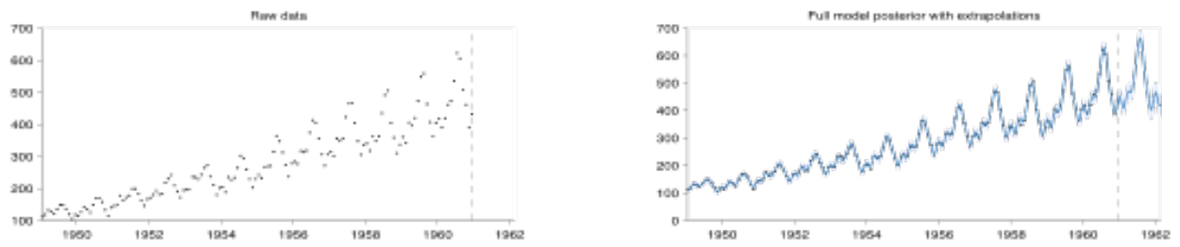
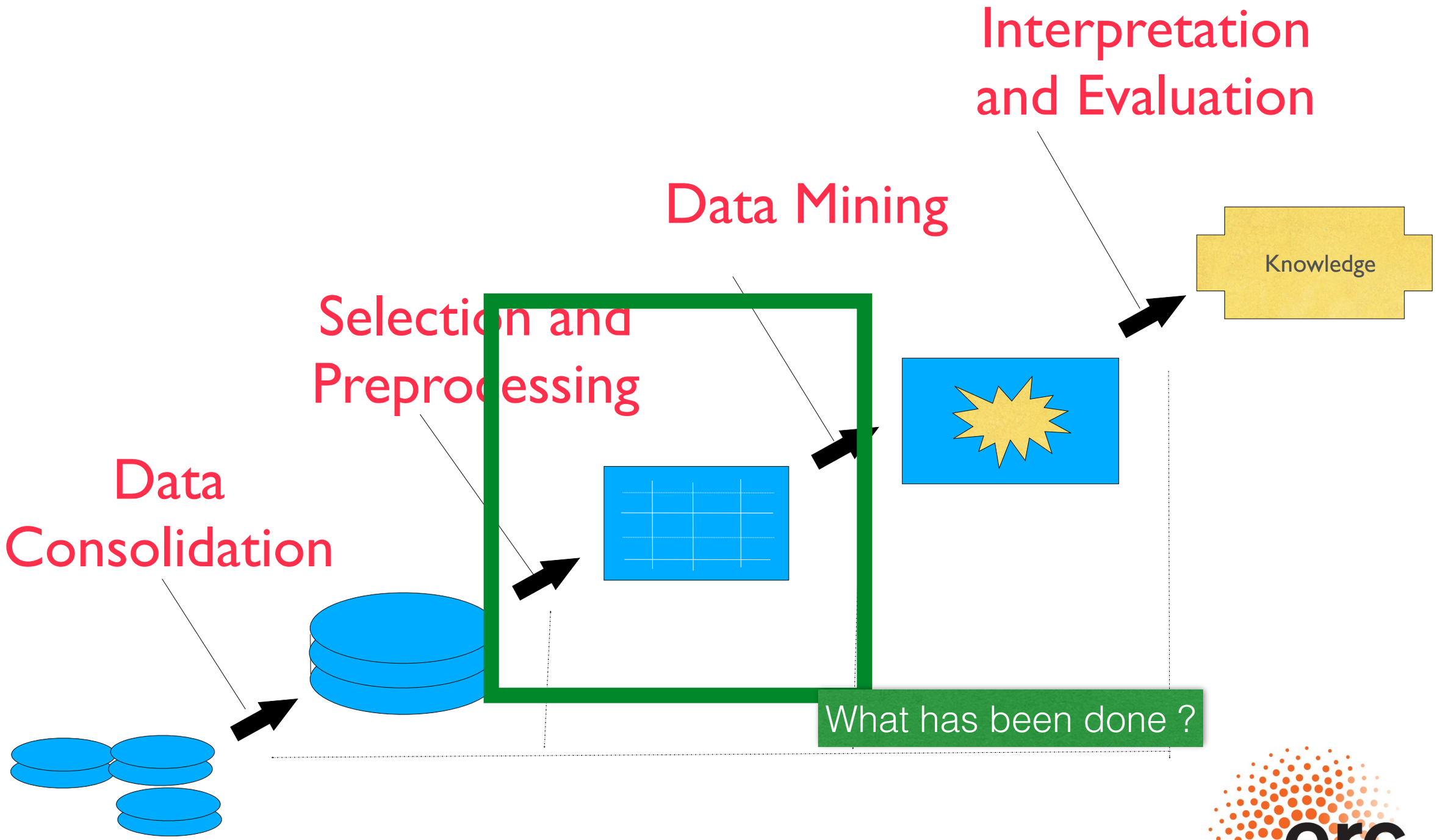


Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.8% of the variation in the data. After the first 3 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with approximately linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.

The KDD process



Fayyad et al. 96

3. **data wrangling**

Data Wrangling Challenges

[Slide by Chris Willams, from his excellent tutorial at Dagstuhl]

- **DP** Data parsing, e.g. converting csv's or tables
- **DD** Obtaining (or inferring) a data dictionary: basic types + semantics
- **DI** Data integration: Combining data from multiple sources
- **ER** Entity resolution: Recognising that two distinct pieces of information in the data concern the same entity. Includes deduplication and record linkage
- **FV** Format variability: e.g. for dates, but also for variability in names (e.g. IBM, I.B.M.).
- **SV** Coping with structural variability in the data, e.g. wide vs tall format. Also variation over time.
- **MD** Identifying and repairing missing data
- **AD** Anomaly detection and repair

What has been done ?

data wrangling

- Preparing data for usage (a lot of time required)

	A	B	C	D	E	F
1	Country Code	Indicator	1962	...	2013	2014
2	TKI	AG.LND.TOTL.K2	139960,00		139960,00	
3		AG.LND.PRCP.MM				691,00
4		EG.USE.PCAP.KG.OE			310,74	306,49

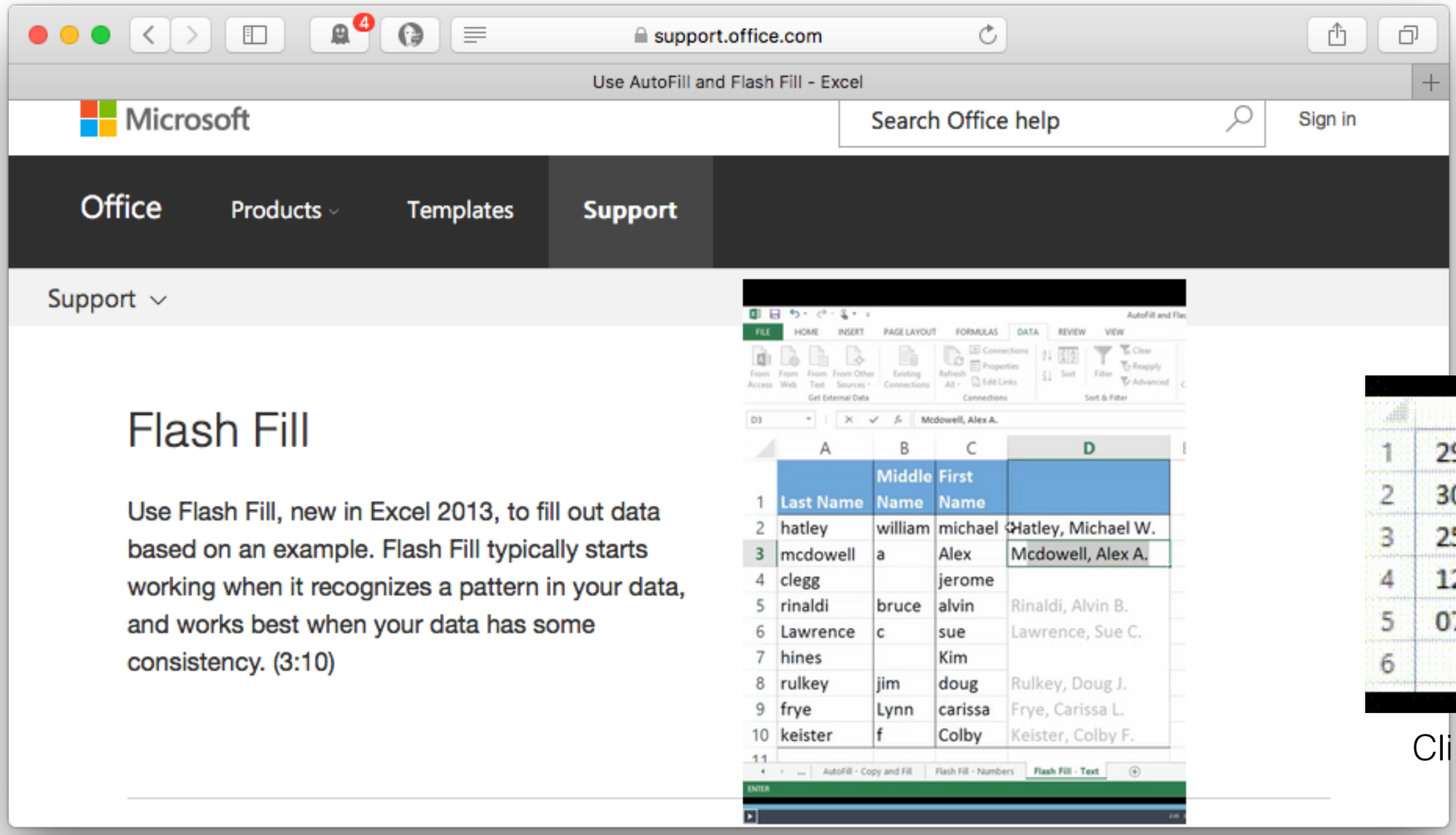
↓

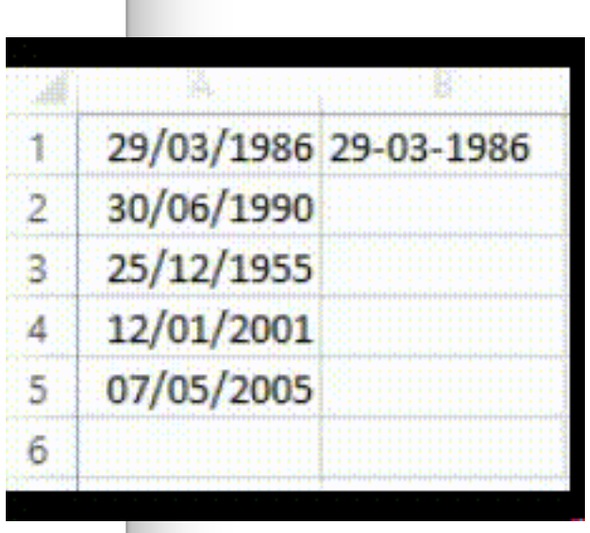
	A	B	C	D
7	Country Code	Indicator	Year	Value
8	TKI	AG.LND.TOTL.K2	1962	139960,00
9	TKI	AG.LND.TOTL.K3	2013	139960,00
10	TKI	AG.LND.PRCP.MM	2014	691,00
11	TKI	EG.USE.PCAP.KG.OE	2013	310,74
12	TKI	EG.USE.PCAP.KG.OE	2014	306,49

- Can we automatically synthesise the programs ?
- Can we program by example ?

FlashFill

Microsoft





Clip:Lidia Contreras

[Sumit Gulwani, Popl 11 and subsequent publications]

erc

Data Wrangling

FOOFAH

Given some input → output examples

	A	B	C	D	E	F
1	**Country Code**	**Indicator**	**1962**	**...**	**2013**	**2014**
2	TKI	AG.LND.TOTL.K2	139960,00		139960,00	
3		AG.LND.PRCP.MM				691,00
4		EG.USE.PCAP.KG.OE			310,74	306,49

↓

	A	B	C	D
7	**Country Code**	**Indicator**	**Year**	**Value**
8	TKI	AG.LND.TOTL.K2	1962	139960,00
9	TKI	AG.LND.TOTL.K3	2013	139960,00
10	TKI	AG.LND.PRCP.MM	2014	691,00
11	TKI	EG.USE.PCAP.KG.OE	2013	310,74
12	TKI	EG.USE.PCAP.KG.OE	2014	306,49

learn a **transformation program** that generalises to the whole table, using a modified A* algorithm.

Data Wrangling

FlashExtract

Given a spreadsheet

	A	B	C	D	E	F
1	**Country Code**	**Indicator**	**1962**	**...**	**2013**	**2014**
2	TKI	AG.LND.TOTL.K2	139960,00		139960,00	
3		AG.LND.PRCP.MM				691,00
4		EG.USE.PCAP.KG.OE			310,74	306,49

and some +/- examples

TKI	AG.LND.TOTL.K2	1962	139960,00
2013	AG.LND.TOTL.K3		139960,00

learn a program that **extracts** the desired relation.

Learning cell and spatial constraints that satisfy positive examples, combining them and searching for combinations that exclude negative examples.

still requires a user knowledgeable of data mining/wrangling



What if

1.AutoML assumes the learning task is known.

what if we do not know the learning task ?

can we automatically determine the right one ?

2. Model learning

can we automatically determine the type of models to consider ?

3. Data wrangling: what if the data still need to be pre-processed ?

can we automatically select the right features ? the dependent from the independent variables ?

can we automatically transform the data in the right form ?

What if

1.AutoML assumes the learning task is known.

what if we do not know the learning task ?

can we automatically determine the ri

KEY OPEN QUESTIONS

2. Model learning

can we automatically he the type of models to consider ?

3. Data wrangling t if the data still need to be pre-processed ?

can we automatically select the right features ? the dependent from the independent variables ?

can we automatically transform the data in the right form ?

erc

Our Setting

The SYNTH setting in a nutshell

What can one discover in a given dataset ?

Given

- a dataset (a set of relational / excel tables)
- some partial new input to that dataset

Find

- autocomplete … (fill out / predict possible values for a naive end-user / correct errors)

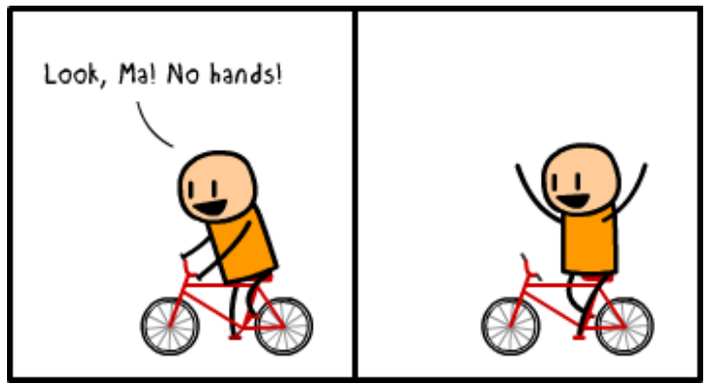
The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

slide by Samuel Kolb



Our setting



Autocompletion :

- Democratizing data science … bring it to the (naive) end-user;
- It is not just BIG data that matters; most datasets are small;

Assumptions :

- we have a *systematic* user of a relational database / a spreadsheet
- there are some regularities and generalisation / data analysis makes sense

Desiderata :

- the user just sees his spreadsheet / database, hide the rest
- the no hands approach (the less interaction, the better)

Human-Like computation

- how would people do this ?

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

slide by Samuel Kolb



The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

Learning constraints and formulas

TACLE SYSTEM (Kolb et al, ML Journal 2017)

Total = SUM(June, July, August)

FK from Type (Table 2) to Type (Table 1)

slide by Samuel Kolb

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

Building predictive models

slide by Samuel Kolb

Predictors for August

Predictors for Profit

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

Predictive auto-completion

unpublished

slide by Samuel Kolb

Jointly complete August, Total, Profit

Formulas + constraints + (specialized) predictors



Students in a master program

- Say you are the program director …
- You accumulate data about students applying in spreadsheets
- The university accumulates data about their results (a year later)
- Can we automatically predict / recommend
  - good courses to follow ? what if questions ?
  - whether to admit or not ? whether students would pass or fail ?
  - provide any other advice to the program director ?

Synth's approach

1. **Predictive autocompletion** as one answer to …

   what if we do not know the learning task ?

   can we automatically determine the right one ?

2. Model learning

   can we automatically determine the type of models to consider ?

   learn constraints + predictive models

   reasoning with these in a probabilistic model

3. Data wrangling: what if the data still need to be pre-processed ?

   can we automatically select the right features ? the dependent from the independent variables ?

   can we **automatically transform the data** in the right form ?



Synth's approach

1. **Predictive autocompletion** as one answer to …

   what if we do not know the learning task ?

   can we automatically determine the right one ?

2. Model learning

   can we automatically determine the type of models to consider ?

   learn constraints + predictive models

   reasoning with these in a probabilistic model



Learning constraints and formulas

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

slide by Samuel Kolb

Learning constraints and formulas

The Magical Ice Cream Factory

T1

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?

Type	Cost
Chocolate	60
Banana	40
Speculaas	70
Vanilla	40

T2

Detecting tables

slide by Samuel Kolb



Learning constraints and formulas

The Magical Ice Cream Factory

B1 B2 B3

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?

Type	Cost
Chocolate	60
Banana	40
Speculaas	70
Vanilla	40

B4 B5

Detecting blocks of the same type

Properties: Type, vector count, vector length, table, …

slide by Samuel Kolb



Learning constraints and formulas

The Magical Ice Cream Factory

B1 B2 B3

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?

Type	Cost
Chocolate	60
Banana	40
Speculaas	70
Vanilla	40

B4 *B5*

Finding input-block assignments for constraint templates

e.g. row-wise SUM

B2 = SUM(B2)

e.g. FK

FK(*B4*, B1) FK(*B4*, B3) FK(*B5*, B2) FK(B1, *B4*) FK(B1, *B5*) FK(B3, *B5*)

Learning constraints and formulas

The Magical Ice Cream Factory

T1[1] T1[3:5] T1[6]

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

T2[1]

Finding subblock assignments for constraint templates

e.g. row-wise SUM

T1[6] = SUM(T1[3:5])

e.g. FK

FK(T2[1], T1[1])

slide by Samuel Kolb



Predictive auto-completion

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

Ordering

e.g. August - Total - Profit

August = f_{A1} Total = f_{T1} Profit = f_{P1}

August = f_{A2} Total = f_{T2} Profit = f_{P2}

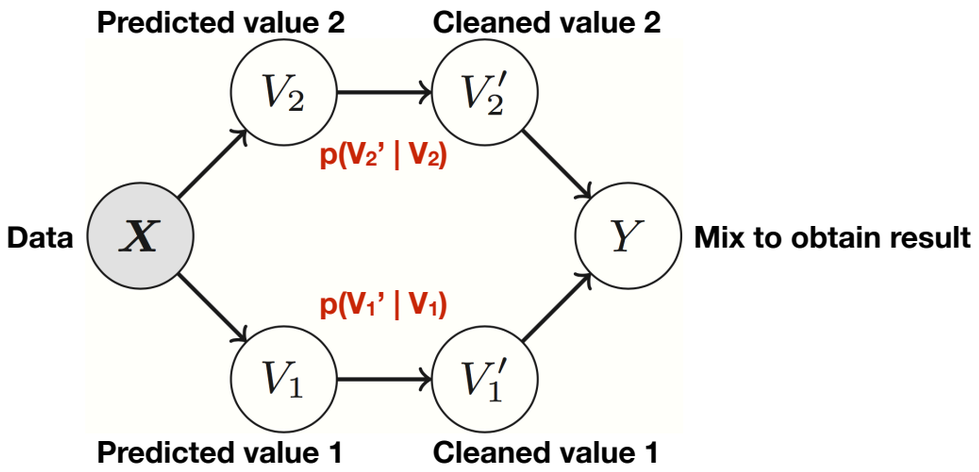
slide by Samuel Kolb

Predictive auto-completion

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	?	?	?
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

Clean up predictions

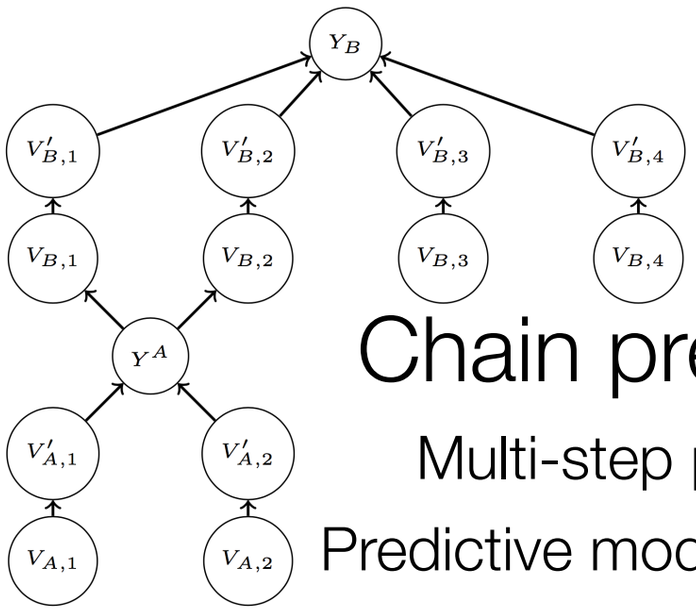


slide by Samuel Kolb

Predictive auto-completion

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	270	1050	NO
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				



Chain predictions

Multi-step predictions

Predictive models + formulas

Obtain joined probability distribution

Condition on constraints

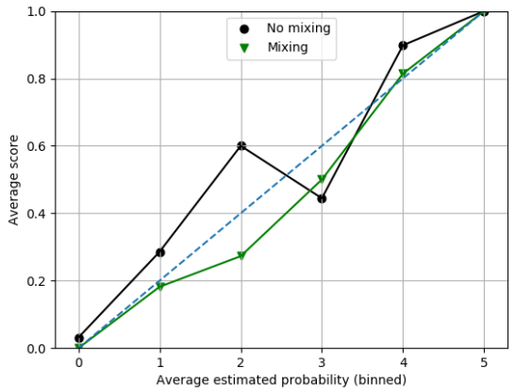
slide by Samuel Kolb

Predictive auto-completion

The Magical Ice Cream Factory

Type	Country	June	July	August	Total	Profit
Vanilla	BE	610	190	670	1470	YES
Banana	BE	170	690	520	1380	YES
Chocolate	BE	560	320	140	1020	YES
Banana	DE	610	640	320	1570	NO
Speculaas	BE	300	270	290	860	NO
Chocolate	FR	430	350	270	1050	NO
Banana	DE	250	650	?	?	?
Chocolate	NL	210	280	?	?	?
	Type	Cost				
	Chocolate	60				
	Banana	40				
	Speculaas	70				
	Vanilla	40				

Estimating correctness



slide by Samuel Kolb

Predictive Autocompletion

Modeling part

- Constraint learning
- Predictive models X,Y,Z -> W,U to fill (sets of) cells
- Autocompletion = reasoning with these two

Constraint Learning

Constraint Learning

- Everybody uses constraints in AI but only few learn them from data
- Learning constraint theories =adapt standard ML to new settings, results in new challenges
- Yields explainable AI models
- New opportunities for constraint optimisation
- Human like -computation — any results on constraint learning ?

See AAAI 18, IJCAI 18, Tutorial Slides &

also [De Raedt et al. AAAI 18] for more details and ref



Constraint Learning

ID	Salesperson	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter	Total	Rank	Label	Items sold total	Max items sold
1	Diana Coolen	353	378	396	387	1514	2	Great	34	20
2	Marc Desmet	370	408	387	386	1551	1	Great	29	10
3	Kris Goossens	175	146	167	203	691	3	Low	19	19
4	Birgit Kenis	93	98	96	105	392	4	Low	17	15

T1

B1 = T1[:, 1] B2 = T1[:, 2] B3 = T1[:, 3:8] B4 = T1[:, 9] B5 = T1[:, 10:11]

T2

Total	991	1030	1046	1081	4148
Average	247.75	257.5	261.5	270.25	1037
Max	370	408	396	387	1551
Min	93	98	96	105	392

B6 = T2[1:4, :]

Salesperson	Items sold
Diana Coolen	5
Marc Desmet	10
Marc Desmet	8
Diana Coolen	9
Birgit Kenis	15
Marc Desmet	8
Birgit Kenis	2
Diana Coolen	20
Marc Desmet	3
Kris Goossens	19

T3

B8 = T3[:, 1] B9 = T3[:, 2]

Quarter	Income	Expenses	Total
Q1	991	212	779
Q2	1030	710	1099
Q3	1046	137	2008
Q4	1081	240	2849

T4

B10 = T4[:, 1] B11 = T4[:, 2:4]

Customer	Contact	Contact Name
Frank	1	Diana Coolen
Sarah	3	Kris Goossens
George	3	Kris Goossens
Mary	2	Diana Coolen
Tim	4	Birgit Kenis

T5

B12 = T5[:, 1] B13 = T5[:, 2] B14 = T5[:, 3]

Row-wise sum

- Syntax: $B_r = SUM_{row}(\mathbf{B_x})$
- Signature: B_r and $\mathbf{B_x}$ are *numeric*; $columns(\mathbf{B_x}) \geq 2$; and $rows(\mathbf{B_x}) = length(B_r)$
- Definition: $B_r[i] = \sum_{j=1}^{columns(\mathbf{B}_x)} row(i, \mathbf{B}_x)[j]$

reminiscent of [Gulwani et al.]
TACLE system
see [Kolb et al, MLJ 17]

$$SERIES(T_1[:, 1])$$
$$T_1[:, 1] = RANK(T_1[:, 5])^*$$
$$T_1[:, 1] = RANK(T_1[:, 6])^*$$
$$T_1[:, 1] = RANK(T_1[:, 10])^*$$
$$T_1[:, 8] = RANK(T_1[:, 7])$$
$$T_1[:, 8] = RANK(T_1[:, 3])^*$$
$$T_1[:, 8] = RANK(T_1[:, 4])^*$$
$$T_1[:, 7] = SUM_{row}(T_1[:, 3{:}6])$$
$$T_1[:, 10] = SUMIF(T_3[:, 1], T_1[:, 2], T_3[:, 2])$$
$$T_1[:, 11] = MAXIF(T_3[:, 1], T_1[:, 2], T_3[:, 2])$$
$$T_2[1, :] = SUM_{col}(T_1[:, 3{:}7])$$
$$T_2[2, :] = AVERAGE_{col}(T_1[:, 3{:}7])$$
$$T_2[3, :] = MAX_{col}(T_1[:, 3{:}7]),$$
$$T_2[4, :] = MIN_{col}(T_1[:, 3{:}7])$$
$$T_4[:, 2] = SUM_{col}(T_1[:, 3{:}6])$$
$$T_4[:, 4] = PREV(T_4[:, 4]) + T_4[:, 2] - T_4[:, 3]$$
$$T_5[:, 2] = LOOKUP(T_5[:, 3], T_1[:, 2], T_1[:, 1])^*$$
$$T_5[:, 3] = LOOKUP(T_5[:, 2], T_1[:, 1], T_1[:, 2])$$

collaboration with Patrick De Causmaecker

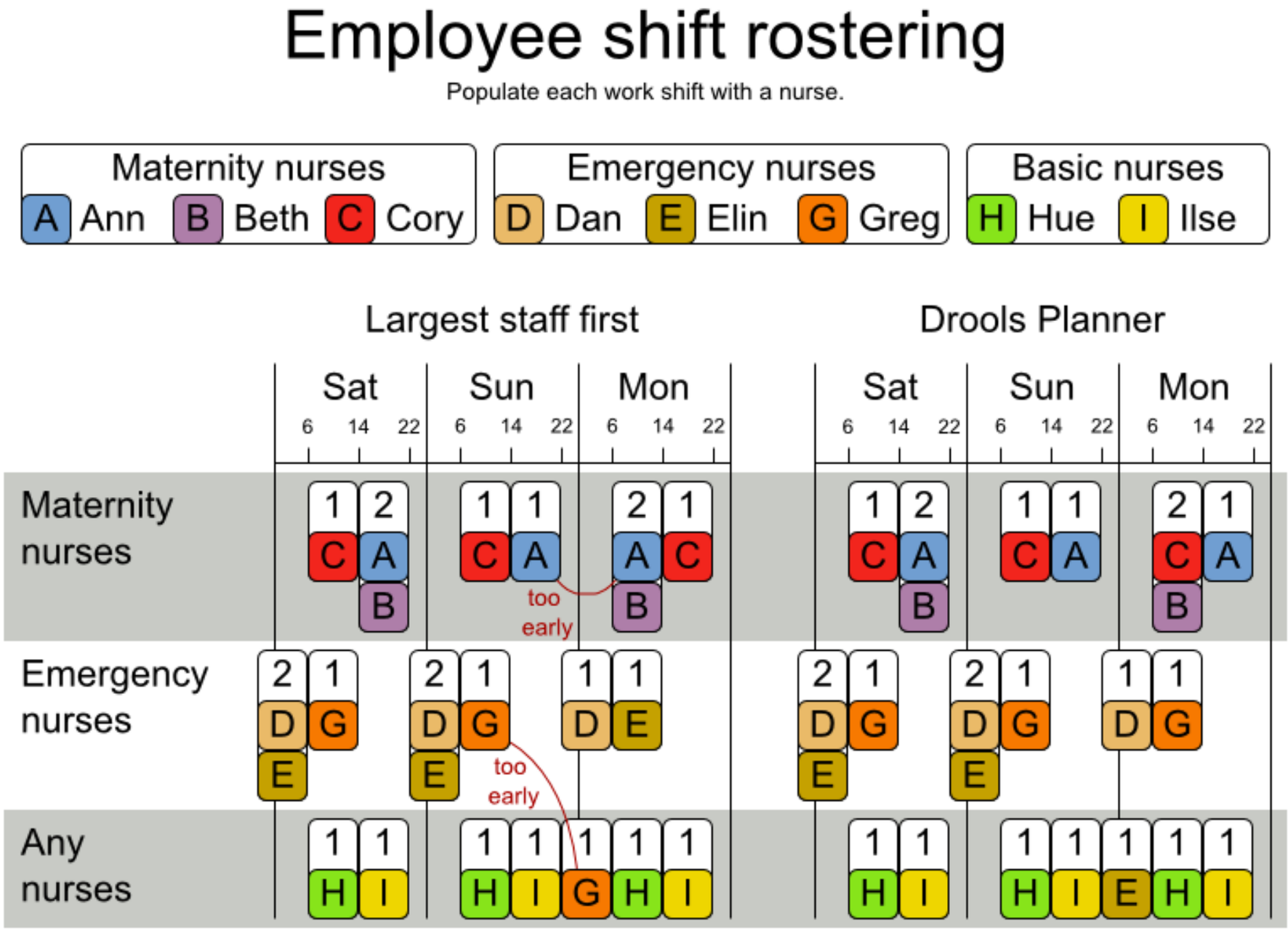


Figure G. De Smet

collaboration with Patrick De Causmaecker

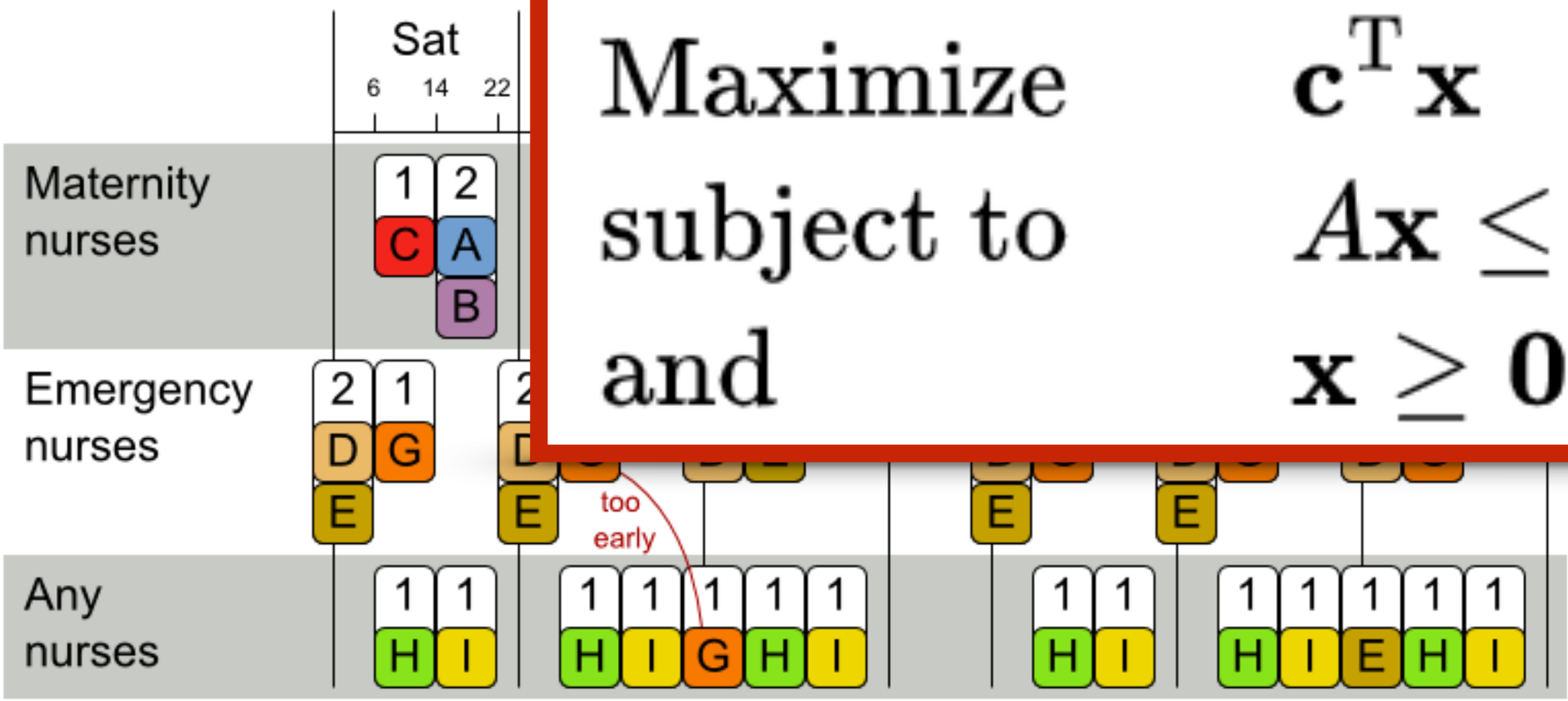
Employee shift rostering

Populate each work shift with a nurse.

Maternity nurses	Emergency nurses	Basic nurses
A Ann, B Beth, C Cory	D Dan, E Elin, G Greg	H Hue, I Ilse

Largest staff first

Drools Planner



$$\text{Maximize} \quad \mathbf{c}^{\mathrm{T}}\mathbf{x}$$

$$\text{subject to} \quad A\mathbf{x} \leq \mathbf{b}$$

$$\text{and} \quad \mathbf{x} \geq \mathbf{0}$$

Figure G. De Smet

Learning constraints and optimisation criteria

- Two aspects
- constraints
- optimisation function

$$\begin{aligned} &\text{Maximize} && \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ &\text{subject to} && A\mathbf{x} \leq \mathbf{b} \\ &\text{and} && \mathbf{x} \geq \mathbf{0} \end{aligned}$$

Two key approaches to constraint learning

1.general search procedures

- for TACLE — clever generate and test
- Other cases — general to specific search [Kumar et al IJCAI 19]

2. encode the problem as a constraint satisfaction problem and call a solver

- Incal — SMT(LRA) — [Kolb et al. IJCAI 18]

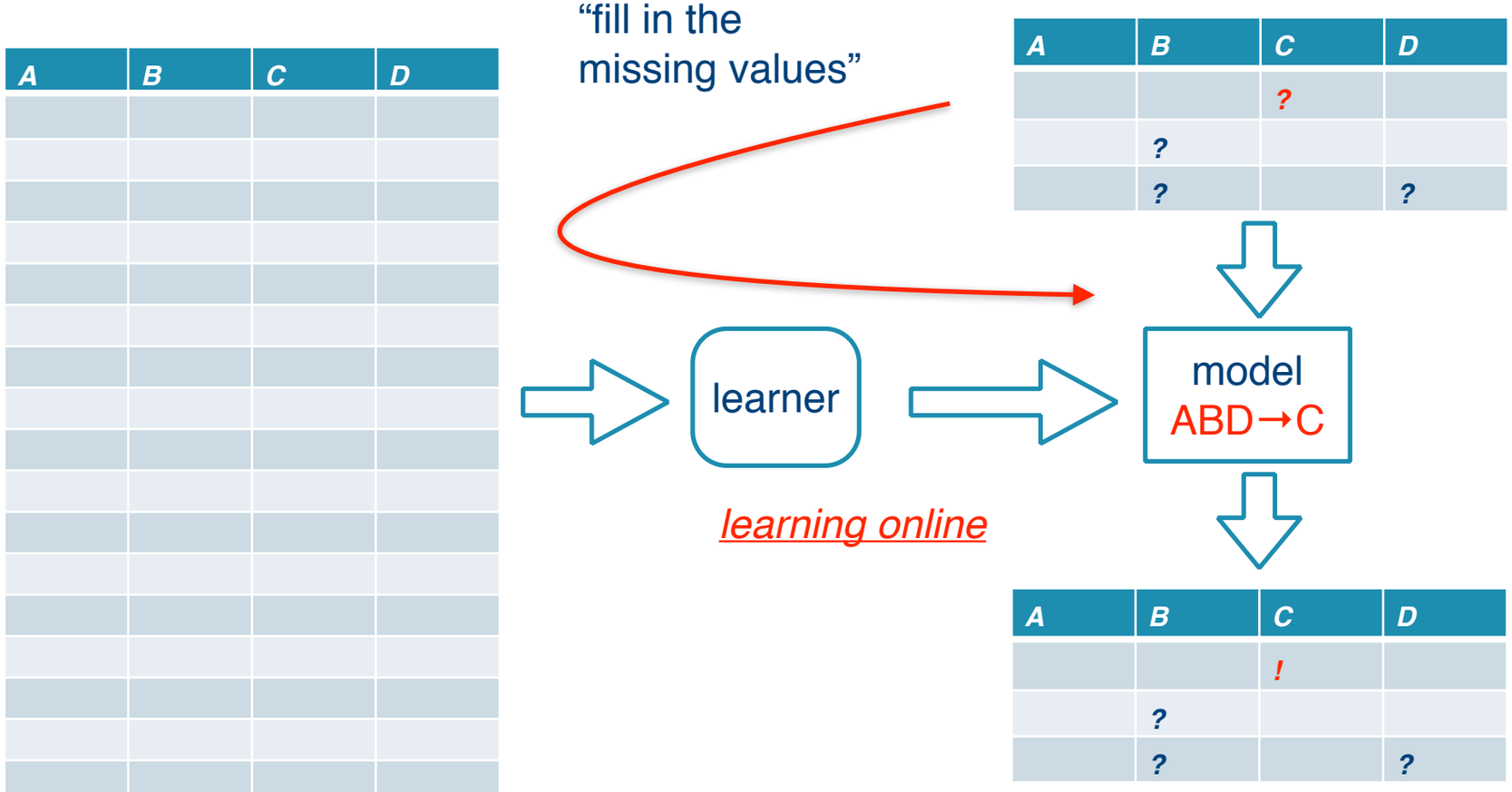


Predictive Modeling



Predictive Models

Predictive models in multiple directions …

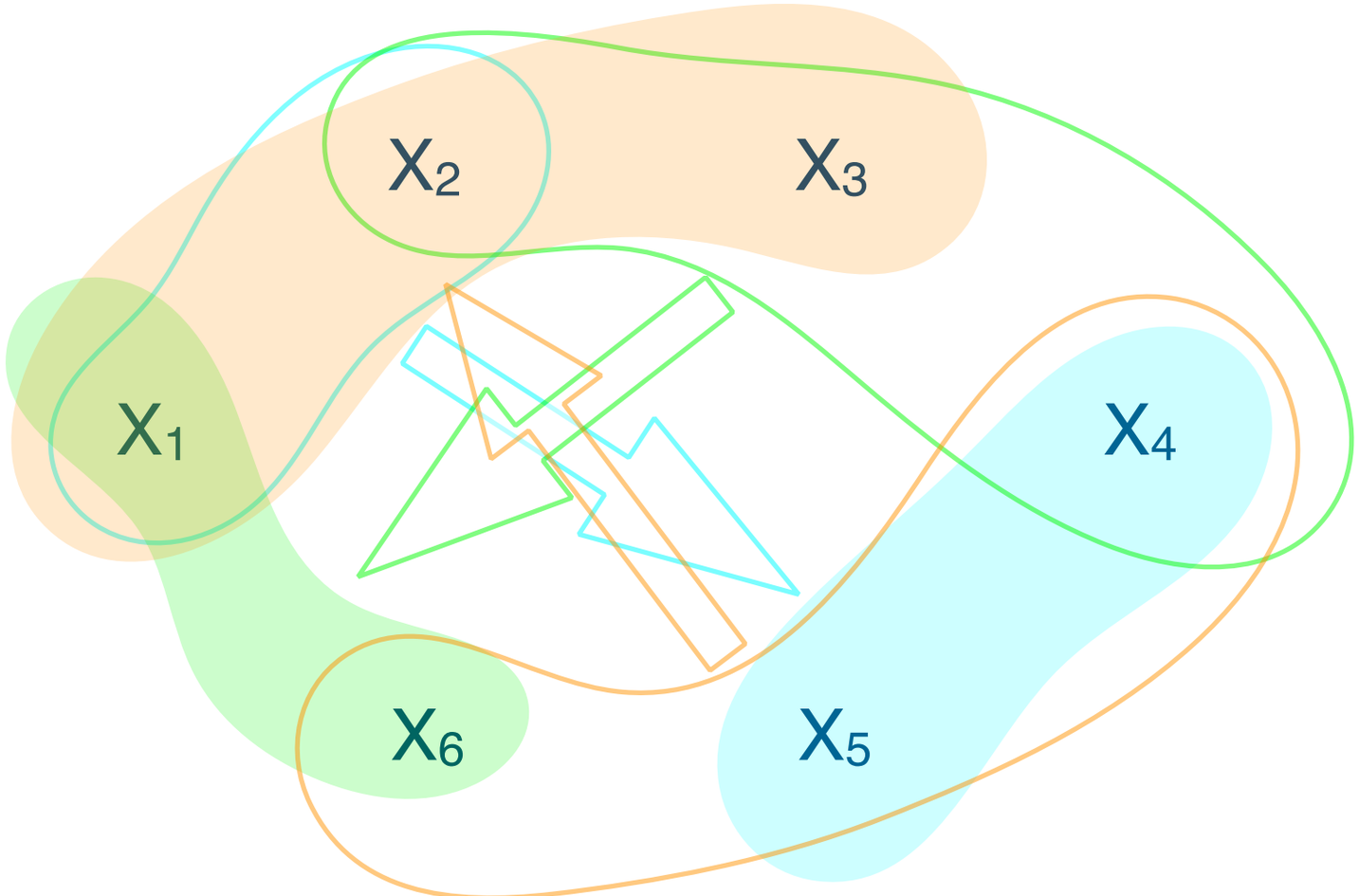


MERCS, see van Wolputte et al. IJCAI 18



Multi-directional ensembles of multi-target decision trees

- Basic idea: include in a single ensemble, trees predicting many variables from many other variables



Data Wrangling

Data Wrangling

t_1

Tim	Junior
Megan	Senior

t_2

Audi		
A1	Tim	no
A1	Megan	no
A4	Tim	yes
BMW		
1	Megan	?



Data Wrangling

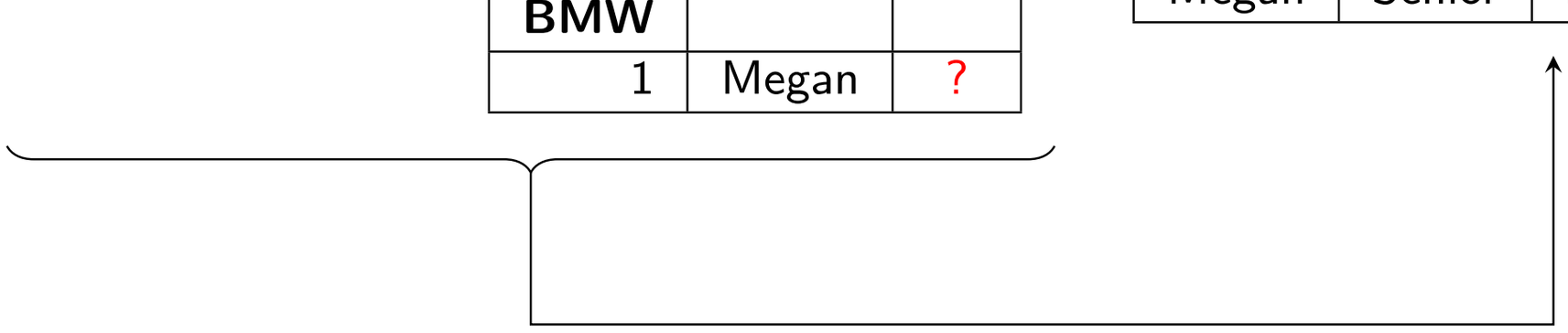
t_1

Tim	Junior
Megan	Senior

t_2

Audi		
A1	Tim	no
A1	Megan	no
A4	Tim	yes
BMW		
1	Megan	?

Tim	Junior	Audi	A1	no
Tim	Junior	Audi	A4	yes
Megan	Senior	Audi	A1	no
Megan	Senior	BMW	1	?



Data Wrangling

t_1

Tim	Junior
Megan	Senior

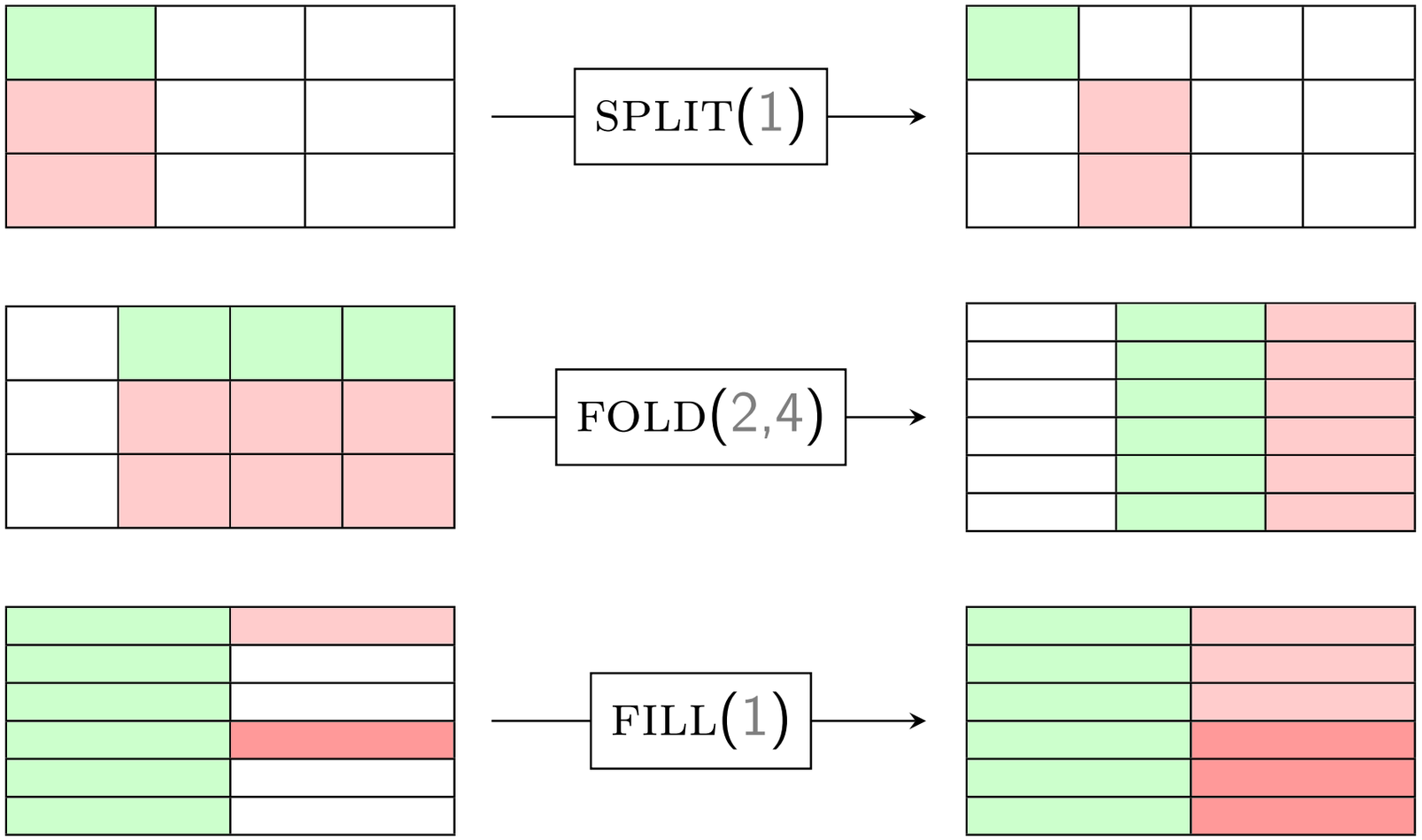
t_2

Audi		
A1	Tim	no
A1	Megan	no
A4	Tim	yes
BMW		
1	Megan	?

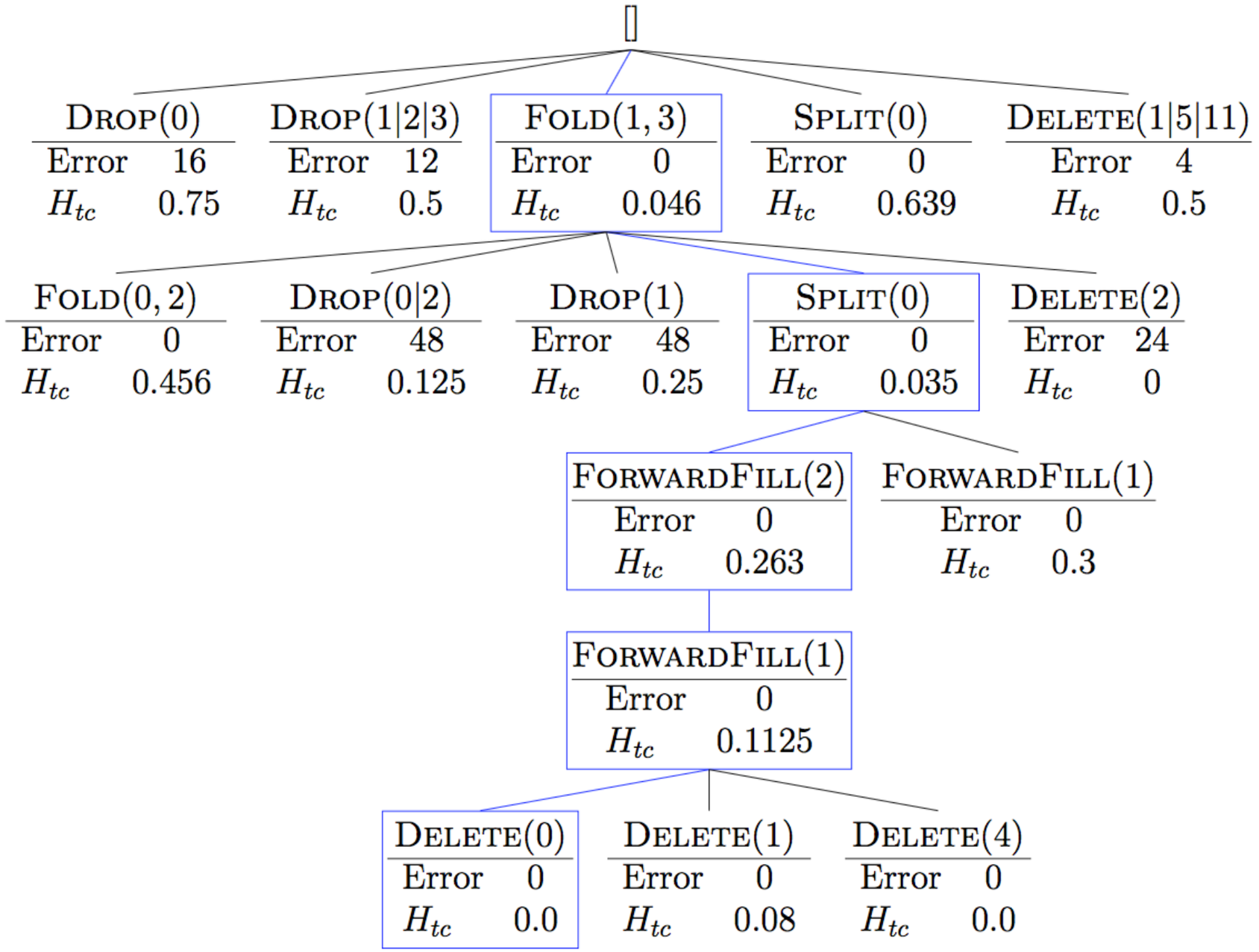
Tim	Junior	Audi	A1	no
Tim	Junior	Audi	A4	yes
Megan	Senior	Audi	A1	no
Megan	Senior	BMW	1	?

$\textsc{split}(t_2,\ 1)$
$\textsc{forwardfill}(t_2,\ 1)$
$\textsc{delete}(t_2,\ 2)$
$\textsc{join}(t_1,\ 1,\ t_2,\ 3)$

Transformations



SYNTH-A-SIZER



Gust Verbruggen & DR, IDA 18

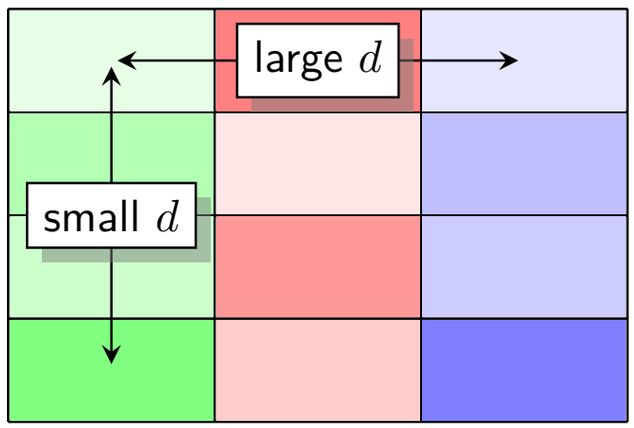
Data Wrangling

Scoring function computes how well the table fits our intuition about the desired format.

For **attribute-value**, we require that elements

1. within a column are similar
2. between columns are different

according to some distance measure d.



Semantic Segmentation

Given a spreadsheet, **find** a partitioning in rectangular groups of cells of the same semantic type,

Tim	Junior
Megan	Senior

Audi		
A1	Tim	no
A1	Megan	no
A4	Tim	yes
BMW		
1	Megan	

which—amongst other things—provides an easy way of scoring transformed tables.

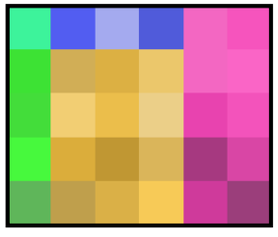
Tim	Junior	Audi	A1	no
Tim	Junior	Audi	A4	yes
Megan	Senior	Audi	A1	no
Megan	Senior	BMW	1	

We view it as a **constrained clustering** problem where clusters of cells are required to be rectangular.

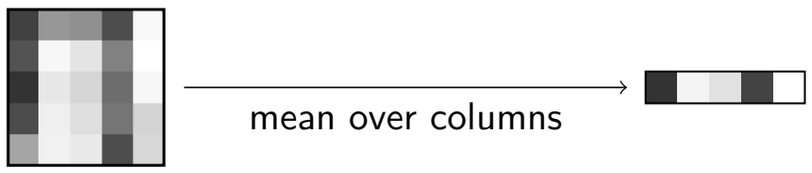
Verbruggen, Conteras-Ochando, Ferri, Hernandez-Orallo, DR, submitted



Semantic Segmentation



Compute **weight** of hypothetical edges



Pairwise distance between
horizontally adjacent cells

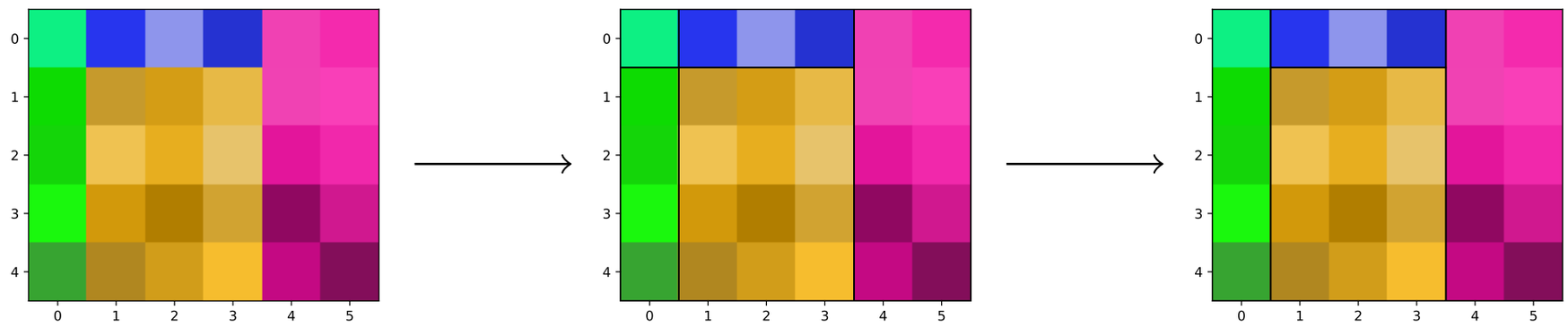
and recursively split edge with highest weight ..

Verbruggen, Conteras-Ochando, Ferri, Hernandez-Orallo, DR, submitted

Semantic Segmentation

Borrow ideas from **boundary extraction** for image segmentation.

1. Recursively split on segment with largest mean between two rows or columns to get an **oversegmentation**.
2. Join segments until optimality criterion achieved.



Verbruggen, Conteras-Ochando, Ferri, Hernandez-Orallo, DR, submitted

Semantic Segmentation

Given a spreadsheet, **find** a partitioning in rectangular groups of cells of the same semantic type,

Tim	Junior
Megan	Senior

Audi		
A1	Tim	no
A1	Megan	no
A4	Tim	yes
BMW		
1	Megan	

which—amongst other things—provides an easy way of scoring transformed tables.

Tim	Junior	Audi	A1	no
Tim	Junior	Audi	A4	yes
Megan	Senior	Audi	A1	no
Megan	Senior	BMW	1	

We view it as a **constrained clustering** problem where clusters of cells are required to be rectangular.

Verbruggen, Conteras-Ochando, Ferri, Hernandez-Orallo, DR, submitted



SYNTH

- Front End
  - a kind of intelligent Excel (we are integrating with Excel)
  - using colours to mark cells … + further annotations
- Back End
  - SYNTH-LOG - language
    - ako. language for learning, reasoning and data science

SynthLog

- Based on the inductive database idea [Iemielinski and Mannila CACM 95]

- Synthesising **inductive data models:**
  - **data models** (as in a data / knowledge base)
  - **inductive models** (learned models)
    - constraint, predictive models, probabilistic models …

- we need a language that supports inference and learning for these types of models



SynthLog

- Key idea —
  - a theory/scope is a **ProbLog** model
  - consists of
    - (probabilistic) facts
    - rules (definite clauses)
    - constraints
    - virtual views …

supports (ako)
answer set programming -
probabilistic programming
inductive learning
reasoning about all this
interface with Excel

- Closure property : the result of any operation is a theory



Conclusions

- Automated data science is an interesting challenge and I believe we can automate it to a large extent

- A small non-trivial “autocompletion” setting that is appealing

- Some contributions on constraint learning, versatile models and data wrangling

- A language for data science that is under development

- Still many open questions — much more work needed

Synth Team

- **Luc De Raedt**
- ***Hendrik Blockeel***
- **Jesse Davis**
- **Jan Ramon**
- **Yann Dauxais**
- **Patrick Decausmaekcer**
- **Anton Dries**
- **Clement Gautrais**
- **Wannes Meert**
- ***Stefano Teso***
- ***collab Andrea Passerini***
- ***Evgenia Korneva***
- **Robin Manhaeve**
- ***Elia Wolfsputte***
- ***Sergey Paramonov***
- ***Samuel Kolb***
- **Gust Verbruggen**
- **Nitesh Kumar**
- ***Mohit Kumar***
- **Arcchit Jain**

