Understanding Data Quality in linked administrative data

Florian Endel[1]

[1]Vienna University of Technology, florian@endel.at

IHDL Conference Vancouver,
2014-04-30

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

Outline

1. Introduction & Background
   - Data Linkage in Austria
   - Motivation
2. Approaching Data Quality
   - Data Quality: Dimensions & Metrics
   - Supporting Software
3. Solutions in GAP-DRG
   - Profiling
   - Database
   - DQA Reports
   - Interactive QA

Outline

Data Linkage in Austria

- Health insurance in **Austria**
  - public & mandatory
  - different providers, fragmentation
  - social security number: Unique Patient Identifier
- Covered years:
  - 2006 – 2007 whole country
  - 2008 – 2011/12 one province ($\sim \frac{1}{8}$)
- Content: claims data
  - inpatient, outpatient (GP, specialist), pharmacies
  - socio-economic factors, demography
  - "imputed" information: ATC $\leftrightarrow$ ICD(10)
  - master data, meta data
- $\sim$ 8 million people, 2.5 billion datasets

Data Quality: Open questions

- Many open and eligible questions:
  - What is the quality of the data?
  - Variation of quality? Depending variables?
  - Differences between data quality and system effects?
  - How can quality be improved?
- Analysis and interpretation of results
  - "intuitive" knowledge about quality
  - **appropriate & usable documentation**
- What are dimensions of data quality?
- How can they be measured and communicated?

Outline

Dimensions of data quality

- Various possibilities & publications
- e.g.
  - accuracy, completeness, consistency, timeliness
- No general agreement on dimensions, meaning, metrics
  - e.g.: missing vs. not existing vs. unknown
- Is it worth the effort? What does it cost?

Supporting Software

- No structured assessment!
- Many solutions: dataqualitypro.com/software-directory
- Various properties to be accounted
  - licensing, platform, customizability, documentation, support, ...
- Restrictions (in our project)
  - price, platform (Linux), no internet connection
- Common problems
  - resource consumption & amount of data
  - flexibility
  - (graphics & statistics)

Outline

Solutions in GAP-DRG: Overview

- Raw data: before the import/integration
- Monitoring data integration
- Ensuring minimal quality in the database
  - constraints
  - referential integrity
- Reporting on quality aspects: profiling
- Interactive exploration
- Data Quality Assessment reports
- From quality assessment to data analysis

Profiling

- Automatic summary of data
- Per variable / data type
- Only simple interactions, partitions and aggregations
- Large amount of information
- Easy & fast to create and read
- Repeat when
  - new data arrives
  - data is transformed / cleaned
- SQL & R & LaTeX

Profiling: example 1

abrvtr

n	missing	unique	Mean
22193	277345	1	12

kat

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
22193	277345	926	27436	1070	1300	5080	11020	32020	99999	99999

```
lowest :     0    60    70   999  1001
highest: 85030 85033 99031 99032 99999
```

posbez

n	missing	unique
299161	377	154469

```
lowest : -                                  0.25 x 6 mm / 31 G 100 Stk.          0.30 x 8 mm / 30 G 100 Stk.
highest: ZYVOXID ILSG 2MG/ML     10ST       ZYVOXID ILSG 2MG/ML B   10ST (J01XX) ZYVOXID ILSG 2MG/ML BTL.10ST
```

fgr

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
22193	277345	37	23.72	1	3	6	11	32	84	85

```
lowest :  1  3  4  5  6, highest: 86 87 91 92 99
```

abt

n	missing	unique
1757	297781	2

```
BAD (1399, 80%), ZAH (358, 20%)
```

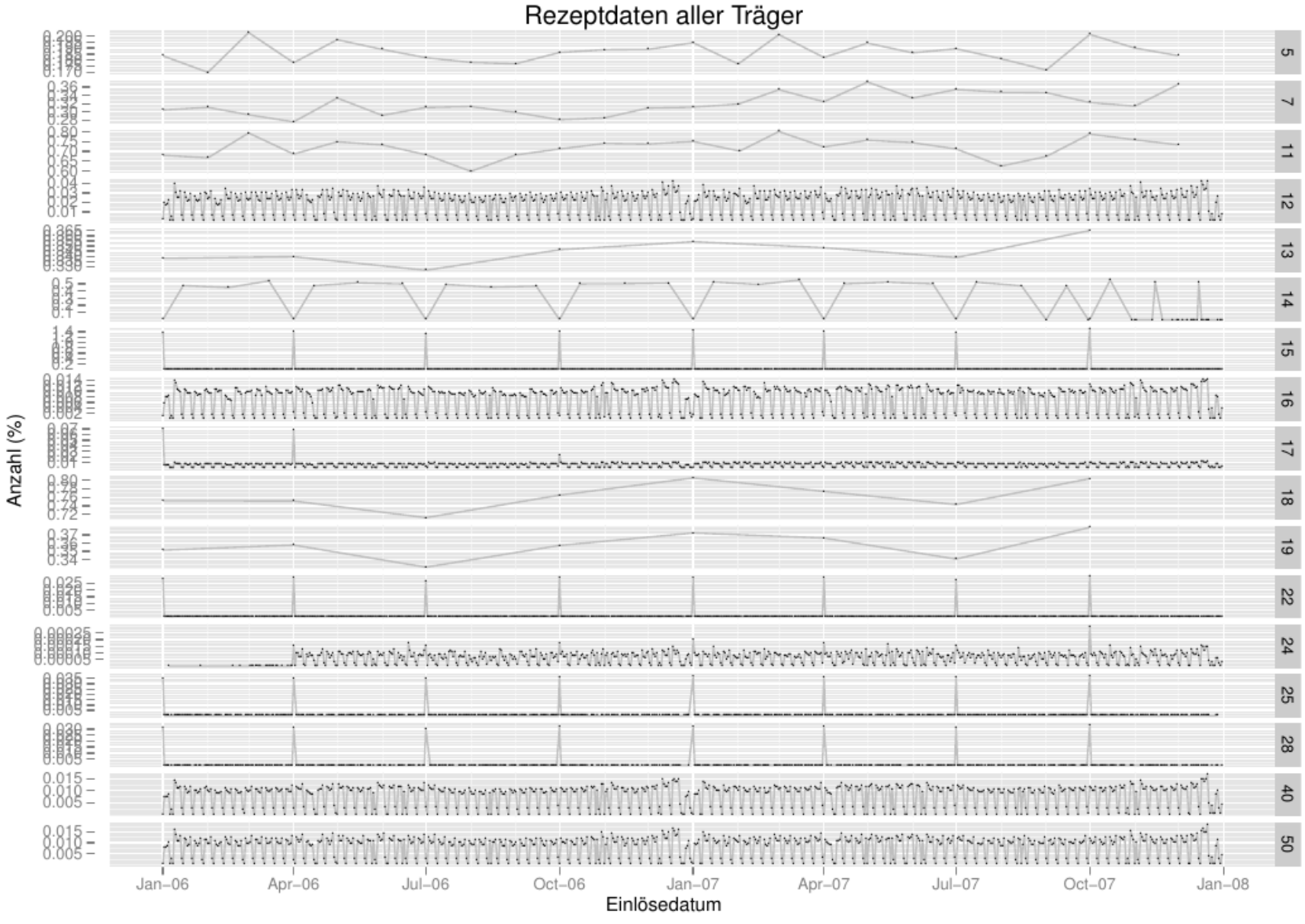
Variables with all observations missing: `zeitraum`

Profiling: example 2

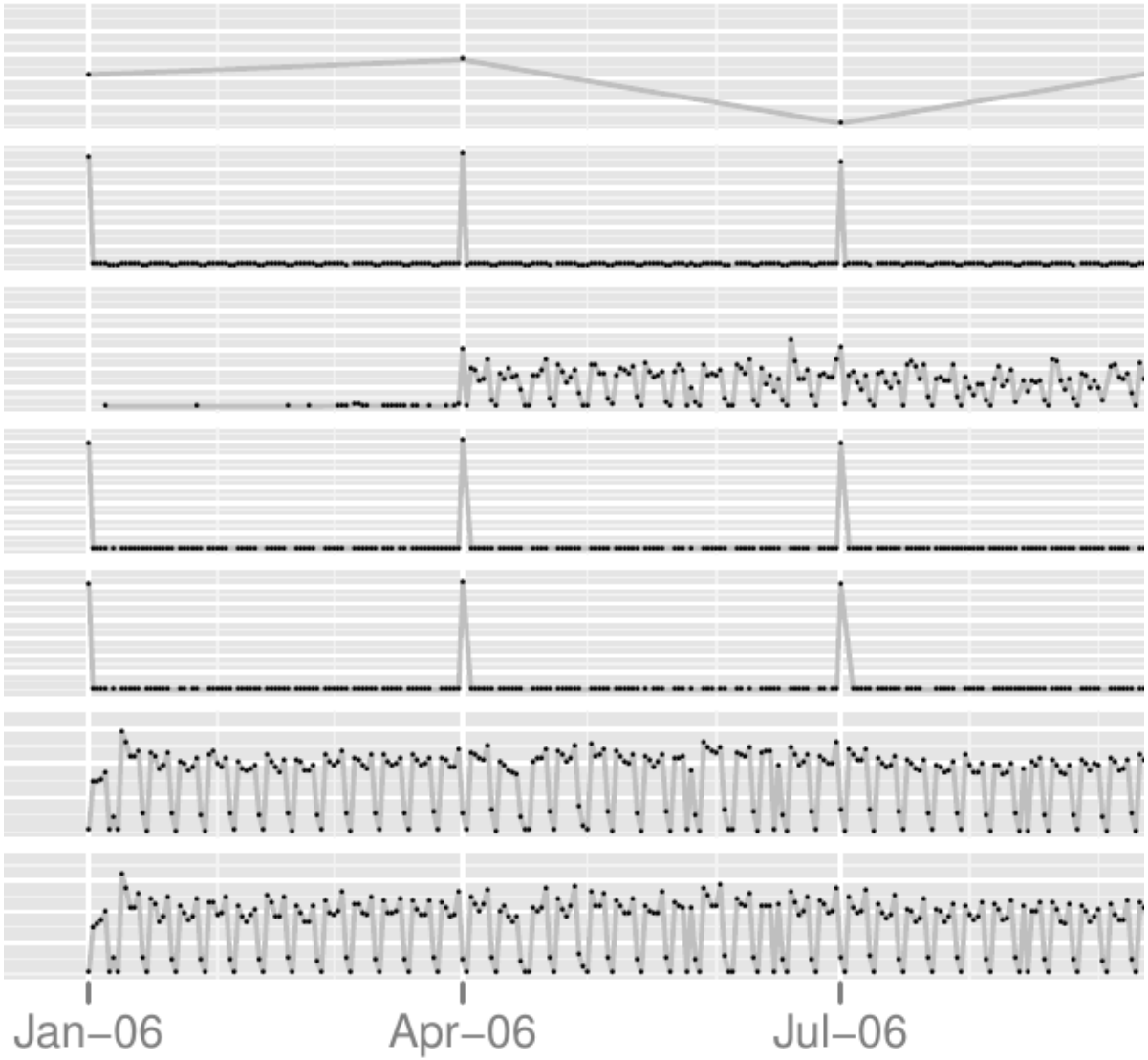
5.8.4 gebjahr

	herkunft	count	NULL	(%)
1	5	832.074	204.039	25 %
2	7	2.291.750	2.215.411	97 %
3	11	10.463.982	170.042	2 %
4	12	7.594.892	7.594.892	100 %
5	13	760.921	164.407	22 %
6	14	4.020.708	10.542	0 %
7	15	5.845.224	23.786	0 %
8	16	1.487.245	211.079	14 %
9	17	2.858.063	150.738	5 %
10	18	2.372.839	2.372.839	100 %
11	19	796.453	116.656	15 %
12	22	39.953	10.376	26 %
13	24	6.524	2.290	35 %
14	25	29.224	7.431	25 %
15	28	24.202	6.151	25 %
16	40	6.897.964	59.950	1 %
17	50	1.029.344	7.822	1 %

Profiling: prescriptions / insurer



Profiling: prescriptions / insurer (detail)



Database Schema

- Common problems:
  - Are diagnosis (ICD10) valid? What is their meaning?
  - One patient, two places of residence at the same time: multiplication of associated records?
  - Varying date of birth
- $\Rightarrow$ **Normalisation & Referential Integrity**

Database Schema

- Common problems:
  - Date of admission before separation?
  - Several hospital stays at the same time
  - Death before birth
  - Encoding of sex
- $\Rightarrow$ **Data Types & Constraints**

DQA Reports

- Special “studies” conducted by experts
- Highly sophisticated
- Analyzing data with emphasis on quality
- Exploratory & knowledge driven (blind spots?)
- e.g. comparison of linked data with
  - demographic reports
  - health surveys
  - reported costs

Interactive Quality Assessment

- Quality profiles are static
- Researchers are interested in
  - details
  - partitions of data
  - views without outliers
  - different levels of aggregation
  - fast (immediate) results
- 2 solutions:
  - interactive profiling
  - analysis of prescriptions
- Prototype: classification with perceptrons & ada boost
- R & shiny

Interactive profiling: user interface

GAP-DRG2 Dates

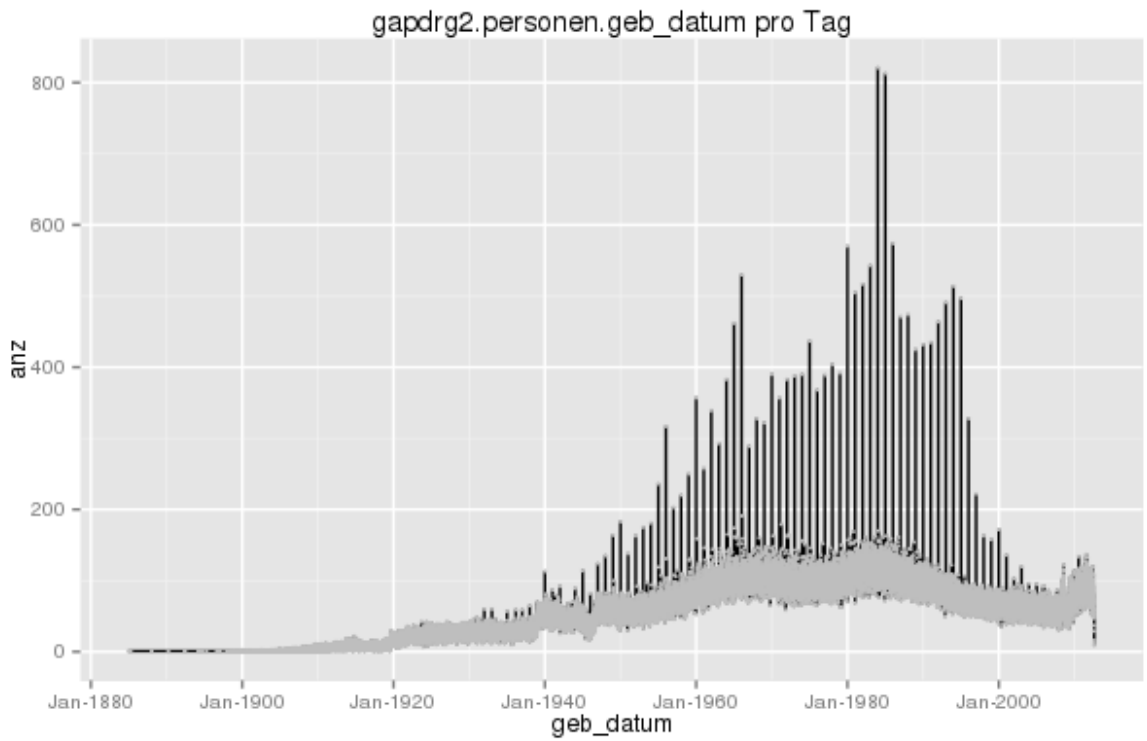
Tabelle:
personen

Spalte:
geb_datum

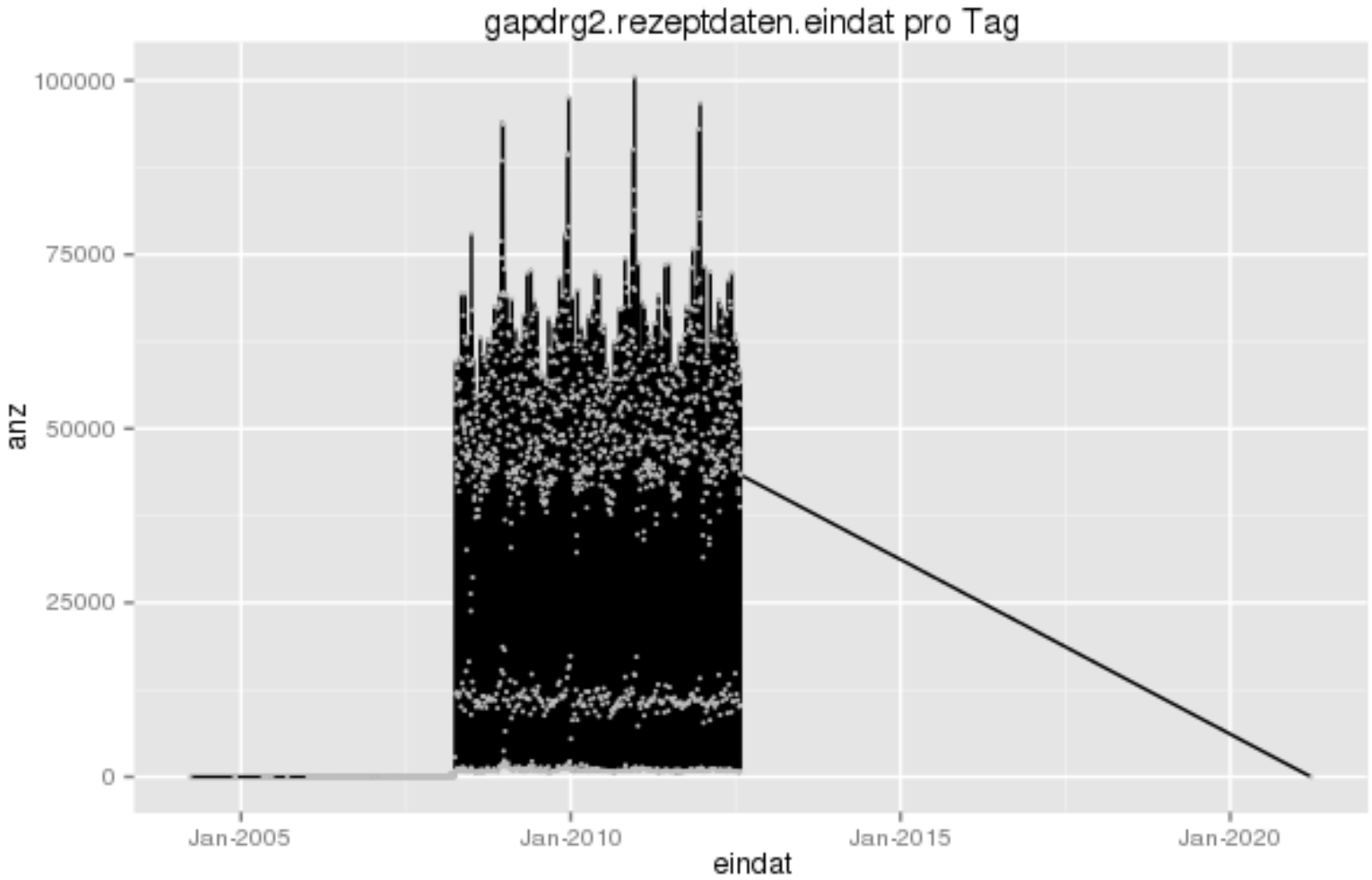
Split:
kein

Range:
1885 2012

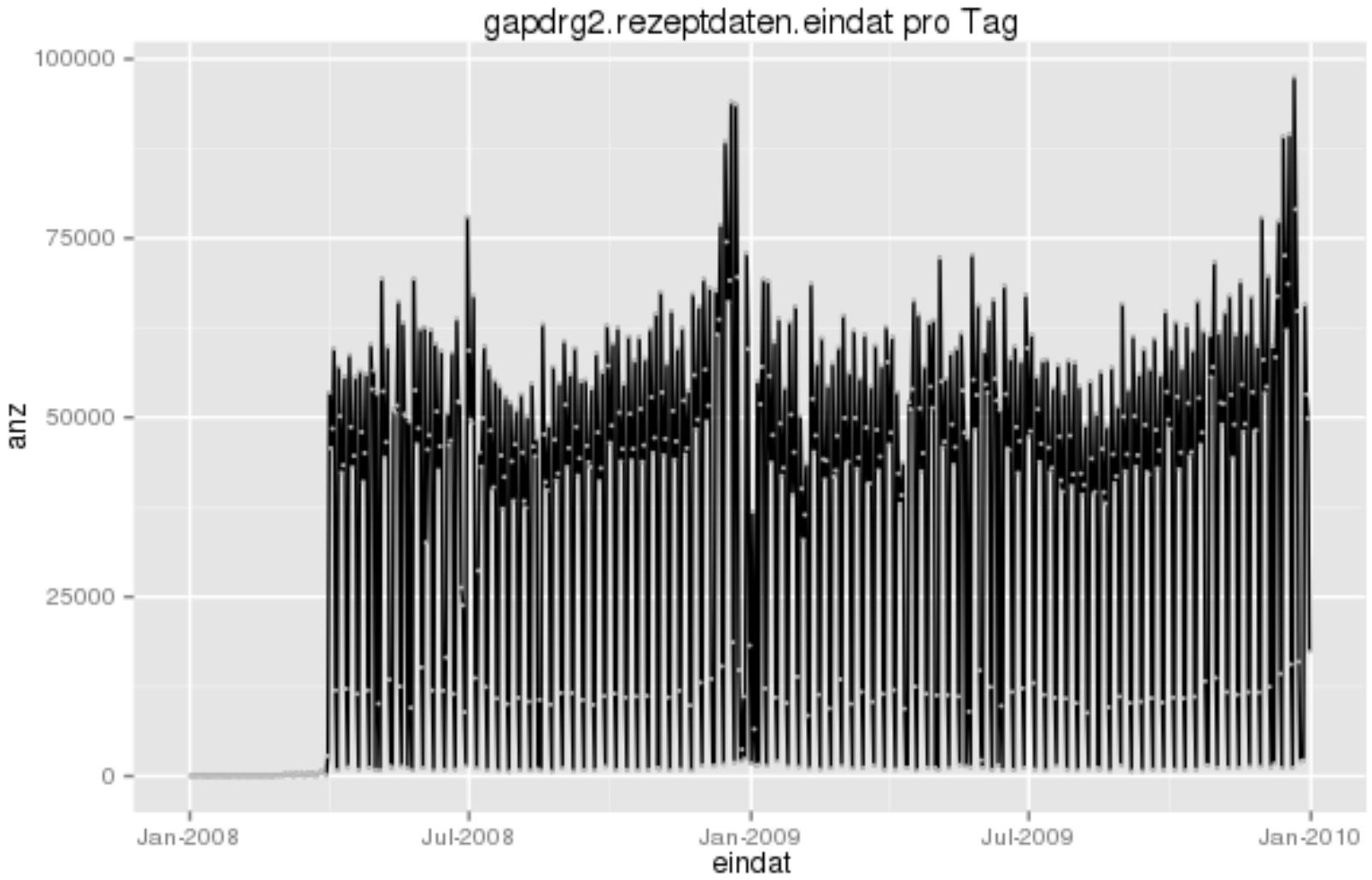
personen: geb_datum



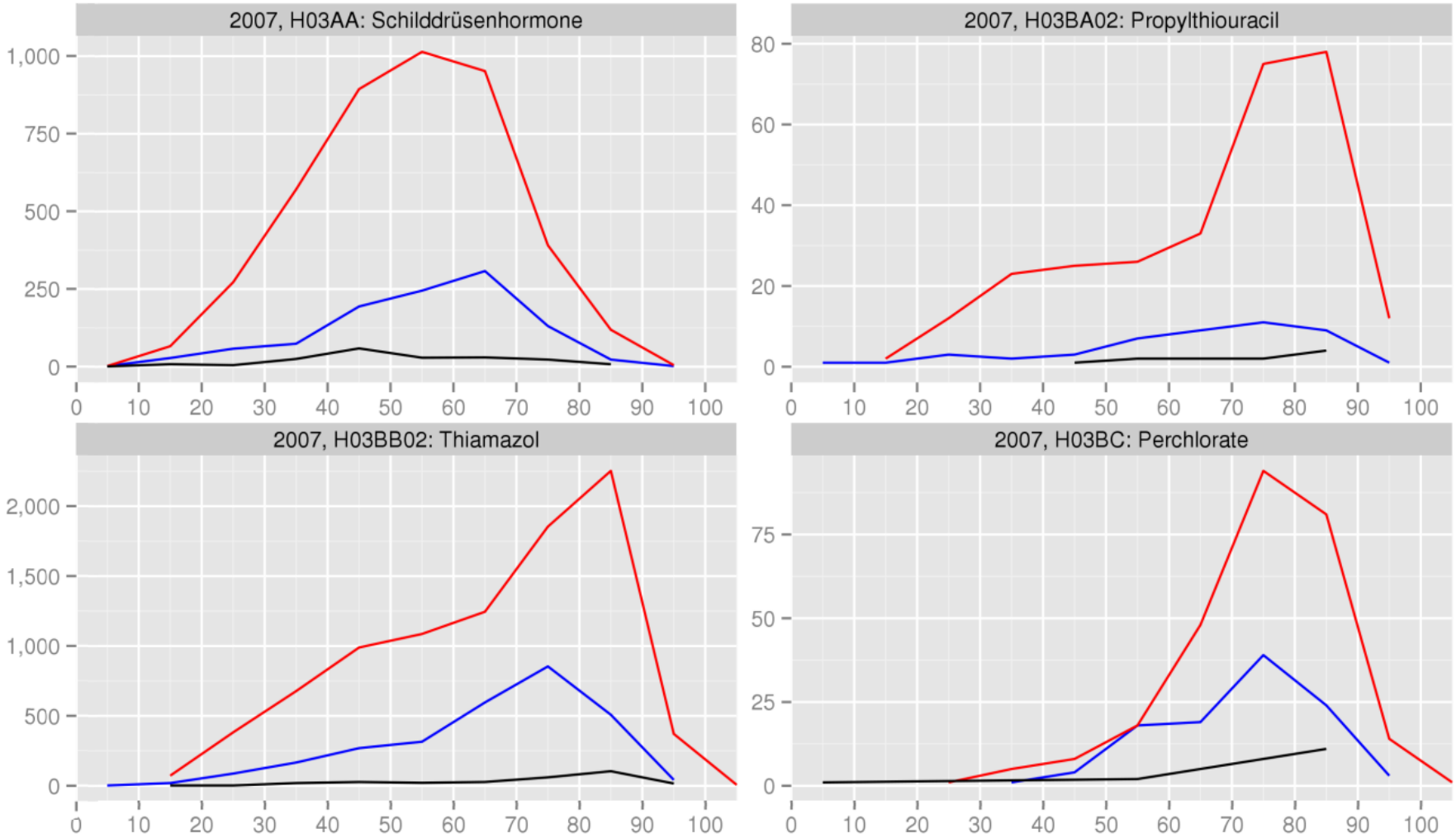
Interactive profiling: example



Interactive profiling: example



Analysis of prescriptions: example



Classification (prototype!)

Dataset & Variables

Ada Boost Options

Perceptron Options

Decay:
FALSE

hidden Layers:
a

Learning Rate:
0.01 0.3 1

Momentum:
0 0.2 1

convert Nominal to Binary:
TRUE

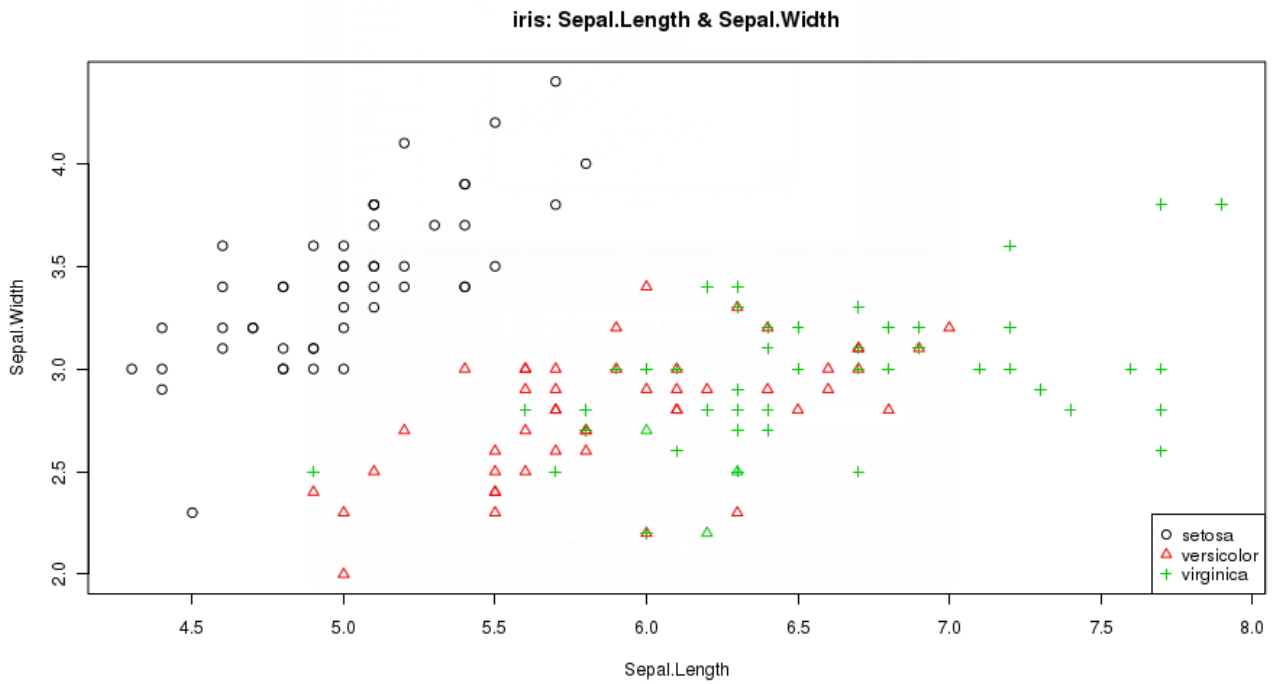
Normalize:
TRUE

Normalize Class:
TRUE

Reset:
TRUE

iris

Data: raw | Data: interactive | Data: summary | Ada: summary | Classification: Result



Understanding Data Quality in linked administrative data

Florian Endel[1]

[1]Vienna University of Technology, florian@endel.at

IHDL Conference Vancouver,
2014-04-30

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology