UNIVERSITY OF TARTU
Institute of Computer Science

Data extraction from PDF invoices

Kristjan Veskimäe, kristjan.veskimae@gmail.com
Institute of Computer Science, University of Tartu

IT Akadeemia
toetab Skype™

https://github.com/kveskimae/eazyfill

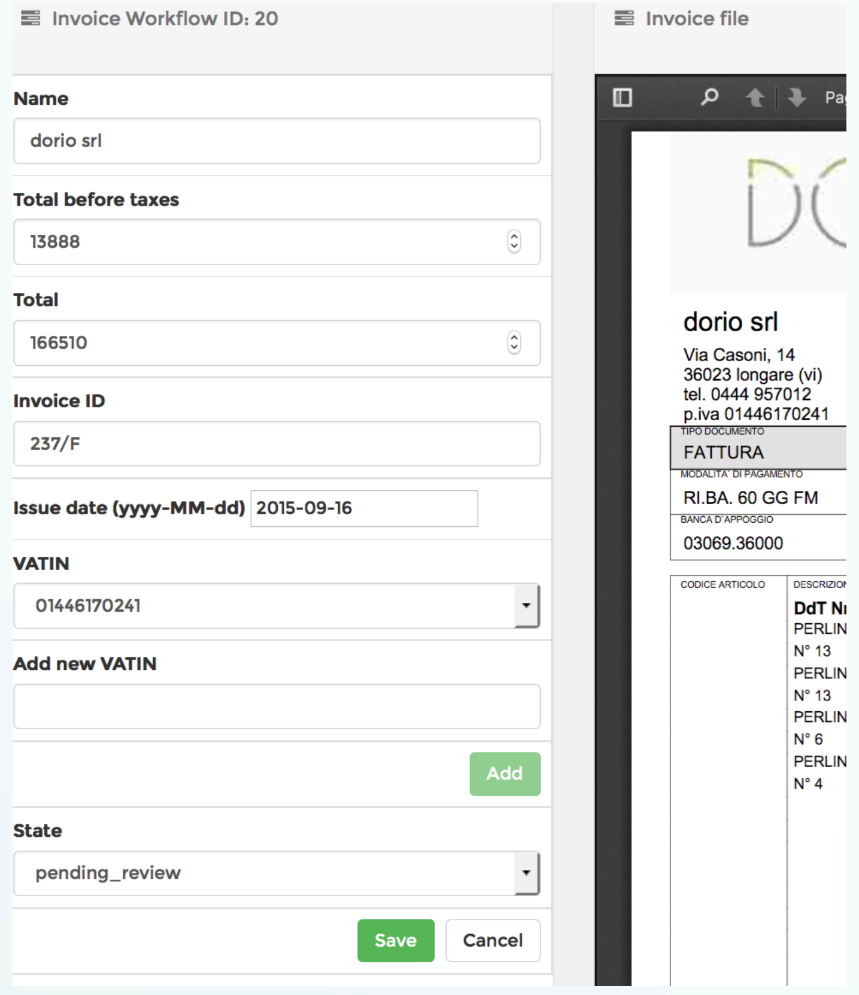
http://www.eazyfill.com/?lang=en

Introduction Accountants need to manually copy-paste payment data from invoices into accountancy software. While PDF invoice files have no clear format, it is still quite possible to extract this data automatically using data like position of different phrases on invoice.

Objectives Goal is to automate data insertion from PDF invoices into accountancy software.

Solution Our first task was to train our data extractor in finding all the possible candidates for a payment field (supplier company name, total to be paid, taxes etc.). Sample invoices were collected and data points for payment field values were extracted – including text font properties, position, proximity to their corresponding key phrases, regex format they confirm to. Then algorithm selects the most likely value from different candidates.

Custom implicit k-d tree As an example of techniques that are used in selecting the correct value, below is demonstrated using positional data to train an implicit k-d tree. Each tree node stores the data point locations in its associated cell. Iterative updating propagates a new data point top-down and can split a cell. As can be seen from below, most of the data points got cluttered in the upper right cell for reference number in Estonian invoices. It is therefore the most likely place to find the value for reference number.


Invoice Workflow ID: 20

Name: dorio srl

Total before taxes: 13888

Total: 166510

Invoice ID: 237/F

Issue date (yyyy-MM-dd): 2015-09-16

VATIN: 01446170241

Add new VATIN:

State: pending_review

Invoice file

dorio srl
Via Casoni, 14
36023 longare (vi)
tel. 0444 957012
p.iva 01446170241

TIPO DOCUMENTO: FATTURA
MODALITA' DI PAGAMENTO: RI.BA. 60 GG FM
BANCA D'APPOGGIO: 03069.36000

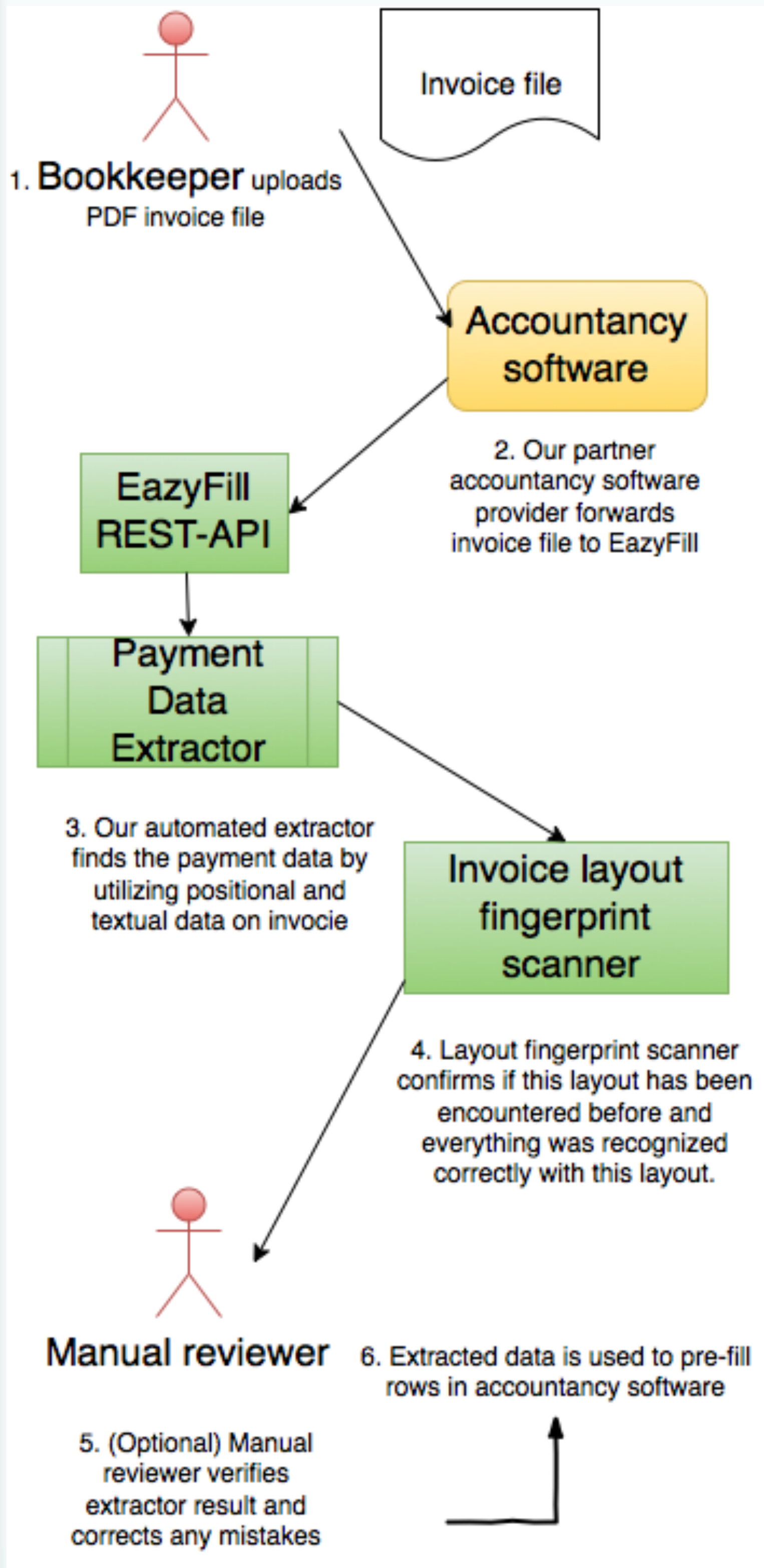
CODICE ARTICOLO | DESCRIZIONE
DdT N
PERLIN
N° 13
PERLIN
N° 13
PERLIN
N° 6
PERLIN
N° 4

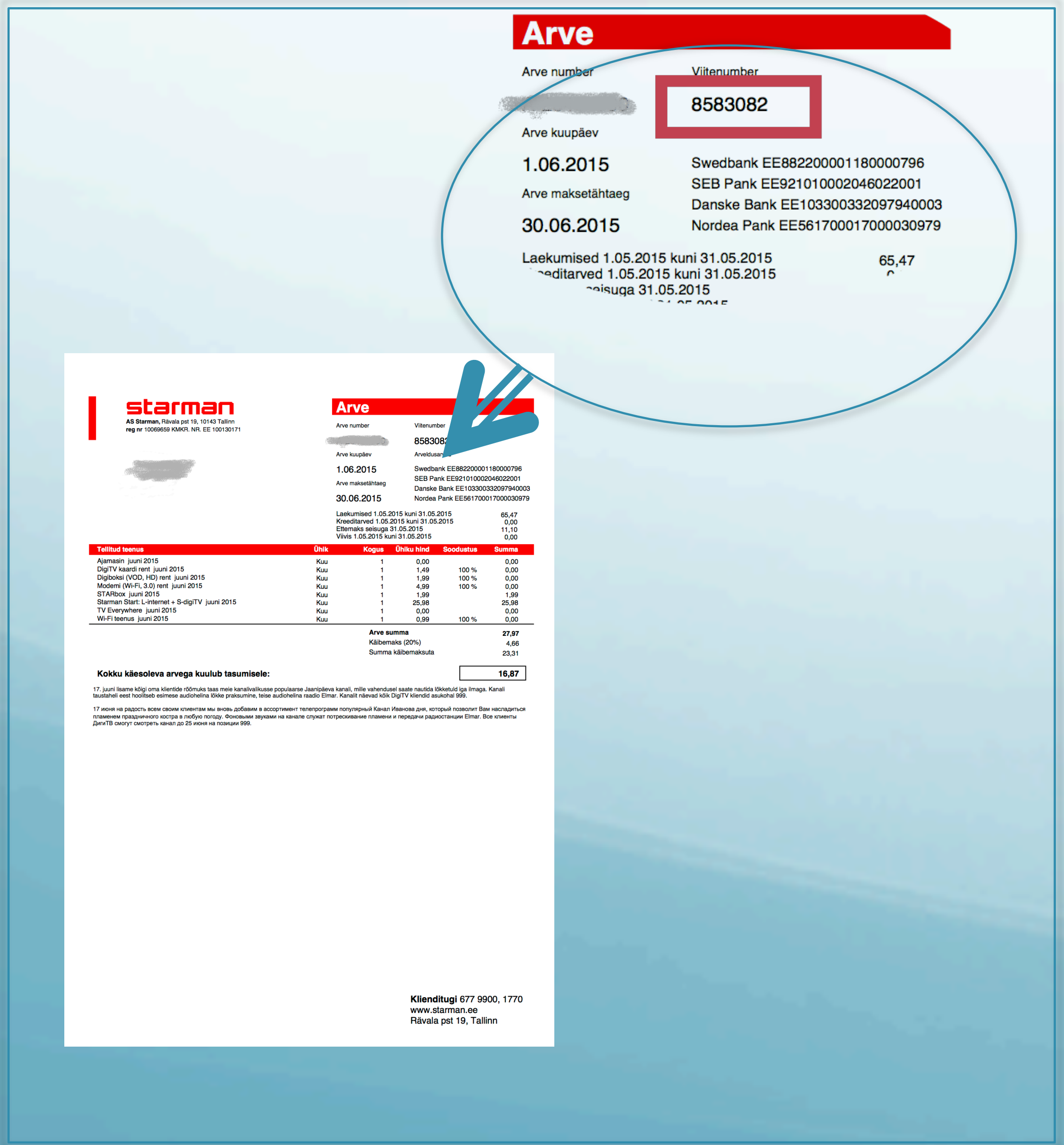
Results Extracting some fields, e.g. invoice ID and total, can be automated to a high degree and we are already achieving good accuracy levels. On the business side, we are validating the idea, providing extraction service to one hundred Italian companies. All are clients of accountancy software provider Adamo. We are using semi-automated process with manual review step for unrecognized invoices (see partial screenshot on left).

Future Work Some fields still need more work. For example, finding name is oftentimes a challenging endeavor. Name can be in a sentence, while some companies also do not include company type abbreviation (AS/OÜ for Estonia, SRL/SOC/etc. for Italy). A good solution is still to be worked out, possibly querying a name database by value added tax ID, instead of defining precise position to look for name.

Unrecognized invoices are still vast majority for us, but we see our competitive advantage against possible future competitors the ability to fingerprint and recognize layouts and extract data instantly without review, which would also greatly lower the operational costs (see right).

Extraction service is currently a desired solution also for other accountancy software providers and there are three more partners waiting for integrations, which are planned in the upcoming months. This is together with Adamo wanting to expand to all two thousand of its clients.


Invoice file
1. Bookkeeper uploads PDF invoice file
Accountancy software
2. Our partner accountancy software provider forwards invoice file to EazyFill
EazyFill REST-API
Payment Data Extractor
3. Our automated extractor finds the payment data by utilizing positional and textual data on invocie
Invoice layout fingerprint scanner
4. Layout fingerprint scanner confirms if this layout has been encountered before and everything was recognized correctly with this layout.
Manual reviewer
5. (Optional) Manual reviewer verifies extractor result and corrects any mistakes
6. Extracted data is used to pre-fill rows in accountancy software

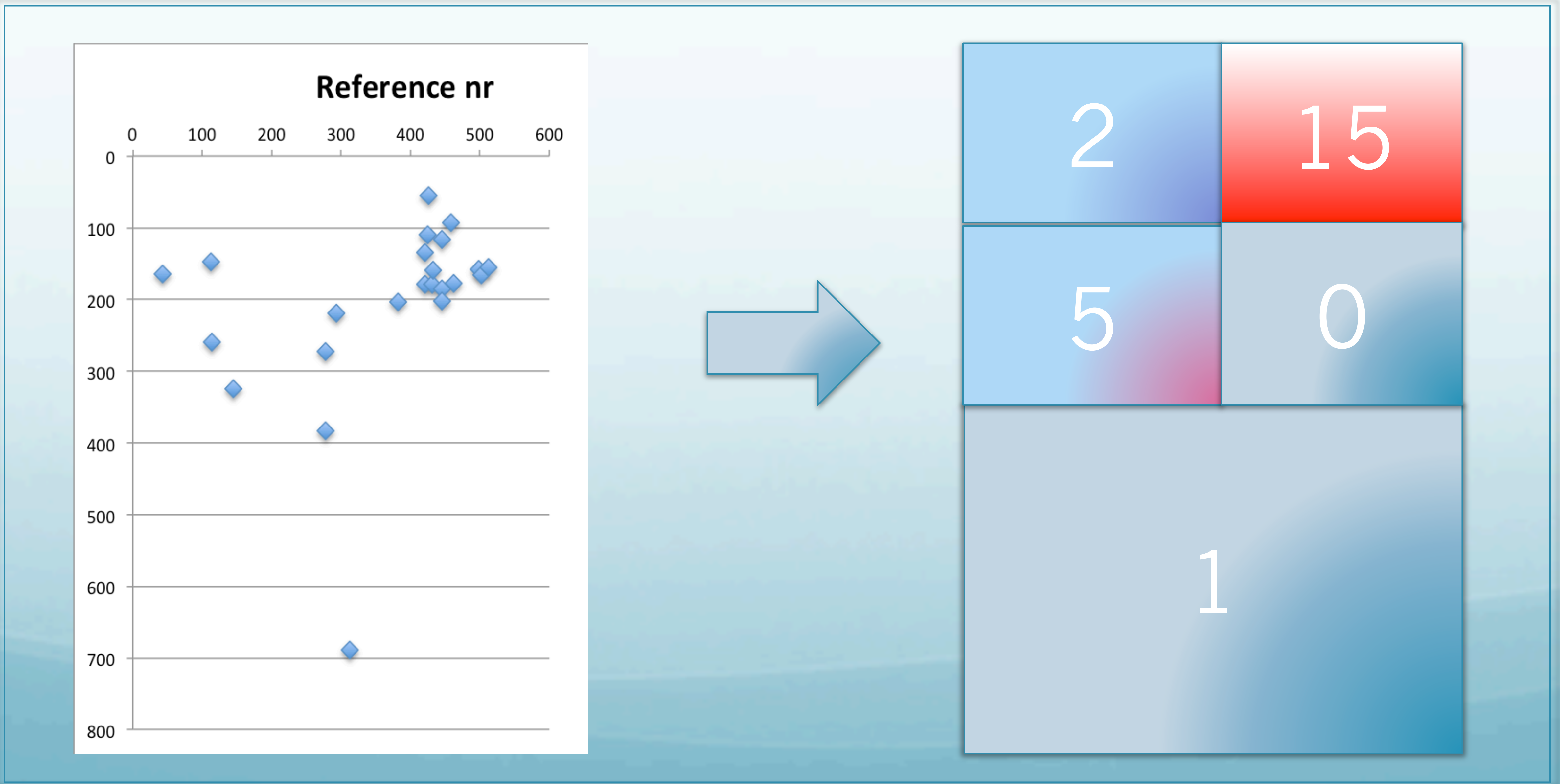

Arve

Arve number: [redacted]
Viitenumber: 8583082

Arve kuupäev: 1.06.2015
Arve maksetähtaeg: 30.06.2015

Swedbank EE882200001180000796
SEB Pank EE921010002046022001
Danske Bank EE103300332097940003
Nordea Pank EE561700017000030979

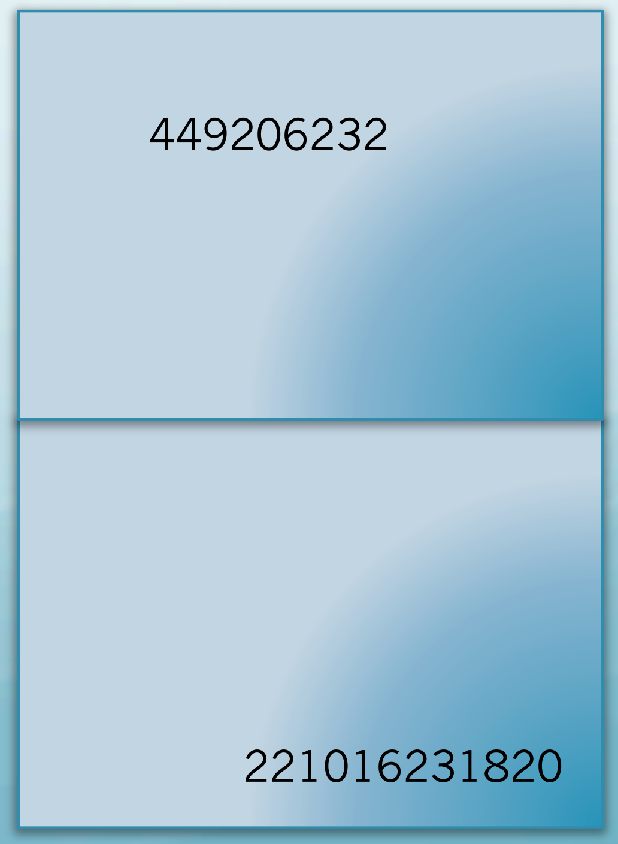
Laekumised 1.05.2015 kuni 31.05.2015 65,47

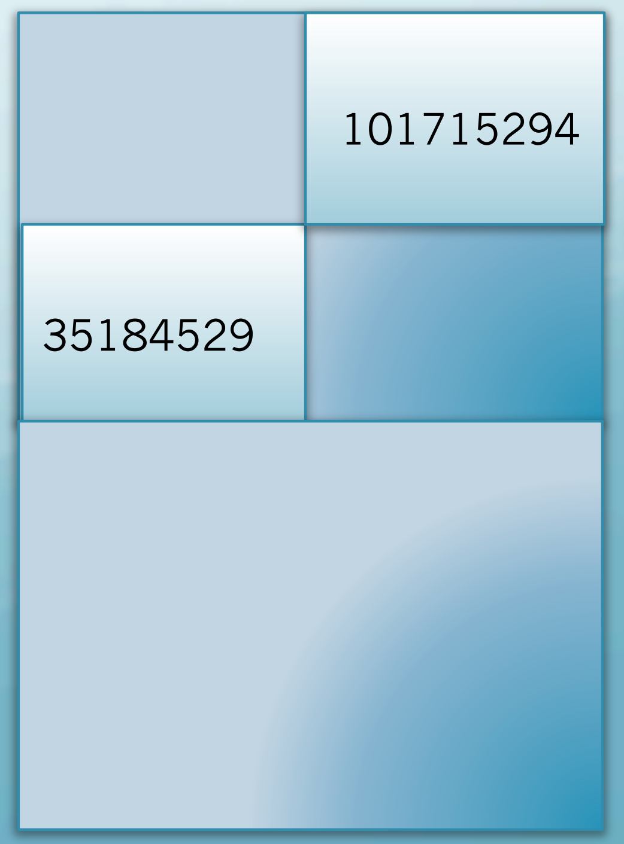


Comparing

Algorithm estimates 449206232 to be 22 times more likely to be the correct value

Algorithm estimates 101715294 to be 3 times more likely to be the correct value


449206232
221016231820


101715294
35184529