

# Working as a data scientist with real world clinical data

13 October 2015

# Filmpje

<https://www.youtube.com/watch?v=UoYl7eCesqw&feature=youtu.be&a>

# Clinical data scientist

- Data preparation
- Collaboration
- Advanced Programming
- Statistics
- Scientific rigour
- Visualisations
- Hacker mindset
- Understanding clinical data

# Data preparation

- Data from different sources
  - GP database
  - Pharmacy claims database
  - Hospital admissions
  - Hospital pharmacies
  - Laboratories
  - Additional data sources



# Data preparation

- Extracted
- Standardized
- Linked
- Coded
- Explored
- Creation of analysis datasets

# Collaboration

- Different departments
  - Information Management
  - Research
  - Reporting
- Engage with senior management
- Explore customer needs
- Influence without authority

# Advanced programming

- SAS Base
- SQL queries
- SAS Macro
- SAS Graph
  - GTL
- SAS VA
- Other programming languages



# Good Programming Practices

- Efficient
  - What is the most straightforward method (less code)
  - What is the fastest method (less time)
- Repeatable
  - Macro language
  - No re-programming of same code
- Clear and transferable
  - Debugging must be easy
  - Better to understand by colleagues

# Example:

## SAS SQL combined with MACRO

```
PROC SQL;  
    CREATE TABLE __allPat AS  
        SELECT &repmonc COUNT(Distinct &pat) AS TotPat  
        FROM &DsIn  
        WHERE &geslacht IN ('M','V') AND age ne . AND age<98  
        %IF &bymon=Y %THEN %STR(GROUP BY repmon);  
    ;  
QUIT;
```

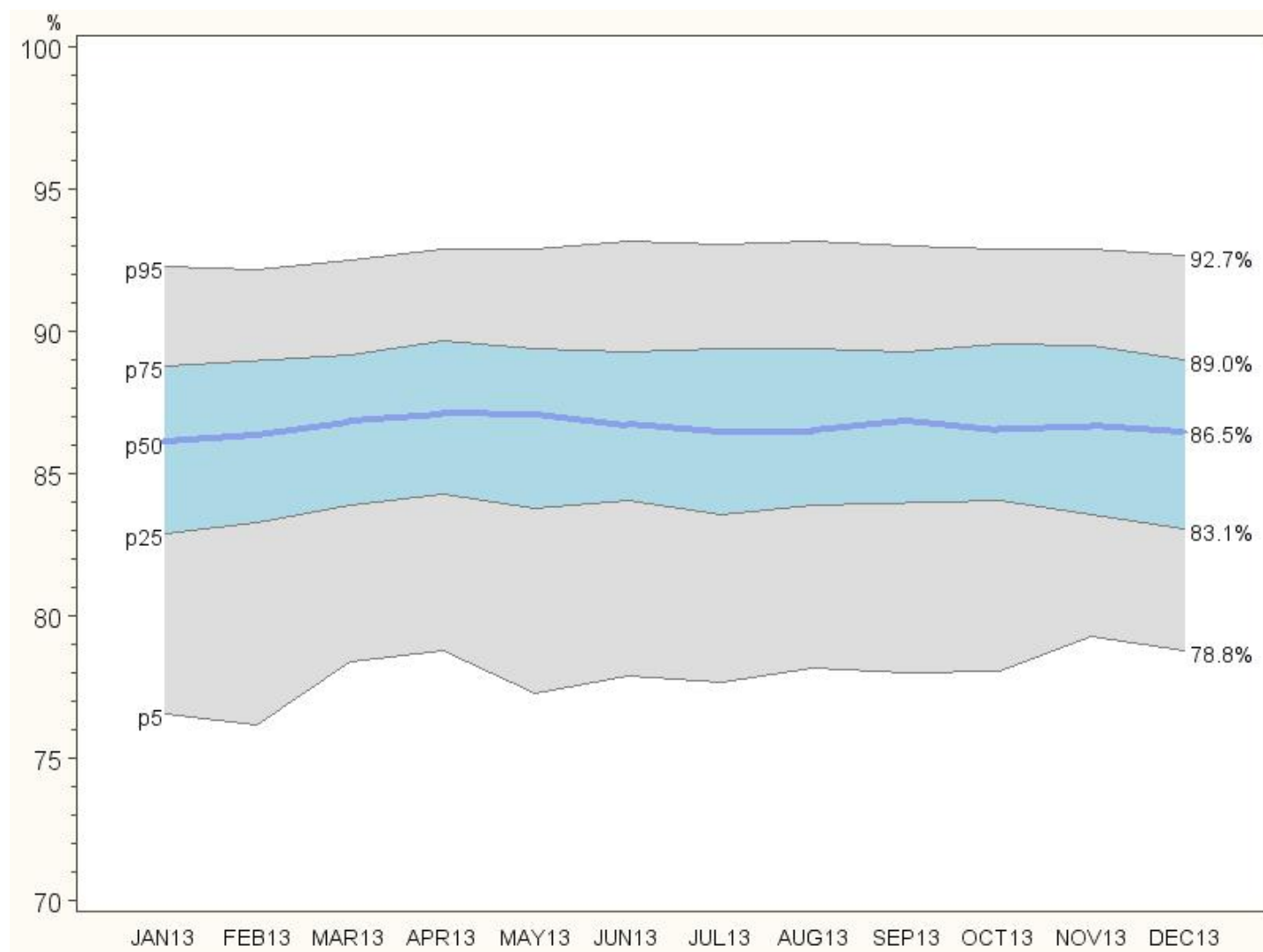


# Statistics

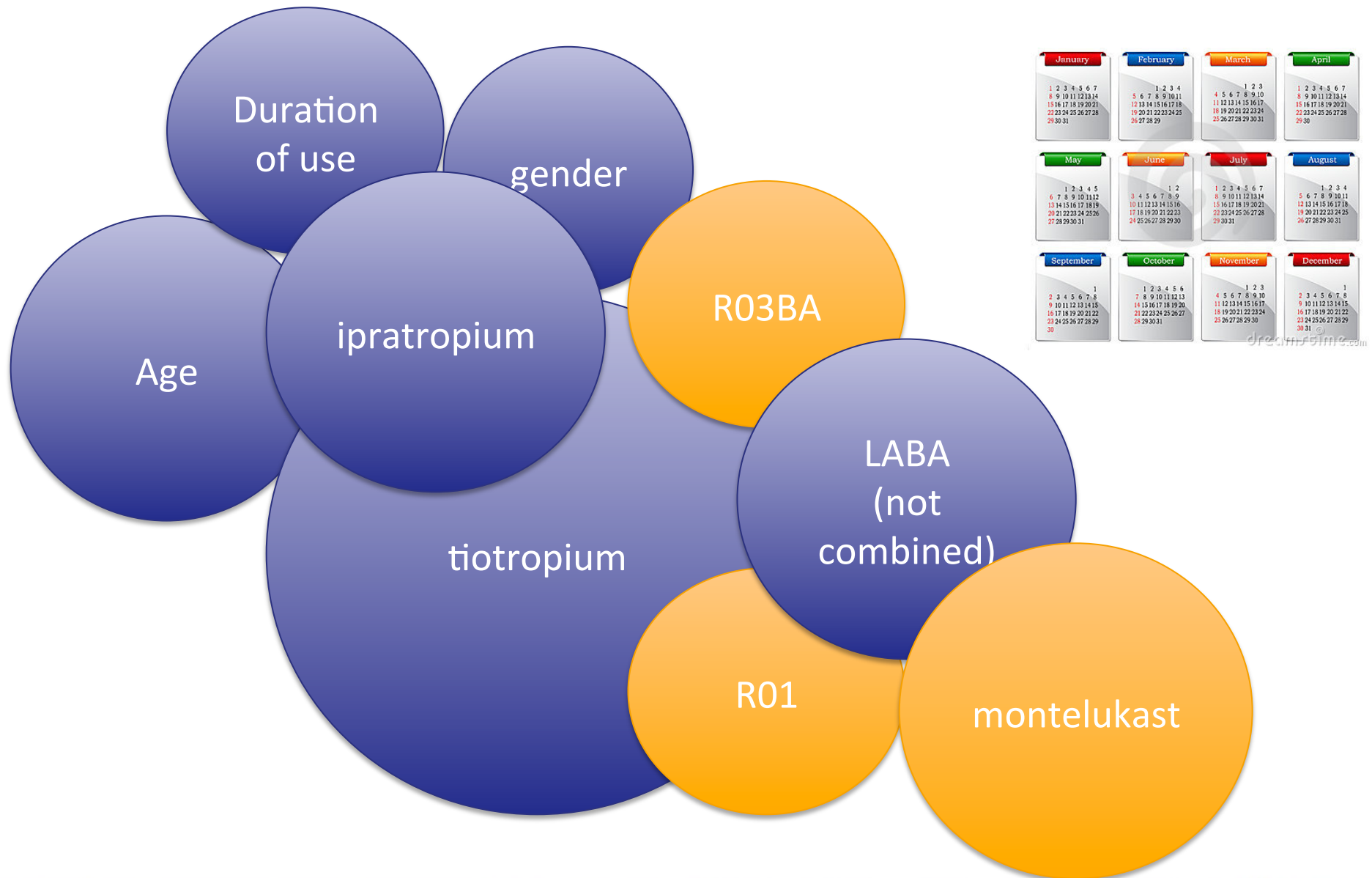
- Descriptive
  - Mean, median, percentiles etc
- Modelling
  - Influence of surrounding factors
- Forecasting
- Explorative
- Pattern seeking



# Diabetes adherence – descriptive

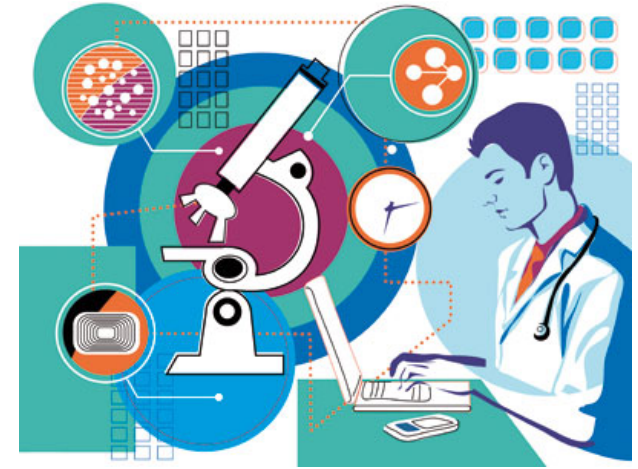


# COPD drug predictors of indication



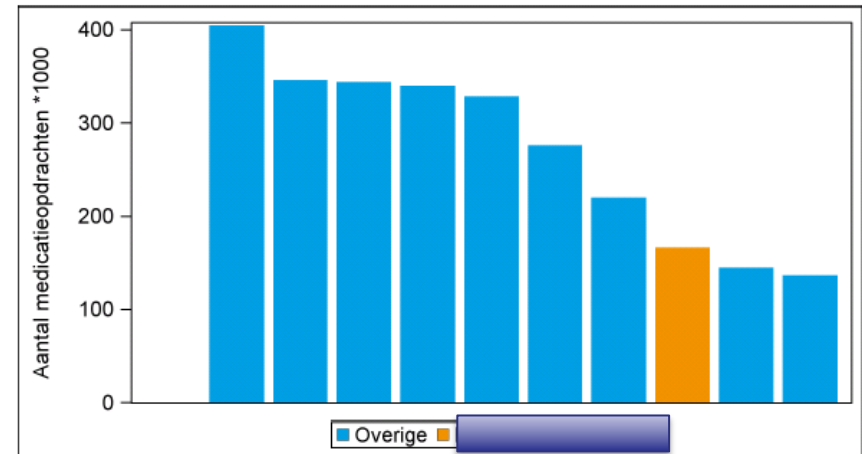
# Scientific rigour

- Understand data
- Distinguish sense from nonsense
- Understand need of customers  
(caretakers and pharma companies)
- Be able to fund and discuss the obtained results

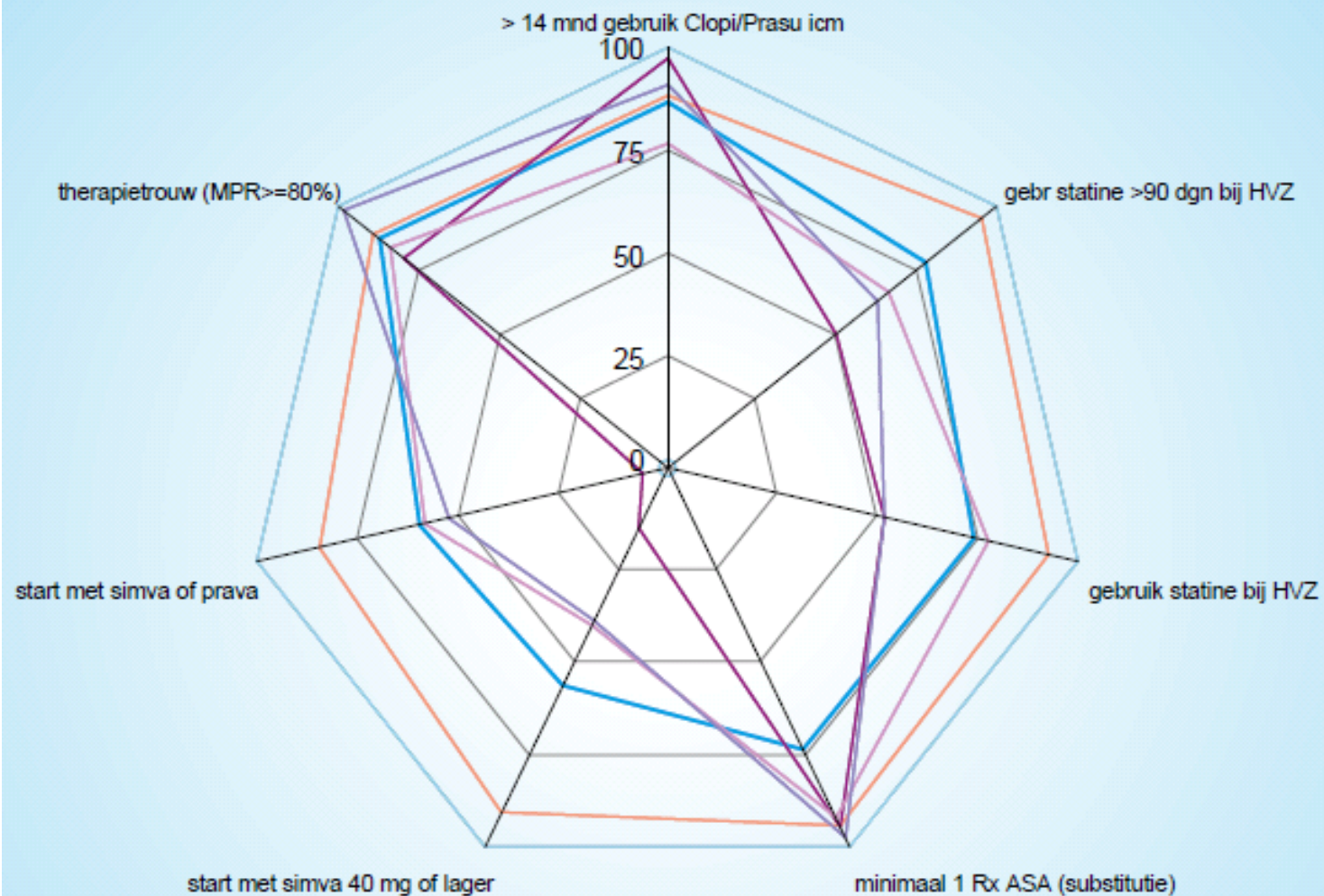


# Visualisations

- Two types
  - Explorative
  - Explanatory
- Give insight in a glance
- Insight in data
  - Understandable / Intuitive
  - Not too much information



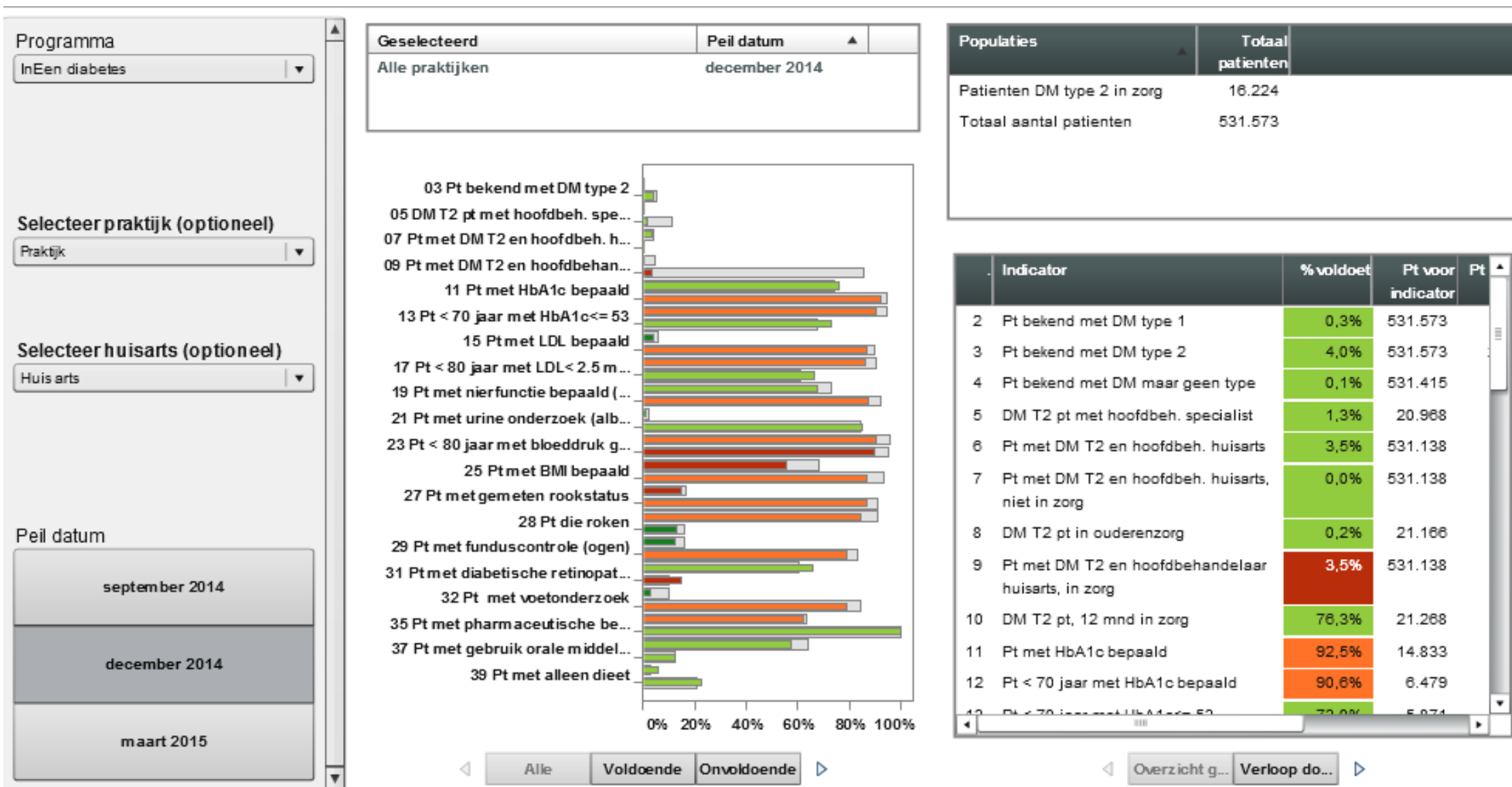
## Percentielen ten opzichte van STIZON apotheek benchmark



Legenda



100% is de beste apotheek, 0% is de minste apotheek, 50% is de mediaan.



# Hacker mindset

- Curious about the data
- Problem solver
- Creative
- Out of the box
  - Try to find what is not obvious...
  - Find abberant patterns ...



# Using real world clinical data

- **Privacy**
  - Permission of patient and the owner of the data (caretaker)
  - Data security
  - Anonymizing
  - Documentation of process
- **Big data**
  - Millions of records
  - At least 7 years history
  - Lot of possible links to other databases

# MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

## MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

## PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

## DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

## COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



**MarketingDistillery.com** is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

*Marketing*  
**DISTILLERY**  
(c) Krzysztof Zawadzki

[WWW.PHARMO.COM](http://WWW.PHARMO.COM)

