Data warehousing and analytics for healthcare

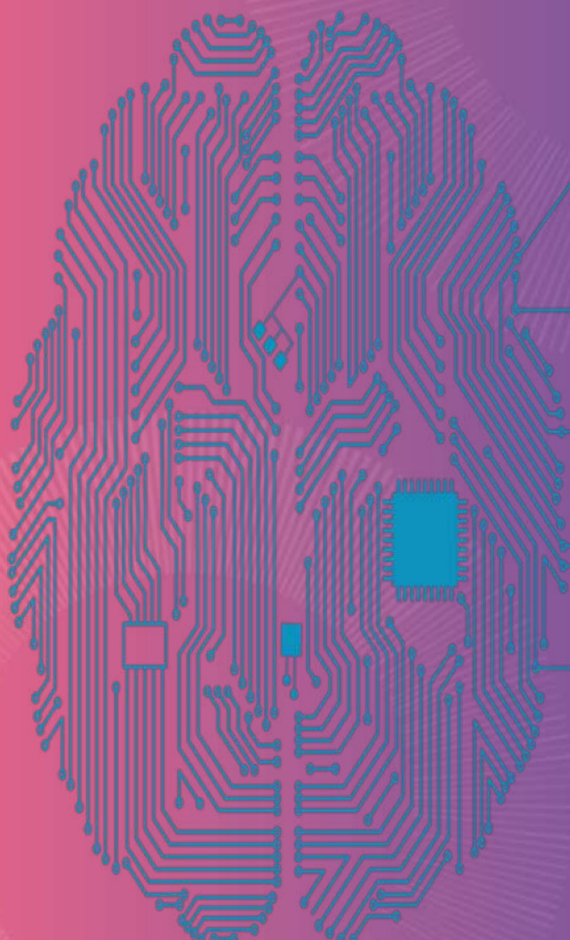
Dr. Daniel Kapitan | Chief Data Scientist | Mediquest

TU/e Data Science Center

Eindhoven, 29 May 2019

mediquest

The promise of AI: will computers be able to do all this?



natural language processing

knowledge representation

automated reasoning

machine learning

computer vision

robotics

Norig & Russel: *Artificial Intelligence: A Modern Approach* (first edition 1995, third edition 2009)

mediquest

Real life is less sexy, and more hard work

Sanders (2012), *Healthcare Analytics Adoption Model*

mediquest

Today’s agenda

5 – 8: Healthcare analytics

Case study: predicting outcomes of surgery

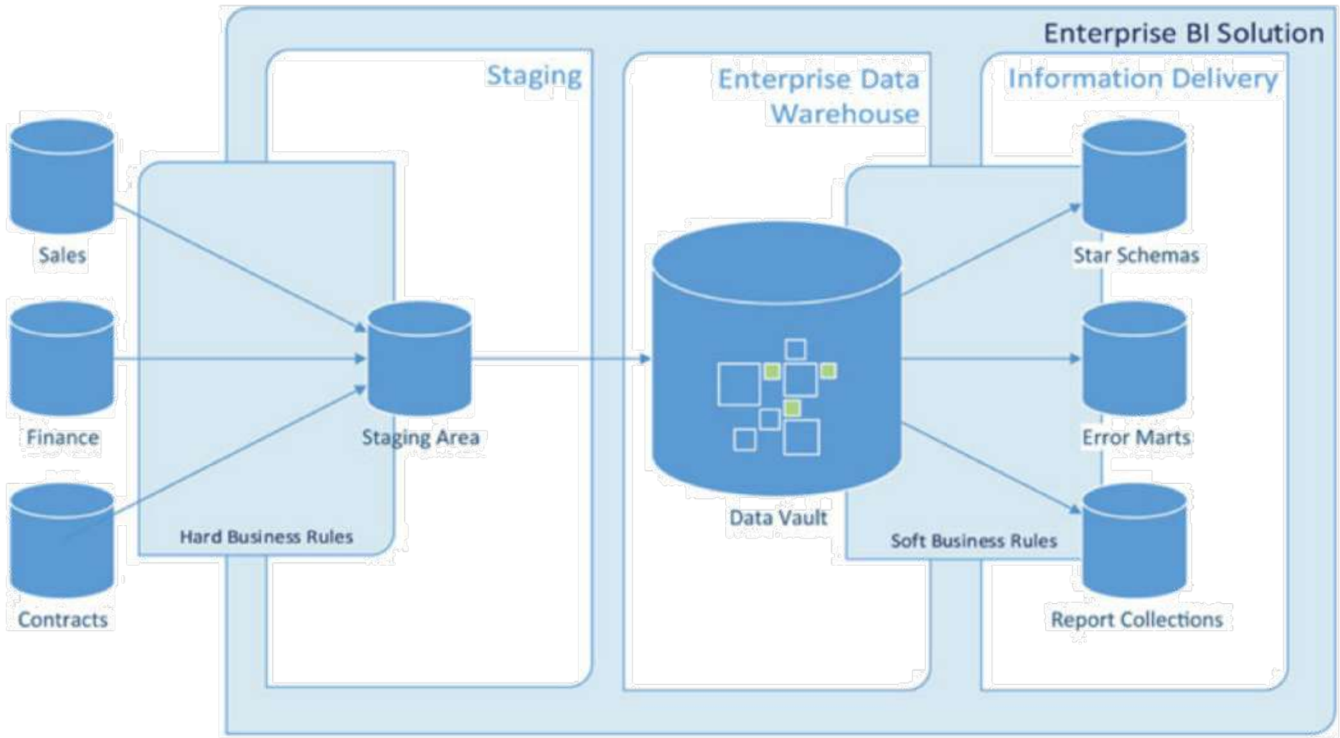
1 – 4: Laying the foundation

- Data warehousing
- Data integration
- Semantic modelling
- Business intelligence

Part I: Laying the Foundation

DATA WAREHOUSING IN HEALTHCARE

The main components of a data warehouse



Sources

Staging area / Source of Record (SoR) data lake

Data vault

Data marts

Hultgren (2012), *Modelling the Agile Data Warehouse With Data Vault*

mediquest

The 'Supplier Landscape': choices, choices …

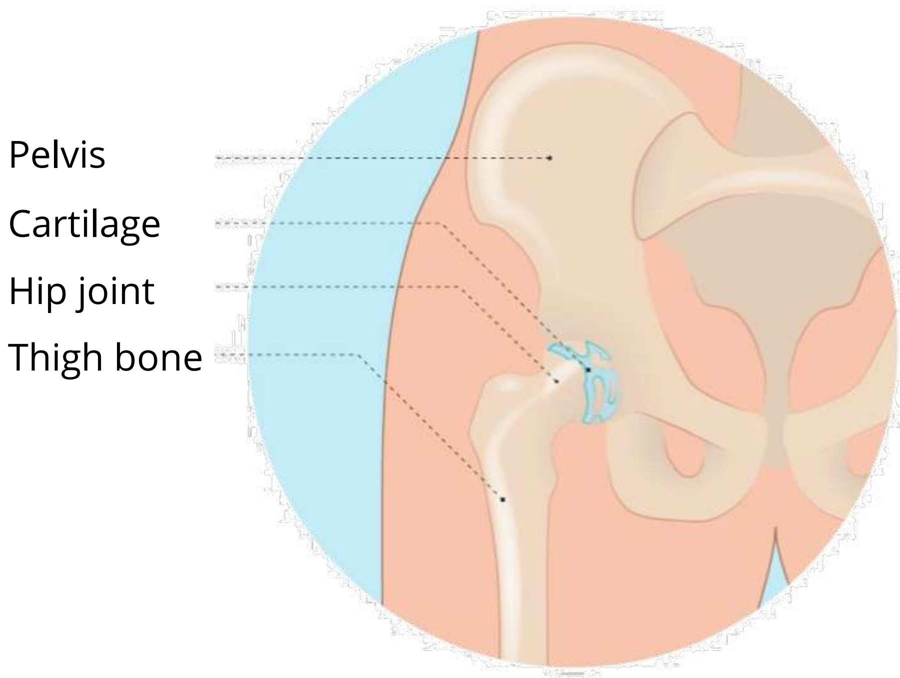
http://mattturck.com/bigdata2018/

- **How many layers:**
  decoupling vs. simplicity?
- **Storage platform:**
  RDBMS, NoSQL, cloud …?
- **Data integration:**
  Semantic modelling, dimensional modeling … ?
- **Way of working:**
  hand-coded vs. graphical tools?
- **Dashboarding & data viz:**
  Yellowfin, Tableau, PowerBI …

Choosing the storage engine

	PostgreSQL	Microsoft SQL Server	Google BigQuery
Daily use (from a data scientist's perspective)	• good CSV handling • Unicode support • Regular expressions • ANSI SQL compliance	• native Excel • Nice GUI IDE for querying, maintenance and workflow	• Cloud-native, low maintenance • Easy to use with API
Platform	• All major OS-es	• Windows, Linux added recently	• Cloud (lock-in)
Extensibility	• Many languages built-in (Python, Javascript, R)	• R built-in (acquisition RStudio) • .NET framework	• Javascript • Tight integration GCP
Specific for data analytics	• Best-in-class for GIS • jsonb format for unstructured data storage • MADlib for built-in machine learning	• Integrates well with Power BI stack • Hybrid solution cloud – on-premise possible with Azure SQL	• StandardSQL performs well even on petabytes • BigQuery ML
TCO	• Forever free with community version • Enterprise DB for paid support (same pricing as MS SQL)	• Value for money with Standard Edition • Gets expensive when Enterprise features are needed	• Pay only for queried volume

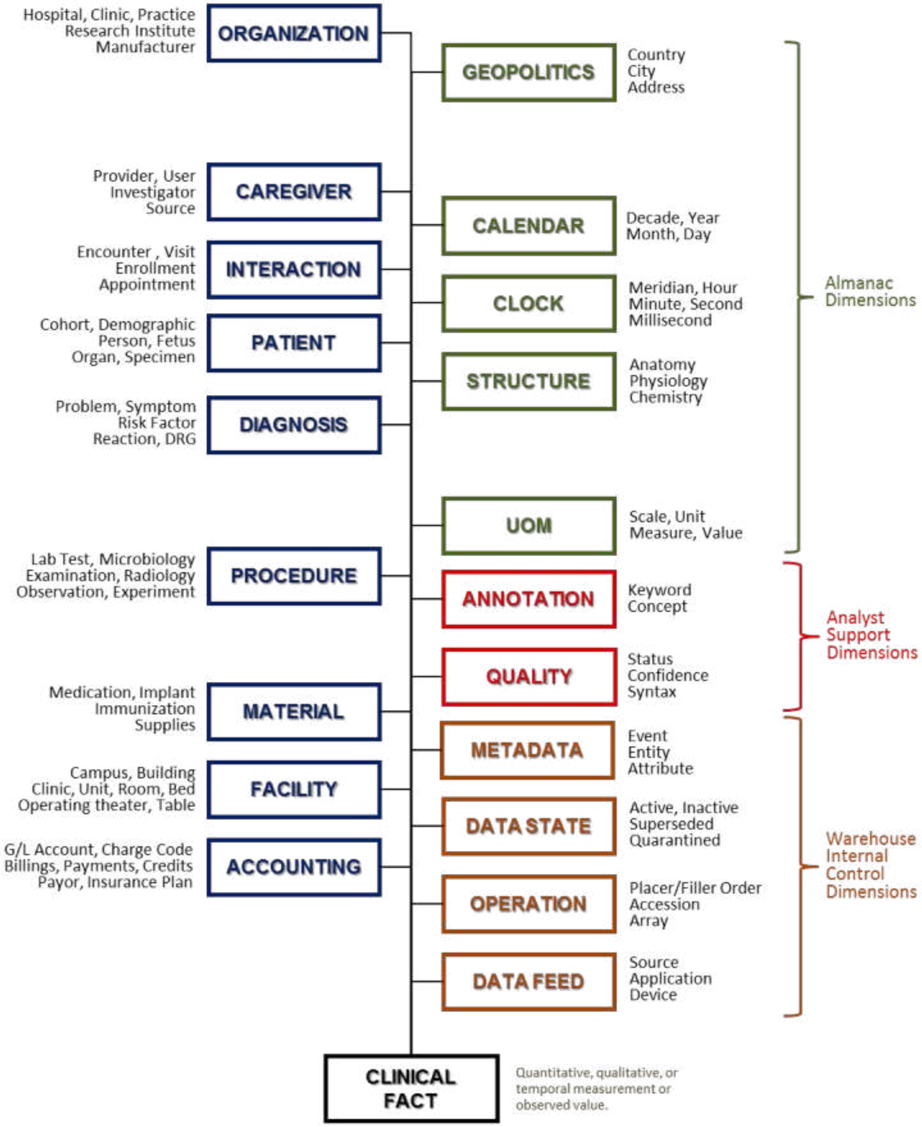
Data integration with semantic dimensional modelling



- Open a webbrowser and go to: https://zibs.nl/wiki/HCIM_Mainpage
- **Assignment: design a data model for a orthopedic clinics**
- Each data element should be captured/mapped to a 'zorginformatiebouwsteen'

Think about all the data you need to integrate so you can do funky machine learning on it

Data integration with semantic dimensional modelling



- Concept of star schema:
  - Typically 5 to 10 fact-tables
  - 20 to 30 dimension tables
- Modelling challenges:
  - Uniformity of dimensions (using same codes)
  - Uniformity of business keys (how to uniquely identify a hip implant)
  - Privacy-sensitive data (SSN, identifiable data)
- Engineering challenges:
  - Dealing with changes in data
  - Speed of batch processing

Biehl (2015), *Data Warehousing for Biomedical Informatics*

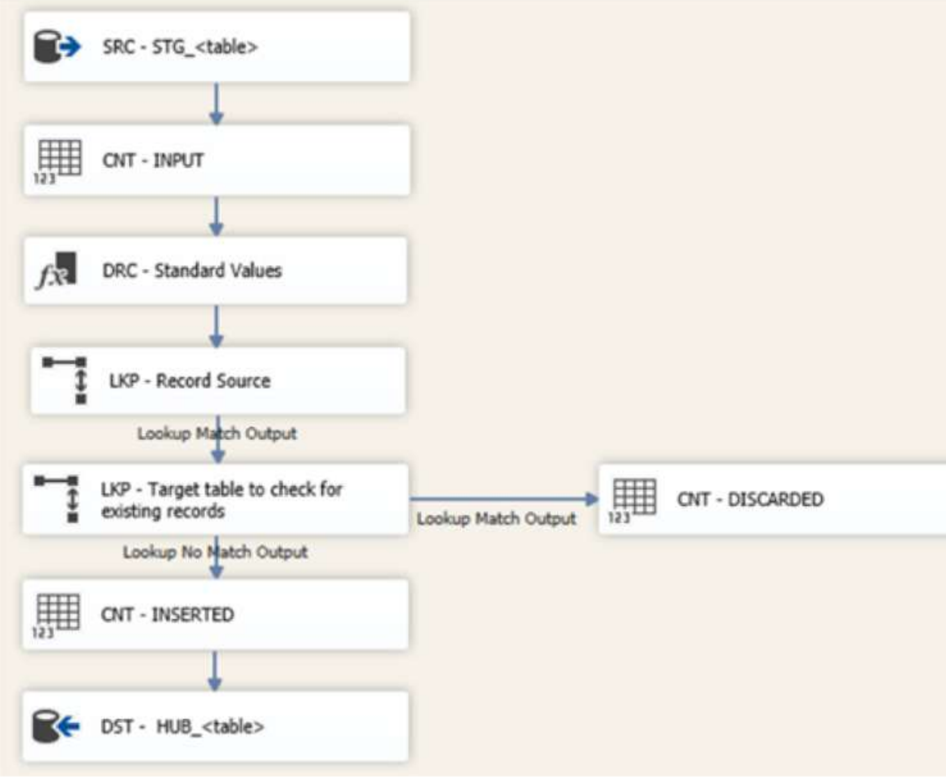
mediquest

Choose your way of working

Hard/hand-coded scripts

```
def init_sor_persoon_hstage_to_patient_mappings(sor):
    mappings = []

    mapping = SorToEntityMapping('persoon_hstage', Patient, sor)
    mapping.map_bk(["'timeff'", 'ifct_relatienr'])

    # SAT Patient.Identificatie
    mapping.map_field("extern_patientnummer", Patient.Identificatie.extern_nummer)
    mapping.map_field("ifct_relatienr", Patient.Identificatie.nummer)
    mapping.map_field("ifct_id", Patient.Identificatie.bron_id)

    # SAT Patient.Default
    mapping.map_field("ifct_geboortedtm::date", Patient.Default.geboortedatum)
    mapping.map_field("ifct_geslacht", Patient.Default.geslacht_code)
    mapping.map_field("", Patient.Default.meerling_indicator)
    mapping.map_field("", Patient.Default.overlijdens_indicator)
    mapping.map_field("", Patient.Default.datum_overlijden)

    # SAT Patient.IdentificatieBewijs
    mapping.map_field("ifct_bsn", Patient.IdentificatieBewijs.nummer, type=Patient.IdentificatieBewij
    mapping.map_field("", Patient.IdentificatieBewijs.geldig_tot, type=Patient.IdentificatieBewijs.Ty
    mapping.map_field("ifct_legitimatieid", Patient.IdentificatieBewijs.nummer, type=Patient.Identifi
    mapping.map_field("", Patient.IdentificatieBewijs.geldig_tot, type=Patient.IdentificatieBewijs.Ty
    mapping.map_field("ifct_rijbewijsnummer", Patient.IdentificatieBewijs.nummer, type=Patient.Identi
    mapping.map_field("", Patient.IdentificatieBewijs.geldig_tot, type=Patient.IdentificatieBewijs.Ty

    # SAT Patient.Adres
    mapping.map_field("ifct_straat_b", Patient.Adres.straat, type=Patient.Adres.Types.woonadres)
    mapping.map_field('(sor_timeff.split_huisnummer(ifct_huisnr_b)).huisnummer', Patient.Adres.huisnu
    mapping.map_field("", Patient.Adres.huisnummerletter, type=Patient.Adres.Types.woonadres)
    mapping.map_field('(sor_timeff.split_huisnummer(ifct_huisnr_b)).huisnummer_toevoeging', Patient.A
    mapping.map_field("", Patient.Adres.aanduiding_bij_nummer_code, type=Patient.Adres.Types.woonadre
    mapping.map_field("ifct_plaats_b", Patient.Adres.woonplaats, type=Patient.Adres.Types.woonadres)
    mapping.map_field("", Patient.Adres.gemeente, type=Patient.Adres.Types.woonadres)
    mapping.map_field("", Patient.Adres.land_code, type=Patient.Adres.Types.woonadres)
    mapping.map_field("ifct_postcode_b", Patient.Adres.postcode, type=Patient.Adres.Types.woonadres)
    mapping.map_field("", Patient.Adres.additionele_informatie, type=Patient.Adres.Types.woonadres)
```

Graphical user interface



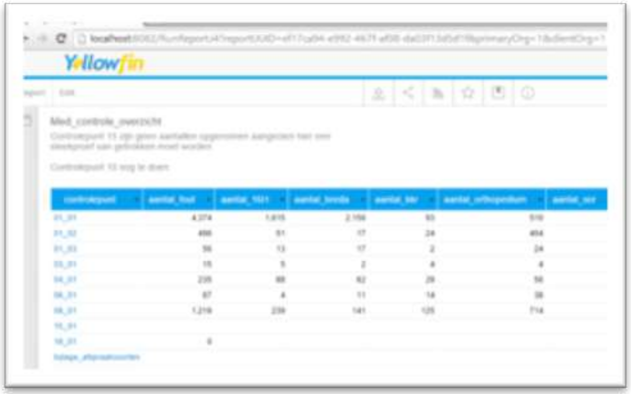
mediquest

Design data marts for flexibility and human-readability

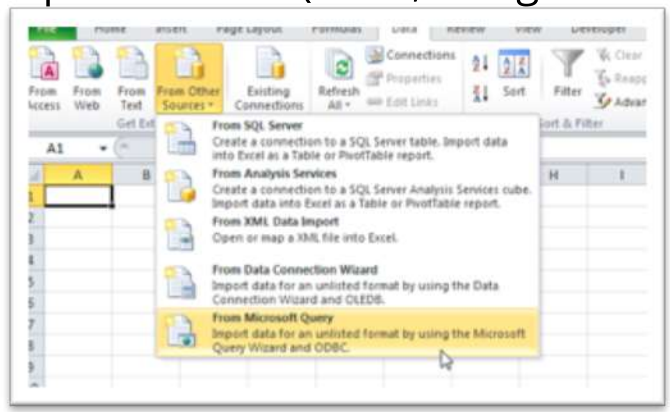
mediquest

Let end-users choose their own tool

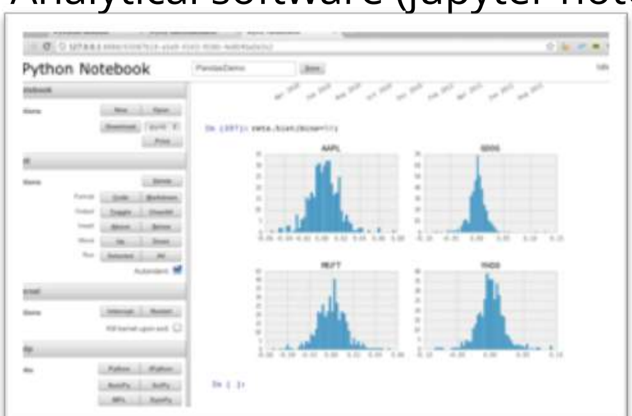
Dedicated BI tool (Tableau, PowerBI, Yellowfin)



Spreadsheets (Excel, Google Sheets, Libre Office)



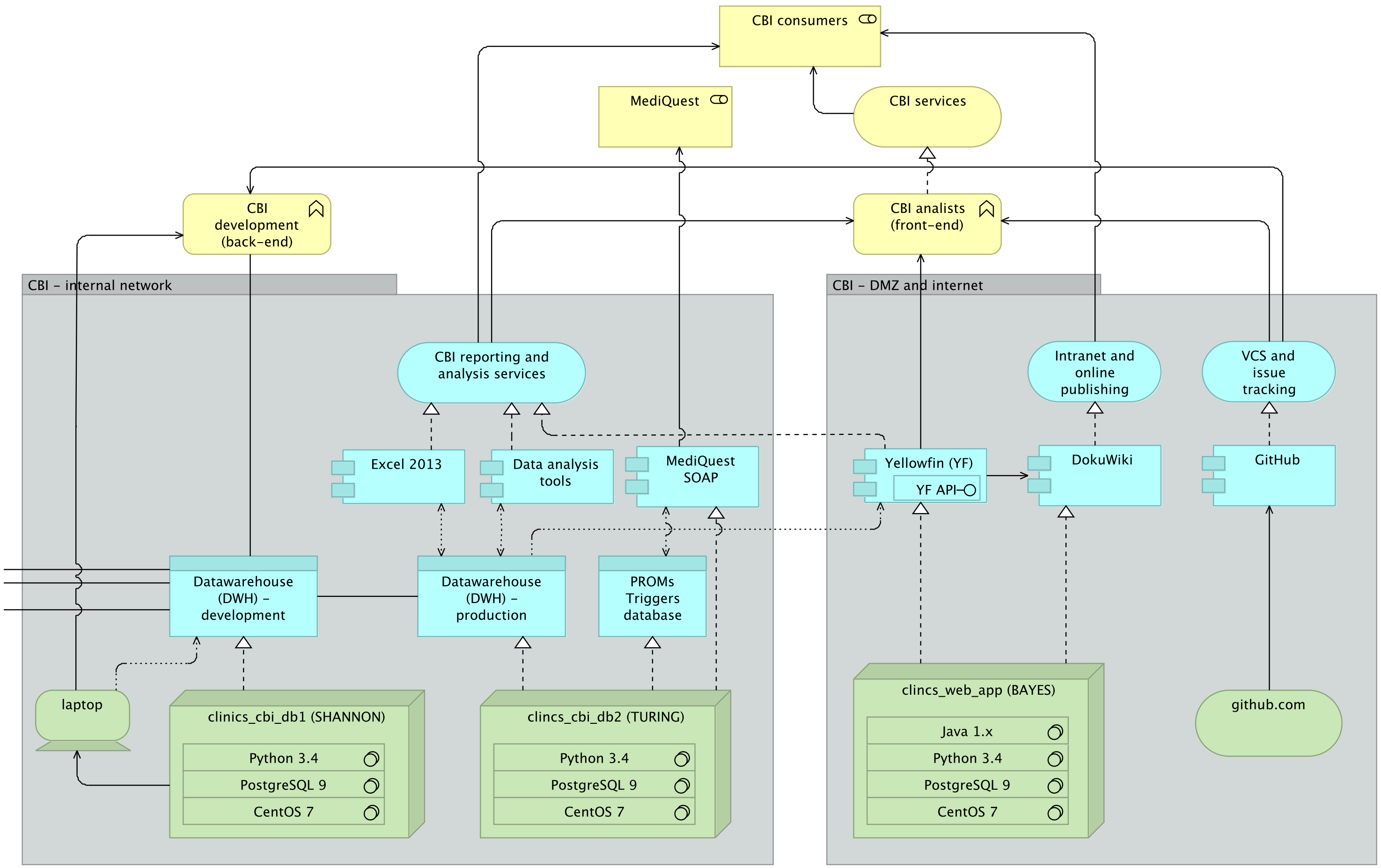
Analytical software (jupyter notebooks, SAS, SPSS)



(... and many more)

mediquest

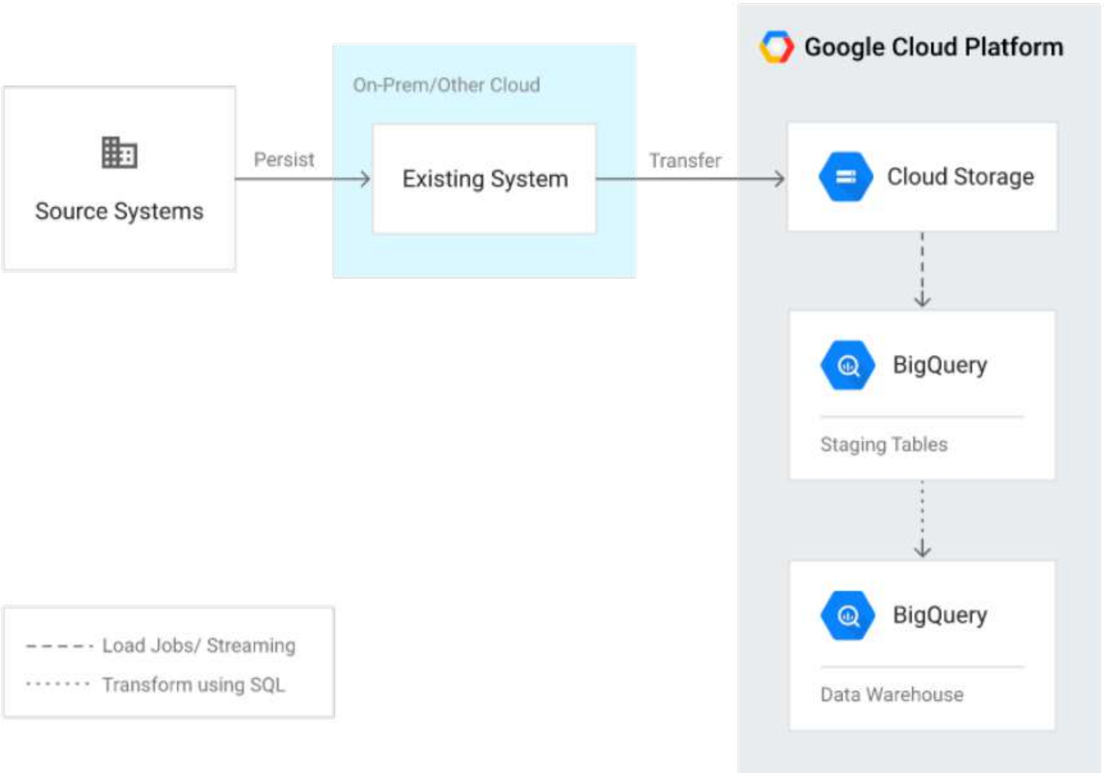
End-result: with own hardware



- Python as the core language for building datavault and analytics
- PostgreSQL database as storage engine
- Infrastructure: virtualized CentOS with SSD SAN

mediquest

End-result: on Google Cloud Platform



- Python and Clojure as the core languages
- BigQuery and Cloud storage as storage engines
- see https://cloud.google.com/solutions/bigquery-data-warehouse for more details

mediquest

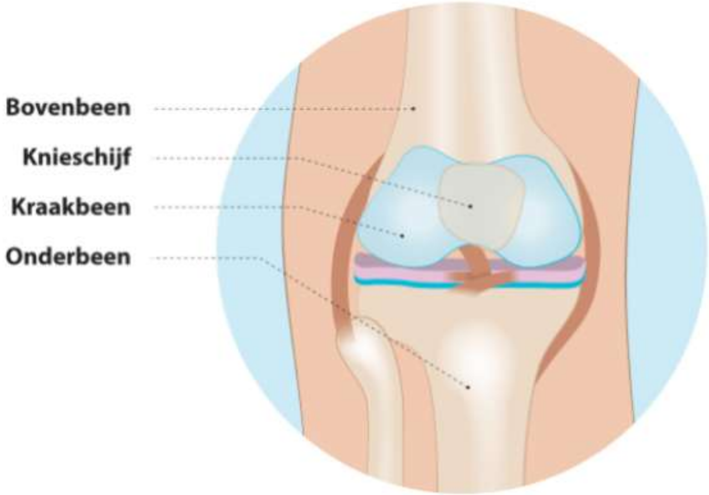
Part II: Healthcare Analytics

PREDICTING OUTCOMES

Imagine you are consulting an orthopedic surgeon because of knee osteoarthritis

Kniegewricht

Het kniegewricht bestaat uit het bot in het bovenbeen, het grote bot in het onderbeen en de knieschrijf. Tussen de botten zit kraakbeen (afbeelding). Het zachtere kraakbeen zorgt ervoor dat de botten makkelijk langs elkaar kunnen bewegen.

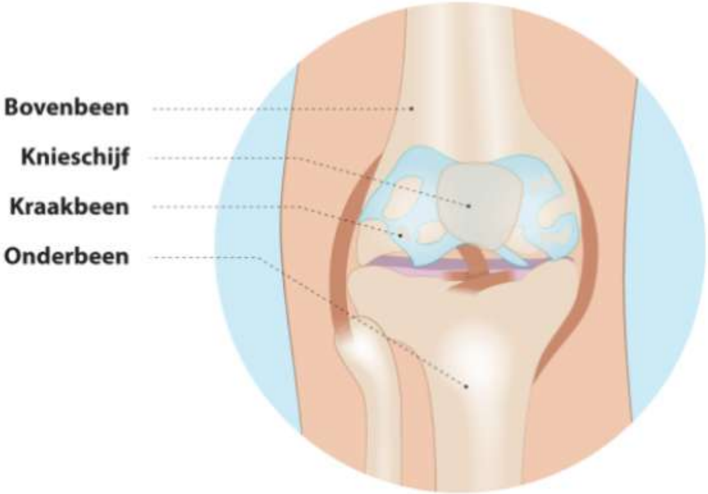


Afbeelding: een gezond kniegewricht.

bron: https://www.keuzehulp.info/

Een versleten knie

Als je ouder wordt, verslijt je kraakbeen. Dit wordt een versleten knie genoemd, of knie-artrose. Bij een versleten knie is het kraakbeen bijna helemaal verdwenen (afbeelding). De botten kunnen niet makkelijk meer langs elkaar bewegen. Dit zorgt voor een pijnlijke en stijve knie. Een versleten knie komt veel voor. [1]



Afbeelding: een versleten kniegewricht.

You have a choice:
operate (new knee) or conservative treatment

Operation:
90 out of 100 patient have less pain.
They are also more active.

Conservative treatment:
Half of all patient have less pain after physiotherapy and taking painkillers.

bron: https://www.keuzehulp.info/

mediquest

What if an algorithm says that in your case, chance of succesful operation is also 50%

Operation:

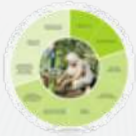
Chance of success of 50%

Conservative treatment:

Half of all patient have less pain after physiotherapy and taking painkillers.

mediquest

So, what are outcomes in healthcare?

	Condition			
	Cataract	Macular Degeneration	Low Back Pain	Hip & Knee Osteoarthiritis
Outcome category				
PACs*	• Re-operation • Endopthalmitis • Corneal oedema	• Endopthalmitis	• Mortality • Readmissions • Postop infections	• Mortality • Readmissions • Postop infections
Patient-reported	• Catquest-9SF	• Brief IVI	• EQ-5D • Oswestry Disability Index • NRS painscore • Work status	• EQ-5D • KOOS/HOOS • NRS painnscore • Satisfaction • Work status
Clinical reported	• Best corrected visual acuity • Refraction	• Best corrected visual acuity • Refraction		• Timed-Up and Go

*Potentially avoidable complications

mediquest

Project Nightingale.

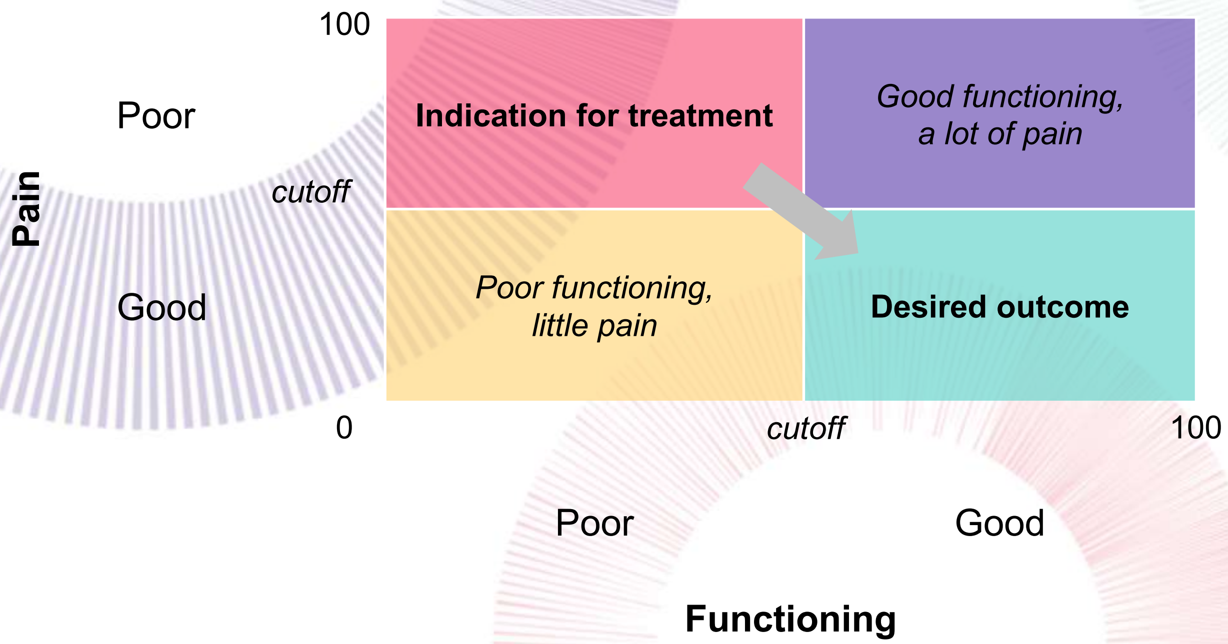
1. Compounded outcome measures relevant for shared-decision making, using existing data dictionairies (ICHOM, national registries)

2. Supervised learning for e.g. identifying high-risk patients prior to an intervention

3. ‘Unboxing the black box’ by relating results of machine learning to existing epidemiological research

mediquest

A simple idea: choose the two most relevant indicators and determine cutoff values for each



mediquest

Outcome total knee replacement in the UK

(data NHS Digital | n=140.000 | period 2011 – 2017 | 284 providers)

Prior to operation

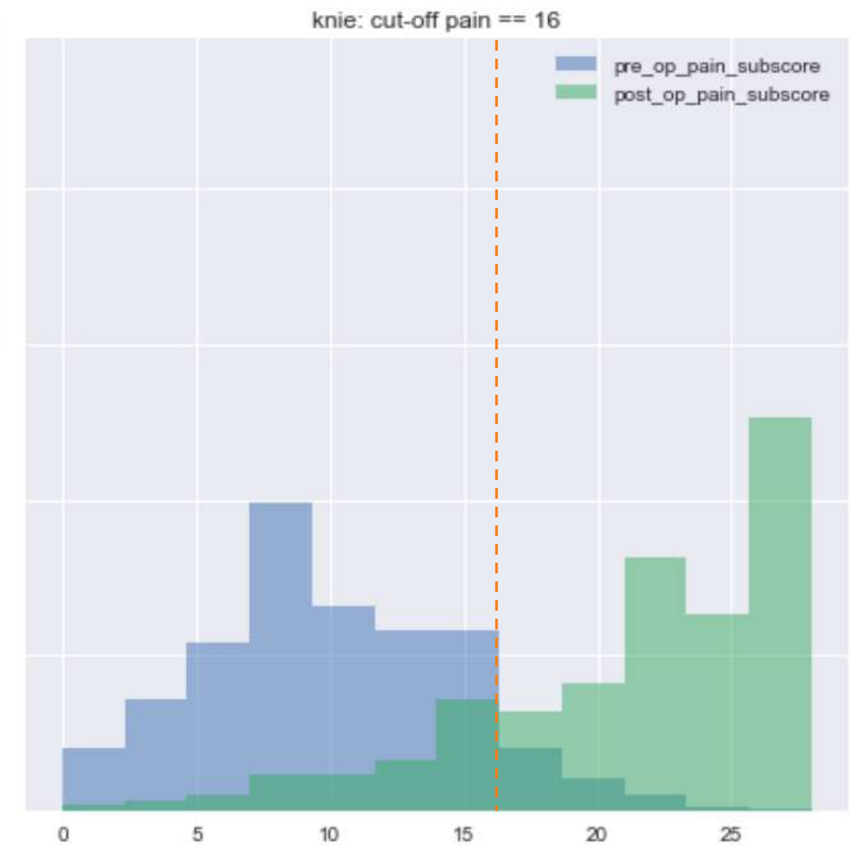
Pre-operative (T0)		Functioning: Poor	Functioning: Good
Pain	Poor	83%	9%
	Good	3%	5%

After operation

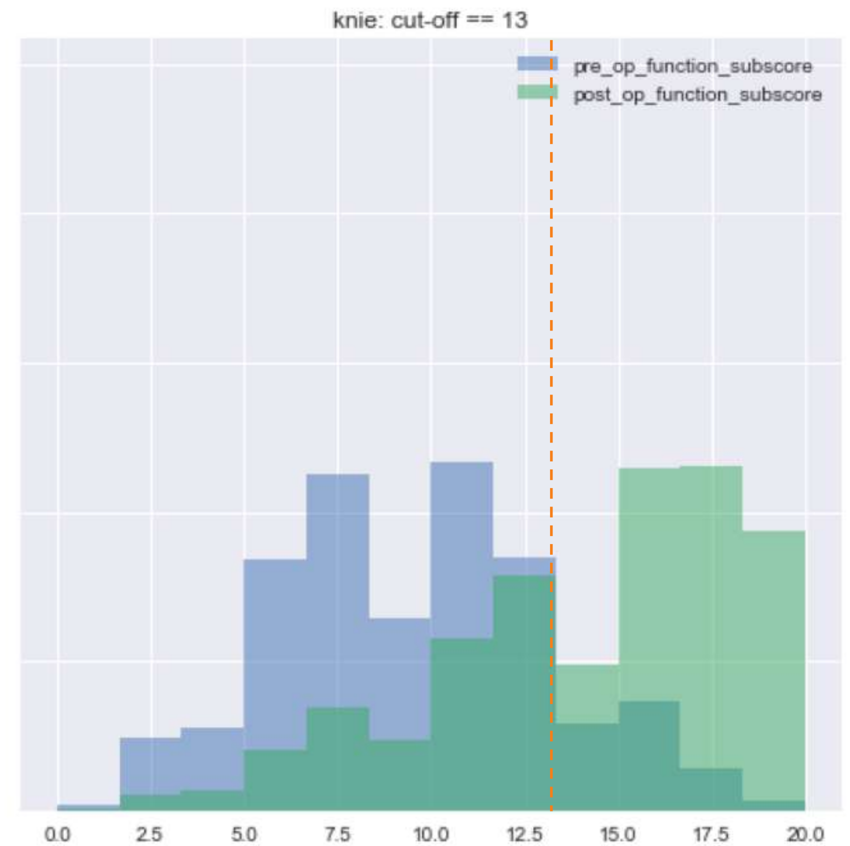
Post-operative (T1)		Functioning: Poor	Functioning: Good
Pain	Poor	18%	1%
	Good	20%	61%

Measuring outcomes is not always straightforward

Pain



Functioning



mediquest

Prior to cataract surgery.

Pre-operative (T0)		Visual acuity: Poor	Visual acuity: Good
PROMs	Poor	50%	24%
	Good	15%	11%

- 50% of patients have consistent indication (poor-poor)
- … but how about the other half?

mediquest

After cataract surgery.

Pre-operative (T0)		Visual acuity: Poor	Visual acuity: Good
PROMs	Poor	50%	24%
	Good	15%	11%

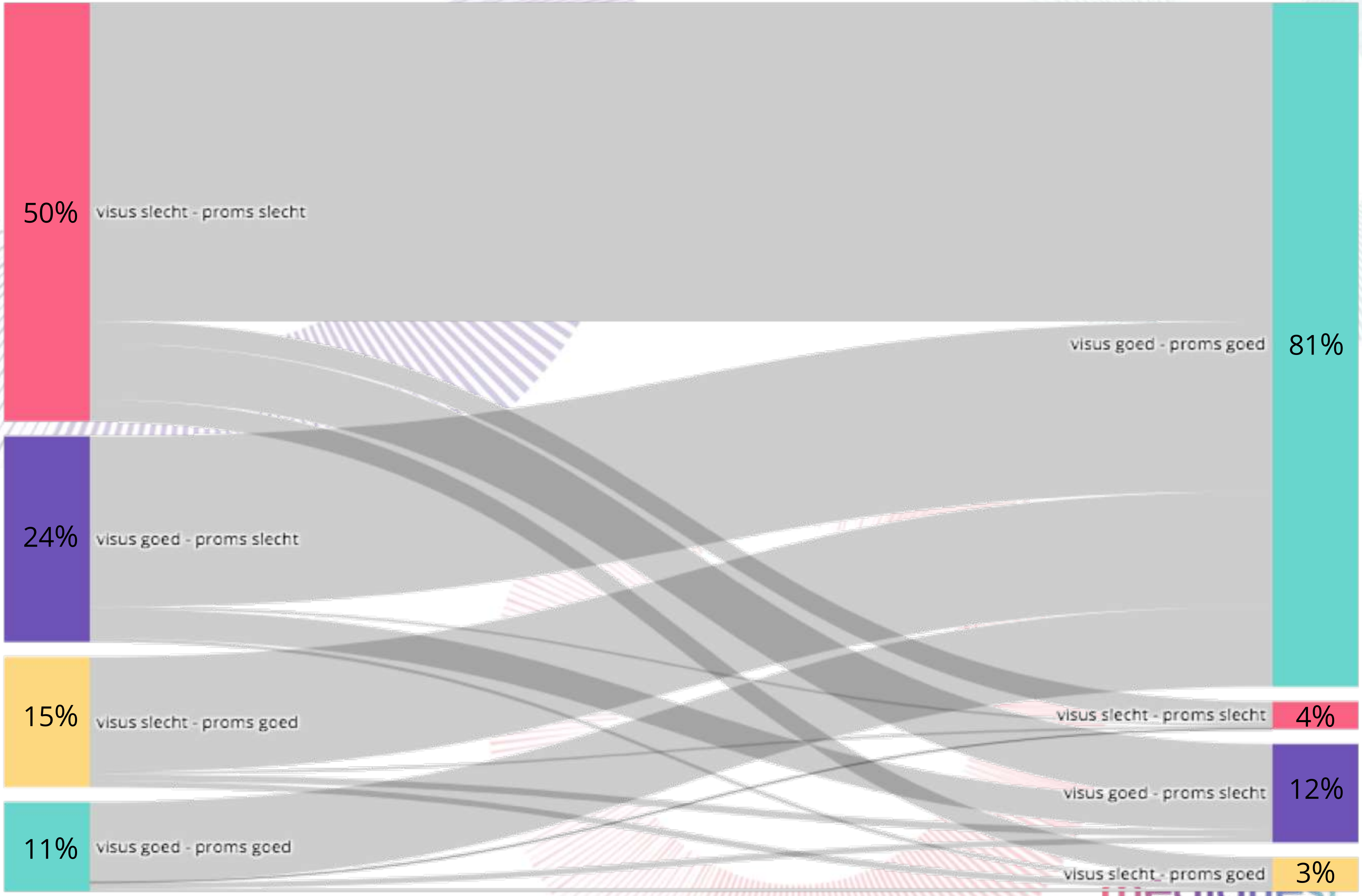
Post-operative (T1)		Visual acuity: Poor	Visual acuity: Good
PROMs	Poor	3%	12%
	Good	4%	**81%**

- 50% of patients have consistent indication (poor-poor)
- … but how about the other half?

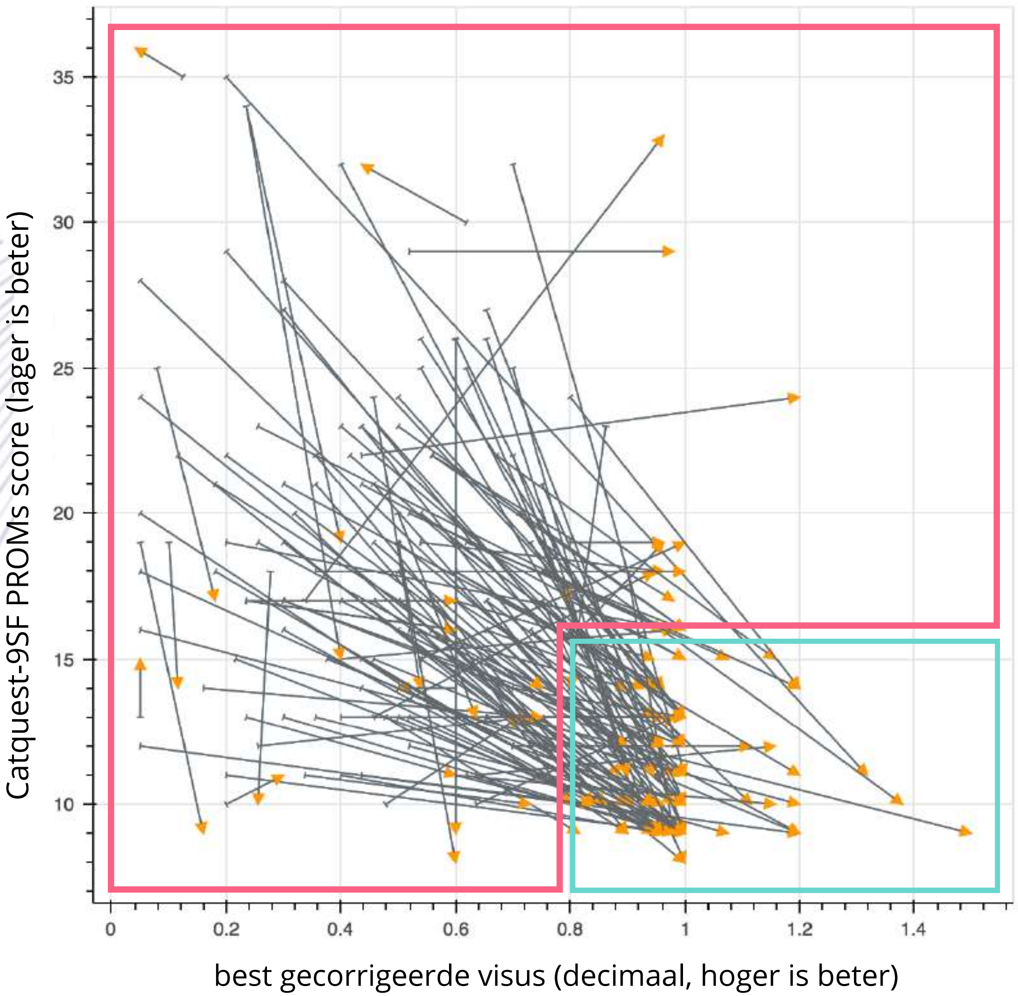
- Good outcome for 81% of all patients
- Remaining 'outliers' of 19% require more detailed inspection

mediquest

No simple, linear mapping between pre to post



Can we predict the outcome prior to surgery?



- **Sensitivity 0.5**
  Half of the arrows that end up in the red quadrants kan be identified prior to surgery; 9% of all patient receive correct warning signal

- **Positive predictive value 0.58**
  I.e. 42% of warning signals is false-positive. Good enough?

mediquest

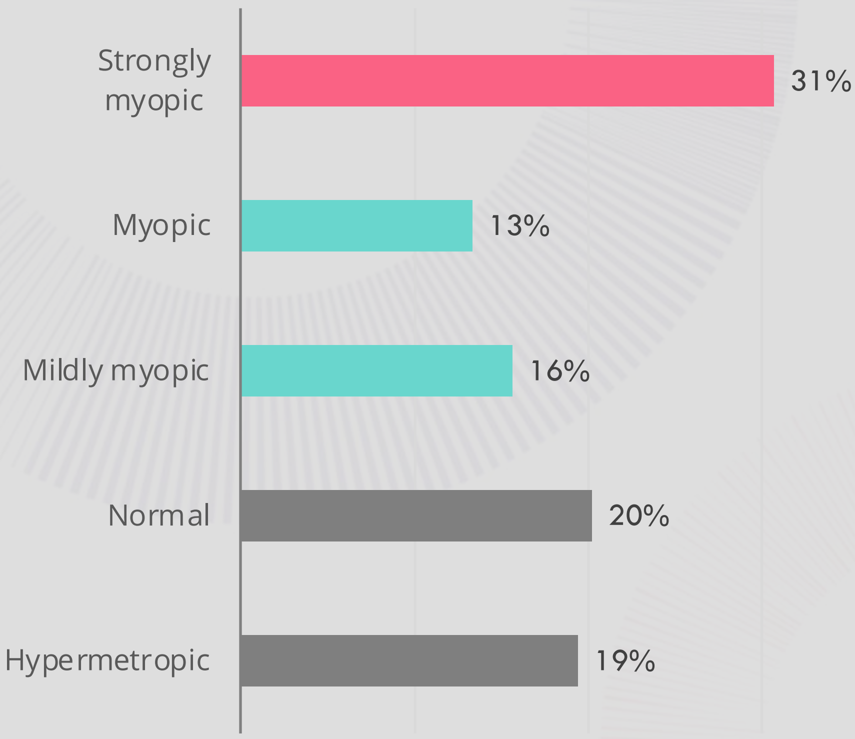
Do we understand what the algorithm does?

Risk factors known from literature
Post-operative complications
Best corrected visual acuity
Target refraction
Capsule complications
Ocular co-morbidities
PROMs sumscore
Gender
Age

Relative feature importance in random forest
Target refraction
Age
Best gecorrected visual acuity
PROMs sumscore
PROMs sub-score near-vision
PROMs sub-score far-vision
PROMs sub-score general satisfaction
Individual PROMs items
Gender
Previous surgery other eye
Macular degeneration
Other ocular co-morbidities

- Lundström, M. and Stenevi, U., *Analyzing Patient-Reported Outcomes to Improve Cataract Care,* Optometry and Vision Science 2013: vol. 90 no. 8: 754-759
- Grimfors et al., *Ocular comorbidity and self-assessed visual function after cataract surgery,* J. Cataract Refract Surg 2014; 40:1163-1169
- Lundström et al., *Visual outcome of cataract surgery,* J. Cataract Refract Surg 2013: 39:673-679
- Mollazadegan, K. and Lundström, M., *A study of the correlation between patient-reported outcomes and clinical outcome after cataract surgery in ophthalmic clinics*, Acta Ophthalmol. 2015: 93: 293-298

Significant effect size in outcome by target refraction (strength of glasses)



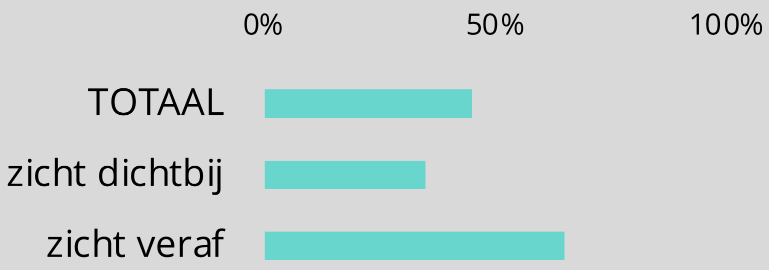
Percentage poor outcome by target refraction, i.e. chosen strength of glasses post-surgery

Target refraction group in diopters (n observations)

- Strongly myopic:
  < -4.0 (n=13)
- Myopic:
  between -4.0 and -2.0 (n=771)
- Mildly myopic:
  between -2.0 and -0.5 (n=319)
- Normal:
  between -0.5 and 0.5 (n=3993)
- Hypermetropic:
  > 0.5 (n=36)

Staaroperatie keuzehulp

Bent u tevreden met uw zicht?



Wat vindt u belangrijk?

Activiteiten en hobbies met:

- ☐ zicht dichtbij
- ☐ zicht veraf

Ik wil dit **met/zonder** bril kunnen

Uw ogen op dit moment

	L	R
zicht met bril:	0.6	0.4
brilsterkte:	-1.5	-2.0

andere aandoeningen: macula degeneratie

Verwachte uitkomst operatie

	L	R
zicht met bril:	0.6	1.0
brilsterkte:	-1.5	-0.5

aandachtspunt: ⚠️ risico eindresultaat

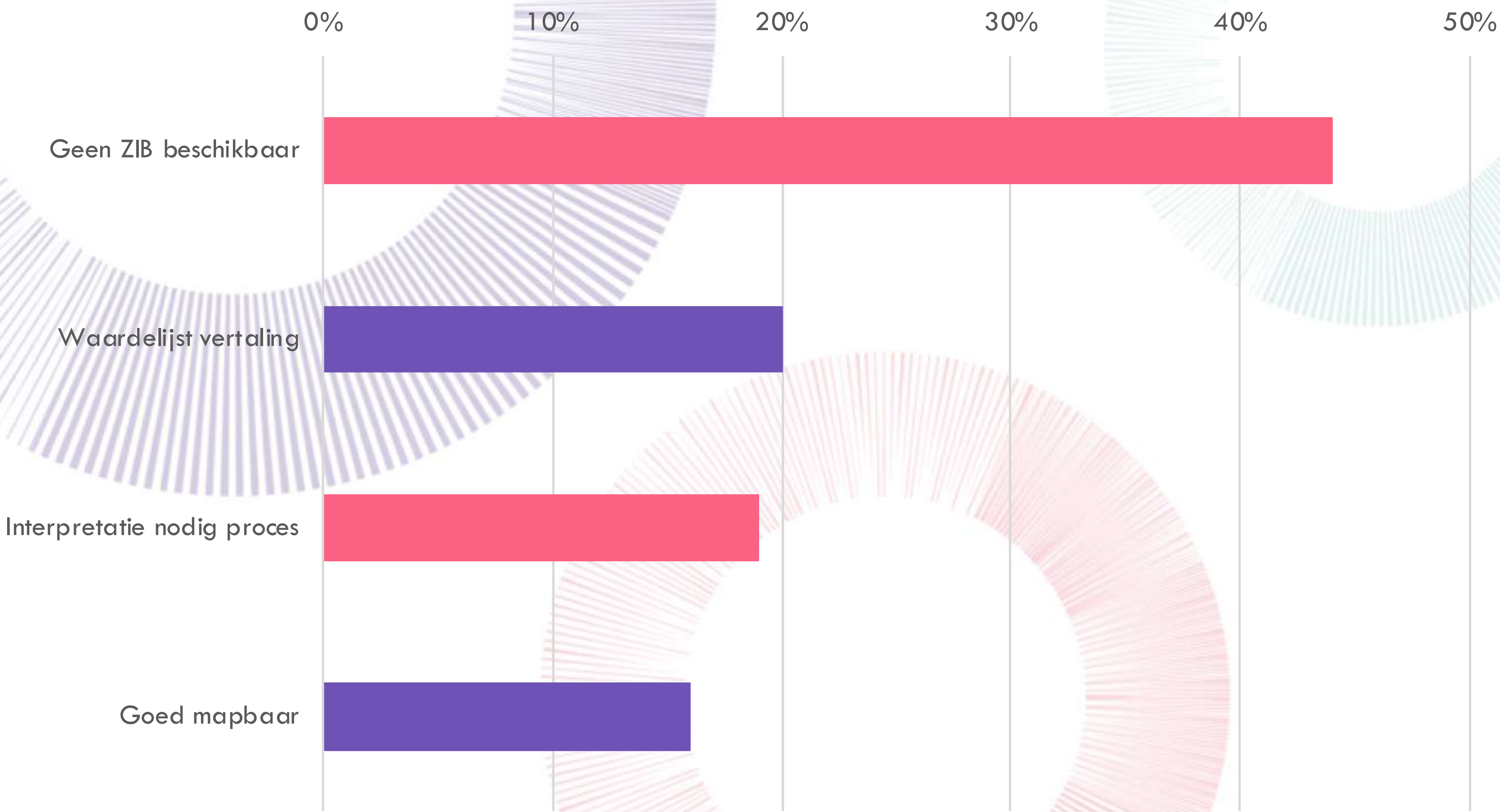
mediquest

Lessons learned from applying machine learning in daily operations

1. Data quality (*registratie aan de bron*)
2. Harmonisation and semantic integration of different registries (e.g. recent analysis Nictiz)
3. Open sourcing trained models for clinical decision support

Mapping ICHOM to Dutch standards

(Analysis Nictiz, August 2018)



mediquest

More info?

- Drop me an email at **dkapitan@mediquest.nl**

- Connect on LinkedIn **https://www.linkedin.com/in/dkapitan/**

mediquest