Data warehousing and healthcare analytics

Dr. Daniel Kapitan | data scientist in healthcare and public domain

JADS Jheronimus Academy of Data Science

Please allow me ...

Associate Data Science

VERDONCK KLOOSTER & ASSOCIATES

sinds april 2020

Verdonck, Klooster & Associates maken kiezen mogelijk. Juist als het moeilijke keuzes zijn. Als data expert werk ik op projecten waarin data-gedreven werken en data science centraal staan.

Academic Director Healthcare

JADS

sinds april 2020

Samen met Joran Lokkerbol ben ik het Data Science in Healthcare programma verder aan het ontwikkelen en vormgeven.

Associate Data

WIELINQ

sinds januari 2020

Met digitalisering wil Wielinq de klantbeleving drastisch verbeteren. Want: klanten, studenten, patiënten, huurders of medewerkers verwachten dat zij altijd en overal toegang hebben tot informatie en diensten. Als Associate richt ik mij o.a. op data-gedreven werken, implementatie van dataplatformen en vraagstukken op het gebied van data governance.

Data onderzoeker

Timefflabs

sinds november 2019

Als data onderzoeker van Timeff Labs leid ik de ontwikkeling van nieuwe, machine-learning gebaseerde functionaliteit voor het Emma EPD.

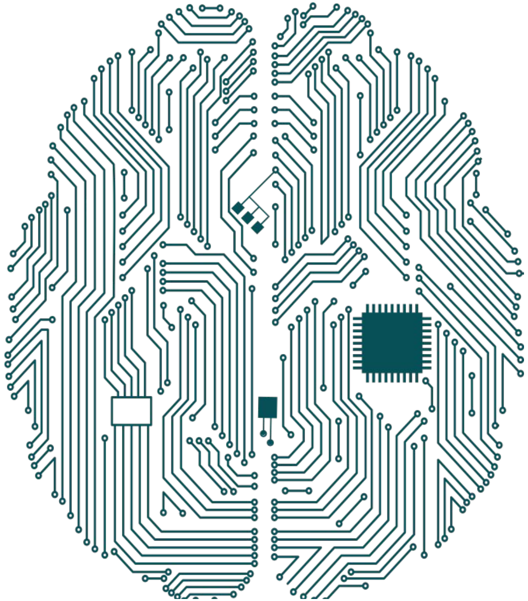
Chief Data Scientist

mediquest

sinds februari 2018

Als Chief Data Scientist werk ik met het team bij Mediquest aan de ontwikkeling van modellen voor uitkomstgerichte zorg en beslisondersteuning.

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



- natural language processing
- knowledge representation
- automated reasoning
- machine learning
- computer vision
- robotics

Norvig & Russel: *Artificial Intelligence: A Modern Approach* (third edition 2009, fourth edition forthcoming 2020)

Real life is less sexy, and more hard work

Level 8	Personalized Medicine & Prescriptive Analytics	Tailoring patient care based on population outcomes and genetic data. Fee-for-quality rewards health maintenance
Level 7	Clinical Risk Intervention & Predictive Analytics	Organizational processes for intervention are supported with predictive models. Fee-for-quality includes fixed per capita payment.
Level 6	Population Health Management & Suggestive Analytics	Tailoring patient care based upon population metrics. Fee-for-quality includes bundled per case payment.
Level 5	Waste & Care Variability Reduction	Reduction variability in care processes. Focusing on internal optimization and waste reduction.
Level 4	Automated External Reporting	Efficient, consistent production of reports and adaptability to changing requirements
Level 3	Automated Internal Reporting	Efficient, consistent production of reports and widespread availability in the organization.
Level 2	Standardized Vocabulary & Patient Registries	Relating and organizing the core data content.
Level 1	Enterprise Data Warehouse	Collecting and integrating the core data content.
Level 0	Fragmented Point Solutions	Inefficient, inconsistent versions of the truth. Cumbersome internal and external reporting.

Sanders (2012), Healthcare Analytics Adoption Model.

Today's agenda

5 - 8: Healthcare analytics

Case study on predicting outcomes

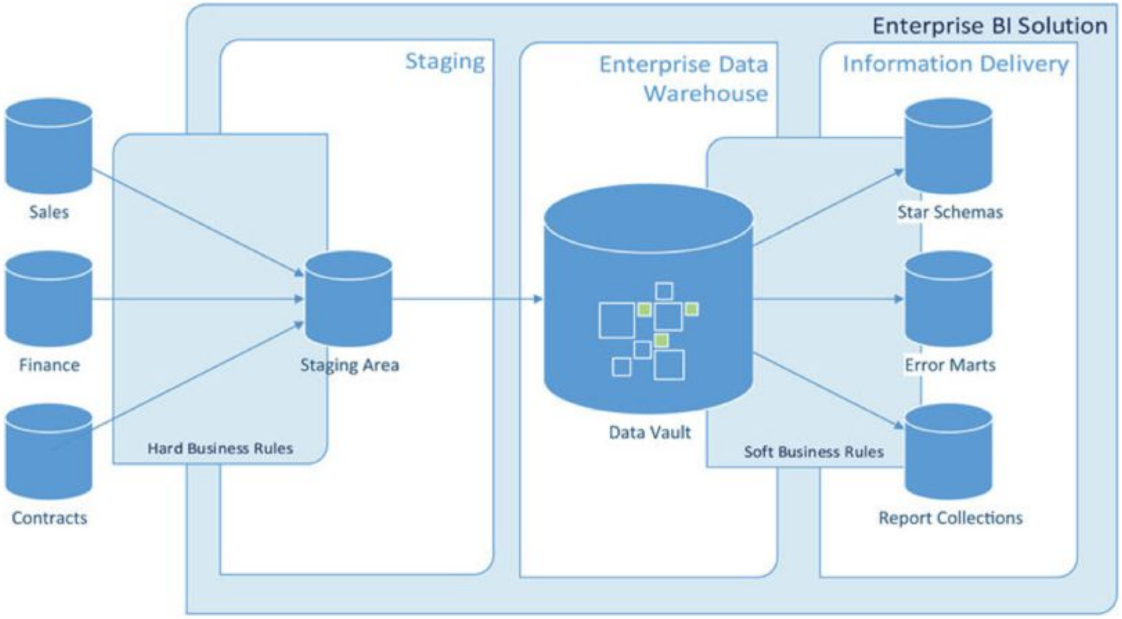
0 - 4: Laying the foundation

- 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*

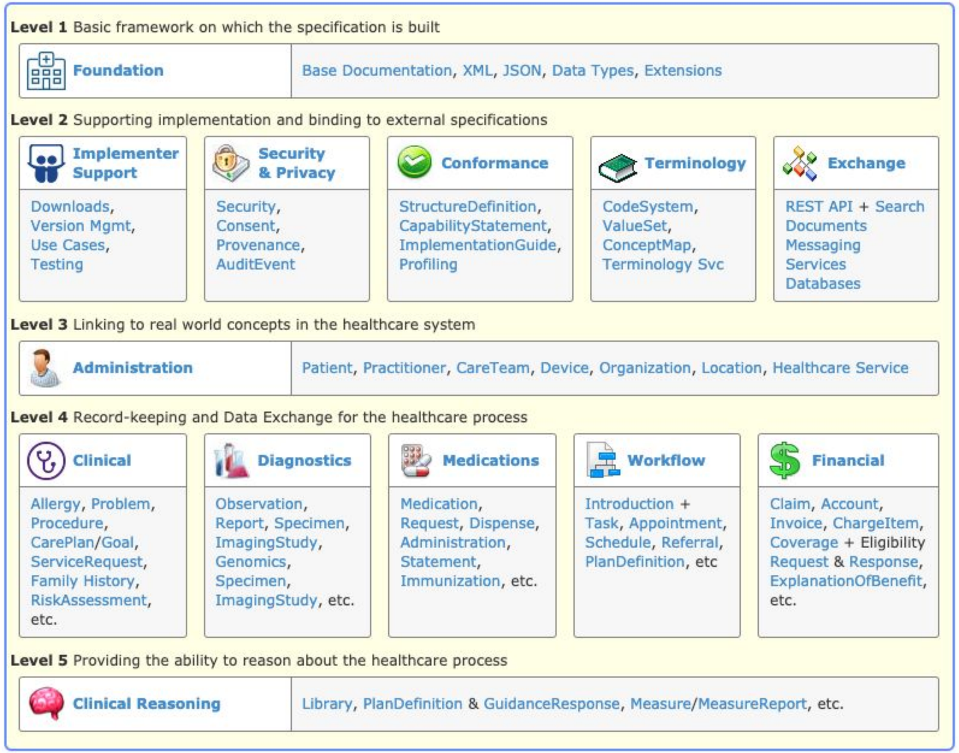
The supplier landscape: choices, choices

- **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:**
  Tableau, PowerBI, Qlik …

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 model



https://hl7.org/fhir

Quiz: how many resources are defined in the FHIR standard?

A. 50

B. 100

C. 150

D. 250

The Dutch equivalent: zorginformatiebouwstenen

Health and care information models prerelease 2019-2

group: Administrative, count: 6

ContactPerson-v3.3	HealthcareProvider-v3.3	Patient-v3.1.1
Encounter-v4.0	HealthProfessional-v3.4	Payer-v3.1

group: Basic elements, count: 1

BasicElements-v1.0.1

group: Clinical context, count: 21

Alert-v4.0	FeedingPatternInfant-v1.0	NutritionAdvice-v3.2	Vaccination-v4.0
AllergyIntolerance-v3.3	FeedingTubeSystem-v3.3	Pregnancy-v3.1.1	VisualFunction-v3.1
BladderFunction-v3.2	FunctionalOrMentalStatus-v3.1.1	PressureUlcer-v3.3	Wound-v3.2
BowelFunction-v3.1.1	HearingFunction-v3.2	Problem-v4.3	
Burnwound-v3.3	Infusion-v3.3	SkinDisorder-v3.2	
DevelopmentChild-v1.2	MedicalDevice-v3.3	Stoma-v3.2	

group: Measurements, count: 13

BloodPressure-v3.2	FluidBalance-v1.0	LaboratoryTestResult-v4.5	TextResult-v4.3
BodyHeight-v3.1	GeneralMeasurement-v3.0	O2Saturation-v3.1	
BodyTemperature-v3.1.1	HeadCircumference-v1.2	PulseRate-v3.3	
BodyWeight-v3.1	HeartRate-v3.3	Respiration-v3.2	

group: Medication, count: 6

AdministrationAgreement-v1.0.2	MedicationAdministration2-v1.1	MedicationDispense-v2.0.1
DispenseRequest-v1.0.2	MedicationAgreement-v1.1	MedicationUse2-v1.1

group: Partial information models, count: 7

AddressInformation-v1.1	InstructionsForUse-v1.2	PharmaceuticalProduct-v2.1.1	TimeInterval-v1.0
ContactInformation-v1.1.1	NameInformation-v1.0.1	Range-v1.0.1	

group: Patient context, count: 17

AdvanceDirective-v3.1	FamilySituation-v3.2	LegalSituation-v1.0	ParticipationInSociety-v3.1
AlcoholUse-v3.1	FamilySituationChild-v1.2	LifeStance-v3.2	TobaccoUse-v3.2
DrugUse-v3.2	HelpFromOthers-v3.01	LivingSituation-v3.2	
Education-v3.1	IllnessPerception-v3.1	MaritalStatus-v3.1	
FamilyHistory-v3.1	LanguageProficiency-v3.1	Nationality-v3.0	

group: Scales en screening tools, count: 13

ApgarScore-v1.0	DOSScore-v1.0	PainScore-v3.2	StrongKidsScore-v1.1
BarthelADLIndex-v3.1	FLACCpainScale-v1.1	SNAQ65+Score-v1.2	
ChecklistPainBehavior-v1.1	GlasgowComaScale-v3.1	SNAQrcScore-v1.1	
ComfortScale-v1.1	MUSTScore-v3.1	SNAQScore-v3.2	

group: Selfcare, count: 10

AbilityToDressOneself-v3.1	AbilityToGroome-v1.0	AbilityToPerformNursingActivities-v1.0	Mobility-v3.2
AbilityToDrink-v3.1	AbilityToManageMedication-v1.0.1	AbilityToUseToilet-v3.1	
AbilityToEat-v3.1	AbilityToPerformMouthcareActivities-v3.1	AbilityToWashOneself-v3.1	

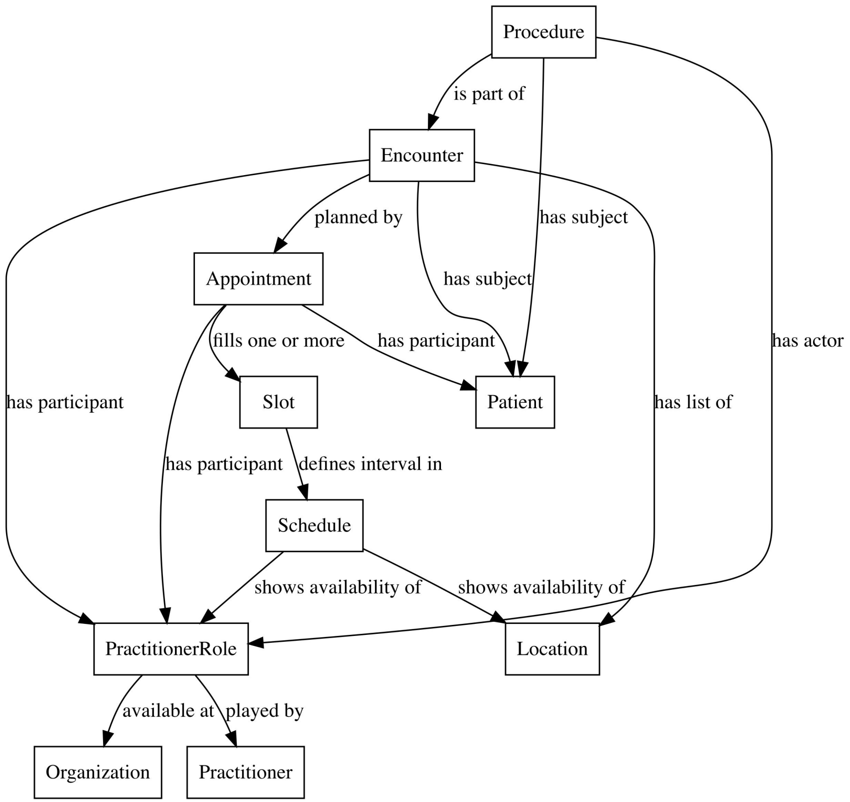
group: Treatment, count: 6

FreedomRestrictingMeasures-v3.2	OutcomeOfCare-v3.1	TreatmentDirective-v3.2
NursingIntervention-v3.2	Procedure-v5.1	TreatmentObjective-v3.1

https://zibs.nl/wiki/HCIM_Release_2019(EN)

- Part of MedMij standard for healthcare information exchange in the Netherlands
- Currently being implemented for personalized healthcare portals (PGOs):
  - VIPP: hospitals
  - OPEN: general practitioners
  - VIPP GGZ: mental care
  - … elderly care, home care to follow?

Example: patient scheduling in specialty clinics

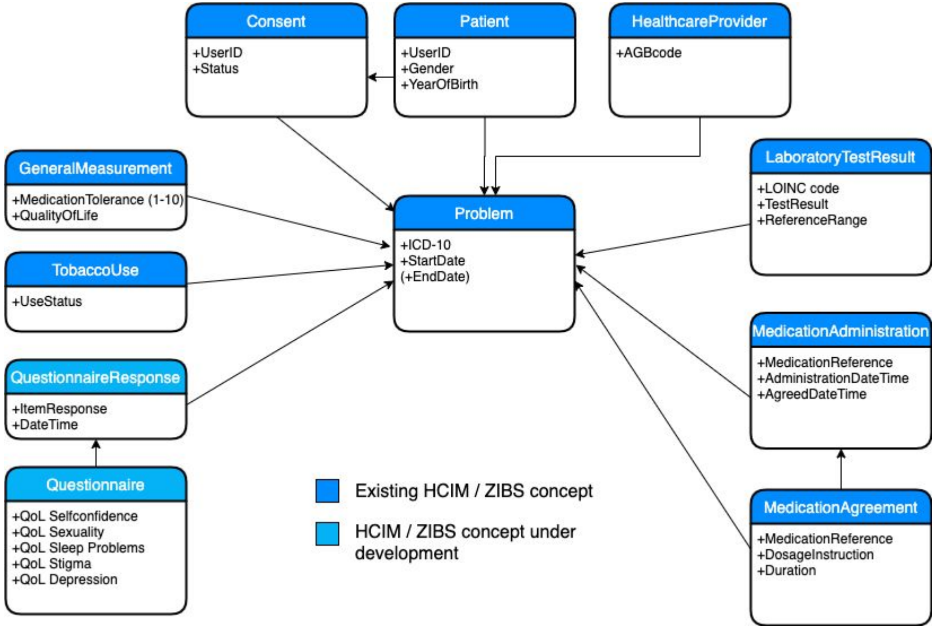


Is it possible to have free-walk in at specialty clinics for ophthalmology?

- Research project at Timeff Labs

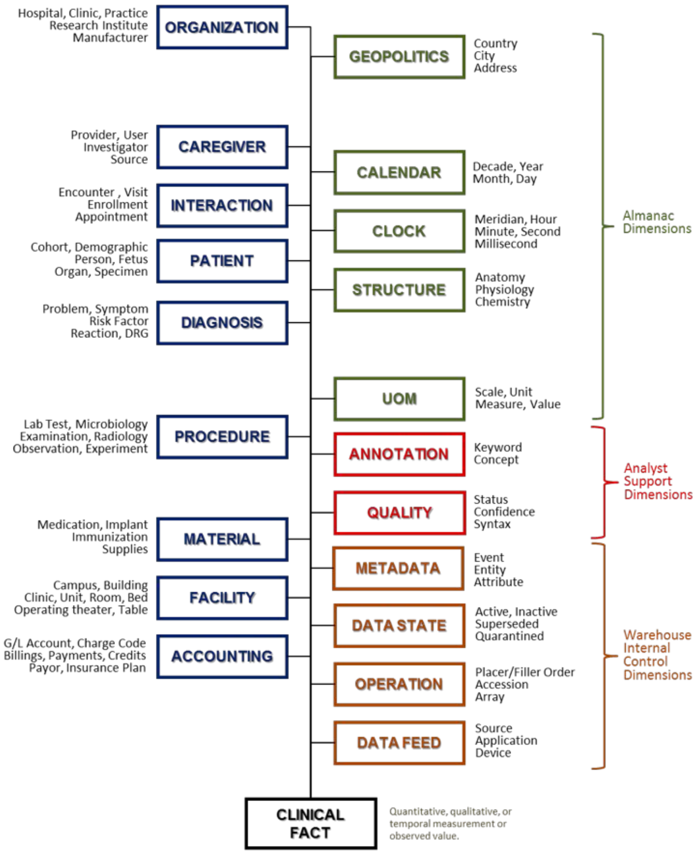
- Thesis Paulien Koeleman (VU 2012)
  A careful solution: patient scheduling in healthcare

Example: the Happi App … and the Happi Datalab



https://happiapp.nl

Harmonize dimensions across facts



- 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

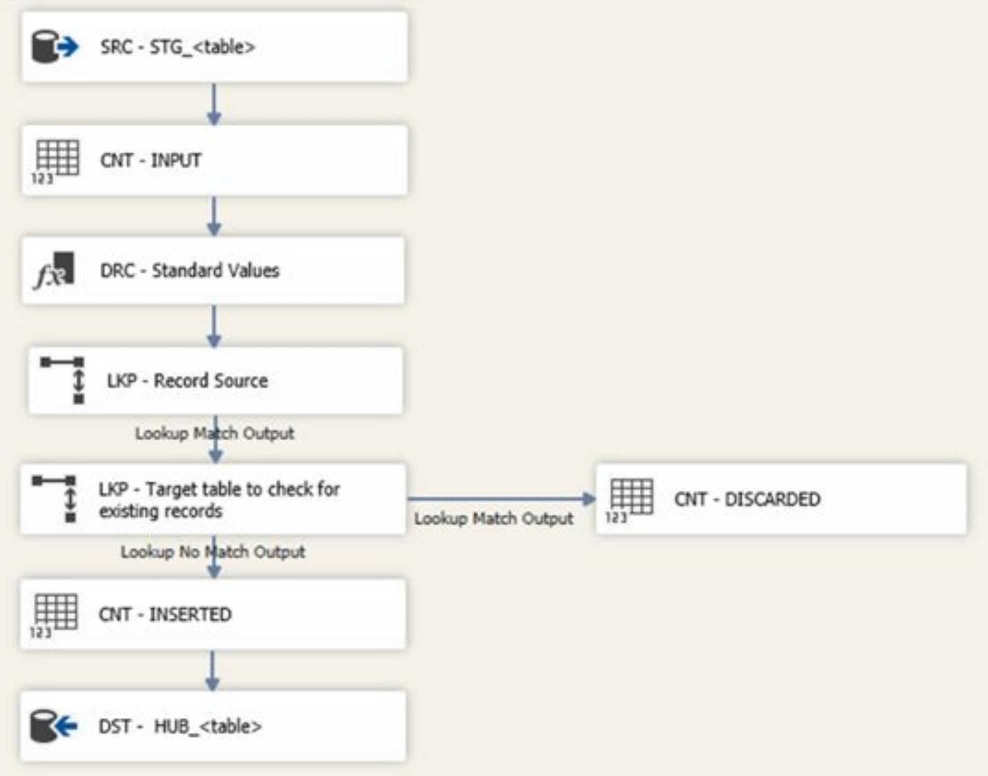
Biehl (2015), *Data Warehousing for Biomedical Informatics*

Choose your way of working

Coding (Apache Beam, Airflow, Prefect)

```python
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)
```

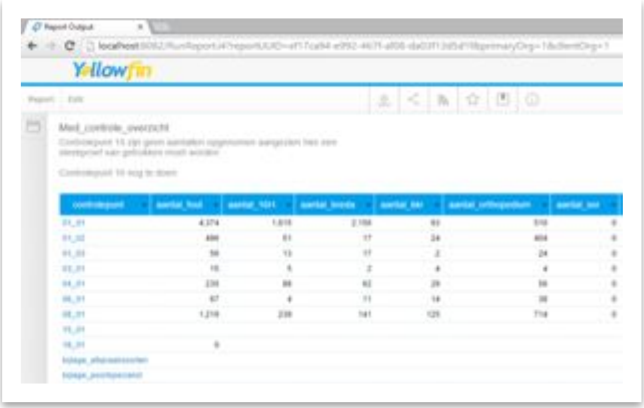
Graphical ETL tool (Microsoft SSIS, Talend)



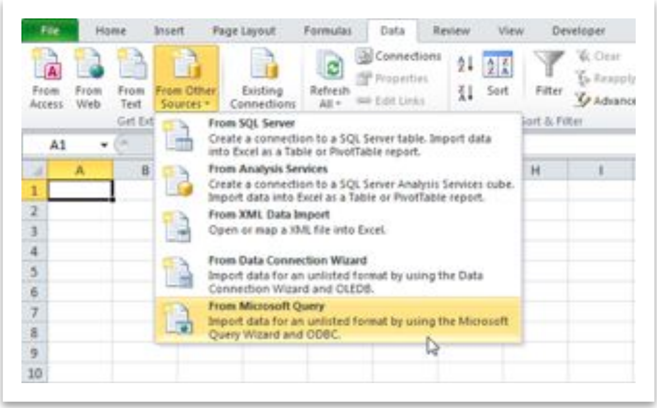
Design data marts for human readability

Let end-users choose their own tool

Dedicated BI tool: Tableau, PowerBI etc.

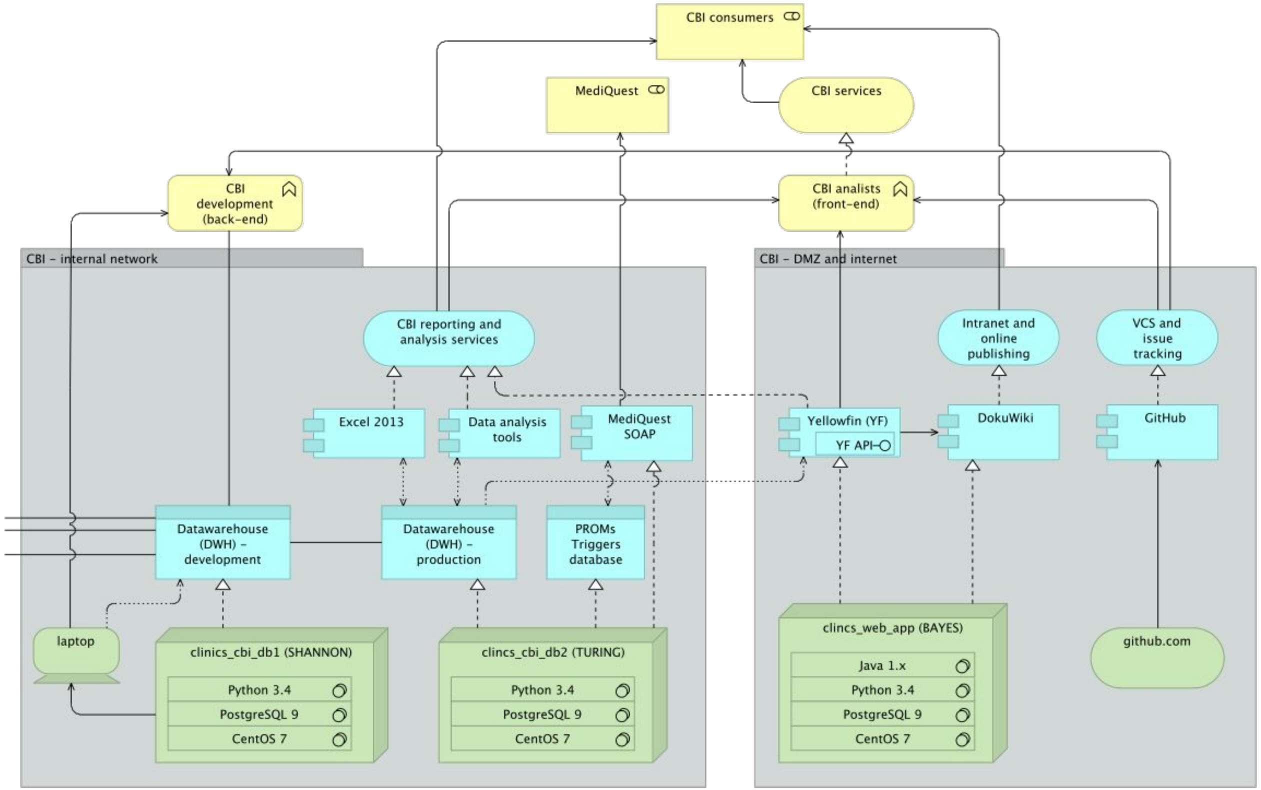


Spreadsheet: Excel, Google Sheets etc.



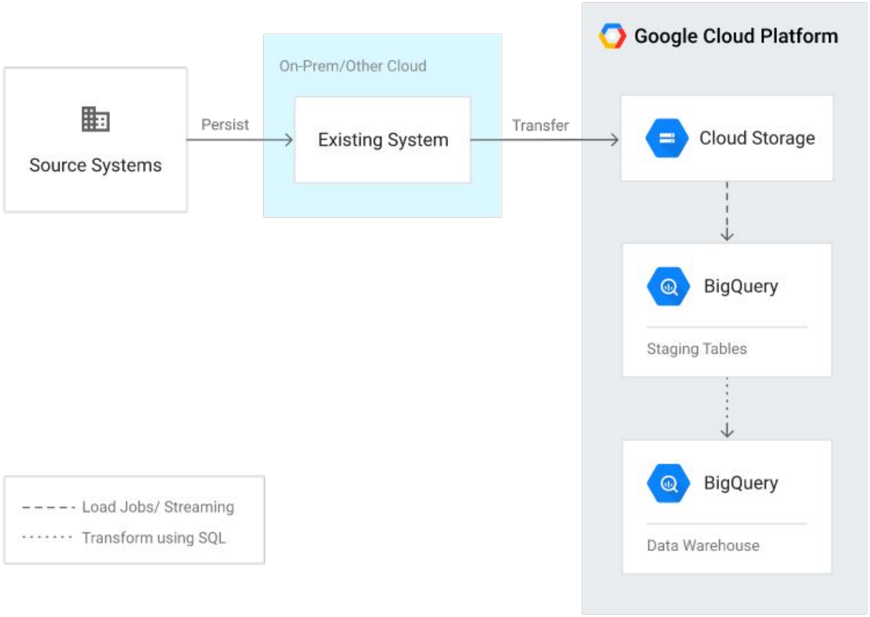
Analytical software (notebooks, SAS, SPSS)
… and lots more

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

End result: on the Google Cloud Platform



Functional data engineering with Google BigQuery and Prefect

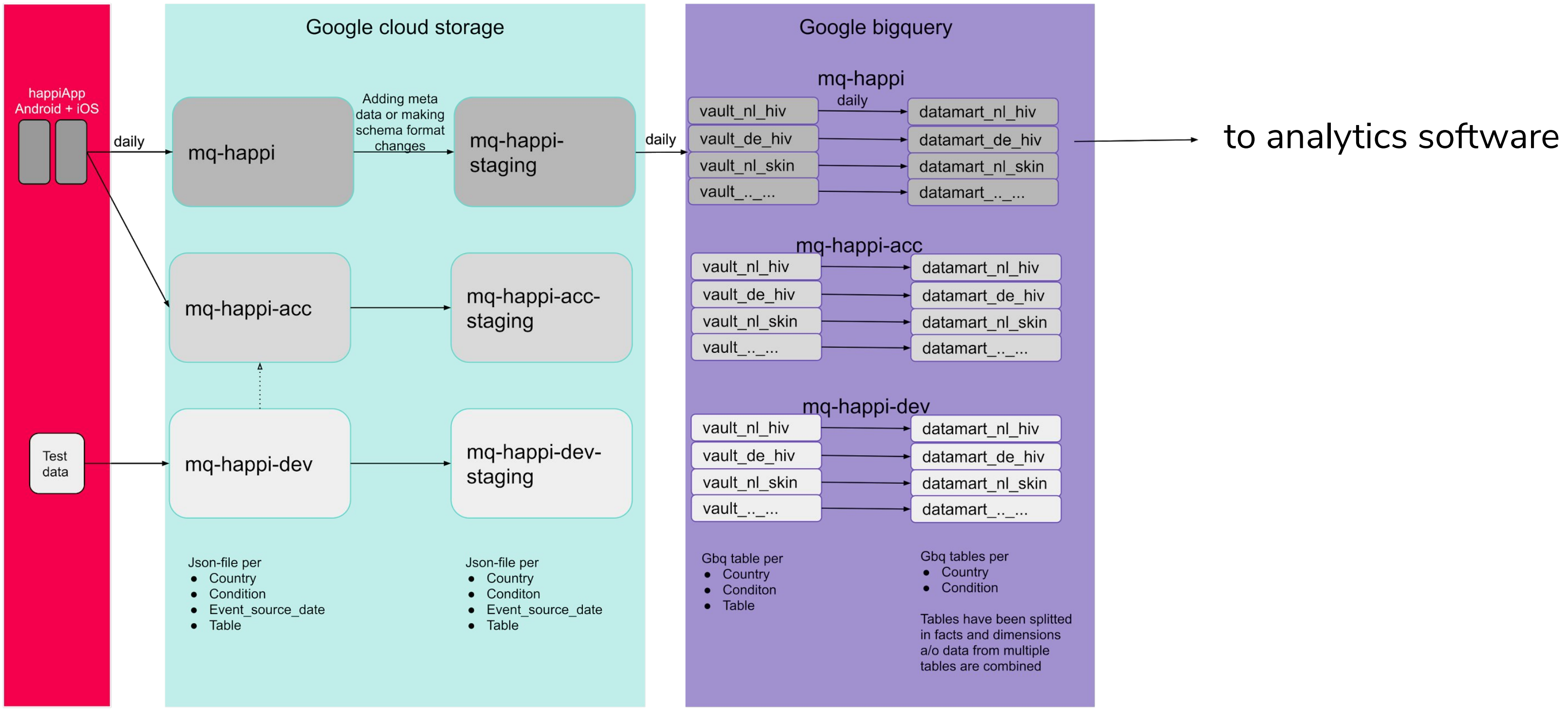
Principles

- use pure tasks in your data pipeline
- regard table partitions as immutable objects
- using a persistent and immutable staging areas

Read more

- Prefect workflow engine
- Work in progress, see recent blog post and nl-open-data project

Example: Happi dataflow

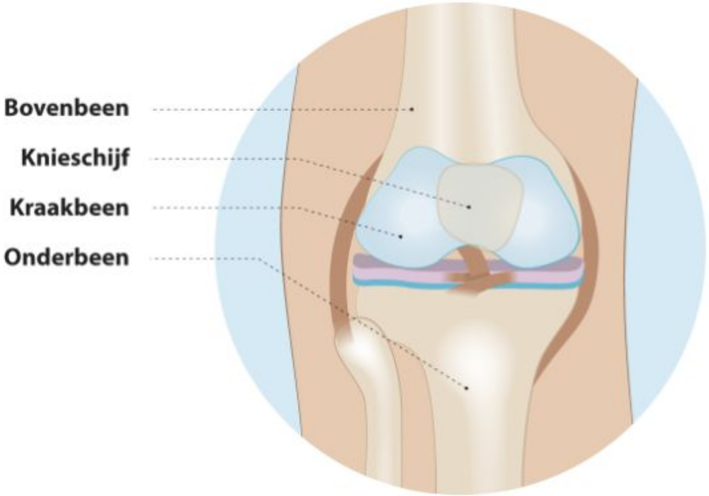


Part II: Healthcare analytics

PREDICTING OUTCOMES

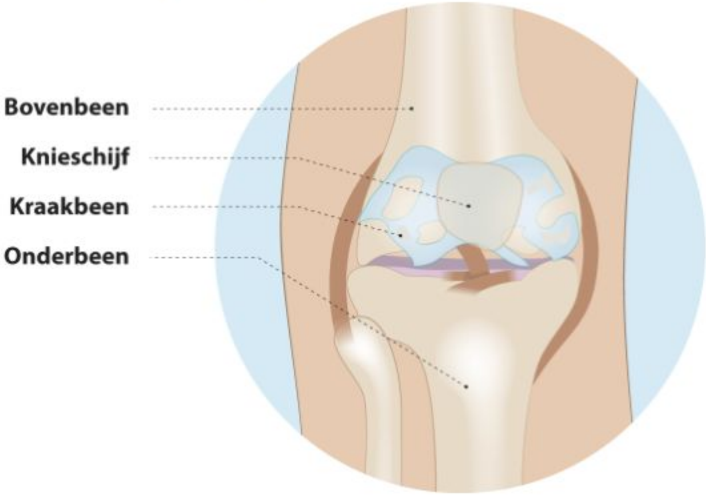
Imagine you are consulting an orthopedic surgeon because of knee osteoarthritis

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.

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.

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

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/

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

Operation:

Your chance of success is also 50%

Conservative treatment:

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

So, what are outcome in healthcare?

Outcome category	Condition: Cataract	Condition: Macular Degeneration	Condition: Low Back Pain	Condition: Hip & Knee Osteoarthiritis
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

*PACs: Potentially Avoidable Complications

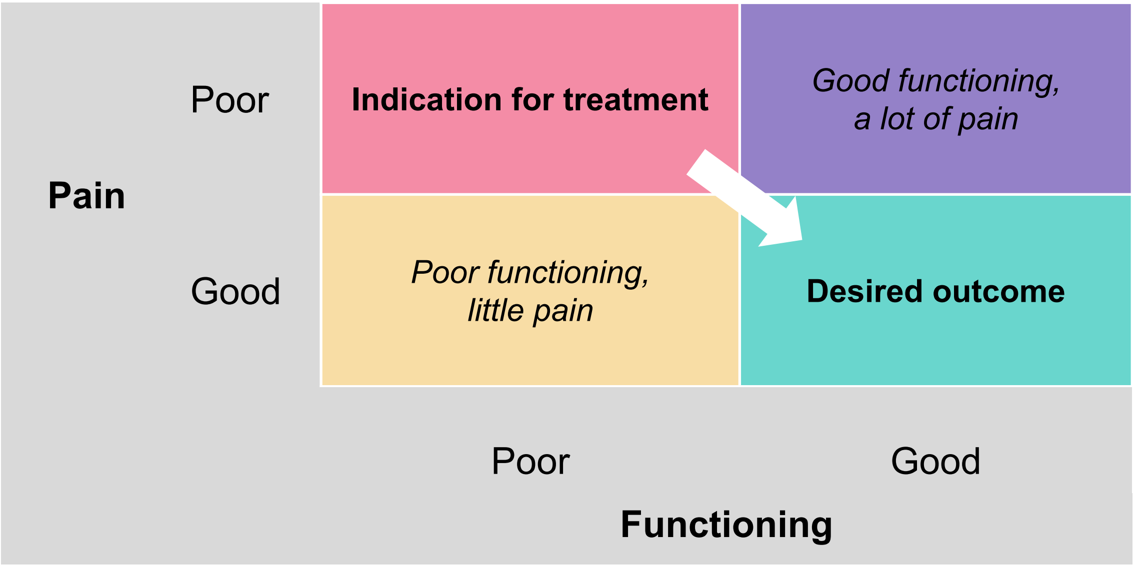
Project Nightingale

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

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



Quiz: what percentage of all knee replacements in the UK have the desired outcome?

A. 50% or less

B. 60%

C. 70%

D. 80% or more

Outcome total knee replacement in the UK

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

Prior to operation

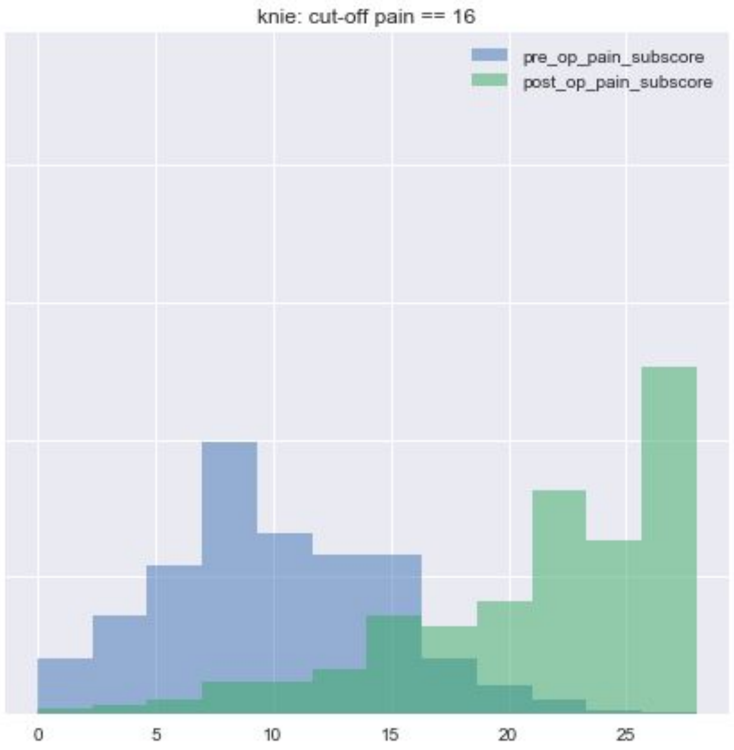
Pre-operative (T0)		Functioning: Poor	Functioning: Good
Pain	Poor	83%	9%
	Good	3%	5%

After operation

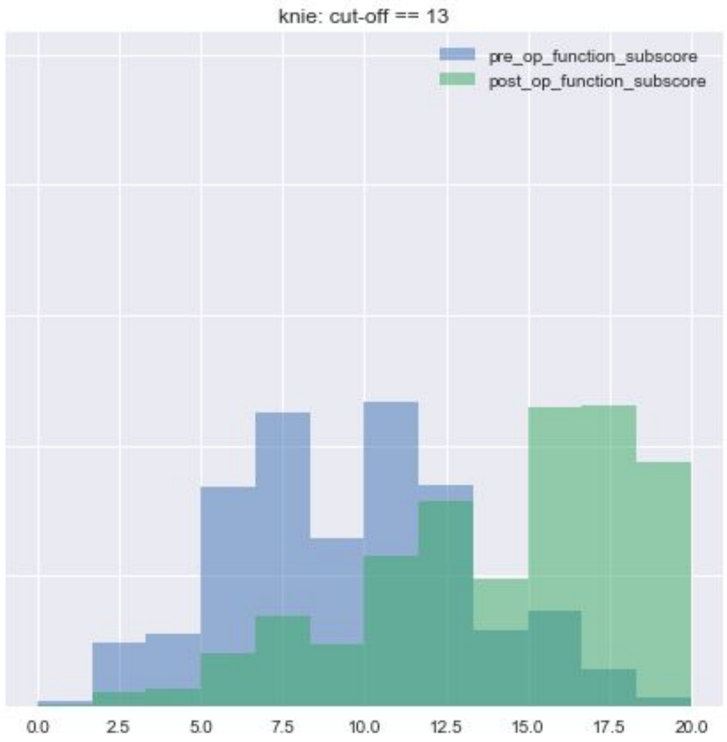
Post-operative (T1)		Functioning: Poor	Functioning: Good
Pain	Poor	18%	1%
	Good	20%	61%

Measuring outcomes is not always straightforward

Pain



Functioning



Prior to cataract surgery

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

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

After cataract surgery

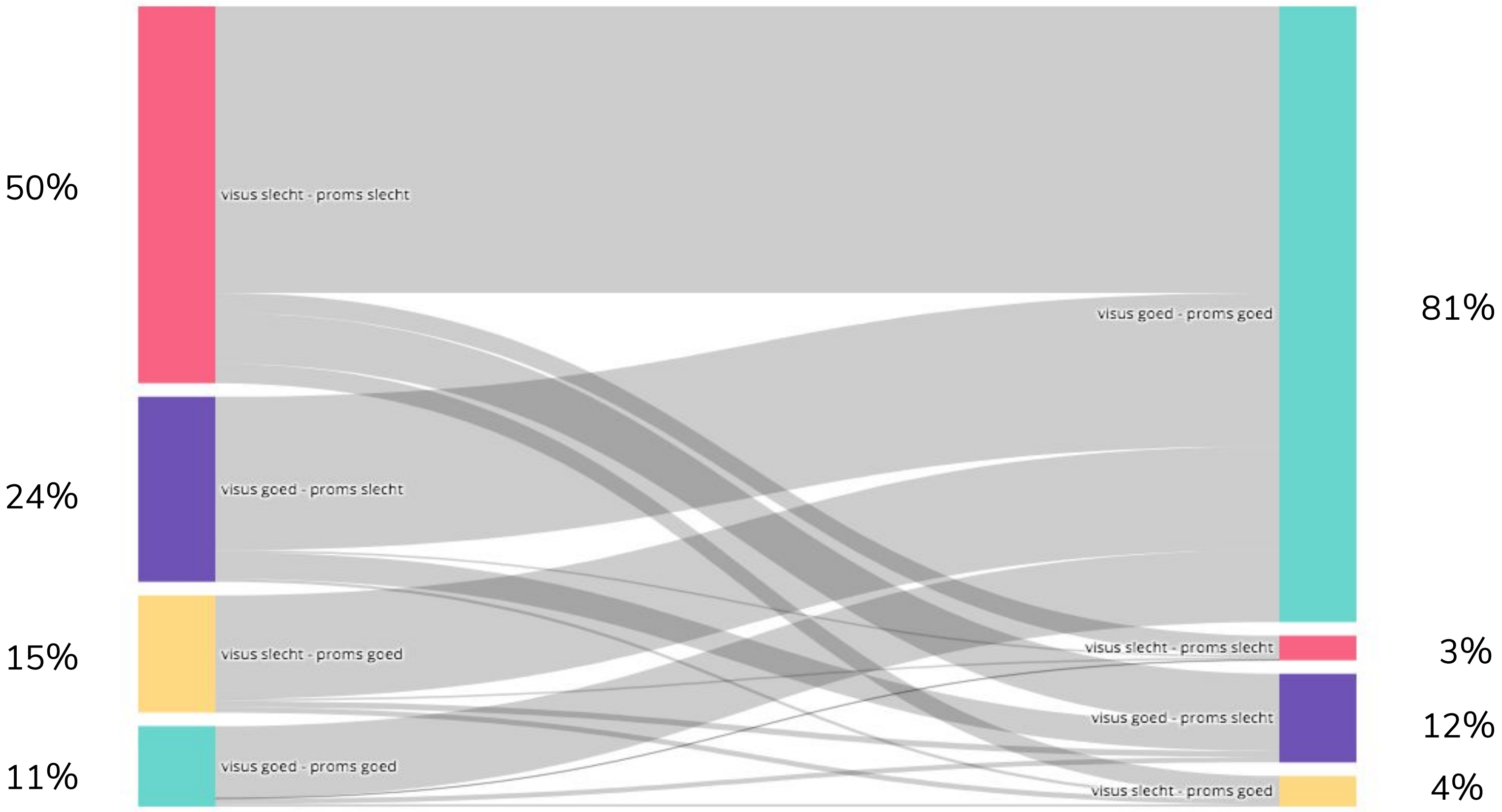
Pre-operative (T0)		Visual acuity: Poor	Visual acuity: Good
PROMs	Poor	50%	24%
	Good	15%	11%

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

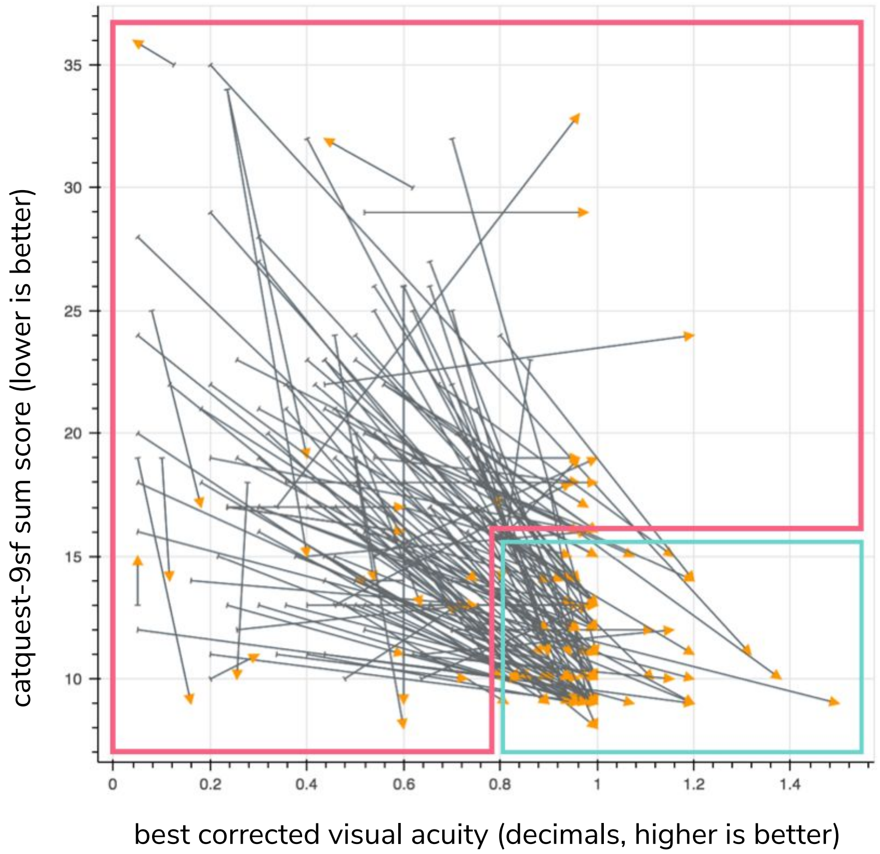
Post-operative (T1)		Visual acuity: Poor	Visual acuity: Good
PROMs	Poor	3%	12%
	Good	4%	**81%**

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

No simple 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?

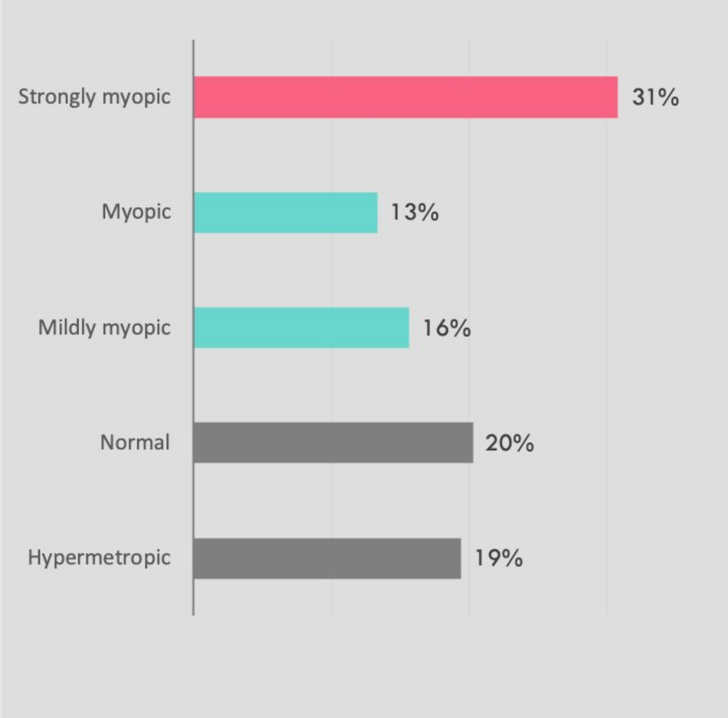
Do we understand what the algorithm does?

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

- Lundström, M. and Stenevi, U., Optometry and Vision Science 2013: vol. 90 no. 8: 754-759
- Grimfors et al., J. Cataract Refract Surg 2014; 40:1163-1169
- Lundström et al., J. Cataract Refract Surg 2013: 39:673-679
- Mollazadegan, K. and Lundström, M., Acta Ophthalmol. 2015: 93: 293-298

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

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

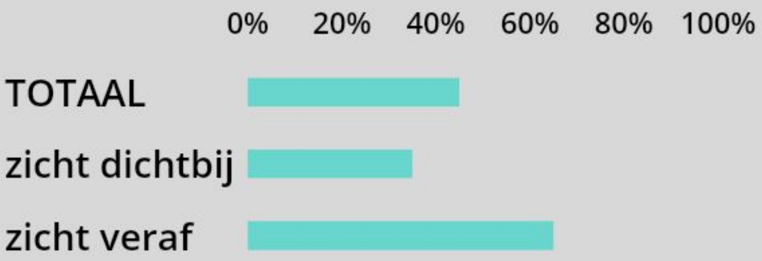


Percentage poor outcome by target refraction, i.e. chosen strength of glasses post-surgery

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)

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

Lessons learned from applying machine learning

- Data quality (registratie aan de bron)
- Harmonisation and semantic integration of different registries
- Open validation trained models for clinical decision support (MRDR)

Questions, more info?

- Drop me an email at
  daniel@kapitan.net

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