



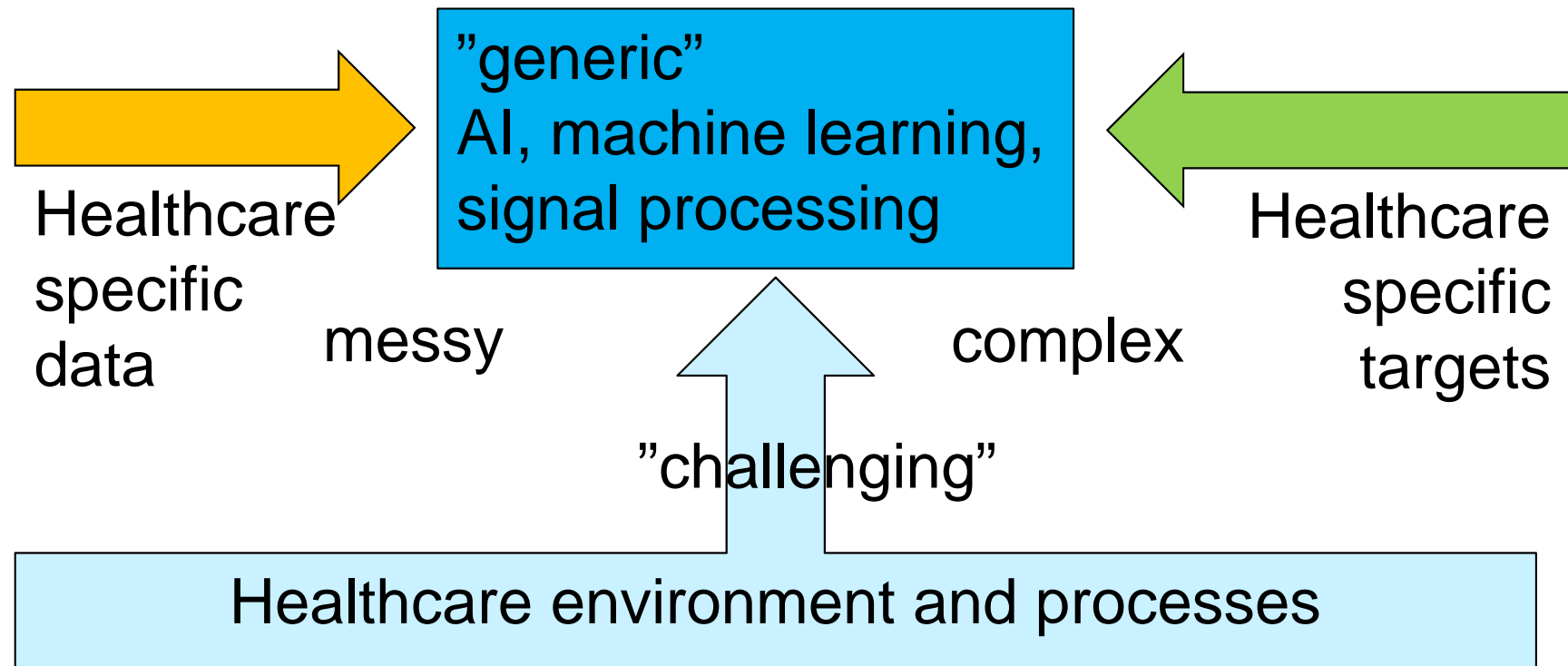
AI for decision support in Health – how to make it work

Mark van Gils, Ph.D., Adjunct Professor
Principal Scientist, Smart Health

mark.vangils@vtt.fi

Goal of this tutorial

- Get an understanding of why AI applied to healthcare is not 'business as usual'
- Get acquainted with the main practical problems and understand how to address them





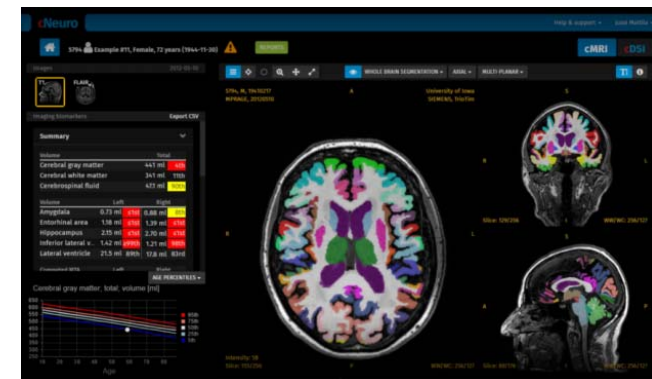
Personalised care: health care treatments to match the unique characteristics of individuals. Genomic and other ‘omics data with other health and behavioral data including data from sensors.

AI is needed to process huge amounts of data.

Automated health data analytics:

Automated analysis of complex health data - imaging, electronic health records, sensor data – reliable quantification and interpretation.

AI to pre-process, and analyse data.



Continuous citizen-centric care: Improve continuous preventive management of health of individuals by automatically monitoring and integrating information.

Use AI to analyse, interpret changes in health status. Engage. Motivate.

The Data – measured from humans

- Every person is different (their own reference) [**inter-subject** variability]
 - What is a low heart rate for me, could be normal for you

- Your personal 'normal' may change over time [**intra-subject** variability]
 - My blood pressure today may be higher than yesterday, it doesn't mean I suddenly need to see a doctor (nervousness for giving a presentation? coffee usage? Short night sleep?)

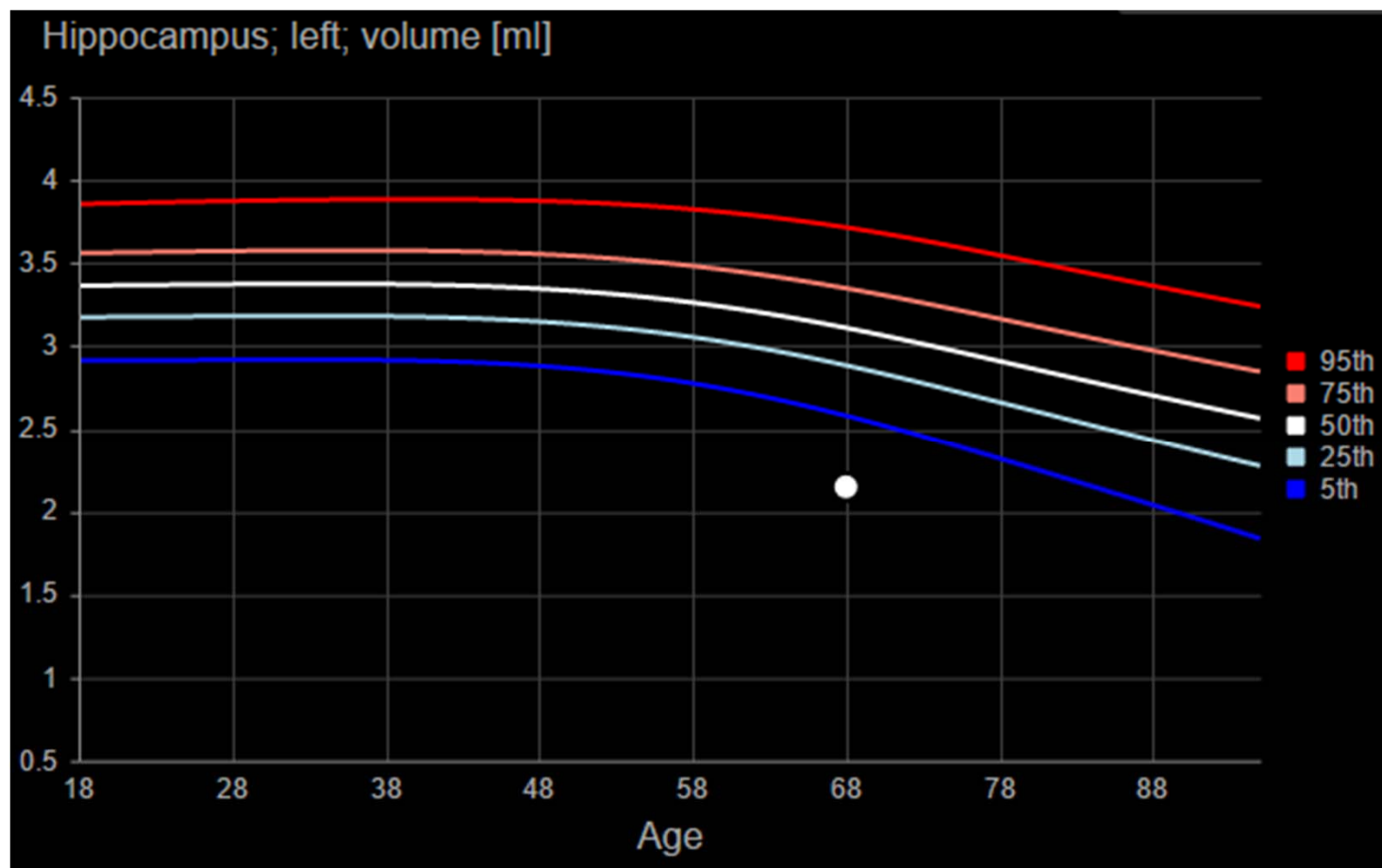
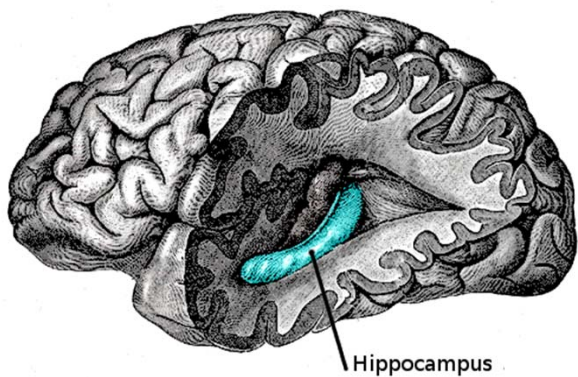
- How to deal with these variations in decision making?

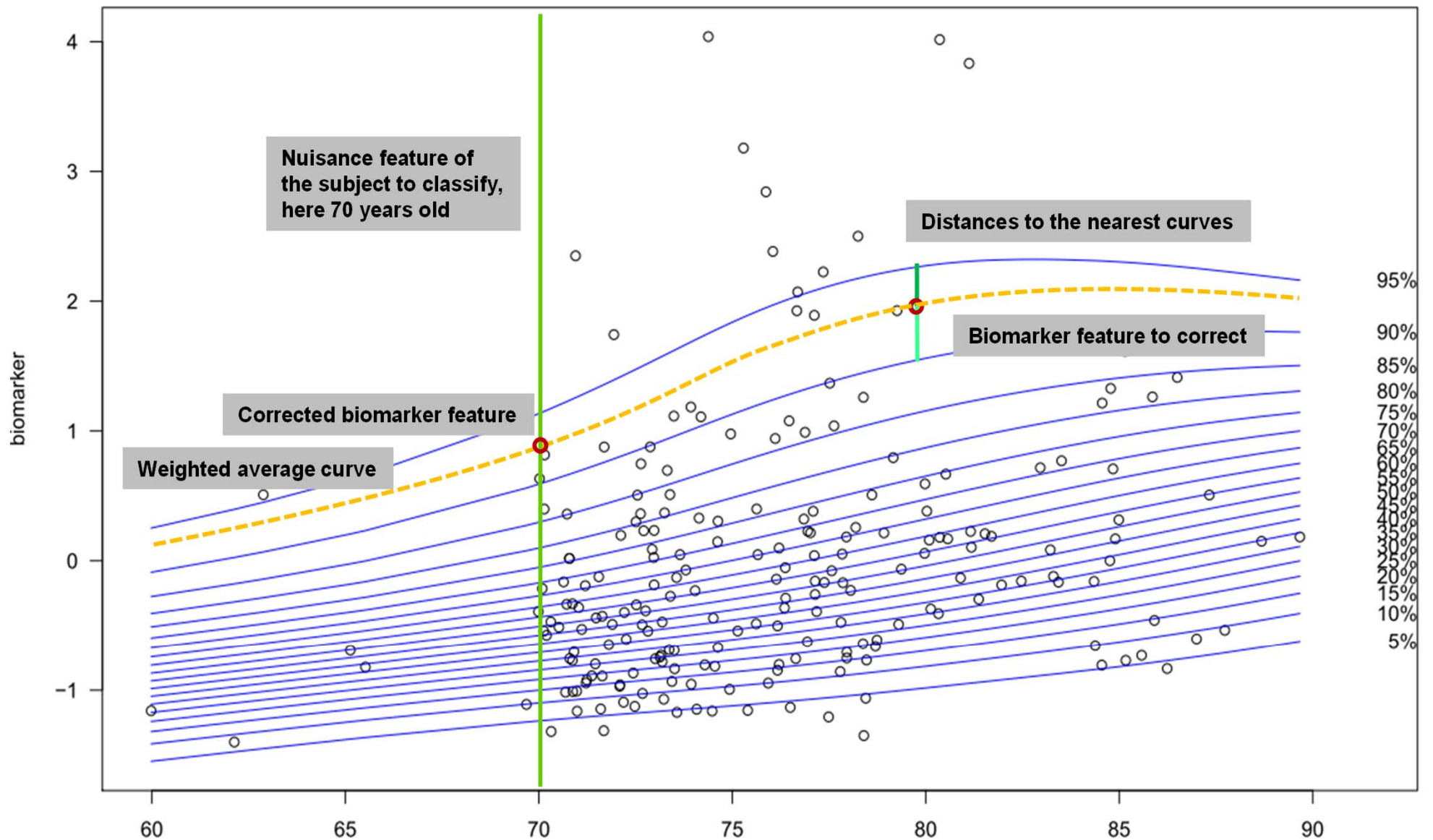
Strategy A – Just use all data as one batch

- Ignore any individual variation, just add all the data in one big batch
- Motto: "The AI will sort it out itself if we have enough data"
- Might work – IF we have lots of representative data. Not optimal though (and difficult to sell to clients (clinicians))

Strategy B – personalisation, stratification

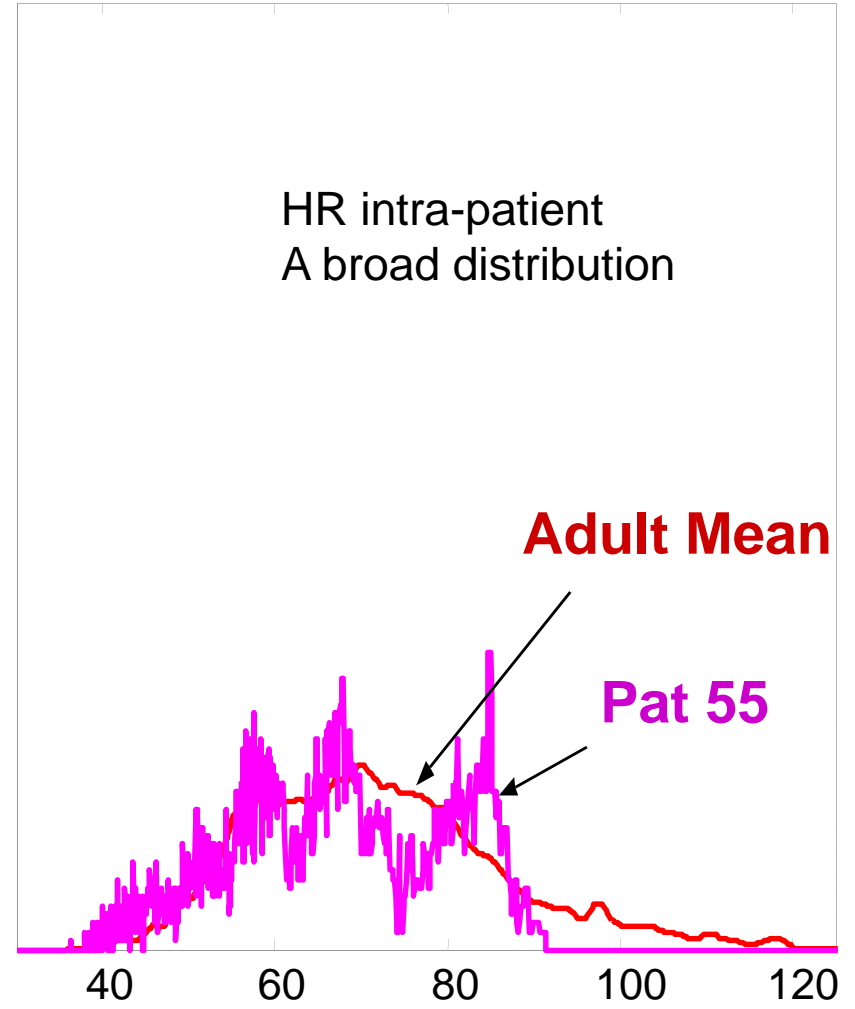
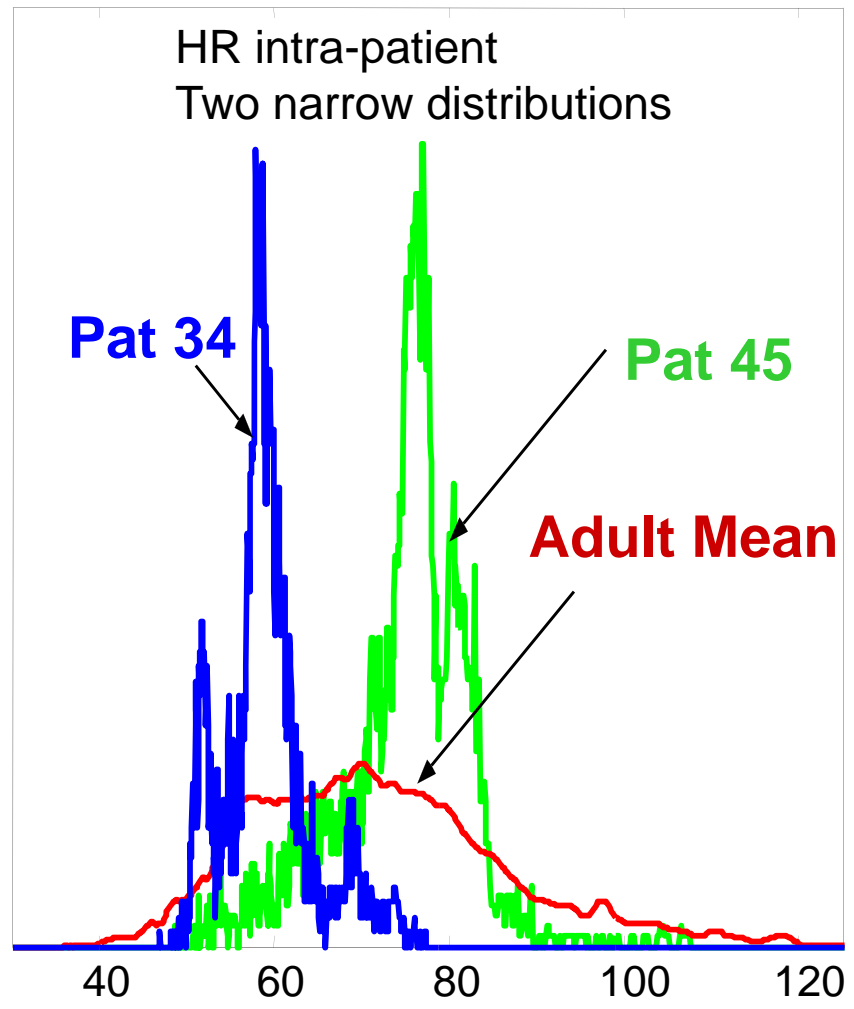
- Use statistical methods to account for variation in different subjects, e.g. use partial correlations, patient ID's as dummy variable
- Normalise all data to personalised baseline values, MinMax(); StandardScaler - Gaussian with 0 mean and unit variance
- Stratification, binning into groups of similar persons (based on age, gender, education,...)
- Correct for 'nuisance variables' such as age (e.g. age has an effect on brain size, regardless of diseases like dementia)
- Dedicated methods like histogram transform





Adaptive centile correction algorithm. Each feature is remapped according to the feature-wise centile curve (yellow) to the age of the subject (green vertical line). The feature-wise centile is computed as a weighted average from the nearest centile curves acquired with LMS method with Yeo-Johnson power transformation.

Histogram transformation example: variability of heart rate between patients



How do we create an individual normalisation transformation?

estimate group distribution

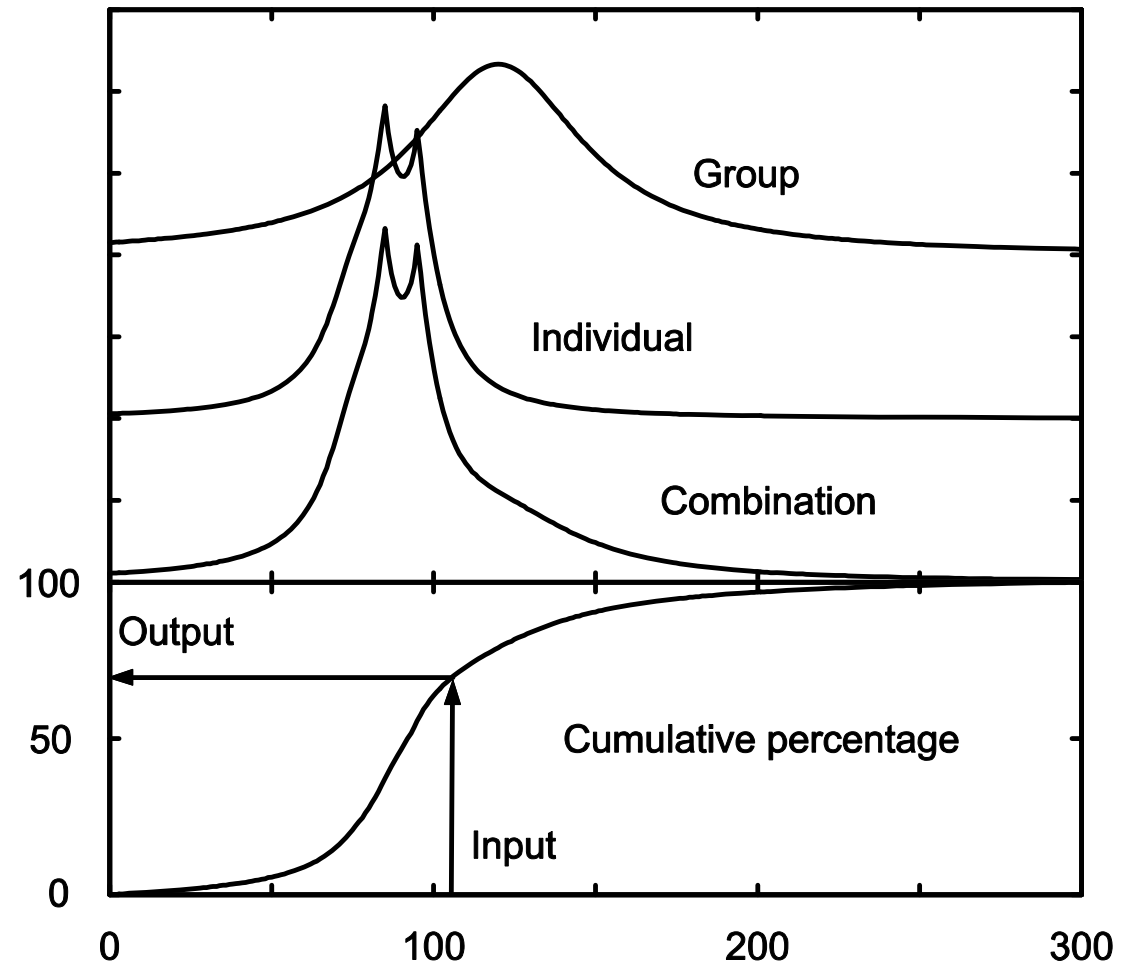
estimate individual distribution
data either

1. since the beginning of recording, or
2. over running time window

calculate combined distribution
 $A * \text{group} + B * \text{individual}$
(eg $A=0.7$, $B=0.3$)

calculate cumulative combined distribution

>> histogram transformation



Missing, incomplete data

- In multi-variate/multi-modal data sets
- In time series

- Most classifiers don't work correctly when data input vectors have empty elements or NaNs – what to do?


Dealing with Missing data

- In the majority of applications where we combine lots of modalities (images, -omics, text, patient monitors), some elements are missing from the input vector
- Only use complete cases?
 - Leaves only a fraction of the data, we throw away too much valuable data
- Impute all missing data? With averages, interpolations etc.
 - May work sometimes, but defeats eg the purpose of clustering
- **Compromise: Only use cases that have at least eg 70% of all elements present, impute the rest**
- Consider data analysis methods that may work with incomplete data, eg c-means clustering, DSI*

Missing data in time series

- Imputation (splines, regressions etc), within reason (and in agreement with domain experts).
A curve that is visually pleasing is not necessarily clinically valid.
- **BUT NOTE:** the fact that there is missing data is information in itself
- For example, decreased frequencies of self-measurement may indicate changes in the person's state

Are Breaks in Daily Self-Weighing Associated with Weight Gain?

Elina E. Helander, Anna-Leena Vuorinen , Brian Wansink, Ilkka K. J. Korhonen

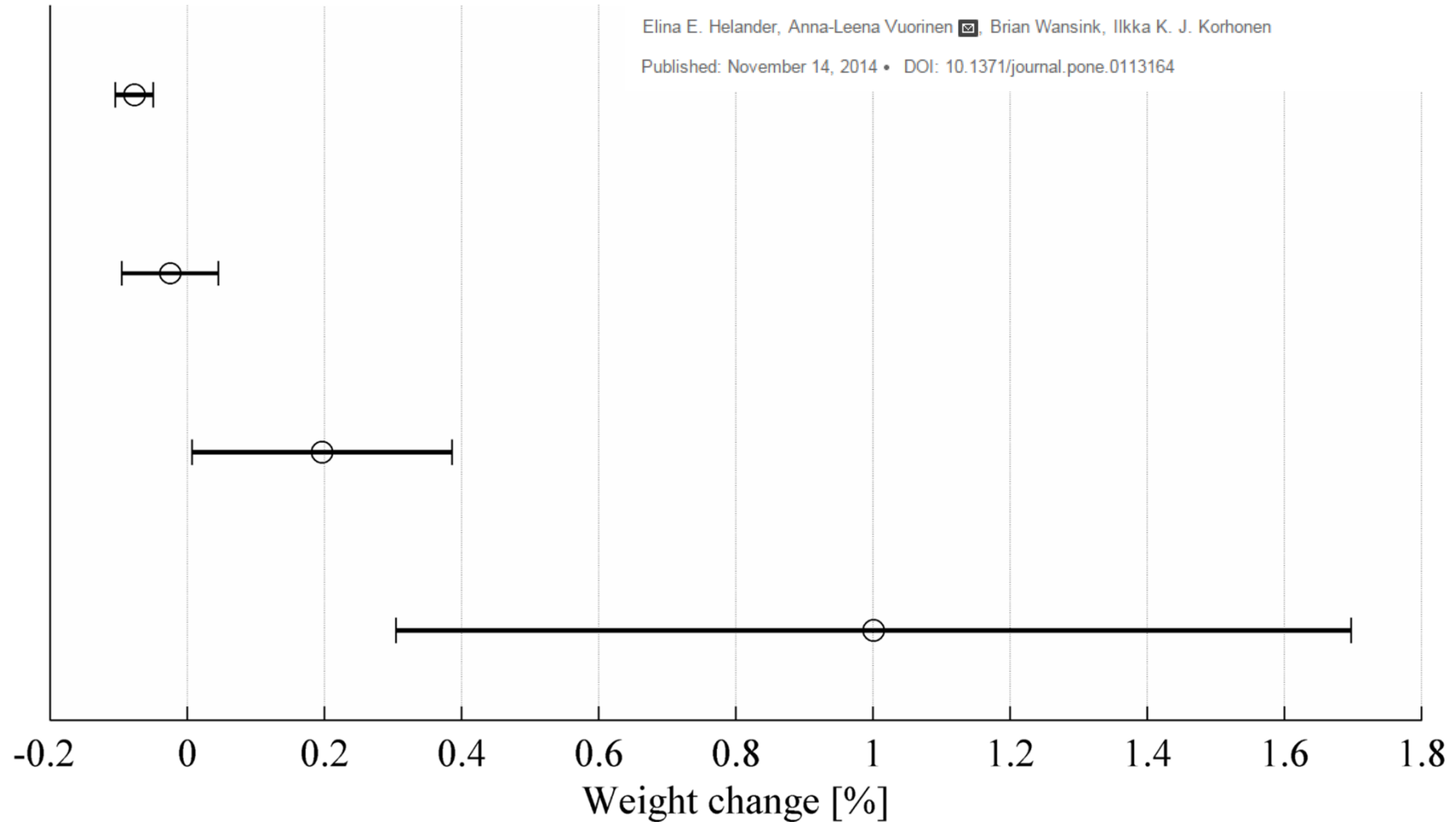
Published: November 14, 2014 • DOI: 10.1371/journal.pone.0113164

Daily
 $n=1951$
 $N=37$

At least
weekly
 $n=637$
 $N=38$

At least
monthly
 $n=163$
 $N=33$

Less than
monthly
 $n=47$
 $N=23$

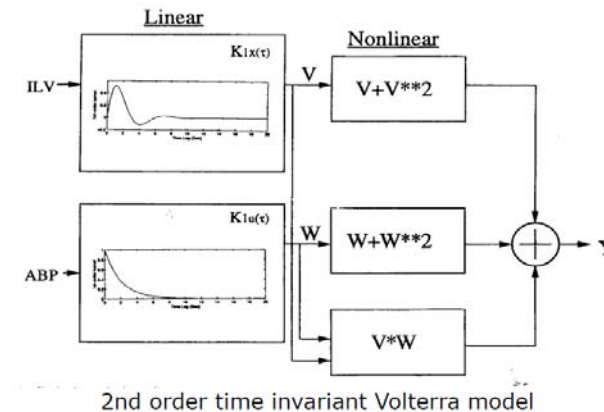


Weight loss took place during periods of daily self-weighing, whereas breaks longer than one month posed a risk of weight gain.

Missing data in weight management studies with a weight-monitoring component may be associated with non-adherence to the weight loss programme and an early sign of weight gain

From model-based to black-box approaches

- Rule-based approaches
- Mathematical models
- Physiological models
- System dynamic models
- Probabilistic models
- Syntactic pattern recognition
- Decision trees
- ...
- Black-box approaches

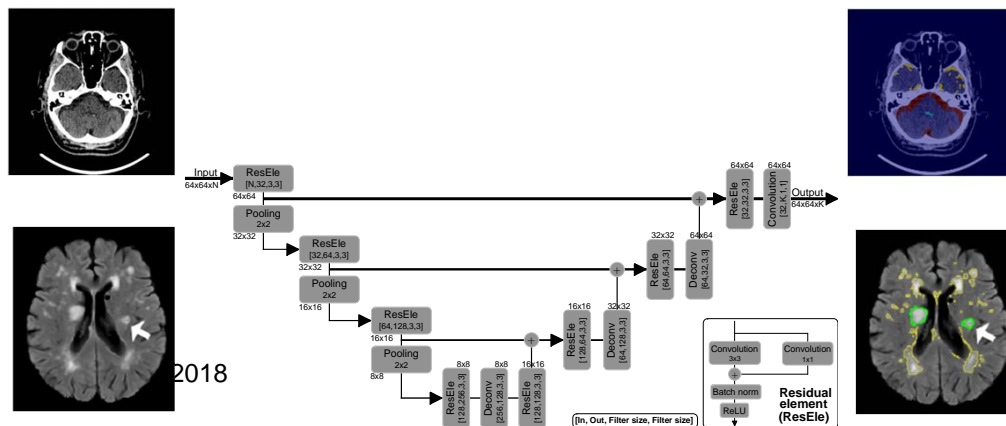


		LIDO										MPET						
Case #	MRI (T1w)	DTI	Perm. map	Arterial flow	Arterial flow	Aur	Apr	Hand	Wheel	Stair	ICP	ICP	CSF	Perfu. map	Perfu. map	δ	δ	
24						M	66	67	No	> 1 hr	0.75							
05						F	73	88	Daily	N/A	0.82							
64						F	77	80	No	N/A	N/A							
09						F	81	85	No	> 1 hr	0.82							
06						F	82	85	Daily	> 1 hr	0.42							
72						M	66	80	No	< 15 min	N/A							
30						F	73	80	No	> 1 hr	0.3							
07						F	77	73	No	30 - 50 min	0.22							
12						F	81	80	No	15 - 20 min	N/A							
44						F	82	82	No	> 1 hr	N/A							

= High Activity/Low Activity
 = Daily Leisure
 = ICP
 = CSF clearance
 = Displacement

Black-box approaches?

- Traditionally, the argument has been that if the functioning of an algorithm is not explainable it cannot be used for making decision support in healthcare
- Explainability does improve confidence in decision making, and is **HIGHLY VALUED** by healthcare professionals
- But, its strict necessity may be decreasing for certain tasks, especially the more routine subtasks
 - E.g. a CNN that segments images very well does not necessarily need explainability



Arterys Receives First FDA Clearance for Broad Oncology Imaging Suite with Deep Learning



FDA clearance covers all solid tumors. Initial launch will include Liver AI and Lung AI oncology software to empower clinicians to quickly measure and track lesions and nodules in MRI and CT scans

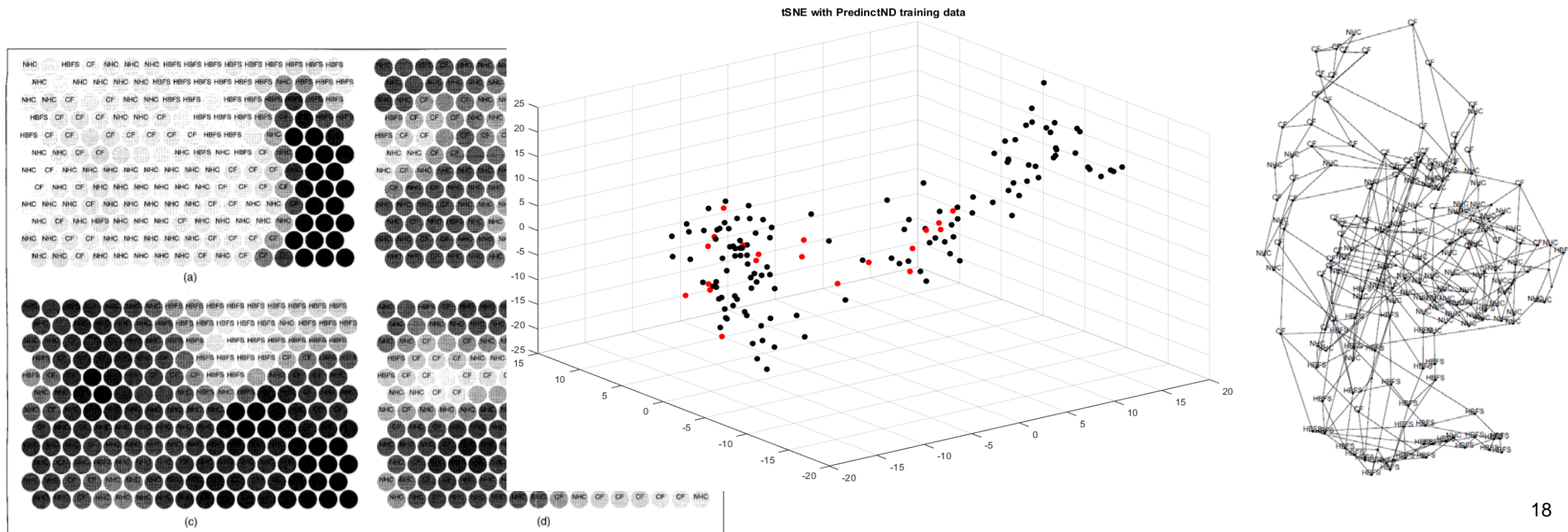
NEWS PROVIDED BY
Arterys Inc. →
Feb 15, 2018, 08:07 ET



SAN FRANCISCO, Feb. 15, 2018 /PRNewswire/ -- Arterys Inc., the leader in intelligent, cloud-based medical imaging

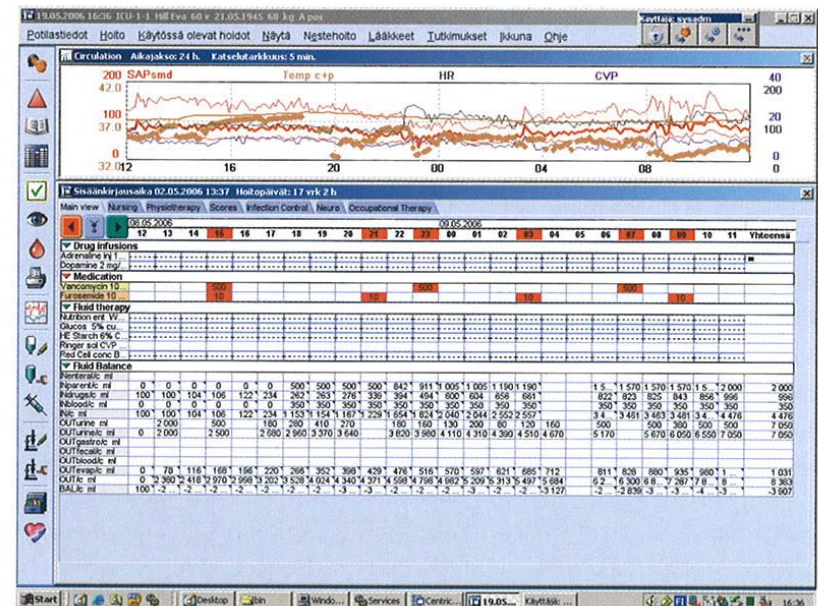
Results presentation and interaction

- Patient data can be very high-dimensional
- We as data scientists love to find clusters and visualise interesting structures in the data
- Tools: PCA plots, t-SNE, SOMs, Sammon mappings,



Clinicians' preferences

- Overall, high-dimensional visualisations and (non-)linear mappings are difficult concepts to communicate to non-data scientists
- In addition, clinicians are often conservative, they prefer
 - Raw data, time series – old fashioned timelines that they are used to seeing
 - Quick overview + details on demand



Clinical Platform for decision support in dementia



Back to Patient Selection

Layout: Entry Preview Fingerprint Analysis

Patient Overview Entry Analysis

Patient

Imagets processed Study Id: Vtt5003

Identifier: 20141_2011-03-29_5009

Name: 20141_2011-03-29_5009

Gender: Female

Baseline age: 76

Entries

Imaging

- MRI Manual Medical tempor...
March 4, 2015

Neuropsychology

TMT A 75 s March 4, 2015	MMSE Total Score 25 March 4, 2015	VAT Naming set A 2 March 4, 2015	CDR Global Score 1 March 4, 2015	RAVLT Learning 31 March 4, 2015
--------------------------------	-----------------------------------------	----------------------------------------	----------------------------------------	---------------------------------------

Animal Fluency Score
10
March 4, 2015

Background

Demographic Years of Educa... 13 March 4, 2015	ADL DAD Total Score 90 March 4, 2015	GDS Total Score 1 March 4, 2015
------------------------------------------------------	--------------------------------------------	---------------------------------------

CSF

Amyloid beta...
479
March 4, 2015

Multiclass classification statistics

Disease State Index

Total: 0.92 ± 0.05 Acc: 0.91 Sens: 0.91 Spec: 0.91 Rel: 0.82

Influence

Chart Overlays: Confidence Training Data Fitness Probability

Select Diseases to Compare

Etiology: AD: 0.76 FTD: 0.68 VaD: 0.51 LBD: 0.51 SMC: 0.09

Progression: AD AD AD AD

Compare to: SMC FTD VaD LBD

Disease State Fingerprint

- Total: 0.92 ± 0.05
- Neuropsychological tests: 0.77 ± 0.08
- RAVLT: 0.25 ± 0.14
- MMSE: 25 ± 0.02
- Visual association test: 0.99 ± 2.00
- Category fluency: 1.00 ± 1.00
- Trail making test: 0.90 ± 0.30
- Cerebrospinal fluid: 0.80 ± 0.01
- Amyloid beta 1-42: 479 ± 0.30
- Total tau: 1059 ± 0.99
- Tau phosphorylated at threonine-181: 56 ± 0.44
- Background: 0.83 ± 0.02
- Clinical Dementia Rating: 0.90 ± 0.90
- Activities of daily living: 0.76 ± 0.76
- Geriatric Depression Scale: 1 ± 0.72
- Structural MRI (T1): 0.71 ± 0.07
- VBM: 1.00 ± 2.00
- Volumes: 0.17 ± 0.15
- TBM: 0.98 ± 0.30
- Manual MRI analysis: 0.92 ± 0.99
- Medial temporal lobe atrophy left: 2 ± 0.94
- Medial temporal lobe atrophy right: 3 ± 1.00
- Global cortical atrophy: 2 ± 0.94
- White matter hyperintensities: 1 ± 0.59
- Vascular burden MRI (FLAIR): 0.88 ± 0.02
- WMH: 0.88 ± 0.02

Assessment of performance

- We need to measure how well our AI method does

- Commonly used measures:
 - Accuracy, sensitivity, specificity, positive prediction probability (PPV), negative prediction probability (NPV)

 - Precision, Recall, F-score (=harmonic mean of precision and recall)

 - ROC curve and AUC-ROC

*Note: Precision and recall are related to PPV/NPV and sensitivity respectively if we have a 2-class problem

Alarm overload



1. Alarm Hazards **& 2013** **& 2012**
2. Infusion Pump Medication Errors
3. CT Radiation Exposure in Pediatric Patients
4. Data Integrity Failures in EHRs and other Health IT Systems
5. Occupational Radiation Hazards in Hybrid ORs
6. Inadequate Reprocessing of Endoscopes and Surgical Instruments
7. Neglecting Change Management for Networked Devices and Systems
8. Risks to Pediatric Patients from "Adult" Technologies
9. Robotic Surgery Complications due to Insufficient Training
10. Retained Devices and Unretrieved Fragments

	Clinical Priority	Overall rate (/hr)	Relevant Alarms %	False Alarms %	Nuisance Alarms %
ECG	1	0.7	9	41	50
SpO2	2	0.9	5	63	32
InvPress	3	0.6	20	27	54
ImpResp	4	0.4	17	45	37
Other mon	5	0.2	19	22	59



Source: Görges, et al. Improving alarm performance in the medical Intensive care unit using delays and clinical context. Anesth Analg 2009. See Table 1.

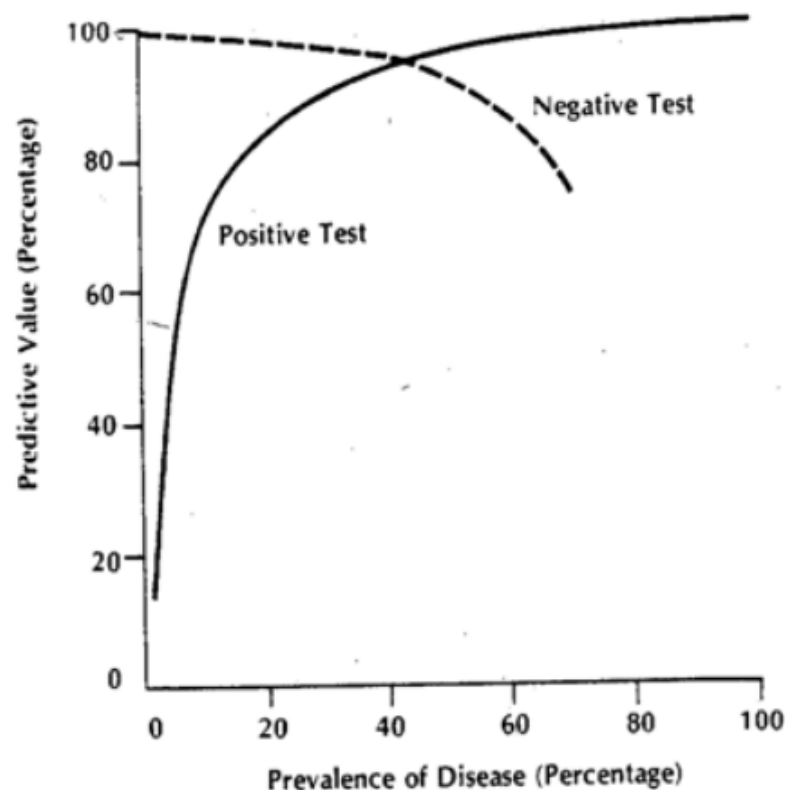
Erno Muuranto,
GE Healthcare

Prevalences (percentage of disease cases in population)

- Health data are often highly imbalanced. With the number of disease cases (luckily) much smaller than the number of healthy cases.
- PPV (or, Precision) is especially intuitive for health tests: "If the test says that I have a disease, what is the probability I actually have that disease"
- `sklearn.metrics.classification_report()` gives precision (PPV) as one of its main outputs
- However, **precision (or, PPV) is dependent on prevalence** (next slide)
- Hence, we often prefer sensitivity, specificity, confusion matrices, ROC curves

Prevalence vs PPV

Using the same test in a population with higher prevalence increases positive predictive value. Conversely, increased prevalence results in decreased negative predictive value.



Relationship between disease prevalence and predictive value in a test with 95% sensitivity and 85% specificity.
 (From Mausner JS, Kramer S: Mausner and Bahn Epidemiology: An Introductory Text. Philadelphia, WB Saunders, 1985, p. 221.)

<https://onlinecourses.science.psu.edu/stat507/node/71>

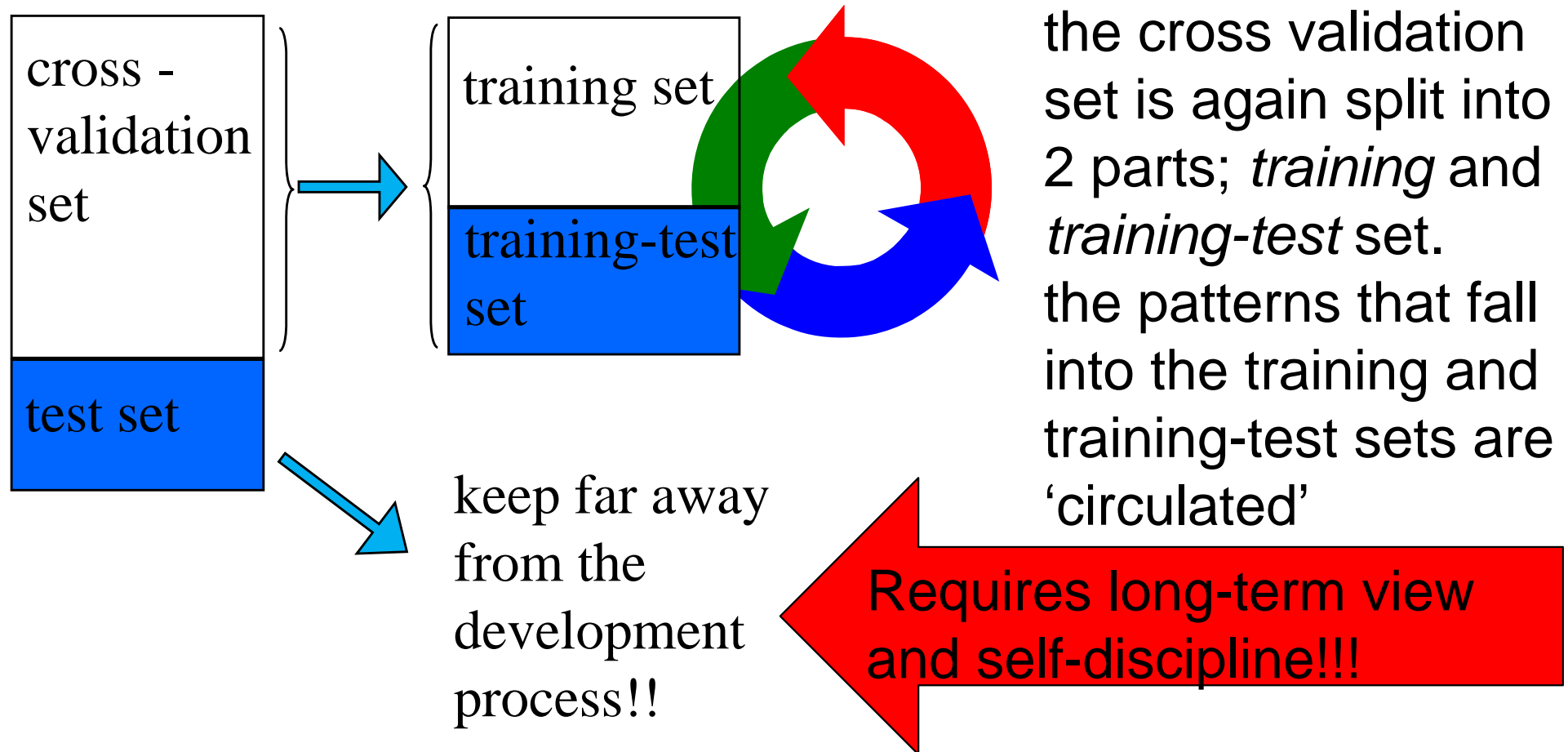
In other words

- If I want to make a classifier to detect a "rare" disease, e.g. with prevalence 1% (1 in 100 cases is 'disease', 99 are 'healthy')
- And I use some scheme to have a training&test set that has 50% of the cases as disease and 50% as healthy, to make my "training easier"
- The PPV values that are reported after cross-validation may be enormously inflated vs what I will encounter in real-life use
- Hence: rely more on confusion matrices, sens, spec

Cross Validation

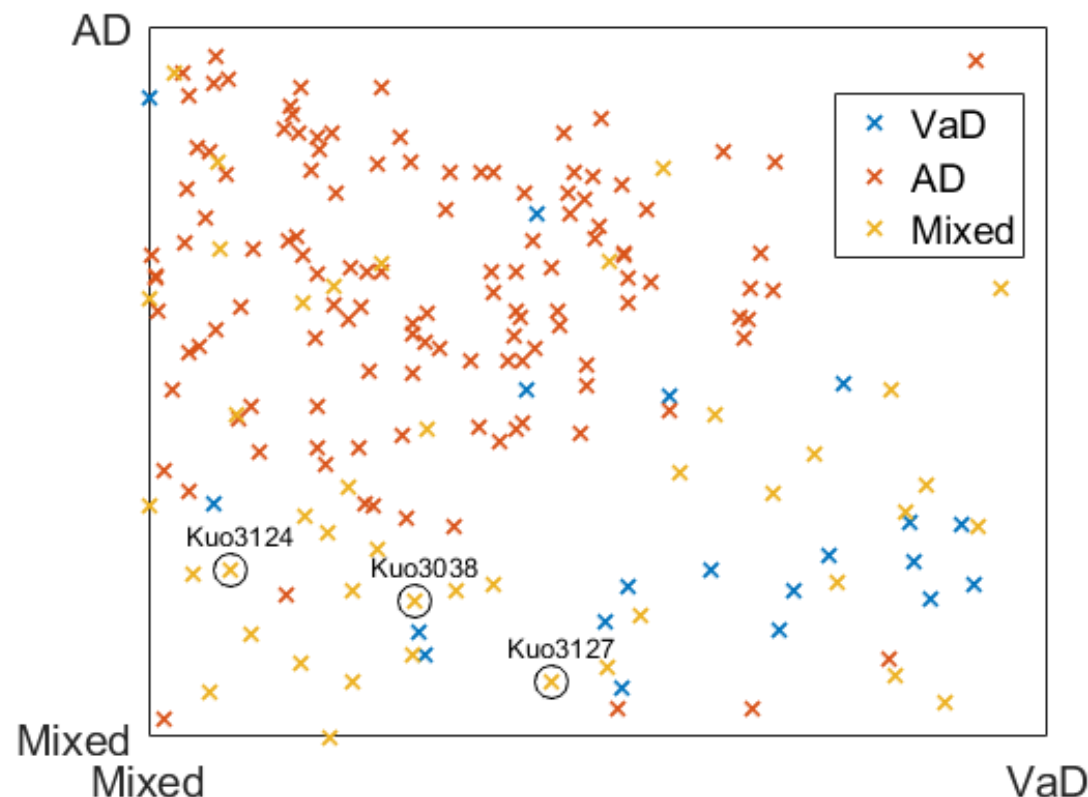
it is common practice to divide the total data set into 2 parts: the *cross-validation* (CV) set (this is what we work with) and the *test set*.

The test set is left absolutely untouched during the entire development process, only when we have finalised our classifier we can use it for performance assessment.



Complex cases, co-morbidities

- Most people (at older age) have multiple diseases at the same time (diabetes&cardiac disease; dementia&high blood pressure, etc)
- This makes classification more difficult, as we have no clear class 0 vs class 1 data anymore, but more complex mixes
- This is enormously tricky – for clinicians as well as data scientists, but one of the Grand Challenges to solve



Cost functions for training an AI

- `loss = tf.losses.mean_squared_error(...`
- `loss = tf.losses.softmax_cross_entropy(..`
- Etc..

- Traditionally we have been using RMSE or other convenient mathematical functions on straightforward data.

- Other measures might be more relevant
- Clinical relevance? Confidence? Speed? Quality of Life? Financial cost?
- Moving towards reinforcement learning?

Incorporating AI into real clinical practice takes considerable efforts, but we will succeed



- If we keep in mind
 - Things like data collection and curation take a long time
 - Your toolbox is big, think not only deep learning, but also old-school statistics, linear filters, logistic regression, ... – they can serve you very well. Try to see the bigger picture and use all tools that may be useful
 - Co-operation: data-driven and modelling approaches
 - Healthcare professionals are your friends, users and guides: visit and discuss with them – often