Introduction to applied **data mining** in medicine

Andrea Chincarini

INFN - Genova

XVI INFN School on Software for Nuclear, Subnuclear and Applied Physics, Alghero 2019

summary

this talk **won't** cover:

- general machine learning
  - you already had plenty of it in these days
- deep/shallow/convolutional/adversarial/ ... neural networks
  - same as above
- clustering
  - very important but no time...

SIGNAL & NOISE IN MEDICINE

SOME DATA ANALYSIS STEPS

MINING EXAMPLES

WRAP-UP

data mining is ...

the short answer:

- TO FIND **MEANING** WITHIN A DATASET

which often includes

- TO **FORECAST** FUTURE IMPLICATIONS
- TO FIND **AFFINITIES** AND DIFFERENCES

and always imply

- **DECISIONS** ON METHODS & PARAMETERS
- HYPOTHESES **TESTING**

DATA are not just a [large] bunch of numbers

They ...

- have qualities/provenance
- have heterogeneity
- are related to a model
  - implicit / explicit / qualit. / quant.
- come with extra-DATA knowledge
  - metadata

physics

- **Observations**
  - Direct / indirect
  - Derived from previous experiments / better estimates of current theories
- **Theory**
  - One or more models, depend on free parameters
  - Few parameters = happy physicist
- **Experiment**
  - Designed to verify key aspects of theory, prove/disprove models
  - Typical paradigm: **Out = signal + noise**
  - Reproducibility is a key factor
- **Data analysis**
  - Designed to extract “signal” from “noise” [filters]
  - Experiment characterization [noise]
  - Estimate model parameters [from signal]
  - Error estimation relatively simple

medicine

- **Observations**
  - Direct: Clinical practice
- **Theory**
  - **No comprehensive models**
  - Highly complex system
  - Subsystem interactions and history not negligible
- **Experiment**
  - Clinical trials (in vitro, in vivo, ….)
  - Typical paradigm: **improvement / no-improvement**
  - Reproducibility is rarely achieved
- **Data analysis**
  - Designed to extract “improvement probability”
  - Strong a-priori assumptions
  - What is “noise”?
  - Error estimation generally difficult

signal & noise

Noise

Random fluctuations that obscure or do not contain meaningful data or other information

Signal

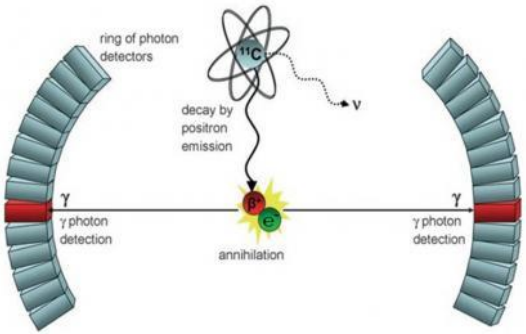
Those **meaningful** data or other information, which are interesting to us

the **meaning** depends on the **goal**

Always **ask** yourself what is the relevant information that might be present in your data
Only then can you define what signal and noise really are ...

example / Positron Emission Tomography

line of response (LOR) for unscattered photons

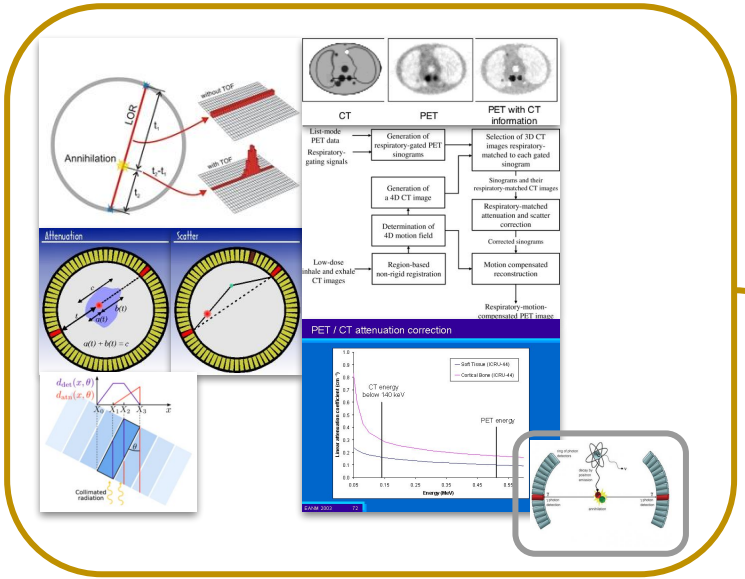


Physical process: 511 keV γ photons

Signal → num. of events (counts in coincidence received on the detector)

Noise → dark current, quantum efficiency, alignment, crystal uniformities, impurities, ...

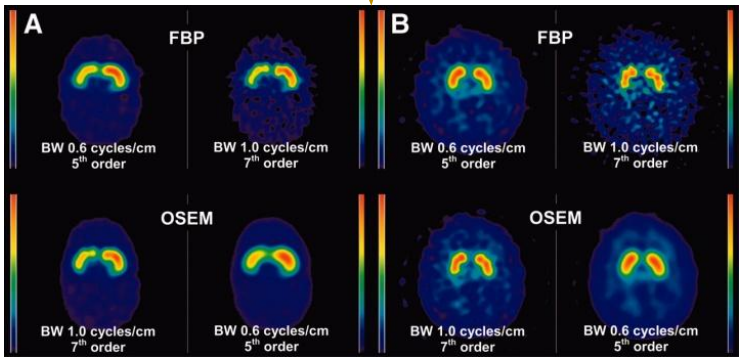
ToF, scatter correction, spatially variant PSF compensation



CT (X-ray) tissue dependent 3D attenuation, motion compensation, reconstructing algorithm

Signal → 3D intensity map (image), the nicest one … but which one?

Noise → electronics, calibration issues, algorithm parameters, models, displaced intensity



injection protocol,
scanner acquisition settings

Image analysis,
clinical evaluation,
metadata

Signal → likelihood of showing a pathological pattern

Noise → comorbidities, pathological models, templates, human experience, ...

noises in medicine

- ACQUISITION
  - Protocol (resolution, calibration, …)
  - Scanner/site quality issues (B-field inhomogeneities, electronic noise…)
  - Patient artefacts (movements, implants, medications, …)
- PROCESSING
  - Image reconstruction algorithm
  - Signal is deduced by comparison among cohorts → method selection is important
  - Information degradation due to sub-optimal processing
  - Depends on assumptions on “signal”
- PHYSIOLOGICAL
  - Confounding variables (age, sex, education, general anamnesis,…)
  - History (comorbidities, unrecalled events, …)
- STANDARDS
  - What is our standard? Clinical evaluation? Autoptic studies?
  - Group mixing (clinical assessment is not 100% accurate)
  - Group purity (comorbidity, who is a “Normal/healthy control”)
  - Data provenance / population sampling
- MODEL
  - Data interpretation depends on pathology model
  - Critical decision about the prognosis
  - Analysis validation, inclusion/exclusion criteria

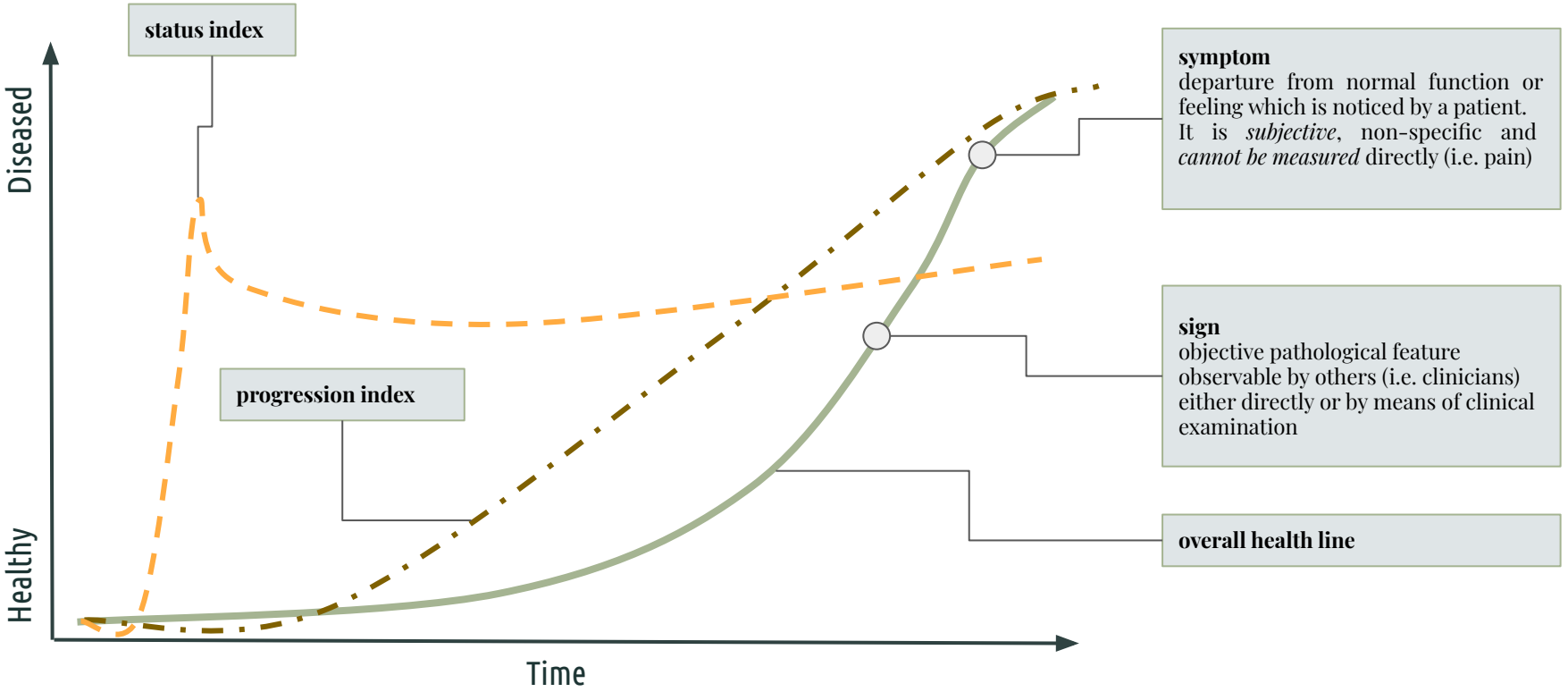
stochastic-like

bias

systematic

what about the signal?

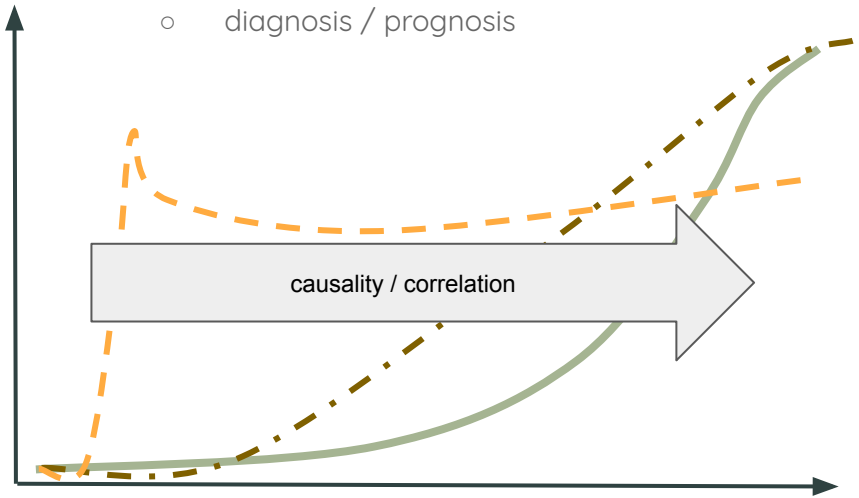
symptoms, signs & markers



biomarkers

If ...

- there are a sufficient number of observations
- the statistical evidence is strong (cohorts, sensitivity, specificity,...)
- it works within a comprehensive pathological model
- longitudinal studies show at least correlation
    - diagnosis / prognosis



def.1: an objective indication of medical state observed from outside the patient which can be **measured accurately** and **reproducibly**.

def.2: any substance, structure, or process that can be measured in the body or its products and **influence** or **predict** the incidence of outcome or **disease**

biomarkers stand in **contrast** to medical symptoms, which are limited to those indications of health or illness perceived by patients themselves or read by trained personnel.

assumptions & pathology models

- pathology models are the medical counterpart of theories in physics
- unfortunately, they are mostly qualitative assessment relying on several assumptions and often limited data
- Yet, it is possible (and useful) to integrate them into our data analysis

A **good** analysis plan includes **models** and **assumptions** in it. This approach allows to test deviations from the theory and allows a more informed analysis

assumptions example:

- **Space**
  - Pathology manifestation is characterized by a “common signature” in the data and throughout the subjects
- **Time**
  - Pathology development is slow [quick] with respect to other physiological variabilities
- **Linearity**
  - Comorbidity is additive
- **Survival**
  - Comorbidity is multiplicative
- **Derivative**
  - The path from normalcy to pathological state can be modeled as a “smooth, continuous” transition so that we can use the two extremes as reference
- **Sampling**
  - Our sample is a good/bad representative of the whole population

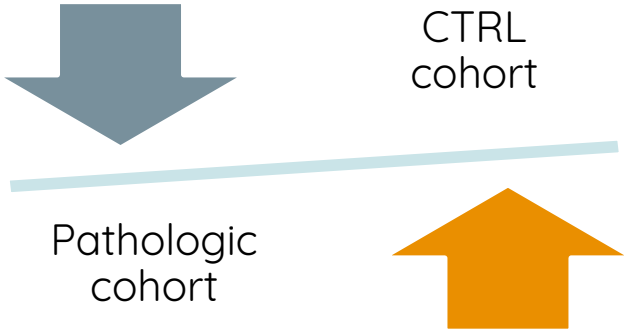
abstraction

Measuring is the core concept for a biomarker. This is where data mining comes into play

Raw information is often too coarse and "dirty" to be useful

Abstraction is really important. It is the approach to the data where we embed extra knowledge (models, qualitative information, etc.) and clean our data so that we can properly apply analysis techniques.

Abstraction is often implemented as multiple pre-processing steps, which often include feature extraction, dealing with missing data and typically imply dimensionality reduction techniques.



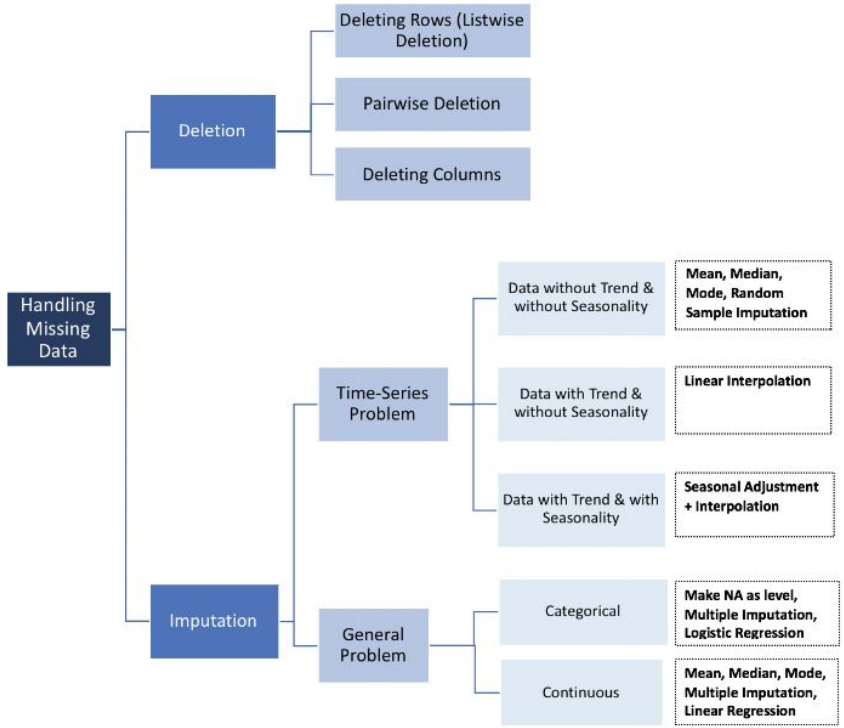
1. Clean data
2. Embed pathology models and extra info
3. Make data commensurable
4. Find common traits within cohorts
5. Find differences between them
6. Test and validate

missing data

fill in holes

- **Missing at Random** (MAR)
  - missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- **Missing Completely at Random** (MCAR)
  - the fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- **Missing not at Random** (MNAR)
  - missing value depends on the hypothetical value or on another variable (e.g. females generally don't want to reveal their ages! missing value in age variable is impacted by gender variable)

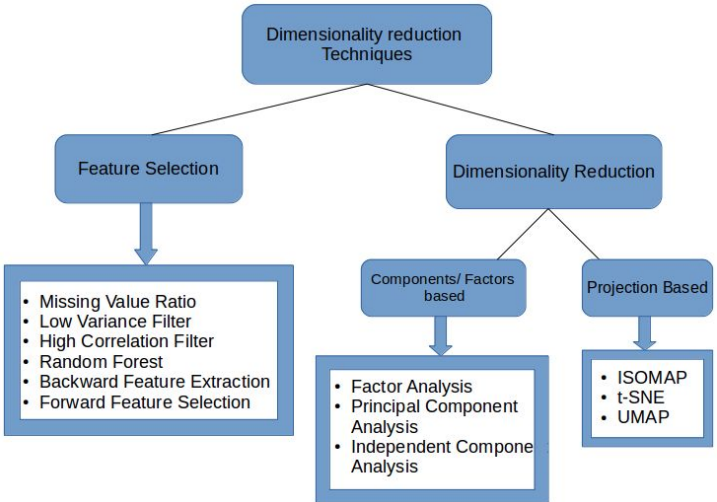
In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations. Note that imputation does not necessarily give better results.



dimensionality reduction

Find the data representation (space) such that the contrast between signal and noise is maximized.

The selection of the appropriate space is usually the key to a successful data mining

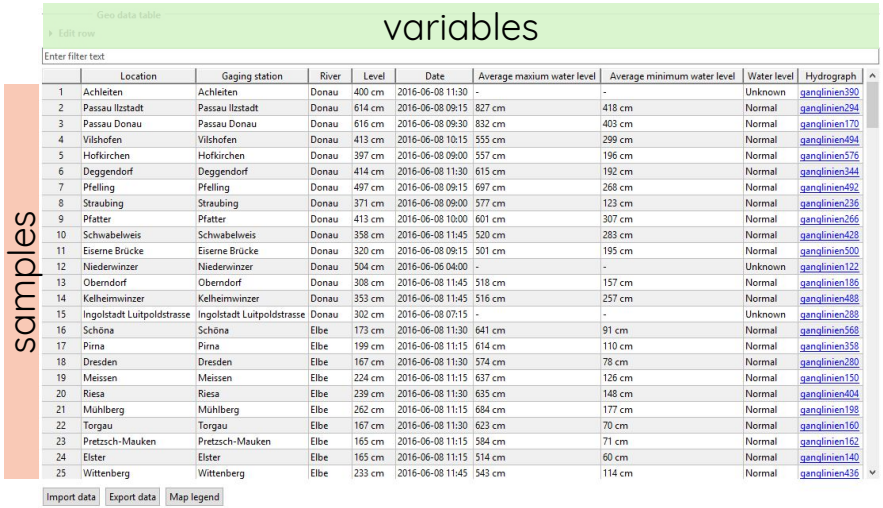


some notable examples

- Linear
  - SVD, **PCA**
  - Factor Analysis, **Ind. Component An.**
- Non linear
  - **manifold embedding**
  - t- Distributed Stochastic Neighbor Embedding (t-SNE)
  - **autoencoders**
- Feature selection
  - **random forest** (RF)
  - greedy algorithms
  - **correlation filters**

data as table

after the abstraction layer we are often left with a table. here is where core data analysis techniques live



	Location	Gaging station	River	Level	Date	Average maxium water level	Average minimum water level	Water level	Hydrograph
1	Achleiten	Achleiten	Donau	400 cm	2016-06-08 11:30	-	-	Unknown	ganglinien390
2	Passau Ilzstadt	Passau Ilzstadt	Donau	614 cm	2016-06-08 09:15	827 cm	418 cm	Normal	ganglinien294
3	Passau Donau	Passau Donau	Donau	616 cm	2016-06-08 09:30	832 cm	403 cm	Normal	ganglinien170
4	Vilshofen	Vilshofen	Donau	413 cm	2016-06-08 10:15	555 cm	299 cm	Normal	ganglinien494
5	Hofkirchen	Hofkirchen	Donau	397 cm	2016-06-08 09:00	557 cm	196 cm	Normal	ganglinien576
6	Deggendorf	Deggendorf	Donau	414 cm	2016-06-08 11:30	615 cm	192 cm	Normal	ganglinien344
7	Pfelling	Pfelling	Donau	497 cm	2016-06-08 09:15	697 cm	268 cm	Normal	ganglinien492
8	Straubing	Straubing	Donau	371 cm	2016-06-08 09:00	577 cm	123 cm	Normal	ganglinien236
9	Pfatter	Pfatter	Donau	413 cm	2016-06-08 10:00	601 cm	307 cm	Normal	ganglinien266
10	Schwabelweis	Schwabelweis	Donau	358 cm	2016-06-08 11:45	520 cm	283 cm	Normal	ganglinien428
11	Eiserne Brücke	Eiserne Brücke	Donau	320 cm	2016-06-08 09:15	501 cm	195 cm	Normal	ganglinien500
12	Niederwinzer	Niederwinzer	Donau	504 cm	2016-06-06 04:00	-	-	Unknown	ganglinien122
13	Oberndorf	Oberndorf	Donau	308 cm	2016-06-08 11:45	518 cm	157 cm	Normal	ganglinien186
14	Kelheimwinzer	Kelheimwinzer	Donau	353 cm	2016-06-08 11:45	516 cm	257 cm	Normal	ganglinien488
15	Ingolstadt Luitpoldstrasse	Ingolstadt Luitpoldstrasse	Donau	302 cm	2016-06-08 07:15	-	-	Unknown	ganglinien288
16	Schöna	Schöna	Elbe	173 cm	2016-06-08 11:30	641 cm	91 cm	Normal	ganglinien568
17	Pirna	Pirna	Elbe	199 cm	2016-06-08 11:15	614 cm	110 cm	Normal	ganglinien358
18	Dresden	Dresden	Elbe	167 cm	2016-06-08 11:30	574 cm	78 cm	Normal	ganglinien280
19	Meissen	Meissen	Elbe	224 cm	2016-06-08 11:15	637 cm	126 cm	Normal	ganglinien150
20	Riesa	Riesa	Elbe	239 cm	2016-06-08 11:30	635 cm	148 cm	Normal	ganglinien404
21	Mühlberg	Mühlberg	Elbe	262 cm	2016-06-08 11:15	684 cm	177 cm	Normal	ganglinien198
22	Torgau	Torgau	Elbe	167 cm	2016-06-08 11:30	623 cm	70 cm	Normal	ganglinien160
23	Pretzsch-Mauken	Pretzsch-Mauken	Elbe	165 cm	2016-06-08 11:15	584 cm	71 cm	Normal	ganglinien162
24	Elster	Elster	Elbe	165 cm	2016-06-08 11:15	514 cm	60 cm	Normal	ganglinien140
25	Wittenberg	Wittenberg	Elbe	233 cm	2016-06-08 11:45	543 cm	114 cm	Normal	ganglinien436

"DATA AS TABLE" is the most common processed format in data mining

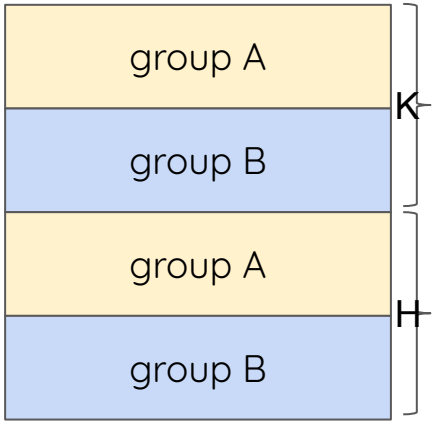
- descriptive statistics
- ROC analysis
- feature selection
- classifiers
- clustering
- linear / non-linear multivariate analysis
- predictions
- ...

the provenance systematic

Systematic error due to data acquisition, treatment, internal quality and pre-processing protocols that is related to a categorical variable (typically the acquisition site).

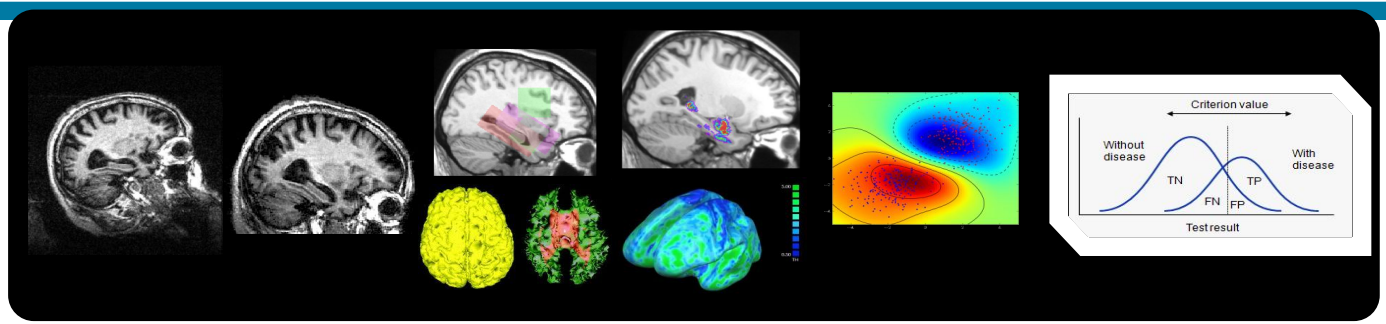
Provenance systematic is very difficult to eliminate “a-priori” and it must always be considered in the analysis as a co-factor.

In medical data, the typical provenance error is much greater than the signal



$(A_K - A_H) \sim (B_K - B_H) > (B_j - A_j)$

typical process



Signal contribution	Physics, scanner, protocols	Image modality	Clinical input	Information theory	Data set
Processing	Quality assessment	Normaliza tion	Characteriza tion	Feature extraction	Classifier
Noises	Acquisition noises	Physiological	Processing, Gold standard	Data processing	Gold standard

Raw data 10^7 - 10^8 DoF

abstraction / model 10^1 – 10^3 DoF

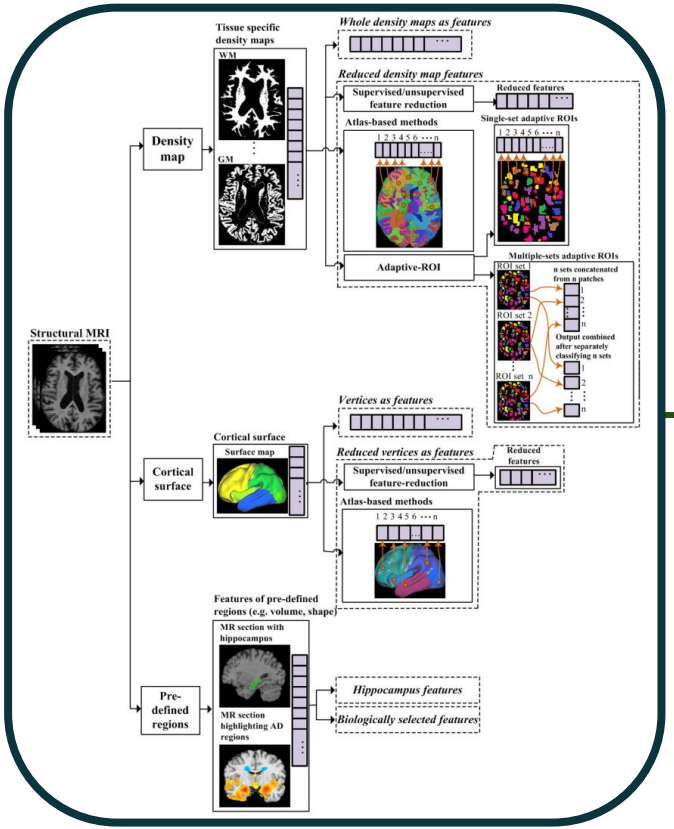
Marker 1-2 DoF

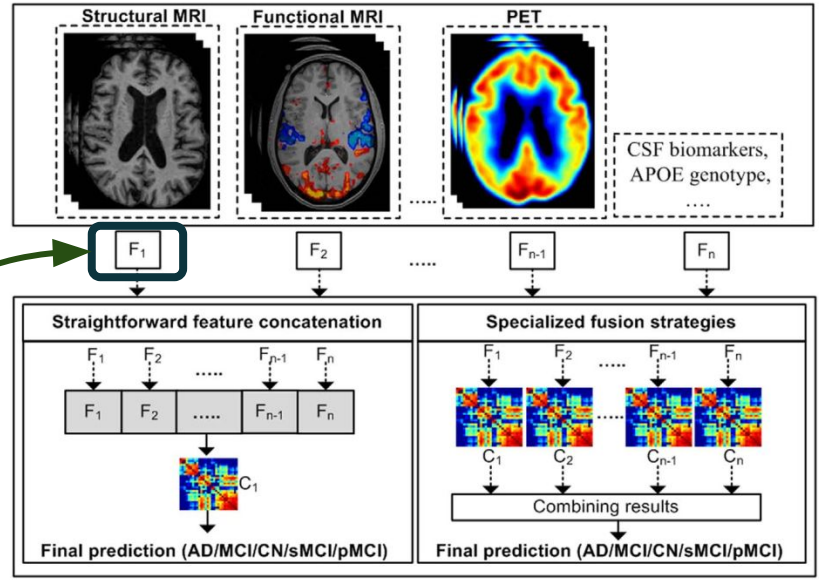
Raw data — Synthesis

MARKER

Abstraction

which can be complicated as needed...





example #1

working with mixed data types

The European DLB* dataset

N=183 samples (patients) diagnosed in **9** European clinical centers

- brain FDG-PET scans
- various metadata

P.S. for this disease, 183 patients in Europe is the largest dataset to date ...

*Dementia with Lewy Bodies

the E-DLB dataset

4 “core clinical features” 0/1

- [PARK] parkinsonism
- [VH] visual hallucinations
- [CFL] cognitive fluctuations
- [RBD] REM-behavior disorder

no “pure” samples (patients always show mixed symptoms)

Annals of NEUROLOGY
An Official Journal of the American Neurological Association and the Child Neurology Society
ANA AMERICAN NEUROLOGICAL ASSOCIATION

Research Article

Metabolic patterns across core features in dementia with lewy bodies

Silvia Morbelli MD, PhD, Andrea Chincarini PhD, Matthias Brendel MD, Axel Rominger MD, Rose Bruffaerts MD, PhD, Rik Vandenberghe MD, PhD, Milica G. Kramberger MD, PhD ... See all authors

First published: 25 February 2019 | https://doi.org/10.1002/ana.25453

analysis questions:

- is there a significant relationship between PET uptake and the core clinical features?
- if so, does this relationship has a distinct spatial characteristics (pattern)?
- can we find common traits to a single core clinical feature?
- can we use this trait as a way to discriminate / diagnose patients?

step 1

clean & normalize

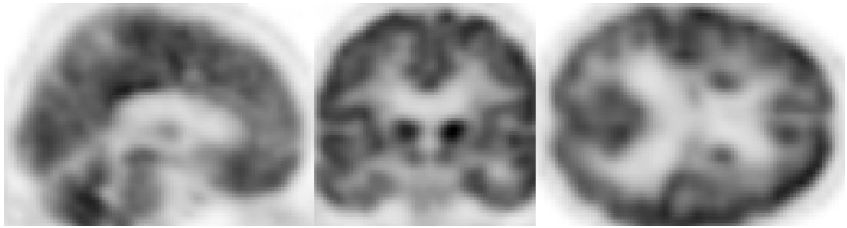
assess data properties

look for missing data/outliers and decide how to handle them

embed hypotheses

transform data into a normative space and intensity

properties



Code	Centre	eyes	Gender	Age	Education	MMSE	Parkinsonism
'GEN-DLB-30'	genova	closed	m	70	8	26	1
'GEN-DLB-33'	genova	open	f	81	5	5	1
'GEN-DLB-43'	genova	open	f	78	5	25	0
'GEN-DLB-05'	genova	open	f	75	5	17	1
'GEN-DLB-07'	genova	open	f	86	5	26	1
'GEN-DLB-10'	genova	open	f	71	8	22	1
'GEN-DLB-11'	genova	open	m	68	8	27	1
'GEN-DLB-13'	genova	open	f	72	8	28	0

dataset consists in N=183 samples (patients) diagnosed in 9 European clinical centers

heterogeneous data / sample consists of

- a 3D matrix
  - FDG-PET image, each center with its own dimension and SNR
- 6 categorical items
  - center, gender, protocol, presence of core clin. features
- 1 discretized item
  - MMSE neuropsychological test
- 2 continuous variables
  - age, education

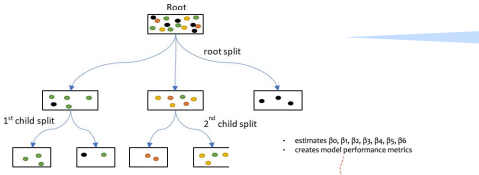
filling missing data with imputation

assumed missing data model: MNAR (missing not at random) because of strong link with center

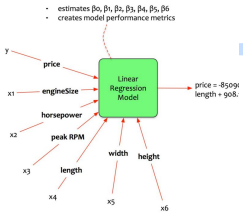
set a limit on data reconstruction

skip samples with too many missing

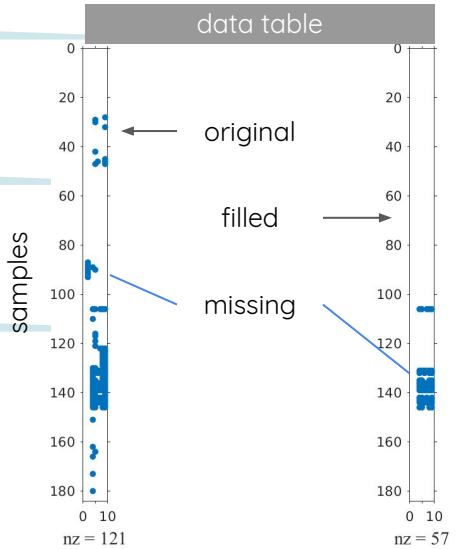
validate final results with the original subset of complete data



decision tree for categorical var.



multivariate linear regression for cont. var.



train on complete data subset

estimate the [model dependent] most likely outcome for missing data

spatial & intensity normalization

spatial registration: iterative process mapping two domains

- The map is a transformation matrix depending on a set of free parameters (d.o.f.)
- A metric is defined to measure how similar is the mapped domain (moving) to the target domain (fixed, template)
- Metric is minimized over d.o.f.

Final space is that of the template.

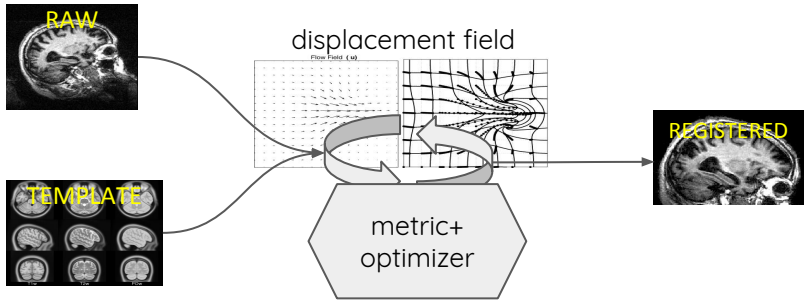
For instance we have now ~5 x 10^5 voxels for all images

normalization is akin to resample and scale your data on a uniform grid

intensity normalization: linear/non-linear scaling of the data to calibrate values on a reference norm

it sets the “unit of measures” for all data

it typically requires a reference measured in the same condition as the data



step 2

abstract

embed extra knowledge

map into feature space

model information

embed knowledge

we want to embed the following notions into the analysis

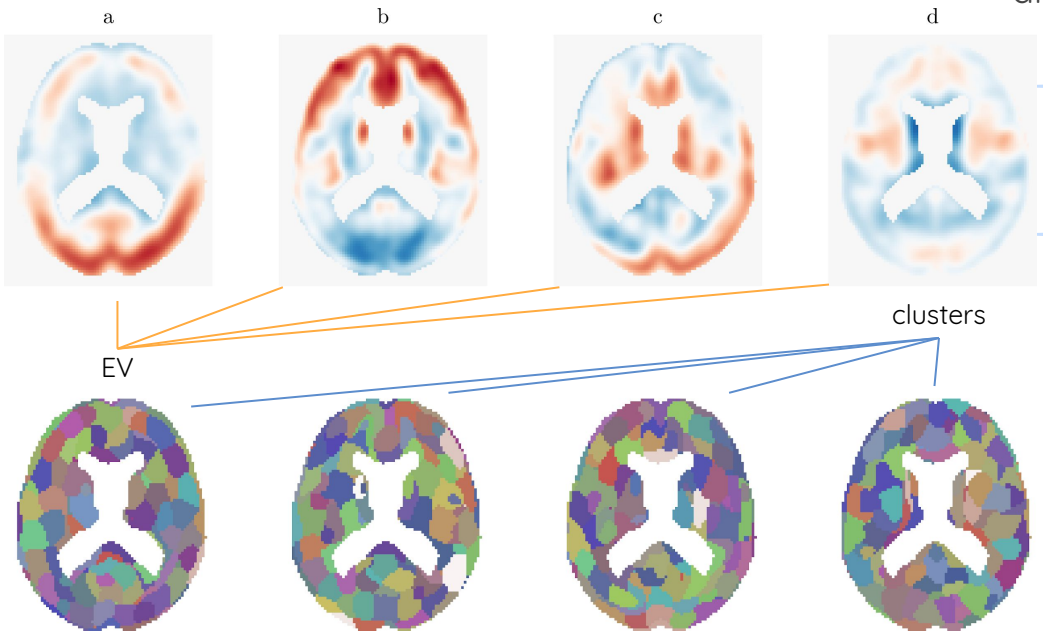
- PET information has a typical spatial coherence length
  - due to PSF and - more importantly - brain regions anatomy
- Intensity variation pattern *should* be related with clinical features
- ~ 5 x 10^5 voxels are too unbalanced with respect to 171 samples

solution:

- use PCA eigenvectors as guide for relevant intensity-range
- partition volume using coherence length

mapping into feature space

first 20 eigenvectors (then spatial frequencies become higher than the inverse length)



for each eigenvector:

zscore normalization to provide compactness and link to PCA variability

embed EV info v into coordinates [x y z v] for k-means clustering

compact clusters which follow the EV gradient information

~ 500 partitions per EV (due to spatial coherence)

model

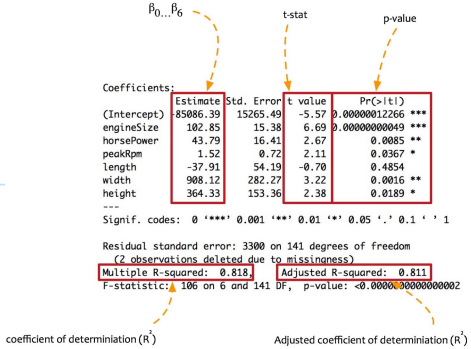
we apply now the multivariate linear model **for each cluster** in a partition

$$\langle \text{mean intensity in ROI} \rangle = \sum_j \text{effect}_j \times \text{covariate}_j + \text{residual}$$

dependent variable

clinical core features, demographics, etc.

output: **significance** (p-values), **effect** size* for each covariate



no quadratic / interaction effect for now.
a simpler model is more robust (keep in mind n. of samples). test residuals for Gaussianity...

*rate of change in the dependent variable for a unit change in the covariate

step 3

analyze

estimate parameters

extract knowledge

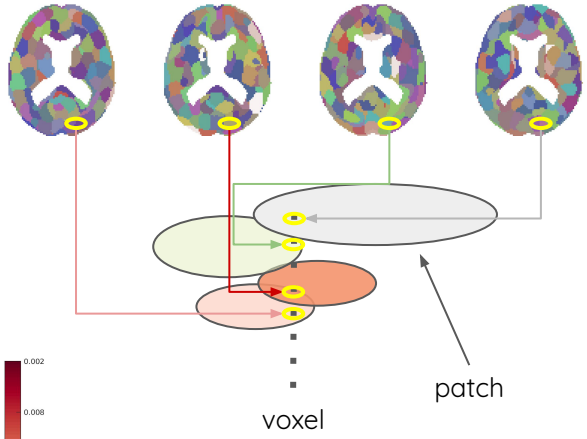
validate assumptions

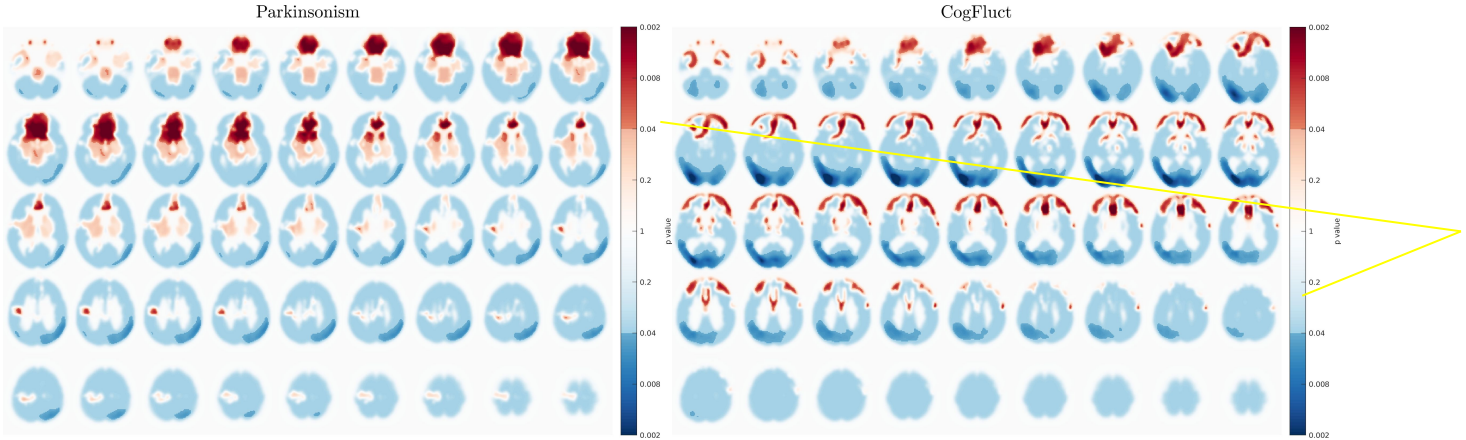
estimate significant ROI

spatial mean of effects and p-values

for each voxel, average p-values and effects over all partitions to get a robust map

partition average avoid fluctuations due to multiple comparison (i.e. Bonferroni correction)



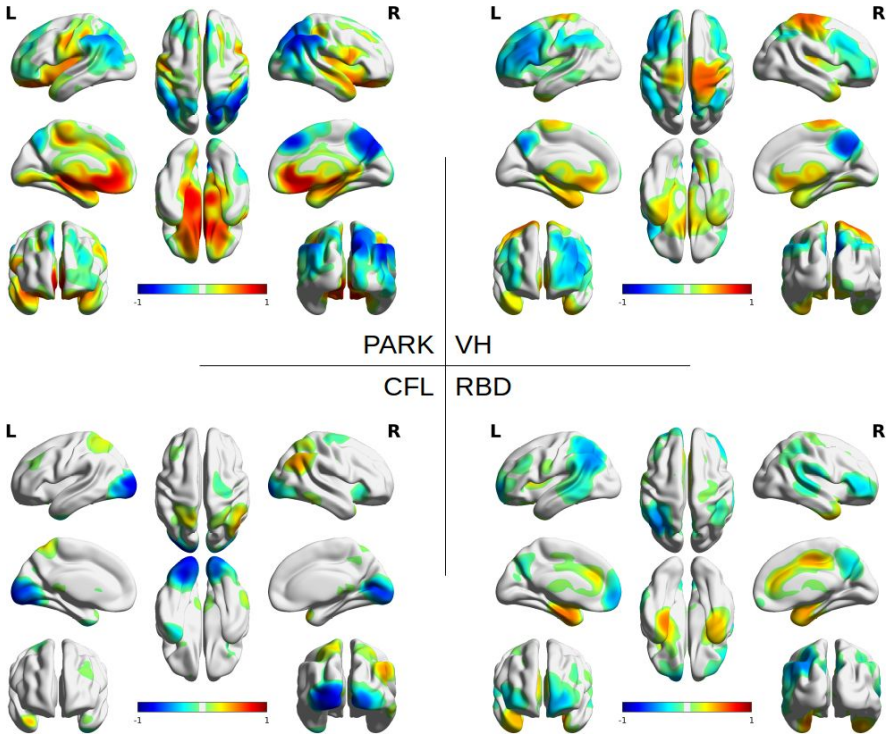


patterns

this is the axis onto which we project our data

patterns = [effect / const] $|_{\text{p-value} < \text{threshold}}$

the pattern is a normalized vector with dimensionality = 3D image

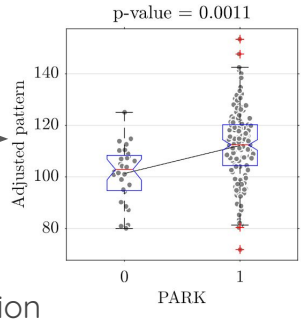


discriminators and classifiers

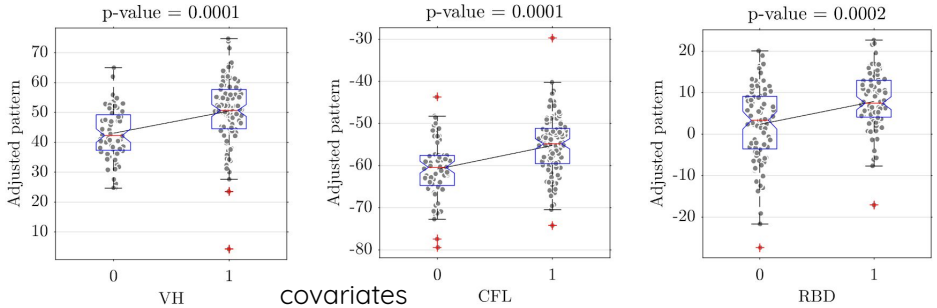
we can think of images and patterns as vectors in space

their scalar product gives the projection

projection adjusted for all other covariates

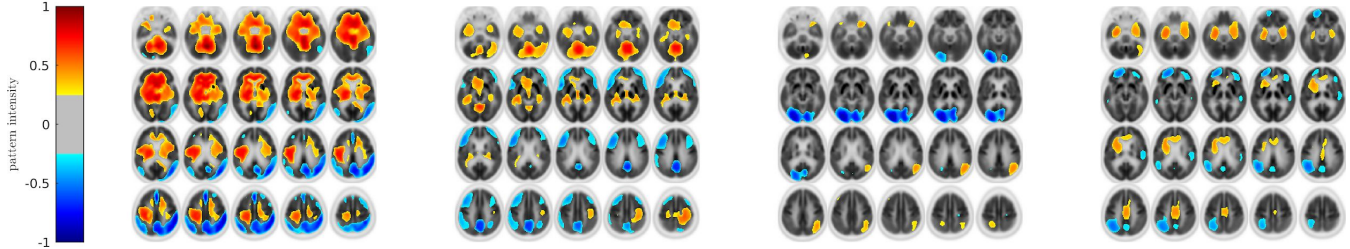


significance of a t-test over the projection

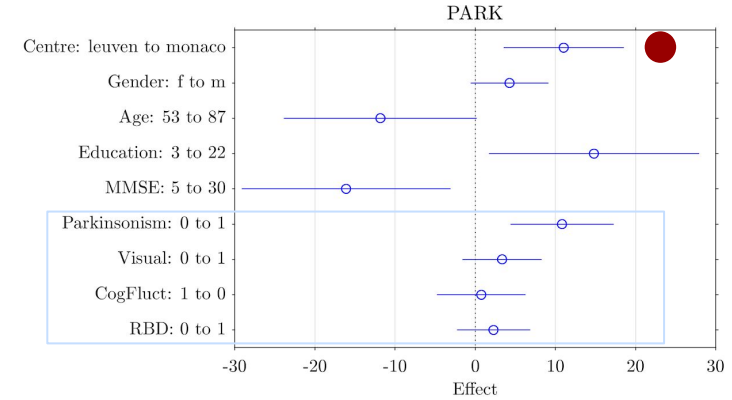


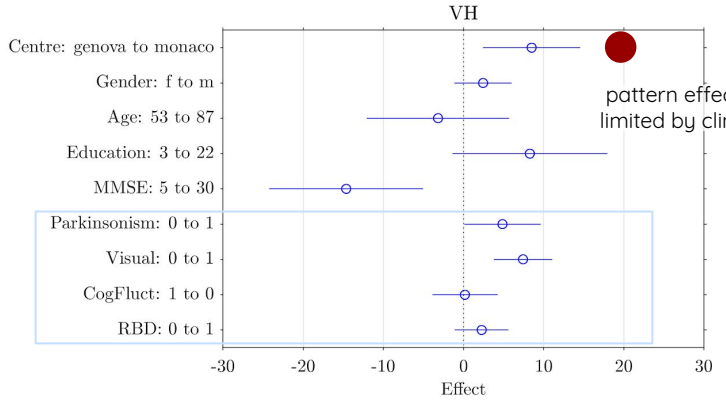
covariates

patterns

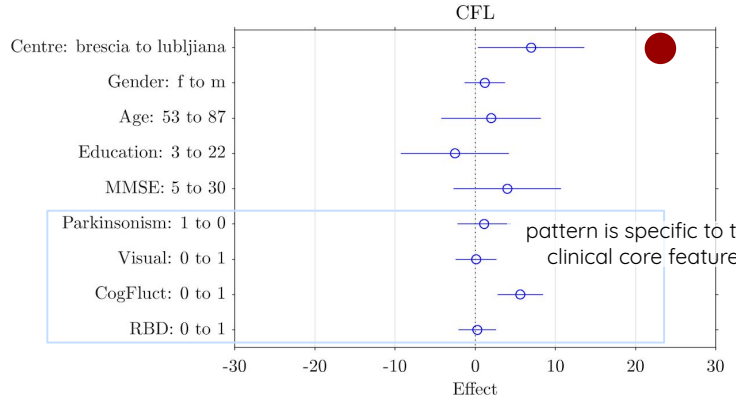


gauge the effect size

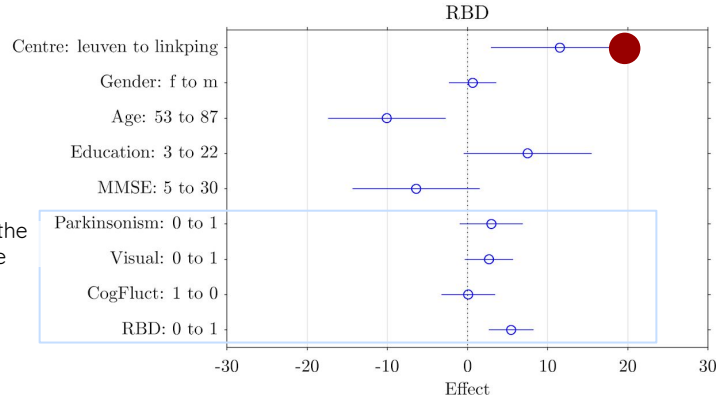




pattern effect size severely limited by clinical covariates

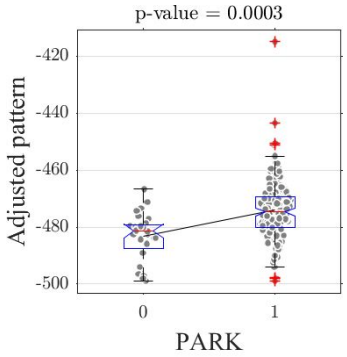


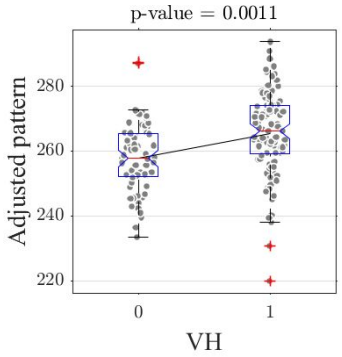
pattern is specific to the clinical core feature

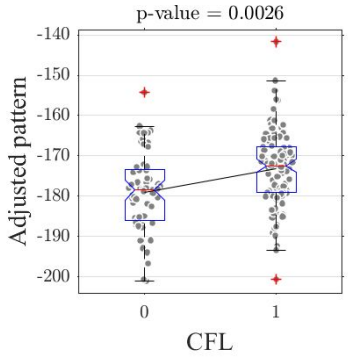


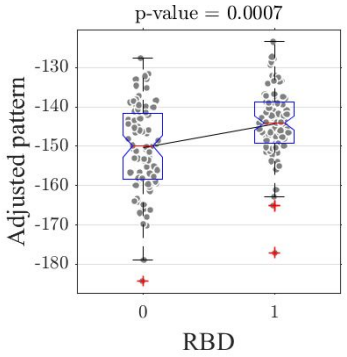
inference on new data

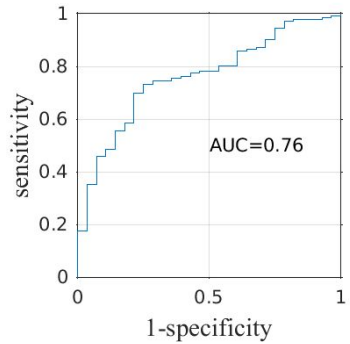
cohort statistics vs. single sample prediction

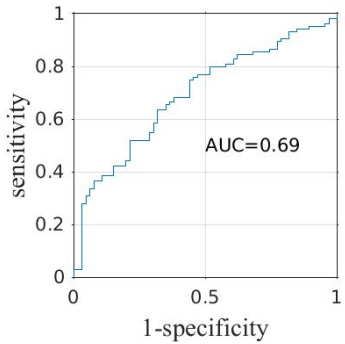


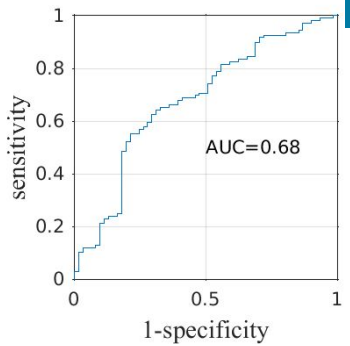




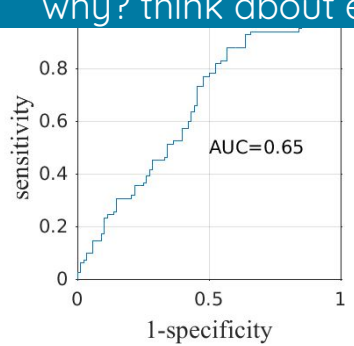








why? think about effects



more useful techniques

a glimpse on...

texture

relationship

causality

texture

scalar values are not the only interesting property. **relationships** (**textures**, pattern) can be much more **meaningful**

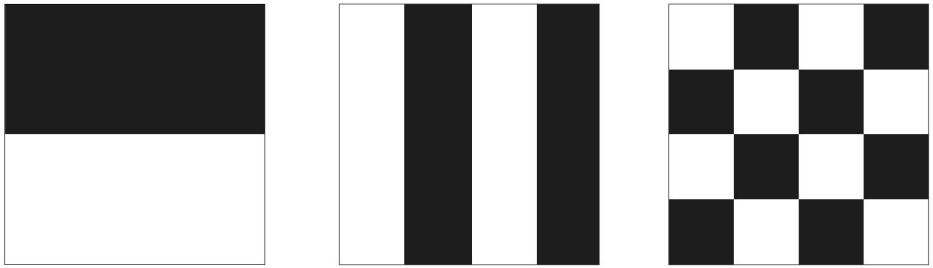
here again the metric use to define the relationship is arbitrary

Texture provides information in the **spatial arrangement** of colours or intensities in an image.

Texture is characterized by the spatial distribution of intensity levels in a neighborhood.

50% black and 50% white distribution of pixels

Three different images with the same intensity distribution, but with different textures

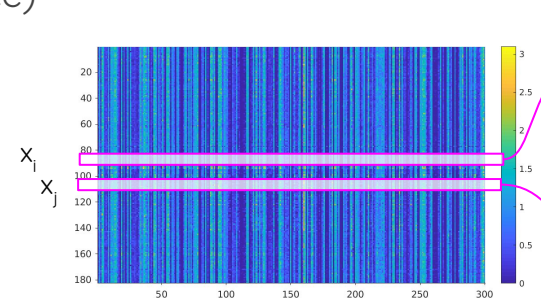


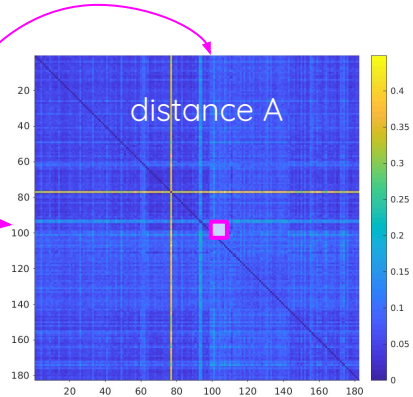
relationship

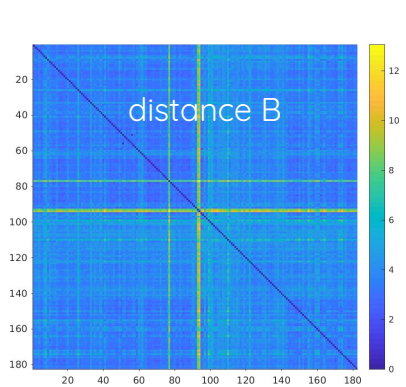
relationship (distance) matrix

$M_{ij}=d(x_i,x_j)$

very useful for patterns, networks, clustering, ...





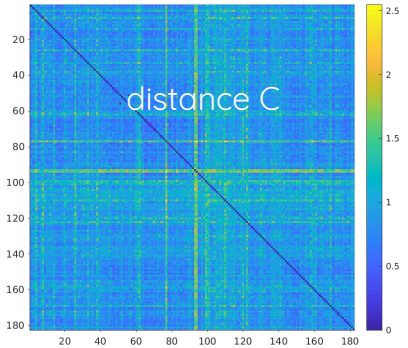


depends on the distance:

A: euclidean $d_{sl}^2 = (x_s - x_t)(x_s - x_t)'$

B: correlation $d_{ss} = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'}}$

C: chebyshev $d_{st} = \max_j\{|x_{sj} - x_{tj}|\}$



causality

relationship based on a distance is symmetrical

$d(x_i,x_j)=d(x_j,x_i)$

causality analysis can infer dominance in the dynamic of a variable over another

$c(x_i,x_j)\neq c(x_j,x_i)$

can test assumptions and models

very powerful! but use with caution

Most common methods:

- Structural equations
- Granger Causality
- Convergent Cross Mapping

PLOS ONE

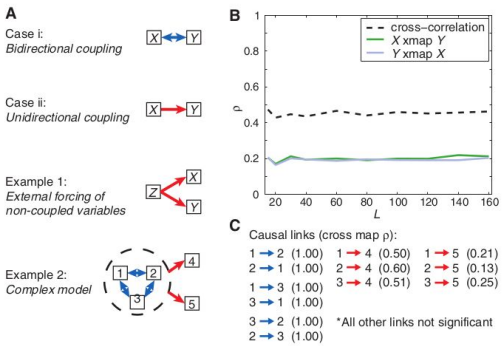


Fig. 4. Model causal networks. (**A**) Schematics of causal networks: two base cases and two model examples showing external forcing of noncoupled variables. (**B**) Cross-map results for example 1: external forcing of noncoupled variables. Cross-correlation erroneously suggests that *X* and *Y* are interacting, whereas cross mapping correctly shows that there is no interaction. (**C**) Cross-map results for the complex five-species model example. All significant ($P < 0.05$) mappings are given and indicate that species 1, 2, and 3 (the subsystem in the circle) all interact mutually (case i), but interact only asymmetrically as external forcing variables with respect to 4 and 5 (case ii), which do not interact directly themselves.

RESEARCH ARTICLE

Causality Analysis: Identifying the Leading Element in a Coupled Dynamical System

Amir E. BozorgMagham[1]*, Safa Motesharrei[2,3,4], Stephen G. Penny[1,5], Eugenia Kalnay[1,4]

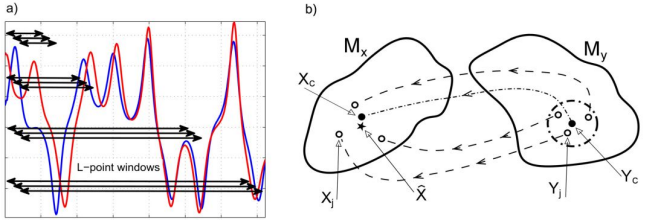


Fig 2. Schematics of *L*–point windows and reconstructed phase spaces. (a) Two time–series $x(t)$ and $y(t)$ and a schematic of the *L*–point windows with different lengths sweeping the whole span of the time–series. (b) A schematic of the reconstructed phase spaces of the *L*–point windows corresponding to two time–series. For each *E*-dimensional Y–central point, Y_c, in the response reconstructed phase space, M_y, a sufficient number of nearest neighbor points are selected (empty circles, Y_j, right) and their distances, d_j, to Y_c are determined. For each neighbor point, its contemporaneous point in the driver reconstructed phase space, M_x, is determined (empty circles, X_j, left). The weighted average of these points, $\hat{X}$, is compared with X_c, the true contemporaneous point in M_x corresponding to Y_c. The CCM coefficient, $\rho(L)$, is defined as the correlation coefficient between $\hat{X}$ and X_c, averaged over all possible *L*–point windows.

example #2

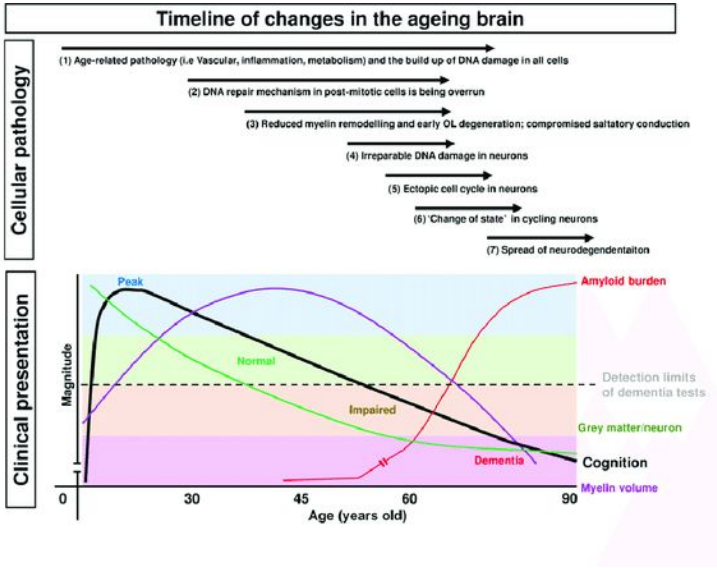
application of relationship matrix to graph & clustering

amyloid accumulation patterns

amyloidosis patterns

accumulation patterns in amyloidosis

- qualitative model indicates amyloid accumulation in the brain as a monotonic function of time
- histopathological studies show that amyloid load slowly grows in the brain from the most central parts towards the periphery
- we can't follow a subject throughout his life with amyloid scans ⇒ we have only cross-sectional data → ergodic theorem



embed hypotheses

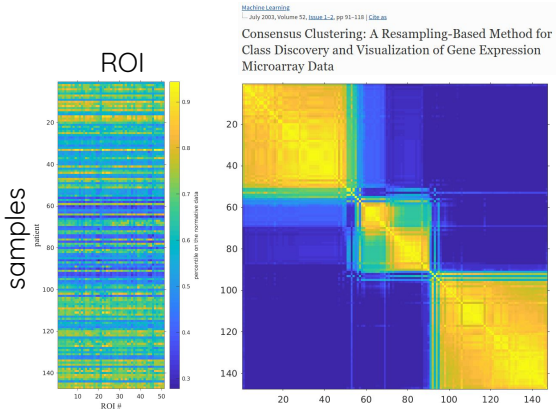
parcellate brain into ROIs and measure the amyloid load for each ROI

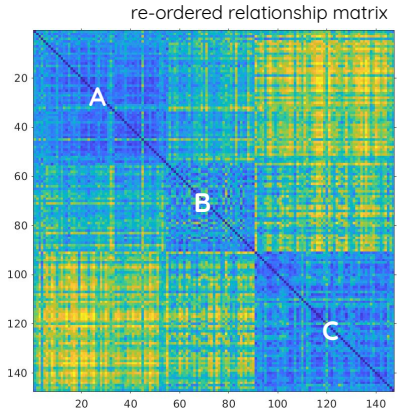
do this for a number of samples (patients) that include all amyloid loads (from the most negative to the most positive)

PCA scores on 1st eigenvector to determine the transition

data matrix

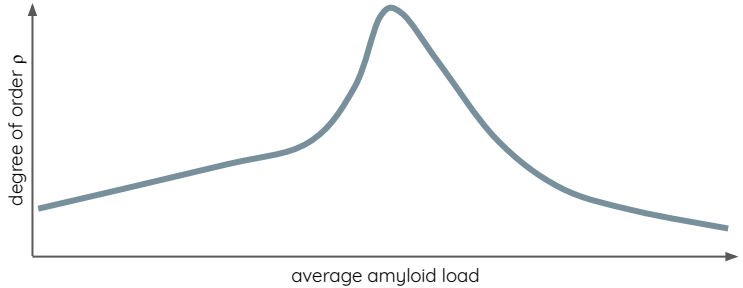
consensus clustering

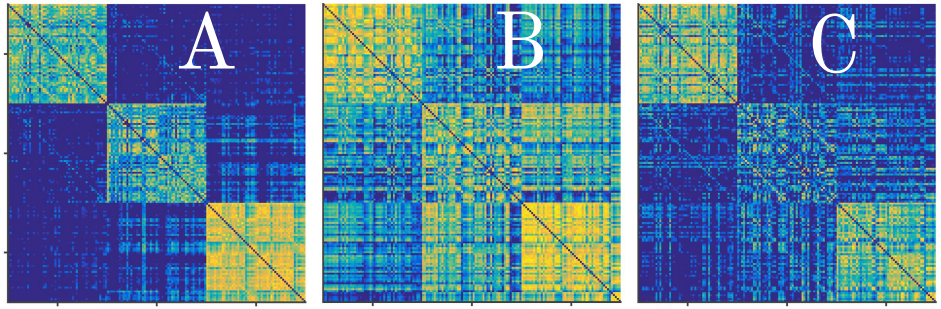


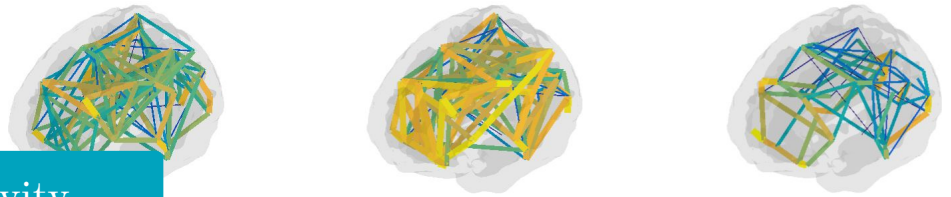


networks

degree-of-order peaks in the transition
several graphs properties to investigate







amyloid uptake transition dominates the connectivity graph. almost all connections lost after plateau!

now we can look for specific/unique patterns

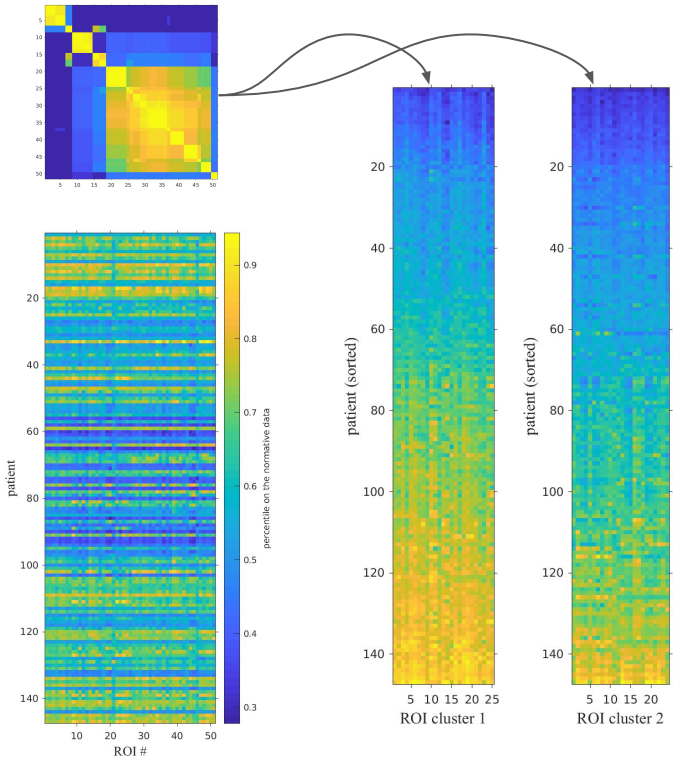
path

accumulation path as rank distance

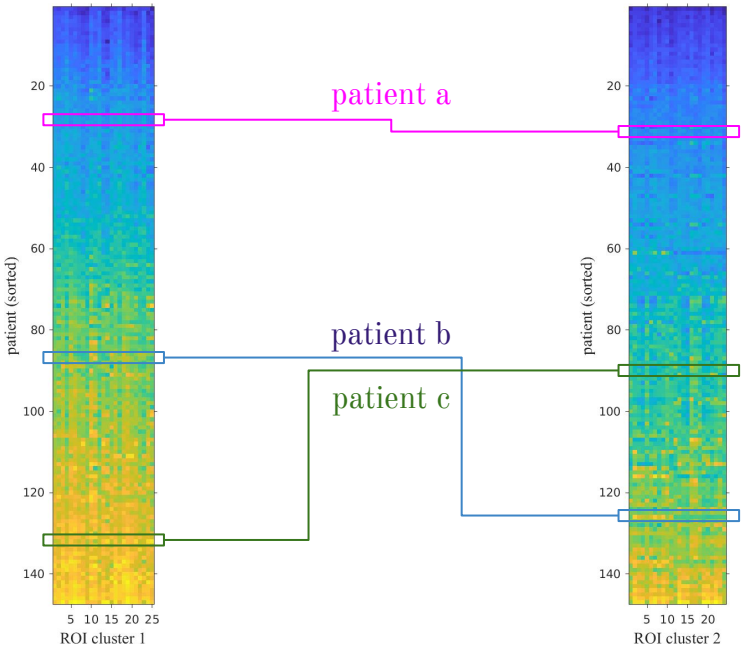
consensus clustering applied to
the **transposed** data matrix

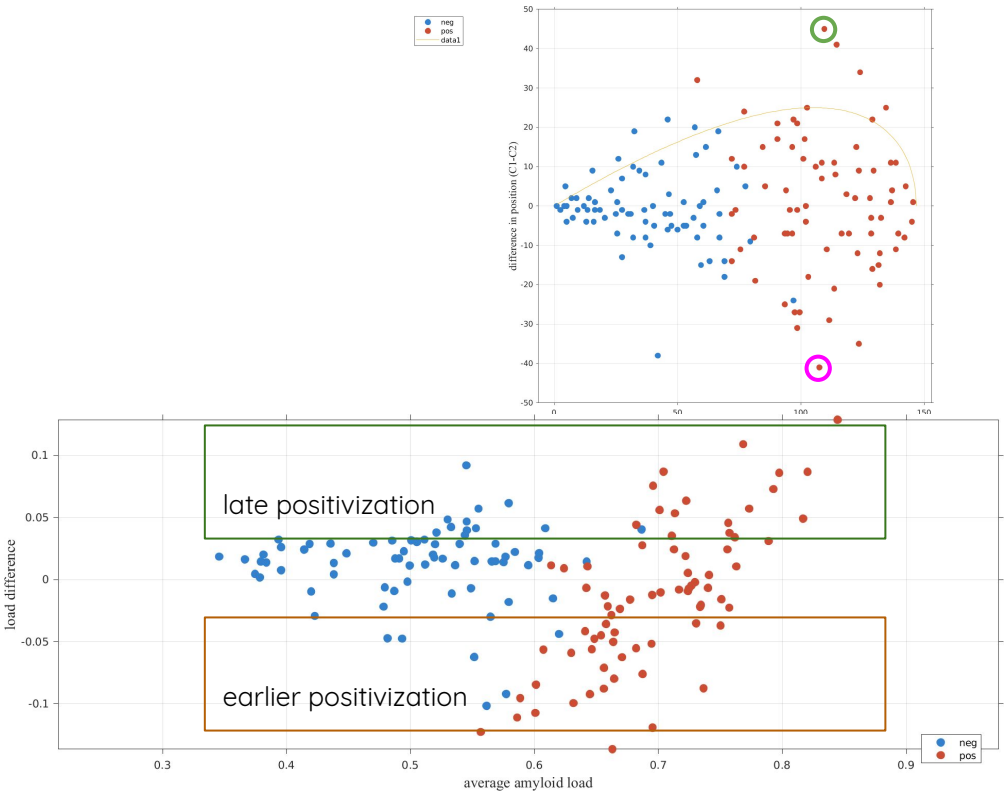
metric: **Spearman correlation** (rank)

we can test possible positivization paths:
do all patients become amyloid-positive in the same way?



if we couple it with the sigmoid model...



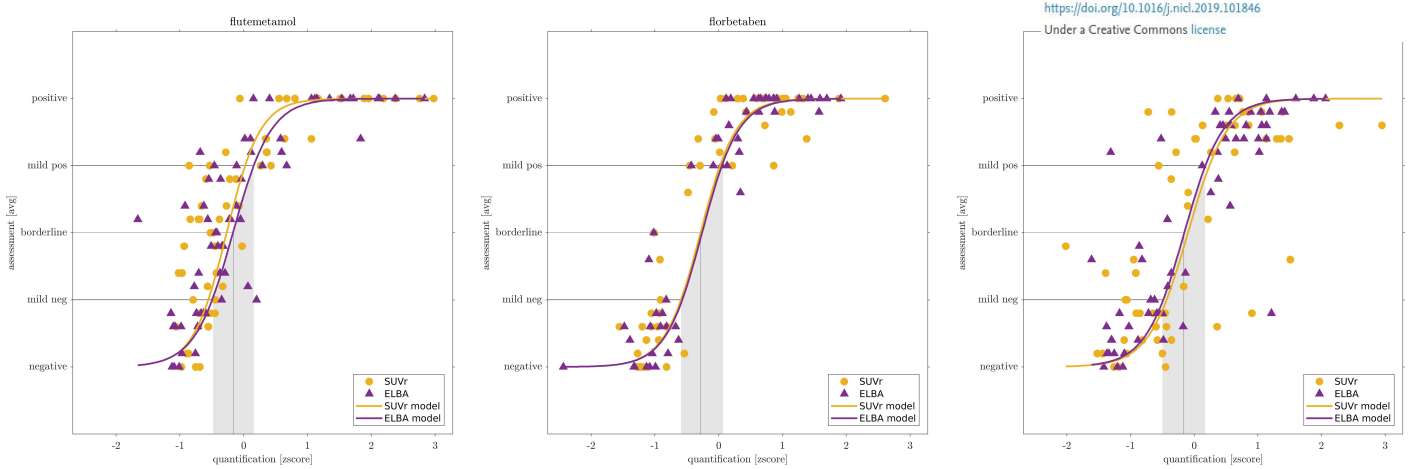


model validation

you can verify models and assumptions, for instance:
take the amyloid accumulation model...

tested with all three 18F amyloid tracers, ~500 scans, 2 quantification methods

gold standard: consensus visual reading (5 ind. clin.)



ELSEVIER
NeuroImage: Clinical
Volume 23, 2019, 101846

Semi-quantification and grading of amyloid PET: A project of the European Alzheimer's Disease Consortium (EADC)

A. Chincarini [a], E. Peira [a, d], S. Morbelli [b, c], M. Pardini [c, d], M. Bauckneht [c], J. Arbizu [e], M. Castelo-Branco [f], K.A. Büsing [g], A. de Mendonça [h], M. Didic [i], M. Dottorini [j], S. Engelborghs [k, l], C. Ferrarese [m], G.B. Frisoni [n, ac], V. Garibotto [o], E. Guedj [p, q], L. Hausner [r, s], J. Hugon [t] ... F. Nobili [c, d]

conclusion

learn to know your data

embed as much information as possible into your analysis

don’t skip on cleaning, normalizing and dim. reduction (data abstraction)

aim for the most informative analysis technique

keep it simple, don’t just fall into the most fashionable technique