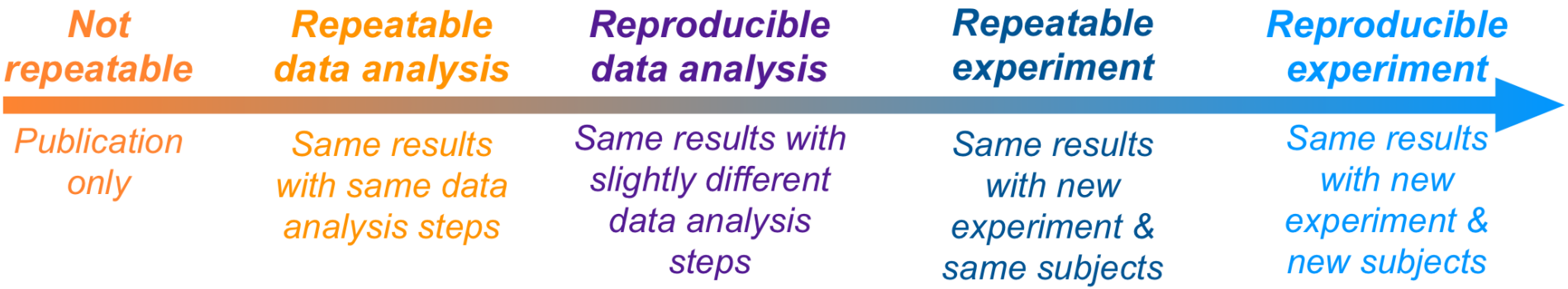
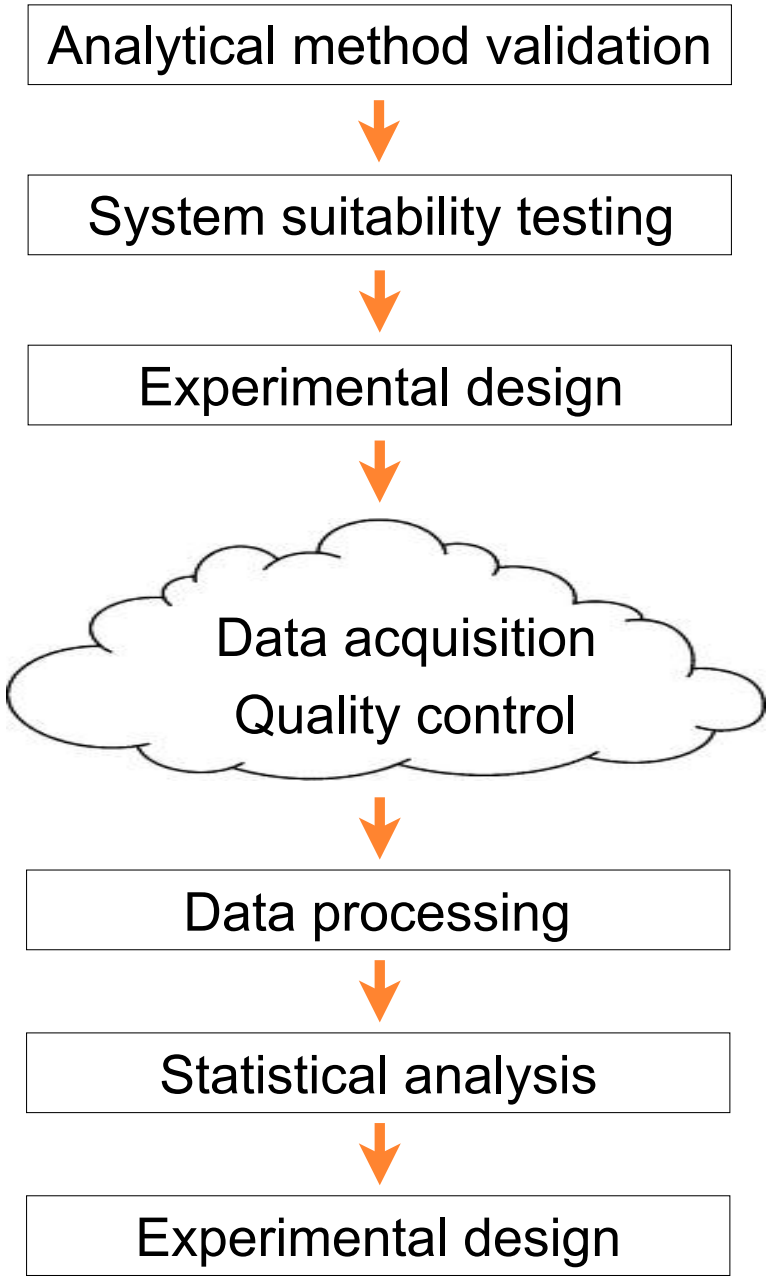
WHY STATISTICS?

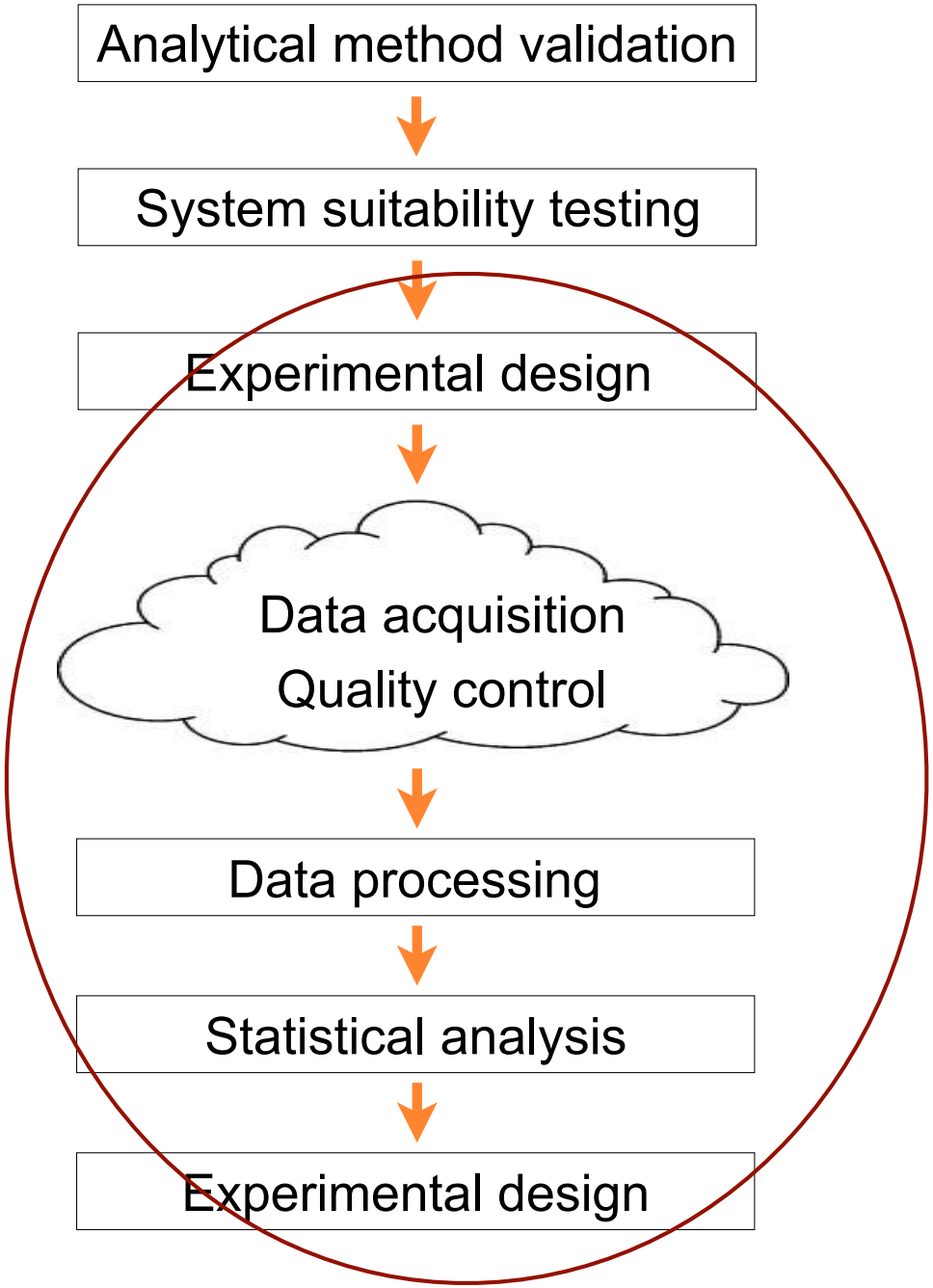
- Variation and uncertainty are unavoidable
  - *Technical variation:* sampling handling, storage, processing
  - *Instrumental variation:* matrix effects, ion suppression
  - *Signal processing:* peak boundaries, identity, intensity
  - *Biological variation:* variation in protein abundance
- Overall goal: effective, reproducible research



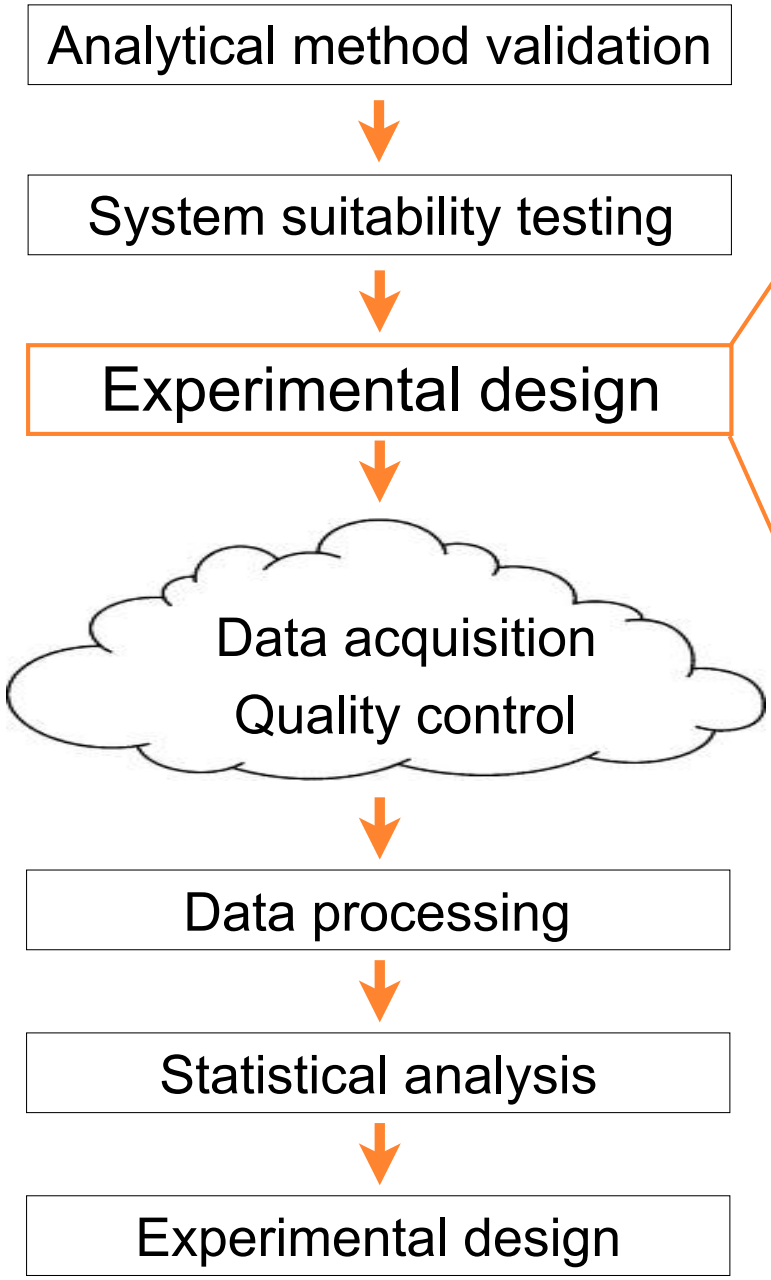
MSI EXPERIMENT: STATISTICIAN'S VIEW



MSI EXPERIMENT: STATISTICIAN'S VIEW



MSI EXPERIMENT: STATISTICIAN'S VIEW



Study goals

- Image segmentation
- Biomarker discovery
- Differential analysis

Biological aspects

- Selection of conditions
- Selection of replicates

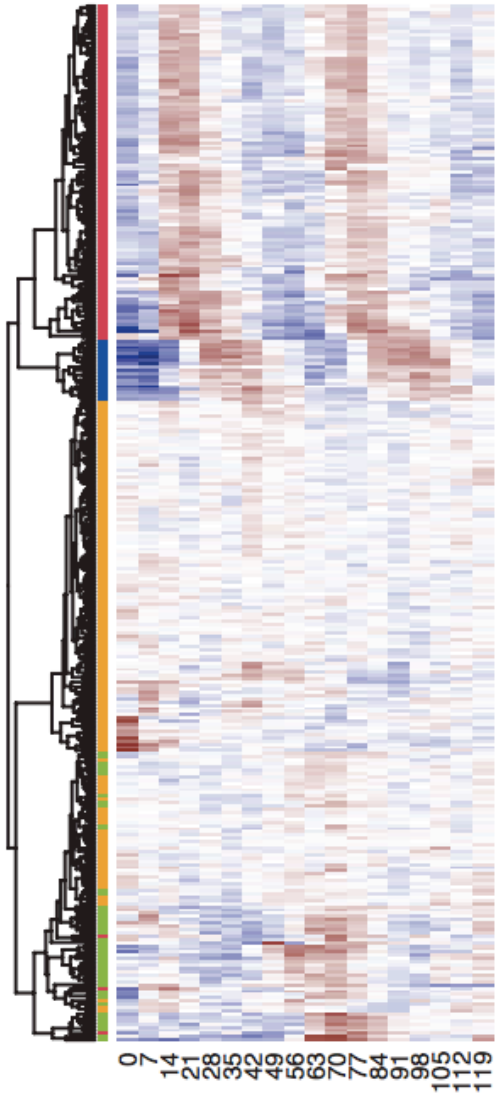
Technological aspects

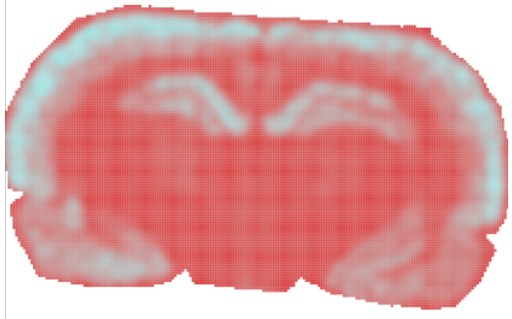
- Sample prep
- Data acquisition
- Randomization

STATISTICAL GOAL 1: CLASS DISCOVERY

Discover analytes or subjects with similar patterns

- No known class labels
  - E.g., no 'healthy' or 'disease'
  - No error rates
- Can't find something meaningful if unsure what we look for
  - Best used for visualization

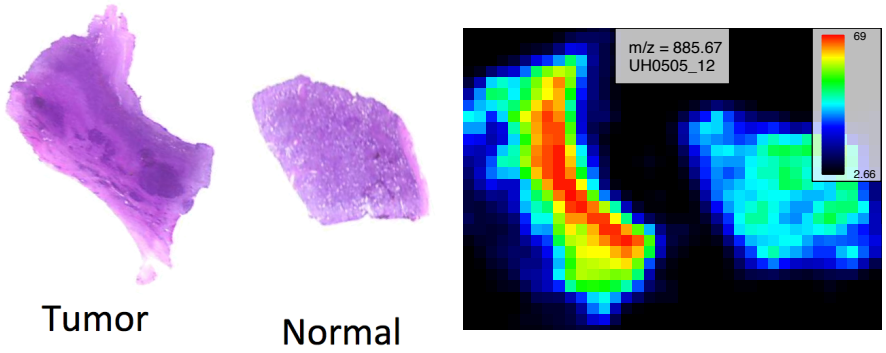


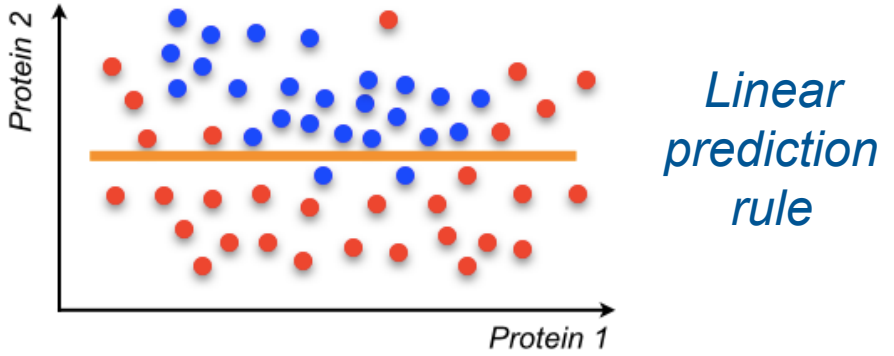


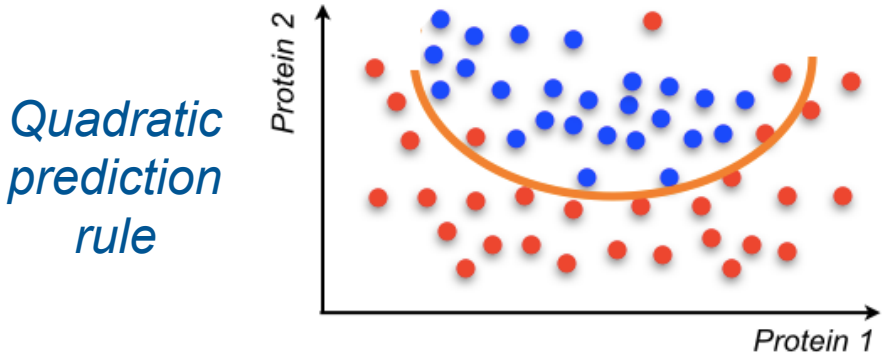
STATISTICAL GOAL 2: CLASS PREDICTION

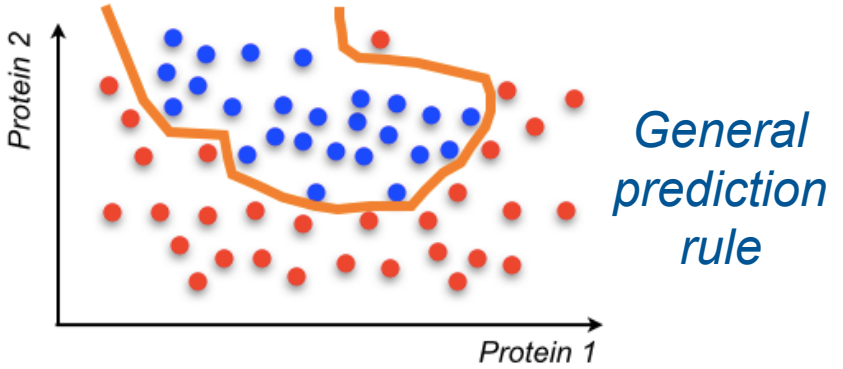
Classify each subject into a known group

- Known class labels
  - Classify pixels/subjects
  - Report misclassification error (sensitivity, specificity…)
- Required for biomarker discovery
  - Distinct training/validation set





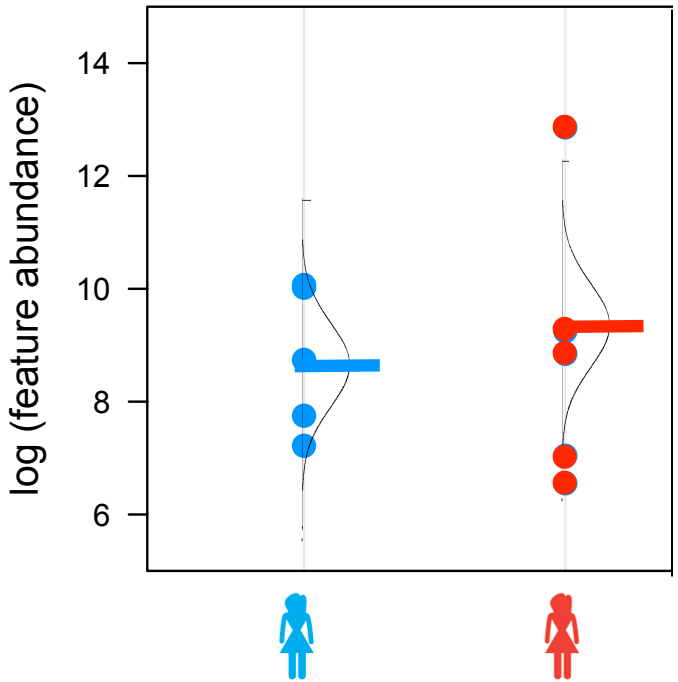


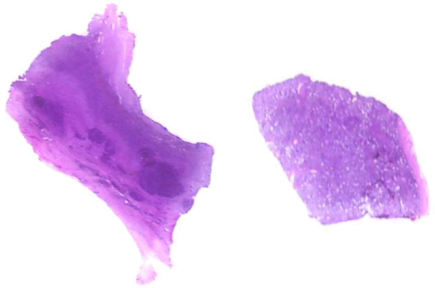


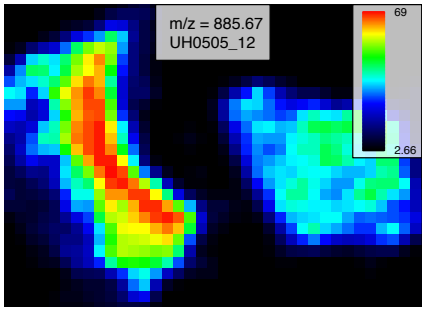
STATISTICAL GOAL 3: CLASS COMPARISON

Compare analyte abundances across locations or subjects from known groups

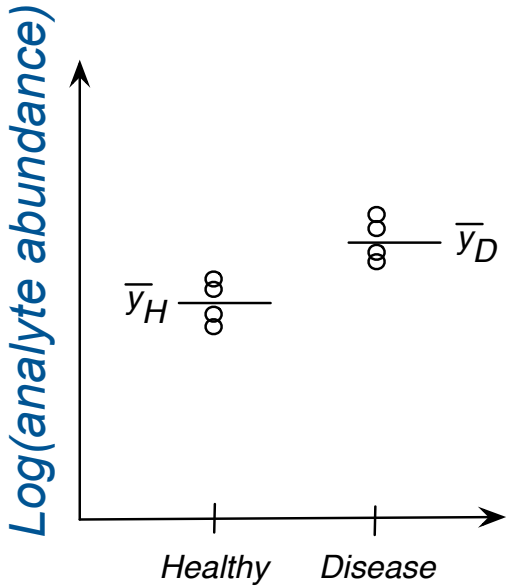
- Known class labels
  - Which analytes co-localize with pre-defined conditions?
  - Report p-values, posterior probabilities etc
  - Control FDR
- Useful in
  - Exploratory clinical studies
  - Basic biology studies



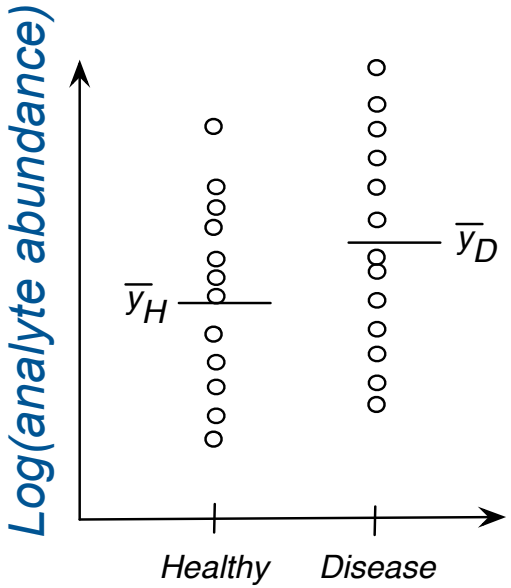




DIFFERENTIALLY ABUNDANT ANALYTES ARE NOT ALWAYS BIOMARKERS



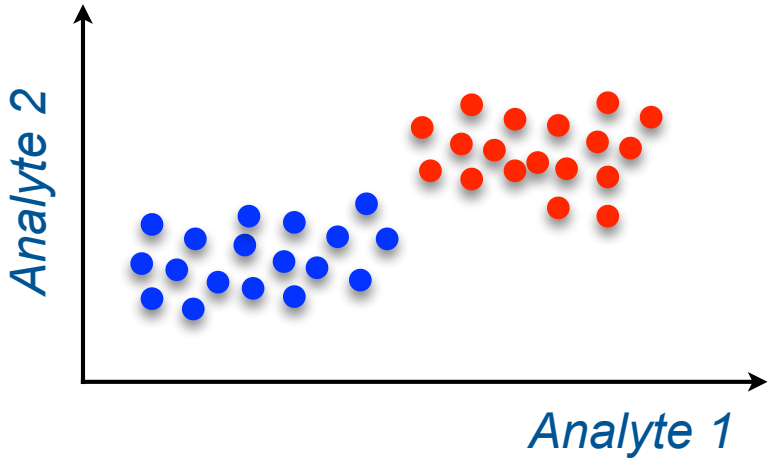
Differentially abundant and predictive



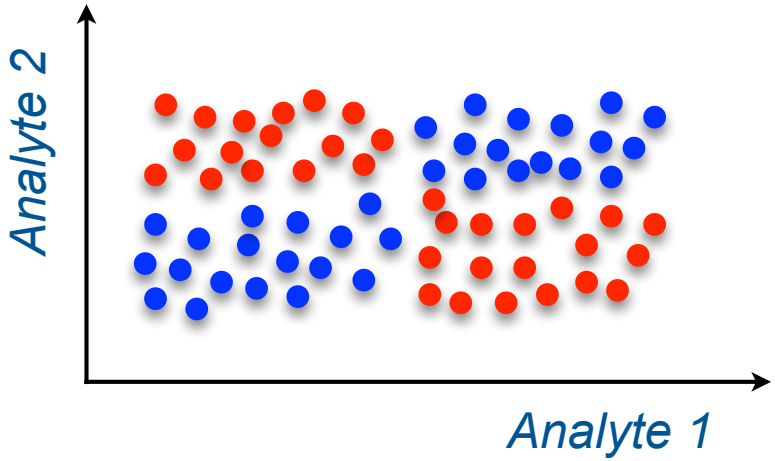
Differentially abundant and not predictive

Single analyte: *Differentially abundant* → ✗ *Predictive*

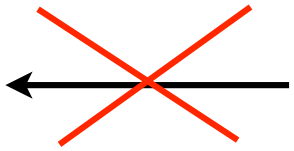
BIOMARKERS ARE NOT ALWAYS DIFFERENTIALLY ABUNDANT



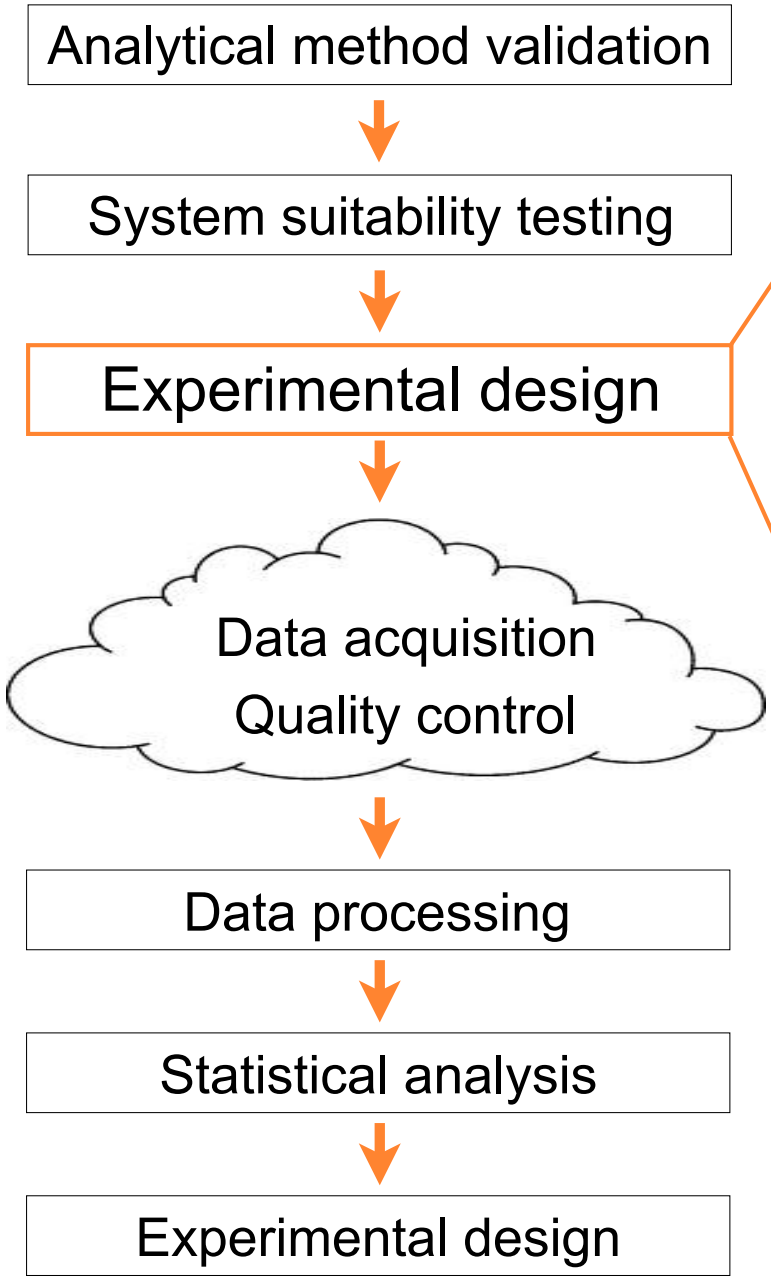
Differentially abundant and predictive



Not differentially abundant but predictive

Single analyte: *Differentially abundant*  *Predictive*

MSI EXPERIMENT: STATISTICIAN'S VIEW



Study goals

- Image segmentation
- Biomarker discovery
- Differential analysis

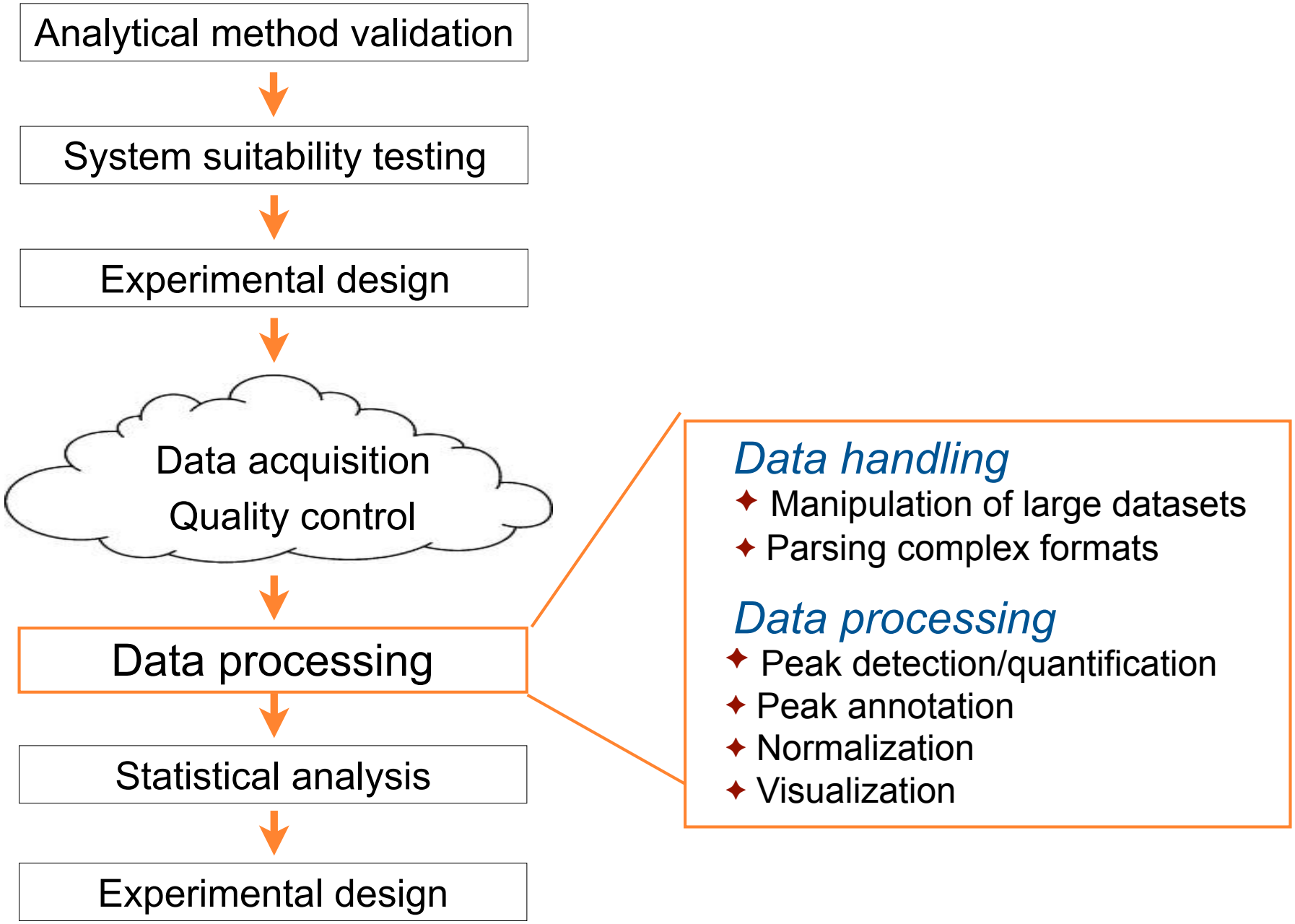
Biological aspects

- Selection of conditions
- Biological replicates

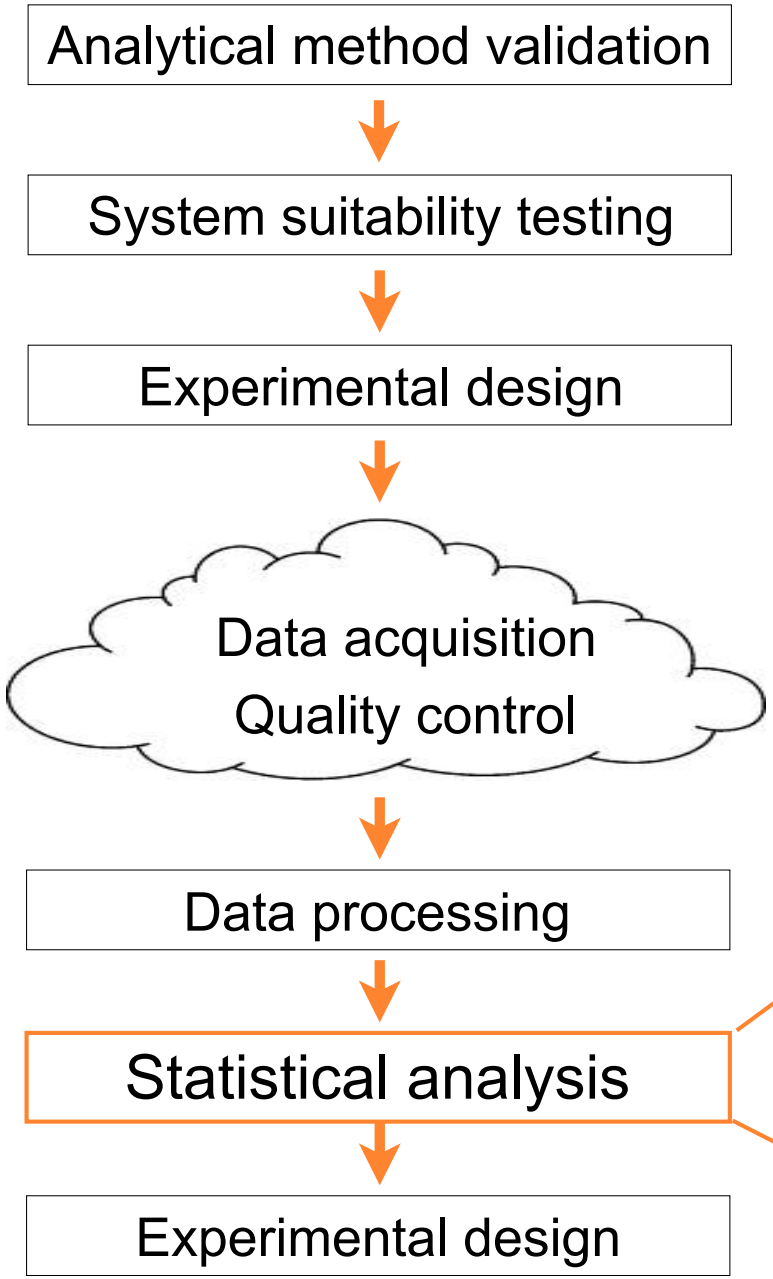
Technological aspects

- Sample prep
- Data acquisition
- Randomization

MSI EXPERIMENT: STATISTICIAN'S VIEW



MSI EXPERIMENT: STATISTICIAN'S VIEW

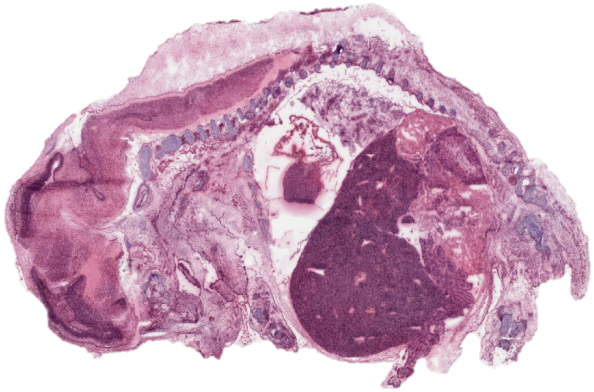


Good statistical practice:

- Match sources of variation
- Match experimental design
- Match research goal
- Follow pre-defined protocol

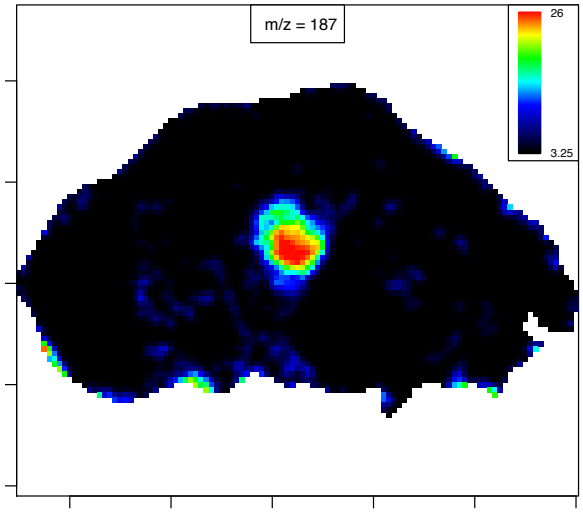
STATISTICAL GOAL 1: CLASS DISCOVERY

Discover analytes or subjects with similar patterns

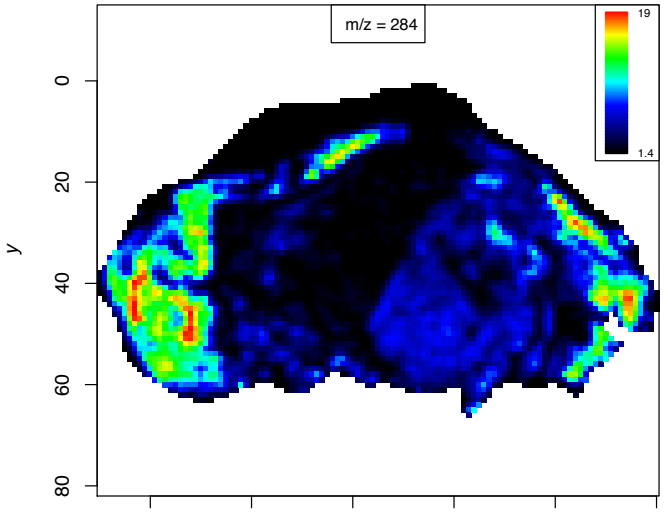


Optical image

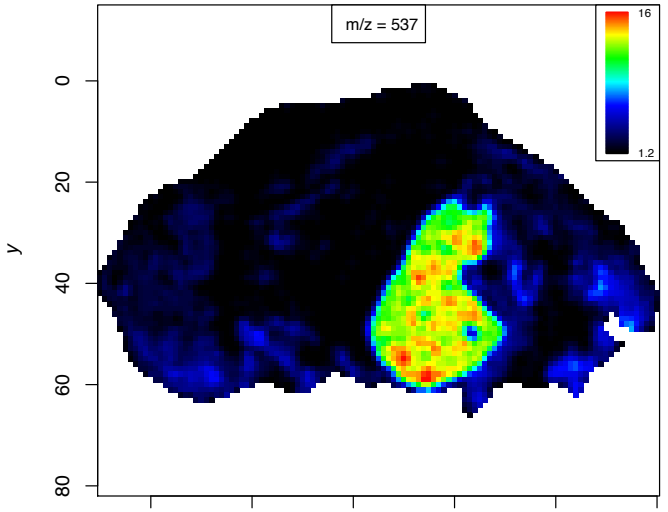
- Pig fetus section
- 4959 pixels
- 10,200 spectral features



Peak associated with heart



Peak associated with brain

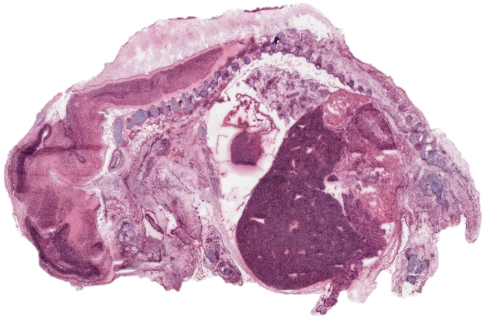


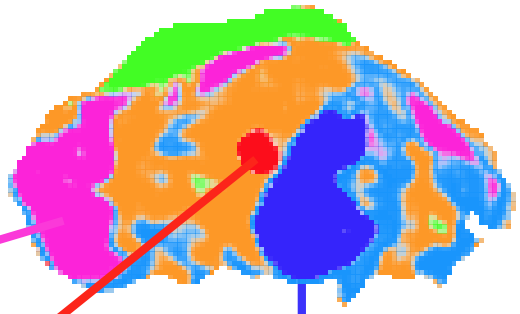
Peak associated with liver

SPATIAL SHRUNKEN CENTROIDS

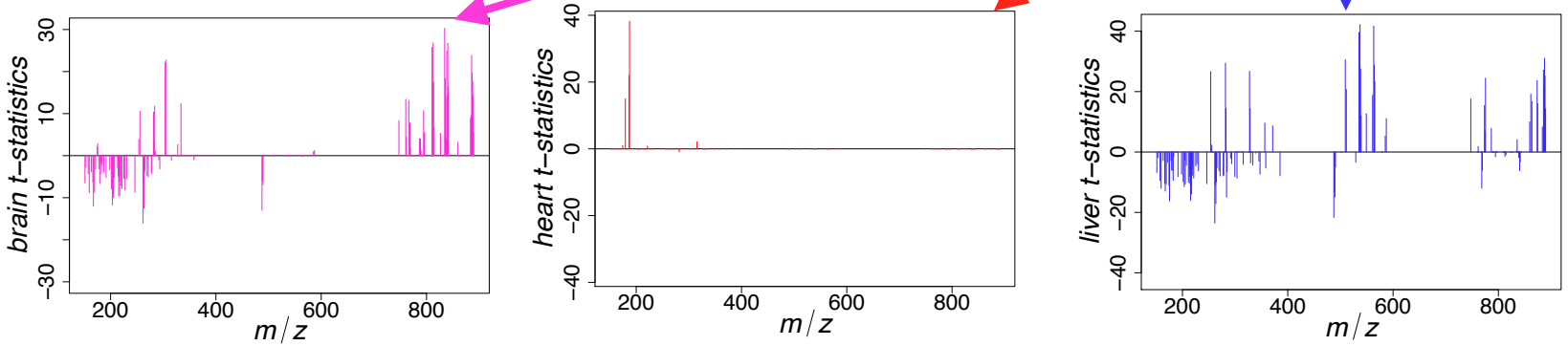
spatially-aware classification/segmentation with feature selection

- Combines spatial information & feature selection
  - Spatially-aware distance from *spatially-aware clustering* (Alexandrov and Kobarg, 2011)
  - Statistical regularization from *nearest shrunken centroids* (Tibshirani, Hastie, et al., 2013)
- Improved image classification & segmentation
  - Data-driven selection of appropriate number of segments
  - Selects most important ions for distinguishing class/segment
  - Probability model characterizes uncertainty





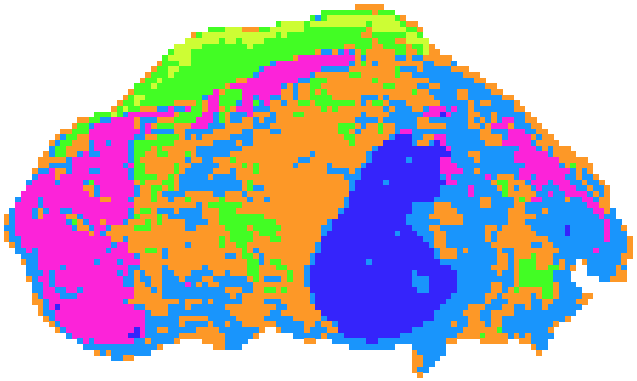
t-statistics show important ions for the brain, heart, and liver segments



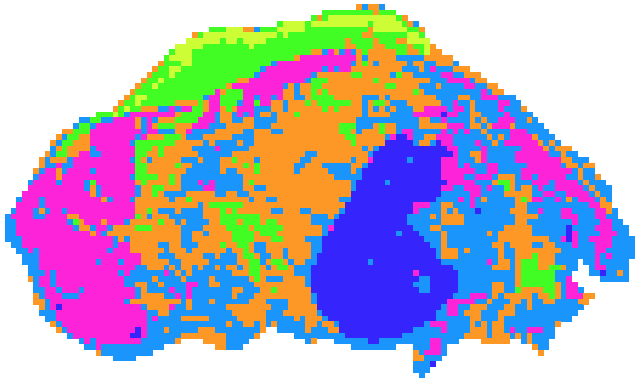
- K. Bemis, A. Harry, L. S. Eberlin, C. Ferreira, S. M. van de Ven, P. Mallick, M. Stolowitz, O. Vitek. "Probabilistic segmentation of mass spectrometry images helps select important ions and characterize confidence in the resulting segments ". Molecular & Cellular Proteomics, 2016

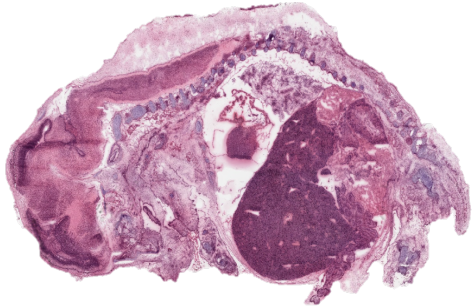
IMPROVED SEGMENTATION
from statistical regularization and spatial information

K–means

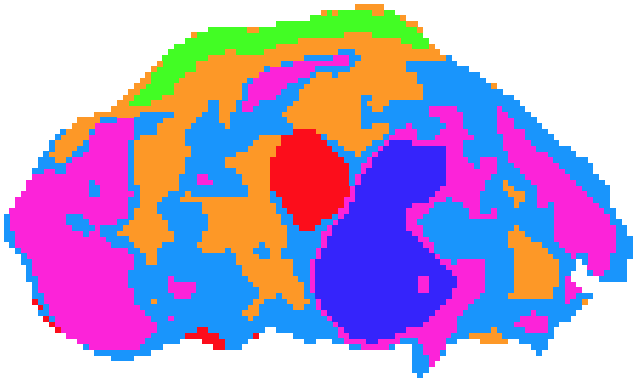


PCA + K–means

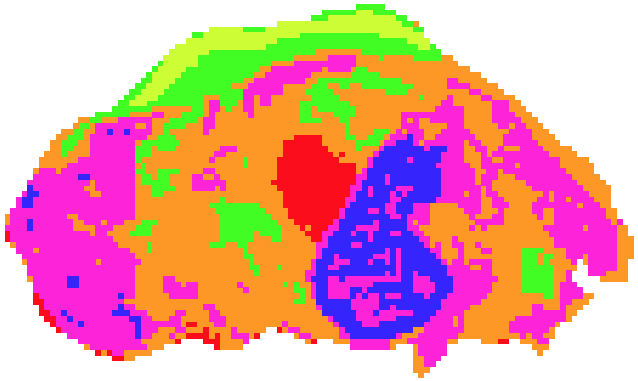




SA + K–means



SASA + K–means



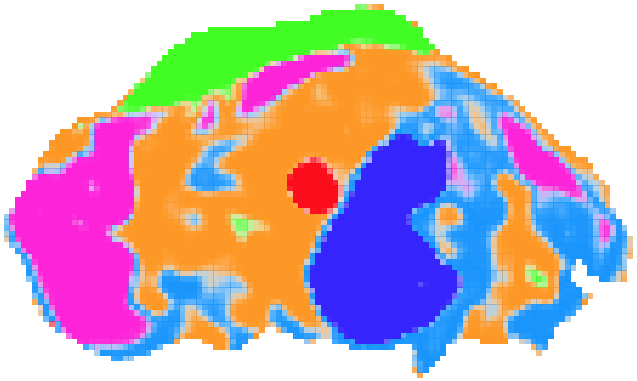
Alexandrov & Kobarg, Bioinformatics, 2011

r=2, k=6

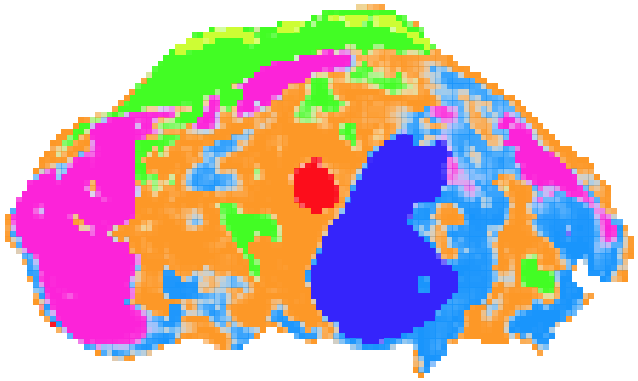
SA=Spatially Aware

SASA=Spatially Aware Structurally Adaptive

SA + Shrunken Centroids



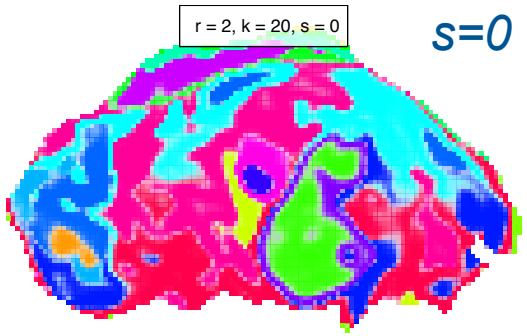
SASA + Shrunken Centroids



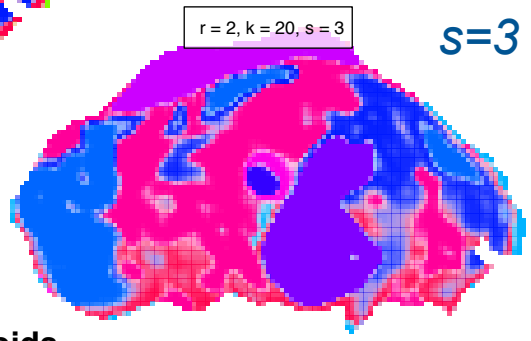
Spatial shrunken centroids

r=2, k=20, s=6
6 segments

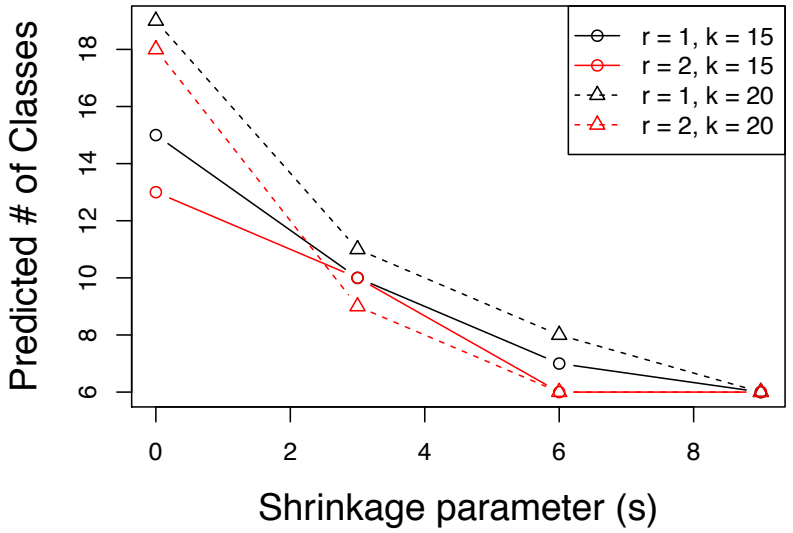
DATA-DRIVEN MODEL SELECTION
for unsupervised experiments through statistical regularization



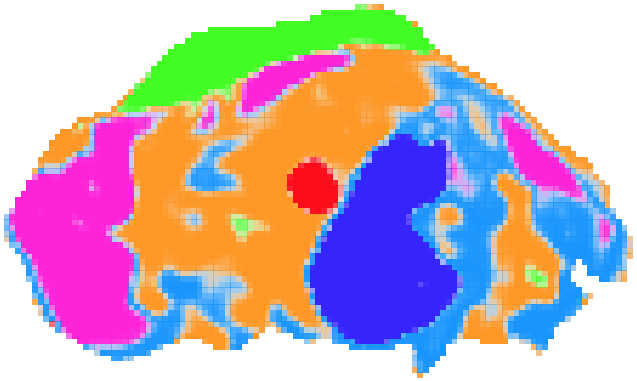
s=0



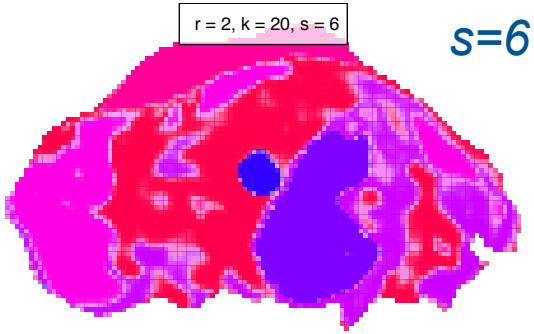
s=3



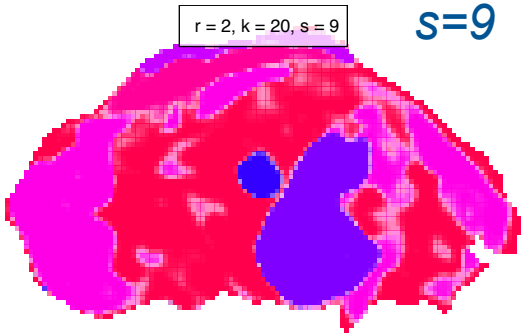
SA + Shrunken Centroids



r=2, k=20, s=6
6 segments



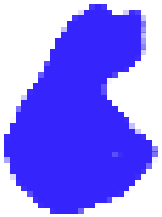
s=6

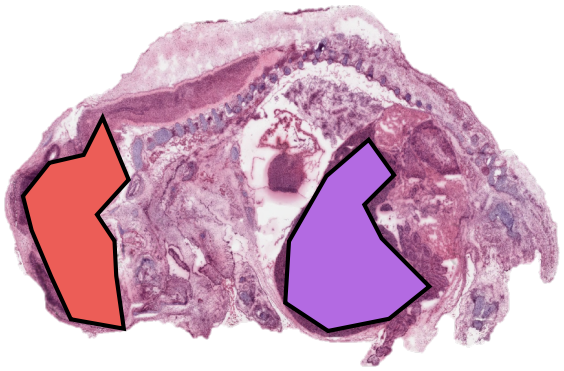


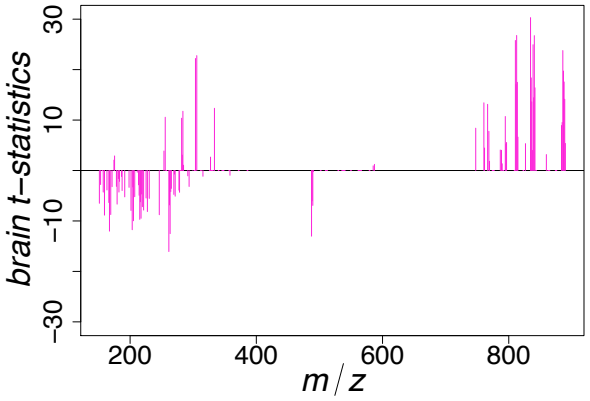
s=9

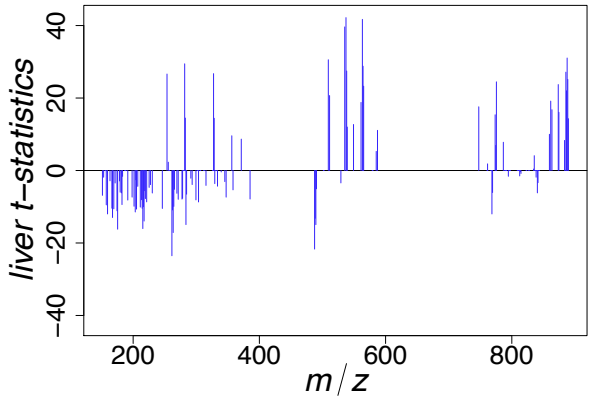
Empirical relationship exists between sparsity in the # of features and # of segments

SELECTION OF MOLECULAR FEATURES
that distinguish each segment for improved interpretability

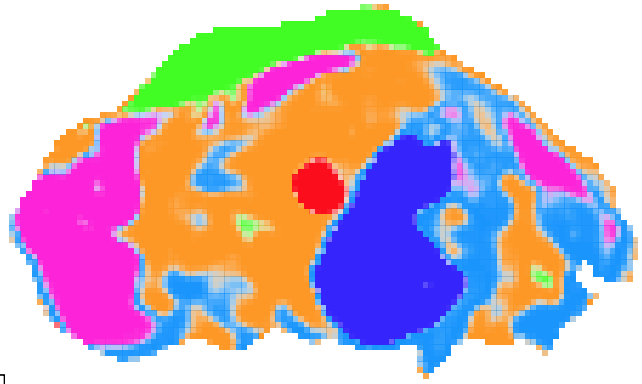


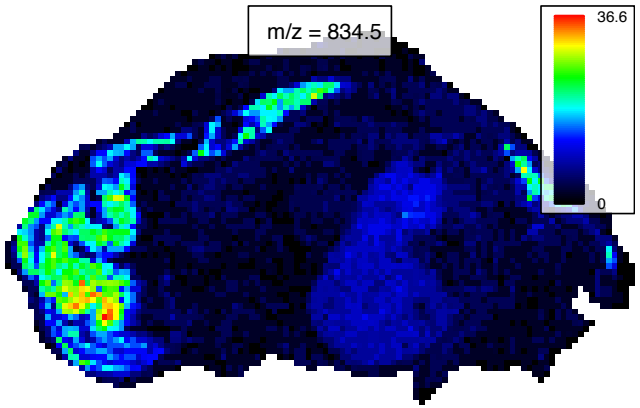


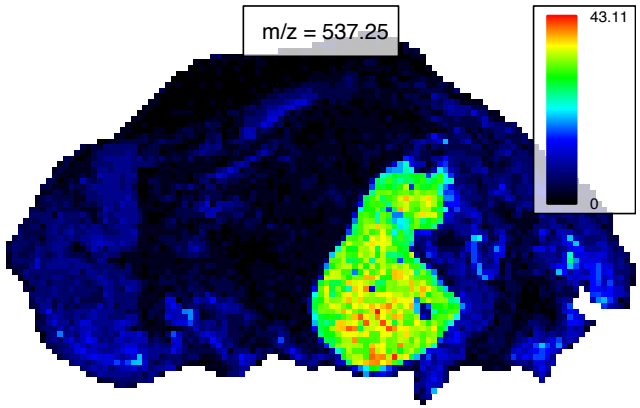




SA + Shrunken Centroids







r=2, s=6, k=20
6 segments

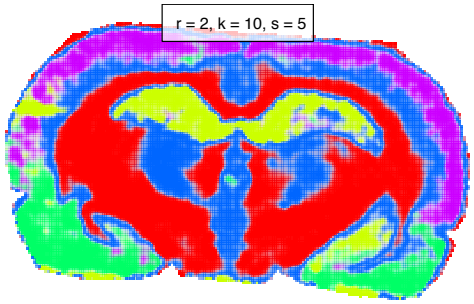
VISUALIZE UNCERTAINTY

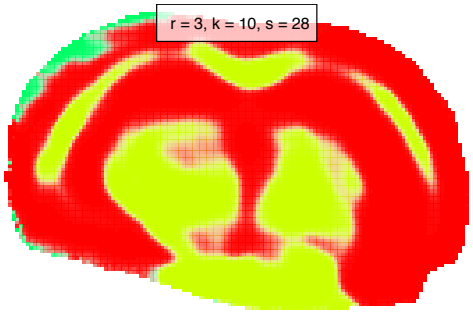
probabilistic model characterizes uncertainty in segmentation

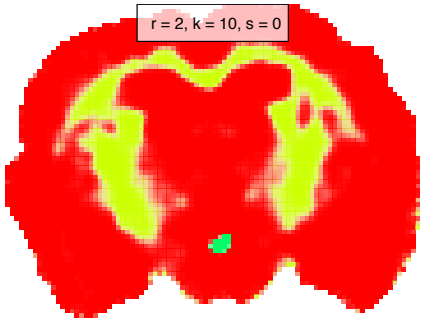
Low noise (MALDI rat) | **Medium noise (DESI mouse)** | **High noise (MALDI mouse)**

Optimal # of segments

Optimal sparsity

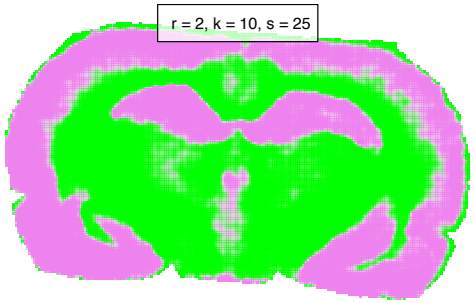


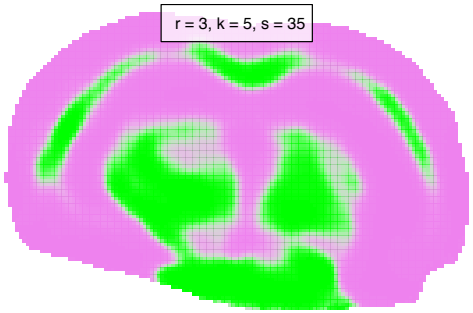


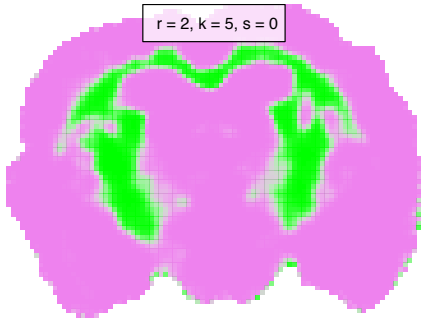


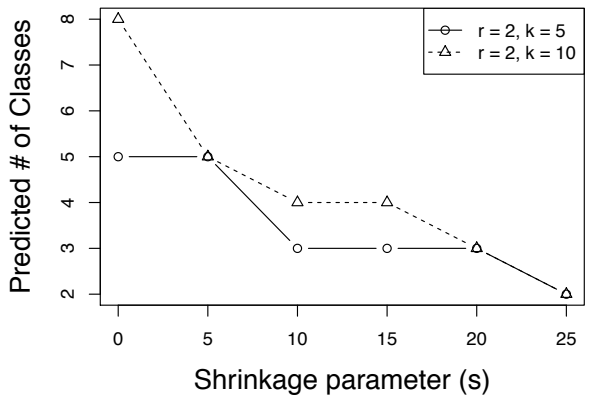
Reduced to 2 segments

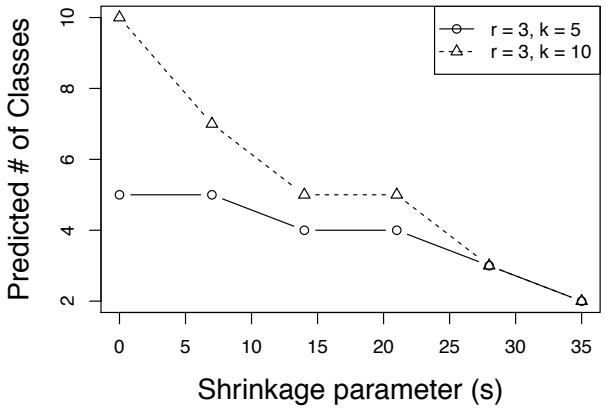
Higher sparsity

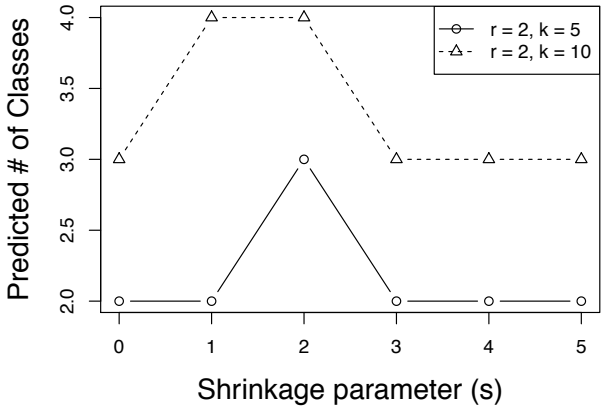










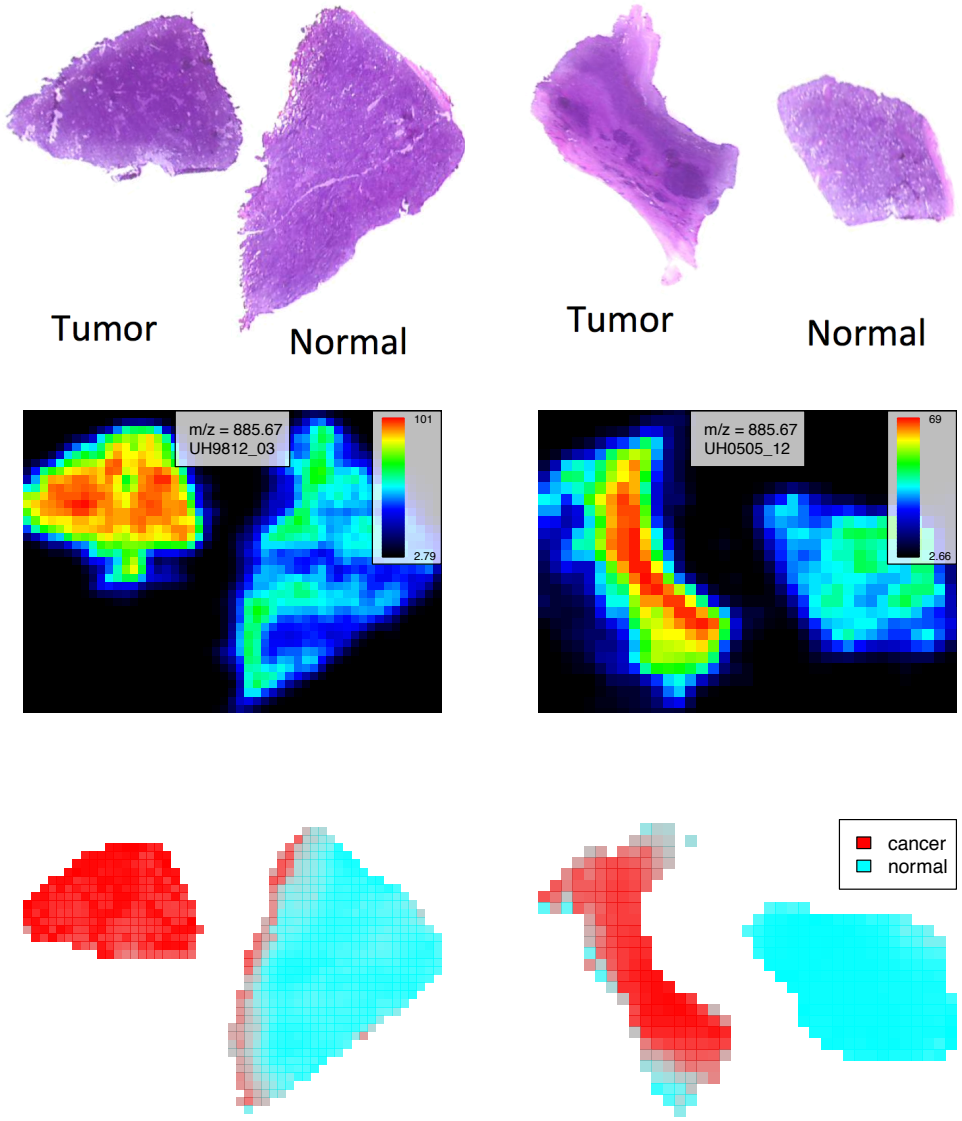


STATISTICAL GOAL 2: CLASS PREDICTION

Classify each subject into a known group

A. L. Dill et all Analytical and Bioanalytical Chem., 2010.

Renal cell carcinoma (RCC)



r=3, s=20

Selected by cross-validation