Scientific computation and anomaly recognition

Chen Zhang [1,2]

[1]RTL CRI_SZ UIH @ Shenzhen, [2]PMJL RUJC-CRI UIH @ Shanghai

Git URL: CubicZebra

UNITED IMAGING

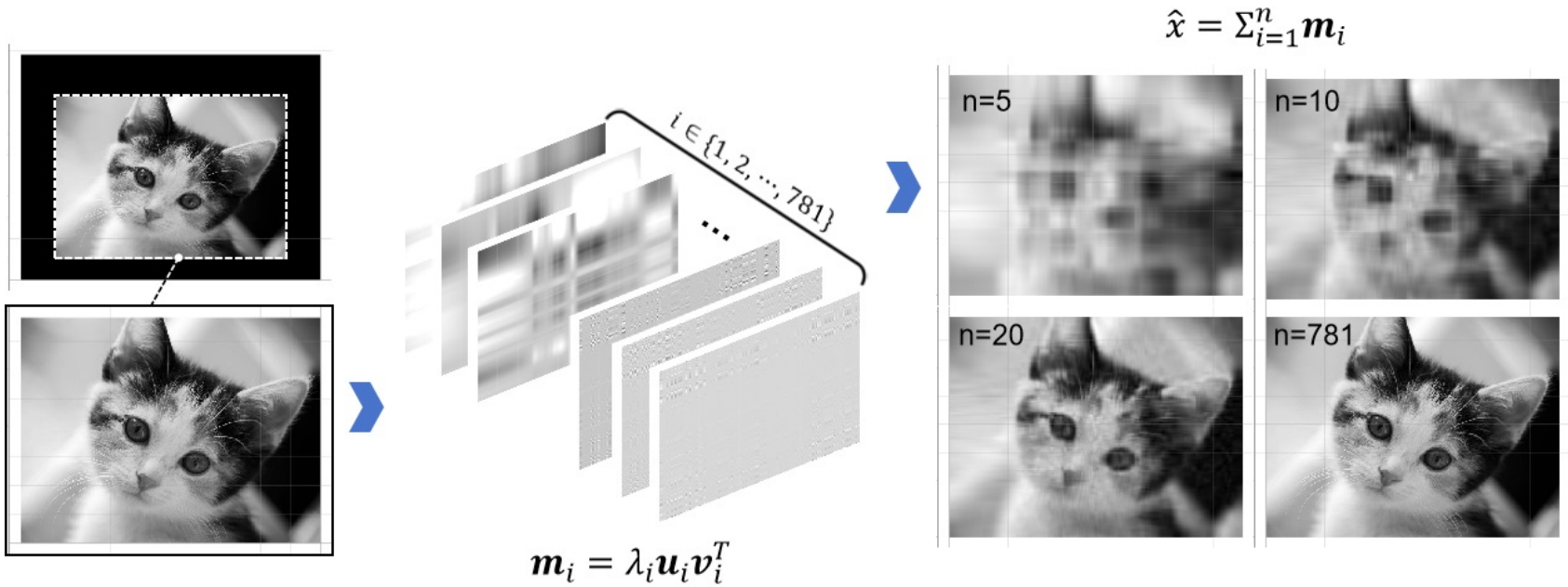
Essence: utilizing the valuable pattern from data

- Program in CS:
  data (digital) + algorithms (business)
- Data mining: (purification, critical)
  data (conceptual) + algorithms (simplication)
- AI application: (utilization)
  data (representation) + algorithms (criterion)

UNITED IMAGING

Data mining

- Pattern: minimal representation of data[1]



[1] Pattern in high dimensional data

AI application

- Mathematical statistics:

  hypotest, ANOVA, Bayesian stats, statistical learning, ...
- Machine learning:
  - analytical: PCA, SVM, conjugate gradient descent, ...
  - randomness: RF, ensemble, stochastic gradient descent, ...
- Deep learning:
  - Conv+Pool (+randomness & variation, data aug & mining)
  - MLP ($p = f(\boldsymbol{x}) \rightarrow f$, set criterion of regressor)

UNITED IMAGING

Scientific computation: base implementation of algorithms

- Signal decomposition
  - eigen and sigular value decomp.,
  - tensor decomp. & synthesis (CP, tucker, train, ring);
- Optimization
  - linear or nonlinear OLS,
  - gradient descent: (quasi-)newton iter, conjugate, stochastic,
  - stochastic process;
- Statistics
  - inference: MLE, Bayes (posteriori & prediction dis),
  - measurement: univariate/multivariate hypothesis tests,
  - sampling: MH, MCMC, stat simulation;
- Linear algebra, ODR, PDE, ...

UNITED IMAGING

For rational AI approaches

- Comprehension on algorithm frame
  - What: base components of scientific computation methods,
  - Why: reason of frame architecture design,
  - How: concrete computation steps of each component;
- Modification on algorithm frame
  - Parameter: inherent adaption support of frame,
  - Component: change features of frame for certain purpose;
- Creating customized frame
  - Analysis: data => concerned info => sc methods,
  - Architect: sc implements => atomic ops => pipe;

UNITED IMAGING

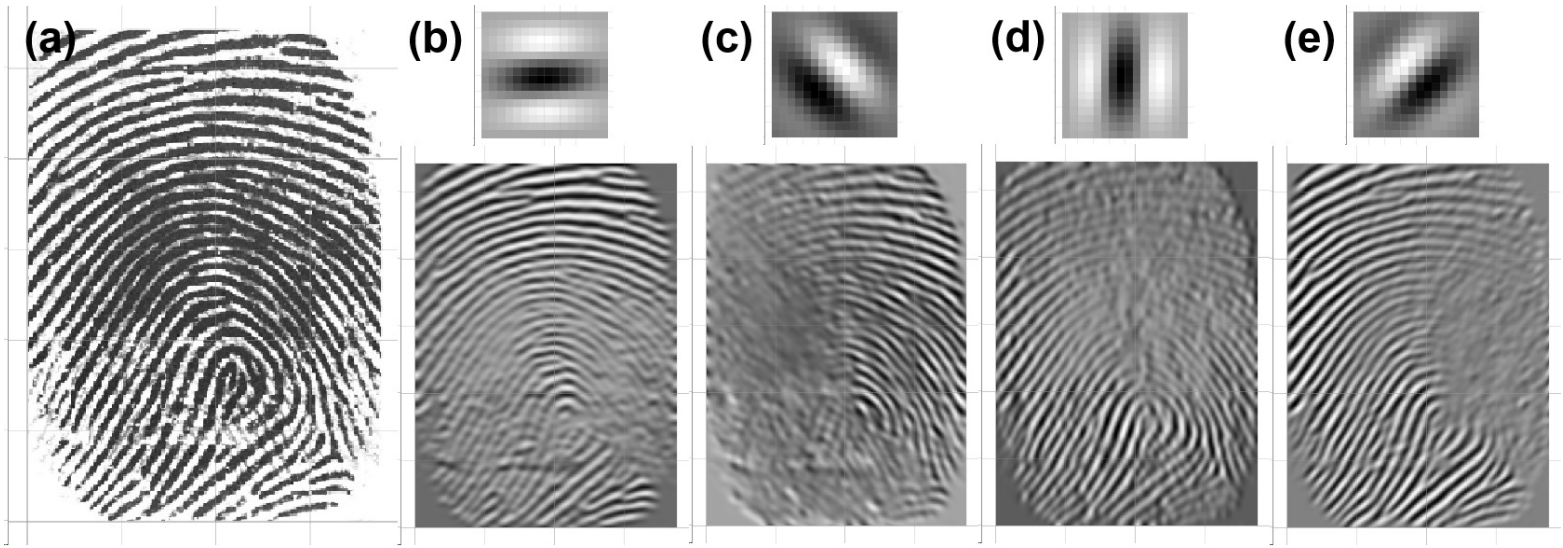
For rational AI approaches

Two suggestions in practice:

1. **Deep into principles:** despite autonomous vehicles, a driver inside should have license at least.
2. **Don't against the current:** modification/design based on *concerned info* and *task objective*.

The objective-oriented methodology

- Case 1: fingerprint recognition[2]
  - interested info: pattern of texture
  - sc methods: spatial gabor filtering (2D)

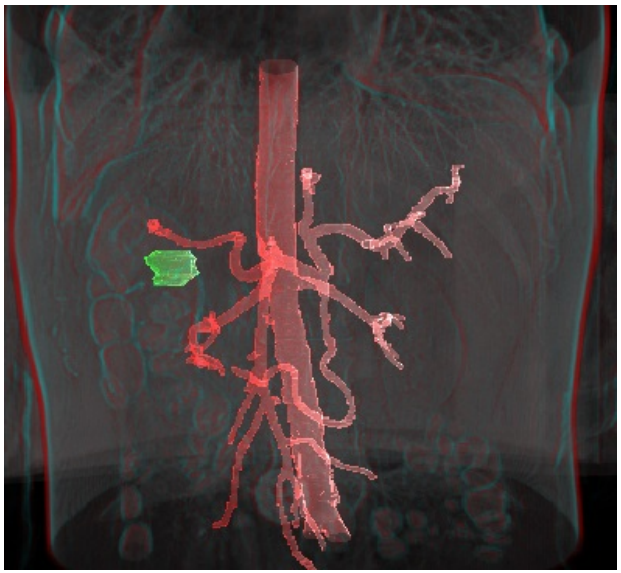


UNITED IMAGING

[2] Determine appropriate transformation

The objective-oriented methodology

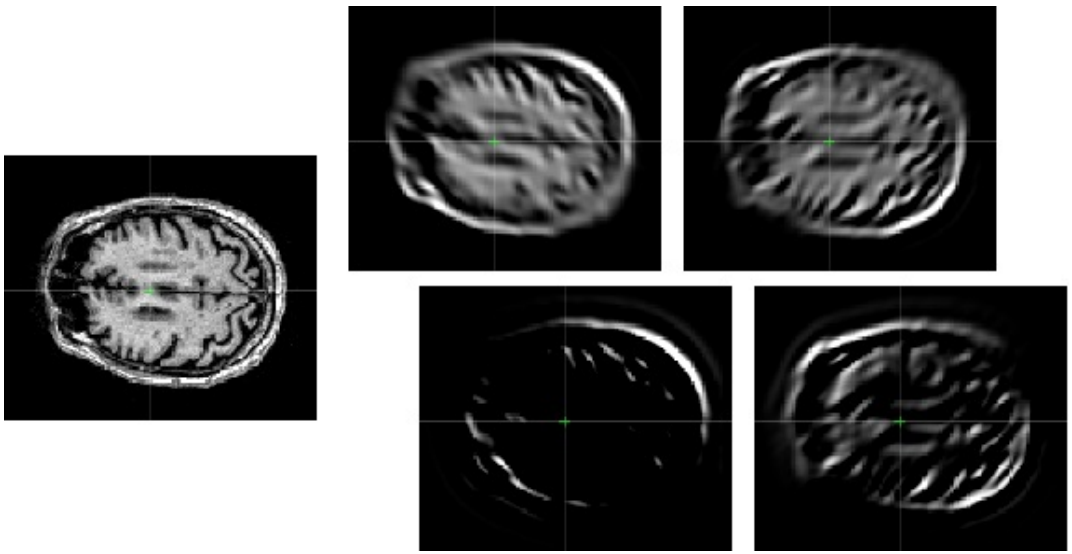
- Case 2: vessel measurement in pancreatic carcinoma study
  - interested info: space angle
  - sc methods: morphological ops + linalg computation



UNITED IMAGING

The objective-oriented methodology

- Case 3: brain-related study (pipe reusability in Case 1)
  - interested info: texture pattern of brain
  - sc methods: spatial gabor filtering (3D)



UNITED IMAGING

Primer concepts in statistics

- Why we need statistics
  - data: listing all acquired observations
  - statistics: description on data via sth. (e.g. dis)
- Samples & population
  - estimation: study on samples => conclusion on population
  - bias: diff between samples & population[3]

UNITED IMAGING

[3] Unbias estimation

Primer concepts in statistics

- Parametric or non-parametric[4]
  - parametric: at least info of a dis., strong assumption,
  - -diff, the statisical illusion (e.g. median/mean)

$A = \{1, 2, 3, 4, 5\}$ and $B = \{1, 2, 3, 4, 5000\}$

Data does not lie. People do.

—Lee Baker, *Truth, Lies & Statistics: How to Lie with Statistics*

UNITED IMAGING

[4] Parametric and non-parametric

Primer concepts in statistics

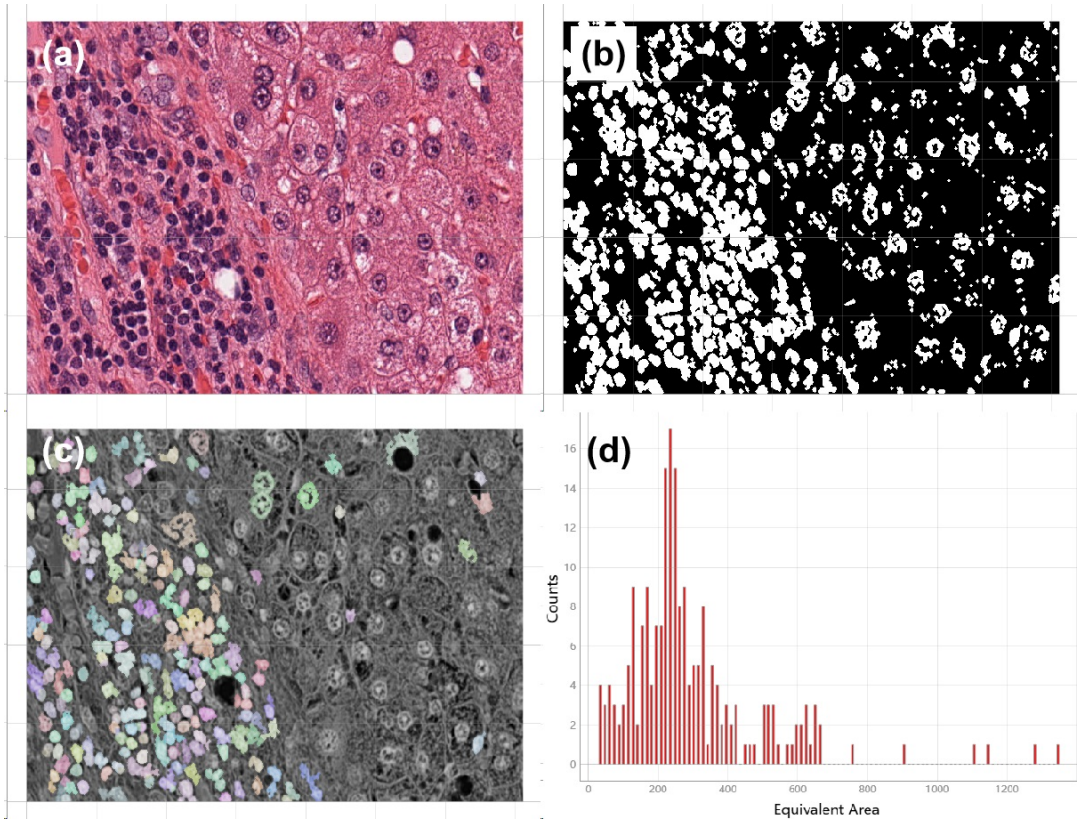
- Consideration of sufficiency
  - def: $f(\boldsymbol{x}|\boldsymbol{\theta}) = g(T(\boldsymbol{X}|\boldsymbol{\theta}))h(\boldsymbol{x})$
  - design intention of `StratifiedKFold` in `scikit-learn`
  - interpretation of bootstrap (CS) in statistics (convergency)

Proposals in managing data practice

- Diagnosis on samples (assume $\boldsymbol{A} \subset \boldsymbol{B} \subset \boldsymbol{C}$)
  - $s(\boldsymbol{A})$ diff from $s(\boldsymbol{B})$ $=>$ increase nums. | test variants,
  - $\boldsymbol{A} \subset \boldsymbol{B} \rightarrow f$ be ineffective from $\boldsymbol{B}$ to $\boldsymbol{C}$ $=>$ data aquisition;
- Applicability of algorithm
  - evaluate applicability from frame to task-objective,
  - modification | design $=>$ parameterization;
- Further works
  - cleaning, preprocessing, training, validation, etc.

Determination of algorithm design

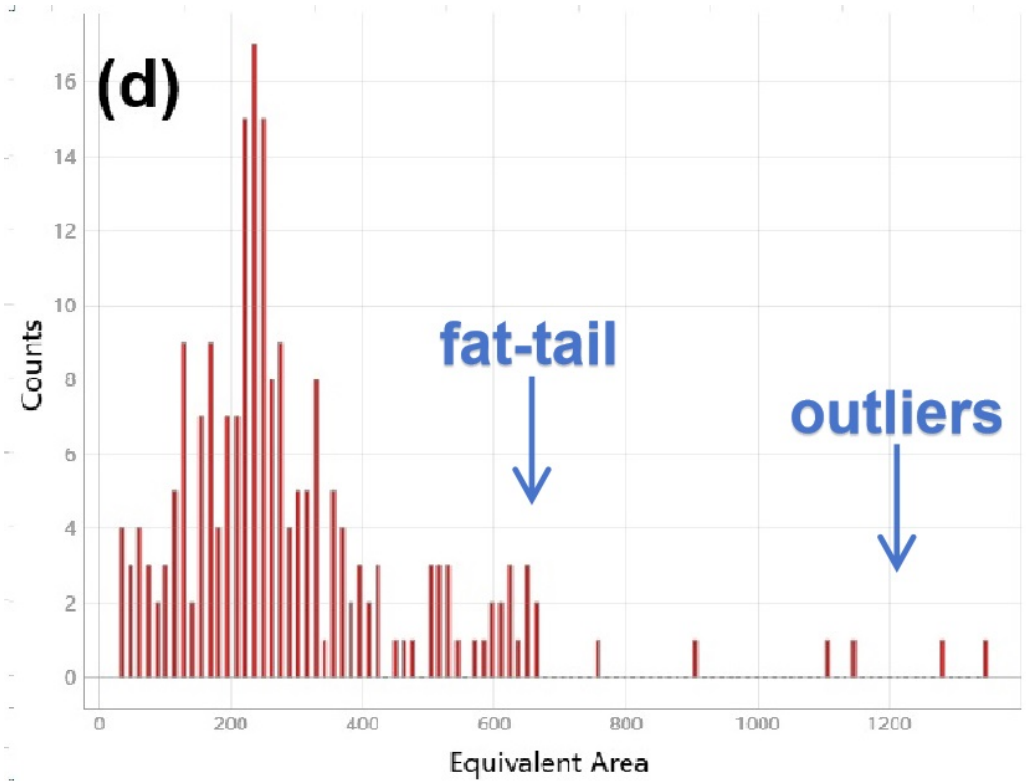
- Case 4: statistics on nuclei of cancer cells
  - objective: $\mathcal{N}(\boldsymbol{x}, \boldsymbol{\Sigma})$ dis of nuclei, statistical approahces



UNITED IMAGING

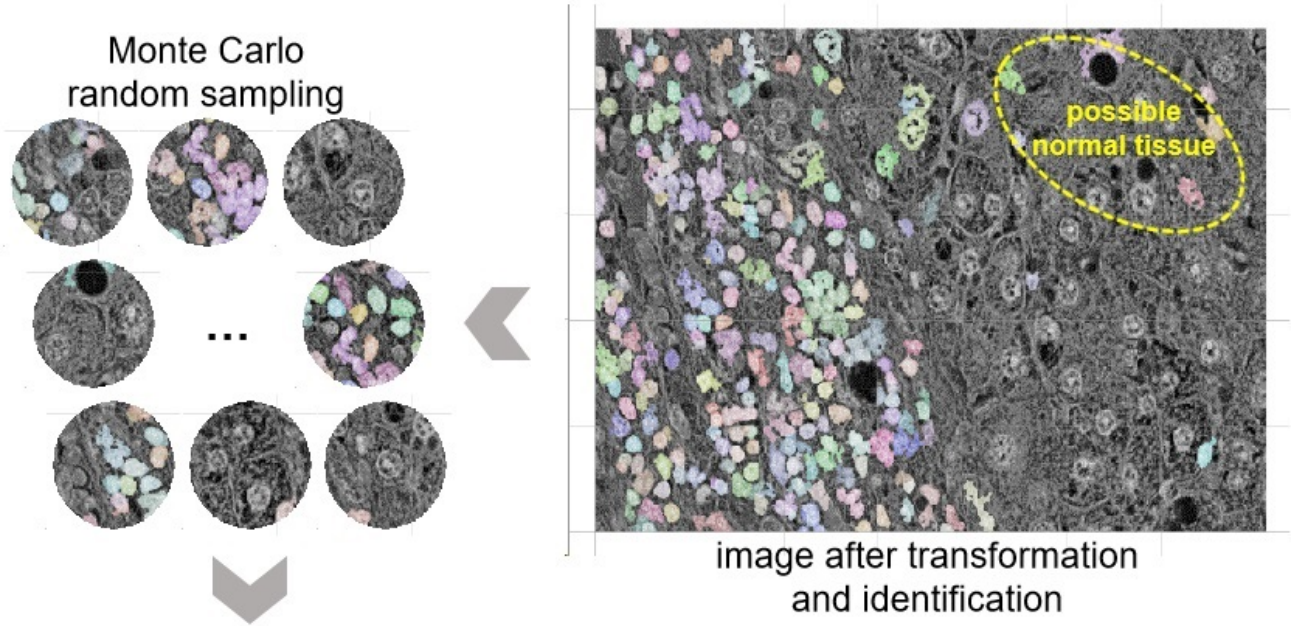
Determination of algorithm design

- Case 4: statistics on nuclei of cancer cells
  - diagnosis on dis => fat-tailed & outliers



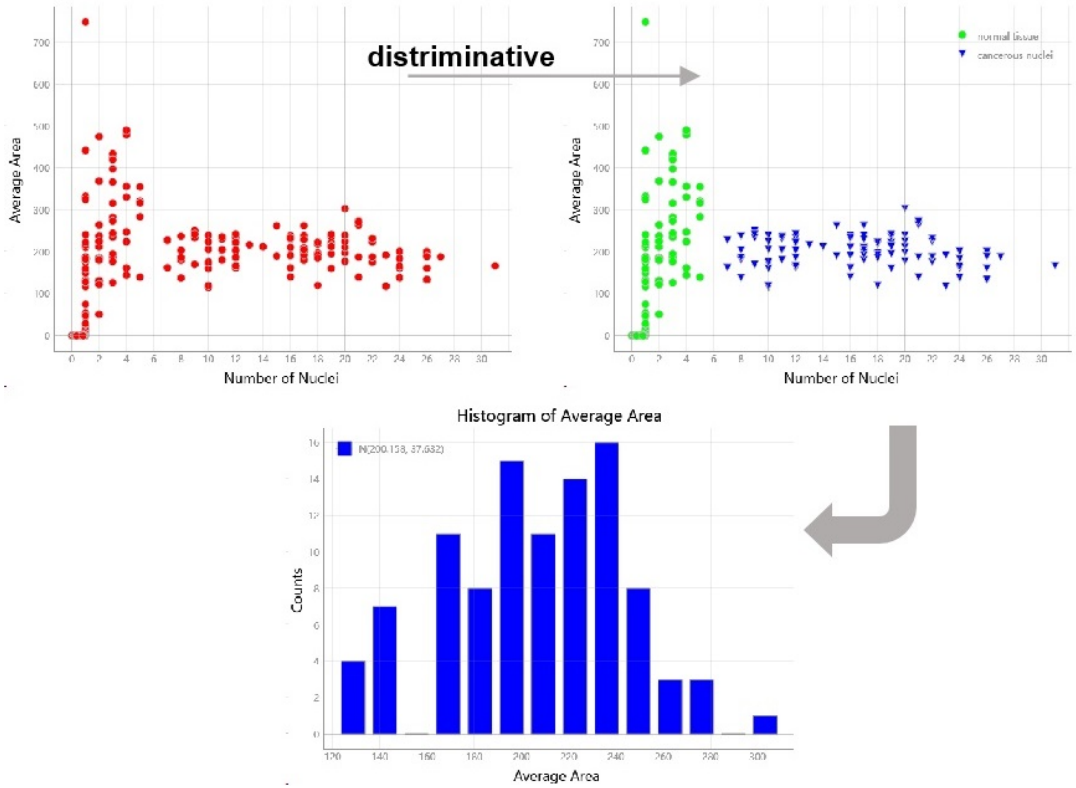
Determination of algorithm design

- Case 4: statistics on nuclei of cancer cells
  - +MCMC aug => params converge in prob



Determination of algorithm design

- Case 4: statistics on nuclei of cancer cells
  - +MCMC aug => params converge in prob



Comb. & App. of multi sc meta implementations

- Dis-related methods (e.g. T2, Naive Bayes)
  - methematical statistics, hypothesis tests
  - linear algebra (supporting multivariate)
  - Bayes statistics and computation
- Data-related methods (e.g. neighbors)
  - data structure (CS), `KDTree` for query
  - optimizations, for numeric solution
  - signal decomposition
  - linear algebra as well
- Other methods...

UNITED IMAGING

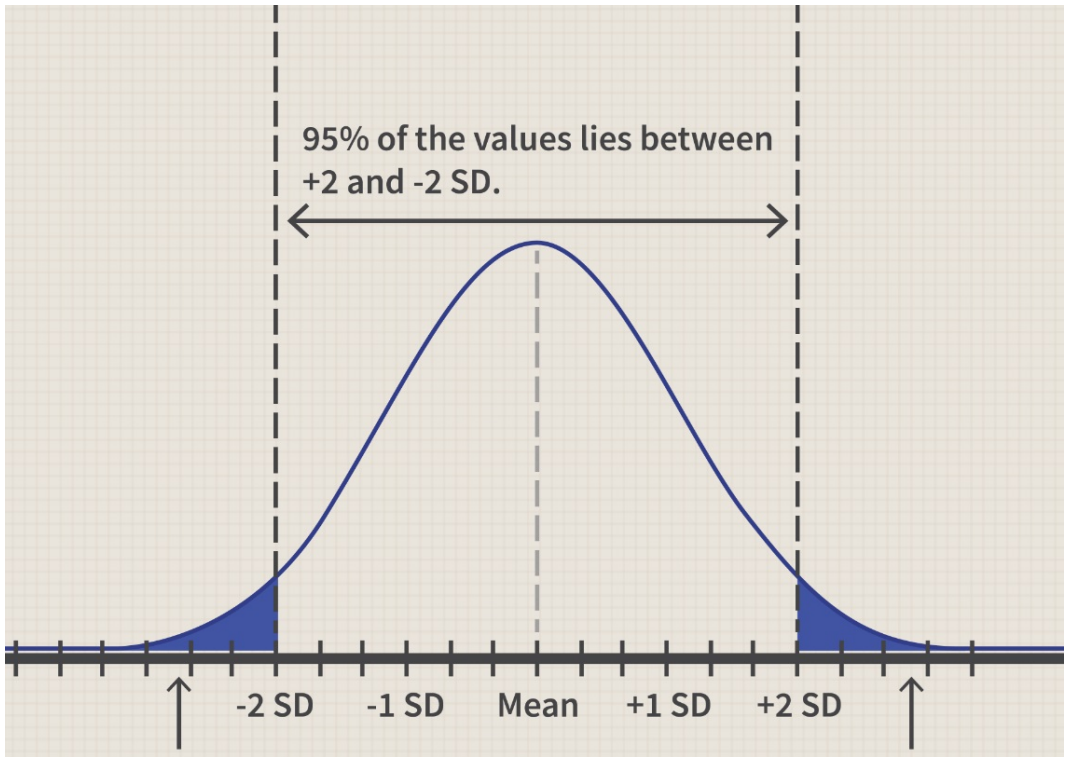
Comb. & App. of multi sc meta implementations

- Security:
  - risky transaction, anti-fraud, ...
- Quality ensurance:
  - quality testing, mechanical fault monitoring, ...
- Networks:
  - spam recog., attack monitoring, ...
- Medical:
  - **less reported**

UNITED IMAGING

Essentials: conventional hypothesis testing

- One-side modeling: stats on one-class $\rightarrow$ normal/anomaly



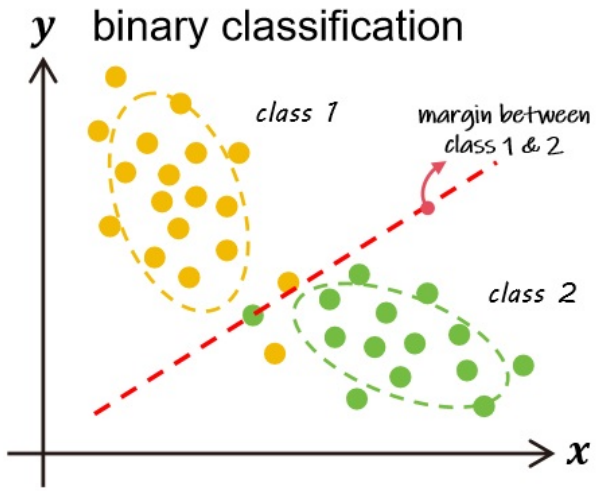
accept/rejection regions[5]

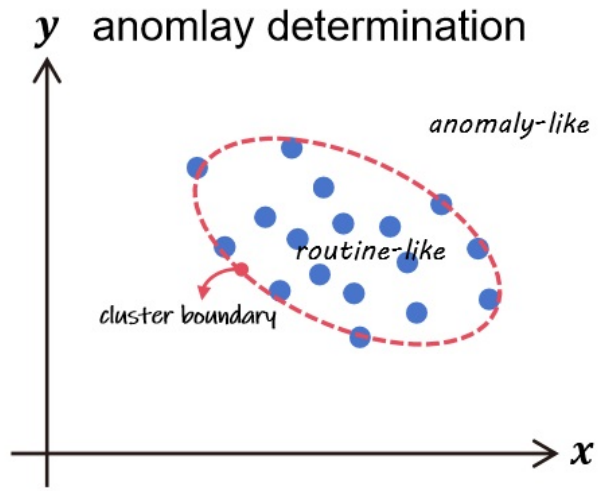
UNITED IMAGING

[5] Hypothesis testing concepts and examples

Essentials: concept and applicable scope

- Anomaly detection[6]:
  - stats on one-class → accept/rejection → routine/anomaly-like





[6] Anomaly and change

Essentials: concept and applicable scope

Why not binary classification?

- **binary classification**
  1. narrow vs. narrow
  2. large data required
  3. low generalization
- **anomaly detection**
  1. narrow vs. generic
  2. low data required
  3. high generalization

Source: limited capability for unknown pattern

UNITED IMAGING

Hotelling T2:[7] multivariate student-T

- **Hypothesis:**
  - null: case is derived from the identical population,
  - alternative: case is not derived from the identical population;
- **T2 statistic:**

$$T^2 = \frac{N-M}{(N+1)M}(\boldsymbol{x}' - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}' - \hat{\boldsymbol{\mu}}) \sim F(M, N-M)$$

- **Criterion:**

$$(\boldsymbol{x}' - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}' - \hat{\boldsymbol{\mu}}) \sim \chi^2(x|M, 1)$$

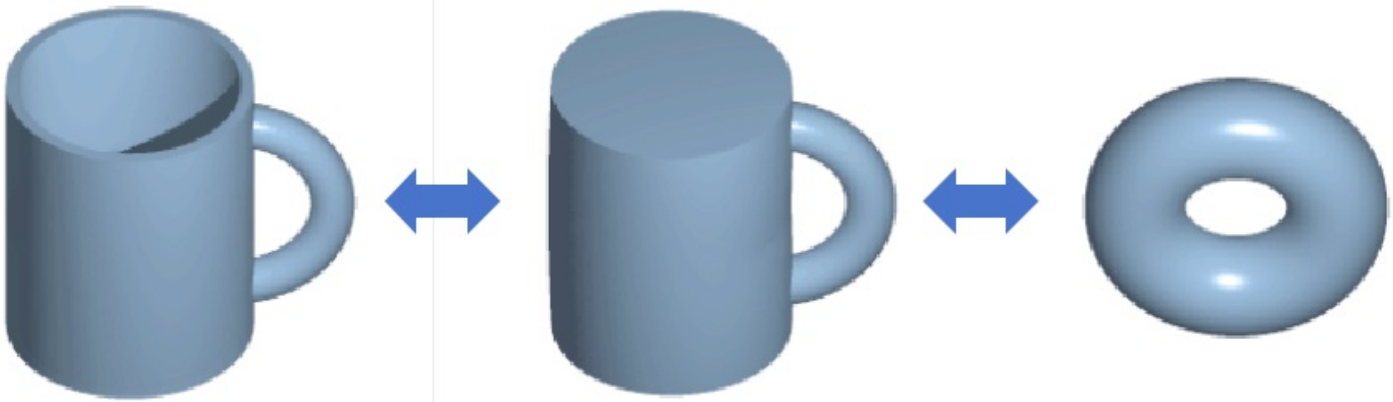
calculation on basis of $\chi^2(x|M, 1)$

UNITED IMAGING

[7] Hotelling T-squared

Large margin nearest neigbors (LMNN)

- Essential concept: *homeomorphism*[8]

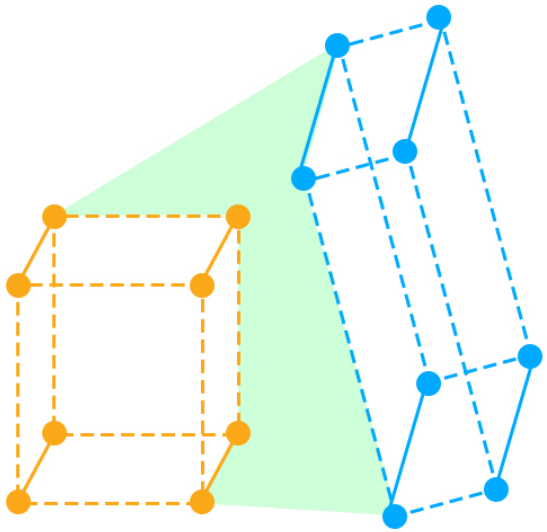


[8] Topological isomorphism

LMNN: role of Riemannian space

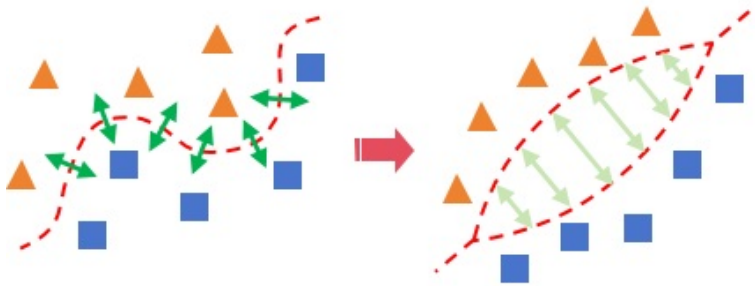
Euclidean (global ops)

$$x' = f(x) = Ax$$



Riemannian (local ops)

$$x' = f(x) = Bx$$



- The determination of B?

UNITED IMAGING

LMNN: problem description[9]

- **Optimization objective**:

$$\Psi(\boldsymbol{R}) = \frac{1}{N}\sum_{c=1}^{s}\sum_{n=1}^{N}\left[w_c \cdot \psi_1^{(n)}(\boldsymbol{R}) + \sum_{m\in\{c\}^C} w_m \cdot \psi_2^{(n)}(\boldsymbol{R})\right]$$

- **Constraints**:

$$\text{s.t. } \boldsymbol{R} \succeq 0$$

- **Supports**:
  `linalg`, matrix decomp., data structure, ...

UNITED IMAGING

[9] Empirical distribution and neighbors

LMNN: optimization & Riemannian space

$$\arg\min_{\boldsymbol{R}} \Psi(\boldsymbol{R}) \rightarrow \boldsymbol{R}^*$$

- matrix decomposition: LDLt

$$\boldsymbol{R}^* = \boldsymbol{L}_m \boldsymbol{\Lambda}_m \boldsymbol{L}_m{}^\top = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top \boldsymbol{L}^\top$$

$$\therefore \boldsymbol{L}' = \boldsymbol{L}\boldsymbol{\Lambda}$$

- Cartesian to Riemannian space:

$$f(\boldsymbol{x}) = \boldsymbol{L}'\boldsymbol{x}$$

UNITED IMAGING

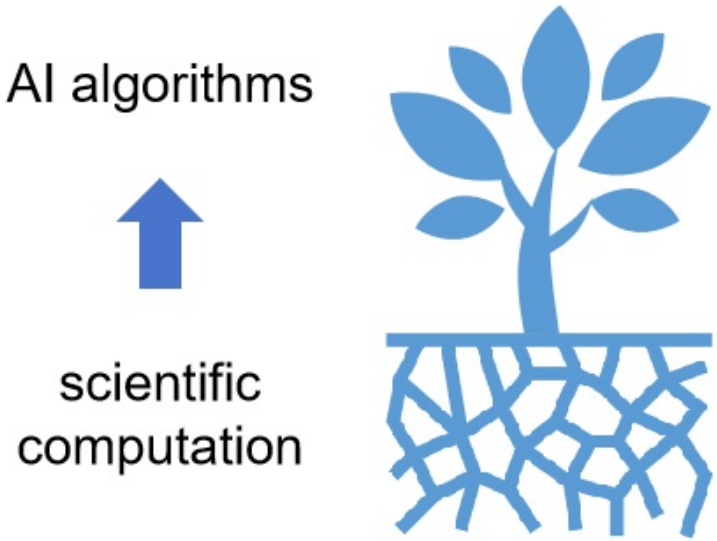
Anomaly detection: decoupling the AI algorithms

- As for general case:
  - normal/anomaly determination;
- As for anomaly cases:
  - expertise model for subtypes (most researchers mainly do);
- As for interested subtypes:
  - further investigation for principles/mechanisms;
- Well recoginized $\rightarrow$ full utilization:
  - model to export (suggested) decision, etc.

UNITED IMAGING

AI algorithm in practice

- The relations between AI algorithm and scientific computation



1. **SC** $\rightarrow$ AI $\rightarrow$ App.
   - technical methods set
   - rational combination
2. objective $\rightarrow$ AI frame
   - high generalization
   - expertise for problem
   - interpretability

UNITED IMAGING

Mottos:

- There is neither elixir for all diseases in this world, nor generic solution for all questions.[10]
- Invocation without coprehension is just like tree without root, stream without source.
- Scientific computation and rationality: the motivation and reason of finding the information underlying the data.

UNITED IMAGING

[10] Pattern in high dimensional data

Thanks.