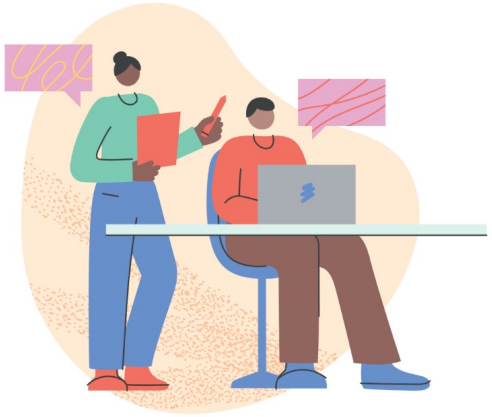
MSB
TRAINING SCHOOL
2023

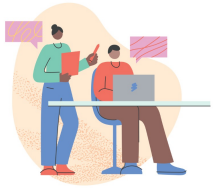


Dimensionality reduction and feature selection

Marta Belchior Lopes

UNIVERSITETI I TIRANËS

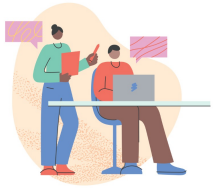
NOVA SCHOOL OF
SCIENCE & TECHNOLOGY



Dimensionality reduction

Why reducing data dimension?

- Not all features are important (e.g., less impact on the outcome or noise features)
- High number of features increase model complexity ➡ need to simplify
- Measuring can be expensive and time-consuming
- **Preprocessing** step for reducing **data dimension** and **model complexity**



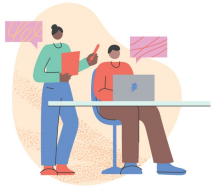
Dimensionality reduction

Why reducing data dimension?

- Not all features are important (e.g., less impact on the outcome or noise features)
- High number of features increase model complexity ➡ need to simplify
- Measuring can be expensive and time-consuming
- **Preprocessing** step for reducing **data dimension** and **model complexity**

FEATURE SELECTION

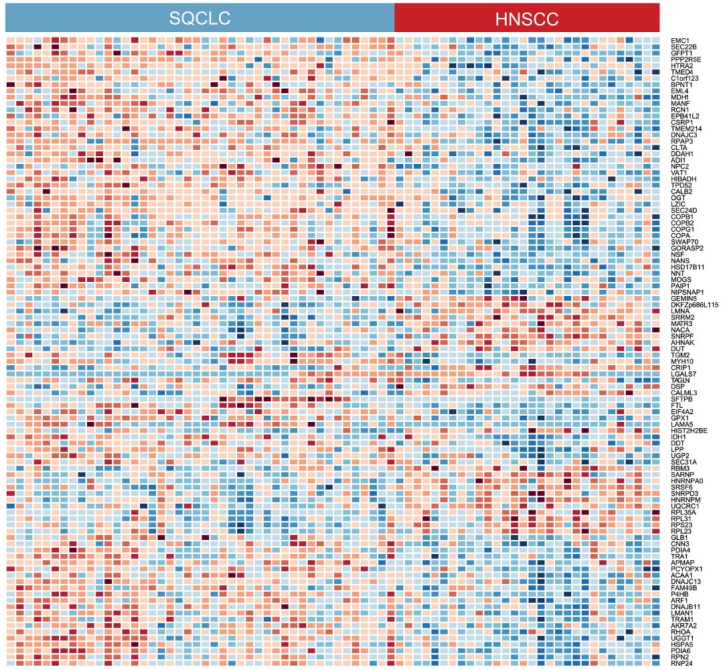
➡ **Improve model performance, interpretability** and **efficiency**



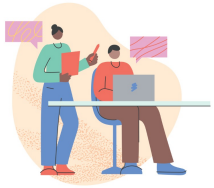
Feature selection

Examples

- **Genomics and Proteomics** in disease modeling



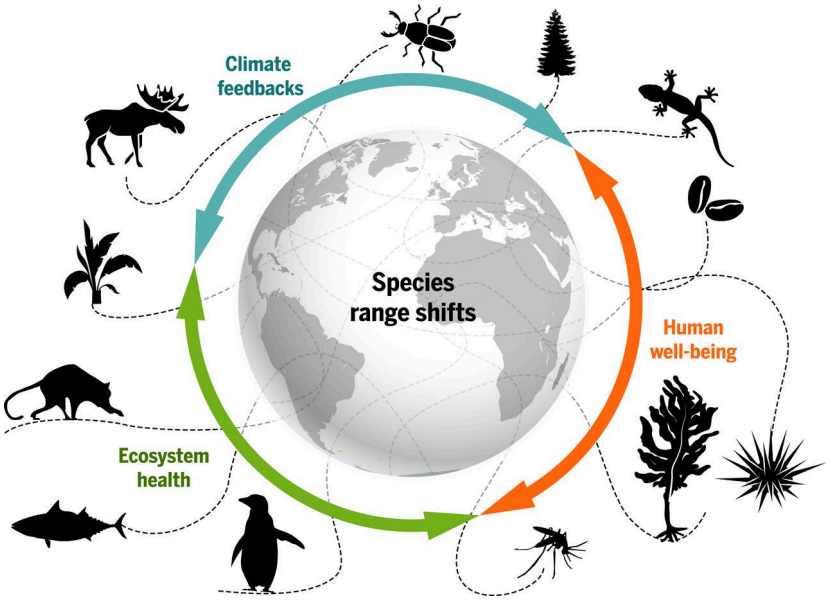
EMBO Mol Med (2018)10:e8428



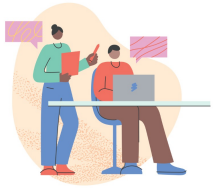
Feature selection

Examples

- **Genomics and Proteomics** in disease modeling
- **Ecological** studies
  - Species distribution
  - Biodiversity



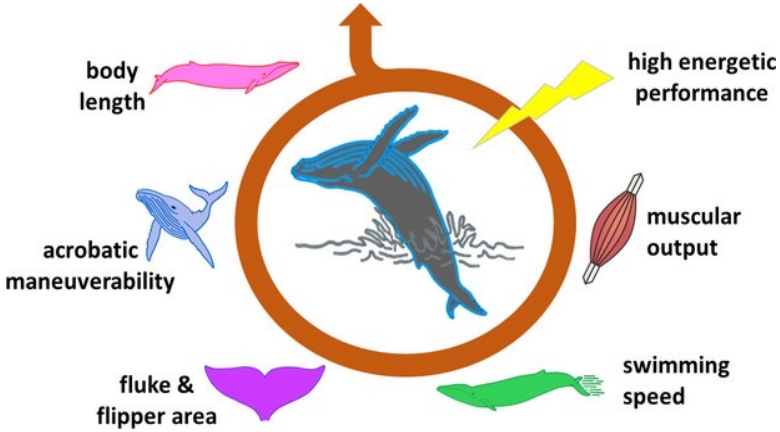
Science **355**, eaai9214 (2017)



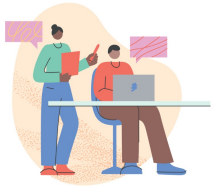
Feature selection

Examples

- **Genomics and Proteomics** in disease modeling
- **Ecological** studies
  - Species distribution
  - Biodiversity
  - Behaviour ecology



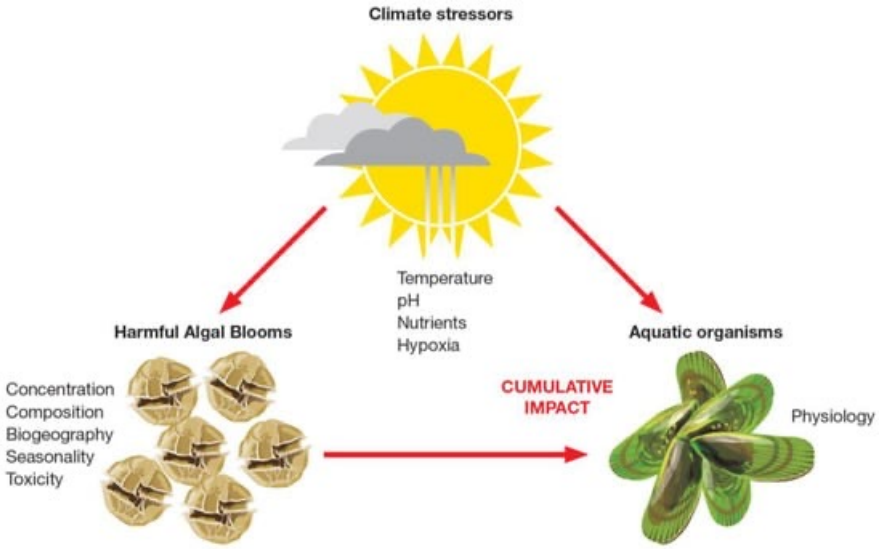
eLife **9**: e55722 (2022)



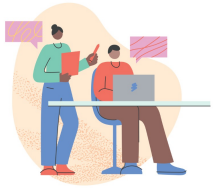
Feature selection

Examples

- **Genomics and Proteomics** in disease modeling
- **Ecological** studies
- **Toxicology** studies



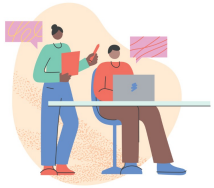
Toxins **2022**, *14*, 341



Feature selection

Strategies

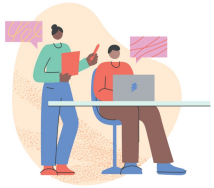
- **Filter** methods: independent of the learning algorithm used for model building (e.g., statistical properties of the variables)
- **Wrapper** methods: evaluate feature subsets using a learning algorithm and select features based on the performance of the algorithm
- **Embedded** methods: combine feature selection and model training, selecting features as part of the model-building process



Feature selection

Filter methods

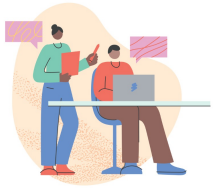
- Independence on the learning algorithm, and used as a **preprocessing** setp
- Computationally **efficient**
- Features assessed based on **statistical** measures



Feature selection

Filter methods

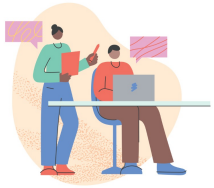
- Univariate (unsupervised and supervised)
    - Variance
    - Correlation
    - Mutual information
    - Chi-square test
    - Analysis of variance (ANOVA)



Feature selection

Filter methods

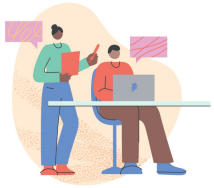
- Univariate (unsupervised and supervised)
    - Variance
    - Correlation
    - Mutual information
    - Chi-square test
    - Analysis of variance (ANOVA)
- Multivariate
    - Correlation-based



Feature selection

Wrapper methods

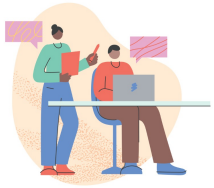
- **Dependence** on the learning algorithm
- **Iterative** process
- **Model performance** metrics as criterion (e.g., accuracy)
- Computationally **expensive** (inadequate for high-dimenisonal data)



Feature selection

Wrapper methods

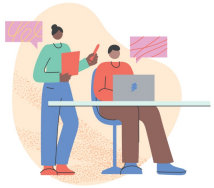
- **Dependence** on the learning algorithm
- **Iterative** process
- **Model performance** metrics as criterion (e.g., accuracy)
- Computationally **expensive** (inadequate for high-dimenisonal data)
    - Forward selection
    - Backward elimination
    - Recursive feature elimination



Feature selection

Embedded methods

- Combined **feature selection** and **model training**
- More **efficient** than wrapper methods
    - L1 regularization (LASSO)
    - Tree-based methods



Feature selection

Embedded methods

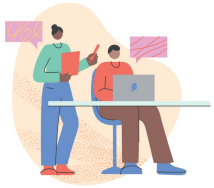
- Combined **feature selection** and **model training**
- More **efficient** than wrapper methods
    - L1 regularization (LASSO)
    - Tree-based methods

Sparse logistic regression

$$l(\beta_0, \beta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right] + \lambda F_\alpha(\beta)$$

$$F_\alpha(\beta) = \sum_{i=1}^{p}\left(\alpha|\beta_i| + \frac{1-\alpha}{2}\beta_i^2\right)$$

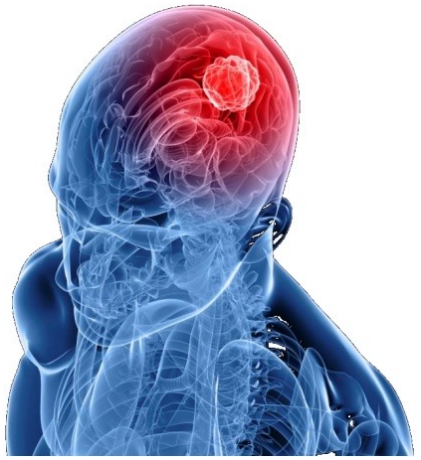
Elastic net & LASSO

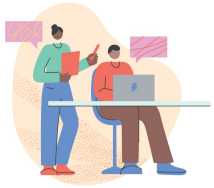


Feature selection

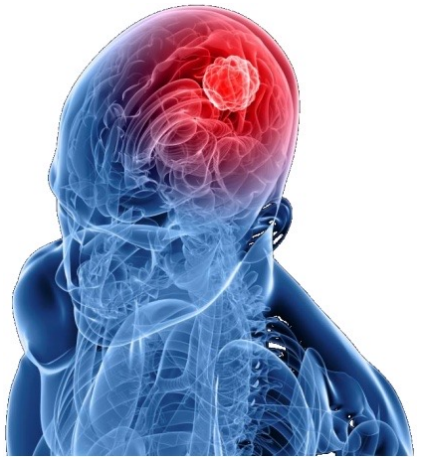
Single-cell gene expression data from gliomas

Gliomas – the most common brain tumors





Feature selection

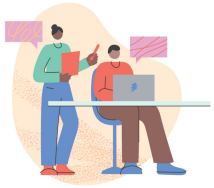


Single-cell gene expression data from gliomas

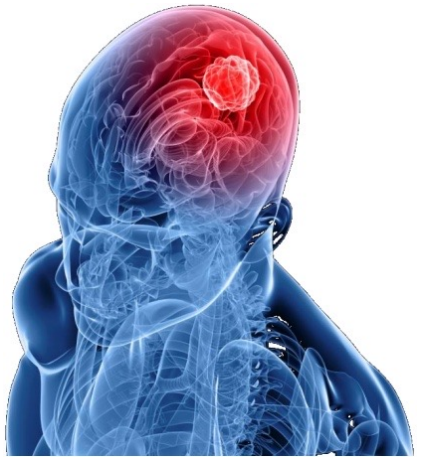
Gliomas – the most common brain tumors

Intertumoral heterogeneity

- Astrocytoma
- Oligodendroglioma
- Glioblastoma



Feature selection



Single-cell gene expression data from gliomas

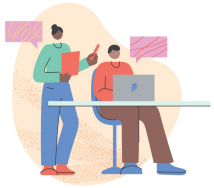
Gliomas – the most common brain tumors

Intertumoral heterogeneity

- Astrocytoma
- Oligodendroglioma
- Glioblastoma

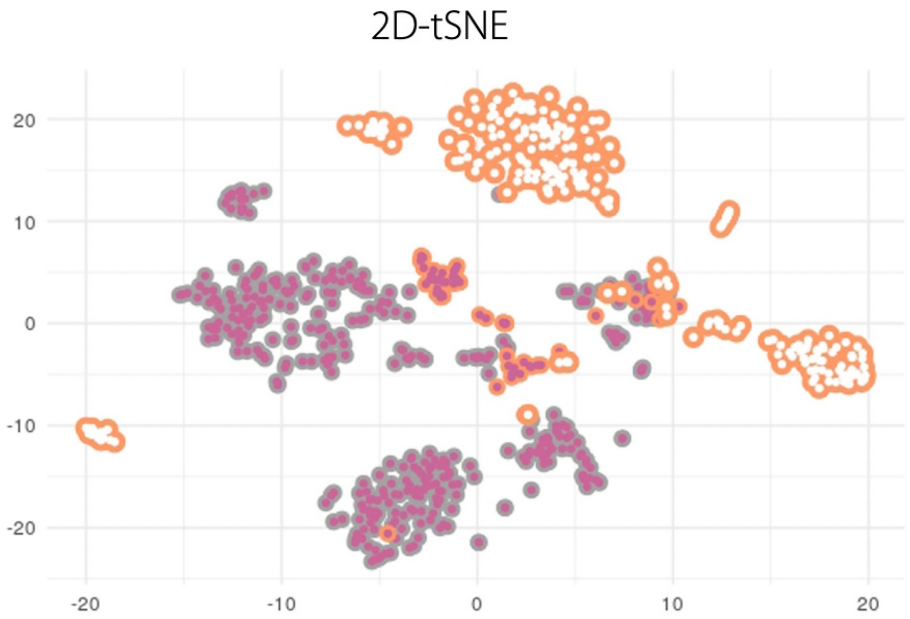
Intratumoral heterogeneity

Therapy failure and tumor relapse



Feature selection

Glioblastoma single-cell gene expression data



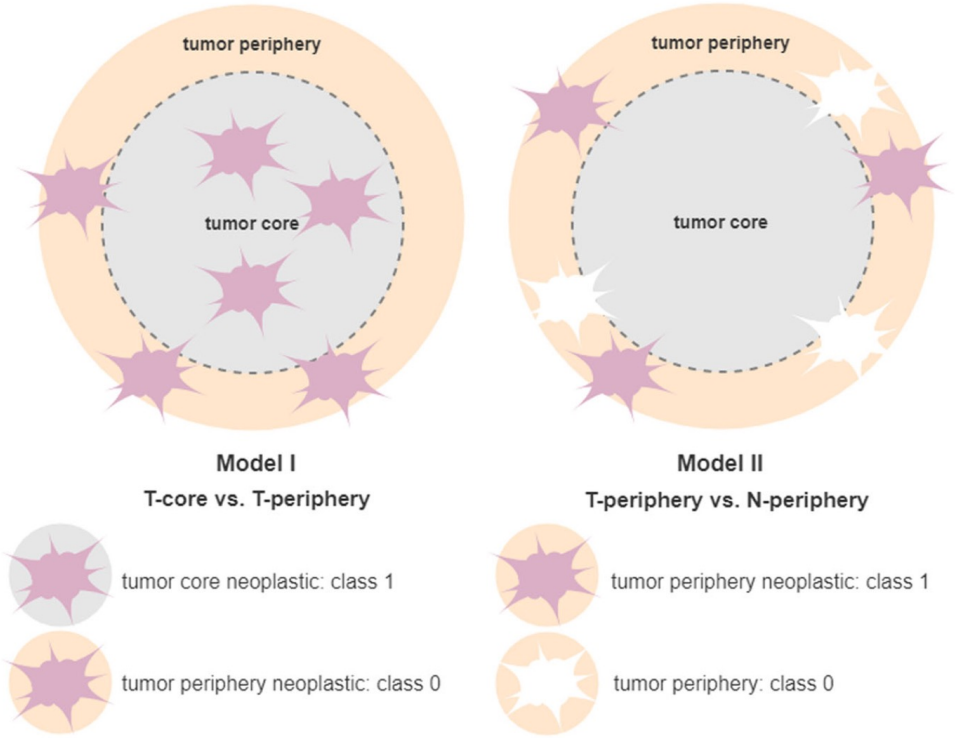
- Single-cell RNA-seq data
- Four primary glioblastoma patients
- **3,589 cells** and **23,368 genes**

tumor core neoplastic
tumor periphery neoplastic
tumor periphery normal

Lopes and Vinga *BMC Bioinformatics* (2020) 21:59
https://doi.org/10.1186/s12859-020-3390-4

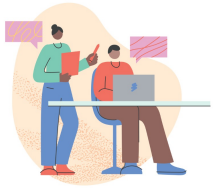
Feature selection

Glioblastoma single-cell gene expression data



- **Sparse** logistic regression
- **Elastic net** penalty

Lopes and Vinga *BMC Bioinformatics* (2020) 21:59
https://doi.org/10.1186/s12859-020-3390-4

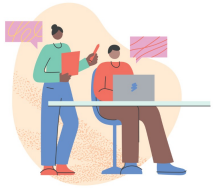


Feature selection

Glioblastoma single-cell gene expression data

Model performance

Classes	Vars	Miscl		AUC	
		Train	Test	Train	Test
I - T-core vs. T-periphery	83	10	7	0.97	0.94
II - T-periphery vs. N-periphery	85	3	4	0.99	0.96



Feature selection

Glioblastoma single-cell gene expression data

Model performance

Classes	Vars	Miscl		AUC	
		Train	Test	Train	Test
I - T-core vs. T-periphery	83	10	7	0.97	0.94
II - T-periphery vs. N-periphery	85	3	4	0.99	0.96

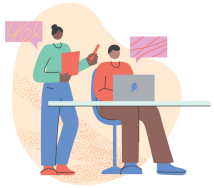
Genes selected

Model I - T-core vs. T-periphery

***ATP1A2**	***CLDN10**	***ECHDC2**	***FGFR3**	***GRM3**
***HERC6**	***HIF3A**	***HSPB8**	***NPL**	***PCSK1N**
***PPM1K**	***PRODH**	***SCG3**	***SPARCL1**	***TMSB10**

Model II - T-periphery vs. N-periphery

***ADAMTS3**	***ADAMTSL1**	***ANXA1**	***COL28A1**	***CRNDE**
***EGFR**	***EMP1**	***F2R**	***GNG5**	***HES6**
***HLA-A**	***HOXB3**	***HSPB6**	***HTRA1**	***ID3**
***IFI44L**	***IGFBP2**	***IQCE**	***LINC00475**	***MGLL**
***PSPH**	***PTGDS**	***SEC61G**	***SPOCK1**	***VIM**



Summary

- **Need** for data dimensionality reduction
- **Methods** for feature selection
- **Embedded** feature selection
- **Example** of application in Biomedicine