IST 2020

25 JUN

Feature Selection Methods based on Mutual Information

M. ROSÁRIO OLIVEIRA

CEMAT AND DEPT. MATEMÁTICA

Everest and papers about Feature Selection

Pedro Larrañaga.

The team:

Francisco Macedo

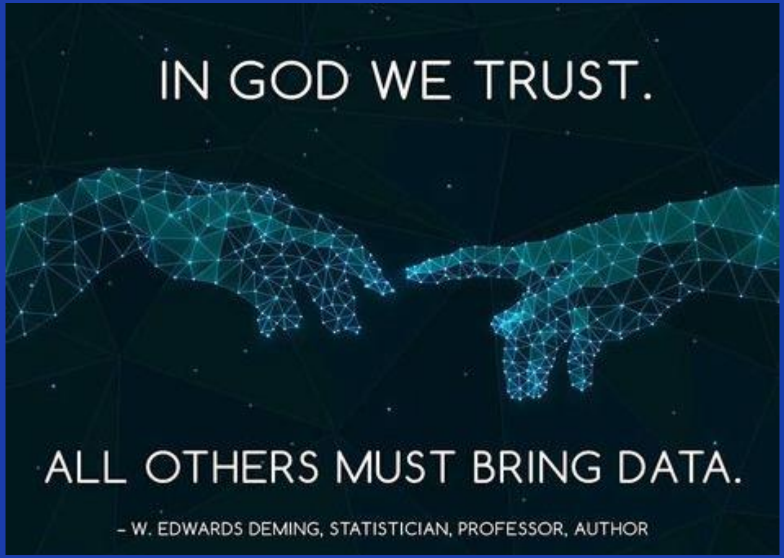
António Pacheco

Rui Valadas

Cláudia Pascoal

Eunice Carrasquinha

THE DATA BOOM



W. Edwards Deming

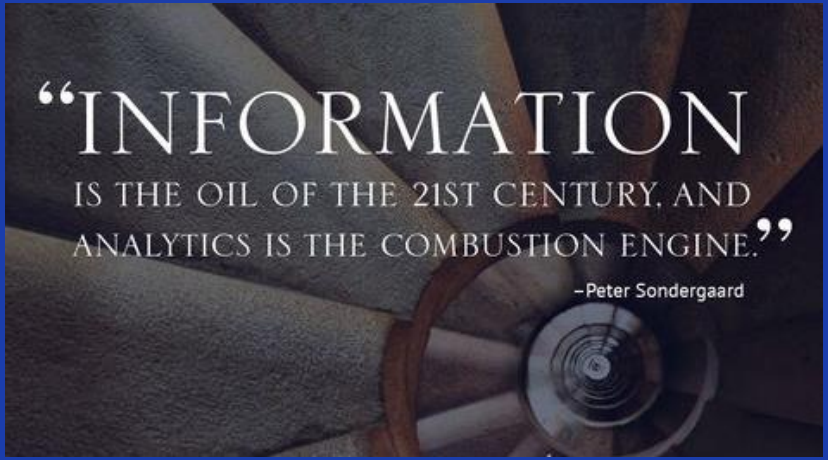
Jim Barksdale, former Netscape CEO

*"It's easy to **lie** with statistics. It's hard to **tell the truth** without statistics."*

Andrejs Dunkels, Swedish mathematics teacher, mathematician, and writer

*"**Data** is not **information**, **information** is not **knowledge**, **knowledge** is not **understanding**, **understanding** is not **wisdom**."*

Clifford Stoll, American astronomer, author and teacher



Stephen Few, Information Technology innovator, teacher, and consultant

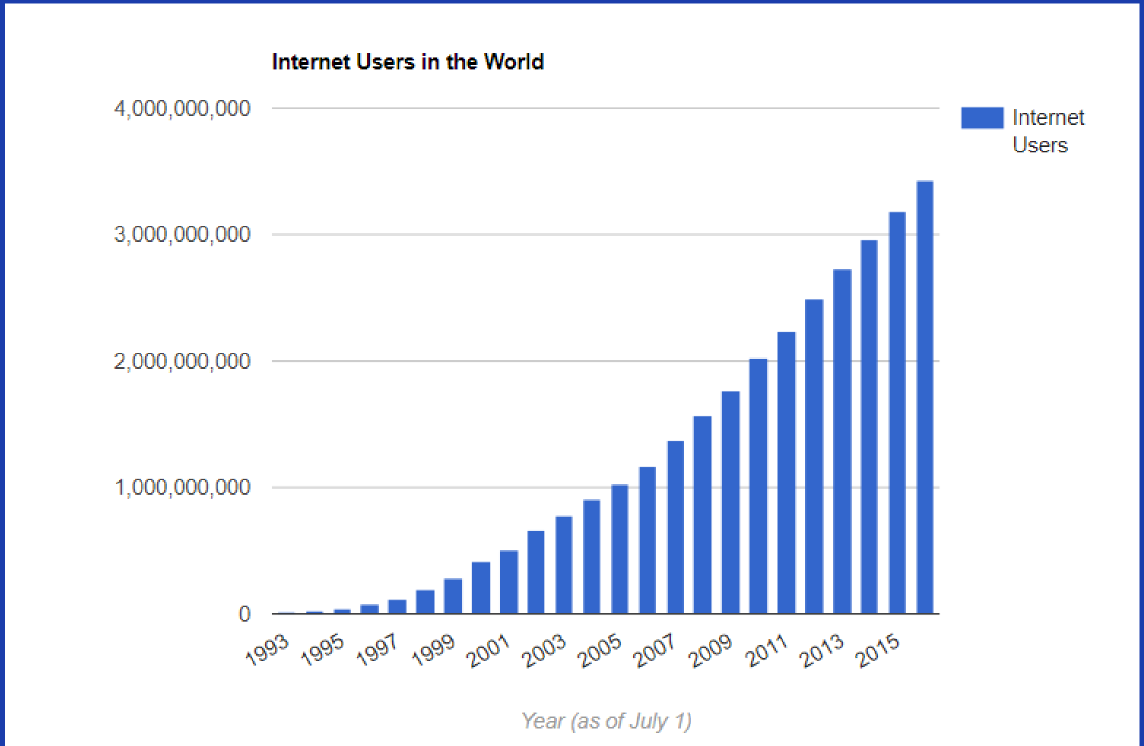
*"Big Data is like **teenage sex**:*
-everyone talks about it,
-nobody really knows how to do it,
-everyone thinks everyone else is doing it,
-so everyone claims they are doing it."

Dan Ariely, Duke University

Ronald Coase, British economist and author

Source: https://www.edvancer.in/50-amazing-big-data-and-data-science-quotes-to-inspire-you/

Internet



According with ***Internet live stats:***

- ≈40% of the world population has an Internet connection today
- In 1995, < 1%
- The number of Internet users has increased tenfold from 1999 to 2013
- 2005: 1st **billion**
- 2010: 2nd **billion**
- 2014: 3rd **billion**

Internet Live Stats

Source: https://www.internetlivestats.com//

Internet

- It **changed** the way we live and interact
- We are **generating data** according with our:
  - business, professional and social preferences
  - habits and activities

Google

Tweeter

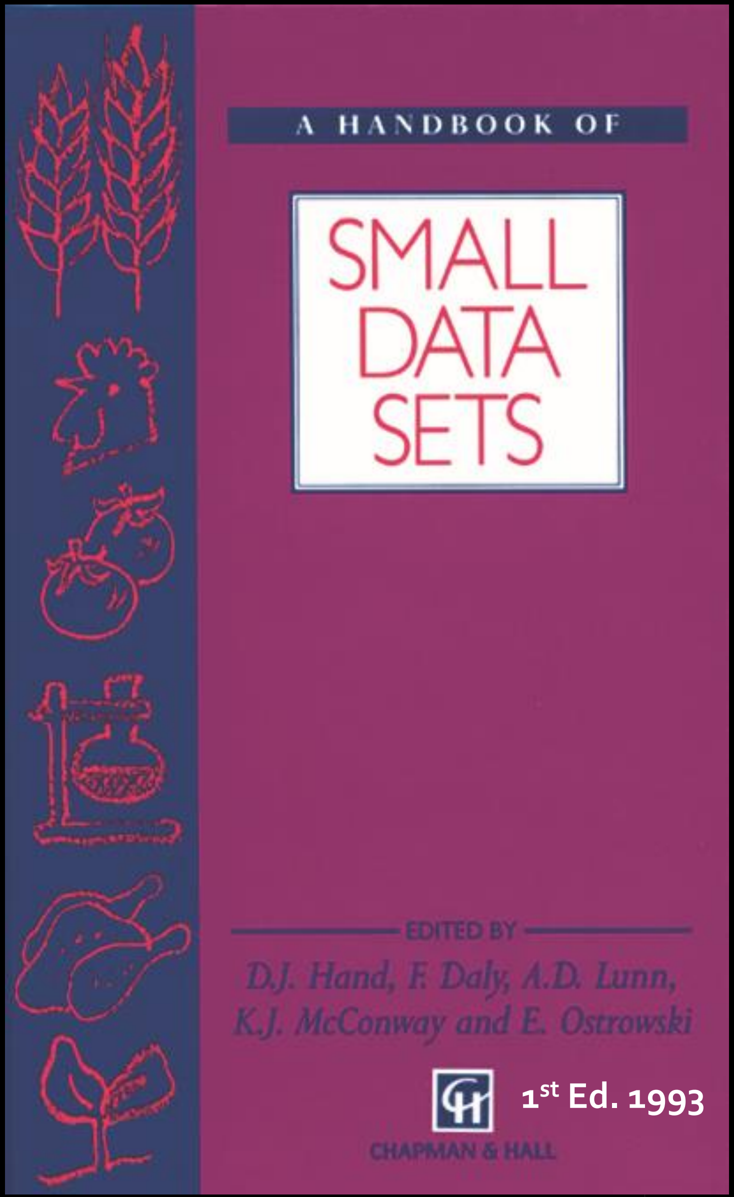
YouTube

Instagram

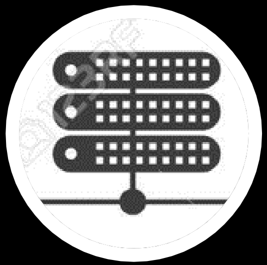
Skype

Facebook

SMALL DATA
BIG DATA

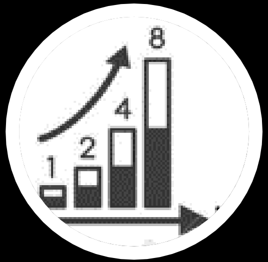


1st Ed. 1993



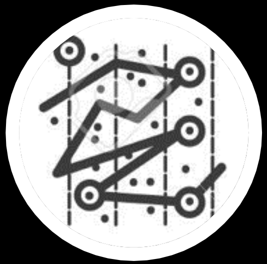
Volume

- Amount of Data
- Dimensionality
- Size



Velocity

- Data in motion
- Streaming
- Sensors



Variety

- Log files
- Text
- Video

Veracity

- Data in doubt
- Correctness
- Quality

Data Seen as Value

- "Big Data … belief that d
- "Scientists data coul but now th including government an management in particular, has realized that data create value."

S. P. Murphy, 2013

- "Suddenly it makes **economic sense** to … extract value from all this data out

S. Owen, 2014

"But People don't want data! They want answers!"

David Hand

New Ways to Collect Data

- **Online/web surveys**
- **Mobile phone surveys**
- **GPS tracking**
- **Web tracking technologies** (like cookies or meters)
- **Social media monitoring/listening**
- Crowdsourcing

- **IoT - Internet of things**
- **Chatbot** is an artificially intelligent software program that uses natural language processing to hold a conversation with its users
- Web Scraping from websites

iNaturalist
Explore Learn Record

- **Volunteer Monitoring/Citizen Science**
- Satellite data
- Invented/fake data

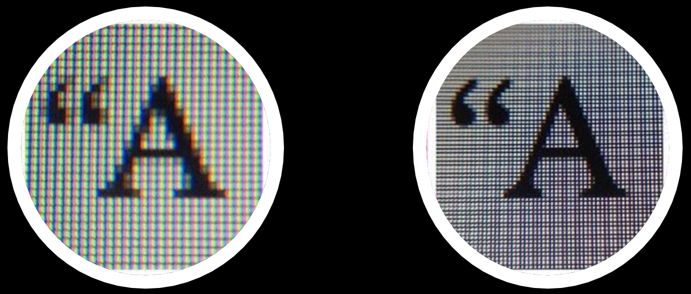
Trump's celebration of its election as USA president

https://i2ifacility.org/insights/blog/15-innovations-in-data-collection-methods-broadening-the-financial-inclusion-survey-toolkit?entity=blog

FEATURE SELECTION

FEATURE SELECTION:
THE RIGHT DATA

More data is not necessarily **more information**…



Feature Selection:

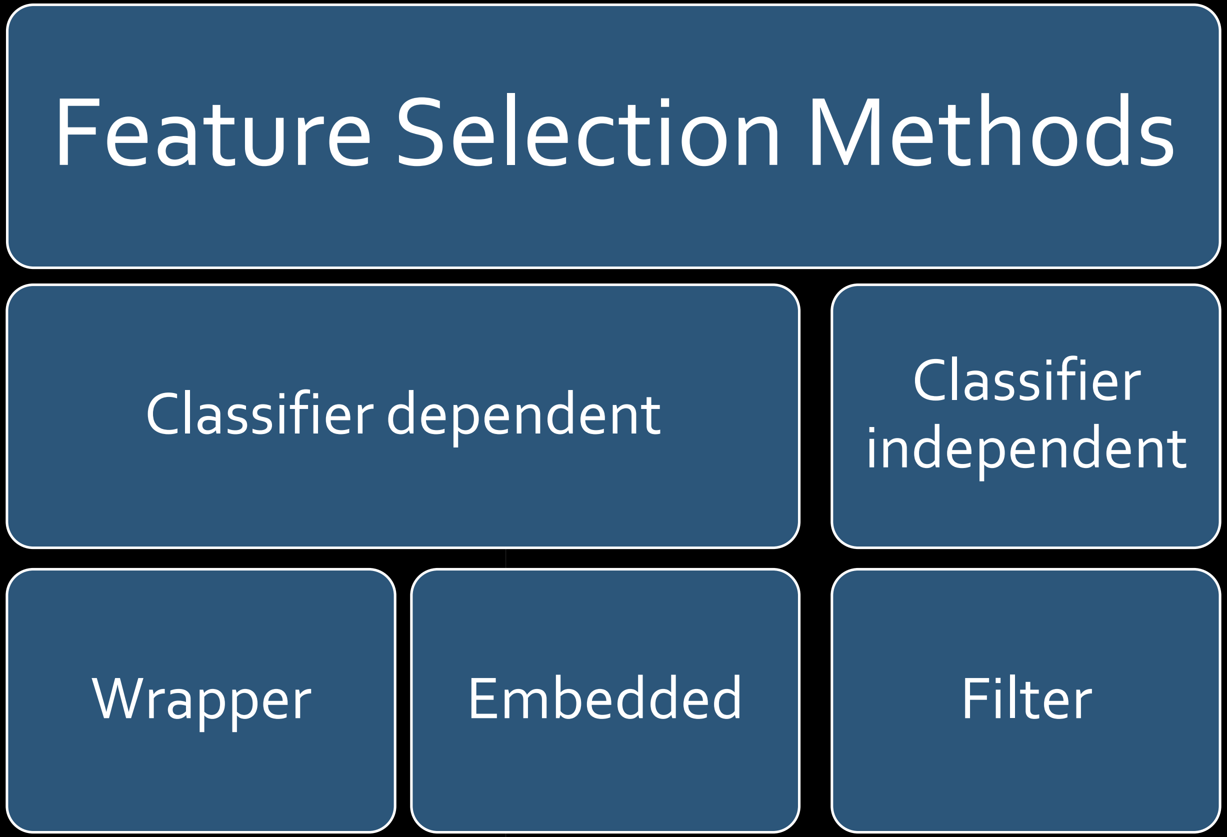
- Extract from the data **useful** and **valuable** knowledge for **real problem solving**

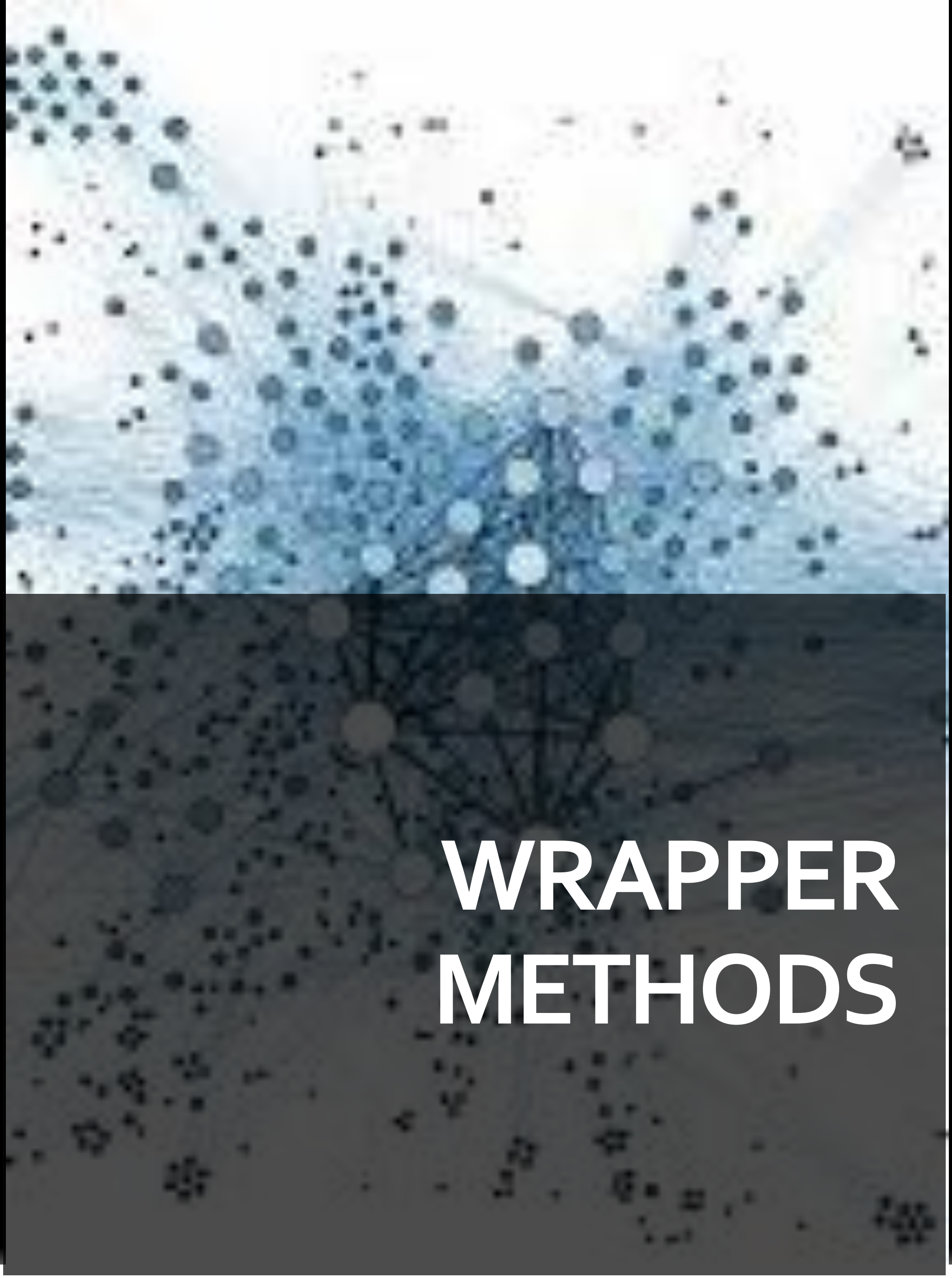
FEATURE SELECTION:
THE RIGHT DATA

- Select a **small subset** of the original features

- Designed to remove **irrelevant** and **redundant** features

- Reduce **computational complexity**
- Improve model **accuracy**
- Increase model **interpretability**


Feature Selection Methods
Classifier dependent
Classifier independent
Wrapper
Embedded
Filter

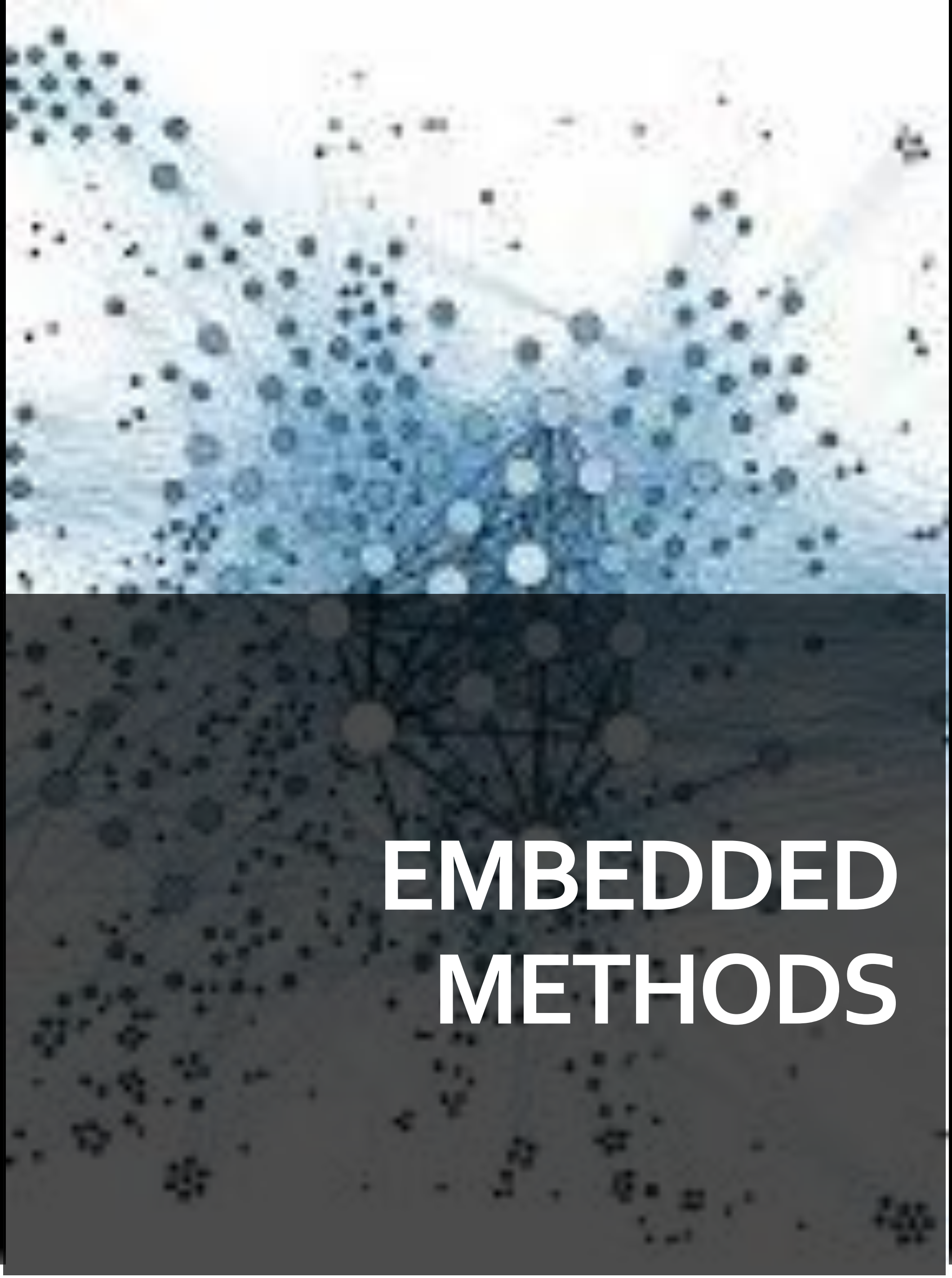


WRAPPER METHODS

- **Idea:** search for feature subsets, using the classifier accuracy as the measure of utility for a candidate subset

- **Disadvantages:**
  - computational cost
  - selected features are classifier specific

- **Example:**
  - Stepwise regression



EMBEDDED METHODS

- **Idea:** Classifier estimations and feature selection are not separated and interact
- **Disadvantages:**
  - Selected features are classifier specific
  - Regularized_OF=OF+λregularization_penalty
- **Example:**
  - Regularization methods

FILTER METHODS

- **Idea:** Classifier estimations and feature selection are separated and depend on a specific measure of benefit

- Most popular ones: rely on **Mutual Information** and **Entropy**
- **Mutual Information:** measures linear and non-linear associations among features

- **Example:**
  - Forward feature selection methods based on MI

ENTROPY, MUTUAL INFORMATION

Entropy

Entropy

Motivated by problems in the field of telecommunications

A Mathematical Theory of Communication*

C. E. Shannon (1948)

- A measure of uncertainty
- One formula that changed the world…

Entropy
Discrete rv

$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x}\in\mathcal{X}} P(\boldsymbol{X}=\boldsymbol{x}) \ln P(\boldsymbol{X}=\boldsymbol{x}).$$

- Does not depend on the values of X, only on its prob.
- H(a)=0
- H(X)≥0, Non-negative
- H(X)=ln(n), X~Unif{$a_1,\ldots,a_n$}, maximum

Differential Entropy
Continous rv

$$h(\boldsymbol{X}) = -\int_{\boldsymbol{x}\in\mathcal{X}} f_X(\boldsymbol{x}) \ln f_X(\boldsymbol{x}) d\boldsymbol{x}.$$

- Does not depend on the values of X, only on its prob.
- Can be negative
- h(X)=ln(a), X~Unif(0,a),
  - a=1, h(X)=0
  - a<1, h(X)<0

Mutual Information

Mutual Information
Discrete rv

$$\mathrm{MI}(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(X=x, Y=y) \ln \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)}.$$

- Measures linear and non-linear associations between X and Y
- $MI(X,Y) \geq 0$
- Symmetric
- $MI(X,Y)=0$ iff $X \perp\!\!\!\perp Y$
- $MI(X,X)=H(X)$

Mutual Information
Continuous rv

$$\mathrm{MI}(X,Y) = \int_{y\in\mathcal{Y}}\int_{x\in\mathcal{X}} f_{X,Y}(x,y) \ln \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} dx\,dy.$$

- Measures linear and non-linear associations between X and Y
- All properties hold, except
- $MI(X,X)=+\infty$

Generalizations of MI:

Triple Mutual Information

TMI(X,Y,Z)

- The generalization to more than 2 rv -- not unique
- Measures association among **X, Y,** and **Z**
- Not necessarily non-negative

$$\mathrm{TMI}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) - \mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}).$$

Conditional Mutual Information

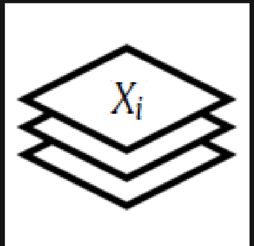
MI(X,Y|Z)

- Measures association between **X** and **Y** given **Z**
- MI(X,Y|Z)=0 iff X ⫫ Y | Z, conditional independence

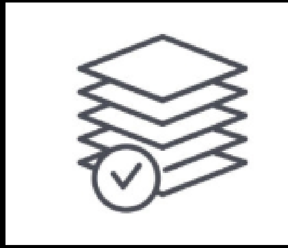
FORWARD FEATURE SELECTION

FORWARD FEATURE SELECTION

Goal: Select a **small subset** of the original features, excluding **irrelevant** and **redundant** features



$$X_j = \arg\max_{X_i \in F} \mathrm{MI}(C, \boldsymbol{S} \cup \{X_i\}).$$

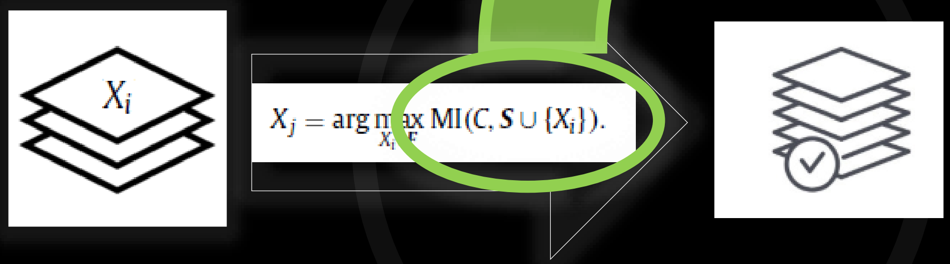


F= Candidate features

S= Selected features

C= Class-variable

FORWARD FEATURE SELECTION



Problem:

$$\mathrm{MI}(C, \boldsymbol{S} \cup \{X_i\}) = \mathrm{MI}(C, \boldsymbol{S}) + \mathrm{MI}(C, X_i | \boldsymbol{S})$$

Calculation / Estimation

Method

FORWARD FEATURE SELECTION

1^{ST} GROUP

First Group of Methods

- **Goal:** Obtain subset of features leading to:
  - **maximum *relevance*** between the candidate feature and the class

Method	Objective function evaluated at X_i
MIM	$MI(C, X_i)$

- Maximizes ***association with the class***

FORWARD FEATURE SELECTION

2^{ND} GROUP

Second Group of Methods

- **Goal:** Obtain subset of features leading to:
  - **maximum *relevance*** between the candidate feature and the class
  - **minimum *redundancy*** of the candidate feature with respect to the already selected ones

FORWARD FEATURE SELECTION

2^{ND} GROUP

Second Group of Methods

Method	Objective function evaluated at X_i
MIFS	$\mathrm{MI}(C, X_i) - \beta \sum_{X_s \in S} \mathrm{MI}(X_i, X_s)$
mRMR	$\mathrm{MI}(C, X_i) - \frac{1}{\lvert S \rvert} \sum_{X_s \in S} \mathrm{MI}(X_i, X_s)$
maxMIFS	$\mathrm{MI}(C, X_i) - \max_{X_s \in S} \mathrm{MI}(X_i, X_s)$

- ***Inter-feature redundancy*** : association between the candidate **feature** and the **selected ones**
- Avoids ***collinearity*** in the classifier estimation

FORWARD FEATURE SELECTION

2ND GROUP

Third Group of Methods

- Adds a third term:
  - **Complementary** accommodates possible dependencies among the features given the class

FORWARD FEATURE SELECTION

3RD GROUP

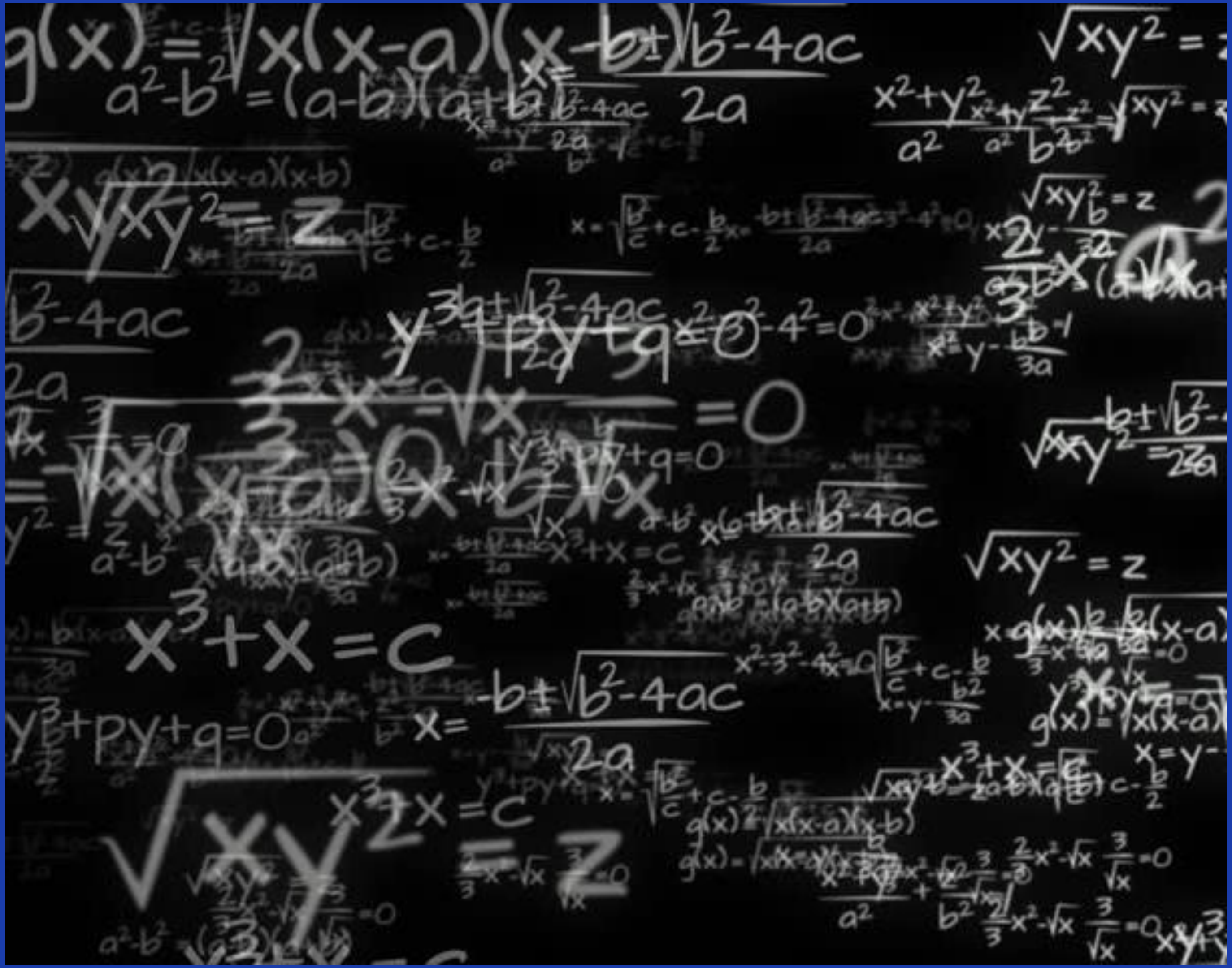
Third Group of Methods

Method	Objective function evaluated at X_i
CIFE	$MI(C,X_i) - \sum_{X_s \in S}(MI(X_i,X_s) - MI(X_i,X_s \mid C))$
JMI	$MI(C,X_i) - \frac{1}{\lvert S \rvert}\sum_{X_s \in S}(MI(X_i,X_s) - MI(X_i,X_s \mid C))$
CMIM	$MI(C,X_i) - \max_{X_s \in S}\{MI(X_i,X_s) - MI(X_i,X_s \mid C)\}$
JMIM	$MI(C,X_i) - \max_{X_s \in S}\{MI(X_i,X_s) - MI(X_i,X_s \mid C) - MI(C,X_s)\}$

- ***Class-relevant redundancy***: contribution of a candidate feature to the explanation of the class, when taken together with already selected features

FEATURE SELECTION METHODS

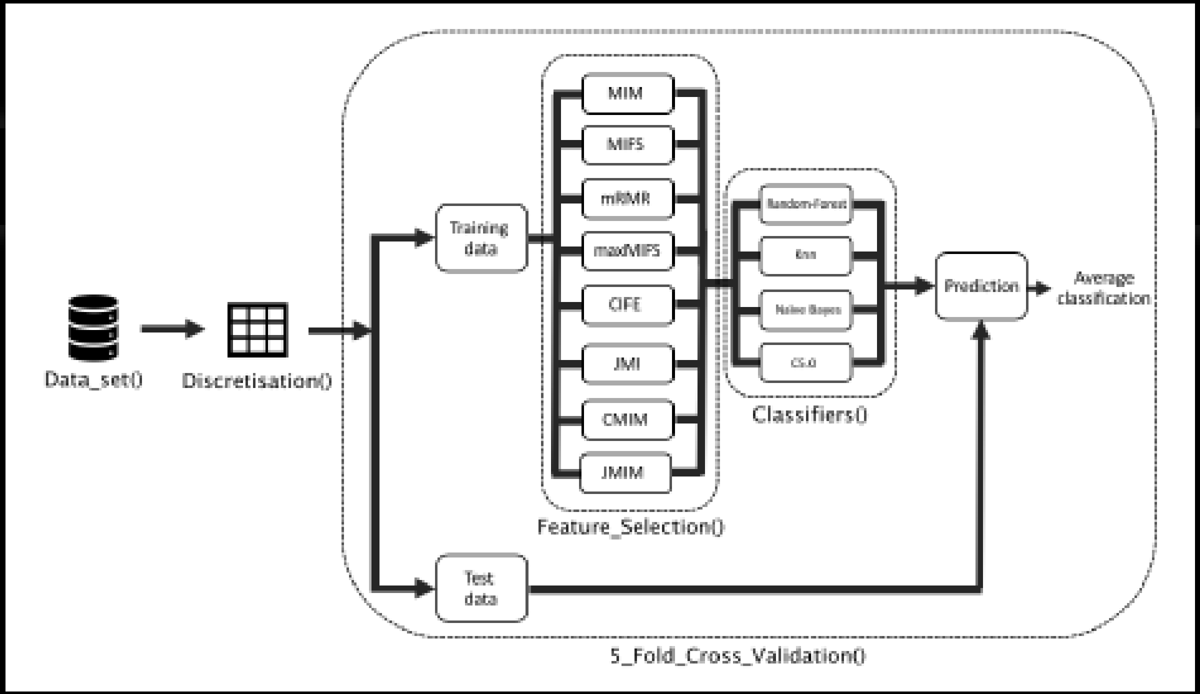
A THEORETICAL COMPARISON



NUMERICAL COMPARISON:

HOW THINGS ARE USUALLY DONE

How comparisons are usually done:



Source: Botelho (2020), Study project about feature selection.

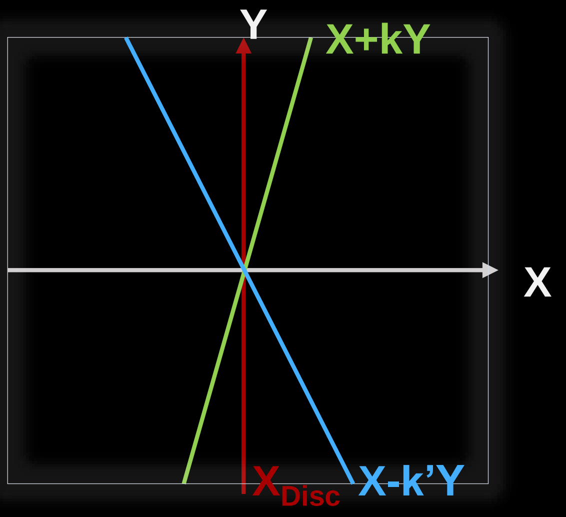
THEORETICAL COMPARISON:

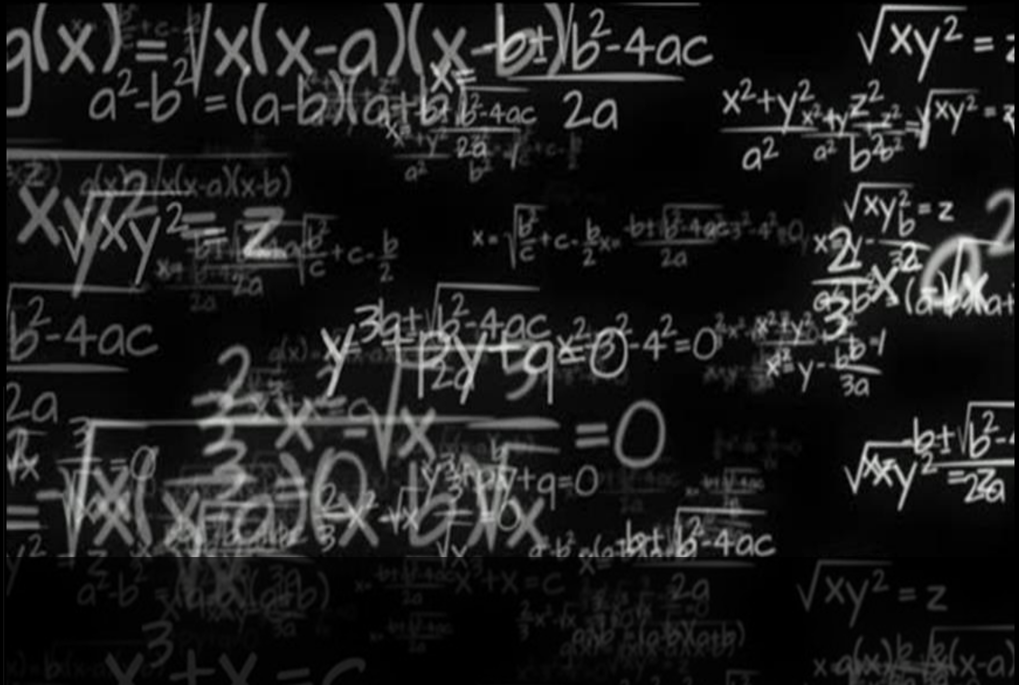
USING A DISTRIBUTIONAL SETTING

Theoretical Setup:

Class-Variable: **C**=Sgn(X+kY)

Candidate Features: **X, X-k'Y**, Sgn(X), Z





THEORETICAL COMPARISON:

USING A DISTRIBUTIONAL SETTING

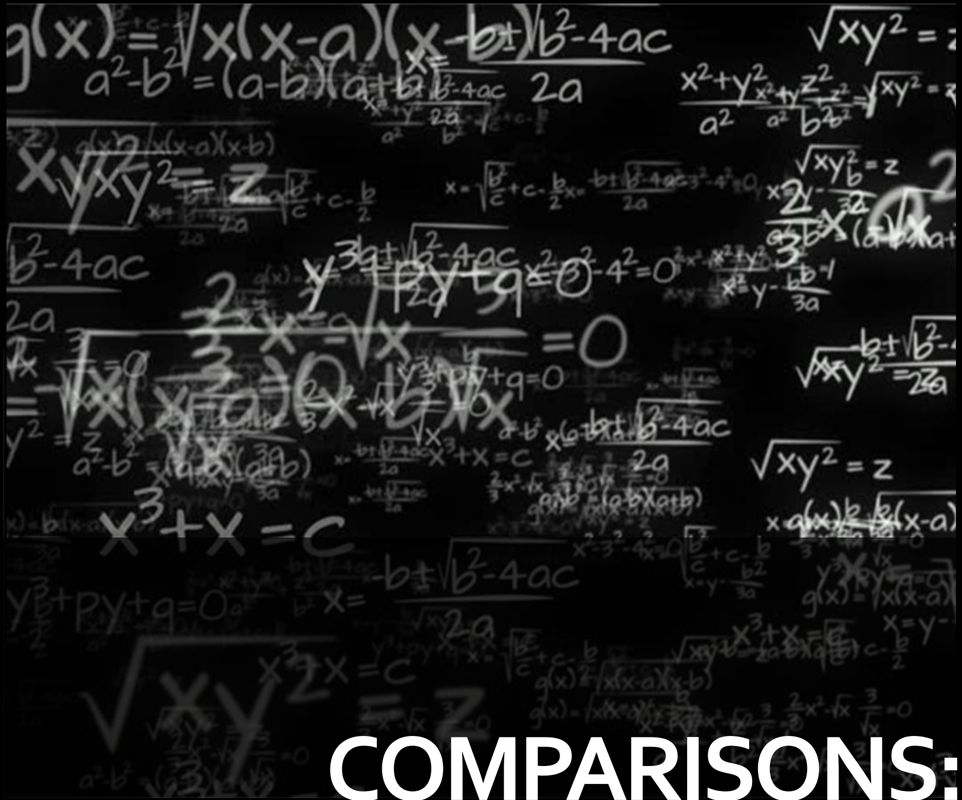
Features Order: Objective functions were calculated theoretically assuming X, Y, and Z are N(0,1)

Performance Measure:

Minimum Bayes Risk

=

Minimum Probability of Misclassifiction



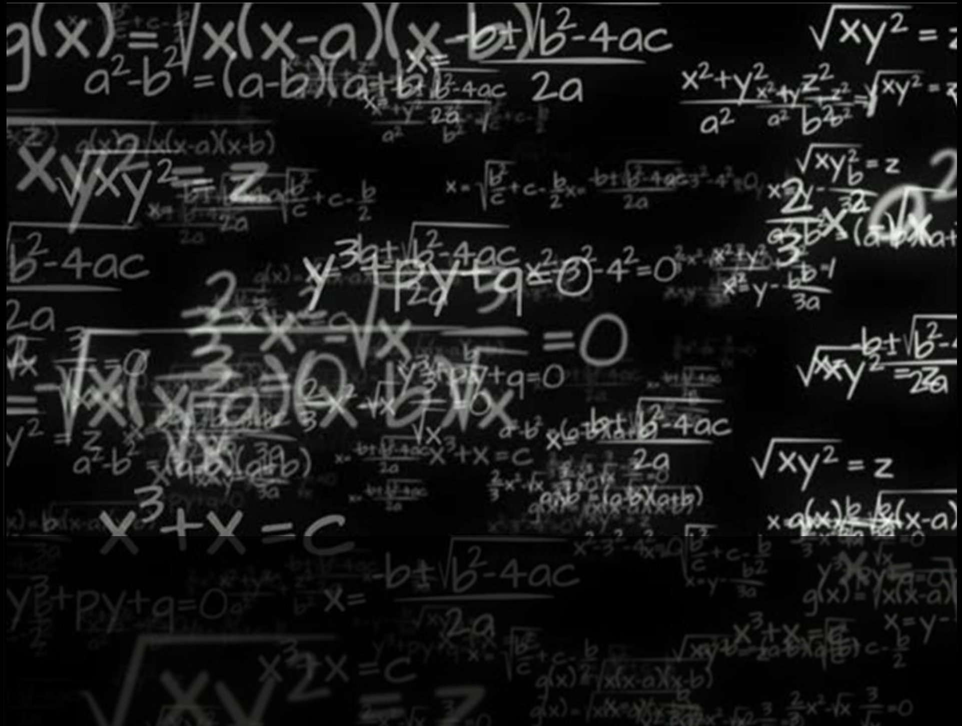
COMPARISONS:

FORWARD FEATURE SELECTION METHODS

Advantages and drawbacks:

Advantage	MIM
Class association	X
Feature redundancy	
Feature complementarity	
Drawback	**MIM**
Redundancy undervalued	
Redundancy overscaled	
Complementarity penalized	
Unimportant term approximated	

Ignores Redundancy



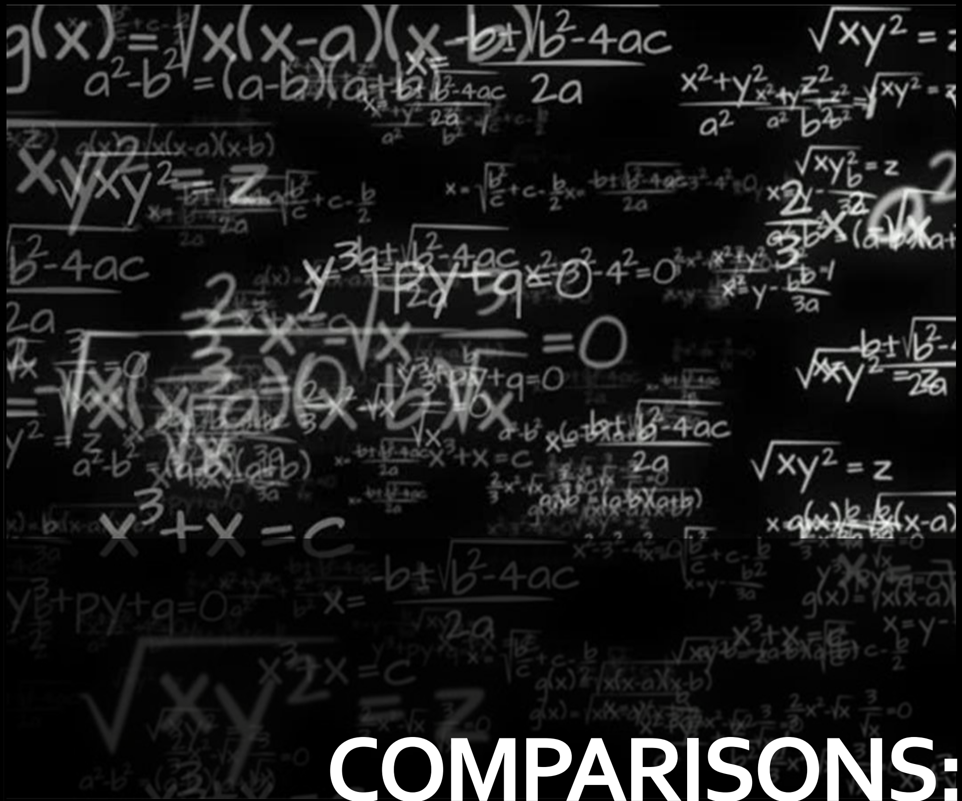
COMPARISONS:

FORWARD FEATURE SELECTION METHODS

Advantages and drawbacks:

Advantage	MIM	MIFS	mRMR	maxMIFS
Class association	X	X	X	X
Feature redundancy		X	X	X
Feature complementarity				
Drawback	**MIM**	**MIFS**	**mRMR**	**maxMIFS**
Redundancy undervalued			X	X
Redundancy overscaled		X		
Complementarity penalized				
Unimportant term approximated				

Cannot guarantee that relevant are selected before redundant and irrelevant features



COMPARISONS:

FORWARD FEATURE SELECTION METHODS

Advantages and drawbacks:

Advantage	JMI	CMIM
Class association	X	X
Feature redundancy	X	X
Feature complementarity	X	X
Drawback	**JMI**	**CMIM**
Redundancy undervalued	X	X
Redundancy overscaled		
Complementarity penalized		X
Unimportant term approximated		

Recommended,
But there are room for
improvements!

FEATURE SELECTION METHODS

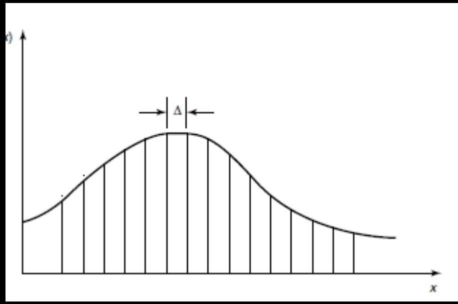
SOLVING REAL PROBLEMS - ESTIMATION

ENTROPY ESTIMATION

Empirical estimators:

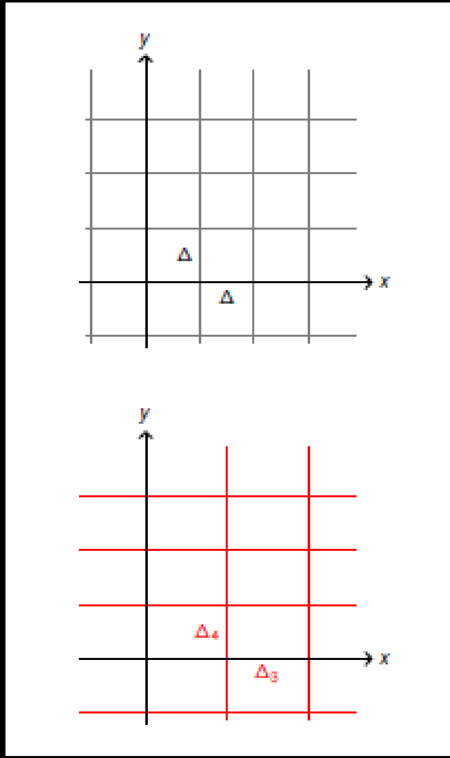
- **Discretize continuous** rv
- **Estimate entropy and MI** (discrete case)

$$\mathrm{H}(X^{\Delta}) + \ln \Delta \to h(X)\,, \quad \text{as} \quad \Delta \to 0$$



MUTUAL INFORMATION

Empirical estimators:



$$\mathbf{MI}\left(\mathbf{X}^{\Delta}, \mathbf{Y}^{\Delta}\right) \rightarrow \mathbf{MI}(\mathbf{X}, \mathbf{Y})$$

$$\mathbf{MI}\left(\mathbf{X}^{\Delta}, \mathbf{Y}^{\Delta}\right) - \ln \Delta_1 - \ln \Delta_2 + \ln \Delta_3 + \ln \Delta_4 \rightarrow \mathbf{MI}(\mathbf{X}, \mathbf{Y})$$

ESTIMATION

Example:

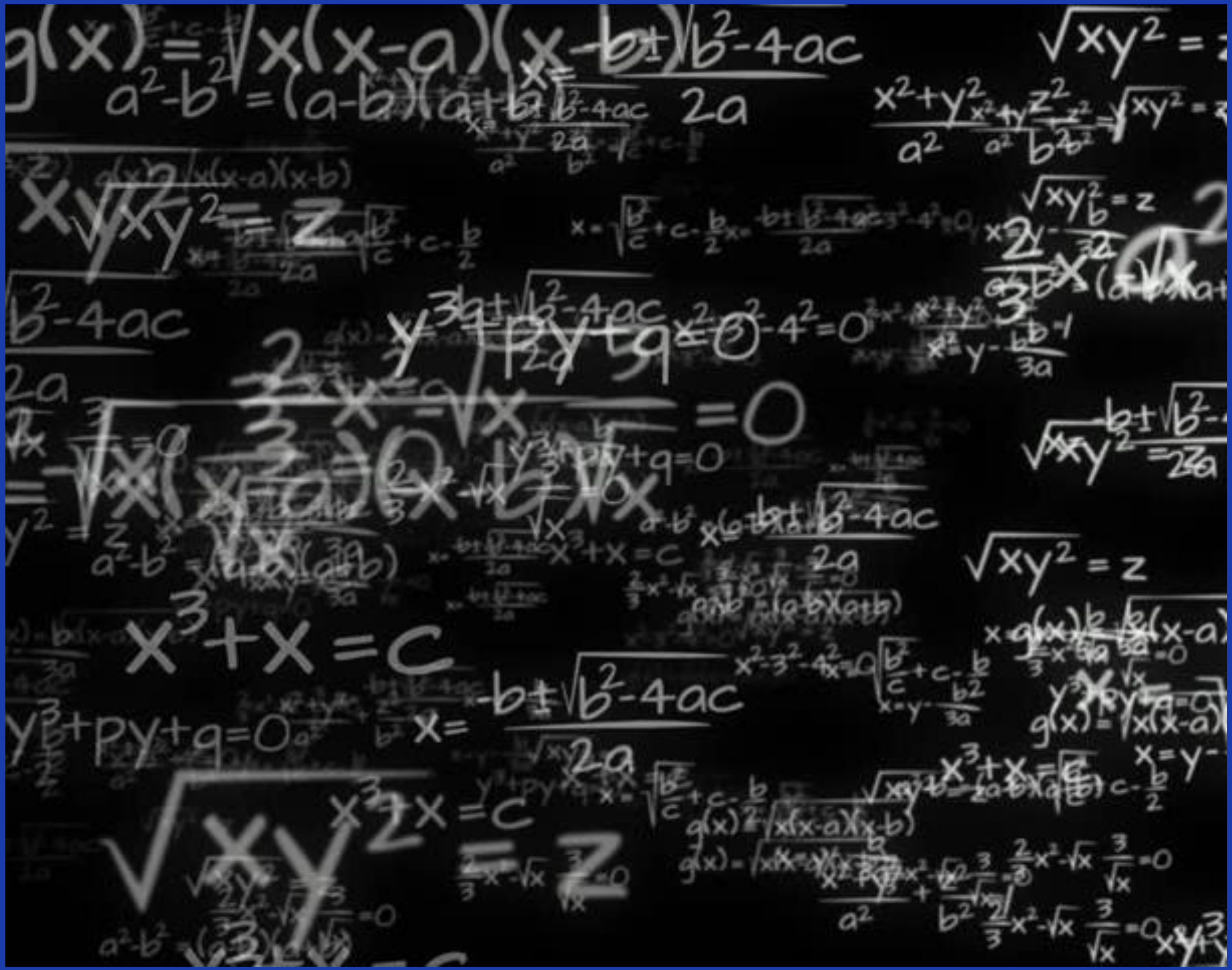
	MI(C,X)	MI(C,X-Y)	MI(X,X-Y)
Kwak, Choi (2002)	0.8459	0.2621	0.6168
Huang et al. (2008)	0.8438	0.2807	0.6099
Pascoal (2014)	0.5932	0.1779	0.5004
TRUE Value	**0.5932**	**0.1785**	**0.5**

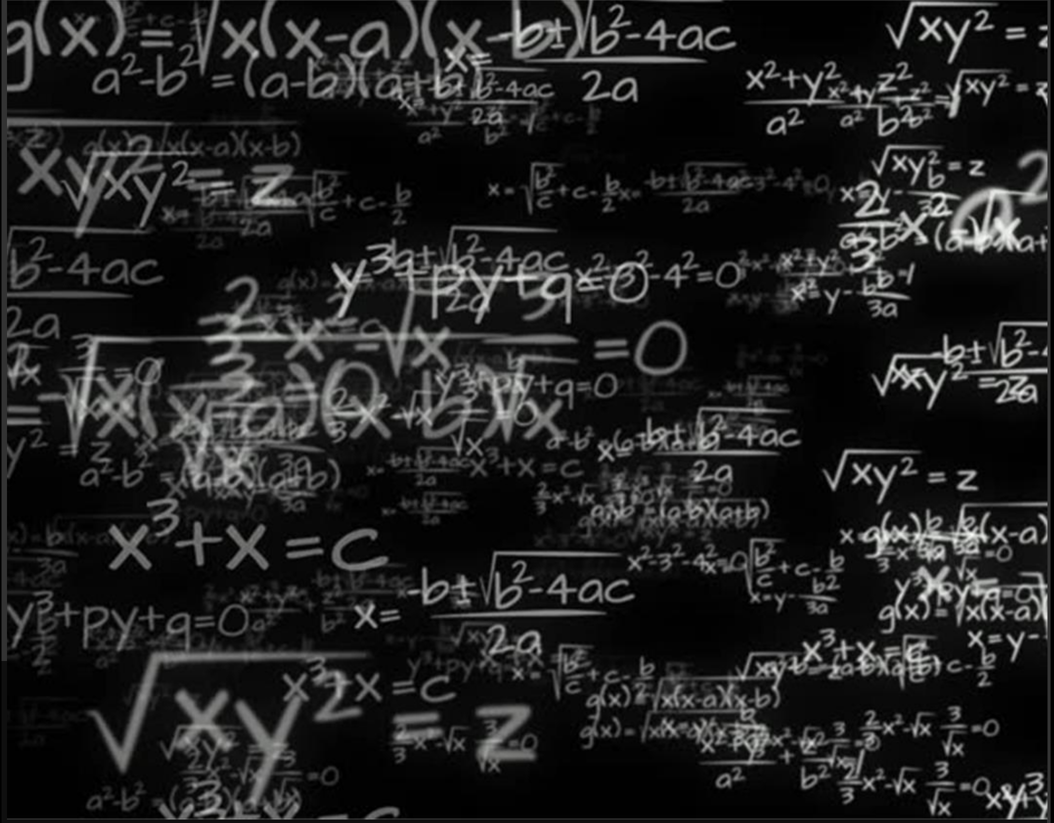
Challenges:

- How to discretize
- Number of classes
- Continuity corrections

FEATURE SELECTION

THE FUTURE

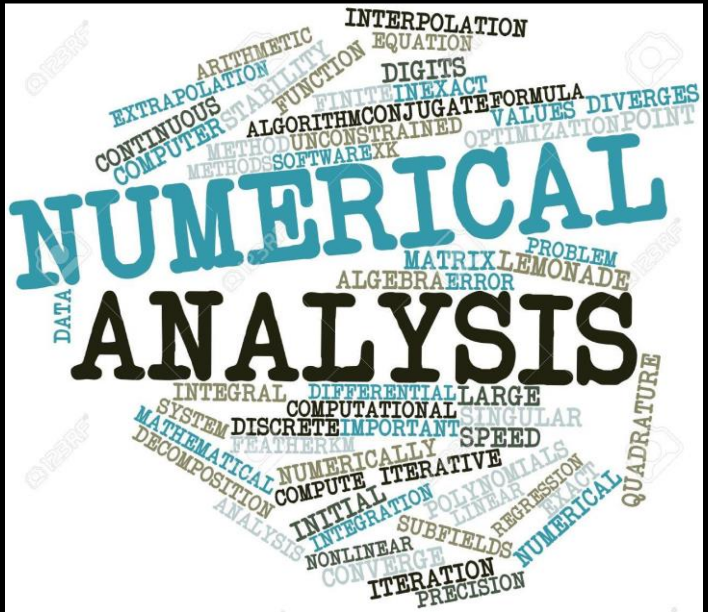




FEATURE SELECTION

THE FUTURE

Principal Component Analysis



Estimation: $\hat{\Sigma}$

Eigen Problem

THE TEAM:

Giulia Ferrandi

Igor Kravchenko

Michiel Hochstenbach

Main References

Theoretical foundations of forward feature selection methods based on mutual information

Francisco Macedo[a,b], M. Rosário Oliveira[a,*], António Pacheco[a], Rui Valadas[c]

[a] CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa 1049–001, Portugal
[b] EPF Lausanne, SB-MATHICSE-ANCHP, Station 8, Lausanne CH-1015, Switzerland
[c] IT and Departament of Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa 1049–001, Portugal

Theoretical evaluation of feature selection methods based on mutual information

Cláudia Pascoal[a], M. Rosário Oliveira[a,*], António Pacheco[a], Rui Valadas[b]

[a] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[b] IT and Dep. Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal