Introduction to Statistical Learning

Santiago Marco

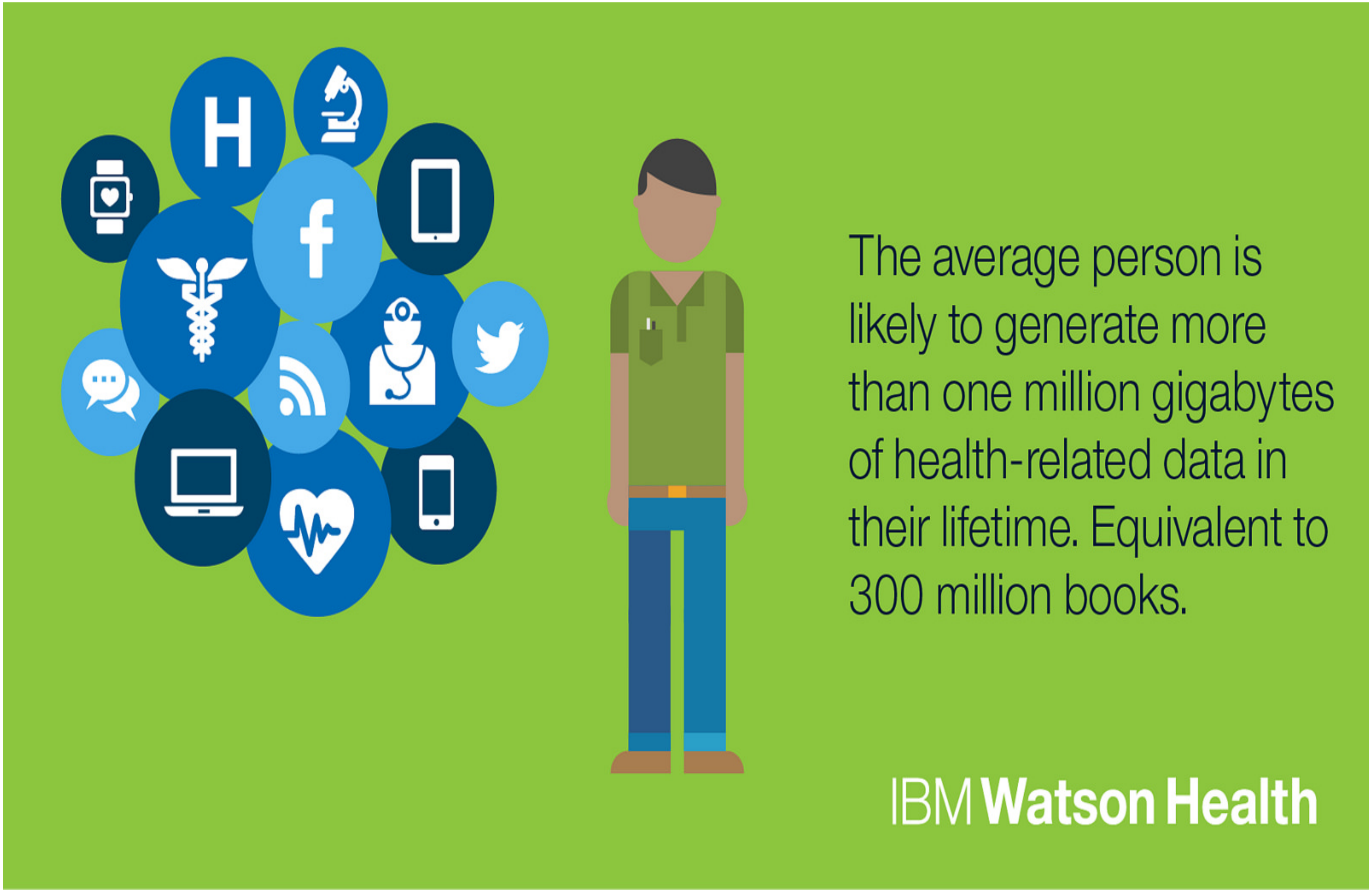
Department of Electronics and Biomedical Engineering, UB
Institute for Bioengineering of Catalonia

santiago.marco@ub.edu
smarco@ibecbarcelona.eu

Introduction to Statistical Learning

- **Motivation and Basic Concepts**
- **Application Examples**
- **Introduction to very basic classifiers**
  - Minimum Distance Classifier
  - K-Nearest Neighbours
- **Complexity Control:**
  - Dimensionality Reduction
  - Regularization
- **Loss functions and figures of merit.**
- **Summary**

The Data Avalanche

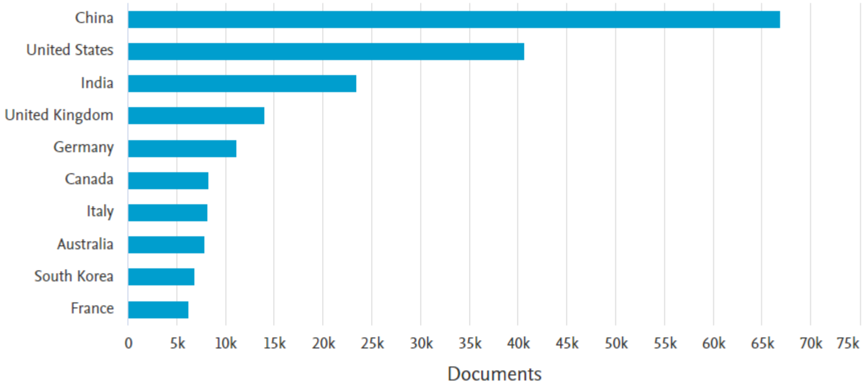


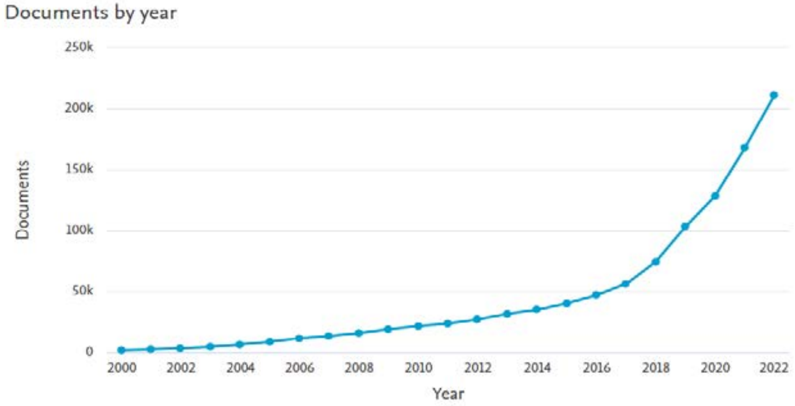
The average person is likely to generate more than one million gigabytes of health-related data in their lifetime. Equivalent to 300 million books.

IBM Watson Health

ML and (Health or Bioinformatics)

SCOPUS

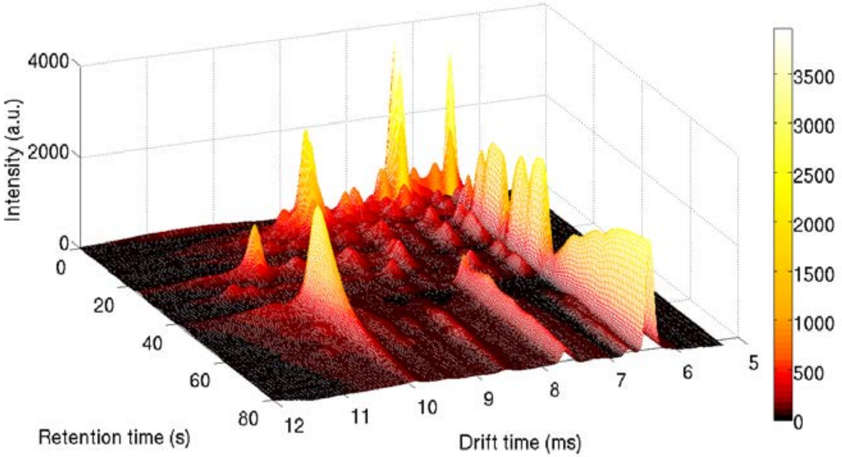


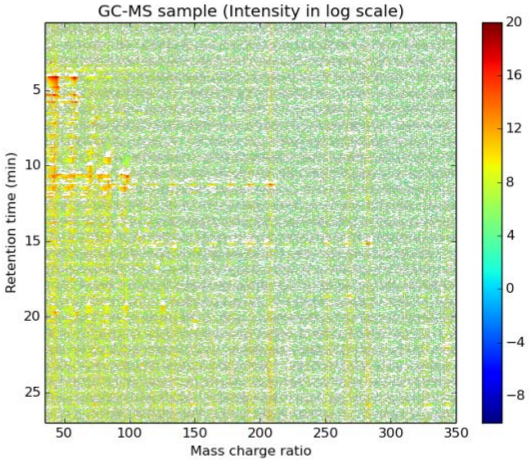


- **Relevant scientific journals (IF as 2022):**
  - Journal of Machine Learning Research (JMLR. Inc, IF=4.1)
  - Arificial Intelligence in Medicine (Elsevier, IF=5.0)
  - IEEE Journal of Biomedical and Health Informatics (IEEE, IF=5.2)
  - Bioinformatics (Oxford, IF=5.6)
  - Medical Image Analysis (Elsevier, IF=11.1)
  - PLOS Computational Biology (PLOS, IF=5.8)
  - Journal of Biomedical Informatics (Elsevier, IF=4.5)

Signal and Data Processing

- **Signal processing and data analysis are of increasing importance in:**
  - -omics data, electronic health records
  - General analytical chemistry,
  - chemical sensing/detection.
- **Modern Chemical Instrumentation offers enourmous capabilities for signal recording and storage**





- **Then signal / data analysis and interpretation may become the bottleneck in the process.**
- **The ultimate goal is to extract hidden information in the data**

Machine Learning Definition

- **Machine:**
  - A mechanically, electrically or electronically operated device for performing a task
- **Learning:**
  - The activity or process of gaining knowledge or skill by studying, practicing, being taught, or experience something
- **From Wikipedia:**
  - "“…is a subfield of computer science, evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions”

Machine Learning

vs

Statistical Learning??.

Multivariate description of samples:

Features and patterns

- **Pattern**
  - A composite of traits or features characteristic of an individual/sample
  - In classification tasks, a pattern is a pair of variables $\{x, \omega\}$ where
    - ***x*** is a collection of observations or features (feature vector)
    - ω is the concept behind the observation (label)

Features and patterns

- **Pattern**
  - A composite of traits or features characteristic of an individual/sample
  - In classification tasks, a pattern is a pair of variables $\{x,\omega\}$ where
    - $\boldsymbol{x}$ is a collection of observations or features (feature vector)
    - ω is the concept behind the observation (label)

- **Body Temperature ?**
- **Heart Rate ?**
- **Respiratory Rate ?**
- **Blood Pressure ?**
- **Cough ?**
- **Fatigue Level ?**
- **Gastrointestinal symptoms ?**

Feature Table

	BT	HR	RR	BP	Cough	Fatigue	Vomits	Diarreah	Y
Alex	37.5	60	0.5	150	1	3	0	0	sick
Maria									Sick
Julia	36	90	2	120	1	1	0	0	healthy
Jan									Healthy
Jose									Sick
Eric									Sick
Michael									sik

X block

Y block

Eva?	37	80	0.2	90	1	2	0	0	?

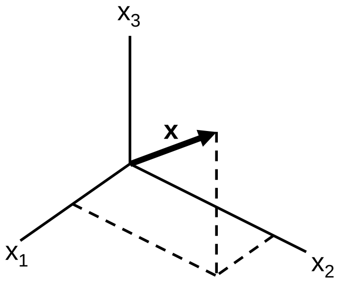
Features, patterns and classifiers

Feature

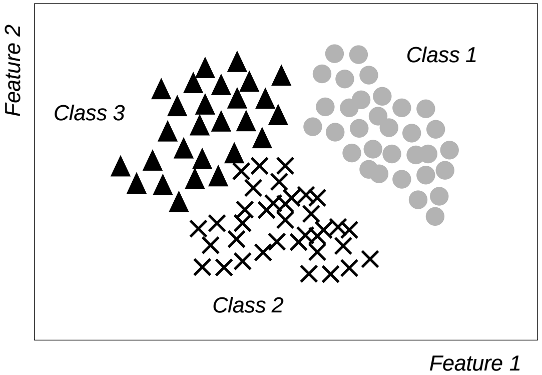
- The combination of *d* features is represented as a *d*-dimensional column vector called a **feature vector**
  - The d-dimensional space defined by the feature vector is called **feature space**
  - Objects are represented as points in feature space. This representation is called a **scatter plot**

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Feature vector



Feature space (3D)

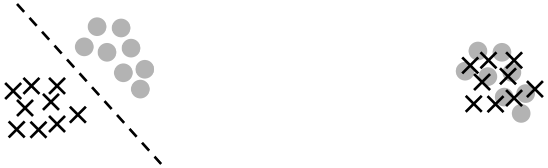


Scatter plot (2D)

Adapted from R. Gutierrez-Osuna

Features, patterns and classifiers

- **What makes a "good" feature vector?**
  - The quality of a feature vector is related to its ability to discriminate examples from different classes
    - Examples from the same class should have similar feature values
    - Examples from different classes have different feature values

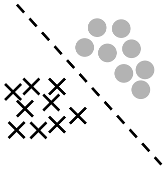


"Good" features ***"Bad" features***

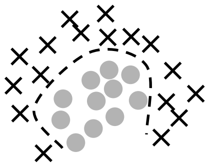
Adapted from R. Gutierrez-Osuna

Features, patterns and classifiers

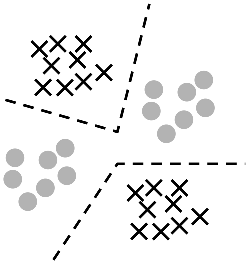
More feature properties



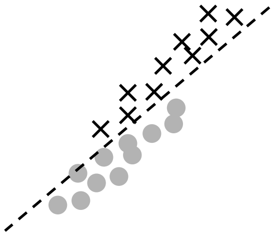
Linear separability



Non-linear separability



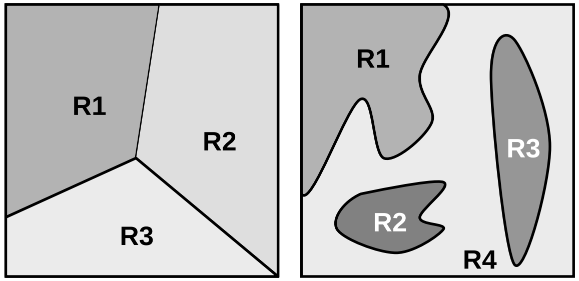
Multi-modal



Highly correlated features

Classifiers

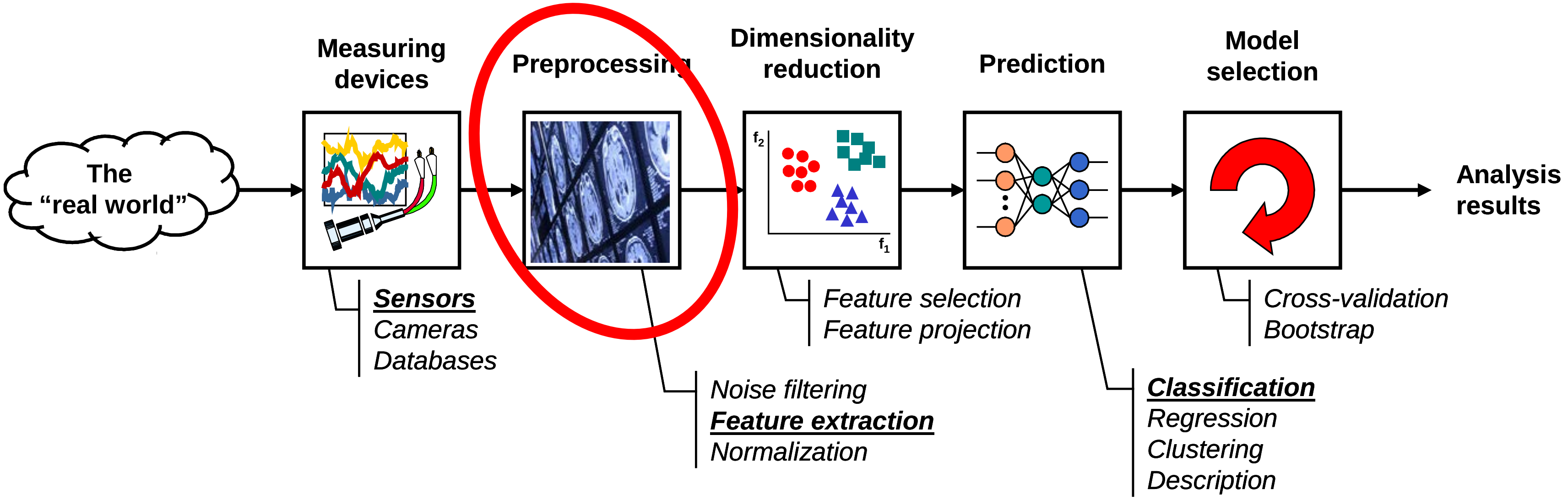
- The goal of a classifier is to partition feature space into class-labeled **decision regions**
- Borders between decision regions are called **decision boundaries**



Adapted from R. Gutierrez-Osuna

Components of Machine Learning Solution

- **A basic system contains**
  - A Measuring Devices (GeneChip, LC-MS, Sequencing Machine, Imaging technique)
  - A preprocessing mechanism **(VIP!!!)**
  - A feature extraction mechanism **(VIP!!)**
  - A classification algorithm (or quantitative predictor)
  - A set of examples (training set) already classified or described



The "real world"

Measuring devices

Preprocessing

Dimensionality reduction

Prediction

Model selection

Analysis results

Sensors
Cameras
Databases

Noise filtering
Feature extraction
Normalization

Feature selection
Feature projection

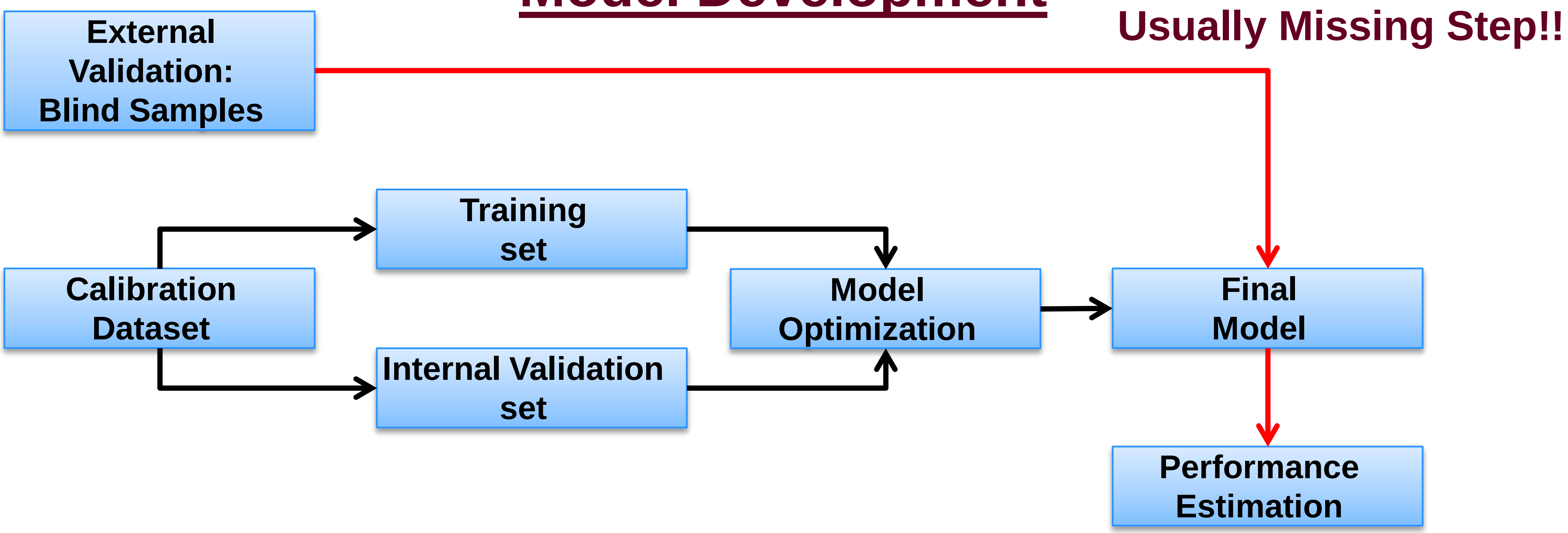
Classification
Regression
Clustering
Description

Cross-validation
Bootstrap

Adapted from R. Gutierrez-Osuna

Philosophy of Predictive Models

Model Development



Usually Missing Step!!

External Validation: Blind Samples

Training set

Calibration Dataset

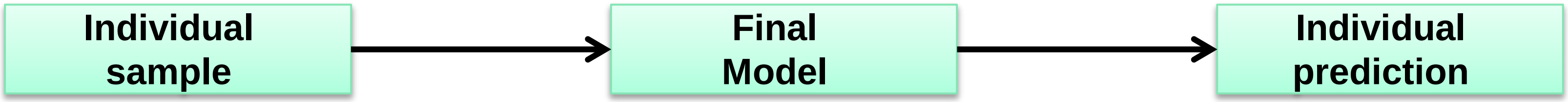
Model Optimization

Final Model

Internal Validation set

Performance Estimation

Deployment



Individual sample

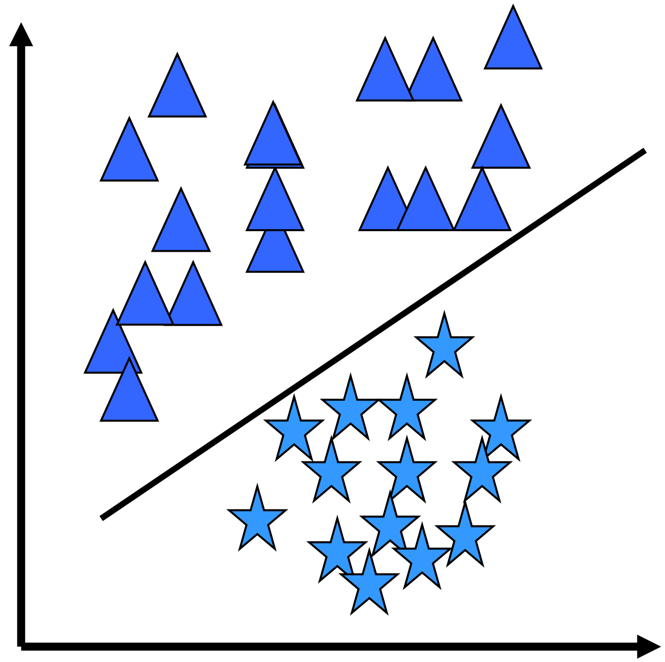
Final Model

Individual prediction

To ensure model quality (generalization) external validation is a must

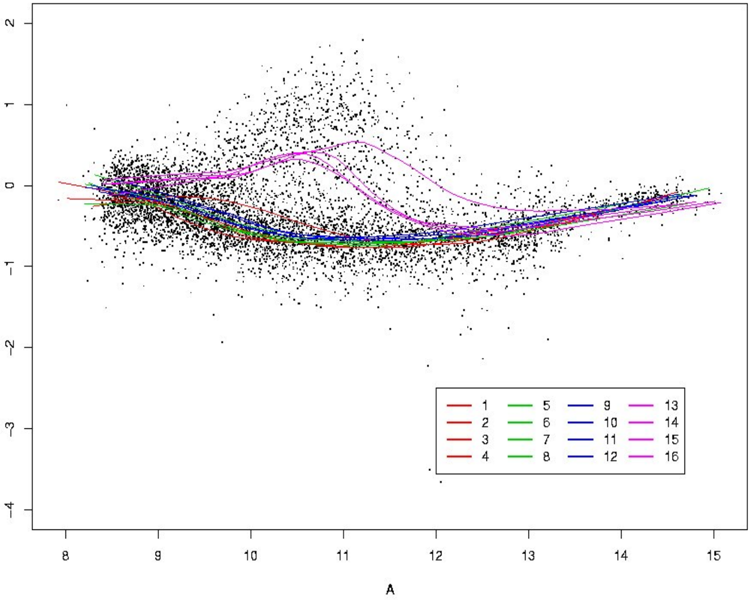
Introduction to Statistical Learning: Classification

- **Given:**
  - input/output (x,t) data pairs, (x continuous, t categorical)
- **Question:**
  - What‘s the best label t given x_{new}?
  - What‘s the probability of t given x_{new}?
- **Figures of merit (binary)**
  - Accuracy
  - Sensitivity
  - AUC
- **Main algorithms:**
  - PLS-DA
  - Random Forest
  - SVM
  - ANN



Supervised Learning: Regression

- **Given:**
  - input/output (**x**,t) data pairs
  - (**x**, t continuous)
  - (**x** will be in general a vector)

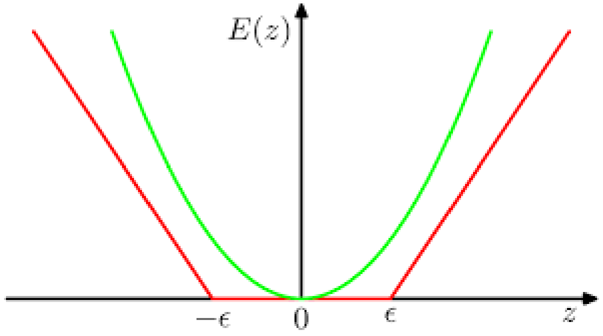


- **Generated by an underlying t=f(x)+nois**

Question:

How the function f(x) look like?
What‘s the best value of t given x
What‘s the probability of t given x? p(t/x)

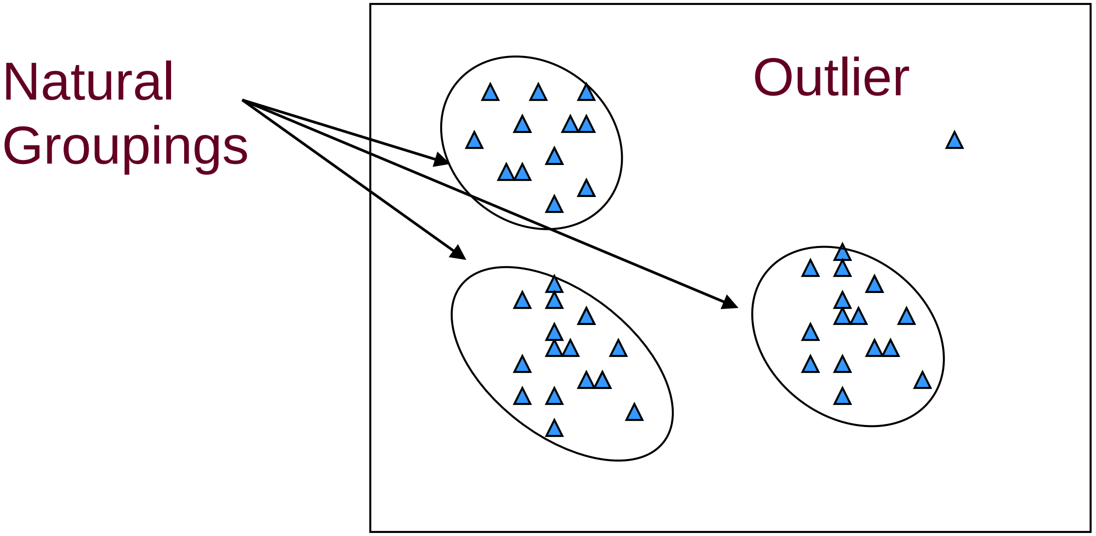
- **Figure of merit**
  - Root Mean Square Error in Prediction (L2 loss)
  - Absolute Error (L1 loss)



- **Main Algorithms**
  - Support Vector Regression
  - Partial Least Squares
  - ANN

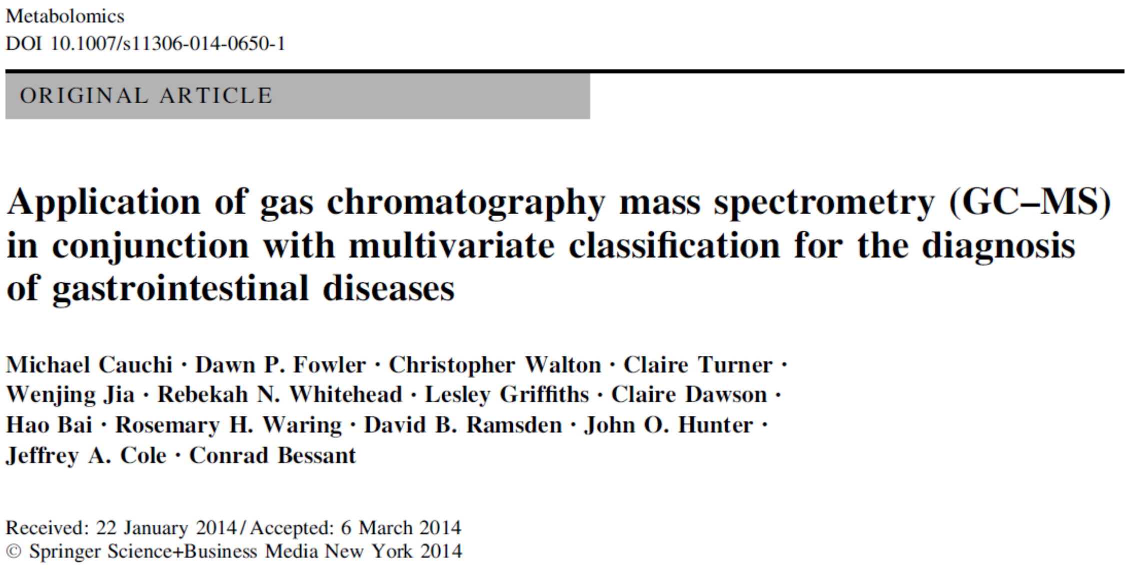
Introduction to Statistical Learning: Clustering

- **Given some data points z_i (dataset)**
- **We need a *model* that captures the *important structure***
- **Typically the model is a probability distribution p(z)**
- **This is an** Unsupervised Learning **problem**



STATISTICAL LEARNING EXAMPLES IN HEALTH

One Example in Metabolomics



Metabolomics
DOI 10.1007/s11306-014-0650-1

ORIGINAL ARTICLE

Application of gas chromatography mass spectrometry (GC–MS) in conjunction with multivariate classification for the diagnosis of gastrointestinal diseases

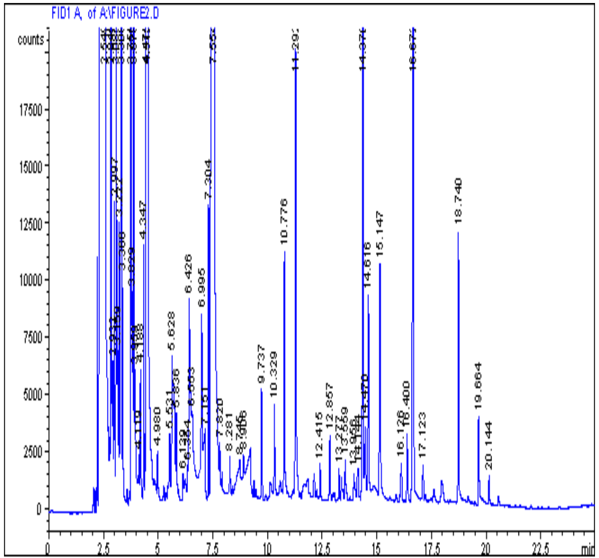
Michael Cauchi · Dawn P. Fowler · Christopher Walton · Claire Turner · Wenjing Jia · Rebekah N. Whitehead · Lesley Griffiths · Claire Dawson · Hao Bai · Rosemary H. Waring · David B. Ramsden · John O. Hunter · Jeffrey A. Cole · Conrad Bessant

Received: 22 January 2014 / Accepted: 6 March 2014

- Samples: Faeces
- 91 patients: 20 healthy, 24 Chron Desease,

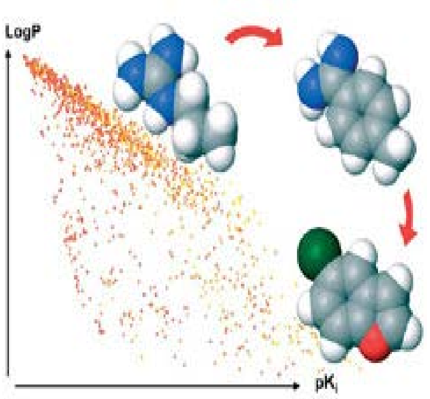
19 Ulcerative Colitis, 28 Irritable Bowel Syndrome

- Technique: Headspace-GC-MS
- Pre-processing: Norm. +Alignment + Outlier Detection
- Feature extraction: Total Ion Chromatogram
- Predictive Model: PLS-DA (Multivariate)
- Internal Validation + Permutation test.

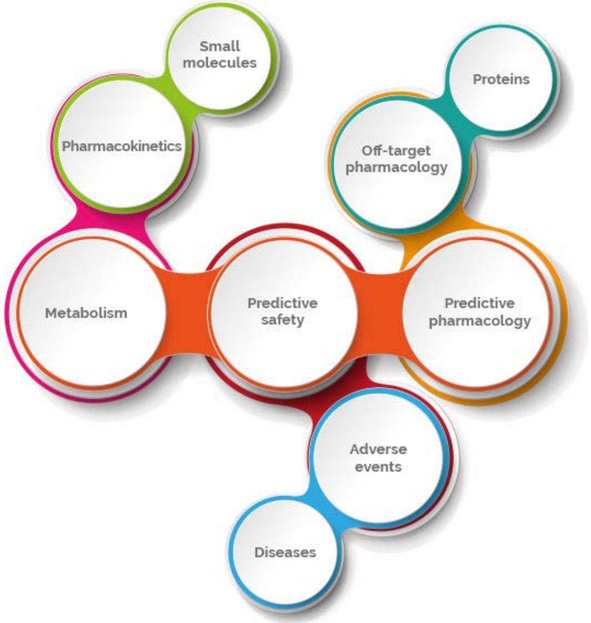


Statistical Learning in Pharma

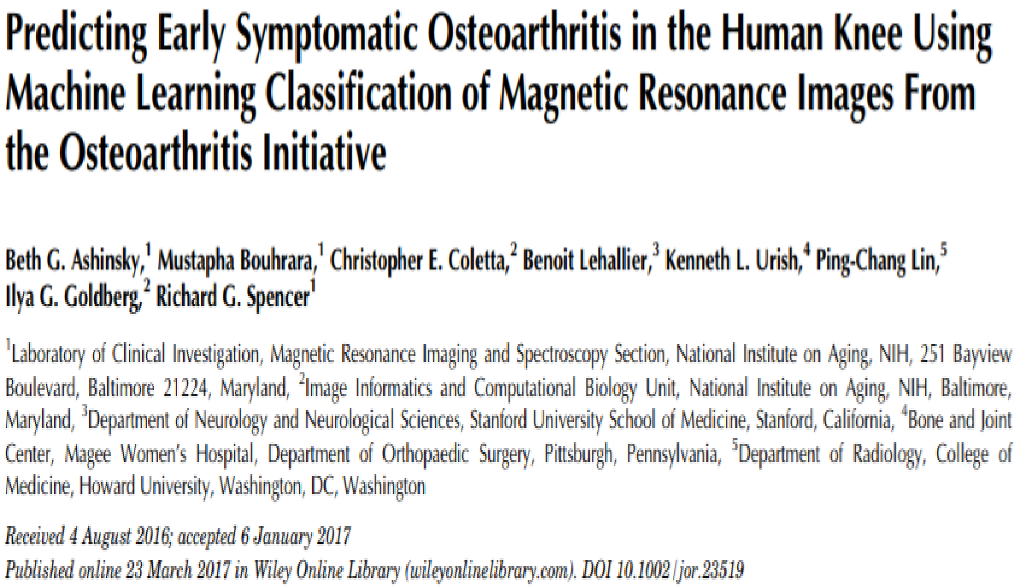
- **Drug Discovery (Boehringer Ingelheim)**
  - The objective is to predict the bioactivity of a molecule from thousands of molecular descriptors. In the Biberach site, they work on small molecule drug discovery programs for cardio-metabolic, CNS, and respiratory diseases. The group is equipped with industry standard Chemoinformatics and Computational Chemistry software and has access to high-performance computing resources.



- **Predicting Toxicity of new Drugs (Chemotargets)**
  - Based on proprietary methodology and an expertly curated database of metabolic transformations (Chemotargets MetDB), Chemotargets CLARITY is able to predict metabolites with a high degree of accuracy and precision. The generated metabolites can be assessed for target activity and safety and their profiles can be compared with substrates.



Statistical Learning Applications

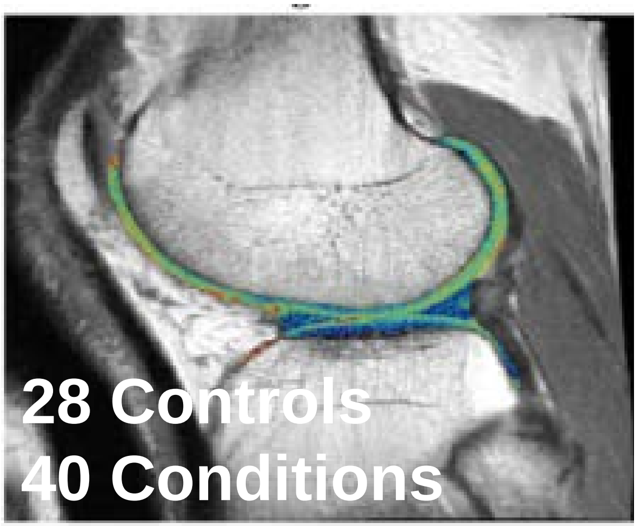


Predicting Early Symptomatic Osteoarthritis in the Human Knee Using Machine Learning Classification of Magnetic Resonance Images From the Osteoarthritis Initiative

Beth G. Ashinsky,[1] Mustapha Bouhrara,[1] Christopher E. Coletta,[2] Benoit Lehallier,[3] Kenneth L. Urish,[4] Ping-Chang Lin,[5] Ilya G. Goldberg,[2] Richard G. Spencer[1]

[1]Laboratory of Clinical Investigation, Magnetic Resonance Imaging and Spectroscopy Section, National Institute on Aging, NIH, 251 Bayview Boulevard, Baltimore 21224, Maryland, [2]Image Informatics and Computational Biology Unit, National Institute on Aging, NIH, Baltimore, Maryland, [3]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, California, [4]Bone and Joint Center, Magee Women's Hospital, Department of Orthopaedic Surgery, Pittsburgh, Pennsylvania, [5]Department of Radiology, College of Medicine, Howard University, Washington, DC, Washington

Received 4 August 2016; accepted 6 January 2017
Published online 23 March 2017 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/jor.23519



28 Controls
40 Conditions

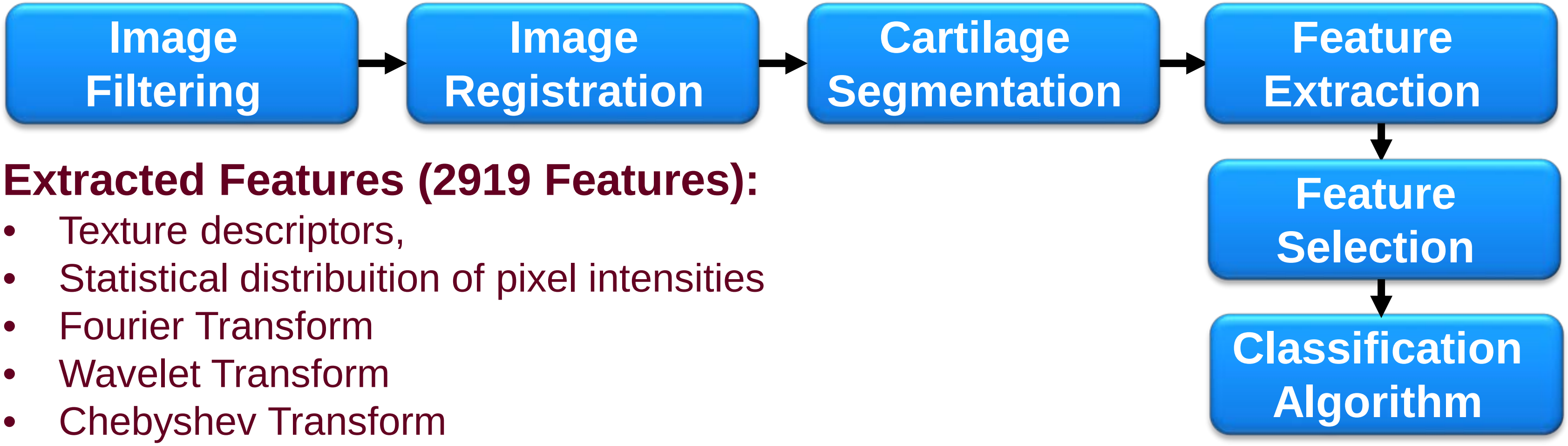


Image Filtering → Image Registration → Cartilage Segmentation → Feature Extraction → Feature Selection → Classification Algorithm

Extracted Features (2919 Features):

- Texture descriptors,
- Statistical distribuition of pixel intensities
- Fourier Transform
- Wavelet Transform
- Chebyshev Transform

Statistical Learning Applications

Classifying Histopathology Images

- A pathologist's report after reviewing a patient's biological tissue samples is often the gold standard in the diagnosis of many diseases.
- Today Histopathology Images at 40X have Gigapixel size.
- Google has developed a Deep learning approach that won an international contest (ISBI-Camelyon) for the classification of cancer lesions.
- the prediction heatmaps produced by the algorithm had improved so much that the localization score (FROC) for the algorithm reached 89%, which significantly exceeded the score of 73% for a pathologist with no time constraint

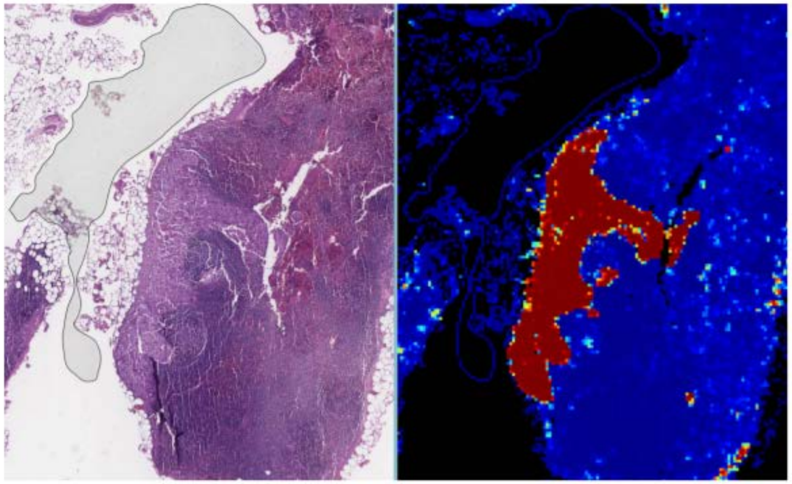
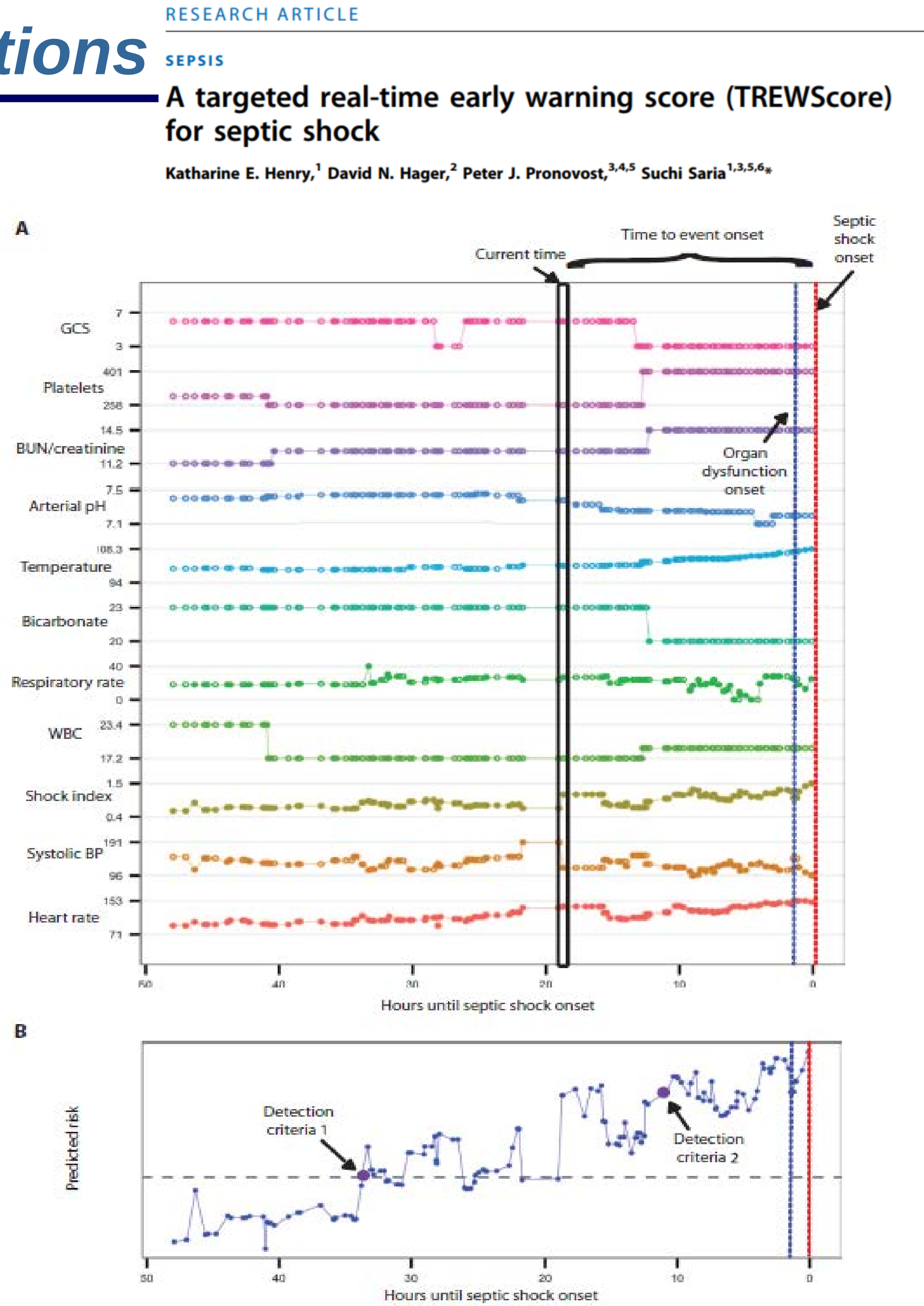


Fig. 8. Left: a patch from a H&E-stained slide in our additional validation set, NHO-1. The tumor cells are a lighter purple than the surrounding cells. A variety of artifacts are visible: the dark continuous region in the top left quadrant is an air bubble, and the white parallel streaks in the tumor and adjacent tissue are cutting artifacts. Furthermore, the tissue is hemorrhagic, necrotic and poorly processed, leading to color alterations to the typical pink and purple of a H&E slide. **Right**: the corresponding predicted heatmap that accurately identifies the tumor cells while ignoring the various artifacts, including lymphocytes and the cutting artifacts running through the tumor tissue.

Statistical Learning Applications

- **Sepsis is a major cause of death. Arises when body reaction to infection damages organs and tissues.**
- **Sepsis early symptons are unespecific:**
  - Fever, increased heart rate, increased breath rate
- **TREWSscore a machine learning prediction algorithm identified patients before Sepsis Shock with AUC=0.83 > 0.73 current algorithm.**
- **TREWSscore detected sepsis with a median of 28 h before sepsis shock onset.**
- **(Mortality incresase 7% per each hour delay).**



RESEARCH ARTICLE

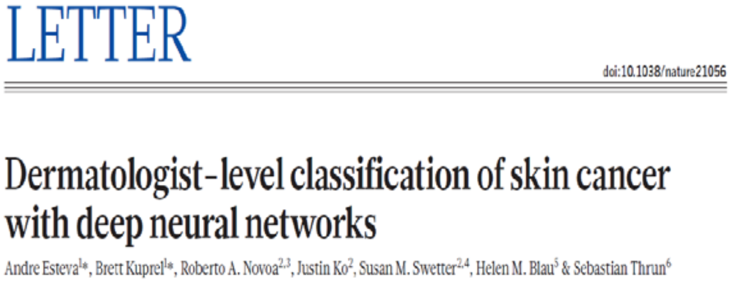
SEPSIS

A targeted real-time early warning score (TREWScore) for septic shock

Katharine E. Henry,[1] David N. Hager,[2] Peter J. Pronovost,[3,4,5] Suchi Saria[1,3,5,6]*

Statistical Learning Applications

- Deep convolutional Neural Networks were used to detect skin cancer lessions.
- A main challange was to assamble and annotate a database of 130.000 images.
- Lession labels extended to 2032 diseases.

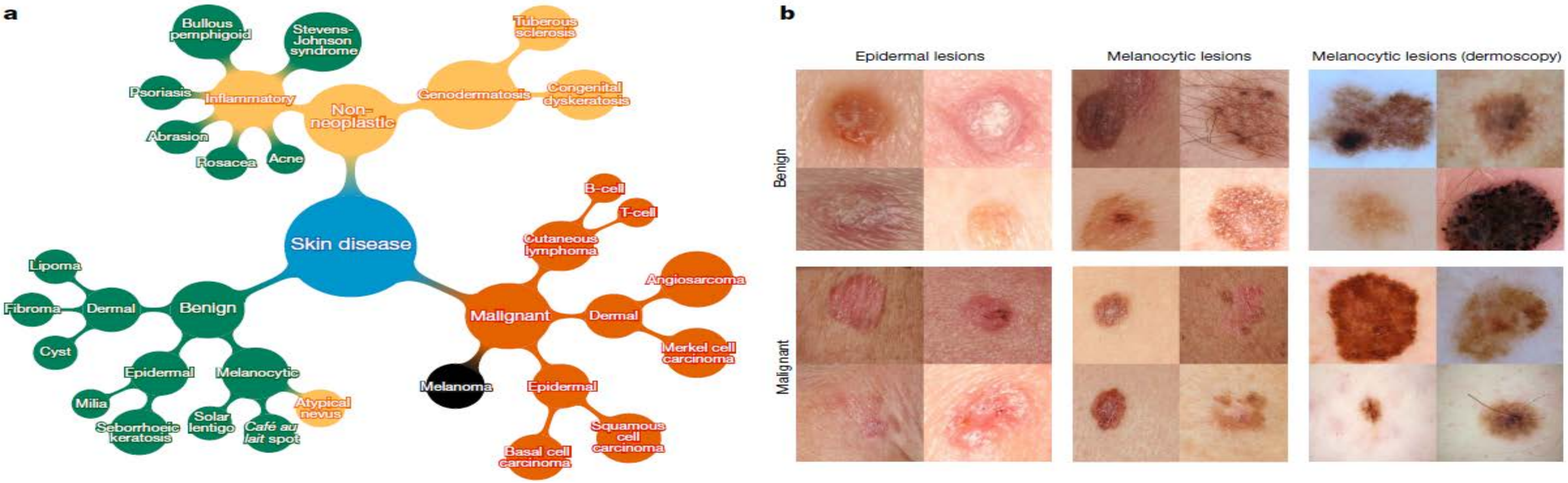


LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1*], Brett Kuprel[1*], Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]



Statistical Learning Applications in Bioinformatics

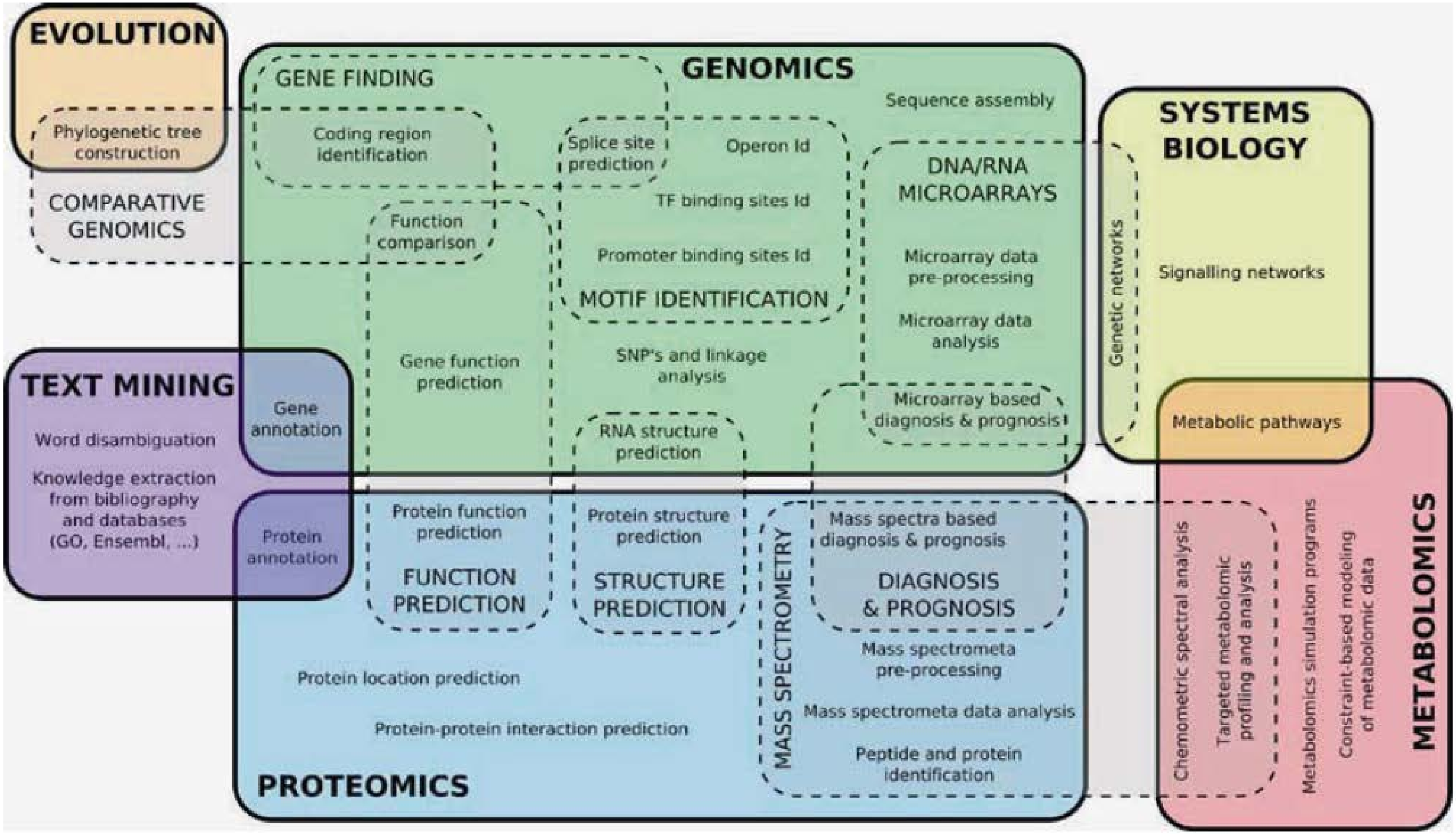

EVOLUTION
Phylogenetic tree construction
COMPARATIVE GENOMICS
GENE FINDING
GENOMICS
Coding region identification
Splice site prediction
Operon Id
Sequence assembly
Function comparison
TF binding sites Id
Promoter binding sites Id
MOTIF IDENTIFICATION
DNA/RNA MICROARRAYS
Microarray data pre-processing
Microarray data analysis
Microarray based diagnosis & prognosis
Genetic networks
SYSTEMS BIOLOGY
Signalling networks
Metabolic pathways
TEXT MINING
Word disambiguation
Knowledge extraction from bibliography and databases (GO, Ensembl, ...)
Gene annotation
Protein annotation
Gene function prediction
SNP's and linkage analysis
RNA structure prediction
Protein function prediction
FUNCTION PREDICTION
Protein structure prediction
STRUCTURE PREDICTION
MASS SPECTROMETRY
Mass spectra based diagnosis & prognosis
DIAGNOSIS & PROGNOSIS
Mass spectrometa pre-processing
Mass spectrometa data analysis
Peptide and protein identification
Protein location prediction
Protein-protein interaction prediction
PROTEOMICS
Chemometric spectral analysis
Targeted metabolomic profiling and analysis
Metabolomics simulation programs
Constraint-based modeling of metabolomic data
METABOLOMICS

Fig. 2.2. General scheme of the current applications of machine learning techniques in bioinformatics.

From “Bioinformatics methods in Clinical Reseach”, Springer (2010)

ML for Bioinformatics and Health Data Science

- **No programming skills required**
- **KNIME (https://www.knime.com/)**
  - Graphical User Interface
  - Flexible & Extensible
  - Bioinformatics-friendly nodes
- **Weka**
  - Comprehensive Algorithms
  - Data Analysis & Prediction
  - Not Bioinformatics-specific
- **RapidMiner**
  - User-friendly
  - Advanced Analytics Processes
  - Healthcare Templates
- **Orange**
  - Visual Workflow
  - Data Visualization & Analysis
  - Bioinformatics Add-on

ML for Bioinformatics and Health Data Science

- **No programing skills required**
- **KNIME (https://www.knime.com/)**
  - Graphical User Interface
  - Flexible & Extensible
  - Bioinformatics-friendly nodes
- **Solo**
  - Comprehensive Algorithms
  - Data Analysis & Prediction
  - Geared towards Chemometrics
- **RapidMiner**
  - User-friendly
  - Advanced Analytics Processes
  - Healthcare Templates
- **Orange**
  - Visual Workflow
  - Data Visualization & Analysis
  - Bioinformatics Add-on

ML for Bioinformatics and Health Data Science

- **R-studio (Posix) :**
  - https://posit.co/download/rstudio-desktop/
- **Bioconductor:**
  - Open source software for Bioinformatics
  - https://www.bioconductor.org/
- **CRAN:**
  - The Comprehensive R Archive Network
  - https://cran.r-project.org/
- **Relevant packages in the R ecosystem:**
  - General purpose ML: *caret, mixOmics*
  - Regression: *glmnet, lars*….
  - Classification: *e1071, randomForest*,
  - Clustering: *cluster, dbscan*
  - Neural Networks: *nnet, keras*…..

ML for Bioinformatics and Health Data Science

- **Python has a rich ecosystem of libraries and frameworks geared towards data analysis:**

- **General purpose ML**
  - Scikit-learn
  - XGBoost

- **Deep Learning Frameworks**
  - TensorFlow
  - Pytorch

- **Data Manipulation and Visualization**
  - Pandas
  - Matplotlib/Seaborn

- **Specialized libraries:**
  - BioPython

ML for Bioinformatics and Health Data Science

- **MATLAB:**
  - Bioinformatics toolbox:
    - Includes basic machine learning/feature selection
  - Statistics and Machine Learning toolbox
    - Full suite of methods for advanced statistics and machine learning
- **Third party toolboxes**
  - PLS toolbox
    - Full suite for Building Predictive Models
    - Provides also a user interface

Summary

- Today we are able to gather a huge amoung of health related data.
- This data can be complex, heterogenous and incomplete
- Machine Learning methods aim to analyze this data in order to extract hidden information
- Machine Learning methods use a variety of tolos
- High Level languages (R, Python, MATLAB, etc) offer possibilities to develop custom solutions.