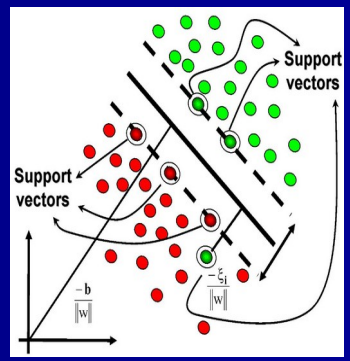
ADVANCED MACHINE LEARNING

-- A PERSONAL VIEW --

LEARNING SCENARIOS AND RESOURCES

Iñaki Inza
Intelligent Systems Group, www.sc.ehu.es/isg
Computer Science Faculty
University of the Basque Country, Donostia - San Sebastian

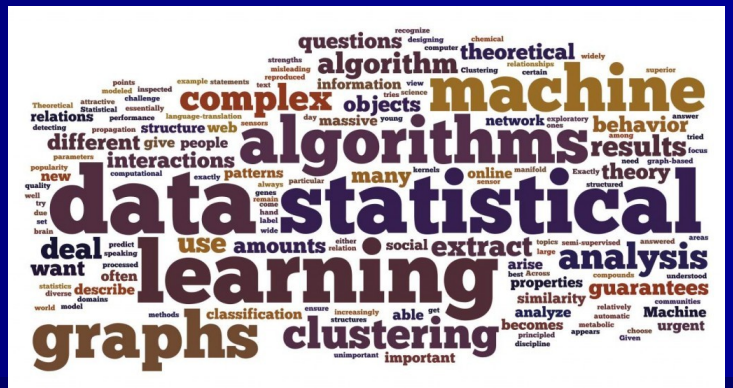


OUTLINE

- Data science: the term
- Learning scenarios:

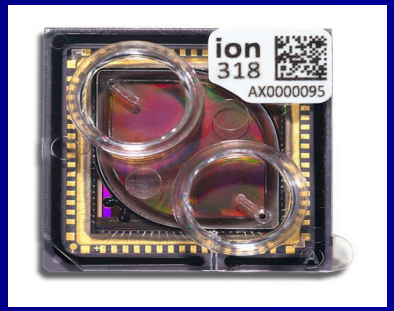
  Type of data matrix ~ Type of data analysis

- Real-life applications for each type of analysis
- Resources: business oportunities, software tools, BigData
- Avoiding formalisms...



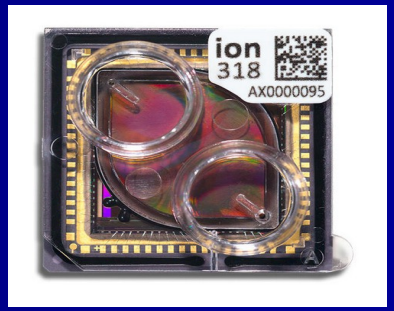
THE AGE OF DATA

- Data collected at enormous speeds in **everyday practice**:
  - text - social networks' activity
  - shop - electronic purchases
  - medicine - monitored patients
  - bioinformatics - gene expression data
  - telescopes scanning the skies
  - industry - Internet of Things
  - ...



THE AGE OF DATA

- Computers and storage systems have become **cheaper** and more **powerful**
- World's technological capacity to store info: **doubled** every 40 months since 80's
- Since 90's, much more data is being stored **than** analyzed (around 5-10%)
- By 2020 estimated... 30.6 exabytes through networks, 35 trillion GB stored, 11.6 billion connected devices
- **Traditiona**l data analysis techniques (70s-80s) **unfeasible** for "modern" data



THE AGE OF DATA

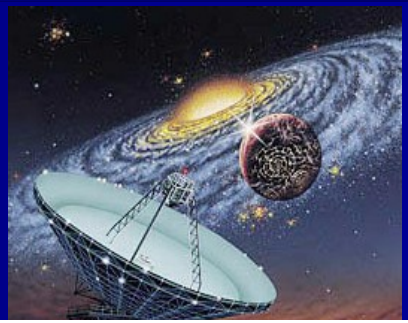
- Computers and storage systems have become **cheaper** and more **powerful**

- Hidden big data. Large quantities of useful data are in fact useless because they are untagged, file-based, and unstructured. The 2012 IDC study on big data [117] explained that, in 2012, 23% (643 exabytes) of the digital universe would be useful for big data if tagged and analyzed. However, at that time only 3% of the potentially useful data was tagged, and even less was analyzed. The figures have probably gotten worse in recent years. The Open Data and Semantic Web movements have emerged, in part, to make us aware and improve on this situation. No comments

- **Traditional** data analysis techniques (70s-80s) **unfeasible** for "modern" data

THE AGE OF DATA

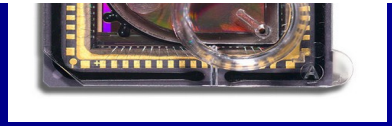
- Computers and storage systems have become **cheaper** and more **powerful**



- World's technological capacity to store info: **doubled** every 40 months since 80's

Labelomania

We are witnessing a data labeling market explosion: labeling platforms have hit prime time. S&P Global released an October 11 report entitled *Avoiding Garbage in Machine Learning* in which it termed unlabeled data "garbage data" to highlight the importance of labeling in AI. The Economist

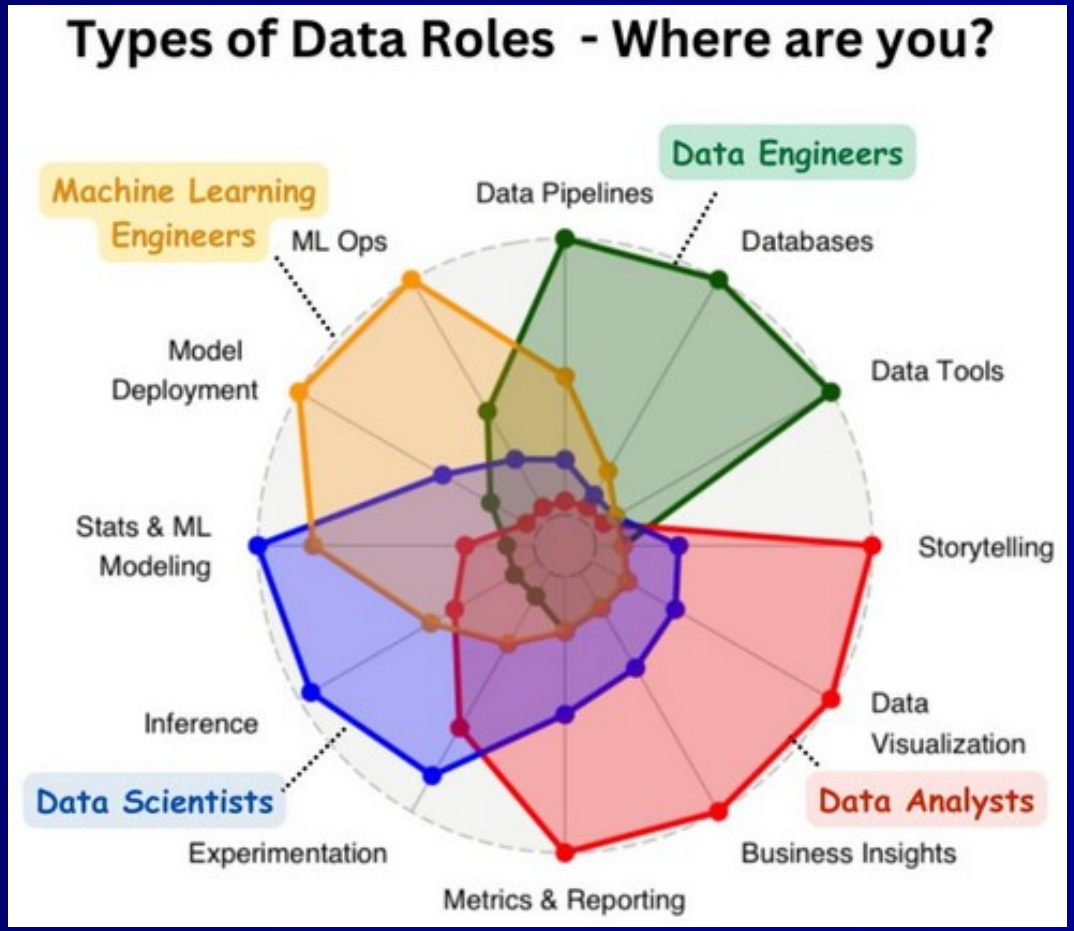


- By 2020 estimated... 30.6 exabytes through networks, 35 trillion GB stored, 11.6 billion connected devices

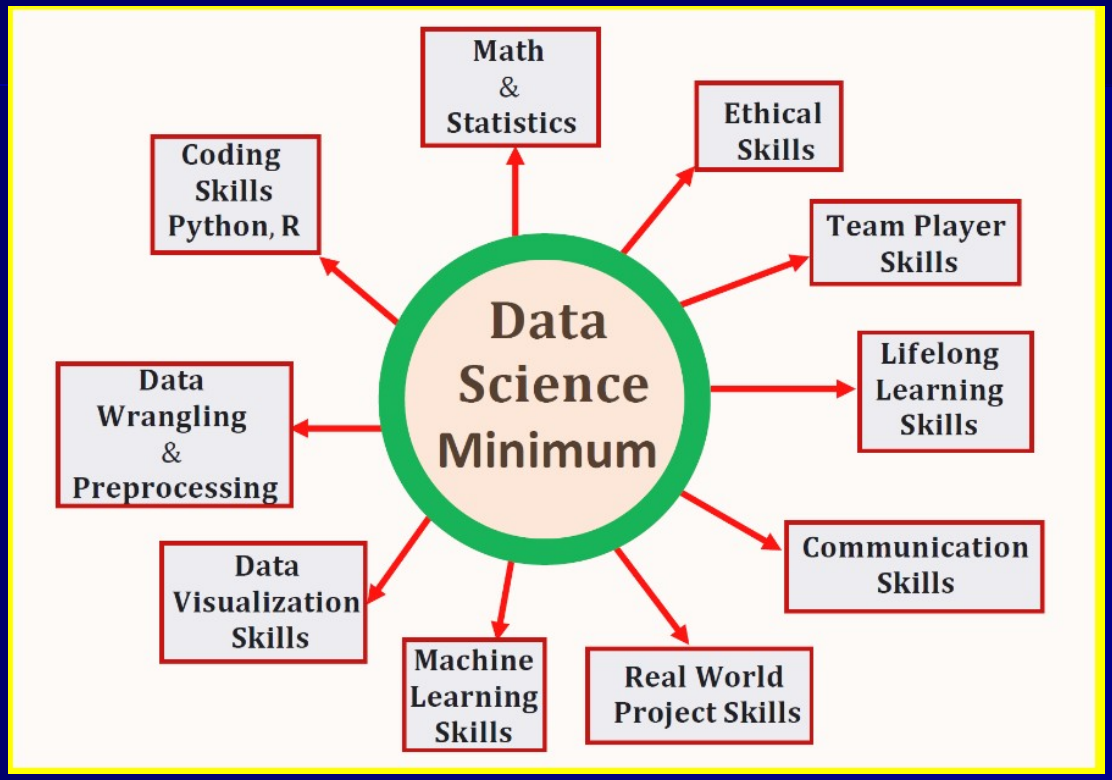
- **Traditiona**l data analysis techniques (70s-80s) **unfeasible** for "modern" data

THE AGE DATA SCIENCE

DATA SCIENCE'S ECOSYSTEM

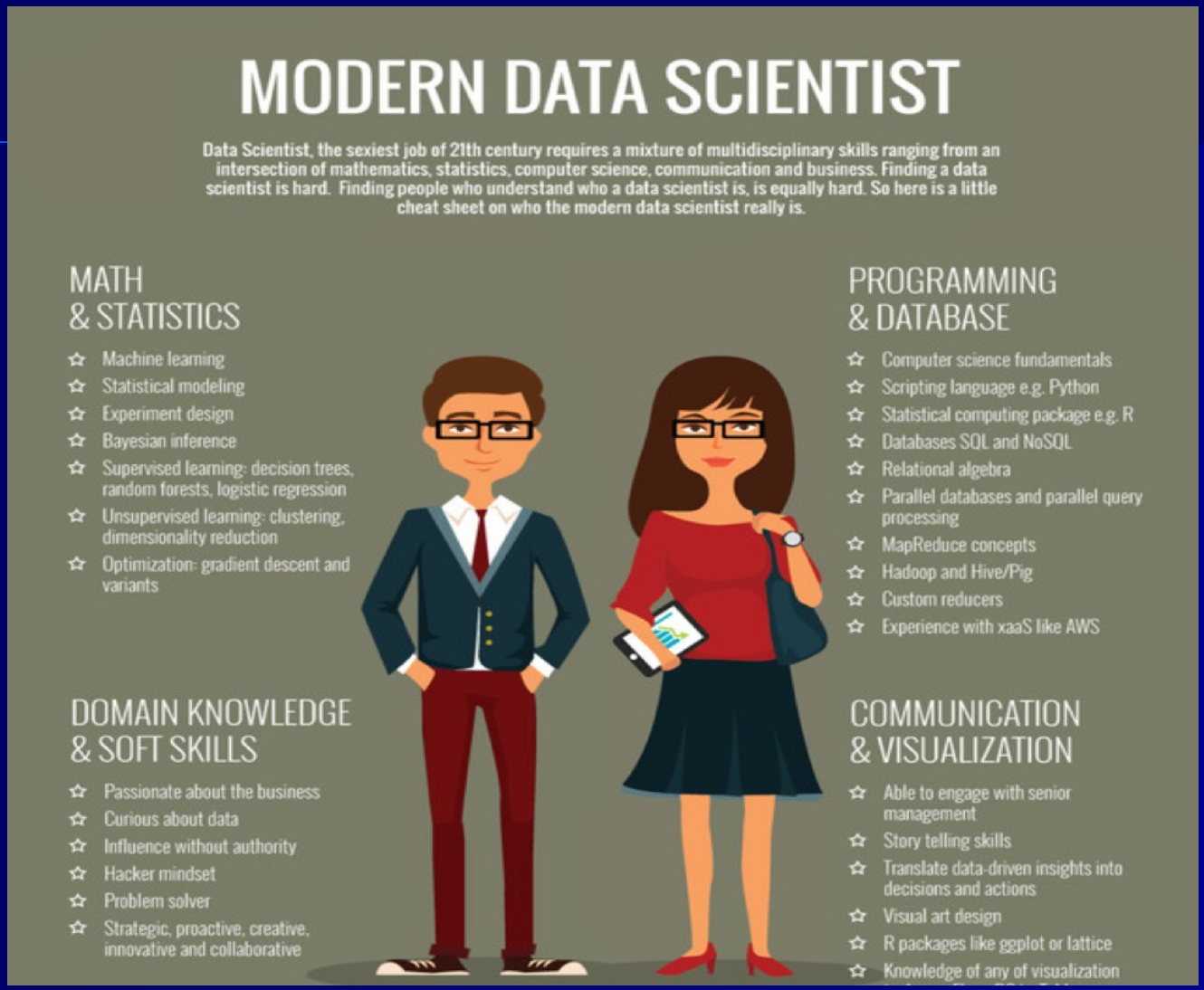


DATA SCIENCE'S ECOSYSTEM

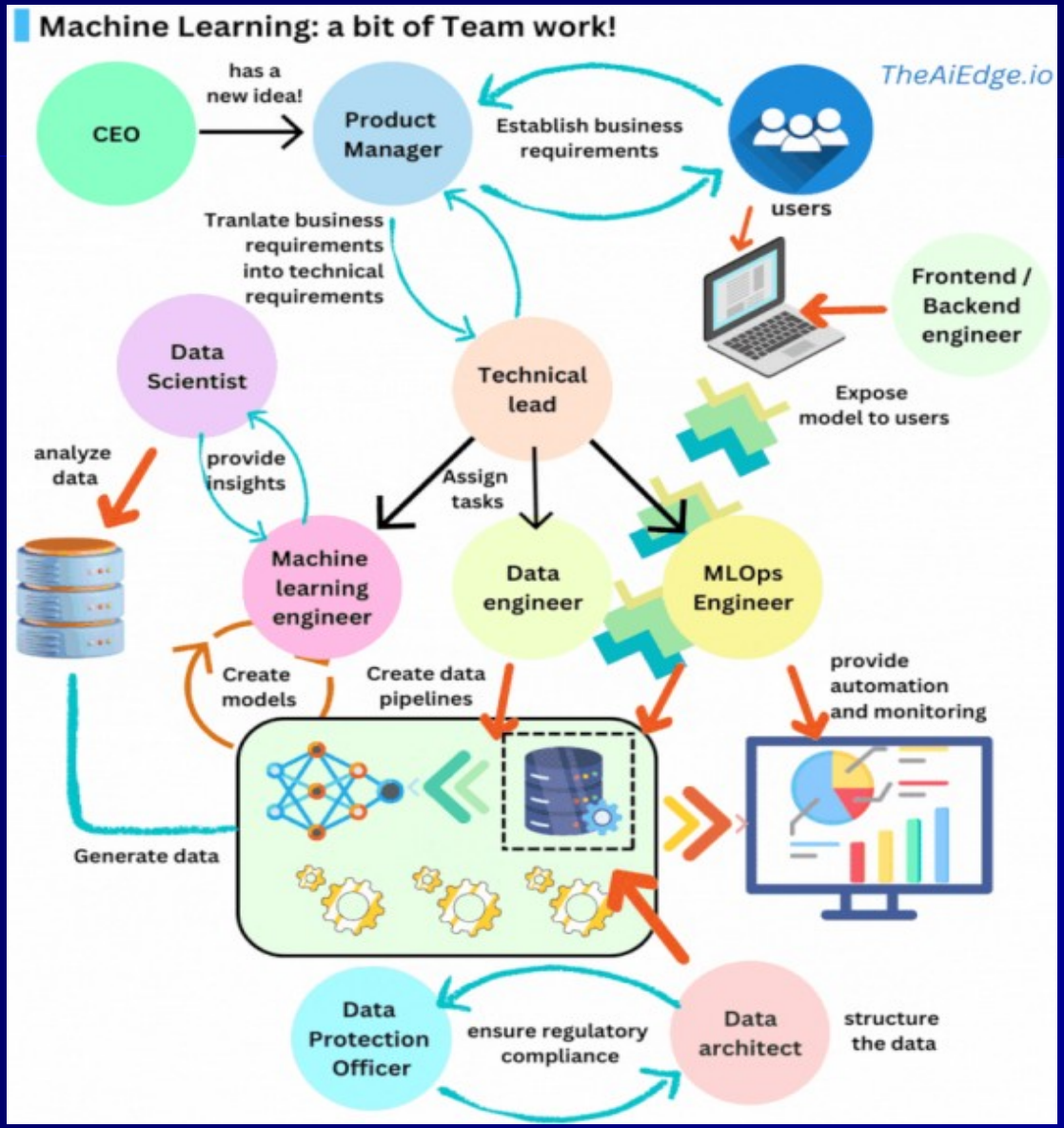


- 7 steps for learning data mining and data science
- 10 "best and free" ML online courses
- Data sciencists versus Statisticians
- Most viewed YouTube videos on data mining
- Tour of real-world machine learning problems
- Top-10 data science videos in Youtube

DATA SCIENCE'S ECOSYSTEM



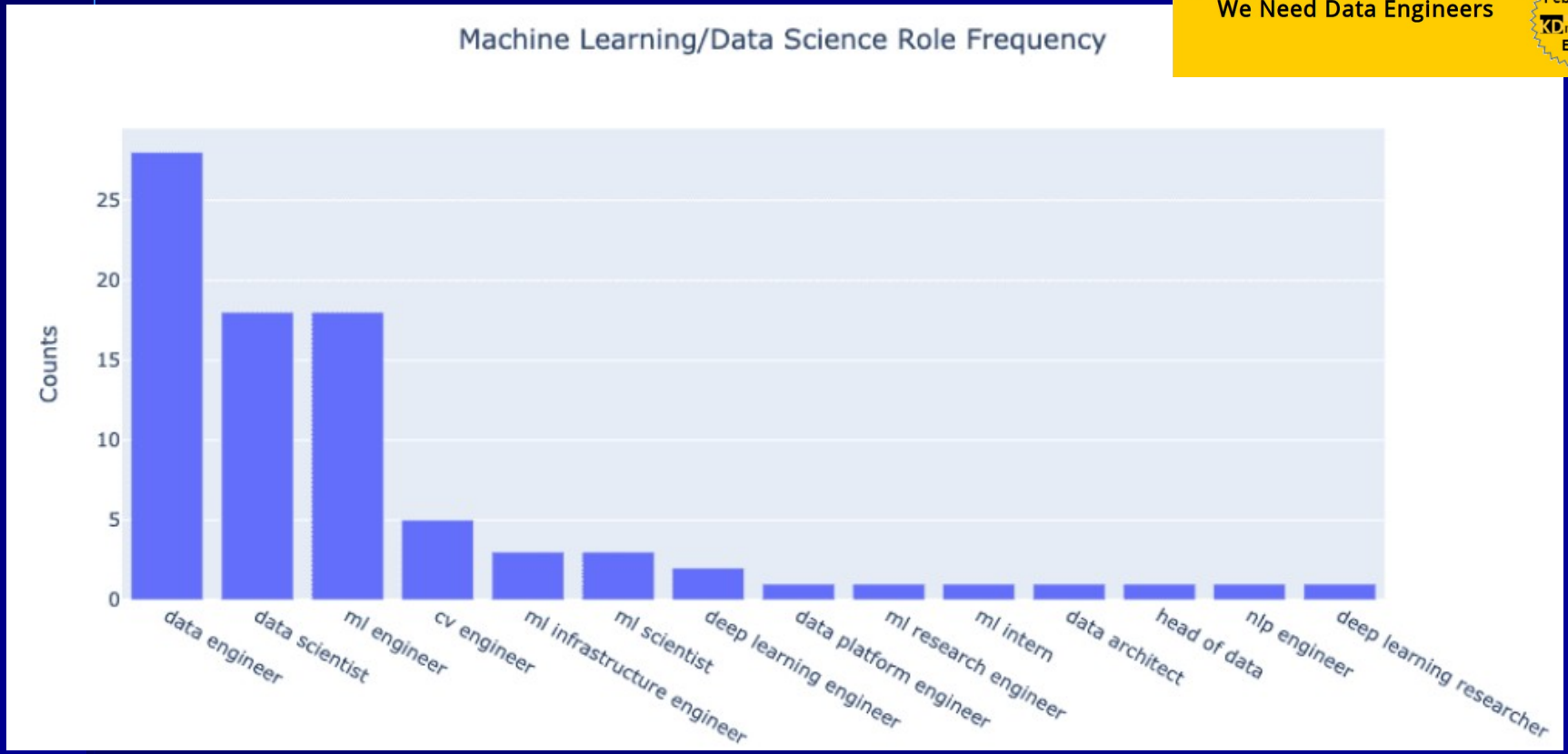
DATA SCIENCE'S ECOSYSTEM



DATA SCIENCE – ROLES

We Don't Need Data Scientists, We Need Data Engineers

Feb 2021 Gold KDnuggets Blog



ARTIFICIAL INTELLIGENCE ~ DATA

"Almost current 95% of AI is based on machine learning from data"

Data: a cornerstone for AI – Toward a Common European Data Space

> *" Good quality shared data is essential to develop socially responsive AI "*

For an application of artificial intelligence (AI) to be ready for market entry it has to learn on the basis of training data. Once in use on the market, it should generate a sufficient amount of data as part of its use.

ARTIFICIAL INTELLIGENCE ~ DATA

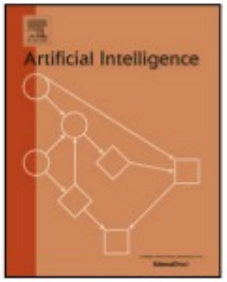
"Almost current 95% of AI is based on machine learning from data"

Artificial Intelligence 289 (2020) 103386

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Artificial Intelligence requires more than deep learning – but what, exactly?

Michael Wooldridge

Department of Computer Science, University of Oxford, United Kingdom of Great Britain and Northern Ireland

ARTIFICIAL INTELLIGENCE ~ IN THE WAY TO CONSCIOUNESS?

EL PAÍS

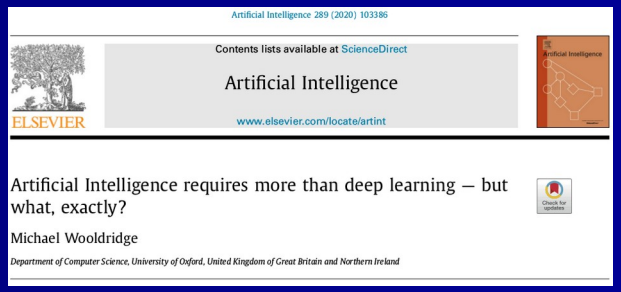
Ciencia / Materia

ASTROFÍSICA · MEDIO AMBIENTE · INVESTIGACIÓN MÉDICA · MATEMÁTICAS · PALEONTOLOGÍA

INTELIGENCIA ARTIFICIAL >

Marcus du Sautoy, matemático: "Existe la posibilidad de que la inteligencia artificial se vuelva consciente"

El catedrático de la Universidad de Oxford plantea el surgimiento de "una nueva especie" a partir de los algoritmos



Artificial Intelligence 289 (2020) 103386

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

ELSEVIER

Artificial Intelligence requires more than deep learning — but what, exactly?

Michael Wooldridge

Department of Computer Science, University of Oxford, United Kingdom of Great Britain and Northern Ireland

EL PAÍS

Ciencia / Materia

ASTROFÍSICA · MEDIO AMBIENTE · INVESTIGACIÓN MÉDICA · MATEMÁTICAS · PALEONTOLOGÍA · ÚLTIMAS NOTICIAS

NOBEL >

Premio Nobel de Física 2024 a John Hopfield y Geoffrey Hinton por poner las bases de la inteligencia artificial

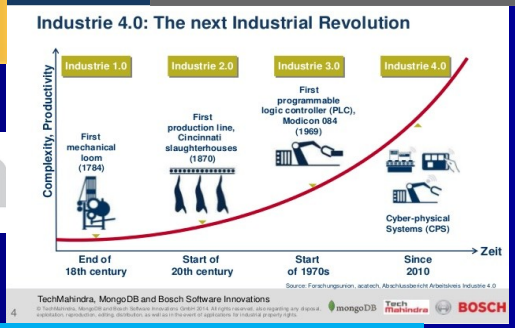
La Academia de Ciencias sueca concede el galardón a los considerados 'padrinos' del aprendizaje de máquinas

BUSINESS OPORTUNITIES

KDnuggets

OPEN DATA

kaggle.com



UNITED NATIONS GLOBAL PULSE
Harnessing big data for development and humanitarian action

GLOBAL PULSE

Big Data, Big Impact:
New Possibilities for International Development

BUSINESS INTELLIGENCE COMPANIES – THE VOCABULARY!

- Data mining → a (one more) tool for BI
- Saving, analysis, sharing
- Objective: help the company in its decisions

BIG DATA: OPORTUNITIES FOR BUSINESS AND DEVELOPMENT

Big Data, Big Impact: New Possibilities for International Development

Executive Summary

A flood of data is created every day by the interactions of billions of people using computers, GPS devices, cell phones, and medical devices. Many of these interactions occur through the use of mobile devices being used by people in the developing world, people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realise the potential for channelling these torrents of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises for the benefit of low-income populations. Concerted action is needed by governments, development organisations, and companies to ensure that this data helps the individuals and communities who create it.

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

BUSINESS OPORTUNITIES

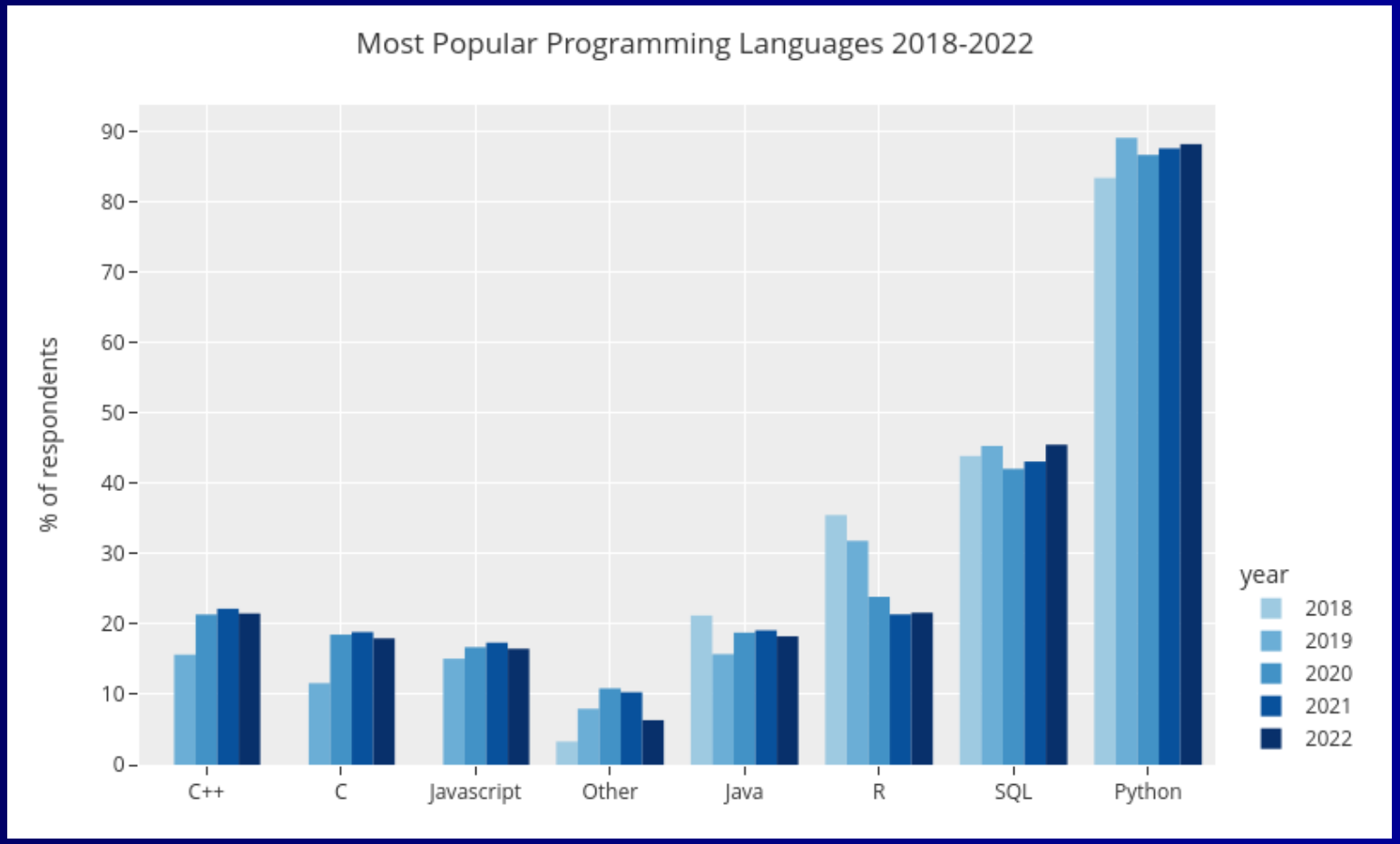
kaggle.com

 **Online Product Sales** Predict the online sales of a consumer product based on a data set of product features.	$22,500	365	14 months ago
Boehringer Ingelheim **Predicting a Biological Response** Predict a biological response of molecules from their chemical properties	$20,000	703	14 months ago
Ford **Stay Alert! The Ford Challenge** Driving while not alert can be deadly. The objective is to design a classifier that will detect whether the driver is alert or not alert, employing data that are acquired while driving.	$950	176	2 years ago

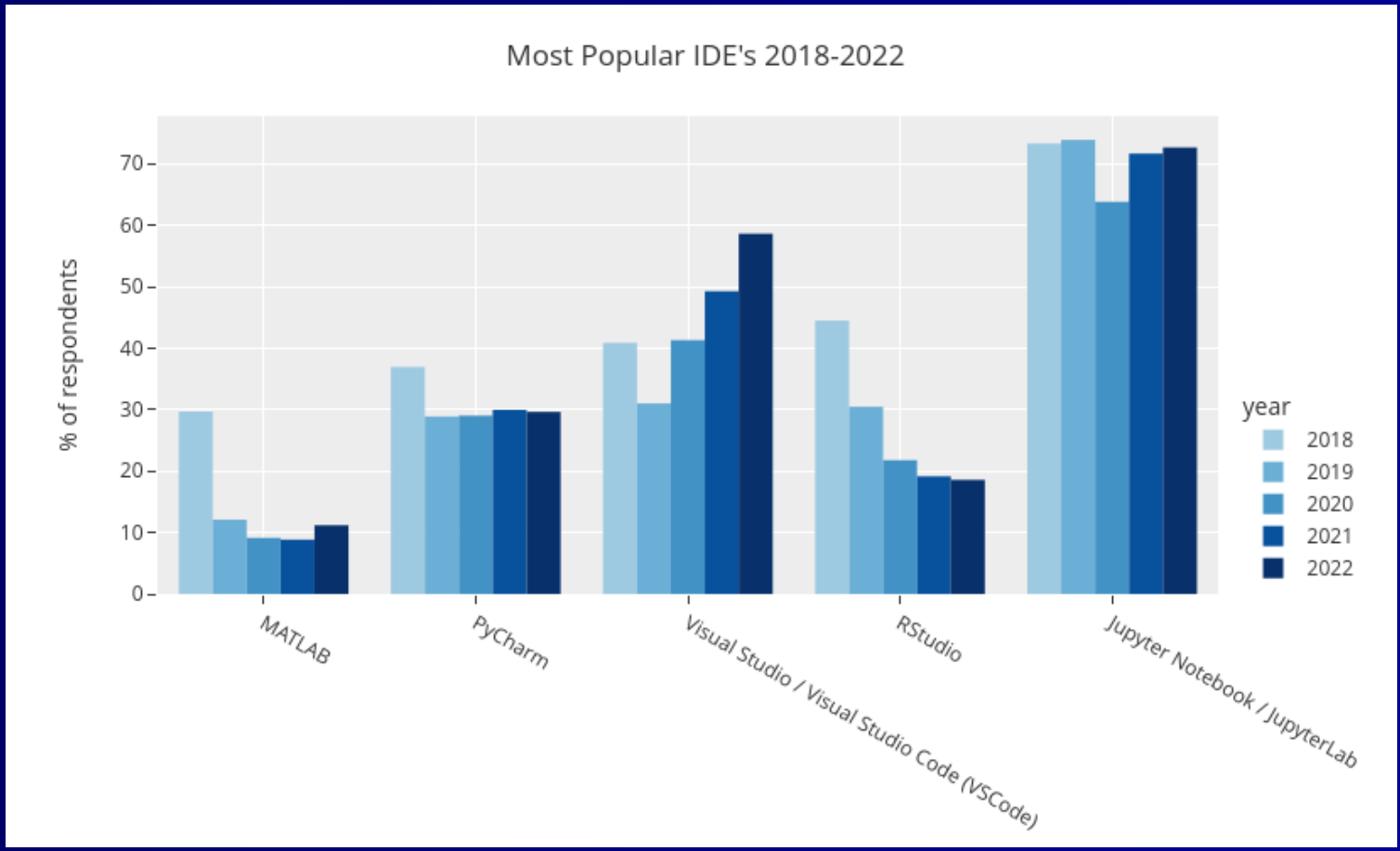
Kaggle competitions

Kaggle Open Datasets

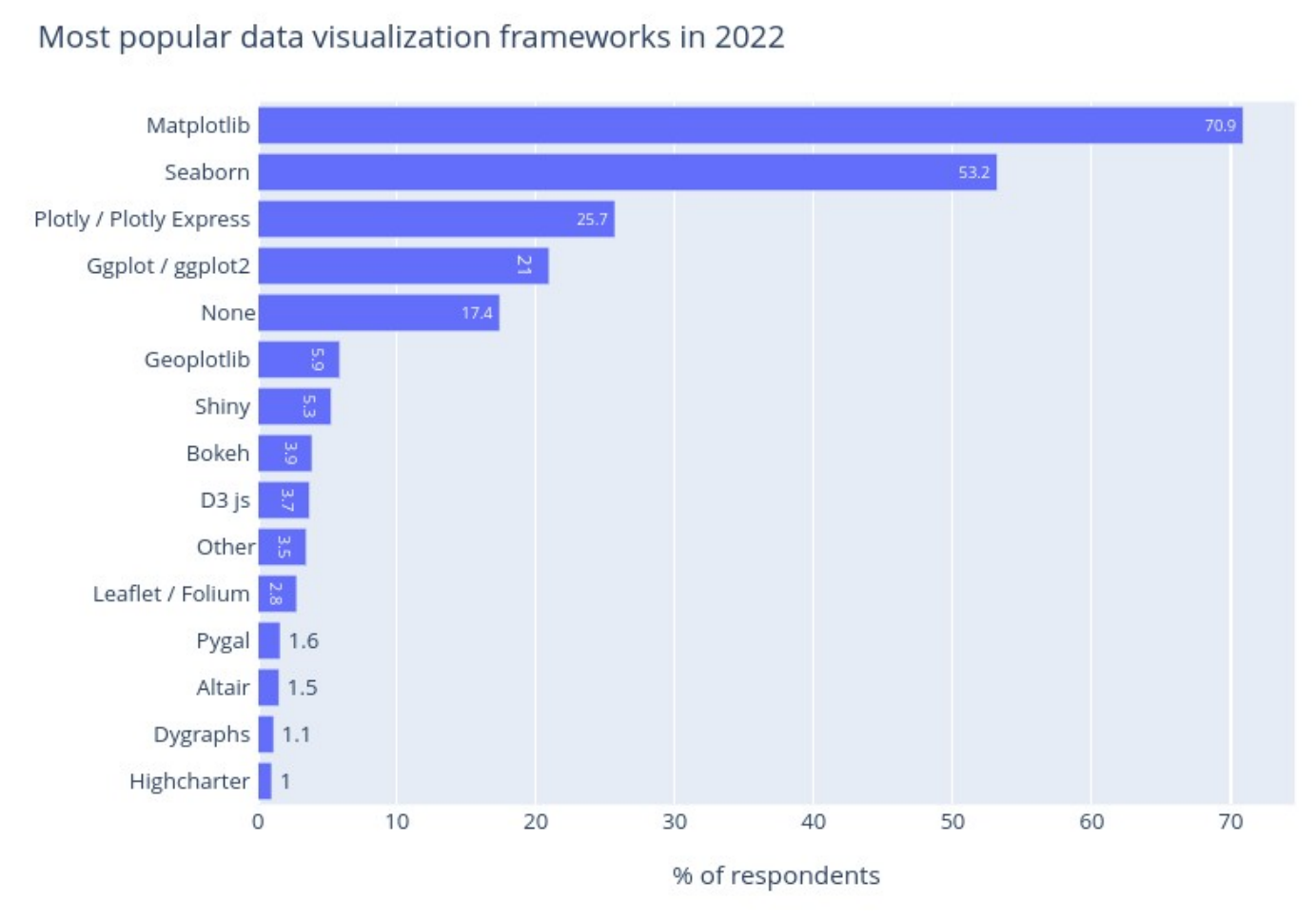
KAGGLE 2022 SURVEY
– 23,997 responses –



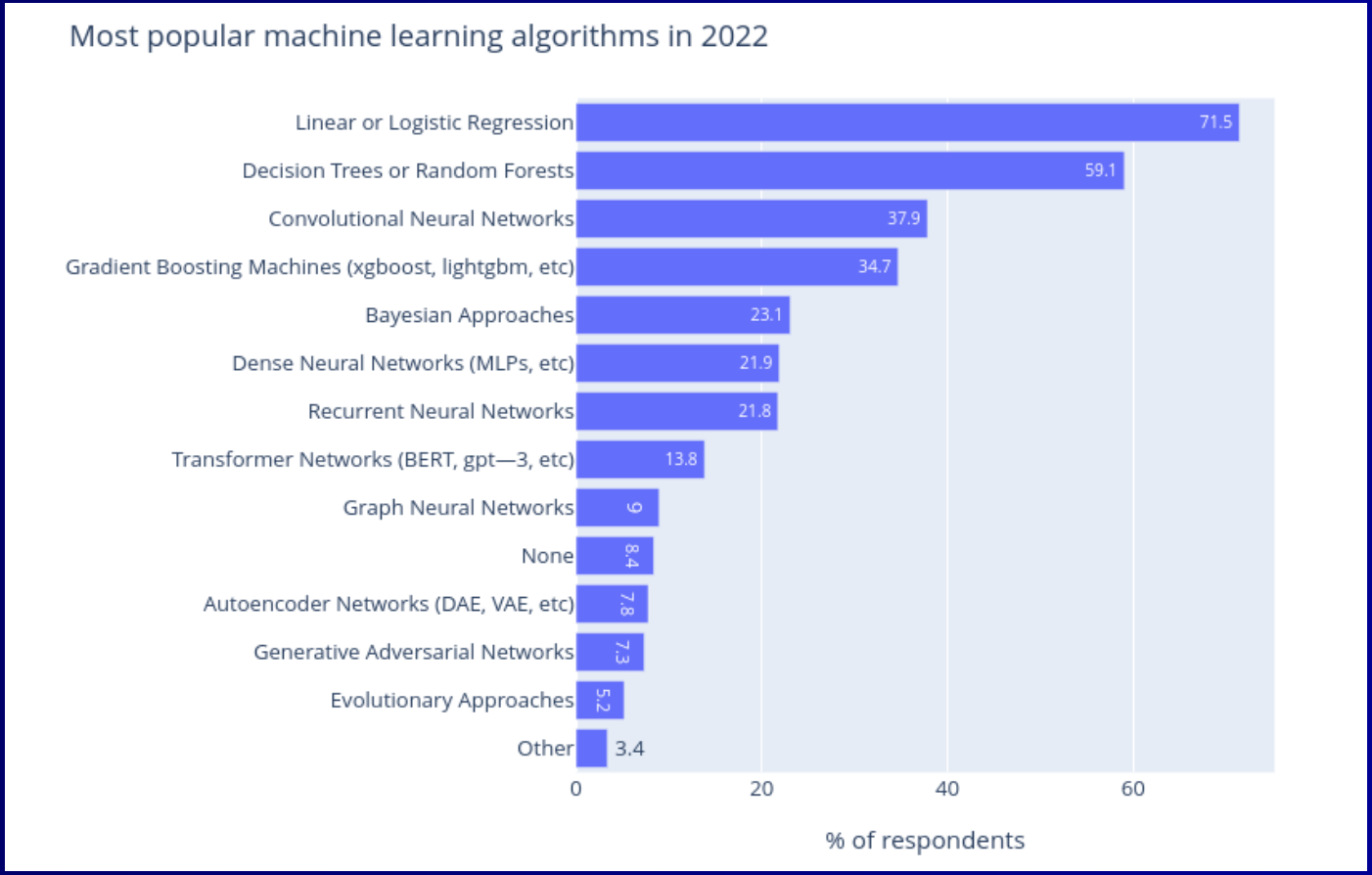
KAGGLE 2022 SURVEY



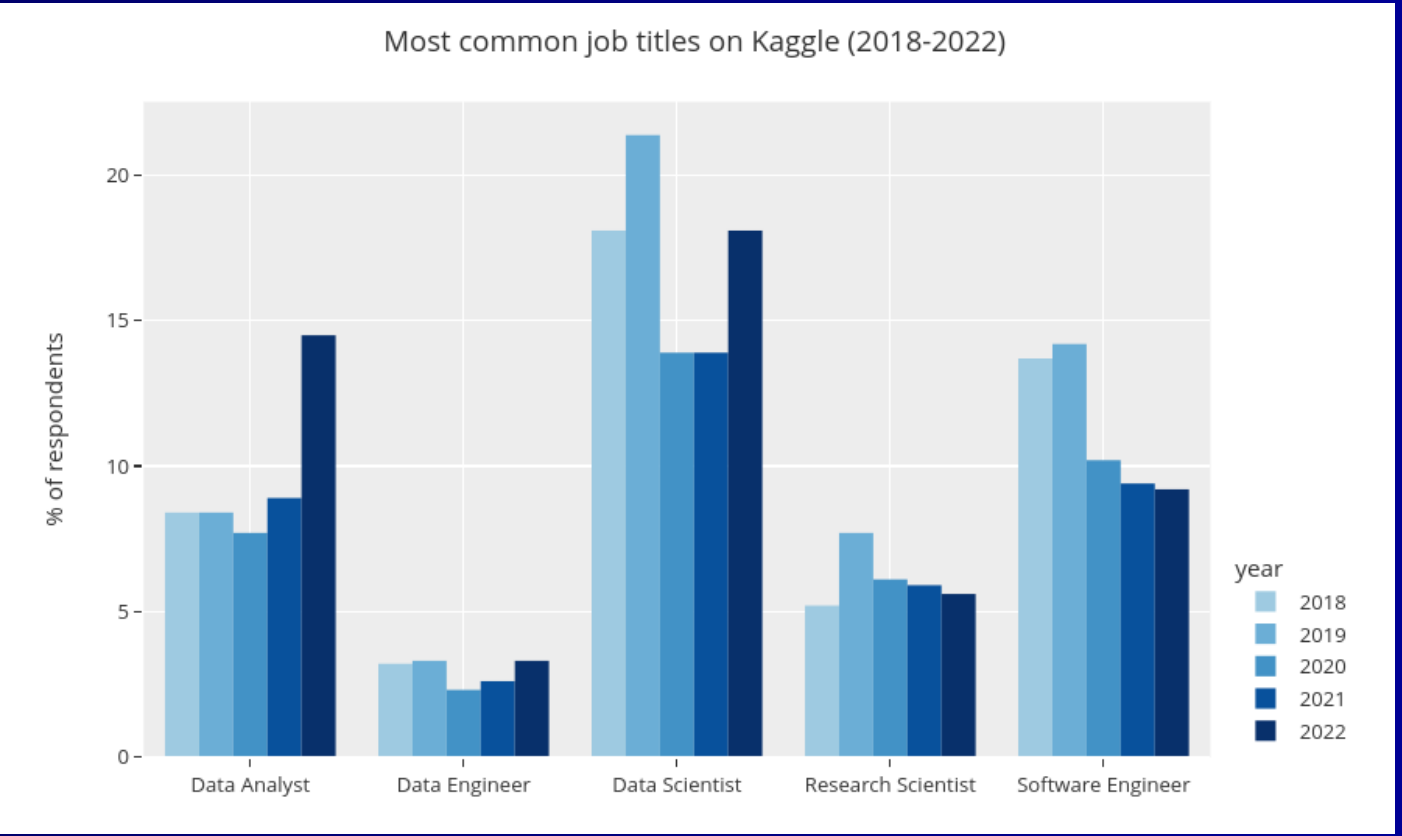
KAGGLE 2022 SURVEY



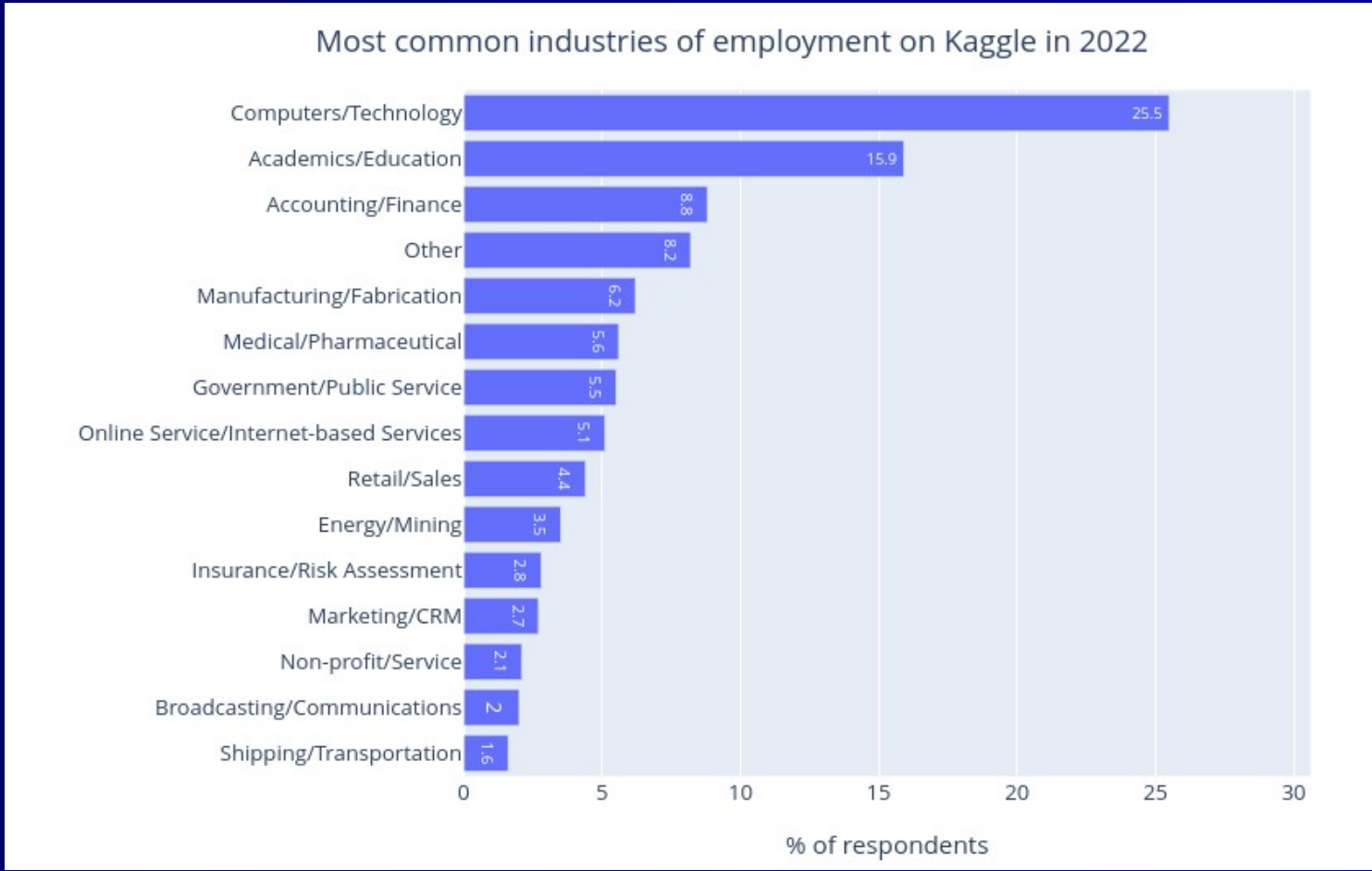
KAGGLE 2022 SURVEY



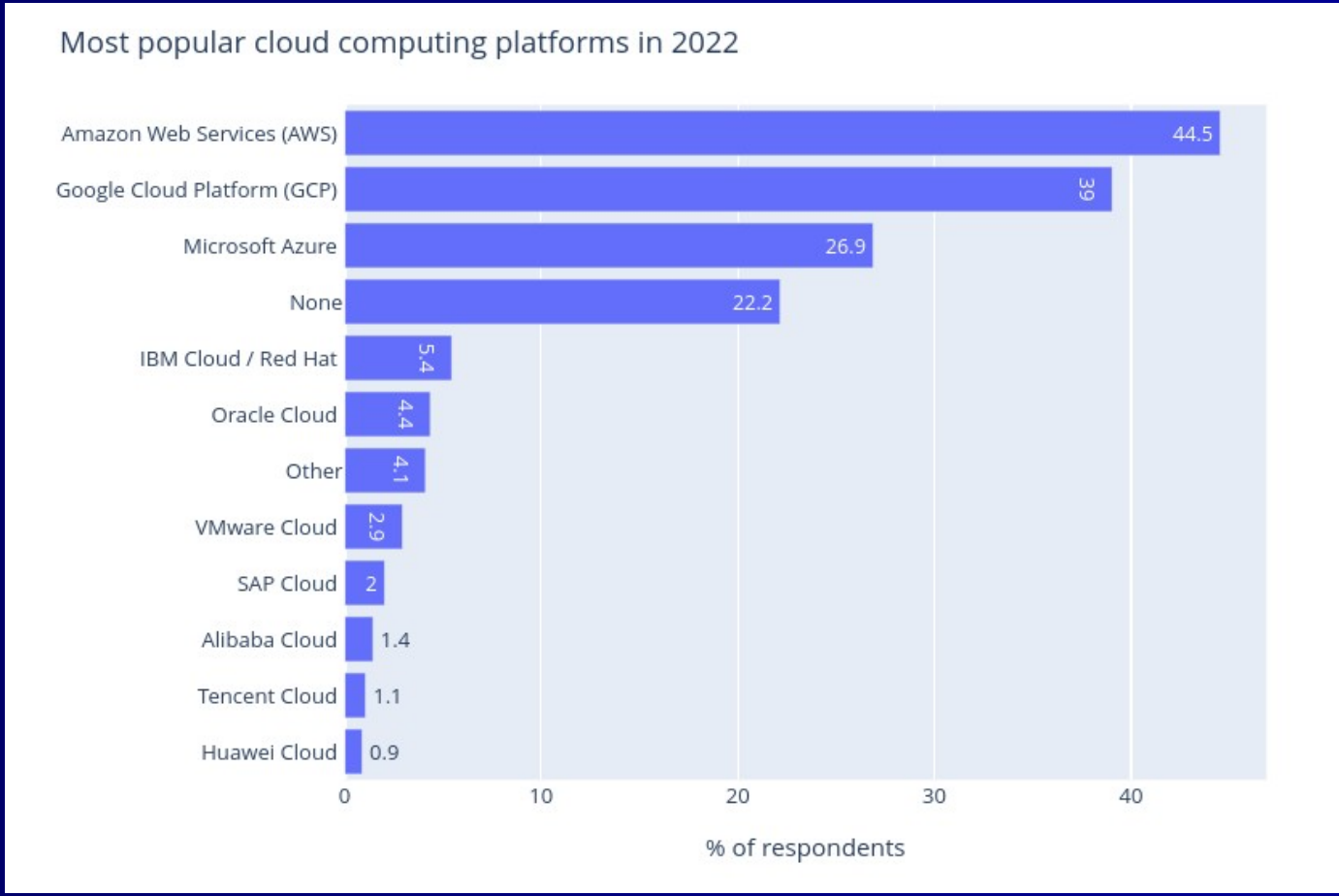
KAGGLE 2022 SURVEY



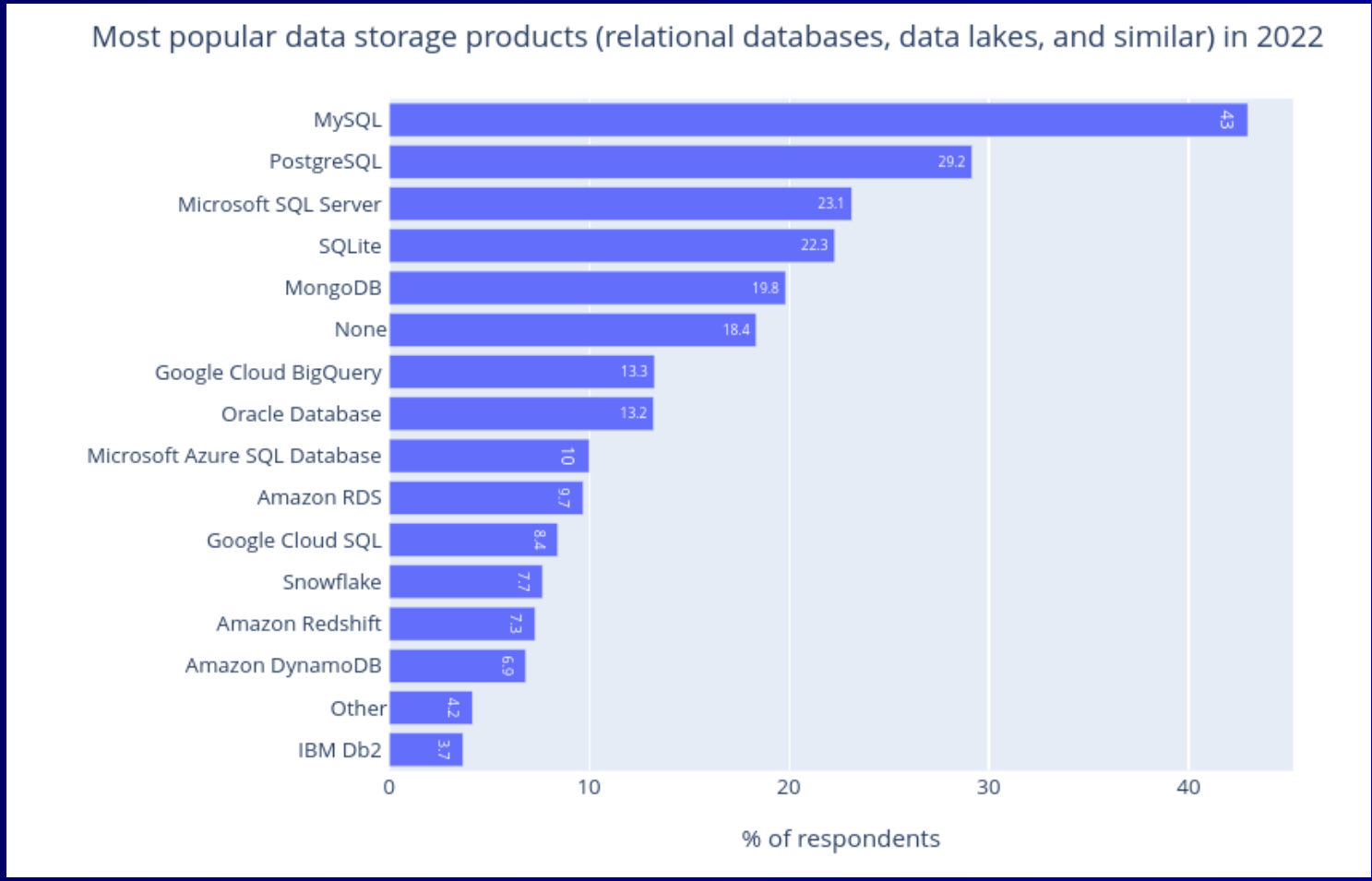
KAGGLE 2022 SURVEY



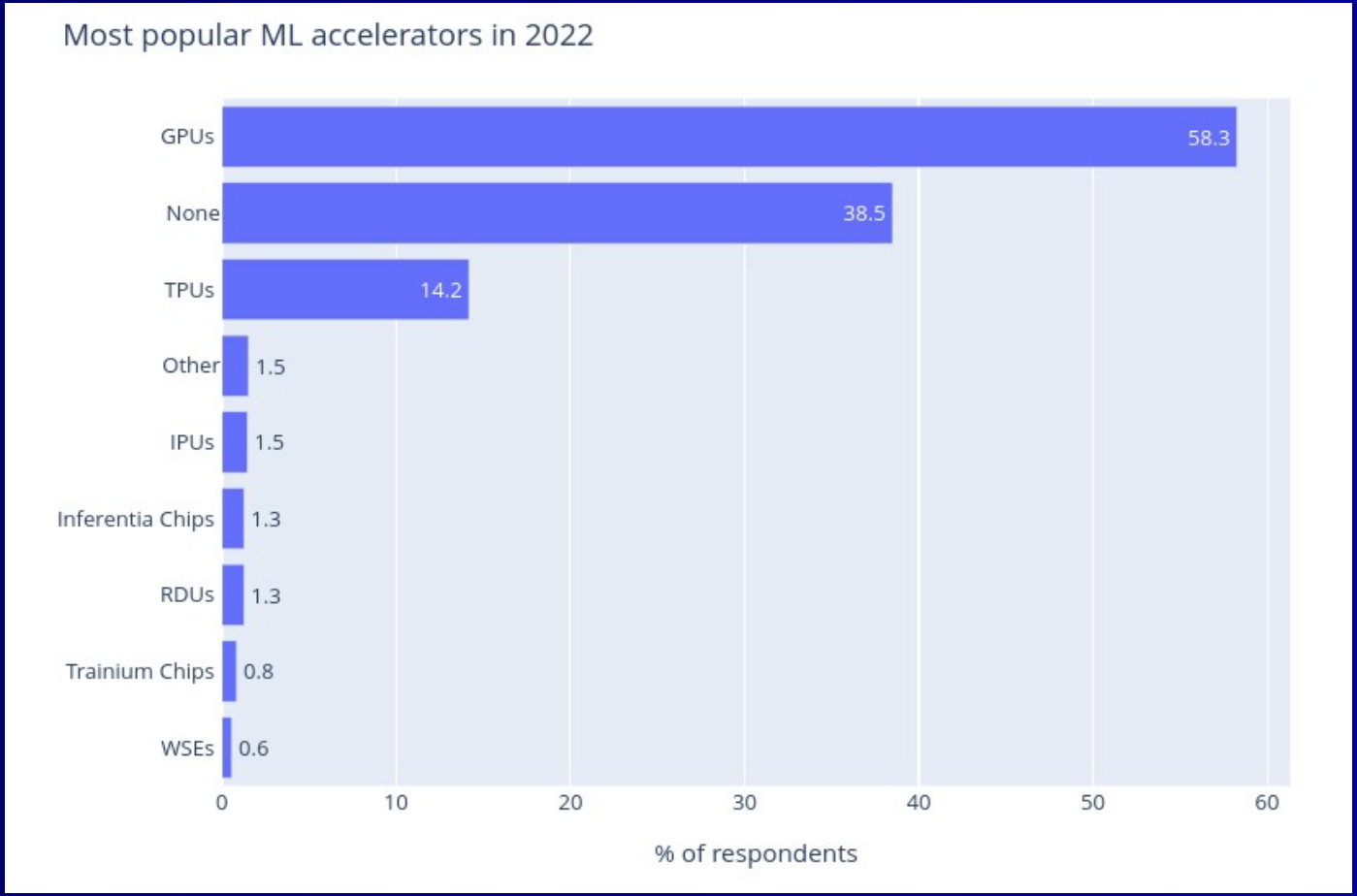
KAGGLE 2022 SURVEY



KAGGLE 2022 SURVEY



KAGGLE 2022 SURVEY



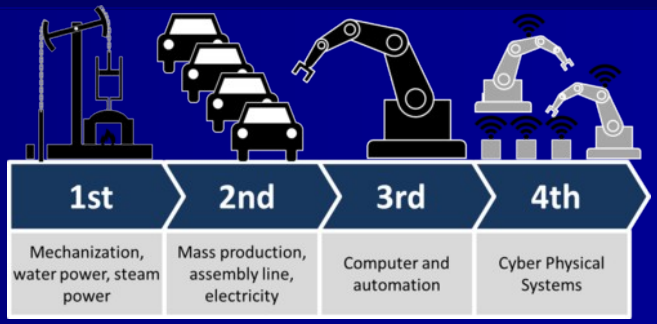
KAGGLE 2022 SURVEY

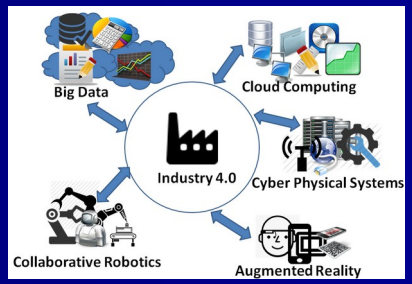
FULL SURVEY RESULTS IN [THIS LINK]

Kaggle Survey 2022: All Results

INDUSTRY 4.0

- 4th industrial revolution
- Role of new technologies
- "Smart factories"
- Role of big data and analytics
- 'Basque Industry 4.0'
- Machine Tool sector





DATA SCIENCE FOR A BETTER WORLD

- UN initiative
- Data: phones, meteo, networks...

- Humanitarian emergencies
- Pandemic diseases
- Sustainable development

- Quick answer
- Global answer

UNITED NATIONS GLOBAL PULSE

Harnessing big data for development and humanitarian action

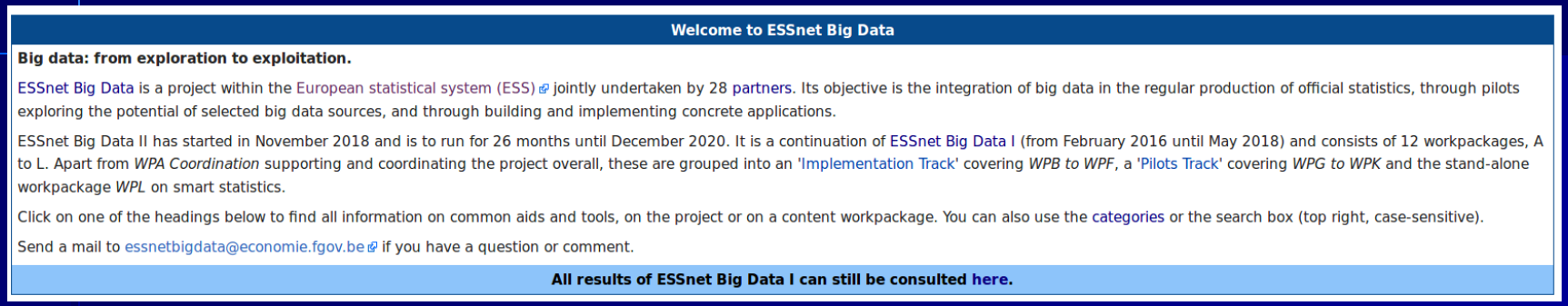
Global Pulse is a flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action.
The initiative was established based on a recognition that digital data offers the opportunity to gain a better understanding of changes in human well-being, and to get real-time feedback on how well policy responses are working.

To this end, Global Pulse is working to promote awareness of the opportunities Big Data presents for sustainable development and humanitarian action, forge public-private data sharing partnerships, generate high-impact analytical tools and approaches through its network of Pulse Labs, and drive broad adoption of useful innovations across the UN System.

DATA-POP ALLIANCE

Data-Pop Alliance is a global coalition on Big Data and development created by the Harvard Humanitarian Initiative, MIT Media Lab, and Overseas Development Institute that brings together researchers, experts, practitioners, and activists to promote a people-centered Big Data revolution through collaborative research, capacity building, and community engagement. As of February 2016, Flowminder Foundation has joined Data-Pop Alliance as its fourth Core Member.

EUROPEAN STATISTICS



Welcome to ESSnet Big Data

Big data: from exploration to exploitation.

ESSnet Big Data is a project within the European statistical system (ESS) jointly undertaken by 28 partners. Its objective is the integration of big data in the regular production of official statistics, through pilots exploring the potential of selected big data sources, and through building and implementing concrete applications.

ESSnet Big Data II has started in November 2018 and is to run for 26 months until December 2020. It is a continuation of ESSnet Big Data I (from February 2016 until May 2018) and consists of 12 workpackages, A to L. Apart from *WPA Coordination* supporting and coordinating the project overall, these are grouped into an 'Implementation Track' covering *WPB to WPF*, a 'Pilots Track' covering *WPG to WPK* and the stand-alone workpackage *WPL* on smart statistics.

Click on one of the headings below to find all information on common aids and tools, on the project or on a content workpackage. You can also use the categories or the search box (top right, case-sensitive).

Send a mail to essnetbigdata@economie.fgov.be if you have a question or comment.

All results of ESSnet Big Data I can still be consulted here.

Workpackages

Implementation Track

WPB Online job vacancies	WPC Enterprise characteristics	WPD Smart energy
WPE Tracking ships	WPF Process and architecture	

Pilots Track

WPG Financial transactions data	WPH Earth observation	WPI Mobile networks data
WPJ Innovative tourism statistics	WPK Methodology and quality	

Smart statistics

WPL Preparing smart statistics

- EuroStat
- National Institutes on Official Statistics

- Traditionally → surveys
- Compute alternative statistics from new big data sources

OPEN DATA GOVERNMENT

Data: Government, State, City, Local and Public

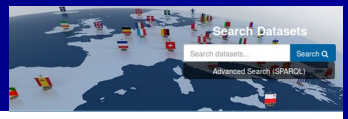
This is a directory of government, federal, state, city, local and other public datasets. See also Data APIs, Hubs, Marketplaces, Platforms, Portals, and Search Engines.

Portals | Global | USA | Canada | Europe | Asia | Australia, NZ and Pacific | Latin America | Africa | Middle East

Public data catalogs, portals, and services

- AWS (Amazon Web Services) Public Data Sets, provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications.
- Datacatalogs.org, open government data from US, EU, Canada, CKAN, and more.
- DataMarket, visualize the world's economy, societies, nature, and industries, with 100 million time series from UN, World Bank, Eurostat and other important data providers.
- datamob, Public data put to good use.
- Enigma, "Google for public data", provides easy access to government, NGO, and other public domain datasets.
- Freebase, a community-curated database of well-known people, places, and things.
- Google Public Data, with dynamic visualization and exploration tools.
- Knoema World Data Atlas, over 1000 indicators on all countries
- National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including

EUROPEAN DATA PORTAL



The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries. Information regarding the provision of data and the benefits of re-using data is also included.

OPEN DATA EUSKADI

Donostia data
Unean-unean

OUR WORLD IN DATA

The project, produced at the University of Oxford, is made available in its entirety as a **public good**. Visualizations are licensed under CC BY-SA and may be freely adapted for any purpose. Data is available for download in CSV format. Code we write is open-sourced under the MIT license and can be found on GitHub. Feel free to make use of anything you find here!

SPANISH "AI" STRATEGY

mercadofinanciero / economía finanzas

El Gobierno crea un Grupo de Sabios para elaborar un libro blanco sobre Inteligencia Artificial y Big Data

Publicado 14/11/2017 13:39:20 CET

MADRID, 14 Nov. (EUROPA PRESS) -

El Gobierno ha puesto en marcha un Grupo de Sabios formado por nueve expertos del mundo académico, empresarial e institucional que abordará las implicaciones sociales, jurídicas y éticas de la utilización de la Inteligencia Artificial (IA) y el Big Data con el objetivo de elaborar un Libro Blanco sobre la materia.

Las recomendaciones que incluya el documento elaborado por el equipo de expertos servirán para que el Ejecutivo impulse la elaboración de un código ético sobre el uso de los datos en las administraciones públicas, así como un código de buenas prácticas para las empresas en el uso de la IA y los datos.

EU versus USA
THE "AI" STRATEGIES

The Economist Topics Current edition More

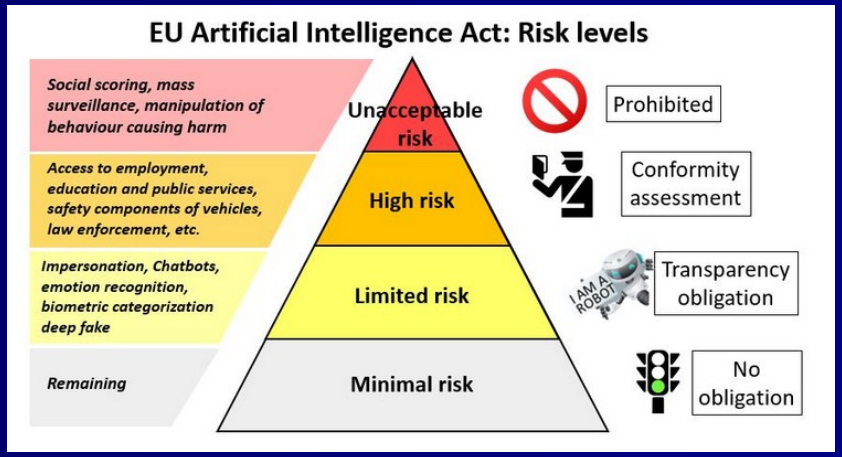
Big data, small politics

Can the EU become another AI superpower?

Taking on America and China will be hard

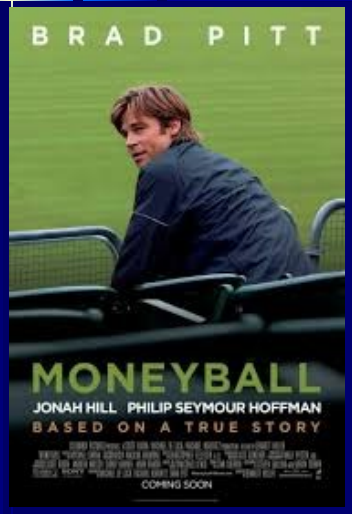
"The region also has a structural disadvantage: a lack of scale. Benefiting from huge, homogeneous home markets, America's and China's tech giants have a surfeit of the most vital resource for AI: data."

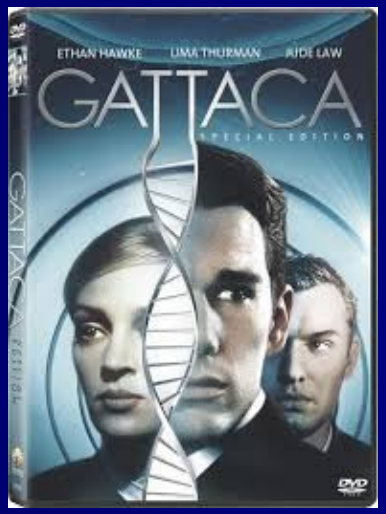
EU – AI LAW REGULATION



Resumen Ley Europea I [link]

FILMS AND "AI"


BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
COMING SOON


ETHAN HAWKE UMA THURMAN JUDE LAW
GATTACA
SPECIAL EDITION
DVD


DESCIFRAR EL CÓDIGO. GANAR LA GUERRA
★★★★★ INTENSO, IMPECABLE, ÚNICO. DE ÓSCAR. VARIETY
★★★★★ GANADORA Y CONMOVEDORA. UNA GRAN PELÍCULA. HOLLYWOOD REPORTER
★★★★★ UNA MAGISTRAL INTERPRETACIÓN DE BENEDICT CUMBERBATCH. NEW YORK POST
EXCEPCIONAL. ELLE
★★★★★ THE INDEPENDENT
★★★★★ GLAMOUR
THE IMITATION GAME
BENEDICT CUMBERBATCH
KEIRA KNIGHTLEY
(DESCIFRANDO ENIGMA)

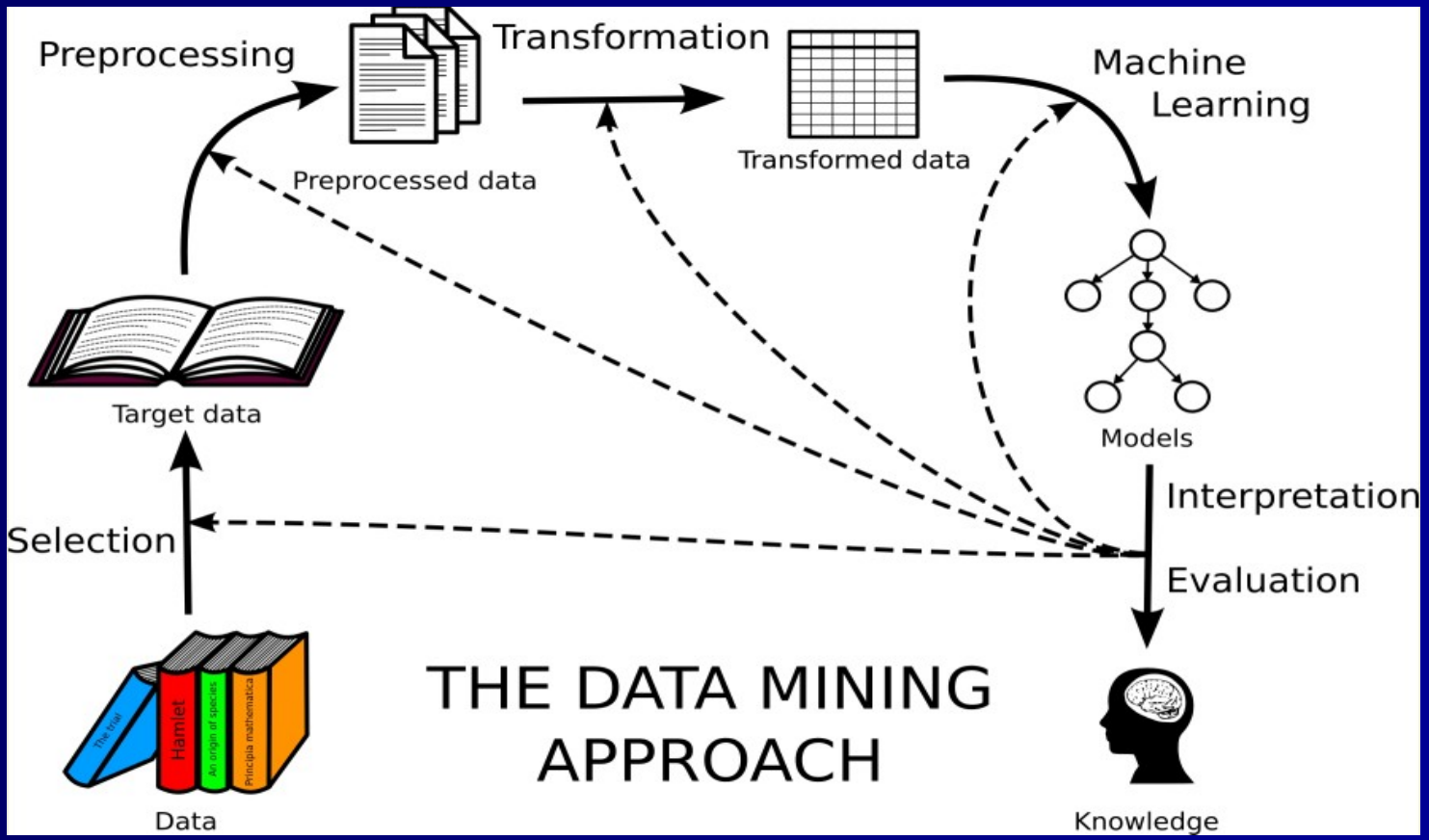

JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY
IN CINEMAS FEBRUARY 14TH


28 DAYS LATER
ES GIBT NICHTS MENSCHLICHERES
ALS DEN WILLEN ZU ÜBERLEBEN
EX_MACHINA
DEMNÄCHST IM KINO

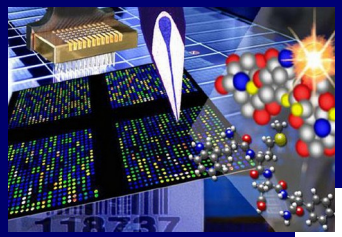

TOM CRUISE
MINORITY REPORT

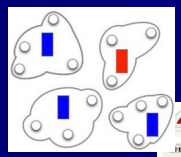
DATA MINING PROCESS: THE PIPELINE

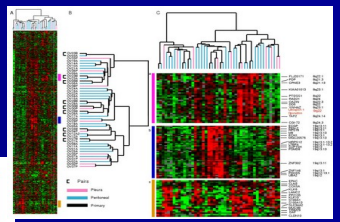
KDD PROCESS: KNOWLEDGE DISCOVERY IN DATABASES

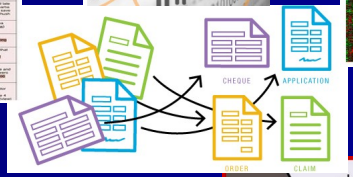


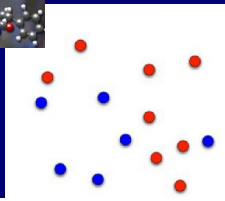
(a subset of) LEARNING SCENARIOS AND APPLICATIONS

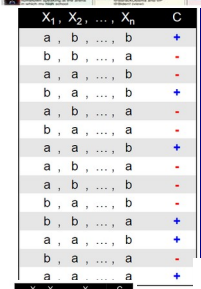


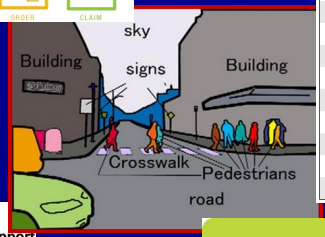


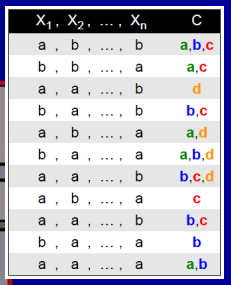


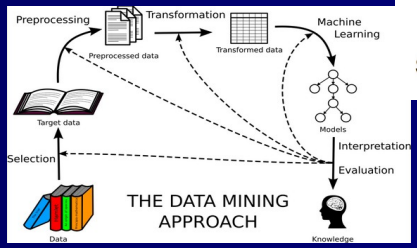


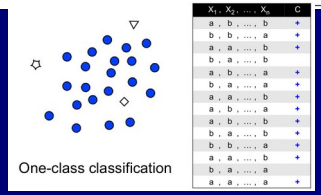


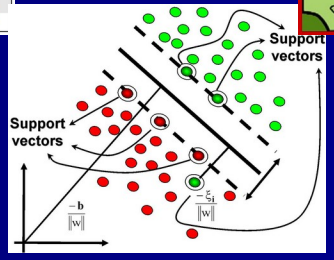








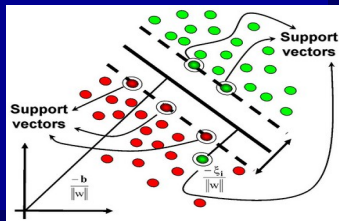




DATA MINING: MAIN TASKS

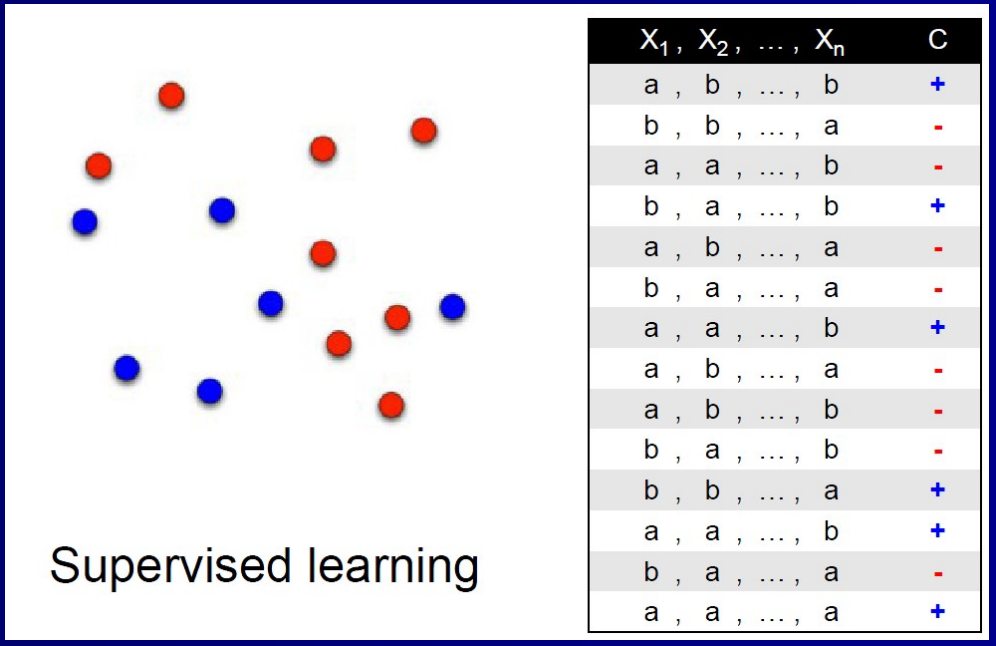
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
    - *Supervised classification: nominal variable to be predicted*
    - *Regression: quantitative variable to be predicted*

- Description Methods
  - Find human-interpretable patterns that describe the data
    - *Clustering – unsupervised classification*
    - *ANOVA – groups differences by variance analysis*
    - *Association rule discovery*
    - *Feature selection: discover the key predictors*
    - *Outlier detection*



SUPERVISED CLASSIFICATION

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-predictors*
  - Each record belongs to a *class, our variable of interest (variable to be predicted)*



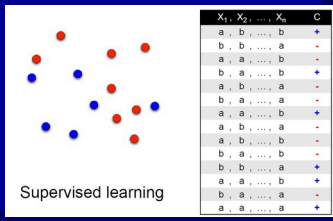
X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	-
a , b , … , b	-
b , a , … , b	-
b , b , … , a	+
a , a , … , b	+
b , a , … , a	-
a , a , … , a	+

SUPERVISED CLASSIFICATION

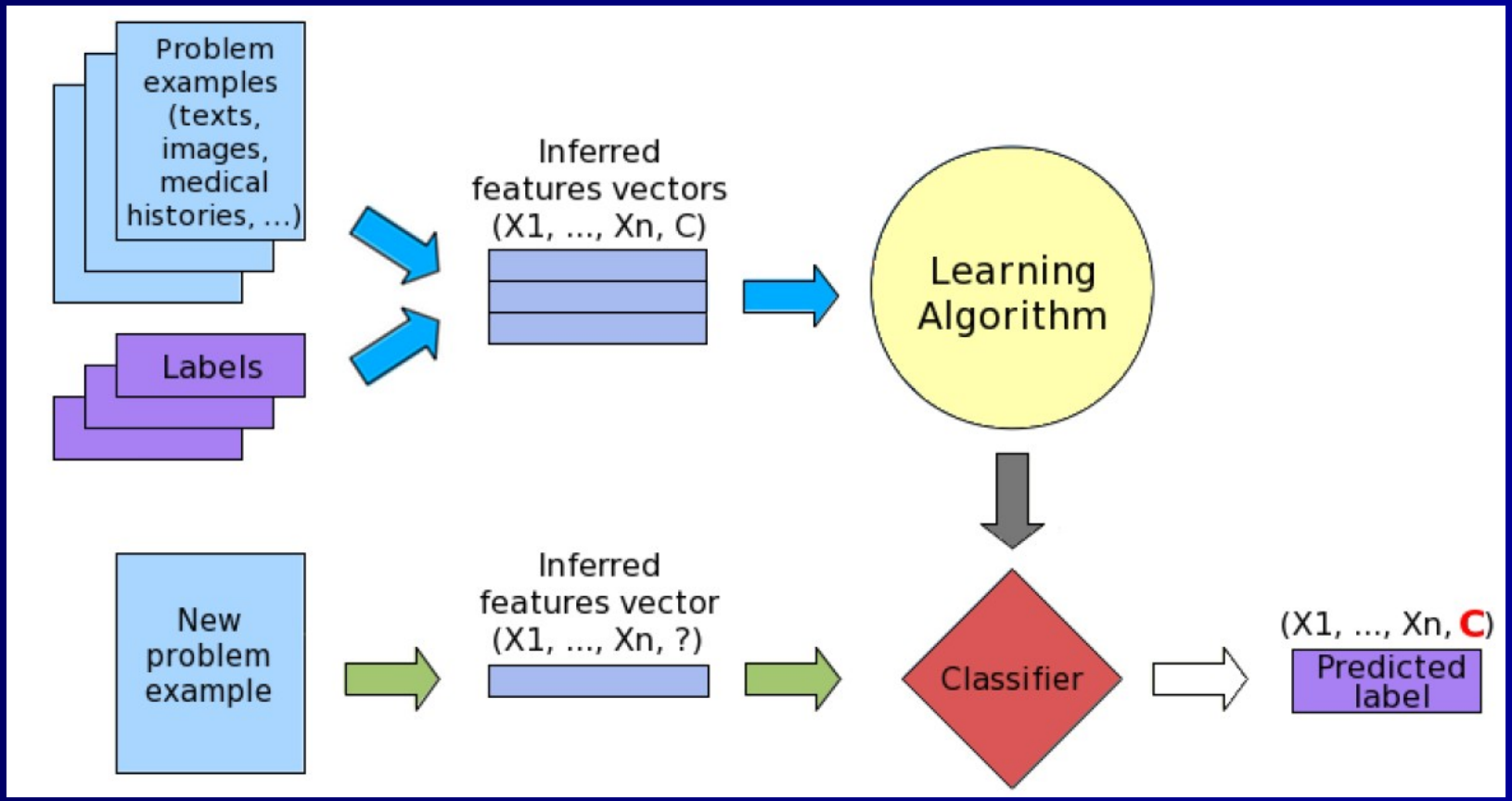
One of the most important tasks in data mining [116,121,240] is *supervised classification*, which seeks procedures for classifying objects in a set Ω into a set $\mathcal{C}$ of classes. Each object $u \in \Omega$ has associated a pair $(\mathbf{x}^u, y^u)$, where $\mathbf{x}^u$, the *predictor vector*, takes values on a set X, usually assumed to be a subset of $\mathbb{R}^p$, and $y^u \in \mathcal{C}$ is the class membership of u. Hereafter, we will simply use the term *variable* to refer to each component of the predictor vector.

Not all the information about the objects in Ω is available: the class membership c^u is only known for those objects u in some subset $I \subset \Omega$, called the *training sample*.

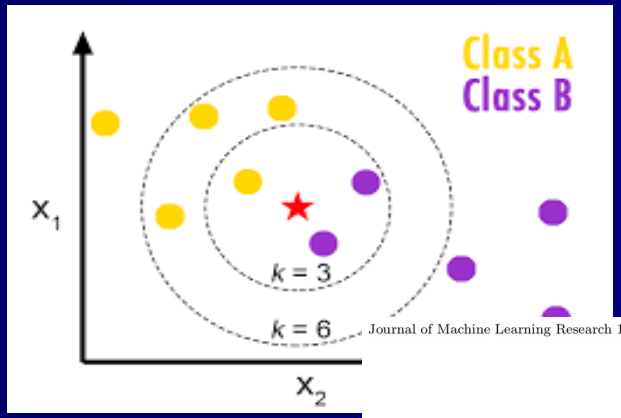
With this information, a classification rule is sought, i.e., a function $y : X \to \mathcal{C}$, which assigns label $y(\mathbf{x}) \in \mathcal{C}$ to predictor vector $\mathbf{x}$, $\forall \mathbf{x}$.

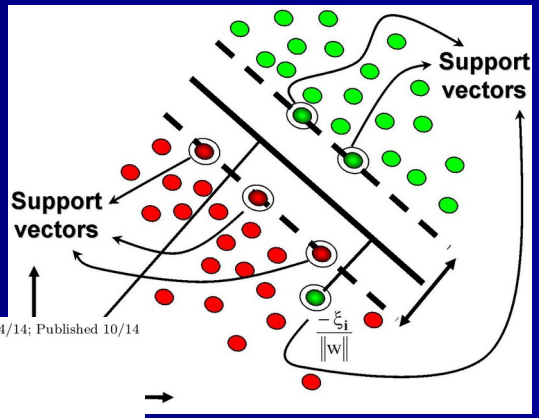


SUPERVISED CLASSIFICATION: the standard scenario



SUPERVISED CLASSIFICATION: models



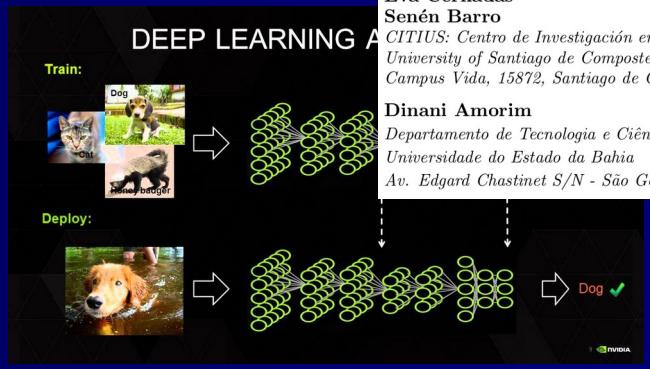


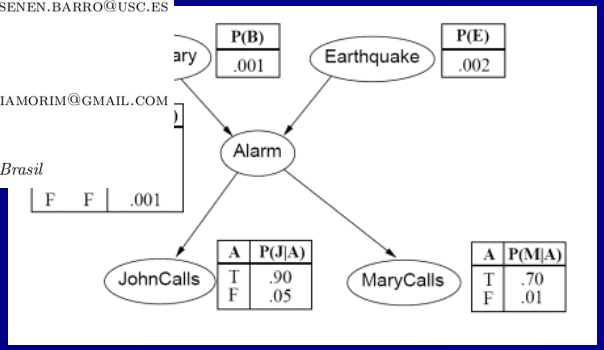
Journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

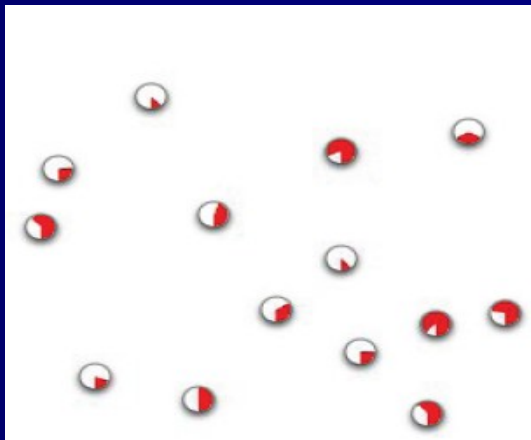
Dinani Amorim DINANIAMORIM@GMAIL.COM
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil





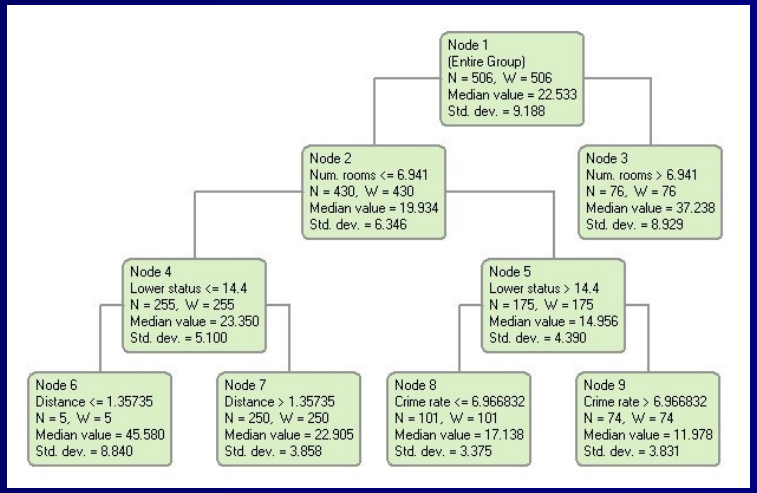
REGRESSION

- The *variable of interest* to be predicted is *quantitative*

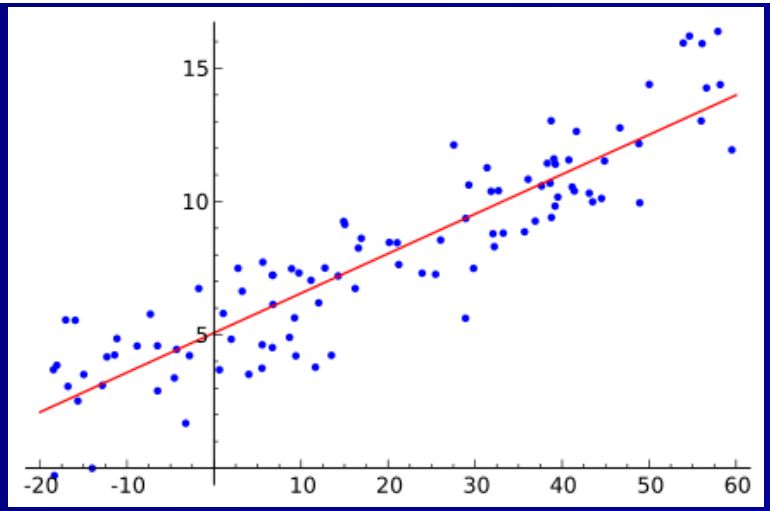


X_1 , X_2 , … , X_n	C
a , b , … , b	1.1
b , b , … , a	0.1
a , a , … , b	6.3
b , a , … , b	2.0
a , b , … , a	3.3
b , a , … , a	2.1
a , a , … , b	5.5
a , b , … , a	4.0
a , b , … , b	0.5
b , a , … , b	5.6
b , b , … , a	4.3
a , a , … , b	2.2
b , a , … , a	1.2
a , a , … , a	3.1

REGRESSION: models



$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$



REGRESSION

Consider the problem of approximating the set of data,

$$\mathcal{D} = \left\{ (x^1, y^1), \ldots, (x^l, y^l) \right\}, \quad x \in \mathbb{R}^n, y \in \mathbb{R}, \tag{5.1}$$

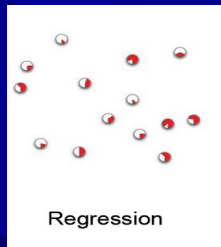
with a linear function,

$$f(x) = \langle w, x \rangle + b. \tag{5.2}$$

the optimal regression function is given by the minimum of the functional,

$$\Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C \sum_i (\xi_i^- + \xi_i^+), \tag{5.3}$$

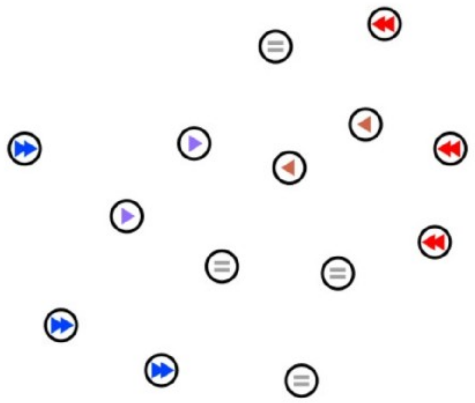
where C is a pre-specified value, and ξ^-, ξ^+ are slack variables representing upper and lower constraints on the outputs of the system.



X_1 , X_2 , ... , X_n	C
a , b , ... , b	1.1
b , b , ... , a	0.1
a , a , ... , b	6.3
b , a , ... , b	2.0
a , b , ... , a	3.3
b , a , ... , a	2.1
a , a , ... , b	5.5
a , b , ... , a	4.0
a , b , ... , b	0.5
b , a , ... , b	5.6
b , b , ... , a	4.3
a , a , ... , b	2.2
b , a , ... , a	1.2
a , a , ... , a	3.1

ORDINAL CLASSIFICATION

- The *variable of interest* to be predicted is *discrete, but ordered*

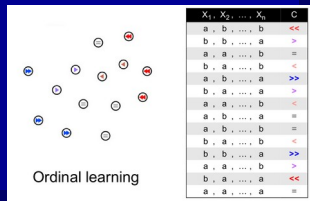


Ordinal learning

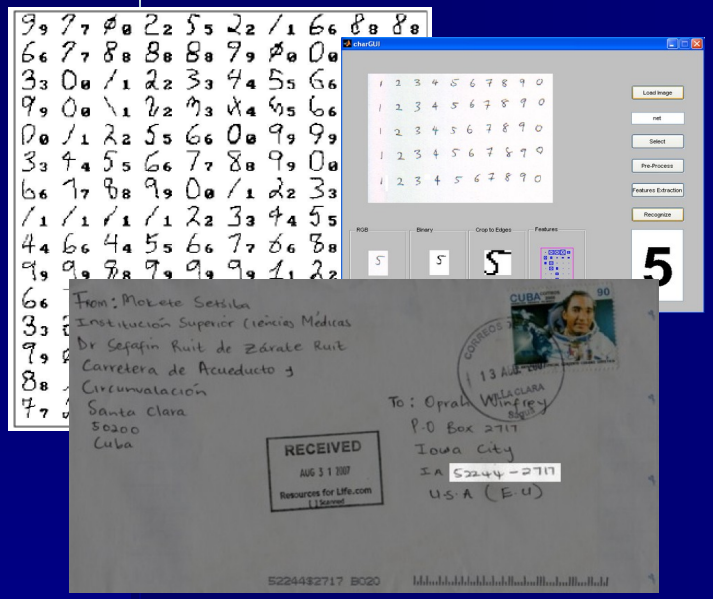
X_1, X_2, ... , X_n	C
a , b , ... , b	<<
b , b , ... , a	>
a , a , ... , b	=
b , a , ... , b	<
a , b , ... , a	>>
b , a , ... , a	>
a , a , ... , b	<
a , b , ... , a	=
a , b , ... , b	=
b , a , ... , b	<
b , b , ... , a	>>
a , a , ... , b	>
b , a , ... , a	<<
a , a , ... , a	=

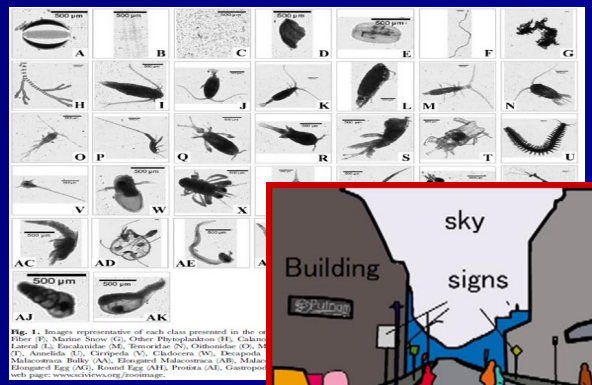
ORDINAL CLASSIFICATION

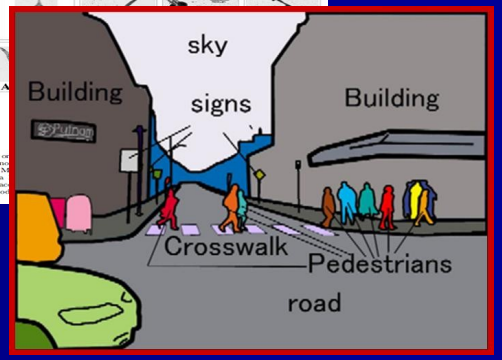
The ordinal regression problem consists on predicting the label y of an input vector $\mathbf{x}$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$ and $y \in \mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_Q\}$, i.e. $\mathbf{x}$ is in a K-dimensional input space and y is in a label space of Q different labels. These labels form categories or groups of patterns, and the objective is to find a classification rule or function $r : \mathcal{X} \to \mathcal{Y}$ to predict the categories of new patterns, given a training set of N points, $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$. A natural label ordering is included for ordinal regression, $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_Q$, where $\prec$ is an order relation. Many ordinal regression



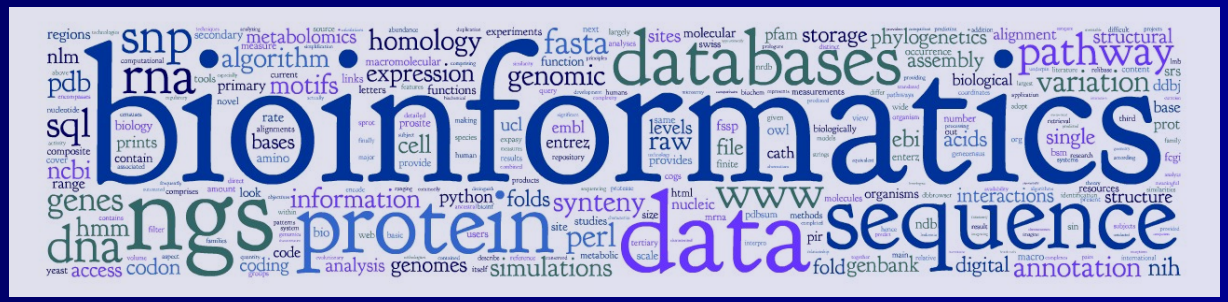
SUPERVISED CLASSIFICATION and REGRESSION:
APPLICATIONS
PATTERN RECOGNITION

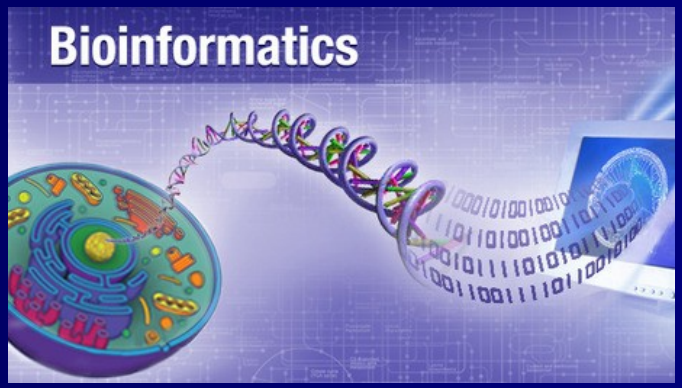


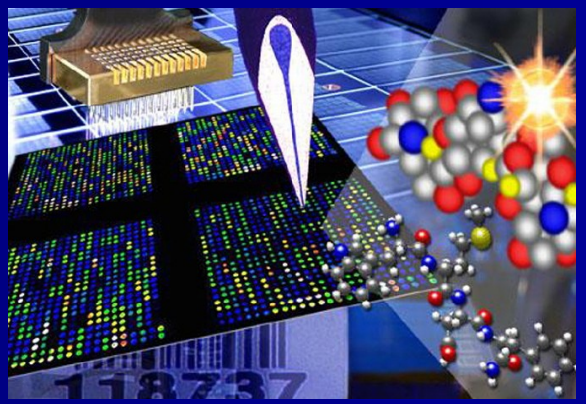




BIOINFORMATICS
DIAGNOSIS AND PROGNOSIS OF DISEASES
BIOMARKER DISCOVERY

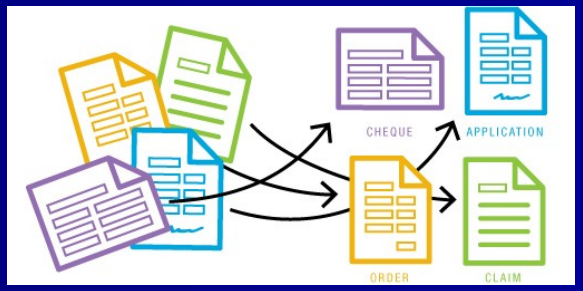




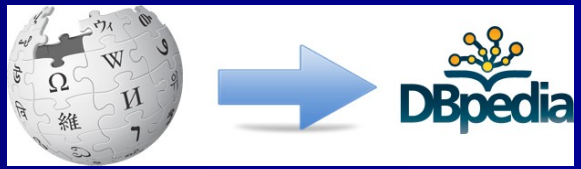


DOCUMENT CLASSIFICATION

- "Natural Language Processing" (NLP)



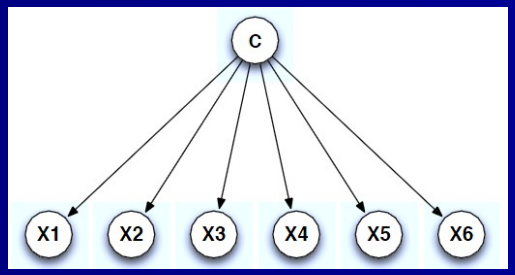
- Topic - category
- Level of difficulty
- Author's genre

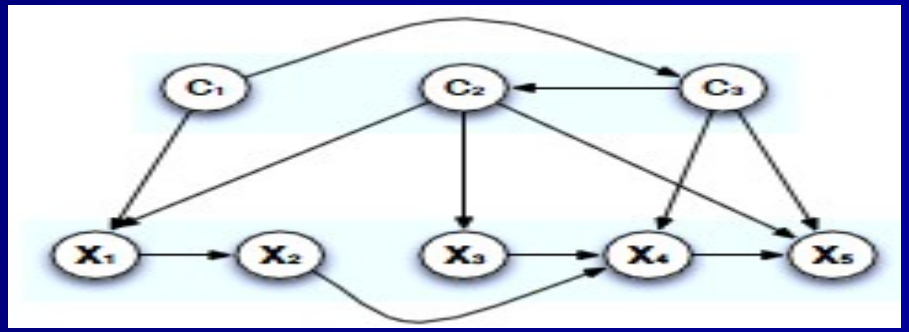


BEYOND SINGLE CLASS VARIABLE…
MULTIDIMENSIONAL CLASSIFICATION

- Several class variables to be jointly predicted
- Learn relationships between class variables
- New term: Joint accuracy

X_1	X_2	…	X_n	C_1	C_2	…	C_m
$x_1^{(1)}$	$x_2^{(1)}$	…	$x_n^{(1)}$	$c_1^{(1)}$	$c_2^{(1)}$	…	$c_m^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	…	$x_n^{(2)}$	$c_1^{(2)}$	$c_2^{(2)}$	…	$c_m^{(2)}$
…	…	…	…	…	…	…	…
$x_1^{(N)}$	$x_2^{(N)}$	…	$x_n^{(N)}$	$c_1^{(N)}$	$c_2^{(N)}$	…	$c_m^{(N)}$





BEYOND SINGLE CLASS VARIABLE...
MULTIDIMENSIONAL CLASSIFICATION

In this paper we are interested in classification problems where there are multiple class variables $C_1, \ldots, C_d$. Therefore the *multi-dimensional classification* problem consists of finding a function h that assigns to each instance given by a vector of m features $\mathbf{x} = (x_1, \ldots, x_m)$ a vector of d class values $\mathbf{c} = (c_1, \ldots, c_d)$:

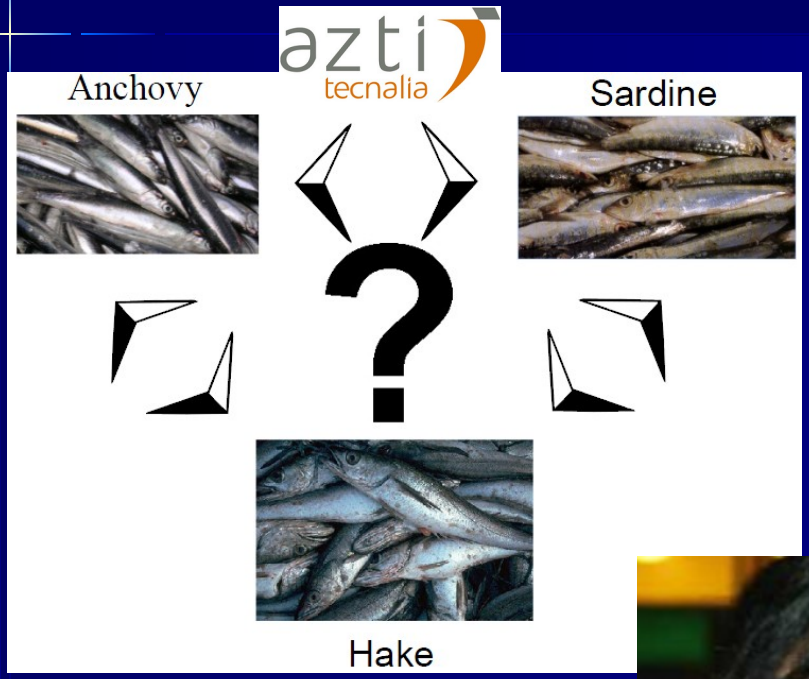
$$h : \Omega_{X_1} \times \cdots \times \Omega_{X_m} \rightarrow \Omega_{C_1} \times \cdots \times \Omega_{C_d}$$

$$(x_1, \ldots, x_m) \mapsto (c_1, \ldots, c_d)$$

We assume that C_i is a discrete variable, for all $i = 1, \ldots, d$, with Ω_{C_i} denoting its sample space and $\mathcal{I} = \Omega_{C_1} \times \cdots \times \Omega_{C_d}$, the space of joint configurations of the class variables. Analogously, Ω_{X_j} is the sample space of the discrete feature variable X_j, for all $j = 1, \ldots, m$.

X_1	X_2	...	X_n	C_1	C_2	...	C_m
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_n^{(1)}$	$c_1^{(1)}$	$c_2^{(1)}$	...	$c_m^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_n^{(2)}$	$c_1^{(2)}$	$c_2^{(2)}$	...	$c_m^{(2)}$
...	...	...	...	...	...	...	...
$x_1^{(N)}$	$x_2^{(N)}$	...	$x_n^{(N)}$	$c_1^{(N)}$	$c_2^{(N)}$	...	$c_m^{(N)}$

MULTIDIMENSIONAL CLASSIFICATION APPLICATIONS



Contents lists available at SciVerse ScienceDirect

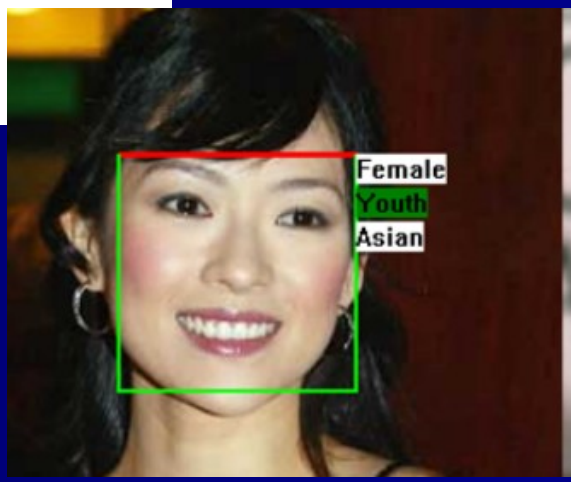
Environmental Modelling & Software

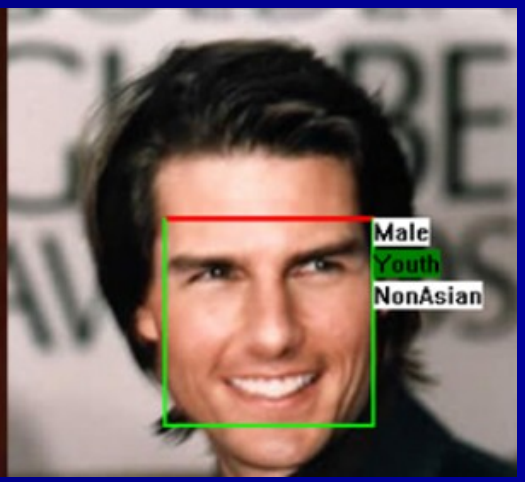
journal homepage: www.elsevier.com/locate/envsoft

Supervised pre-processing approaches in multiple class variables classification for fish recruitment forecasting

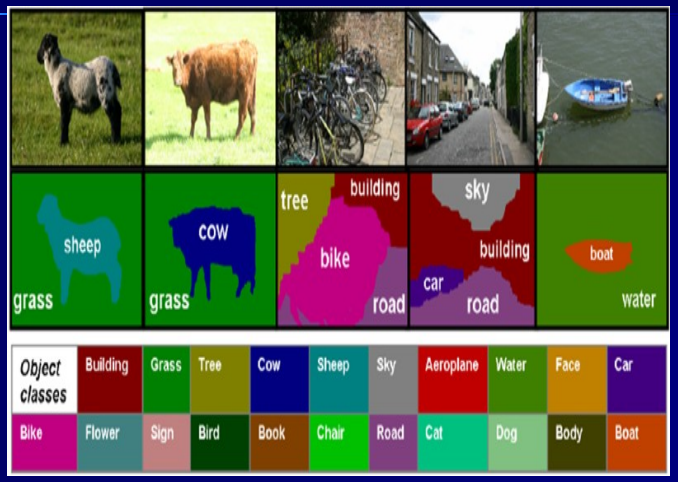
Jose A. Fernandes [a,b,*], Jose A. Lozano [b], Iñaki Inza [b], Xabier Irigoien [a,c], Aritz Pérez [b], Juan D. Rodríguez [b]

[a] AZTI-Tecnalia, Marine Research Division, Herrera Kaia z/g, E-20110 Pasaia (Gipuzkoa), Spain
[b] University of the Basque Country, Department of Computer Science and AI, Intelligent Systems Group (ISG), Paseo Manuel de Lardizabal, 1. E-20018 Donostia – San Sebastián, Spain
[c] King Abdullah University of Science and Technology (KAUST), Chemical and Life Sciences and Engineering, Red Sea Research Center, Thuwal 23955-6900, Saudi Arabia





MULTILABEL CLASSIFICATION

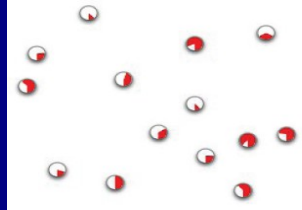


N.	Film	Year	Genre
1	Cadena perpetua	1994	Crime, Drama
2	El padrino	1972	Crime, Drama
3	El padrino. Parte II	1974	Crime, Drama
4	El bueno, el feo y el malo	1966	Adventure, Western
5	Pulp Fiction	1994	Crime, Thriller
6	12 hombres sin piedad	1957	Drama
7	La lista de Schindler	1993	Biography, Drama, History, War
8	El caballero oscuro	2008	Action, Crime, Drama, Thriller
9	El señor de los anillos: El ret...	2003	Action, Adventure, Drama, Fantasy
10	El club de la lucha	1999	Drama

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

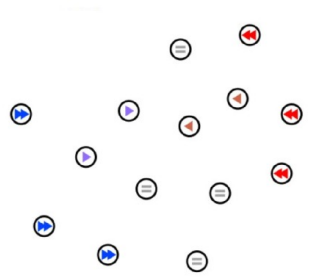
FULL SUPERVISION

X_1	X_2	...	X_n	C_1	C_2	...	C_m
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_n^{(1)}$	$c_1^{(1)}$	$c_2^{(1)}$	...	$c_m^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_n^{(2)}$	$c_1^{(2)}$	$c_2^{(2)}$	...	$c_m^{(2)}$
...	...	...	...	...	...	...	...
$x_1^{(N)}$	$x_2^{(N)}$	...	$x_n^{(N)}$	$c_1^{(N)}$	$c_2^{(N)}$	...	$c_m^{(N)}$



Regression

X_1 , X_2 , ... , X_n	C
a , b , ... , b	1.1
b , b , ... , a	0.1
a , a , ... , b	6.3
b , a , ... , b	2.0
a , b , ... , a	3.3
b , a , ... , a	2.1
a , a , ... , b	5.5
a , b , ... , a	4.0
a , b , ... , b	0.5
b , a , ... , b	5.6
b , b , ... , a	4.3
a , a , ... , b	2.2
b , a , ... , a	1.2
a , a , ... , a	3.1



Ordinal learning

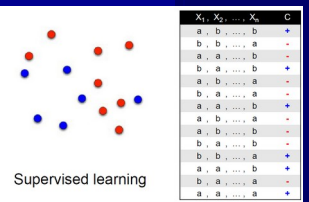
X_1 , X_2 , ... , X_n	C
a , b , ... , b	<<
b , b , ... , a	>
a , a , ... , b	=
b , a , ... , b	<
a , b , ... , a	>>
b , a , ... , a	>
a , a , ... , b	<
a , b , ... , a	=
a , b , ... , b	=
b , a , ... , b	<
b , b , ... , a	>>
a , a , ... , b	>
b , a , ... , a	<<
a , a , ... , a	=

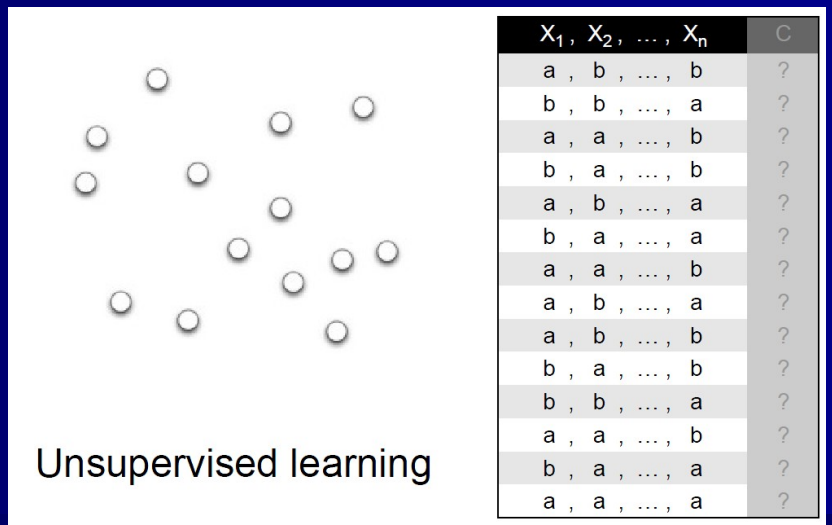
X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

UNSUPERVISED CLASSIFICATION – CLUSTERING –

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-~~predictors~~*
  - *No "target feature" (class)* which supervises the learning process

- Find groups of cases with:
  - Large intra-group homogeneity: clustering similar samples
  - Large inter-groups heterogeneity

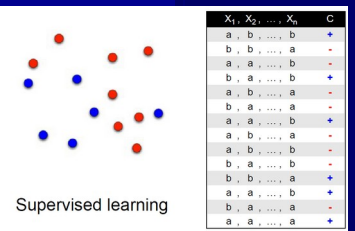




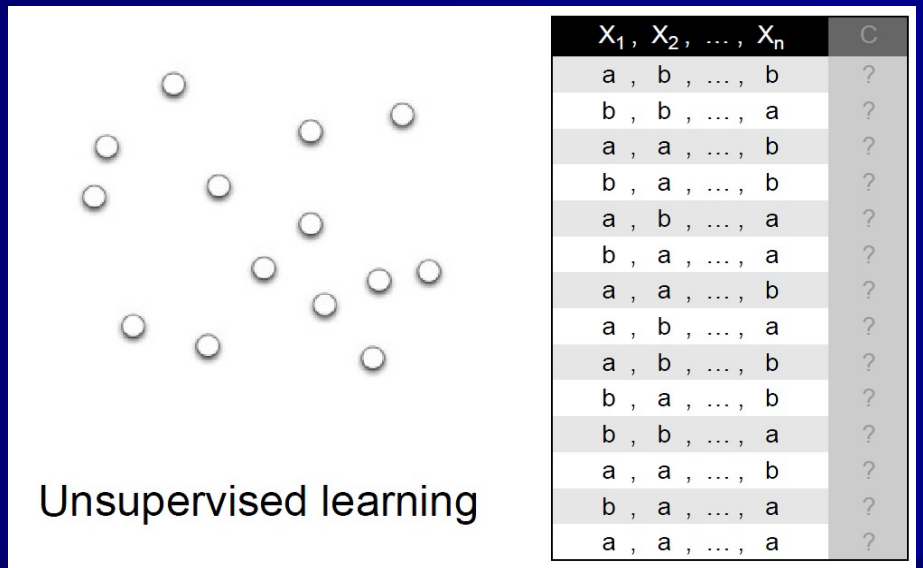
X_1 , X_2 , … , X_n	C
a , b , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

UNSUPERVISED CLASSIFICATION – CLUSTERING –

- Difficult evaluation-measure of these properties --› no recognition rate
- Number of groups... deciding before-hand, difficult decision
- "Distance"-"similarity" function → numerical features? nominal predictors?



X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , a , ... , b	+
a , b , ... , b	-
b , a , ... , b	-
b , b , ... , a	+
a , a , ... , b	+
b , a , ... , a	-
a , a , ... , a	+



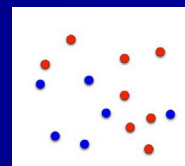
X_1 , X_2 , ... , X_n	C
a , b , ... , b	?
b , b , ... , a	?
a , a , ... , b	?
b , a , ... , b	?
a , b , ... , a	?
b , a , ... , a	?
a , a , ... , b	?
a , b , ... , a	?
a , b , ... , b	?
b , a , ... , b	?
b , b , ... , a	?
a , a , ... , b	?
b , a , ... , a	?
a , a , ... , a	?

UNSUPERVISED CLASSIFICATION – CLUSTERING –

Suppose we are given a data set $X = \{x_1, \ldots, x_N\}, x_n \in R^d$. The M-clustering problem aims at partitioning this data set into M disjoint subsets (clusters) $C_1, \ldots, C_M$, such that a clustering criterion is optimized. The most widely used clustering criterion is the sum of the squared Euclidean distances between each data point x_i and the centroid m_k (cluster center) of the subset C_k which contains x_i. This criterion is called clustering error and depends on the cluster centers $m_1, \ldots, m_M$:

$$E(m_1, \ldots, m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k) \|x_i - m_k\|^2, \qquad (1)$$

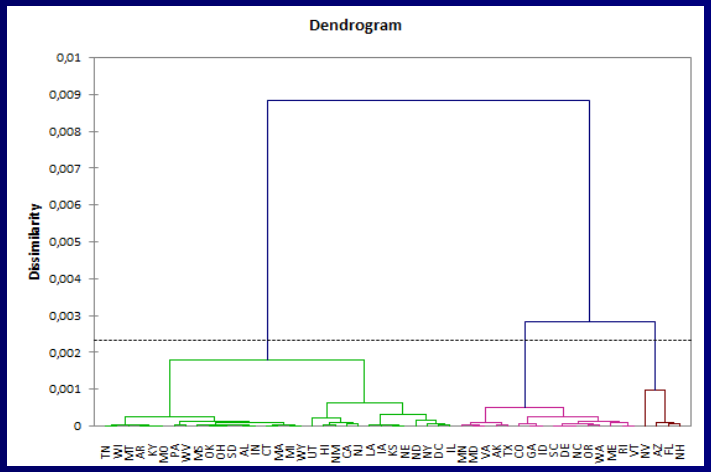
where $I(X) = 1$ if X is true and 0 otherwise.

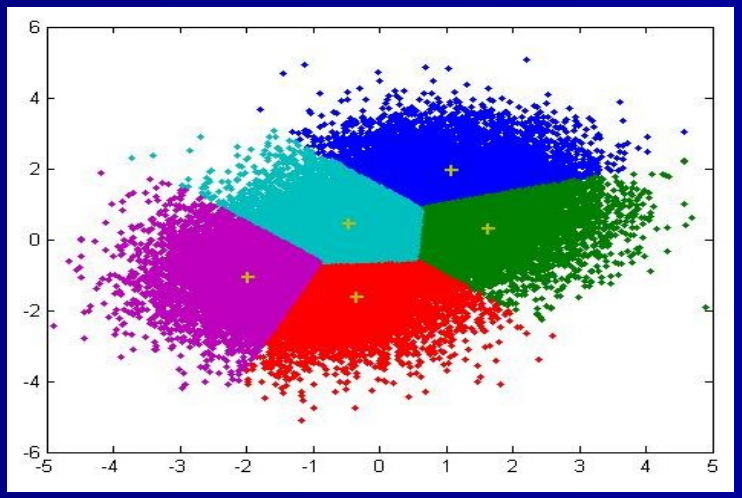


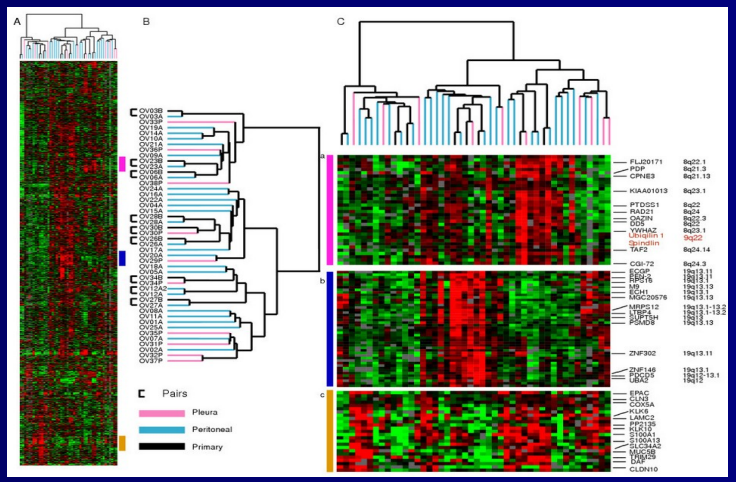
Supervised learning

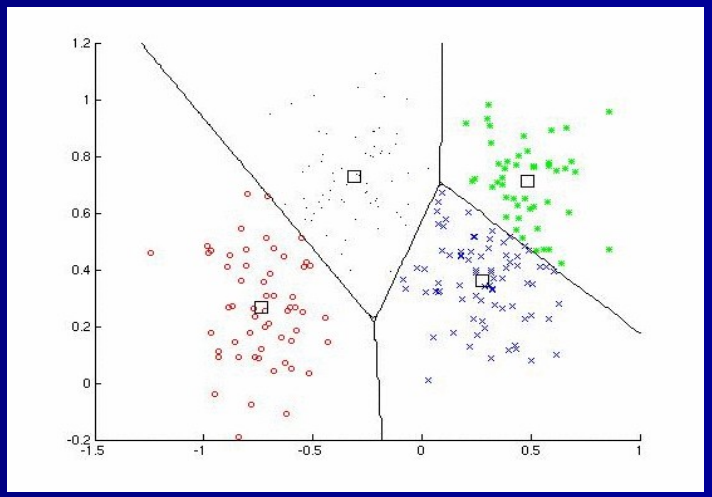
X_1	X_2	...	X_n	C
a	b	...	b	+
b	b	...	a	-
a	a	...	b	-
b	a	...	b	+
a	b	...	a	-
b	a	...	a	-
a	a	...	b	+
a	b	...	a	-
a	b	...	b	-
b	a	...	b	-
b	b	...	a	+
a	a	...	b	+
b	a	...	a	-
a	a	...	a	+

CLUSTERING: MODELS







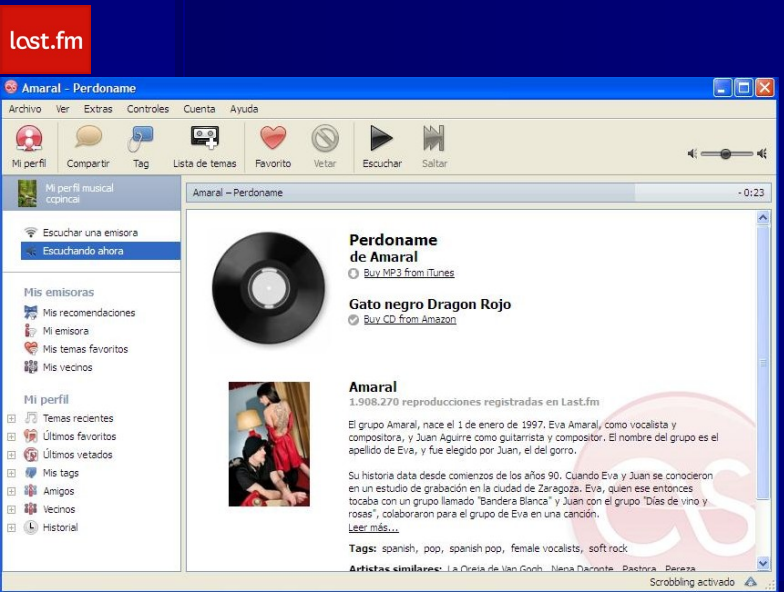


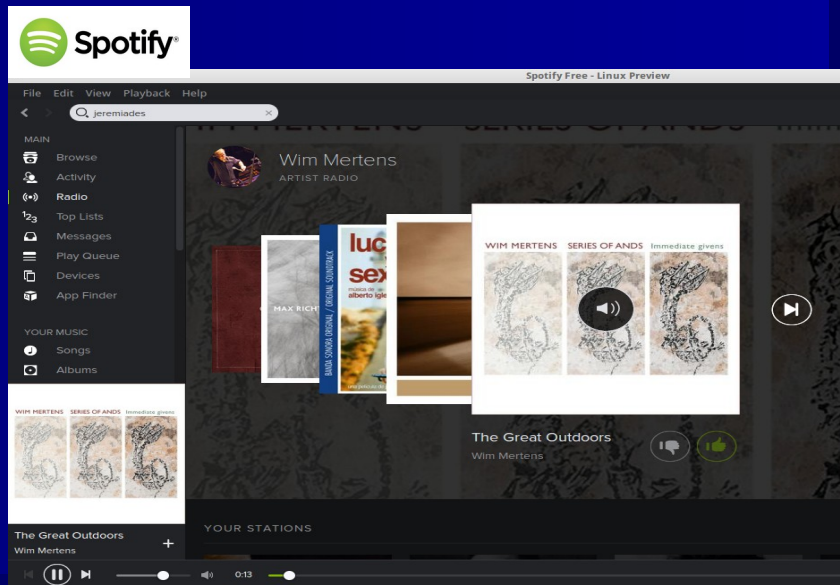
CLUSTERING: APPLICATIONS
CUSTOMER SEGMENTATION

- Identify micro-markets and develop policies for each
- Targeted marketing
- Similar customers are grouped in the same cluster

COLLABORATIVE FILTERING RECOMMENDER SYSTEMS

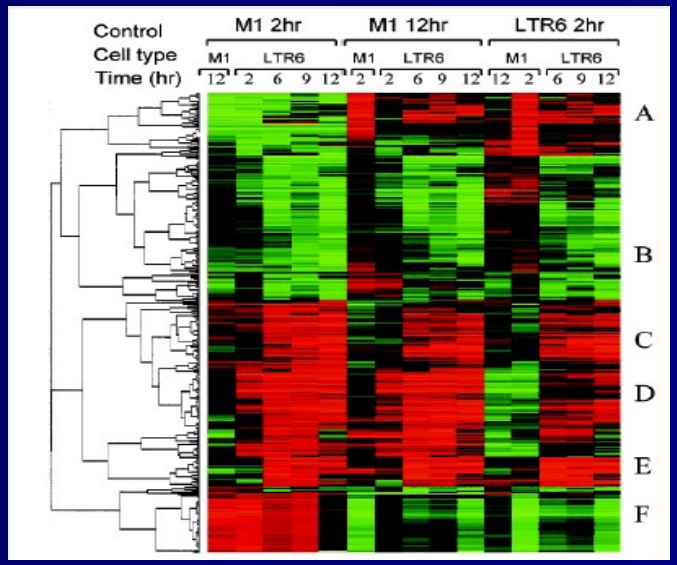
Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	6
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?

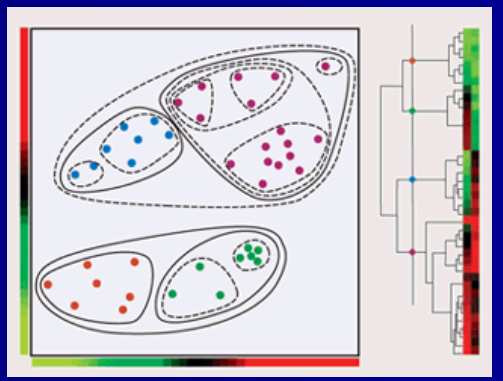




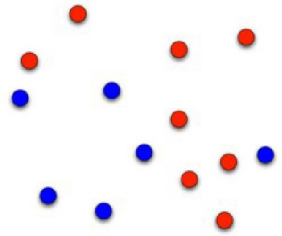
DNA MICROARRAY CLUSTERING

- Find genes with similar expression profiles ~ a way to infer the function of genes whose function is unknown
- Biclustering... a classic concept in fashion again:
  - Finding a subgroup of samples with a similar pattern in a subgroup of variables (not in all the variables)



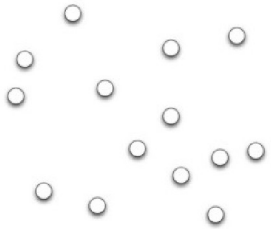


IS THERE SOMEONE IN THE MIDDLE?



Supervised learning

X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	-
a , b , … , b	-
b , a , … , b	-
b , b , … , a	+
a , a , … , b	+
b , a , … , a	-
a , a , … , a	+



Unsupervised learning

X_1 , X_2 , … , X_n	C
a , b , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

Pattern Recognition Letters 69 (2016) 49–55

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Weak supervision and other non-standard classification problems: A taxonomy☆

Jerónimo Hernández-González*, Iñaki Inza, Jose A. Lozano

Intelligent Systems Group, University of the Basque Country UPV/EHU, P. Manuel Lardizabal 1, 20018 Donostia, Spain

ARTICLE INFO

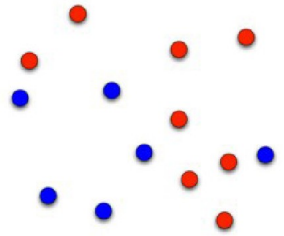
Article history:
Received 10 May 2015
Available online 24 October 2015

Keywords:
Weakly supervised classification
Partially supervised classification
Degrees of supervision

ABSTRACT

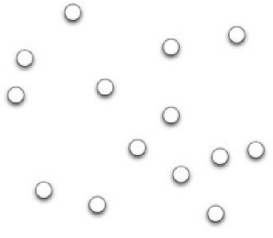
In recent years, different researchers in the machine learning community have presented new classification frameworks which go beyond the standard supervised classification in different aspects. Specifically, a wide spectrum of novel frameworks that use partially labeled data in the construction of classifiers has been studied. With the objective of drawing up a description of the state-of-the-art, three identifying characteristics of these novel frameworks have been considered: (1) the relationship between instances and labels of a problem, which may be beyond the *one-instance one-label* standard, (2) the possible provision of partial class information for the training examples, and (3) the possible provision of partial class information also for the examples in the prediction stage. These three ideas have been formulated as axes of a comprehensive taxonomy that organizes the state-of-the-art. The proposed organization allows us both to understand similarities/differences among the different classification problems already presented in the literature as well as to discover unexplored frameworks that might be seen as further challenges and research opportunities. A representative set of state-of-the-art problems has been used to illustrate the novel taxonomy and support the discussion.

IS THERE SOMEONE IN THE MIDDLE?



Supervised learning

X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	-
a , b , … , b	-
b , a , … , b	-
b , b , … , a	+
a , a , … , b	+
b , a , … , a	-

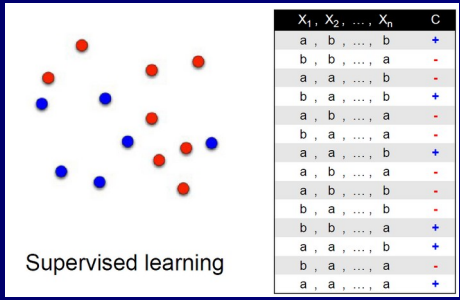


Unsupervised learning

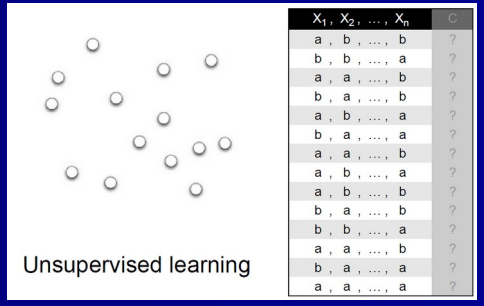
X_1 , X_2 , … , X_n	C
a , b , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?

- Hidden big data. Large quantities of useful data are in fact useless because they are untagged, file-based, and unstructured. The 2012 IDC study on big data [117] explained that, in 2012, 23% (643 exabytes) of the digital universe would be useful for big data if tagged and analyzed. However, at that time only 3% of the potentially useful data was tagged, and even less was analyzed. The figures have probably gotten worse in recent years. The Open Data and Semantic Web movements have emerged, in part, to make us aware and improve on this situation. No comments

IS THERE SOMEONE IN THE MIDDLE?



X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	-
a , b , … , b	-
b , a , … , b	-
b , b , … , a	+
a , a , … , b	+
b , a , … , a	-
a , a , … , a	+



X_1 , X_2 , … , X_n	C
a , b , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

Pattern Recognition Letters 69 (2016) 49–55

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Weak supervision and other non-standard classification problems: A taxonomy☆

Jerónimo Hernández-González*, Iñaki Inza, Jose A. Lozano

Intelligent Systems Group, University of the Basque Country UPV/EHU, P. Manuel Lardizabal 1, 20018 Donostia, Spain

ARTICLE INFO

Article history:
Received 10 May 2015
Available online 24 October 2015

Keywords:
Weakly supervised classification
Partially supervised classification
Degrees of supervision

ABSTRACT

In recent years, different researchers in the machine learning community have presented new classification frameworks which go beyond the standard supervised classification in different aspects. Specifically, a wide spectrum of novel frameworks that use partially labeled data in the construction of classifiers has been studied. With the objective of drawing up a description of the state-of-the-art, three identifying characteristics of these novel frameworks have been considered: (1) the relationship between instances and labels of a problem, which may be beyond the *one-instance one-label* standard, (2) the possible provision of partial class information for the training examples, and (3) the possible provision of partial class information also for the examples in the prediction stage. These three ideas have been formulated as axes of a comprehensive taxonomy that organizes the state-of-the-art. The proposed organization allows us both to understand similarities/differences among the different classification problems already presented in the literature as well as to discover unexplored frameworks that might be seen as further challenges and research opportunities. A representative set of state-of-the-art problems has been used to illustrate the novel taxonomy and support the discussion.

International Journal of Approximate Reasoning 150 (2022) 258–272

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

On the relative value of weak information of supervision for learning generative models: An empirical study

Jerónimo Hernández-González [a,*], Aritz Pérez [b]

[a] *Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), Gran Via de les Corts Catalanes 585, Barcelona, Spain*
[b] *Basque Center for Applied Mathematics, Al. Mazarredo 14, Bilbao, Spain*

ARTICLE INFO

Article history:
Received 2 July 2022
Received in revised form 8 August 2022
Accepted 22 August 2022
Available online 31 August 2022

Keywords:
Weak supervision
Model learning
Generative models
Empirical study

ABSTRACT

Weakly supervised learning is aimed to learn predictive models from partially supervised data, an easy-to-collect alternative to the costly standard full supervision. During the last decade, the research community has striven to show that learning reliable models in specific weakly supervised problems is possible. We present an empirical study that analyzes the value of weak information of supervision throughout its entire spectrum, from none to full supervision. Its contribution is assessed under the realistic assumption that a small subset of fully supervised data is available. Particularized in the problem of learning with candidate sets, we adapt Cozman and Cohen [1] key study to learning from weakly supervised data. Standard learning techniques are used to infer generative models from this type of supervision with both synthetic and real data. Empirical results suggest that weakly labeled data is helpful in realistic scenarios, where fully labeled data is scarce, and its contribution is directly related to both the amount of information of supervision and how meaningful this information is.

THE TERM
- "WEAKLY SUPERVISED LEARNING" -
- RESEARCH OPORTUNITIES -

- GoogleScholar – number of "search results":
  - Since 2015 → 24,600
  - Since 2018 → 21,200
  - Since 2020 → 18,900

A Brief Introduction to Weakly Supervised Learning

Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
zhouzh@nju.edu.cn

WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation

Thibaut Durand[(1)⋆], Taylor Mordan[(1,2)⋆], Nicolas Thome[(3)], Matthieu Cord[(1)]
(1) Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu, 75005 Paris
(2) Thales Optronique S.A.S., 2 Avenue Gay Lussac, 78990 Élancourt, France
(3) CEDRIC - Conservatoire National des Arts et Métiers, 292 rue St Martin, 75003 Paris, France
{thibaut.durand, taylor.mordan, nicolas.thome, matthieu.cord}@lip6.fr

Jain *et al. BMC Bioinformatics* 2015, **17**(Suppl 1):1
DOI 10.1186/s12859-015-0844-1

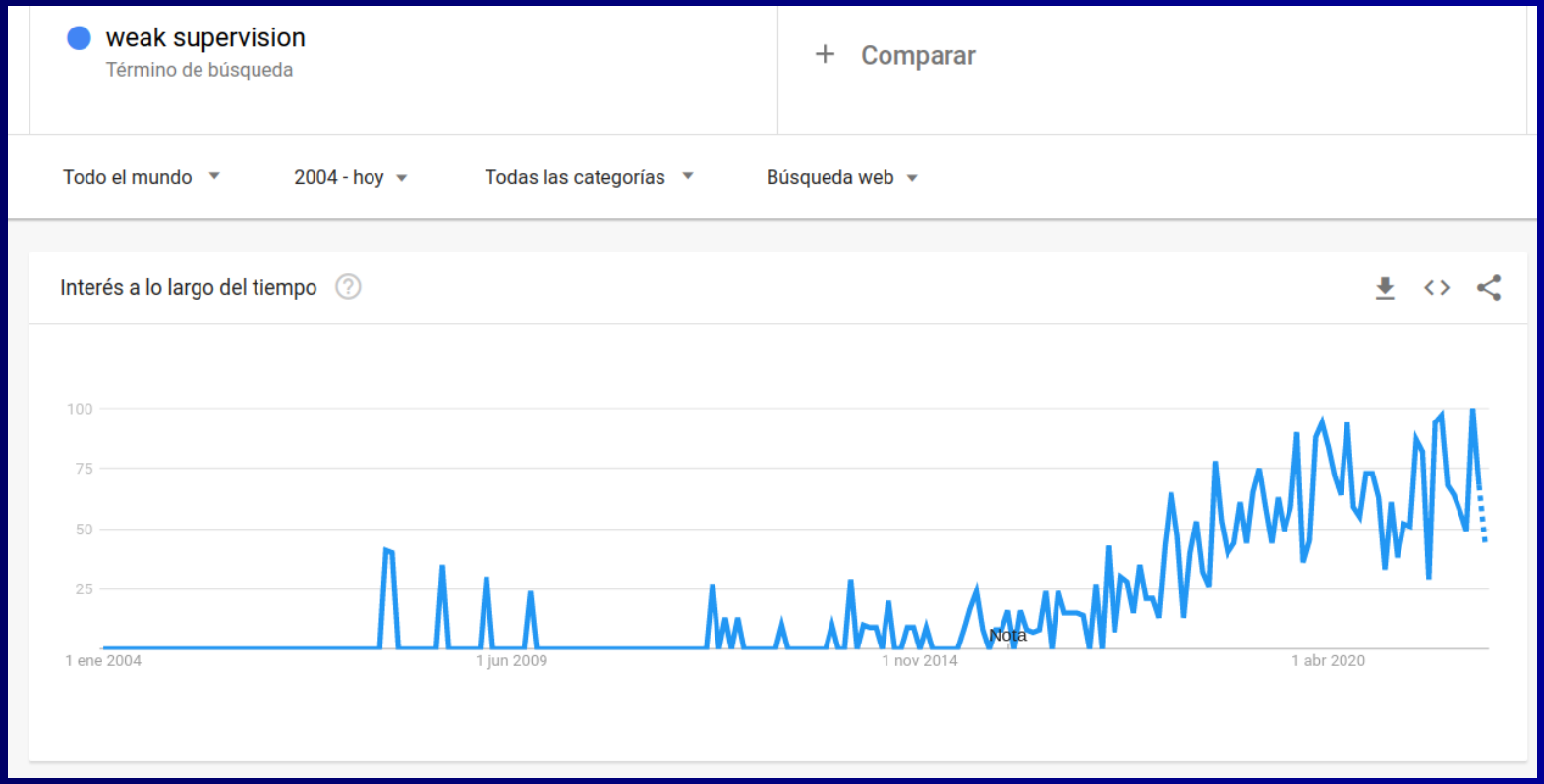
BMC Bioinformatics

PROCEEDINGS Open Access

Weakly supervised learning of biomedical information extraction from curated data

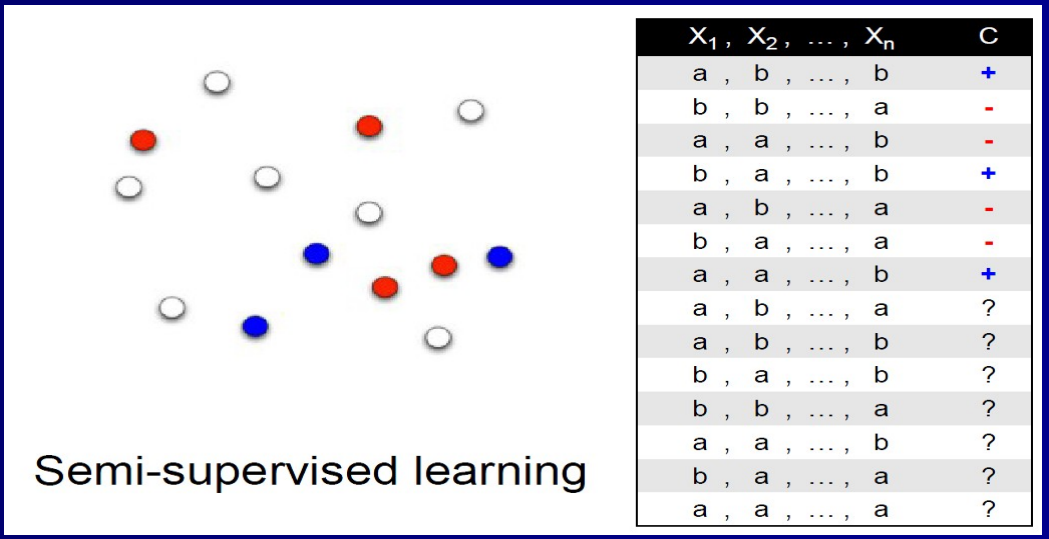
Suvir Jain[1†], Kashyap R.[1†], Tsung-Ting Kuo[2], Shitij Bhargava[1], Gordon Lin[1] and Chun-Nan Hsu[2*]

THE TERM – GOOGLE TRENDS

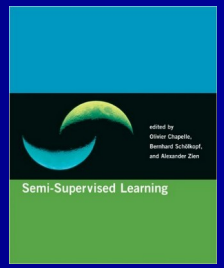


SEMI SUPERVISED CLASSIFICATION

- Most of the samples do not show a class value. Why?
  - Categorization: human-time consuming task
  - No knowledge to categorize the samples
- Objective: learn a supervised model
- Can a learning process which takes advantage of unlabeled samples, construct a better supervised classification model?



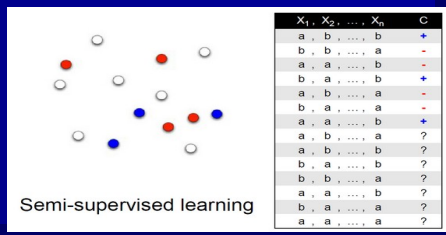
X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , a , ... , b	+
a , b , ... , a	?
a , b , ... , b	?
b , a , ... , b	?
b , b , ... , a	?
a , a , ... , b	?
b , a , ... , a	?
a , a , ... , a	?



SEMI SUPERVISED CLASSIFICATION

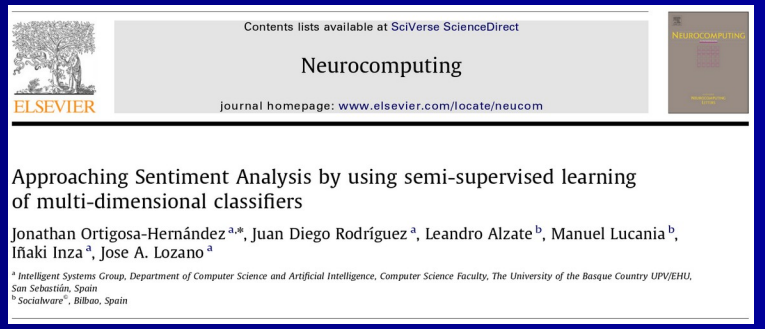
In traditional supervised learning problems, we are presented with an ordered collection of l labelled data points $D_L = ((x_i, y_i))_{i=1}^{l}$. Each data point (x_i, y_i) consists of an object $x_i \in \mathcal{X}$ from a given input space $\mathcal{X}$, and has an associated label y_i, where y_i is real-valued in regression problems and categorical in classification problems. Based on a collection of these data points, usually called the *training data*, supervised learning methods attempt to infer a function that can successfully determine the label y^* of some previously unseen input x^*.

In many real-world classification problems, however, we also have access to a collection of u data points, $D_U = (x_i)_{i=l+1}^{l+u}$, whose labels are unknown. For instance, the data points for which we want to make predictions, usually called the *test data*, are unlabelled by definition. Semi-supervised classification methods attempt to utilize unlabelled data points to construct a learner whose performance exceeds the performance of learners obtained when using only the labelled data. In the remainder of this survey, we denote with X_L and X_U the collection of input objects for the labelled and unlabelled samples, respectively.[1]



SEMI SUPERVISED LEARNING SENTIMENT ANALYSIS

- Companies: reputation
- Opinions about its products:
  - social networks
  - blogs
  - forums...
- Automatically classify the written opinion: {+, -, neutral}
- NLP: “Natural Language Processing”



Contents lists available at SciVerse ScienceDirect

Neurocomputing

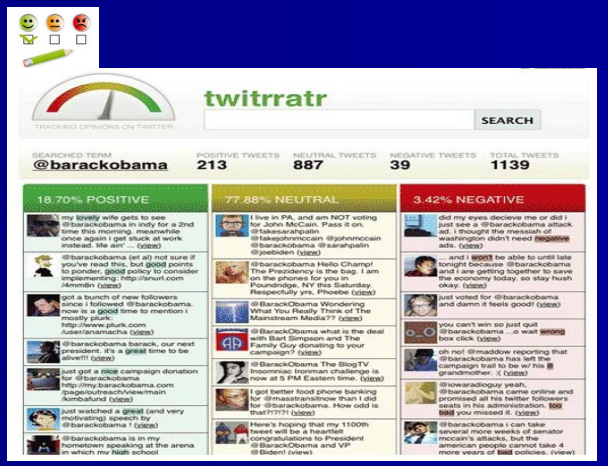
journal homepage: www.elsevier.com/locate/neucom

Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers

Jonathan Ortigosa-Hernández [a,*], Juan Diego Rodríguez [a], Leandro Alzate [b], Manuel Lucania [b], Iñaki Inza [a], Jose A. Lozano [a]

[a] Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, Computer Science Faculty, The University of the Basque Country UPV/EHU, San Sebastián, Spain

[b] Socialware®, Bilbao, Spain



SEMI SUPERVISED LEARNING

Machine Learning, 39, 103–134, 2000.

Text Classification from Labeled and Unlabeled Documents using EM

KAMAL NIGAM knigam@cs.cmu.edu
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ANDREW KACHITES MCCALLUM mccallum@justresearch.com
Just Research, 4616 Henry Street, Pittsburgh, PA 15213, USA; School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

SEBASTIAN THRUN thrun@cs.cmu.edu
TOM MITCHELL tom.mitchell@cmu.edu
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Int. J. Mach. Learn. & Cyber. (2017) 8:355–370
DOI 10.1007/s13042-015-0328-7

ORIGINAL ARTICLE

Semi-supervised self-training for decision tree classifiers

Jafar Tanha · Maarten van Someren · Hamideh Afsarmanesh

Combining Labeled and Unlabeled Data with Co-Training*†

Avrim Blum
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891
avrim+@cs.cmu.edu

Tom Mitchell
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891
mitchell+@cs.cmu.edu

AMBIGUOUS TRAINING DATA

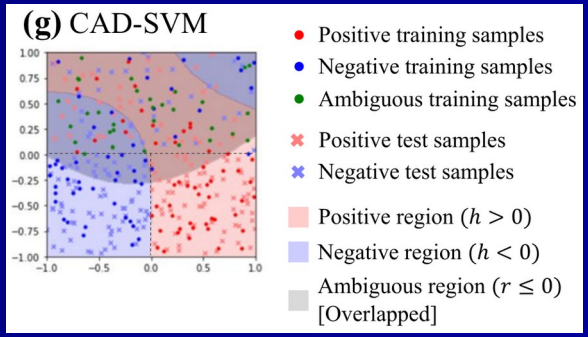
We consider three class labels, i.e,. positive, ambiguous, and negative: $y \in \mathcal{Y}_0 = \{1, 0, -1\}$. Suppose that we are given a set of positive, ambiguous, and negative samples $\{(x_i, y_i)\}_{i=1}^{N}$ drawn independently from the probability distribution with density $p_0(x, y)$ defined on $\mathcal{X} \times \mathcal{Y}_0$. Our goal is still to learn a discriminant function that classifies test samples into either the positive or negative class (not in the ambiguous class). Our key question in this scenario is if we can utilize the ambiguous training data to improve the classification accuracy of the discriminant function.

X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , a , ... , b	+
a , b , ... , a	?
a , b , ... , b	?
b , a , ... , b	?
b , b , ... , a	?
a , a , ... , b	?
b , a , ... , a	?
a , a , ... , a	?

Machine Learning (2020) 109:2369–2388
https://doi.org/10.1007/s10994-020-05915-2

Binary classification with ambiguous training data

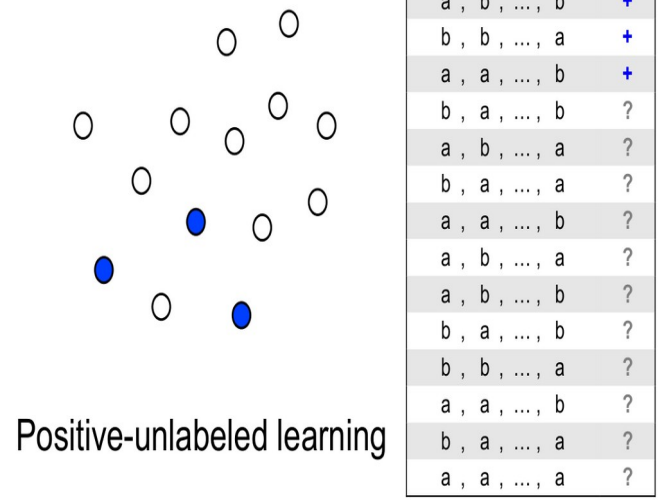
Naoya Otani[1] · Yosuke Otsubo[1] · Tetsuya Koike[1] · Masashi Sugiyama[2,3]



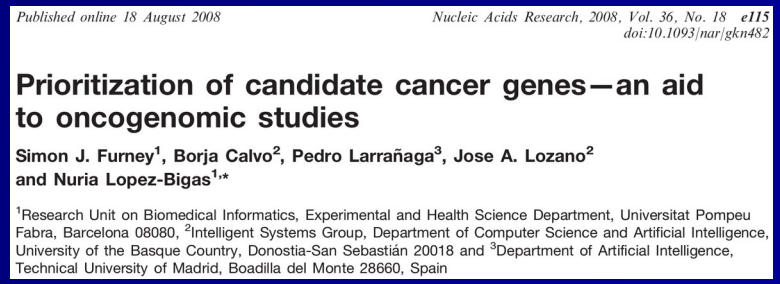
POSITIVE UNLABELED LEARNING

- More difficult than semi-supervised classification
- Prediction: "+" or "-"

- Application → prediction of genes related to cancer
- Web page visiting prediction → personalized ads



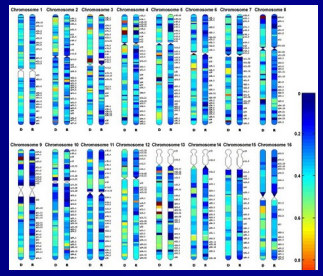
$X_1, X_2, \ldots, X_n$	C
a , b , … , b	+
b , b , … , a	+
a , a , … , b	+
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?


Published online 18 August 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 18* **e115**
doi:10.1093/nar/gkn482

Prioritization of candidate cancer genes—an aid to oncogenomic studies

Simon J. Furney[1], Borja Calvo[2], Pedro Larrañaga[3], Jose A. Lozano[2] and Nuria Lopez-Bigas[1,*]

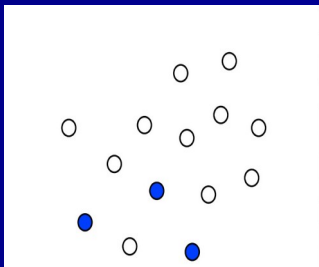
[1]Research Unit on Biomedical Informatics, Experimental and Health Science Department, Universitat Pompeu Fabra, Barcelona 08080, [2]Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country, Donostia-San Sebastián 20018 and [3]Department of Artificial Intelligence, Technical University of Madrid, Boadilla del Monte 28660, Spain



POSITIVE UNLABELED LEARNING

Problem settings Let $X \in \mathbb{R}^d$ and $Y \in \{\pm 1\}$ ($d \in \mathbb{N}$) be the input and output random variables. Let $p(x, y)$ be the *underlying joint density* of (X, Y), $p_{\mathrm{p}}(x) = p(x \mid Y = +1)$ and $p_{\mathrm{n}}(x) = p(x \mid Y = -1)$ be the *P and N marginals* (a.k.a. the P and N class-conditional densities), $p(x)$ be the *U marginal*, $\pi_{\mathrm{p}} = p(Y = +1)$ be the *class-prior probability*, and $\pi_{\mathrm{n}} = p(Y = -1) = 1 - \pi_{\mathrm{p}}$. π_{p} is assumed known throughout the paper; it can be estimated from P and U data [23, 24, 25, 26].

Consider the *two-sample problem setting* of PU learning [5]: two sets of data are sampled independently from $p_{\mathrm{p}}(x)$ and $p(x)$ as $\mathcal{X}_{\mathrm{p}} = \{x_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}} \sim p_{\mathrm{p}}(x)$ and $\mathcal{X}_{\mathrm{u}} = \{x_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}} \sim p(x)$, and a classifier needs to be trained from $\mathcal{X}_{\mathrm{p}}$ and $\mathcal{X}_{\mathrm{u}}$.[2] If it is PN learning as usual, $\mathcal{X}_{\mathrm{n}} = \{x_i^{\mathrm{n}}\}_{i=1}^{n_{\mathrm{n}}} \sim p_{\mathrm{n}}(x)$ rather than $\mathcal{X}_{\mathrm{u}}$ would be available and a classifier could be trained from $\mathcal{X}_{\mathrm{p}}$ and $\mathcal{X}_{\mathrm{n}}$.



Positive-unlabeled learning

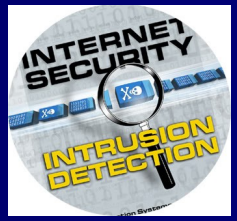
$X_1, X_2, \ldots, X_n$	C
a , b , … , b	+
b , b , … , a	+
a , a , … , b	+
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

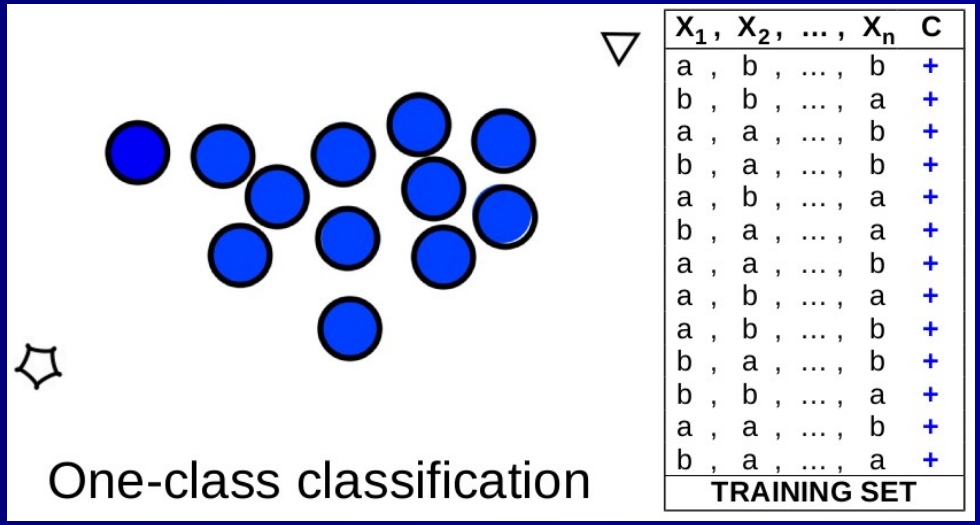
ONE CLASS CLASSIFICATION – OUTLIER DETECTION –

- One category: forms a representative sample
- Only “normal behaviour” samples in training time

- Training phase: model the “normal” behaviour
- Prediction phase → detect “deviations” from the “normal” model

- Model the “dominant” class + “isolate” outliers in “operation phase”



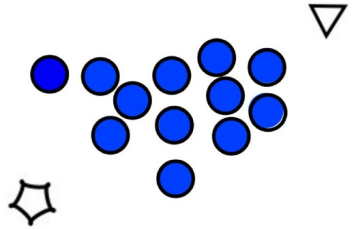


X_1	X_2	...	X_n	C
a	b	...	b	+
b	b	...	a	+
a	a	...	b	+
b	a	...	b	+
a	b	...	a	+
b	a	...	a	+
a	a	...	b	+
a	b	...	a	+
a	b	...	b	+
b	a	...	b	+
b	b	...	a	+
a	a	...	b	+
b	a	...	a	+

TRAINING SET

ONE CLASS CLASSIFICATION – OUTLIER DETECTION –

"normal" patterns $\boldsymbol{X}$ are available for training, while "abnormal" ones are relatively few. A model of normality $M(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the free parameters of the model, is inferred and used to assign novelty scores $z(\mathbf{x})$ to previously unseen test data $\mathbf{x}$. Larger novelty scores $z(\mathbf{x})$ correspond to increased "abnormality" with respect to the model of normality. A novelty threshold $z(\mathbf{x}) = k$ is defined such that $\mathbf{x}$ is classified "normal" if $z(\mathbf{x}) \leq k$, or "abnormal" otherwise. Thus, $z(\mathbf{x}) = k$ defines a decision boundary.

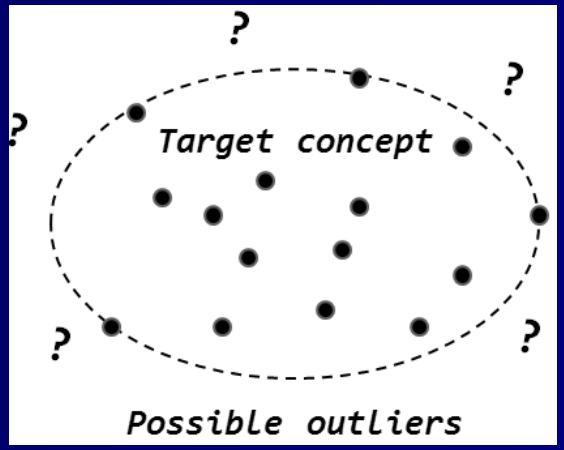


One-class classification

X_1	X_2	...	X_n	C
a	b	...	b	+
b	b	...	a	+
a	a	...	b	+
b	a	...	b	+
a	b	...	a	+
b	a	...	a	+
a	a	...	b	+
a	b	...	a	+
a	b	...	b	+
b	a	...	b	+
b	b	...	a	+
a	a	...	b	+
b	a	...	a	+

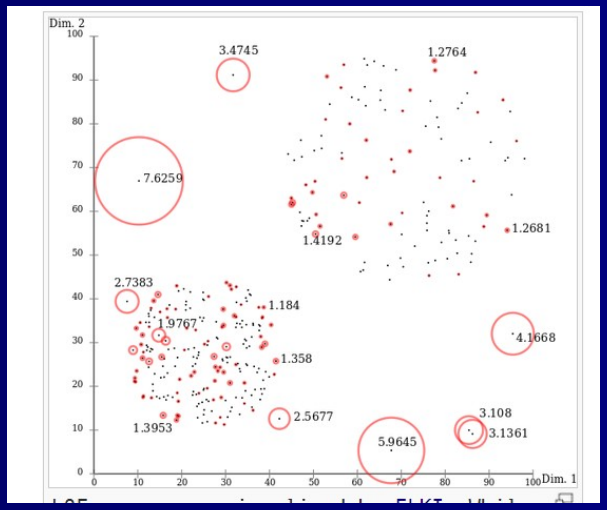
TRAINING SET

ONE CLASS CLASSIFICATION – OUTLIER DETECTION –



OneClass SVM
AutoEncoders

1-Class data



Local Outlier Factor
Isolation Forests

MultiClass data

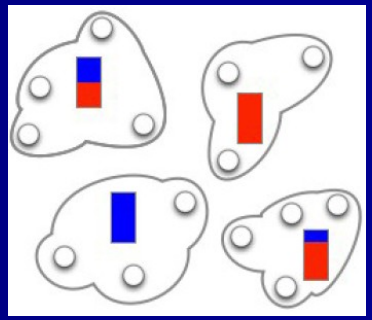
LEARNING with LABEL PROPORTIONS

$X_1, X_2, \ldots, X_n$	C
a , b , … , b	
b , b , … , a	0.5
a , a , … , b	0.5
b , a , … , b	
a , b , … , a	
b , a , … , a	0.0
a , b , … , a	1.0
a , a , … , b	
a , b , … , b	0.25
b , a , … , b	0.75
b , a , … , a	
a , a , … , b	
b , b , … , a	1.0
a , a , … , a	0.0

Supervised Learning by Training on Aggregate Outputs

David R. Musicant, Robert Atlas, Janara M. Christensen, Jamie F. Olson, Jeffrey M. Rzeszotarski, Emma R. D. Turetsky

Abstract—Supervised learning is a classic data mining problem where one wishes to be able to predict an output value associated with a particular input vector. We present a new twist on this classic problem where, instead of having the training set contain an individual output value for each input vector, the output values in the training set are only given in aggregate over a number of input vectors. This new problem arose from a particular need in learning on mass spectrometry data, but could easily apply to situations when data has been aggregated in order to maintain privacy. We provide a formal description of this new problem for both classification and regression. We then examine how k-nearest neighbor, neural networks, support vector machines, and decision trees can be adapted for this problem.



LEARNING with LABEL PROPORTIONS

The dataset D of a LLP problem is composed of m unlabeled examples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$. In this paper, we assume that D has been sampled i.i.d. from some underlying probability distribution. The examples are provided grouped in b *bags* ($D = \mathbf{B}_1 \cup \mathbf{B}_2 \cup \cdots \cup \mathbf{B}_b$ where $\mathbf{B}_i \cap \mathbf{B}_j = \emptyset, \forall i \neq j$). Each bag $\mathbf{B}_i$ groups m_i instances, where $\sum_{i=1}^{b} m_i = m$, and m_{ic} denotes the number of instances in $\mathbf{B}_i$ which have the label c. These m_{ic} values, called *counts* of the bag $\mathbf{B}_i$, sum up to m_i; i.e. $\sum_{c \in C} m_{ic} = m_i$. Similarly, bag class information can be provided in terms of *proportions* [6], $p_{ic} = m_{ic}/m_i \in [0, 1]$, with $\sum_{c \in C} p_{ic} = 1$.

$X_1, X_2, \ldots, X_n$	C
a , b , ... , b	
b , b , ... , a	0.5
a , a , ... , b	0.5
b , a , ... , b	
a , b , ... , a	
b , a , ... , a	0.0
a , b , ... , a	1.0
a , a , ... , b	
a , b , ... , b	0.25
b , a , ... , b	0.75
b , a , ... , a	
a , a , ... , b	
b , b , ... , a	1.0
a , a , ... , a	0.0

LABEL PROPORTIONS – APPLICATIONS –

Embryo selection in Assisted Reproductive Technologies (ART)

Two steps:

- **Transfer**: step in which one or several embryos are placed into the uterus of the patient.
- **Implantation**: step in which pregnancy is established (by one or several embryos).

Application	MILIp problem
Transfered embryos	Dataset
Implanted or not	Class labels
ART process	Bag
Number of children	Label proportions

Information Sciences 481 (2019) 381–393

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Aggregated outputs by linear models: An application on marine litter beaching prediction

Jerónimo Hernández-González[a,*], Iñaki Inza[a], Igor Granado[b], Oihane C. Basurko[b], Jose A. Fernandes[b], Jose A. Lozano[a,c]

[a] Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia, Spain
[b] Marine Research Division at AZTI-Tecnalia, Pasaia, Spain
[c] Basque Center for Applied Mathematics, Bilbao, Spain

SMMR STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Fitting the data from embryo implantation prediction: Learning from label proportions

Jerónimo Hernández-González,[1] Iñaki Inza,[1] Lorena Crisol-Ortíz,[2] María A Guembe,[2] María J Iñarra[2] and Jose A Lozano[1,3]

Statistical Methods in Medical Research
0(0) 1–11

DOI: 10.1177/0962280216651098
smm.sagepub.com
SAGE

2017 IEEE International Conference on Data Mining

A Probabilistic Approach for Learning with Label Proportions Applied to the US Presidential Election

Tao Sun[1], Dan Sheldon[1,2], Brendan O'Connor[1]
[1]College of Information and Computer Sciences, University of Massachusetts Amherst
[2]Department of Computer Science, Mount Holyoke College
Email: {taosun, sheldon, brenocon}@cs.umass.edu

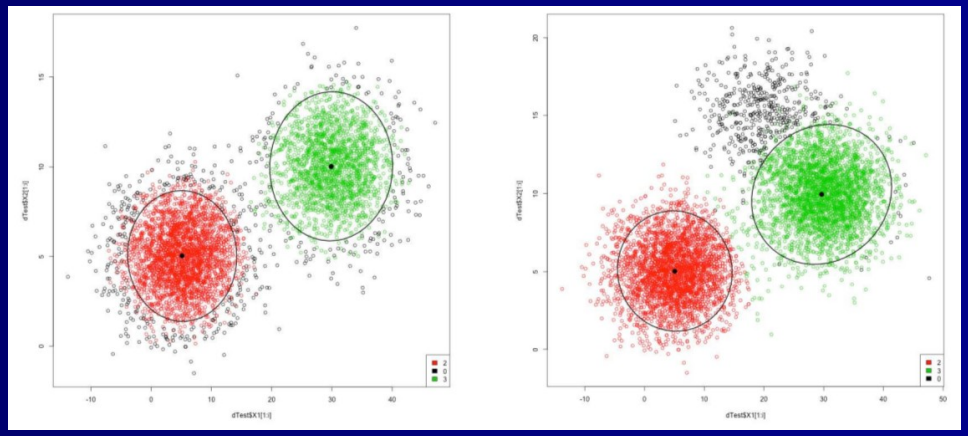
Possible voters based on previous election results

- It involves any situation related with an election that can be organised as follows:

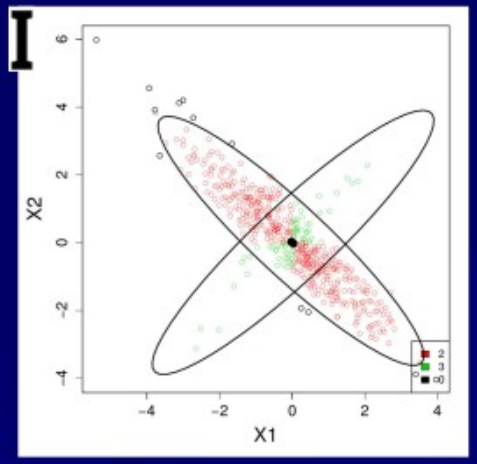
Application	MILIp problem
Census	Dataset
Candidates	Class labels
Polling station	Bag
Election results	Label proportions

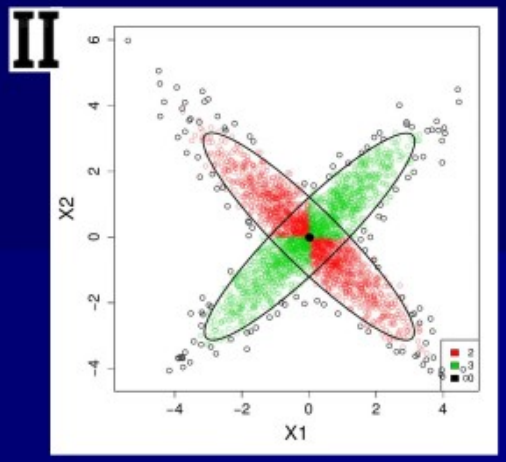
NOVELTY DETECTION

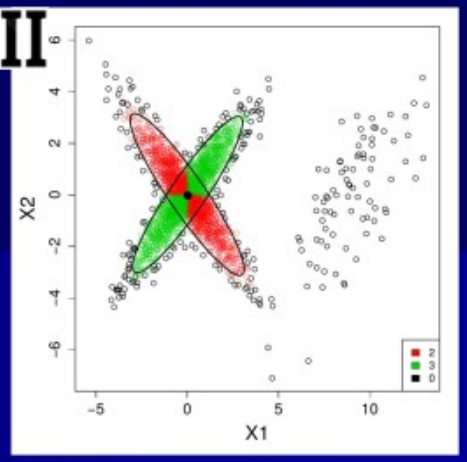
- Initially labeled dataset → train a model
- Unlabeled samples arrive → in 2nd dataset – or streaming
- 2nd dataset → an "emergent" class appears?
- "Novel class"? → "detect + baptise"
- Separation + cohesion
- Re-train the model with the "baptised class" samples

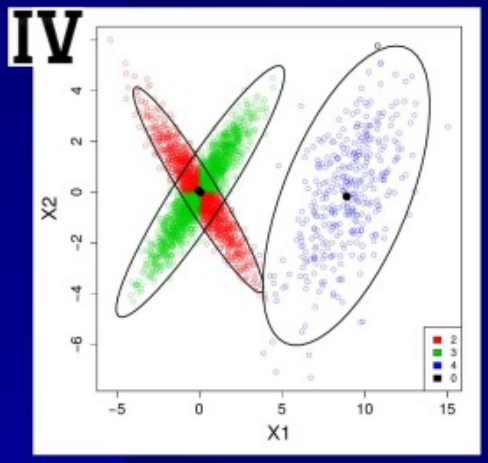


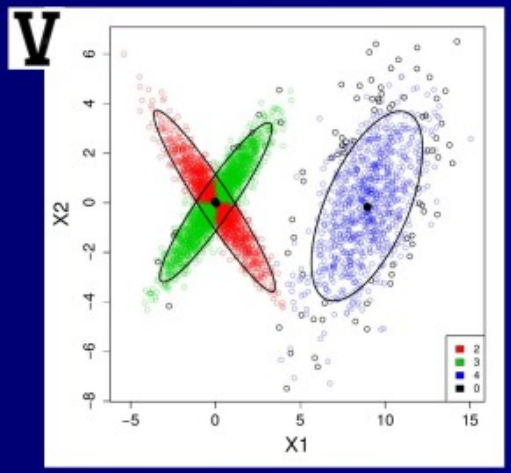
NOVELTY DETECTION

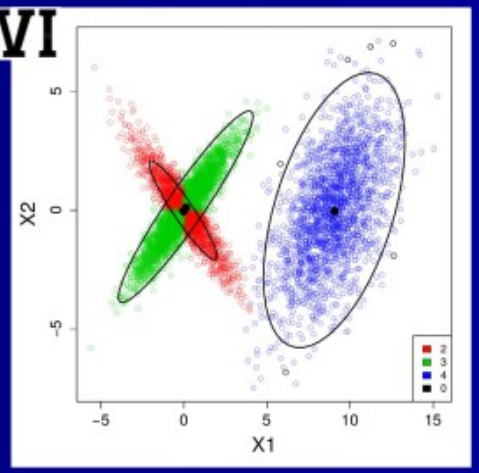






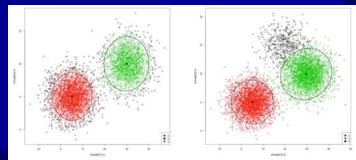






NOVELTY DETECTION

Suppose we are given a training dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\} \in \{\mathcal{X}, \mathcal{Y}\}$, and m unlabeled examples $\{\mathbf{x}_1^U, \ldots, \mathbf{x}_m^U\} \in \mathcal{X}$. $\mathcal{X}$ is a d-dimensional space. $\mathcal{Y}$ denotes the label space for the training set and $y_i \in \{1, 2, \ldots, l\}$, where l is the number of classes. Different from traditional learning problems, the label space of the unlabeled examples are larger than that of the labeled examples, and is denoted as $\mathcal{Y}^{'}$, $\mathcal{Y}^{'} \supseteq \mathcal{Y}$. Suppose $\mathcal{Y}^{'} = \{1, 2, \ldots, l + k\}$, where k is the number of new categories in the unlabeled examples. The main objective of *Serendipitous Learning (SL)* is to infer the labels of the unlabeled examples, so that the examples belonging to $\mathcal{Y}$ can be correctly classified, and the examples belonging to the novel categories, i.e., $\mathcal{Y}^{'} - \mathcal{Y} = \{l + 1, l + 2, \ldots, l + k\}$, can be clustered to the correct group.



CLASSIFICATION WITH PARTIAL LABELS

Journal of Machine Learning Research 12 (2011) 1501-1536 Submitted 10/10; Revised 2/11; Published 5/11

Learning from Partial Labels

Timothee Cour TIMOTHEE.COUR@GMAIL.COM
NEC Laboratories America
10080 N Wolfe Rd # Sw3350
Cupertino, CA 95014, USA

Benjamin Sapp BENSAPP@CIS.UPENN.EDU
Ben Taskar TASKAR@SEAS.UPENN.EDU
Department of Computer and Information Science
University of Pennsylvania
3330 Walnut Street
Philadelphia, PA 19107, USA

In the standard supervised multiclass setting, we have labeled examples $S = \{(x_i, y_i)_{i=1}^m\}$ from an unknown distribution $P(X,Y)$ where $X \in \mathcal{X}$ is the input and $Y \in \{1,\ldots,L\}$ is the class label. In the partially supervised setting we investigate, instead of an unambiguous single label per instance we have a set of labels, one of which is the correct label for the instance. We will denote $\mathbf{y}_i = \{y_i\} \cup \mathbf{z}_i$ as the ambiguity set actually observed by the learning algorithm, where $\mathbf{z}_i \subseteq \{1,\ldots,L\} \setminus \{y_i\}$ is a set of additional labels, and y_i the latent groundtruth label which we would like to recover.

Each instance comes annotated with several class labels but only one of them is valid.

X_1 , X_2 , ... , X_n	C
a , b , ... , b	a,b,c
b , b , ... , a	a,c
a , a , ... , b	d
b , a , ... , b	b,c
a , b , ... , a	a,d
b , a , ... , a	a,b,d
a , a , ... , b	b,c,d
a , b , ... , a	c
a , a , ... , b	b,c
b , a , ... , a	b
a , a , ... , a	a,b

CLASSIFICATION UNDER PARTIAL MULTI-LABEL

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 44, NO. 7, JULY 2022

Partial Multi-Label Learning With Noisy Label Identification

Ming-Kun Xie and Sheng-Jun Huang

For each partially labeled training example, we denote by $\mathbf{x}_i \in \mathbb{R}^d$ a feature vector and its corresponding label vector $\mathbf{y} \in \{0,1\}^q$ with q class labels. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \{0,1\}^{q \times n}$ denote the noise-corrupted label matrix. In this setting, $\mathbf{y}_{ji} = 1$ means the j-th label is a candidate label to the i-th instance. We further denote by $\widetilde{\mathbf{y}} \in \{0,1\}^q$ the unknown ground-truth label vector.

Each instance is assigned with a candidate label sets, which contains multiple relevant labels and some noisy labels.

X_1 , X_2 , ... , X_n	C
a , b , ... , b	a,b,c
b , b , ... , a	a,c
a , a , ... , b	d
b , a , ... , b	b,c
a , b , ... , a	a,d
b , a , ... , a	a,b,d
a , a , ... , b	b,c,d
a , b , ... , a	c
a , a , ... , b	b,c
b , a , ... , a	b
a , a , ... , a	a,b

PROBABILISTIC LABELS

LABEL DISTRIBUTIONS

X_1 , X_2 , … , X_n	c_1	c_2	c_3
a , b , … , b	0.3	0.3	0.4
b , b , … , a	0.4	0.2	0.4
a , a , … , b	0	1	0
a , b , … , a	0.7	0.2	0.1
b , a , … , a	0.5	0.5	0
a , a , … , b	0.3	0.1	0.6
a , b , … , a	0.4	0.2	0.4
a , b , … , b	0.7	0.2	0.4
b , a , … , b	0.9	0.1	0
b , b , … , a	0.6	0.2	0.2

Learning from data with uncertain labels by boosting credal classifiers

Benjamin Quost
HeuDiaSyC laboratory
deptartment of Computer Science
Compiègne University of Technology
Compiègne, France
quostben@hds.utc.fr

Thierry Denœux
HeuDiaSyC laboratory
deptartment of Computer Science
Compiègne University of Technology
Compiègne, France
tdenoeux@hds.utc.fr

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 7, JULY 2016

Label Distribution Learning

Xin Geng, *Member, IEEE*

PROBABILISTIC LABELS

LABEL DISTRIBUTIONS

First of all, the main notations used in this paper are listed as follows. The instance variable is denoted by $\boldsymbol{x}$, the particular i-th instant is denoted by $\boldsymbol{x}_i$, the label variable is denoted by y, the particular j-th label value is denoted by y_j, the description degree of y to $\boldsymbol{x}$ is denoted by $d_{\boldsymbol{x}}^{y}$, and the label distribution of $\boldsymbol{x}_i$ is denoted by $D_i = \{d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \cdots, d_{\boldsymbol{x}_i}^{y_c}\}$, where c is the number of possible label values.

represents a general case of label distribution, which satisfies the constraints $d_{\boldsymbol{x}}^{y} \in [0, 1]$ and $\sum_y d_{\boldsymbol{x}}^{y} = 1$. Such examples illustrate that label distribution is more general than both single-label annotation and multi-label annotation, and thus can provide more flexibility in the learning process.

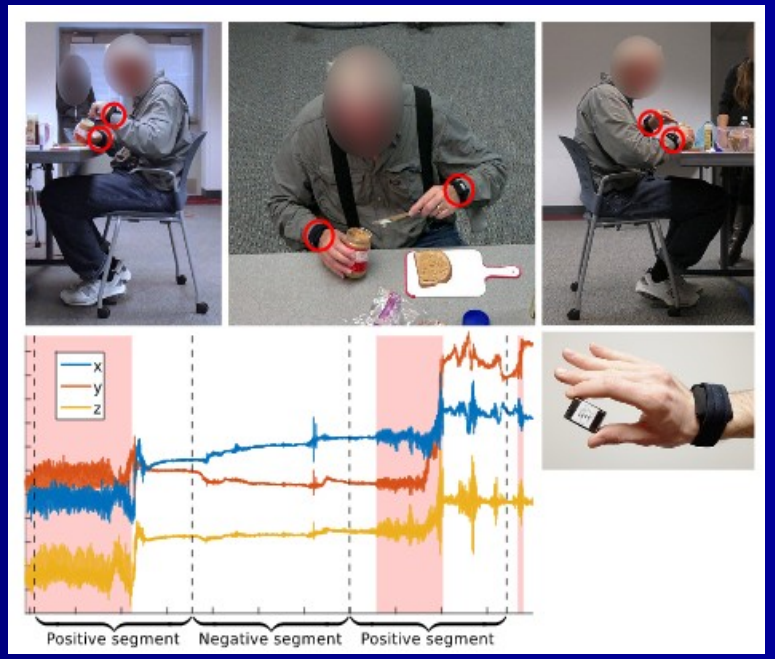
STRATIFIED LEARNING

- Approximate proportion of each label is provided
- Labeling → intervals of labels' proportions

Weakly-supervised Learning for Parkinson's Disease Tremor Detection

Ada Zhang[1], Alexander Cebulla[2], Stanislav Panev[1], Jessica Hodgins[1], and Fernando De la Torre[1]
[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2] ETH Zurich, Switzerland

algorithms degrades as labels become less precise. (2) We provide a simple modification to existing weakly-supervised learning algorithms that allow them to take advantage of labels containing the approximate percentage of tremor (*e.g.*, 0-24%, 25-49%, 50-74%, 75-100%) within a segment.



FULL-CLASS SET CLASSIFICATION

X_1 , X_2 , ... , X_n	C
a , b , ... , b	b
b , b , ... , a	c
a , a , ... , b	c
b , a , ... , b	b
a , b , ... , a	a
b , a , ... , a	a
a , a , ... , b	c
a , b , ... , a	c
a , a , ... , b	b
b , a , ... , a	b
a , a , ... , a	a

Training

X_1 , X_2 , ... , X_n	C
b , b , ... , a	Y_1
a , a , ... , b	
b , a , ... , b	
a , b , ... , a	Y_2
b , a , ... , a	
a , b , ... , a	
a , a , ... , b	Y_3
a , b , ... , b	
b , a , ... , b	
a , a , ... , b	Y_4
b , b , ... , a	
a , a , ... , a	

Test

$Y_i = (y_{i1}, y_{i2}, y_{i3}) =$ PermutationOf{a,b,c}

Permutations						
y_{i1}	a	a	b	b	c	c
y_{i2}	b	c	a	c	a	b
y_{i3}	c	b	c	a	b	a

Int. J. Mach. Learn. & Cyber. (2010) 1:53–61
DOI 10.1007/s13042-010-0002-z

ORIGINAL ARTICLE

Full-class set classification using the Hungarian algorithm

Ludmila I. Kuncheva

- identify different known objects in a group
- identify the location of the students in the classroom
- Supervision degree in prediction time!! Class-info extra in prediction time!! → permutation of labels

FULL-CLASS SET CLASSIFICATION

Consider a classification problem where an instance $\mathbf{x}$ may come from one of the c classes in the set $\Omega = \{\omega_1, \ldots, \omega_c\}$. Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_c\}$ be a set containing exactly one instance from each class. A set-classifier, D_{set}, will label any set X with a permutation of the class indices.

The accuracy of D_{set} is the probability that the *whole* set is labelled correctly. D_{set} can be constructed using the output of a base individual classifier D_{ind}. We assume that D_{ind} outputs estimates of $P(\omega_i|\mathbf{x}_j)$, the posterior probability that instance $\mathbf{x}_j$ belongs to class ω_i. Then D_{set} is defined as

$$D_{\text{set}} : \mathcal{M} \rightarrow \mathcal{I}, \tag{1}$$

where $\mathcal{I}$ is the set of all permutations of the class indices, and $\mathcal{M}$ is the set of all square matrices of size $(c \times c)$ with entries $m_{(i,j)} \in [0, 1]$, such that $\sum_{j=1}^{c} m_{(i,j)} = 1$.

$Y_i = (y_{i1}, y_{i2}, y_{i3}) =$ PermutationOf{a,b,c}

RESTRICTED SET CLASSIFICATION

Pattern Recognition 63 (2017) 158–170

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Restricted set classification: Who is there?

Ludmila I. Kuncheva[a,*], Juan J. Rodríguez[b], Aaron S. Jackson[c]

[a] Bangor University, Dean Street, Bangor Gwynedd LL57 1UT, United Kingdom
[b] University of Burgos, Escuela Politécnica Superior, Avda. de Cantabria s/n, 09006 Burgos, Spain
[c] The University of Nottingham, NG8 1BB, United Kingdom

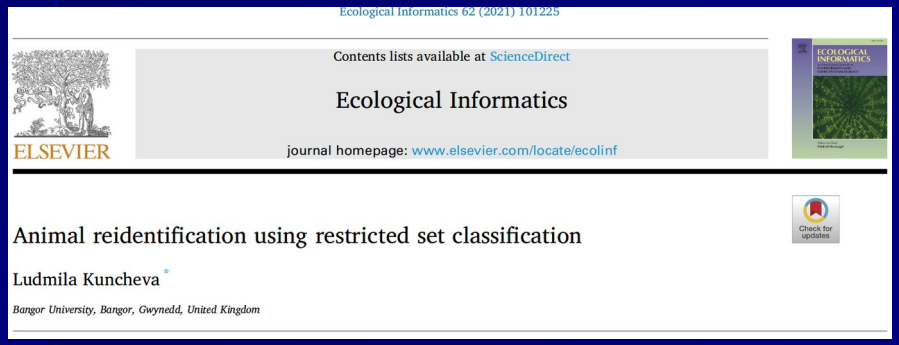
Classes in the chess-pieces recognition problem, and the limit number for each class in a standard chess game.

Class #:	1	2	3	4	5	6	7	8	9	10	11	12	13
	King	Queen	Rook	Bishop	Knight	Pawn	King	Queen	Rook	Bishop	Knight	Pawn	Empty
Numbers allowed:	1	1	2	2	2	8	1	1	2	2	2	8	62

This paper extends the above model to the more general case where X consists of m instances, and it is known that at most k_i instances may belong to class ω_i, $i = 1, \ldots, c$. Denoting $k = k_1 + \cdots + k_c$, we require that $m \leq k$. The who-is-who task is a special case where k_i=1, $i = 1, \ldots, c$, and m=c.

- When Predicting → Maximum number of samples per class is upper-bounded:
  - "Supervision degree" in prediction time !!
- Illustrative application → recognition of chess pieces

RESTRICTED SET CLASSIFICATION



Ecological Informatics 62 (2021) 101225

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Animal reidentification using restricted set classification

Ludmila Kuncheva *

Bangor University, Bangor, Gwynedd, United Kingdom

An individual animal cannot be present more than once in the same image.

- When Predicting → Maximum number of samples per class is upper-bounded:
  - "Supervision degree" in prediction time
- Illustrative application → identification of individual animals in images
  → one individual animal per image

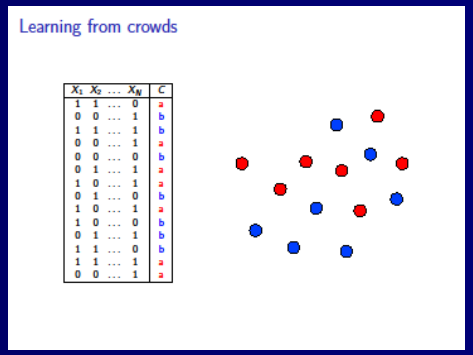
AND WHEN ANNOTATIONS ARE NOT FULLY RELIABLE?...

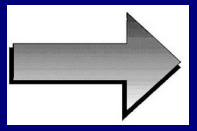
LEARNING FROM CROWDS

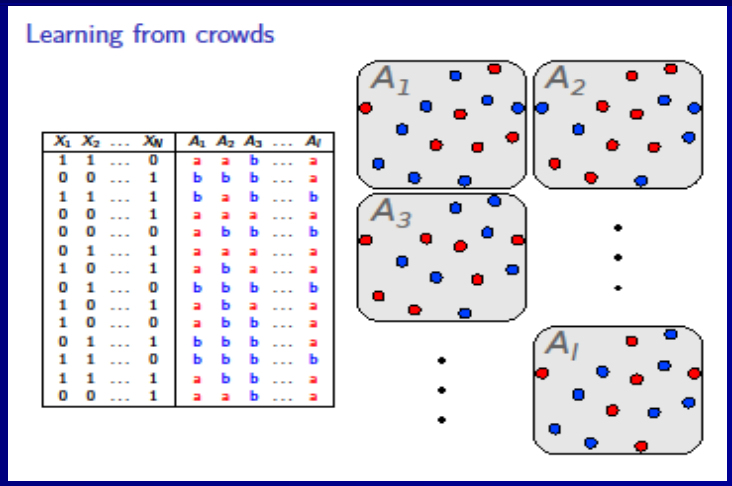
empresas que hacen la anotacion de los datos, anotan los c_i

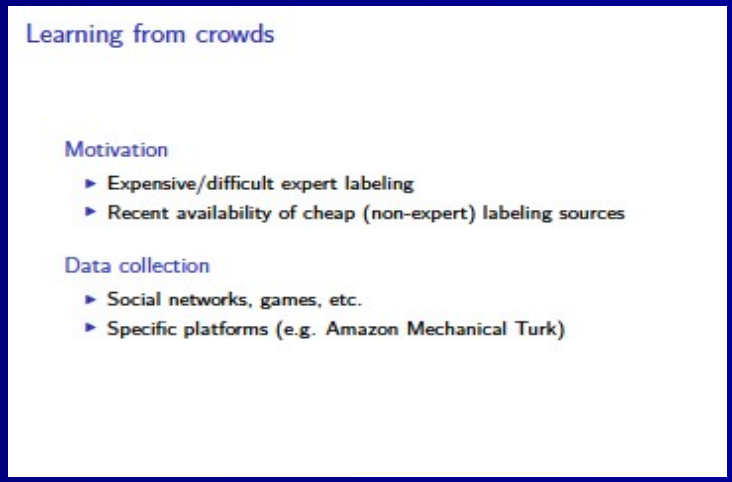
- real class for each object is not known → no "golden truth"

- domain experts (A_i) annotate their opinion about the label of each object







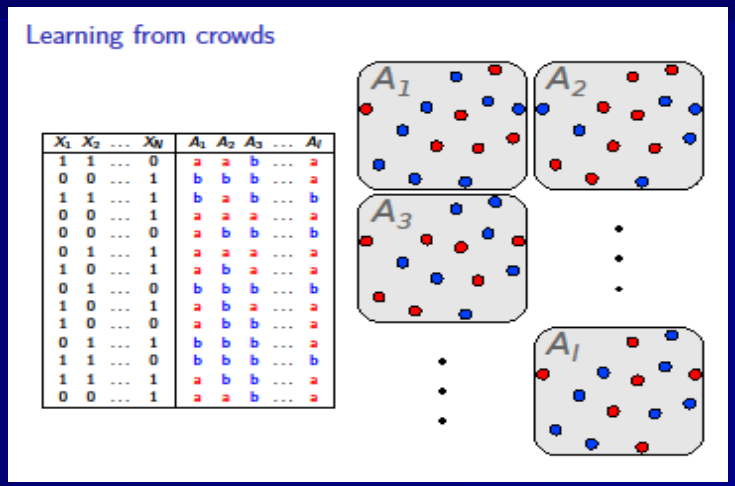


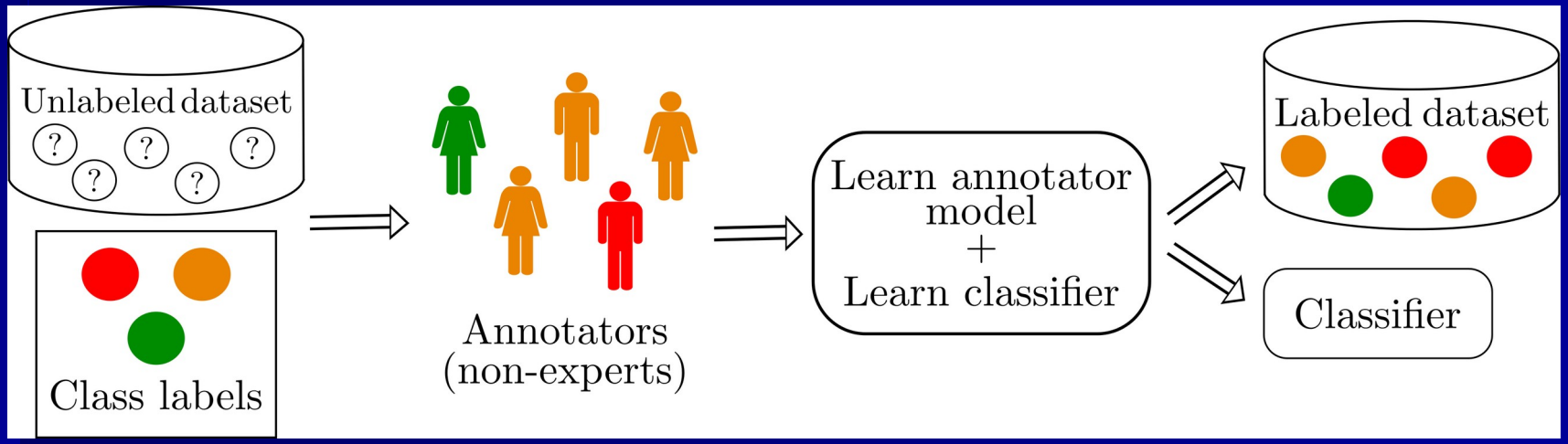
- Crowd annotation platforms

- Prolific.co, TolokaAI, AI Crowd, Amazon Mechanial Turk...

AND WHEN ANNOTATIONS ARE NOT FULLY RELIABLE?...

LEARNING FROM CROWDS





AND WHEN ANNOTATIONS ARE NOT FULLY RELIABLE?...

LEARNING FROM CROWDS

Journal of Machine Learning Research 11 (2010) 1297-1322 Submitted 9/09; Revised 2/10; Published 4/10

Learning From Crowds

Vikas C. Raykar VIKAS.RAYKAR@SIEMENS.COM
Shipeng Yu SHIPENG.YU@SIEMENS.COM
CAD and Knowledge Solutions (IKM CKS)
Siemens Healthcare
Malvern, PA 19355 USA

Linda H. Zhao LZHAO@WHARTON.UPENN.EDU
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104 USA

Gerardo Hermosillo Valadez GERARDO.HERMOSILLOVALADEZ@SIEMENS.COM
Charles Florin CHARLES.FLORIN@SIEMENS.COM
Luca Bogoni LUCA.BOGONI@SIEMENS.COM
CAD and Knowledge Solutions (IKM CKS)
Siemens Healthcare
Malvern, PA 19355 USA

Linda Moy LINDA.MOY@NYUMC.ORG
Department of Radiology
New York University School of Medicine
New York, NY 10016 USA

Learning from crowds [2] considers a training dataset without expert supervision. By contrast, a set of noisy labelers annotates each example: $D = \{(\boldsymbol{x}^1, \boldsymbol{a}^1), \ldots, (\boldsymbol{x}^n, \boldsymbol{a}^n)\}$, where a^j is a t-tuple with $a_l^j \in \mathcal{C}$ indicating the class label assessed by labeler L_l for x^j.

SUPERVISION MODELS

Table 2
Collection of supervision models.

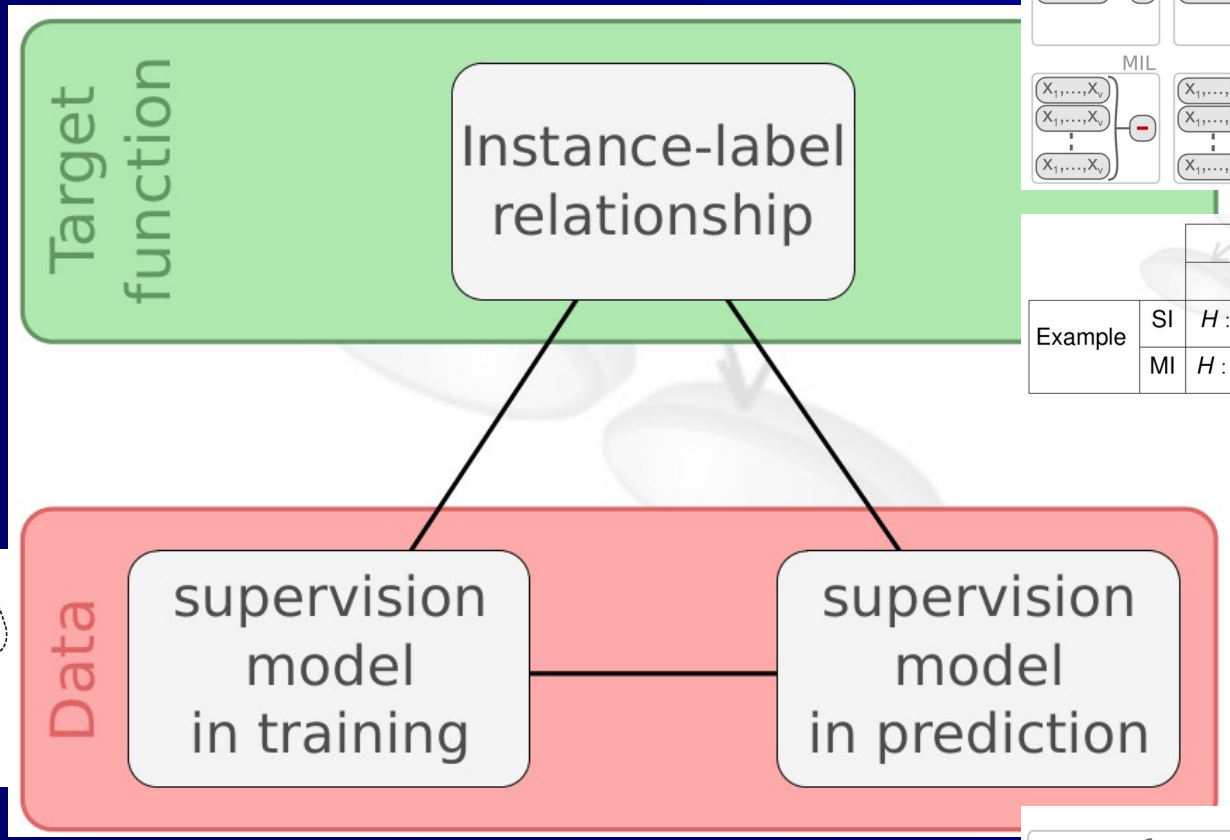
Model	References	Description
Full-supervision	[9,24,34,43]	For each example, complete class information is provided.
Unsupervision	[24]	No class information is provided with the examples.
Semi-supervision	[5]	Part of the examples are provided fully supervised. The rest are unsupervised.
Positive-unlabeled	[4,10,21,32]	Part of the examples are provided fully supervised, all of them with the same categorization. The rest are unsupervised.
Candidate labels	[7,13,16]	For each example, a set of class labels is provided. In this set, the class label(s) that compose the real categorization of the example are included.
Probabilistic labels	[18]	For each example, the probability of belonging to each class label is provided. This probability distribution is expected to assign high probability to the real label(s).
Incomplete	[3,33,42]	For each example, a subset of the labels that compose its real categorization is provided (SIML or MIML, Table 1).
Noisy labels	[2,44]	For each example, complete class information is provided, although its correctness is not guaranteed.
Crowd	[30,40]	For each example, many different non-expert annotators provide their (noisy) categorization.
Mutual label constraints	[19,20,31]	For each **group** of examples, an explicit relationship between their class labels is provided (e.g., all the examples have the same categorization).
Candidate labeling vectors	[22]	For each **group** of examples, a set of labeling vectors (including the real one) is provided. A labeling vector provides a class label for each examples of a group.
Label proportions	[15,25,28]	For each **group** of examples, the proportion of examples belonging to each class label is provided.

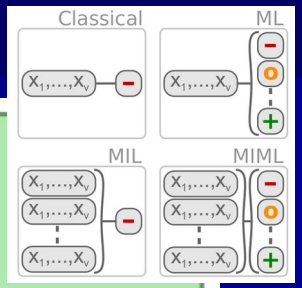
LEARNING SCENARIOS

Table 3
Brief description of classification problems and characterization according to the three axes of the taxonomy.

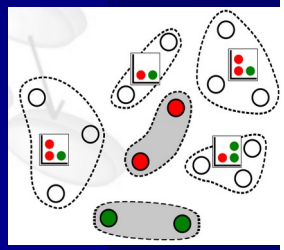
Problem	Description	Application (e.g.)	IL rel.	SUPERVISION MODEL	
				Learning	Prediction
Standard problem [24]	Learning with full categorized examples	Handwritten digit recogn.	SISL	Full-supervision	Unsupervision
Semi-supervised [5]	Learning with categorized and uncategorized examples	Text classification	SISL	Semi-supervision	Unsupervision
Positive-unlabeled [4]	Learning with examples of a category and other uncategorized examples	Spam detection, Gene prediction	SISL	Positive-unlabeled	Unsupervision
Mislabeled data [2]	Learning with maybe wrong-categorized examples	Subjective labeler	SISL	Noisy Labels	Unsupervision
Ambiguous labels [44]					
Partial labels [7]	Learning and prediction with uncategorized examples that have a set of possible categorizations	Classifying photographs with captions	SISL	Candidate labels	Unsupervision / Candidate labels
Multiple labels [18]	Learning with uncategorized examples that, with some probability, belong to a certain categorization	Bioinformatics	SISL	Probabilistic labels	Unsupervision
Partial equivalence relations [19]	Learning with groups of examples of the same/different categorization	Computer vision	SISL	Mutual label constraints	Unsupervision
Full-class set [20]	Prediction for a group of examples, all of them with a different categorization	Automatic attendance recording	SISL	Full-supervision	Mutual label constraints
Label proportions [15]	Learning with groups of examples only knowing how many of them belong to each categorization	Embryo Selection, Polls prediction	SISL	Label proportions	Unsupervision
Aggregate outputs [25]					
Candidate labeling sets [22]	Learning with groups of examples and sets of possible categorizing vectors	Classifying photographs with captions	SISL	Candidate labeling vectors	Unsupervision
Learning from crowds [30,40]	Learning with examples categorized with many candidate noisy categorizations	Image annotation	SISL	Crowd	Unsupervision
Multi-label [34]	Learning with examples that belong to several categorizations at the same time	Film genre prediction	SIML	Full-supervision	Unsupervision
Semi-supervised multi-label [6]	Learning with examples categorized with multiple labels or uncategorized	Text categorization	SIML	Semi-supervision	Unsupervision
ML with weak label [33]	Learning with examples categorized with a subset of the real multiple labels	Image annotation	SIML	Incomplete	Unsupervision
ML incomplete class [3]					
Set classification [26]	Prediction for a group of examples, all of them with the same categorization	Face recognition with multiple photos	SIML	Full-supervision	Mutual label constraints
MIL [9]	Learning with multiple-instance examples that are positive if at least one of their instances is	Molecule activation prediction	MISL	Full-supervision	Unsupervision
G-MIL [39]	Learning with examples represented by several instances with generalized function for positives	Key-and-lock prediction problem	MISL	Full-supervision	Unsupervision
MISSL [29]	Learning with categorized and uncategorized multiple-instances examples	Content-based image retrieval	MISL	Semi-supervision	Unsupervision
MIML [43]	Learning with examples represented with several instances that belong to several categorizations	Classifying texts, images or videos	MIML	Full-supervision	Unsupervision
SSMIML [41]	Learning with multiple-instance examples categorized with multiple labels or uncategorized	Video annotation	MIML	Semi-supervision	Unsupervision
MIML with weak labels [42]	Learning with multiple-instance examples categorized with a subset of the real multiple labels	Image annotation	MIML	Incomplete	Unsupervision

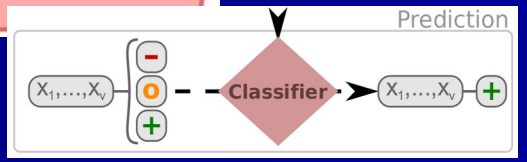
TAXONOMY OF WEAKLY SUPERVISED SCENARIOS





		Categorization	
		SL	ML
Example	SI	$H : \mathcal{X} \to \mathcal{C}$	$H : \mathcal{X} \to 2^{\mathcal{C}}$
	MI	$H : 2^{\mathcal{X}} \to \mathcal{C}$	$H : 2^{\mathcal{X}} \to 2^{\mathcal{C}}$





ALGORITHMS FOR WEAKLY SUPERVISED LEARNING

Document Classification Using Expectation Maximization with Semi Supervised Learning

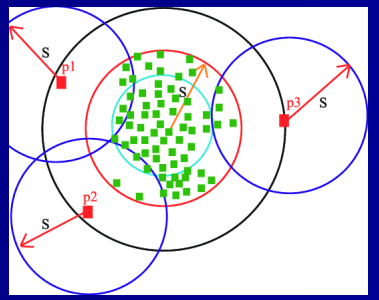
Bhawna Nigam[1], Poorvi Ahirwal[2], Sonal Salve[3], Swati Vamney[4]

1. Start from MLE $\theta = \{w, \mu, \Sigma\}_{1:2}$ on (X_l, Y_l),
   - w_c=proportion of class c
   - μ_c=sample mean of class c
   - Σ_c=sample cov of class c

   repeat:
2. The E-step: compute the expected label $p(y|x,\theta) = \frac{p(x,y|\theta)}{\sum_{y'} p(x,y'|\theta)}$ for all $x \in X_u$
   - label $p(y=1|x,\theta)$-fraction of x with class 1
   - label $p(y=2|x,\theta)$-fraction of x with class 2
3. The M-step: update MLE θ with (now labeled) X_u

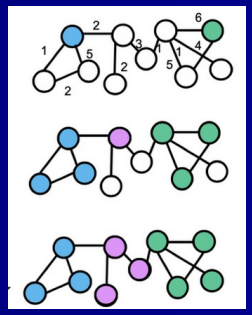
LOF: Identifying Density-Based Local Outliers

Markus M. Breunig†, Hans-Peter Kriegel†, Raymond T. Ng‡, Jörg Sander†



Label Propagation for Deep Semi-supervised Learning

Ahmet Iscen[1] Giorgos Tolias[1] Yannis Avrithis[2] Ondřej Chum[1]
[1]VRG, FEE, CTU in Prague [2]Univ Rennes, Inria, CNRS, IRISA



SOFTWARE LIBRARIES FOR WEAKLY SUPERVISED LEARNING

RSSL: Semi-supervised Learning in R

Jesse H. Krijthe[1,2]

[1] Pattern Recognition Laboratory, Delft University of Technology
[2] Department of Molecular Epidemiology, Leiden University Medical Center
jkrijthe@gmail.com

`sklearn.neighbors`.LocalOutlierFactor

AdaSampling

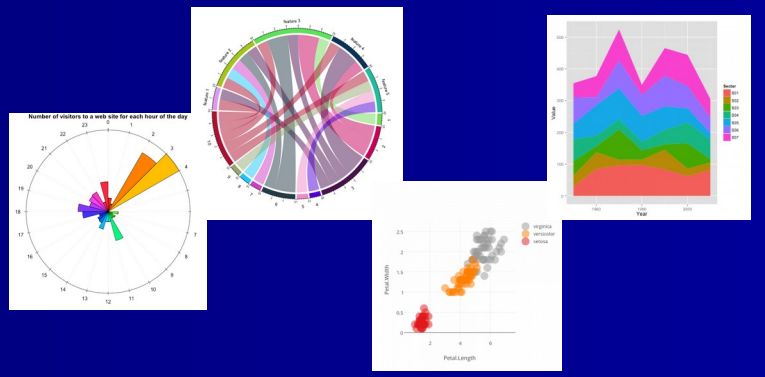
An R implementation of the AdaSampling algorithm for positive unlabeled and label noise learning

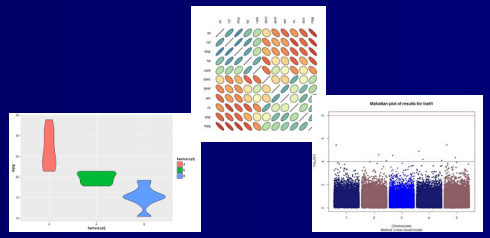
pickLabel: Pick the optimal label from candidate labels

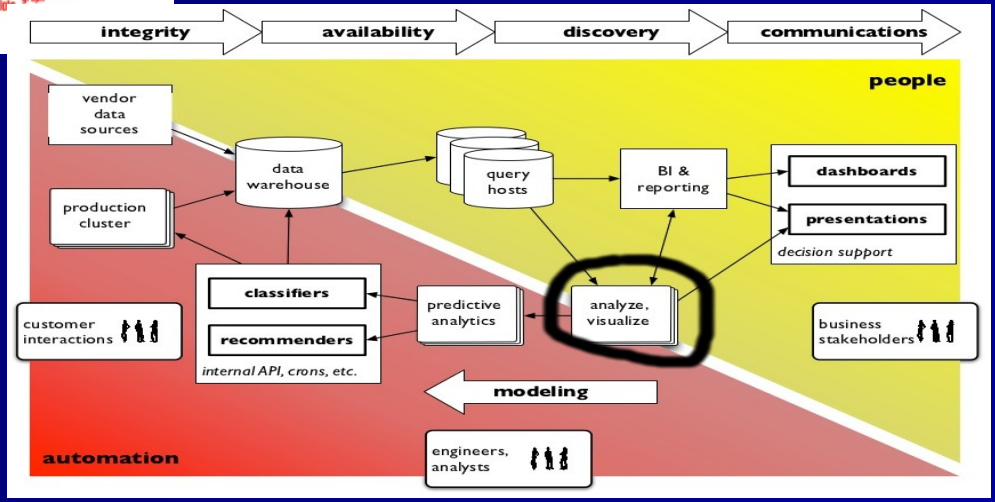
In Luwei-Ying/validateIt: Validating Topic Coherence and Topic Labels

`sklearn.svm`.OneClassSVM

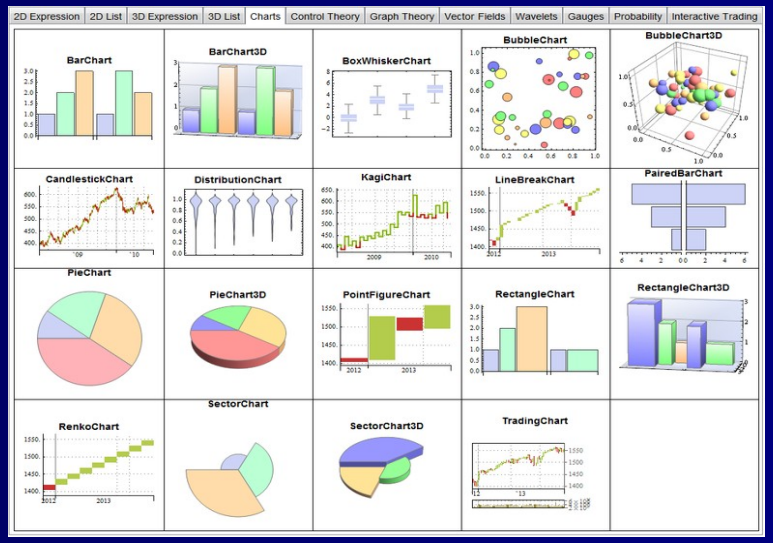
IS NEEDED TO LEARN "A CRYSTAL BALL"? OR JUST VISUALIZE MY DATA?

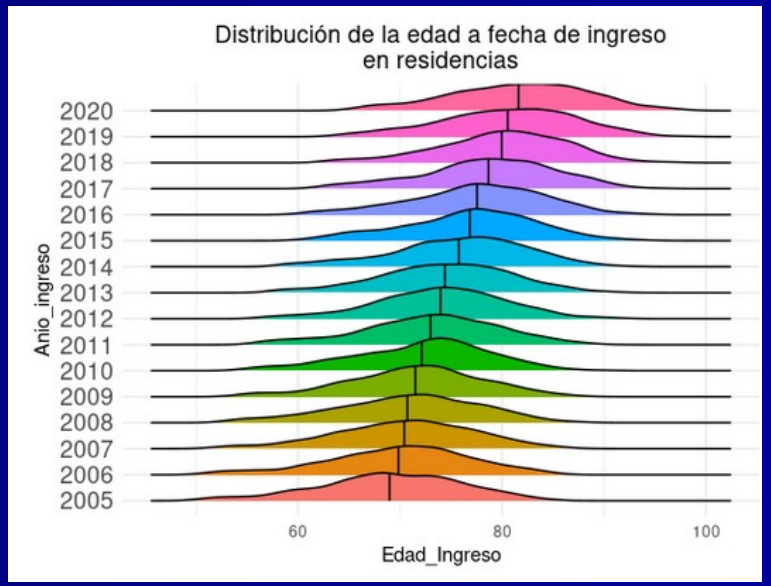






DATA VISUALIZATION
DATA EXPLORATION





- Top 10 data visualization tools
- List of commercial and free visualization tools
- 11 steps for data exploration in R: (with codes)
- The R graph gallery: a collection of 200 graphs in R
- 7 simple visualizations you should know in R

RESOURCES

SOFTWARE & DATASETS

UCI Machine Learning Repository

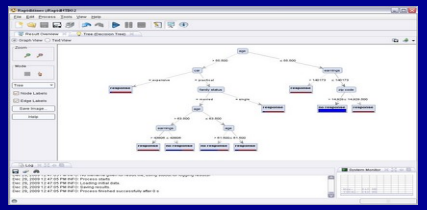
scikits learn machine learning in Python

kaggle

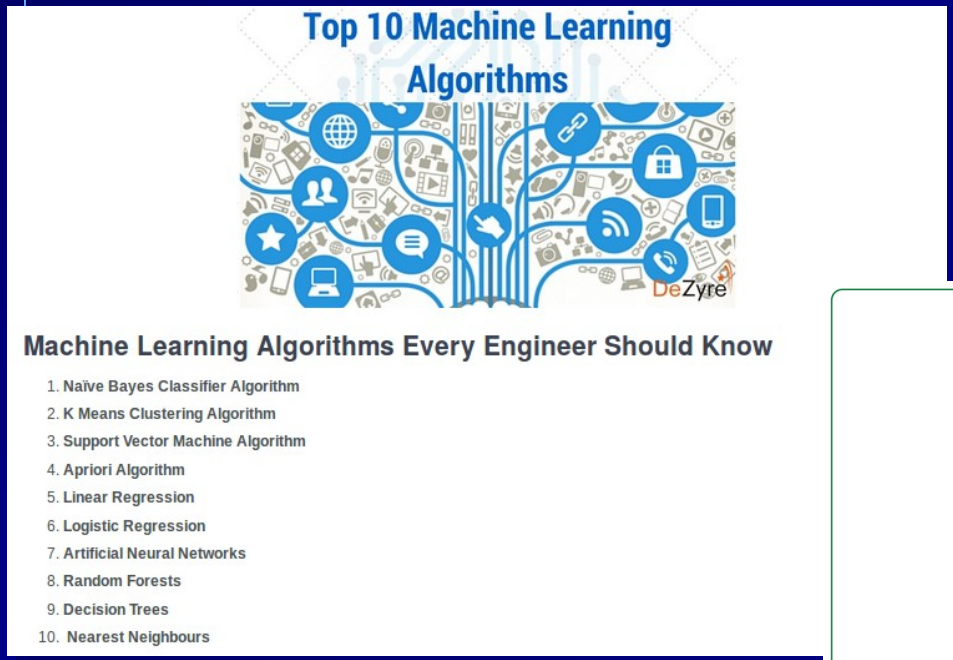
github SOCIAL CODING

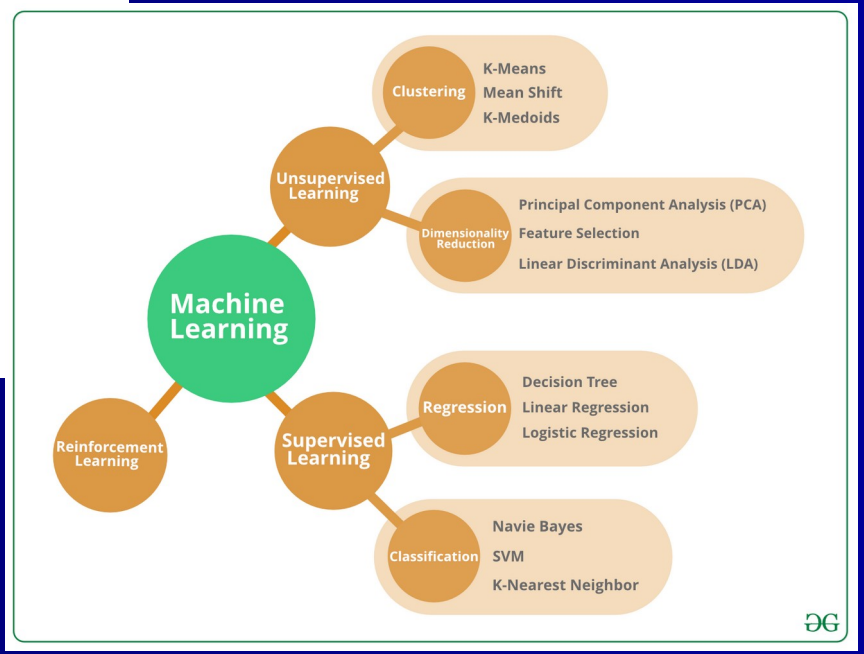
useR!

KDnuggets

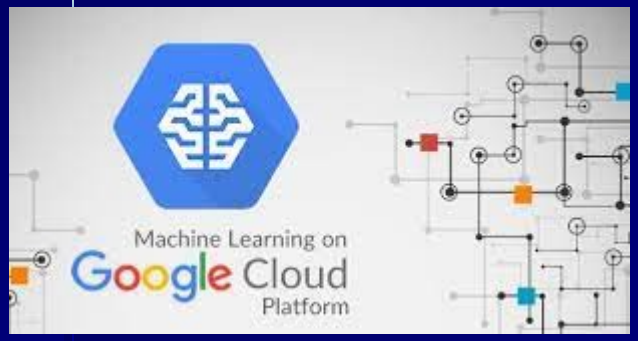


TOP-10 DATA MINING ALGORITHMS

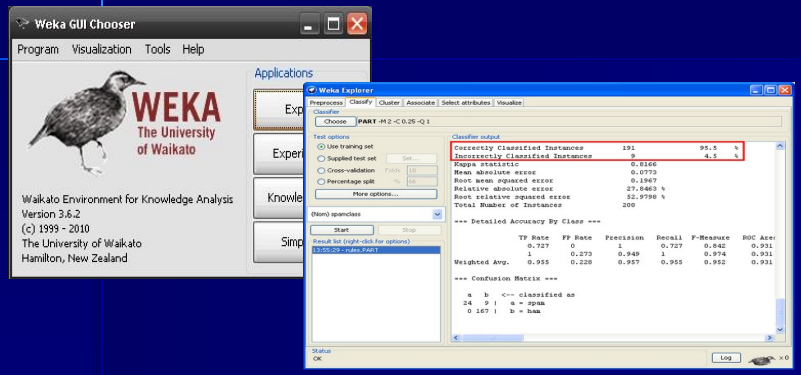


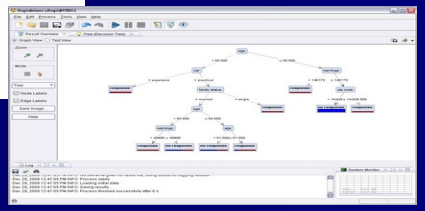


COMERCIAL SOFTWARE FOR DATA MINING – "as a service"



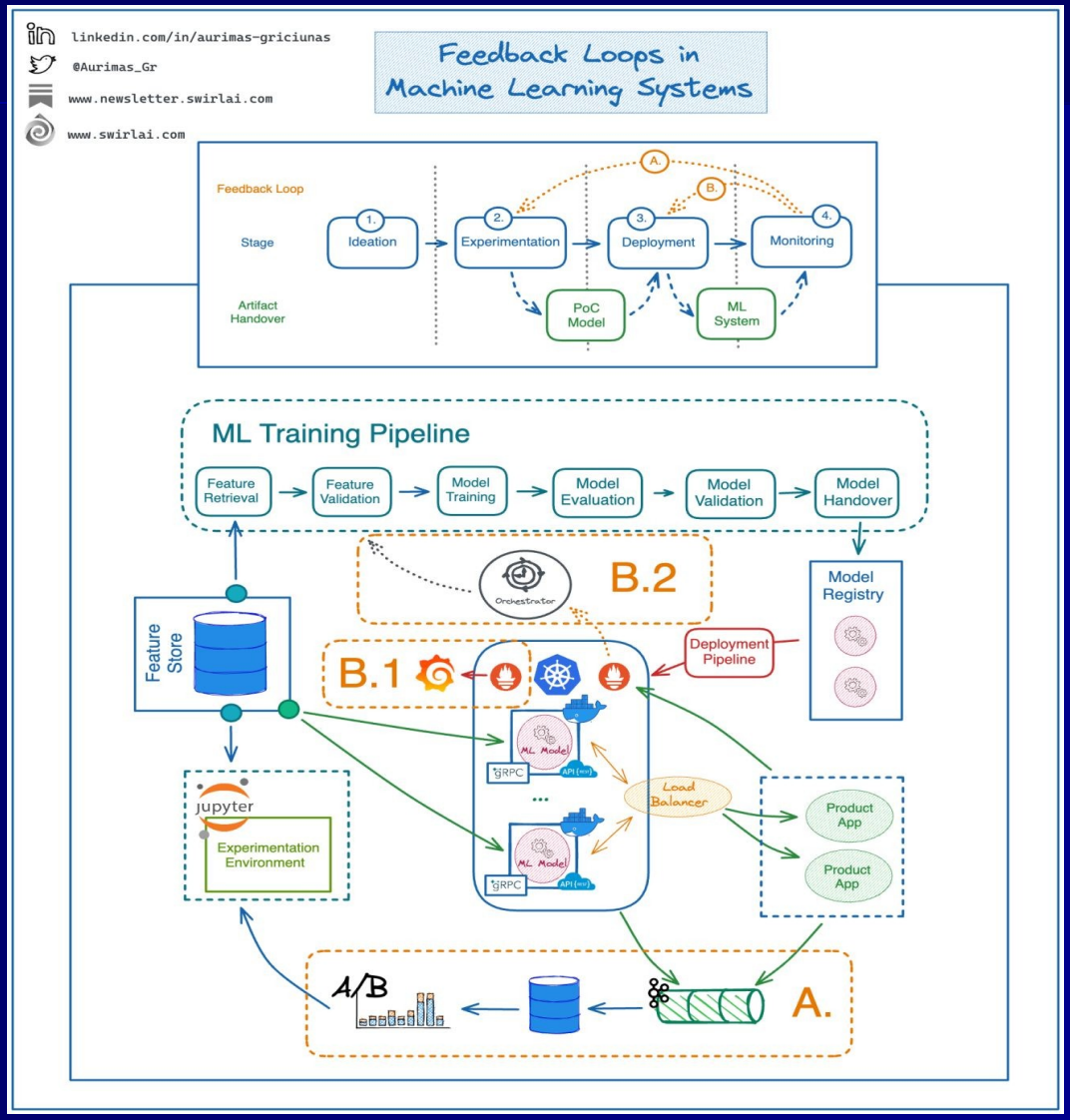
FREE SOFTWARE FOR DATA MINING





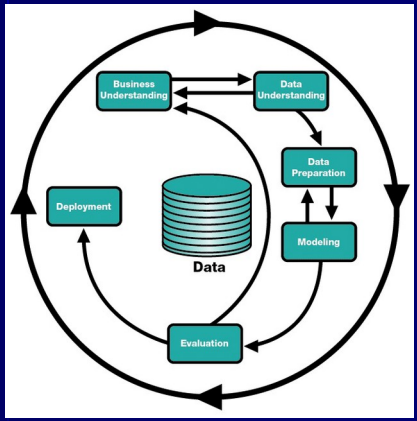
- Software suites for data mining, analytics and knowledge discovery
- 11 open source tools to make the most of machine learning
- Top 10 machine learning projects in GitHub
- 50 useful machine learning & prediction APIs
- Classification software: a list
- Top 15 frameworks for machine learning experts
- Bayesian networks and Bayesian classifier software

PROJECT MANAGEMENT



PROJECT MANAGEMENT
CRIPS-DM METHODOLOGY

- "Cross-Industry Standard Process for Data Mining" (CRISP-DM)
- kdnuggets.com Poll'2014: What methodology in DM projects?



What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

2014 poll 2007 poll

CRISP-DM (86)	43% / 42%
My own (55)	27.5% / 19%
SEMMA (17)	8.5% / 13%
Other, not domain-specific (16)	8% / 4%
KDD Process (15)	7.5% / 7.3%
My organizations' (7)	3.5% / 5.3%
A domain-specific methodology (4)	2% / 4.7%
None (0)	0% / 4.7%

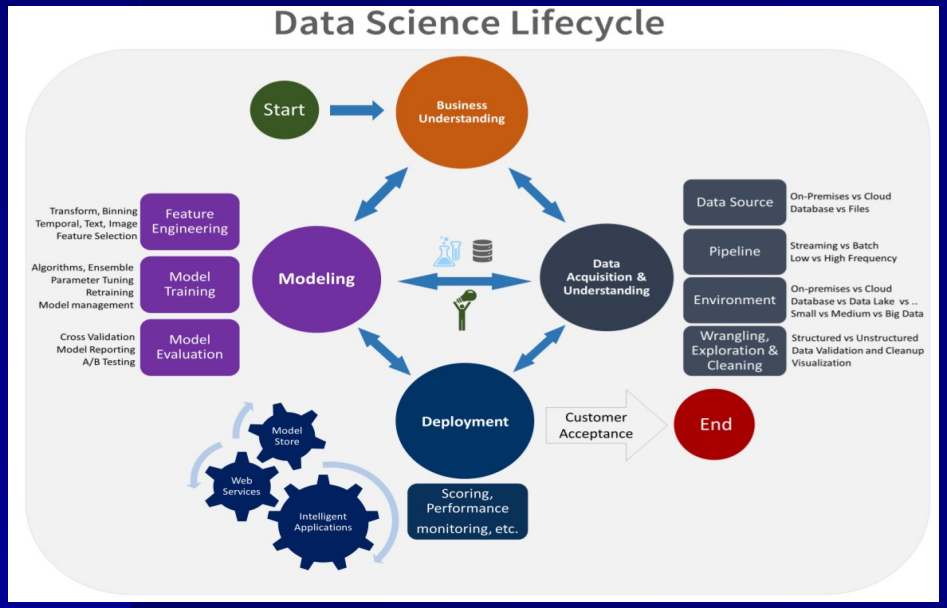
Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

- New items to be considered

- Infraestructure
- Staff training
- Outsourcing
- Contingencies plan and risks management
- Post-development:
  - - Installation
  - - Support
  - - Maintenance
  - - Discover defects
  - - User training

PROJECT MANAGEMENT TEAM DATA SCIENCE PROCESS

- A recent proposal by Microsoft
- It incorporates recent issues related to Big Data projects



Infrastructure and resources for data science projects

TDSP provides recommendations for managing shared analytics and storage infrastructure such as:

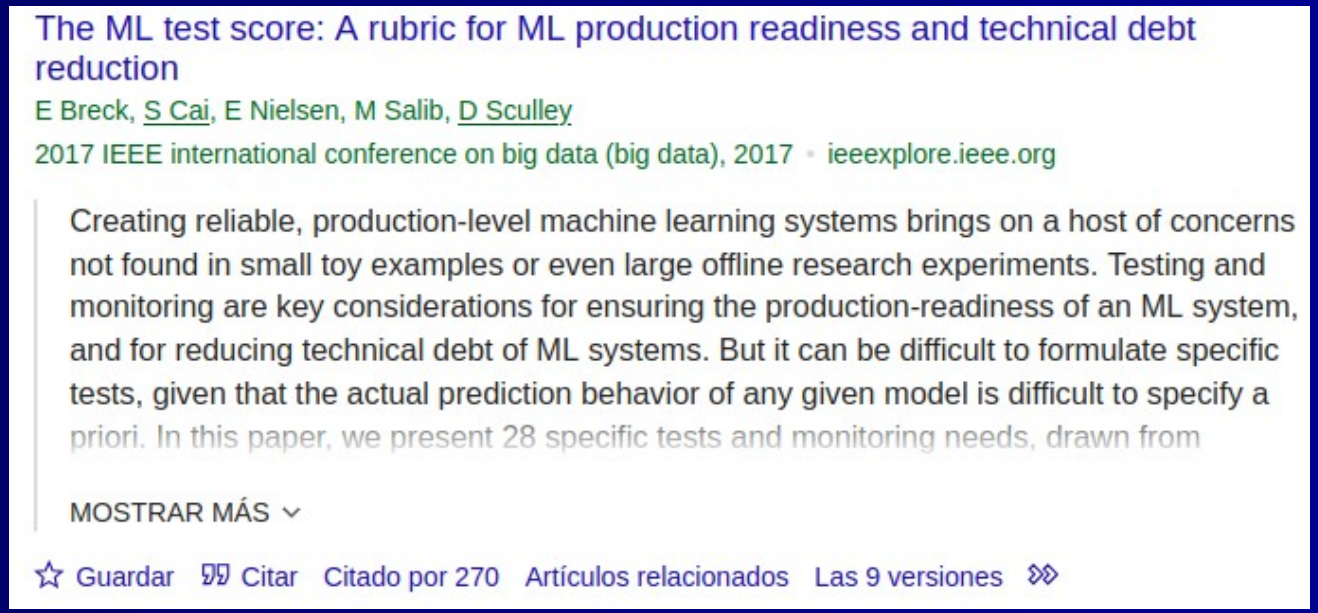
- cloud file systems for storing datasets,
- databases
- big data (Hadoop or Spark) clusters
- machine learning services.

Standardized project structure

- a project charter to document the business problem and scope of the project
- data reports to document the structure and statistics of the raw data
- model reports to document the derived features
- model performance metrics such as ROC curves or MSE

PROJECT MANAGEMENT
DEVOps – MLOps (by Google)

- "Good practices" → until product development
- 4 areas → 7 checkbox per area → final score



PROJECT MANAGEMENT
DEVOps – MLOps

1- Data and characteristics
2- Model Developing
3- Infraestructure
4- Monitoring

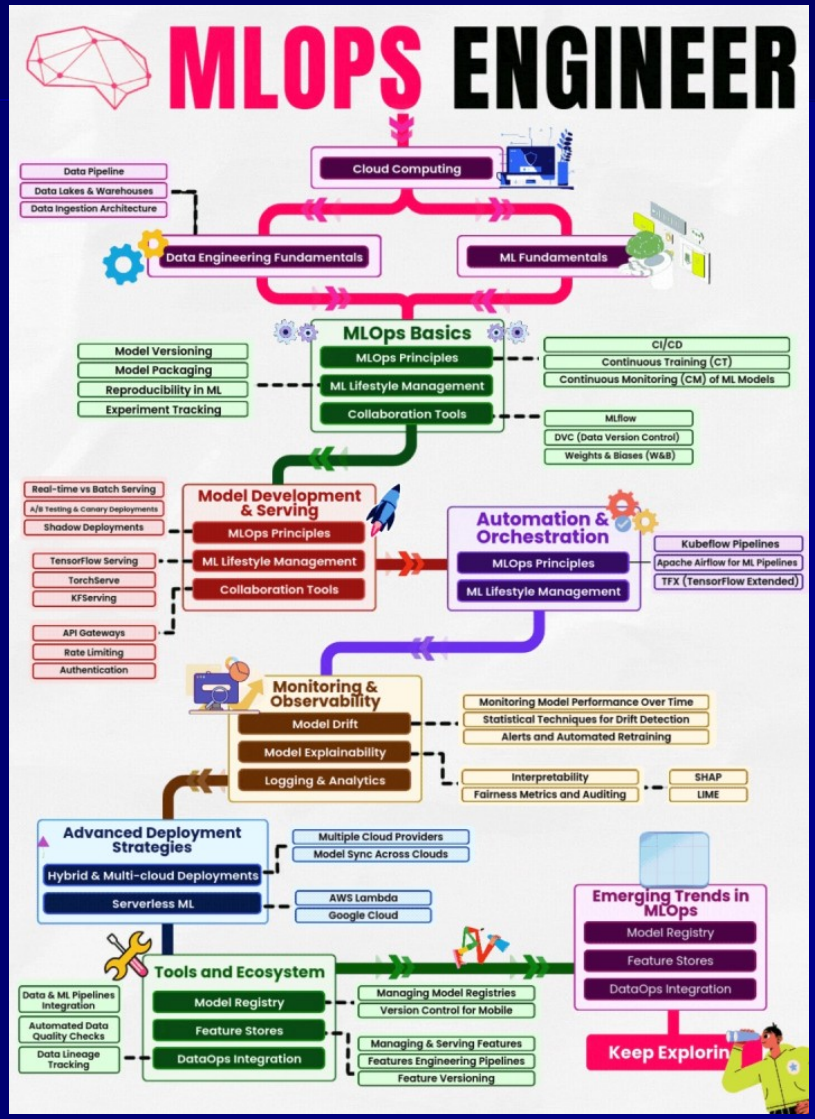
1	Feature expectations are captured in a schema.
2	All features are beneficial.
3	No feature's cost is too much.
4	Features adhere to meta-level requirements.
5	The data pipeline has appropriate privacy controls.
6	New features can be added quickly.
7	All input feature code is tested.

1	Model specs are reviewed and submitted.
2	Offline and online metrics correlate.
3	All hyperparameters have been tuned.
4	The impact of model staleness is known.
5	A simpler model is not better.
6	Model quality is sufficient on important data slices.
7	The model is tested for considerations of inclusion.

1	Training is reproducible.
2	Model specs are unit tested.
3	The ML pipeline is Integration tested.
4	Model quality is validated before serving.
5	The model is debuggable.
6	Models are canaried before serving.
7	Serving models can be rolled back.

1	Dependency changes result in notification.
2	Data invariants hold for inputs.
3	Training and serving are not skewed.
4	Models are not too stale.
5	Models are numerically stable.
6	Computing performance has not regressed.
7	Prediction quality has not regressed.

PROJECT MANAGEMENT
DEVOps – MLOps



PROJECT MANAGEMENT
DEVOps – MLOps