

Python en Químioinformática

curso

“Introducción a la Químioinformática”

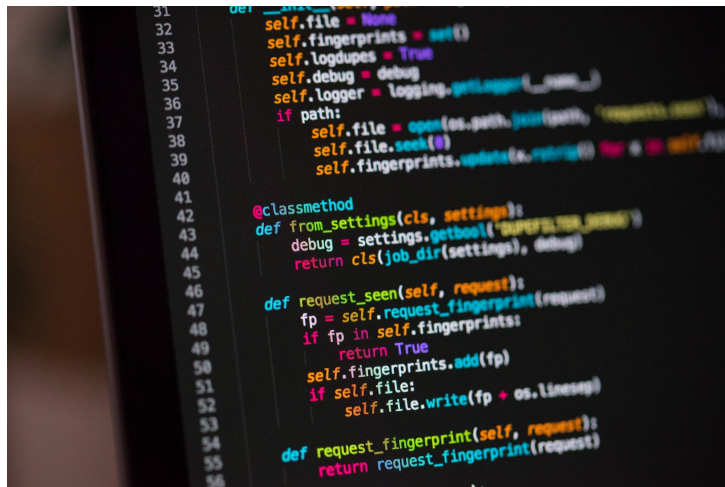
Bárbara I. Díaz Eufracio

- **Python (generalidades)**
- **Python y Servidores web
quimioinformáticos**
- **Python y aprendizaje
automático**

Outline

Programar

Escribir código que la computadora entienda.



```
31 def __init__(self, settings):
32     self.file = None
33     self.fingerprints = set()
34     self.logdupes = True
35     self.debug = debug
36     self.logger = logging.getLogger(__name__)
37     if path:
38         self.file = open(os.path.join(path, "requests.log"),
39                           "a")
40         self.file.seek(0)
41         self.fingerprints.update([x.strip() for x in self.file])
42
43 @classmethod
44 def from_settings(cls, settings):
45     debug = settings.getbool("DEBUG_LOG_REQUESTS")
46     return cls(job_dir(settings), debug)
47
48 def request_seen(self, request):
49     fp = self.request_fingerprint(request)
50     if fp in self.fingerprints:
51         return True
52     self.fingerprints.add(fp)
53     if self.file:
54         self.file.write(fp + os.linesep)
55
56 def request_fingerprint(self, request):
57     return request_fingerprint(request)
```

Código:

un texto escrito en lenguaje formal que utilizamos para interactuar con las máquinas.

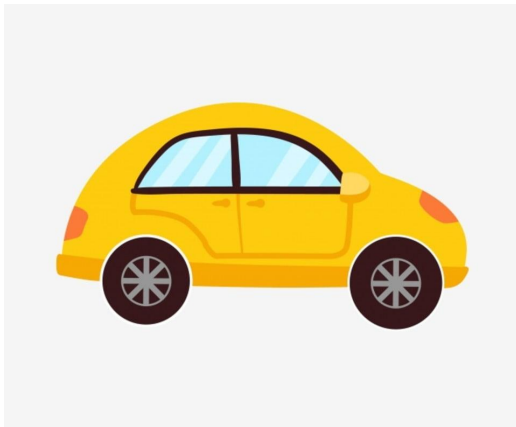
Programar

Escribir código que la computadora entienda. Utilizando un lenguaje formal y respetando la sintaxis del lenguaje de programación empleado.

Python

- Interpretado
- Tipado dinámico
- Fuertemente tipado
- Orientado o objetos

Objetos



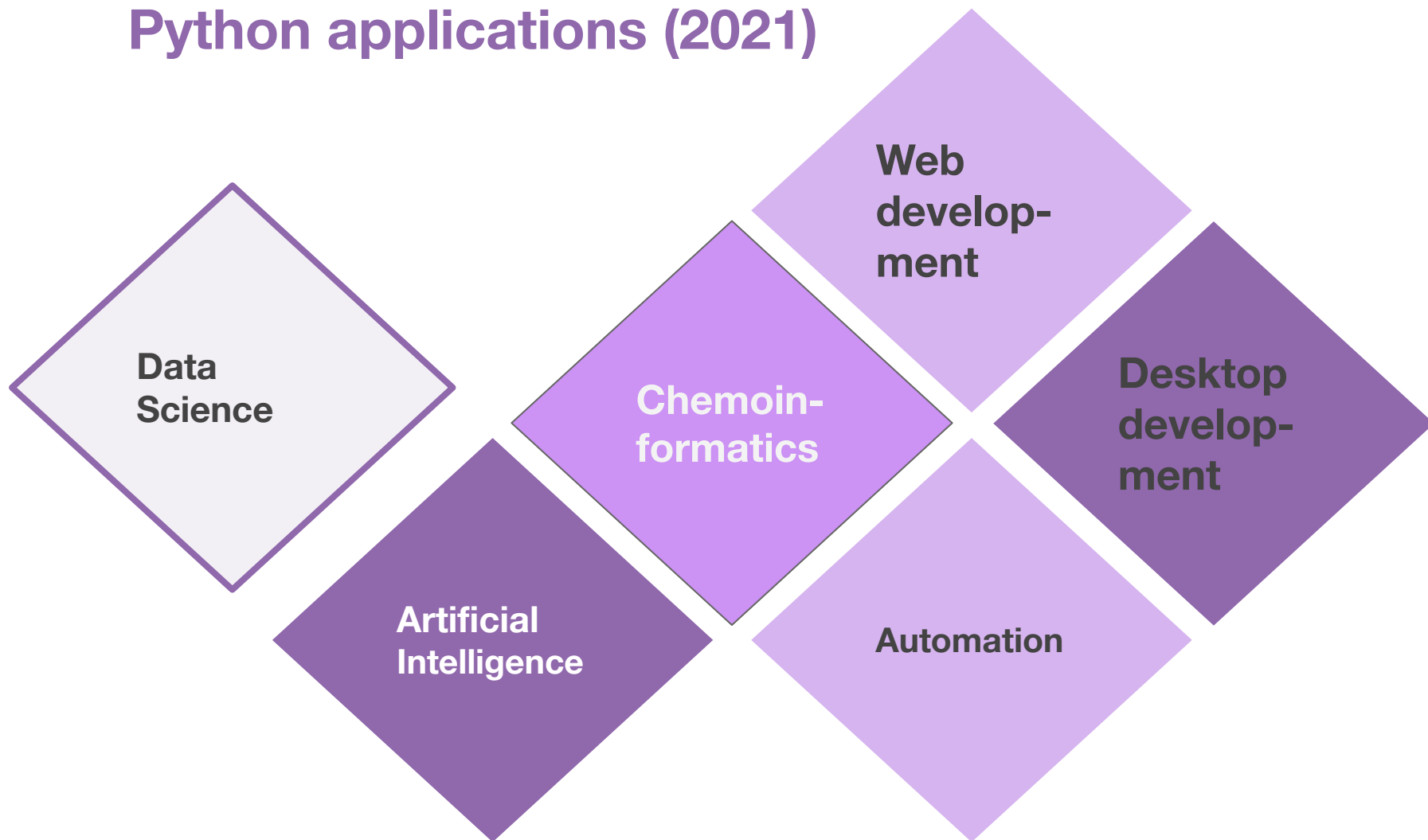
una entidad que agrupa un
estado y una **funcionalidad**

relacionada

Se definen a través
de variables
llamadas **atributos**

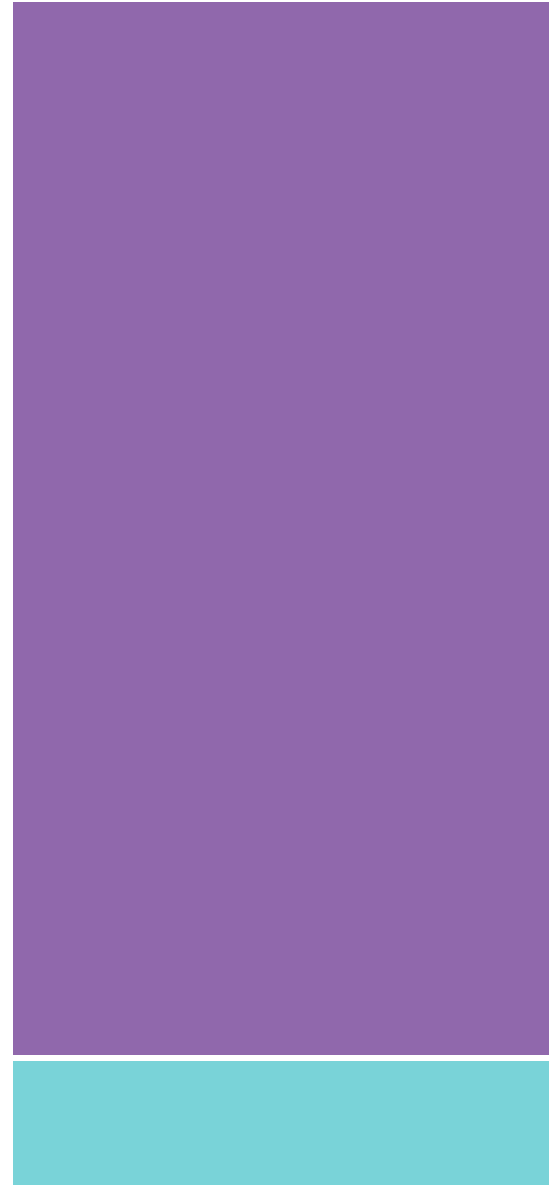
se moldea a través
de **funciones**

Python applications (2021)





Python y Servidores web quimioinformáticos



Discipline that employs **equations, models** and **computational techniques**.

Apply chemical tools to solve chemical problems.

Chemoinformatics

Developed at industry to handle databases and chemical **structure representations**.



Gasteiger J. The central role of chemoinformatics. *Chemometr Intell Lab Syst* **2006**, 82:200–209

Chemoinformatic

Disciplina que conjunta la ciencia de la computación y la química para resolver problemas

1

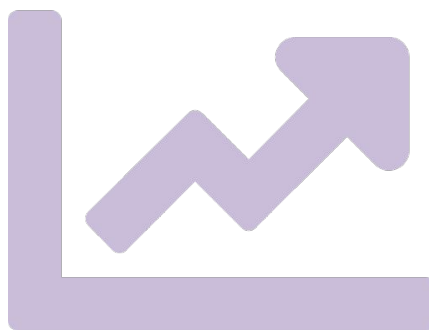
Extracción

2

Procesamiento

3

Extrapolación



Obtener información
valiosa a partir de
estructuras químicas

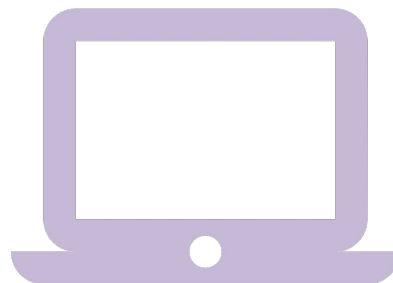
Servidores Web

El objetivo es:

Promover la disponibilidad de herramientas e información de acceso libre para la investigación.



Aumento en el número de servidores web



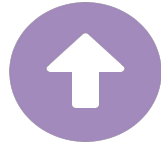
¿Todos podemos desarrollar herramientas web?

Allarakhia, M. Expert Opin. Drug Discov. **2014**, 9, 459–465.

González-Medina, M.; Naveja, J. J.; Sánchez-Cruz, N.; Medina-Franco, J. L. RSC Adv. **2017**, 7, 54153–54163.

Web Servers

Ventajas

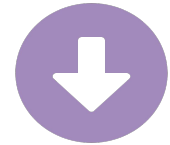


Open source.

Gratuitos.

**Usuarios sin
conocimiento de
programación.**

Desventajas



“Black Boxes”

**Habilidades de
programación y
desarrollo.**

Herramientas

Programming languages



Python



R

Frameworks

django

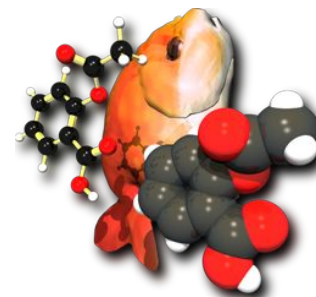


Shiny

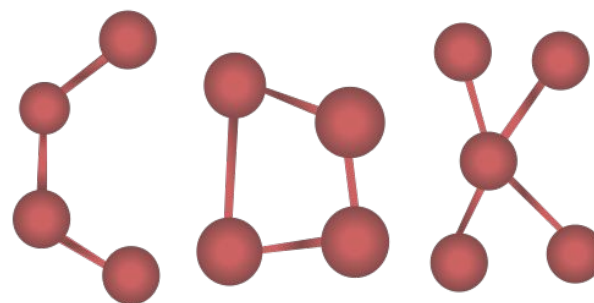
Cheminformatics toolkits



Open-Source Cheminformatics
and Machine Learning



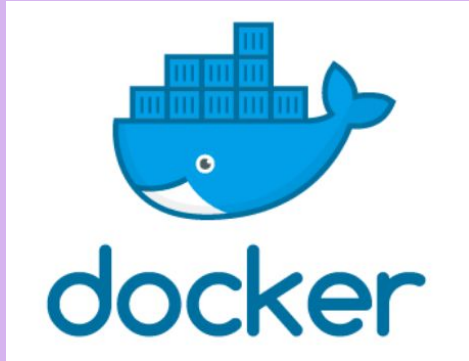
Open Babel



Chemistry Development Kit
(CDK)

Herramientas

Contenedores



Frontend



Databases



SQL

Bases de Datos

- Información específica
- Public

PubChem

Extensive records of **compounds**, bioactivity, assays and targets.

pubchem.ncbi.nlm.nih.gov

DrugBank

a curated **pharmaceutical knowledge base**.

drugbank.com

ZINC15

Free database of **commercially-available** compounds.

zinc-docking.org

ChEMBL

1.1 million **Protein - Ligand** complexes.

www.ebi.ac.uk/chembl

D-TOOLS

www.difacquim.com/d-databases

Developed at
DIFACQUIM,
UNAM



D-DATABASE (D-DB)

Curated high quality data on target annotations of small molecules.



Epigenomics chemical database

54 dianas epigenéticas.



Biofacquim

Productos naturales caracterizados en México.

Chemoinformatic Tools

Chemical space

PCA physico-chemical properties

Diversity
Tanimoto similarity

PUMA⁽¹⁾

Analyze **structure–activity** relationships of compound data sets.

Activity Landscape Plotter⁽²⁾

Represent in low dimensions the **diversity of chemical** libraries. Simultaneously multiple molecular representations.

Consensus Diversity Plots⁽³⁾

www.difacquim.com/d-tools/

(1)González-Medina, M.; Medina-Franco, J. L. Platform for Unified Molecular Analysis: PUMA. *J. Chem. Inf. Model.* **2017**, 57 (8), 1735–1740.

(2) González-Medina, M.; Méndez-Lucio, O.; Medina-Franco, J. L. *J. Chem. Inf. Model.* **2017**, 57, 397.

(3) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.; Medina-Franco, J. L. Consensus Diversity Plots: A Global Diversity Analysis of Chemical Libraries. *J. Cheminform.* **2016**, 8, 63.

Chemoinformatic Tools

This website allows you to estimate the bioactivity profile of a small molecule over a panel of 55 human epigenetic targets.

**Epigenetic Target
Profiler**

Peptide libraries
numeration with different
topologies
Diversity Analysis
Chemical Space
Visualization (PCA, tSNE)

Peptide Builder

www.difacquim.com/d-tools/

How much python is needed to learn django?

Variables, Basic Data
Types, Operators
Control Flow: Loops,
Conditionals,
Comprehensions
Basic Boolean Expressions
Conditionals
Loops
Comprehensions
Functions, Arguments
Classes
Packages, modules, pip
Exception handling

Other Skills Besides Python

Basic Command Line Skills
Css, Html
Client-Server Model, Frontend
vs. Backend Code
Basics of Http

Is It Possible to Learn Django Without Knowing Python?

Is It Worth The Effort?

Why is important write code?

Resolve a particularly task

Generate a tool te resolve a similar task in the future
(Automatation)

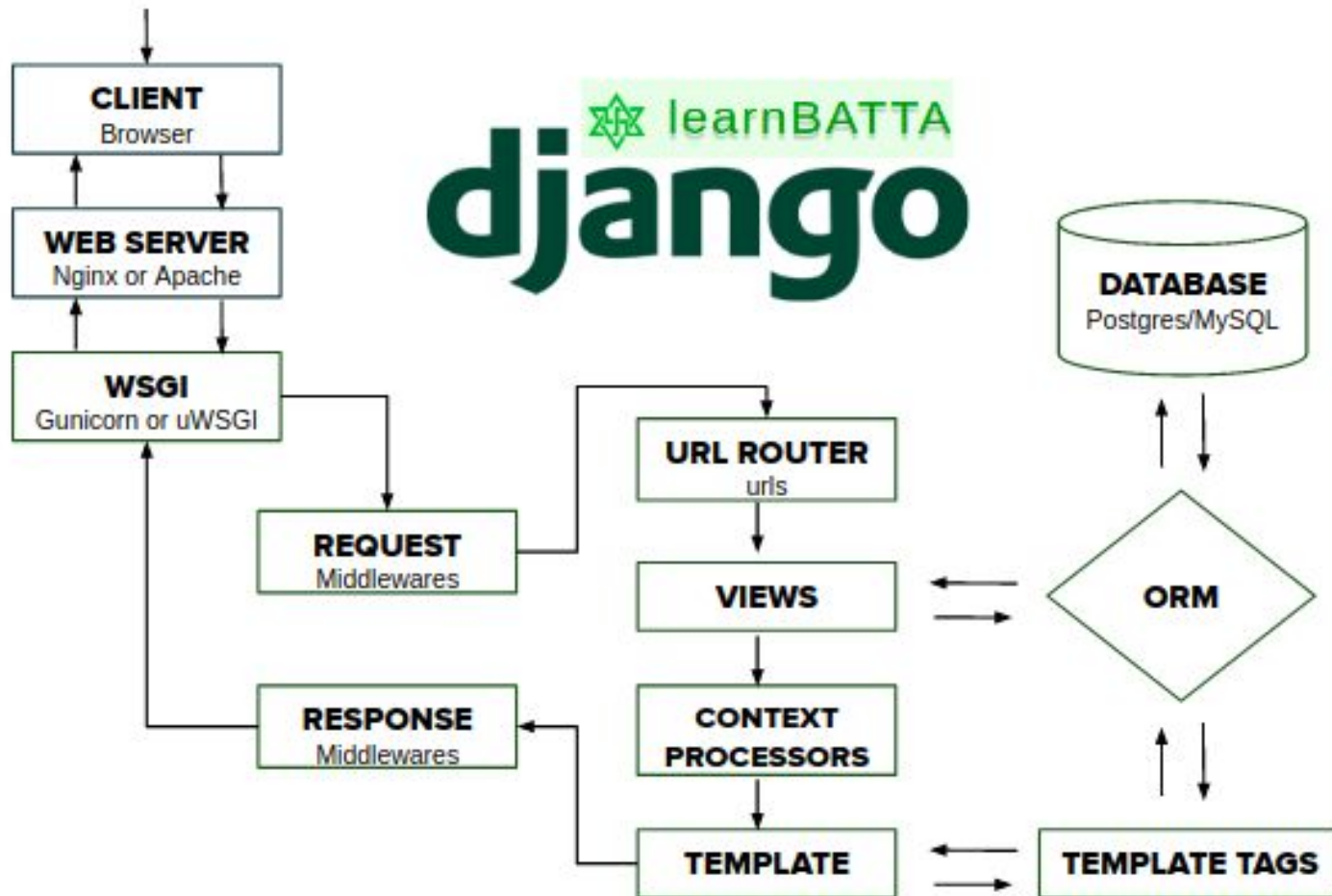
Why is important document my code?

Make code useful for other programmers

Some Dayli tasks on chemoinformatic performance with python:

- **Descriptor calculation**
- **Diversity analysis**
- **Statistical Analysis**
- **Exploratory data analysis**
- **ML model development**
- **Data visualization**
 - **Chemical Space exploration**
 - **Descriptor Analysis**
 - **Trajectory analysis (MD)**

Request response lifecycle in Django



slido.com with #689775

slido



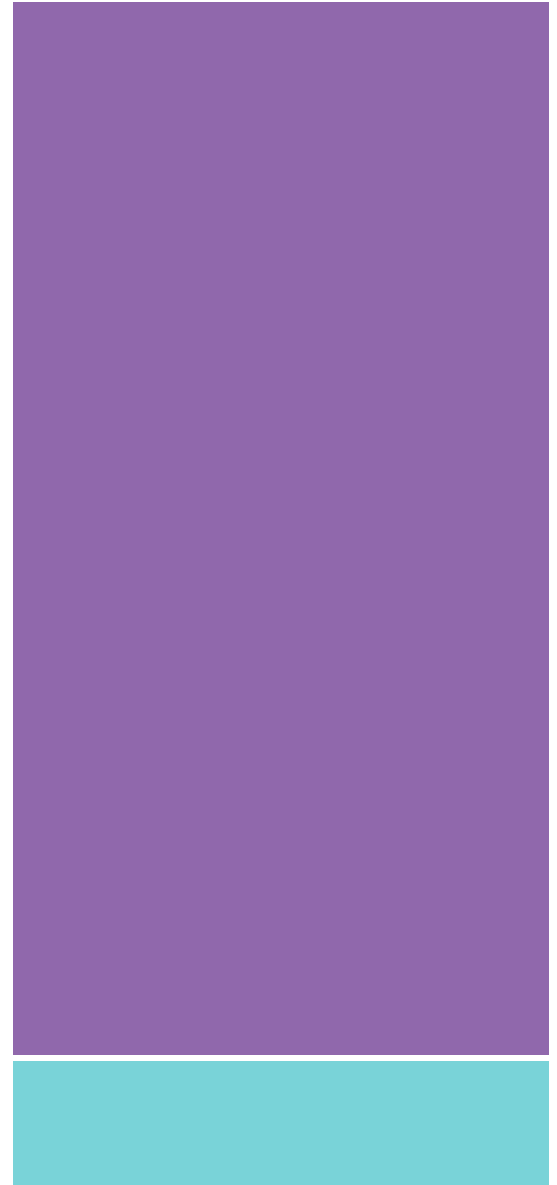
Herramientas utiles para desarrollar servidores web

① Start presenting to display the poll results on this slide.

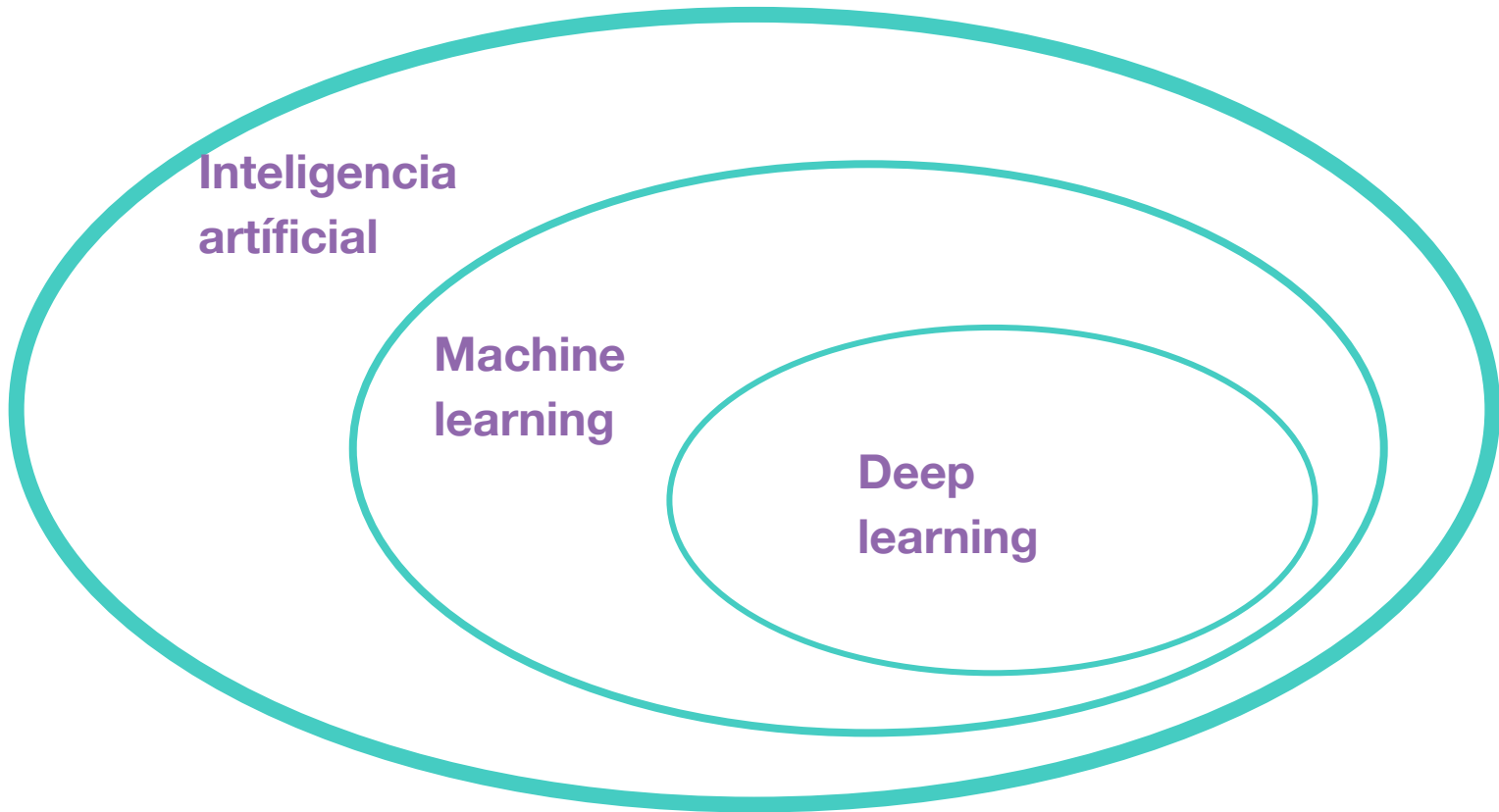


**Python y
automático**

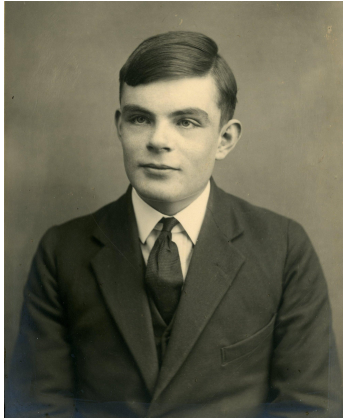
aprendizaje



¿Cómo se relacionan?



Machine learning (ML)



(Aprendizaje automático)

“Computing Machinery and Intelligence”

- *prueba de Turing*
- concepto clave para sentar las bases de AI

Nuevo paradigma de programación

Reglas
Datos

Programación
clásica

Respuestas

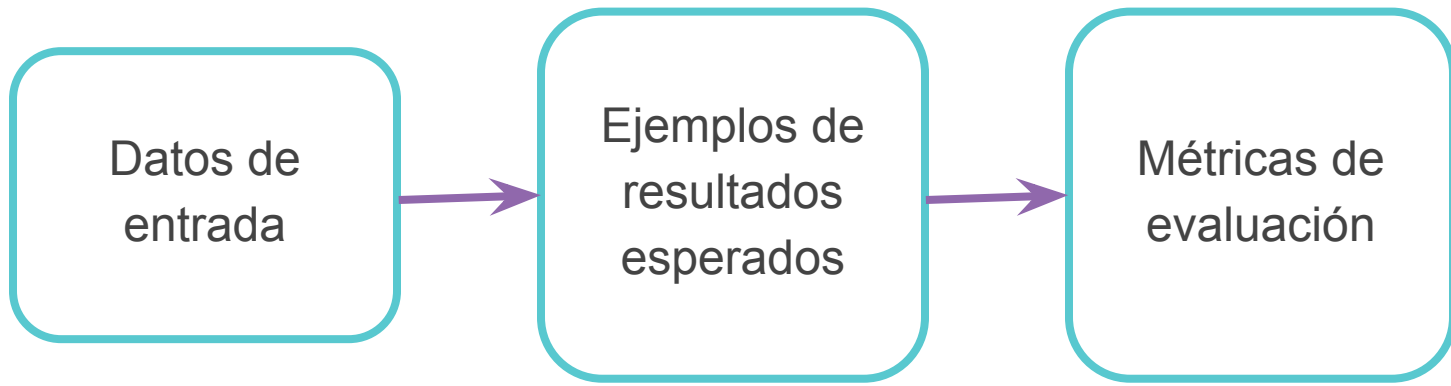
Datos
Respuestas

Machine
Learning

Reglas

Un sistema de
“aprendizaje
automático” es
entrenado en lugar de
programado

¿Qué necesitamos para entrenar modelos de ML?

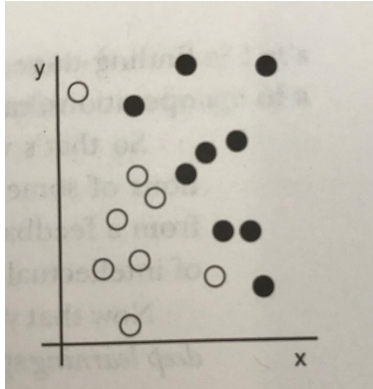


¿Cómo funciona el ML?

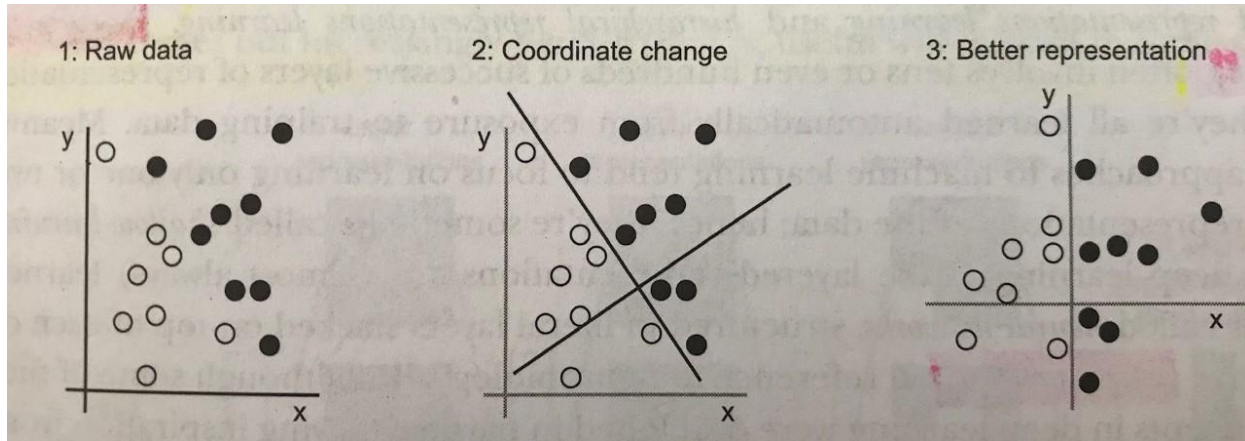


Transformar los datos de entrada en resultados (outputs) significativos

Ejemplo de algunos datos



Cambio de coordenadas



Herramientas que facilitan implementación de ML

- Hardware
- Bases de datos
- Librerías que permiten implementar algoritmos

Democratización de ML

- Simplicity
- Scalability
- Versatility and reusability

Se necesita un nivel intermedio de conocimientos de programación en python para entrenar modelos de ML

Librerías de python para AI

ML



scikit-learn.org

DL



tensorflow.org



pytorch.org

Clasificación de los sistemas de ML

Aprendizaje Supervisedo

Etiquetas y dianas **son conocidas**

- Regresión
- Clasificación

Aprendizaje Supervisedo

NO

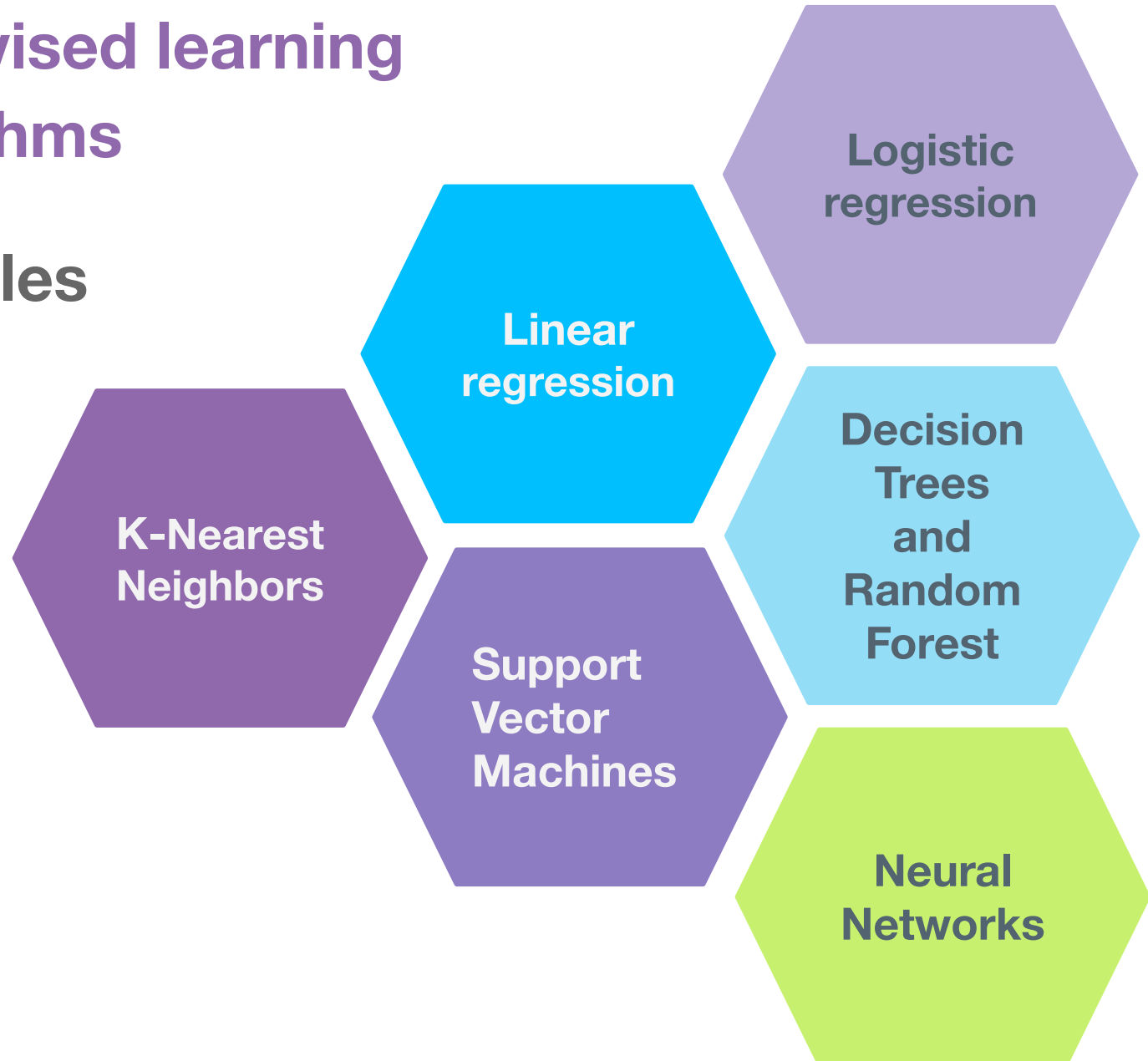
Una vez entrenado, el modelo **predice las etiquetas.**

Etiquetas no conocidas

- Clustering
- Visualization and dimensionality reduction
- Association rule learning

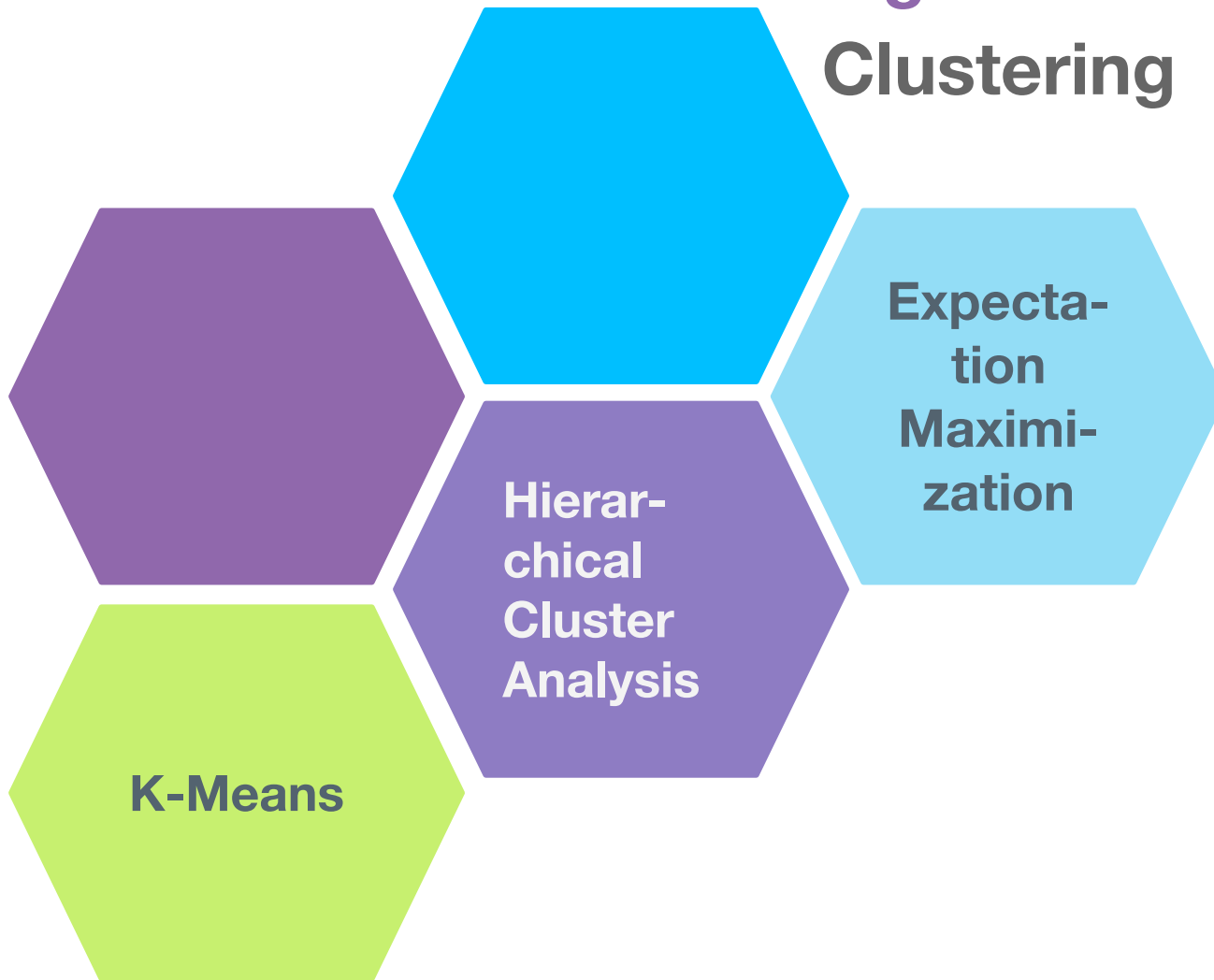
Supervised learning algorithms

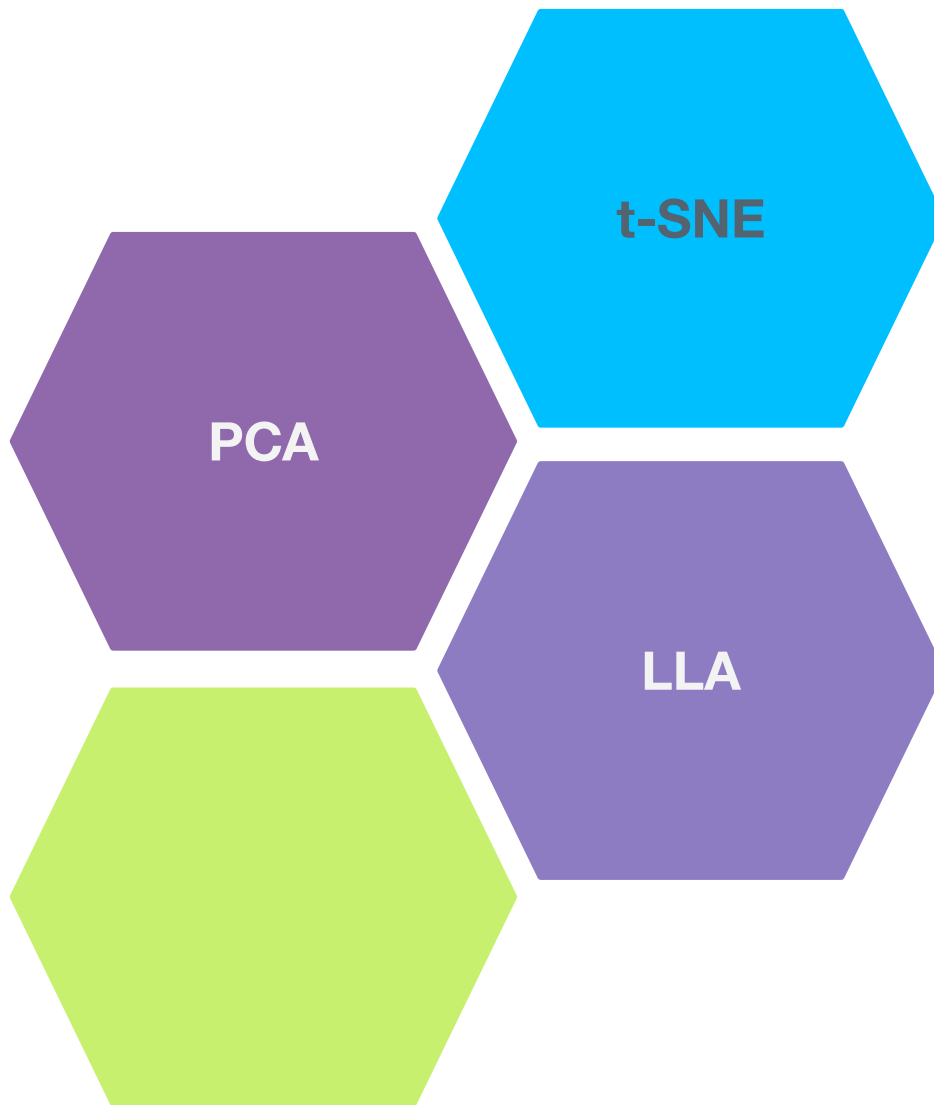
examples



Unsupervised learning algorithms (1)

Clustering





Unsupervised learning algorithms (2)

Visualization and dimensionality reduction

t-SNE, t-distributed Stochastic Neighbor Embedding

PCA, Principal Component Analysis

LLA, Locally-Linear Embedding

**Diferentes
algoritmos de ML
para diferentes
datos**

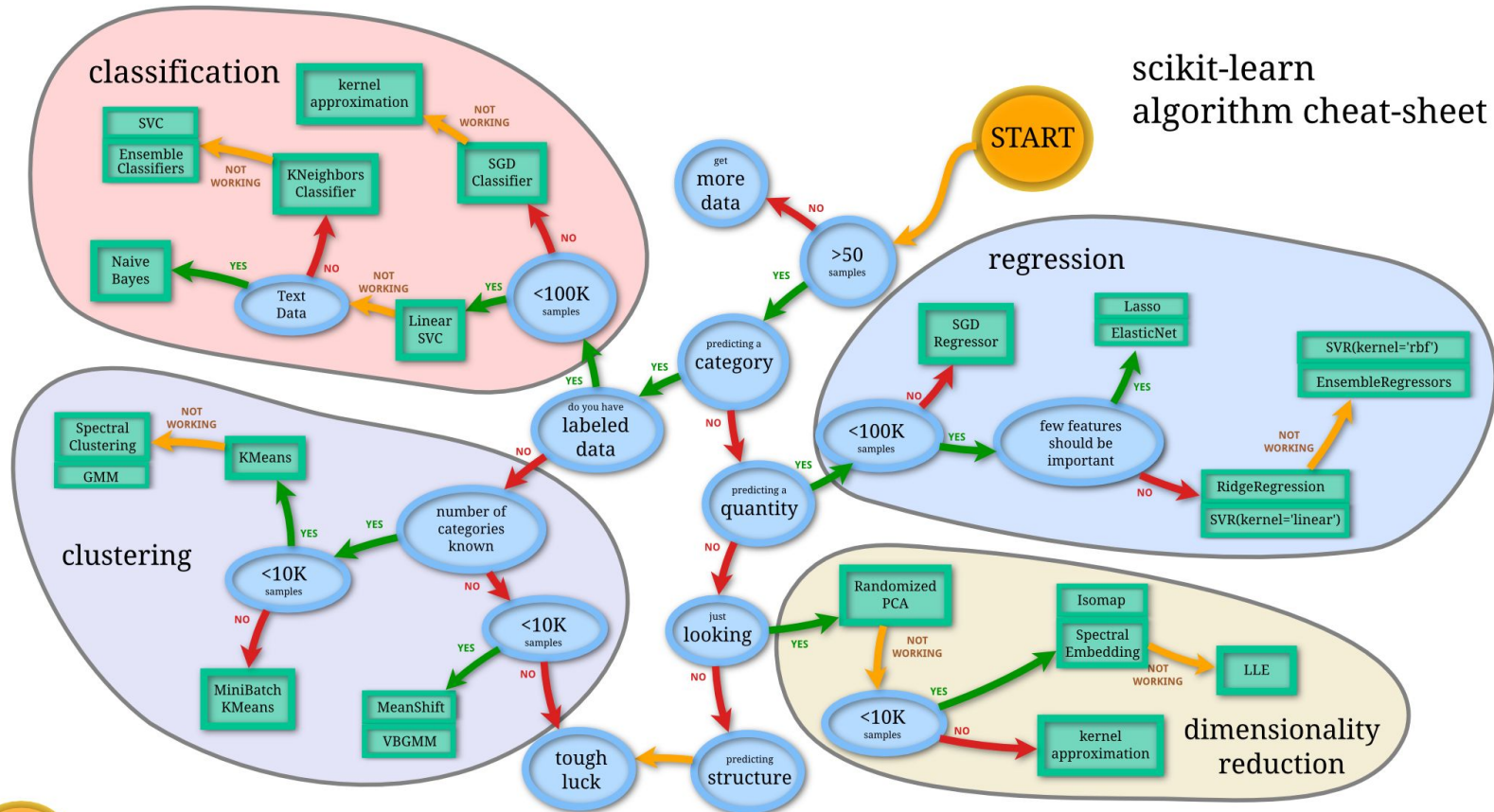
¿Cómo elegir el algoritmo
adecuado?

Depende

- ¿Qué **estamos buscando?**
- ¿Qué **información tenemos a nuestra disposición?**

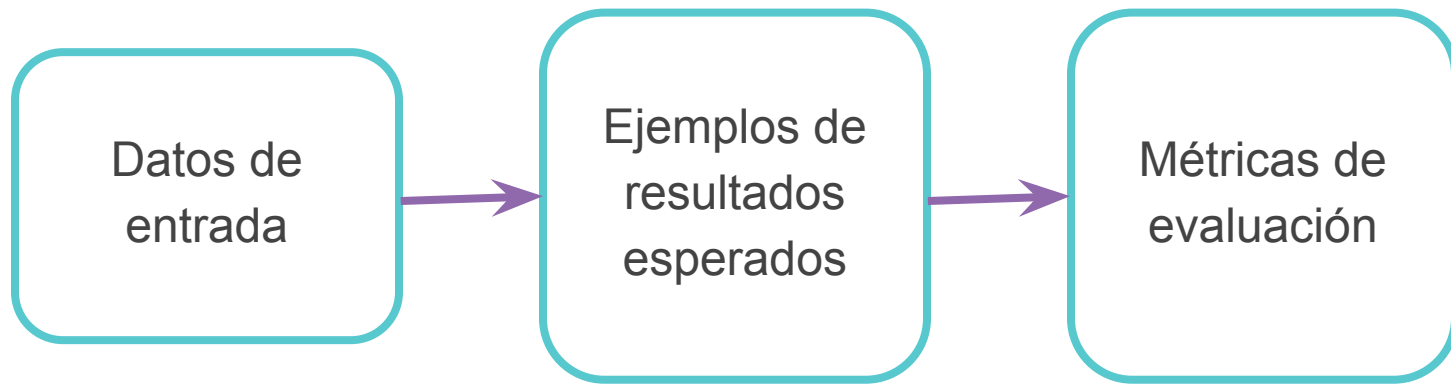
ML Roadmap

scikit-learn
algorithm cheat-sheet



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Métricas



Diferentes algoritmos para diferentes respuestas

¿Las métricas de evaluación son iguales para todos los algoritmos?

No

Métricas

Regresión

- Mean Absolute Error
- Mean Squared Error
- R2 Score

Clasificación

- Accuracy Score
- Recall
- Confusion Matrix
- F1

The increase of publication and study cases that employ machine learning to chemical analysis confirms the utility of this tool.

Machine learning in chemoinformatics

Efforts

- Industry
- Academia

Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Drug Discov. Today **2018**, 23, 1538–1546.

Machine learning in chemoinformatics and drug discovery

Use pattern recognition algorithms to discern mathematical relationships between empirical observations of small molecules and extrapolate them to predict chemical, biological and physical properties of novel compounds.

Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Drug Discov. Today **2018**, 23, 1538–1546.

QSAR and QSPR (One of the primary ML applications)

- Powerful tool on Drug Discovery
 - Biological Activity
 - ADME descriptors
 - Toxicity descriptors
 - Interactions

Necessary tools:

- Advanced chemoinformatics and machine learning techniques capable of modeling nonlinear datasets
- Big Data, and ML algorithms

QSAR, quantitative structure activity relationships

QSPR quantitative structure-property relationships

ML and Molecular Dynamics (MD)

MD simulations are an important tool for describing the evolution of a chemical system with time.

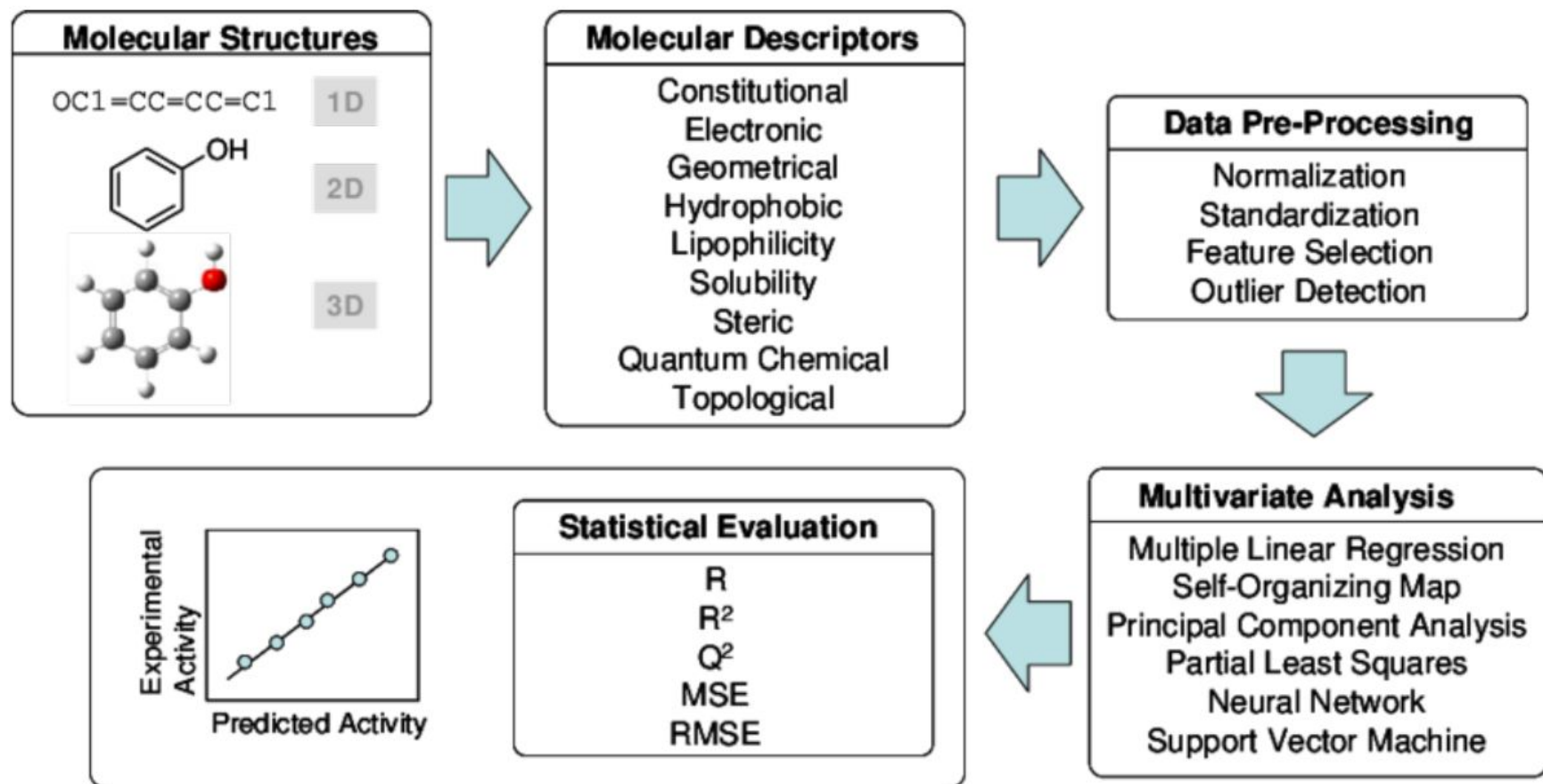
ML techniques can help to overcome computational limitations by providing access to

- potential energies
- forces
- molecular properties modeled directly after an accurate electronic structure reference

At only a fraction of the original computational cost.

Gastegger M., Marquetand P. (2020) Molecular Dynamics with Neural Network Potentials. In: Schütt K., Chmiela S., von Lilienfeld O., Tkatchenko A., Tsuda K., Müller KR. (eds) Machine Learning Meets Quantum Physics. Lecture Notes in Physics, vol 968. Springer, Cham. https://doi.org/10.1007/978-3-030-40245-7_12

QSAR Workflow



Casos de éxito (1)

Click2Drug | SwissDock | SwissParam | SwissSidechain | SwissBioStere | SwissTargetPrediction | **SwissADME** | SwissSimilarity | About us

SIB
Swiss Institute of Bioinformatics

SwissADME

Home | FAQ | Help | Disclaimer

This website allows you to compute physicochemical descriptors as well as to predict ADME parameters, pharmacokinetic properties, druglike nature and medicinal chemistry friendliness of one or multiple small molecules to support drug discovery.

The main article describing the web service and its underlying methodologies is **SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules**. *Sci. Rep.* (2017) 7:42717.
For details about development and validation of iLOG, please refer to this article: **iLOGP: a simple, robust, and efficient description of *n*-octanol/water partition coefficient for drug design using the GB/SA approach**. *J. Chem. Inf. Model.* (2014) 54(12):3284-3301.
For details about development and validation of the BOILED-Egg, please refer to this article: **A BOILED-Egg to predict gastrointestinal absorption and brain penetration of small molecules**. *ChemMedChem* (2016) 11(11):1117-1121.

Developed and maintained by the **Molecular Modeling Group** of the SIB | Swiss Institute of Bioinformatics.

Enter a list of SMILES here:
CC(=O)NC1=CC=C(O)C=C1

Fill with an example | Clear | Run!

Show BOILED-Egg

Retrieve data:

Powered by ChemAxon

Molecule 1

CC(=O)NC1=CC=C(O)C=C1

Physicochemical Properties

Formula	C8H9NO2
Molecular weight	151.16 g/mol
Num. heavy atoms	11
Num. arom. heavy atoms	6
Fraction Csp3	0.12
Num. rotatable bonds	2
Num. H-bond acceptors	2
Num. H-bond donors	2
Molar Refractivity	42.78
TPSA	49.33 Å²

Lipophilicity

Log <i>P</i> _{ow} (iLOGP)	1.21
Log <i>P</i> _{ow} (XLOGP3)	0.46
Log <i>P</i> _{ow} (WLOGP)	1.16
Log <i>P</i> _{ow} (MLOGP)	0.91
Log <i>P</i> _{ow} (SILICOS-IT)	0.89
Consensus Log <i>P</i> _{ow}	0.93

Water Solubility

Log S (ESOL)	-1.34
Solubility	6.93e+00 mg/ml ; 4.59e-02 mol/l
Class	Very soluble
Log S (All)	-1.06
Solubility	1.30e+01 mg/ml ; 8.62e-02 mol/l
Class	Very soluble
Log S (SILICOS-IT)	-2.19
Solubility	9.72e-01 mg/ml ; 6.43e-03 mol/l
Class	Soluble

Pharmacokinetics

GI absorption	High
BBB permeant	Yes
P-gp substrate	No
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	No
Log <i>K</i> _p (skin permeation)	-6.90 cm/s

Druglikeness

Lipinski	Yes; 0 violation
Ghose	No; 1 violation: MW<160
Veber	Yes
Egan	Yes
Muegge	No; 1 violation: MW<200
Bioavailability Score	0.55

Medicinal Chemistry

PAINS	0 alert
Brenk	1 alert: hydroquinone
Leadlikeness	No; 1 violation: MW<250
Synthetic accessibility	1.00

CYP1A2 inhibitor

Si o No

**Máquina de Soporte
Vectorial**

AUC 0.90

Water Solubility

Valor Continuo

Regresión líneal

Accuracy

SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. (2017) 7:42717.

Take home message

- Python permite resolver y automatizar un gran número de tareas en quimioinformática.
- En los últimos años el número de servidores y sus funcionalidades ha incrementado.
- El empleo de **algoritmos de aprendizaje automático** se ha consolidado como **una herramienta importante** en el diseño de fármacos y la quimioinformática.

Material disponible en:

<https://github.com/BarbaraDiazE/PythonEnQuimioinformatica>

Contact:



dieb@comunidad.unam.mx



@bdiazeufracio

