Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Predicting Drug Function Using Small-Molecule Structure Information

S. Ravichandran, Ph.D
BIDS, FNLCR

July 16, 2020

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Acknowledgements

- NCI-DOE Pilot, ATOM Teams
  - NCI-DOE: https://datascience.cancer.gov/collaborations/joint-design-advanced-computing
  - ATOM: https://pubmed.ncbi.nlm.nih.gov/32243153/

- BIDS
  - Drs. George Zaki, Andrew Weissman and Eric Stahlberg
  - Hue Reardon, Anney Che
  - Colleagues who reviewed the material

https://github.com/ravichas/ML-predict-drugclass

Introduction

- This is part of the NCI-DOE knowledge/capability transfer efforts
  - Knowledge-transfer of reproducible Machine-Learning frameworks for modeling drug-discovery problems
  - Reproducible Notebook that supports multiple languages

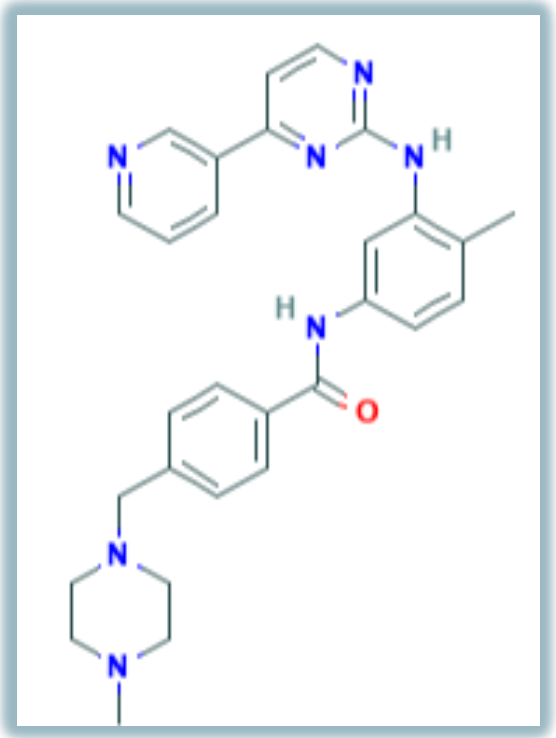
- We often take a test-case and provide tools (code/scripts) to accomplish the tasks.
  - We hope you can tune it to your needs
  - We believe the test-cases can be extended/modified to address advanced topics

- Please send us your feed-back
  - ravichandrans@mail.nih.gov

Objectives

- Identify drugs that belong to a Pharmacological Class using chemical properties (*in-silico*)

- Example
  - Given a drug (ex. Imatinib) in the form of chemical structure, can we predict, its function?



Antineoplastic Agents: Substances that inhibit or prevent the proliferation of NEOPLASMS.

CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5.CS(=O)(=O)O

Test case example

Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests

Jesse G. Meyer*, Shengchao Liu, Ian J. Miller, Joshua J. Coon and Anthony Gitter

Cite this: *J. Chem. Inf. Model.* 2019, 59, 10, 4438-4449
Publication Date: September 13, 2019
https://doi.org/10.1021/acs.jcim.9b00236

Why are we interested in this problem?

- Closely related to another area in Drug-Discovery called Drug repurposing

> "*Drug repurposing is a strategy for identifying new uses for approved or investigational drugs that are outside the scope of the original medical indication*" *Nat. Rev. 18, 41, 2019*

Drug-repurposing/Repositioning/Reprofiling/Re-tasking

"*Drug repurposing is a strategy for identifying new uses for approved or investigational drugs that are outside the scope of the original medical indication*"
Nat. Rev. 18, 41, 2019

Drug-repurposing/Repositioning/Reprofiling/Re-tasking

Zhou et al. *Cell Discovery* (2020)6:14
https://doi.org/10.1038/s41421-020-0153-3

Cell Discovery
www.nature.com/celldisc

ARTICLE Open Access

Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2

Yadi Zhou[1], Yuan Hou[1], Jiayu Shen[1], Yin Huang[1], William Martin[1] and Feixiong Cheng[1,2,3]

Drug repurposing: progress, challenges and recommendations

*Sudeep Pushpakom[1], Francesco Iorio[2], Patrick A. Eyers[3], K. Jane Escott[4], Shirley Hopper[5], Andrew Wells[6], Andrew Doig[7], Tim Guilliams[8], Joanna Latimer[9], Christine McNamee[1], Alan Norris[1], Philippe Sanseau[10], David Cavalla[11] and Munir Pirmohamed[1]**

NATURE REVIEWS | DRUG DISCOVERY VOLUME 18 | JANUARY 2019 | 41

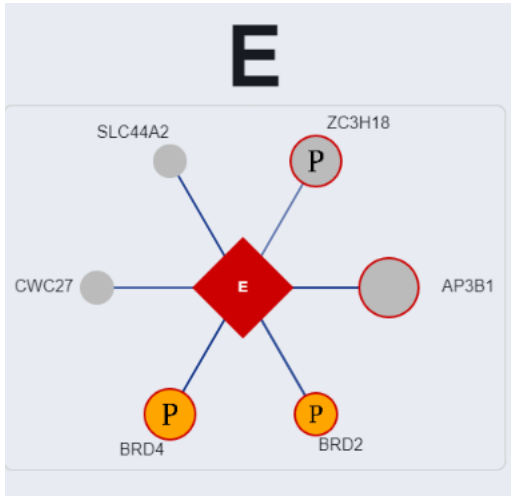
Drug-repurposing/Repositioning/Reprofiling/Re-tasking

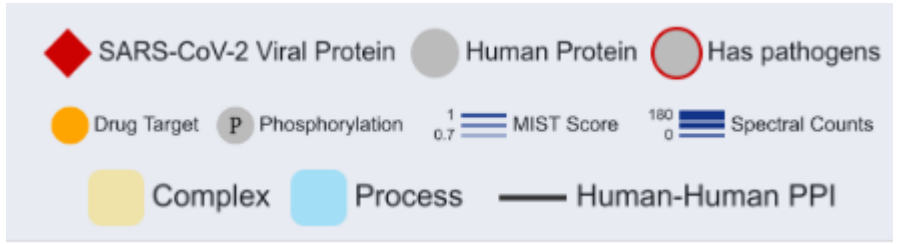
NATURE REVIEWS | DRUG DISCOVERY VOLUME 18 | JANUARY 2019 | 41

Drug Name	Original Indication	New Indication	Date of Approval	Repurposing approach used	Comments
Zidovudine	Cancer	HIV/AIDS	1987	In vitro screening of compound libraries	First anti-HIV drug to be approved by the FDA
Minoxidil	Hypertension	Hair-loss	1988	Retrospective clinical analysis (identification of hair growth as an adverse effect)	Global sale for minoxidil were US $860 million in 2016

Choice of host receptors gleaned via P-P network

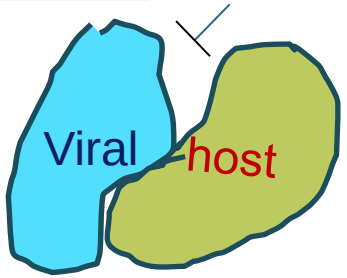
- D. E. Gordon *et al.,* https://www.nature.com/articles/s41586-020-2286-9 (2020).
  - "cloned, tagged and expressed 26 of the 29 viral proteins in human cells and identified the human proteins physically associated with each using Affinity-Purification Mass Spectrometry (AP-MS)"
  - we identify 67 druggable human proteins or host factors targeted by 69 existing FDA-approved drugs, drugs in clinical trials and/or preclinical compounds,





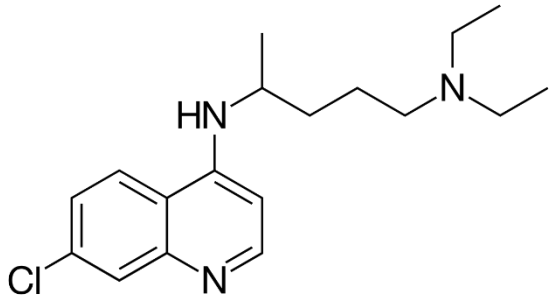
https://ppi.zoiclabs.io/#/

Here is an example of how targeting BRD2/4 can possibly interrupt virus infections



Drugs that can target either one of them

Chloroquine, a malarial drug, against Coronavirus?



The New York Times

MATTER

Scientists Identify 69 Drugs to Test Against the Coronavirus

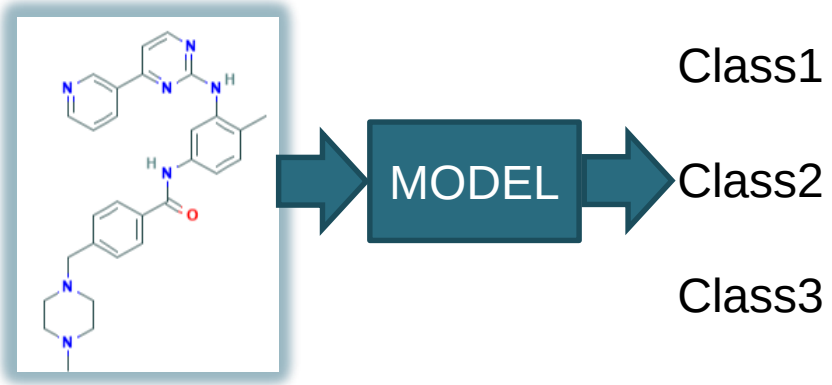
Two dozen of the medicines are already under investigation. Also on the list: chloroquine, a drug used to treat malaria.

https://www.nytimes.com/2020/03/22/science/coronavirus-drugs-chloroquine.html

Given a drug (ex. Imatinib) can we predict, its function?

Objectives → ?s

- Where do we begin and ?s
  1. We want to predict outcome (Class), what estimator will be appropriate?
     - Imatinib ←→ Antineoplastic agent

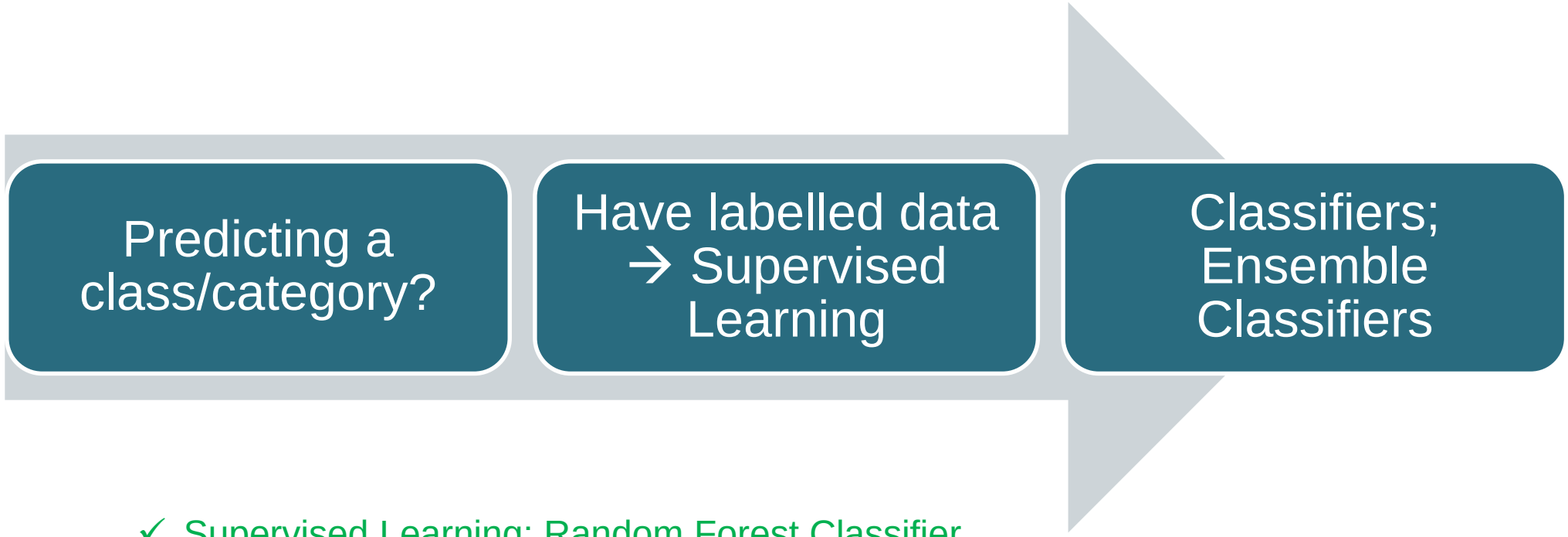


  2. Where do we get drug-class (outcome) and the chemical structure of compounds?

  3. How can we calculate Feature/property for drug molecules?

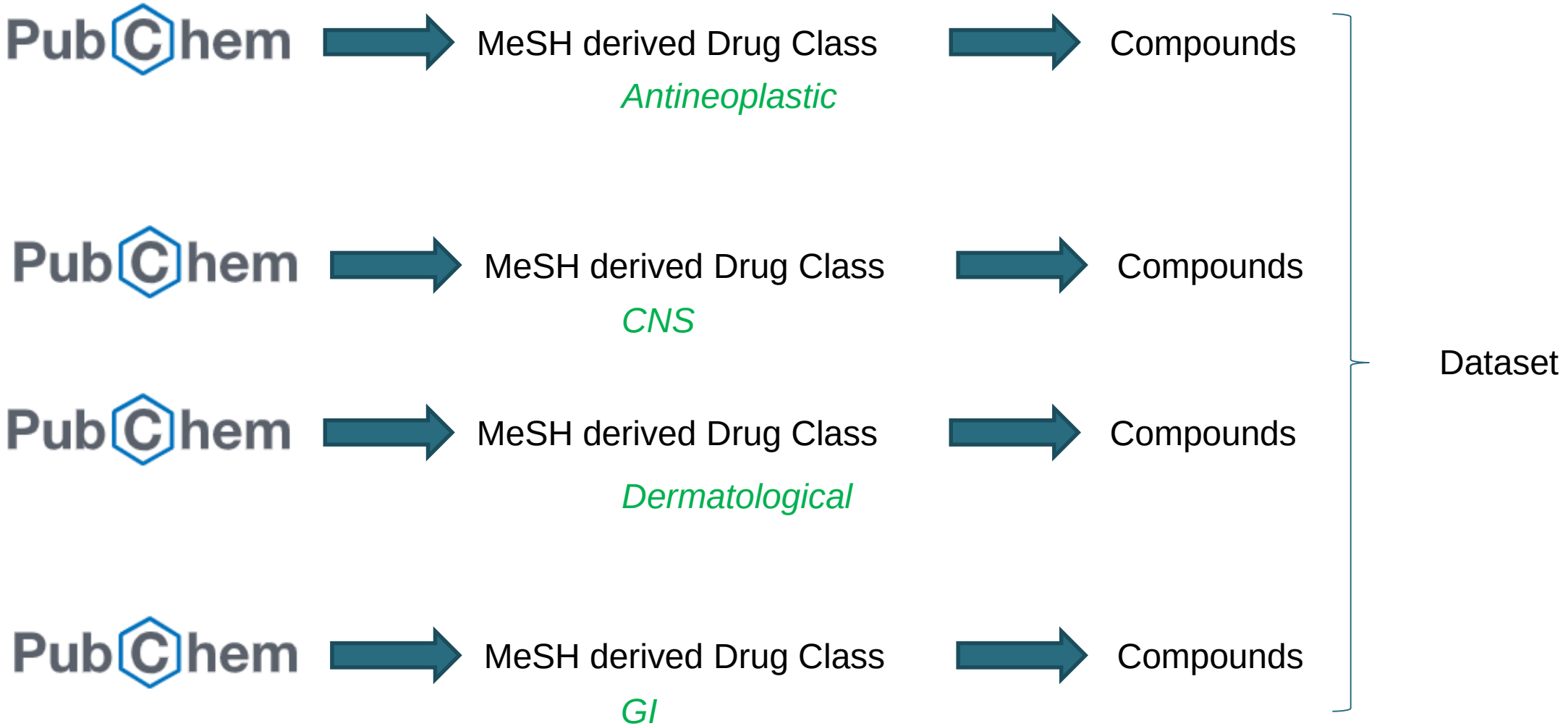
- We will provide some tools (software libraries) that can accomplish the above tasks

1. We want to predict outcome (Class), what estimator will be appropriate?



✓ Supervised Learning; Random Forest Classifier

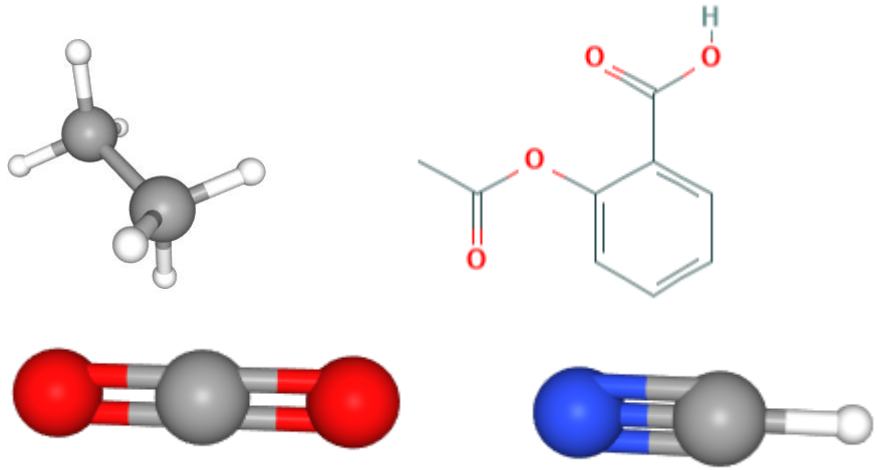
2. Where do we get drug-class (outcome) and the chemical structure of compounds?



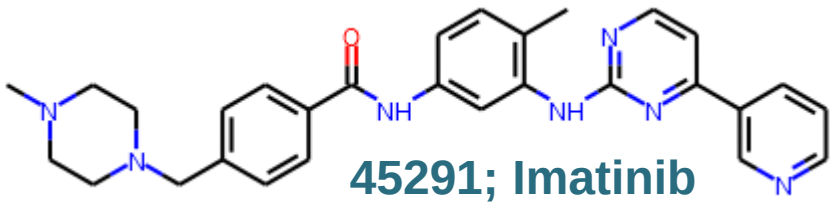
2. Where do we get drug-class (outcome) and the chemical structure of compounds?

- We need structure to compute chemical properties.
- SMILES (Simplified Molecular Input Line Entry System)
- *“ SMILES is a line notation* (a typographical method using printable characters) for entering and representing molecules and reactions.”

SMILES	Names
CC	Ethane
O=C=O	Carbon dioxide
C#N	Hydrogen Cyanide
CC(=O)OC1=CC=CC=C1C(=O)O	Aspirin

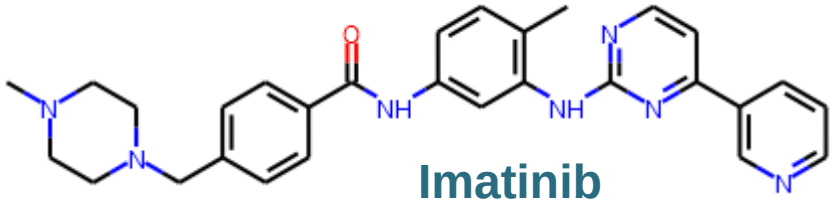


CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5.CS(=O)(=O)O



45291; Imatinib

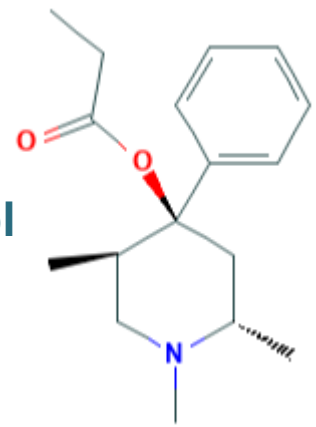
2. Where do we get drug-class (outcome) and the chemical structure of compounds?

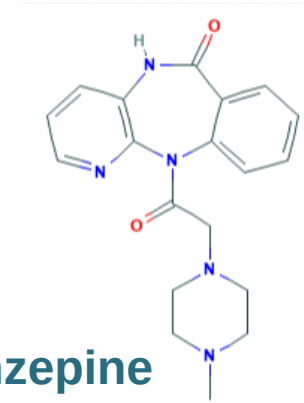


Imatinib

CID	Name	SMILES	Class
45291	Imatinib	CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5	Antineoplastic
20055107	gamma-Promedol	CCC(=O)OC1(CC(N(CC1C)C)C)C2=CC=CC=C2	CNS
5362119	lisinopril	C1=CC(=CC(=C1)C(=O)NCCO)C2=CC(=NC=N2)NC3=CC=C(C=C3)OC(F)(F)F	Cardio
4848	Pirenzepine	CN1CCN(CC1)CC(=O)N2C3=CC=CC=C3C(=O)NC4=C2N=CC=C4	GI

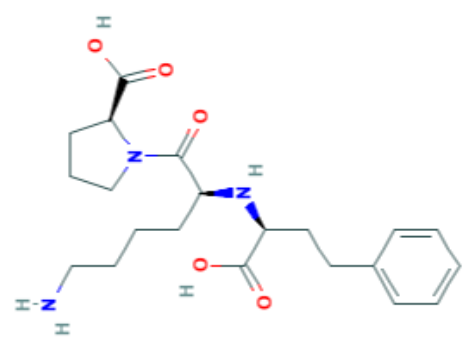
Gamma-Promedol



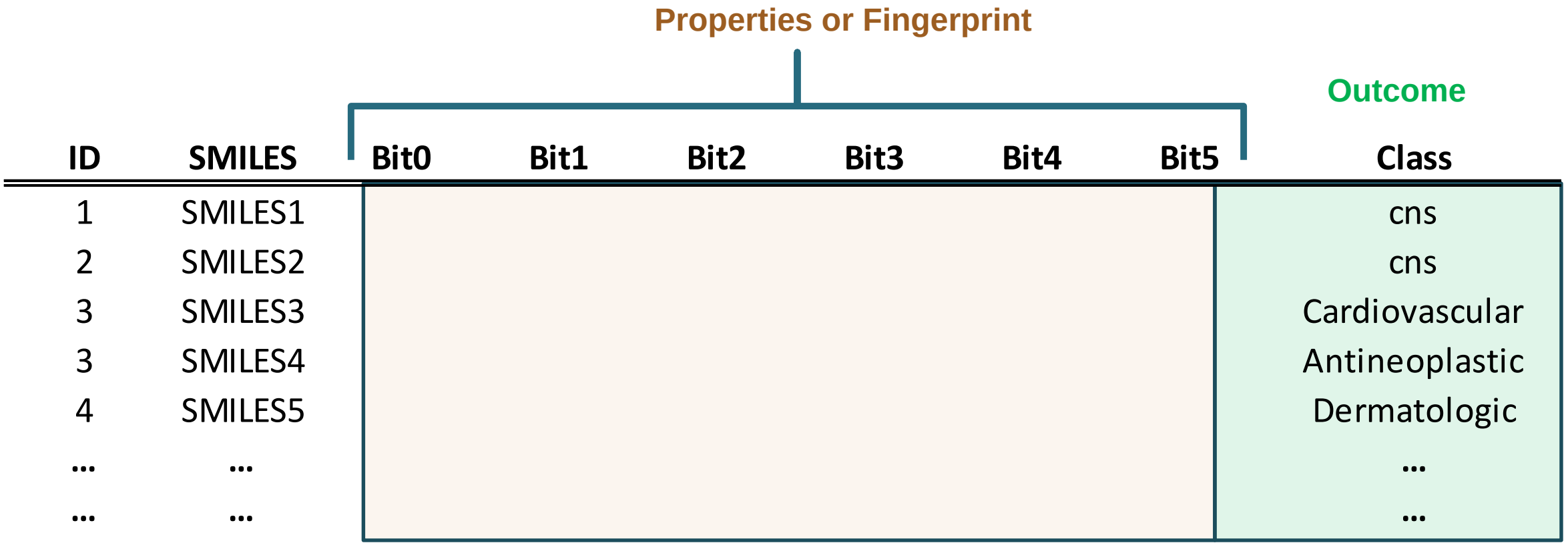


Pirenzepine

Lisinopril



3. How can we calculate Feature/property for drug molecules?



ID	SMILES	Bit0	Bit1	Bit2	Bit3	Bit4	Bit5	Class
1	SMILES1							cns
2	SMILES2							cns
3	SMILES3							Cardiovascular
3	SMILES4							Antineoplastic
4	SMILES5							Dermatologic
…	…							…
…	…							…

- ✓ Supervised Learning; Random Forest Classifier
- ✓ PubChem to gather data
- ○ Fingerprints for descriptors

3. How can we calculate Feature/property for drug molecules?

Properties or Fingerprint

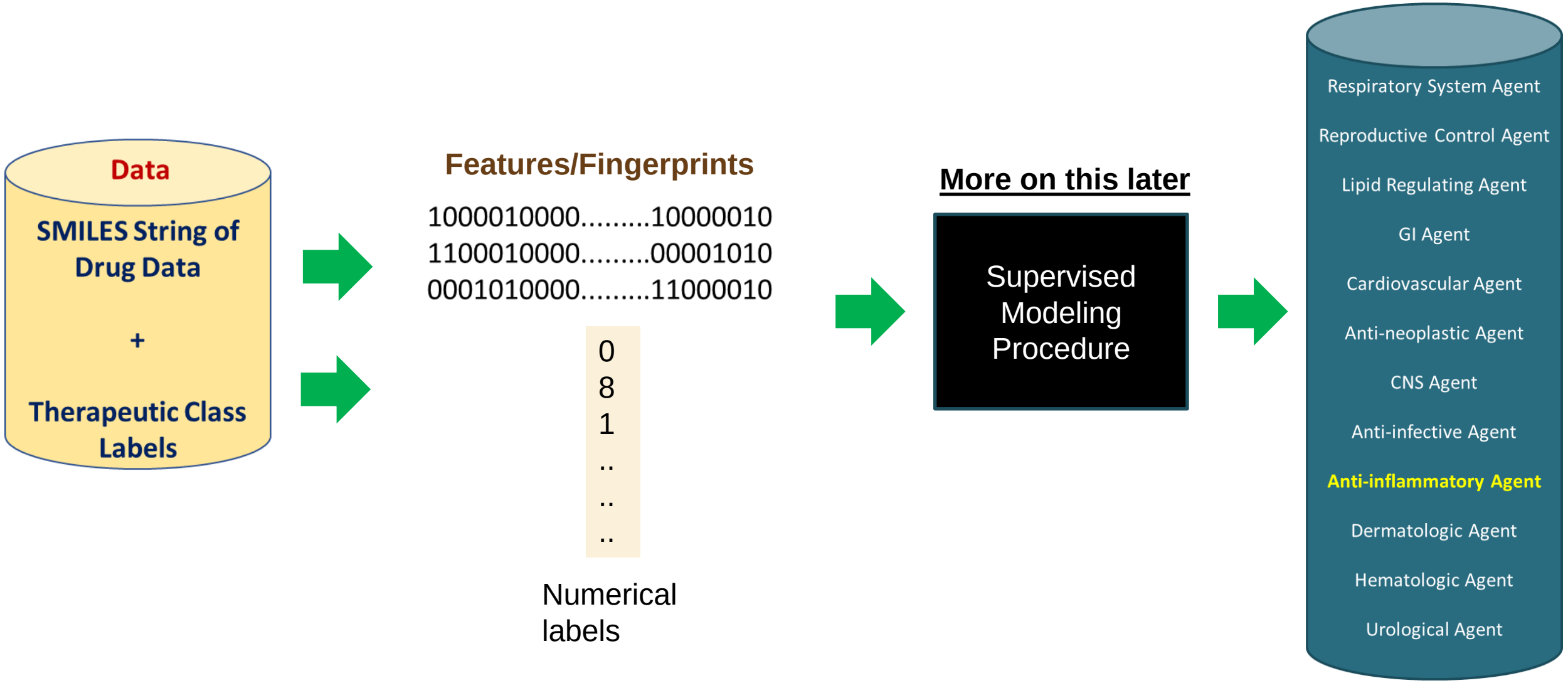
Only the first 6 bits for 5 molecules are shown

Outcome

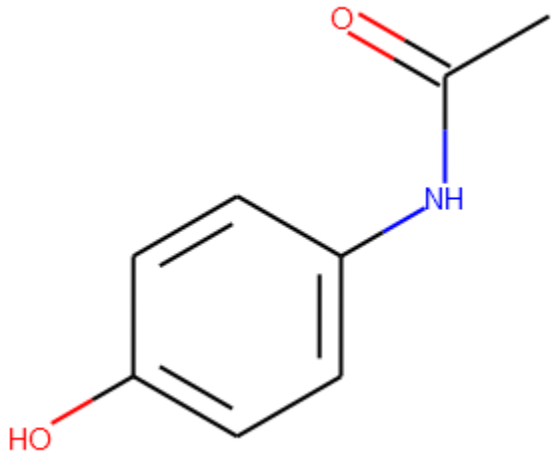
ID	SMILES	Bit0	Bit1	Bit2	Bit3	Bit4	Bit5	Class
1	SMILES1	1	1	0	1	0	1	cns
2	SMILES2	0	0	0	1	1	0	cns
3	SMILES3	1	0	0	1	0	0	Cardiovascular
3	SMILES4	1	0	0	1	1	0	Antineoplastic
4	SMILES5	1	1	0	1	1	1	Dermatologic
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

- ✓ Supervised Learning; Random Forest Classifier
- ✓ PubChem to gather data
- ✓ Fingerprints for descriptors

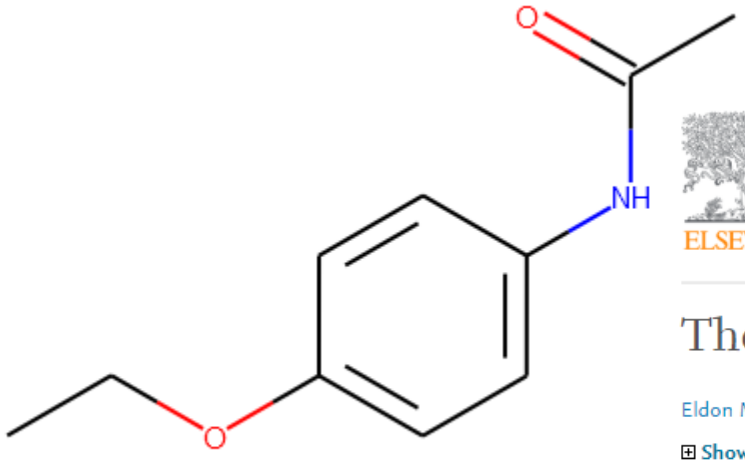
Recent efforts have showed that Molecular Fingerprints can Serve an Effective Feature Set for Machine-Learning



Molecular Fingerprints



Paracetamol



Phenacetin

01111000000000000000000000111

.....

.....

01111000000000011000000000111

.....

.....

Toxicology and Applied Pharmacology
Volume 1, Issue 3, May 1959, Pages 240-249

The acute oral toxicity of phenacetin

Eldon M. Boyd [1]

Show more

https://doi.org/10.1016/0041-008X(59)90108-5

Get rights and content

J Chem Inf Model, 50 (5), 742-54 2010 May 24

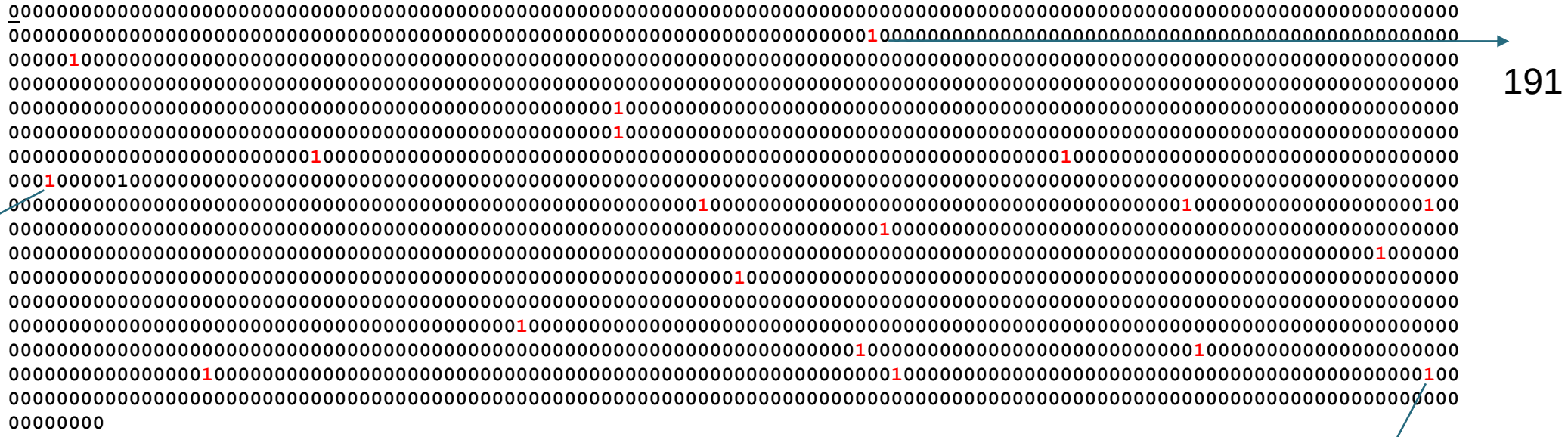
Extended-connectivity Fingerprints

David Rogers [1], Mathew Hahn

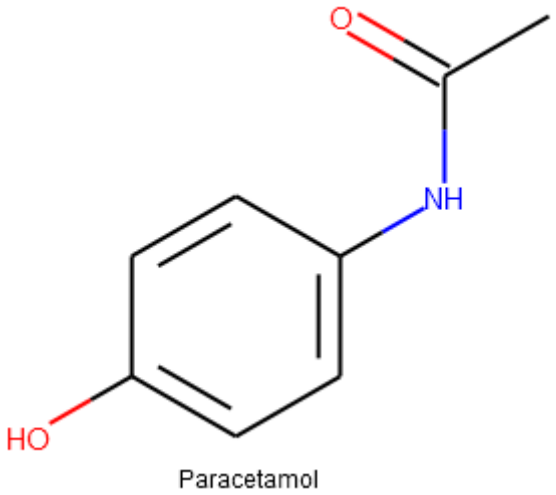
Affiliations + expand

PMID: 20426451 DOI: 10.1021/ci100050t

Morgan 2048-bit FingerPrint for Paracetamol



[**191**, 245, 530, 650, 745, 807, **843**, 849,
1017, 1057, 1077, 1152, 1313, 1380,
1602, 1750, 1778, 1816, 1873, **1917**]

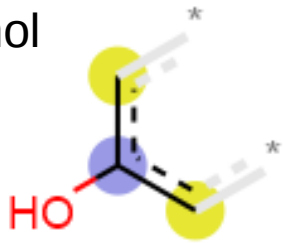


Paracetamol

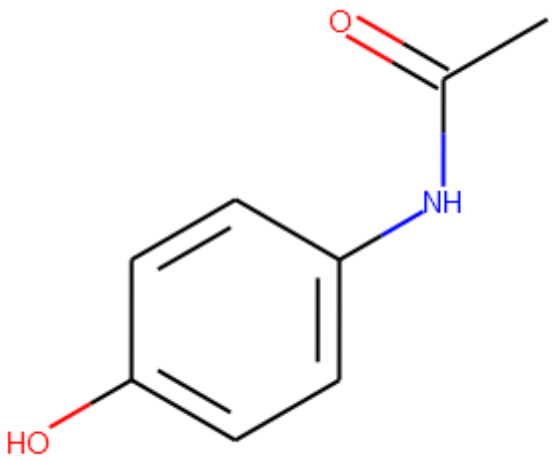
Morgan 2048-bit FingerPrint

Turned-on bits for Paracetamol [191, 245, 530, 650, 745, 807, 843, 849, 1017, 1057, 1077, 1152, 1313, 1380, 1602, 1750, 1778, 1816, 1873, 1917]

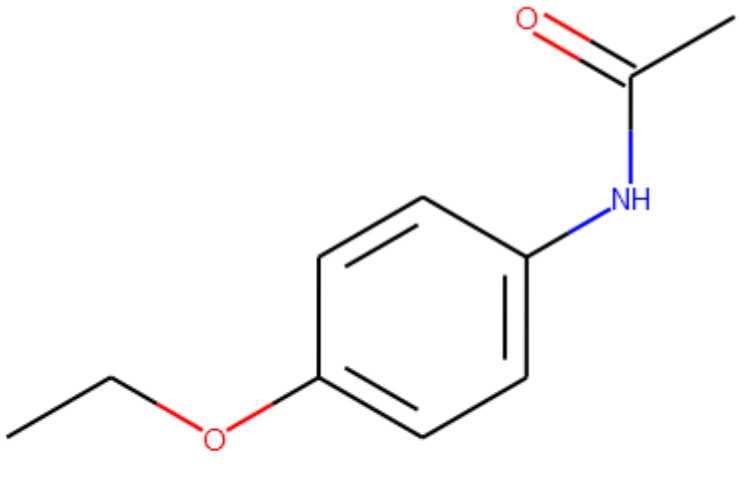
Paracetamol unique



745

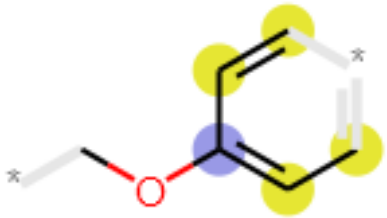


Paracetamol

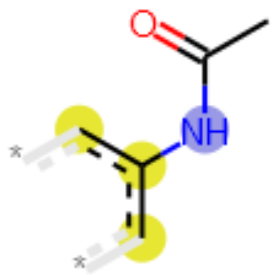


Phenacetin

Phenacetin unique



102



530

blue: the central atom in the environment

yellow: aromatic atoms

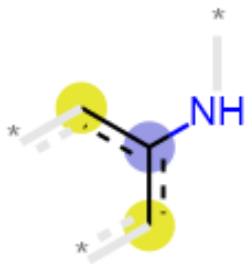
gray: aliphatic ring atoms

Underline marks common bits

Turned-bits for Pheacetin [69, 80, 102, **191**, 237, **245**, 294, 322, **530**, **650**, 695, 718, **807**, **843**, **849**, **1017**, **1057**, **1077**, **1152,** 1238, **1380**, 1452, **1750**, **1816**, **1873**, **1917**]

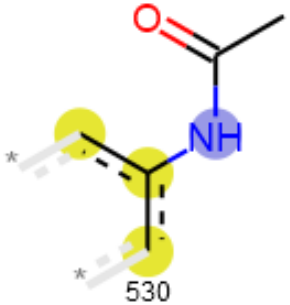
Paracetamol fingerprint collection

Turned-on bits for Paracetamol [191, 245, 530, 650, 745, 807, 843, 849, 1017, 1057, 1077, 1152, 1313, 1380, 1602, 1750, 1778, 1816, 1873, 1917]



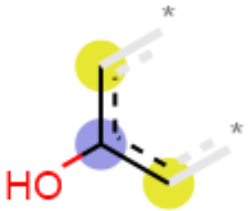
191

245



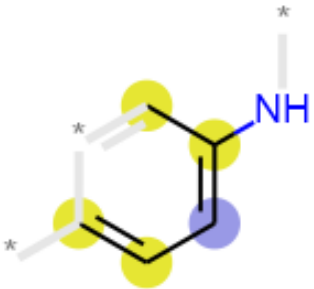
530

650

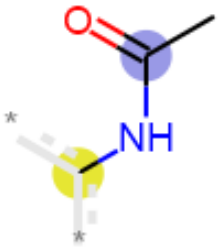


745


807



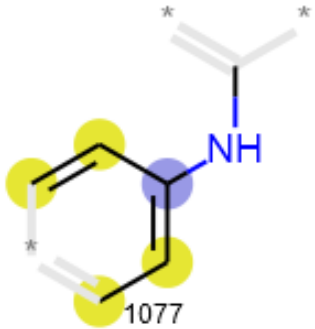
843



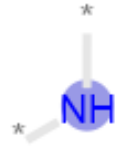
849


1017

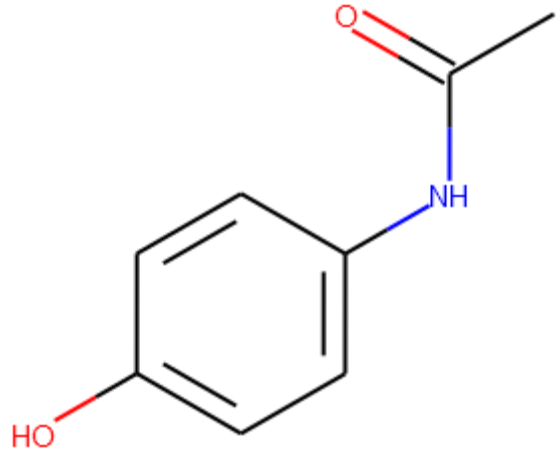

1057



1077



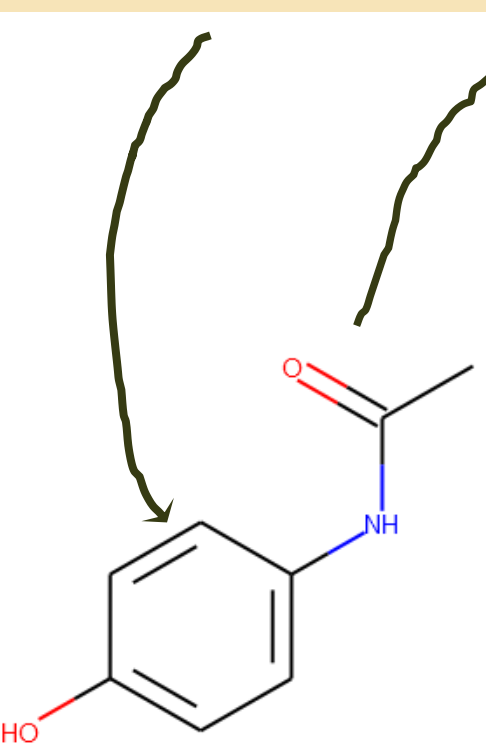
1152



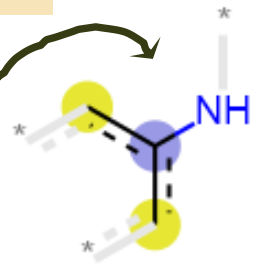
Paracetamol

Part-1: SMILES → Structure → SubFragments → Fingerprint

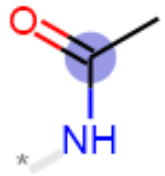
CC(=O)NC1=CC=C(C=C1)O



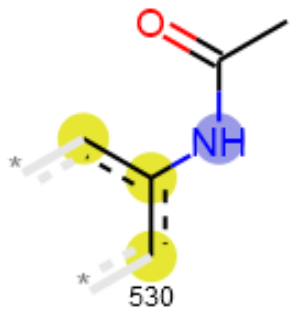
Paracetamol



191



245

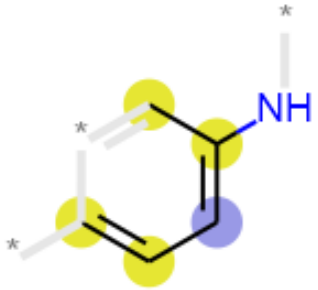


530

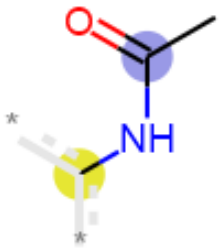
650

745

807



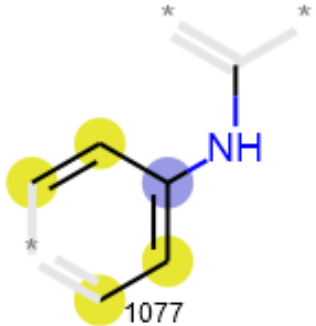
843



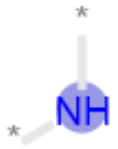
849

1017

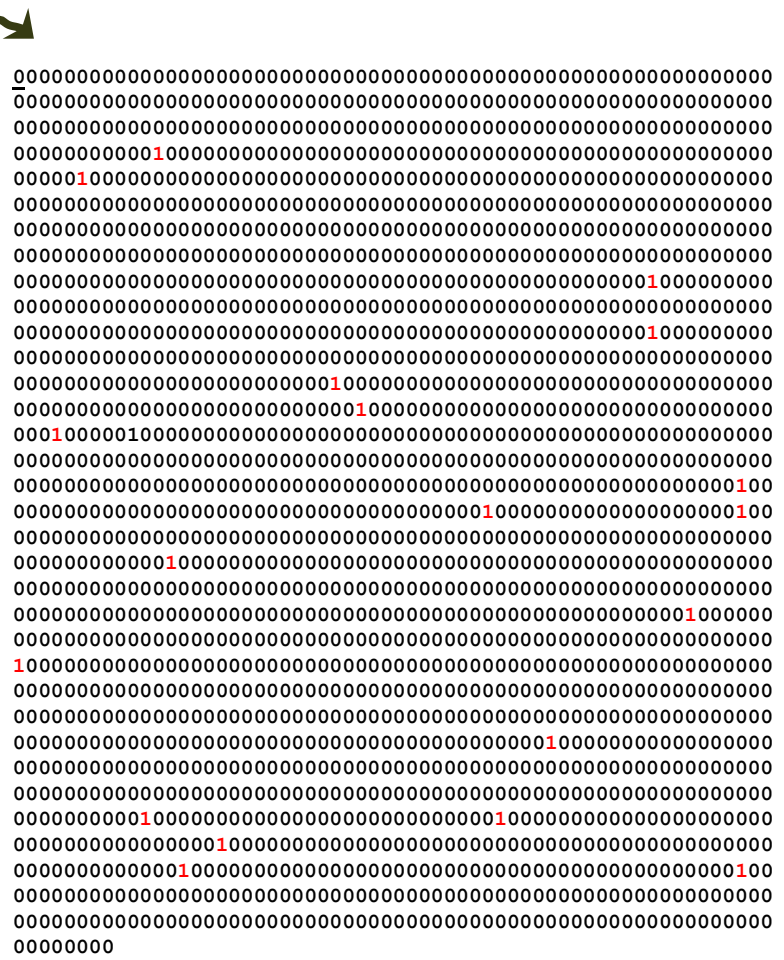
1057



1077



1152



Part-II: Outcome labels to numerical quantities

Antineoplastic Agents (antineoplastic)
Cardiovascular Agents (cardio)
Central Nervous Systems (cns)

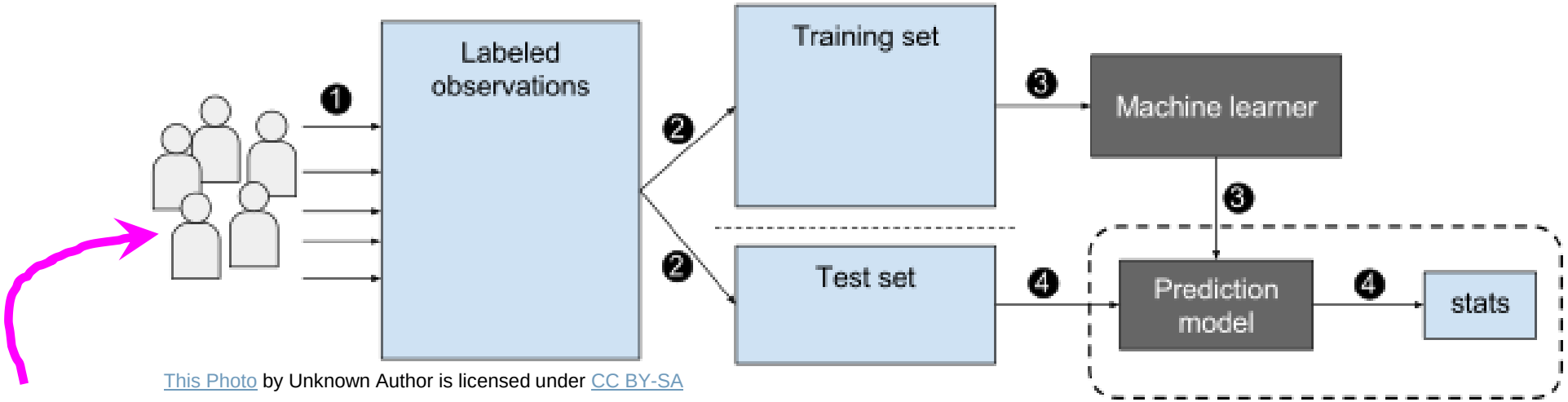
0: Antineoplastic Agents (antineoplastic)
1: Cardiovascular Agents (cardio)
2: Central Nervous Systems (cns)

Tools Review

https://github.com/ravichas/ML-predict-drugclass

Machine Learning using Supervised Learning as an example

- Randomly split the data into Training (ex. 60%) and Test set (ex. 40%)
- Using the training dataset we would like to:
  - Accurately(??) predict new or unseen case labels
  - Try to understand which inputs affect (& how) the outcome (i.e. Cancer or not)
  - Evaluate (using test set) the correctness of our predictions and inferences



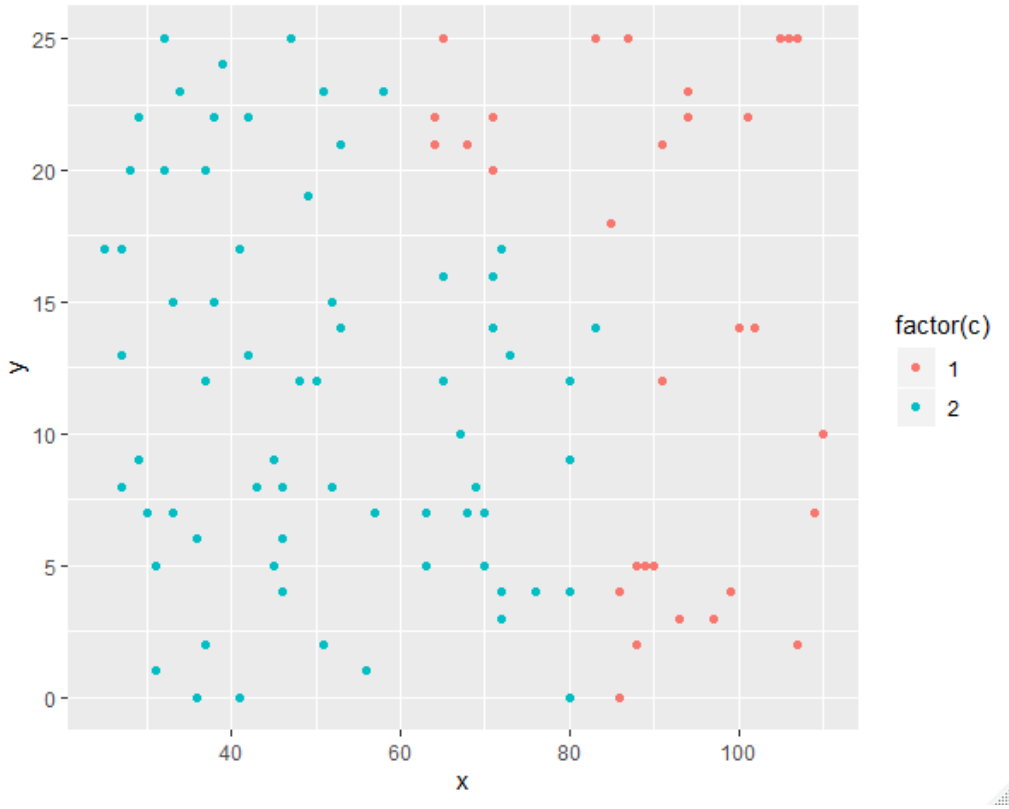
ID Bit1 Bit2 Bit3 ….. Bit2048 OUTCOME

Machine-learner: Classification

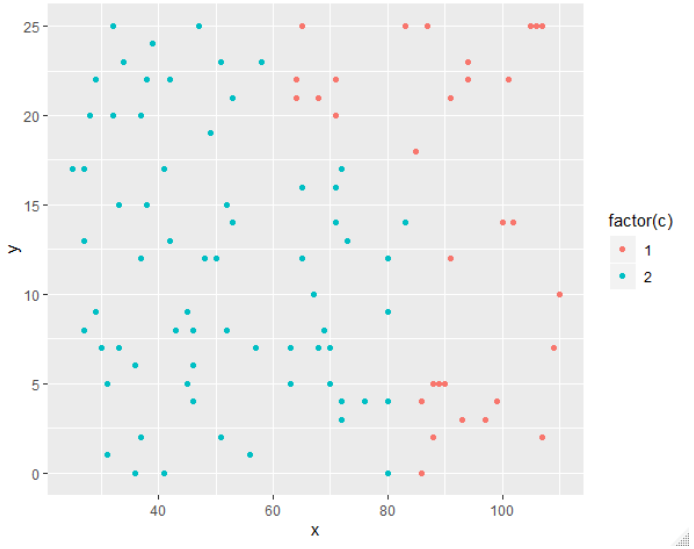
- Qualitative variables: unordered set, C,
  - Eye-color ∈ { **black, brown, blue** }
  - Given a feature vector X (fingerprint) and a qualitative outcome Y (taking values from the set in C, the task is to identify a function C(X) that predicts a value for Y
    - C(X) ∈ C
    - Takes in X and outputs one of the elements of C
  - One can also compute the probabilities of what X belongs to C

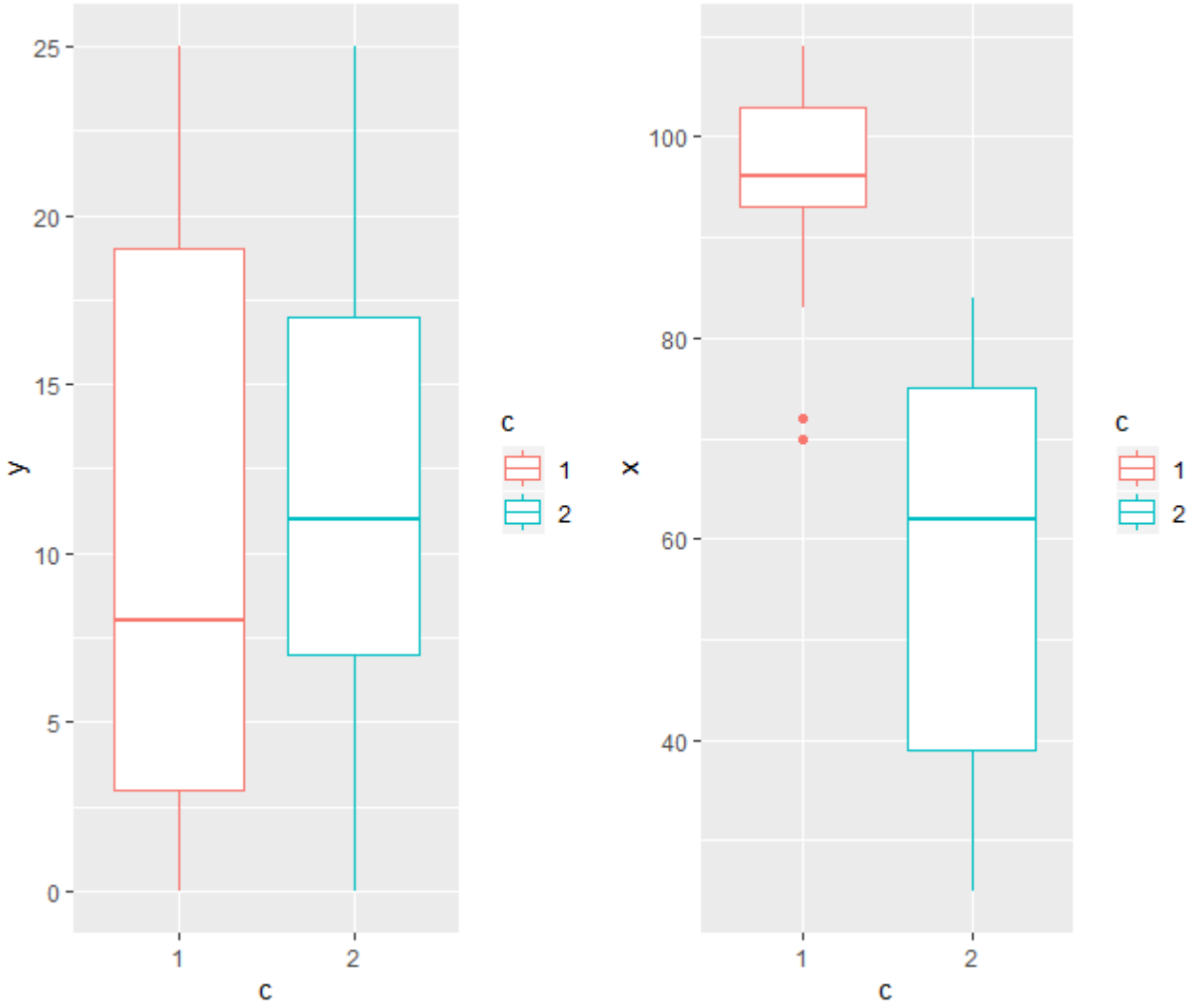
CART: Classification motivation

ID	SMILES	sFP1	sFP2	Class
1	SMILES1	1	1	cns
2	SMILES2	0	0	cns
3	SMILES3	1	0	Cardiovascular
3	SMILES4	1	0	Cardiovascular
4	SMILES5	1	1	cns
…	…	…	…	…
…	…	…	…	…



CART/binary-trees: Classification motivation



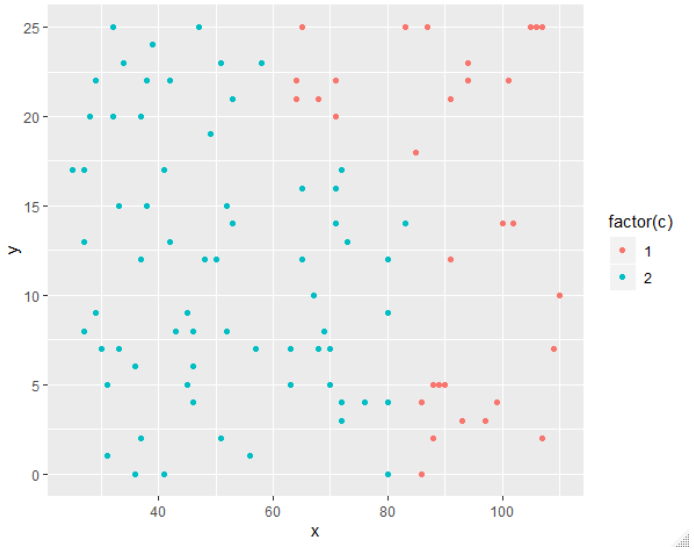


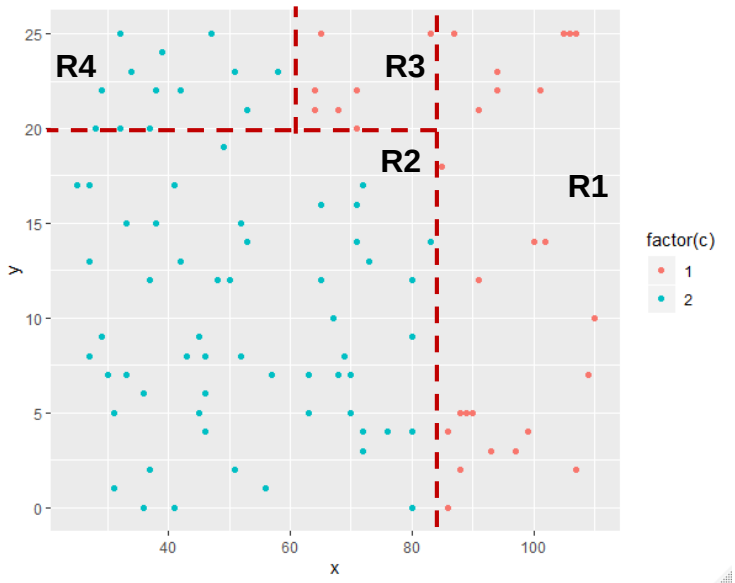
Split using single feature (x) and a cut-off (t_x >= 84)

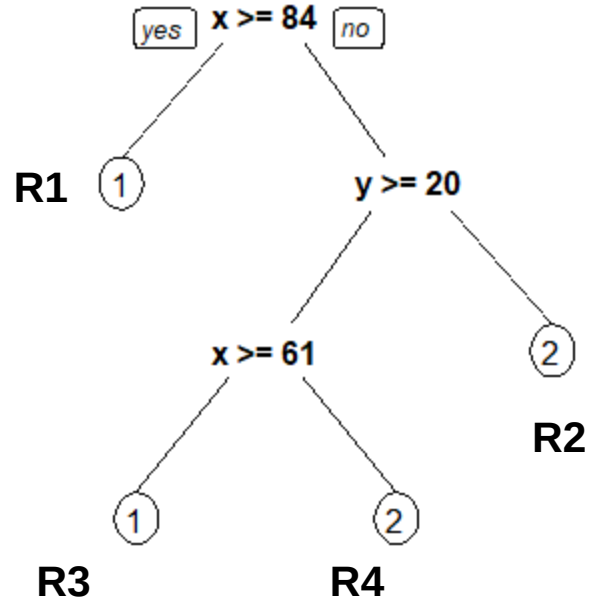
Choice of feature/cutoff is based on a COST function (that attempts of find pure nodes)

John Tukey (Box-plots)

Decision-trees: Classification motivation





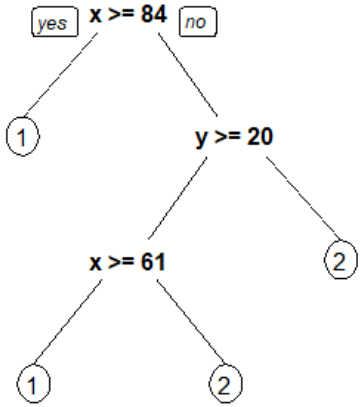


Terminal/Leaf nodes: R1, R2, R3
Internal nodes: Predictor space splitting points

CART produces binary tree; other algorithms (not common) can produce more than 2 children

Decision-trees: Important points

- Does trees explore all possible combinations
  - No? It uses something called **top-down**, **greedy** (best split at the particular step) approach; Not optimal but reasonably good
    - Recursive binary splitting

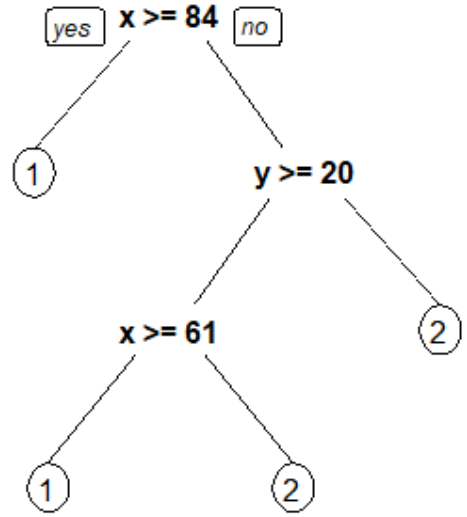


- Trees identify a predictor X_j and a cut-point c such that the splitting the predictor space leads to lower variance
  - Region1: **$\{X|X_j < c\}$;** Region2: **$\{X|X_j \geq c\}$**

- Tree splitting will continue from the above cut regions and will continue to split

Decision-trees: Important points

- Trees usually are grown bushier and pruned back to find a best sub-tree
  - Check Rob Tibshirani book or Google for details.



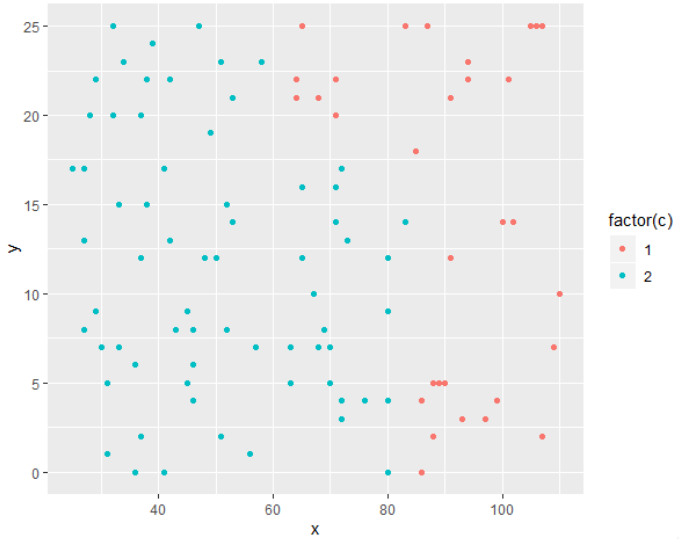
- Process will continue will our criteria is reached
  - We call this Hyperparameters
  - Bucket size, depth of tree etc.

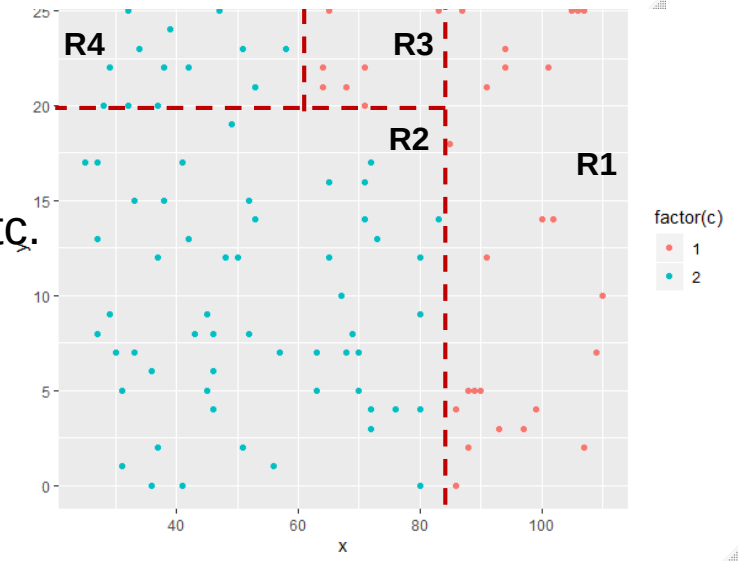
CANDLE (Cancer Distributed Learning Environment)

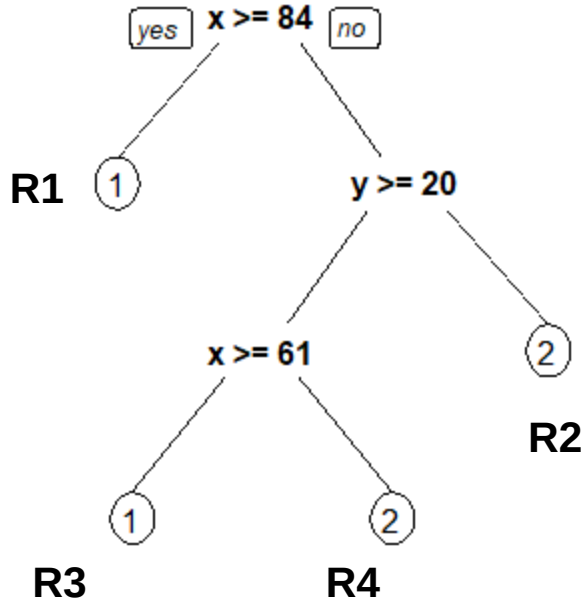
https://hpc.nih.gov/apps/candle/index.html

Hyperparameters

- Hyperparameters
  - Set before training a model
  - Drive the training process
  - Tuned between training iteratior
- Examples
  - RandomForest
    - # of observations in a leaf etc.
  - KNN
    - Number of neighbors

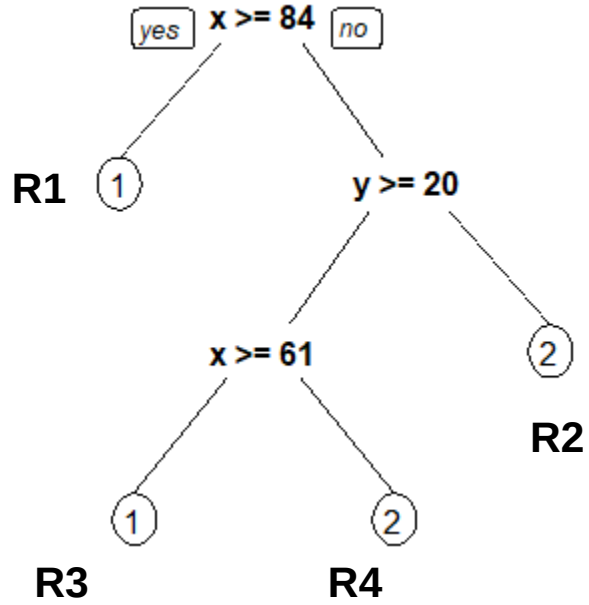






Decision-tree: Summary

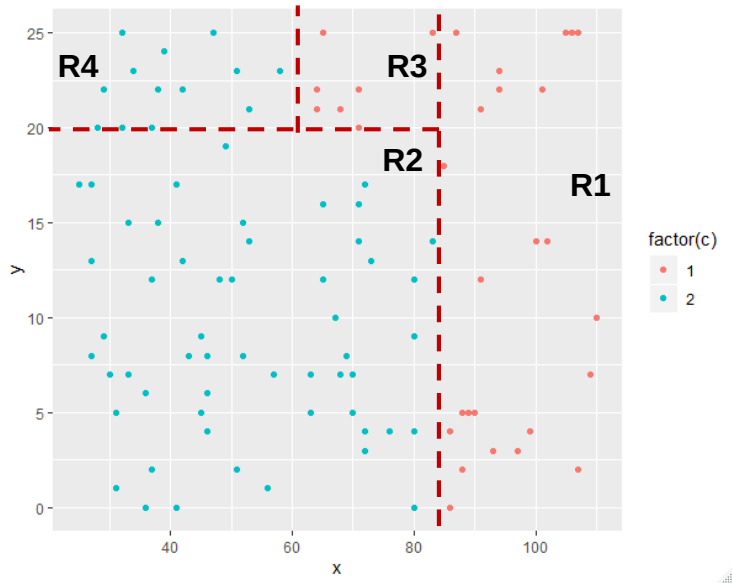
- Outcomes: Categorical (unordered) variables
- Tree or decision tree is a set of **Yes/No** questions
- Predictions are given by the nodes (or ends)
  - That is which class is most common within the partition



- Trees work by partitioning/segmenting the predictor space (lines or boxes) with the hope of getting a pure space (ie of one class)

Tree building metrics: Measure of the quality of a split

- **Gini index** or node purity for a node, m
  - $G_m = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk})$
    - mth region and kth class
  - A small value indicates a region that contains predominantly of one class
  - Pure node: G = 0; Mixed class node: G = 0.5 (equal proportions)



- **Cross-entropy** a similar measure to Gini index
  - $D = -\sum_{k=1}^{K} \hat{P}_{mk}\, log(\hat{P}_{mk})$

Decision-Tree

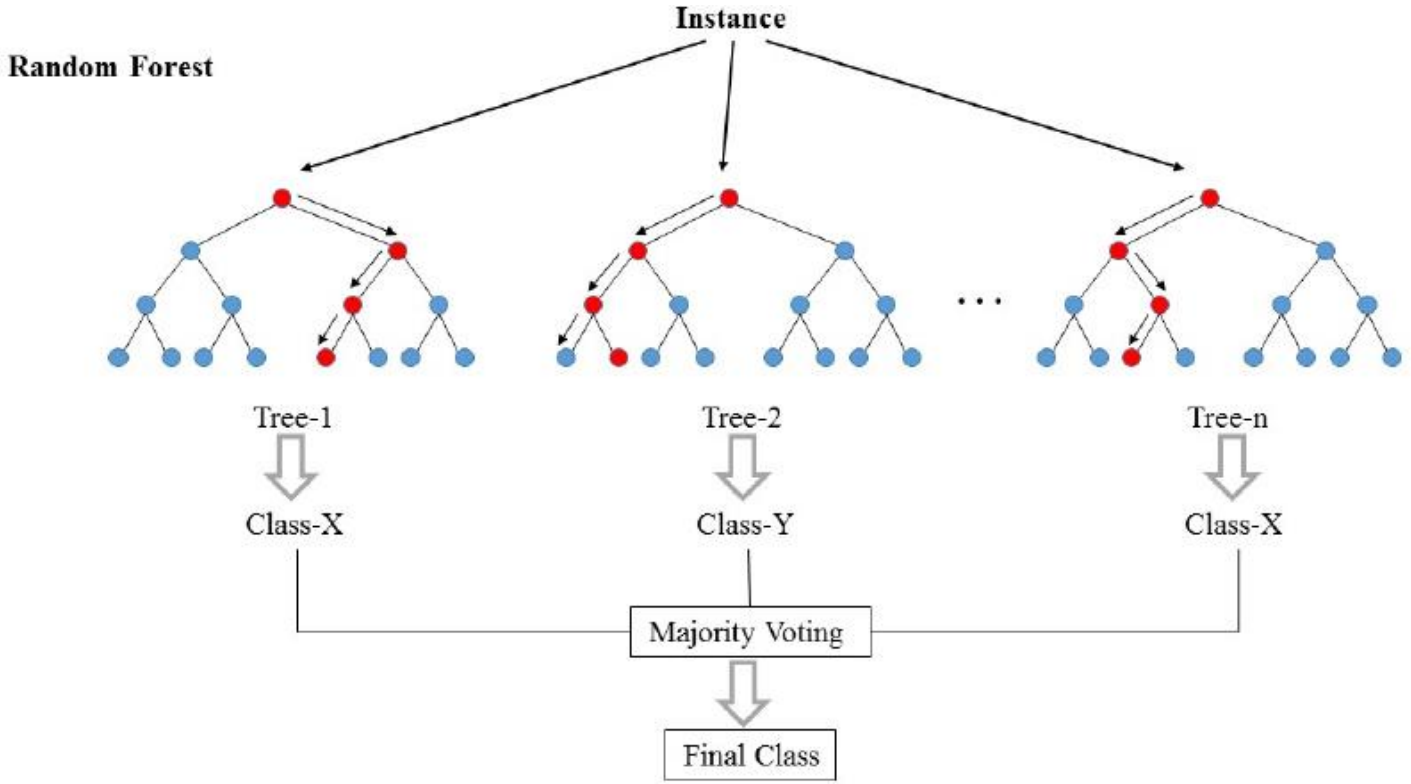
- **Pros**:
  - Easy to explain
  - Interpretable
  - No preparation/scaling/centering
  - Non-linear method

- Cons
  - High variance
  - Unstable with small changes in the data
  - Perform poorly when compared to ensemble based methods (Random Forest)

Ensemble Methods: Random Forest

Law of Large Numbers

Wisdom of the crowd



For each observation, record the class prediction from each of the B trees and take a majority vote

Random Forest

- Combining a large number of trees result in improvements in accuracy
  - n independent measurements Z_1, …, Z_n each with variance of σ^2, variance of the mean $\bar{Z}$ will be $\frac{\sigma^2}{n}$

- Scikit-learn
  - Random sampling (with replacement; bootstrapping) of training data points when building trees
  - Random subsets
    - Usually $\sqrt{(n\_\text{features})}$ considered when splitting nodes

Evaluation of binary classifiers

- Confusion Matrix and **B**alanced **A**ccuracy (**BA**) Score
  - Count the number of times Class A is predicated as A or not (or other classes)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathbf{BA} = \frac{\mathrm{TPR} + \mathrm{TNR}}{2}$$

n = 1000	Predicted: Yes	Predicted:No	
Actual: Yes	890	10	900
Actual: No	90	10	100
	980	20	

Accuracy = 0.9000

TPR = 0.9082
TNR = 0.5000
BA = 0.7041

Overview of Classification Process

Training Data

Therapeutic Class Labels (MeSH Classification)

SMILES String of Drug Data

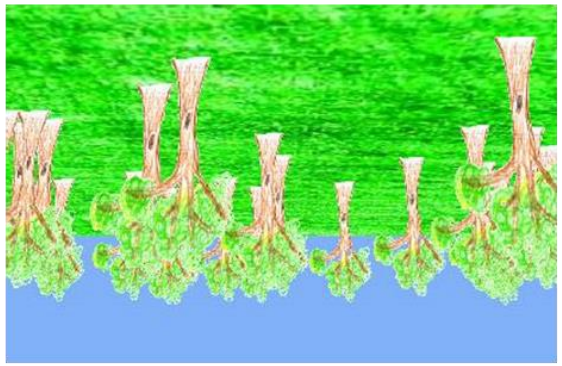
Fingerprints

1000010000.....10000010
1100010000.....00001010
0001010000.....11000010

Test Data

SMILES String of Drug Data

Random Forest



- Respiratory System Agent
- Reproductive Control Agent
- Lipid Regulating Agent
- GI Agent
- Cardiovascular Agent
- Anti-neoplastic Agent
- CNS Agent
- Anti-infective Agent
- **Anti-inflammatory Agent**
- Dermatologic Agent
- Hematologic Agent
- Urological Agent

Thanks

- Contact Info
  - ravichandrans@mail.nih.gov

Extra

Evaluating Trees

- Classification Error Rate
  - $E = 1 - \max_k(\hat{P}_{mk})$
  - $\hat{P}_{mk}$: proportion of training observations in the mth region that belong to kth class
  - But, this is not sensitive, noisy and not commonly used