



Predicting Drug Function Using Small-Molecule Structure Information

S. Ravichandran
BIDS, FNLCR

(announcement coming soon)

Acknowledgements

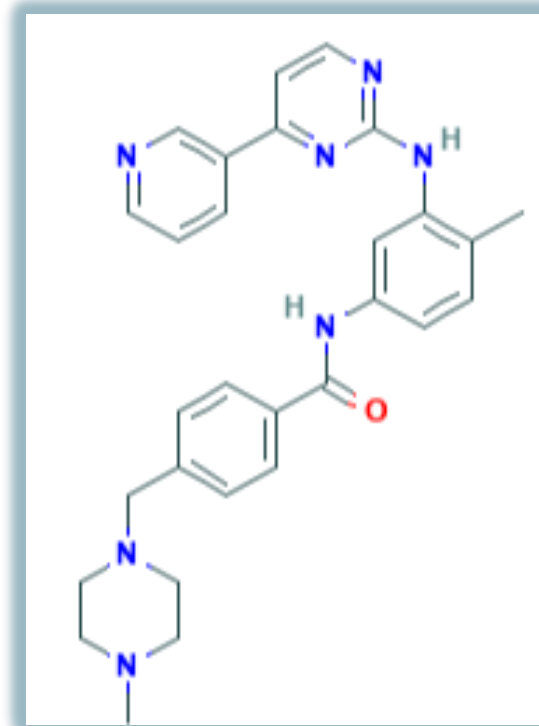
- NCI-DOE Pilot-1 Team
- BIDS
 - Drs. George Zaki, Andrew Weissman and Eric Stahlberg
 - Colleagues who reviewed the material

Introduction

- This is part of the NCI-DOE knowledge/capability transfer efforts
- Share tools/techniques/solutions for cancer related problems. We often take a test-case and show how it works
- You will be able to take the test-case (code/scripts) and tune it to your needs
- We want to hear from you, please send us your feed-back

Objectives

- Knowledge-transfer of reproducible Machine-Learning frameworks for modeling drug-discovery problems
 - User-friendly, run across multiple-OS etc.
 - A Notebook that supports multiple languages
- 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 drug function?

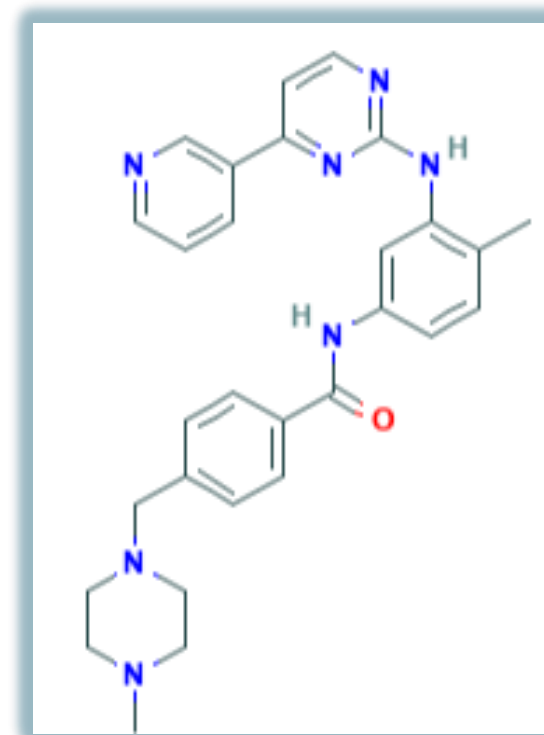


Antineoplastic Agents:
Substances that inhibit or prevent the proliferation of NEOPLASMS.

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

Objectives → ?s

- Where do we begin and ?s
 - Drug molecules: List of compounds
 - A definition of “How to define drug function?”
 - A list of properties associated with molecules



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

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 Zhou¹, Yuan Hou¹, Jiayu Shen¹, Yin Huang¹, William Martin ¹ and Feixiong Cheng^{1,2,3}

Drug repurposing: progress, challenges and recommendations

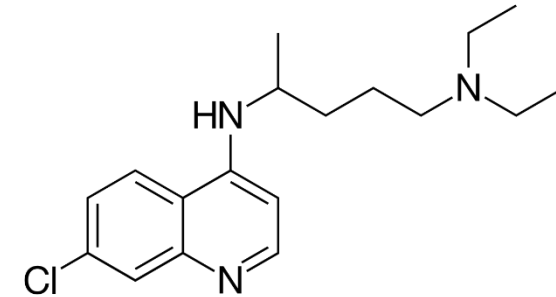
Sudeep Pushpakom¹, Francesco Iorio², Patrick A. Eyers³, K. Jane Escott⁴, Shirley Hopper⁵, Andrew Wells⁶, Andrew Doig⁷, Tim Williams⁸, Joanna Latimer⁹, Christine McNamee¹, Alan Norris¹, Philippe Sanseau¹⁰, David Cavalla¹¹ and Munir Pirmohamed¹

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

Frederick National Laboratory for Cancer Research

Chloroquine, a malarial drug, against Coronavirus?

The New York Times



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

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.



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

Test case example

[RETURN TO ISSUE](#) | [< PREV](#) **ARTICLE** [NEXT >](#)

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>

Copyright © 2019 American Chemical Society

[RIGHTS & PERMISSIONS](#)

 ACS AuthorChoice
with CC-BY license

Article Views

1662

Altmetric

21

Citations

-

[LEARN ABOUT THESE METRICS](#)

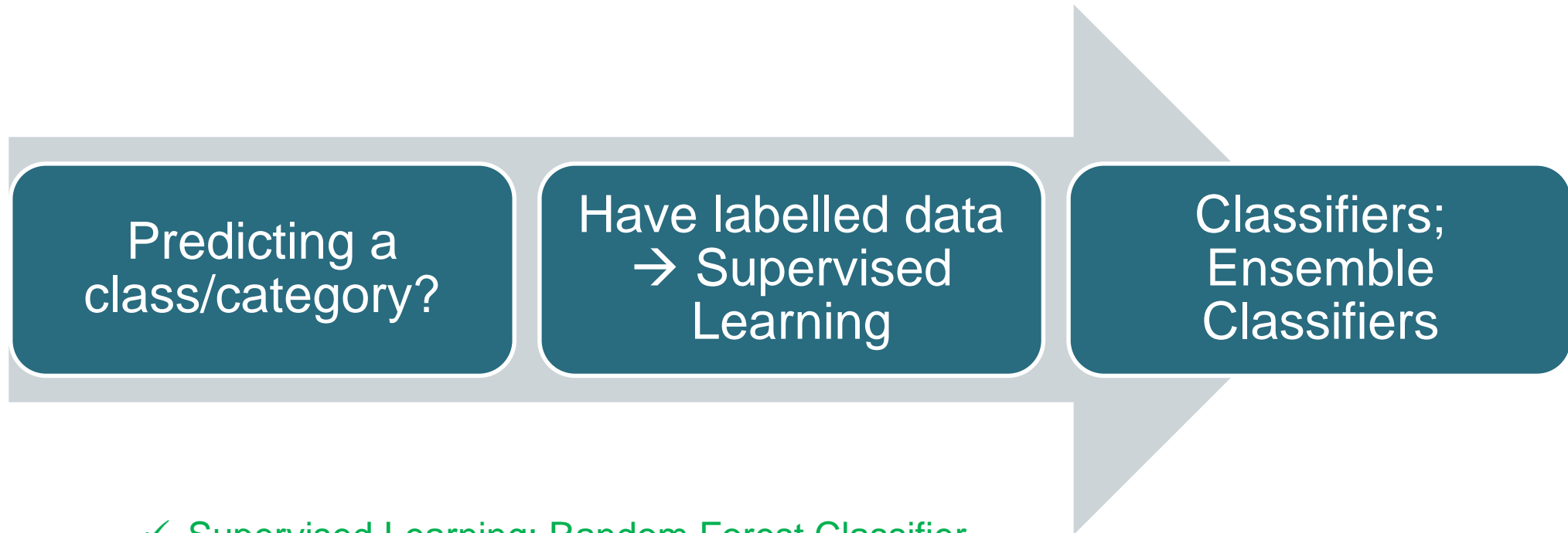
[Share](#) [Add to](#) [Export](#)



Goals/Questions

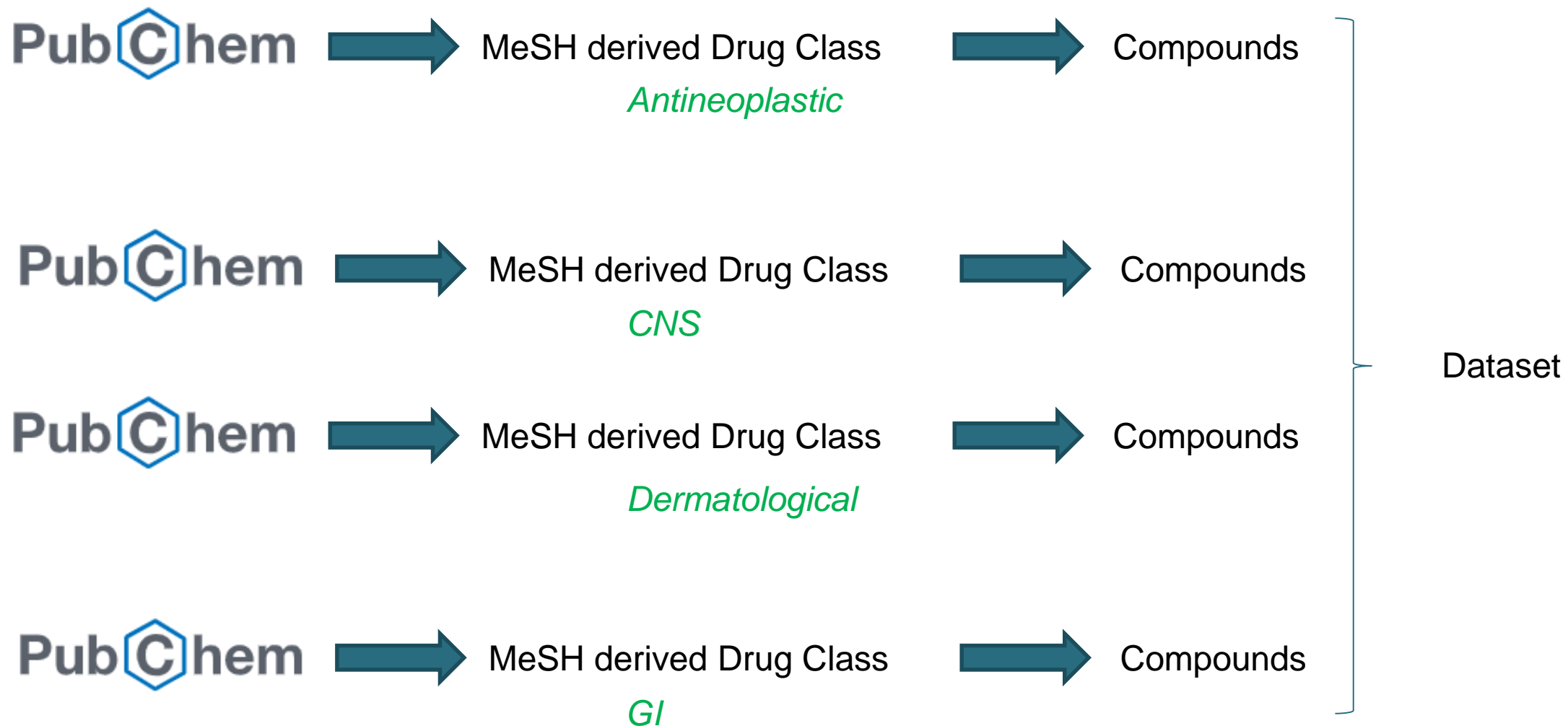
- To identify drugs that belong to a Pharmacological Class using chemical properties (in-silico)
 1. We want to predict outcome (Class), what estimator will be appropriate?
 2. Where do we get drug-class (outcome) and the chemical structure of compounds?
 3. How can we calculate Feature/property for drug molecules?
 4. 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

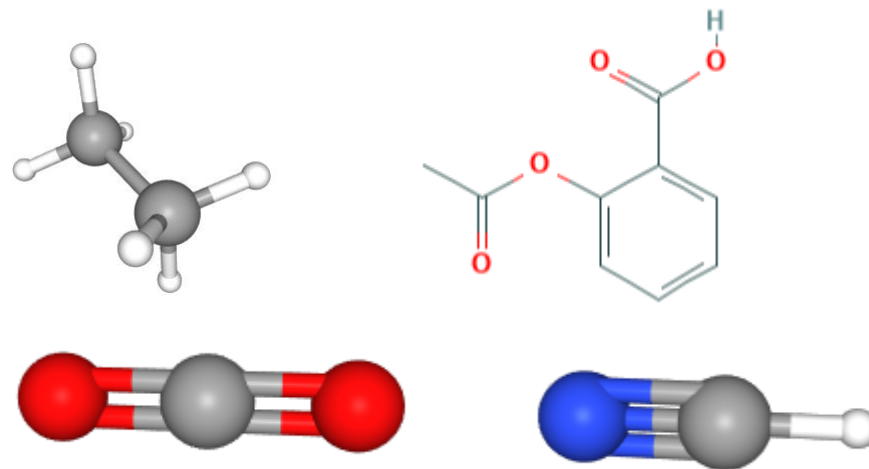
2. Where do we get drug-class (outcome) and the chemical structure of compounds?



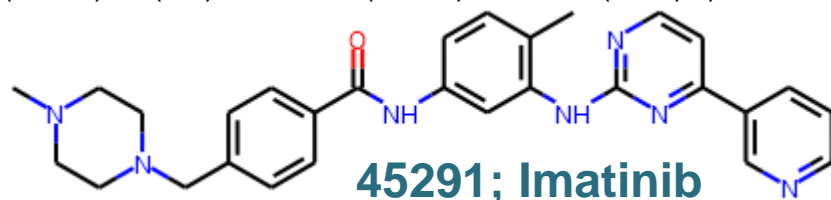
2. Where do we get drug-class (outcome) and the chemical structure of compounds?

- 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.”

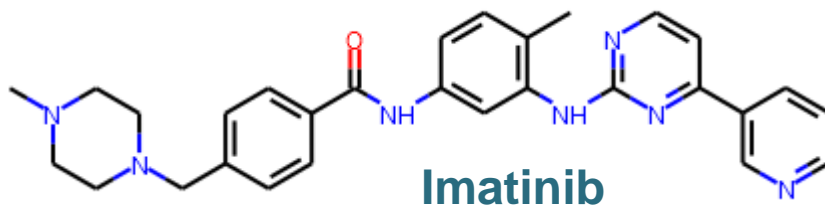
SMILES	Names
CC	Ethane
O=C=O	Carbon dioxide
C#N	Hydrogen Cyanide
CC(=O)OC1=CC=CC=C1C(=O)O	Aspirin



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



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

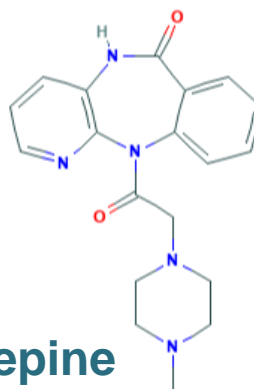


CID	Name	SMILES	Class
45291	Imatinib	<chem>CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5</chem>	Antineoplastic
20055107	gamma-Promedol	<chem>CCC(=O)OC1(CC(N(CC1C)C)C)C2=CC=CC=C2</chem>	CNS
5362119	lisinopril	<chem>C1=CC(=CC(=C1)C(=O)NCCO)C2=CC(=NC=N2)NC3=CC=C(C=C3)OC(F)(F)F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

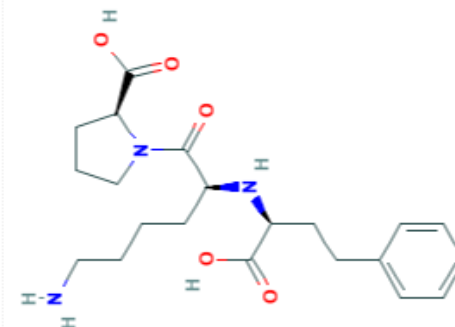
Gamma-Promedol



Pirenzepine



Lisinopril



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

		Properties or Fingerprint						Outcome
ID	SMILES	Bit0	Bit1	Bit2	Bit3	Bit4	Bit5	Class
1	SMILES1							cns
2	SMILES2							cns
3	SMILES3							Cardiovascular
3	SMILES4							Antineoplastic
4	SMILES5							Dermatologic
...
...

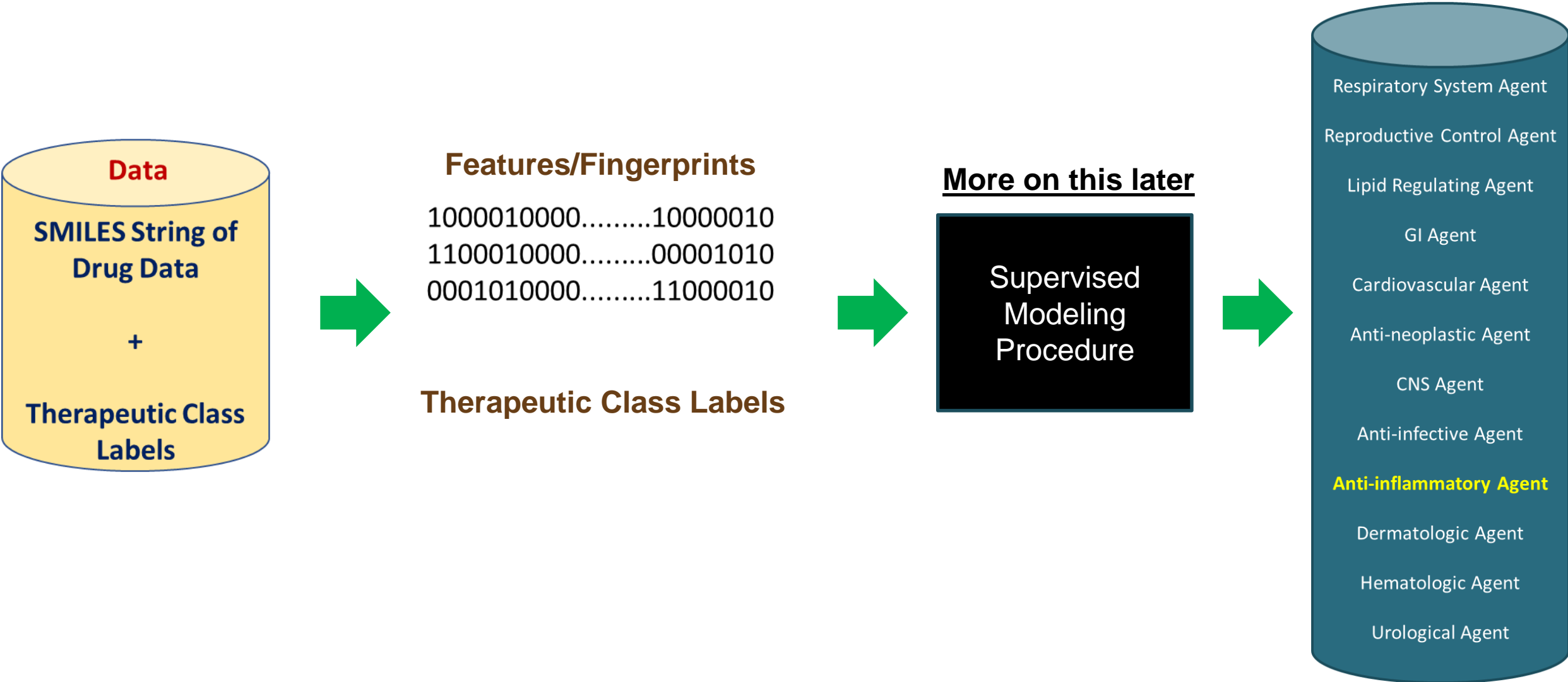
- ✓ Supervised Learning; Random Forest Classifier
- ✓ PubChem to gather data
- Fingerprints for descriptors

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

		Properties or Fingerprint						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



The property we will compute is called Molecular Fingerprint

> [J Chem Inf Model](#), 50 (5), 742-54 2010 May 24

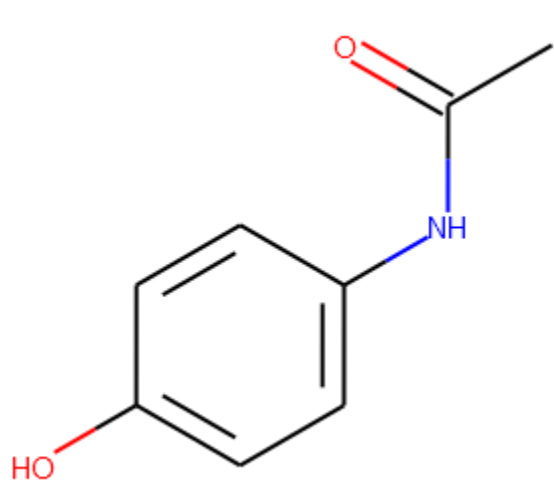
Extended-connectivity Fingerprints

[David Rogers](#) ¹, [Mathew Hahn](#)

Affiliations + expand

PMID: 20426451 DOI: [10.1021/ci100050t](#)

Molecular Fingerprints

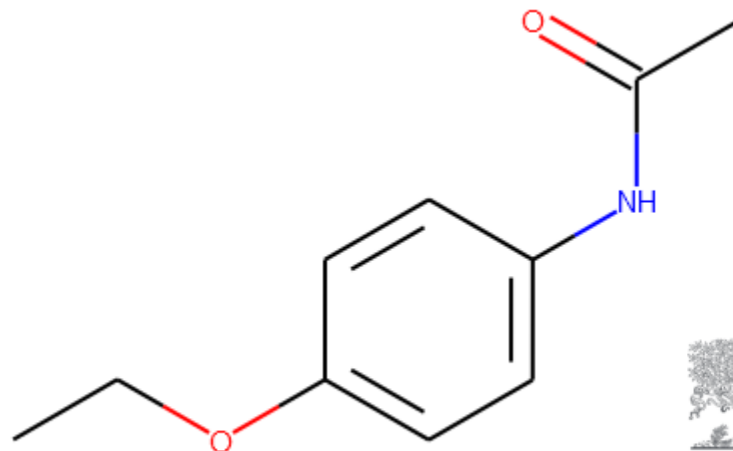


Paracetamol



0111100000000000
0000000000000111

.....
.....



Phenacetin



0111100000000000
1100000000000111

.....
.....



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

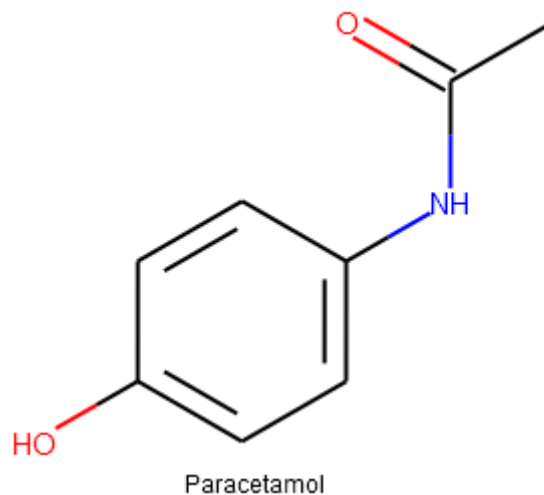
[Get rights and content](#)

Morgan 2048-bit FingerPrint for Paracetamol

191

843

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

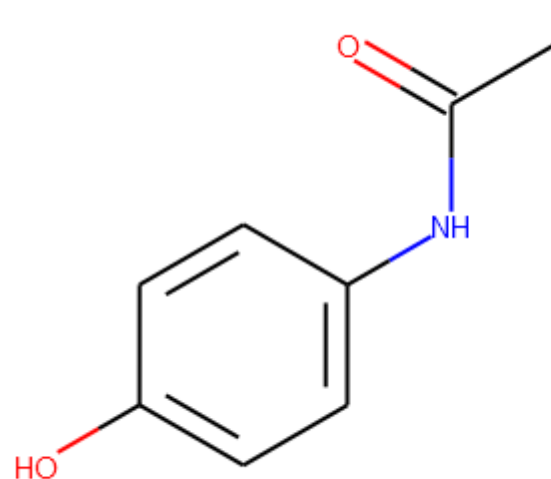


1917

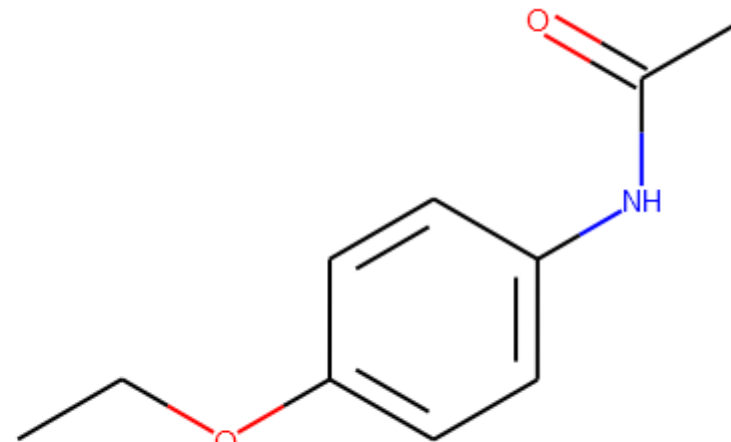
Morgan 2048-bit FingerPrint

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

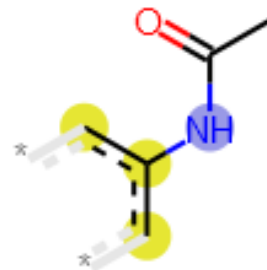
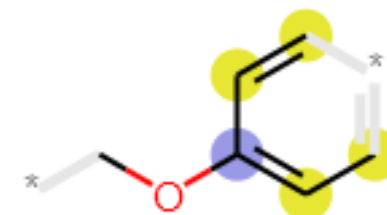


Paracetamol



Phenacetin

Phenacetin
unique



blue: the central atom in the environment

yellow: aromatic atoms

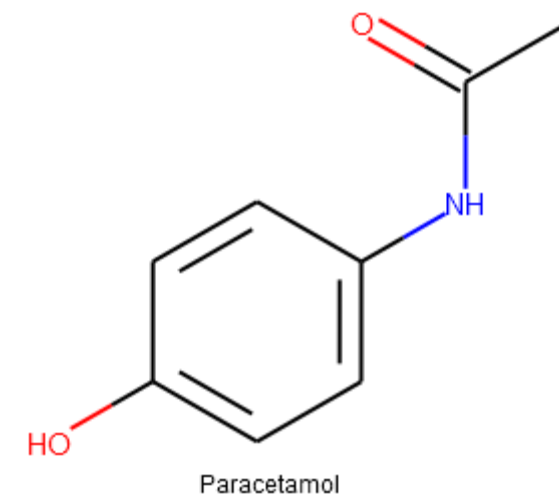
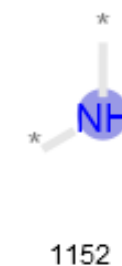
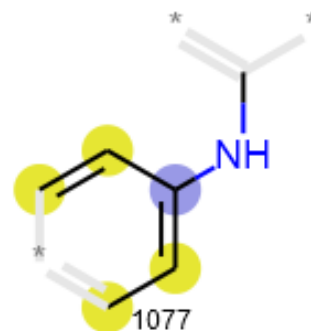
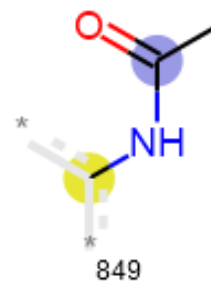
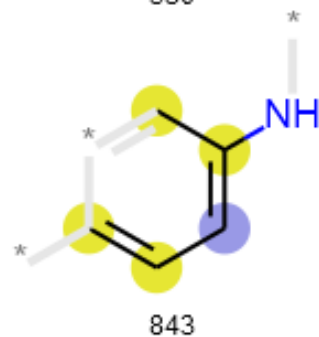
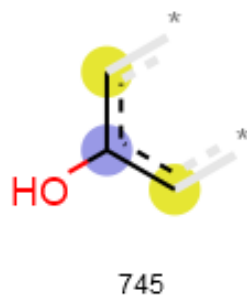
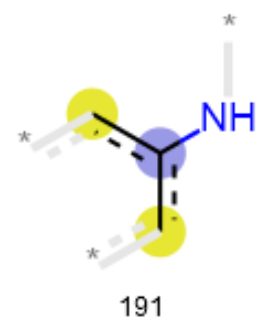
gray: aliphatic ring atoms

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]

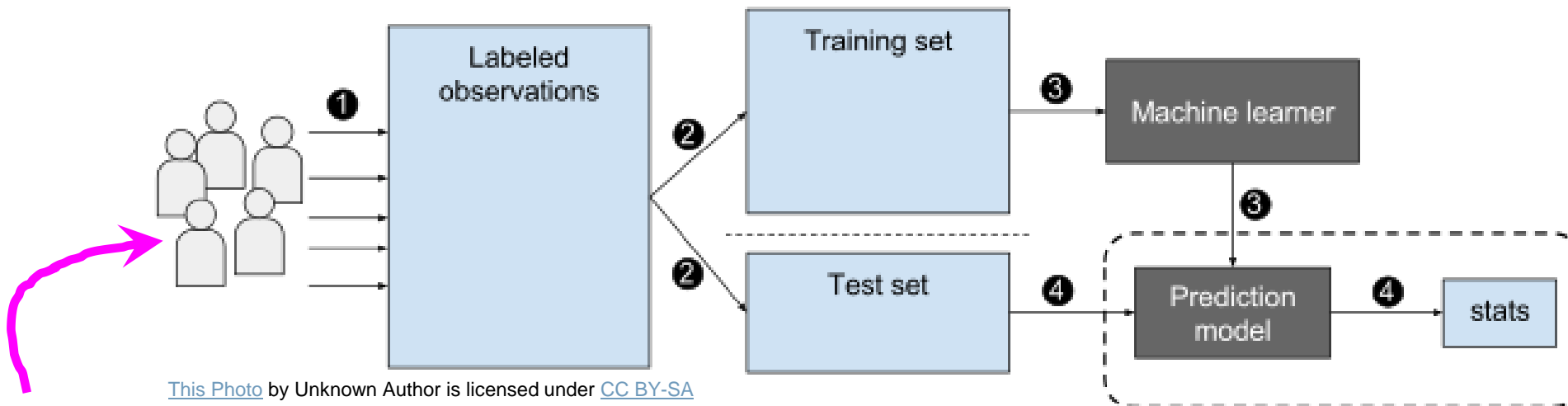


Tools Review

<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



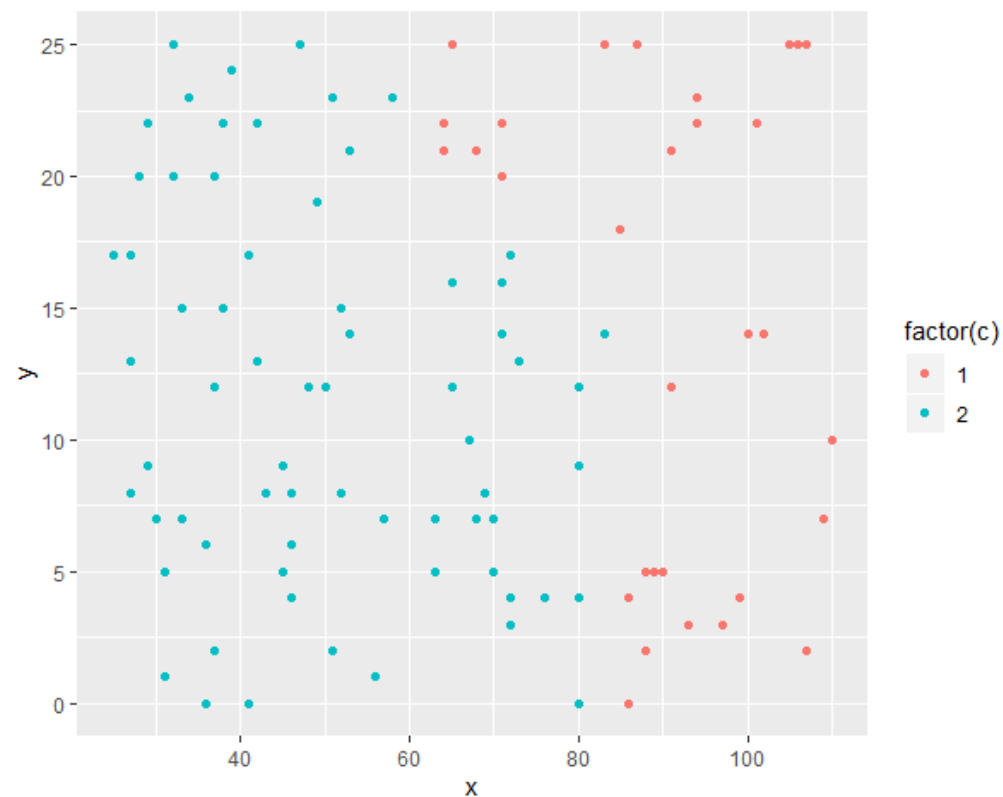
ID Bit1 Bit2 Bit3 Bit2048 OUTCOME

Machine-learner: Classification

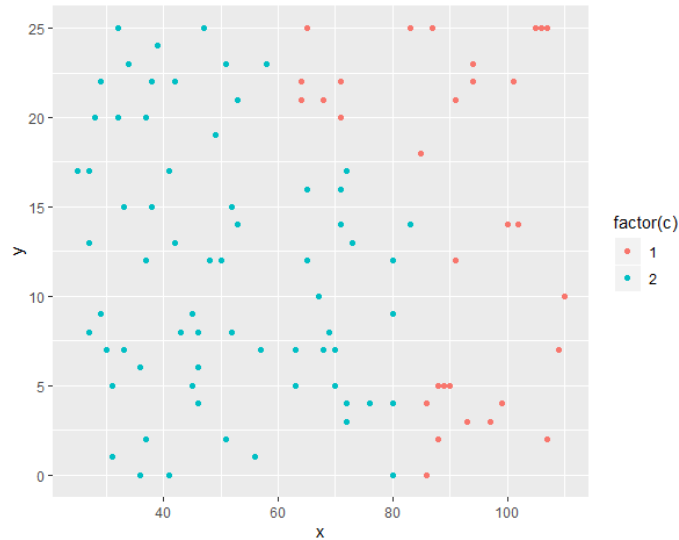
- Qualitative variables: unordered set, C ,
 - Eye-color $\in \{ \text{black, brown, blue} \}$
- Given a feature vector X (fingerprint) and a qualitative outcome Y (taking values from the set in C , the task is to identify a function $C(X)$ that predicts a value for Y
 - $C(X) \in C$
 - Takes in X and 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
...
...

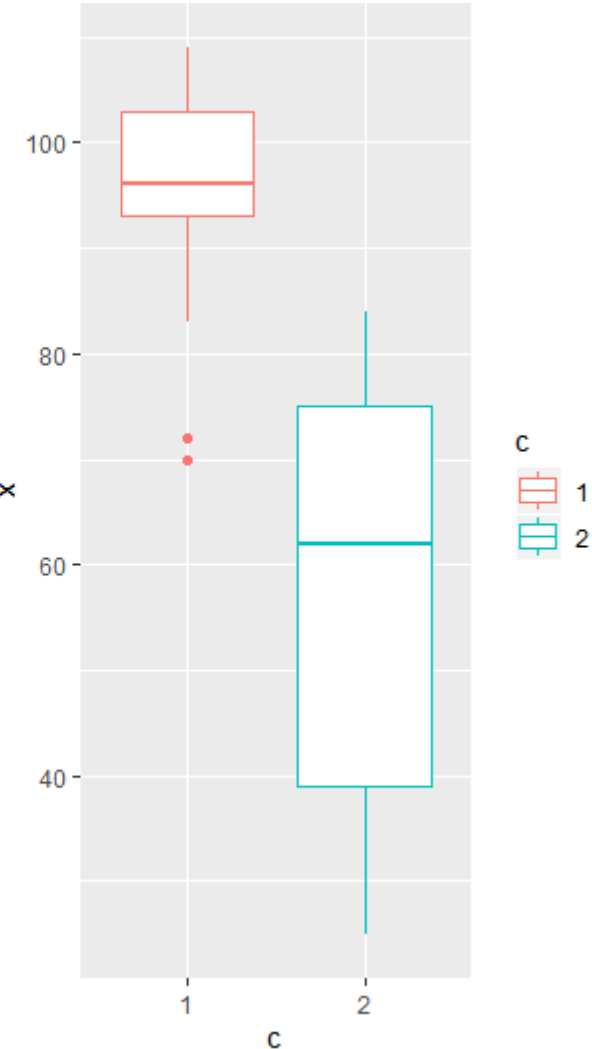
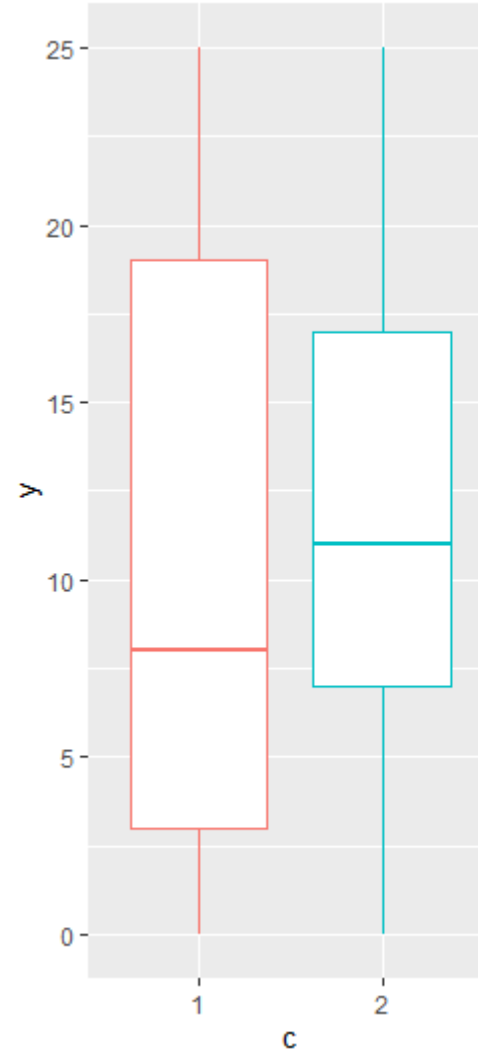


CART/binary-trees: Classification motivation



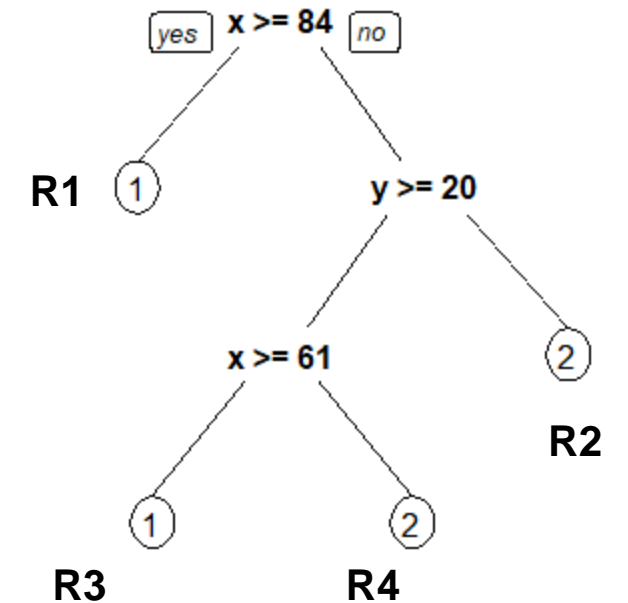
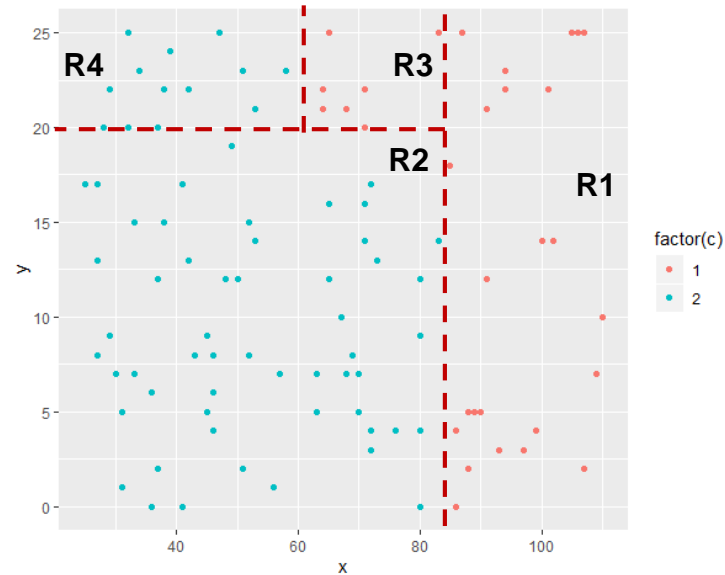
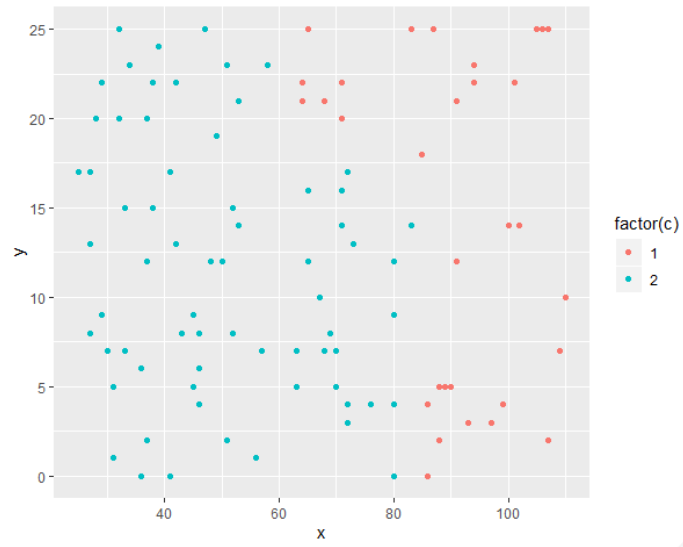
Split using single feature (x) and a cut-off ($t_x \geq 84$)

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



John Tukey (Box-plots)

Decision-trees: Classification motivation



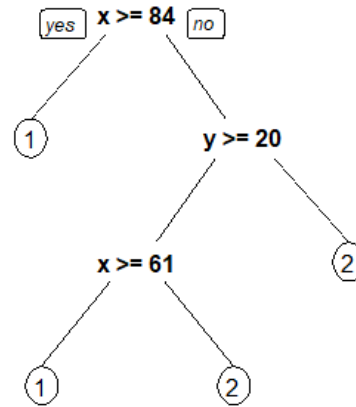
Terminal/Leaf nodes: R1, R2, R3

Internal nodes: Predictor space splitting points

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

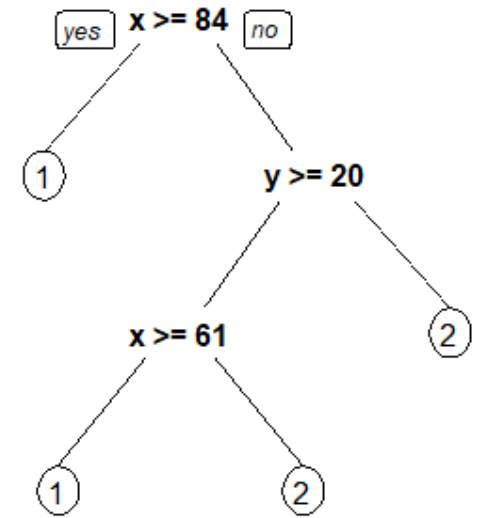
Decision-trees: Important points

- 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



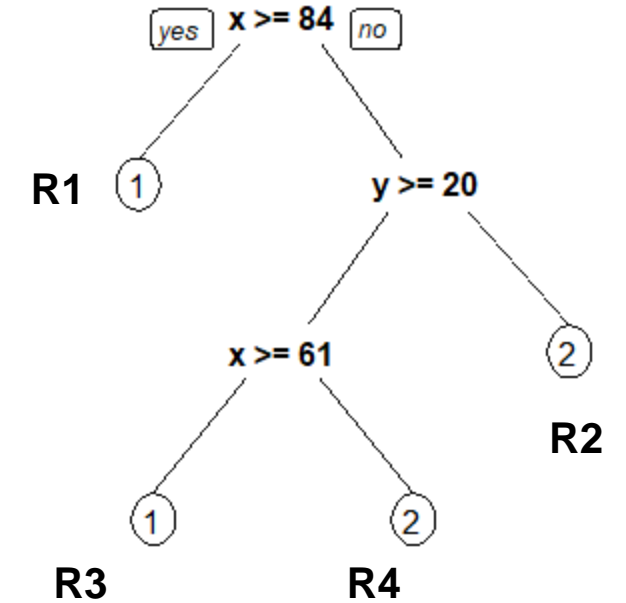
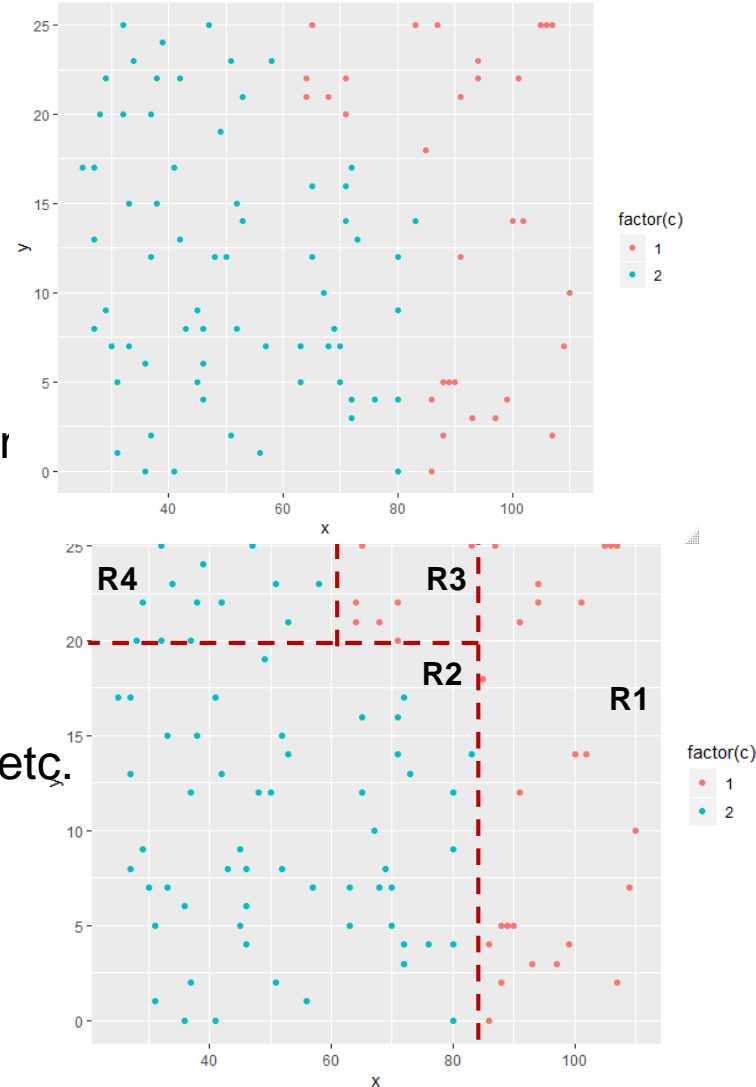
Decision-trees: Important points

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



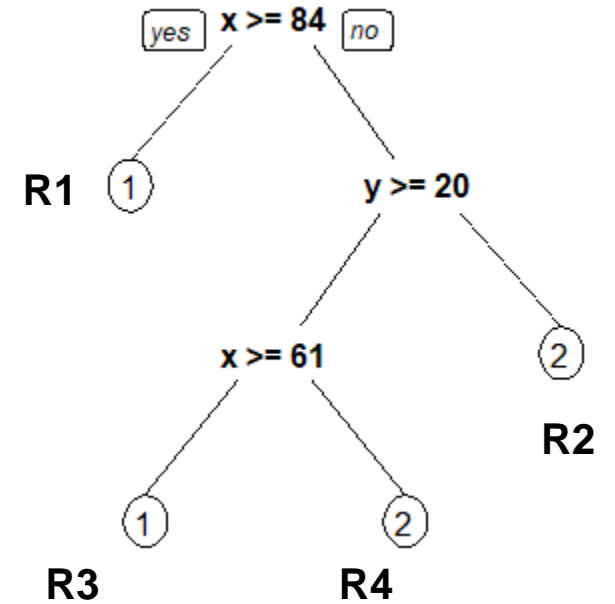
Hyperparameters

- Hyperparameters
 - Set before training a model
 - Drive the training process
 - Tuned between training iteration
- 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

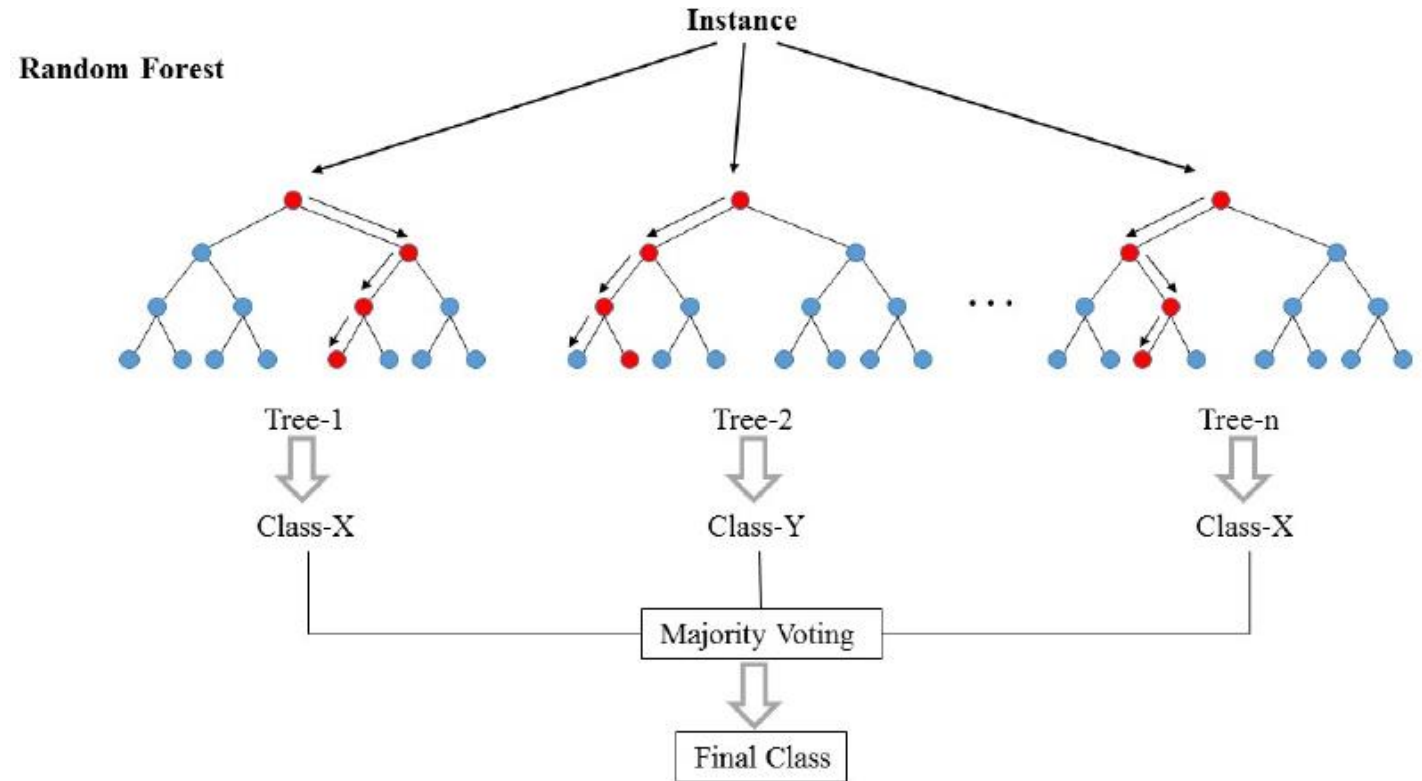
Law of Large Numbers

Wisdom of the crowd



[This Photo](#) by Unknown Author is licensed under [CC BY](#)

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



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Random Forest

- Combining a large number of trees result in improvements in accuracy
 - n independent measurements Z_1, \dots, 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 $\sqrt{(n_features)}$ considered when splitting nodes

Evaluation of binary classifiers

- Confusion Matrix and Balanced Accuracy (BA) Score
 - Count the number of times Class A is predicated as A or not (or other classes)

$$\text{TPR} = \frac{\text{TP}}{\text{P}}$$

$$\text{TNR} = \frac{\text{TN}}{\text{N}}$$

$$\text{BA} = \frac{\text{TPR} + \text{TNR}}{2}$$

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

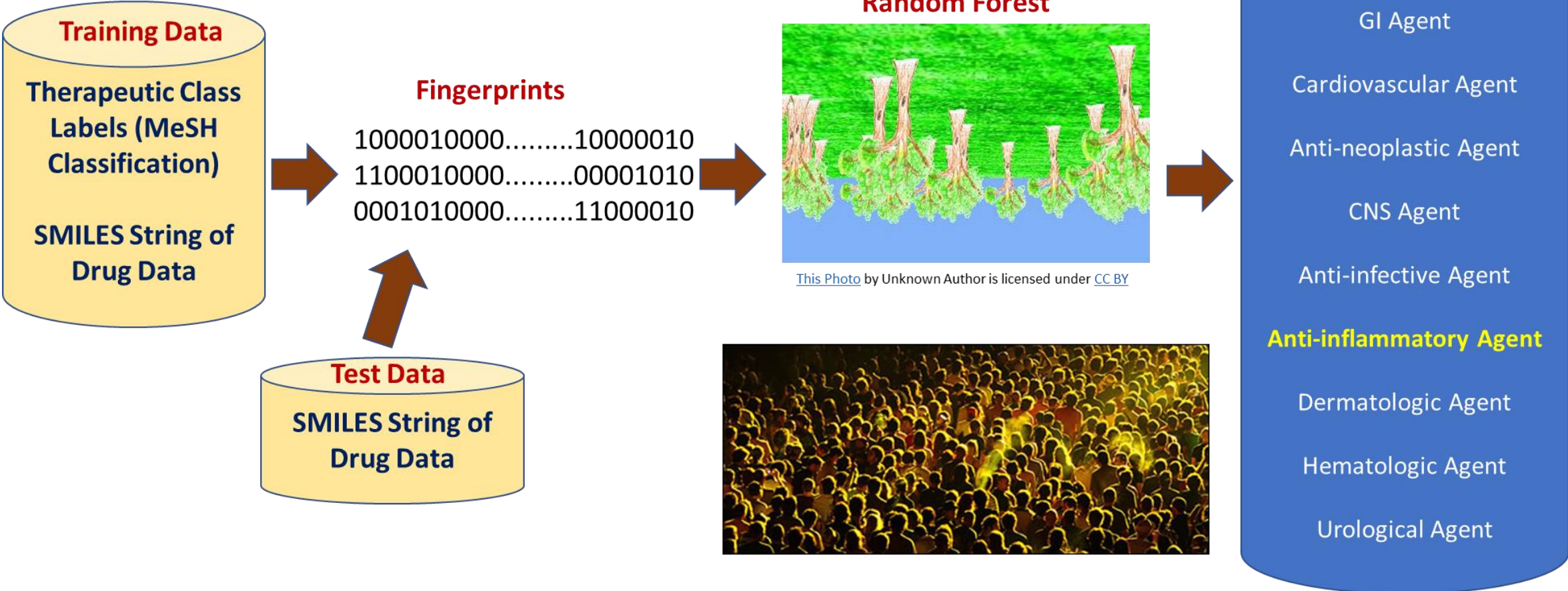
Accuracy = 0.9000

TPR = 0.9082

TNR = 0.5000

BA = 0.7041

Overview of Classification Process



Thanks

- Contact Info
 - ravichandrans@mail.nih.gov



[This Photo](#) by Unknown Author is licensed under [CC BY-SA-NC](#)

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

