

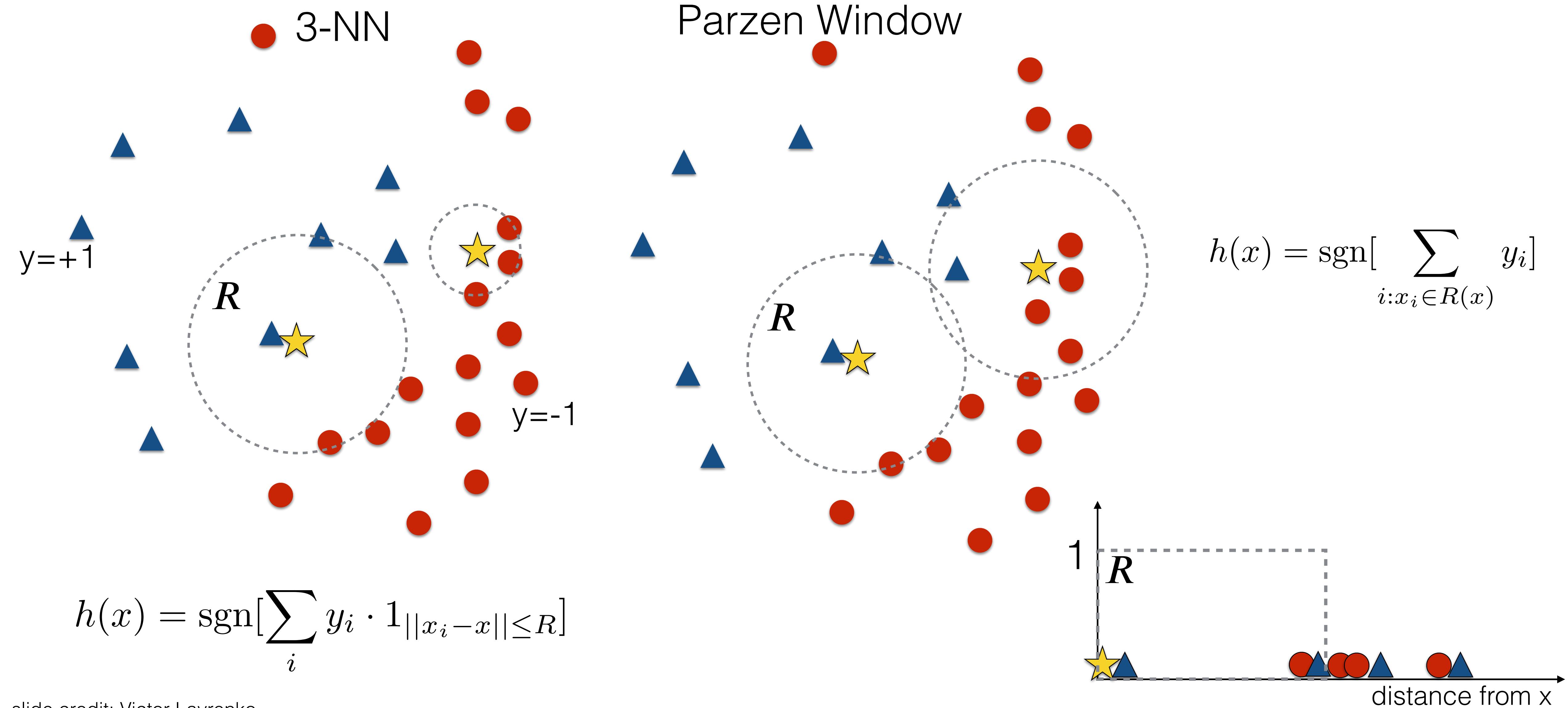
# Statistical Methods in AI (CSE 471)

## Lecture5: Decision Trees

Vineet Gandhi  
Centre for Visual Information Technology (CVIT)



# kNN, Parzen Windows and Kernels



# Decision Trees

- One of the most intuitive classifiers
- Easy to understand and construct
- Surprisingly, also works very (very) well

# Decision Trees

File Edit View Favorites Tools Help

Back Search Favorites Media Mail Answers Fantasy Sports Hockey Music Personals

Address http://www.sfgate.com/cgi-bin/article.cgi?f=/n/a/2006/10/01/financial/f210530D02.DTL Go Links

Y! netscape webmail Search Web Upgrade Now! Mail My Yahoo! Answers Fantasy Sports Hockey Music Personals

Google Go Bookmarks 5 blocked Check AutoLink AutoFill Send to netflix million prize Settings

**Unlimited DVD's Delivered**  
**FREE In-Store Rentals**

Starting at only **\$9.99** per month  
2 WEEKS FREE! GO

**SFGate.com**

[SFGATE HOME](#) • [BUSINESS](#) • [SPORTS](#) • [ENTERTAINMENT](#) • [TRAVEL](#)      [JOBS](#) • [REAL ESTATE](#) • [CARS](#)

Search SFGate News Web by Google™

**AP Breaking News**

**Netflix offers \$1 million prize for better movie recommendations**

By MICHAEL LIEDTKE, AP Business Writer  
Sunday, October 1, 2006

(10-01) 21:05 PDT San Francisco (AP) --

Online DVD rental pioneer Netflix Inc. wants recommendations on how to improve its movie-recommendation system so badly that it's dangling a \$1 million reward as an incentive.

The prize, offered in a contest scheduled to begin Monday, is part of Netflix's effort to sharpen its competitive edge as it continues a bitter duel with Blockbuster Inc. and prepares for an anticipated onslaught of services that make it easier to download movies on to computer hard drives.

[Printable Version](#)  
[Email This Article](#)

**Business & Finance**

Get Quote:  
Detailed  
Submit  
[Symbol Lookup](#)

**Commission on the Regulation of U.S. Capital Markets in the 21st Century**  
Examining the Competitive Environment in the Global Marketplace: A Town Hall Meeting

**THURSDAY, OCTOBER 12**  
The City Club of San Francisco

**REGISTER TODAY**

**ChronicleJobs TOP JOBS**

► **NURSING**  
Clinical Application

► **PLANNING**  
MTC Metropolitan

**HAPPENING SAN FRANCISCO**

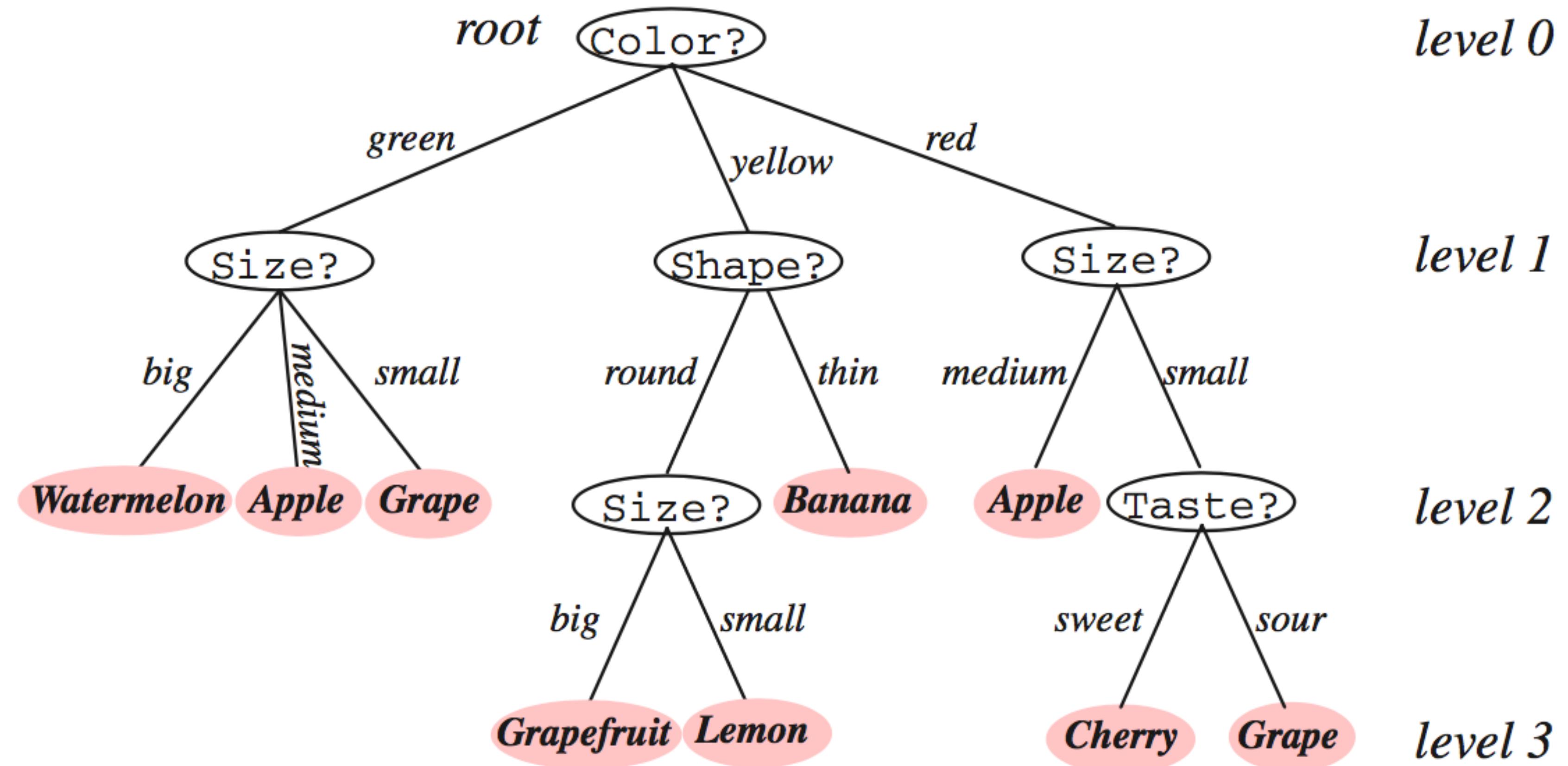
Error on page. Internet

# Decision Trees

Attributes (features)					Label
Movie	Type	Length	Director	Famous actors	Liked?
m1	Comedy	Short	Adamson	No	Yes
m2	Animated	Short	Lasseter	No	No
m3	Drama	Medium	Adamson	No	Yes
m4	animated	long	Lasseter	Yes	No
m5	Comedy	Long	Lasseter	Yes	No
m6	Drama	Medium	Singer	Yes	Yes
m7	animated	Short	Singer	No	Yes
m8	Comedy	Long	Adamson	Yes	Yes
m9	Drama	Medium	Lasseter	No	Yes

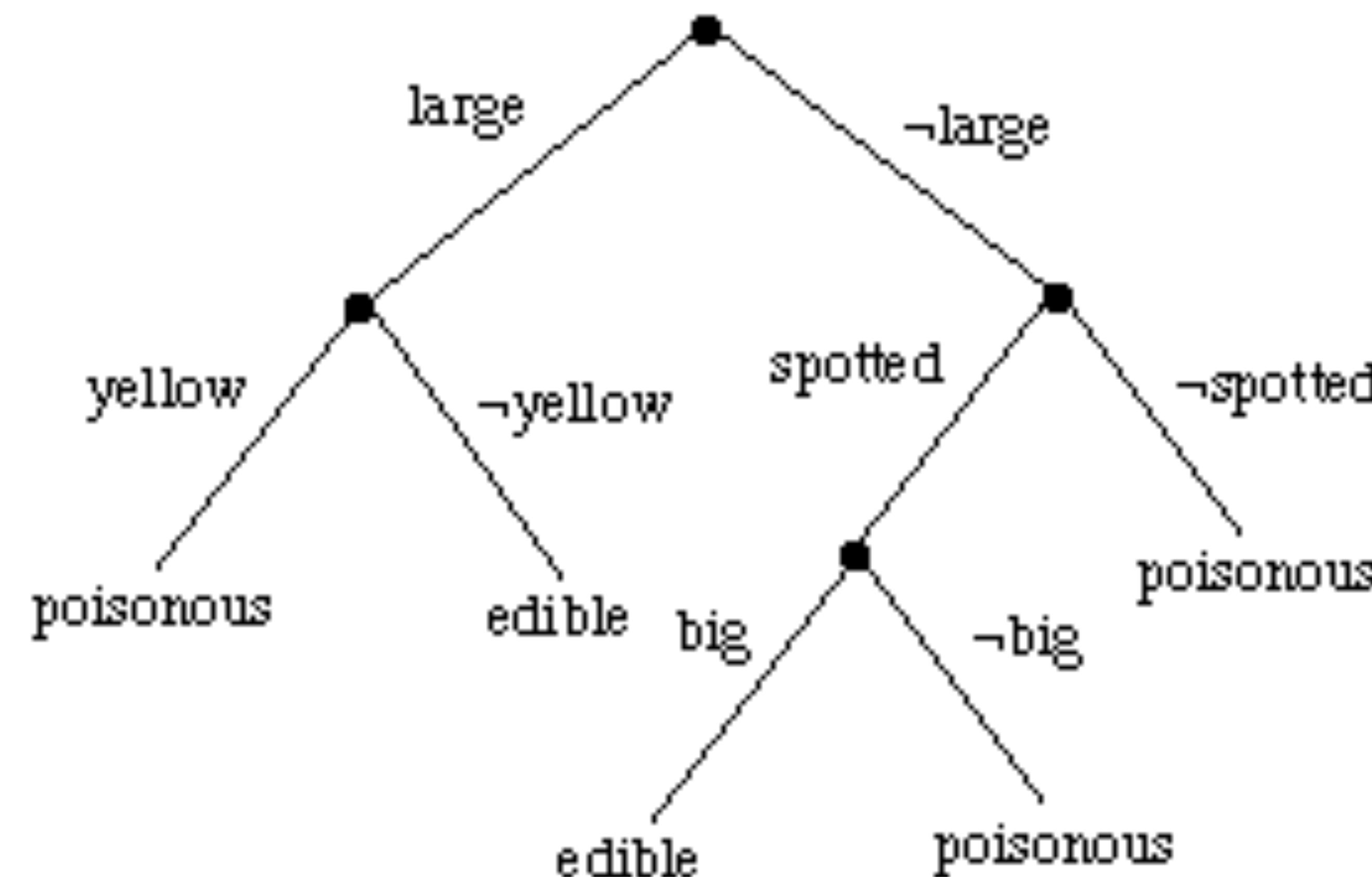
100,480,507 ratings that  
480,189 users gave to  
17,770 movies

# Decision Trees

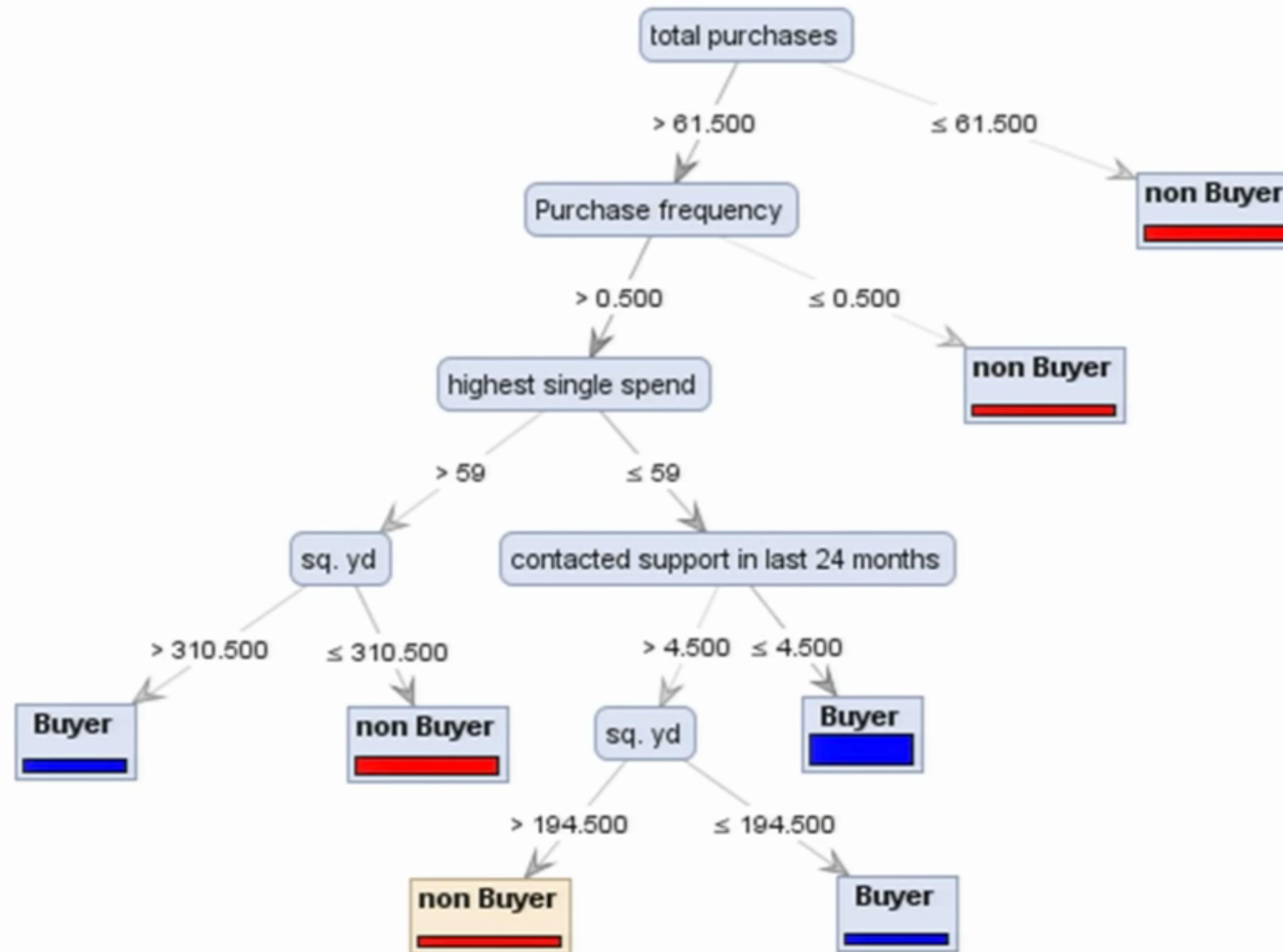


fruit described using 4 tuple: {color, shape, taste, size}

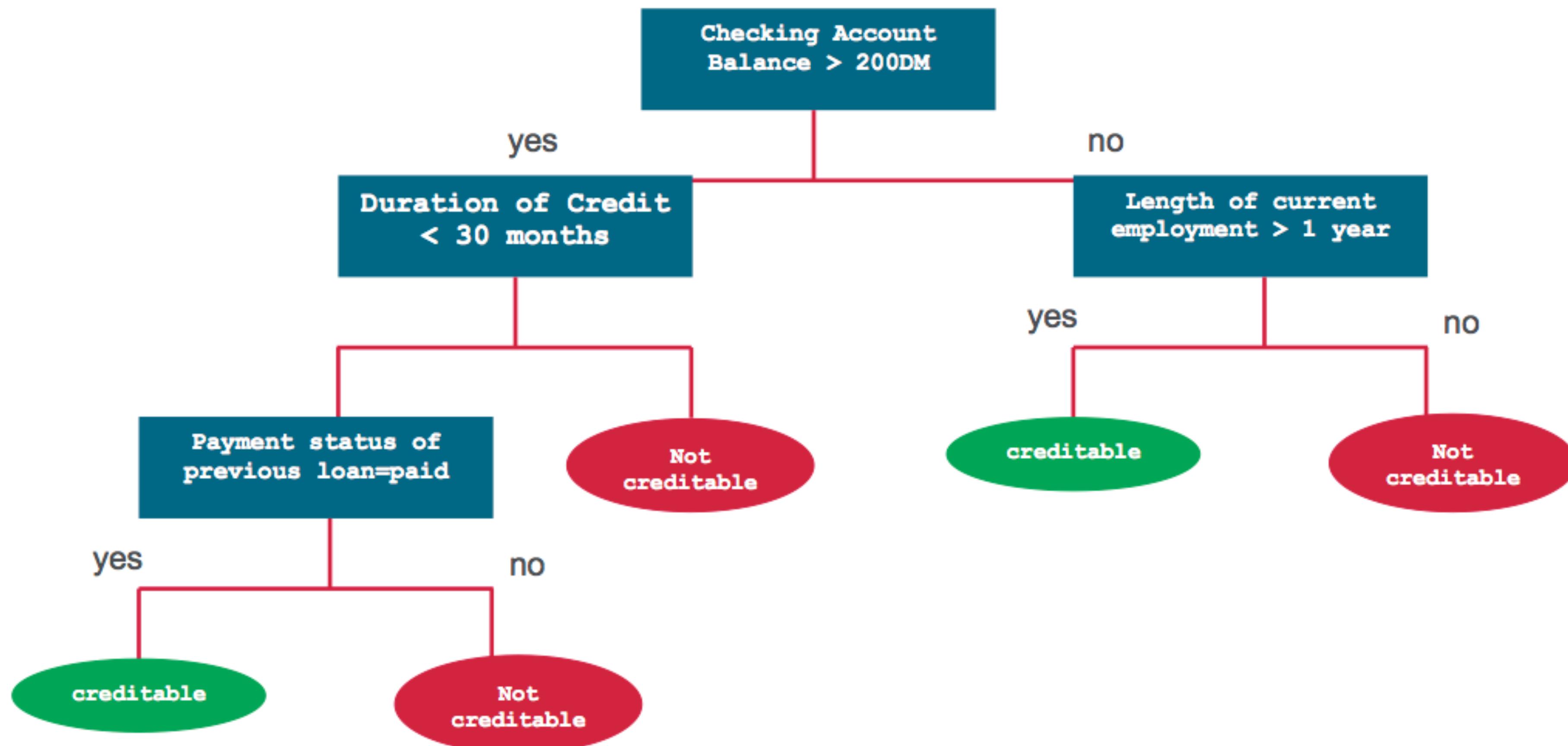
# Edible Mushroom



# Identify buyers



# Credit approval



# Vampire or not

<b>Vampire?</b>	<b>Shadow?</b>	<b>Garlic?</b>	<b>Complexion?</b>	<b>Accent?</b>
No	?	Yes	Pale	None
No	Yes	Yes	Ruddy	None
Yes	?	No	Ruddy	None
Yes	No	No	Average	Heavy
Yes	?	No	Average	Odd
No	Yes	No	Pale	Heavy
No	?	Yes	Ruddy	Odd
No	Yes	No	Average	Heavy

# Vampire or not

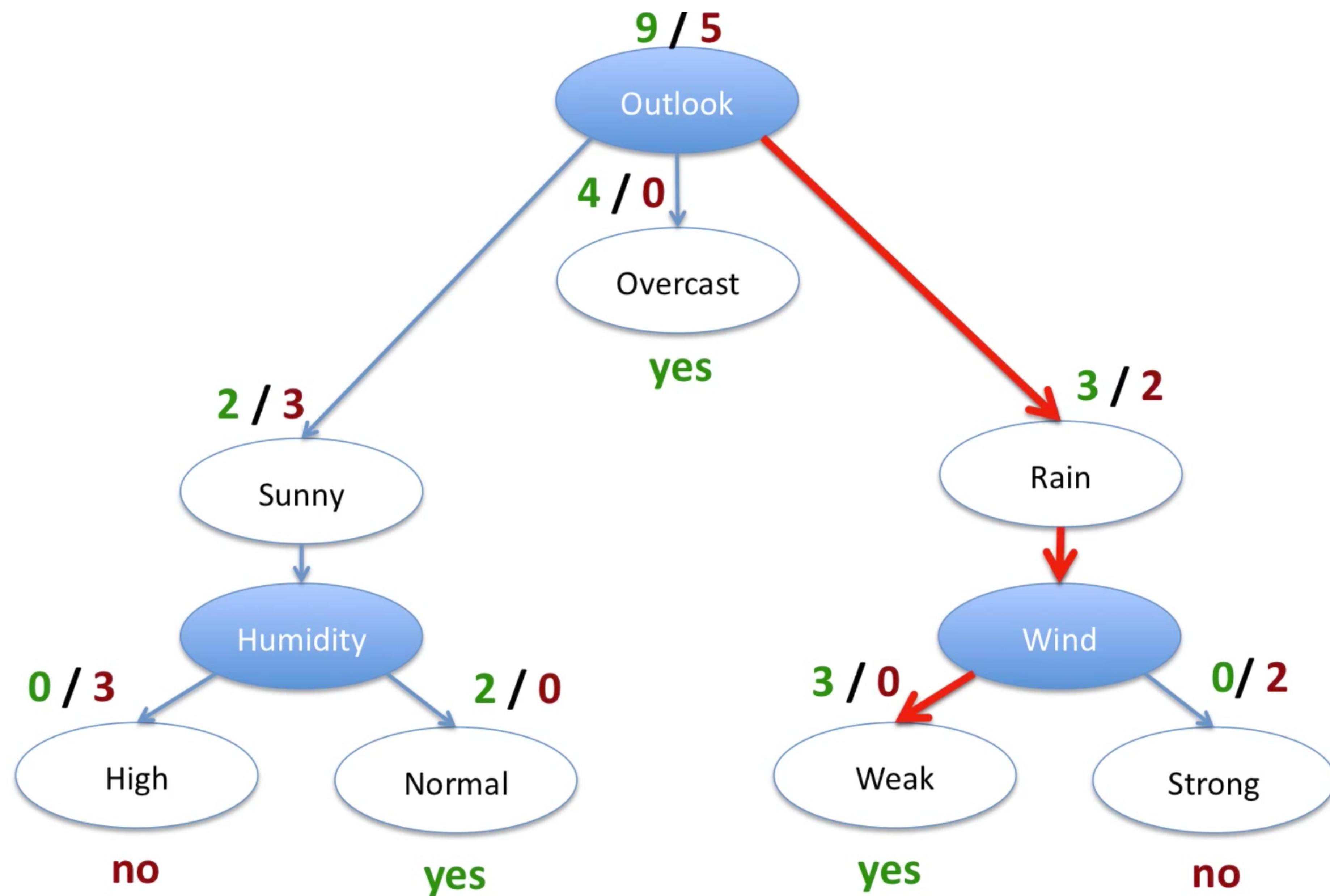
<b>Vampire?</b>	<b>Shadow?</b>	<b>Garlic?</b>	<b>Complexion?</b>	<b>Accent?</b>
No	?	Yes	Pale	None
No	Yes	Yes	Ruddy	None
Yes	?	No	Ruddy	None
Yes	No	No	Average	Heavy
Yes	?	No	Average	Odd
No	Yes	No	Pale	Heavy
No	?	Yes	Ruddy	Odd
No	Yes	No	Average	Heavy

dataset credit: "Artificial Intelligence", Patrick Winston

# Another example

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

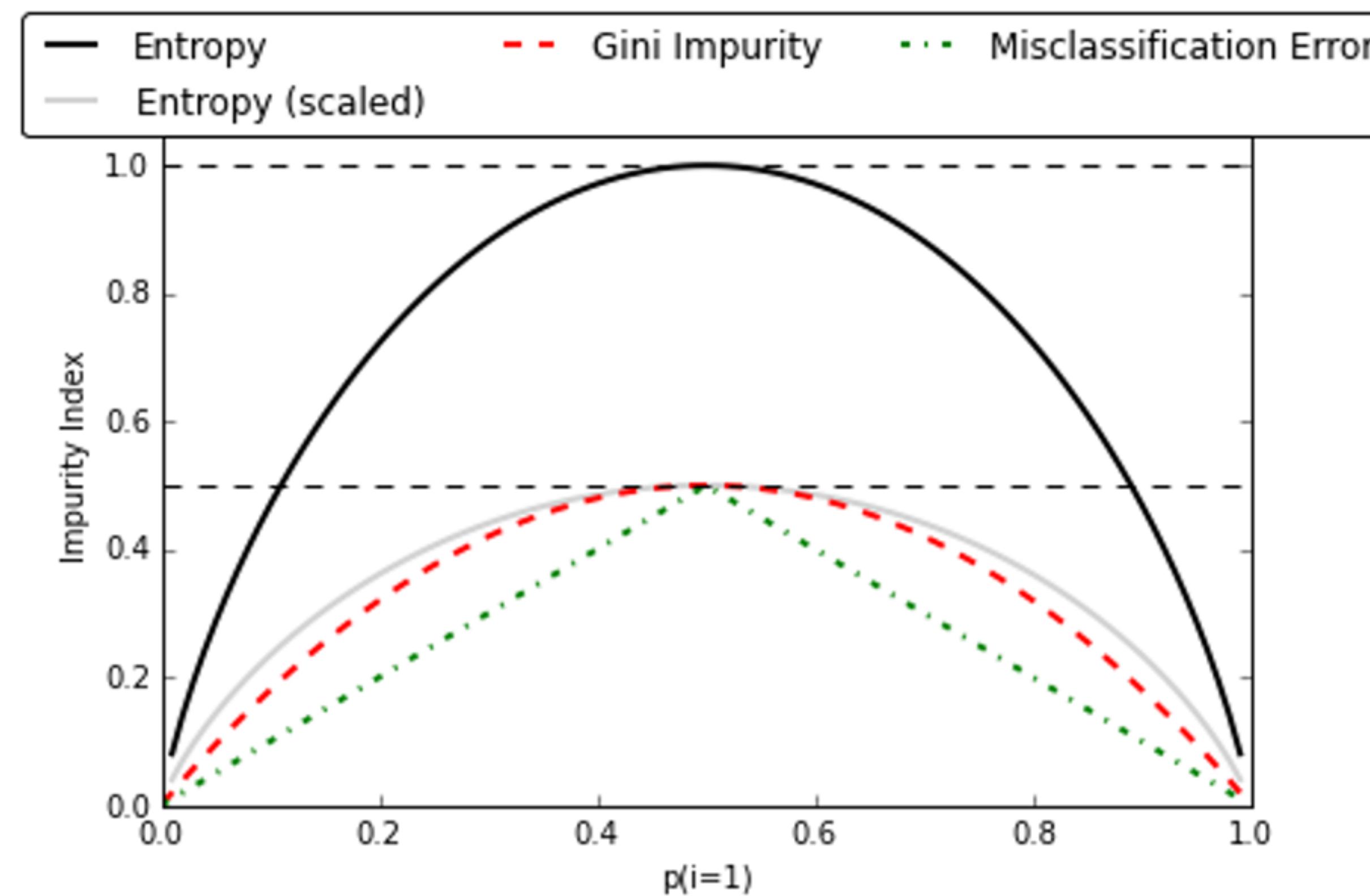


# Impurity functions

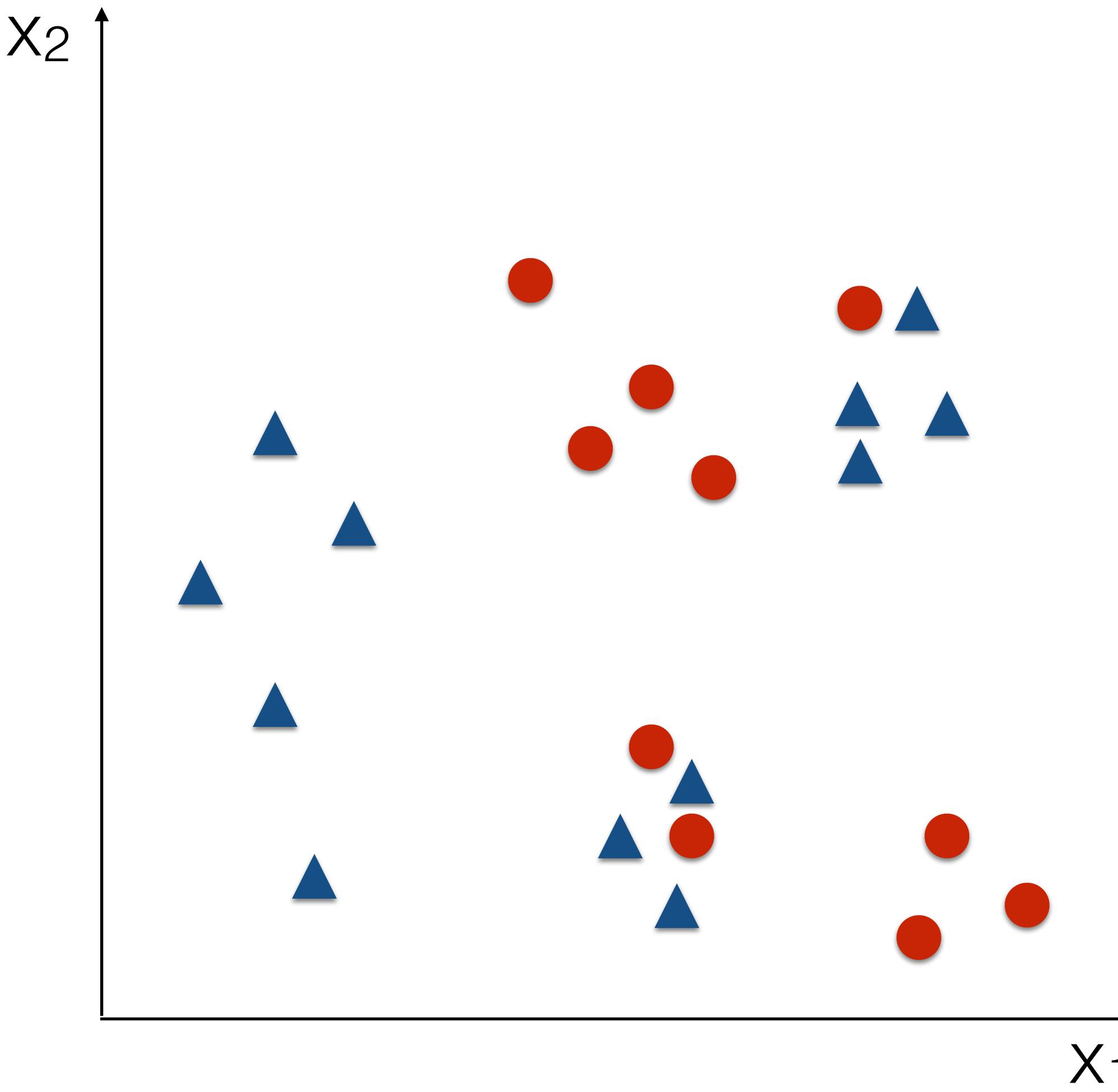
**Entropy:**  $i(V) = -(q \log q + (1 - q) \log(1 - q))$

**Gini index:**  $i(V) = 2q(1 - q)$

**Misclassification rate:**  $i(V) = \min(q, 1 - q)$

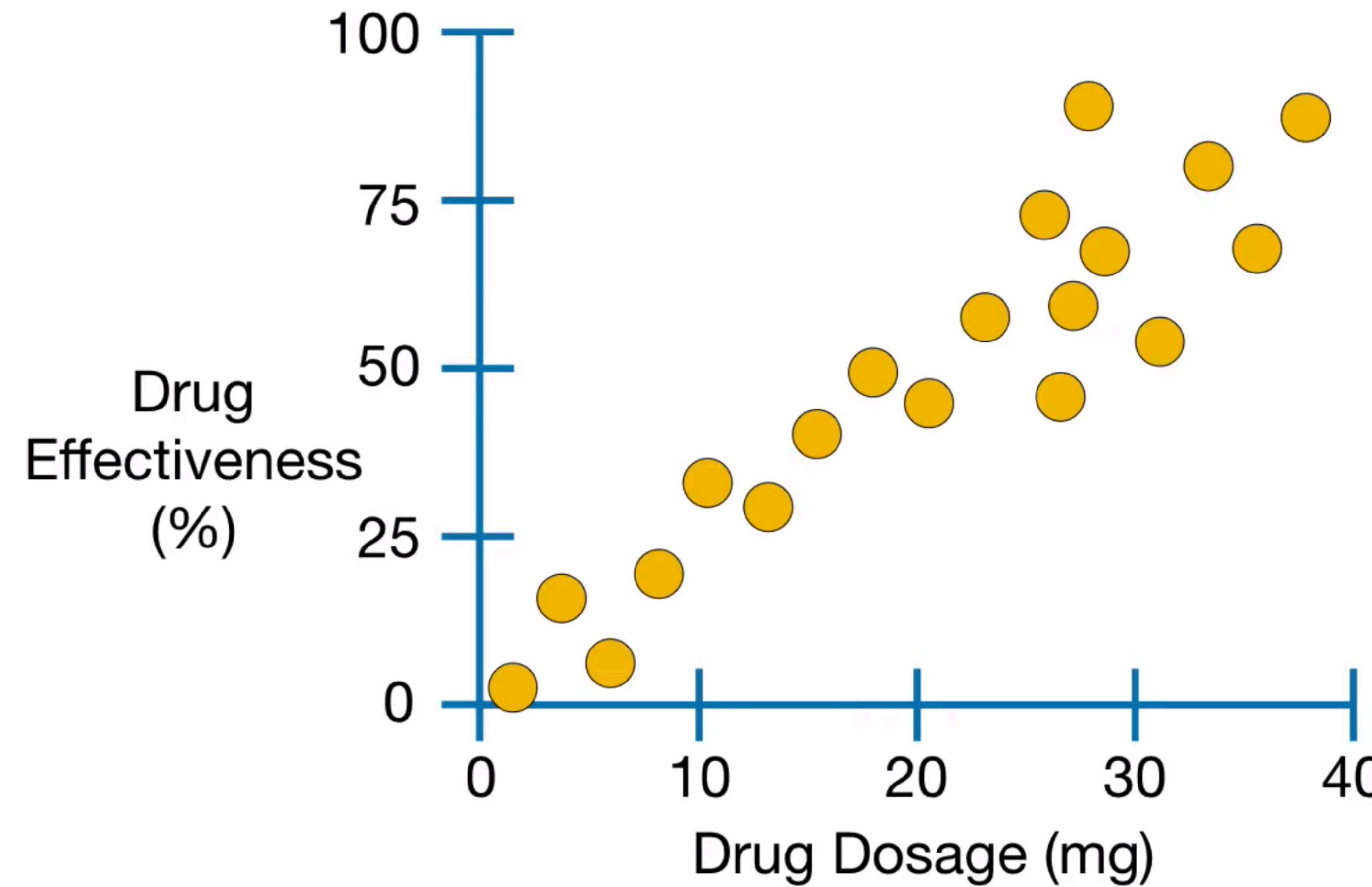


# Numeric data?

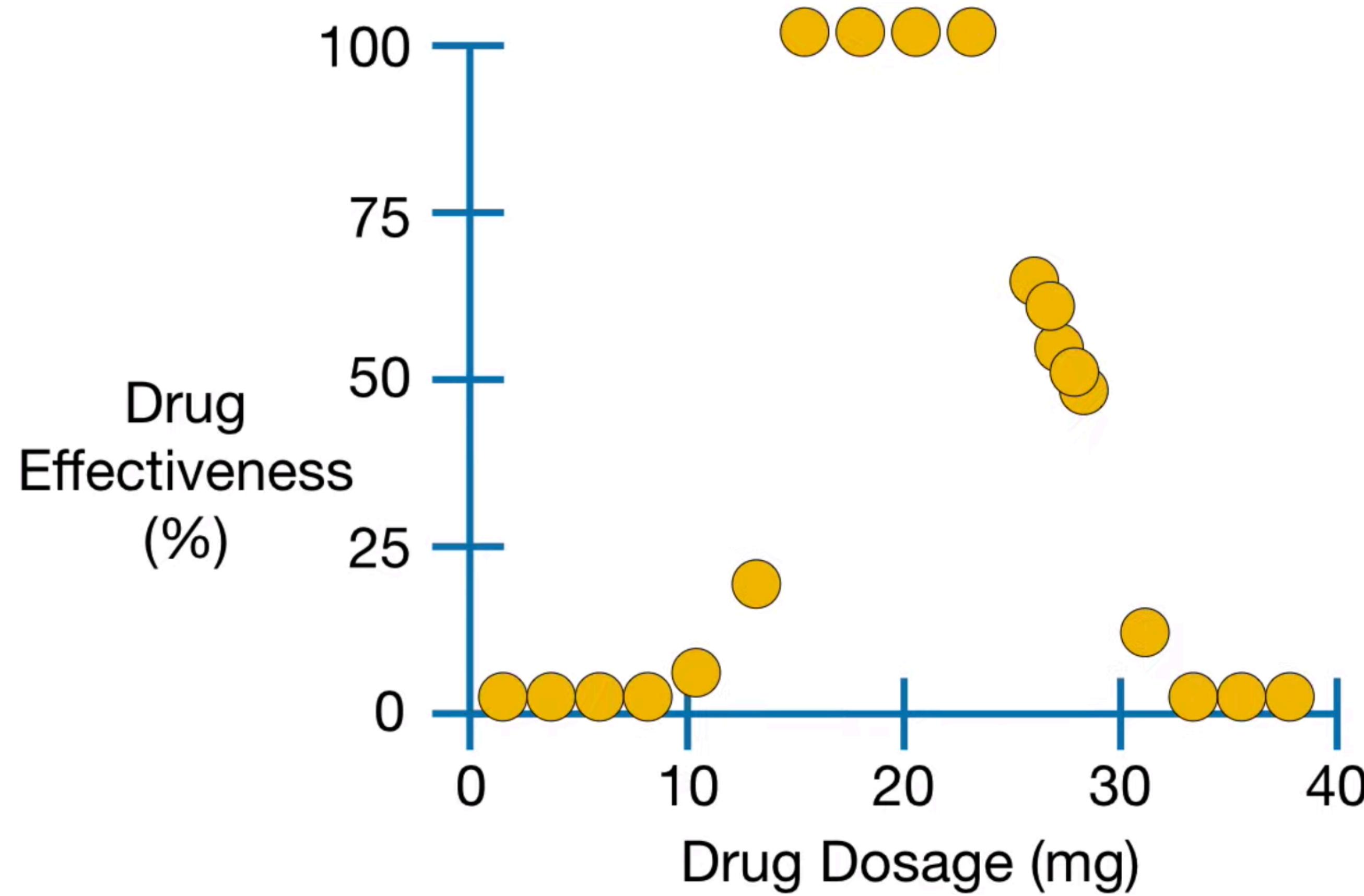


# Regression

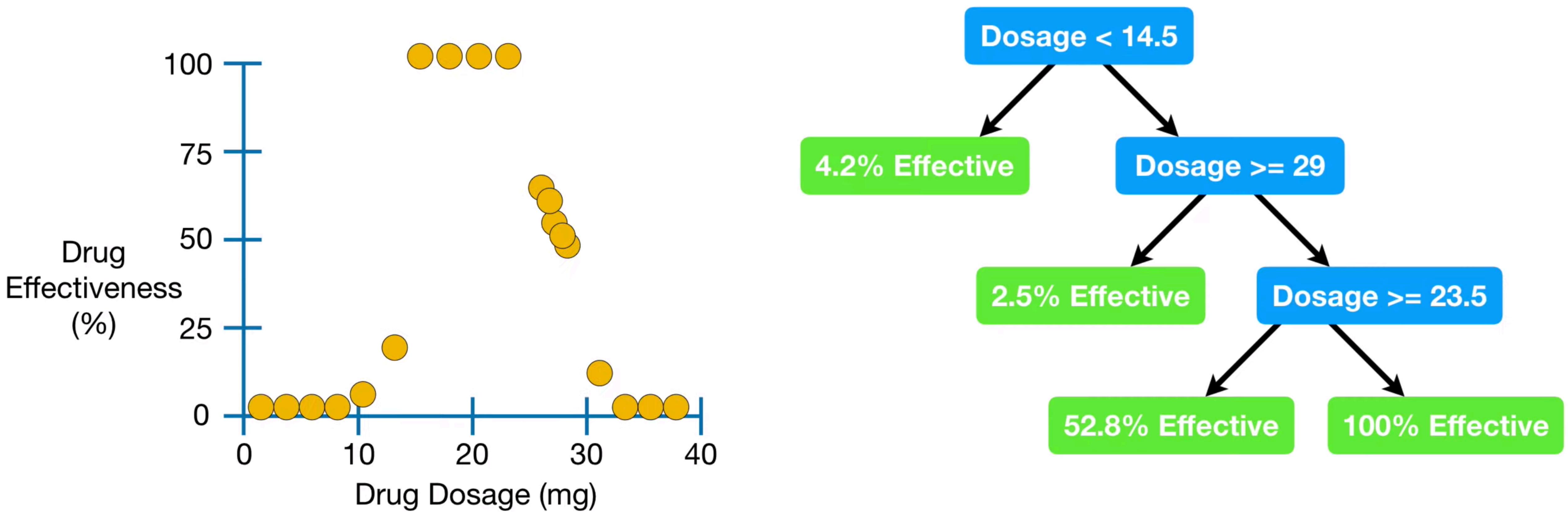
- Drug effectiveness vs dosage



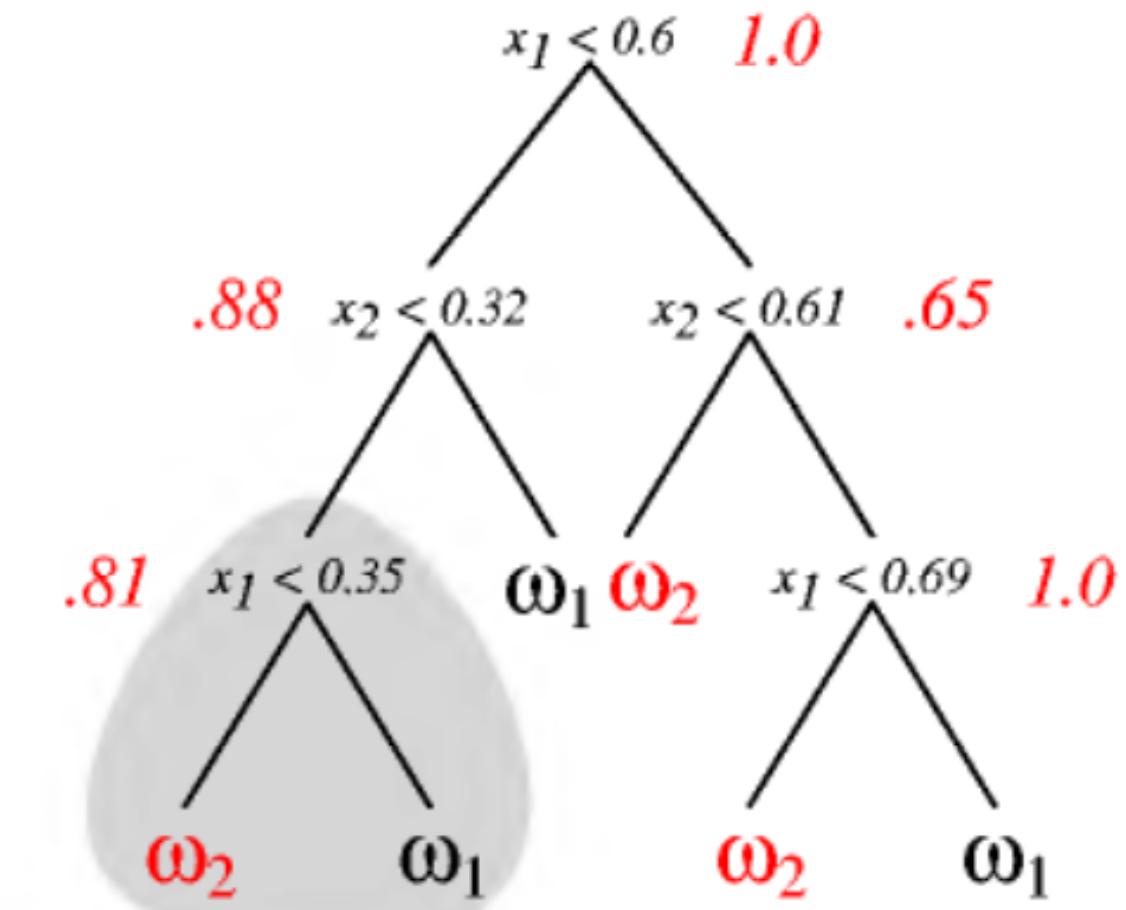
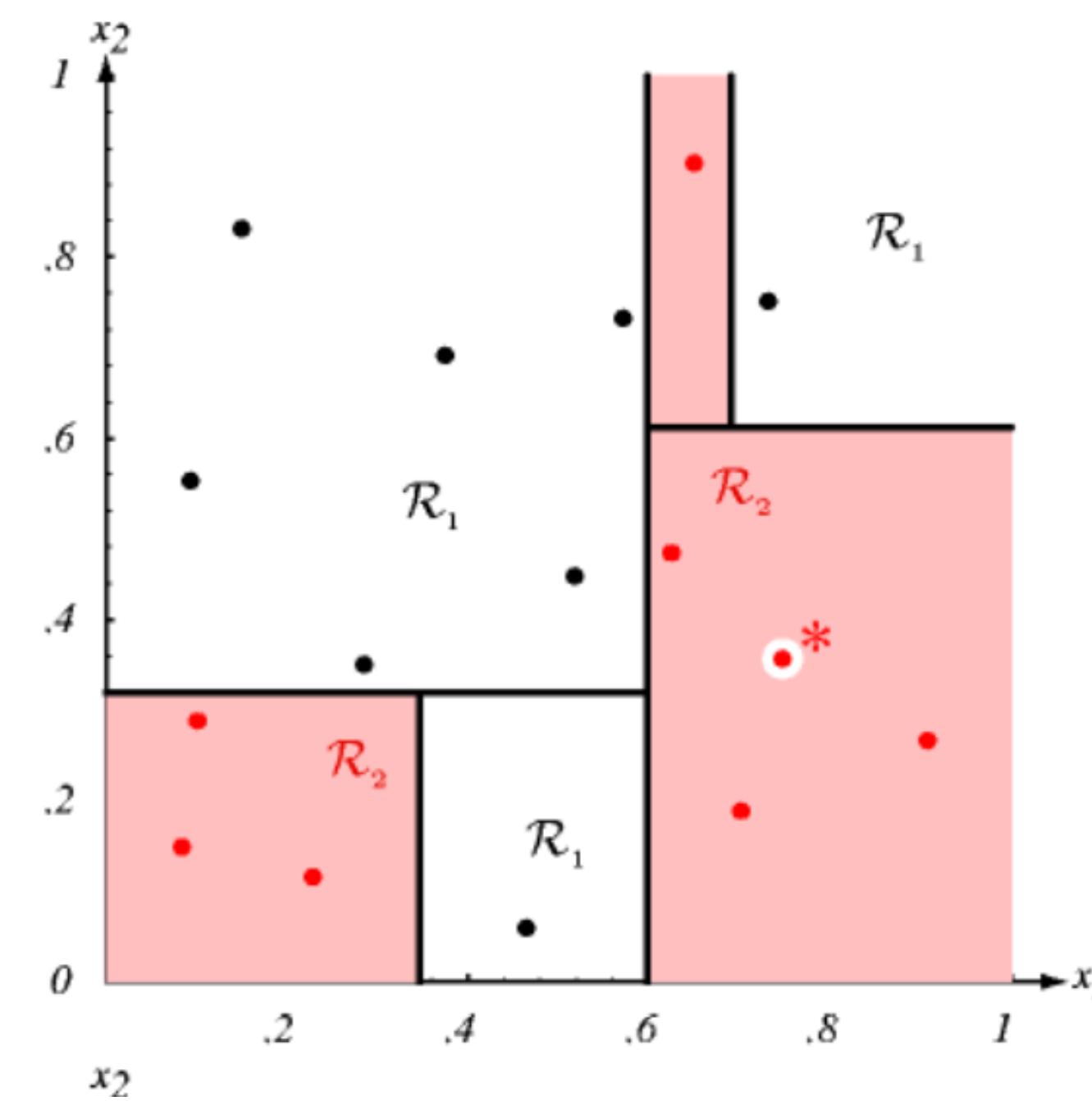
# Regression



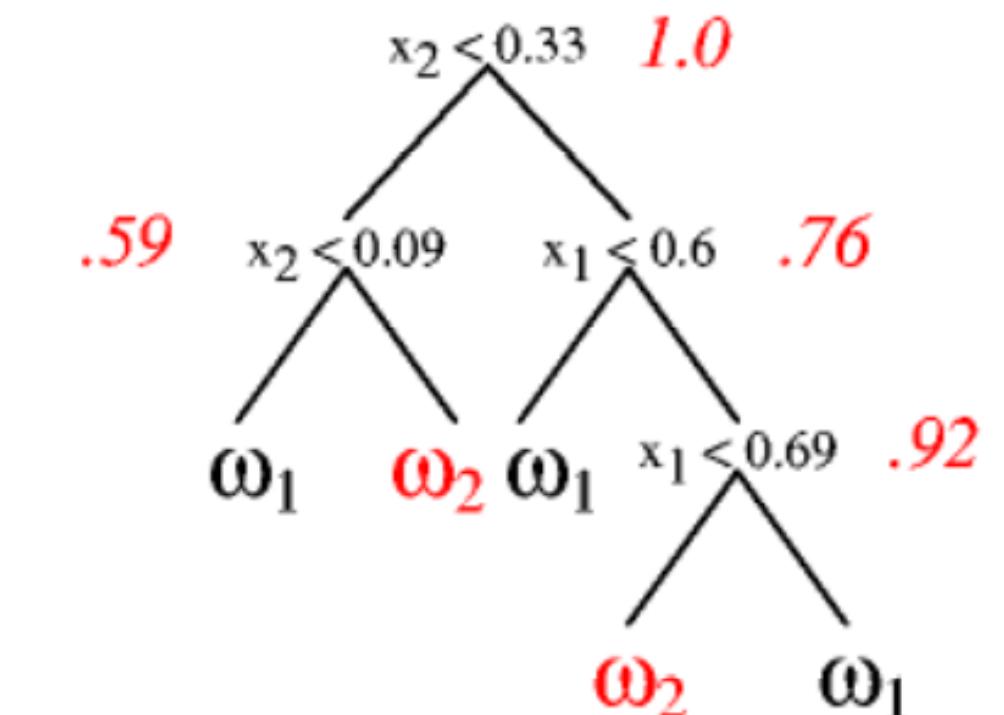
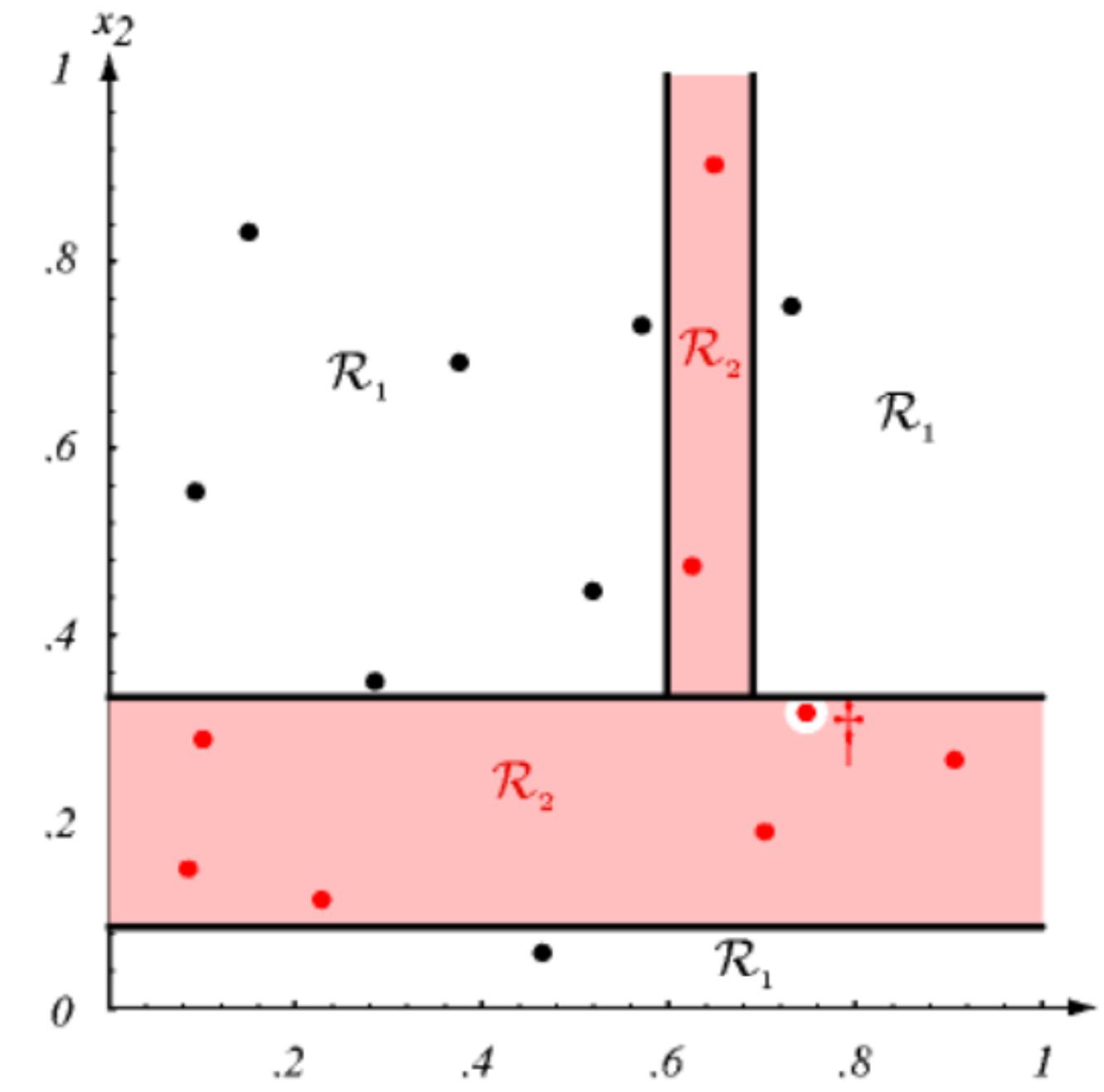
# Regression



$\omega_1$ (black)		$\omega_2$ (red)	
$x_1$	$x_2$	$x_1$	$x_2$
.15	.83	.10	.29
.09	.55	.08	.15
.29	.35	.23	.16
.38	.70	.70	.19
.52	.48	.62	.47
.57	.73	.91	.27
.73	.75	.65	.90
.47	.06	.75	.36* (.32 <sup>†</sup> )



The instability is consequence of the discrete and greedy nature of tree creation



# Random Forest

- Grow K different Trees
  - pick a random subset of training examples
  - pick d random attributes (compute gain based on subset instead of full set)
  - repeat K times