

Memorial Sloan Kettering
Cancer Center

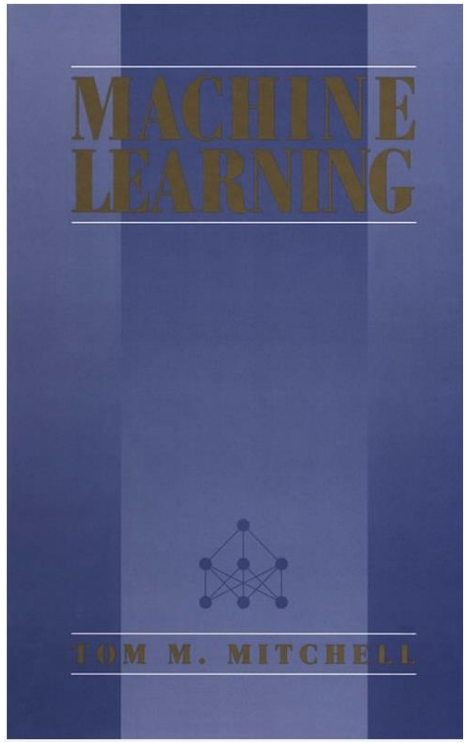
Lecture 1: Statistical Methods Foundations I

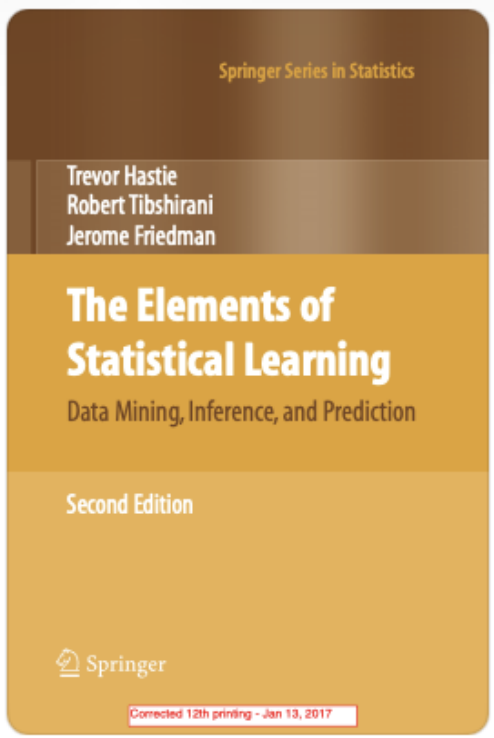
Harini Veeraraghavan, PhD
Director, AI for Image Guided Therapies
Memorial Sloan Kettering Cancer Center

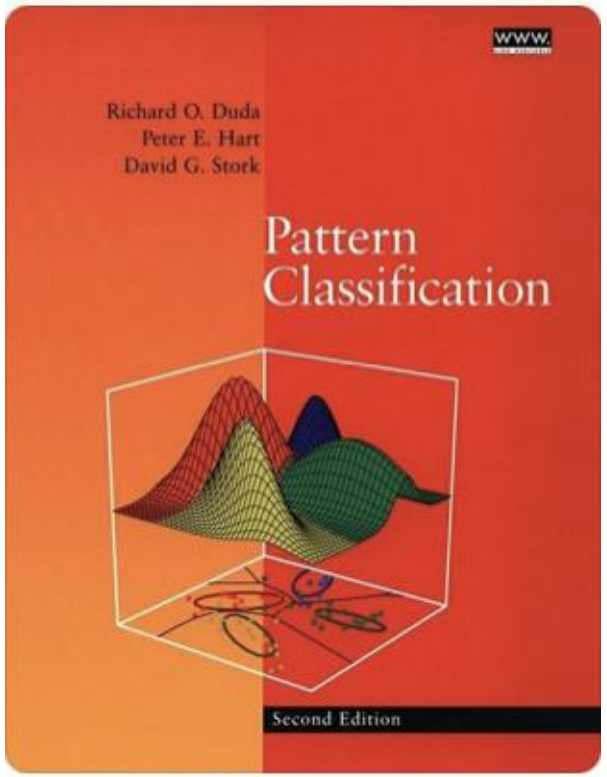
OUTLINE

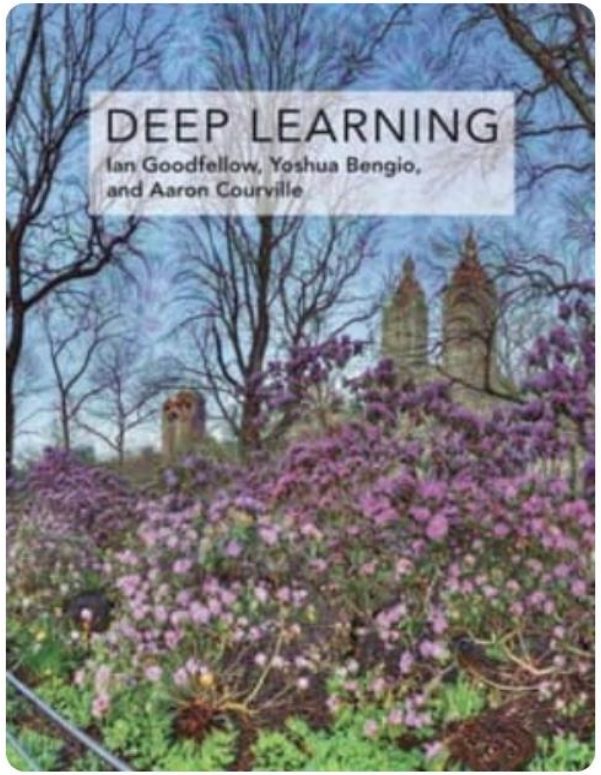
- Data types and distributions (Binomial, Normal)
- Descriptive statistics:
  - Mean, median, standard deviation, skewness, kurtosis
- Samples vs. populations; estimators and bias
- Bayesian learning
- Confidence intervals brief introduction

Books that might be helpful references in this course if you are more curious





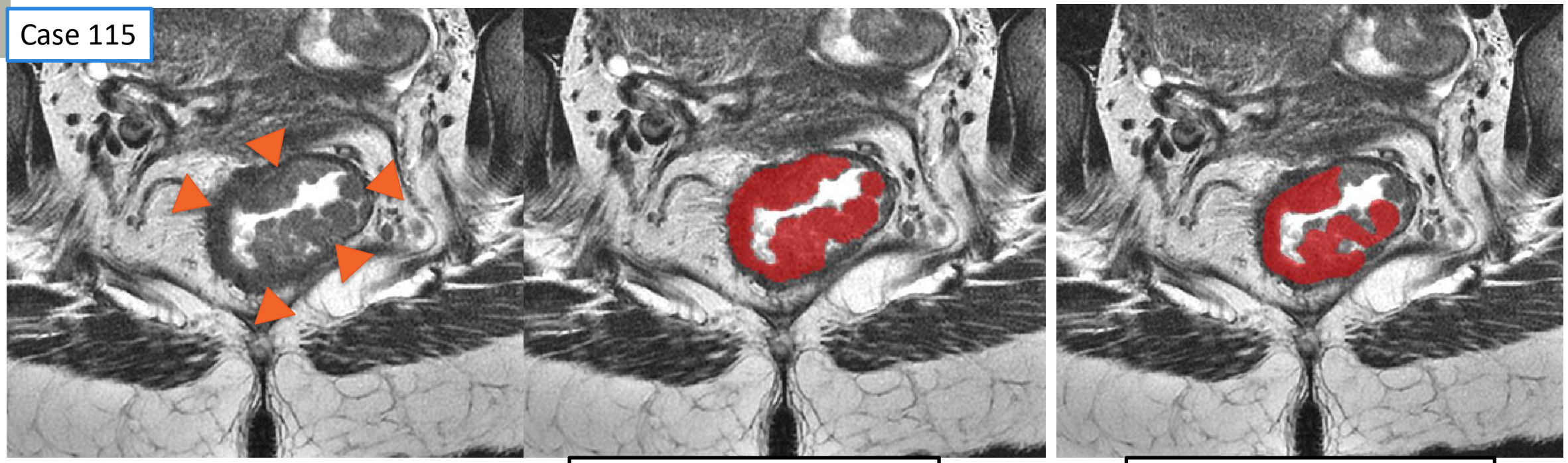




You don’t have to read these books – the lectures are roughly based on these

Why do we need to understand distributions?

Which one of these two tumor segmentation models is more accurate?



Magnetic resonance image with rectal cancer

Model A (SWIFTe)

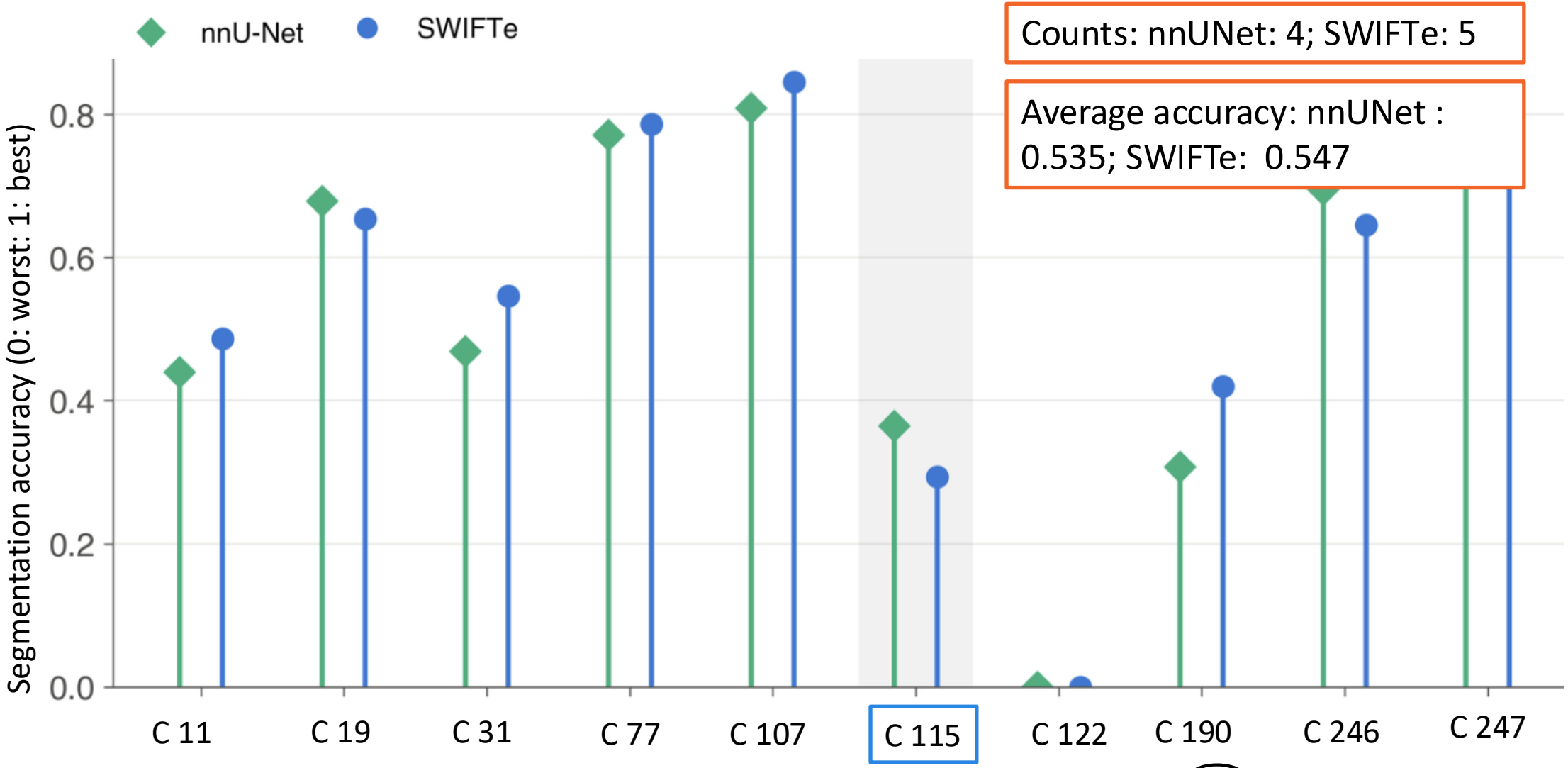
Model B (nnUNet)

SWIFTe: Swin Transformer with efficient Fine Tuning

nnUNet: no new U-Net

Memorial Sloan Kettering Cancer Center

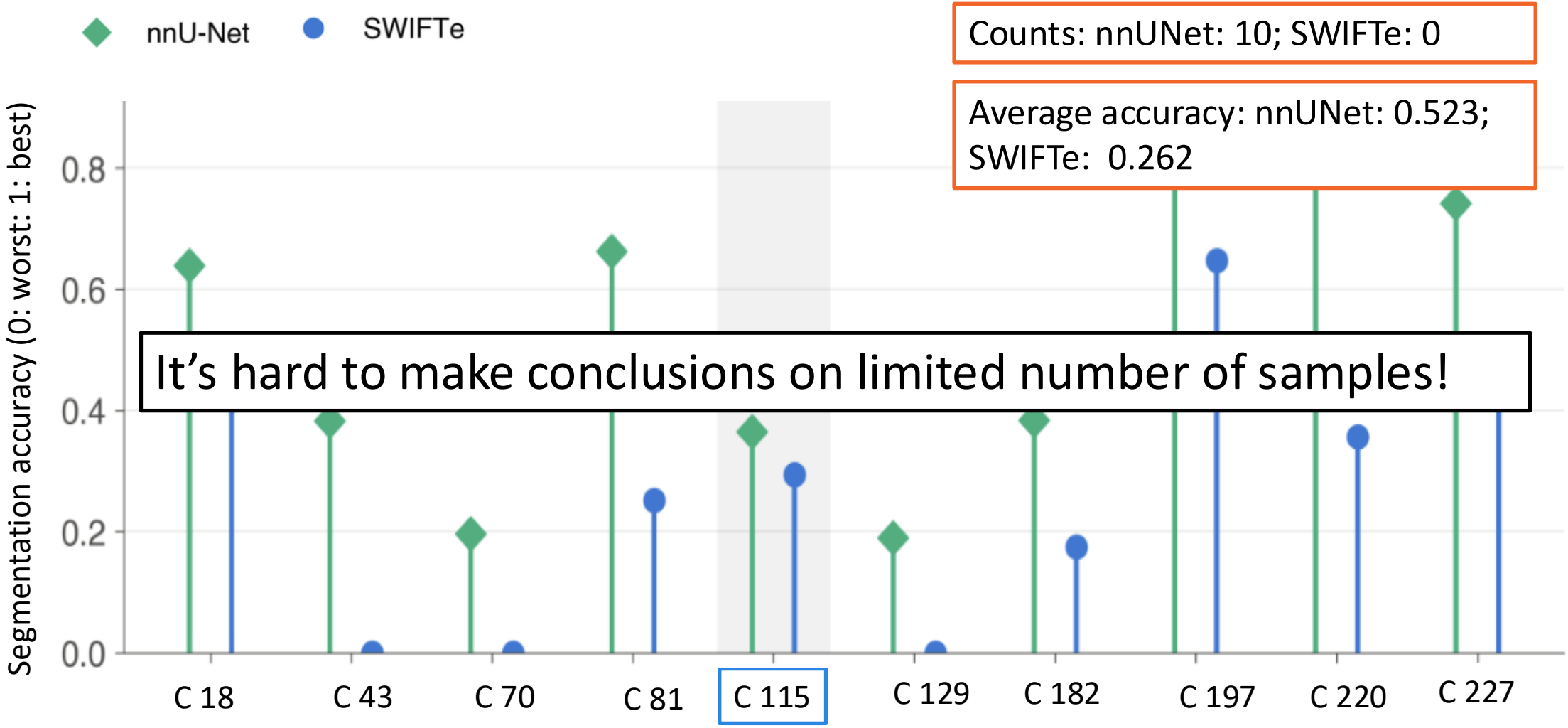
If you could look at performance for few samples, can you tell which model is better?



Counts: nnUNet: 4; SWIFTe: 5

Average accuracy: nnUNet : 0.535; SWIFTe: 0.547

How about now with different samples? Which model is better?



Counts: nnUNet: 10; SWIFTe: 0

Average accuracy: nnUNet: 0.523; SWIFTe: 0.262

It's hard to make conclusions on limited number of samples!

Understanding samples:
Types of data and distribution

Categorical data

Values represent labels or groups — not measured quantities.

Nominal

No inherent order

Examples:

Tumor site	**Lung, prostate, brain**
Response	**No response, partial, complete response**
Scanner manufacturer	**GE, Siemens, Phillips**

Ordinal

Categories have a natural order

Examples:

Toxicity grade	**I < II < III < IV**
T-stage	**T1 < T2 < T3 ..**
Pain severity	**None < Mild < Moderate < Severe**
Likert response	**Strongly disagree --> Strongly agree**

Binary data are categorical with only two levels (e.g., recurrence: yes / no).

Numerical data

Discrete

Countable values – fixed integers

Examples:

Number of prior treatments	**1, 2, ….**
Smoking years	**1, 10, 0**
Radiotherapy fractions	**2, 10, 25**

Continuous

Typically floating point measurements

Examples:

Tumor volume	**5 cc, 3.8 cc**
Segmentation accuracy	**0.8, 0.75, 0.637**
Time to recurrence	**3.8 months, 249 days,**

Numerical data support meaningful arithmetic – e.g. mean, median, standard deviation

Binomial distribution

- Used for modeling binary outcomes with fixed number of independent trials
- Each trial has only two possible results
- Probability of success (p) is fixed

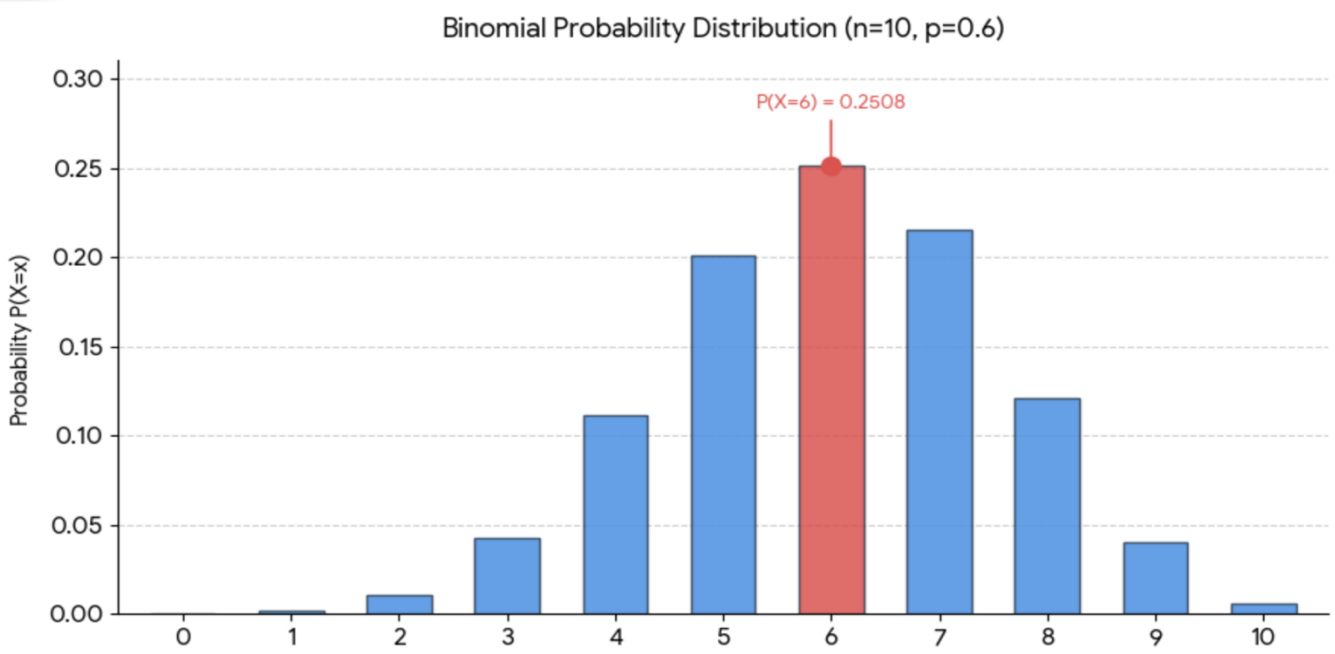
Probability of getting **x** successes in **n** trials $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

Binomial coefficient $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Binomial distribution Example

Treatment response

Number of patients in trial: 10
Probability of response p = 0.6
Number of responding x = 6



Probability of 6 patients responding to trial:

$$P(X=6) = \underbrace{\text{210 ways to choose response}}_{} \times 0.6^{6} \times (1-0.6)^{(1-0.6)}$$

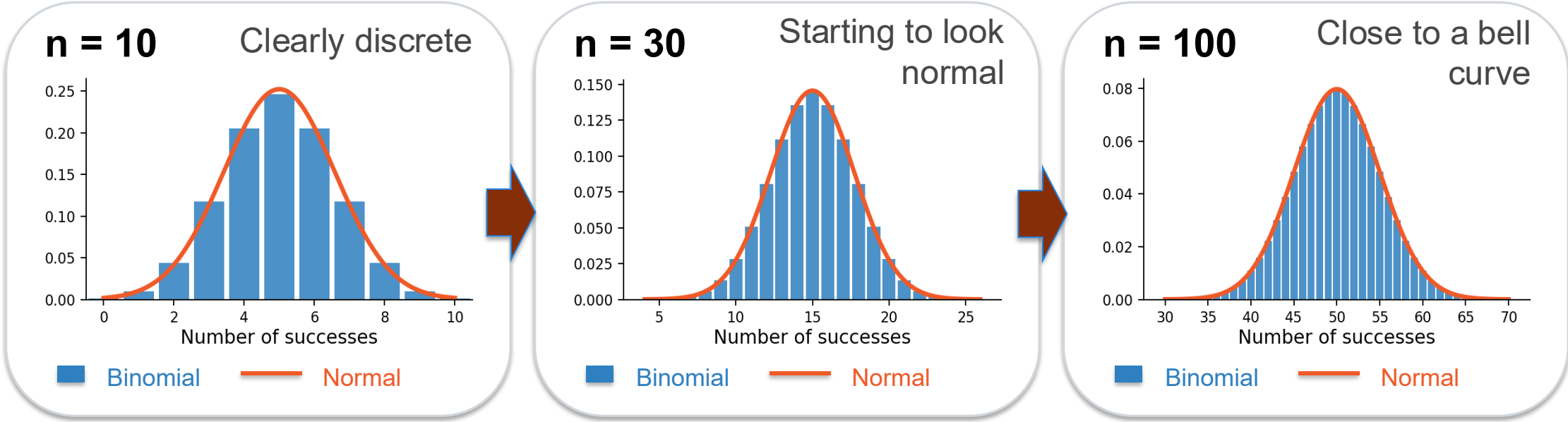
Probability of no response

Probability of response

$$P(X=6) = 0.2508$$

Binomial distribution shifts to normal distribution with increasing number of trials

As the number of independent trials increases, a binomial distribution becomes increasingly bell-shaped.



Why? A binomial count is a sum of many independent Bernoulli trials.

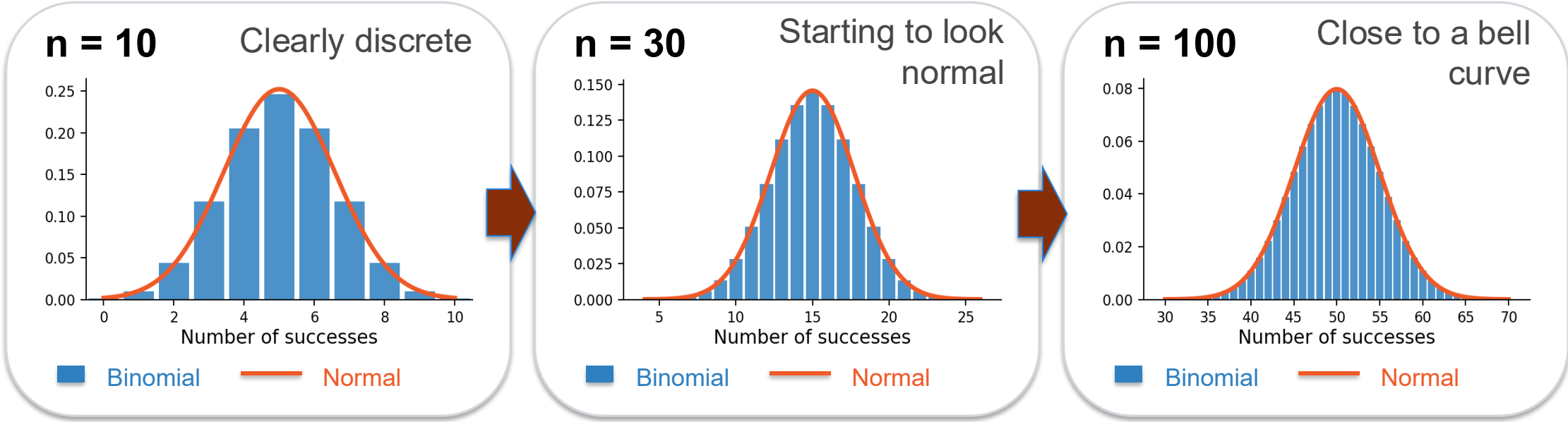
$X = X_1 + X_2 + \cdots + X_n$ where each X_i is 0 or 1

Central Limit Theorem

For many independent observations, the distribution of their sum or mean approaches a normal distribution as n grows.

Binomial distribution shifts to normal distribution with increasing number of trials

As the number of independent trials increases, a binomial distribution becomes increasingly bell-shaped.



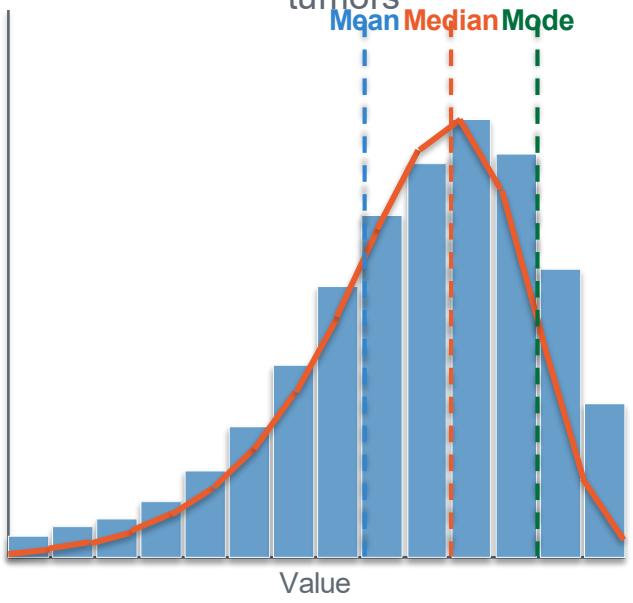
Normal distribution can be summarized by mean (average) and standard deviation

Normal distribution is perfectly symmetric – mean (average), median (exact middle value), and mode (most frequent value) are all equal

Tendency of distribution can shift resulting in different skewness

Left-skewed

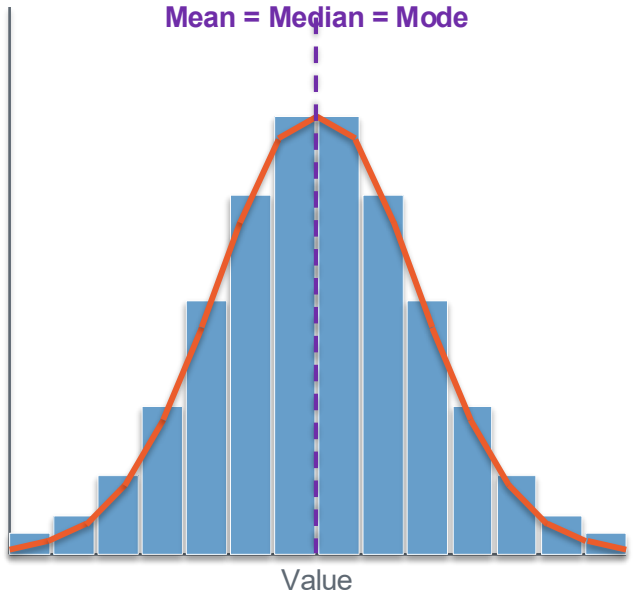
Example: Planned radiation dose to tumors



Mean < Median < Mode

Normal

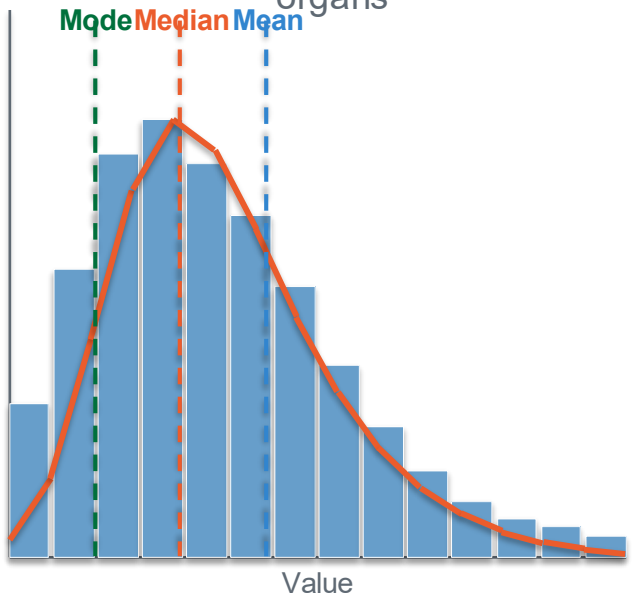
Example: Adult heights



Mean = Median = Mode

Right-skewed

Example: Planned radiation dose to organs



Mode < Median < Mean

Poisson Distribution

Models the number of events occurring in a fixed interval of time or space

Probability mass function

$$P(X = k) = {(e^{-\lambda}\lambda^{k})}/{k!}$$

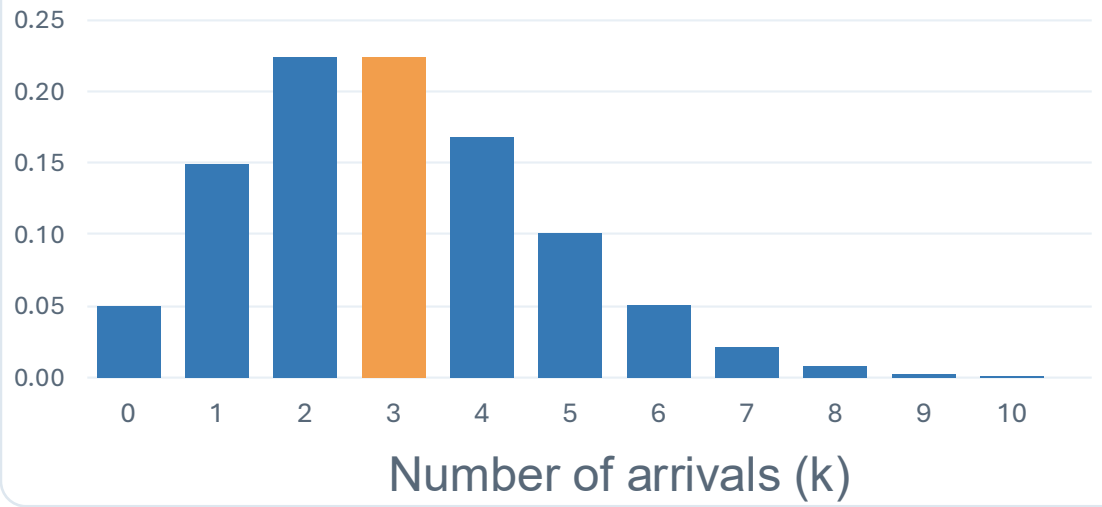
k = 0, 1, 2, … λ = expected event count

Key assumptions

- Events occur independently
- Average event rate is constant
- Events are counted in a fixed interval

Example: Patient arrivals per hour

Average rate λ = 3 patients/hour



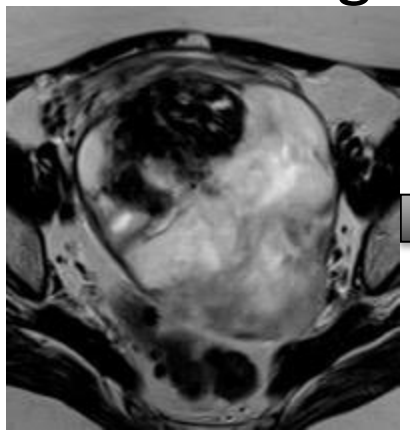
Interpretation:

P(X = 3) ≈ 0.224 → 22.4% chance of exactly 3 arrivals in one hour

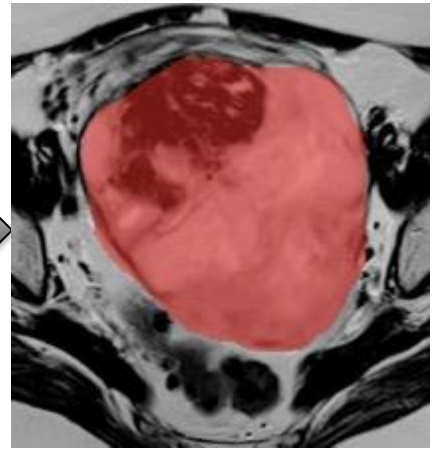
First order moments of a histogram: Radiomics features

Radiomics: Extraction of quantitative features from images

Source Image



Segmented

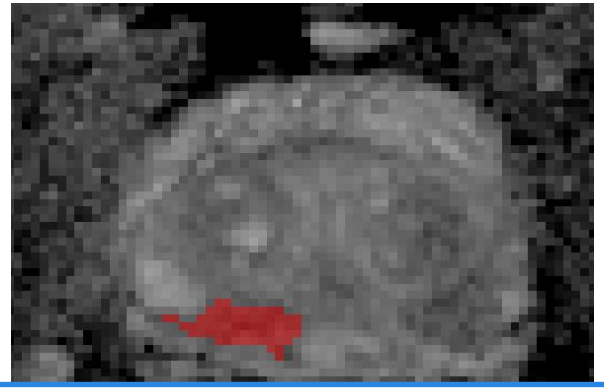


Represent the intensity statistics within tumor as a histogram

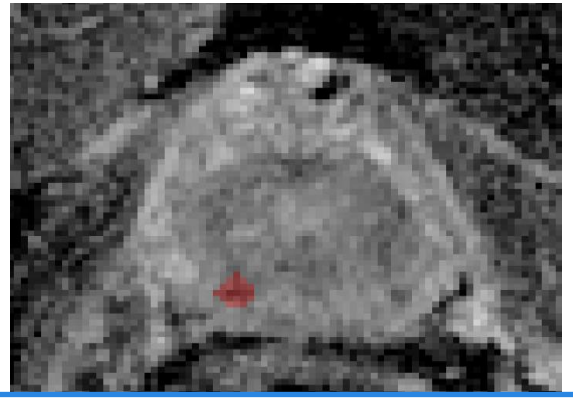
Moments of the histogram: Mean, standard deviation, kurtosis, skewness can be used as features within machine learning model

Real world example where skewness has meaning

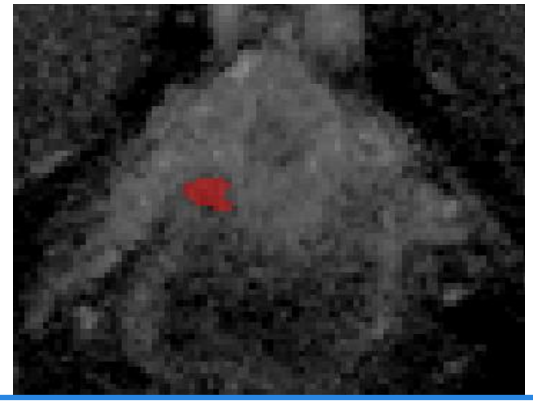
Baseline



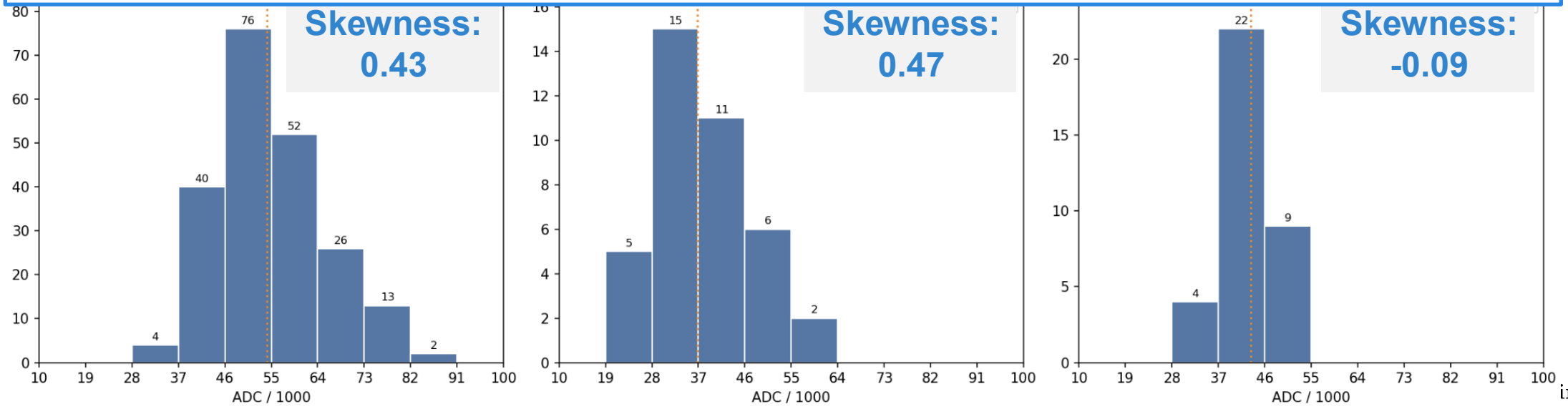
3 months after Tx



6 months after Tx

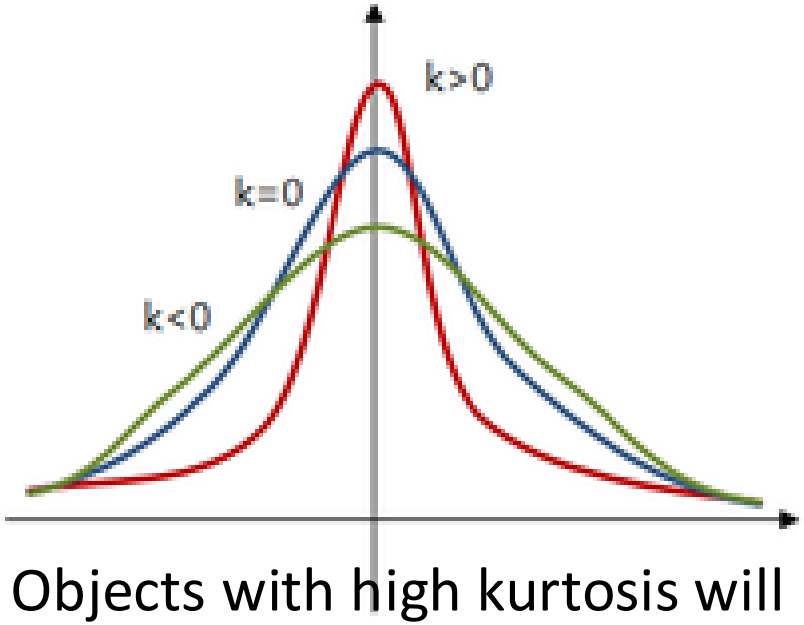


Skewness of apparent diffusion coefficient (how easily water molecules move between cells) decreases as prostate cancer responds to treatment



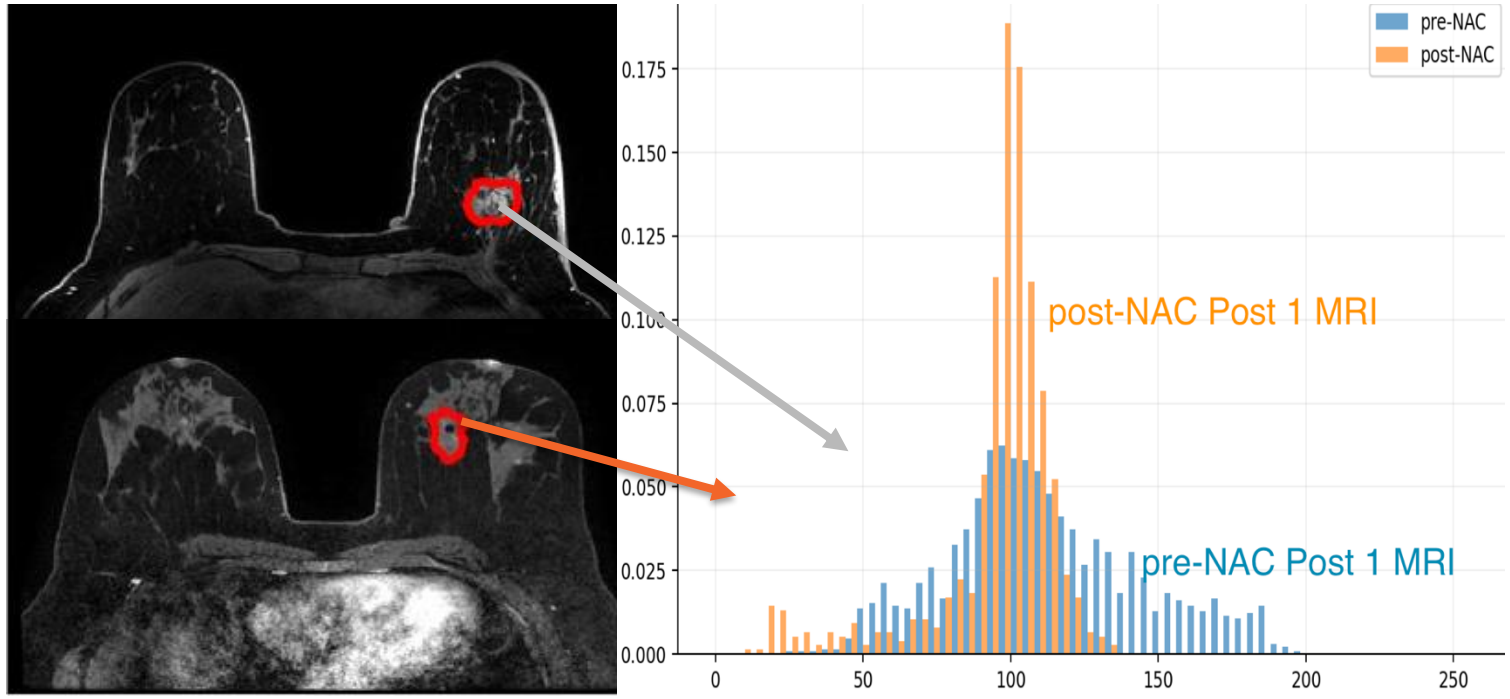
Another useful measure: Kurtosis

Kurtosis measures how peaked a distribution is and emphasizes the tails and peaks of a distribution



Objects with high kurtosis will be isointense

Low kurtosis objects will be diffuse

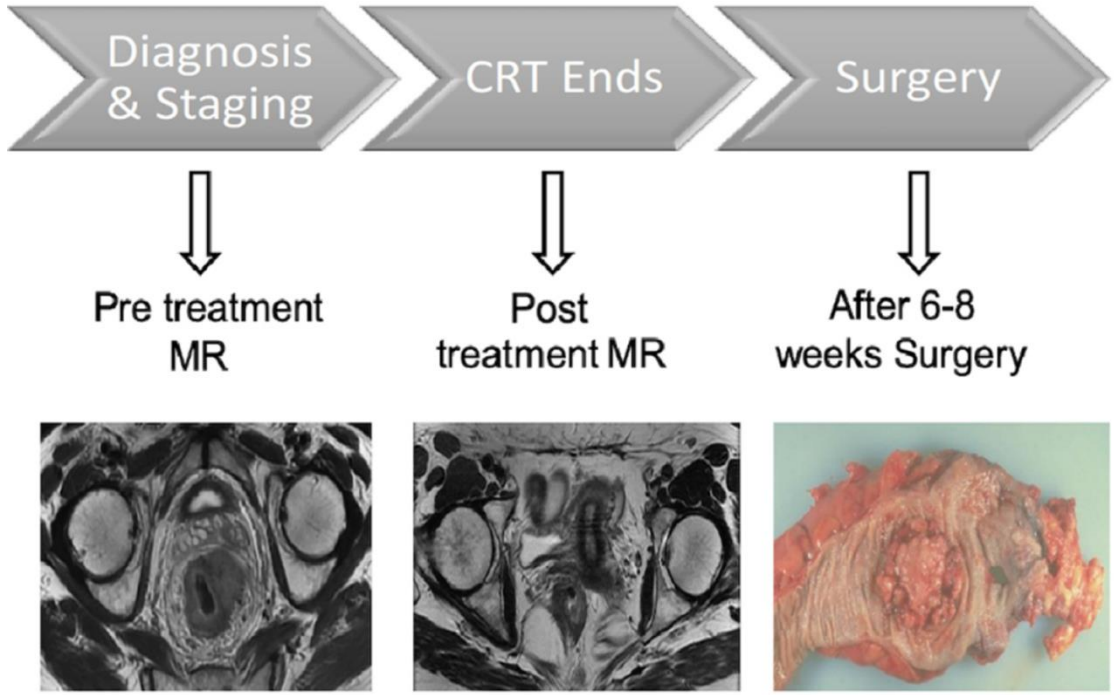


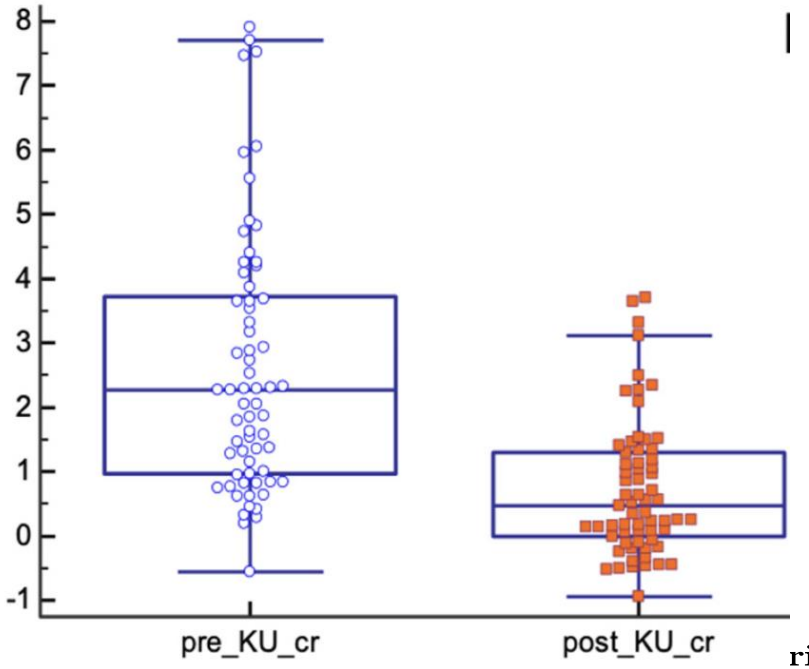
Intensity histogram of breast cancer for a patient who did not respond to treatment

Kurtosis can be a useful predictor of response

Performance of Machine Learning and Texture Analysis for Predicting Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer with 3T MRI

Davide Bellini [1], Iacopo Carbone [1,*], Marco Rengo [1], Simone Vicini [1], Nicola Panvini [1], Damiano Caruso [2], Elsa Iannicelli [2], Vincenzo Tombolini [3] and Andrea Laghi [2]





ring

From Samples to Populations: Estimator basics

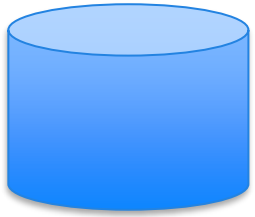
Memorial Sloan Kettering Cancer Center

What is an Estimator?

- Rule of formula describing statistical properties of a population based on observed data samples

Height of students in class

Data samples

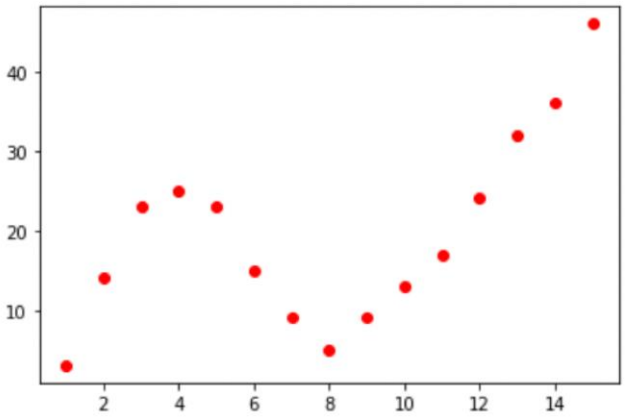


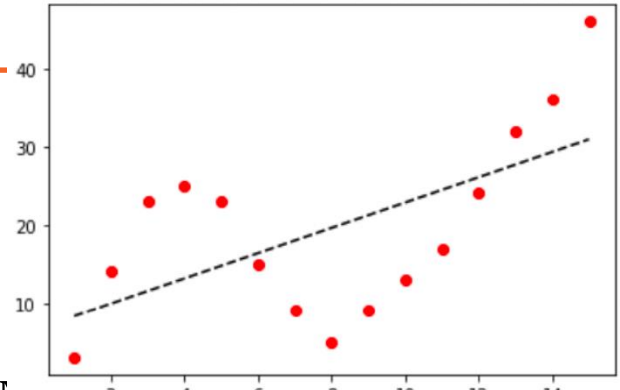
Estimator

Mean height of students in the class

$$f(x) = mx + c$$

Estimator can also be a function





1884
Cancer Center

Parametric distributions can be estimated using descriptive statistics

Distribution	Data	Parameters	
		Parameter	What it captures
Normal	Continuous	Mean (μ)	Center of data
		Variance (σ^2)	Squared distance from mean
Binomial	# success in n trial	Number of trials (n)	Probability (p)
Poisson	Event counts	λ	Expected event count

Parametric distributions provide a convenient way to summarize samples from a population

Parametric machine learning: Logistic Regression

Instead of estimating point approximations of data, we can estimate the entire relationship between input and output!

Fits a logistic sigmoid to the data to estimate output

$$y = \frac{1}{1 + e^{-x}}$$

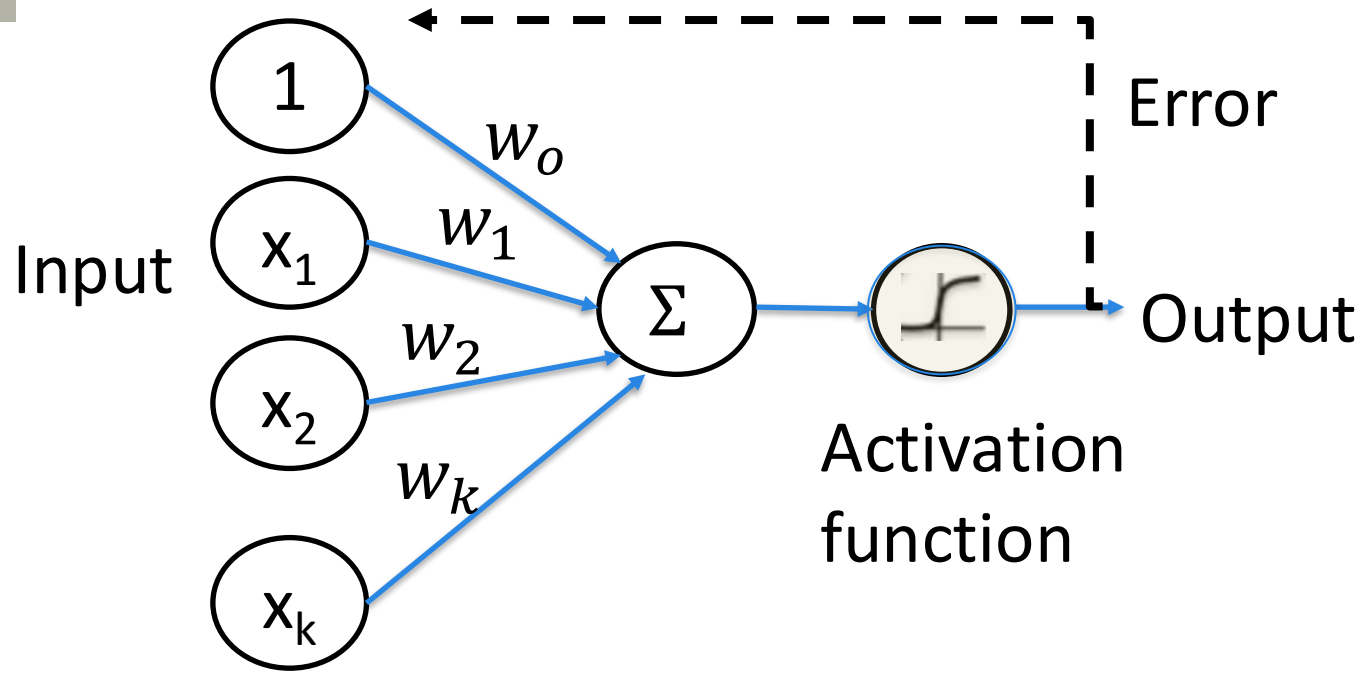
X = Input data;

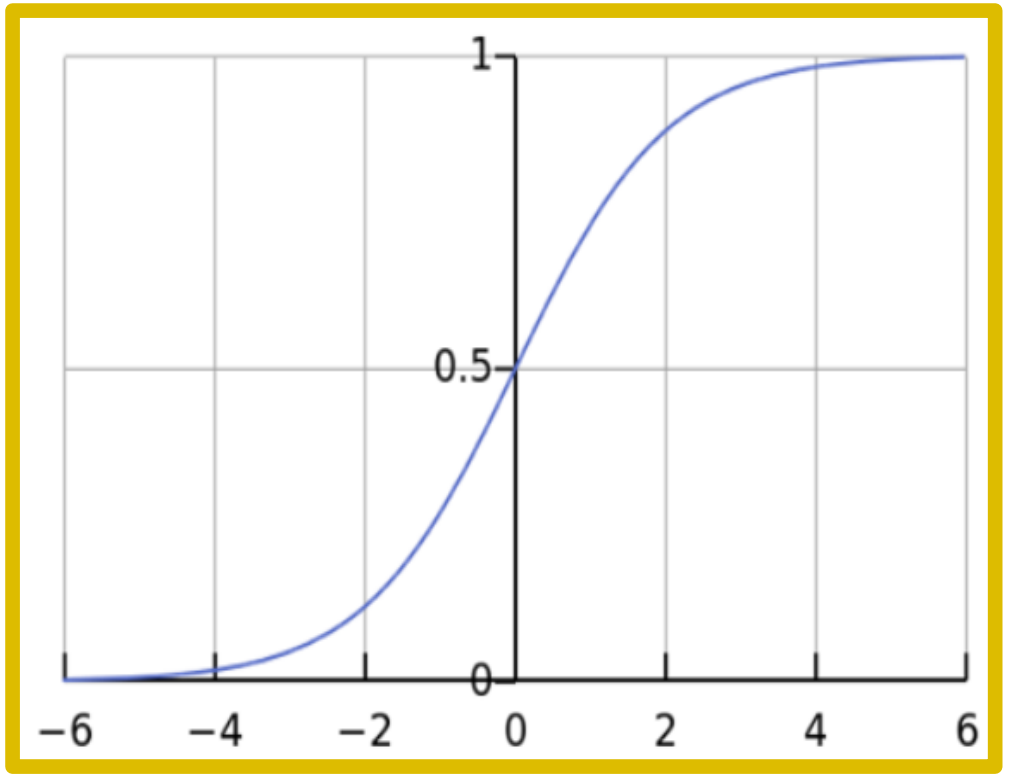
dose_max	dose_mean	Volume	HPV	Age	Sex
66.0	2.200000	0.079261	Positive	0.299436	M
70.0	2.121212	-0.867136	Positive	0.774731	M
70.0	2.121212	2.800150	Positive	1.250026	M

y = output (e.g Patient with osteoradionecroris or dying bone tissue)

Memorial Sloan Kettering Cancer Center

How logistic regression works?





$$log\text{-}odds(Y) = w_o.1 + w_1.x_1 + w_2.x_2 + w_3.x_k$$

A logistic sigmoid function combines inputs and calculates outputs

Memorial Sloan Kettering Cancer Center
1884

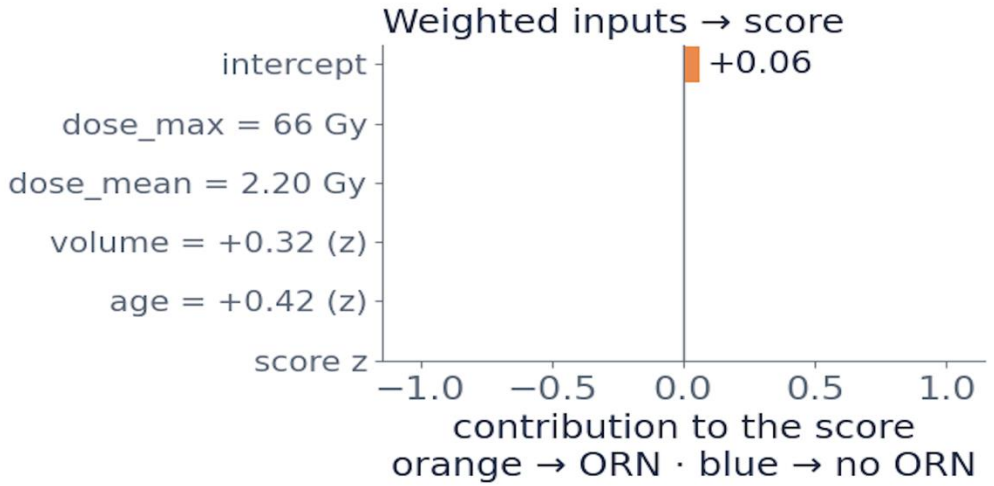
Example classification using logistic regression

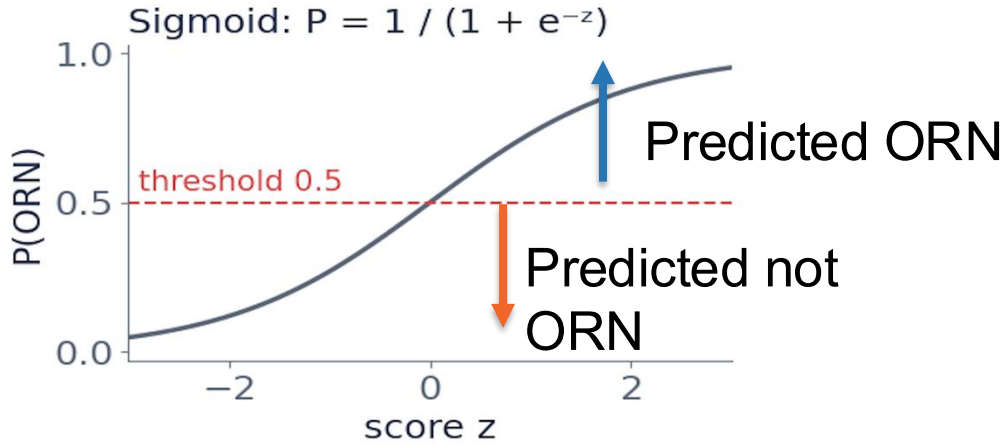
Predict patients with head and neck cancer treated with radiation therapy likely to develop osteoradionecrosis or (dying bone tissue)

Model was trained with 35 patients

log-odds(ORN) = +0.06 − 0.123·dose_max + 0.165·dose_mean + 0.022·volume + 0.010·age (standardized)

original units (Gy): log-odds(ORN) = −3.08 − 0.096·dose_max + 4.602·dose_mean + 0.021·volume + 0.010·age





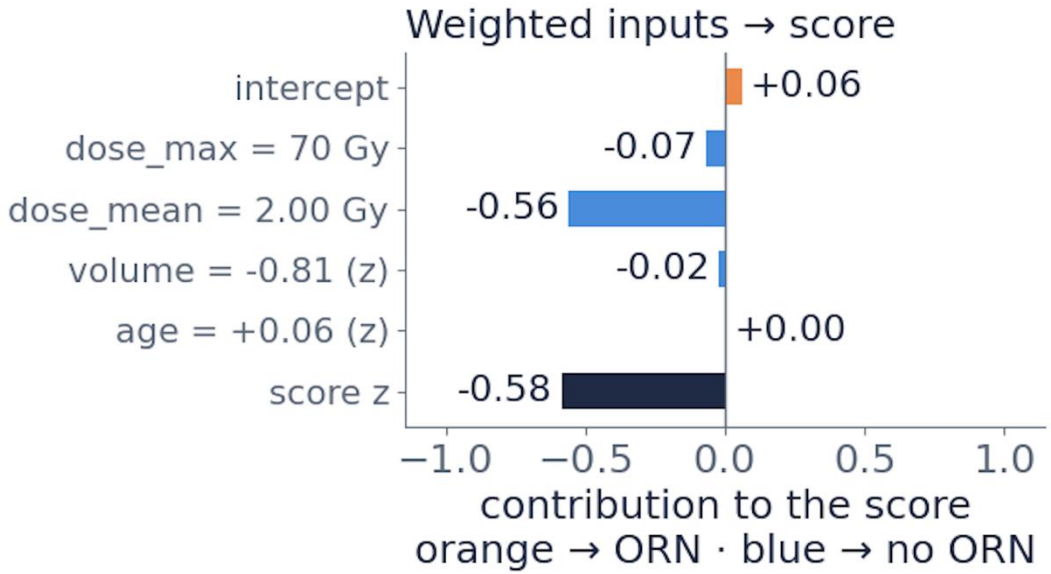
patient 1	patient 2	patient 3	patient 4	patient 5	patient 6

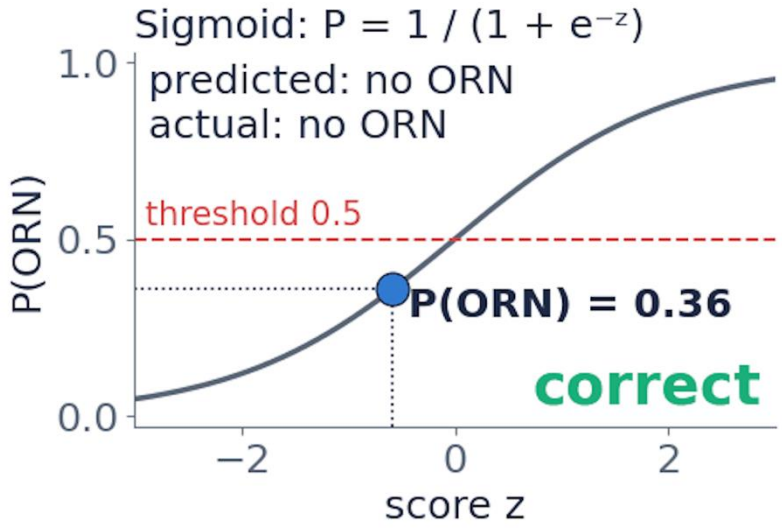
Example classification using logistic regression

Notice log-odds close to the selected cut-offs can lead to wrong predictions

log-odds(ORN) = +0.06 − 0.123·dose_max + 0.165·dose_mean + 0.022·volume + 0.010·age (standardized)

original units (Gy): log-odds(ORN) = −3.08 − 0.096·dose_max + 4.602·dose_mean + 0.021·volume + 0.010·age



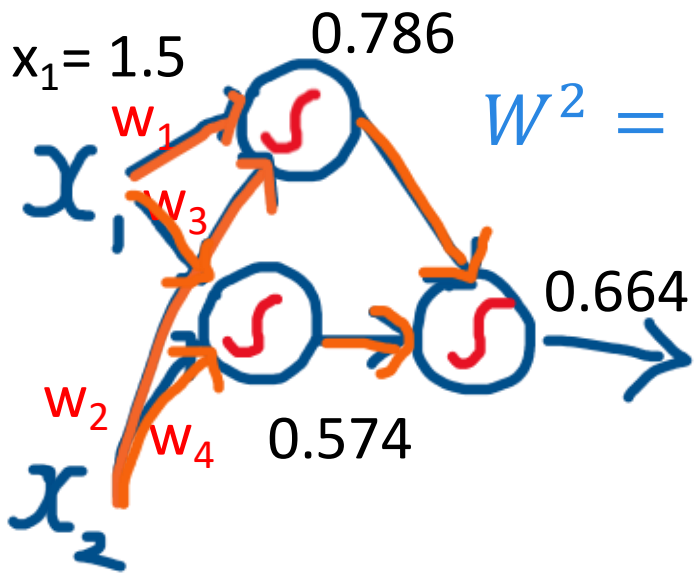


P = 0.68	P = 0.49	P = 0.49	P = 0.51	P = 0.49	P = 0.36
true: ORN	true: ORN	true: ORN	true: no ORN	true: no ORN	true: no ORN
correct	**wrong**	**wrong**	**wrong**	**correct**	**correct**

Bridge to Deep learning

- Logistic regression: f(x) = y; We learn the parameters for f
- Deep learning in simplest form: f(f(f(f(f(x)))))
- What distinguishes deep learning is there are many functions or what we call **hidden layers** and each of these **functions can take different sets of parameters and activations**

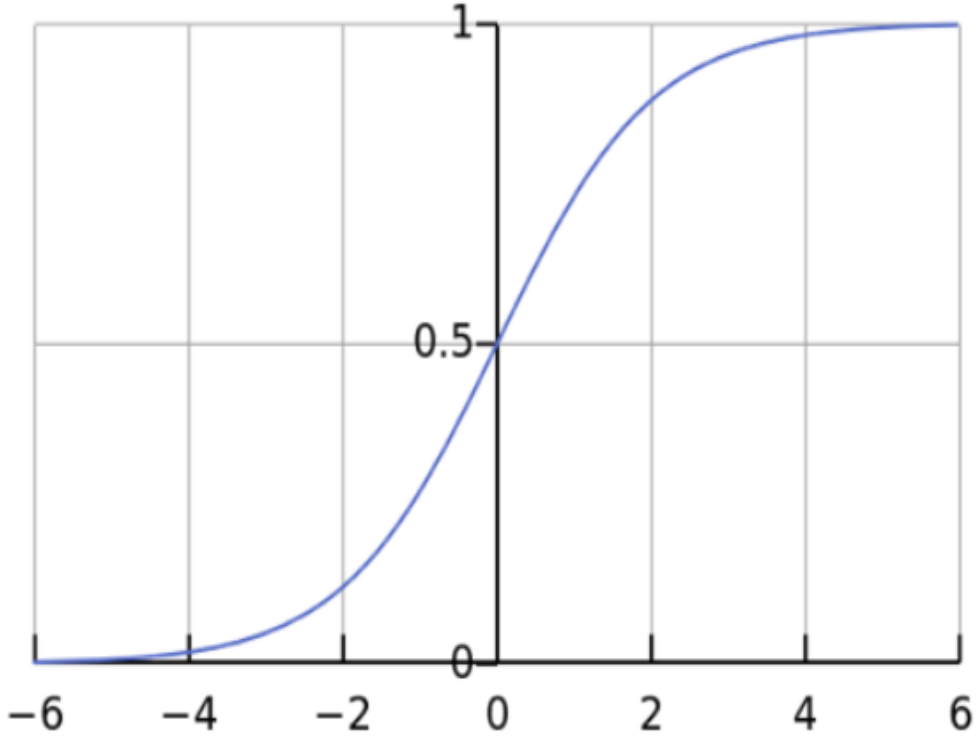
SEQUENCE OF ACTIVATION FUNCTIONS DRIVE A NN



$W^2 = \{0.5, 0.5\}$

$W^1 = \{0.6, 0.8, 0.1, 0.3\}$

- Sigmoid function: $\frac{1}{1+e^{-x}}$

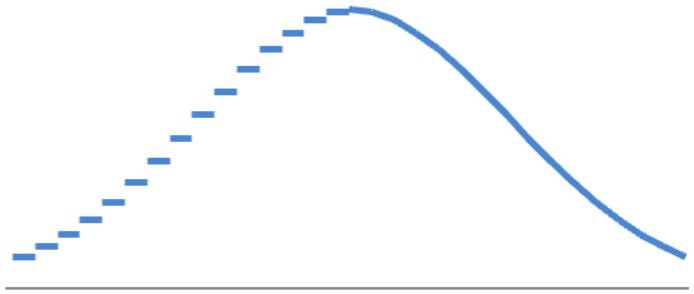


Memorial Sloan Kettering Cancer Center

Non-parametric distributions and estimators

- Often occur due to limited samples
- Estimator estimates shape of distribution from the data

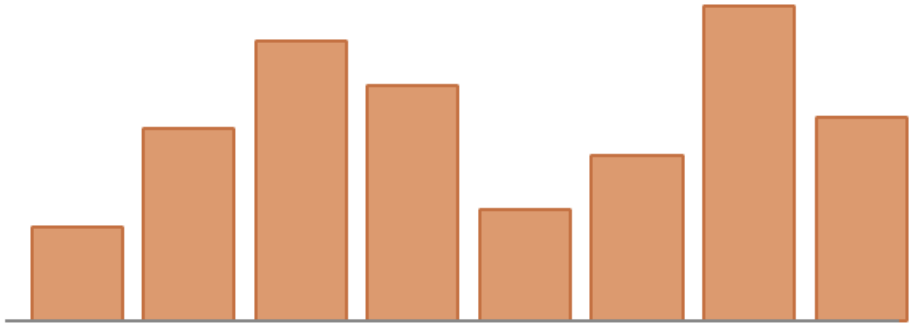
Parametric



Example: Normal distribution
Estimate a few parameters
(e.g., mean and standard deviation)

Nonparametric

Do not assume a fixed distribution shape



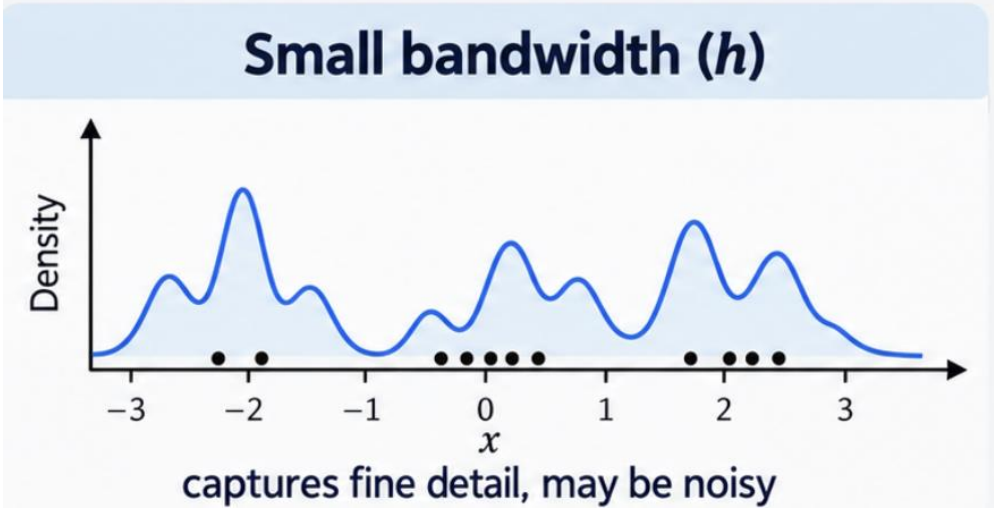
Estimate frequencies in a histogram

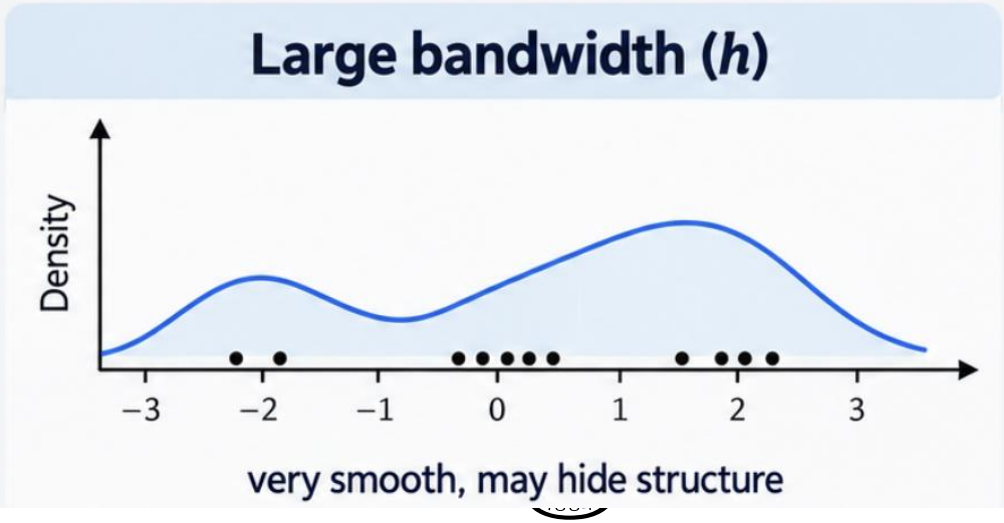
Estimating nonparametric distribution using Kernel Density Estimation

- Places windows or kernels centered at each data point and smooths the distribution using a bandwidth parameter
- Most common kernel is a Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- n = number of observations
- K = kernel (smooth bump)
- h = bandwidth (controls smoothness)





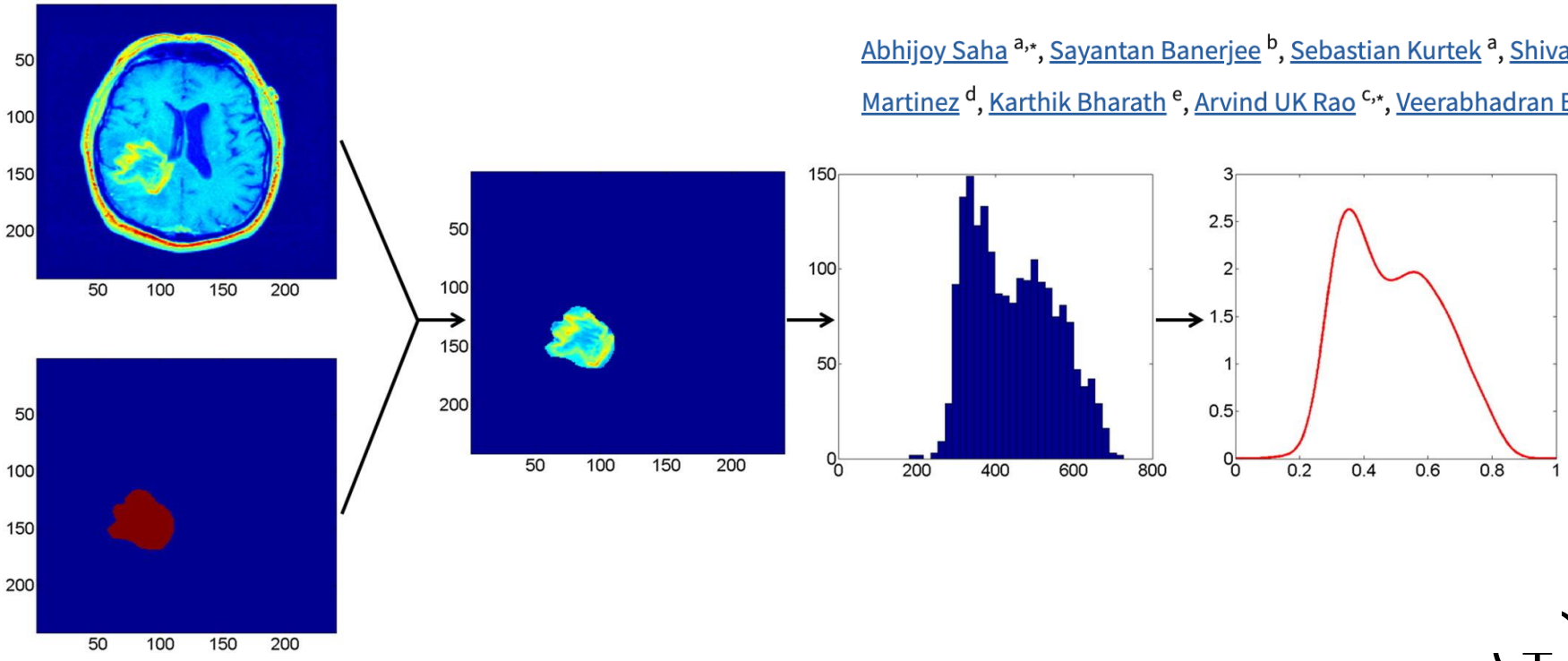
Kernel density estimate used to characterize brain tumor heterogeneity

Tumors vary in appearance and the degree of variation (or heterogeneity) is often associated with how aggressive a tumor is

Neuroimage Clin. 2016 May 27;12:132–143. doi: 10.1016/j.nicl.2016.05.012

DEMARCATE: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer

Abhijoy Saha [a,*], Sayantan Banerjee [b], Sebastian Kurtek [a], Shivali Narang [c], Joonsang Lee [c], Ganesh Rao [d], Juan Martinez [d], Karthik Bharath [e], Arvind UK Rao [c,*], Veerabhadran Baladandayuthapani [f,*]



Bayesian statistics

More math formulas but now we can now ask how probable events are

Concept of uncertainty is important in most fields including medicine

Basics of Bayesian learning

Bayes' Theorem is the basis for updating probability of a hypothesis θ conditioned on observed data *D*

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Prior: indicates how likely the hypothesis is

$p(\theta|D)$ increases when data is observed conditioned on hypothesis

$p(\theta|D)$ ↓ when data is observed independent of hypothesis

Typically, we evaluate from a set of candidate hypotheses and want to know which is most likely given what we observe

E.g. Hypotheses (Rain, ¬ Rain)
Data(Air temperature: 85F, Dew point: 50F, Humidity: 20%)

E.g. Hypotheses (Flu, Lyme disease, appendicitis)
Data (high fever, elevated WBC, flu season, no stomach pain)

How do we evaluate from a set of candidates using Bayes' rule?
We apply maximum a posteriori (MAP) hypothesis: In plain speak select the most likely out of all the available hypotheses

Maximum a Posteriori (MAP) Hypothesis

Given observed data D, choose the hypothesis $\theta_i \epsilon \{\theta_1, \ldots \theta_N\}, i \leq N$ that has the highest posterior probability.

MAP hypothesis

$$\boldsymbol{\theta_{MAP}} = \underset{\theta}{\operatorname{argmax}}\, \boldsymbol{P(\theta|D)}$$

Just a mathy way to say we select the hypothesis that has maximum value (in our case probability)

Using Bayes' rule:

$$\boldsymbol{\theta_{MAP}} = \underset{\theta}{\operatorname{argmax}}\, \boldsymbol{P(D|\theta)} \times \boldsymbol{P(\theta)}$$

Prior $\boldsymbol{P(\theta)}$
Belief before observing data

Likelihood $\boldsymbol{P(D|\theta)}$
How well hypothesis explains the data

Posterior $\boldsymbol{(\theta|D)}$
Updated belief after data

Using Bayesian learning to detect cancer example

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Prior: indicates how likely the hypothesis is

$p(\theta|D)$ increases when data is observed conditioned on hypothesis

$p(\theta|D)$ ↓ when data is observed independent of hypothesis

+: positive test; −: negative test
P(cancer) = 0.008,
P(¬ cancer) = 1 – P(cancer): 0.992
P(+|cancer) = 0.98, P(− |cancer) = 0.02
P(+ |¬ cancer) = 0.03, P(−|¬cancer) = 0.97

Data: Lab test returns a positive result
P(+|cancer) P(cancer) = 0.98 x 0.008
= 0.0078
P(+ |¬ cancer) P(¬ cancer) = 0.03 x 0.992
= 0.0298

$$\text{P(cancer)} = \frac{0.0078}{(0.0078+0.0289)} = 0.21$$

$$\text{P(¬ cancer)} = \frac{0.0289}{(0.0078+0.0289)} = 0.73$$

Problem with solving Bayes when we have multiple attributes

E.g. Hypotheses (Flu, Lyme disease, appendicitis)
Data (high fever, elevated WBC, flu season, no stomach pain, chills)

a_1 a_2 a_3 a_4 a_5

Computing MAP hypothesis now involves evaluating it with respect to all available attributes:

$$\boldsymbol{\theta_{MAP}} = \underset{\theta_j \epsilon \theta}{\operatorname{argmax}} \boldsymbol{P(\theta_j | a_1, a_2, a_3, a_4, a_5)}$$

Using Bayes theorem this is

$$\boldsymbol{\theta_{MAP}} = \underset{\theta_j \epsilon \theta}{\operatorname{argmax}} \frac{\boldsymbol{P(\theta_j | a_1, a_2, a_3, a_4, a_5) P(v_j)}}{\boldsymbol{P(a_1, a_2, a_3, a_4, a_5)}}$$

Problem with solving Bayes when we have multiple attributes

E.g. Hypotheses (Flu, Lyme disease, appendicitis)
Data (high fever, elevated WBC, flu season, no stomach pain, chills)

a_1 a_2 a_3 a_4 a_5

Problem is that estimating the first term using training data would require a lot of data due to lot of different ways the attributes can vary across patients

$$\boldsymbol{\theta_{MAP}} = \underset{\theta_j \epsilon \theta}{\operatorname{argmax}} \boldsymbol{P}\left(\boldsymbol{\theta_j} \middle| \boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}, \boldsymbol{a_4}, \boldsymbol{a_5}\right) \boldsymbol{P}(\boldsymbol{v_j})$$

Naïve Bayes Classifier to the rescue

Key Idea: Treat each attribute as **conditionally independent** given the target value

Data (high fever, elevated WBC, flu season, no stomach pain, chills)

a_1 a_2 a_3 a_4 a_5

This means we solve with $P(a_i|\theta_j)$ instead of $\boldsymbol{P(\theta_j|a_1, a_2, a_3, a_4, a_5)}$

Naïve Bayes: $\theta_{NB} = \underset{\theta_j \epsilon \theta}{\text{argmax}} \; P(\theta_j) \prod_i P(a_i|\theta_j)$

$$= \underset{\theta_j \epsilon \theta}{\text{argmax}} \; P(\theta_j) \, P(a_1|\theta_j) . P(a_2|\theta_j) . P(a_3|\theta_j) . P(a_4|\theta_j) . P(a_5|\theta_j)$$

This we can train with reasonably sized data!

Example to predict if our agent will play tennis

New day x: Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong

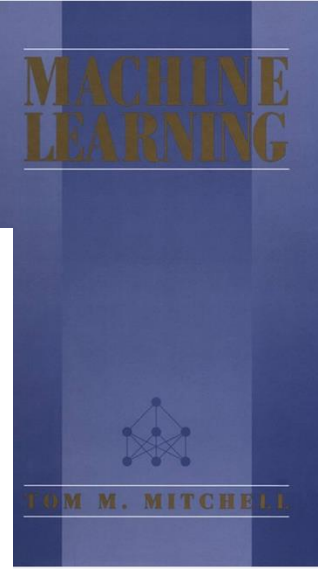
Day	Outlook	Temp	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10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

Step 7 · Normalize: divide by the sum

$P(\text{class} \mid x) \propto P(\text{class}) \cdot \prod P(\text{feature value} \mid \text{class})$

	Yes	No
P(class)	9/14 = 0.643	5/14 = 0.357
P(Sunny | ·)	2/9 = 0.222	3/5 = 0.600
P(Cool | ·)	1/3 = 0.333	1/5 = 0.200
P(High | ·)	1/3 = 0.333	4/5 = 0.800
P(Strong | ·)	1/3 = 0.333	3/5 = 0.600
product	**0.0053**	**0.0206**
P(class | x)	**0.205**	**0.795**

Prediction: PlayTennis = No



Tom Mitchell, “Machine Learning” Book

Frequentionist versus Bayesian approach

Frequentist learning

- Estimate outcome for test data instances based on statistics of the training samples
- No need for prior knowledge as relevant information is learned directly from data

Cons

- Learned knowledge is only as good as the quality and quantity of the training samples
- Uncertainty in prediction is not directly available

Bayesian learning

- Update probability of hypothesis based on prior knowledge
- Explicitly captures the uncertainty in predictions
- Does not require a lot of data for learning

Cons

- Computationally expensive as hypotheses increase
- Capturing prior domain knowledge is not always easy

Confidence intervals

What are they and why do we need them?

Confidence intervals allow to control for sampling variations

95% Confidence Interval for a Mean

Sample mean ± Margin of error

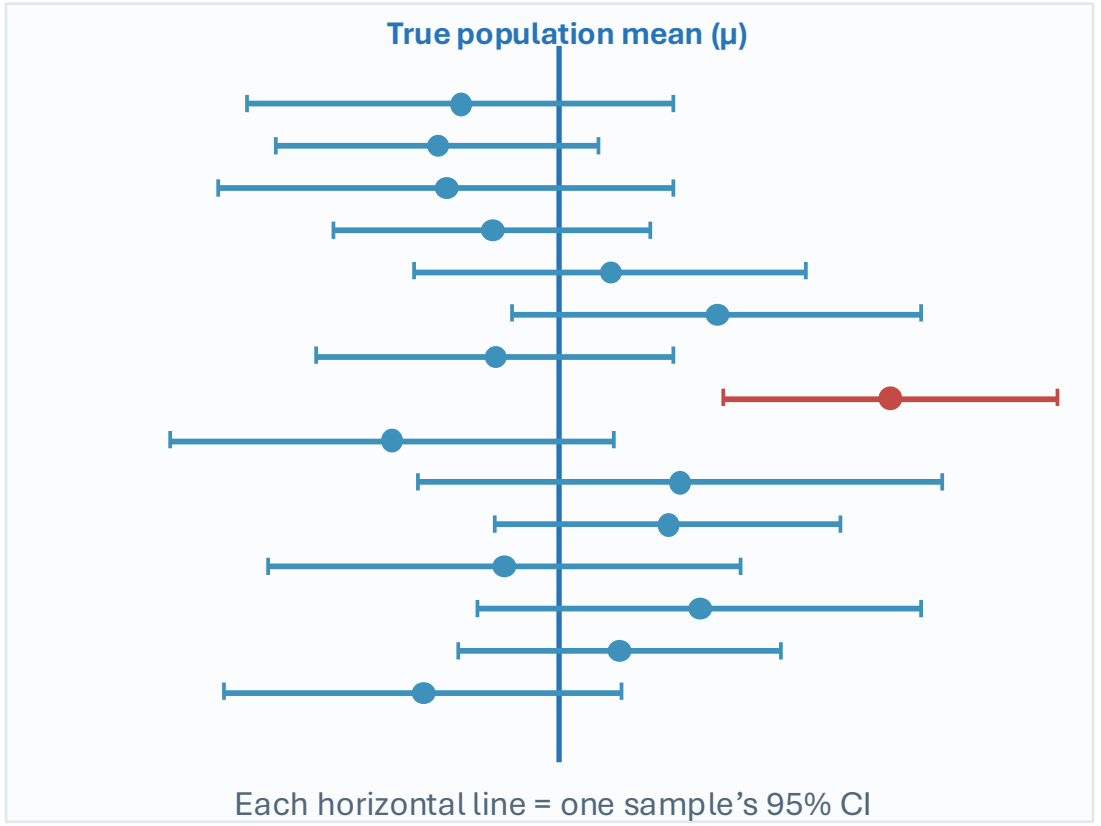
$$\bar{x} \pm 1.96 \times (s / \sqrt{n})$$

Approximate 95% CI for large samples; use a t critical value with estimated SD for small samples.

How to interpret 95% confidence

If we repeated the sampling process many times, about 95% of intervals constructed this way would contain the true population mean.

Repeated samples → different intervals



Key insight: 95% confidence describes the method's long-run coverage—not a 95% probability that μ is in this particular interval.

Take home points

- Hard to make assessment with limited samples
- Samples might misrepresent the general distribution hiding true accuracy
- Depending on the statistical measure (count, mean, median) different methods come out on the top
- When we have less training data, but can model domain knowledge as **prior** information, Bayesian learning can help