BAYESIAN CLASSIFIER

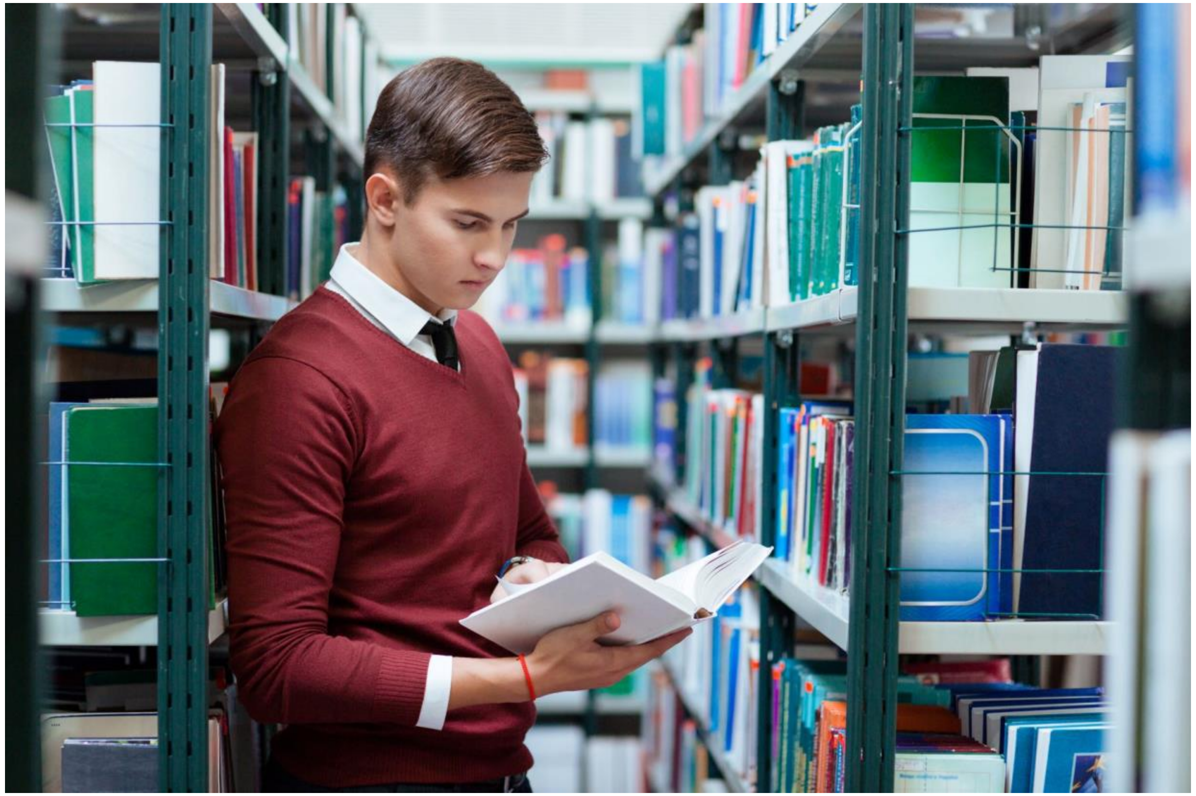
DR. FARHAD RAZAVI

OUTLINE

- Bayesian Inference
  - Conditional Probability
  - Joint Probability
  - Marginal Probability
  - Probability Distribution
- Bayesian Classifier

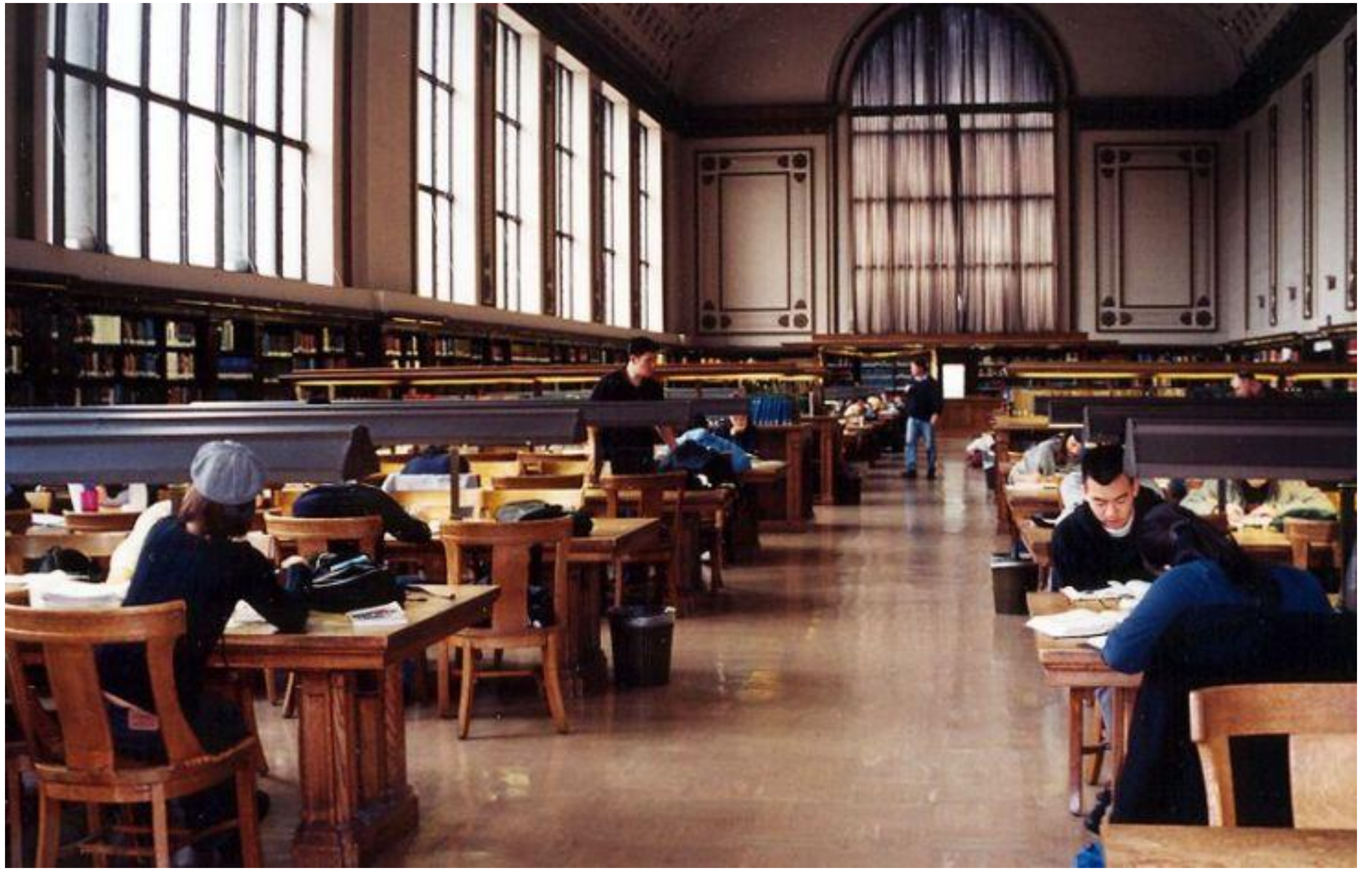
FIRST DAY OF SCHOOL DILEMMA

- Assume your school has only two departments. Business department and Math department.
- In the first day at school while checking the library, you bump into Danny for the first time.
- You open a conversation and ask a few questions. From your short interaction, you figure out Danny to be a very "shy" person.
- Do you give higher chance for Danny being a student from Business department or him being from the Math department?



FIRST DAY OF SCHOOL DILEMMA

- What if I tell you the following information?
  - The number of students who have enrolled to the Department of Mathematics is 250.
  - The number of students who have enrolled to the School of Business is 1200.
  - What changed?!



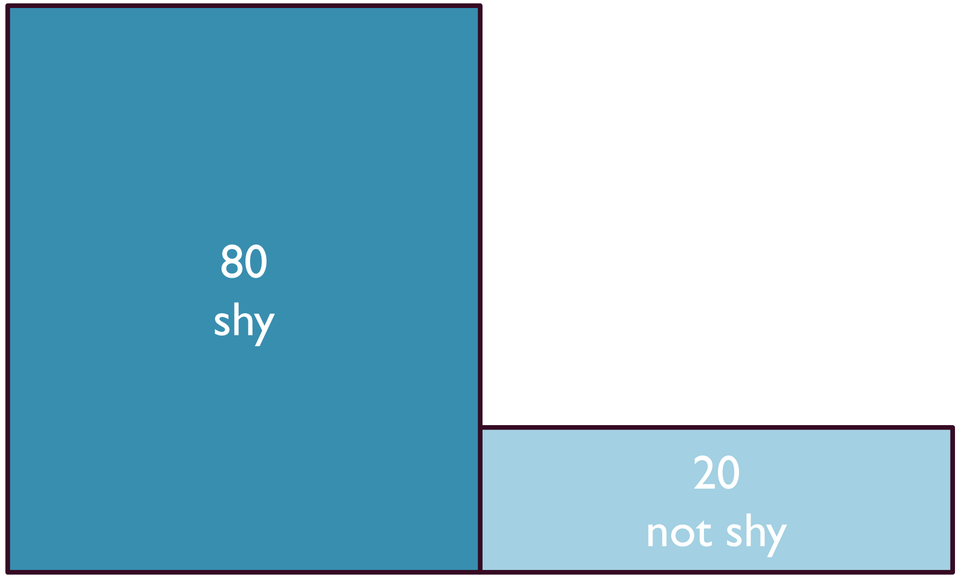
BAYESIAN INFERENCE

- Bayesian inference is a way to capture common sense.
- It helps you use what you know to make better guess.

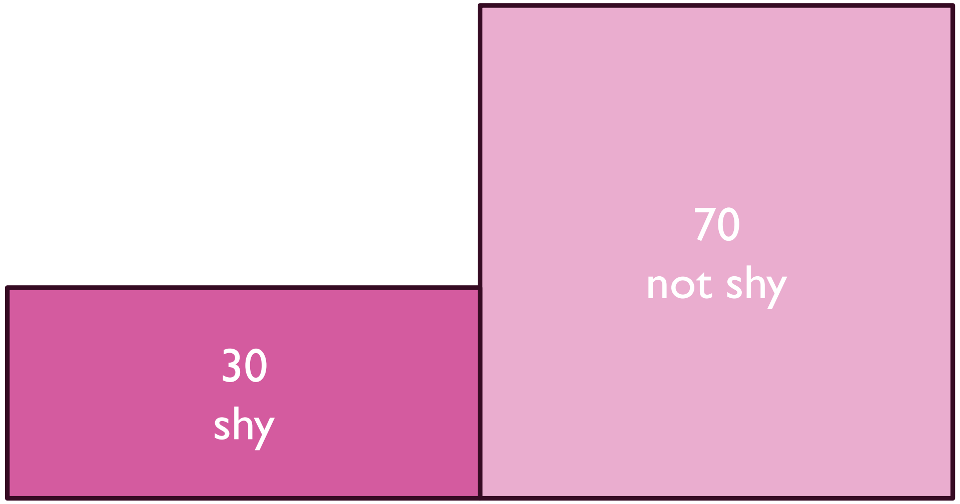
FIRST DAY OF SCHOOL DILEMMA

- Let's put numbers to our dilemma

Out of 100 students from
Department of Mathematics

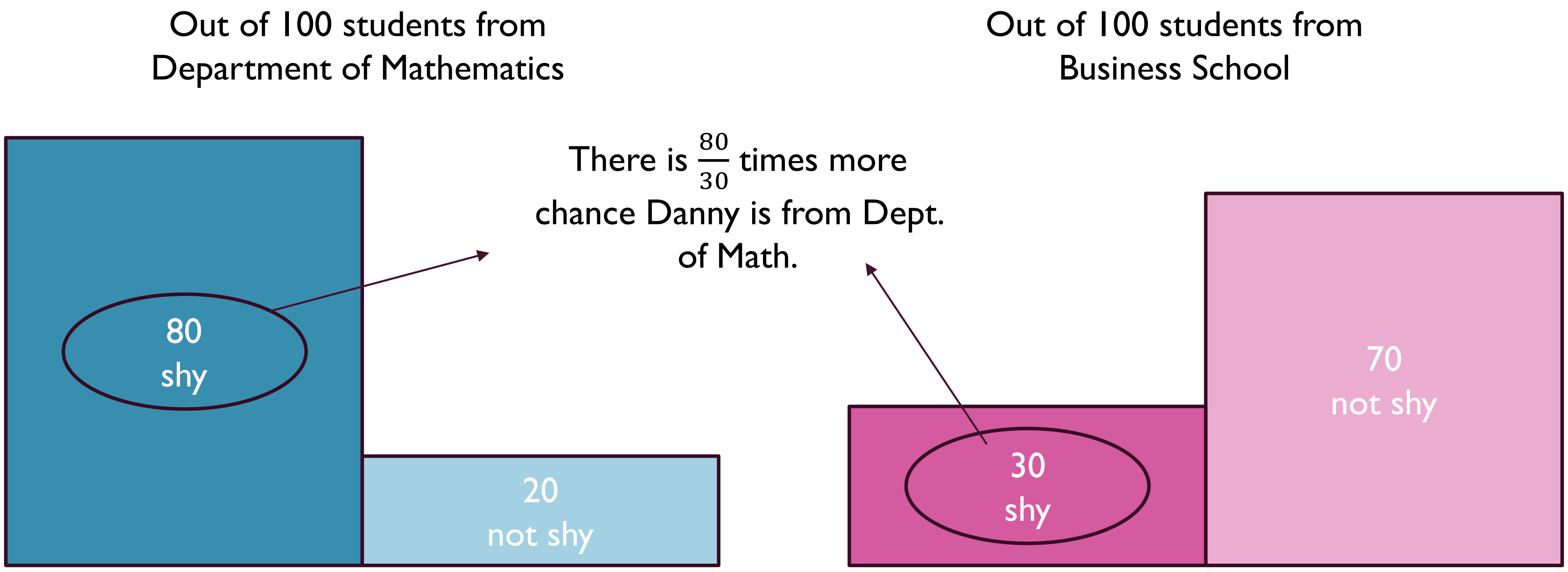


Out of 100 students from
Business School



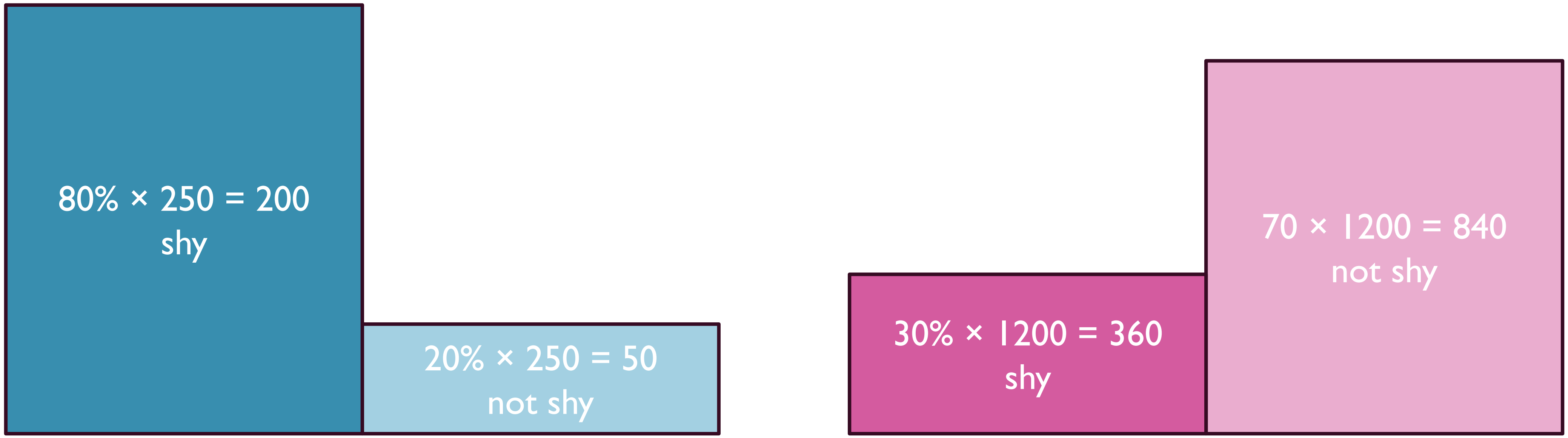
FIRST DAY OF SCHOOL DILEMMA

- If we had assumed there is the same number of students from both departments at school, then:



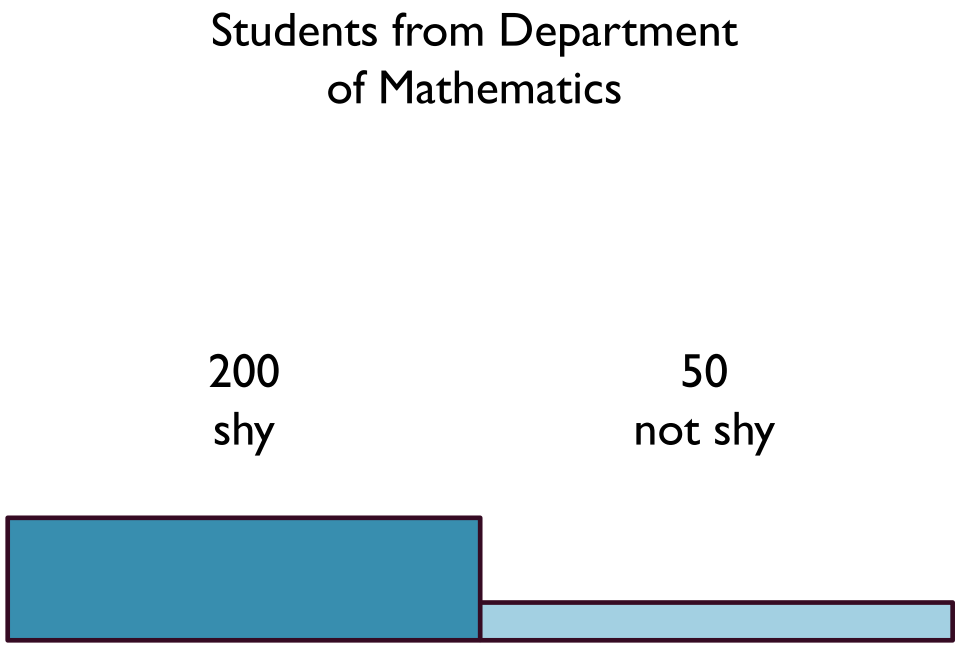
FIRST DAY OF SCHOOL DILEMMA

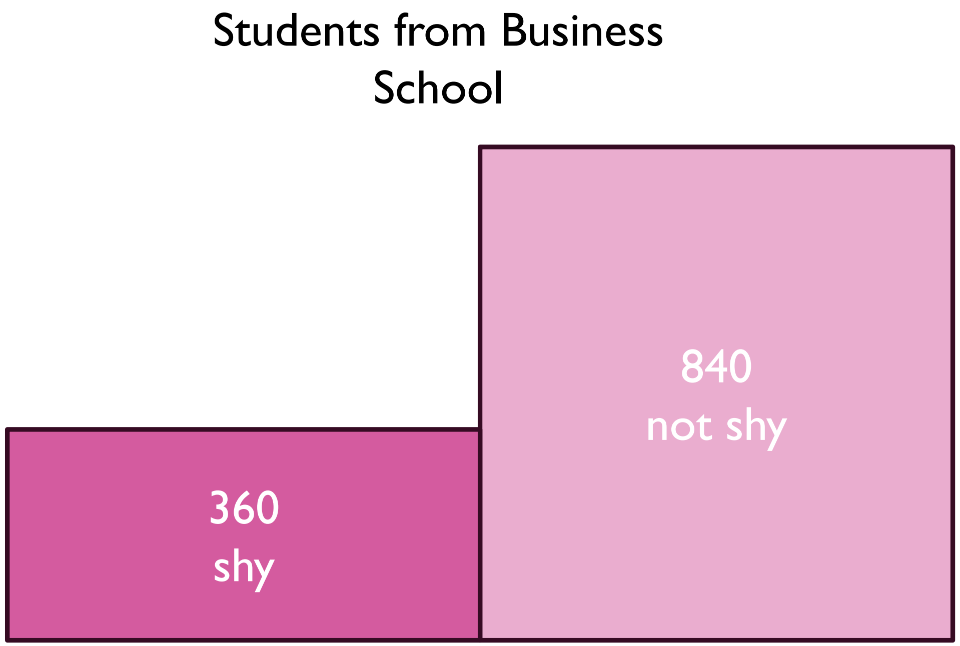
- Considering the new information of business school having 1200 students vs. 250 students on Math Dept.
- The total space of possibilities is not the same between two groups



FIRST DAY OF SCHOOL DILEMMA

- Considering the new information of business school having 1200 students vs. 250 students on Math Dept.
- The total space of possibilities is not the same between two groups

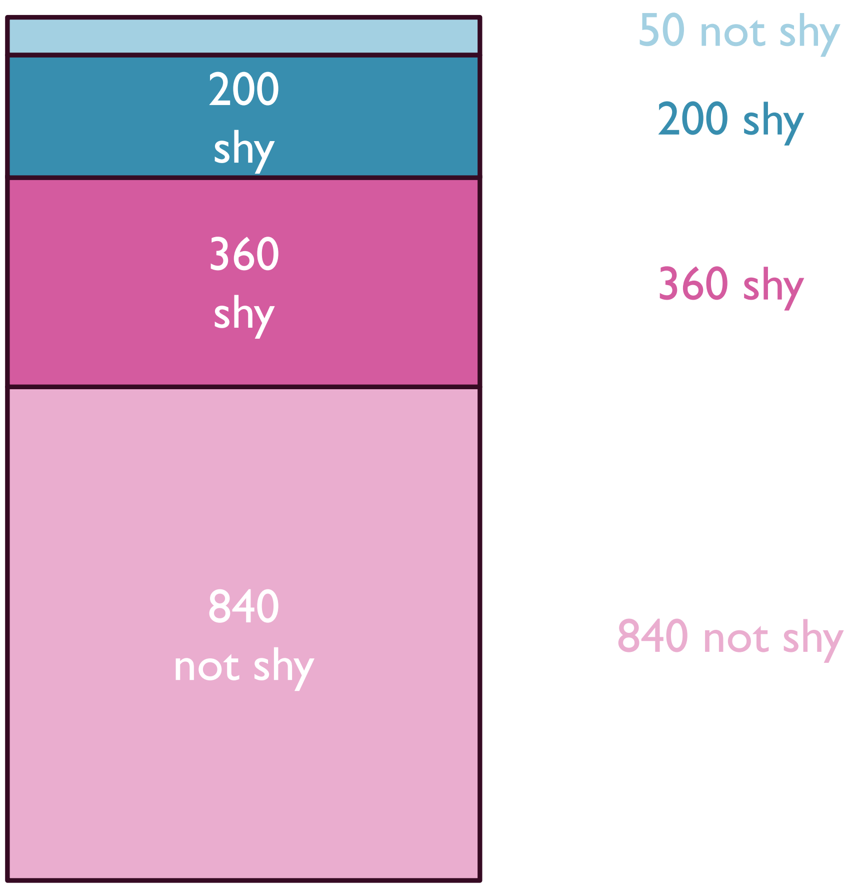




THE TRUE SPACE OF THE PROBLEM

- The total space is consisting of 1200 + 250 = 1450 students

- Next, we will see how the Bayesian Theorem could have been applied. But before that we will need a refresher for conditional and joint probabilities.



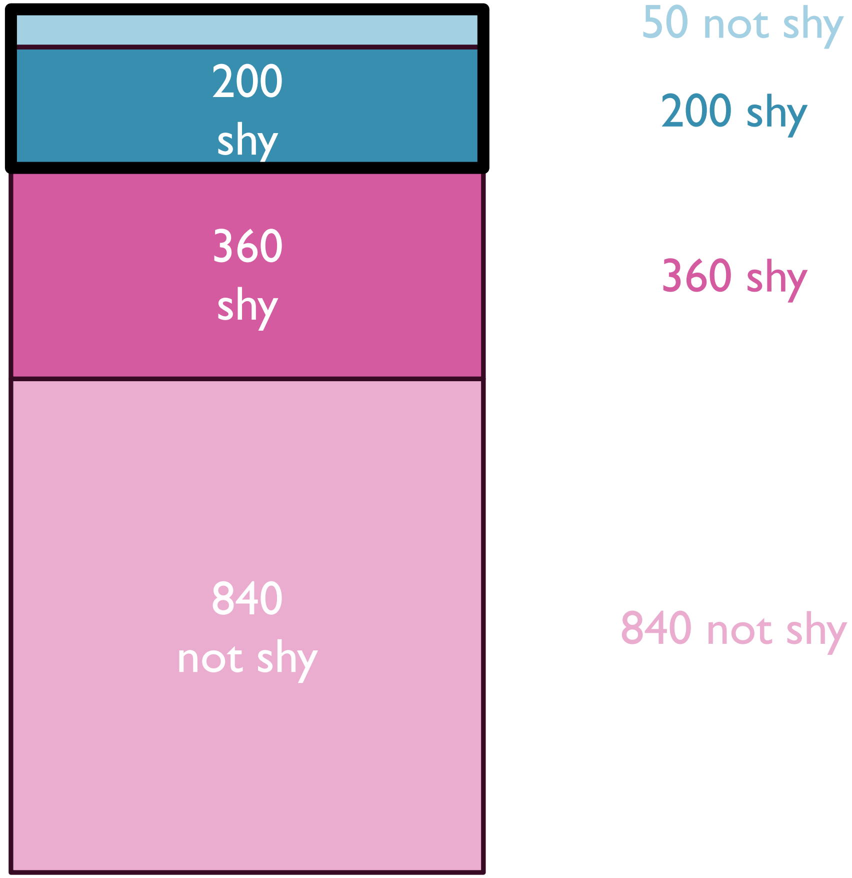
CONDITIONAL PROBABILITIES

- P(shy | Math. Dept.)
- If I know that Danny is from department of mathematics, what is the probability that he is shy?

P(shy | Math. Dept.)

$$= \frac{\text{\# students from math dept who are shy}}{\text{\# students from math dept}}$$

$$= \frac{200}{250} = 0.80$$



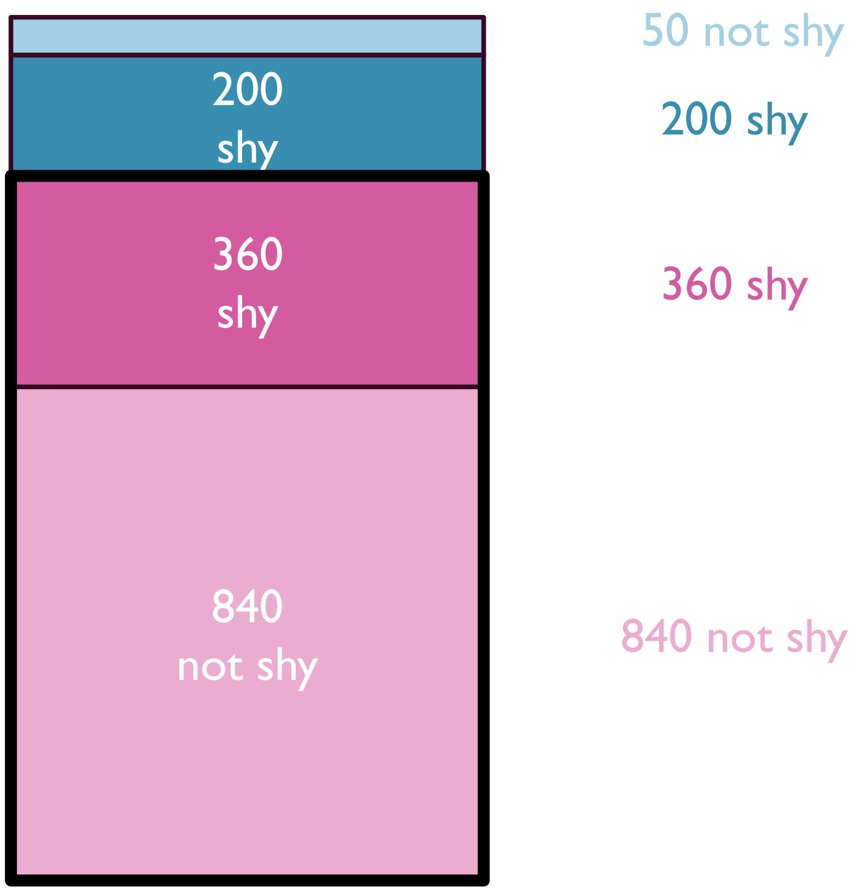
CONDITIONAL PROBABILITIES

- P(shy | Biz. School)
- If I know that Danny is from department of mathematics, what is the probability that he is shy?

P(shy | Biz. School)

$$= \frac{\text{\# students from biz. school who are shy}}{\text{\# students from biz. school}}$$

$$= \frac{360}{1200} = 0.30$$



CONDITIONAL PROBABILITIES

- P(A | B) is the probability of A, given B.
- If I know B is the case, what is the probability that A is also the case?
- P(A | B) is not the same as P(B | A).
- P(cute | puppy) is not the same as P (puppy | cute)
- If I know the thing, I’m holding is a puppy, what is the probability it is cute?
- If I know the thing, I’m holding is cute, what is the probability it is a puppy?

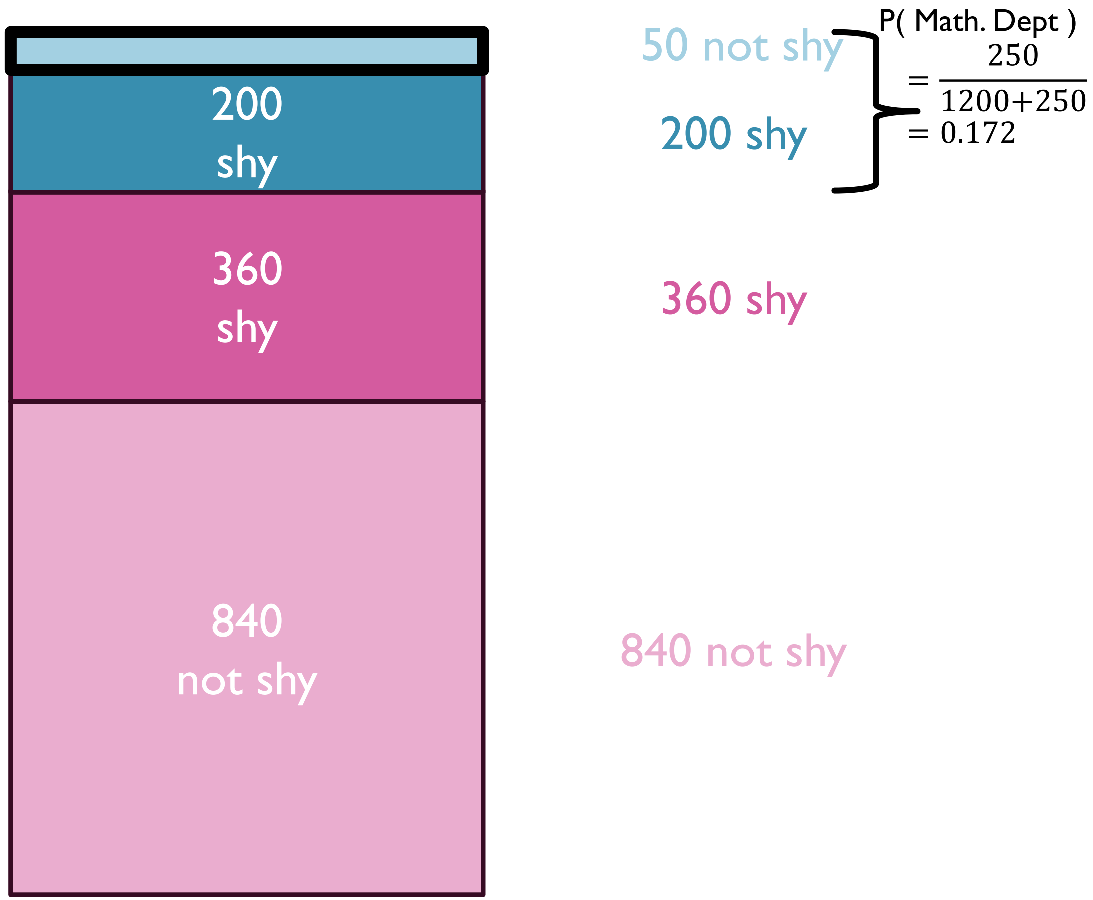
JOINT PROBABILITY

- What is the probability that a student is both from math. Dept. and is not shy?

P(Math. Dept. & not shy)

$$= P(\text{ Math. Dept. }) \times P(\text{ not shy } | \text{ Math. Dept. })$$

$$= \frac{250}{1200+250} \times \frac{50}{250} \approx 0.035$$



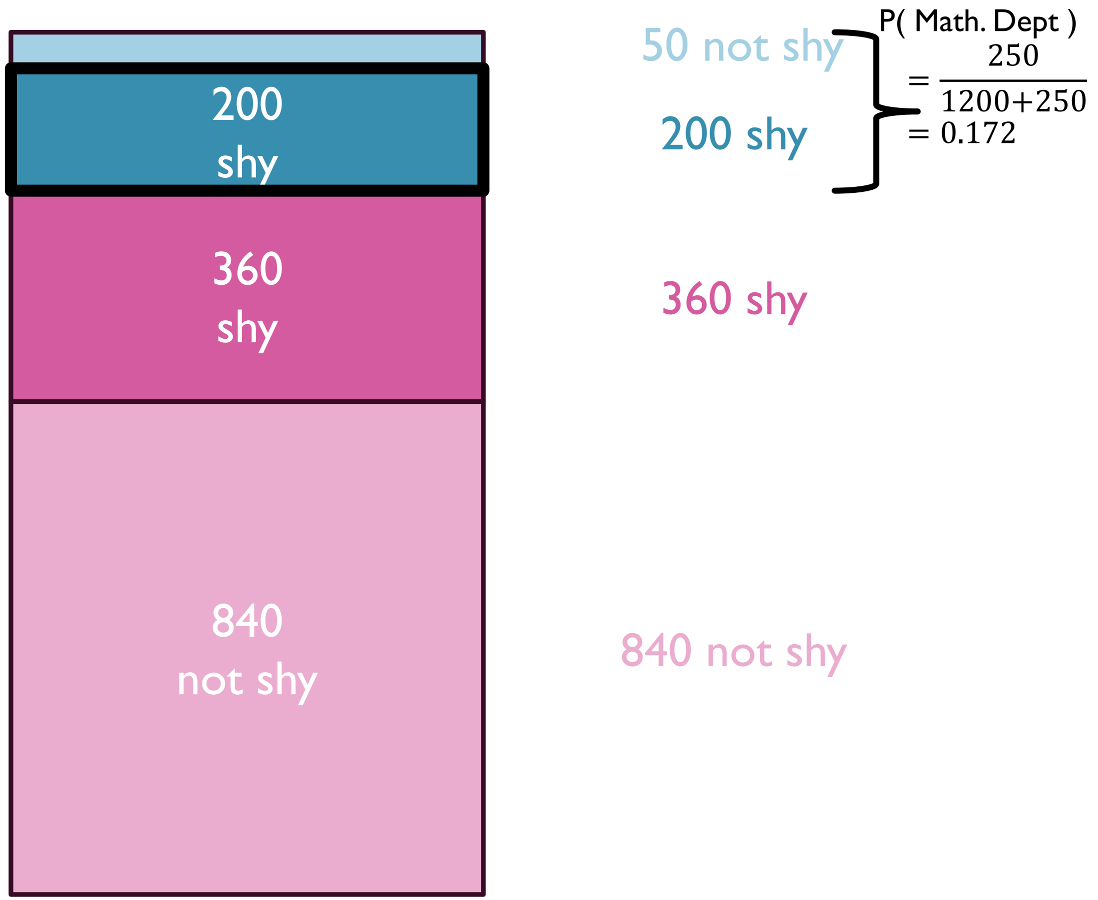
JOINT PROBABILITY

- What is the probability that a student is both from math. Dept. and is shy?

P(Math. Dept. & shy)

= P(Math. Dept.) × P(shy | Math. Dept.)

$$= \frac{250}{1200+250} \times \frac{200}{250} \approx 0.138$$



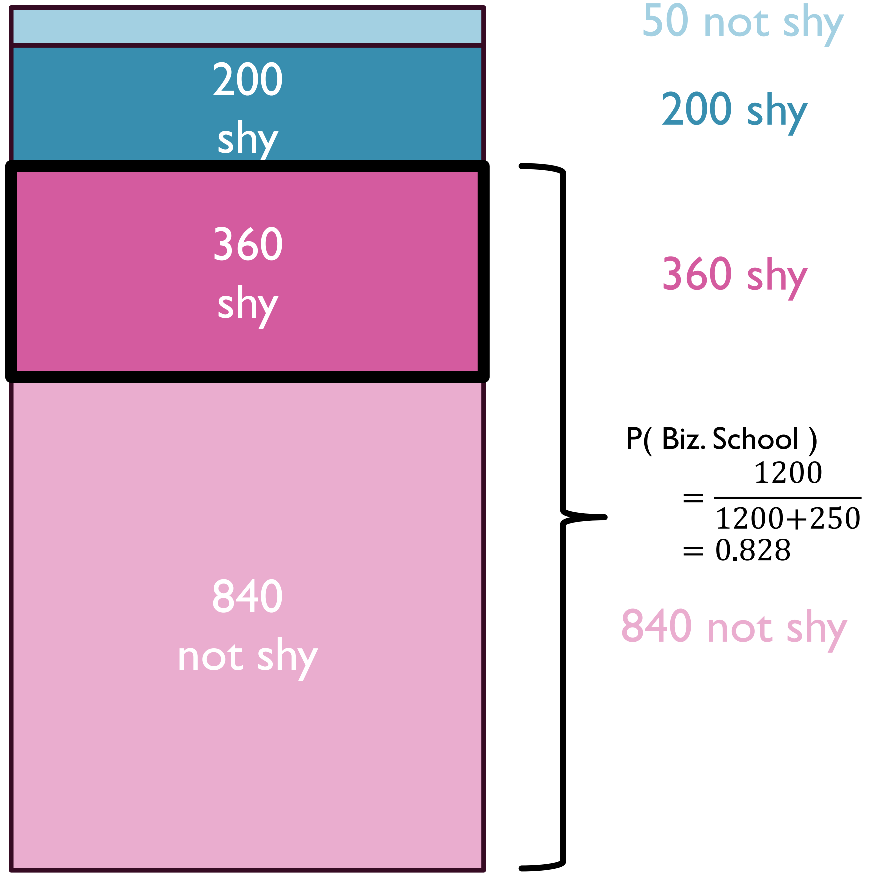
JOINT PROBABILITY

- What is the probability that a student is both from Biz. School and is shy?

P(Biz. school & shy)

= P(Biz. School) × P(shy | Biz. School)

$$= \frac{1200}{1200+250} \times \frac{360}{1200} \approx 0.248$$



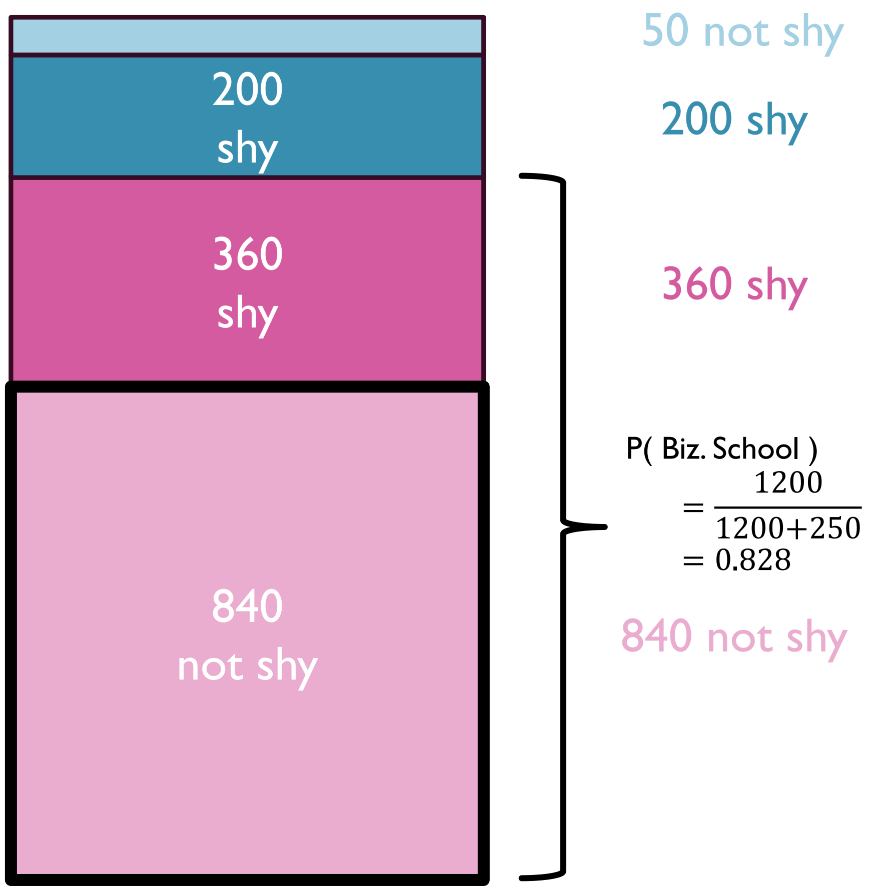
JOINT PROBABILITY

- What is the probability that a student is both from Biz. School and is not shy?

P(Biz. school & not shy)

= P(Biz. School) × P(not shy | Biz. School)

$$= \frac{1200}{1200+250} \times \frac{840}{1200} \approx 0.579$$

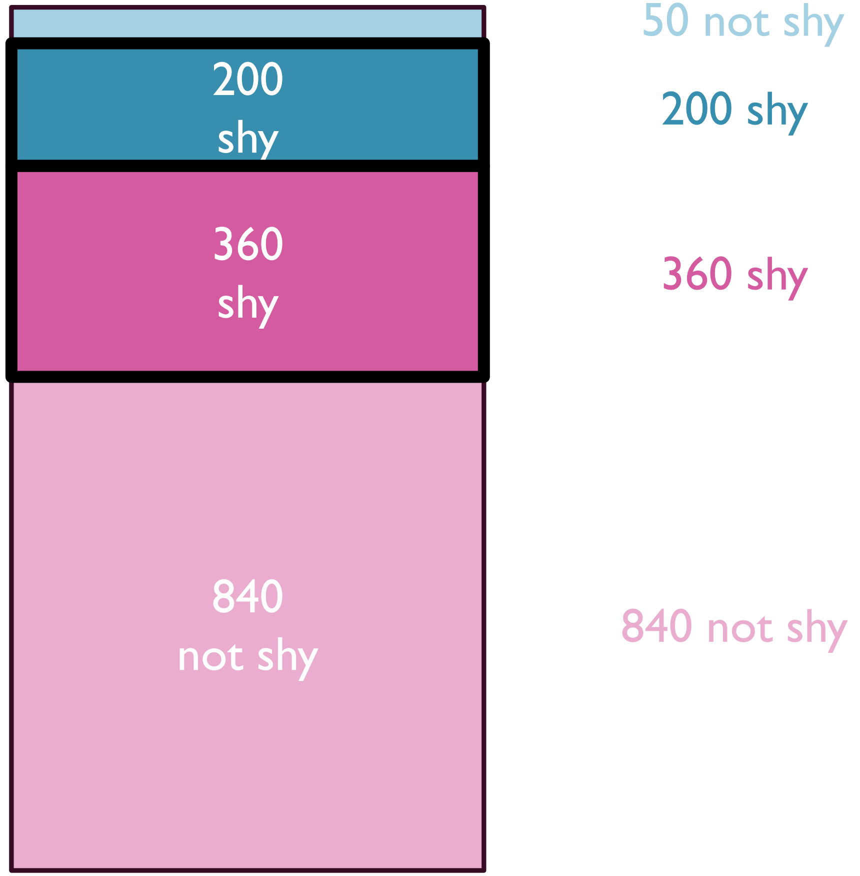


JOINT PROBABILITY

- P (A & B) is the probability that both A and B are the case.
- Also written P (A, B) or P (A ∩ B)
- P (A & B) is the same as P (B & A)
- The probability that I am having tea with biscuits is the same as the probability of I am having biscuits with tea.
- P (tea & biscuits) = P (biscuits & tea)

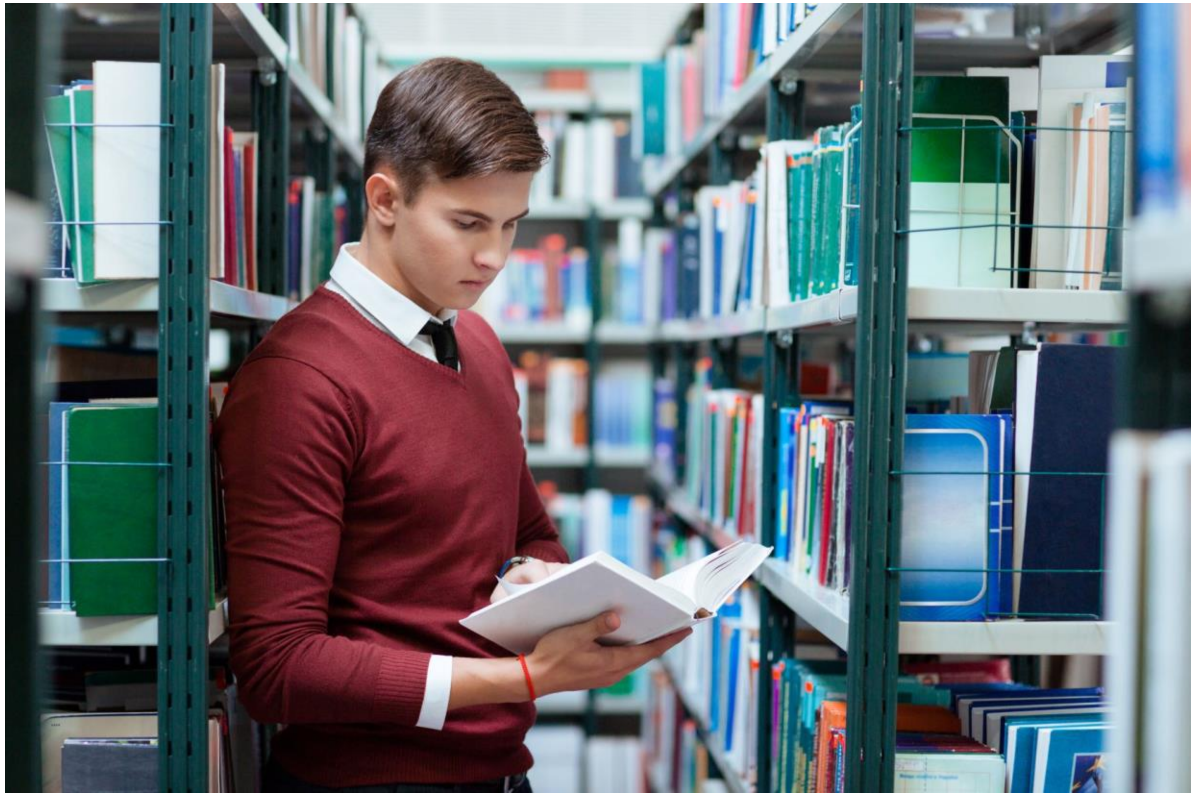
MARGINAL PROBABILITY

- P (shy) = P (Biz. School & shy) + P (Math. Dept. & shy)
  = 0.138 + 0.248 = 0.386



BACK TO OUT DILEMMA

- Knowing that Danny is shy what is the probability that he is from Math. Dept. ?
- P (Math. Dept | shy) ?
- You might have a hunch but what is the probability value?



BAYESIAN THEOREM TO RESCUE

P(Math. Dept. & shy)

= P(Math. Dept.) × P(shy | Math. Dept.)

P(shy & Math. Dept.)

= P(shy) × P(Math. Dept. | shy)

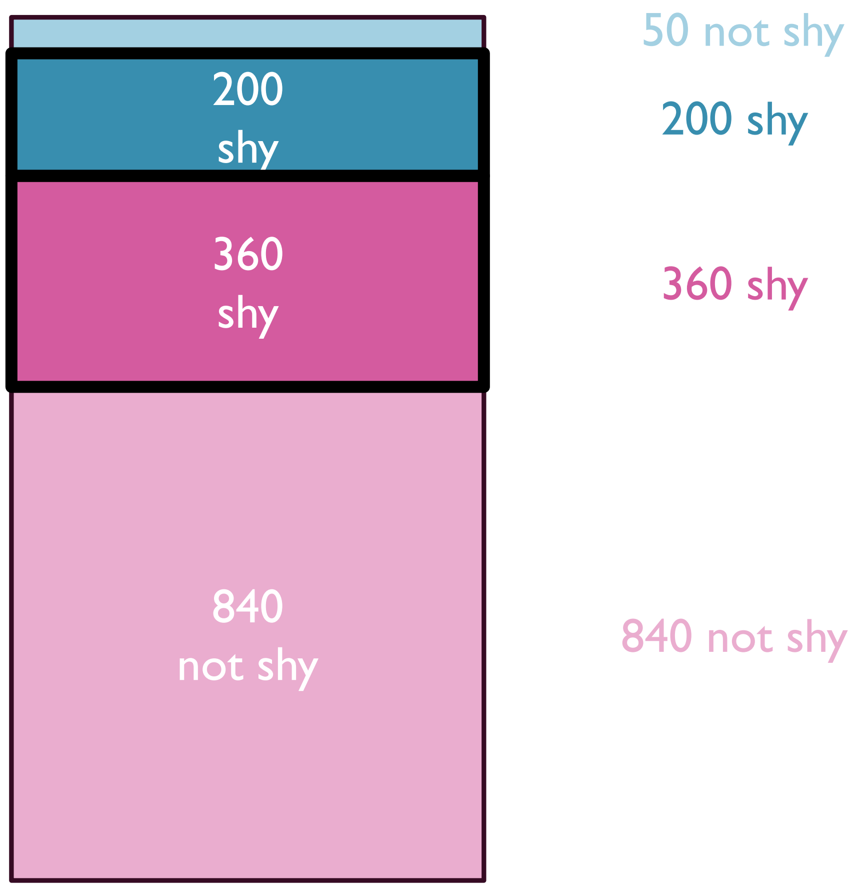
P(Math. Dept.) × P(shy | Math. Dept.)

= P(shy) × P(Math. Dept. | shy)

P(Math. Dept. | shy)

= P(Math. Dept.) × P(shy | Math. Dept.) / P(shy)

P(A | B) = P(B | A) × P(A) / P (B)



BAYES' THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

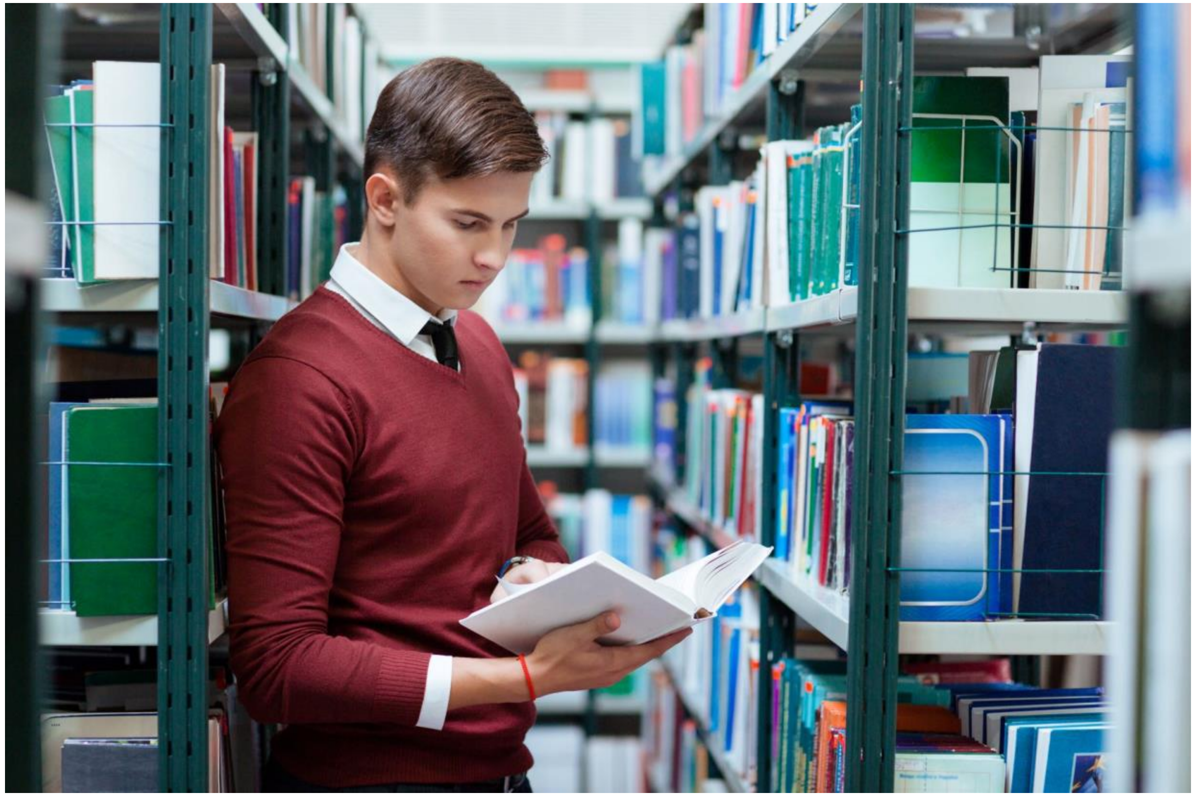
BACK TO OUR DILEMMA

- P(Math. Dept. | shy) =

  P(Math. Dept.) × P(shy | Math. Dept.) / P(shy)

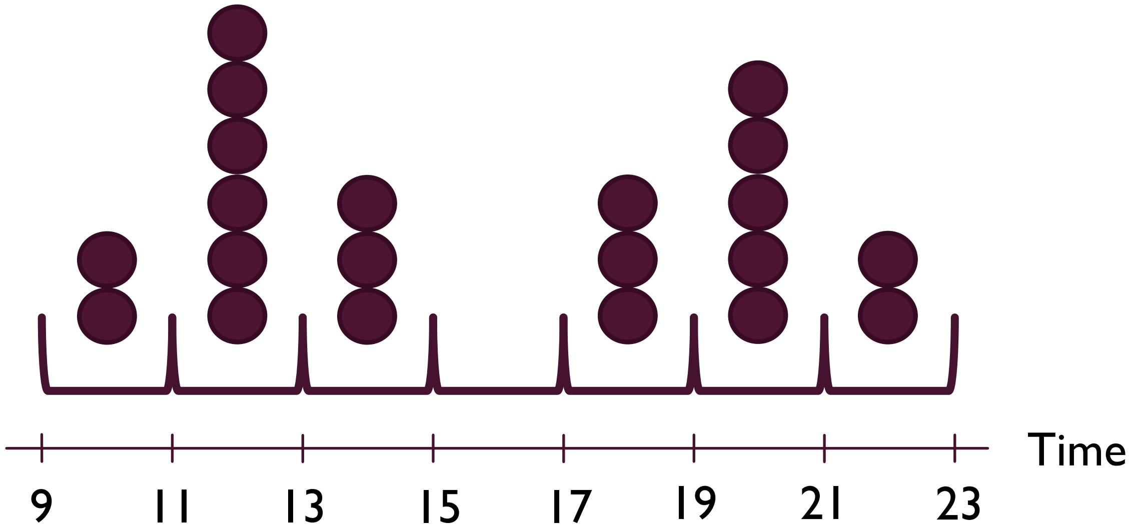
- P(Math. Dept. | shy) =

  $$\frac{0.172\times0.8}{0.386} = 0.352$$



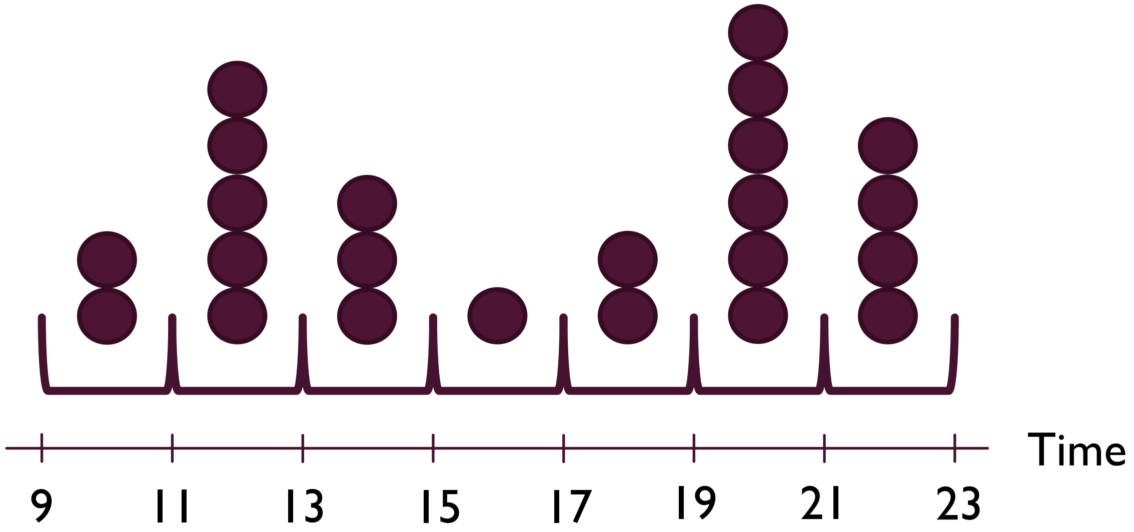
PROBABILITY DISTRIBUTIONS

- Let's say you want to keep track of the customers arrival to your restaurant.
- On 1st day



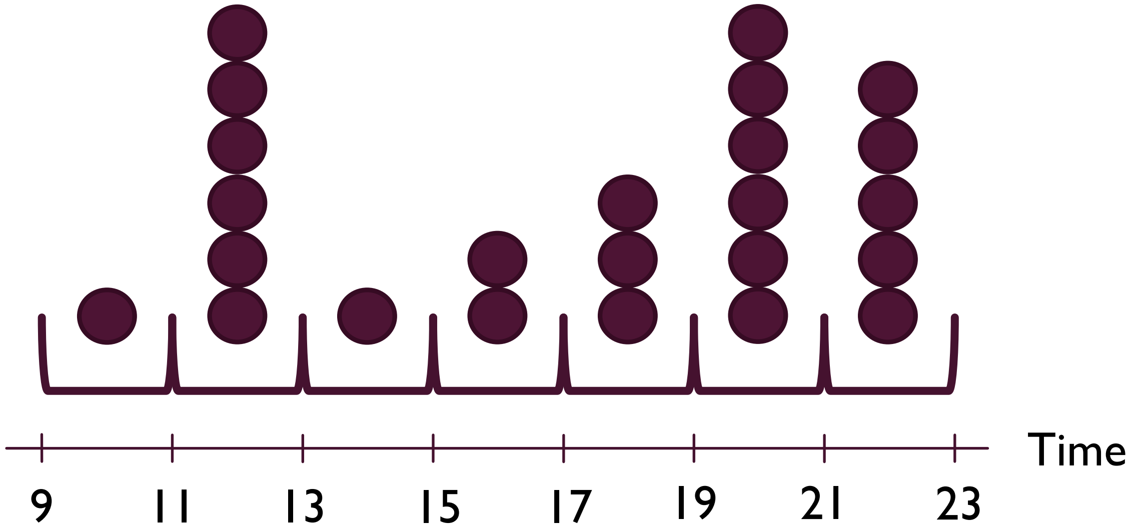
PROBABILITY DISTRIBUTIONS

- Let's say you want to keep track of the customers arrival to your restaurant.
- On 2nd day



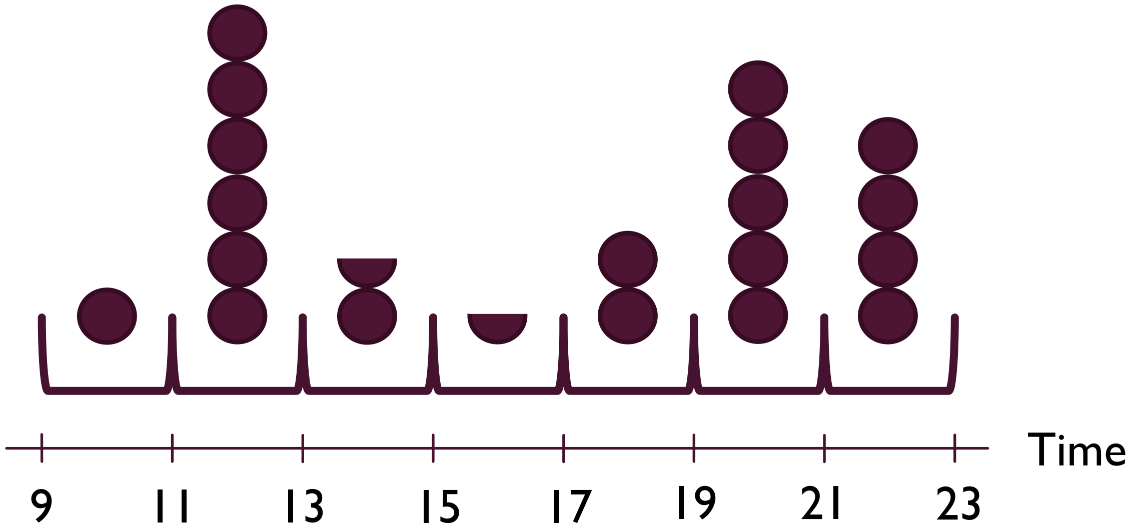
PROBABILITY DISTRIBUTIONS

- Let's say you want to keep track of the customers arrival to your restaurant.
- On 360th day



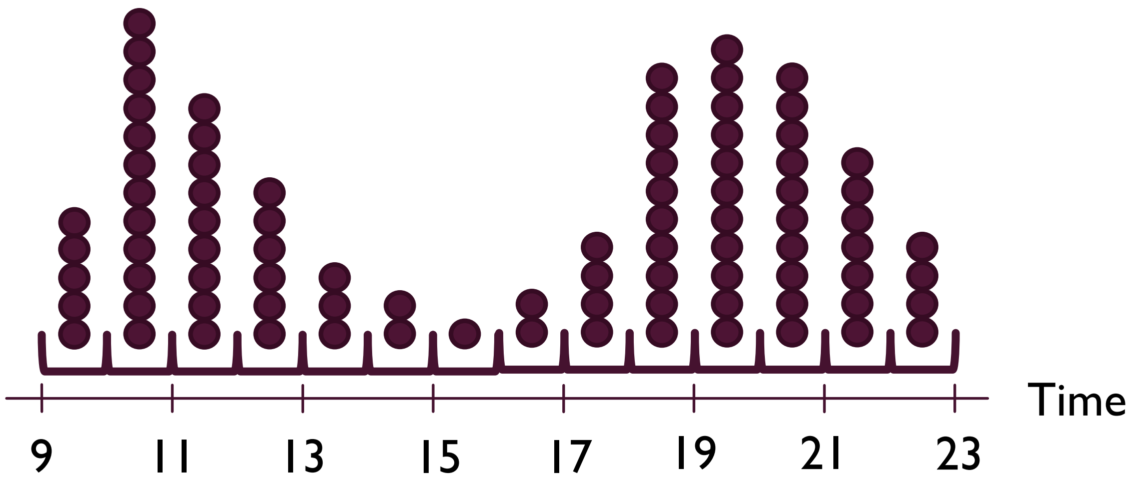
PROBABILITY DISTRIBUTIONS

- Let’s say you want to keep track of the customers arrival to your restaurant.
- Average over all days



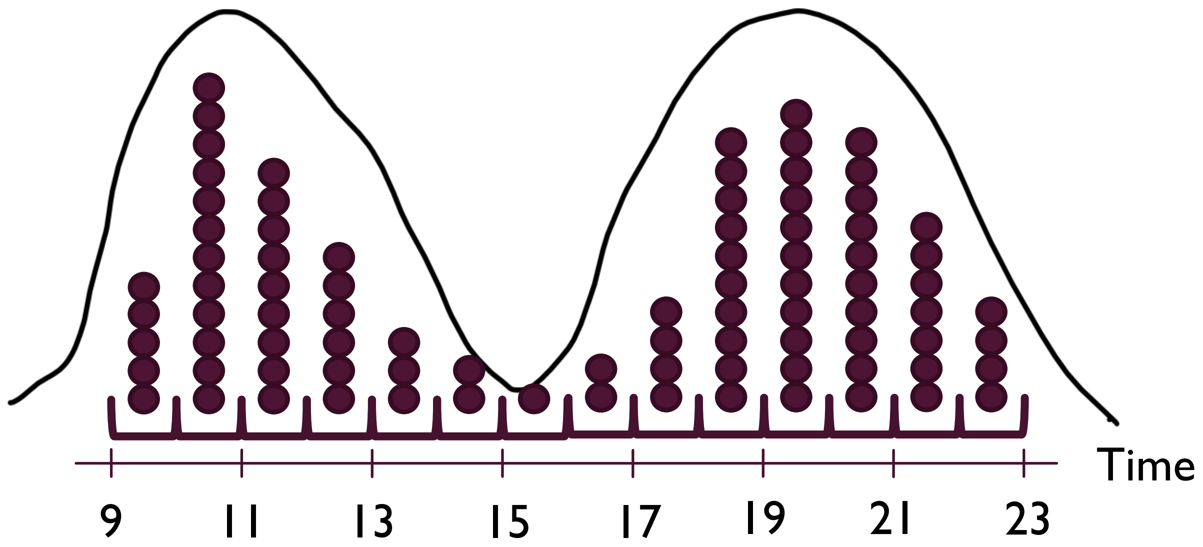
PROBABILITY DISTRIBUTIONS

- You already have collected the entire data over the 360 days.
- Now what if we place them on smaller bins?



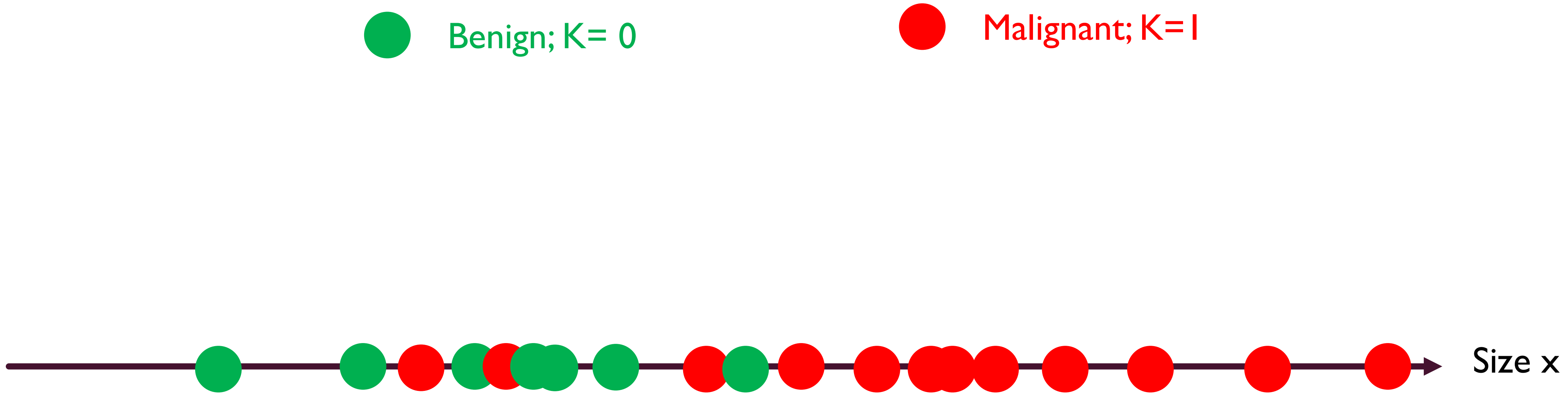
PROBABILITY DISTRIBUTIONS

- If you keep getting the bins smaller and smaller in the limit you will get a continuous distribution.



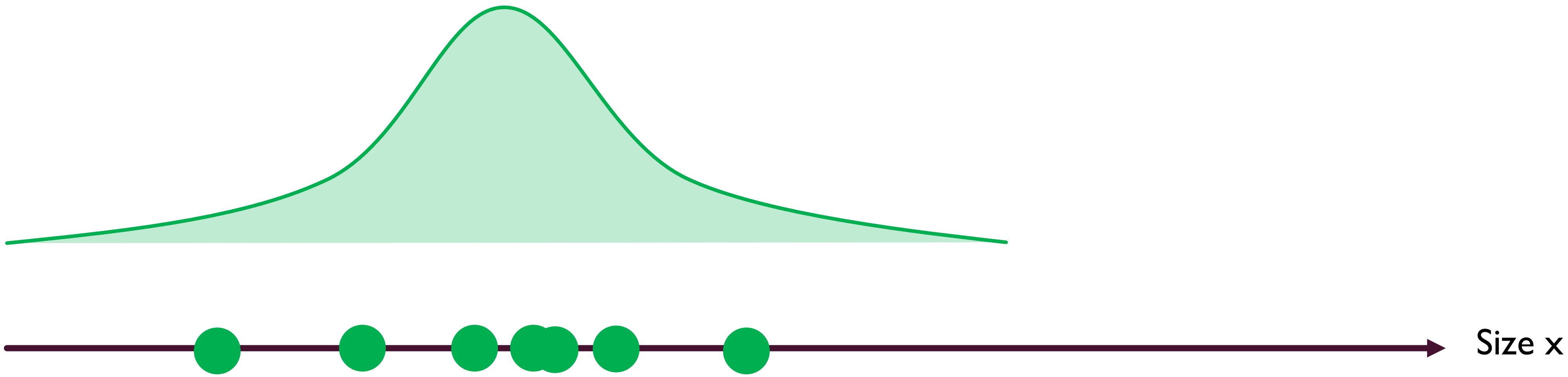
BAYESIAN CLASSIFIER

- Remember our tumor example where the malignant and benign tumors were classified based on their size?



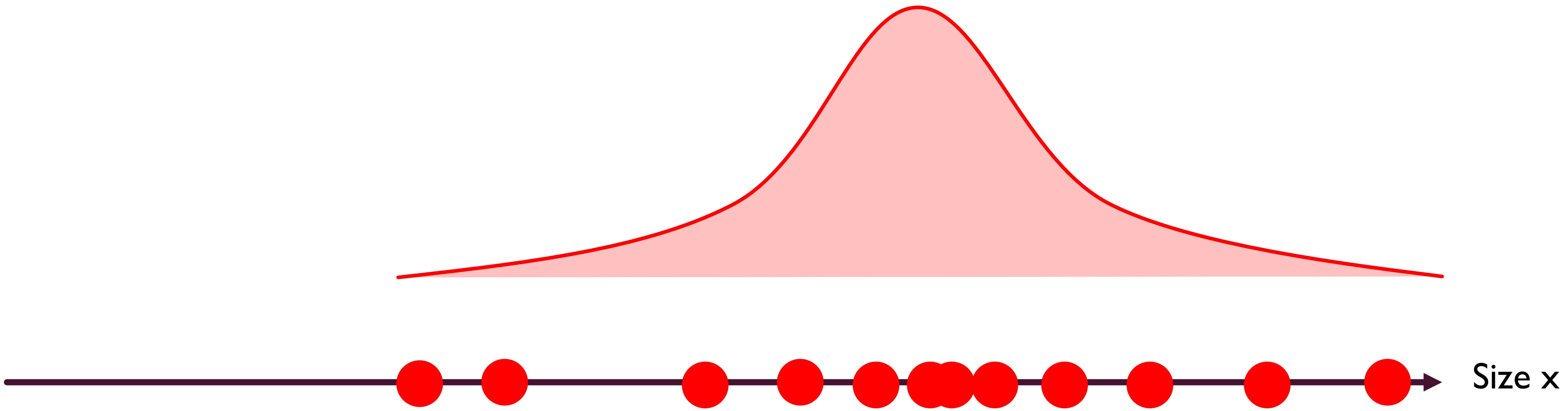
BAYESIAN CLASSIFIER

- We can try find the underlying probability density associate with each class separately.
- Assuming normal distribution here seems reasonable.



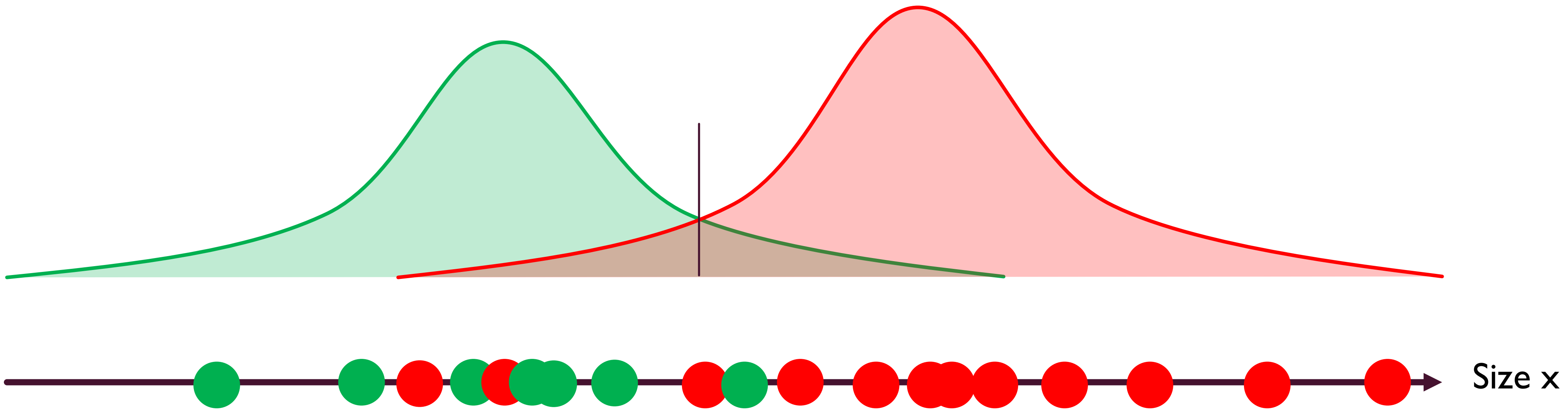
BAYESIAN CLASSIFIER

- We can try find the underlying probability density associate with each class separately.
- Assuming normal distribution here seems reasonable.



BAYESIAN CLASSIFIER

- We can try find the underlying probability density associate with each class separately.
- Assuming normal distribution here seems reasonable.



BAYESIAN CLASSIFIER MATHEMATICALLY DEFINED

- If we want to follow a probabilistic approach, we could use the following prediction model:

$$f_{\vec{W}}(\vec{x}) = \underset{K}{\operatorname{argmax}}\, P(y = K|\vec{x})$$

- To use this model, we need to know $P(y = K|\vec{x})$ for each class K. We can use Bayes' theorem:

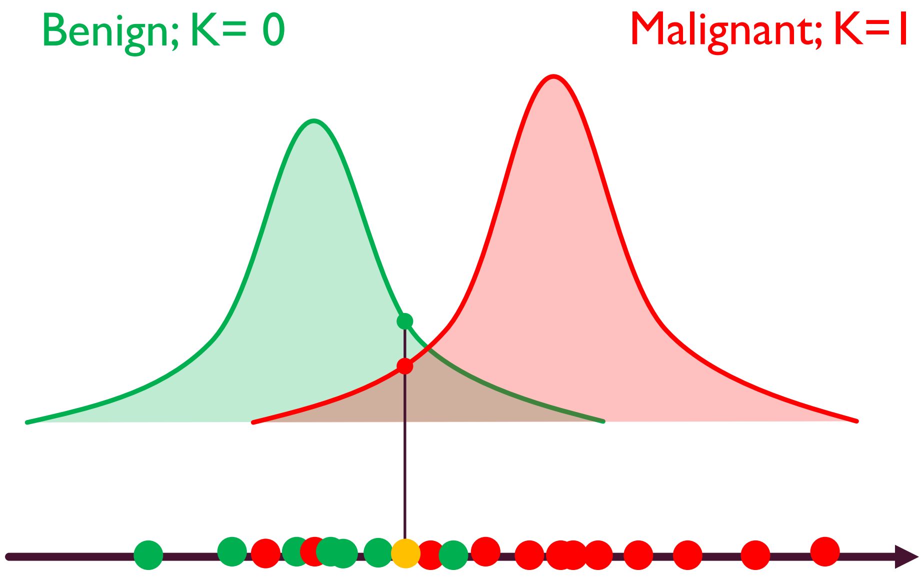
$$P(y = K|\vec{x}) = \frac{P(\vec{x}|y = K)P(y = K)}{P(\vec{x})}$$

- $P(y = K|\vec{x})$, $P(\vec{x}|y = K)$, $P(y = K)$ and $P(\vec{x})$ are called posterior, likelihood, Prior and Evidence probabilities.
- Since $P(\vec{x})$ is the same for all K and we are only interested in the max, we can throw away the denominator:

$$P(y = K|\vec{x}) \propto P(\vec{x}|y = K)P(y = K)$$

- To use this equation, we need to decide on forms for $P(\vec{x}|y = K)$ and $P(y = K)$ and figure out how we will learn their parameters $\vec{W}$ from the training data $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$.

ONE DIMENSIONAL BAYESIAN CLASSIFIER



- Given a new feature x what is the probability it is from class 0?

$$P(y=0|x) = ?$$

- Based on Base's Theorem this is equivalent to:

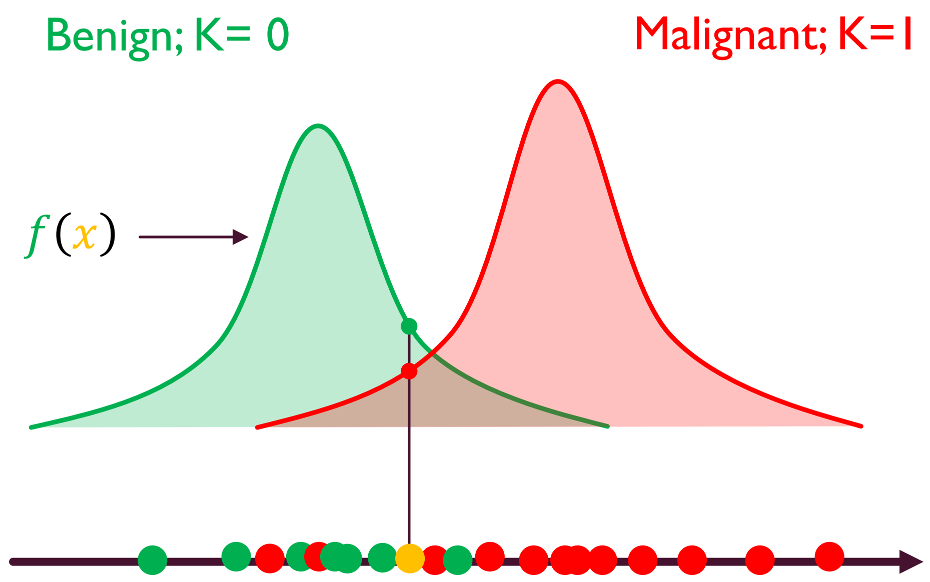
$$P(y=0|x) = \frac{P(x|y=0)P(y=0)}{P(x)}$$

- For now, lets ignore $P(x)$

$$P(y=0|x) \propto P(x|y=0)P(y=0)$$

- Then we need to find these two quantities, $P(x|y=0)$ and $P(y=0)$.

ONE DIMENSIONAL BAYESIAN CLASSIFIER



- If for instance, in our training samples we have 7 cases of benign tumors (K=0) and 12 cases of malignant tumors (K=1) it makes sense to assume probability of a new sample belonging to class 0 as follow:

$$P(y=0)=\frac{7}{7+12}=\frac{7}{19}$$

- $P(x|y=0)$ is the probability of the new feature x "knowing" that it belongs to benign tumor class!
- Now the question becomes: what is the probability density function of class 0?
- One assumption is normal distribution!

$$P(x|y=0)=f(x)=\frac{1}{\sigma_0\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma_0}\right)^2}$$

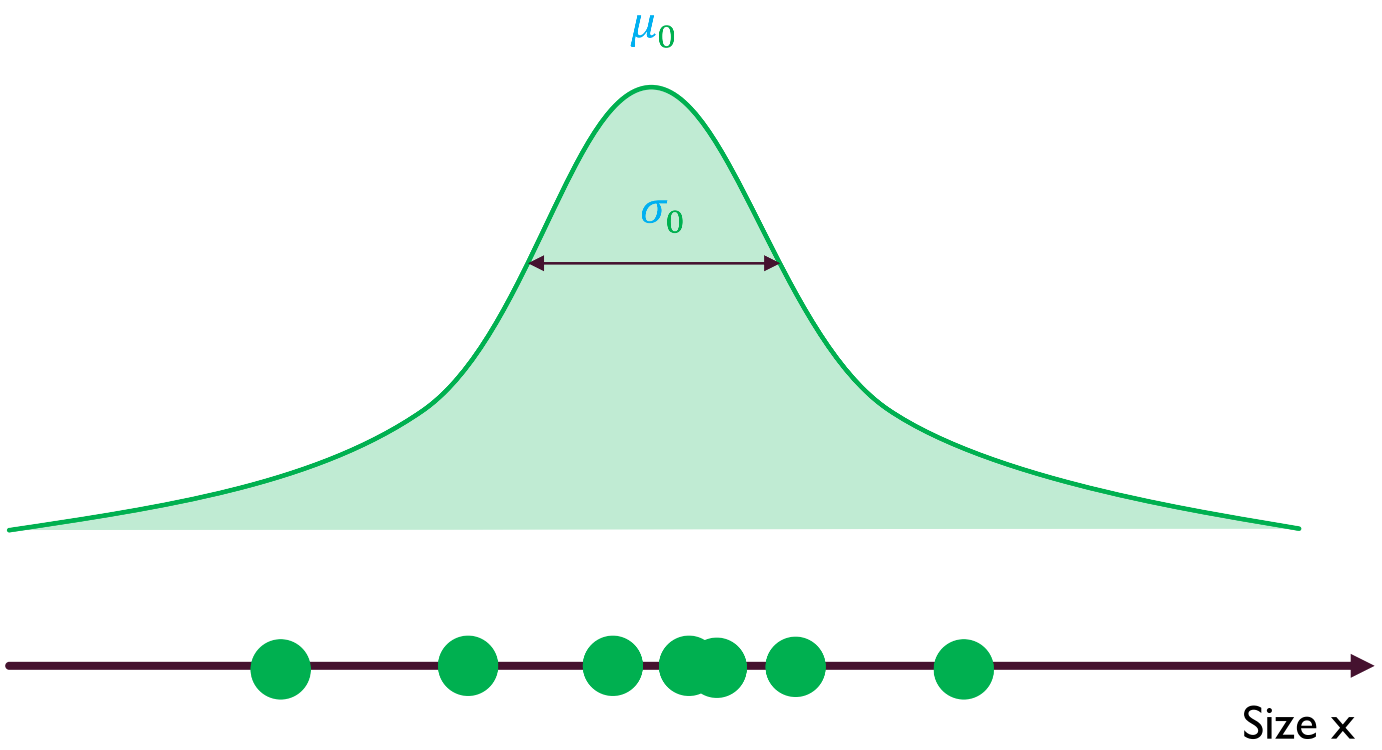
ESTIMATING NORMAL DISTRIBUTION PARAMETERS

- Now that we assumed our model has normal distribution (Gaussian), we need to estimate its parameters μ_0 (mean) and σ_0 (variance).
- Using the training samples one can use the maximum likelihood estimations to approximate these parameters.

$$\mu_0 \approx \hat{\mu}_0 = \frac{1}{m_0} \sum_i x^{(i)}$$

$$\sigma_0{}^2 \approx \hat{\sigma}_0{}^2 = \frac{1}{m_0} \sum_i \left(x^{(i)} - \hat{\mu}_0\right)^2$$

- $P(y = 0|x) \propto \frac{1}{\hat{\sigma}_0\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\hat{\mu}_0}{\hat{\sigma}_0}\right)^2} \times \frac{7}{19}$



ESTIMATING NORMAL DISTRIBUTION PARAMETERS

- Now that we assumed our model has normal distribution (Gaussian), we need to estimate its parameters μ_0 (mean) and σ_0 (variance).
- Using the training samples one can use the maximum likelihood estimations to approximate these parameters.

$$\mu_0 \approx \hat{\mu}_0 = \frac{1}{m_0}\sum_i x^{(i)} \qquad ; {\sigma_0}^2 \approx {\hat{\sigma}_0}^2 = \frac{1}{m_0}\sum_i \left(x^{(i)} - \hat{\mu}_0\right)^2$$

$$\mu_1 \approx \hat{\mu}_1 = \frac{1}{m_1}\sum_i x^{(i)} \qquad ; {\sigma_1}^2 \approx {\hat{\sigma}_1}^2 = \frac{1}{m_1}\sum_i \left(x^{(i)} - \hat{\mu}_1\right)^2$$

- $P(y = 0|x) \propto \frac{1}{\hat{\sigma}_0\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\hat{\mu}_0}{\hat{\sigma}_0}\right)^2} \times \frac{7}{19}$
- $P(y = 1|x) \propto \frac{1}{\hat{\sigma}_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\hat{\mu}_1}{\hat{\sigma}_1}\right)^2} \times \frac{12}{19}$

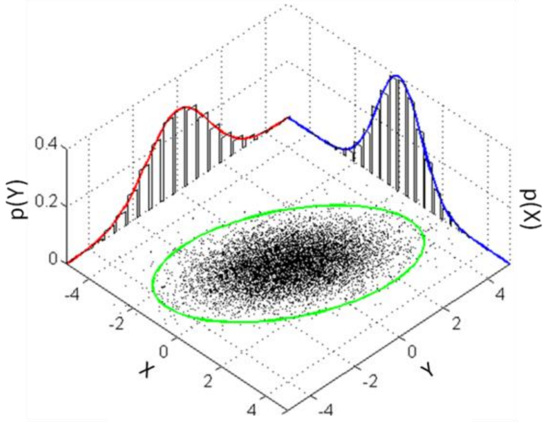
- If $\hat{y} = \begin{cases} 0 & \text{if } P(y = 0|x) \geq P(y = 1|x) \\ 1 & \text{if } P(y = 0|x) < P(y = 1|x) \end{cases}$

MULTIVARIATE NORMAL DISTRIBUTION

- One common way to calculate $P(y = K)$ is to simply count the number of training points assigned to each class and divide it to the total number of training samples.

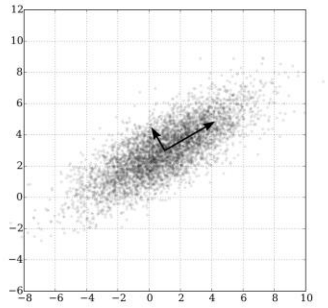
$$P(y = K) = \frac{\sum_{i=1}^{m} \mathrm{I}\left(y^{(i)} = K\right)}{m} = \frac{m_K}{m}$$

- For $P(\vec{x}|y = K)$ a common choice is to use a Multivariate Normal Distribution



$$P(\vec{x}|y = K) = \mathcal{N}\left(\vec{x}^{(i)};\ \vec{\mu}_K, \mathbf{\Sigma}_K\right)$$

$$\mathcal{N}(\vec{x};\ \vec{\mu}_K, \mathbf{\Sigma}_K) = \frac{1}{\sqrt{(2\pi)^2|\mathbf{\Sigma}_K|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_K)^T \mathbf{\Sigma}_K{}^{-1}(\vec{x}-\vec{\mu}_K)}$$



QUADRATIC AND LINEAR DISCRIMINANT

$$P(\vec{x}|y=K) = \mathcal{N}\left(\vec{x}^{(i)};\ \vec{\mu}_K, \boldsymbol{\Sigma}_K\right)$$

- Parameters $\overrightarrow{\boldsymbol{W}} = \vec{\mu}_K, \boldsymbol{\Sigma}_K$ are the mean vector and covariance matrix associated with normal distribution in N dimensional space and must be fitted to the data in each class K. This can be done using a maximum likelihood estimator. This is called quadratic discriminant analysis (QDA).
- This means the total number of parameters $\overrightarrow{\boldsymbol{W}}$ that need to be found is $K(N+N^2)$.
- One way is to assume all the classes have the same covariance matrix $\boldsymbol{\Sigma}_K$ and then only fit the mean values $\vec{\mu}_K$. This is called linear discriminant analysis (LDA).

GAUSSIAN NAÏVE BAYES

- In Gaussian naïve Bayes, we assume that each feature is independent, i.e. each dimension of $\vec{x}$ is independent.

$$P(\vec{x}|y = K, \vec{w}) = \prod_{n=1}^{N} P(x_n|y = K; \vec{w})$$

- Naïve Bayes assumption can be made for any distribution, not just Gaussians. For the Gaussian case, it leads to

$$P(\vec{x}|y = K, \vec{w}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n; \mu_{K,n}, \sigma_{K,n}^2\right)$$

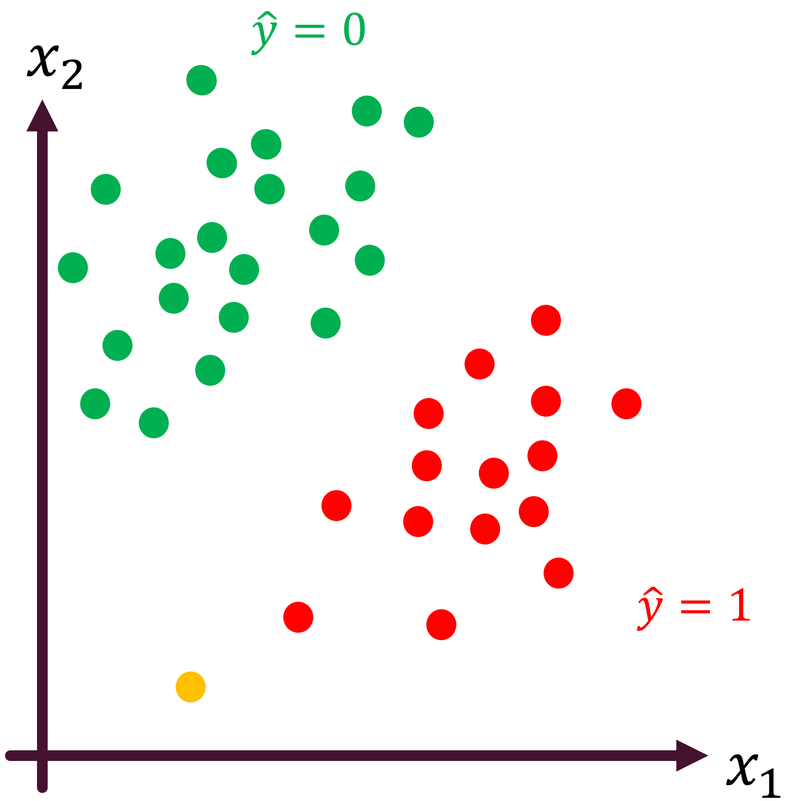
- For the case of Gaussian Naïve Bayes, we can use the simple formulas based on maximum likelihood estimate:

$$\mu_{K,n} = \frac{1}{m_K} \sum_{i=1}^{m_K} x_n^{(i)}$$

$$\sigma_{K,n}^2 = \frac{1}{m_K} \sum_{i=1}^{m_K} \left(x_n^{(i)} - \mu_{K,n}\right)^2$$

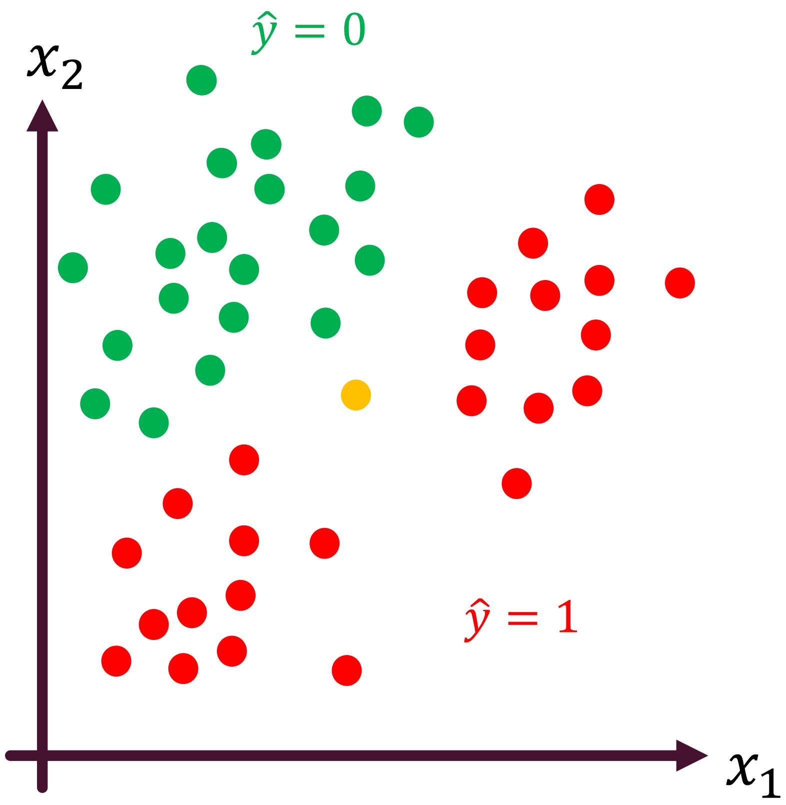
- m_K is the total number of training data which belong to class K.

BAYESIAN CLASSIFIER INTUITION



- $P(y = K|\vec{x}) \propto P(\vec{x}|y = K)P(y = K)$
- Need to fit $P(y = K)$ and $P(\vec{x}|y = K)$ from these data which have two features each.
- What is the intuitive way to calculate the following?
  - $P(y = 0)$
  - $P(y = 1)$
  - $P(\vec{x}|y = 0)$
  - $P(\vec{x}|y = 1)$

BAYESIAN CLASSIFIER INTUITION

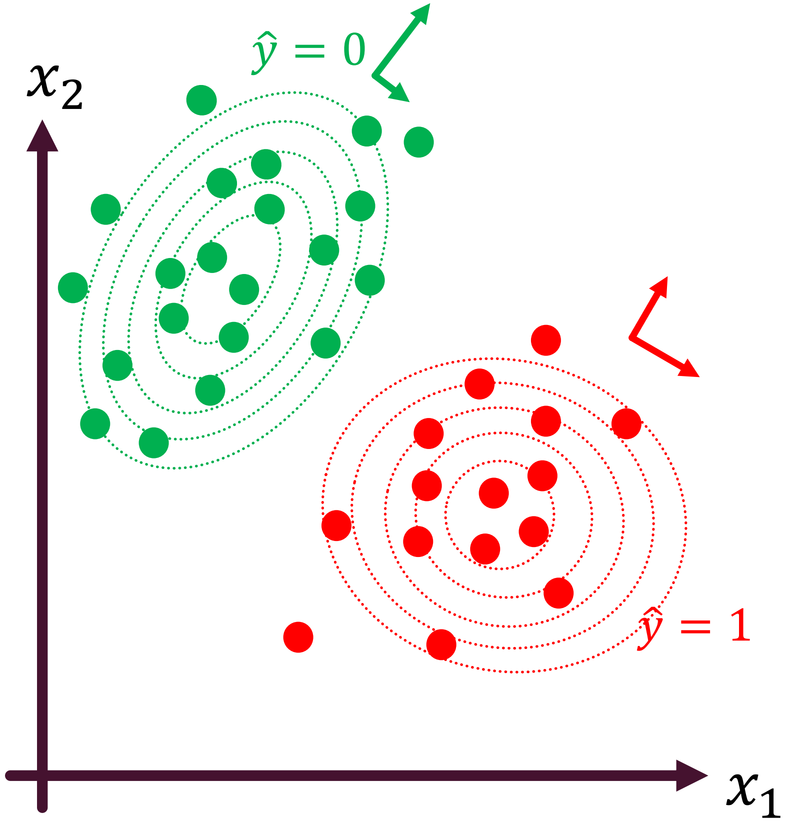


$P(\vec{x}|y = 0)$
$P(\vec{x}|y = 1)$

- $P(y = K|\vec{x}) \propto P(\vec{x}|y = K)P(y = K)$
- Need to fit $P(y = K)$ and $P(\vec{x}|y = K)$ from these data which have two features each.

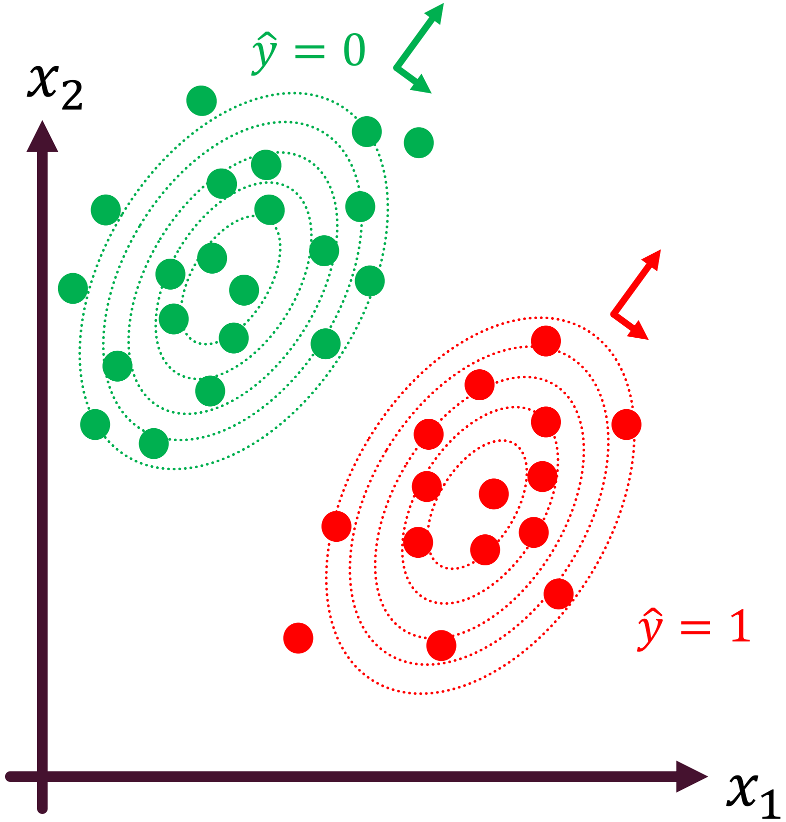
- What is the intuitive way to calculate the following?
  - $P(y = 0)$
  - $P(y = 1)$
  - $P(\vec{x}|y = 0)$
  - $P(\vec{x}|y = 1)$

MULTIVARIATE BAYESIAN CLASSIFIER (QDA)



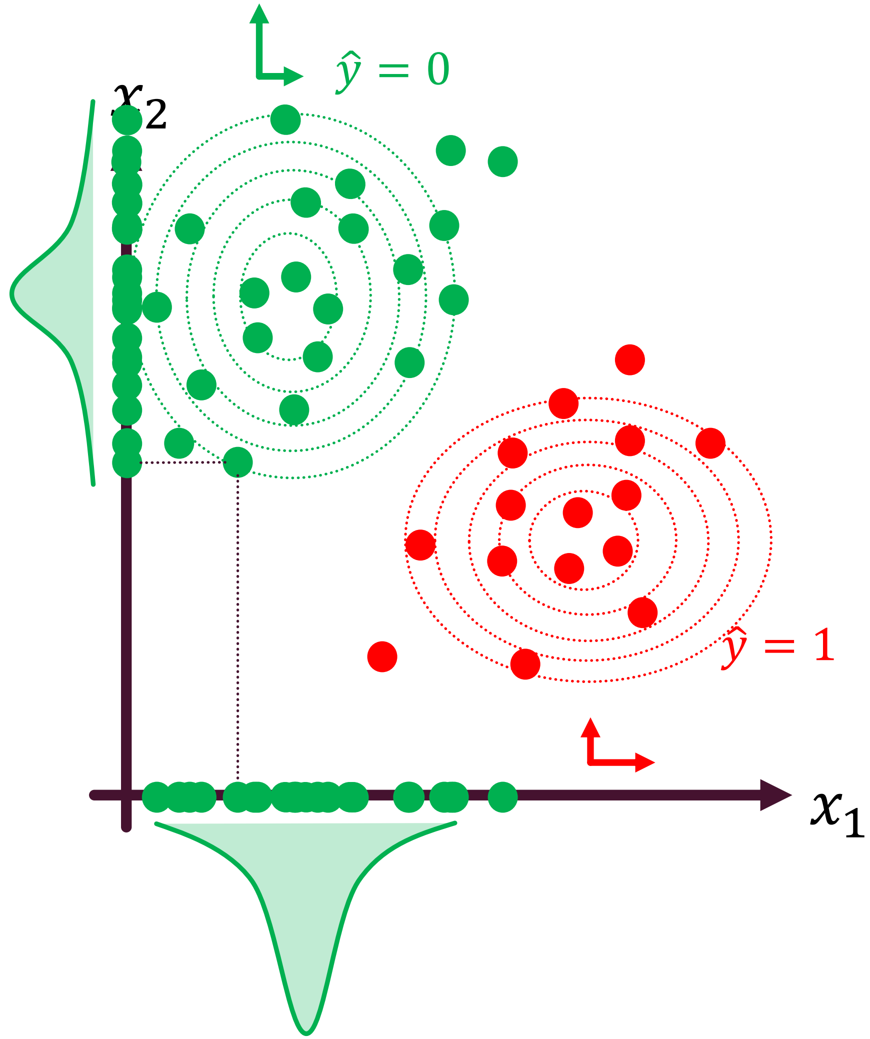
- What is the intuitive way to calculate the following?
  - $P(y = 0) = \frac{21}{21+15}$
  - $P(y = 1) = \frac{15}{21+15}$
  - $P(\vec{x}|y = 0)$, contours are concentric ellipses
  - $P(\vec{x}|y = 1)$, contours are concentric ellipses
  - The ellipse have different angles and formation

MULTIVARIATE BAYESIAN CLASSIFIER (LDA)



- What is the intuitive way to calculate the following?
  - $P(y = 0) = \frac{21}{21+15}$
  - $P(y = 1) = \frac{15}{21+15}$
  - $P(\vec{x}|y = 0)$, contours are concentric ellipses
  - $P(\vec{x}|y = 1)$, contours are concentric ellipses
  - The ellipse have the same angles and formation

NAÏVE BAYESIAN CLASSIFIER



- What is the intuitive way to calculate the following?
  - $P(y = 0) = \frac{21}{21+15}$
  - $P(y = 1) = \frac{15}{21+15}$
  - $P(\vec{x}|y = 0)$, contours are concentric ellipses
  - $P(\vec{x}|y = 1)$, contours are concentric ellipses
  - Both ellipses have angle of zero, but their formation is different

REFERENCES

- E2EML 191. How Selected Models and Methods Work, Brandon Rohrer, https://end-to-end-machine-learning.teachable.com/p/machine-learning-signal-processing-statistics-concepts/
- DatA414. Yet another introduction to machine, Herman Kamper, https://www.kamperh.com/data414/

https://towardsdatascience.com/animations-of-logistic-regression-with-python-31f8c9cb420