ECE 5984: Introduction to Machine Learning

Topics:

- Probability Review
- Statistical Estimation (MLE)

Readings: Barber 8.1, 8.2

Dhruv Batra

Virginia Tech

Administrativia

- HW1
  - Due on Sun 02/15, 11:55pm
  - http://inclass.kaggle.com/c/VT-ECE-Machine-Learning-HW1

Project

- Groups of 1-3
  - we prefer teams of 2

- Deliverables:
  - Project proposal (NIPS format): 2 page, due Feb 24
  - Midway presentations (in class)
  - Final report: webpage with results

Proposal

- 2 Page (NIPS format)
  - http://nips.cc/Conferences/2013/PaperInformation/StyleFiles
- Necessary Information:
  - Project title
  - Project idea.
    - This should be approximately two paragraphs.
  - Data set details
    - Ideally existing dataset. No data-collection projects.
  - Software
    - Which libraries will you use?
    - What will you write?
  - Papers to read.
    - Include 1-3 relevant papers. You will probably want to read at least one of them before submitting your proposal.
  - Teammate
    - Will you have a teammate? If so, what's the break-down of labor? Maximum team size is 3 students.
  - Mid-sem Milestone
    - What will you complete by the project milestone due date? Experimental results of some kind are expected here.

Project

- Rules
  - Must be about machine learning
  - Must involve real data
    - Use your own data or take from class website
  - Can apply ML to your own research.
    - Must be done this semester.
  - OK to combine with other class-projects
    - Must declare to both course instructors
    - Must have explicit permission from BOTH instructors
    - Must have a sufficient ML component
  - Using libraries
    - No need to implement all algorithms
    - OK to use standard SVM, MRF, Decision-Trees, etc libraries
    - More thought+effort => More credit

Project

- Main categories
  - Application/Survey
    - Compare a bunch of existing algorithms on a new application domain of your interest
  - Formulation/Development
    - Formulate a new model or algorithm for a new or old problem
  - Theory
    - Theoretically analyze an existing algorithm

- Support
  - List of ideas, pointers to dataset/algorithms/code
    - https://filebox.ece.vt.edu/~s15ece5984/project.html
    - We will mentor teams and give feedback.

Administrativia

- HW1
  - Due on Sun 02/15, 11:55pm
  - http://inclass.kaggle.com/c/VT-ECE-Machine-Learning-HW1

- Project Proposal
  - Due: Tue 02/24, 11:55 pm
  - <=2pages, NIPS format

Procedural View

- Training Stage:
  - Raw Data → x (Feature Extraction)
  - Training Data { (x,y) } → f (Learning)

- Testing Stage
  - Raw Data → x (Feature Extraction)
  - Test Data x → f(x) (Apply function, Evaluate error)

Statistical Estimation View

- Probabilities to rescue:
  - x and y are *random variables*
  - $D = (x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N) \quad \sim P(X,Y)$

- IID: Independent Identically Distributed
  - Both training & testing data sampled IID from P(X,Y)
  - Learn on training set
  - Have some hope of *generalizing* to test set

Plan for Today

- Review of Probability
  - Discrete vs Continuous Random Variables
  - PMFs vs PDF
  - Joint vs Marginal vs Conditional Distributions
  - Bayes Rule and Prior

- Statistical Learning / Density Estimation
  - Maximum Likelihood
  - Maximum A Posteriori
  - Bayesian Estimation

- We will discuss simple examples (like coin toss), but these SAME concepts will apply to sophisticated problems.

Probability

- The world is a very uncertain place
- 30 years of Artificial Intelligence and Database research danced around this fact
- And then a few AI researchers decided to use some ideas from the eighteenth century

Probability

- A is non-deterministic event
  - Can think of A as a boolean-valued variable

- Examples
  - A = your next patient has cancer
  - A = Rafael Nada wins French Open 2015

Interpreting Probabilities

- What does P(A) mean?

- Frequentist View
  - limit $N \rightarrow \infty$ #(A is true)/N
  - limiting frequency of a repeating non-deterministic event

- Bayesian View
  - P(A) is your “belief” about A

- Market Design View
  - P(A) tells you how much you would bet

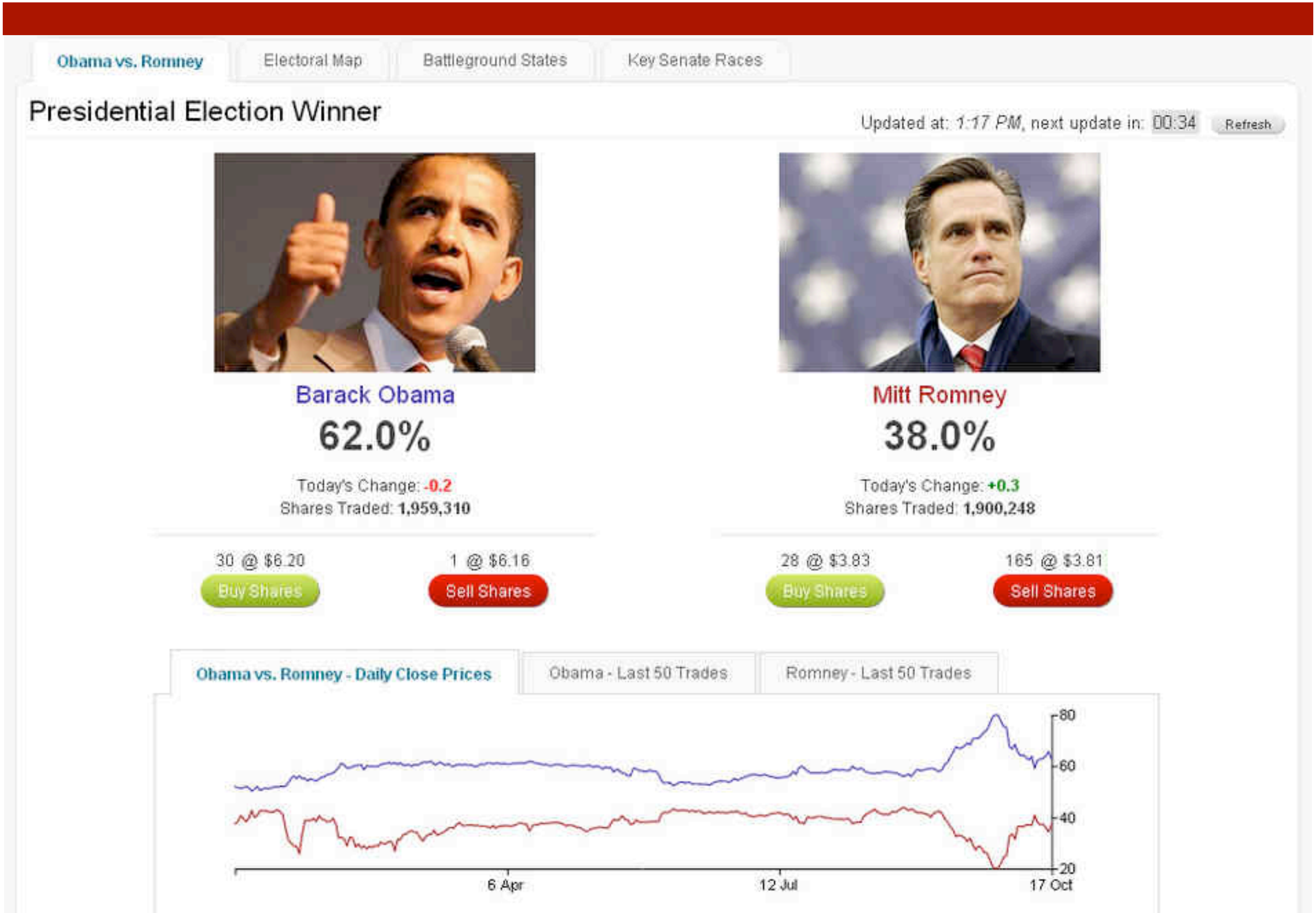


Image Credit: Intrade / NPR

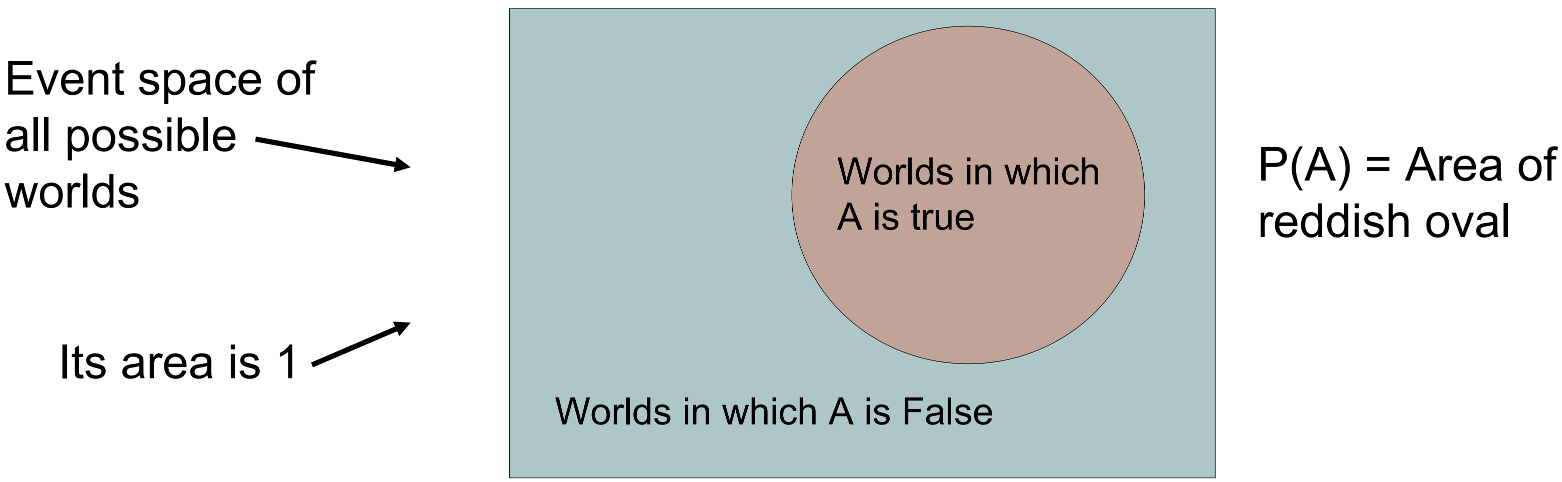

The
Axioms
Of
Probabi
lity

Axioms of Probability

- $0 <= P(A) <= 1$
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)

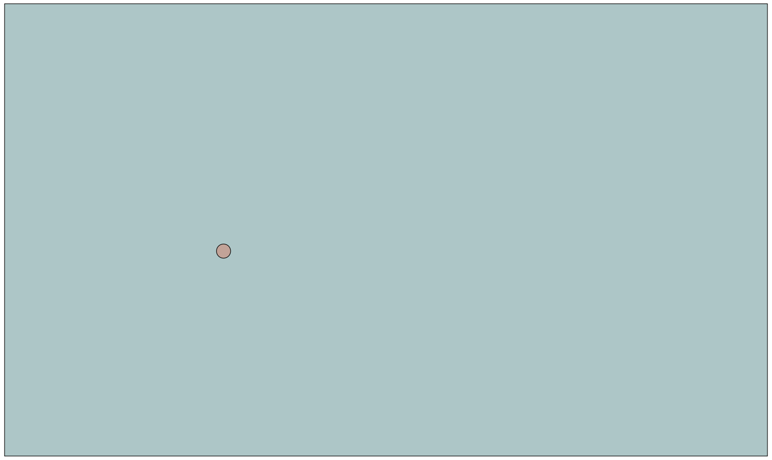
Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)



Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)



The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Image Credit: Andrew Moore

Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)

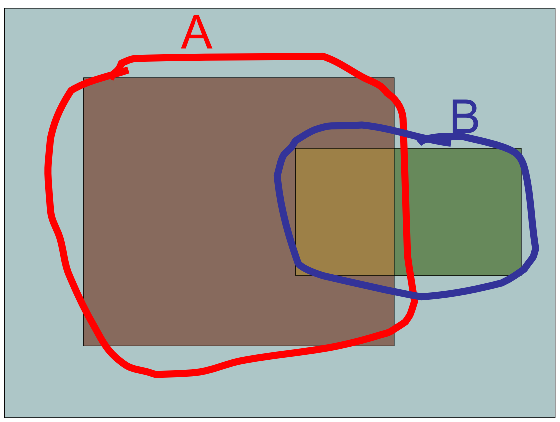
The area of A can't get any bigger than 1

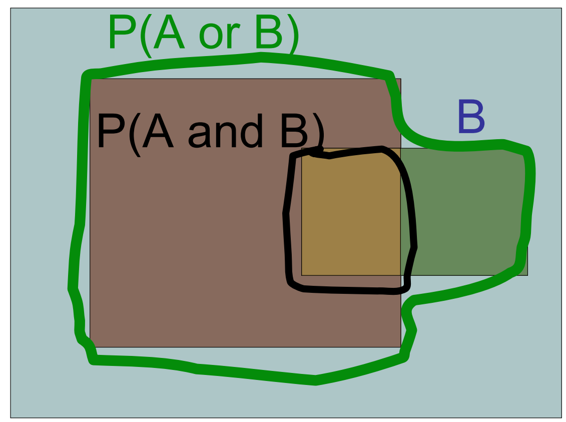
And an area of 1 would mean all worlds will have A true

Image Credit: Andrew Moore

Interpreting the Axioms

- 0<= P(A) <= 1
- P(empty-set) = 0
- P(everything) = 1
- P(A or B) = P(A) + P(B) – P(A and B)





Simple addition and subtraction

Concepts

- Sample Space
  - Space of events
- Random Variables
  - Mapping from events to numbers
  - Discrete vs Continuous
- Probability
  - Mass vs Density

Discrete Random Variables

X → discrete random variable

$\mathcal{X}$ or Val(X) → sample space of possible outcomes, which may be finite or countably infinite

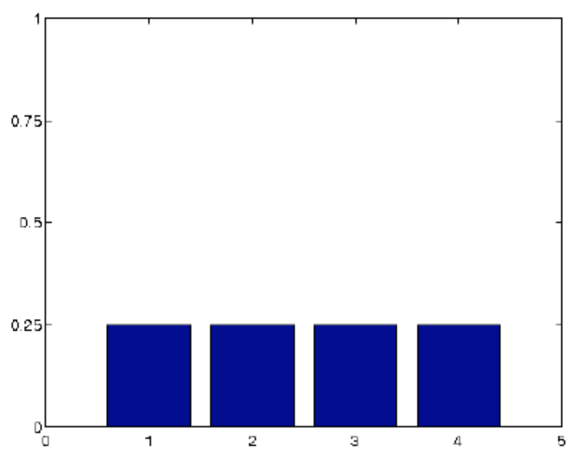
$x \in \mathcal{X}$ → outcome of sample of discrete random variable

$p(X = x)$ → probability distribution (probability mass function)

$p(x)$ → shorthand used when no ambiguity

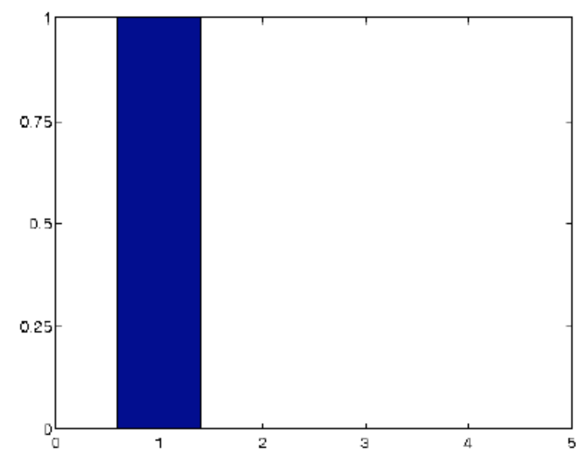
$$0 \leq p(x) \leq 1 \text{ for all } x \in \mathcal{X}$$

$$\sum_{x \in \mathcal{X}} p(x) = 1$$



uniform distribution

$$\mathcal{X} = \{1, 2, 3, 4\}$$



degenerate distribution

Continuous Random Variables

- On board

Concepts

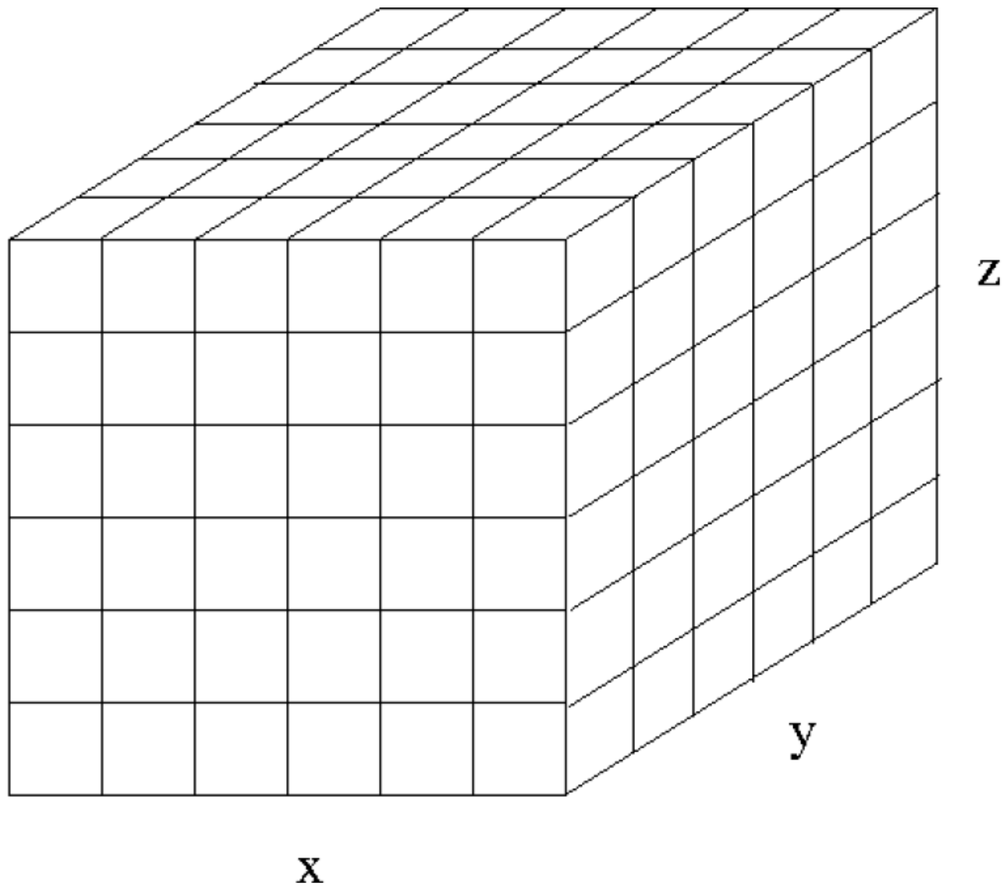
- Expectation

- Variance

Most Important Concepts

- Marginal distributions / Marginalization
- Conditional distribution / Chain Rule
- Bayes Rule

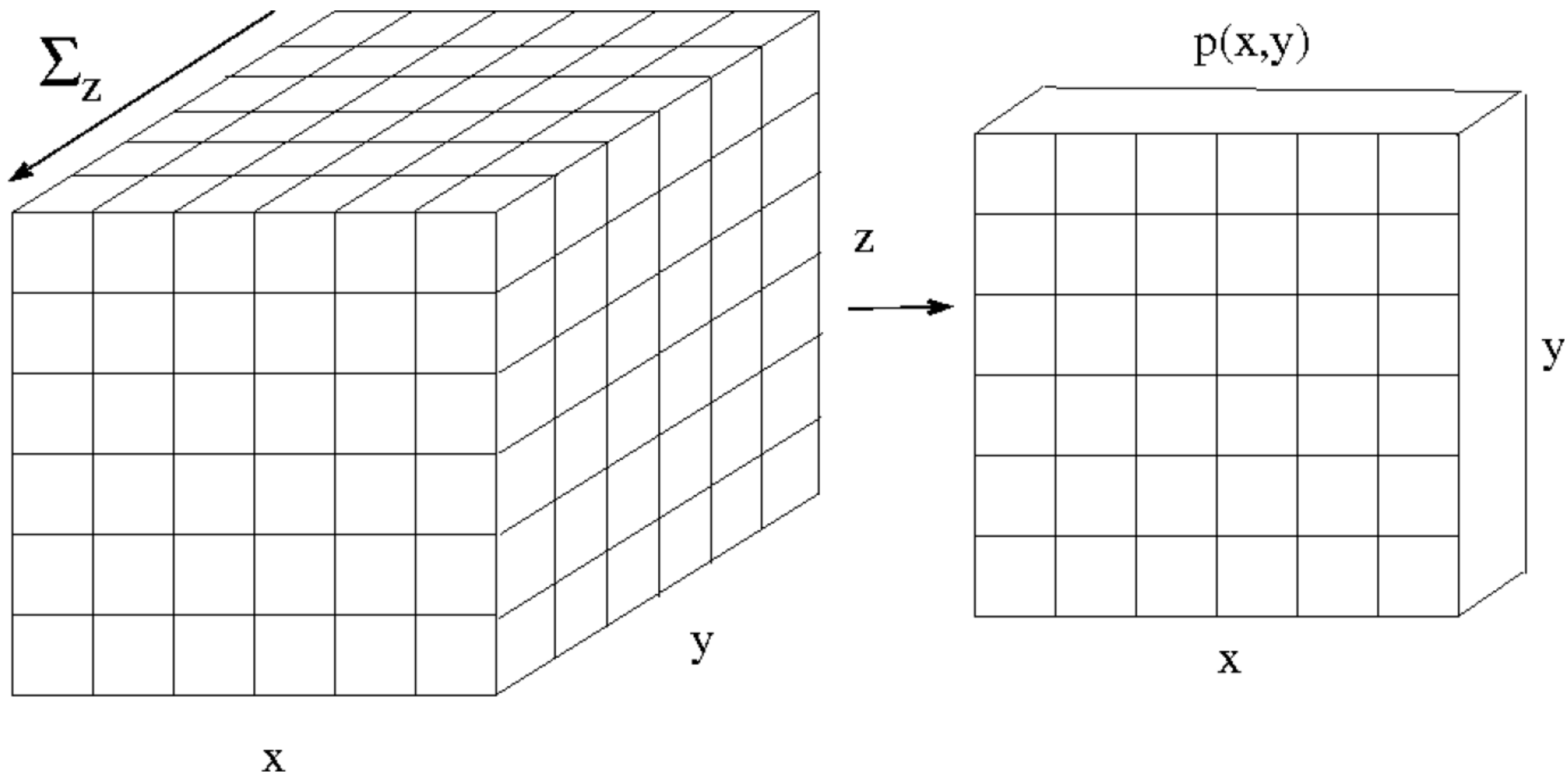
Joint Distribution



Marginalization

- Marginalization
  - Events: P(A) = P(A and B) + P(A and not B)
  - Random variables $P(X=x)=\sum_{y} P(X=x,Y=y)$

Marginal Distributions



$$p(x,y) = \sum_{z \in \mathcal{Z}} p(x,y,z) \qquad p(x) = \sum_{y \in \mathcal{Y}} p(x,y)$$

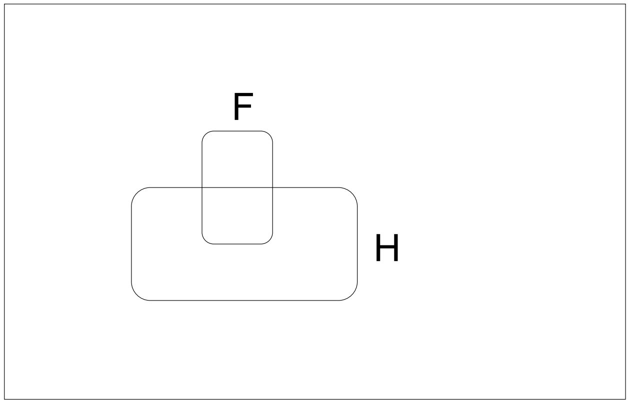
Conditional Probabilities

- $P(Y=y \mid X=x)$
- What do you believe about Y=y, if I tell you X=x?
- P(Rafael Nadal wins French Open 2015)?
- What if I tell you:
  - He has won the French Open 9/10 he has played there
  - Novak Djokovic is ranked 1; just won Australian Open
  - I offered a similar analysis last year and Nadal won

Conditional Probabilities

- P(A | B) = In worlds that where B is true,
  fraction where A is true

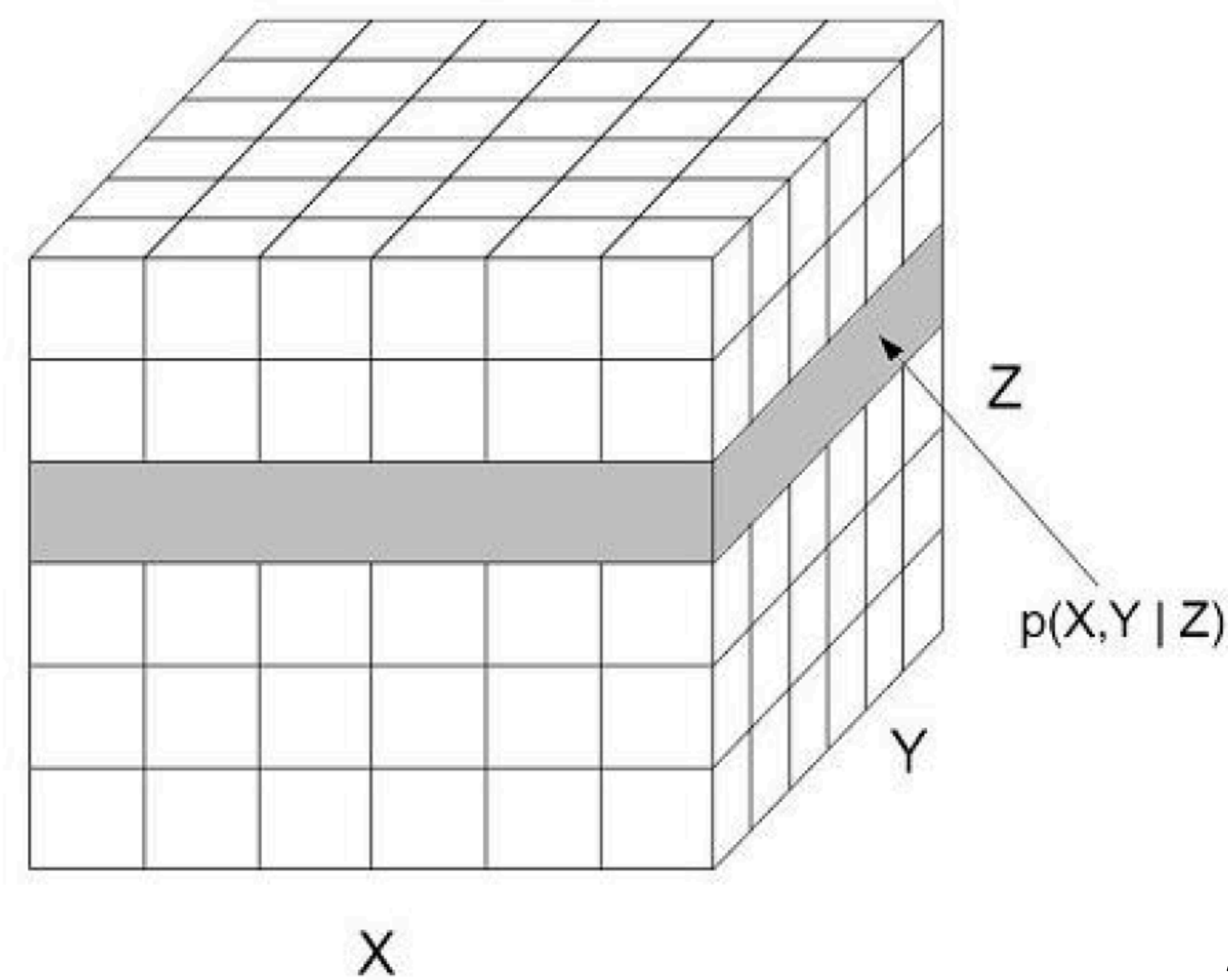
- Example
  - H: "Have a headache"
  - F: "Coming down with Flu"



P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

Conditional Distributions



$$p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)}$$

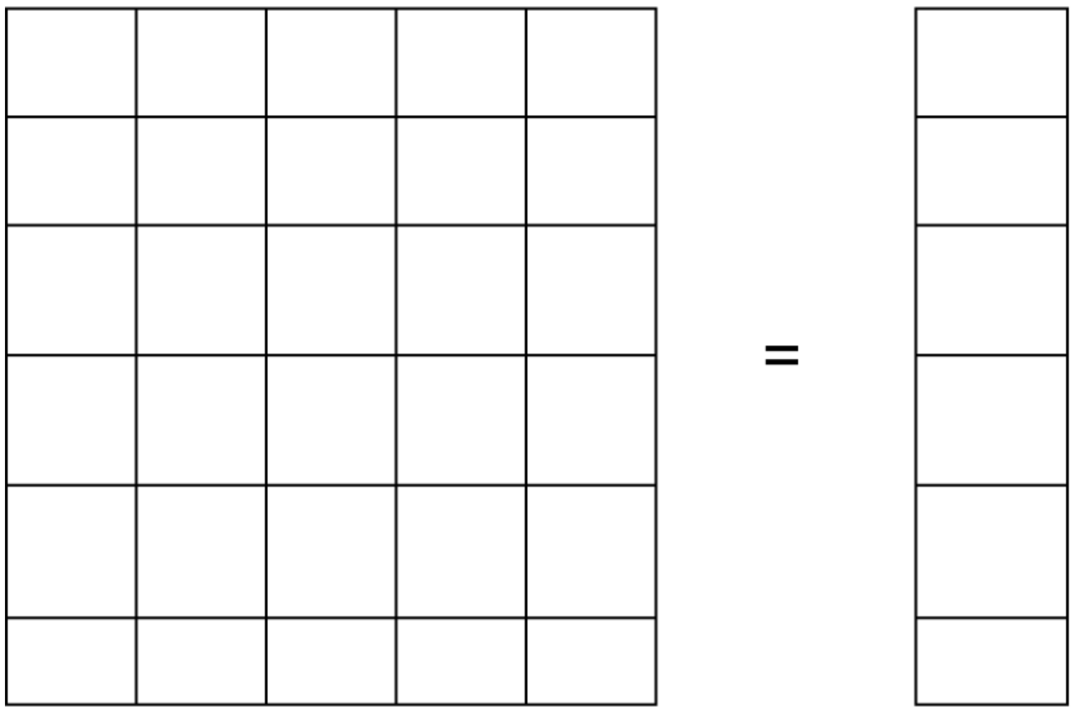
Conditional Probabilities

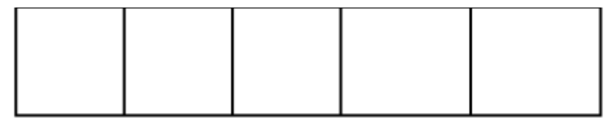
- Definition

- Corollary: Chain Rule

Independent Random Variables

P(x,y)





$$X \perp Y$$

$$\Updownarrow$$

$$p(x, y) = p(x)p(y)$$

$$\text{for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

Marginal Independence

- **Sets** of variables **X**, **Y**

- **X** is independent of **Y**
  - Shorthand: $P \vdash (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** *P* satisfies $(\mathbf{X} \perp \mathbf{Y})$ if and only if
  - $P(\mathbf{X=x},\mathbf{Y=y}) = P(\mathbf{X=x})\, P(\mathbf{Y=y})$, $\forall \mathbf{x} \in Val(\mathbf{X}),\ \mathbf{y} \in Val(\mathbf{Y})$

Conditional independence

- **Sets** of variables **X**, **Y**, **Z**

- **X** is independent of **Y** given **Z** if
  - Shorthand: $P \vdash (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - For $P \vdash (\mathbf{X} \perp \mathbf{Y} \mid \emptyset)$, write $P \vdash (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** *P* satisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\, P(\mathbf{Y}|\mathbf{Z})$, $\quad \forall \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y}), \mathbf{z} \in Val(\mathbf{Z})$

Concept

- Bayes Rules
  - Simple yet fundamental

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

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Image Credit: Andrew Moore

Bayes Rule

- Simple yet profound
  - Using Bayes Rules doesn’t make your analysis Bayesian!

- Concepts:
  - Likelihood
    - How much does a certain hypothesis explain the data?
  - Prior
    - What do you believe before seeing any data?
  - Posterior
    - What do we believe after seeing the data?

Entropy

- Measures the amount of ambiguity or uncertainty in a distribution:

$$H(p) = -\sum_x p(x) \log p(x)$$

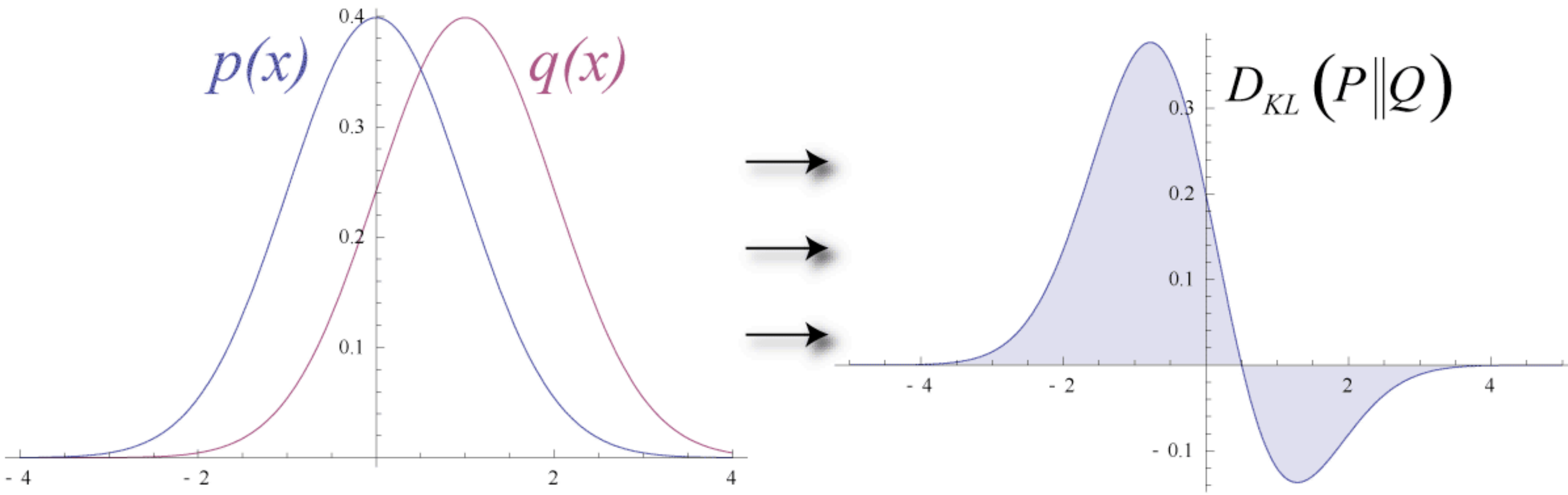
- Expected value of $-\log p(x)$ (a function which depends on p(x)!).
- $H(p) > 0$ unless only one possible outcomein which case $H(p) = 0$.
- Maximal value when p is uniform.
- Tells you the expected "cost" if each event costs $-\log p(\text{event})$

KL-Divergence / Relative Entropy

- An assymetric measure of the distancebetween two distributions:

$$KL[p||q] = \sum_x p(x)[\log p(x) - \log q(x)]$$

- $KL > 0$ unless $p = q$ then $KL = 0$
- Tells you the extra cost if events were generated by $p(x)$ but instead of charging under $p(x)$ you charged under $q(x)$.



Original Gaussian PDF's

KL Area to be Integrated

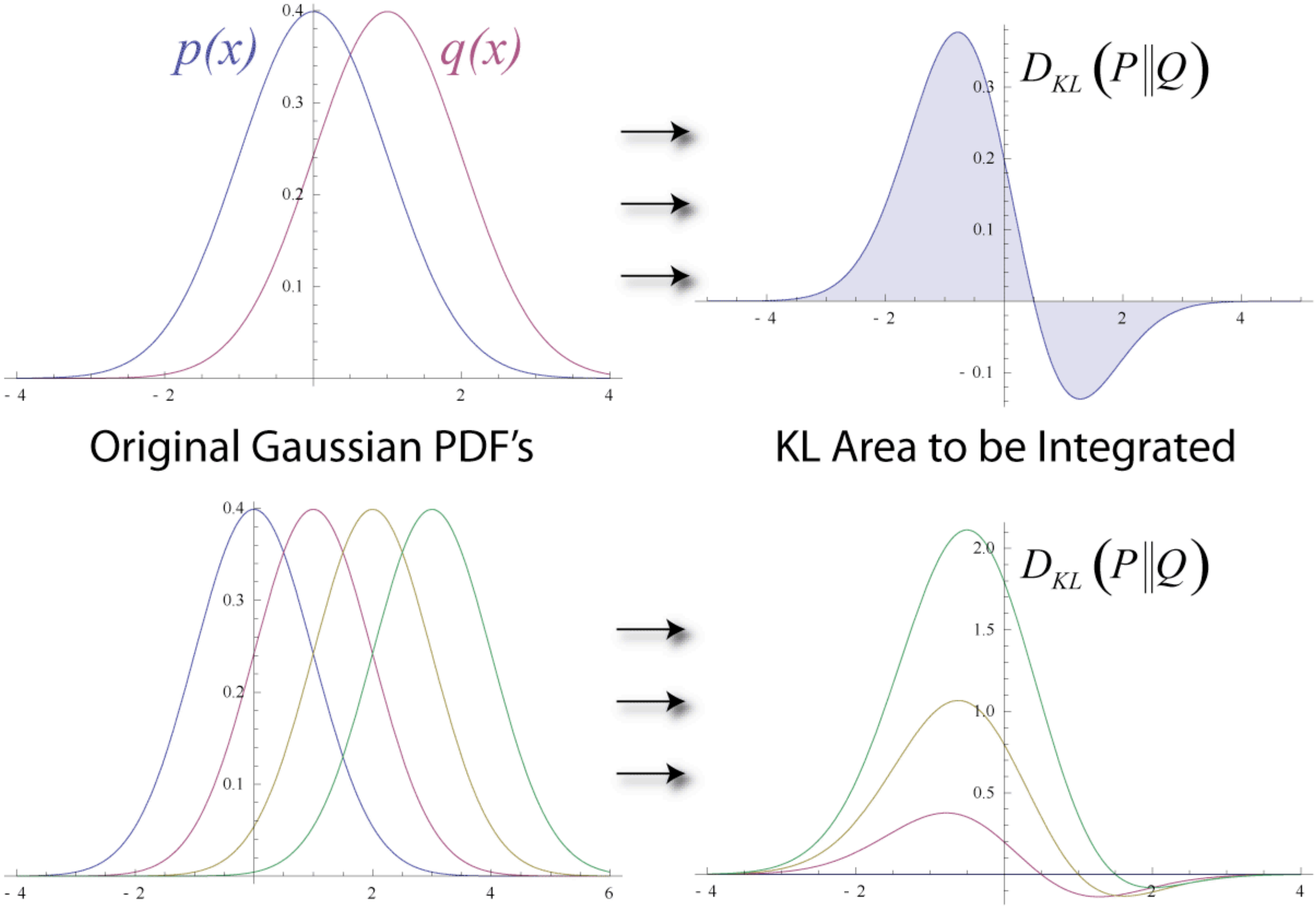


Image Credit: Wikipedia

- End of Prob. Review

- Start of Estimation