Teaching probability and statistics from a purely Bayesian point of view

Sanjoy Mahajan
Olin College of Engineering

sanjoy@olin.edu
streetfightingmath.com

MIT ESME, Cambridge, MA, 05 Mar 2019

Philosophy of mathematics from the start

$\Pr(10^{13}\text{th digit of } \pi \text{ is a } 7) =$

a. 0 or 1

b. It's a nonsense question.

c. 1/10

d. 1/5

Philosophy of mathematics from the start

$\Pr(10^{13}\text{th digit of } \pi \text{ is a } 7) =$

a. 0 or 1 (objective or frequentist probability)

b. It's a nonsense question. (objective or frequentist probability)

c. 1/10 (subjective or Bayesian probability)

d. 1/5 (crazy or have special knowledge about π!)

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Course is a first course in ProbStat for engineers

- 50 students/year (offered once/year)
- All students are engineering majors
- Satisfies probability/statistics requirement (one choice among 5 or 6 courses)
- 2×100-minute lectures / week
- No recitations, no TA.
- Homework: twice / week (problems plus reading)

Who am I?

PhD in physics (but including a physics model for the density of primes), so I believe that

$2 = 1 = -1 = e$.

I wrote a mathematics textbook: *Street-Fighting Mathematics: The Art of Educated Guessing and Opportunistic Problem Solving* (MIT Press, 2010).

One equation rules them all

$$\underbrace{\Pr(H \mid D)}_{\text{posterior prob.}} \propto \underbrace{\Pr(D \mid H)}_{\text{likelihood}} \times \underbrace{\Pr(H)}_{\text{prior prob.}}$$

Posterior: New belief in theory H

Likelihood: Explanatory power of theory H

Prior: Old belief in theory H (before considering data or evidence D)

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Bayesian probability generalizes logic beyond true/false

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B \,|\, A).$$

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B).$$

Bayesian probability generalizes logic beyond true/false

$\Pr(A \text{ and } B) = \Pr(A)\,\Pr(B \,|\, A).$ generalizes logical AND

$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B).$ generalizes logical OR

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Standard Monty Hall is the worked example

How it goes (WLOG):

1. You pick door 1.
2. Monty shows you door 2, and it is empty. (Monty will show you an empty door that is not the one you picked.)
3. You choose whether to stay with door 1 or switch to door 3.

An organized table reinforces the big idea

H	:	$\Pr(H)$	$\times$	$Pr(D \mid H)$	$=$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
1								
2								
3								

An organized table reinforces the big idea

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

An organized table reinforces the big idea

H	:	$\Pr(H)$	$\times$	$Pr(D \mid H)$	$=$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
1		1/3						
2		1/3						
3		1/3						

An organized table reinforces the big idea

H	:	$\Pr(H)$	$\times$	$Pr(D \mid H)$	$=$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
1		1/3						
2		1/3		0				
3		1/3						

An organized table reinforces the big idea

H	:	$\Pr(H)$	$\times$	$Pr(D \mid H)$	$=$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
1		1/3						
2		1/3		0				
3		1/3		1				

An organized table reinforces the big idea

H	:	$\Pr(H)$	$\times$	$Pr(D \mid H)$	$=$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
1		1/3		1/2				
2		1/3		0				
3		1/3		1				

An organized table reinforces the big idea

H	:	$\Pr(H)$	$\times$	$Pr(D \mid H)$	$=$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
1		1/3		1/2		1/6		
2		1/3		0		0		
3		1/3		1		1/3		

An organized table reinforces the big idea

H :	$\Pr(H)$ ×	$Pr(D\,\vert\,H)$ =	$\Pr(H)\Pr(D\,\vert\,H)$ ∝	$\Pr(H\,\vert\,D)$
1	1/3	1/2	1/6	
2	1/3	0	0	
3	1/3	1	1/3	
			$\sum = 1/2$	

An organized table reinforces the big idea

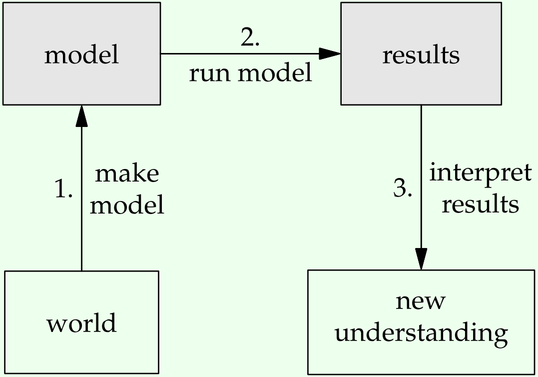
H :	$\Pr(H)$ $\times$	$Pr(D \mid H)$ $=$	$\Pr(H)\Pr(D \mid H)$ $\propto$	$\Pr(H \mid D)$
1	1/3	1/2	1/6	1/3
2	1/3	0	0	0
3	1/3	1	1/3	2/3
			$\sum = 1/2$	

Modeling can be practiced in the small for spaced repetition

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Modeling can be practiced in the small for spaced repetition



Deciding whether to switch doors practices step 3.

Drunk Monty Hall is the faded example

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$ ×	$Pr(D \mid H)$ =	$\Pr(H)\Pr(D \mid H)$ ∝	$\Pr(H \mid D)$
1	1/3	1/2	1/6	1/3
2	1/3	0	0	0
3	1/3	1	1/3	2/3
			$\sum = 1/2$	

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$ ×	$Pr(D\mid H)$ =	$\Pr(H)\Pr(D\mid H)$ ∝	$\Pr(H\mid D)$
1	1/3	1/2	1/6	1/3
2	1/3	0	0	0
3	1/3	1	1/3	2/3
			$\sum = 1/2$	

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$ ×	$Pr(D\|H)$ =	$\Pr(H)\Pr(D\|H)$ ∝	$\Pr(H\|D)$
1	1/3	1/3	1/6	1/3
2	1/3	0	0	0
3	1/3	1	1/3	2/3
			$\sum = 1/2$	

Q: How did $\Pr(D|H)$ *change if D and H did not?*

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$ ×	$Pr(D \mid H)$ =	$\Pr(H)\Pr(D \mid H)$ ∝	$\Pr(H \mid D)$
1	1/3	1/3	1/6	1/3
2	1/3	0	0	0
3	1/3	1	1/3	2/3
			$\sum = 1/2$	

A: Really $\Pr(D \mid H, \mathit{Background})$. *All probability is conditional!*

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$	$\times$ $Pr(D \mid H)$	$=$ $\Pr(H)\Pr(D \mid H)$	$\propto$ $\Pr(H \mid D)$
1	1/3	1/3	1/6	1/3
2	1/3	0	0	0
3	1/3	1	1/3	2/3
			$\sum = 1/2$	

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$	$\times$ $Pr(D \mid H)$	$=$ $\Pr(H)\Pr(D \mid H)$	$\propto$ $\Pr(H \mid D)$
1	1/3	1/3	1/9	1/3
2	1/3	0	0	0
3	1/3	1	1/9	2/3
			$\sum = 1/2$	

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$	$\times$ $Pr(D \mid H)$	$=$ $\Pr(H)\Pr(D \mid H)$	$\propto$ $\Pr(H \mid D)$
1	1/3	1/3	1/9	1/3
2	1/3	0	0	0
3	1/3	1	1/9	2/3
			$\sum = 2/9$	

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$	$\times$ $Pr(D\|H)$	$=$ $\Pr(H)\Pr(D\|H)$	$\propto$ $\Pr(H\|D)$
1	1/3	1/3	1/9	1/2
2	1/3	0	0	0
3	1/3	1	1/9	1/2
			$\sum = 2/9$	

Drunk Monty Hall is the faded example

How it works: Thanks to drink, Monty has forgotten where the prize is. He staggers on stage and into door 2, which springs open and is empty.

Q: How do you modify the regular-Monty Bayesian table?

H :	$\Pr(H)$	$\times$ $Pr(D \mid H)$	$=$ $\Pr(H)\Pr(D \mid H)$	$\propto$ $\Pr(H \mid D)$
1	1/3	1/3	1/9	1/2
2	1/3	0	0	0
3	1/3	1/3	1/9	1/2
			$\sum = 2/9$	

Interpret (modeling!): Switching doesn't help or hurt.

Maybe-drunk Monty is the traditional problem

How it works: Monty could be drunk or sober. Based on his behavior, you judge each possibility at 50–50. He wobbles on stage and opens door 2, which is empty.

Q: How do you modify the drunk-Monty Bayesian table?

H :	$\Pr(H)$ ×	$Pr(D\|H)$ =	$\Pr(H)\Pr(D\|H)$ ∝	$\Pr(H\|D)$
1	1/3	1/3?	1/9	1/2
2	1/3	0	0	0
3	1/3	1/3?	1/9	1/2
			$\sum = 2/9$	

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Online tutor fosters focused processing

- Called CAT-SOOP (cat-soop.org), a cousin of the EdX tutor and written by Adam Hartz at MIT for MIT's 6.01 course
- Students enter answers online and get instant feedback on correctness.
- Solutions available right when ideas about problem are fresh (in working-memory cache and ready for correction and elaboration).

Online tutor fosters focused processing

PROBLEM 1. MONTY MIGHT BE DRUNK

You are playing on the Monty Hall show and, for variety, select door 3 (instead of our usual door 1). Before you open it, Monty, looking somewhat worse for the wear, opens door 2 and shows you that it is empty. He might be sober (but tired) or drunk. Based on his appearance, but before you know which door he has opened, you think that $P(\text{Monty is drunk}) = 0.5$and $P(\text{Monty is sober}) = 0.5$

In this problem, you'll work out the probability of your winning the prize if you switch to door 1. First, here are two preliminary questions:

(a) What is the symbolic expression for the probability of your winning the prize if you switch to door 1?

- $P(D)$
- $P(H_1)$
- $P(H_1|D)$
- $P(H_1, D)$
- $P(D|H_1)$

Save | Submit | View Answer

As staff, you are always allowed to submit. If you were a student, you would see the following:

You have 1 submission remaining.

Online tutor fosters focused processing

(a) What is the symbolic expression for the probability of your winning the prize if you switch to door 1?

- ○ $P(D)$
- ◉ $P(H_1)$
- ○ $P(H_1|D)$
- ○ $P(H_1, D)$
- ○ $P(D|H_1)$

Save | Submit | View Explanation | Clear Answer | 0.00%

As staff, you are always allowed to submit. If you were a student, you would see the following:

You have 0 submissions remaining.

Solution: $P(H_1|D)$

✖

Online tutor fosters focused processing

- ○ $P(D)$
- ◉ $P(H_1)$
- ○ $P(H_1|D)$
- ○ $P(H_1, D)$
- ○ $P(D|H_1)$

Save | Submit | Clear Answer | 0.00%

As staff, you are always allowed to submit. If you were a student, you would see the following:

You have 0 submissions remaining.

Solution: $P(H_1|D)$

Explanation:

If you switch to door 1, you win if and only if the prize is behind door 1 (i.e. if H_1 is true). So, you want to know the probability of H_1. But you don't want the prior probability $P(H_1)$, because it ignores the data D that Monty, who might be drunk, opened door 2 and showed you that it was empty. To account for the data, you use the posterior probability $P(H_1|D)$.

✖

Tutor allows chained parts still to be educational

Students still have to think about the mathematical goal ($\Pr(H_1 \mid D)$).

They cannot game the system by extracting it from the statement of a subsequent part.

But they still have scaffolding. Thus, they can solve, and learn from, the subsequent part even if they didn't figure out that $\Pr(H_1 \mid D)$ is the goal.

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Bayes' theorem and the law

Federal Rules of Evidence

Rule 401. Evidence is relevant if:

a. it has any tendency to make a fact more or less probable than it would be without the evidence; and

b. the fact is of consequence in determining the action.

Q: What is the probabilistic translation of FRE 401(a)?

Whole courtroom arguments can also be translated into mathematics

Background information for a murder trial:

1. The victim wounded the killer.
2. The killer's blood was found at the scene. It is a rare type found in 1 out of 100 people in the general population.
3. The detective questions neighbors of the victim and notices one wearing a bandage. Based on his overall impression of this man, the detective estimates the probability of the man's guilt to be 0.1.
4. *Blood-test evidence.* The detective then learns that the suspect has the same rare blood type as the killer.

Whole courtroom arguments can also be translated into mathematics

[Source: W. C. Thompson and E. L. Shumann, (1987). "Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy." *Law and Human Behavior* 2(3): 167.]

Whole courtroom arguments can also be translated into mathematics

Argument 1

The blood test evidence is highly relevant. The suspect has the same blood type as the attacker. This blood type is found in only 1% of the population, so there is only a 1% chance that the blood found at the scene came from someone other than the suspect. Since there is only a 1% chance that someone else committed the crime, there is a 99% chance the suspect is guilty.

Q: Is this reasoning correct or incorrect? Should the detective revise his probability estimate in light of the blood-test evidence? If the detective should revise it, what should the new probability be?

Whole courtroom arguments can also be translated into mathematics

Argument 2

The evidence about blood types has very little relevance for this case. Only 1% of the population has the “rare” blood type, but in [a city like the one where the crime occurred], with a population of 200,000, this blood type would be found in approximately 2000 people. Therefore the evidence merely shows that the suspect is one of 2000 people in the city who might have committed the crime. A one-in-2000 chance of guilt (based on the blood test evidence) has little relevance for proving this suspect is guilty.

Q: Is this reasoning correct or incorrect? Should the detective revise his probability estimate in light of the blood-test evidence? If the detective should revise it, what should the new probability be?

Other real-world applications were included

1. If you newborn has a positive PKU screening test, how worried should you be?
2. Analysis of raw(ish) polling data
3. Does Sanjoy have Marfan's syndrome?
4. When my daughters play the card game War and one is ahead 14–4, is it due to cheating or is it just luck?
5. Does sickle-cell anemia protect against malaria?
6. Crisis of reproducibility
7. and many others.

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Careful continuum limit is a form of scaffolding and makes calculus come alive

Q: What kind of coin is being flipped (what is its probability of turning up heads)?

$D =$ came up heads

$x \equiv \Pr(\text{coin turns up heads})$

Careful continuum limit is a form of scaffolding and makes calculus come alive

H	$\Pr(H)$	$Pr(D \mid H)$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
$0.0 \le x < 0.1$	0.1	0.05	0.005		0.01
$0.1 \le x < 0.2$	0.1	0.15	0.015		0.03
$0.2 \le x < 0.3$	0.1	0.25	0.025		0.05
$\vdots$					
$0.9 \le x \le 1.0$	0.1	0.95	0.095		0.19
			$\sum = 1/2$		

Careful continuum limit is a form of scaffolding and makes calculus come alive

H	$\Pr(H)$	$Pr(D \mid H)$	$\Pr(H)\Pr(D \mid H)$	$\propto$	$\Pr(H \mid D)$
$0.0 \leq x < 0.1$	0.1	0.05	0.005		0.01
$0.1 \leq x < 0.2$	0.1	0.15	0.015		0.03
$0.2 \leq x < 0.3$	0.1	0.25	0.025		0.05
x	Δx	x	$x\Delta x$		$2x\Delta x$
$0.9 \leq x \leq 1.0$	0.1	0.95	0.095		0.19
			$\int = 1/2$		

Probability density is the probability without (per) the Δx.

$p(x) = 1$

$p(x \mid D) = 2x$

Frequentist ideas can be included

Q: Toss a coin 7 times and get `HHHHHHT`*. Is the coin fair* (H_0)*?*

$\Pr(H_0 \mid D)$ Illegal in frequentism

Frequentist ideas can be included

Q: Toss a coin 7 times and get `HHHHHHT`*. Is the coin fair* (H_0)*?*

$\Pr(H_0 \mid D)$ Illegal in frequentism

$\Pr(D \mid H_0)$ Is 2^{-7} for any 7-toss sequence

Frequentist ideas can be included

Q: Toss a coin 7 times and get `HHHHHHT`*. Is the coin fair* (H_0)*?*

$\Pr(H_0 \,|\, D)$ Illegal in frequentism

$\Pr(D \,|\, H_0)$ Is 2^{-7} for any 7-toss sequence

$\Pr(\underbrace{D}_{\text{without order}} \,|\, H_0)$ Limits to zero even when H_0 is true

Frequentist ideas can be included

Q: Toss a coin 7 times and get `HHHHHHT`. *Is the coin fair* (H_0)*?*

$\Pr(H_0 \mid D)$	Illegal in frequentism
$\Pr(D \mid H_0)$	Is 2^{-7} for any 7-toss sequence
$\Pr(\underbrace{D}_{\text{without order}} \mid H_0)$	Limits to zero even when H_0 is true
$\Pr(D \text{ or } \underbrace{D^\star}_{\text{more extreme data}} \mid H_0)$	The p-value

It has many problems, as Bayesian framework makes clear

p-value $\equiv \Pr(D \text{ or } D^\star \,|\, H_0)$

It has many problems:

1. doesn't take into account other hypothesis.
2. requires knowing a stopping rule.
3. isn't what you expect.
4. doesn't use prior → crisis of reproducibility

Bayes is your north star.

$$\underbrace{\Pr(H\,|\,D)}_{\text{posterior prob.}} \propto \underbrace{\Pr(D\,|\,H)}_{\text{likelihood}} \times \underbrace{\Pr(H)}_{\text{prior prob.}}$$

The transferable lessons are several

1. Philosophy of mathematics from the start
2. One huge idea throughout
3. Scaffolding: worked examples to faded examples to traditional problems
4. Modeling
5. Online tutor
6. Real-world data and arguments
7. Giving meaning to the great idea of calculus

Students found the course valuable

“This class did more to change my perspective on math and statistics in the world than practically any other class I have taken[.]”

“Even though I knew nothing about the subject going in, I was able to learn a ton and became genuinely interested in the topic and feel like it is something that I will have with me for the rest of my life.”

“[M]ost of the material was focused on understanding the significance of statistics in your view of the world.”

“It is no exaggeration to say that taking your class really did change how I look at the world.”

Goal of teaching

The goal [of teaching] should be, not to implant in the students' mind every fact that the teacher knows now;

but rather to implant a *way of thinking* that enables the student, in the future, to learn in one year what the teacher learned in two years.

Only in that way can we continue to advance from one generation to the next.

—*Edwin T. Jaynes* (*1922–1998*)

My favorite references

Hartz, Adam, et al. *CAT-SOOP Online Tutor*.
`catsoop.mit.edu/website`
Online homework-entry system used for the course.

Jaynes, Edwin T. (2003). *Probability theory: The Logic of Science*. Cambridge University Press. *A classic on Bayesian inference by a pioneer of the so-called objective Bayesian viewpoint (at the graduate level in philosophy and mathematics).*

Lindley, Dennis V. (2013). *Understanding Uncertainty*. John Wiley & Sons. *An introductory textbook, by a pioneer of the so-called subjective Bayesian viewpoint.*

Renkl, Alexander (2014). Learning From Worked Examples: How to Prepare Students for Meaningful Problem Solving.
`teachpsych.org/ebooks/asle2014/index.php`
My favorite introduction to using worked and faded examples.

Teaching probability and statistics from a purely Bayesian point of view

Sanjoy Mahajan
Olin College of Engineering

sanjoy@olin.edu
streetfightingmath.com

Produced with free software:
Emacs, LuaTeX, and ConTeXt

MIT ESME, Cambridge, MA, 05 Mar 2019