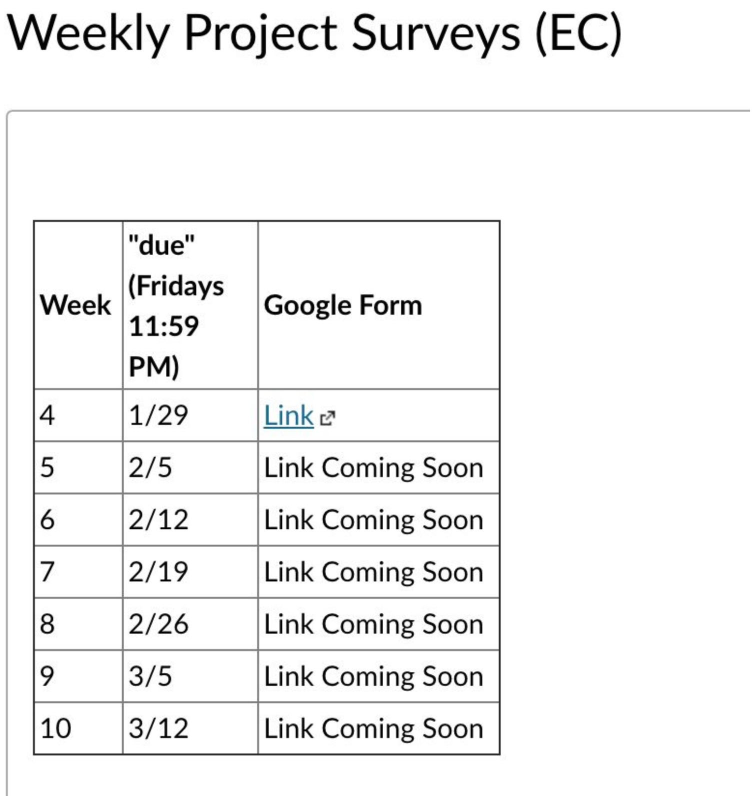
Course Reminders

- Due Friday
  - D3
  - Q3
  - Project Proposal
  - Weekly Project Survey
    - *optional*; for EC (1%)
    - 7 weeks (weeks 4-10)
    - Links posted on Canvas



Weekly Project Surveys (EC)

Week	"due" (Fridays 11:59 PM)	Google Form
4	1/29	Link
5	2/5	Link Coming Soon
6	2/12	Link Coming Soon
7	2/19	Link Coming Soon
8	2/26	Link Coming Soon
9	3/5	Link Coming Soon
10	3/12	Link Coming Soon

Data Science Questions & Intuition

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

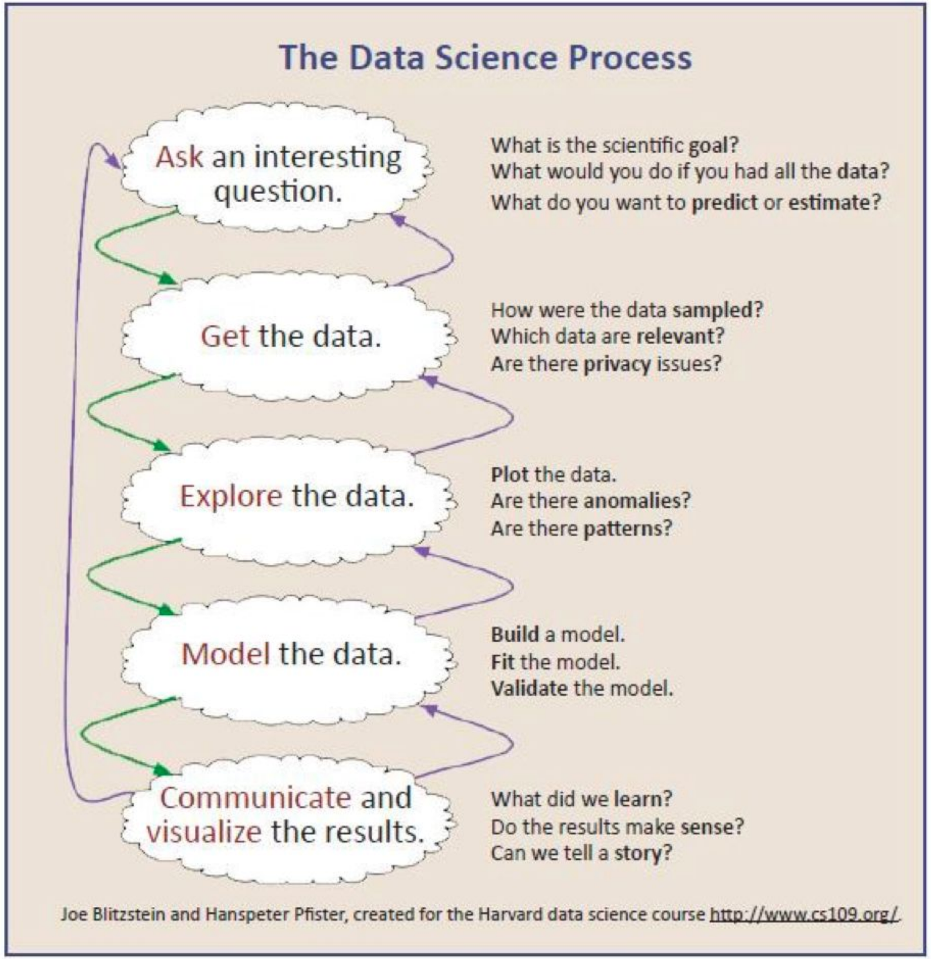
Formulating Data Science Questions

When you and your group sit down to figure out what you're going to do for your final project in this class, you'll have to formulate a strong question - one that is specific, can be answered with data, and makes clear what exactly is being measured.

Nature of a data scientist

- data-driven.
- care about answers. They analyze data to discover something about how the world works.
- care about whether the results make sense, because they care about what the answers mean.
- are comfortable with the idea that data have errors.
- know nothing is ever completely true or false in science, while everything is either true or false in computer science or mathematics.

adapted from Chris Keown



The Data Science Process

Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://www.cs109.org/.

adapted from Chris Keown

If I had an hour to solve a problem and my life depended on it, I would use the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes. —Einstein

adapted from Chris Keown

Data Science questions should...

- Be specific
- Be answerable with data
- Specify what's being measured

What makes a question a good question?

Nailing down the right question: politics

Too-vague question: What impacts politics in America?

Improving: Does pop culture have an impact on American politics?

… Do American TV shows have an impact on American politics?

… Does South Park affect American politics?

… Is there a relationship between words in South Park episodes and American politics?

… Is there a relationship between the sentiment of political words in South Park and American politics?

… Is there a relationship between the sentiment of political words in South Park and America’s presidential approval rating?

https://github.com/COGS108/FinalProjects-Wi18/blob/master/071-FinalProject.ipynb

Nailing down the right question: flight delays

Too-vague question: Why are the flights I take never on time?

Improving: What causes delays during travel?

...Do certain airports have more flight delays than others?

...Does the likelihood of a flight delay depend upon where the flight starts or ends?

...Does the likelihood of a flight delay, cancellation, or diversion depend upon where the flight starts or ends?

https://github.com/COGS108/FinalProjects-Wi20/blob/master/FinalProject_group036.ipynb

Nailing down the right question: cause of death

Too-vague question: What gets attention in the news?

Improving: Do terrorist attacks get reported too much?

... Is there a relationship between the number of people who die relative to the amount of media attention a story gets?

... What causes of death are over reported in the news relative to CDC death data? Underreported?

... Is there a relationship over time between cause of death terms in the *NYT*, The Guardian, and Google trends data relative to data from the CDC?

https://github.com/COGS108/FinalProjects-Wi18/blob/master/010-FinalProject.ipynb

Nailing down the right question: policing

Too-vague question: Why isn’t police response time always the same?

Improving: How can we improve police response time?

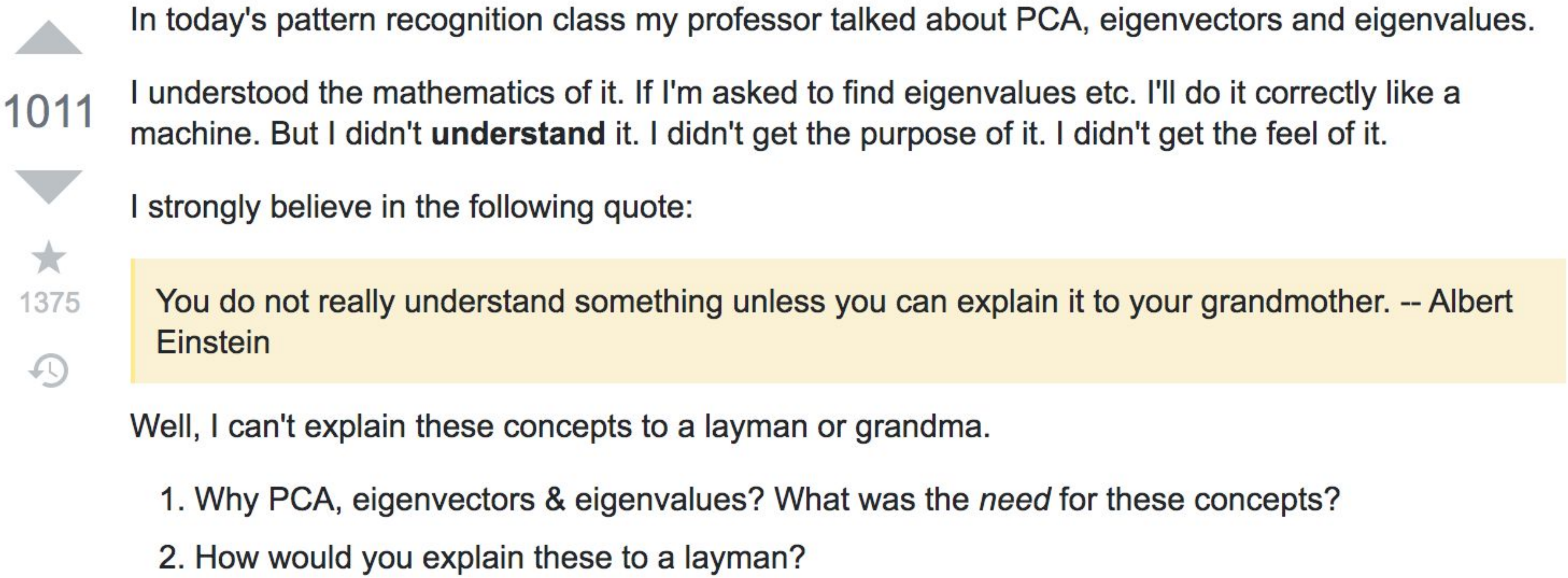
... Do crime levels and time of day affect response time?

... Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable?

... Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable throughout San Diego?

https://github.com/COGS108/FinalProjects-Sp17/blob/master/029-FinalProject.ipynb

Data Intuition



In today's pattern recognition class my professor talked about PCA, eigenvectors and eigenvalues.

I understood the mathematics of it. If I'm asked to find eigenvalues etc. I'll do it correctly like a machine. But I didn't **understand** it. I didn't get the purpose of it. I didn't get the feel of it.

I strongly believe in the following quote:

> You do not really understand something unless you can explain it to your grandmother. -- Albert Einstein

Well, I can't explain these concepts to a layman or grandma.

1. Why PCA, eigenvectors & eigenvalues? What was the *need* for these concepts?
2. How would you explain these to a layman?

https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Theory vs. Practice: “Tai’s model”

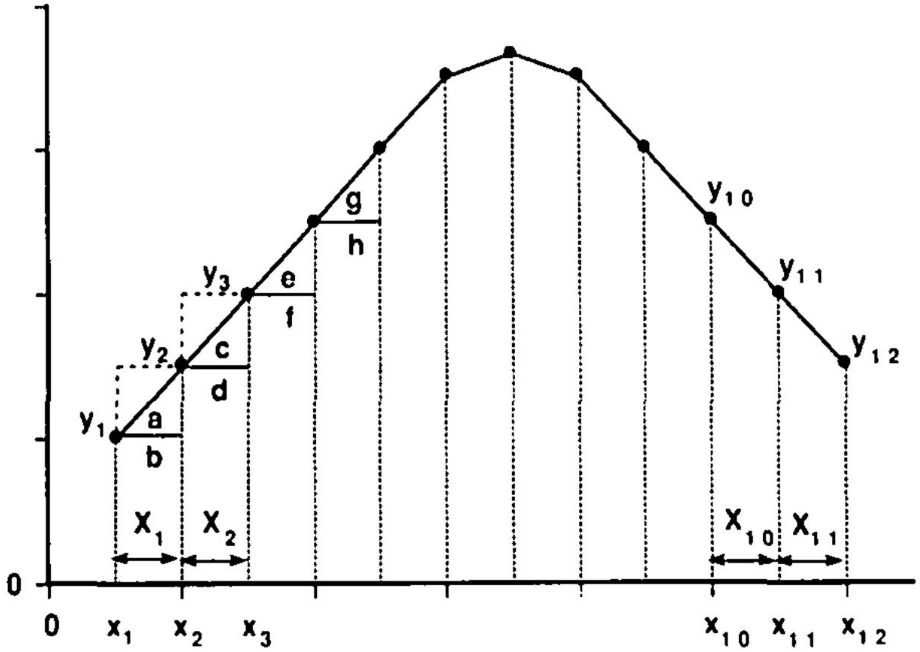
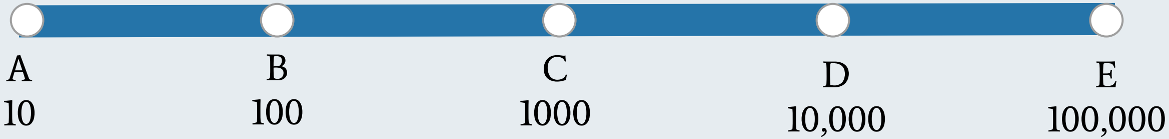


Figure 1—*Total area under the curve is the sum of individual areas of triangles a, c, e, and g and rectangles b, d, f, and h.*

Tai, *Diabetes Care* (1994)

Fermi Estimation

Approximately how many piano tuners do you think there are in the city of Chicago?



A 10

B 100

C 1000

D 10,000

E 100,000

343
299,792,458 m/s

https://www.youtube.com/watch?v=0YzvupOX8Is

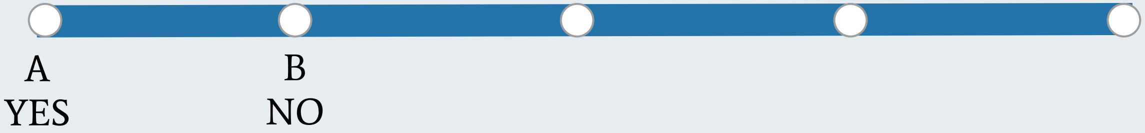
Has humanity produced enough paint to cover the entire land area of the Earth?

—Josh (Bolton, MA)

https://what-if.xkcd.com/84/

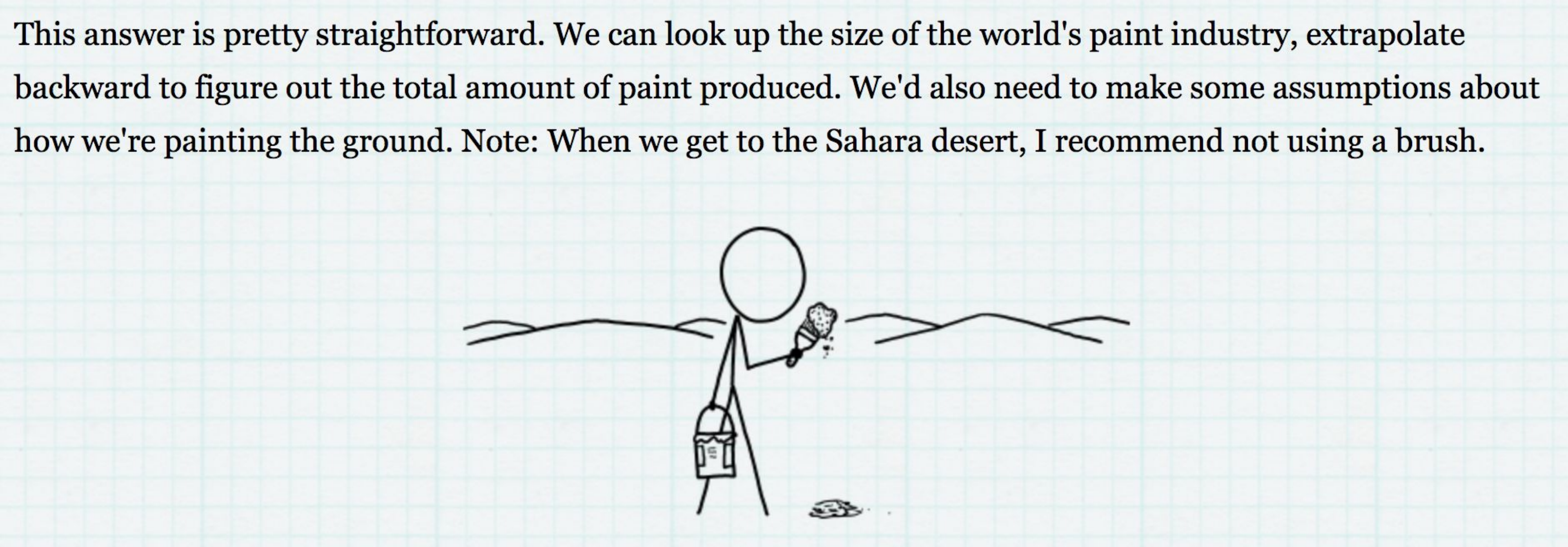
Fermi Estimation

Has humanity produced enough paint to cover the entire land area of the Earth?



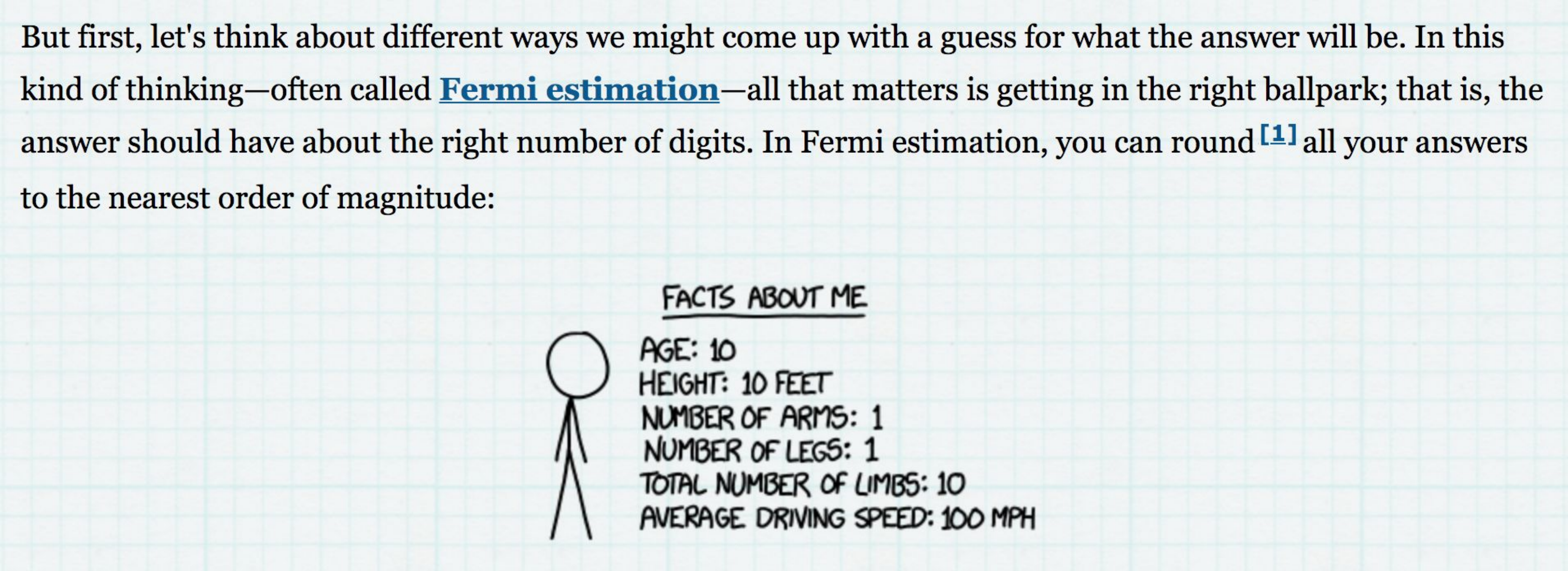
A
YES

B
NO



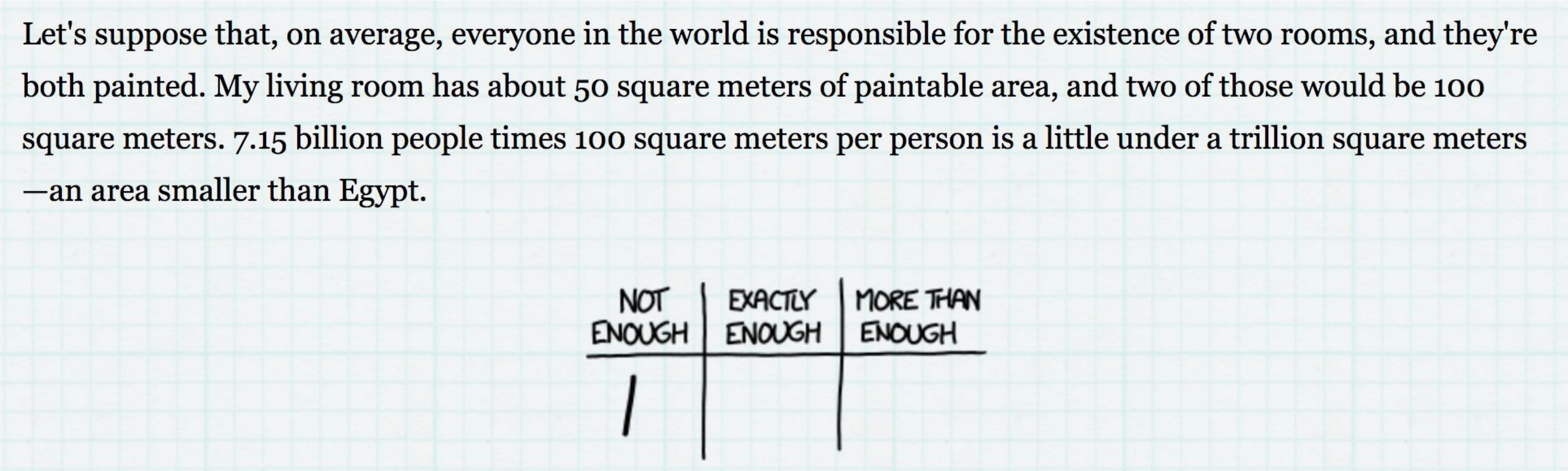
This answer is pretty straightforward. We can look up the size of the world's paint industry, extrapolate backward to figure out the total amount of paint produced. We'd also need to make some assumptions about how we're painting the ground. Note: When we get to the Sahara desert, I recommend not using a brush.

https://what-if.xkcd.com/84/



But first, let's think about different ways we might come up with a guess for what the answer will be. In this kind of thinking—often called **Fermi estimation**—all that matters is getting in the right ballpark; that is, the answer should have about the right number of digits. In Fermi estimation, you can round[1] all your answers to the nearest order of magnitude:

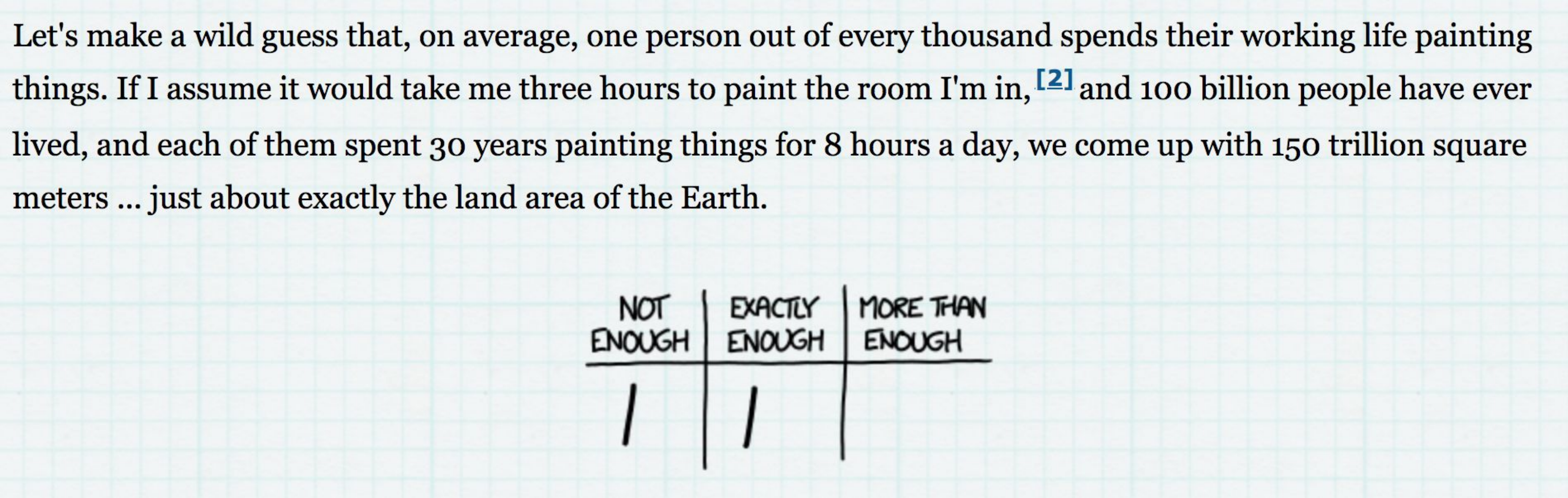
https://what-if.xkcd.com/84/



Let's suppose that, on average, everyone in the world is responsible for the existence of two rooms, and they're both painted. My living room has about 50 square meters of paintable area, and two of those would be 100 square meters. 7.15 billion people times 100 square meters per person is a little under a trillion square meters —an area smaller than Egypt.

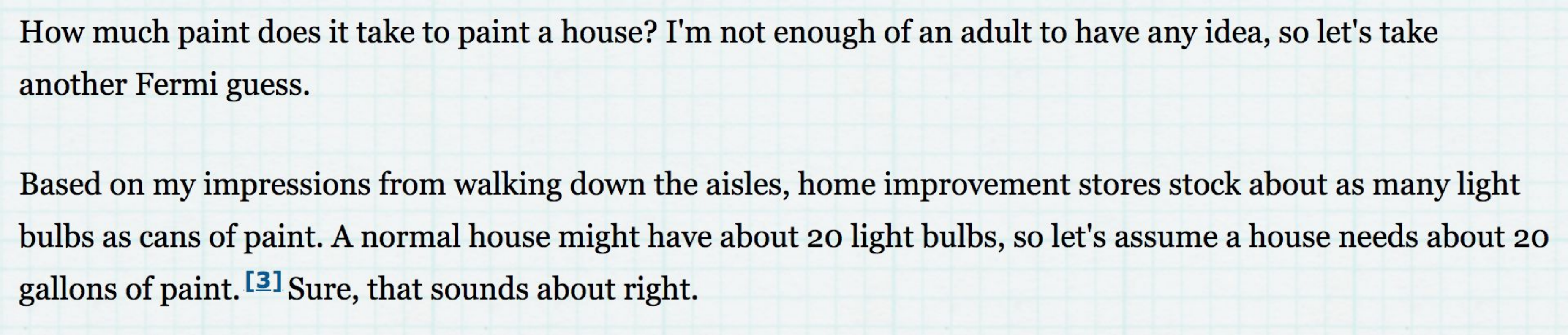
NOT ENOUGH	EXACTLY ENOUGH	MORE THAN ENOUGH
I		

https://what-if.xkcd.com/84/



Let's make a wild guess that, on average, one person out of every thousand spends their working life painting things. If I assume it would take me three hours to paint the room I'm in,[2] and 100 billion people have ever lived, and each of them spent 30 years painting things for 8 hours a day, we come up with 150 trillion square meters ... just about exactly the land area of the Earth.

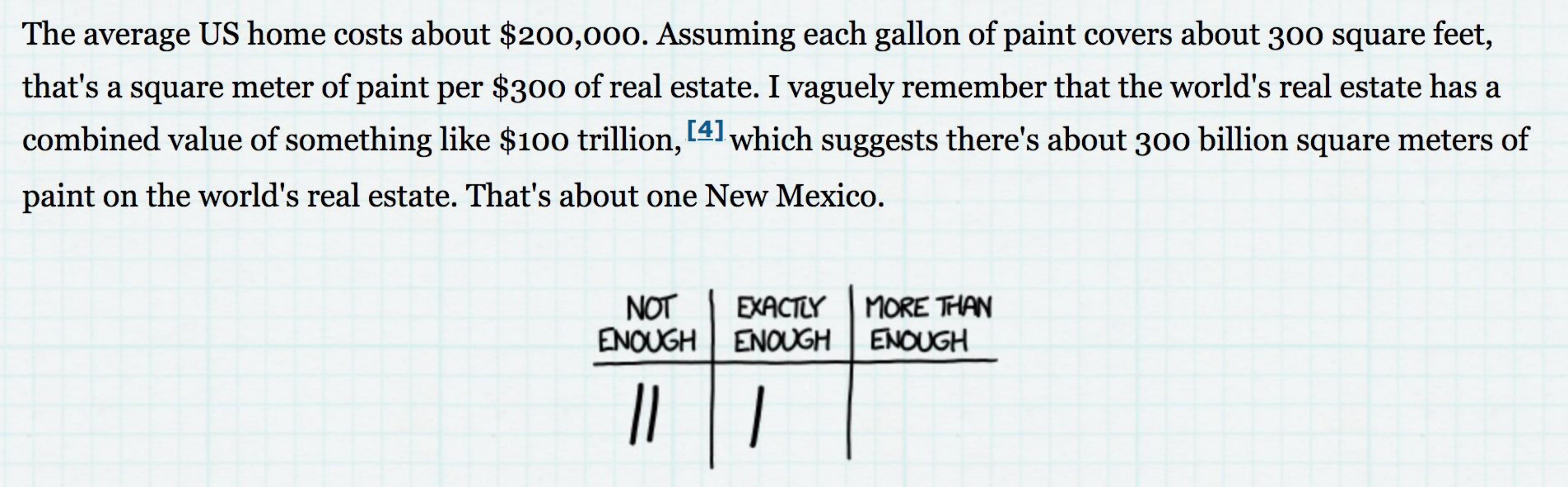
https://what-if.xkcd.com/84/



How much paint does it take to paint a house? I'm not enough of an adult to have any idea, so let's take another Fermi guess.

Based on my impressions from walking down the aisles, home improvement stores stock about as many light bulbs as cans of paint. A normal house might have about 20 light bulbs, so let's assume a house needs about 20 gallons of paint.[3] Sure, that sounds about right.

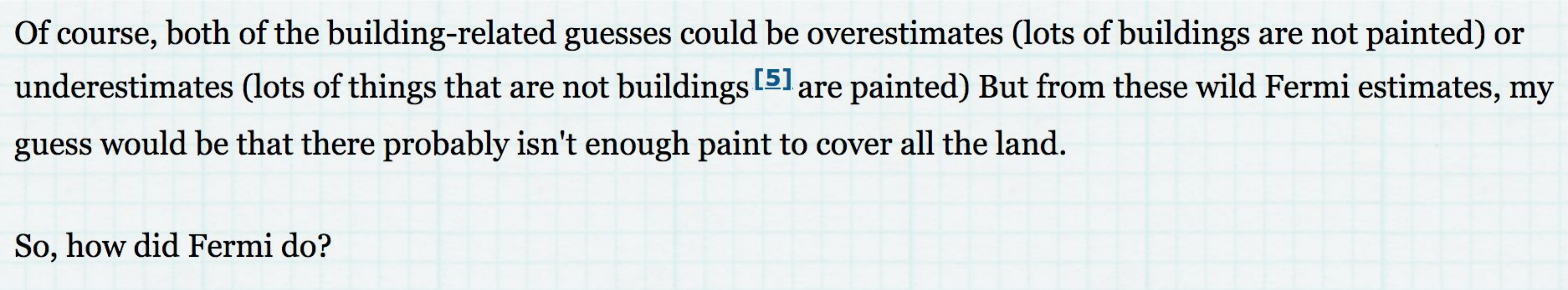
https://what-if.xkcd.com/84/



The average US home costs about $200,000. Assuming each gallon of paint covers about 300 square feet, that's a square meter of paint per $300 of real estate. I vaguely remember that the world's real estate has a combined value of something like $100 trillion,[4] which suggests there's about 300 billion square meters of paint on the world's real estate. That's about one New Mexico.

NOT ENOUGH	EXACTLY ENOUGH	MORE THAN ENOUGH
//	/	

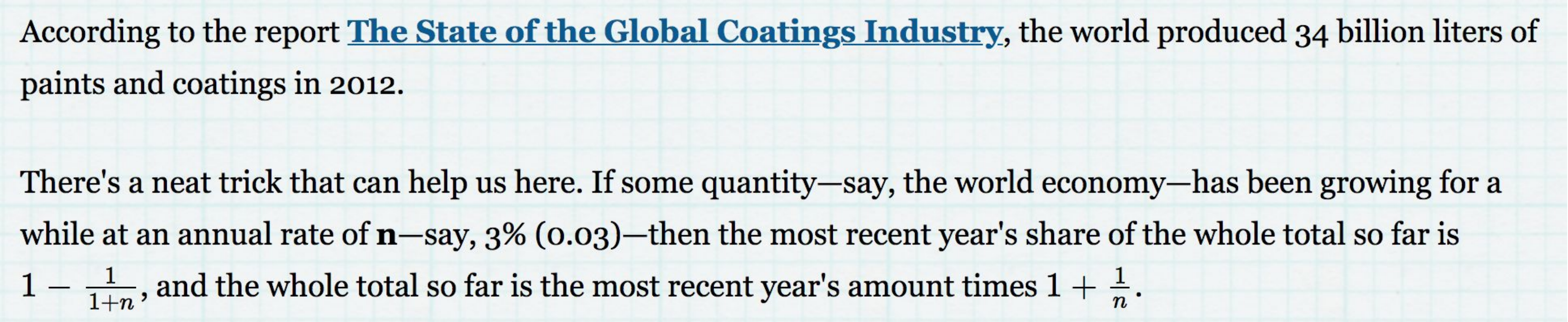
https://what-if.xkcd.com/84/



Of course, both of the building-related guesses could be overestimates (lots of buildings are not painted) or underestimates (lots of things that are not buildings [5] are painted) But from these wild Fermi estimates, my guess would be that there probably isn't enough paint to cover all the land.

So, how did Fermi do?

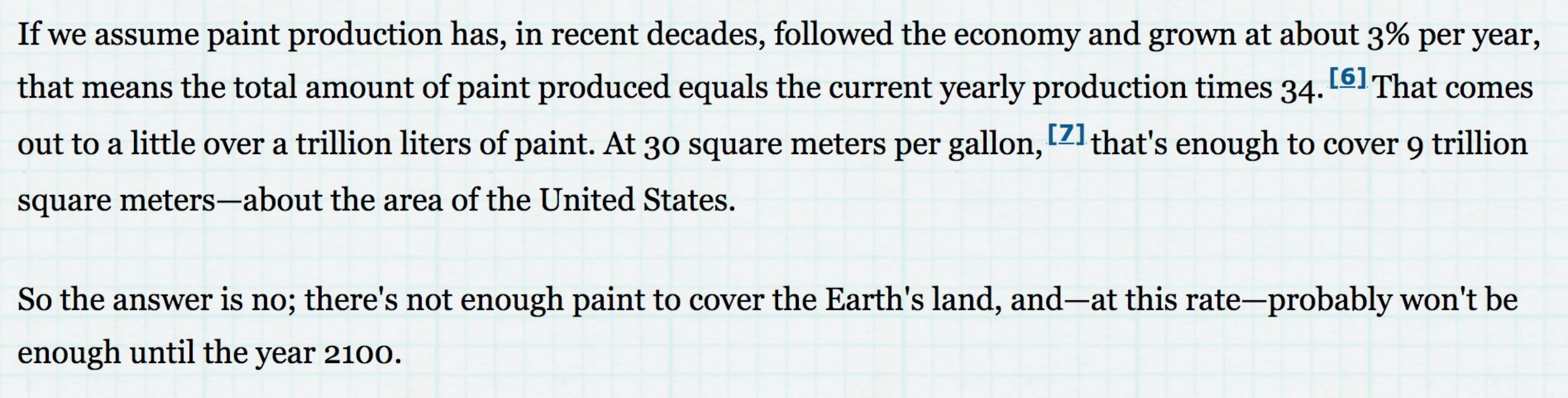
https://what-if.xkcd.com/84/



According to the report **The State of the Global Coatings Industry**, the world produced 34 billion liters of paints and coatings in 2012.

There's a neat trick that can help us here. If some quantity—say, the world economy—has been growing for a while at an annual rate of **n**—say, 3% (0.03)—then the most recent year's share of the whole total so far is $1 - \frac{1}{1+n}$, and the whole total so far is the most recent year's amount times $1 + \frac{1}{n}$.

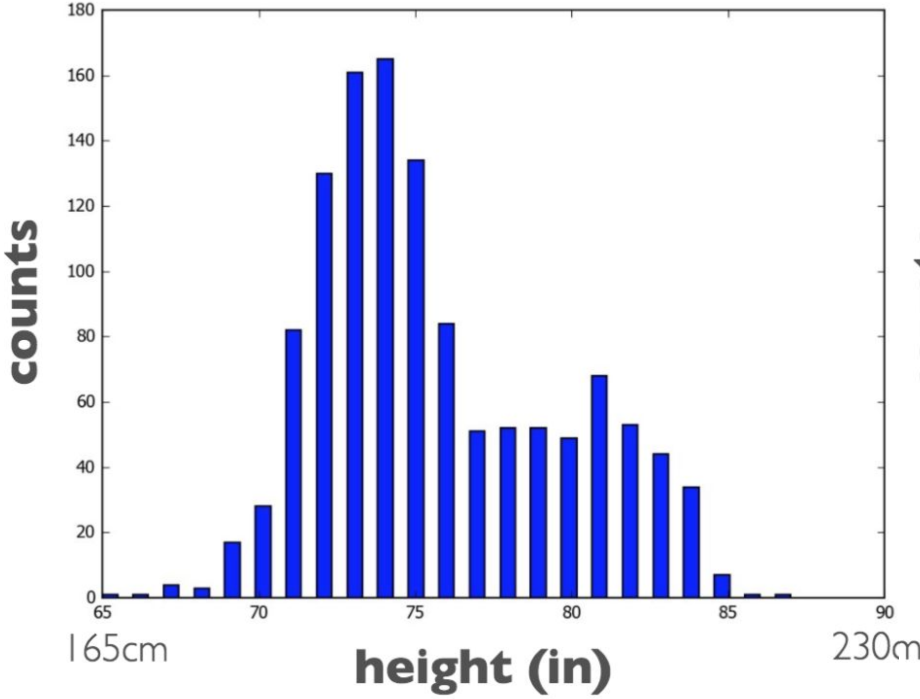
https://what-if.xkcd.com/84/

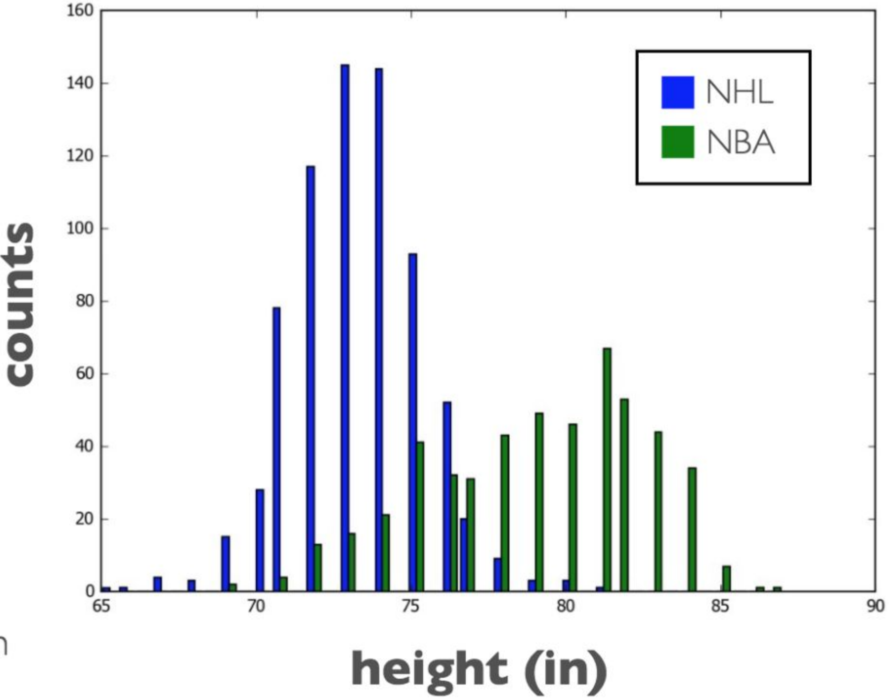


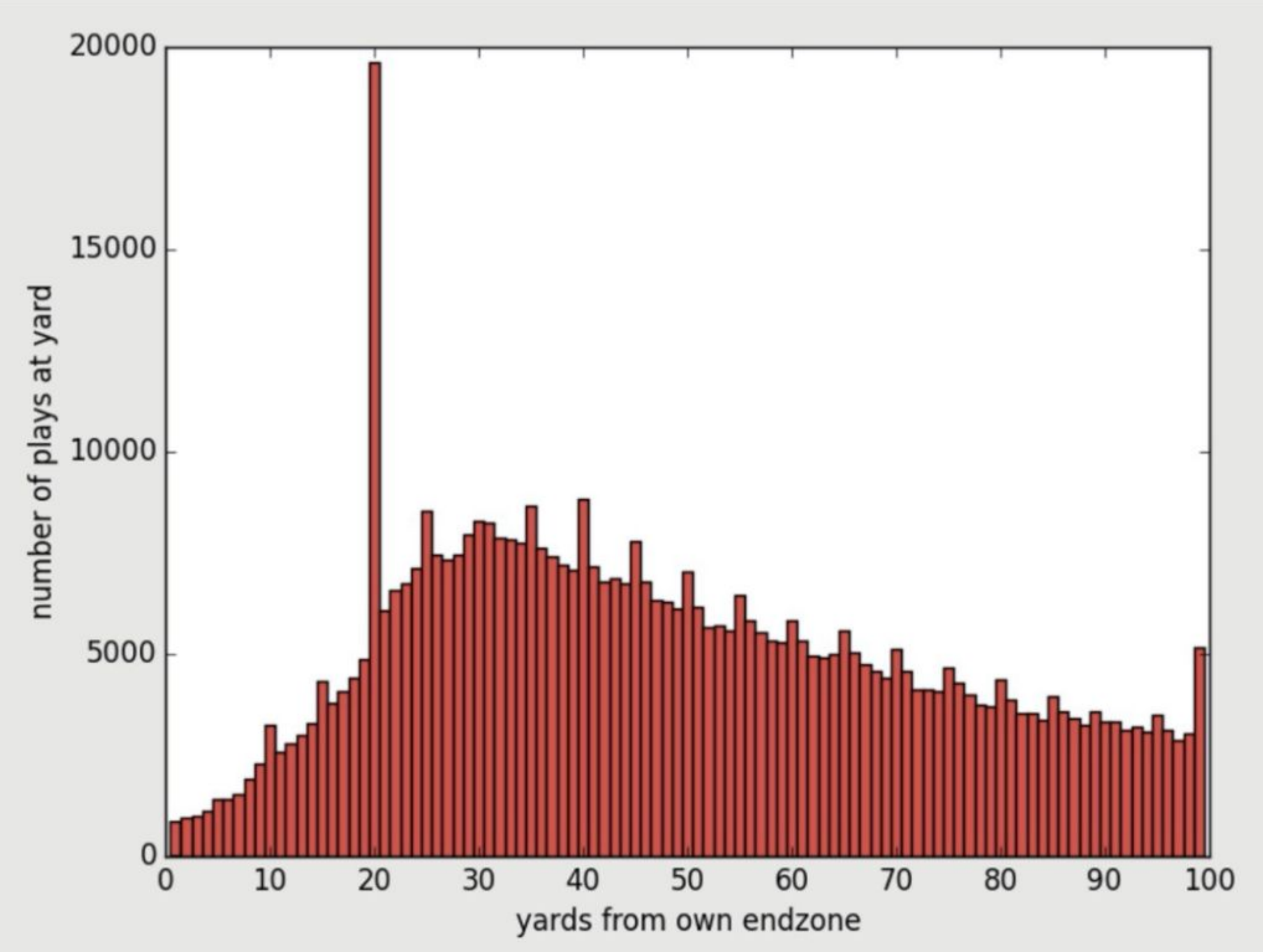
If we assume paint production has, in recent decades, followed the economy and grown at about 3% per year, that means the total amount of paint produced equals the current yearly production times 34.[6] That comes out to a little over a trillion liters of paint. At 30 square meters per gallon,[7] that's enough to cover 9 trillion square meters—about the area of the United States.

So the answer is no; there's not enough paint to cover the Earth's land, and—at this rate—probably won't be enough until the year 2100.

https://what-if.xkcd.com/84/


180
160
140
120
100
80
60
40
20
0
counts
65
70
75
80
85
90
165cm
height (in)
230m

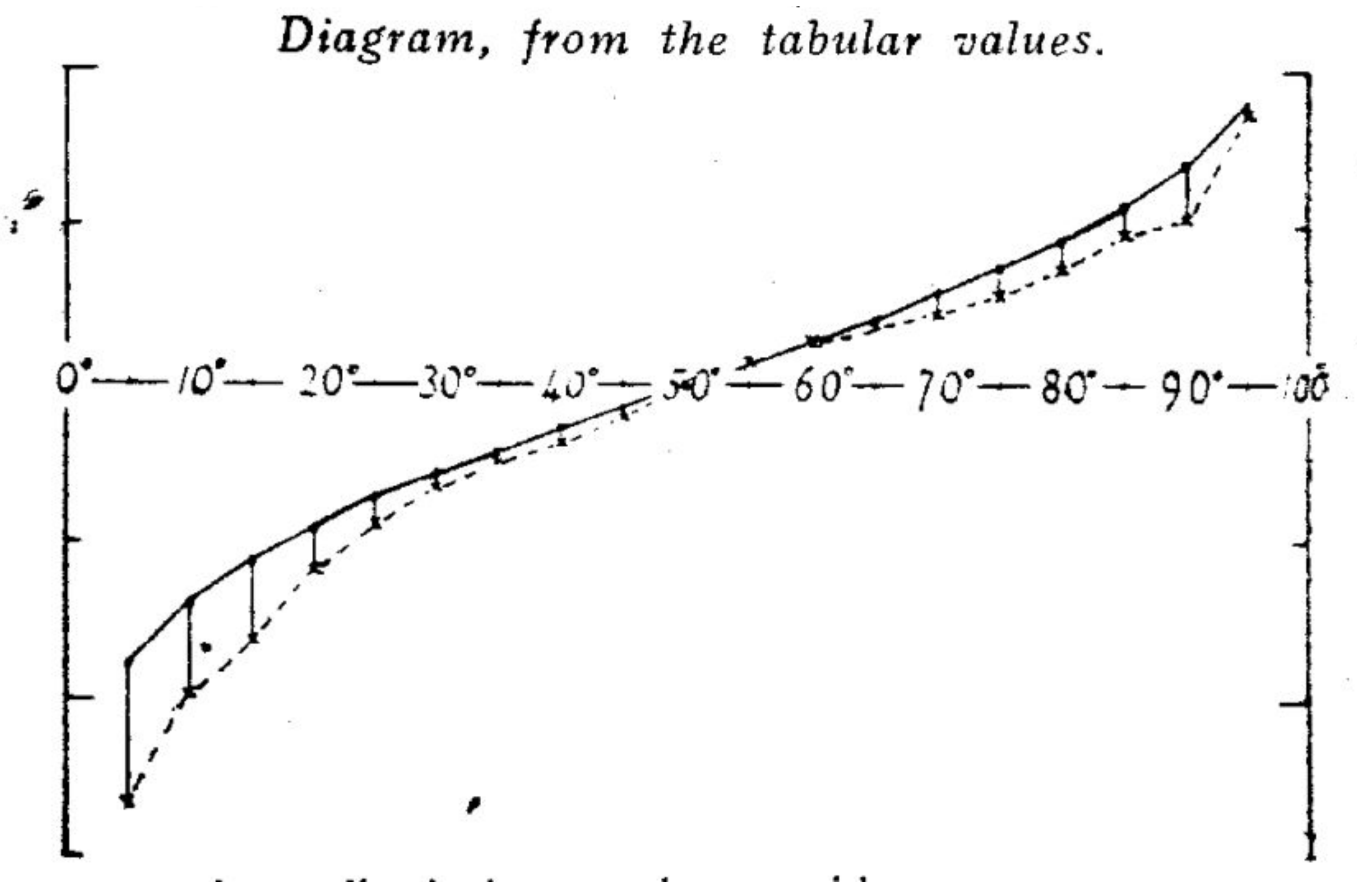

160
140
120
100
80
60
40
20
0
counts
NHL
NBA
65
70
75
80
85
90
height (in)


number of plays at yard
20000
15000
10000
5000
0
0
10
20
30
40
50
60
70
80
90
100
yards from own endzone

Data Intuition

1. Think about your question and your expectations
2. Do some Fermi calculations (back of the envelope calculations)
3. Write code & look at outputs <- think about those outputs
4. Use your gut instinct / background knowledge to guide you
5. Review code & fix bugs

Vox Populi



Galton, *Nature* (1907)

The Wisdom of the Crowds

- Diversity of opinion: Each person should have private information….even if it's just an eccentric interpretation of the known facts
- Independence: People's opinions aren't determined by the opinions of those around them
- Decentralization: People are able to specialize and draw on local knowledge
- Aggregation: Some mechanism exists for turning private judgements into a collective decision

On your own (meaning w/o Googling), please fill out quickly:

http://bit.ly/fermi_wi21