

# Designing data science

***November 2019***

Hadley Wickham  
[@hadleywickham](https://twitter.com/hadleywickham)  
Chief Scientist, RStudio



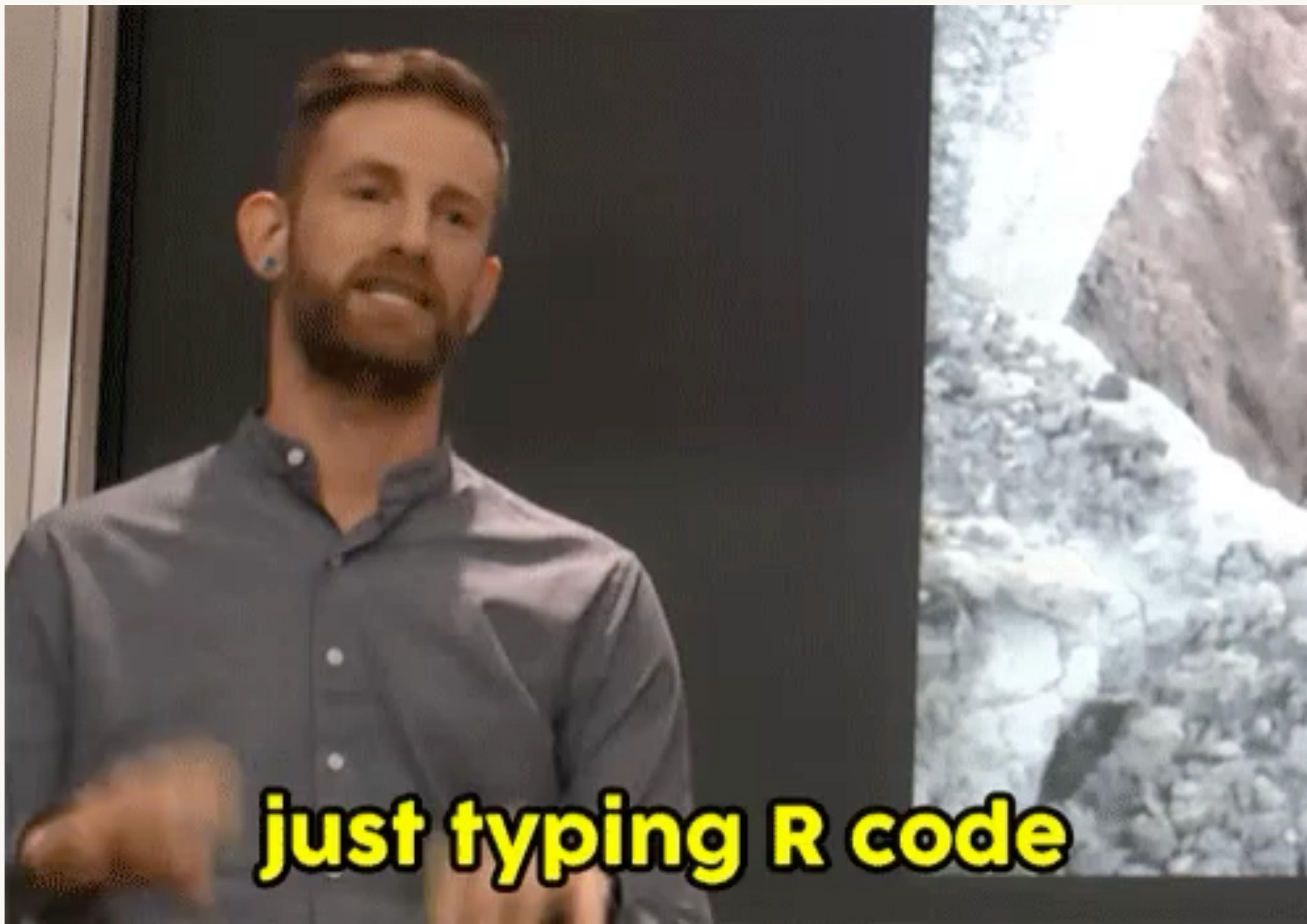
What do I do?

$$\sigma\sqrt{2\pi}$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



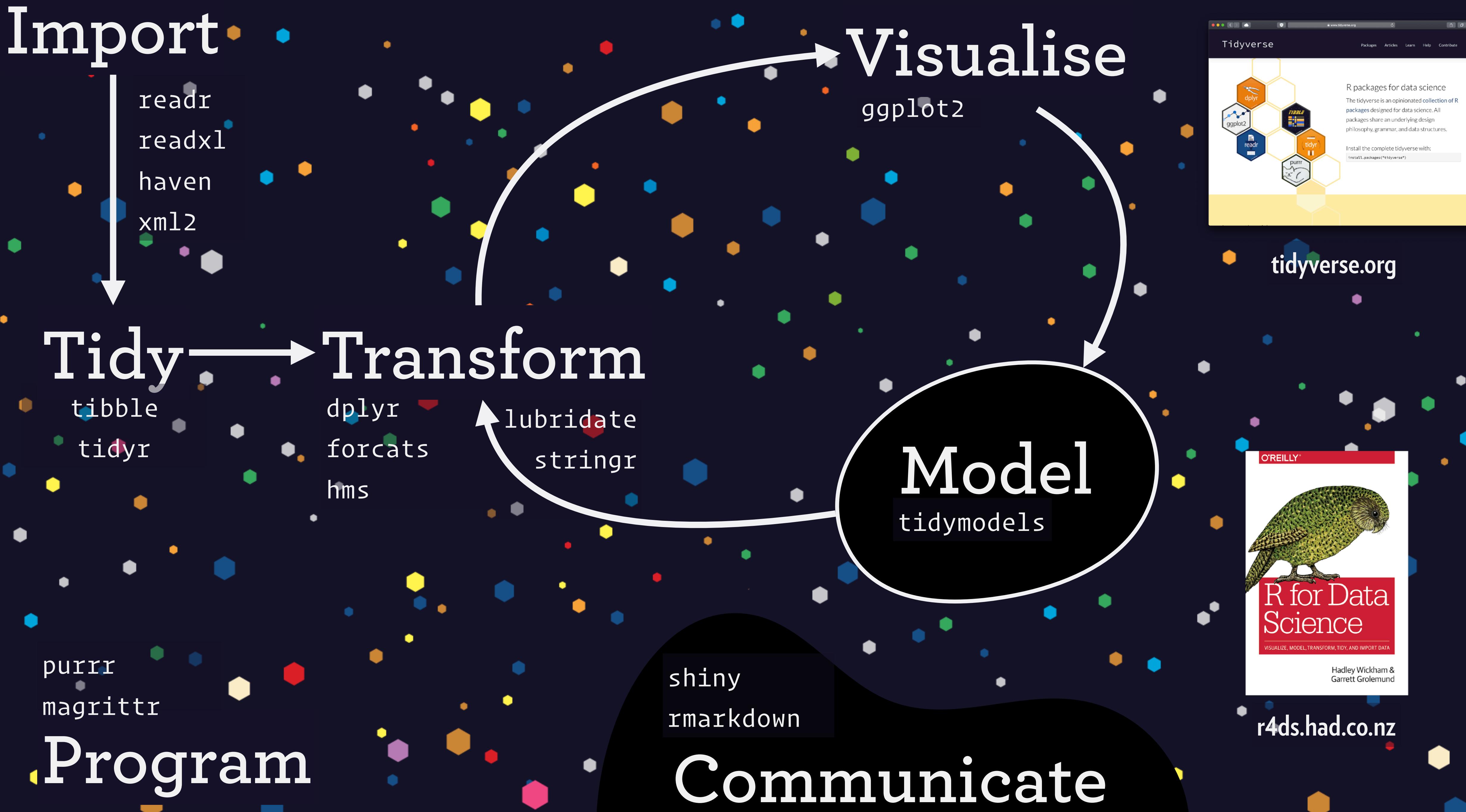
Pit of success



**just typing R code**

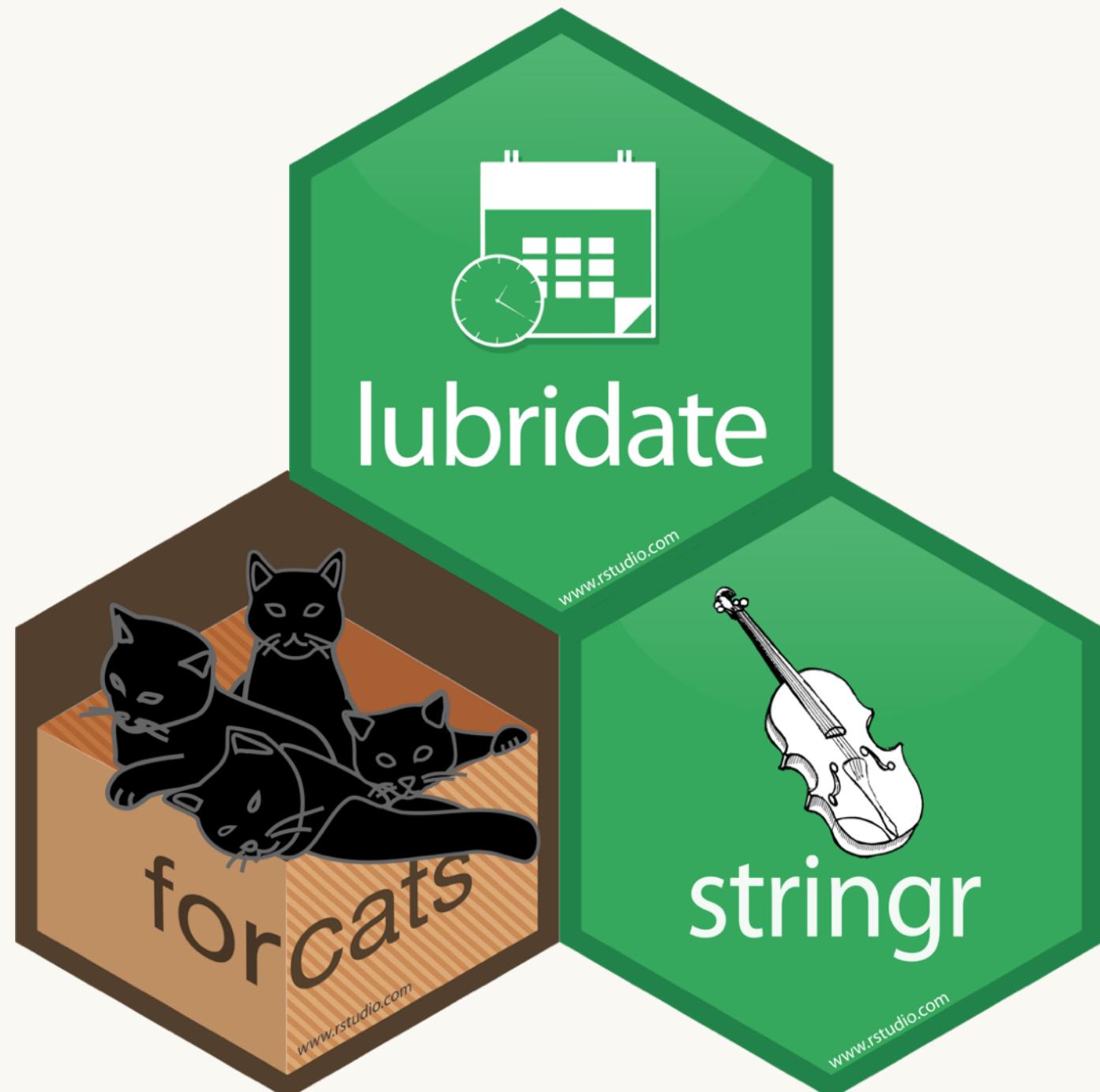
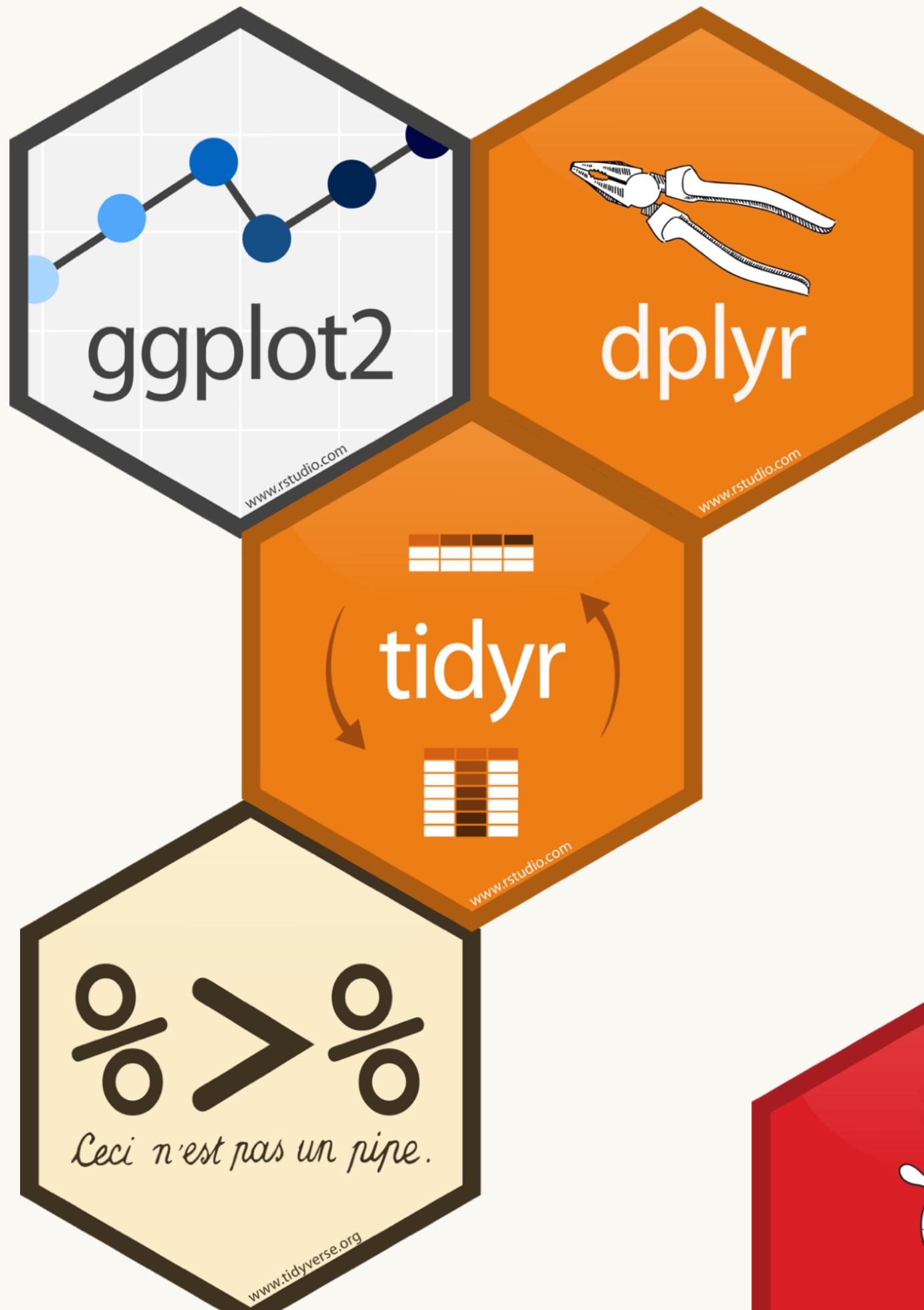
We wanted users to be able to begin in an interactive environment, where they did not consciously think of themselves as programming. Then as their needs became clearer and their sophistication increased, they should be able to slide gradually into programming, when the language and system aspects would become more important.

— John Chambers, “Stages in the Evolution of S”



Design

Software  
engineering



# What is design?

<https://99percentinvisible.org/article/norman-doors>

# Designerly ways of knowing by Nigel Cross

Education in any of these cultures entails the following three aspects:

- the transmission of **knowledge** about a phenomenon of study
- a training in the appropriate **methods** of enquiry
- an initiation into the **belief systems** and values of the culture

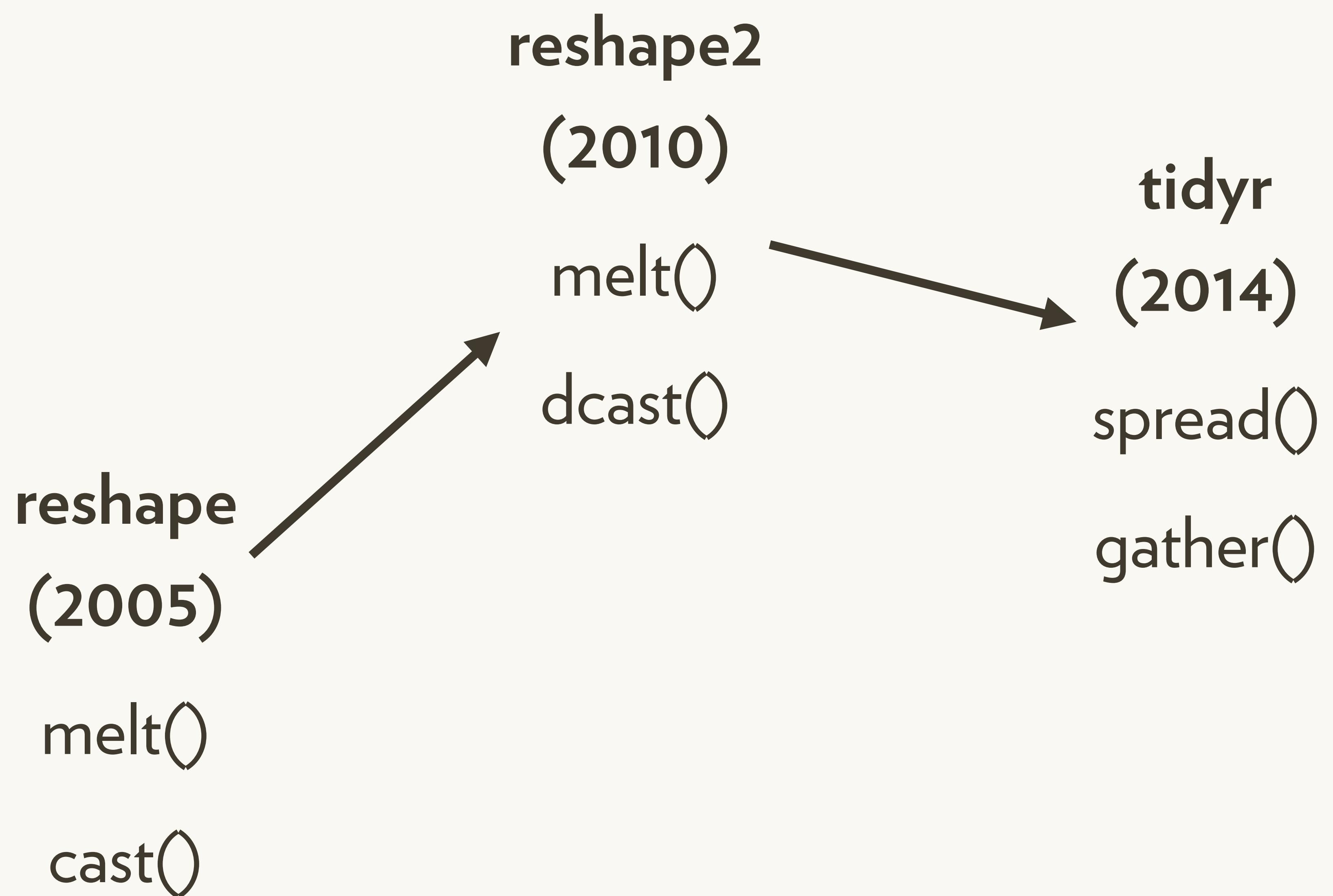
	Phenomenon of study	Values	Appropriate methods
Humanities	human experience	subjectivity, imagination, commitment, and a concern for ‘justice’	analogy, metaphor, criticism, evaluation
Sciences	the natural world	objectivity, rationality, neutrality, and a concern for ‘Truth’	controlled experiment, classification, analysis
Design	the constructed world	practicality, ingenuity, empathy, and a concern for ‘appropriateness’	modelling, pattern-formation, synthesis

1. Design is iterative
2. Design embraces constraints
3. Design is affective
4. Design is taking the blame

Design is iterative

# What is tidy data?

1. Each variable in a column
2. Each observation in a row
3. Each cell is one value



twitter.com/irismwang/status/994349937944158208

Home Notifications Messages

Iris Wang @irismwang

Follow

Any R functions you find you have to google every time to make sure you've got it right? For me it's gather and spread. Still.

5:54 PM - 9 May 2018

3 Likes

Tweet your reply

Privacy policy Imprint

**Trends for you**

#Tatort #Montagslaecheln #EESmicrobiome #bitcoin #Mannheim Marc Benioff  
#Mangkhut Frankfurt #blackandwhitechallenge #Brexit

© 2018 Twitter About Help Center Terms Privacy policy Imprint Cookies Ads info

<https://twitter.com/irismwang/status/994349937944158208>

twitter.com/BrockTibert/status/1009543843690278920

Home Notifications Messages  Search Twitter  Tweet  X

 **Brock Tibert**  
@BrockTibert

**Follow**

tidyr is fantastic, but I'd be lying if I didn't have to do ?gather and ?spread at least once a week #rstats

4:09 PM - 20 Jun 2018

1 Like 

2  1  

 **Tanya Cashorali** @tanyacash21 · Jun 20  
Replying to @BrockTibert  
Dude the things I still google... I wish more people scared of coding knew how much we do this.

2   2 

 **Brad Weiner** @brad\_weiner · Jun 20  
Replying to @BrockTibert  
That sure beats the 3 times weekly that I had to Google melt and cast.

1   



**Trends for you**

#Tatort #Montagslaecheln #EESmicrobiome #bitcoin #Mannheim Marc Benioff  
#Mangkhut Frankfurt #blackandwhitechallenge #Brexit

<https://twitter.com/BrockTibert/status/1009543843690278920>

twitter.com/mattfrost/status/768078660200898560

Home Notifications Messages Search Twitter Tweet X

**Matt Frost** @mattfrost Follow

Will the tidyR syntax for gather() and spread() ever become as intuitive as the reshape syntax for melt() and cast()? Please say yes.

8:33 AM - 23 Aug 2016

1 reply 1 retweet 0 likes

**Azad AF** @azadag · 23 Aug 2016 Replying to @mattfrost

I can't really get my head around it despite getting it to work successfully... but others say that it's intuitive to them

1 reply 1 retweet 0 likes

**Matt Frost** @mattfrost · 23 Aug 2016 I keep expecting it to click, then I keep typing ?gather

1 reply 1 retweet 0 likes

**Azad AF** @azadag · 23 Aug 2016 the documentation is not helpful, bc it acts differently given groupings of variables. The vignette is a little better. But..

1 reply 1 retweet 0 likes

**Matt Frost** @mattfrost · 23 Aug 2016 On another note, you know about kable(), right? That is so rad.

1 reply 1 retweet 0 likes

**Azad AF** @azadag · 23 Aug 2016 I wish I could just produce markdown products. I'm knee deep in ReporteRs making notes

<https://twitter.com/mattfrost/status/768078660200898560>

twitter.com/gshotwell/status/675344503566417921

Home Notifications Messages Search Twitter Tweet X

Gordon Shotwell @gshotwell Following

Any **#rstats** people have a good mnemonic device for remembering which tidyverse function does what? I reliably confuse gather and spread.

10:00 AM - 11 Dec 2015

Reply Retweet Like Email

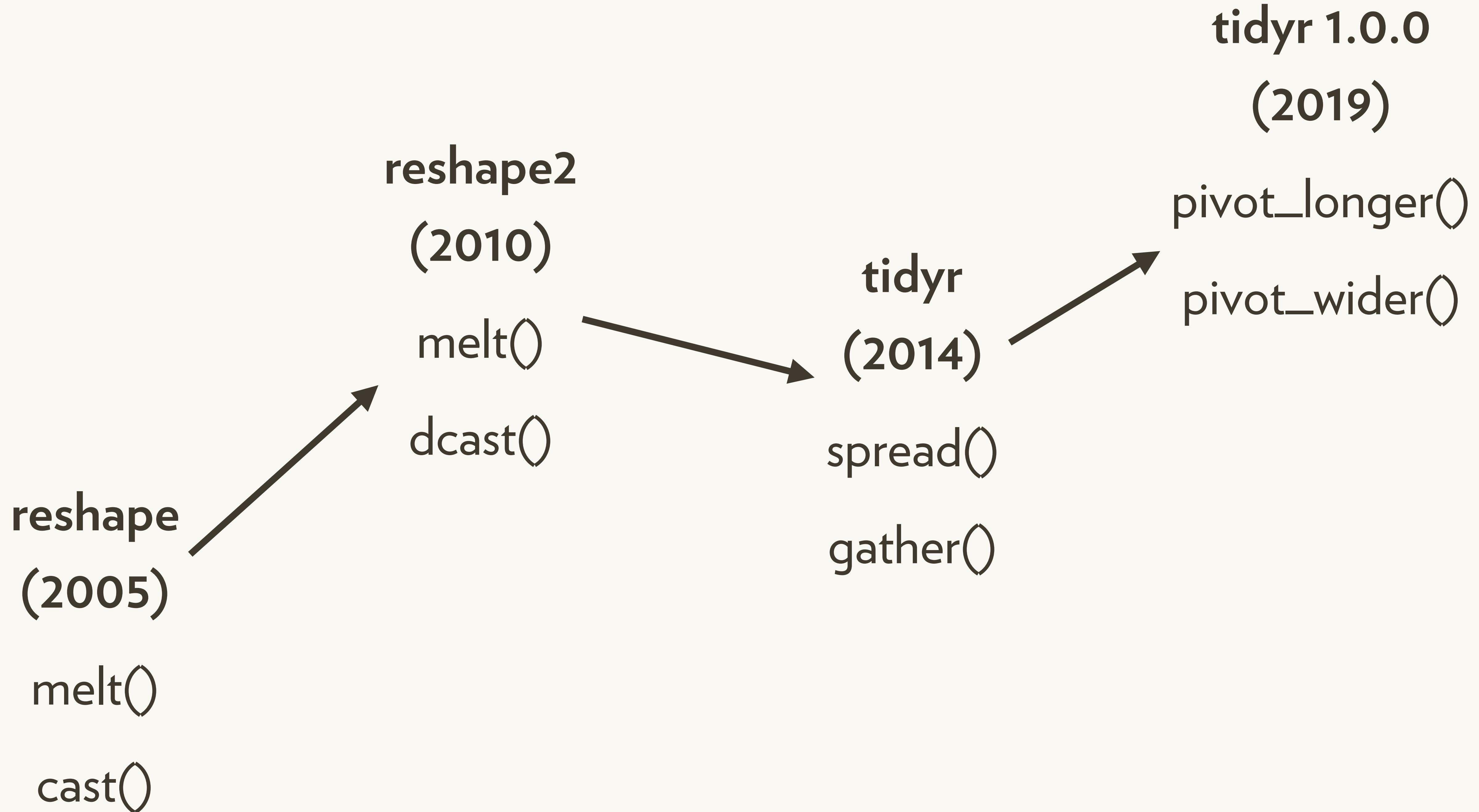
Tweet your reply

**Trends for you**

#Tatort #Montagslaecheln #EESmicrobiome #bitcoin #Mannheim Marc Benioff  
#Mangkhut Frankfurt #blackandwhitechallenge #Brexit

© 2018 Twitter About Help Center Terms Privacy policy Imprint Cookies Ads info

<https://twitter.com/gshotwell/status/675344503566417921>





**JD Long**  
@CMastication



I finally held my breath and updated {tidyr} just to play with the new pivot functions. I read the help file on `pivot\_wider` and that damn thing worked \*on the first try\*. That's preposterous.

[#rstats](#)[tidyverse.org/articles/pivot...](https://tidyverse.org/articles/pivot.html)



**Pivoting**  
[tidyverse.org](https://tidyverse.org)

♥ 222 2:07 PM - Oct 31, 2019



↪ 41 people are talking about this





**Travis Dawry**  
@tdawry



Some of my students had problems with `gather()` and `spread()`,  
but with `pivot_longer()` and `pivot_wider()` the tables have turned.

#rstats

182 2:14 PM - Oct 12, 2019



22 people are talking about this

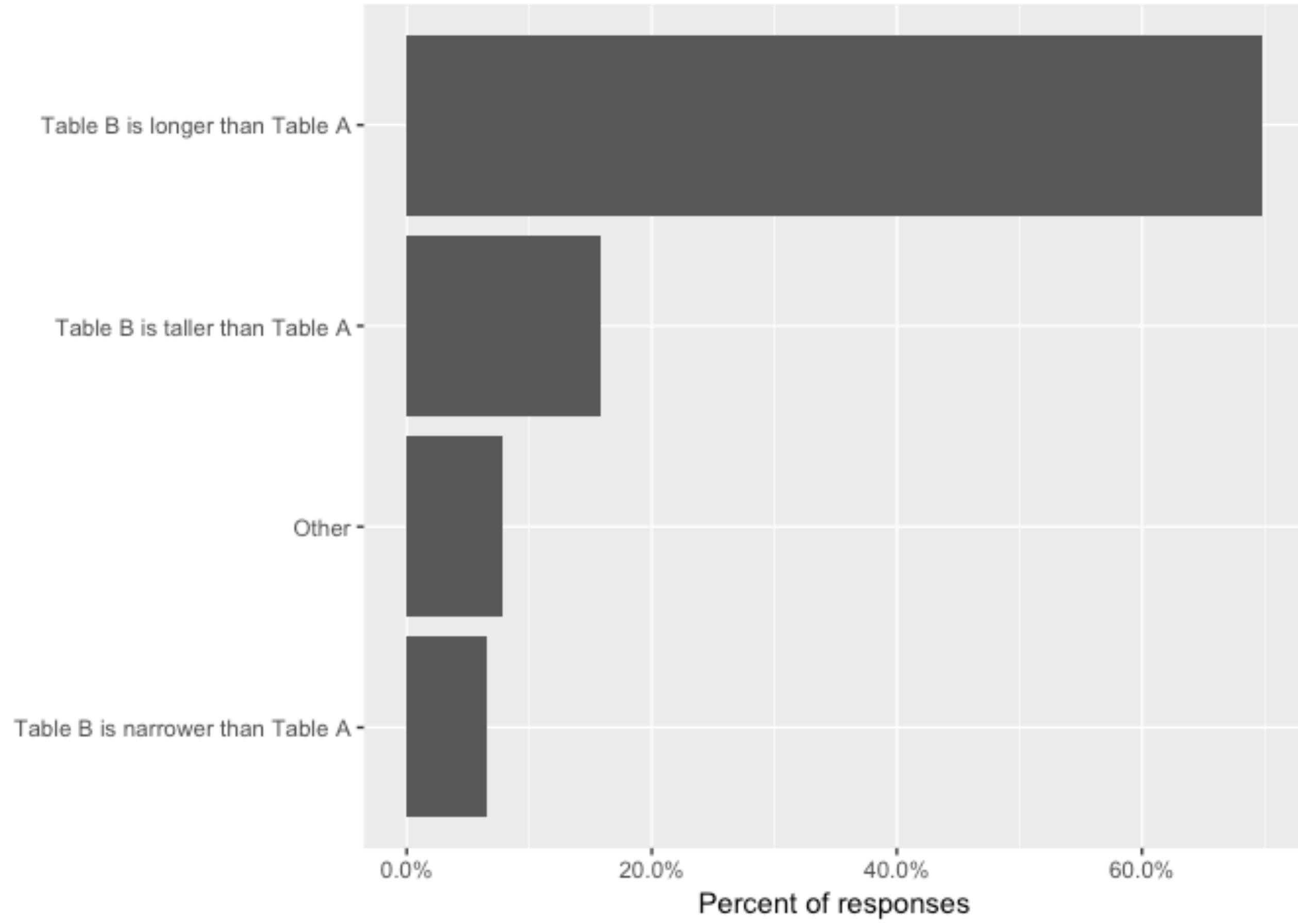
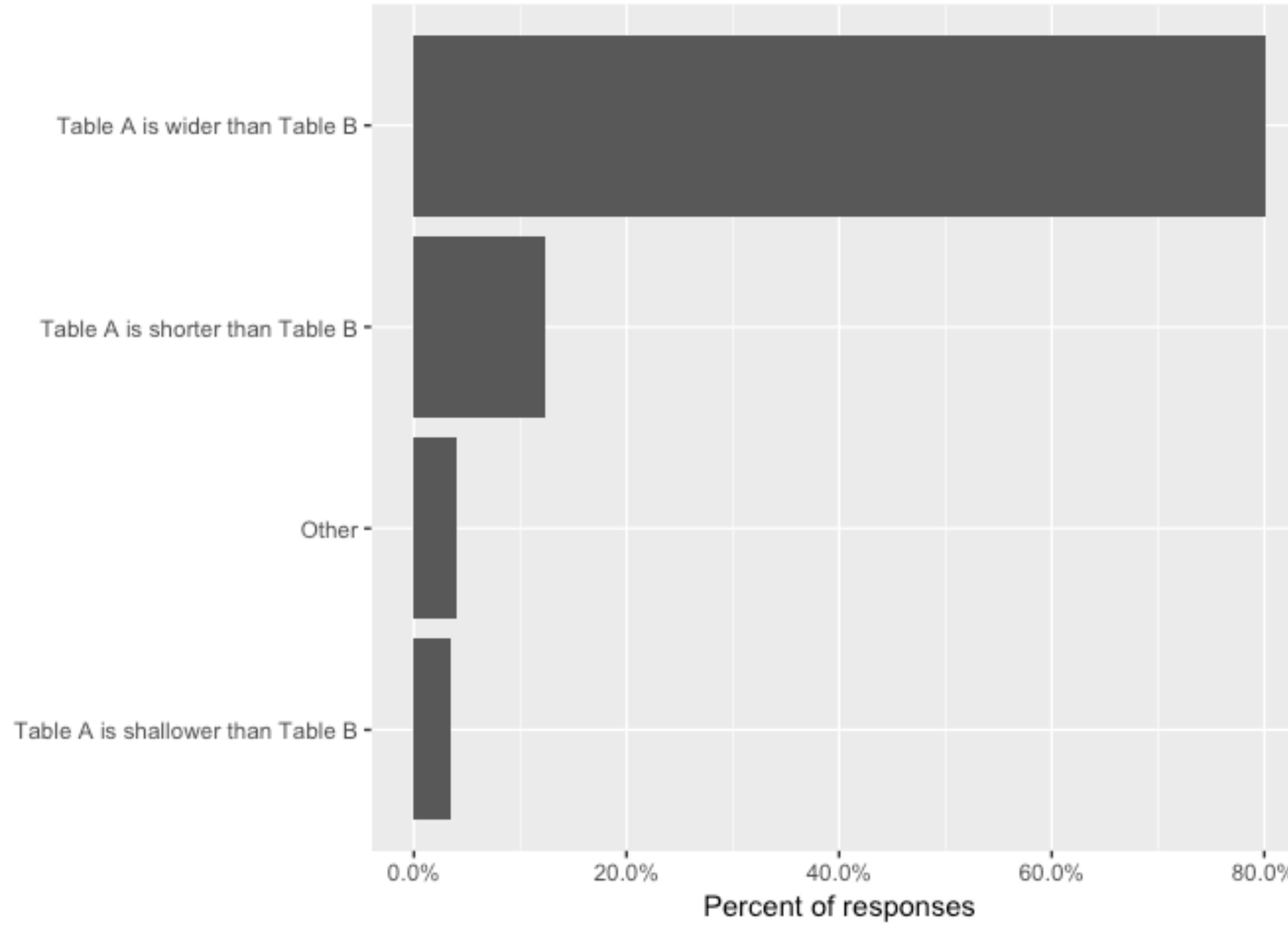


**Table A**

<b>id</b>	<b>x</b>	<b>y</b>	<b>z</b>
1	a	b	c
2	d	e	f

**Table B**

<b>id</b>	<b>n</b>	<b>v</b>
1	x	a
1	y	b
1	z	c
2	x	d
2	y	e
2	z	f



<https://github.com/hadley/table-shapes>

“The cost of never making a  
mistake is very often never making  
a change. It’s just too incredibly  
hard to be sure.”

— GeePaw Hill <http://geepawhill.org/try-different-not-harder/>

Design accepts  
constraints

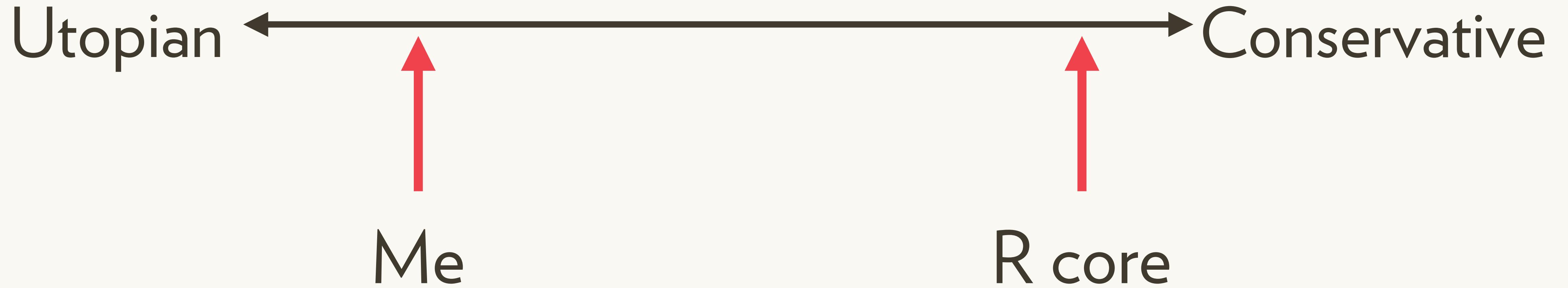
# What happens if you make a design mistake?



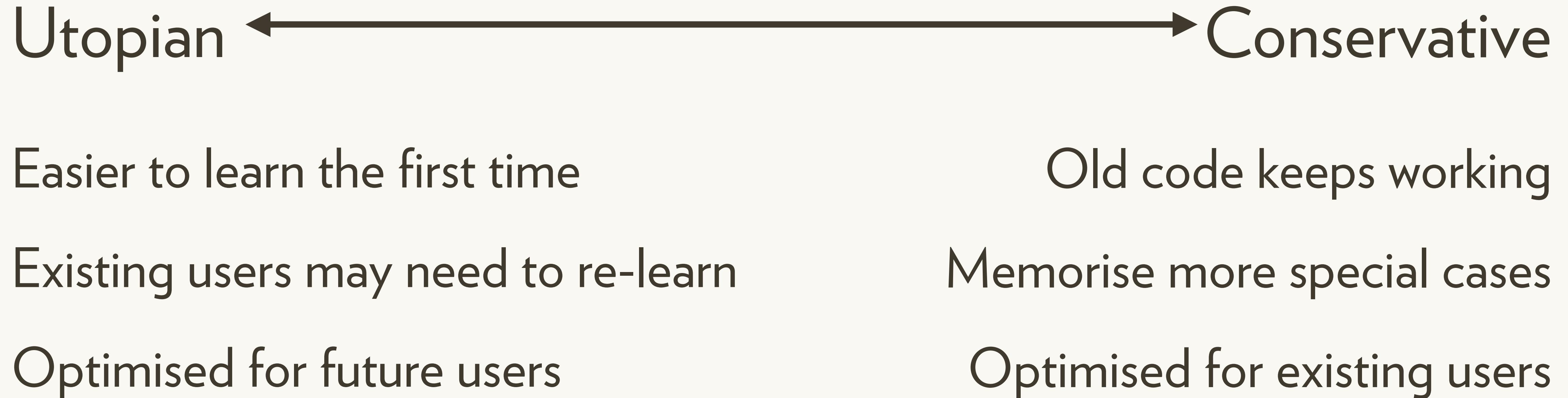
I think this is the greatest challenge in programming



I think this is the greatest challenge in programming



# Both have pros and cons



# Can you predict what this returns?

```
df <- data.frame(xyz = "a")  
df$x
```

# Data frames

```
df <- data.frame(xyz = "a")
df$x
#> [1] a
#> Levels: a
```

Can't change 20 year behaviour, can but introduce new classes

```
df <- tibble(xyz = "a")  
df$x  
#> NULL  
#> Warning message:  
#> Unknown or uninitialised column: 'x'.
```

# Tibbles

```
df <- tibble(xyz = "a")  
df$xyz  
#> [1] "a"
```

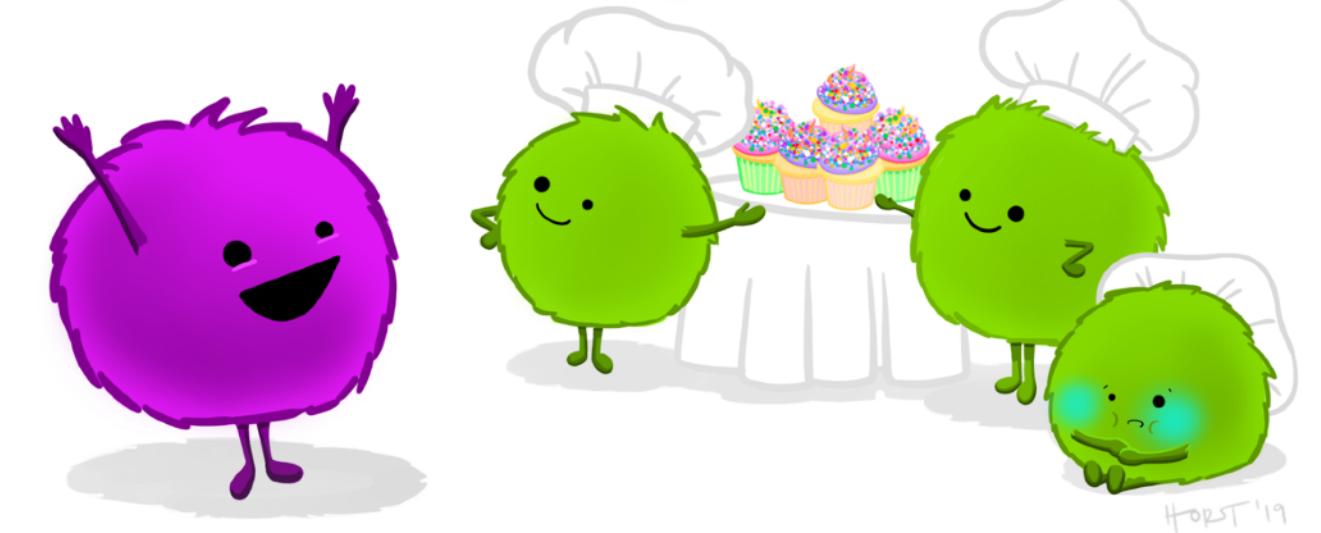
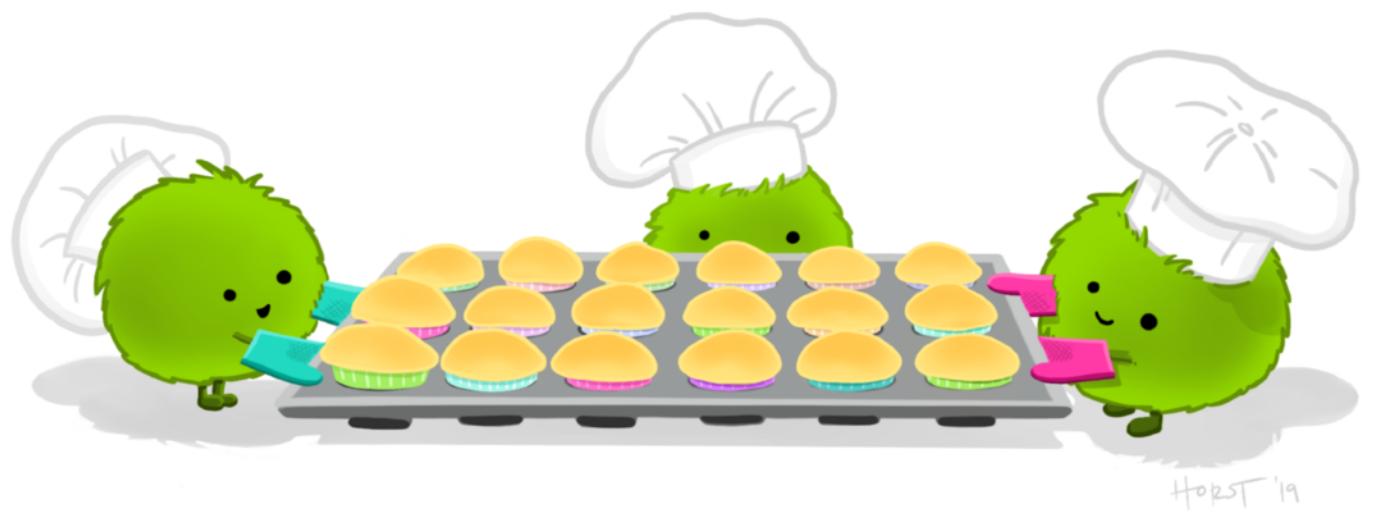
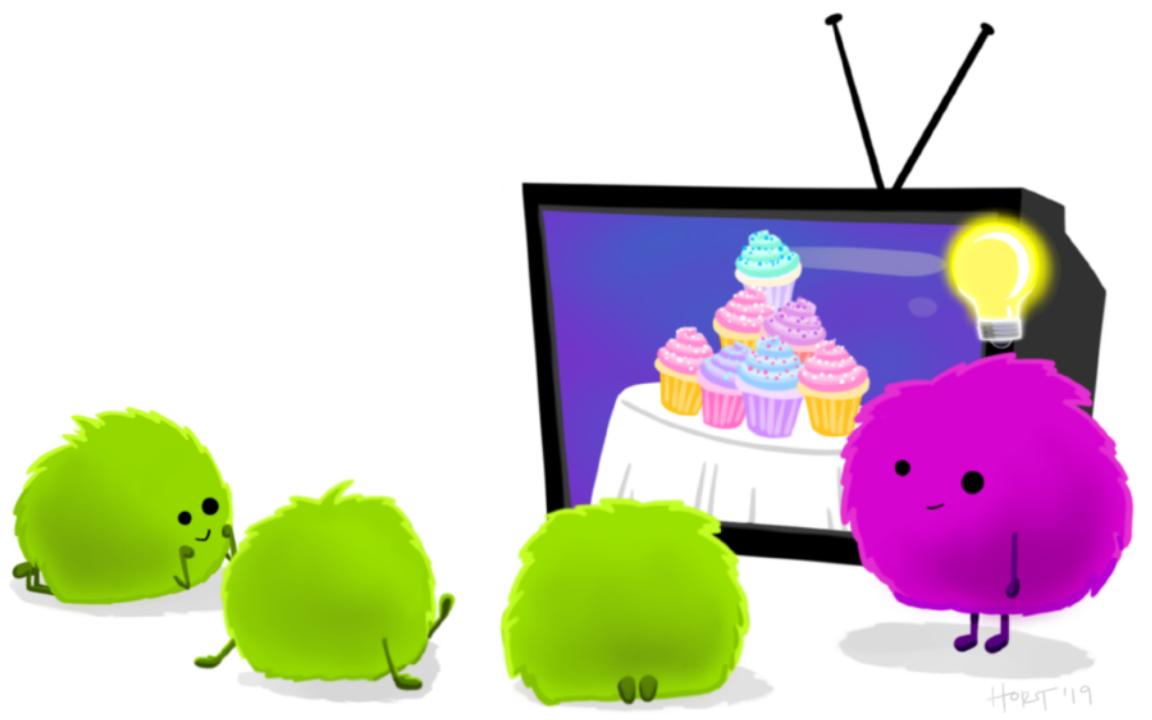
Is it worth it?

Design is affective



# The art of Allison Horst

<https://github.com/allisonhorst>



# Teacups, Giraffes, and Statistics

Hasse Walum and Desirée de Leon



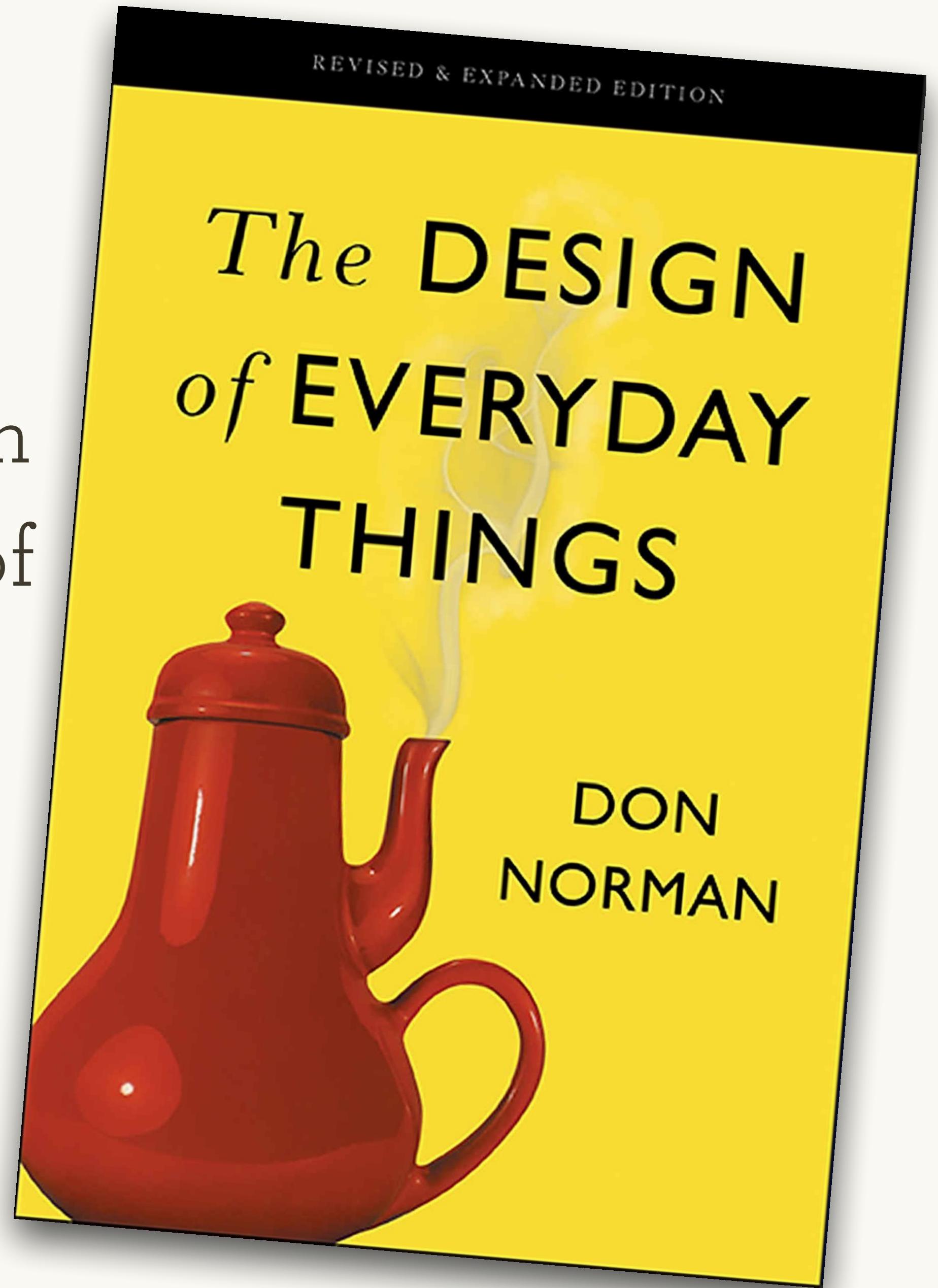
Design is taking the  
blame



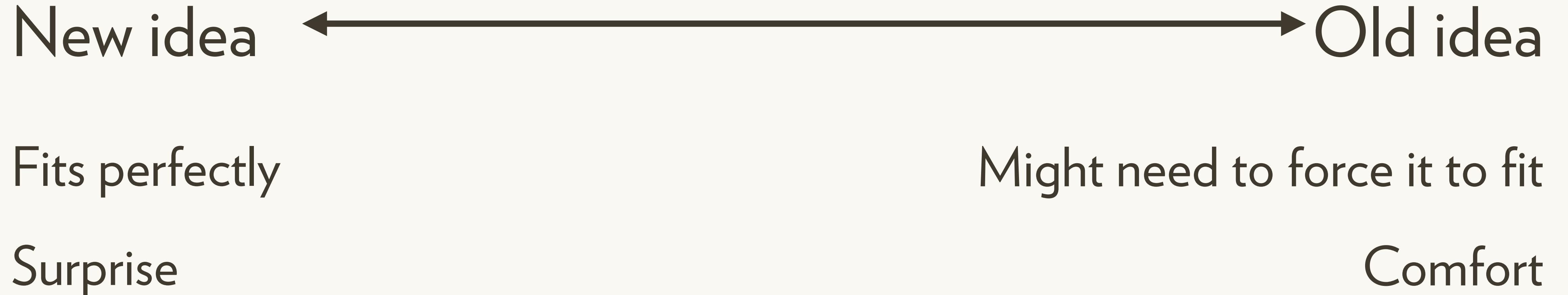
There's

“When you have trouble with things—whether it’s figuring out whether to push or pull a door or the arbitrary vagaries of the modern computer and electronics industries—it’s not your fault. Don’t blame yourself: blame the designer...”

— Donald A. Norman



But the idea of a door  
already exists in your head.  
How do new metaphors get  
in there?



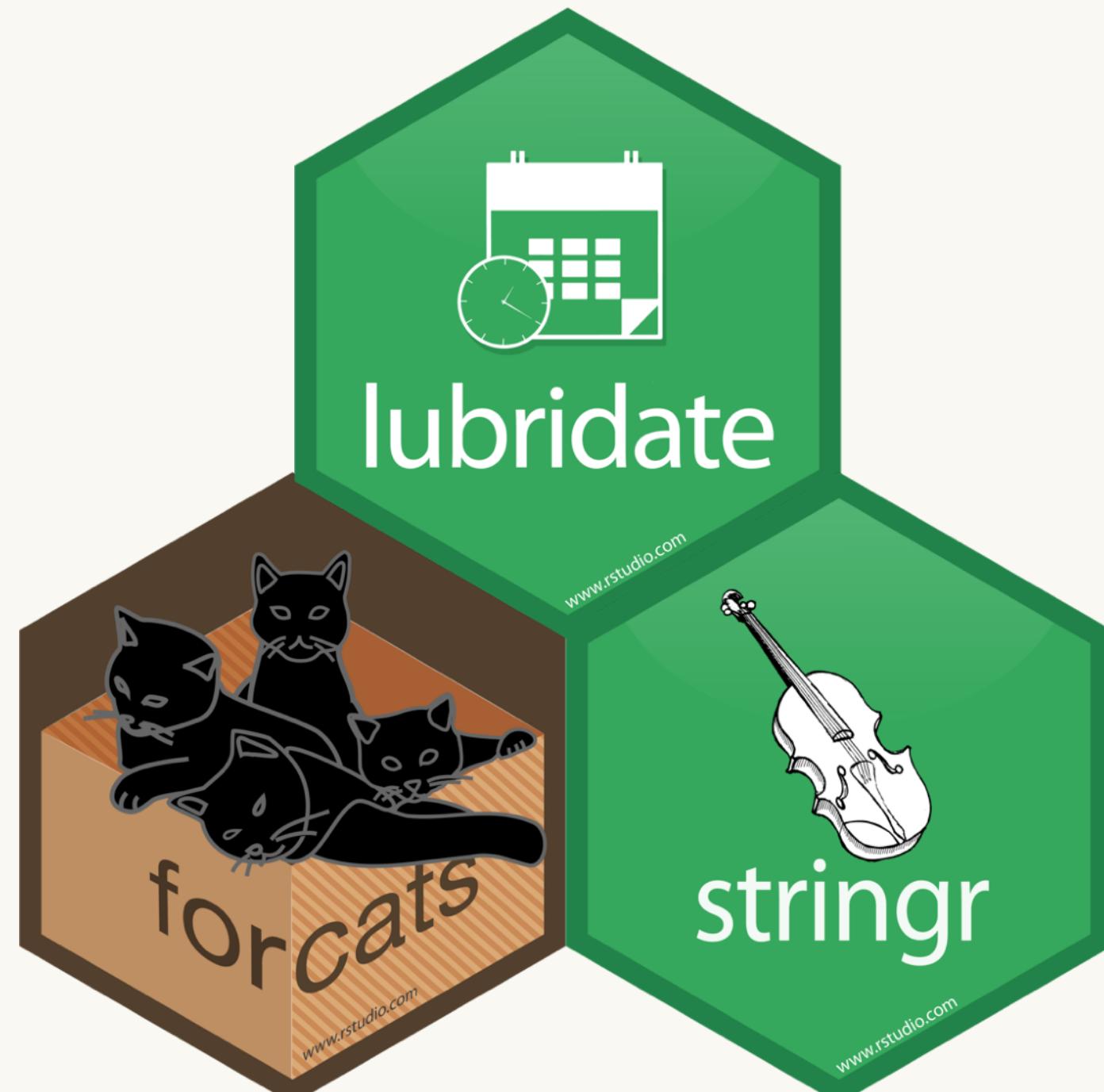
# Conclusion



Pit of success

Design

Software  
engineering



1. Design is iterative
2. Design embraces constraints
3. Design is affective
4. Design is taking the blame

This work is licensed as

Creative Commons  
Attribution-ShareAlike 4.0  
International

To view a copy of this license, visit

<https://creativecommons.org/licenses/by-sa/4.0/>

R	Python
Apple	Linux
CRAN	pip
Specialised	General purpose
Exploration	Automation
REPL + Rmd	Jupyter notebooks