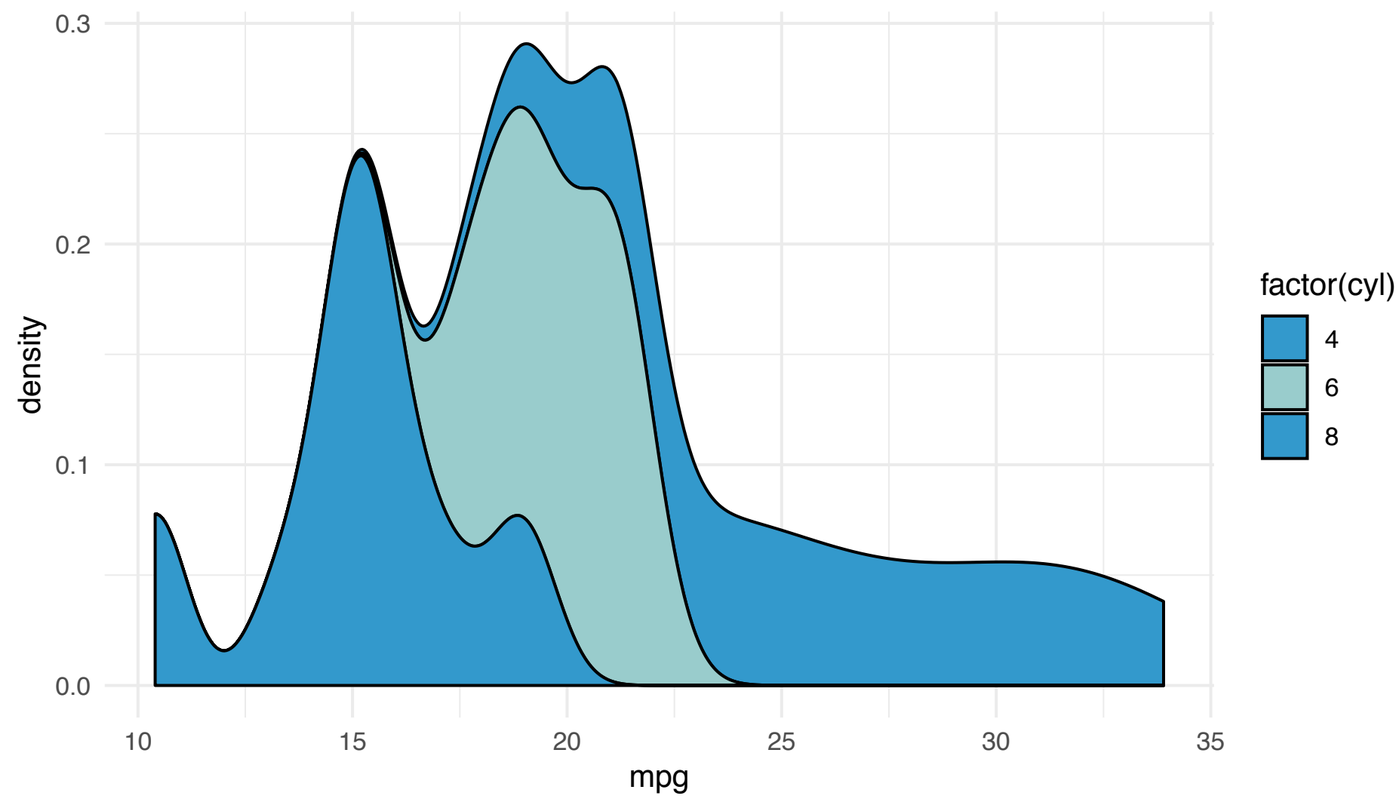
A Probabilistic Grammar of Graphics

Xiaoying Pu
Prelim presentation

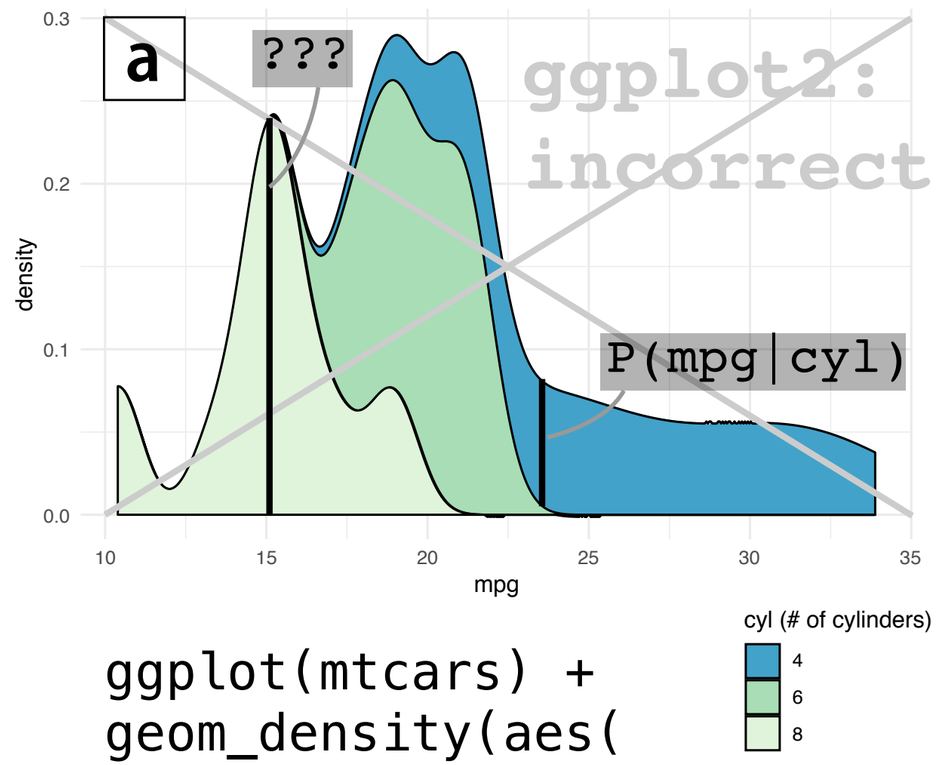
What could possibly go wrong?

	mpg	cyl	am
Mazda RX4	21.0	6	1
Mazda RX4 Wag	21.0	6	1
Datsun 710	22.8	4	1
Hornet 4 Drive	21.4	6	0
Hornet Sportabout	18.7	8	0
Valiant	18.1	6	0

```
ggplot(mtcars) +
geom_density(aes(
        x = mpg,
     fill = cyl),
 position = “stack”)
```

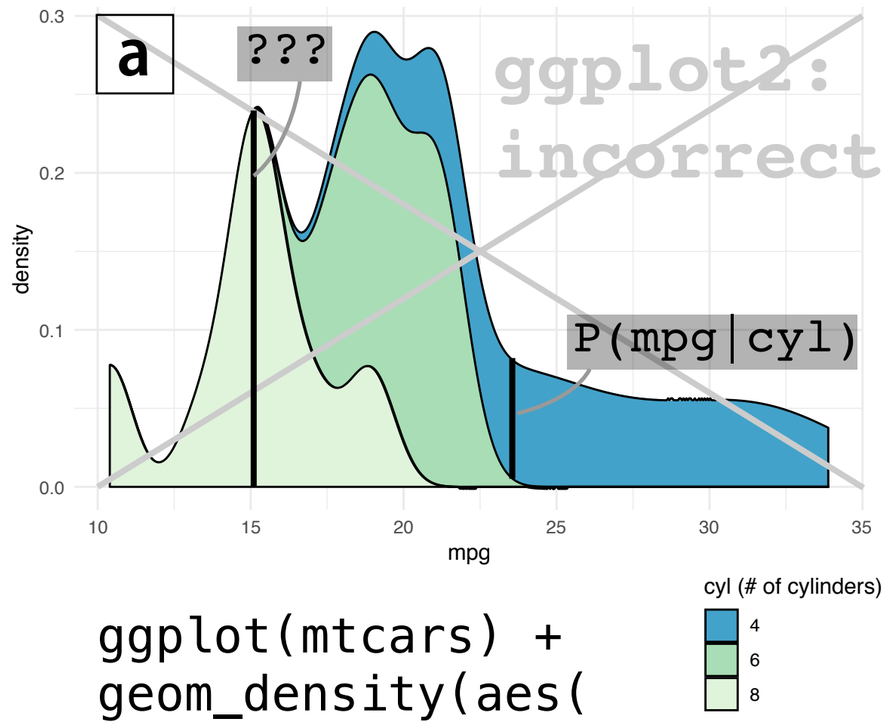


Problem 1: visualization shows incorrect probability distrib

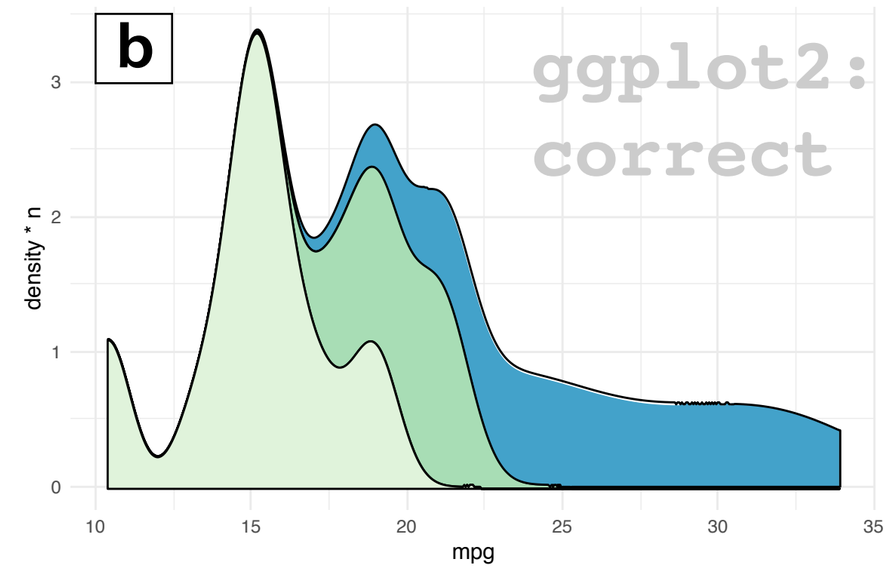



```
ggplot(mtcars) +
geom_density(aes(
        x = mpg,
     fill = cyl),
 position = "stack")
```

Wait we can fix this density plot

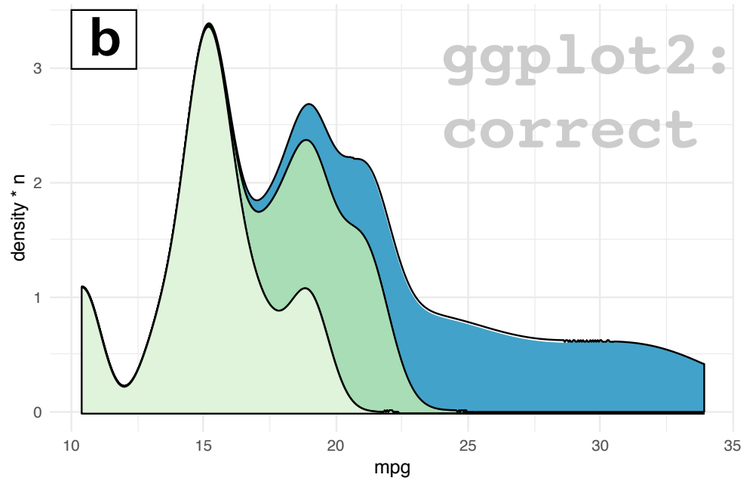



```
ggplot(mtcars) +
geom_density(aes(
        x = mpg,
     fill = cyl),
 position = "stack")
```

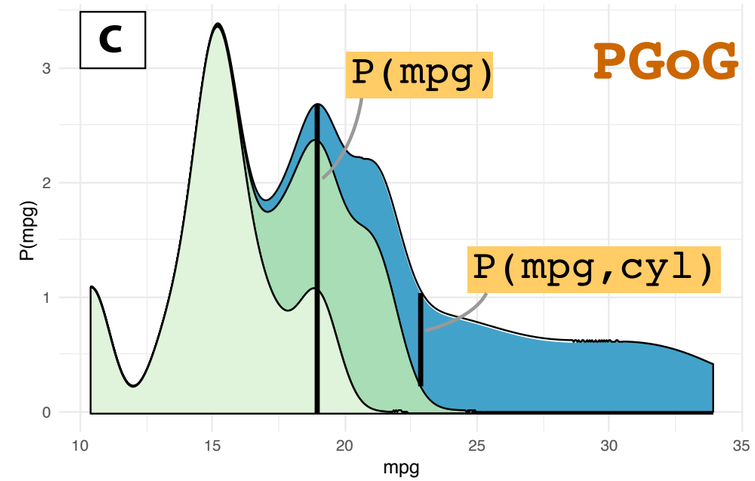



```
ggplot(mtcars)+
geom_density(aes(
        x = mpg,
        y = stat(density*n),
     fill = cyl)) +
position = "stack")
```

Problem 2: specifying probability distributions is convoluted




```
ggplot(mtcars)+
geom_density(aes(
       x = mpg,
       y = stat(density*n),
    fill = cyl)) +
position = "stack")
```




```
ggplot(mtcars) +
geom_bloc(aes(
     x = mpg,
height = P(mpg|cyl) P(cyl),
  fill = cyl)
```

But what are `stat(density*n)` and `position`?

$P(A|B)$

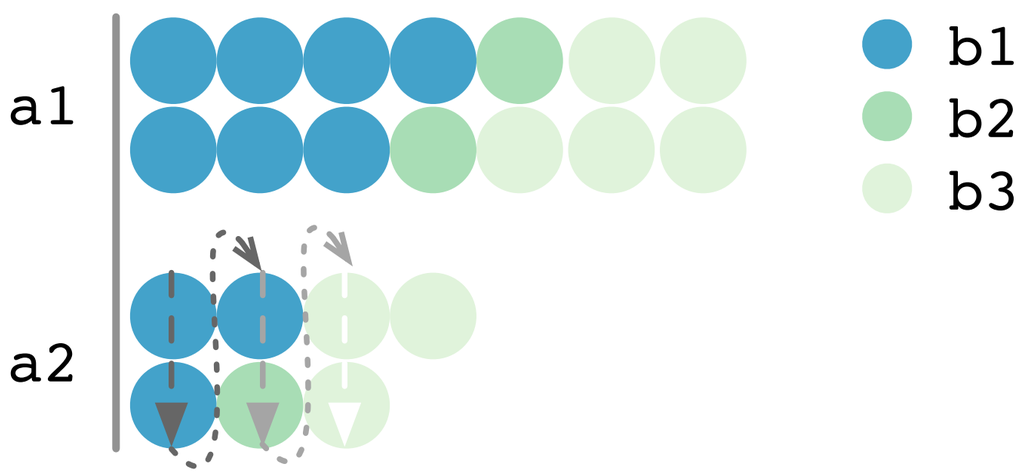
Given

1. visualization shows incorrect *probability distribution*
2. specifying *probability distribution* is convoluted

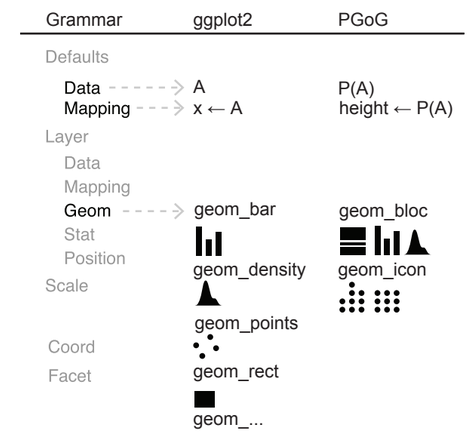
A Probabilistic Grammar of Graphics

- A high-level visualization grammar
- Makes probability distributions first-class citizens (thus solving the two problems)
- Covers a meaningful set of probabilistic visualizations

aes		P(cyl, gear|am)	marg	cond
x.cond	←	P(1|am)	1	am
x.height	←	P(gear|am)	gear	am
height	←	P(cyl|gear,am)	cyl	gear,am

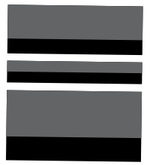



```
    y <- A
width <- P(A) P(B|A)
 fill <- B
```


Grammar
ggplot2
PGoG
Defaults
Data
A
P(A)
Mapping
x ← A
height ← P(A)
Layer
Data
Mapping
Geom
geom_bar
geom_bloc
Stat
Position
geom_density
geom_icon
Scale
geom_points
Coord
geom_rect
Facet
geom_...

Existing ggplot2 packages

Changes



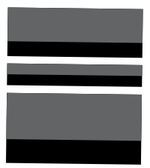
```
geom:mosaic→bar
+fill +y
-divider
```

Syntax

```
geom_mosaic
      x = cyl,
          mpg*
divider = hspine,
          hspine
```

Probabilistic Grammar of Graphics (PGoG)

Changes

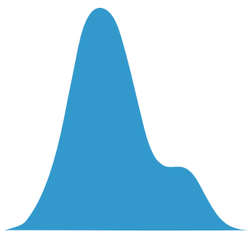
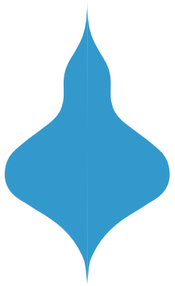
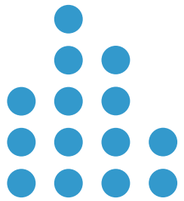
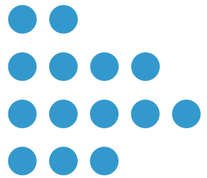
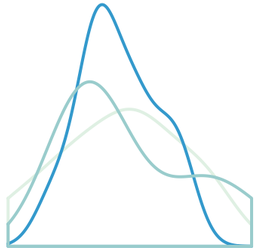
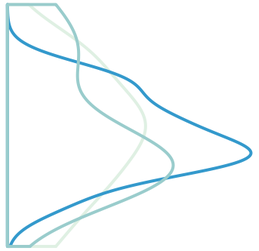
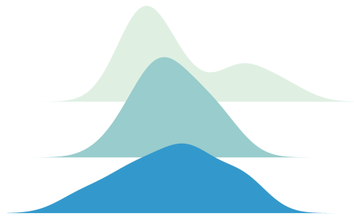
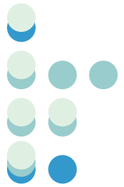
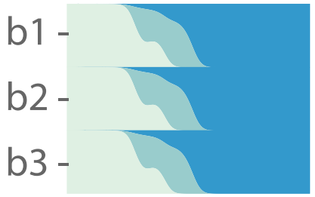
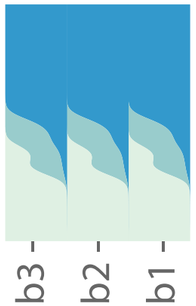
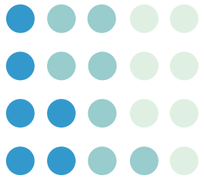
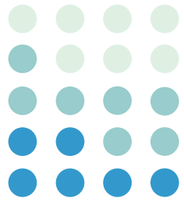


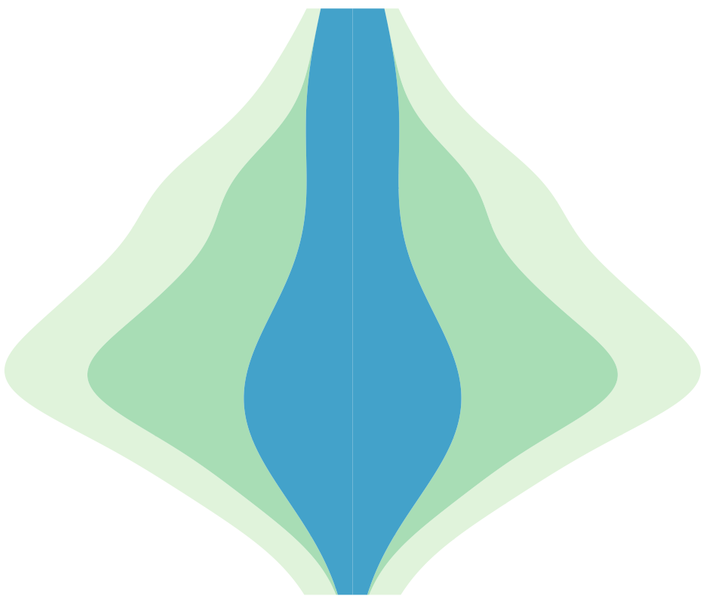
```
+x
```

Syntax

```
geom_bloc
  h <- P(mpg*)
       P(cyl|mpg*)
fill<- cyl
```

`mpg*`: discretized miles per gallon

	geom_bloc				ge	
	x ← A h ← ...	y ← A w ← ...	x ← A y ← B h ← ...	y ← A x ← B w ← ...	x ← A h ← ...	y ← A w ← ...
P(A)	 density plot		NA	NA	 dotplot	
P(A|B)			 ridge plot	 violin plot		
P(B|A) (B|A, C)			 	 		



Onion plot

```
geom_bloc:
        y ← mpg
    width ← P(mpg) P(cy
direction ← both
```

Data	+	Aesthetics	+	Geometry	+ ... =	A plot

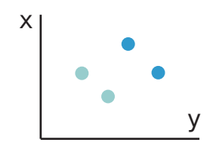
A	B	C
1	2	a
2	1	a
3	4	b
4	2	b

Tidy format, 3 columns

x	y	color
1	2	a
2	1	a
3	4	b
4	2	b

```
x <- A
y <- B
color <- C
```

Point;
Needs x, y aesthetics



a
b

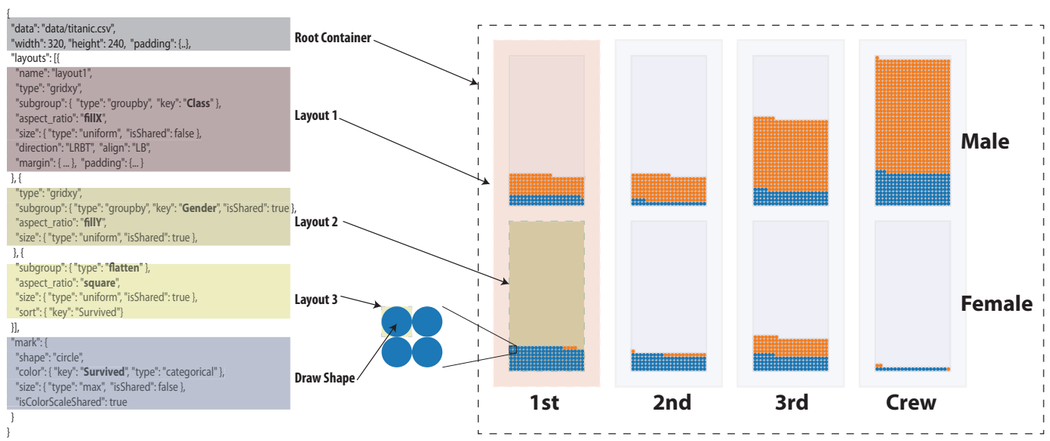
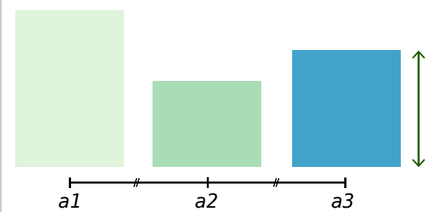
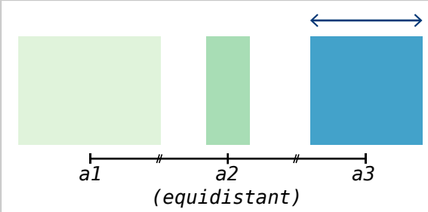
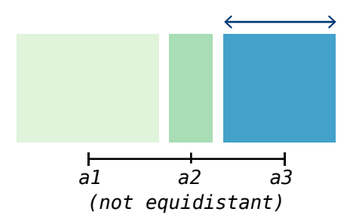


Fig. 6. Example grammar to generate a unit column chart for survivors of the Titanic by passenger class.

		PGoG	Product Plot
Normal bar		`x <- A` `height <- P(A)`	`~ A` `hbar`
"Lying down"		`x <- A` `width <- P(A)`	Does not exist
Spine plot		`width <- P(A)`	`~ A` `hspine`