

Department of Statistics

ST 310: Machine Learning (for Data Science)

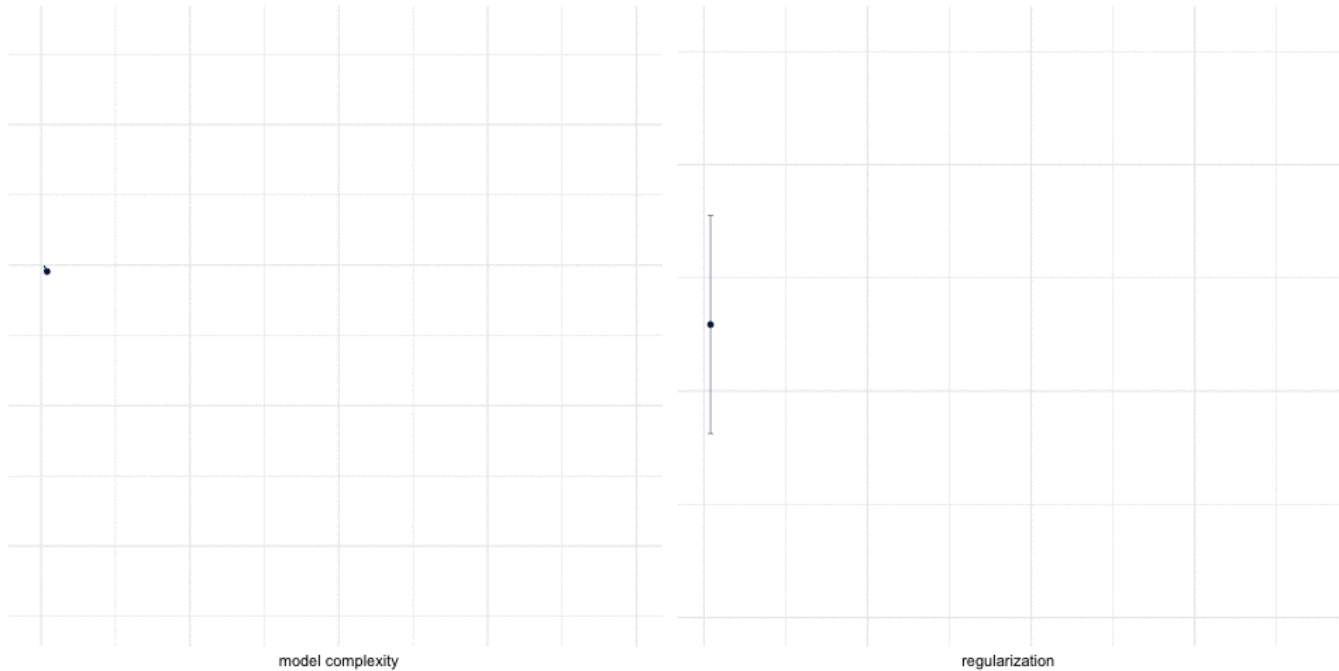
Lecturer: [Joshua Loftus](#)

Teachers: [Phil Chan](#), [Shakeel Gavioli-Akilagun](#), [Domenico Mergoni](#)

Website: ml4ds.com

About the course

- Course info
- Quick preview
- Teaching/course philosophy



Course info

Format

- *Mostly* self-contained -- ask if you need help!
- Seminars: pre-work, Zoom links on Moodle
- Course page <http://ml4ds.com/> (links, slides, misc. notes)
- Participation, active learning
- Assigned weekly readings, attendance of lectures and seminars
- Readings / supplemental references
 - **ISLR** Introduction to Statistical Learning
 - **ESL** Elements of Statistical Learning
 - **CASI** Computer Age Statistical Inference
 - **Mixtape** Causal Inference: The Mixtape
 - **R4DS** R for Data Science

Assessments

As described in the [course listing](#)

- Problem sets: 4 total, 2 are summative
- Summative
 - Problem sets (30%)
 - Individual project (40%) in LT
 - Group project (30%) in LT

Quick preview

Don't worry about following all the details now

We'll introduce R coding gradually

Data	Code	Plot	Modify	Plot again
------	------	------	--------	------------

```
library(gapminder)
gapminder %>% head() %>% kable()
```

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134

Data

Code

Plot

Modify

Plot again

```
gdp_data <- gapminder %>%  
  filter(year == max(year))  
  
life_exp_plot <-  
  ggplot(gdp_data, aes(x = gdpPercap, y = lifeExp)) +  
  geom_point(aes(color = continent,  
                 shape = continent,  
                 size = pop))  
  
life_exp_plot +  
  stat_smooth(formula = y ~ x, method = "loess", span = 1)
```

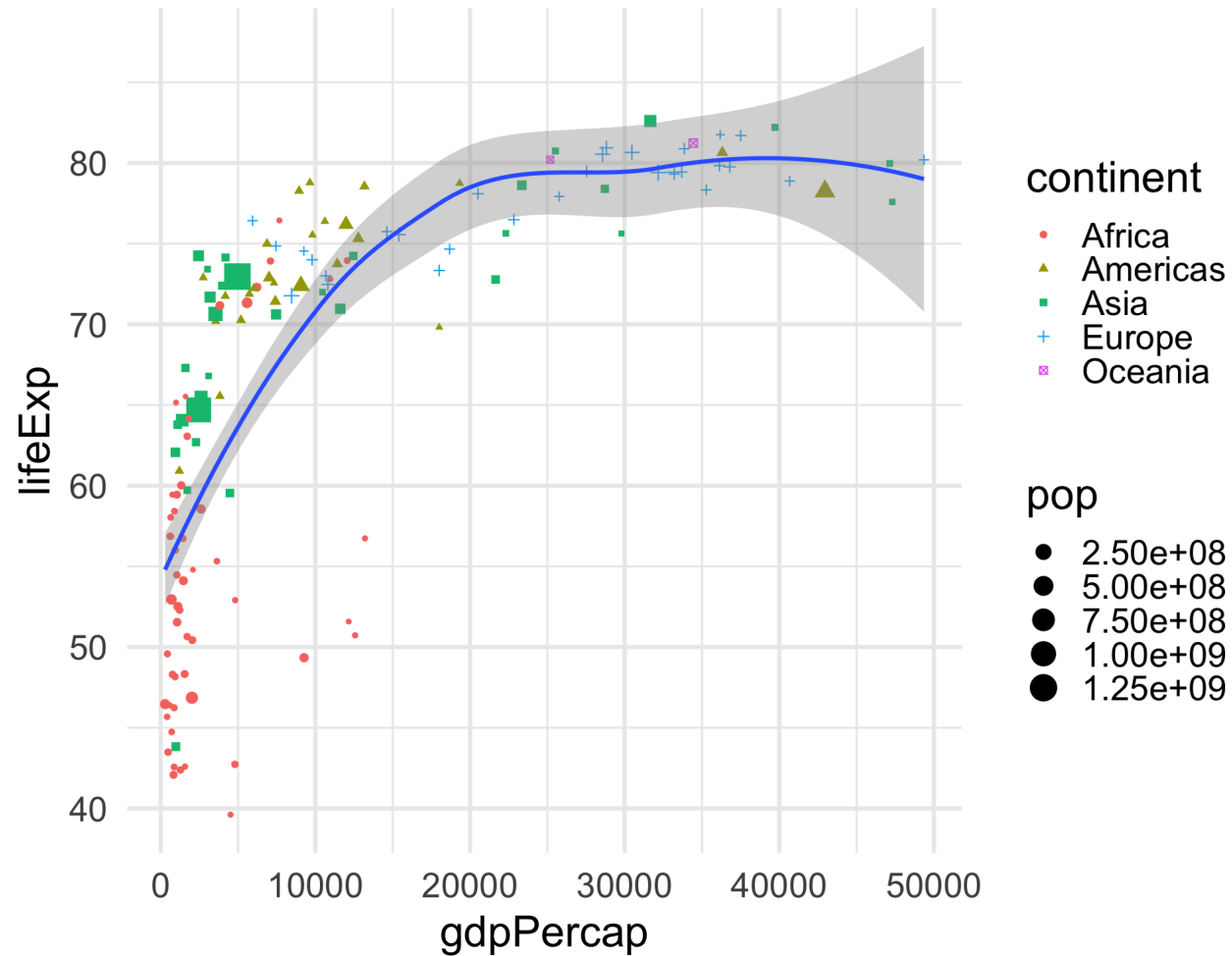

Data

Code

Plot

Modify

Plot again



Data

Code

Plot

Modify

Plot again

```
life_exp_plot +  
  scale_x_log10() +  
  stat_smooth(formula = y ~ x, method = "lm") +  
  xlab("GDP per capita") +  
  ylab("Life expectancy")
```

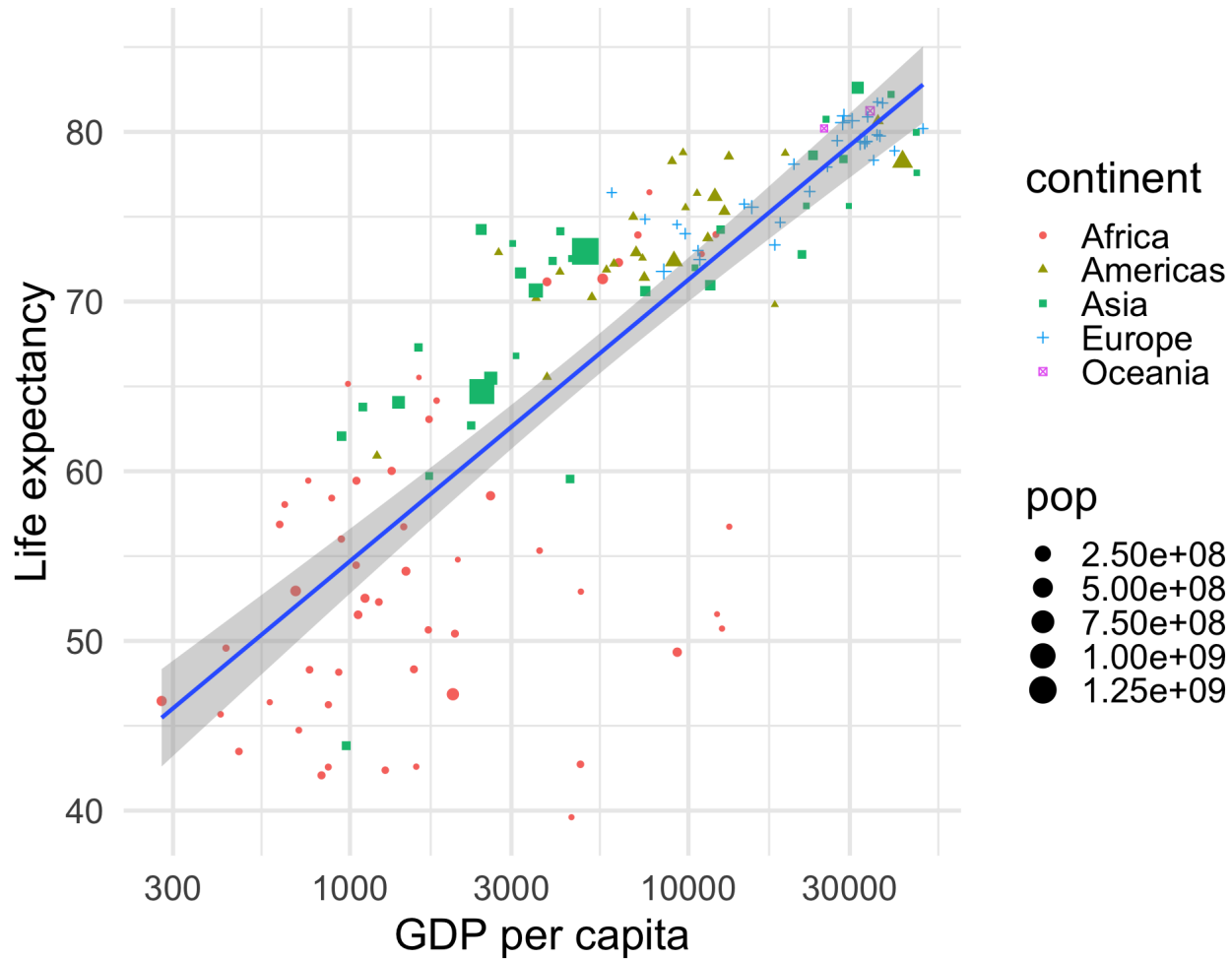
Data

Code

Plot

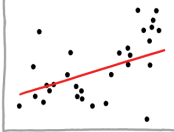
Modify

Plot again



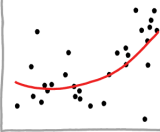
CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



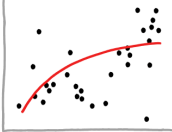
"HEY, I DID A
REGRESSION!"

QUADRATIC



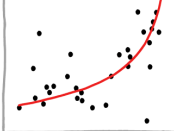
"I WANTED A CURVED
LINE, SO I MADE ONE
WITH MATH."

LOGARITHMIC



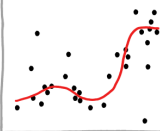
"LOOK, IT'S
TAPERING OFF!"

EXPONENTIAL



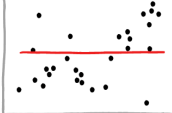
"LOOK, IT'S GROWING
UNCONTROLLABLY!"

LOESS



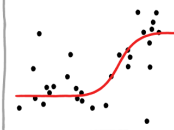
"I'M SOPHISTICATED, NOT
LIKE THOSE BUMBLING
POLYNOMIAL PEOPLE."

LINEAR,
NO SLOPE



"I'M MAKING A
SCATTER PLOT BUT
I DON'T WANT TO."

LOGISTIC



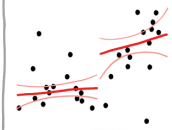
"I NEED TO CONNECT THESE
TWO LINES, BUT MY FIRST IDEA
DIDN'T HAVE ENOUGH MATH."

CONFIDENCE
INTERVAL



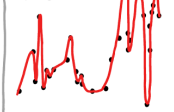
"LISTEN, SCIENCE IS HARD.
BUT I'M A SERIOUS
PERSON DOING MY BEST."

PIECEWISE



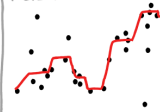
"I HAVE A THEORY,
AND THIS IS THE ONLY
DATA I COULD FIND."

CONNECTING
LINES



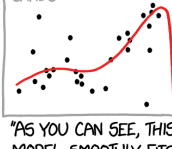
"I CLICKED 'SMOOTH
LINES' IN EXCEL."

AD-HOC
FILTER



"I HAD AN IDEA FOR HOW
TO CLEAN UP THE DATA.
WHAT DO YOU THINK?"

HOUSE OF
CARDS



"AS YOU CAN SEE, THIS
MODEL SMOOTHLY FITS
THE— WAIT NO NO DON'T
EXTEND IT AAAAAA!!!"

Yes, there will be memes,
videos, etc...

Source: [xkcd](#)

Teaching/course philosophy

Recurring themes

- Human-centric ML
 - Tools for us to control, not conversely...
 - Ethics of data science
- Interpretability
 - Philosophy of science
 - Causality vs "curve fitting"
- Social learning
 - Come post on the forum!
 - (More on this later)