



From PhD to Industry

Data Science in action

Sarah Stolle

04-12-18

Stereotype of programmers

pro·gram·mer

[**proh**-gram-er], *n*,

1. an organism that converts caffeine and pizza into software, usually late at night. 2. a human that has a deeper and more meaningful relationship with a computer than with other humans.

DATA

Data Scientist: The Sexiest Job of the 21st Century

by **Thomas H. Davenport** and **D.J. Patil**

FROM THE OCTOBER 2012 ISSUE

It's never too late

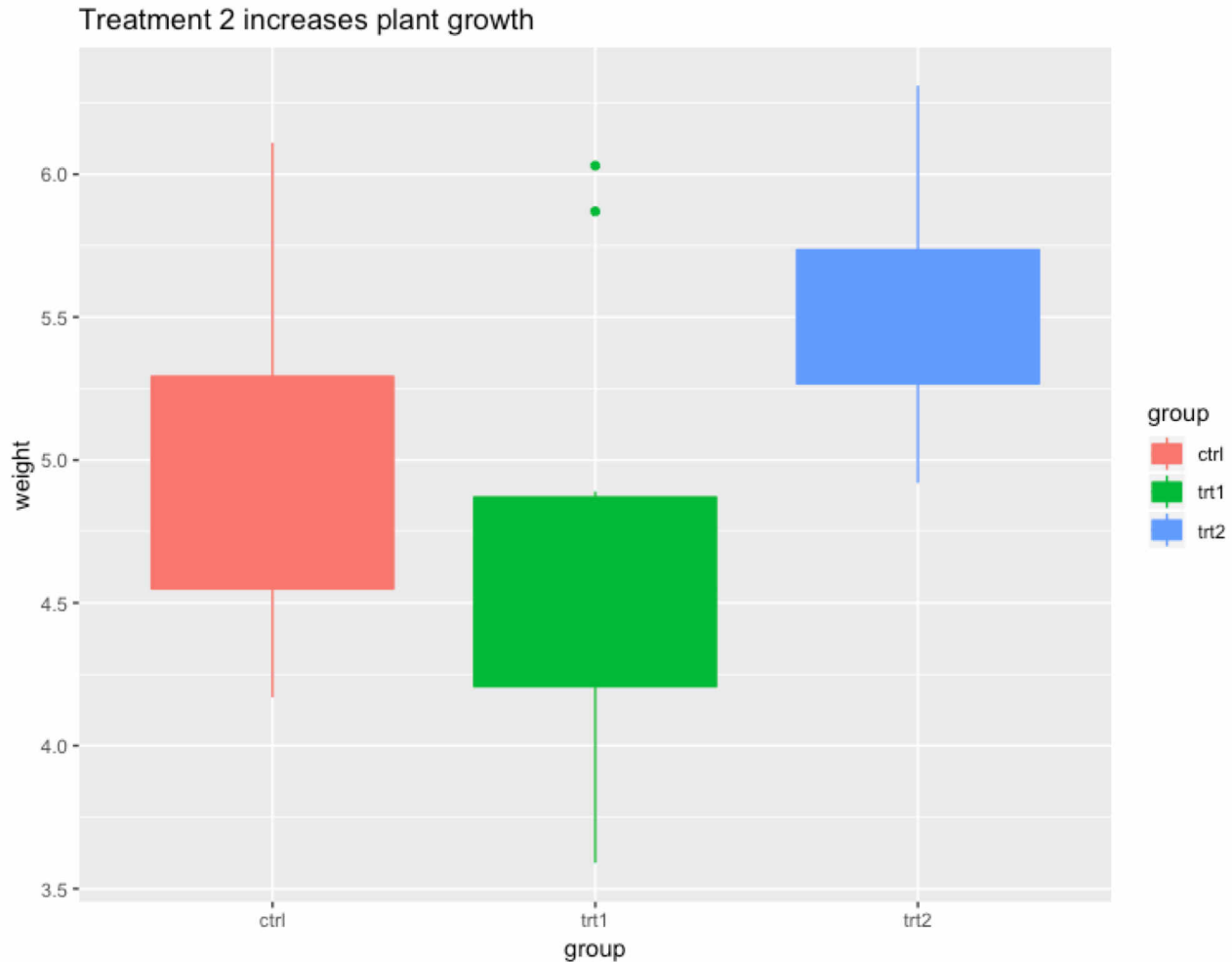
to start coding!

Workflow in R

```
# load data
pg <- PlantGrowth

# process data
pg_bar <- pg %>%
  group_by(group) %>%
  summarise(avg_weight = mean(weight))

# create plot
ggplot(pg_bar, aes(y=avg_weight, x=group, color=group,
  geom_bar(stat="identity", position = "dodge") +
  ggtitle("Treatment 2 increases plant growth")
```



Benefits of coding


Reproducibility

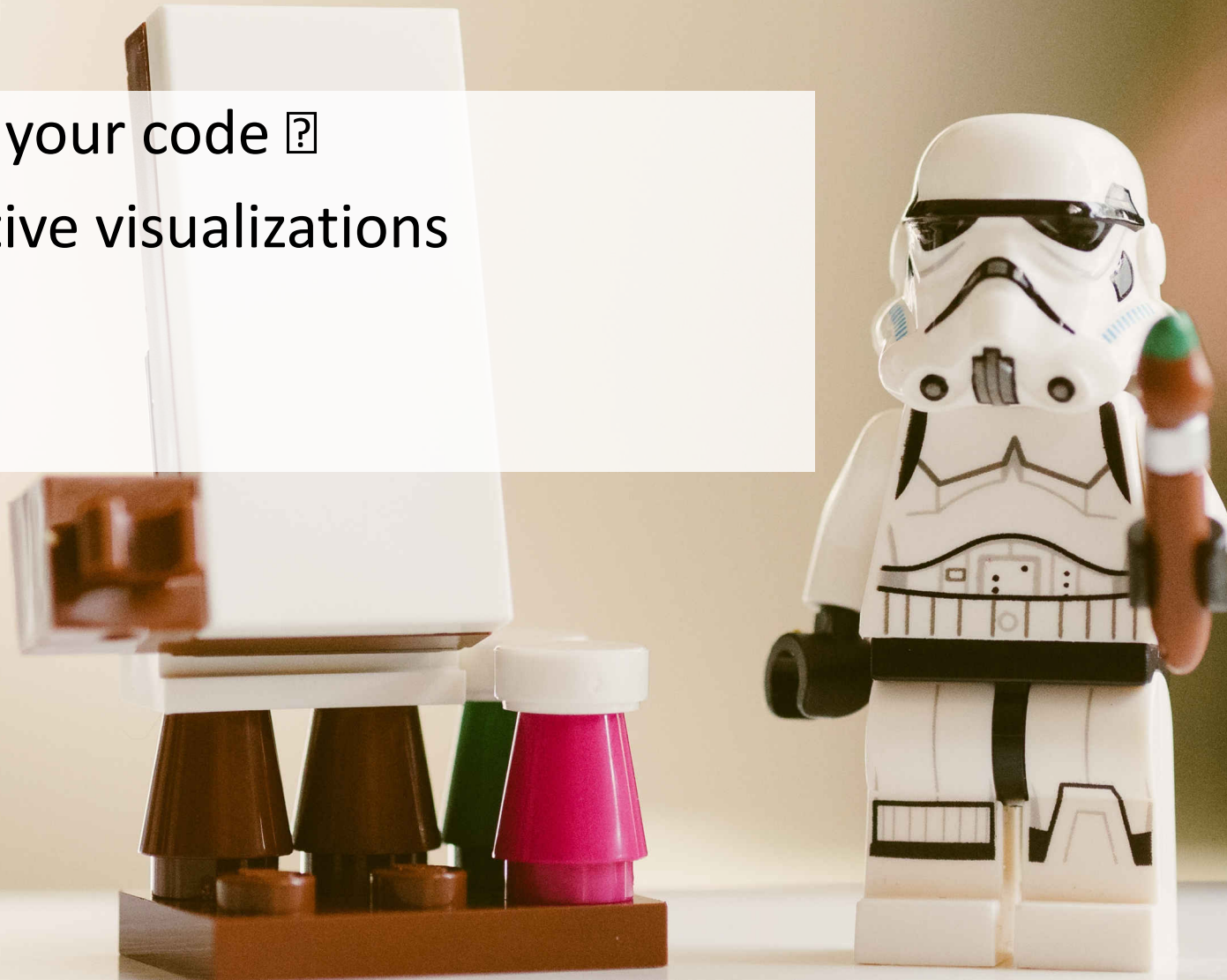
Documentation


Less copy-paste errors

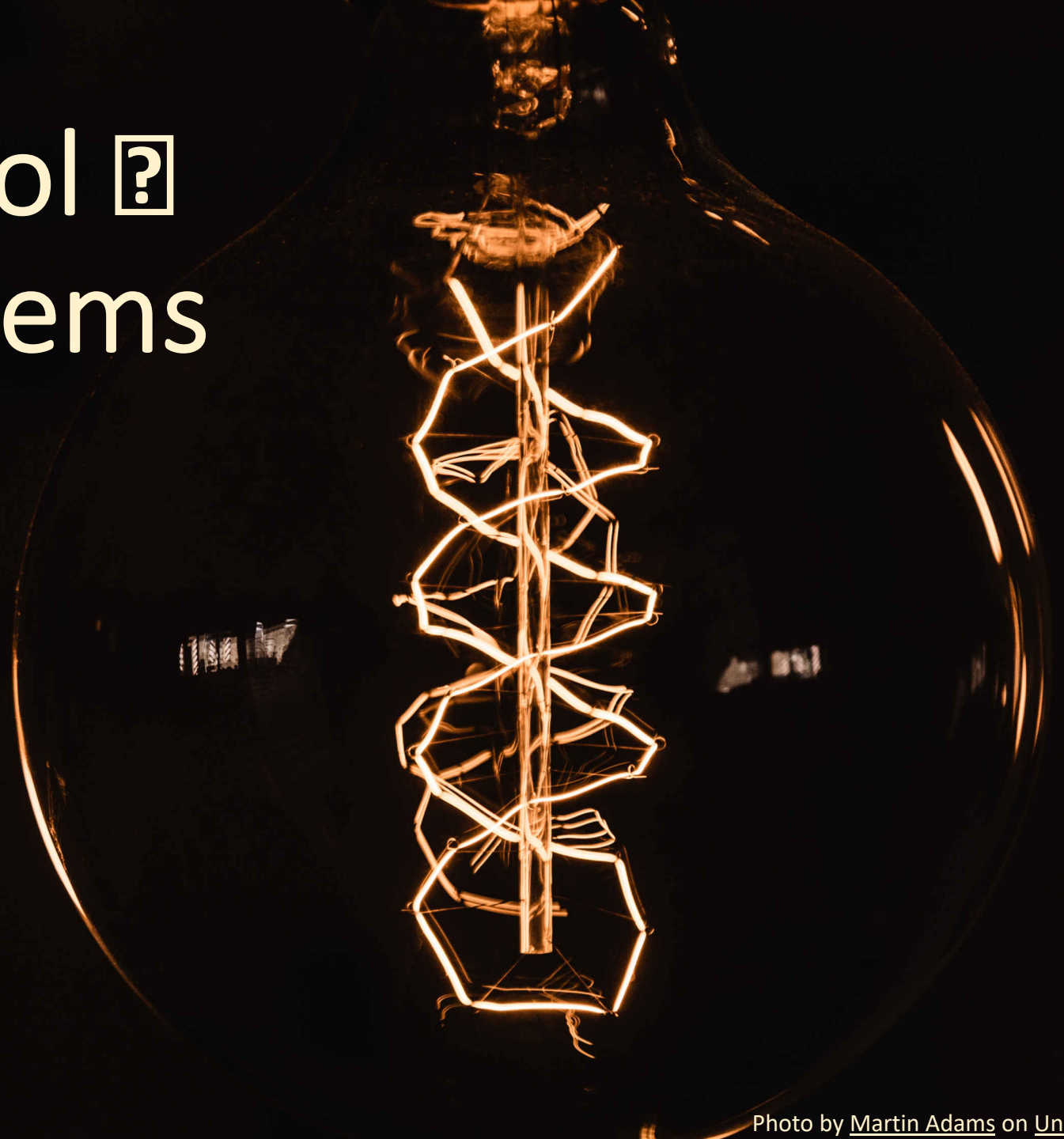
Big data analysis

Communicate your discoveries

- Publish your code 
- Interactive visualizations



Coding is a tool 
to solve problems



Literacy in data analysis is becoming a basic requirement

- Data-driven insights
- Creating value

R vs Python

- Coding principles are similar
- Functional vs OOP
- R is popular in biological field
- Python is popular in industry

Resources

- The pragmatic programmer

<https://pragprog.com/book/tpp/the-pragmatic-programmer>

Resources

- R tutorials and webinars
<https://cran.r-project.org/doc/manuals/R-intro.html>
- R on twitter (large community!) (e.g. [@hadleywickham](#), [@dataandme](#) & [@drob](#))
- R 4 data science book & community
<https://r4ds.had.co.nz/>
- Tidyverse (R packages for data wrangling)

Python Resources

- Jupyter notebook gallery
<https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks>
- Pydata (conference talks on YouTube)
- Most important packages for data analysis:
 - Numpy
 - Pandas
 - Jupyter notebooks
 - Matplotlib/ seaborn/ plotnine/ plotly (visualizations)
 - Sklearn / statsmodels (machine learning or modelling)