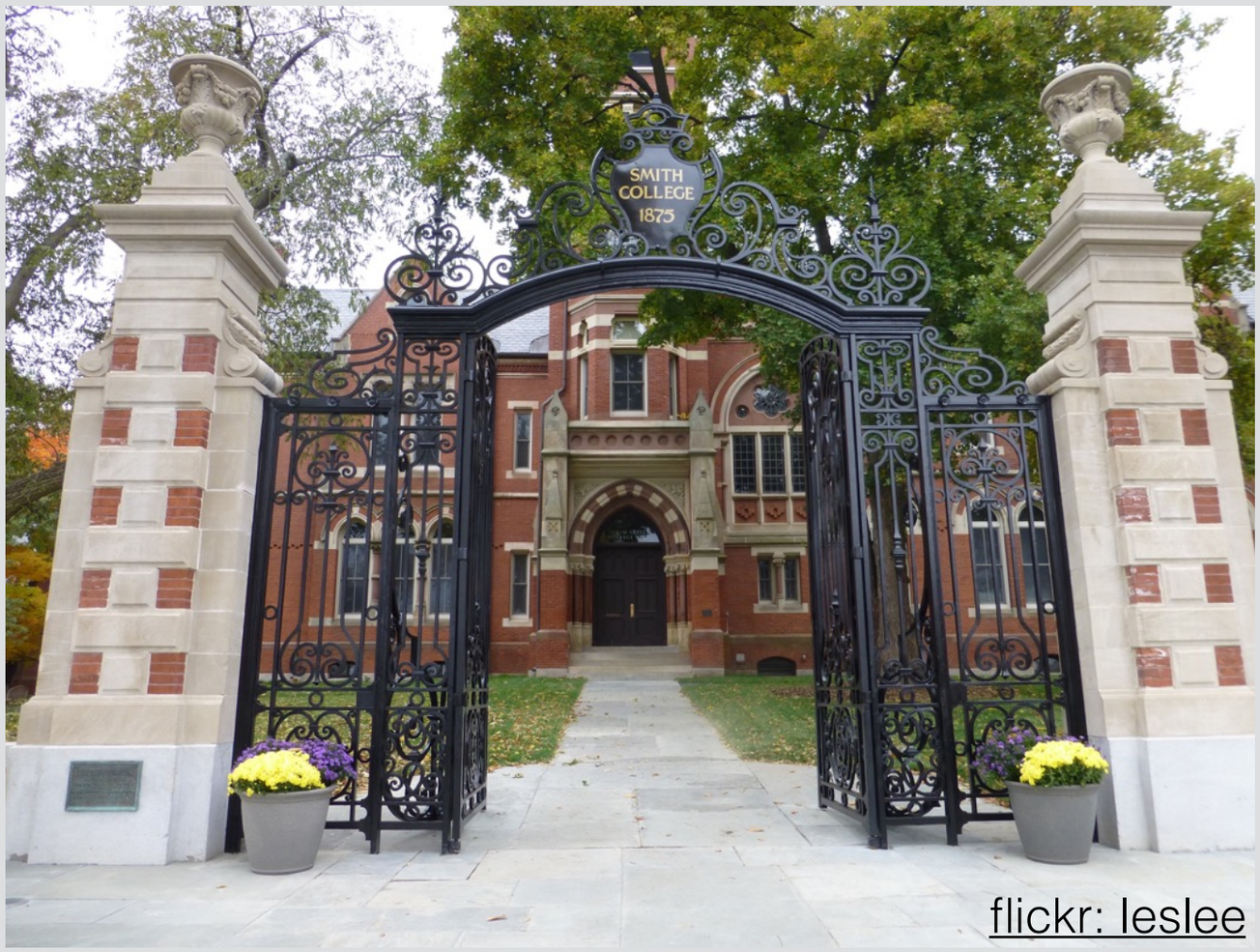
Data Science for Statisticians

Amelia McNamara @AmeliaMN

About me

- BA in Mathematics and English from Macalester College
- PhD in Statistics from UCLA
- Visiting Assistant Professor of Statistical and Data Sciences at Smith College
- Research at the intersection of statistical computing, statistics education, data visualization

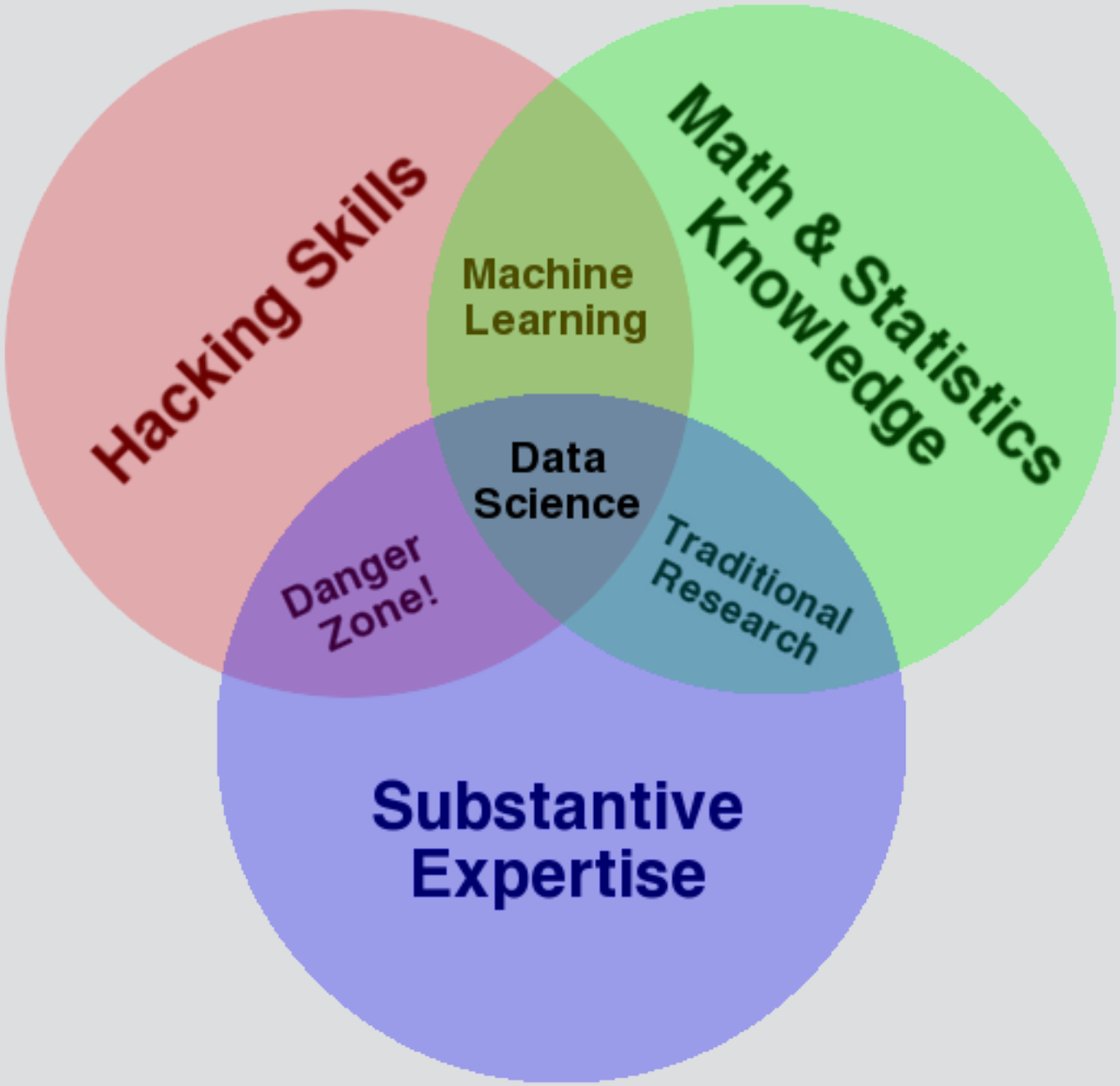


flickr: leslee

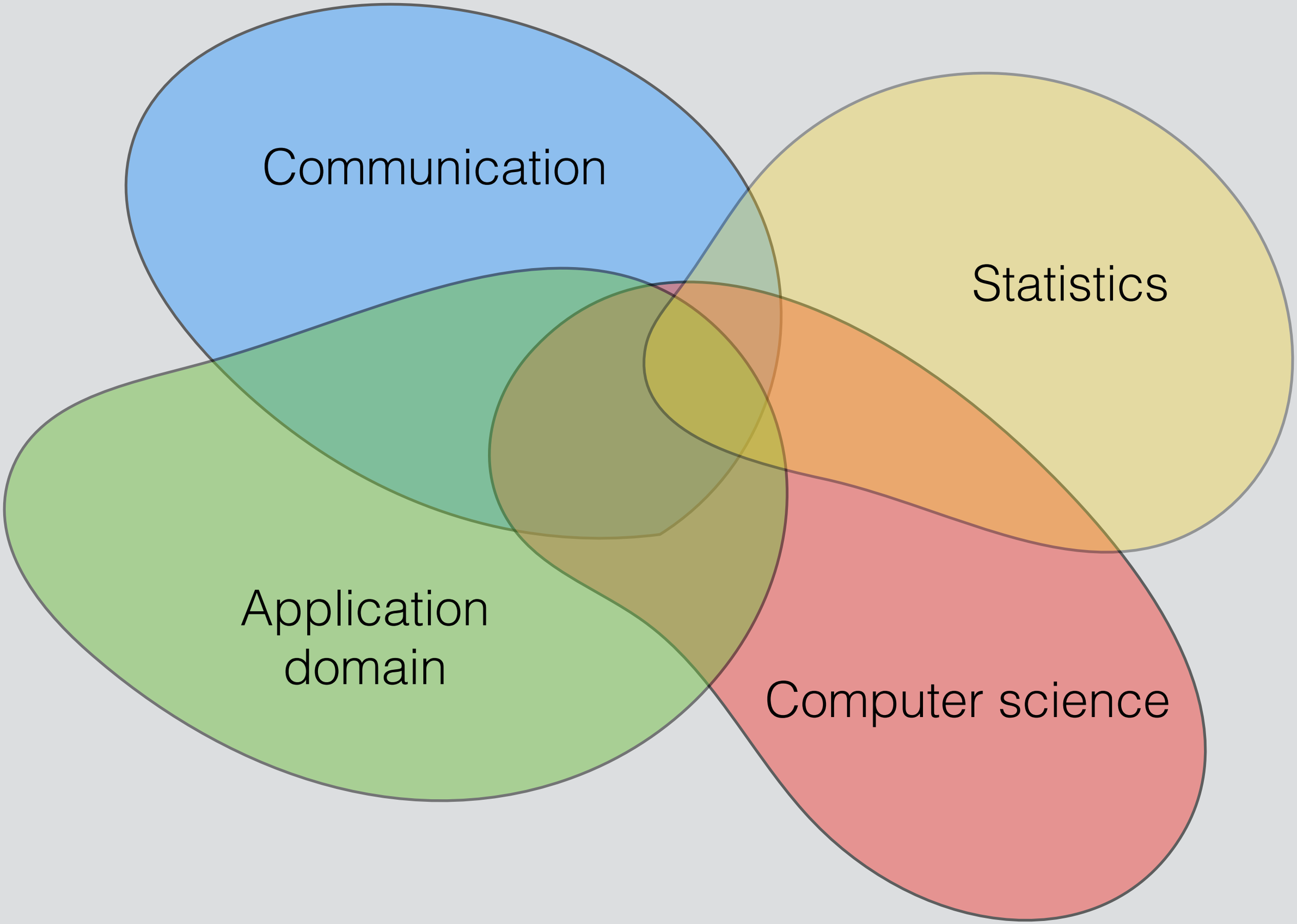
What is data science?

“Data science is statistics on a Mac”

-@BigDataBorat

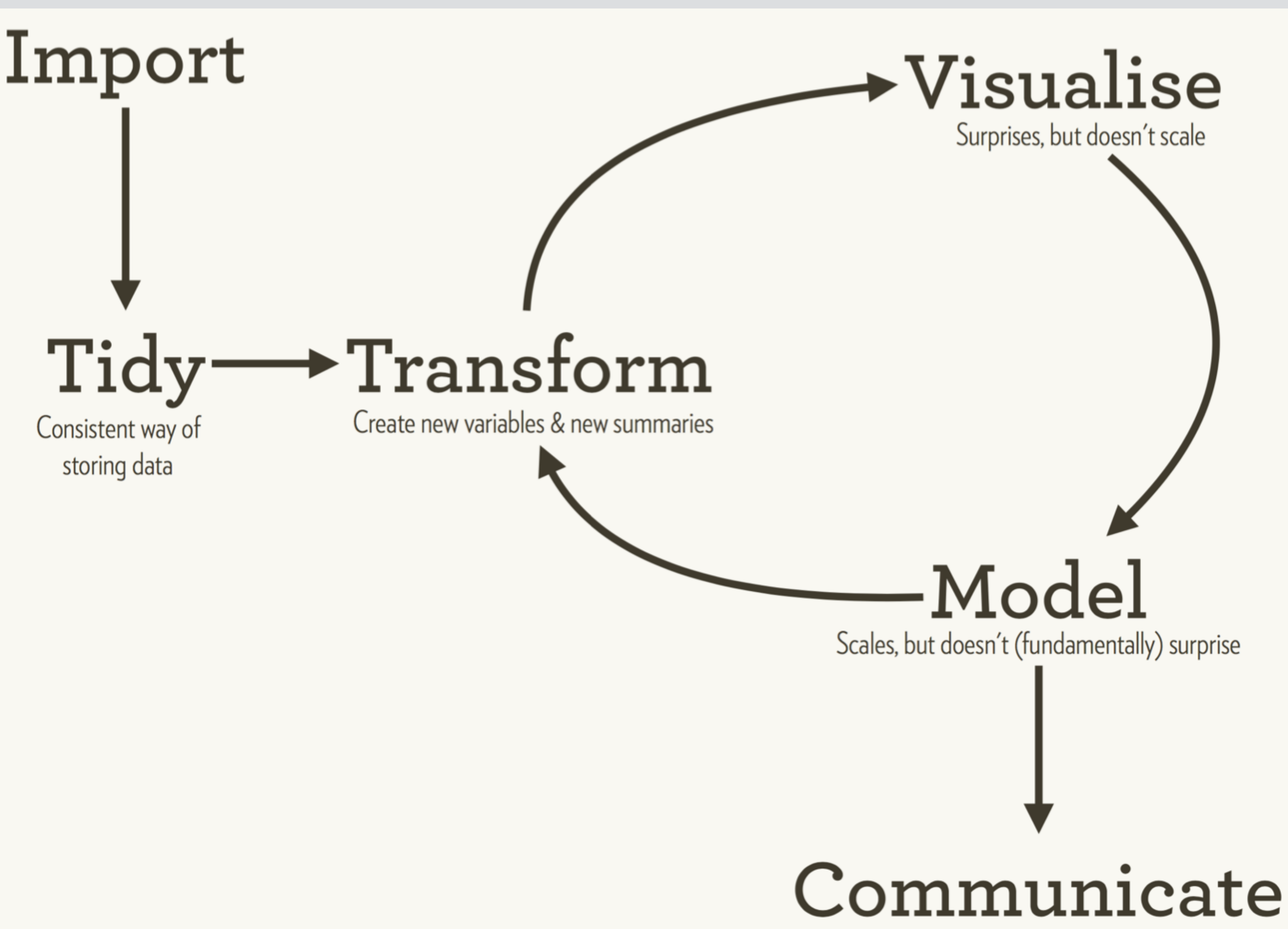

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

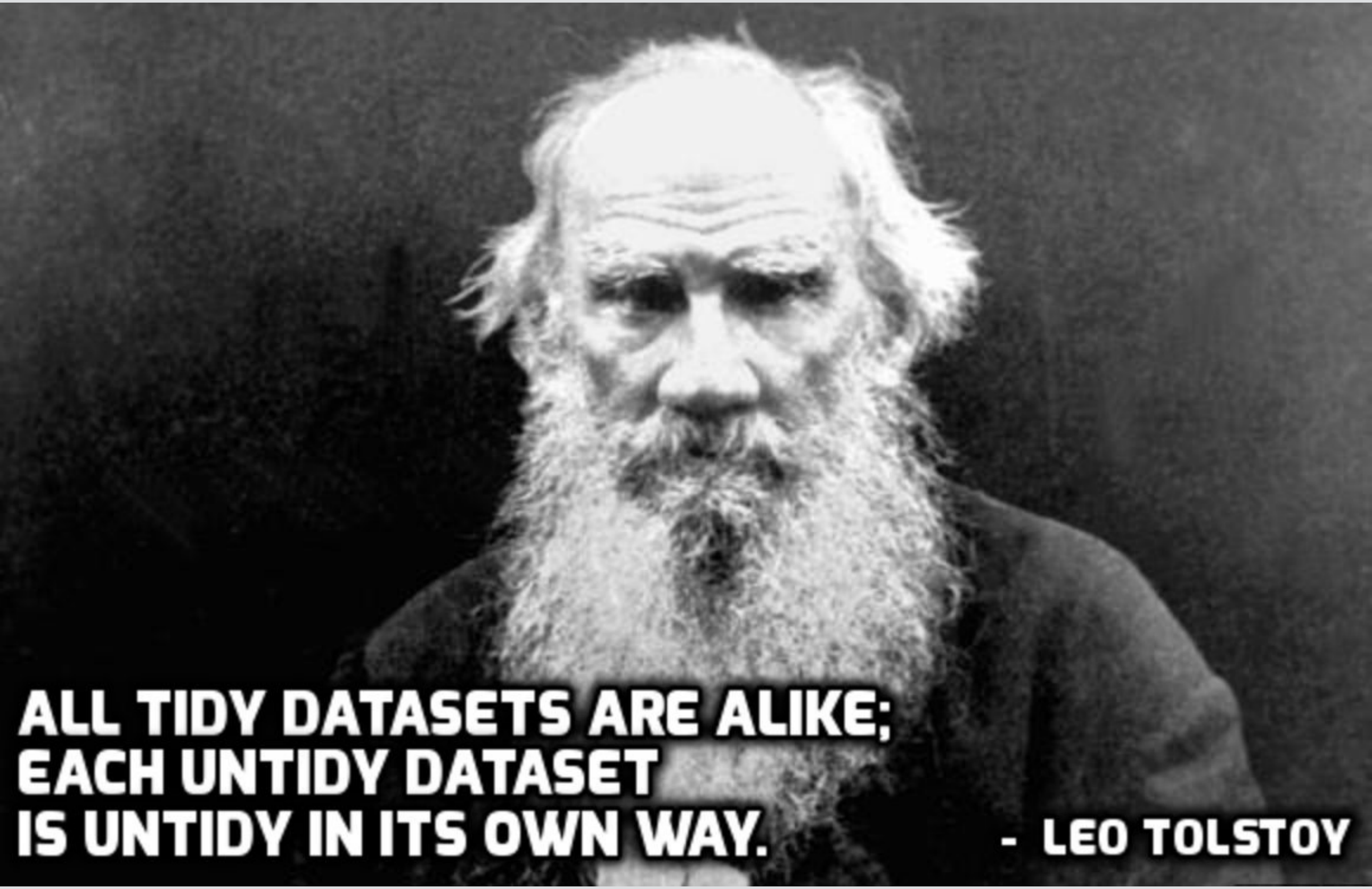
Drew Conway



Smith SDS major

Hadley Wickham


Import
Visualise
Surprises, but doesn't scale
Tidy
Consistent way of storing data
Transform
Create new variables & new summaries
Model
Scales, but doesn't (fundamentally) surprise
Communicate


ALL TIDY DATASETS ARE ALIKE;
EACH UNTIDY DATASET
IS UNTIDY IN ITS OWN WAY.
- LEO TOLSTOY

Sean Kross

un-tidy data

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

tidy data

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Today's (tentative) schedule:

8:00-9:00	Introduction to R and the tidyverse
9:00-10:15	More dplyr and ggplot2
10:15-10:30	Coffee break
10:30-12:00	Reproducible research and version control
12:00-1:00	Lunch
1:00-2:00	APIs and web scraping
2:00-3:15	Interactivity with shiny, manipulate and leaflet
3:15-3:30	Coffee break
3:30-4:30	Learning more, finding help, becoming a data scientist
4:30-5:00	Wrap up