What is Data Science?

Some answers:

Some answers:

» "A data scientist is a statistician who lives in San Francisco"

Some answers:

» "A data scientist is a statistician who lives in San Francisco"

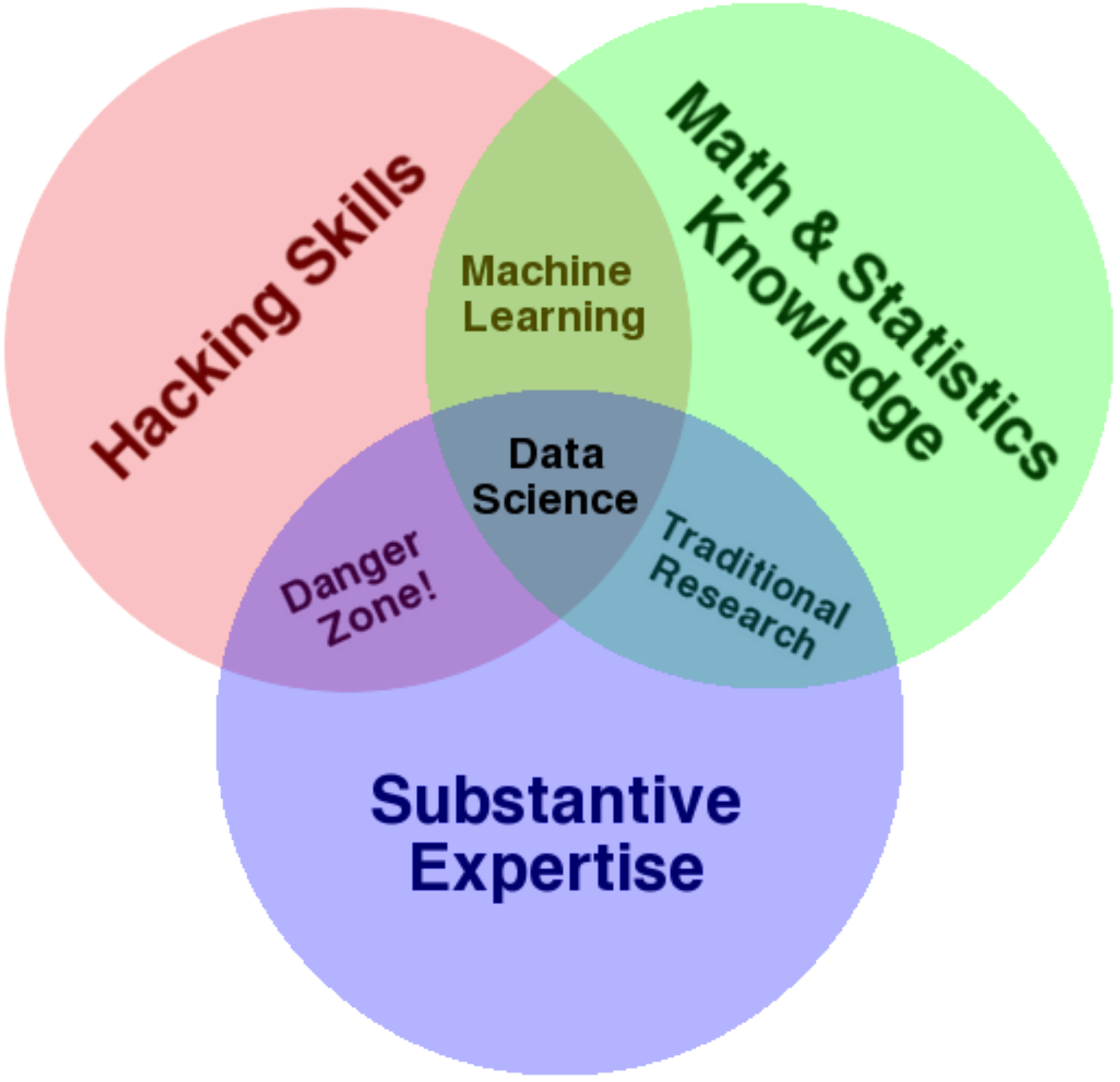
» "Data Science is statistics on a Mac."

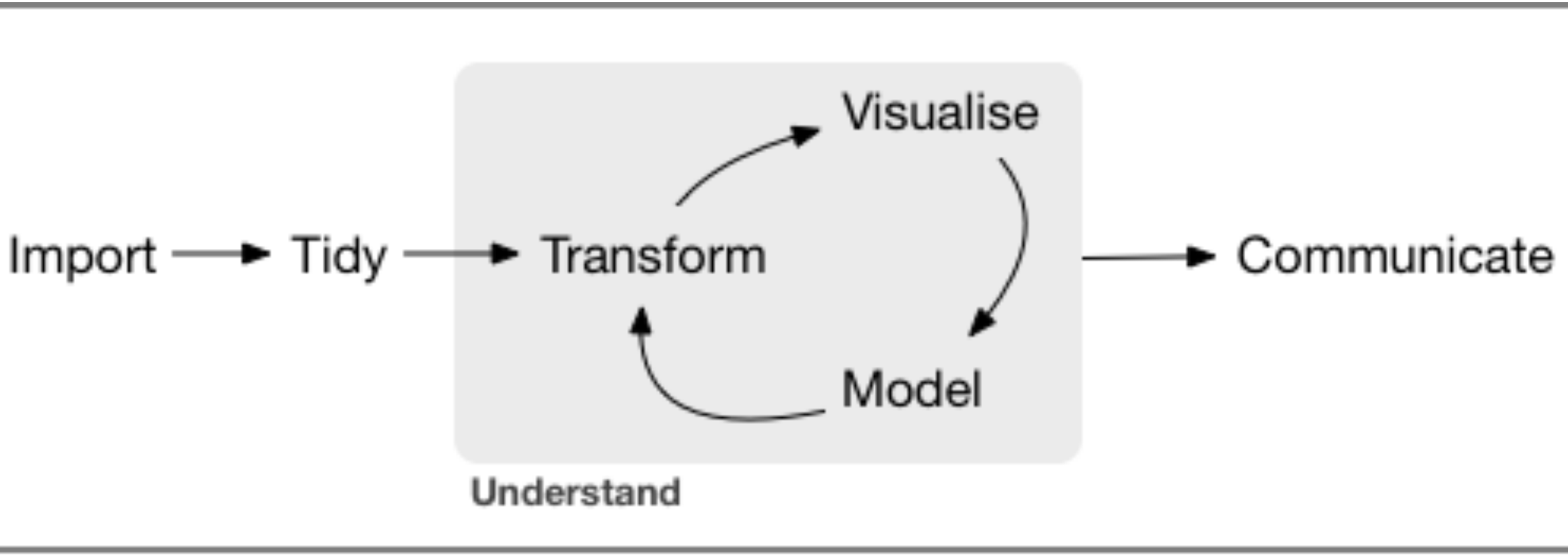
Some answers:

» "A data scientist is a statistician who lives in San Francisco"

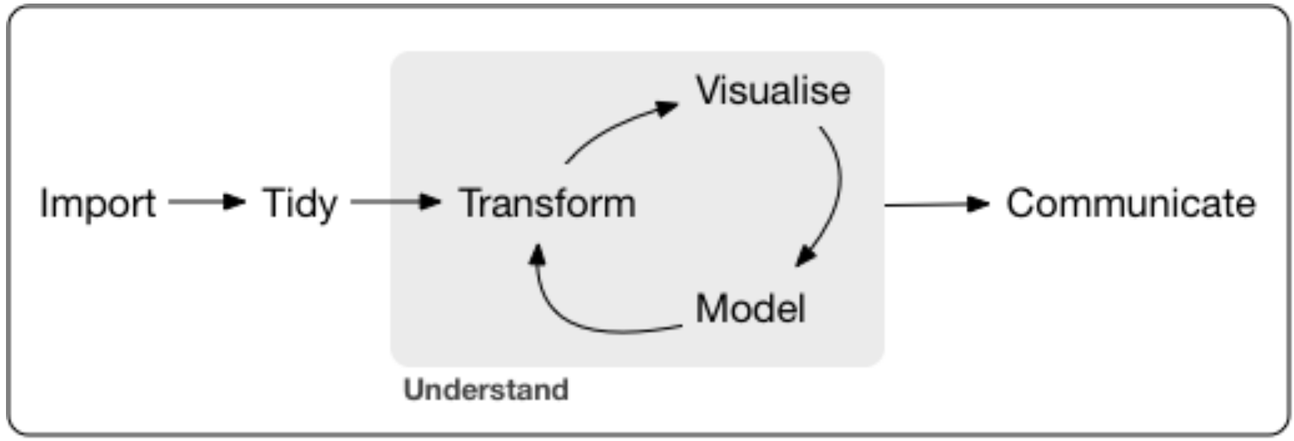
» "Data Science is statistics on a Mac."

» "A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."


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

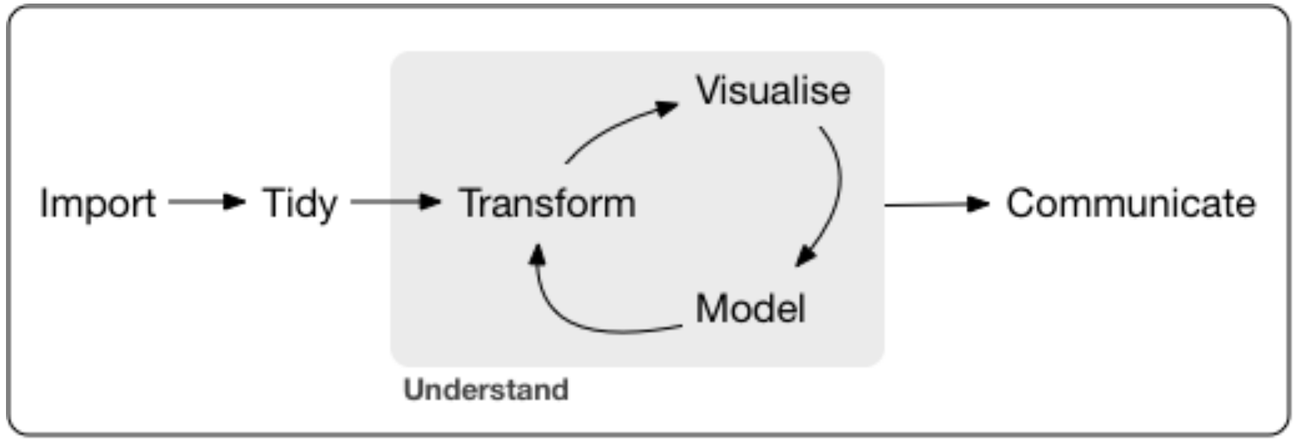

Visualise
Import
Tidy
Transform
Communicate
Model
Understand

This course


Import
Tidy
Transform
Visualise
Model
Communicate
Understand

This course

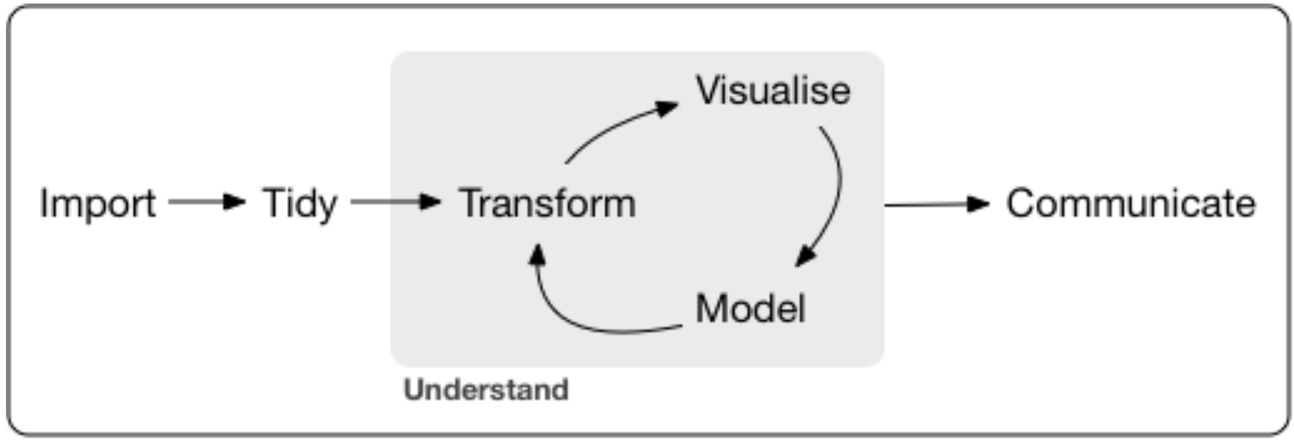
» Statistics courses mostly focus on the central part, and mostly on the model part



This course

» Statistics courses mostly focus on the central part, and mostly on the model part

» The steps on the left can easily take 80% of the time

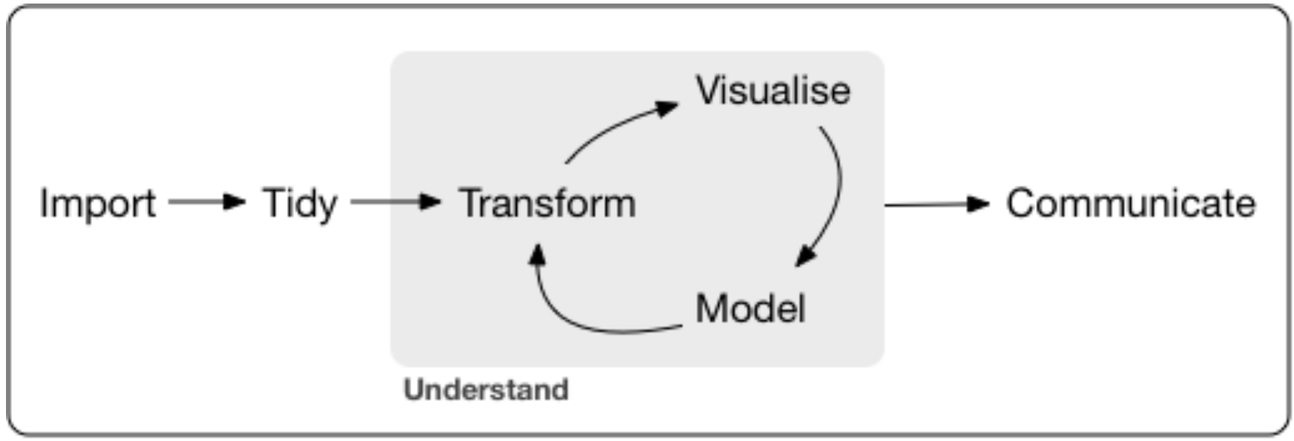


This course

» Statistics courses mostly focus on the central part, and mostly on the model part

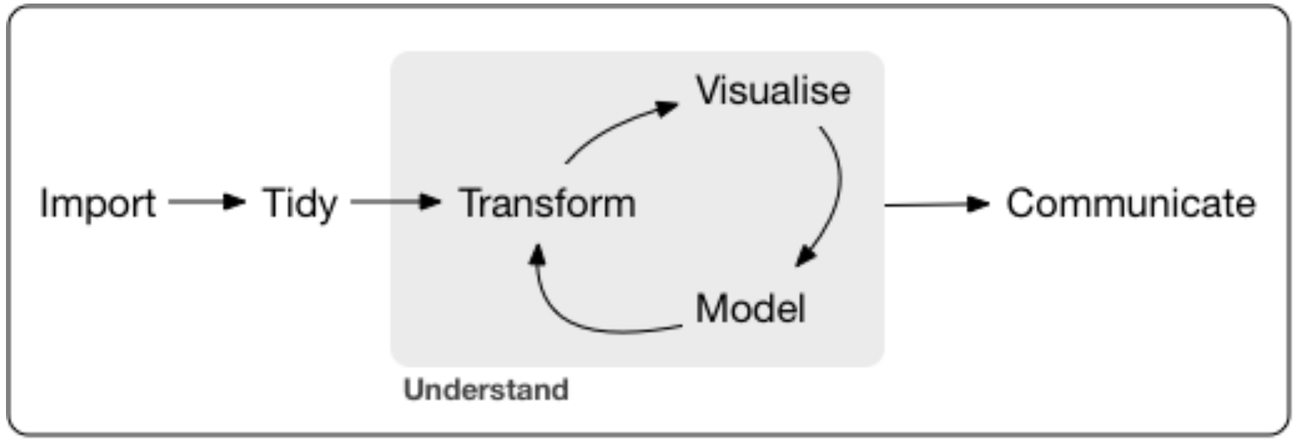
» The steps on the left can easily take 80% of the time

» This course is focused on tools for that 80%



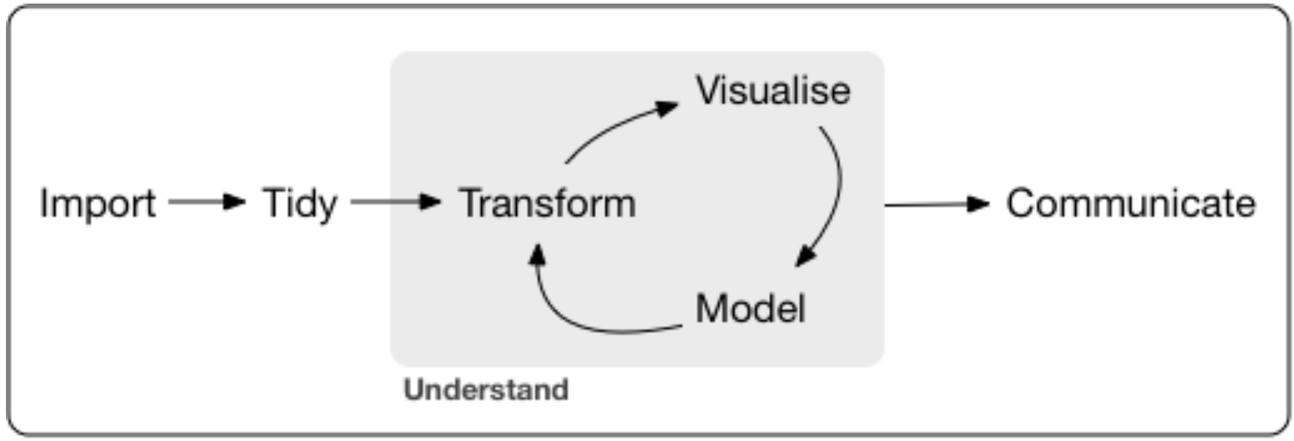
This course

- Statistics courses mostly focus on the central part, and mostly on the model part
- The steps on the left can easily take 80% of the time
- This course is focused on tools for that 80%
- The course is mostly about tools, somewhat about principles, and not at all about statistical theory



This course

- Statistics courses mostly focus on the central part, and mostly on the model part
- The steps on the left can easily take 80% of the time
- This course is focused on tools for that 80%
- The course is mostly about tools, somewhat about principles, and not at all about statistical theory
- (Diagram from Hadley Wickham)



TECHNOLOGY

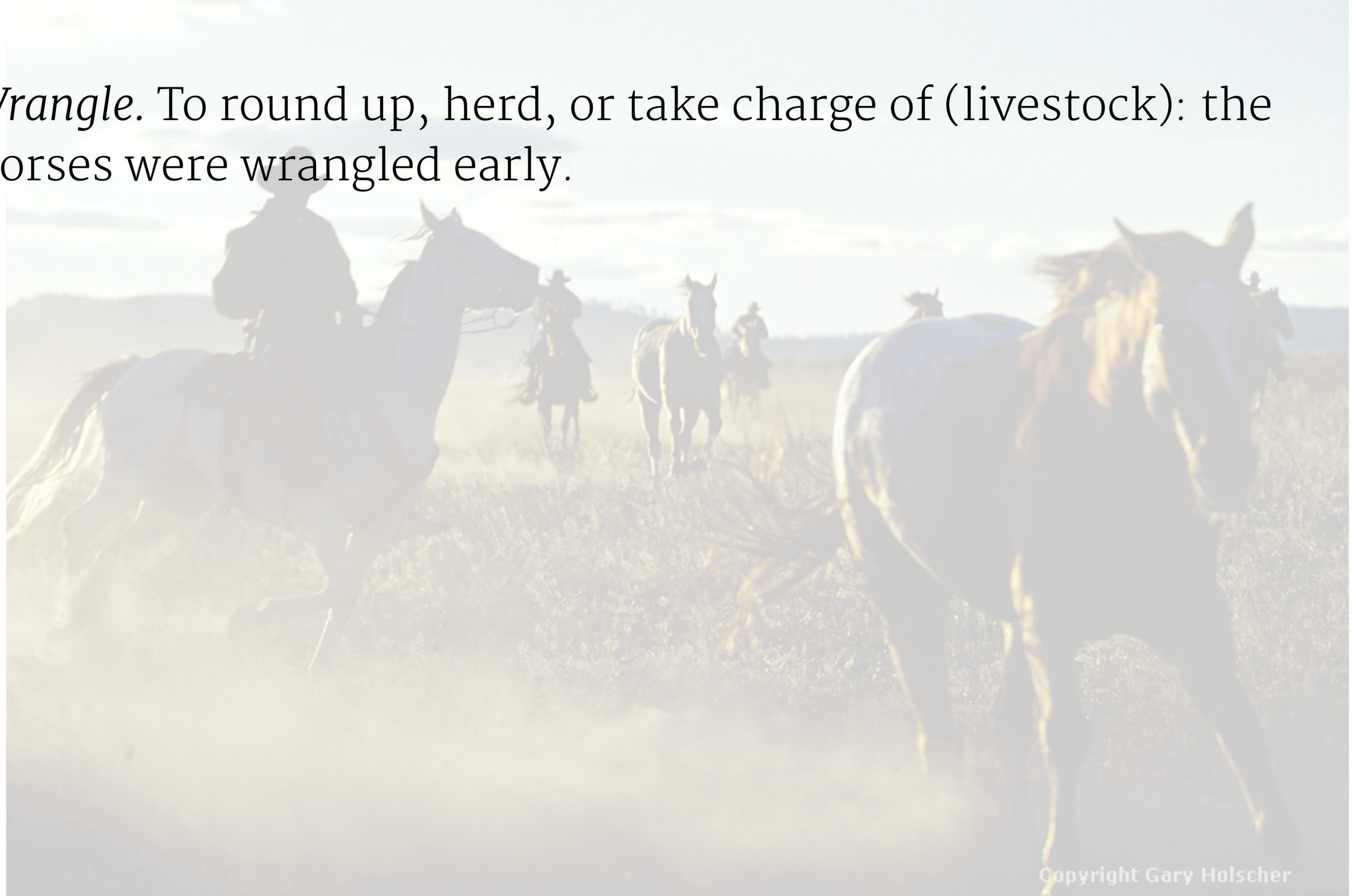
For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

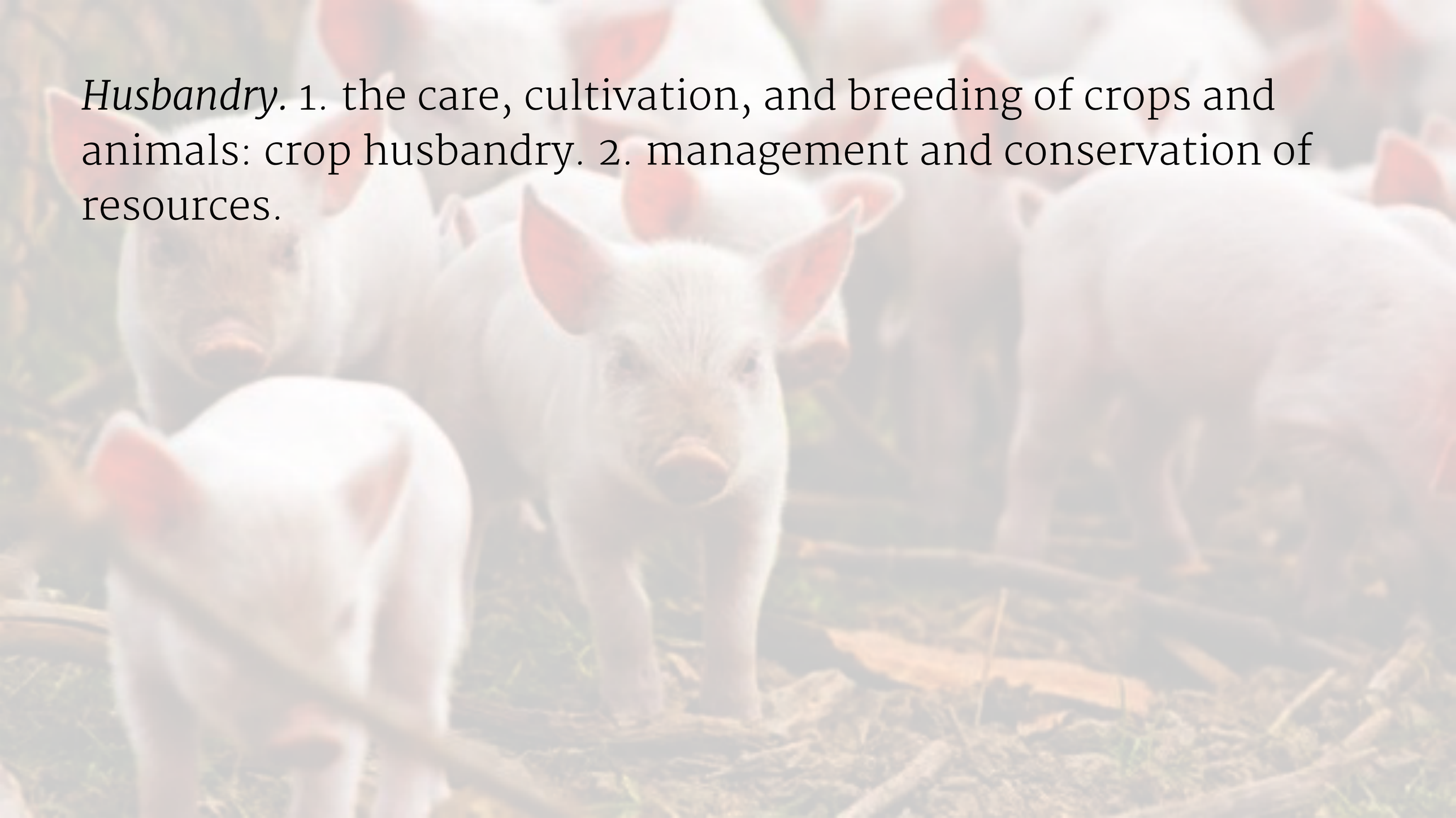
By **STEVE LOHR** AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.

Peter DaSilva for The New York Times

Wrangle. To round up, herd, or take charge of (livestock): the horses were wrangled early.





Husbandry. 1. the care, cultivation, and breeding of crops and animals: crop husbandry. 2. management and conservation of resources.

"What I Am Looking for in a New Data Science Hire"

(from: http://www.datajujitsu.co.uk/blog/2016/09/03/what-i-am-looking-for-in-a-new-data-science-hire/)

"What I Am Looking for in a New Data Science Hire"

» Good conceptual knowledge of maths, statistics and machine learning

(from: http://www.datajujitsu.co.uk/blog/2016/09/03/what-i-am-looking-for-in-a-new-data-science-hire/)

"What I Am Looking for in a New Data Science Hire"

» Good conceptual knowledge of maths, statistics and machine learning

» *Data processing and visualisation expertise*

(from: http://www.datajujitsu.co.uk/blog/2016/09/03/what-i-am-looking-for-in-a-new-data-science-hire/)

"What I Am Looking for in a New Data Science Hire"

» Good conceptual knowledge of maths, statistics and machine learning

» *Data processing and visualisation expertise*

» Scientific method

(from: http://www.datajujitsu.co.uk/blog/2016/09/03/what-i-am-looking-for-in-a-new-data-science-hire/)

"What I Am Looking for in a New Data Science Hire"

» Good conceptual knowledge of maths, statistics and machine learning

» *Data processing and visualisation expertise*

» Scientific method

» An interest in programming

(from: http://www.datajujitsu.co.uk/blog/2016/09/03/what-i-am-looking-for-in-a-new-data-science-hire/)

"What I Am Looking for in a New Data Science Hire"

» Good conceptual knowledge of maths, statistics and machine learning

» *Data processing and visualisation expertise*

» Scientific method

» An interest in programming

» Domain knowledge

(from: http://www.datajujitsu.co.uk/blog/2016/09/03/what-i-am-looking-for-in-a-new-data-science-hire/)

This is perhaps the only absolute essential thing I am looking for and is why ‘pure’ statisticians with experience only in SPSS, Stata or similar are really not likely to make it past an interview. I need people who are data manipulation cyborgs, for whom filtering, subsetting, merging and transforming data is second nature and can do it using their preferred tool (be it dplyr, Pandas, SQL or whatever) with flow, without being hindered by those tools. Think of the difference between a new driver, who has to think consciously about every single decision (signal, gears, brakes, accelerator, mirrors etc.) and yet is still overloaded with information and someone who has been driving for years for whom many of these tasks are handled by muscle memory and the unconscious brain. . . .

» *Data processing and visualisation expertise*

This is perhaps the only absolute essential thing I am looking for and is why ‘pure’ statisticians with experience only in SPSS, Stata or similar are really not likely to make it past an interview. I need people who are data manipulation cyborgs, for whom filtering, subsetting, merging and transforming data is second nature and can do it using their preferred tool (be it dplyr, Pandas, SQL or whatever) with flow, without being hindered by those tools. Think of the difference between a new driver, who has to think consciously about every single decision (signal, gears, brakes, accelerator, mirrors etc.) and yet is still overloaded with information and someone who has been driving for years for whom many of these tasks are handled by muscle memory and the unconscious brain. . . .

. . . Data manipulation is not the chore to get through to get to the proper data science – it is data science. The same is true for visualisation. Exploring the data always comes before understanding it and visualising data is a hugely important part of this. Personally I spent a lot of time getting familiar enough with ggplot2 to be able to crank out quality plots on demand but if you have another favourite that you work efficiently with, that is equally fine. Related to this is report building, my workflow has been drastically improved by learning to generate automated reports using tools like Rmarkdown and Emacs org-mode. If new hires have already spent the time getting to grips with these tools, they can get to grips with the data that much more quickly.

A few general comments about R

A few general comments about R

» R is now 24 years old (production quality version is 17 years old)

A few general comments about R

- R is now 24 years old (production quality version is 17 years old)
- Starting 10 years ago, there has been an effort led by Hadley Wickham to improve the data handling and visualization aspects of R (once known as the "Hadleyverse" but now known as the "tidyverse")

A few general comments about R

- R is now 24 years old (production quality version is 17 years old)
- Starting 10 years ago, there has been an effort led by Hadley Wickham to improve the data handling and visualization aspects of R (once known as the "Hadleyverse" but now known as the "tidyverse")
- Old-timers tend to use the older, though less convenient, base R commands.

A few general comments about R

- R is now 24 years old (production quality version is 17 years old)
- Starting 10 years ago, there has been an effort led by Hadley Wickham to improve the data handling and visualization aspects of R (once known as the "Hadleyverse" but now known as the "tidyverse")
- Old-timers tend to use the older, though less convenient, base R commands.
- The tidyverse approach is rapidly winning out

A few general comments about R (continued)

A few general comments about R (continued)

- We will focus on the tidyverse approach with some base R as well

A few general comments about R (continued)

- We will focus on the tidyverse approach with some base R as well
- Style
  - Style is important in R as in any code
    - "Thissentenceisunderstandablebutnoteasily"
  - Page 24 of *Data Wrangling with R* has a good style guide
  - There's a detailed style guide at style.tidyverse.org. Pay particular attention to the advice on spacing and indentation.