

You code to build models

You build models to make decisions

R for Data Science

Author: LDG

Is coding valuable?

Everybody says coding is valuable

Salaries (and framing!) **seem** to provide supporting evidence

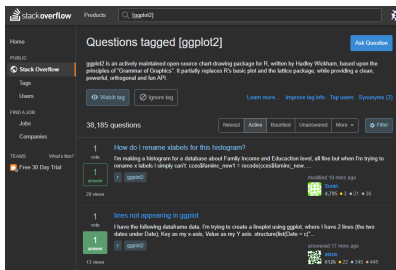


But on the other hand...

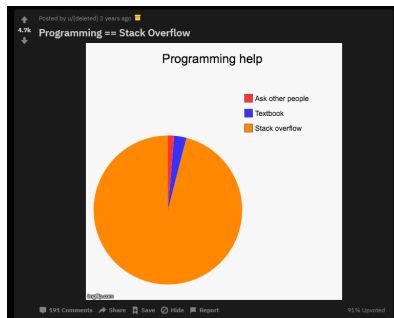
Is coding valuable?

Wisdom of Crowds

... code can just be copy-pasted from Stack Overflow



The screenshot shows the Stack Overflow website interface. The search bar at the top contains the text 'ggplot2'. Below the search bar, the page title is 'Questions tagged [ggplot2]'. A description of the 'ggplot2' package is visible: 'ggplot2 is an actively maintained open source chart drawing package for R, written by Hadley Wickham, based upon the principles of "Grammar of Graphics". It partially replaces R's base plot and the lattice package, while providing a clean, powerful, orthogonal and fun API.' There are 38,185 questions listed. Two questions are highlighted with green '1' icons, indicating they have one answer. The first question is 'How do I rename labels for this histogram?' and the second is 'lines not appearing in ggplot2'.



Is coding valuable?

Technological advancements

... and complex procedures can be done in a single line of code...¹

```
nn <- neuralnet(  
  Species == "setosa" ~ Petal.Length + Petal.Width,  
  data = iris,  
  linear.output = FALSE)
```

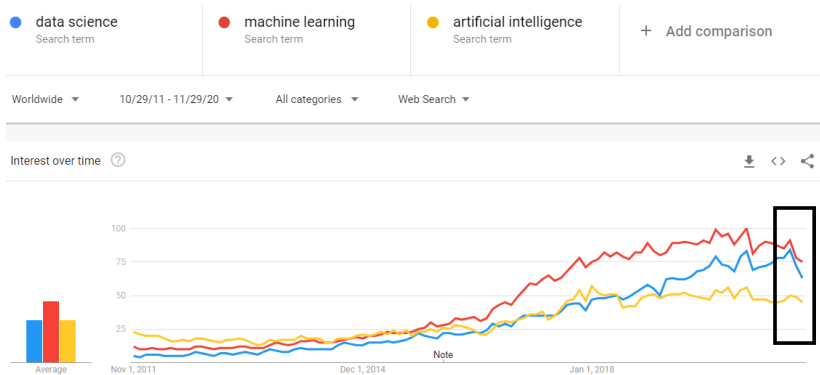
¹<https://www.rdocumentation.org/packages/neuralnet/versions/1.44.2/topics/neuralnet>

Is coding valuable?

zero-profit condition

... so if there were wage premiums simply because few people knew how to code...

... over time those premiums should dissipate (especially if low-level programming can be done by AIs)



People don't value information

They value **stories** – or more generally, **models** that **summarize** and **compress** information

RESEARCH ARTICLE



Using narratives and storytelling to communicate science with nonexpert audiences

Michael F. Dahlstrom

PNAS September 16, 2014 111 (Supplement 4) 13614-13620; first published September 15, 2014;
<https://doi.org/10.1073/pnas.1320645111>

Edited by Dietram A. Scheufele, University of Wisconsin-Madison, Madison, WI, and accepted by the Editorial Board April 7, 2014 (received for review November 1, 2013)

Article

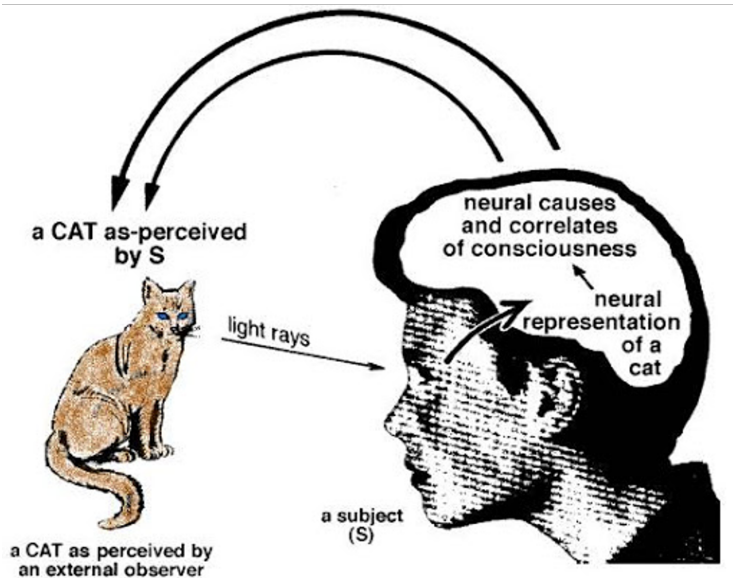
Info & Metrics

PDF

Abstract

Although storytelling often has negative connotations within science, narrative formats of communication should not be disregarded when communicating science to nonexpert audiences. **Narratives offer increased comprehension, interest, and engagement.** Nonexperts get most of their science information from mass media content, which is itself already biased toward narrative formats. Narratives are also intrinsically persuasive, which offers science communicators tactics for persuading otherwise resistant audiences, although such use also raises ethical considerations. Future intersections of narrative research with ongoing discussions in science communication are introduced.

Humans are pattern seekers



Humans are pattern seekers

'Virgin Mary' toast fetches \$28,000

A decade-old toasted cheese sandwich said to bear an image of the Virgin Mary has sold on the eBay auction website for \$28,000.

An internet casino confirmed it had purchased the sandwich, saying it had become a "part of pop culture".

Goldenpalace.com says it will take the sandwich on world tour before selling it and donating the money to charity.

Diane Duyser, from Florida, says the sandwich has never gone mouldy since she made it 10 years ago.



The toast is not intended for consumption

Modeling = storytelling (with data)

People often say more information is better

But models go in the opposite direction: **throw out** information to **gain** information

For example, the estimator of the population mean:

$$E[X] = \frac{1}{n} \sum_{i=1}^n x_i$$

Take n data points and **compress** them into one data point (the sample mean \bar{x})

Ditto linear regression, logistic regression, neural nets, etc.

You code to build models

Code itself is not valuable – nobody pays for Stack Overflow answers

What's valuable is what code **creates**

An app, a website – or to the data scientist, a **model**

A model of what?

- ▶ a model that **predicts** (e.g., “**what** will people buy?”)
- ▶ a model that **infers** (e.g., “**why** will people buy it?”)

You build models to make decisions

And even then the model is only so valuable!

Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction

Ajay Agrawal

Joshua S. Gans

Avi Goldfarb

JOURNAL OF ECONOMIC PERSPECTIVES

VOL. 33, NO. 2, SPRING 2019

(pp. 31-50)

“Prediction is useful because it is an **input into decision-making**.

Prediction has no value in the absence of a decision.”

- ▶ “**what** will people buy?” \implies **decide** what to make
- ▶ “**why** will people buy it?” \implies **decide** how to price it (or how to market it, or . . .)

Our focus

So we are here to introduce model building with R

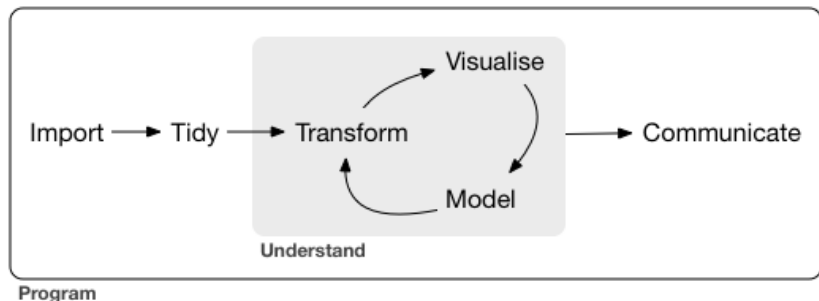
Focus is not on “Stack Overflow questions” (i.e., nuts and bolts stuff you can get from the internet)

Focus on the tidyverse packages:²

- ▶ transforming and summarizing (dplyr)
- ▶ visualizing (ggplot2)
- ▶ modeling (base R, broom, modelr)
- ▶ reproducing (RMarkdown)

²<https://www.tidyverse.org/>

The tidyverse



Why R?

Programming languages, like spoken languages, emerge to solve communication problems

R emerged from S (Bell Labs) to make statistical programming easier³

But unlike spoken languages, programming languages have evolved to become **complements** rather than **substitutes**

The market for programming languages is **not zero-sum**

So why do people learn different languages?

path dependence: people randomly exposed to one language or another (class, friend, whatever) and anchor in it because of a) high fixed costs to learning a language and b) strong network effects of mastering one language

public goods problem: who would pay the substantial cost of developing, distributing and coordinating users around a **free**, one-ring-to-rule-them-all language? (governments historically solved this problem with spoken languages – force people to speak one language or else!)

³[https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))

R vs and Python

R package `library(reticulate)` was designed to make it **easier** (not harder!) to use another language inside R...

...especially in **interactive notebooks** – which is what we will use

```
13
14 ▾ ```{python}   
15 import pandas
16 flights = pandas.read_csv("flights.csv")
17 flights = flights[flights['dest'] == "ORD"]
18 flights = flights[['carrier', 'dep_delay', 'arr_delay']]
19 flights = flights.dropna()
20 ```
21
22 ▾ ```{r, fig.width=7, fig.height=3}   
23 library(ggplot2)
24 ggplot(py$flights, aes(carrier, arr_delay)) + geom_point() + geom_jitter()
25 ```
26
```