Data extraction, transformation, and loading in R

Ernest Guevarra

2022-02-07

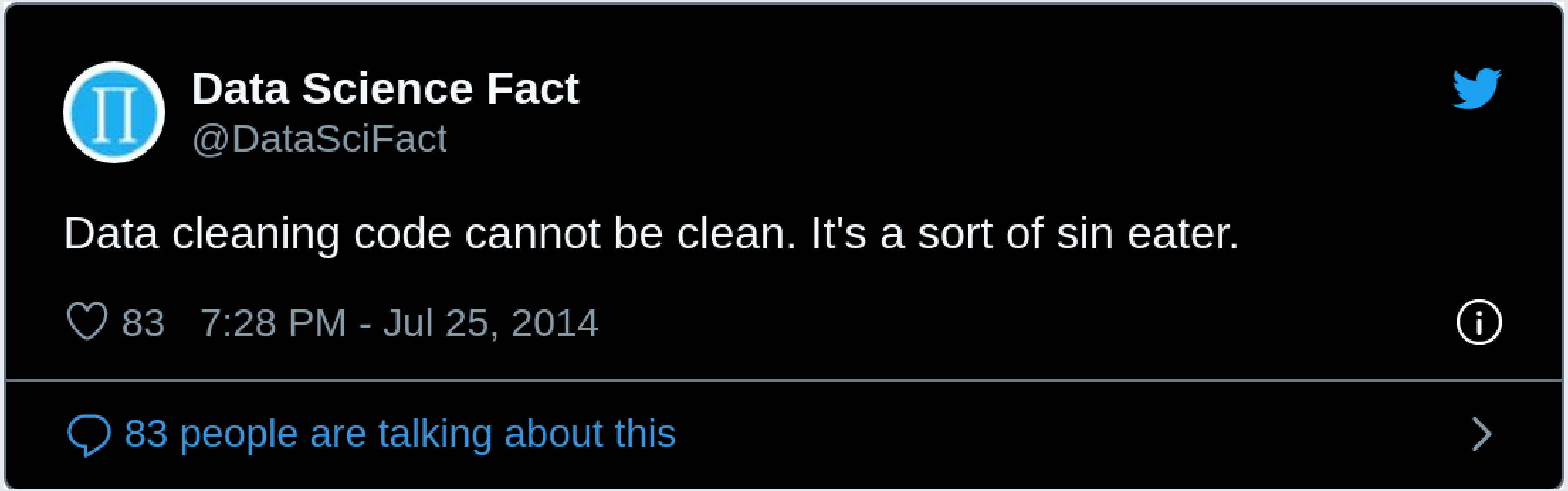
Outline

- Concepts and considerations about data
- Practical session

Concepts and considerations about data in tweets and quotes

Consideration #1:

Data cleaning/processing/checking is an important, but very challenging, step when working with data



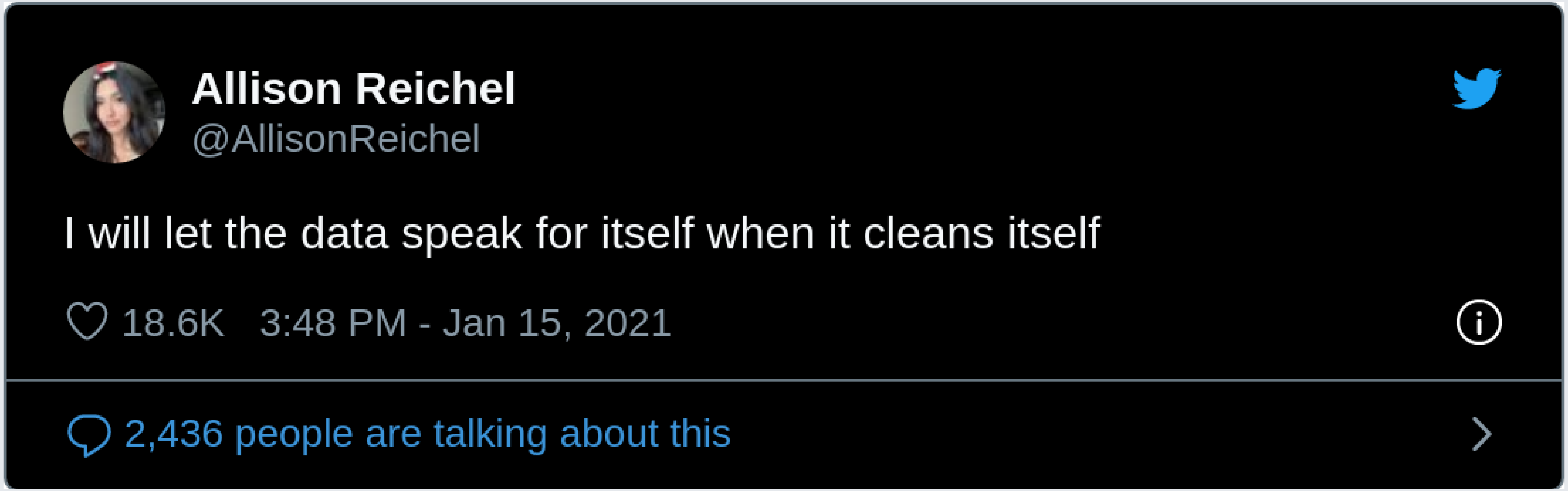
Data Science Fact
@DataSciFact

Data cleaning code cannot be clean. It's a sort of sin eater.

♡ 83 7:28 PM - Jul 25, 2014

83 people are talking about this

@StatFact tweet



Allison Reichel
@AllisonReichel

I will let the data speak for itself when it cleans itself

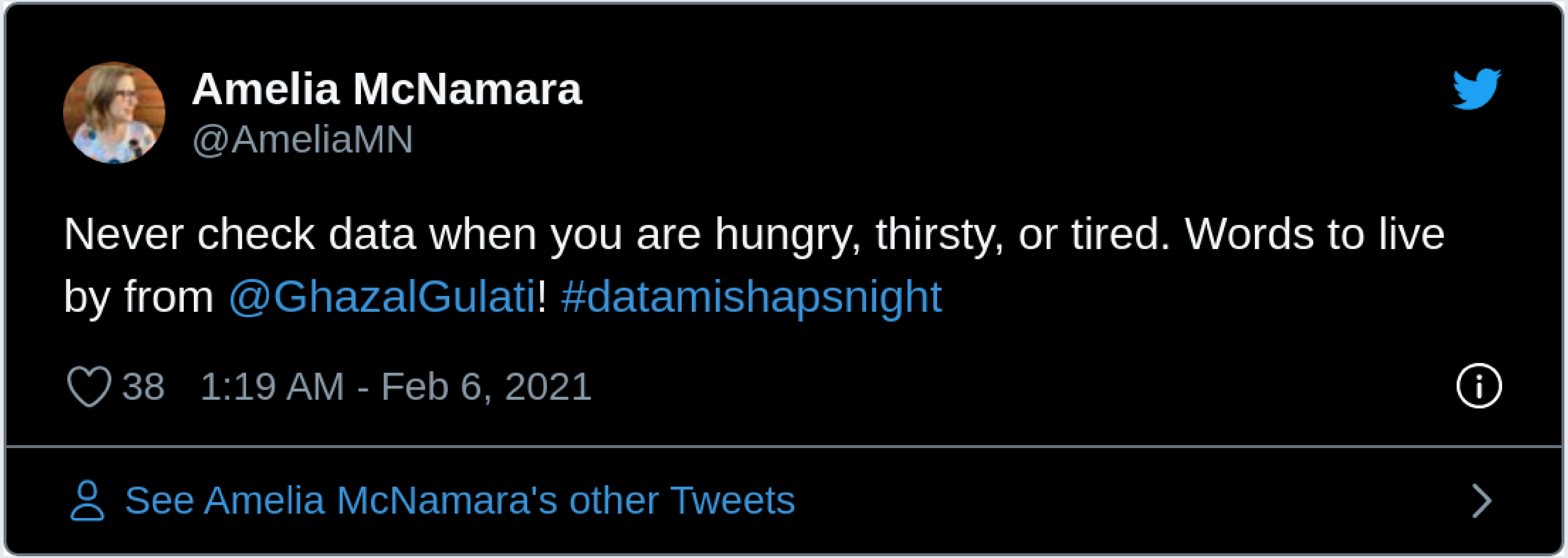
♡ 18.6K 3:48 PM - Jan 15, 2021

2,436 people are talking about this

@AllisonReichel tweet

and I’m still pretty sure some of the data is missing, but it could still be here, in this ONE HUNDRED SHEET excel file

@RallidaeRule tweet



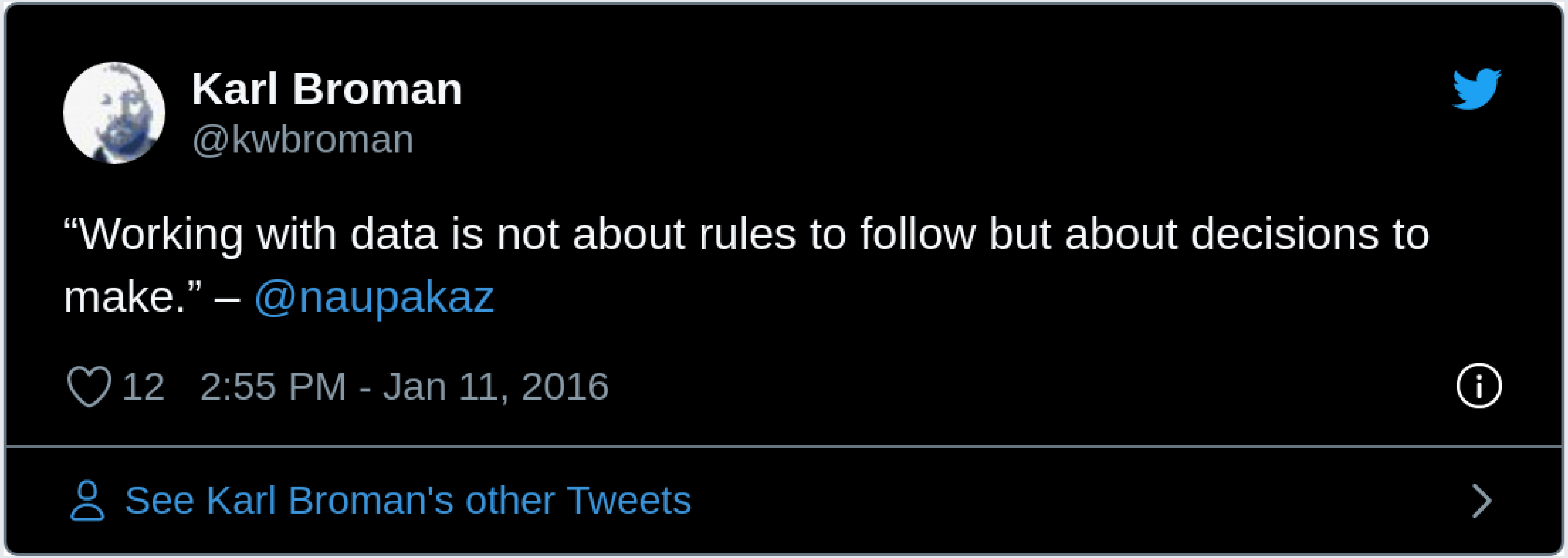
Amelia McNamara
@AmeliaMN

Never check data when you are hungry, thirsty, or tired. Words to live by from @GhazalGulati! #datamishapsnight

@GhazalGulati, via @AmeliaMN tweet

Classroom data are like teddy bears; real data are like a grizzly with salmon blood dripping out its mouth.

@JennyBryan, via @sgrifter tweet



Karl Broman
@kwbroman

“Working with data is not about rules to follow but about decisions to make.” – @naupakaz

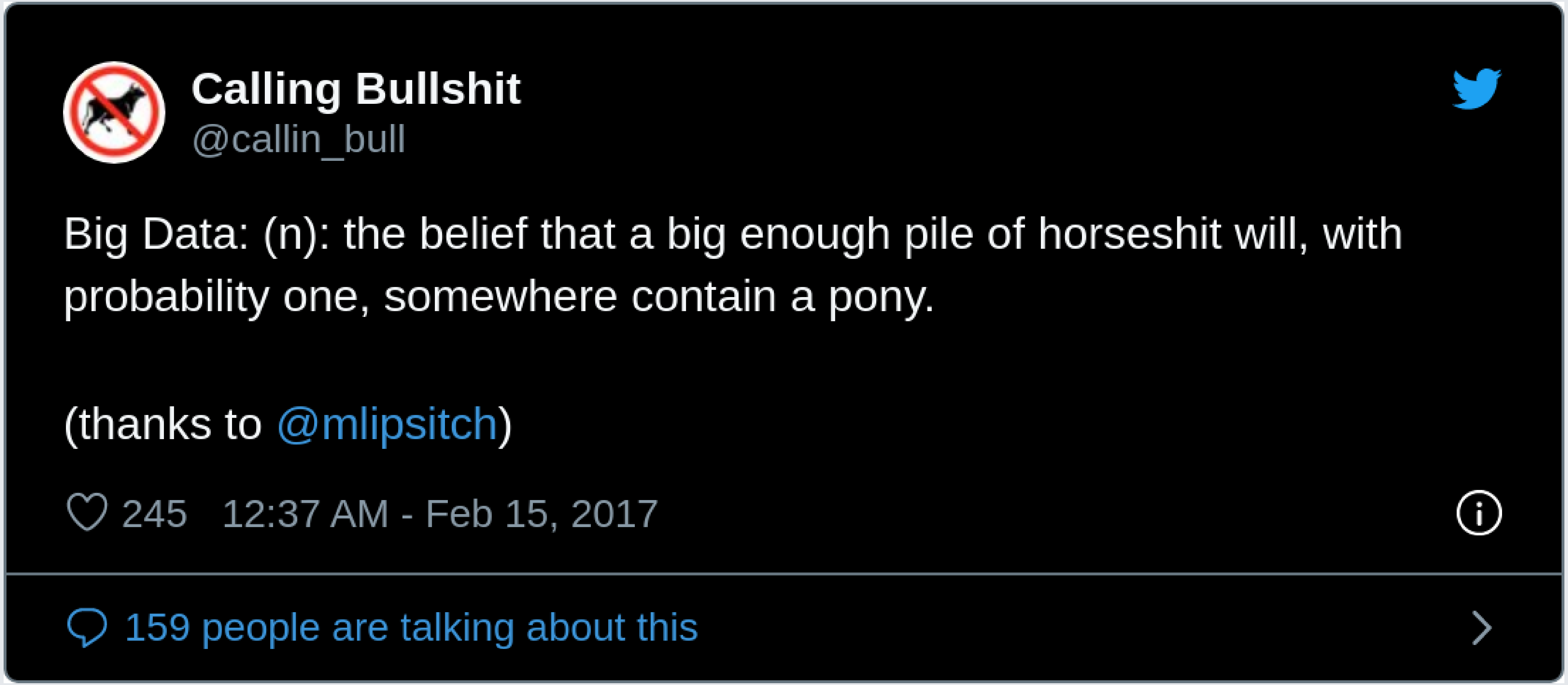
12 2:55 PM - Jan 11, 2016

See Karl Broman's other Tweets

@naupakaz, via @kwbroman tweet

Consideration #2

Big data is what everyone is talking about



Calling Bullshit
@callin_bull

Big Data: (n): the belief that a big enough pile of horseshit will, with probability one, somewhere contain a pony.

(thanks to @mlipsitch)

♡ 245 12:37 AM - Feb 15, 2017

159 people are talking about this

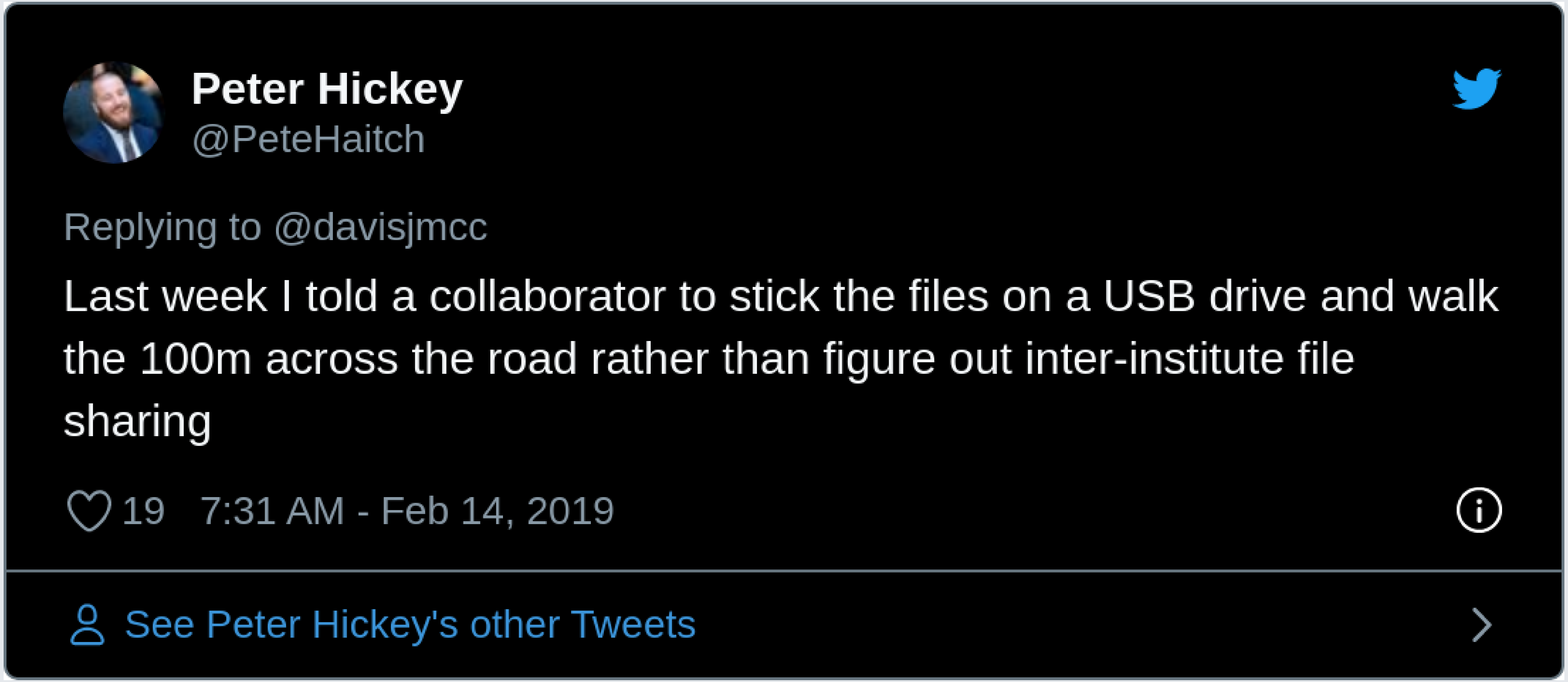
@mlipsitch, via @callin_bull tweet

Let's start the "titanic data" movement. Data too big to fail.

@neilfws with assist from @aaronquinlan, tweet

Consideration #3

Data sharing requires a well-thought out process that everyone can follow



Peter Hickey
@PeteHaitch

Replying to @davisjmcc

Last week I told a collaborator to stick the files on a USB drive and walk the 100m across the road rather than figure out inter-institute file sharing

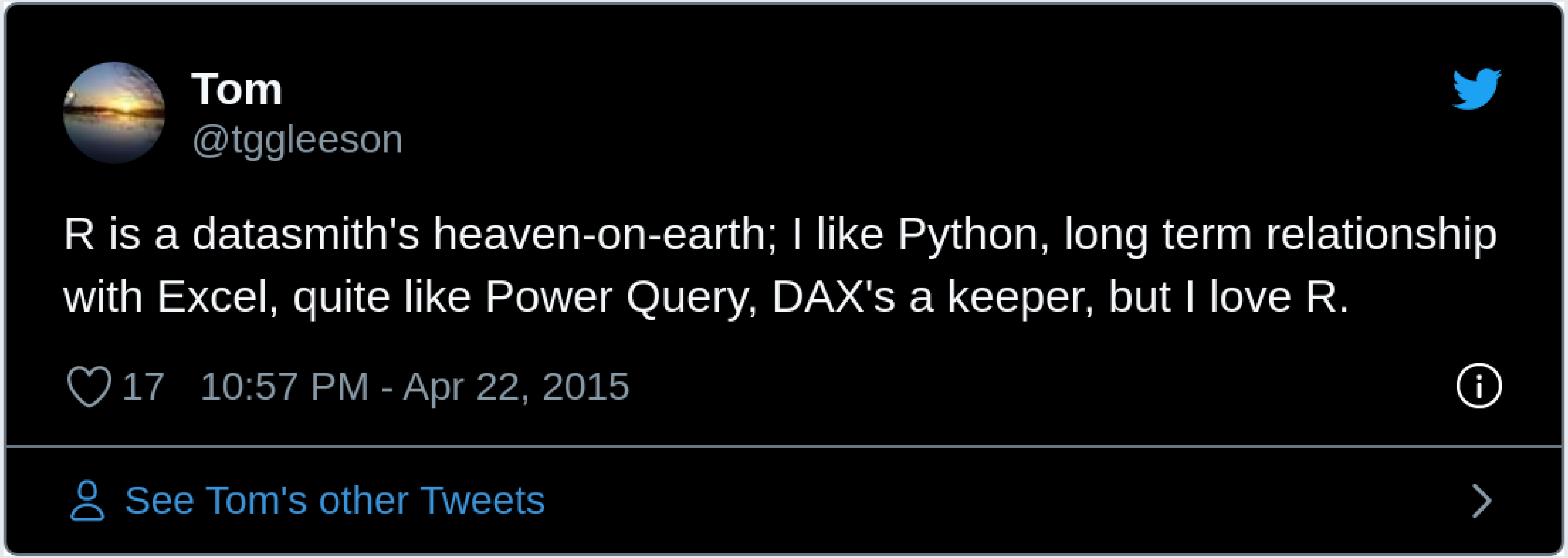
♡ 19 7:31 AM - Feb 14, 2019

See Peter Hickey's other Tweets

@PeteHaitch tweet

Consideration #4

Working with data in R is great!



Tom
@tggleeson

R is a datasmith's heaven-on-earth; I like Python, long term relationship with Excel, quite like Power Query, DAX's a keeper, but I love R.

17 10:57 PM - Apr 22, 2015

@tggleeson tweet

Practical session

Practical session

- We will all go through Exercise #1 in the Practical R for Epidemiologists book
- This is also one of your next assignments for this lecture series. This can be accessed via our GitHub Classroom through this link - https://classroom.github.com/a/xQNlWMp-
- This link has also been emailed to you.

Questions?

Thank you!

Slides can be viewed at https://OxfordIHTM.github.io/open-reproducible-science/session3.html

PDF version of slides can be downloaded at https://OxfordIHTM.github.io/open-reproducible-science/pdf/session3-working-with-data-in-r.pdf

R scripts for slides available here