Reproducible Research: Why and How

SER Pre-Conference Workshop

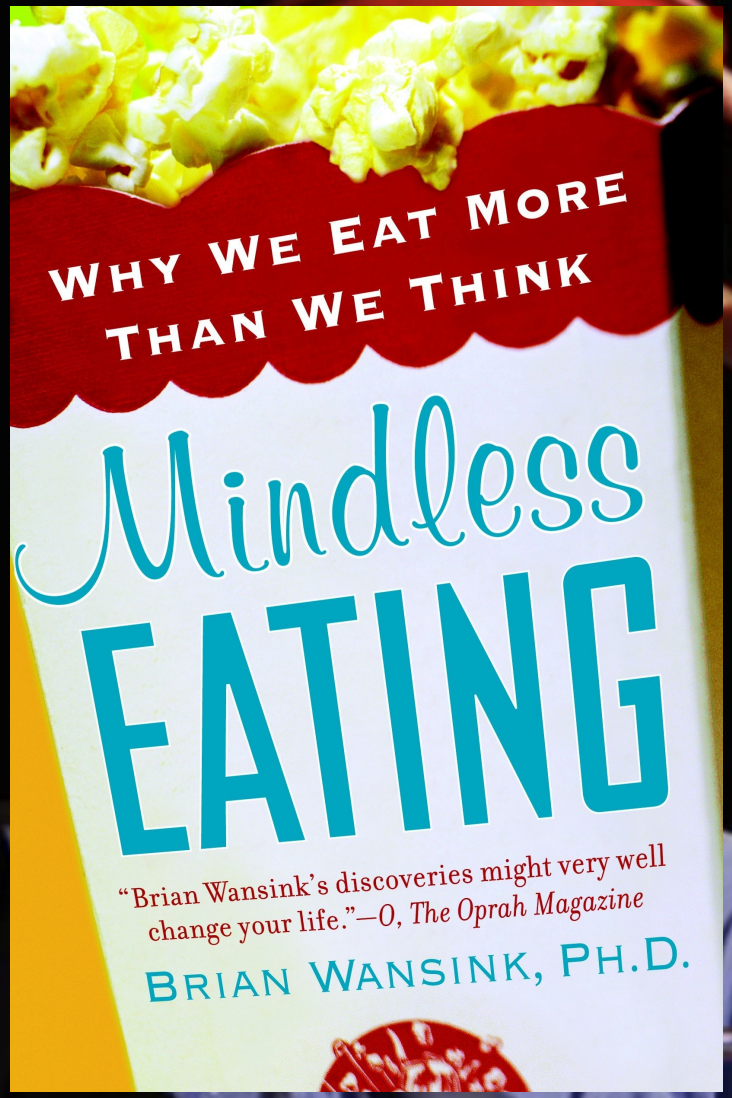
Sam Harper

2020-10-30

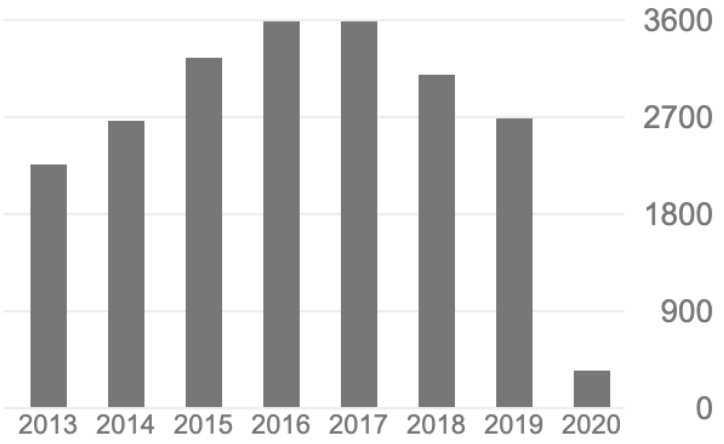
I am a social epidemiologist at McGill University.

I work mainly on evaluating programs and policies on social inequalities in health.

I have nothing to disclose, other than a strong commitment to open science



Cited by		VIEW ALL
	All	Since 2015
Citations	31622	16656
h-index	83	64
i10-index	273	232



NOV
20
2007

Brian Wansink! At the USDA!

Every now and then something incredible happens and here it is. Brian Wansink, Cornell Professor and author of Mindless Eating, has been appointed executive director of the USDA Center for Nutrition Policy and Promotion. This is the piece of USDA responsible for dietary advice to the public. Wansink is the guy who does the terrific research on environmental determinants of overeating showing that large portions, wide drinking glasses, foods close by, and health claims encourage everyone to eat more calories than they need or want. Will he be able to anything good at USDA? Let's hope so. In the meantime, cheers to USDA for making a brilliant appointment.

https://www.foodpolitics.com/2007/11/brian-wansink-at-the-usda/

"I gave her a data set of a self-funded, failed study which had null results... I said, 'This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set.' I had three ideas for potential Plan B, C, & D directions (since Plan A had failed)." -blog, 2016

Archived post here.

"I gave her a data set of a self-funded, failed study which had null results... I said, 'This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set.' I had three ideas for potential Plan B, C, & D directions (since Plan A had failed)." -blog, 2016

Enterprising grad students found:

- impossible values
- incorrect ANOVA results
- dubious p-values

Wansink denied requests for access to the original data.

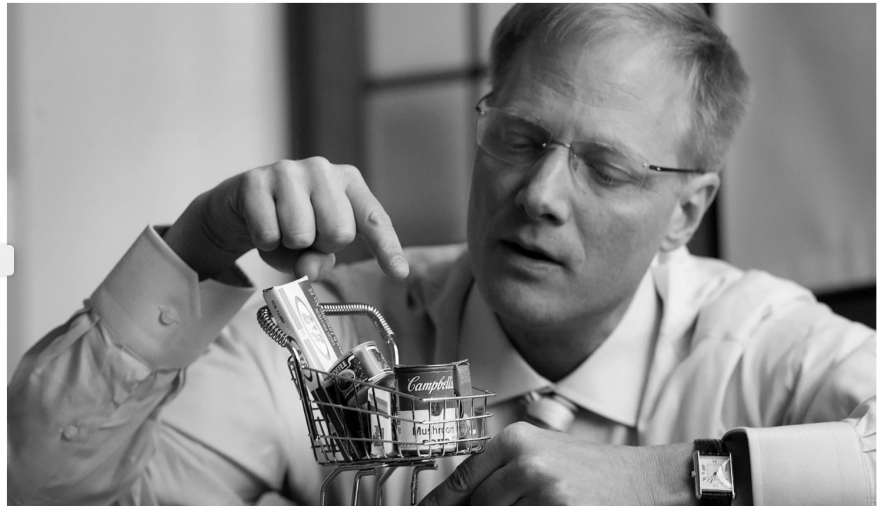
Archived post here.

A top Cornell food researcher has had 15 studies retracted. That's a lot.

Brian Wansink is a cautionary tale in bad incentives in science.

By Brian Resnick and Julia Belluz | Updated Oct 24, 2018, 2:25pm EDT

SHARE



Wansink resigned from Cornell in 2019.

Tools have consequences

SEPT2 gene

👇

👇

2-Sep

Boddy (2016), Ziemann (2016)

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl ver-

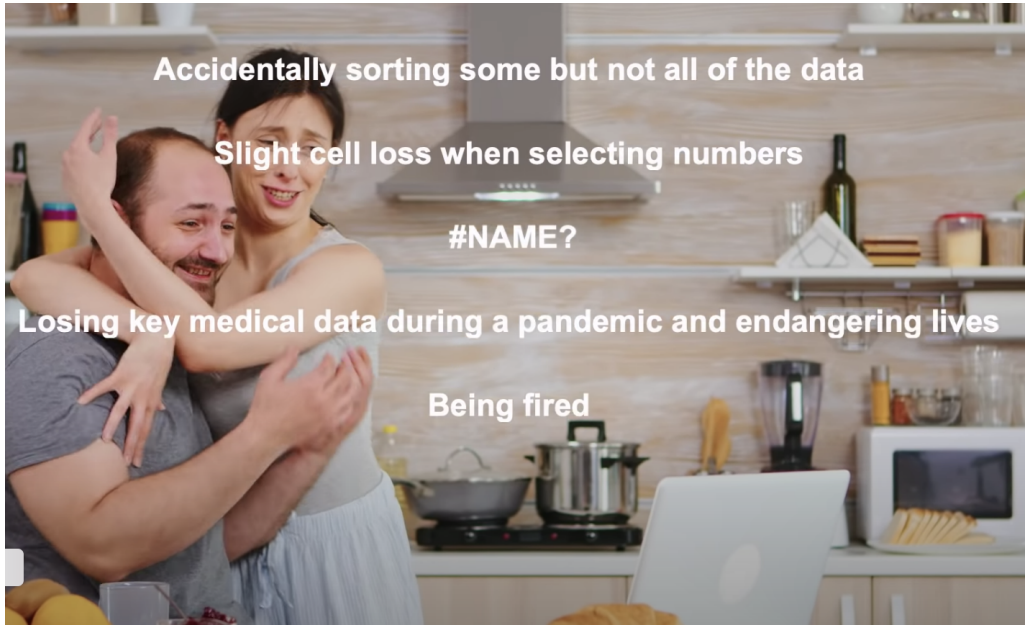
More recently...

Covid: how Excel may have caused loss of 16,000 test results in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

Sources: The Guardian (2020-10-06), YouTube

Are Spreadsheets® right for you? Side effects may include:



The integrity of science is compromised by non-reproducible research.

There are tools to help you.

Setting expectations

Today is not about:

- Mastering software
- Learning to code
- Mastering version control
- Mastering statistical analysis

Today is about:

- *Why* to do reproducible research.
- Understanding concepts of *how* to do it.
- Getting familiar with tools to help.
- Learning where to find out more.

What is the plan?

Plan for today

1. Scientific Integrity Problems (1220h-1250h) ☕
2. Design Solutions (1300h-1330h)
3. Analytic Solutions (1330h-1350h, ☕ 1400h-1450h ☕)
4. Dissemination Solutions (1500h-1530)
5. Reproducible Example (1530h-1600h)

Code of Conduct

Do

- Be respectful.
- Ask questions in the chat.
- Use the 'raise your hand' feature to ask a question or make a comment.
- Interrupt me if I didn't notice your chat or 'hand'.
- Feel free to turn your camera on (if you are comfortable).

Don't

- Worry about taking notes (but feel free to do so). You will have access to all of the material for the workshop when we are finished.
- Be disrespectful or rude.

Let's go!