Target Shuffling
by Scott Nestler
to MRUG, 21 AUG 2020

Hello
my name is

Scott Nestler

snestler@nd.edu

scottnestler

@scottnestler

snestler

Family:

- From Harrisburg, PA
- Kristin, Anna(18), Sophia(15)

Professional:

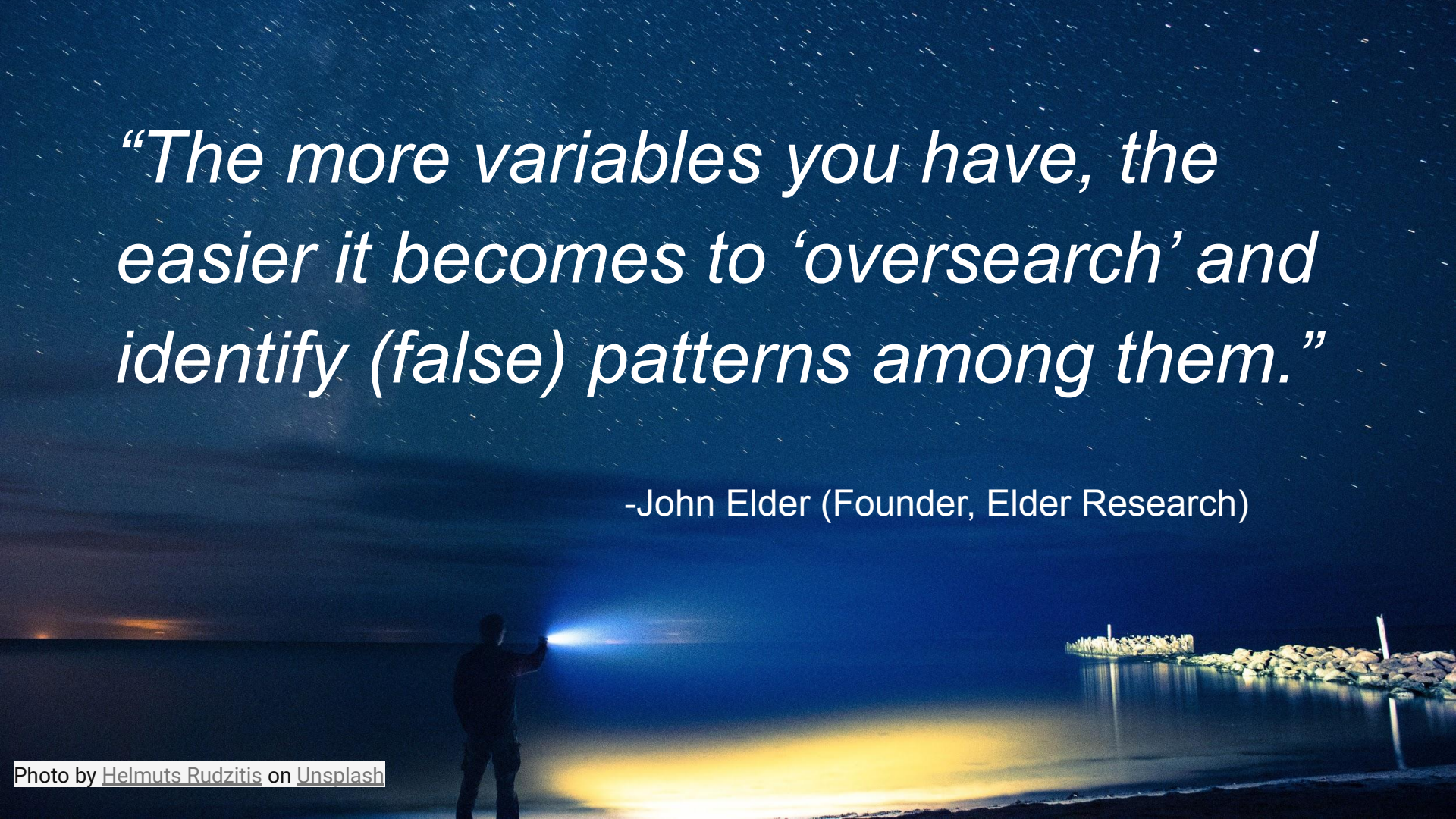
- Univ. of Notre Dame, 5 years
- U.S. Army, 25 years

Academic:

- Univ. of Maryland - College Park (PhD, Management Science)
- Army War College (MSS, Strategy)
- Naval Postgraduate School (MS, Applied Math / OR)
- Lehigh University (BS, Civil Engineering)

Volunteer:

- Certified Analytics Professional (CAP)
- INFORMS Pro Bono Analytics
- Rebuilding Together


“The more variables you have, the easier it becomes to ‘oversearch’ and identify (false) patterns among them.”
-John Elder (Founder, Elder Research)
Photo by Helmuts Rudzitis on Unsplash

“The Redskins Rule”

IF:

The Washington Redskins won their last home football game,

THEN:

The incumbent party would win the U.S. Presidential election.

Year	Redskins Win?	Incumbent Win?	Rule Held?
2016	W	L	No
2012	L	W	Yes
2008	L	L	Yes
2004	L	W	Yes
2000	L	L	Yes
1996	W	W	Yes
1992	L	L	Yes
1988	W	W	Yes
1984	W	W	Yes
			Yes
			Yes
			Yes
			Yes
			Yes
			Yes
1956	W	W	Yes
1952	L	L	Yes
1948	W	W	Yes
1944	W	W	Yes
1940	W	W	Yes
1936	W	W	Yes
1932	W	L	No

21 times in a row!
21 / 23 = 91% accuracy

IMAGE CREDIT: https://redskinswire.usatoday.com/2019/10/19/redskins-keys-to-victory-49ers/

What's the Problem?

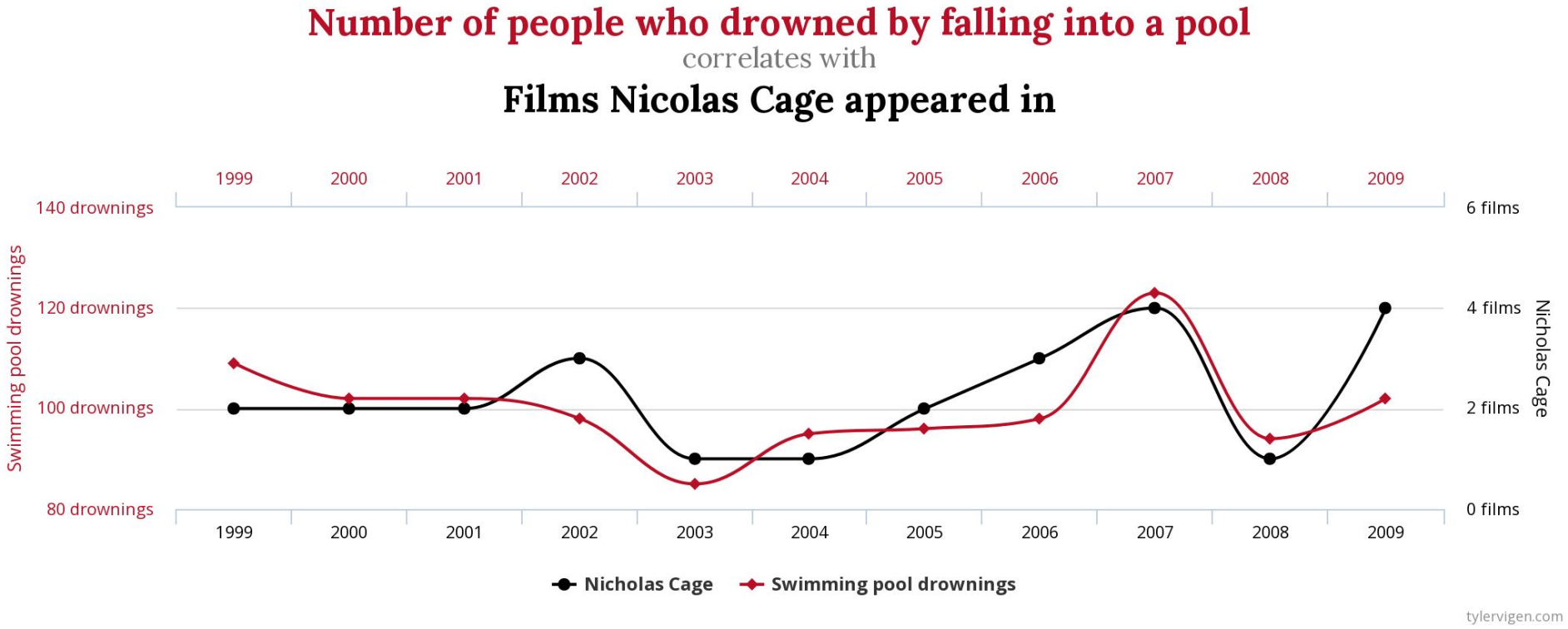
Can make inferences that are:

1. Wrong, and
2. Misleading

Especially if we forget that whole "correlation is not causation" mantra.

Photo by Olav Ahrens Røtne on Unsplash

Another Spurious Correlation



SOURCE: http://tylervigen.com/spurious-correlations

Photo by Olav Ahrens Røtne on Unsplash

Target Shuffling* - A Solution?

- A process that “reveals how likely it is that results occurred by chance.”
- Particularly helpful for detecting the presence of false positives.
- It’s *“essentially a computer simulation that does what long-standing statistical tests were designed to do.”*

-John Elder

* Also called Y scrambling, randomization testing or permutation testing by some.

Isn’t that the Bootstrap?*

No; but it’s related.

* https://projecteuclid.org/euclid.aos/1176344552

Photo by Rostyslav Savchyn on Unsplash

Photo by Sheri Hooley on Unsplash

The “Target Shuffling” Process

1. Randomly shuffle the output (target variable) on the training data.
2. Search for combinations of variables with high concentration of interesting output.
3. Save the “most interesting” result.

REPEAT

4. Look at distribution of the bogus “most interesting results” to see how much you can get from random data.
5. Evaluate where on (beyond) this distribution your actual results stand.
6. Use this as your “significance” measure.

A Simple Example

Imagine you have a class of students ready to take a quiz.

Before beginning, you have them complete a card like this:

Name: Scott

Sex: Male

Age: 16th prime number minus one

Sisters: 1 (or 3, if include step-sisters)

Brothers: 0 (or 1, if incl. step-brothers)

Hair Color: Blonde / Brown/ Gray

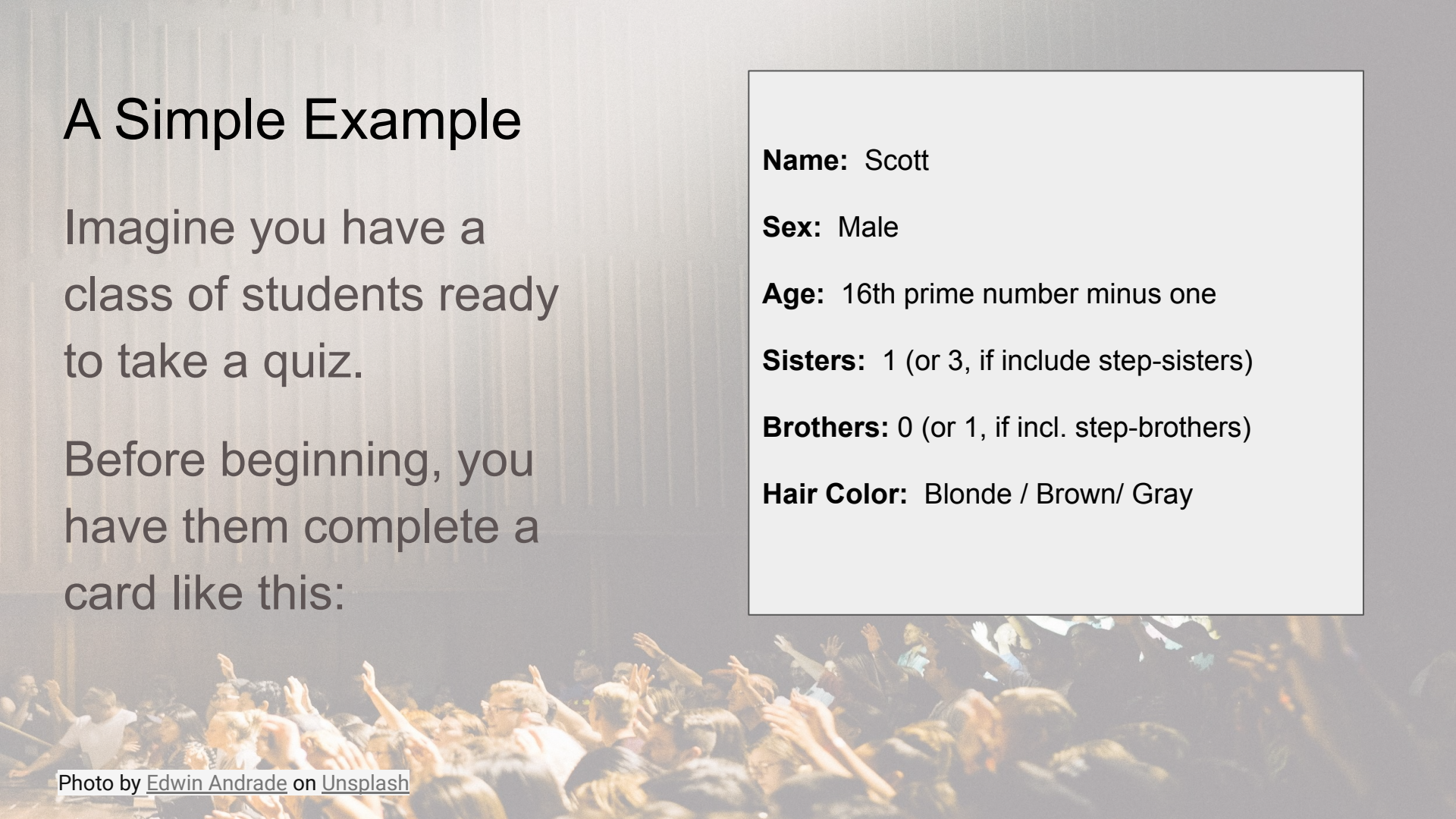


Photo by Edwin Andrade on Unsplash

A Simple Example (cont.)

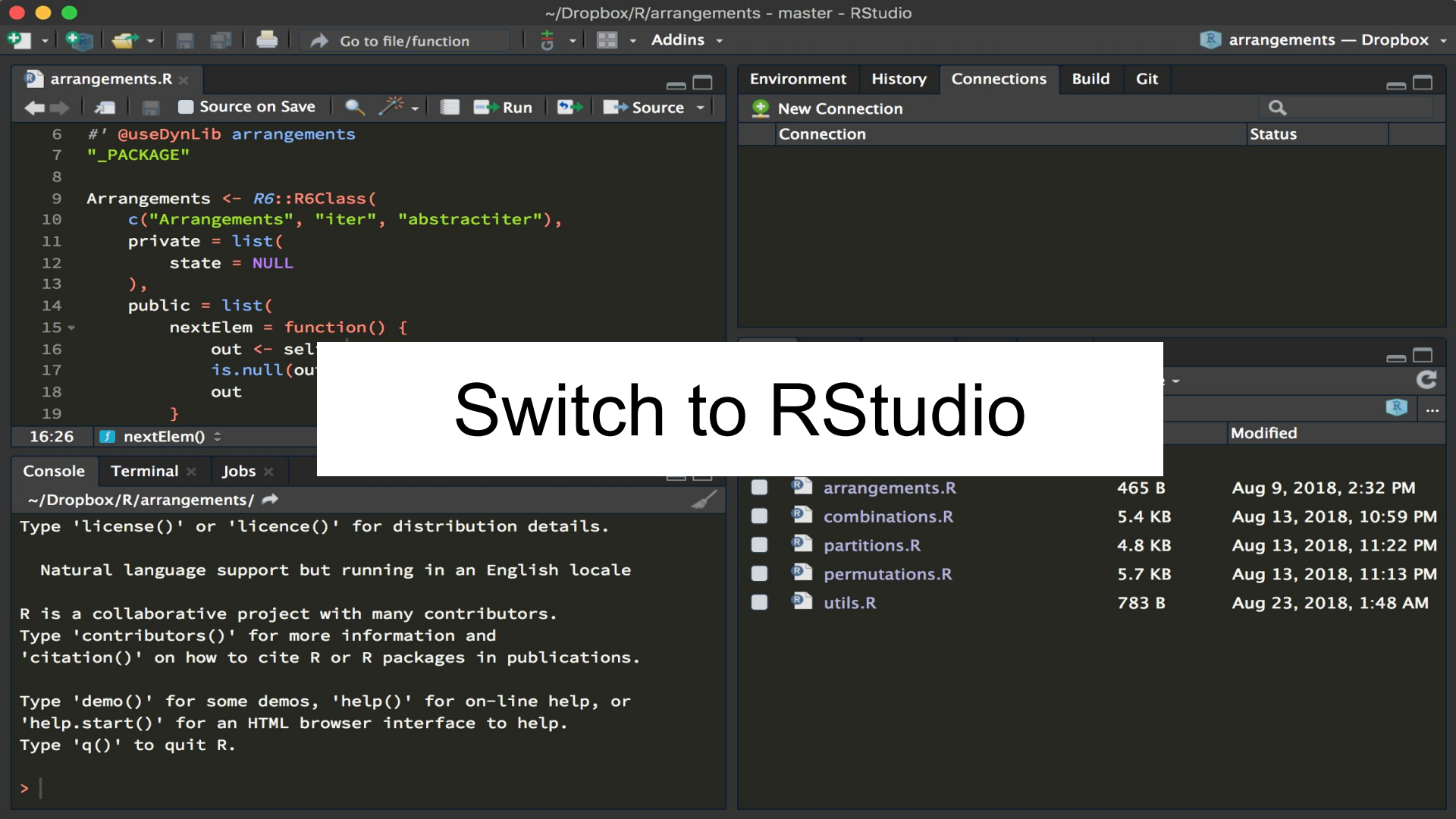
Quiz Score	Name	Sex	Age	Sisters	Brother	Hair Color	Eye Color	Pets
79	Scott	M	52	1	0	Brown	Hazel	Dog
87	Jen	F	50	1	0	Red	Blue	None
92	Luke	M	44	2	1	Brown	Green	None
94	Fariba	F	48	1	1	Brown	Brown	Cat
	...	...	...	...	...	...	...	...
74	Tony	M	46	0	3	Blonde	Blue	Dog

A Simple Example (cont.)

Quiz Score	Name	Sex	Age	Sisters	Brother	Hair Color	Eye Color	Pets
79 87	Scott	M	52	1	0	Brown	Hazel	Dog
87 94	Jen	F	50	1	0	Red	Blue	None
92 92	Luke	M	44	2	1	Brown	Green	None
94 74	Fariba	F	48	1	1	Brown	Brown	Cat
	...	...	...	...	...	...	...	...
74 79	Tony	M	46	0	3	Blonde	Blue	Dog

Now, “shuffle” these quiz scores, so that people are randomly assigned someone else’s.

This is an R Meetup, right???


~/Dropbox/R/arrangements - master - RStudio
Go to file/function
Addins
arrangements — Dropbox
arrangements.R
Source on Save
Run
Source
#' @useDynLib arrangements
"_PACKAGE"
Arrangements <- R6::R6Class(
c("Arrangements", "iter", "abstractiter"),
private = list(
state = NULL
),
public = list(
nextElem = function() {
out <- sel
is.null(ou
out
}
16:26
nextElem()
Environment
History
Connections
Build
Git
New Connection
Connection
Status
Switch to RStudio
Modified
arrangements.R 465 B Aug 9, 2018, 2:32 PM
combinations.R 5.4 KB Aug 13, 2018, 10:59 PM
partitions.R 4.8 KB Aug 13, 2018, 11:22 PM
permutations.R 5.7 KB Aug 13, 2018, 11:13 PM
utils.R 783 B Aug 23, 2018, 1:48 AM
Console
Terminal
Jobs
~/Dropbox/R/arrangements/
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>

(Hopefully) Really Rare Events

What if we are trying to detect something that really doesn't happen all that often?

If we change `n <- 1000`

to `n <- 1000000`

we might have some issues.

(Or at least need to be patient.)

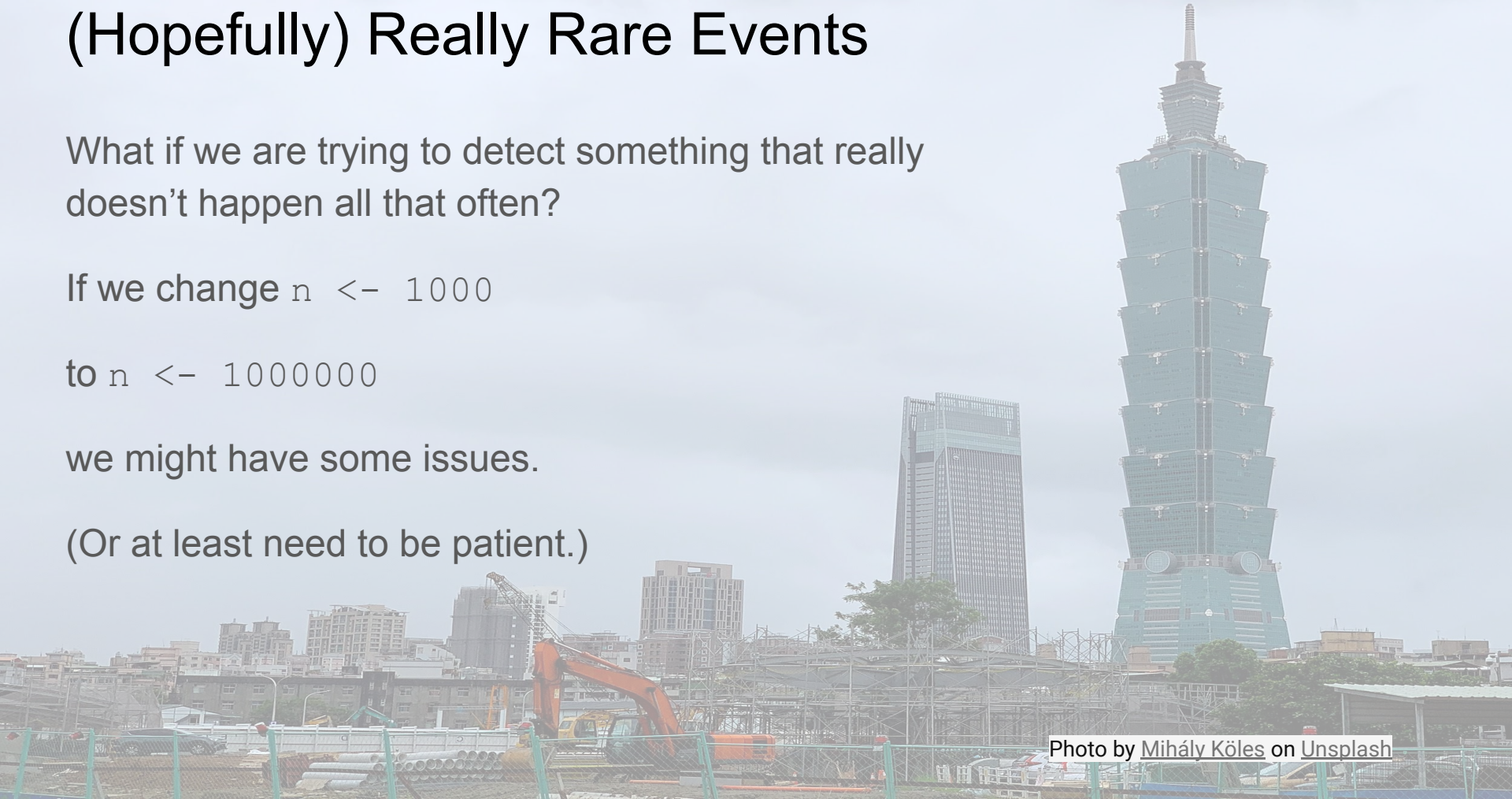


Photo by Mihály Köles on Unsplash

Parallel Options in R

Many packages to choose from in 2020:

- ’parallel’ (now comes with R, includes updated versions of earlier ‘multicore’ and ‘snow’
- ‘foreach’ (from Revolution Analytics)
- NOTE: Much easier on Linux or Mac than in Windows
- More options in CRAN Task View on High Performance and Parallel Computing

https://cran.r-project.org/web/views/HighPerformanceComputing.html

Two Lessons Learned

“Target shuffling is a very good way to test non-traditional statistical problems.”

“More importantly, it’s a process that makes sense to a decision maker. Statistics is not persuasive to most people…”

-John Elder

Elder Research