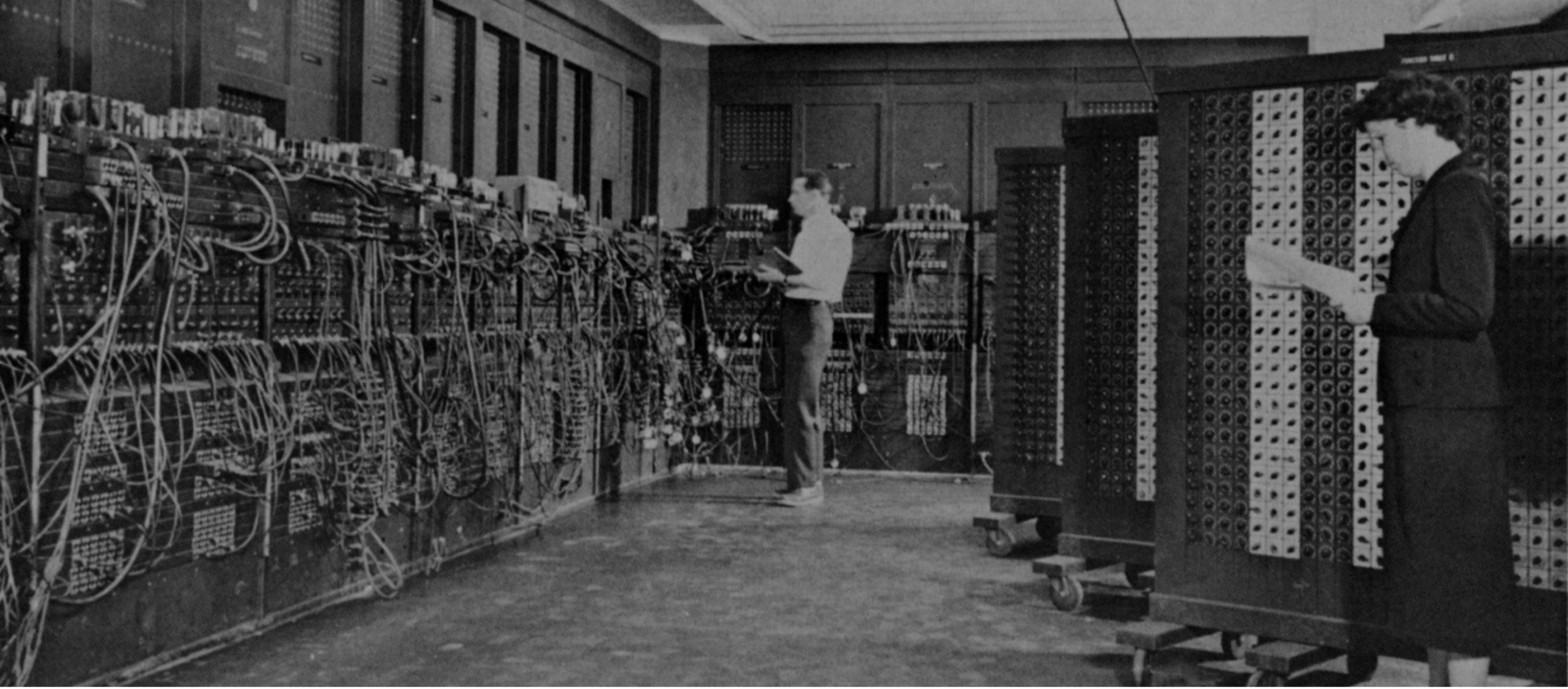


A Practical Guide to Your Computer

for Social and Policy Analysts

What are we up to?

- Where we've are: We want to get to data for analysis, but to do it in a structured and reliable way, we need to take care of the underlayment first
  - The very first part of the course: a lot of housekeeping
  - Tools, computational concepts, light programming
- Where we're going today:
  - A bit of history, a bit of science, a bit of data life advice

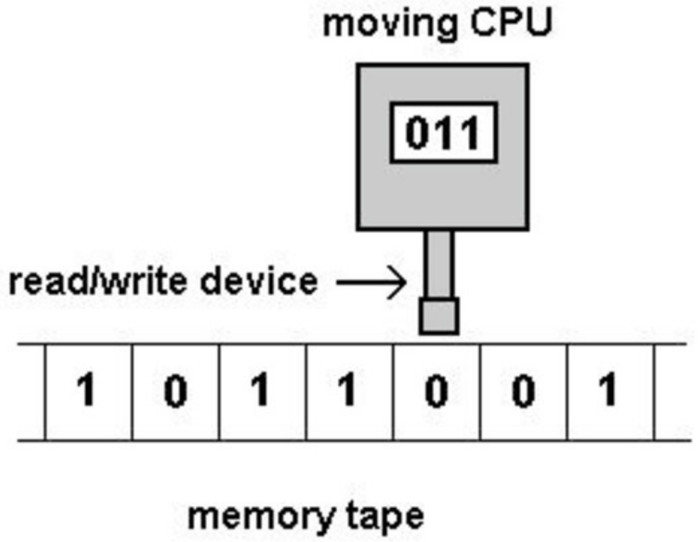
title page: ENIAC, the first programmable digital computer

What is a computer?

com·put·er | kəm'pyo͞odər |

noun

an electronic device for storing and processing data, typically in binary form, according to instructions given to it in a variable program: *my computer is frozen | the computer handles the entire process | [as modifier] : a computer program | there is a hugely expensive new computer system.*



a Turing machine

Course Housekeeping

- What are we doing?
  - We're on the road to data work - but we're taking care of some computational preliminaries first. If you are:
    - not terribly interested in the inner workings of a computer *and/or*
    - not that interested in doing arithmetic with R or learning about the differences between string, boolean, and numeric variables . . .
  - **That's OK!** Just bear with us for a little bit, because these things do become *useful* for the kinds of things you do want to do

Useful?

Verbatim, from job ads in public administration, policy, and campaigns:

- 1 “Quantitative, analytical and research skills. Ability to distill analysis into key results for broad consumption. A desire to grow expertise in a new subject area through research and analysis. **Skilled at analyzing reports and datasets to solve specific research objectives**.”
- 2. “Effective written and oral communication skills. Comfort with public speaking. Skill at **packaging and explaining complex concepts to a wide range of audiences through clear written materials**.”
- 3. “Experience using **GIS mapping** and/or Tableau for d**ata visualisation**”
- 4. “Ability to apply concepts from statistics and explain your decisions clearly to others”
- 5. “**Proficiency in R**, Python, Stata, or another statistical computing language”
- 6. “Experience in **data standardization and data manipulation**”
- 7. “Facility with research design, particularly survey design or experiment design”
- 8. “Experience in **natural language processing and textual analysis**”
- 9. “Proficiency with **GIS or other mapping software**”
- 10. “Conducting quantitative analysis using SPSS and/or **R statistical packages and other tools**”
- 11. “Building **data visualizations** and graphic representations of data”
- 12. “Writing analysis memos and summaries that provide strategic guidance to clients”
- 13. “Solid knowledge of **tidyverse R** and SQL”

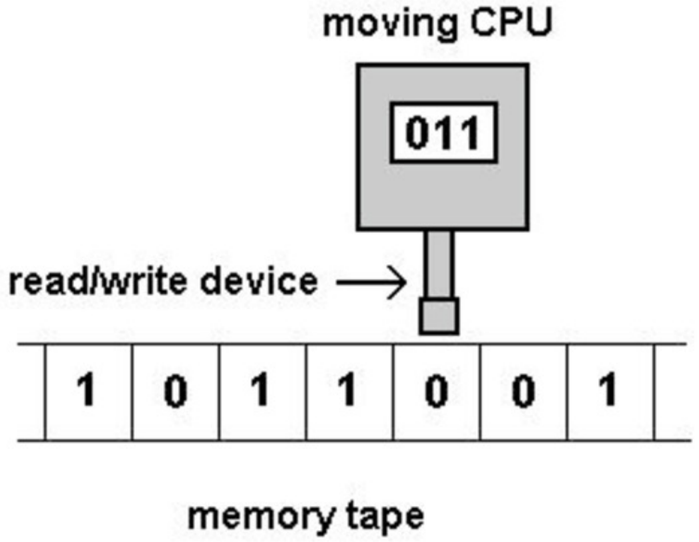
Source Ads: Data science, The Messina Group, Energy Campaigns Analyst, Sierra Club,Research Associate, WPA Intelligence, Data science, WPA Intelligence

What is a computer?

com·put·er | kəm'pyo͞odər |

noun

an electronic device for storing and processing data, typically in binary form, according to instructions given to it in a variable program: *my computer is frozen | the computer handles the entire process | [as modifier] : a computer program | there is a hugely expensive new computer system.*



An illustration of a Turing Machine

title page: ENIAC, the first programmable digital computer

What is a computer?

- Really two things:
  - **Memory** - stored data that gets manipulated
  - **Processor** - thing that reads data, manipulates it, and writes it back to memory
    - according to a series of **instructions**

Today

- Talk about three things
  - Hardware: what a computer is
  - Software: how we give a computer commands
  - Tools: the software we use to give computers commands

Hardware

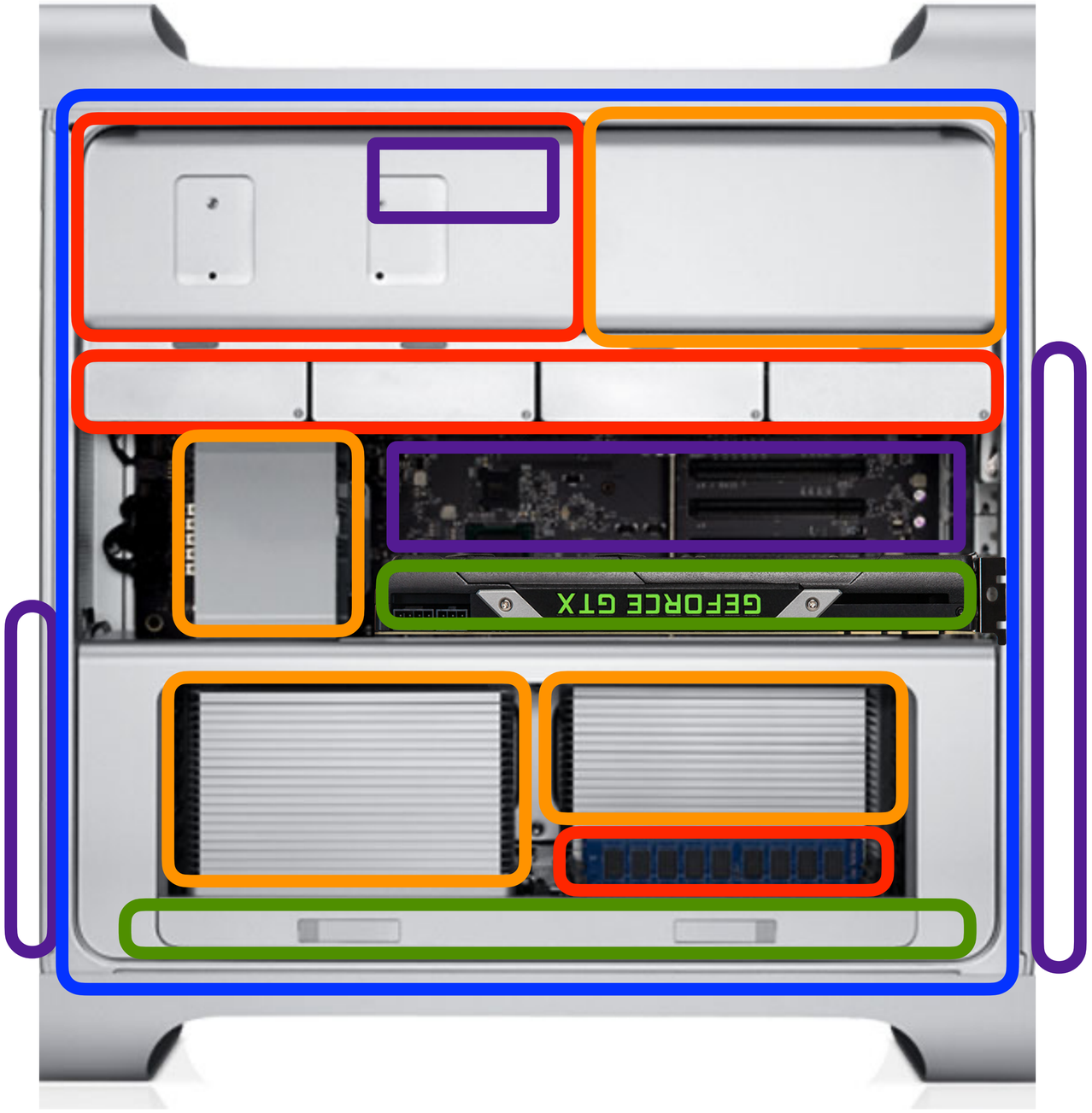
Processing

Core components

- **Processing**
  - The thing that performs operations
- **Memory**
  - Where data from those operations is read and stored

It must be more complex, right? (It is.)

- Each computer has a variety of different kinds of processing units:
  - Central Processing Unit (CPU), Graphics Processing Unit (GPU), Neural Processing Unit (NPU), Tensor Processing Unit (TPU), Media Processors, other *hard-coded* computational engines for specific purposes
- And each can have a variety of different kinds of memory:
  - Primary memory (RAM), secondary memory (drives/hard disks/solid state drives), tertiary/removable storage (usb sticks, floppy drives, CD/DVD), off-site (not under direct control of the processing units)
- All of which must be properly connected together through a databus (or just bus) that then must be powered and cooled while allowing for other I/O and connectivity to other computers and humans



- processing
- memory & storage
- motherboard/ bus
- power & cooling
- I/O and connectivity

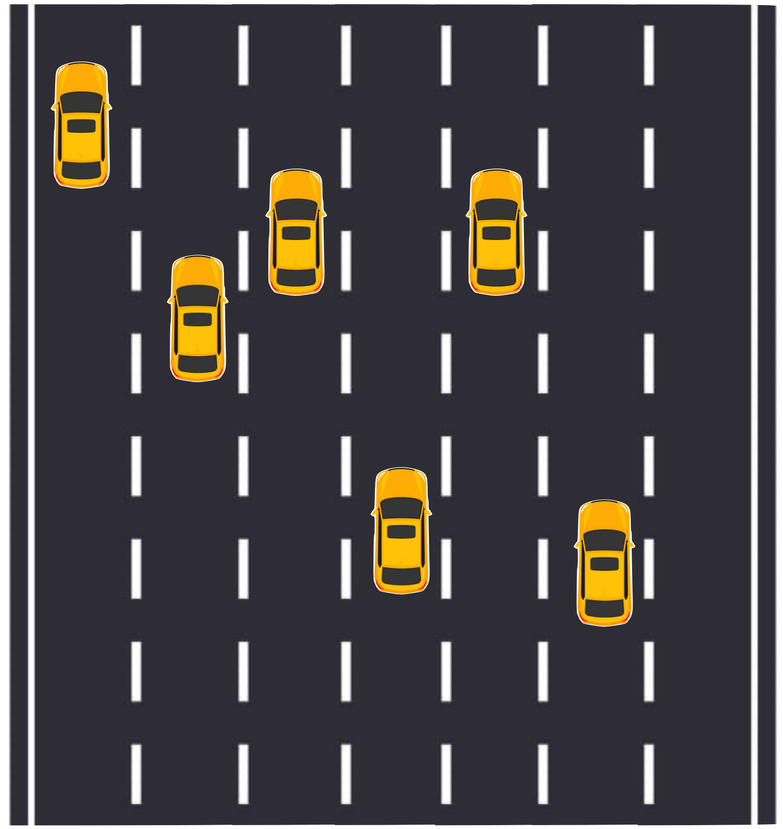
Different processing units?

- Fundamental theory in computing (Universal Turing Machine) states that any sufficient computing machine can, in theory, simulate another machine
- In other words, any computable sequence can be run by a UTM
  - But it does not say anything about **efficiency**
- We care about two forms of efficiency, as well:
  - Computing time
  - Energy cost

The very basics of computer processing

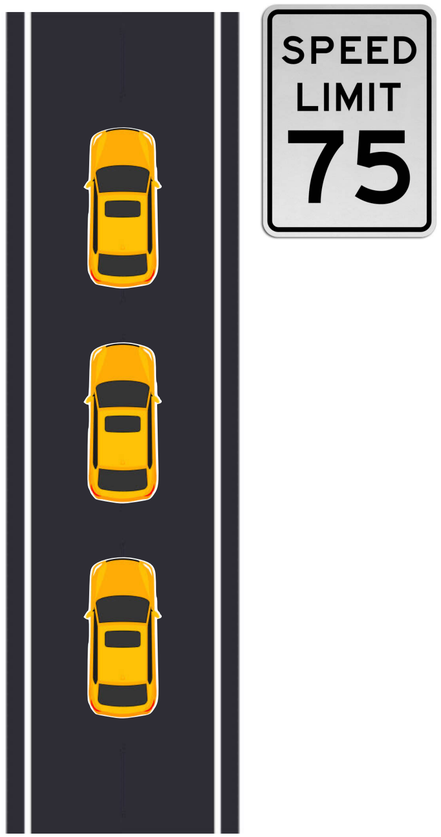
Serial

Parallel

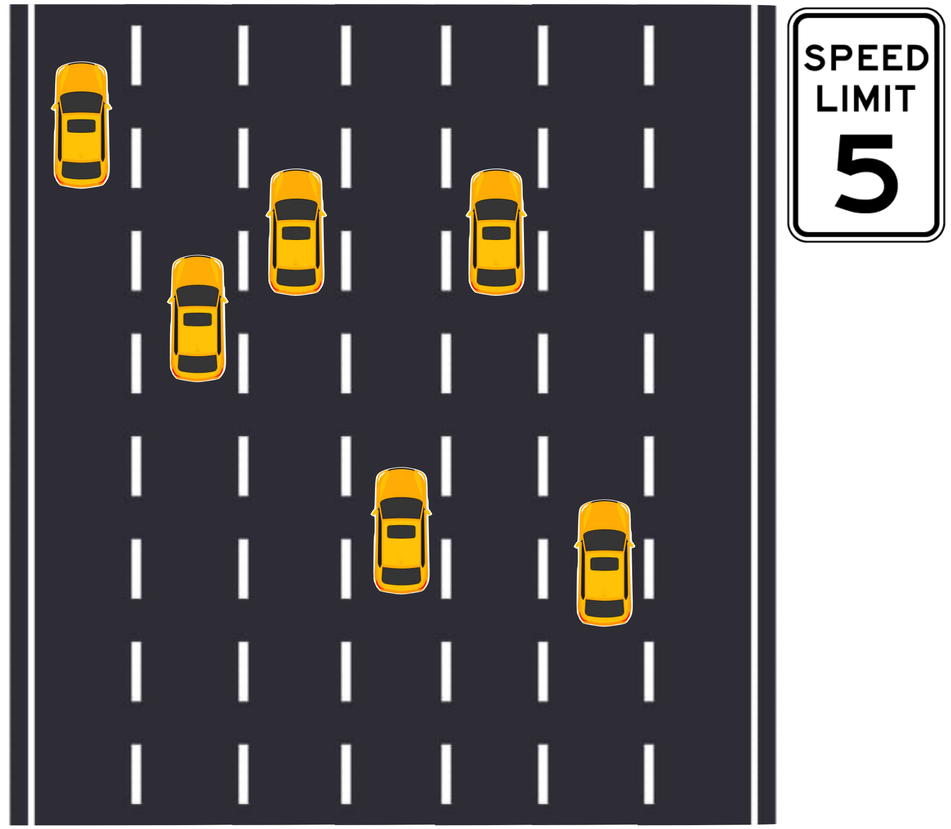


Why wouldn’t you always use parallel processing?

Serial

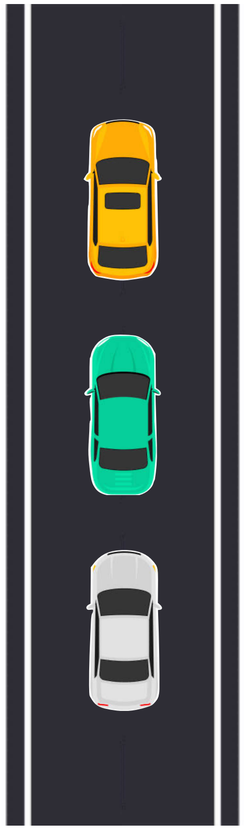


Parallel



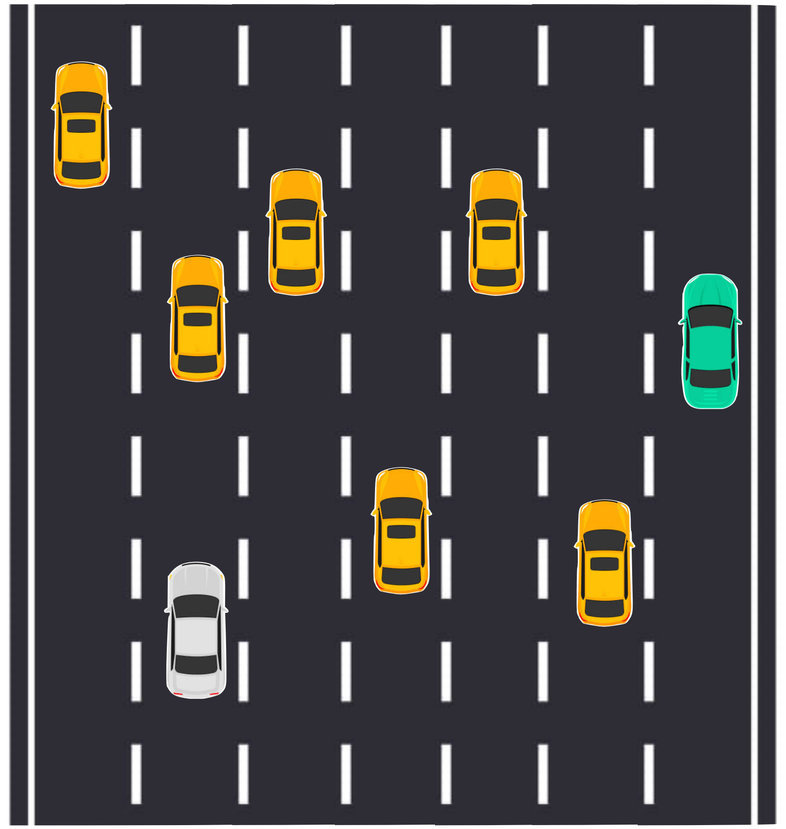
Why wouldn’t you always use parallel processing?

Serial



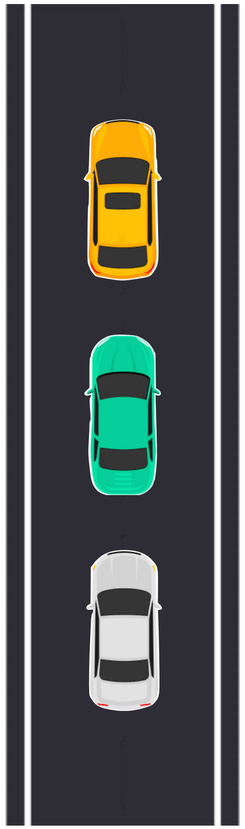
What if we have *contingent* processes?

Parallel



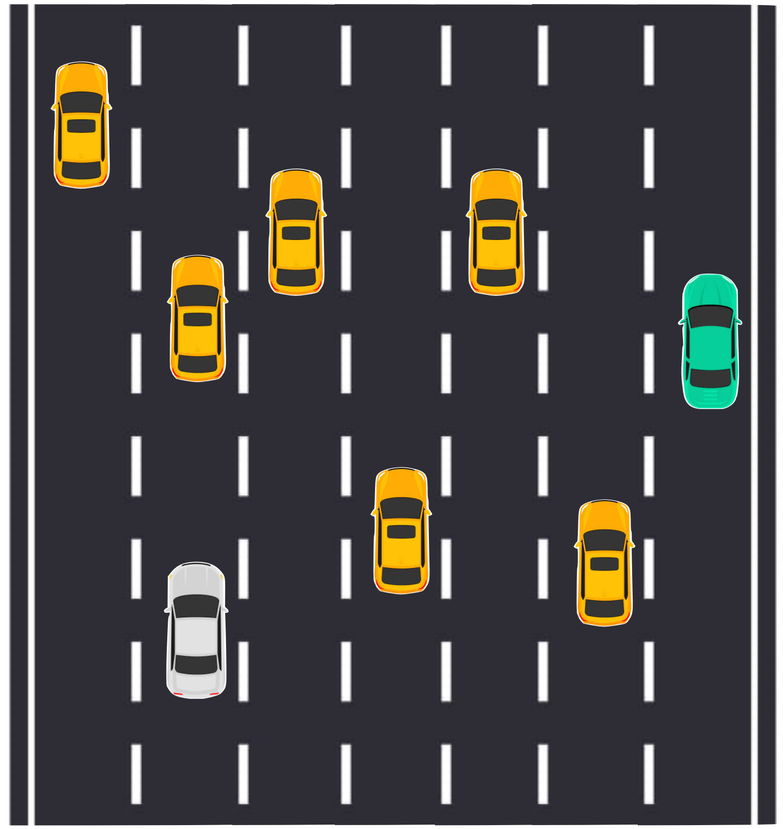
Why wouldn’t you always use parallel processing?

Serial



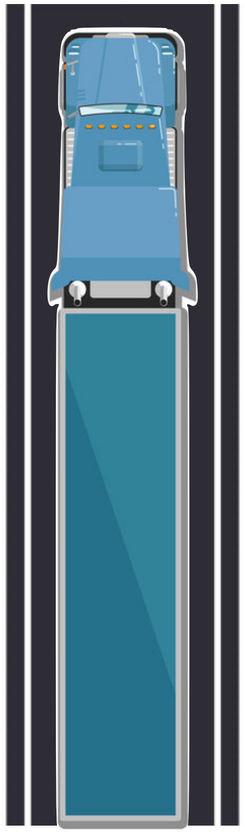
How do we get all these cars on and off efficiently?

Parallel



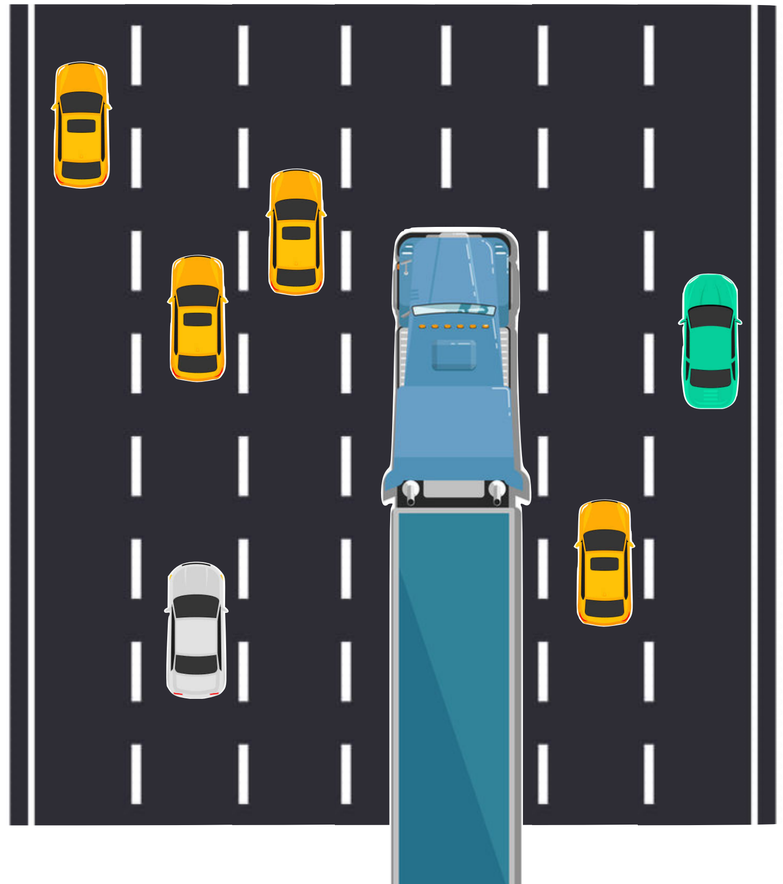
Why wouldn’t you always use parallel processing?

Serial



Serial process lines can often fit bigger trucks, more complex tasks

Parallel



Different computing problems, different solutions

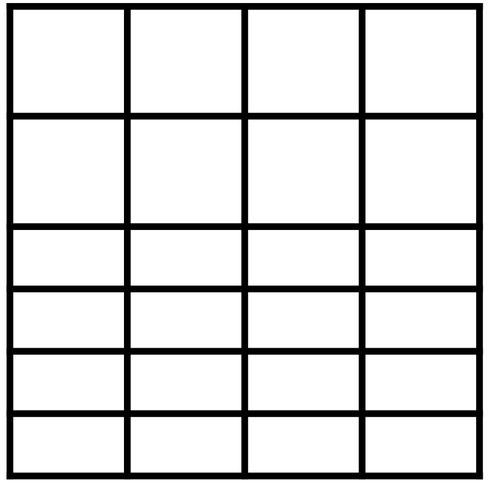
- Some computing problems are easily *parallelizable*
  - Image and video decoding, many AI tasks, some data tasks
- Some computing problems are more naturally *serial*
  - Contingent processes, simple processes with little practical benefit to complex organizational efforts, many kinds of user interaction

CPUs and GPUs are generally targeted at these two different spaces

- The CPU has a small number of **very fast**, **very big**, driving lanes (**processing cores**)
- The GPU has many **slower, smaller, less capable** driving lanes (processing cores)
  - But has **so many** it makes up for lack of capability through brute force
- Other specialized processing units are even more specialized, able to take only a very particular kind of car going in a very particular direction (NPU/TPUs, media encoders, etc)

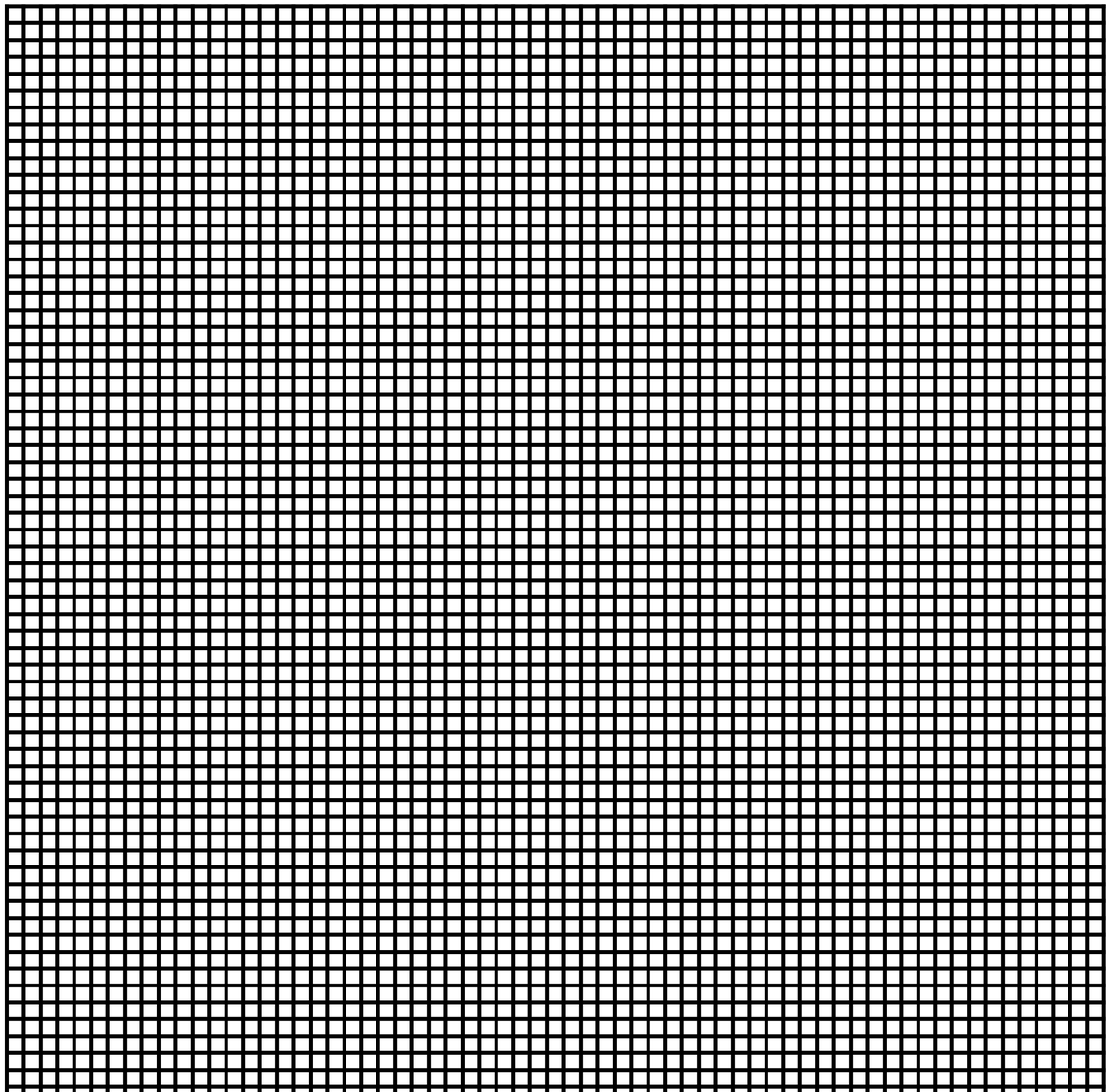
Intel i9-14900K

8 Performance Cores, 16 Efficiency Cores



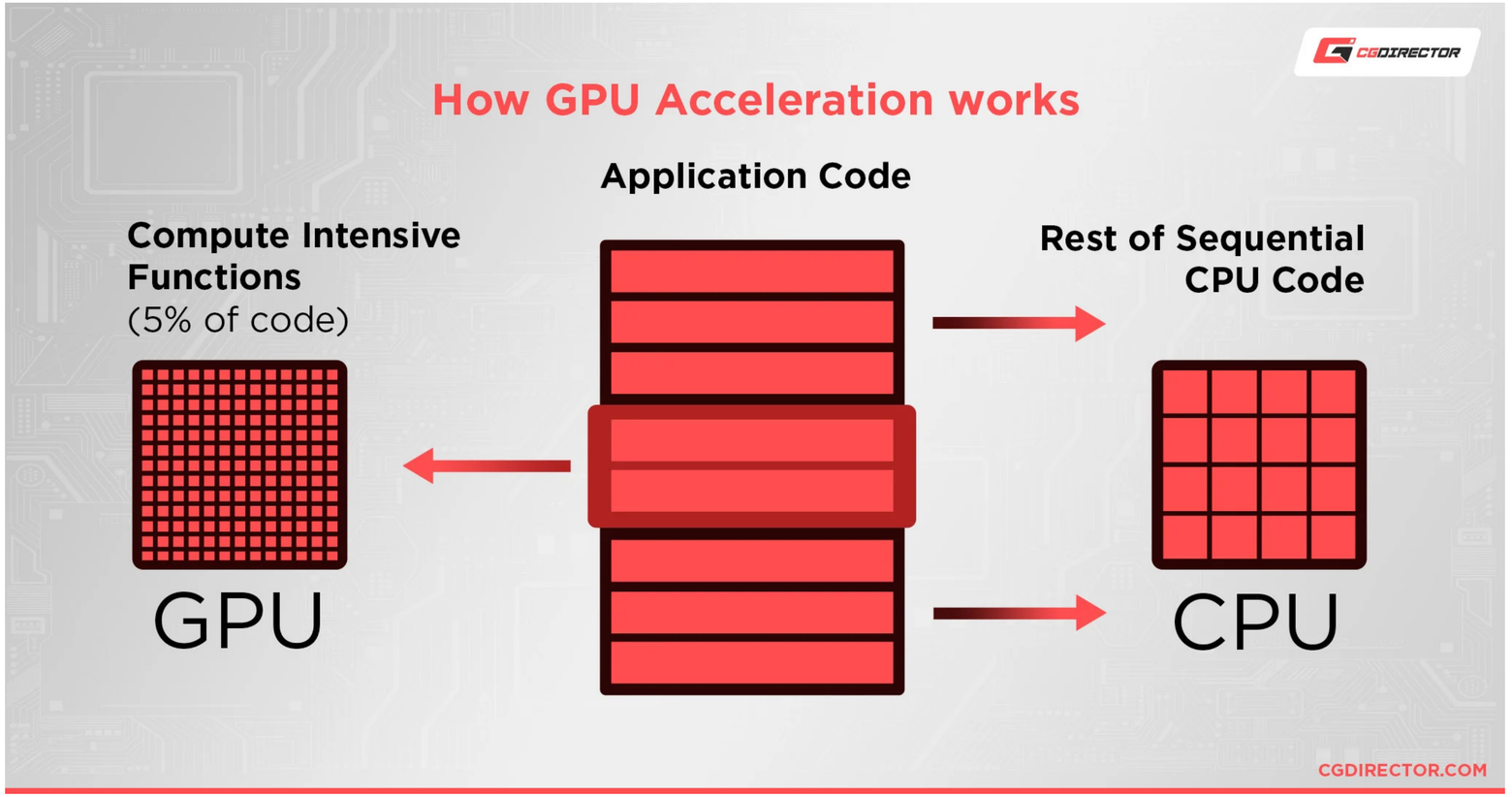
Nvidia GeForce GTX 5090

21,760 CUDA Cores



In practice

- The CPU controls the “computer”, while handing off special tasks that are particularly valuable to happen at greater speed in parallel to specialized processing units (like the GPU)



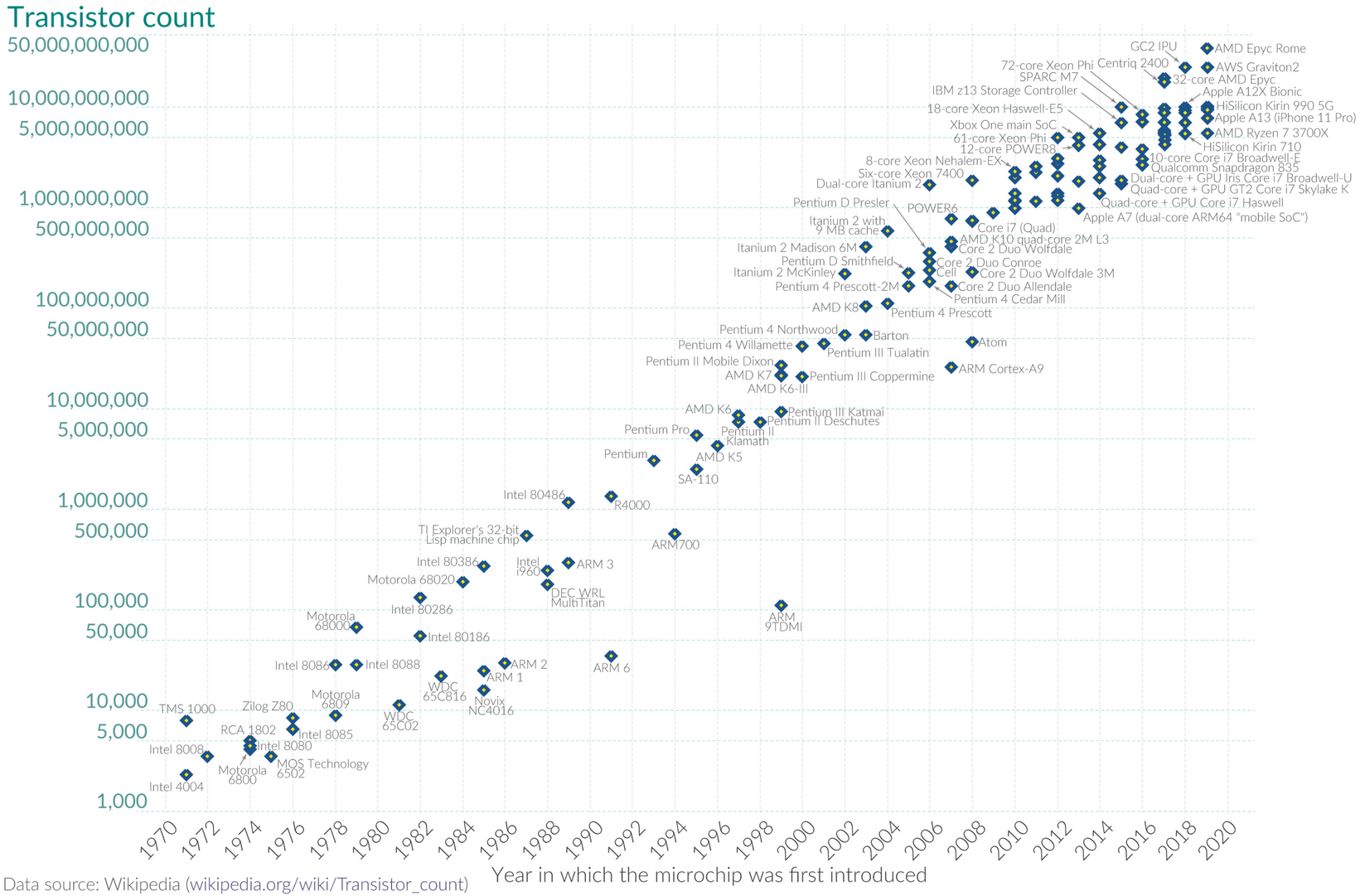
Computers get more powerful over time

- Moore’s Law: the number of transistors doubles every two years
- This process, however, is not automatic
- It is **hard** to make a processor, requiring exceptionally specialized equipment and technology
- Speed increases come from **die shrinks** along with design optimizations

Moore’s Law: The number of transistors on microchips doubles every two years

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world’s largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

Software

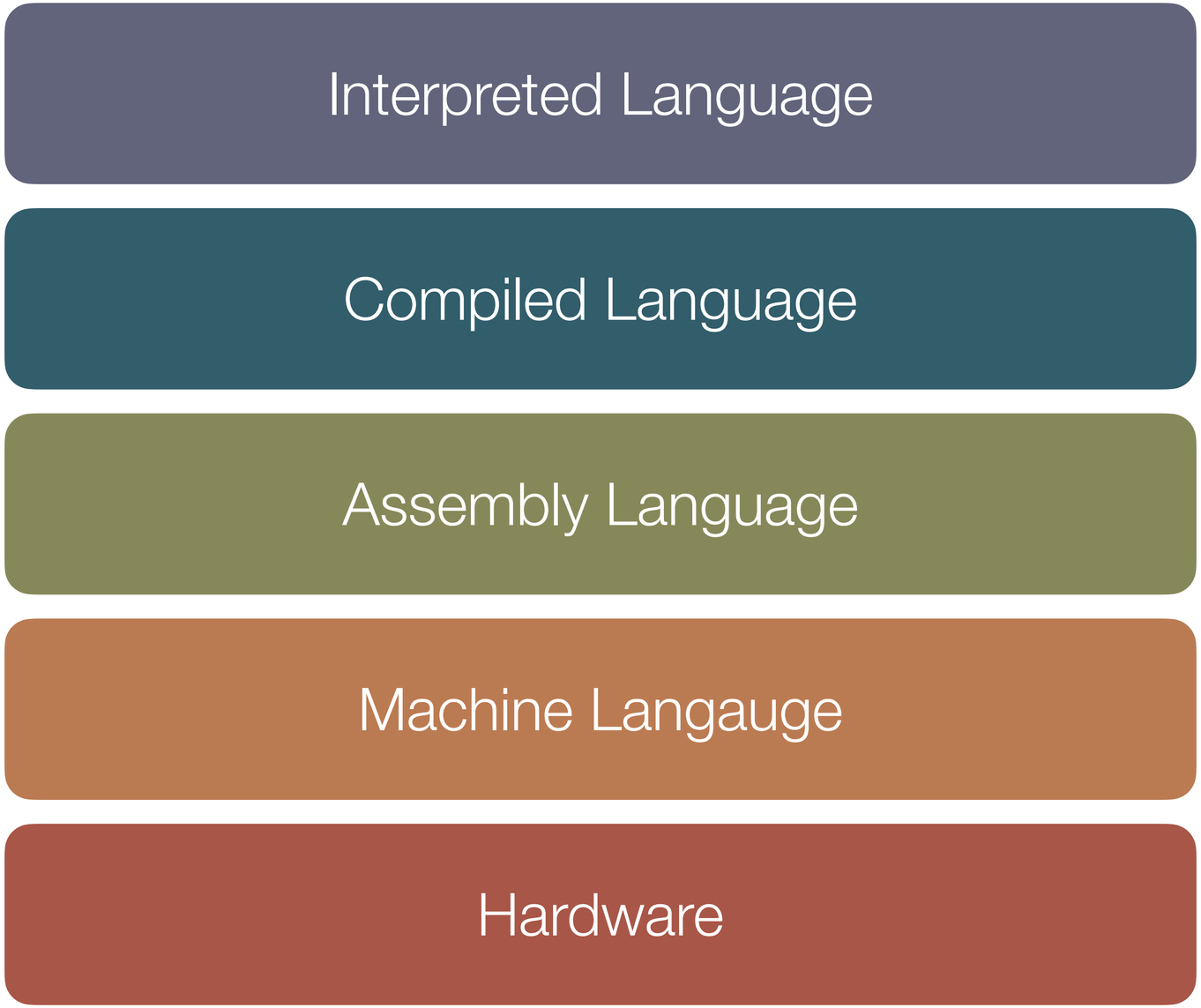
Programming

Programming

- How we tell a computer what we want it to do: a sequence of commands and instructions for the computer to execute
- Hardware takes a very limited number of commands: fundamentally, it all breaks down to manipulating little 0s and 1s
- We program a computer by writing **programs** and calling **libraries** and **APIs** within programming environments
  - **Algorithm** - a process or series of rules for a computer to follow
  - **Computer Program** - takes certain inputs and returns certain outputs

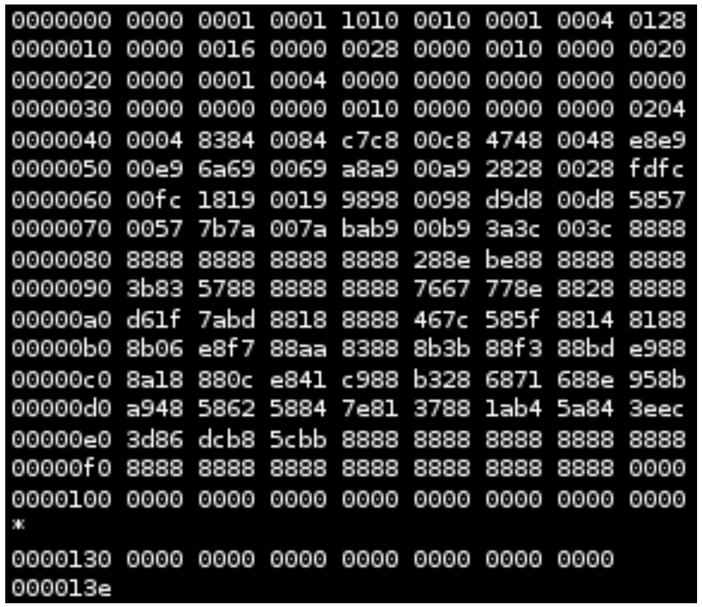
An exceptionally simple program

```
15 library(tidyverse)
16 library(osmdata) # package for working with streets
17 library(showtext) # for custom fonts
18 library(ggmap)
19 library(sf)
20 #library(rvest) #not needed here
21
22 #you can find all the available features of OSM here:
23 #https://wiki.openstreetmap.org/wiki/Map_features
24 #or here (can be slow, uncomment to see):
25 available_features()
26
27 #Also use openstreetmap.org to inspect features
28
29 available_tags("highway")
30
31 #getbb = "Get Bounding Box"
32 srqbounds<-getbb("Onondaga County New York")
33
34 #Now we're going to "pipe" commands together to build our map
35 #Piping is another way to pass commands directly to other commands
36 #Essentially: get our bounding box, <pass>, build an openstreemap query (opq), <pass>,
37 #add features, <pass>, return query as formatted SF object
38
39 srq_big_streets <- getbb("Onondaga County New York") %>%
40   opq() %>%
41   add_osm_feature(key = "highway",
42                   value = c("motorway", "trunk", "primary", "motorway_link", "trunk_link", "primary_lin
43   osmdata_sf()
44
45 #Now let's see what we have!
46 srq_big_streets
47
48 #Remember - the core components of graphing.  Points, Lines, Polygons (and iterations on those)
49   # osm_lines, osm_points, osm_polygons
50
51 #Time to do some plotting with our old friend, GGPlot
52 #geom_sf - plot "simple feature" objects (the kind used in GIS)
53
54 ggplot() +
55   geom_sf(data = srq_big_streets$osm_lines,
56           inherit.aes = FALSE,
57           color = "black",
58           size = 1)
59
60 #Can play with aesthetics, like always
61
62 ggplot() +
63   geom_sf(data = srq_big_streets$osm_lines,
64           inherit.aes = FALSE,
65           color = "purple",
66           size = 8)
```


Interpreted Language
Compiled Language
Assembly Language
Machine Langauge
Hardware

Machine code

Binary



Hexdecimal

Assembly

```
MONITOR FOR 6802 1.4          9-14-80  TSC ASSEMBLER  PAGE    2


 C000                    ORG    ROM+$0000 BEGIN MONITOR
 C000 8E 00 70  START    LDS    #STACK

                *************************************
                * FUNCTION: INITA - Initialize ACIA
                * INPUT: none
                * OUTPUT: none
                * CALLS: none
                * DESTROYS: acc A

 0013           RESETA  EQU    %00010011
 0011           CTLREG  EQU    %00010001

 C003 86 13     INITA   LDA A  #RESETA   RESET ACIA
 C005 B7 80 04          STA A  ACIA
 C008 86 11             LDA A  #CTLREG   SET 8 BITS AND 2 STOP
 C00A B7 80 04          STA A  ACIA

 C00D 7E C0 F1          JMP    SIGNON    GO TO START OF MONITOR

                *************************************
                * FUNCTION: INCH - Input character
                * INPUT: none
                * OUTPUT: char in acc A
                * DESTROYS: acc A
                * CALLS: none
                * DESCRIPTION: Gets 1 character from terminal

 C010 B6 80 04  INCH    LDA A  ACIA      GET STATUS
 C013 47                ASR A            SHIFT RDRF FLAG INTO CARRY
 C014 24 FA             BCC    INCH      RECIEVE NOT READY
 C016 B6 80 05          LDA A  ACIA+1    GET CHAR
 C019 84 7F             AND A  #$7F      MASK PARITY
 C01B 7E C0 79          JMP    OUTCH     ECHO & RTS

                *************************************
                * FUNCTION: INHEX - INPUT HEX DIGIT
                * INPUT: none
                * OUTPUT: Digit in acc A
                * CALLS: INCH
                * DESTROYS: acc A
                * Returns to monitor if not HEX input

 C01E 8D F0     INHEX   BSR    INCH      GET A CHAR
```

A program in C (“mid” level; compiled)

```c
#include <stdlib.h>
#include <stdio.h>
#include <stdbool.h>

struct stats { int count; int sum; int sum_squares; };

void stats_update(struct stats * s, int x, bool reset) {
    if (s == NULL) return;
    if (reset) * s = (struct stats) { 0, 0, 0 };
    s->count += 1;
    s->sum += x;
    s->sum_squares += x * x;
}

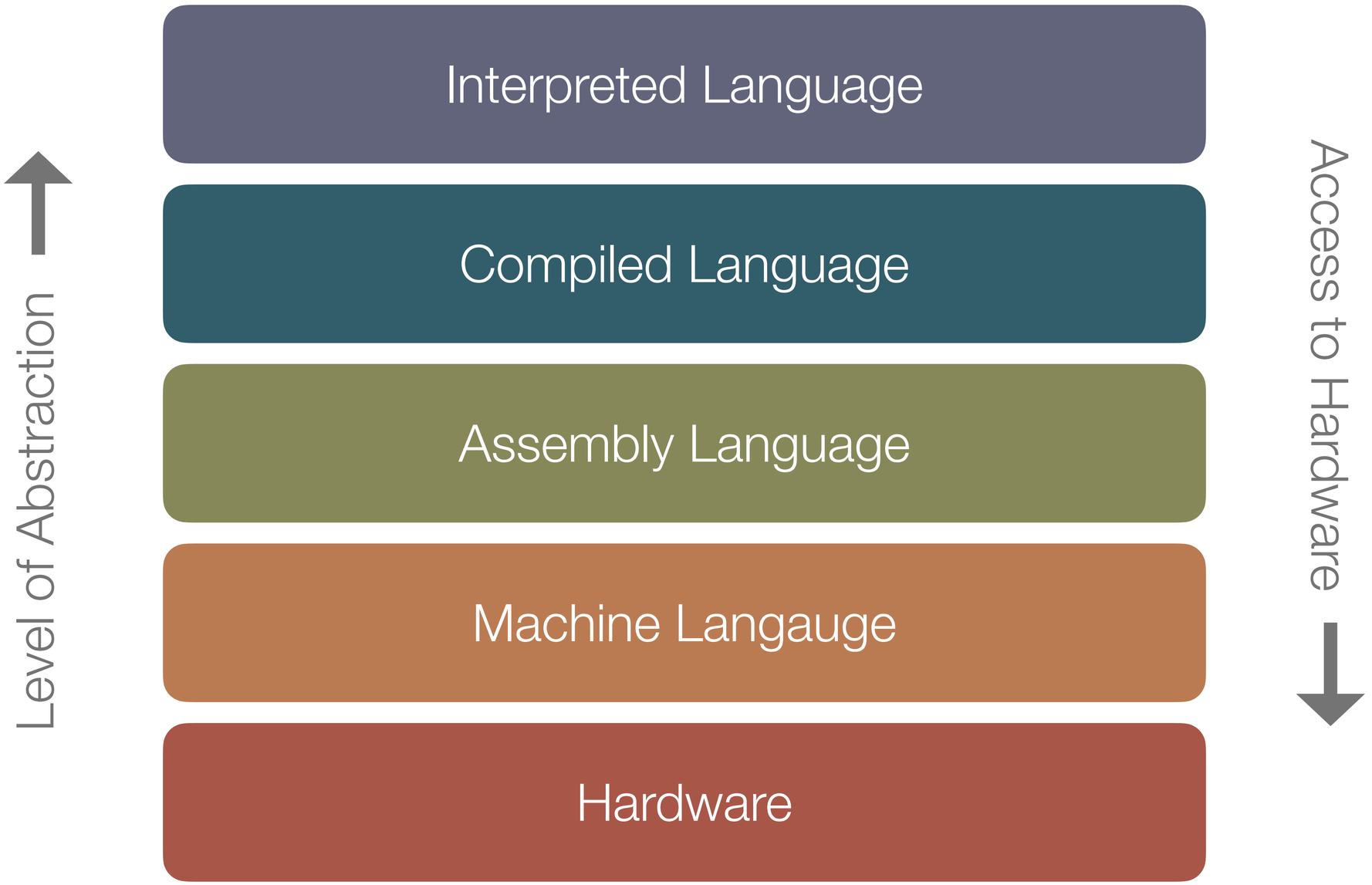
double mean(int data[], size_t len) {
    struct stats s;
    for (int i = 0; i < len; ++i)
        stats_update(&s, data[i], i == 0);
    return ((double)s.sum) / ((double)s.count);
}

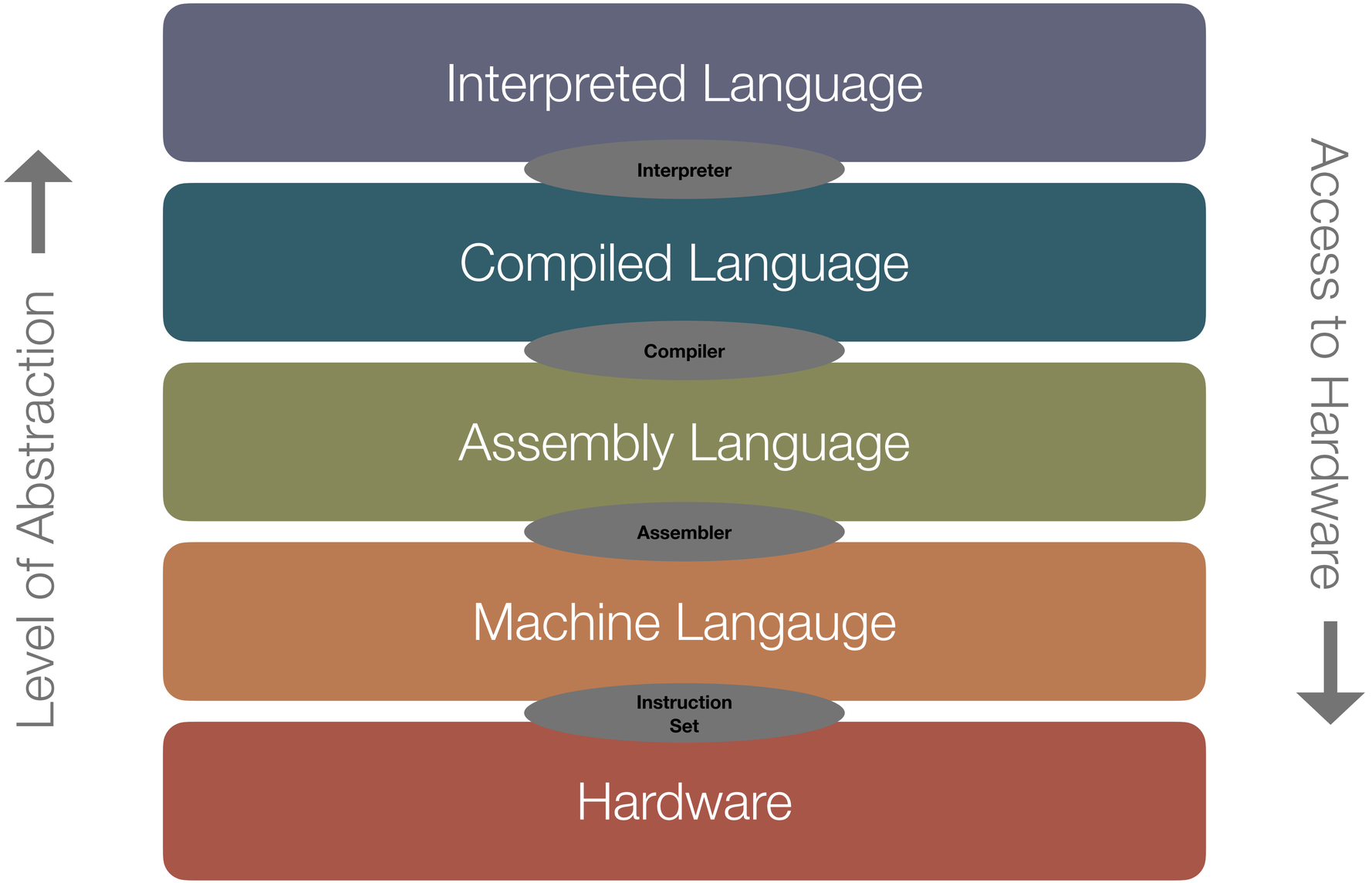
void main() {
    int data[] = { 1, 2, 3, 4, 5, 6 };
    printf("MEAN = %lf\n", mean(data, sizeof(data) / sizeof(data[0])));
}
```

this code calculates a mean

A program in R ("high" level, interpreted)

```
15  library(tidyverse)
16  library(osmdata) # package for working with streets
17  library(showtext) # for custom fonts
18  library(ggmap)
19  library(sf)
20  #library(rvest) #not needed here
21
22  #you can find all the available features of OSM here:
23  #https://wiki.openstreetmap.org/wiki/Map_features
24  #or here (can be slow, uncomment to see):
25  available_features()
26
27  #Also use openstreetmap.org to inspect features
28
29  available_tags("highway")
30
31  #getbb = "Get Bounding Box"
32  srqbounds<-getbb("Onondaga County New York")
33
34  #Now we're going to "pipe" commands together to build our map
35  #Piping is another way to pass commands directly to other commands
36  #Essentially: get our bounding box, <pass>, build an openstreemap query (opq), <pass>,
37  #add features, <pass>, return query as formatted SF object
38
39  srq_big_streets <- getbb("Onondaga County New York") %>%
40    opq() %>%
41    add_osm_feature(key = "highway",
42                    value = c("motorway", "trunk", "primary", "motorway_link", "trunk_link", "primary_link")) %>%
43    osmdata_sf()
44
45  #Now let's see what we have!
46  srq_big_streets
```

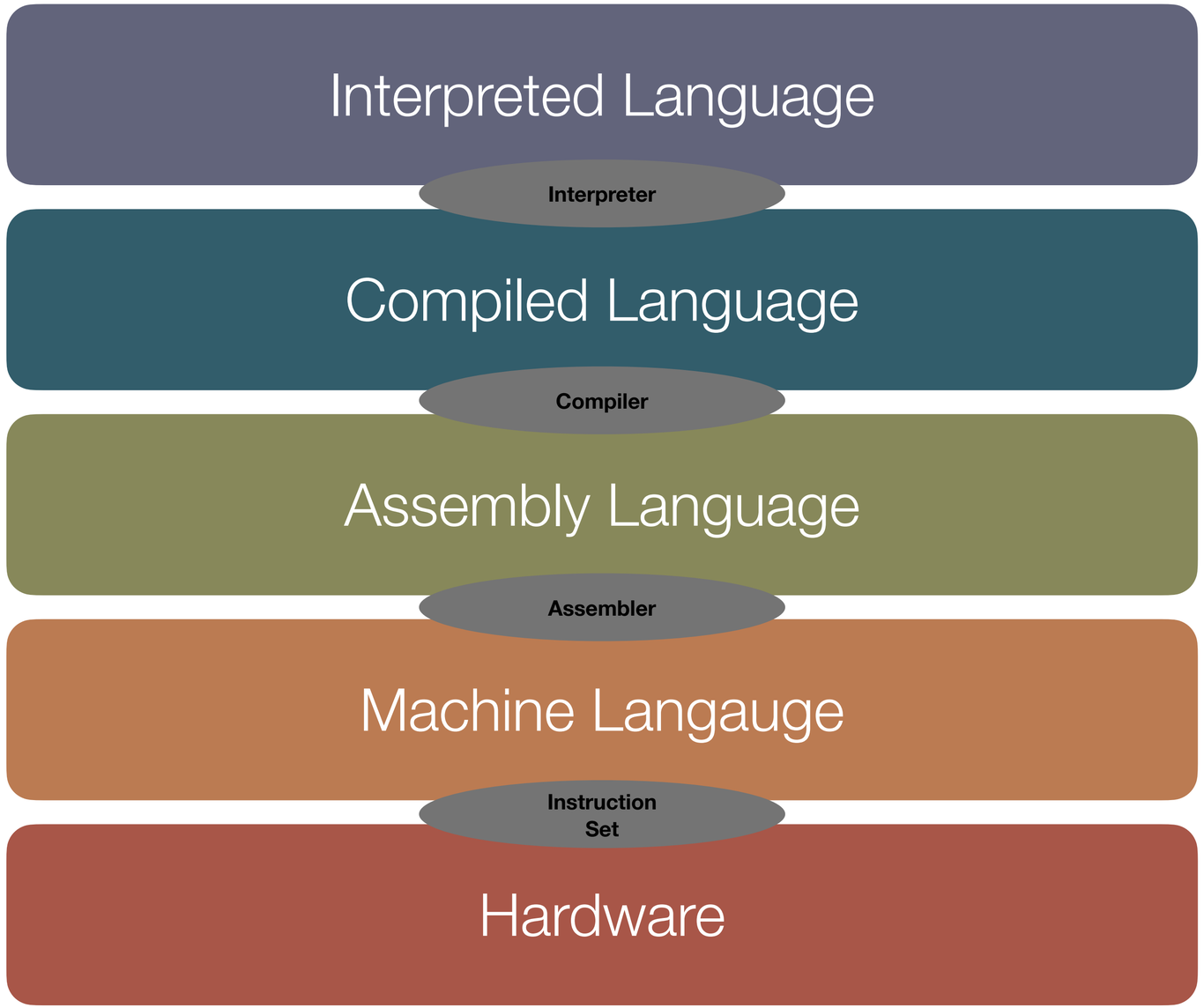

Interpreted Language
Compiled Language
Assembly Language
Machine Langauge
Hardware
Level of Abstraction
Access to Hardware



*heavily simplified

Langauge tradeoffs

- Higher level languages and APIs (application programming environments), in general on average, **increase** human readability at a cost of **speed** and **hardware access**
- **“Understand your stack”** - understand all the **dependencies** your code has
- Dependencies - not just hardware levels, but software levels and dependences as well
  - Software can call other software to perform functions



*heavily simplified

What do we do when we compute?

- We’re - to a first approximation - just doing the Turing machine thing
  - **Command** (through software or user interface)
    - a computer to **perform an operation** (analyze data, play a movie, edit a picture)
      - on **data** (tabular data, video file, important cat .gif)
- **How do we command our computers to do things?**

so much power to show us cute pictures of kittehs

The great computing divorce

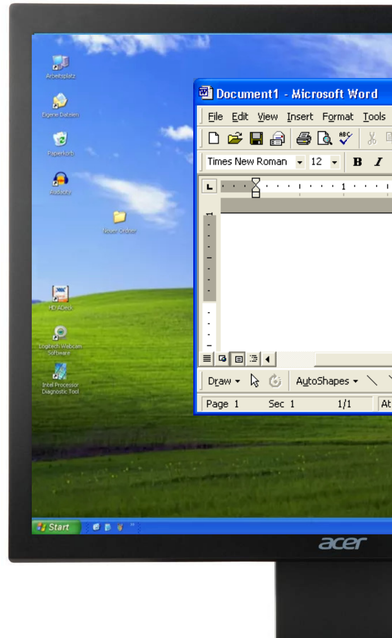
- How you probably like to use a computing device

- How you probably use a computing device for work

The great computing divorce

- Snazzy and colorful!
- Touch the screen!
- Apps!
- One application at a time
- File system? What file system?
- It’s the 2020s!

- Boxy desk-bound
- Click the mouse
- Big applications
- Lots of windows
- Files in folders
- Y2k is calling



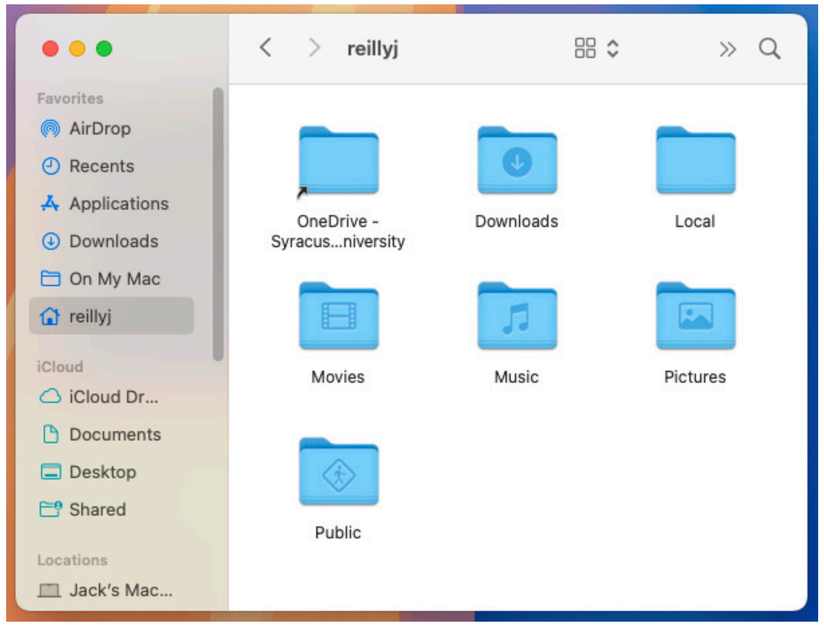
Within the work computer side of the divide, however, lies another divide

A deeper, darker past . . .

Two interface paradigms

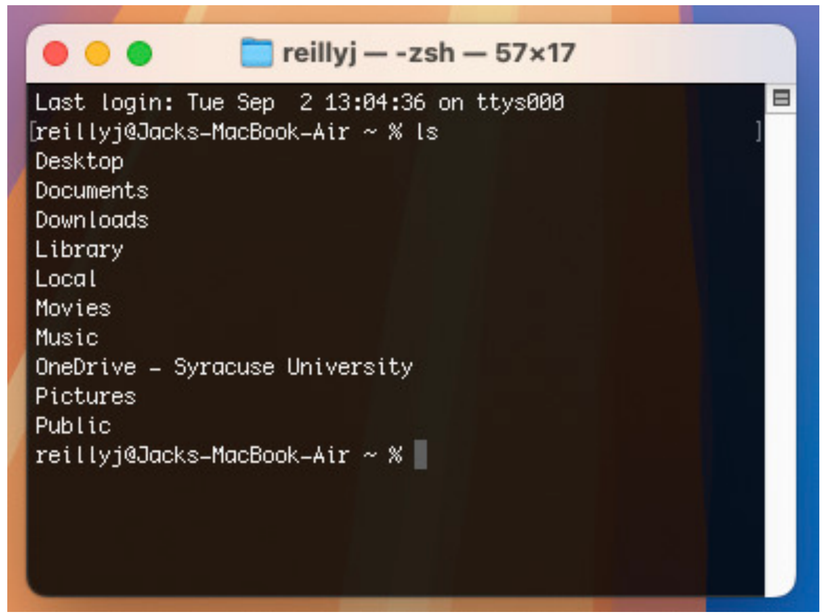
GUIs

Graphical User Interfaces



CLIs

Command Line Interfaces



This is where the heart of technical computing (still) resides

- UNIX, the shell, and the command line
  - UNIX is an operating system used for mainframes originally developed in 1969
- Its derivatives live on today, at the heart of (most) computing systems
  - MacOS, Linux, iOS, Android, ChromeOS, Kindles, even your watch . . . all UNIX
    - Even Windows has a Linux subsystem
- The vast majority of people that use their computers for code live a great portion of their computing lives in software inspired by the 1970s

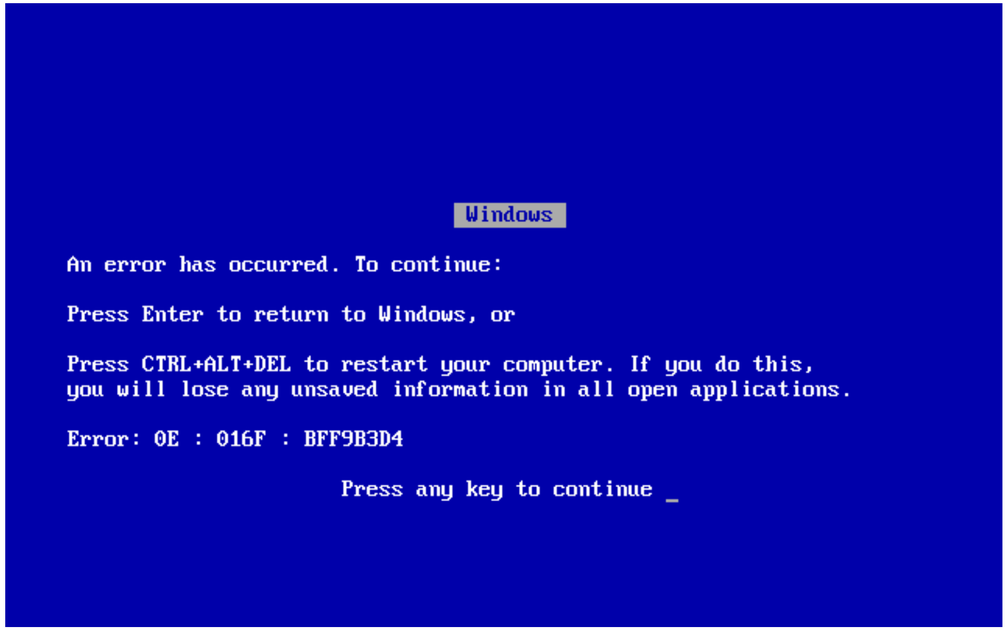
IBM 1401 mainframes, 1969
https://www.computerhistory.org/revolution/mainframe-computers/7/166/663

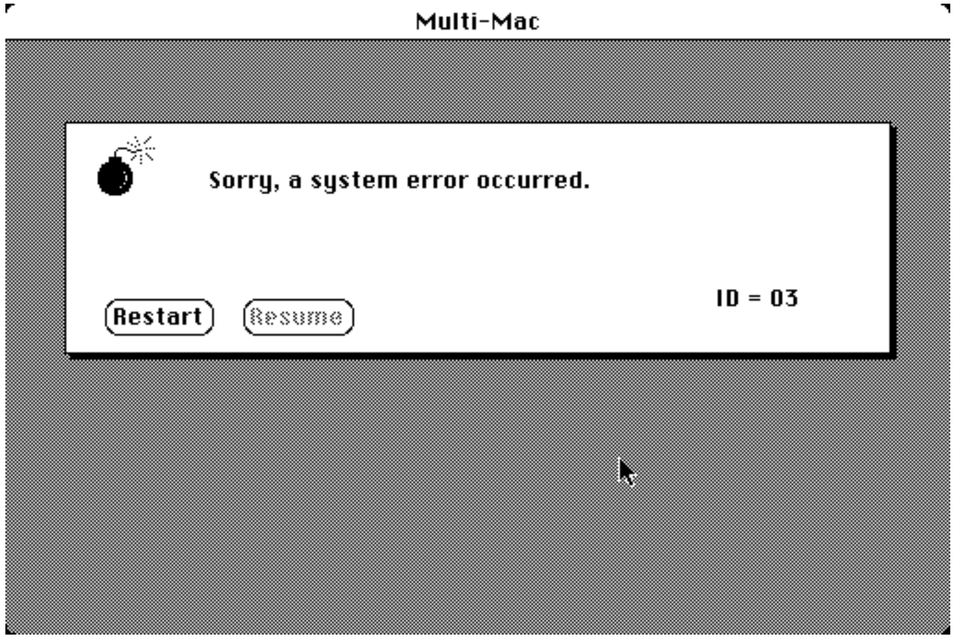
Why would this happen?

- Well, there was a push to make computers smaller and more personal

- And, for a while, smaller meant that computers couldn’t handle UNIX-like operating systems

Which came with a lot of tradoffs





But then, a funny thing happened

- Computers got smaller still, but powerful enough run UNIX at speed

Runs UNIX!!

The **Good** News

We're **not** going to exclusively use the shell (CLI)

The **Complicating** News

but the kinds of work we do is going to be heavily influenced by it, and it's going to be useful to drop down to it every now and then

Two models of work computing

- **Office** model
  - Formatted documents, application-centric files
  - Intermediate products are cut and pasted, often without link backs or references to source
  - Changes are tracked by application
  - Editing window = formatted window (WYSIWYG)
- **Engineering** model
  - Open, plain text documents, readable with many applications
  - Intermediate products produced via code, linked with reference to source
  - Changes tracked outside files, at the project level
  - Final outputs programmatically *assembled* to final document

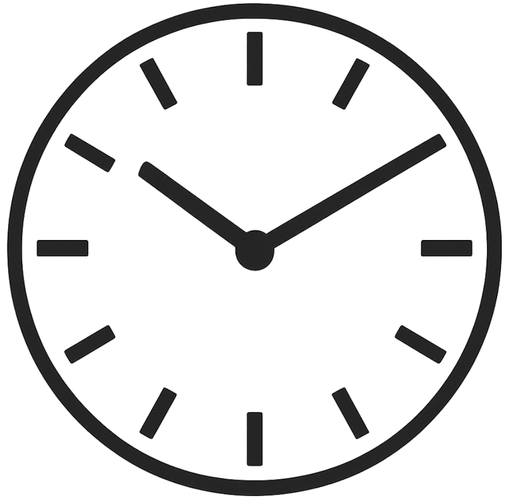
Oooh, a new way to use the computer! Is the engineering model fun?

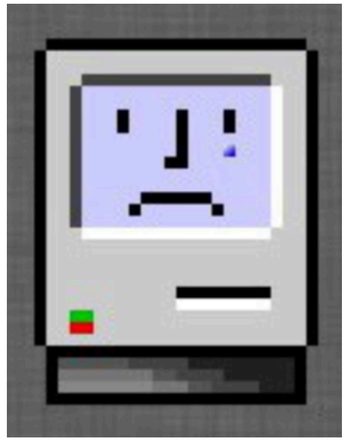
Oooh, a new way to use the computer! Is the engineering model fun?

- Well, no. At times, it can feel more like this.

Oh, OK, but can we make it better?

- Yes! But it takes a long time and a lot of practice.





Uh, well, why do we do it?

- Two **very, very good** reasons

CONTROL

- and -

Reproducibility

(also, it's weirdly fun when you get used to it)

Also, once you get used to it, you may like it

- And it kinda starts to feel like this

CONTROL

- and -

Reproducibility

So why do we care so much about control and reproducibility?

- Research is difficult.
- It’s easy to make small mistakes that become large errors.
- Doing it badly - or fraudulently - is too easy.
- And even when we don’t make mistakes, and put in our best effort, sometimes we’re . . . just wrong

The Replication Crisis

- A genuinely surprisingly number of scientific and policy articles do not **replicate**
  - **Replicate:** obtain the same, or a similar, finding upon conducting the **same study**
- **Behavioral sciences** are particularly at risk (pesky humans)

News | Published: 27 August 2015

Over half of psychology studies fail reproducibility test

Monya Baker

Nature (2015) | Cite this article

53k Accesses | **123** Citations | **1347** Altmetric | Metrics

Largest replication study to date casts doubt on many published positive results.

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Copyright: © 2005 John P. A. Ioannidis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

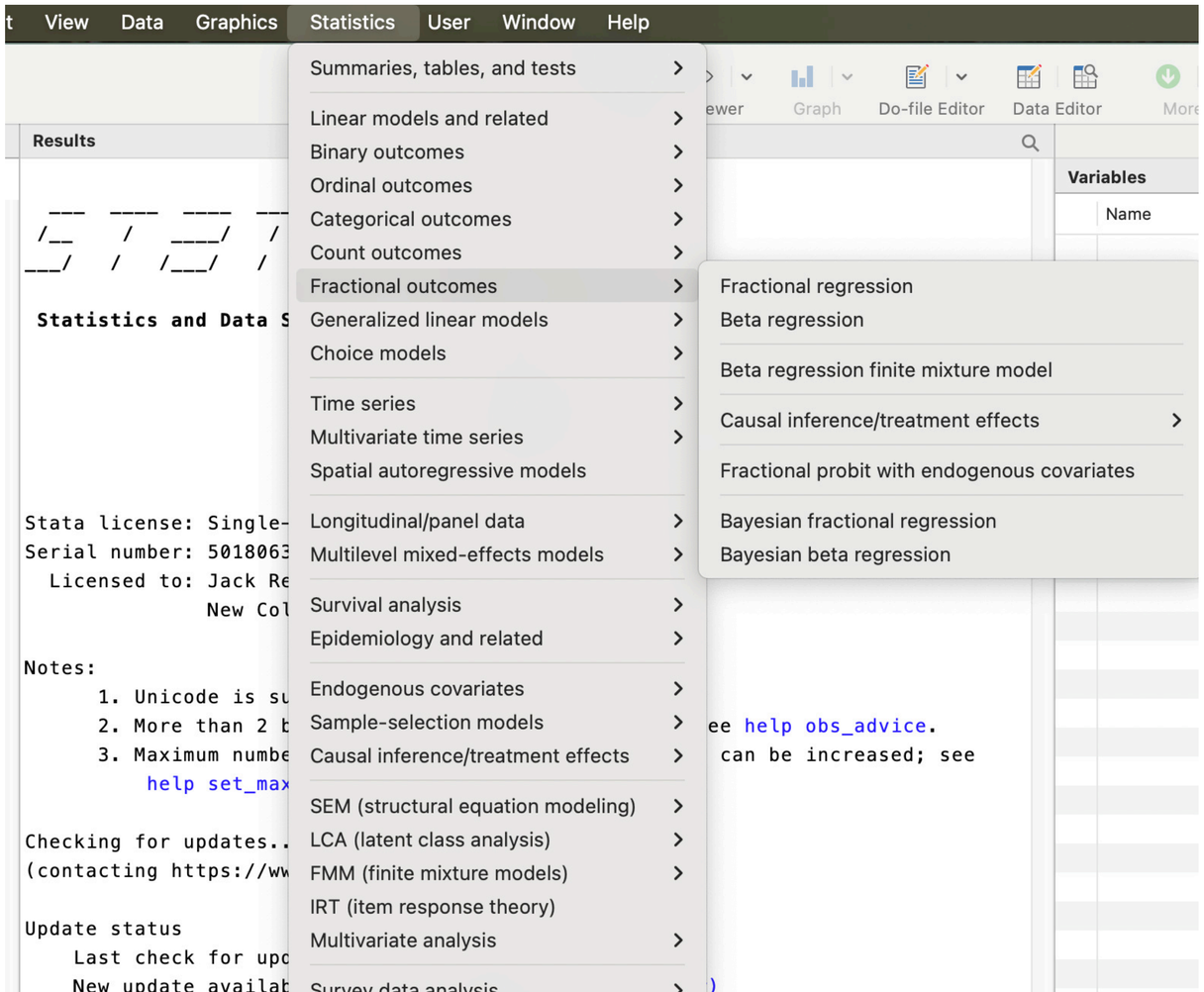
DOI: 10.1371/journal.pmed.0020124

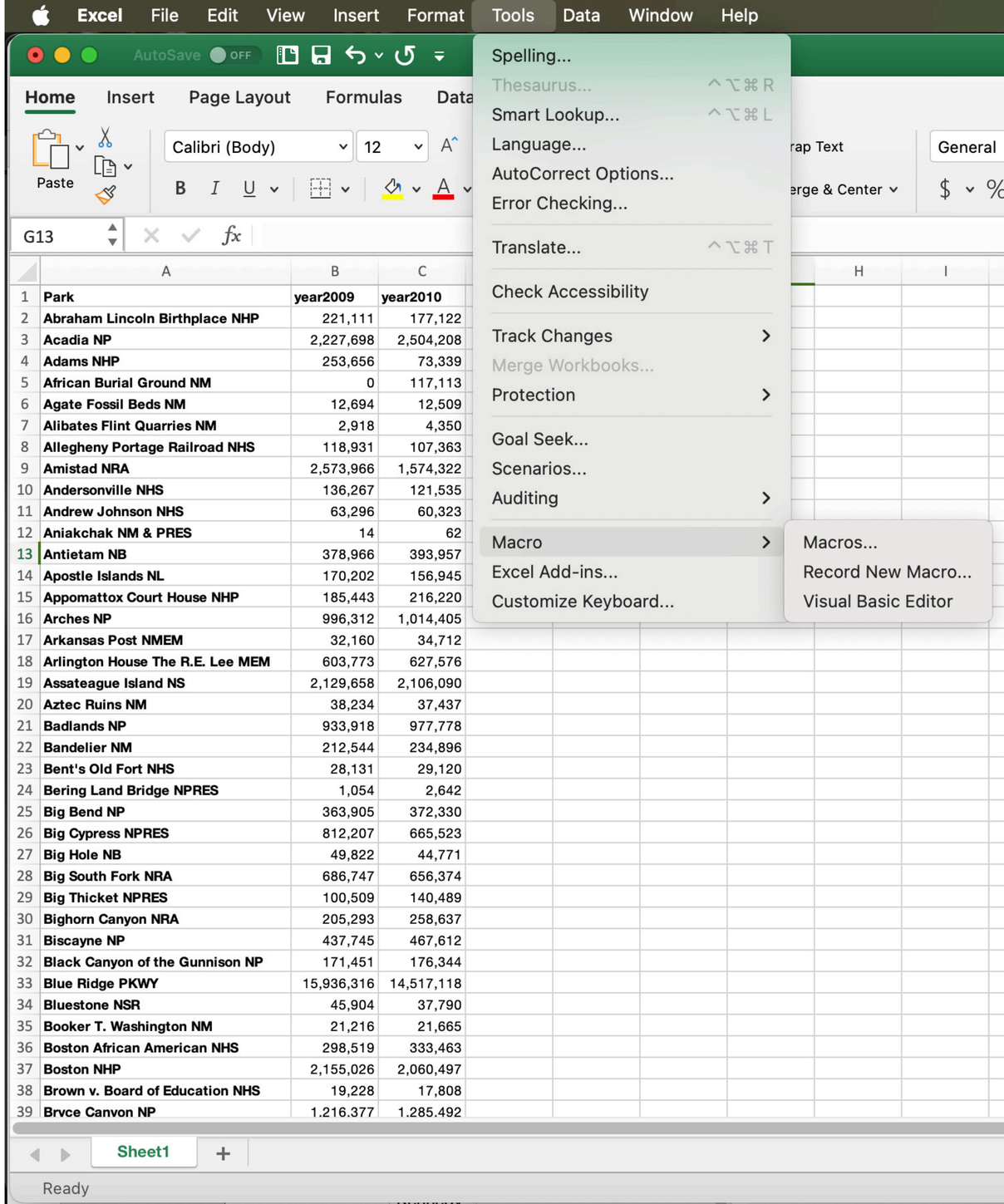
So what do we do?

- **Control:**
  - ensure we know *exactly* what is happening at all points in our study
- **Reproducibility:**
  - ensure that we (and others) can precisely reproduce all of our findings, from the same data, as easily as possible

The weaknesses of the office model

- It’s just easy to make mistakes





The weaknesses of the office model

- The tools of the office model are ill suited to our task
  - Sawitzki (1994)

Testing Numerical Reliability of Data Analysis Systems

Günther Sawitzki
StatLab Heidelberg
Im Neuenheimer Feld 294
6900 Heidelberg

Summary
We discuss some typical problems of statistical computing, and testing strategies for the testing data anlysis systems. For one special area, regression, we illustrate these problems and strategies.

Keywords
Regression, Statistical computing, Software quality control

Introduction
Reliable performance is the main reason to use well-established statistical packages. One aspect of this reliability is correctness of the results. The quest for correctness must take into account practical limitations. Any computer is finite, and reliability necessarily will have its limitations: with finite resources, we cannot perform any calculation on any data set with arbitrary precision. It is not the question whether there are limitations of a system. The question is where these limitations are and how the system behaves at these limitations.

For any application area, it is crucial that the limitations of a system are wide enough to include the usual data sets and problems in this area. If the limitations are too restricting, the system simply is not usable. If the limits are known, some rough considerations usually will show whether they are wide enough. Failing systems should be discarded. Once the limits are well known, we can take special precautions not to exceed these limits for ordinary cases. Problems occur when limits are not known, or cannot be inferred from the behaviour of the program. The system can have several ways to fail if its limits are exceeded:

– fail catastrophically
– diagnose the problem and fail gracefully
– diagnose the problem and recover
– provide a plausible side exit
– run into an arbitrary solution

We will not like the first way. A catastrophic failure under certain conditions will force us to find out the limits. So we are doing the work the software developers could have done, and eventually we will ask whether the license fee we have paid is more like a license fine[1]. Not particularly dangerous, but expensive. We have to accept the second way of failure. Since we are working with finite machines and finite resources, we have to accept limitations, and having a clear failure indication is the best information we can hope for. The third way, failure diagnostics and recovery, may be acceptable depending on the recovery. If it is fail-safe, we may go along with it. If the recovery is not fail-safe, it is the same as the worst solution: The Ultimate Failure: run into an arbitrary solution, the worst kind of failure. If the arbitrary solution is catastrophic, or obviously wrong, we are back in the first case. We will notice the error. We have to spend work we could use for better purposes, but still we are in control of the situation. But the solution could as well be slightly wrong, go unnoticed, and we would work on wrong assumptions. We are in danger of being grossly misled.

The Design of a Test Strategy
Data analysis systems do not behave uniformly over all data sets and tasks. This gives the chance and the task to choose

[1]Special thanks to P. Dirschedl for pointing out the difference between a license fee and a license fine.

The weaknesses of the office model

- The tools of the office model are ill suited to our task
  - Sawitzki (1994)

Testing Numerical Reliability of Data Analysis Systems

Günther Sawitzki
StatLab Heidelberg
Im Neuenheimer Feld 294
6900 Heidelberg

Summary
We discuss some typical problems of statistical computing, and testing strategies for the testing data anlysis systems. For one special area, regression, we illustrate these problems and strategies.

Keywords
Regression, Statistical computing, Software quality control

EXCEL

Microsoft EXCEL, and other spreadsheet programs, claim to be systems for data analysis. They should be taken by their word and included in this test. EXCEL 4.0 failed to calculate even the summary statistics (task II.C). It returned errors starting at the first significant digit. It was not even stable enough to produce the same standard deviation for X and ROUND using the standard EXCEL functions AVERAGE(), VAR() and STDEV().

	X	BIG	LITTLE	HUGE	TINY	ROUND
mean	5	99999995	0.99999995	5E+12	5E-12	4.5
var	7.5	6	8.88178E-16	7.5E+24	7.5E-24	7.5
stddev	2.73861279	2.449489743	2.98023E-08	2.73861E+12	2.73861E-12	2.738612788

These results are most surprising, particularly on the Macintosh. The Macintosh operation system, going beyond more classical systems, supports a full extended precision IEEE arithmetics as part of the operating system. Using these operating system facilities, even the most crude algorithm would give a correct value for variance and standard deviation in our test.

Although EXCEL did not manage to calculate the variance correctly, they got correct results for all the correlations (task II.D), using CORREL(). EXCEL gave satisfactory results for the polynomial regression problem (task IV.A) using the "Regression" analysis tool packed with EXCEL (residual sum of square: 1.093E-13; integrated square error=0.049).

EXCEL failed to diagnose the singularity when regressing X on BIG and LITTLE (task IV.C)and gave a regression

	Coefficients	Standard Error	t Statistic	P-value
Intercept	-125829115	11828363.26	-10.637914	5.3396E-06
BIG	1.258291259	0.236567273	5.31895746	0.00071193
LITTLE	0	19379590.84	0	1

atistical packages. One aspect of this reliability is
account practical limitations. Any computer is finite,
urces, we cannot perform any calculation on any data
e limitations of a system. The question is where these

m are wide enough to include the usual data sets and
tem simply is not usable. If the limits are known, some
ough. Failing systems should be discarded. Once the
ed these limits for ordinary cases. Problems occur
our of the program. The system can have several ways

conditions will force us to find out the limits. So we
eventually we will ask whether the license fee we have
xpensive. We have to accept the second way of
urces, we have to accept limitations, and having a clear
d way, failure diagnostics and recovery, may be
along with it. If the recovery is not fail-safe, it is the
rary solution, the worst kind of failure. If the arbitrary
rst case. We will notice the error. We have to spend
of the situation. But the solution could as well be
mptions. We are in danger of being grossly misled.

nd tasks. This gives the chance and the task to choose

se fee and a license fine.

The weaknesses of the office model

- Surely, though, those were fixed, right?
  - McCullogh & Wilson (1999)

ELSEVIER

Computational Statistics & Data Analysis 31 (1999) 27–37

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 97

B.D. McCullough *, Berry Wilson

Federal Communications Commission, 445 12th St. SW, Room 2C-134, Washington, DC 20554, USA

Received 1 June 1998; received in revised form 1 December 1998

Abstract

The reliability of statistical procedures in Excel are assessed in three areas: estimation (both linear and nonlinear); random number generation; and statistical distributions (e.g., for calculating *p*-values). Excel's performance in all three areas is found to be inadequate. Persons desiring to conduct statistical analyses of data are advised not to use Excel. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: DIEHARD; ELV; Numerical accuracy; Software testing; StRD

1. Introduction

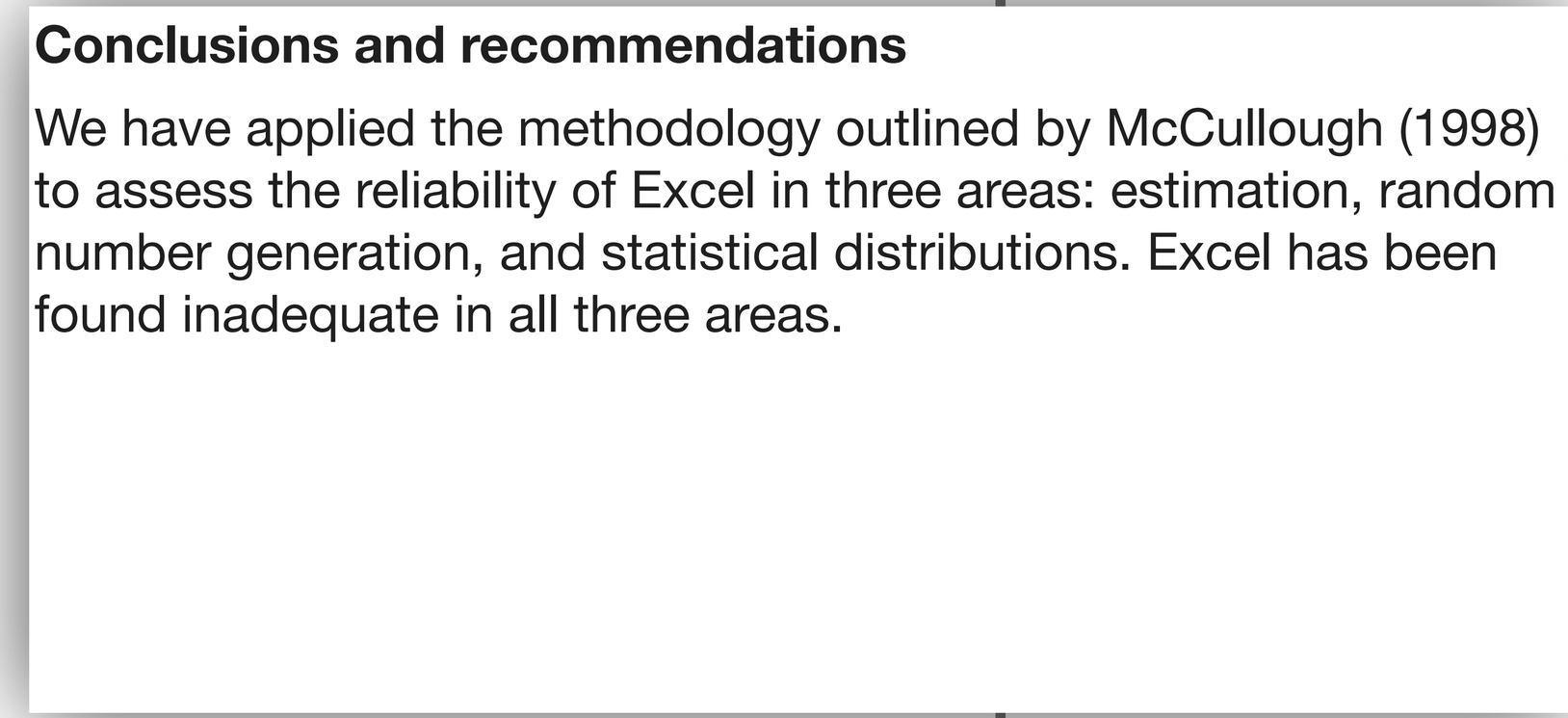
Sawitzki (1994b) documented the failure of many statistical packages to pass entry-level tests of accuracy known as the Wilkinson Tests (Wilkinson, 1985). The need exists to know how statistical packages fare on more substantial tests of numerical accuracy (Sawitzki, 1994a). Recently McCullough (1998) proposed a collection of intermediate-level tests which assesses the numerical reliability of a package in three areas: estimation (both linear and nonlinear); random number generation; and statistical distributions (e.g., for calculating *p*-values). Estimation is assessed via the *Statistical Reference Datasets* (StRD), which recently was released by the American "National Institute of Standards and Technology" (NIST). The output of the random number generator (RNG) is assessed using statistical tests for randomness. The accuracy of statistical distributions is assessed by comparing the results of Excel to those from a specialized package such as Knüsel's (1989) ELV program.

* Corresponding author
E-mail address: bmccullo@fcc.gov (B.D.McCullough)

0167-9473/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S 0167-9473(99)00004-3

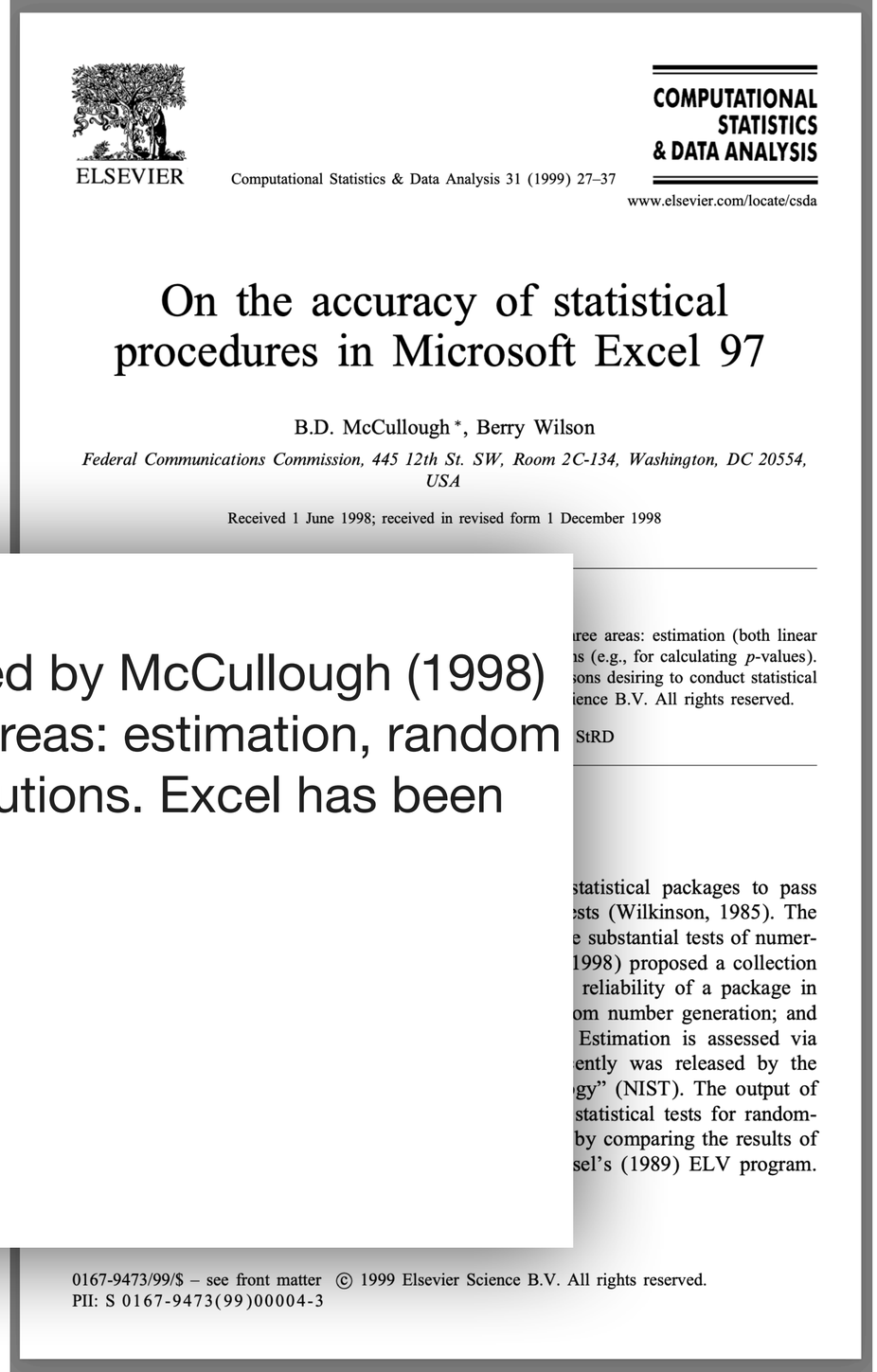
The weaknesses of the office model

- Surely, though, those were fixed, right?
  - McCullogh & Wilson (1999)



Conclusions and recommendations

We have applied the methodology outlined by McCullough (1998) to assess the reliability of Excel in three areas: estimation, random number generation, and statistical distributions. Excel has been found inadequate in all three areas.



ELSEVIER

Computational Statistics & Data Analysis 31 (1999) 27–37

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 97

B.D. McCullough *, Berry Wilson

Federal Communications Commission, 445 12th St. SW, Room 2C-134, Washington, DC 20554, USA

Received 1 June 1998; received in revised form 1 December 1998

ree areas: estimation (both linear
s (e.g., for calculating *p*-values).
ons desiring to conduct statistical
ience B.V. All rights reserved.

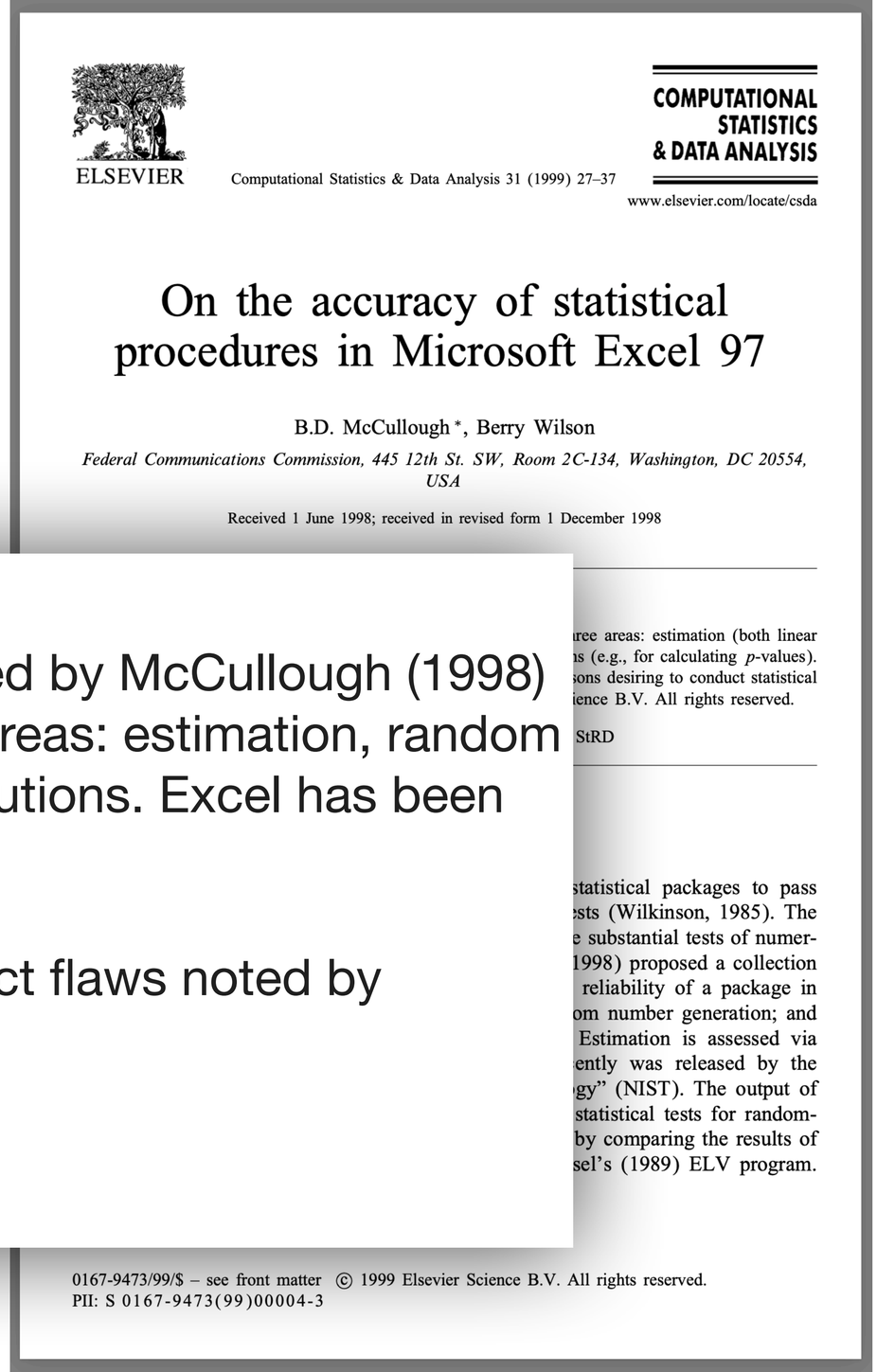
StRD

statistical packages to pass
sts (Wilkinson, 1985). The
e substantial tests of numer-
1998) proposed a collection
reliability of a package in
om number generation; and
Estimation is assessed via
ently was released by the
gy" (NIST). The output of
statistical tests for random-
by comparing the results of
sel's (1989) ELV program.

0167-9473/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S 0167-9473(99)00004-3

The weaknesses of the office model

- Surely, though, those were fixed, right?
  - McCullogh & Wilson (1999)



ELSEVIER

Computational Statistics & Data Analysis 31 (1999) 27–37

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 97

B.D. McCullough *, Berry Wilson

Federal Communications Commission, 445 12th St. SW, Room 2C-134, Washington, DC 20554, USA

Received 1 June 1998; received in revised form 1 December 1998

ree areas: estimation (both linear
s (e.g., for calculating *p*-values).
sons desiring to conduct statistical
ience B.V. All rights reserved.

StRD

statistical packages to pass
sts (Wilkinson, 1985). The
e substantial tests of numer-
1998) proposed a collection
reliability of a package in
om number generation; and
Estimation is assessed via
ently was released by the
gy" (NIST). The output of
statistical tests for random-
by comparing the results of
sel's (1989) ELV program.

Conclusions and recommendations

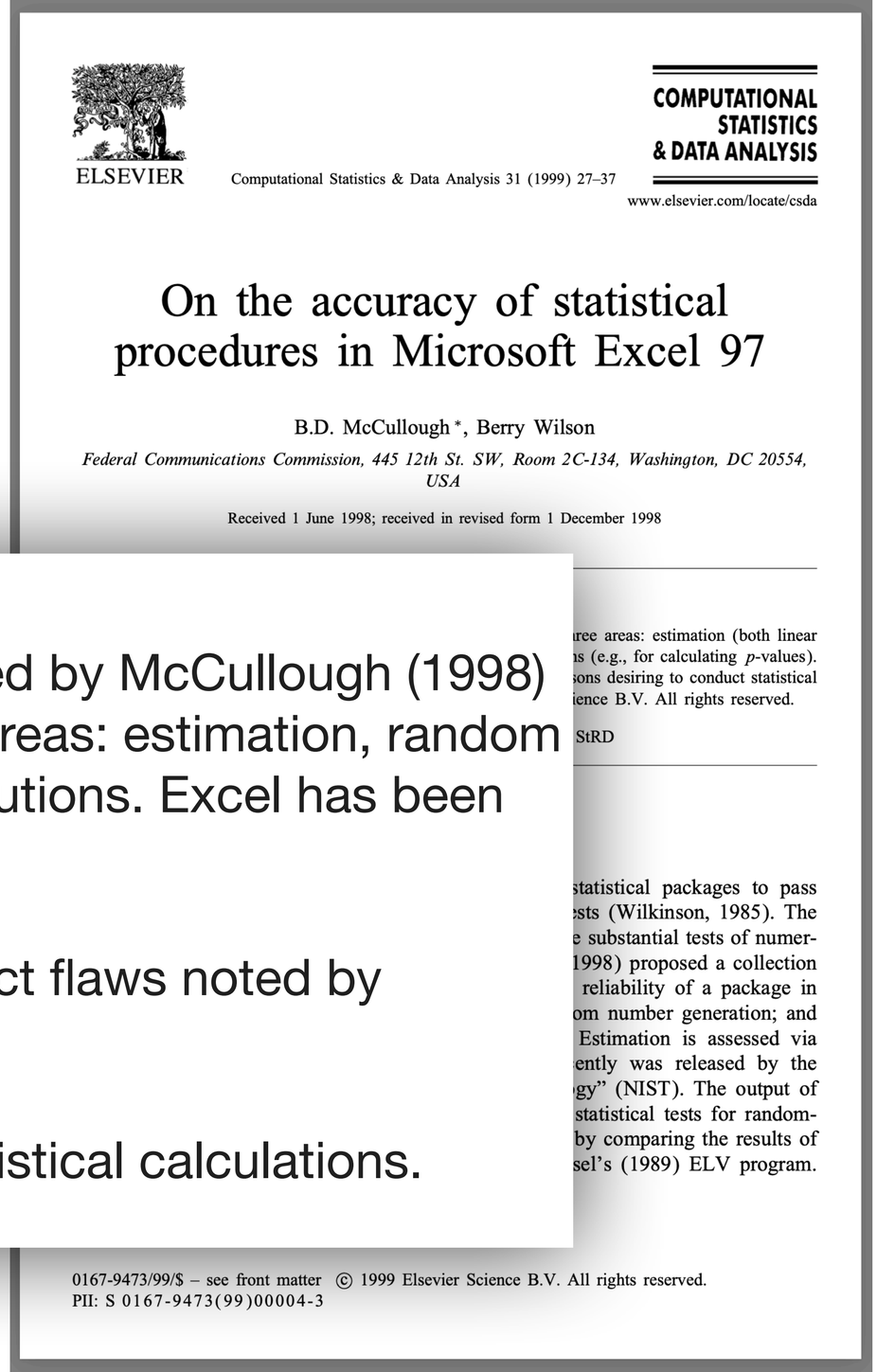
We have applied the methodology outlined by McCullough (1998) to assess the reliability of Excel in three areas: estimation, random number generation, and statistical distributions. Excel has been found inadequate in all three areas.

We also note that Microsoft did not correct flaws noted by Sawitzki (1994b).

0167-9473/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S 0167-9473(99)00004-3

The weaknesses of the office model

- Surely, though, those were fixed, right?
  - McCullogh & Wilson (1999)



ELSEVIER

Computational Statistics & Data Analysis 31 (1999) 27–37

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 97

B.D. McCullough *, Berry Wilson

Federal Communications Commission, 445 12th St. SW, Room 2C-134, Washington, DC 20554, USA

Received 1 June 1998; received in revised form 1 December 1998

ree areas: estimation (both linear
s (e.g., for calculating *p*-values).
ons desiring to conduct statistical
ience B.V. All rights reserved.

StRD

statistical packages to pass
sts (Wilkinson, 1985). The
e substantial tests of numer-
1998) proposed a collection
reliability of a package in
om number generation; and
Estimation is assessed via
ently was released by the
gy" (NIST). The output of
statistical tests for random-
by comparing the results of
sel's (1989) ELV program.

Conclusions and recommendations

We have applied the methodology outlined by McCullough (1998) to assess the reliability of Excel in three areas: estimation, random number generation, and statistical distributions. Excel has been found inadequate in all three areas.

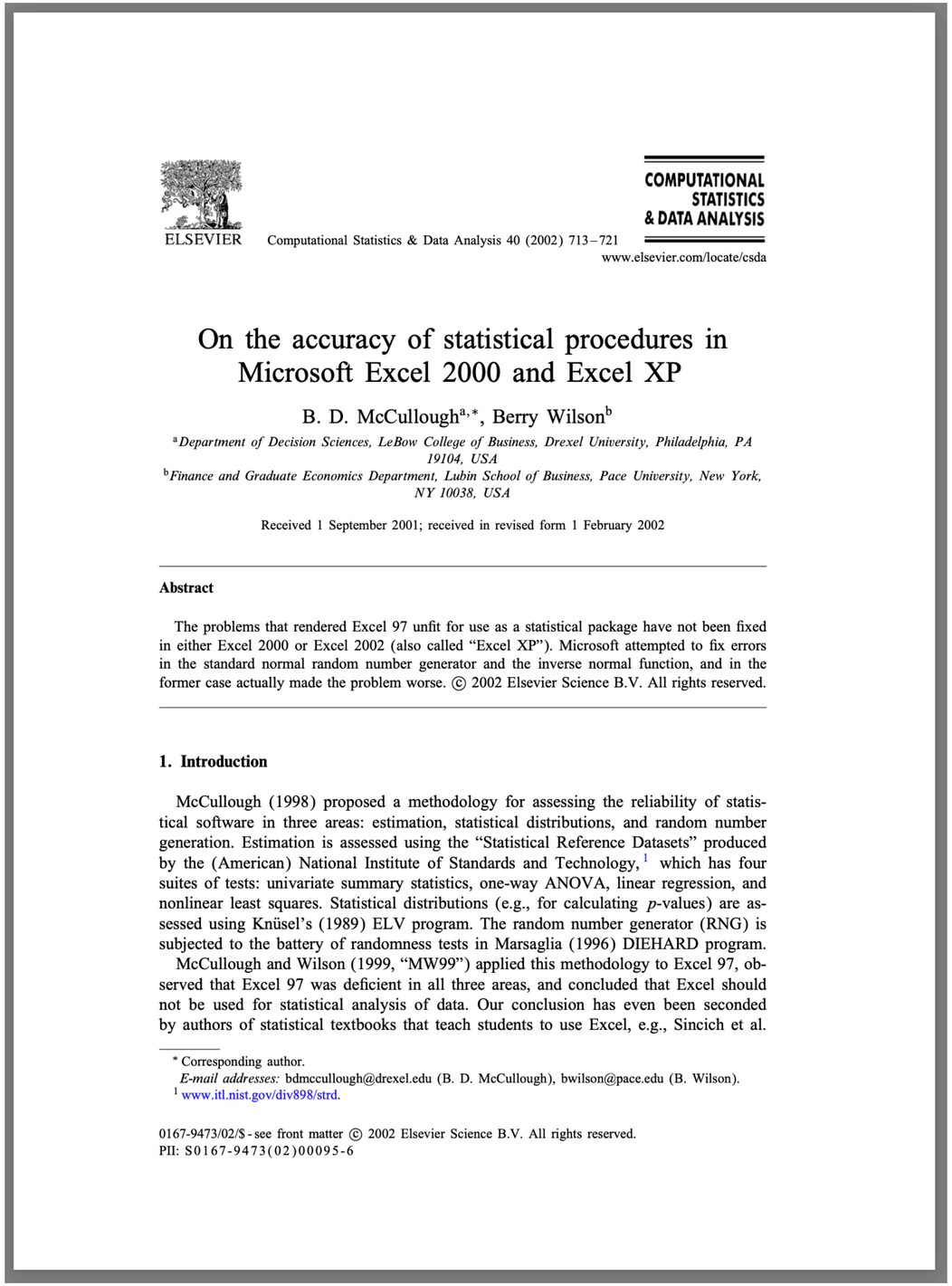
We also note that Microsoft did not correct flaws noted by Sawitzki (1994b).

We advise that Excel not be used for statistical calculations.

0167-9473/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S 0167-9473(99)00004-3

The weaknesses of the office model

- Well, come on, I'm sure Microsoft got around to it next time
  - McCullogh & Wilson (2002)



ELSEVIER
Computational Statistics & Data Analysis 40 (2002) 713–721
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP

B. D. McCullough[a,*], Berry Wilson[b]

[a] *Department of Decision Sciences, LeBow College of Business, Drexel University, Philadelphia, PA 19104, USA*
[b] *Finance and Graduate Economics Department, Lubin School of Business, Pace University, New York, NY 10038, USA*

Received 1 September 2001; received in revised form 1 February 2002

Abstract

The problems that rendered Excel 97 unfit for use as a statistical package have not been fixed in either Excel 2000 or Excel 2002 (also called "Excel XP"). Microsoft attempted to fix errors in the standard normal random number generator and the inverse normal function, and in the former case actually made the problem worse. © 2002 Elsevier Science B.V. All rights reserved.

1. Introduction

McCullough (1998) proposed a methodology for assessing the reliability of statistical software in three areas: estimation, statistical distributions, and random number generation. Estimation is assessed using the "Statistical Reference Datasets" produced by the (American) National Institute of Standards and Technology,[1] which has four suites of tests: univariate summary statistics, one-way ANOVA, linear regression, and nonlinear least squares. Statistical distributions (e.g., for calculating *p*-values) are assessed using Knüsel's (1989) ELV program. The random number generator (RNG) is subjected to the battery of randomness tests in Marsaglia (1996) DIEHARD program.

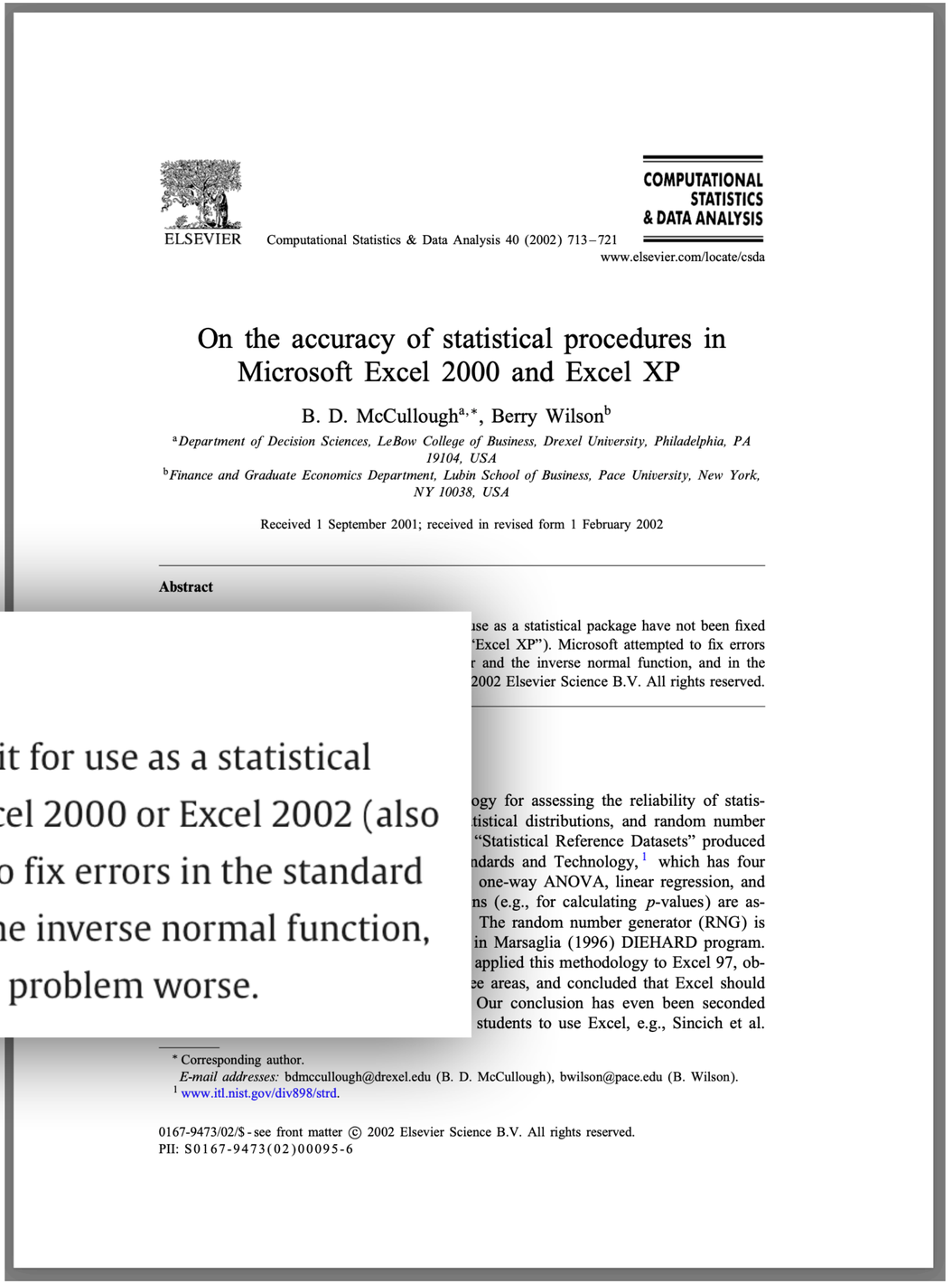
McCullough and Wilson (1999, "MW99") applied this methodology to Excel 97, observed that Excel 97 was deficient in all three areas, and concluded that Excel should not be used for statistical analysis of data. Our conclusion has even been seconded by authors of statistical textbooks that teach students to use Excel, e.g., Sincich et al.

* Corresponding author.
E-mail addresses: bdmccullough@drexel.edu (B. D. McCullough), bwilson@pace.edu (B. Wilson).
[1] www.itl.nist.gov/div898/strd.

0167-9473/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9473(02)00095-6

The weaknesses of the office model

- Well, come on, I'm sure they got around to it next time
  - McCullogh & Wilson (2002)



ELSEVIER Computational Statistics & Data Analysis 40 (2002) 713–721
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP

B. D. McCullough[a,*], Berry Wilson[b]
[a] *Department of Decision Sciences, LeBow College of Business, Drexel University, Philadelphia, PA 19104, USA*
[b] *Finance and Graduate Economics Department, Lubin School of Business, Pace University, New York, NY 10038, USA*

Received 1 September 2001; received in revised form 1 February 2002

Abstract

The problems that rendered Excel 97 unfit for use as a statistical package have not been fixed in either Excel 2000 or Excel 2002 (also called "Excel XP"). Microsoft attempted to fix errors in the standard normal random number generator and the inverse normal function, and in the former case actually made the problem worse.

2002 Elsevier Science B.V. All rights reserved.

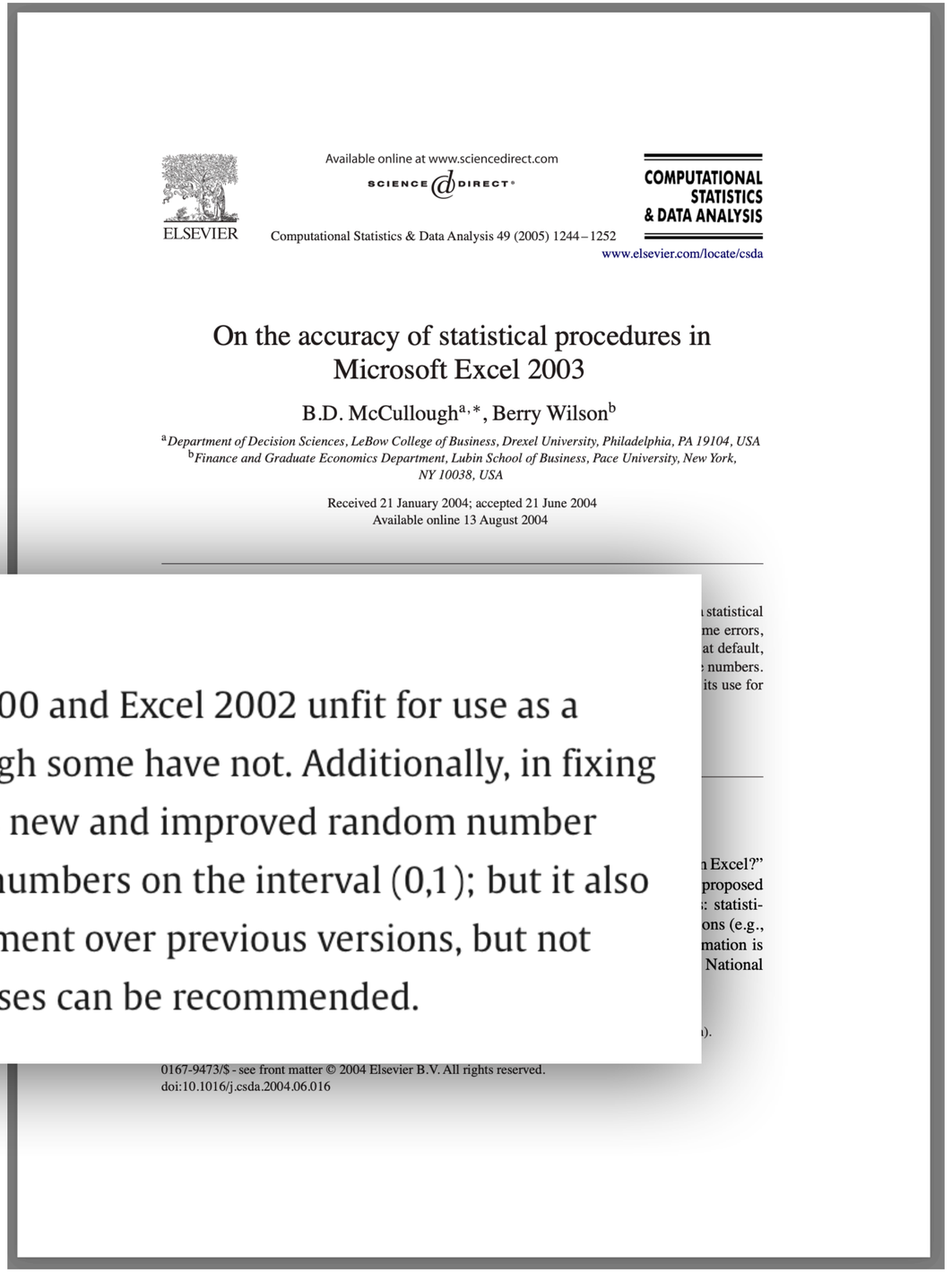
ogy for assessing the reliability of statis-
tistical distributions, and random number
"Statistical Reference Datasets" produced
ndards and Technology,[1] which has four
one-way ANOVA, linear regression, and
ns (e.g., for calculating *p*-values) are as-
The random number generator (RNG) is
in Marsaglia (1996) DIEHARD program.
applied this methodology to Excel 97, ob-
ee areas, and concluded that Excel should
Our conclusion has even been seconded
students to use Excel, e.g., Sincich et al.

* Corresponding author.
E-mail addresses: bdmccullough@drexel.edu (B. D. McCullough), bwilson@pace.edu (B. Wilson).
[1] www.itl.nist.gov/div898/strd.

0167-9473/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9473(02)00095-6

The weaknesses of the office model

- Third time's the charm???
  - McCullogh & Wilson (2002)



Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 49 (2005) 1244–1252
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 2003

B.D. McCullough[a,*], Berry Wilson[b]

[a]*Department of Decision Sciences, LeBow College of Business, Drexel University, Philadelphia, PA 19104, USA*
[b]*Finance and Graduate Economics Department, Lubin School of Business, Pace University, New York, NY 10038, USA*

Received 21 January 2004; accepted 21 June 2004
Available online 13 August 2004

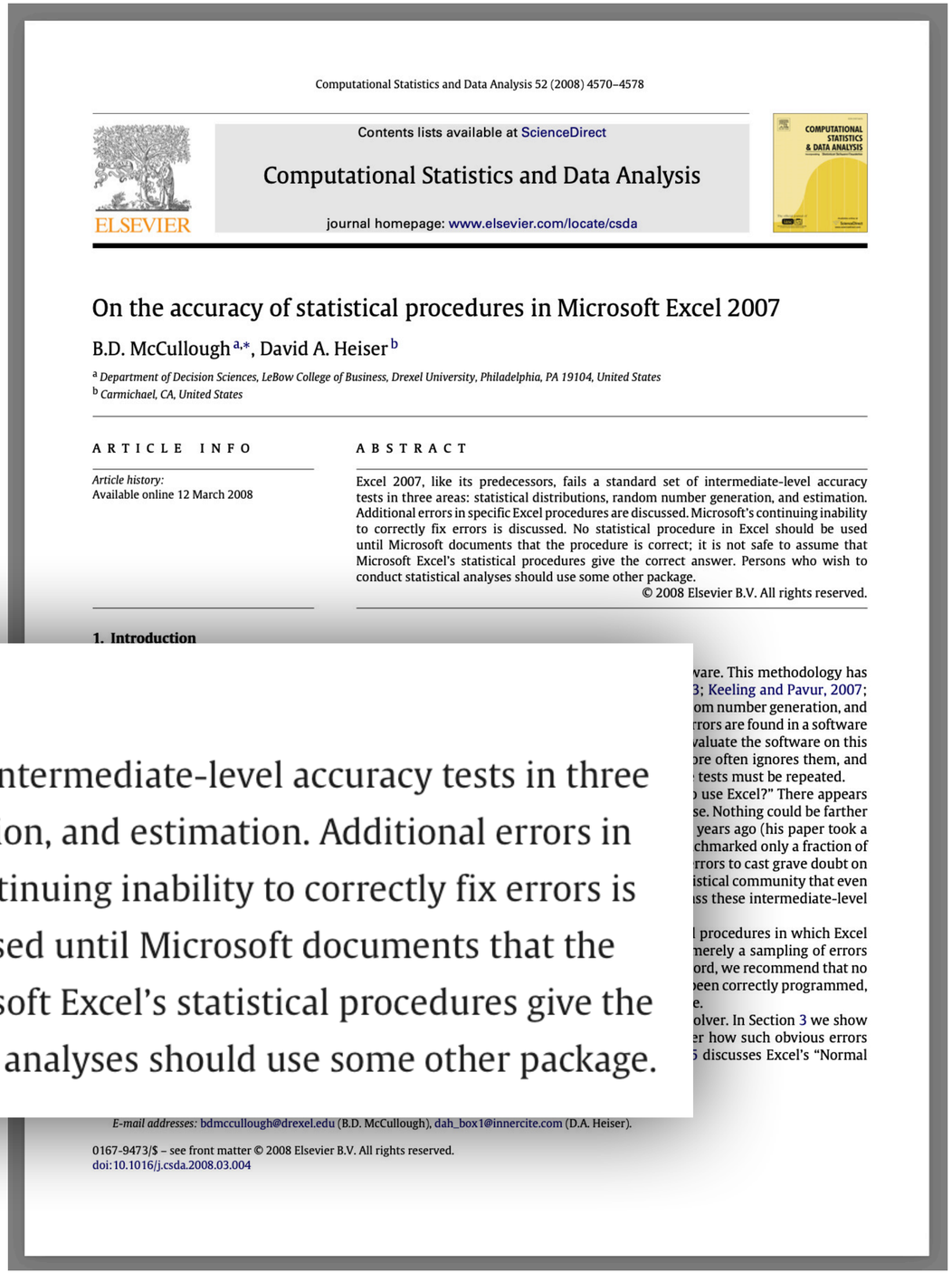
Abstract

Some of the problems that rendered Excel 97, Excel 2000 and Excel 2002 unfit for use as a statistical package have been fixed in Excel 2003, though some have not. Additionally, in fixing some errors, Microsoft introduced other errors. Excel's new and improved random number generator, at default, is supposed to produce uniform numbers on the interval (0,1); but it also produces negative numbers. Excel 2003 is an improvement over previous versions, but not enough has been done that its use for statistical purposes can be recommended.

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.06.016

The weaknesses of the office model

- You know, just don't tell me. I don't want to know.
  - McCullogh & Wilson (2008)



Computational Statistics and Data Analysis 52 (2008) 4570–4578

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

On the accuracy of statistical procedures in Microsoft Excel 2007

B.D. McCullough [a,*], David A. Heiser [b]

[a] Department of Decision Sciences, LeBow College of Business, Drexel University, Philadelphia, PA 19104, United States
[b] Carmichael, CA, United States

ARTICLE INFO

Article history:
Available online 12 March 2008

ABSTRACT

Excel 2007, like its predecessors, fails a standard set of intermediate-level accuracy tests in three areas: statistical distributions, random number generation, and estimation. Additional errors in specific Excel procedures are discussed. Microsoft's continuing inability to correctly fix errors is discussed. No statistical procedure in Excel should be used until Microsoft documents that the procedure is correct; it is not safe to assume that Microsoft Excel's statistical procedures give the correct answer. Persons who wish to conduct statistical analyses should use some other package.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

E-mail addresses: bdmccullough@drexel.edu (B.D. McCullough), dah_box1@innercite.com (D.A. Heiser).

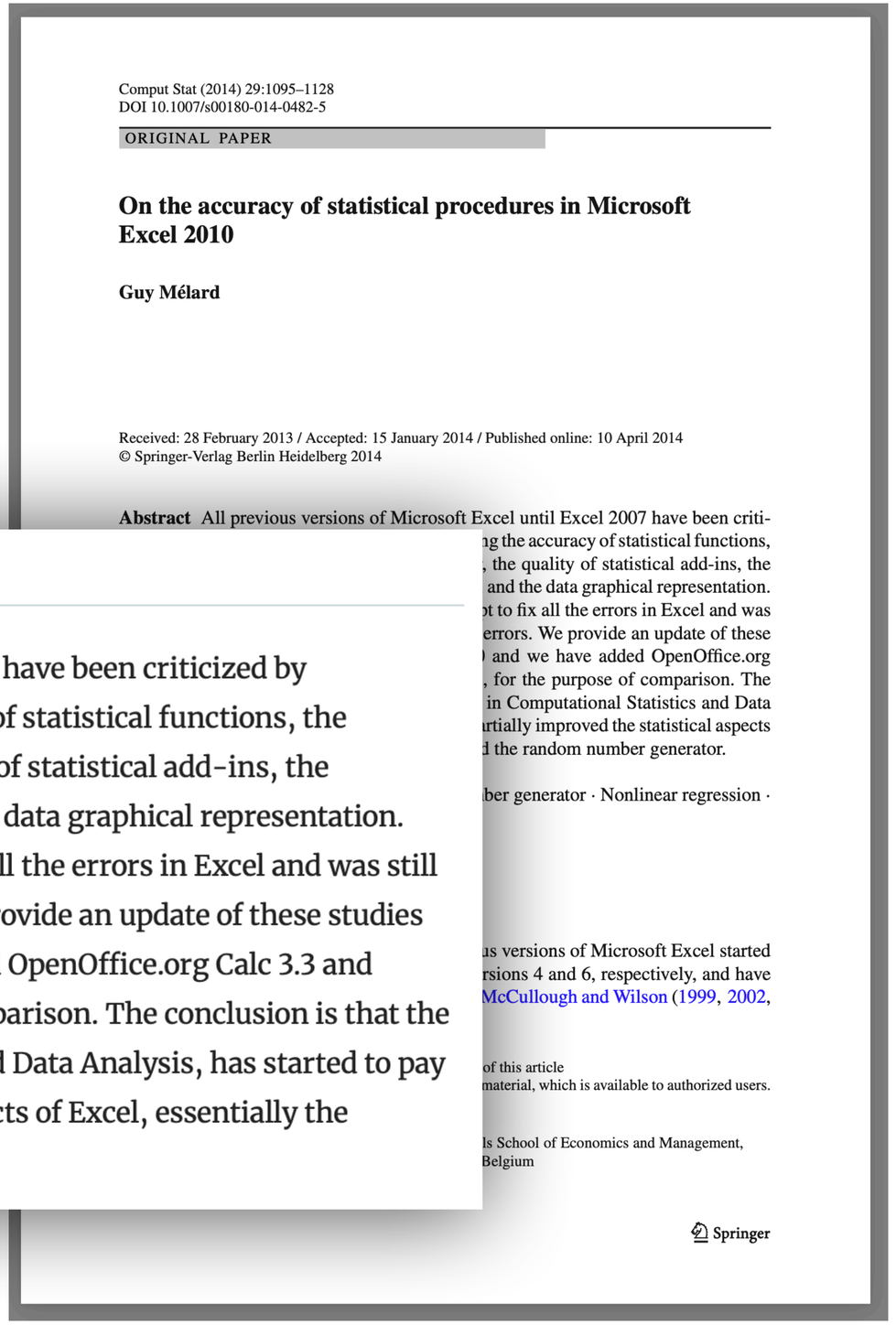
0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.03.004

Abstract

Excel 2007, like its predecessors, fails a standard set of intermediate-level accuracy tests in three areas: statistical distributions, random number generation, and estimation. Additional errors in specific Excel procedures are discussed. Microsoft's continuing inability to correctly fix errors is discussed. No statistical procedure in Excel should be used until Microsoft documents that the procedure is correct; it is not safe to assume that Microsoft Excel's statistical procedures give the correct answer. Persons who wish to conduct statistical analyses should use some other package.

The weaknesses of the office model

- Do we . . . really have to keep doing this?
  - Melard (2014)



Comput Stat (2014) 29:1095–1128
DOI 10.1007/s00180-014-0482-5

ORIGINAL PAPER

On the accuracy of statistical procedures in Microsoft Excel 2010

Guy Mélard

Received: 28 February 2013 / Accepted: 15 January 2014 / Published online: 10 April 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract All previous versions of Microsoft Excel until Excel 2007 have been criti-
ng the accuracy of statistical functions,
, the quality of statistical add-ins, the
and the data graphical representation.
pt to fix all the errors in Excel and was
errors. We provide an update of these
and we have added OpenOffice.org
, for the purpose of comparison. The
in Computational Statistics and Data
rtially improved the statistical aspects
l the random number generator.

ber generator · Nonlinear regression ·

us versions of Microsoft Excel started
rsions 4 and 6, respectively, and have
McCullough and Wilson (1999, 2002,

of this article
material, which is available to authorized users.

ls School of Economics and Management,
Belgium

Abstract

All previous versions of Microsoft Excel until Excel 2007 have been criticized by statisticians for several reasons, including the accuracy of statistical functions, the properties of the random number generator, the quality of statistical add-ins, the weakness of the Solver for nonlinear regression, and the data graphical representation. Until recently Microsoft did not make an attempt to fix all the errors in Excel and was still marketing a product that contained known errors. We provide an update of these studies given the recent release of Excel 2010 and we have added OpenOffice.org Calc 3.3 and Gnumeric 1.10.16 to the analysis, for the purpose of comparison. The conclusion is that the stream of papers, mainly in Computational Statistics and Data Analysis, has started to pay off: Microsoft has partially improved the statistical aspects of Excel, essentially the statistical functions and the random number generator.

Everybody! Everybody!

- IT GOT BETTER!
- IT GOT BETTER!
- **EXCEL DID IT!**
- Well . . . kind of.

You knew we weren't done

- There's . . . more?
  - Ziemann, Eren, and El-Osta (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

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.

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) …n was applied to identify gene …e symbols from Ensembl ver…mber 2015, were obtained for …enorhabditis elegans, Drosoph…rerio, Escherichia coli, Gallus …s musculus, Oryza sativa and [2]. The regex search used was …previously by Zeeberg and col…d screen for dates in other for…nd MM-DD-YY). To expedite …y files from multi-disciplinary …articles screened to those that …in the title or abstract (*Science*, …cel files (.xls and.xlsx) deposited … Omnibus (GEO) [3] were also … (files released 2005–2015). All …d scripts used in this study are …urceForge (https://sourceforge. …rsscreen/). Scripts were run on …GNU bash, version 4.3.11. These …anually by downloading and checking Excel files from every paper and GEO file suspected to include gene name errors.

Supplementary files in Excel format from 18 journals published from 2005 to 2015 were programmatically screened for the presence of gene name errors. In total, we screened 35,175 supplementary Excel files, finding 7467 gene lists attached to 3597 published papers. We

* Correspondence: Assam.El-Osta@bakeridi.edu.au
[1]Baker IDI Heart & Diabetes Institute, The Alfred Medical Research and Education Precinct, Melbourne, Victoria 3004, Australia
[3]Central Clinical School, Faculty of Medicine, Monash University, Clayton, Victoria 3168, Australia
Full list of author information is available at the end of the article

BioMed Central

© 2016 The Author(s). **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

“. . . one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.”

Classes of quantitative errors

- Research design errors
- Calculation errors (looking at you, Excel 97)
- Data errors (looking at you, Excel SEPT2)
- Intentionally malicious work
  - ***So what do we do?***

CONTROL

- and -

Reproducibility

we want to know the **exact code**

that produces the **exact result**

that we **exactly report**

also, we'd like to **keep records**
preferably **open**

if we make a **mistake**, we want to

know exactly where it came from

also, we'd like to use software tailored to our purposes

This “engineering” approach

Is it perfect?

No.

Is it easy?

No.

Will I like it?

I hope so!

That “office” approach

Is it bad?

No.

Is it easy?

Kind of.

Can I go back?

Not entirely.

Strengths and weaknesses

- **Office** model
  - Documents look like documents
  - Everyone knows Word, Excel, Google docs, etc
  - “Track changes” is super useful in a single document
  - *Hm, I can’t remember how I made this figure.*
  - *Where did this table of results come from?*
  - `Paper_edits_FINAL_really_lastone-6.docx`
- **Engineering** model
  - Plain text is highly portable.
  - Push button, recreate analysis.
  - Can track changes across entire project
  - Tables and figures produced and integrated programatically.
  - For the love of God, why can’t I do this simple #(*&@#$ thing?
  - `Object of type 'closure' is not subsettable`

A Tale of Two ~~Cities~~ Computers

- Each approach generates its own problems
  - **Engineering** grants **control** of code, **reproducibility,** and **advanced reporting options**
    - *at the expense of a learning curve and some continued frustration*
  - **Office** grants ease, speed, and nice GUI interfaces
    - *at the expense of control, reproducibility, and reporting options*

Principles | Tools

O'Hara-Wild, Mitchell. 2018. "useR! 2018 Feature Wall." July 11, 2018. https://mitchelloharawild.com/blog/user-2018-feature-wall.

The to-do list

we want to know the **exact code**
that produces the **exact result**
that we **exactly report**

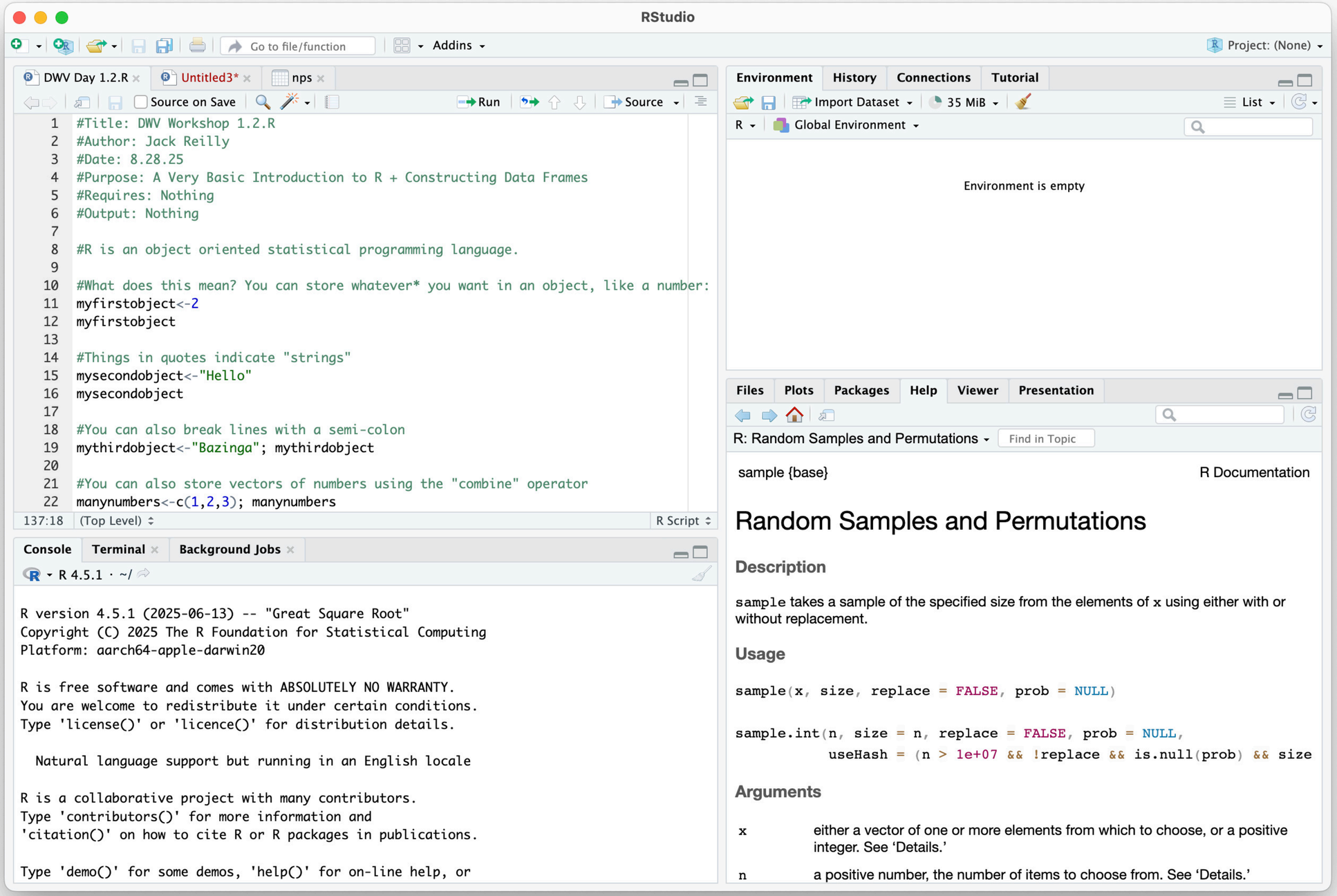
also, we'd like to **keep records**
preferably **open**

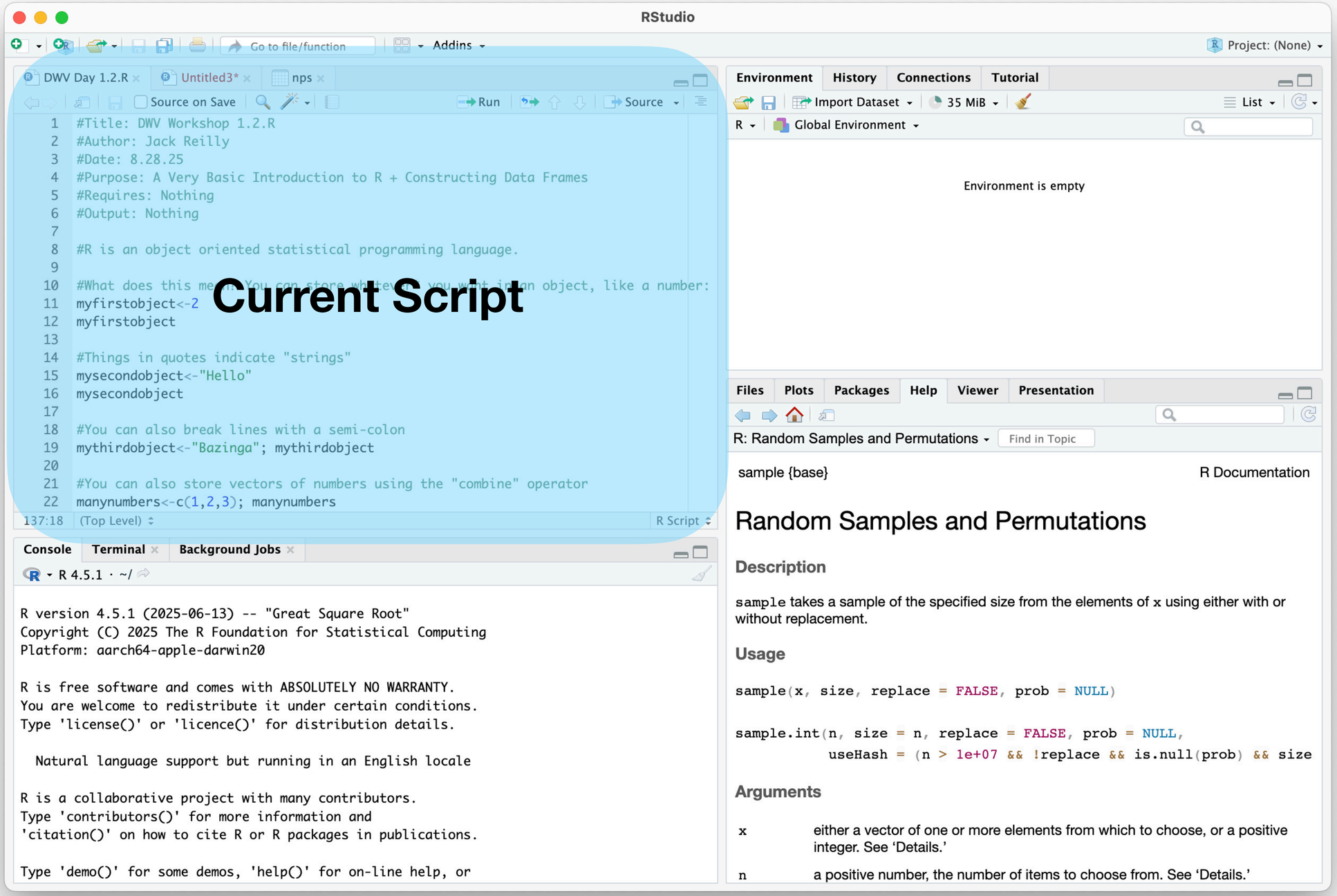
The Open Science Toolkit

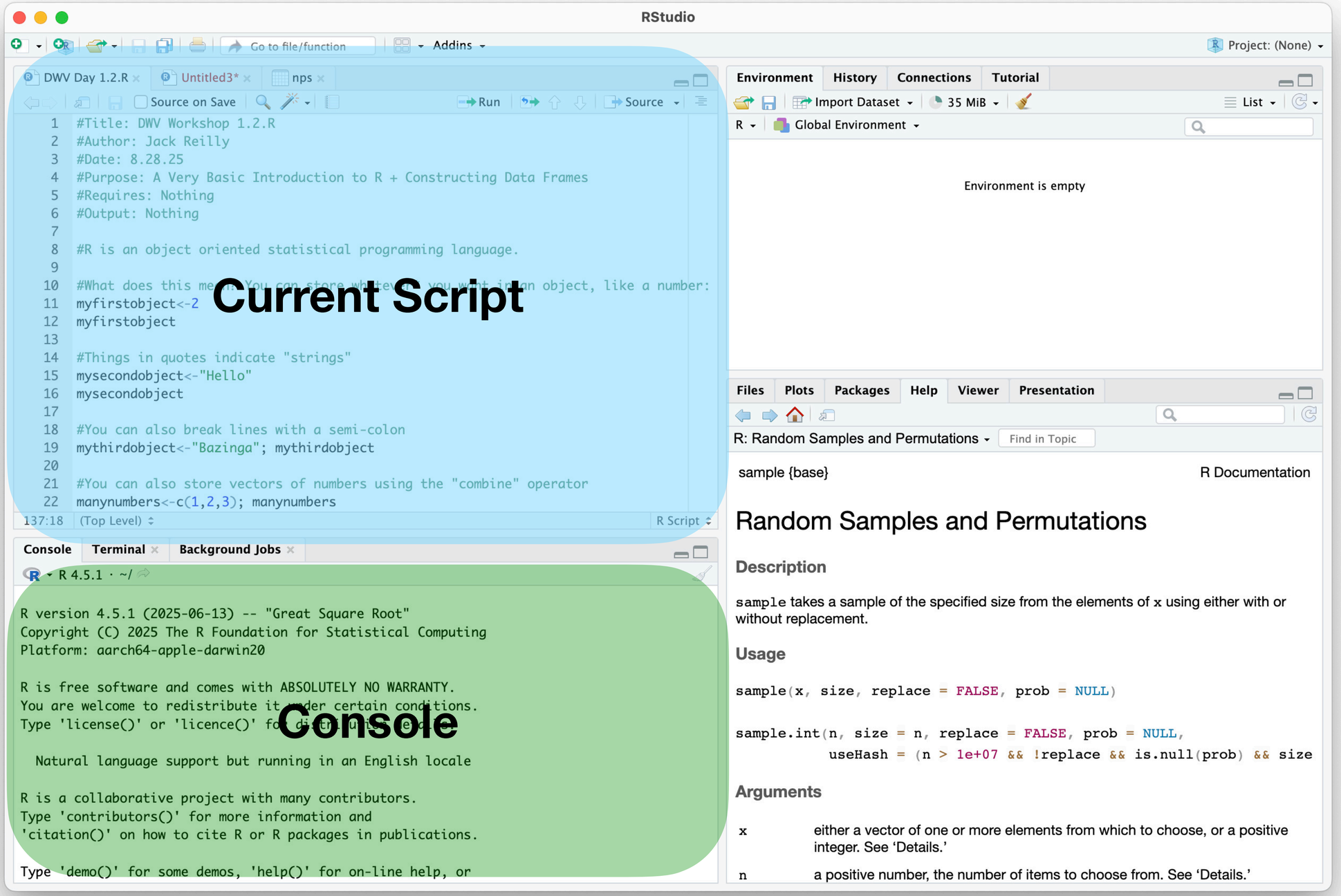
- So, first, we want to do our work **openly** and **reproducibly**
  - First, we write **in code**, rather than point and click
  - Second, we use open software (R, in our case) and save our *scripts*
  - In an IDE (integrated development environment) known as Rstudio

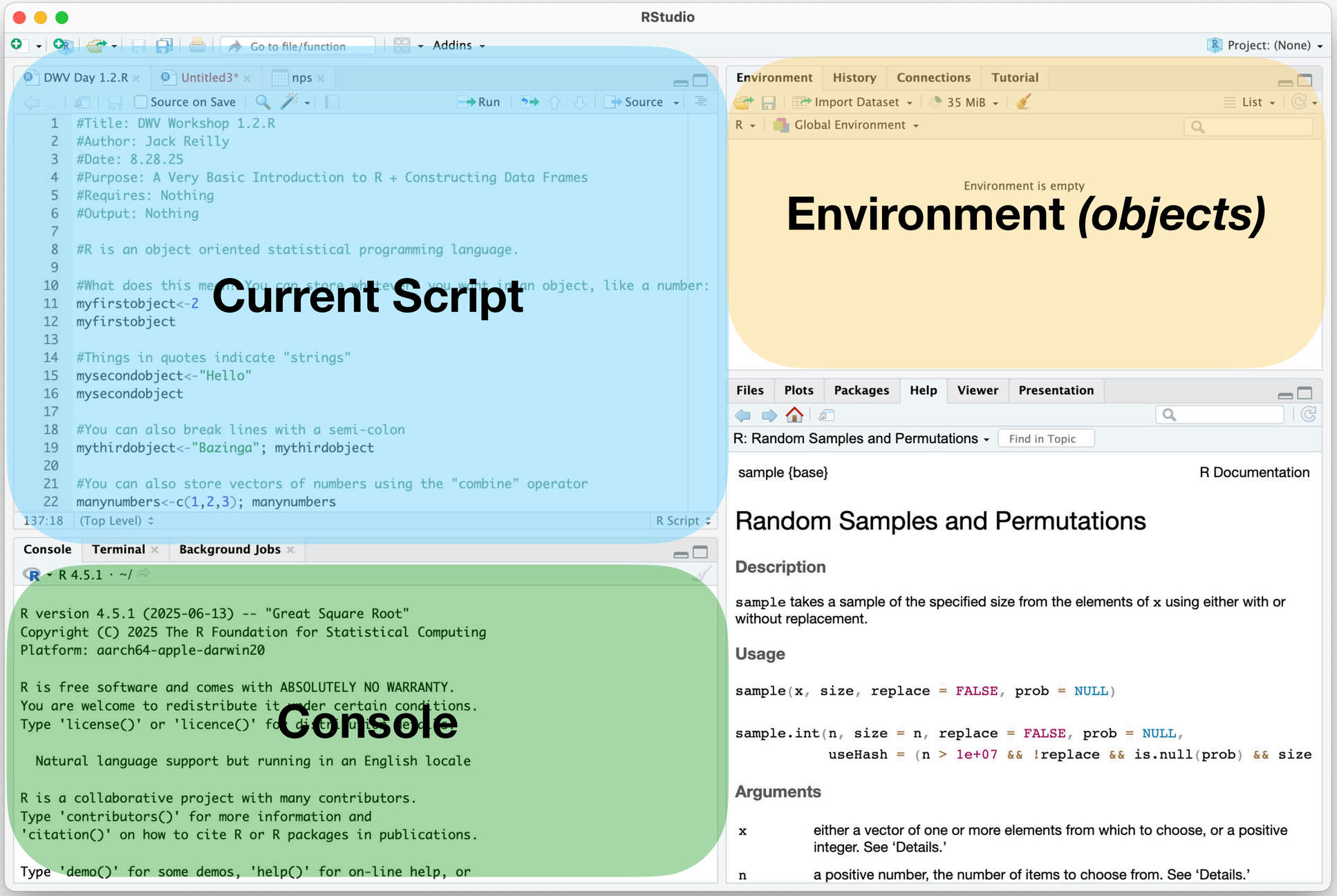
An IDE?

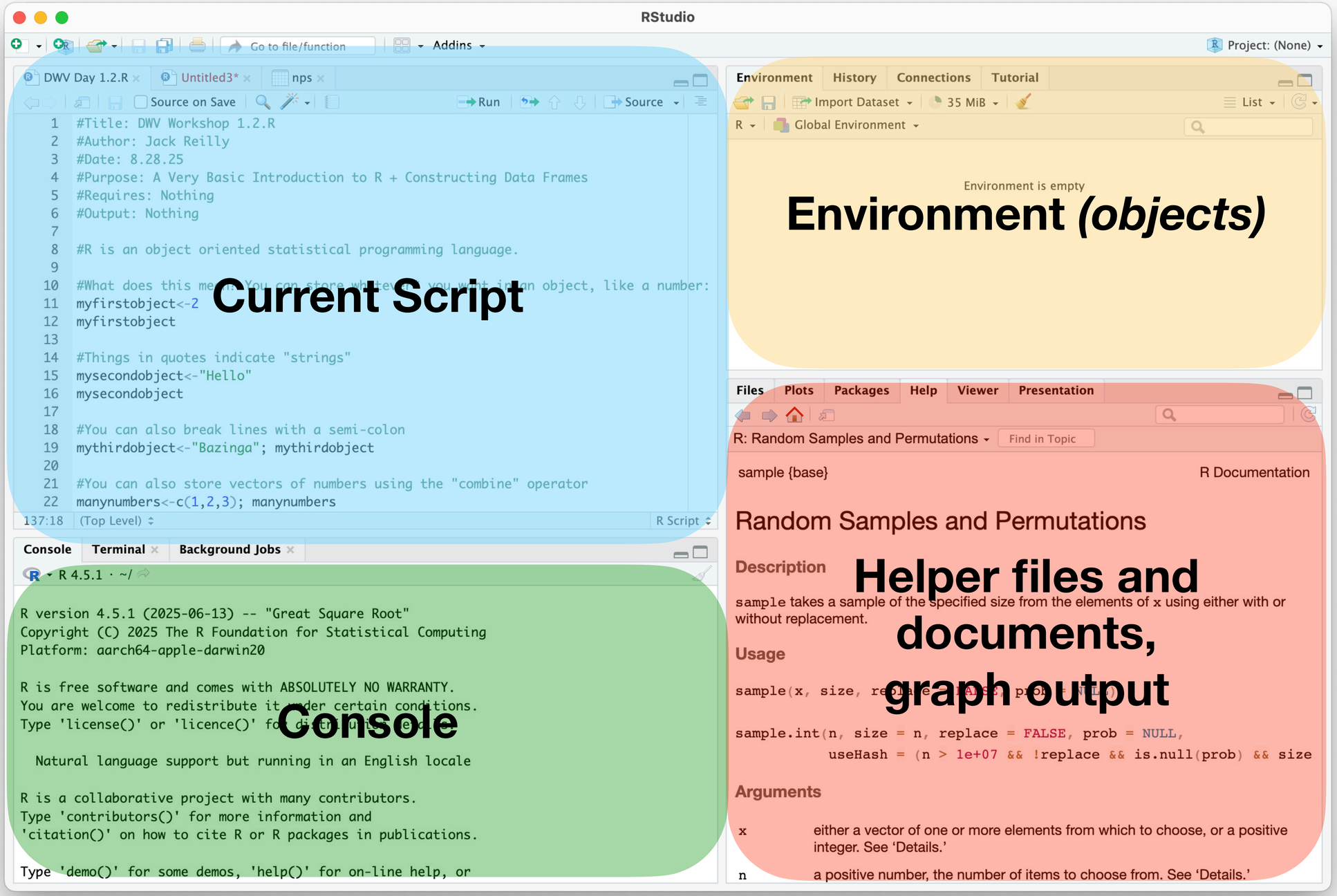
- *Integrated Development Environment*
  - Think of a kitchen
  - You’re the chef. The kitchen (IDE) has everything you need in it make the recipe (the script file) and produce the food (data analysis)


RStudio
Go to file/function
Addins
Project: (None)
DWV Day 1.2.R
Untitled3*
nps
Source on Save
Run
Source
#Title: DWV Workshop 1.2.R
#Author: Jack Reilly
#Date: 8.28.25
#Purpose: A Very Basic Introduction to R + Constructing Data Frames
#Requires: Nothing
#Output: Nothing
#R is an object oriented statistical programming language.
#What does this mean? You can store whatever* you want in an object, like a number:
myfirstobject<-2
myfirstobject
#Things in quotes indicate "strings"
mysecondobject<-"Hello"
mysecondobject
#You can also break lines with a semi-colon
mythirdobject<-"Bazinga"; mythirdobject
#You can also store vectors of numbers using the "combine" operator
manynumbers<-c(1,2,3); manynumbers
137:18
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.5.1 · ~/
R version 4.5.1 (2025-06-13) -- "Great Square Root"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
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
Environment
History
Connections
Tutorial
Import Dataset
35 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
R: Random Samples and Permutations
Find in Topic
sample {base}
R Documentation
Random Samples and Permutations
Description
sample takes a sample of the specified size from the elements of x using either with or without replacement.
Usage
sample(x, size, replace = FALSE, prob = NULL)
sample.int(n, size = n, replace = FALSE, prob = NULL,
useHash = (n > 1e+07 && !replace && is.null(prob) && size
Arguments
x
either a vector of one or more elements from which to choose, or a positive integer. See 'Details.'
n
a positive number, the number of items to choose from. See 'Details.'


RStudio
Go to file/function
Addins
Project: (None)
DWV Day 1.2.R
Untitled3*
nps
Source on Save
Run
Source
#Title: DWV Workshop 1.2.R
#Author: Jack Reilly
#Date: 8.28.25
#Purpose: A Very Basic Introduction to R + Constructing Data Frames
#Requires: Nothing
#Output: Nothing
#R is an object oriented statistical programming language.
#What does this mean? You can store whatever you want in an object, like a number:
myfirstobject<-2
myfirstobject
#Things in quotes indicate "strings"
mysecondobject<-"Hello"
mysecondobject
#You can also break lines with a semi-colon
mythirdobject<-"Bazinga"; mythirdobject
#You can also store vectors of numbers using the "combine" operator
manynumbers<-c(1,2,3); manynumbers
137:18
(Top Level)
R Script
Current Script
Console
Terminal
Background Jobs
R 4.5.1 · ~/
R version 4.5.1 (2025-06-13) -- "Great Square Root"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
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
Environment
History
Connections
Tutorial
Import Dataset
35 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
R: Random Samples and Permutations
Find in Topic
sample {base}
R Documentation
Random Samples and Permutations
Description
sample takes a sample of the specified size from the elements of x using either with or without replacement.
Usage
sample(x, size, replace = FALSE, prob = NULL)
sample.int(n, size = n, replace = FALSE, prob = NULL,
useHash = (n > 1e+07 && !replace && is.null(prob) && size
Arguments
x
either a vector of one or more elements from which to choose, or a positive integer. See 'Details.'
n
a positive number, the number of items to choose from. See 'Details.'


RStudio
Go to file/function
Addins
Project: (None)
DWV Day 1.2.R
Untitled3*
nps
Source on Save
Run
Source
#Title: DWV Workshop 1.2.R
#Author: Jack Reilly
#Date: 8.28.25
#Purpose: A Very Basic Introduction to R + Constructing Data Frames
#Requires: Nothing
#Output: Nothing
#R is an object oriented statistical programming language.
#What does this mean? You can store whatever you want in an object, like a number:
myfirstobject<-2
myfirstobject
#Things in quotes indicate "strings"
mysecondobject<-"Hello"
mysecondobject
#You can also break lines with a semi-colon
mythirdobject<-"Bazinga"; mythirdobject
#You can also store vectors of numbers using the "combine" operator
manynumbers<-c(1,2,3); manynumbers
137:18
(Top Level)
R Script
Current Script
Console
Terminal
Background Jobs
R 4.5.1 · ~/
R version 4.5.1 (2025-06-13) -- "Great Square Root"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
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
Console
Environment
History
Connections
Tutorial
Import Dataset
35 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
R: Random Samples and Permutations
Find in Topic
sample {base}
R Documentation
Random Samples and Permutations
Description
sample takes a sample of the specified size from the elements of x using either with or without replacement.
Usage
sample(x, size, replace = FALSE, prob = NULL)
sample.int(n, size = n, replace = FALSE, prob = NULL,
useHash = (n > 1e+07 && !replace && is.null(prob) && size
Arguments
x
either a vector of one or more elements from which to choose, or a positive integer. See 'Details.'
n
a positive number, the number of items to choose from. See 'Details.'


RStudio
Go to file/function
Addins
Project: (None)
DWV Day 1.2.R
Untitled3*
nps
Source on Save
Run
Source
#Title: DWV Workshop 1.2.R
#Author: Jack Reilly
#Date: 8.28.25
#Purpose: A Very Basic Introduction to R + Constructing Data Frames
#Requires: Nothing
#Output: Nothing
#R is an object oriented statistical programming language.
#What does this mean? You can store whatever you want in an object, like a number:
myfirstobject<-2
myfirstobject
#Things in quotes indicate "strings"
mysecondobject<-"Hello"
mysecondobject
#You can also break lines with a semi-colon
mythirdobject<-"Bazinga"; mythirdobject
#You can also store vectors of numbers using the "combine" operator
manynumbers<-c(1,2,3); manynumbers
137:18
(Top Level)
R Script
Current Script
Console
Terminal
Background Jobs
R 4.5.1 · ~/
R version 4.5.1 (2025-06-13) -- "Great Square Root"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
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
Console
Environment
History
Connections
Tutorial
Import Dataset
35 MiB
List
R
Global Environment
Environment is empty
Environment (objects)
Files
Plots
Packages
Help
Viewer
Presentation
R: Random Samples and Permutations
Find in Topic
sample {base}
R Documentation
Random Samples and Permutations
Description
sample takes a sample of the specified size from the elements of x using either with or without replacement.
Usage
sample(x, size, replace = FALSE, prob = NULL)
sample.int(n, size = n, replace = FALSE, prob = NULL,
useHash = (n > 1e+07 && !replace && is.null(prob) && size
Arguments
x
either a vector of one or more elements from which to choose, or a positive integer. See 'Details.'
n
a positive number, the number of items to choose from. See 'Details.'


RStudio
Go to file/function
Addins
Project: (None)
DWV Day 1.2.R
Untitled3*
nps
Source on Save
Run
Source
#Title: DWV Workshop 1.2.R
#Author: Jack Reilly
#Date: 8.28.25
#Purpose: A Very Basic Introduction to R + Constructing Data Frames
#Requires: Nothing
#Output: Nothing
#R is an object oriented statistical programming language.
#What does this mean? You can store whatever you want in an object, like a number:
myfirstobject<-2
myfirstobject
#Things in quotes indicate "strings"
mysecondobject<-"Hello"
mysecondobject
#You can also break lines with a semi-colon
mythirdobject<-"Bazinga"; mythirdobject
#You can also store vectors of numbers using the "combine" operator
manynumbers<-c(1,2,3); manynumbers
137:18
(Top Level)
R Script
Current Script
Console
Terminal
Background Jobs
R 4.5.1 · ~/
R version 4.5.1 (2025-06-13) -- "Great Square Root"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
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
Console
Environment
History
Connections
Tutorial
Import Dataset
35 MiB
List
R
Global Environment
Environment is empty
Environment (objects)
Files
Plots
Packages
Help
Viewer
Presentation
R: Random Samples and Permutations
Find in Topic
sample {base}
R Documentation
Random Samples and Permutations
Description
sample takes a sample of the specified size from the elements of x using either with or without replacement.
Usage
sample(x, size, replace = FALSE, prob = NULL)
sample.int(n, size = n, replace = FALSE, prob = NULL,
useHash = (n > 1e+07 && !replace && is.null(prob) && size
Arguments
x
either a vector of one or more elements from which to choose, or a positive integer. See 'Details.'
n
a positive number, the number of items to choose from. See 'Details.'
Helper files and documents, graph output

Do I need to use RStudio? Or an IDE?

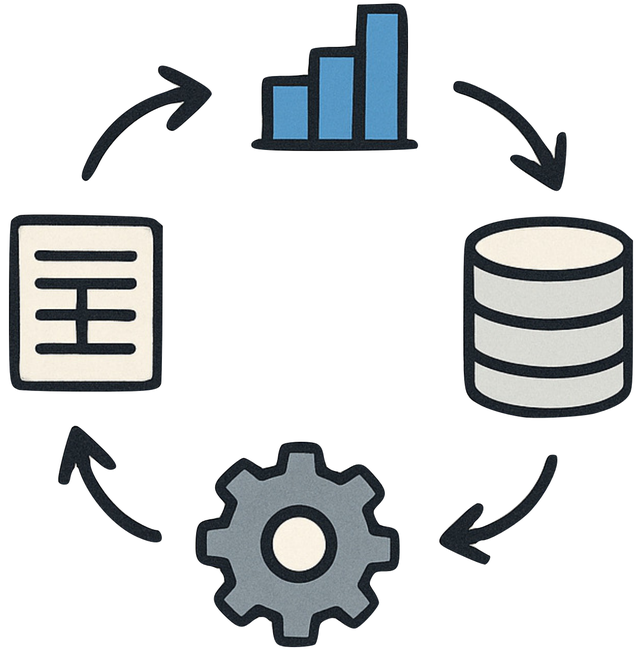
- No, actually!
    - A script file is just a plain text file. You can edit it anywhere. (*Although you **do** have to use a script file to make things reproducible*)
- There are other IDEs and text editors that others use
    - Visual Studio Code is getting very popular; Positron is kind of a next-generation RStudio; lots of people just write things in text editors like Sublime Text, TextMate, BBEdit, etc
- But it's handy to have everything together

RStudio does all of the following for you

- A **text editor** for writing code and documents
- A **console** or REPL (Read-Eval-Print Loop) for running code interactively
- A **terminal** to talk to the operating system
- A **debugger** to help find problems in your code
- A **file manager** to navigate your project
- A **version control interface** to manage changes to your code
- A **viewer** for plots, tables, and other outputs
- An **inspector** to see what's in your environment

Two (more) rough philosophies of computer usage

- The **omnipresent, monolithic** application approach
  - *“Do everything”*



- The **modular, interoperable** application approach
  - *“Do one thing, and do it well”*

Neither is perfect

- *When done well:*
  - One application for everything!
- *When done poorly:*
  - Jack of all trades, master of none

- *When done well:*
  - Symphony of parts.
- *When done poorly:*
  - Intractable coordination

Rstudio is kind of both! A monolithic wrapper around a bunch of smaller utilities

Two (more) rough philosophies of computer usage

- Most people lean towards the first
    - The second has some virtues, too
- Even if you kind of like the modular approach, odds are RStudio (or another IDE) has enough niceties you'll find yourself there pretty often for at least one or two of them
    - And, because the filetypes are all **open**, you can switch to other applications pretty easily

The to-do list

we want to know the **exact code**
that produces the **exact result**
that we **exactly report**

also, we'd like to **keep records**
preferably **open**

*we want to know the **exact code***
*that produces the **exact result***
*that we **exactly report***
*also, we'd like to **keep records***
*preferably **open***

The to-do list

- OK, so we have **code**, that hopefully produces **results**, and we're **open**.
  - **EXCELLENT.** Thanks R and RSTUDIO!
- What about **reporting** and **record-keeping**?
  - We have two important additional tools:
    - **Reporting**: Quarto (and Rmarkdown, and/or LaTeX)
    - **Record-keeping** (aka **version tracking**): git and GitHub

Reporting

- Wait, can't I just write this thing in Word? *What new devilry are you going to foist upon me?*
  - Well, you can.
  - And unless you work in a pretty modern, tech-forward environment, most people do.
  - The replacement of Excel with more powerful (and open) tools is more complete than the replacement of Word
- But once you've started down the open science path, you'll find aspects of Word to be . . . less than desirable

Word

- Do you want to manually copy/paste a figure or table into word every time you re-create it?
- Would you like better interoperability with different file formats?
- Would you like better control over where figures go, rather than them jumping all over pages?
- Declarative formatting, indexing, and the like?

Quarto and Rmarkdown: a notebook approach

- We're going to write an plain-text document that is interspersed with two things:
  - Analysis code *(writing for machines)*
  - Interpretation and explanation *(writing for humans)*
- Imagine we start like this
  - Script files have some nice features, but: we want it to look **pretty**. Namely:
    - No code! Prefer to see results
    - Comments ugly! Prefer nice looking prose

```
#Title: DWV Workshop 1.2.R
#Author: Jack Reilly
#Date: 8.28.25
#Purpose: A Very Basic
Introduction to R +
Constructing Data Frames
#Requires: Nothing
#Output: Nothing

#R is an object oriented
statistical programming
language.

#What does this mean? You
can store whatever* you
want in an object, like a
number:
myfirstobject<-2
myfirstobject

#Things in quotes indicate
"strings"
mysecondobject<-"Hello"
mysecondobject
```

````
---
title: "My First Quarto"
author: "Jack Reilly"
format: html
---

## Quarto

Quarto enables you to weave together content and
executable code into a finished document. To learn
more about Quarto see <https://quarto.org>.

## Running Code

When you click the **Render** button a document
will be generated that includes both content and
the output of embedded code. You can embed code
like this:

```{r}
1 + 1
```

You can add options to executable code like this

```{r}
#| echo: false
2 * 2
```

The `echo: false` option disables the printing of
code (only output is displayed).
````

My First Quarto

AUTHOR
Jack Reilly

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
1 + 1
```

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

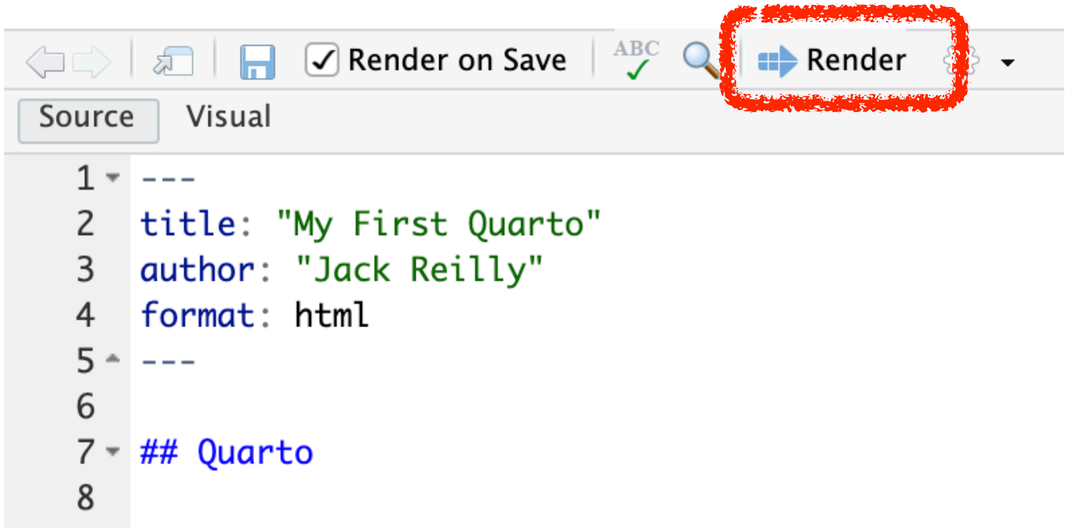
The `echo: false` option disables the printing of code (only output is displayed).

Ooo! Ooo! That's pretty!

- Yes, yes it is. And easy!
- How does it work?
  - Write in plain text with appropriate *markup* indicators
    - Including indicators to let the computer know when you're writing human language to communicate and when you're writing computer language to calculate.
  - **knit** or **render** the document together
    - **Run** all calculations and **typeset** the text, pretty-like

Rough outline

- File>>New File>>Quarto Document



- Declare format, and render it to html, pdf, or docx (or others)

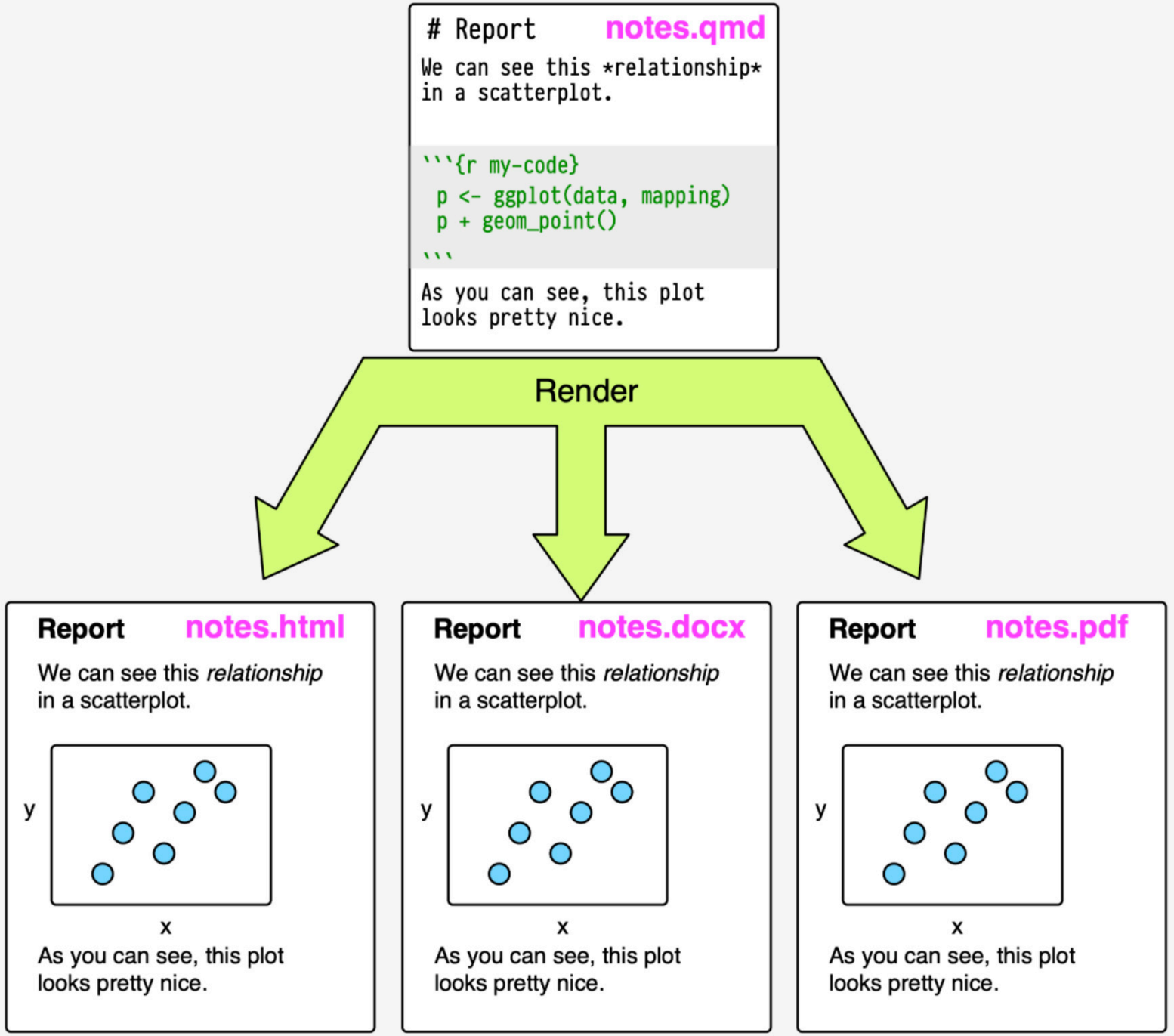


Image Credit: Kieran Healy

Rules of the road

- Basic markdown formatting (for text)
- Quarto *assumes* text unless you specifically declare code with ticks, like so:

```
```{r}
1 + 1
```
```

- If code, will just display result

Desired style	Use the following Markdown annotation
Heading 1	`# Heading 1`
Heading 2	`## Heading 2`
Heading 3	`### Heading 3` (Actual heading styles will vary.)
Paragraph	Just start typing
Bold	`**Bold**`
Italic	`*Italic*`
Images	`[Alternate text for image](path/image.jpg)`
Hyperlinks	`[Link text](https://www.visualizingsociety.com/)`
Unordered Lists	
- First	`- First`
- Second.	`- Second`
- Third	`- Third`
Ordered Lists	
1. First	`1. First`
2. Second.	`2. Second`
3. Third	`3. Third`
Footnote.[1]	`Footnote[^notelabel]`
[1]The note's content.	`[^notelabel] The note's content.`

Image Credit: Kieran Healy

In practice

- Notebooks work **smoothly** when
  - Your document or report is small and self-contained.
  - Your analysis is quick or lightweight.
  - You are making slides.
  - You are making a lot of similar reports from a template.
  - You regularly refer to calculated items in the text of your analysis.
- Notebooks can get **awkward** when
  - Your analysis has many pieces.
  - Your project has many authors.
  - Your analysis needs a lot of cleaning, scaffolding, and other prep-work before you can produce the tables and figures in your output document.
  - You have a lot of different outputs

Monolith vs modular

- Quarto is more of a monolith approach
- Naturally suited to small projects
- Substantial advantages for large, as well, but complexity rises
  - The great thing about plain text files is that you can always ***refactor*** them (split up functionality/purpose across files) to make them ***more modular***
  - *Course website, for instance, is written in quarto files!*
- Other options, as well: LaTeX, etc

Last, but not least

- Our final major tool: **Record Keeping (version tracking)**

Last, but not least

- **Record Keeping (version tracking)**
  - Next week

Keeping the right frame of mind

- This is like learning how to drive a car, or how to cook in a kitchen … or learning to speak a language.
- After some orientation to what's where, you will learn best by doing.
- Software is a pain . . .
  - but at least you won't crash the car or burn your house down.

For now

- Time (hopefully) to troubleshoot any R problems you're having
- Workshop this week: getting used to Quarto and Markdown files