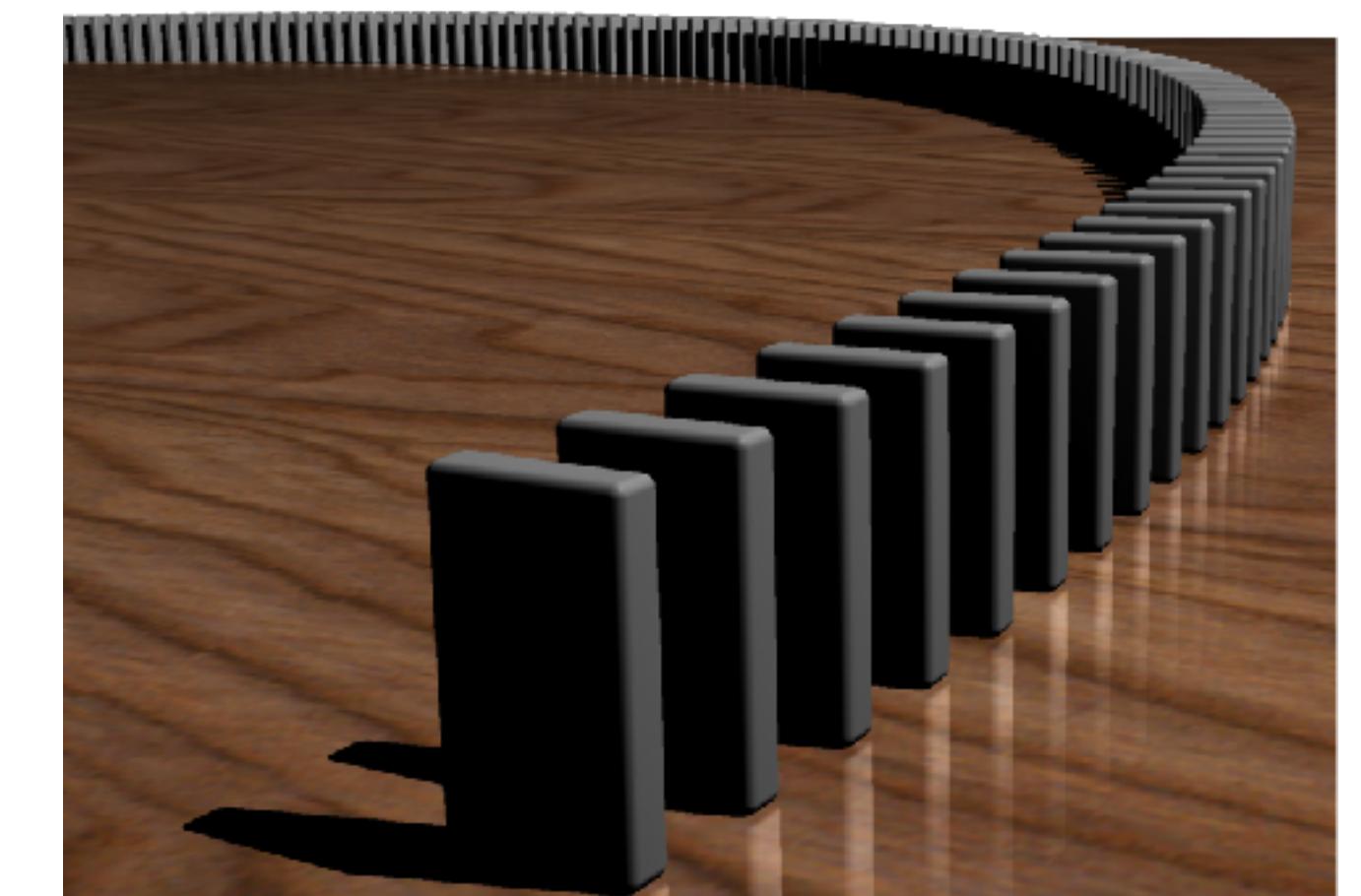


## Lecture 18: Induction and command line tools

Instructor: Michael Szell

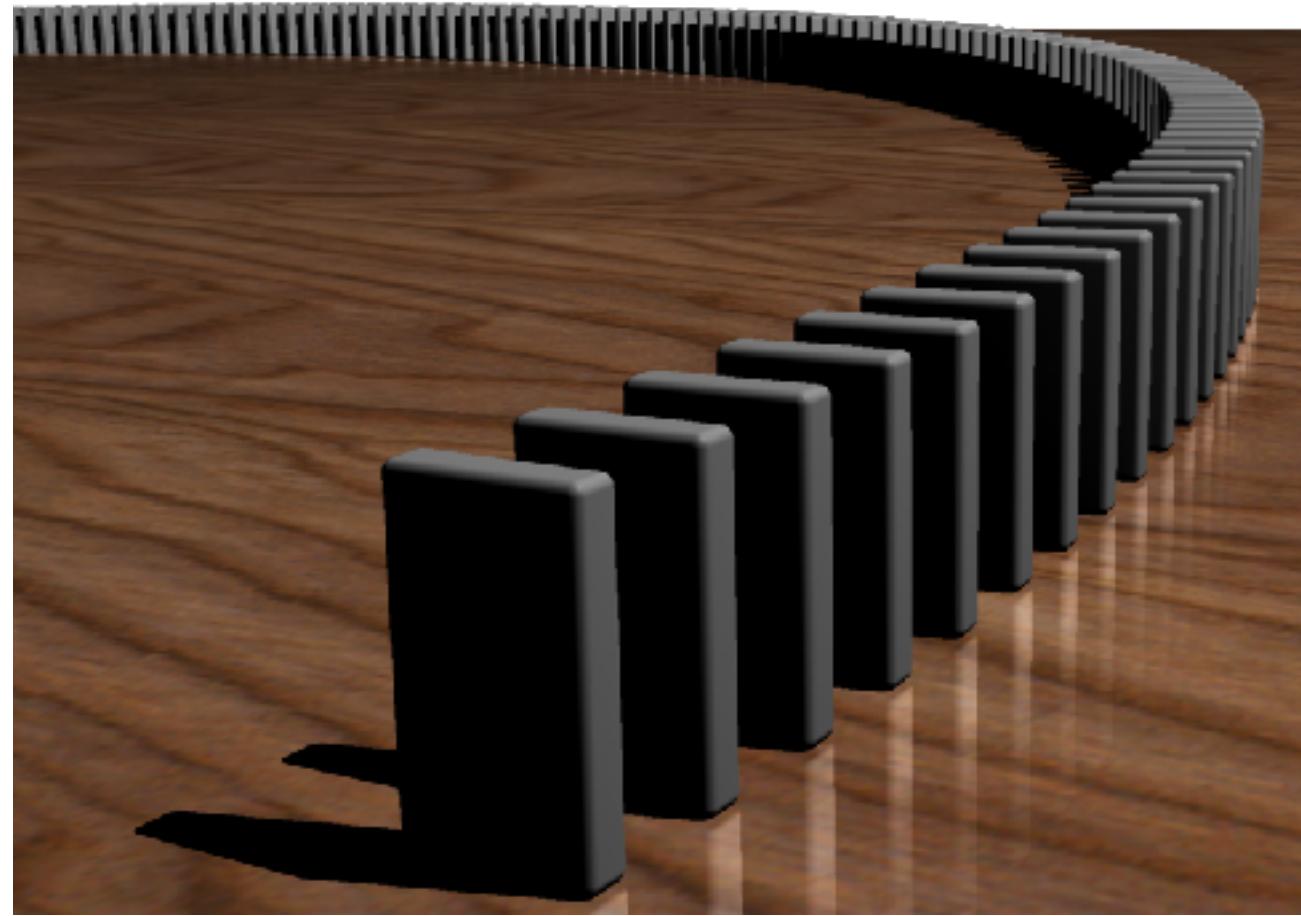
Nov 3, 2023

```
> >> awk cat cd cp  
cut diff du grep  
head less ls man mv  
nl rm sed sort tr  
uniq wc |
```



# Today we learn induction proofs and command line exploration

## Induction



## Exploratory data analysis

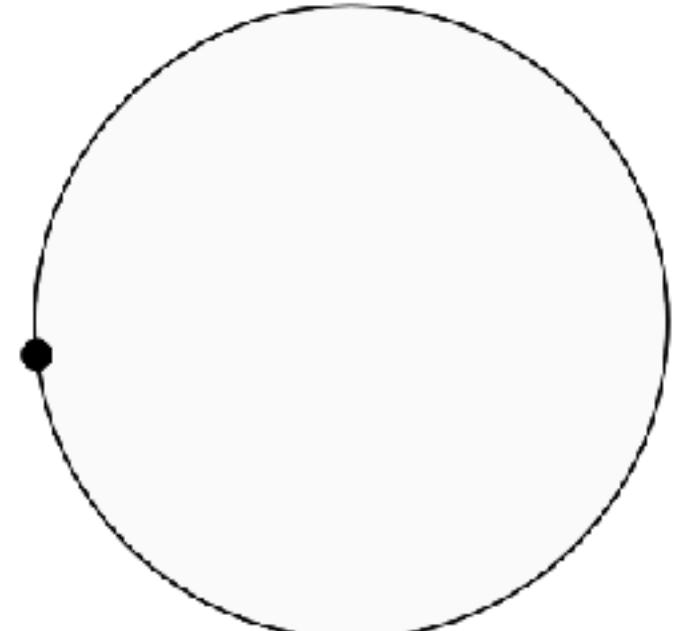


## Command line tools

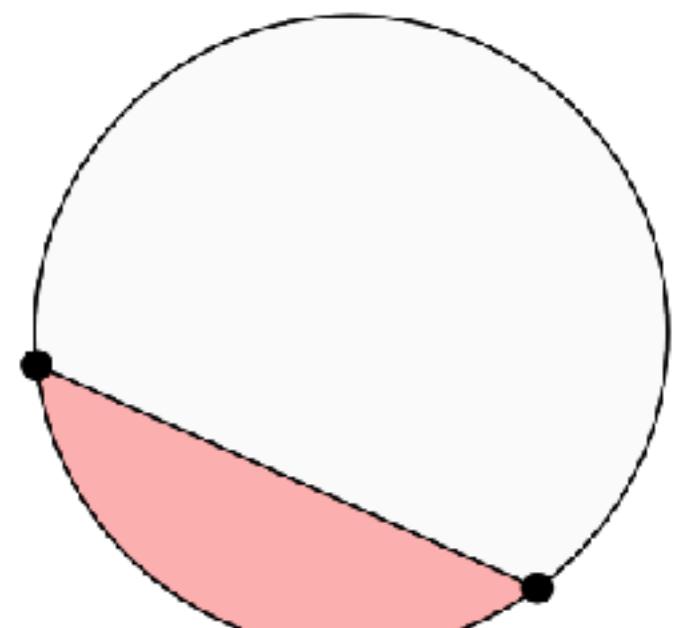
```
> >> awk cat cd cp  
cut diff du grep  
head less ls man mv  
nl rm sed sort tr  
uniq wc |
```

# Induction

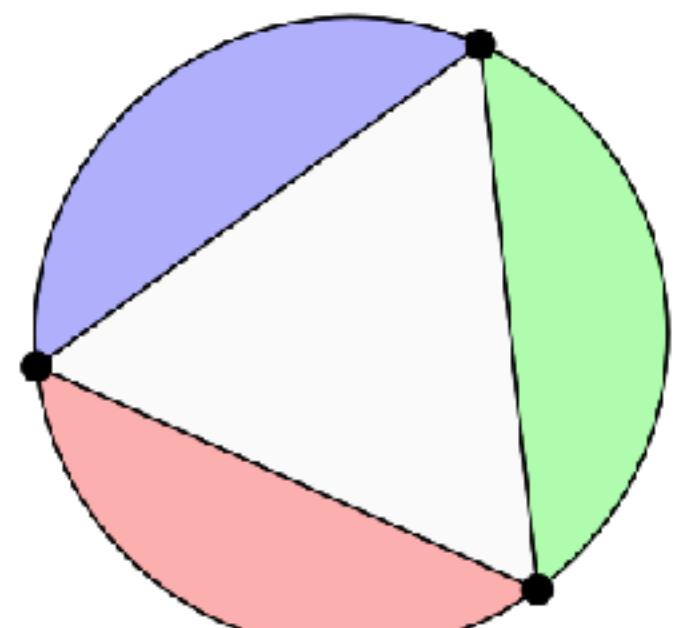
# Let's explore this mathematical problem



$$n=1, c=0, r_g=1$$



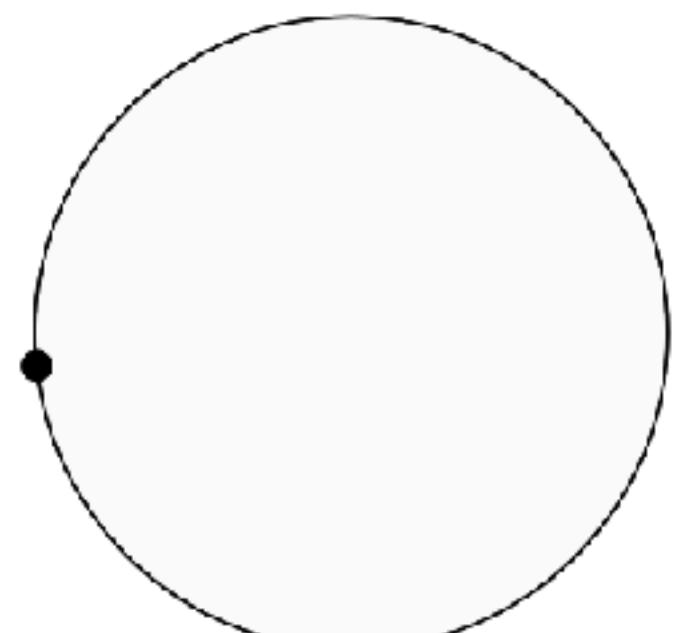
$$n=2, c=1, r_g=2$$



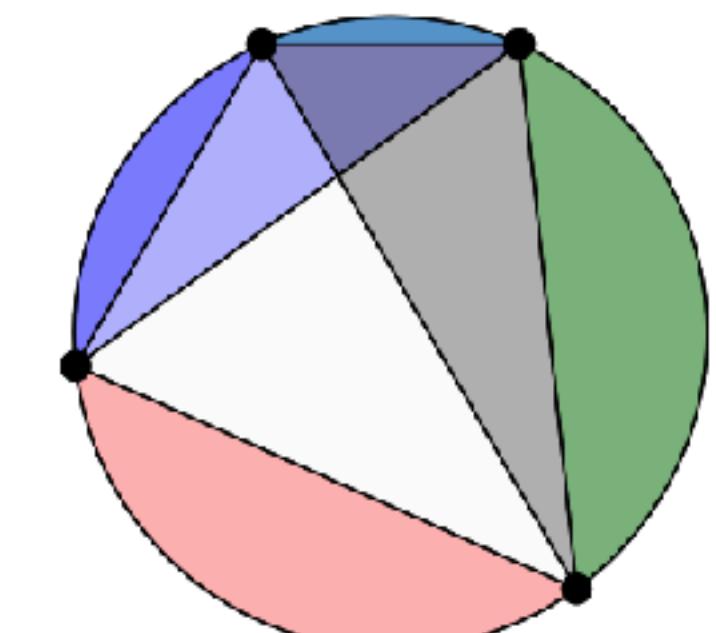
$$n=3, c=3, r_g=4$$

What is the maximum number of areas  $r$  we can divide a circle into by connecting chords between  $n$  points?

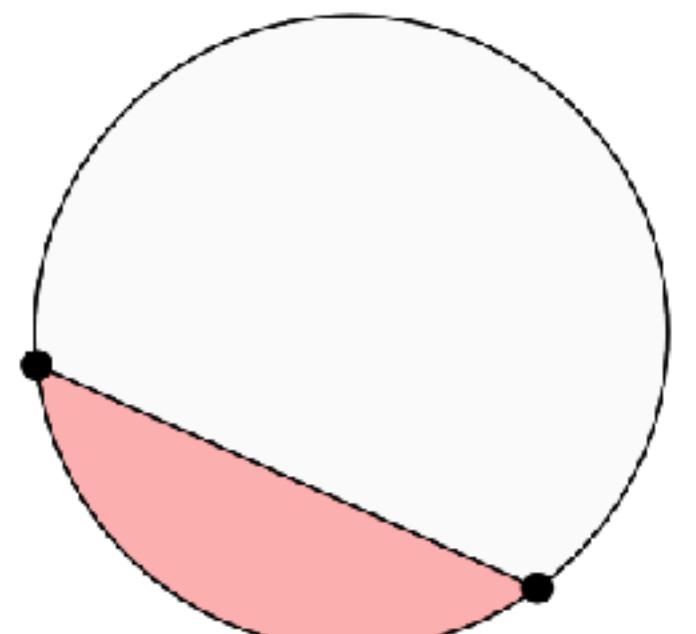
# Let's explore this mathematical problem



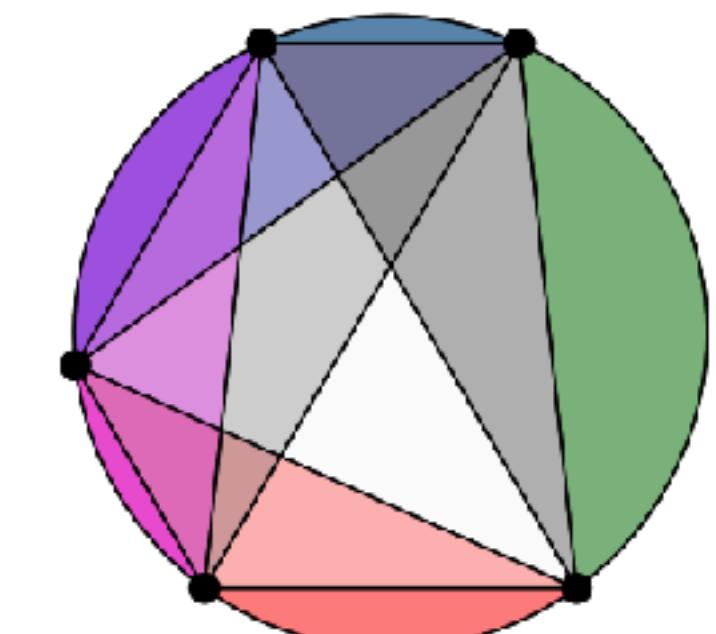
$$n=1, c=0, r_G=1$$



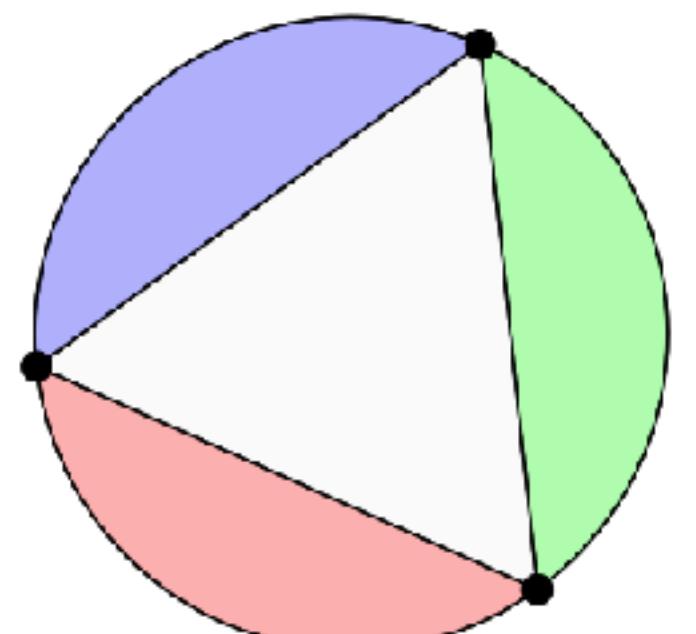
$$n=4, c=6, r_G=8$$



$$n=2, c=1, r_G=2$$



$$n=5, c=10, r_G=16$$

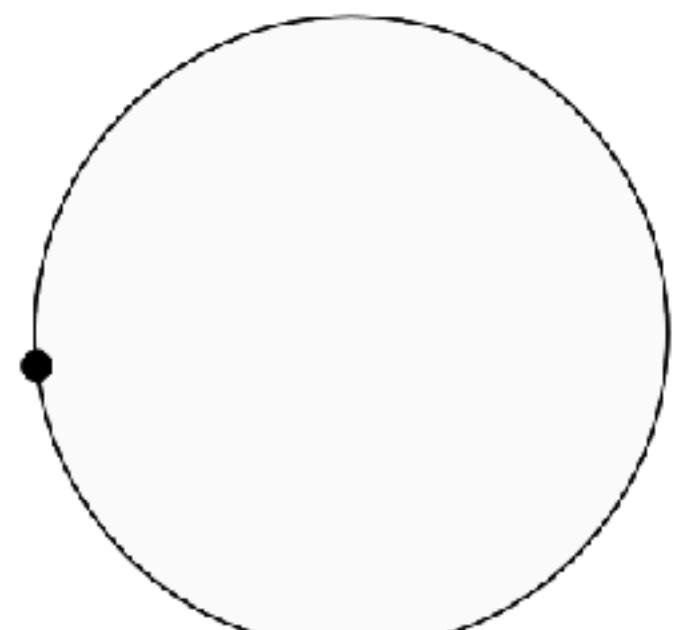


$$n=3, c=3, r_G=4$$

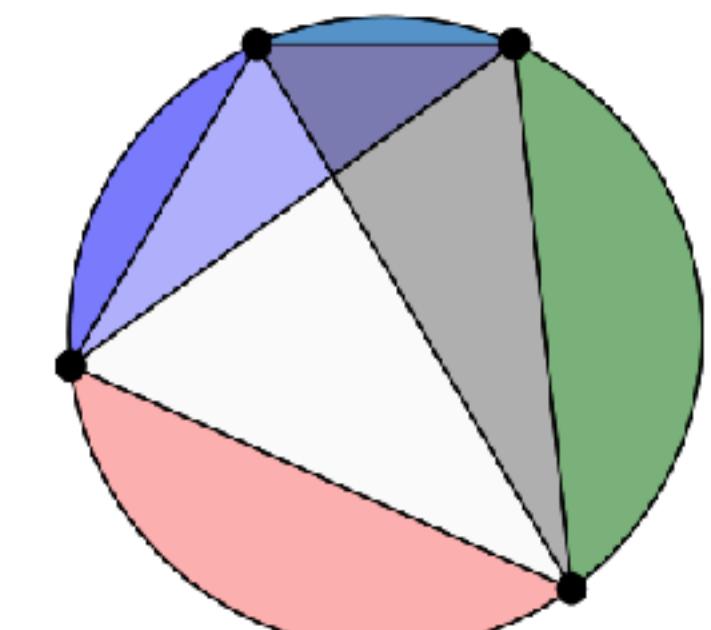
What is the maximum number of areas  $r$  we can divide a circle into by connecting chords between  $n$  points?

Ideas for a general relation?

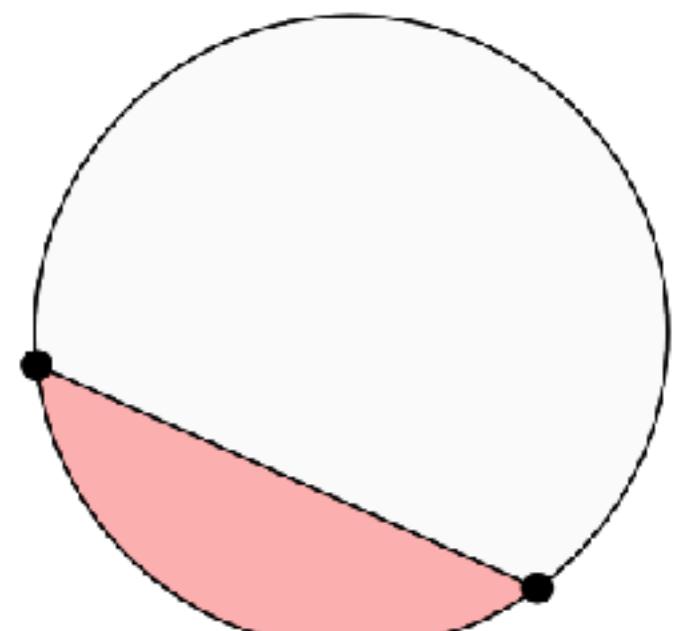
# Let's explore this mathematical problem



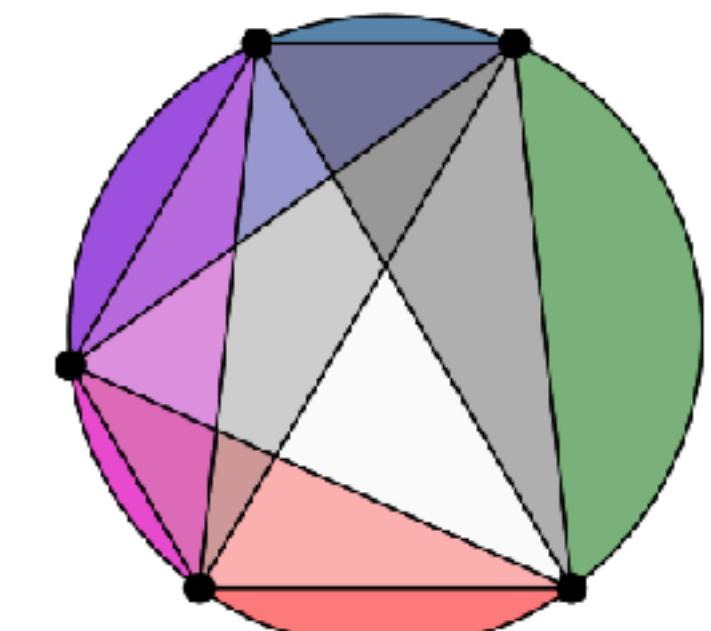
$n=1, c=0, r_G=1$



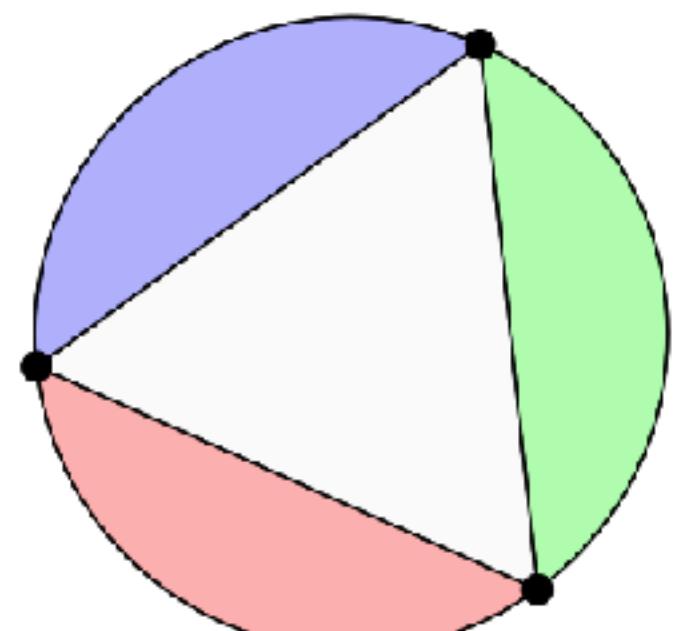
$n=4, c=6, r_G=8$



$n=2, c=1, r_G=2$



$n=5, c=10, r_G=16$

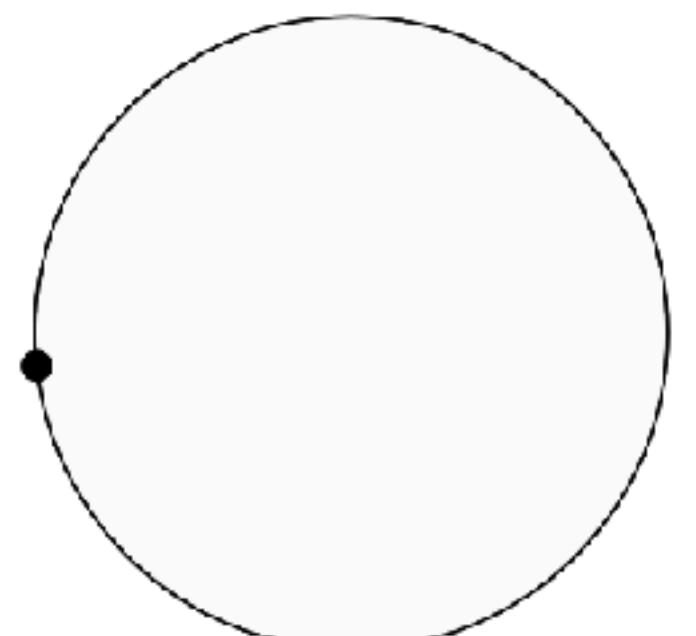


$n=3, c=3, r_G=4$

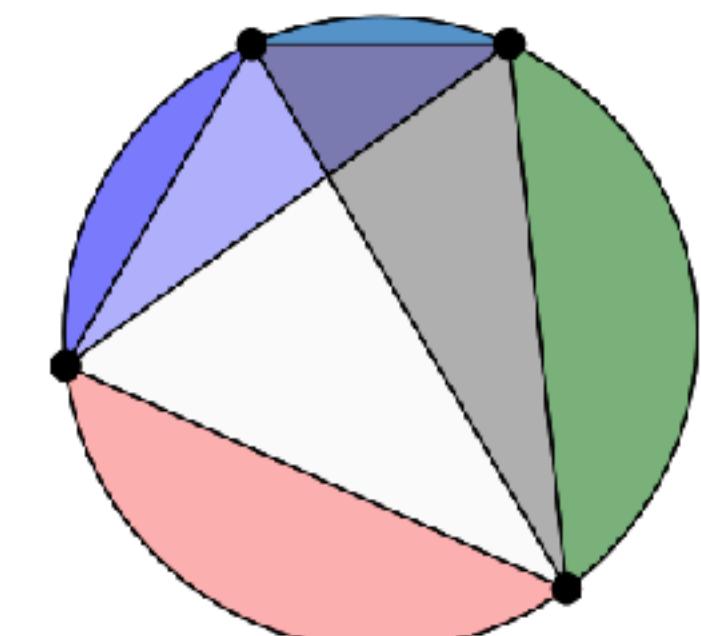
What is the maximum number of areas  $r$  we can divide a circle into by connecting chords between  $n$  points?

Looks like:  $r = 2^{n-1}$

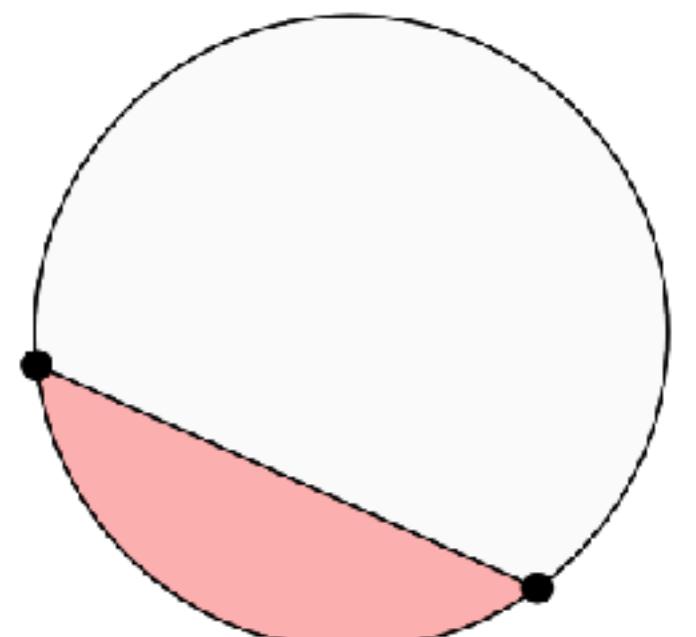
# Let's explore Moser's circle problem



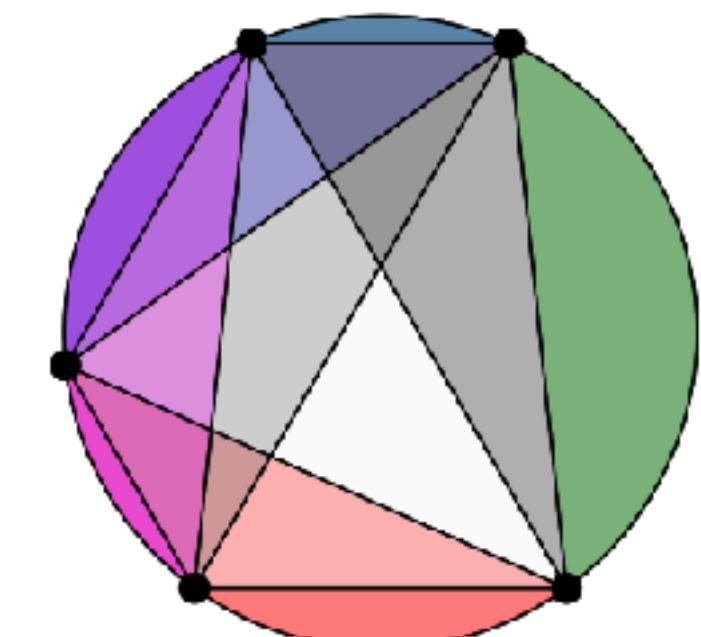
$n=1, c=0, r_G=1$



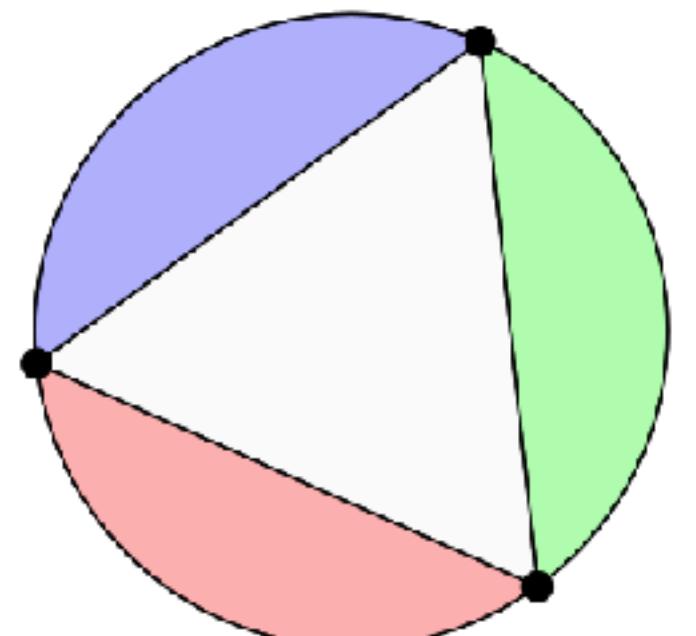
$n=4, c=6, r_G=8$



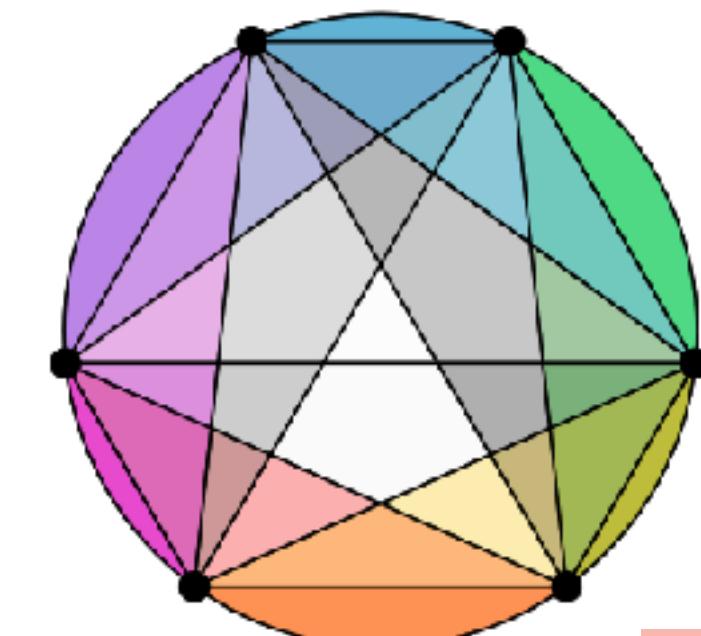
$n=2, c=1, r_G=2$



$n=5, c=10, r_G=16$



$n=3, c=3, r_G=4$



$n=6, c=15, r_G=31$

What is the maximum number of areas  $r$  we can divide a circle into by connecting chords between  $n$  points?

~~Looks like:  $r = 2^{n-1}$~~

# Borwein integral

$$\int_0^\infty \frac{\sin x}{x} = \frac{\pi}{2}$$

# Borwein integral

$$\int_0^\infty \frac{\sin x}{x} = \frac{\pi}{2}$$

# Borwein integral

$$\int_0^\infty \frac{\sin x}{x} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} = \frac{\pi}{2}$$

# Borwein integral

$$\int_0^\infty \frac{\sin x}{x} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} \cdot \frac{\sin \frac{x}{13}}{\frac{x}{13}} = \frac{\pi}{2}$$

# Borwein integral

$$\int_0^\infty \frac{\sin x}{x} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} \cdot \frac{\sin \frac{x}{13}}{\frac{x}{13}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} \cdot \frac{\sin \frac{x}{13}}{\frac{x}{13}} \cdot \frac{\sin \frac{x}{15}}{\frac{x}{15}} =$$

$$\frac{467807924713440738696537864469\pi}{935615849440640907310521750000} \approx 0.9999999999852937186\frac{\pi}{2}$$

[https://math.stackexchange.com/  
questions/111440/examples-of-  
patterns-that-eventually-fail](https://math.stackexchange.com/questions/111440/examples-of-patterns-that-eventually-fail)

# Borwein integral

$$\int_0^\infty \frac{\sin x}{x} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} \cdot \frac{\sin \frac{x}{13}}{\frac{x}{13}} = \frac{\pi}{2}$$

$$\int_0^\infty \frac{\sin x}{x} \cdot \frac{\sin \frac{x}{3}}{\frac{x}{3}} \cdot \frac{\sin \frac{x}{5}}{\frac{x}{5}} \cdot \frac{\sin \frac{x}{7}}{\frac{x}{7}} \cdot \frac{\sin \frac{x}{9}}{\frac{x}{9}} \cdot \frac{\sin \frac{x}{11}}{\frac{x}{11}} \cdot \frac{\sin \frac{x}{13}}{\frac{x}{13}} \cdot \frac{\sin \frac{x}{15}}{\frac{x}{15}} =$$

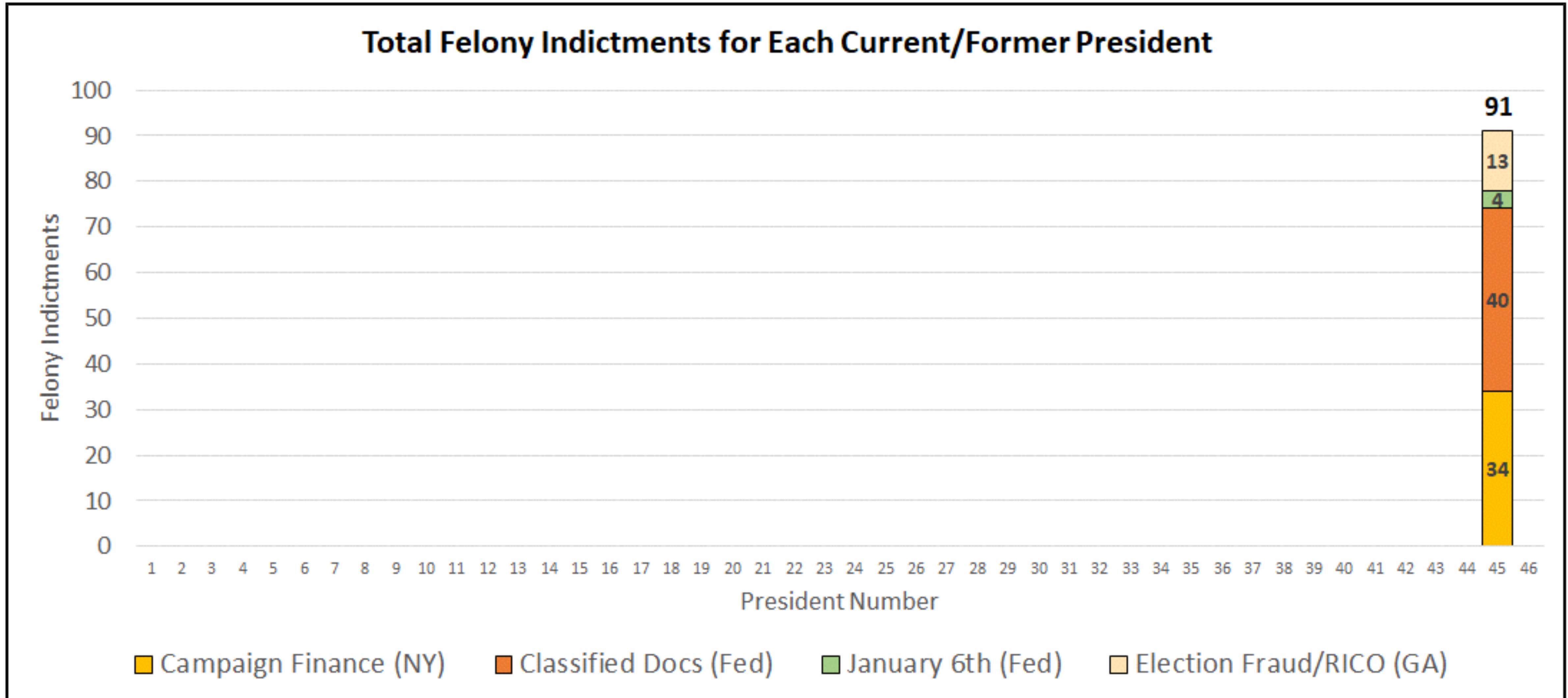
$$\frac{467807924713440738696537864469\pi}{935615849440640907310521750000} \approx 0.9999999999852937186\frac{\pi}{2}$$

Out of **infinite** cases, if you find just **one**, you disproved the claimed rule

Counterexample

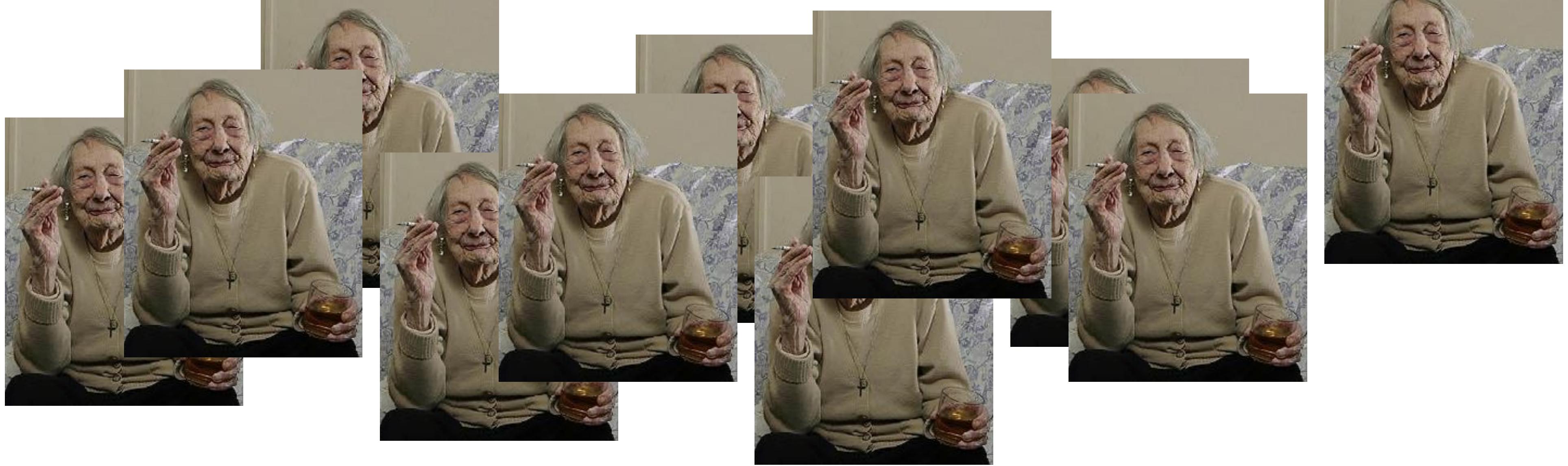
<https://math.stackexchange.com/questions/111440/examples-of-patterns-that-eventually-fail>

# "Not happened" is not proof for "never happening"

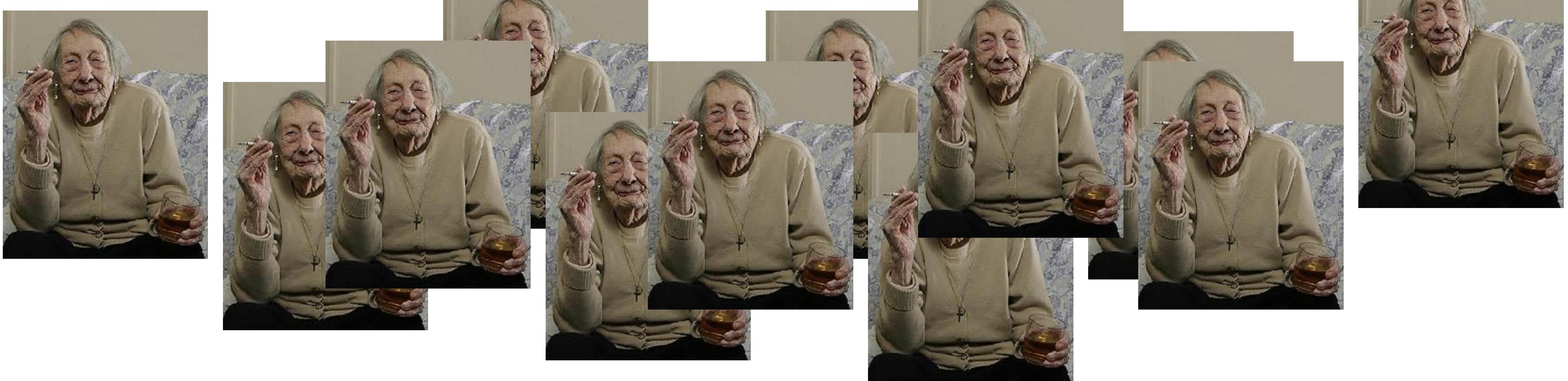




How much data do we need to collect to *prove* a relation?



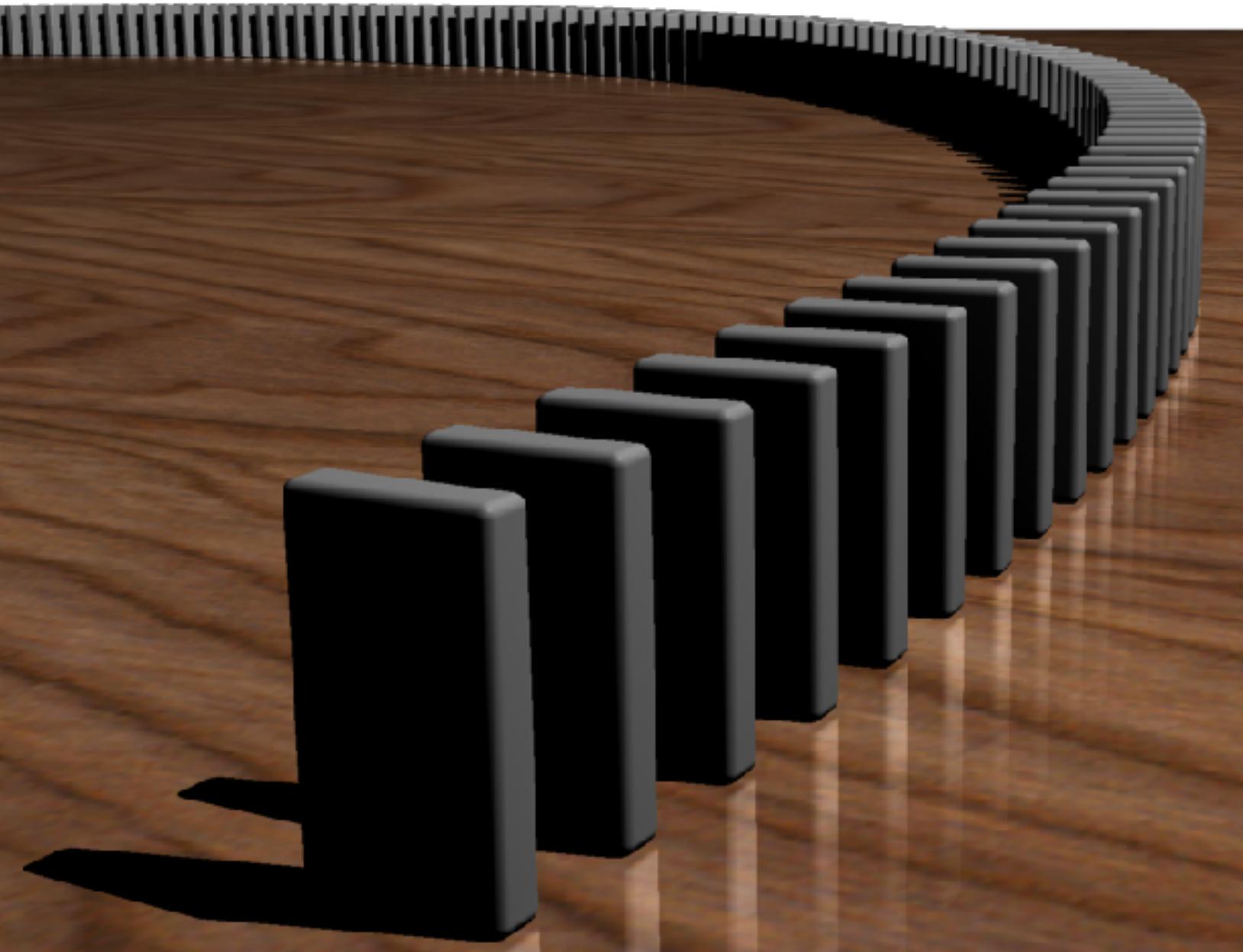
How much data do we need to  
collect to *prove* a relation?



# How much data do we need to collect to *prove* a relation?

In statistics, there are rules of thumb but they are arbitrary.

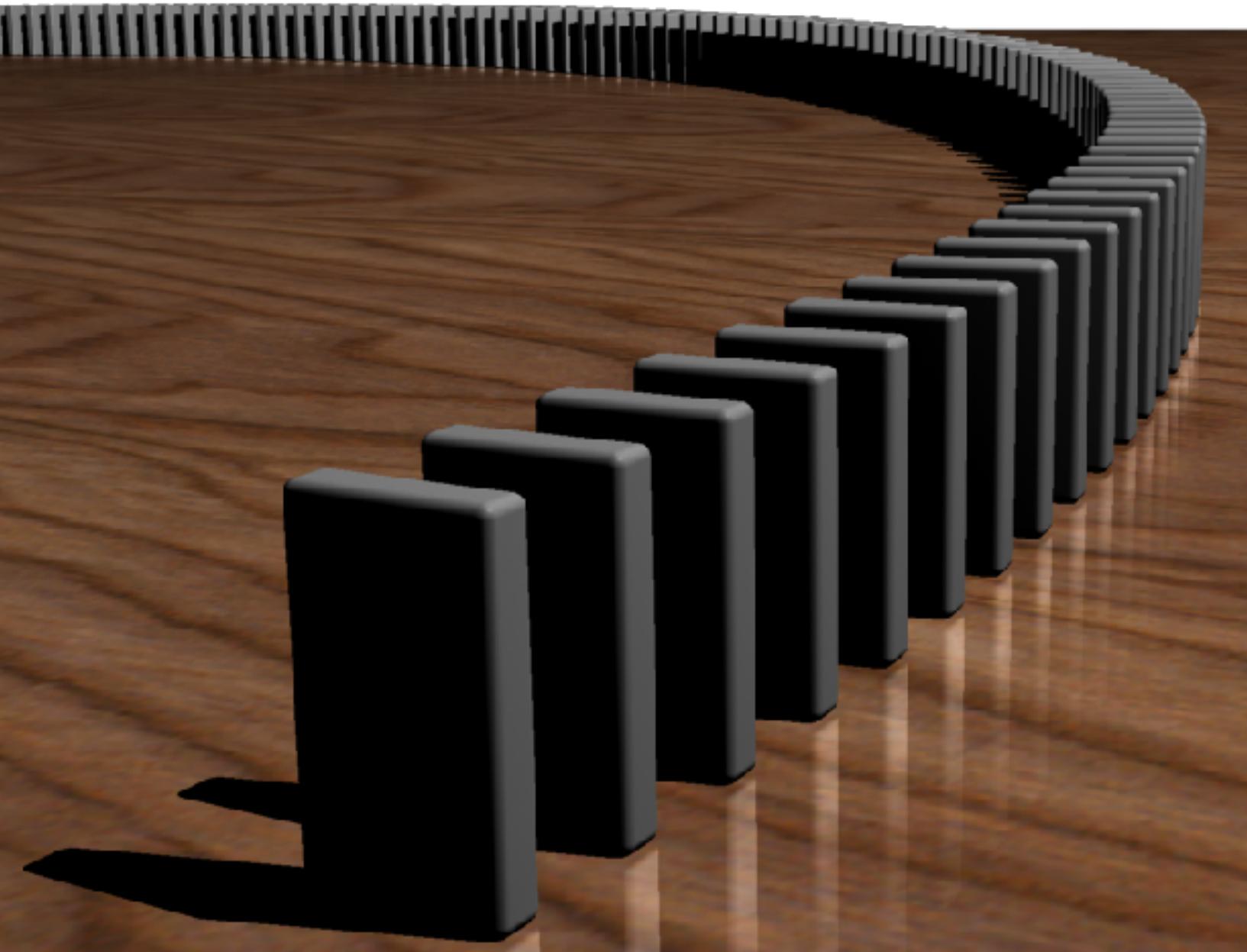
In maths, we can prove things *without doubt*, with induction



## 1) Base case

Show that a statement holds for  $n = 0$  (or a fixed number)

In maths, we can prove things *without doubt*, with **induction**



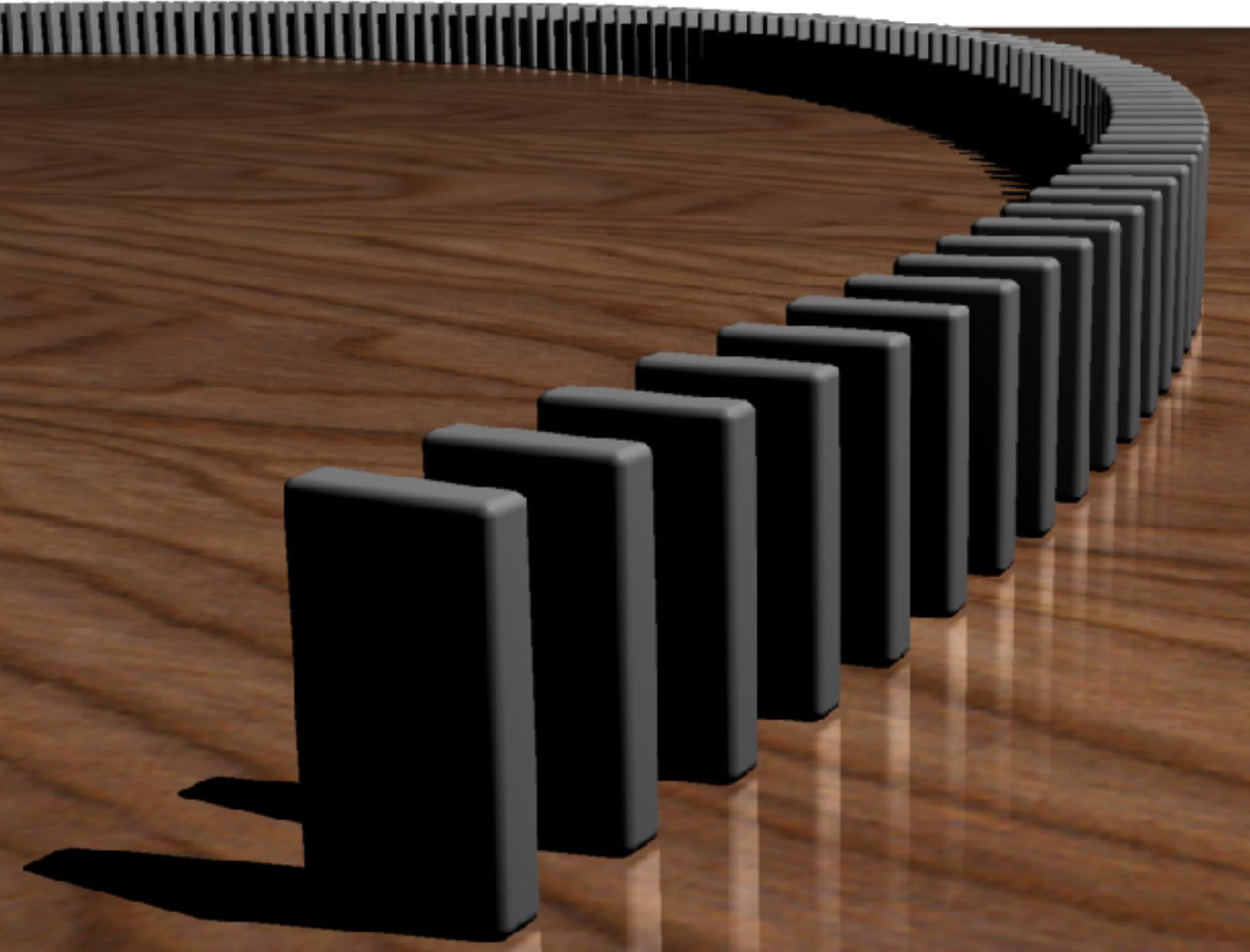
### 1) Base case

Show that a statement holds for  $n = 0$  (or a fixed number)

### 2) Induction step

Show that IF a statement holds for any given  $n = k$   
THEN it holds for  $n = k + 1$

In maths, we can prove things *without doubt*, with **induction**



### 1) Base case

Show that a statement holds for  $n = 0$  (or a fixed number)

### 2) Induction step

Show that IF a statement holds for any given  $n = k$   
THEN it holds for  $n = k + 1$

If you do that, you *Prove* the statement for **all. infinite. numbers. forever.**

Works because numbers are well-ordered: 0,1,2,3,...

# Example: Gauss' sum formula

$$1 + 2 + 3 + \cdots + 100 = ?$$



# Example: Gauss' sum formula

$$1 + 2 + 3 + \cdots + 100 = 5050$$



# Example: Gauss' sum formula

$$1 + 2 + 3 + \cdots + 100 = 5050$$

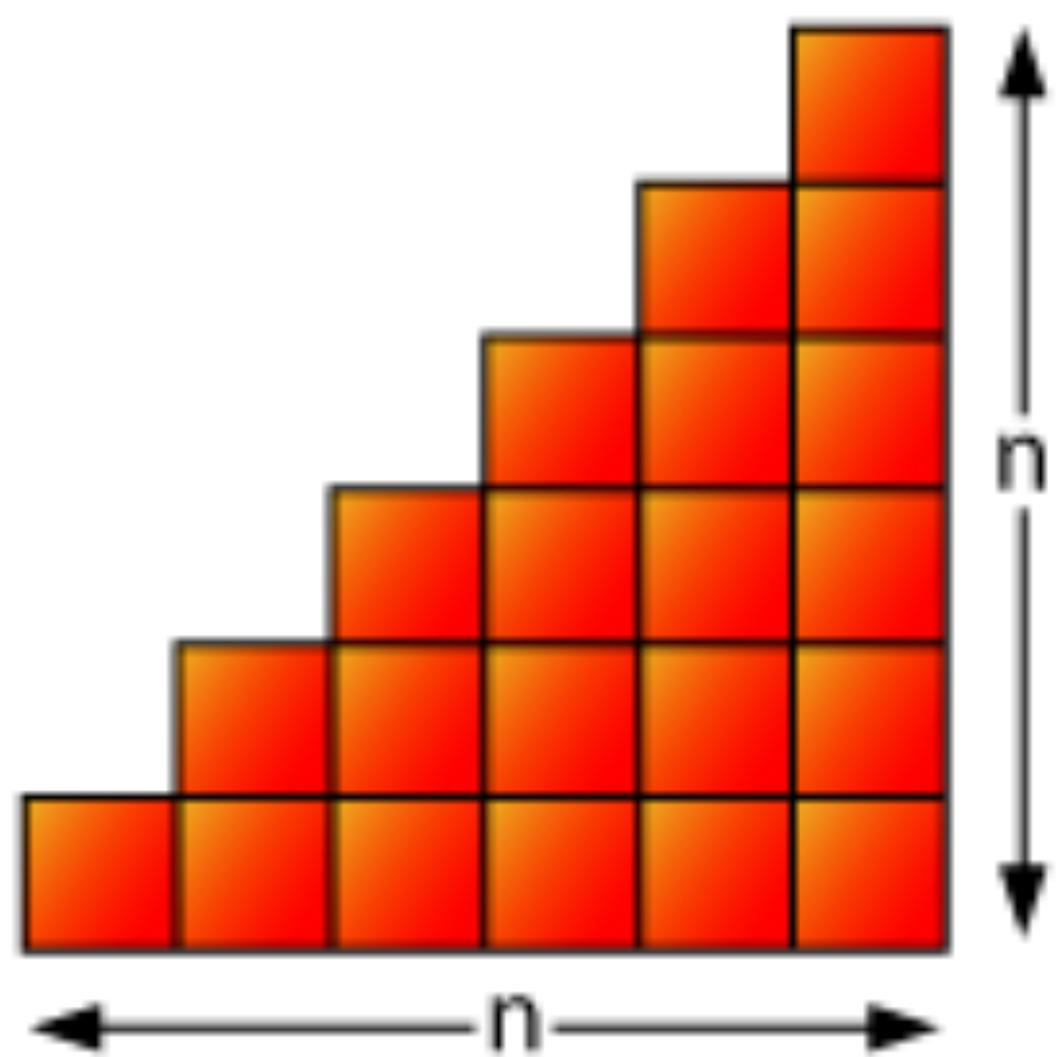
$$1 + 2 + 3 + \cdots + n = ?$$



# Example: Gauss' sum formula

$$1 + 2 + 3 + \cdots + 100 = 5050$$

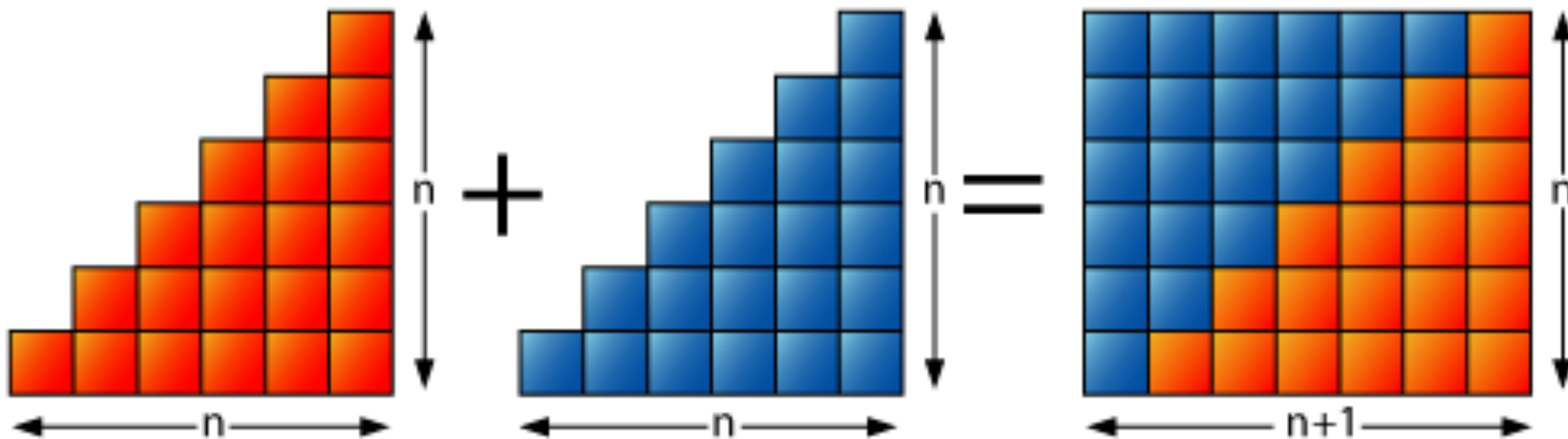
$$1 + 2 + 3 + \cdots + n = ?$$



# Example: Gauss' sum formula

$$1 + 2 + 3 + \cdots + 100 = 5050$$

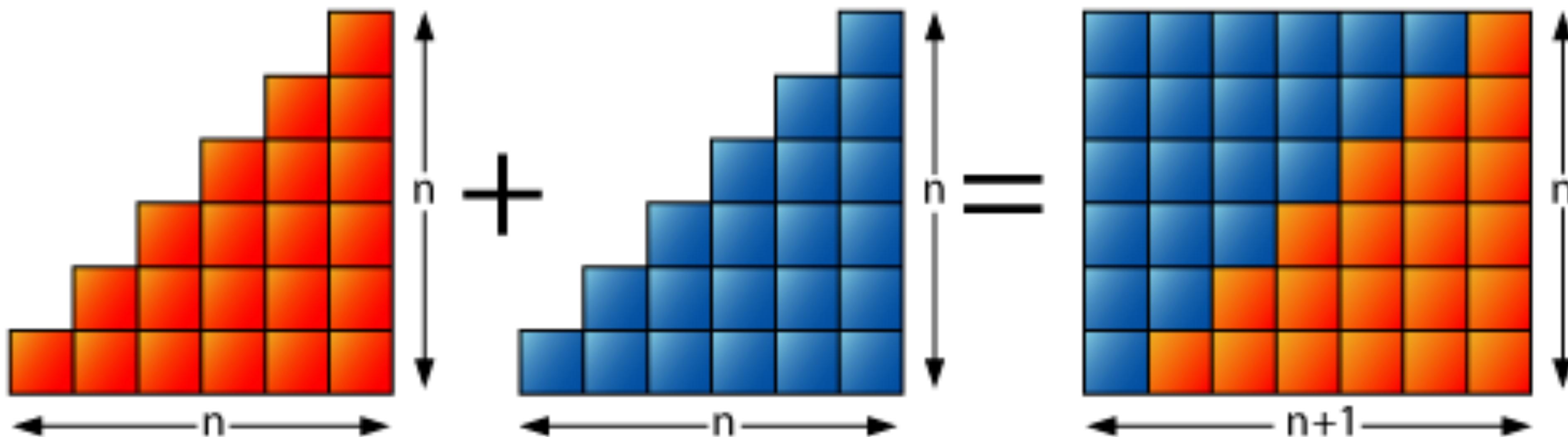
$$1 + 2 + 3 + \cdots + n = ?$$



# Example: Gauss' sum formula

$$1 + 2 + 3 + \cdots + 100 = 5050$$

$$1 + 2 + 3 + \cdots + n = \frac{n(n + 1)}{2}$$

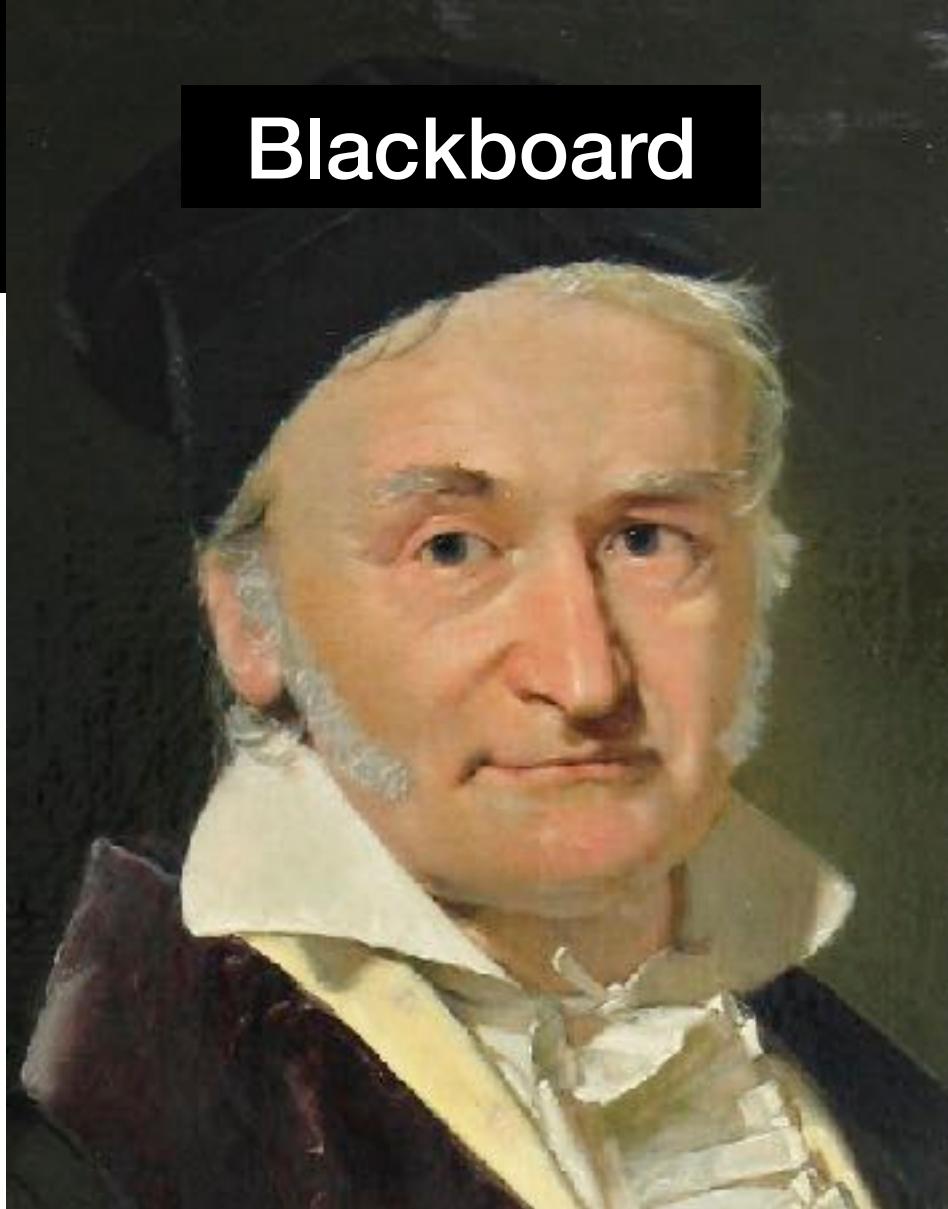


Visual proof

# Example: Gauss' sum formula

$$1 + 2 + 3 + \cdots + 100 = 5050$$

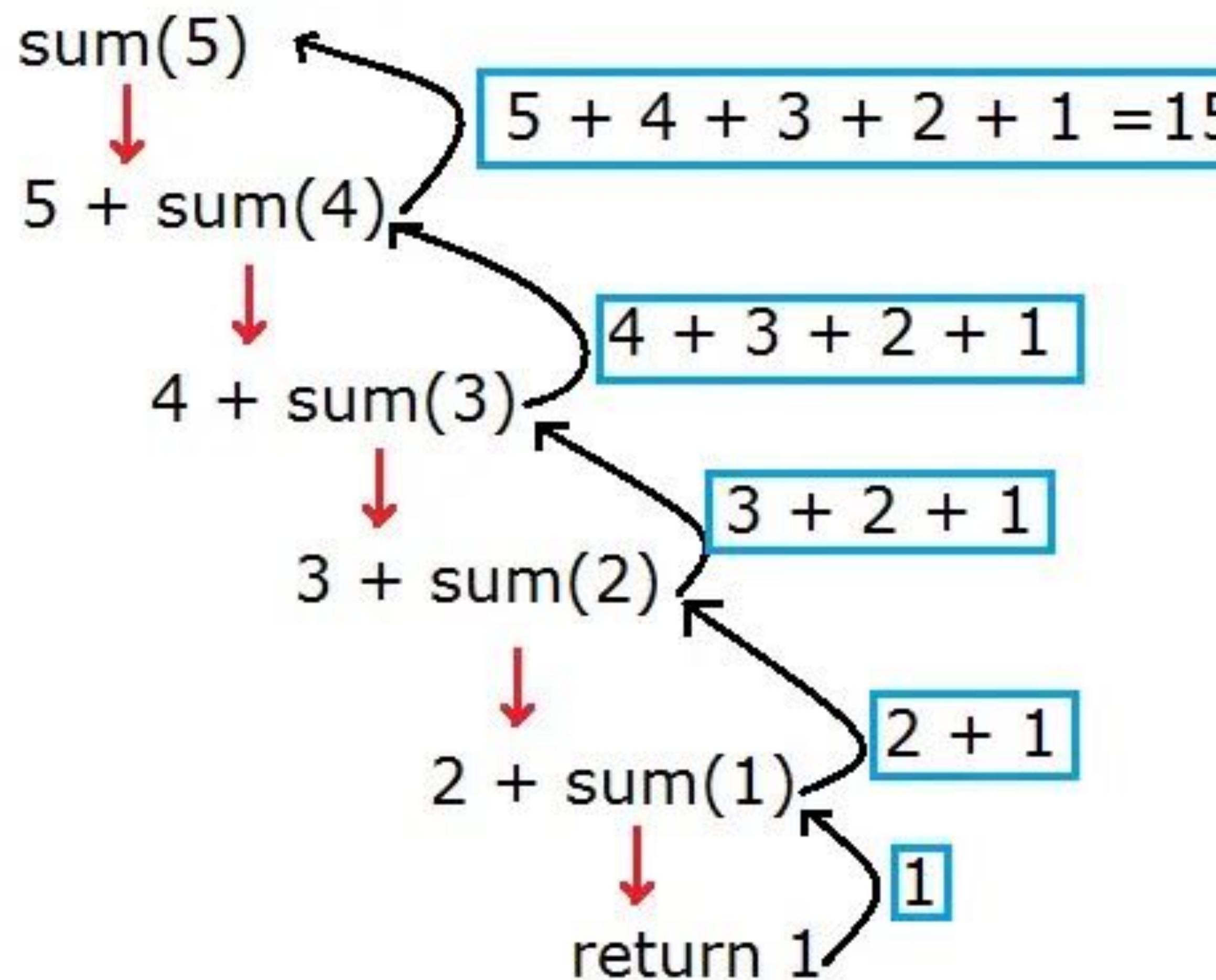
$$1 + 2 + 3 + \cdots + n = \frac{n(n + 1)}{2}$$



Induction proof..

# Induction in maths is what recursion is in computer science

Blackboard



## Question 4: Induction (55 points)

Prove the following by induction. Explicitly state the base case and the induction step.

Prove for all  $n \geq 4$ , that:

$$2n < n!$$

# Data Science at the command line

# Command line tools are really fast

Agile

Augmenting

Scalable

Extensible

Ubiquitous

# Command line tools are really fast

Agile

Augmenting

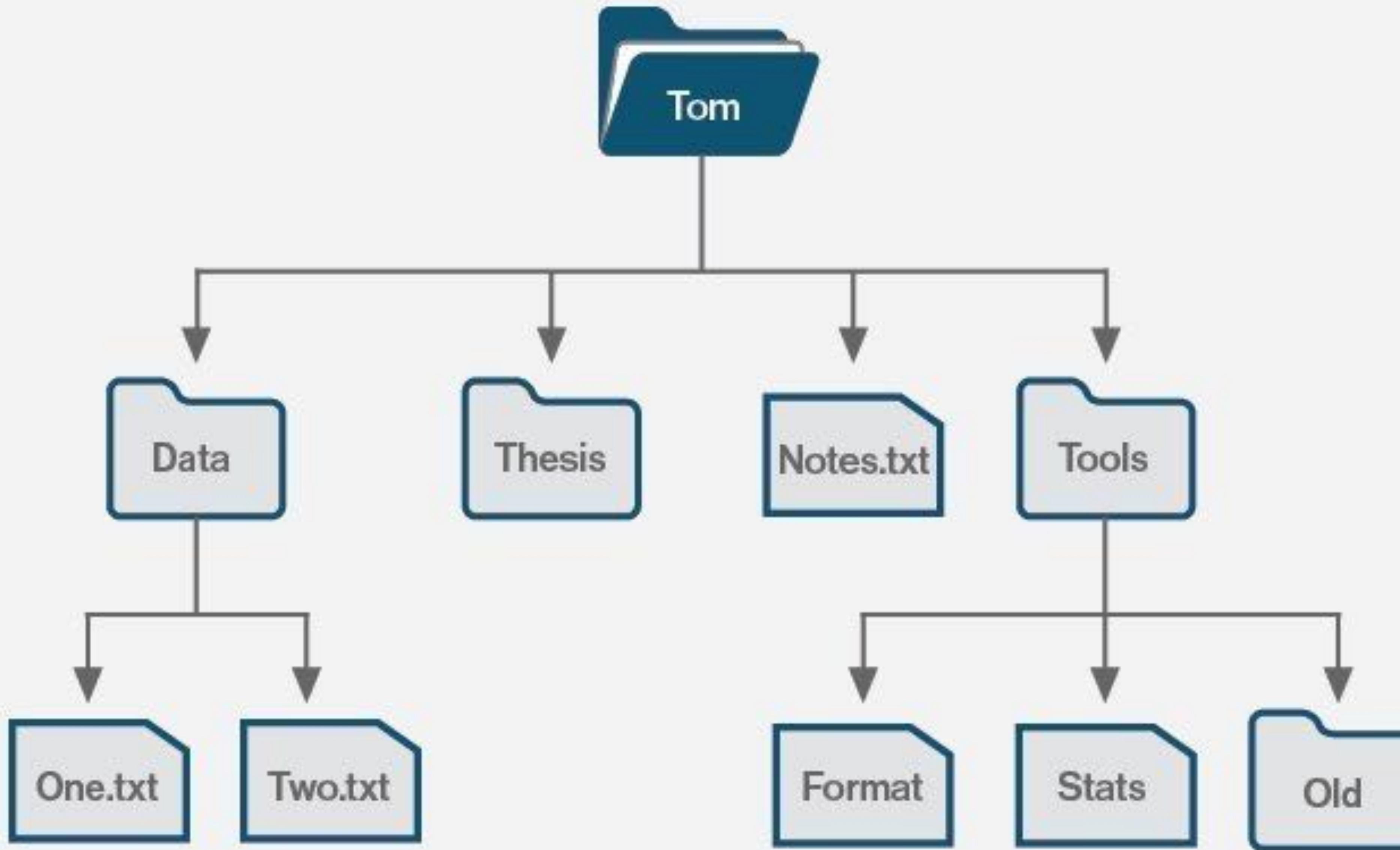
Scalable

Extensible

Ubiquitous

Choice between command line, Python, or even Excel depends on size of your data and what you want to do.

# The file system is the way how files are stored logically



# File system operations you should become familiar with

Terminal

cd change directory

mv move

rm remove

cp copy

ls list directory

du disk usage

cat concatenate, or just output content

# Getting help for a command cmd

Terminal

man cmd

cmd --help

cmd -h

# Many commands have useful options

home



ls



ls -a

# Many commands have useful options

Terminal

home



ls

Show all files  
(including hidden ones)



ls -a

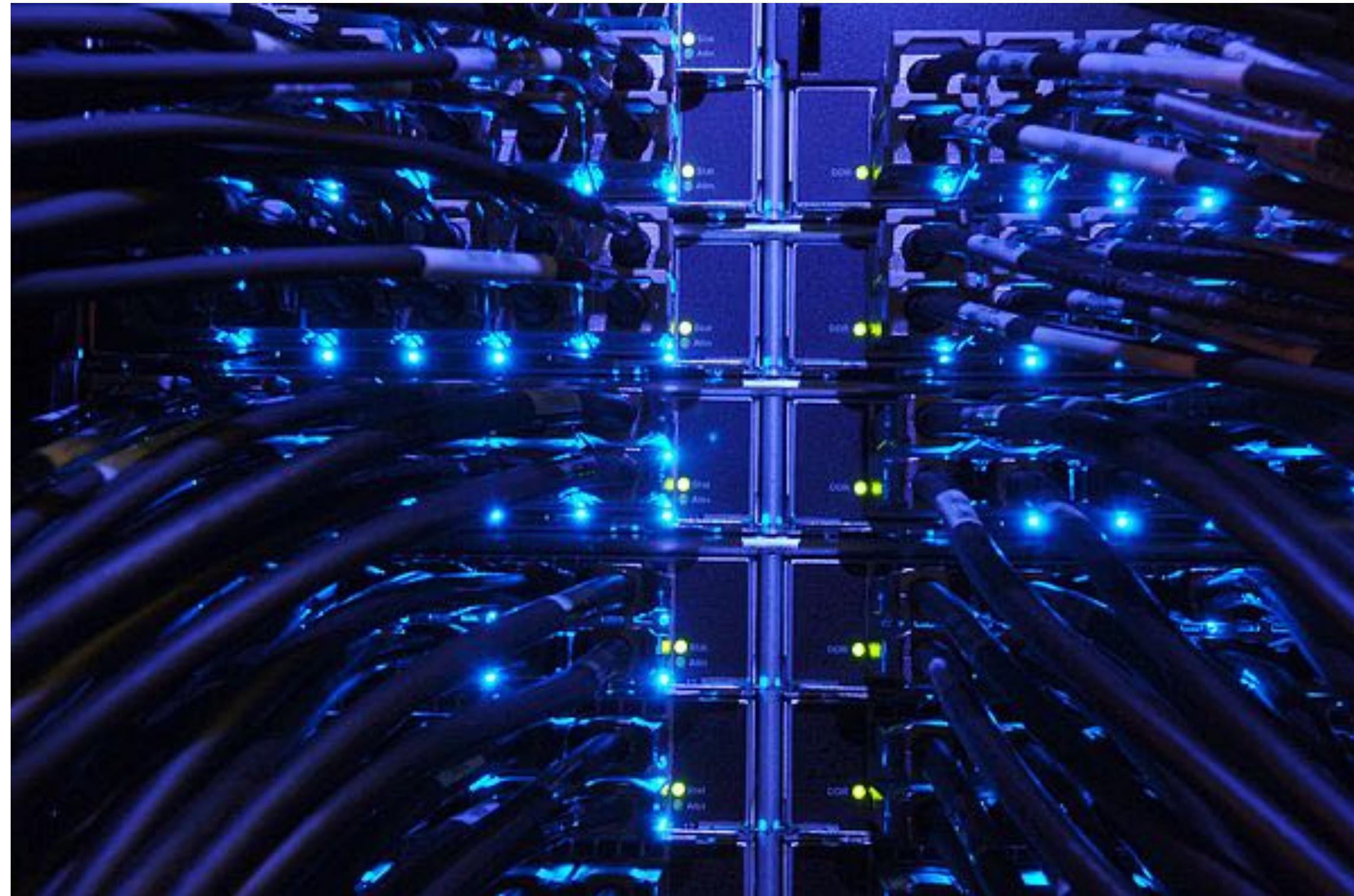
long format  
ls -ahl  
human-readable

# You access most servers with ssh



# Use high-performance computing (HPC) for big computations

Terminal



The HPC cluster offers computational resources for researchers for research projects and students in courses with computational needs.

# Most important tools for data exploration

head

returns top lines

less

navigate file content

wc, wc -l

word count, line count

tr

string replacement

sort

sort

uniq

find unique (adjacent!) lines

cut

cut

diff

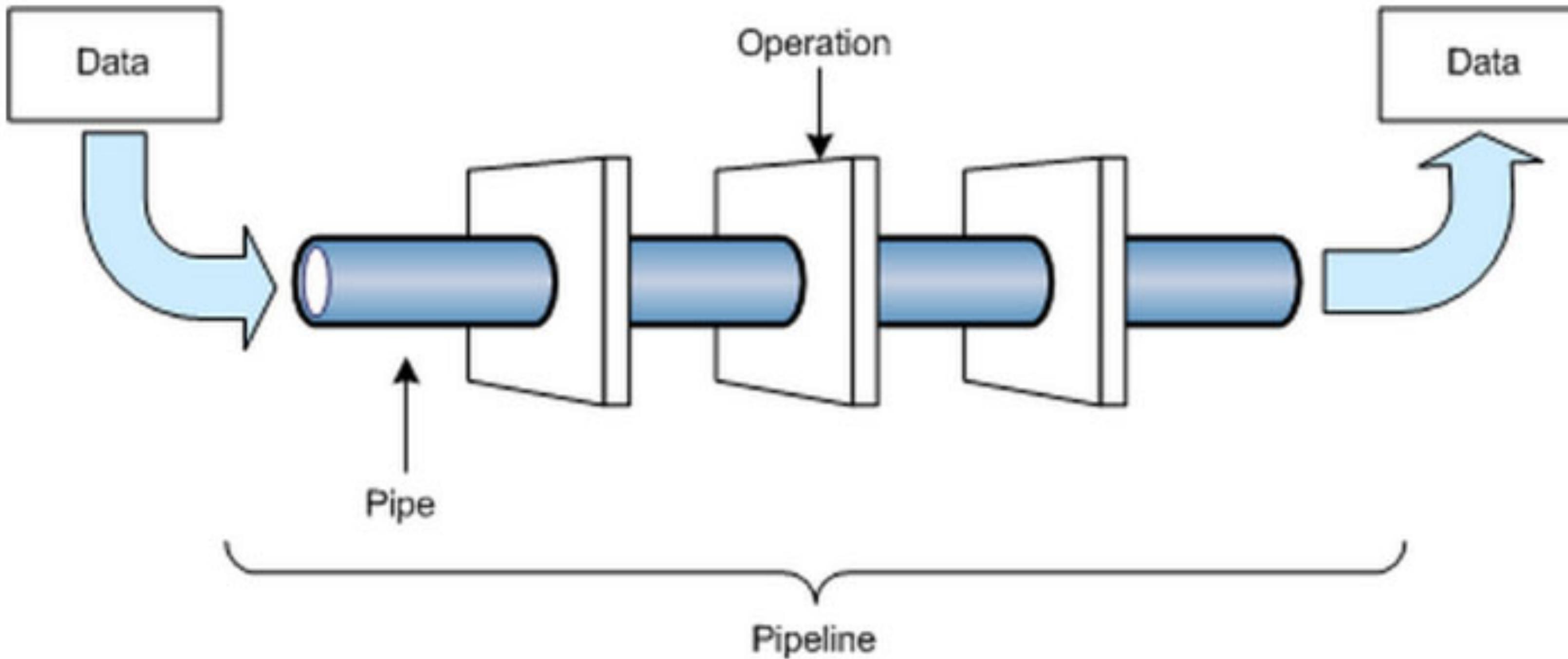
file difference

nl

line number

# The pipe | combines commands into a pipeline

cmd1 | cmd2 | cmd3 | ...



The > redirects your output, >> appends

```
cmd infile.txt > outfile.txt
```

```
cmd infile.txt >> outfile.txt
```

# Tools you should have heard about

grep

globally search for a regular expression and print matching lines

sed

stream editor

additionally to grep, good for replacing strings

# awk is a scripting language, powerful data extraction tool

Example: print the second column in a CSV file

```
awk -F',' '{print $2}' mycsv.csv
```

# awk

JULIA EVANS  
@b0rk

awk is a tiny programming language for manipulating columns of data



I only know how to do 2 things with awk but it's still useful!

basic awk program structure

```
BEGIN{ ... }  
CONDITION {action}  
CONDITION {action}  
END { ... }  
                    ↑  
                    do action on  
                    lines matching  
                    CONDITION
```

extract a column of text with awk

awk -F, '{print \$5}'  
                    ↑  
                    single quotes!  
                    ↑  
column separator     print the 5<sup>th</sup> column



this is 99% of what I do with awk

SO MANY unix commands print columns of text (ps! ls!)

so being able to get the column you want with awk is GREAT

A few more awk programs →

sum the numbers in the 3<sup>rd</sup> column

```
-----  
            action  
{s += $3 }  
-----  
END {print s}  
-----  
            at the end, print  
            the sum!
```

print every line over 80 characters

```
-----  
length($0) > 80  
-----  
            condition
```

(there's an implicit {print} as the action)

# Your Data Science command line alphabet

```
cd mv rm cp ls du cat grep sed man awk head  
less wc tr sort uniq cut diff nl | > >>
```

```
cd mv rm cp ls du cat grep sed man awk head  
less wc tr sort uniq cut diff nl | > >>
```

**Exercise (together):** Make a pipeline that takes `dsalphabet.txt` and creates `dsalphabet_ordered.txt`:

```
> >> awk cat cd cp cut diff du grep head less  
ls man mv nl rm sed sort tr uniq wc |
```

```
cd mv rm cp ls du cat grep sed man awk head  
less wc tr sort uniq cut diff nl | > >>
```

**Exercise (together):** Make a pipeline that takes `dsalphabet.txt` and creates `dsalphabet_ordered.txt`:

```
> >> awk cat cd cp cut diff du grep head less  
ls man mv nl rm sed sort tr uniq wc |
```

**Solution:** `cat dsalphabet.txt | tr " " "\n" | sort | tr "\n" " " > dsalphabet_ordered.txt`

# Data set: Road collisions in the UK in 2019

## Road Safety Data

**Published by:** Department for Transport

**Last updated:** 08 January 2021

**Topic:** Transport

**Licence:** [Open Government Licence](#)

### Summary

[Road Safety Statistics releases](#)



Exploring the data

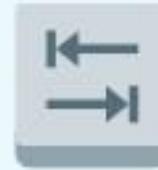
Total number of casualties?

Which are the LSOAs with most accidents?

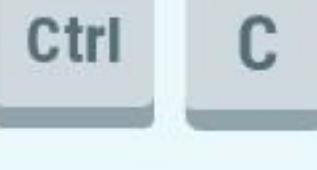
# Get familiar with terminal shortcuts to navigate fast

Terminal

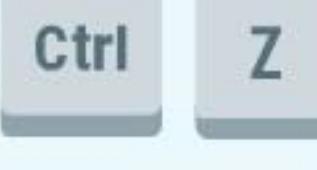
## LINUX TERMINAL SHORTCUTS CHEATSHEET



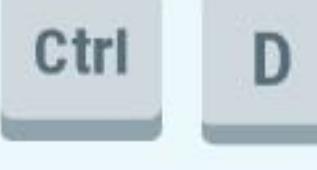
Automatically complete the file, directory, or command you're typing.



Kill the current foreground process running in terminal.



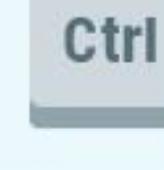
Suspend the current foreground process running in terminal.



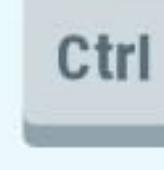
Delete the character at the cursor location.



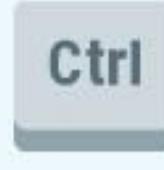
Erases the complete line.



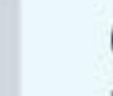
Erase the part of the line after the cursor.



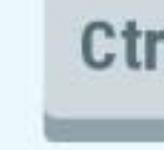
Erase the word before the cursor.



Paste the last thing you cut from the clipboard.



Clears the screen just like clear command.



Move the cursor to the beginning of the line.



Move the cursor to the end of the line.



Go to the previous command in the command history.



Go to the next command in the command history.



LINUX HANDBOOK



<https://linuxhandbook.com/linux-shortcuts/>

<https://linuxhandbook.com/linux-shortcuts/>

<https://hackertyper.net/>

# Jupyter

# In the next year, explore all "mandatory" Unix commands

## List of Unix commands

From Wikipedia, the free encyclopedia

"Unix command" redirects here. For other uses, see [Command \(computing\)](#).

This is a list of [Unix](#) commands as specified by [IEEE Std 1003.1-2008](#), which is part of the [Single UNIX Specification](#) (SUSv4) and can be found on Unix operating systems and most [Unix-like](#) operating systems.

**Contents** [show]

### List [edit]

IEEE Std 1003.1-2008 utilities

Name	Category	Status (Option code)	Description
<a href="#">alias</a>	Misc	Mandatory	Define or display aliases
<a href="#">ar</a>	Misc	Mandatory	Create and maintain <a href="#">library</a> archives
<a href="#">at</a>	Process management	Mandatory	Execute commands at a later time
<a href="#">awk</a>	Text processing	Mandatory	Pattern scanning and processing language
<a href="#">basename</a>	Filesystem	Mandatory	Return non-directory portion of a pathname; see also <a href="#">dirname</a>
<a href="#">batch</a>	Process management	Mandatory	Schedule commands to be executed in a batch queue
<a href="#">bc</a>	Misc	Mandatory	<a href="#">Arbitrary-precision arithmetic</a> language
<a href="#">cat</a>	Filesystem	Mandatory	Concatenate and print files
<a href="#">cd</a>	Filesystem	Mandatory	Change the working directory
<a href="#">chgrp</a>	Filesystem	Mandatory	Change the file group ownership
<a href="#">chmod</a>	Filesystem	Mandatory	Change the file modes/attributes/permissions
<a href="#">chown</a>	Filesystem	Mandatory	Change the file ownership
<a href="#">cksum</a>	Filesystem	Mandatory	Write file <a href="#">checksums</a> and sizes

# Why we generally do not use Excel in Data Science



Lack of:

- Reproducibility
- Version control
- Testing
- Maintainability
- Accuracy

# Why we generally do not use Excel in Data Science

Comment | [Open Access](#) | Published: 23 August 2016

## Gene name errors are widespread in the scientific literature

[Mark Ziemann](#), [Yotam Eren](#) & [Assam El-Osta](#) 

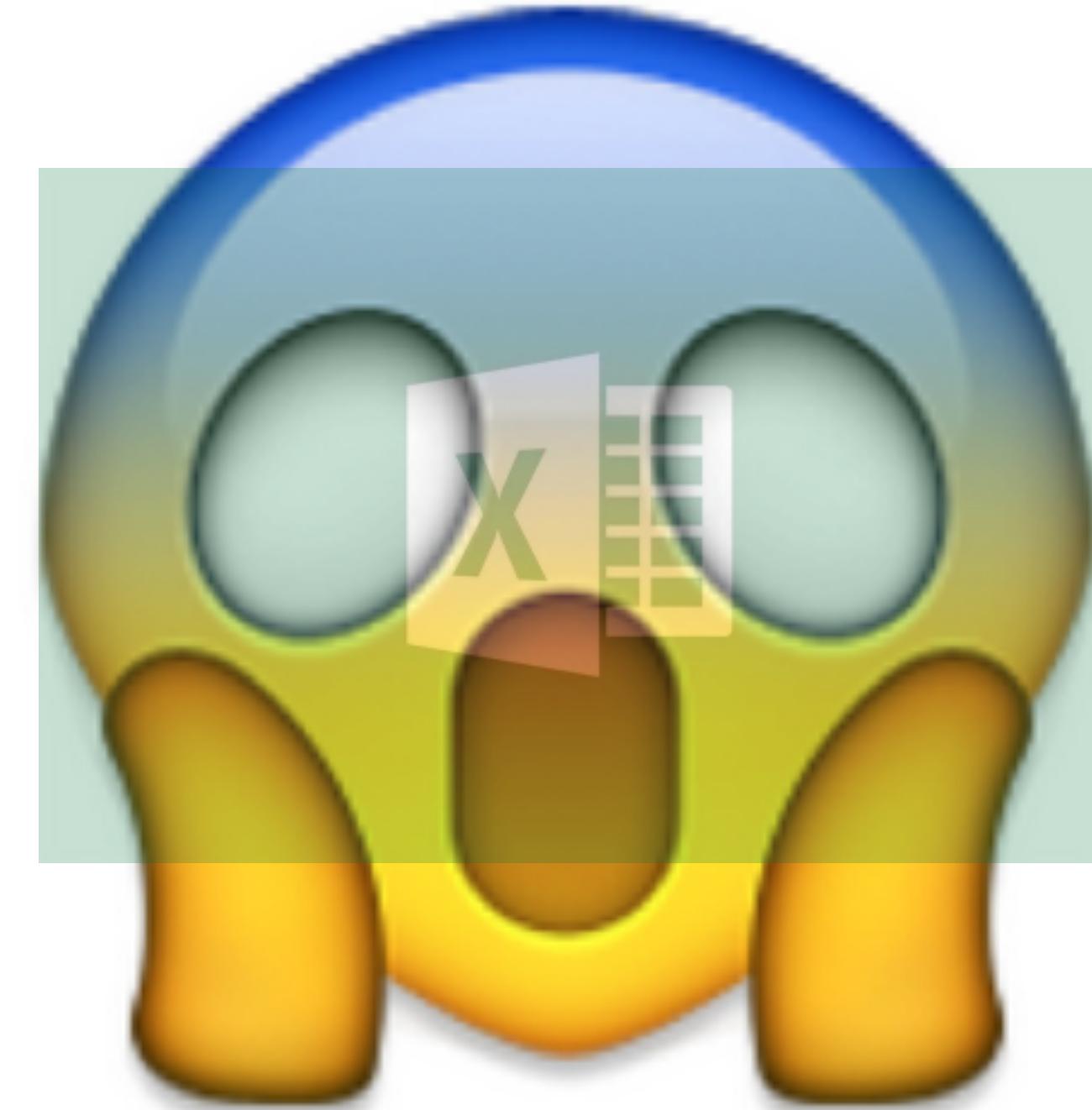
[Genome Biology](#) 17, Article number: 177 (2016) | [Cite this article](#)

127k Accesses | 45 Citations | 2567 Altmetric | [Metrics](#)

### 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.

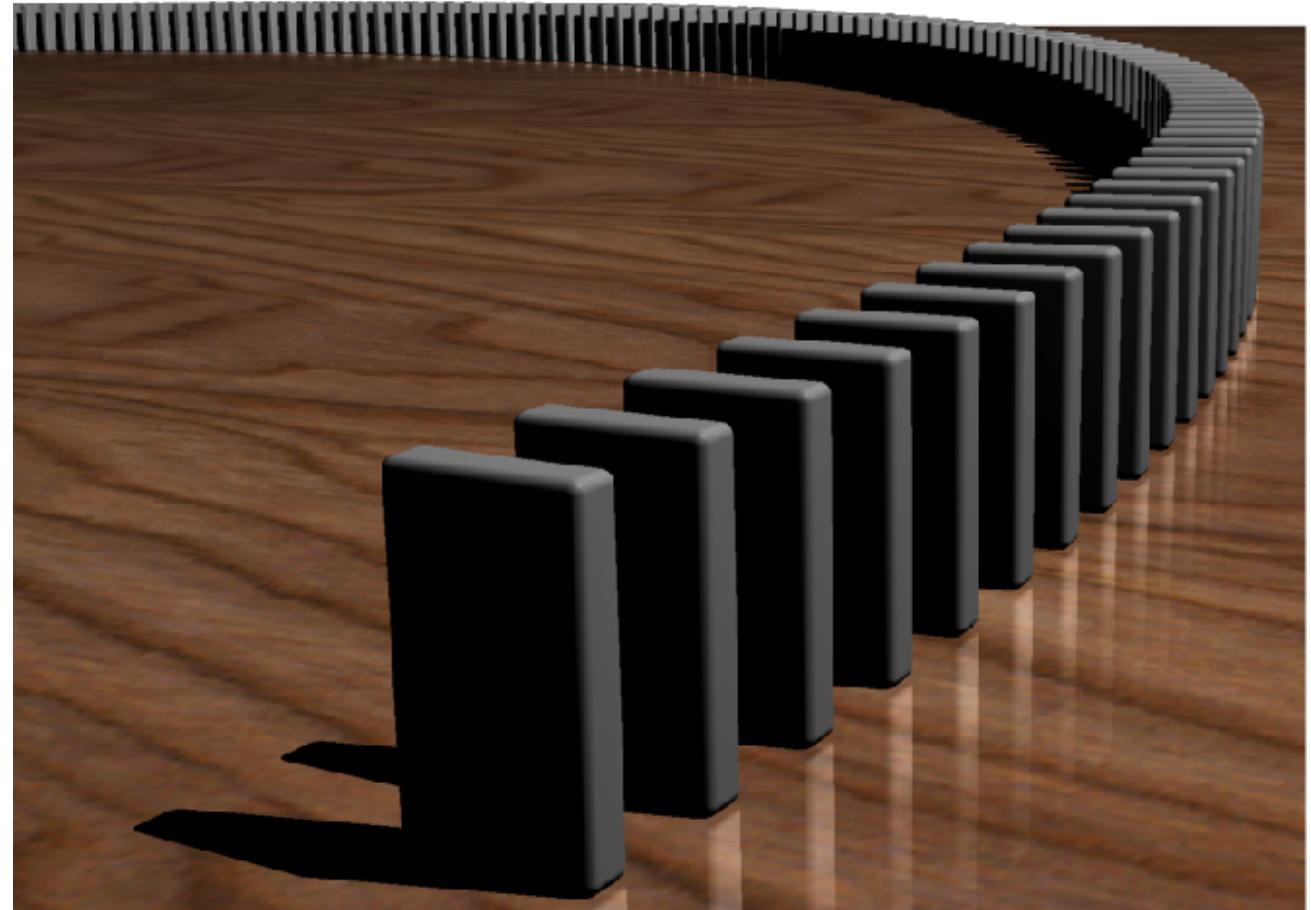
The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since that report, we have uncovered further instances where gene symbols were converted to dates in supplementary data of recently published papers (e.g. '*SEPT2*' converted to '2006/09/02'). This suggests that gene name errors continue to be a problem in supplementary files accompanying articles. Inadvertent gene symbol conversion is problematic because these supplementary files are an important resource in the genomics community that are frequently reused. Our aim here is to raise awareness of the problem.



# Jupyter

# Take home message for today

Induction truly proves mathematical relationships, not probabilistically

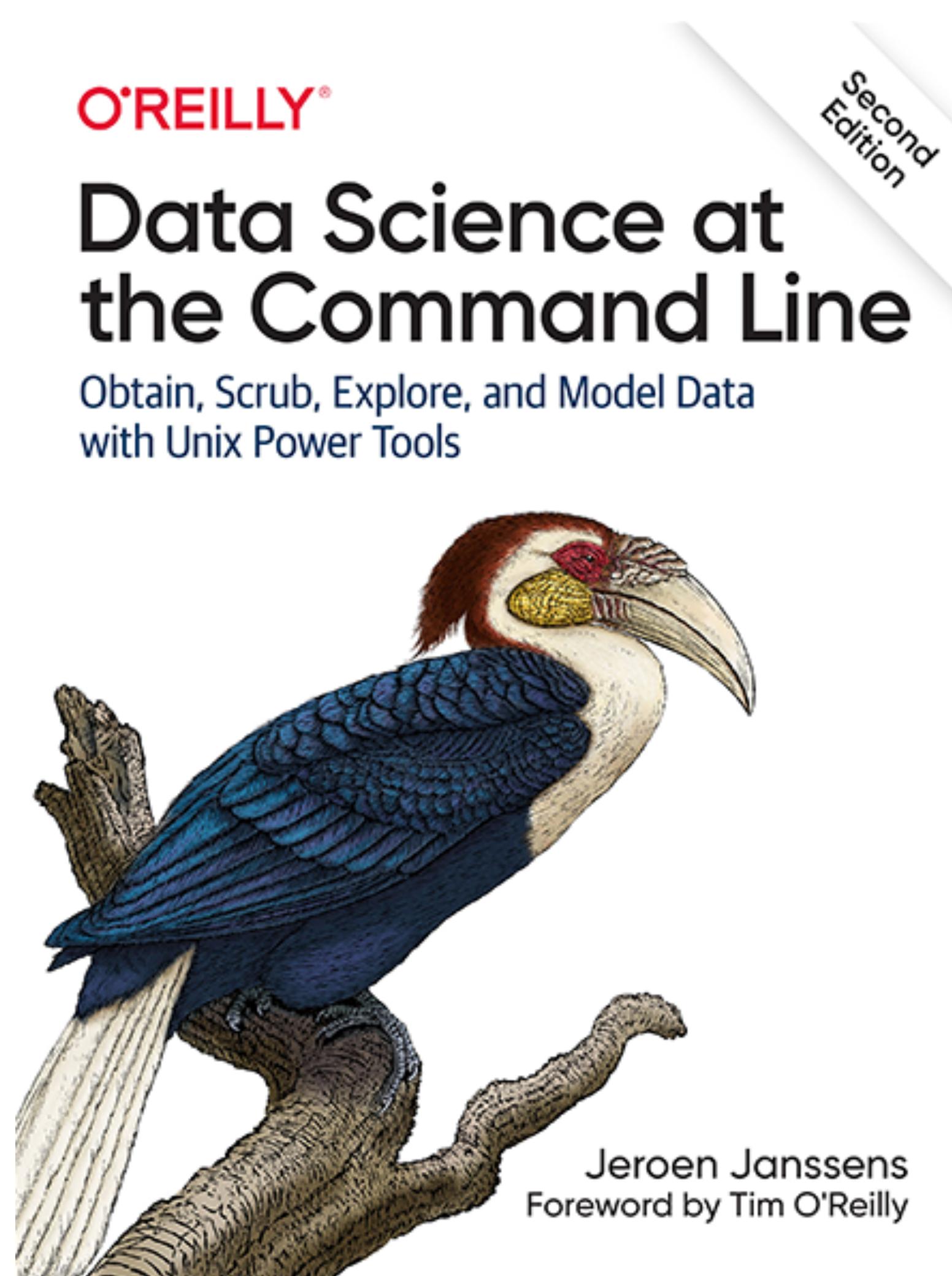


Command-line tools take learning but are the fastest way to explore big data  
(Beware of Excel)



Working with servers requires command line know-how

# Sources and further materials for today's class



<https://jeroenjanssens.com/dsatcl/>

<http://mywiki.wooledge.org/BashGuide/CommandsAndArguments>

<https://en.wikipedia.org/wiki/AWK>

[https://en.wikipedia.org/wiki/Pipeline\\_\(Unix\)](https://en.wikipedia.org/wiki/Pipeline_(Unix))

<https://linuxhandbook.com/linux-shortcuts/>

<https://datascience.stackexchange.com/questions/5443/do-data-scientists-use-excel>

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7>