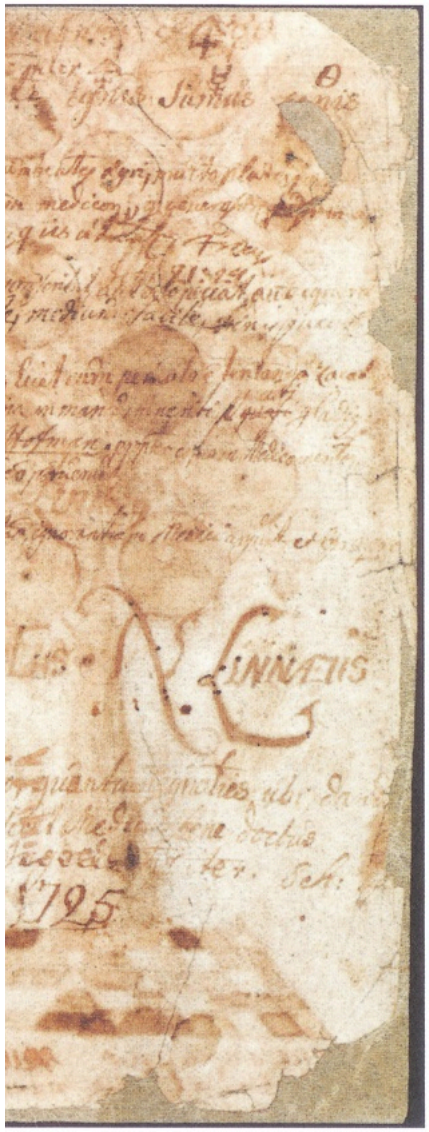
Own Your Data

Power of plain text for automating reproducible research

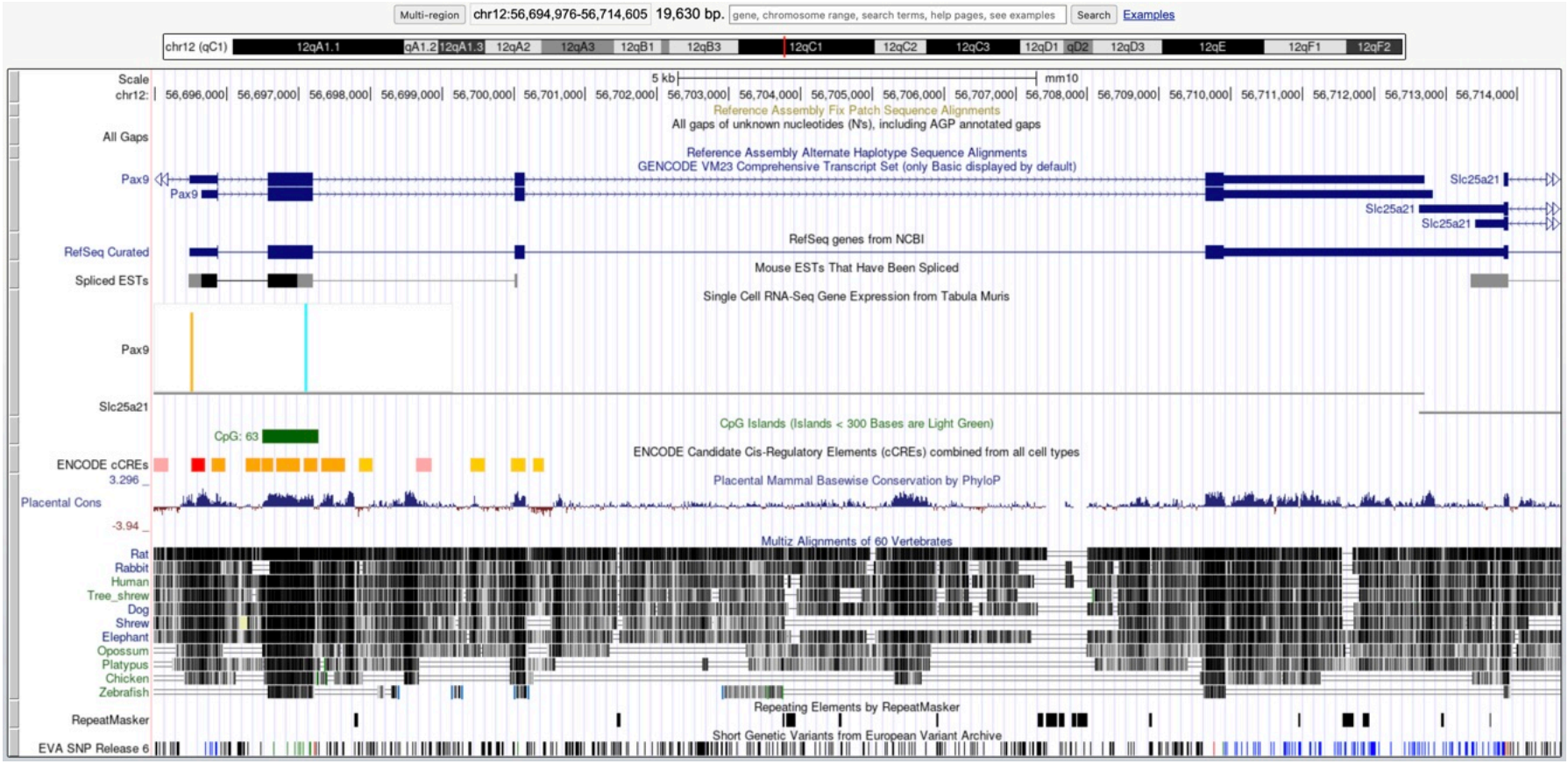
Leon Han
Streets Lab, Computational Biology Graduate Group
May 1st, 2025

Örtaboken (Herb book), an early Linnaeus manuscript, 1725

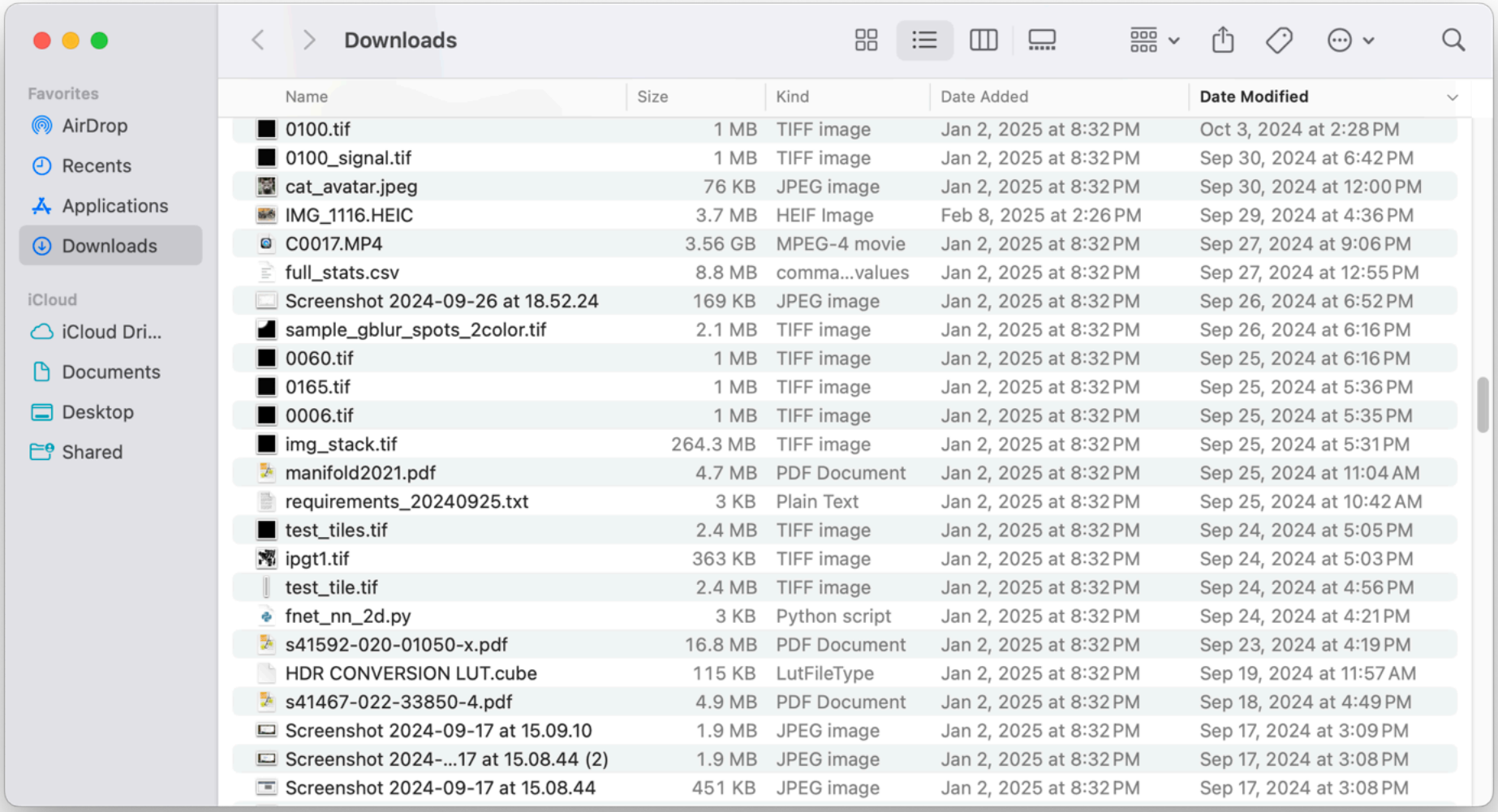
5268×3717, 7.6MB

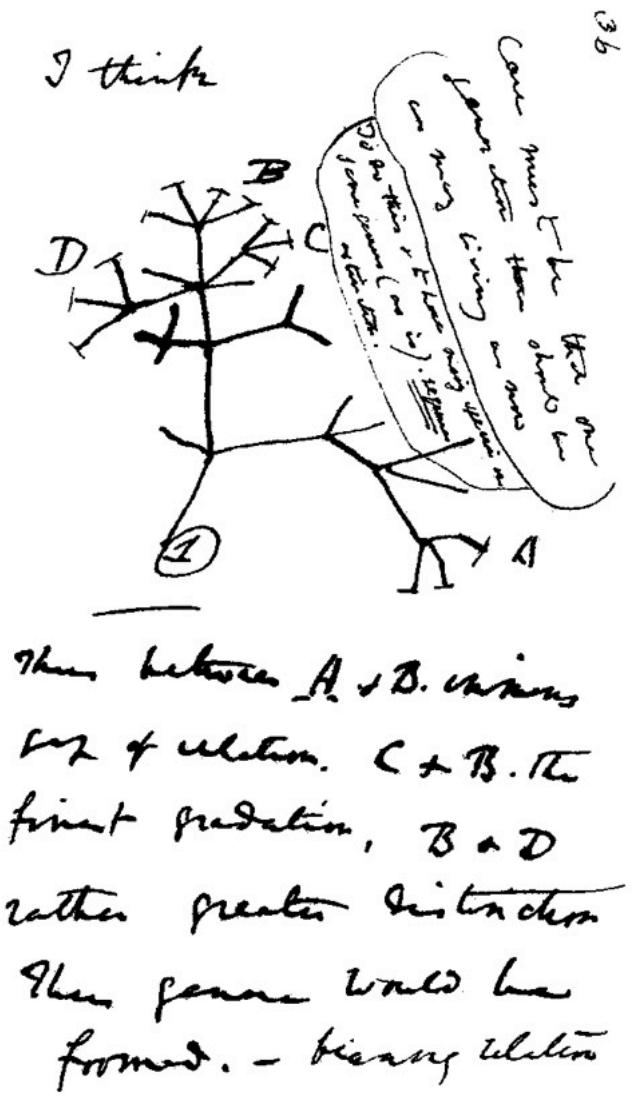


A World Flooded with Data



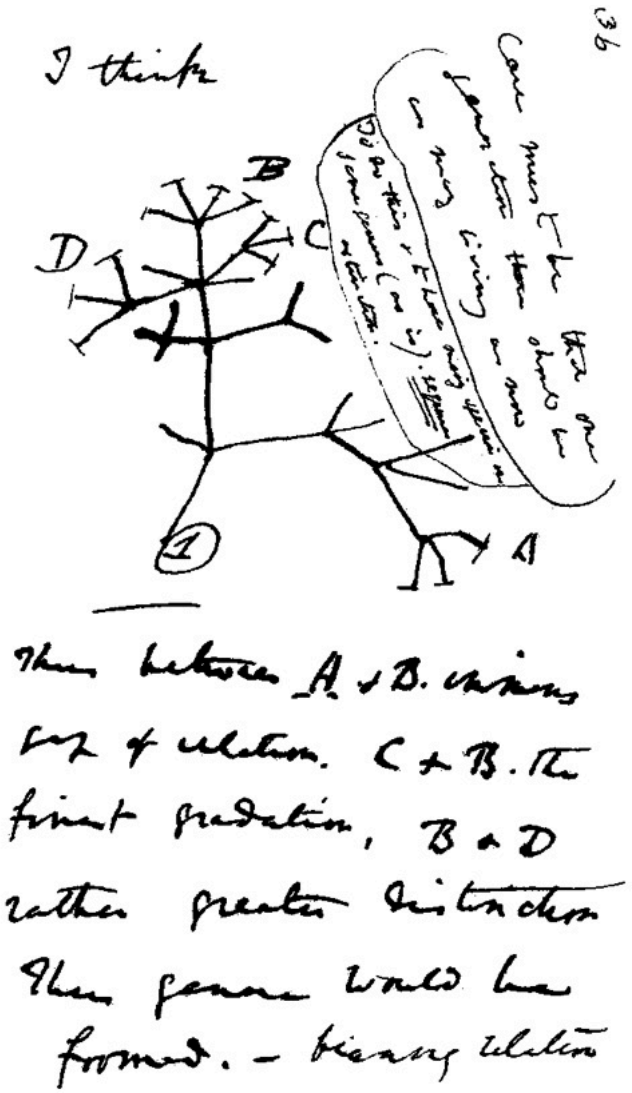
UCSC Genome Browser, genome of *Mus musculus*, 6.3GB, 2.7 billion base pairs


Downloads
Favorites
AirDrop
Recents
Applications
Downloads
iCloud
iCloud Dri...
Documents
Desktop
Shared
Name
Size
Kind
Date Added
Date Modified
0100.tif 1 MB TIFF image Jan 2, 2025 at 8:32 PM Oct 3, 2024 at 2:28 PM
0100_signal.tif 1 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 30, 2024 at 6:42 PM
cat_avatar.jpeg 76 KB JPEG image Jan 2, 2025 at 8:32 PM Sep 30, 2024 at 12:00 PM
IMG_1116.HEIC 3.7 MB HEIF Image Feb 8, 2025 at 2:26 PM Sep 29, 2024 at 4:36 PM
C0017.MP4 3.56 GB MPEG-4 movie Jan 2, 2025 at 8:32 PM Sep 27, 2024 at 9:06 PM
full_stats.csv 8.8 MB comma...values Jan 2, 2025 at 8:32 PM Sep 27, 2024 at 12:55 PM
Screenshot 2024-09-26 at 18.52.24 169 KB JPEG image Jan 2, 2025 at 8:32 PM Sep 26, 2024 at 6:52 PM
sample_gblur_spots_2color.tif 2.1 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 26, 2024 at 6:16 PM
0060.tif 1 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 25, 2024 at 6:16 PM
0165.tif 1 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 25, 2024 at 5:36 PM
0006.tif 1 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 25, 2024 at 5:35 PM
img_stack.tif 264.3 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 25, 2024 at 5:31 PM
manifold2021.pdf 4.7 MB PDF Document Jan 2, 2025 at 8:32 PM Sep 25, 2024 at 11:04 AM
requirements_20240925.txt 3 KB Plain Text Jan 2, 2025 at 8:32 PM Sep 25, 2024 at 10:42 AM
test_tiles.tif 2.4 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 24, 2024 at 5:05 PM
ipgt1.tif 363 KB TIFF image Jan 2, 2025 at 8:32 PM Sep 24, 2024 at 5:03 PM
test_tile.tif 2.4 MB TIFF image Jan 2, 2025 at 8:32 PM Sep 24, 2024 at 4:56 PM
fnet_nn_2d.py 3 KB Python script Jan 2, 2025 at 8:32 PM Sep 24, 2024 at 4:21 PM
s41592-020-01050-x.pdf 16.8 MB PDF Document Jan 2, 2025 at 8:32 PM Sep 23, 2024 at 4:19 PM
HDR CONVERSION LUT.cube 115 KB LutFileType Jan 2, 2025 at 8:32 PM Sep 19, 2024 at 11:57 AM
s41467-022-33850-4.pdf 4.9 MB PDF Document Jan 2, 2025 at 8:32 PM Sep 18, 2024 at 4:49 PM
Screenshot 2024-09-17 at 15.09.10 1.9 MB JPEG image Jan 2, 2025 at 8:32 PM Sep 17, 2024 at 3:09 PM
Screenshot 2024-...17 at 15.08.44 (2) 1.9 MB JPEG image Jan 2, 2025 at 8:32 PM Sep 17, 2024 at 3:08 PM
Screenshot 2024-09-17 at 15.08.44 451 KB JPEG image Jan 2, 2025 at 8:32 PM Sep 17, 2024 at 3:08 PM



Not All Data Speaks the Same Language

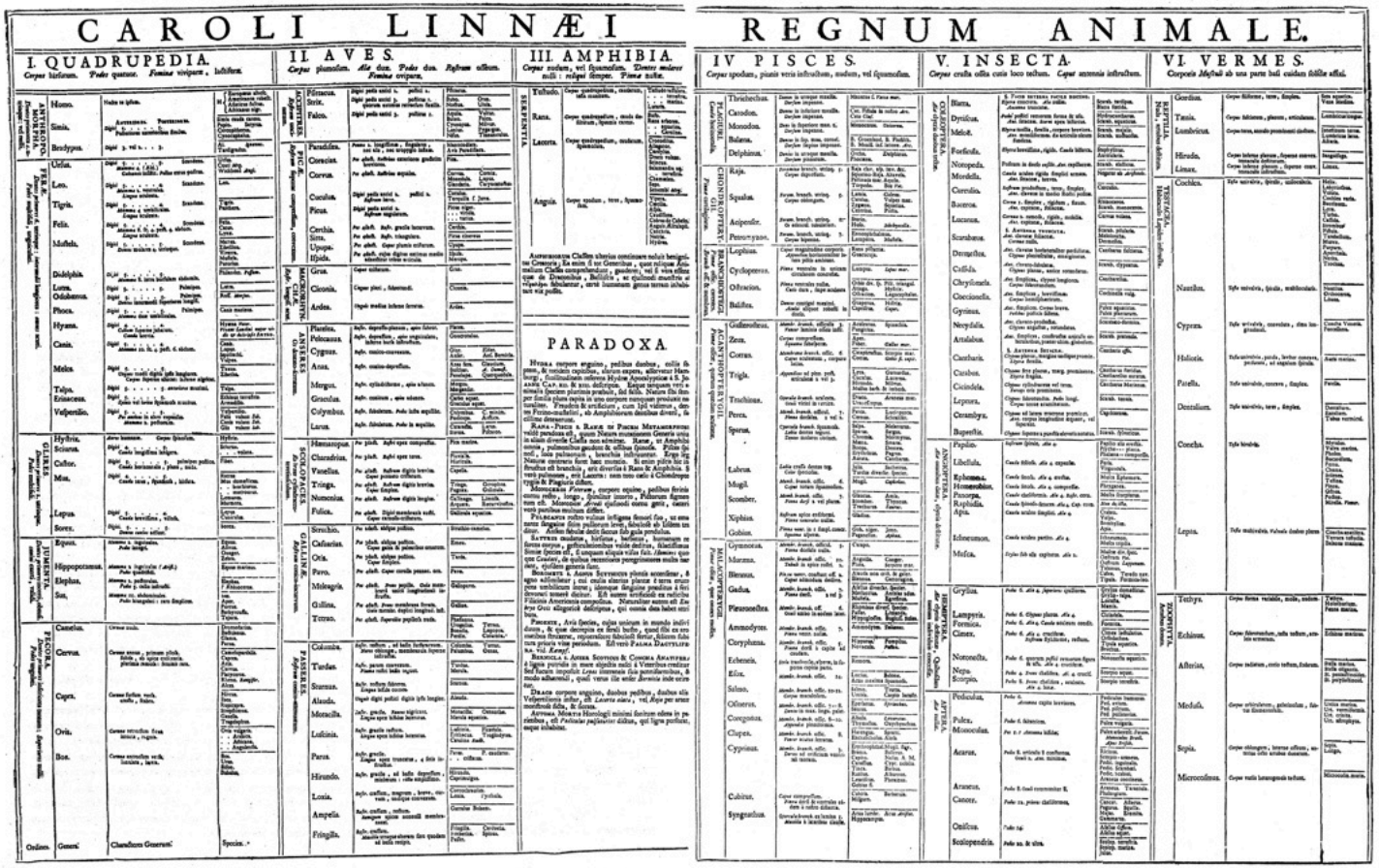
Tree of Life, Charles Darwin, 1837
695×1181, 15KB



Not All Data Speaks the Same Language

> I think ... Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.
>
> Thus between A & B ... Thus genera would be formed.

On the Origin of Species has approximately 893,878 characters, or 150,144 words.



Regnum Animale, Linnaeus, 1735

Not All Data Speaks the Same Language

- **Directly interpretable**
  - Documents, slides, spreadsheets, figures
- Machine-oriented
  - cryoEM images, viral genome, RNA-seq, neural networks, ...

Not All Data Speaks the Same Language

- Directly interpretable
  - Documents, slides, spreadsheets, figures
- **Machine-oriented**
  - cryoEM images, viral genome, RNA-seq, neural networks, ...

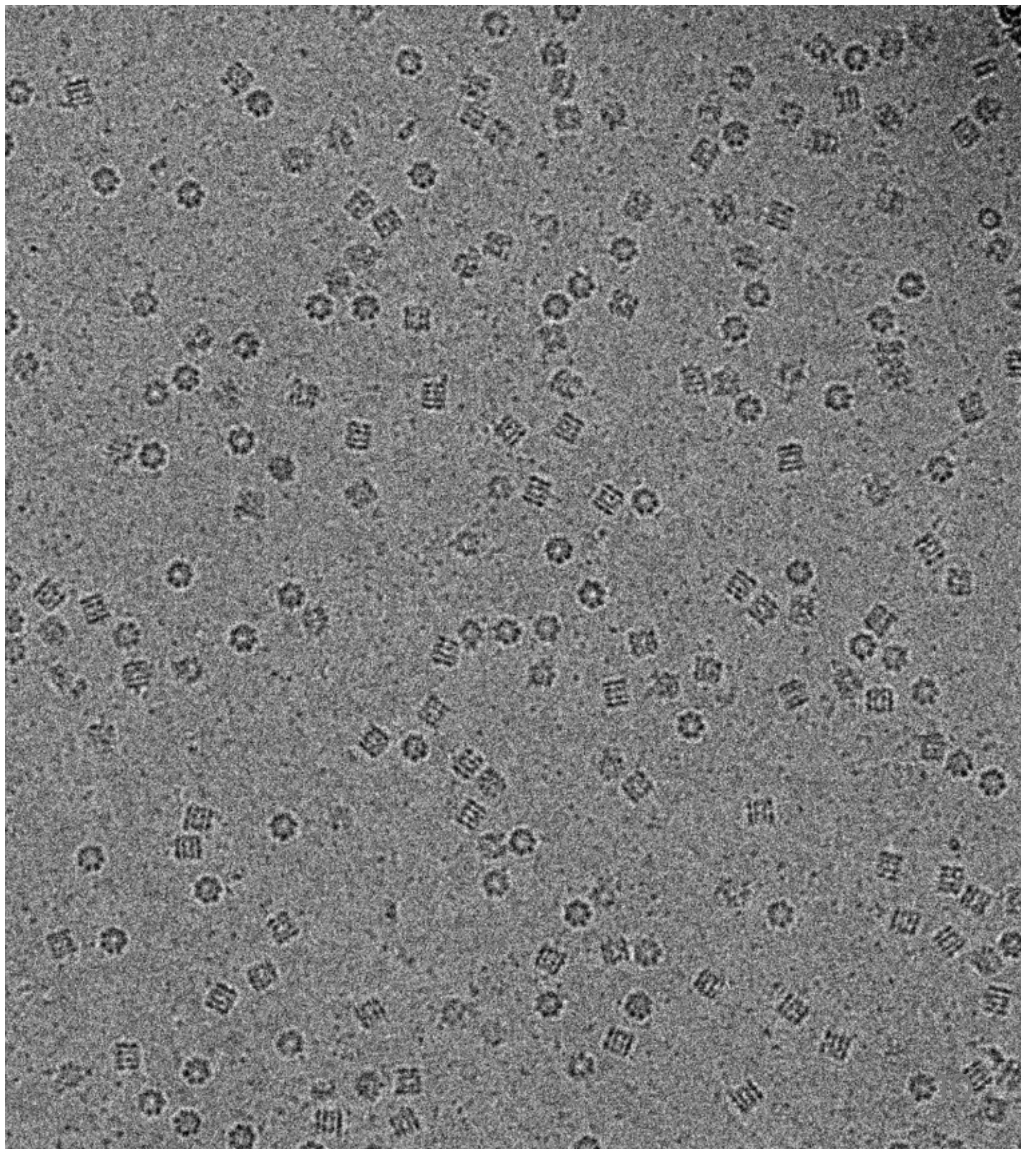


Image By Vossman CC BY-SA 4.0,

What Makes Data Readable?

Spatial Information

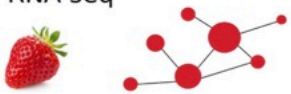
Microscopy v. Transcriptomics

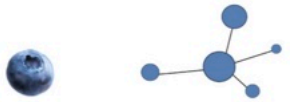
https://www.thepioneerwoman.com/food-cooking/recipes/a39752886/fresh-fruit-tart-recipe/

organ/ tissue

bulk RNA seq

single-cell RNA seq

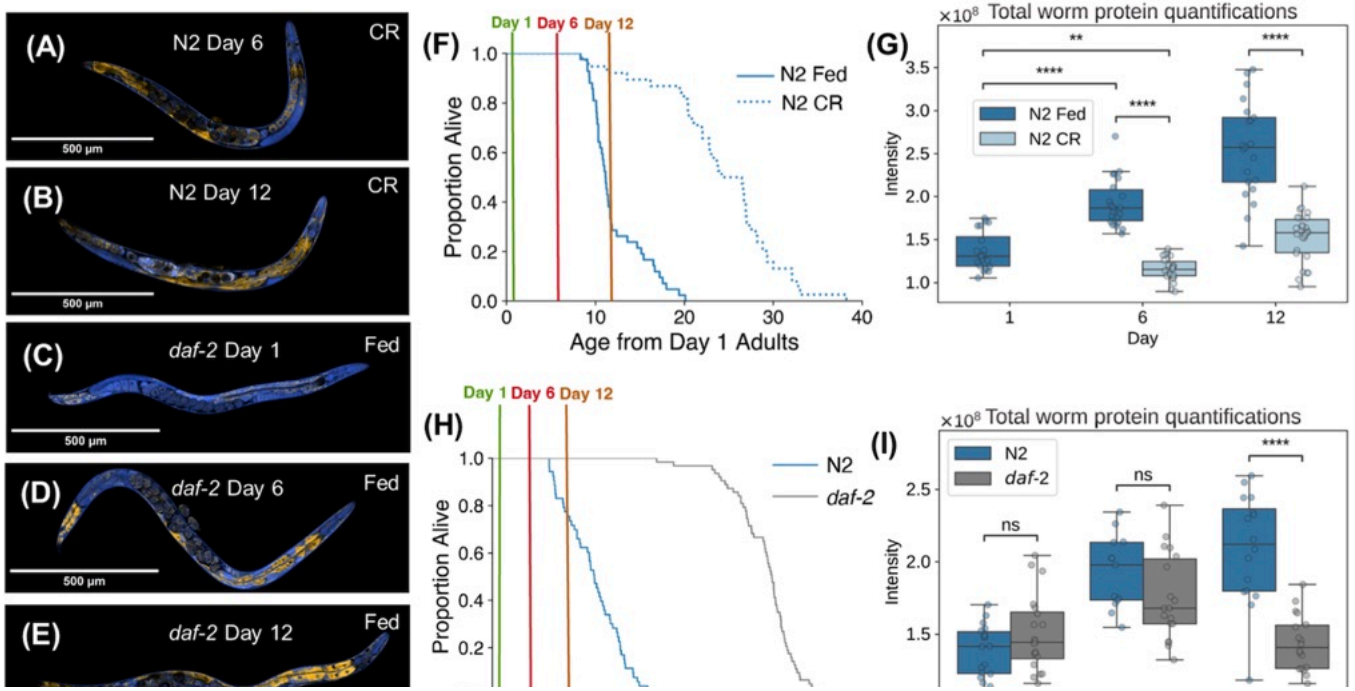


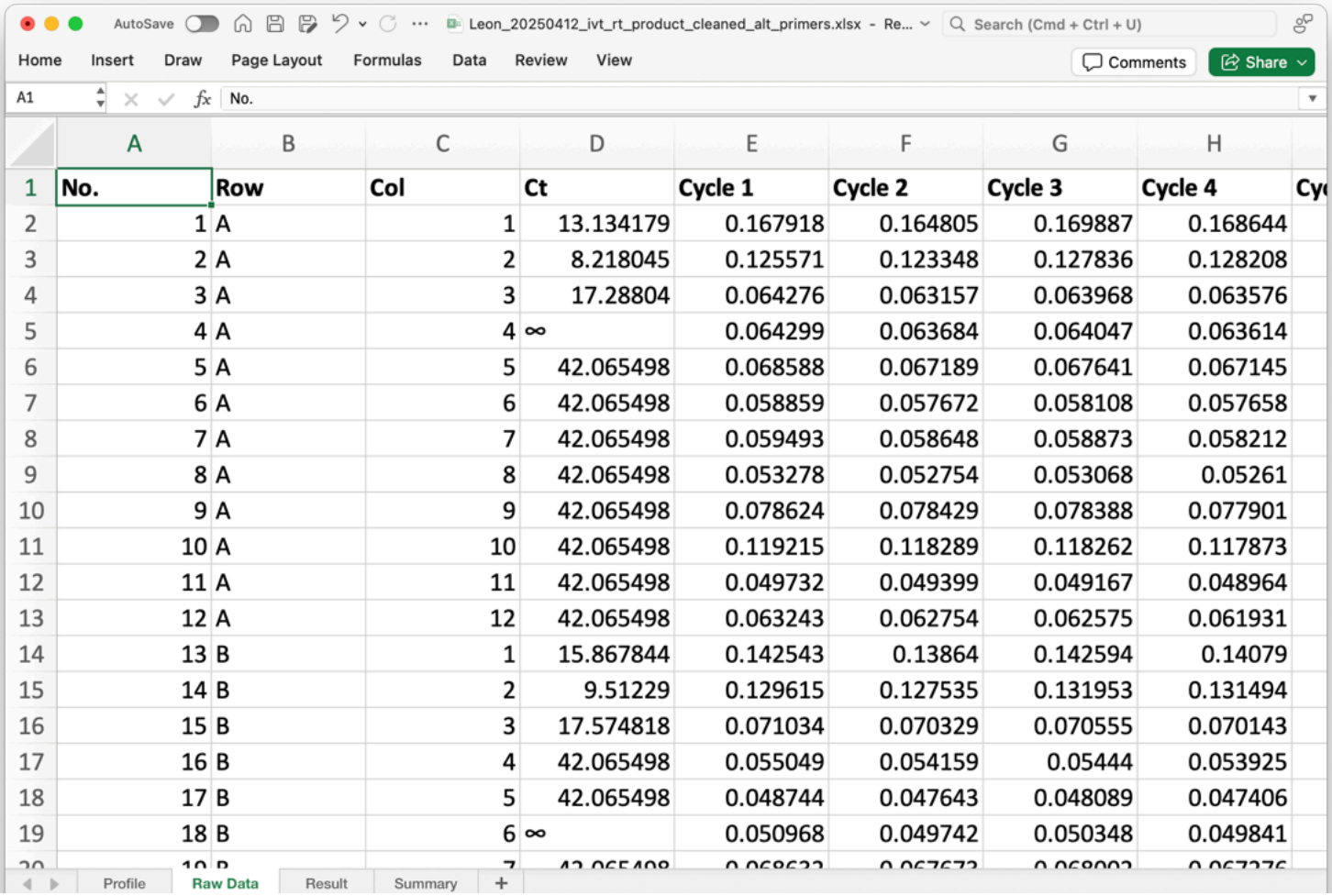


Steinheuer, Lisa Maria, Sebastian Canzler, and Jörg Hackermüller. "Benchmarking ScRNA-Seq Imputation Tools with Respect to Network Inference Highlights Deficits in Performance at High Levels of Sparsity." Preprint. (2021)

5

longer compared to their wildtype counterparts (Figure 3H). Then, using these same conditions, we performed SRS imaging at day 1, day 6, and day 12 of adulthood. Consistent with previous studies, *daf-2* mutant worms displayed increased lipid levels in their intestine by day 6 of adulthood (Figures 2B, 3D, and S1)[21,36,39–42]. Importantly, similar to CR, *daf-2* worms showed significantly lower levels of protein buildup by day 12 of adulthood compared to their N2 counterparts (Figure 3D, 3E, and 3I).



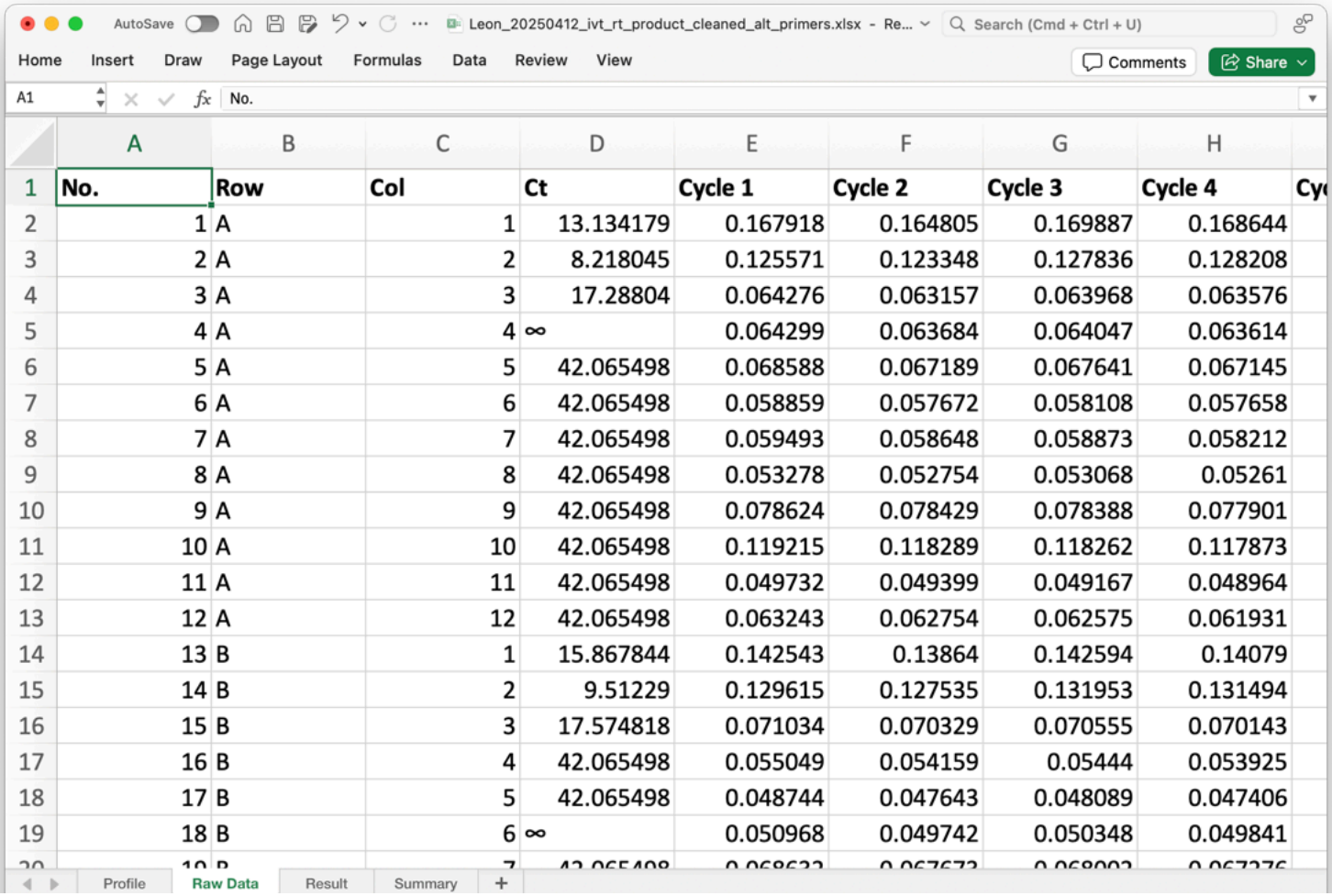


	A	B	C	D	E	F	G	H
1	No.	Row	Col	Ct	Cycle 1	Cycle 2	Cycle 3	Cycle 4
2	1	A	1	13.134179	0.167918	0.164805	0.169887	0.168644
3	2	A	2	8.218045	0.125571	0.123348	0.127836	0.128208
4	3	A	3	17.28804	0.064276	0.063157	0.063968	0.063576
5	4	A	4	∞	0.064299	0.063684	0.064047	0.063614
6	5	A	5	42.065498	0.068588	0.067189	0.067641	0.067145
7	6	A	6	42.065498	0.058859	0.057672	0.058108	0.057658
8	7	A	7	42.065498	0.059493	0.058648	0.058873	0.058212
9	8	A	8	42.065498	0.053278	0.052754	0.053068	0.05261
10	9	A	9	42.065498	0.078624	0.078429	0.078388	0.077901
11	10	A	10	42.065498	0.119215	0.118289	0.118262	0.117873
12	11	A	11	42.065498	0.049732	0.049399	0.049167	0.048964
13	12	A	12	42.065498	0.063243	0.062754	0.062575	0.061931
14	13	B	1	15.867844	0.142543	0.13864	0.142594	0.14079
15	14	B	2	9.51229	0.129615	0.127535	0.131953	0.131494
16	15	B	3	17.574818	0.071034	0.070329	0.070555	0.070143
17	16	B	4	42.065498	0.055049	0.054159	0.05444	0.053925
18	17	B	5	42.065498	0.048744	0.047643	0.048089	0.047406
19	18	B	6	∞	0.050968	0.049742	0.050348	0.049841

Text: The Universal Interface

- **Document**: text, images, links, tables, equations
- **Slides**: text, images, videos, tables, equations
- **Figures**: text + images (a specialized “document”)
- **Spreadsheet**: text + numbers

Behind the Scenes: It's Still Text!



	A	B	C	D	E	F	G	H
1	No.	Row	Col	Ct	Cycle 1	Cycle 2	Cycle 3	Cycle 4
2	1	A	1	13.134179	0.167918	0.164805	0.169887	0.168644
3	2	A	2	8.218045	0.125571	0.123348	0.127836	0.128208
4	3	A	3	17.28804	0.064276	0.063157	0.063968	0.063576
5	4	A	4	∞	0.064299	0.063684	0.064047	0.063614
6	5	A	5	42.065498	0.068588	0.067189	0.067641	0.067145
7	6	A	6	42.065498	0.058859	0.057672	0.058108	0.057658
8	7	A	7	42.065498	0.059493	0.058648	0.058873	0.058212
9	8	A	8	42.065498	0.053278	0.052754	0.053068	0.05261
10	9	A	9	42.065498	0.078624	0.078429	0.078388	0.077901
11	10	A	10	42.065498	0.119215	0.118289	0.118262	0.117873
12	11	A	11	42.065498	0.049732	0.049399	0.049167	0.048964
13	12	A	12	42.065498	0.063243	0.062754	0.062575	0.061931
14	13	B	1	15.867844	0.142543	0.13864	0.142594	0.14079
15	14	B	2	9.51229	0.129615	0.127535	0.131953	0.131494
16	15	B	3	17.574818	0.071034	0.070329	0.070555	0.070143
17	16	B	4	42.065498	0.055049	0.054159	0.05444	0.053925
18	17	B	5	42.065498	0.048744	0.047643	0.048089	0.047406
19	18	B	6	∞	0.050968	0.049742	0.050348	0.049841

```
<?xml version="1.0" ?>
<ns0:worksheet xmlns:ns0="[http://schemas.openxmlformats.org/spreadsheetml/2006/main]
(http://schemas.openxmlformats.org/spreadsheetml/2006/main)">
<ns0:sheetData> <ns0:row r="1">
<ns0:c r="A1" s="1" t="str">
<ns0:v>No.</ns0:v> </ns0:c>
<ns0:c r="B1" s="1" t="str">
<ns0:v>Row</ns0:v> </ns0:c>
<ns0:c r="C1" s="1" t="str">
<ns0:v>Col</ns0:v> </ns0:c>
<ns0:c r="D1" s="1" t="str">
<ns0:v>Ct</ns0:v> </ns0:c>
<ns0:c r="E1" s="1" t="str">
<ns0:v>Cycle 1</ns0:v> </ns0:c>
<ns0:c r="F1" s="1" t="str">
<ns0:v>Cycle 2</ns0:v> </ns0:c>
<ns0:c r="G1" s="1" t="str">
<ns0:v>Cycle 3</ns0:v> </ns0:c>
...
</ns0:row> </ns0:sheetData> </ns0:worksheet>
```


Cal

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 1.1//EN"
"http://www.w3.org/Graphics/SVG/1.1/DTD/svg11.dtd">
<svg width="100%" height="100%" viewBox="0 0 365 292"
version="1.1" xmlns="http://www.w3.org/2000/svg"
xmlns:xlink="http://www.w3.org/1999/xlink"
xml:space="preserve" xmlns:serif="http://www.serif.com/"
style="fill-rule:evenodd;clip-rule:evenodd;
stroke-linejoin:round;stroke-miterlimit:2;">
    <g>
        <path d="M345.763,55.486C343.428,54.198 337.196,
        58.793 337.196,58.793C325.262,68.27 316.292,80.
        011 308.562,93.539C288.138,129.28 290.022,157.364
        290.022,157.364C290.022,157.364 300.676,145.235
        303.19,141.262C306.617,135.845 319.512,117.712
        326.767,104.928C334.546,91.221 343.485,72.249 345.
        701,65.679C345.701,65.679 349.289,57.431 345.763,
        55.486ZM195.013,140.13C189.087,140.312 174.954,
        147.607 163.653,168.51C161.016,173.388 152.84,193.
        678 157.27,203.806C158.608,206.868 161.029,208.
        968 164.086,209.714C164.876,209.909 165.691,210.
        006 166.503,210.006C176.472,210.008 183.454,196.06
```

$$M = \begin{bmatrix} m_{11} & \cdots & m_{1g} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{ng} \end{bmatrix}, \ \Lambda = \underbrace{\left.\begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1g} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{ng} \end{bmatrix}\right\} n,}_{g}$$

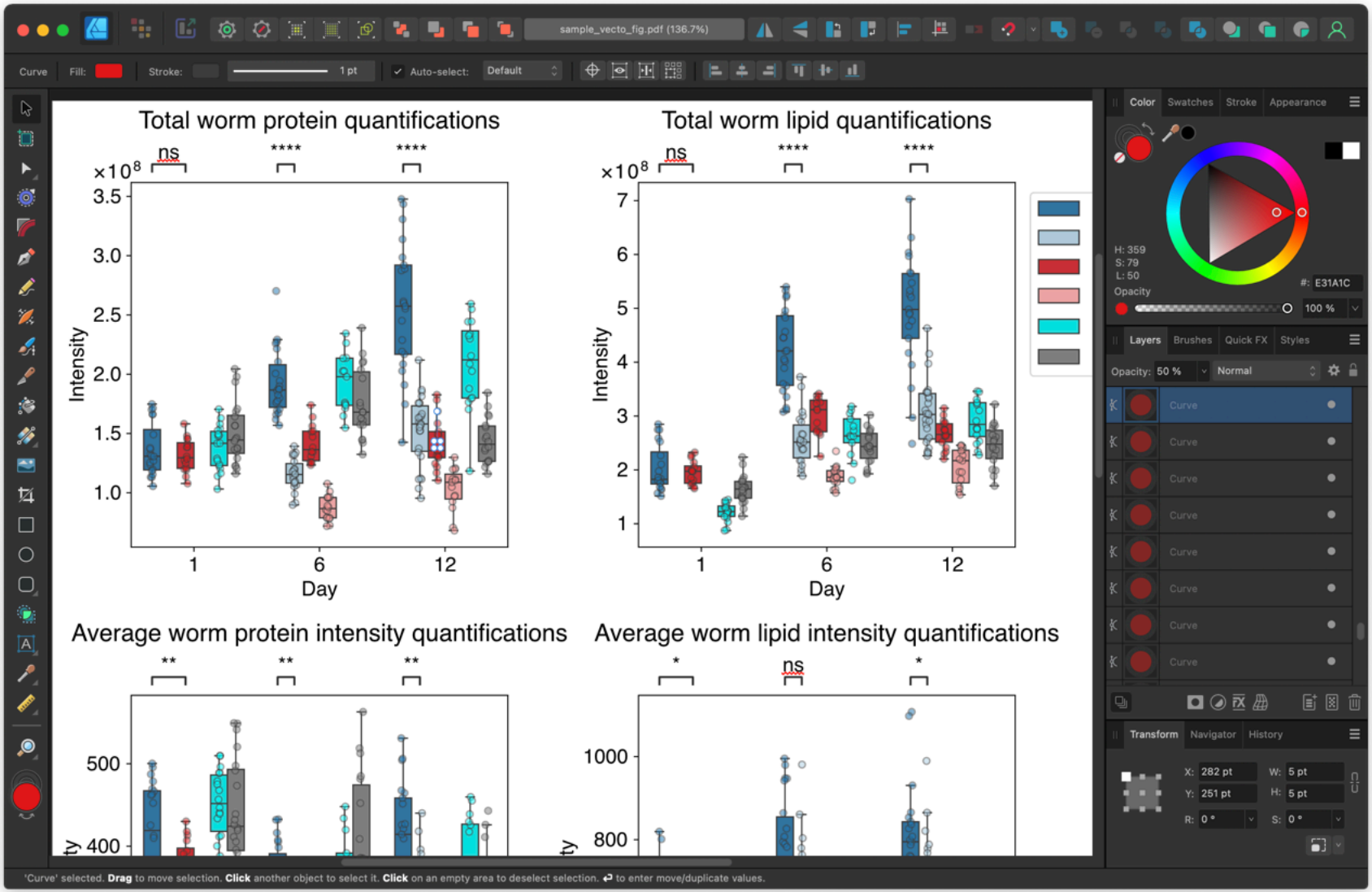
```
$$
M =  \begin{bmatrix}
m_{11} &  \cdots &  m_{1g} \\
\vdots & \ddots  &  \vdots \\
m_{n1}  &  \cdots  & m_{ng}
\end{bmatrix},\ \ \Lambda =\underbrace{\begin{bmatrix}
...
\end{bmatrix}}_{\displaystyle g}\left.\vphantom{\begin{bmatrix}
...
\end{bmatrix}}\right\}n,
$$
```

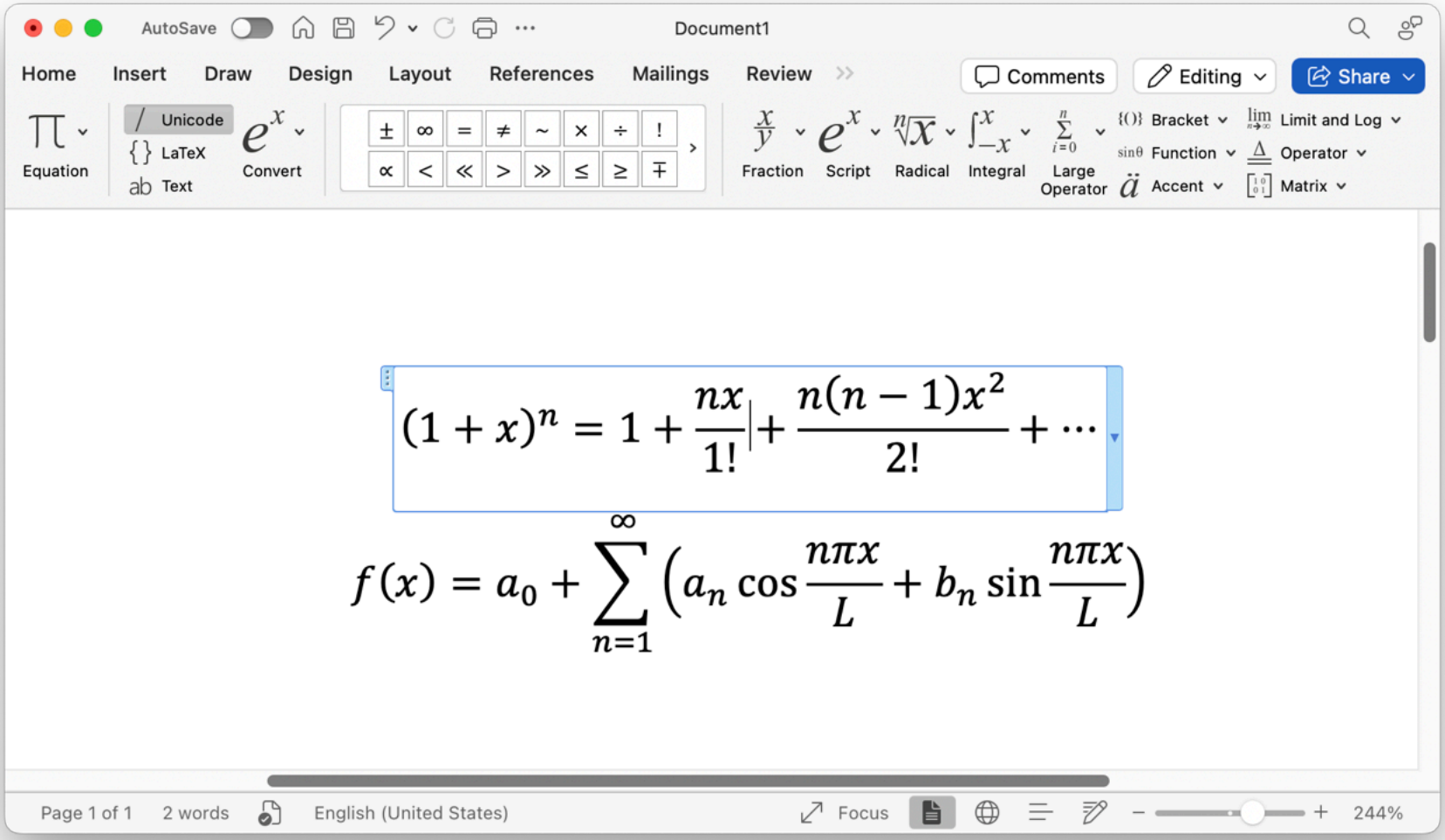
Scientific Research is Both Dynamic and Scalable

Small Tasks, Big Headaches

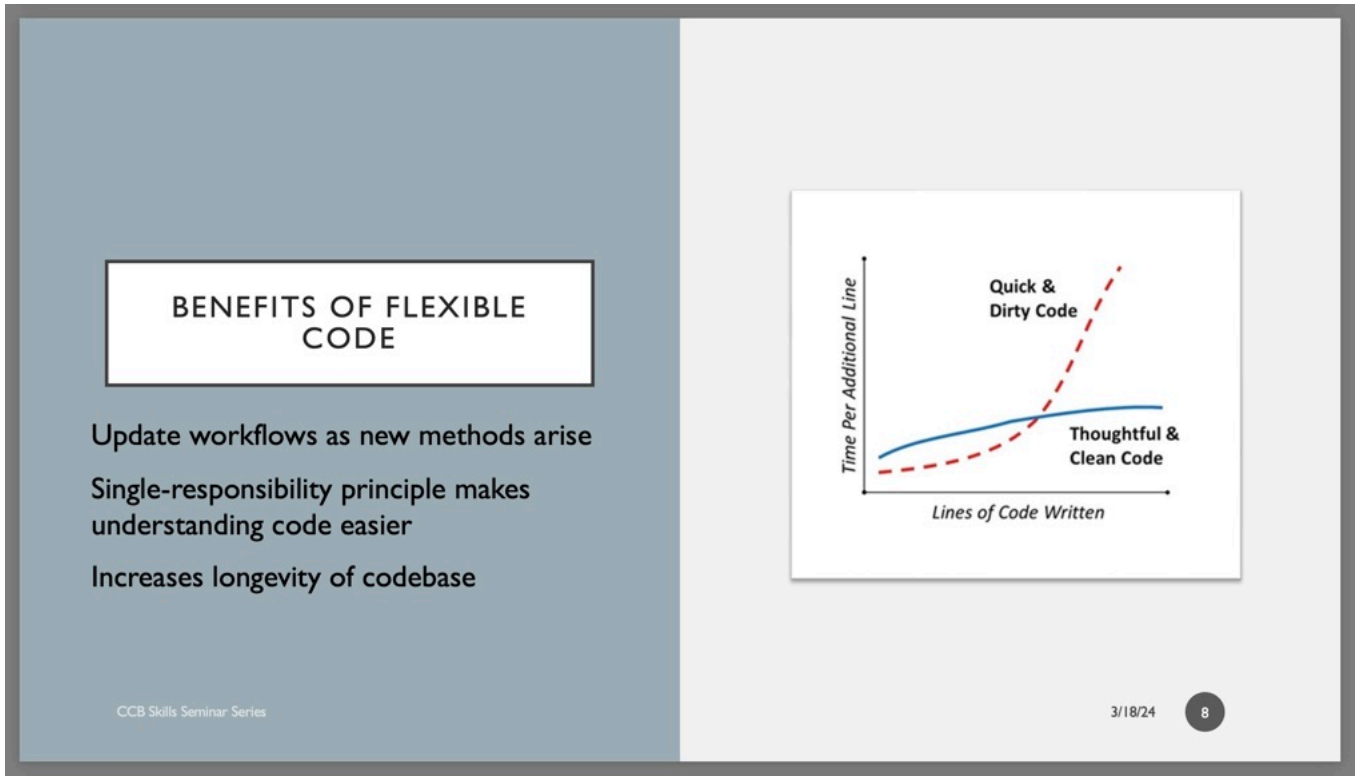
> Dilute the 100 µM stock to a concentration of **80 ng/µL** for each oligo. Label as probe, T7.

Oligo	Stock Conc.	Final Conc.	Volume of Stock	Volume of Water	Final Volume
T7 promoter	100.00 µM	12.89 µM	2.58 µL	17.42 µL	20.00 µL


sample_vecto_fig.pdf (136.7%)
Total worm protein quantifications
Total worm lipid quantifications
Average worm protein intensity quantifications
Average worm lipid intensity quantifications
Intensity
Day
ns
****
**
*
Color
Swatches
Stroke
Appearance
Layers
Brushes
Quick FX
Styles
Transform
Navigator
History
'Curve' selected. Drag to move selection. Click another object to select it. Click on an empty area to deselect selection. ↵ to enter move/duplicate values.


AutoSave
Document1
Home
Insert
Draw
Design
Layout
References
Mailings
Review
Comments
Editing
Share
Equation
Unicode
LaTeX
Text
Convert
Fraction
Script
Radical
Integral
Large Operator
Bracket
Function
Accent
Limit and Log
Operator
Matrix
$(1+x)^n = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots$
$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$
Page 1 of 1
2 words
English (United States)
Focus
244%

When Effort Scales Poorly

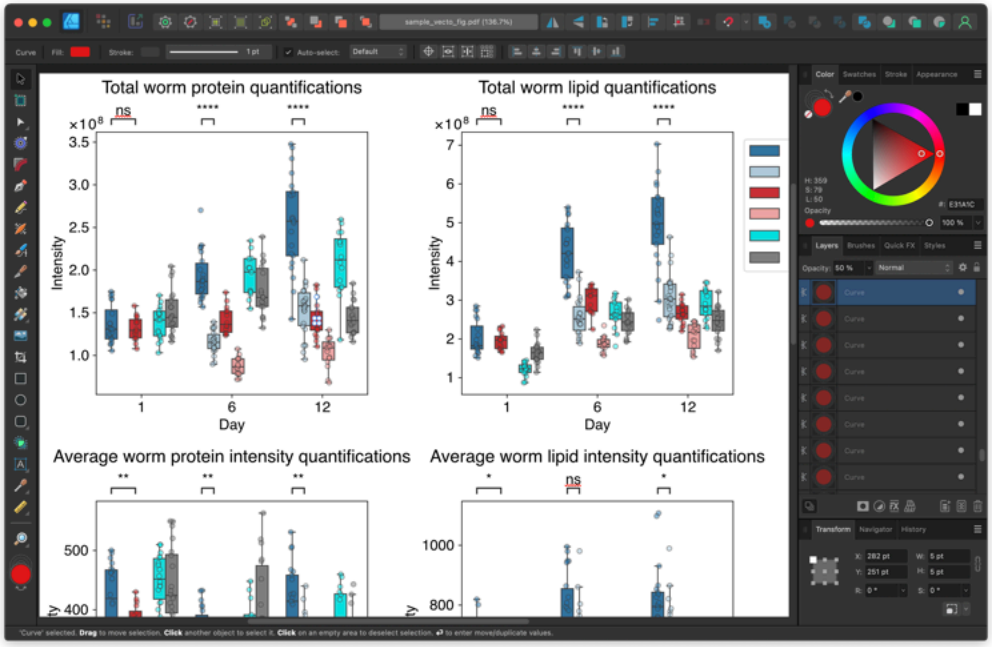


Credit: Matthew Giammar, CCB

Text as a Tool for Automation and Reuse

```
def calculate_mw(sequence, is_duplex):
        ...
def convert_conc(conc, unit, mw):
        ...
def dilution_ratio(start_conc, target_conc):
        ...
def determine_dilution_steps(dil_ratio):
    steps = []
    while dil_ratio > 10:
        if dil_ratio >= 10:
            steps.append(10)
            dil_ratio /= 10
        elif dil_ratio >=5:
            steps.append(5)
            dil_ratio /=5
    steps.append(dil_ratio)
    return steps
```

Step	Oligo	Stock Conc.	Final Conc.	Volume of Stock	Volume of Water	Final Volume
1	T7 promoter	100.00 µM	12.89 µM	2.58 µL	17.42 µL	20.00 µL
1	Probe	100.00 µM	10.00 µM	1.00 µL	9.00 µL	10.00 µL
2	Probe	10.00 µM	1.91 µM	3.81 µL	16.19 µL	20.00 µL
1	Probe+T7	100.00 µM	10.00 µM	1.00 µL each	8.00 µL	10.00 µL
2	Probe+T7	10.00 µM	1.66 µM	8.31 µL	41.69 µL	50.00 µL
...						




```
hue_order = [
    'N2_Fed', 'N2_CR',
    'CEH60_Fed', 'CEH60_CR',
    'N2_(daf2_control)_Fed',
    'daf2_Fed',
]

my_colors = [
    '#1f78b4',
    '#a6cee3',
    '#e31a1c',
    '#fb9a99',
    '#00FFFF',
    '#808080',
]

hue_dict = dict(zip(
    hue_order,
    my_colors
))
```

```
\begin{frame}
\frametitle{From Discrete to Continuous}
\[N_k=N_0\lambda^{k}, k\in\mathbb{N}\]
\pause
\begin{description}
\item[Comparison]The geometric sequence
\pause
\item[Expansion]{What if the cells do not divide simultaneously?
What about a non-integer time (e.g. $0.7\tau$)?}
\end{description}
\pause
\begin{align*}
&N_t=N_0\lambda^t\\
&t=C\tau, C\in\mathbb{R}
\end{align*}
\end{frame}
```

From Discrete to Continuous

$$N_k = N_0\lambda^k, k \in \mathbb{N}$$

Comparison The geometric sequence

Expansion What if the cells do not divide simultaneously? What about a non-integer time (e.g. 0.7τ)?

$$N_t = N_0\lambda^t$$
$$t = C\tau, C \in \mathbb{R}$$

Key Takeaways

1. Text is the foundation of scientific data and communication
2. A lot of digital content is text under the hood
3. Using text early enables flexibility, scalability, and reproducibility

Structured Text for Machines and Humans

JSON:

- Created in early 2000s for JavaScript-based data exchange
- Now a **universal format** for web APIs, cloud apps, scientific tools
- Simple key–value pairs, supports arrays and nesting
- **Readable by both humans and machines**

```
{
  "name": "Taiyaki",
  "age": 1,
  "favoriteFood": "tuna"
}
```

Structured Text for Machines and Humans

YAML:

- Introduced in 2001 as **YAML Ain't Markup Language**
- Prioritizes **human readability** over strict syntax
- Widely used in configuration files (Docker, GitHub Actions, CI/CD tools)
- Indentation defines structure (no braces or commas)

```
name: taiyaki
age: 1
favoriteFood: tuna
```

Text Formatting Tools

HTML5 - the Heart of Modern Internet

> HTML5 (Hypertext Markup Language 5) is a markup language used for structuring and presenting hypertext documents on the World Wide Web. It was the fifth and final major HTML version that is now a retired World Wide Web Consortium (W3C) recommendation. The current specification is known as the HTML Living Standard.

— Wikipedia

Text Formatting Tools

HTML5 - the Heart of Modern Internet

Structured:

- Separates content and style

Easy Navigation:

- Support interconnected links and embedded contents

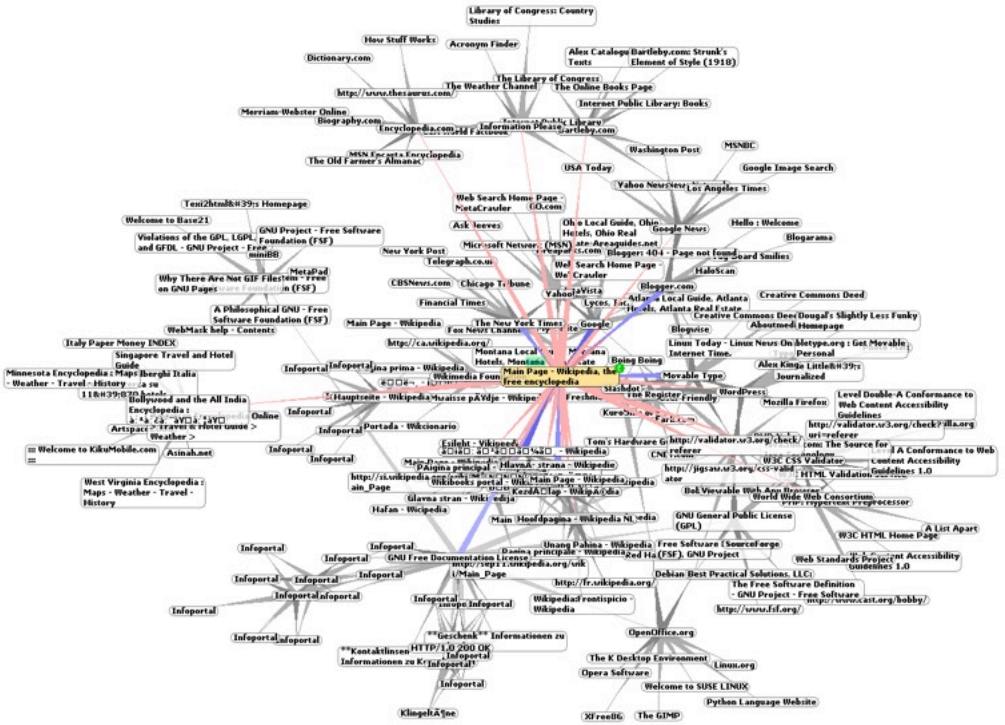
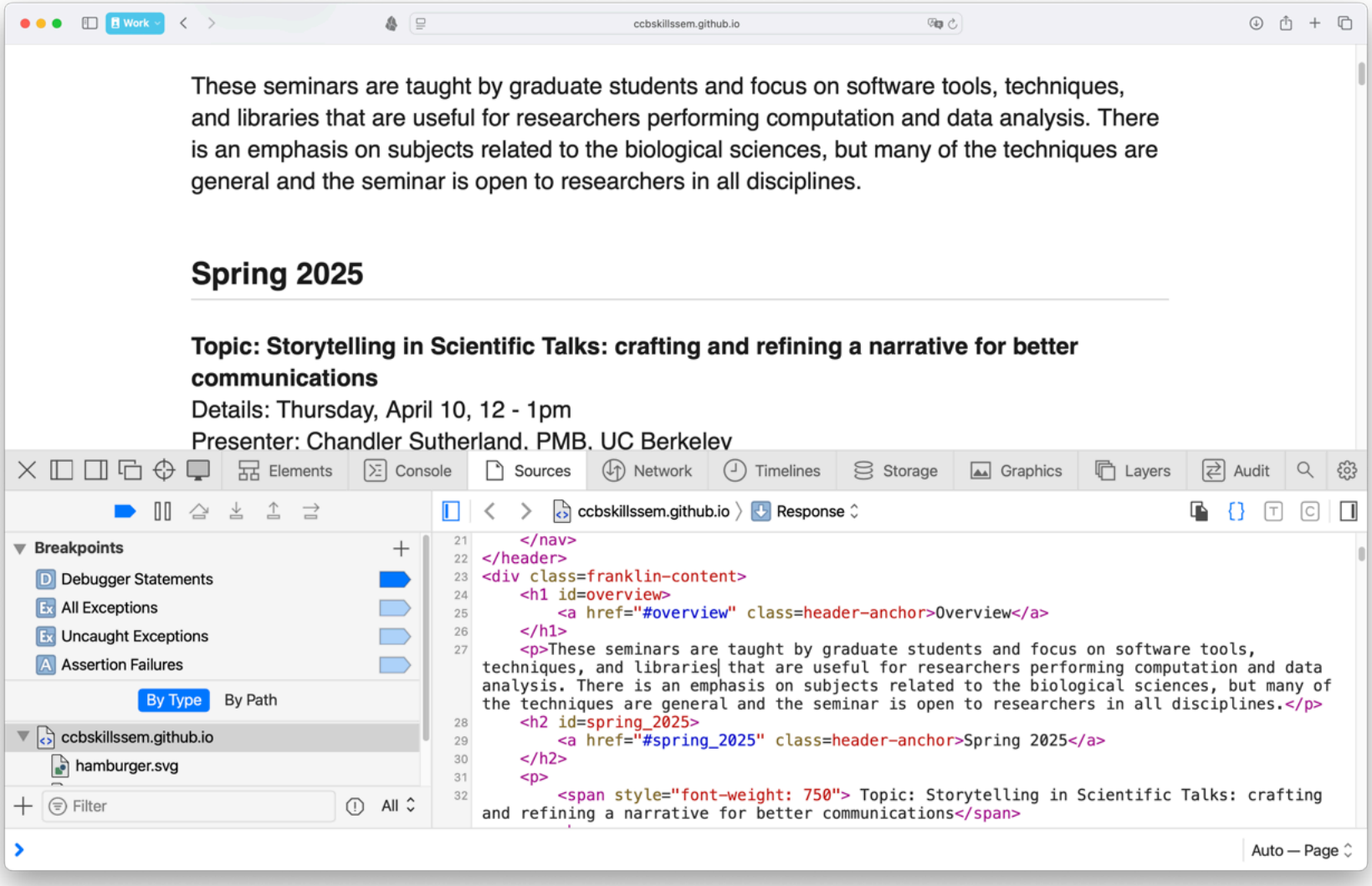


Image credit: Chris 73


Work
ccbskillssem.github.io
These seminars are taught by graduate students and focus on software tools, techniques, and libraries that are useful for researchers performing computation and data analysis. There is an emphasis on subjects related to the biological sciences, but many of the techniques are general and the seminar is open to researchers in all disciplines.
Spring 2025
Topic: Storytelling in Scientific Talks: crafting and refining a narrative for better communications
Details: Thursday, April 10, 12 - 1pm
Presenter: Chandler Sutherland, PMB, UC Berkelev
Elements
Console
Sources
Network
Timelines
Storage
Graphics
Layers
Audit
Breakpoints
Debugger Statements
All Exceptions
Uncaught Exceptions
Assertion Failures
By Type
By Path
ccbskillssem.github.io
hamburger.svg
Filter
All
ccbskillssem.github.io
Response
</nav>
</header>

<h1 id=overview>
Overview
</h1>
These seminars are taught by graduate students and focus on software tools, techniques, and libraries that are useful for researchers performing computation and data analysis. There is an emphasis on subjects related to the biological sciences, but many of the techniques are general and the seminar is open to researchers in all disciplines.
<h2 id=spring_2025>
Spring 2025
</h2>

 Topic: Storytelling in Scientific Talks: crafting and refining a narrative for better communications
Auto — Page

"Less is more."

Markdown Unifies Readability and Text Structuring

> Markdown is a lightweight markup language for creating formatted text using a plain-text editor. Markdown is widely used for blogging and instant messaging, online forums, collaborative software, documentation pages, and readme files.

Heading

Sub-heading

Alternative heading

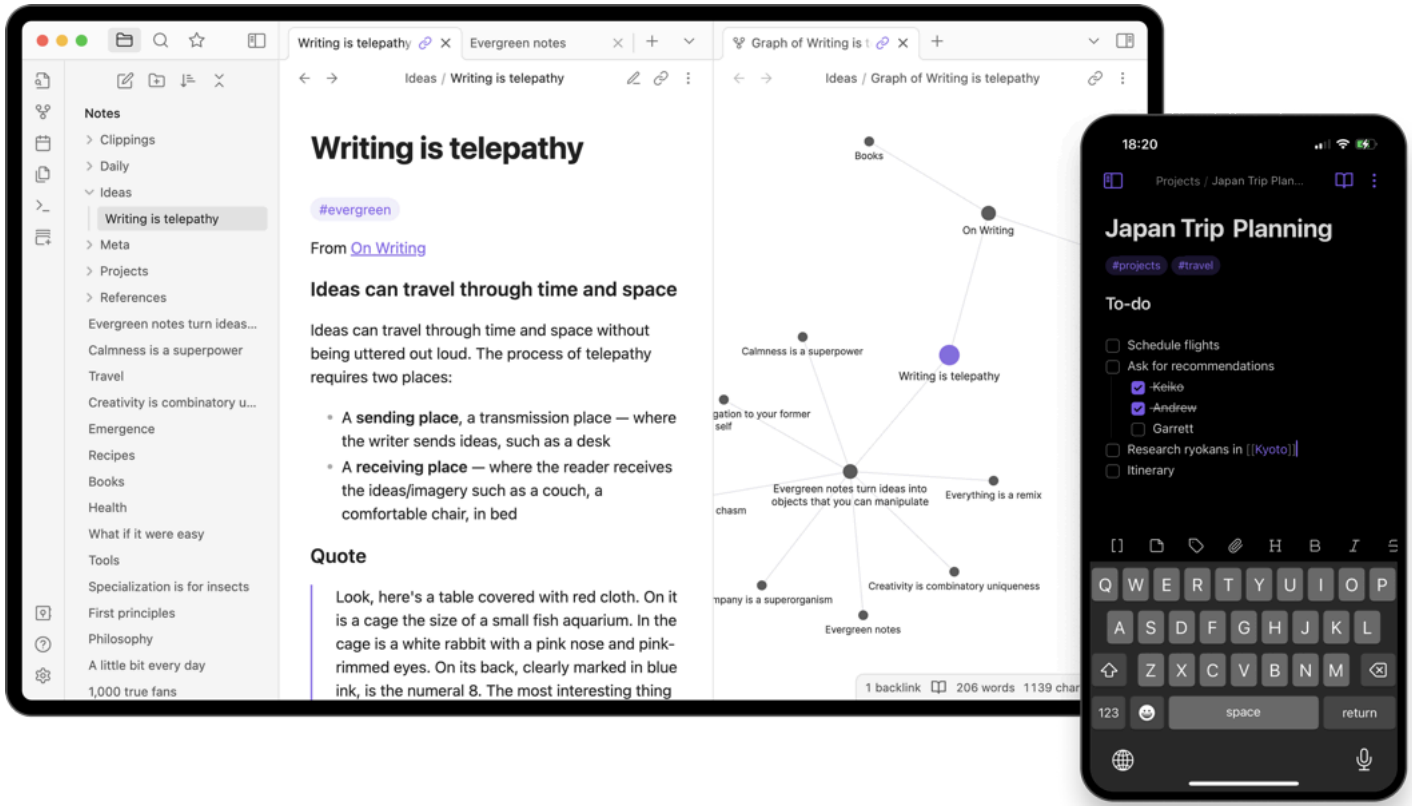
Alternative sub-heading

1. fruits
    - *apple*
    - **banana**
2. link to vegetables

> Rabbits like carrots.

```
---
Heading
=======

Sub-heading
-----------

# Alternative heading

## Alternative sub-heading

1. fruits
     * *apple*
     * **banana**
2. [link to vegetables](https://foo.bar)
> Rabbits like carrots.
```

Honorable Mention: Obsidian



Summary

1. You own the data as long as you know about it

2. Text is not only the abstraction of data, but also the cohesive of your knowledge base