[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

A UNICODE DATA STORY

PROCESS BOOK

CS 1710 / Visualization / Fall 2025

—

Chi Le
chile@college.harvard.edu

Rain YeYang
xyeyang@college.harvard.edu

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Team Agreement

1. Both team members will be involved with the technical aspects of the project, as well as brainstorming/layout design. All code will be documented well.
2. Final design decisions will be discussed among both members of the team, and will be agreed upon before proceeding. We will meet each other in the middle and find fair compromises when necessary. If any changes arise that may affect the other's workflow, that team member will notify their partner before moving forward with that option.
3. Work hours, and the 'weight' of delegated tasks, will be split as evenly as possible, though actual task output may differ based on varied background, experience, and previously learned skills. For this reason, we will aim to play to each other's strengths, while also ensuring fairness and a fruitful learning opportunity for both members.
4. We will both play an equal role in keeping the team accountable. If for whatever reason, a team member is not meeting expectations, both members will communicate transparently with each other and find a reasonable solution forward.
5. We will meet every week around Thursdays, in addition to in-class milestones on Mondays. We may work asynchronously, virtually, and/or side-by-side, depending on what we both agree is suitable for the week in question. Mondays are for brainstorming and critique, while Thursdays will be for ensuring we are on track for the upcoming deadlines.
6. We will communicate throughout the week as the project progresses and keep each other informed of ideas, developments, and other updates!

Signed,

Chi Le *Rain YeYang*

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Team Name

The Typos

[Original] Project Proposal

Abstract (TYPEFACE PARADIGMS THROUGH THE DIGITAL AGES)

Our project aims to explore the intersection of visual design and linguistics through typeface paradigms in the digital age (or multiple thereof, should we find appropriate delimiters). We will analyze historical trends and patterns in typeface usage across various applications and industries, including but not limited to high-level changes in brand logo fonts (and other marketing areas), popularity of typefaces in mass media, and creations of new fonts alongside relevances of old ones. Our ultimate goal is to visualize the many nuances of typography, language, and cultural impact/perception, the last of which we view is a cornerstone of visual design and communication. We would also like to explore theory-adjacent facets of our topic, namely the evolution of typeface development with respect to limitations posed by software and hardware (e.g., impacts of ASCII, UTF-8/16, Unicode, etc.) Time and/or scope permitting, we might like to orient our "digital age" analysis in a broader timeline (possibly tracing back to the inception of the printing press).

The project will integrate various forms of data: user-generated preferences across different eras/industries (e.g., advertising/marketing, entertainment, publishing, technology, etc.), the development of branding typefaces over time, popularity of typefaces by decade (mass media analytics, open source library data, case studies), and survey results. We envision our final website as a presentation of visualization snapshots/slides and information in roughly-chronological order. We want to tell an engaging (hi)story that combines static data visualizations (timelines, stacked bar charts, word-font mappings) with gamified and/or interactive elements (e.g., *Shape Type*-inspired mini-games, font quizzes).

Background and Motivation

Both of us are deeply interested in UI/UX and graphic design, which is what we bonded over in our initial conversation. (We were strangers prior to project formation, but we had a very animated discussion right off the bat about where we could take this project idea!) We are interested in the topic on a more abstract level as well, especially its implications on the study of aesthetics and human communication — fonts are not neutral; they influence trust, legibility, emotional tone, and cultural resonance. Typography sits at the intersection of design, linguistics, and cultural studies, but the 'visual design of letters' (typeface design) is

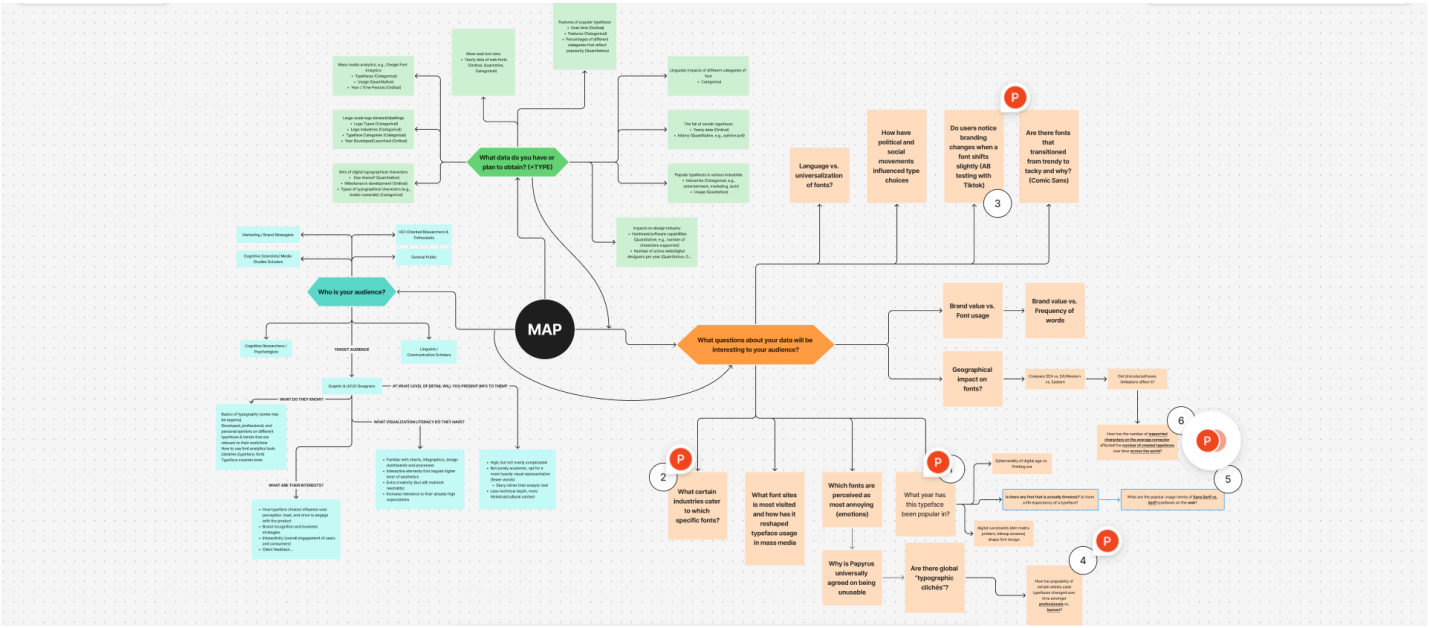
sometimes considered too subtle to be at the very forefront of design considerations. This project allows us to explore that layer which underpins design and make it visible through data.

Data

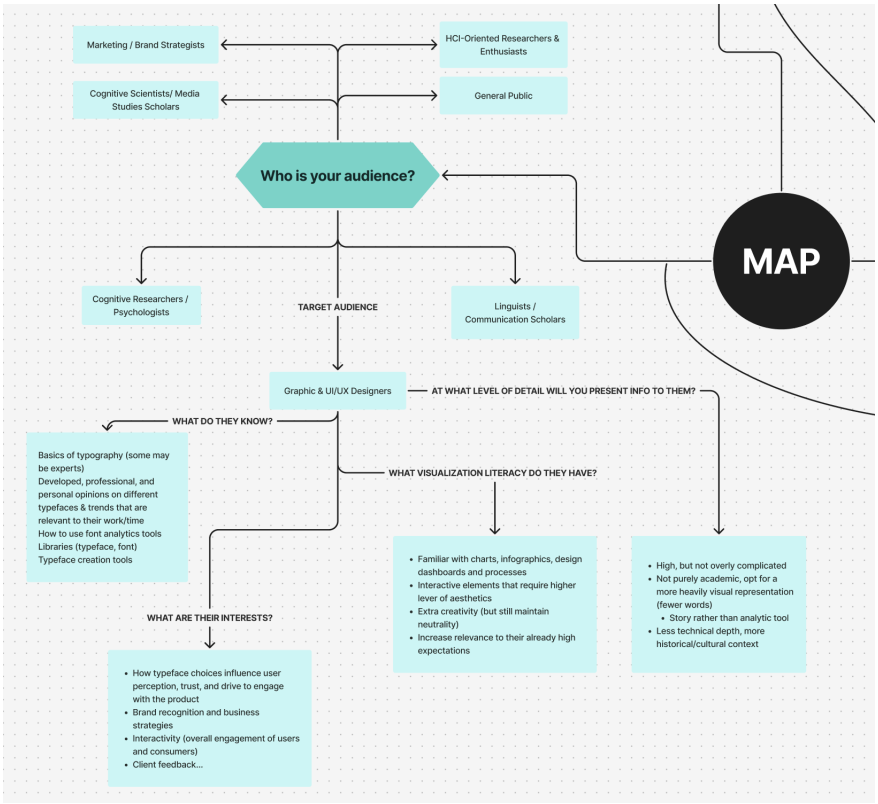
- Mass media analytics, open source library data, case studies
    - Google Font Analytics: *https://fonts.google.com/analytics*
    - Most popular web fonts each year: *https://www.typewolf.com/all-fonts*
- Large-scale logo dataset/labellings: *https://github.com/msn199959/Logo-2k-plus-Dataset* *https://datasetninja.com/logodet-3k*, *https://lhf-labs.github.io/tm-dataset/*
- *Lelis, C., Leitão, S., Mealha, Ó., & Dunning, B. (2020). Typography: the constant vector of dynamic logos. Visual Communication, 21(1), 146-170. https://doi.org/10.1177/1470357220966775* (Original work published 2022*)*
- Eye-tracking study (psychological analysis)
    - *https://etheses.whiterose.ac.uk/id/eprint/22135/*
- Trends
    - Emi, R. A. and Adekoya, S. (2019) “Computer, letters and typography in the twenty first century: antecedent, trend and expectations”, Gateway Information Journal, 20(2), *https://www.gatewayinfojournal.org/index.php/gij/article/view/20*
- Yearly data of web fonts
    - https://almanac.httparchive.org/en/2024/fonts

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

[ON-OLD NARRATIVE] Map

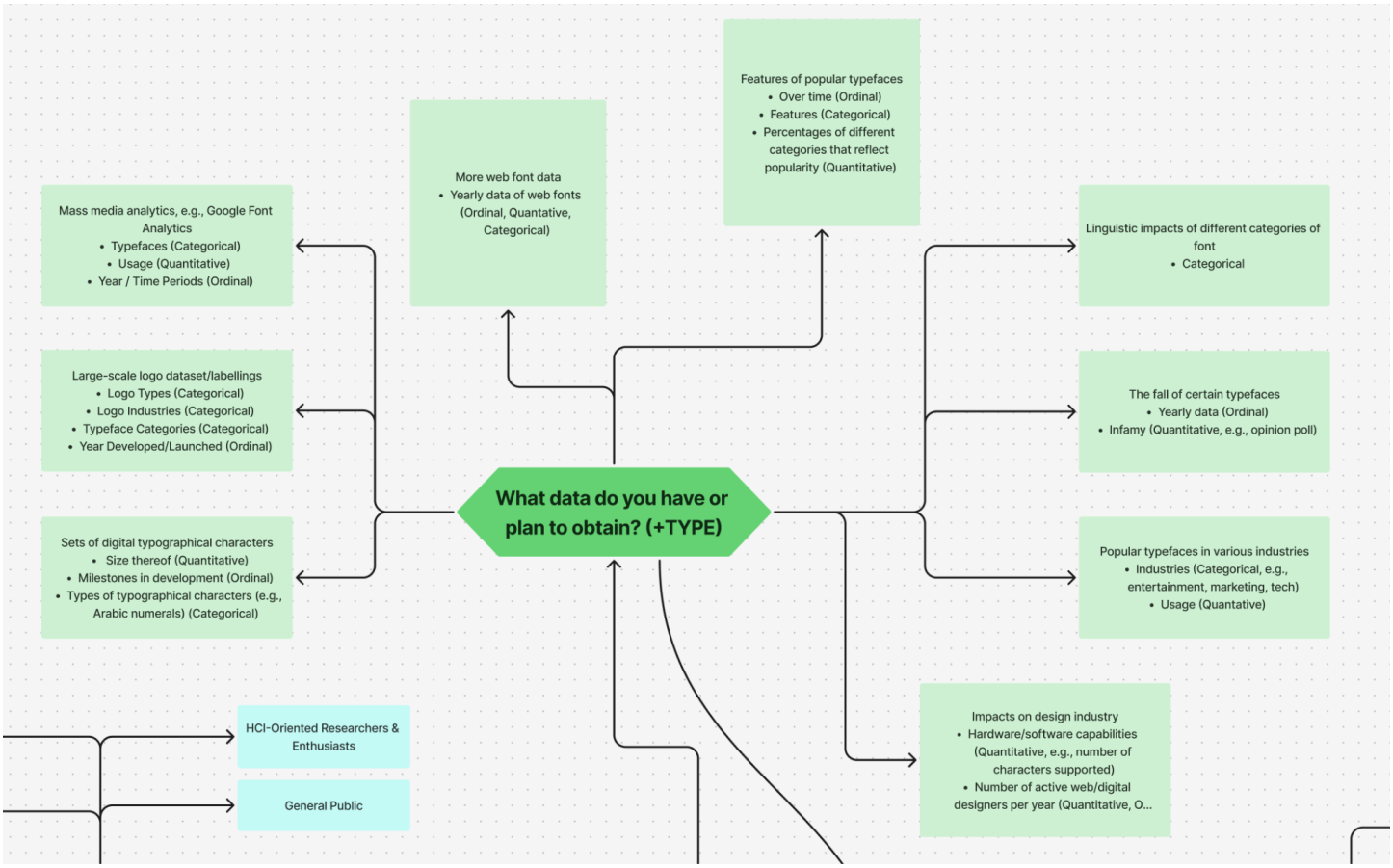


Mindmap (Figma Jamboard) link

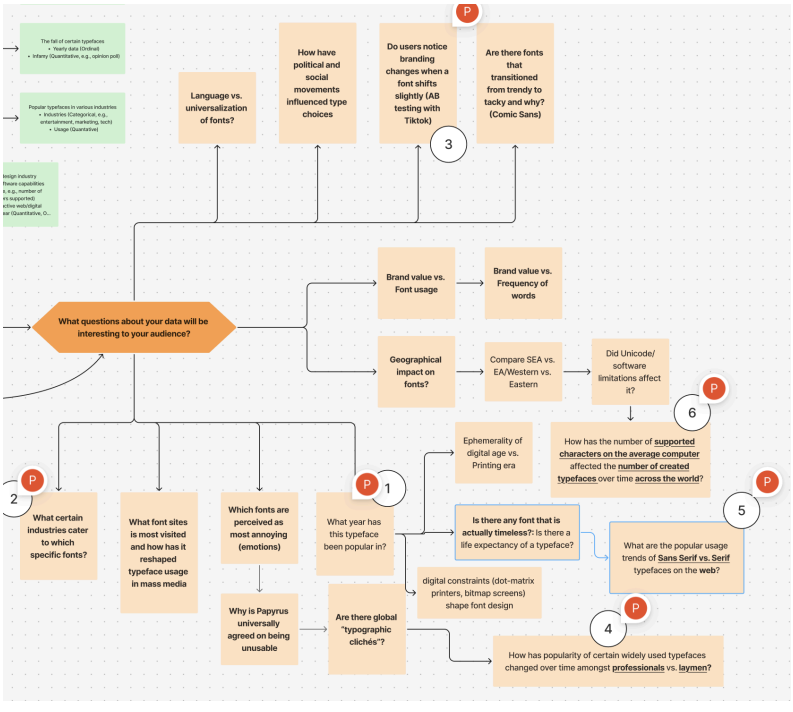


For our audience, we initially considered several directions, including people with backgrounds in typography, media history, and cognitive psychology. These groups would all have valid stakes in the questions of typeface perception and cultural impact. Here, in our mindmap, we focused on the three big framing questions: Who is our target audience? What data could we use? How can we analyze it for them? After discussion, we decided to orient our project toward graphic and UI/UX designers. This choice reflects both our own backgrounds and interests, as well as the fact that designers represent one of the largest and most active demographics engaging directly with typography in practice. Designers not only consume type but also deploy it strategically in branding, interfaces, and communication, which means

they are naturally attuned to the interplay of aesthetic choices and audience perception. We therefore see them as the most impactful audience for our visualizations.



For data, we continue to draw on the resources in our project proposal and categorize them. Although it has not been particularly designed for any set of questions yet, I do believe that narrowing down the scope would definitely help later on.



For questions, we wrote down whatever came to mind first that might be of interest to the general public. The numbers attached to the questions are our choice for visualization, with 1-3 being Chi's choices and 4-6 being Rain's choices. Comments are ways to visualize the data that we discussed at our meetup (to avoid overlapping), which is subject to change. Initially, these comments were our prompts for Claude, which turned out to have payment processing problems. In the end, we did not

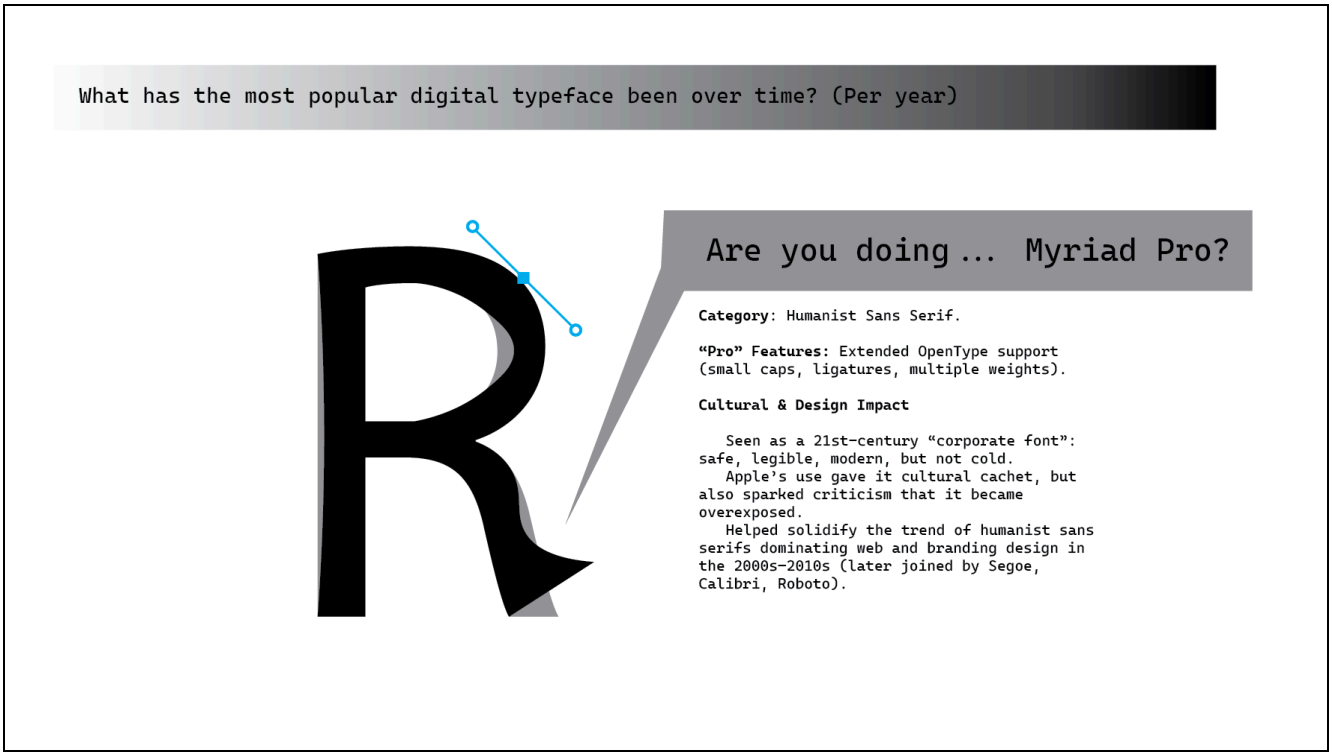
use AI; Chi used Adobe Illustrator and Rain used Figma for their visualizations.

**Note that the data in the visualizations below is placeholder data*

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

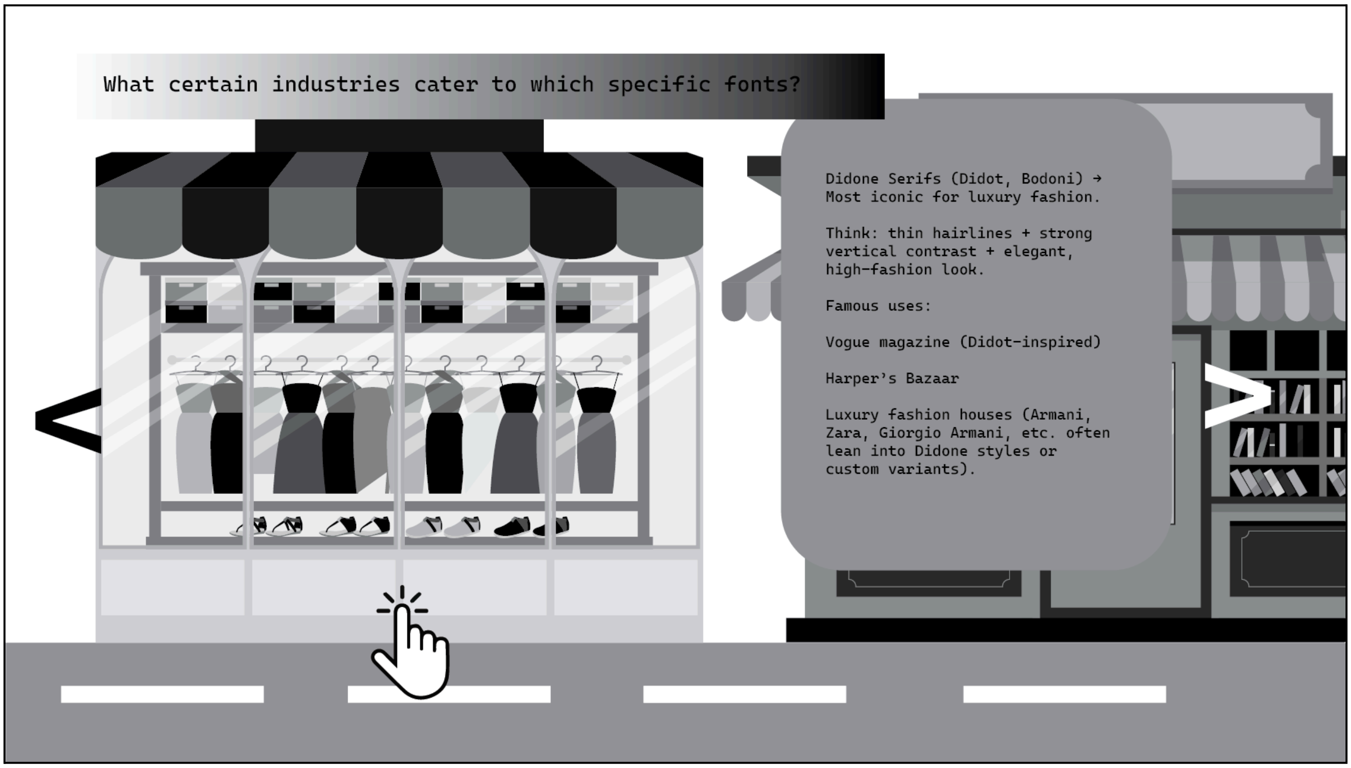
Chi

Visualizations 1–3 ideated from scratch & created in Adobe Illustrator

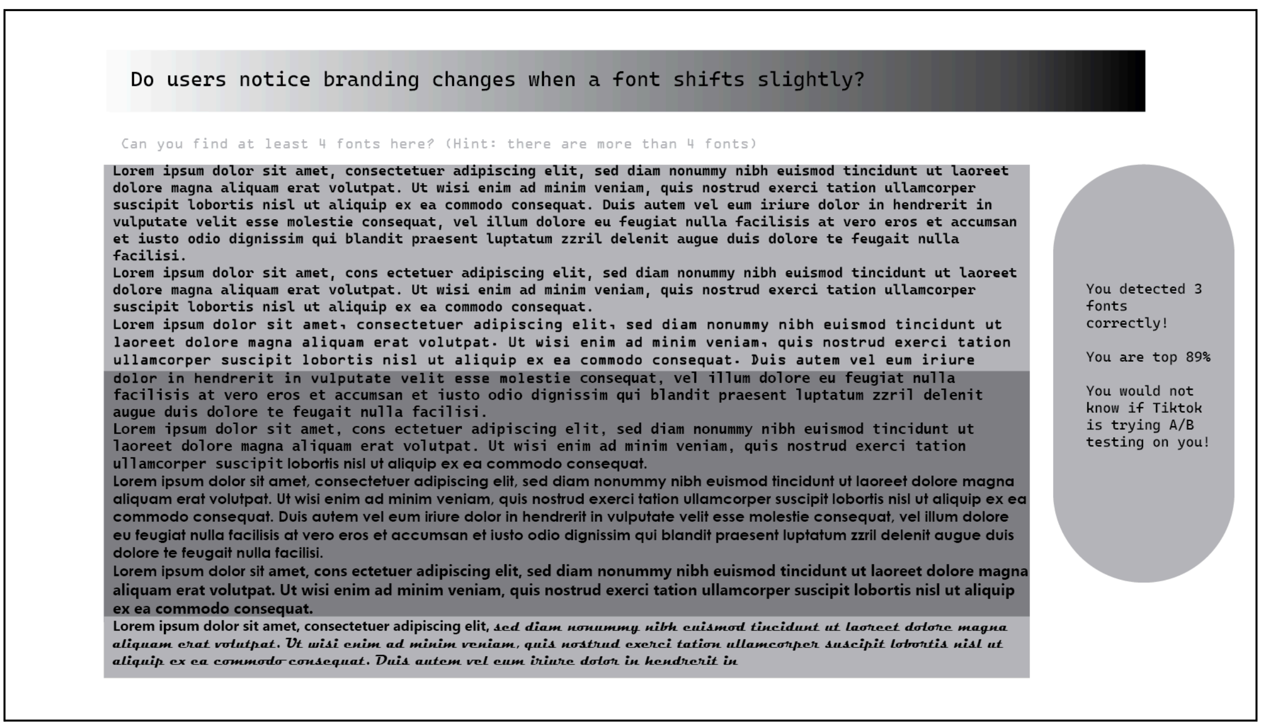


Visualization 1

[Process Book Link](#) / [Github Repository](#) / precideer.github.io/cs-1710-unicode



Visualization 2

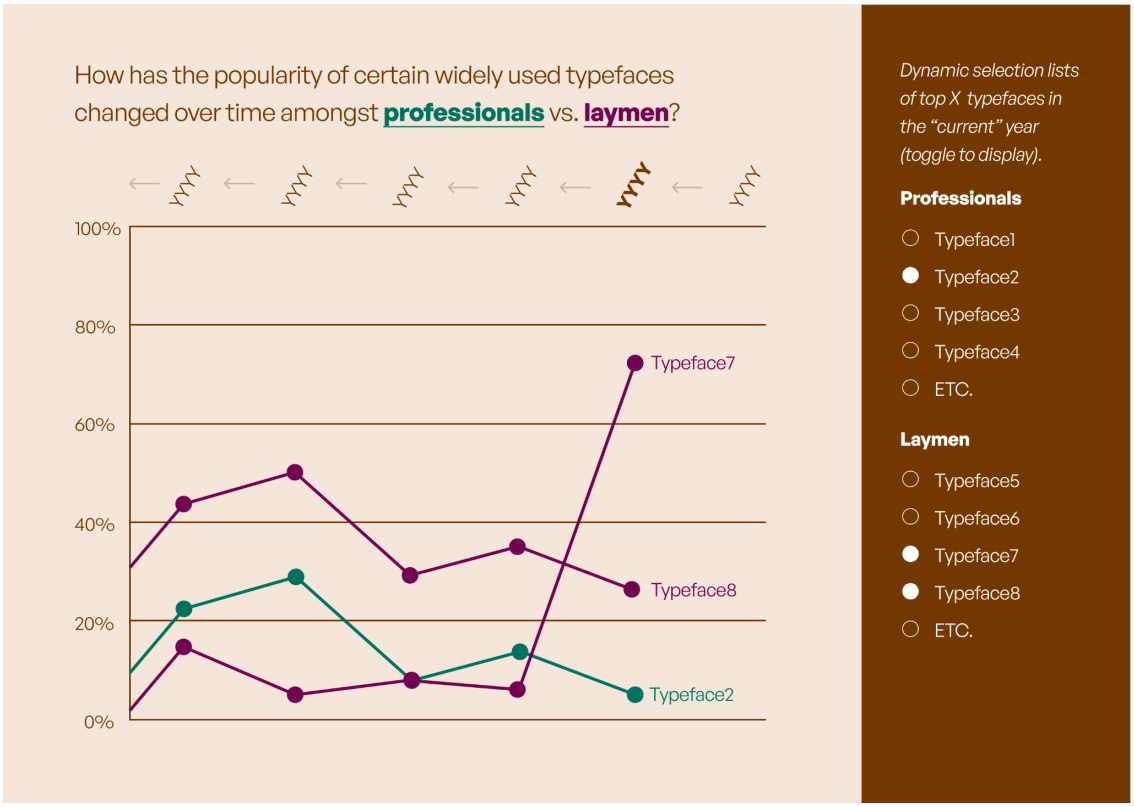


Visualization 3

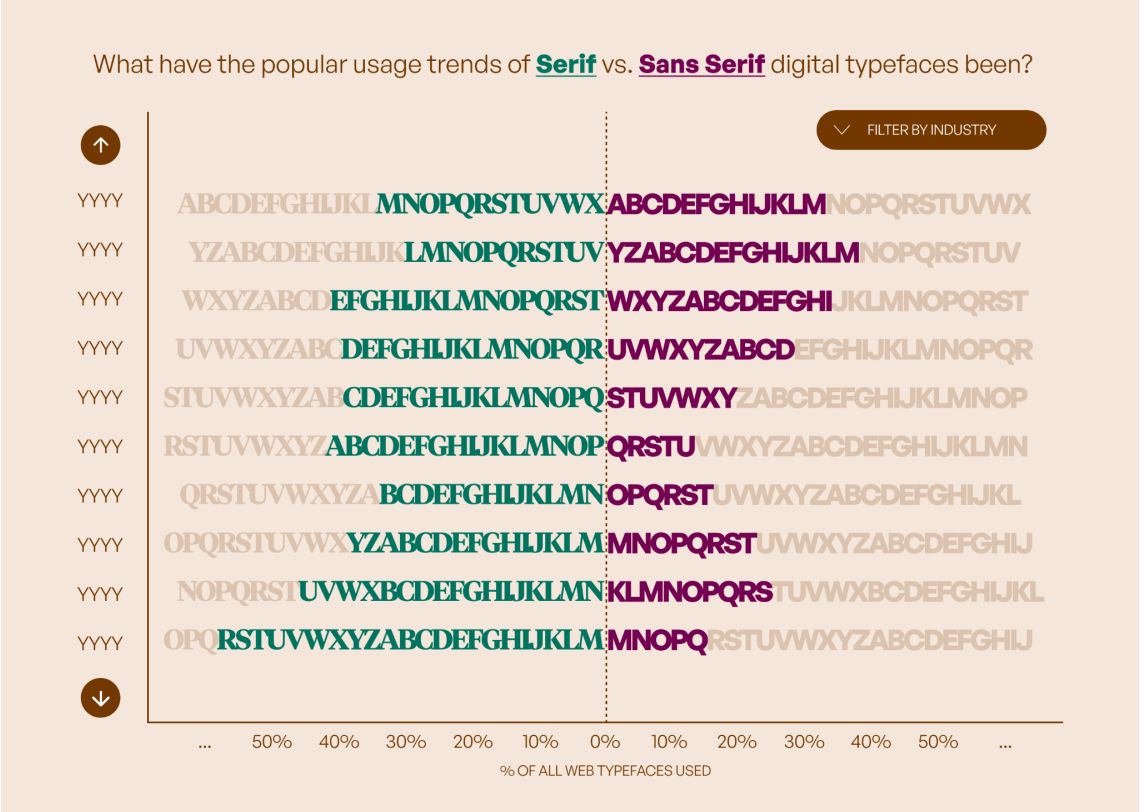
Rain

Visualizations 4–6 ideated from scratch & created in [Figma](#)

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

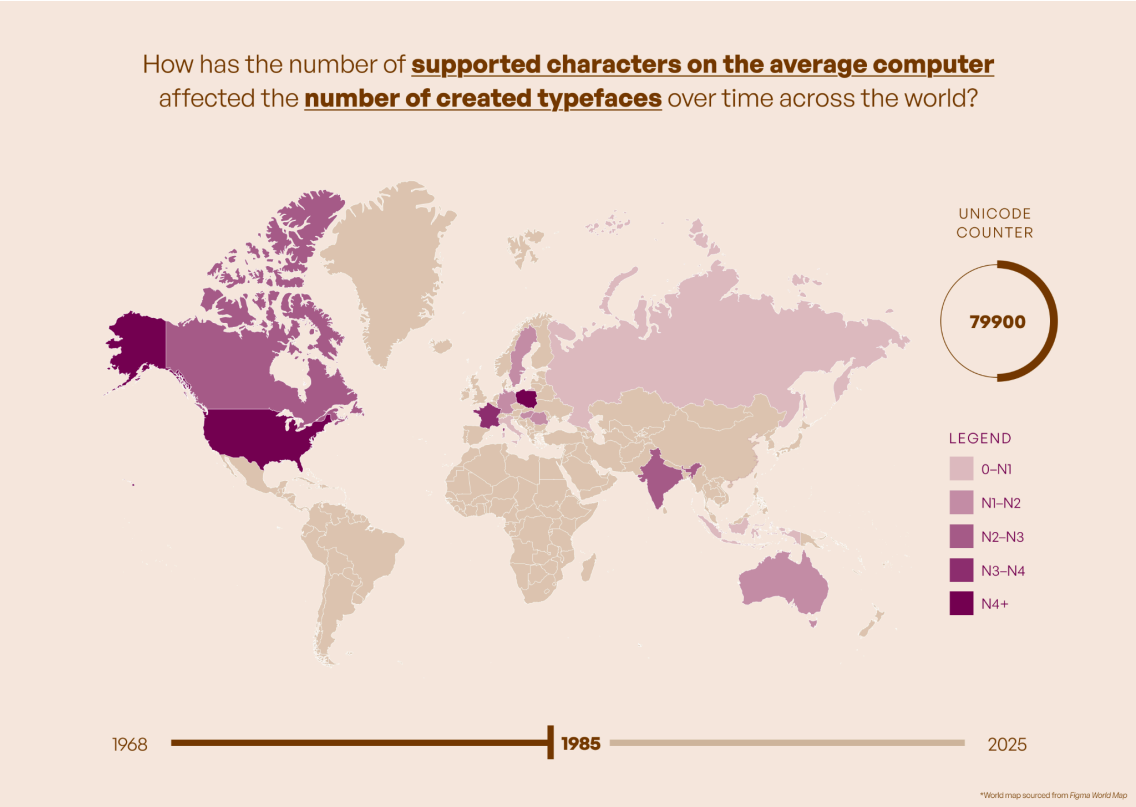


Visualization 4



[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Visualization 5



Visualization 6

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Reflection — Chi

The questions answered in my sketches build directly on the team's original questions because they already framed essential issues in typeface history and perception, but I wanted to adapt them into more interactive and metaphor-rich visualizations that align with our target audience of graphic and UI/UX designers. The original set leaned toward broad research inquiries and a more general public (e.g., "What fonts are popular across industries?"), which are important starting points. My sketches instead focus on translating those inquiries into experiences that resonate with how designers think. As a graphic designer myself, I would want something that is engaging visually, experientially, and contextually. For instance, rather than representing the "most popular digital typeface over time" as a static timeline, I proposed a Shape Type–inspired tool where users manipulate letterforms themselves, directly connecting the act of design to historical and statistical information. Drawing inspiration from a familiar reference also lowers the barrier for engaging with a novel visualization while maintaining its uniqueness. Similarly, for the question of whether users notice subtle branding changes, I built on the original idea by incorporating perceptual testing (e.g., detecting font shifts within a paragraph), which both embeds the data in the text itself and reflects designers' interests in user testing and feedback.

Reflection — Rain

The questions addressed in my sketches are derivative of our team's initial questions (especially since time is a ubiquitous data attribute throughout), but my focus was on building a framework centered around design-driven narratives that highlight contrast, competing trends, and a global perspective. Since our overall topic is so broad, we have a plethora of directions we could go in (as is probably obvious from our mind map), and the scope of questions under the same category can also vary significantly. I think some (new) questions are better than others (original) insofar as they are far more conducive to interesting, creative, and potentially engaging (*especially interactive*) visualizations. For instance, Visualization 4 gives the user a chance to compare their perspectives of popular typefaces against those of another group, and select/deselect typefaces of interest as the timeline goes by. Visualization 5 lends itself to convey an interesting competition (*because* the categories were narrowed down) and visually incorporate elements of typefaces as well. Visualization 6 paints a relatively more targeted and broad story of the impact of typefaces, which our original questions did not necessarily set us up to achieve.

[ON] Data

Source & Inventory

Dataset Name	Link	Inventory	Description	Restriction
Google Fonts Analytics (Download by #2)	Link	**Categorical:** font_name, category, designer, release_date, license, version, scripts_supported **Ordinal:** popularity_rank, **Quantitative:** views, downloads	Provides pageview/download counts and popularity ranks of Google Fonts over time.	Open data (Google Fonts API Terms)
Google Fonts Metadata API (JSON)	Link	**Categorical:** kind, family, version, color_capabilities, tags, subsets, variants, tags, **Ordinal:** variants (when numeric weights like 100–900) **Quantitative:** axes.min, axes.max	Contains font family metadata.	Open Source (SIL Open Font License / Apache 2.0)
Logo-2k+ Dataset LogoDet-3k Dataset (CSV + Image files)	Link 3D Link	**Categorical:** brand_name, industry, font_style_class , color_code,	Dataset of brand logos annotated by industry; usable to examine typographic style in corporate branding.	CC-BY-NC 4.0 (Non-commercial)
2024 Fonts queries (JSON)	Link	**Categorical:** font_family, host_type, file_format, font_display_value, color_capability, script **Quantitative:** page_count, usage_share **Ordinal:** Popularity_rank	The HTTP Archive crawls ~8.5 million URLs monthly using Chrome Lighthouse, recording each site's requests and CSS rules.	

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Dataset Name	Link	Inventory	Description	Restriction
TonerBuzz Font Statistics	Link	HTML / manually extractable table	Aggregated SimilarWeb data on font usage in industries (e.g., law websites: Open Sans 24.5%).	Informational / citation allowed
Creative Review Typeface Survey (2018)	Link	HTML	1,161 designers ranked their most-used and favorite typefaces.	Editorial reuse allowed
Typewolf All Fonts Dataset	Link	**Categorical:** font_name, foundry, year, classification **Ordinal:** year_featured, ranking_position	Annual rankings of the most popular web fonts by year	Fair use for research and citation
Typography study (by Lelis et al.)	Link	**Categorical:** brand_name, industry_sector, logo_type, typeface_used **Quantitative:** number_of_variants **Ordinal:** year_of_redesign	Academic study data on typeface changes in dynamic branding	Academic citation required
Typographic emphasis and contrastive focus: an eye tracking study	Link	**Categorical:** font_type, participant_id, task_type **Quantitative:** fixation_duration, reading_speed, comprehension_score **Ordinal:** readability_rating	Psychological study on typeface perception alongside legibility	Academic citation required (due to this being thesis data)

Tools used: Excel, ChatGPT Code Interpreter, Claude AI, and Google Sheets

Statistical summaries:

FOR GOOGLE FONT ANALYTICS:

- Distribution analysis
  - ~1,400 font families analyzed
  - Range of downloads: 10K to 50B+ (Roboto)
  - Mean downloads: ~500M
  - Median downloads: ~80M (right-skewed)
  - Top 10 fonts account for ~60% of all font usage

- Outlier detection
    - Roboto is an extreme outlier (3+ standard deviations) at 50B+ downloads
    - Open Sans, Lato, Montserrat have downloads in the 10–20B range
    - More than 70% of fonts have <100M downloads
- Category distribution
    - Sans Serif (~58%) — Serif (~28%) — Display (~10%) — Handwriting (~3%) — Monospace (~1%)

FOR HTTP ARCHIVE FONTS DATA:

- Trends over time
    - 2019–2024 showed a 340% increase in font file requests per page
    - Variable fonts grew from 0.2% to 8.5% between 2020 and 2024
    - WOFF2 format adoption grew from 45% to 87% from 2019 to 2024
- Missing data patterns
    - Around 12% of records lack font_display values
    - Color capability data is only available from 2022 and onwards
    - Script coverage is incomplete for fonts added before 2018

FOR LOGO DATASETS:

- Industry breakdown
    - Technology (~35% sans-serif geometric fonts) — finance (~45% serif fonts [traditional trust markers]) — fashion (~60% custom/modified typefaces)
- Data quality shortcomings
    - Only ~40% of logos have font classifications (many will require manual font identification)
    - Time period coverage is also somewhat inconsistent (most of the data is from 2010–2023)

Examples of AI interaction prompts:

- *Summarize font popularity distributions from the Google Fonts metadata CSV and detect outliers*
- *Group fonts by category and visualize frequency*
- *"Load the Google Fonts metadata CSV and show me the first 10 rows with basic statistics for the 'downloads' column." → Generated pandas DataFrame summary showing mean, median, std dev, and identified Roboto as extreme outlier*
- *"Help me standardize the font category labels. I have variations like 'Sans Serif', 'sans-serif', 'SANS_SERIF'. Create a cleaning function." → Provided Python function using .str.lower().str.replace() to normalize all variations to 'sans-serif'*

- *"Analyze missing values pattern in the HTTP Archive dataset. Which variables have the most missing data and is there a pattern by year?" → Could generate heatmap showing font_display_value missing primarily in 2019-2020 data (pre-standardization), color_capability missing pre-2022*

Data Cleaning

Issue	Description	Action
Missing release years	Lack explicit release year	Imputed by designer's earliest work date or API "last modified" field
Inconsistent categories	Variants like "Sans Serif", "sans-serif"	Standardized labels (lowercase + hyphen)
Duplicate font families	Roboto, Roboto Condensed	Consolidated families under canonical name
Non-ASCII characters	Fonts with international names	Normalized UTF-8 encoding
Multi-value fields	"scripts_supported" has comma-separated lists	Split into list objects or binary flags
Inconsistent units	Views in trillions (Google), % (Web Almanac)	Normalized to percentages or log-scaled where appropriate
Nested JSON structures	Google Fonts API returns nested objects for variants/axes	Can flatten using json_normalize() and created separate rows for each variant weight
Logo image quality variance	Logo-2k+ images range from 50px to 4000px	Can filter for minimum res (e.g. 300 px) and exclude low-quality samples
Survey response bias	Creative Review survey sees UK designers over-represented (68%)	Document geographic skew (can segment by region if more data can be collected)
Categorical ambiguity	Some fonts can be classified as both Display and Serif fonts	Create hierarchical classifications (e.g., primary category vs. secondary category)
Gaps in industry classification categories	Logo dataset uses more than 20, survey uses only 8	Generalize these categories, or merge as we see fit to create an appropriate umbrella taxonomy

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Reflection

Our original goal was to tell a cohesive story of how (primarily digital) typography has evolved with technology and culture over time, visualizing the relationships between design trends, technological constraints, and perception trends. The actual datasets exceeded our expectations in quantity but also suffered from varied consistency: web-based data sources (e.g., Google Font Analytics, HTTP Archive) provided rich and well-structured quantitative data with excellent timeline coverage from 2015 to 2024, enabling us to track clear patterns like market shares of certain font classifications. However, qualitative dimensions of our datasets were more inconsistent: brand and survey data were semi-manual and required cleaning. Most of our data came from digging into web reports' citations to find if we could access the data. There are some measurable correlations — such as sans-serifs peaking in 2019 web usage and the rise of variable fonts since 2021 — that directly answer earlier questions about typographic "epochs."

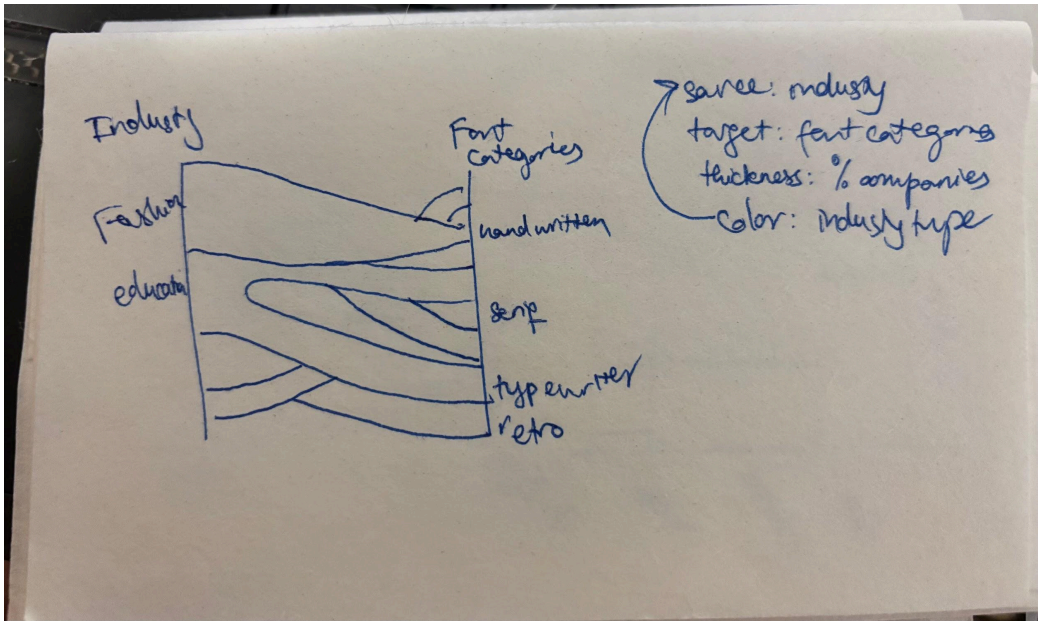
Some questions we can now definitively answer include ones that pertain to temporal popularity trends (Which fonts dominated 2015–2020?), technological adoption of certain typefaces ("How quickly did variable fonts spread?"), and certain subsets of industry patterns ("Do finance brands prefer Sans Serif or Serif fonts?"). However, not all audience questions could be answered purely quantitatively. Topics like "emotional perception" or "branding authenticity" required qualitative interpretation of survey reports and case examples (e.g., Comic Sans vs. Helvetica trust metrics). The biggest challenges were data fragmentation (no unified schema across sources), inconsistent time coverage (especially pre-2000), and manual extraction for older or industry-specific statistics. Yet, these limitations underscored a key insight for our visualization: the incompleteness of data itself mirrors how typography's evolution is nonlinear, diverse, and continually reinterpreted.

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

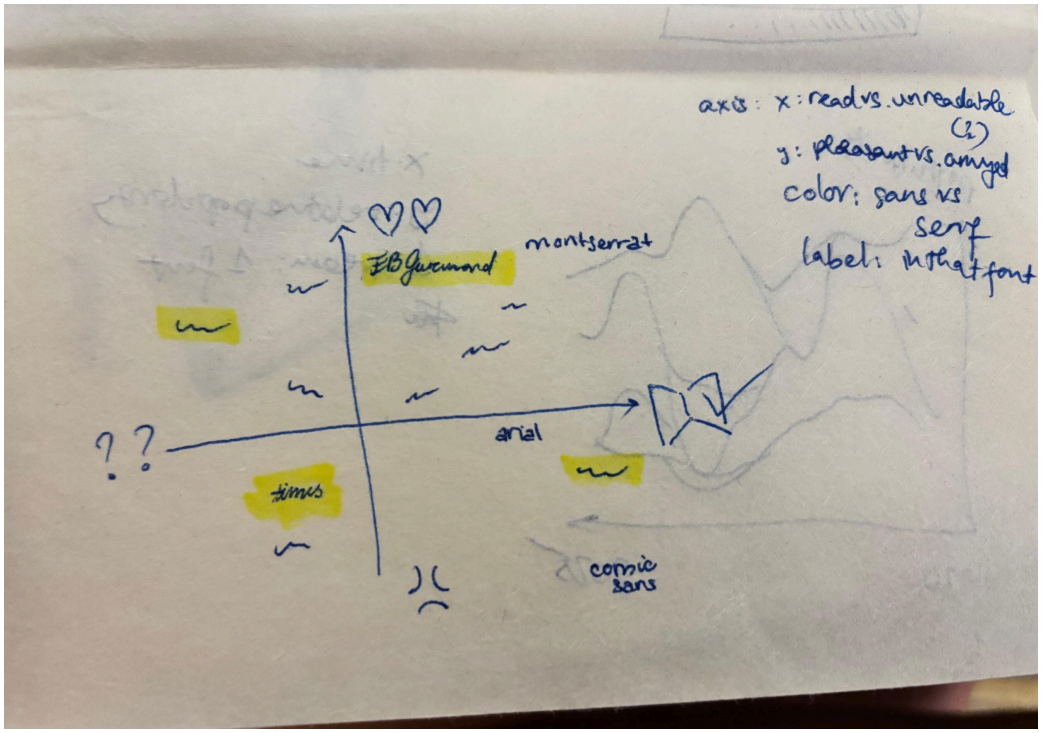
[ON] Sketch

Chi

1. Which certain industries cater to which specific fonts?

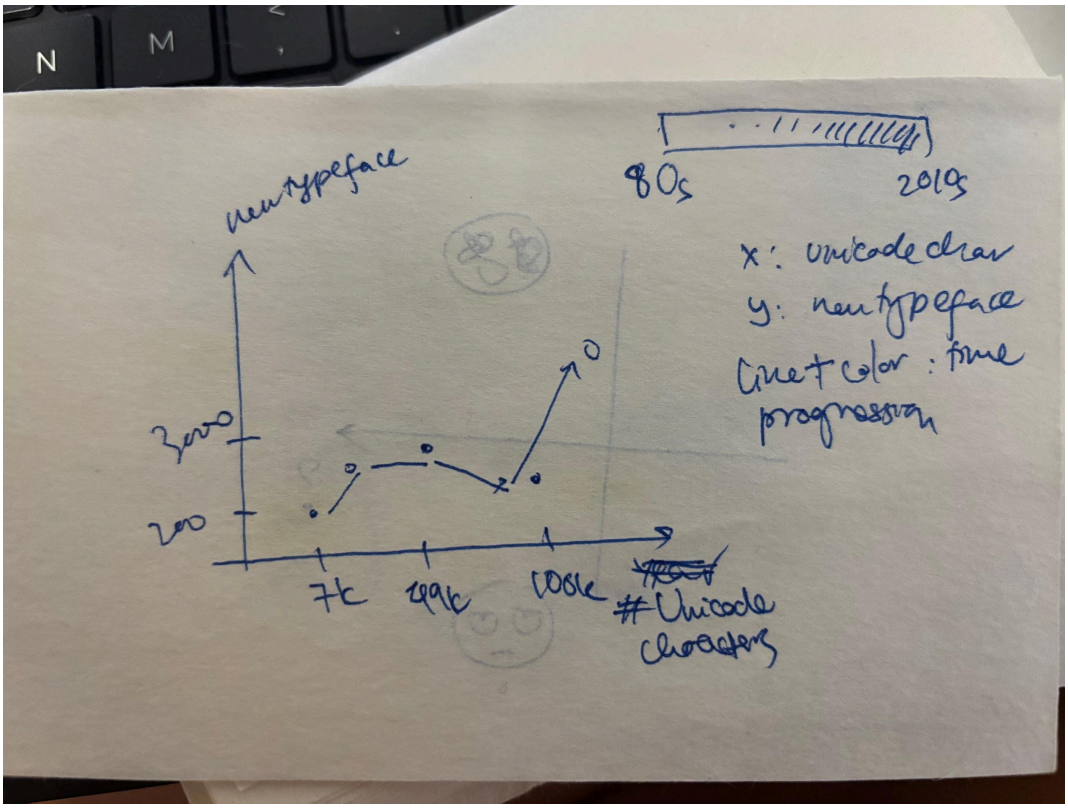


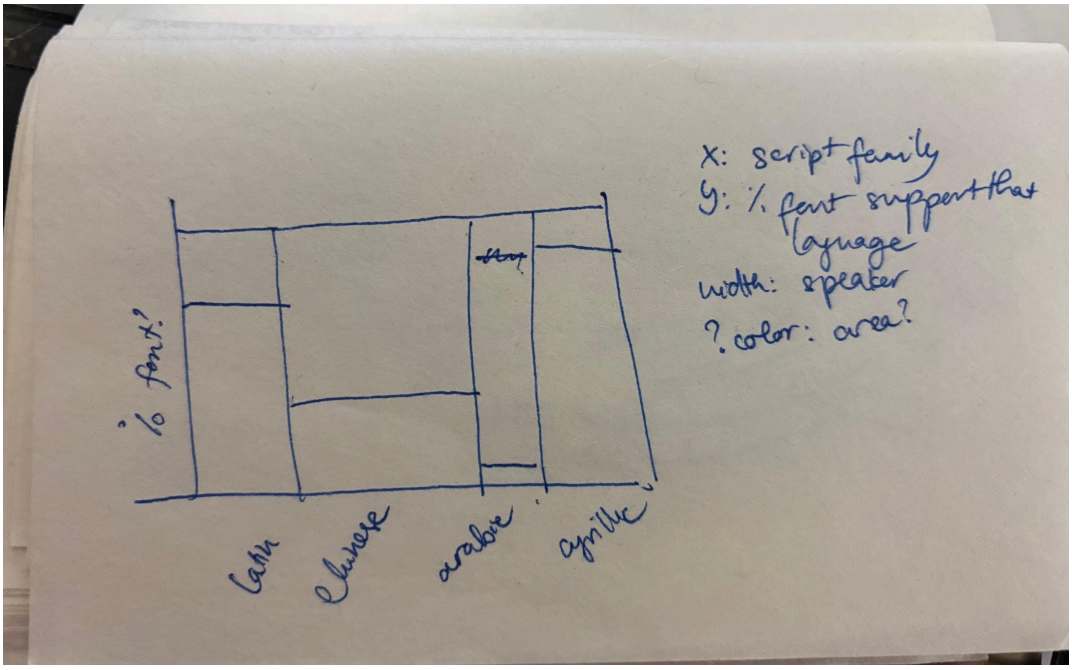
2. Which fonts are perceived as most annoying?



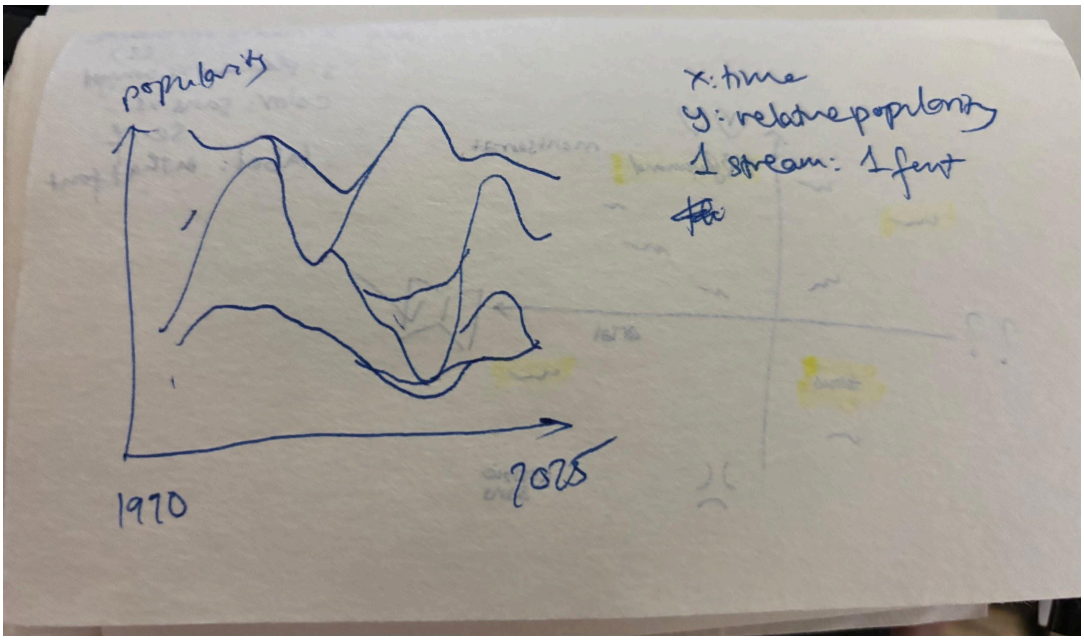
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

3. How has the number of supported characters on the average computer affected the number of created typefaces over time across the world?



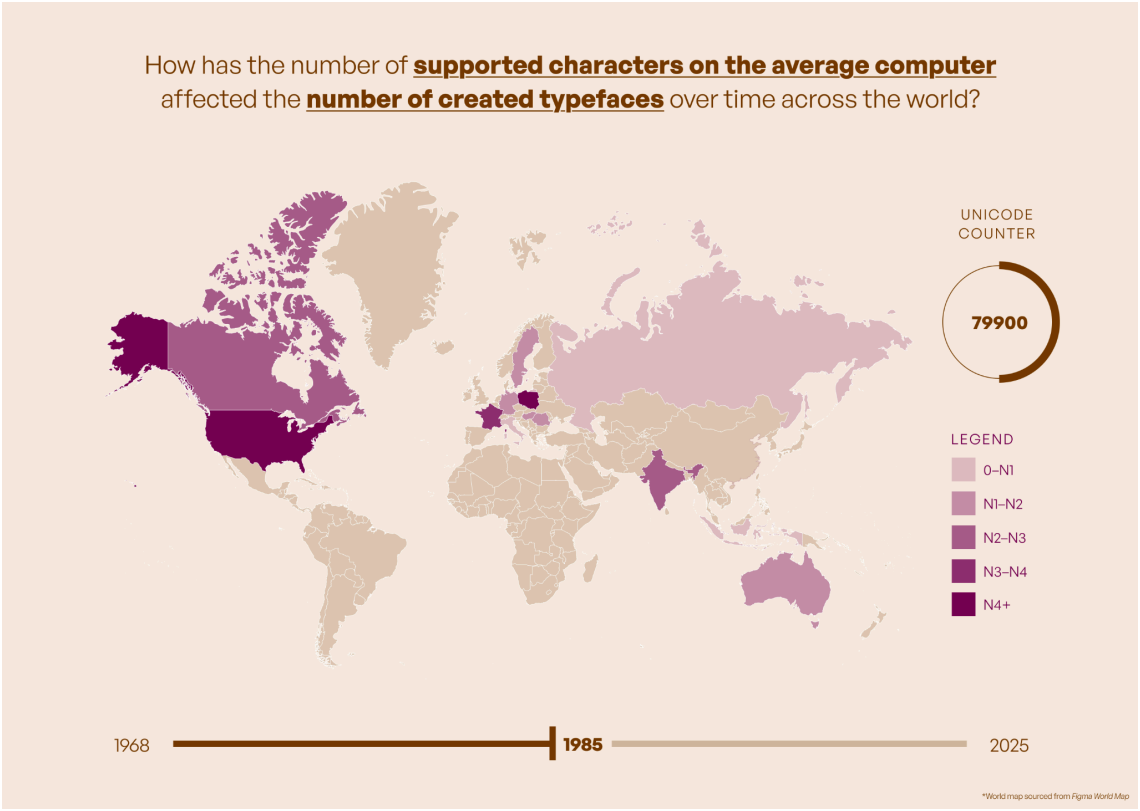


4. Are there fonts that transitioned from trendy to tacky?



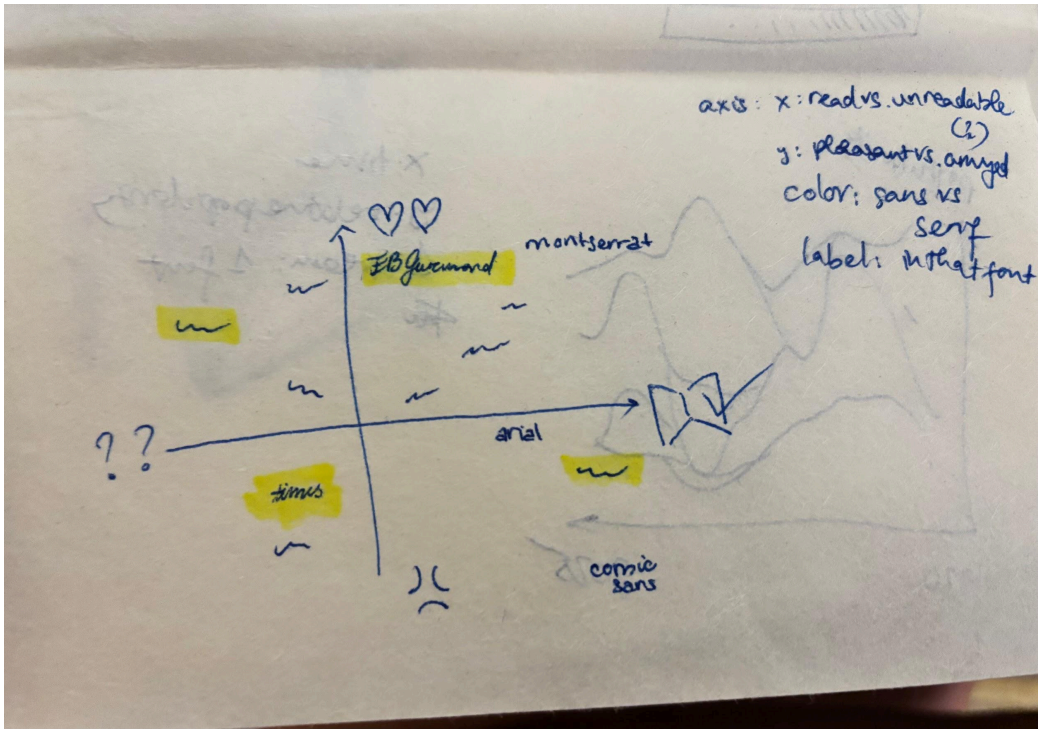
[ON] Decide

First sketch



[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Second sketch: Which fonts are perceived as most annoying?



Since we are quite lacking in sketches, I decided to combine both visualizations and my sketches since they also head towards the same direction of answering the questions we proposed. I also emailed Richard and he said it is fine to move forward with any of my 2 decisions. For my final implementation, I chose these two sketches because they together reveal both macro-level patterns and micro-level perceptions within typography. The first visualization of the world map makes it intuitive to interact and the different opacity allows users to easily understand the narrative of the global expansion of type design in relation to the increasing number of supported characters on computers. It also allows viewers to easily identify regions where communication and typographic development may have progressed more slowly due to limited character support. The second sketch shifts focus to emotional and subjective responses by visualizing which fonts are perceived as most or least annoying based on readability and pleasantness. I find this abstract representation to be a quirky yet insightful way to give users a quick, engaging understanding of what qualities they might consider incorporating into their own type designs.

[ON] Storyboard

I approached this project differently from our initial plan. Since our available data sources turned out to be less diverse than expected, I decided to reverse-engineer the story. Instead

of forcing a theme upon the insights from the data, I began by studying our existing visualizations to understand what narratives they naturally revealed, and only afterward found more datasets to support those directions. After a while, I categorized all of our Miro questions and visuals into three thematic groups: **current trends**, **historical trends**, and **user-centered insights.**

Originally, my plan was to structure the story chronologically, moving from *historical → current → user insights*. However, when brainstorming the hook, I realized that this linear approach didn't feel engaging since it needs to be something relatable to the viewer. Therefore, I decided to invert the order to *user insights → current → historical → user insights,* which creates a circular narrative: beginning with personal perception, expanding outward to data and history, and ending back with reflection and implications for users and designers.

Message/Question

We all know/can infer that typography reflects our cultural/technological/etc. development. However, is it also accidentally producing an aesthetic monoculture?

Hook

Our story begins with an interactive prompt: "**Click if you trust this button,**" shown in three fonts (one of my initial ideas) — most likely to be Baskerville, Comic Sans, and Computer Modern (because I found this article later on). This quick interaction either reveals our central premise, that we don't consciously *see* fonts, but we instinctively *feel* them (algorithm's power), or shows a shift in aesthetic recognition.
https://github.com/radames/google-fonts-analytics-archive?tab=readme-ov-file

Rising Insights

As said in the above outline, we will follow this structure:

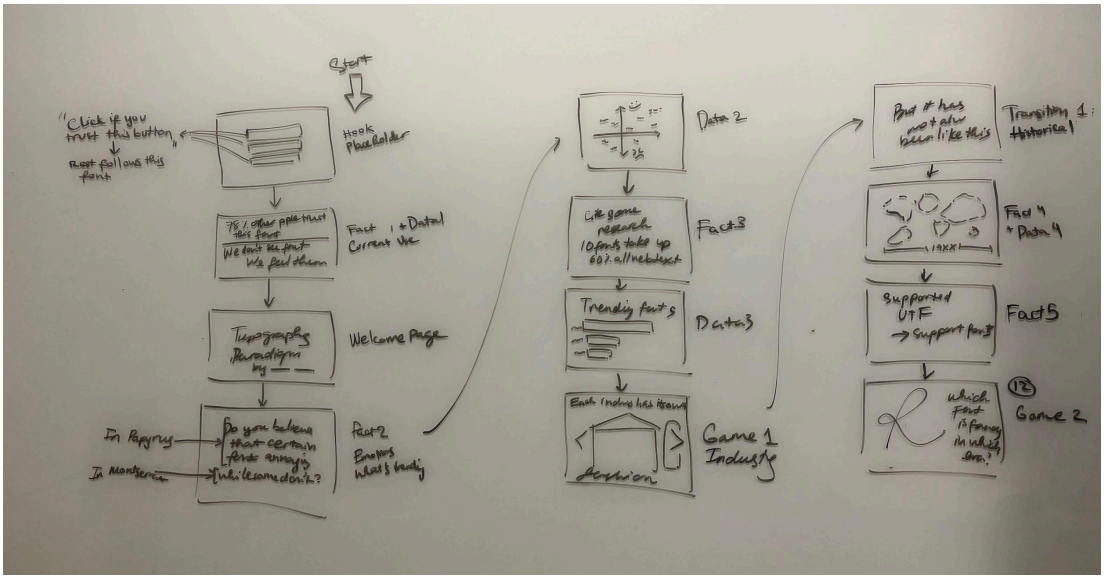
1. User Insights: Using affective data (FontLex) and readability metrics (Rello et al.), we visualize how fonts balance emotion and legibility.
2. Current Trends: Google Fonts Analytics and HTTP Archive data reveal that just ten fonts account for over 60% of all web text.
3. Historical Trends: Unicode expansion and the evolution of web formats (WOFF2, variable fonts) illustrate how technology broadened what's possible yet narrowed what's visible.
4. User Insights (Return): Comparing professional vs. layperson preferences highlights that designers value nuance while the general public gravitates toward defaults — showing how human perception and algorithmic systems converge.

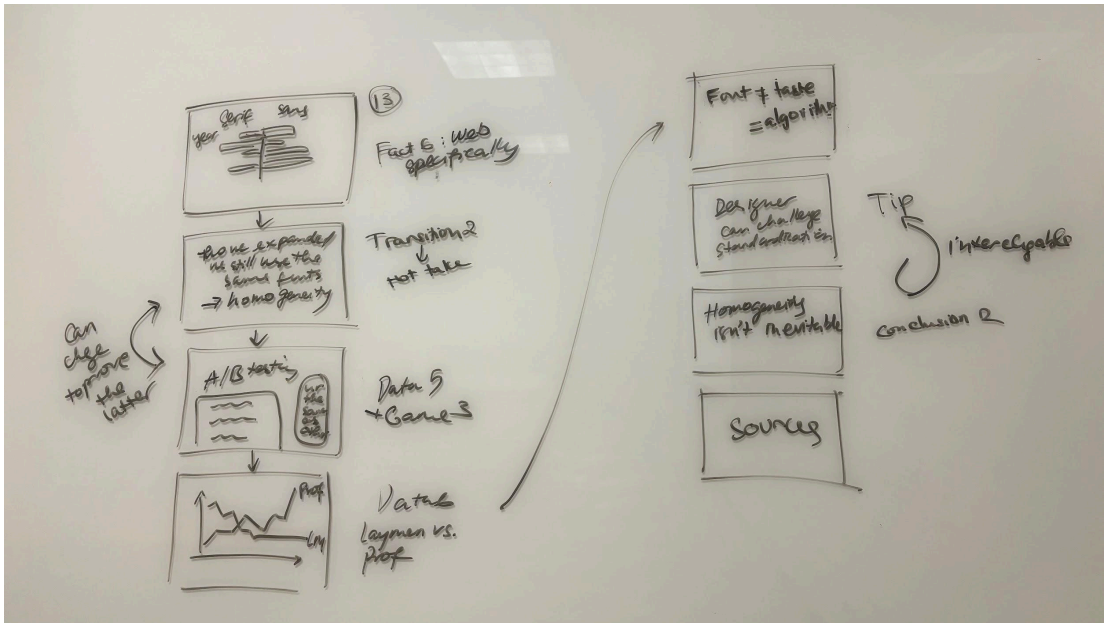
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Solution

Our conclusion emphasizes agency, that homogeneity is not inevitable. Designers can consciously resist algorithmic defaults by reintroducing friction, selecting fonts that reflect local voices, cultural context, or emotional resonance rather than relying solely on performance metrics. We end our visualization with something like “every letter can write your story if we let it”.

Storyboard Sketch





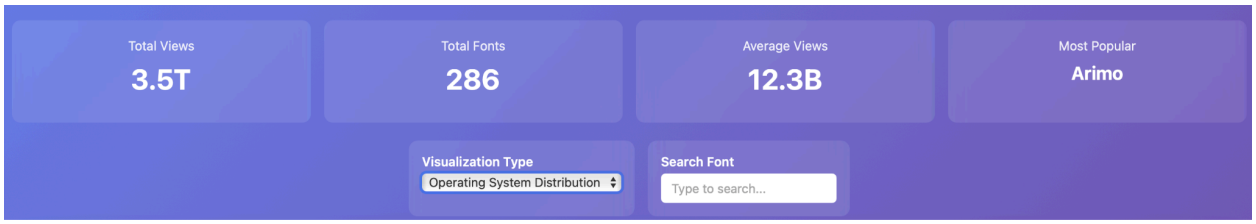
[ON+ NN] Prototype V1

[Link to the full Drive](#)

Currently, we are redirecting our project toward a new narrative, returning to the Mapping stage to reframe our focus. Since Rain is still catching up with her workload and we were only able to briefly discuss our direction, we decided that, for this milestone, the data pipeline, additional visualization drafts, and D3 visualizations will continue to follow our previous narrative. Meanwhile, any sketching and conceptual work will shift toward our new narrative on the development of character encoding and communication.

[ON] Visualization

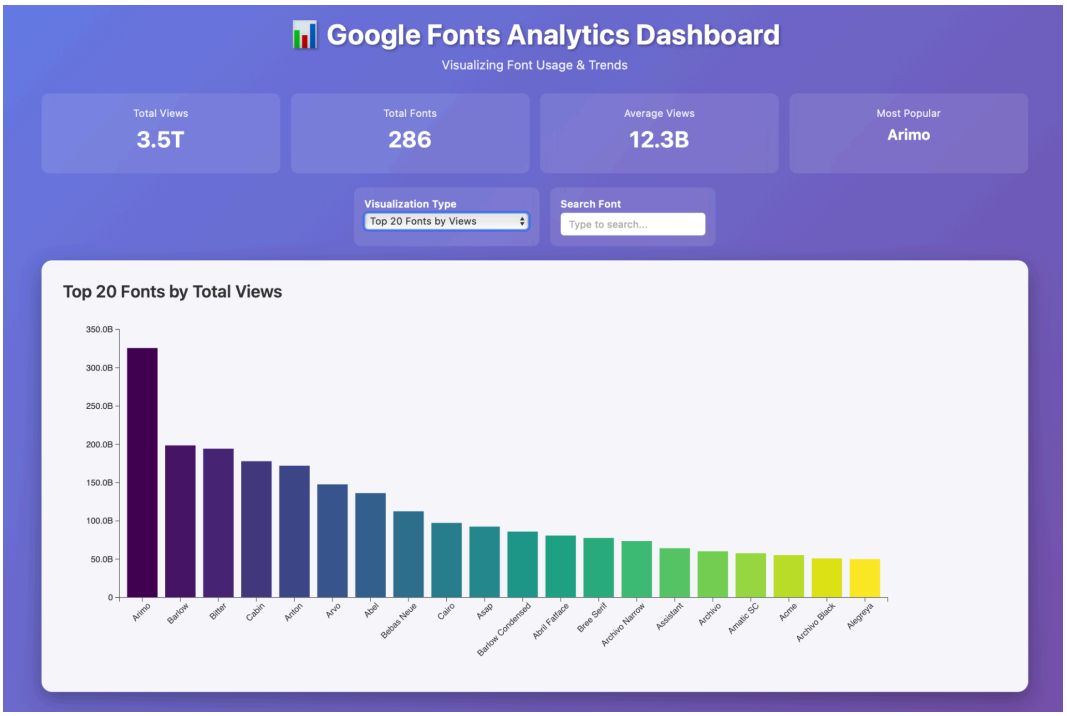
In either of our visualizations, users can view the total views, fonts displayed, and average view per font on the Google Fonts Analytics. For each, they can also search fonts to view individually if it is not on display. Any fonts outside of Google Fonts are not in this dataset.



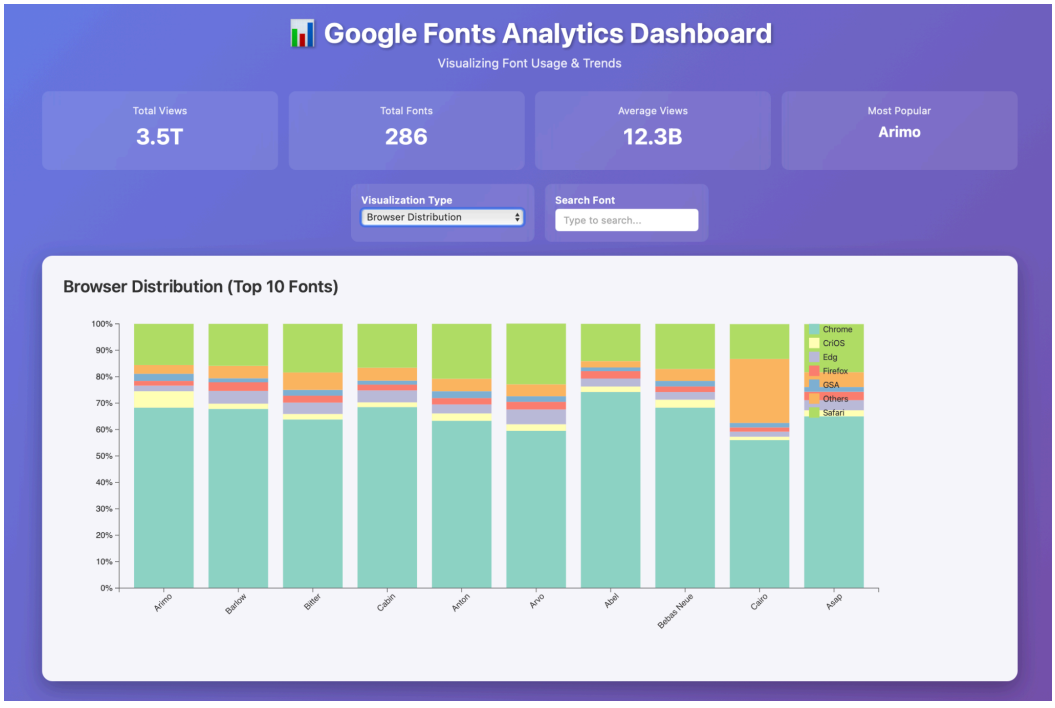
Users have the option to view different modes for these visualizations. Any of the visualizations are in the same dropdown box. In particular, we have 2 different visualizations which are bar charts for the first 2 (more about comparison among other fonts), and stacked bar charts for the last 2 (more about distribution within itself).

In our Drive, the folder *Data Scrapped* contains datasets associated with our earlier topic. However, as we move forward, we agreed that exploring Character Encoding vs. Timeline offers a more public-facing and literal framework—narrower in scope but clearer in storytelling potential. To support this transition, I experimented with ChatGPT to generate a summary of findings related to this new direction.

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



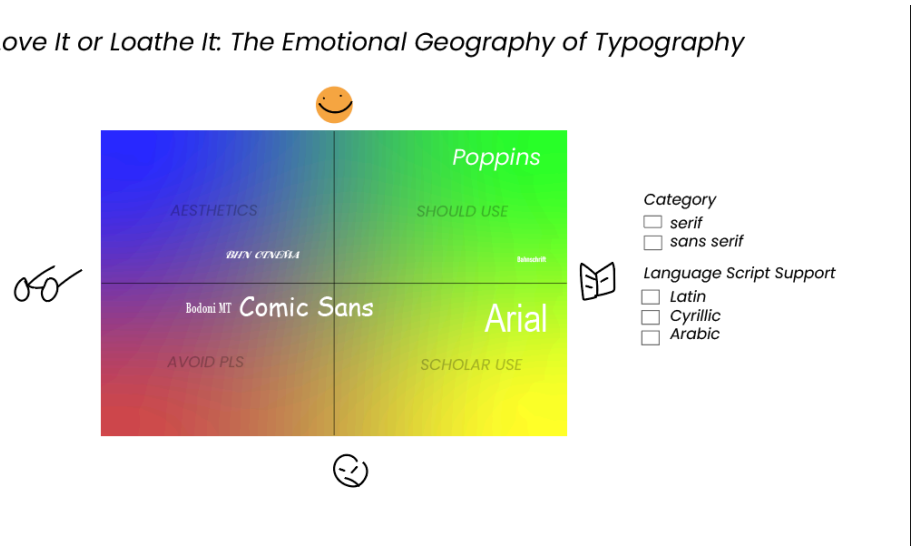
The first visualization is top 20 views for fonts with data scraped from Google Fonts.



The second visualization is browser distribution, on which users view these fonts through, for top 10 most viewed fonts.

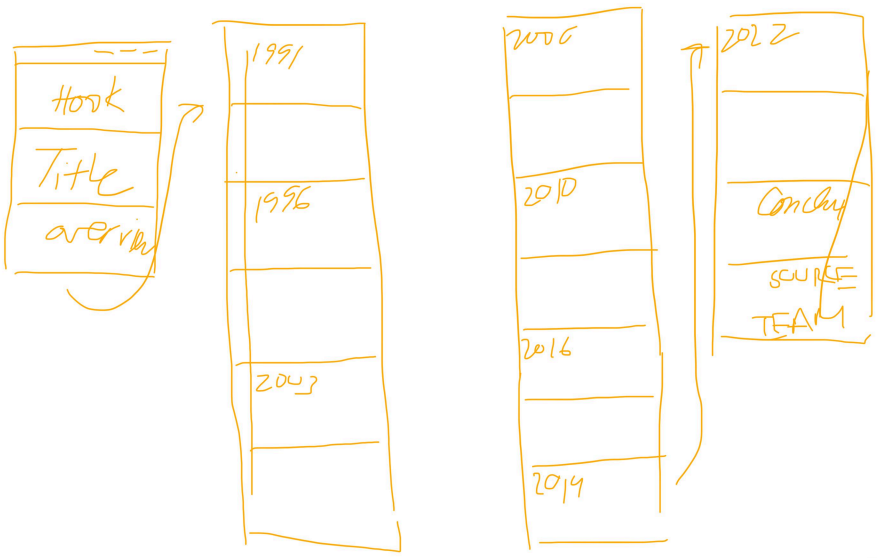
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

[ON] Additional visualization drafts



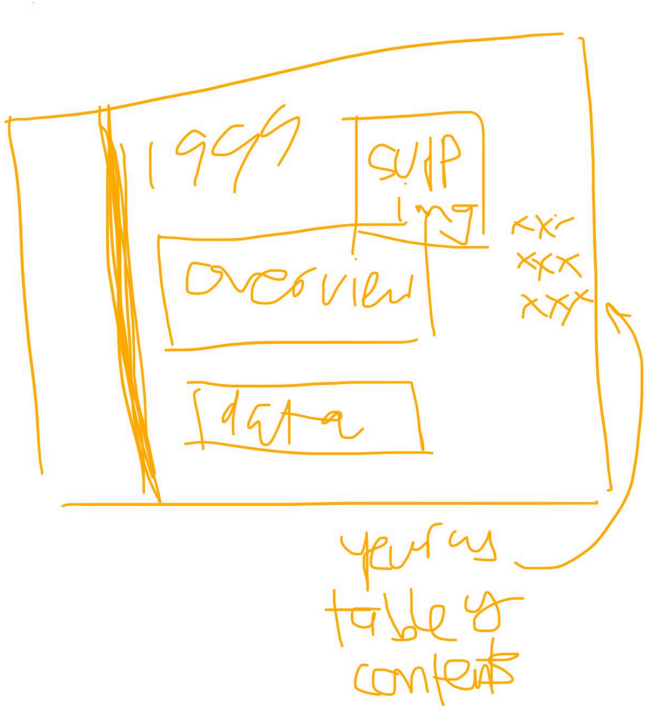
[NN-NEW NARRATIVE] Website structure

The general structure of the website will take the form of a timeline



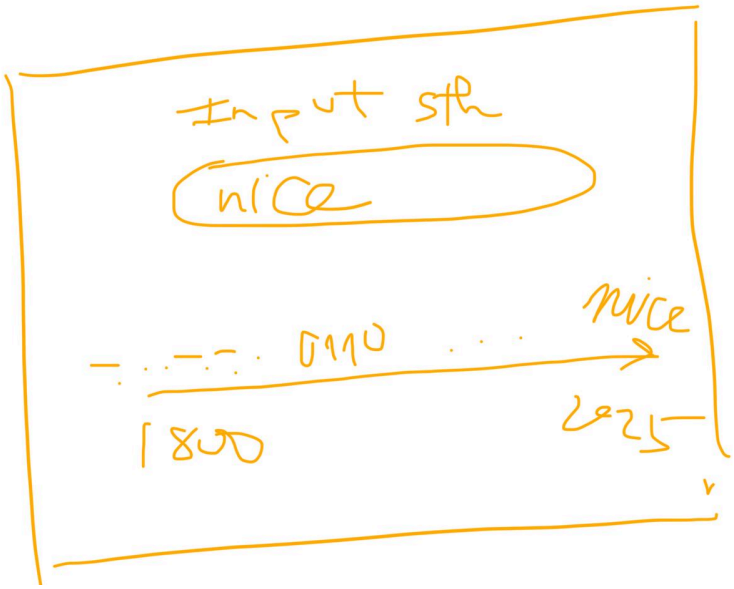
Here, on the General Page onwards, there will be a navigation bar that includes options to display particular information. We will decide later if we should do separate timelines for separate supported sets of characters (story of emojis, story of Chinese letters, etc.) or maybe display as that era's encoding environment. Visualizations will be on currently blank pages. An average timeline page will look something like this:

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



One innovative visualization concept is a “Character Flow Interface”, which will animate the translation of language through encoding layers. Users will be able to:

- Input a letter or phrase and see how it transforms through different encoding systems (e.g., Morse → ASCII → UTF-8 → Unicode). Particularly, we will try to add transitions and animations through phases to put more emphasis on each era. Observe how each system’s limitations and structure affect what can or cannot be represented.



[NN] Next Step

Since we decided to take on the timeline as the general skeletal structure, we want to make sure there is an additional layer of story told besides the popularity and what is used by which era. Our goal is to move past a purely historical record and instead highlight the human, cultural, and linguistic implications of these technological transitions. Possible questions/visualizations include:

- When did your language become part of the internet?
- Now that every language can be represented digitally, what happens to the ones we stop using? What happens to a culture when its language arrives late — or never — in the digital world?
- ASCII: When computers learned English, but not the rest of the world/Why couldn't your grandmother type her name online before 2005?
- Hook (?): Why do emojis look different on your iPhone vs your friend's Android?
- Factual > Data: When Morse turned into ASCII, and ASCII into Unicode — what got lost in translation?
- The future of bits and pieces: New keyboard design → Would there be emoji/Chinese URLs?

[NN] Data: Chi

Dataset 1: World Bank — Adult Literacy Rate (% ages 15+)

URL: https://data.worldbank.org/indicator/SE.ADT.LITR.ZS
File format: CSV or Excel (available directly via the download button)
Source: The World Bank, compiled from UNESCO Institute for Statistics
Usage restrictions: Open access under the World Bank Open Data License.

Number of items: ~200 countries × ~55 years (≈ 11,000 records)
Number of variables: 5

Variable	Description	Type	Value Range	Missing Data
`Country Name`	Country name	Categorical	200+ values	None
`Country Code`	ISO 3-letter code	Categorical	3-char string	None
`Indicator Name`	Literacy rate (% ages 15+)	Categorical	Single value	None
`Indicator Code`	SE.ADT.LITR.ZS	Categorical	Fixed	None
`Year`	Year of observation	Temporal	1970–2024	Sparse pre-1990

Dataset 2: UNESCO UIS Literacy Gender Parity Index

URL: https://databrowser.uis.unesco.org/view
File format: CSV (export via Data Browser)
Usage restrictions: Open for academic use under UNESCO data terms.

Number of items: ~180 countries × 4 indicators × 50 years ≈ 36,000 records
Indicators included:

- `LR.GALP.AG25T64.GPIA` – Adult literacy GPI, ages 25–64
- `LR.GALP.AG15T99.GPIA` – Adult literacy GPI, ages 15+
- `LR.GALP.AG15T24.GPIA` – Youth literacy GPI, ages 15–24
- `LR.GALP.AG65T99.GPIA` – Senior literacy GPI, ages 65+

Variable schema:

Variable	Description	Type	Range

`Country`	Name	Categorical	180+
`ISO_Code`	ISO 3-letter	Categorical	3-char
`Year`	Year of record	Temporal	1970–2024
`Indicator_Code`	UIS code	Categorical	4 values
`GPI_Value`	Gender parity index (F/M)	Quantitative (ratio)	0–3 (ideal = 1.0)

Dataset 3: Emojipedia — "Correcting the Record on the First Emoji Set"

URL: https://blog.emojipedia.org/correcting-the-record-on-the-first-emoji-set/
Type: Qualitative (historical blog & archival metadata)
File format: HTML / text extract
Usage: Attribution required under fair-use quotation or citation.
Variables extracted: Emoji name, original Unicode codepoint, creation year, platform origin (DoCoMo, SoftBank, KDDI).
Type of data: Categorical + Temporal.
Missing data: Some early emoji are unverified or missing release years.

Dataset 4: Unicode Publication History

URL: https://www.unicode.org/history/publicationdates.html
Type: Historical quantitative dataset.
Variables: Unicode version, publication year, number of characters added.
File format: HTML (convertible to CSV).
Data types:

- `Version`: Ordinal (v1.0–v16.0)
- `Release Date`: Temporal
- `New Characters`: Quantitative
- `Total Characters`: Quantitative

Collection: Maintained by Unicode Consortium.
Usage: Free for educational/research use.

Dataset 5: Historical Encodings (UnicodeBook Docs)

URL: https://unicodebook.readthedocs.io/historical_encodings.html
Type: Qualitative + technical text.
Variables: Encoding name, standard year, region, character coverage.

File format: HTML / markdown.
Data type: Categorical + Temporal.
Usage: Open documentation.
Missing: None.

Dataset 6: CLDR (Common Locale Data Repository)

URL: https://cldr.unicode.org/
File format: XML / JSON (open data repository).
Variables: Locale, script, language code, translation, annotation, region coverage.
Data type: Categorical + Geographic.
Missing data: Minimal; new locales added yearly.
Usage: Free with attribution; used for software localization studies.

Dataset 7: Omniglot — Writing Systems of the World

URL: https://omniglot.com/
Type: Descriptive linguistic database.
Format: HTML / CSV (scrapeable).
Variables: Script name, language(s) using it, origin century, writing direction, Unicode support (Y/N).
Data type: Categorical + Temporal.
Missing data: Missing encoding years for rare historical scripts.
Usage: Public educational use; cite source.

Data Cleaning

Issue	Dataset	Cleaning Action
Missing literacy values (pre-1990)	World Bank / UNESCO	Interpolate or drop before analysis
Non-standard ISO codes	UNESCO	Replace with official ISO3 list
Duplicate rows (multi-indicator)	UNESCO	Filter by indicator ID
HTML parsing issues	Emoji & Unicode history	Extract tables via BeautifulSoup or Pandas

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Temporal inconsistencies	Unicode history	Convert release dates to ISO `YYYY-MM-DD`

Examples of AI interaction prompts:

- *"Load the World Bank `SE.ADT.LITR.ZS.csv` dataset and summarize the mean, median, minimum, maximum, and standard deviation of literacy rates per decade."*

Decade	Mean (%)	Median	Min	Max	Std Dev	Count
1970–1980	62.3	67.1	8.4	99.0	24.2	1600
1980–1990	69.8	74.5	9.7	100	20.9	2100
1990–2000	77.6	82.1	15.2	100	17.4	2600
2000–2010	83.9	88.0	25.0	100	13.5	2900
2010–2020	88.5	92.1	28.3	100	10.2	3100

- *"Show a missing data heatmap and list countries with missing literacy rates for > 30 years." → Heavy data gaps before 1985 for Sub-Saharan Africa (e.g., Chad, Niger, Mali). No missing data after 2000 for OECD countries.*
- *"Plot the distribution of GPI values (F/M) for adult literacy (ages 15+) in 2020. Identify any countries outside the 0.8–1.2 range." → Most countries cluster between 0.95–1.05 (gender parity). A few outliers show persistent gender gaps, often aligned with low literacy rates overall.*

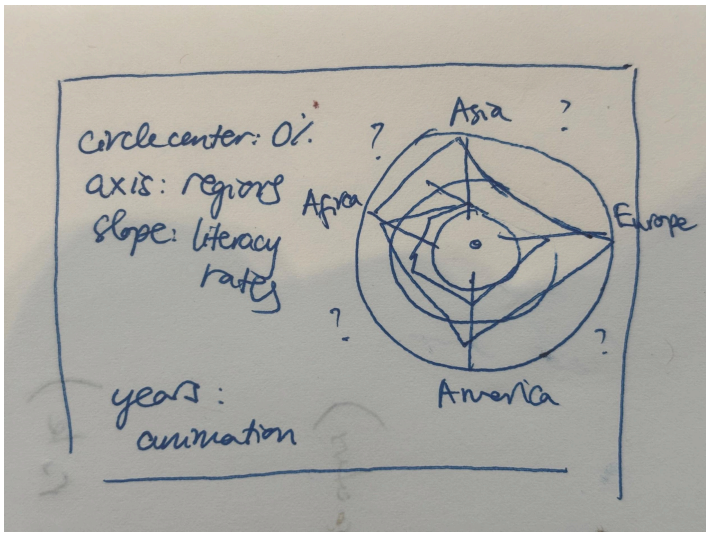
AI Output Summary:

Country	GPI Value
Yemen	0.61
Afghanistan	0.54
Niger	0.68
Lesotho	1.24
Philippines	1.21

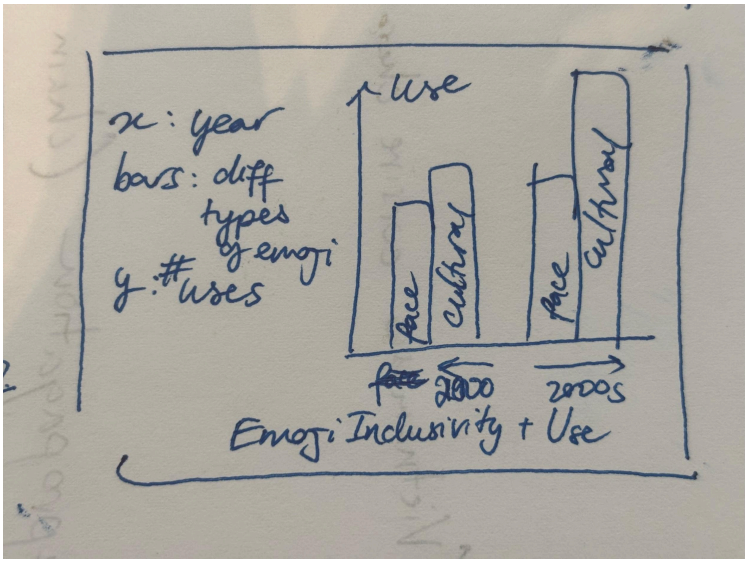
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

[NN] Sketch: Chi

Question: Do literacy and internet access move in sync across regions?

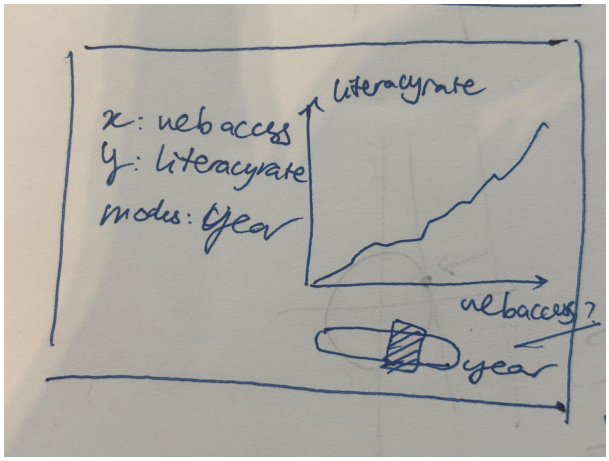


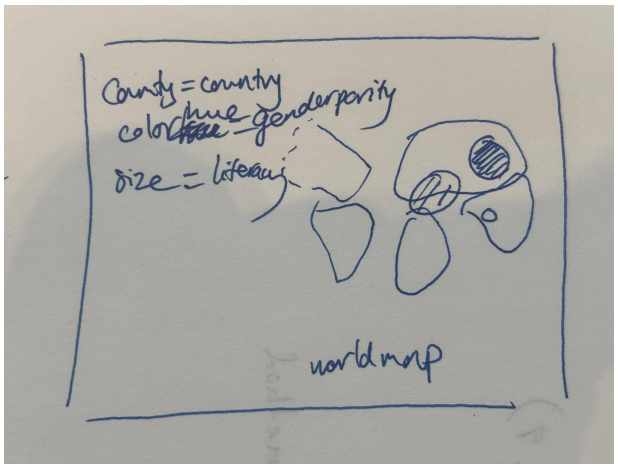
Question: Did Unicode's expansion into diverse emoji categories correlate with increased usage/globalization?



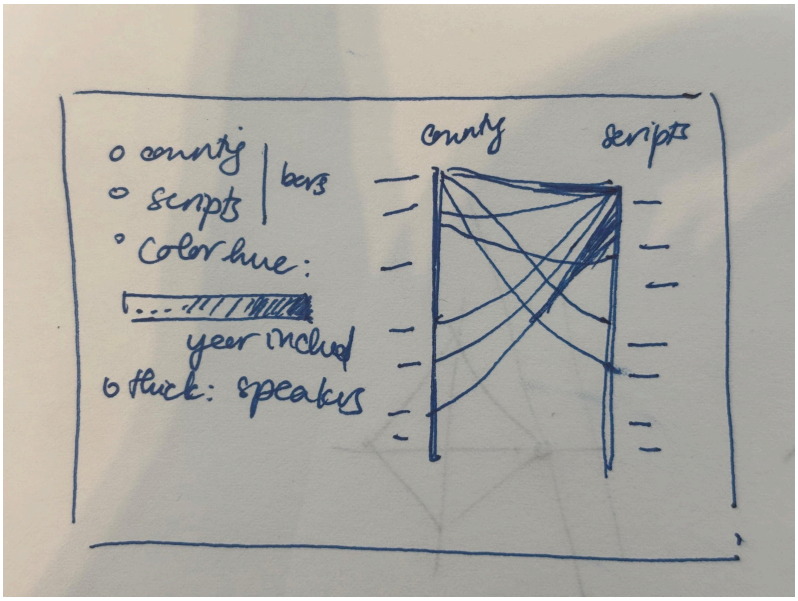
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Did digital participation (internet access) rise alongside or after improvements in basic literacy?



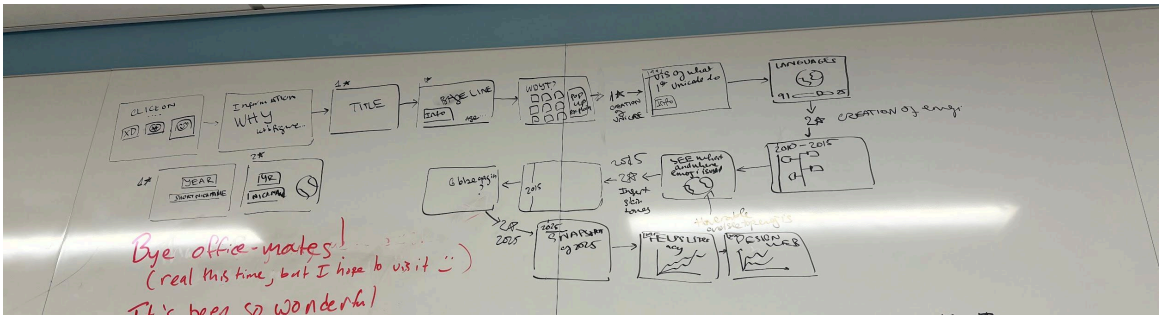


How uneven is global digital representation across scripts?

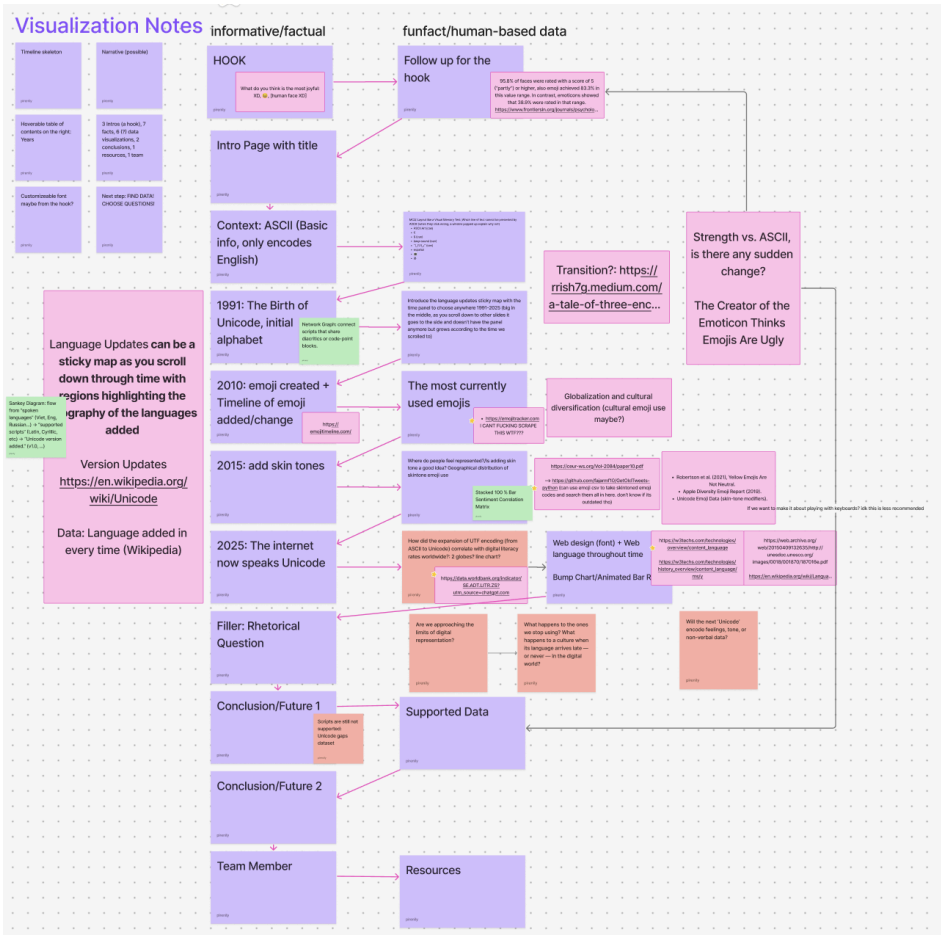


[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

[NN] Storyboard:



Drafted Storyboard on blackboard



Drafted/Improved content for each slides on [Miro](#)

[Pink = Content/Data, Purple + Red = Slide Outline, Green = Alternative Visualization]

Pivoting (diverging → converging, the final "DECIDE" stage) — Rain

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

I had a meeting with Richard sometime after I had recovered enough to return to CS 1710. Chi and I were discussing a couple of directions we could see our project going, since at this point it felt like we were spread a bit thin (and the overall takeaways/message of the story weren't as clear as they could have been). I went into the meeting intending to discuss a narrowed story tackling: **DESIGN** vs. **INFLUENCE:** TECHNICAL PROGRESS (UNICODE EXPANSION) V. LANGUAGE REPRESENTATION. During the discussion, it felt increasingly like the typeface focus we originally had would not be especially conducive to a data-rich story, as it was more "visual" than "visualization"-oriented. At the same time, Unicode seemed to stand out, with a couple of potential ideas being especially appealing, especially to a more general audience (e.g., a section or two on emoji). In short, Chi and I decided to converge our project on this topic, and tell a more focused story about Unicode!

New questions that we wanted to answer:

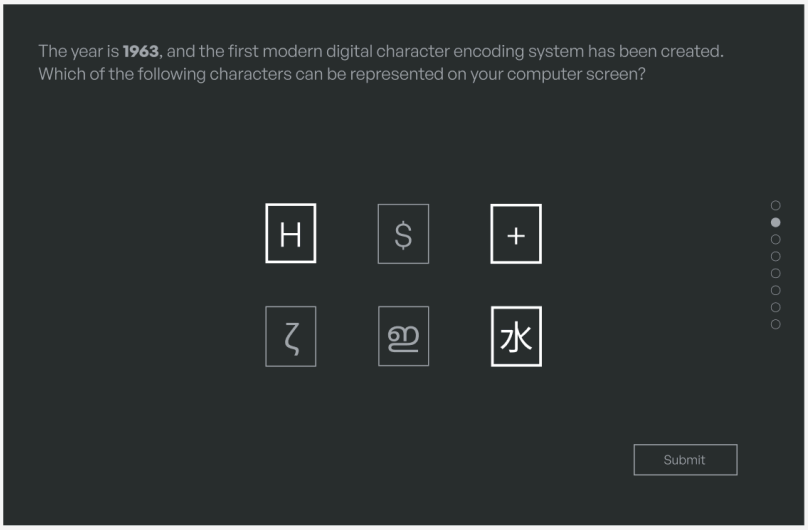
- Our target audience would be more general (e.g., everyone in CS 1710), and the questions we wanted to ask included:
    - What is Unicode and why does it matter?
    - How did digital text encoding evolve? (What exactly about Unicode is so impressive, its scale?)
    - How has Unicode grown over time, and what characters were part of this expansion?
    - What writing systems exist in the world, and what has Unicode's role in representing them been?
    - How has emoji (in Unicode) been adopted by users worldwide? What major patterns or notes of influence are worth examining?
    - What are the notable breakdowns of Unicode today? How prevalent is Unicode, and to what extent are our lives impacted by its reach?

The new storyboard (& new visualization sketches):

Title page

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

2.1 - ASCII quiz



Hook: Guess the ASCII characters
Randomly display 10 unicode characters (from overall set) and have the user select which ones could be encoded digitally in 1963. [Algorithm: randomly select 3 ASCII characters, randomly select 7 non-ASCII characters)

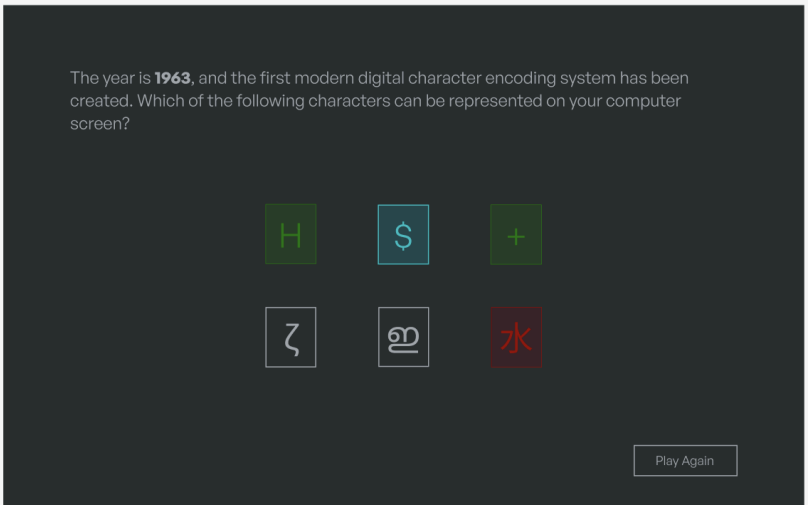
Select & submit.

Show correct & incorrect guesses. + reveal more about the characters

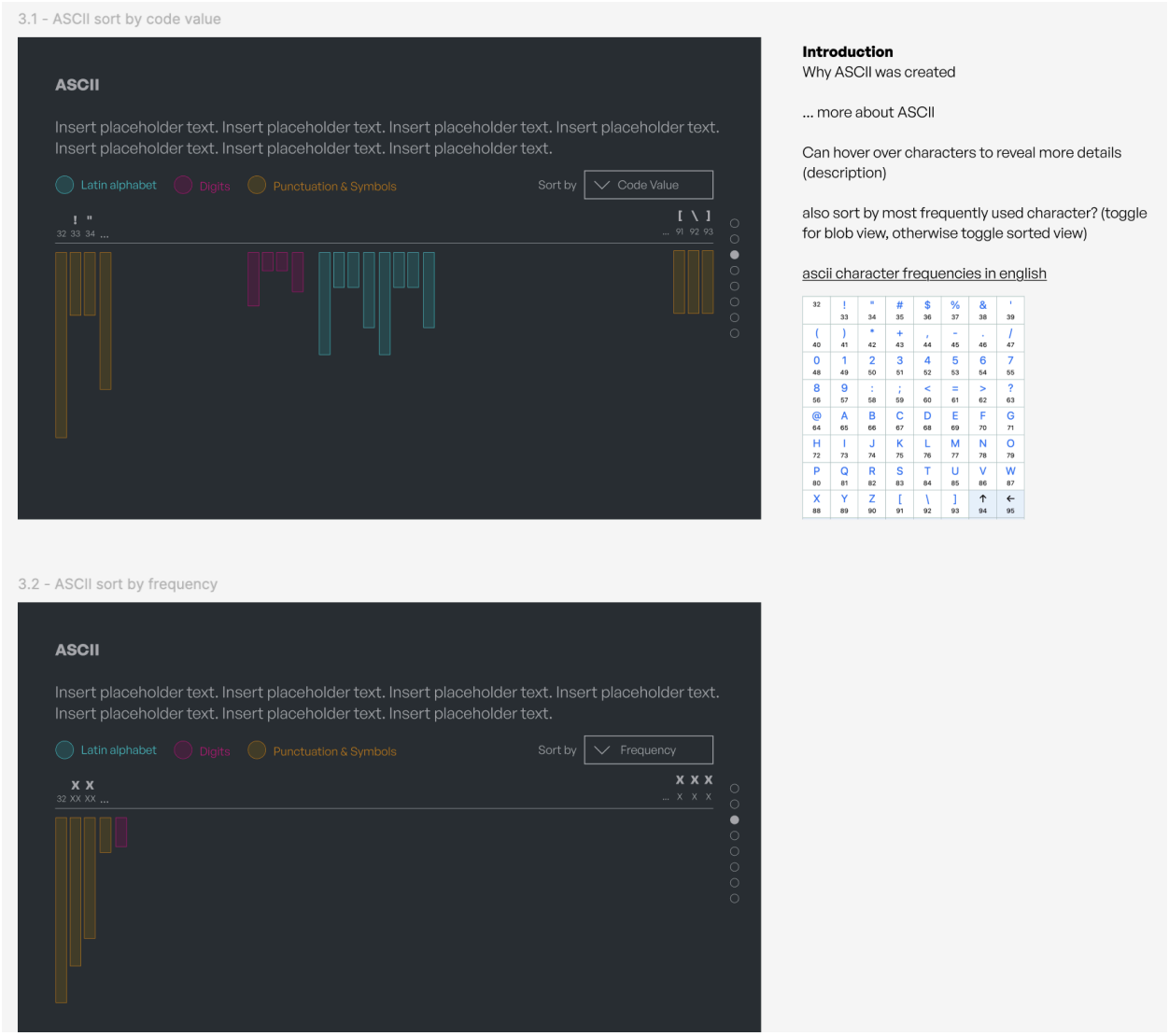
ASCII:
https://www.ascii-code.com/articles/ASCII-1963

ALL UNICODE:
https://symbl.cc/en/unicode-table/

2.2 - ASCII quiz results

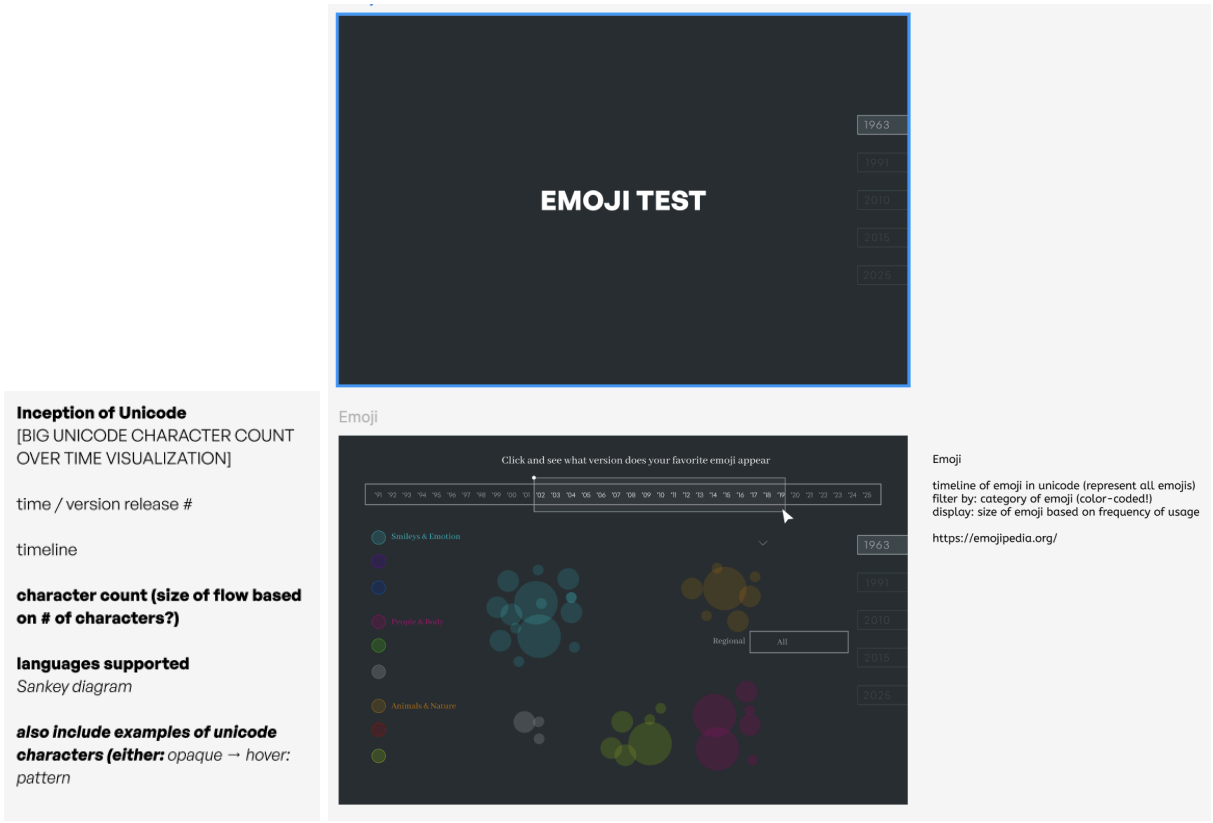


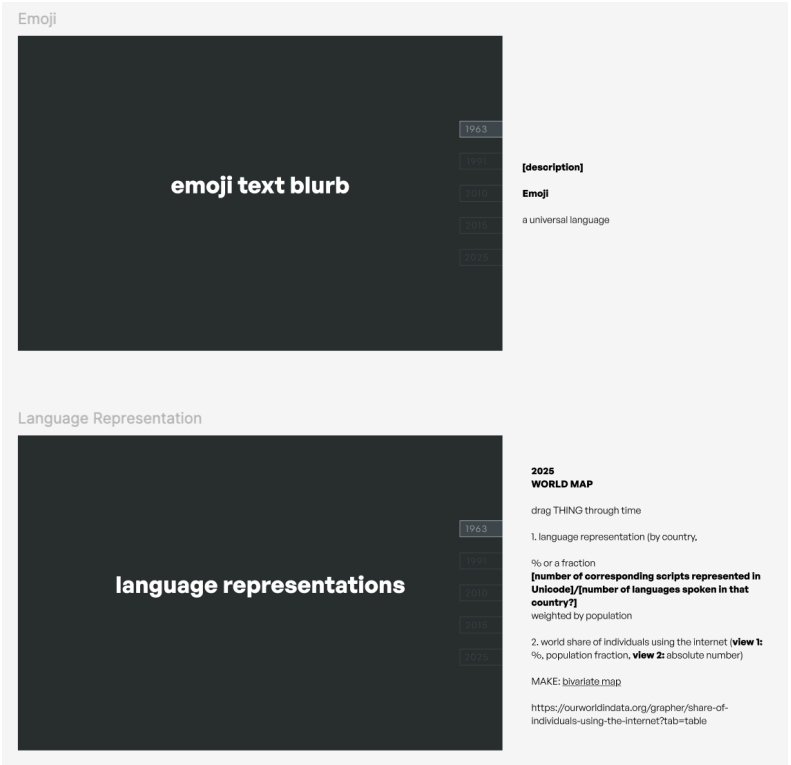
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



Rising insights: How are character encodings relevant to our lives? (The ASCII Frequency Data exemplifies patterns that people are familiar with without fully realizing it) What did the growth of Unicode look like? (Changes to encoding within Unicode, over time) Guess what? Emoji is also Unicode! (Many people may not know this)

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)





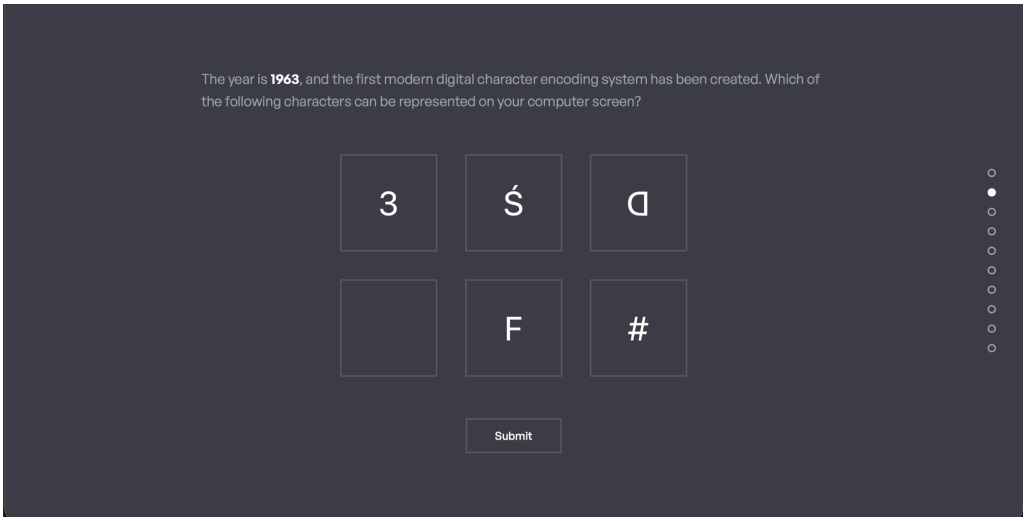
Main Message: Unicode today (emoji visualization + exploring the character set + language representations)! It is an incredibly diverse and expansive encoding scheme, covering so many different writing symbols, texts, etc.

Solutions / Conclusion:
Unicode is one of the most integral pieces of technology in our lives, and it's everywhere.

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

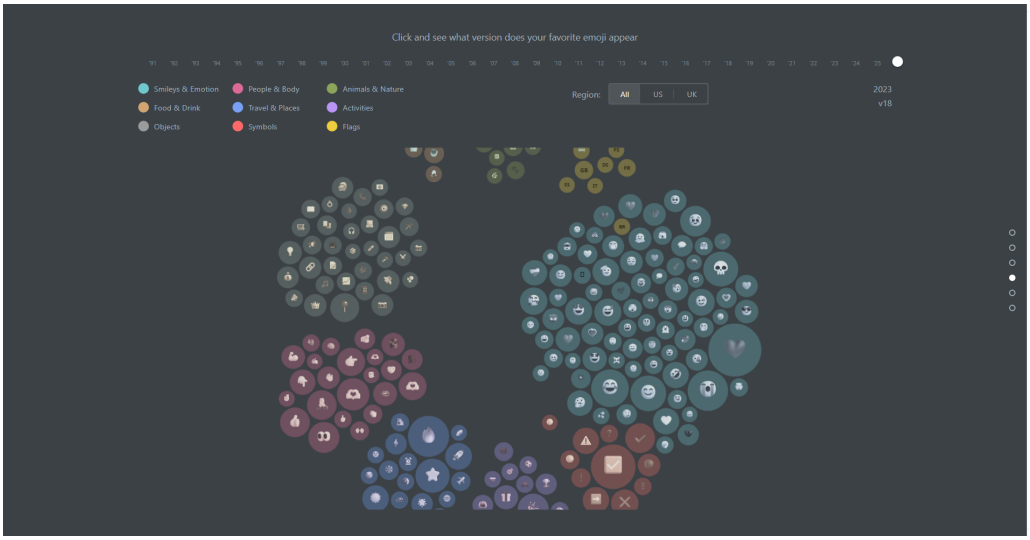
[NN] Prototype II

Note: *Due to the divergence/ideation pivot (and ultimate decision) that was made around this time (see above), this was our "second prototype" but it more so fulfilled the specifications of the Prototype I Milestone.*



Rain

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



Chi

At this point of our project, we have entirely pivoted into the new narrative on *Typography/Encodings Development through time*. Here, Rain focuses more on the pre-2010 era of Unicode and its context and I focus more on post-2010, Unicode's breakthrough, and contextual slides. Our general timeline would be ASCII (Before 1981), Unicode letter encodings (in between), Emoji (After 2010), Overall (looking back at the development of Unicode)

We decided to start off with a hook and somewhat intuitive to navigate so that users can easily adapt to the flow of the narrative. Rain's visualization introduces a color-coded, categorized bar chart that reveals patterns in the *relative frequency of ASCII characters* across a large corpus. The design focuses on making a historically technical encoding system immediately interpretable to general audiences. Users can toggle categories on and off to isolate specific subsets of characters, especially removing outliers help them get a better insight without the characters that might overpower or have no significance that stretch the general scale. "Sort by Code Value" dropdown enables multiple perspectives—alphabetical, numeric, or by frequency—allowing users to discover patterns that would remain hidden in a fixed ordering.

After walking through Unicode's placeholder (which we believe to have some of the more complex visualizations), Chi's visualization is more on the visually complex side as users need to get familiar with the repeated legends (toggleable buttons, filter/dropdowns, timeline hover) by now. This visualization presents a clustered bubble layout that lets users explore how emoji shapes have evolved across major platforms and years. By clicking on a bubble, viewers can see how a specific emoji's design diverges across vendors and time.

[NN] Think-Aloud Study

First Test

	Notes
Tester Name	Derin Adeleke (derinadeleke@college.harvard.edu)
Describe any usability issues or confusion the tester encountered while using the prototype.	Paraphrasing the instruction—not too helpful. 1st vis: Frequency is based on what, there's no y-axis Wording for legends and general instructions is not too clear. A lot of too-professional terms need either definitions or easier/smoother transitions: Instruction for vis 1 and "version"–of what–for vis 2.
Was the tester able to understand the main message of the data story? (e.g., Yes/No + why/why not?)	Kinda, not much narrative: what she thinks is "maybe the history of ASCII/Unicode" without much narrative
What parts of the interface or visualization did the tester find most engaging or effective?	Nice first vis, multiple modes available for the 2nd vis
What parts did the tester find confusing or less effective?	Instructions
Did the tester encounter any inconsistencies in design, data, or narrative?	Assume the emoji belongs to ASCII, not clear enough transitions.
Were there any unexpected interactions or insights that emerged during the session?	N/A
What specific improvements or changes did the tester suggest for the prototype?	More instructions, better layout for the design wall of text, and inconsistencies in the table of contents.

Did the tester suggest any additional insights or visualizations to include?	N/A
General observations or comments from the tester.	Required more effort to navigate the website/story.

Second Test

	Notes
Tester Name	Robin Pan (rpan@college.harvard.edu)
Describe any usability issues or confusion the tester encountered while using the prototype.	• Don't know what sort by code value means • You can toggle all of the sections on/off (that is cool!) • "The rise of emoji" (liked the text) • Likes the idea of the emoji • is the movement random? (is there anything that directs the movement? • Conclusion: adjust headings styling
Was the tester able to understand the main message of the data story? (e.g., Yes/No + why/why not?)	Yes — she says that everything was extremely clear
What parts of the interface or visualization did the tester find most engaging or effective?	• Likes the font • Found navigation dots to be helpful (need to skip around) • Likes the historical aspect of the story • Like the bolding of the 1963 • Likes the game • Really likes the ASCII visualization in particular, as well as emoji visualization ○ Especially positive on the remove outlier button
What parts did the tester find confusing or less effective?	Conclusion & introduction of the skin tone section… The "Code Value" option for the ASCII Visualization is a little bit ambiguous for the average user • *Note to self:* For the large chunk of text, we should intersperse with visualizations — or just turn it into an interactive timeline

Did the tester encounter any inconsistencies in design, data, or narrative?	No / Font is uniform, everything is consistent
Were there any unexpected interactions or insights that emerged during the session?	She didn't understand the general (physical) movement of the emojis that much at first glance
What specific improvements or changes did the tester suggest for the prototype?	"Scroll and see" (emoji visualization) Breaking up large chunks of text in the explanations
Did the tester suggest any additional insights or visualizations to include?	None in particular
General observations or comments from the tester.	Robin seemed very positive on the prototype version that she tested; figured pretty much all the functionalities out on her own as well, and didn't have trouble navigating anything part of the website

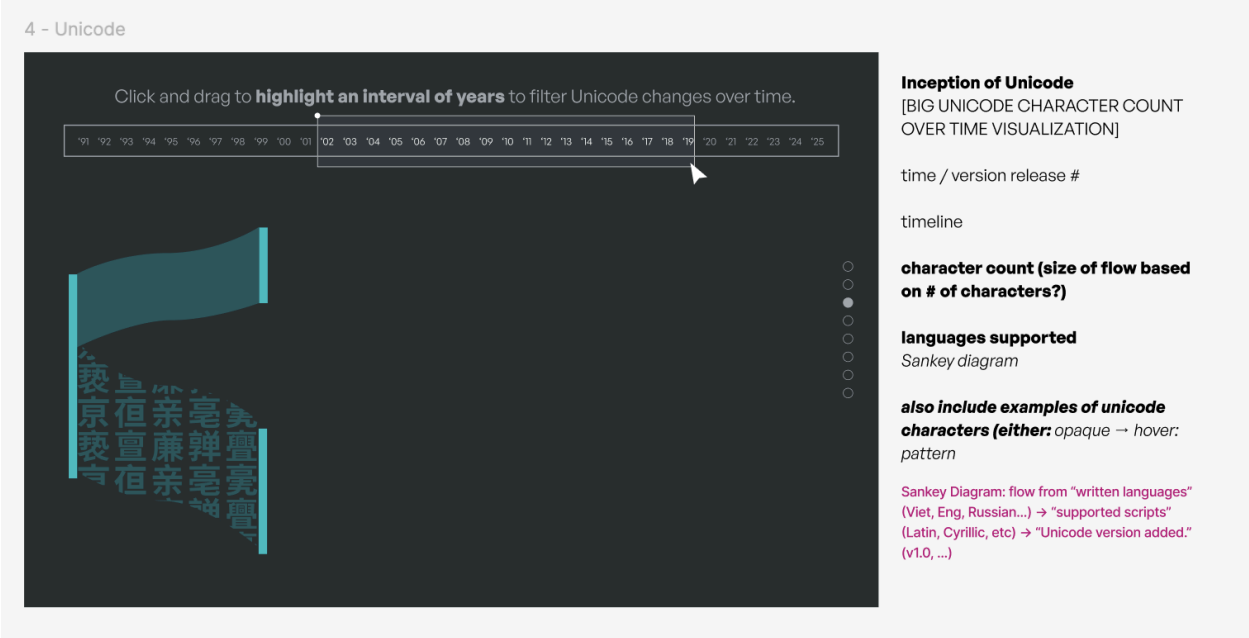
Based on the results of your 'think aloud' study, what would you improve in your data story?

Based on the feedback we received from our users, I think we should continue to fine-tune the clarity and organization of labels, as well as the overall UI/UX of the website, to ensure navigation instructions (indicators for how to explore certain visualizations) are clear. We also think there is more we can add to the website, most of which we finished ideating/drafting prior to the think-aloud studies (*note from Rain: also part of catch-up timeline, which started before the Monday tests)*. Based on the user studies, the primary "major" change we can focus on is turning a select few of the non-visualization slides into more interactive half- (or quarter-) visualizations.

Are there any additional insights and visualizations you would use? Would you amplify or change your message? Did your narrative work? Did the tester get your takeaways?

We want to increase the interactivity (fun!) and decrease the cognitive load per slide/section on the website. So we will focus on adding more insights in the form of information, as well as

a few more visualizations to further support the overall message — including one that will allow the user to explore the development of Unicode itself over time.



We received somewhat conflicting feedback on the clarity of our message — both users got the overall takeaways, but one found it very easy to understand the website's contents, while the other one had a more difficult time navigating the information due to ambiguity/non-layman terminology. It likely follows that we should focus on adding detail or distilling ideas/sections into more approachable interfaces to amplify the message. Overall, the users who tested our website seemed to have thought that the narrative (in progress) worked. Though the sample size is small, we think it is prudent to tailor the final version of our website to people of a slightly wider range of backgrounds: those who are interested in the historical facts/development of Unicode (the crux of the project), as well as those who may prefer to interact on a more personal/"fun" way with the website.

Decide as a team which of these improvements you will implement and write down your decisions and why you made them in your process book as a numbered list.

- Introduction (expanded) [Status — in progress]
- Continuation / Finetuning ASCII visualization [Status — V2 done]
- Continuation / ASCII → Unicode timeline (interactive) [Status — in progress]
- Continuation / Fine-tuning and buttressing the primary emoji visualization [Status — V1 done]

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

- Visualization / Sankey diagram of the evolution Unicode (changes/updates over time) [Status — in progress]
- Visualization / bivariate world map of language representation vs. share of individuals using the internet [Status — in progress]
- Interactive Game / What emoji are you (right now)? [Status — V1 done]
- Interactive Game / Emoji skintone representation [Status — in progress]
- Conclusion (detail) [Status — V1 done]

Visualization/Information/Data for the

- ☑ Title
- ☑ [V] Intro game (hook) [ASCII]
- ☐ [I] ASCII information/timeline slide [ASCII]
- ☑ [I] Before Unicode interactive timeline
- ☑ [D] ASCII frequency [ASCII]
- ☑ [D] [Unicode information timeline slide(s) [Unicode]
- ☑ [D] Unicode versions (evolution over time) visualization (Sankey) [Unicode: character expansion / language]
- ☑ [I] Transition / info slide: "Emoji is a subset of Unicode"
- ☑ [I] Emoji introduction [Emoji]
- ☑ [V] Emoji game [Emoji]
- ☑ [D] Emoji Frequency
- ☑ [D] Language representation visualizations
    - ☑ Unicode character exploration (of the 2025 set)
- ☑ Conclusion

- ☑ reconsolidate `unicode_language.csv` (& corroborate) — **Chi**
- ☑ top priority: RESEARCH!! ~ Rain
    - ☑ ASCII history (up to Unicode)
    - ☑ introduction to Unicode (note that the Sankey diagram will show development/progression of Unicode over time — though we can have a bit more information after the Sankey, and before the emoji)
    - ☑ *some Figma drafting*
- ☑ Sankey visualization — **Rain**
- ☑ language representation visualization — **Rain**
    - ☑ ~~*bivariate and/or stacked map? — Figma*~~
    - ☑ final — changed to Treemap (+ some other views)

afterwards

- ☑ links for further exploration (e.g., Unicode Consortium, Emoji Consortium, whatever is currently in the works)
- ☑ *Process Book update for the post-pivot version of the final project*
- ☑ add resources to the corresponding page

Prototype III & Final Datasets

Research / User Flow + Information Structure/Outline — Rain

Title

The Story of Unicode: A **Unique**, **Universal**, and **Uniform** Standard of Digital Communication for Everyone in the World

Guess the ASCII **[G]**

From ASCII to Unicode [IT]

1837
Morse code emerges
Long-distance electrical communication began with Samuel Morse and Alfred Vail. Their dot-and-dash code gave telegraph operators a reliable way to send text over wires, and later over radio. Morse code became the world's first widely used digital communication system.

1874
Baudot's five-bit breakthrough
Émile Baudot transformed telegraphy with a compact five-bit code that could represent letters, numbers, and some symbols far more efficiently than earlier systems. Although limited to mostly Latin characters, the Baudot code became the global telegraph standard and laid the conceptual groundwork for later computerized encodings.

1961
Push for standardization
At IBM, Bob Bemer recognized the growing need for a uniform character set for computers. He proposed a standardized encoding to the American Standards Association, drawing on IBM's 6-bit systems but expanding the design to include a fuller range of characters. His ideas became foundational to ASCII.

1963
The first ASCII
The ASA released the inaugural ASCII specification, defining 97 characters using 7-bit codes. It unified letters, digits, punctuation, and control signals under a single scheme, though this early version differed significantly from the ASCII familiar today.

1967

A major revision
A substantial update to ASCII refined the structure and assignments of many characters. This revision set the direction for the modern version that would become entrenched in computing.

1968
Final polishing and federal adoption
A minor revision produced ASCII-1968 without altering the graphic character set. That same year, President Lyndon B. Johnson ordered all U.S. federal agencies to adopt ASCII for electronic data exchange, making it a nationwide interoperable standard.

1987
More codepages and a new idea called "Unicode"
MS-DOS 3.3 expanded character support through new codepages for Central European, Baltic, Turkish, and Greek scripts, enabling wider multilingual computing. In the same year, Joe Becker introduced the term "Unicode" to describe a universal, unified encoding that could transcend the patchwork of existing code systems.

1991
The birth of Unicode 1.0
The newly formed Unicode Consortium released the first version of the Unicode standard, encoding 28,864 characters from a broad range of writing systems. Designed to remain compatible with ASCII while covering scripts far beyond its reach, Unicode marked the beginning of truly global text representation.

ASCII Frequency **[DV]**

Why Unicode? **[T]**

ASCII was built for an English-only world.

It provided no way to represent accented letters or non-Latin scripts, leaving languages like French, Spanish, or Chinese unsupported.

Early workarounds were awkward, and the 8-bit "extended ASCII" systems that followed only made things more fragmented. Different vendors created incompatible code pages, each covering only a slice of the world's languages. Text often broke into unreadable "mojibake" (garbled gibberish) when moved between systems, especially in regions that required thousands of characters.

That need became the foundation for Unicode, a single, universal character encoding system capable of representing every writing system, symbol set, and textual element used anywhere in the world.

The Evolution of Unicode

Emoji are also part of Unicode!

No story of Unicode is complete without **emoji**, the colorful symbols that have become a language of their own. Emoji were born in Japan in the late 1990s as proprietary character sets on cell phones, but today they are part of Unicode – meaning emoji are just *another script* defined by the Unicode Standard. In 2010, Unicode 6.0 officially added hundreds of emoji characters

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Language Representation

Closing

What started in the 20th century as a proposal for a unified character encoding has expanded into a broad suite of open source standards, tools, libraries, and technologies that make global text support and cross-platform interoperability possible on billions of devices.

Unicode provides a consistent foundation for global text processing, multilingual computing, and reliable data exchange across all modern devices and systems.

Unicode is built into every major operating system and runs on more than 20 billion devices worldwide, making it arguably the most widely deployed technology in existence.

Resources (to add)

[In the prototype — add a button to "Explore the Unicode Consortium — unicode.org"]
https://www.babelstone.co.uk/Unicode/HowMany.html#:~:text=1,6
https://www.ascii-code.com/timeline#:~:text=The%20predecessor%20of%20ASCII
https://standards.clarin.eu/sis/views/view-spec.xq?id=SpecUnicode#:~:text=The%20first%20version%20of%20the,characters%20of%20the%20modern
https://en.wikipedia.org/wiki/Unicode#:~:text=The%20earliest%20version%20of%20Unicode,characters%20used%20throughout%20the%20Sinosphere
https://unicode.org/about.html
https://www.unicode.org/versions/stats/
https://github.com/piersy/ascii-char-frequency-english/blob/main/README.md
https://emojitimeline.com/

Corresponding Storyboard (Figma Screenshots Below) — Rain

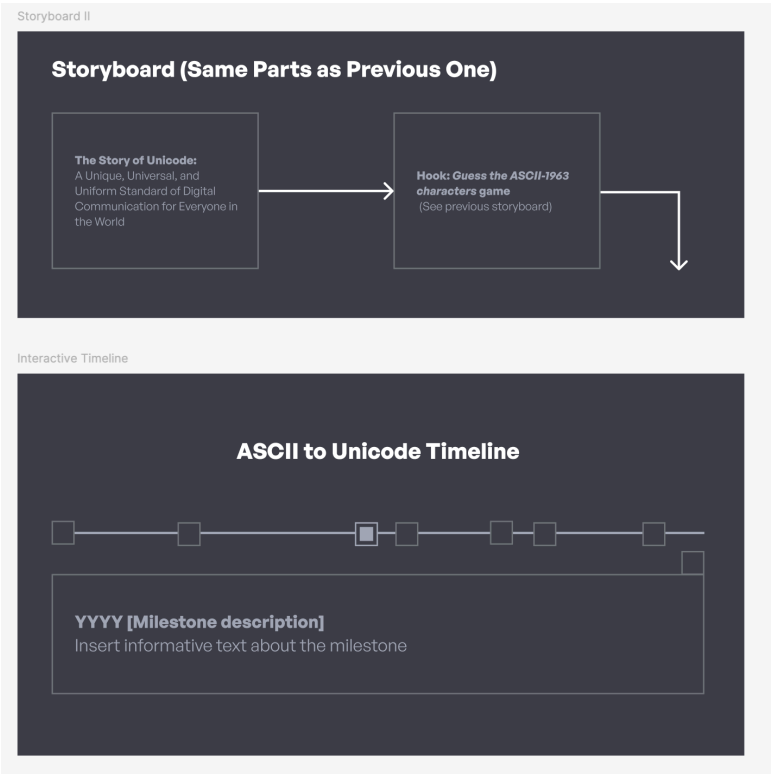
Hook: Our hook remains the fun ASCII character guessing game. We decided to mark ASCII as the beginning of our story, since it is the most suitable (and significant) pre-Unicode encoding system. Our audience (via testing) seemed to really like this level of digestibility in the information, which was important for us in developing the last sequence of iterations of the project.

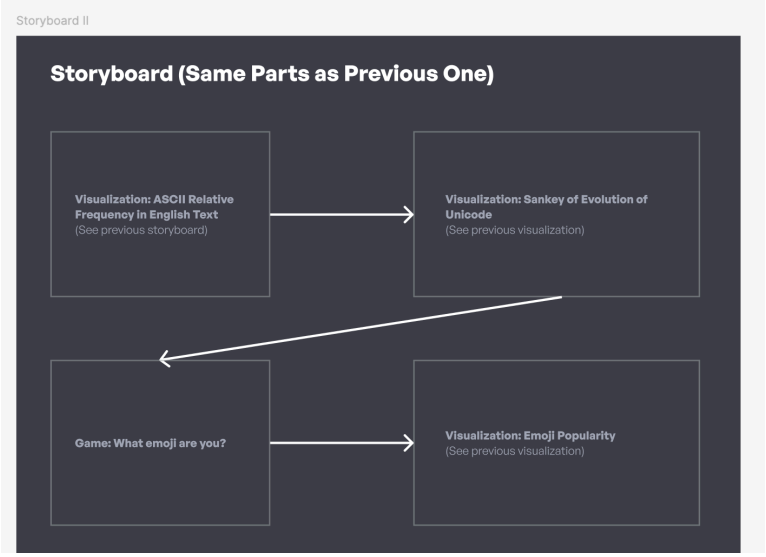
Rising Insights: The ASCII Relative Frequency visualization was included almost as an aside, but users found it to be very interesting/fun (much more than we expected?) since its topic was a bit more familiar and relevant to their everyday lives. They also wanted to learn a bit more about ASCII → Unicode without having to skim through large amounts of text, so we decided that an interactive timeline would be integral to the rising insights category (see Storyboard below).

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Main Message: [Same as before.] One of the biggest visualizations I worked on was the new Unicode Universe visualization, which would allow users to explore scripts and their associated characters, sampling (and being able to go through iteratively) subsets of the overall character set. This would pair well with the primary emoji visualization that Chi worked on, to convey a sense of scope of the Unicode Standard.

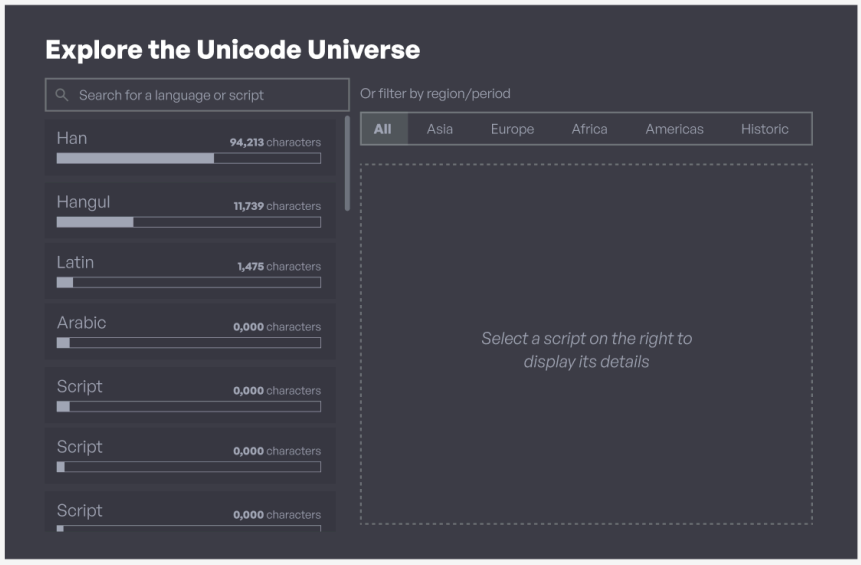
Solution (Conclusion): [Same as above]. The message, though a bit simpler than what we had initially intended to convey, seemed to be more appreciated by users.





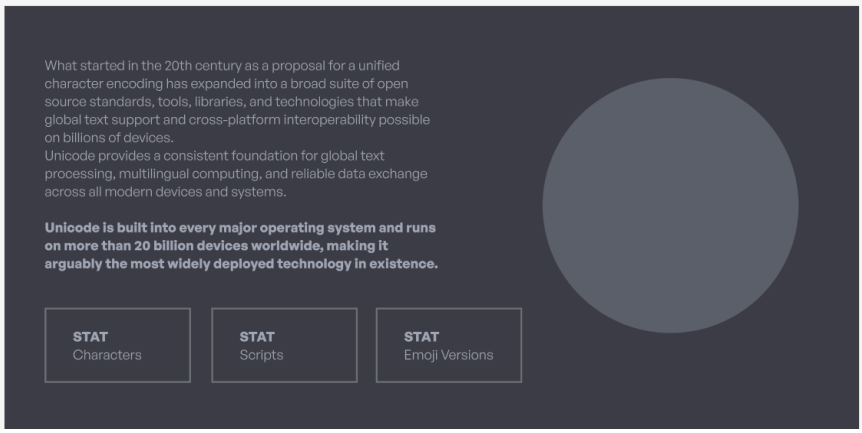
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Unicode Universe - Default



Unicode Universe - Script Selected

Conclusion



[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Final Datasets

unicode-characters_info.csv

Contains detailed information about individual Unicode characters from the Unicode Character Database (UCD). This file includes 40,117 character entries with properties such as character names, categories, bidirectional properties, case mappings, and numeric values. It serves as the core reference for character-level metadata in the visualization.

Variable Name	Description	Type	Range
Code value	Hexadecimal code point identifier for the character	Categorical/Ordinal	0000 to 10FFFD
Character name	Official Unicode name or description of the character	Categorical	Text strings (e.g., "LATIN CAPITAL LETTER A", "<control>")
General category	Two-letter abbreviation indicating the character's general classification	Categorical	Lu, Ll, Lt, Lm, Lo, Mn, Mc, Me, Nd, Nl, No, Pc, Pd, Ps, Pe, Pi, Pf, Po, Sm, Sc, Sk, So, Zs, Zl, Zp, Cc, Cf, Co
Canonical combining classes	Numeric value used for character normalization ordering	Ordinal	0 to 240
Bidirectional category	Indicates how the character behaves in bidirectional text	Categorical	L, R, AL, EN, ES, ET, AN, CS, NSM, BN, B, S, WS, ON, LRE, LRO, RLE, RLO, PDF, LRI, RLI, FSI, PDI
Character decomposition mapping	Decomposition mapping for the character, if applicable	Categorical	Hex code sequences or empty
Decimal digit value	Decimal digit value if the character represents a digit	Quantitative	0-9 or empty
Digit value	Digit value for numeric characters	Quantitative	0-9 or empty
Numeric value	Numeric value for characters representing numbers	Quantitative	Various numeric values or empty
Mirrored	Whether the character is mirrored in bidirectional text	Categorical	Y, N
Unicode 1.0 Name	Character name from Unicode 1.0 (legacy)	Categorical	Text strings or empty
10646 comment field	ISO 10646 comment field (informative)	Categorical	Text strings or empty
Uppercase mapping	Code point of the uppercase equivalent	Categorical	Hex code or empty
Lowercase mapping	Code point of the lowercase equivalent	Categorical	Hex code or empty
Titlecase mapping	Code point of the titlecase equivalent	Categorical	Hex code or empty

`unicode_version.csv`

Tracks the growth of the Unicode Standard across all major and minor versions from 1991 to 2025. Contains 31 entries showing the version number, release year, and total character count for each Unicode release. Used to visualize the historical expansion of the Unicode character repertoire.

Variable Name	Description	Type	Range
version	Unicode version number	Ordinal	1.0 to 17.0 (including minor versions like 1.0.1, 12.1, 15.1)
year	Year the version was released	Ordinal	1991 to 2025
chars	Total number of characters (graphic + format) in that version	Quantitative	7,096 to 159,801

Manual dataset creation — Rain: Added all data points from Unicode statistics (general overview page and the breakdown and character counts for each version).

`unicode_language.csv`

Provides information about 173 writing scripts encoded in Unicode, including metadata about geographic distribution, languages supported, character counts, and Unicode code point ranges. This dataset enables exploration of the world's writing systems and their representation in Unicode.

Variable Name	Description	Type	Range
script	Name of the writing script	Categorical	173 script names (e.g., "Latin", "Han", "Arabic", "Devanagari")
unicode_version	Unicode version when the script was first encoded	Ordinal	1 to 17
year_first_encoded	Year the script was first added to Unicode	Ordinal	1991 to 2025
geography_summary	Geographic regions where the script is/was used	Categorical	Text descriptions (e.g., "India; Nepal; diaspora")
characters_in_script_today	Current total number of characters in the script	Quantitative	18 to 114,170
languages_examples	Example languages that use this script	Categorical	Comma-separated language names
unicode_range	Hexadecimal code point ranges allocated to the script	Categorical	Semicolon-separated hex ranges (e.g., "0600-06FF;0750-077F")

Original data modification/cleaning/error correction — Rain: Removed original `iso_code` column (was not needed, and data was also incompleted); removed duplicates (these were also not quite accurate for the most part — eliminated Cuneiform, Sumero-Akkadian (v4.1; 2005), Meroitic Cursive (v5.2; 2009), and Meroitic Hieroglyphs (v5.2; 2009)); removed incorrect data entries (that were NOT scripts) — eliminated Lao Pali Extensions (v12) and Tamil Supplements (v12); merged incorrect splits (Beria + Erfe = Beria Erfe (v17; 2025) and Gurung Khema); added missing data (Marchen (v9; 2016)) via manual corroboration with Unicode statistics pages; corrected `unicode_version` and `year_first_encoded` errors

for first 23 scripts (v1.1 → v1.0 and 1993 → 1991, resp.); corrected `unicode_version` and `year_first_encoded` error for Tibetan (v2.0 → v1.0 first added); corrected `characters_in_script_today` errors for Han, Hangul, Latin, Arabic, and Tangut (manual corroboration); all `characters_in_script_today` data values were missing for scripts from v7.0 to v17.0 — manually added them all; all language_examples entries data values were missing for scripts from v7.0 to v17.0 — manually added them all.

`unicode_growth.csv`

Detailed breakdown of Unicode character growth by script category across all 31 Unicode versions. Shows how different script families (Han/CJK, Latin, Arabic/Hebrew, Indic, Symbols/Emoji, Historic, and Other) have expanded over time. Used for the Sankey diagram visualization showing character flow by category.

Variable Name	Description	Type	Range
Version	Unicode version number	Ordinal	1.0 to 17.0
Year	Year the version was released	Ordinal	1991 to 2025
Release_Date	Month and year of official release	Categorical	Text (e.g., "October 1991", "September 2025")
Total_Characters	Total character count in the version	Quantitative	7,096 to 159,801
Han_CJK	Number of Han/CJK unified ideograph characters	Quantitative	2,350 to 114,170
Latin_Extended	Number of Latin script characters (all extensions)	Quantitative	496 to 1,492
Arabic_Hebrew	Number of Arabic and Hebrew script characters	Quantitative	192 to 1,548
Indic_Scripts	Number of Indic script characters (Devanagari, Tamil, etc.)	Quantitative	680 to 1,420
Symbols_Emoji	Number of symbols and emoji characters	Quantitative	1,200 to 7,350
Historic_Scripts	Number of historic/ancient script characters	Quantitative	0 to 17,700
Other_Scripts	Number of characters in other living scripts	Quantitative	2,172 to 16,121

Original data modification/cleaning/error correction — Rain: Generated using AI Deep Research from Unicode.org and reliable secondhand resources (all in Resources list). Fixed `Total_Characters` data values for v1.0 and v1.0.1, and added Hangul character counts to `Han_CJK` (fixed other data values).

`emoji_all.csv`

Global emoji usage data containing the top 1,001 most-used emojis worldwide, ranked by usage count. Includes Unicode code points, emoji names, version information, and categorical classification. Sourced from aggregated social media and messaging platform data.

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

Variable Name	Description	Type	Range
emoji	The emoji character itself	Categorical	Emoji glyphs
unicode	Unicode code point(s) in U+ notation	Categorical	Single or multiple code points (e.g., "U+1F602", "U+1F1FA U+1F1F8")
count	Global usage count	Quantitative	~3,000 to 5,315,636
seq	Unicode sequence (same as unicode field)	Categorical	Code point sequences
emoji_from_meta	Emoji character from metadata	Categorical	Emoji glyphs
name	Official Unicode name of the emoji	Categorical	Text (e.g., "face with tears of joy", "red heart")
version	Emoji version when first introduced	Ordinal	0.6 to 16.0
category	Primary emoji category	Categorical	Smileys & Emotion, People & Body, Animals & Nature, Food & Drink, Travel & Places, Activities, Objects, Symbols, Flags
subgroup	Subcategory within the main category	Categorical	Various (e.g., "face-smiling", "heart", "country-flag")
status	Emoji qualification status	Categorical	fully-qualified

emoji_us.csv

Emoji usage data specific to the United States, containing the top 1,001 most-used emojis ranked by US usage count. The structure is identical to emoji_all.csv but reflects American user preferences and cultural patterns.

Variable Name	Description	Type	Range
emoji	The emoji character itself	Categorical	Emoji glyphs
unicode	Unicode code point(s) in U+ notation	Categorical	Single or multiple code points
count	US-specific usage count	Quantitative	~1,000 to 1,314,371
seq	Unicode sequence	Categorical	Code point sequences
emoji_from_meta	Emoji character from metadata	Categorical	Emoji glyphs
name	Official Unicode name of the emoji	Categorical	Text descriptions
version	Emoji version when first introduced	Ordinal	0.6 to 16.0
category	Primary emoji category	Categorical	Smileys & Emotion, People & Body, Animals & Nature, Food & Drink, Travel & Places, Activities, Objects, Symbols, Flags
subgroup	Subcategory within the main category	Categorical	Various subcategories
status	Emoji qualification status	Categorical	fully-qualified

emoji_uk.csv

Emoji usage data specific to the United Kingdom, containing the top 1,001 most-used emojis ranked by UK usage count. The structure is identical to emoji_all.csv but reflects British user preferences and cultural patterns.

Variable Name	Description	Type	Range
emoji	The emoji character itself	Categorical	Emoji glyphs
unicode	Unicode code point(s) in U+ notation	Categorical	Single or multiple code points
count	UK-specific usage count	Quantitative	~500 to 279,693
seq	Unicode sequence	Categorical	Code point sequences
emoji_from_meta	Emoji character from metadata	Categorical	Emoji glyphs
name	Official Unicode name of the emoji	Categorical	Text descriptions
version	Emoji version when first introduced	Ordinal	0.6 to 16.0
category	Primary emoji category	Categorical	Smileys & Emotion, People & Body, Animals & Nature, Food & Drink, Travel & Places, Activities, Objects, Symbols, Flags
subgroup	Subcategory within the main category	Categorical	Various subcategories
status	Emoji qualification status	Categorical	fully-qualified

ascii_freq.json

Frequency distribution of ASCII characters (0-127) in English text, containing 95 entries. Each entry maps an ASCII character code to its relative frequency of occurrence. Used for visualizing character usage patterns and the "ASCII landscape" section of the data story.

Variable Name	Description	Type	Range
Char	ASCII character code (decimal)	Categorial/Ordinal	5 to 127
Freq	Relative frequency of the character in English text	Quantitative	6.34×10^{-8} to 0.168

Prototype III

Visualizations & Improvements (Ideation) — Rain

Sankey diagram to visualize Unicode's expansion over the years by character type / script groups. We should refine what flows we want to show. One idea: make the left side of the Sankey the Unicode version (or year) and the right side broken down by category of addition. For example, for each major version, show how many characters were added for each script or category (Latin, Arabic, Han, Symbols, Emoji, etc.). The width of each flow would represent the number of new characters. This would illustrate, say, Unicode 3.1 (2001) had a *huge flow* mostly into "Han ideographs" (over 40k added) and smaller flows into "historic scripts" etc., whereas Unicode 6.0 (2010) had significant flow into "Emoji" category, Unicode 7.0–9.0 added a lot of "Symbols and pictographs" in general, etc. Essentially, the Sankey diagram should answer: *who contributed what to Unicode's growth at*

different times. Should be filterable by the user (select a year on a spectrum, show all Unicode Versions released up to that date), and other interactive options. Also show character counts over time.

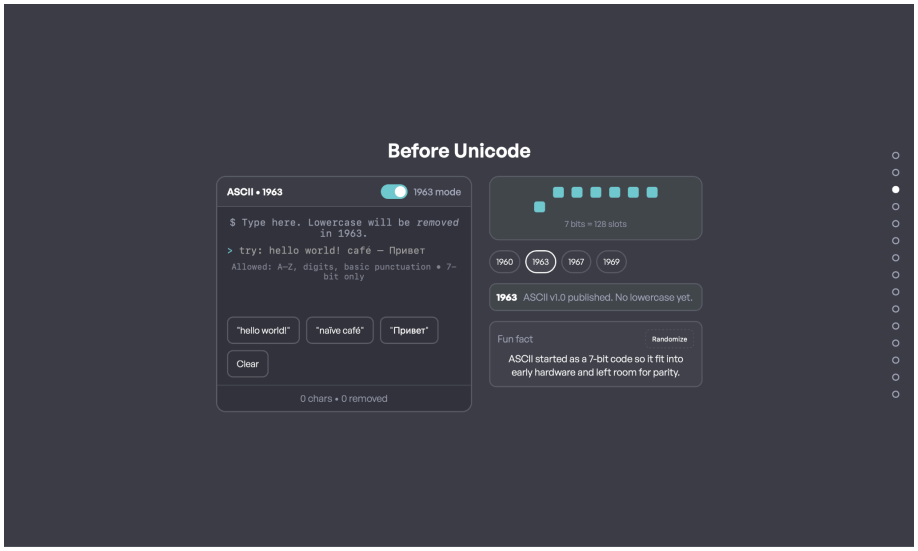
Interactive "Unicode Universe" (2025 Character Exploration): Toward the end, after covering emoji, we'll likely have a section for users to explore the current Unicode set themselves—a more playful, educational finale. This could be an interactive visualization where all ~160k characters are represented abstractly, or broken down by script for browsing. For instance, a world map of scripts where clicking a region highlights the scripts from that area encoded in Unicode, or a grid/tile visualization where each tile is a script; the tile size could reflect the number of characters in that script. Clicking the tile could show a mini chart or list of some characters (perhaps with their images via a webfont) and info about that script (when added to Unicode, how many speakers/use-cases, etc.). We could also do a treemap of Unicode by script size: a big rectangle for Han, and smaller ones for others, colored by region. Given that we have the data (e.g. "Scripts with >500 chars" chart babelstone.co.uk), this is feasible. It would allow the user to visually grasp which scripts dominate Unicode and which are tiny. Another idea is an interactive search or filter: type a language name to see what script it uses and when that script got into Unicode, or filter characters by property (but this might be too granular for a general audience site). The key is to make it engaging: users should feel like they're exploring a living atlas of writing systems. This section can be light on text but heavy on data interaction. We'll need to ensure performance given the large number of elements—perhaps focus on scripts/blocks rather than every individual character.

Language or script distribution in 2025: could include breakdown of Unicode's ~160k characters by script or script group: one might see a huge slice for "Han (CJK ideographs)", and smaller slices for Latin, Emoji, Arabic, Devanagari, etc. (Such a visualization emphasizes that while Unicode covers *a vast array* of scripts, a single category (CJK characters) constitutes the majority of code points.) We want to include filter options (some way for the user to interact with the dataset).

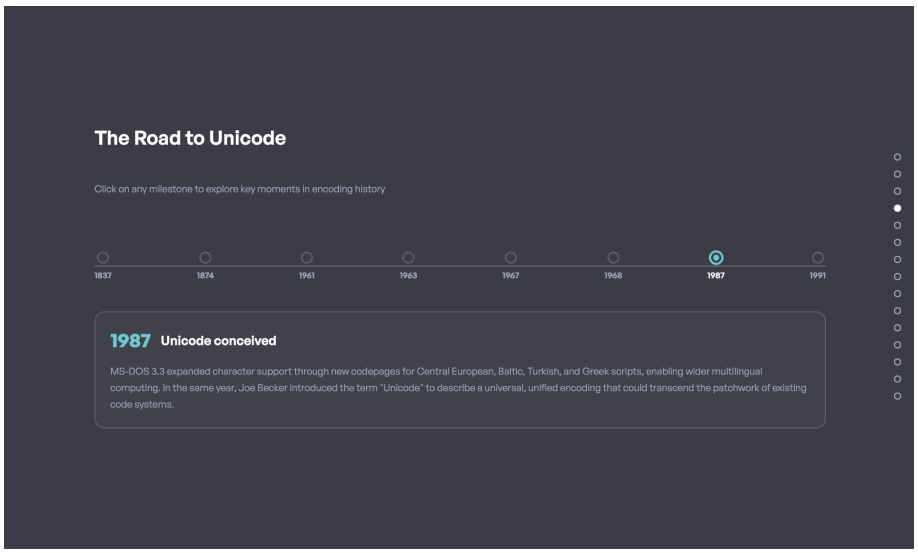
The 95%-completion stage:

Revised title/subtitle (to encompass thesis)

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)

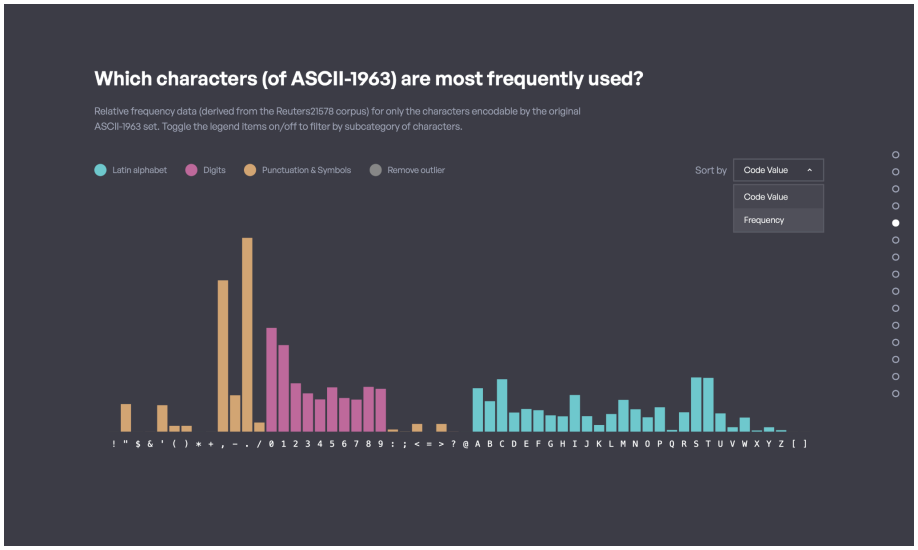


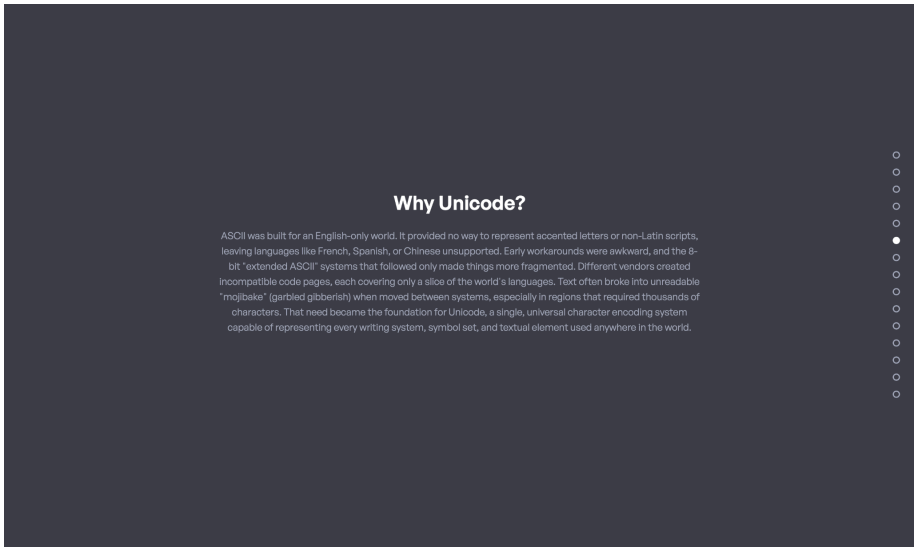
An interactive simulator of old ASCII, with fun fact generator and timeline info on the side (alongside a new visual for ASCII bit space)



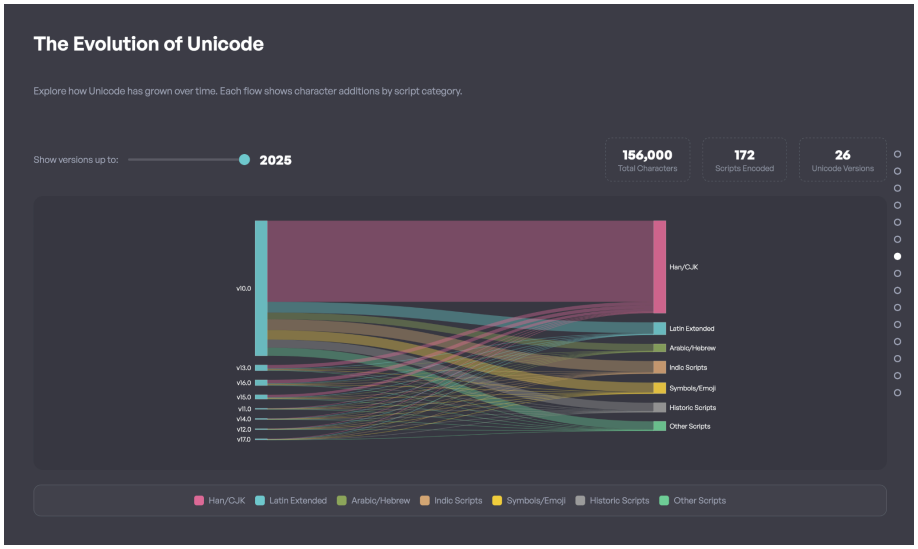
Turned the text block into a clickable timeline, per users' preferences. Also pinpointed the major milestone years and information we want to focus on — informative but also digestible and quick (also per the users' preference).

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



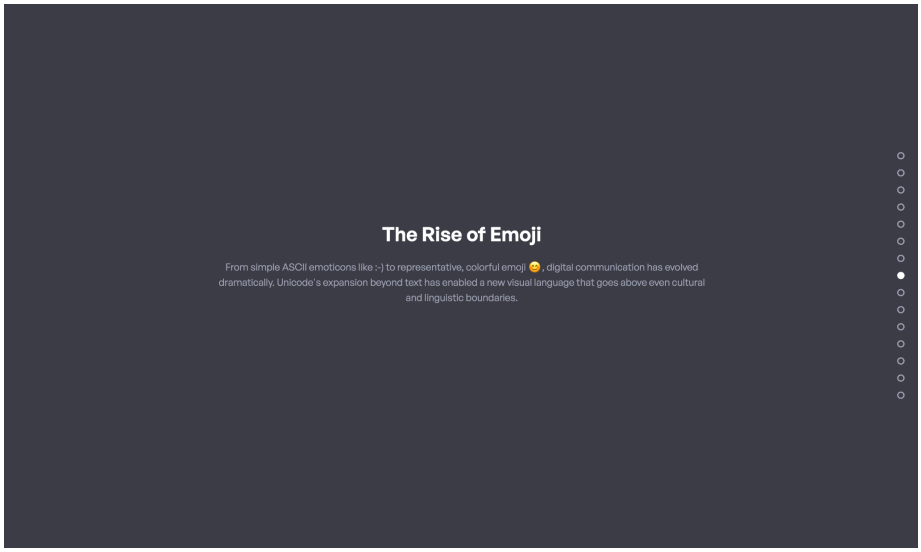


New information to introduce and motivate Unicode (leading up into the core visualizations).

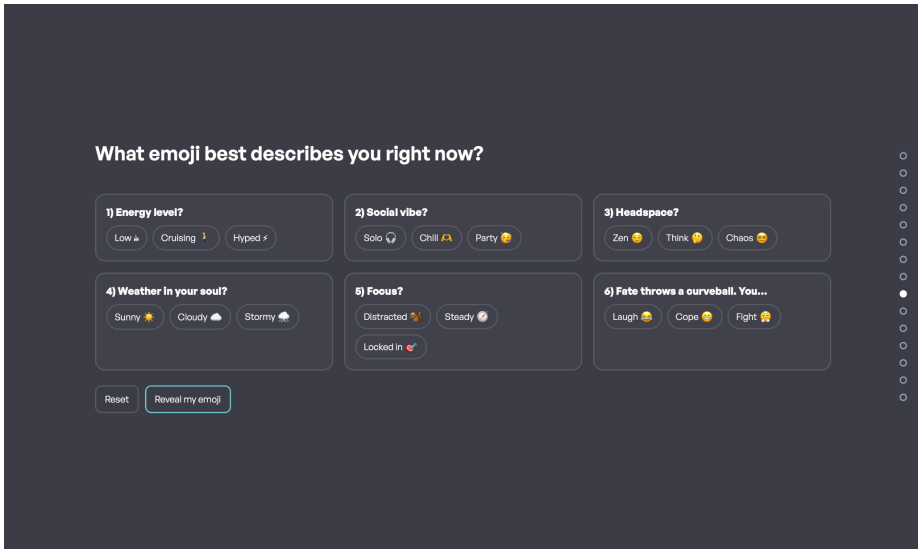


Sankey visualization that allows users to explore Unicode versions through time (from 1991 to 2025). Added an overall character counter, total script counter, and number of total versions (including minor updates). Flows can be hovered over to see more precise character counts and distributions.

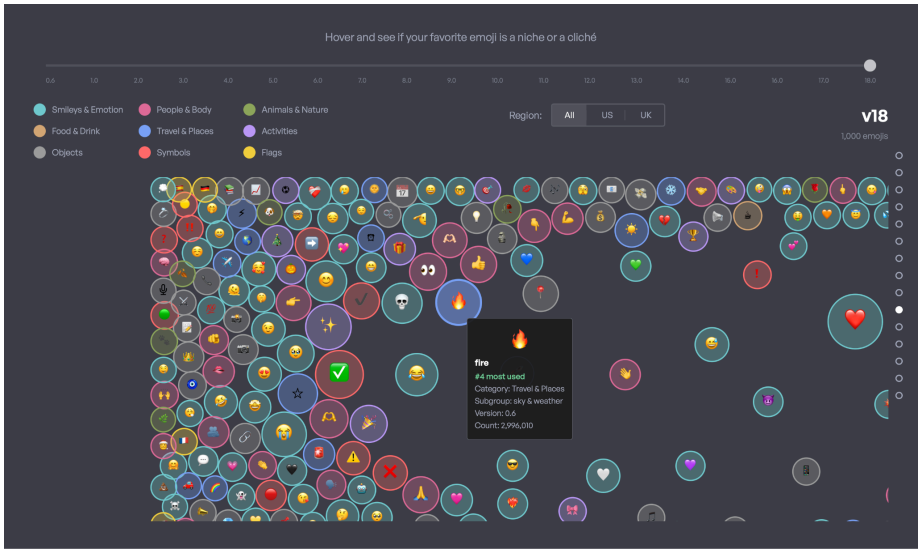
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



We added a smoother introduction into the Emoji section (for flow).

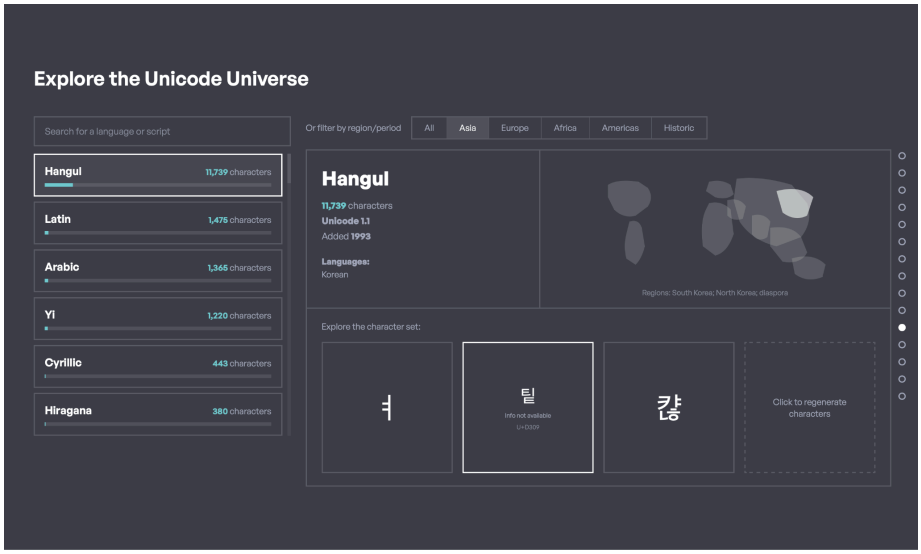


Emoji game!

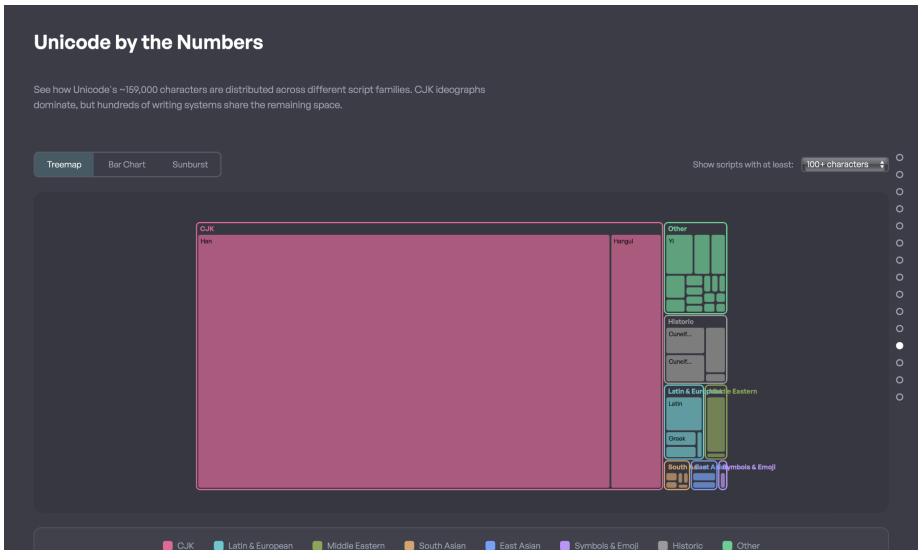


Emoji visualization, allowing users to explore the popularity data of different emoji in the world/US/UK. Legend items are all toggle-able, and can be used to filter. The slider on the top acts as a version timeline.

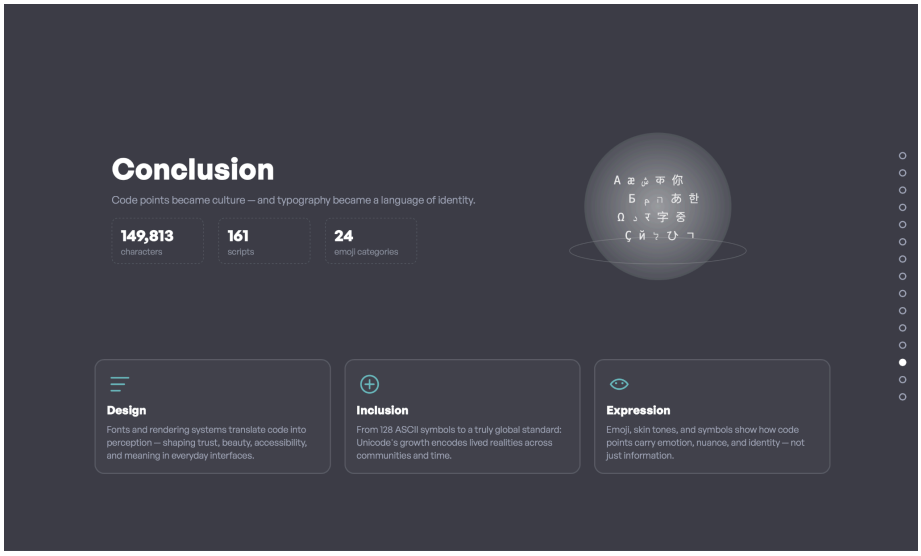
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



Implemented the Unicode Universe interactive visualization, which allows users to explore the writing system character set for all supported scripts. The left hand side has a list of scripts (sorted by character count), filterable by region/period. Info box shows languages covered, character count, version history, map (to-be-implemented), and the bottom is a randomly generated character sample. I decided to allow users to explore the set this way, since a number of characters cannot be supported due to limited font/keyboard support in various operating systems (including Mac/Windows), and this also reduces cognitive load for users (can seek out script categories on their own). The search bar enables language searching, which should be more familiar for users.

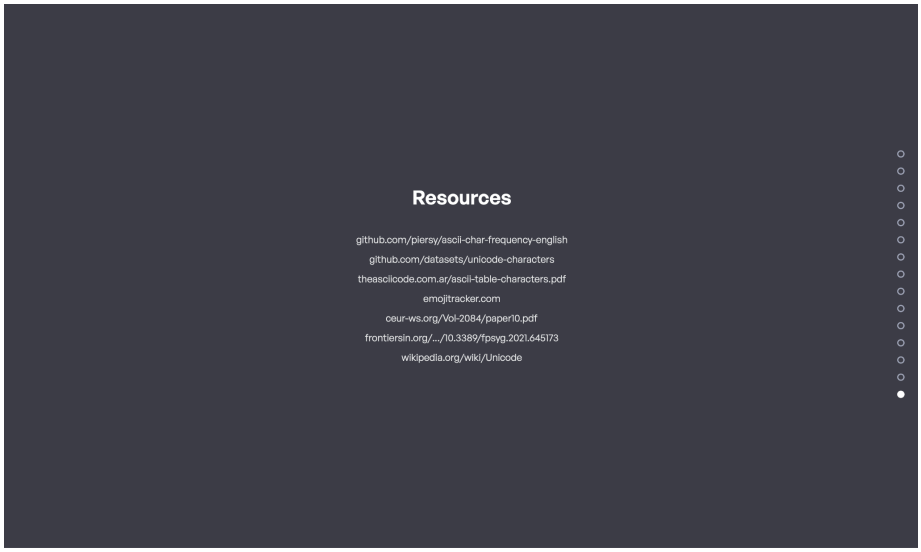


Wanted to show the *size* of script representations as well — a very interesting takeaway is how big the Chinese/Japanese/Korean/Han character set is. Allows users to filter the treemap by scripts of size X, see the data in bar chart/sunburst form. I wanted to include this visualization since it puts into perspective how limited our everyday symbol usage is, compared to what Unicode is capable of representing.



A proper conclusion slide! We wanted to summarize some key statistics, and have a nice visual at the end to wrap everything up.

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



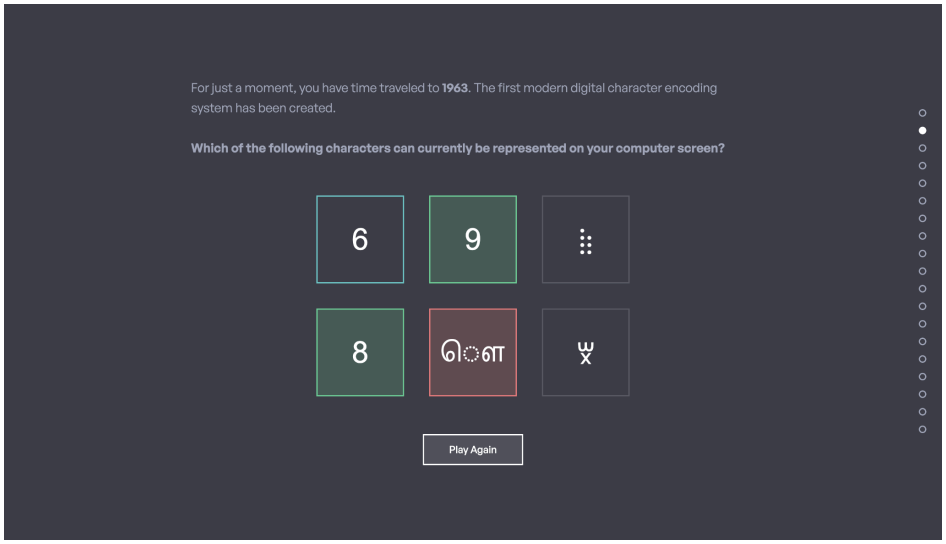
Resources page (bare bones)

Final Prototype

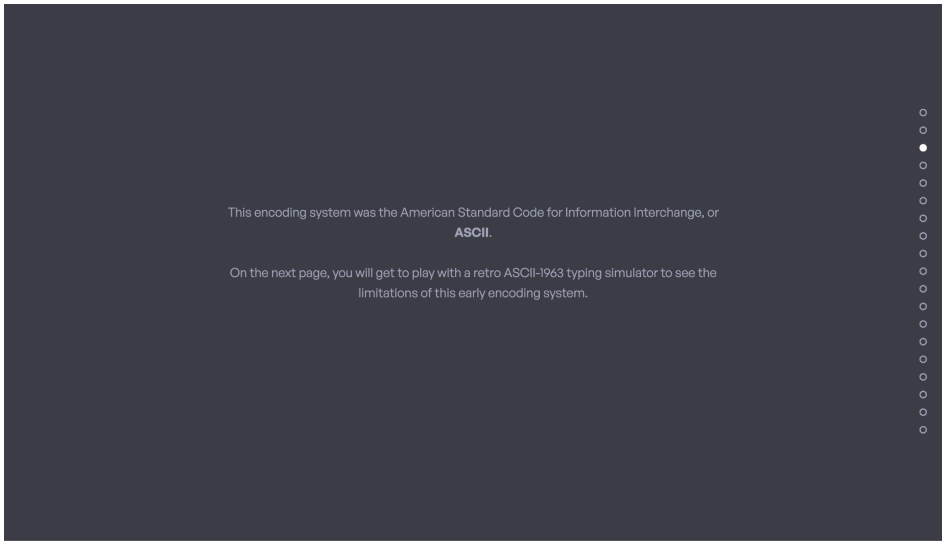
These final changes were made in several iterations; results combined & summarized below.
General: Revised and revamped general styling to be uniform/clean; fixed and clarified headings, subheadings, descriptions, in addition to all text and information throughout the website.

Redesigned the title page (dynamic background)

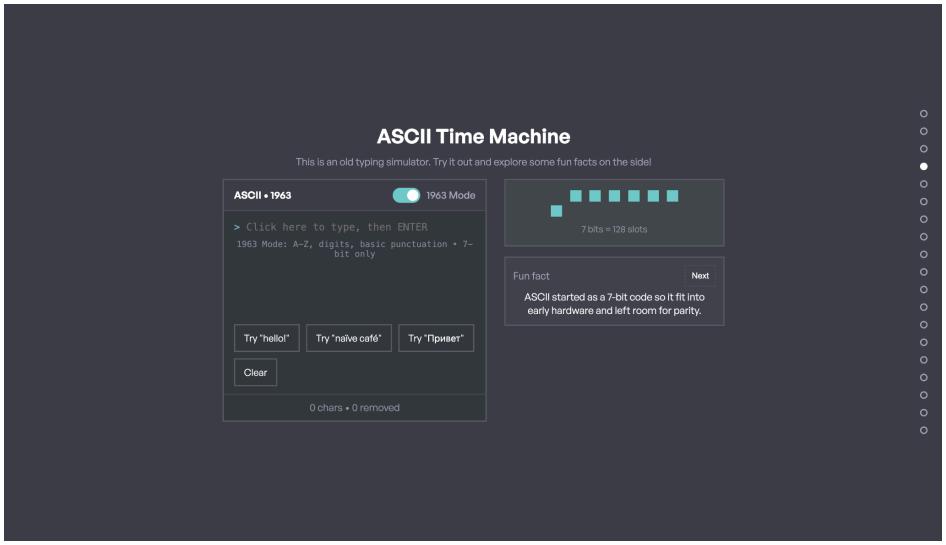
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



Improved spacing/text readability, and rephrased to be more conversational.

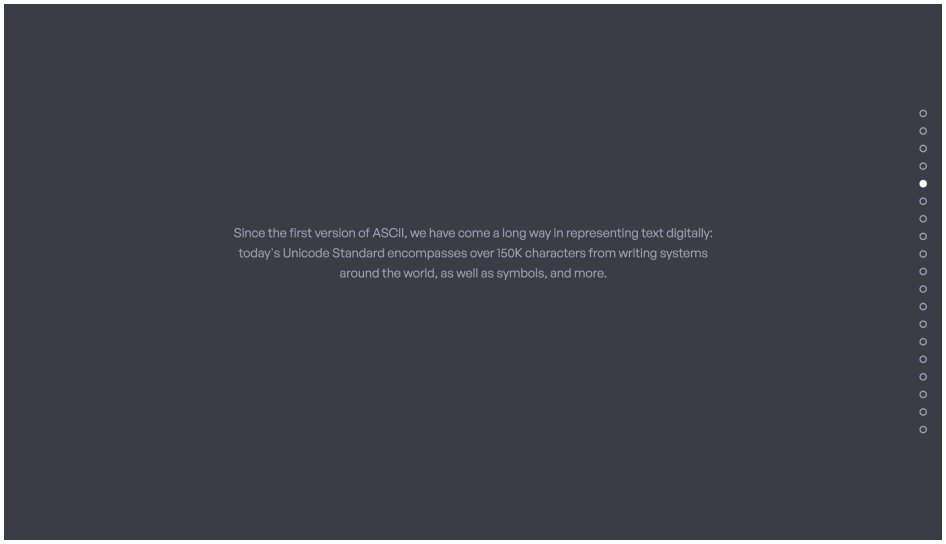


Added a clearer transition slide to introduce ASCII right off the bat, and flagging the simulator on the next page.

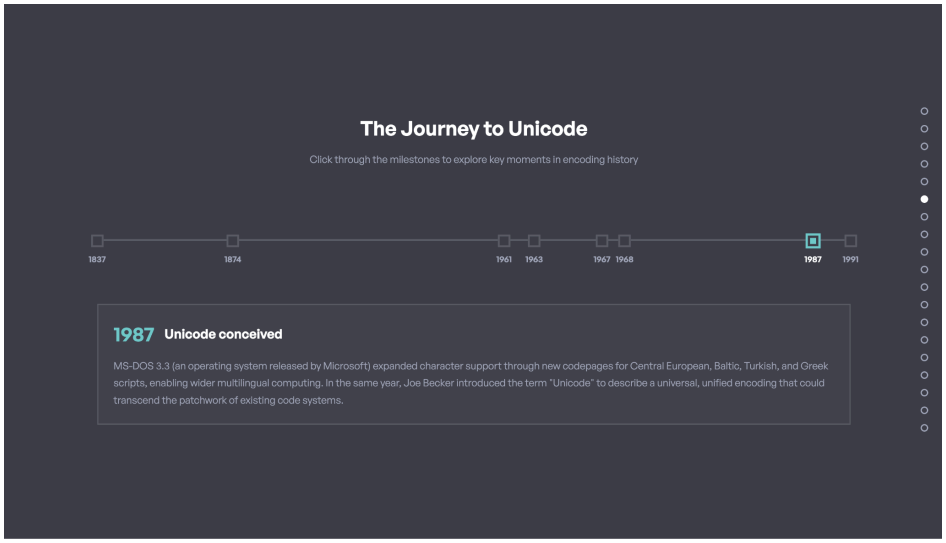


UI cleanup / removed extraneous timeline (repetitive and it cluttered the interface). Cleaned up the interface instructions, the off-centered blinking cursor that was confusing to users, and differentiated the trial options by including “Try” in the button.

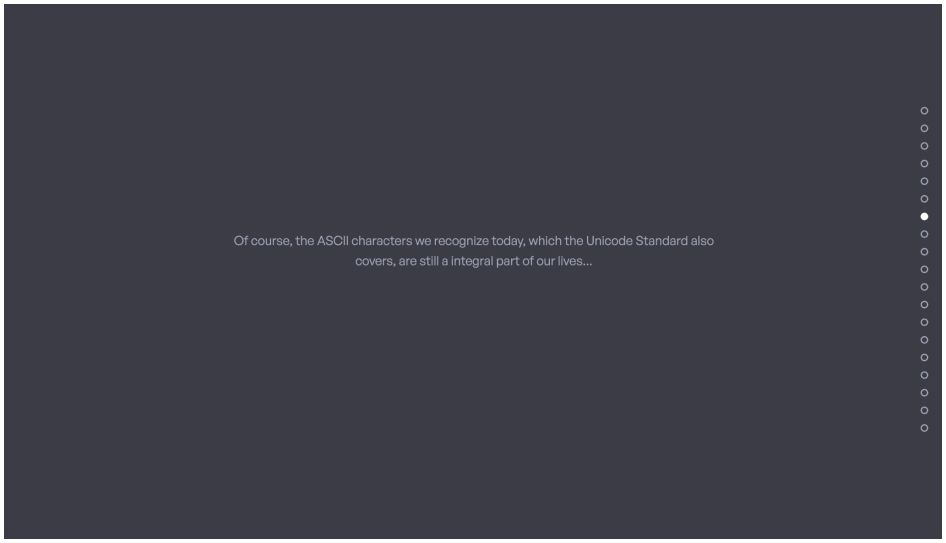
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



Added a smoother transition between the “ASCII sections” to the timeline (users tended to ask questions prematurely).

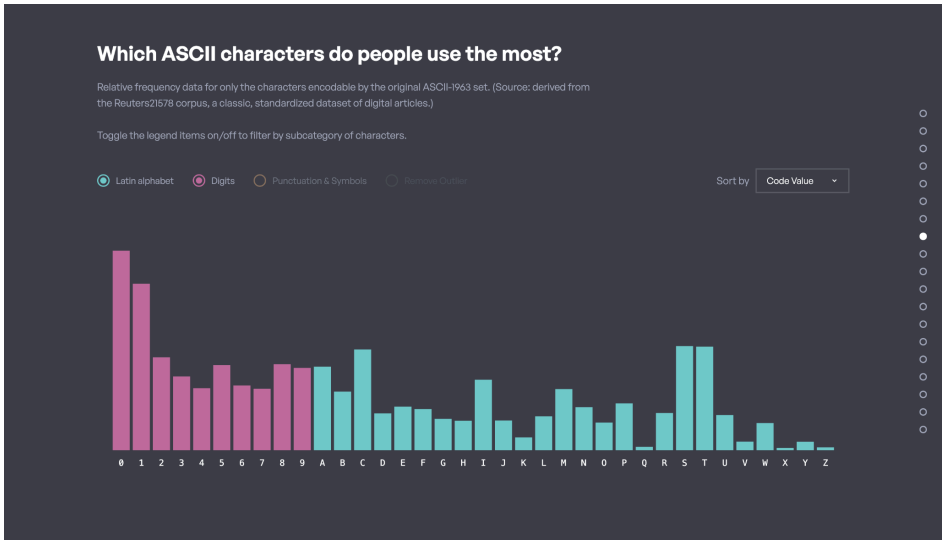


Revamped timeline. Besides UI fixes, the text was made clearer (acronyms were expanded and/or briefly explained), and the timeline was made to be more “to scale.”



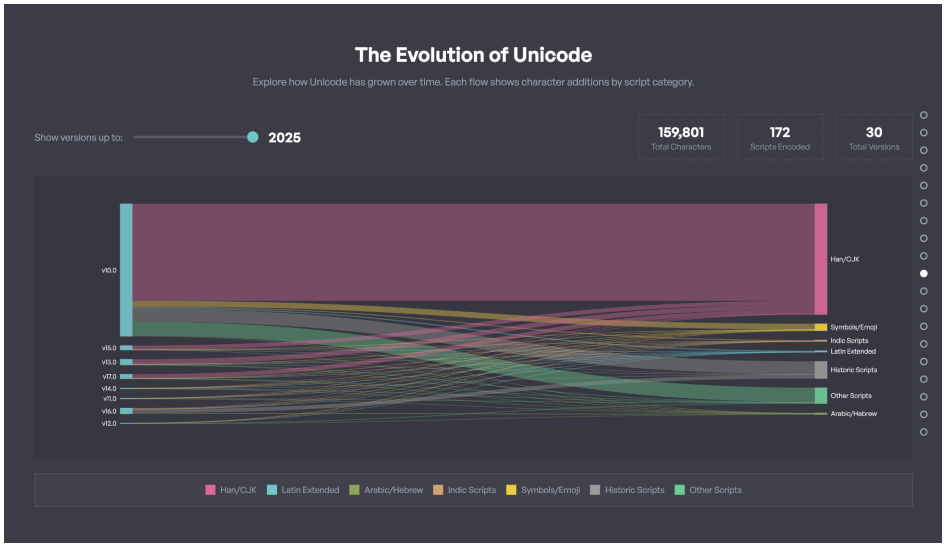
Clarifies relationship between ASCII and Unicode today, and segues into a more relatable visualization.

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



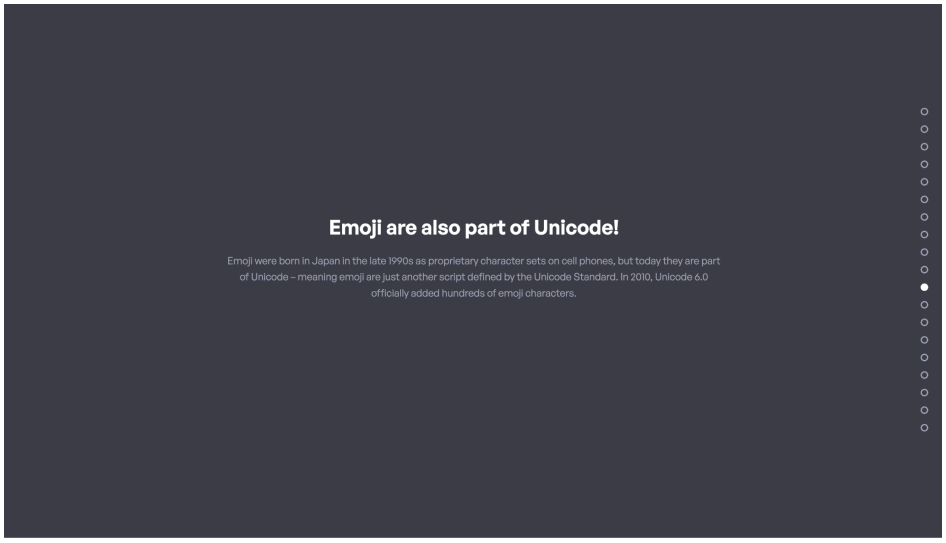
Improved spacing/text clarification, and changed button/legend styling to radio so that users can tell right off the bat that they can be toggled on and off. Also disabled “Remove Outlier” when the Punctuation & Symbols were filtered out (no outlier to remove there).

Animated the title (slightly more engaging for one of the longer text pages). Motivated Unicode, and explained the core issue underlying encoding.

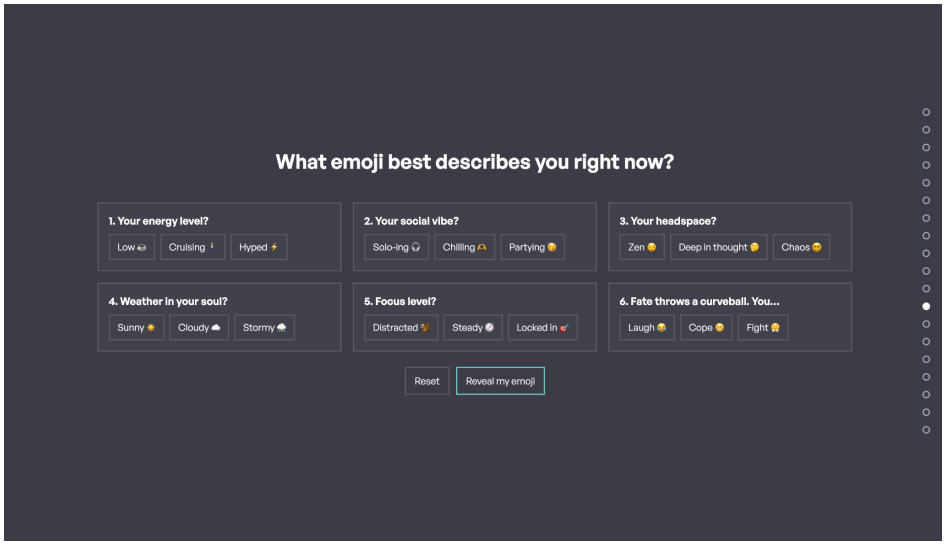


Revamped/added/corroborated Sankey diagram data.

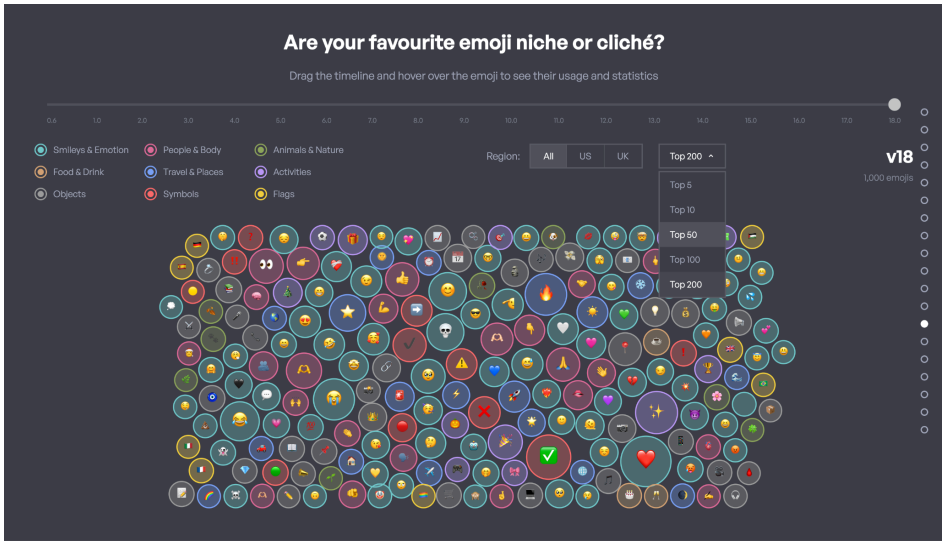
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



A smoother transition into emoji, with a description that is also more informative.

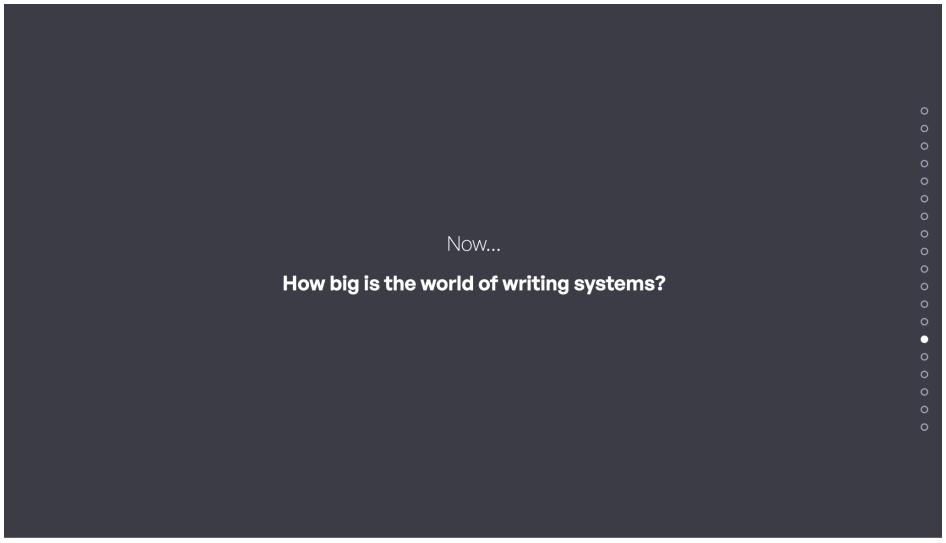


UI revamp.

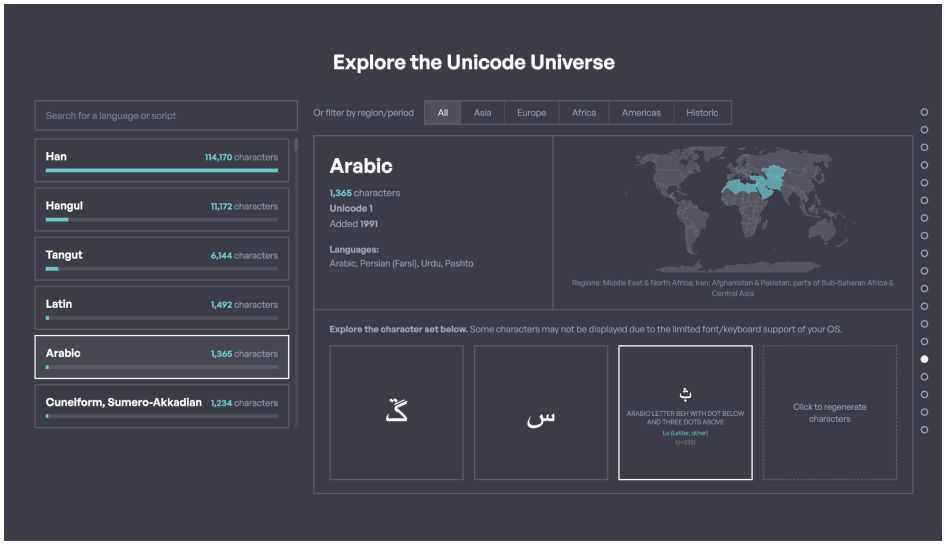


Modified the force/positioning of all the individual emoji, updated button styling to radio (so that users know right off the fact that the emojis are filterable), and added a new filter to change the number of emojis on the screen at a time!

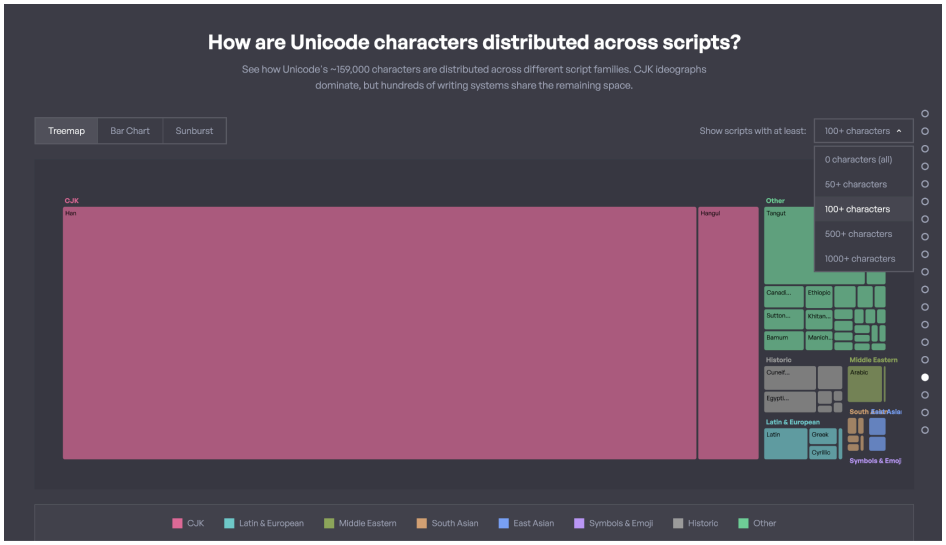
[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



Segue from symbols/visuals → writing systems.

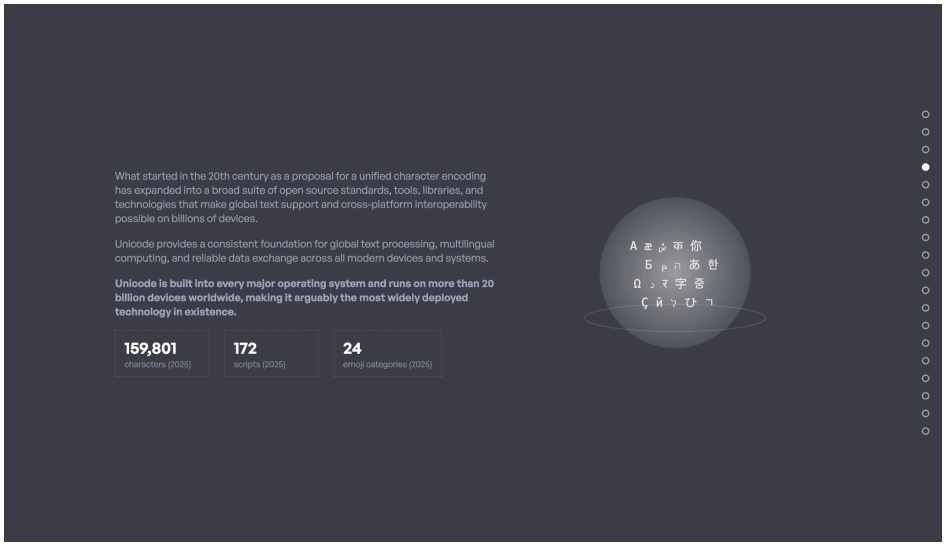


Updated error messages to decrease cognitive load on the user's part (interface simplified as much as possible). Fixed region mapping as well as filtering accuracy. Datawise, we added a column to the unicode_language.csv file to check/filter Unicode sampling ranges.



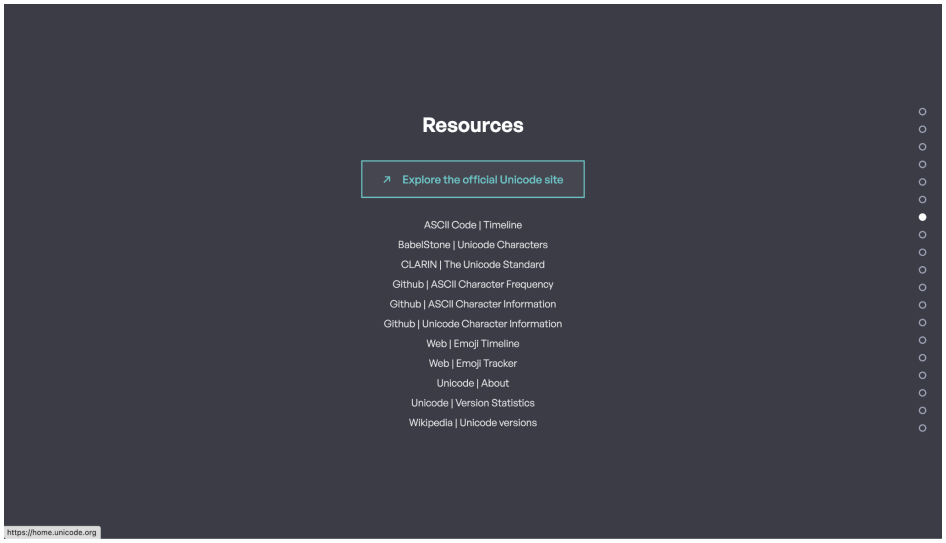
UI expanded/fixed.

[Process Book Link](#) / [Github Repository](#) / [precideer.github.io/cs-1710-unicode](#)



Specified year for the statistics on the concluding slide, and imported an informative conclusion.

Team slide.



A button to encourage users to explore Unicode on their own!