Assignment P5 (Summer 2021)

Emily Woska
ewoska3@gatech.edu

1 ANALYTICS: INFORMATION VISUALIZATION

Figure 1 (Martinčević, 2020) highlights the major players in the global economy, specifically the 42 countries with the highest gross domestic product (GDP) and share of the global economy. Each piece is proportional in size to the country's GDP market share, and its color corresponds to one of the 7 different continents (or the "Rest of the World" bucket). Each continent's member nations are also clustered together, as shown below. Further, each segment features a label with the country name, its revenue (in USD), and the percentage of global GDP.

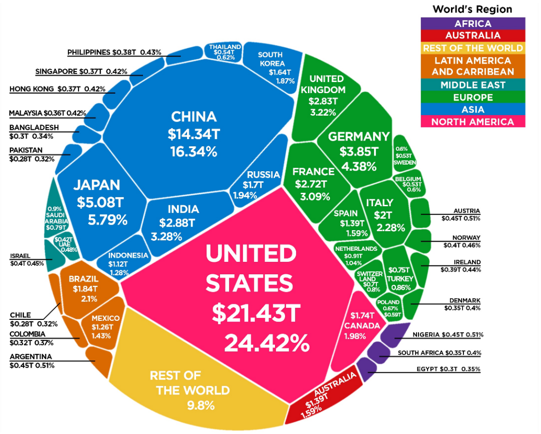

Figure 1—The World Economy: Global Domestic Product (GDP) by Country, 2019

This graphic highlights the economic powerhouses — namely the United States, China, and Japan — but obscures information about smaller countries (some of which are not even shown). The use of a circular representation presents a challenge because viewers will immediately think of the spherical shape of our planet; if we assume no prior knowledge of geography, viewers may interpret this visualization a bit too literally and draw some incorrect conclusions, like the relative sizes of neighboring nations (within a continent, especially) and the location/proximity of continents with respect to one another.

Figure 2 is a redesigned visualization with the same source data from the World Bank, but the size of each segment is determined by GDP per capita whereas the blue hue depicts the country's share (%) of global GDP.

Note: The page width means the labels are small, but on a dashboard or web page, readability would not be an issue. I created this treemap in Tableau, where tooltips provide access to additional details, such as country name, GDP in total and per capita, population size, and relative ranks.

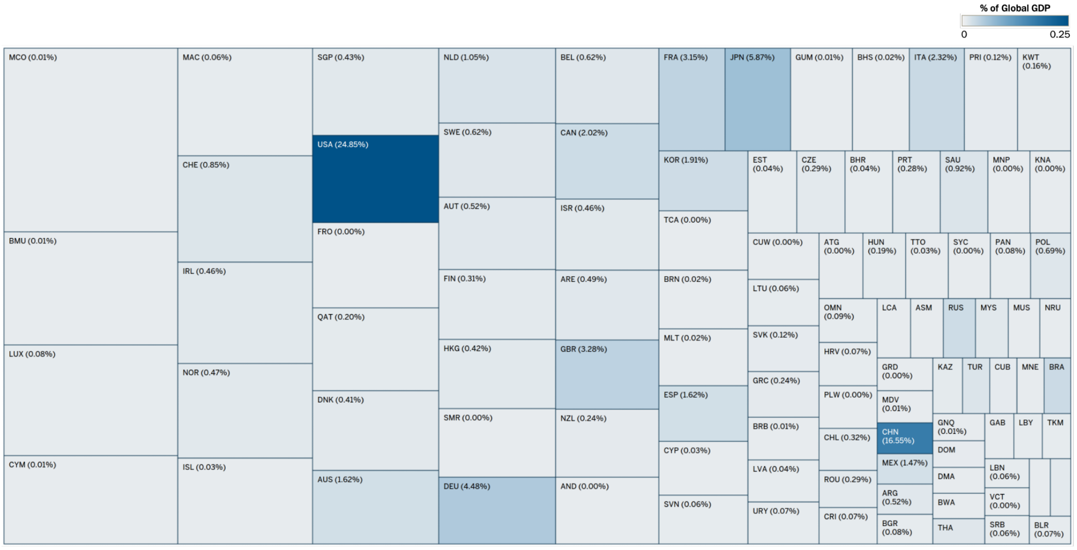

Figure 2—World Economy Treemap: Global Domestic Product (GDP) per Capita by Country, 2019

Organization is the primary difference between the two graphics. The original clusters countries together by continent, which the new design does not achieve (but could, as shown in **Figure** 3 within the Appendix, if the rectangles need not be shaded based on % of total GDP). The advantage gained has to do with the relative sizes and standardized shapes of the rectangles. Since the creators of the original chose to use a more physical, geographic representation of the data, I followed suit but redesigned it such that the shapes are all rectangles, which the human eye can more easily compare and detect differences. The location of each rectangle within the treemap also bears significance; the country with the highest GDP per capita rests in the upper left-hand corner while the countries with the lowest values fall towards the lower right-hand corner. The intensity of the blue hue maintains the same information that is presented in the original visualization, but this is now paired with key contextual details (at least, for a general audience of non-economists). From the original, one might conclude the U.S. is the richest country in the world; it is certainly the country with the highest GDP. For the layperson, however, the definition of "richest" may be different. The U.S., as the third most populous country, is simply more likely to have a sizeable GDP footprint. GDP per capita provides more information about the state of an economy. For example, China is the second largest segment in the original graphic but did not crack the top 75 for GDP per capita in 2019. The countries that the layperson would likely consider the "richest" or the most pleasurable to live in (i.e., Monaco, Bermuda, Luxembourg, etc.) do not even appear in the original representation, as they are lumped in with the "Rest of the World."

Technically, these are different interpretations of the same data, but the redesign considers an additional variable (population by country) and takes the shape of a more predictable structure, which is less "sexy" but more useful to a broader audience. Because the original visualization conveys a geographic element and the source data includes relevant features like total area, population size, etc., the approach with GDP per capita (despite being a slightly different reporting metric) does not require any changes to the data, and in the eyes of the consumer, may ultimately be a "more faithful" representation.

2 ANALYTICS: DEDUCTIVE DISCLOSURE

Prevention of deductive disclosures is a shared responsibility. Creators of the data set must ensure appropriate measures are taken to mitigate the risk of deductive disclosures, but custodians and researchers also play a role. The data set can be thought of like any tool; the creator(s) should do their best to engineer safeguards against misuse, but custodians and researchers can (and will) always find ways to misuse tools, especially because of the ease with which technology can rapidly evolve to overcome even the best of current security protocols. Furthermore, there has to be a balance between (A) the resources needed to engineer the safeguards, and (B) the risk that the information in the data set could be compromise, plus the harm that would come if that were to happen. This particular data set carries an extremely low level of risk (i.e., it is unlikely anyone, including you, cares about that C in Calculus 2), so many of the below protections would not be necessary.

Suppression techniques (i.e., the removal of "riskier" features in the data set before publication) could be used to prevent deductive disclosures in this case. The year could be stripped from the semester column, which would then only indicate whether the term occurred in the fall, spring, or summer. Notably, this approach would render any sort of time series analysis impossible. The student identifier could also be removed from the publicly available data set. This would hinder within-student analysis (e.g., if a researcher wanted to examine how a single student performed in multiple classes to determine the appropriateness of prerequisites). Suppression can also be applied to certain records with riskier information; for example, if any grade below a B- is considered higher risk, then the corresponding years in those rows could be eliminated, or the grade itself could be eliminated (thereby implying that blank values represent a category of grades ranging from C+ to F). This last example branches into another type of risk mitigation methods, known as generalization. Sometimes researchers will make only abbreviated versions of a data set available to the public. For example, semesters could be aggregated into years, or even two/five-year groupings. Alternatively, courses could be labeled by department and level of difficulty (e.g., "entry level" and "computer science"). In all of these cases, 2+ data sets must be managed by the researchers/curators, but risk is significantly reduced.

These safeguards limit the usefulness of the data, so custodians typically implement additional precautions, like restricted (e.g., on-premise only) access to the data. That does not seem necessary here. At minimum, however, database logins should be traced, so if re-identification does occur, the appropriate authorities can trace the problem to its source. Another reasonable safeguard is the implementation of a requirement for requesters of the data to submit a research proposal for approval by the creator and/or a community review board, which continues to become more popular at healthcare institutions across the nation. One notable disadvantage here is the increased burden of maintenance and oversight required by data custodians (and imposed on researchers). These methods limit the potential benefits because fewer people will have access to the data set, but increased accountability for researchers should ultimately reduce — **but not eliminate** — the risk of deductive disclosures. The risk of misuse remains, and the burden lies with the researcher(s) to use the data only for legitimate purposes.

3 CHI CONFERENCE PAPERS

3.1 Understanding Walking Meetings: Drivers and Barriers

Author(s): Ida Damen, Carine Lallemand, Rens Brankaert, Aarnout Brombacher, Pieter van Wesemael, and Steven Vos

Understanding Walking Meetings: Drivers and Barriers (Damen et al., 2020) explores the concept of "walking meetings" as a way to limit sedentary behavior in traditional office environments. Their primary driver is the non-new realization that a sedentary lifestyle is detrimental to human health and well-being.

The researchers established a standard "WorkWalk" route, placed signs outside faculty buildings to mark group meeting points, and updated the university's existent conference room reservation system to include the ability to schedule meetings along the route (instead of a normal, in-office meeting). Usage of the "WorkWalk" was evaluated over the course of 14 months; sixteen participants contributed to the study through individual, semi-structured interviews conducted in motion. The collective results focus on key areas, such as social dynamics, feasibility, context-setting, and time management.

They identified several barriers to "WorkWalk" implementation. First, some employees were reluctant to request a walking meeting with their leadership; the presence/absence of a relationship with the other participant(s) and the ability to anticipate their feelings played a role in the decision-making process. Second, some participants considered walking meetings to be distractions that may adversely impact meeting outcomes. Third, the "WorkWalk" simply felt out of the ordinary because walking meetings were not a part of everyone's daily routines, and certain external factors, such as bad weather, disrupted patterns of consistency. Finally, meetings that required larger groups or presentations/note packages could not be conducted in motion due to impracticality.

The researchers also found positives. "WorkWalk" meetings were perceived as more informal, relaxed, and natural — thereby enabling better communication between more junior employees and leadership. The standard path fostered a shared sense of space and time, for specific locations and landmarks could be interpreted by meeting attendees as markers of progress (i.e., how long was left); notably, these context clues improved information recall, as details could be associated with physical memories.

I have struggled with the consequences of an office-induced sedentary lifestyle for years. Pre-COVID, I frequently advocated for walking (or at least standing) meetings when appropriate/possible because people seem to enjoy the change in pace, which typically boosts productivity and collaboration. Now at home, I have attempted to introduce more physical activity into my work day; even in quarantine, I have identified new opportunities and recently invested in miscellaneous exercise equipment for my home office, to include an under-the-desk elliptical and a vibration plate. My experience with all the usual work activities (e.g., strategy sessions, code reviews, stakeholder interviews) has been completely transformed, and it makes my days much more enjoyable.

3.2 Why Johnny Can't Unsubscribe: Barriers to Stopping Unwanted Email

Author(s): Jayati Dev, Emilee Rader, and Sameer Patil

Why Johnny Can't Unsubscribe: Barriers to Stopping Unwanted Email (Dev, Rader, and Patil, 2020) explores mechanisms employed by senders to implement "unsubscribe" functionalities and their relative ease to use. The researchers focus on 3 topics: the overwhelming nature of institutional email, ineffective email filtering/boundary definitions, and users' inability to cope with unwanted email.

In total, 18 participants of different ages and backgrounds took part in the study. They were asked to save 10+ unwanted emails beforehand, and their attempts to unsubscribe were observed and recorded. Semi-structured interviews were also conducted to better understand their motivations. This revealed 3 layers of difficulty. The most straightforward and convenient methods involved 1-2 clicks, which navigated the recipient either directly to a confirmation page or to a page where the user simply clicks a button to confirm the unsubscribe action. Other methods were labeled as "somewhat inconvenient and difficult" and either involve (A) multiple or more complex confirmation pages (e.g., prompts for an email, required justification) or (B) broken/suspicious web pages. In these cases, users generally felt uncomfortable and/or annoyed. The most difficult and cumbersome mechanisms resulted in the worst user experiences. Some required composition of an email to unsubscribe (with or without a confirmation in response), site login for manual adjustment of settings, or navigation of a "subscription center" landing page with information overload. Each of these methods aims to obfuscate the unsubscribe process and prevent users from achieving their goal.

Spam and subscription emails became the bane of my existence at some point. My personal internet setup now employs Pi-hole (https://pi-hole.net) to block unwanted advertisements and internet traffic. Most companies collect data on customer behavior, so it is not uncommon for a merchant to "phone home" to an analytics server to try and capture a customer's attempt to unsubscribe. In response, Pi-hole blocks the traffic, as well as the entire unsubscribe flow. This means, to unsubscribe, I unfortunately either have to add their particular DNS entry to my allowed-list or wait until I disconnect from my home network.

4 OTHER CONFERENCE PAPERS

4.1 Is Faster Better? A Study of Video Playback Speed

Author(s): David Lang, Guanling Chen, Kathy Mirzaei, and Andreas Paepcke
Conference: International Learning Analytics & Knowledge Conference

Is Faster Better? A Study of Video Playback Speed (Lang et al., 2020) confronts a potential challenge in online education: how higher playback speeds impact students' comprehension of lecture material and overall course performance.

The researchers cite previous studies that revealed (1) human comprehension benefits from engagement of multiple communication channels (i.e., audio, video, and textual components), and (2) students dislike 2.0x video playback speeds but do not suffer adverse impacts in terms of comprehension. Research also suggests massive open online course (MOOC) delivery has adapted to meet student needs/expectations; this is evidenced by professors' attempts to chunk content into short, targeted videos that allow for quick absorption.

The hypothesis is that students try to optimize their educational productivity within certain (time and effort) constraints. The researchers touch on multiple production functions, with derivatives taken to determine the existence of comprehension optimality with respect to different budget and indifference factors. The analysis shows that higher playback speeds correlate with decreased returns due to the asymptotic curve.

Experiments involved different students, who watch videos at 1.0x and 1.25x speed, across multiple course types and video duration times. Their findings indicate students who consume videos at 1.25x the normal speed experience more time-savings (obviously). The notion that these students, who speed up lecture videos, may consume more content was only marginally supported, however (understandable given their objectives and constraints). Interestingly, self-regulatory behaviors, such as pausing and rewinding, were split at higher playback speeds — with fewer pauses, but more rewinds, observed.

The main conclusion here is the most important: students, who speed up video content intake, do indeed perform better in courses, as compared to those students, who prefer the normal playback speed. From my point of view, the study lacks heterogeneity in their sample of student participants (i.e., little variation across age, demographics, and educational background), so it is entirely possible that these results do not hold for all student populations. Nevertheless, this paper delivers some promising numbers in support of increased course intake and a foundation for additional research on online learning best practices.

Over the course of my 9 previous OMSA courses (and this one, to date), I have experimented with different video playback speeds and identified 1.5X as my sweet spot. I was curious as to whether this research would validate my decision or reveal that I have made things harder for myself over the last 2 years...

4.2 Student Performance Prediction Using Dynamic Neural Models

Author(s): Marina Delianidi, Konstantinos Diamantaras, George Chrysogonidis, and Vasileios Nikiforidis
Conference: International Conference on Educational Data Mining

Student Performance Prediction Using Dynamic Neural Models (Delianidi et al., 2021) explores the possibility of predicting a student's answer to their next exam question(s) based on previous course performance and evaluation. The architecture of their approach centers on a multi-layer model, comprised of an initial dynamic neural network (either recurrent or time-delay) that is followed by a multi-layer feed-forward network for the answer prediction(s). Through use of different embedding, input skills and historical responses were encoded into the overall model, and different arrangements and initializations were tested.

Other studies have attempted to use different neural architectures, like hidden states in LSTM (Long Short-Term Memory) models; Dynamic Key-Value Memory Networks (DKVMN); and Deep-IRT models with an element of Item Response Theory to extend DKVMN and measure student ability and question difficulty.

The researchers employed (1) time-delay neural networks (TDNN) that used only feed-forward connections and had finite memory, and (2) recurrent neural-networks (RNN) with feedback connections that can have potentially infinite memory. Two different "treatments" existed for the initial embeddings, which represented skills in these models: pre-trained or randomized. Model performance remained fairly consistent between builds with either pre-trained or randomized input embedding initializations. Both the RNN and TDNN experiments outperformed the others (likely due to the use of embeddings, which generally outperform one-hot data encodings). The RNN model outperformed its TDNN counterpart, but the TDNN results are exciting because the technique is fairly novel within this field.

Most of the data science community is excited about the potential to simulate human understanding! Consider the infinite applications of neural models — computer vision, fraud detection, NLP, etc. If these same concepts could be applied to predictions of student performance, I wonder about the implications for how students would then be evaluated, given feedback, coached, and re-evaluated.

5 REFERENCES

[1] Damen, I., Lallemand, C., Brankaert, R., Brombacher, A., Wesemael, P. van, and Vos, S. (2020). “Understanding Walking Meetings: Drivers and Barriers”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, New York: Association for Computing Machinery, pp. 1–14. URL: https://doi.org/10.1145/3313831.3376141.

[2] Delianidi, M., Diamantaras, K. I., Chrysogonidis, G., and Nikiforidis, V. (2021). “Student Performance Prediction Using Dynamic Neural Models”. In: *CoRR* abs/2106.00524. URL: https://arxiv.org/abs/2106.00524.

[3] Dev, J., Rader, E., and Patil, S. (2020). “Why Johnny Can’t Unsubscribe: Barriers to Stopping Unwanted Email”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, New York: Association for Computing Machinery, pp. 1–12. URL: https://doi.org/10.1145/3313831.3376165.

[4] Lang, D., Chen, G., Mirzaei, K., and Paepcke, A. (2020). “Is Faster Better? A Study of Video Playback Speed”. In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. New York, NY, USA: Association for Computing Machinery, pp. 260–269. URL: https://doi.org/10.1145/3375462.3375466.

[5] Martinčević, I. (2020). *The World Economy in One Chart: GDP by Country*. HowMuch.net. URL: https://howmuch.net/articles/the-world-economy-2019.

6 APPENDIX

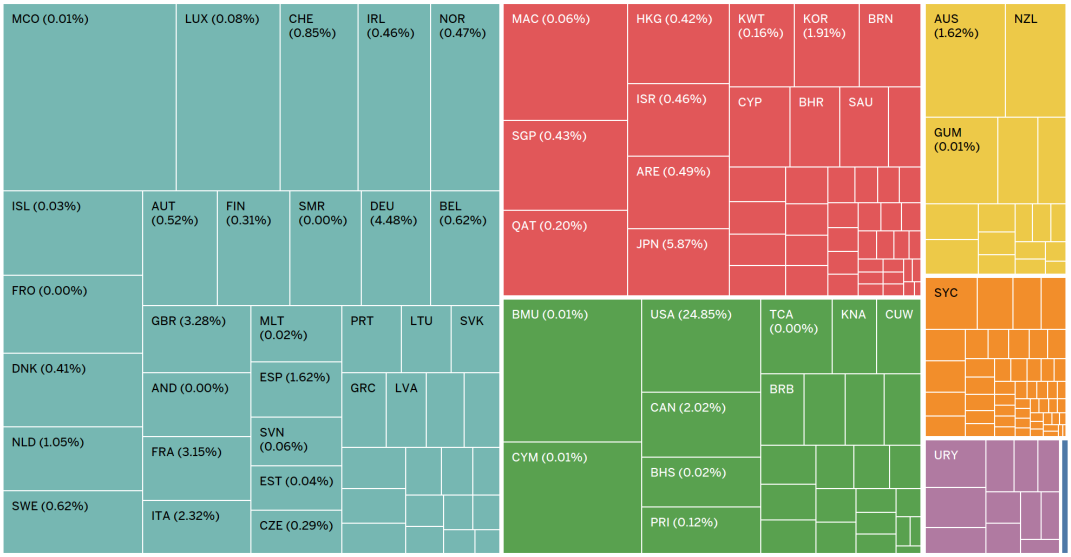

Figure 3—World Economy Treemap: Global Domestic Product (GDP) per Capita by Continent and Country, 2019