Five Steps for Using Legal Data to Improve Drafting

bryan wilson

Published in Towards Data Science · 7 min read · Jan 3, 2018

Legal data is incredibly valuable. In the insurance field, data about the average similarity score of the clauses in a policy — a measure of the relative similarity or divergence for one clause in relation to a set of other similar clauses that are meant to accomplish the same function — can be used to instantly show how one policy stacks up against another. Data about the composition of cyber policies can be used to reveal which clauses or information a given policy might be missing.

In order to unlock the value of this data, however, legal professionals need to understand how to properly leverage it. One of the most straightforward contexts for doing this is in the drafting of legal language. And the key here, I believe is making the data easily digestible at a basic level.

A couple months ago, when I was fortunate enough to have the opportunity to speak on a panel at KC Cyber — a local cybersecurity conference — about cost effective strategies for managing cyber risk. Our panel specifically sought to answer questions about the interplay between information security and cyber liability insurance. Since the event ended, however, a question from an audience member stuck with me and can be credited as the inspiration for this post. The question was, "*What are the essential components of a cyber liability insurance policy?"* The reason the question has stuck with me is because it helped contextualize an issue I had been wrestling with for a long time — how to effectively use policy data to make this process of drafting a legal document more effective and understandable.

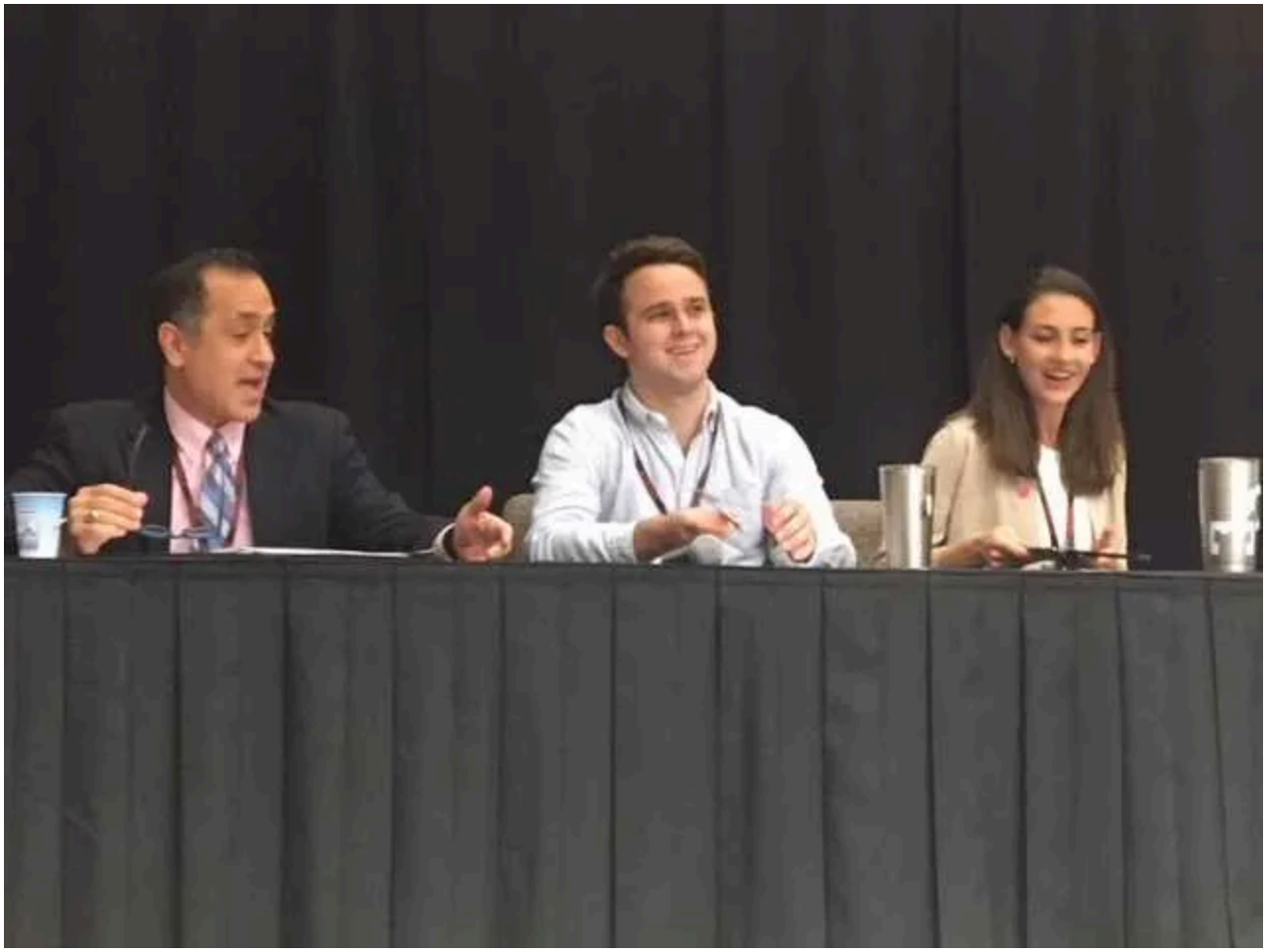


photo credit: Heather Otto

My immediate answer to the question referred to some findings from Woodruff, Sawyer and Co. on the core components of cyber risk. At a conceptual level, a cyber policy should cover the four general types of cyber risk: Privacy, Network Security, Errors & Omissions, and Media Liability. The law student in me would argue that this is more of a substance over form issue; the way that these general risk concepts are managed can vary from group to group, but as a reference point these general concepts can be used as a step on the path toward data-driven analysis. After there has been a general understanding of the risks that need to be managed, the same set of data can be used to help identify more specific pieces of policy language that would provide an organization with an even more robust framework for managing risk.

These strategies, however, are not limited to cyber liability insurance. In fact, the use of data — the end result of combining numerous policies, contracts, or documents, with machine learning algorithms, natural language processing, and other computer science techniques — can help identify the missing and divergent areas of any set of documents.

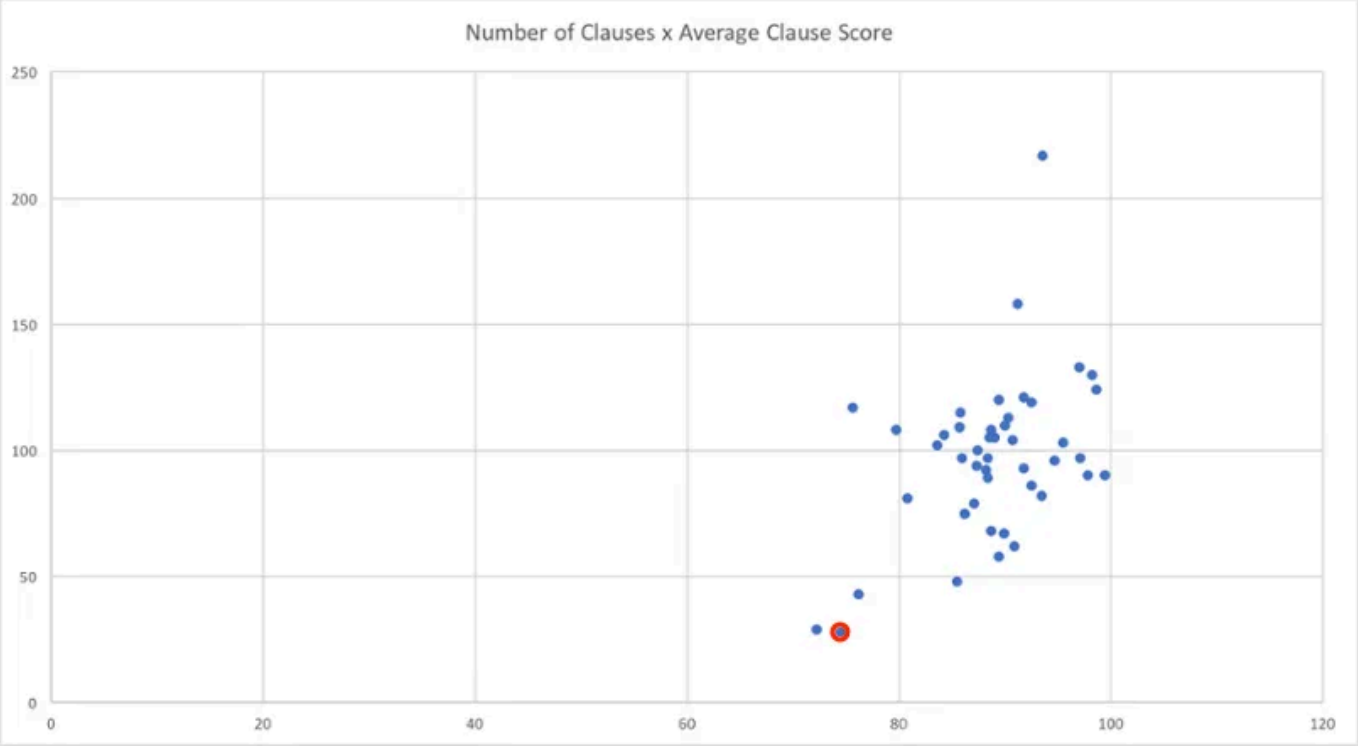
Looking at this from a more general perspective, I want to suggest a strategy for using policy data to identify divergent policy language, missing coverage parts, and bringing a higher deree of transparency to the process of identifying and understanding cyber risk.

To draft a more comprehensive cyber liability policy, underwriters, lawyers, and other risk management professionals should 1) inventory types of risk that need to be managed, 2) inventory the language of a policy (list clauses present, clauses not present, similarity scores, and types of risk to be managed), 3) use data about missing clauses and types of risk to add missing clauses to policy, 4) use data about low similarity score and types of risk to be

managed to improve weak areas of policy, and 5) evaluate the new policy data in light of changes to the policy.

> 5 Steps for Using Data to Improve Drafting
> 1) Risk inventory
> 2) Policy language inventory
> 3) Use data about missing clauses to match types of risk
> 4) Use data about higher scoring clauses to comprehensively improve similarity scores
> 5) Reevaluate updated policy

Continuing to use cyber liability policy language as a use case, let's walk through these steps and explain what this looks like in practice by analyzing the policy below circled in red, Aviva's Engineering Product "Cyber" Extension Wordings. One important fact to take note of throughout this exercise is that, like other legal texts, certain cyber policy wordings may be constructed for narrow purposes and integrated into a larger policy or set of documents.



Risk Inventory

Starting with the basics, let us only assume that the types of risk that are meant to be covered are the types of cyber risk outlined in the Woodruff, Sawyer, and Co. overview of cyber risk — Privacy, Network Security, Errors & Omissions, and Media Liability. It is foreseeable that this step would be more robust depending on how nuanced the Risk Inventory is.

Policy Language Inventory

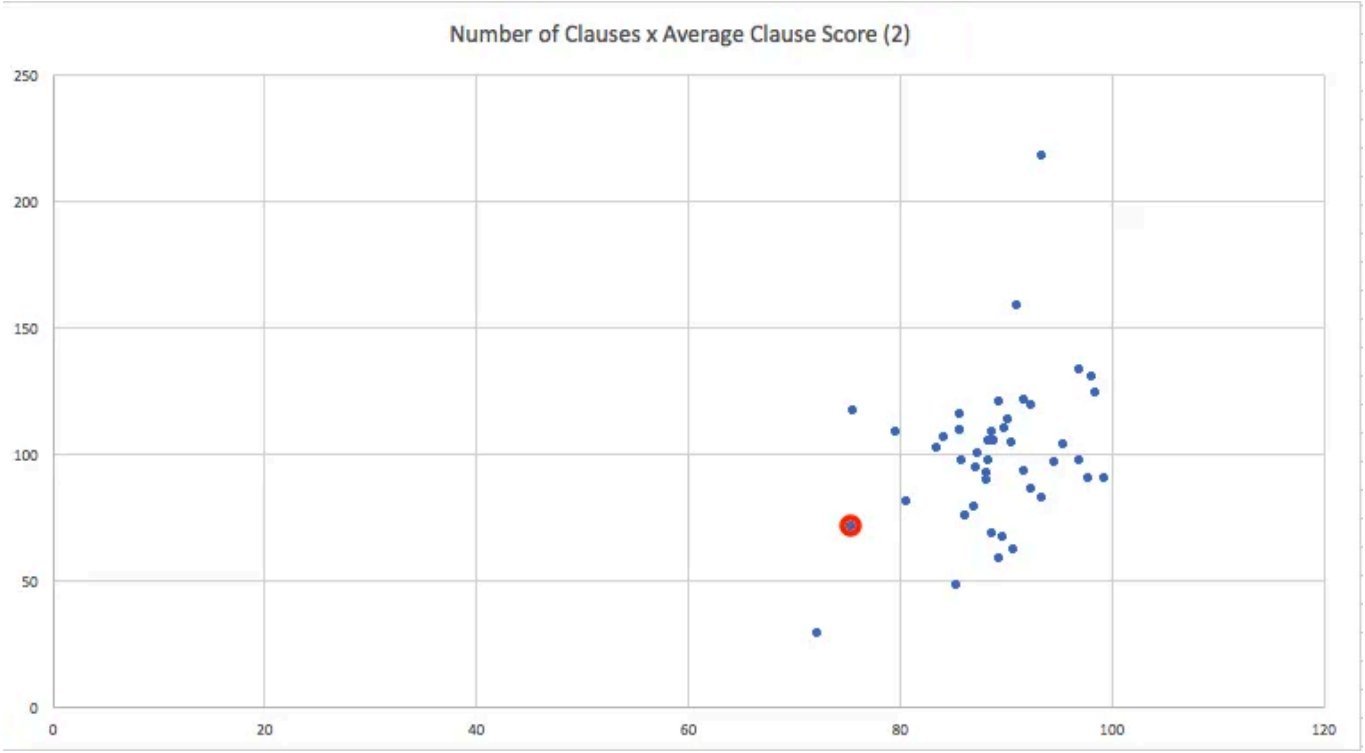
To understand what types of risk a policy is meant to cover, it is helpful to look at the insuring agreement sections of the policy. The insuring agreement sections are truly the core of a policy. Everything else in a policy is used to explain how to interpret these sections, define the words in insuring agreement sections, show what conditions must be met for the insuring agreement to apply, identify exclusions to the coverage, establish limits for the coverage, and anything else the drafters might have decided to add or exclude.

Looking at the previously mentioned Aviva document — the Cyber Extension Wordings — it is clear that this particular policy is only meant to insure the event management relating to Data Security Breaches and other obligations that result from the loss, theft, or accidental release of certain personal data. For example, the insuring agreement sections in this document cover things like indemnification in respect to "costs incurred… arising out of a Data Security Breach discovered during the Period of Insurance…"

Identifying Missing Clauses

At a granular level, there are many types of clauses that could be absent from a policy and cause it to be an outlier. It does not take a supreme amount of statistical knowledg to see that the average cluster of clauses is somewhere between 75 and 125 clauses with an average similarity score around 85.

Then, by taking the data about which policies are scoring well, which clauses are most common to the policies that are scoring well, and which clauses are outliers or appear less frequently, drafters can gain real-time feedback about other clauses that are usually found in policies of this type.



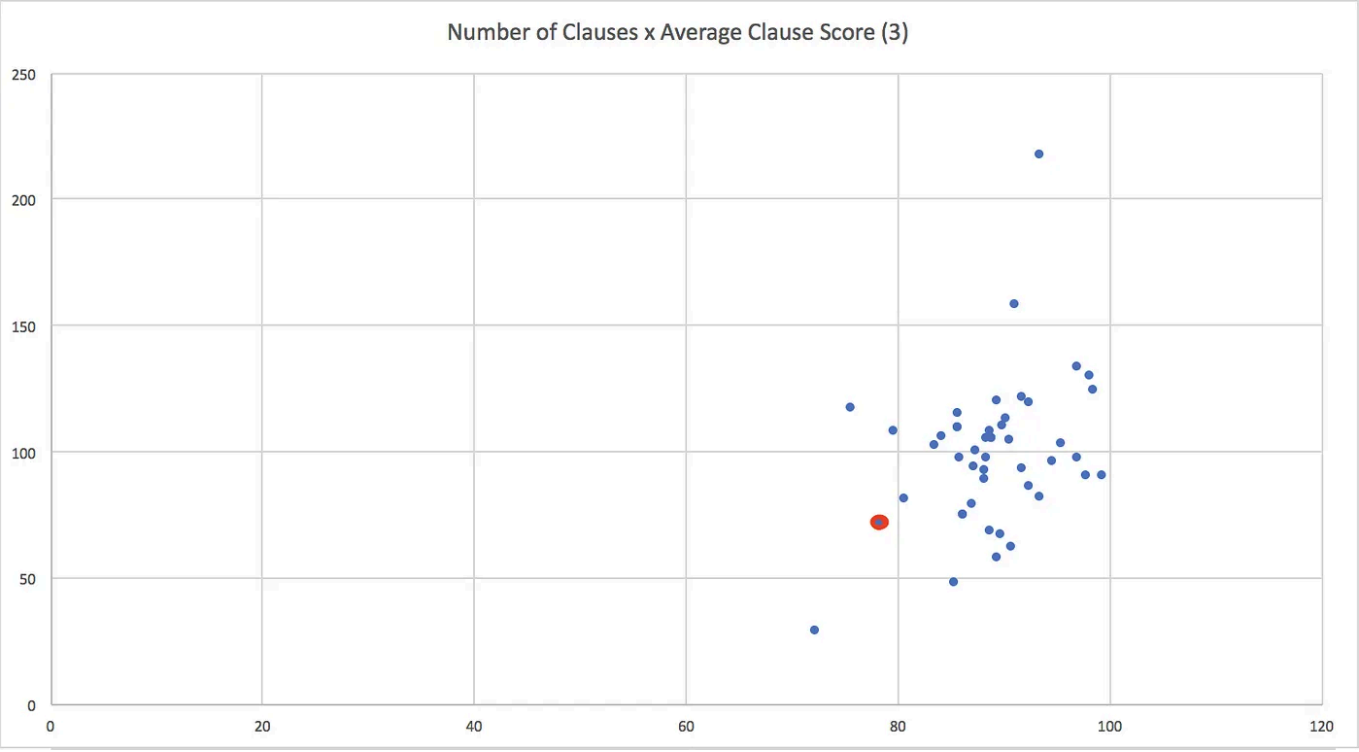
As an organization shopping for a policy, it would be helpful to understand this specific document could have actually function as an extension of Aviva's Cyber Electronic Media Core Policy. By combining the number of clauses of the two documents and averaging the new set of scores from Aviva's Electronic Media Core Policy and their Engineering Product "Cyber" Extension Wordings, the result (red circle in the graph above) is that the new policy is more in line with other policies in the industry in terms of length, but still has an average of under 80% similarity to the rest of our cyber index.

Taking into account the rest of the context, that there is really nothing in this combined policy that thoroughly covers Privacy or Media Liability, it would not be difficult to go through a library of clauses, find clauses for those sections, and continue building a more comprehensive policy that covers all of the risks we identified with the Risk Inventory.

Reviewing and Modifying Low Scoring Clauses

In doing a brief look at the lowest scoring clauses in the document, I

identified the areas of the policy that were weakest in comparison to the rest of the cyber liability marketplace (Exclusion — Prior Acts, Exclusion — Intentional Acts, Exclusion — Fees, Fines, and Penalties, and Limit — Deductible), and updated them to include higher scoring policy language from different policies. The result is that the average similarity score for the entire policy went up from 76% to 78%. What excites me about this analytical exercise is that this increase was achieved by only changing five of seventy one clauses. If the other 66 clauses were reviewed and modified, then the gains could possibly improve its ranking compared to the other policies that are out there.



Reevaluating the Updated Policy

Looking at what I did with this exercise, there are still some things that I would want to change before trying to create a robust, general cyber liability policy. For example, I would want to add clauses relating to Privacy and Errors and Omissions and substitute other lower scoring clauses for higher scoring clauses. By doing this, it would not be inconceivable to increase the

similarity score above 90% and have something that could be valuable and usable to a number of people.

The continued use and creation of data throughout the drafting process will facilitates a more iterative, transparent, and open process to reviewing and managing different types of risk. For those drafting these sorts of policies, it can cut down on time spent doing research and provide a measuring stick for less experienced drafters. And for those looking to learn about what policies actually should have, data can be a tool to teach people about when certain clauses or provisions apply, and under what circumstances other clauses or provisions should apply.

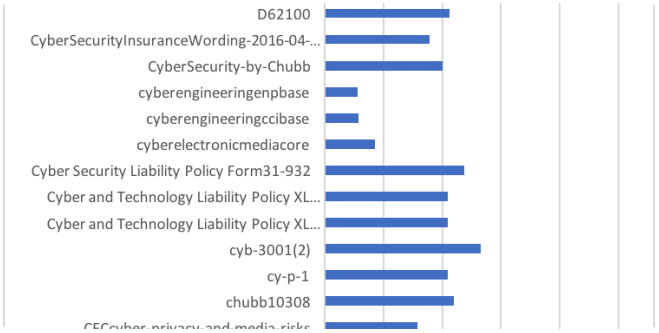
Cybersecurity Insurance Legal Insurtech

Written by bryan wilson

Edit profile

42 Followers · Writer for Towards Data Science

More from bryan wilson and Towards Data Science

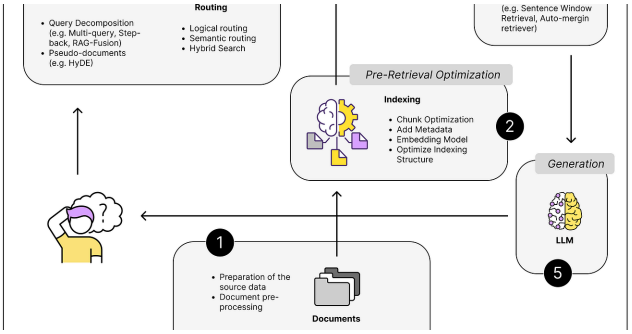


bryan wilson

Size *Sometimes* Matters: Using cyber policy length and similarity...

The following series of posts survey how the seemingly nascent features of an insurance...

Nov 29, 2017 115



Dominik Polzer in Towards Data Science

17 (Advanced) RAG Techniques to Turn Your LLM App Prototype into...

A collection of RAG techniques to help you develop your RAG app into something robus...

Jun 26 1.99K 20

Mauro Di Pietro in Towards Data Science

GenAI with Python: RAG with LLM (Complete Tutorial)

Build your own ChatGPT with multimodal data and run it on your laptop without GPU

Jun 28 744 13

bryan wilson

A screaming comes across the sky.

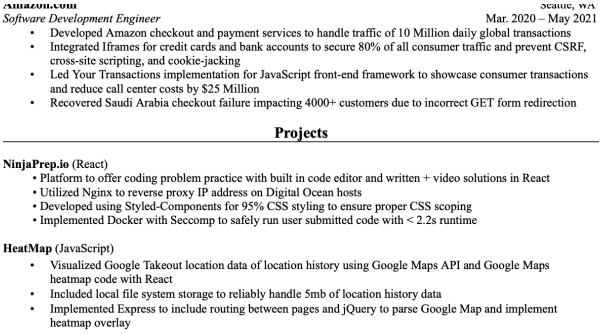
What follows will trace along the familiar storylines of Joseph Campbell's construction...

Aug 20, 2019 204

See all from bryan wilson

See all from Towards Data Science

Recommended from Medium


Amazon.com Seattle, WA
Software Development Engineer Mar. 2020 – May 2021
- Developed Amazon checkout and payment services to handle traffic of 10 Million daily global transactions
- Integrated Iframes for credit cards and bank accounts to secure 80% of all consumer traffic and prevent CSRF, cross-site scripting, and cookie-jacking
- Led Your Transactions implementation for JavaScript front-end framework to showcase consumer transactions and reduce call center costs by $25 Million
- Recovered Saudi Arabia checkout failure impacting 4000+ customers due to incorrect GET form redirection

Projects

NinjaPrep.io (React)
- Platform to offer coding problem practice with built in code editor and written + video solutions in React
- Utilized Nginx to reverse proxy IP address on Digital Ocean hosts
- Developed using Styled-Components for 95% CSS styling to ensure proper CSS scoping
- Implemented Docker with Seccomp to safely run user submitted code with < 2.2s runtime

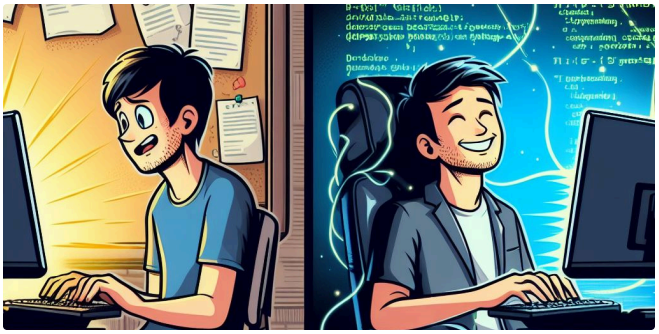
HeatMap (JavaScript)
- Visualized Google Takeout location data of location history using Google Maps API and Google Maps heatmap code with React
- Included local file system storage to reliably handle 5mb of location history data
- Implemented Express to include routing between pages and jQuery to parse Google Map and implement heatmap overlay

Alexander Nguyen in Level Up Coding

The resume that got a software engineer a $300,000 job at Google.

1-page. Well-formatted.

Jun 1 14.5K 219



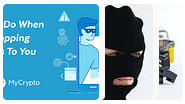
Abhay Parashar in The Pythoneers

17 Mindblowing Python Automation Scripts I Use Everyday

Scripts That Increased My Productivity and Performance

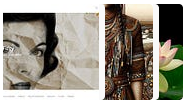
1d ago 4K 32

Lists



Tech & Tools

17 stories · 272 saves



Medium's Huge List of Publications Accepting...

334 stories · 3149 saves

Staff Picks

694 stories · 1158 saves

Natural Language Processing

1592 stories · 1151 saves

Barack Obama

My Statement on President Biden's Announcement

Joe Biden has been one of America's most consequential presidents, as well as a dear...

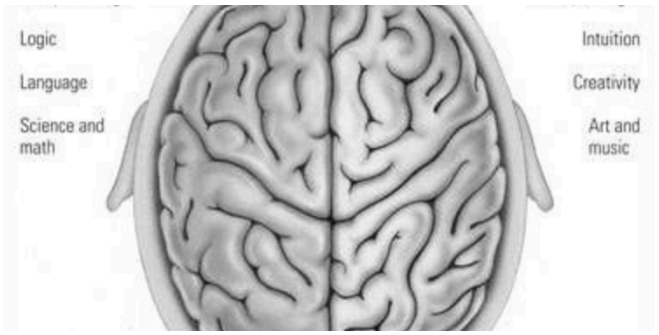
2d ago 18.6K 207

Karolina Kozmana

Common side effects of not drinking

By rejecting alcohol, you reject something very human, an extra limb that we have...

Jan 21 41K 1108

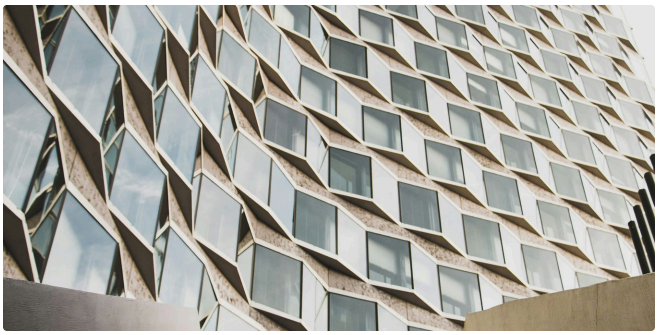


Sufyan Maan, M.Eng in ILLUMINATION

What Happens When You Start Reading Every Day

Think before you speak. Read before you think. — Fran Lebowitz

Mar 11 27K 606



Bernd Wessely in Towards Data Science

Modern Enterprise Data Modeling

How to address the shortcomings of shallow, outdated models and future-proof your...

4d ago 59 1

See more recommendations