Anti-Privacy Anti-Patterns

Sarah Harvey
BSidesSF 2019

@worldwise001

Hi, welcome to my talk. Thank you all for coming to BSides! As one of the organizers, it is really exciting to see so much of the SF security community come out to listen to and support all of these lovely talks.

Today we are going to talk about privacy. My name is Sarah, you just heard my bio, but in case you missed it, I'm currently a privacy engineer at Square, with about 8 years combined of industry and academic experience.

Some of you are into twitter. Please feel free to take pictures of slide and tweet.

If you’re not sure you have the right twitter account, here you go.

What is Privacy?

I think it's important first to establish here what I mean by privacy, so that we're all on the same page. Specifically I think the term privacy is extremely fluid and can have multiple meanings, so I am really narrowing the scope so that it's easier to see the specific problems, and thus we can come up with tangible solutions.

Since I am the speaker, and you're all forced here to listen to whatever I say, I am going to establish my set of assumptions here for the rest of the talk about privacy. Specifically…

Privacy is about control of data about an individual

https://reprodux.com/record-management/information-management-r2-data-control-image-2/

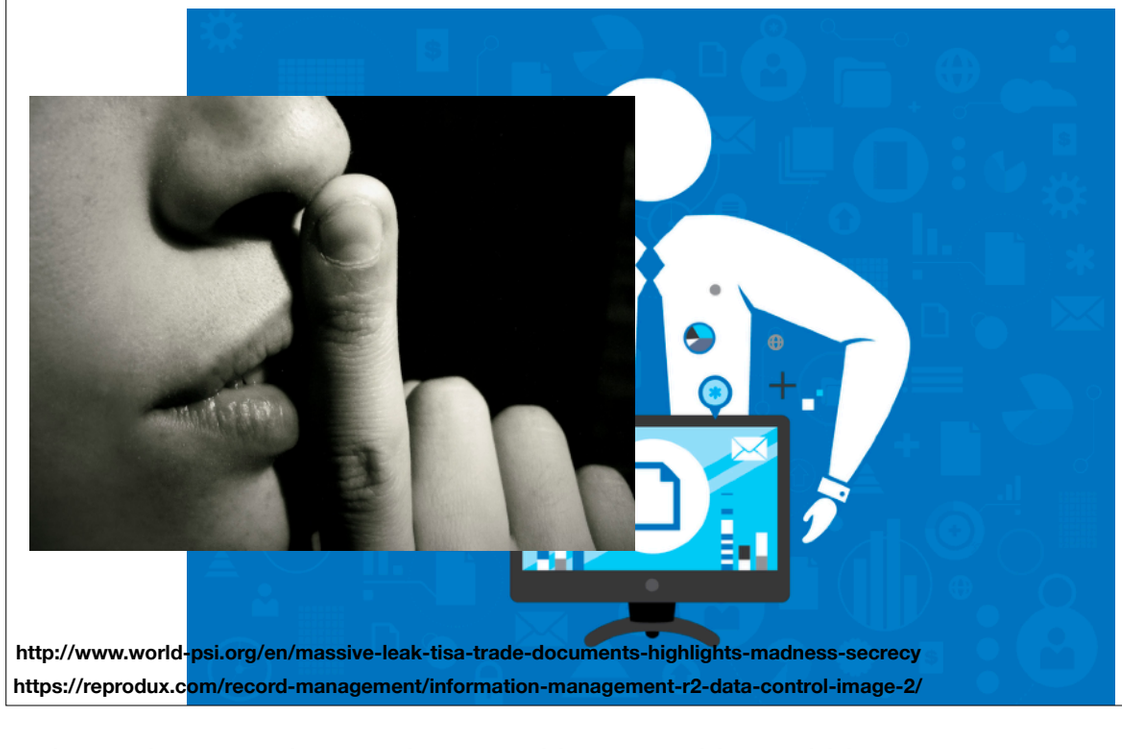
Privacy is about control of data about an individual. So I can say I am more digitally private because I have more control over information about me. This is a much more rigid definition and doesn’t suffer from the subjectivity of what a person considers to be private or not.

Privacy is about control of data about an individual

https://reprodux.com/record-management/information-management-r2-data-control-image-2/

Privacy is different from secrecy. I can be private about my data not because I have anything to hide.
Privacy is different from anonymity. I can be private about my data but still pieces of data identifiable to me, as a specific individual.

Privacy is about control of data about an individual



http://www.world-psi.org/en/massive-leak-tisa-trade-documents-highlights-madness-secrecy
https://reprodux.com/record-management/information-management-r2-data-control-image-2/

Privacy is different from secrecy. I can be private about my data not because I have anything to hide.
Privacy is different from anonymity. I can be private about my data but still pieces of data identifiable to me, as a specific individual.

Privacy is about control of data about an individual

https://bitcoinist.com/mit-riffle-anonymity-king/
http://www.world-psi.org/en/massive-leak-tisa-trade-documents-highlights-madness-secrecy
https://reprodux.com/record-management/information-management-r2-data-control-image-2/

Privacy is different from secrecy. I can be private about my data not because I have anything to hide.
Privacy is different from anonymity. I can be private about my data but still pieces of data identifiable to me, as a specific individual.

Data in scope of privacy includes inferred data

$$\frac{d\,\text{data}}{dx}$$

https://ericbrown.com/the-data-way.htm

Data privacy includes inferred data. This includes behavioral data we derive from the existing data that may be collected about an individual. An example is say, information about my sleep schedule based on when I turn on and off lights. Not just tangible pieces such as my name, my e-mail address, etc.

Privacy is not dead

https://twitter.com/tvgrimreaper

Let’s assume privacy is not fully dead yet.

Everyone has a right to privacy

https://theamericanpolitikos.com/issue/justice-reforms/

Let’s also assume everyone has the right and therefore ability to maintain privacy.

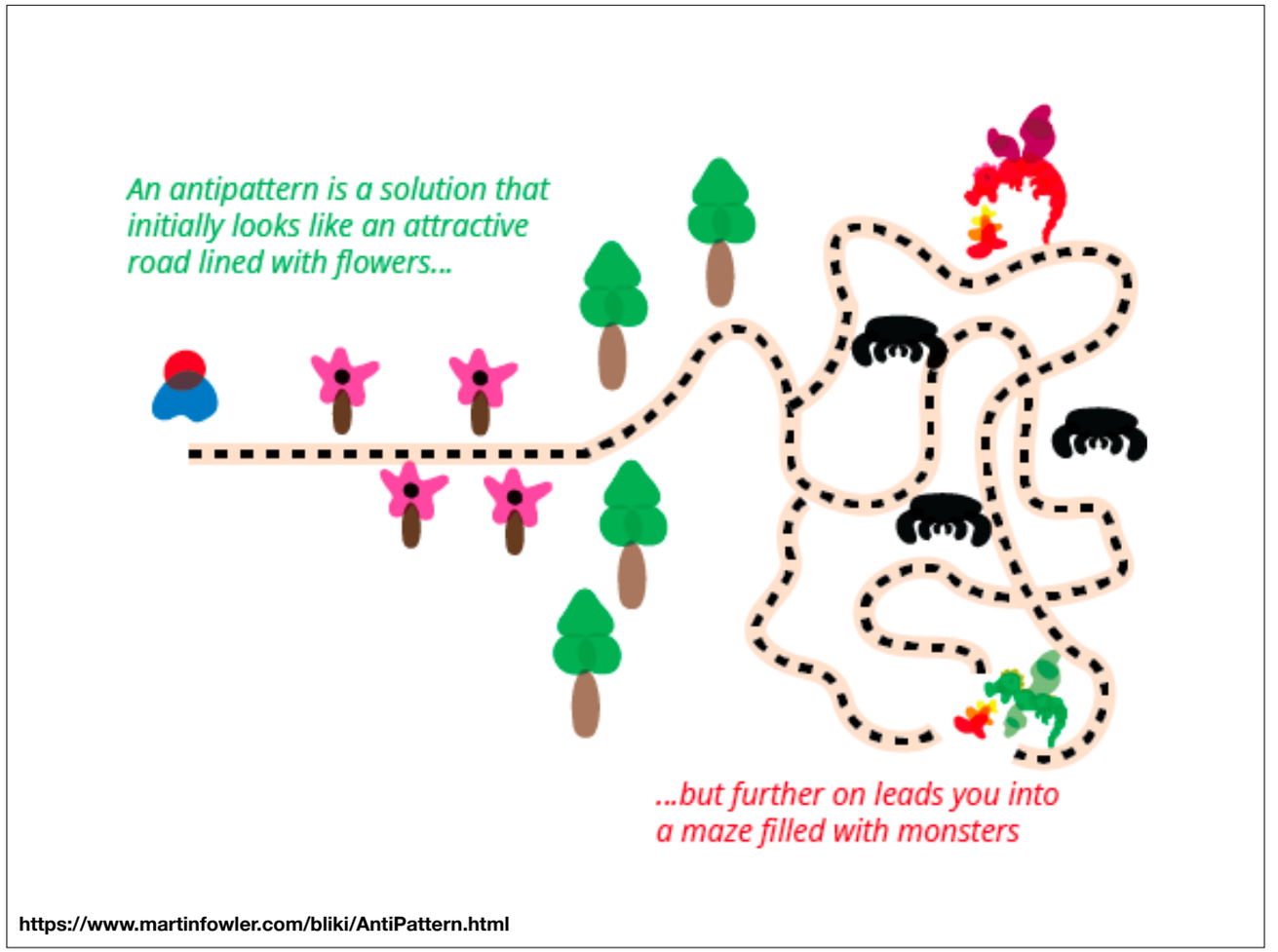
I’m not going to talk about ads

https://landerapp.com/blog/creative-image-advertising-examples/

This talk is not about ads.

What is a Pattern?

I probably should also define the word pattern.



https://www.martinfowler.com/bliki/AntiPattern.html

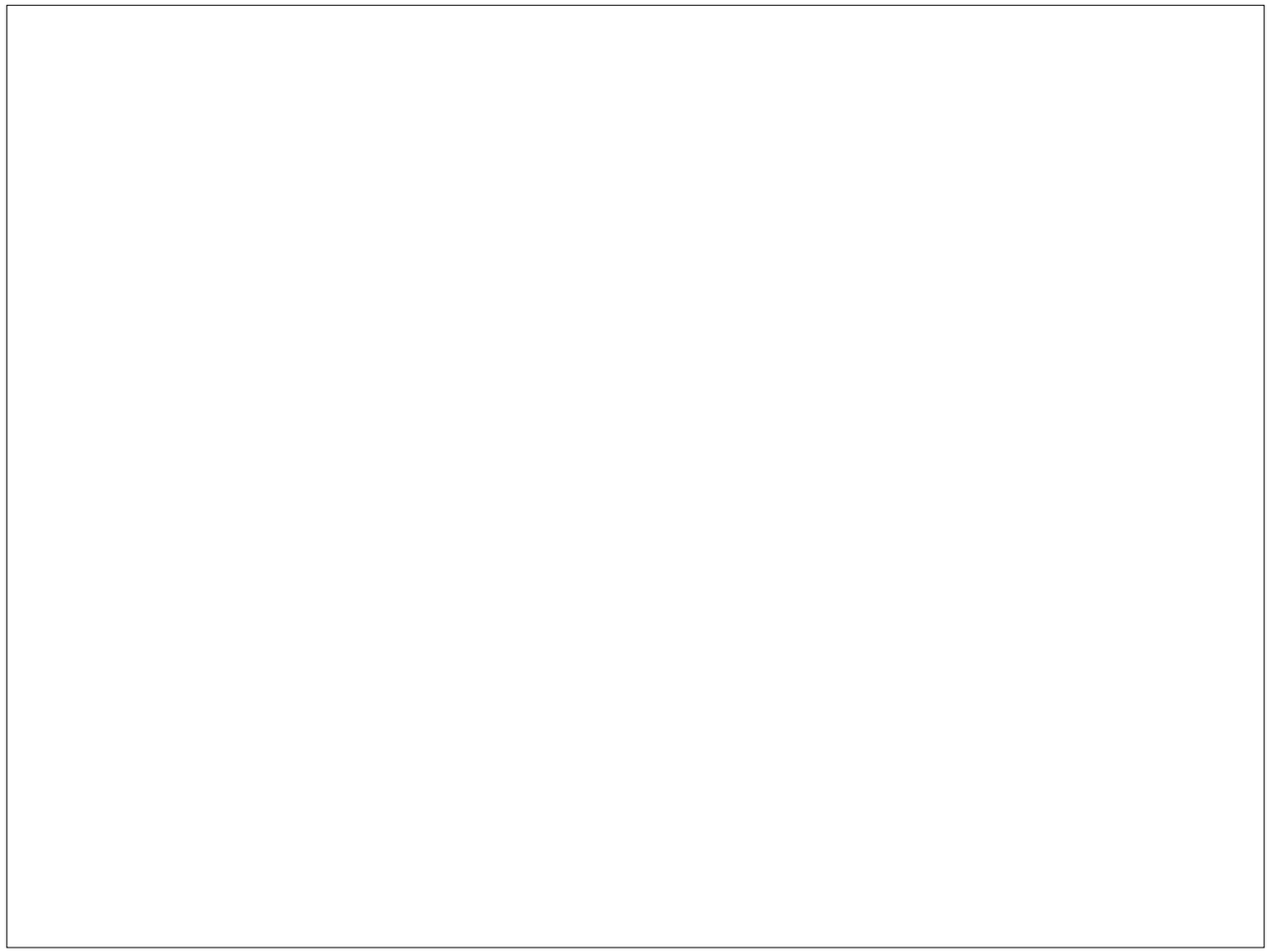
A design pattern is a software engineering phrase. It basically means it’s a behavior you should be using to solve specific problems, because they’re really effective. Therefore Anti-pattern, means that it’s a response to a problem that is counterproductive or ineffective.

https://shop.nationalgeographic.com/products/national-geographic-allanson-globe

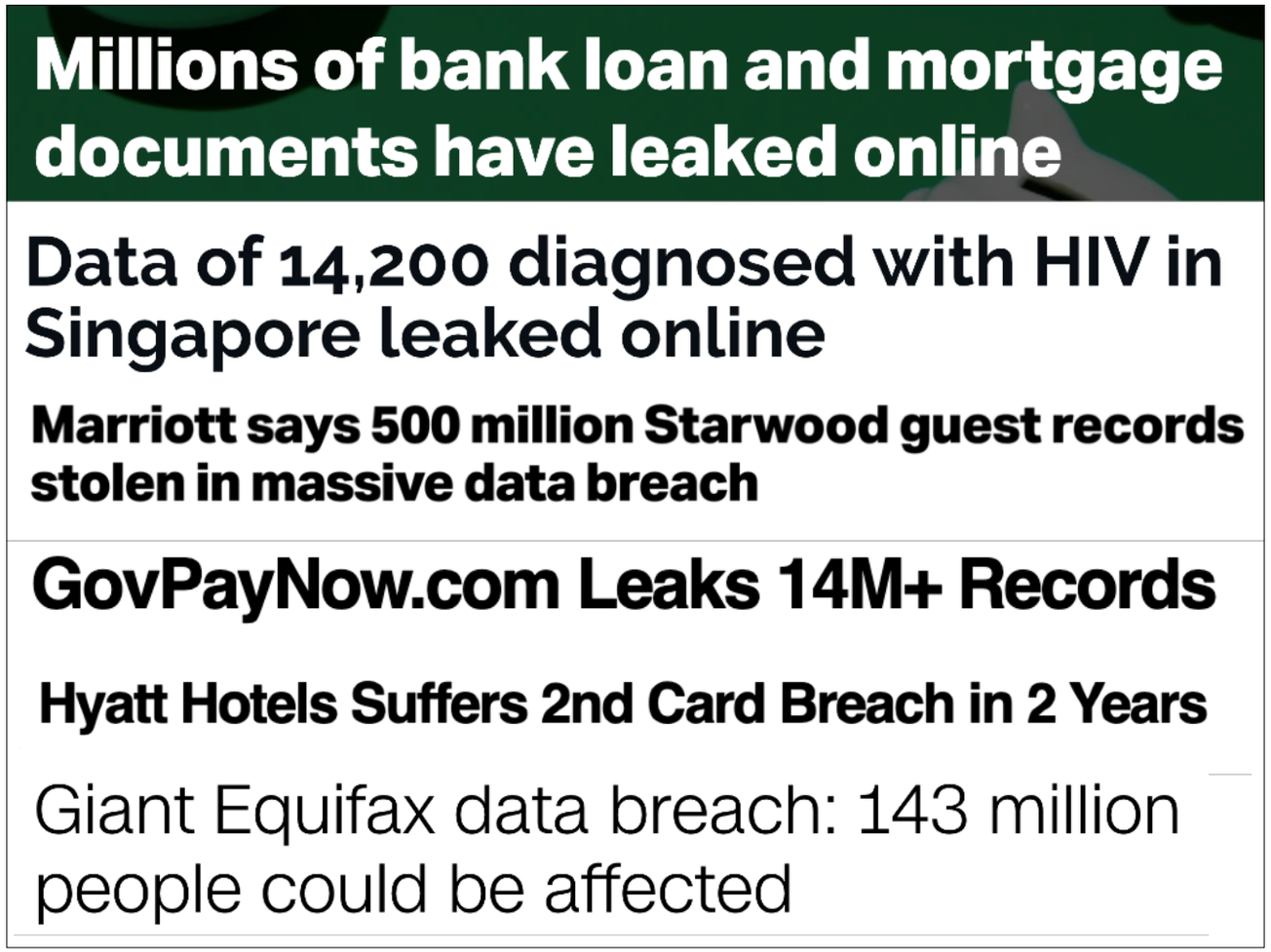
Now that we have had definitions out of the way, let’s talk about the state of the world today

https://en.wikipedia.org/wiki/Cetacean_surfacing_behaviour

Breaches! They are everywhere! Oh that’s not actually the breach I mean.



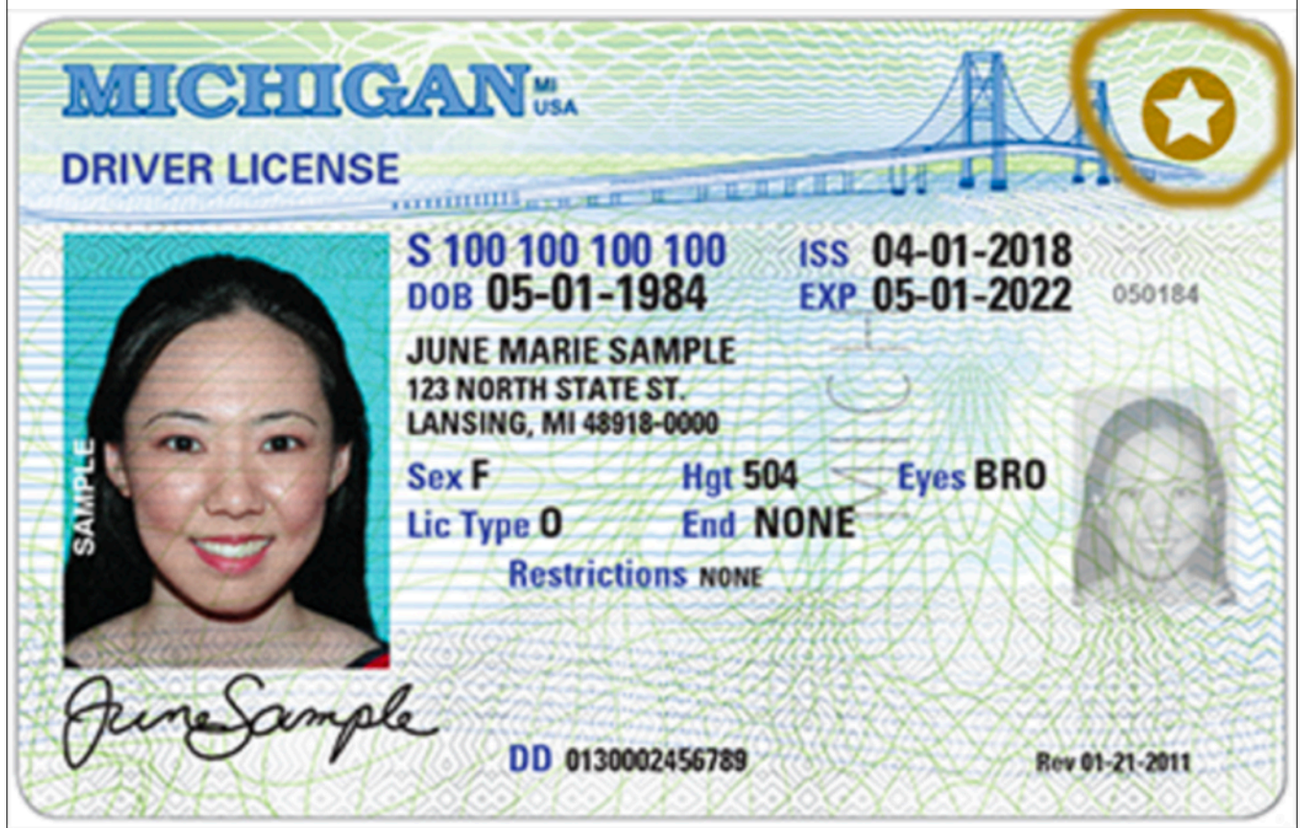
These are the breaches I mean, data breaches.



These are the breaches I mean, data breaches.

https://upgradedpoints.com/instant-approval-credit-cards

https://www.agweb.com/article/how-to-tell-if-a-social-security-card-is-fake/



https://psmag.com/social-justice/michigan-immigrant-rights-advocates-score-a-major-win-in-the-drivers-license-standoff

The New York Times

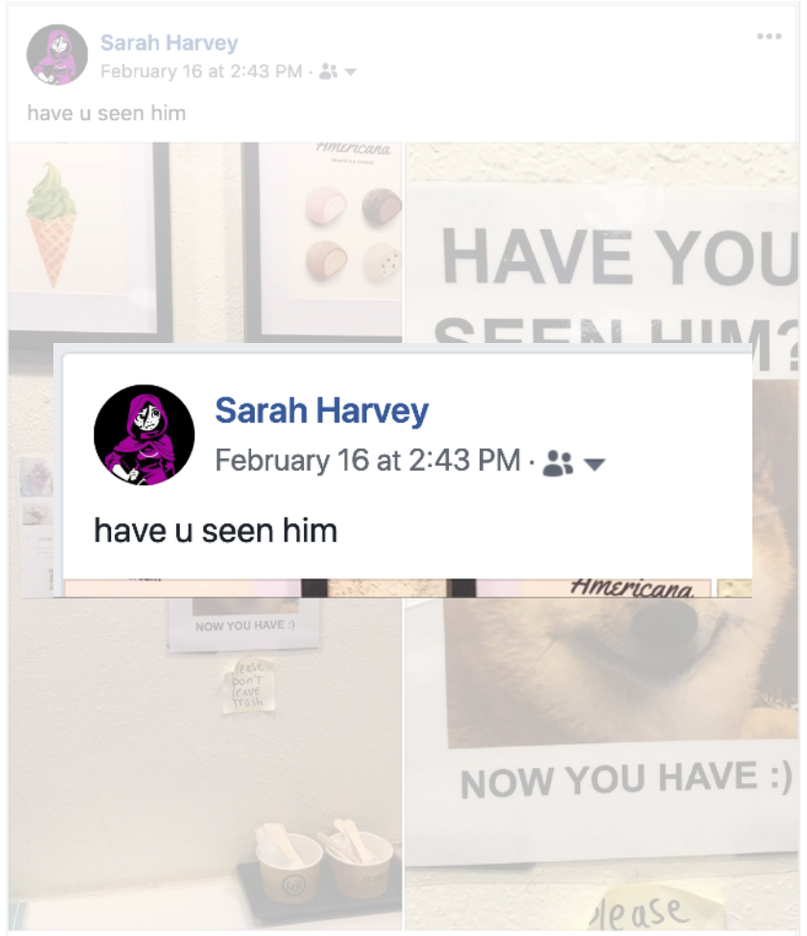
Facebook Says Cambridge Analytica Harvested Data of Up to 87 Million Users

So then what do we do with news like this?

- Data?
- ????
- My shitposts?
- ???????

What was taken?

- Data?
- ????
- My shitposts?
- ???????



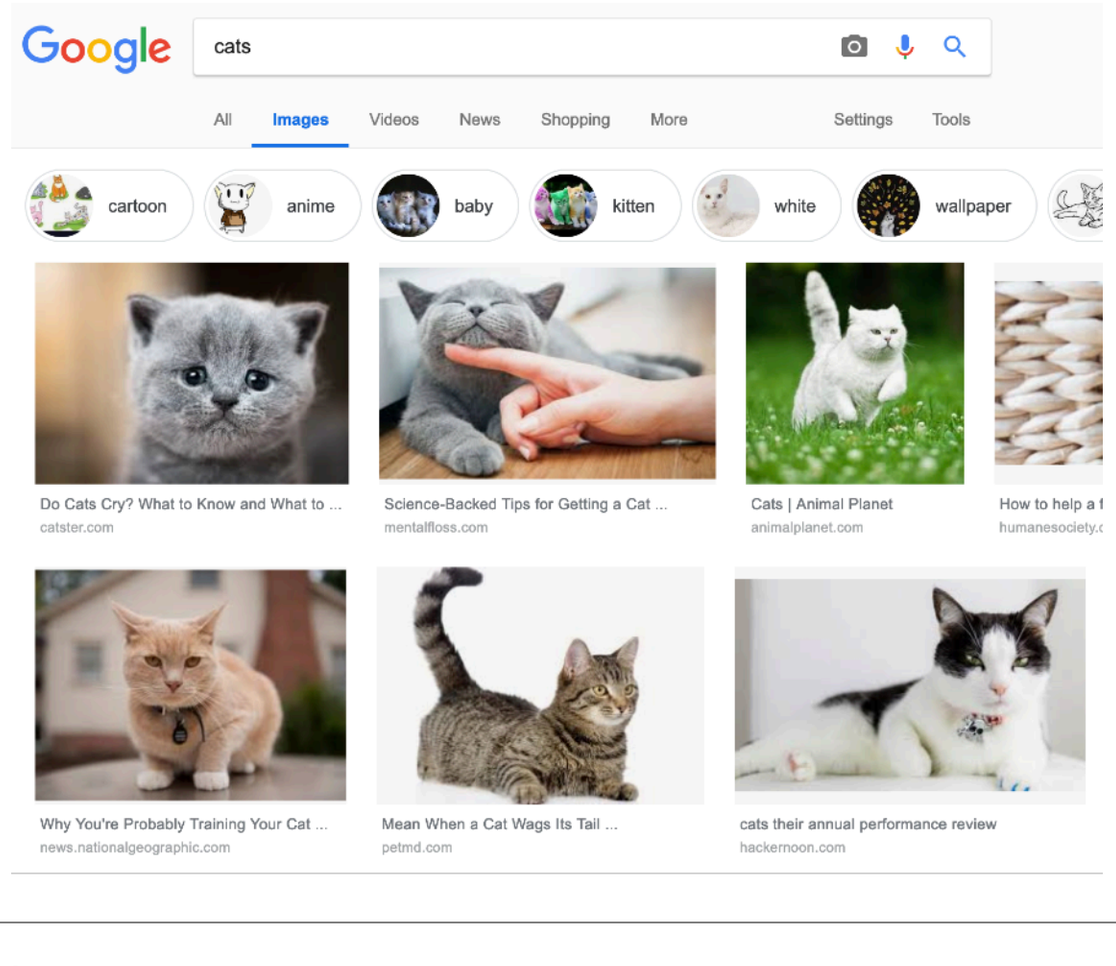
What was taken?

Do I care? Actually I do. I don’t understand what these news mean.

More importantly it still feels bad. It feels wrong. It feels like a violation. It feels like a breach of trust.

To talk about why it feels wrong, we need to talk about the internet.

How do we use the internet?



I like to use the internet to search for pictures of cats.

What is Search?

So what is search?

I come from academia. Search for us is an entire field of research called Information Retrieval.

What is Information Retrieval?

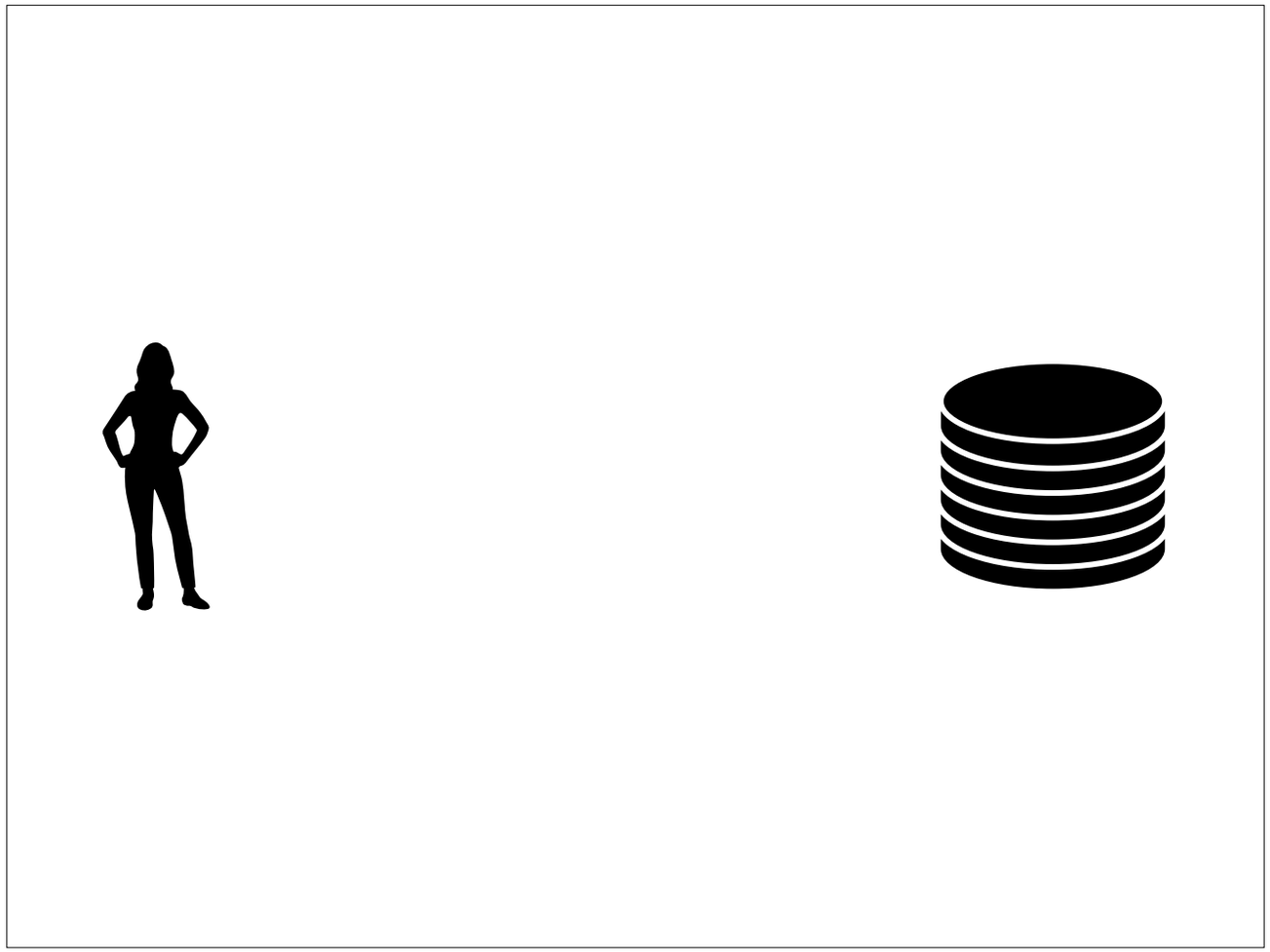
Ok… so what is Information Retrieval?

Simplified, an Information Retrieval system works as follows.

We have a user and a system.

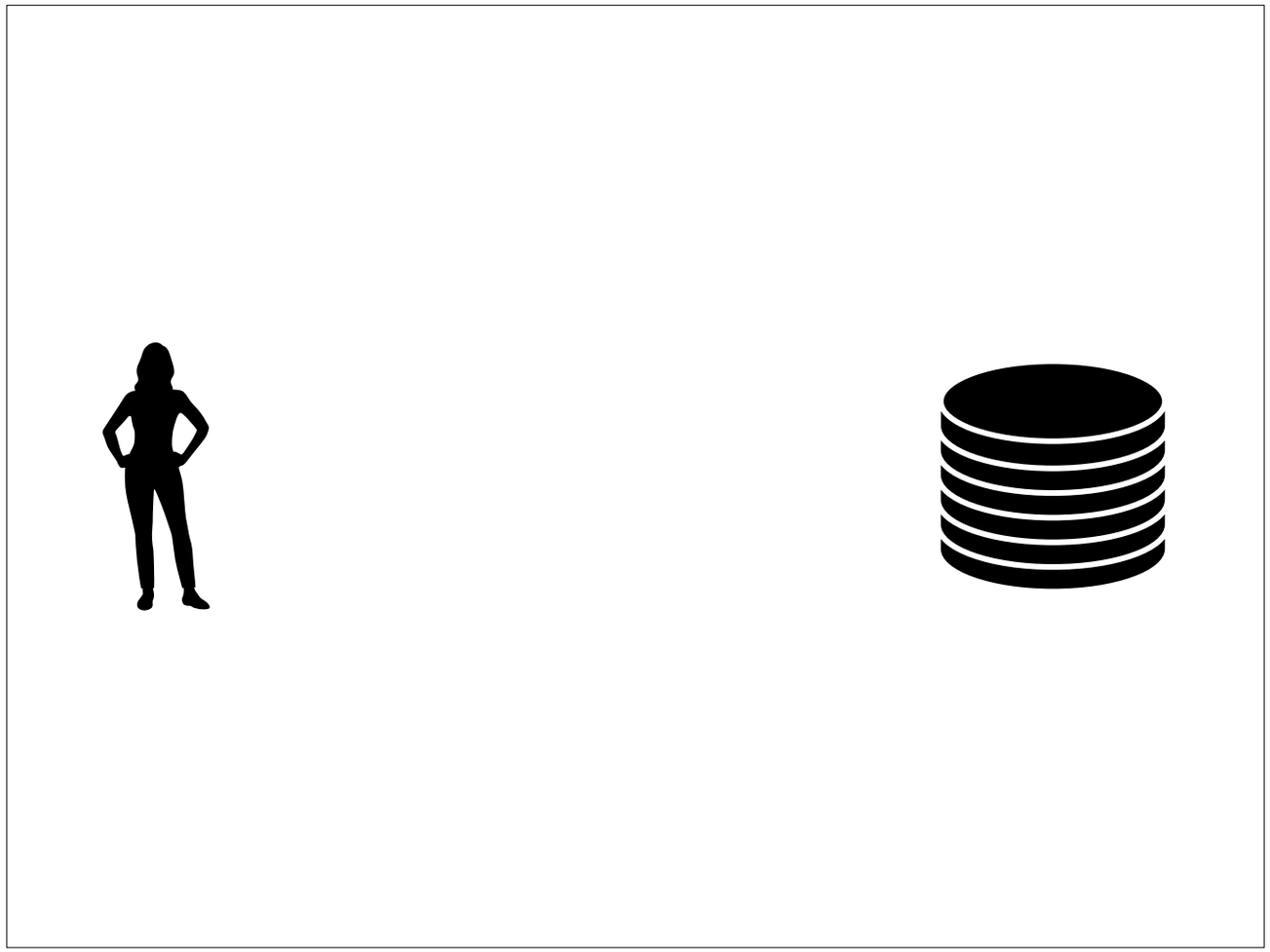
Simplified, an Information Retrieval system works as follows.

We have a user and a system.

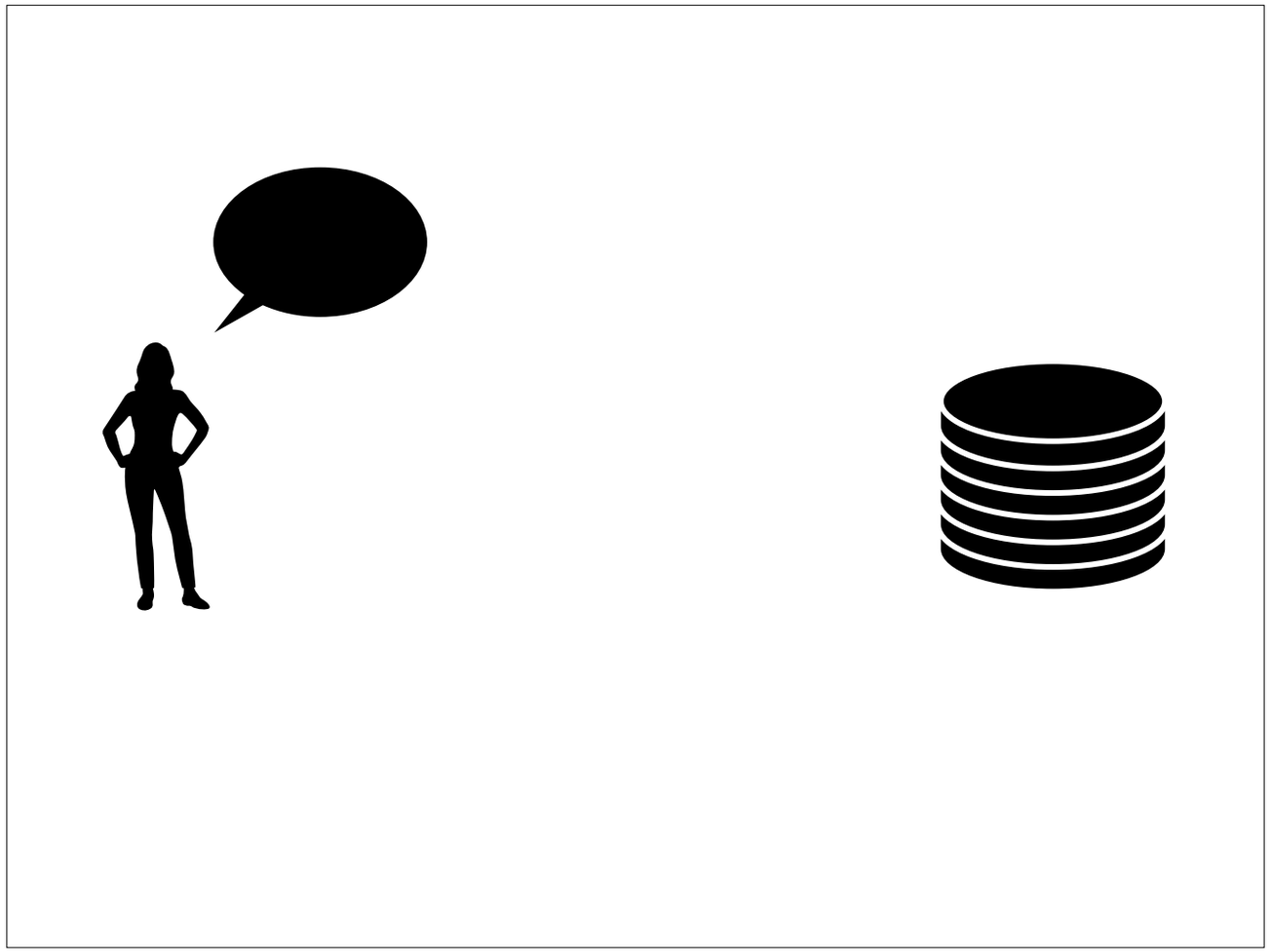


Simplified, an Information Retrieval system works as follows.

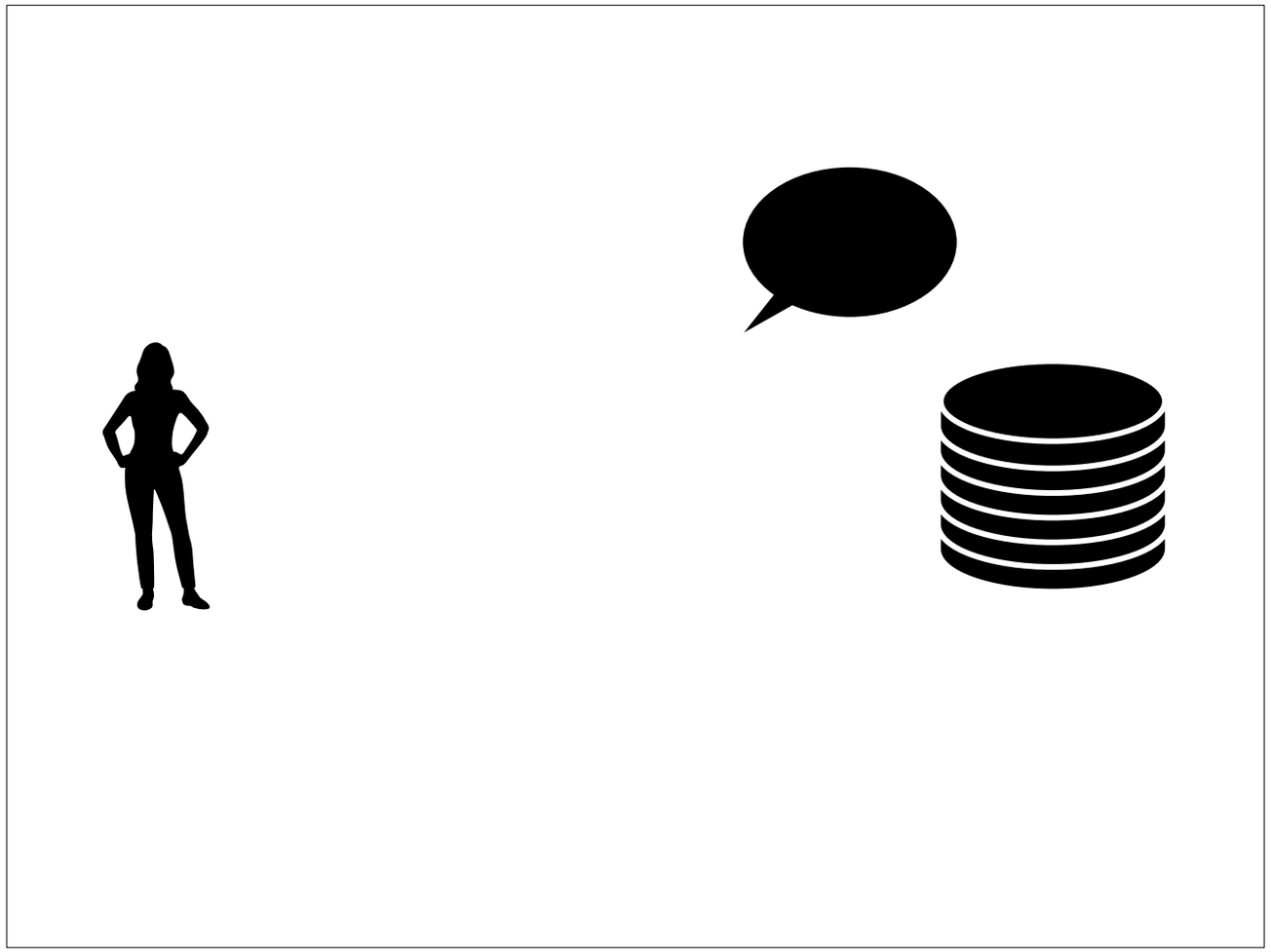
We have a user and a system.



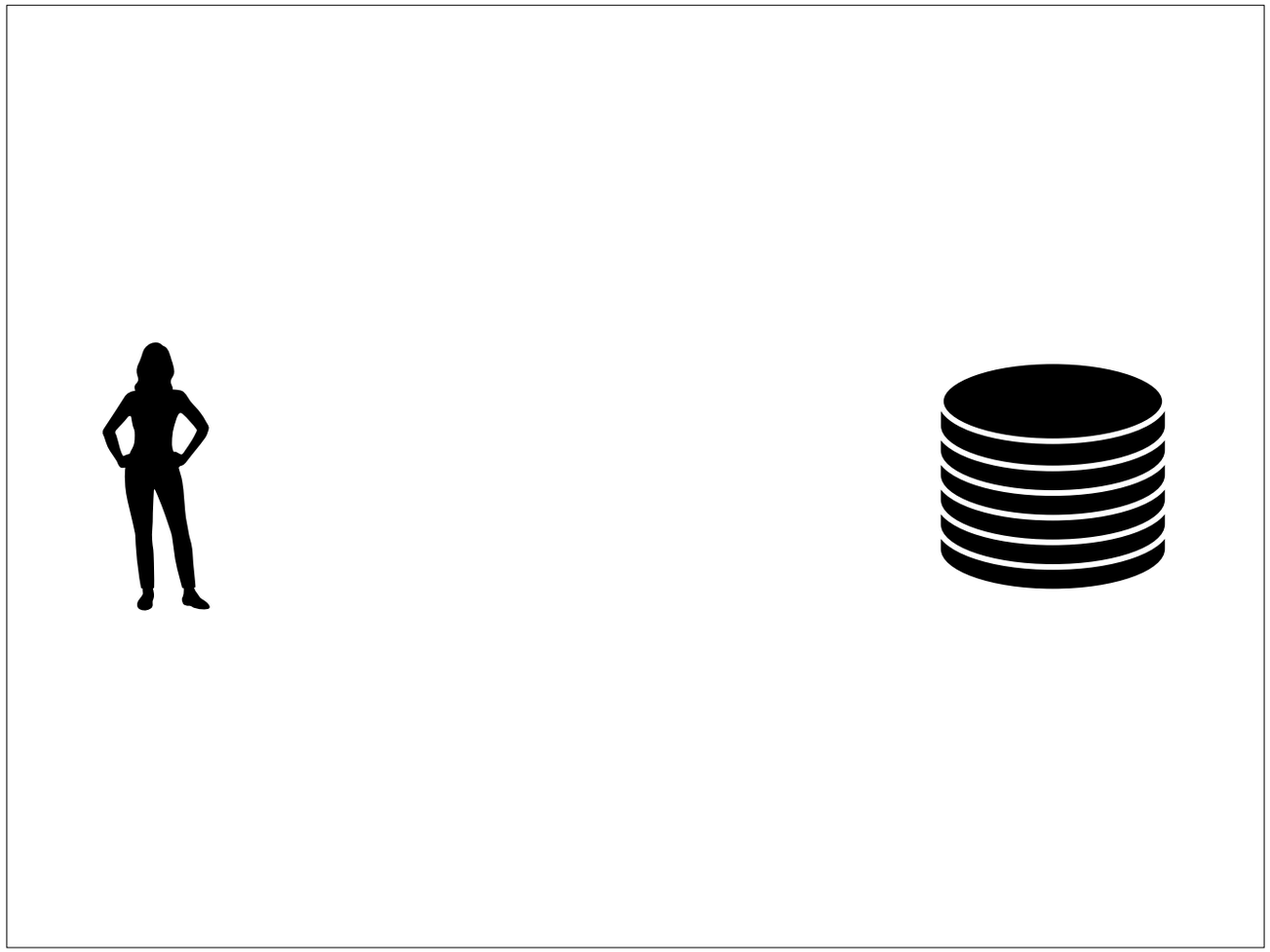
The user makes a query to the system.



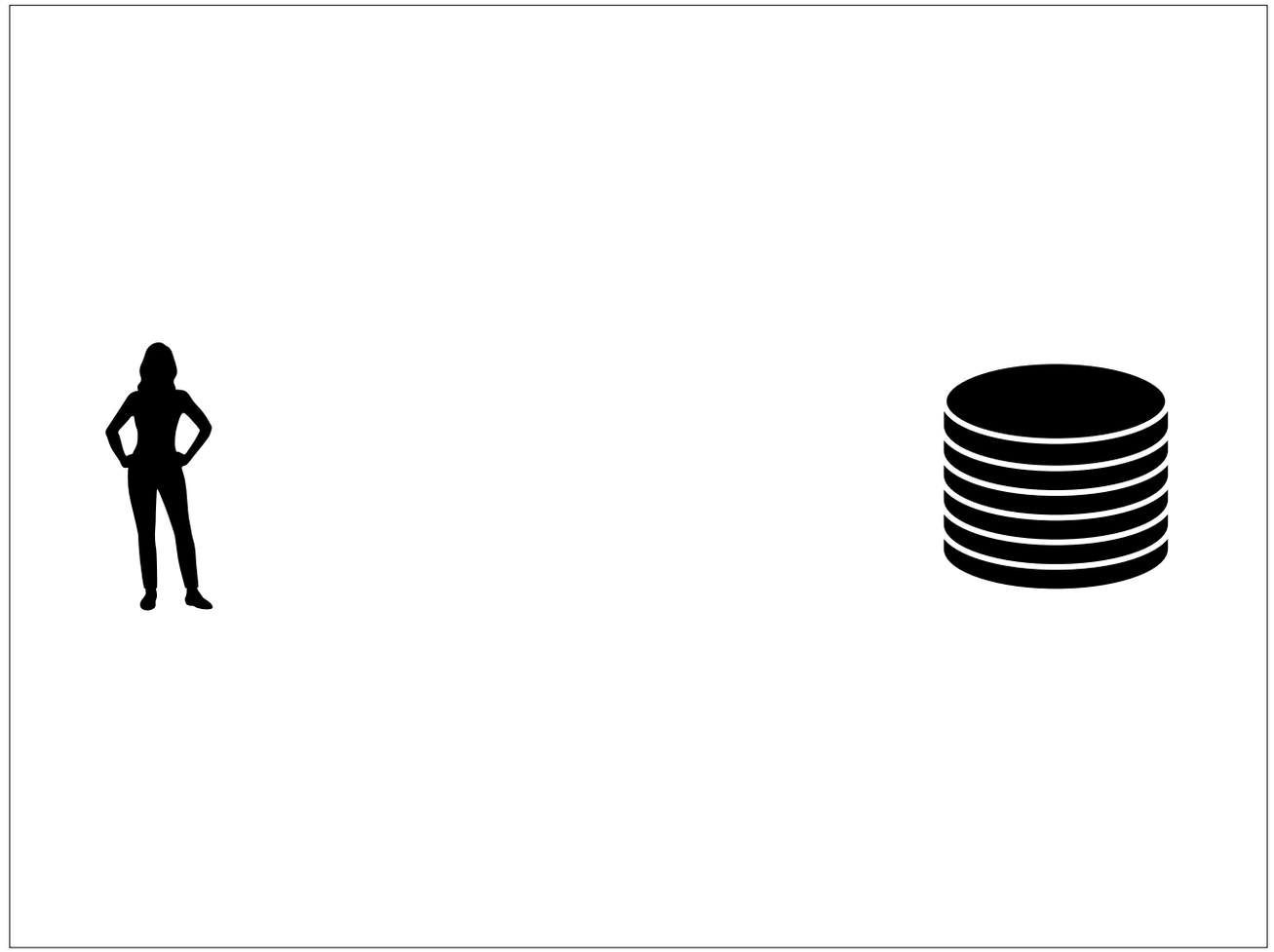
The user makes a query to the system.



The user makes a query to the system.

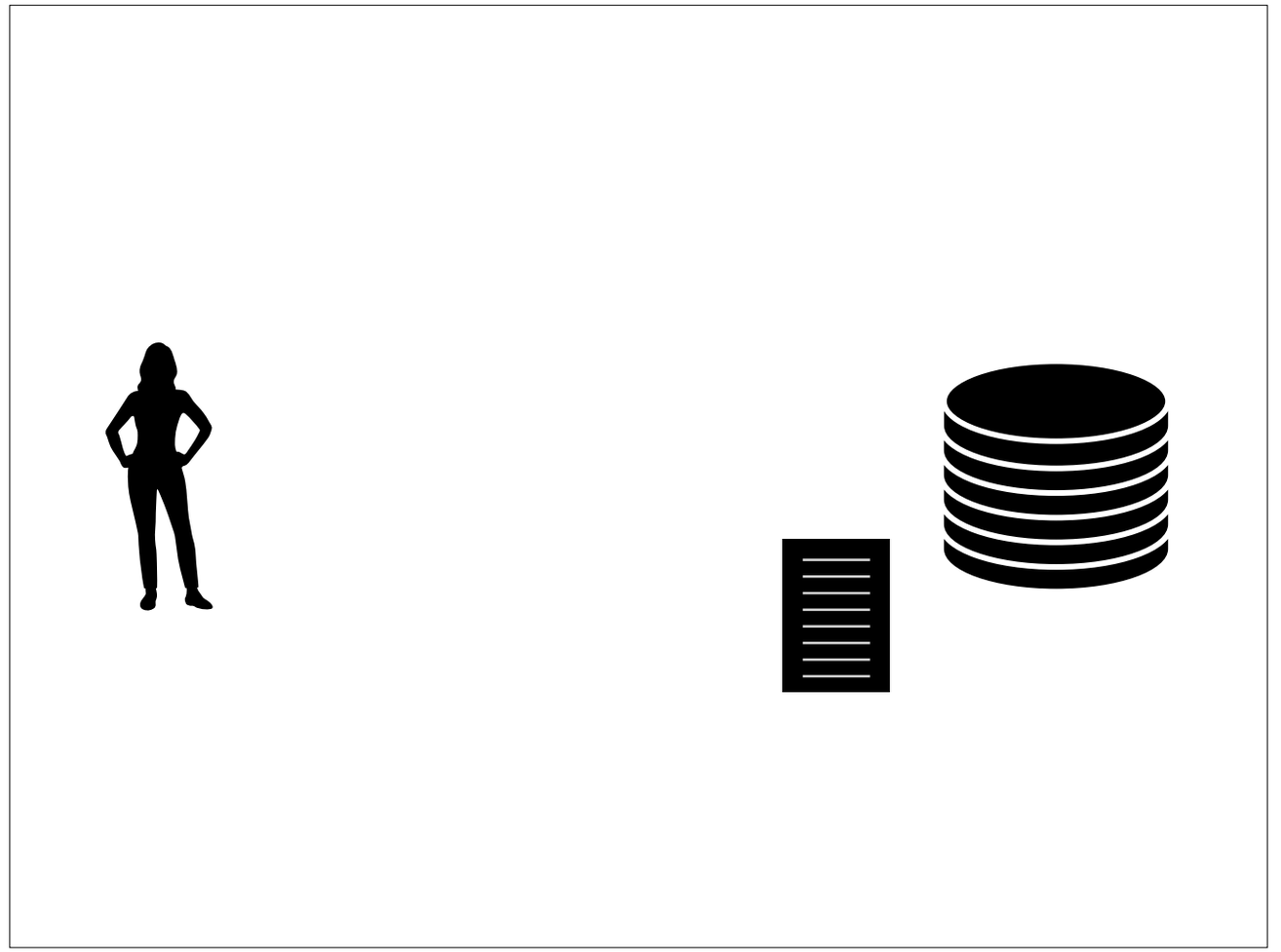


The user makes a query to the system.



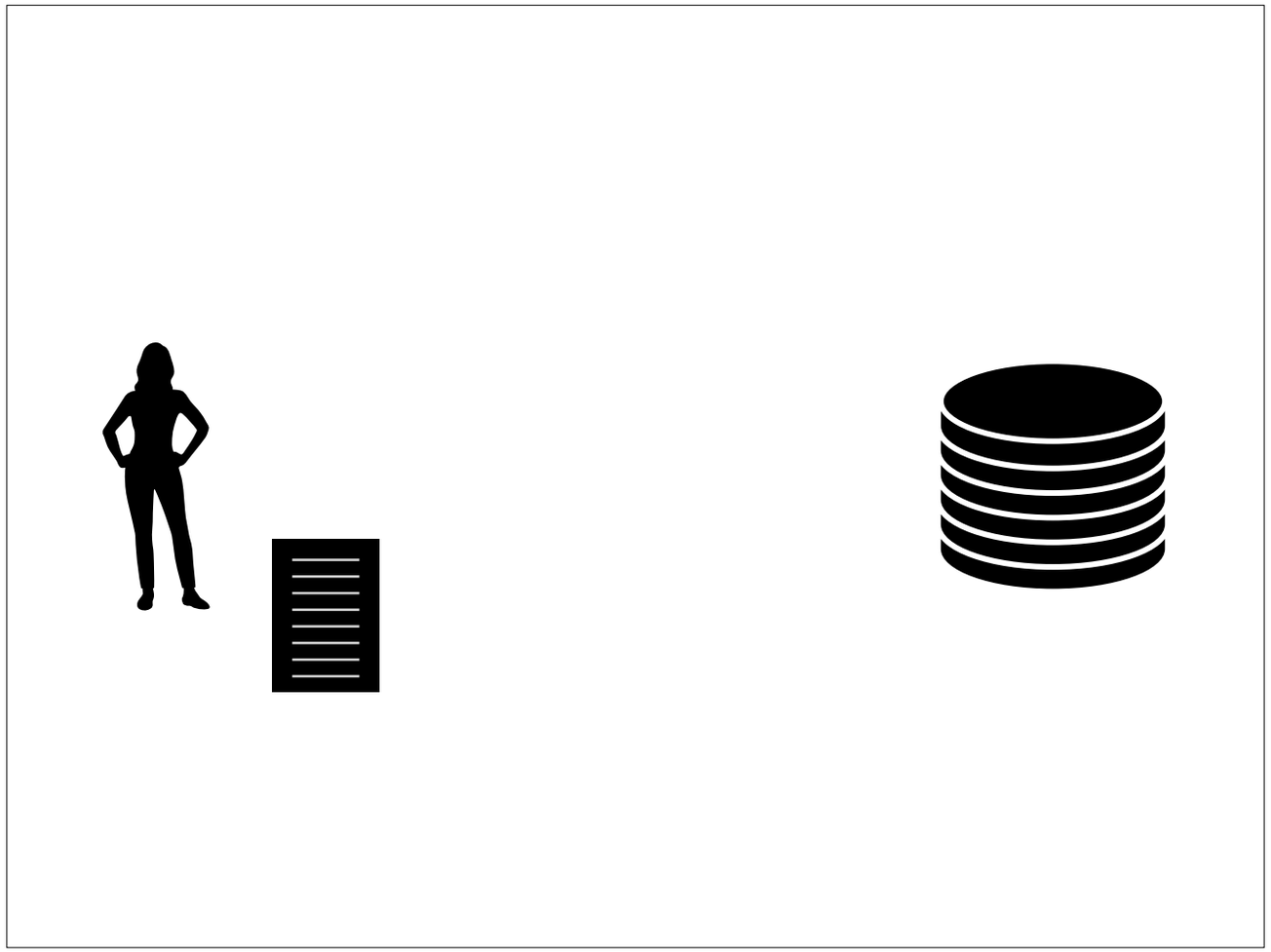
And then the system makes a decision to return a document, or series of documents back.

And a document here could be a webpage, or a some images



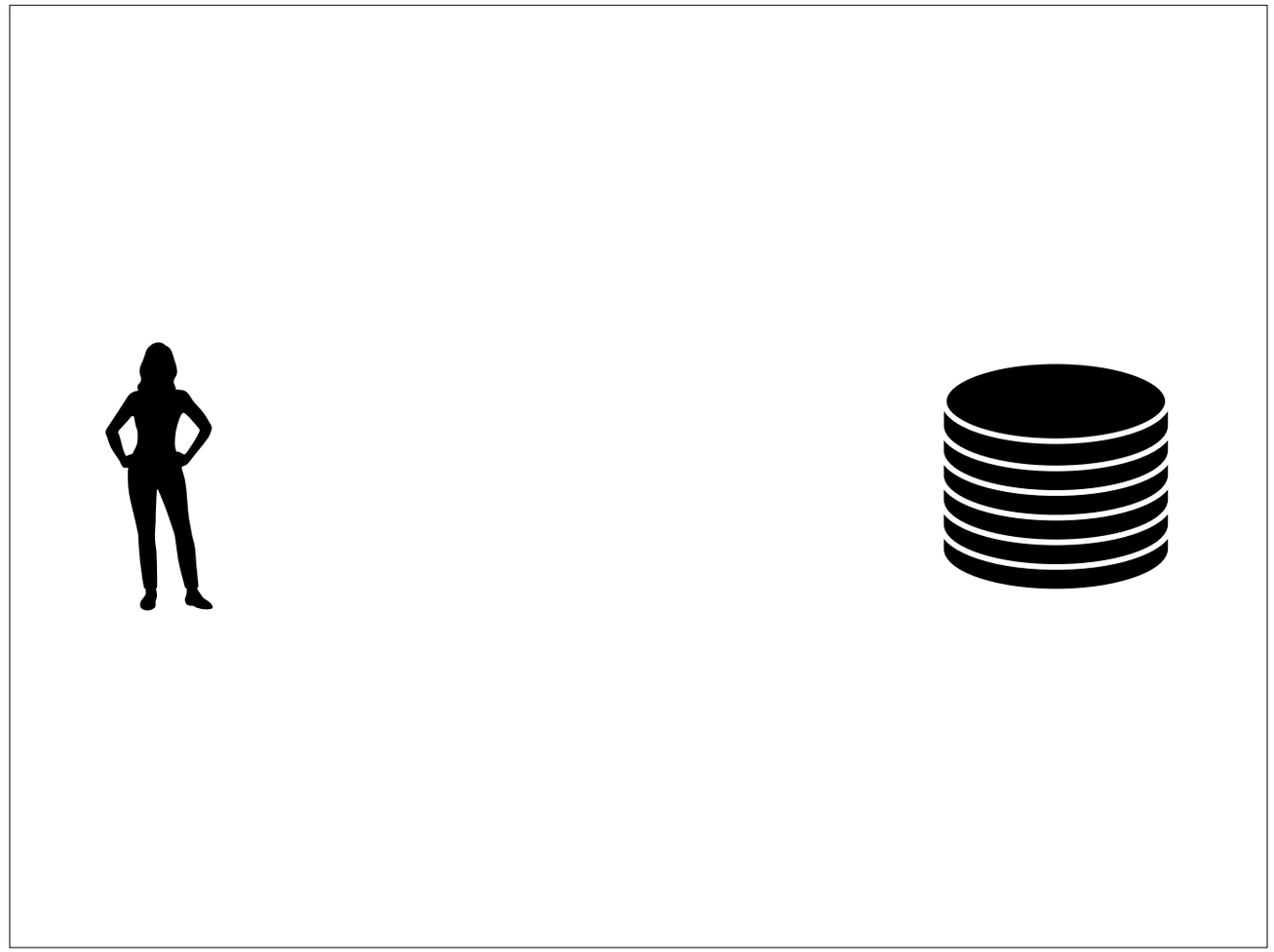
And then the system makes a decision to return a document, or series of documents back.

And a document here could be a webpage, or a some images



And then the system makes a decision to return a document, or series of documents back.

And a document here could be a webpage, or a some images

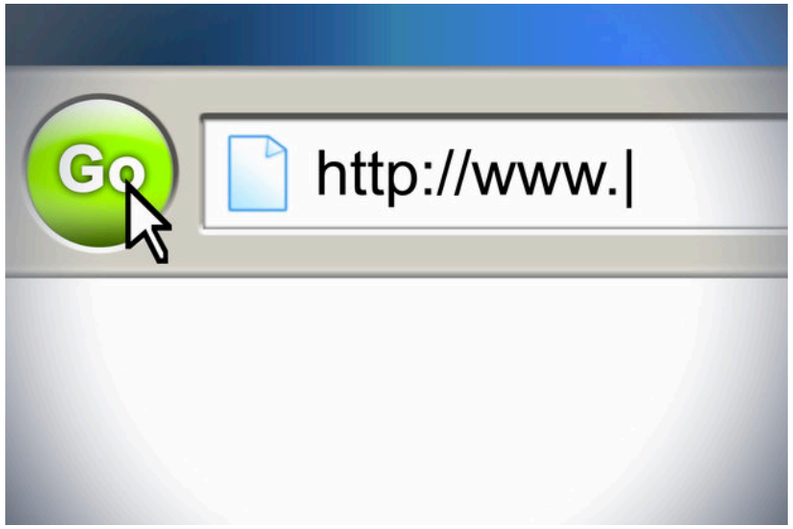


And then the system makes a decision to return a document, or series of documents back.

And a document here could be a webpage, or a some images

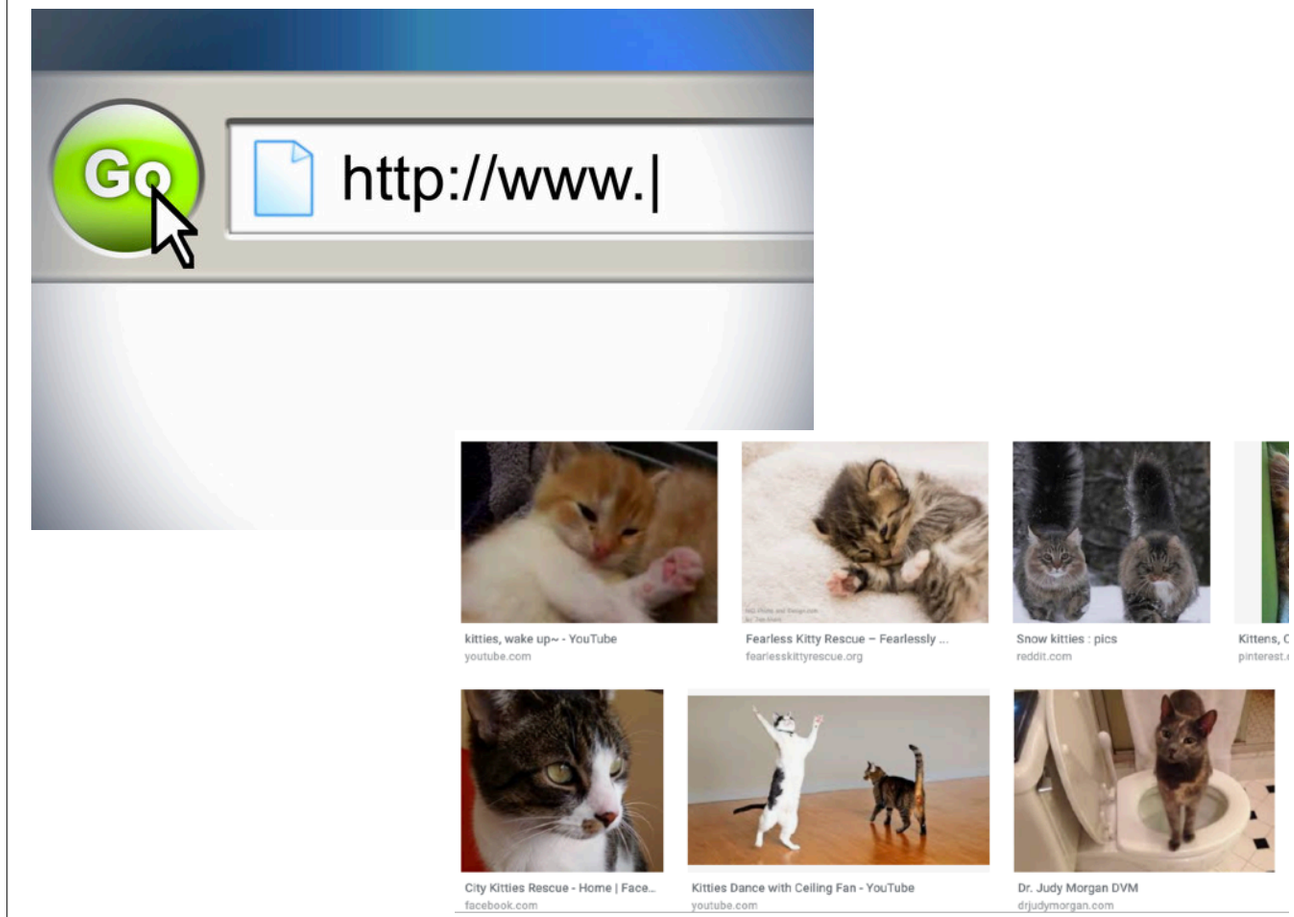
https://www.computerworld.com/article/2992179/why-do-websites-take-so-long-to-load.html

And a document could be anything, it could be a webpage, an image, videos, anything that is stored in the information retrieval system.



https://www.computerworld.com/article/2992179/why-do-websites-take-so-long-to-load.html

And a document could be anything, it could be a webpage, an image, videos, anything that is stored in the information retrieval system.

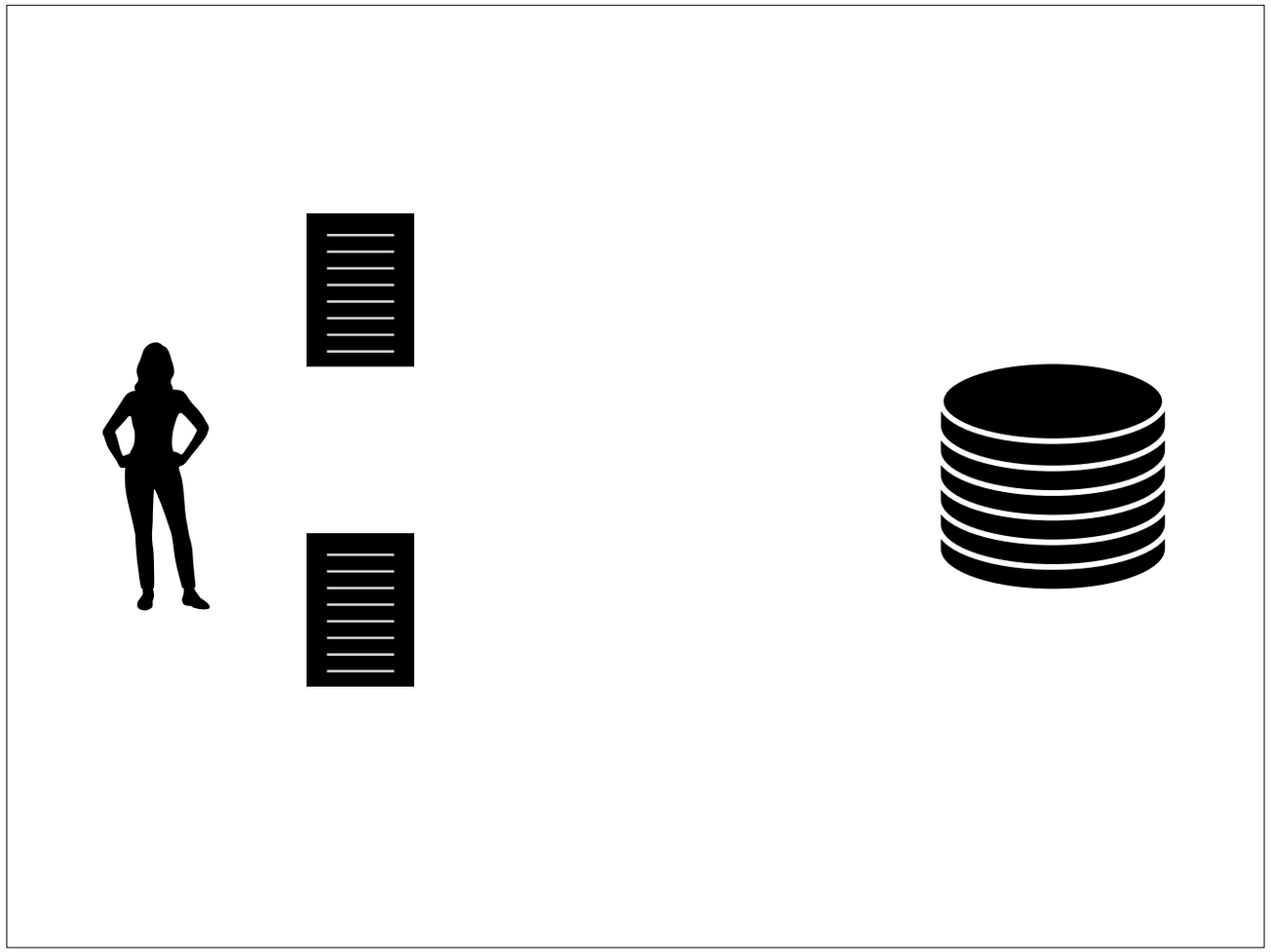


https://www.computerworld.com/article/2992179/why-do-websites-take-so-long-to-load.html

And a document could be anything, it could be a webpage, an image, videos, anything that is stored in the information retrieval system.

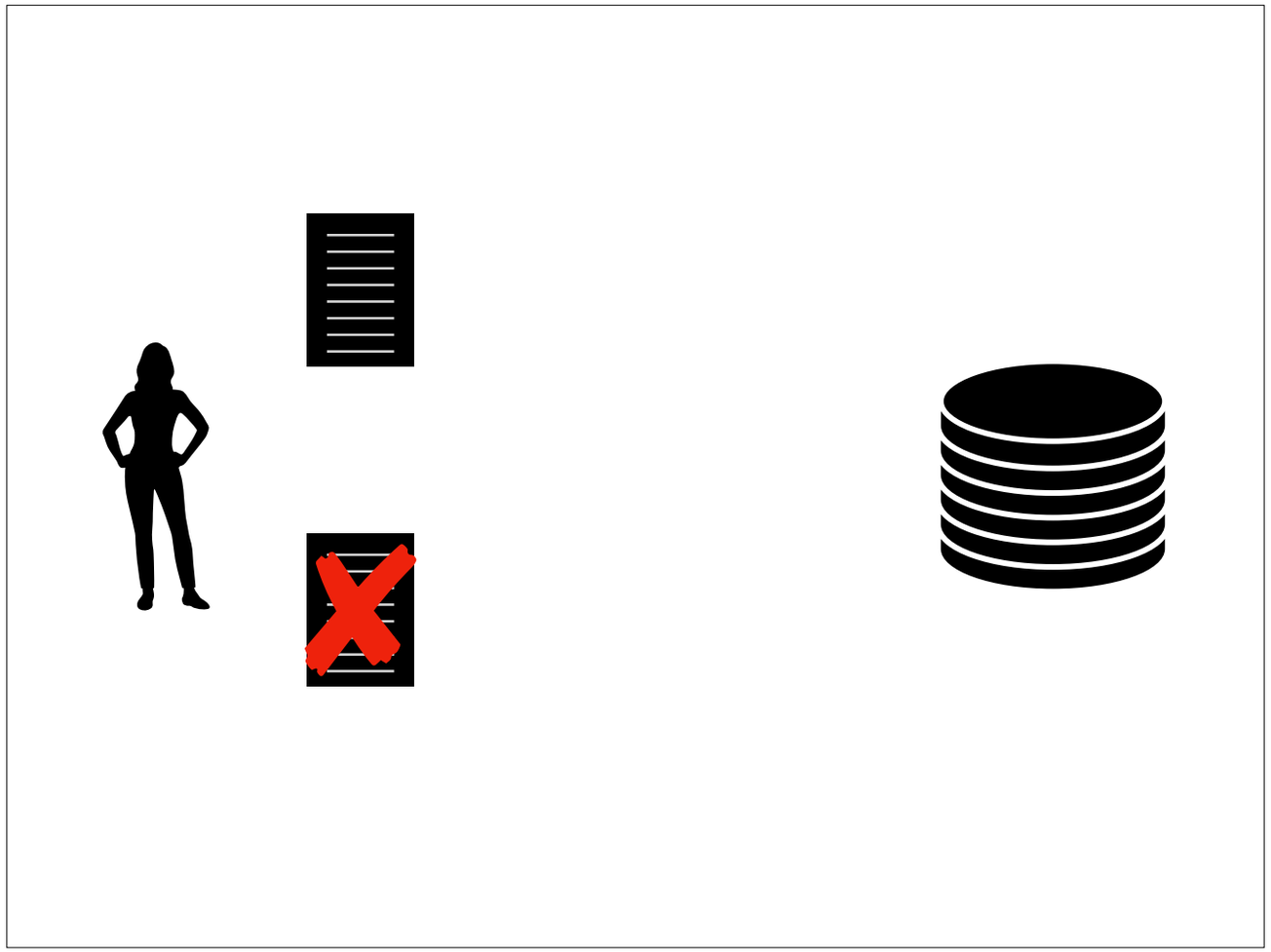
Some of you may say, well this sounds awfully like a database. And you would be right, except for one key difference.

Important in this system is the concept of relevancy. Note that I say relevancy, which is not the same as accuracy. Relevancy is the measure of whether or not the document result is of interest to the user. This means relevancy is subjective! So information retrieval systems, unlike database systems, present a ranked list of documents based on what the system thinks is important to the user.



Some of you may say, well this sounds awfully like a database. And you would be right, except for one key difference.

Important in this system is the concept of relevancy. Note that I say relevancy, which is not the same as accuracy. Relevancy is the measure of whether or not the document result is of interest to the user. This means relevancy is subjective! So information retrieval systems, unlike database systems, present a ranked list of documents based on what the system thinks is important to the user.



Some of you may say, well this sounds awfully like a database. And you would be right, except for one key difference.

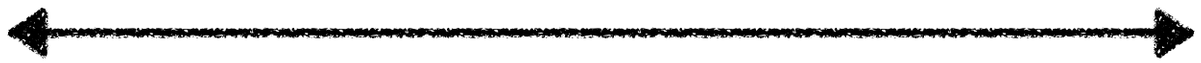
Important in this system is the concept of relevancy. Note that I say relevancy, which is not the same as accuracy. Relevancy is the measure of whether or not the document result is of interest to the user. This means relevancy is subjective! So information retrieval systems, unlike database systems, present a ranked list of documents based on what the system thinks is important to the user.

Information Retrieval

Now Information Retrieval as a field has been around for a long time.

But I think it's most interesting to talk about what's happened in the past 20-30 years.

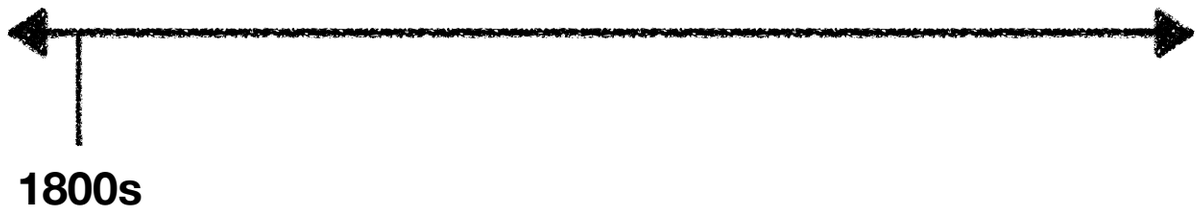
Information Retrieval



Now Information Retrieval as a field has been around for a long time.

But I think it’s most interesting to talk about what’s happened in the past 20-30 years.

Information Retrieval



Now Information Retrieval as a field has been around for a long time.

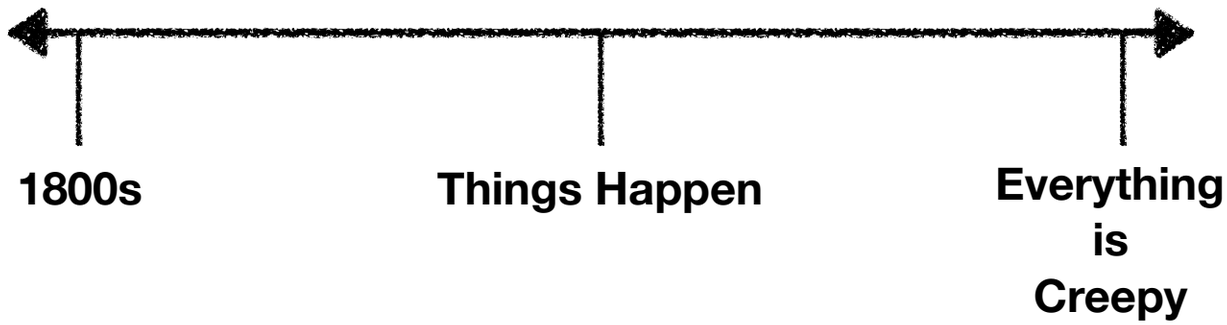
But I think it’s most interesting to talk about what’s happened in the past 20-30 years.

Information Retrieval

Now Information Retrieval as a field has been around for a long time.

But I think it’s most interesting to talk about what’s happened in the past 20-30 years.

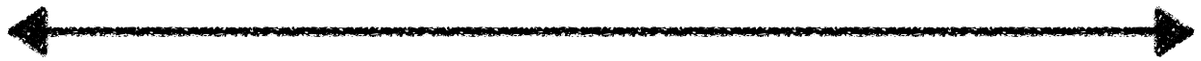
Information Retrieval



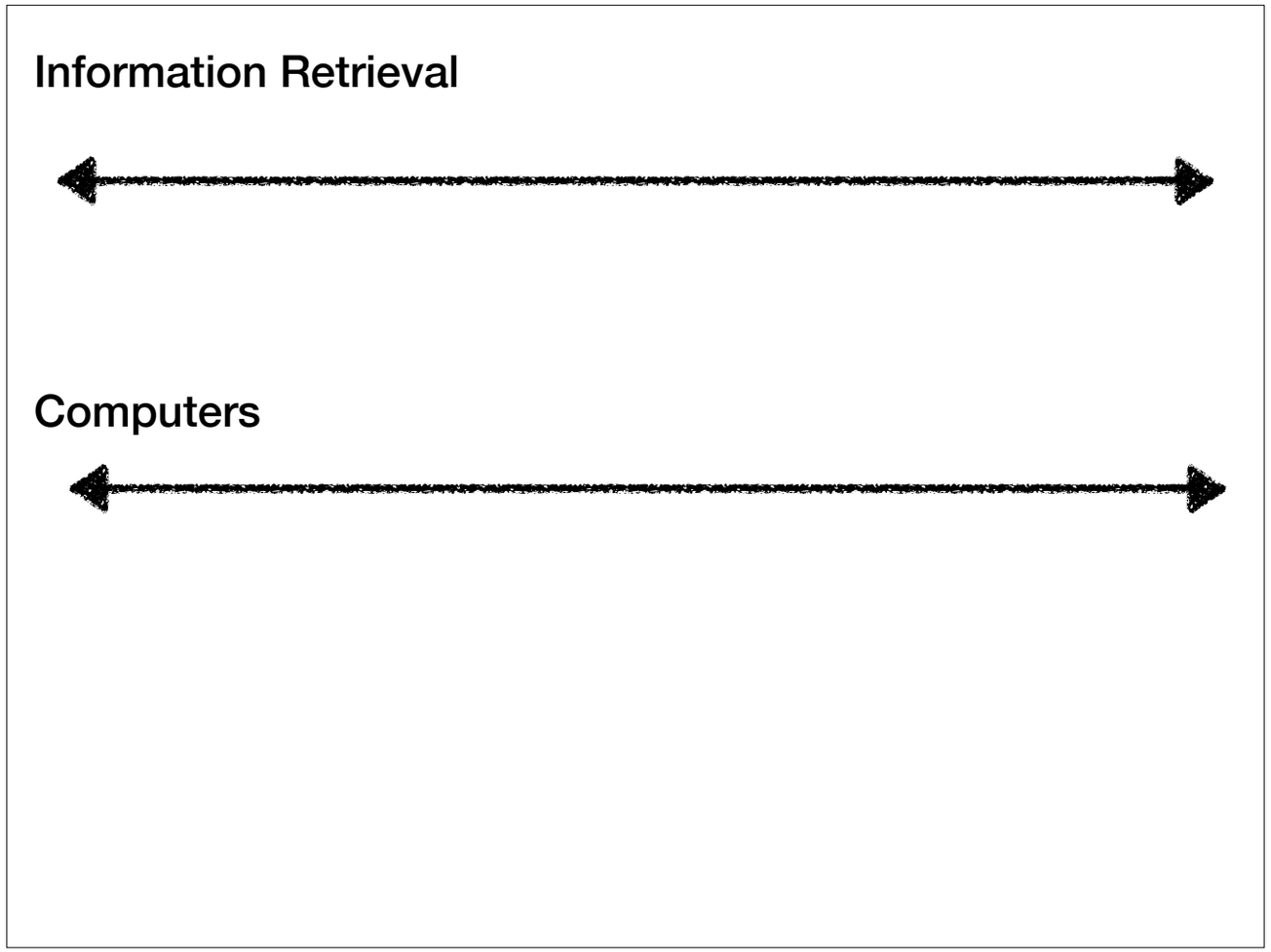
Now Information Retrieval as a field has been around for a long time.

But I think it’s most interesting to talk about what’s happened in the past 20-30 years.

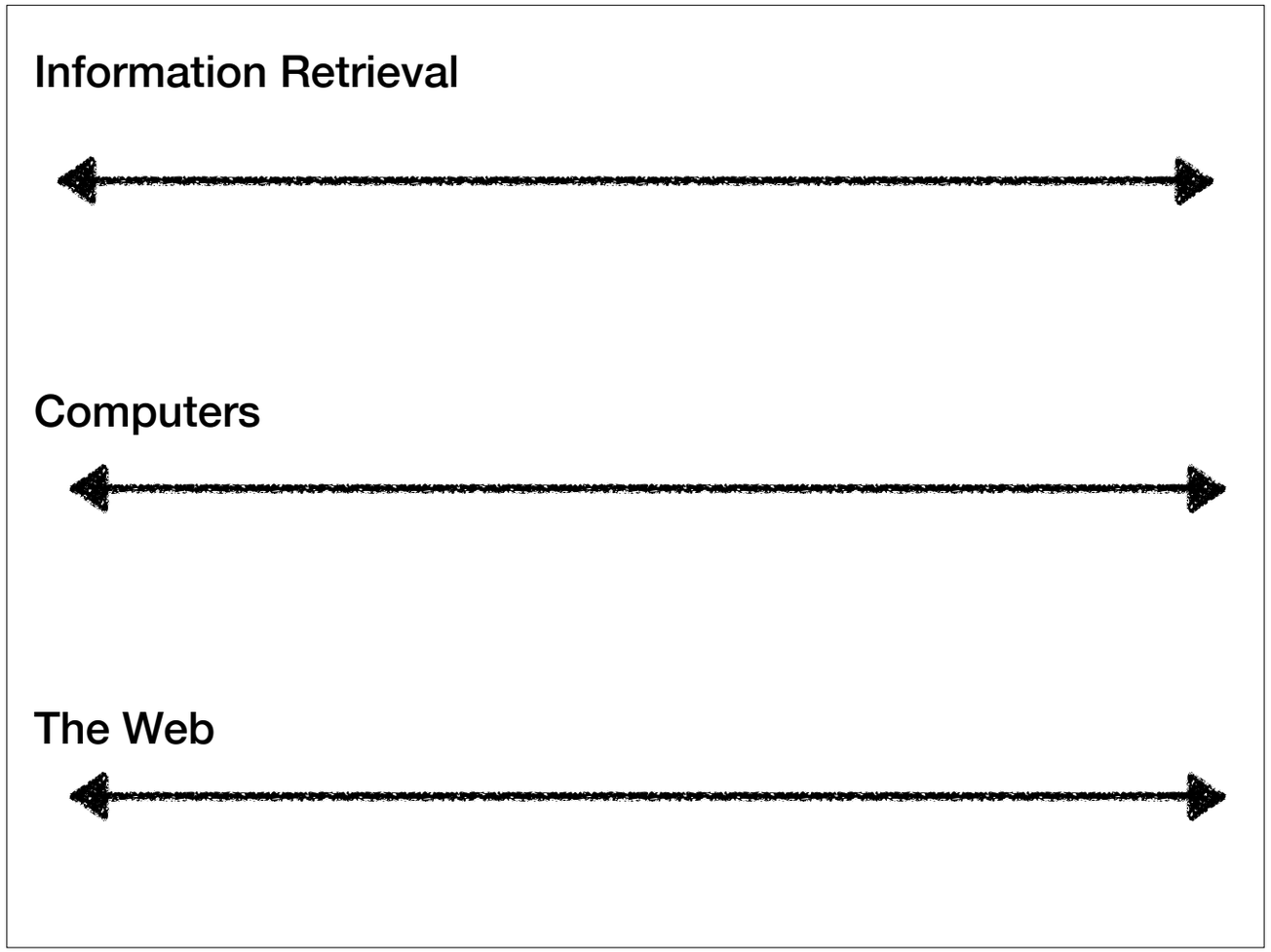
Information Retrieval



We are going to look at some things on the timeline as compared to the state of Computers, and the state of The World, i.e. what everyone else manages to see.



We are going to look at some things on the timeline as compared to the state of Computers, and the state of The World, i.e. what everyone else manages to see.

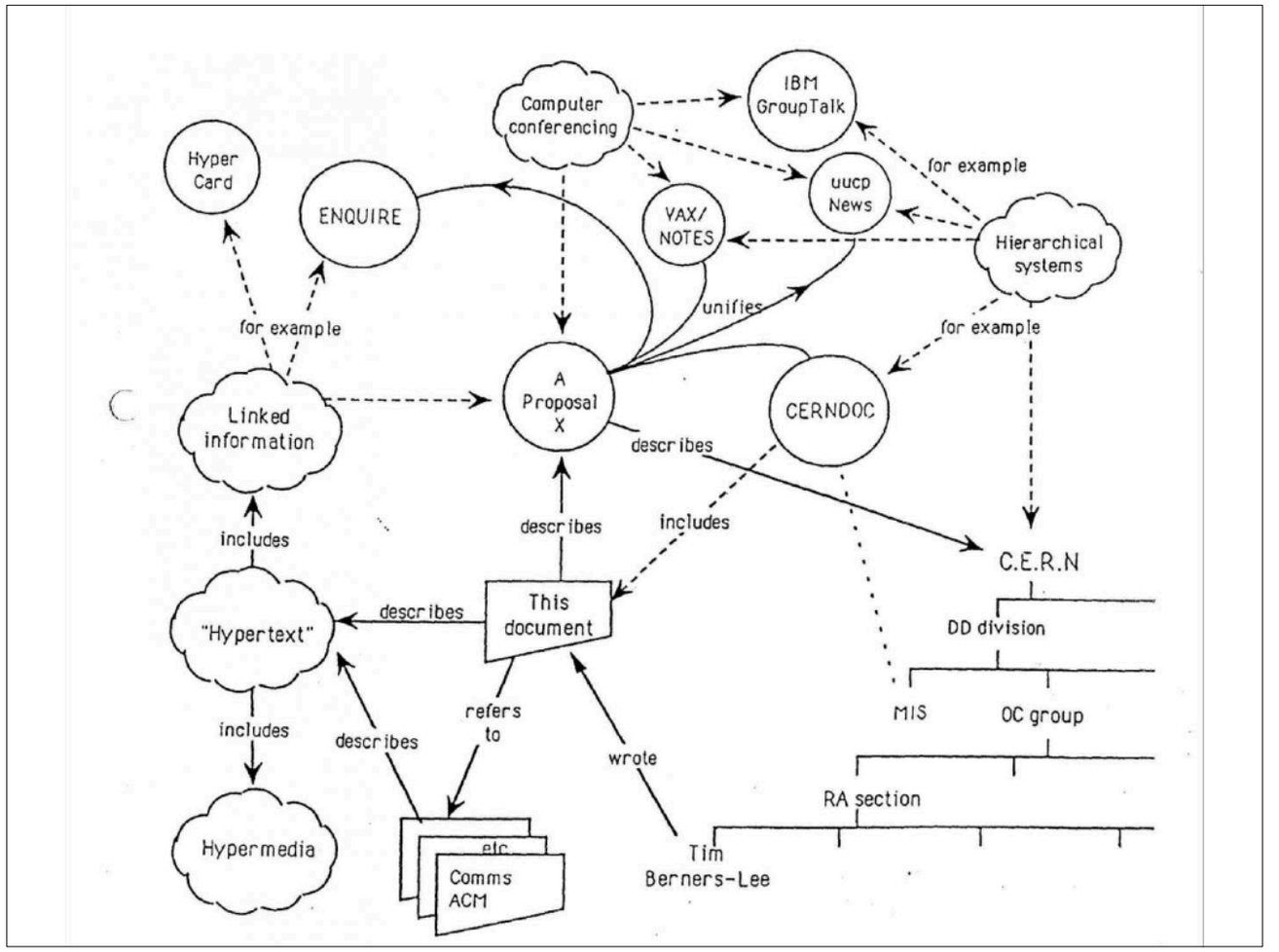


We are going to look at some things on the timeline as compared to the state of Computers, and the state of The World, i.e. what everyone else manages to see.

https://www.indiewire.com/2015/09/7-highlights-from-the-20th-anniversary-celebration-of-hackers-including-sequel-talk-and-fashion-drama-58080/

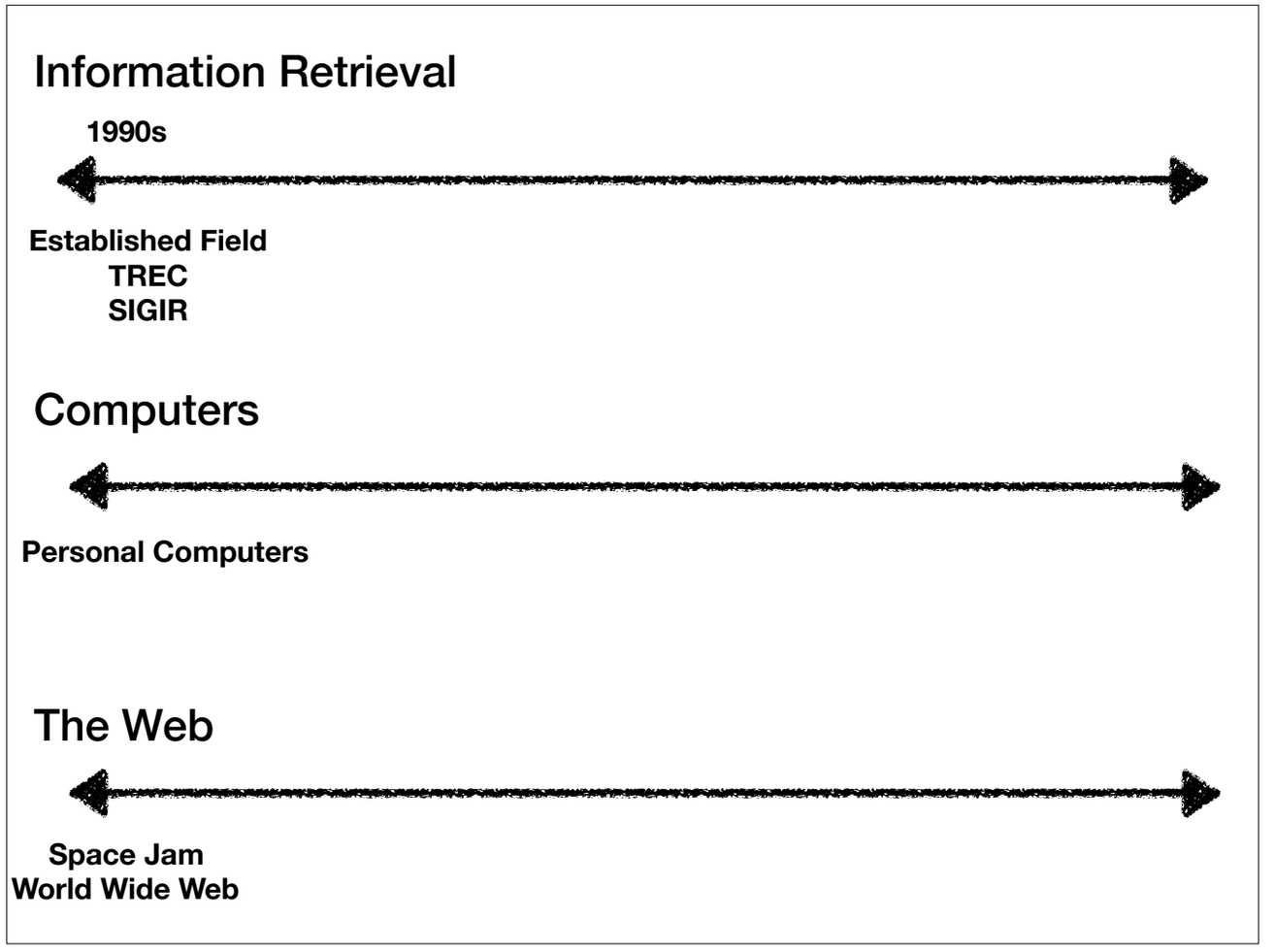
Let’s talk about the 90s. What was happening in the 90s?

Computers were starting to look a bit like this. We were already moving away from the big hulking mainframes of the past, and computing power was rapidly growing with smaller and smaller form factors.



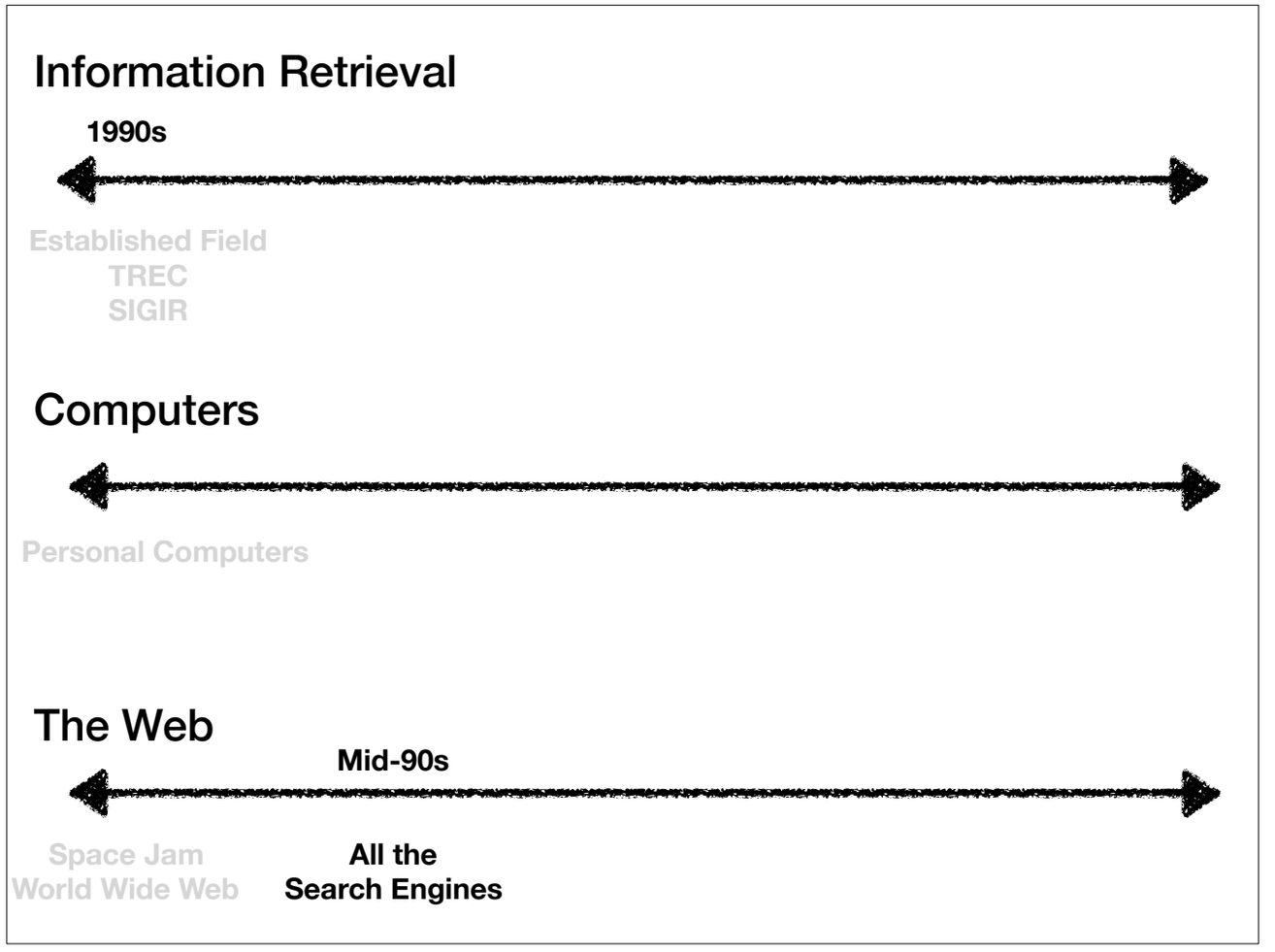
In 1989, Tim Berners-Lee had proposed the idea of the World Wide Web. So by the 90s, we had a really nascent internet or web presence.

For instance, the website for Space Jam.

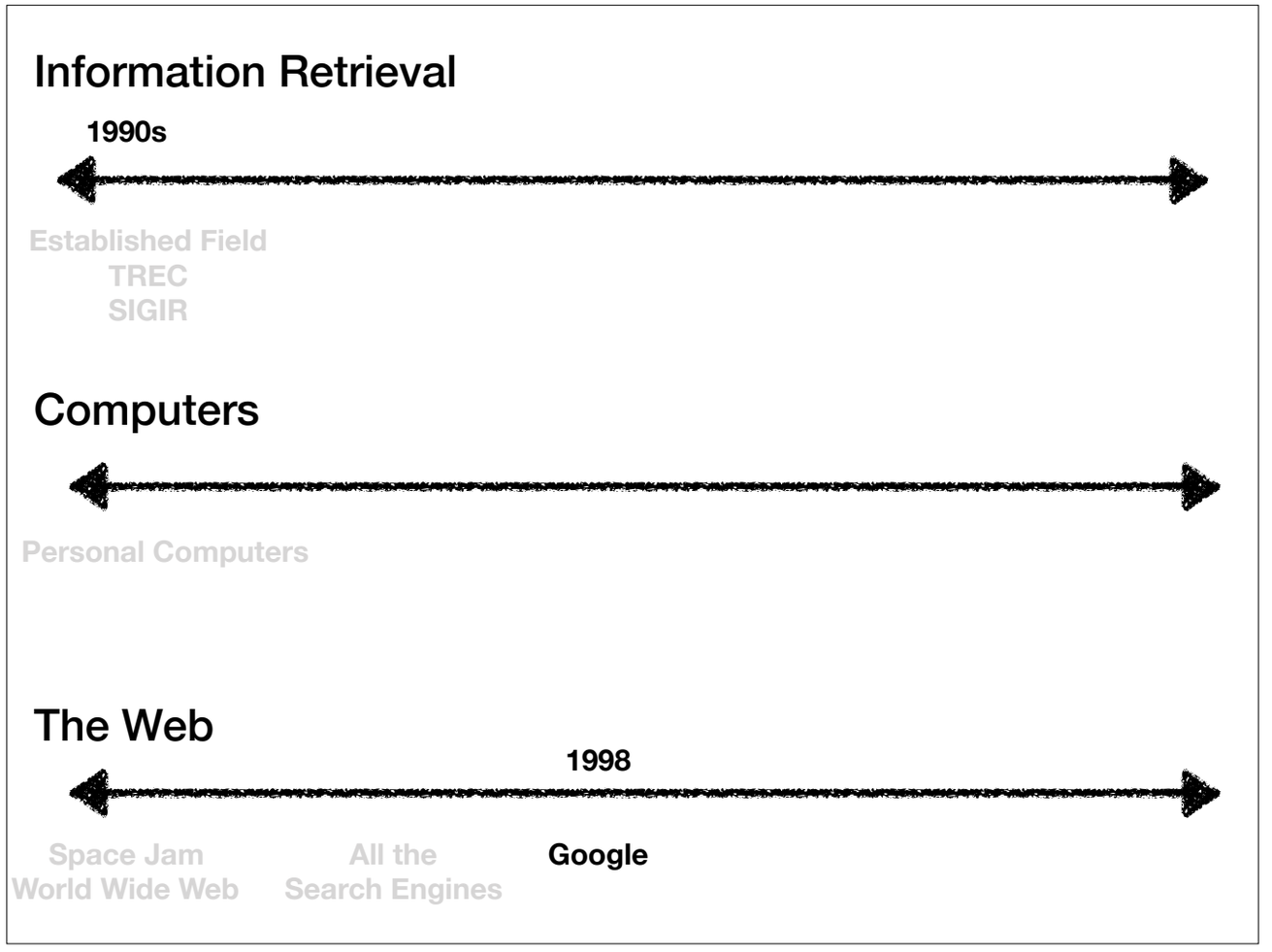


However Information retrieval was becoming a pretty established field. There was already theoretical research into the idea of how to rank results, how to translate free-form queries, to the point that two established conference venues were created.

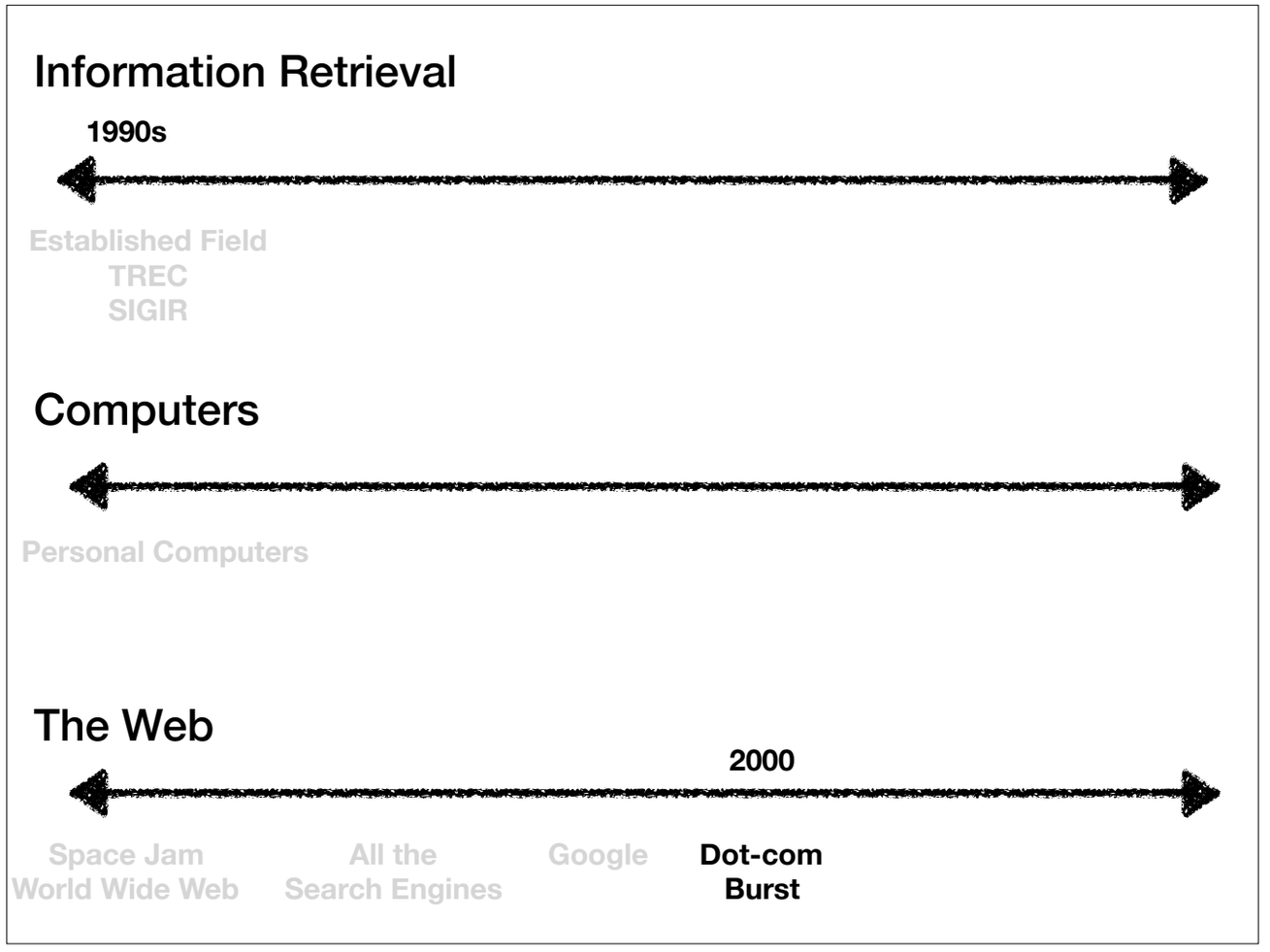
It was probably not a coincidence then that it was in the 90s that we started seeing IR research start getting more applied. TREC for instance is one of those applied conferences, where researchers come together in a “researcher hackathon” to brainstorm and try new tactics of search on computers.



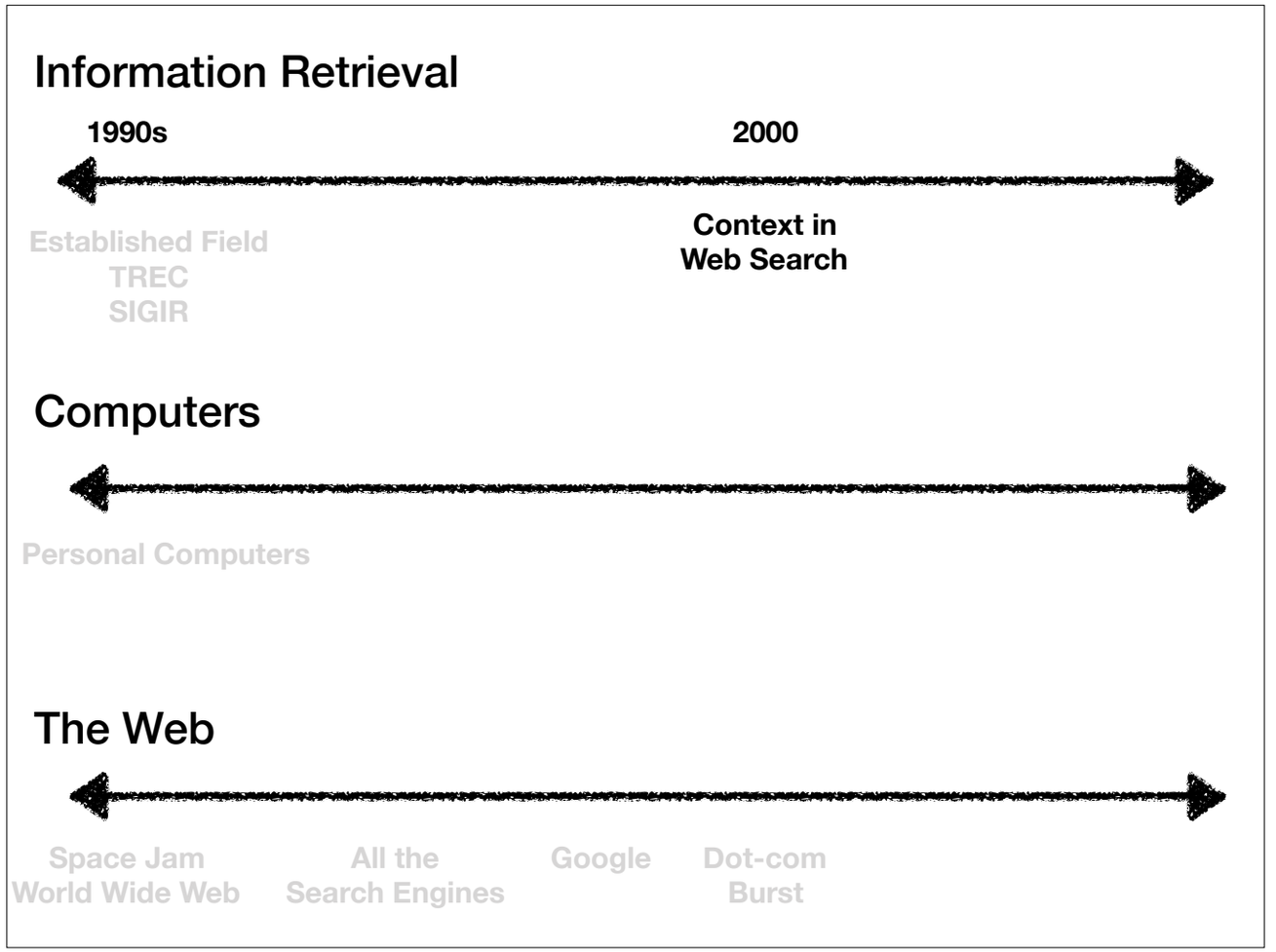
So in the mid-90s, we are seeing all the search engine companies come into being.



Google was formed in 1998.



And so by the early 2000s, we hit about the max that search and the web was capable of, which coincides with the dot-com burst.



At about the same time, in the research world, we got this paper starting to tie in some ideas about context in web search.

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.6478&rep=rep1&type=pdf#page=27

Context in Web Search

Steve Lawrence
NEC Research Institute
Princeton, New Jersey
http://www.neci.nec.com/~lawrence
lawrence@research.nj.nec.com

Abstract

Web search engines generally treat search requests in isolation. The results for a given query are identical, independent of the user, or the context in which the user made the request. Next-generation search engines will make increasing use of context information, either by using explicit or implicit context information from users, or by implementing additional functionality within restricted contexts. Greater use of context in web search may help increase competition and diversity on the web.

1 Introduction

As the web becomes more pervasive, it increasingly represents all areas of society. Information on the web is authored and organized by millions of different people, each with different backgrounds, knowledge, and expectations. In contrast to the databases used in traditional information retrieval systems, the web is far more diverse in terms of content and structure.

Current web search engines are similar in operation to traditional information retrieval systems [57] – they create an index of words within documents, and return a ranked list of documents in response to user queries. Web search engines are good at returning long lists of *relevant* documents for many user queries, and new methods are improving the ranking of search results [8, 10, 21, 36, 41]. However, few of the results returned by a search engine may be *valuable* to a user [6, 50]. Which documents are valuable depends on the context of the query – for example, the education, interests, and previous experience of a user, along with information about the current

This paper is one of the earliest papers to suggest that Personalization is the way to go to improve web search.

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.6478&rep=rep1&type=pdf#page=27

Context in Web Search

Steve Lawrence
NEC Research Institute
Princeton, New Jersey
http://www.neci.nec.com/~lawrence

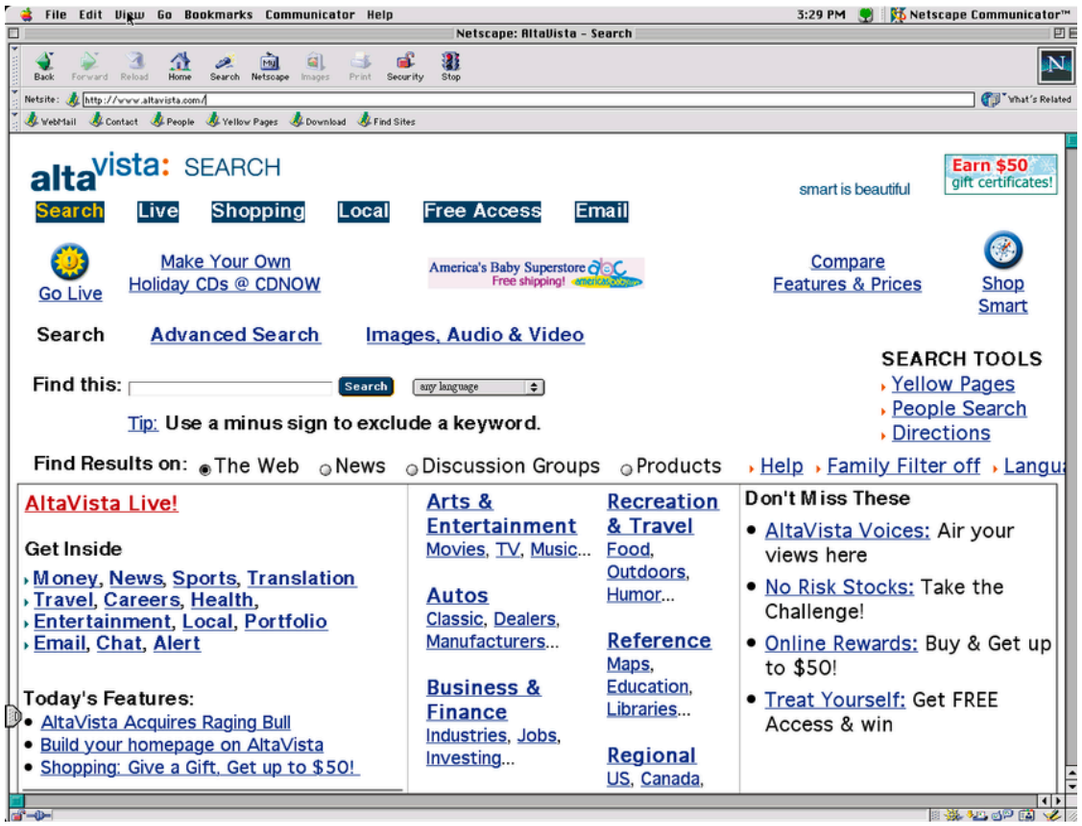
In addition to providing a keyword query, users choose a category such as “personal homepages”, “research papers”, or “general introductory information”.

1 Introduction

As the web becomes more pervasive, it increasingly represents all areas of society. Information on the web is authored and organized by millions of different people, each with different backgrounds, knowledge, and expectations. In contrast to the databases used in traditional information retrieval systems, the web is far more diverse in terms of content and structure.

Current web search engines are similar in operation to traditional information retrieval systems [57] – they create an index of words within documents, and return a ranked list of documents in response to user queries. Web search engines are good at returning long lists of *relevant* documents for many user queries, and new methods are improving the ranking of search results [8, 10, 21, 36, 41]. However, few of the results returned by a search engine may be *valuable* to a user [6, 50]. Which documents are valuable depends on the context of the query – for example, the education, interests, and previous experience of a user, along with information about the current

This is because, the latest that technology could do in 2000 was ask users to input their preferences to improve search.



Remember search in the 90s? It was very categorical. This is what the home page of a search engine looked like.

https://people.csail.mit.edu/teevan/work/publications/papers/tochi10.pdf

Potential for Personalization

…there will always be many cases where people are unable to clearly articulate their needs because they lack the knowledge or vocabulary to do so, or where search engines cannot take good advantage of the additional information…

Personalizing Search via Automated Analysis

Jaim
MIT
32 Vassa
Cambridge,
teevan@

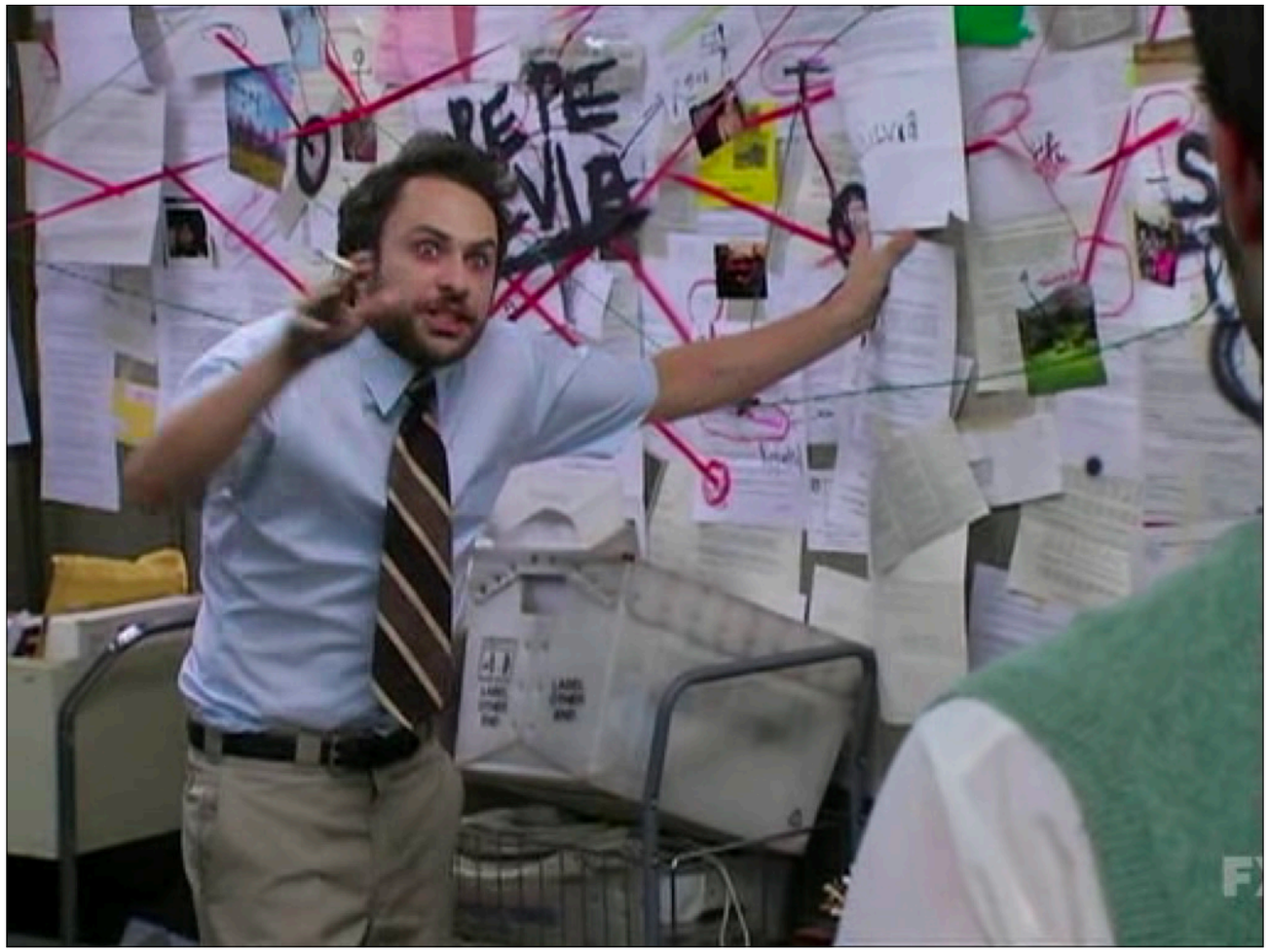
ISA
om

…while current Web search tools do a good job of retrieving results to satisfy the range of intents people have for a given query, they do not do a very good job of discerning individuals' search goals…

http://erichorvitz.com/SIGIR2005_personalize.pdf

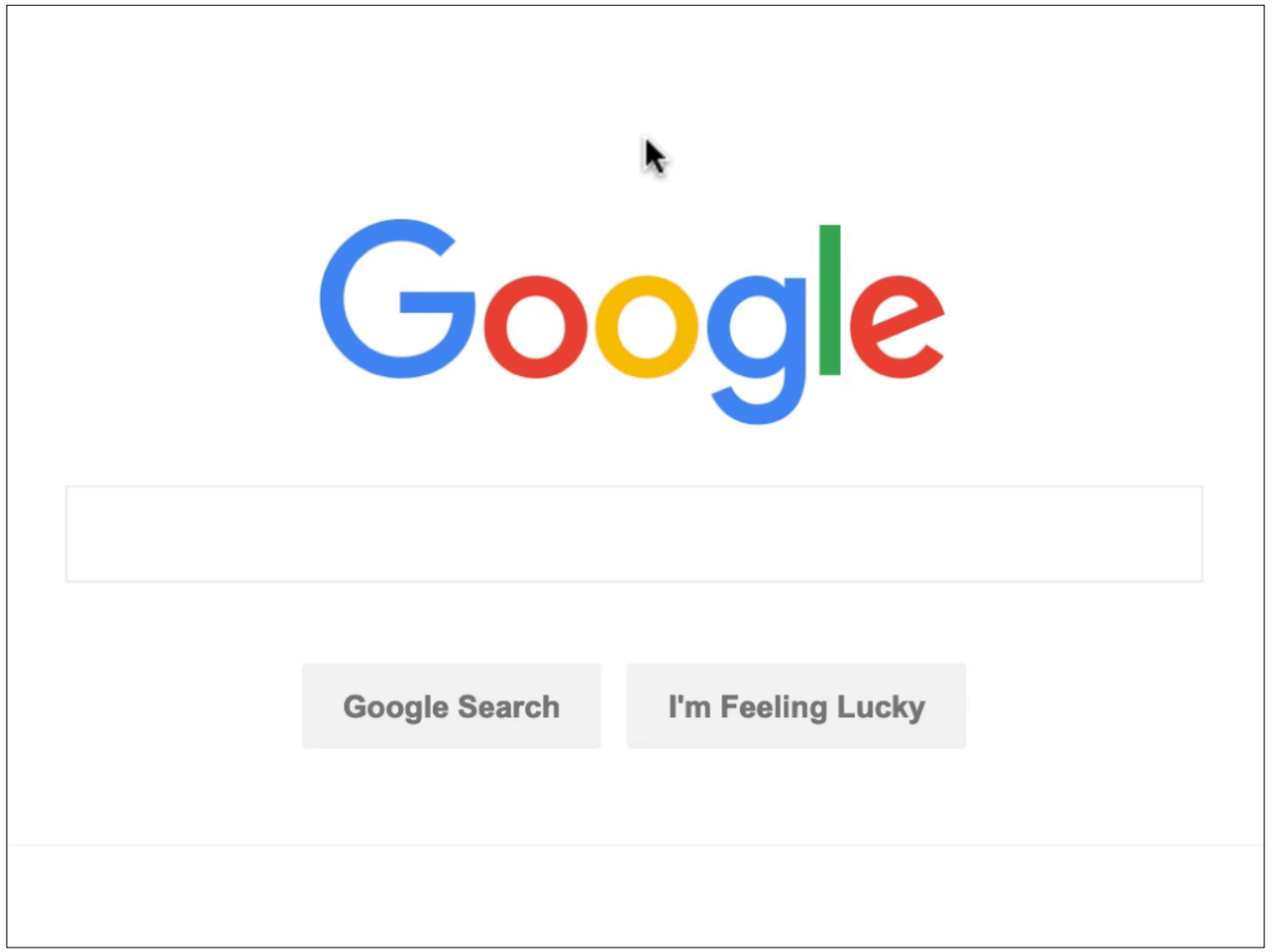
That's because it turns out people are really bad at knowing how to search. If you read all the personalization papers after the year 2000, you'll find that they all have the same introductory theme.

Let's look at an example.

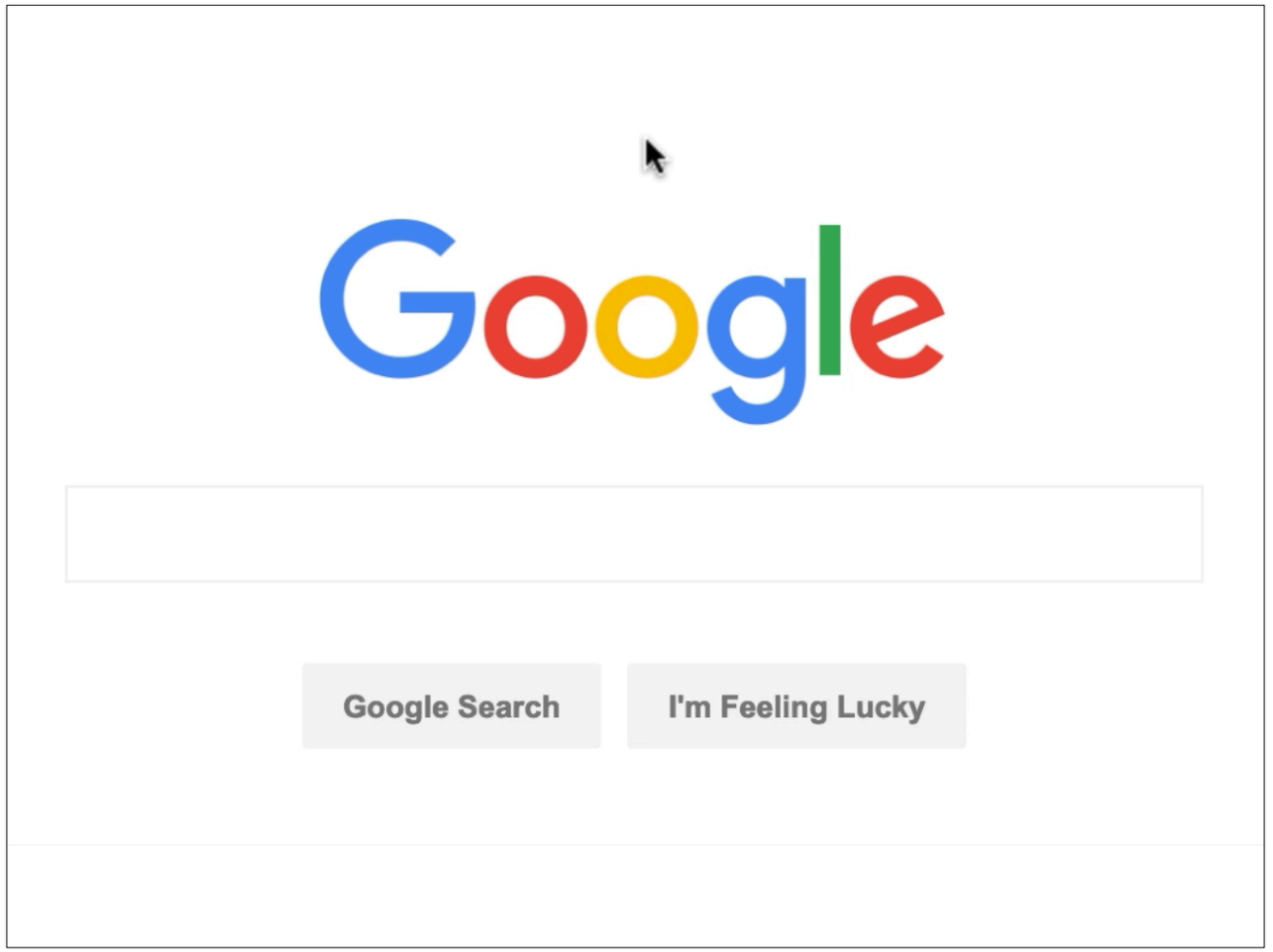


This is one of my favorite pictures to really describe my state of work at any given time, so I often use it for presentations like this one.

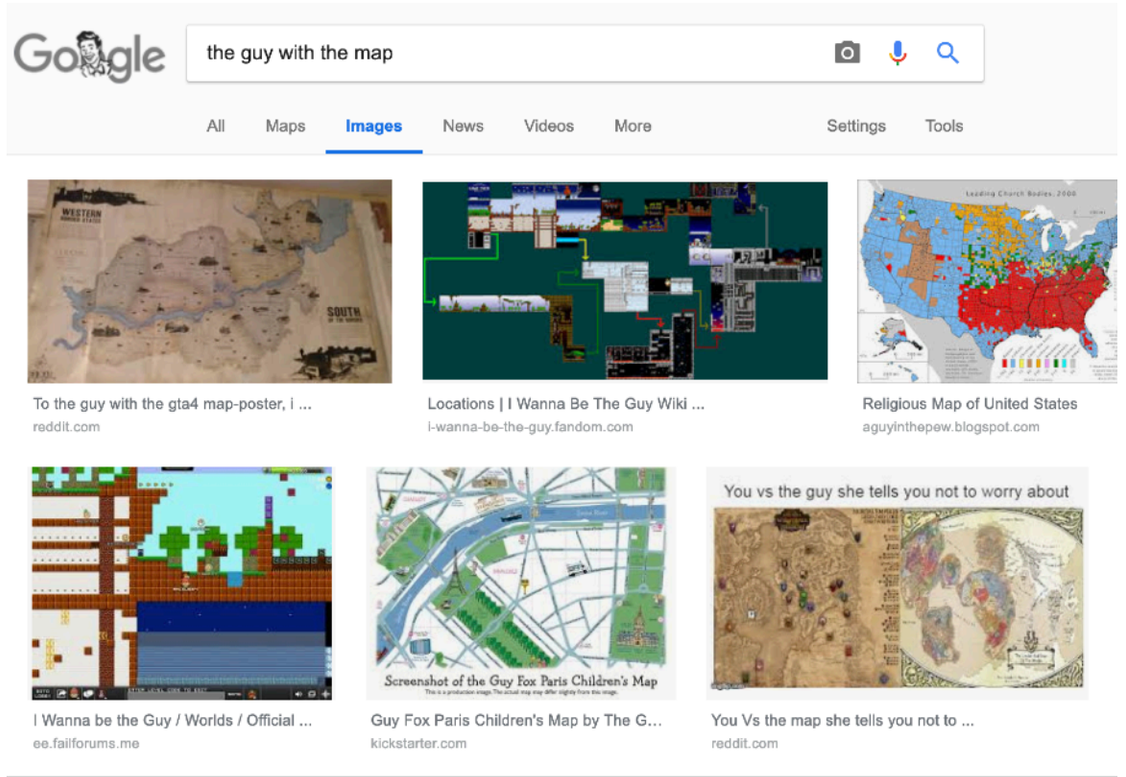
However I never remember where it’s from, so, I often start off by googling



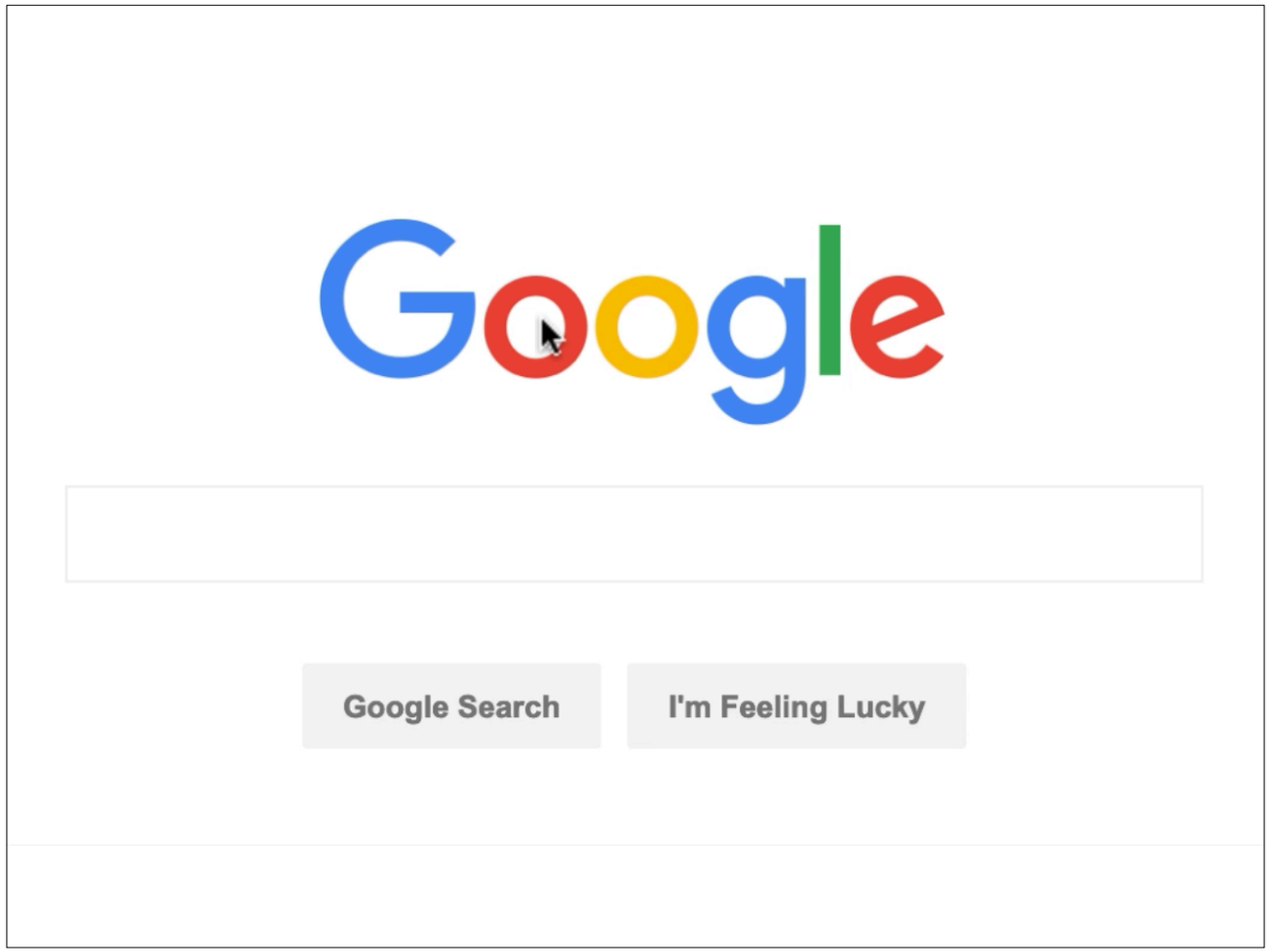
The guy with the map



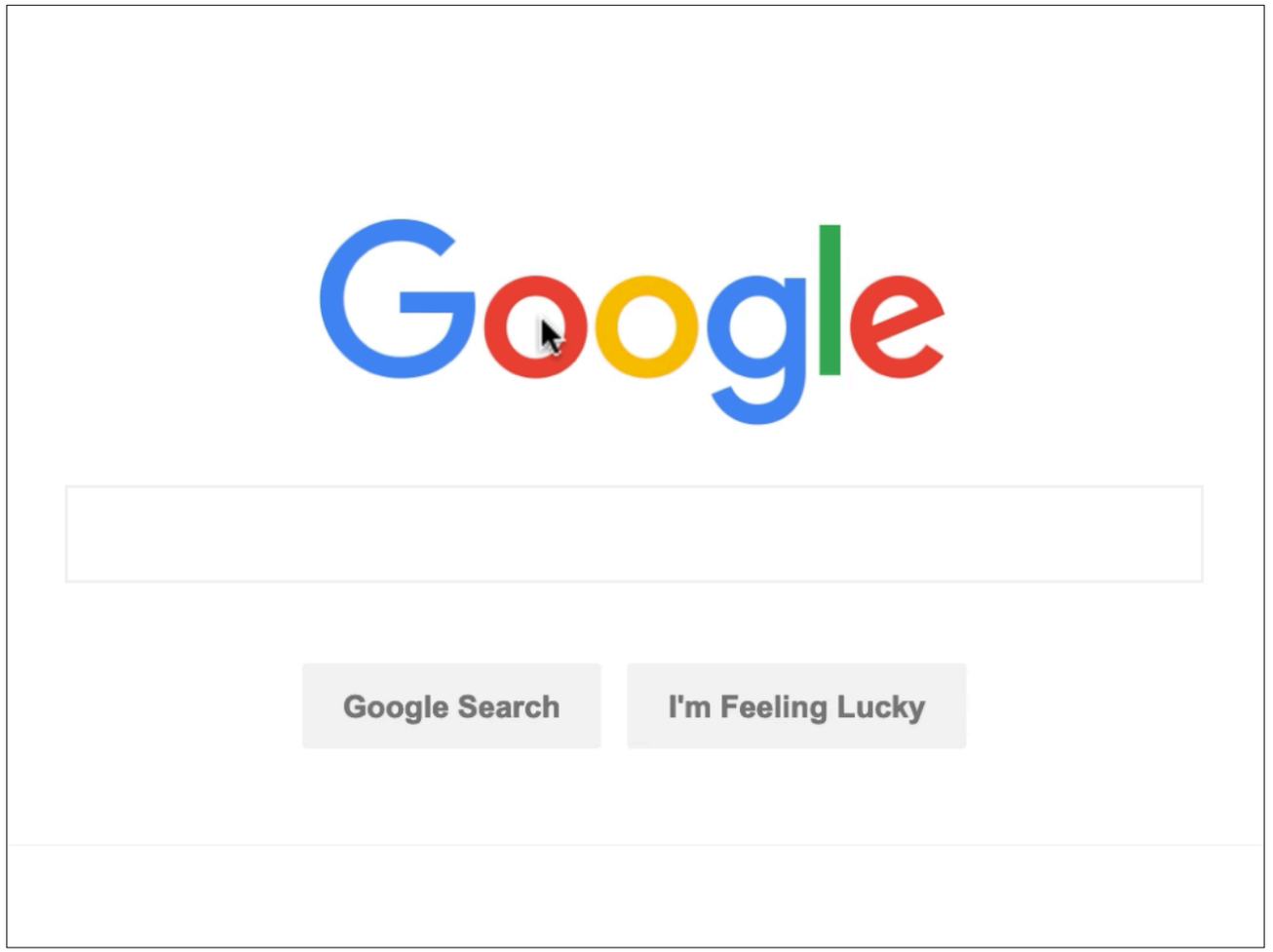
The guy with the map



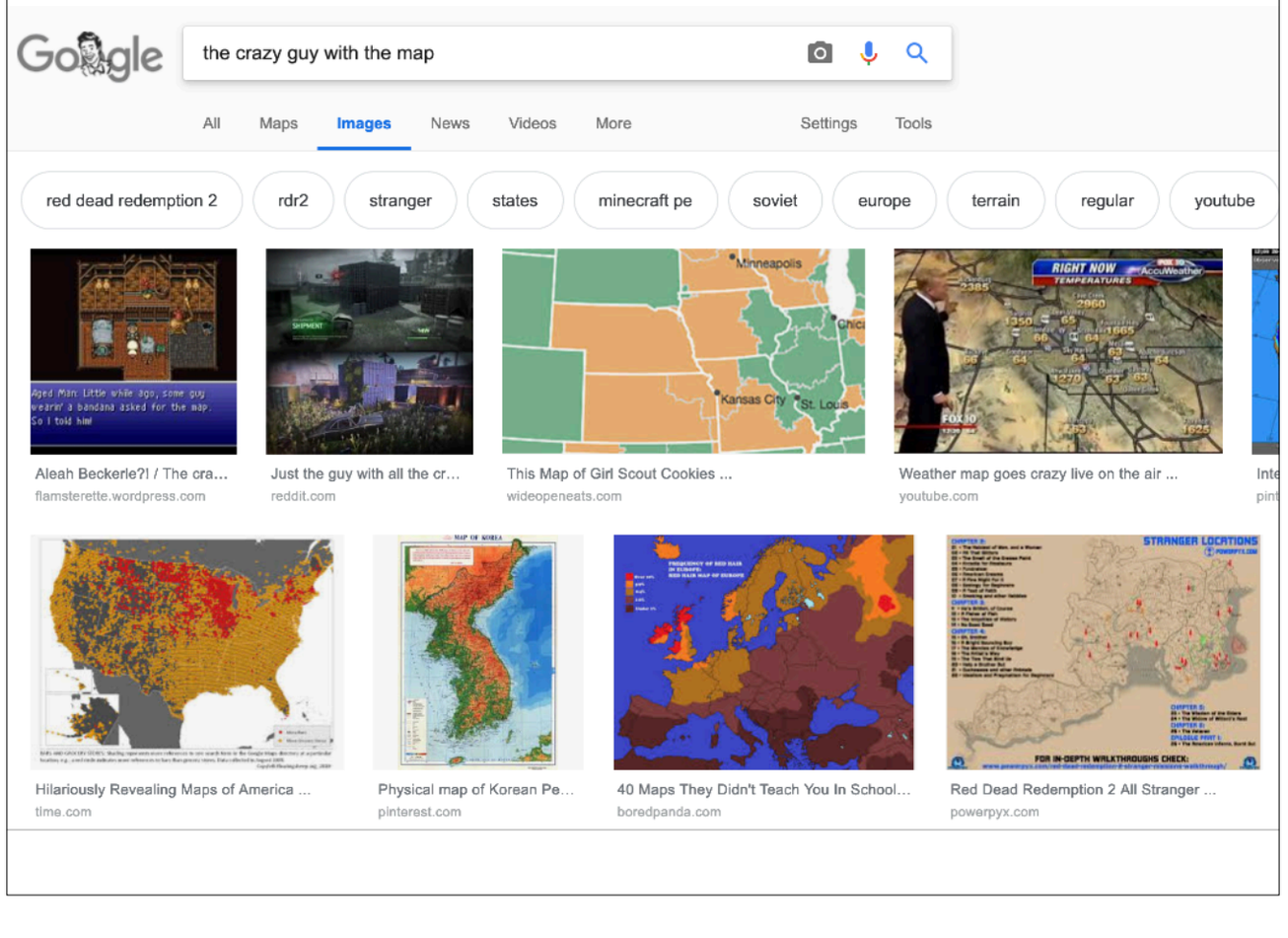
Ok that didn’t help me



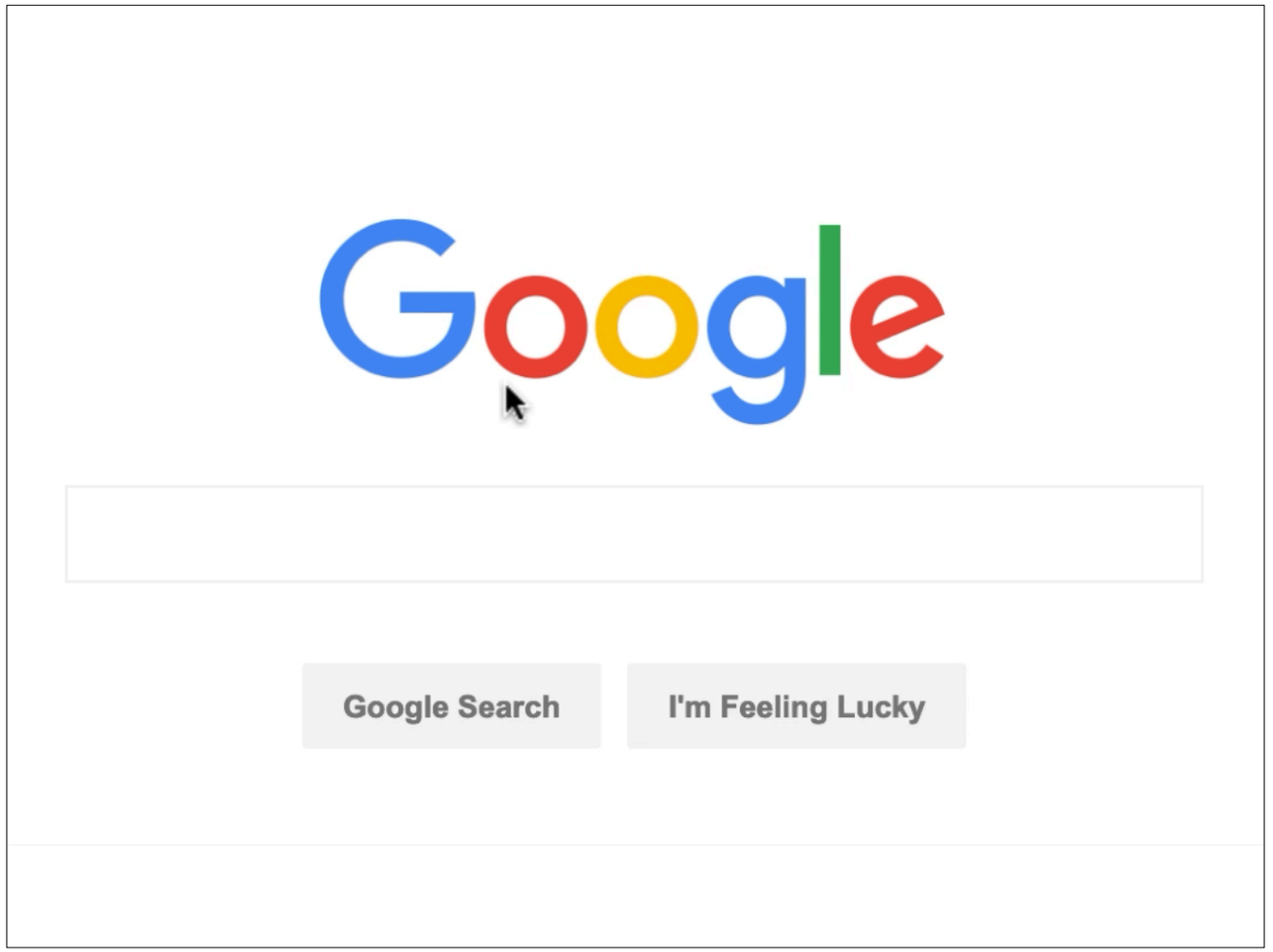
ok, let’s revise our search. I know it’s a crazy looking guy.



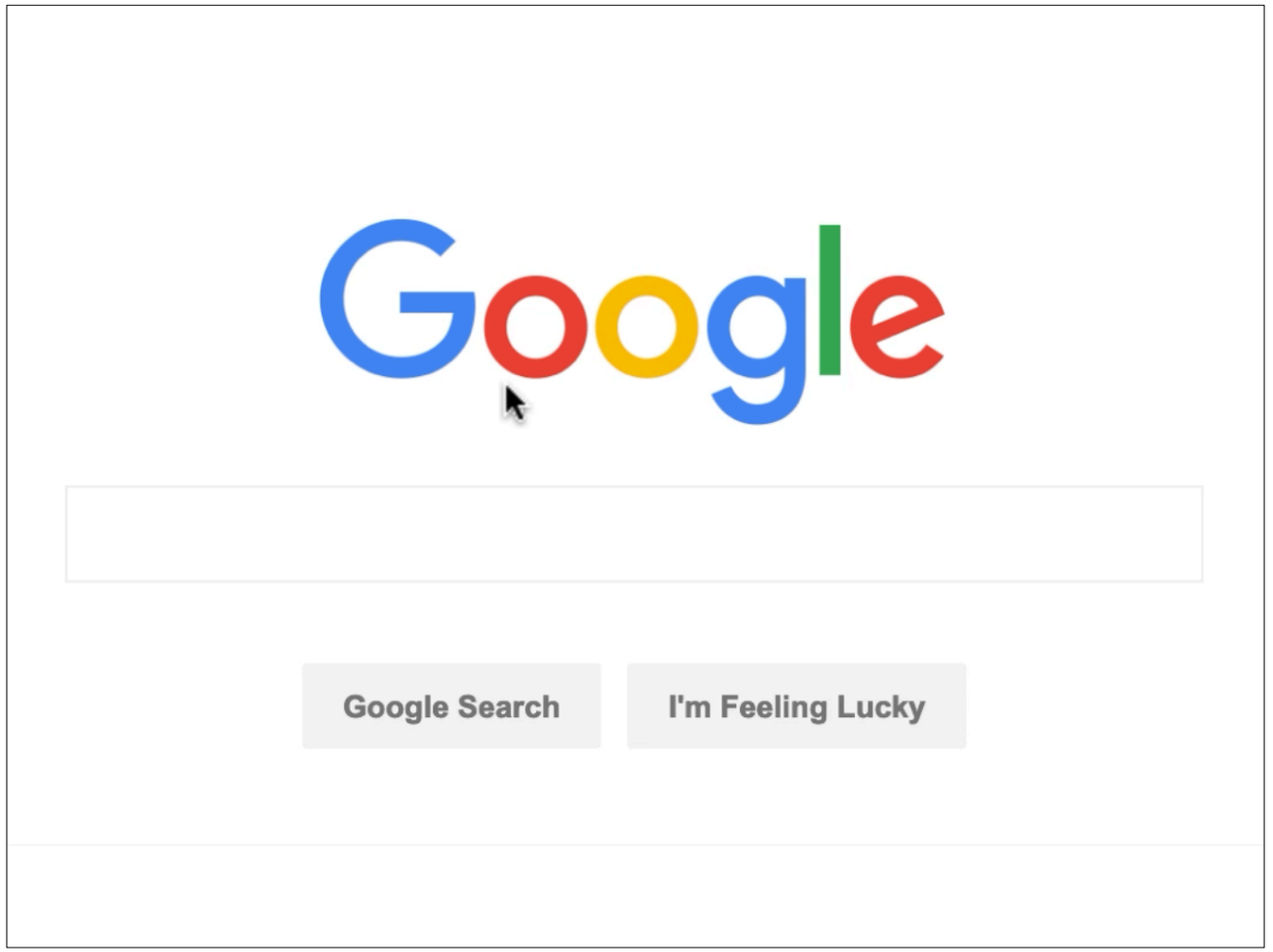
ok, let’s revise our search. I know it’s a crazy looking guy.



Which still doesn’t help me

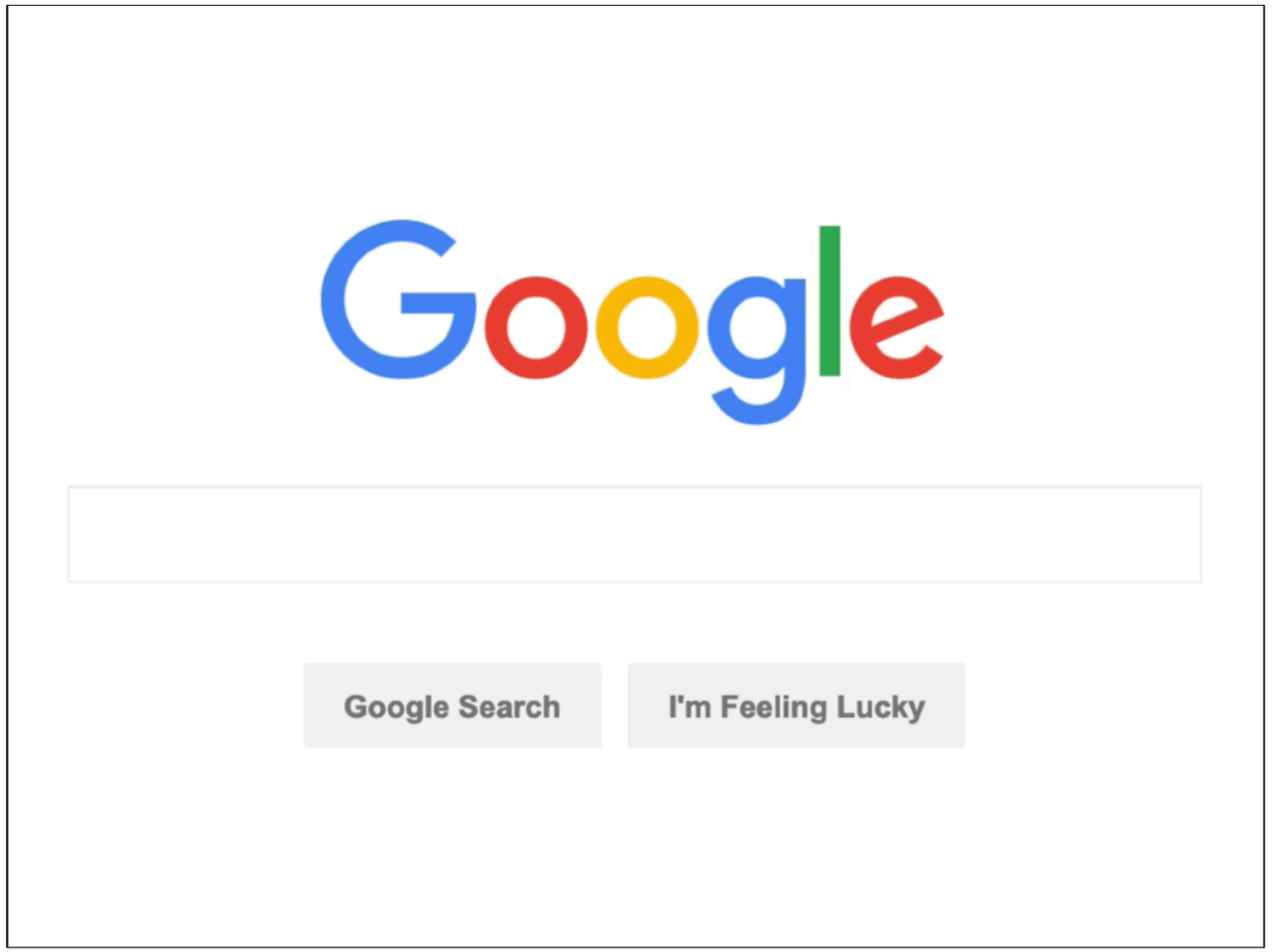


Oh but then I remember that it’s a meme which is why I like using it in my slides



Oh but then I remember that it’s a meme which is why I like using it in my slides

And bam I get what I want.



So, you might remember Google was revolutionary because it had none of those. Just a text bar and two buttons and a logo. Over time Google has dominated the market because it was a company that started adopting personalization.

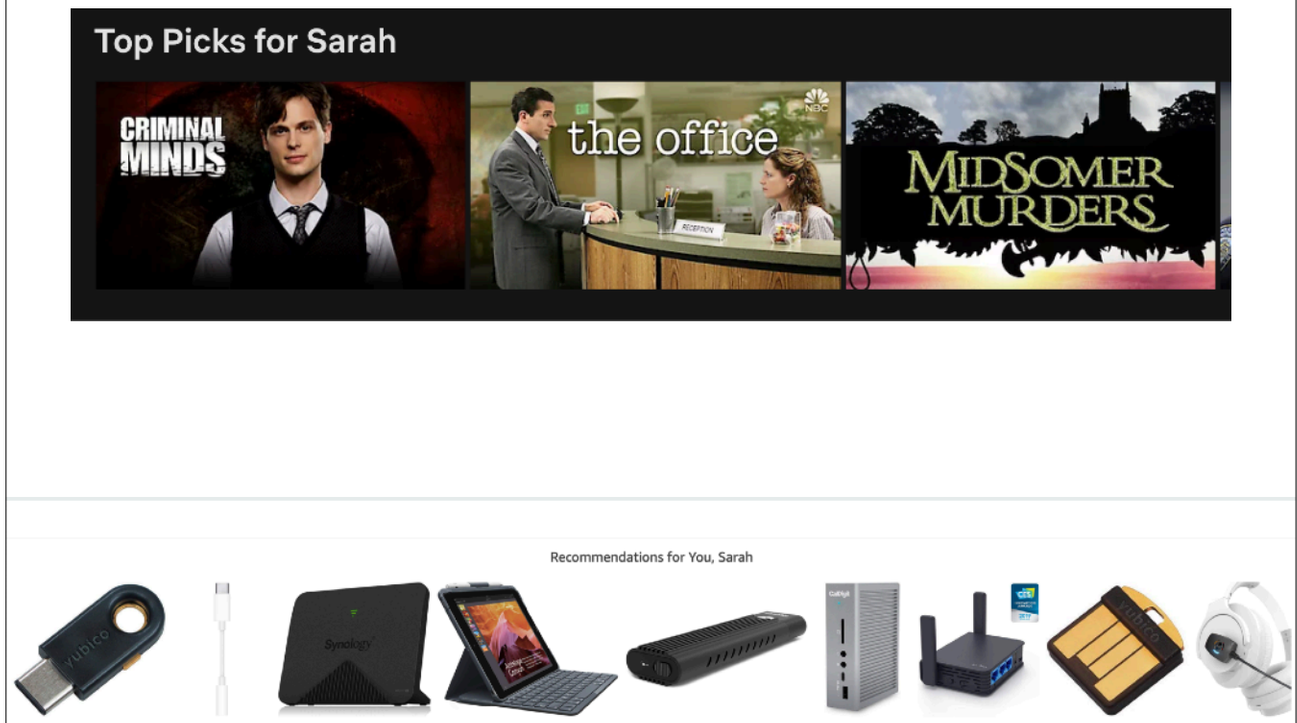
And here’s a hint: you look at the trend of the success of any major information retrieval company, it’s probably because they adopted personalization.

Personalization

So what is personalization?

https://www.flipkart.com/avs-stuffed-spongy-hugable-cute-teddy-bear-heart-just-you-red-color-45-cm/p/itmexekzf7xumj5z

Personalization is used to cater things just to you.



Examples of personalization! Any recommendation system is a good example of personalization. It’s based on building a profile of you. So here Netflix recommends videos based on what I watched, and Amazon recommends products based on what I purchased or looked at.

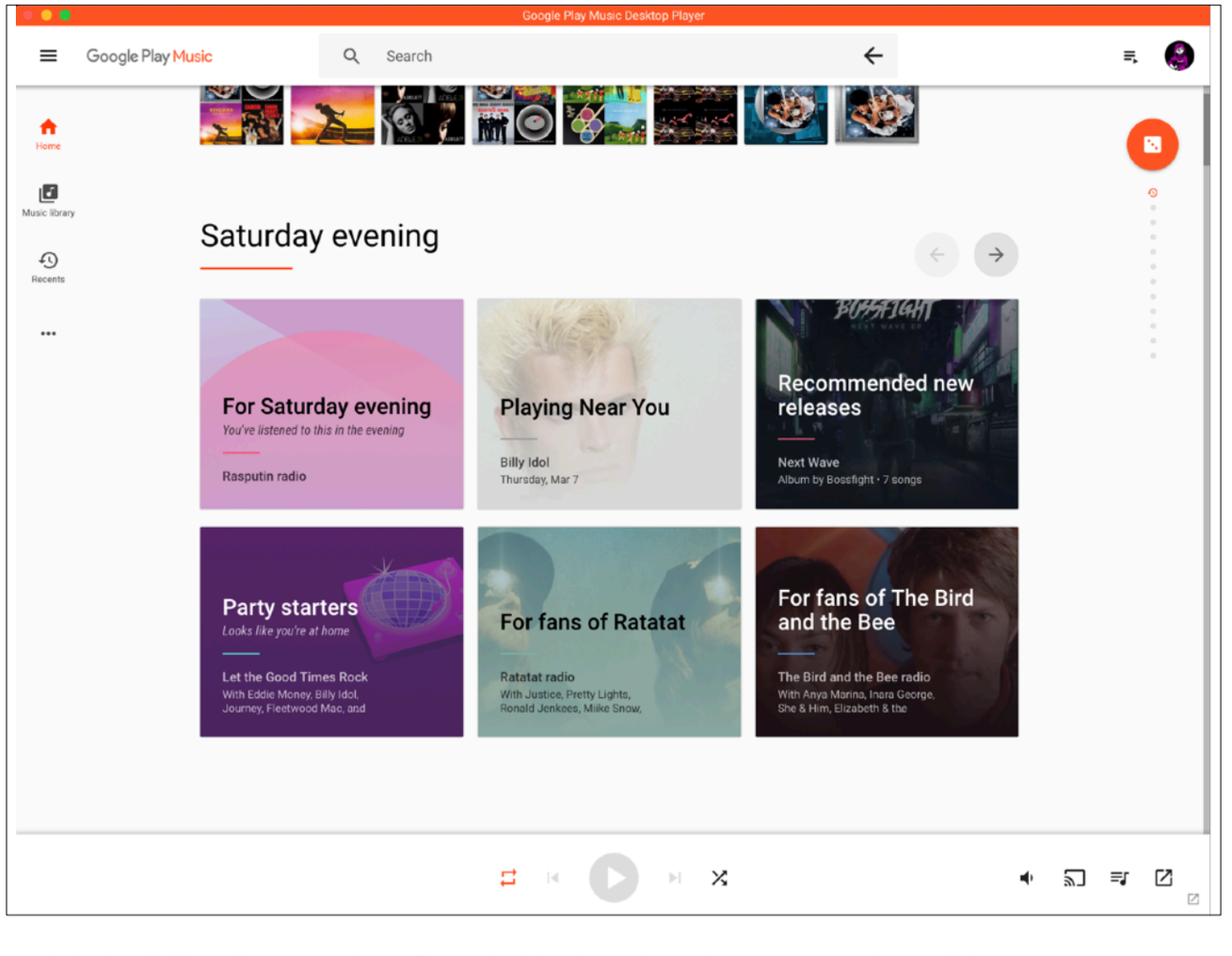
Because I am a security nerd, that’s why not one, but two yubikey products show up in this list.

Now personalization is made up of two parts, and the example you see here is an example of Individualization, I.e. catering to me personally based on my preferences.

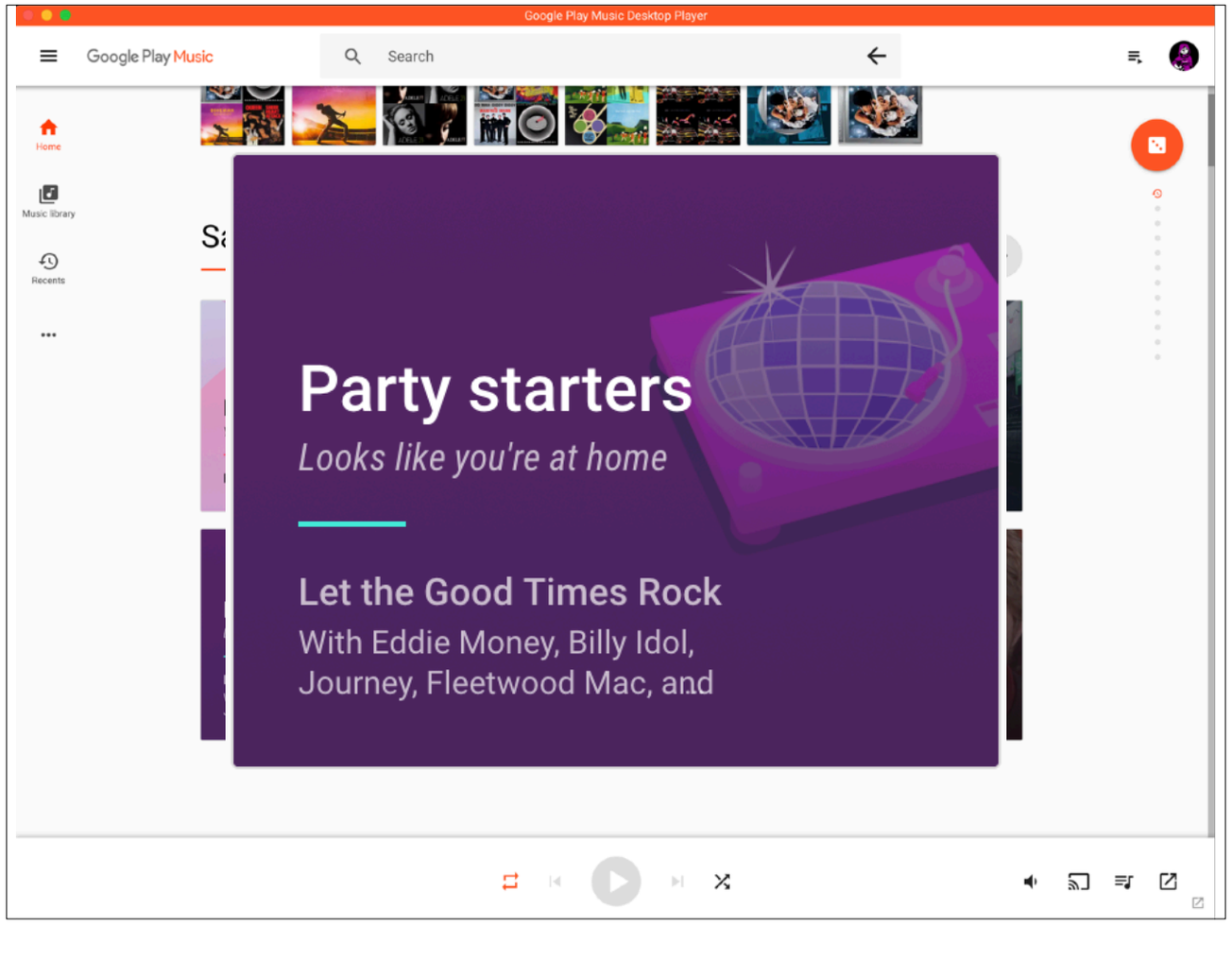
Individualization

Contextualization

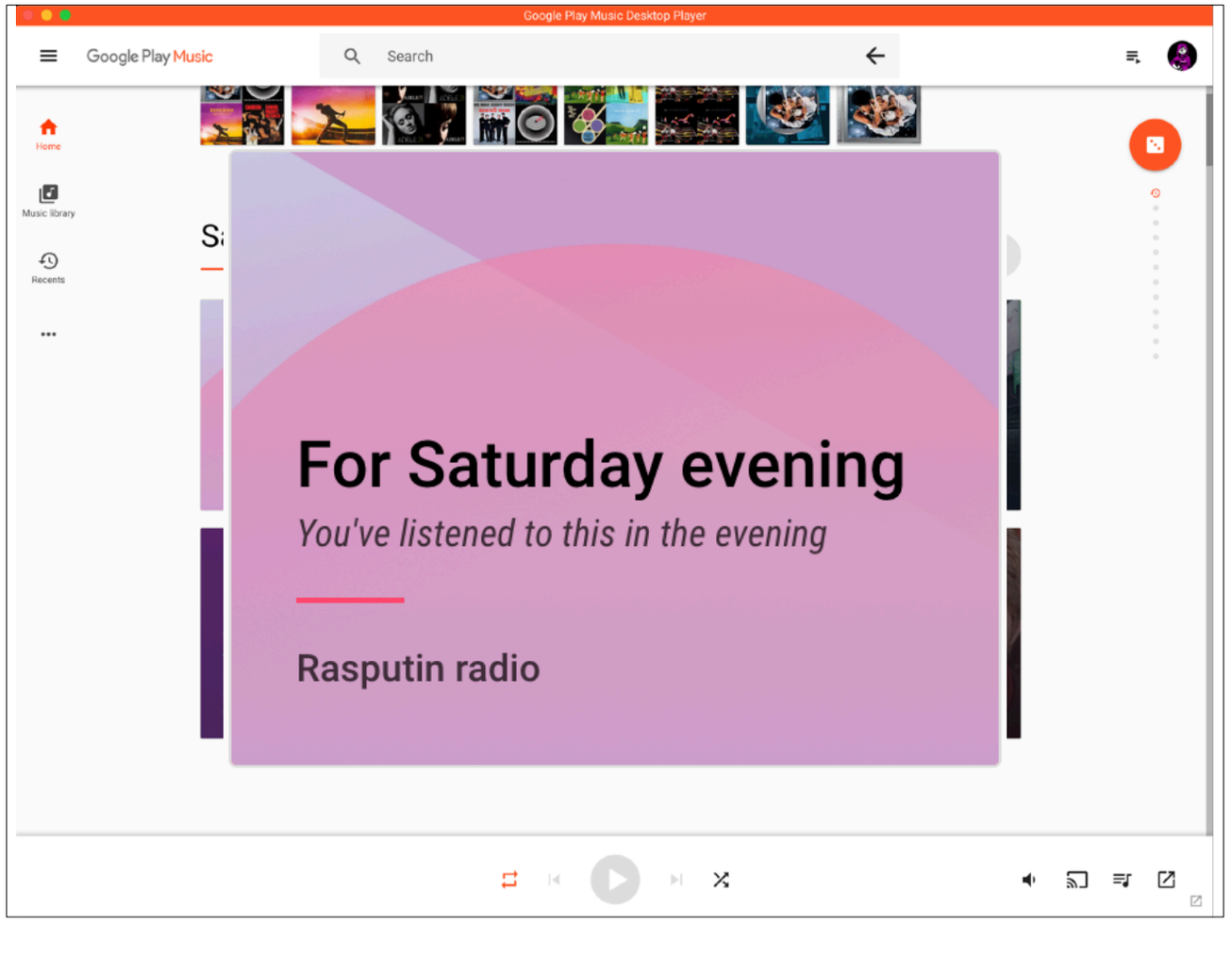
The other part of personalization is contextualization, which is the act of making a scenario have context.



So let's talk about what contextualization might look like. I opened up Google Music a few days ago and got this.



“Looks like you’re at home”



“You’ve listened to this in the evening”.

How do they know?

So how do they know?

Why does this feel weird?

Behavior Inference

(i.e. Lack of Consent)

The reason this feels weird is that this is all behavior inference. Different services are making guesses about what you want and what you are searching for. And it wouldn’t be terrible if the guesses were wrong, but the fact is, the guesses are right, and are creepily right. And so it feels like another violation.

Anti-Pattern #1:
Inference without Consent

So here is anti-pattern 1: we have inferences without explicit consent.

Why are we inferring context?

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.6478&rep=rep1&type=pdf#page=27

Context in Web Search

Steve Lawrence
NEC Research Institute
Princeton, New Jersey
http://www.neci.nec.com/~lawrence
lawrence@research.nj.nec.com

Inquirus 2 can greatly improve search precision, but requires the user to explicitly enter context information. What if search context could be automatically inferred?

1 Introduction

As the web becomes more pervasive, it increasingly represents all areas of society. Information on the web is authored and organized by millions of different people, each with different backgrounds, knowledge, and expectations. In contrast to the databases used in traditional information retrieval systems, the web is far more diverse in terms of content and structure.

Current web search engines are similar in operation to traditional information retrieval systems [57] – they create an index of words within documents, and return a ranked list of documents in response to user queries. Web search engines are good at returning long lists of *relevant* documents for many user queries, and new methods are improving the ranking of search results [8, 10, 21, 36, 41]. However, few of the results returned by a search engine may be *valuable* to a user [6, 50]. Which documents are valuable depends on the context of the query – for example, the education, interests, and previous experience of a user, along with information about the current

Remember we talked about how users don’t know what they’re searching for. Remember that paper back from 2000? It proposed, what if we could automatically infer search context?

Context in Web Search

…attempts to model the context of user information needs based on the content of documents being edited in Microsoft Word, or viewed in Internet Explorer.

Abstract

…indexes specified files such as email messages and research papers, and continually searches for related documents while a user edits a document in the Emacs editor.

As the web becomes more pervasive, it increasingly represents all areas of society. Information on the web is authored and organized by millions of different people, each with different backgrounds, knowledge, and expec-

…automatically suggests content from the web or local files, based on the documents a user is reading or editing.

eligible may be valuable to a user [6, 36]. Which documents are valuable depends on the context of the query – for example, the education, interests, and previous experience of a user, along with information about the current

Here are year 2000 era ideas on inferring context.

However in all these cases, the limitation is that, the search engine only has information on current, right-now context. This is otherwise known as short-term context. There is no information about long-term context.

Note that this is all client-based! Why is this a problem?

Remember this was a computer in the 90s.

https://www.bridgemi.com/talent-education/college-costs-2000

By the year 2000, this is what a personal computer looked like.

2000	On January 5, 2000, AMD released the 800 MHz Athlon processor.
2000	Intel released the Celeron 533 MHz with a 66 MHz bus processor on January 4, 2000.
2000	AMD first released the Duron processor on June 19, 2000, with speeds of 600 MHz to 1.8 GHz and bus speeds of 200 MHz to 266 MHz. The Duron was built on the same K7 architecture as the Athlon processor.
2000	Intel announces on August 28th that it will recall its 1.3 GHz Pentium III processors due to a glitch. Users with these processors should contact their vendors for additional information about the recall.

https://www.computerhope.com/history/processor.htm

Some specs on speeds.

I couldn’t find an article on disk drives that went into a lot of detail, but the largest hard disks of the era were maybe at 100GB. Most consumers probably had disk drives in the tens of GB.

So we’re faster than before, but not really that much faster.

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.6478&rep=rep1&type=pdf#page=27

Context in Web Search

Steve Lawrence
NEC Research Institute
Princeton, New Jersey

However, these services do not have local access to a large scale index of the web, which limits their functionality. For example, such a service could not rank the homepage of the computer scientist highly for the query “Michael Jordan”, unless a search service returns the page within the maximum number of results that the client retrieves.

authored and organized by millions of different people, each with different backgrounds, knowledge, and expectations. In contrast to the databases used in traditional information retrieval systems, the web is far more diverse in terms of content and structure.

Current web search engines are similar in operation to traditional information retrieval systems [57] – they create an index of words within documents, and return a ranked list of documents in response to user queries. Web search engines are good at returning long lists of *relevant* documents for many user queries, and new methods are improving the ranking of search results [8, 10, 21, 36, 41]. However, few of the results returned by a search engine may be *valuable* to a user [6, 50]. Which documents are valuable depends on the context of the query – for example, the education, interests, and previous experience of a user, along with information about the current

So we have this situation of we are well in the era of personal computing, and everyone has their own device, but trying to store the entire index of the web locally is basically impossible.

For comparison, wikipedia is 100GB compressed, and 10TB uncompressed.

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.6478&rep=rep1&type=pdf#page=27

Context in Web Search

Steve Lawrence
NEC Research Institute
Princeton, New Jersey
http://www.neci.nec.com/~lawrence

With the cost of running a large scale search engine already very high, it is likely that server-based full-scale personalization is currently too expensive for the major web search engines… However, advances in computer resources should make large scale server-based personalized search more feasible over time.

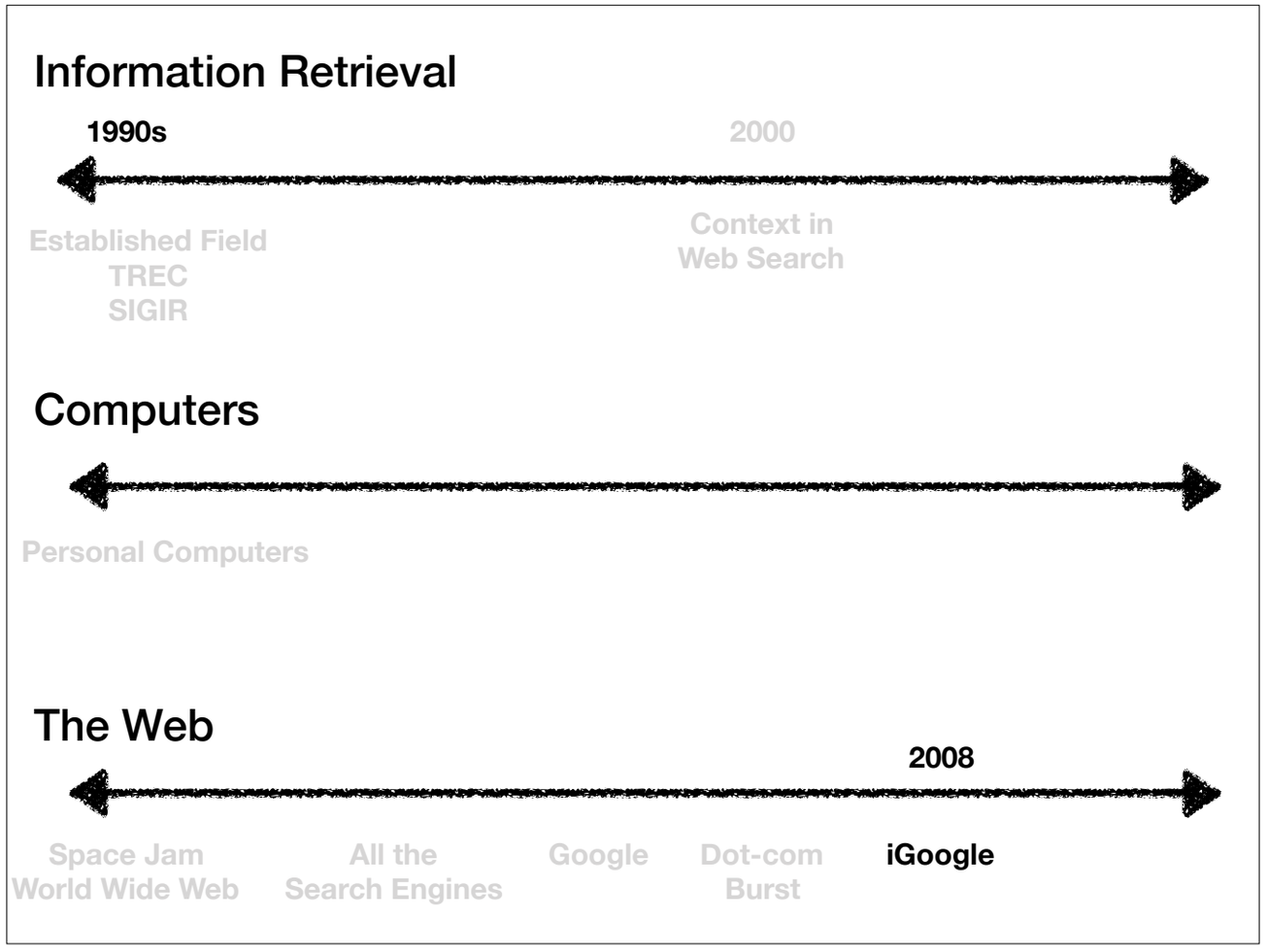
As the web becomes more pervasive, it increasingly represents all areas of society. Information on the web is authored and organized by millions of different people, each with different backgrounds, knowledge, and expectations. In contrast to the databases used in traditional information retrieval systems, the web is far more diverse in terms of content and structure.

Current web search engines are similar in operation to traditional information retrieval systems [57] – they create an index of words within documents, and return a ranked list of documents in response to user queries. Web search engines are good at returning long lists of *relevant* documents for many user queries, and new methods are improving the ranking of search results [8, 10, 21, 36, 41]. However, few of the results returned by a search engine may be *valuable* to a user [6, 50]. Which documents are valuable depends on the context of the query – for example, the education, interests, and previous experience of a user, along with information about the current

So basically a prediction was made, back even in 2000! That personalized search would eventually be feasible in the future when cpu and storage got cheaper.

- Pure search methods don’t capture actual intent
- Humans are bad at expressing intent
- Context is valuable for deriving intent
- Client-side methods are not powerful enough

So to summarize, the research community decided that…



What did that look like on the outside? Well by 2008, Google started rolling out “personalization” as an opt-in basis. You could opt-in to have this customization.

Google Now Notifies Of “Search Customization” & Gives Searchers Control

Danny Sullivan on July 30, 2008 at 4:33 pm

Google is now showing "search customization" messages to inform searchers when their search results have been modified from "normal" due to a searcher’s geographic location, previous query or web surfing and search history. It’s a nice move to help searchers know what exactly is going on "under the hood" at Google and override it if the wrong choices are being made. I’d like to see more of it.

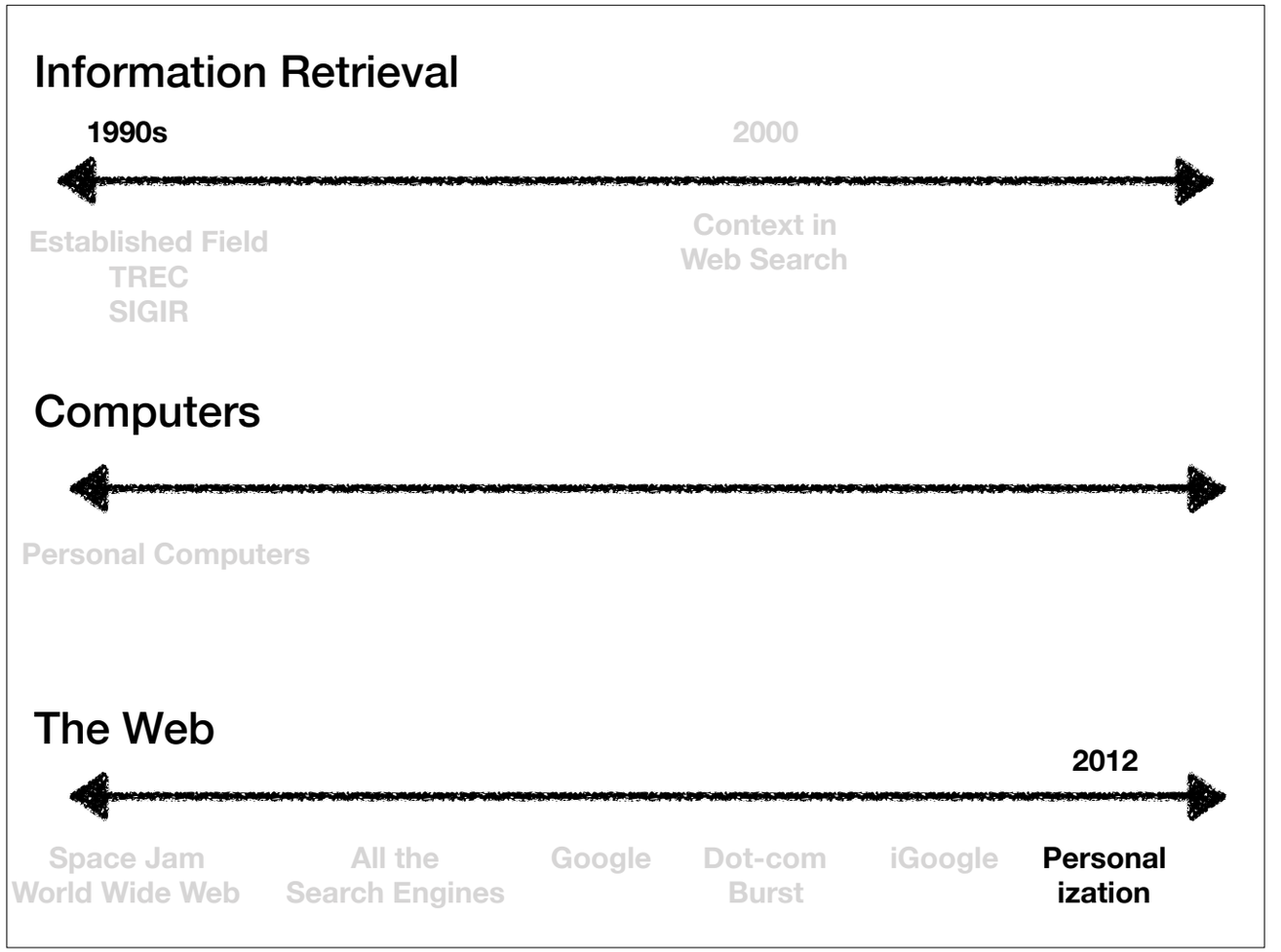
What’s Normal?

There was a time when everyone saw exactly the same search results at the major search engines. Over time, this has changed. In particular, geographic targeting has meant that searchers in different countries often see different results (see
How Search Engines
Redirect Users To Country-Specific Sites for more on that). In addition, personalized search results on Google mean that, more and more, searchers are seeing things differently than what others get. "Normal" search results are becoming an endangered species.

Still, there are times when you want Google to get back to as normal as possible. Today’s rollout is designed to inform searchers when Google is customizing their results in three key areas and allow them to regain control. Note that this only works for English queries on Google.com, at the moment. Not everyone will see it immediately, either — but it should go fully live over the next few days.

https://searchengineland.com/google-now-notifies-of-search-customization-gives-searchers-control-14485

Here’s some guy writing about that. You might have remembered it being called iGoogle.



What did that look like on the outside? Well by 2008, Google started rolling out “personalization” as an opt-in basis. You could opt-in to have this customization.

Of “Magic Keywords” & Flavors Of Personalized Search At Google

Danny Sullivan on November 9, 2012 at 2:10 pm

When is personalized search not personalized search? A recently discovered shift on how Google may alter your search results based on what you — and others in aggregate — previously have searched for may have you wondering how to answer that question.

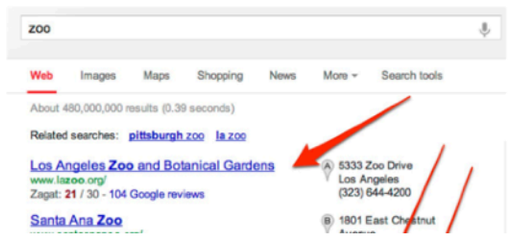
To understand the latest development, I think it’s helpful to go back and review the “flavors” of personal search that Google has, flavors that often all get mixed together. Let’s dive in.

Personalization Based On Geography

Google has long had personalized search, where various factors are used to provide results tailored to the individual.

In fact, Google really hasn’t had “normal” results for years, since for over a decade, it has personalized based on geographic location. These days, it’s gone well past being country-based and works to the city level. What someone sees in one city can be radically different than someone in another.

For example, here’s what I get for a “zoo” search:



https://searchengineland.com/flavors-of-google-personalized-search-139286

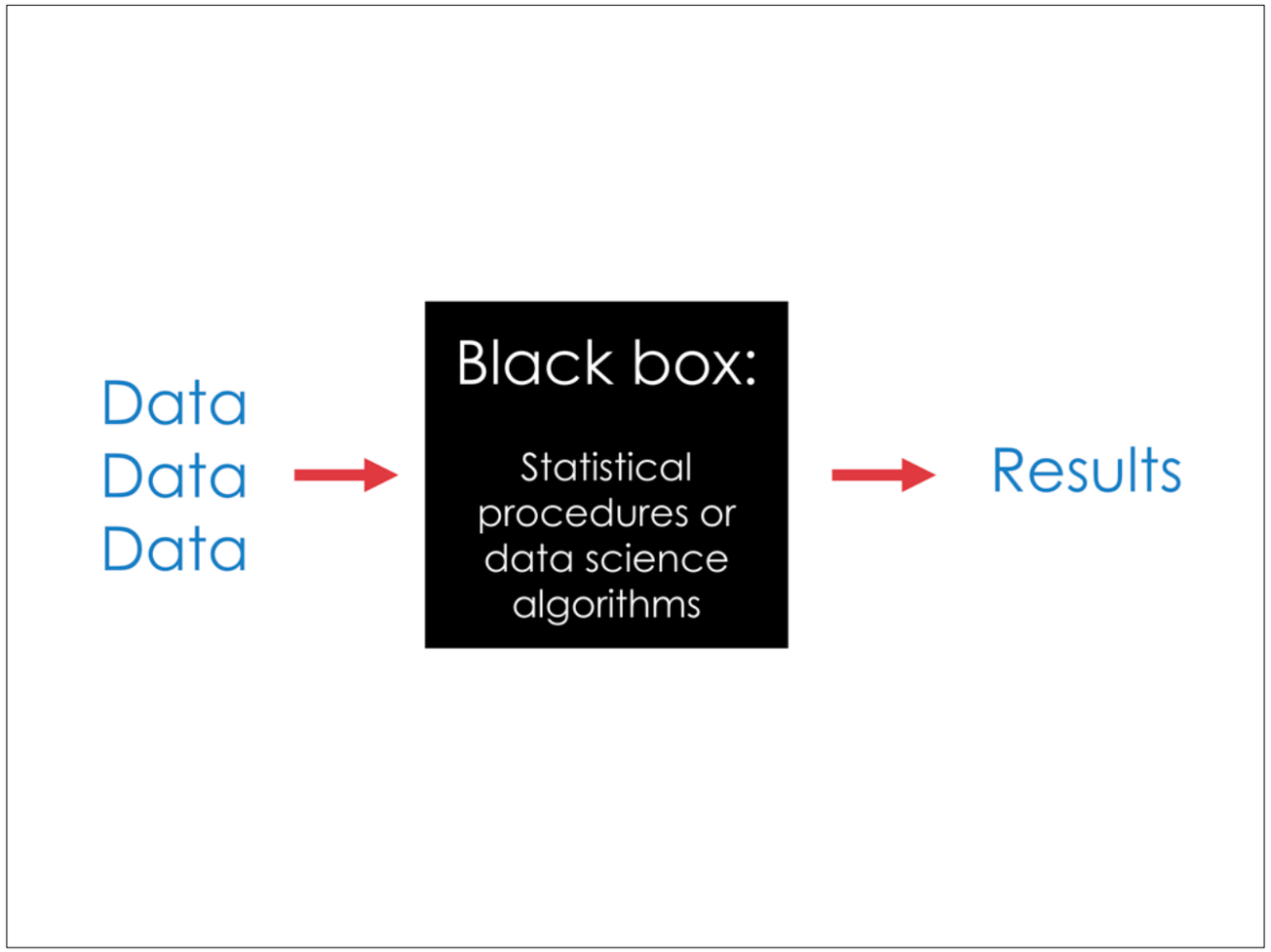
However, by 2012, this kind of setting started going away, and becoming more opaque. Folks noticed that personalization features were getting rolled out without any sort of notification to the user.

Anti-Pattern #2:
Lack of Transparency

So here is anti-pattern 2: Lack of Transparency

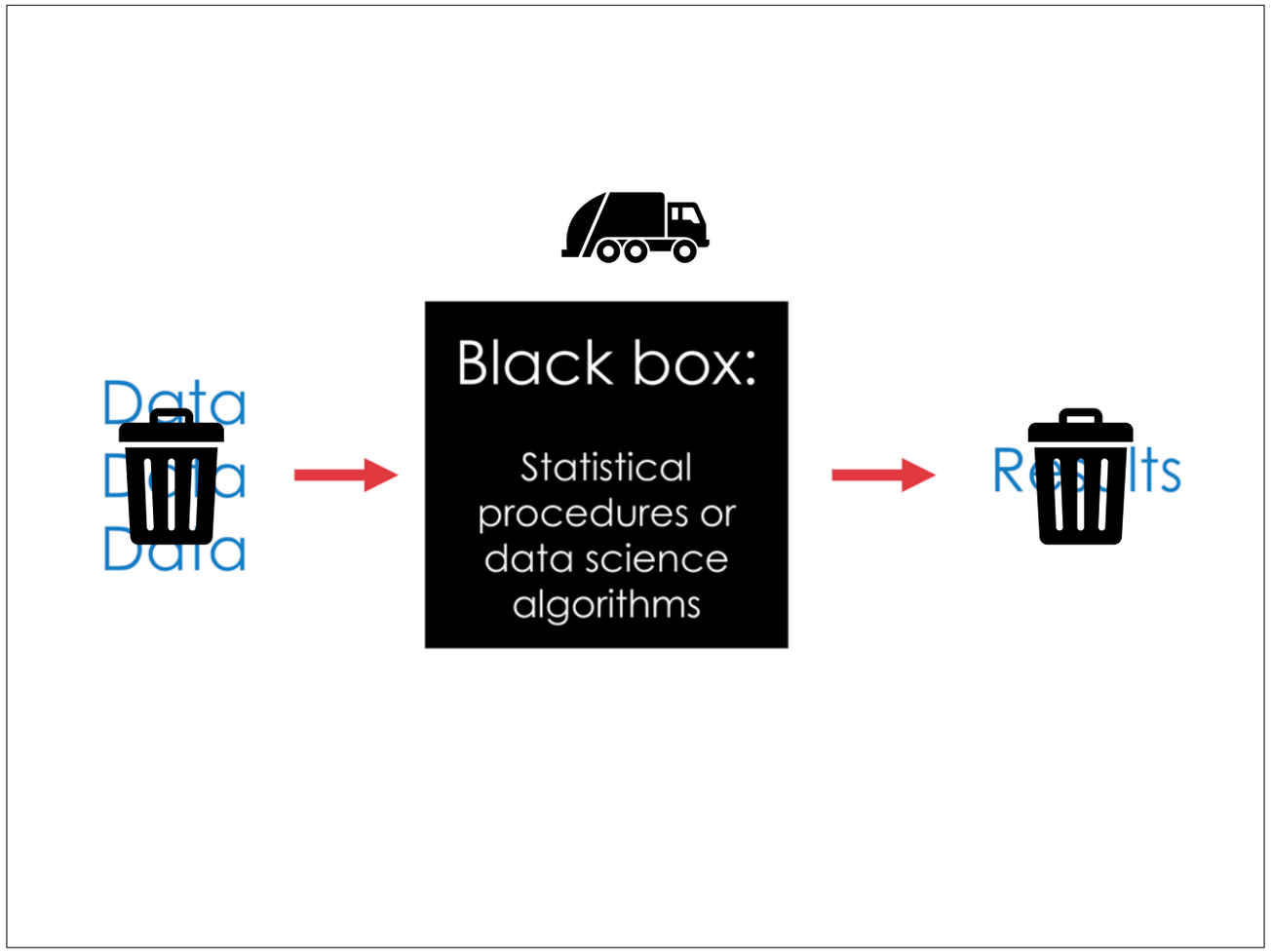
Anti-Pattern #3:
Erosion of Control

And also Anti-pattern 3: erosion of control!



Sort of at the same time, we started to see the democratization of AI/ML techniques. A lot of this is due to the intertwined relationship between the search community and the ML community. So any developments in search would likely translate to developments in ML techniques.

If you don’t know what Machine learning is, here is a helpful diagram. Data goes in to mysterious black box derived by computer scientists, and results come out. The problem with this model though is that it suffers from the Garbage in Garbage out problem.



Sort of at the same time, we started to see the democratization of AI/ML techniques. A lot of this is due to the intertwined relationship between the search community and the ML community. So any developments in search would likely translate to developments in ML techniques.

If you don’t know what Machine learning is, here is a helpful diagram. Data goes in to mysterious black box derived by computer scientists, and results come out. The problem with this model though is that it suffers from the Garbage in Garbage out problem.

Applications [edit]

This list has no **precise inclusion criteria** as described in the Manual of Style for standalone lists. Please improve this article by adding inclusion criteria. (Discuss) *(April 2018)*

Applications for machine learning include:

- Agriculture
- Anatomy
- Adaptive websites
- Affective computing
- Bioinformatics
- Brain–machine interfaces
- Cheminformatics
- Computer Networks
- Computer vision
- Credit-card fraud detection
- Data quality
- DNA sequence classification
- Economics
- Financial market analysis
- General game playing
- Handwriting recognition
- Information retrieval
- Insurance
- Internet fraud detection
- Linguistics
- Machine learning control
- Machine perception
- Machine translation
- Marketing
- Medical diagnosis
- Natural language processing
- Natural language understanding
- Online advertising
- Optimization
- Recommender systems
- Robot locomotion
- Search engines
- Sentiment analysis
- Sequence mining
- Software engineering
- Speech recognition
- Structural health monitoring
- Syntactic pattern recognition
- Telecommunication
- Theorem proving
- Time series forecasting
- User behavior analytics

And with the applications of machine learning being just about everything…

https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/

You end up with something like Tay.

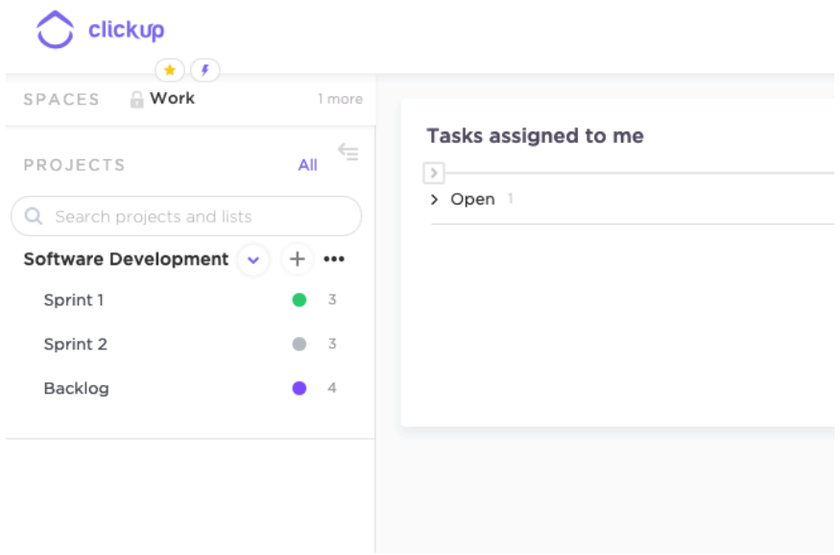
Anti-Pattern #4:
Trust in Black Box Algorithms

So here is anti-pattern 4: trust in black box algorithms

- Personalization is maybe doable in today’s resources
- Personalization research has sped up ML development
- Personalization research has democratized ML
- ML is useful for products, which require data
- There is precedent for collecting context data
- There is a general move to personalize everything

So let’s recap.

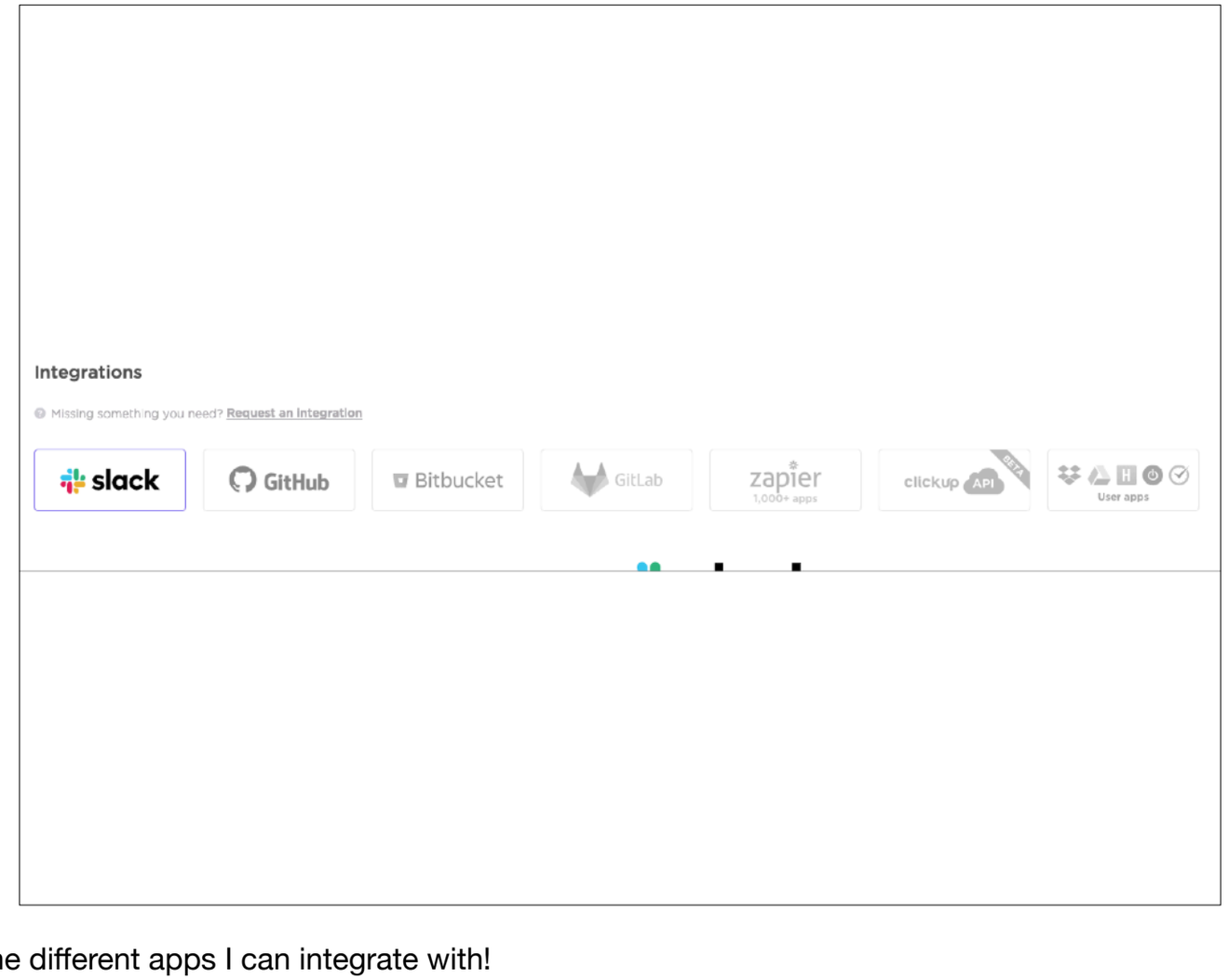
So where does that leave us today?



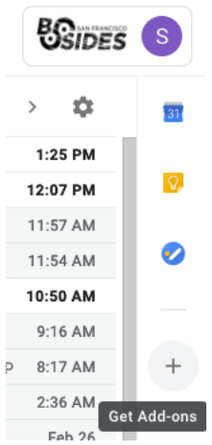
Let’s talk about how apps/services are today.

This kind of scenario may seem familiar to you. I am trying out this productivity app called Clickup. It has some means of handling things.

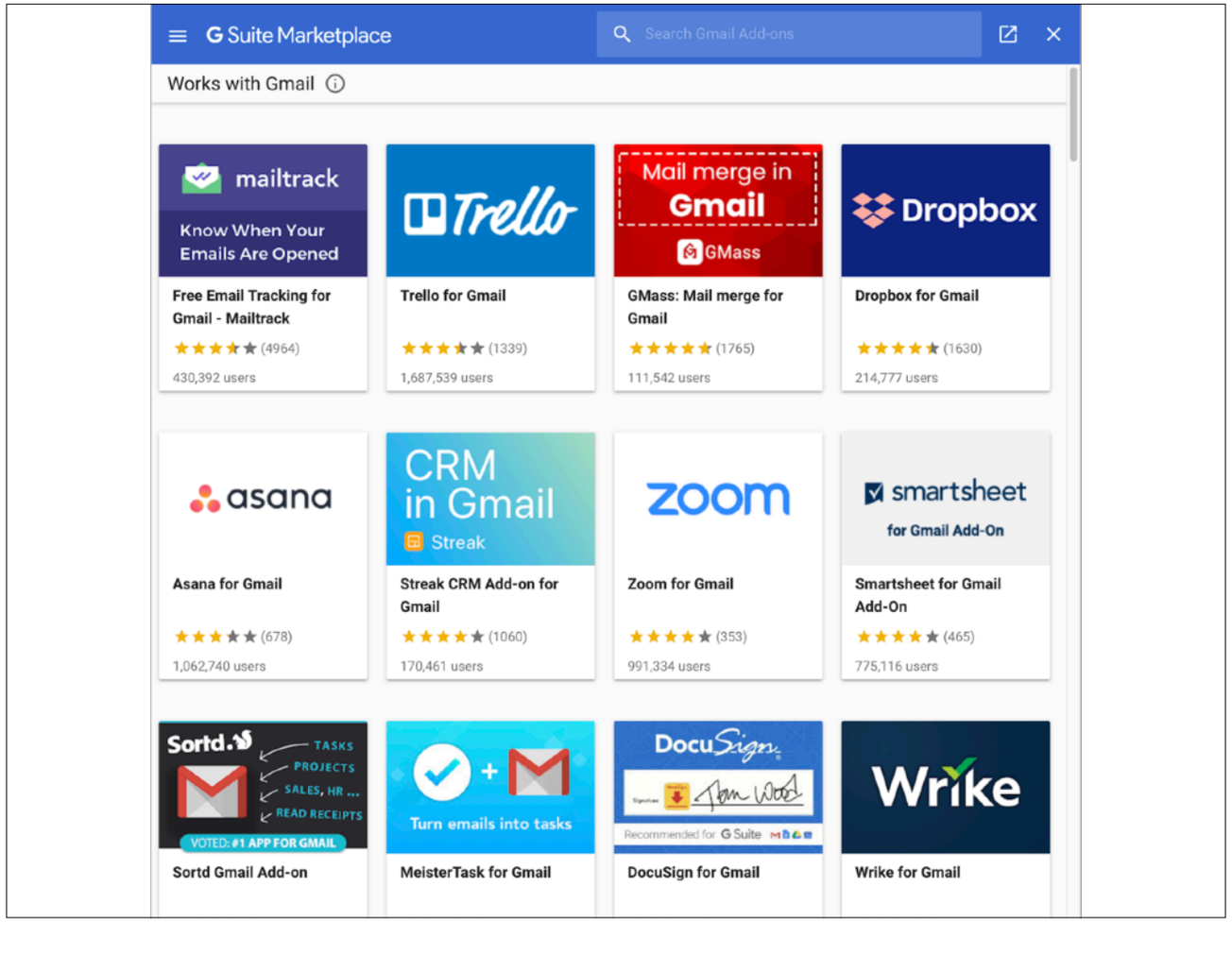
But it would be a lot more useful if like, I got it to talk to other things.



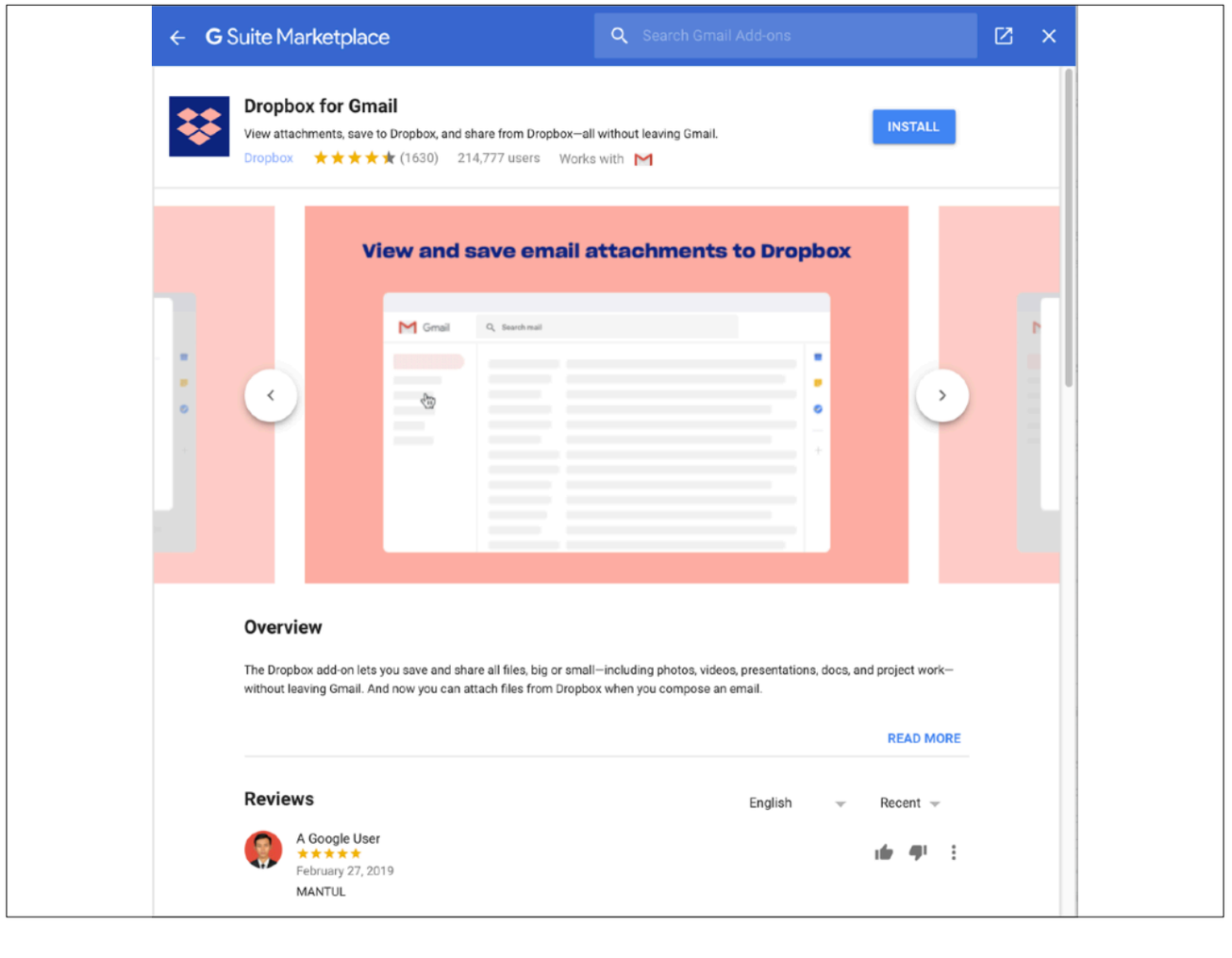
This maybe looks familiar! Look at all the different apps I can integrate with!



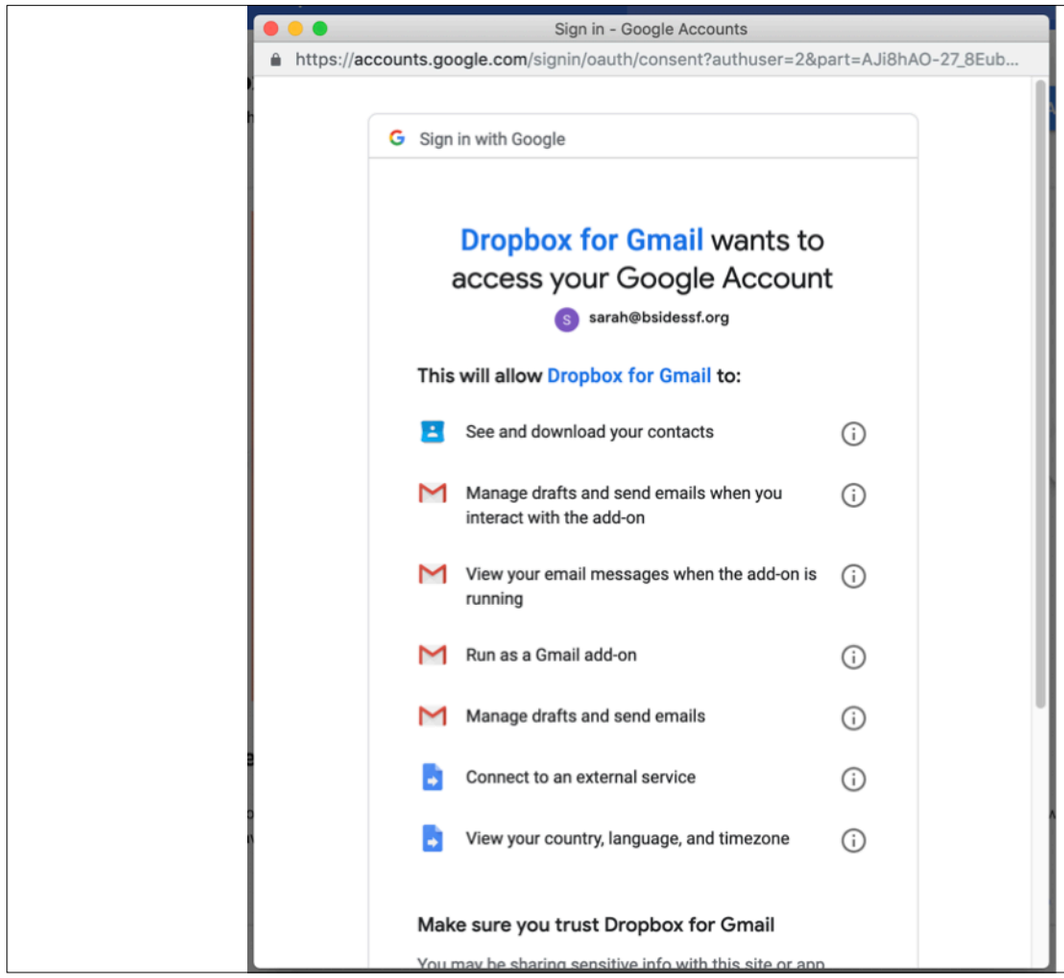
Even within gmail, I can click get add-ons, and integrate with a third party.



Look at all these products, what do they do?



Let’s choose dropbox.

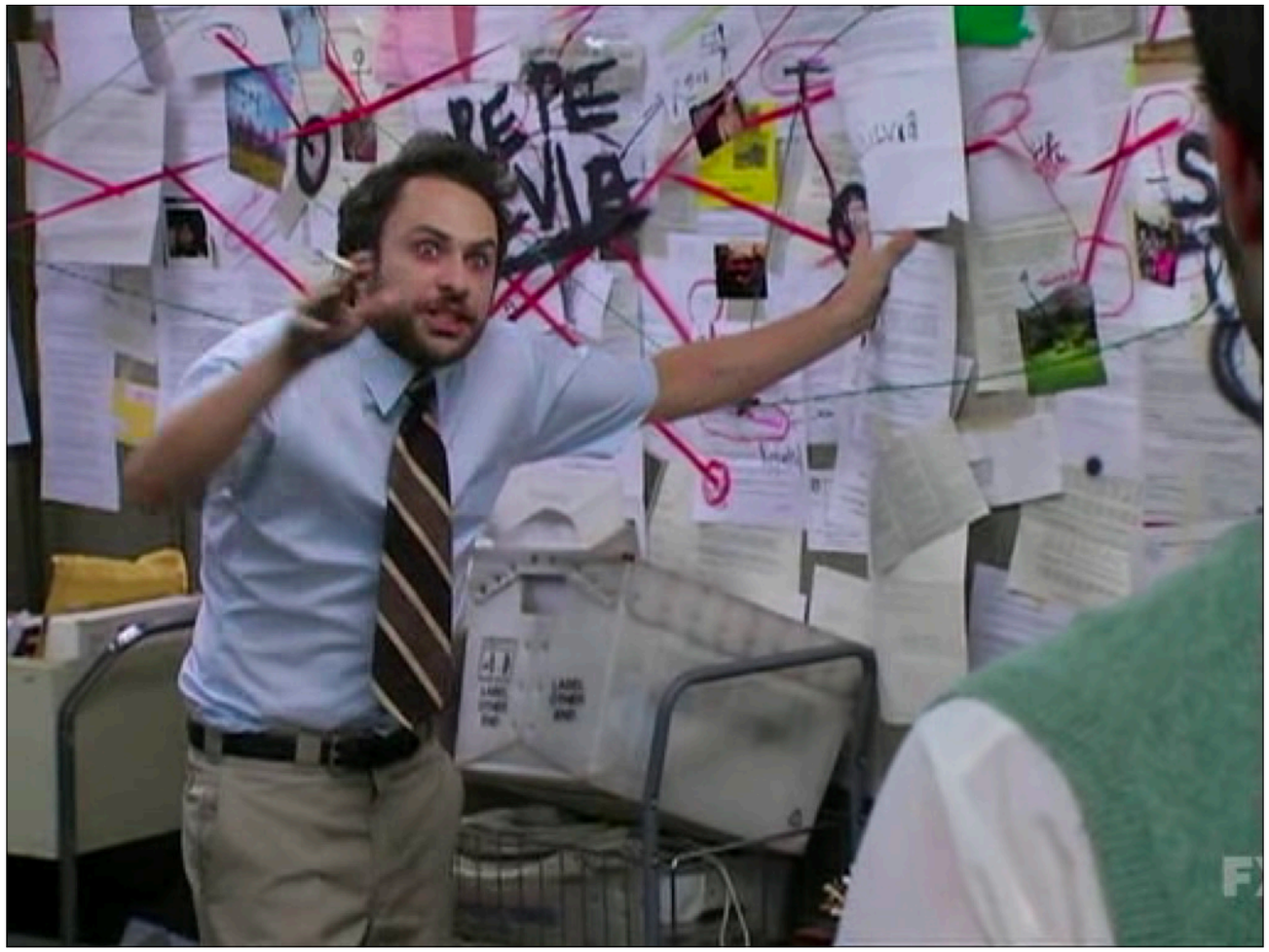


But if we go through the integration, we see that Dropbox requires a ton of permissions. Such as viewing your e-mails, managing your e-mails, connecting to some service, and seeing and downloading all your contacts.

Anti-Pattern #5:
Internet of Integrations

So here is anti-pattern 5: internet of integrations.

We want to personalize and make better products, but that can’t be done locally. So we do it on server, but all of the features we’d like, prompt us to integrate with other services.

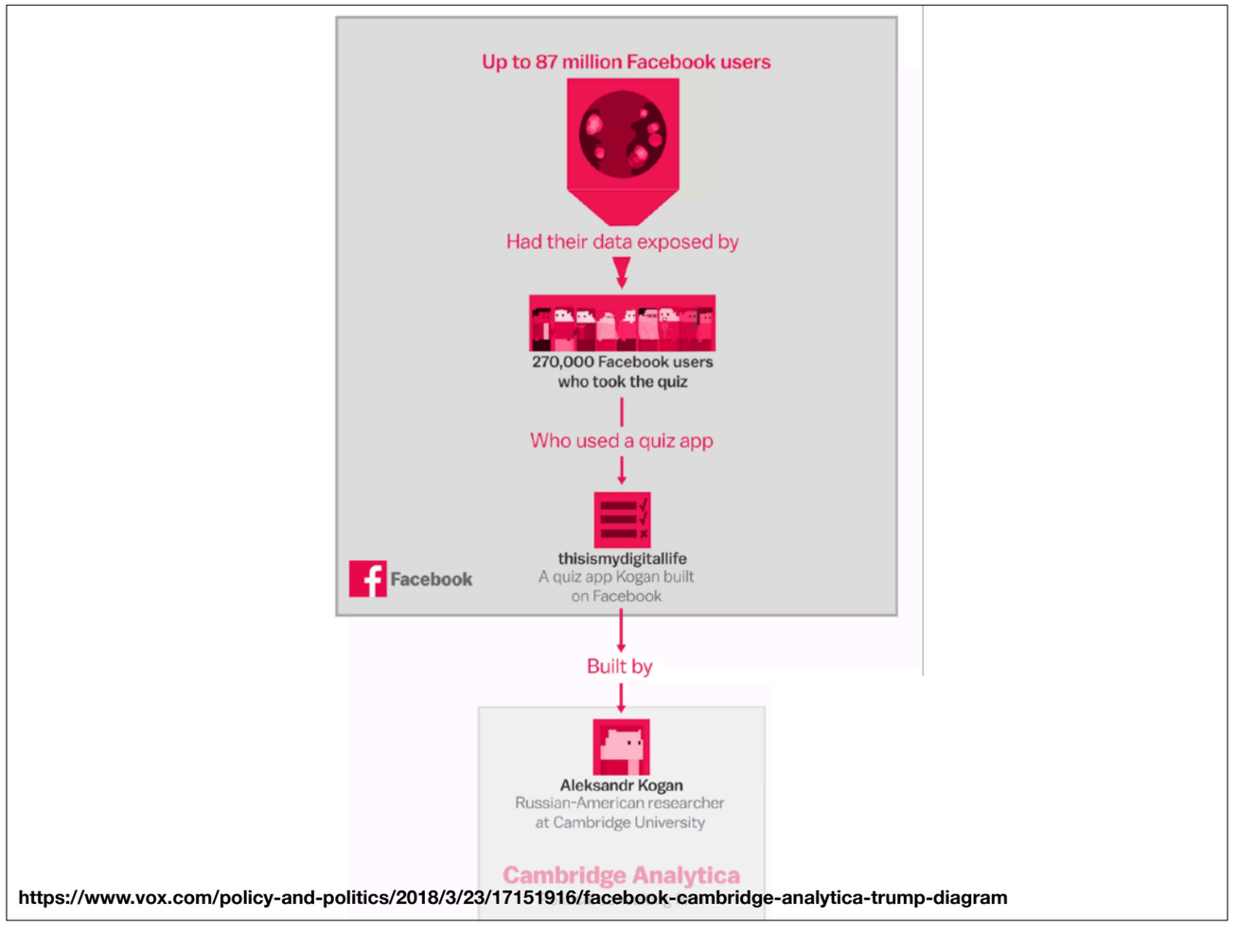


Which brings us back to this.

The New York Times

Facebook Says Cambridge Analytica Harvested Data of Up to 87 Million Users

I mean this.



https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram

What happened with Cambridge Analytica?

270,000 users took a quiz, which recorded their data to a third party that the users agreed to, and then that third party used that data for something else.

Point being, it's the past 20-30 years that has led us to this situation, where we have a third party need to collect all the data, but needs facebook to get the targeting and that viewership. And then now that they have that data, because it's theirs, they decide they can do ANYTHING ELSE with that data.

Anti-Pattern #6:
Unclear Data Ownership

So here is anti-pattern 6: Unclear data owner.

- Inference without Consent
- Lack of Transparency
- Erosion of Control
- Trust in Black-Box Algorithms
- Internet of Integrations
- Unclear Data Ownership

So let’s recap again!

We got here because the technology with good intentions brought us here. So these are the things we need to re-evaluate and re-examine before we can truly get a grasp on privacy.

FIN

(questions????)

@worldwise001