Human interaction and collaboration with ML models

Hendrik Strobelt, PhD
hen_str

@ Virtual Linz

IBM **Research**

Thank You,

Hanspeter Pfister
An Wang Professor of Computer Science, Harvard University
Verified email at seas.harvard.edu

Oliver Deussen
Professor of Computer Science, University of Konstanz
Verified email at uni-konstanz.de

Nils Gehlenborg
Assistant Professor - Department of Biomedical Informatics @ Harvard University
Verified email at hms.harvard.edu

Sebastian Gehrmann
Research Scientist, Google AI Language
Verified email at google.com

Alexander Lex
Associate Professor, SCI Institute, School of Computing, University of Utah
Verified email at sci.utah.edu

Alexander M. Rush
Associate Professor, Cornell University
Verified email at seas.harvard.edu

David Bau
PhD Student at MIT
Verified email at mit.edu

Antonio Torralba
Professor of Computer Science, MIT
Verified email at csail.mit.edu

Jun-Yan Zhu
Adobe Research, Carnegie Mellon University
Verified email at cs.cmu.edu

Bolei Zhou
Assistant Professor at The Chinese University of Hong Kong
Verified email at ie.cuhk.edu.hk

Daniel Keim
Professor of Computer Science, Data Analysis and Visualization, University of Konstanz
Verified email at uni-konstanz.de

Daniela Oelke
Verified email at freenet.de

Andreas Stoffel
University of Konstanz
Verified email at astoffel.de

Kasper Dinkla
IBM Research
Verified email at zurich.ibm.com

Romain Vuillemot
Ecole Centrale Lyon, LIRIS
Verified email at ec-lyon.fr

Adam Perer
Carnegie Mellon University
Verified email at cmu.edu

Christian Rohrdantz
University of Konstanz, Germany
Verified email at uni-konstanz.de

Johanna Beyer
Postdoctoral Fellow, SEAS, Harvard University
Verified email at seas.harvard.edu

Michael Schroeder
Professor in Bioinformatics, TU Dresden
Verified email at biotec.tu-dresden.de

Jörg Hakenberg
Illumina, Inc.
Verified email at illumina.com

Loic Alain Royer
Group Leader
Verified email at czbiohub.org

Marc Streit
Professor for Visual Data Science @ Johannes Kepler University Linz
Verified email at jku.at

Markus Hadwiger
Associate Professor of Computer Science, KAUST
Verified email at kaust.edu.sa

Dr. Marc Spicker
University of Konstanz
Verified email at uni-konstanz.de

Ulrik Brandes
ETH Zürich, Social Networks Lab
Verified email at ethz.ch

Enrico Bertini
NYU Tandon School of Engineering
Verified email at nyu.edu

Payel Das
Manager and Principal Research Staff Member, AI, IBM Watson Research

hendrik.strobelt.com

Research Scientist, IBM Research
Explainability Lead, MIT-IBM Watson AI Lab
Visiting Researcher, MIT CSAIL

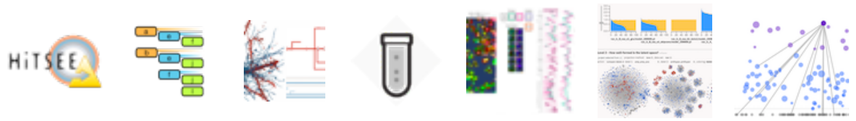
Vis for ML/AI + Collaborative AI

LSTMVis | GLTR | ExBERT | GANPaint

Vis for Text & Documents

DocumentCards

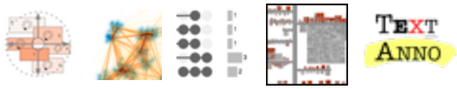
Vis for the Sciences



MolExplorer

HITSee | vials.io | IVELS

Visual Encoding



Upset Plots

hendrik.strobelt.com

Research Scientist, IBM Research
Explainability Lead, MIT-IBM Watson AI Lab
Visiting Researcher, MIT CSAIL

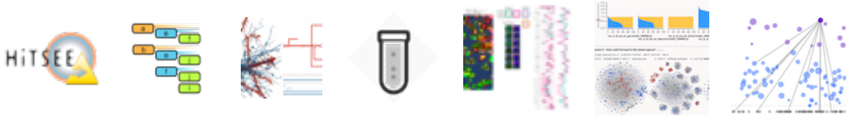
Vis for ML/AI + Collaborative AI

LSTMVis | GLTR | ExBERT | GANPaint

Vis for Text & Documents

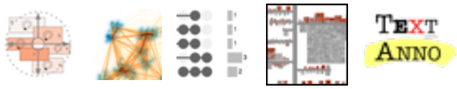
DocumentCards

Vis for the Sciences



HITSee | vials.io | IVELS | MolExplorer

Visual Encoding



Upset Plots

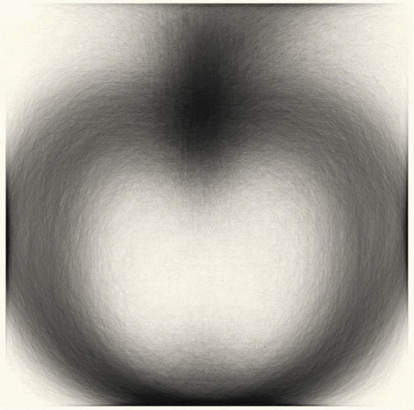
Forma Fluens

Each period of culture produces an art of its own which can never be repeated. We are living through a widely distributed amateur creativity. We are in the age of sharing, in the age of user-generated content. In Forma Fluens (Latin: Flowing Form) you are not a passive observer or consumer. With our DoodleMaps, you can be the author of one of the stories that emerge from the exploration of millions of drawings. Or you can generate new icons from the overlap of thousands of drawings with IconoLap. Finally, in the video Points in Movement you can observe an overlap of millions of drawings and find out how all humanity draws some forms in the same way.

DoodleMaps

Explore millions of drawings.

Points in Movement

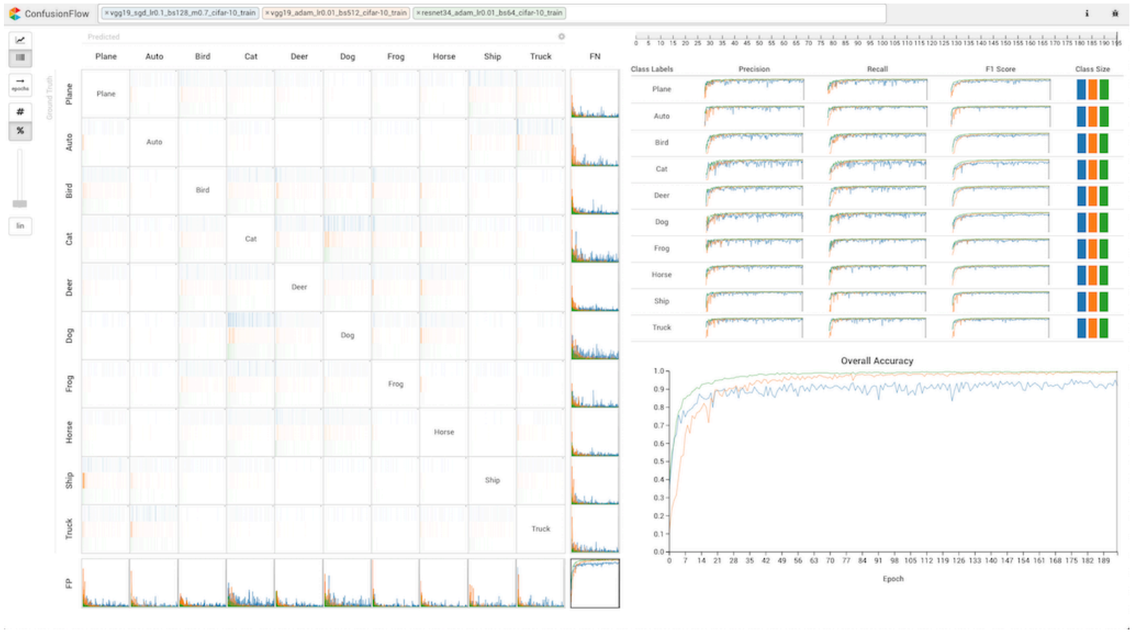


Observe the human cognitive perception reflected in drawings.

Icono Lap

Generate new icons from the overlap of thousands of drawings.

http://formafluens.io



https://github.com/ConfusionFlow/confusionflow

What is light ?

wave?
photon?

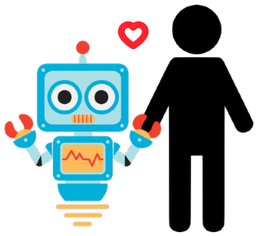
model(nature)

model()

How does a ANN work? 

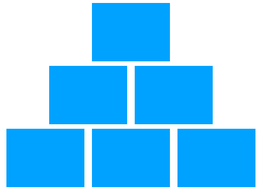
observe

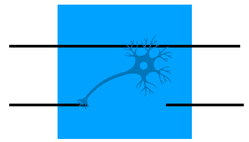
interact



collaborate

model() =





non-linear statistical method?
generalization?
memorization?
magic ?
...

model(model(nature))

Three ideas

- Dissecting a GAN
- Gaentidodes: Detecting machine-generated text
- Become an ExBert to understand chemical reactions

Human-AI Collaboration for Generation: GANpaint

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba.
"GAN Dissection: Visualizing and Understanding Generative Adversarial Networks."
In *International Conference on Learning Representations*. 2018.

Bau, David, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba.
"Semantic photo manipulation with a generative image prior."
ACM Transactions on Graphics (TOG) 38, no. 4 (2019): 1-11.

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba.
"Understanding the role of individual units in a deep neural network."
Proceedings of the National Academy of Sciences (2020).

z

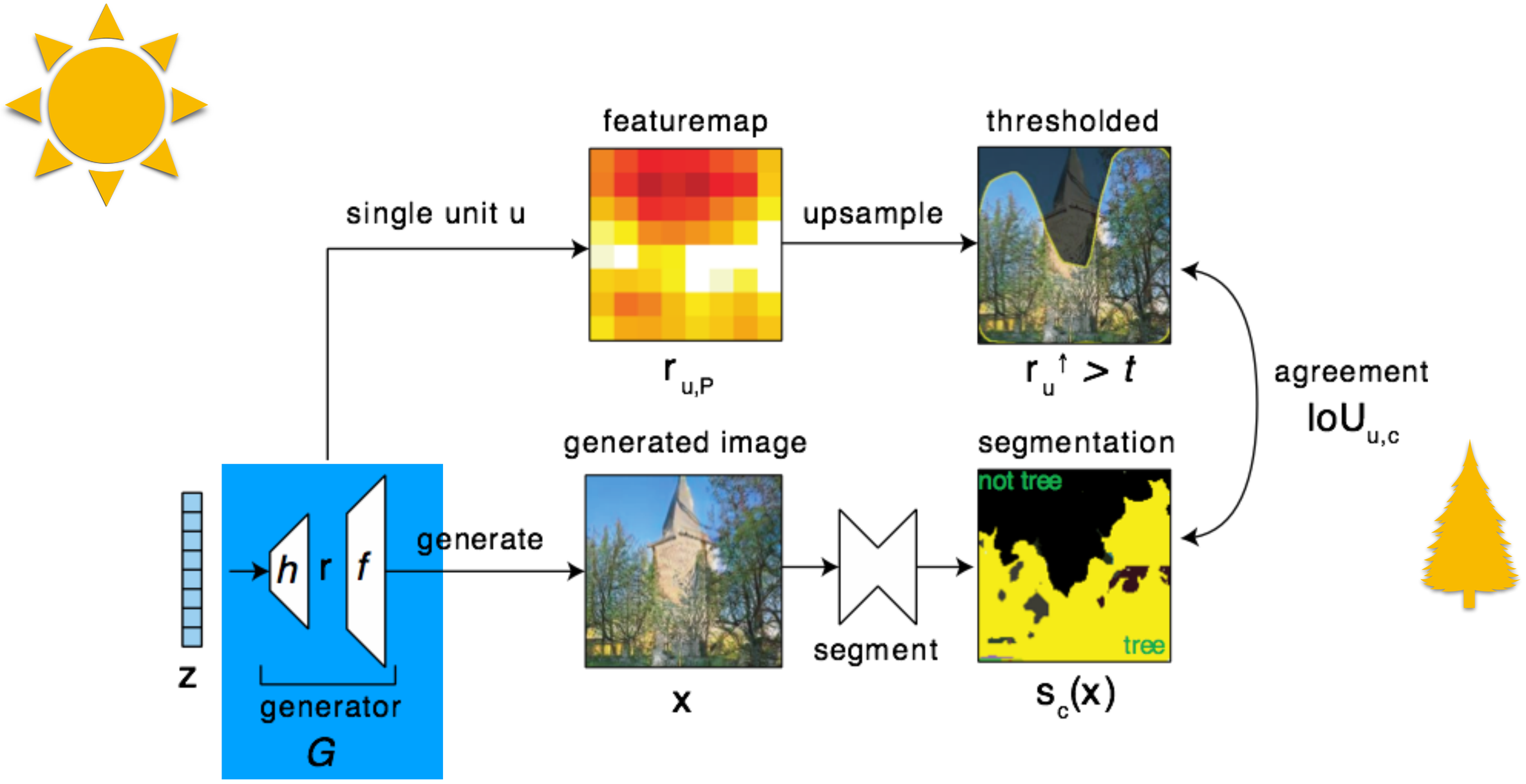
Generator

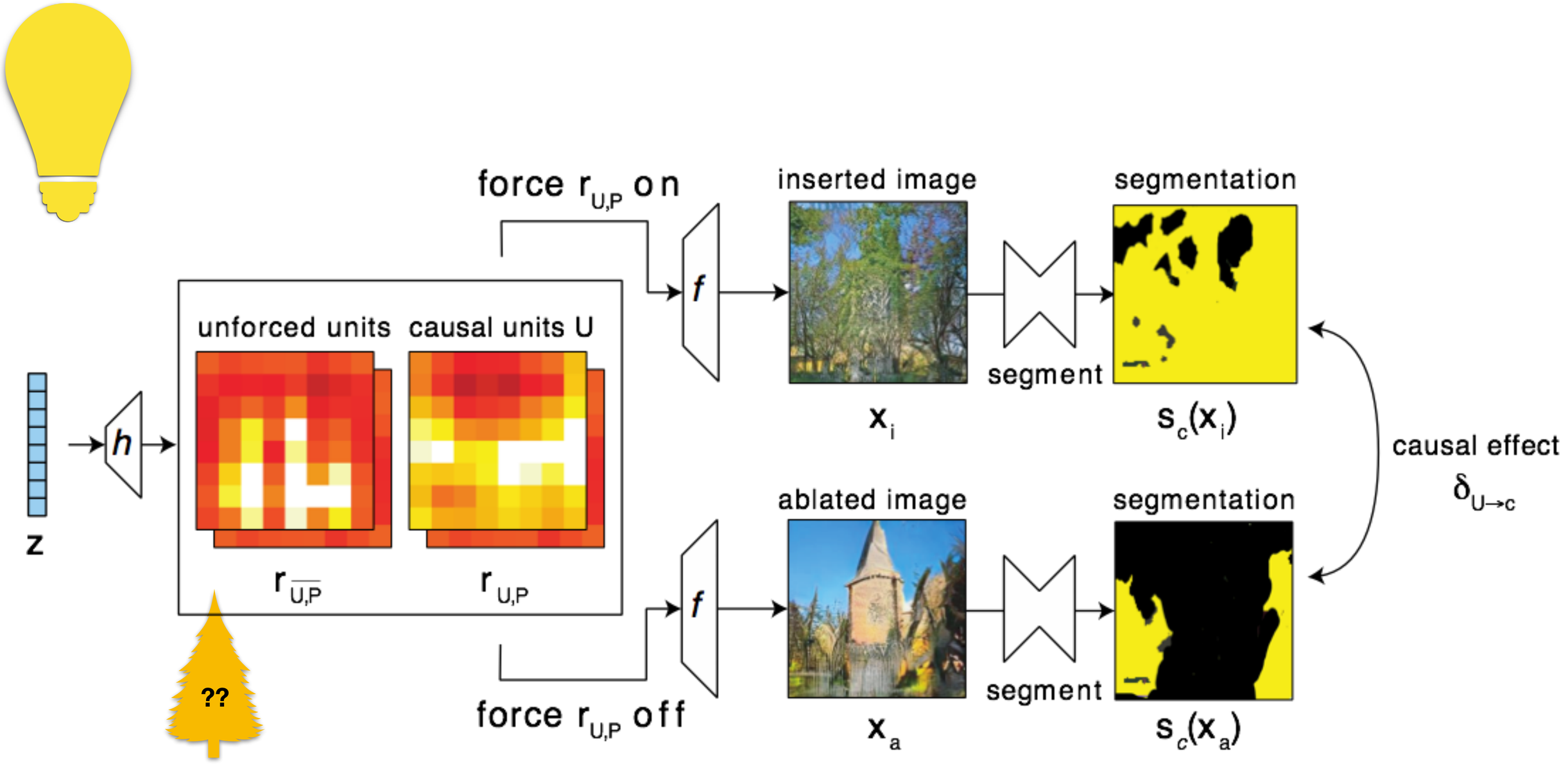
GAN invented photo

z

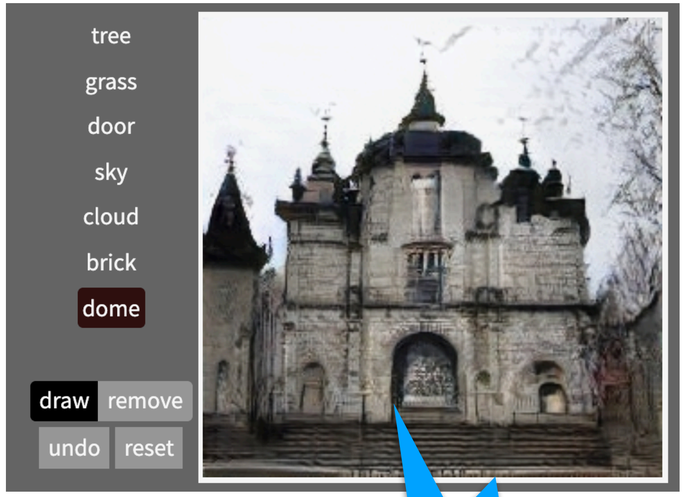
Generator

GAN invented photo


featuremap
thresholded
single unit u
upsample
$r_{u,P}$
$r_u^{\uparrow} > t$
agreement
$IoU_{u,c}$
generated image
segmentation
generate
not tree
tree
h
r
f
z
generator
G
segment
x
$s_c(x)$


force $r_{U,P}$ on
inserted image
segmentation
unforced units
causal units U
f
segment
causal effect
$\delta_{U \to c}$
h
z
$r_{\overline{U,P}}$
$r_{U,P}$
x_i
$s_c(x_i)$
ablated image
segmentation
??
f
segment
force $r_{U,P}$ off
x_a
$s_c(x_a)$

Tree	Door	Dome
3	5	99
45	77	14
12	8	33
6	41	74



z

Generator

GAN invented photo

- edited -

Collaborative Generation

Gæntidotes


DCGAN 2014

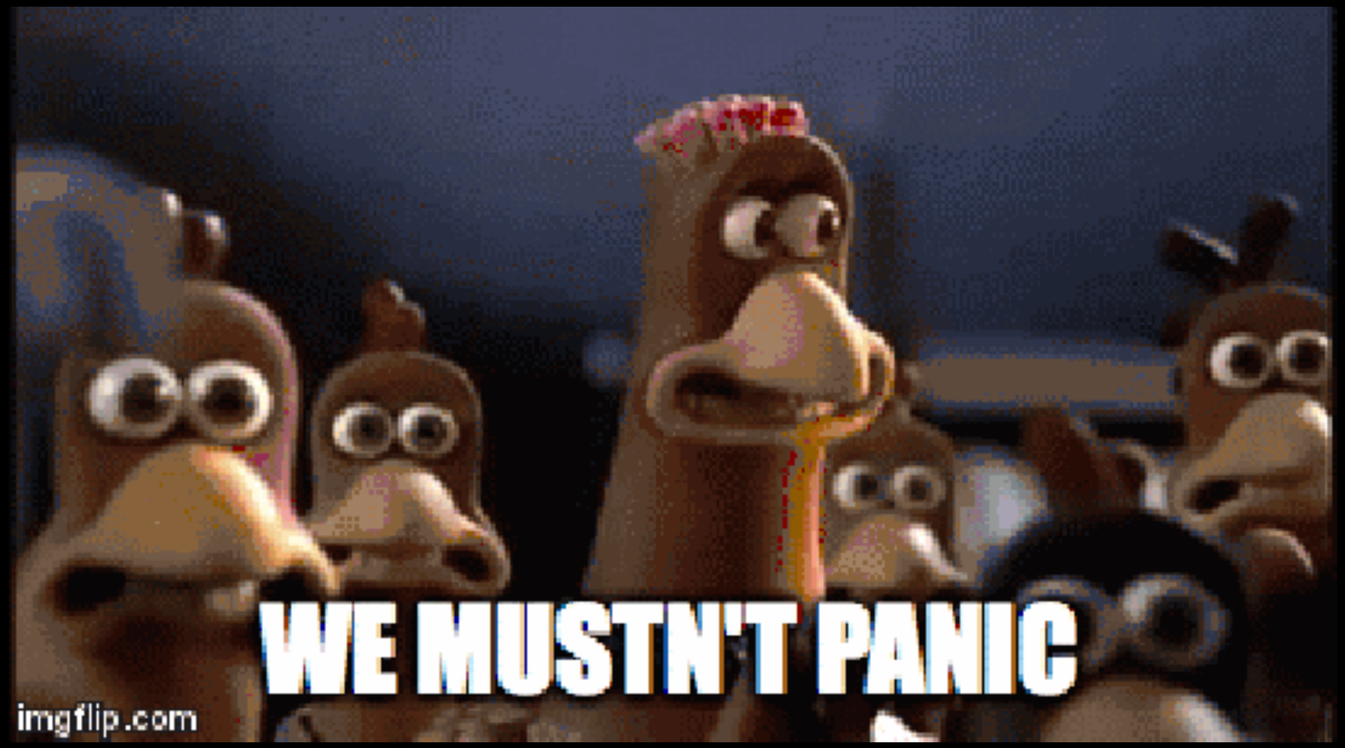

ProGAN 2017


StyleGAN 2018

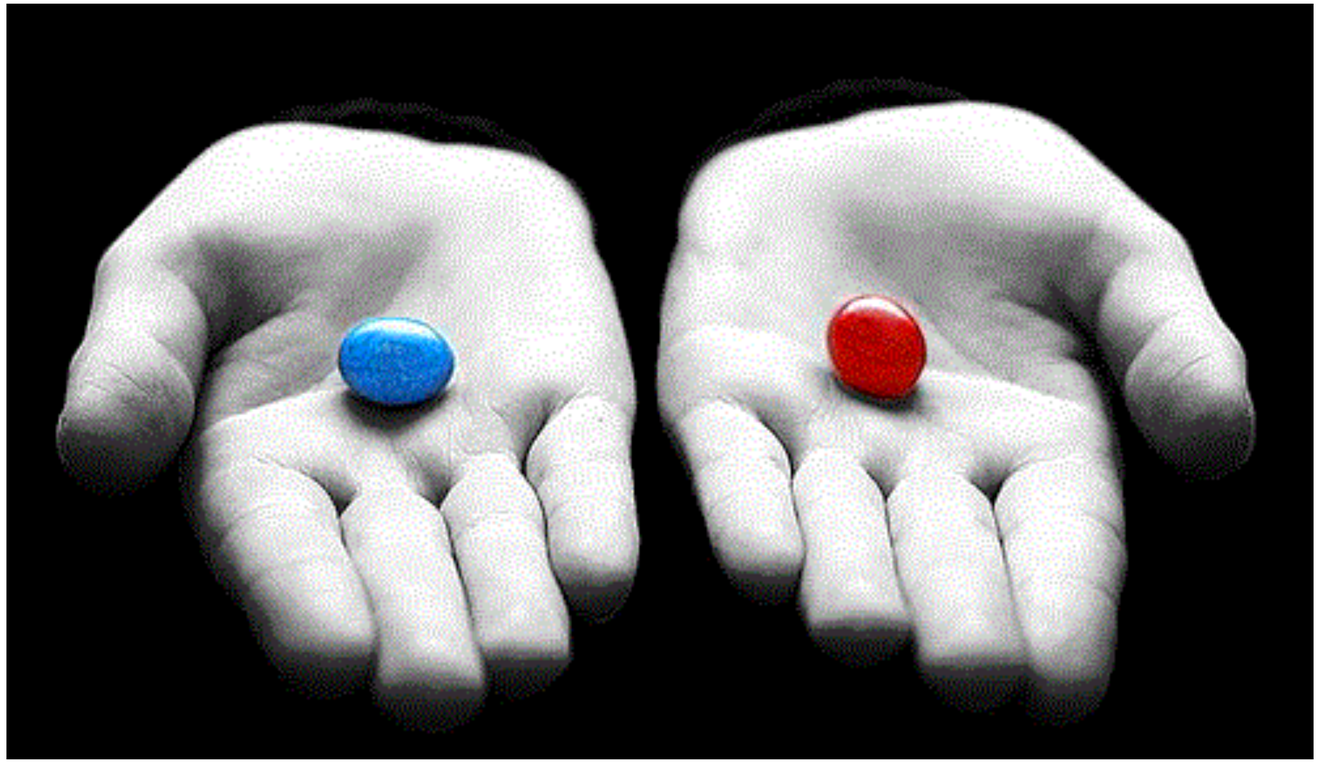
Technology

Elon Musk's OpenAI builds artificial intelligence so powerful it must be kept locked up for the good of humanity

February 15, 2019

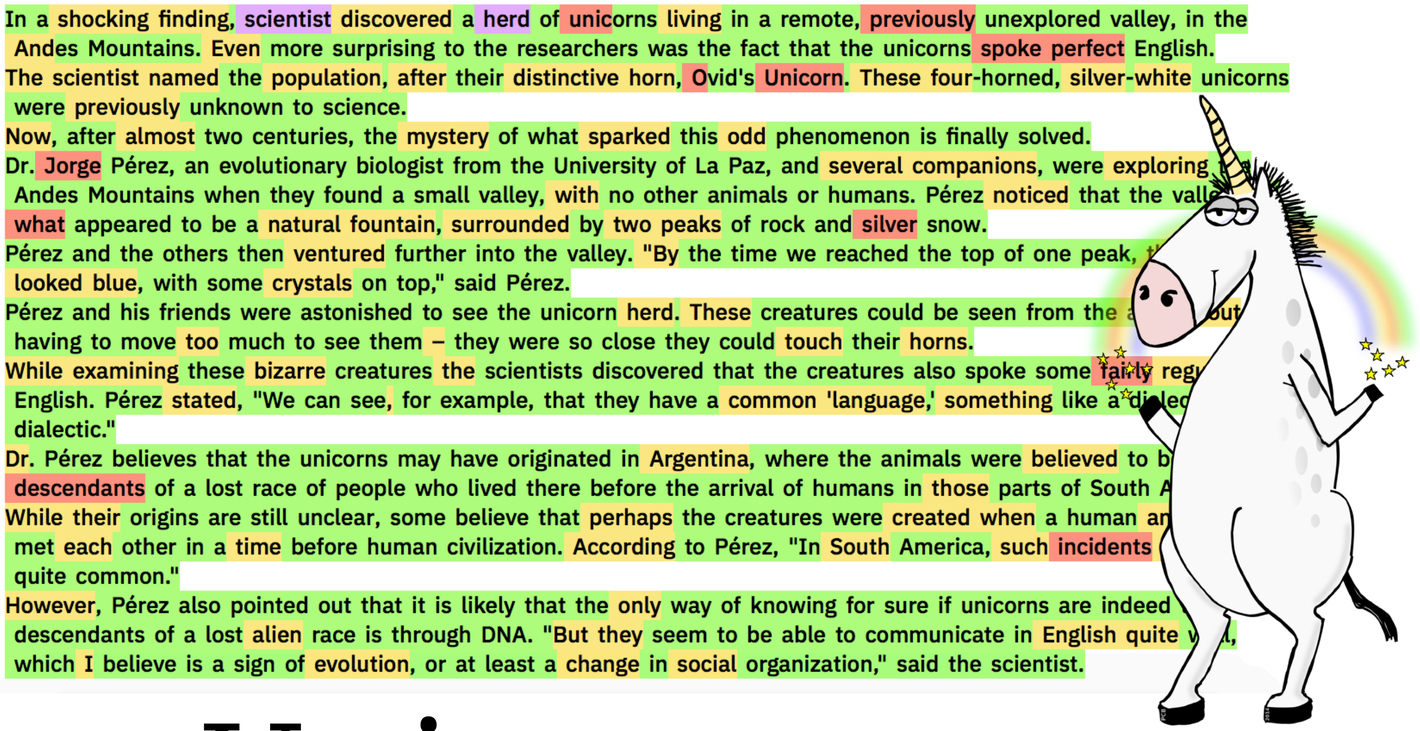

WE MUSTN'T PANIC
imgflip.com

Sign Non-Machine



Detect Machine

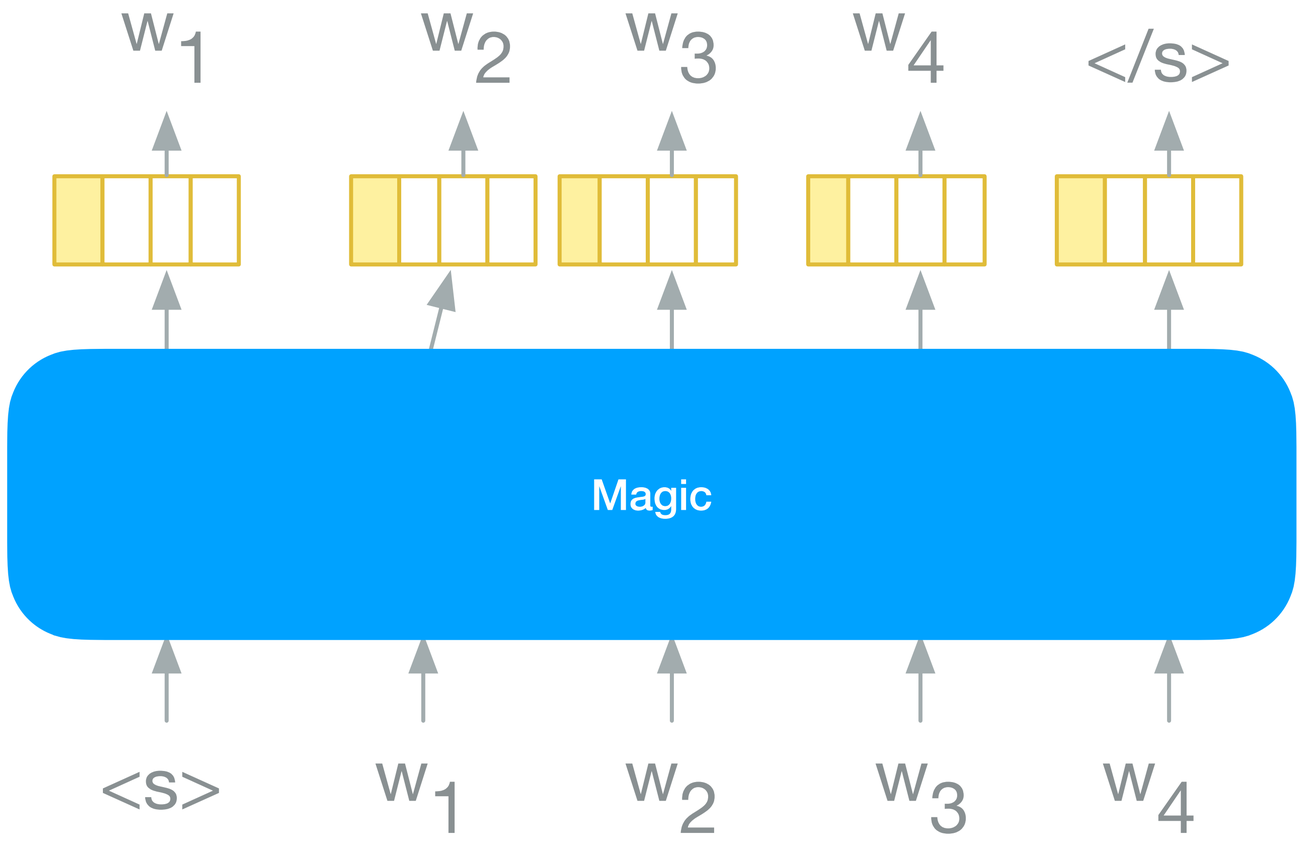
Detect Machine



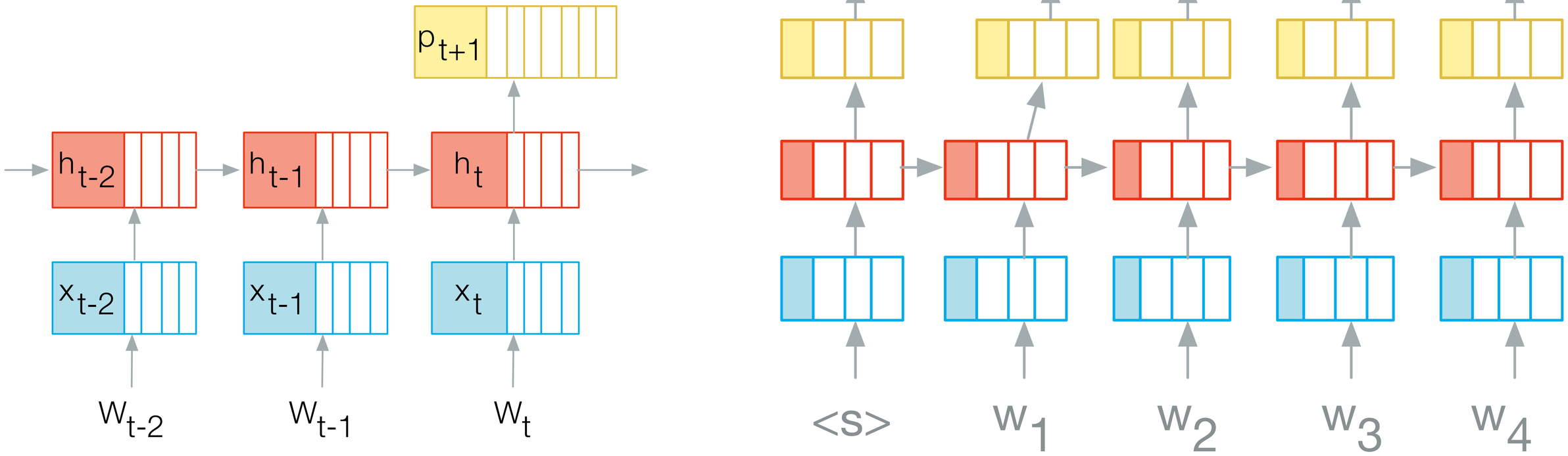
Catching Unicorns with GLTR

Gehrmann, Sebastian, Hendrik Strobelt, and Alexander M. Rush.
"GLTR: Statistical Detection and Visualization of Generated Text."
In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 111-116. 2019.

Sampling / NLG



Sampling / NLG



Sampling from a language model

- at each sampling-step there are many (~50k) differently probable events (words) to chose from 

- greedy strategy: take the most likely event all the time
- top-n strategy: randomly sample from top-n events
- beam search: keep the top-n events to see which one will be likely including the next sample steps

Sampling from a language model

- at each sampling-step there are many (~50k) differently probable events (words) to chose from 

- greedy strategy: take the **most likely** event all the time
- top-n strategy: randomly sample from **top-n events**
- beam search: keep the **top-n events** to see which one will be likely including the next sample steps

Idea

Can we detect how likely a text is sampled from a model?

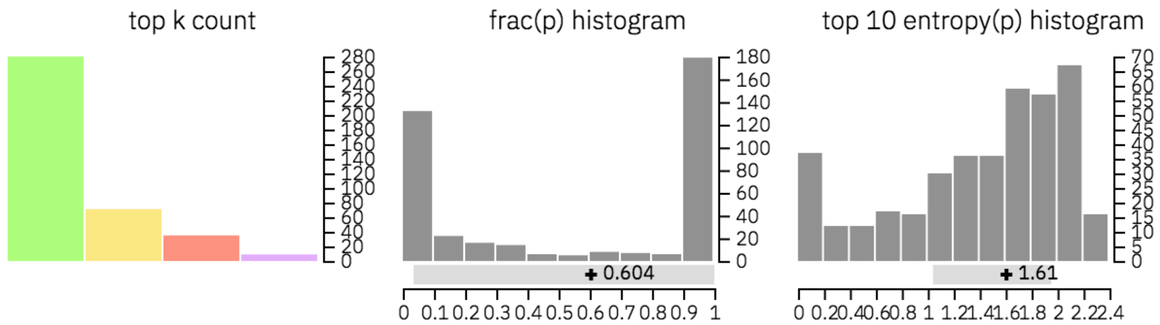
—> Kind of. We can indicate for each word the rank of this word in model predictions given the left-side context

Sample - Sanity Check

Top K | Frac P | Colors (top k): 10 | 100 | 1000

The following is a transcript from The Guardian's interview with the British ambassador to the UN, John Baird.
Baird: The situation in Syria is very dire. We have a number of reports of chemical weapons being used in the country. The Syrian opposition has expressed their willingness to use chemical weapons. We have a number of people who have been killed, many of them civilians. I think it is important to understand this.
There are many who are saying that the chemical weapons used in Syria are not only used to destroy people but also to destroy the Syrian people. The Syrian people have been suffering for many years. The regime is responsible for that suffering. They have been using chemical weapons. They have killed many people, and they continue to kill many more.
I think that the international community has to take a position that the Assad regime has a responsibility for that suffering. It must take a stand that we are not going to allow the Syrian government to use chemical weapons on civilians, that we are not going to allow them, and that we do not condone their use.
We have a lot of people who believe that the regime is responsible for this suffering, and that they are responsible for this suffering, and that they are responsible for the use of chemical weapons. I think that we need to be clear about that.
We must be clear that the use of chemical weapons by any country, including Russia and Iran, is a violation of international law. We are not going to tolerate that. We do not tolerate that. And we have the responsibility to ensure that the world doesn't allow the Assad regime to use chemical weapons against civilians.
Baird: It seems that there are a range of people that are saying that we are not allowed to use chemical weapons in Syria. There are many who say we are not allowed to use chemical weapons in Syria.
I think there are a lot of people that are saying that we are not allowed to use chemical weapons in Syria. I think that we have to take a stand that we are not going to allow the Assad regime to use chemical weapons on civilians, that we are not going to tolerate that. We have to take a stand that we are not going to allow Russia and Iran to use chemical weapons on civilians.
Baird: I think it is important for us to understand that the use of chemical weapons in Syria is an extremely dangerous situation. I think there has been very little information from the UN that the regime has used any chemical weapons. We have not seen any evidence that they are using them.
We have to understand that the use of chemical weapons is very dangerous.

Sample - NYTimes



Top K | Frac P | Colors (top k): 10 | 100 | 1000

MONEY, Miss. – Along the edge of Money Road, across from the railroad tracks, an old grocery store rots.

In August 1955, a 14-year-old black boy visiting from Chicago walked in to buy candy. After being accused of whistling at the white woman behind the counter, he was later kidnapped, tortured, lynched and dumped in the Tallahatchie River.

The murder of Emmett Till is remembered as one of the most hideous hate crimes of the 20th century, a brutal episode in American history that helped kindle the civil rights movement. And the place where it all began, Bryant's Grocery & Meat Market, is still standing. Barely.

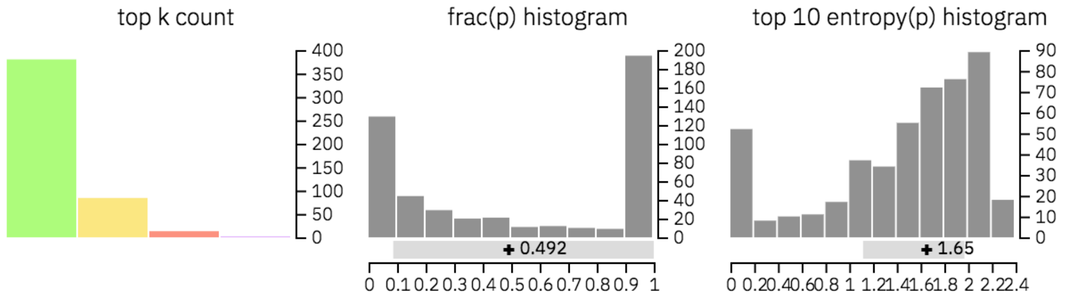
Today, the store is crumbling, roofless and covered in vines. On several occasions, preservationists, politicians and business leaders – even the State of Mississippi – have tried to save its remaining four walls. But no consensus has been reached.

Some residents in the area have looked on the store as a stain on the community that should be razed and forgotten. Others have said it should be restored as a tribute to Emmett and a reminder of the hate that took his life.

As the debate has played out over the decades, the store has continued to deteriorate and collapse, even amid frequent cultural and racial reckonings across the nation on the fate of Confederate monuments. At stake in Money and other communities across the country is the question of how Americans choose to acknowledge the country's past.

“It's part of this bigger story, part of a history that we can learn from,” said the Rev. Wheeler Parker, 79, a pastor in suburban Chicago and a cousin of Emmett's who went with him to Bryant's Grocery that day. –“The store should be one of the places we share Emmett's story.”

Sample - Unicorn



Top K | Frac P | Colors (top k): 10 | 100 | 1000

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

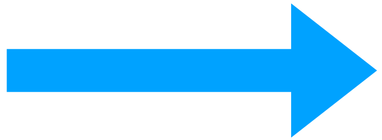
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

Human Subjects Study

54.2%

acc. without highlights



72.3%

acc. with GLTR highlights

35 students, 5 texts in 90 sec, mix of generated and non-generated text

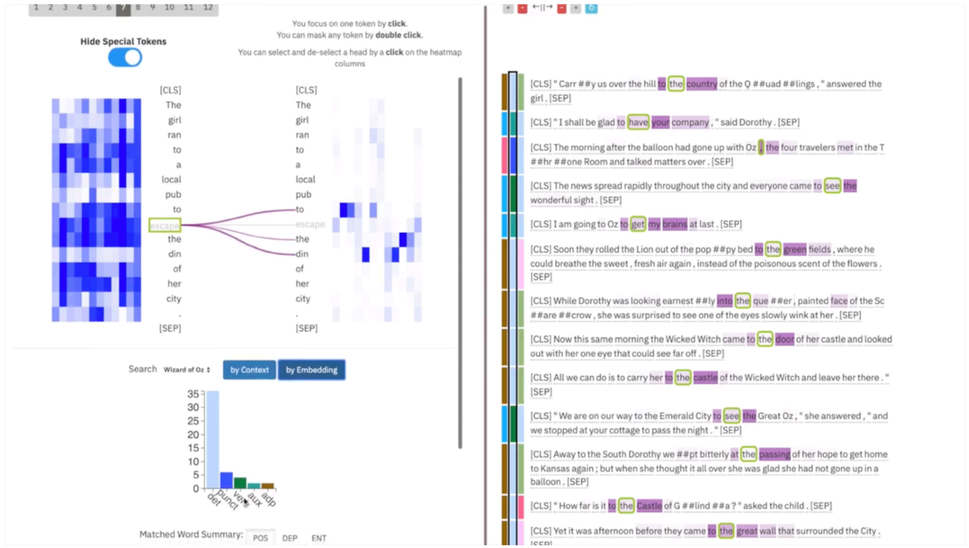
Real-World Impact of GLTR

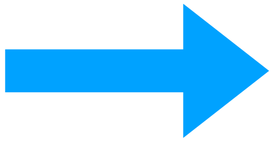
credit: Sebastian Gehrmann

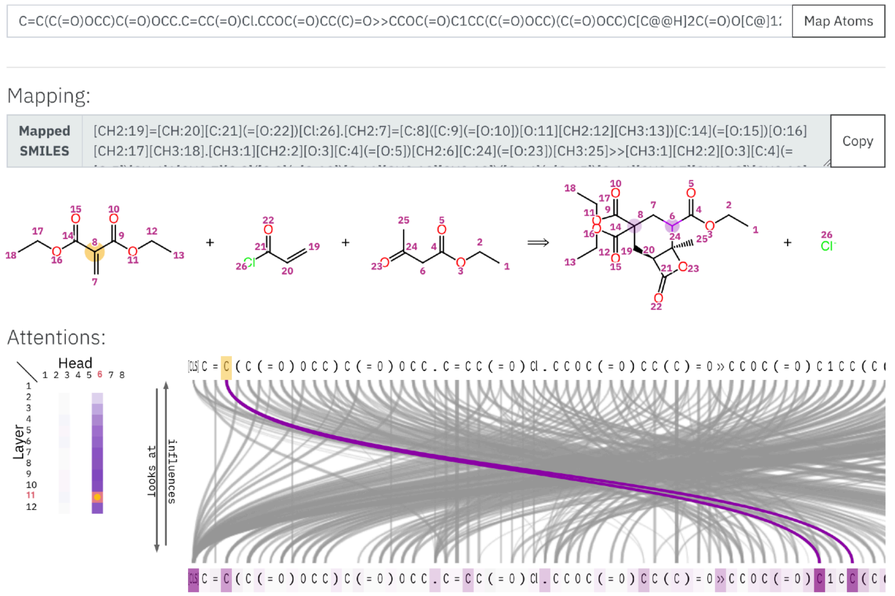
Researchers need to be thought leaders

- GLTR to date has almost 100,000 page views and was covered by 20+ news outlets, we used this opportunity to try to paint an accurate picture of the state of AI for the public
- Discussions with OpenAI lead to thought exchange investigating the prevention of abuse of large models
- We have been invited to DARPA events and advised the government-led global engagement center on prevention of foreign propaganda, we are collaborating on deploying and testing GLTR-like methods at scale

Attention is (nearly) all you need







ExBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models.
B. Hoover, H. Strobelt, and S. Gehrmann.
In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 187–196, Online, July 2020. ACL.

Unsupervised Attention-Guided Atom-Mapping.
P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, and T. Laino.
Chemarxiv 2020
ML Interpretability for Scientific Discovery, Workshop@ICML 2020

ExBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models.

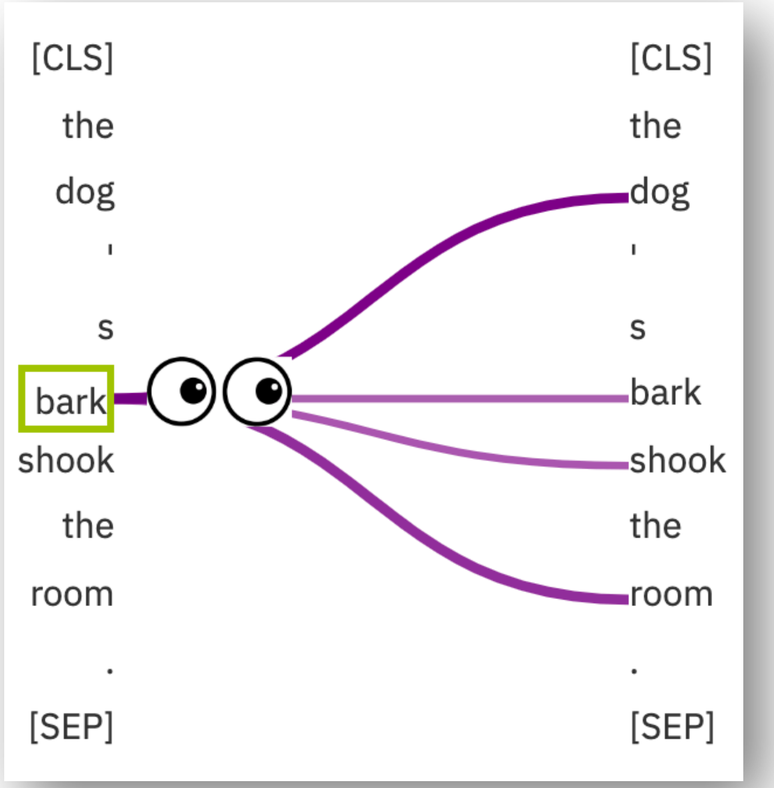
B. Hoover, H. Strobelt, and S. Gehrmann.

In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 187–196, Online, July 2020. ACL.

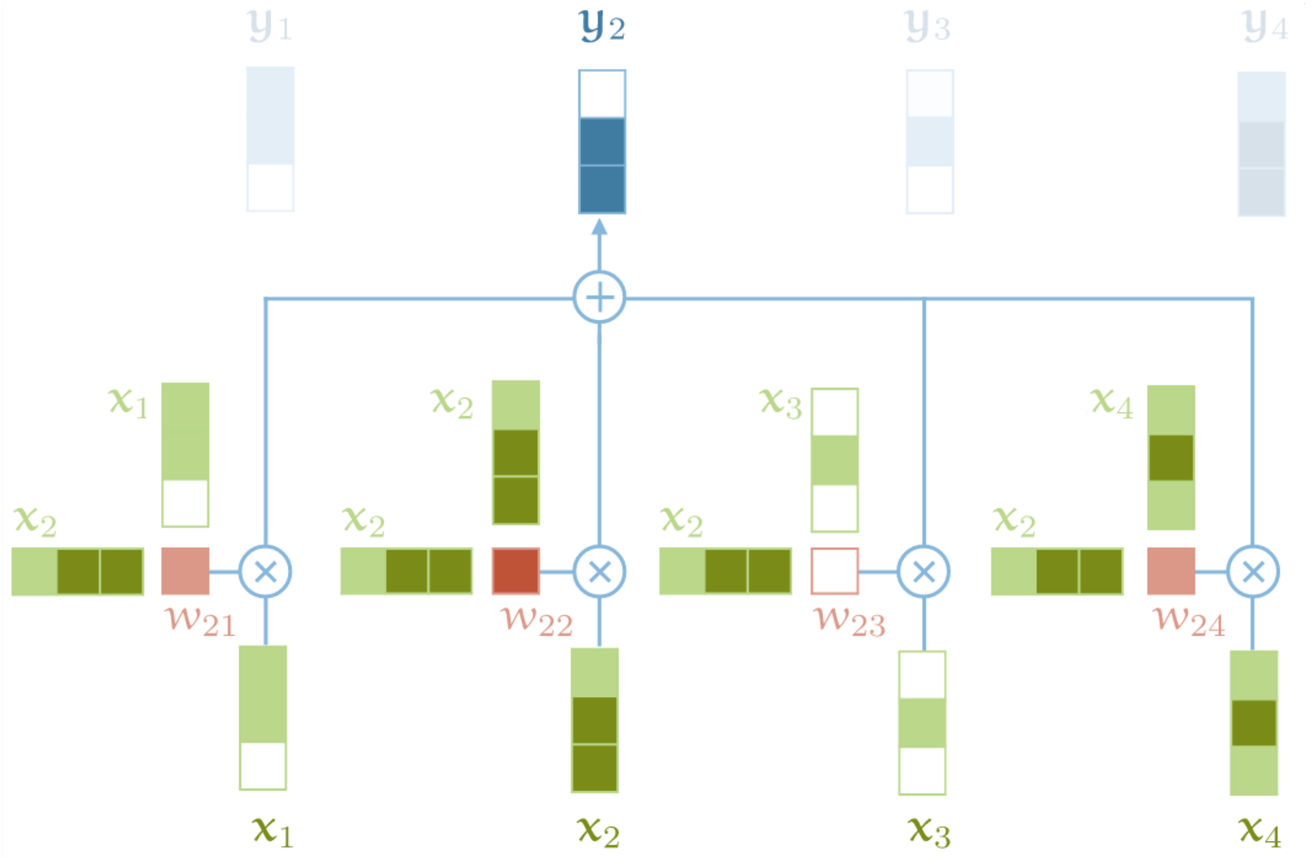
Background

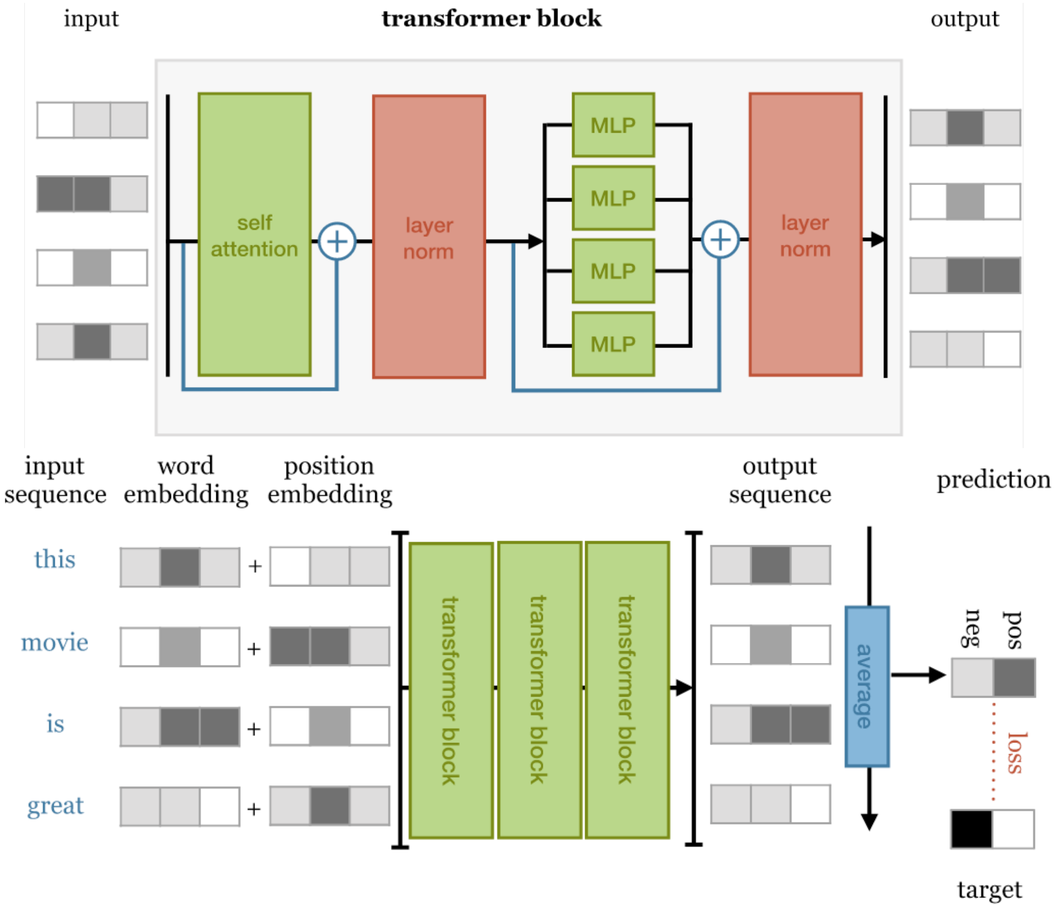
- What are Transformers?
  - A Deep Learning Architecture that uses **Self-Attention** to be SOTA in many NLP tasks
- What is Self-Attention?
  - Mechanism that allows each token to refine its **Representation** (or **Embedding**) by looking at each other token in its context
- Representations? Embeddings?
  - High dimensional vectors that computers use to assign meaning tokens

Self-Attention

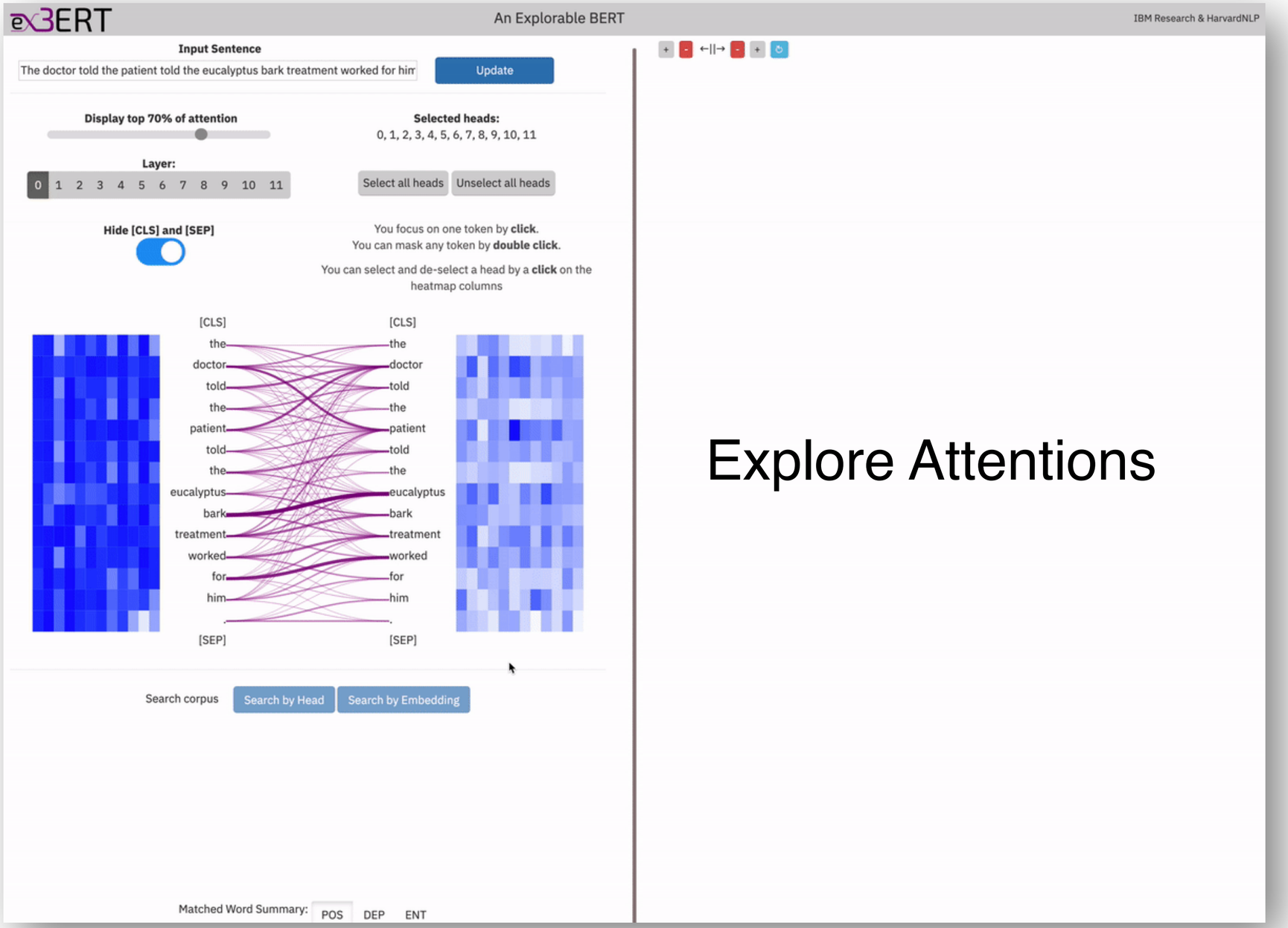


Self Attention in Transformer Models (NLP)



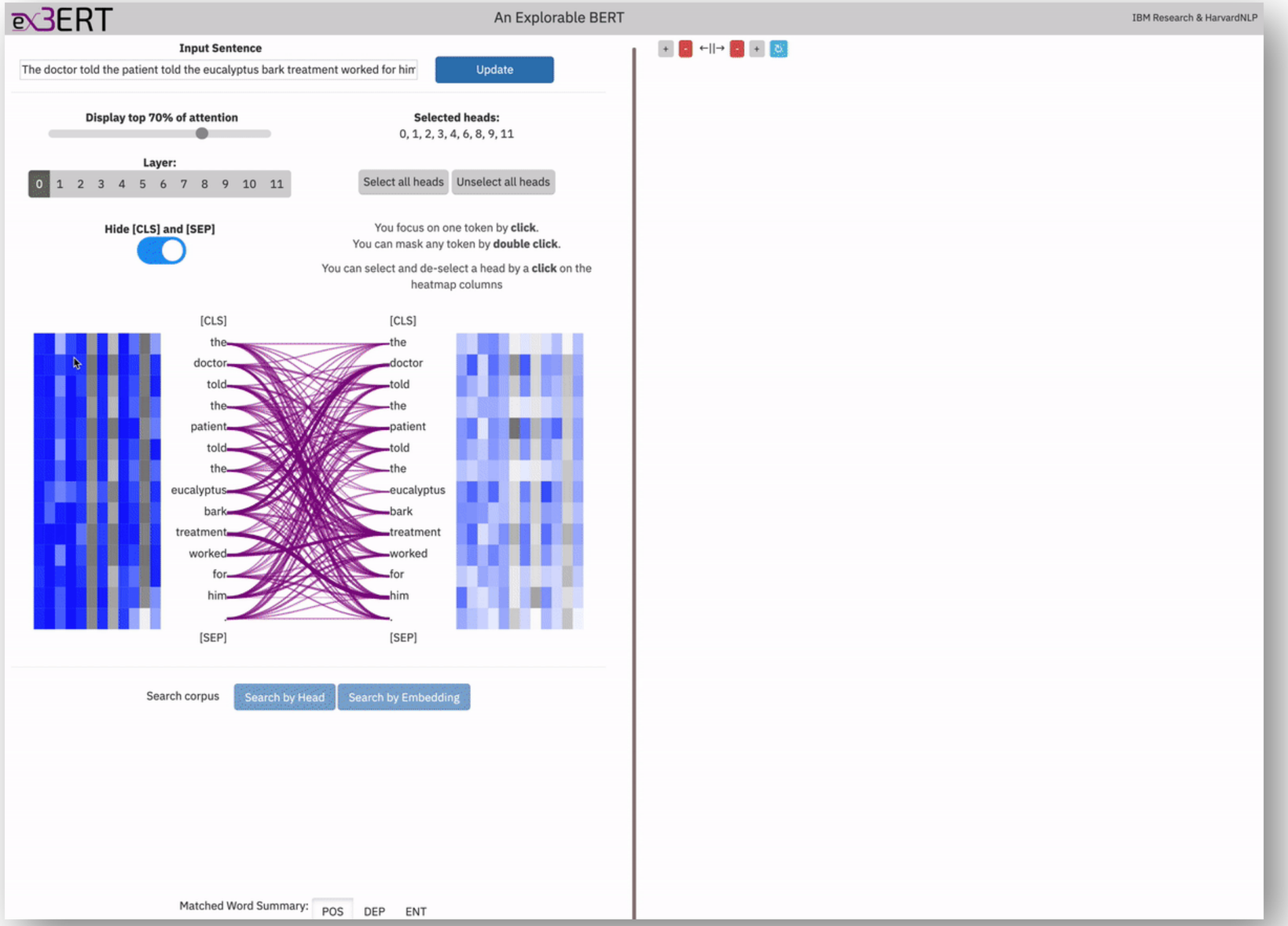


http://www.peterbloem.nl/blog/transformers



Explore Attentions

Explore Representations

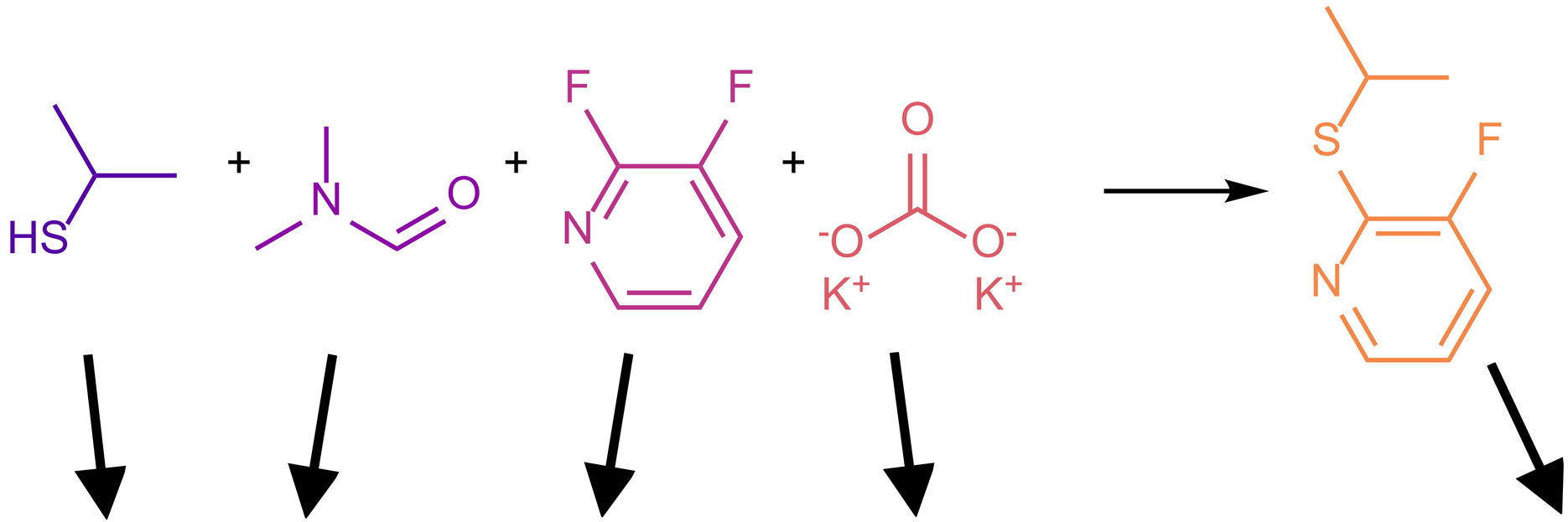


RXNMapper: Unsupervised Attention-Guided Atom-Mapping

Unsupervised Attention-Guided Atom-Mapping.
P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, and T. Laino.
Chemarxiv 2020
ML Interpretability for Scientific Discovery, Workshop@ICML 2020

Chemical reactions can be represented as **text**.

precursors >> products

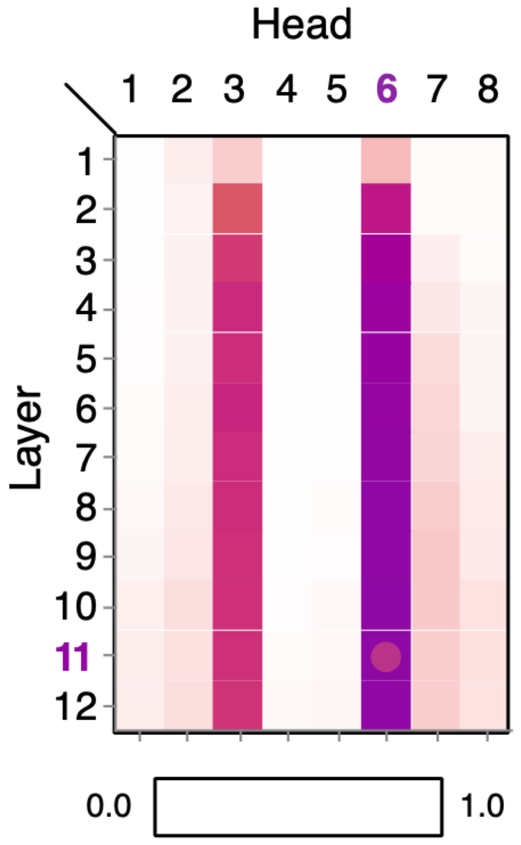


SMILES: CC(C)S.CN(C)C=O.Fc1cccnc1F.O=C([O-])[O-].[K+].[K+]>>CC(C)Sc1ncccc1F

SmilesTokenizer

Tokens: C C (C) S . C N (C) C = O . F c 1 c c c n c 1 F . O = C ([O-]) [O-] . [K+] . [K+] >> C C (C) S c 1 n c c c c 1 F

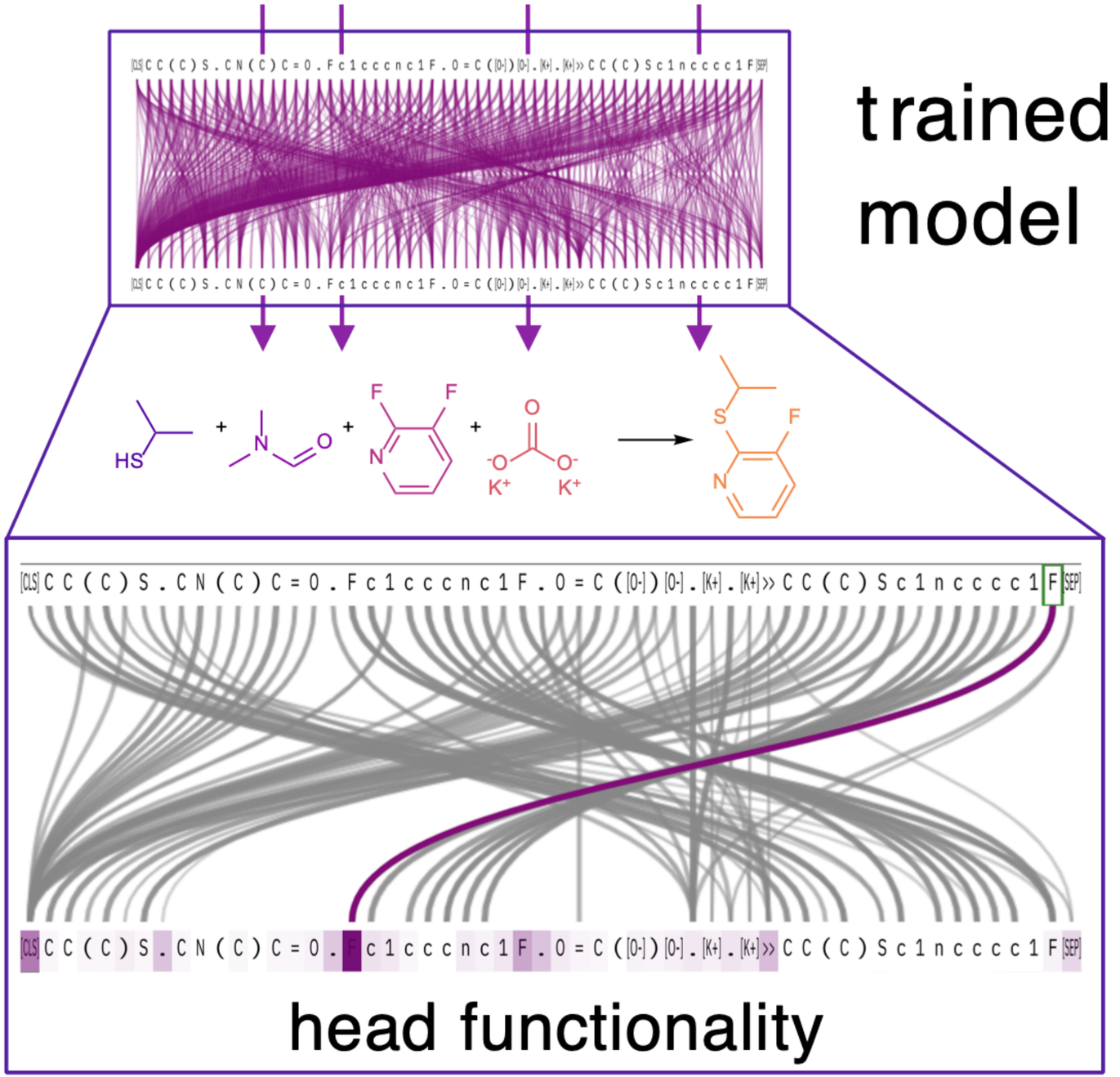
Visual Analysis
of Attention Weights

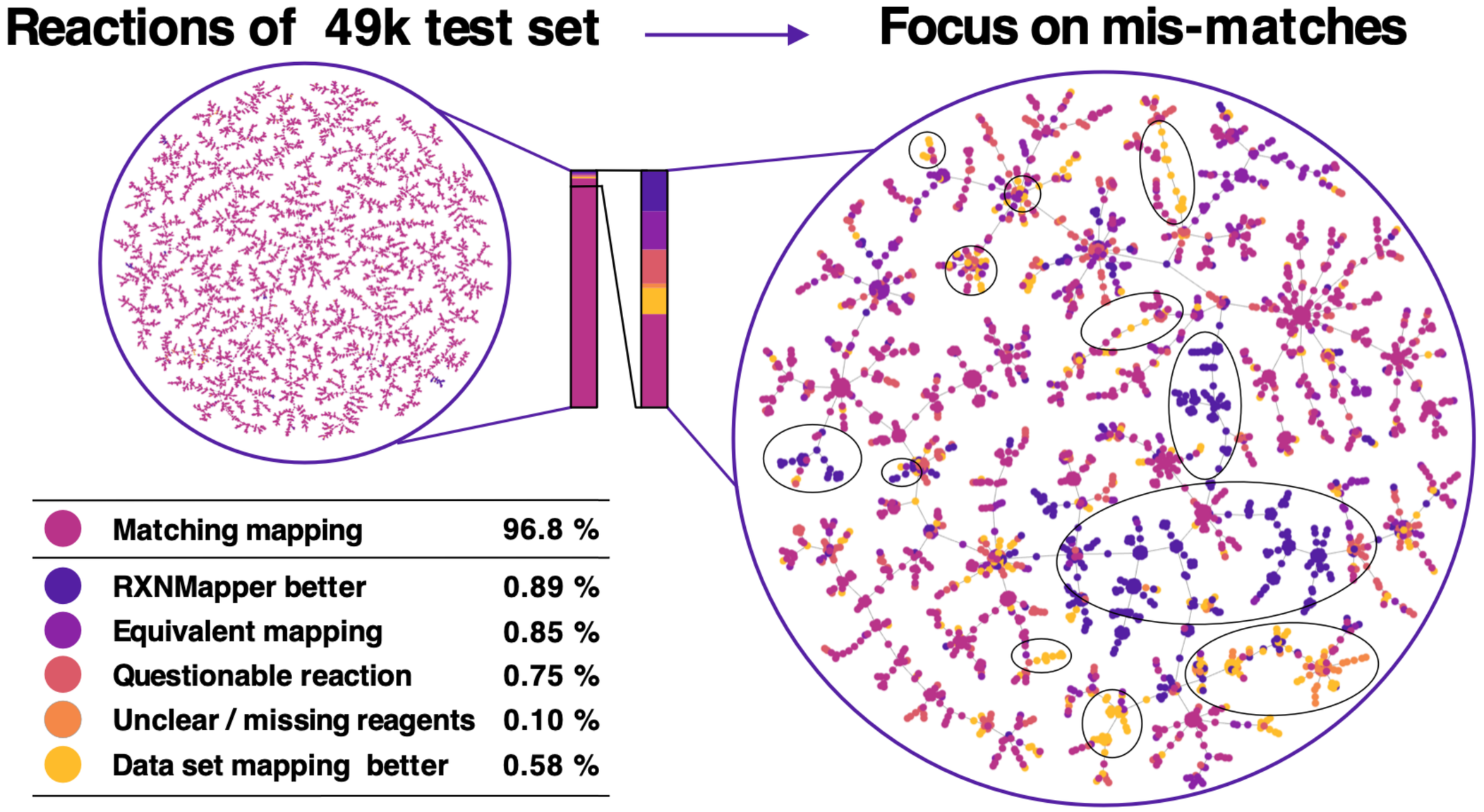


Discovery: Atom-mapping

→ **RXNMapper**

trained model





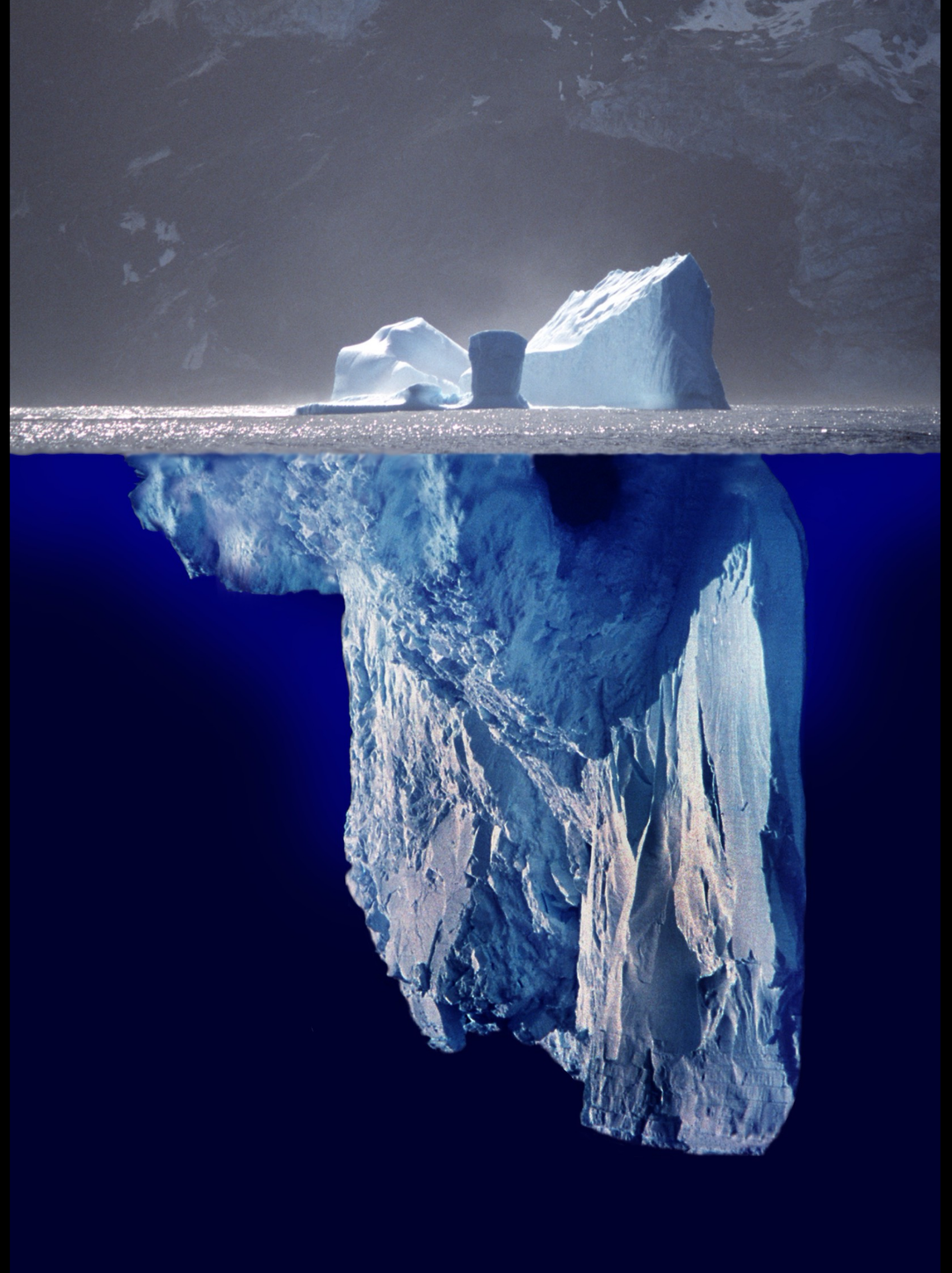
99.4% atom-mappings correct

Three stories with lessons learned

- GanPaint
  Interactive tools can enable collaboration between human and models for creation and building intuition

- GLTR
  Don’t be afraid to build a tool - maybe someone else finds a solution.

- ExBERT => RxnMapper
  DL can be a complex statistical tool to discover latent characteristics.
  Visualization can help reading these statistics.



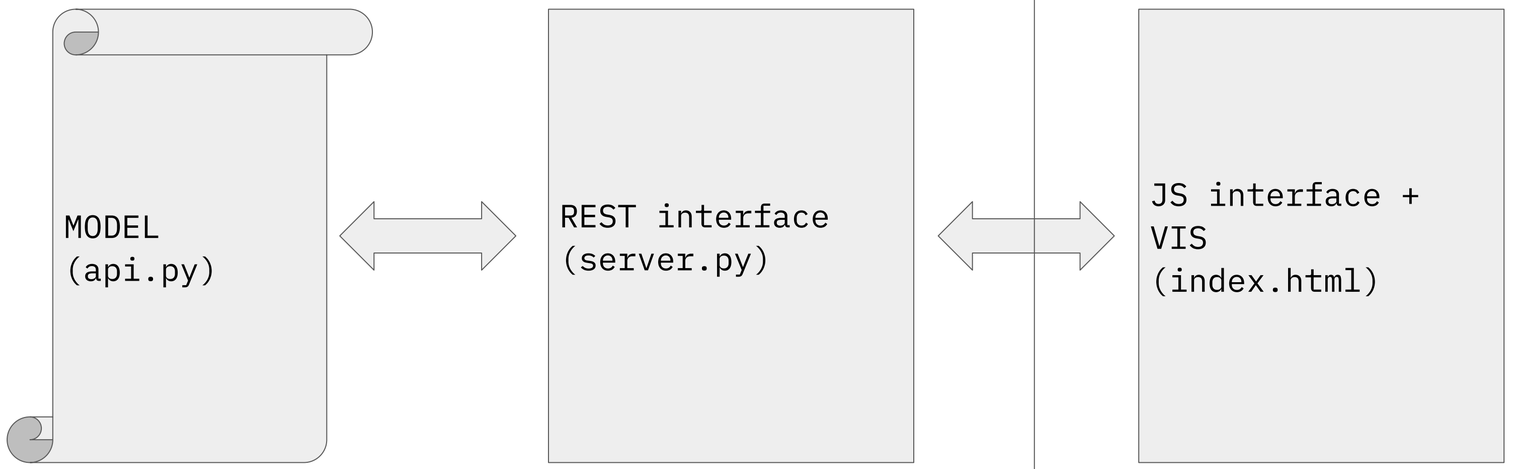
One more thing

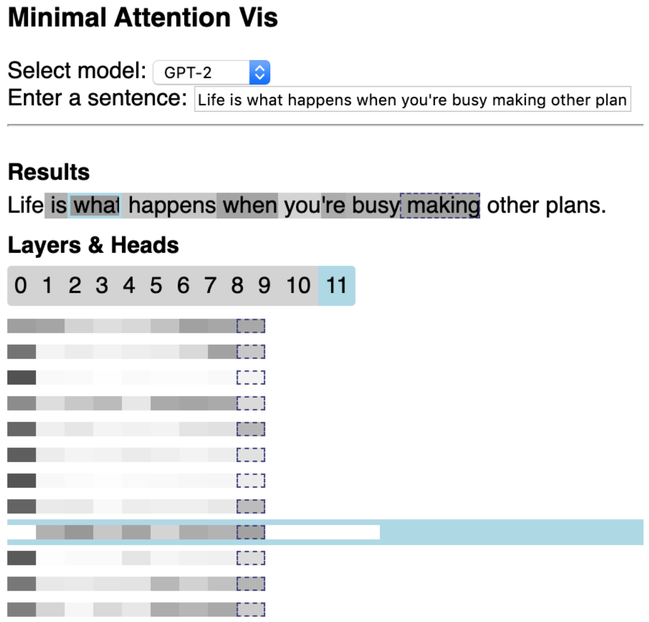
The 1-day JS Prototype

https://github.com/SIDN-IAP/attnvis
https://github.com/HendrikStrobelt/sentimenter_minimal_hai

Python

Javascript / HTML / CSS

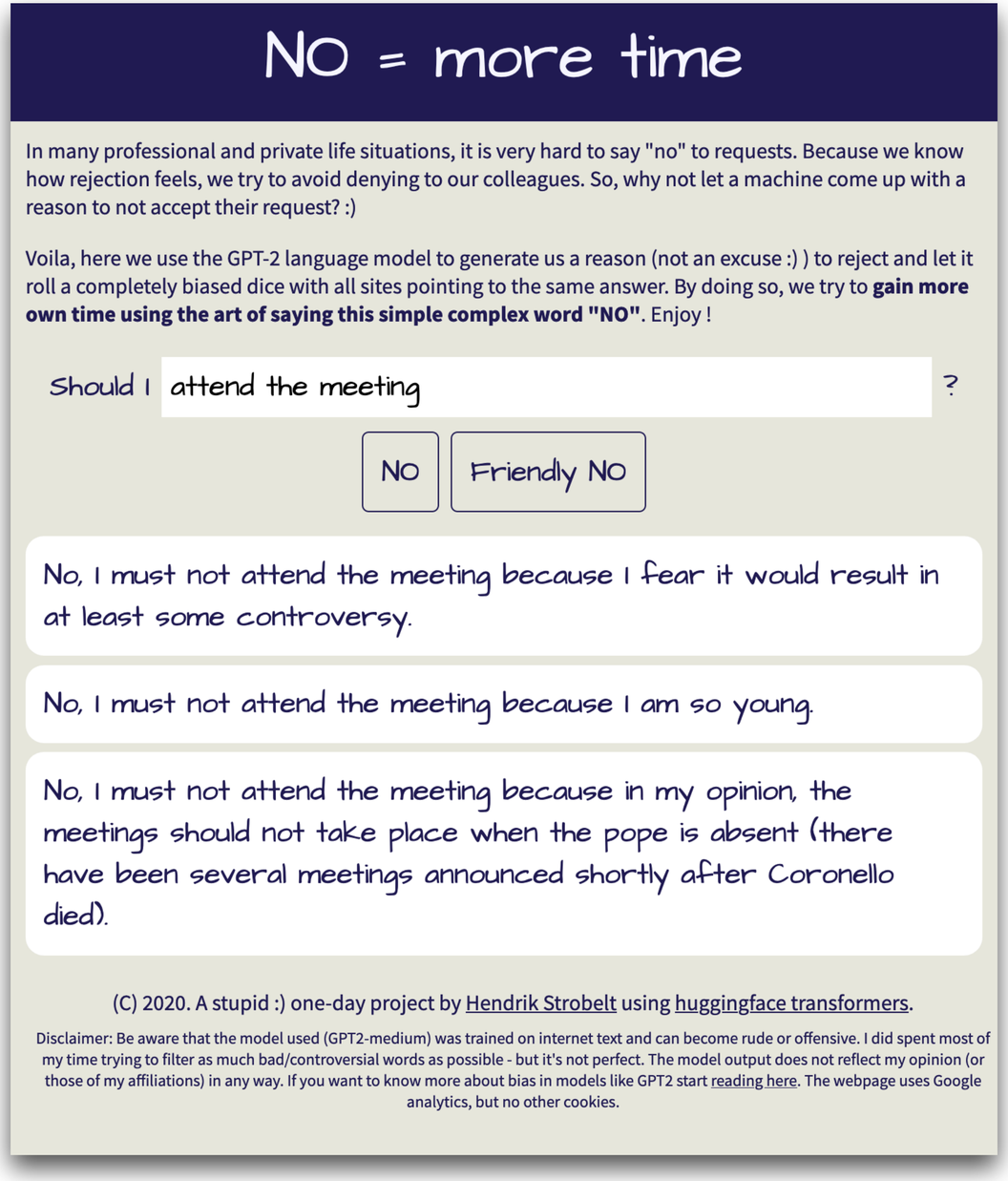




huggingface
pytorch

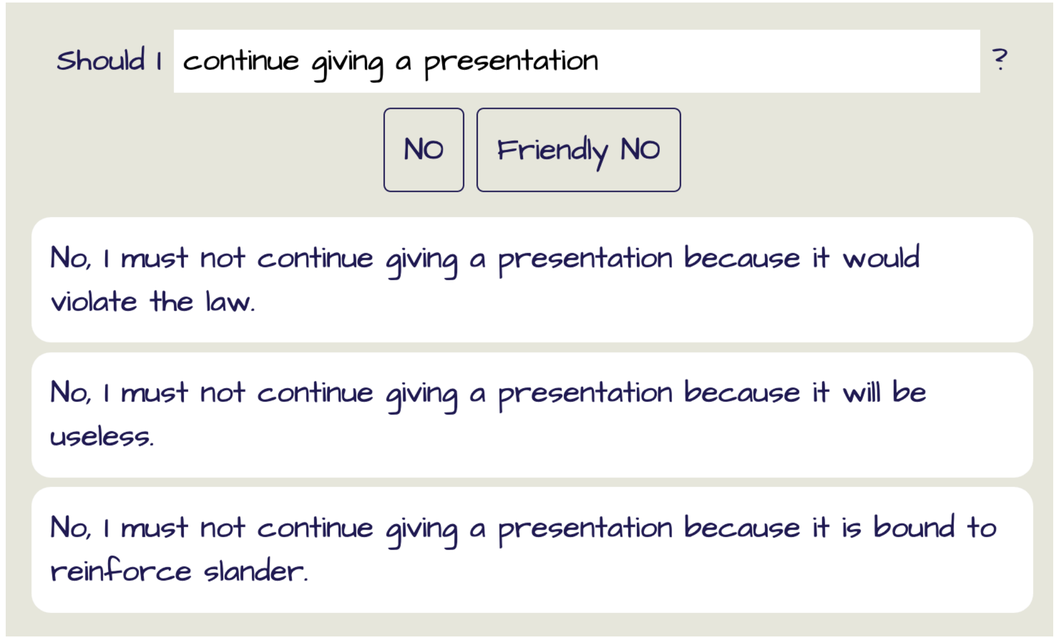
flask

html/css/js
d3.js



https://no-moretime.net/

Hendrik Strobelt.
“No = more time. The art of saying ‘No’”
NeurIPS art gallery, 2020.



Human interaction and collaboration with ML models

Hendrik Strobelt, PhD
hen_str

THANK YOU !

MIT-IBM Watson AI Lab

IBM **Research**