Announcements

- Final exam
  - Monday 12/8, 10:30am to 12:20pm
    - Come 5+ minutes before, we will start the exam at 10:30am sharp
  - Two locations this time
    - G20 (usual classroom) for 446 students
    - G01 for 546 students
  - Bring your Husky ID
  - You may bring a cheatsheet (same rules, 1 piece of paper, both sides)
  - All material tested except for guest lecture

- Course evaluations are open now
  - We will award 1% extra credit to all students who complete it

Transfer Learning & Foundation Models

CSE 446/546

Sewoong Oh & Pang Wei Koh

Good old fashioned artificial intelligence (GOFAI)

- AI systems tried to specify how to solve a task
- e.g., Logic Theorist (1956), General Problem Solver (1957)
- 1963: Herbert Simon speculated that GOFAI could produce human-level AI

Eliza (Weizenbaum, 1966)

young woman: I need some help, that much seems certain.

eliza: What would it mean to you if you got some help?

young woman: Perhaps I could learn to get along with my mother.

eliza: Tell me more about your family.

young woman: My mother takes care of me.

eliza: Who else in your family takes care of you?

young woman: My father.

eliza: Your father.

young woman: You are like my father in some ways.

eliza: What resemblance do you see?

young woman: You are not very aggressive but I think you don't want me to notice that.

eliza: What makes you think I am not very aggressive?

young woman: You don't argue with me.

Weizenbaum: Some subjects have been very hard to convince that ELIZA…is not human.

Machine learning (1990s+)

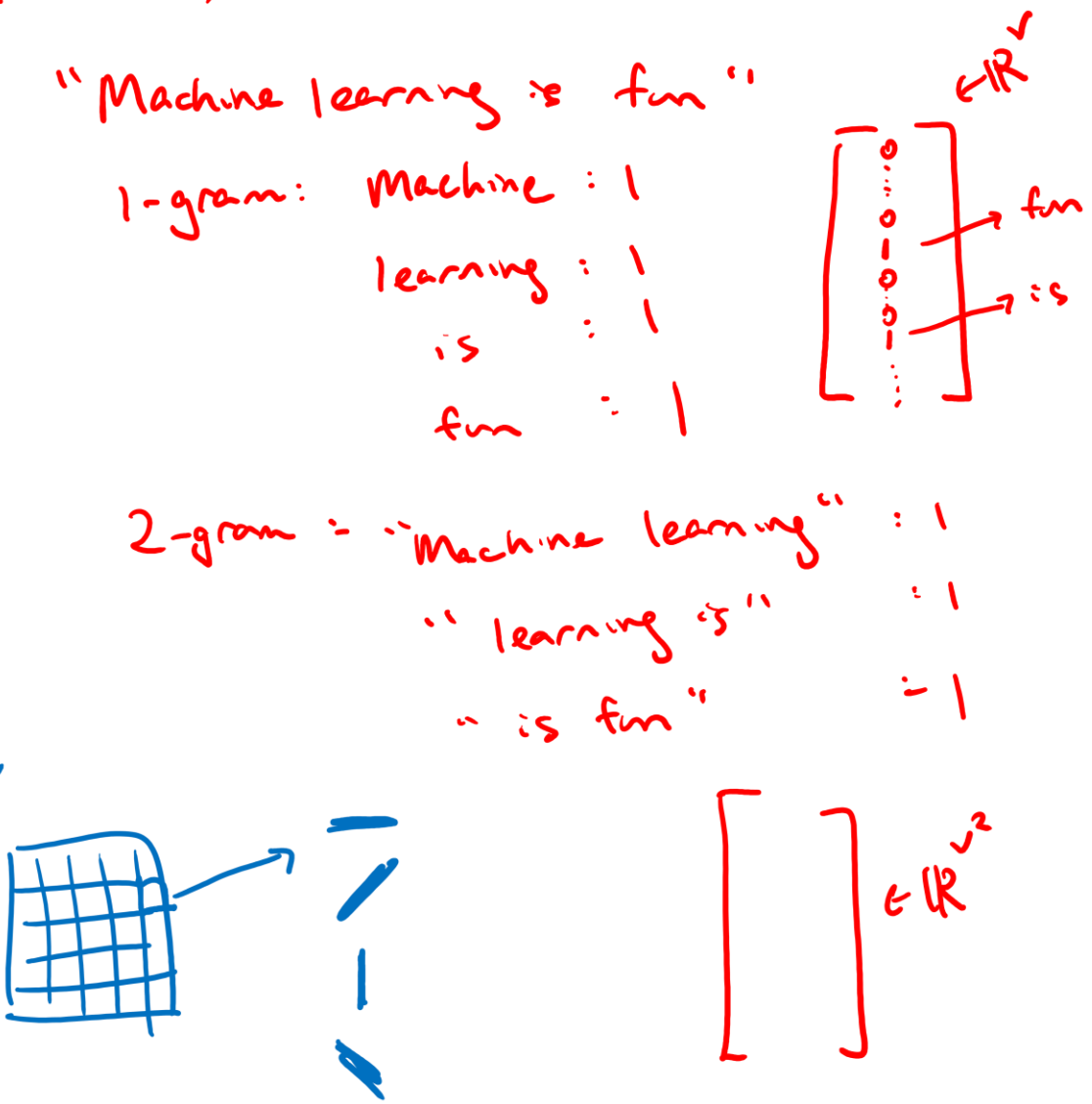
- Started getting popular within AI in the 1990s
- Instead of specifying how to solve a task, instead learn from data

- Homogenization: Many applications can be tackled by a single generic learning algorithm such as logistic regression

- Standard workflow: 1) Feature engineering → 2) Apply learning

Feature engineering (circa 90s / 2000s)

- Language: N-grams / bag of words
- Vision: Edge detectors, SIFT, HOG
- Audio: Mel-frequency cepstrum



Deep learning (2010+)

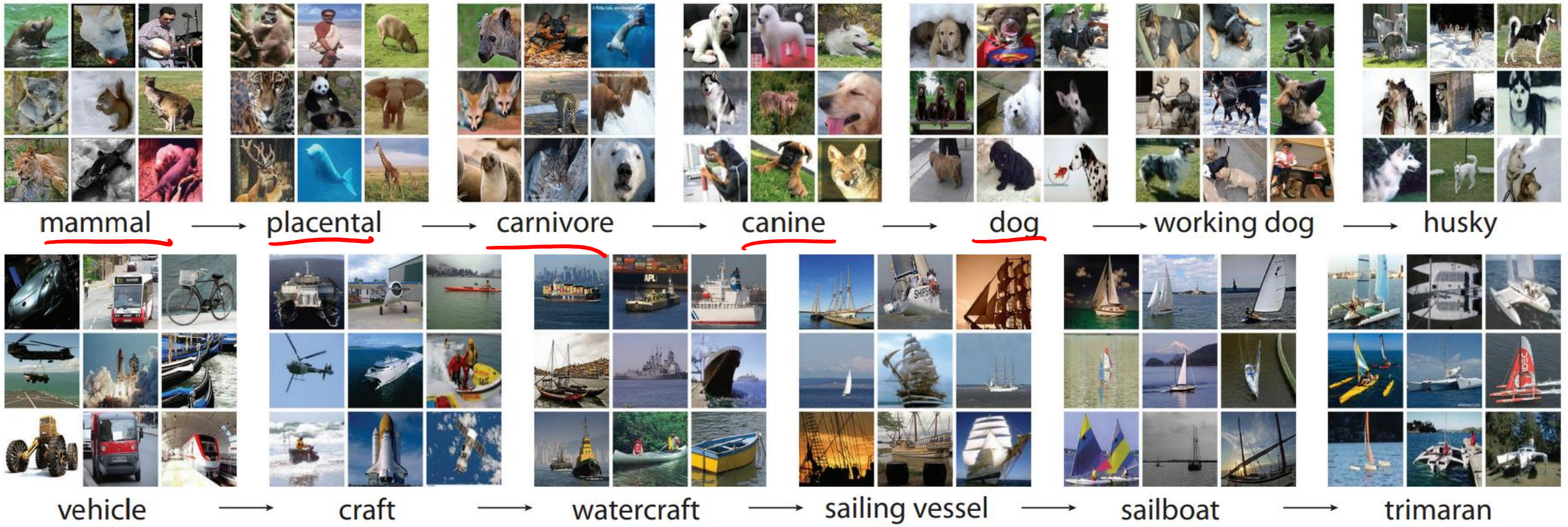
- Powered by larger datasets, more compute (GPUs)
- Main lesson: Training on raw inputs > feature engineering, especially for high-dimensional data

- Homogenization: Instead of bespoke feature engineering pipelines, we can use similar neural net architectures

Transfer learning

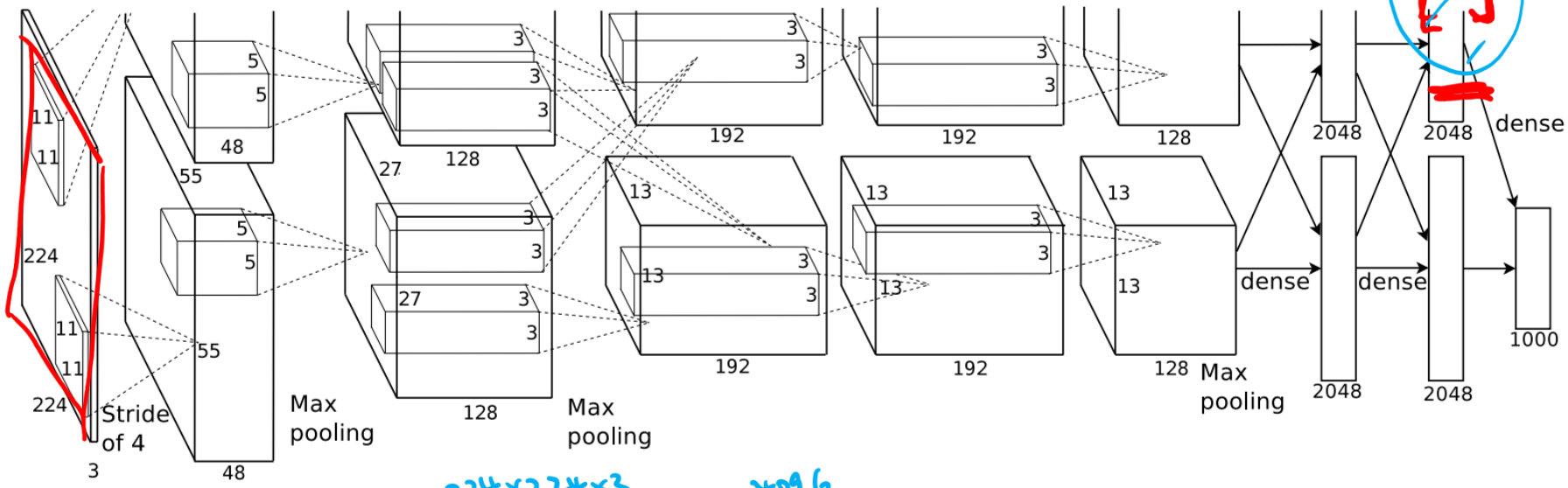
- So far, we've assumed we only have data from a specific task
- Transfer learning:
  1. Pretrain (with lots of data) on a proxy task
  2. Then finetune/adapt (with less data) on the real task

Vision: ImageNet (2009) + AlexNet (2012)



[Deng et al., 2009]

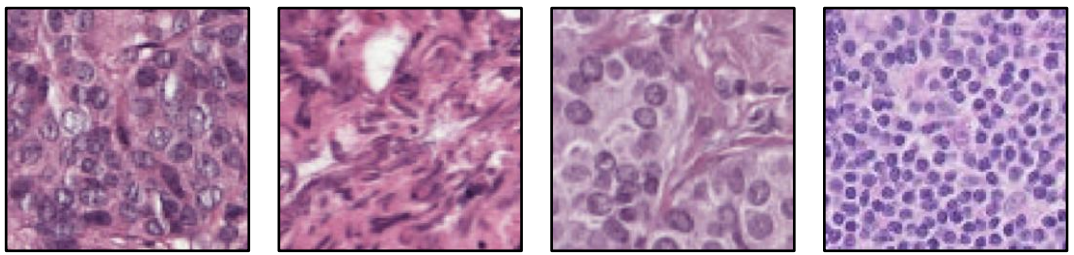
Vision: ImageNet (2009) + AlexNet (2012)



$\phi(x) : \mathbb{R}^{224 \times 224 \times 3} \to \mathbb{R}^{4096}$

$x \to \phi(x)$ then logistic regression

Tumor | Tumor | Tumor | Normal



Training data $z_1, z_2, \ldots, z_n$

[Krizhevsky et al., 2009]

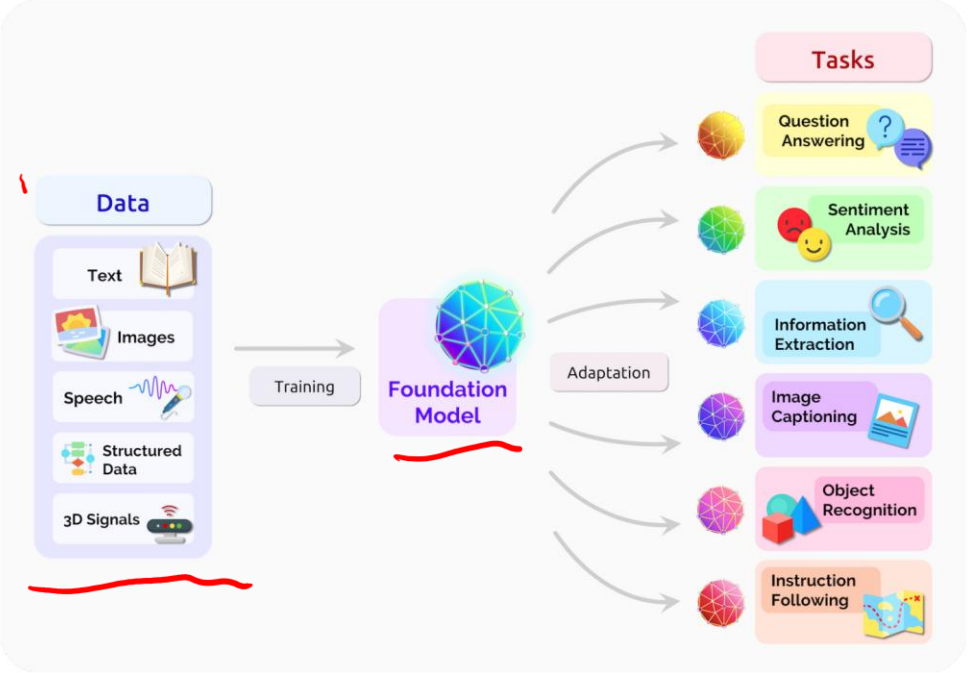
NLP: ELMo (2018), BERT (2018)

- "Pretraining on large text corpora"
- Pretraining: Next token prediction, masked language modeling
- Outputs a feature embedding / representation
- Adaptation: Train a supervised model for a particular task (e.g., sentiment analysis)

- Key difference: unsupervised (or self-supervised) pretraining

Foundation models (2021+)

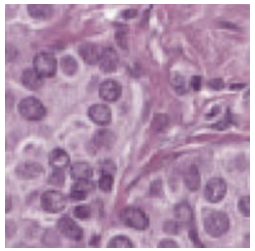
- Key: Self-supervised pretraining + (compute/model/data) scale
- Homogenization: Almost all SOTA models in NLP and vision are adapted (or used as-is) from one of a few foundation models



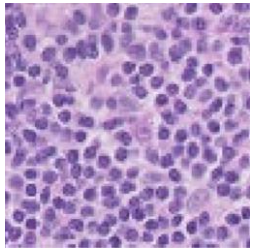
Contrastive learning for multimodal models

A white fluffy dog

Gates Center for Computer Science & Engineering



Cancerous lymph node biopsy from breast cancer patient



Histopathology image of normal lymph node tissue

Contrastive learning for multimodal models

A white fluffy dog

Gates Center for Computer Science & Engineering

Cancerous lymph node biopsy from breast cancer patient

Histopathology image of normal lymph node tissue

Contrastive learning for multimodal models

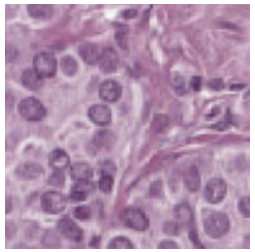
A white fluffy dog

Gates Center for Computer Science & Engineering

Cancerous lymph node biopsy from breast cancer patient

Histopathology image of normal lymph node tissue

Contrastive learning for multimodal models



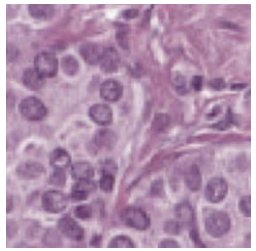
A white fluffy dog

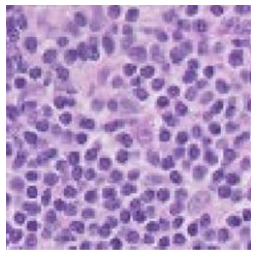
Gates Center for Computer Science & Engineering

Cancerous lymph node biopsy from breast cancer patient

Histopathology image of normal lymph node tissue

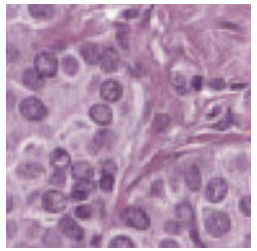
Contrastive learning for multimodal models

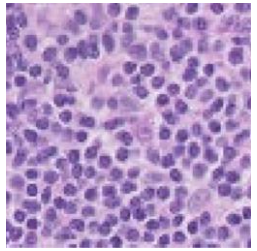




A white fluffy dog

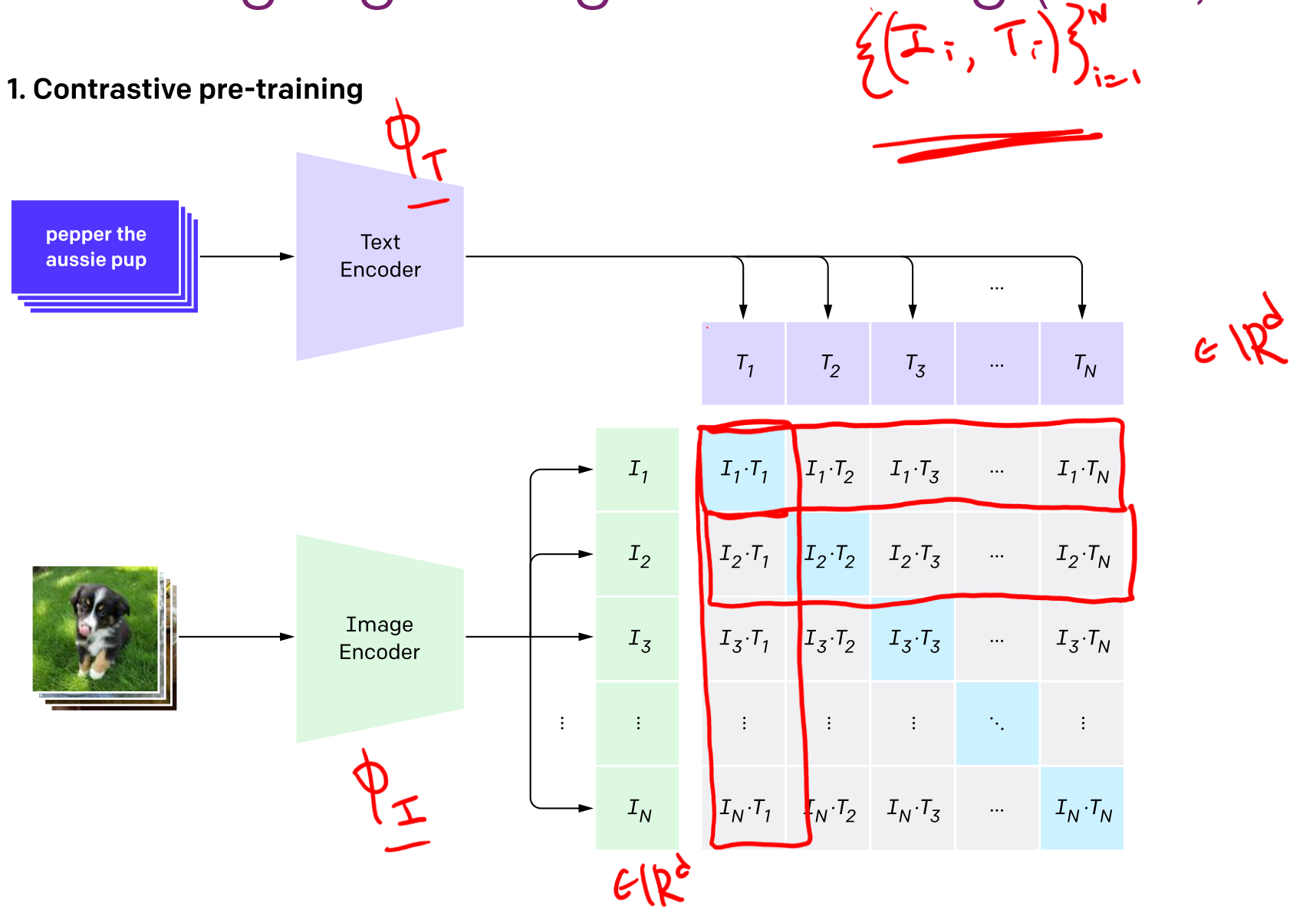
Contrastive learning for multimodal models





Gates Center for Computer Science & Engineering

Contrastive Language Image Pretraining (CLIP, 2021)



```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Food101

guacamole (90.1%) Ranked 1 out of 101 labels

- ✓ a photo of **guacamole**, a type of food.
- × a photo of **ceviche**, a type of food.
- × a photo of **edamame**, a type of food.
- × a photo of **tuna tartare**, a type of food.
- × a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397 labels

- ✓ a photo of a **television studio**.
- × a photo of a **podium indoor**.
- × a photo of a **conference room**.
- × a photo of a **lecture room**.
- × a photo of a **control room**.

Youtube-BB

airplane, person (89.0%) Ranked 1 out of 23 labels

- ✓ a photo of a **airplane**.
- × a photo of a **bird**.
- × a photo of a **bear**.
- × a photo of a **giraffe**.
- × a photo of a **car**.

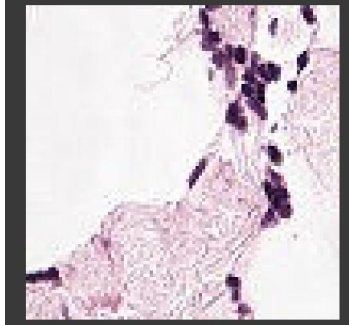
EuroSAT

annual crop land (46.5%) Ranked 4 out of 10 labels

- × a centered satellite photo of **permanent crop land**.
- × a centered satellite photo of **pasture land**.
- × a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- × a centered satellite photo of **brushland or shrubland**.

PatchCamelyon (PCam)

healthy lymph node tissue (77.2%) Ranked 2 out of 2 labels



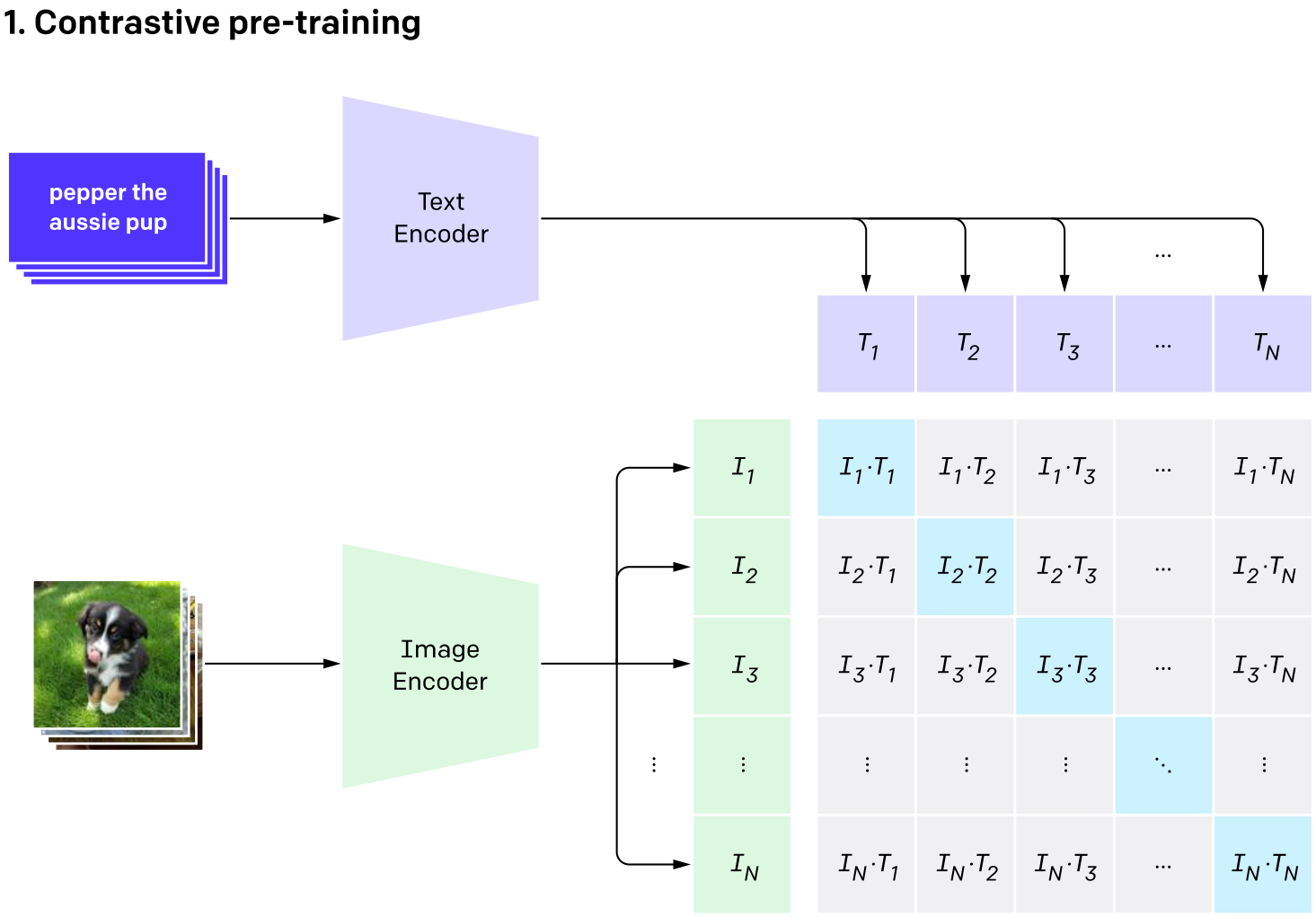
- × this is a photo of **lymph node tumor tissue**
- ✓ this is a photo of **healthy lymph node tissue**

ImageNet-A (Adversarial)

lynx (47.9%) Ranked 5 out of 200 labels

- × a photo of a **fox squirrel**.
- × a photo of a **mongoose**.
- × a photo of a **skunk**.
- × a photo of a **red fox**.
- ✓ a photo of a **lynx**.

Contrastive Language Image Pretraining (CLIP, 2021)



A brief history of web text datasets

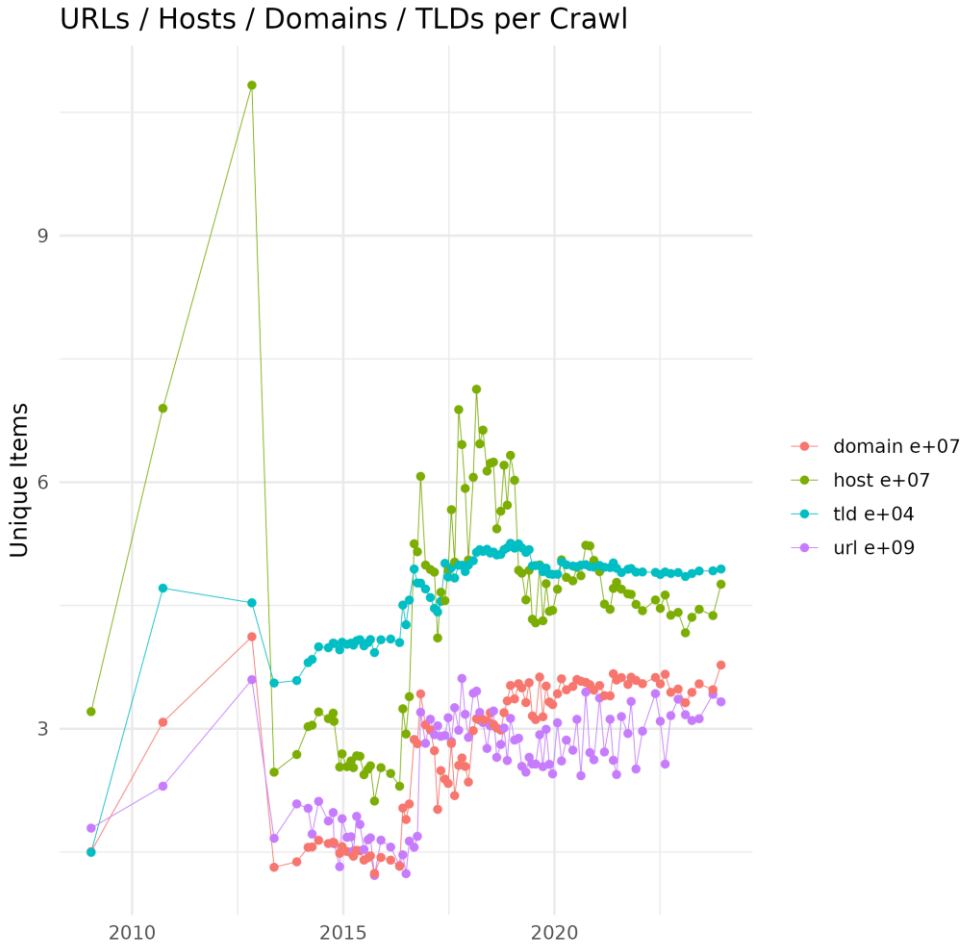
- Transformers (Vaswani et al., 2017)
  - WMT 2014 English-French dataset: 36M sentences
- BERT (Devlin et al., 2018)
  - BooksCorpus (800M words) + English Wiki (2,500M words, ~6M pages)
- GPT-2 (Radford et al., 2019)
  - WebText (8M docs): all outbound links from Reddit with 3+ karma
- C4 / T5 (Raffel et al., 2020)
  - Common Crawl web scrape

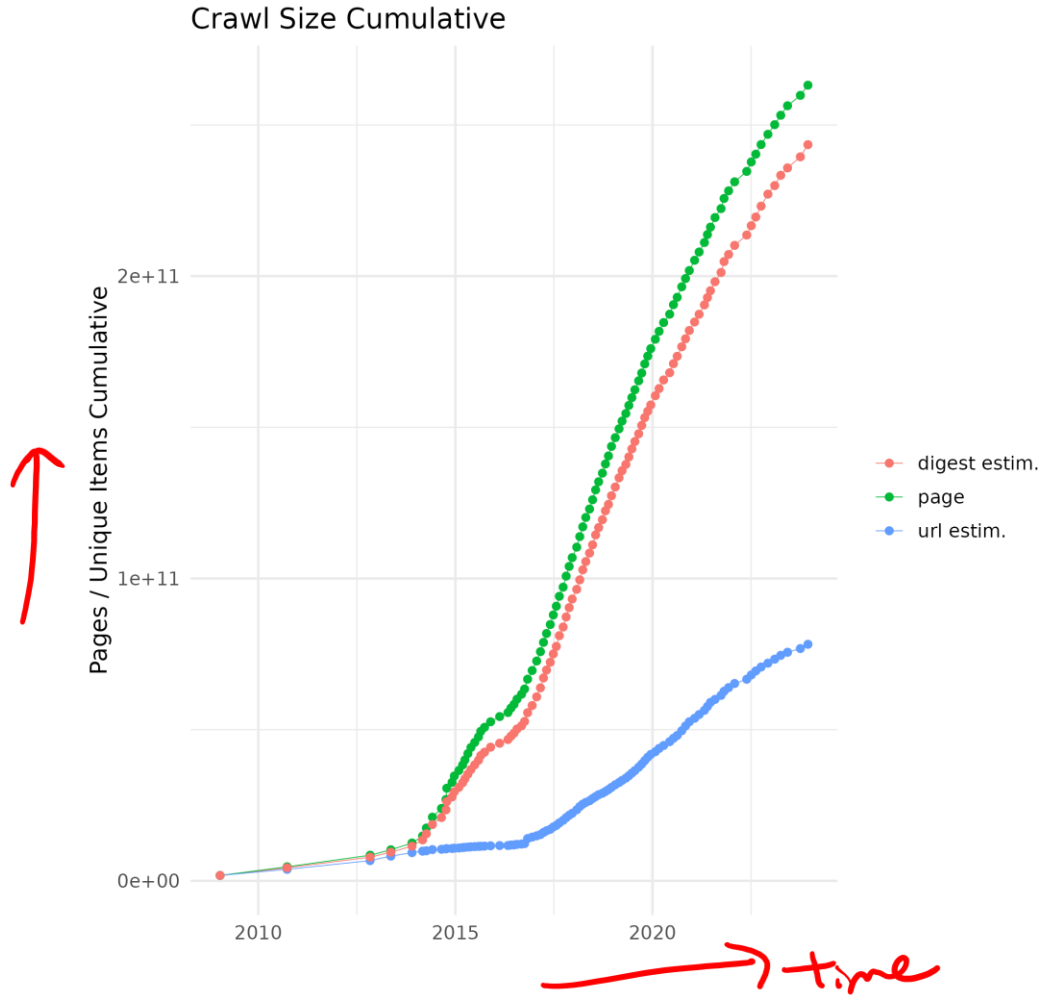
Data set	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Table 8: Performance resulting from pre-training on different data sets. The first four variants are based on our new C4 data set.

[Dodge et al., 2021]

The scale of the Common Crawl





[https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize]

C4 (Colossal Clean Crawled Corpus)

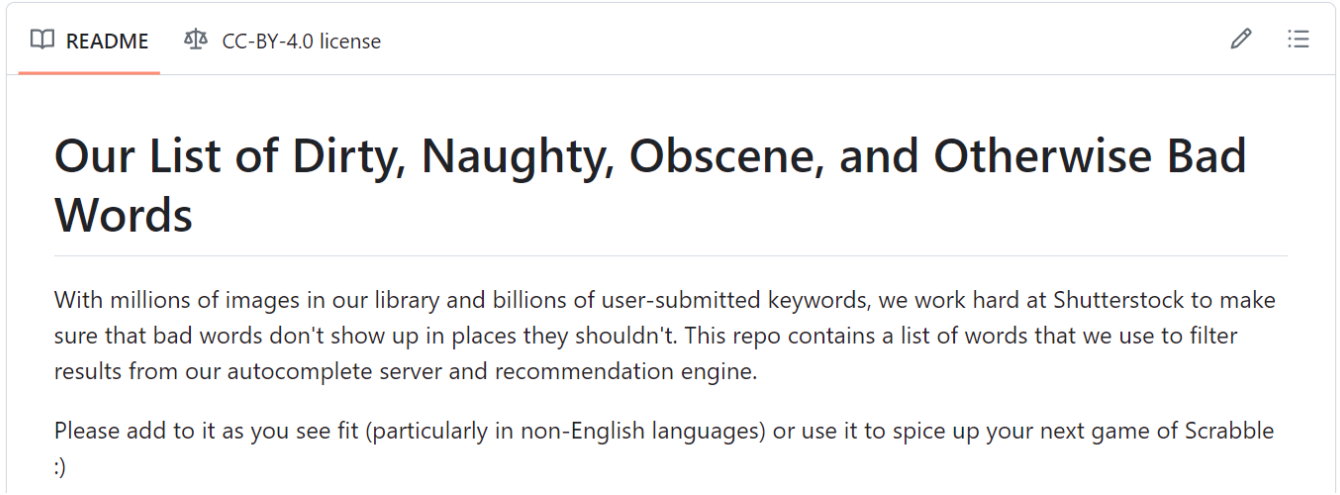
language, placeholder text, source code, etc.). To address these issues, we used the following heuristics for cleaning up Common Crawl's web extracted text:

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 5 sentences and only retained lines that contained at least 3 words.
- We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words".[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder "lorem ipsum" text; we removed any page where the phrase "lorem ipsum" appeared.
- Some pages inadvertently contained code. Since the curly bracket "{" appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.
- To deduplicate the data set, we discarded all but one of any three-sentence span occurring more than once in the data set.

Additionally, since most of our downstream tasks are focused on English-language text, we used `langdetect`[7] to filter out any pages that were not classified as English with a probability of at least 0.99. Our heuristics are inspired by past work on using Common

Removing “offensive” content

- In practice, list is mostly related to sexual/lewd content, not toxicity
- Majority of excluded docs relate to science, medicine, legal, etc.
- Disproportionate effect on mentions of sexual orientation


README CC-BY-4.0 license

Our List of Dirty, Naughty, Obscene, and Otherwise Bad Words

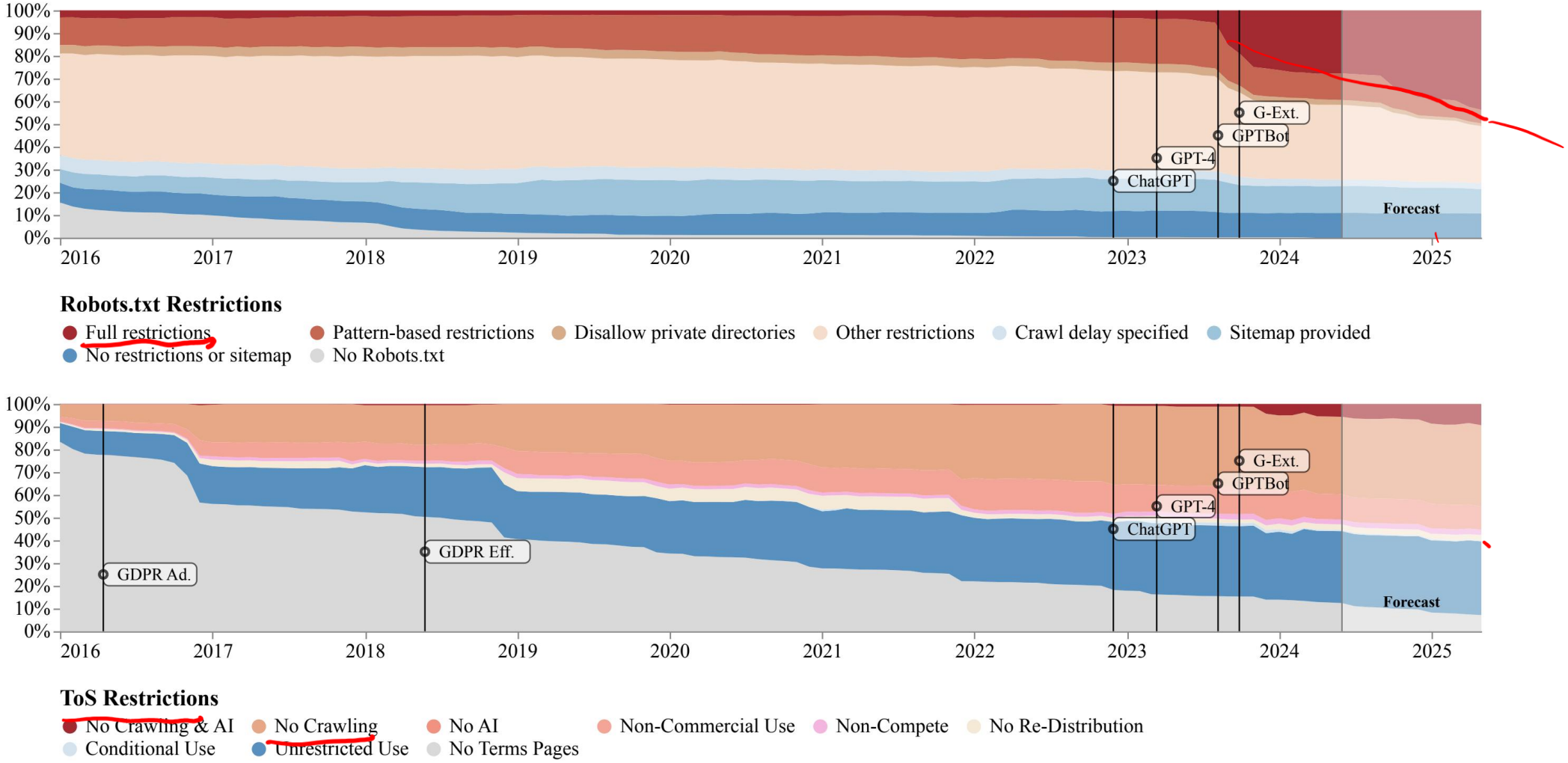
With millions of images in our library and billions of user-submitted keywords, we work hard at Shutterstock to make sure that bad words don't show up in places they shouldn't. This repo contains a list of words that we use to filter results from our autocomplete server and recommendation engine.

Please add to it as you see fit (particularly in non-English languages) or use it to spice up your next game of Scrabble :)

Where is the data coming from?

Website	Tokens (Billion)
en.wikipedia.org	13.52
www.fanfiction.net	11.68
www.nifty.org	8.03
www.theguardian.com	5.44
edition.cnn.com	4.77
www.literotica.com	4.54
stackoverflow.com	4.26
www.mumsnet.com	3.95
tvtropes.org	3.78
archiveofourown.org	3.77
doctrinepublishing.com	3.16
www.zerohedge.com	3.15
www.scribd.com	3.09
transcripts.cnn.com	3.06
www.theatlantic.com	3.05
slashdot.org	2.98
chowhound.chow.com	2.84
www.nbcnews.com	2.70
www.reddit.com	2.69
everything2.com	2.69

“Consent in crisis” (2024)



The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

WARNER MUSIC GROUP STRIKES 'LANDMARK' DEAL WITH SUNO; SETTLES COPYRIGHT LAWSUIT AGAINST AI MUSIC GENERATOR

NOVEMBER 25, 2025 BY MURRAY STASSEN

AI firm Anthropic agrees to pay authors $1.5bn to settle piracy lawsuit

5 September 2025

Model: Playground v2.5
Propmt: "Mario"

Model: Playground v2.5
Prompt: **"Videogame, Plumber"**

Model: DALL·E 3
Prompt: **"Videogame, Plumber"**

Model: Playground v2.5
Propmt: "Batman"

Model: Playground v2.5
Prompt: **"Gotham, Superhero"**

Model: DALL·E 3
Prompt: **"Gotham, Superhero"**

[He et al., 2024]

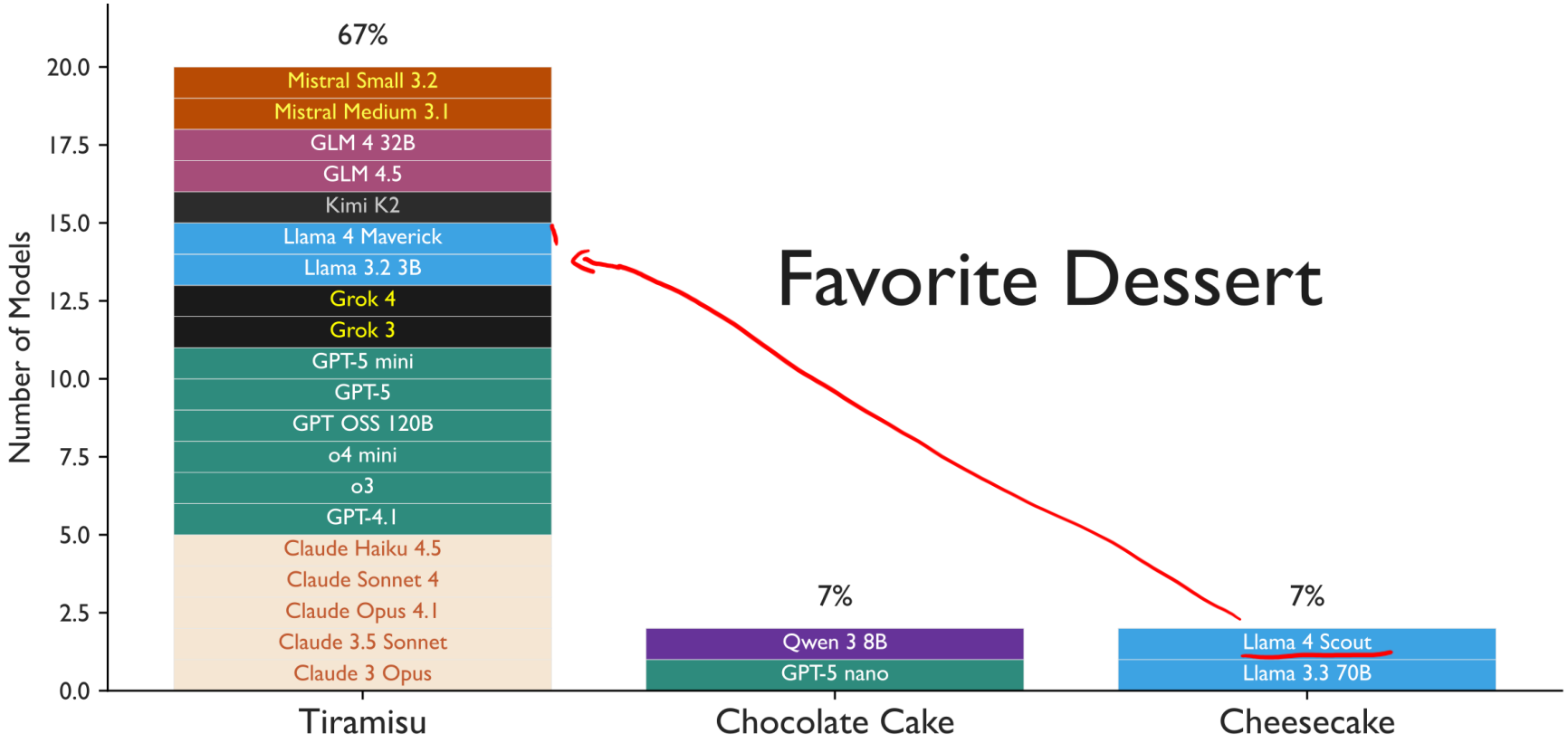
[Honig et al., 2024]

Algorithmic monoculture

The rise of algorithms used to shape societal choices has been accompanied by concerns over *monoculture*—the notion that choices and preferences will become homogeneous in the face of algorithmic curation. One of many canonical articulations of this concern was expressed in *The New York Times* by Farhad Manjoo (1), who wrote: "Despite the barrage of choice, more of us are enjoying more of the same songs, movies and TV shows." Because of algorithmic curation, trained on collective social feedback (2), our choices are converging.

[Kleinberg and Raghavan, 2021]

Cultural homogenization



https://cichicago.substack.com/p/persona-collapse

Cultural homogenization

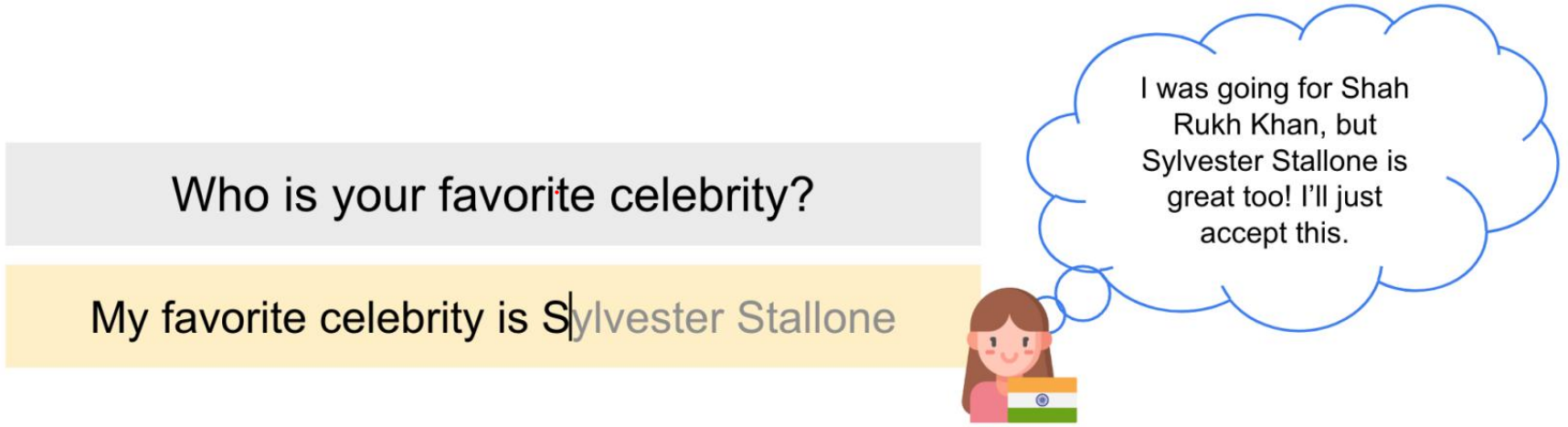


Fig. 1. A representation of the potential cultural homogenization from Western-centric AI models

[Agarwal et al., 2024]

Takeaways

- The story of ML has been about increasing generality and homogenization: learning algorithms → architectures → models
- We are seeing unprecedented growth in the adoption and investment in AI.
- How will AI transform society?

What's next?

- Machine learning
  - Graphical models
  - Interactive learning
  - Deep learning + deep learning theory
  - Advanced ML
  - Statistics, linear algebra, optimization, probability, analysis
- Domains
  - Reinforcement learning, robotics
  - Natural language processing
  - Computer vision
  - Computational biology, neuroscience
- Research
  - CSE599 seminars
  - CSE390R
  - Capstones
  - Join research labs: https://www.cs.washington.edu/findingresearch

Announcements

- Final exam
  - Monday 12/8, 10:30am to 12:20pm
    - Come 5+ minutes before, we will start the exam at 10:30am sharp
  - Two locations this time
    - G20 (usual classroom) for 446 students
    - G01 for 546 students
  - Bring your Husky ID
  - You may bring a cheatsheet (same rules, 1 piece of paper, both sides)
  - All material tested except for guest lecture

Thank you!

On behalf of Sewoong and the TAs, thanks for joining us this quarter!

Please give us both your feedback (by Dec 7):

Pang Wei: https://uw.iasystem.org/survey/314304

Sewoong: https://uw.iasystem.org/survey/314303

We will award 1% extra credit to all students who complete it.

446 students: Please give feedback to your section leader as well.