Multimodal / Image Search

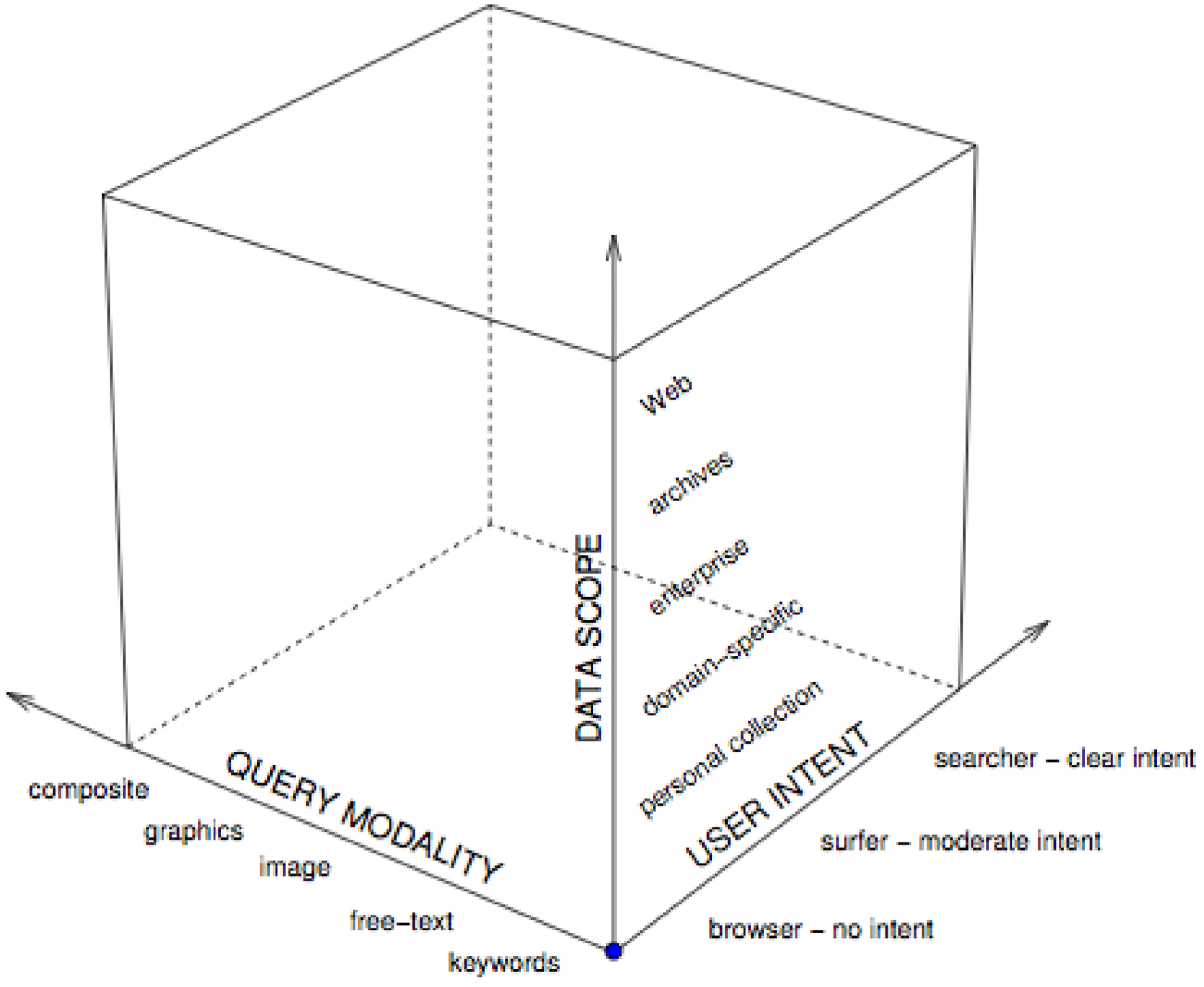
SI 650 / EECS 549
Nov. 5 2025

Slides adapted from David Jurgens

Today's goals

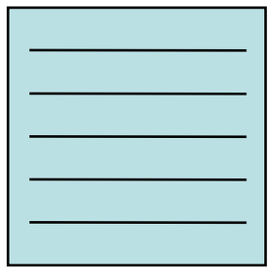
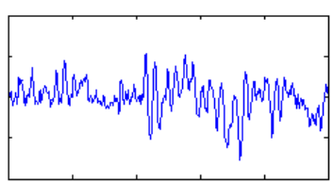
- Introduce different problems in image retrieval
- Discuss how to represent images as queries and documents
- Present a range of IR techniques (old and new) for workin with images

Different Systems



http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta_TR.pdf

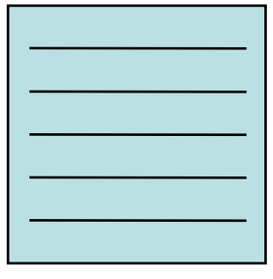
Information retrieval: data

	amount of data	data characteristics
Text retrieval 	trillions of web pages within an order of magnitude in “private” data	• user generated • some semi-structured • link structure
Audio retrieval 	order of a few billion? E.g., Spotify has >80M songs	• mostly professionally generated • co-occurrence statistics
Image retrieval 	somewhere in between text and audio	• user generated • some tagging • incorporated into web pages (context)
Code retrieval import os.path os.path.isfile(fname)	billions of code files + millions of Q&A answers	• user generated • sort of like text • can have comments

Information retrieval: challenges

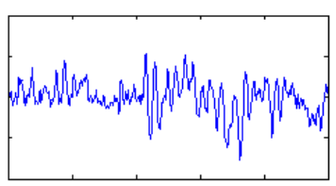
challenges

Text retrieval



- scale
- ambiguity of language
- link structure
- spam

Audio retrieval



- query language
- user interface
- features/pre-processing

Image retrieval

- query language
- user interface
- features/pre-processing
- ambiguity of pictures

Code retrieval

```
import os.path
os.path.isfile(fname)
```

- mixed query language
- content interpretation
- features/pre-processing

Document-as-Image Search

What's in a document?

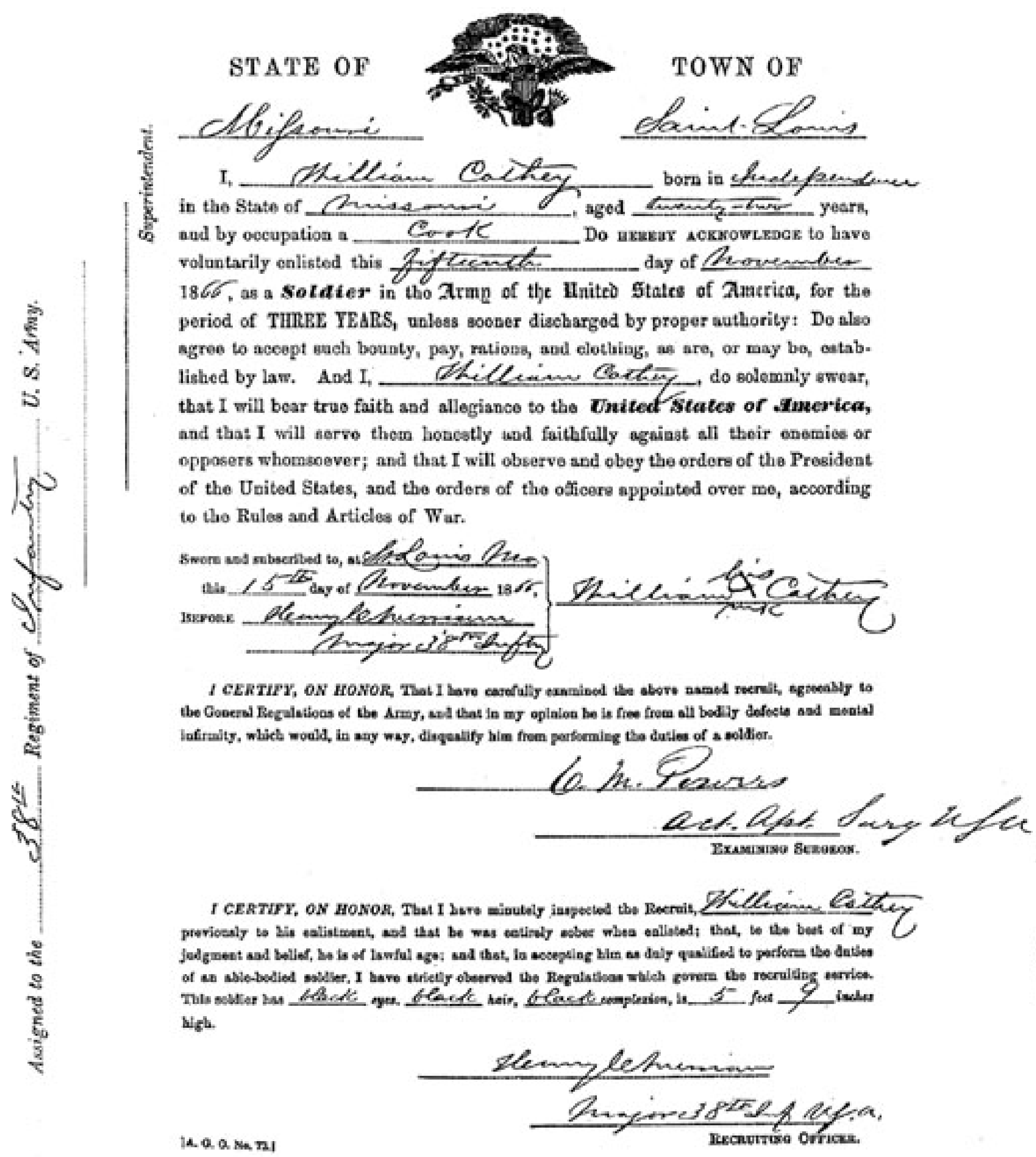
- I give you a file I downloaded
- You know it has text in it
- What are the challenges in determining what characters are in the document?
- File format:

File types indexable by Google

Google can index the content of most types of pages and files. The most common file types we index include:

- Adobe Portable Document Format (.pdf)
- Adobe PostScript (.ps)
- Autodesk Design Web Format (.dwf)
- Google Earth (.kml, .kmz)
- GPS eXchange Format (.gpx)
- Hancom Hanword (.hwp)
- HTML (.htm, .html, other file extensions)
- Microsoft Excel (.xls, .xlsx)
- Microsoft PowerPoint (.ppt, .pptx)
- Microsoft Word (.doc, .docx)
- OpenOffice presentation (.odp)
- OpenOffice spreadsheet (.ods)
- OpenOffice text (.odt)
- Rich Text Format (.rtf)
- Scalable Vector Graphics (.svg)
- TeX/LaTeX (.tex)
- Text (.txt, .text, other file extensions), including source code in common programming languages:
  - Basic source code (.bas)
  - C/C++ source code (.c, .cc, .cpp, .cxx, .h, .hpp)
  - C# source code (.cs)
  - Java source code (.java)
  - Perl source code (.pl)
  - Python source code (.py)
- Wireless Markup Language (.wml, .wap)
- XML (.xml)

What is a document?



STATE OF Missouri TOWN OF Saint Louis

I, William Cathey, born in Independence in the State of Missouri, aged twenty-two years, and by occupation a Cook, Do HEREBY ACKNOWLEDGE to have voluntarily enlisted this fifteenth day of November 1866, as a **Soldier** in the Army of the United States of America, for the period of THREE YEARS, unless sooner discharged by proper authority: Do also agree to accept such bounty, pay, rations, and clothing, as are, or may be, established by law. And I, William Cathey, do solemnly swear, that I will bear true faith and allegiance to the *United States of America*, and that I will serve them honestly and faithfully against all their enemies or opposers whomsoever; and that I will observe and obey the orders of the President of the United States, and the orders of the officers appointed over me, according to the Rules and Articles of War.

Sworn and subscribed to, at St Louis Mo this 15th day of November 1866, William Cathey

BEFORE Major 38th Infty

I CERTIFY, ON HONOR, That I have carefully examined the above named recruit, agreeably to the General Regulations of the Army, and that in my opinion he is free from all bodily defects and mental infirmity, which would, in any way, disqualify him from performing the duties of a soldier.

Act. Asst. Surg. USA

EXAMINING SURGEON.

I CERTIFY, ON HONOR, That I have minutely inspected the Recruit, William Cathey previously to his enlistment, and that he was entirely sober when enlisted; that, to the best of my judgment and belief, he is of lawful age; and that, in accepting him as duly qualified to perform the duties of an able-bodied soldier, I have strictly observed the Regulations which govern the recruiting service. This soldier has black eyes, black hair, black complexion, is 5 feet 9 inches high.

Major 38th Inf. USA

RECRUITING OFFICER.

[A. G. O. No. 73.]

Assigned to the 38th Regiment of Infantry U.S. Army.

Superintendent.

2833

The Washington Post

DISTRICT FINAL

BEING A BLACK MAN

After Zarqawi, No Clear Path In Weary Iraq

Difficult Questions Surround Legacy of Insurgent Leader

The Young Apprentice

Marcus's Parents Agonize Over How to Protect — and Prepare — Him

How U.S. Forces Found Iraq's Most-Wanted Man

Kaine Delays Execution Of Inmate for 6 Months

Inquiry Into Killer's Mental State Ordered

INSIDE

FDA Approves Vaccine That Should Prevent Most Cervical Cancers

Document Images

- A document image is a document that is represented as an image, rather than some predefined format
- Like normal images, contain pixels
  - often binary-valued (black, white)
  - But greyscale or color sometimes
- 300 dots per inch (dpi) gives the best results
  - But images are quite large (1 MB per page)
  - Faxes are normally 72 dpi
- Usually stored in TIFF or PDF format

Want to be able to process them like text files

Issue: Our images are actually documents in disguise

SEARCHING

INDEXING

Doc Rep

Query Rep

query

User

Ranking

results

INTERFACE

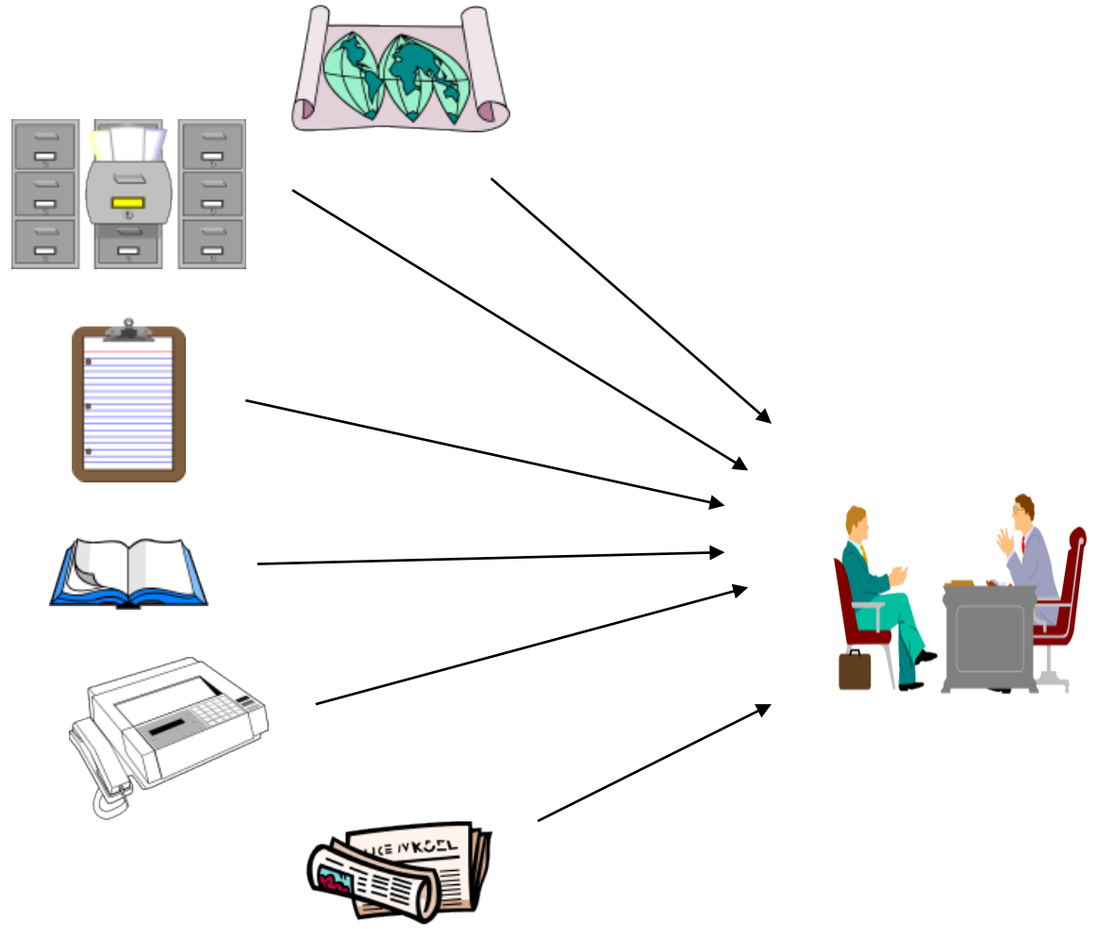
Feedback

judgments

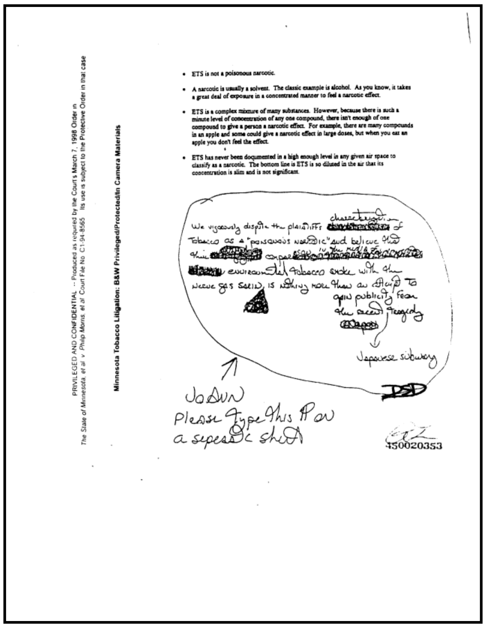
QUERY MODIFICATION

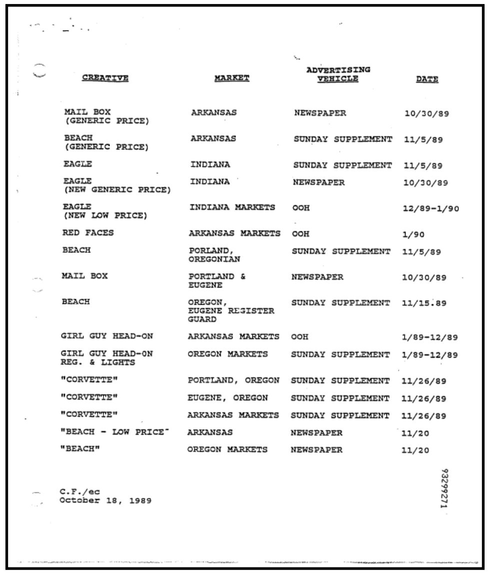
Where do all these document image come from?

- Web
  - Patents
  - Some Arabic news stories are GIF images
  - Code screen shots
  - Google Books, Project Gutenberg (though these are a bit different)
- Library archives
- Other
  - Tobacco Litigation Documents
    - 49 million page images



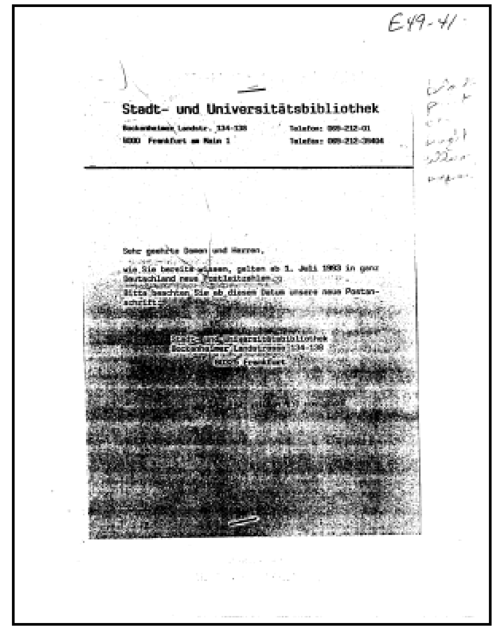
Some of them come from U-M!





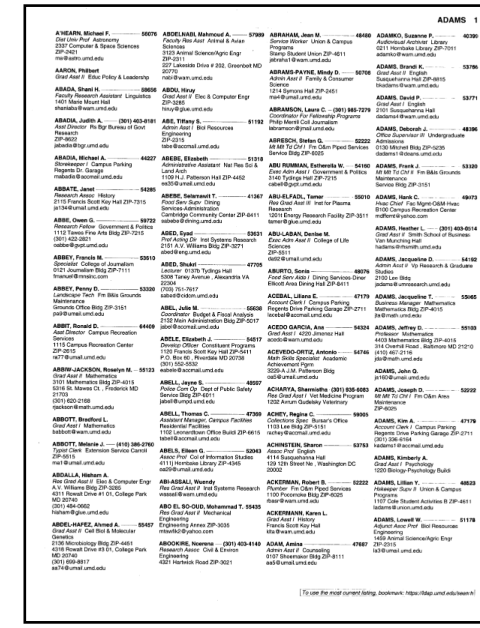
CREATIVE	MARKET	ADVERTISING VEHICLE	DATE
MAIL BOX (GENERIC PRICE)	ARKANSAS	NEWSPAPER	10/30/89
BEACH (GENERIC PRICE)	ARKANSAS	SUNDAY SUPPLEMENT	11/5/89
EAGLE	INDIANA	SUNDAY SUPPLEMENT	11/5/89
EAGLE (NEW GENERIC PRICE)	INDIANA	NEWSPAPER	10/30/89
EAGLE (NEW LOW PRICE)	INDIANA MARKETS	OOH	12/89-1/90
RED FACES	ARKANSAS MARKETS	OOH	1/90
BEACH	PORTLAND, OREGONIAN	SUNDAY SUPPLEMENT	11/5/89
MAIL BOX	PORTLAND & EUGENE	NEWSPAPER	10/30/89
BEACH	OREGON, EUGENE REGISTER GUARD	SUNDAY SUPPLEMENT	11/15.89
GIRL GUY HEAD-ON	ARKANSAS MARKETS	OOH	1/89-12/89
GIRL GUY HEAD-ON REG. & LIGHTS	OREGON MARKETS	OOH	1/89-12/89
"CORVETTE"	PORTLAND, OREGON	SUNDAY SUPPLEMENT	11/26/89
"CORVETTE"	EUGENE, OREGON	SUNDAY SUPPLEMENT	11/26/89
"CORVETTE"	ARKANSAS MARKETS	SUNDAY SUPPLEMENT	11/26/89
"BEACH - LOW PRICE"	ARKANSAS	NEWSPAPER	11/20
"BEACH"	OREGON MARKETS	NEWSPAPER	11/20

C.F./ec
October 18, 1989



Stadt- und Universitätsbibliothek

A a



The Washington Post

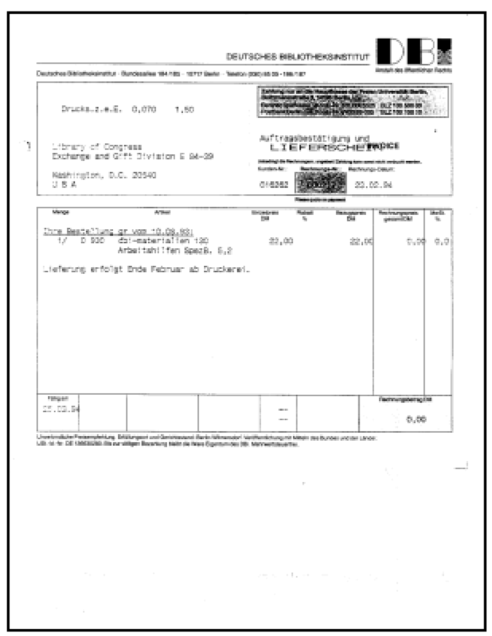
After Zarqawi, No Clear Path In Weary Iraq

The Young Apprentice

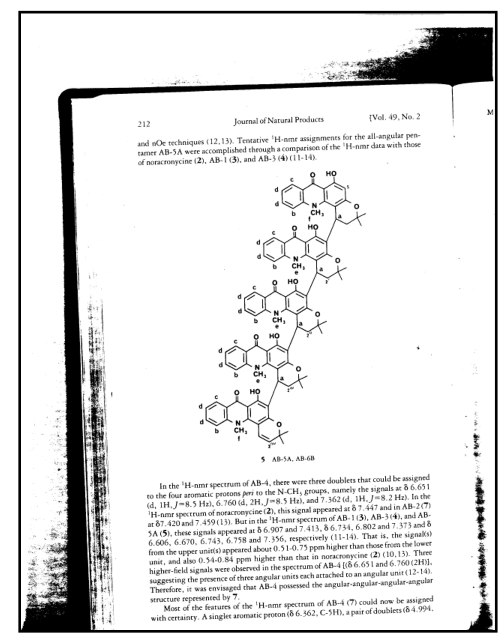
How U.S. Forces Found Iraq's Most-Wanted Man

Kaine Delays Execution Of Inmate for 6 Months

FDA Approves Vaccine That Should Prevent Most Cervical Cancers



DEUTSCHES BIBLIOTHEKSINSTITUT



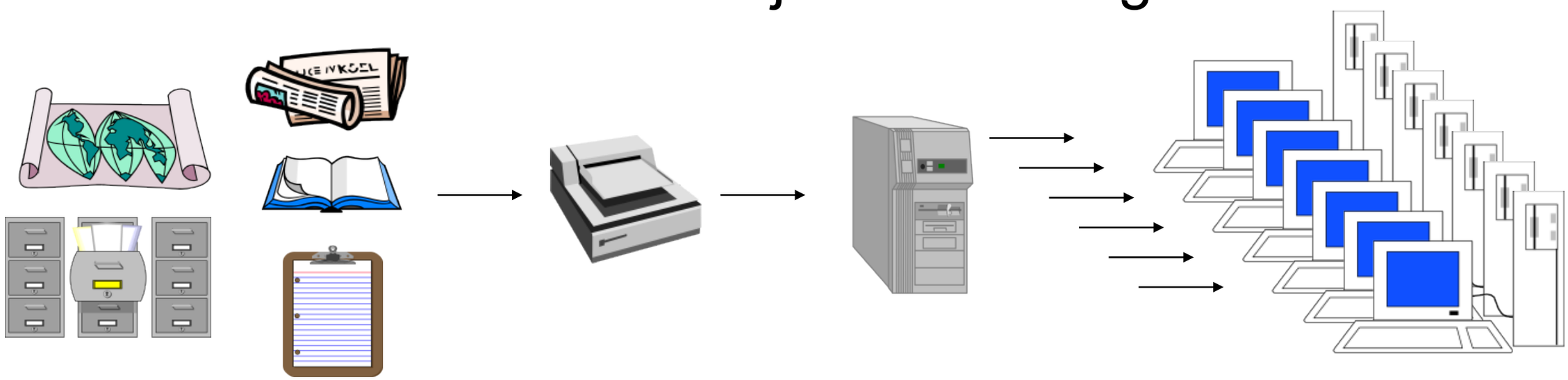
FINANCIAL TIMES

Iraq braced as Saddam rejects Bush ultimatum

Fed holds rates but signals uncertain future

Document Image Database

- Collection of scanned images
- Need to be available for indexing and retrieval, abstracting, routing, editing, dissemination, interpretation
- NOTE: more needs than just searching!



Text Search on Document-Images

What are the challenges?

What are the sub-problems?

Document images

- So far, we've only been interested in documents as strings of text
- Document images potentially contain additional information
  - embedded images
  - formatting
  - handwritten annotations
  - figures/diagrams/tables
- Classes of documents
  - memo
  - newspaper article
  - book page
  - …

Challenges

- The document is now an image
- Quality
  - scan orientation
  - noise
  - contrast
- Hand-written text
- Hand-written diagrams

Sub-problems

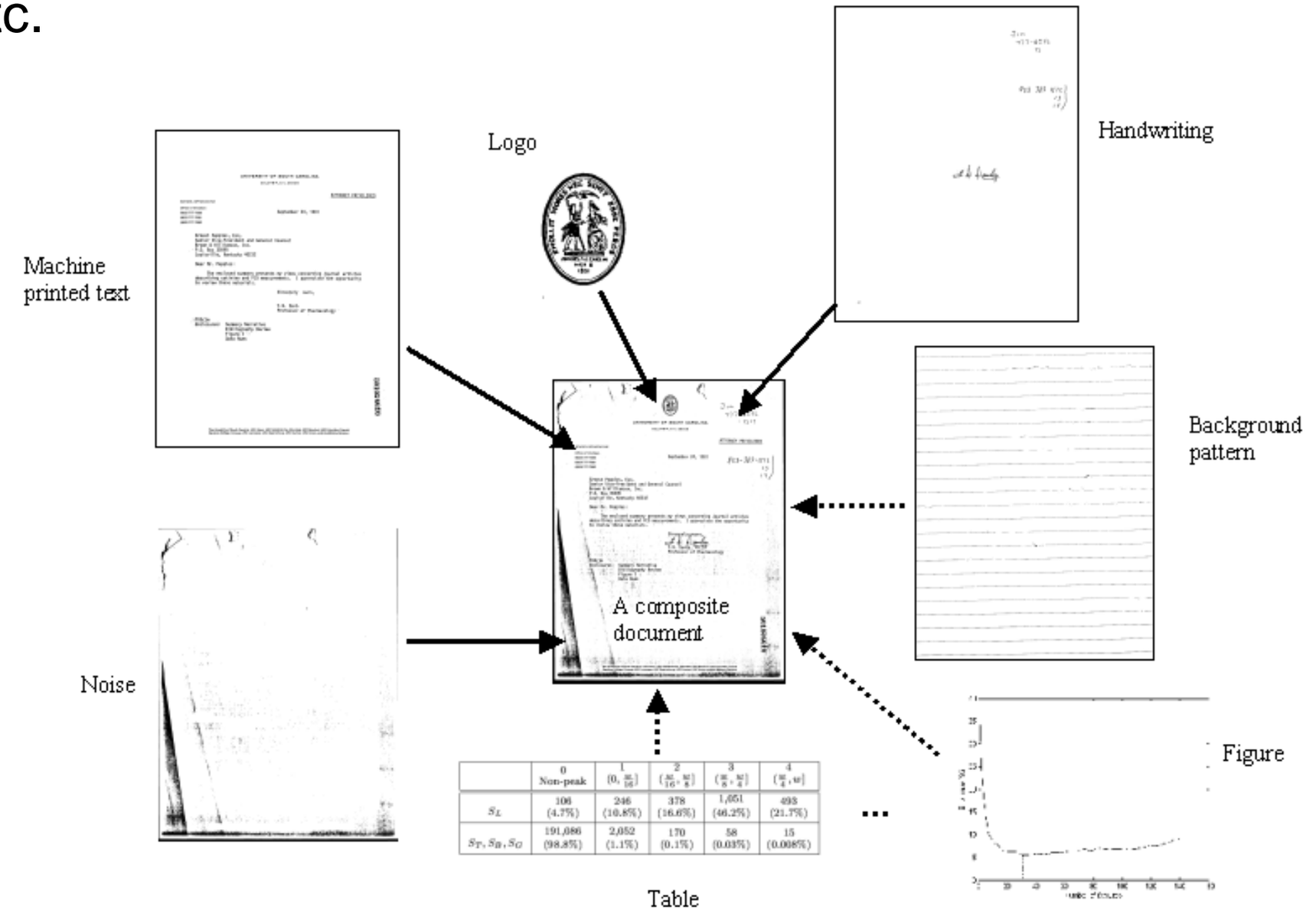
- Classification - what type of document image is this?
- Page segmentation
  - structure
  - identify images
  - identify text
  - identify handwritten text
  - diagram identification
- Meta-data identification
  - title, author
  - language
- OCR
- Reading ordering
- Indexing

Problems we'll discuss today…

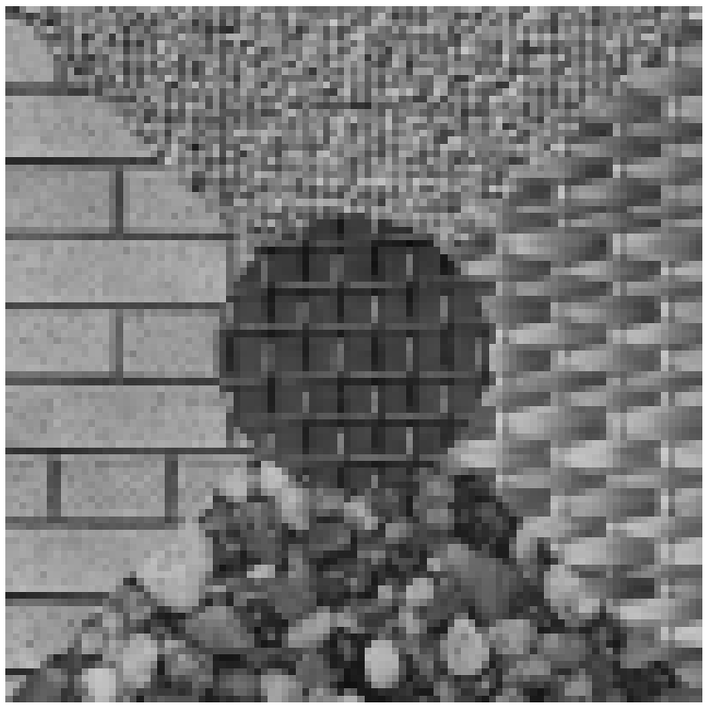
- Preprocessing issues
  - Page Layer Segmentation
  - OCR
  - Reading order
- IR issues

Problem: Page Layer Segmentation

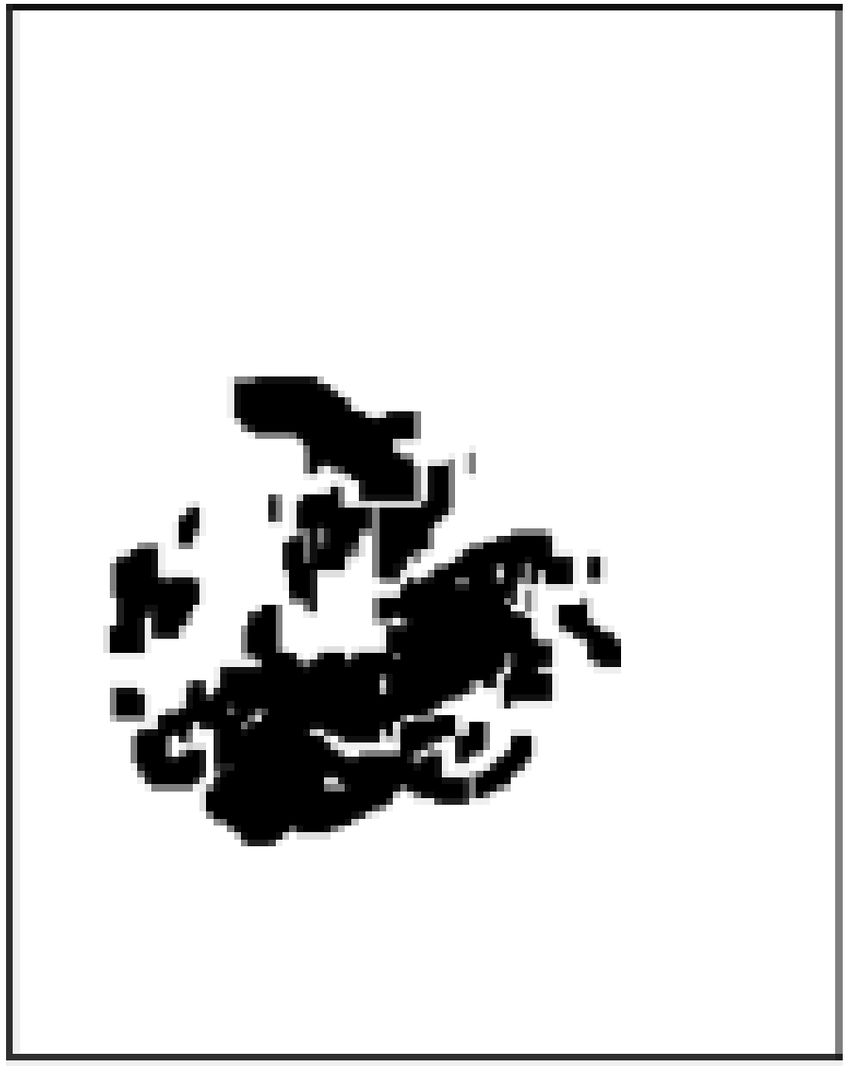
- A document consists of many layers, such as handwriting, machine printed text, background patterns, tables, figures, noise, etc.

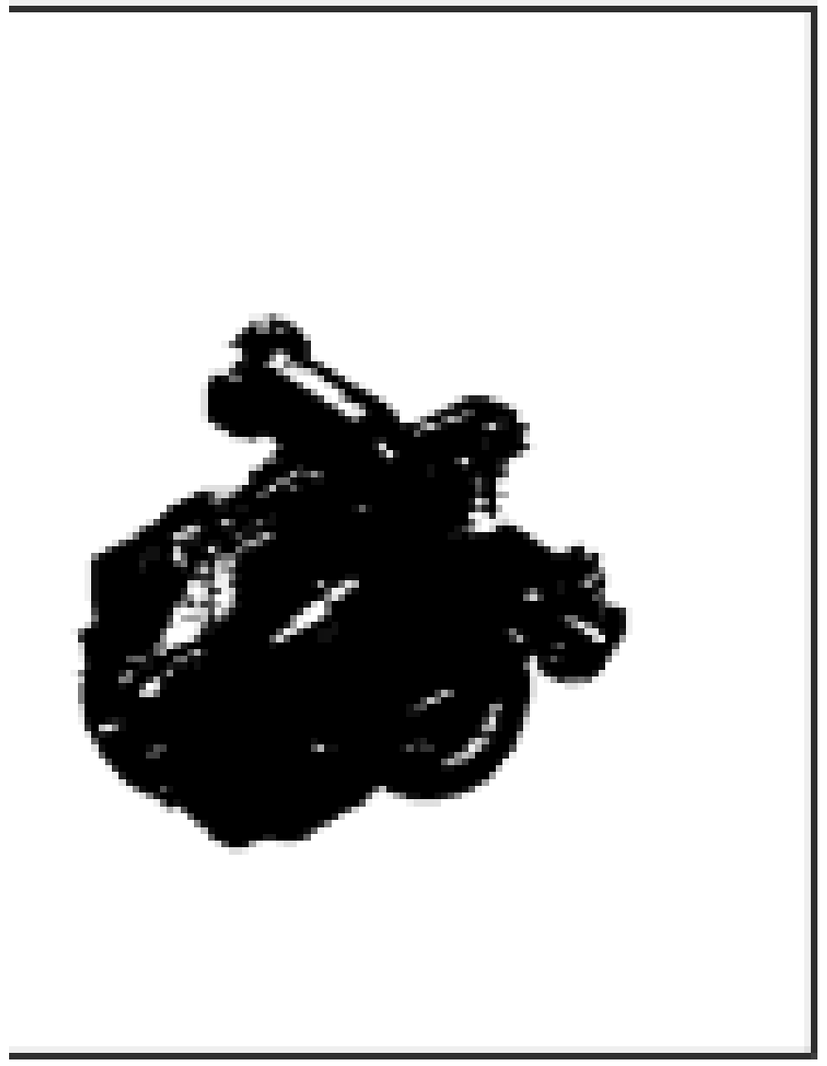


Step 1 - segmentation: Which part is the document?

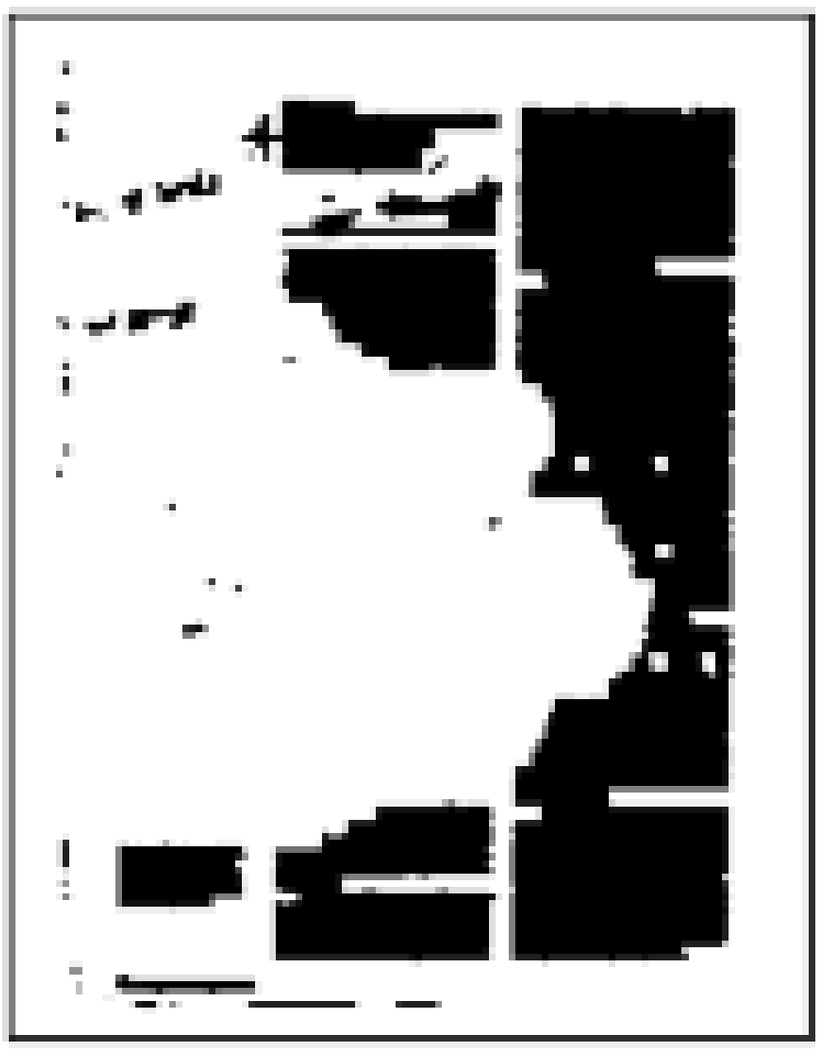


Segmentation





Segmentation

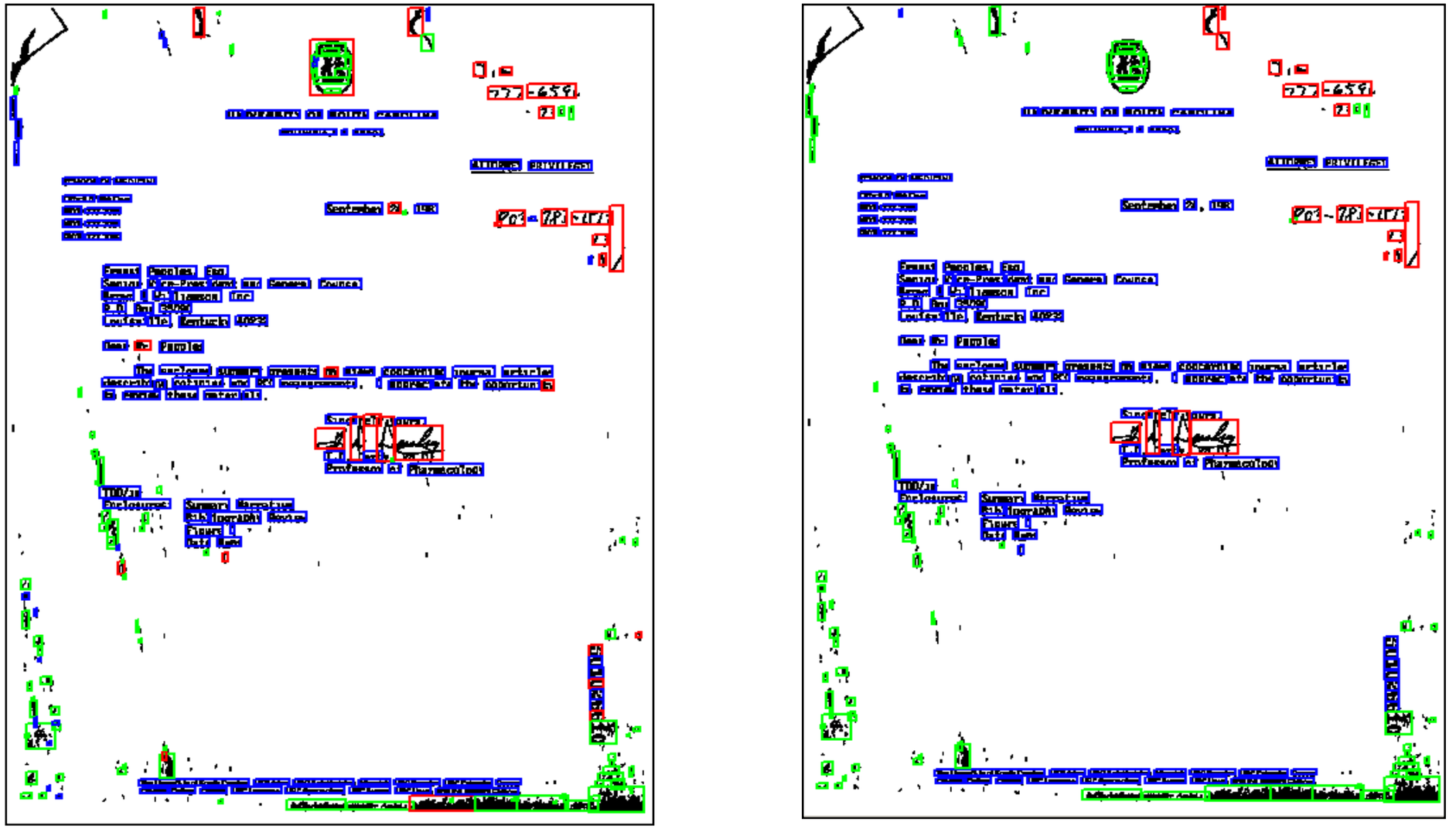


Segmentation

- **Preprocessing is crucial**
  - **Eg.** Convert gray/color image → black-and-white to binarize
  - Noise removal for speckles, dust, scan dust, etc.
- **Early top-down (recursive "X-Y cut" & whitespace analysis) segmentation strategy**
  - Compute **horizontal** and **vertical projection profiles** (sum of black pixels per row/column).
  - Find **valleys** (long runs of zeros/low counts) → pick the biggest whitespace gap.
  - **Cut** the page along that gap; recurse in each sub-region until no big valleys remain.
  - Each leaf region ≈ a block (column, figure, etc.).

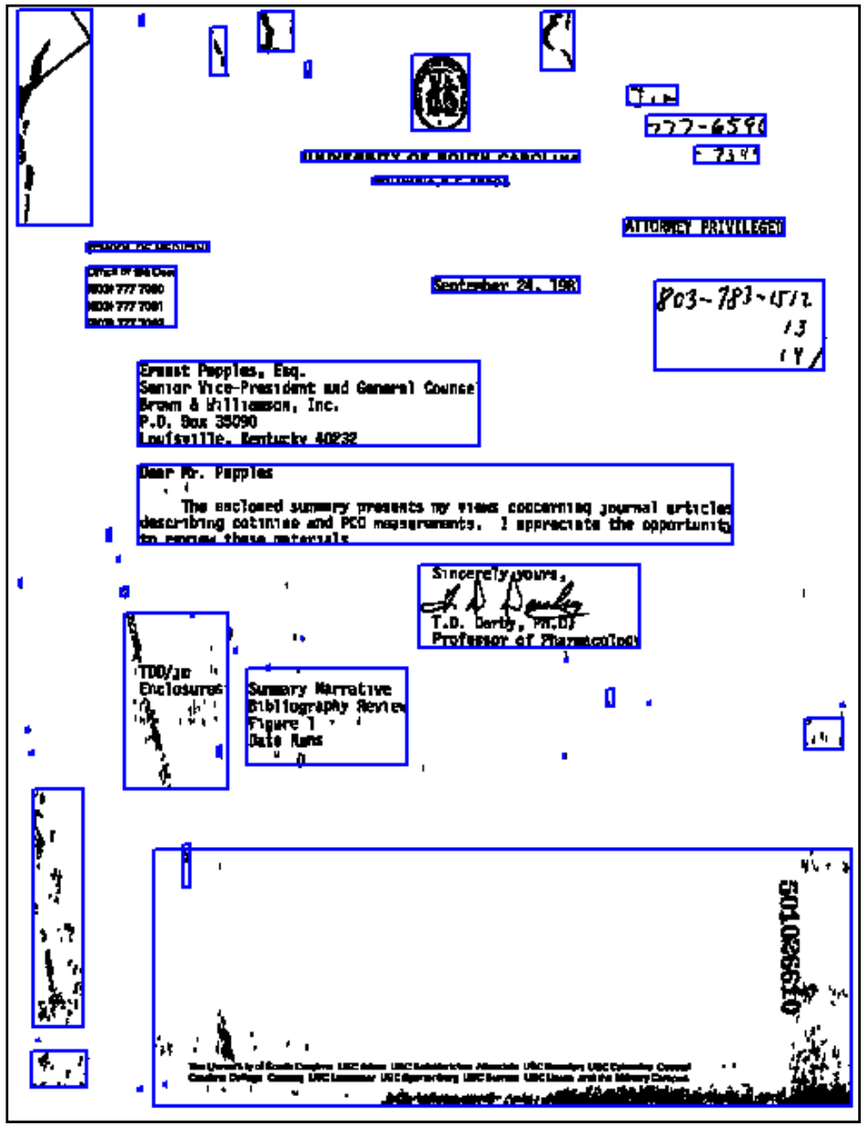
Step 2 – classify the segments

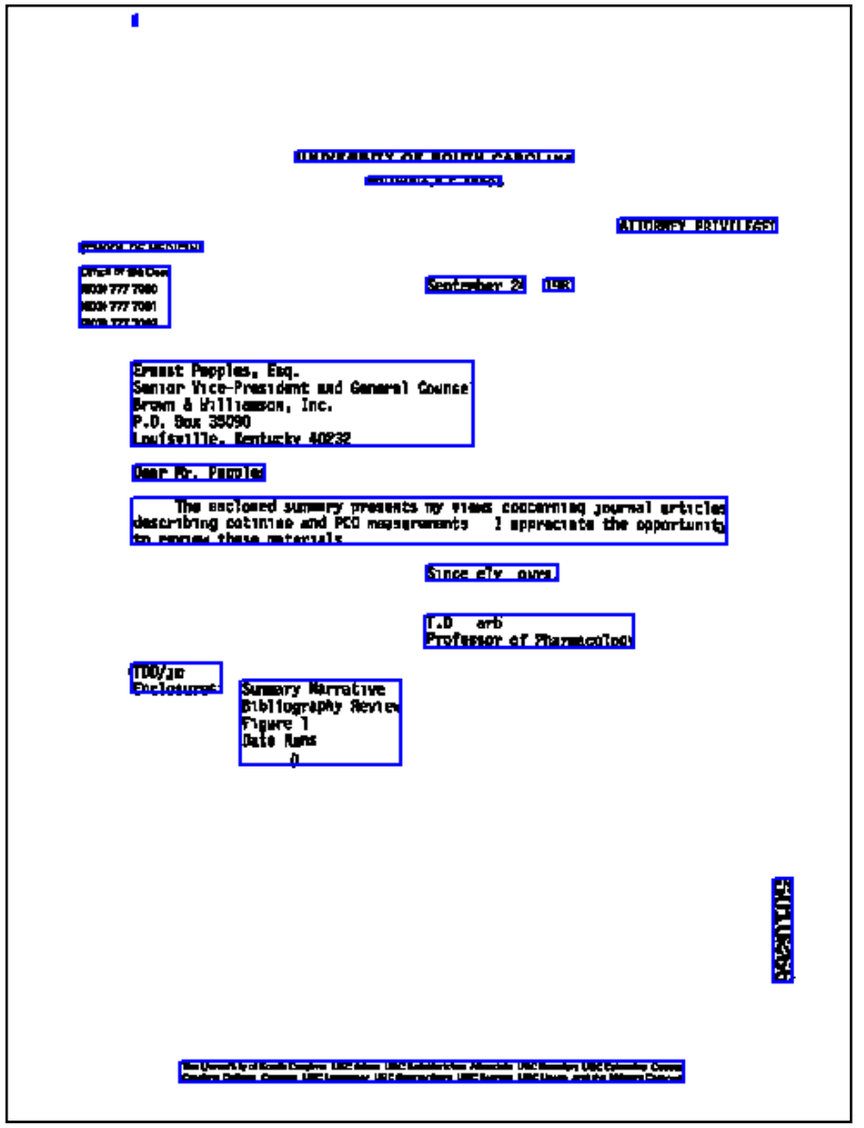
Printed text
Handwriting
Noise



We can use features of the “segment” as well as positional information about the other segments

Segmentation Classification





Before enhancement

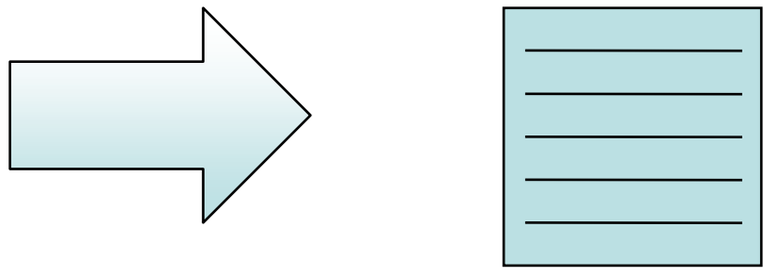
After enhancement

Segment Classification

- **Simplest classification: Text vs not.**
- **Early approaches (1980s): Rule based (**if density > 0.6 and variance small → text.”**)**
- **Late 1990s: Start of use of Supervised ML methods + texture geometry features**
- **Hierarchical models:** Stage 1: Text vs non-text. Stage 2: Printed vs handwritten. Stage 3: Figure vs table.
- Later, around 2005–2015, models like **LayoutLMv1**, **DiT**, etc. replaced these handcrafted features with learned ones

Problem: OCR

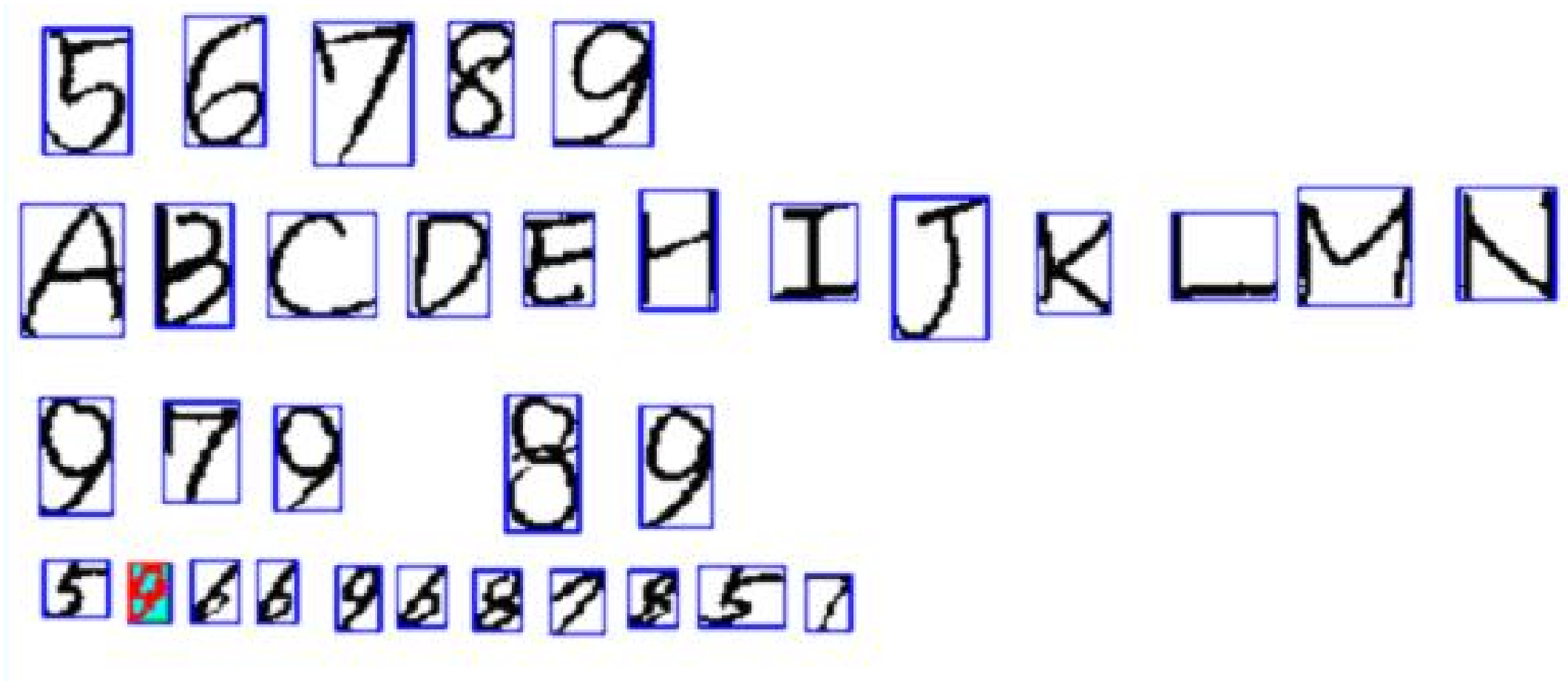
- One of the more successful applications of computer vision



How do pixels turn into text?

OCR: One solution

- Pattern-matching approach
  - Standard approach in commercial systems
  - Segment individual characters
  - Recognize using a neural network classifier



OCR isn’t just for English!

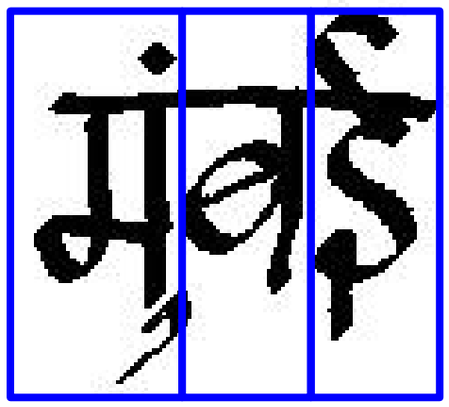
एक बार एक सज्जन आश्रम आये थे, उन्होने बातचीत
के सिलसिले में श्री श्री ठाकुर से कहा कि मुझ पर कृपा
कीजियेगा इस पर श्री श्री ठाकुर ने कहा : 'कृपा ' में
पहला अक्षर 'कृ' है तब 'पा' यानी करने से पाओगे ।"
पुनः उस सज्जन ने कहा, सुनते हैं कि दुर्गापाठ करने

مرحبا بكم على موقع مجلس ايسلنجتون – هو افضل موقع للحصول على معلومات
عن الزيارة والعيش والعمل في ايسلنجتون. يمكنكم ايضا معرفة اين تقع اقرب
صالة سينما وكيفية دفع ضريبة المجلس. اقرؤا عن الخدمات التي يقدمها المجلس
للراشدين والأطفال واقرؤا ايضا عن الديمقراطية في ايسلنجتون وكيف يمكنكم
اعطاء آرائكم بخصوص قرارات المجلس.

Optical Character Recognition

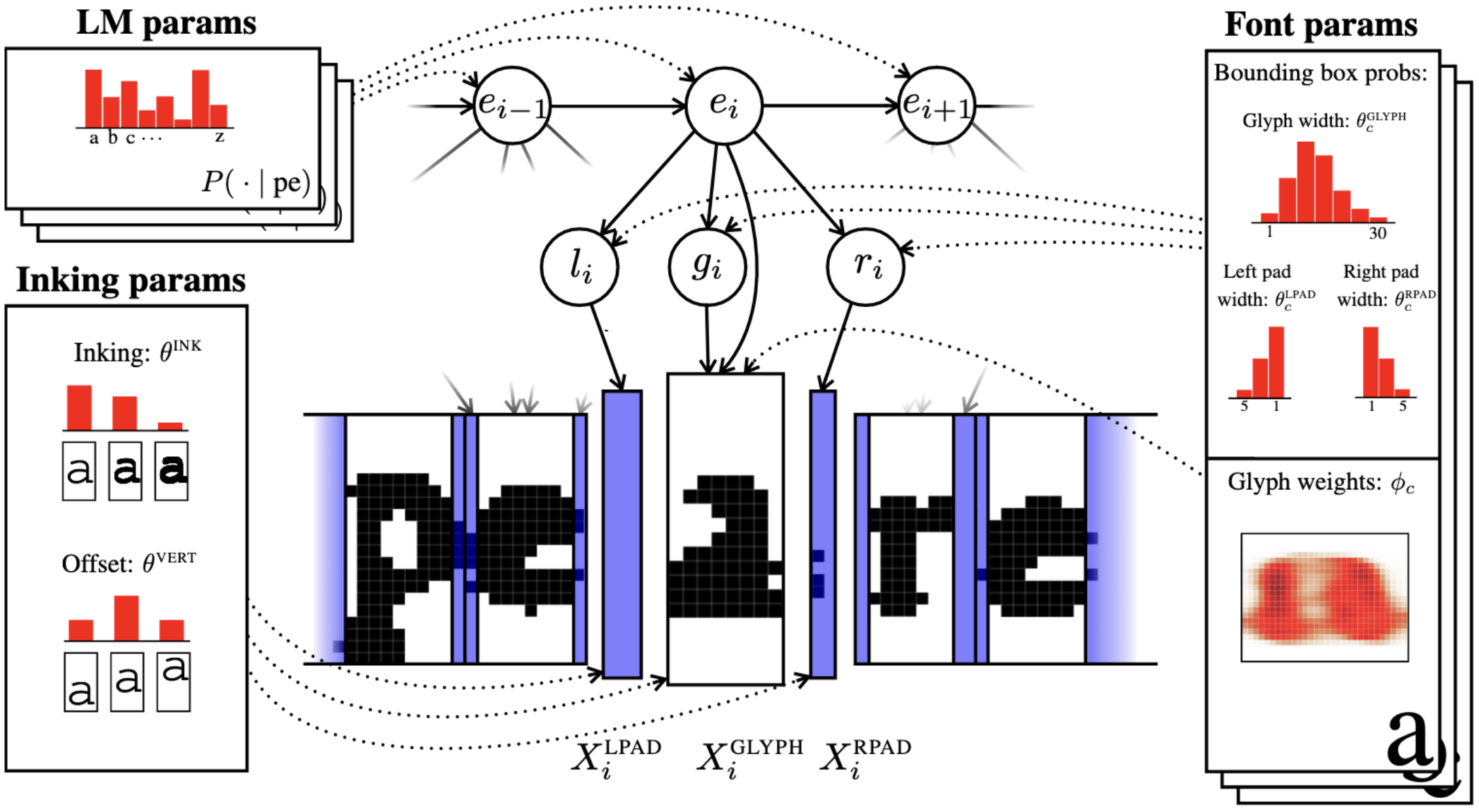
- Hidden Markov model approach
  - Experimental approach developed at BBN
  - Segment into sub-character slices
  - Limited lookahead to find best character choice

Determining character segmentation is difficult!



- Uniform slices
- View as a sequential prediction problem

Fancier approaches try to model multiple aspects of the text



Berg-Kirkpatrick et al. (2013)

OCR Accuracy Problems

- Character segmentation errors
  - In English, segmentation often changes “m” to “rn”
- Character confusion
  - Characters with similar shapes often confounded
- OCR on copies is much worse than on originals
  - Pixel bloom, character splitting, binding bend
- Uncommon fonts can cause problems
  - If not used to train a neural network

Improving OCR Accuracy

- Image preprocessing
  - Mathematical morphology for bloom and splitting
  - Particularly important for degraded images
- "Voting" between several OCR engines helps
  - Individual systems depend on specific training data
- Linguistic analysis can correct some errors
  - Use confusion statistics, word lists, syntax, …
  - But more harmful errors might be introduced

OCR Speed

- Neural networks take about 10 seconds a page
  - Hidden Markov models are often slower
- Voting can improve accuracy
  - But at a substantial speed penalty
- Easy to speed things up with several machines
  - For example, by batch processing - using desktop computers at night

Challenge with OCR is there is a often a trade-off between speed and accuracy

Problem: Reading Order

The Washington Post

DISTRICT FINAL

Friday, June 9, 2006

BEING A BLACK MAN

After Zarqawi, No Clear Path In Weary Iraq

Difficult Questions Surround Legacy of Insurgent Leader

The Young Apprentice

Marcus's Parents Agonize Over How to Protect -- and Prepare -- Him

How U.S. Forces Found Iraq's Most-Wanted Man

Kaine Delays Execution Of Inmate for 6 Months

Inquiry Into Killer's Mental State Ordered

INSIDE

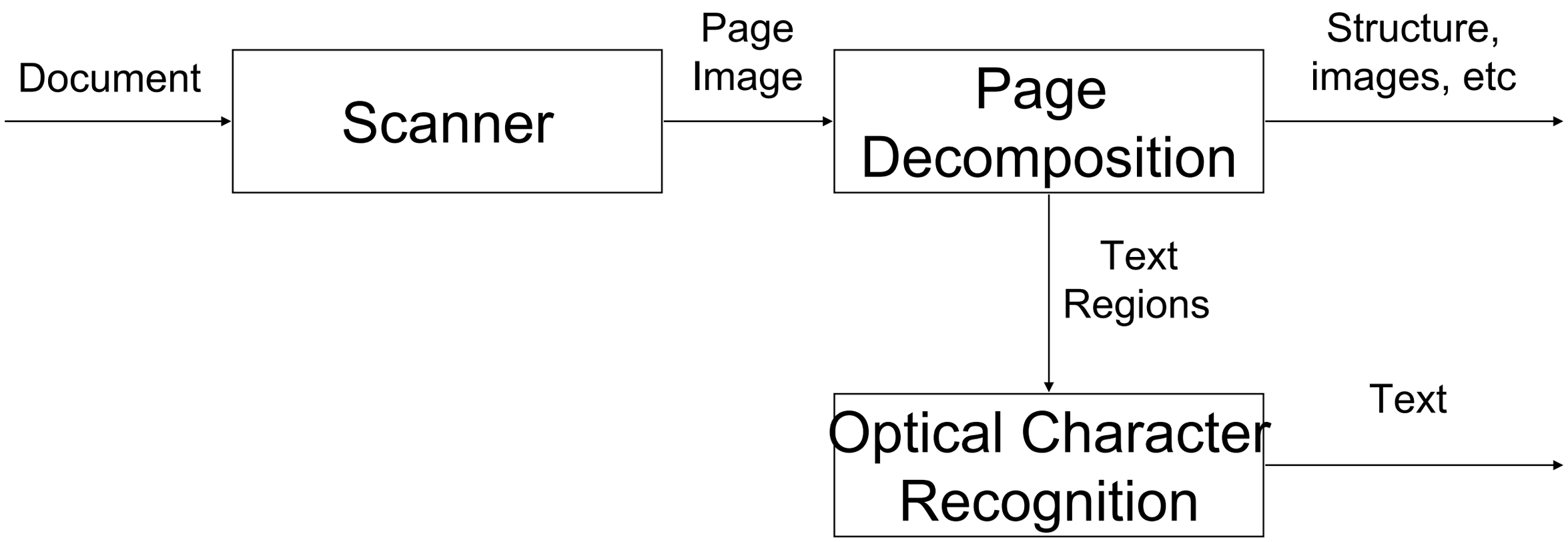
FDA Approves Vaccine That Should Prevent Most Cervical Cancers

What is the sequence of words from this document?

Logical Page Analysis

- Can be hard to guess in some cases
  - Newspaper columns, figure captions, appendices, …
- Sometimes there are explicit guides
  - “Continued on page 4” (but page 4 may be big!)
- Structural cues can help
  - Column 1 might continue to column 2
- Content analysis is also useful
  - Word co-occurrence statistics, syntax analysis

Traditional Approach



Remember our goal

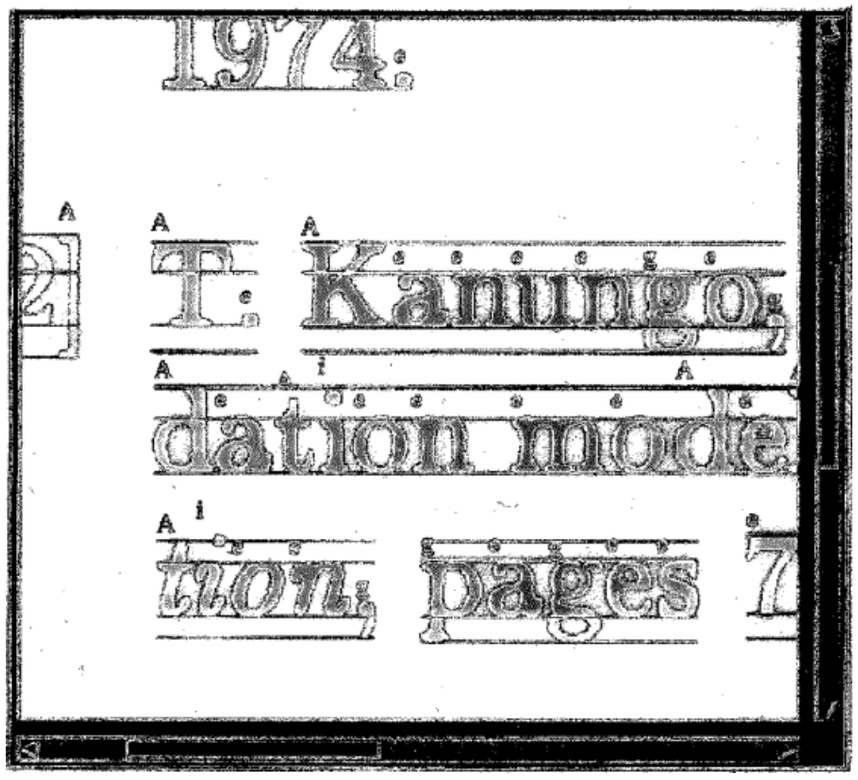
- Create an IR system over image documents
- Challenge: OCR is not perfect
  - Success for high quality OCR (Croft et al 1994, Taghva 1994)
  - Limited success for poor quality OCR (1996 TREC, UNLV)

Proposed Solutions

- Improve OCR
- Again, speed is always a concern
- Similar to spelling correction
  - Automatic Correction
  - Characters N-grams
    - Statistically robust to small numbers of errors
    - Rapid indexing and retrieval
    - Works from 70%-85% character accuracy where traditional IR fails

Shape Coding

- Approach
  - Use of Generic Character Descriptors
  - Map Character based on Shape features including ascenders, descenders, punctuation and character with holes



•Presence of **ascenders** (strokes going above midline — e.g., *b, d, h, k*).
•Presence of **descenders** (below baseline — *g, p, q, y*).
•Presence of **holes** (closed loops — *a, b, d, e, o, p, q*).

Shape Codes

- Group all characters that have similar shapes
  - {a, c, e, n, o, r, s, u, v, x, z}
  - {b, d, h, k, }
  - {f, t}
  - {g, p, q, y}
  - {i, j, l, 1, I}
  - {m, w}
- Shape codes whether a subset of an image belongs to a given character set
- Sub-process later based on linguistic and/or OCR

Why Use Shape Codes?

- Can recognize shapes faster than characters
  - Seconds per page, and very accurate
- Preserves recall, but with lower precision
  - Useful as a first pass in any system
- Easily extracted from JPEG-2 images
  - Because JPEG-2 uses object-based compression

Evaluation

- The usual approach: Model-based evaluation
  - Apply confusion statistics to an existing collection

- A bit better: Print-scan evaluation
  - Scanning is slow, but availability is no problem

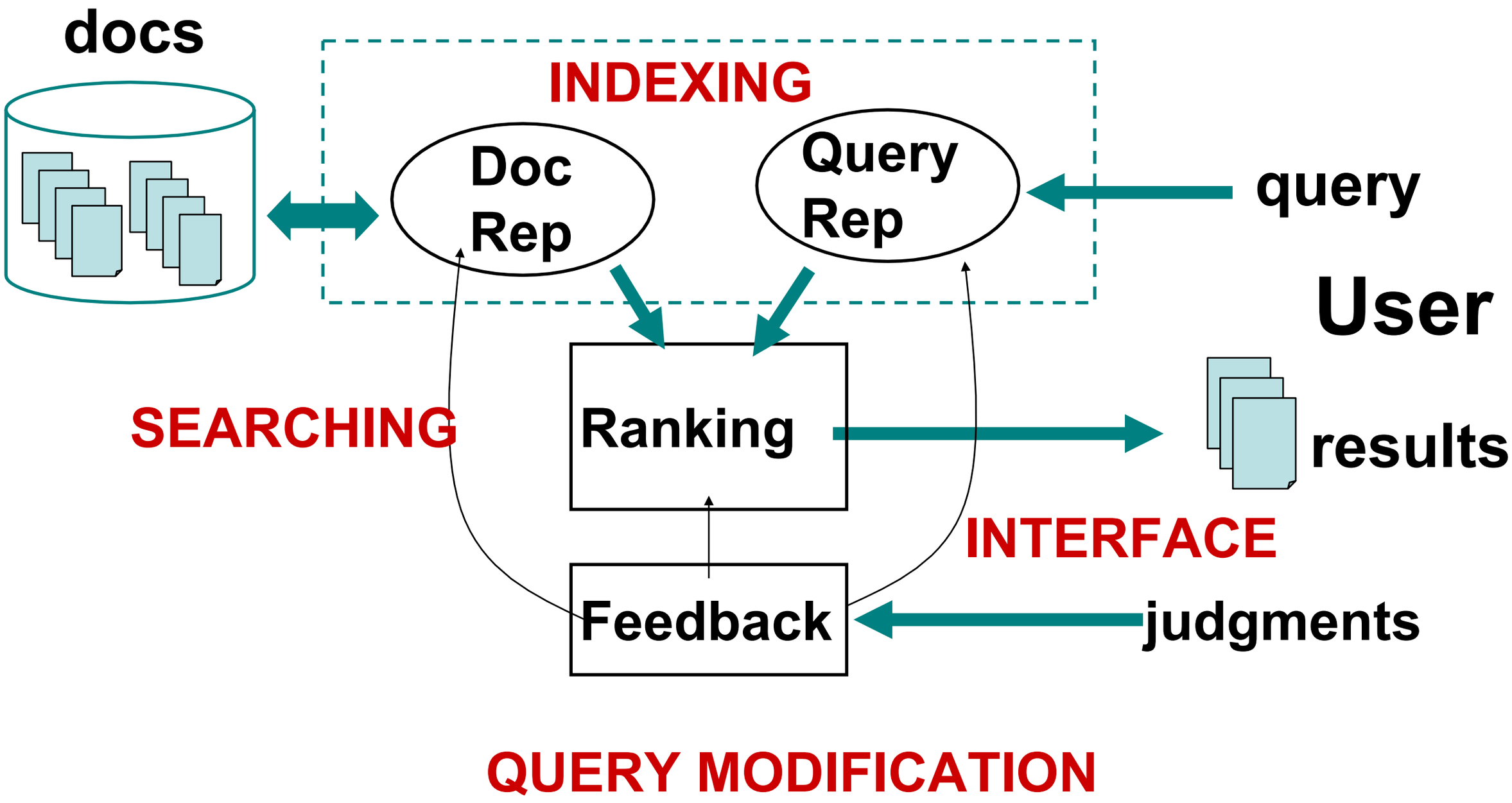
- Best: Scan-only evaluation
  - Few existing IR collections have printed materials

Summary

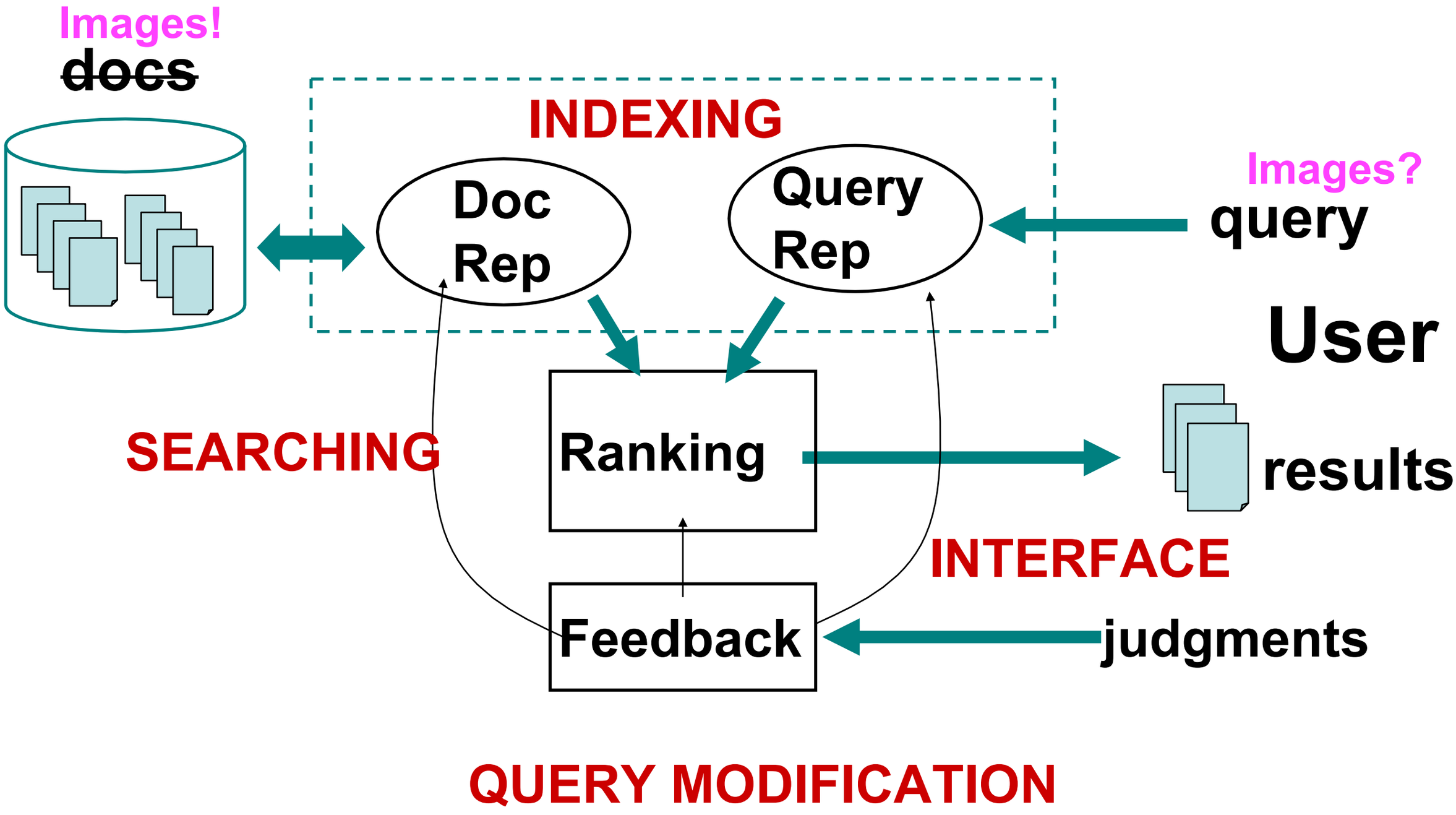
- Many applications benefit from image based indexing
  - Less discriminatory features
  - Features may therefore be easier to compute
  - More robust to noise
  - Often computationally more efficient
- Many classical IR techniques have application for DIR
- Structure as well as content are important for indexing
- Preservation of structure is essential for in-depth understanding

Image Search

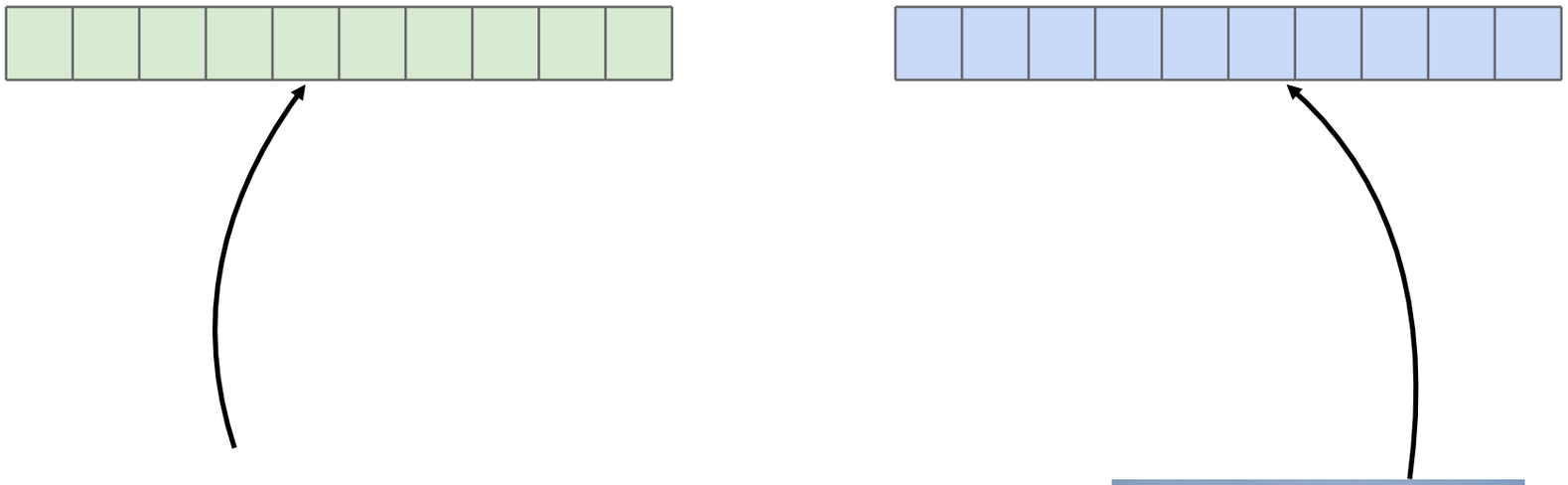
IR System Architecture



Core issue: Different representations in images and text

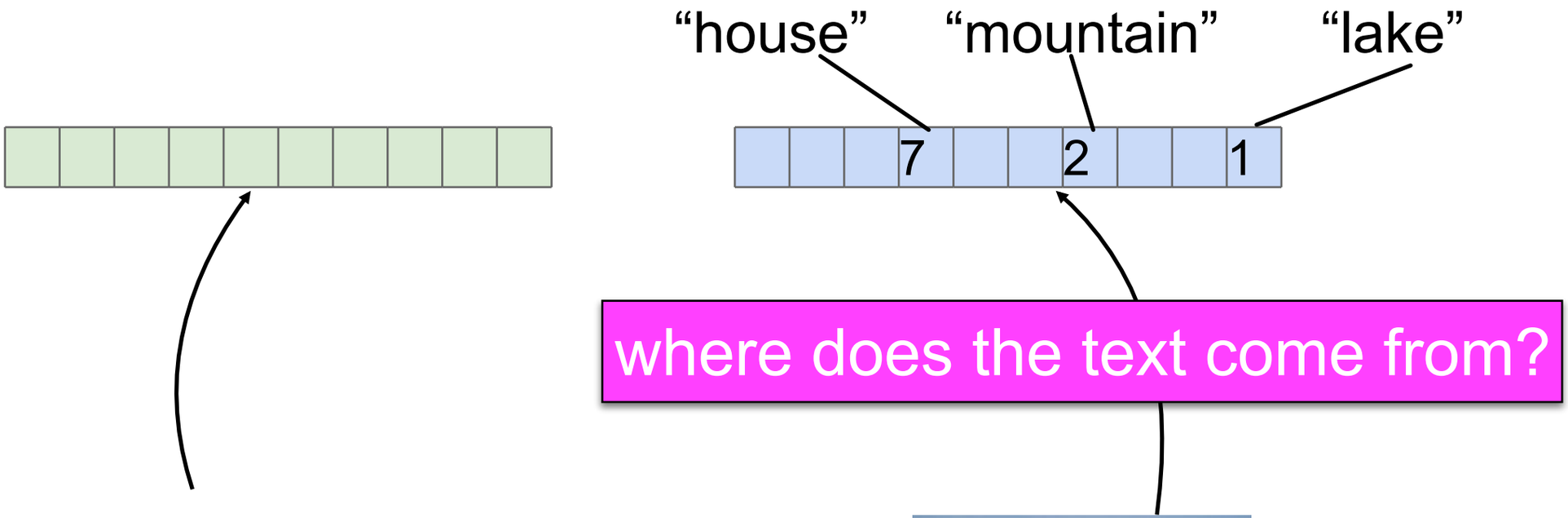


Core Question: How to represent an image?



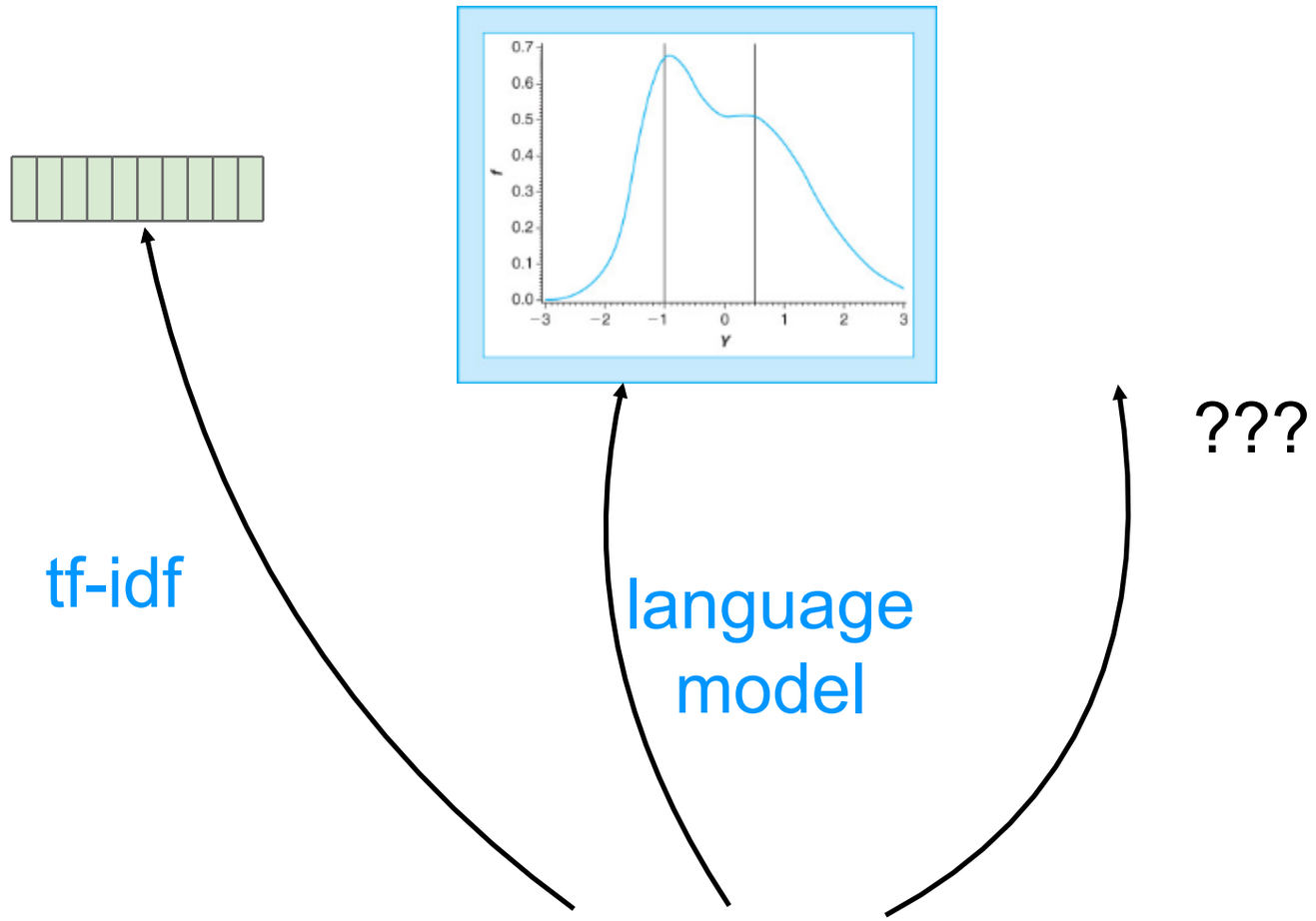
Query: “Scene of houses in front of mountains and next to a lake”

Option 1: Turn the image into “text”

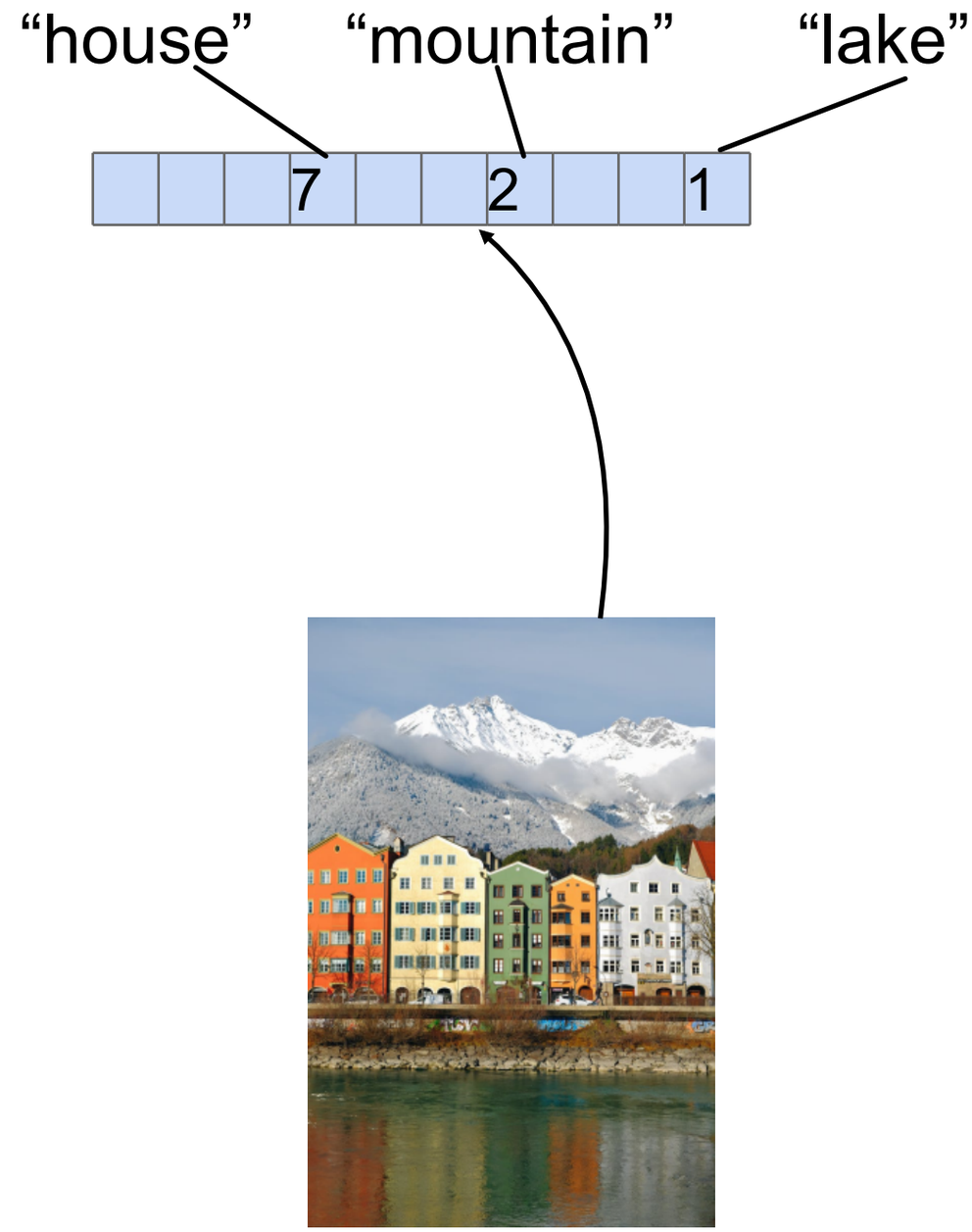


Query: “Scene of houses in front of mountains and next to a lake”

Image Search requires rethinking what is an “encoder” for queries and documents



How might we design an encoder to get text features for images on web pages



Collecting Image Data on the Web for IR

- Autonomous “spiders”
  - Traversal Spider – “assembles lists of candidate Web documents that may include images, videos, or hyperlinks to them”
  - Hyperlink Parser – “which extracts the Web addresses of images and videos”
  - Content Spider – “which retrieves, analyzes, and iconifies the images and videos”

Image Data Collection Process

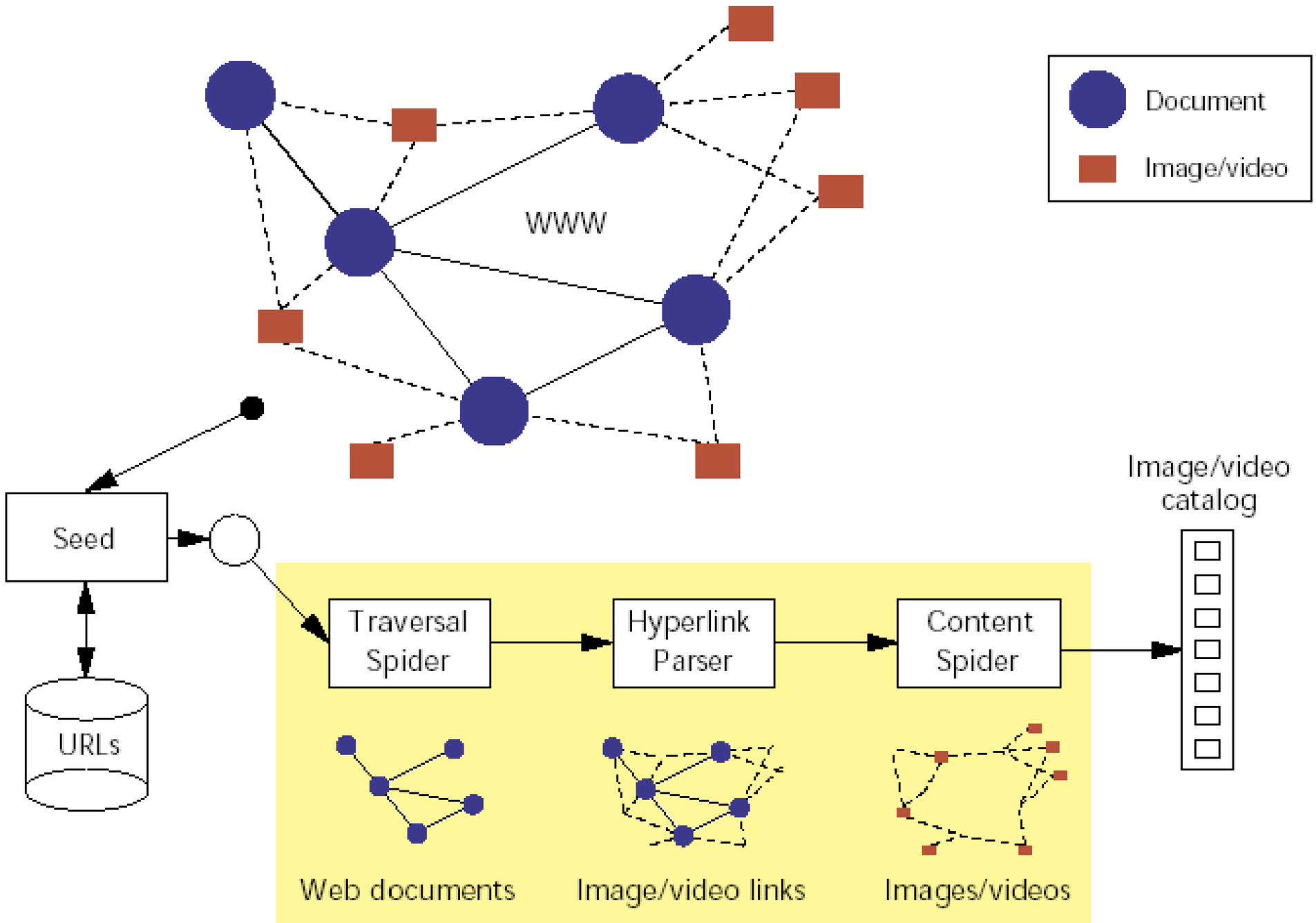
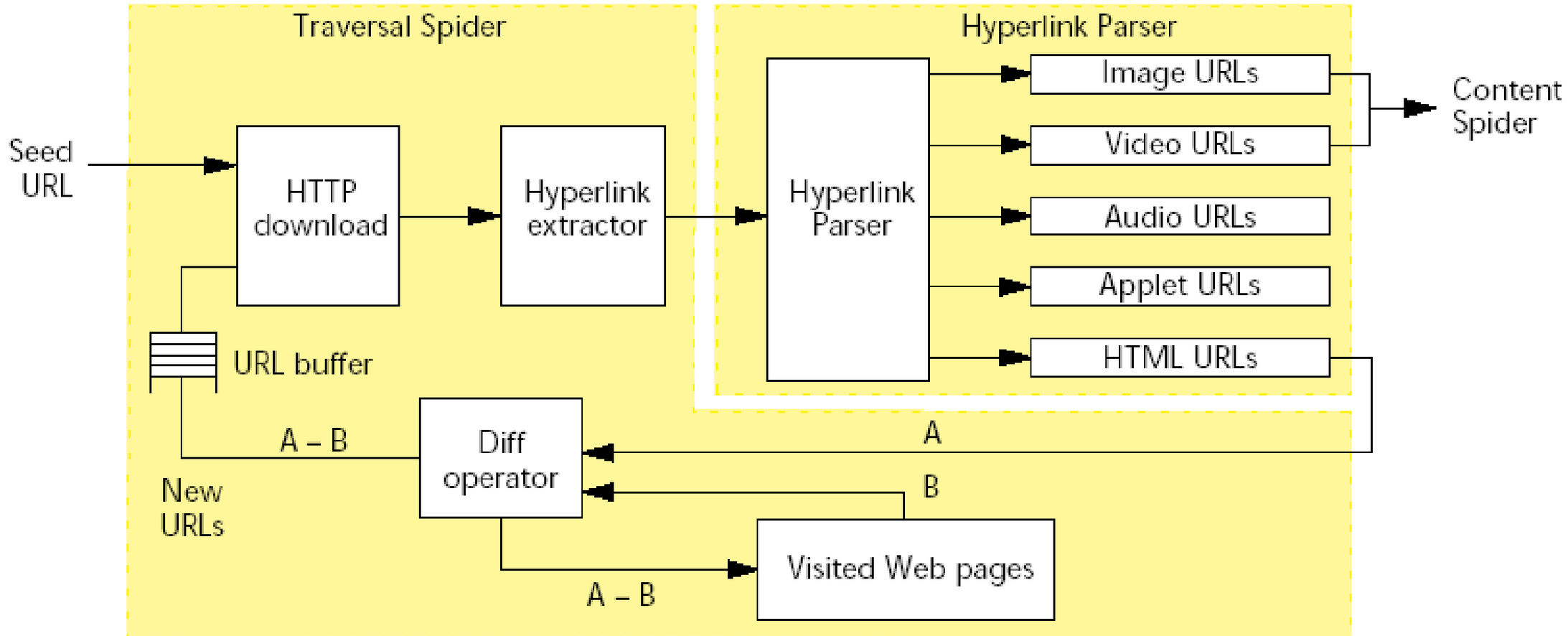


Image Data Collection Process

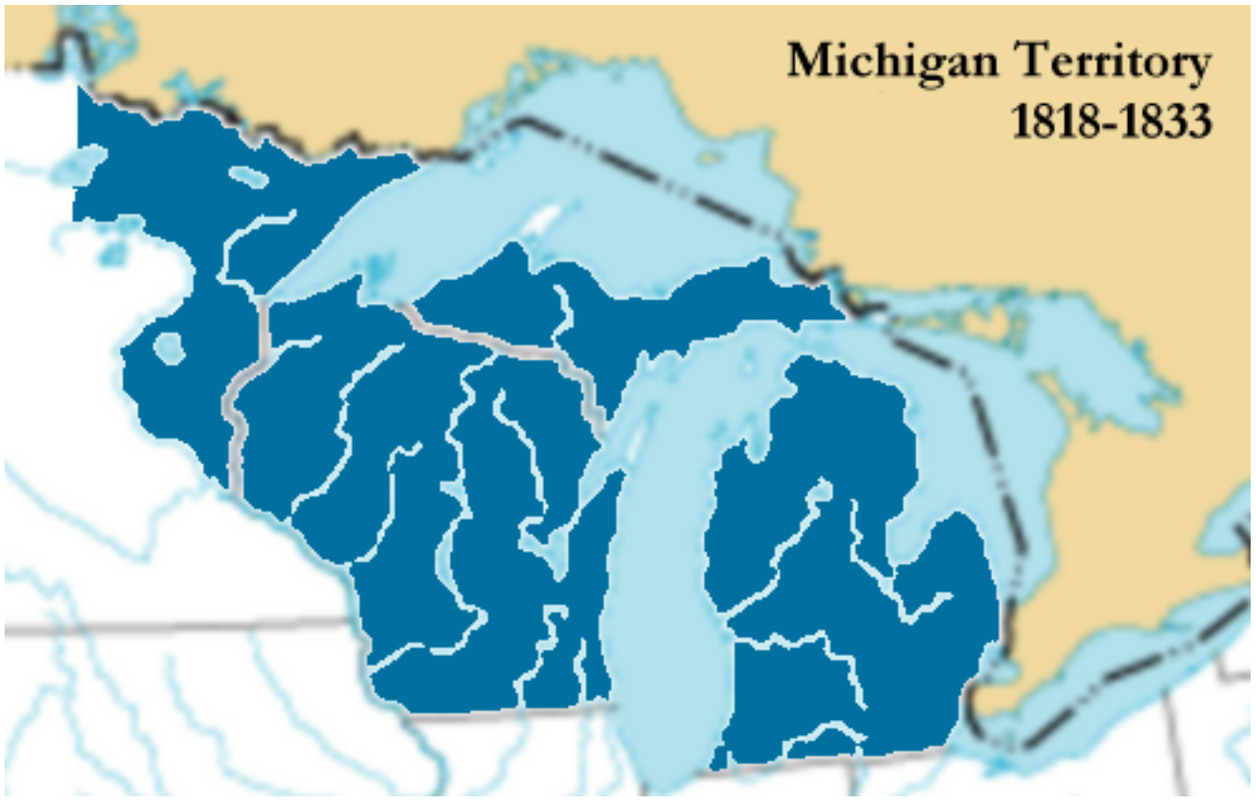


Data Collection Process

- Content Spider Functions
  - “extracts visual features that allow for content-based searching, browsing, and grouping”
  - “extracts other attributes such as width, height, number of frames, type of visual data”
    - Color histogram
  - “generates an icon, or motion icon, which sufficiently compacts and represents the visual information to be used for browsing and displaying query results”
    - Compression algorithms

Idea: Image Subject Classification Process

- Text provides clues about the semantic content of visual information
  - URL
  - File name



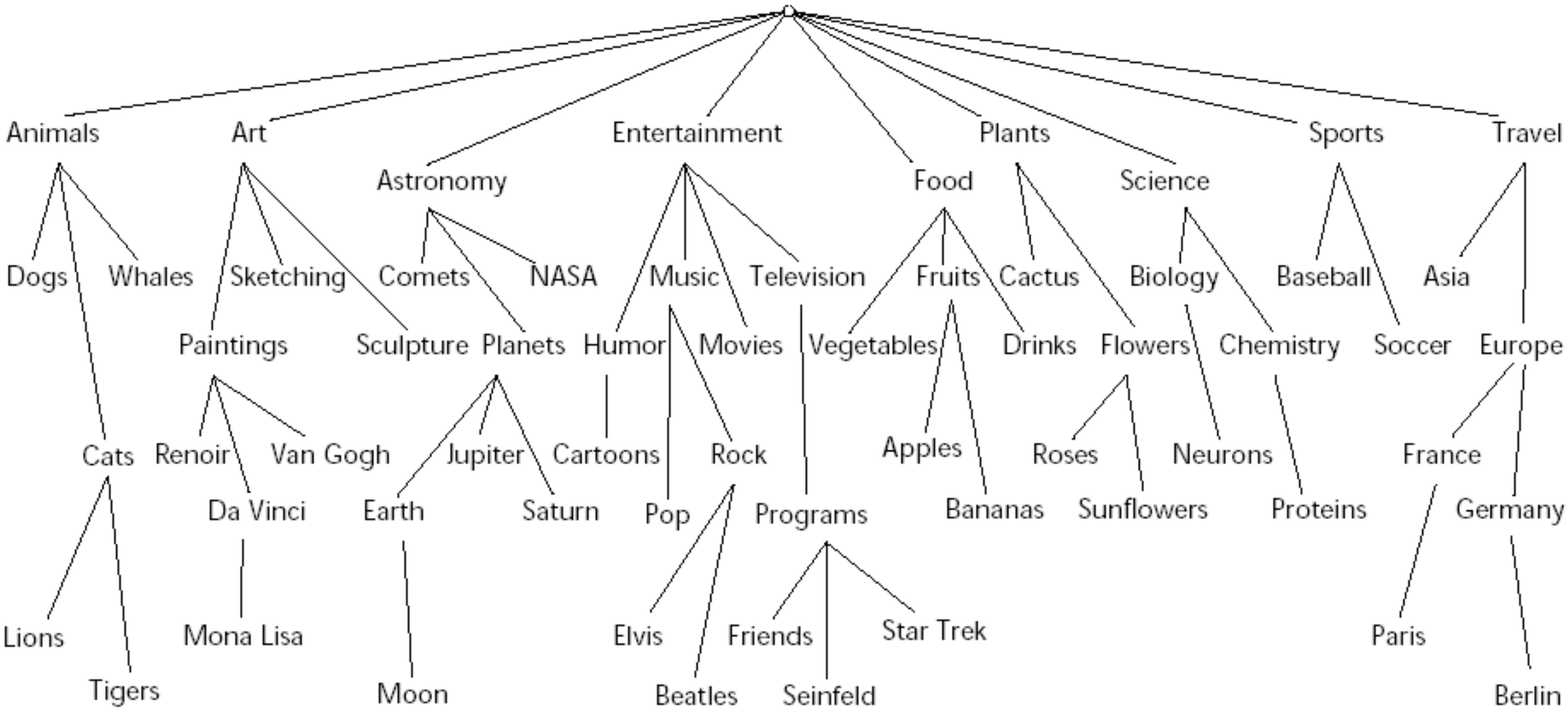
- Text clues can be found in HTML syntax
  - <img src=URL alt=[alt text]>
  - [hyperlink text]

tps://upload.wikimedia.org/wikipedia/commons/3/3e/Michigan-territory-1830-blue.pr

Image Subject Classification Process

- Term extraction
  - Extracted from URLs, alt tags, hyperlink text by removing non-alpha characters
  - F_{key} (URL) = F_{chop} (“animals/domestic-beasts1/dog37”) = “animals,” “domestic,” “beasts,” “dog.”
- Dictionary name extraction
  - F_{dir} (URL) = “animals/domestic-beasts.”
- Key-term dictionary
  - Terms and Dictionary names are used to create t^*_k terms
  - t^*_k terms identified semantically related to subject classes s_m
  - M_{km}: $t^*_k \rightarrow s_m$

Taxonomies provide structure to the Subject Classification Process



Tags can provide usable metadata

https://www.flickr.com/photos/peeblespair/23672999691/in/photolist-C4Uj46-h9iPQA-29FrQoy-289x1qQ-tQbTYe-2kHzJzE-ivWzxr-dukpTQ-bW98zj-dukpFU-erQiWq-a4gE8i-dEBB6f-u2RM8o-2komCaA-doXUXX-ArGLVD-svnAQy-xvtqM4-xNaJd9-h9k4Ha-bW98tY-23LY1Qe-oEwKx7-dueP6p-dT12Dm-gN2FBo-dptS8u-fPuWR1-dptRPu-ebpx91-pRCoSH-ivWsyE-h9iN8g-ndSV2Z-22YMeYq-ffLB7k-L9so9-t8mbuc-ivWa6H-B2aW9j-bW98pj-fPuVBy-bW98K7-dKALmR-2kifRMf-wFdkDh-2hoVP1j-h9iPMQ-erQnHu

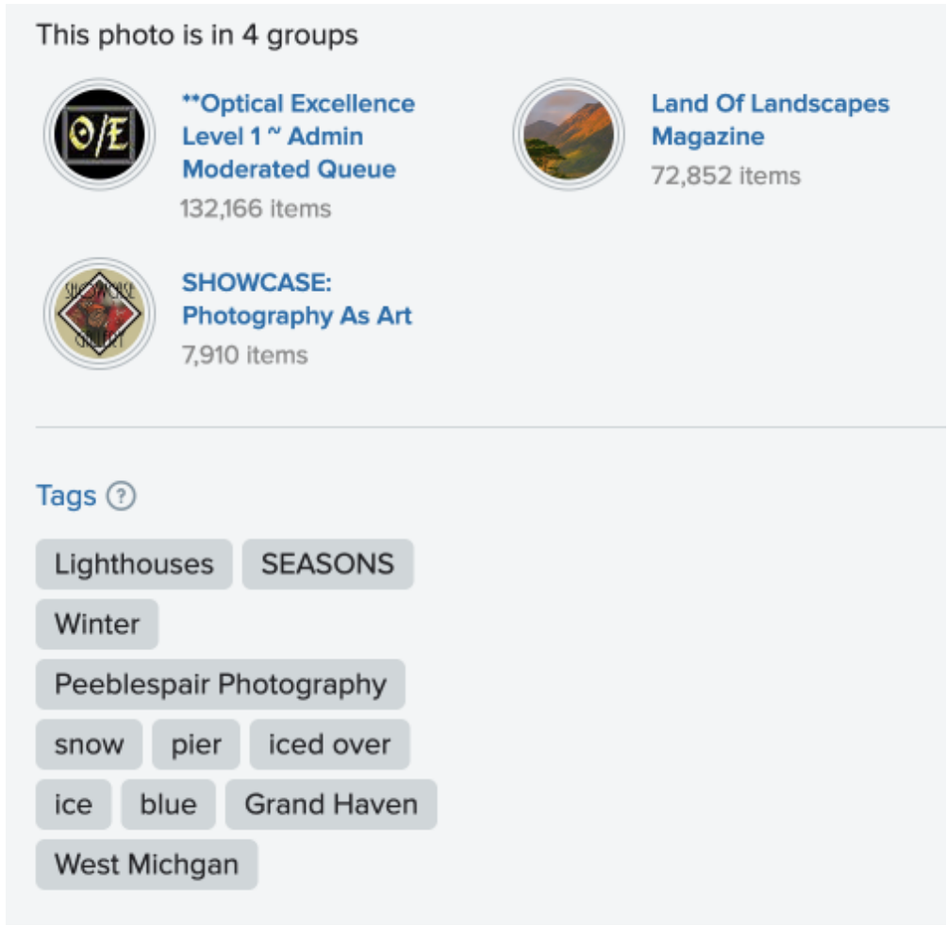


Image Search with Traditional IR

- Extract all textual content from image URL, related text, and metadata
- Put content in index and use traditional IR methods

Governor Gretchen Whitmer (D) speaking at a National Guard ceremony in 2019

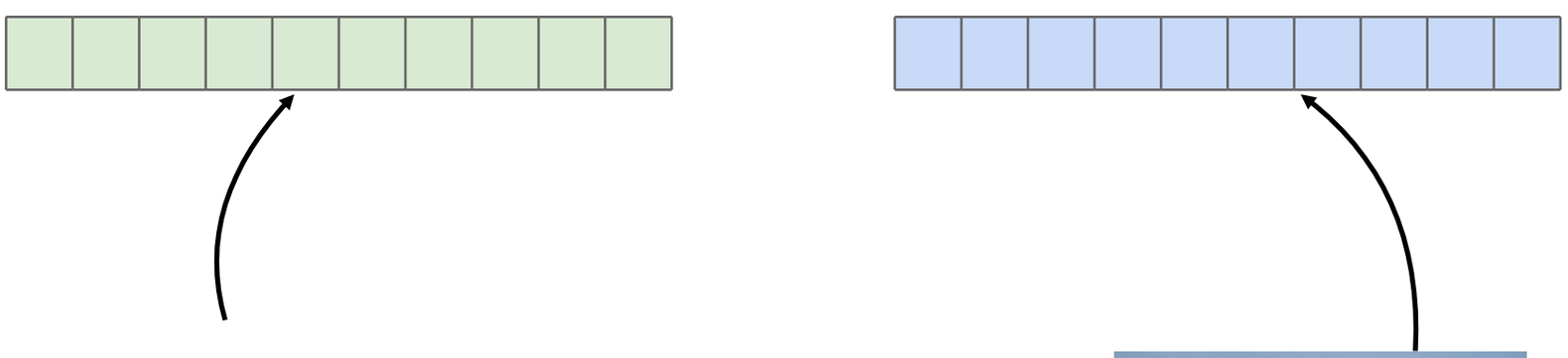
http://salinejournal.com/wp-content/uploads/2019/01/Michigan-Governor-Gretchen-Whitmer-2019-Inauguration-Address-Lansing_8108466brc0c-Saline-Journal-678x381.jpg

What could go wrong with image agnostic approach?

- No related content in URL, text, metadata
- Wrong image descriptions (or worse)
- Some images are hard to describe!
- Other issues?

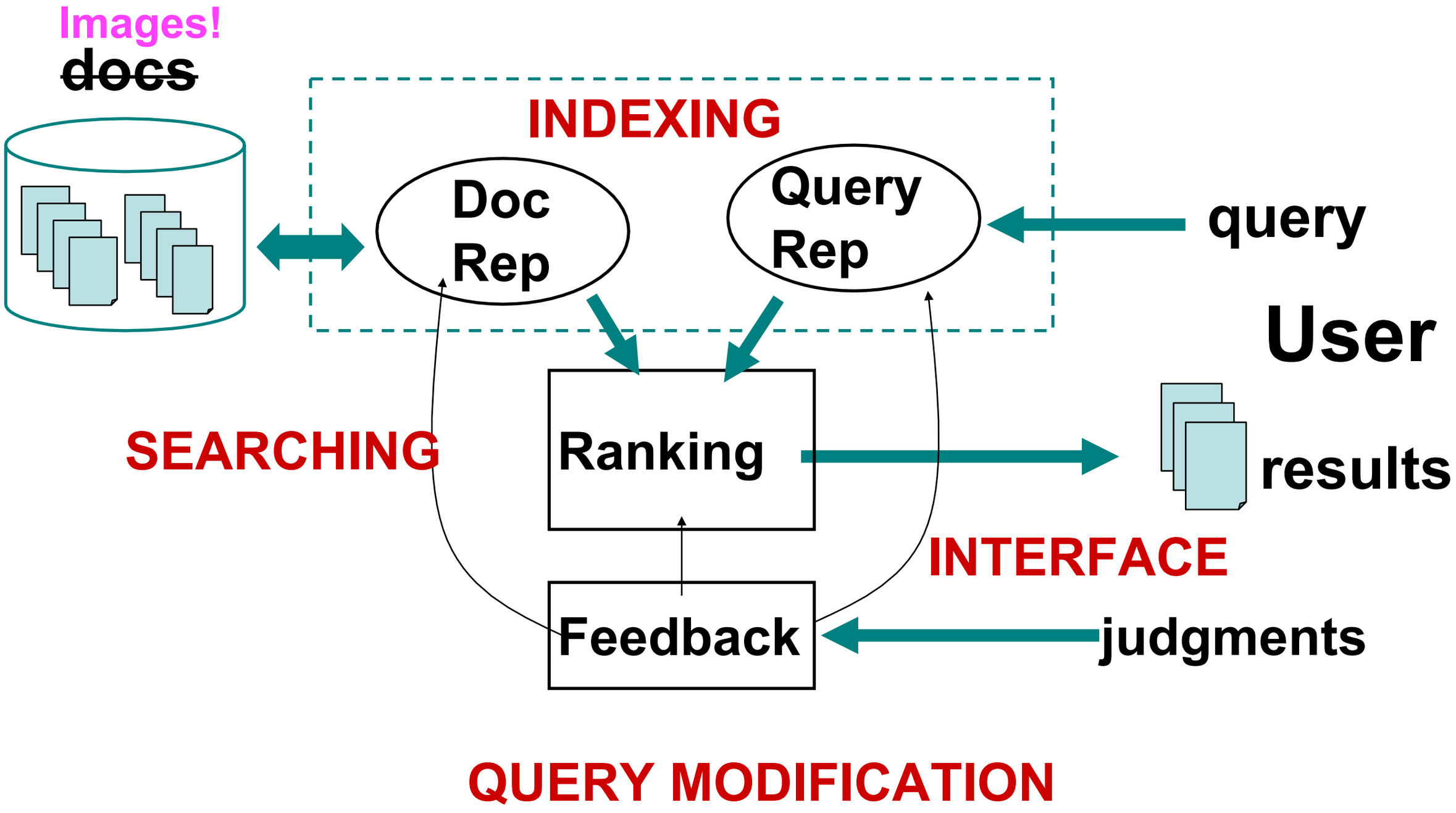
Option 2: Turn the query and image into something else…

Could we make these have the same representation?

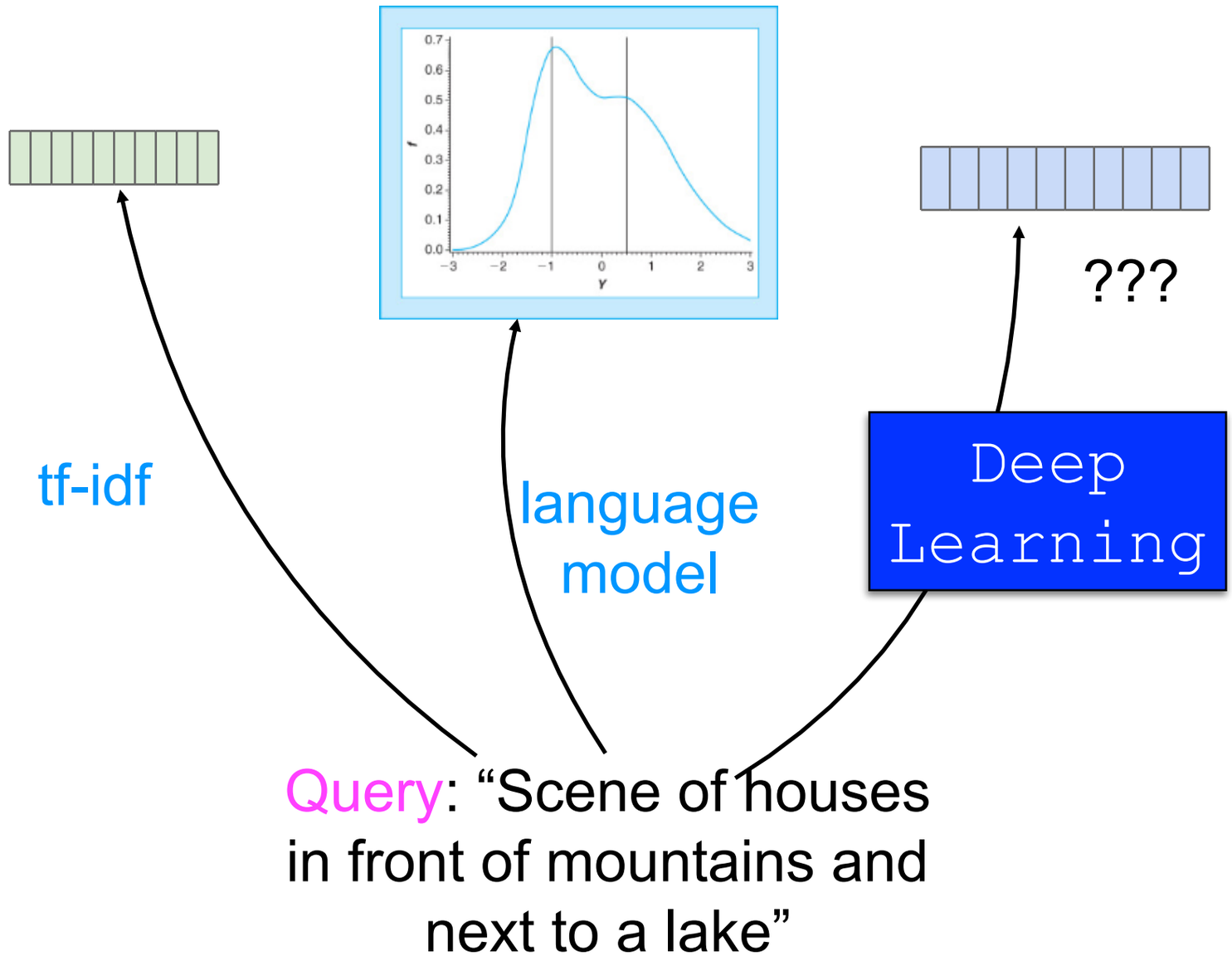


Query: “Scene of houses in front of mountains and next to a lake”

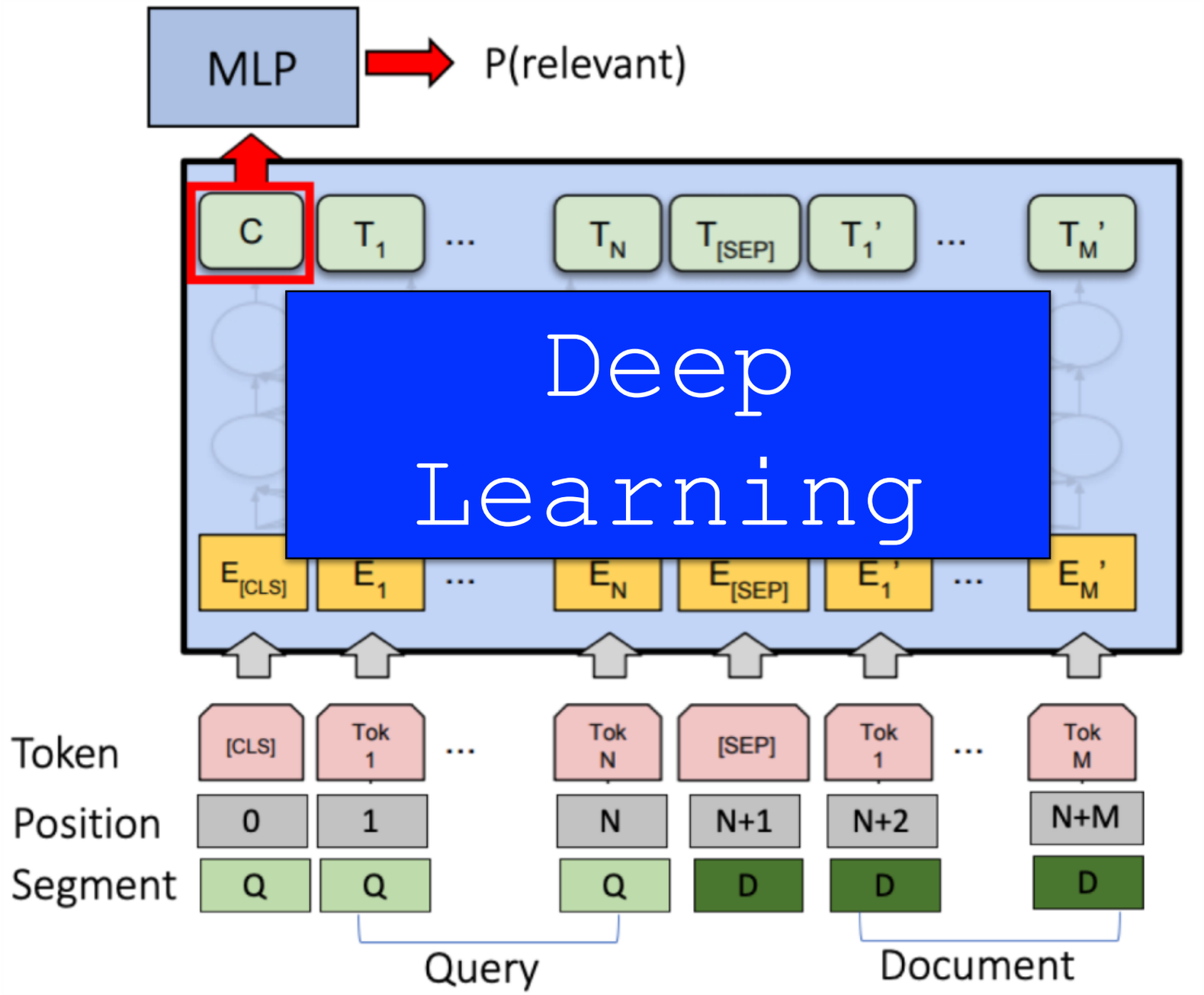
Core issue: Different representations in images and text



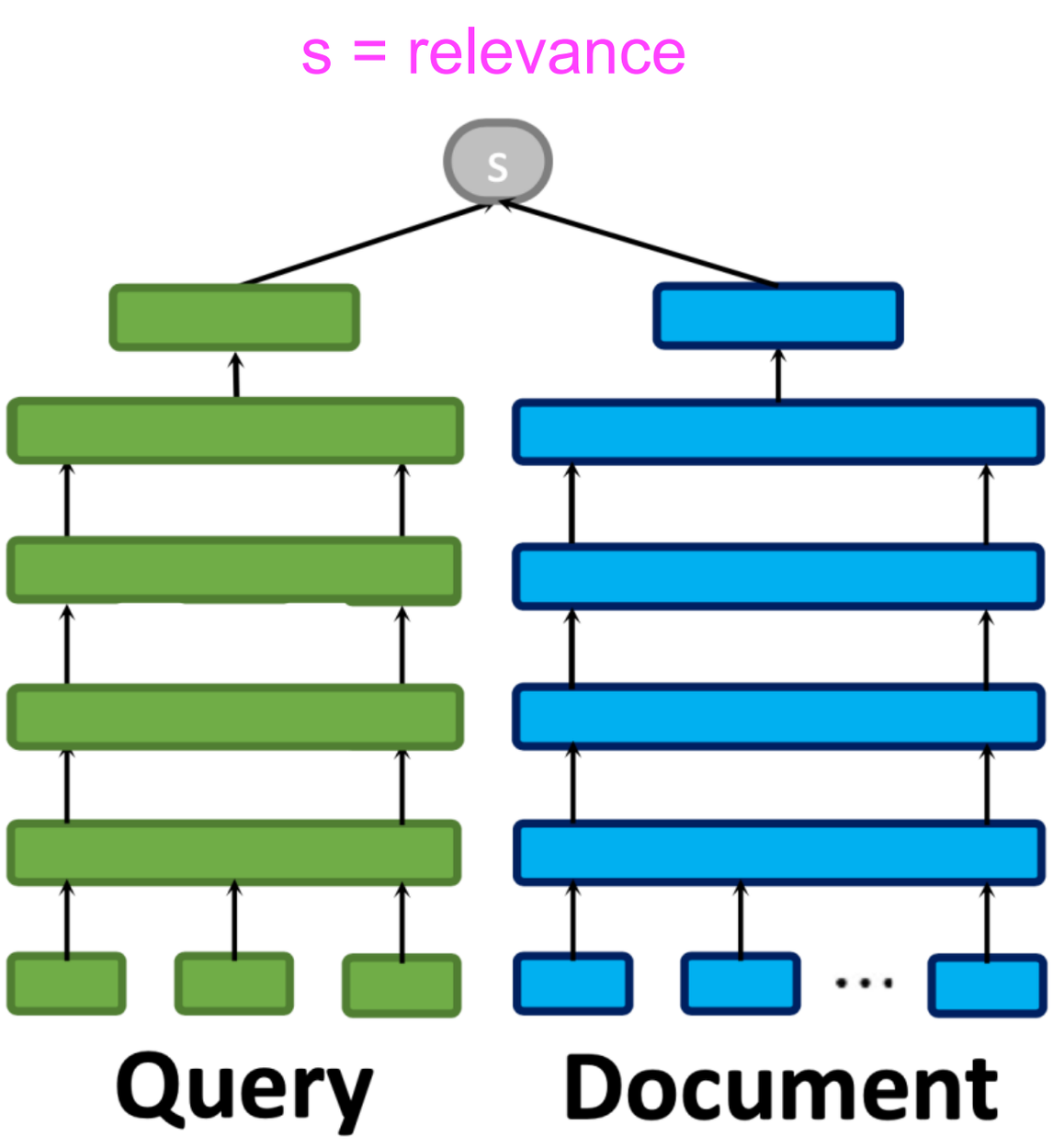
We can think of an encoder as a function of some input to a vector output



We've already seen one deep learning encoder!

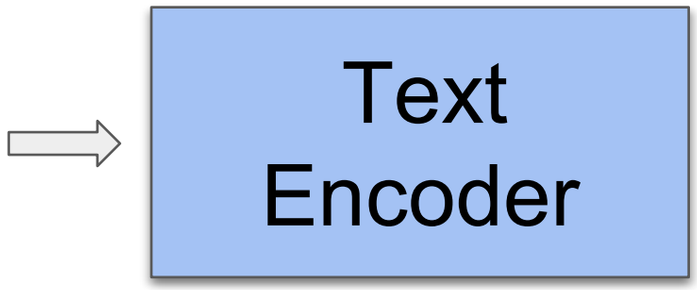


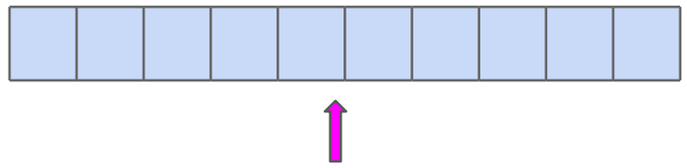
We could try another formulation that *learns* how to encode queries and documents



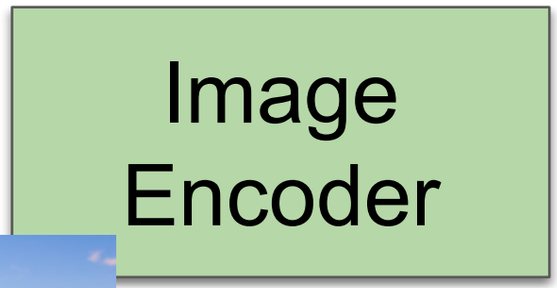
How to retrieve an image?
Encode into same space!

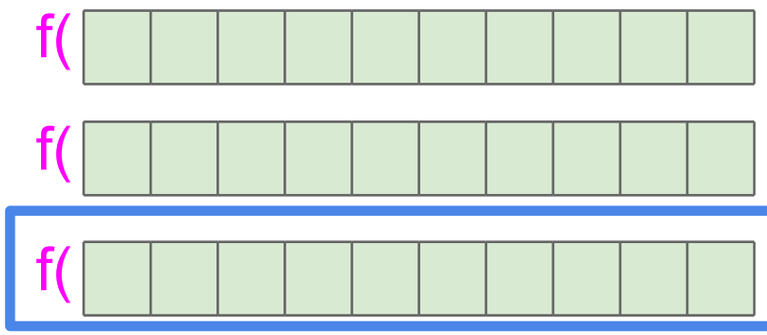
Michigan winter



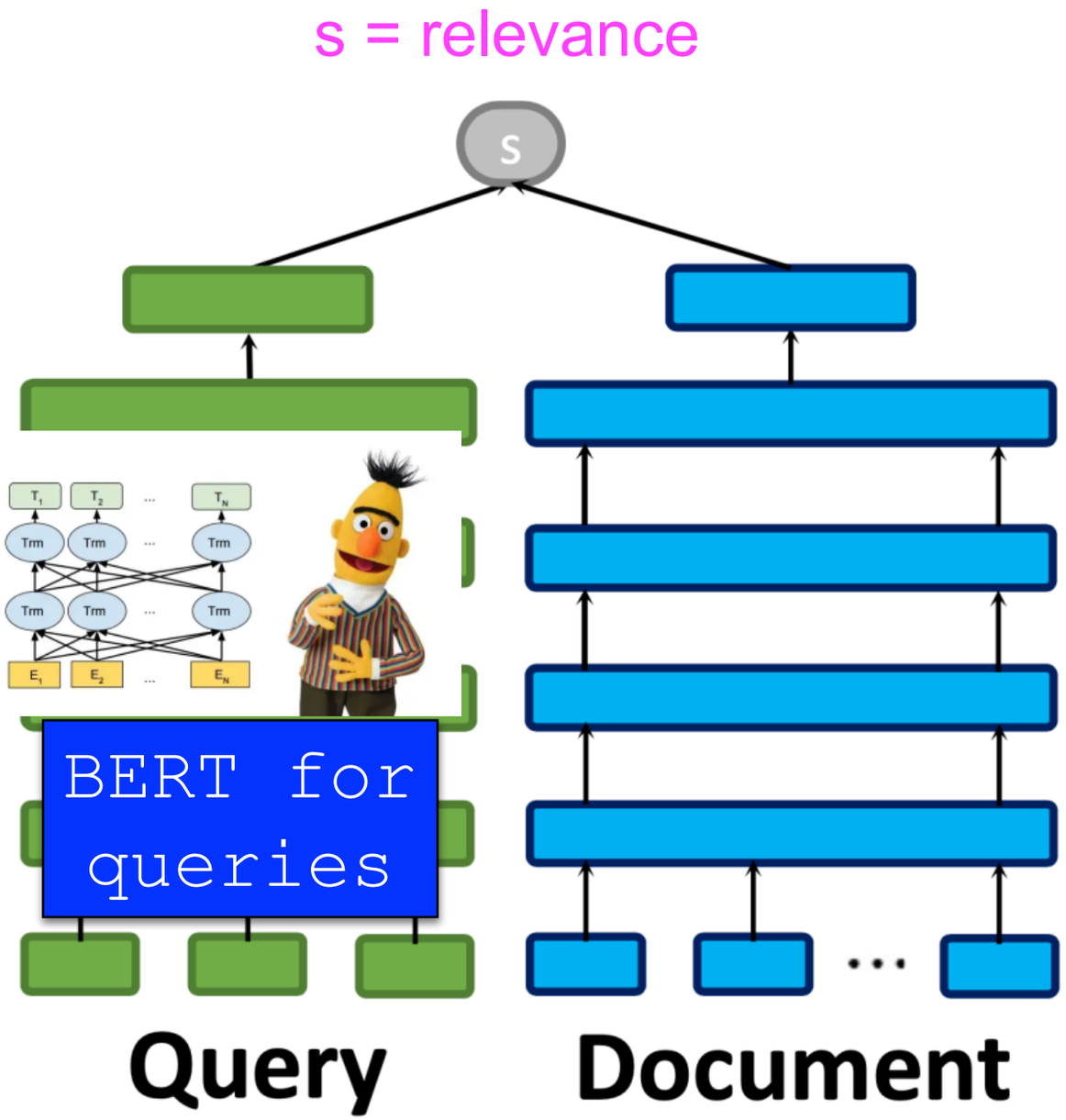


Compare these two representations to estimate the image's relevance

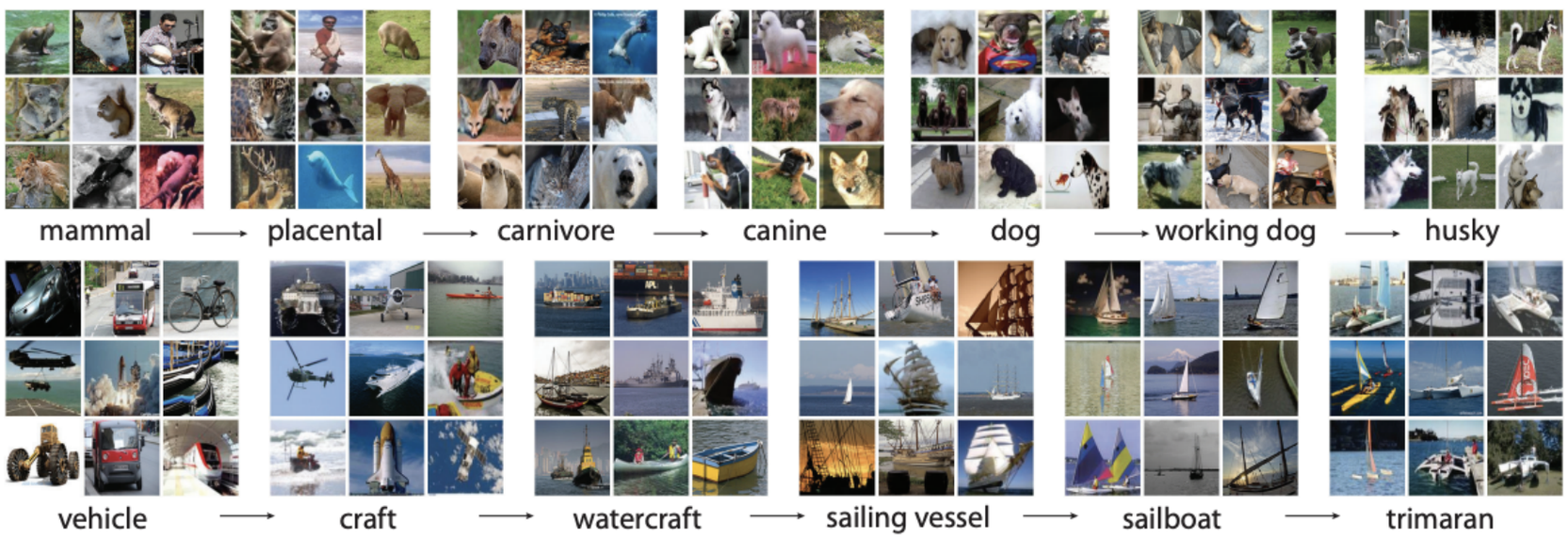




We could try another formulation that *learns* how to encode queries and documents



Where do the image encoders come from?



Critical view of Imagenet

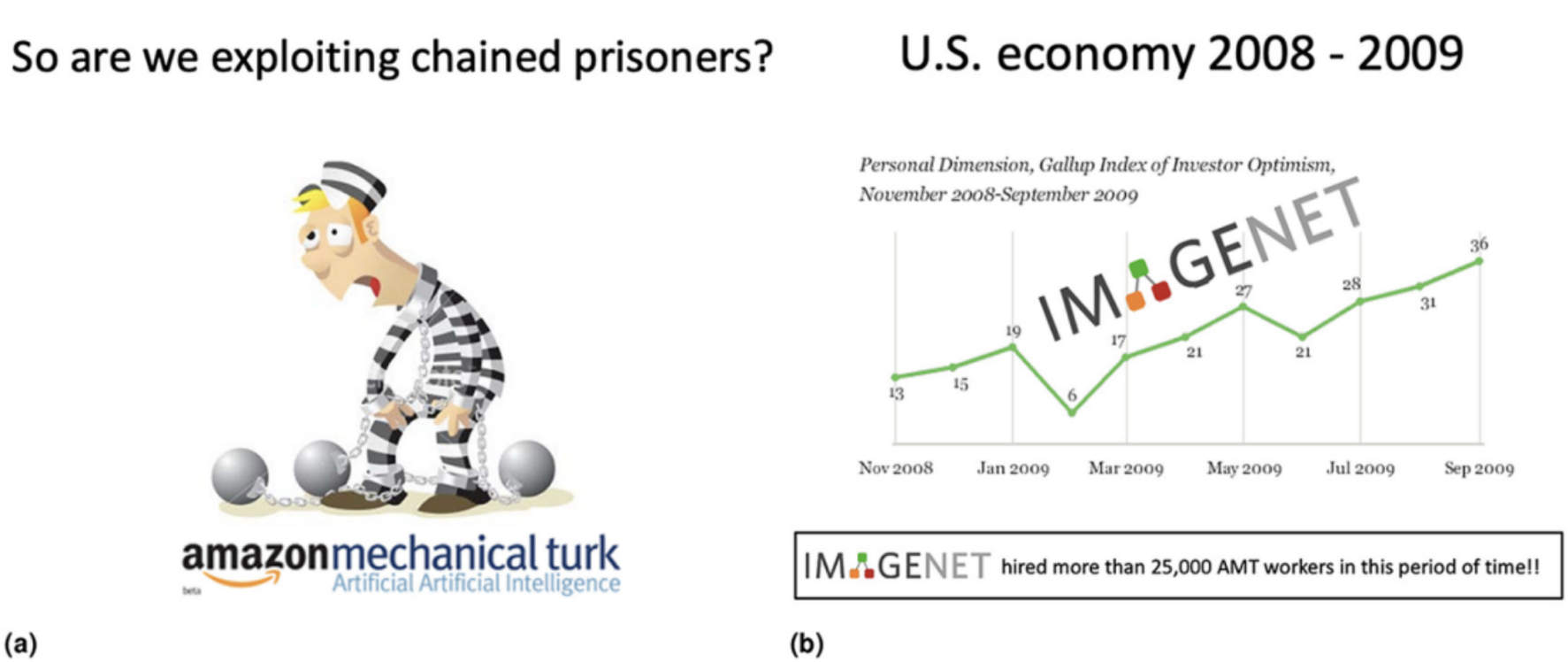
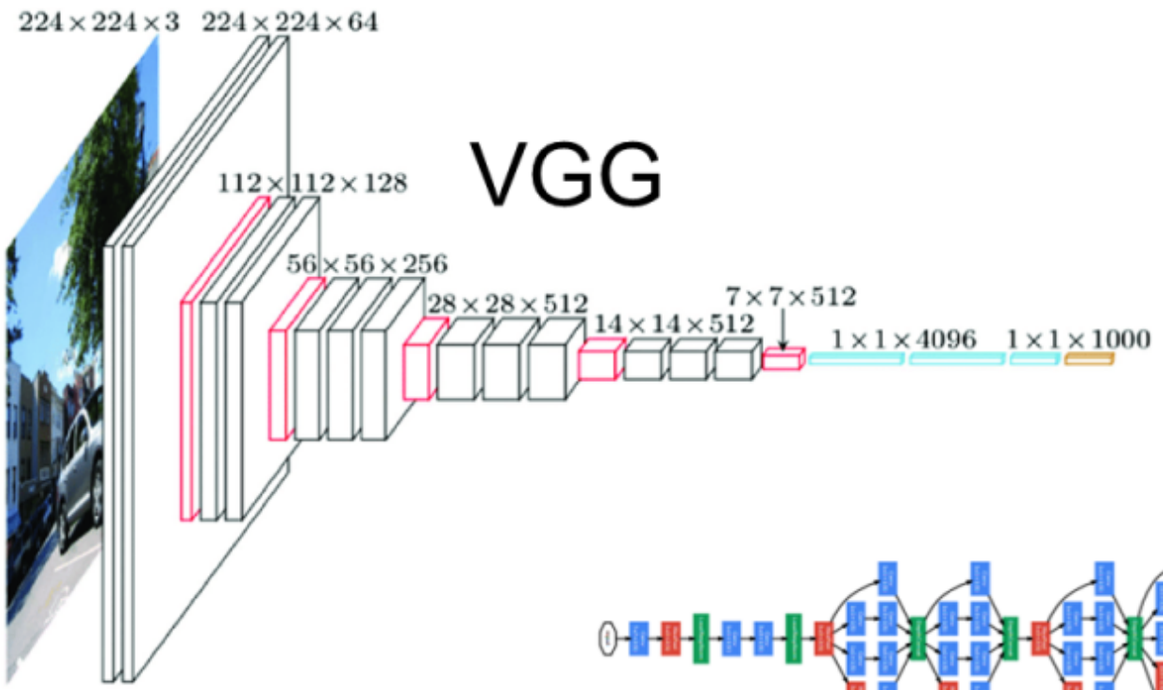


Figure 2. Slides from talk "ImageNet: Crowdsourcing, Benchmarking, & Other Cool Things" (Fei-Fei, 2010).

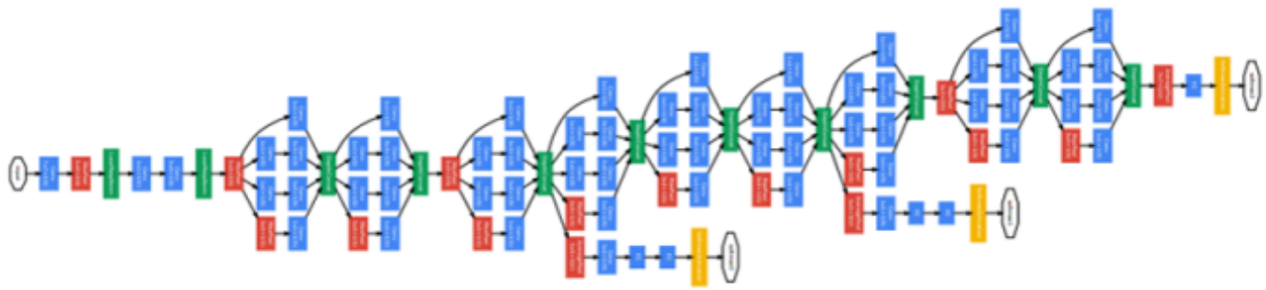
Image credit: Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211035955

Pretrain Image Classifiers on ImageNet (like BERT on text)



VGG

GoogLeNet



ResNet

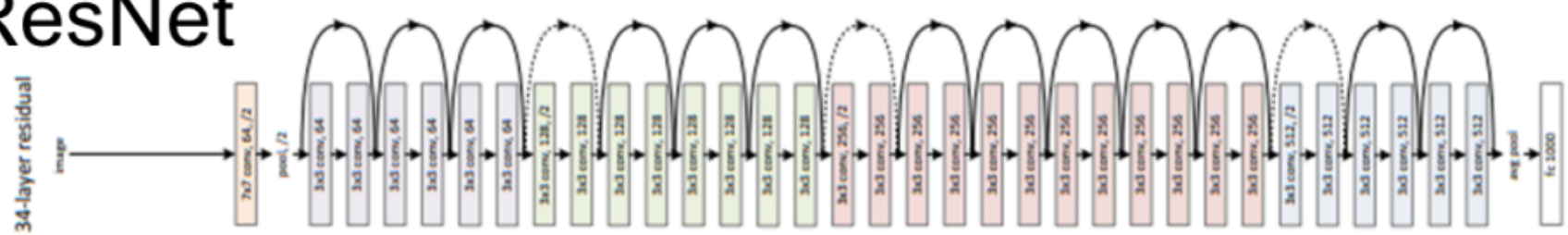
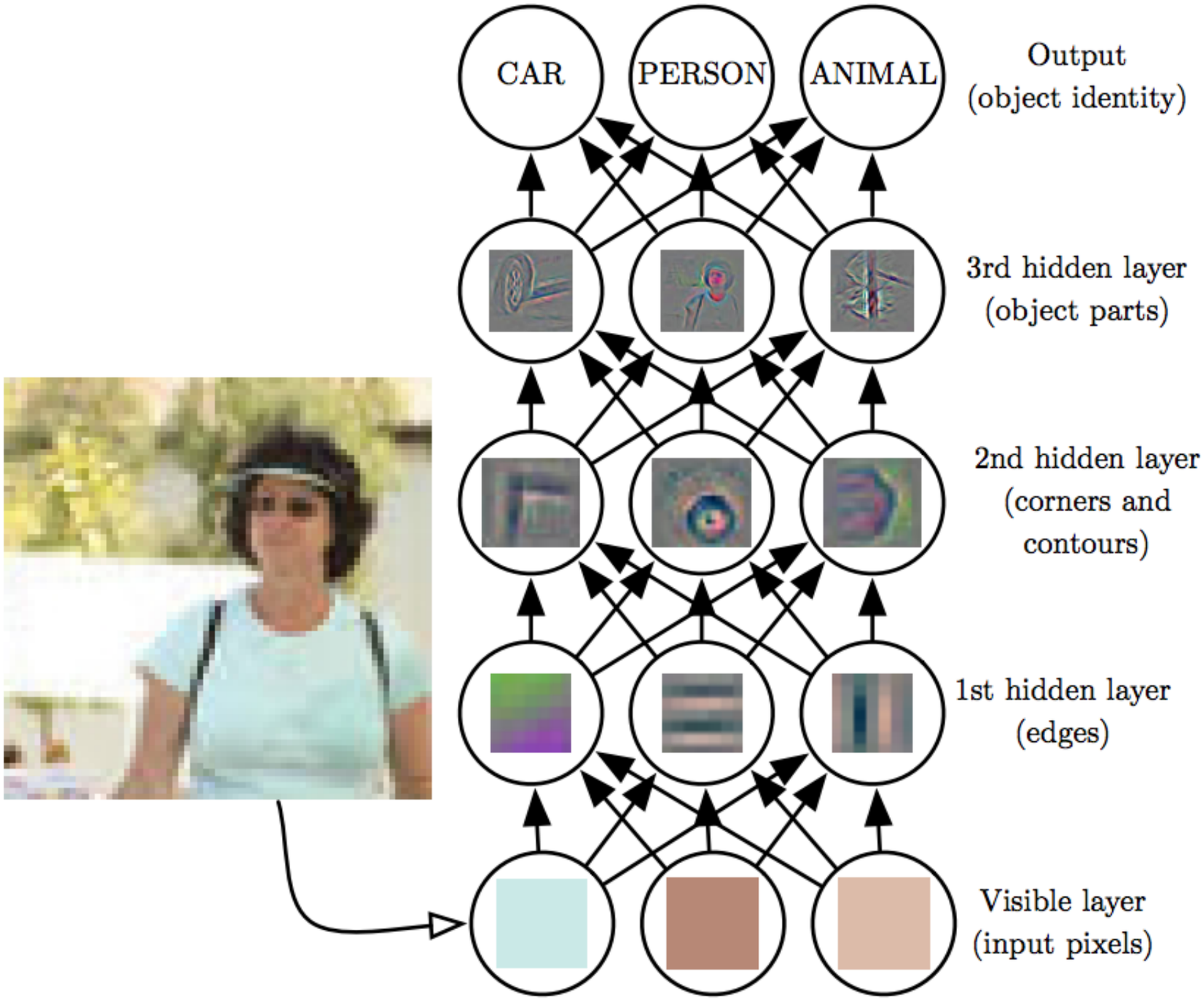
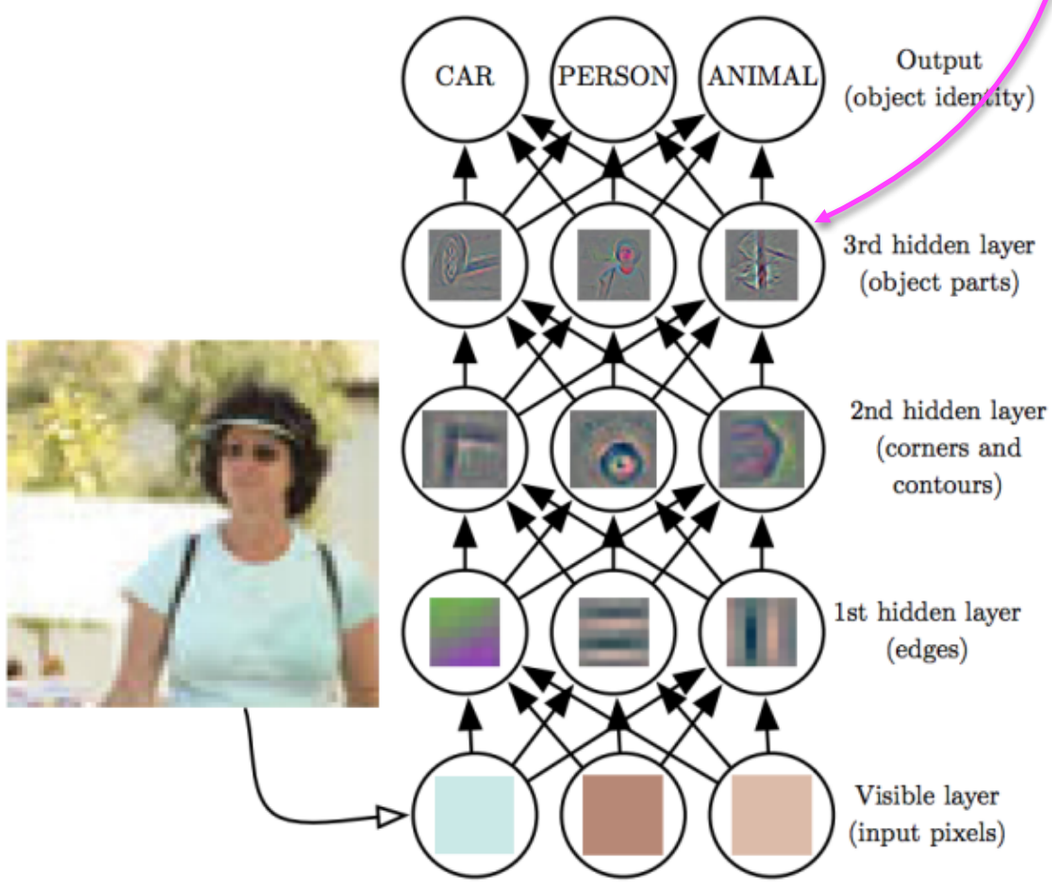


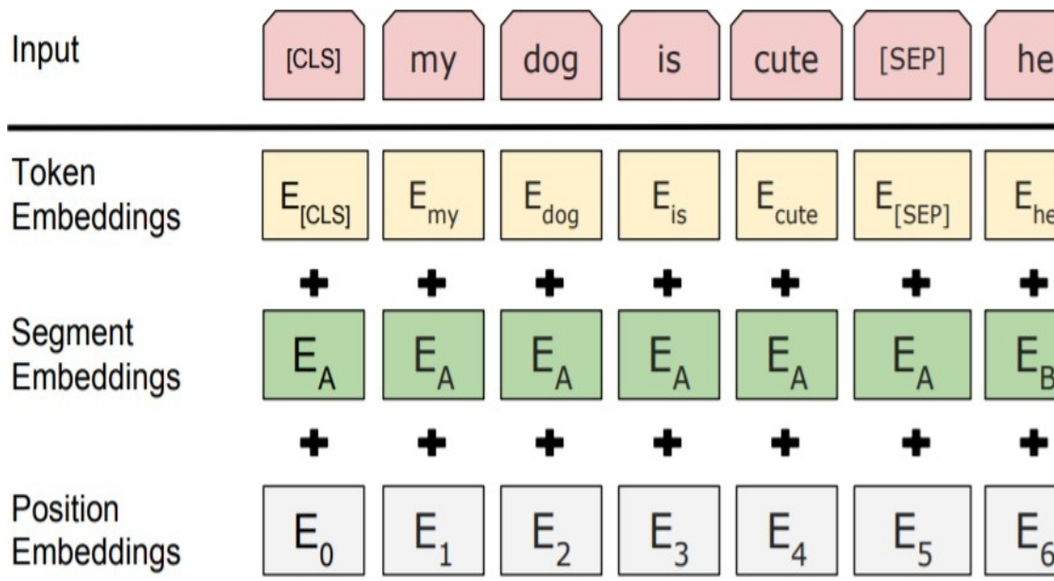
Image Encoders learn hierarchical representations of features



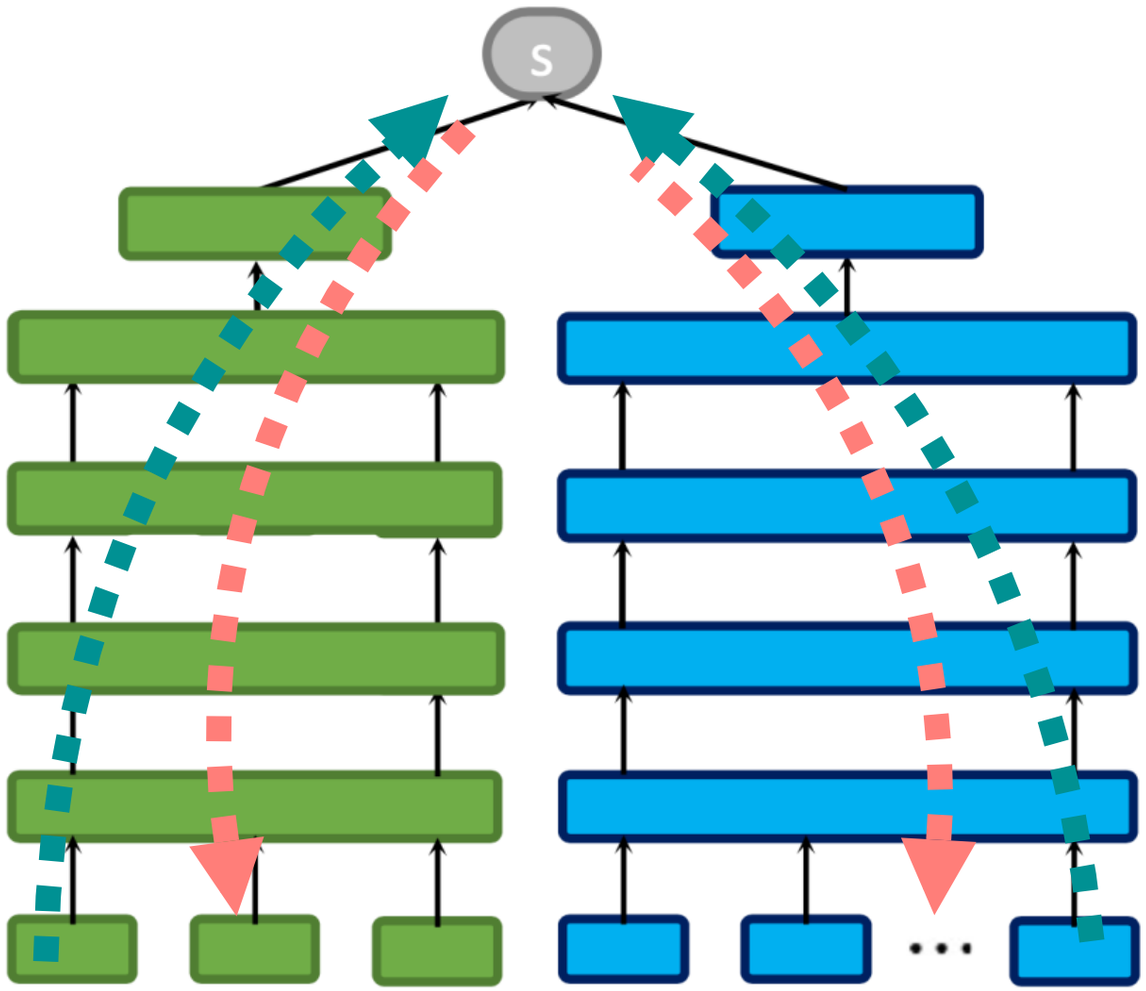
Our goal: have the two models use the same latent feature space!

Train models to have these representations become similar if the text is describing an image





By training to predict relevance, we update how the models learn to represent text and images



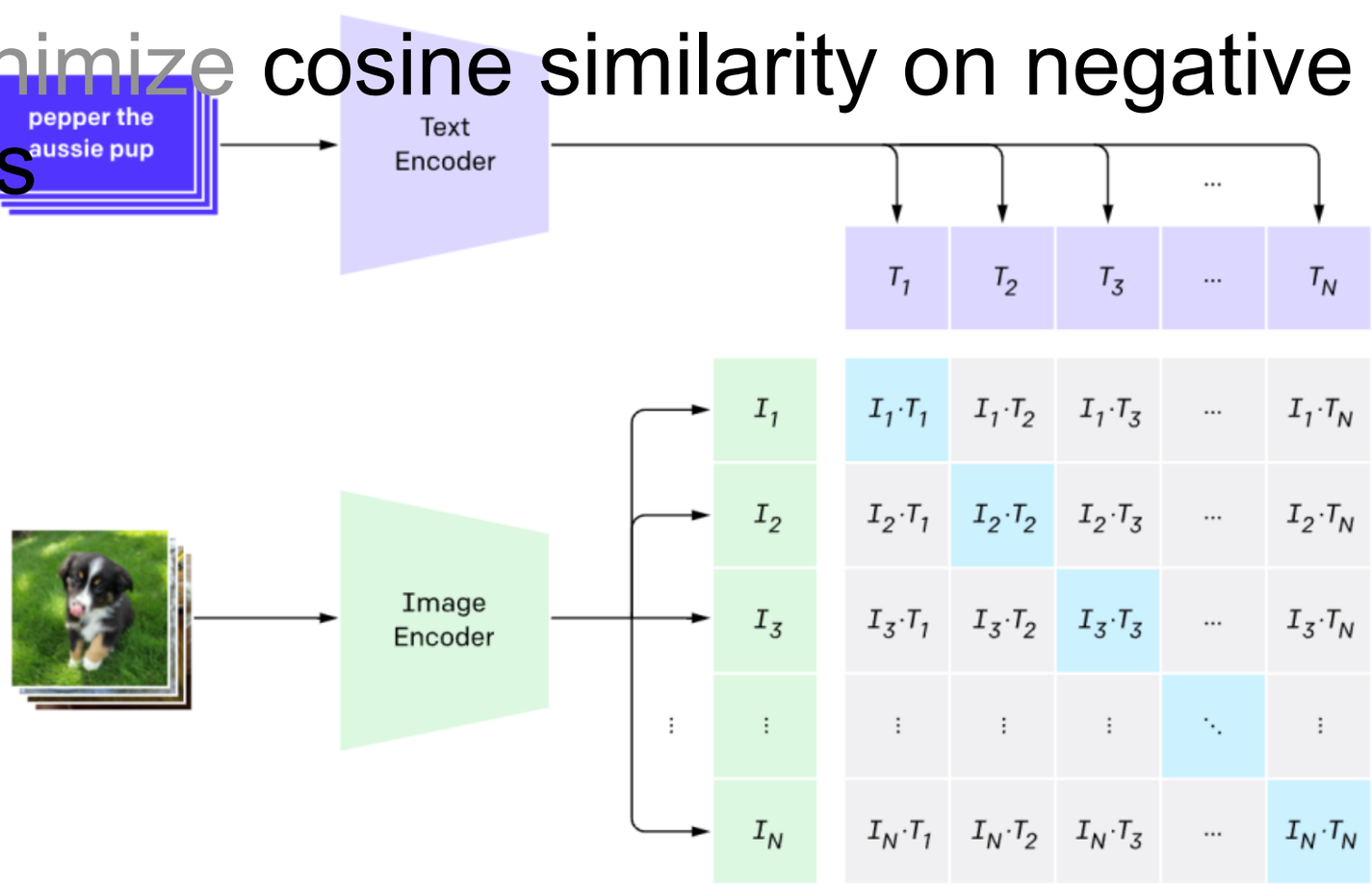
Predict

Update weights with Stochastic Gradient Descent

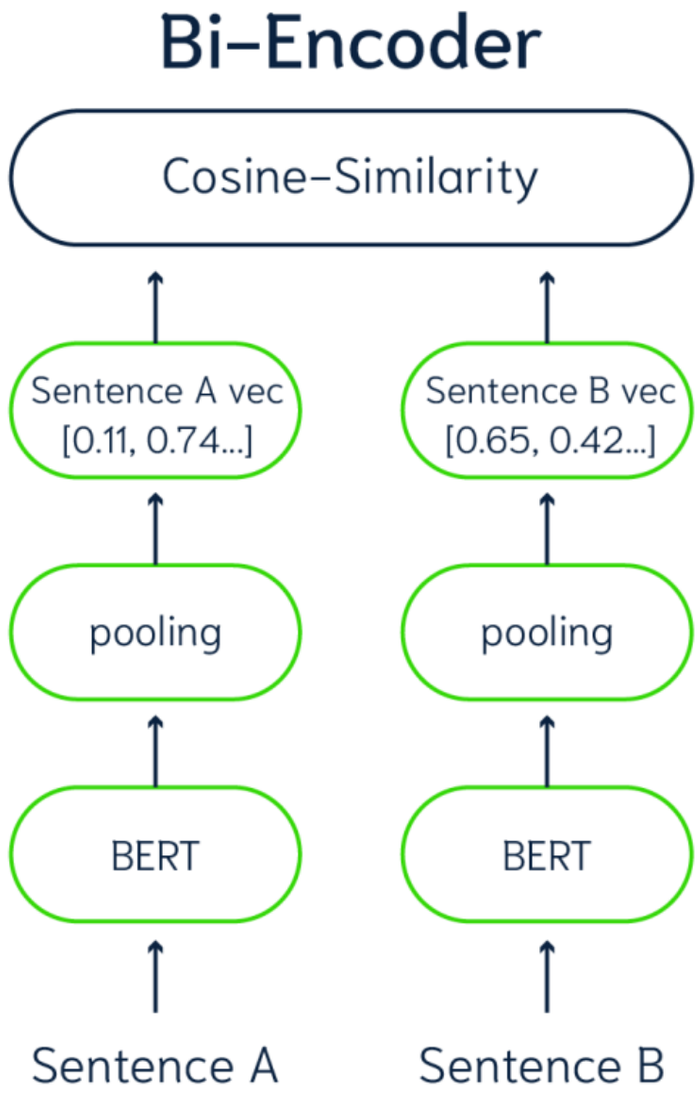
Query: “Scene of houses in front of mountains and next to a lake”

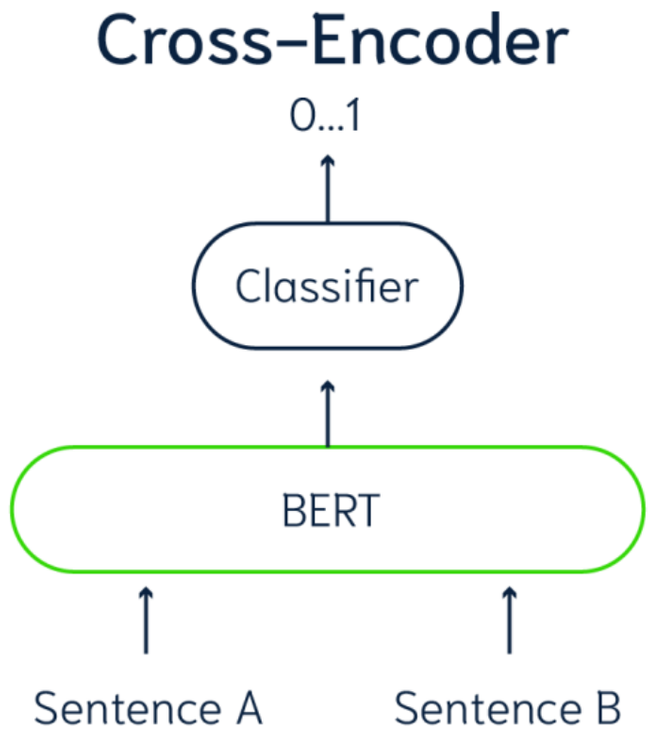
Training Multimodal Encoders

- Use contrastive loss
- We maximize cosine similarity between text encoding and image encodings on positive samples
- And minimize cosine similarity on negative samples

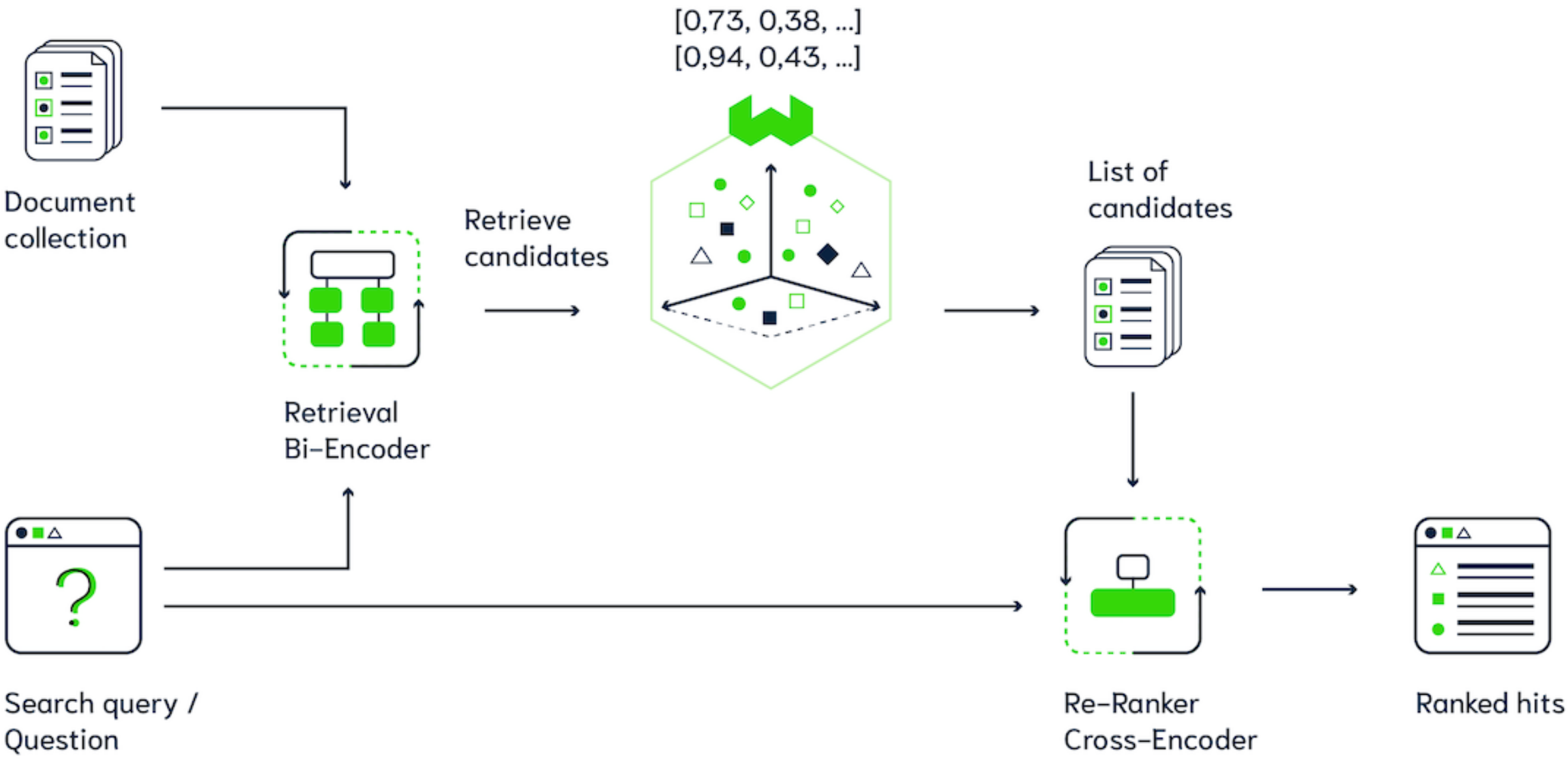


Quick aside: Bi-Encoder vs. Cross-Encoder



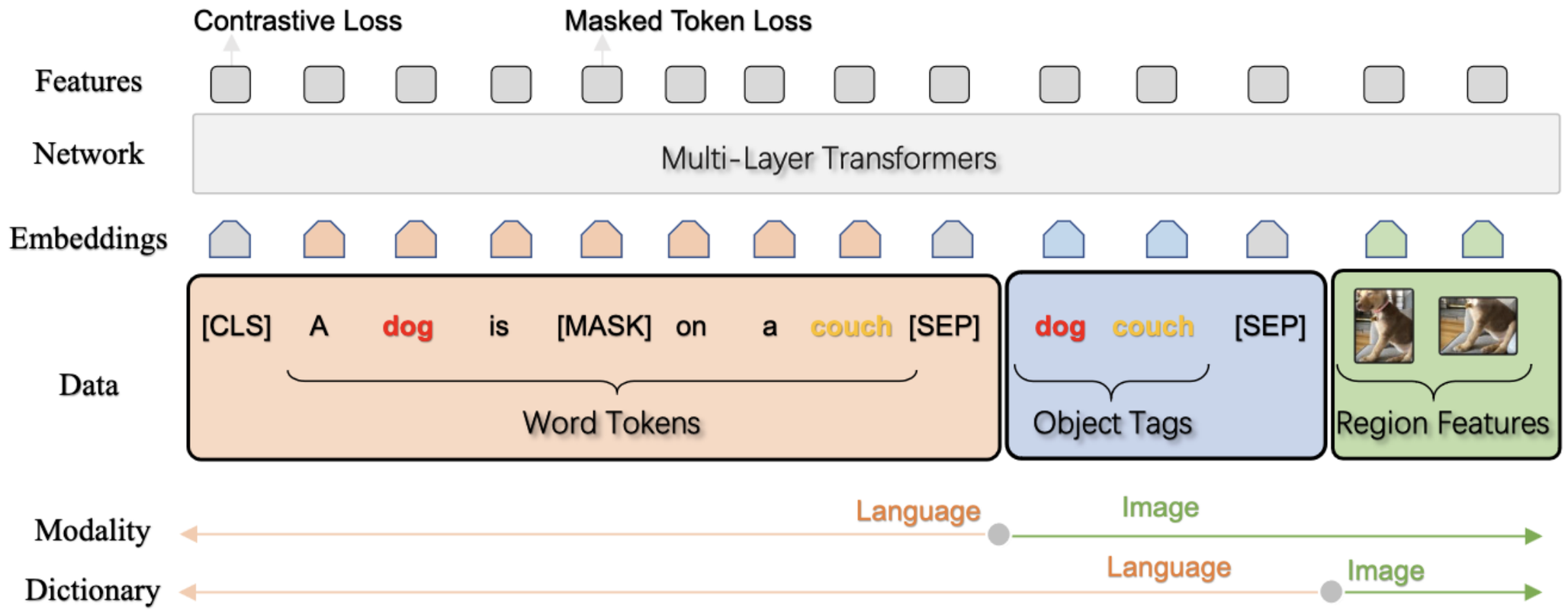


Both setups are are typically used in practice



What does a cross-encoder look like for text and images?

- It varies, but parts of or all of an image get encoded as input “tokens”



Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. Springer International Publishing, 2020.

Deep Learning for Multimodal IR

- Very new and fast changing
  - Vision Transformers (ViT): Use parts of images as “words”
  - Lots of multimodal transformers out there like CLIP
- High value in domains with limited data
- Expensive to do for all images
  - But can precompute image representations
  - Fast search among representations becomes bottleneck
- Potentially less biased (?)
  - No need to rely on meta data or text
  - But what are these models pretrained on??

Discussion Sections for Demos!

```
# Load the CLIP model, which knows how to embed both text *and* images!
model = SentenceTransformer('clip-ViT-B-32')
```

```
Downloading (…)d52eb/.gitattributes:   0%|          | 0.00/690 [00:00<?, ?B/s]
Downloading (…)LIPModel/config.json:   0%|          | 0.00/4.03k [00:00<?, ?B/s]
Downloading (…)CLIPModel/merges.txt:   0%|          | 0.00/525k [00:00<?, ?B/s]
Downloading (…)rocessor_config.json:   0%|          | 0.00/316 [00:00<?, ?B/s]
Downloading pytorch_model.bin:   0%|          | 0.00/605M [00:00<?, ?B/s]
Downloading (…)cial_tokens_map.json:   0%|          | 0.00/389 [00:00<?, ?B/s]
Downloading (…)okenizer_config.json:   0%|          | 0.00/604 [00:00<?, ?B/s]
Downloading (…)CLIPModel/vocab.json:   0%|          | 0.00/961k [00:00<?, ?B/s]
Downloading (…)859cad52eb/README.md:   0%|          | 0.00/1.88k [00:00<?, ?B/s]
Downloading (…)ce_transformers.json:   0%|          | 0.00/116 [00:00<?, ?B/s]
Downloading (…)cad52eb/modules.json:   0%|          | 0.00/122 [00:00<?, ?B/s]
```

```
def get_image(url):
    return Image.open(requests.get(url, stream=True).raw)
```

```
# It's a dog
dog_img_emb = model.encode(get_image('https://live.staticflickr.com/2109/2203669161_da6400d66f_b.jpg'))

# It's Ann Arbor
a2_url = 'https://upload.wikimedia.org/wikipedia/commons/thumb/9/91/Ann_Arbor_Skyline_2021.jpg/640px-Ann_Arbor_Skyline_2021.jpg'
a2_img_emb = model.encode(get_image(a2_url))
```

```
#Encode text descriptions
text_emb = model.encode(['A dog standing on a sidewalk', 'Late lunch with friends', 'A picture of Ann Arbor at night'])
```

```
# Rank the dog picture according to our queries
cos_scores = util.cos_sim(dog_img_emb, text_emb)
print(cos_scores)
```

```
tensor([[26.3228, 20.1940, 18.3339]])
```

```
# Rank the city picture according to our queries
cos_scores = util.cos_sim(a2_img_emb, text_emb)
print(cos_scores)
```

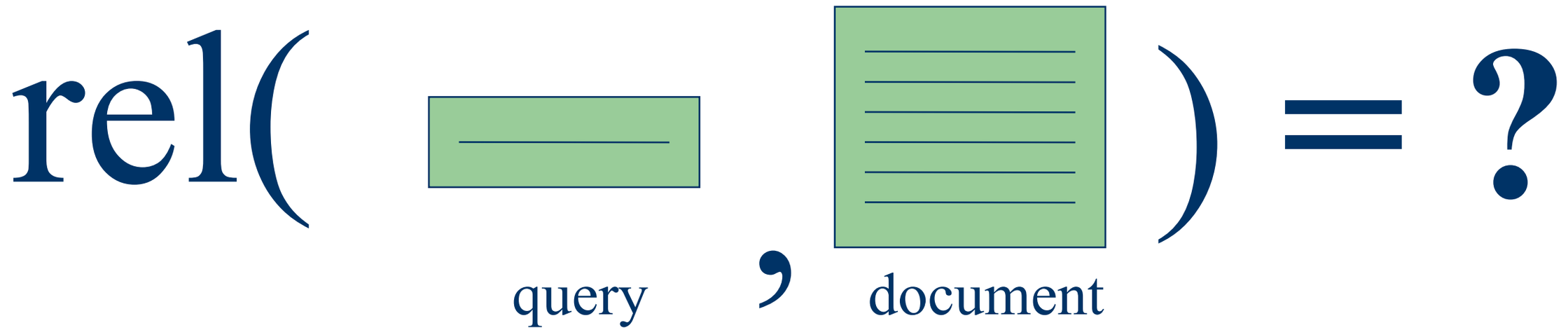
```
tensor([[0.1900, 0.1970, 0.2163]])
```

But you will cover a lot more, including discussions on fairness!

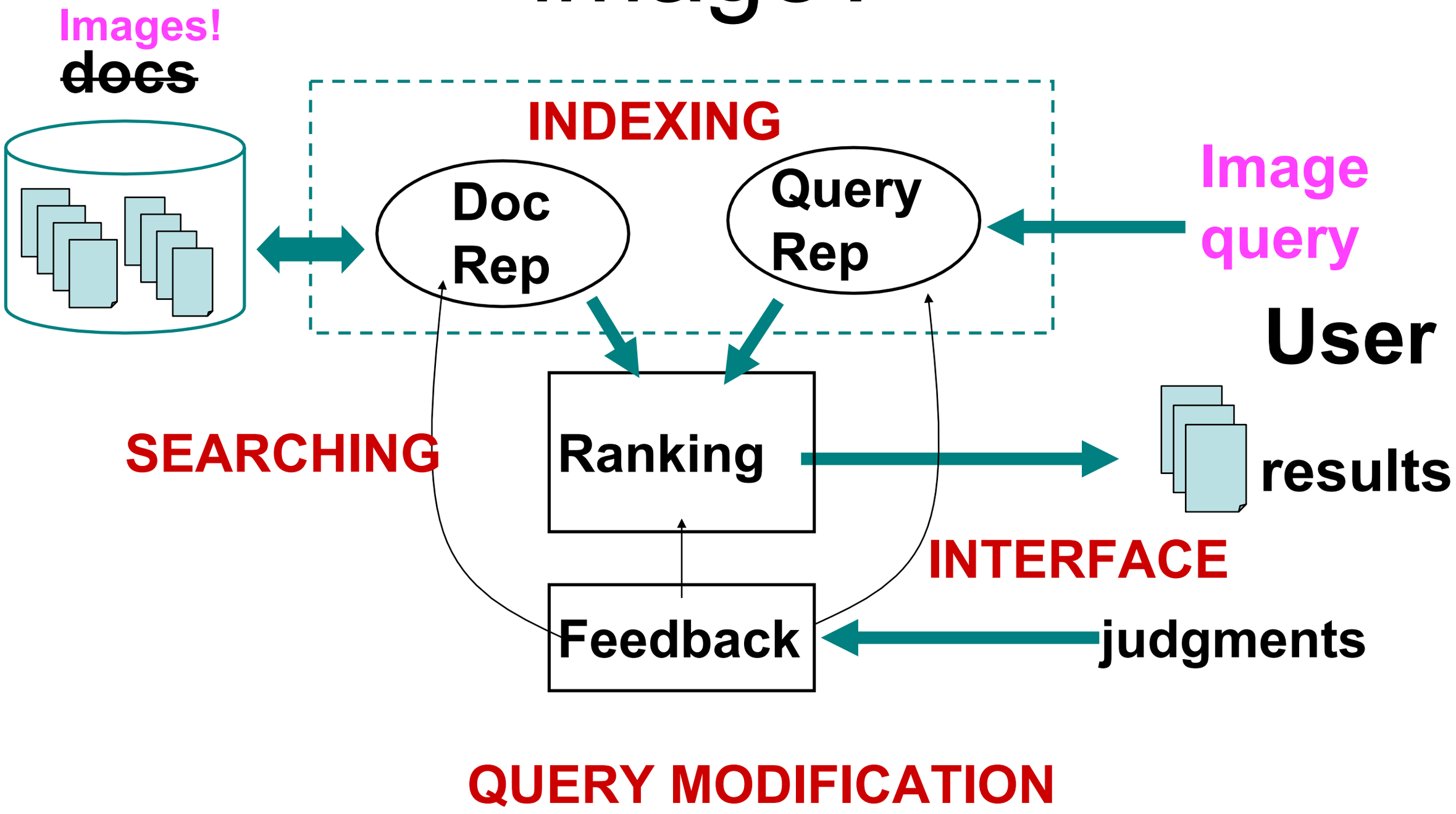
Content-Based Image Retrieval (CBIR)

or Image-to-Image Search

What we know: Text retrieval



What if the query is an image?



The Problem: Image Similarity

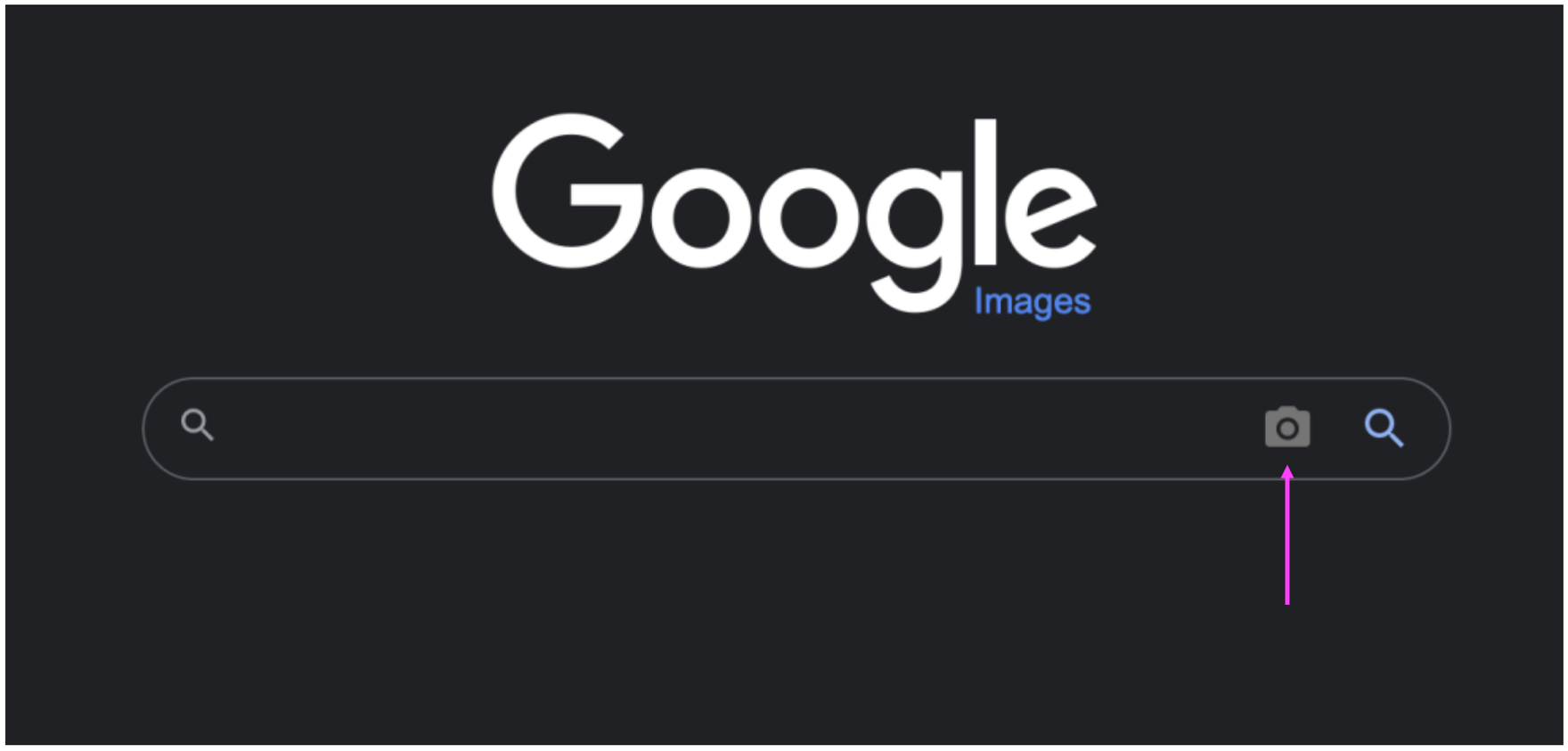
sim( , ) = ?

Where does this problem arise in computer vision?

- Image Classification
- Image Retrieval
- Image Segmentation

Content-Based Image Retrieval (CBIR)

- Retrieves images on the basis of automatically-derived features such as color, texture and shape.
  - Given: a reference database of unlabeled images
  - No provided textual descriptions or metadata



Content-based image retrieval (CBIR)

- Challenges
  - To be fast (efficient indexing structures)
  - To be accurate (rich image descriptions)
  - Avoiding tedious manual adaptations specific to a task

How do we compare images?

?

?

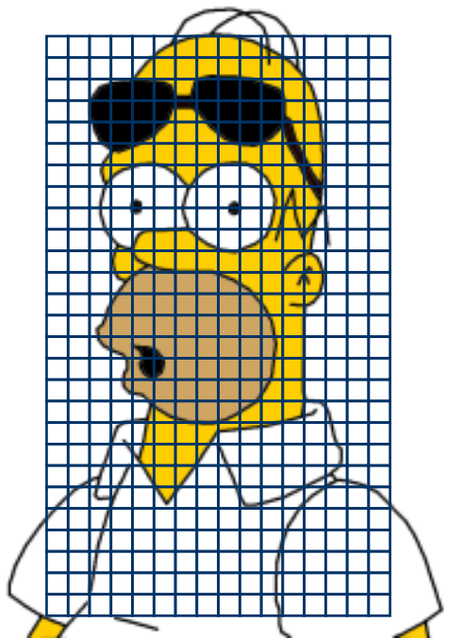
?

Retrieval

Jeremy S. De Bonet, Paul Viola (1997). Structure Driven Image Database Retrieval. Neural Information Processing 10 (1997).

How is an image represented?

How is an image represented?

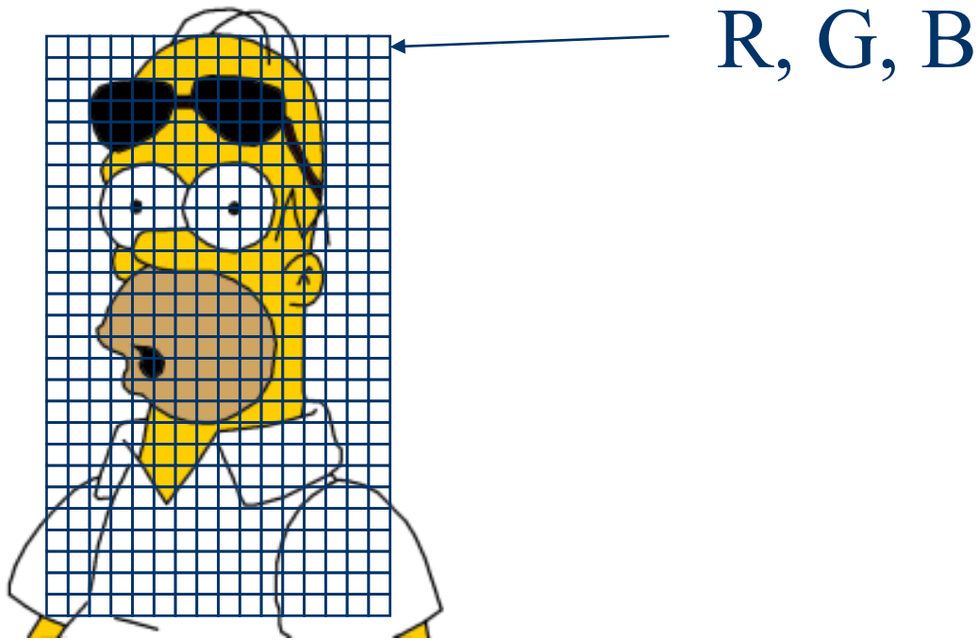


- images are made up of pixels
- for a color image, each pixel corresponds to an RGB value (i.e. three numbers)

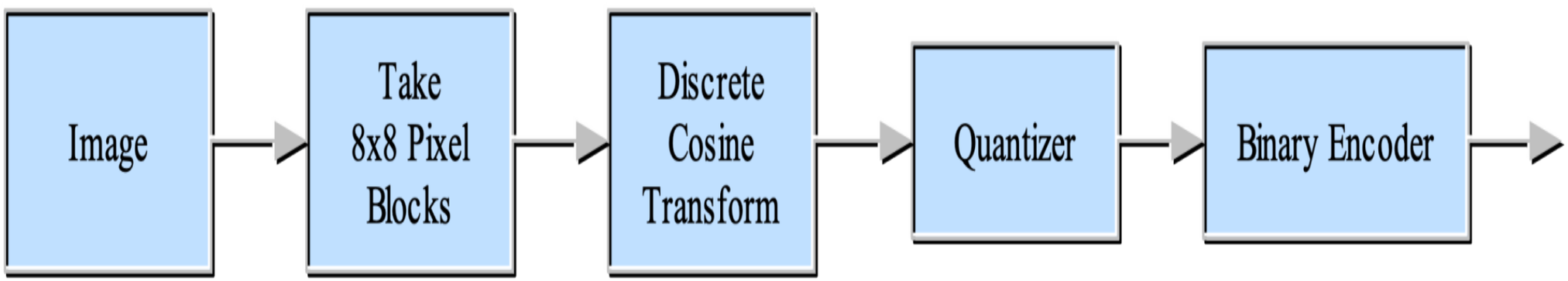
Image file formats

- BitMaP
- JPEG
- TIFF
- Gif
- Png
- …

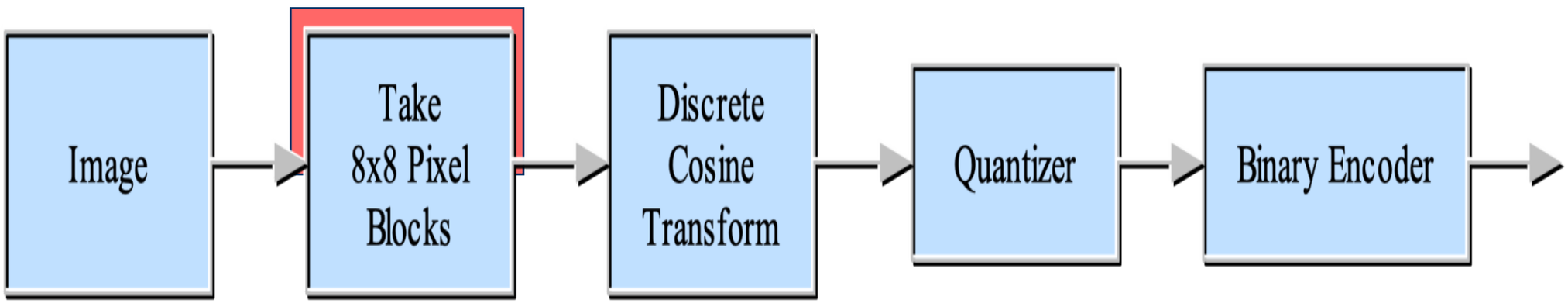
Bitmap

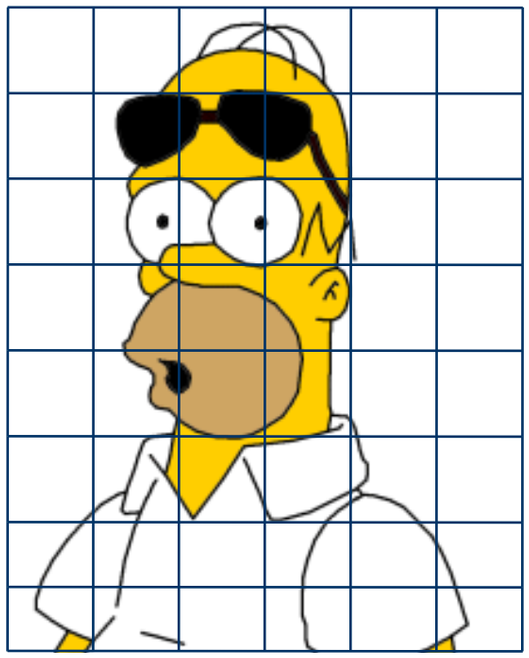


JPEG Compression Process

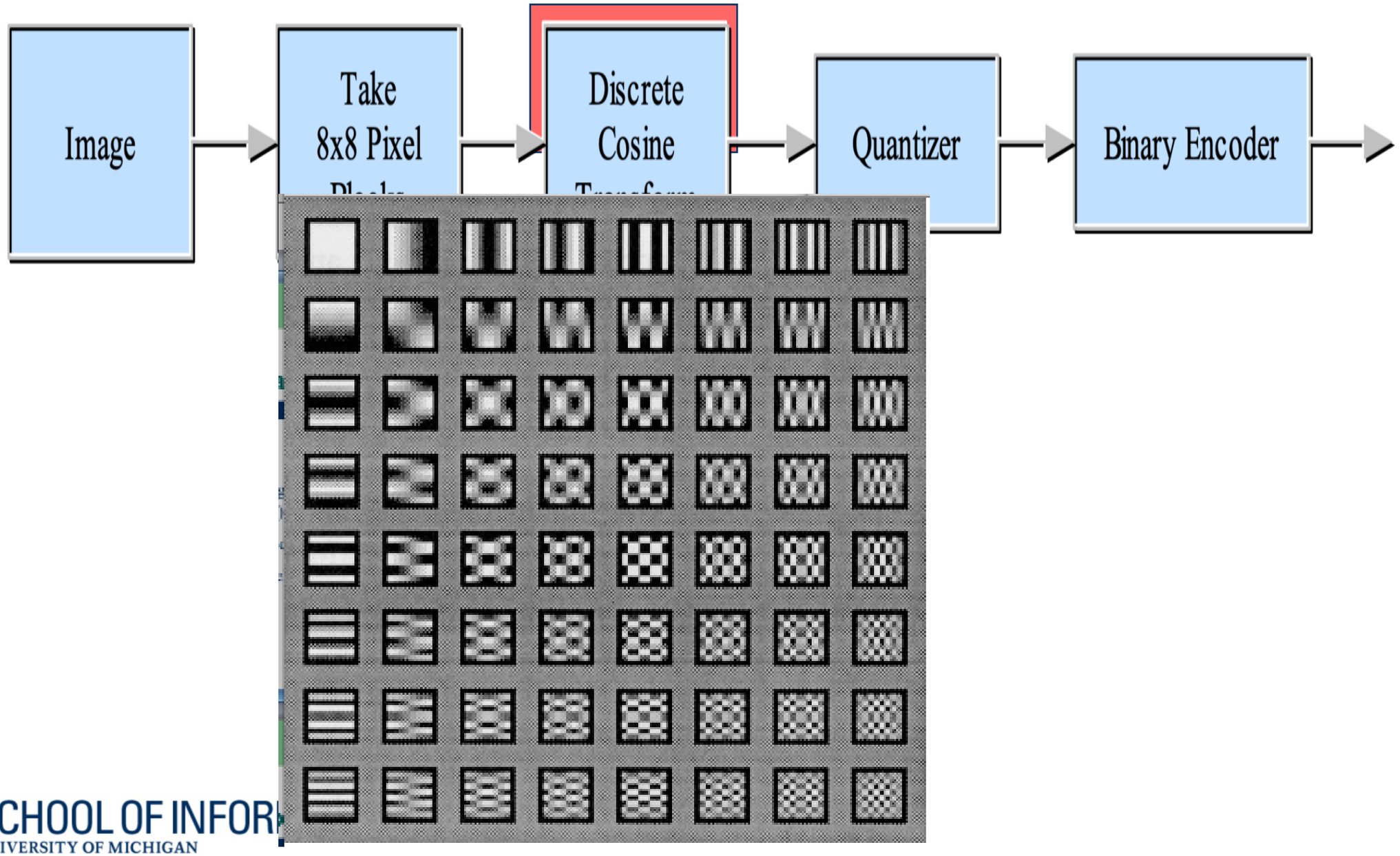


JPEG Compression Process

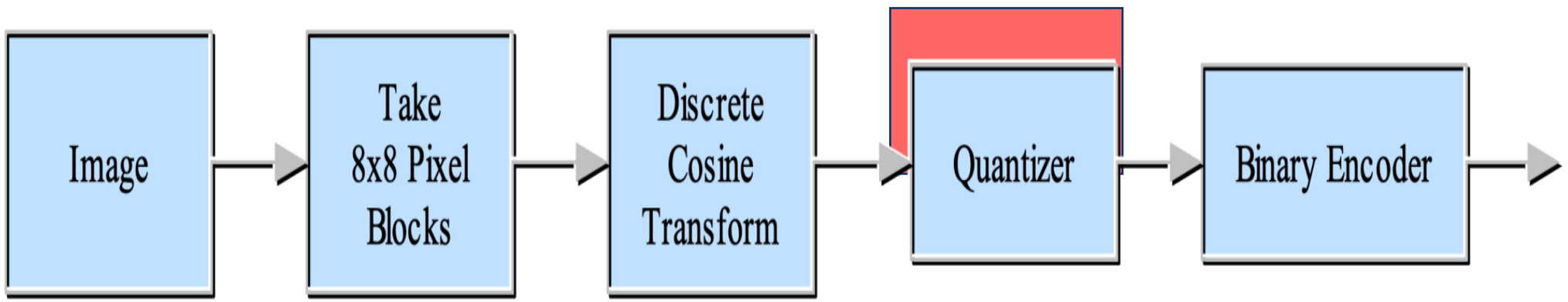




JPEG Compression Process



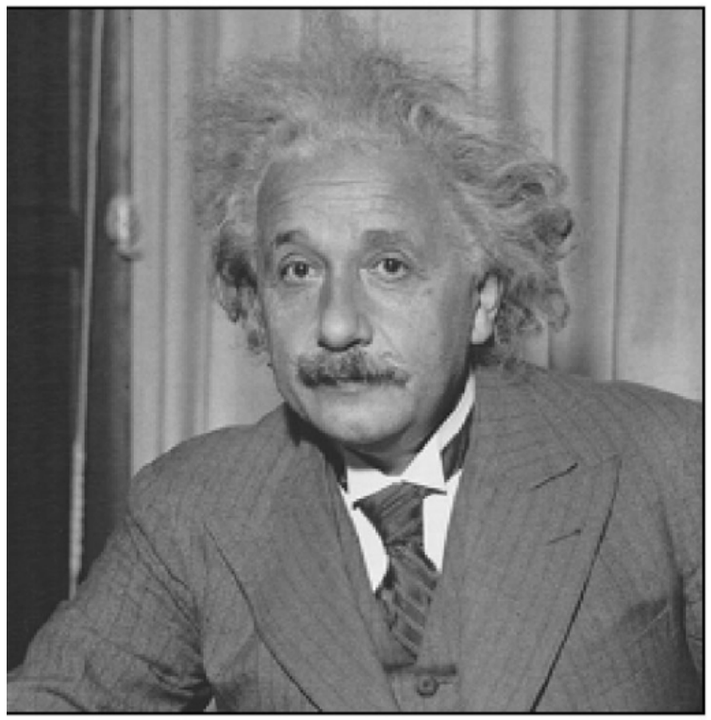
JPEG Compression Process



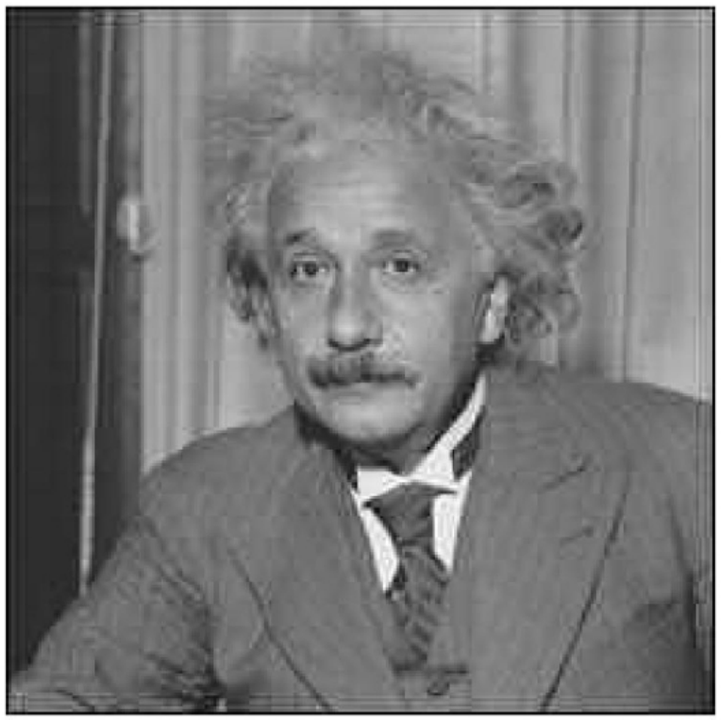
Quantizer: Weights the various spectral coefficients according to their importance, with respect to the human visual system.

JPEG Compression

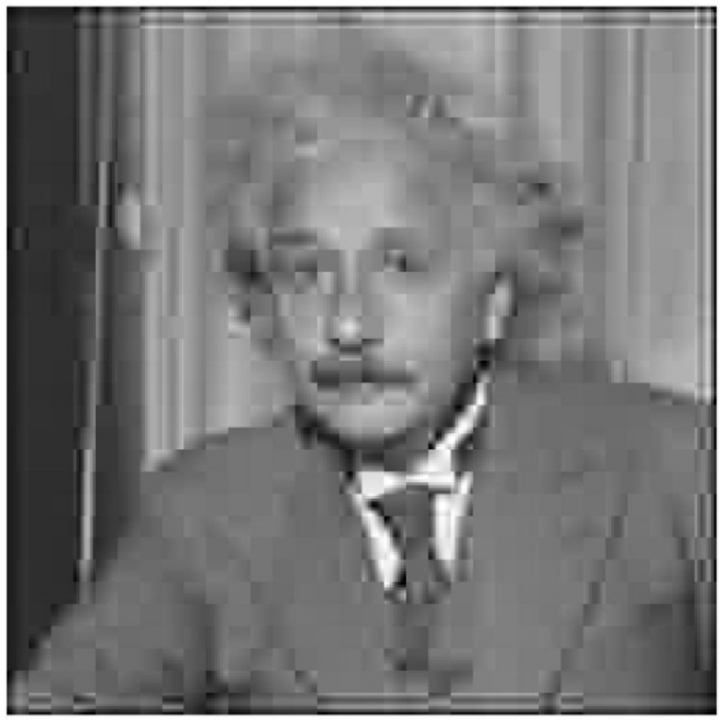
JPEG Image with no Lossy Compression



JPEG Image with Lossy Compression Ratio of ~3



JPEG Image with Lossy Compression Ratio of ~9

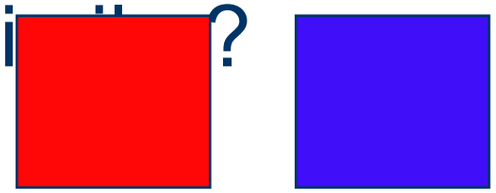


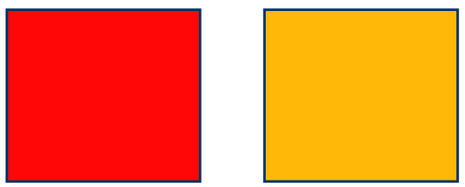
How to describe images?

- Color
- Texture
- ???

Color

Which is more similar?





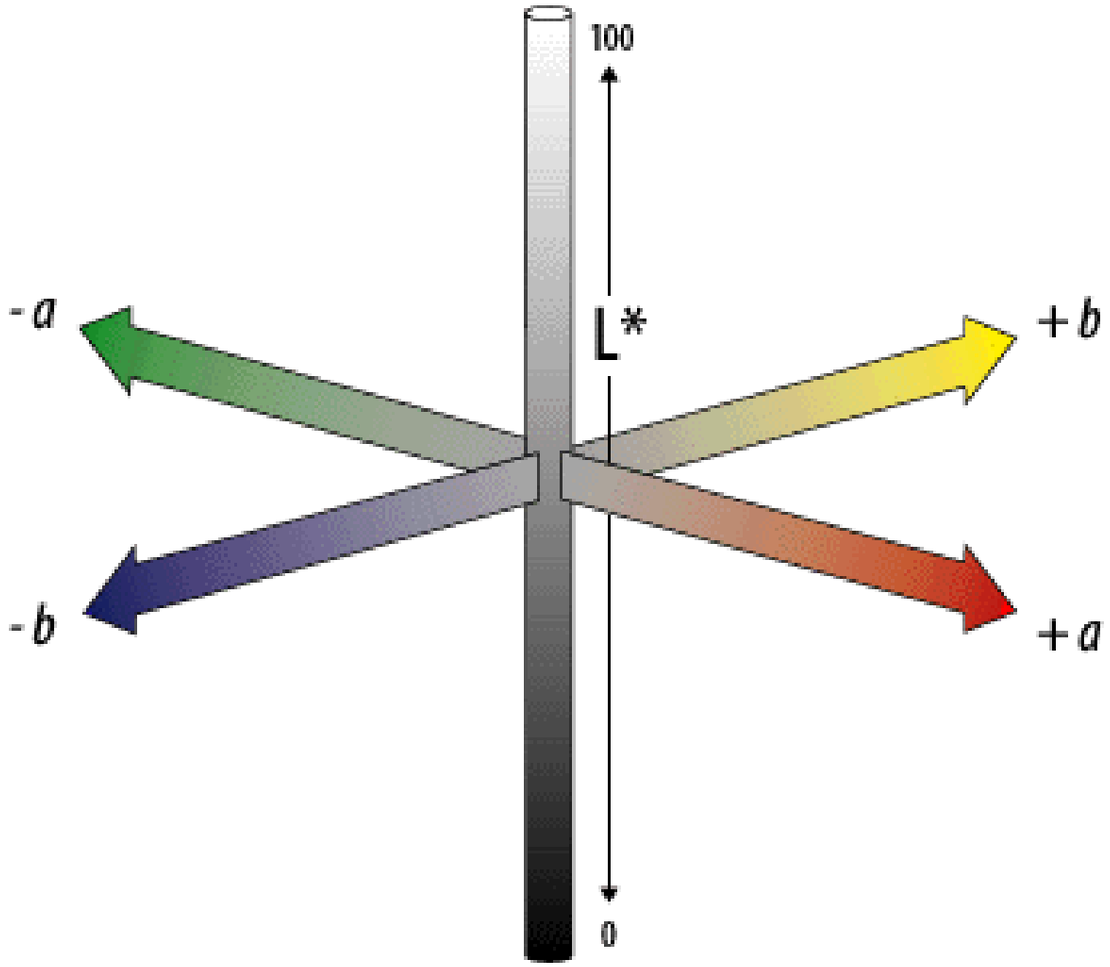
L*a*b* was designed to be uniform in that perceptual “closeness” corresponds to Euclidean distance in the space.

L*a*b*

L – lightness (white to black)

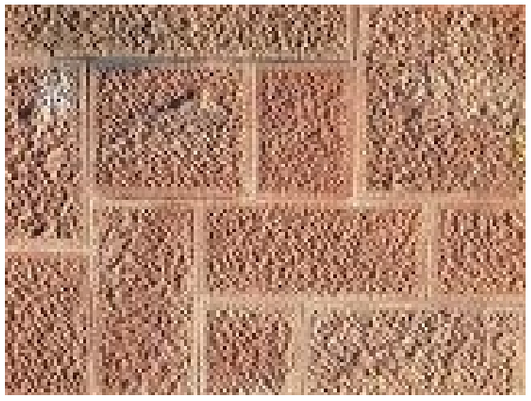
a – red-greeness

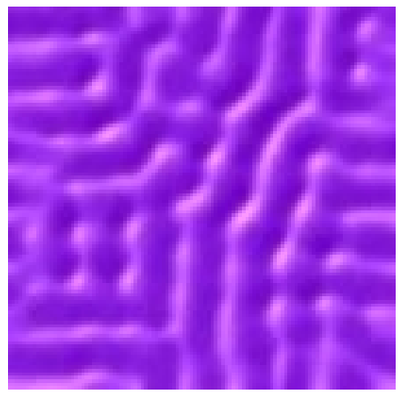
b – yellowness-blueness



Texture

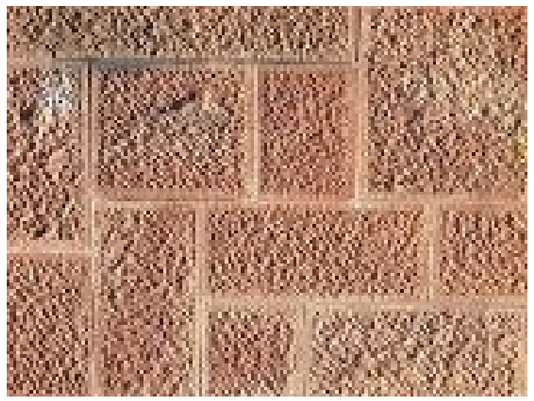
How is texture different than color?

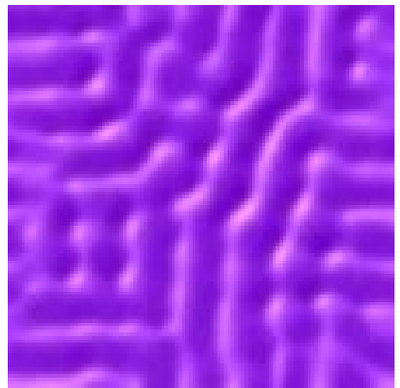




Texture

- Texture is not pointwise like color
- Texture involves a local neighborhood

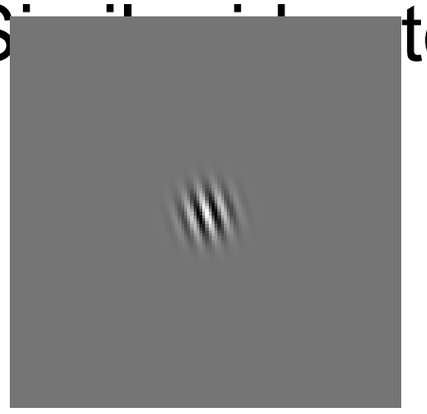




How can we capture texture?

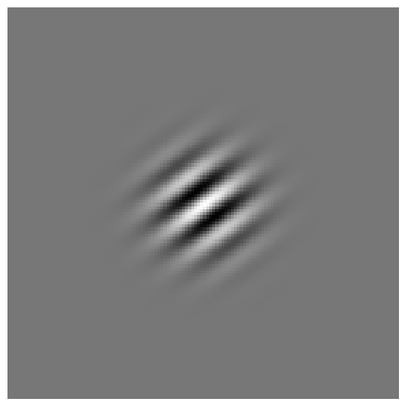
Gabor Filters

- Gabor filters are Gaussians modulated by sinusoids
- They can be tuned in both the scale (size) and the orientation
- A filter is applied to a region and is characterized by some feature of the energy distribution (often mean and standard deviation)
- Similar idea to wavelets (Gabor wavelet)!



Scale: 3 at 72°

Scale: 4 at 108°

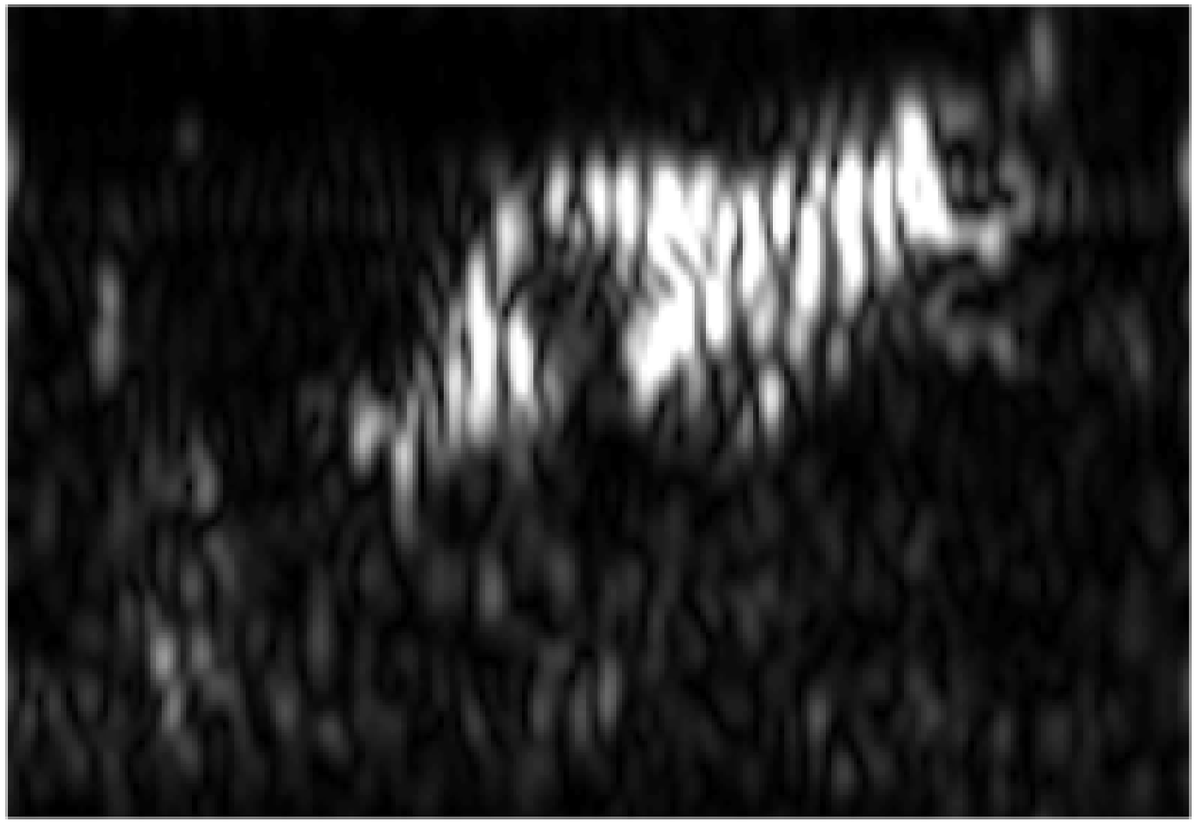


Scale: 5 at 144°

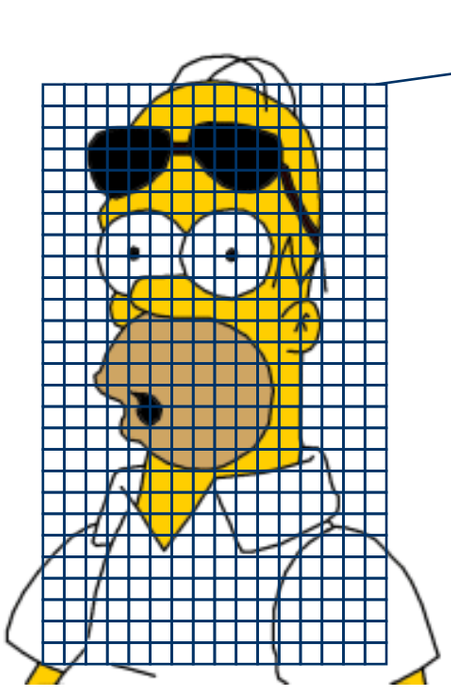
Gabor filters

What would the response look like to a vertical filter

Gabor filters



Features

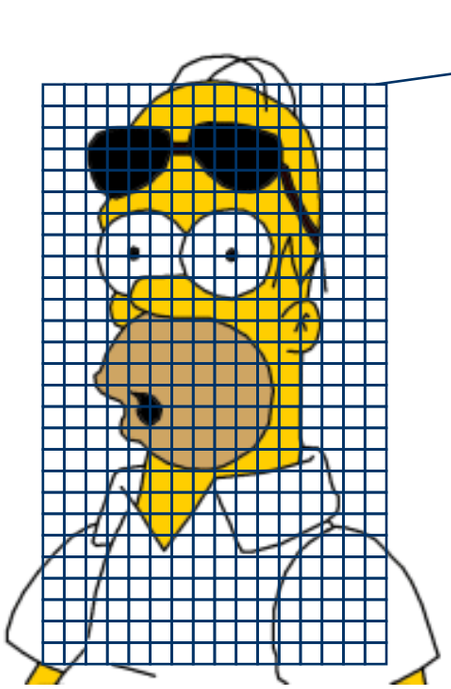


For each pixel:

- set of color features
- set of texture features (i.e. responses to different filters)
- …

Any problems with this approach?

Features



For each pixel:

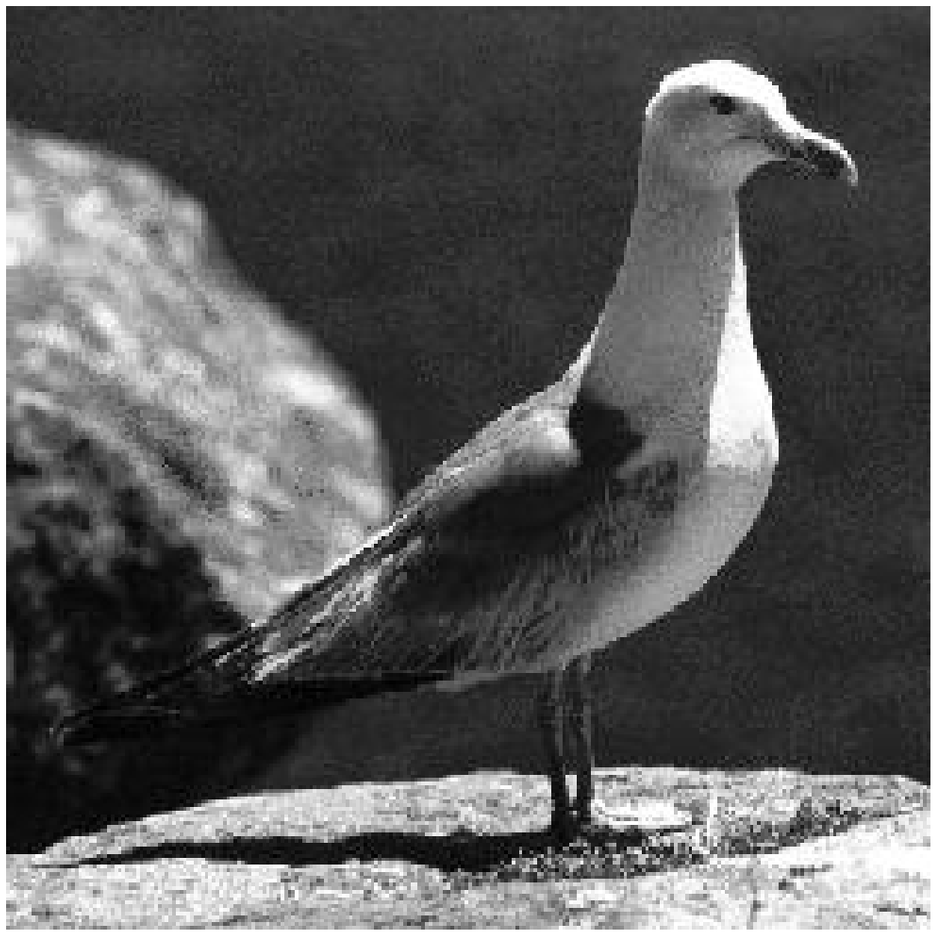
- set of color features
- set of texture features (i.e. responses to different filters)
- …

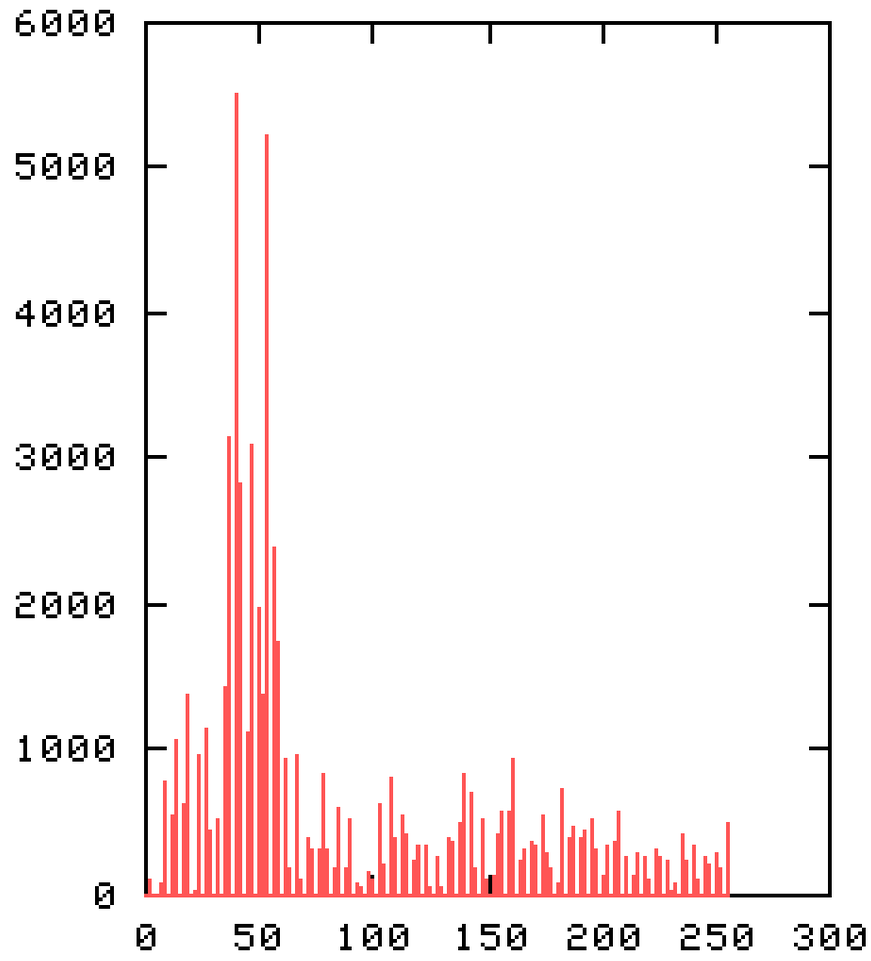
- Lots of features!
- Extremely sparse
- Features are position dependent

One approach: histograms

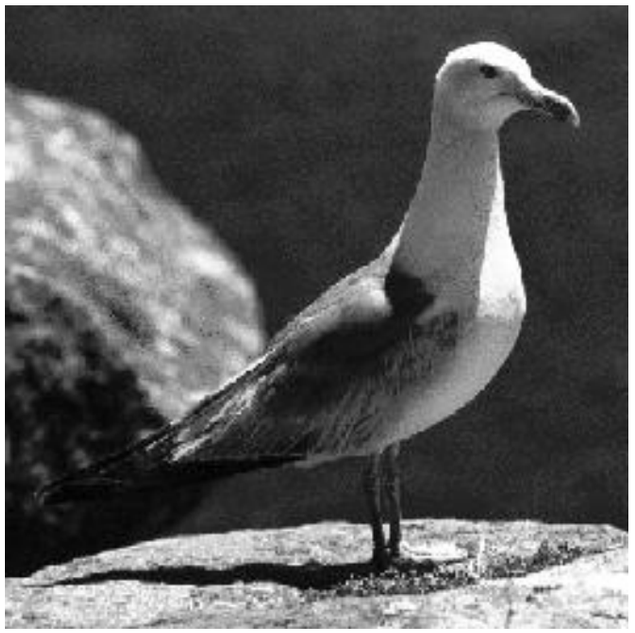
- Examine the distribution of features, rather than the features themselves
- General purpose (i.e. any distribution of features)
- Resilient to variations (shadowing, changes in illumination, shading, etc.)
- Can use previous work in statistics, etc.

Histogram Example

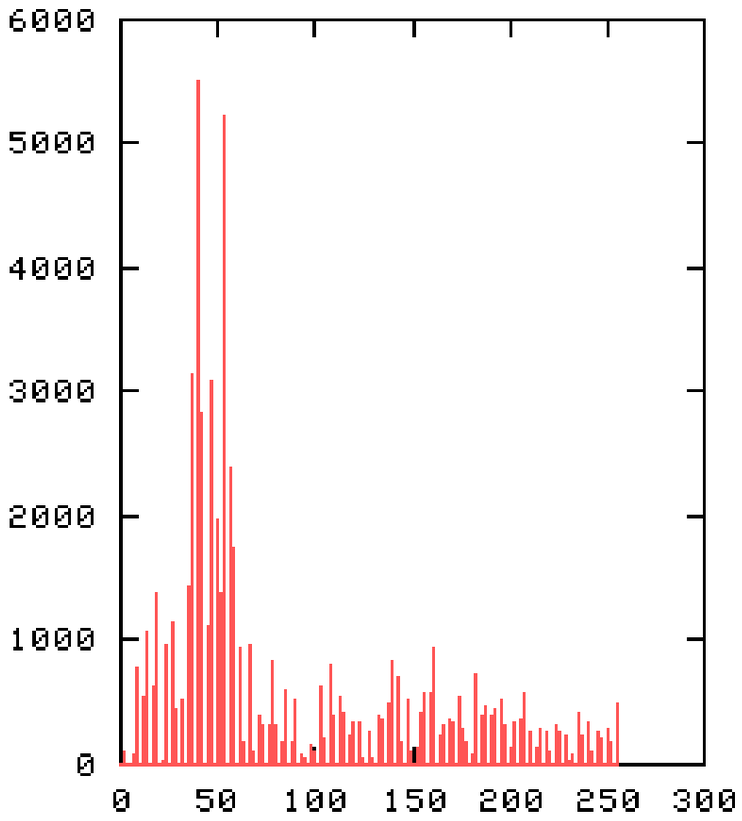




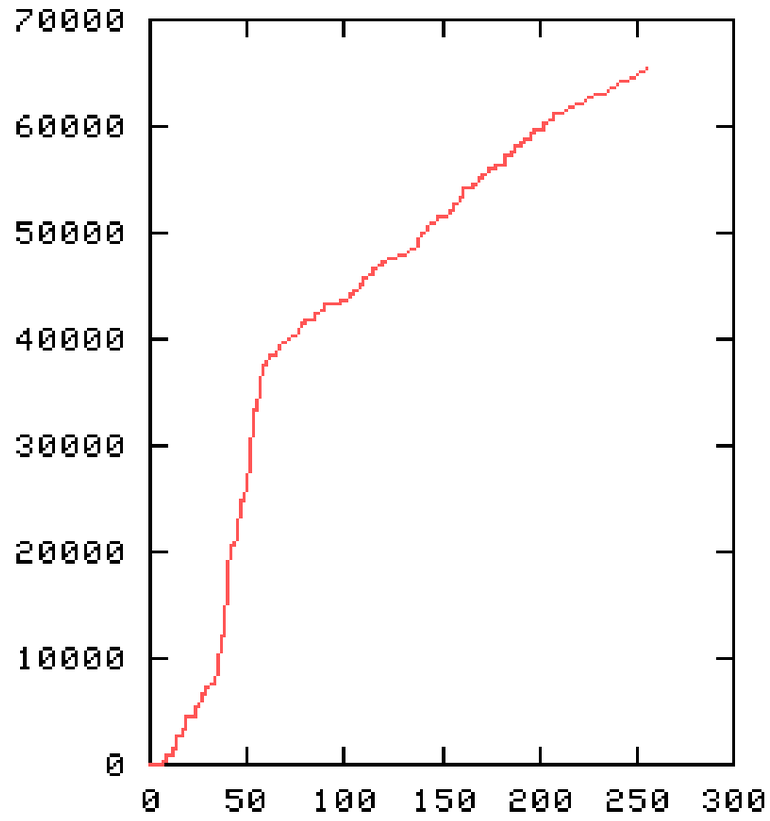
Cumulative Histogram



Normal Histogram

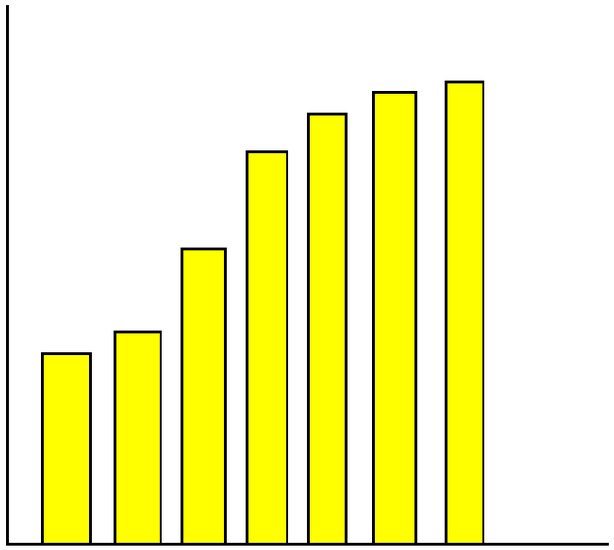


Cumulative Histogram

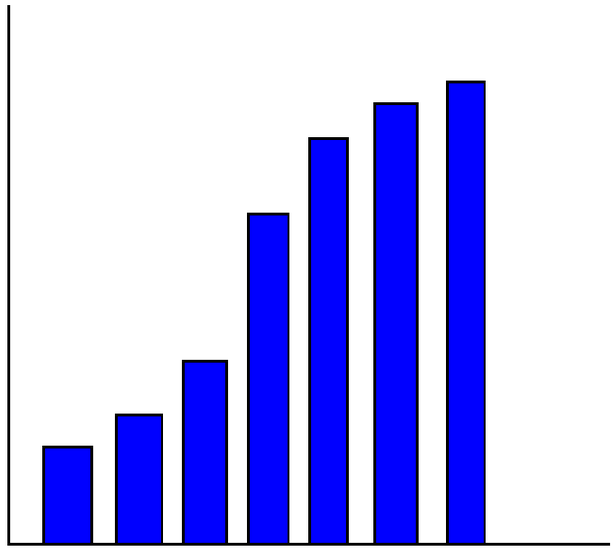


Similarity Measures Using the Histograms

Histogram 1

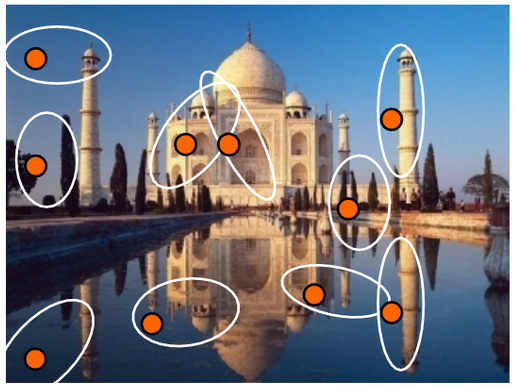


Histogram 2



Need to quantify how similar two histograms are

CBIR Idea: Treat Images Like a Bag Of "Image Feature Words"

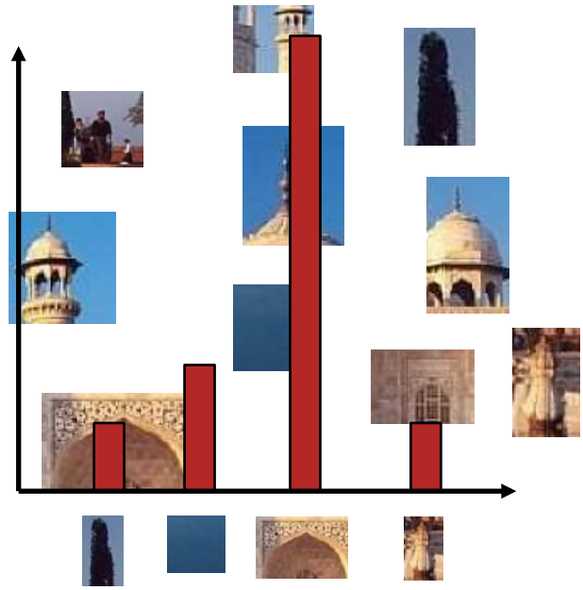


Feature Extraction

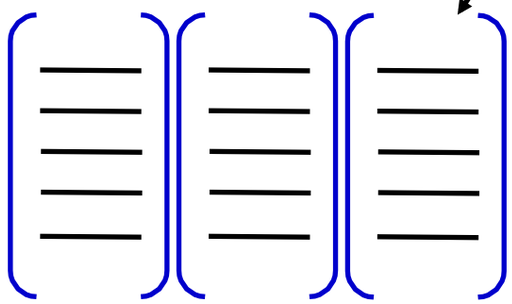
Vector Quantization

"clustering"

Index



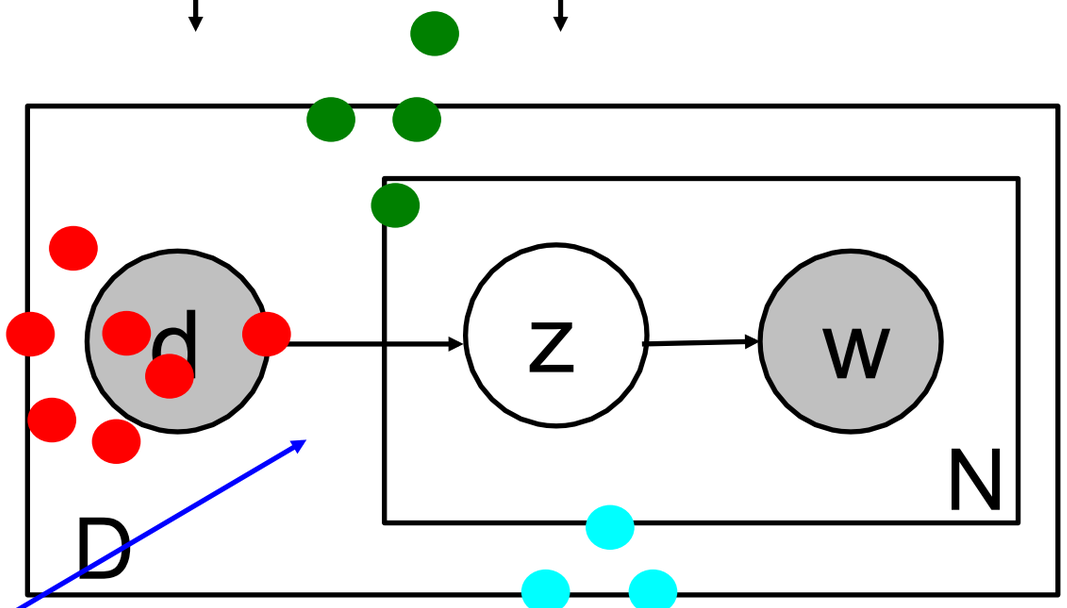
Words



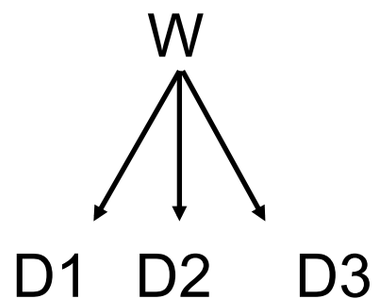
Compute SIFT descriptors

("image feature words")

a [Lowe'99]



PLSA, Hoffman, 2001



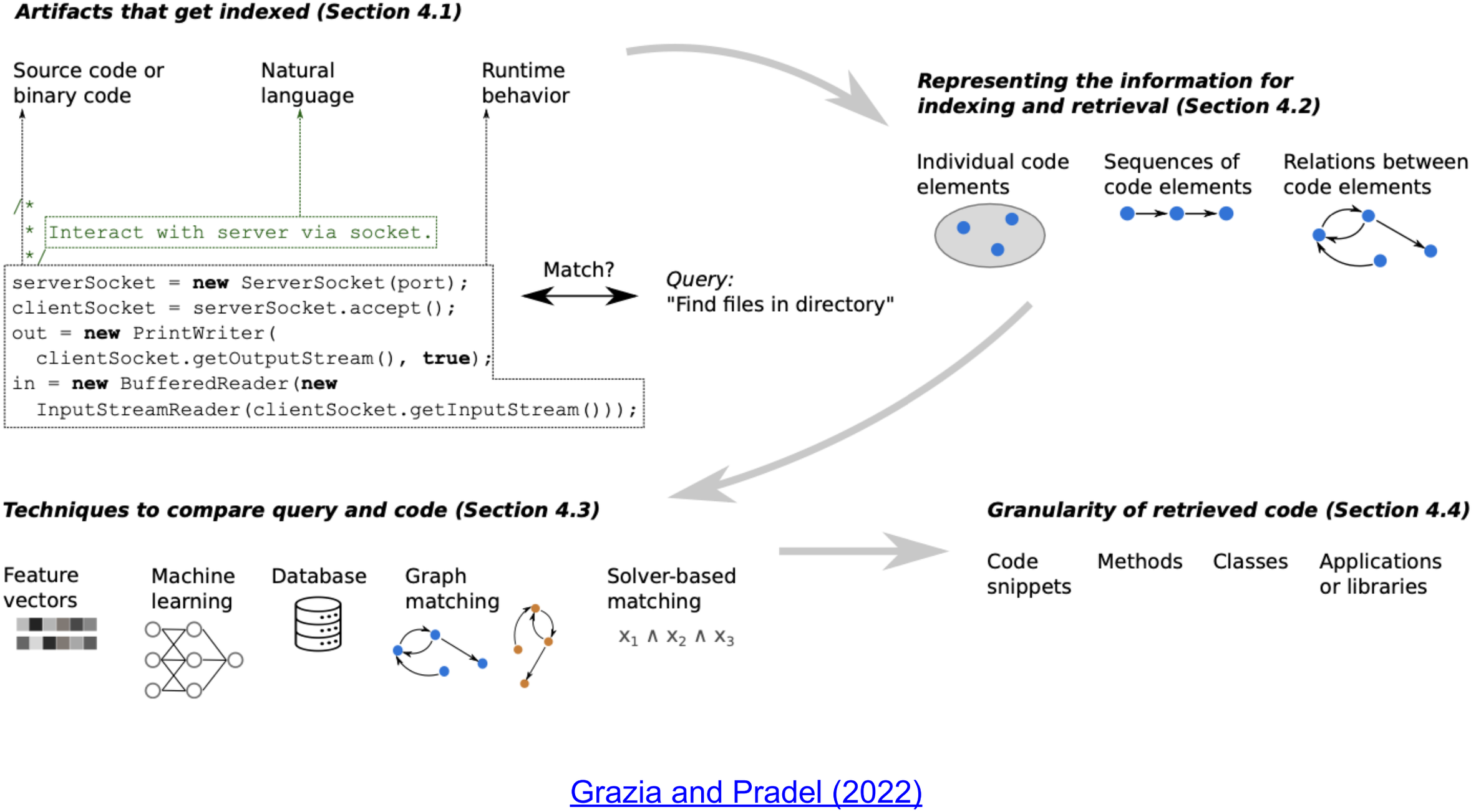
Inverted Index

CBIR Summary

- Content-based image retrieve needs to leverage computer vision techniques to represent the query and images
- Many features are possible:
  - low-level features (colors, textures)
  - mid-level features like SIFT
  - high-level features from image classification
    - Requires having lots of classes!
- Deep learning techniques are increasingly automating this feature-engineering

Code Search

Many, many ways to do code search



Grazia and Pradel (2022)

Future of Multimodal Search?

- Lots of exciting opportunities
- Deep learning makes it possible to rethink what can be encoded
  - code
  - images
  - voice
  - video?
  - gestures?

Sound to song

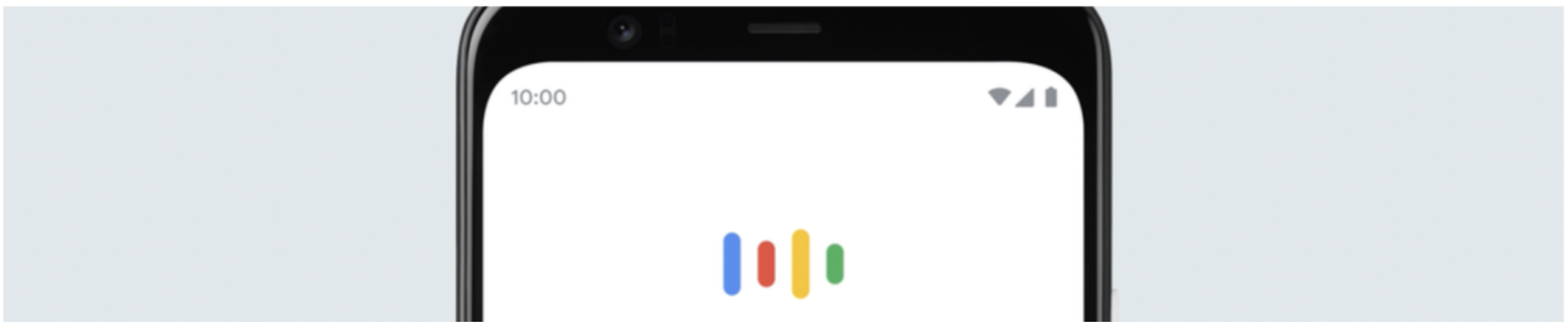
SEARCH

Song stuck in your head? Just hum to search

Oct 15, 2020 · 3 min read

Krishna Kumar
Senior Product Manager, Google Search

Share



Speech to speech

- Capture nuance in speech like…
  - Emphasis in how something is said
  - Speaker demographics
  - Emotion?
  - Speech rate?

INFERRING CLINICAL DEPRESSION FROM SPEECH AND SPOKEN UTTERANCES

Meysam Asgari, Izhak Shafran, and Lisa B. Sheeber

Author information Copyright and License information Disclaimer

Text to Video

- Search for particular sequences of events in long movies
  - “action scene where villain wins”
  - “people crying with joy”
  - “confusion”
  - “lots of dinosaurs running around”
- Challenges:
  - Video data is *big*.
  - Unclear what to extract as meaningful structure/segments at scale

What you need to know

- Text-to-image retrieval often uses non-image content (URL, caption, alt text) to represent query-and-image in the same space
- Document-as-Image retrieval is a special case where content extraction is essential
- Content-based Image Retrieval (CBIR) uses all image-derived features for representation
- Deep learning is rapidly making progress in this space