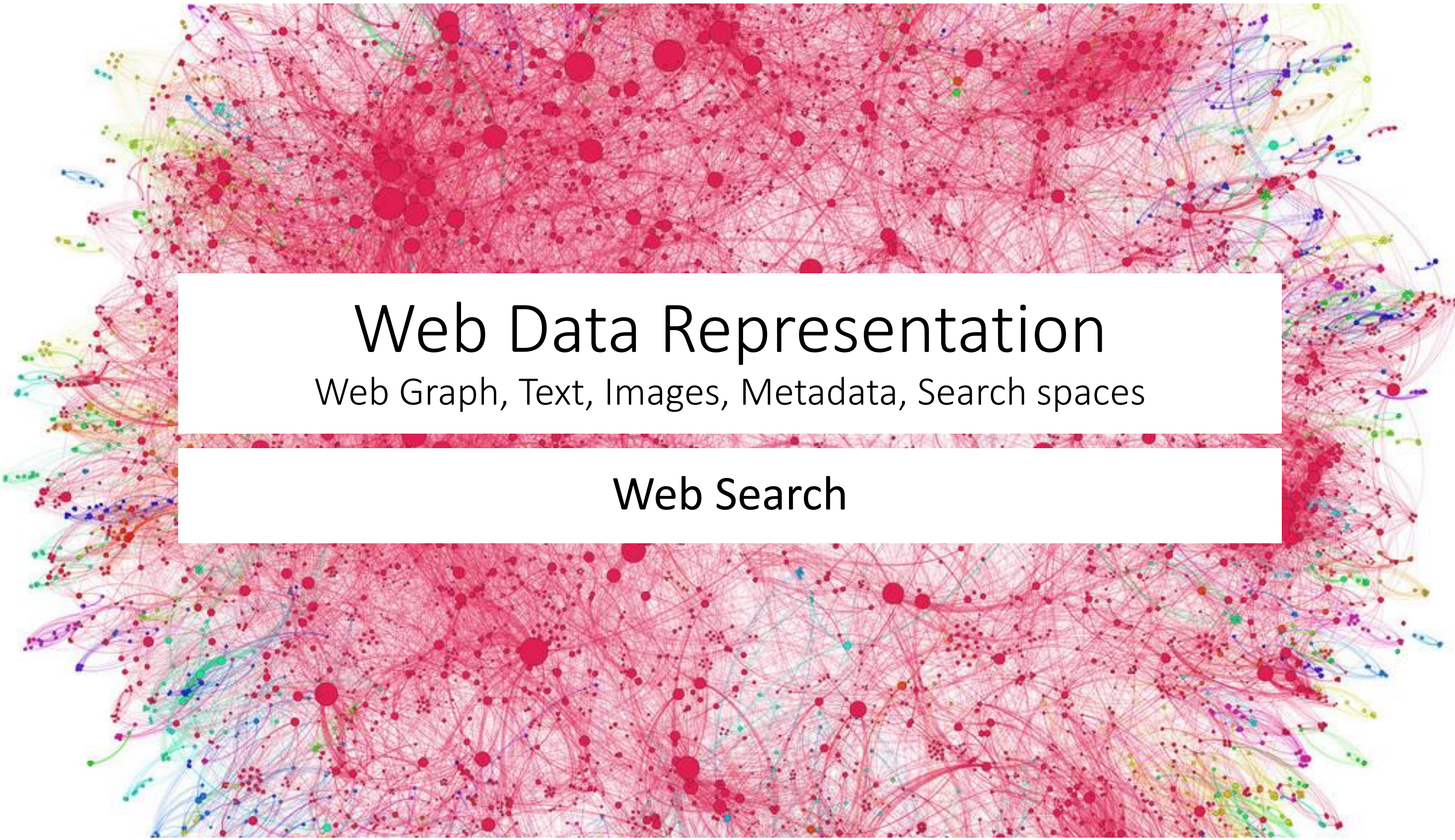
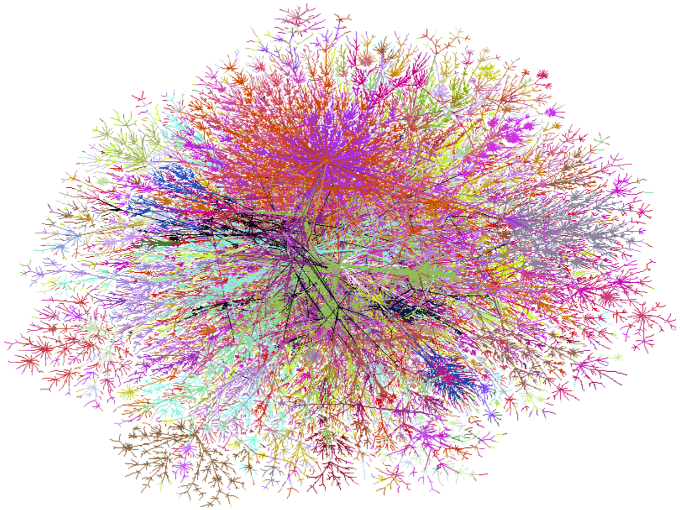


Web Data Representation

Web Graph, Text, Images, Metadata, Search spaces

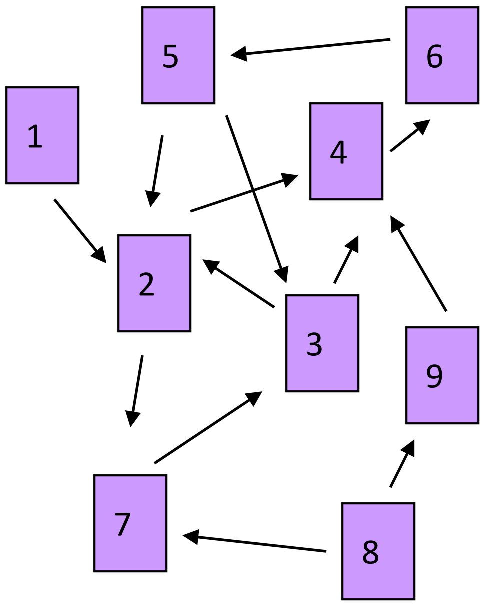
Web Search

The Web corpus

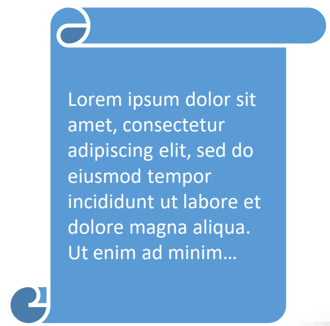


- No design/coordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text corpora… but corporate records are catching up.
- Content can be dynamically generated

Web data

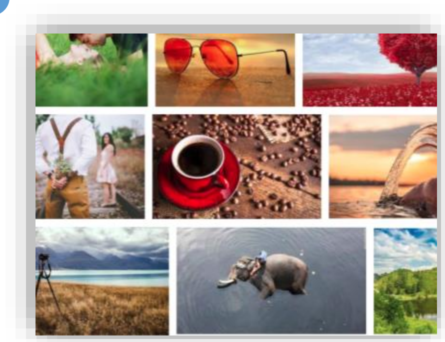


Links



Text

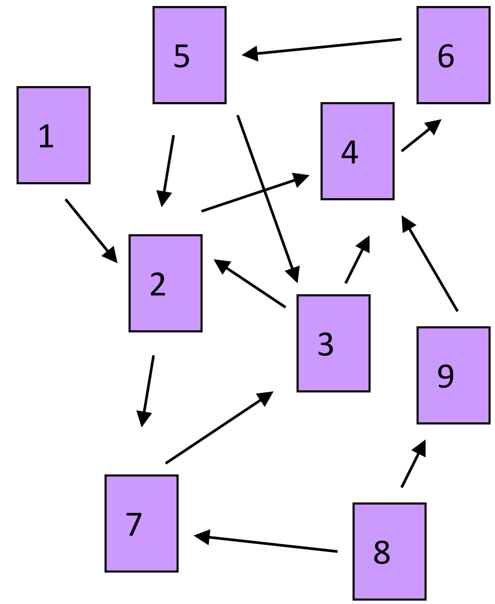
Ratings and clicks

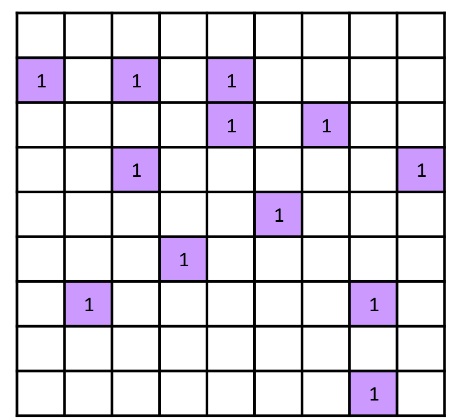


Images/videos

The Web graph

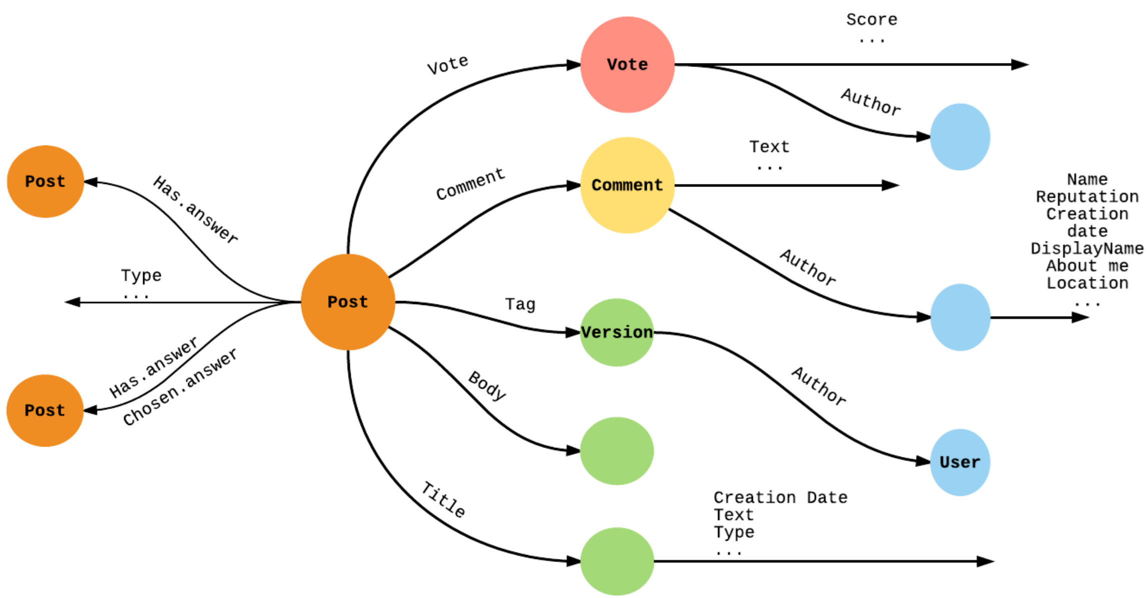
- Generally, the links can be explicit or computed by some function.
- The links can also be weighted by the similarity between pages (i.e. graph nodes in this case)
- Graphs are generally represented as a sparse matrix.
- There are many problems that make use of graph representations:
  - page importance, recommendation, reputation analysis...





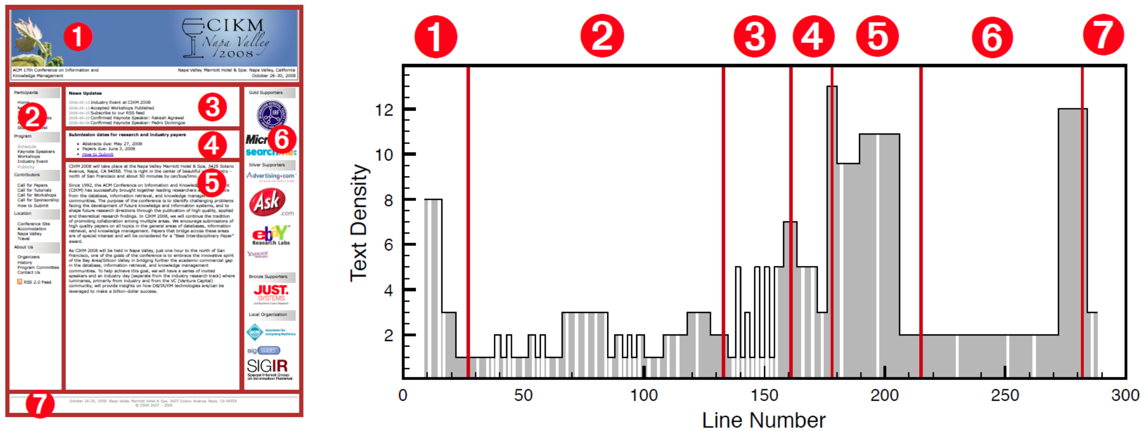
Graphs on the Web

- There are many types of graphs, besides hyperlinks.
- A single web page can be the source of many different information graphs:
  - Followers/followees, graph of named entities, authors, reply-to, ...



Web pages

- Web pages are divided into different parts (title, abstract, body, etc)
- Each part has a specific relevance to the main content
- A Web page can be divided by its HTML structure (e.g., <div> tags) or by its visual aspect.



Web page segmentation methods

- Segmenting visually
  - Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2003). VIPS: A vision-based page segmentation algorithm.
- Linguistic approach
  - Kohlschütter, C. , Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. ACM Web Search and Data Mining.
- Densitometric approach
  - Kohlschütter, C., and Nejdl, W., (2008). A densitometric approach to web page segmentation. ACM Conference on Information and Knowledge Management (CIKM '08).

https://boilerpipe-web.appspot.com/ https://github.com/kohlschutter/boilerpipe

Text data

- Instead of aiming at fully understanding a text document, traditional search engines take a pragmatic approach and look at the most elementary text patterns
  - e.g. a simple histogram of words, also known as “bag-of-words”.
- Heuristics capture specific text patterns to improve search effectiveness
  - Enhances the simplicity of word histograms
- The most simple heuristics are stop-words removal and stemming

Character processing and stop-words

- Term delimitation
- Punctuation removal
- Numbers/dates
- Stop-words: remove words that are present in all documents
  - *a, and, are, as, at, be, but, by, for, if, in, into, is, it, no, not, of, on, or, such, that, the, their, then, there, these, they, this, to, was, will…*

Stemming and lemmatization

- Stemming: Reduce terms to their “roots” before indexing
  - “Stemming” suggest crude affix chopping
  - **e.g., automate(s), automatic, automation all reduced to automat.**
    - http://tartarus.org/~martin/PorterStemmer/
    - http://snowball.tartarus.org/demo.php

- Lemmatization: Reduce inflectional/variant forms to base form, e.g.,
  - *am, are, is* $\rightarrow$ *be*
  - *car, cars, car's, cars'* $\rightarrow$ *car*

Chapter 2: “**Introduction to Information Retrieval**”, Cambridge University Press, 2008

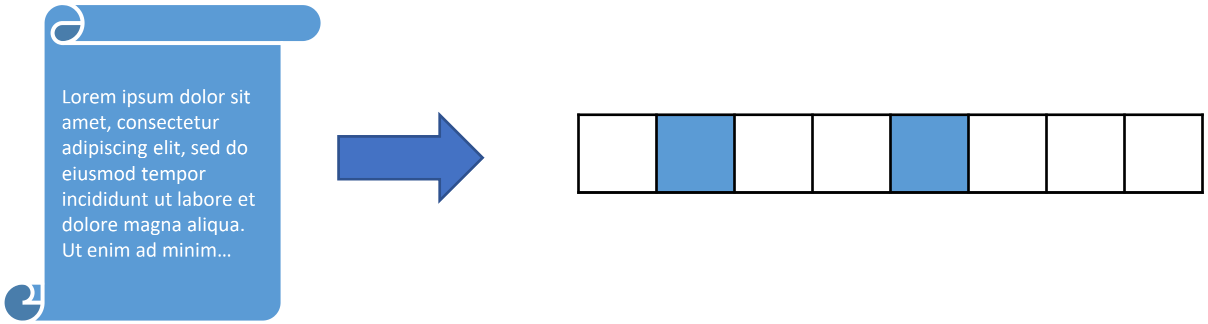
N-grams

- An n-gram is a sequence of items, e.g. characters, syllables or words.
- Can be applied to text spelling correction
  - *“interactive meida” >>>> “interactive media”*
- Can also be used as indexing tokens to improve Web page search
  - You can order the Google n-grams (6DVDs):
    - http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
- N-grams were under some criticism in NLP because they can add noise to information extraction tasks
  - ...but are widely successful in IR to infer document topics.

"Bag of Words" representation

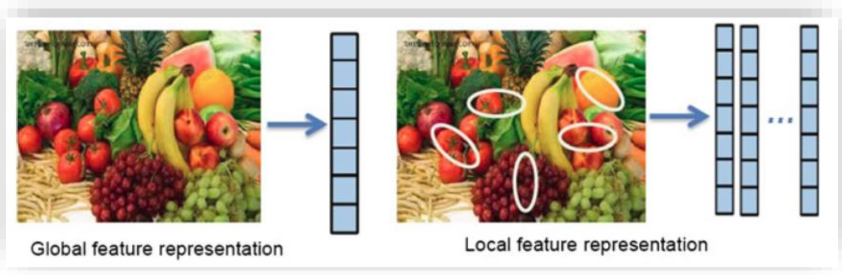
- After the text analysis steps, a document (e.g. Web page) is represented as a vector of terms and n-grams.
  - More complex low-level representations can be used

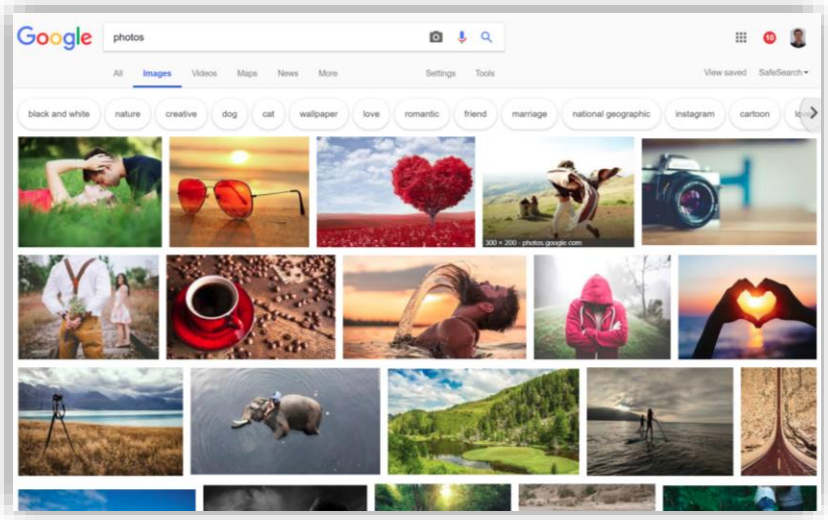
$$d = (w_1, \dots, w_L, ng_1, \dots, ng_M)$$



Visual data

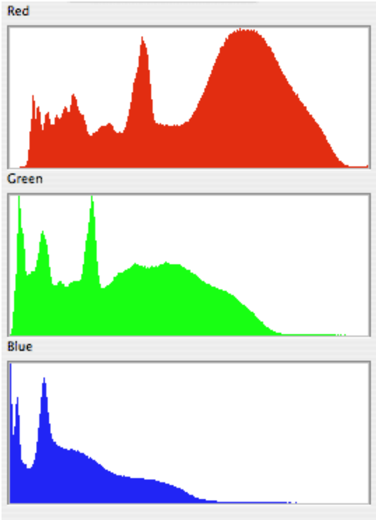
- Visual information also needs to be processed and analysed.
- A compact representation of the image/video content is computed from it.
- This compact representation is then used to accomplish several tasks, e.g. search, categorization.



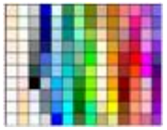


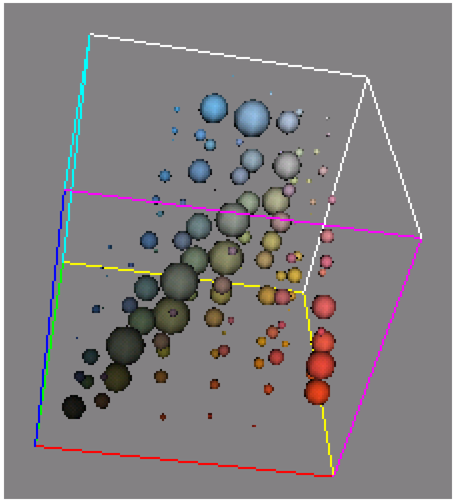
Histograms of colors

- Marginal color histograms consider color channels independently
  - The number of bins define the dimensionality of the space



- 3D colour histograms divide the space into small 3D boxes
  - The numbers of bins per dimension define the number of 3d bins





Color moments

- Color moments measure the statistical properties of the histogram:
  - Mean and variance (1st and 2nd moments)

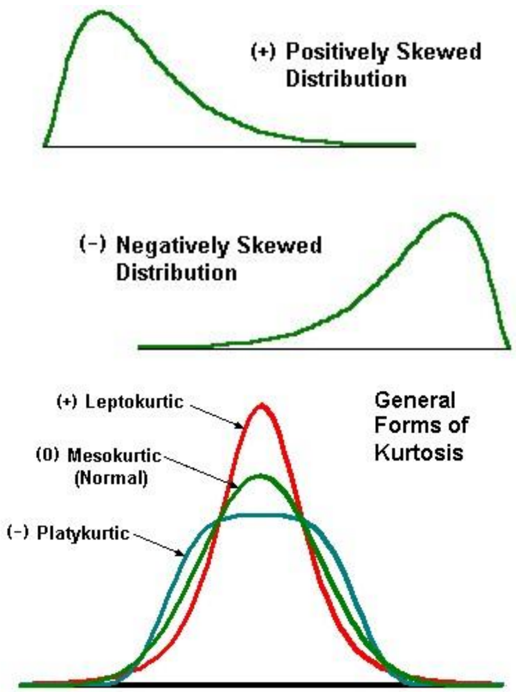
$$m_r = \sum \frac{(X_i - \overline{X})^r}{n}$$

  - Skewness (3rd moment)

$$Skewness = \left[\frac{\sqrt{n(n-1)}}{n-2}\right] \times \frac{m_3}{m_2^{3/2}}$$

  - Kurtosis (4th moment)

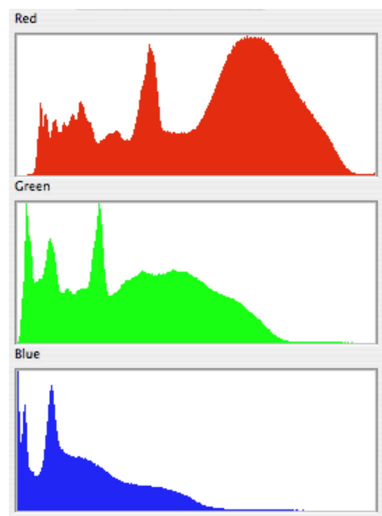
$$Kurtosis = \left[\frac{(n-1)(n+1)}{(n-2)(n-3)}\right] \times \frac{m_4}{(m_2)^2} - 3\left[\frac{(n-1)^2}{(n-2)(n-3)}\right]$$



Example

Color moments

Marginal color histograms



$$d_{hR} = (bin_1, bin_2, \ldots, bin_{16})$$

$$d_{hG} = (bin_1, bin_2, \ldots, bin_{16})$$

$$d_{hB} = (bin_1, bin_2, \ldots, bin_{16})$$

$$d_{cm} = \left(\overbrace{m_R, s_R^{\,2}}, \overbrace{m_G, s_G^{\,2}}, \overbrace{m_B, s_B^{\,2}}\right)$$

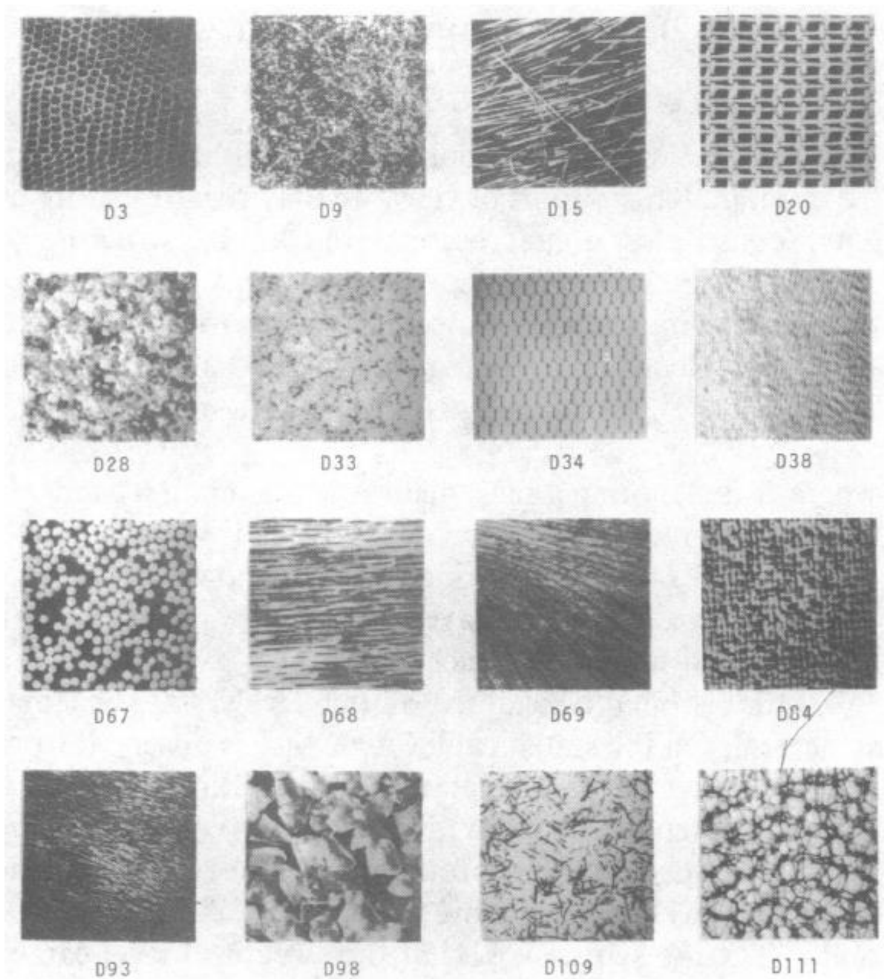
Textures

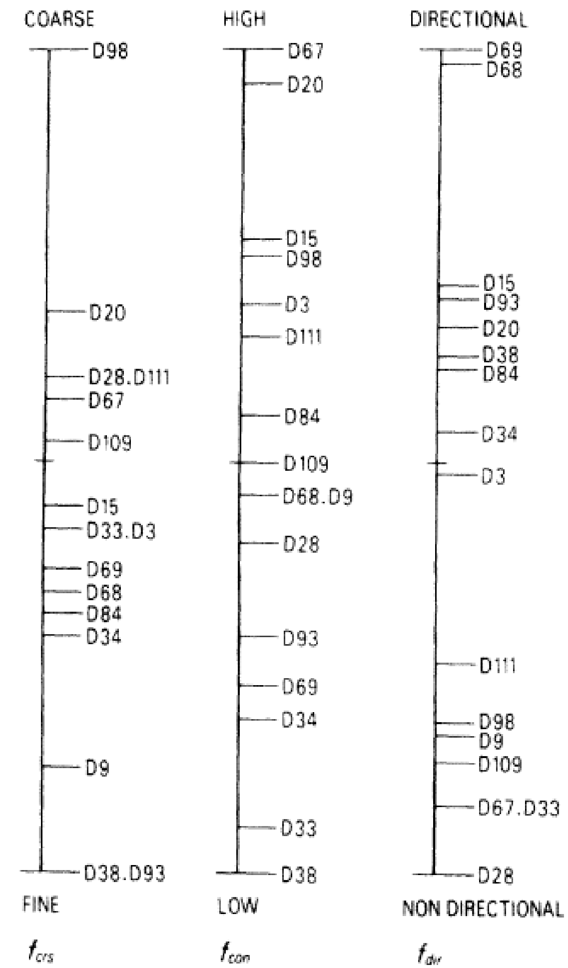
Psychological based textures (Tamura)

- **Coarseness** measures the size of the primitive elements forming the texture
- **Contrast** measures variation in gray levels between black and white
- **Directionality** measures the orientation of the texture
- **Line-likeliness** measures the similarity of the texture to lines
- **Regularity** measures the repetetiveness of the texture pattern
- **Roughness** “we do not have any good ideas for describing the tactile sense of roughness”

Tamura, H., Mori, S., Yamawaki, T., “Textural features corresponding to visual perception,” IEEE Trans on Systems, Man and Cybernetics 8 (1978) 460–472

Psychological based textures (Tamura)

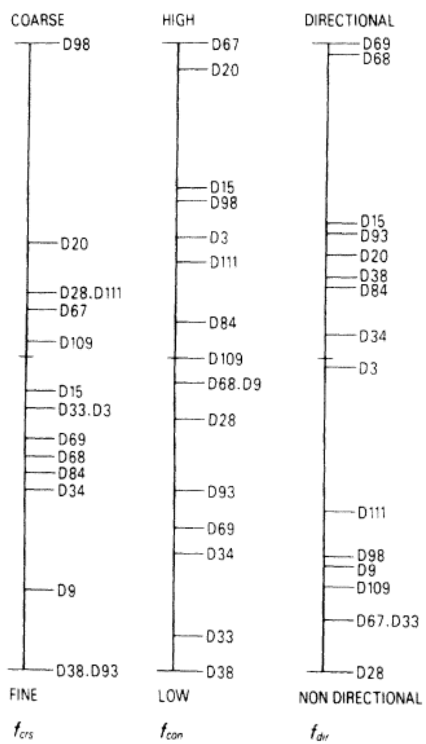




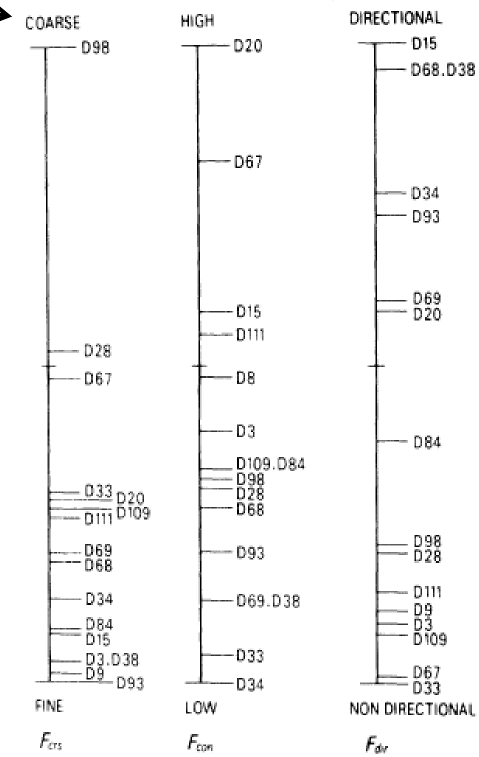
Tamura, H., Mori, S., Yamawaki, T., “Textural features corresponding to visual perception,” IEEE Trans on Systems, Man and Cybernetics 8 (1978) 460–472

Comparing psychological relevance to algorithms

Humans



Algorithm



Ranked relevance metrics

Frequency based textures

- Frequency based texture decompose images according to their frequencies
    - Similar to audio filtering or color filter lenses
- The number of repetitions per area in a texture is related to the frequency of a texture
- Based on the Fourier Transform
- A set of 2 dimensional filters will decompose images into their natural frequencies

Manjunath, B., Ma, W., "Texture features for browsing and retrieval of image data," IEEE Trans on Pattern Analysis and Machine Intelligence 18 (1996) 837–842

Edge detection

J. Canny, “A Computational Approach to Edge Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, Nov. 1986.

Edge detection

- Filter image with a low pass filter

- Apply vertical and horizontal filters to compute Gx and Gy:

+1	+2	+1
0	0	0
-1	-2	-1

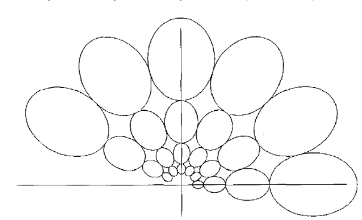
-1	0	+1
-2	0	+2
-1	0	+1

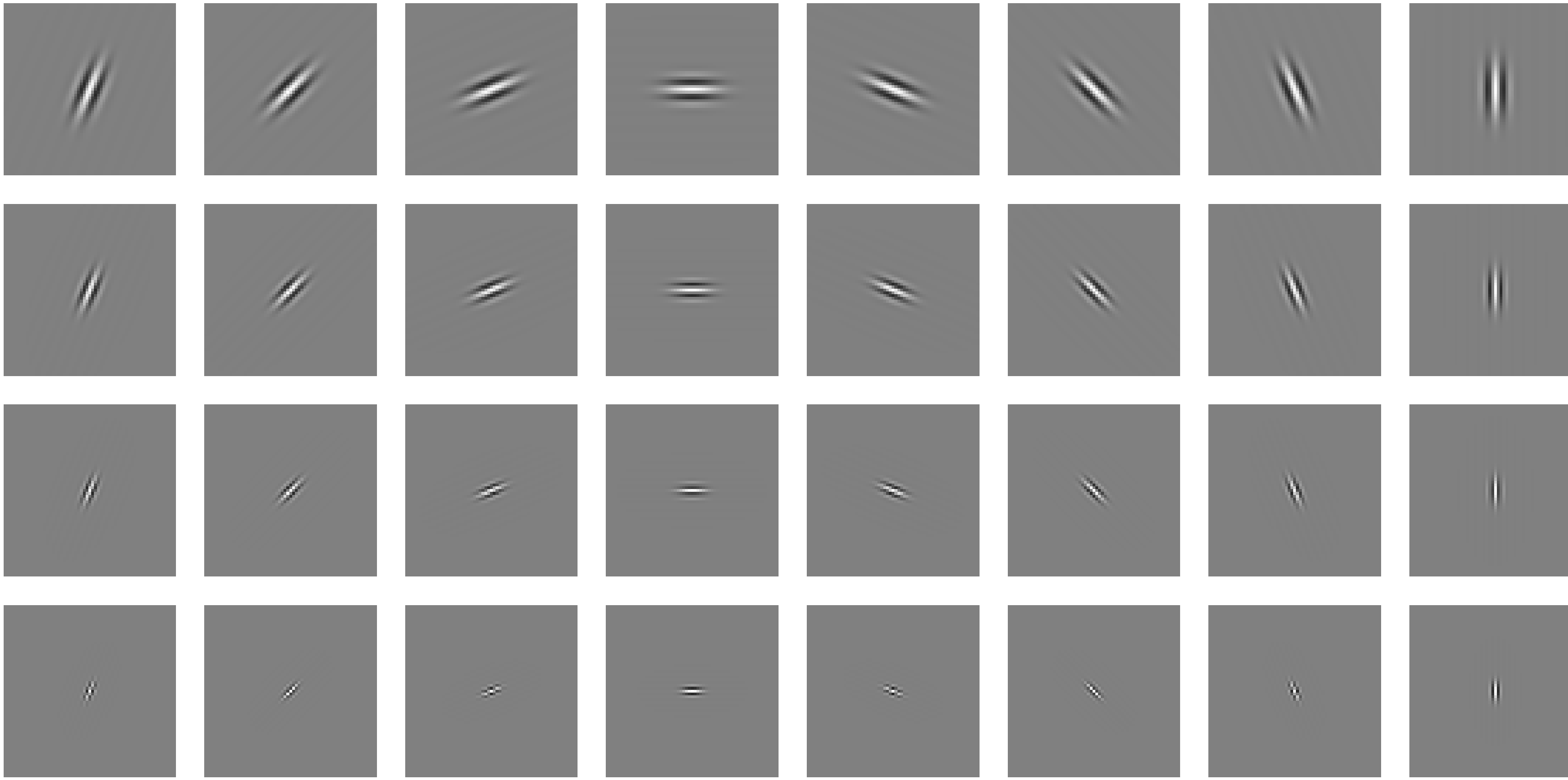
- Compute the gradients as $\mathbf{G} = \sqrt{\mathbf{G}_x{}^2 + \mathbf{G}_y{}^2}$
  - Reduce it to one of the 4 possible directions (0º, 45º, 90º, 135º)

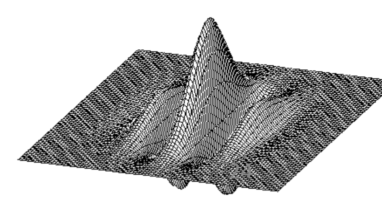
- Compute the orientation of the edges as: $\mathbf{\Theta} = \operatorname{arctan}\left(\frac{\mathbf{G}_y}{\mathbf{G}_x}\right)$

J. Canny, "A Computational Approach to Edge Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, Nov. 1986.

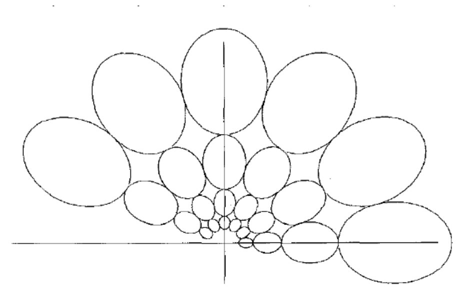
Gabor filters

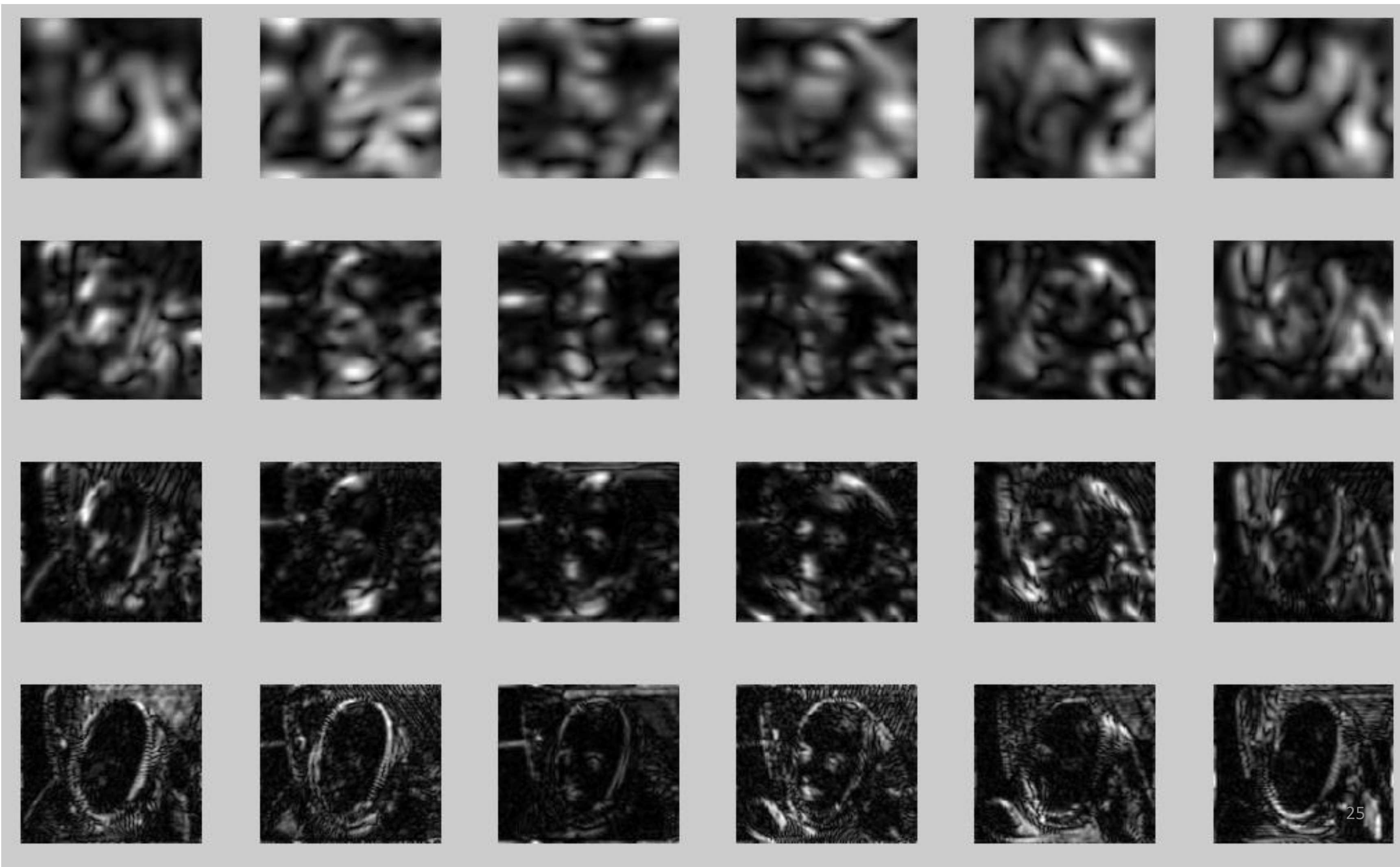






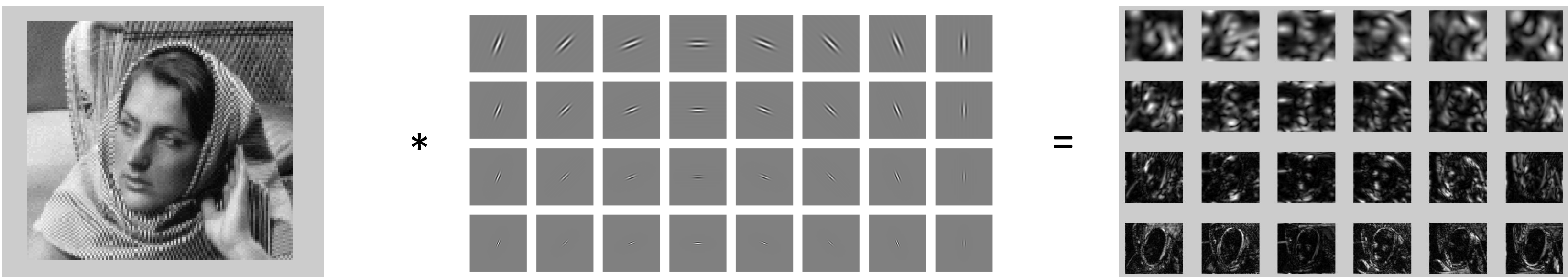
Manjunath, B., Ma, W., “Texture features for browsing and retrieval of image data,” IEEE Trans on Pattern Analysis and Machine Intelligence 18 (1996) 837–842





Gabor texture feature

- Images are convolved (operator *) with each filter individually:



A widely used descriptor corresponds to the mean and variance of the output of each filter:

$$d_{texture} = (m_1, v_1, \ldots, m_k, v_k)$$

Manjunath, B., Ma, W., “Texture features for browsing and retrieval of image data,” IEEE Trans on Pattern Analysis and Machine Intelligence 18 (1996) 837–842

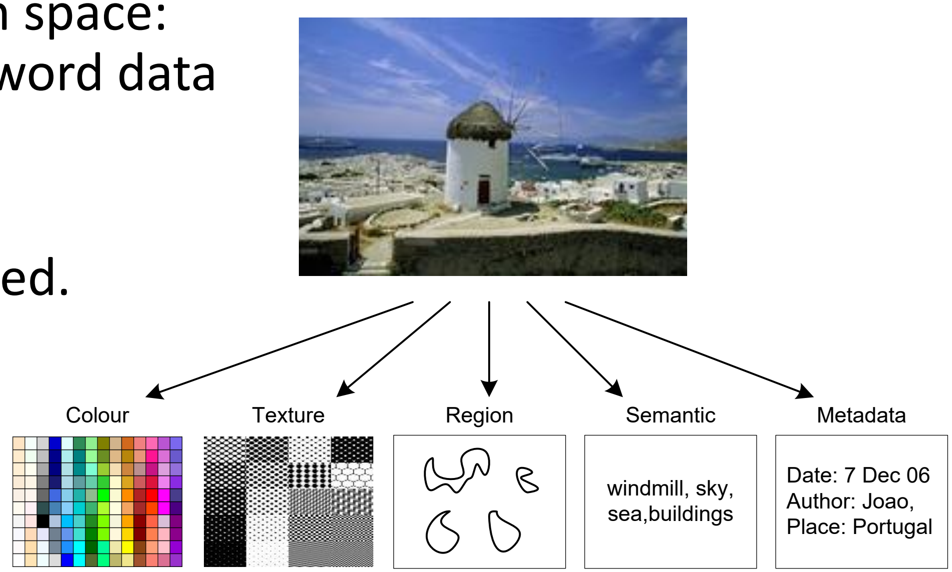
Multiple representations of the same data

- Documents are represented as the set of vectors

$$d = (d_{links}, d_{text}, d_{color}, d_{texture}, d_{metadata}, d_{tags}, \ldots)$$

each one for a different search space: text data, visual data, and keyword data respectively.

- Other search spaces can be used.



Data representations

- Link data

$$d_{links} = (0,0,\ldots,0,1,0,\ldots,0,1,0,\ldots,0)$$

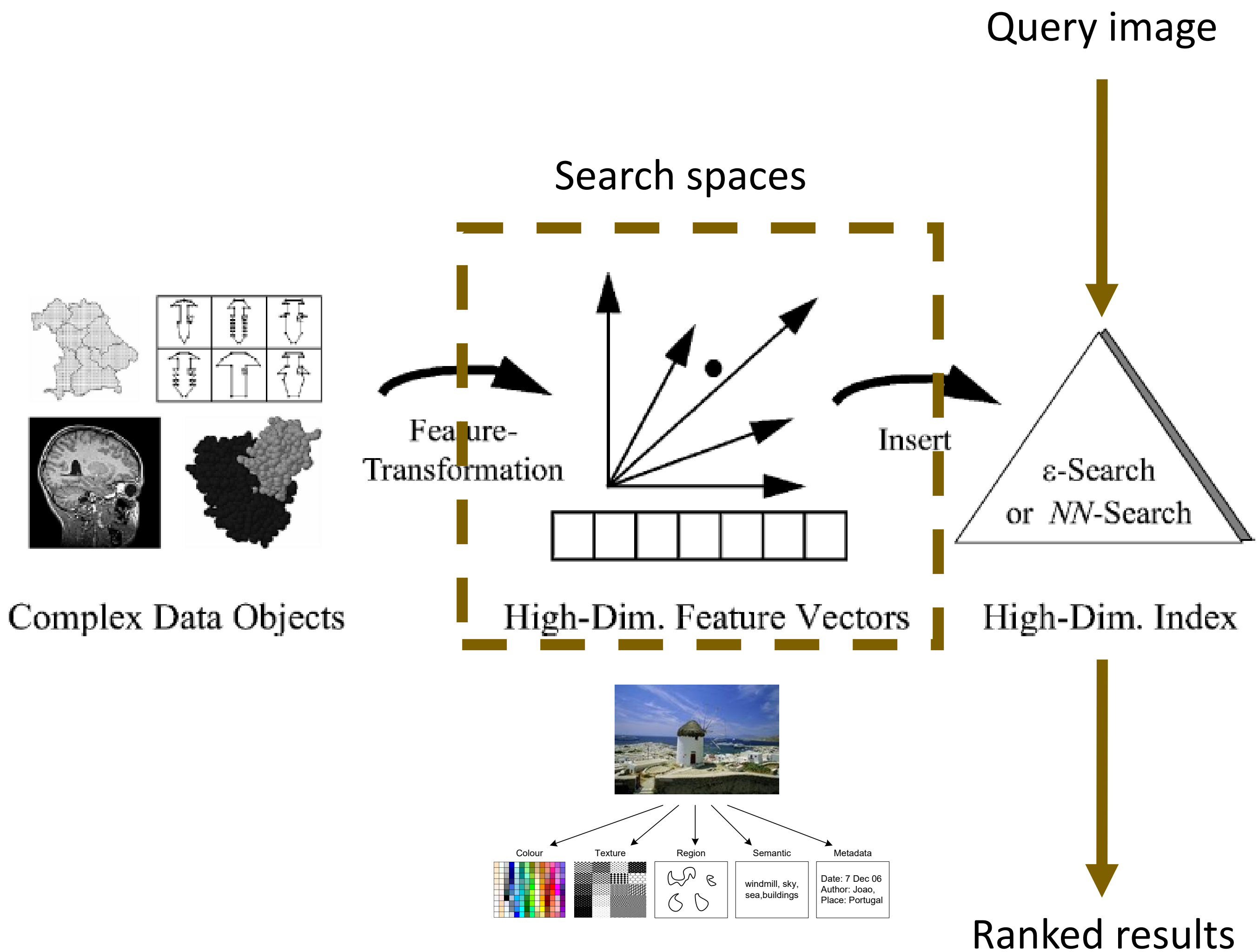
- High-dimensional data
  - Sparse
    - Bag of words
  - Dense
    - Color histograms and moments
    - Textures and edges

$$d_{bow} = (w_1,\ldots,w_L,ng_1,\ldots,ng_M)$$

$$d_{color} = (bin_1,bin_2,\ldots,bin_k)$$

$$d_{texture} = (m_1,v_1,\ldots,m_k,v_k)$$

Search high-dimensional spaces



Definition: metric spaces

- Let $\mathfrak{D}$ be an *n* dimensional space, where each data point is defined as

$$d \in \mathfrak{D}:\ \ d = (d_1, \ldots, d_n),\ \ \ d_i \in \mathbb{R}$$

- The *n* dimensional space $\mathfrak{D}$ is a metric space iff exists a distance function $dist(a, b)$ in $\mathfrak{D}$.

- A distance function has the following properties:

  - Non-negative: $dist(a, b) \quad\quad \forall a, b \in \mathfrak{D}$

  - Indentity: if $dist(a, b) = 0 \quad\quad then\ a = b$

  - Symmetry: $dist(a, b) = dist(b, a) \quad \forall a, b \in \mathfrak{D}$

  - Triangle inequality $dist(a, b) \leq dist(a, c) + dist(c, b) \quad \forall a, b, c \ \in \mathfrak{D}$

Distance vs similarity

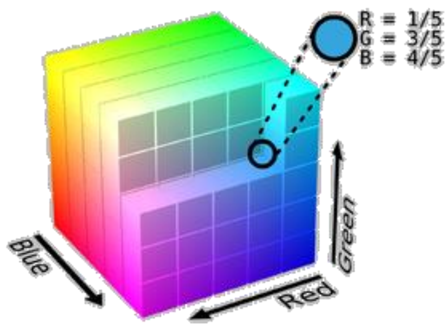
- Distances in a given search space must be meaningful.

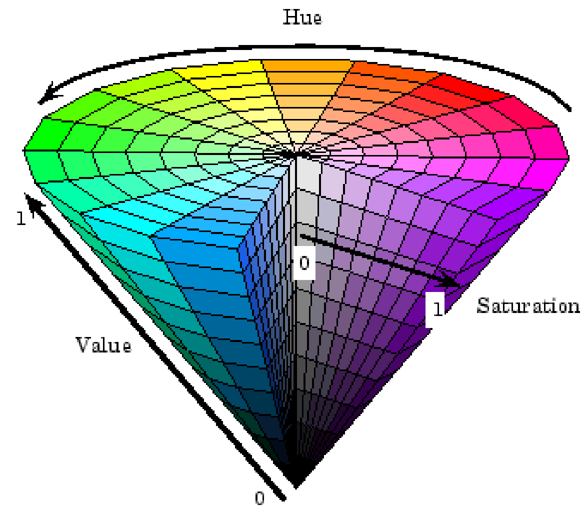
- Distances are used as proxies for similarity.
    - distance = 1-similarity

- Vector spaces and probability spaces are common spaces in Web search.

- The goal is that the similarity/distance between a query and candidate documents will reflect the relevance of the document to the user query.

Example: Distance in the RGB vs HSV color spaces



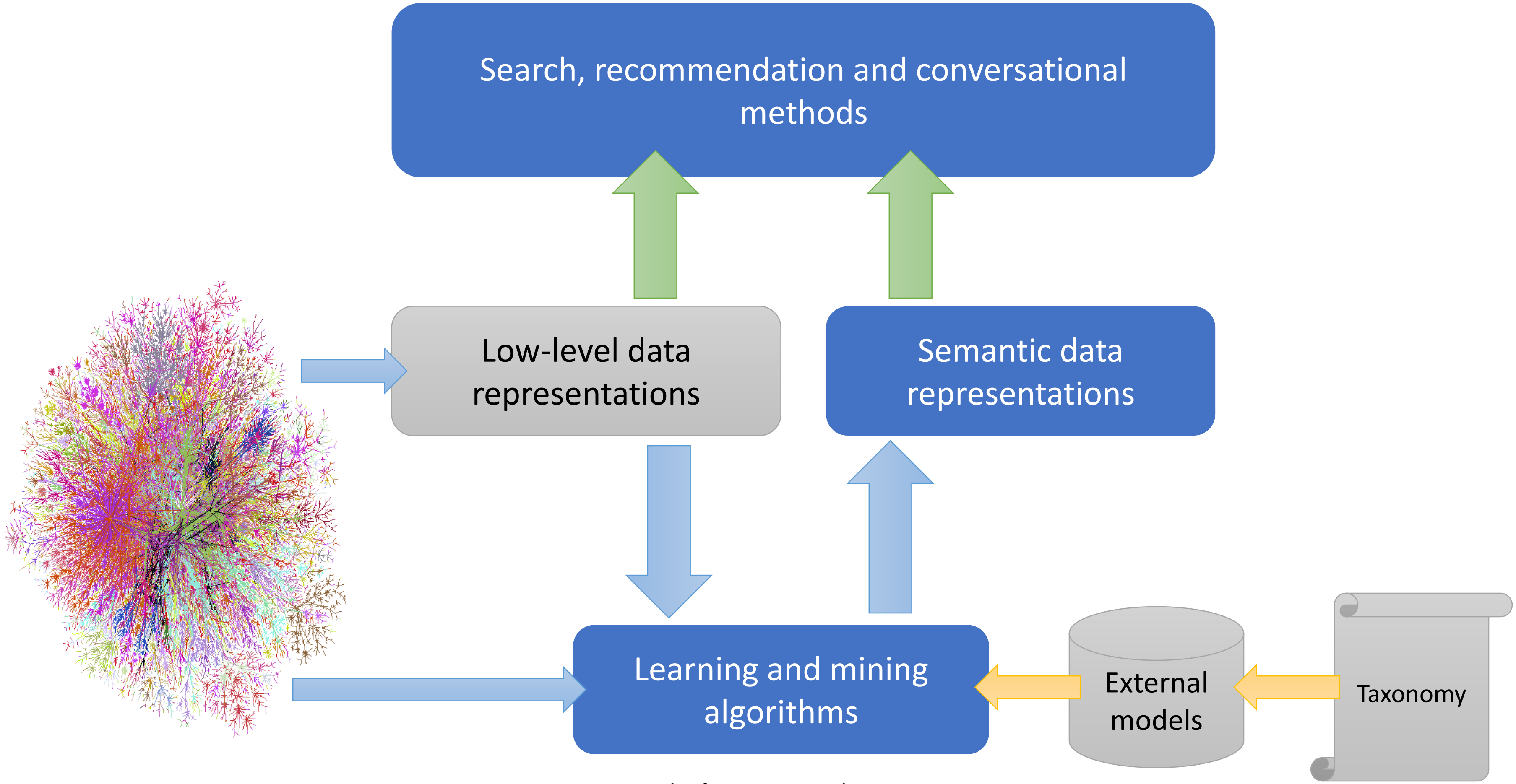


- Euclidean distance in the HSV color space is more meaningful!
  - Hue (H), the color type (such as red, green). It ranges from 0 to 360 degree.
  - Saturation (S) of the color ranges from 0 to 100%. Also sometimes it called the "purity".
  - Value (V), the Brightness (B) of the color ranges from 0 to 100%.

Searching Web content

- Processing real-world information is challenging!!!
- The aim is to search any unstructured data by its content
  - Textual, visual, audio, semantic, etc.
- Data contains very complex information patterns.
- Information needs can be very complex.
  - Queries can be *keywords*, *examples* or *questions*.
  - Finding related trends (consumption patterns)
  - Search images with text and vice-versa

Web Mining and Search course scope



- Multi-feature search
- Learning data representations
- Graph data (Labels and PageRank)
- Recommendation

Summary and readings

- Understanding the **diversity of Web data**
- Understanding the need to have different **data representations**
- **Spaces and similarity functions**
- References:
  - Chapter 2: C. D. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval", Cambridge University Press, 2008.
  - Hassaballah, M., Abdelmgeid, A. A., & Alshazly, H. A. (2016). Image features detection, description and matching. In *Image Feature Detectors and Descriptors* (pp. 11-45). Springer, Cham.