Future Trends in Similarity Searching

Pavel Zezula
Masaryk University
Brno, Czech Republic

Outline of the talk

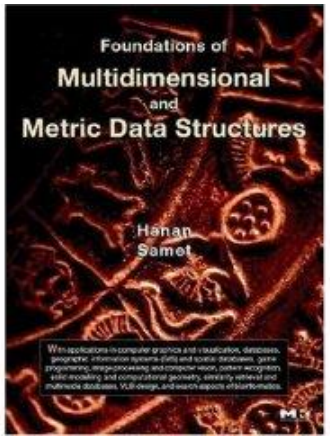
- Introduction
- Many faces of similarity
- Real life and digital similarity
- On the importance of searching
- Current technology and its limitations
- Similarity search computing services
- SISAP after 5 years

Introduction

- Honor and responsibility of Keynote talks
- Technical talk versus a conceptual one
- SISAP is a narrow community – I am a part of it
- Most of my technical knowledge is summarize in the Similarity Search book
- My objectives: current experience, where to go, how to proceed
- Presented observations reflect my personal experience with: **similarity**, **search**, **applications**

Metric Searching technology

Hanan Samet
Foundation of Multidimensional and Metric Data Structures
Morgan Kaufmann, 2006



P. Zezula, G. Amato, V. Dohnal, and M. Batko
Similarity Search: The Metric Space Approach
Springer, 2006

Real-life Similarity

- Are they similar?

Real-life Similarity

- Are they similar?

Real-life Similarity

- Are they similar?

Real-life Similarity

- Are they similar?

Real-Life Motivation

The social psychology view

- Any event in the history of organism is, in a sense, **unique**.

- *Recognition*, *learning*, and *judgment* presuppose an ability to categorize stimuli and classify situations by **similarity**.

- Similarity (*proximity*, *resemblance*, *communality*, *representativeness*, *psychological distance*, etc.) is **fundamental** to theories of *perception*, *learning*, *judgment*, etc.

Contemporary Networked Media

The digital data view

- Almost **everything** that we *see*, *read*, *hear*, *write*, *measure*, or *observe* can be **digital**.
- Users **autonomously** *contribute* to production of global media and the growth is **exponential**.
- Sites like Flickr, YouTube, Facebook host user contributed content for a variety of **events**.
- The elements of networked media are related by numerous multi-facet **links of similarity**.

Challenge

- Networked media is getting close to the human “fact-bases
  - the gap between physical and digital has blurred

- **Similarity data management** is needed to *connect*, *search*, *filter*, *merge*, *relate*, *rank*, *cluster*, *classify*, *identify*, or *categorize* objects across various collections.

WHY?

It is the *similarity* which is in the world *revealing.*

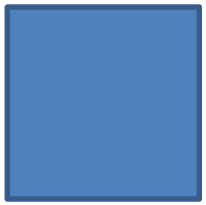
Similarity & Geometry

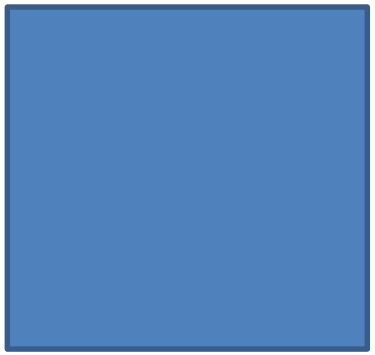
- Figures that have the same shape but not necessarily the same size are *similar figures*:
- Any two line segments are similar:

- Any two circles are similar:

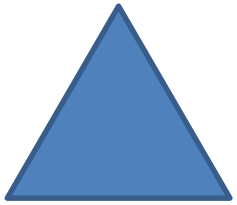
Similarity & Geometry

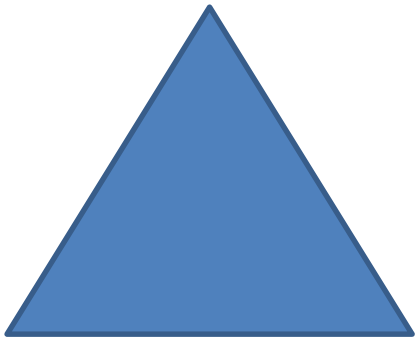
- Any two squares are similar:





- Any two equilateral triangles are similar:





Similarity & Geometry

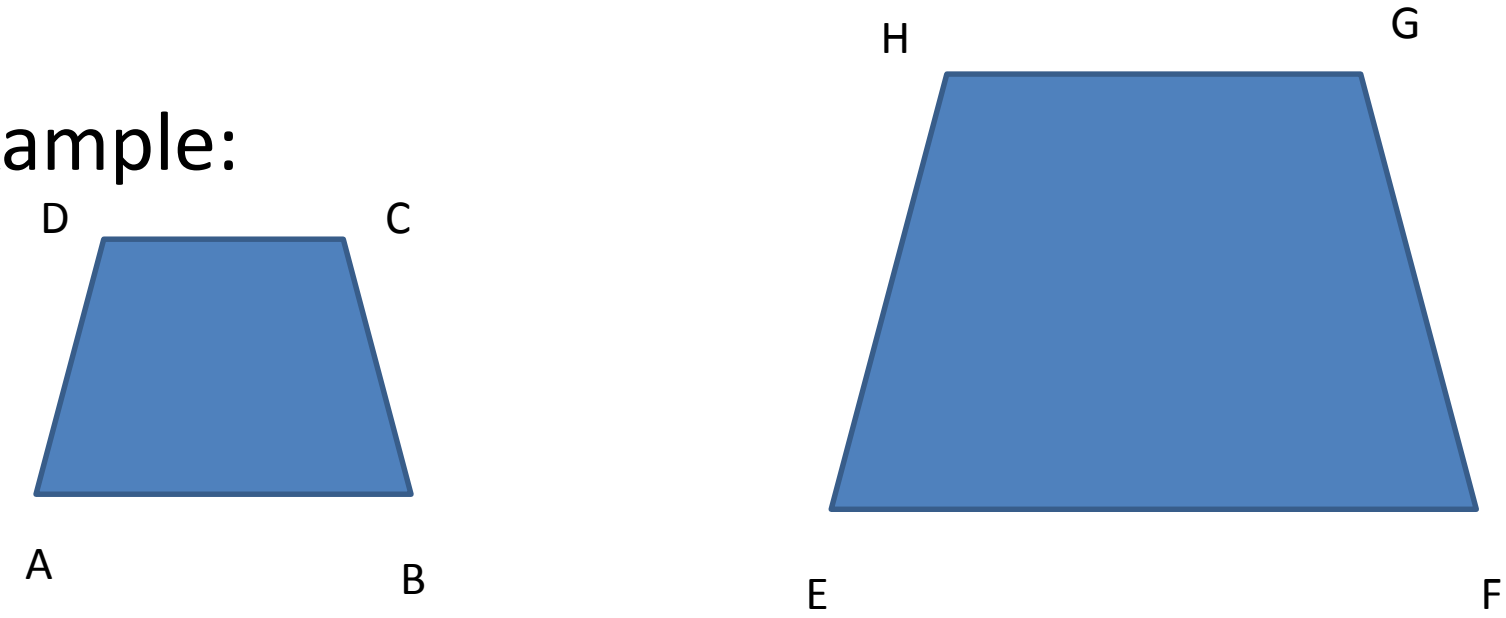
- Definition:

- Two polygons are similar to each other, if:

1. Their corresponding angles are equal
2. The lengths of their corresponding sides are proportional

Similarity & Geometry

- Example:



- <A = <E; <B = <F; <C = <G; <D = <H, and also
- AB/EF=BC/FG=CD/GH=DA/HE

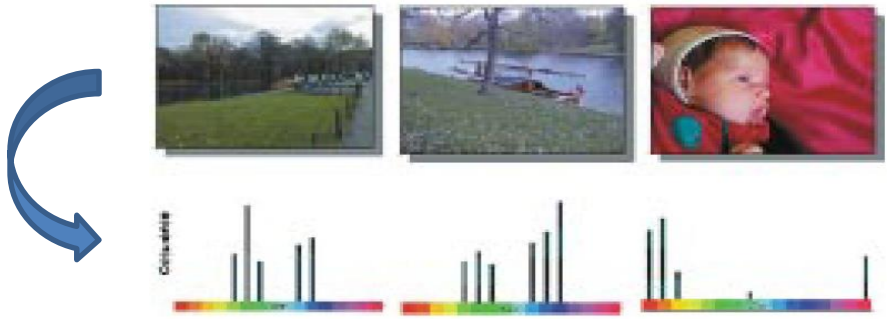
Similarity & Geometry

- If one polygon is similar to a second polygon, and the second polygon is similar to the third polygon, the first polygon is similar to the third polygon.
- In any case:

Two geometric figures are either similar or they are not similar at all

Visual Similarity

- MPEG-7 multimedia content desc. standard
- Global feature descriptors:
  - Color, shape, texture, …



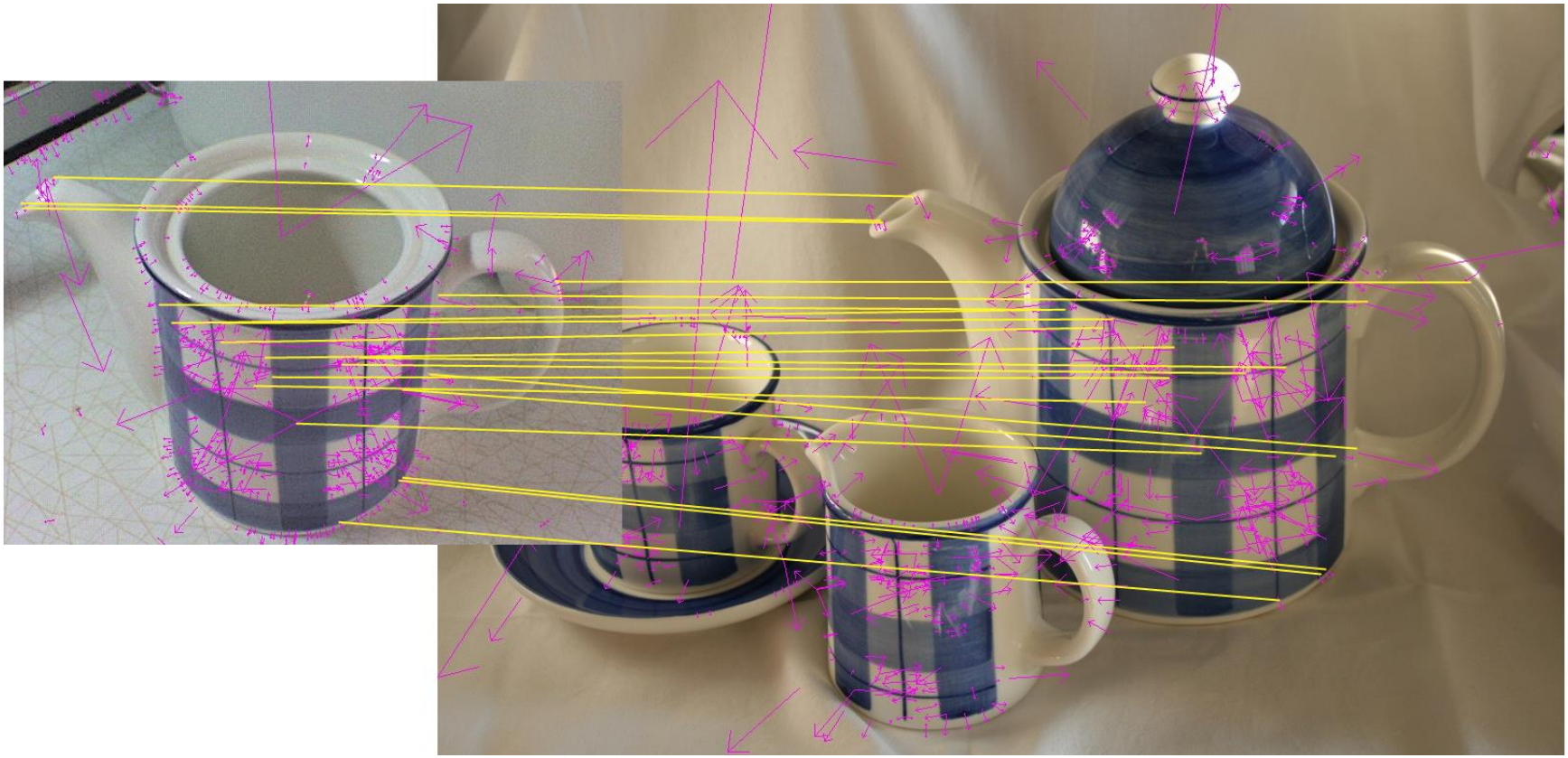
  - One high-dimensional vector per image

Multiple Visual Aspects

Visual Similarity

- Local feature descriptors – SIFT, SURF, etc.
- Invariant to image scaling, small viewpoint change, rotation, noise, illumination

Visual Similarity - finding coresspondence



Biometric Similarity

- Biometrics:
  - methods of recognizing a person based on physiological and behavioral characteristics
- Two types of recognition problems:
  - Verification – authenticity of a person
  - Identification – recognition of a person
- Examples:
  - Finger prints, face, iris, retina, speech, gait, etc.

Biometrics: Fingerprint

- Minutiae detection:
  - Detect ridges (endings and branching)
  - Represented as a sequence of minutiae
    - $P=((r_1,e_1,\theta_1), \ldots, (r_m,e_m,\theta_m))$
    - Point in polar coordinates (r,e) and direction θ
- Matching of two sequences:
  - Align input sequence with database one
  - Compute weighted edit distance
    - $w_{ins,del}=620$
    - $w_{repl}=[0;26]$ - depending on similarity of two minutiae

Biometrics: Hand Recognition

- Hand image analysis
  - Contour extraction, global registration
    - Rotation, translation, normalization
  - Finger registration
  - Contour represented as a set of pixels $F=\{f_1,\ldots,f_{N_F}\}$
- Matching: modified Hausdorff distance

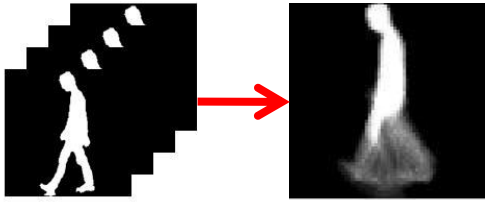
$$H(F,G)=\max\left(h(F,G),h(G,F)\right)$$

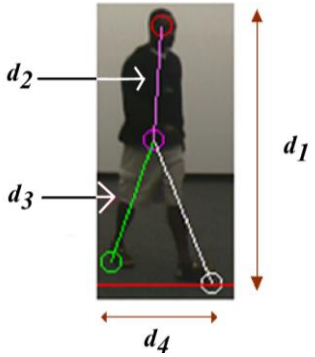
$$h(F,G)=\frac{1}{N_F}\sum_{f\in F}\min_{g\in G}\|f-g\| \qquad h(G,F)=\frac{1}{N_G}\sum_{g\in G}\min_{f\in F}\|f-g\|$$

Remote Biometrics: Approaches

- Detection, normalization, extraction, recognition
- Face recognition
  - Methods:
    - Appearance-based – analyze the face as a whole
    - Model-based – compare individual features (e.g., eyes, mouth)

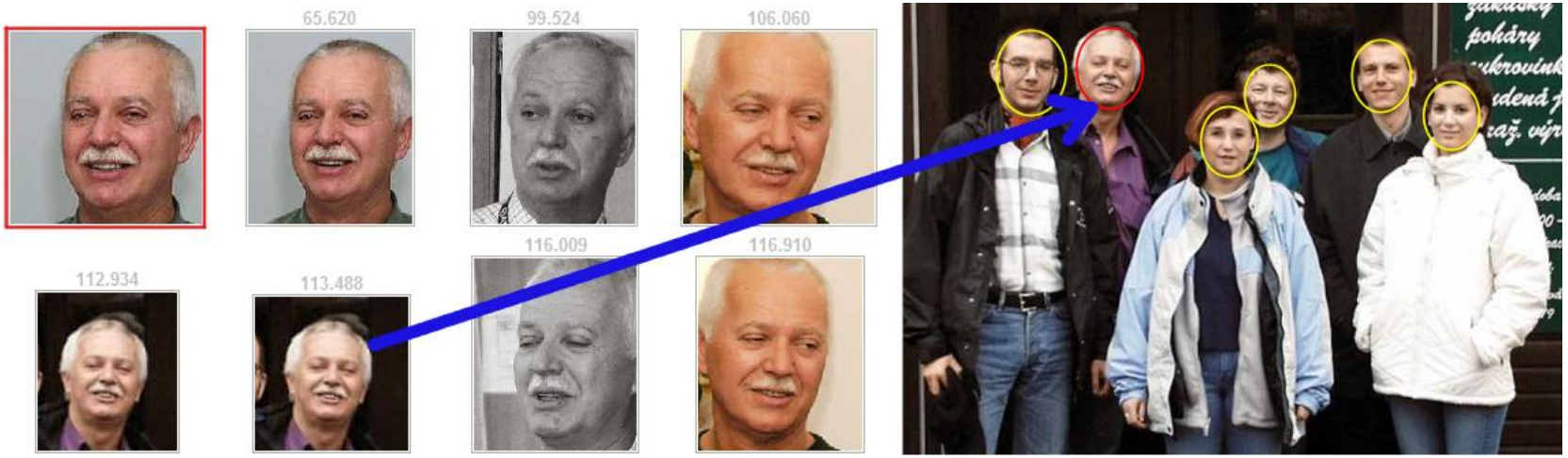
- Gait recognition
  - Less likely to be obscured, low resolution suffices
  - Methods are based on shape or dynamics of the person:
    - Appearance-based – analyze person's silhouettes
    - Model-based – compare features (e.g., trajectory, angular velocity)





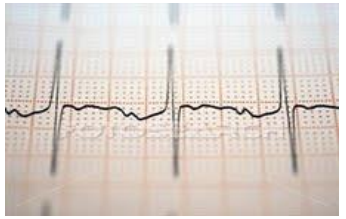
Face similarity

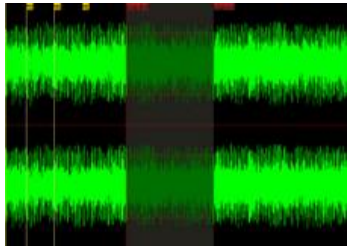
- Face detection
- Face recognition
- face detection – MPEG-7



Signal Processing

- Vast amount of signals produced:
  - Biomedicine data – ECG, CT
  - Biometric data – personal identification
  - Audio data – audio similarity, recognition
  - Sub-image searching
  - Financial time series – analysis, forecasting
  - Time series streams
- Demand for
  - a graceful handling of this data
  - flexible reactions to new application needs





Search – the goals

1. We search to get **results**
2. We ask to find **answers**
3. We use filters so that the right staff **finds us**
4. We **browse** while wandering and way-finding in restricted space

- In reality, we move fluidly between modes of

ask, ***browse***, ***filter***, and ***search***

Search – the traditional way

- Defined by software
- Buy engine, then figure out what it is good for

- It often fails because
  - It is not easy to use
  - It is not able to handle needed content types

Search – some quantitative facts

- 85% of all web traffic comes from search engines
- 450+ million searches/day are performed in North America alone
- 70%+ of all searches are done on Google sites

Search is the **most popular** application

(second to E-mail??)

Search – the best first

- 60% of searchers NEVER go past 1st page of search results
- The top three results draw 80% of the attention
- The first few results inordinately influence query reformulation.

Search - as an interaction

- When we search, our next actions are reactions to the stimuli of a previous search
- What we find is changing what we seek
- In any case, search must be:

fast, ***simple***, and ***relevant***

Search – basic components

- Elements of global search:

Users – goals, psychology, behavior

Interface – interaction, affordances, language

Engine – features, technology, algorithms

Content – indexing structure, metadata

Creators – tools, process, incentives

Search – changes our cognitive habits

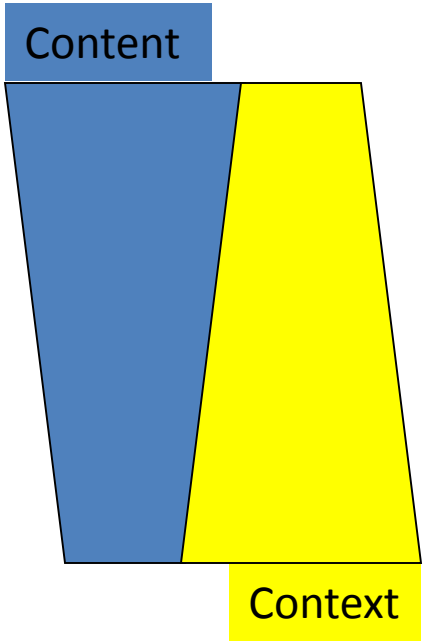
- Assuming information continually and instantaneously available on the web:
  1. We are increasingly handing off the job of remembering to search engines
  2. When we need answer, we do not think, we go immediately to a nearest Web connection
  3. When we expect information to be easily found again, we do not remember it well
  4. Our original memory of facts is changing to a memory of ways to find the facts

Users and their Intent

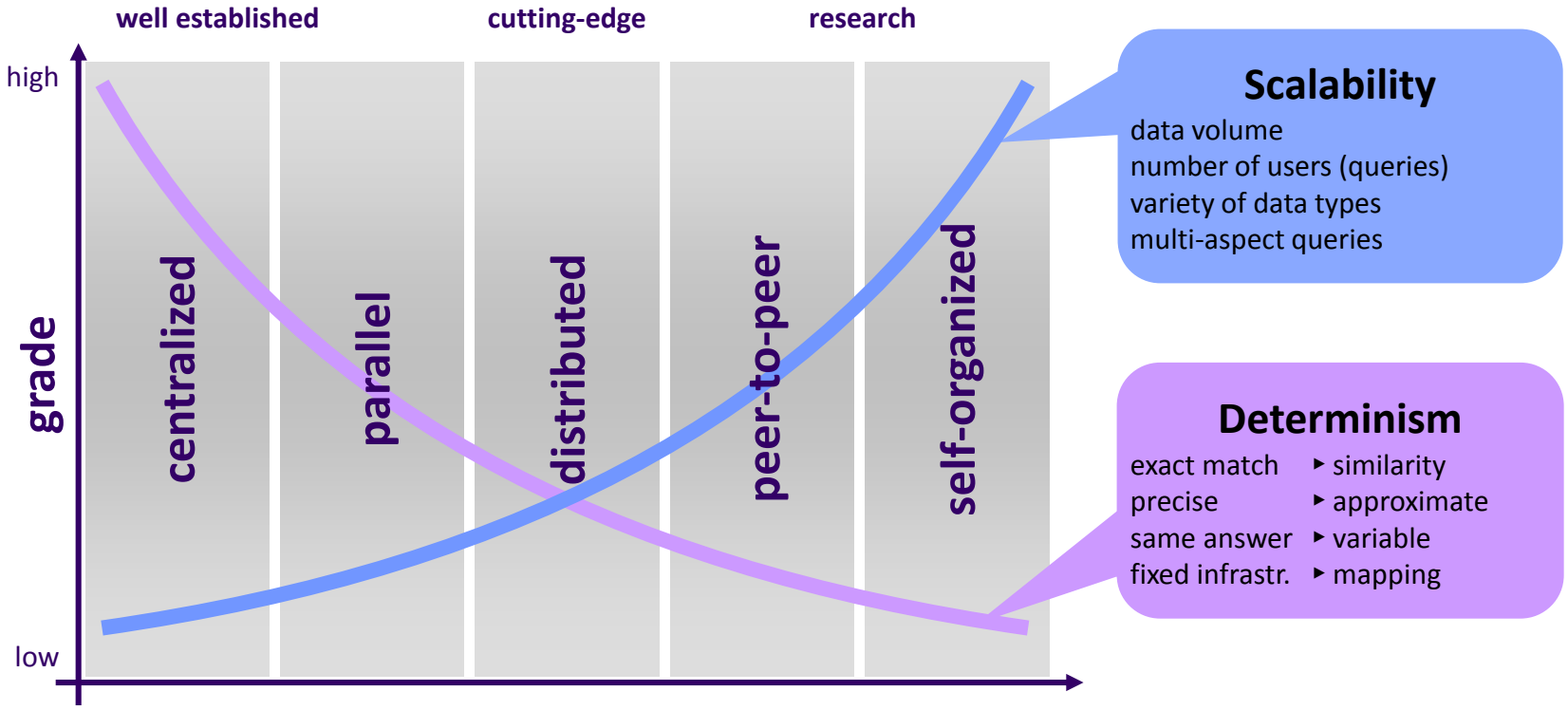
Search is **subjective** and also depends on **visual** and **emotional** attributes, e.g. *shocking*, *funny*, etc.

- **Browser**
  - not clear end-goal; series of unrelated searches; jump across unrelated topics; expects surprises and random search hints
- **Surfer**
  - moderate clarity of end-goal; exploratory actions at the beginning; e.g. planning a holiday
- **Searcher**
  - very clear about what is searching for; completeness and clarity of results are important

Prevalent strategy

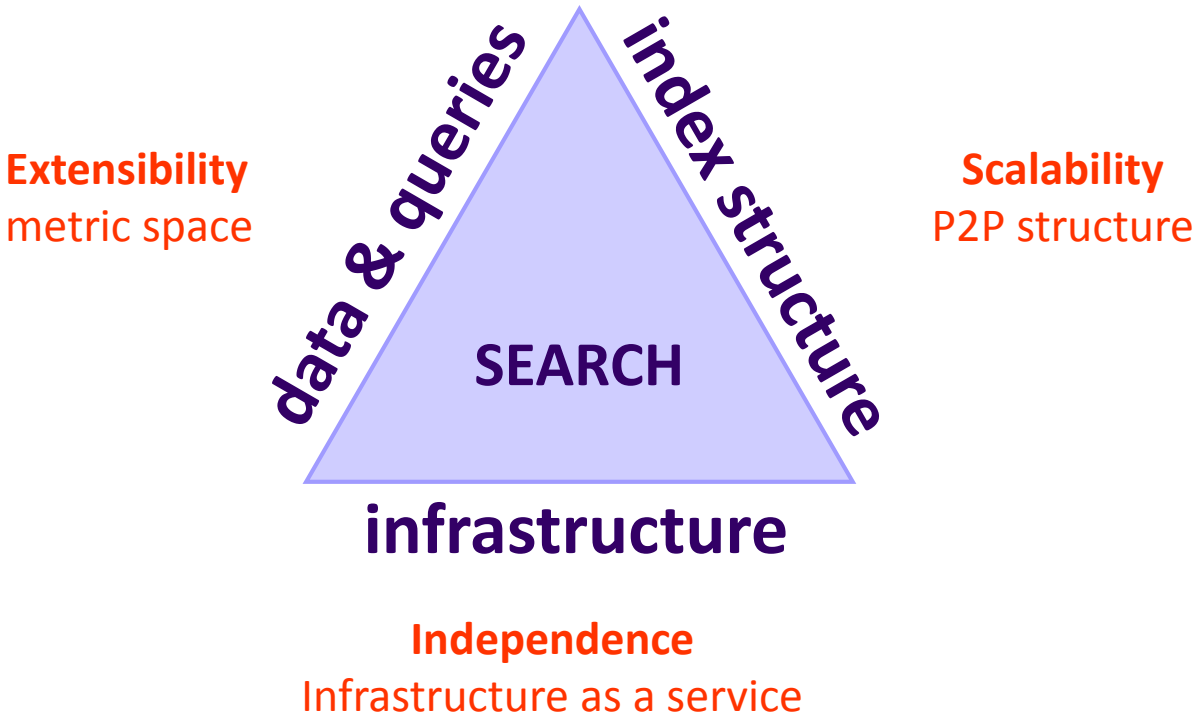


Evolution of Search Engine Strategies



The MUFIN Approach

MUFIN: MUlti-Feature Indexing Network



Infrastructure Independence: MESSIF

Metric Similarity Search Implementation Framework

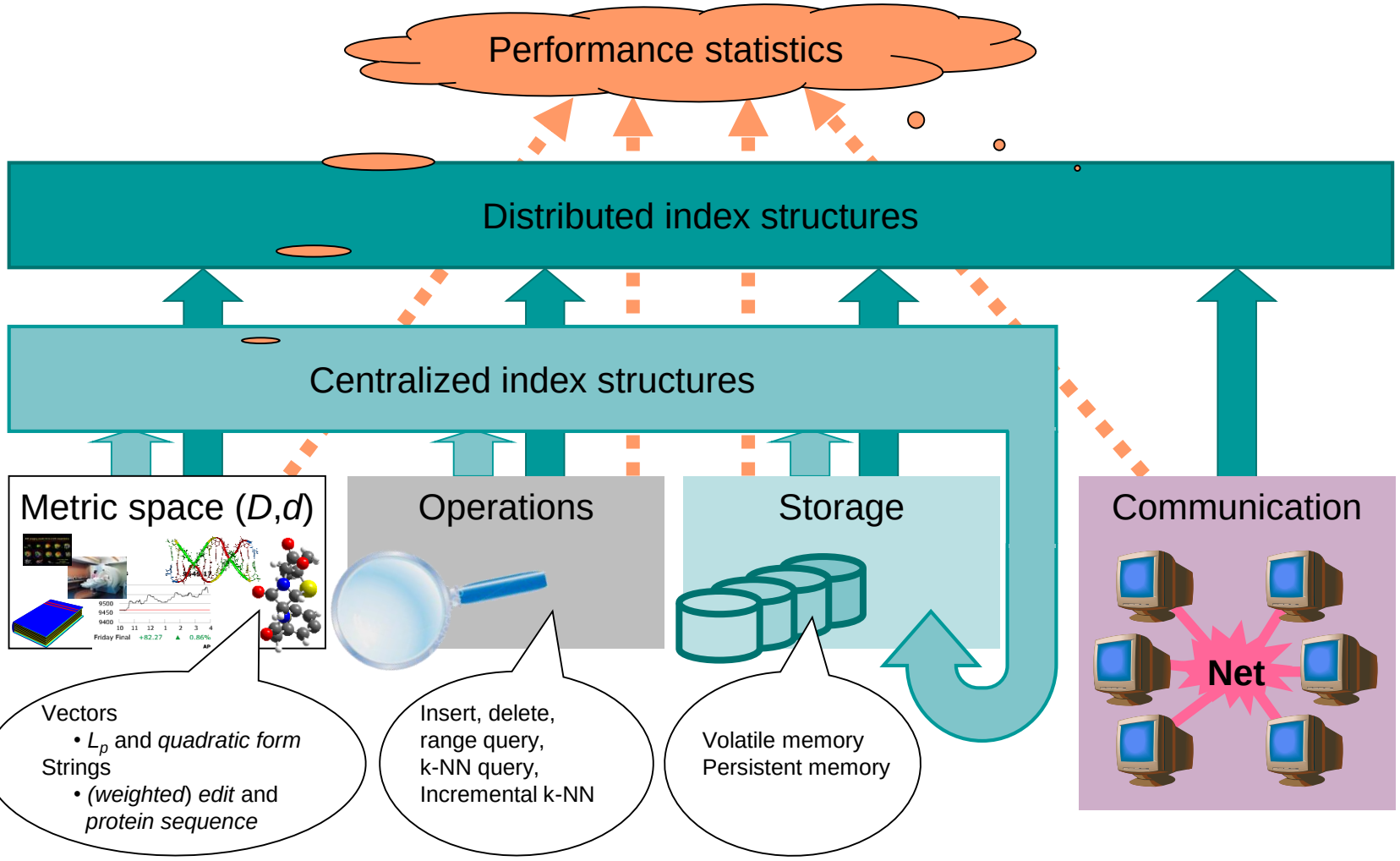


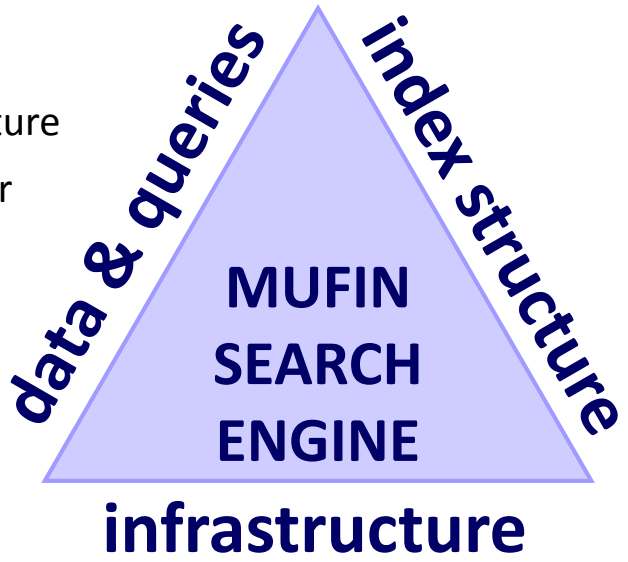
Image Search Demo

http://mufin.fi.muni.cz/imgsearch/

Extensibility

COPHIR

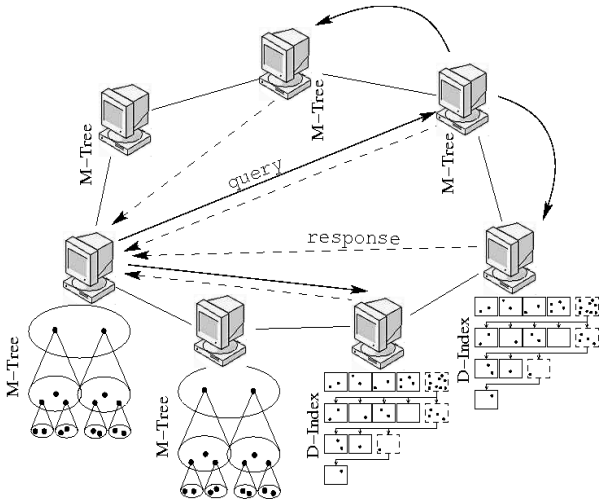
- color structure
- scalable color
- color layout
- edge histogram
- homogeneous texture



6 x IBM server x3400 – 2 servers used

Scalability

M-Chord + M-Index



MUFIN demos

- http://mufin.fi.muni.cz/imgsearch/similar
- http://www.pixmac.com/
- http://mufin.fi.muni.cz/twenga/random
- http://mufin.fi.muni.cz/fingerprints/random
- http://mufin.fi.muni.cz/subseq/random
- http://mufin.fi.muni.cz/mma-faces-extended/
- http://mufin.fi.muni.cz/plugins/annotation

Limitations: Data Types

We know

- Attributes
  - Numbers, strings, etc.
- Text (text-based)
  - Documents, annotations

We need

- Multimedia
  - Image, video, audio
- Security
  - Biometrics
- Medicine
  - EKG, EEG, EMG, EMR, CT, etc.
- Scientific data
  - Biology, chemistry, physics, life sciences, economics
- Others
  - Motion, emotion, events, etc.

Limitations: Models of Similarity

We know

- Simple geometric models, typically vector spaces (metric spaces)

We need

- More complex model
- Non metric models
- Asymmetric similarity
- Subjective similarity
- Context aware similarity
- Complex similarity
- Etc.

Limitations: Queries

We know

- Simple query
  - Nearest neighbor
  - Range

We need

- More query types
  - Reverse NN, distinct NN, similarity join
- Other similarity-based operations
  - Filtering, classification, event detection, clustering, etc.
- Similarity algebra
  - May become the basis of a “Similarity Data Management System”

Limitations: Implementation Strategies

We know

- Centralized or parallel processing

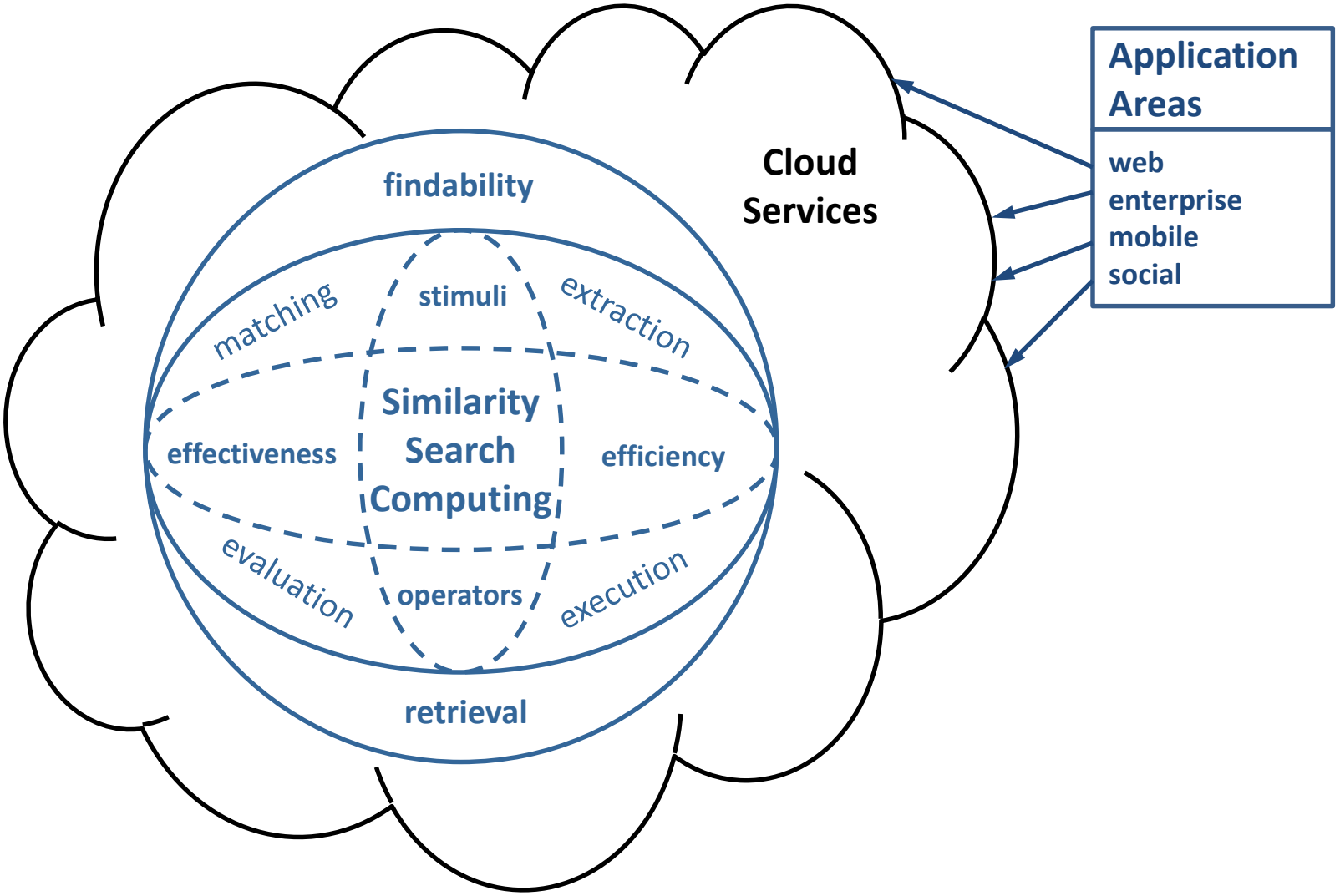
We need more

- Scalable and distributed architectures
- MapReduce like approaches
- P2P architectures
- Cloud computing
- Self-organized architectures
- Etc.

Problems with Current Applications

1. Current applications are implemented as *complex software projects*; it is costly, highly qualified specialists are needed
2. They fitfully need *massive infrastructure* to build and run multiple indices
3. Applications are much more complex than a *search*, which is an important supporting *service*

Similarity Data Management System



Similarity Searching in Clouds

- Retrieval – effectiveness, evaluation, operations, execution, and efficiency
- Findability – effectiveness, matching, stimuli, extraction, and efficiency
- Cloud way of computing:
  - Scalability – must balance load across servers and avoid bottlenecks
  - Elasticity – to allow adding (reducing) capacity to a running system
  - Availability – to provide high levels of usability and fault tolerance
  - Privacy – to safeguard data that is valuable or sensitive against unauthorized access

Five Years of SISAP Conferences

strong points

- Organized every year
- SISAP home page
- A workshop has turned into an international conference
- We are leaders in metric search theory and technology
- Prestigious conf. proceedings publisher (LNCS)
- Journal publications from SISAP conferences

weak points

- Too few submissions
- Narrow and closed community
- Few application papers
- No contacts to related scientific events – e.g. tutorials
- We are little interested in actual search systems
- No cooperation with industry

Next Years of SISAP Conferences

Questions to be discussed:

- Do we need a change?
- Do we want to change?
- What is to be changed and how?
- Are we able to do it?
- Who is willing to do the work?
- … ?