Laboratory of Data Intensive Systems and Applications

DISA

Faculty of Informatics, Masaryk University, Brno

Pavel Zezula - presenter

Content of the talk

- Similarity in our lives and digital data processing
- The metric space model of similarity
- Content based similarity search and feature extraction
- DISA contribution – research, results, awards
- Applied multidisciplinary research
- SWOT a future research directions

DISA members

- Staff:
  - Michal Batko
  - Petra Budikova
  - Vlastislav Dohnal
  - Vladimir Mic
  - Jan Sedmidubsky
  - Pavel Zezula
- Former members: Petr Elias, Filip Nálepa, David Novak, Jakub Valcik

- Current PhD students:
  - Miriama Janosova
  - Iris Kico
  - Jakub Peschel
  - Terezia Slaninakova

Plus, about 10 bachelor and master students

Similarity in our Lives

quotations from the social psychology literature

- 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*, ...) is **fundamental** to theories of *perception*, *learning*, *judgment*, etc.

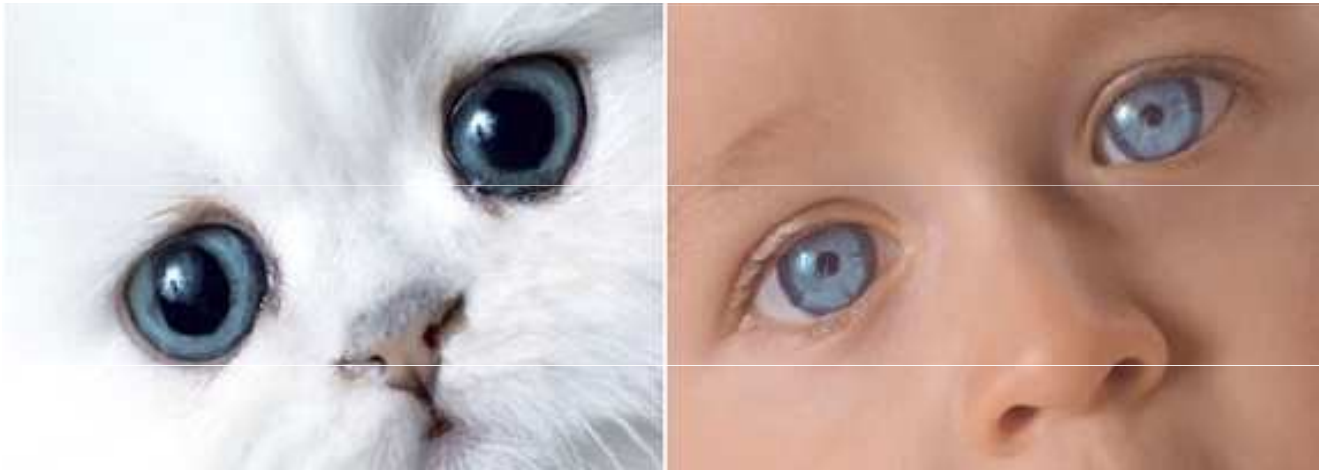
Similarity is **subjective** and **context-dependent**

Real-life Similarity

- Are they similar?

Real-life Similarity

- Are they similar?



Real-life Similarity

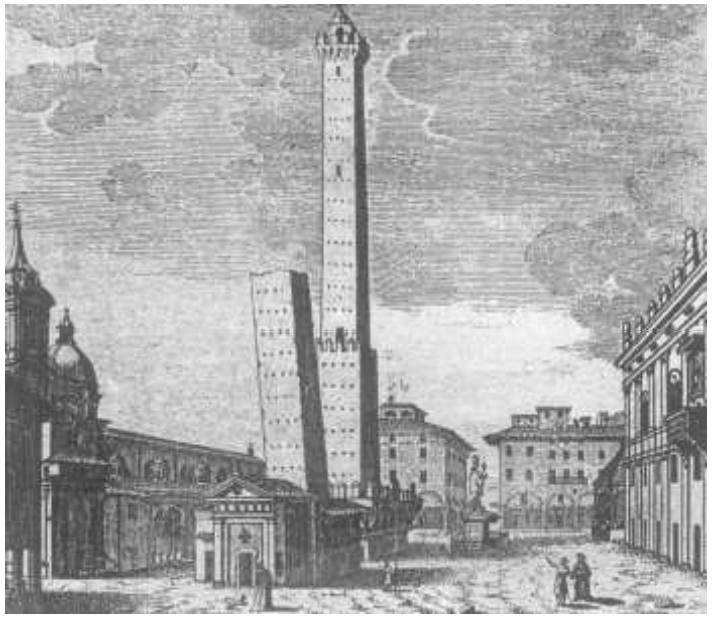
- Are they similar?

Real-life Similarity

- Are they similar?

Prototypicality or Centrality

not symmetric



Context/Data/Environment Dependent

circumstances alter similarities

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**.

Majority of current data is **unstructured**

possibly only structured on display

Challenge

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

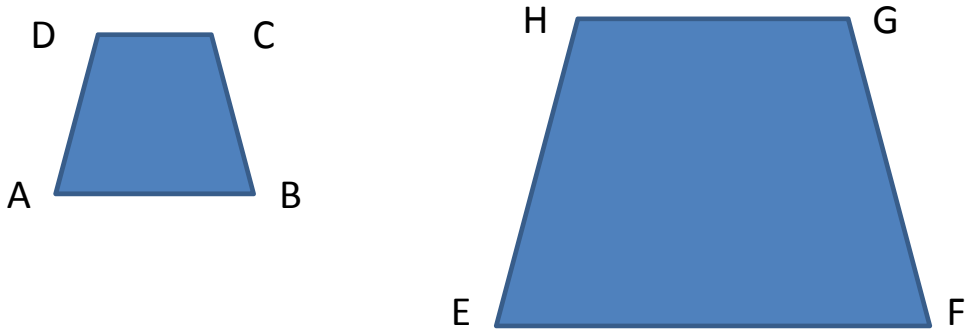
- **Similarity data management** is needed to *connect*, *search*, *filter*, *merge*, *relate*, *rank*, *cluster*, *classify*, *identify*, or *categorize* objects across various collections.

WHY?

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

We learned from School

- GEOMETRY:
- Two polygons are **similar** to each other, if:
  1) Their corresponding angles are **congruent**
     - ∠A = ∠E; ∠B = ∠F; ∠C = ∠G; ∠D = ∠H, and
  2) The lengths of their corresponding sides are **proportional**
     - 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 also **similar** to the third polygon.
- In any case:

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

Metric Space:
A Geometric Model of Similarity

- Metric space: $\mathcal{M} = (\mathcal{D}, d)$
  - $\mathcal{D}$ – domain
  - distance function $d(x,y)$

    $\forall x,y,z \in \mathcal{D}$

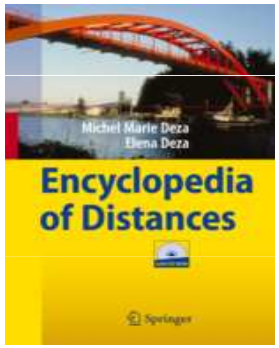
    - $d(x,y) > 0$ - *non-negativity*
    - $d(x,y) = 0 \Leftrightarrow x = y$ - *identity*
    - $d(x,y) = d(y,x)$ - *symmetry*
    - $d(x,y) \leq d(x,z) + d(z,y)$ - *triangle inequality*

Examples of Distance Functions

- L_p **Minkovski distance** (for vectors)
  - L_1 – city-block distance
  - L_2 – Euclidean distance
  - L_∞ – infinity
- **Edit distance** (for strings)
  - minimal number of insertions, deletions and substitutions
  - d('application', 'applet') = 6
- **Jaccard's coefficient** (for sets A,B)

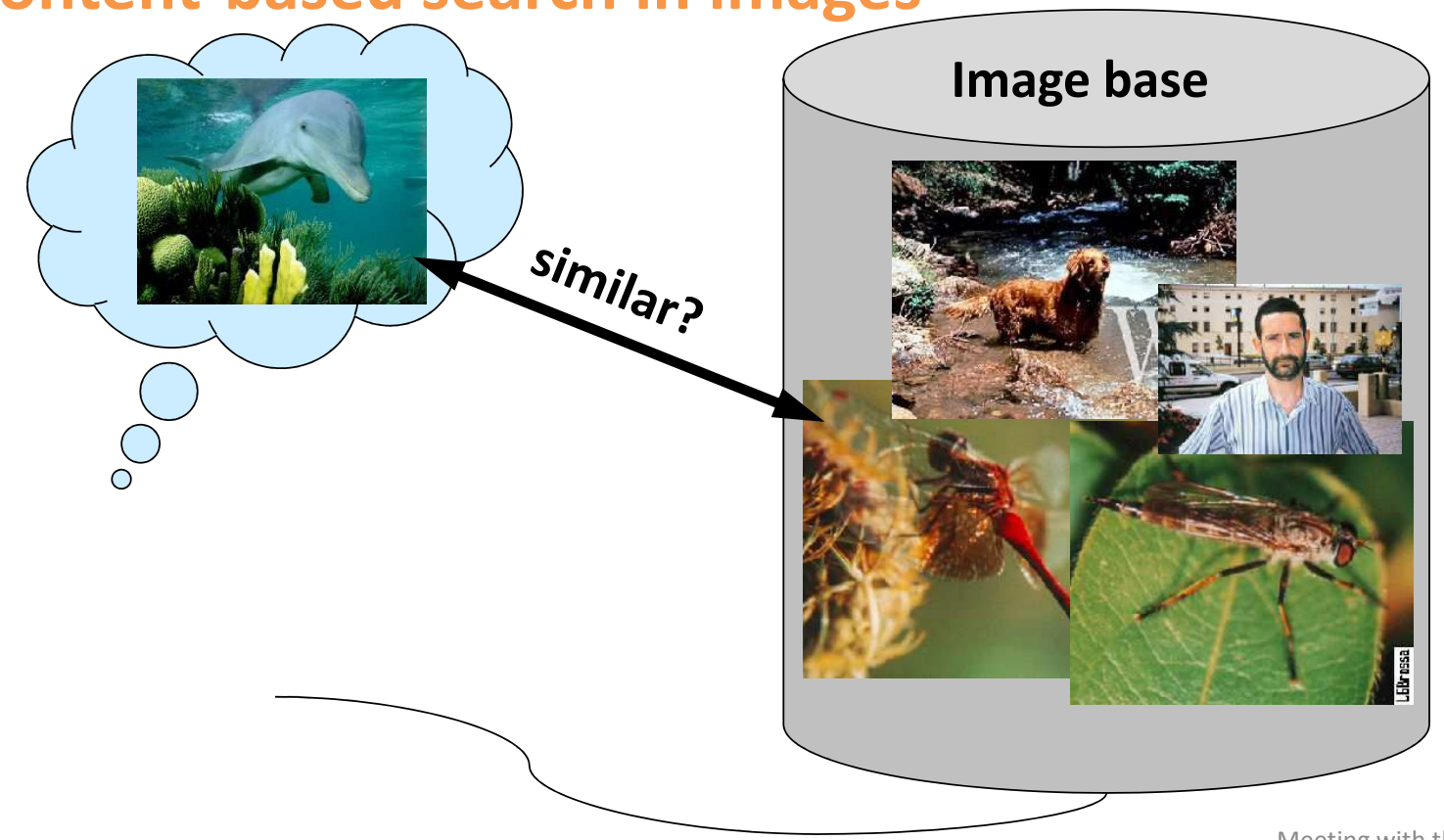
Examples of Distance Functions

- **Mahalanobis distance**
  - for vectors with correlated dimensions
- **Hausdorff distance**
  - for sets with elements related by another distance
- **Earth movers distance**
  - primarily for histograms (sets of weighted features)
- and many others

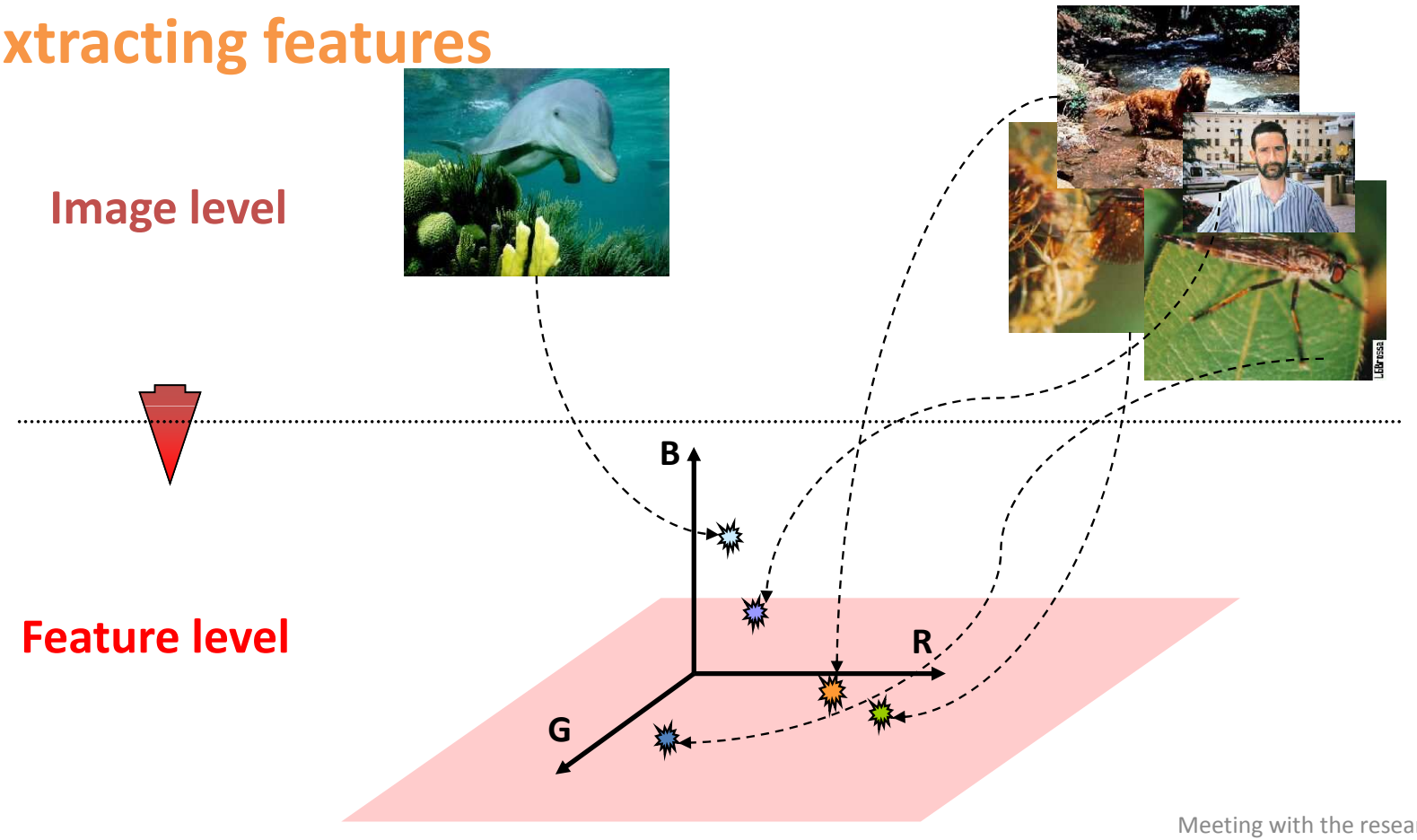


Content-Based Search Objectives

Content-based search in images

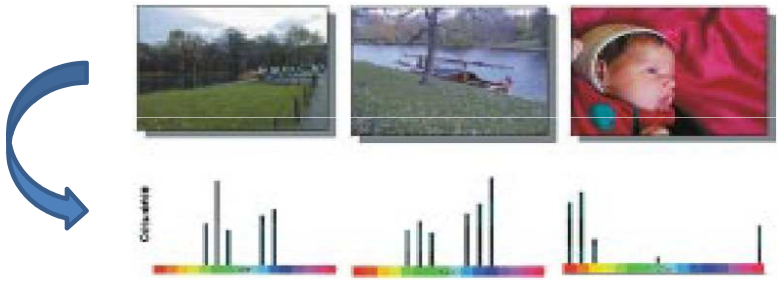


Content-Based Search Implementation



MPEG-7

- Multimedia Content Descriptors Standard ~ 2000
- Global feature descriptors:
  - Color, shape, texture, …



  - One high-dimensional vector per image and feature
  - Minkovski distance used

Visual Similarity

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

Visual Similarity - finding correspondence

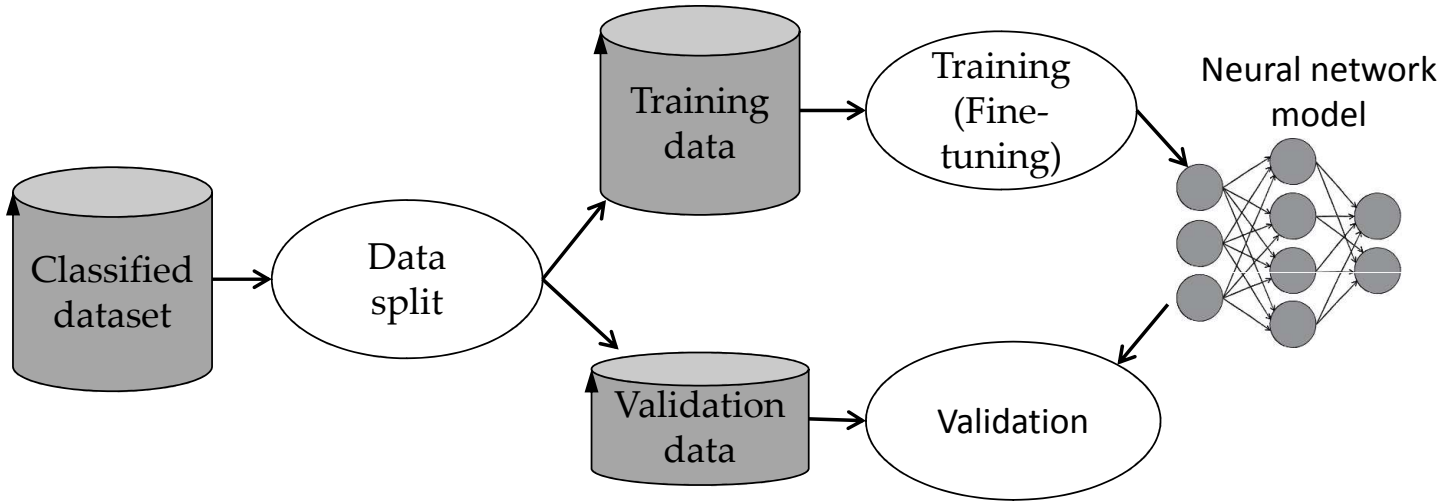
Biometrics: Fingerprint

- 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

Multiple Visual Aspects

Contemporary Approaches to Feature Extraction – Metric Learning

- Neural networks technology
  - Convolutional Neural Networks (CNN)
  - Recurrent Neural Networks (RNN)



Similarity Search Problem

For $X \subseteq \mathcal{D}$ in metric space $\mathcal{M}$,
pre-process X so that the similarity queries
are executed efficiently.

Implementation problems:

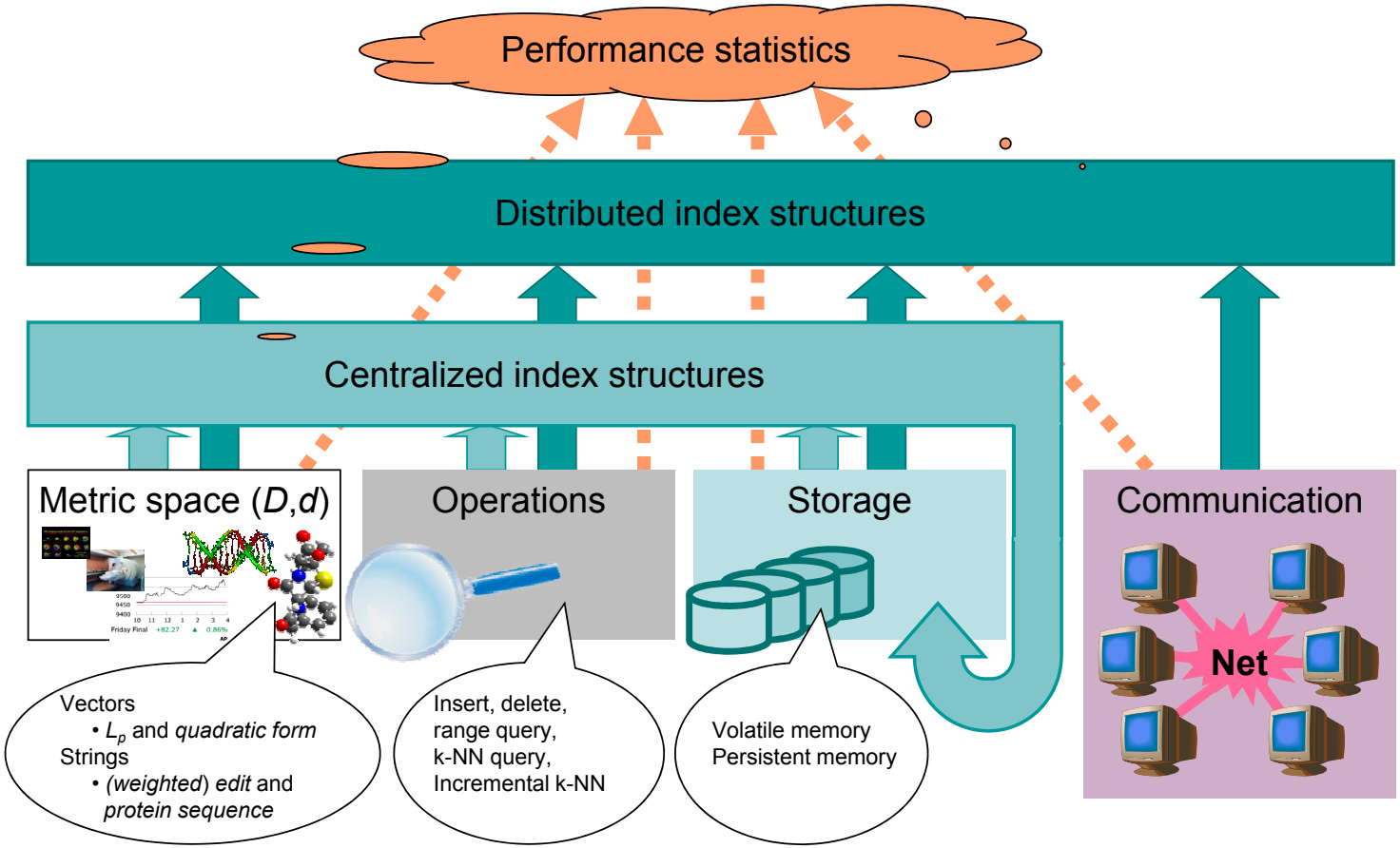
- How to **partition** the data to reduce search space
- How to ask questions - definition of **queries**
- How to **execute** queries – to achieve required performance

The challenge:

In metric space, no total ordering exists!

MESSIF - Metric Similarity Search Implementation Framework

Infrastructure independent



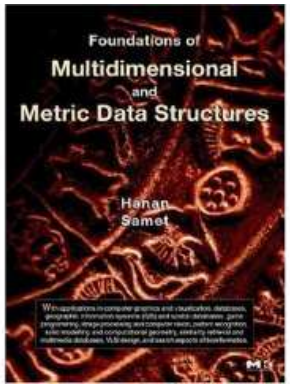
DISA Contribution – grants and partners

- Large spectrum of contributing grants:
  - Academic vs. Industrial
  - National vs. European
  - Focused research vs. Network of Excellence
- Significant cooperating partners:
  - academic (including Max Plant Institute, ETH Zurich, CNR Italy, NII Tokyo, University of St. Andrews, University of Bologna, plus tens of other universities in Europe within networks of excellence)
  - industrial (including IBM Research, Telenor, Telecom Spain, Bull, Athena Security Israel, XEROX SAS Grenoble, Konica-Minolta)

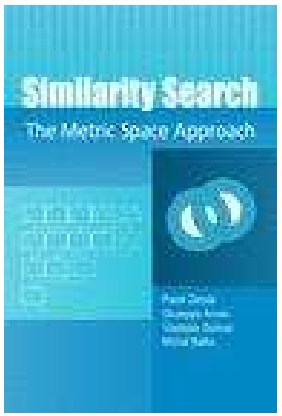
Scientific Achievements

- Most cited works:
  - M-tree 2550; Metric book 1250
- Advanced publication platforms:
  - VLDB, ACM SIGMOD-PODS, ACM SIGIR, ACM TODS, ACM TOIS, VLDB Journal
- Tutorials:
  - ACM SAC, ACM Multimedia, ICMR, ESMAC
- Invited talk and key-notes:
  - ACM SIGIR, ADBIS, MMM, IEEE ISM, SOFSEM, SEDB
- Best paper awards:
  - DEXA, IEEE ISM, SISAP

Textbooks on
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, 2005

Teaching material:
http://www.nmis.isti.cnr.it/amato/similarity-search-book/

SISAP International Conferences

SISAP (Similarity Search and Applications)

International conference series (http://sisap.org/)

2008 Cancun Mexico

2009 Prague Czechia

2010 Istanbul Turkey

2011 Lipari Italy

2012 Toronto Canada

2013 A Coruña Spain

2014 Los Cabos Mexico

2015 Glasgow UK

2016 Tokyo Japan

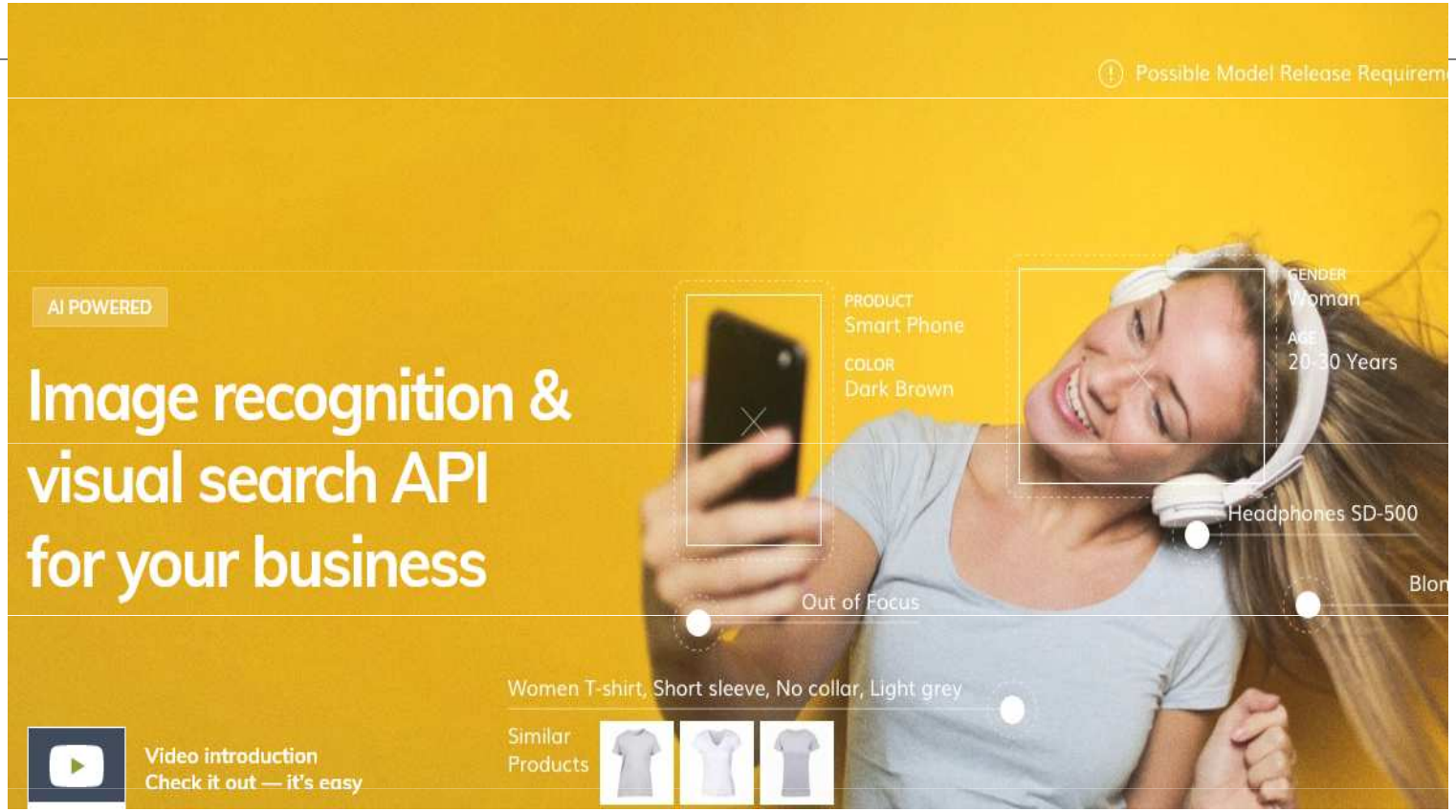
2017 Munich Germany

2018 Lima Peru

2019 Newark NJ USA

XIMILAR – Image Recognition and Visual Search

https://www.ximilar.com/



Appreciation - Awards

- IBM SUR (Shared University Research) Award for

  "Web-scale Similarity Search in Multimedia Data"

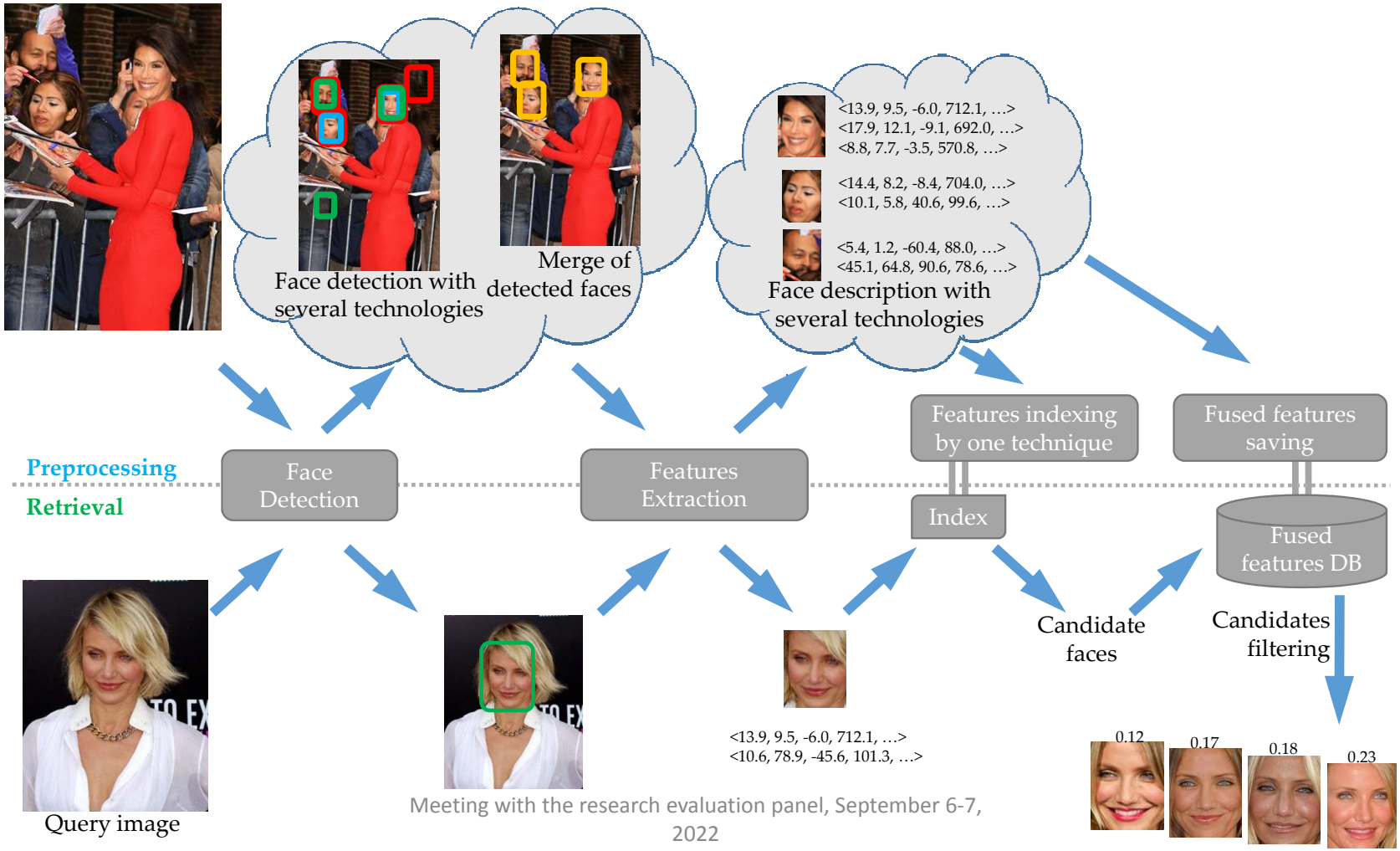
- Top 27 IT Personalities in Czech Republic – Computerworld Magazine

- MU Brno Rector's price 2X

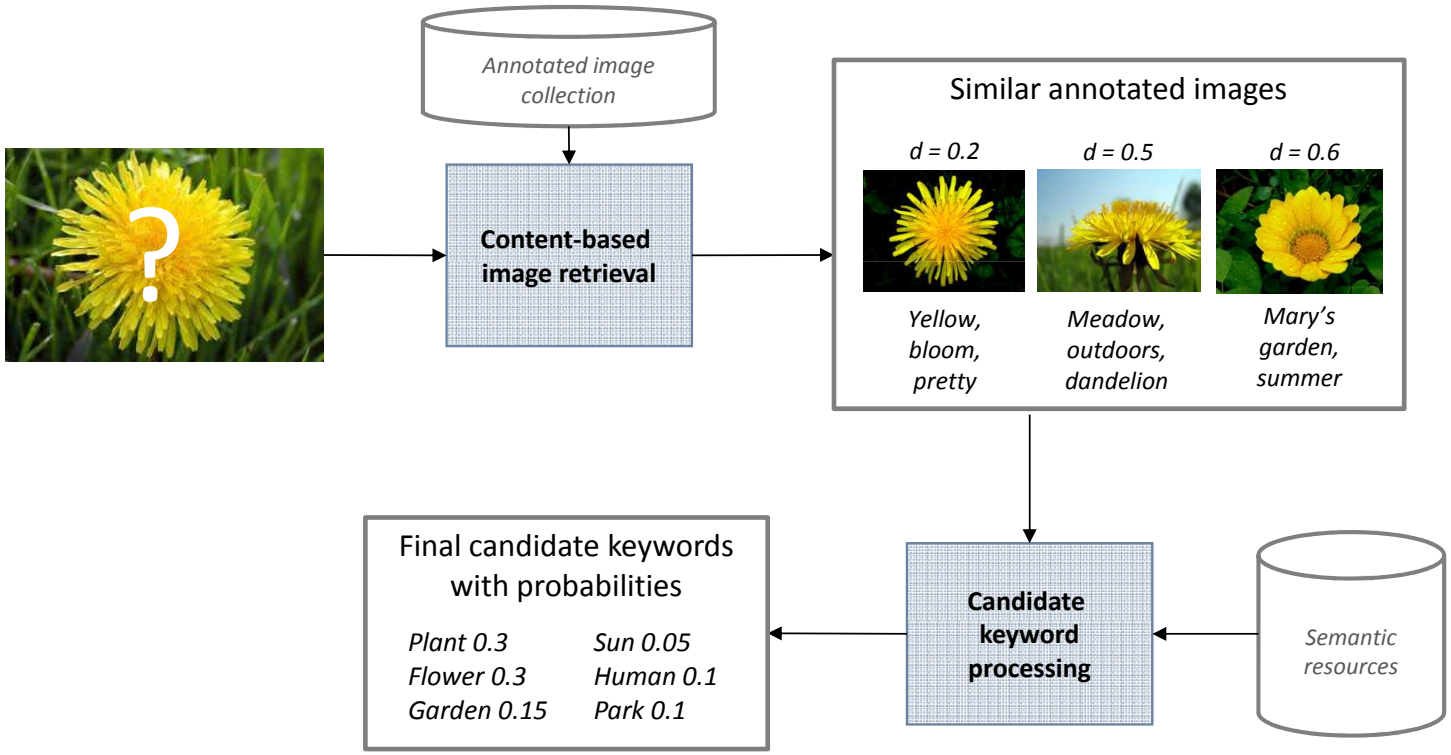
Application Research

- Face Retrieval
- Image annotation
- Motion data management
- Improving Treatments in Cerebral-Palsy
- Protein Similarity Search
- Dyslexia detection

Similarity Search in Collections of Faces



Search-based annotation principles



Example

1. Retrieve 100 similar images from Profiset
2. Merge their keywords, compute frequencies
3. Build the semantic network using WordNet
4. Compute the ConceptRank
5. Apply post-processing & return 20 most probable keywords

Candidate keywords after CBIR
church, architecture, travel, europe, building, religion, germany, buildings, north, churches, christianity, america, religious, exterior, st, historic, world, tourism, united, usa, …

Semantic network
4 relationships: hypernym *(dog → animal)*, hyponym *(animal → dog)*, meronym *(leaf → tree)*, holonym *(tree → leaf)*
270 network nodes, 471 edges

ConceptRank scores
building (2.53), structure (2.41), LANDSCAPE (2.10), BUILDINGS (1.87), OBJECT (1.84), NATURE (1.78), place_of_worship (1.75), church (1.74), Europe (1.68), religion (1.64), continent (1.51), …

Final keywords
building, structure, church, religion, continent, group, travel, island, sky, architecture, tower, person, belief, locations, chapel, christianity, tourism, regions, country, district

Digitization of Human Motion

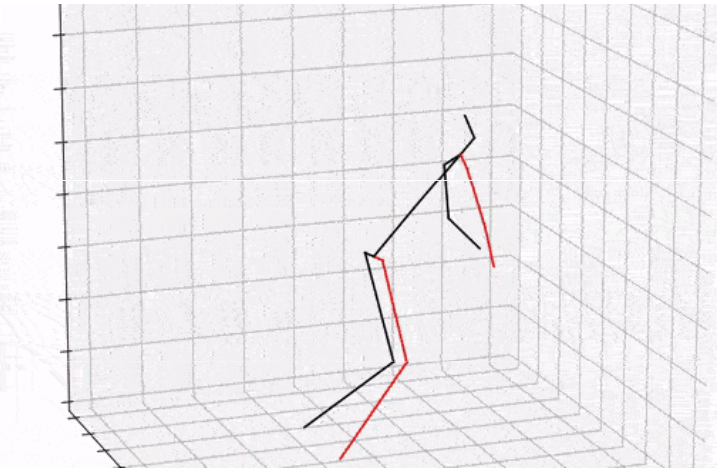
Skeleton-data representation

- Simplified spatio-temporal representation of human motion
  - Sequence of 3D skeletons ~ a set of 3D trajectories of body joints
- Better structured and easier to store than video-based representation

Video-based representation

Source: https://blog.usejournal.com/3d-human-pose-estimation-ce1259979306

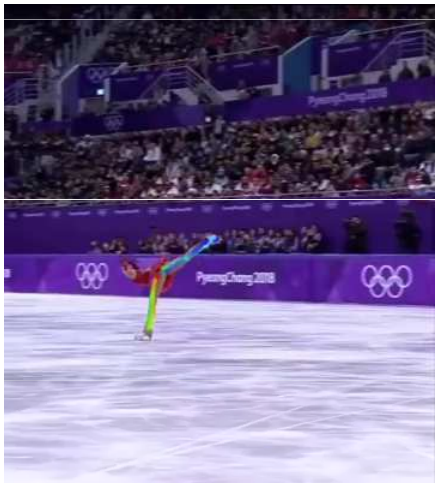
Skeleton-based representation



Great Application Potential

A wide variety of possible applications

- Sports – digital referees assessing the quality of performance
- Virtual reality – recognizing player movements in real time
- Smart-cities – detecting falls of persons crossing a street
- Healthcare – evaluating the rehabilitation progress remotely



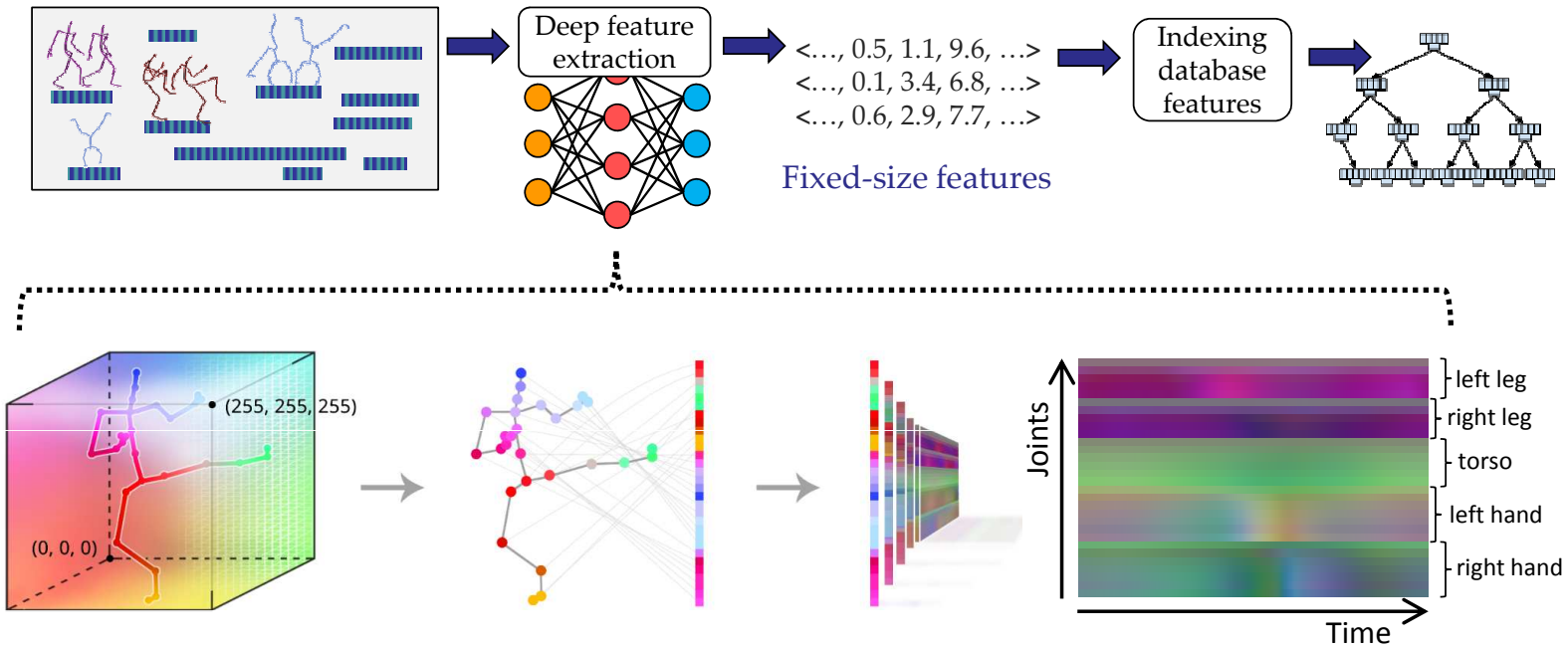
Source: https://www.youtube.com/watch?v=5cI-JibDEMA

Source: https://blog.usejournal.com/3d-human-pose-estimation-ce1259979306

Content-based Processing

Query-by-example searching

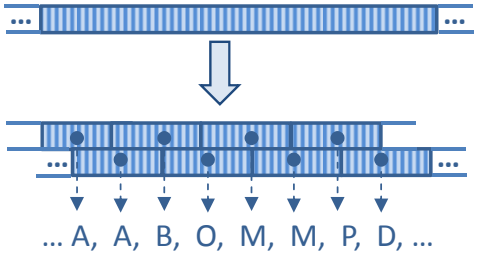
- Transforming complex motions to fixed-size vectors and indexing them by metric-space search methods

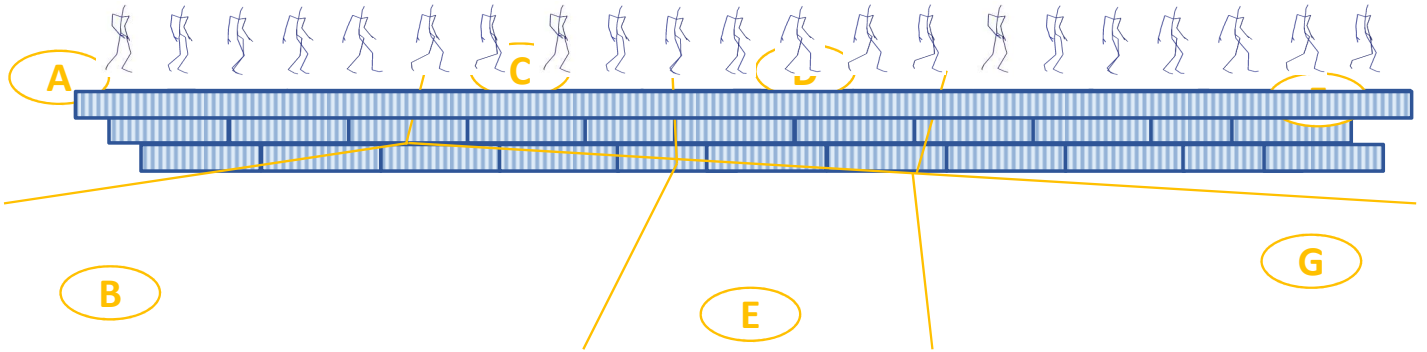


[Sedmidubsky, J., Elias, P., Zezula, P.: Effective and Efficient Similarity Searching in Motion Capture Data. Mult. Tools and Apps. 2018]

Motion Words – idea

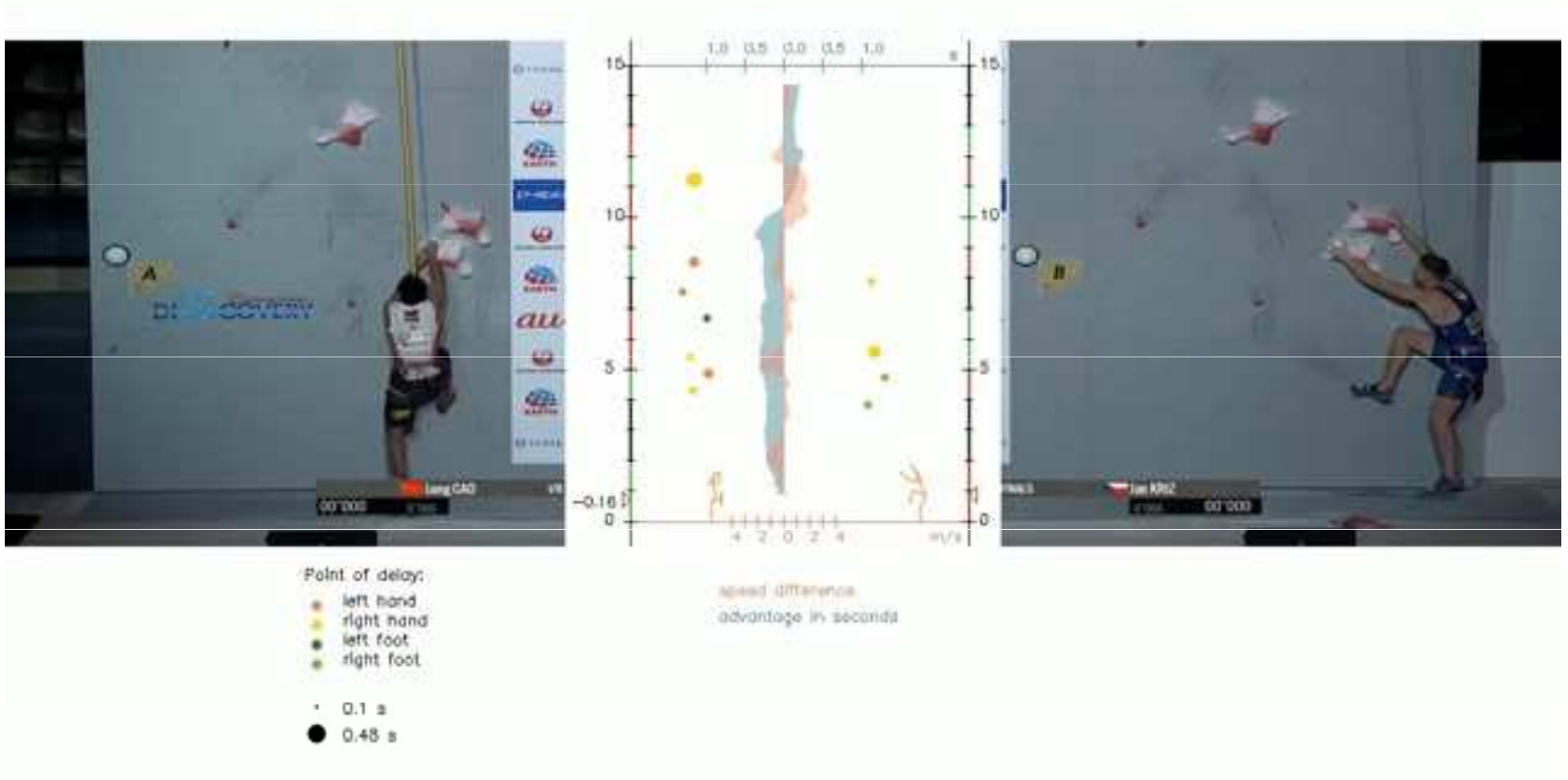
- Cut motion into short, overlapping segments
- Quantize the segment space
- Represent original sequence by identifiers of quantized segments





Content-based Analysis

Comparison of speed-climbing performances



Source: https://www.youtube.com/watch?v=tdxMo11KJGk&t=258s

Similarity Search in Protein Chains

Each protein consists of 1 or more subparts – *protein chains*

Approx. 500,000 chains are known – *Protein Data Bank (PDB)*

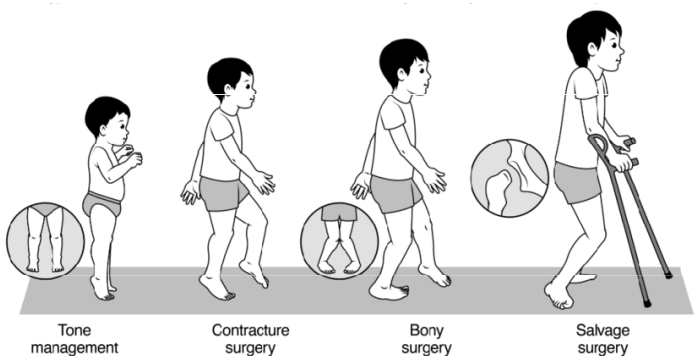
3D models of protein chains are used to define their pairwise similarity

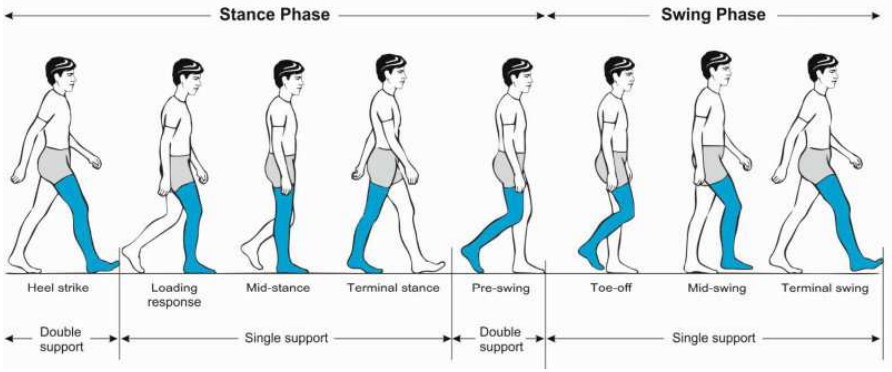
- Similarity evaluation time strongly depends on the size of compared chains
- Distance evaluation time ranges from ms to min.

Model of a protein chain: balls ≈ atoms, sticks ≈ bonds between atoms. Green ribbon ≈ simplification of the main atoms

Recent Applied Research Project #1

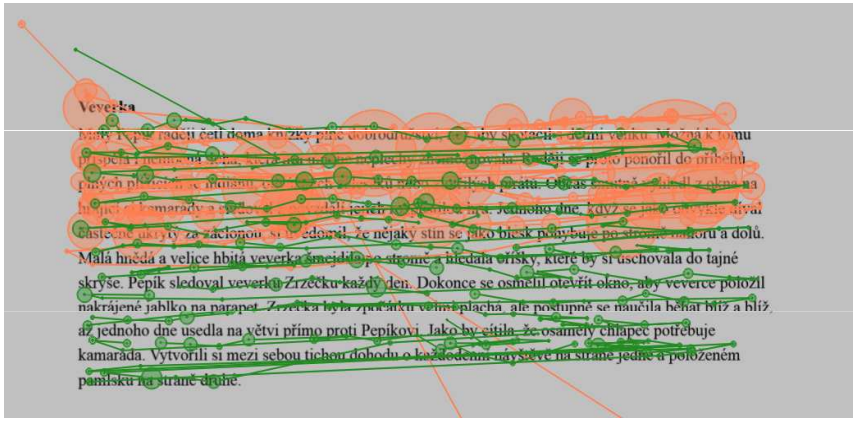
- Project scope:
  - Improving Treatments in Cerebral-Palsy Children using Artificial Intelligence (2020–2022)
  - Cooperation with Children Hospital Brno
  - Main objective – estimate whether a given treatment is suitable for a new child patient suffering from the cerebral-palsy disease
  - Solution – searching for similar gait cycles recorded in the pre-surgery phase and comparing the quality of walking between the pre-surgery and post-surgery phases





Recent Applied Research Project #2

- Project scope:
  - Diagnosis of Dyslexia using Eye-Tracking and Artificial Intelligence (2021–2023)
  - Cooperation with the Faculty of Arts (Masaryk University) and psychological clinics
  - Main objective – estimate how prone the individual is to the dyslexia disease
  - Solution – classifying spatio-temporal eye-tracking data (and their derived features) of dyslexia/intact patients on text-reading tasks



SWOT Strengths

Similarity plays a **central role** in processing contemporary digital data.

We have a **leading position** in this research - most cited papers and the first monograph in the similarity search domain, organize a conference, spin-off

We **teach** corresponding **courses** (even abroad) and have many successful PhD graduates (including foreigners),

Received **prestigious awards** (e.g. IBM SUR, Computerworld magazine, rector's price),

Participated in many **prestigious** national and international **projects** (e.g. European research – Scholnet, Sapir -, European networks of Excellence – DELOS 2X -, GACR Network of Excellence CEMI),

Cooperated with many academic and industrial institutions

Delivered **invited** and **key-note** speeches at important conferences (e.g. ACM SIGIR, SMAC, ADBIS, MMM, IEEE ISM, SEBD),

based on our similarity search technology a spinoff XIMILAR was created by the group's PhD students,

SWOT - Weaknesses

The group is rather **small** with most researchers exclusively supported from **external resources**.

The researchers are **overloaded with teaching** and often must leave the actual research to students - this typically results in routine work, not the best quality.

The endless **fight for grants** consumes too much time and mental capacity of highly qualified researchers.

Not very efficient **communication** with the faculty management.

SWOT - Opportunities

Many open questions/**problems remain** in the similarity search domain thus additional fundamental research is needed – e.g., context dependent, subjective, and adaptable similarity search or explainable similarity data models for AI.

The potential **application area is huge** and opens additional research areas – in medicine, sports, security, game industry, etc.

We can **capitalize on our previous results** in the motion data processing and similarity management in general.

SWOT - Threats

To bring up qualified researchers takes years, but you can lose a skilled person very fast when you do not get grant support in time. Such a system repeatedly alternates periods of too much money for available staff and not enough money for existing staff - making **qualified researchers redundant.**

With the increasing quantity and importance of evaluation indicators, the danger is that researchers will concentrate more on **complying with** required **indicators** rather than the quality of their research work.

Students are not motivated to **study PhD** – it is easy to get a high paid job without a PhD degree.

Our Vision - Future Research Challenges

Challenge No.1 (adaptability):

Respecting continuously changing distance metric – searched collection size as well as up to date collection of known samples – continuously adapt the search indexing mechanisms.

Challenge No. 2 (explainability):

Respecting an application domain – e.g. motion capture data – provide explanation tools that might be requested on demand. Similarity cracks the code of explainable AI.

Research Projects

- Selected basic-research projects:
  - Center of Excellence on Multi-modal Data Interpretation on a Very Large Scale (GBP103/12/G084); Czech Science Foundation (GAČR); 2012–2018
  - Searching, Mining, and Annotating Human Motion Streams (GA19-02033S); Czech Science Foundation (GAČR); 2019–2021

- Selected applied-research (application-oriented) projects:
  - Efficient Searching in Large Biometric Data (VG20122015073); Ministry of the Interior of the Czech Republic; 2012–2015
  - Improving Treatments in Cerebral-Palsy Children using Artificial Intelligence (MUNI/G/1585/2019); GAMU (Interdisciplinary projects); 2020–2022
  - Diagnosis of Dyslexia using Eye-Tracking and Artificial Intelligence (TL05000177); Technology Agency of the Czech Republic (TAČR); 2021–2023