Compromise

Data compression paradigm based on omitting self-evident information

Maribor, 07. 11. 2022

David PODGORELEC & GeMMA Team

University of Maribor

Faculty of Electrical Engineering and Computer Science

Institute of Computer Science

Laboratory for Geospatial Modelling, Multimedia and Artificial Intelligence

Preface – history of cooperation

- UWB and UM collaborate for nearly 30 years.
  - Skala and Žalik first, Kolingerová since early beginnings, Podgorelec since late 90s, Kohout around 2000…
  - Series of bilateral projects with many visits, presentations, discussions, several joint publications, and having fun.
  - Several papers of UM at WSCG (particularly before 2010).
- UWB invited UM to GeoSym at the beginning of 2020.
  - 3-year research project funded by GAČR (lead agency) and ARRS.
  - Successfully approaching the end of the 2nd year.
- UM returned the invitation with Compromise.

Preface – why Compromise

- Attempt to do some fundamental research („for the soul“) after a series of ARRS-funded „applied“ projects.
- We have always liked data compression, we have had some promising results, but we have never had the funding to continue the work.
- More chances to get a bilateral project funded by two agencies and, of course, there is a gratitude for GeoSym.

Project ID card

- Funded by ARRS (lead agency) and GAČR.
  - Approved by ARRS in September 2022, hopefully soon by GAČR.
  - Start in Slovenia 01. 11. 2022, hopefully soon in Czech Republic.
- 3 years
  - ARRS: 300,000.00 € 59.43% 3069 hours (per year)
  - GAČR: 204,805.92 € 40.57%
- Leaders:
  - Borut Žalik (borut.zalik@um.si) and Ivana Kolingerová
- Support:
  - Administrative: David Podgorelec (david.podgorelec@um.si)
  - Technical: Andrej Nerat (andrej.nerat@um.si)

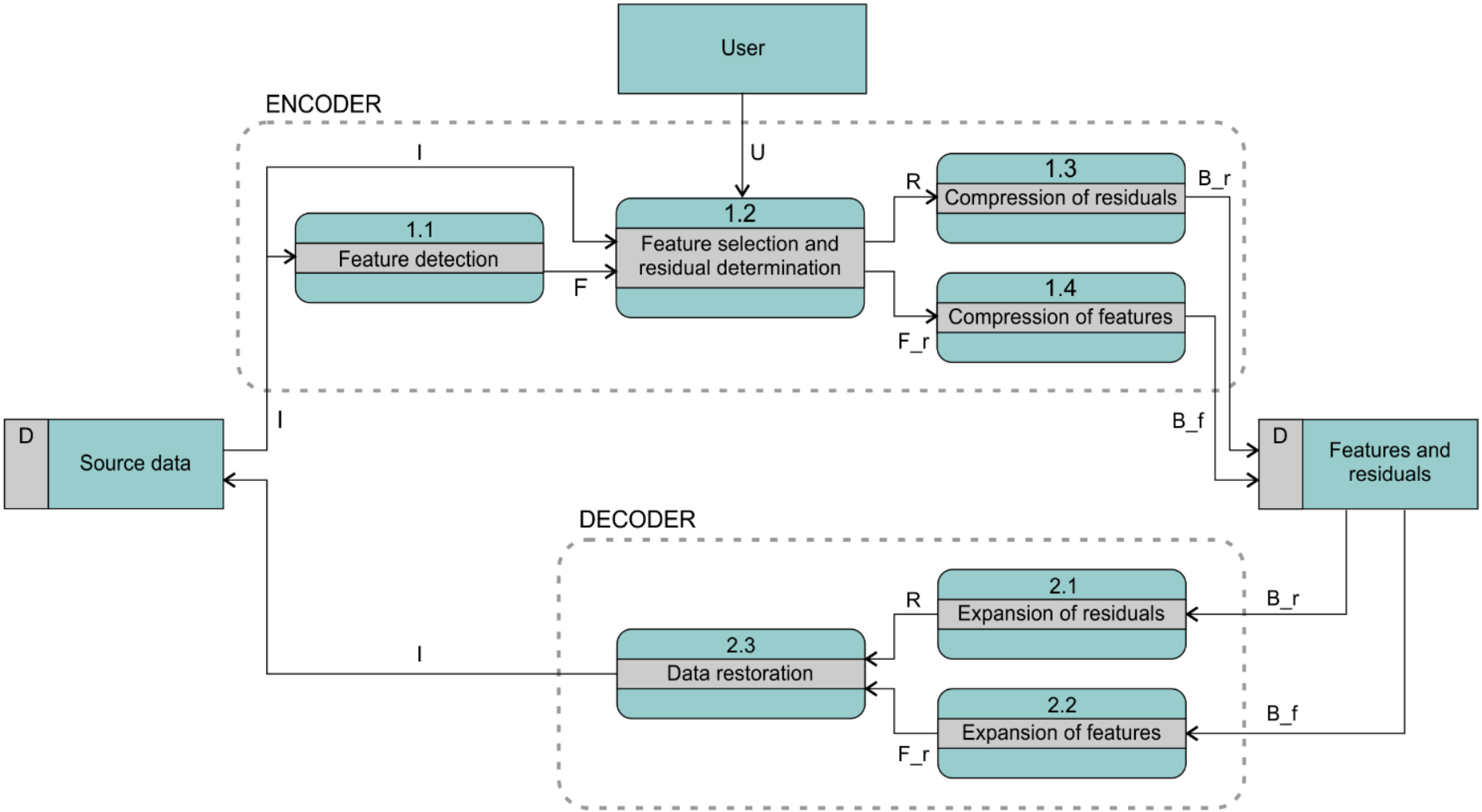
Background

- Difficult to compete with lossy data compression.
  - SOTA (SOA, STAR) methods extremely well elaborated in individual domains.
  - Sophisticated data transformations (frequency analysis) and quantization, both with strong scientific background.
- We thus focus on lossless data compression, where improvements are more likely.
  - Predictions, followed by encoding and compressing errors (residuals).
- Near-lossless methods also diserve a research focus.
  - Lossy because the decompressed data differ from the originals.
  - Derived from the lossless (prediction-based) philosophy.
  - Local error control (global i.e. averaged in lossy methods).

Basic idea

- Feature-based predictions and data restoration.
- Instead of encoding (compressing) individual input primitives (samples, symbols) or pre-defined patterns of primitives, features are extracted from the input stream.
- A feature is a piece of information with high discriminative (predictive) value for human interpretation or machine processing of a data stream.
- After detection, the feature set is optimized, and the selected features are then encoded.
- Residuals are also computed and compressed. Trivial ones (100% correctly predicted) may be omitted. Data restoration instead of data reconstruction or expansion.

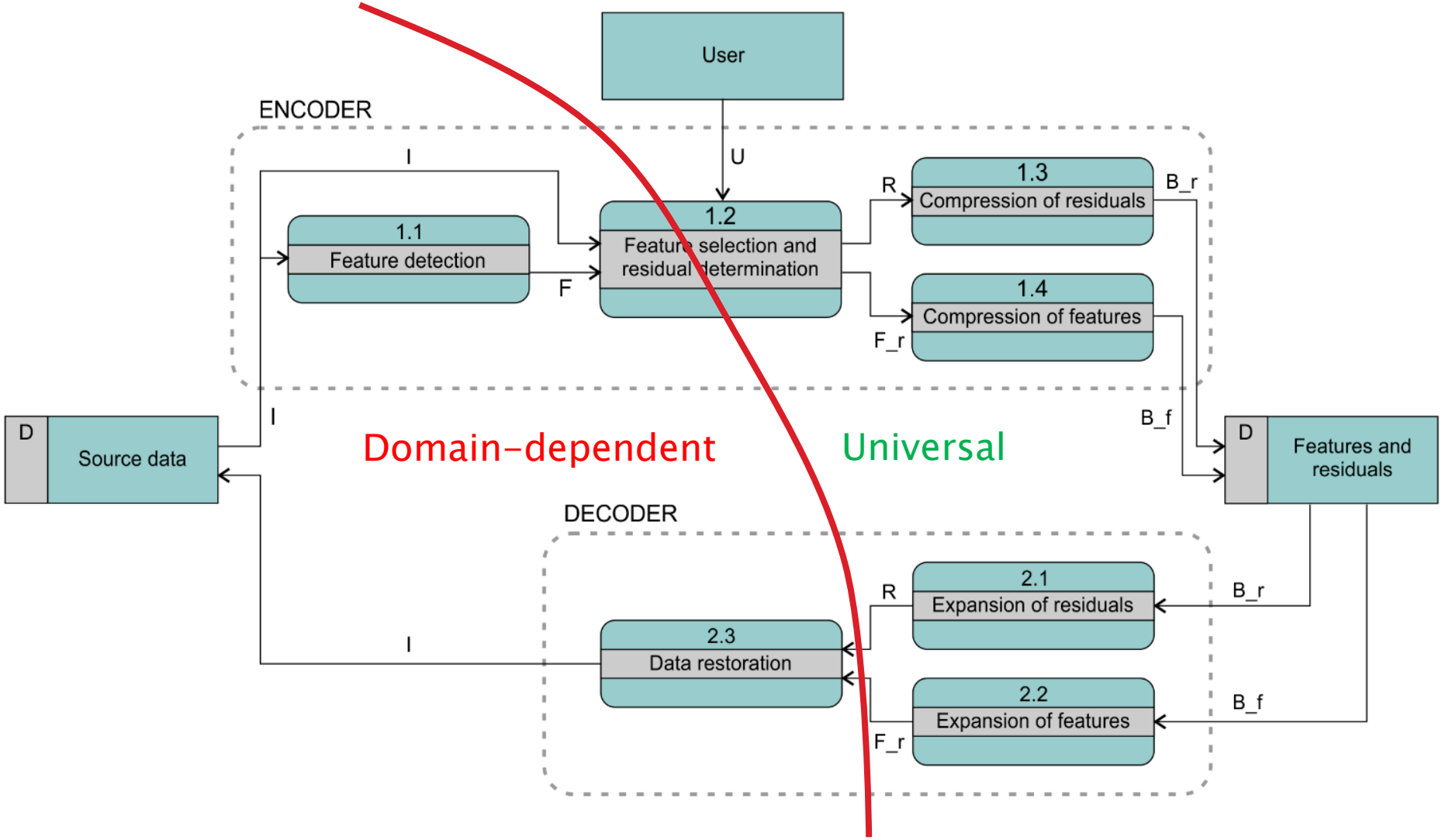
Concept


User
ENCODER
I
U
1.1
Feature detection
1.2
Feature selection and residual determination
F
R
1.3
Compression of residuals
B_r
1.4
Compression of features
F_r
D
Source data
I
B_f
D
Features and residuals
DECODER
R
2.1
Expansion of residuals
B_r
I
2.3
Data restoration
2.2
Expansion of features
F_r
B_f

Concept

- Feature types from the pilot domains will be used to define a unified taxonomy of features.
  - Raster images, audio, biomedical signals, and sparse voxel grids.
- Feature (and residual) compression/expansion wih thus be universal i.e. domain-independent.
- However, feature interpretation (detection, partially optimization, restoration) will remain domain-dependent.
- Role of user: select compression mode (constrain local errors in near-lossless compression). Lossless and lossy compression also considered.
- Compromise – which features (taxonomy and selection) and which mode to use.

Concept



COMPROMISE - Data compression paradigm based on omitting self-evident information

Concept

- Is it feasible?
- How does the universal concept affect the efficiency?
- EXAMPLE (fictional):
  - Universal feature taxonomy: extreme, sequence, border, pattern, ROI (however they are defined and interpreted in each domain).
  - Domain-dependent input data are mapped onto features from these classes. Mapping can also incorporate „pre-compression“ (with domain-dependent methods which we already have).
  - Universal methodology (fictional): compress extremes and patterns with BAC, sequences and borders with BASC, ROIs with Rice codes, and residuals with Deflate. In practice, the same method can be used for all categories.
- Any method can be „sold“if it supports lossless, near-lossless and lossy mode (and is efficient enough).

Objectives

- The overall objective of the proposed project is the development of a new data compression paradigm that is based on the investigation of advanced prediction methods with incorporation of features and restoration methods.

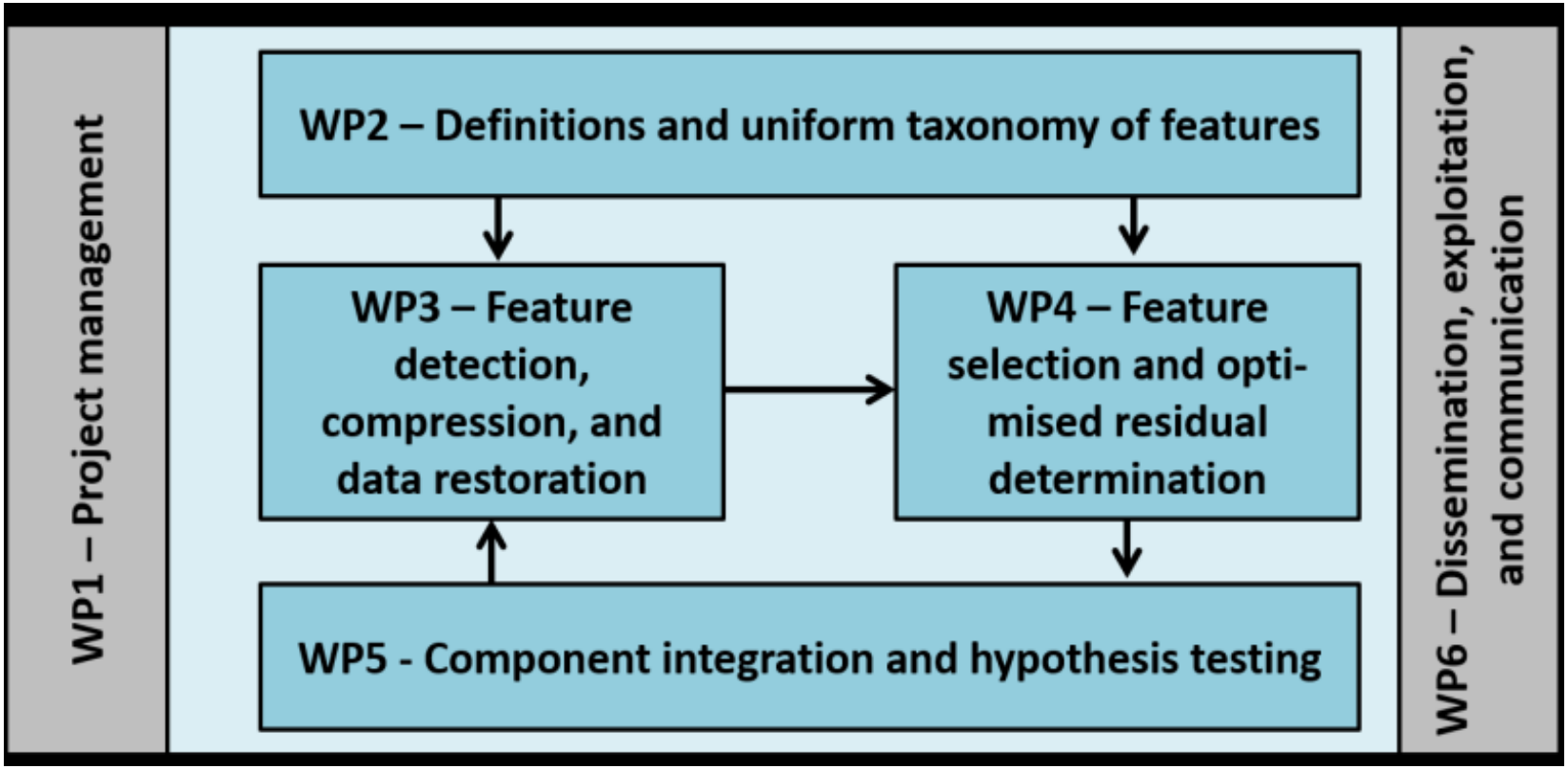
- Hypothesis: The universal methodology of lossless or near-lossless data compression, which will be based on unified feature taxonomy and restoration methods, will be more efficient than the existing compression procedures for raster images, digital audio, biomedical signals, and sparse voxel grids.

Specific objectives

SO1	*To develop a universal data compression methodology with a unified taxonomy of features from diverse domains, and a common framework for lossless, near-lossless, and lossy compression.*
SO2	*To upgrade the prediction of original data by integrating the techniques of feature selection and data restoration.*
SO3	*To improve the compression ratios in lossless and near-lossless mode in comparison with the existing approaches.*
SO4	*To improve the accessibility and reusability of features and feature-based restoration.*
SO5	*To deliver a verification environment for hypothesis testing in four pilot domains: raster images, digital audio, biomedical signals, and sparse voxel grids.*
SO6	*To disseminate the project results.*

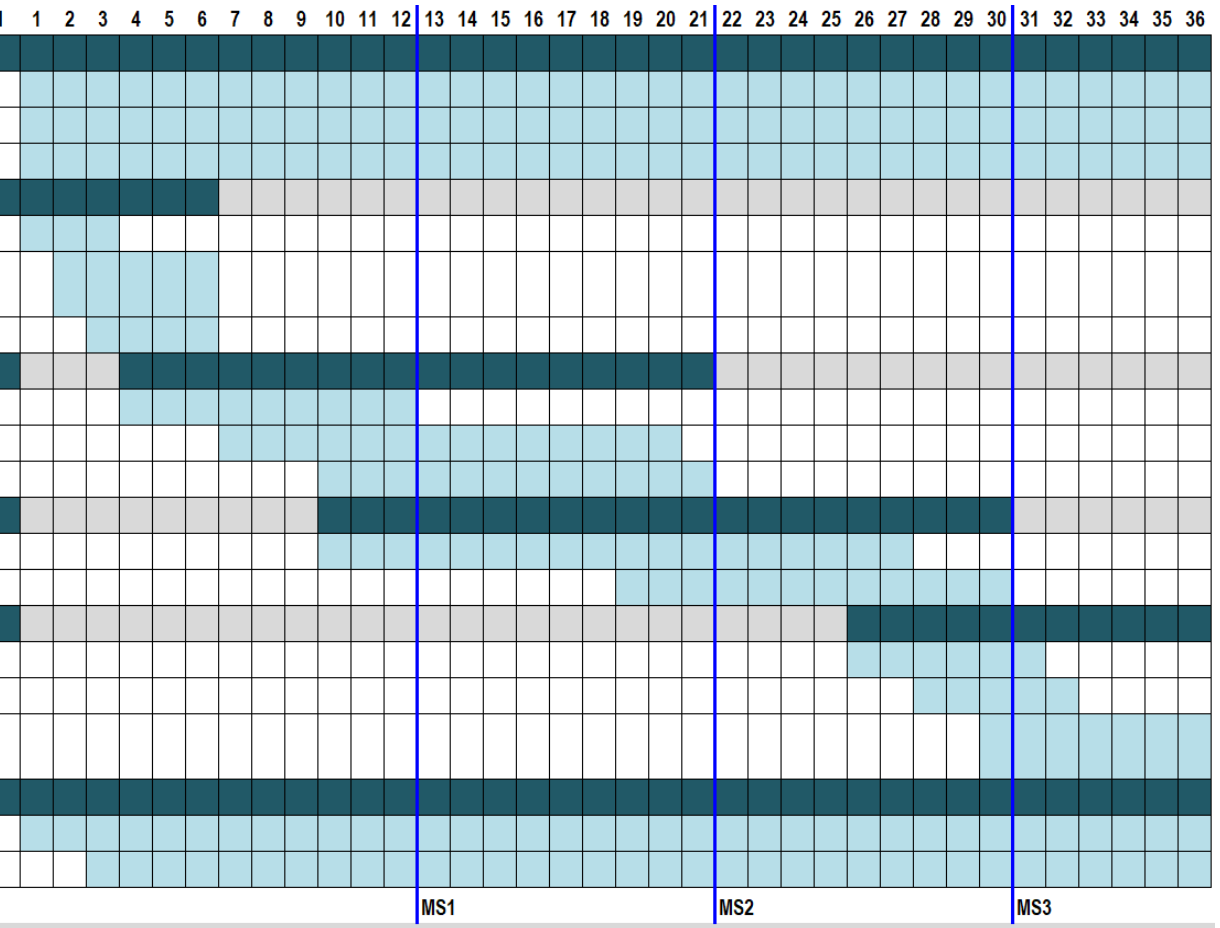
- Means of achieving these objectives and KPIs can be read in the project description document.

Workplan


WP1 – Project management
WP2 – Definitions and uniform taxonomy of features
WP3 – Feature detection, compression, and data restoration
WP4 – Feature selection and opti-mised residual determination
WP5 - Component integration and hypothesis testing
WP6 – Dissemination, exploitation, and communication

Workplan

WP	T	Work package/task title	Start	End
1		**Project management**	1	36
	1	Administrative and financial project management	1	36
	2	Quality assurance and risk mitigation	1	36
	3	Legal, data and knowledge management	1	36
2		**Definitions and unified taxonomy of features**	1	6
	1	Generation of domain-dependent feature repertoires	1	3
	2	Definition of feature descriptions and development of methods for their interpretation	2	6
	3	Specification of domain-independent feature taxonomy	3	6
3		**Feature detection, compression, and data restoration**	4	21
	1	Feature detection	4	12
	2	Data restoration and residual determination	7	20
	3	Lossless compression of features and residuals	10	21
4		**Feature selection and optimised residual determination**	10	30
	1	Feature selection	10	27
	2	Integration of feature selection and residual determination	19	30
5		**Component integration and hypothesis testing**	26	36
	1	Adaptation of SOTA methods for comparison	26	31
	2	Component integration	28	32
	3	Analysis of results, iterative improvements of methodology, and hypothesis testing	30	36
6		**Dissemination, exploitation, and communication**	1	36
	1	Dissemination, exploitation, and communication strategy	1	36
	2	Dissemination activities	3	36



MS1 Proof of concept

MS2 The first operational prototype based on redundant feature set

MS3 Optimized system based on selected features

Deliverables

- At least 3 papers in international open access journals,
- at least 6 conference papers,
- organisation of 2 dedicated presentation events,
- 1 patent application,
- eventual additional requirements from GAČR(?),
- project website and a profile on at least 1 social network (after M6),
- ongoing results at the end of individual WPs (plans, reports, instructions, software, test datasets…).

How to start?

- Read the project description, ask me the questions.
- Take an inventory of your results on data compression so far and think which to use in the project.
- Think about how to make your lossless methods near-lossless or lossy, and vice versa.
- Think about features to be extracted from your input data streams.
  - Domain-dependent feature repertoires till M3, universal feature taxonomy till M6.
  - Raster images, audio, biomedical signals, and sparse voxel grids.

Discussion

- Some other (informal) presentation of previous/ongoing data compression results?
- Questions, ideas, comments?
- Next meeting
  - Date and time
  - Presentation(s)