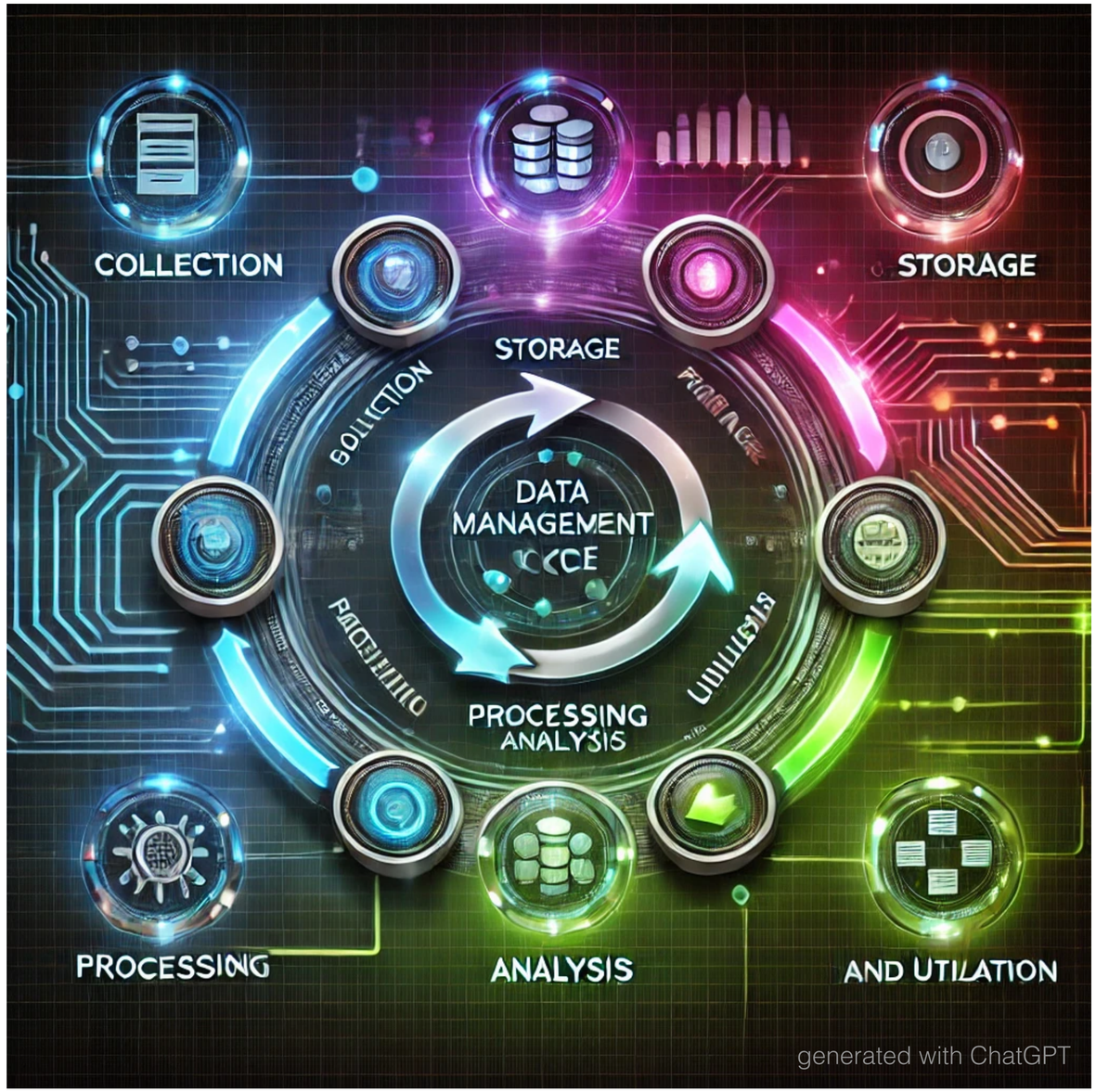



# Data Management and Introduction to AI tools

**Dr. Veselina Kalinova**

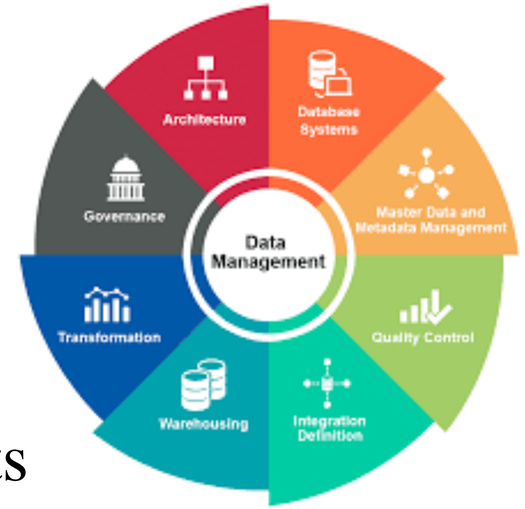
Institute for Sustainable Hydrogen Economy
Forschungszentrum Jülich

Max-Planck-Institut für Radioastronomie

OPINCHARGE

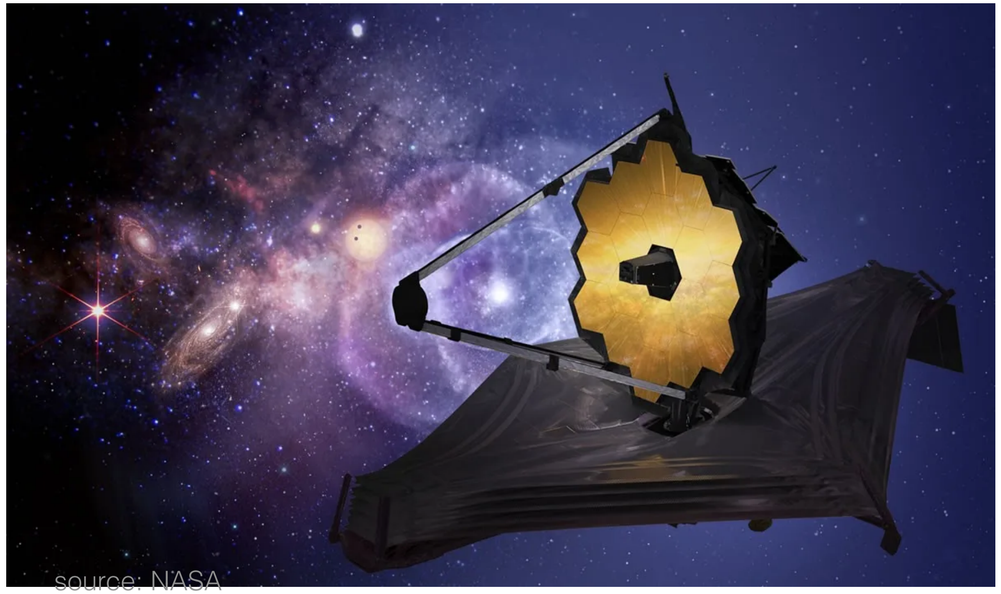
HC·H2 Helmholtz-Cluster Wasserstoff

# Data lifecycle




- **Data Collection:** raw data
- **Data Storage:** where is stored
- **Data Organisation:** properly labelled and structured
- **Data Security:** protecting from unauthorised access and cyber threats
- **Data Governance:** data access and data usage policies
- **Data Integration:** combining data from different sources into a unified view
- **Data Quality:** ensuring that data is accurate
- **Data Access, Retrieval, Publication:** making data easily accessible to users
- **Data Analytics:** data is used for scientific analysis and machine learning applications

# Data Management: key points




- Scientists need to document/label the data in details (e.g. flag quality, units, instrument, background, scientific scope, resolution)

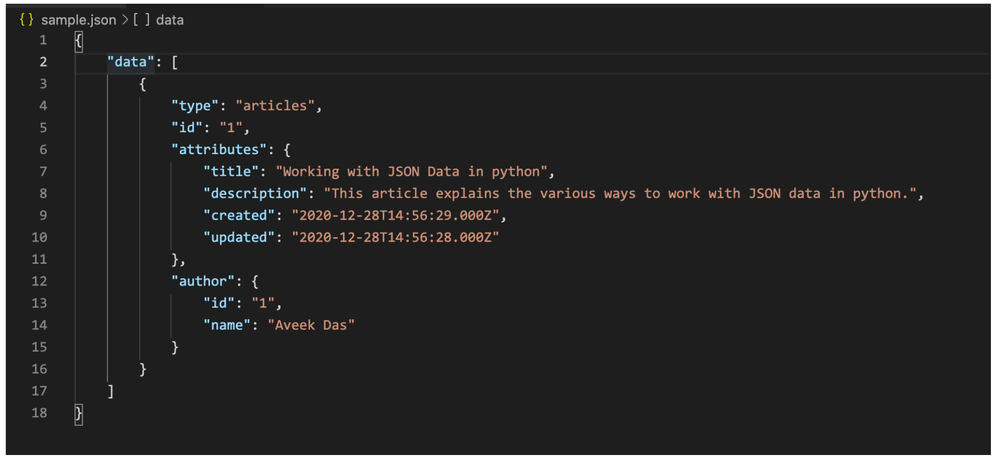
- Test and verification of the created database from collaborators and co-authors

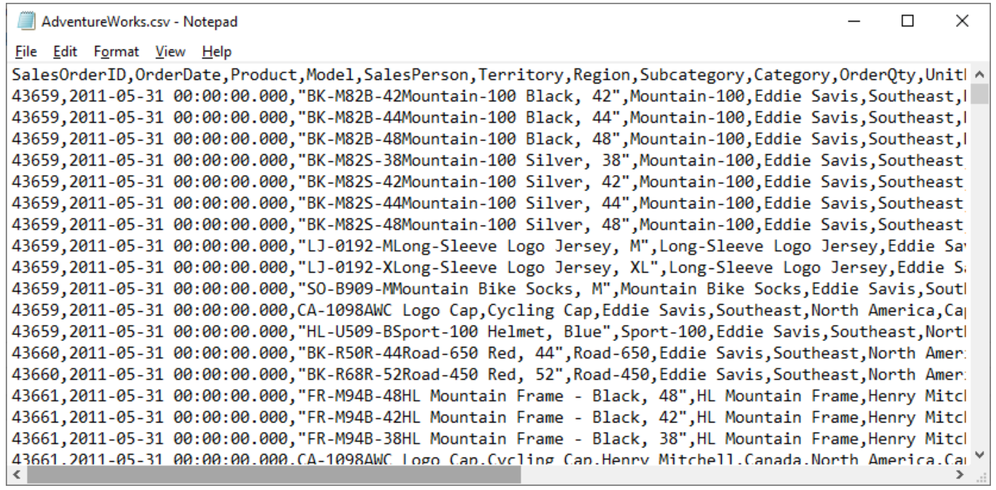
- Re-organising procedures and data on a regular base

- Scientific publications with access to the raw and secondary data products to ensure reproducibility

# Data Formats

***Machine learning research uses various data formats depending on the specific task***




***Some of the most common formats:***

- **Text Formats (TXT, TSV)**: stores sequences of text that can later be transformed into numerical vectors
- **CSV (Comma-Separated Values)**: data is stored in rows and columns, with each column representing a feature
- **JSON (JavaScript Object Notation)**: Flexible for representing complex data structures like dictionaries, lists, and objects
- **Image Formats (JPEG, PNG, TIFF)**: for computer vision tasks; data can be stored as raw pixel values or compressed formats
- **Numpy Arrays (NPY, NPZ)**: arrays or matrices of data that are easy to manipulate with high performance (TensorFlow, PyTorch, and Scikit-learn)
- **HDF5 (Hierarchical Data Format)**: supports complex data hierarchies and allows for easy data access and manipulation; for deep learning tasks

# The best AI tools by category (so far)


credit: Tam Ho

- **Chatbots** (ChatGPT, Claude, Bing AI, Zapier Central)
- **Content creation** (Jasper, Copy.ai, Anyword)
- **Grammar checkers and rewording tools** (Grammarly, Wordtune, ProWritingAid)
- **Video creation and editing** (Descript, Wondershare Filmora, Runway)
- **Image generation** (DALL·E 3, Midjourney, Stable Diffusion)
- **Voice and music generation** (Murf, Splash Pro, AIVA)
- **Knowledge management and AI grounding** (Mem, Notion AI Q&A, Personal AI)
- **Task and project management** (Asana, Any.do, BeeDone)
- **Transcription and meeting assistants** (Fireflies, Airgram, Krisp)
- **Scheduling** (Reclaim, Clockwise, Motion)
- **Email inbox management** (SaneBox, Mailbutler, EmailTree)
- **Slide decks and presentations** (Decktopus, Beautiful.ai, Slidesgo)
- **Automation** (Zapier)
- **Other AI productivity tools**

source: https://zapier.com/blog/best-ai-productivity-tools/

# Usage of AI tools: tips

generated image with ChatGPT-DALL-E

**Example of generated image from text with ChatGPT-DALL-E**

Prompt: "*Would you, please, generate an image of a dragon, reading on an armchair? The dragon look relaxed and he is drinking tea. The style of the image suit a text for children's book (and it is like a paint). The dragon is senior with glasses.*

The details of the generated image are incredible right, based on the input text…

ChatGPT has similar impressive performances for producing computer codes (very useful for scientists), language corrections, summarising contents, etc…

# Usage of AI tools: tips

generated image with ChatGPT-DALL-E

- Do not put personal information in the AI tool (e.g. your CV, names, contacts, addresses, etc)
- Do not fully rely on the generated output -very often some modifications and corrections are needed
- The copy rights of the produce content with AI tools are still debated (e.g. DFG request to be specified if and how an AI tool has been used during writing of the project)
- Be always kind in your request: your answers are further training the models of the AI tool (you never know…)

Did you notice that the tea cup is in the air?…

# OpenAI-Sora: generated video from text

Prompt: "*A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk around.*"

source: https://openai.com/index/sora/ source: Direct link