# CORE147

# DIGITÁLNÍ DATA V HUMANITNÍCH A SOCIÁLNÍCH

Zuzana Nevěřilová
Hana Žižková

2024/25

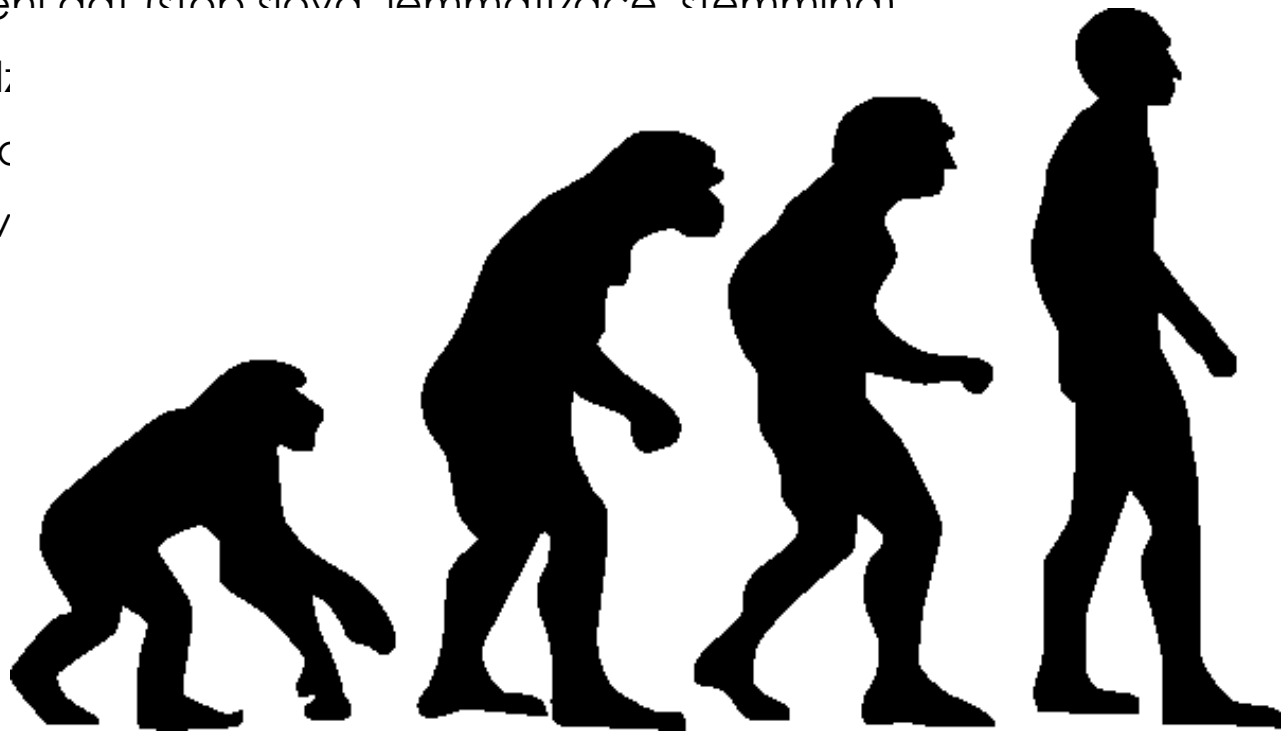
# DATA TEČOU POTRUBÍM
## PROCESSING PIPELINE

„série algoritmů, které řeší opakované procesy“

- Čištění dat (stop slova, lemmatizace, stemming)
- Předzpracování (sjednocení velikosti obrázků)
- Ukládání zpracovaných dat
- Konverze do jiných formátů

# JAK NA TO?

- Čištění dat (stop slova, lemmatizace, stemming)
- Před
- Uklá
- Konv



1. Napíšu si, které programy, v jakém pořadí spustit
2. Vytvořím skript, který programy spustí
3. Kontroluju vstupy, výstupy a verze
4. Vytvořím a uložím workflow

https://commons.wikimedia.org/wiki/File:Human_evolution.svg

# VÝROBNÍ LINKA

**MODULARITA**
**STANDARDIZACE**
**ZAMĚNITELNOST**

- Oddělené kroky
- Návaznosti
- Vyměnitelné části
  - Rychlé úpravy
  - Možnost porovnat

https://www.rte.ie/brainstorm/2018/0312/946759-what-can-the-assembly-line-teach-us-about-innovation/

# EXTRAHUJ, TRANSFORMUJ, ULOŽ

**Komponenty**

- Zdroje dat
- Workflow
- Úložiště
- Transformace
- Business Intelligence

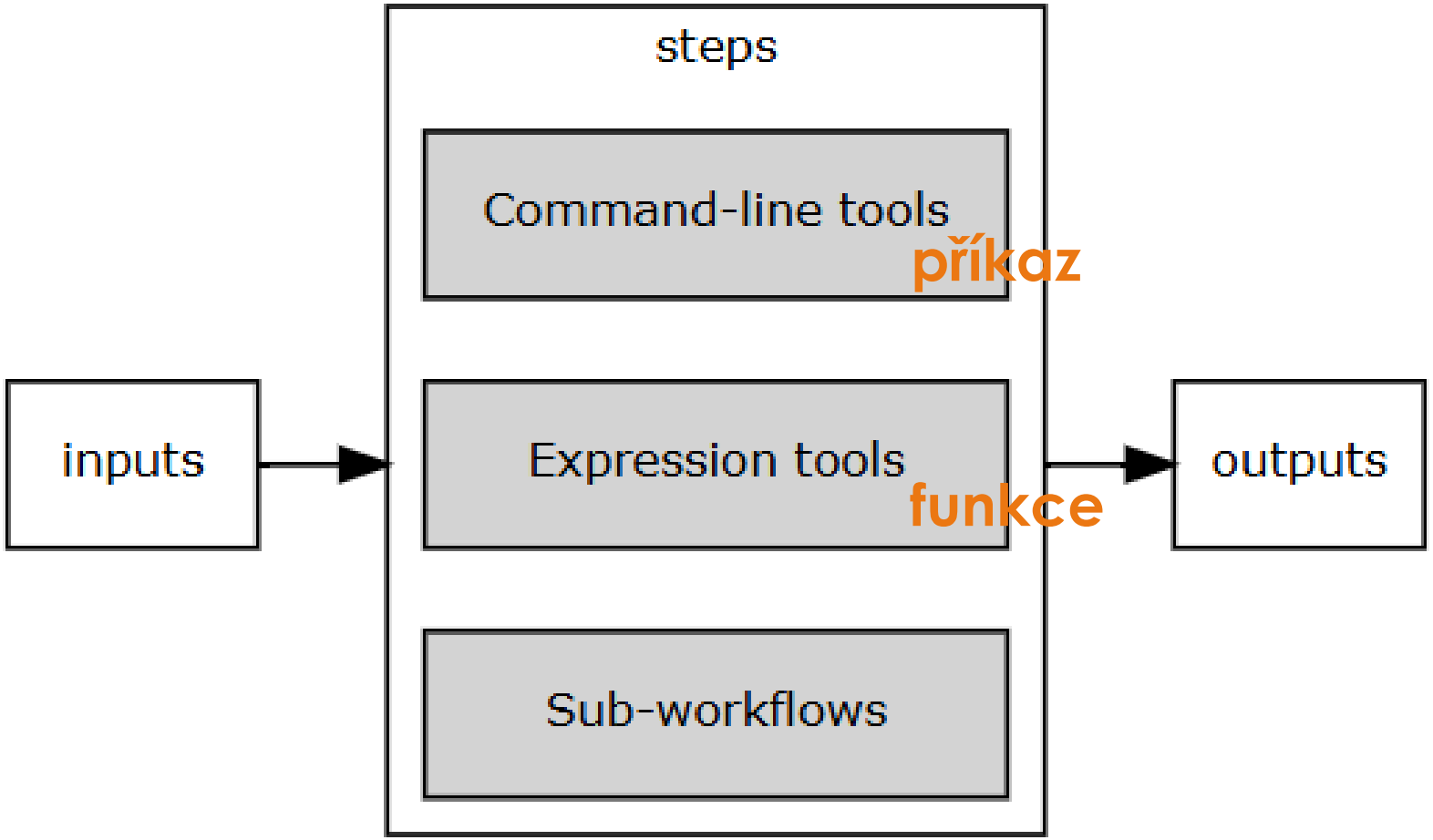
**Aktivity** - Extract, Transform, Load (ETL)

- Proces lze opakovat
- Proces lze automatizovat
- Lze najít chyby a pracovat s nimi
- Průběžné výsledky jsou v bezpečí
- Dají se použít průběžné výsledky

https://lakefs.io/blog/python-data-pipeline/

# WORKFLOW

- **MODULARITA**
- **STANDARDIZACE**
- **ZAMĚNITELNOST**




https://www.commonwl.org/user_guide/topics/workflows.html

# NÁSTROJE PRO POPIS WORKFLOW

- Workflow Description Language
- Common Workflow Language
- Sémantické verzování
  - MAJOR.MINOR.PATCH
  - `1.0.0-alfa`
- Plánování spuštění v čase

- Nástroje pro programátory
  - Airflow, Prefect
- No-code nástroje
  - Hevo Data
  - Intergrate.io

# POTRUBÍ V DIGITÁLNÍCH HUMANITNÍCH A SOCIÁLNÍCH VĚDÁCH

Zpracování textu

- OCR, čištění, stop slova, lemmatizace, stemming, klastrování, hledání témat (topics), rozpoznání pojmenovaných entit a časových údajů

Zpracování obrazu

- Převedení na jednotný formát, změna velikosti, ořez do jednotného tvaru, barevná korekce, filtry (zaostření, rozostření, detekce hran), detekce objektů, rozpoznání lokace

Zpracování videa

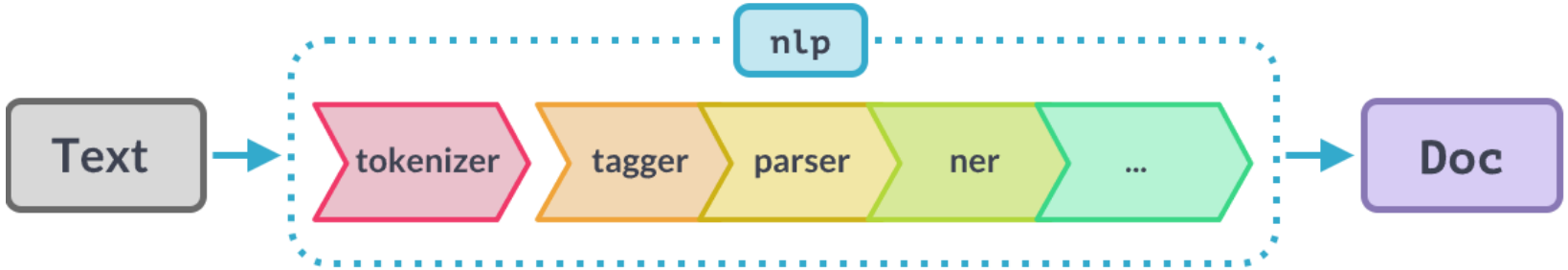
- Tracking objektů ve videu, skupiny objektů

Použití časově nebo finančně náročných kroků

- Jazykové modely, generování obrázků, ...

# WORKFLOW V JEDNOM PROGRAMU

Příklad: SpaCy



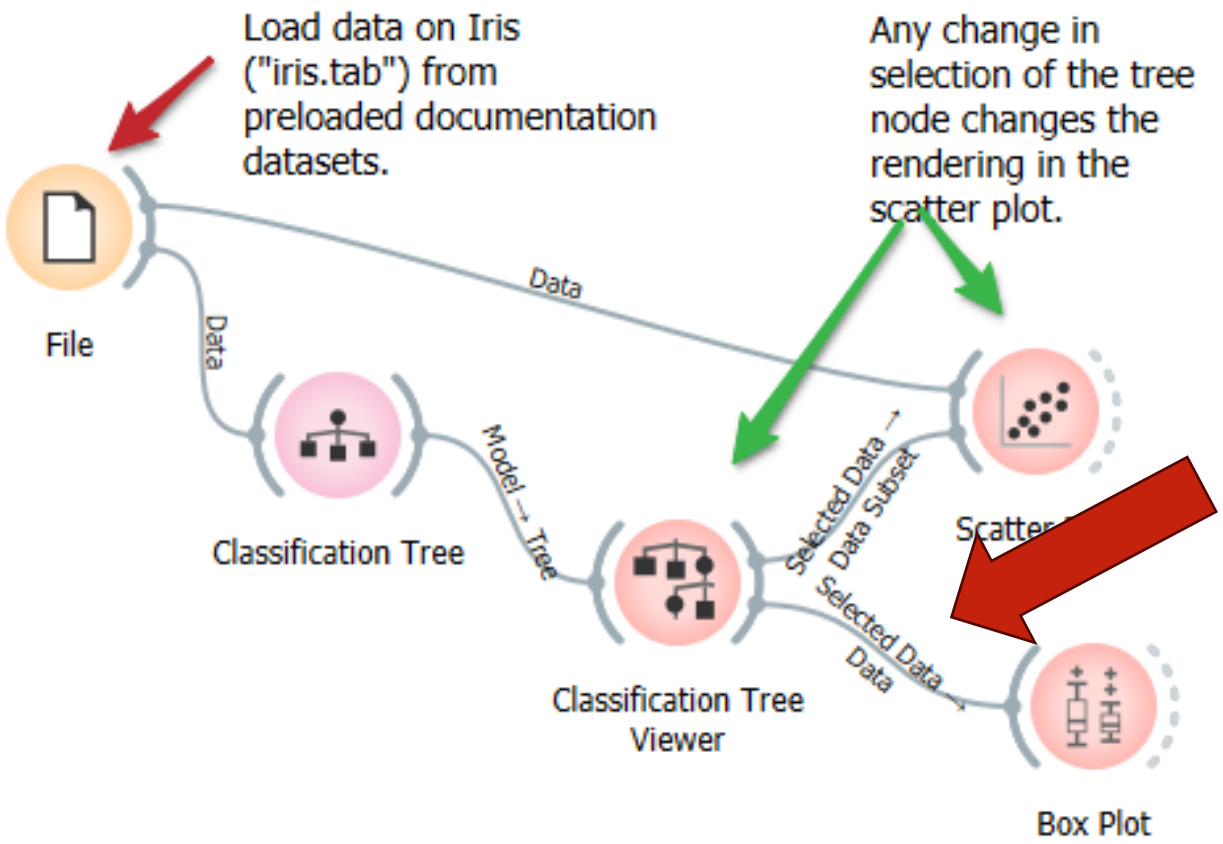

- **MODULARITA**
- **STANDARDIZACE**
- **ZAMĚNITELNOST**

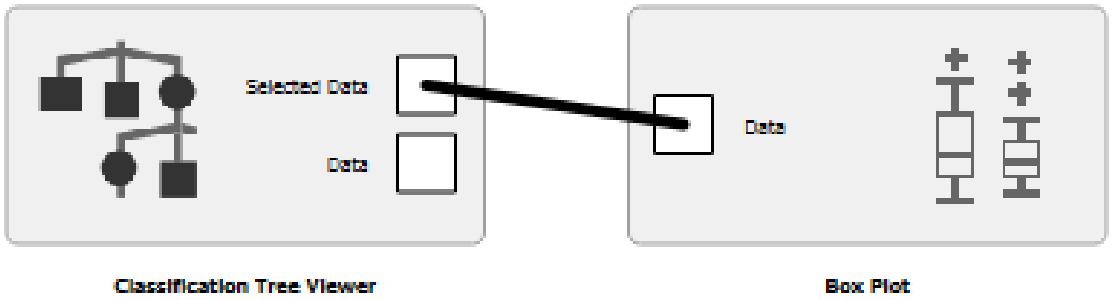
```
pipeline = ["tok2vec", "tagger", "parser", "ner"]
```

https://spacy.io/usage/processing-pipelines

# WORKFLOW NEZÁVISLÉ NA VÝROBCI

Příklad: Orange Data Mining




- MODULARITA
- STANDARDIZACE
- ZAMĚNITELNOST

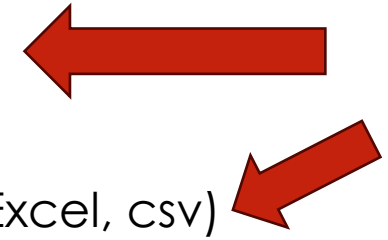
# LINGVISTICKÉ NÁSTROJE JAKO SLUŽBA
## AS A SERVICE

Příklad: WebLicht

- **MODULARITA**
- **STANDARDIZACE**
- **ZAMĚNITELNOST**

# CO MUSÍ PIPELINES UMĚT

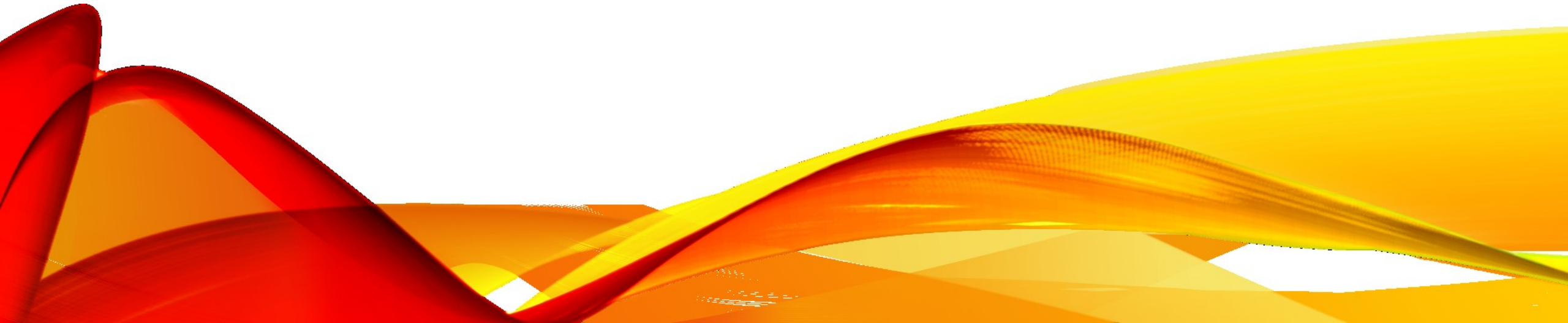
- Uživatel si vybere komponenty a způsob jejich zapojení
- Nástroj umožní spojit jen spojitelné komponenty
- Podpora různých vstupních formátů (plain text, CQL, TCF, Excel, csv)
- Podpora různých výstupních formátů (HTML, TCF, ConLL-U, Excel, csv)
- Konverzní nástroje
- Reportování uvnitř potrubí (export do obrázků, uložení průběžných výsledků)
- Uložení potrubí (processing chain)



- **MODULARITA**
- **STANDARDIZACE**
- **ZAMĚNITELNOST**

# NĚCO (MÁLO) O FORMÁTECH

Text | Obraz



# FORMÁTY

## Text a tabulky

- Plain
- XML
- HTML
- CSV/TSV
- Excel
- …

## Anotace textu

- Tabulka
  - CSV
  - TSV
- Hierarchie
  - XML
- Složitější struktury
  - JSON

# PŘÍKLADY PRO TEXT

## Text Corpus Format (TCF)

XML based format for storing

- Tokens in text corpus
- Metadata
- Annotation layers:
  - Sentences, lemmas, tagging, parsing, morphology, named entities, references, geographical locations, phonetics, …

```
<TextCorpus lang="de">
   <text>Karin fliegt nach New York. </text>
  <tokens>
        <token ID="t_0">Karin</token>
        <token ID="t_1">fliegt</token>
        <token ID="t_2">nach</token>
        …
  </tokens>
  <POStags tagset="stts">
        <tag ID="pt_0" tokenIDs="t_0">NE</tag>
        <tag ID="pt_1" tokenIDs="t_1">VVFIN</tag>
        <tag ID="pt_2" tokenIDs="t_2">APPR</tag>
        …
  </POStags>
</TextCorpus>
```

# PŘÍKLADY PRO TEXT

## ConLL-U

Universal Dependencies Format (text based)

- Universal across languages
- Annotation types: ID, FORM, LEMMA, UPOSTAG, XPOSTAG, FEATS, HEAD, DEPREL, DEPS, MISC

| | | | | |
|---|---|---|---|---|
| 1 | Då | Då | ADV | AB |
| 2 | Var | Vara | VERB | VB.PRET.ACT |
| 3 | han | han | PRON | PN.UTR.SIN.DEF.NOM |
| 4 | elva | elva | NUM | RG.NOM |
| 5 | År | år | NOUN | NN.NEU.PLU.IND.NOM |
| 6 | . | . | PUNCT | DL.MAD |

https://universaldependencies.org/format.html

# PŘÍKLADY PRO TEXT

**XML**

```
<api batchcomplete="">
  <query>
    <normalized>
      <n from=
       "greek_architecture" to=
         "Greek architecture"/>
    </normalized>
    <pages>
      <page _idx= "19493213"
       pageid= "19493213"
       ns="0"
       title= "Greek
         architecture"/>
    </pages>
  </query>
</api>
```

**JSON**

```
{
 "batchcomplete":"",
 "query":{
   "normalized":[{
     "from":"greek_architecture",
     "to":"Greek architecture"}
   ],
   "pages":{
   "19493213":{
     "pageid":19493213,
     "ns":0,
     "title":"Greek architecture„
   }
  }
 }
}
```

# FORMÁTY

## Obraz

- Bitmapa
  - PNG
  - JPG
- Vektorový formát
  - SVG
- Kombinace
  - PDF
  - EPS

## Anotace obrazu

- Tabulka
  - CSV
  - Bounding boxes
- Složitější struktury
  - JSON

https://roboflow.com/formats

# PŘÍKLADY PRO OBRAZ

## Různý pohled na bounding box

Pascal VOC [x_min, y_min, x_max, y_max]

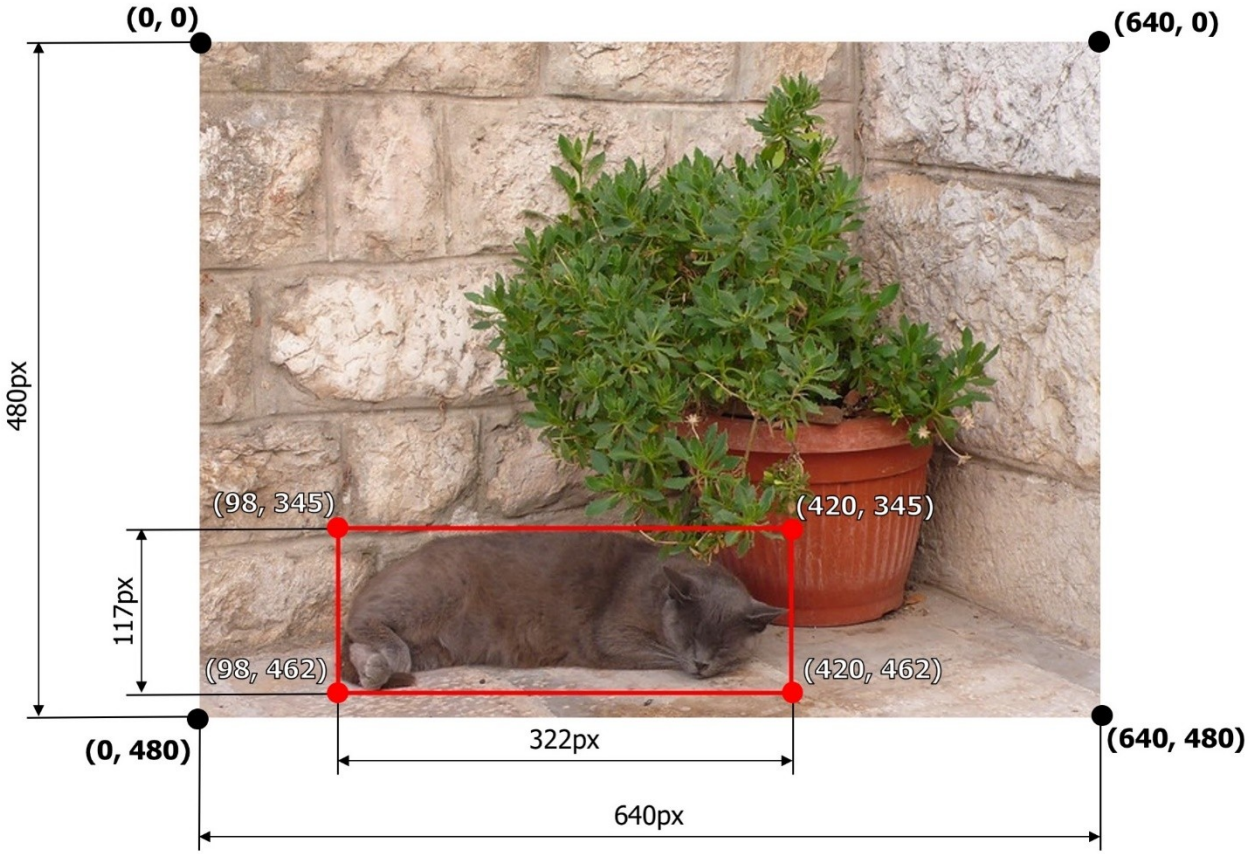
[98, 345, 420, 462]

COCO [x_min, y_min, width, height]

[98, 345, 322, 117]

YOLO [x_center, y_center, width, height]

[0.4046875, 0.840625, 0.503125, 0.24375]



# PIPELINES SHRNUTÍ

**Komponenty**

- Zdroje dat
- Workflow
- Úložiště
- Transformace
- Business Intelligence

**Výsledek**

- Reprodukovatelnost
- Automatizace
- Zotavení z chyb
- Průběžné výsledky
- Bezpečnost dat
- Přehled

# REFERENCES


- Ashley S. Lee , Poom Chiarawongse, Jo Guldi, Andras Zsom: **The Role of Critical Thinking in Humanities Infrastructure: The Pipeline Concept with a Study of HaToRI (Hansard Topic Relevance Identifier)**. DHQ. Volume 14 Number 3. https://digitalhumanities.org/dhq/vol/14/3/000481/000481.html
- CLARIN-D/SfS-Uni. Tübingen. 2012. WebLicht: Web-Based Linguistic Chaining Tool. Online. Date Accessed: 12 Dec 2020. URL https://weblicht.sfs.uni-tuebingen.de/