Leveraging Public APIs for Machine Learning Datasets

Johannes Ahlmann, Sensatus.io

PyCon 2019

2019-10-12

johannes@sensatus.io

About Me

Johannes Ahlmann

- Living in Cork
- Developing in Python since 2002
- Built large-scale Machine Learning solutions using Python, Tensorflow, Kafka, Spark

Sensatus.io

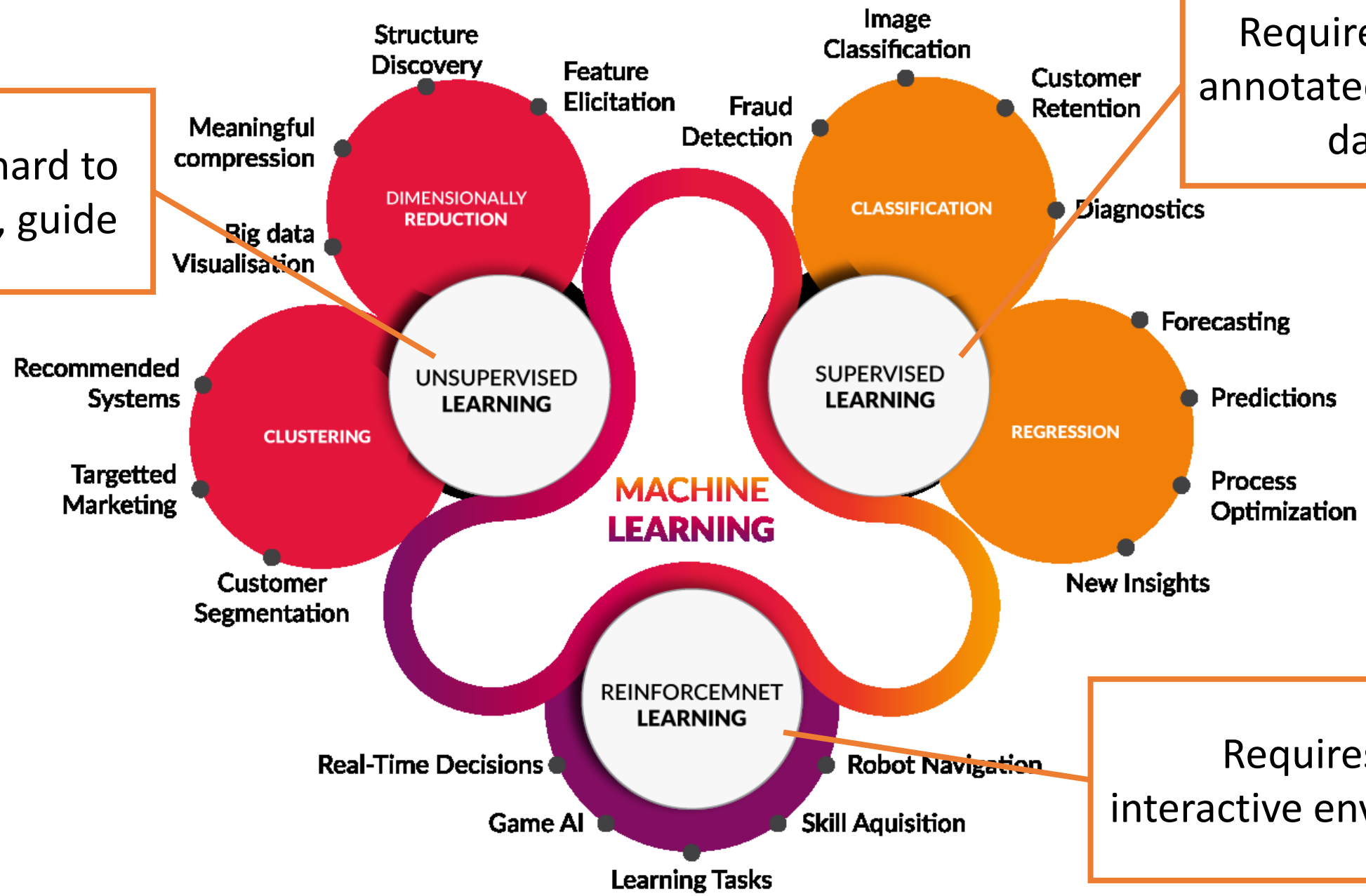
- On-Prem AI Models
- Gathering and Enriching Web Data
- Sales & Client Intelligence

Github: **@codinguncut**

- codinguncut/leveraging_public_apis

Can be hard to validate, guide

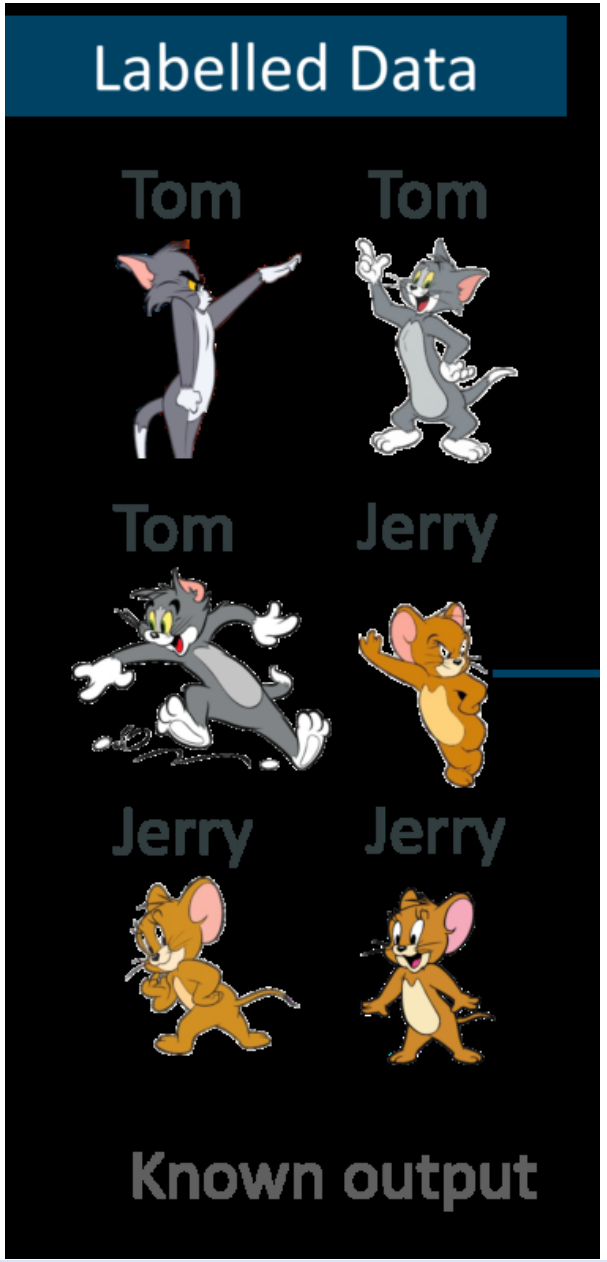
Requires large annotated training data



Requires an interactive environment

Supervised Machine Learning

- Natural Language Processing
  - Machine Translation
  - Sentiment Analysis
  - Document Classification
  - Named Entity Extraction
- Speech
  - Speech-to-Text
  - Text-to-Speech
- Machine Vision
  - Object detection
  - Object classification
  - Facial Recognition
- Semi-Supervised
- Embeddings



Do we need Training Data?

- Customizing a model, Transfer Learning
- Generic models vs. Use Case specific
- Access to Intermediate representations (i.e. for speech-to-text)
- Speech-to-Text
  - Specific to regional accents, dialects, jargon
  - Specific to constrained contexts (i.e. call center calls)
- Sentiment Analysis
  - Cultural differences, Jargon, etc.
- NLP for different languages beyond [English, Spanish, French, German]

Web Scraping

- APIs and existing datasets will only get us so far
- Web Scraping enables us to extract data from publicly accessible websites
- Possible to extract structured data from websites
- Possible to fetch data from millions of pages, as long as we follow best practices and politeness
- Python Tools
  - scrapy/scrapy
  - scrapinghub/frontera
  - scrapinghub/splash