data intelligence
institute of Paris

Data Science

A high level overview

work supported by IdEx Université de Paris, ANR-18-IDEX-0001

Outline

- What is Data Science
- Data Science Workflow
- Tools for Data Science
- Python for Data Science
- Applications of Data Science
- Challenges of Data Science

- Hands-on data science process example

What is Data Science[1]

area of study which involves extracting insights from data using various scientific methods, algorithms, and processes.

- helps to discover hidden patterns in the data.
- can predict unseen facts/events.

What is Data Science[1]

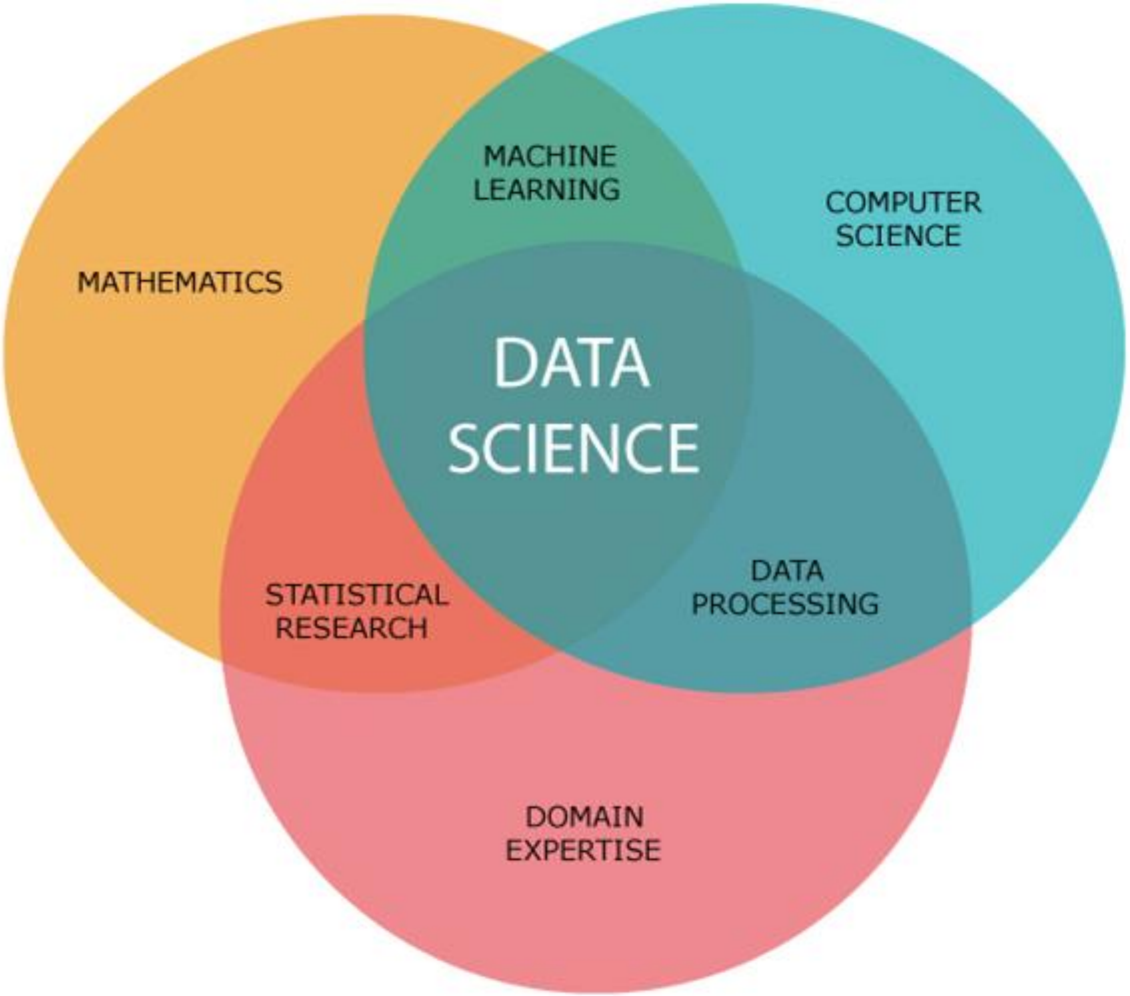
- usually deals with large amounts of data
- allows knowledge extraction from structured or unstructured data
- is an interdisciplinary field
    - mathematics
    - statistics
    - artificial intelligence
    - machine learning
    - data analysis
    - big data management

Why Data Science

- data is one of the most important features of every organization
  - enables making decisions based on facts, statistical numbers and trends
- data formats and sizes have been exploding

Data Science employs and blends methods and techniques from many fields:
artificial intelligence machine learning, visualization, pattern recognition,
probability model, data engineering, signal processing, etc.

Data Science Components[2]



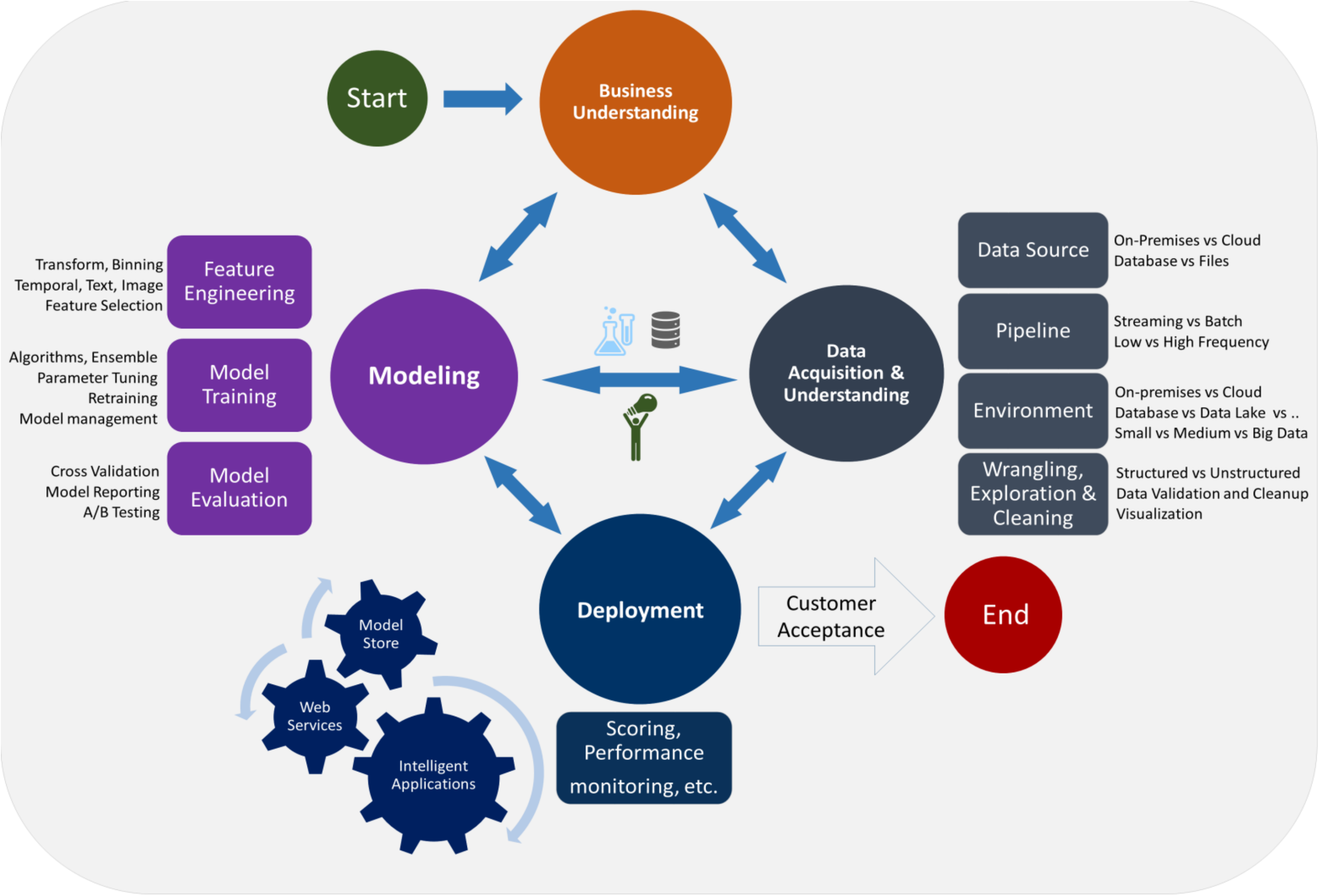
Data Science Components[2]

- **Data engineering:** Involves acquiring, storing, retrieving, and transforming the data and metadata.
- **Statistics:** Is a way to collect and analyze large amounts of (numerical) data and to find meaningful insights from it.
- **Visualization:** Makes it easy to access and inspect (huge amounts of) data.
- **Domain Expertise:** Due to the vast area of fields where data science can be applied.

Data Science Components[2]

- **Machine Learning & Artificial Intelligence:** backbone of data science. In data science, we use various ML/AI algorithms to solve the problems.
- **Mathematics:** Essential to understand the inner-workings of ML/AI techniques.
- **Advanced Computing:** Data science techniques are very resource (computation and memory) intensive

Data Science Workflow[5]



Start

Business Understanding

Data Acquisition & Understanding

Data Source — On-Premises vs Cloud, Database vs Files

Pipeline — Streaming vs Batch, Low vs High Frequency

Environment — On-premises vs Cloud, Database vs Data Lake vs .., Small vs Medium vs Big Data

Wrangling, Exploration & Cleaning — Structured vs Unstructured, Data Validation and Cleanup, Visualization

Modeling

Feature Engineering — Transform, Binning, Temporal, Text, Image, Feature Selection

Model Training — Algorithms, Ensemble, Parameter Tuning, Retraining, Model management

Model Evaluation — Cross Validation, Model Reporting, A/B Testing

Deployment

Scoring, Performance monitoring, etc.

Model Store

Web Services

Intelligent Applications

Customer Acceptance

End

Data Science Workflow[1,5]

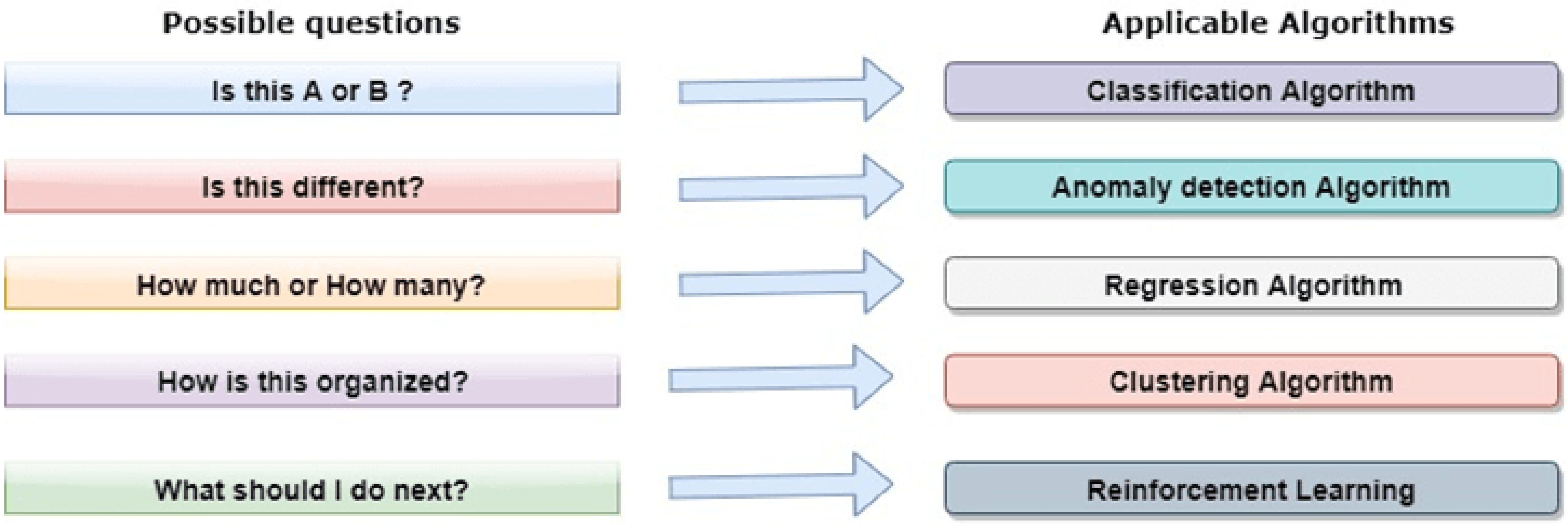
- **Discovery Phase : frame the problem**
  - Understand real world problem
  - Identify relevant ML/AI problem(s) **
  - Identify relevant internal/external data sources
- **Data Identification and Acquisition**
- **Data Preparation**
  - Data cleaning
  - Data transformation
- **Exploratory Data Analysis**
  - understand the information contained within at a high level
    - what kinds of obvious trends or correlations do you see in the data
    - what are the high-level characteristics and are any of them more significant than others?

Data Science Workflow[1,5]

- **Model Building**
  - May involve further data transformations
- **Communicate Results**
- **Model Operationalization (if required)**
  - Design monitoring and update schedule of model
  - Deploy and integrate model into organization’s computing infrastructure

Data Science Workflow

How to map a problem in Data Science to ML/AI tasks?



Possible questions	Applicable Algorithms
Is this A or B ?	Classification Algorithm
Is this different?	Anomaly detection Algorithm
How much or How many?	Regression Algorithm
How is this organized?	Clustering Algorithm
What should I do next?	Reinforcement Learning

Tools for Data Science

- **Data Analysis:** Python/Pandas, Statistics, Jupyter, R/R Studio, MATLAB, Excel, RapidMiner
- **Big Data:** Snowflake, SQL, Apache Spark, AWS Redshift, MemSQL
- **Data Visualization**: R, Python/Pandas/matplotlib, Jupyter, Tableau, Cognos
- **Machine Learning**: Anaconda/Scikit, R, Spark, Mahout, Azure ML studio, Auto ML, NLTK
- **Deep Learning**: Keras/Tensorflow, TensorBoard, Pytorch, MxNet, Caffe, H20, fastai

Applications of Data Science[1,3]

- **Recommendation Systems :**
  - e.g., suggested friends on Facebook, videos on YouTube/Netflix, products on Amazon
  - for users: enable more effective search
  - for providers:
    - increase in sales thanks to personalized offers
    - more time spent on the platform
    - customer retention thanks to users feeling understood
- **Healthcare**
  - medical image analysis
  - genetics and genomics
  - drug development
  - virtual assistance for patients and customer support

Applications of Data Science[1,3]

- **Fraud and Risk Detection**
- **Internet Search**
  - NLP for better quality results
  - semantic search
  - question answering
  - named entity resolution
  - image search
- **Knowledge Bases**
  - product graph construction
  - inference of unknown facts
- **Image, Speech, Video processing**
  - image classification
  - Image segmentation (e.g., seed identification in agriculture)
  - facial recognition
  - virtual assistants, chat bots

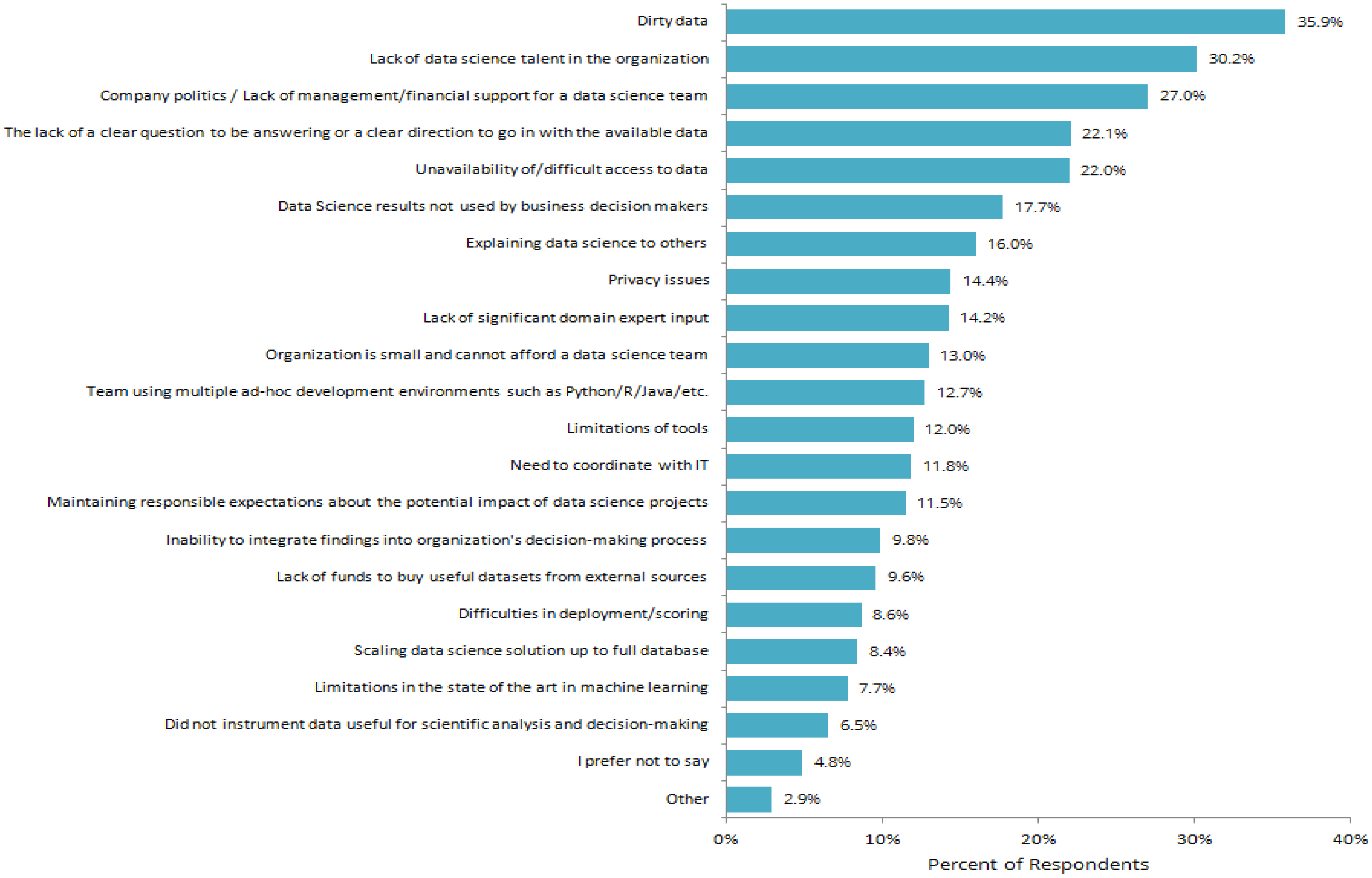
Applications of Data Science[3]

- **Airline Route Planning**
  - predict flight delay
  - decide which class of airplanes to buy
  - decide whether to directly land at the destination or take a halt in between
  - effectively drive customer loyalty programs
- **Targeted Advertising**
- **Gaming**
- **Augmented Reality**
- **Robotics**
- **Self-driving cars**

Challenges of Data Science[1]

- Data issues:
  - High variety of information & data is required for accurate analysis
  - Unavailability of/difficult access to data
  - Privacy issues
- Lack of domain expert
- Explaining data science to non-experts is difficult
- Data Science results not effectively used by business decision makers
- If an organization is very small, they can't have a Data Science team

Challenges of Data Science[4]



Data Science Process Example[6,7]

Example Problem

Leverage ML/AI to support medical image analysis

Discovery Phase

Start by asking a lot of questions:

- What are the major imaging techniques?
  - ➢ magnetic resonance imaging (MRI)
  - ➢ X-ray
  - ➢ computed tomography
  - ➢ mammography
- What are the major tasks in medical image analysis?
  - ➢ deep analysis of organ anatomy
  - ➢ detection or diverse disease conditions
- What part of the image analysis should we prioritize?
  - detect rumors

Discovery Phase

Identify relevant ML/AI tasks

- Can we identify regions of tumors in MRI images?
- Can we detect whether it is malignant or benign?

Discovery Phase

Identify relevant ML/AI tasks

- Can we identify regions of brain tumors in MRI images?
- Can we detect their type (e.g., meningioma, glioma, pituitary tumors)?

“image segmentation to identify, with high probability, regions with tumors”

Data Identification and Acquisition

MRI images are stored in the department's Picture Archiving and Communications System (PACS) in **DICOM** format. The department's **SQL** database contain tables with additional information about the images, i.e., image annotations.

Data Identification and Acquisition

MRI images are stored in the department’s an image archive in **DICOM** format. The department’s **SQL** database contain tables with additional information about the images, i.e., image annotations.

- How should we extract the images from the image archive?
- What format should you store the data to perform your analysis?
- Do you need to anonymize the data?

Data Preparation

- Data cleaning
  - remove corrupted images
  - remove irrelevant images (e.g., non-brain images)
- Data Transformation, e.g.
  - resize images
  - denoise images
- Verify all necessary information exists in the input data
  - What if we wanted to use image annotations

Exploratory Data Analysis

Target exploration to the ML/AI task at hand : segment images to identify regions with brain tumors

Exploratory Data Analysis

Target exploration to the ML/AI task at hand : segment images to identify regions with brain tumors

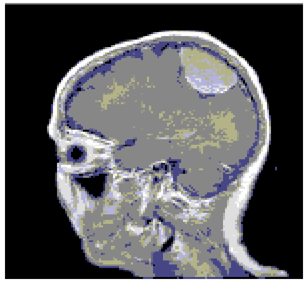
Visualize images to get a feeling of how they look

Exploratory Data Analysis

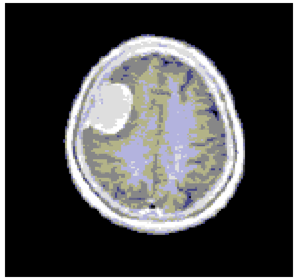
Target exploration to the ML/AI task at hand : segment images to identify regions with brain tumors

Visualize images to get a feeling of how they look

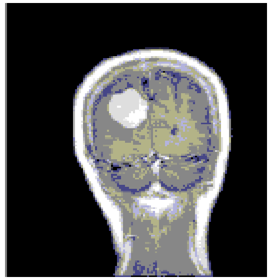
The images look very different depending on their direction (sagittal, axial, coronal)



Sagittal view



Axial view



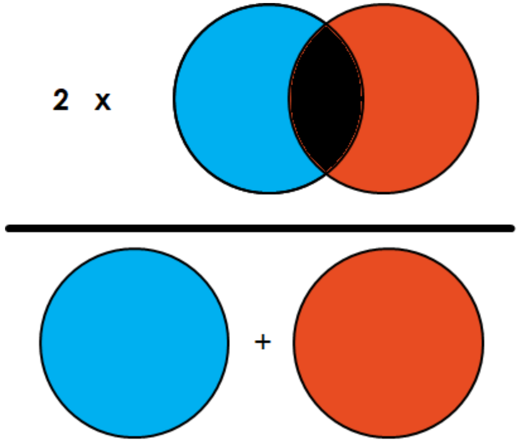
Coronal view

Model Building

- Choose candidate ML/AI model
  - A separate LinkNet[8] for each image angle (sagittal, coronal, axial)
- Generate model input
  - Group images by image angle
  - Normalize
  - Augment
- Create training dataset (e.g., mask images to label the class and area of interest)
- Set aside test images

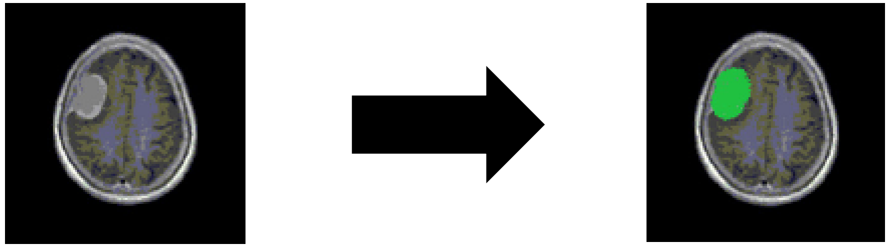
Assume the best model has a Dice score of 0.79

Communicate Results



- Provide details on the performance of the model
  - explain Dice score (computational formula, properties etc.)
  - provide more detailed accuracy numbers (e.g., confusion matrix, components of Dice score)

- Appropriately visualize segmented images, e.g.
  - Include examples of successful and unsuccessful segmentation



Data Science Process Example

https://bitbucket.org/diip20201/tutorials/src/master/data_science_overview/

References

1. https://www.guru99.com/data-science-tutorial.html
2. https://www.javatpoint.com/data-science
3. https://www.edureka.co/blog/data-science-applications/
4. http://businessoverbroadway.com/wp-content/uploads/2018/03/challenges.png
5. https://docs.microsoft.com/fr-fr/azure/machine-learning/team-data-science-process/lifecycle-business-understanding
6. https://medium.com/activewizards-machine-learning-company/top-7-data-science-use-cases-in-healthcare-cddfa82fd9e3
7. https://paperswithcode.com/task/brain-tumor-segmentation/codeless?page=5
8. A. Chaurasia and E. Culurciello: LinkNet: Exploiting encoder representations for efficient semantic segmentation, VCIP 2017