Data Science Projekt:

1. **Data sets from two resources:**

* ([https://www.kaggle.com/kmader/skin-cancer-mnistham10000?select=HAM10000\_images\_part\_2](https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000?select=HAM10000_images_part_2) (MNIST1000)
* <http://www.cs.rug.nl/~imaging/databases/melanoma_naevi/> (datasetWeb)
* Sind sortiert in verschiedene Ordner die Krebsart vorhersagen
* Splitting in Test und Trainingsdaten?
  1. Linked in Github
  2. Wie viele, welches Format, RDF

Haben ImageID, PatientID, Krebsart, Alter, Geschlecht, Stelle (MNIST)

Krebstyp (datasetWeb)

* 1. Welche Krebsarten haben wir?
* Dataset Skin Cancer MNIST: HAM10000: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc)
* Dataset PH2: Melanoma (mel), nevus (nv)
* Preprocessing: Use pretrained model DIP-VAE decoder to get latent features. (Disentangled Inferred Prior Variational Autoencoder)

1. **Data Quality**

* Wie groß sind einzelnen Gruppen von Krebsarten
  1. Completeness
  2. Consistency
  3. Uniqueness
  4. Timeliness
  5. Validity
  6. Accuracy

1. **Algorithms**
   1. Jupyter notebook in github
   2. Mit Google laufen lassen
   3. Radnom Forrest Classifier
   4. Logistic regression classifier
2. **Results**
   1. Accuracy of Models

Wie visualisieren wir Ergebnisse, was sind die Parameter?

* 1. Compare

Welches Model ist besser und warum?

Wo könnten Fehler liegen?

* 1. Ok to use

Kann die Methode in der Realität angewendet werden?

1. **Writing the report**

Aufbau:

1. Einleitung
2. Datensets
3. Algorithmen
4. Results
5. Diskussion

Vorgehen:

* Gemergte Datensätze sind erstellt (bis auf splitting Test/Training)
* RDF erstellen
* Datenqualität messen
* Github erstellen
* Jeder macht eine Methode (benutzen beide das pretrained model vorher)
* Results zusammenfassen
* Bericht schreiben (Aufgeteilt?)