Autoencoders

Die Anatomie einer Steuererklärung

Phillipp Gnan

Wien, 14.11.2024

(c) http://www.garagezero.eu/20170805193813_gartner-quadrant-big-data/

Roadmap

1. Autoencoder Basics

2. Die Anatomie einer Steuererklärung

3. SAS-Case Study

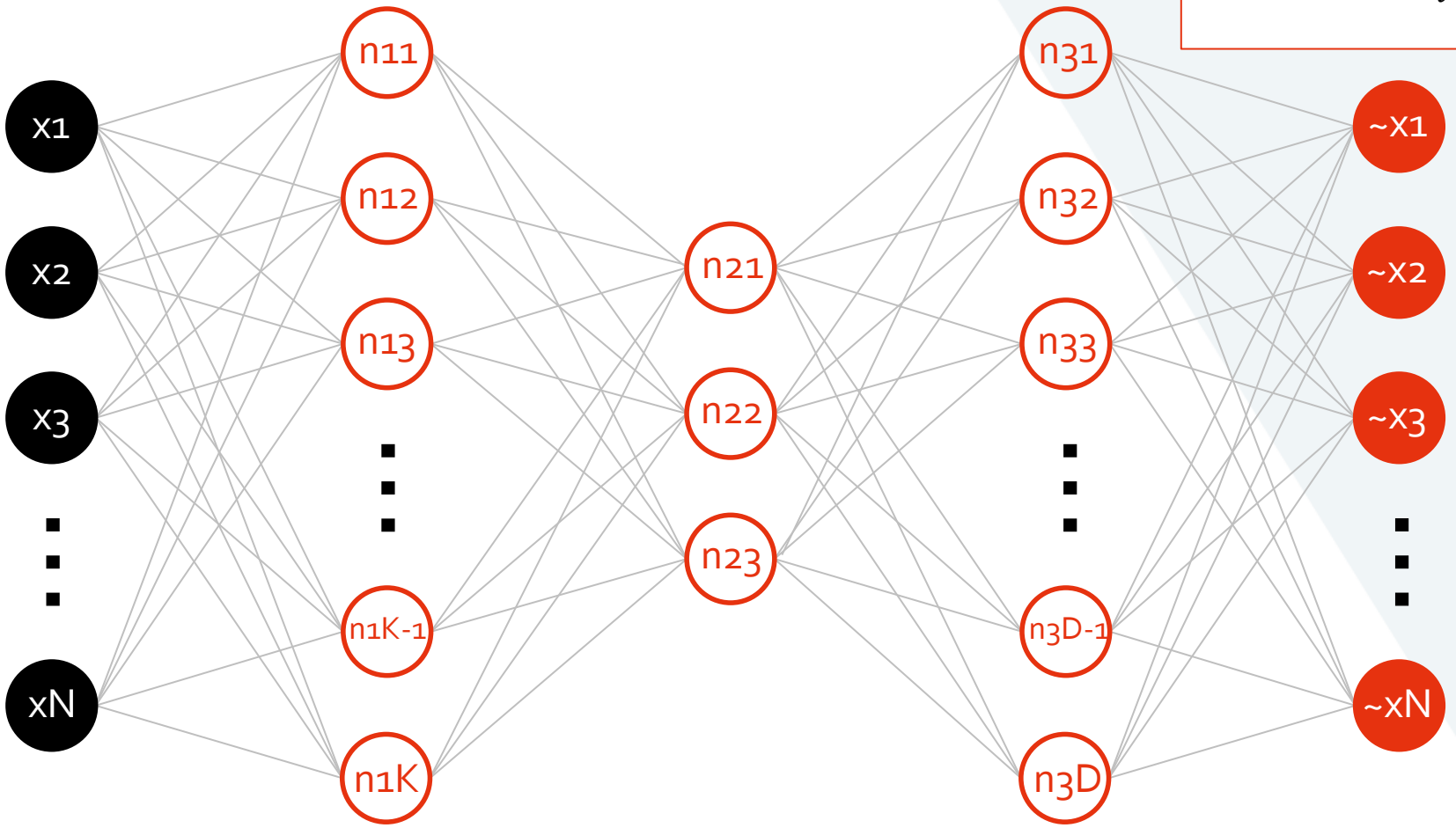
1. Autoencoder Basics

Autoencoder Intuition

Bilder: Gemini / Imagen 3

Autoencoder Architektur

$$Reconstr\ Err_i = \sum_{n=1}^{N} L\ (x_{ni}, {\sim}x_{ni})$$



Input (N)

Hidden Layer 1 (K)

Bottleneck / Code (3)

Hidden Layer 3 (D)

Reconstructed Input (N)

Collage: Twitter/X-User:in @teenybiscuit

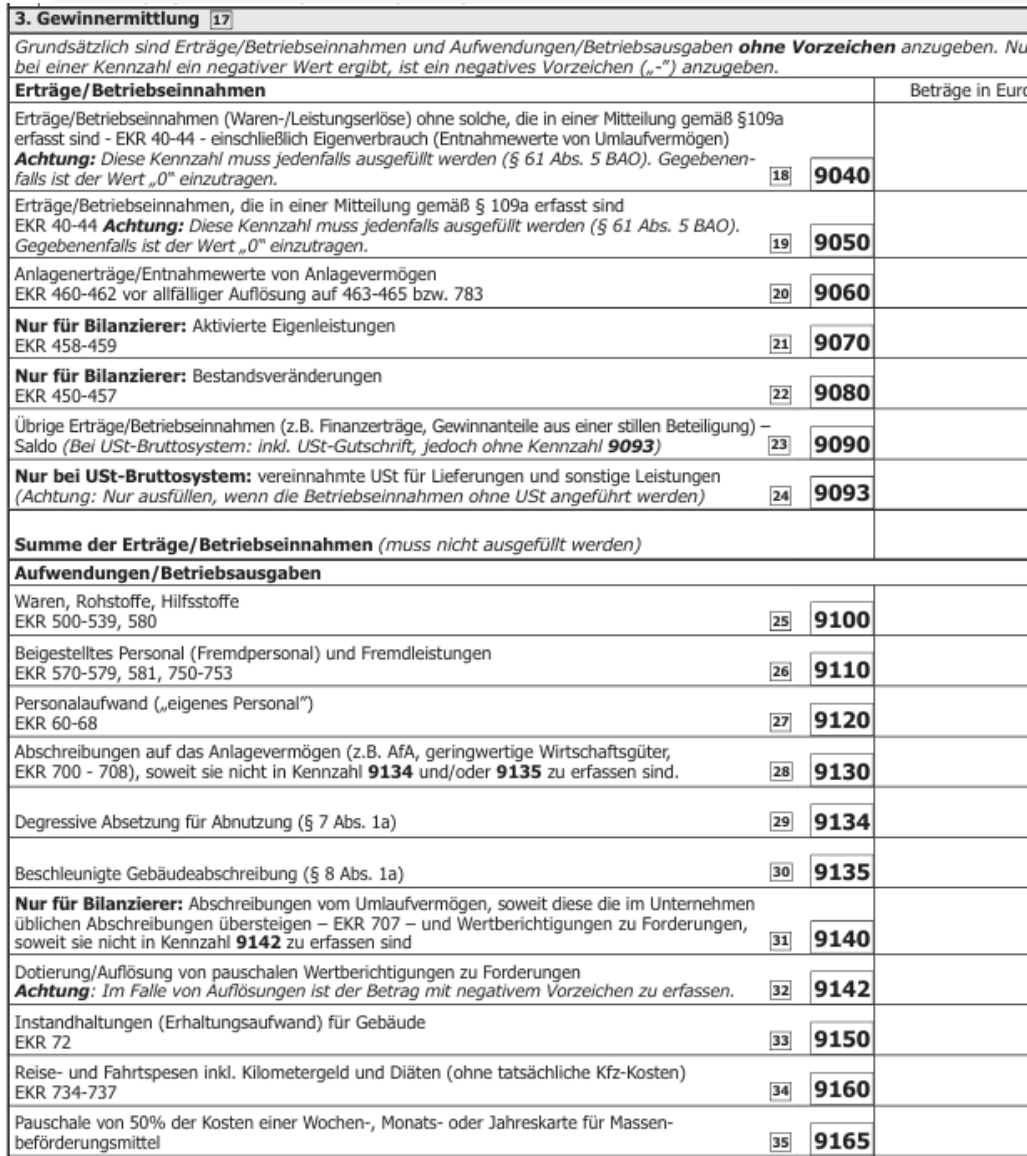
2. Die Anatomie einer Steuererklärung

Fragestellung in der Finanzverwaltung

- Die Finanzverwaltung erhält laufend unzählige „Steuererklärungen“
  - Einkommenssteuererklärungen, Lohnzettel,…
- Nur ein Teil kann überprüft werden

Welche Erklärungen sind „ungewöhnlich“?

Autoencoder



3. Gewinnermittlung 17

Grundsätzlich sind Erträge/Betriebseinnahmen und Aufwendungen/Betriebsausgaben ***ohne Vorzeichen*** *anzugeben. Nu*
bei einer Kennzahl ein negativer Wert ergibt, ist ein negatives Vorzeichen („-“) anzugeben.

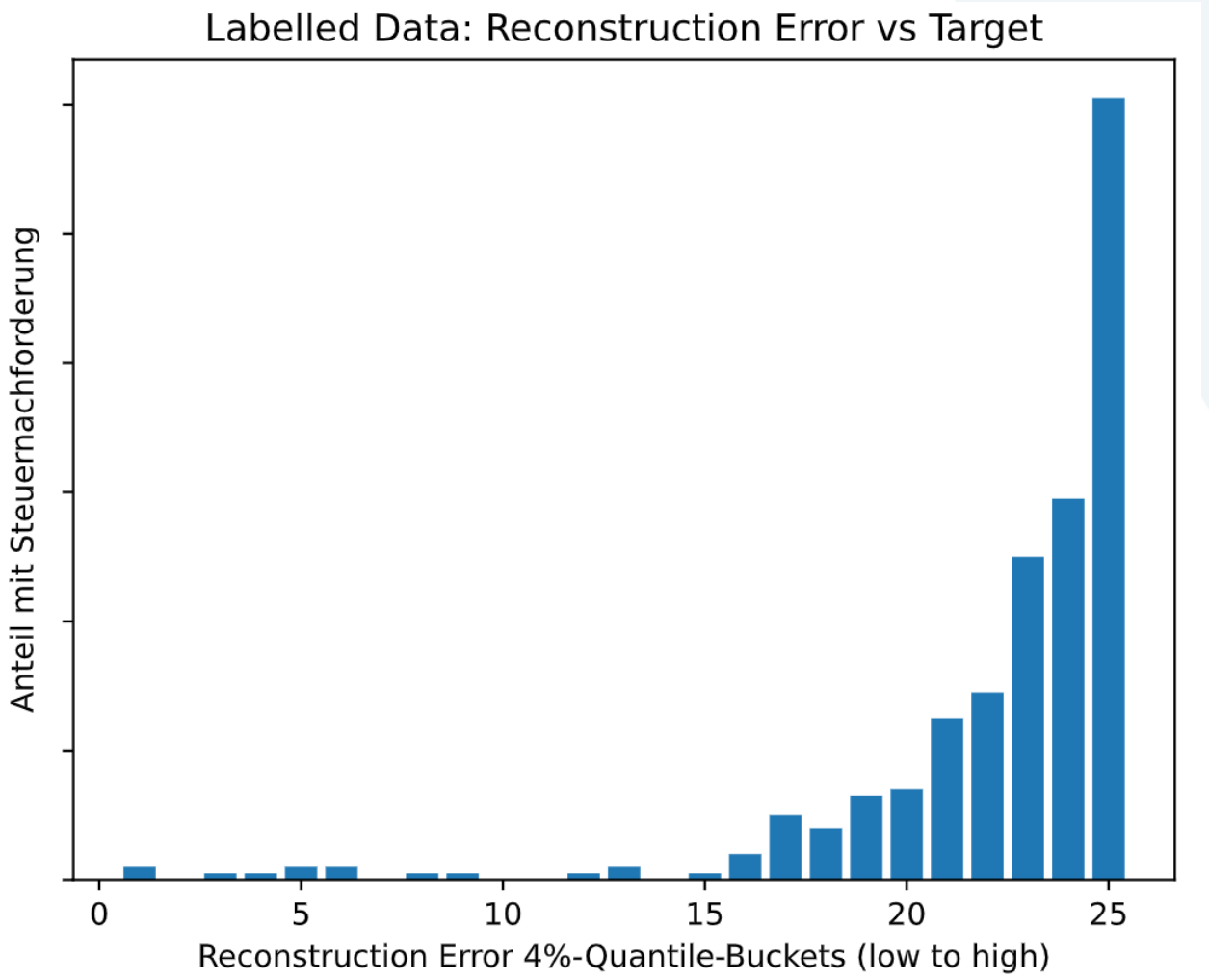
Erträge/Betriebseinnahmen			Beträge in Eur
Erträge/Betriebseinnahmen (Waren-/Leistungserlöse) ohne solche, die in einer Mitteilung gemäß §109a erfasst sind - EKR 40-44 - einschließlich Eigenverbrauch (Entnahmewerte von Umlaufvermögen) ***Achtung:*** *Diese Kennzahl muss jedenfalls ausgefüllt werden (§ 61 Abs. 5 BAO). Gegebenenfalls ist der Wert „0“ einzutragen.*	18	**9040**	
Erträge/Betriebseinnahmen, die in einer Mitteilung gemäß § 109a erfasst sind EKR 40-44 ***Achtung:*** *Diese Kennzahl muss jedenfalls ausgefüllt werden (§ 61 Abs. 5 BAO). Gegebenenfalls ist der Wert „0“ einzutragen.*	19	**9050**	
Anlagenerträge/Entnahmewerte von Anlagevermögen EKR 460-462 vor allfälliger Auflösung auf 463-465 bzw. 783	20	**9060**	
Nur für Bilanzierer: Aktivierte Eigenleistungen EKR 458-459	21	**9070**	
Nur für Bilanzierer: Bestandsveränderungen EKR 450-457	22	**9080**	
Übrige Erträge/Betriebseinnahmen (z.B. Finanzerträge, Gewinnanteile aus einer stillen Beteiligung) – Saldo *(Bei USt-Bruttosystem: inkl. USt-Gutschrift, jedoch ohne Kennzahl* ***9093****)*	23	**9090**	
Nur bei USt-Bruttosystem: vereinnahmte USt für Lieferungen und sonstige Leistungen *(Achtung: Nur ausfüllen, wenn die Betriebseinnahmen ohne USt angeführt werden)*	24	**9093**	
Summe der Erträge/Betriebseinnahmen *(muss nicht ausgefüllt werden)*			
Aufwendungen/Betriebsausgaben			
Waren, Rohstoffe, Hilfsstoffe EKR 500-539, 580	25	**9100**	
Beigestelltes Personal (Fremdpersonal) und Fremdleistungen EKR 570-579, 581, 750-753	26	**9110**	
Personalaufwand („eigenes Personal“) EKR 60-68	27	**9120**	
Abschreibungen auf das Anlagevermögen (z.B. AfA, geringwertige Wirtschaftsgüter, EKR 700 - 708), soweit sie nicht in Kennzahl **9134** und/oder **9135** zu erfassen sind.	28	**9130**	
Degressive Absetzung für Abnutzung (§ 7 Abs. 1a)	29	**9134**	
Beschleunigte Gebäudeabschreibung (§ 8 Abs. 1a)	30	**9135**	
Nur für Bilanzierer: Abschreibungen vom Umlaufvermögen, soweit diese die im Unternehmen üblichen Abschreibungen übersteigen – EKR 707 – und Wertberichtigungen zu Forderungen, soweit sie nicht in Kennzahl **9142** zu erfassen sind	31	**9140**	
Dotierung/Auflösung von pauschalen Wertberichtigungen zu Forderungen ***Achtung****: Im Falle von Auflösungen ist der Betrag mit negativem Vorzeichen zu erfassen.*	32	**9142**	
Instandhaltungen (Erhaltungsaufwand) für Gebäude EKR 72	33	**9150**	
Reise- und Fahrtspesen inkl. Kilometergeld und Diäten (ohne tatsächliche Kfz-Kosten) EKR 734-737	34	**9160**	
Pauschale von 50% der Kosten einer Wochen-, Monats- oder Jahreskarte für Massenbeförderungsmittel	35	**9165**	

Anomaly-Detection mit Autoencoders

- Wir lernen die *typische* „Anatomie“ von bestimmten Erklärungen
  - zB: Beilage Vermietung und Verpachtung der Einkommensteuererklärung, Lohnzettel,…
  - Anatomie: Verhältnis von diversen Angaben in einer Erklärung

1. **Angaben** in einer Erklärung auf eine Handvoll Kennzahlen **komprimieren**
2. **Dekomprimieren** von Kennzahlen: wieder auf ursprüngliche Angaben rückrechnen
3. **Fehler bei Rückrechnung**
   1. klein für „gewöhnliche“ Angaben (Schäferhund)
   2. groß für „ungewöhnliche“ Angaben (Dackel)

Beispiel: Einkommensteuer-Beilage Vermietung und Verpachtung



Challenges

- **Missing Values**
  - Im Steuerkontext: Interpretation als 0 häufig naheliegend
- **Sparse Data**
  - Features mit hoher sparsity / mit wenig Variation häufig weniger aussagekräftig
  - Feature selection: sparsity theshold
- **Schwarze Schafe** in den Trainingsdaten
  - Autoencoder soll möglichst nur von „normalen“ Observationen lernen
    → Fokus auf geprüfte Erklärungen oder Ausschluss von Fällen mit Steuerkorrektur

Challenges

- **Architektur / Bottleneck**
  - Restriktiv ↔ umfangreich:
    normale Observationen schlecht rekonstruiert ↔ Outliers zu gut rekonstruiert
  - „Ökonomische Intuition“ für initial guess
  - Labelled Data für Cross-Check

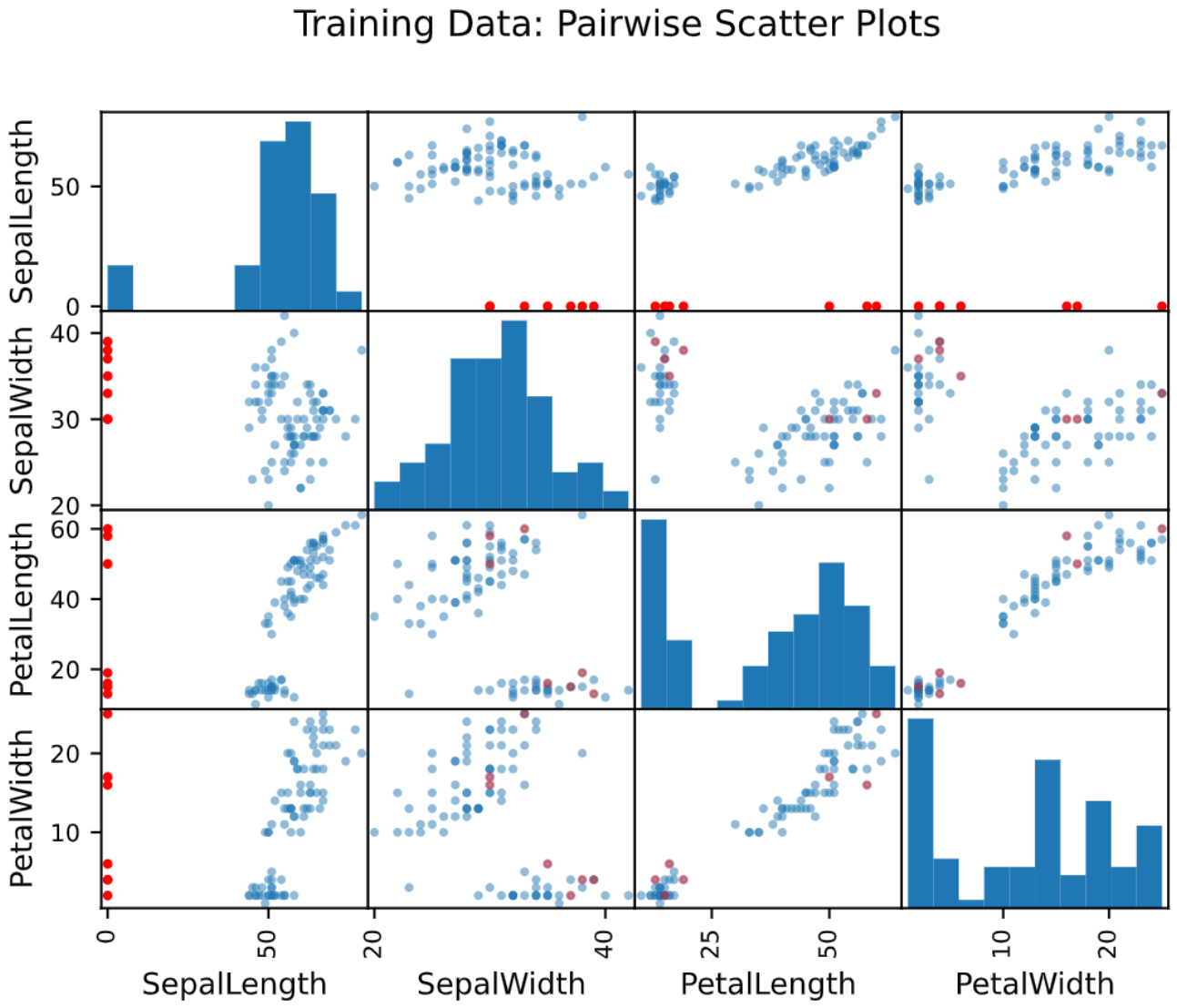
- **Umsetzung in SAS:** nicht alles „out of the box“ implementiert → Siehe Case Study!

3. SAS Case Study

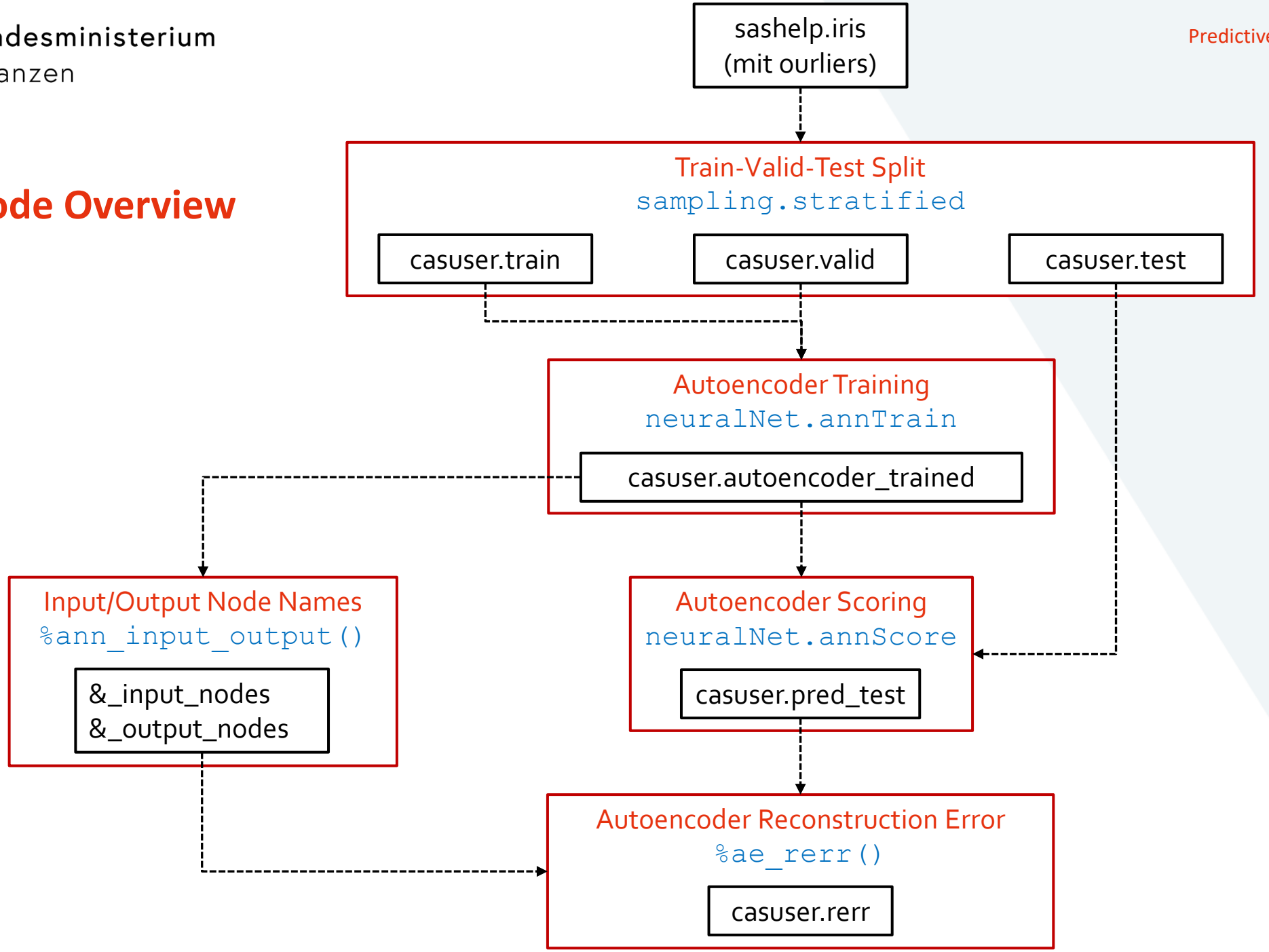
Data – sashelp.iris

- **150 Observationen**: Iris Setosa, Iris Versicolor, Iris Virginica à 50 Samples
- **4 Features**: für Länge und Breite des Kelchblatts bzw des Kronblatts
- **Outliers**: in jeder 10. Zeile → Kelchblattlänge auf 0 gesetzt
- **Train – Validation – Test**
  - 60% Train (90 Obs)
  - 20% Validation (30 Obs)
  - 20% Test (30 Obs)

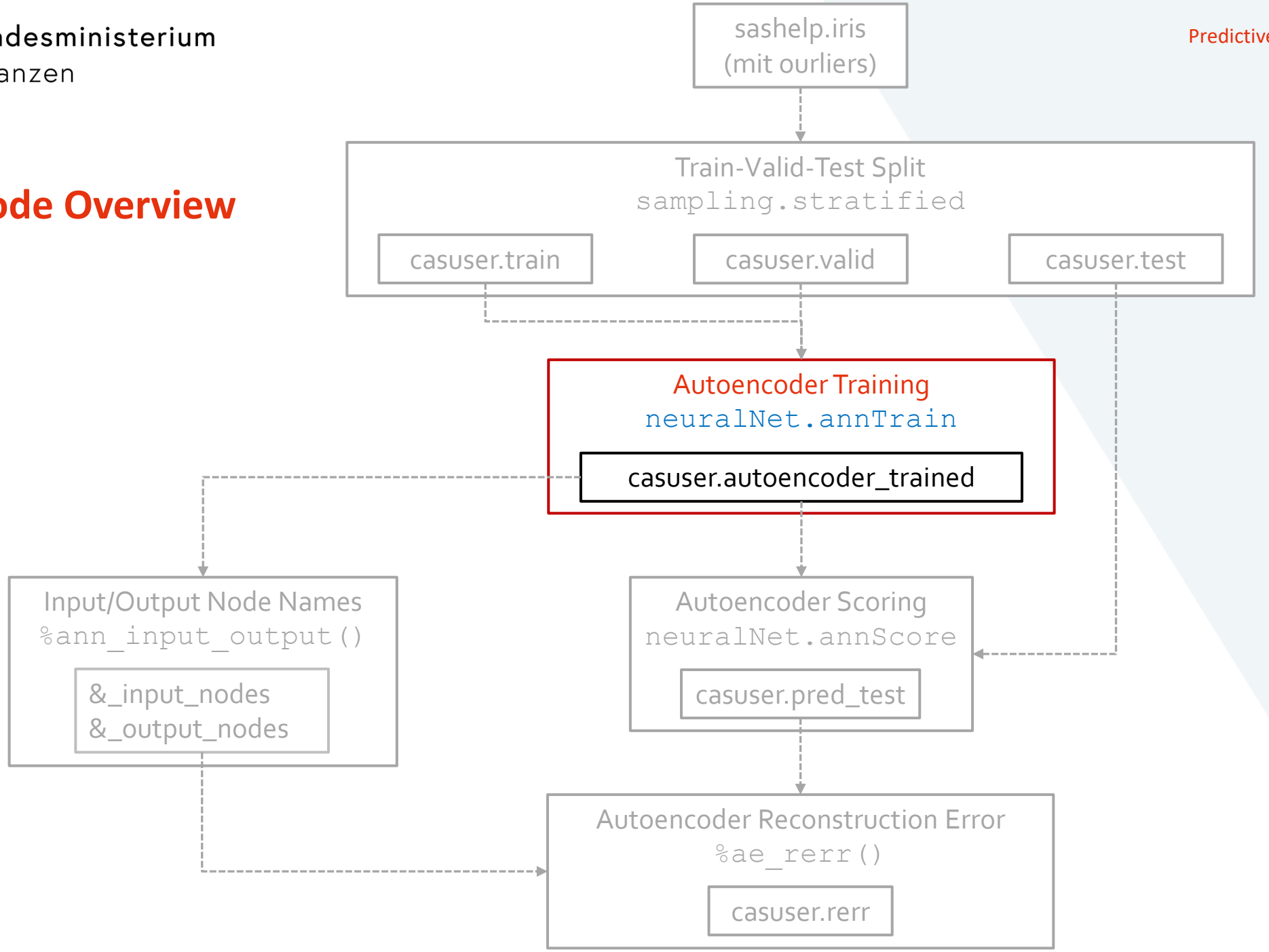
Bild: Gemini / Imagen 3


Training Data: Pairwise Scatter Plots
SepalLength
SepalWidth
PetalLength
PetalWidth
0
50
20
40
25
50
10
20

Code Overview



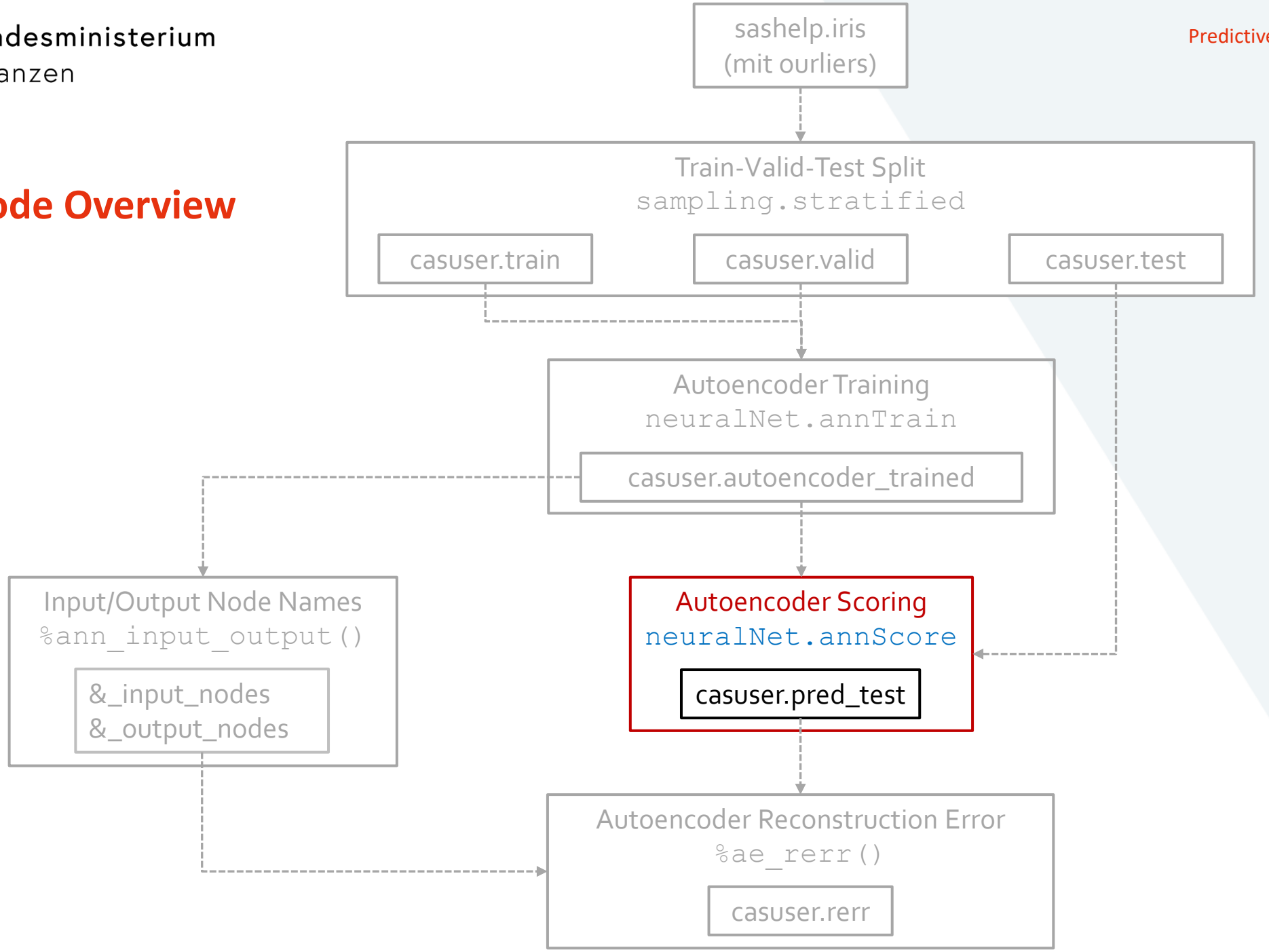
Code Overview



Autoencoder Training – neuralNet.annTrain

```
proc cas;
    neuralNet.annTrain result=r /
    /*INPUTS*/
        table={name="train"}
        inputs={"sepallength","sepalwidth","petallength","petalwidth"}
        std="MIDRANGE"
    /*ARCHITECTURE*/
        hiddens={2, 1}
        combs={"LINEAR"}
        acts={"RECTIFIER"}
        targetAct="IDENTITY"
    /*OPTIMIZER*/
        seed=123
        errorFunc="NORMAL"
        randDist="UNIFORM"
        scaleInit=1
        nloOpts={
            algorithm="SGD",
            optmlOpt={maxIters=200, fConv=1e-10},
/*          lbfgsOpt={numCorrections=6}, */
            sgdOpt={adaptiveDecay=0.99,
                    adaptiveRate=True,
                    learningRate=0.1,
                    miniBatchSize=5,
                    momentum=0.95},
            validate={frequency=1,
                      stagnation=3}}
        validTable={name="valid"}
    /*OUTPUT*/
        casout={name="autoencoder_trained", replace=True}
        encodename=True
        ;
```

Code Overview



Autoencoder Scoring

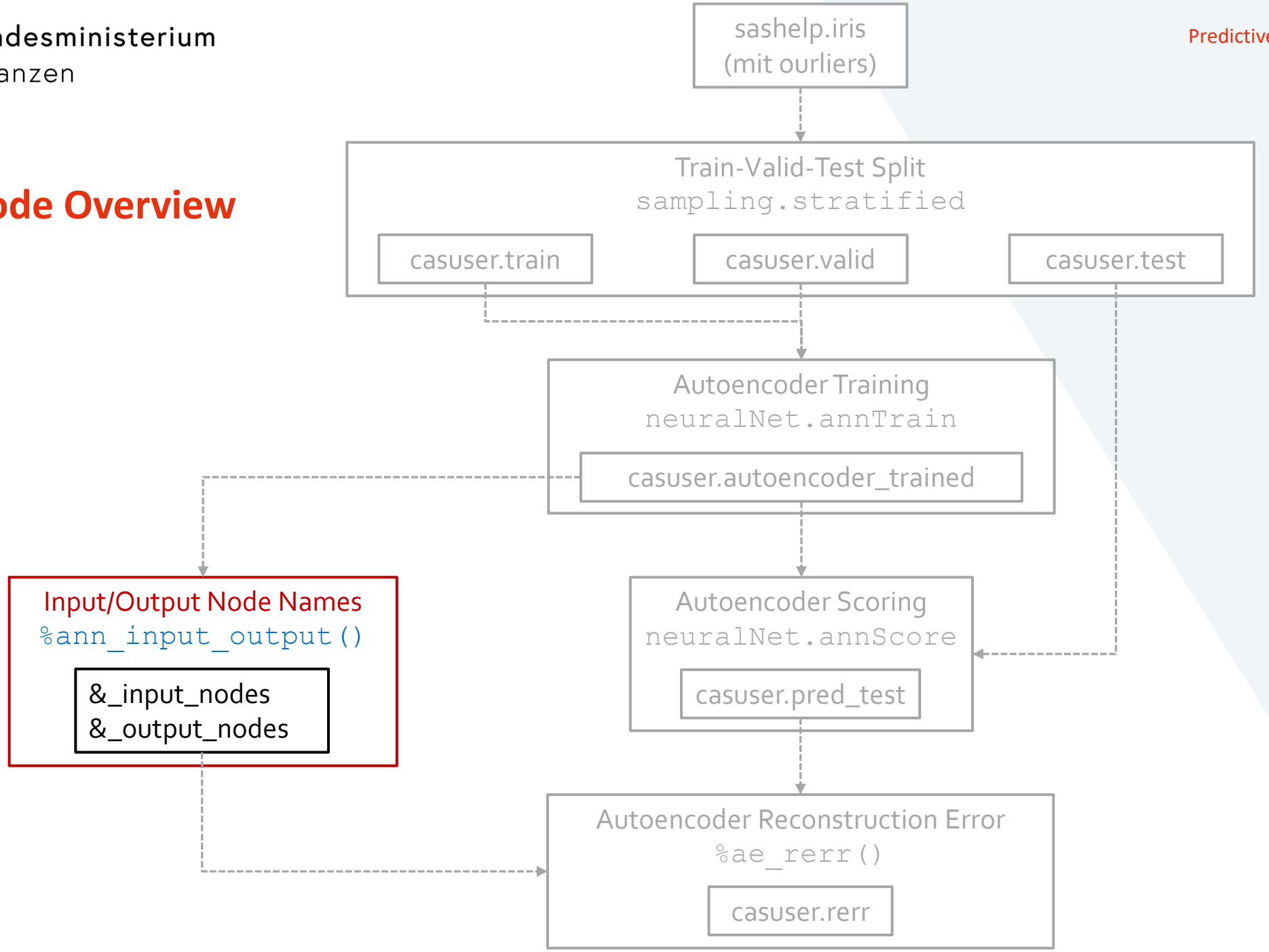
```
proc cas;
    neuralNet.annScore /
        table={name="test"}
        modelTable={name="autoencoder_trained"}
        copyvars={"id"}
        listnode="ALL"

        casOut={name="pred_test", replace=TRUE}
        ;
quit;
```

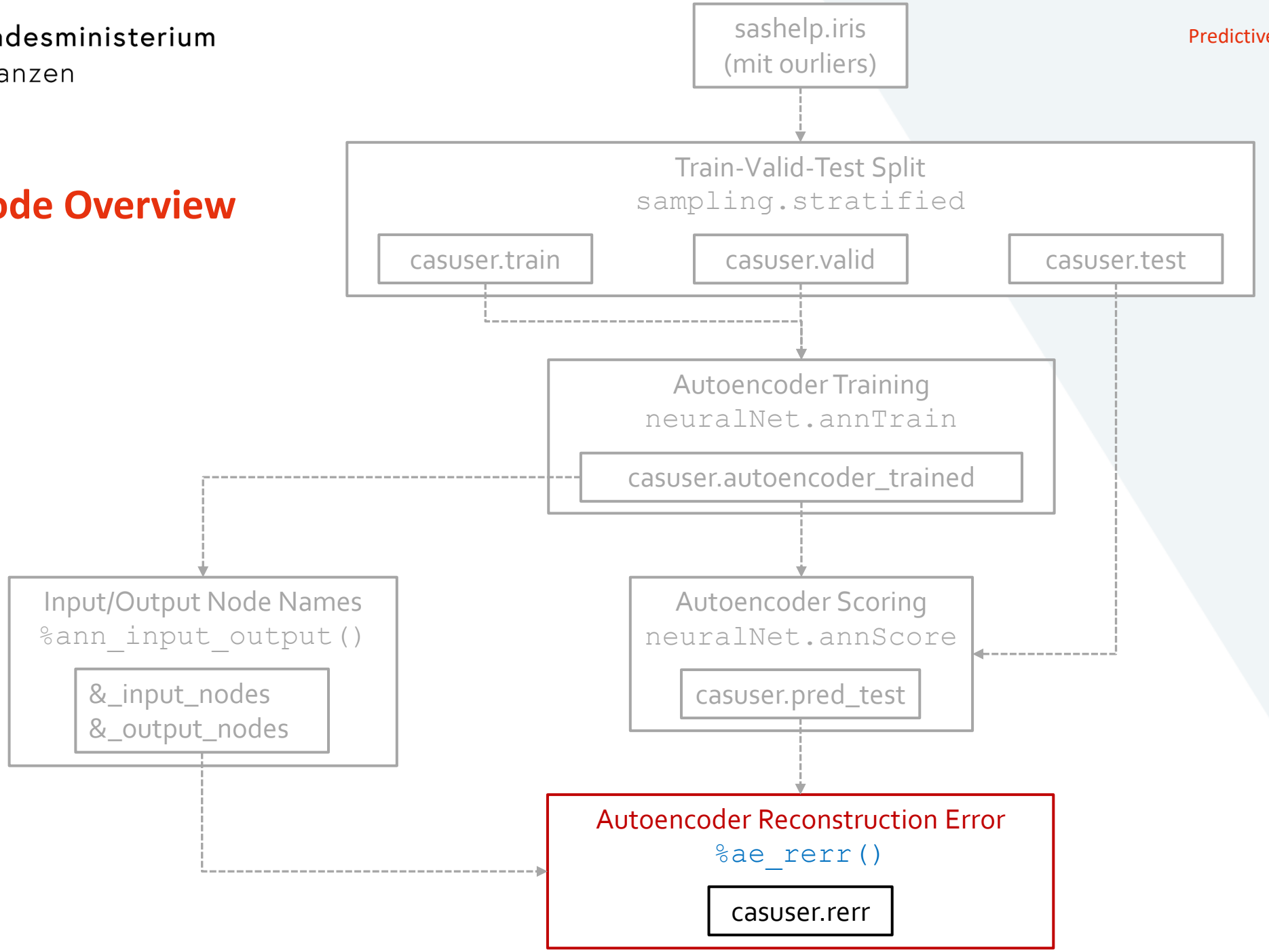
casuser.pred_test

Obs	ID	_Node_0	_Node_1	_Node_2	_Node_3	_Node_4	_Node_5	_Node_6	_Node_7	_Node_8	_Node_9	_Node_10
1	11	0.1139240506	-0.090909091	-0.888888889	-0.916666667	0.5934752457	0.1373982592	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
2	14	0.2911392405	0.6363636364	-0.777777778	-0.916666667	0.5925117685	0.2192327781	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
3	16	0.2658227848	-0.090909091	-0.777777778	-0.916666667	0.5723420919	0.1199868831	0.0053201529	0.0863624451	0.1973638335	-0.802232885	-0.858807189
4	18	0.0886075949	-0.090909091	-0.962962963	-1	0.620137879	0.1472500381	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
5	21	0.2405063291	-0.090909091	-0.851851852	-0.916666667	0.5861758739	0.1202879616	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
6	27	0.3670886076	0.2727272727	-0.814814815	-0.75	0.5633461358	0.1355836283	0.0073509865	0.0884839641	0.1954329822	-0.795824084	-0.852514006
7	30	-1	0	-0.814814815	-0.916666667	0.5859981053	0.3195716052	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
8	32	0.2658227848	0.2727272727	-0.814814815	-0.916666667	0.5893438212	0.1697603699	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
9	34	0.1898734177	0.0909090909	-0.888888889	-0.916666667	0.5982839266	0.1518454362	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
10	38	0.2405063291	0	-0.814814815	-1	0.5946621661	0.1444493817	0	0.0808047246	0.2024220636	-0.819021953	-0.875293372
11	59	0.7721518987	0.0909090909	0.3703703704	0.0833333333	0.2086534429	0.0095056852	0.3528684457	0.4494302556	-0.1330739	0.2945423032	0.2181815546

Code Overview



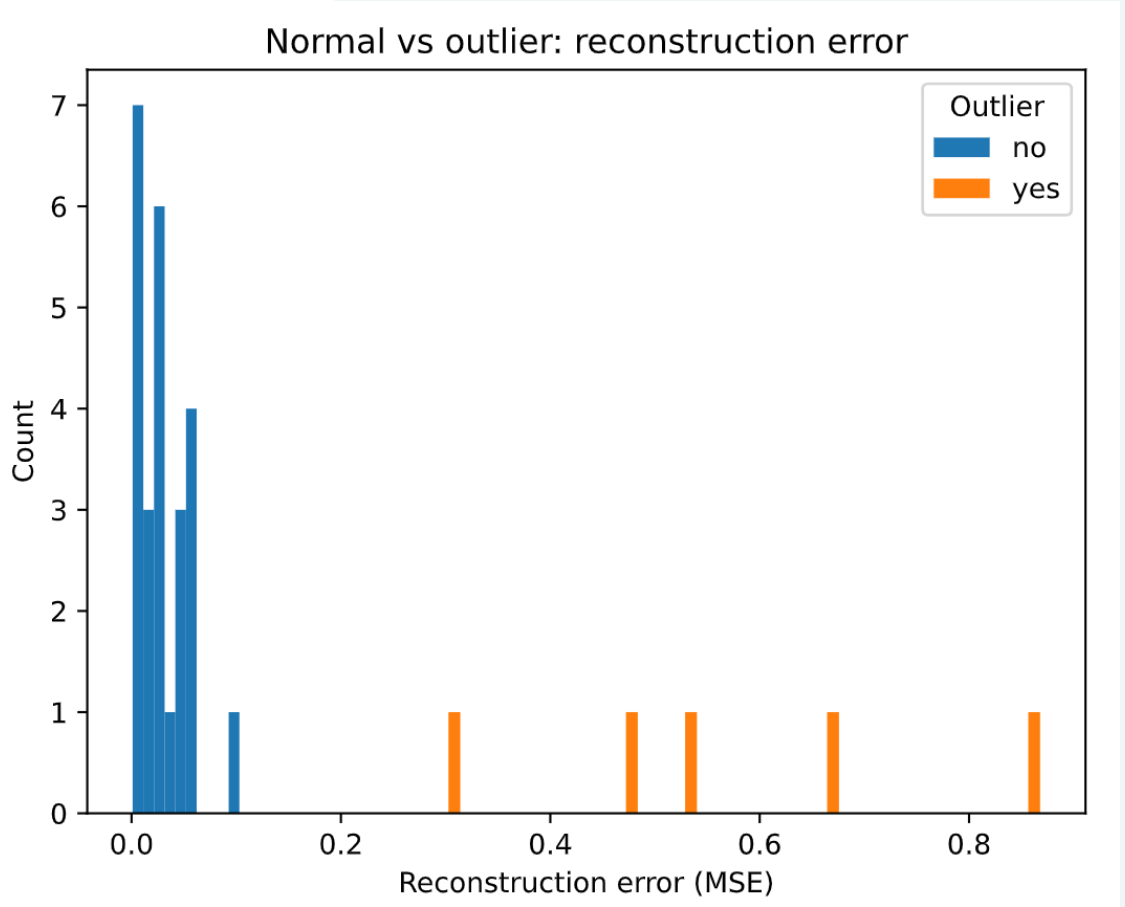
Code Overview



Autoencoder Reconstruction Error

```
%ae_rerr(pred_table=casuser.pred_test,
         input_nodes=&_input_nodes., output_nodes=&_output_nodes.,
         table_out=casuser.rerr,
         keep=id);
```

- Implementierter loss: mean squared error



Also good to know

- neuralNet.anncode
  - Generiert data-step score code
  - Scoring ohne CAS / auch unter SAS 9.4