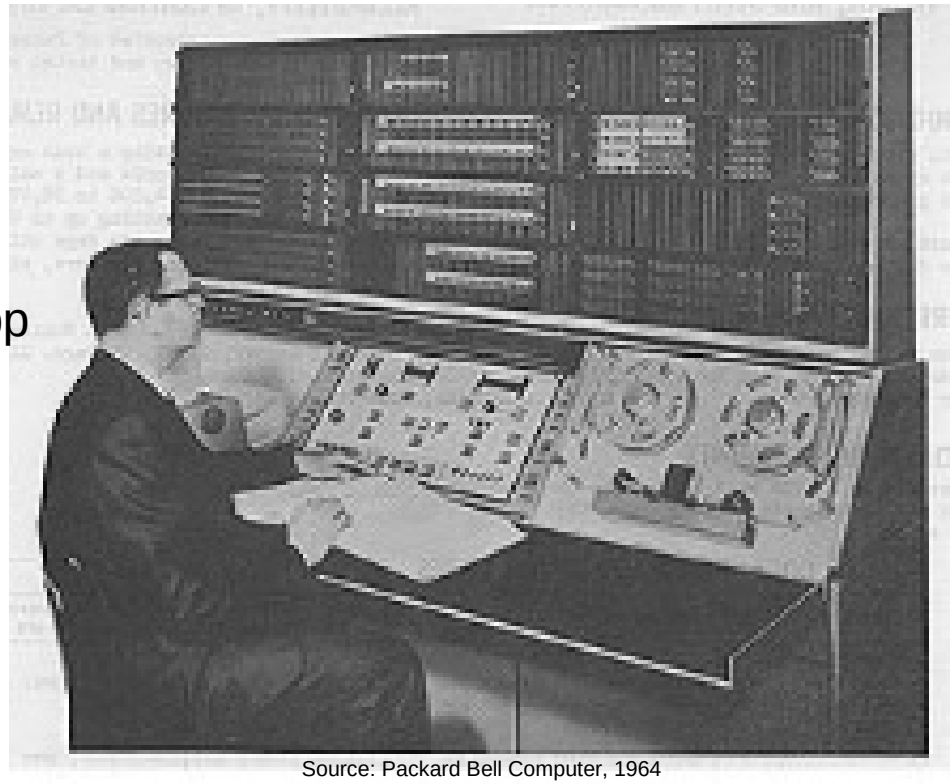
Plausibilitätsprüfung mit Machine Learning
Contrôle de plausibilité avec machine learning

Source: CC0 Public Domain

Team ‘Plausi++’:

Christian Ruiz
Christine Ammann Tschopp
Elisabeth Kuhn
Laurent Inversin
Mehmet Aksözen
Stefan Rüber



Source: Packard Bell Computer, 1964

Abgrenzung / Délimitation

Die folgende Präsentation besteht aus Testelementen, welche vor dem Pilotprojekt «Plausi++» gemacht wurden. Das Pilotprojekt läuft seit April 2018 als Teil der «Dateninnovationsstrategie».

La présentation suivante comporte des tests, qui étaient fait avant le projet pilote «Plausi++». Le projet pilote a débuté en Avril 2018 faisant partie de la «stratégie d'innovation sur les données ».

Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Improving Data Validation using Machine Learning

Prepared by Dr. Christian Ruiz, Swiss Federal Statistical Office, Switzerland[1]

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Switzerland_RUIZ_Paper.pdf

Übersicht / Survol

Teil/Partie I: Überblick über ML / Survol sur ML

Teil/Partie II: Grundidee von Plausi++ / Idée de base de Plausi++

Teil/Partie III: Feedback-Mechanismus / Mécanisme de feedback

«Plausi»

Source: CC0 Public Domain

- Manuell / Manuelle
  (Verschiedene Lösungen / Plusieurs variantes)
- Basierend auf Regeln /Basée sur des règles
  (Verschiedene Lösungen / Plusieurs variantes)
- Idee/Idée: Automatische Erkennung / Reconnaissance automatique?

Machine Learning

Hype? Algorithmen aus 70er? / Vieux algorithmes?

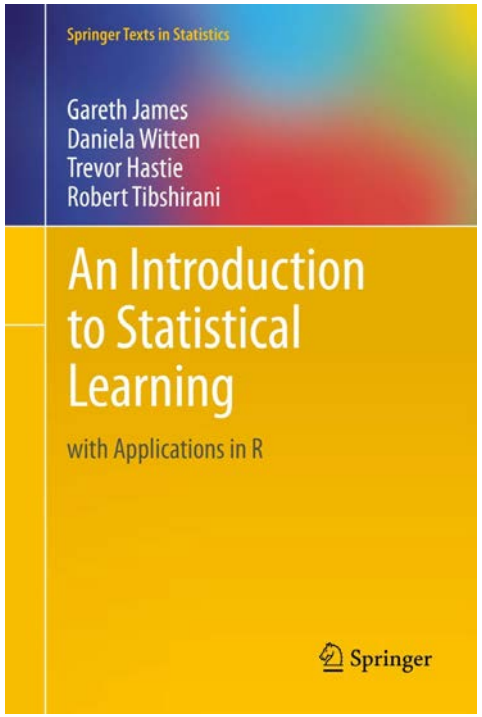
Jein / Oui et non:

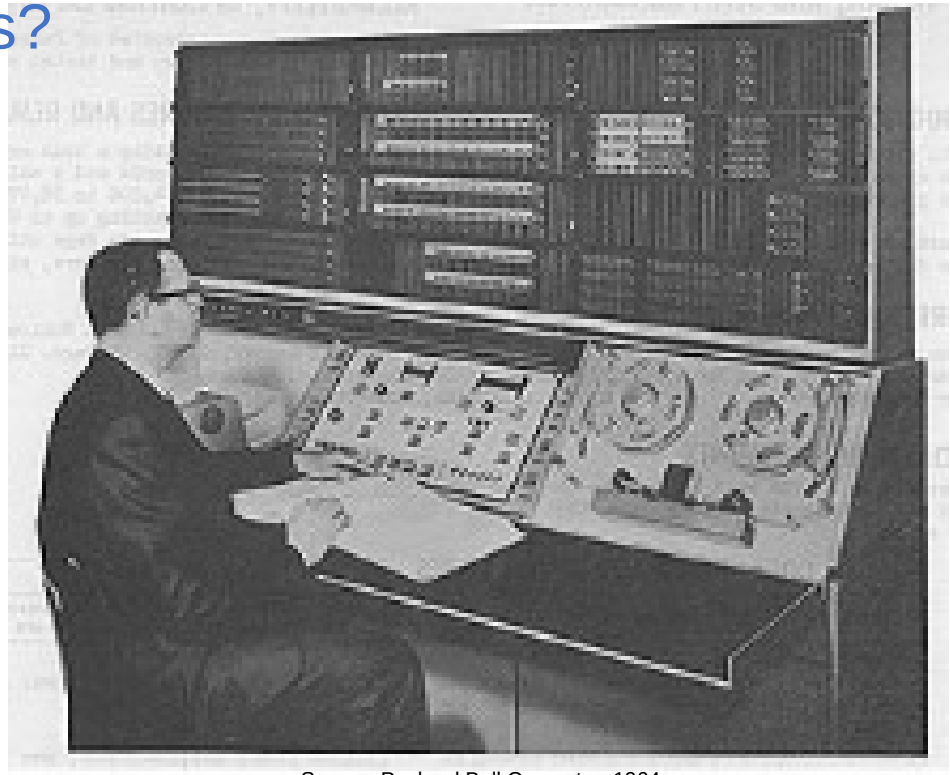
Rechenpower / Pouvoir de calcul

Weiterentwicklung / Développement

Geeignetere Software / Logiciels mieux adaptés

Offene Forschung / Recherche ouverte



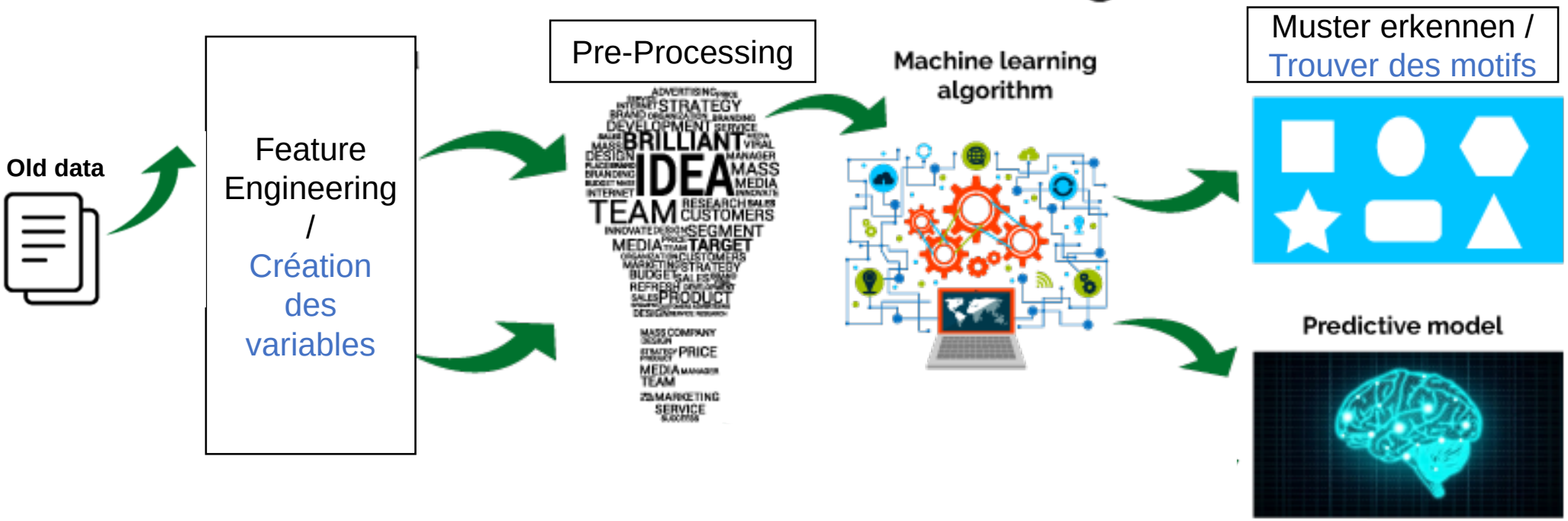


Source: Packard Bell Computer, 1964

PHASE: TRAINING

Machine Learning

Supervised learning



Lernen? / Apprendre?

- Grosse Datenmengen füttern / Insérer une grande quantité des données
- Muster aus Daten erkennen / Reconnaître des motifs
- Vorhersage von Werten / Prédiction des données
- Prédiction n'équivaut pas explication

(Source: CC0 Public Domain with modifications)

PHASE: TESTING

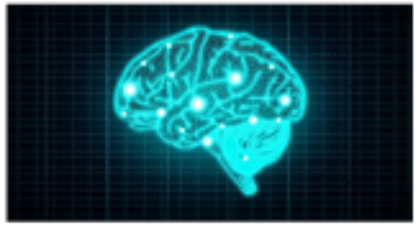
Machine Learning

Supervised learning

Feature Engineering / Création des variables

Pre-Processing

Predictive model



New data

Predicted data

(Source: CC0 Public Domain with modifications)

ML ist hier ein Werkzeug / ML est ici un outil

Source: CC0 Public Domain

Source: CC0 Public Domain

Muss zunächst kreiert werden /
Doit être crée d'abord

Und kann danach benutzt werden /
Et peut être utilisé ensuite

Mehr Mensch als Maschine / Plus d'être humain que de machine

Mensch/être humain (e.h.): Verstehen der Daten! / Comprendre les données!

Mensch/e.h.: Feature Engineering und/et Pre-Processing

Mensch/e.h.: Wahl der **geeigneten** Algorithmen / Choix des algorithmes **appropriés**

Maschine/Machine: Berechnung / Calcul

Maschine/Machine: Aufbereitung für Endnutzer / Préparation pour utilisateur final

Mensch/e.h.: Kalibrierung und Entscheidungen / Calibrations et décisions

Mensch/e.h.: Produktionsintegration / Intégration au système de production

Teil II: Grundidee für Plausi++ / Partie II: Idée de base de Plausi++

Experiment, welches scheinbar so noch nicht existiert / Experience qui semble ne pas encore exister:

1) Auswahl von Variablen in einem BFS-Datenset / Choix des variables dans un dataset de l'OFS
2) Prädiktion durch ML-Algorithmen / Prédiction des données avec des algorithmes ML
3) Vergleich zw. vorhergesagten und neuen Daten / Compar. entre données projetées et nouvelles données

Wenn Abweichungen: Erhebungsfehler oder Ausnahmefälle / Si différence: Faute de relevé ou cas exceptionnel

Beispieldatensatz / Données utilisés comme exemple

Personalkategorie erklärt durch Geschlecht, VZÄ, Fachbereich, Alter, Nationalität, ETH/Uni

Abhängige Variable hat 4 Klassen / Variable dépendante a 4 classes:

P: Professoren / Professeurs

U: Uebrige Dozierende / Autres enseignants

A: Wissenschaftliche Mitarbeiter / Collaborateurs scientifiques

D: Direktion und Administrativpersonal / Direction et personnel administratif

Beispiele/ Exemples

Sex	FTE	Field	Age	Swiss	Uni	p(A|·)	p(D|·)	p(P|·)	p(U|·)	Observed
M	0.75	4.Exact	27	Yes	Yes	0.89	0.11	0.00	0.00	A
F	0.80	5.Med.	26	No	Yes	0.66	0.34	0.00	0.00	A
F	0.56	6.Techn.	57	No	No	0.06	0.07	0.35	0.52	P

Es werden hier nur hypothetische Daten verwendet!
Seules des données hypothétiques sont utilisées!

Bewertung / Interpretation

-Betrachtung der A-posteriori-Wahrscheinlichkeiten des vorhergesagten wie des kodierten Wertes

Analyse de la probabilité a-posteriori entre prédiction et valeur codifiée

-kann sicher verbessert werden

potentiel d’amélioration

Sex	FTE	Field	Age	Swiss	Uni	Observed	Predicted	p(obs|·)	p(pred|·)
F	1.000	5. Medicine	18	TRUE	TRUE	A	D	0.002	0.996
M	1.000	5. Medicine	20	TRUE	TRUE	A	D	0.002	0.996
F	1.000	Other	18	TRUE	TRUE	A	D	0.007	0.989
F	1.000	Other	19	TRUE	TRUE	A	D	0.008	0.988
M	1.000	5. Medicine	21	TRUE	TRUE	A	D	0.009	0.988
F	0.200	8. Central	34	TRUE	FALSE	A	D	0.009	0.987
F	1.000	8. Central	22	FALSE	TRUE	A	D	0.01	0.987
F	0.900	8. Central	61	TRUE	TRUE	P	D	0.007	0.981
M	0.600	6. Technical	26	FALSE	FALSE	D	A	0.011	0.985
F	0.400	8. Central	31	FALSE	FALSE	A	D	0.011	0.984
F	1.000	5. Medicine	30	FALSE	TRUE	A	D	0.012	0.982
F	1.000	2. Economy	56	FALSE	TRUE	A	P	0.005	0.974
M	0.058	2. Economy	72	TRUE	TRUE	A	U	0.004	0.972
F	0.600	4. Exact	25	FALSE	FALSE	D	A	0.014	0.982
F	1.000	2. Economy	55	FALSE	TRUE	A	P	0.005	0.971
F	0.028	Other	55	TRUE	TRUE	D	U	0.008	0.973
M	0.700	4. Exact	25	FALSE	FALSE	U	A	0.002	0.967
F	0.021	Other	56	TRUE	TRUE	A	U	0.004	0.968
M	1.000	2. Economy	49	FALSE	TRUE	A	P	0.006	0.970
M	0.800	8. Central	36	FALSE	FALSE	A	D	0.016	0.980

Es werden hier nur hypothetische Daten verwendet!

Seules des données hypothétiques sont utilisées!

«Fehler» / «faute»?

Heisst nicht, dass Zielvariable falsch ist, sondern dass generell etwas womöglich nicht stimmt.

Ne veut pas dire que la variable dépendante est fausse, mais qu'il y a généralement quelque chose qui n'est pas en ordre

Oder es handelt sich um Sonderfälle (25jähriger Professor?)

Ou il s'agit de cas exceptionnels (Un professeur a 25 ans?)

Benötigt / Requis

- ✓ Eine hohe Anzahl von Fällen / Nombre élevé d'observations
- ✓ Zusammenhänge zwischen Variablen / Relations entre variables
- ✓ 'Sinnvolle' Zusammenhänge / Relations 'avec du sens'
- ✓ Entscheidungen und Anpassungen / Décisions et adaptations

Teil III: Feedback-Mechanismus / Partie III: Mécanisme de Feedback

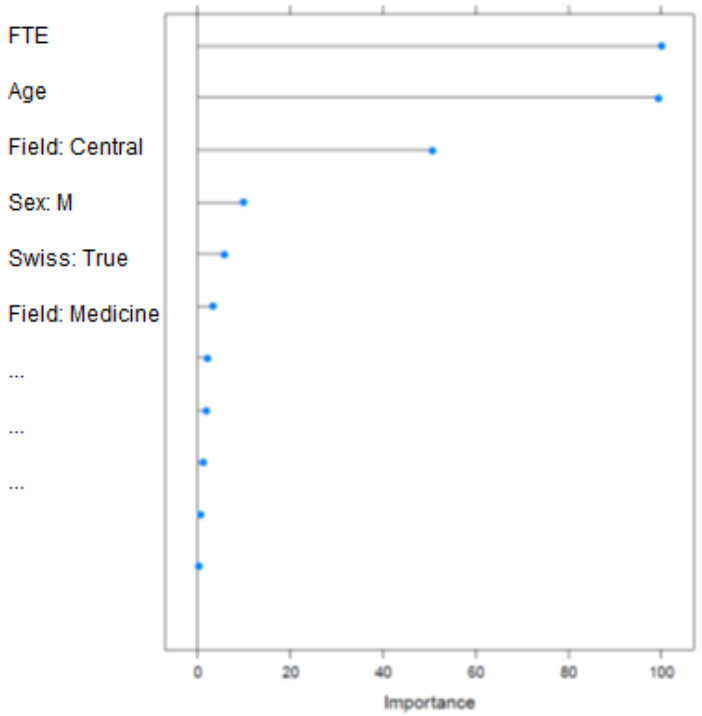
Notwendigkeit von Erklärung und Interpretabilität / Nécessité d'explication et d'interprétabilité

Datenlieferanten sind zentral / Livreurs de données sont centraux

More 'true positives' and less 'false positives'
-> Höhere Datenqualität und weniger Aufwand / Qualité de données plus élevée et moins de charge administrative

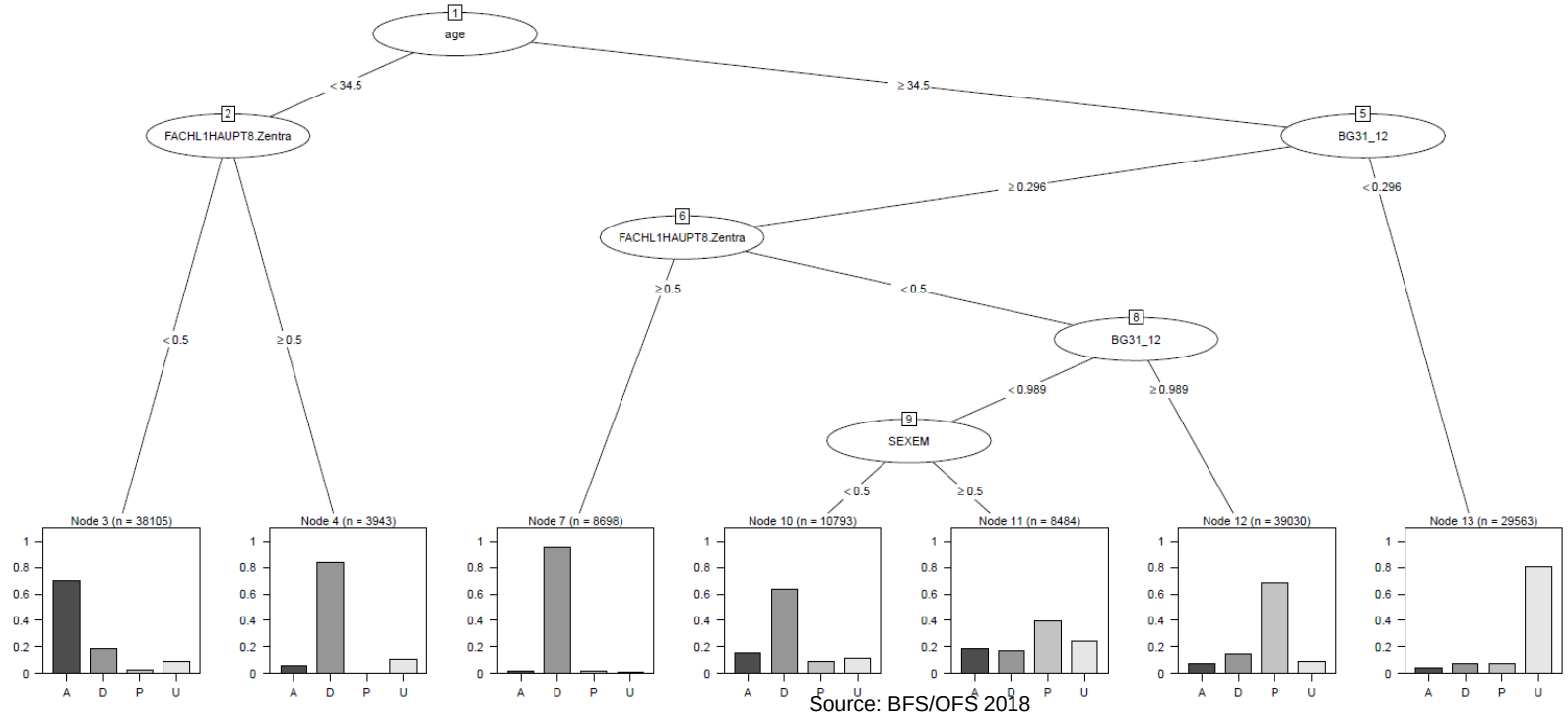
Globale Erklärungen / Explications globales

«Variable importance»



Source: BFS/OFS 2018

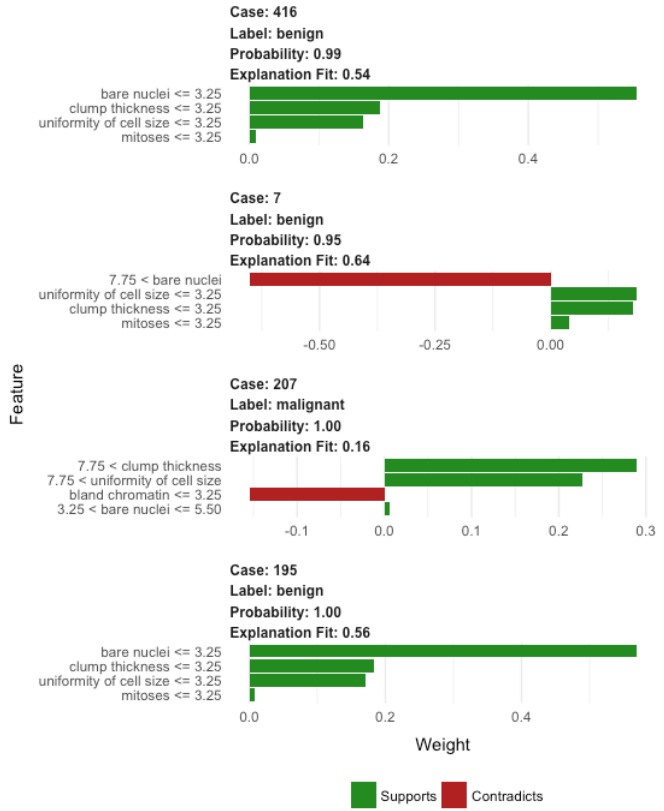
Baumalgorithmen / Algorithmes d'arbres



Source: BFS/OFS 2018

Lokale Erklärungen / Explications locales

Packages «Lime» & «DALEX»



Source: cran.r-project.org, Krebsdaten / données sur des cancers

Distanzmatrix / Matrice de distance

	Case1	Case2	Case3	Case4
Case1	0			
Case2	0.1	0		
Case3	0.7	0.2	0	
Case4	0.9	0.6	0.4	0

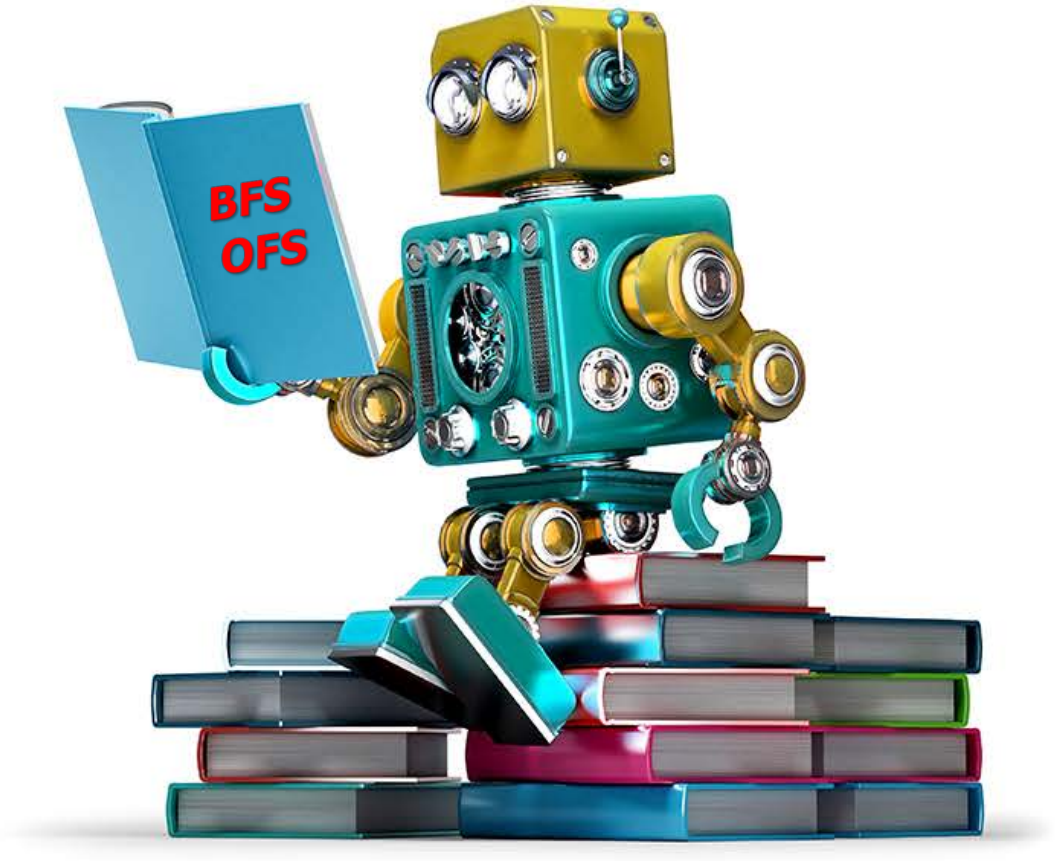
Fazit / Bilan

- Erste positive Tests / Premiers tests positifs
- Pilotprojekt bis 2019 / Projet pilote jusqu'en 2019
- Schwierige Herausforderungen vor uns / Défis importants devant nous

- Working Paper erhältlich / Working paper sur demande
- Feedback willkommen / Feedback bienvenu

Vielen Dank für Ihre Aufmerksamkeit!

Merci beaucoup pour votre attention!



Source: CC0 Public Domain with modifications