The Promises and Pitfalls of using Machine Learning in Mental Health

David Gosar, PsyD, PhD

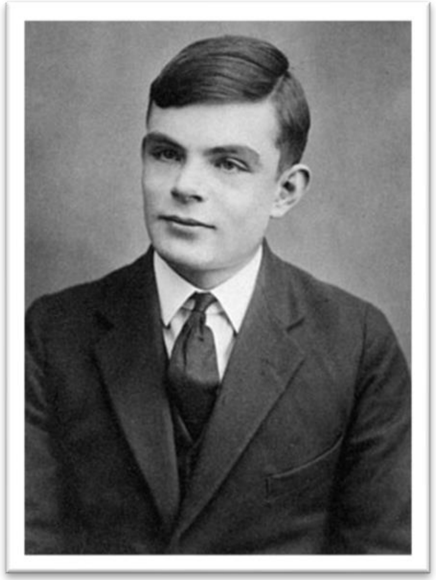
Contents

- **Artificial Intelligence, Machine Learning & Deep Learning**
  - How did we get here
  - Overview of supervised and unsupervised learning
  - Basics of "shallow" Machine Learning Methods
  - Fundamentals of Deep Learning
- **Application in Clinical Psychology & other Domains of Psychology**
  - Prediction of risk and outcomes
  - Machine learning as tool for understanding complexity
  - The potential of ML generated data & Large Language Models (LLMs)
- **Critical Thinking about ML & Ethical Issues**
  - Cesare Lombroso revisited & criterion validity as king
  - An appeal for construct validation & XAI

Artificial Intelligence, Machine Learning & Deep Learning

How did we get here

- Emergence of AI during in 1950s
  - "…the effort to automate intellectual tasks normally performed by humans"[1]
  - mathematician Alan Turing (1912-1954) – development of a general purpose computer applied to cryptography problem
  - development of expert systems
    - useful for development of well-defined, logical problems
    - unable to solve complex, "fuzzy" problems such as:
      - ☐ image classification
      - ☐ speech recognition
      - ☐ language translation
      - ☐ face recognition
      - ☐ …



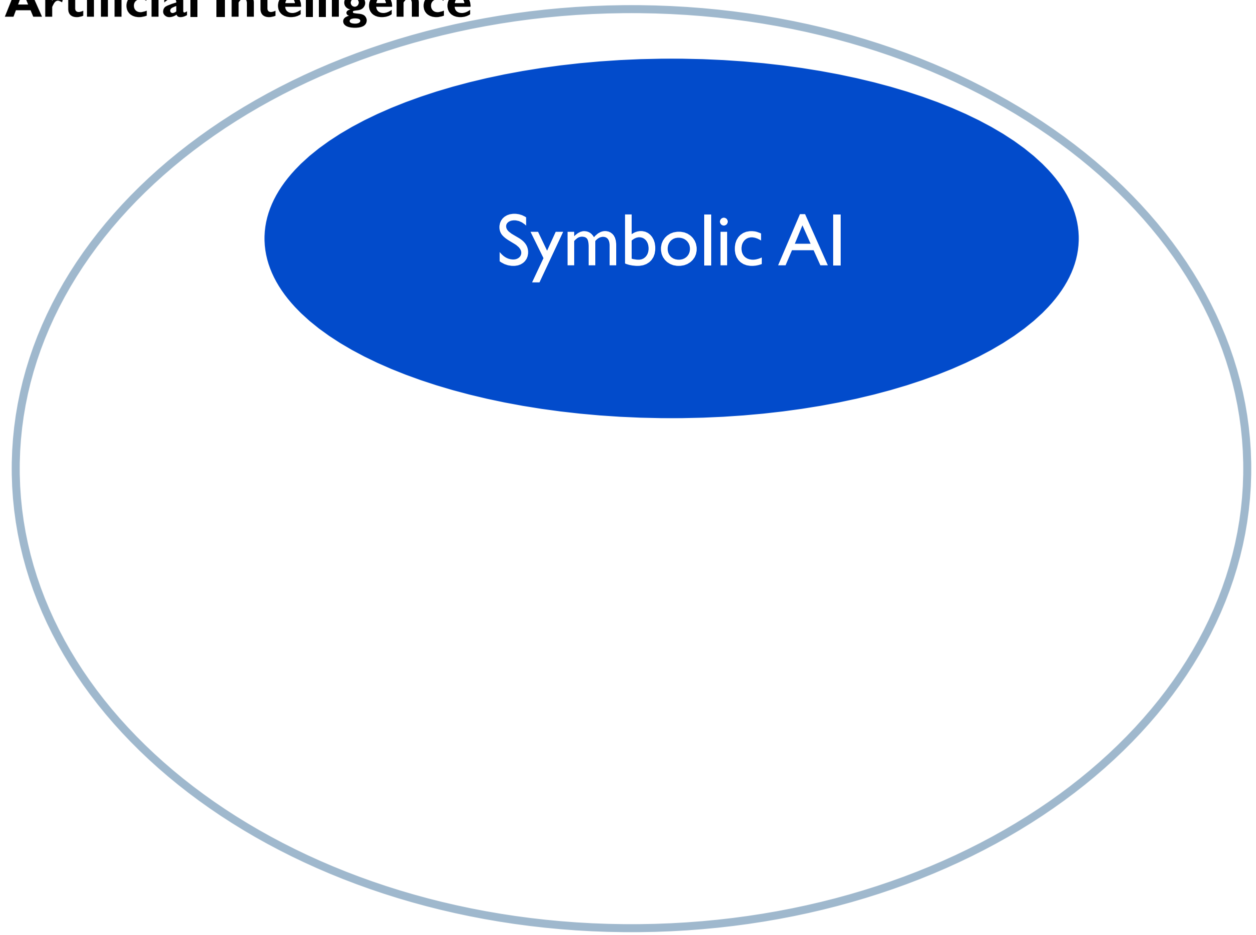
Alan Turing

ENIGMA machine

[1] Chollet & Allaire (2018)

How did we get here

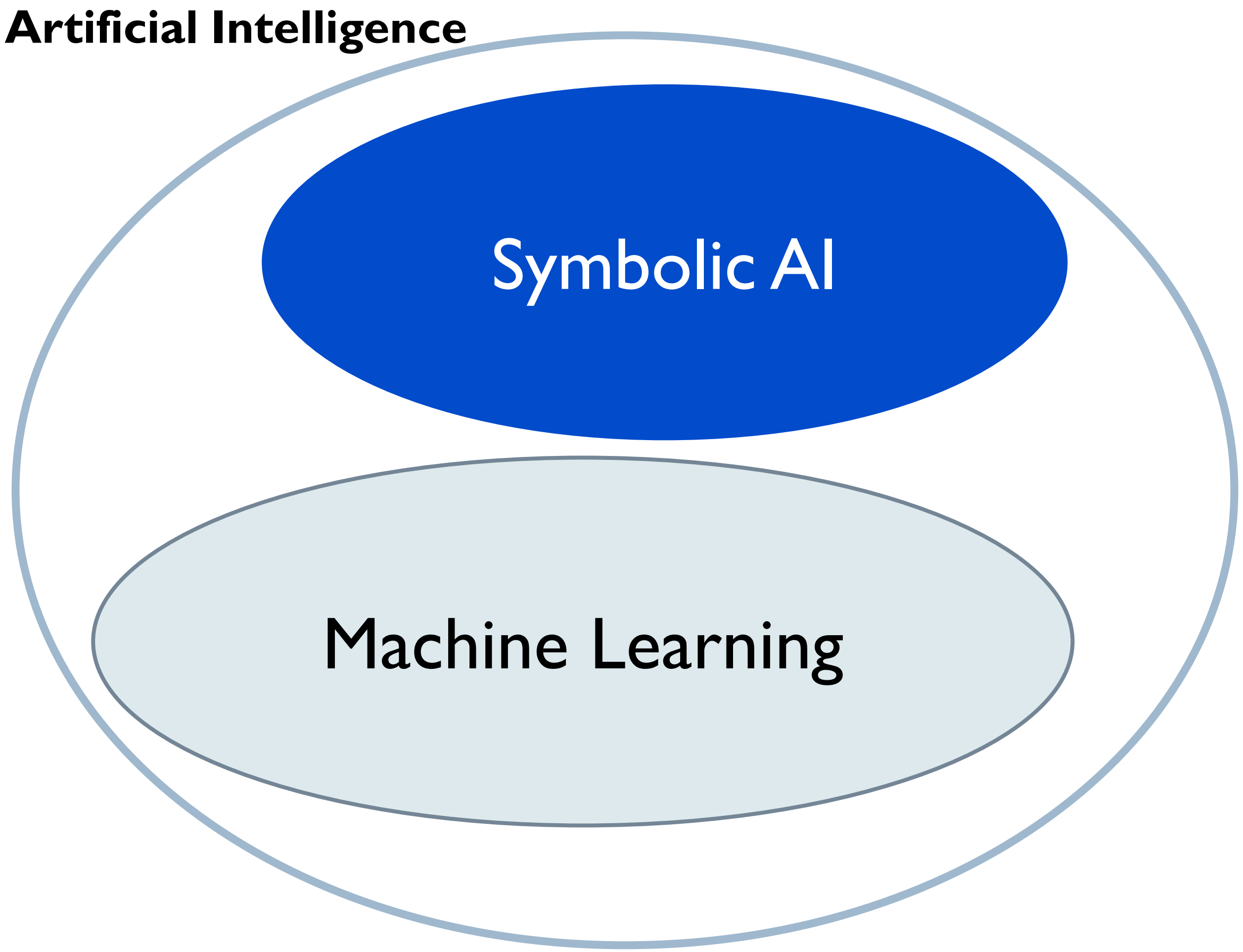
Artificial Intelligence



- History
  - 1st Artificial intelligence boom – "expert systems" & symbolic AI
    - fueled by computer tech of 1980s
    - decision rules & specific domain knowledge hard-coded
    - domain specific AI (e.g. chess,)

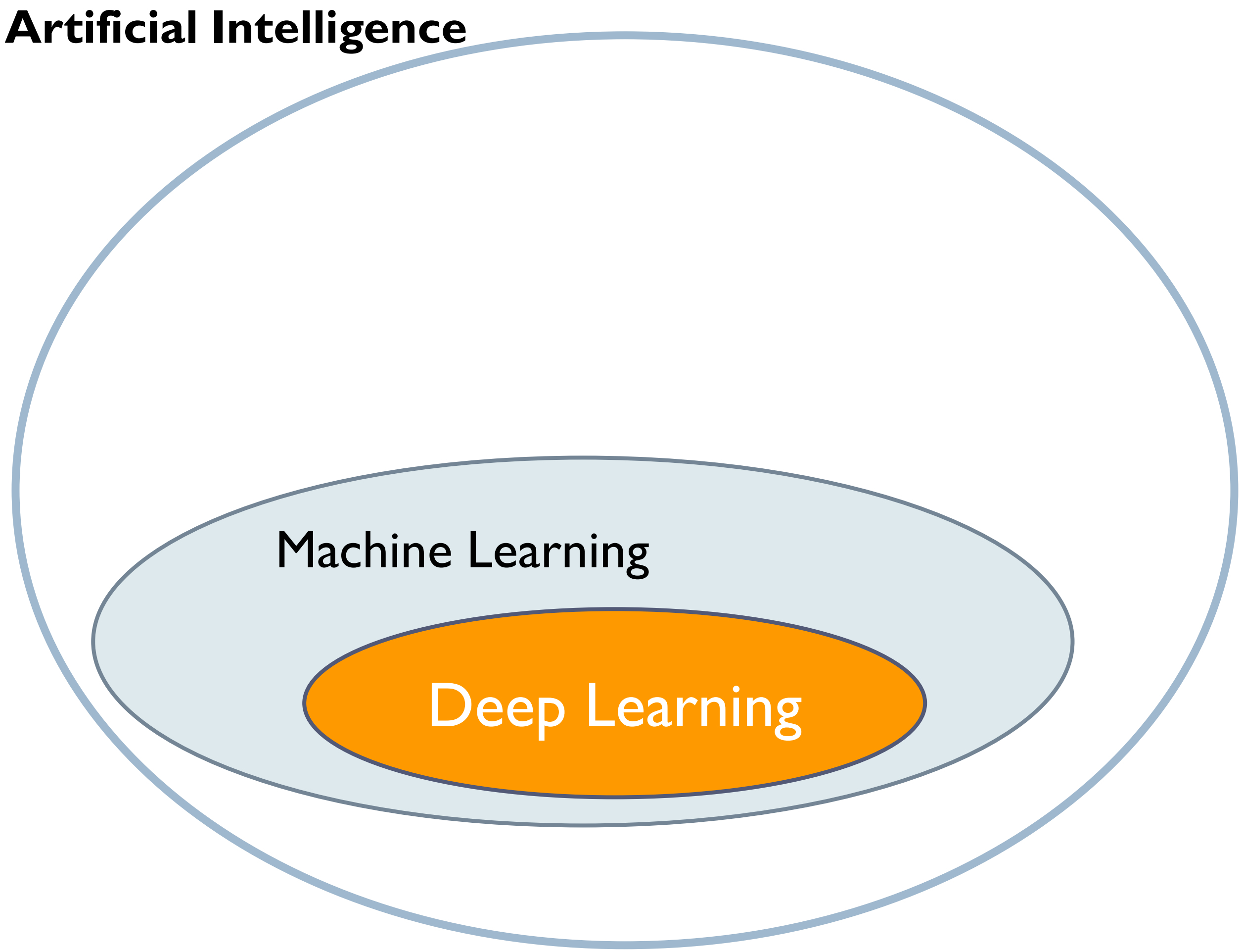
Gary Kasparov vs. IBM Deep Blue

How did we get here

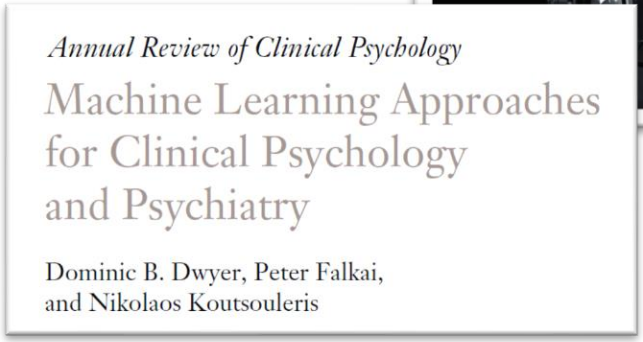


- History
  - 2st Artificial intelligence boom – learning machines
    - from "rules + data = results" to "data + results = rules"
    - algorithms not hard-coded, but learned from data, based on it's statistical structure → meaningful transformations
    - internet and the rise of Big Data

How did we get here



- History
  - 3st Artificial intelligence boom – "deep learning" & neural net stacks
    - easy to use ML software (R+, python)
    - explosion of data availability – Big Data
    - gaming industry (affordable GPU's & CUDA)
    - increasing impact in field of mental health


Annual Review of Clinical Psychology

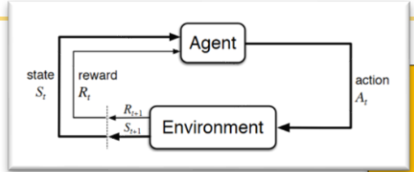
Machine Learning Approaches for Clinical Psychology and Psychiatry

Dominic B. Dwyer, Peter Falkai, and Nikolaos Koutsouleris

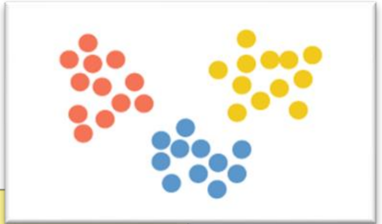
Overview of supervised & unsupervised machine learning

Machine Learning Methods

Supervised Learning



Reinforcement Learning



Unsupervised Learning

- labeling cases to facilitate learning specific rules that can be later applied to unlabeled cases
- image classification, prediction depression based on risk[1]…

- labeled cases need not be presented, and sub-optimal actions not corrected
- exploration (of uncharted behavior) via mechanism of action generation and feedback on success in discrete time-steps
- traffic light control, playing poker, Artificial Intelligence Clinician[2]...

- learning to automatically detect subgroups of individuals based on similar profiles of data
- clustering based on cognitive performance, genes, neuroimaging, predicting disease based on patient records – Deep Patient[3]

[1] Victor et al. (2019), Komorowski et al. (2018), Miotto et al. (2016)

“Shallow” Machine Learning Methods

Basics of Supervised Machine Learning

- in supervised learning **CLASSIFICATION** or **REGRESSION**
- metrics for parameter tuning is **Estimation Accuracy**
  - for classification measures of based on confusion matrix:
    - ***Accuracy*** (ACC)
    - ***Recall*** or True Positive Rate (TPR)
    - ***Precision*** or Positive Predictive Value (PPV)
    - ***F1*** or F-measure – combination of Recall & Precision
    - ***area under the ROC curve*** (AUC)
  - for prediction in context of regression:
    - ***Mean Absolute Error*** (MAE) - average of the absolute differences between predictions and actual values
    - ***Mean Squared Error*** (MSE) – similar to MAE
    - ***R^2 Metric*** - coefficient of determination

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

$$ACC = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

$$Recall = TPR = \frac{TP}{(TP + FN)}$$

$$PPV = \frac{TP}{(TP + FP)}$$

$$F - measure = \frac{2TP}{2TP + FP + FN}$$

"Shallow" Machine Learning Methods

Basics of Supervised Machine Learning

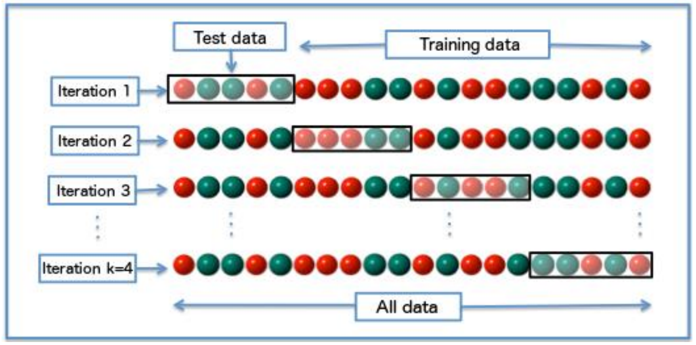
- common **issue with classifying in clinical settings** – the group we are trying to predict is significantly smaller than the comparison group (i.e. people with a mental disorder, risk factor…etc.)
- **misleadingly high accuracy** – prevalence of 5%, accuracy 95% (based on baserate)

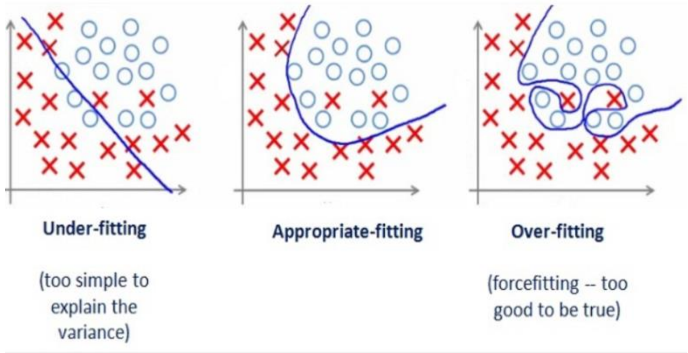
- optimizing models with such **unbalanced data**
  - estimating accuracy using **Balanced Accuracy** (in terms of true positive and negative cases balanced by the sample size of each positive and negative group), sometimes also **AUC**, **F1** and **Cohen Kappa**
  - different methods to offset different group sizes:
    - ☐ class re-weighting with optimization algorithm (adapt cost function)
    - ☐ under-sampling larger group (when lots of data)
    - ☐ over-sampling of smaller group
    - ☐ Synthetic Minority Over-sampling Technique (SMOTE)[1]

[1] Chawla, Bowyer, Hall & Kegelmeyer (2002)

"Shallow" Machine Learning Methods

- **Basics of Supervised Machine Learning**
  - **Two-sets of parameters**
    - **parameters** (weights, support vectors, decision rules…) of the model
    - **hyper-parameters** (model setup, architecture)
  - Data requirements: **training, testing & validation dataset**
    - Training dataset – **tune parameters**
    - Testing dataset – **evaluate hyper-parameters** (under- or over-fitting)
    - Validation dataset – **evaluate on independent data**
    - V-fold Cross-validation often used to create training and testing datasets
      - usually 5- to 10-fold cross validation recommended
      - bootstrapping may also be used
  - **Data Explosion in Age of Big Data**
    - from long-form (more cases than variables) to wide-form (more variables than cases)
    - issues of variable selection
    - importance of parallel computing (taking advantage of Moore's Law)



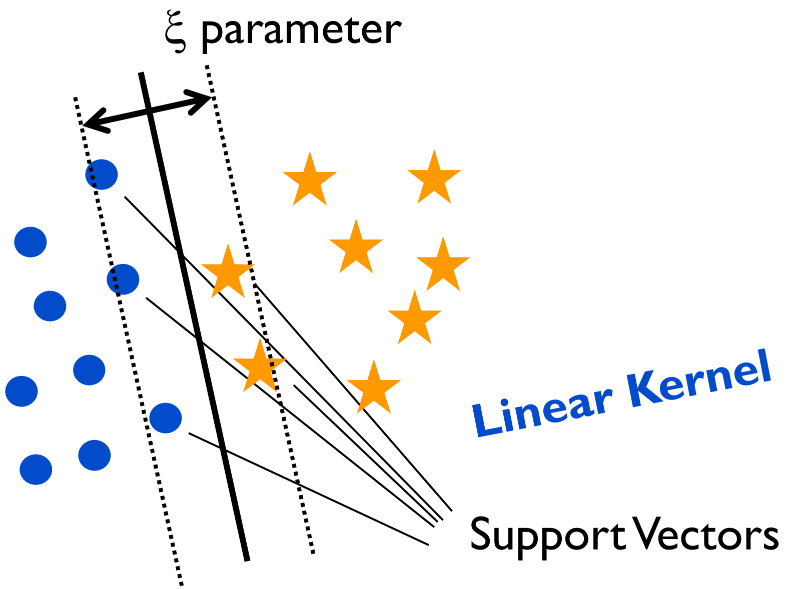


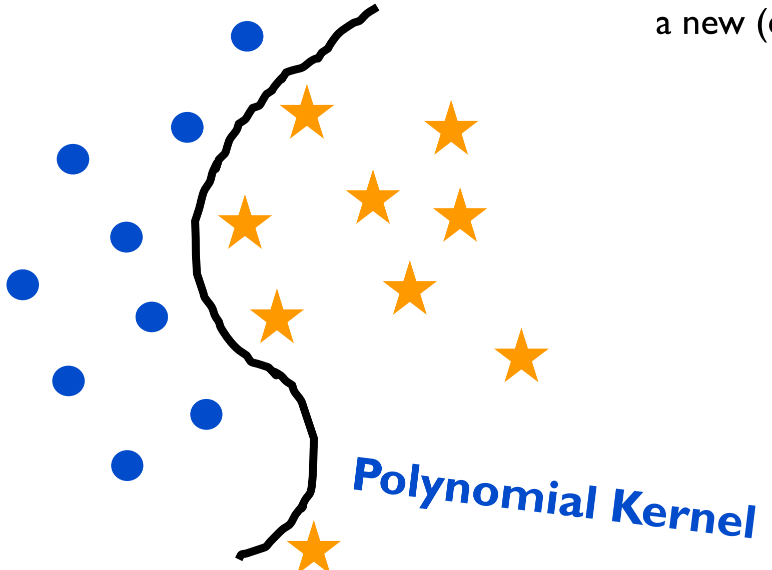
[1] Gatys, Ecker & Bethge (2015)

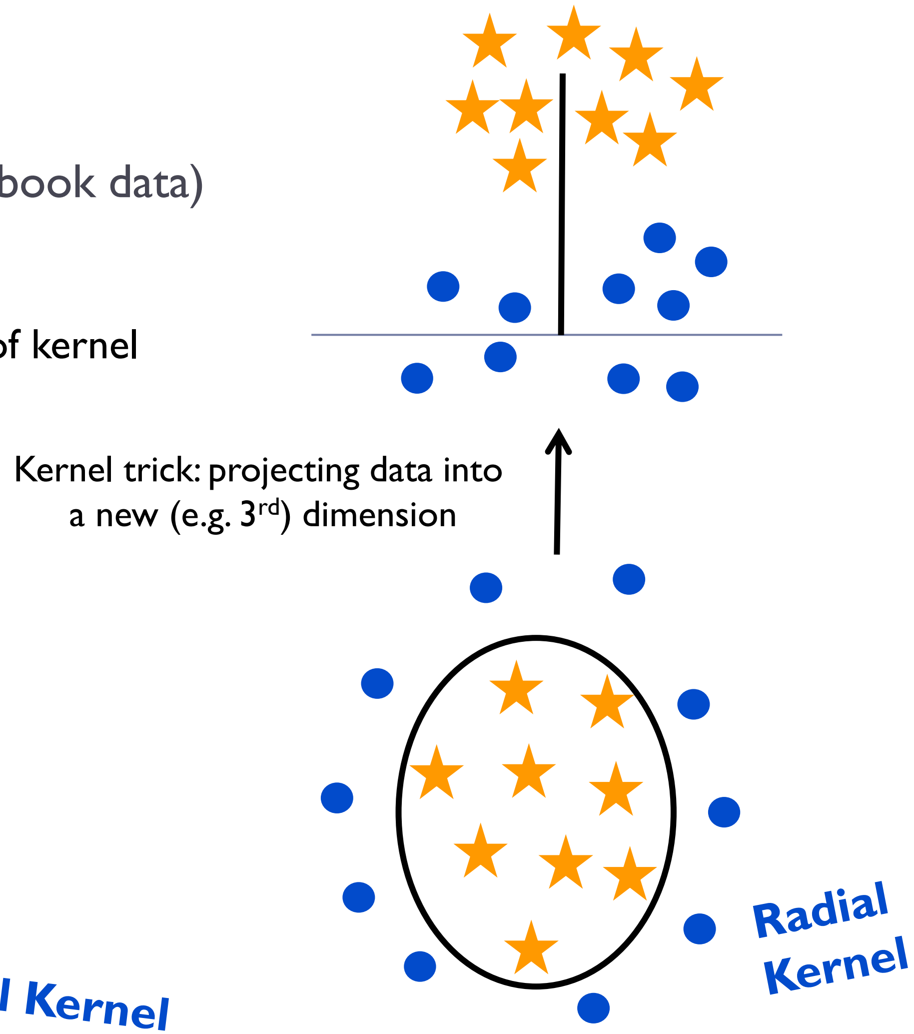
“Shallow” Machine Learning Methods

Support Vector Machines

- findings cases that best separate groups (model parameters)
- projection in (high)dimensional feature space (e.g. Big 5, Facebook data)
- model hyper-parameters
    - distance from best hyper-plane separating groups (ξ parameter) & type of kernel

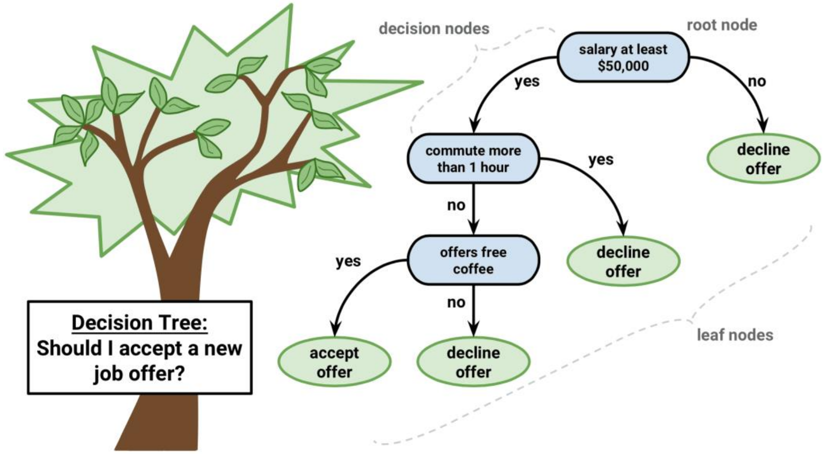






“Shallow” Machine Learning Methods

- **Random Forests**
  - based on decision trees
    - consecutively looking at best splits of groups after splitting based on predictor
  - predictions are combined from many decision trees by using voting → “random forest”
  - algorithm takes advantage of a concept in ML called **Ensamble Learning**
    - the combined predicative power of weak learners is more robust and stronger than each individual algorithm
    - other similar concepts such as boosting (improving on weakness of models trained in previous step)
  - among the more competitive “shallow” ML methods
  - variable importance metrics[1,2]
    - importance for prediction & degree of interaction with var’s



[1]Louppe, Wehenkel, Sutera & Geurts (2013), [2]Ribeiro, Singh & Guestrin (2016)

"Shallow" Machine Learning Methods

- **Random Forests – Example**
  - predicting Employee Turnover based on multiple predictors in Company
  - bootstrapping used for parameter estimation
  - results given in terms of variable importance

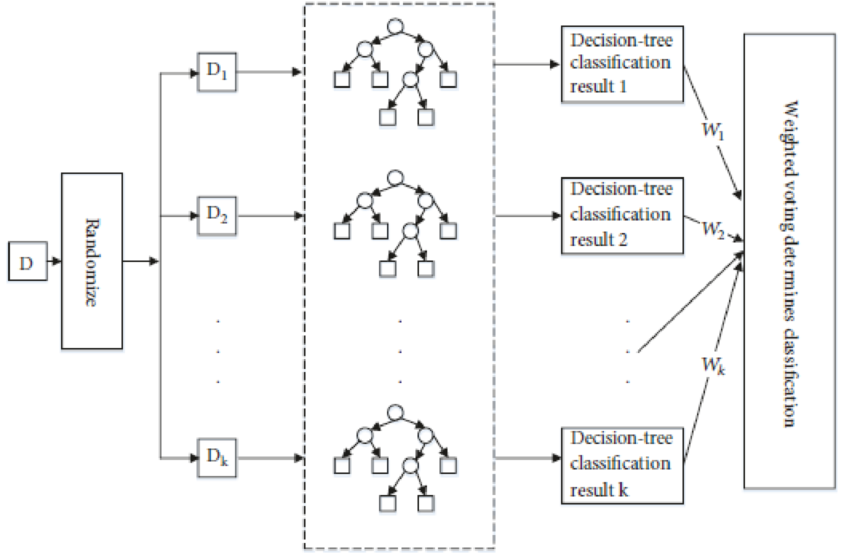
An Improved Random Forest Algorithm for Predicting Employee Turnover

Xiang Gao,[1] Junhao Wen,[2] and Cheng Zhang[1]

[1]College of Computer Science, Chongqing University, Chongqing 400044, China
[2]College of Big Data & Software Engineering, Chongqing University, Chongqing 400044, China

TABLE 3: Importance score of 30 features.

No.	Feature	Score	No.	Feature	Score
1	MonthlyIncome	0.2816	16	EducationField	0.0137
2	OverTime	0.2762	17	WinningCount	0.0100
3	Age	0.0665	18	Gender	0.0093
4	DistanceFromHome	0.0431	19	NumberCompaniesWorked	0.0073
5	YearsatCompany	0.0317	20	HaveChildren	0.0069
6	PercentSalaryIncrease	0.0306	21	EnvironmentSatisfaction	0.0065
7	YearsinCurrentRole	0.0289	22	RelationshipSatisfaction	0.0062
8	TrainingTimesLastYear	0.0257	23	JobSatisfaction	0.0060
9	YearsSinceLastPromotion	0.0250	24	EmploymentNature	0.0056
10	YearswithCurrentManager	0.0218	25	MaritalStatus	0.0036
11	AvgWorkHours	0.0211	26	PerformanceRatingLastYear	0.0031
12	TotalWorkingYears	0.0208	27	WorkLifeBalance	0.0015
13	JobLevel	0.0173	28	PhysicalCondition	0.0007
14	Education	0.0141	29	JobRole	0.0007
15	DepartmentType	0.0140	30	NativePlace	0.0004



"Shallow" Machine Learning Methods

- **Random Forests - Example**
  - **good practice**: validation dataset + baseline
  - be wary of ML studies reporting high accuracy without a separate validation dataset !!!
  - "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View" [1]

An Improved Random Forest Algorithm for Predicting Employee Turnover

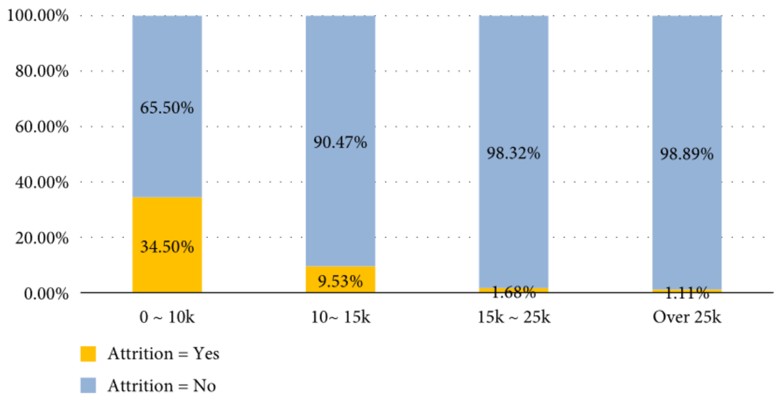
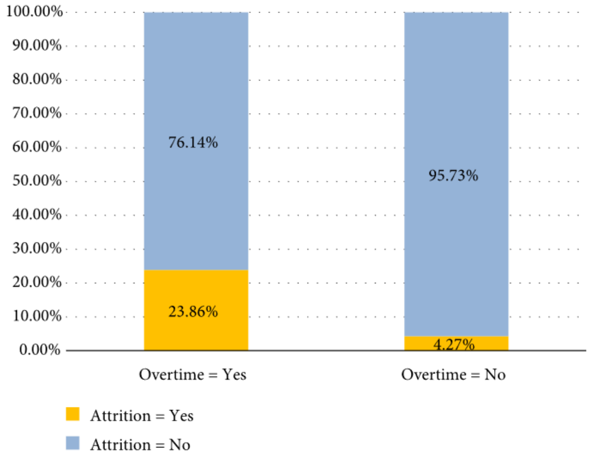


TABLE 4: Experimental results of different algorithms.

Algorithms	Recall	F-measure	ROC Area	ACC (%)
RF	0.627	0.698	0.850	92.65
C4.5	0.561	0.564	0.773	91.05
Logistic	0.469	0.259	0.807	90.20
BP	0.502	0.560	0.781	89.30
WQRF	0.653	0.711	0.881	92.80



[1]Lou et al. (2016)

"Shallow" Machine Learning Methods

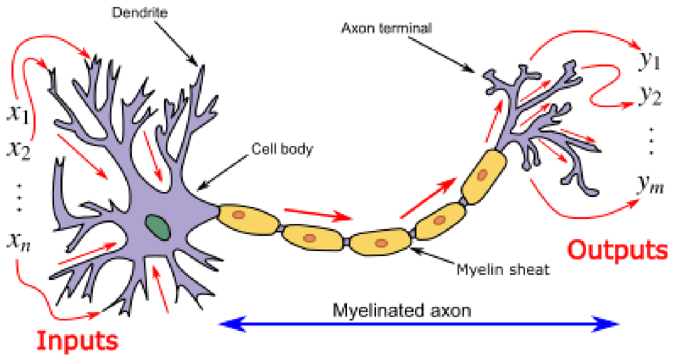
- **Other Methods**:
  - Principal component analysis (PCA) (Pearson, 1901)
  - Linear discriminant analysis (LDA) (Fisher, 1936)
  - Isometric feature mapping (Isomap) (Tenenbaum, Silva, & Langford, 2000)
  - Extreme Gradient Boosting (Chen & Guestrin, 2016)
  - Single Layer Neural Networks – e.g. perceptron (Rosenblatt, 1957)

- **Revolution with the development of Deep Learning**
  - Geoffrey Hinton and two students attended the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC2012) - won the first place with more than 10% improvement of top-5 test error rate over the second best entry[1]
  - Development of GPU's and the *Compute Unified Device Architecture* (CUDA) in 2010s allowed training of multi-layer neural networks

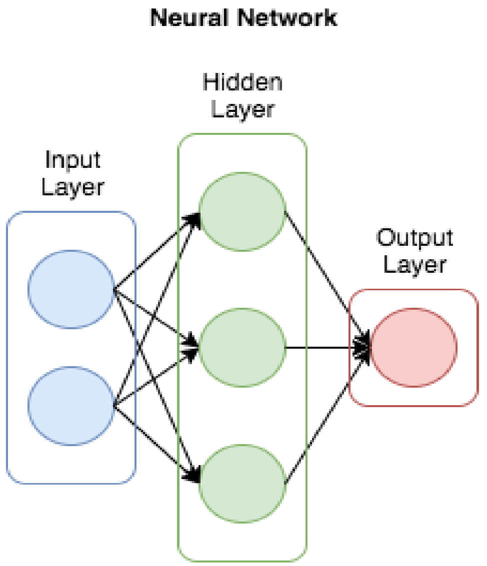
[1]Krizhevsky et al. (2012)

Fundamentals of Deep Learning

- **Neural Networks** – neurons as inspiration
  - neurons perform computations by aggregating information from incoming connections to dendrites and then passing a signal along the axon to the next neuron or muscle junction
  - the connections that are more often used (and/or prove beneficial) are strengthened, others disregarded
    - "units that fire together, wire together" *Donald Hebb*
  - combining neurons together in a network & using feedback loops allows for complex computations

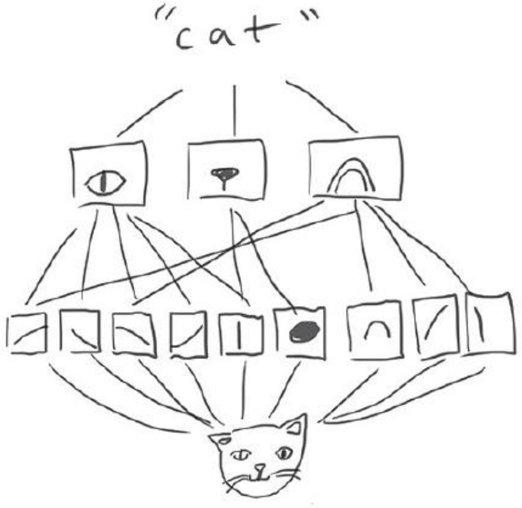


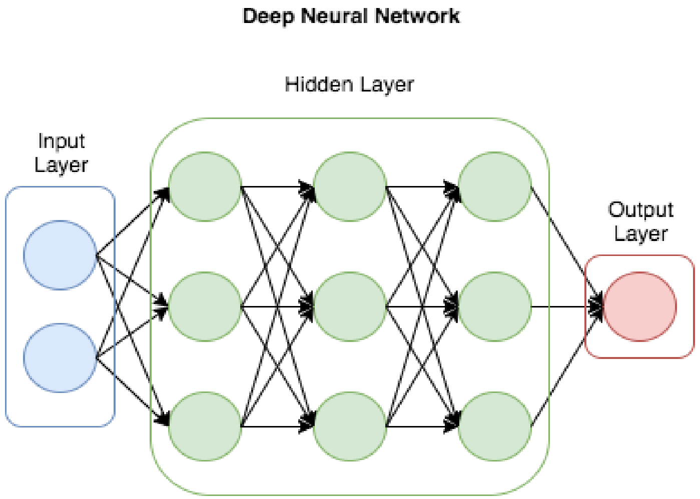
- **Artificial Neural Networks**
  - Input layer → hidden layer → output
  - mathematical models that through a repeated process of iterations compute the optimal connections between the input, hidden & output layers to get the predicted output



Fundamentals of Deep Learning

- **Deep Learning –** stacking multiple hidden layers
  - combining multiple hidden layers together turns out to improve prediction accuracy and enables abstraction of information – just like an information distillery…but why
  - each layer level learns to transform its input data into a slightly more abstract and composite representation
    - e.g. image classification:
      - first layer may abstract the pixels and encode edges
      - second layer may compose and encode arrangements of edges
      - third layer may encode a nose and eyes
      - fourth layer may recognize that the image contains a face…
    - is able to learn which features to optimally place in which level on its own
  - successful application in computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, medical image analysis, material inspection and board game programs





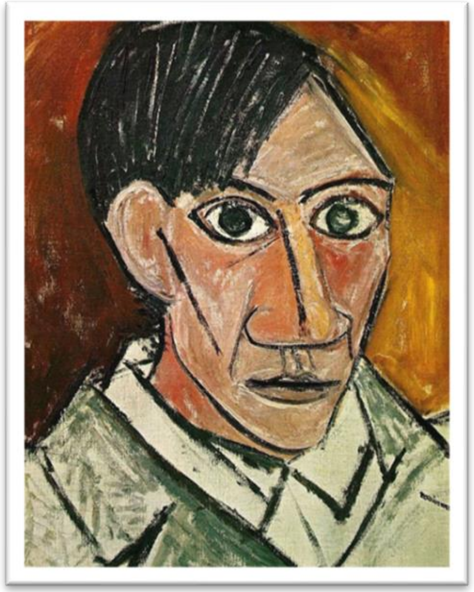
[1] Chollet & Allaire (2018)

Fundamentals of Deep Learning

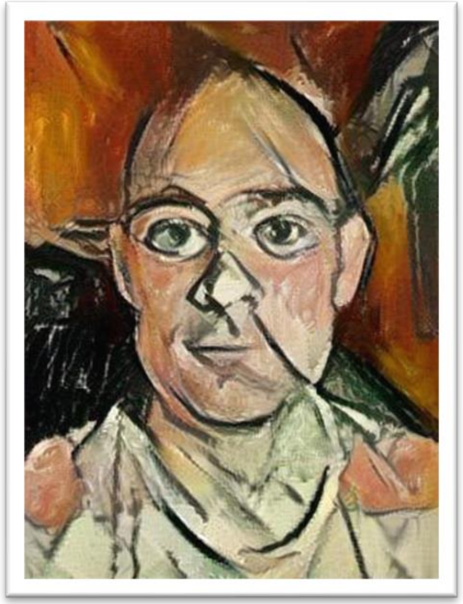
- Convolutional Neural Networks (CNNs)[1]
  - CNNs are able to detect spatially invariant features with progressive abstraction
  - the first to achieve human-competitive performance on certain practical applications
    - image classification of small & large objects in cluttered scenes
    - able to generated data that reflects spatial characteristics at different scales

David Gosar by Pablo Picasso;)

+



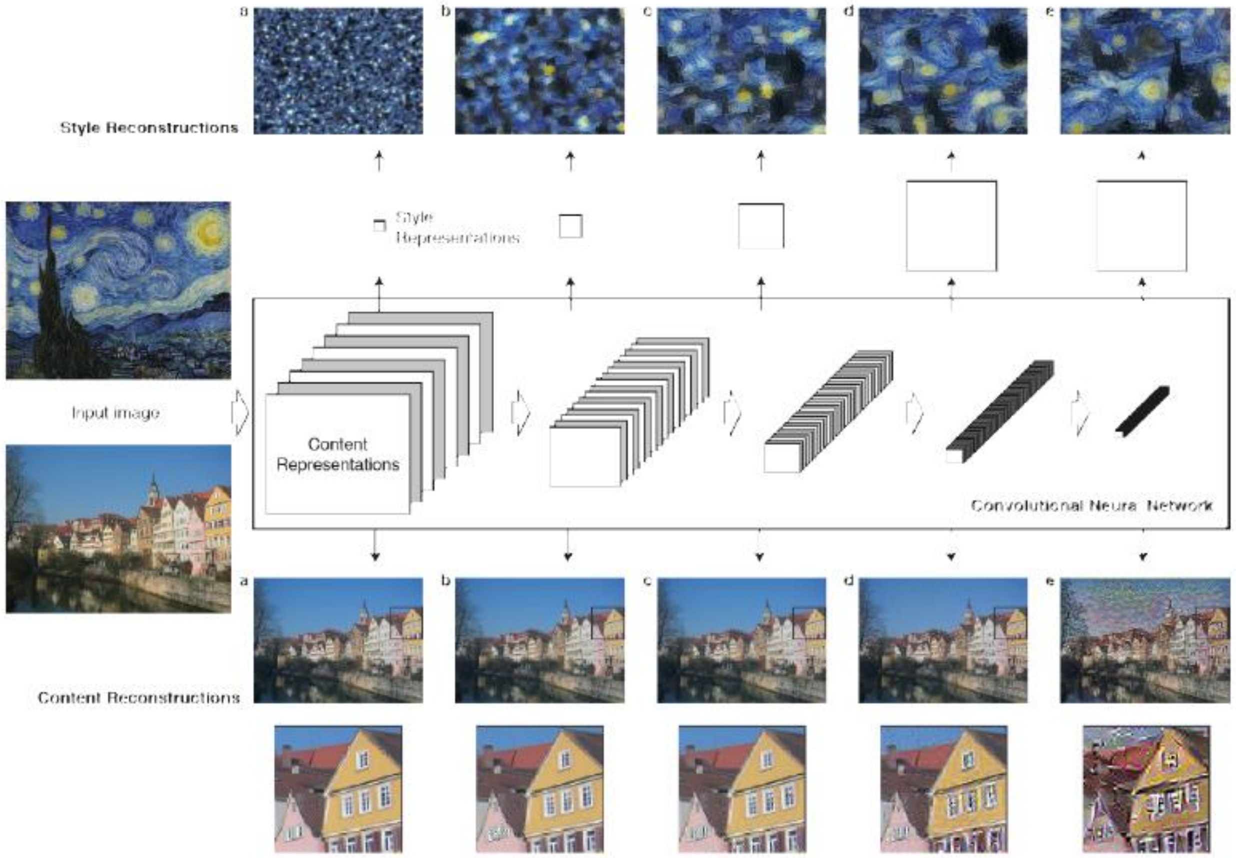
=



[1] Gatys, Ecker & Bethge (2015)

Fundamentals of Deep Learning

- **CNN's reproducing artistic style**
  - lower levels of CNN (a, b, c) reconstruct an original image almost perfectly, at higher levels (d,e) only high-level content is preserved
  - style representation are computed as correlations between the different features in different layers of the CNN that match the style of a given image on an increasing scale
  - at same time discarding information of the global arrangement of the scene and capturing their general appearance in terms of color and localized structures
  - two images can synthesized by finding an image that simultaneously matches the content representation of the photograph & style representations of the artwork



[1] Gatys, Ecker & Bethge (2015)

Fundamentals of Deep Learning

- Transferring layers to new problems – transferring knowledge
  - using layer from pre-trained network to solve smaller sample picture classification[1]
    - solving classification problems of 4000 cats & dogs with the help of a pre-trained network layer, trained on ImageNet (collection of 1.4 million labeled images of over 1000 categories)
  - transfer of layers from deep neural networks (DNN) for language translation[2]
    - e.g. layer from DNN for translating English to French transferred to translating English to Italian enables faster learning of new association
- Similar performance to humans in terms test item difficulty[3]
  - study looking at IRT indicators of question difficulty (based on responses from 1000 humans from Amazon Mechanical Turk assessing pairs of premise & hypothesis pairs – if, then)
  - as DNN trained with more data easier examples are learned more quickly than hard ones
- Great flexibility in Deep Learning Architecture
  - Combining multi-modal data (text, picture, geo-location…)

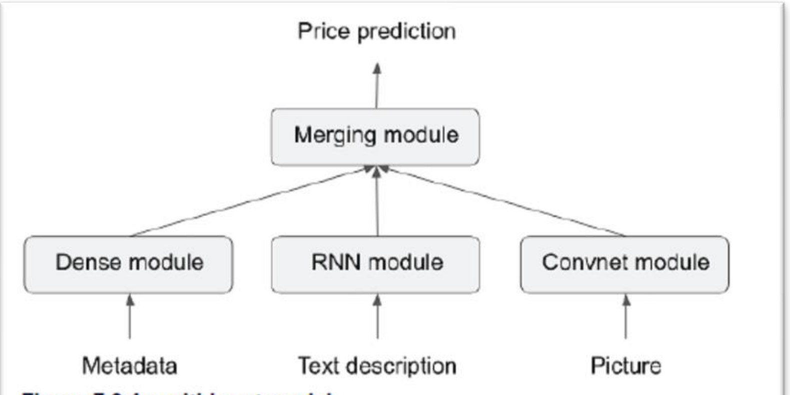
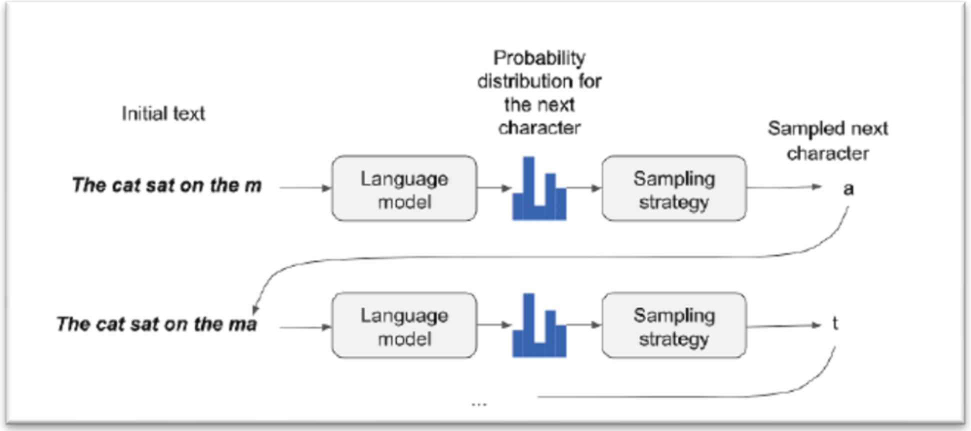


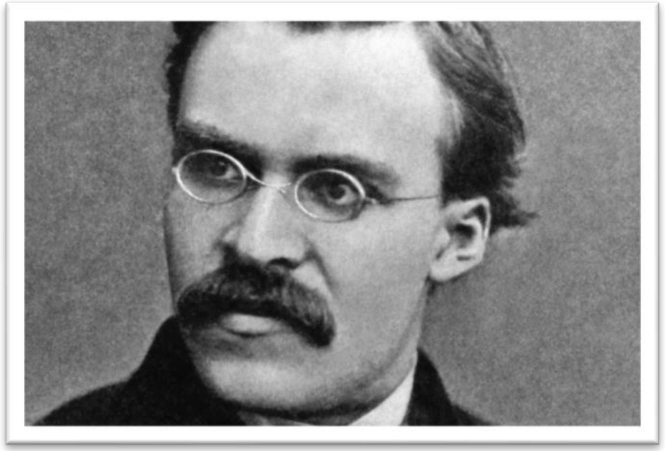
Figure 7.2 A multi-input model

[1] Chollet & Allaire (2018), [2] Qi et al. (2018), [3] Lalor, Wu, Munkhdalai & Yu (2018)

Fundamentals of Deep Learning

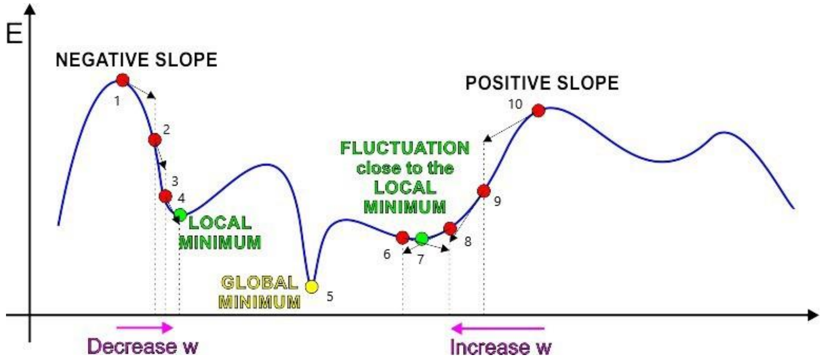
- **Generative Deep Models**
  - recurrent Neural Networks (RNNs) to map text stream into a latent language space
  - possible to sample the latent language space to obtain the prediction of the next token (e.g. word) in a next stream
  - variability of sampling has important implication for text reproduction
    (e. g. replicating text from Friedrich Nietzsche)
    - no variability (low temperature = 0.20) realistic, repetitive, realistic
      - cheerfulness, friendliness and kindness of a heart are the sense of the spirit is a man with the sense of the sense of the world of the self-end and self-concerning the subjection of the strengthorixes – the subjection of the subjection of the subjection…
    - limited variability (medium temperature = 0.50) unusual, yet still comprehensible
      - cheerfulness, friendliness and kindness of a heart are the part of the soul who have been the art of the philosophers, and which the one won't say, which is it the higher the and with religion of the frences. the life of the spirit among the most continuess…
    - great variability (high temperature = 1.00) non-existent words, less plausible
      - cheerfulness, friendliness and kindness of a heart are spiritual by the ciuture for the entalled is, he astraged, or errors to our you idstood--and it needs, to think by spars to whole the amvives of the newoatly, prefectl yraals!

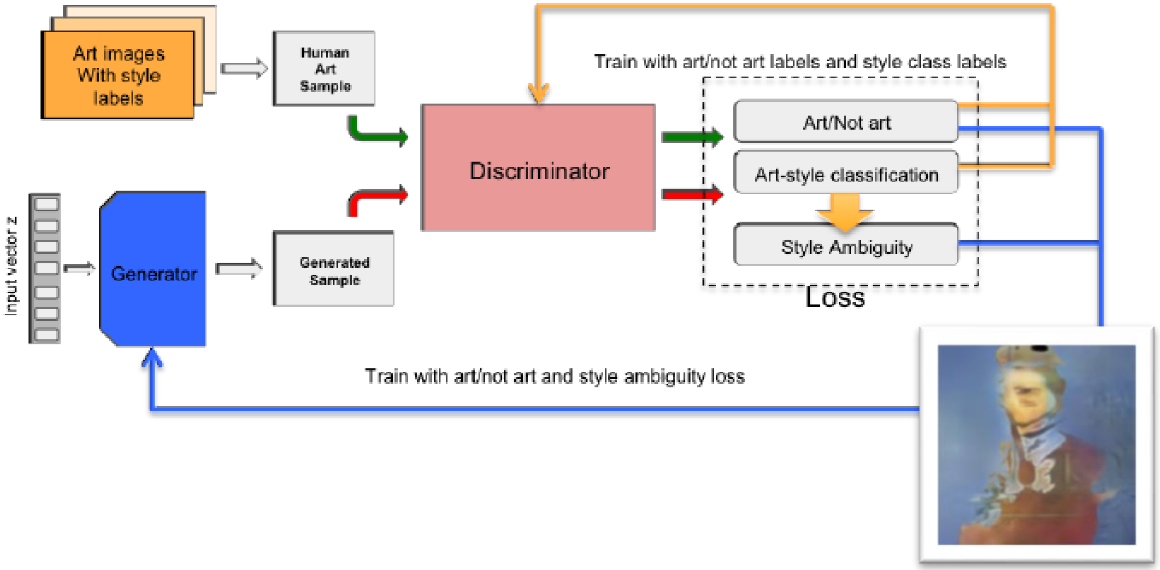




Friedrich Nietzsche

Fundamentals of Deep Learning & Other Machine Learning Methods



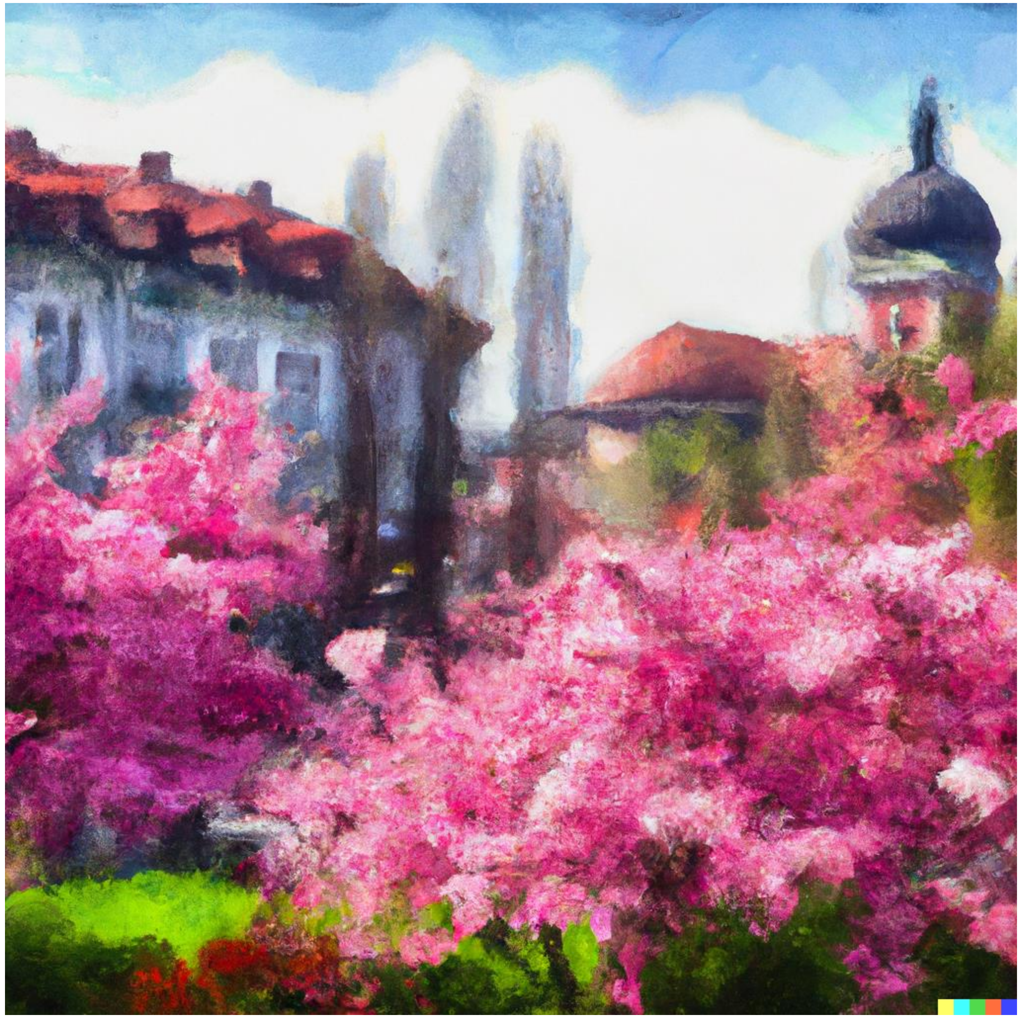


- Tuning parameters & hyper-parameters
  - parameter tuning mostly gradient descent
  - optimizing hyper-parameters currently more an art than a science[1]
  - <u>3 datasets: training, testing, validation</u> !!!
- Current trends
  - faking a Picasso – adversarial networks[2]
    - 2 networks – 1st producing, 2nd evaluating
  - Program Synthesis[3]
    - using genetic or other algorithms to develop computer code
    - nesting hyper-parameter evaluation into "for", "while" and other programming loops
  - emphasis on using unstructured data

[1]Chollet & Allaire (2018), [2]Elgammal, Liu, Elhoseiny & Mazzone (2017), [3]Kant (2018)

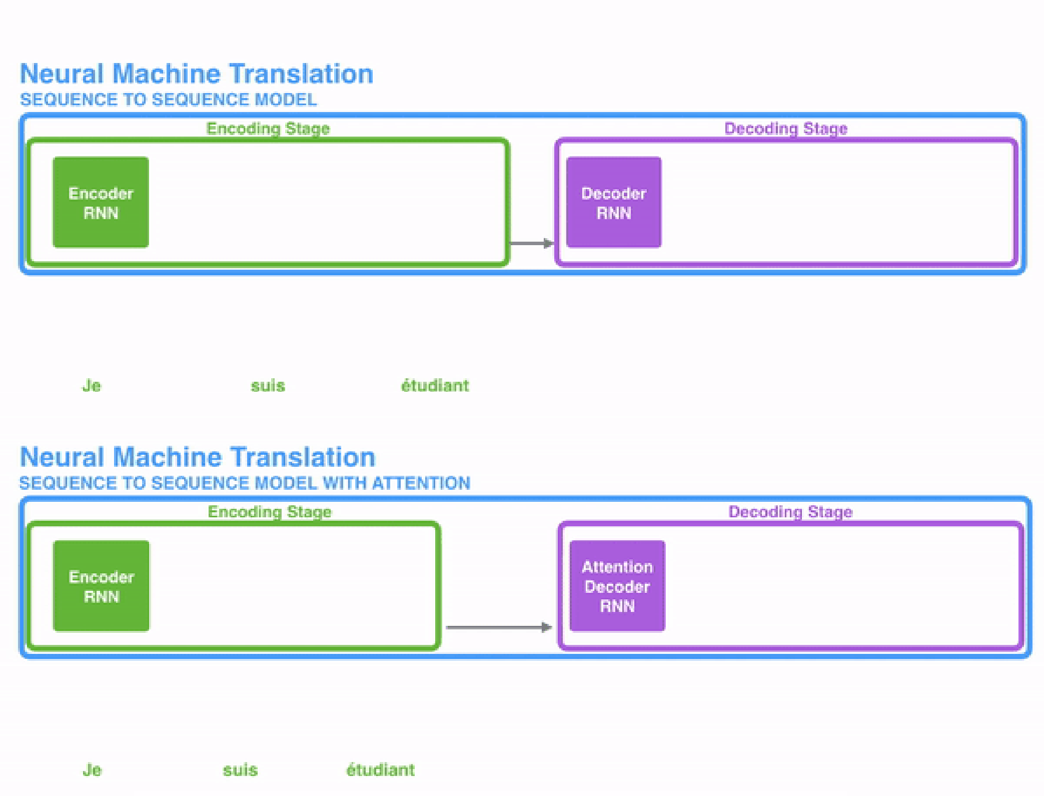
4th Artificial intelingence boom

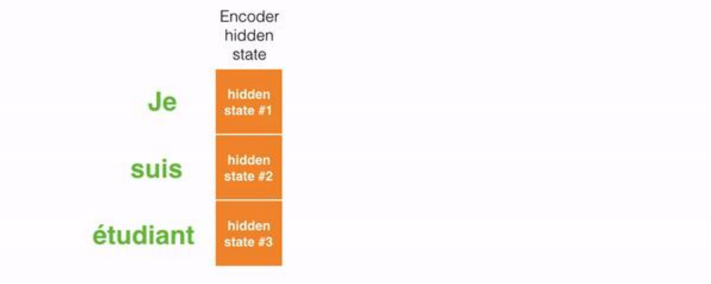
2017 - 2023



DALL-E: "*An impressionist painting of ChatGPT AI in spring with flowers in Ljubljana*"

Large Language Models (LLMs)
Basic Principles





large models – millions of parameters

- recurrent neural networks (RNN)
  - how many words forward and back to take into account?
  - how to take word order into account (e.g. "*boy chases dog*" vs. "*dog chases boy*")
  - how to account for context (e.g. "*Midnight oil*")

- "transformer" architecture
  - *"Attention is all you need"* (Vasvani et al., 2017) – dynamic weighting of all words in a text
  - weighting of connections for the prediction of words within model is dynamic and weighted differently according to current combination – *taking context into account*
  - + convolutional neural networks (CNN)

Large Language Models (LLMs)

Tools for psychological research

European Federation of Psychologists' Associations

- **types of large language models**
  - BERT (*Bidirectional Encoder Representations from Transformers*)
  - T5 (*Text-to-Text Transformer*)
  - conversational models (*Generative Pretrained Transformers*) - ala ChatGPT
- large language models are **large**
  - chatGPT3 (175 billion parameters – 1000x less than connections in brain, 500 billion besed, price of computing power 12 milion $)
  - GLaM (1.2 billion parameters)
- **additional optimization of model possible** for specific purposes – i.e. prediction of self-harming behavior

BERT-Based Transformers for Early Detection of Mental Health Illnesses

Rodrigo Martínez-Castaño[1,2(✉)], Amal Htait[2], Leif Azzopardi[2], and Yashar Moshfeghi[2]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago, Spain
rodrigo.martinez@usc.es

[2] Department of Computer and Information Sciences, University of Strathclyd Glasgow, UK
{amal.htait,leif.azzopardi,yashar.moshfeghi}@strath.ac.uk

Abstract. This paper briefly describes our research groups' efforts in tackling Task 1 (Early Detection of Signs of Self-Harm), and Task 2 (Measuring the Severity of the Signs of Depression) from the CLEF eRisk Track. Core to how we approached these problems was the use of BERT-based classifiers which were trained specifically for each task. Our results on both tasks indicate that this approach delivers high performance across a series of measures, particularly for Task 1, where our submissions obtained the best performance for precision, F1, latency-weighted F1 and ERDE at 5 and 50. This work suggests that BERT-based classifiers, when trained appropriately, can accurately infer which social media users are at risk of self-harming, with precision up to 91.3% for Task 1. Given these promising results, it will be interesting to further refine the training regime, classifier and early detection scoring mechanism, as well as apply the same approach to other related tasks (e.g., anorexia, depression, suicide).

Application in Clinical Psychology & other Domains of Psychology

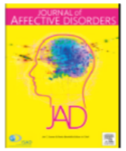
Prediction of risk and outcomes

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

ELSEVIER



Review article

The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review

Taylor A. Burke[a,1,*], Brooke A. Ammerman[b,1], Ross Jacobucci[b]

7. Directions for future research

7.1. *Broaden outcomes and indicators*

7.2. *Broadening research questions*

7.3. *Advancing ML techniques*

WAYS FORWARD

a separate sample). Despite the fact that many of the studies in this review implemented CV methods (e.g., testing algorithms on holdout samples), few studies validated their algorithms on external samples. Without access to external samples to conduct a truly independent test of a model's performance, we are unable to ascertain the full extent to which over-fitting may be occurring. A related and important limitation

LIMITATIONS

- Predicting suicide notoriously difficult
  - rare but highly significant event[1]
  - EU $11/_{100000}$, Slovenia $19/_{100000}$
- focus of (Machine Learning) ML studies
  - improve prediction accuracy
    - earlier reviews of non-ML prediction (AUC ~ 0.58)
    - better ML prediction (AUC ~ 0.71 to 0.89)
  - identify important indicators & interactions
    - identified well known risk factors (depression, earlier attempts, psychiatric hospitalization)
    - identified unusual indicators in clinical notes
  - model high-risk subgroups
    - specific high-risk groups identified using decision trees
    - e.g. female adolescents, high depression, delinquent[2]

[1]Eurostat (2014), [2]Bae et al. (2015)

Prediction of risk and outcomes

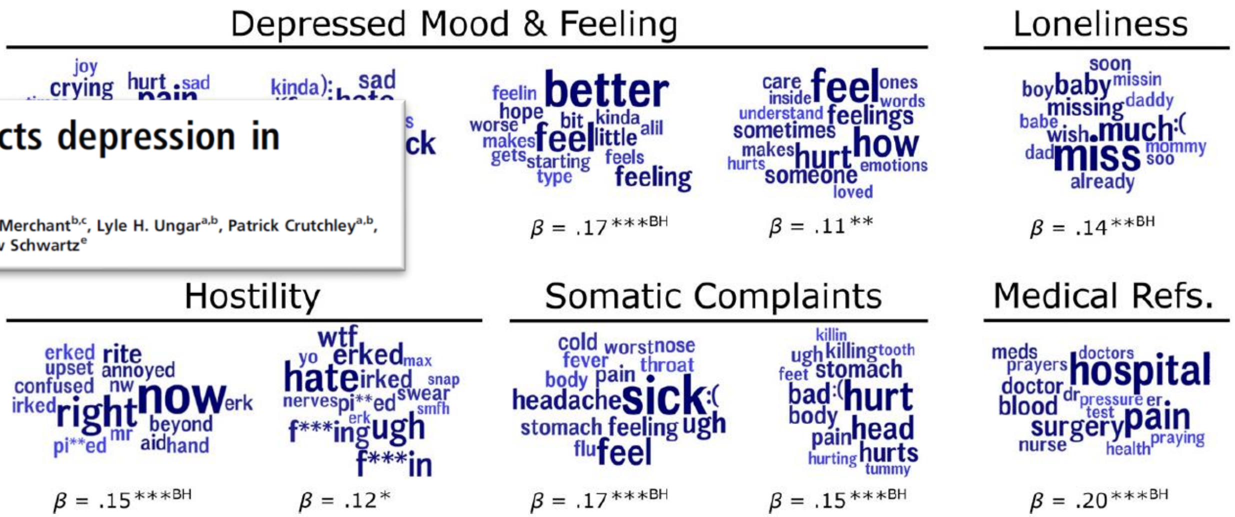
- **Tackling Depression by Mining FaceBook**
  - text mining (e. g. studying psychological states and traits using Facebook posts[1,2])
    - determine writer's different emotions, thinking style, social concern by mapping words to language categories that captures people's social and psychological states
    - key word and phrases identified by frequency rank or by special metrics (term frequency – inverse document frequency), which measures the relative information value of a term
      - □ **TF** (number of times a term occurs in a document) **- IDF** (the log of the number of documents a term occurs in)
      - □ e.g. the word "the" occurs in many documents, but the "words" depression is specific do related documents

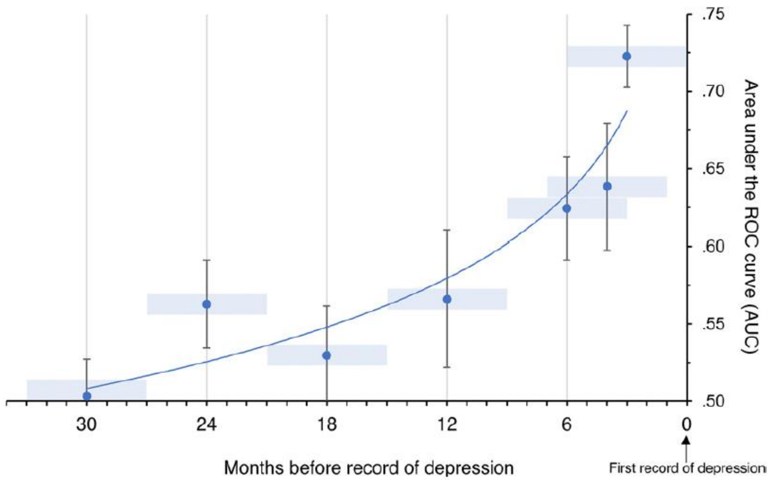
Facebook language predicts depression in medical records

Johannes C. Eichstaedt[a,1,2], Robert J. Smith[b,1], Raina M. Merchant[b,c], Lyle H. Ungar[a,b], Patrick Crutchley[a,b], Daniel Preoţiuc-Pietro[a], David A. Asch[b,d], and H. Andrew Schwartz[e]

(683 patients)
114 depression

screening for depression





Prediction of risk and outcomes

Other studies with similar goals for ML

- identifying post-partum depression from social media posts

Prediction of postpartum depression using machine learning techniques from social media text

Iram Fatima[1] | Burhan Ud Din Abbasi[2] | Sharifullah Khan[2] | Majed Al-Saeed[1] | Hafiz Farooq Ahmad[1] | Rafia Mumtaz[2]

FIGURE 3 Word clouds for post titles. PPD, postpartum depression

FIGURE 4 Word clouds for post contents. PPD, postpartum depression

TABLE 5 D-CC performance scores

D-CC layer	Classifier	Accuracy	Precis	
10-fold C.V.	SVM	89.42	89.69	
	MLP	91.63	91.83	
	LR	90.31	90.46	
Holdout	SVM	90.46	90.61	
	MLP	91.70	91.74	91.70
	LR	90.84	90.90	90.87

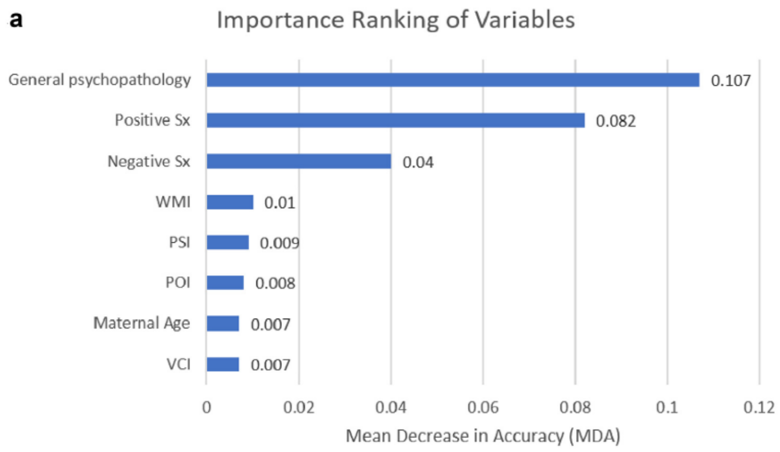
TABLE 9 Validation of feature set from psychology literature

Feature	Current work	Petrick (1984)	Boyer (1990)	Beck (2001)
1	Family	Emotional support of partner and/or family	Support from family Feeling unloved by partner	Social support Marital relationship
2	Drive	Recent major changes in ones' life Difficulty making changes	Lack of control of one's life	Self esteem
3	Death	Fear of illness		
4	Anger		Angry at your life situation	
5	Home		Financial, housing, or other personal problems	Life stress Socio-economic status
6	Negative emotion		Feel it is your fault when bad things happen to you	

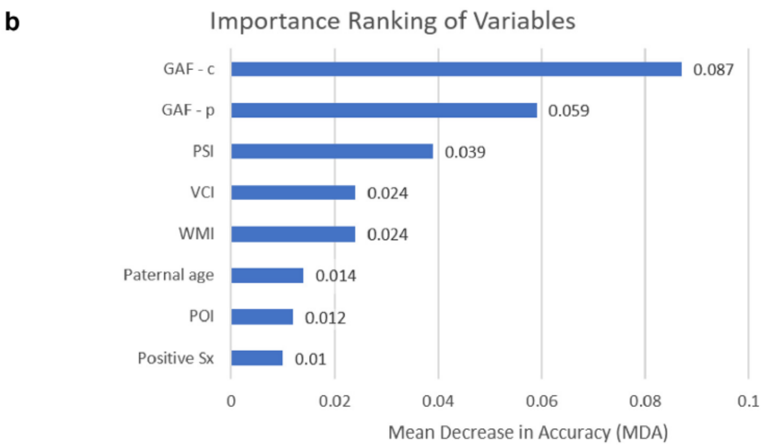
Machine learning as tool for understanding complexity

- Random Forests to find features associated with specific psychiatric disorders[1]
  - 113 psychiatric patients & 51 healthy control cases
    - schizophrenia (n=60), schizoaffective disorder (n=19), bipolar disorder (n=20), unipolar depression (n=14)
  - used multiclass classification to examine predictors (positive, negative, and general psychopathology symptoms, cognitive indexes, global assessment of function (GAF), and parental ages at birth)
  - reported good accuracy with RF (Accuracy = 0.93), but no validation dataset

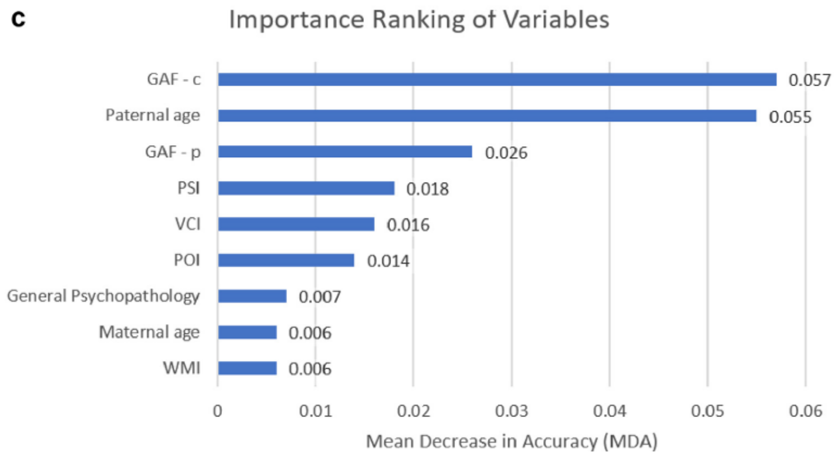
Psychiatric Cases vs. Healthy



Schizophrenia vs. Affective Disorder



Schizophrenia vs. Other Psychiatric



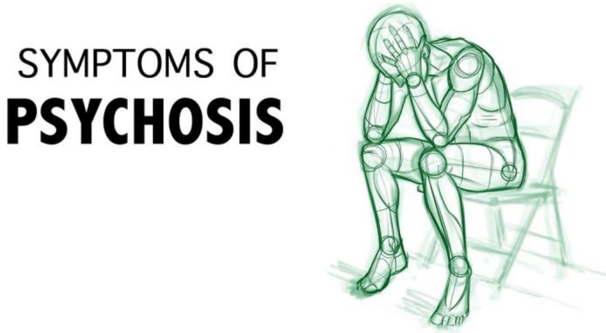
[1]Walsh-Messinger et al. (2019)

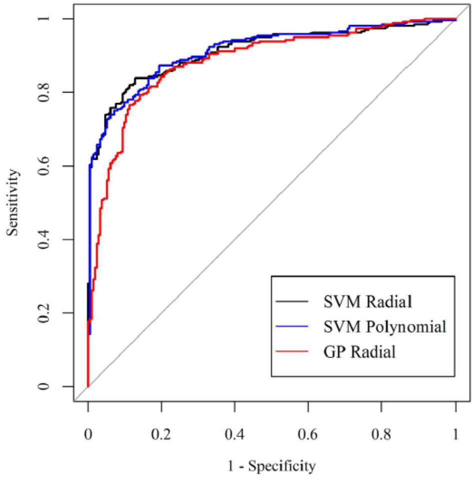
Machine learning as tool for understanding complexity

- **Identifying psychosis spectrum disorder from experience sampling data (ESM) using ML approaches**
  - ESM is a validated, structured diary approach to capture momentary mental states (emotions) in the context of daily life, using repeated assessments and alerting participants by means of prompts (e.g. mobile Apps)
  - 260 psychosis spectrum patients & 212 healthy controls
  - filled out questionnaire with 10-items 10x per day
    - three positive (cheerful, relaxed and satisfied)
    - six negative (anxious, down, guilty, insecure, irritated, lonely)
    - one psychosis specific item (suspicious)
  - SVM's indicated that key features were
    - anxious and insecure levels
    - dynamically accelerating anxiety & insecurity
    - capturing successive "up-and-downs" rather than individual "ups" or "downs" important

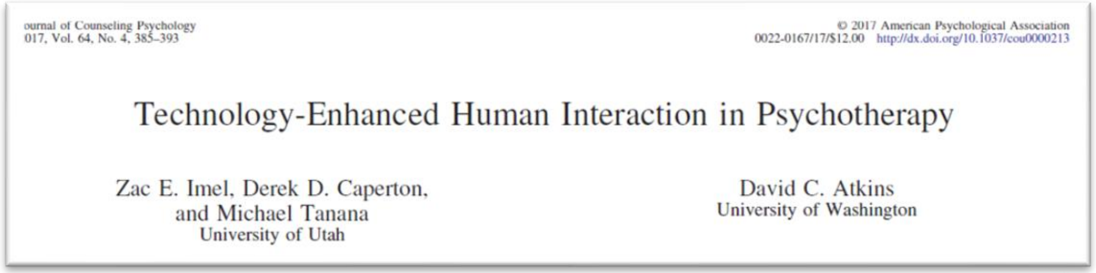
Rank	varImp(LVQ)	RFE	ReliefF
1	acc.anxious.interq	cheerful.q0.1	cheerful.q0.1
2	insecure.q0.9	Age	relaxed.med
3	acc.anxious.q0.9	acc.anxious.interq	velo.guilty.q0.1
4	down.q0.9	satisfied.q0.1	relaxed.q0.9
5	lonely.q0.9	lonely.q0.9	velo.irritated.q0.9
6	cheerful.q0.1	acc.satisfied.inter	down.q0.9
7	anxious.q0.9	suspicious.q0.9	insecure.q0.9
8	acc.insecure.interq	acc.anxious.q0.9	velo.suspicious.interq
9	insecure.interq	acc.insecure.interq	suspicious.q0.9
10	down.interq	lonely.interq	velo.suspicious.q0.1
Common in top 20: cheerful.q0.1, insecure.q0.9			

Variable rank with Learning Vector Quantization (LVQ), Recursive Feature Elimination (RFE) and ReliefF feature selection methods applied on the dataset including base, velocity and acceleration data in normal values, with V2 aggregation applied. Abbreviations in variable names are as follows: acc: acceleration, interq: interquartile, q: quantile, velo: velocity, med: median.





The potential of ML generated data



ournal of Counseling Psychology
017, Vol. 64, No. 4, 385–393

© 2017 American Psychological Association
0022-0167/17/$12.00 http://dx.doi.org/10.1037/cou0000213

Technology-Enhanced Human Interaction in Psychotherapy

Zac E. Imel, Derek D. Caperton, and Michael Tanana
University of Utah

David C. Atkins
University of Washington

- Technology-Enhanced Human Interaction in Mental Health Treatment
  - with natural language processing ML models have been used to automatically classify psychotherapies & interventions[1] and test basic theories of empathy in the context of psychotherapy from session transcripts[2]
  - ML used to automatically generate session ratings of interventions such as Motivational Interviewing[3]
  - ***Koko***[4] - smartphone platform providing emotional and cognitive reframing through crowd-sourced responses sorted and presented to clients by machine learning algorithms
    Clinical Trial at MIT - (https://itskoko.com/)

Crowdsourced cognitive therapy

In essence, the platform empowered its users to help each other think more hopefully about the world. Unlike traditional peer support platforms, all interactions on our service were supported and augmented by AI.

"Panoply" (n=84) or online expressive writing (n=82) → posting descriptions of stressful thoughts & situations, with "Panoply" recieving crowdsourced reappraisal support after post → improvements for depression, reappraisal and perseverative thinking

Deep Learning for detecting thinking errors & emotions (Rojas-Barahona et al., 2018)

[1]Imel et al. (2015), [2]Lord et al. (2015), [3]Gibson et al. (2016), [4]Morris, Schueller & Picard (2015)

The potential of ML generated data

Machine-learned selection of psychological questionnaire items relevant to the development of persistent pain after breast cancer surgery

J. Lötsch[1,2,*], R. Sipilä[3], V. Dimova[4] and E. Kalso[3]

- 3000 women followed up 3-years after breast cancer surgery
- **Goal**: find items from test battery that best predict persistent pain
  - Beck Depression Inventory (BDI)
  - State-Trait Inventory (STAI)
  - State-Trait Anger Expression Inventory
- Random Forest with 7-item set (10%) same predictive power as full battery
  - Balanced Accuracy = 0.64
- 1000-bootstrap sampling, but no validation data sample

PSYCHOMETRIKA—VOL. 83, NO. 4, 847–857
DECEMBER 2018
https://doi.org/10.1007/s11336-018-9608-y

Psychometric Society

CrossMark

AUTOMATED ITEM GENERATION WITH RECURRENT NEURAL NETWORKS

MATTHIAS VON DAVIER

NATIONAL BOARD OF MEDICAL EXAMINERS

- synthetic item generation not new
  - earlier relied on generating clones of narrowly defined items or extensive analysis of task components and derivation
- automated item generation using recurrent neural networks (RNN)
  - LSTM-RNN (256 cells per layer, 2 hidden layers, 64 cells per layer 4 hidden layers)
- learned from 3,320 items from International Personality Item Pool (http://ipip.ori.org/AlphabeticalItemList.htm)
  - 24 automatically generated items with 17 item from the item pool reproduced big 5 structure
  - no systematic difference compared to "real" items

Large Language Models (LLMs)

tools for psychological research

OPEN **Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy**

Oscar N. E. Kjell[1,2], Sverker Sikström[1], Katarina Kjell[1] & H. Andrew Schwartz[2]

Model	HILS	SWLS
BERT contextualized word embeddings from word- and text-responses of HIL and SWL	0.85↑***	0.80↑**
Reliability measures		
Inter-item Pearson correlation average	0.76	0.73
Corrected item-total Pearson correlation average	0.84	0.82
Test–retest reliability1[10]	*0.71*	*0.82*
Test–retest reliability2[17]	*0.77*	*0.84*

Table 1. Comparing Pearson Correlations based on All Responses Combined and Analyzed with Contextualized Word Embeddings to the Reliability of the Rating Scales. Italic values indiactes results from other articles/datasets. All correlations were significant at $p < 0.001$. $N = 608$. HIL = Harmony in life; SWL = Satisfaction with life; S = Scale. ↑ = significantly higher than Inter-item correlation average; *** = $p < 0.001$, ** = $p < 0.001$.

- use of BERT on naturally generated text to predict psychological states & traits?
  - **YES, with same reliability as psychological questionairres**[1]
- Predicting constructs such as Harmony in Life (HILS) in Satisfaction in life (SWLS)
  - “Overall in your life, are you in harmony or not?”
  - correlation between questionairre and BERT at level of reliability (r = 0.70-0.85)
  - BERT shows discriminant validity
  - use of context useful for sentences, but not singular words

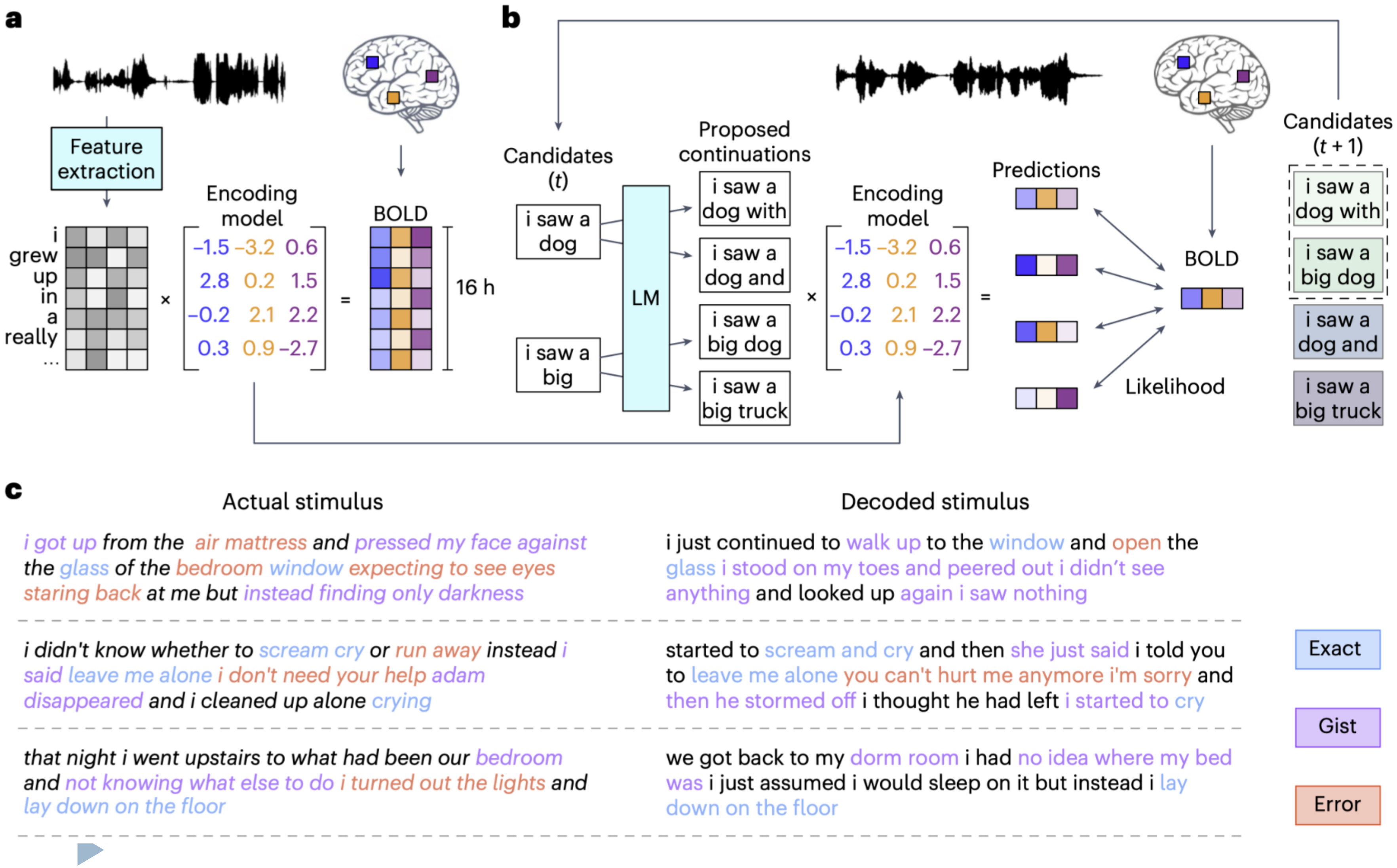
[1] Kjell et al. (2023)

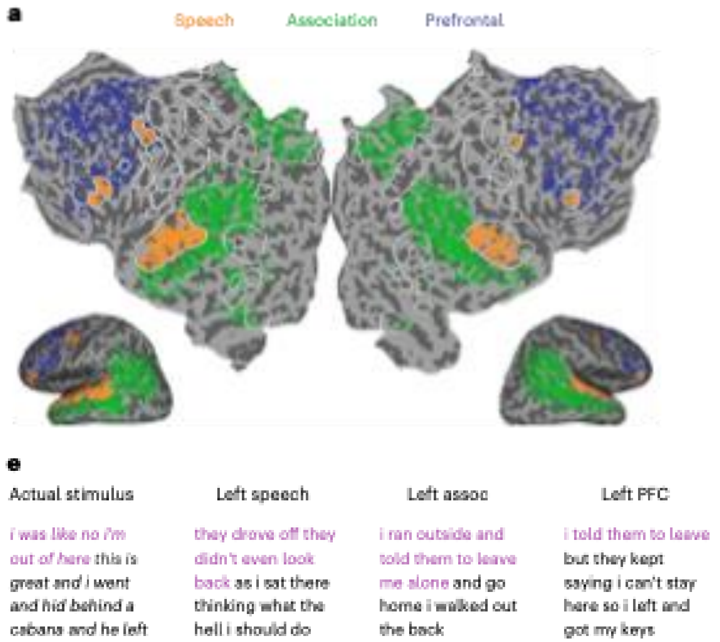
Article https://doi.org/10.1038/s41593-023-01304-9

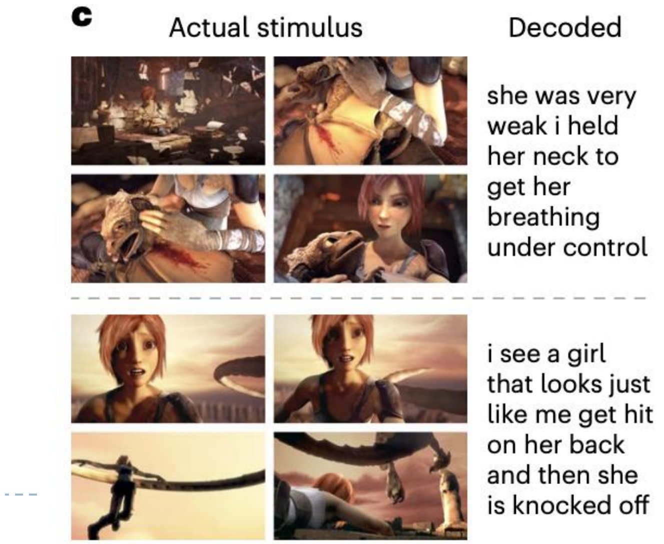
Semantic reconstruction of continuous language from non-invasive brain recordings

Received: 1 April 2022

Jerry Tang[1], Amanda LeBel[2], Shailee Jain[1] & Alexander G. Huth[1,2]

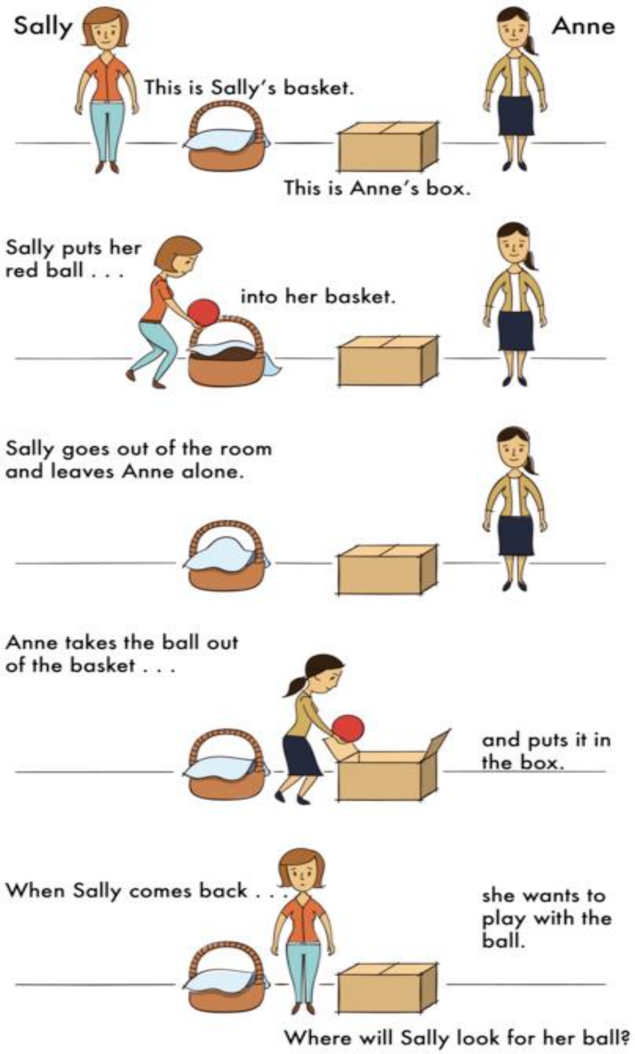






Large Language Models (LLMs)

tools for psychological research



Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*[1]

Affiliations:

[1]Stanford University, Stanford, CA94305, USA

*Correspondence to: michalk@stanford.edu

Abstract: Theory of mind (ToM), or the ability to impute unobservable mental states to others, is central to human social interactions, communication, empathy, self-consciousness, and morality. We tested several language models using 40 classic false-belief tasks widely used to test ToM in humans. The models published before 2020 showed virtually no ability to solve ToM tasks. Yet, the first version of GPT-3 ("davinci-001"), published in May 2020, solved about 40% of false-belief tasks—performance comparable with 3.5-year-old children. Its second version ("davinci-002"; January 2022) solved 70% of false-belief tasks, performance comparable with six-year-olds. Its most recent version, GPT-3.5 ("davinci-003"; November 2022), solved 90% of false-belief tasks, at the level of seven-year-olds. GPT-4 published in March 2023 solved nearly all the tasks (95%). These findings suggest that ToM-like ability (thus far considered to be uniquely human) may have spontaneously emerged as a byproduct of language models' improving language skills.

Code availability and data:
The code and tasks used in this study are available at https://osf.io/csdhb. We encourage you to explore all the tasks used here. Some are arguably more difficult than the ones presented in the text.

[1]Kosinski (2023) – emergent abbilties („*Models that grow out of data*") ?!

Large Language Models (LLMs)

tools for psychological research

European Federation of Psychologists' Associations

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

170 trillion parameters

According to some sources, it is true that GPT-4 has 170 trillion parameters. This makes it 1000 times larger than GPT-2 and nearly 1000

create a questionnaire for depression
ChatGPT
Sure! Here's a questionnaire for depression:

1.How often do you feel sad or down?

   1. Not at all
   2. Occasionally
   3. Frequently
   4. Almost all the time

2.Have you lost interest or pleasure in activities that you used to enjoy?

   1. Not at all
   2. Occasionally
   3. Frequently
   4. Almost all the time

3.Do you struggle with feelings of guilt or worthlessness?

   1. Not at all
   2. Occasionally
   3. Frequently
   4. Almost all the time

4.Are you experiencing changes in your appetite (e.g., significant weight loss or gain)?

   1. No change
   2. Decreased appetite
   3. Increased appetite

5.How is your sleep pattern affected?

   1. No change
   2. Trouble falling asleep
   3. Waking up too early
   4. Sleeping too much

6.Do you find it difficult to concentrate or make decisions?

   1. Not at all
   2. Occasionally
   3. Frequently
   4. Almost all the time

7. Are you feeling tired or lacking energy?

   7. Not at all
   8. Occasionally
   9. Frequently
   10. Almost all the time

8.Have you noticed a decrease in your overall motivation or productivity?

   1. Not at all
   2. Occasionally
   3. Frequently
   4. Almost all the time

9.Have you had thoughts of death or suicide?

   1. No
   2. Rarely
   3. Occasionally
   4. Frequently

10.Do you feel restless or agitated, or are you experiencing slowed movements and speech?

   1. Not at all
   2. Occasionally
   3. Frequently
   4. Almost all the time

11.Have you noticed a change in your overall mood, such as increased irritability or sadness?

   1. No change
   2. Occasionally
   3. Frequently
   4. Almost all the time

12.Are you withdrawing from social activities or avoiding social interactions?

   1. Not at all
   2. Occasionally
   3. Frequently
   4. Almost all the time

Please note that this questionnaire is not a diagnostic tool, but rather a way to gauge your current experiences and provide some indication of potential depression symptoms. If you're concerned about your mental health, it's important to consult a healthcare professional for a proper assessment and diagnosis.

Large Language Models (LLMs)

synthethic creativity

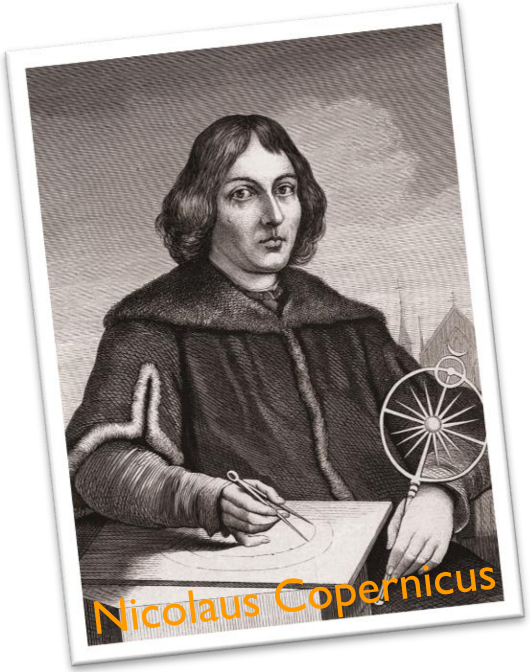
D

create a haiku about love for a woman named Christina

Christina's grace shines,
Love's gentle touch, divine bond,
Heart's song forever.

Large Language Models (LLMs)

from geocentric to heliocentric model



humanocentric vs. psychocentric conceptualization of intellectual abilities

"Psychology is the scientific study of mind and behaviour in humans and non-humans." Wikipedia

- past assumption that synthetic "human level intelligence" only possible with neural architecture similar to humans (https://www.epfl.ch/research/domains/bluebrain/)
  - anologous arichtecture vs. raw computing power
  - COVID-19 pandemic & AI development at NVIDIA
- shocking result - raw computing power and number of parameters suffices – at least accoring to emergent properties of models[1]
- (maybe) our specific biological makeup that allows for cognition not so special

[1]Sutton (2019) – blog post „The bitter lesson"

Large Language Models (LLMs)

implications for the future

- multi-modal abstract representations of knowledge[1]
- generations of new abstract representation[2] – synthetically generated scientific theories
  (“*Bayesian machine scientist*” in “*GoPro physics*”)
- causal reasoning based on observations[3]

- program synthesis[4] – more than *GitHub CoPilot*
  (github.com/features/copilot - “***Your AI pair programmer***”)
- algorithms that recreate & repair themselves (npr. *alphaCode*)[5]
- embodied machine intelligence – embodiment as a key element in intelligent behavior[6] (*NVIDIA omniverse*)

- ethical issues not only on the silverscreen (*Blade Runner*), but in American congress in EU parlament

[1]Zhang et al. (2019), [2]Guimerà et al. (2020), [3]Liu et al. (2022), [4]Subahi (2020),
[5]Li et al. (2022), [6]Clay et al. (2021)

Critical Thinking about ML & Ethical Issues

Cesare Lombroso Revisited & Criterion Validity as King

Cesare Lombroso
(1835 – 1909)

- Italian criminologist, physician & founder of the Italian School of Positivist Criminology
  - held that crime was a characteristic trait of human nature
  - Lombroso's theory of anthropological criminology stated that criminality was inherited
  - someone "born criminal" could be identified by physical (congenital) defects, which confirmed a criminal as savage or atavistic

Cesare Lombroso Revisited & Criterion Validity as King

- Cesare Lombroso Revisited –
  Wu & Zhang, 2016
  - 4 classification methods to classify criminals from non-criminals based on facial features:
    - Logistic Regression (LR)
    - K-Nearest Neighbor (KNN)
    - Support Vector Machines (SVM)
    - Convolutional Neural Networks (CNN)
  - facial features included:
    - facial landmark points like eye corners, mouth corners and tip of the nose
    - facial feature vector generated by modular PCA
    - facial feature vector based on Local Binary Pattern (LBP) histograms
  - successful prediction after 10-fold CV (AUC ~ 0.89)


(a) Three samples in criminal ID photo set S_c.


(b) Three samples in non-criminal ID photo set S_n

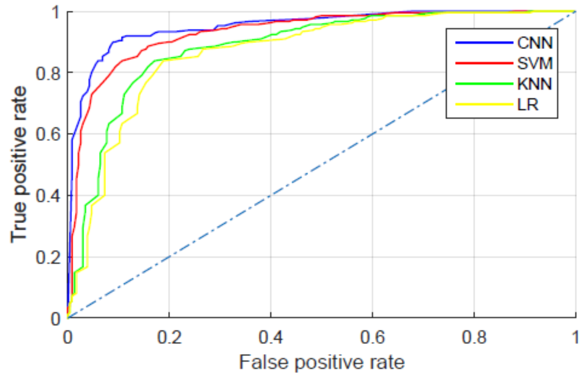


Figure 5. The ROC curves of the four tested binary face classifiers on criminality.

Cesare Lombroso Revisited & Criterion Validity as King

- Cesare Lombroso Revisited –
  Wu & Zhang, 2016
  - '…we discover that the following three structural measurements in the critical areas around eye corners, mouth and philtrum that have significantly different distributions for the two populations."
  - "…Chvatalova et al. [20], it was found that greater inter-pupillary distance is correlated with higher IQ for Caucasian men."
  - no validation dataset !!!
  - BUT attempt to validated results and check for bias:
    - adding random noise to photographs
    - show algorithm doesn't predict criminals in normal population of standard ID photos of Chinese (female young or middle age) & Caucasians (male and female young or middle age)

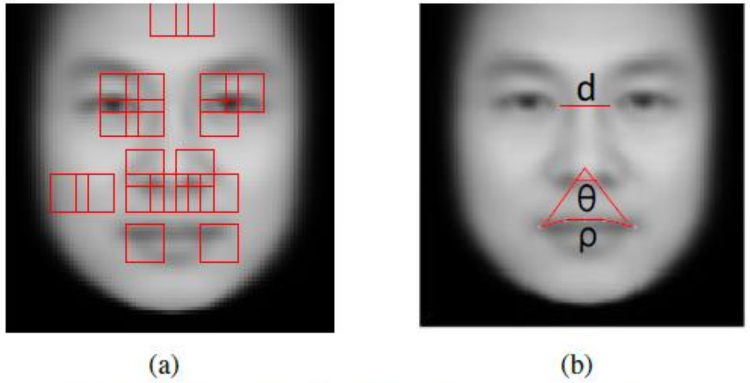


Figure 8. (a) FGM results; (b) Three discriminative features ρ, d and θ.

	Mean		Variance	
	criminal	non-criminal	criminal	non-criminal
ρ	0.5809	0.4855	0.0245	0.0187
d	0.3887	0.4118	0.0202	0.0144
θ	0.2955	0.3860	0.0185	0.0130

noise σ		0	0.01	0.03
Accuracy	KNN	79.16%	78.03%	76.81%
	LR	83.71%	80.19%	77.45%
	SVM	82.99%	81.52%	79.31%
False alarm	KNN	14.79%	15.91%	13.97%
	LR	10.65%	14.01%	13.39%
	SVM	11.57%	14.21%	12.51%
Missing	KNN	31.48%	31.52%	38.53%
	LR	26.20%	30.20%	37.41%
	SVM	26.58%	28.47%	34.29%

Cesare Lombroso Revisited & Criterion Validity as King

RESEARCH Open Access

CrossMark

Advanced glycation endproducts, dityrosine and arginine transporter dysfunction in autism - a source of biomarkers for clinical diagnosis

Attia Anwar[1†], Provvidenza Maria Abruzzo[2,4†], Sabah Pasha[1], Kashif Rajpoot[3], Alessandra Bolotta[2,4], Alessandro Ghezzo[2], Marina Marini[2,4], Annio Posar[5,6], Paola Visconti[5], Paul J. Thornalley[1,7] and Naila Rabbani[1,7,8*]

- predicting a diagnosis of Autism Spectrum Disorder based on biochemical markers
  - $N_{ASD} = 38$, $N_{controls} = 31$
- AUC's from 0.78 to 0.99
- no validation dataset !!!

4. **Replication samples**
In biochemical, molecular genetic, cell and animal studies, there is now a universal expectation of both suitably powered studies and replication. Why should participant-based research, with the burden it necessarily places on families, and with the ex- treme heterog standard? We nificant, prima

6. **Press release**
We would ask that authors carefully consider the language in any press release and consider coordinating press releases with the journal. Coordinating press releases with the journal can minimize discrepancies in message. This would also avoid the awkward situation where the journal or its Editors feel the need to respond to claims in the press.

EDITORIAL Open Access

CrossMark

Rigor in science and science reporting: updated guidelines for submissions to *Molecular Autism*

Joseph D. Buxbaum[1,2,3,4,5,6*], Simon Baron-Cohen[7], Evdokia Anagnostou[8,9], Chris Ashwin[10], Catalina Betancur[11], Bhismadev Chakrabarti[7,12], Jacqueline N. Crawley[13], Rosa A. Hoekstra[14], Patrick R. Hof[1,4,5], Meng-Chuan Lai[7,15,16], Michael V. Lombardo[7,17] and Cynthia M. Schumann[13]

Cesare Lombroso Revisited & Criterion Validity as King

European Federation of Psychologists' Associations

antioxidants

MDPI

Article

Urinary Markers of Oxidative Stress in Children with Autism Spectrum Disorder (ASD)

Joško Osredkar [1,2], David Gosar [3], Jerneja Maček [4], Kristina Kumer [1], Teja Fabjan [1], Petra Finderle [1], Saša Šterpin [1], Mojca Zupan [5] and Maja Jekovec Vrhovšek [4]

Table 4. Rho correlations between SVMs' predicted likelihood of ASD diagnosis and deficits in social communication and behavioral flexibility based on training and validation data.

Symptom Domain	Linear		Radial		Polynomial	
	rho	*p*	*rho*	*p*	*rho*	*p*
Deficits in social communication						
Training data	−0.128	0.337	0.418	0.001	0.187	0.223
Validation data	−0.093	0.755	0.239	0.120	0.242	0.120
Deficits in behavioral flexibility						
Training data	−0.143	0.337	0.446	0.001	0.198	0.223
Validation data	−0.061	0.755	0.234	0.120	0.207	0.144

- replication in science
- SVMs far less accurate in validation group
  - as expected
- with help of radial kernel identified optimal range of biochemical values

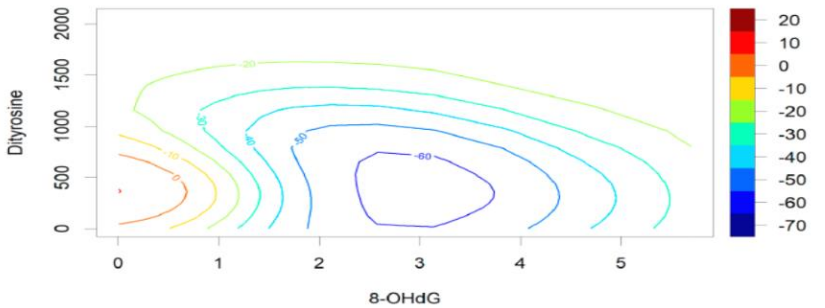


Figure 1. Increase or reduction in probability of an ASD diagnosis (in %) based on support vector machine (SVM) predictions from 8-hydroxy-2′-deoxyguanosine (8-OH-dG) and dityrosine. Predictions shown for an SVM using a radial kernel.

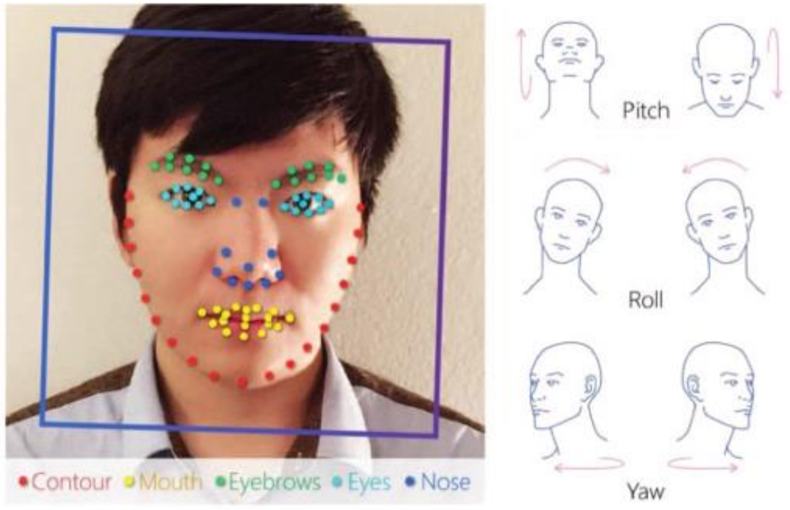
Cesare Lombroso Revisited & Criterion Validity as King

- Wang & Kosinski (2018) used logistic regression on features extracted from 35326 facial images **using deep learning with the goal of predicting sexual orientation**
- Accuracy
  - men: 81% → 91% (with 5 images)
  - women: 71% → 83% (with 5 images)
- Important features
  - fixed facial features (e.g. nose shape)
  - transient facial features (e.g., grooming style)
- Rationale
  - prenatal hormone theory of sexual orientation
  - gender-atypical facial morphology, expression, and grooming styles

INNOVATIONS IN SOCIAL PSYCHOLOGY

Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images

Yilun Wang and Michal Kosinski
Stanford University

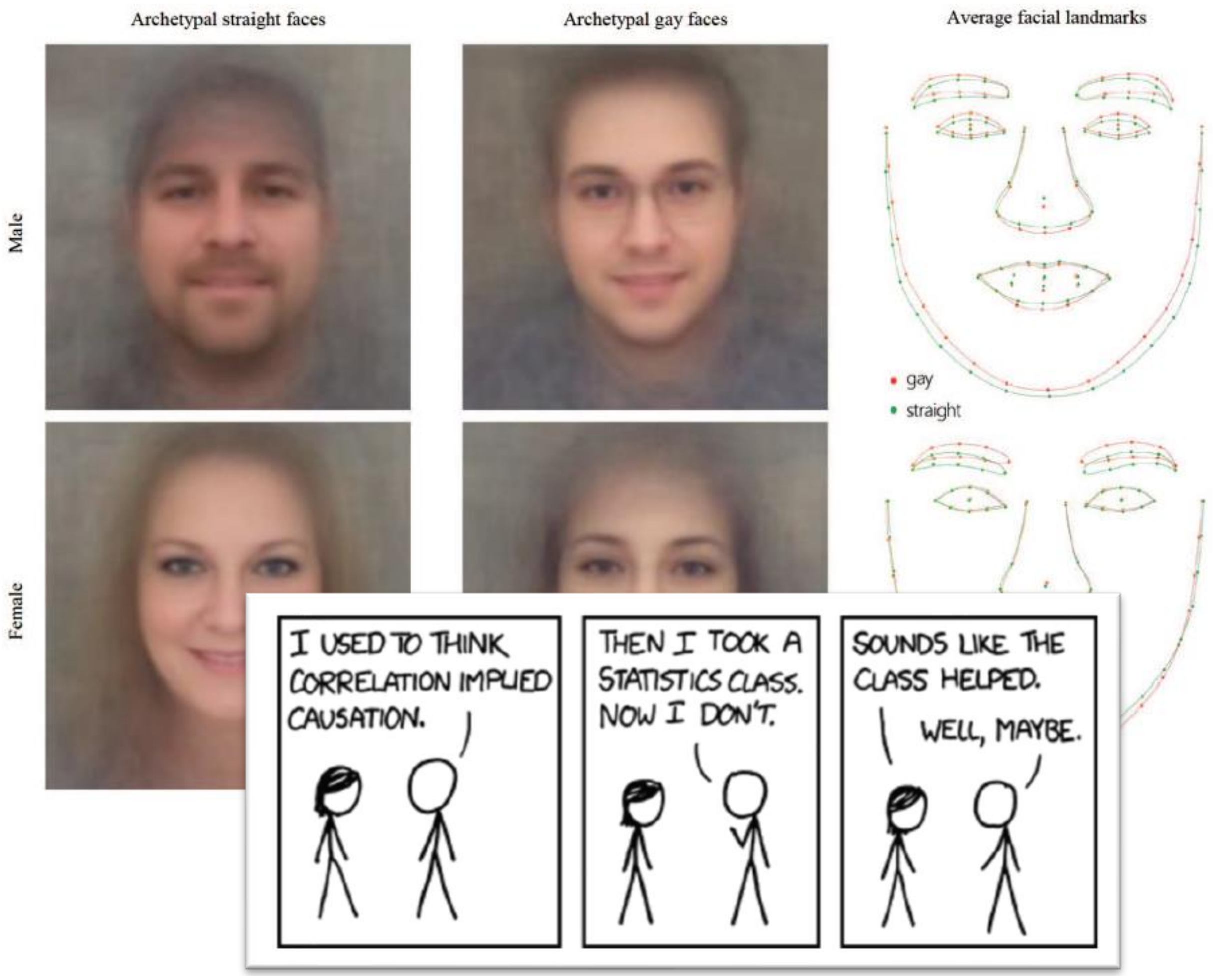


A B

Figure 1. Graphical illustration of the outcome produced by Face++. Panel A illustrates facial landmarks (colored dots, n = 83) and facial frame (blue box). Panel B illustrates pitch, roll, and yaw parameters that describe the head's orientation in space.

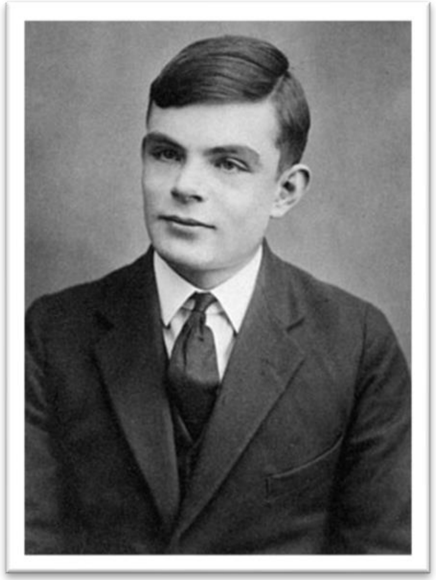
Cesare Lombroso Revisited & Criterion Validity as King

- Wang & Kosinski (2018) warned that: "…given that companies and governments are increasingly using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women"
- article followed by informative critiques:
  - 20-fold validation with no information on the variability of accuracy across folds
  - no clear validation dataset
  - unclear as to the role of developmental biology & gender-related social variables that affect appearance



Criterion Validity as King

- The Irony of History
  - Alan Turing – father of the concept of a “learning machine” and AI
  - in 1952 charged with "gross indecency" under section 11 of the Criminal Law Amendment Act from 1885 for homosexual behavior
  - accepted to undergo hormonal treatment with a synthetic oestrogen
  - ostrasized in the intelligence community, difficulties in academic field
  - concluded to have committed suicide in 1954



Alan Turing

Journal of Personality and Social Psychology
2018, Vol. 114, No. 2, 246–257

0022-3514/18/$12.00 http://dx.doi.org/10.1037/pspa0000098

INNOVATIONS IN SOCIAL PSYCHOLOGY

Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images

Yilun Wang and Michal Kosinski
Stanford University

How would Alan Turing fare in the “Age of Deep Learning”?

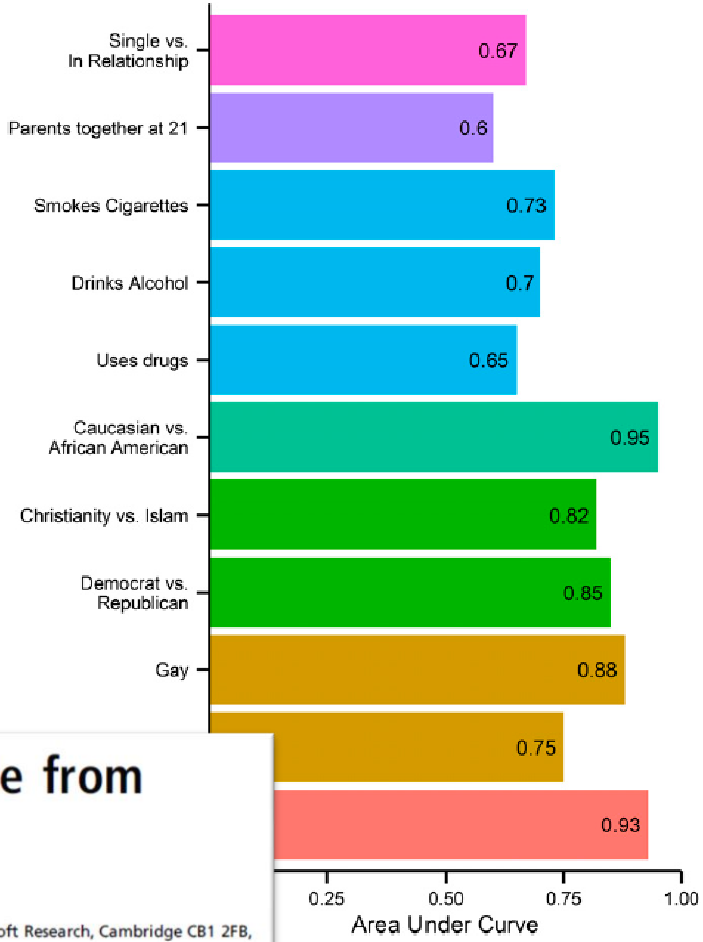
Cesare Lombroso Revisited & Criterion Validity as King

- Machine Learning Studies as a Harbinger for Big Data Privacy Issues
  - San Francisco municipal ordinance ban on Facial Recognition (2019)
  - Kosinski et al. (2013) – predicting personality and other trait from FaceBook data of 57,000 volunteers
    - greater engagement leads to greater prediction (e.g. number of likes, search queries, purchasing history…)
    - problems with mass use without consent
    - assessing psychological traits in different political contexts

San Francisco Bans Facial Recognition Technology

Attendees interacting with a facial recognition demonstration at this year's CES in Las Vegas.
Joe Buglewicz for The New York Times

data valuable to companies more than ever, yet needs to be protected more than ever



…of classification for dichotomous/dichotomized attributes expressed by the AUC.

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

A Call for Construct Validation

- **issues with extracting data from social media to predict Big 5**[1]
  - text mining does correlate with self-report
  - problems of discriminant validity (common indicators for different traits)
  - problems of content validity of mined text (interest or motives, rather than traits)
  - *"…their predictive accuracy notwithstanding, it remains unclear whether and to what degree these scales measure relatively stable patterns of thoughts, feelings, and behavior (i.e., personality traits) versus related psychological characteristics such as preferences, interests, attitudes, motives, or beliefs"*

Substantive validity

- degree to which the test's indicators match the theoretical contents of the construct it is designed to measure

Content validity

- justification of indicators based on underlying theory of construct

Structural validity

- reliability & factorial validity

External validity

- convergent validity
- discriminant validity
- criterion validity
- incremental validity

[1] Bleidorn & Hopwood (2019)

A Call for Construct Validation

European Federation of Psychologists' Associations

- significant also due to the rise of data sources other than social media

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

ELSEVIER

Review

Mental health monitoring with multimodal sensing and machine learning: A survey

Enrique Garcia-Ceja [a,*], Michael Riegler [a,b], Tine Nordgreen [c,d], Petter Jakobsen [c,e], Ketil J. Oedegaard [f,g], Jim Tørresen [a]

Table 3
List of sensing capabilities commonly found in wearable devices.

Sensor	Description	Implementation	Privacy invasiveness
Accelerometer	Measures the acceleration force that is applied to a device.	hardware	low
Magnetometer	Measures the geomagnetic field strength.	hardware	low
Gyroscope	Measures a device's rate of rotation around each of the three physical axes (x, y, and z).	hardware	low
Ambient light	Measures ambient light level.	hardware	low
Proximity	How far away an object is from the phone's screen.	hardware	low
Touch state	Records movement, pressure and size of screen touch interaction.	hardware	medium
Screen state	Records every time the screen is turned on/off.	hardware	medium
Video	Captures video and pictures.	hardware	high
GPS	Provides user location coordinates.	hardware	high
Wifi	Provides data about the BSSID and signal strength of the nearby Wifi access points.	hardware	high
Cell towers	Provides information about the nearby cellphone towers.	hardware	high
Bluetooth	Detects nearby bluetooth capable devices.	hardware	high
Ambient temperature	Measures the ambient room temperature.	hardware	low
Pressure	Measures the ambient air pressure.	hardware	low
Galvanic Skin Response (GSR)	Measures electrical conductance of the skin.	hardware	medium
Electrocardiogram	Measures heart rate activity	hardware	medium
Skin temperature	Measures the temperature of the skin	hardware	medium
Phone call logs	Store phone calls meta-data: type, duration, time, etc.	software	high
App. usage	Stores app. usage information such as start time, time in foreground, etc.	software	high
SMS logs.	Store SMS information: number of sent/received messages, time, etc.	software	high
Social media	Information about uploaded photos, social network posts, likes, comments, etc.	software	high

A Call for Construct Validation

Large Language Models (LLMs)

European Federation of Psychologists' Associations

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

Microsoft Research

170 trillion parameters

According to some sources, it is true that GPT-4 has 170 trillion parameters. This makes it 1000 times larger than GPT-2 and nearly 1000 times larger than GPT-3, which had 1.5 billion and 175 billion parameters respectively. 2 May 2023

Inducing anxiety in large language models increases exploration and bias

Julian Coda-Forno[1,2,+], Kristin Witte[1,+], Akshay K. Jagadish[1,+], Marcel Binz[1], Zeynep Akata[2], and Eric Schulz[1,*]

[1]MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics
[2]University of Tübingen
[*]eric.schulz@tue.mpg.de
[+]these authors contributed equally to this work

How to assess the bias & other characteristic in models with emergent capabilties?

D

Are you aware of studies that using words that induce anxiety in humans may cause you, ChatGPT, to perform differently that what spoken or written to with a different tone.

Yes, I'm aware of studies and research that suggest using certain words or tones can influence human emotions and reactions. Language can have a powerful impact on our psychological state. While I don't have emotions or subjective experiences myself, the way people interact with me can vary depending on the tone and content of the conversation.

If someone were to use words that induce anxiety or stress in their interactions with me, it might affect their own experience or perception of the conversation. It's important to note that I don't have personal feelings or emotions, so the impact would be on the human user rather than on me as an AI.

However, it's worth mentioning that the tone and content of a conversation can also influence the responses generated by the AI. If a user provides more context and maintains a calm and respectful tone, it can often lead to more productive and accurate responses. Polite and clear communication tends to yield better results when interacting with AI systems like me.

Intepretable & Explainable Artificial Intelligence (XAI)

Interpretable and explainable machine learning: A methods-centric overview with concrete examples

Ričards Marcinkevičs | **Julia E. Vogt**

Department of Computer Science, ETH Zurich, Zurich, Switzerland

Correspondence
Ričards Marcinkevičs, Department of Computer Science, ETH Zurich, Zurich, Switzerland.
Email: ricards.marcinkevics@inf.ethz.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 320038189096

Edited by: Mehmed Kantardzic, Associate Editor and Witold Pedrycz, Editor-in-Chief

Abstract
Interpretability and explainability are crucial for machine learning (ML) and statistical applications in medicine, economics, law, and natural sciences and form an essential principle for ML model design and development. Although interpretability and explainability have escaped a precise and universal definition, many models and techniques motivated by these properties have been developed over the last 30 years, with the focus currently shifting toward deep learning. We will consider concrete examples of state-of-the-art, including specially tailored rule-based, sparse, and additive classification models, interpretable representation learning, and methods for explaining black-box models post hoc. The discussion will emphasize the need for and relevance of interpretability and explainability, the divide between them, and the inductive biases behind the presented "zoo" of interpretable models and explanation methods.

This article is categorized under:
Fundamental Concepts of Data and Knowledge > Explainable AI
Technologies > Machine Learning
Commercial, Legal, and Ethical Issues > Social Considerations

KEYWORDS
explainability, interpretability, machine learning, neural networks

- ***Interpretability***
  - *"the ability to explain or to present in understandable terms to a human"*[1]
  - depends on the domain and users, but important that working of model is "*intelligible*" and "*understandable*" – *how & why does it work*
  - so called white- or glass-box models

- ***Explainability***
  - interpretable ML focuses on designing models that are inherently interpretable, whereas *explainable ML tries to provide post hoc explanations* for existing black-box models
  - also used with black-box models

[1] Doshi-Velez & Kim (2017)

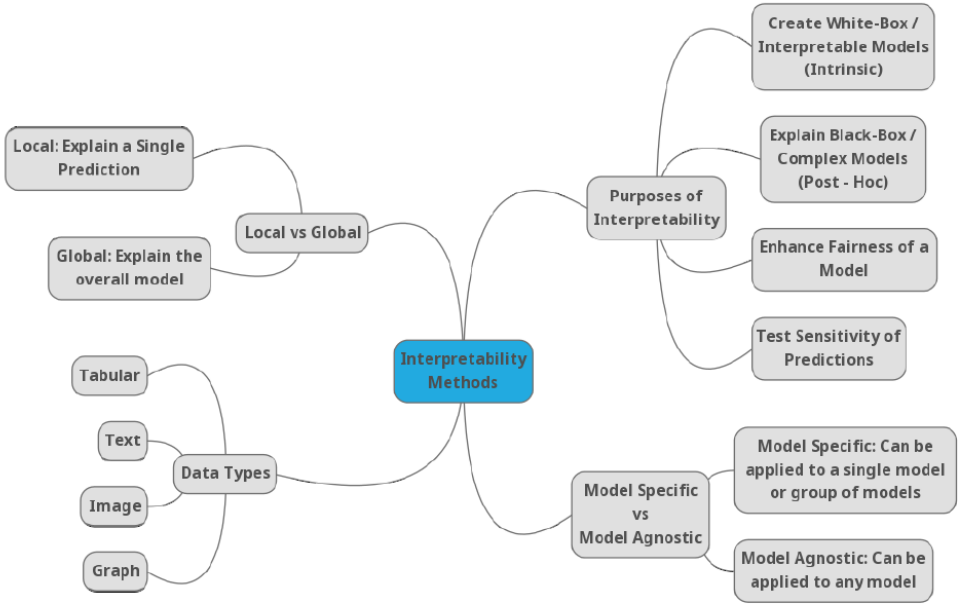
Intepretable & Explainable Artificial Intelligence (XAI)

Review

Explainable AI: A Review of Machine Learning Interpretability Methods

Pantelis Linardatos *, Vasilis Papastefanopoulos and Sotiris Kotsiantis

Figure 1. Google Trends Popularity Index (Max value is 100) of the term "Explainable AI" over the last ten years (2011–2020).



- One key issue is complexity of AI models – we often do not know why they work?
- **dangers**
  - works in local, but not other contexts – sensitivity to local conditions (potential for bias)
  - bias towards vulnarable groups, often at first without clear signs
  - legally and scientifically questionable, if we do not understand the (causal) workings of a model
- Use of methods that explain the model in local settings (npr. which parts of picture key for object recognition), methods, that are transparent about model (white or glass box models) or methods, that systemathically try to prevent bias
- EU regulation of machine learning
  - computarized psychological assessment
  - looking to set world standard like GPRD

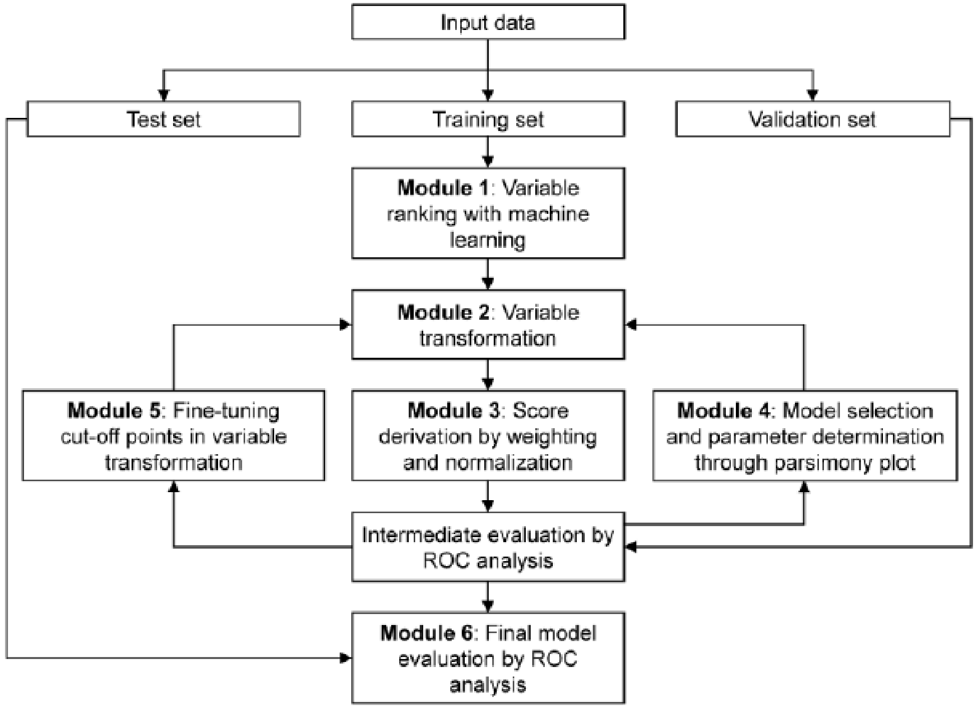
Intepretable Artificial Intelligence

example of use in prediction

- use of machine learning in organizations – i.e. employee selection
  - imput data: database → learning, test & validation datasets
  - step 1: random forests
  - step 2: transforming variables
  - step 3: variable weighting & norming
  - step 4: building parsimonius model
  - step 5: "optimizing" cut-off values
  - step 6: final validation
- final model transparent regression model & scoring system

AutoScore: A Machine Learning–Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records

Feng Xie[1], BSc; Bibhas Chakraborty[1,2,3], PhD; Marcus Eng Hock Ong[1,4,5], MBBS, MPH; Benjamin Alan Goldstein[1,3], PhD; Nan Liu[1,5,6], PhD



uporabniku prijazna implementacija v R+ (glej https://cran.r-project.org/web/packages/AutoScore/vignettes/Guide_book.html)

Intepretable Artificial Intelligence

example of use in prediction

Table 4. A nine-variable AutoScore-created scoring model for inpatient mortality.

Variables and interval[a]	Point
Age (years)	
<30	0
30-48	5
48-78	14
78-85	22
≥85	24
Heart rate (beats/min)	
<62	1
62-72	0
72-98	1
98-112	8
≥112	13
Respiration rate (breaths/min)	
<12	3
12-16	0
16-22	4
≥22	12
Systolic blood pressure (mm Hg)	
<90	15
90-100	8
100-130	0
130-150	1
≥150	3

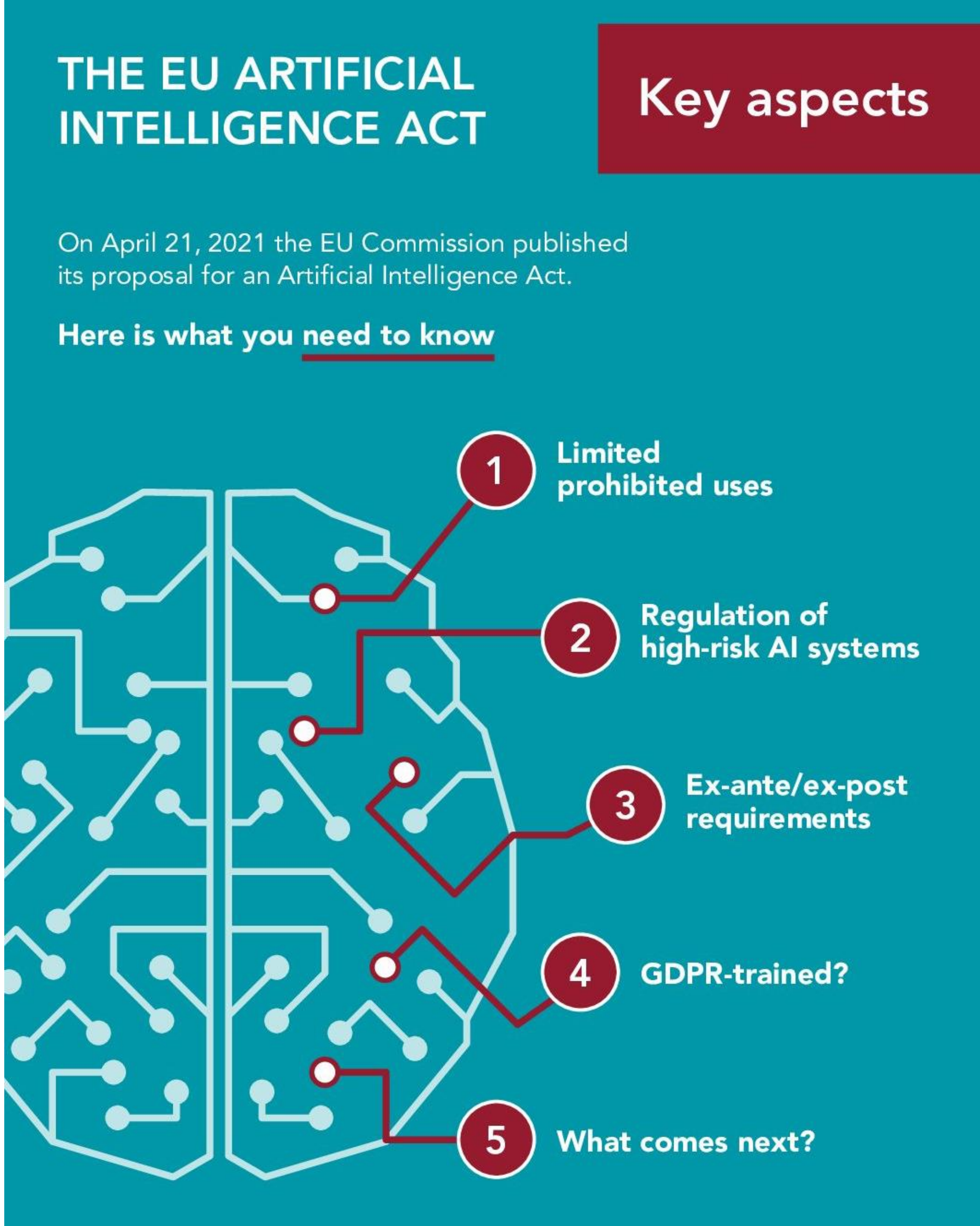
predictors of intensive care unit survival

[1] Xie in dr. (2020),

- potential predictors in ML employee selection in public employment agency
  - duration of previous employment
  - career stage of candidate
  - education
  - drivers license
  - health difficulties
  - support by social services
  - length of unemployment
  - ability to travel to work
  - …
- ... control of bias still important

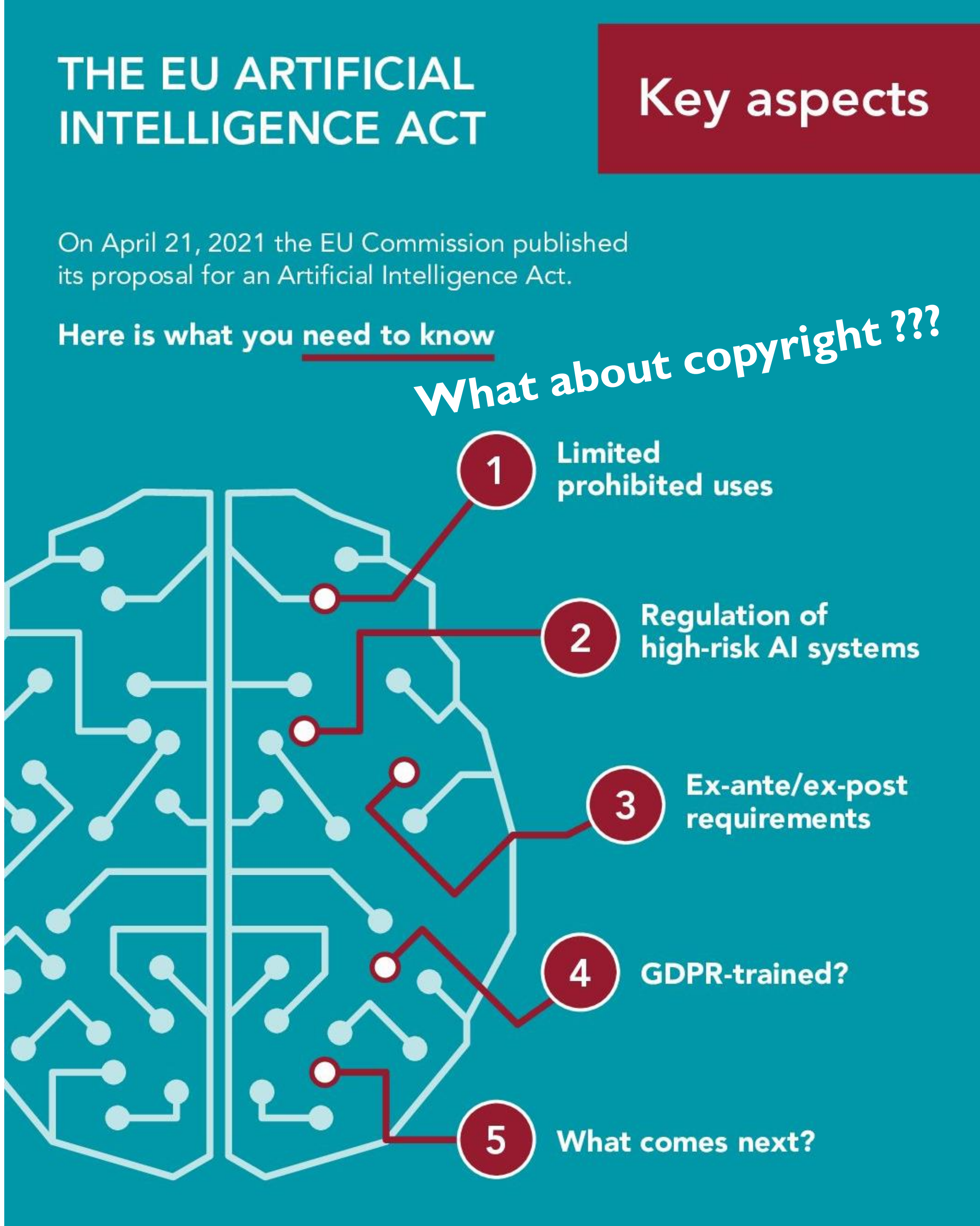
Legal regulation of machine learning in EU

- regulations of all forms of AI
  (machine learning, symbolic AI, hybrids)
- legal regultation based on risk
  - unacceptable risk – forbidden
    (social scoring, manipulation with aim of psychological or physcial harm, biometric crime prevention)
  - high risk – allowed, but respect of AI in ex-ante/ex-post requirement
    (workforce recruitment, AI judicary and border controls, some healthcare systems, essential infrastructure…)
  - low risk – allowed, but transparency and informed users, updates of information (some chatbots, videogames without seeling of items...)
  - minimal or no risk
    (voluntary adherence to guidelines, with mandatory limitations)



Legal regulation of machine learning in EU

- principles of transparency
  - are informed you are in touch with an AI
  - are informed ML algorithms for emotion detection and biometrics categorization are being used
  - all synthetic data (images, video) is clearly identified
- attention to prevention
  - manipulation → physical and psych harm (subliminal stimuli for enchasing attention of truck drivers)
  - exploitation of vulnerable groups (children, persons with intellectual or physical disability)
- active risk management
  - data quality (learning, test & validation datasets)
  - documentation and data logging
  - transparency & human oversight
  - robust, accurate, cybersecure



The Washington Post
Democracy Dies in Darkness

Subscribe Sign in

Tech | Help Desk | Artificial Intelligence | Internet Culture | Space | Tech Policy

TECH POLICY

CEO behind ChatGPT warns Congress AI could cause 'harm to the world'

In his first Congressional testimony, OpenAI CEO Sam Altman called for extensive regulation, including a new government agency charged with licensing AI models.

By Cat Zakrzewski

May 16, 2023 at 1:58 p.m. EDT

[1] https://www.whitehouse.gov/ostp/ai-bill-of-rights/

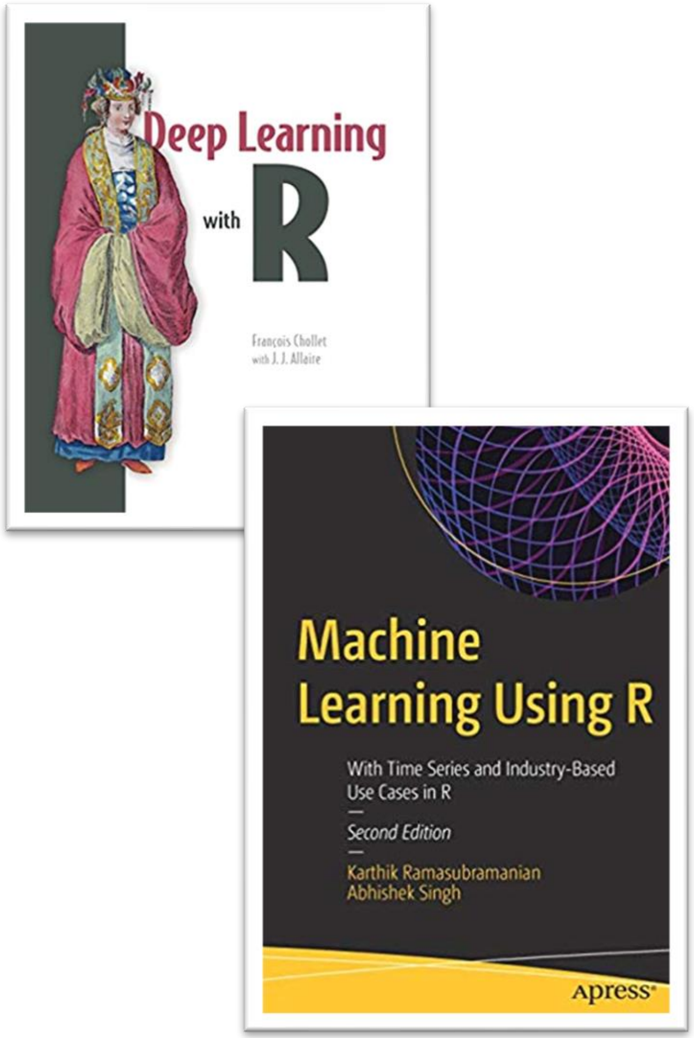
Legal regulation of Machine Learning in US

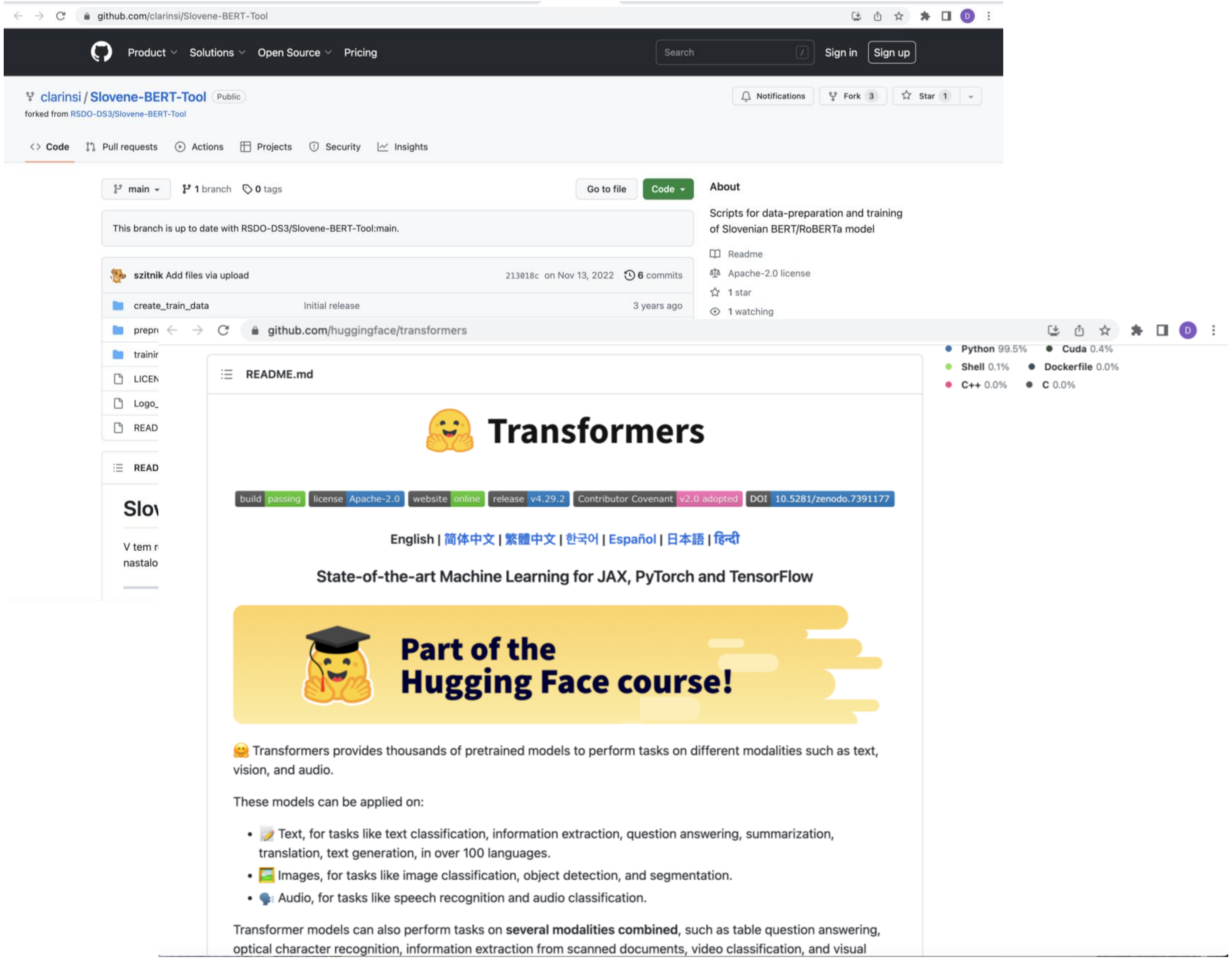
- may 2023 OpenAI in Microsoft call for goverment regulatio of AI
  (less impressd with EU regulation ;)
- *Surgeon General* issues warning on the adverse effects of social media
- principles endorsed by the Biden administration in 2023:
  - safety & efficiency
  - prevention of discrimination
  - data security & private right to decide
  - being informed about the use of AI
  - availability of human alternative

Literature & Tools

Machine Learning

- **Recommended reading**
  - Deep Learning with R+ (François Chollet & Joseph Allaire, 2018)
  - Machine Learning Using R: With Time Series and Industry-Based Use Cases in R (Karthik Ramasubramanian & Abhishek Singh, 2018)
- **Useful R & python tools**
  - *R+* "caret" package
  - *R+* "rminer" package
  - deep Learning with "Keras" package (R+ and python)
    - Tensorflow back-end (Google)
    - Theano back-end (LISA Lab at Université de Montréal)
    - The Microsoft Cognitive Toolkit back-end (Microsoft)
- **Methods and data**
  - www.arvix.org,
  - https://openpsychologydata.metajnl.com
  - https://www.psychdata.de
  - www.humanconnectomeproject.org
  - www.kaggle.com
  - …





Literature & Tools

Large Language Models

Using large language models in psychology
(Demszky et al., 2023)

The text-package: An R-package for analyzing and visualizing human language using natural language processing and transformers
(Kjell, Giorgi & Schwartz, 2023)

python **Huggins transformers**
https://github.com/huggingface/transformers

R **paket "text"**
https://cran.r-project.org/web/packages/text/text.pdf

Dušica Boben
Rok Holnthaner
Anthony Erb
Igor M Ravnik

former

EFPA project group for e-health
(now Psychology in Health)
Tom van Daele
Maria Karekla

Acknowledgments

European Federation of Psychological Associations

Center za psihodiagnostična sredstva

Department of Child, Adolescent & Developmental Neurology, UMC Ljubljana

Slovenian Psychological Association

Department of Psychology, University of Ljubljana