barcelona|gse
graduate school of economics

Master Thesis

Topic:

Unsupervised learning in decision making

Author:	Domagoj Fizulic
	Felix Gutmann
Student number:	125604
	125584
Program:	M.S. Data Science
E-Mail:	domagoj.fizulic@barcelonagse.eu
	felix.gutmann@barcelonagse.eu

I Table of Contents

II List of Figures

III List of Tables

IV List of Listings

V List of mathematical symbols

Symbol	Meaning
a_t	Action at time t
$Q(a)_t$	Value function at time t
ϵ	Probability of exploration in epsilon greedy
α	Learning rate
τ	Softmax parameter
X	Random variable
$H(X)$	Entropy of a discrete random variable X
$d(\cdot, \cdot)$	Distance Function
$S(\cdot, \cdot)$	Similarity Function
$\mathbb{R}_0^+$	Positive real numbers including zero
$\mathcal{X}$	Data set
$\mathbf{W}$	Weighted adjacency matrix
d_i	Degree of node i
$\mathbf{D}$	Diagonal matrix of degrees
$\mathbf{L}$	Graph laplacian

VI List of abbreviations

Abbreviations	Description

1 Introduction and conceptual approach

Learning is a complex procedure. The learning procedure can affected due to social conditions... Due to their mental ability people should show different learning behavior.
In *supervised learning* data are predicted by training a classifier based on features and corresponding categories . In comparison to that in *unsupervised learning* we don't know the ground truth and we try discover natural clustering behavior in the data. [Murphy, 2012, page 9 et. seqq.]. A vast class of clustering algorithms are proposed in the literature. In this paper we are investigating whether we can use unsupervised learning methods to cluster different groups of people with respect to their behavior. This paper is therefore in the intersection of machine learning and cognitive science. To our knowledge this particular setting was not studied before.
We first used a reinforcement learning based approach to simulate data and study theoretical boundaries of our methods and under which conditions they might be applicable. We then apply our methods to two real world data sets. In particular we have data of people with different criminal profile performing a standardized test. Furthermore, we have data of people with different sex, level of education and age doing a simple n-armed bandit experiment.

The report has the following structure. In section two we provide a short overview of related literature in the field.
The following section three is dedicated to the theoretical foundation and the simulation. We first provide a basic knowledge of reinforcement and line out our experiment design. Since this a pioneer work we address the topic in a broad manner. Hence, this section includes an overview of applied similarity concepts and algorithms. Finally we present the results of our simulation and a short interim conclusion.
After that we provide results for the real world data. Since this a first research on this subject we can only address a small part of the problem. Therefore, we close the paper with discussion of possible extensions and summary of found results.

2 Relevant Literature

There exists a rich literature on identifying behavioral groups using the Iowa Gambling Task experiment. It has been shown that individuals with pre-frontal brain damage and decision-making defects continue to choose disadvantageously even after they learned the optimal strategy [Bechara et al., 1997]. A broad overview is given in [Steingroever et al., 2013] Several studies identify specific drug-user groups, e.g. cocaine addicts [Stout et al., 2004], chronic cannabis users [Fridberg et al., 2010], heavy alcohol users(heavy drinkers) [Gullo and Stieger, 2011]. Extensive set of research is focused around particular mental disabilities, e.g. Asperger's disorder [Johnson et al., 2006], psychopathic tendencies [Blair et al., 2001], bipolar disorder [Brambilla et al., 2012], schizophrenia [Martino et al., 2007] pathological gambling disorder [Cavedini et al., 2002], attention-deficit-hyperactivity disorder [Nirit Agaya, Eldad Yechiama, Ziv Carmelb, 2010]. Most popular reinforcement learning models for identifying behavioural differences between different disorders are Expectancy Valence model [Busemeyer and Stout, 2002] and Prospect Valence Learning model [Ahn et al., 2008].

3 Theoretical Background and simulation experiments

This section is dedicated to the a more detailed outline of our analysis approach and the results of our simulation experiments. We find that our data are not immediately suitable for common unsupervised learning methods. Hence, we also discuss how we approach this problem. Besides the introduction to reinforcement learning this section will be rather a qualitative discussion. The appendix provides in more detail a mathematical background to applied algorithms, distance concepts and related clustering evaluation techniques.

3.1 Experiment design and problem formulation

Before analyzing real data we run some simulation experiments to identify adequate approaches to model our data. To generate data artificially we follow a reinforcement learning and multi arm bandit approach where we let an agent learn the distribution of a simulated data set. The procedure is illustrated in figure 1. To keep things simple we draw a set of rewards from a normal distribution. We let several agents with different parameter setting process the reward data. From this procedure we obtain for each agent a time series of choices.

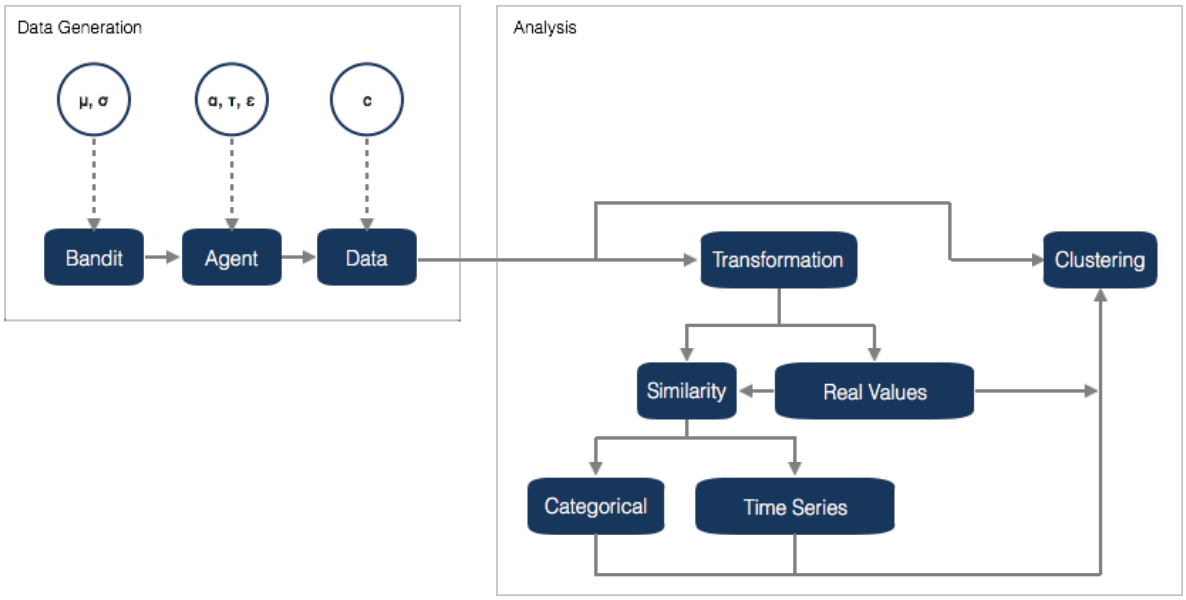


Figure 1: Flowchart experiment design

3.2 Reinforcement Learning background and multi arm bandits

In the following we provide some basics of *Reinforcement Learning* (RL) [Sutton and Barto, 2012, chapter 1 and 2]. RL is a branch of *machine learning* try model how an artificial agents interact with its environment and learns from the process over time.
In particular an agents is confronted with the task of choosing sequentially from a set of choices. In comparison to *supervised learning*, where an agent is learning based on set of examples an agent in REL doesn't have any knowledge about the system apriori. Therefore, it has to learn the nature of the system by sequentially interacting with its environment and keeping tack of the obtained information. Since the agent doesn't have any apriori information about the system it has to explore new possible action and so has to deviate from the optimal action. Furthermore, it has to keep track of value of each action he did so far. So the main task of the agent is to balance exploration and exploitation. There are to basic approaches to model this trade-off; An *"Epsilon-Greedy"* selection method and Soft *"Softmax"* selection method. Before explaining both concepts we introduce the value function for a given action a. Therefore, let $Q_t(a)$ be the value function of action defined as:

$$Q_t(a) := \frac{R_1 + R_2 + \cdots + R_{K_\alpha}}{K_\alpha} \tag{1}$$

The value function is the average over rewards.

Considering now epsilon greedy action selection method: The rule in general is to select the next action as the current current highest value function. However, to model exploration we introduce a random element to deviate from that greedy strategy. Following that the next action

$$a_{t+1} = \begin{cases} \text{random action} & \text{, with probability } \epsilon \\ \arg\max_i Q_t(i) & \text{, with probability } 1-\epsilon \end{cases} \tag{2}$$

where $\epsilon \in [0, 1]$ is a parameter controlling the random behaviour of the agent.

In the softmax action selection method compute for each action a probability (also called *Boltzmann Distribution).* The probability for action a is computed by:

$$P(a_t|X) = \frac{e^{\frac{Q_t(a)}{\tau}}}{\sum_i^K e^{\frac{Q_t(i)}{\tau}}} \tag{3}$$

In each iteration the next action of the agent is drawn with probability p_a:

$$a_{t+1} \sim p_a \tag{4}$$

After selecting an action the agent is updating its believe of the chosen action. Formally the update rule is defined

$$Q(a)_{k+1} = Q(a)_k + \alpha \left[R(a)_k - Q(a)_k \right], \tag{5}$$

where α is a is the non negative *learning rate* defining how much the current action is affecting the believes.

3.3 Unsupervised learning methods, data handling and similarity

We find two main challenges. First, our data have a categorical nature. Furthermore, the learning process also imposes a time series dependence on the data.
A well studied approach clustering such data are *hidden markov models*. For example a decent research on that can be found in [Pamminger, 2007], [Pamminger and Fruhwirth-Schnatter, 2009 and [Pamminger and Fr??hwirth-Schnattery, 2010]. However, we focus our attention only on partition based clustering algorithms.
Some of applied algorithms operate on different distance or similarity concepts. Given our data we have to think carefully think about distance and similarity measures to respect the nature of our data and algorithms.[1] In extension to only considering raw choices we might try to re-express our data to discriminate them according to different purposes. The first approach is to map the series of choices to a real valued series. One way is to consider *Shannon's Entropy* introduced by [Shannon, 1948] based on the empirical probability of the choices. Let X be a discrete random variable with probability p, then the entropy is defined as [MacKay, 2005, page 32]:

$$H(X) := -\sum_{i=1}^{N} p_i \log_2 p_i \tag{6}$$

Entropy gives measure on how random a random variable behaves. Withing the entropy framework we consider to types of entropies. For each time step we compute the entropy using the choices done so far. We call this *cumulative entropy*. Furthermore, we might want to observe more clearly how individuals adapt their behavior over time.

[1] A formal definition of the distance and similarity can be found in the appendix

Considering an experiment with 100 trials. We then compute the entropy for set of ten choices. We call this *blockwise entropy*. Mapping choices to and entropy based data set is therefore aims to discriminate individuals by their level randomness in their behavior.

A second approach is based on the experimental setting. We know within out simulation and our simulation and real data frame work the set of choices which are disadvantageous for the participant. Following e.g. [Yechiam et al., 2008] or [Ahn et al., 2008], we then compute block wise the ratio of disadvantageous choices. We call this *disadvantageous choice*.

Within our analysis we consider a broad selection of several clustering techniques and similarity concepts. Table 1 shows an overview of the algorithms and their corresponding distance/similarity requirement. A technical description for all of them is provided in the appendix.

Algorithm	Input	Datatype
Spectral Clustering	Similarity	-
Affinity Propagation	Similarity	-
K-Means Clustering	L_2 Distance	-
Ward Clustering	Data Matrix	-
PCA + Ward Clustering	Data Matrix	-

Table 1: Overview clustering algorithms

Accoding to our different data types and for our different algorithms we apply the following distance and similarity concepts.[2]

[2]Note that distance and similarity can be converted according to equation (7) in the appendix

Similarity/Distance	Data type
Cosine Similarity	Categorical
Levenstein Distance	Categorical
Lin Similarity	Categorical
Edit Distance on Real Sequence	Time series
Dynamic Time Warp	Time series
RBF Kernel	Real Data

Table 2: Overview similarity and distance measures

3.4 Simulation results

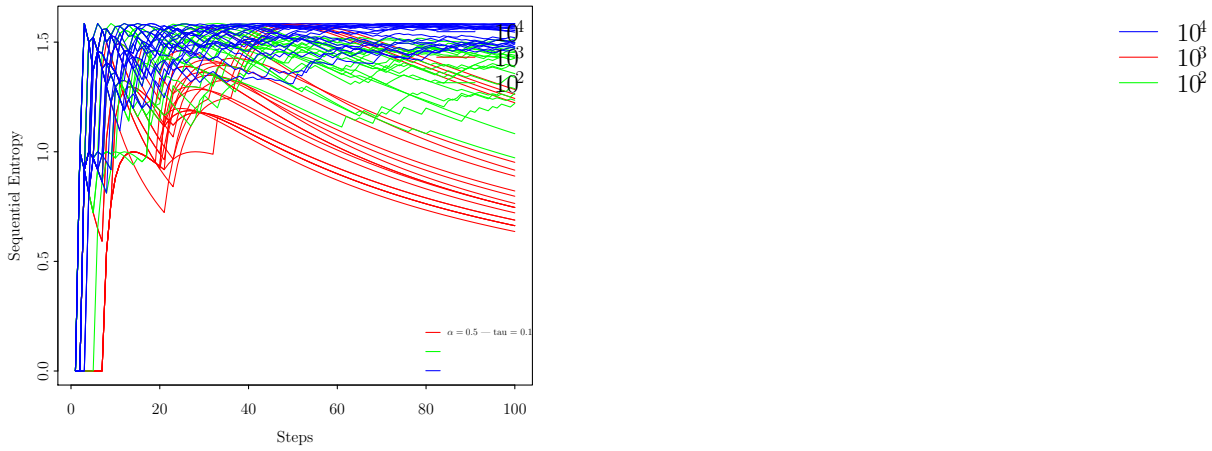


Figure 2: ACF and PACF of MCSI

The following table shows a snippet of our simmulation results. We have a broad range o similarity measures. Some of them are task specific meaning being designed for either categorical or time series data or both.[3]

[3]For example Levenstein distance (or edit distance) is a basic way to measure similarity between categorical sequences [**?**, page 1]Richter or [Gabadinho et al., 2009, page 2], however some authors stating that it might perform poorly in their task [Ren et al., 2011, page 3] or a poor measure at all [Morzy et al., , page 5]

We also considered some specilised articles. [Garcia-Magarinos and Vilar, 2014] provide an algorithm to cluster categorical time series.

Specification	μ	σ	CL Size	SD	Decision	α	τ	N	ALG	TRNS	MI	NMI	AMI	CS	HS	VMS
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		

Table 3: Overview Outcome

4 Data Analysis

Our simulation result suggested that we can cluster the decision behavior under some give constraints. We apply our approach to different real data sets

4.1 Prison data and approach

With acknowledgment to [HERE WE ADD THE GUY] we were provided with experimental data from [Yechiam et al., 2008]. We try to apply our methods to cluster different groups in the data. The participants had to perform the Iowa gambling task (detailed description see later on). [Steingroever et al., 2015] assembled a collection of data of healthy individuals performing the IGT test from various sources. We extended the data set we got with a set of healthy people, which we hope will behave significantly different than the individuals with criminal profile.

4.1.1 Data summary and experimental design

In particular we have data of 96 individuals with different criminal profile. Table 4 gives a broad summary of some demographics.

Crime	**Count**	**Age**	**TABE Score**	**Education**	**Beta IQ**
Theft/Burglary	22	25.36 (7.03)	11.09 (1.29)	7.38 (3.34)	92.91 (14.37)
Robbery	6	24.17 (9.83)	11.00 (0.63)	9.22 (3.30)	96.50 (7.58)
Sex	17	33.41 (13.59)	10.97 (1.47)	9.15 (2.98)	99.65 (11.74)
Drug	22	30.91 (10.11)	11.64 (1.85)	9.06 (2.70)	100.36 (12.92)
OWI	4	38.75 (7.27)	10.88 (1.93)	7.12 (1.17)	94.25 (10.40)
Assault	10	27.20 (8.77)	12.30 (2.41)	7.62 (2.28)	94.50 (11.29)
Escape/ Failure To Appear	4	2.008 (5.60)	11.00 (1.35)	7.78 (3.21)	96.50 (14.18)
Vandalism	1	18.00 (NA)	11.00 (NA)	9.40 (NA)	90.00 (NA)
Forgery	7	34.57 (13.14)	10.93 (5.15)	9.83 (3.82)	100.71 (11.01)
Probabiton	1	38.00 (NA)	12.00 (NA)	6.30 (NA)	92.00 (NA)
Other	2	35.00 (9.90)	11.50 (0.00)	9.20 (4.67)	95.00 (5.66)

Table 4: Summary prision data (means with standard deviation in parenthesis)

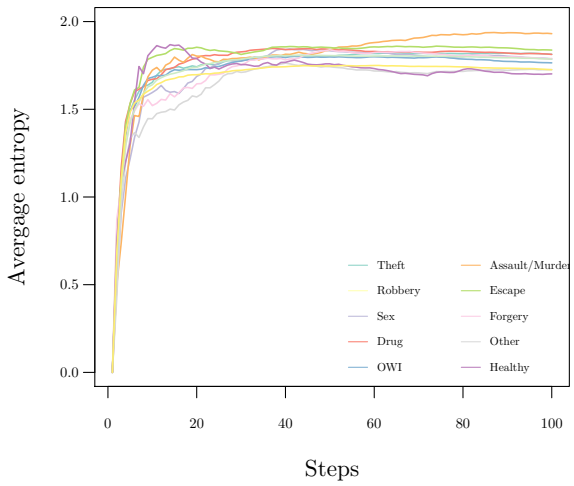
Propositi had to perform the *Iowa gambling tast* (IGT), where they have to pick sequentially a card from four different decks. Two decks have distributions with negative

expectations while two have positive. However, they different according to their variance.
Figure 3 (a) shows the average cumulative entropy averaged across groups. We observe a random behavior independent from group affiliation. However, normal people show the least random behavior. Furthermore, Figure 3 (b) shows the average people of negative choices people chose over steps of ten periods averaged across groups.

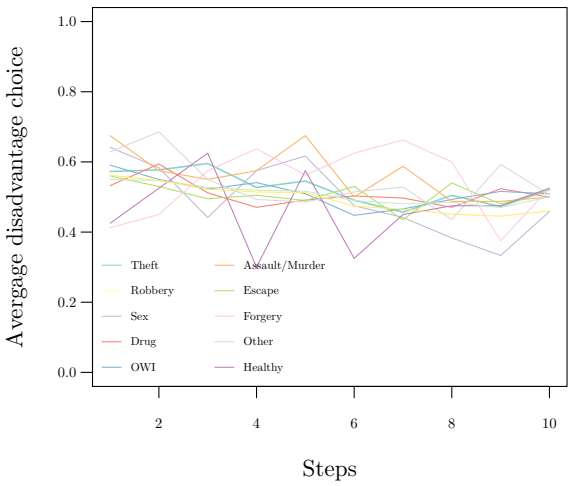
4.1.2 Prison results

Following the last section we provide a summary of our results.

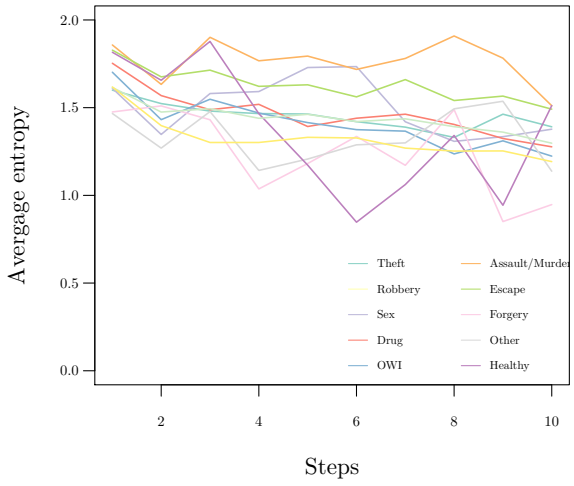
5 Conclusion



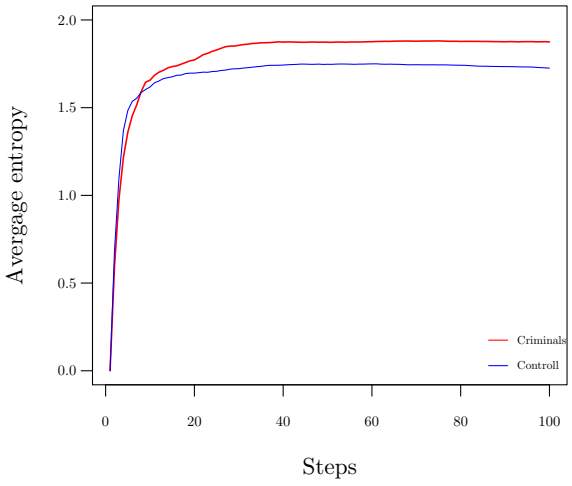
(a) Average Entropy across groups



(b) Blockwise picks from disadvantageous deck



(c) Average Entropy across groups



(d) Blockwise picks from disadvantageous deck

Figure 3: Average py and disadvantageous behavior

6 List of Literature

[Ahn et al., 2008] Ahn, W.-K., Busemeyer, J. R., Wagenmakers, E.-J., and Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, 32(8):1376–1402.

[Bechara et al., 1997] Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1997). Deciding Advantageously Before Knowing the Advantageous Strategy. *Science*, 275:1293–1295.

[Blair et al., 2001] Blair, R. J., Colledge, E., and Mitchell, D. G. V. (2001). Somatic markers and response reversal: is there orbitofrontal cortex dysfunction in boys with psychopathic tendencies? *Journal of Abnormal Child Psychology*, 29(6):499–511.

[Boriah et al., 2008] Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. *SIAM*.

[Brambilla et al., 2012] Brambilla, P., Perlini, C., Bellani, M., Tomelleri, L., Ferro, a., Cerruti, S., Marinelli, V., Rambaldelli, G., Christodoulou, T., Jogia, J., Dima, D., Tansella, M., Balestrieri, M., and Frangou, S. (2012). Increased salience of gains versus decreased associative learning differentiate bipolar disorder from schizophrenia during incentive decision making. *Psychological Medicine*, pages 1–10.

[Brusco and Köhn, 2008] Brusco, M. J. and Köhn, H.-F. (2008). Clustering by passing messages between data points. *Science (New York, N.Y.)*, 319(5864):726; author reply 726.

[Busemeyer and Stout, 2002] Busemeyer, J. R. and Stout, J. C. (2002). L8 A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara gambling task. *Psychological assessment*, 14(3):253–262.

[Cavedini et al., 2002] Cavedini, P., Riboldi, G., Keller, R., D'Annucci, A., and Bellodi, L. (2002). Frontal lobe dysfunction in pathological gambling patients. *Biological Psychiatry*, 51(4):334–341.

[Fratev et al., 1979] Fratev, F., Polansky, O. E., Mehlhorn, A., and Monev, V. (1979). Application of distance and similarity measures. The comparison of molecular electronic structures in arbitrary electronic states. *Journal of Molecular Structure*, 56(C):245–253.

[Fridberg et al., 2010] Fridberg, D. J., Queller, S., Ahn, W. Y., Kim, W., Bishara, A. J., Busemeyer, J. R., Porrino, L., and Stout, J. C. (2010). Cognitive mechanisms

underlying risky decision-making in chronic cannabis users. *Journal of Mathematical Psychology*, 54(1):28–38.

[Gabadinho et al., 2009] Gabadinho, A., Ritschard, G., Studer, M., and Nicolas, S. M. (2009). SUMMARIZING SETS OF CATEGORICAL SEQUENCES Selecting and visualizing representative sequences. *Methods*, pages 6–8.

[Garcia-Magarinos and Vilar, 2014] Garcia-Magarinos, M. and Vilar, J. A. (2014). *A framework for dissimilarity-based partitioning clustering of categorical time series*, volume 29.

[Gullo and Stieger, 2011] Gullo, M. J. and Stieger, A. A. (2011). Anticipatory stress restores decision-making deficits in heavy drinkers by increasing sensitivity to losses. *Drug and Alcohol Dependence*, 117(2-3):204–210.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

[Johnson et al., 2006] Johnson, S. a., Yechiam, E., Murphy, R. R., Queller, S., and Stout, J. C. (2006). Motivational processes and autonomic responsivity in Asperger's disorder: evidence from the Iowa Gambling Task. *Journal of the International Neuropsychological Society*, 12:668–676.

[Luxburg, 2007] Luxburg, U. V. (2007). A Tutorial on Spectral Clustering. (March):1–32.

[MacKay, 2005] MacKay, D. J. C. (2005). *Information Theory, Inference, and Learning Algorithms David J.C. MacKay*, volume 100.

[Martino et al., 2007] Martino, D. J., Bucay, D., Butman, J. T., and Allegri, R. F. (2007). Neuropsychological frontal impairments and negative symptoms in schizophrenia. *Psychiatry Research*, 152(2-3):121–128.

[Morzy et al.,] Morzy, T., Wojciechowski, M., and Zakrzewicz, M. Clustering sequences of categorical values.

[Murphy, 2012] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

[Nirit Agaya, Eldad Yechiama, Ziv Carmelb, 2010] Nirit Agaya, Eldad Yechiama, Ziv Carmelb, Y. L. (2010). Non-specific effects of Methylphenidate (Ritalin) on cognitive ability and decision- making of ADHD and healthy adults. (972).

[Pamminger, 2007] Pamminger, C. (2007). Bayesian Clustering of Categorical Time

Series Using Finite Mixtures of Markov Chain Models Bayesian Clustering of Categorical Time Series Using Finite Mixtures of Markov Chain Models.

[Pamminger and Fr??hwirth-Schnattery, 2010] Pamminger, C. and Fr??hwirth-Schnattery, S. (2010). Model-based clustering of categorical time series. *Bayesian Analysis*, 5(2):345–368.

[Pamminger and Fruhwirth-Schnatter, 2009] Pamminger, C. and Fruhwirth-Schnatter, S. (2009). Bayesian Clustering of Categorical Time Series Using Finite Mixtures of Markov Chain Models. 43(0907).

[Ren et al., 2011] Ren, J., Cao, S., and Hu, C. (2011). A Hierarchical Clustering Algorithm Based on Dynamic Programming for Categorical Sequences. 5:1575–1581.

[Rosenberg and Hirschberg, 2007] Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1(June):410–420.

[Serra and Arcos, 2014] Serra, J. and Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, 67:305–314.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(July 1928):379–423.

[Shirali and Vasudeva, 2005] Shirali, S. and Vasudeva, H. L. (2005). *Metric Spaces*. Springer.

[Shlens, 2014] Shlens, J. (2014). A Tutorial on Principal Component Analysis. *ArXiv*, pages 1–13.

[Steingroever et al., 2015] Steingroever, H., Fridberg, D. J., Horstmann, A., Kjome, K. L., Kumari, V., Lane, S. D., Maia, T. V., Mcclelland, J. L., Pachur, T., Premkumar, P., Stout, J. C., and Wetzels, R. (2015). Data from 617 Healthy Participants Performing the Iowa Gambling Task : A "Many Labs" Collaboration. *Journal of Open Psychology Data*.

[Steingroever et al., 2013] Steingroever, H., Wetzels, R., and Wagenmakers, E.-J. (2013). Absolute Performance of Reinforcement-Learning Models for the Iowa Gambling Task. *Decision*.

[Stout et al., 2004] Stout, J. C., Busemeyer, J. R., Lin, A., Grant, S. J., and Bonson,

K. R. (2004). Cognitive modeling analysis of decision-making processes in cocaine abusers. *Psychonomic Bulletin & Review*, 11(4):742–747.

[Sutton and Barto, 2012] Sutton, R. and Barto, A. (2012). Reinforcement Learning : An Introduction.

[Vinh et al., 2010] Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854.

[Wang et al., 2013] Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309.

[Yechiam et al., 2008] Yechiam, E., Kanz, J. E., Bechara, A., Stout, J. C., Busemeyer, J. R., Altmeier, E. M., and Paulsen, J. S. (2008). Neurocognitive deficits related to poor decision making in people behind bars. *Psychonomic Bulletin & Review*, 15(1):44–51.

Appendix

Metrics and Similarities

This part of the appendix formally defines metrics and similarities and dissimilarities (proximity) used in this paper. We first define some basic general concepts followed by a description of the applied distance and similarity concepts.

Distances vs. Similarities

Let $\mathcal{X}$ be a dataset and let $\boldsymbol{x_i}, \boldsymbol{x_j}$ be two datapoints, such that $\boldsymbol{x_i}, \boldsymbol{x_j} \in \mathcal{X}$.

A distance function assign for pairs a points a non negative real number as distance. $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_0^+$. Formally if the following properties are additionally staisfied the distance is also metric [Shirali and Vasudeva, 2005, page 28].

1. $d(\boldsymbol{x_i}, \boldsymbol{x_j}) \geq 0$
2. $d(\boldsymbol{x_i}, \boldsymbol{x_i}) \geq 0$
3. $d(\boldsymbol{x_i}, \boldsymbol{x_j}) = d(\boldsymbol{x_j}, \boldsymbol{x_i})$
4. $d(\boldsymbol{x_i}, \boldsymbol{x_j}) \leq d(\boldsymbol{x_i}, \boldsymbol{x_j}) + d(\boldsymbol{x_j}, \boldsymbol{x_i})$

A distance can be seen as a measure for dissimilarity of two points. Besides distance some algorithms operate on a *similarity* matrix. Formally a similarity is a function $S : \mathcal{X} \times \mathcal{X} \mapsto [0, 1]$. Also for similarity we can define the following properties [Fratev et al., 1979, page 3]:

1. $0 \leq S(\boldsymbol{x_i}, \boldsymbol{x_j}) \leq 1, \text{for } i \neq j$
2. $S(\boldsymbol{x_i}, \boldsymbol{x_i}) = 1$
3. $S(\boldsymbol{x_i}, \boldsymbol{x_j}) = S(\boldsymbol{x_j}, \boldsymbol{x_i})$

Once we have computed distance or a similarity we can compute for two data points we can use this information to transform it to a similarity or the distance vice versa [Boriah et al., 2008, page 4]:

$$S(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{1}{1 + d(\boldsymbol{x_i}, \boldsymbol{x_j})} \quad \Leftrightarrow \quad d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{1}{S(\boldsymbol{x_i}, \boldsymbol{x_j})} - 1 \tag{7}$$

.0.1 Similarity measures for time series data

We use three different similarity measures for time series. E.g. [Wang et al., 2013] or [Serra and Arcos, 2014] provide an overview and empirical evaluation on common similarity measures for time series. The following definitions are taken from the latter. Empirical research suggest that simple euclidean distance for time series performs quite well and is hard to beat. Hence, the first distance measure for time series is simply euclidean distance between two time series. They might be converted to similarity based on equation 7. Let $\boldsymbol{x}, \boldsymbol{y}$ be two time series over N-periods. Then the L_2 distance between two time series is define as:

$$d_E(\boldsymbol{x}, \boldsymbol{y}) := \sqrt{\left(\sum_{i=1}^{N} (x_i - y_i)^2 \right)} \tag{8}$$

.0.2 Similarity measures for categorical data

Reffering to equation 8 we define a simple measure for the categorical aspect of the data. Note that [SOURCE] defines the simplest measure for categorical data. However, since the probability that two people behave exactly the same in our context is arguably zero we modify this concept slightly. So we relax that and define the overlap similarity just as the count of overlapping instances. This serves as a benchmark similarity for categorical data.

In [Boriah 2008] we find a rich class of further categorical measures. To keep it concise we considered two of them. The first one is *Eskin* similarity measures.

$$d_O(\boldsymbol{x}, \boldsymbol{y}) := \begin{cases} 1 & ,\text{if } \boldsymbol{x} = \boldsymbol{y} \\ \frac{n_k^2}{n_k^2+2} & \text{otherwise} \end{cases} \tag{9}$$

Furthermore, we consider lins similarity

$$d_{lin}(\boldsymbol{x}, \boldsymbol{y}) := \begin{cases} 2\log \hat{p}_k(X_k) & ,\text{if } \boldsymbol{x} = \boldsymbol{y} \\ 2\log\left(\hat{p}_k(x_k) + \hat{p}_k(y_k)\right) & \text{otherwise} \end{cases} \tag{10}$$

Clustering Evaluation

In this section we formally derive and explain the applied clustering metrics. Evaluating clustering performance has some issues. The algorithm is producing labels. Nevertheless we generated the "true" labels the might not be comparable. A simple example we might consider the following situation. Let y denote the labels of data the data and y' the corresponding prediction such that $y, y' \in \{0, 1\}$. In a small example let our data points be like $y = (1, 1, 0, 0)$ and the corresponding prediction $y' = (0, 0, 1, 1)$. Obviously the clustering worked perfectly, however comparing "labels" would produce an accuracy of zero.

There a several clustering metrics, which respect such a situation. We consider a bunch of information based metrics. Most of the measures use some sort of entropy. The following concepts can be found in [Rosenberg and Hirschberg, 2007] and [Vinh et al., 2010] and as additional reading [Hubert and Arabie, 1985].

First we might introduce the contingency table.

	V_1	V_2	$\dots$	V_c	Σ
U_1	$n_{1,1}$	$n_{1,2}$	$\dots$		a_1
U_2	$n_{2,1}$	$\ddots$			a_1
$\vdots$	$n_{1,1}$	$\dots$			a_1
U_R	$n_{1,1}$	$\dots$			a_1
	b_1	b_2	$\dots$	b_c	N

Table 5: Contigency Table

N_{11}**:** Number of pairs in the same cluster

N_{00}**:** Number of pairs that are in different clusters in both v and u

N_{01}**:** Number of pairs that are in the same cluster in both u but different in v

N_{10}**:** Number of pairs that are in the same cluster in both v but different in u

$$RI(u,v) = \frac{N_{00} + N_{11}}{\binom{N}{2}} \tag{11}$$

$$ARI(u,v) = \frac{2\left(N_{00}N_{11} - N_{01}N_{10}\right)}{\left(N_{00} + N_{01}\right)\left(N_{01} + N_{11}\right) + \left(N_{00} + N_{10}\right)\left(N_{10} + N_{11}\right)} \tag{12}$$

$$H(u) = -\sum_{i=1}^{R} \frac{a_i}{N} \log \frac{a_i}{N} \tag{13}$$

$$H(v) = -\sum_{i=1}^{C} \frac{b_i}{N} \log \frac{a_i}{N} \tag{14}$$

$$H(u,v) = -\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{i,j}}{N} \log \frac{n_{i,j}}{N} \tag{15}$$

$$H(u|v) = -\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{i,j}}{N} \log \frac{n_{i,j}/N}{b_j/N} \tag{16}$$

$$H(v|u) = -\sum_{i=1}^{C}\sum_{j=1}^{R} \frac{n_{i,j}}{N} \log \frac{n_{i,j}/N}{b_j/N} \tag{17}$$

$$I(u,v) = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{i,j}}{N} \log \frac{n_{i,j}/N}{a_i b_j/N} \tag{18}$$

$$\tag{19}$$

Normalized Info Score:

This is one example of a normalized version

$$NMI_{max}(u,v) = \frac{I(u,v)}{\max\left(H(u), H(v)\right)} \tag{20}$$

$$\begin{aligned} AMI_{max}(u,v) &= \frac{NMI_{max}(u,v) - \mathbb{E}\left[NMI_{max}(u,v)\right]}{1 - \mathbb{E}\left[NMI_{max}(u,v)\right]} \\ &= \frac{I(u,v) - \mathbb{E}\left[I(u,v)\right]}{\max\left(H(u), H(v)\right) - \mathbb{E}\left[I(u,v)\right]} \end{aligned} \tag{21}$$

$$\mathbb{E}\left[I(u,v)\right] = \sum_{i=1}^{R}\sum_{j=1}^{C} \sum_{n_{i,j}=\max(a_i+b_j-N,0)}^{\min(a,b)} \frac{n_{ij}}{N} \log\left(\frac{N n_{ij}}{a_i b_j}\right) \frac{a_i! b_j! (N-a_i)! (N-b_j)!}{N! n_{ij}! (a_i - n_{ij})! (b_j - n_{ij})! (N - a_i - b_j + n_{ij})!} \tag{22}$$

Homogeneity:

$$h = \begin{cases} 1 & \text{,if } H(u,v) = 0 \\ 1 - \frac{H(u|v)}{H(u)} & \text{otherwise} \end{cases} \tag{23}$$

Completeness:

$$c = \begin{cases} 1 & \text{,if } H(v,u) = 0 \\ 1 - \frac{H(v|u)}{H(v)} & \text{otherwise} \end{cases} \tag{24}$$

V Measure Score:

$$V_\beta = \frac{(1+\beta)hc}{\beta h + c} \tag{25}$$

.1 Algorithms

Principal Components and Multidimensional Scaling

Coming to principal component analysis. [Shlens, 2014] provides an excellent contribution considering both practical applications and theoretical background. The following derivations are from this source.

Spectral Clustering

For the spectral clustering algorithm we formally introduce some graph notation. If not stated otherwise the following derivation follows [Luxburg, 2007]. In the following we consider a weighted and simple undirected graph.

$$G = \{V, E\} \tag{26}$$
$$V = \{v_1, \ldots, v_n\} \tag{27}$$
$$E = \{e_1, \ldots, e_n\} \tag{28}$$

Furthermore let the graph has a weighted and symetric ($|V| \times |V|$) adjacency matrix, such that:

$$\boldsymbol{W} = \begin{cases} w_{i,j} & \text{,if } v_i v_j \in E \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

The *degree* of a node is defined as the sum of edge weights of connected nodes. Formally we denote the degree of node i as:

$$d_i := \sum_{j=1}^{n} w_{ij} = \sum_{i=1}^{n} w_{ij} \tag{30}$$

Using the last expression we define matrix $\boldsymbol{D}$ as the diagonal matrix of the degress

$$\boldsymbol{D} := diag(\boldsymbol{d}) \tag{31}$$

The algorithm works on the *Laplacian* matrix defined by:

$$\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{W} \tag{32}$$

Former versions of the algorithm are applied on the graph laplcian. However, there were proposed newer versions using the so called *normalized laplacian*. Since also the python version is using this package we will focus on this version of the algorithm. Following that the normalized graph laplacian is defined as:

$$\boldsymbol{L}_{norm} := \boldsymbol{D}^{1/2} \boldsymbol{L} \boldsymbol{D}^{1/2} = \boldsymbol{I} - \boldsymbol{D}^{1/2} \boldsymbol{W} \boldsymbol{D}^{1/2} \tag{33}$$

A Affinity Propagation

[Brusco and Köhn, 2008]