SUPPLEMENTARY MATERIAL

Model-based feature construction for multivariate decoding

K.H. Brodersen, F. Haiss, C.S. Ong, F. Jung, M. Tittgemeyer, J.M. Buhmann, B. Weber, K.E. Stephan

S1 Dataset 1 – experimental methods

Surgical preparation and anaesthesia

Experiments were performed in 3 adult male Sprague-Dawley rats weighing 250 g each. The animals were kept in cages in a ventilated cabinet with standardized conditions of temperature and light (night/day-cycle 12h/12h). Free access to food and water was ensured at all times.

Surgical procedures and measurements were performed under isoflurane anaesthesia (2.5-3.5% during surgery and 1-1.5% during data acquisition). Surgery involved the cannulation of the right femoral artery and vein with PE-50 tubing containing saline, as well as a tracheotomy for artificial ventilation of the animal. The arterial catheter was used for the continuous monitoring of the arterial blood pressure and for withdrawal of blood for blood-gas analysis. After fixating an animal's head in a stereotactic frame (Kopf Instruments, Tujunga, CA, USA), bucain injections were administered subcutaneously prior to the scalp incision. The skull above the barrel cortex (1 mm caudal and 3 mm lateral from Bregma) was exposed after a midline incision and after disconnecting the temporal muscle from the skull. Using a dental drill (Bien Air Medical Technologies, Bienne, Switzerland) a craniotomy with a diameter of 4 mm was carried out above barrel cortex, after which the dura was carefully removed. Using a heating blanket, body temperature was kept at 37 °C. Blood gases were maintained within normal ranges by adjusting the ventilation parameters. Upon the completion of data acquisition, animals were euthanized

with a bolus of intravenous pentobarbital (200 mg/kg). All experimental procedures were approved by the veterinary authorities of the Canton of Zurich.

Stimulation and recording

Local field potentials (LFPs) were recorded using multielectrode silicon probes (NeuroNexus Technologies, Ann Arbor, MI, USA). One shank with 16 electrodes (impedance approx. 1 MΩ, spacing 100 µm) was gently inserted into barrel cortex by 1700 µm. Recordings were performed using a multichannel extracellular amplifier (MultiChannelSystems, Reutlingen, Germany; gain x5000, sampling frequency 20 kHz, band pass 1-5000 Hz). Voltage traces were band-pass filtered offline with digital filters (1-200 Hz) to uncover LFP signals.

Experimental stimuli were presented using a glass capillary (length 5 mm) mounted to the tip of a piezo-bending actuator (Q220-A4-303YB, Piezo Systems, Woburn, MA, USA). The actuator was fixed on an articulated arm (Baitella, Zurich, Switzerland) to allow for accurate positioning of the stimulator. Movements of the bending actuator were calibrated using an optical Laser Micrometer (RX 03, Metralight, San Mateo, CA, USA). Two whiskers (dataset A1: whiskers E_1 and D_3; dataset A2: whiskers C_1 and C_3; dataset A3: whiskers D_3 and β) were stimulated independently using two piezo-bending actuators that produced brisk rostral to caudal deflections. Stimuli involved a single cosine wave (frequency 120 Hz, amplitude approx. 500 µm). Each whisker was stimulated 300 times, in randomized order, leading to 600 sweeps of a duration of 2 s each. Electrophysiological recordings were started 100 ms prior to stimulation onsets. Inter-trial intervals were randomly jittered using a uniform distribution between 2200 and 2750 ms. An experimental control (dataset A4) was recorded by following precisely the same procedure, except that whisker stimulators were repositioned to be as close to the original whiskers (D_3 and β) as possible without physically touching them.

S2 Dataset 2 – experimental methods

Surgical preparation and implant

In order to record event-related responses in the awake, unrestrained animal, a telemetric recording system (TSE Systems) was set up using chronically implanted epidural silverball electrodes above the left auditory cortex in 3 Lister hooded rats (cf. Jung et al., 2009). Prior to surgery, rats were placed in an exsiccator that was perfused with isoflurane (5%) mixed with 30% oxygen (O_2) and 70% nitrous oxide (N_2O). Once deeply anaesthetized, rats were transferred into a stereotactic frame and fixated using ear bars and a tooth bar. During surgery, animals were constantly inhaling a similar mixture of gases through a mask (isoflurane reduced to 2-3%). Using a heating pad, feedback-regulated by means of a rectal probe, body temperature was kept constantly at 37.5 °C. Guided by stereotaxic coordinates (Paxinos & Watson, 2007), two electrodes were positioned 5 mm posterior to Bregma and 7 mm (electrode 1) and 8 mm (electrode 2) lateral from the sagittal suture (depth 4 mm), targeting the primary and secondary auditory cortex, respectively (Doron, Ledoux, & Semple, 2002). A reference electrode was placed above the frontal sinus. The telemetry socket, to which electrodes were soldered, was fixed onto the head with dental cement. All experimental procedures were approved by the local governmental and veterinary authorities.

Stimulation and recording

Recordings began one week after surgery. At the beginning of each experiment, in order to allow for wireless data transfer, an EEG telemetry transmitter was attached to the implanted socket. Rats were anaesthetized briefly for this procedure. During the period of data acquisition, rats

were awake and placed in a cage (21 x 35 x 22 cm^3) that ensured a reasonably constrained variance in the distance between the animal and the speakers (± 25 cm).

All recordings were carried out in a sound-attenuated chamber. Stimuli consisted of bandpass-filtered noise of different carrier frequencies (B1: standards 5-7 Hz, deviants 15-17 Hz; B2: standards 15-17 Hz, deviants 5-7 Hz; B3: standards 10-12 Hz, deviants 16-18 Hz). Each stimulus had a length of 50 ms, including a 5 ms ramp on either end, as depicted in Figure 7b. Initially, stimuli were presented in simple, homogeneous sequences. Subsequently, those two stimuli were chosen that evoked the highest amplitudes in the recorded signal. Standard and deviant stimuli were then presented pseudo-randomly with different deviant probabilities (B1: 0.1; B2: 0.2; B3: 0.1). The recording window covered 90 ms before and 300 ms after the stimulus onset, leading to a total sweep length of 390 ms. The inter-trial interval was 210 ms. The three datasets comprised 900, 500, and 900 trials, respectively.

S3 Additional information on analysis methods

DCM specification for dataset 1

In the context of model-based decoding of the first dataset, a single-region dynamic causal model for ERP data was specified and inverted using SPM8. Neural priors were chosen according to SEP settings. The neural model was an LFP model with 1 region. Given a true stimulus onset at 100 ms after the beginning of a sweep, we specified a time window of [90, 390] ms and an onset of 105 ms. (Note that all times were converted to peristimulus times in the main text by shifting them so that the stimulus onset occurred at 0 ms.) Further settings included: detrend 1; subsample 1. The model was fitted individually to each trial.

DCM specification for dataset 2

For the second dataset, given that it comprised 2 recording sites, 3 alternative models for ERP data were specified: (i) a model with forward connections from region 1 to region 2, backward connections from region 2 to region 1, and stimulus input arriving in region 1; (ii) a model with forward connections from region 2 to region 1, backward connections from region 1 to region 2, and stimulus input arriving in region 2; (iii) a model with lateral connections between the two regions and stimulus input arriving in both region 1 and region 2. In all models, neural priors were chosen according to SEP settings. Given a true stimulus onset at 90 ms after the beginning of a sweep, we specified a time window of [80, 400] ms and an onset of 100 ms. (Again, all times were converted to peristimulus times in the main text.) Further settings included: detrend 1; subsample 1. Using the first half of the data only, we assessed which model architecture yielded the highest model-based classification accuracy. We then applied this model to the second half of the data and reported the resulting accuracies.

Classification

All classification analyses were based on a cross-validation scheme that was tailored to the characteristics of the datasets at hand.

Dataset 1 (Section 3.1) comprised 600 trials per experiment (100 trials in the control condition). Overall conventional and model-based classification analyses were based on leave-20-out cross-validation, i.e., 30 folds in the experimental datasets and 5 folds in the control (Figure 5). For the temporal analyses, carried out separately for each time bin, we used a computationally less expensive scheme by randomly splitting the data into 580 trials for training and 20 trials for testing, repeating the process 5 times (Figure 4).

Dataset 2 (Section 3.2) contained 900, 500, and 900 trials in experiments B1, B2, and B3, respectively. Here, due to the larger number of examples, overall classification analyses were based on a randomized cross-validation scheme throughout, training on all but 20 examples and repeating the process 20 times (Figure 9). For the temporal analyses, carried out separately for each time bin, we randomly split the data into 890 trials for training (B2: 490 trials) and 10 trials for testing and repeated the process 30 times (Figure 8).

Prior to classification, all examples were normalized (i.e., their norm was set to unity). In other words, they were represented as points on a d-dimensional sphere of radius $r = 1$, where d is the number of features.

In the case of ordinary (non-random) cross-validation, in order to avoid optimistic accuracy estimates that may result from temporal autocorrelation in the signal, we removed from each training set the two trials surrounding the test set. In addition, in order to prevent the learning algorithm from acquiring a strong bias towards one class (e.g., towards standard tones as

opposed to deviants), we balanced the training set within each cross-validation fold by removing surplus trials until both classes were of the same size.

During the training phase of the support vector machine, we optimized the regularization parameter C by a simple linear search using inner 5-fold cross-validation on the training set. In the case of a nonlinear kernel, we carried out a grid search in $\log_2$ space instead to find a combination of kernel parameters that minimized the error rate on the inner test set. We then used these optimal parameters to train the classifier on the current fold-wise training set and make predictions on the corresponding test set. This nested procedure ensured that information from the test set was neither used when training the classifier nor when finding optimal parameters.

All analyses were implemented in MATLAB 2009a to run in a parallelized fashion on a compute cluster at ETH Zurich using Platform LSF (http://www.platform.com/grids/platform-lsf). Some portions of the analysis used additional code from SPM8 (http://www.fil.ion.ucl.ac.uk/spm/), the Princeton MVPA toolbox v1.0 (http://www.csbmb.princeton.edu/mvpa/), and the LIBSVM library v2.9.1 (Chang & Lin, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm/).

S4 Sensitivity comparison between model-based decoding and conventional DCM analyses

As described in the main text, we propose model-based decoding as a complementary approach to established Bayesian model selection (BMS) in situations where log-evidence based approaches are not applicable. However, as suggested by one of our reviewers, it might also be worth investigating whether model-based decoding offers higher or lower sensitivity than log-evidence based approaches in situations where both could be used. Specifically, one could compare p-values obtained from model-based decoding to (equivalents of) p-values derived from Bayes factors in the context of conventional DCM and BMS. In the DCM analysis, one would model the

differences in class means in terms of changes in specified parameters, and then compare this model to a null model in which no changes in parameters (and thus no differences between class means) are allowed. Here, the equivalent of a p-value can be derived from the posterior model probabilities (i.e., one minus the conditional probability that the alternate model was better than the null model).

Such a comparison is feasible but must be qualified carefully since the two approaches differ in several aspects. BMS-based p-values are the result of a fitting procedure that uses all available data, while classification operates on a strongly reduced feature space. Thus, one might generally expect model-based classification to be less sensitive than evidence-based model comparison. On the other hand, in the case of current DCM implementations for evoked responses, only a few parameters are allowed to change for explaining differences in observed responses (i.e., extrinsic connections strengths and the amplitude of excitatory postsynaptic potentials), whereas classification in a model-based feature space may utilize *all* parameters for identifying differences between trial types. In addition, a nonlinear classifier may allow for trial-type separation when no significant difference is revealed by class means alone. These considerations imply that the relative sensitivity of DCM/BMS vs. model-based classification may vary depending on the particular data set and model in question.

Indeed, when carrying out the comparison on our two datasets, as described below, we obtained mixed results (see Table S4). For the first (somatosensory) dataset, we found decoding-based p-values to be smaller than the p-values derived from the log Bayes factor in the conventional DCM analysis in two out of three cases, and both values were indistinguishable from zero in one case. In contrast, for the second (mismatch negativity) dataset, we found that in all three animals DCM-based p-values were smaller than the p-values provided by our model-based approach.

In summary, the relative sensitivity of DCM/BMS and model-based decoding for establishing differences between trial types (or subject classes) is difficult to determine in full generality, but is likely depend on the data observed and the particular model used. Our results described here are thus of an anecdotal nature and should not be overly generalized.

Animal	Bayesian model comparison (BMS)		Model-based decoding	Comment
A1	0.9445	>	0	decoding more sensitive
A2	0.5002	>	0	decoding more sensitive
A3	0	≈	0	indistinguishable
A4*	0.2193	<	0.589	decoding more specific
B1	0	<	0.0113	BMS more sensitive
B2	0	<	0.0023	BMS more sensitive
B3	0.5046	<	0.9585	BMS more sensitive

Table S4 – Comparison of p-values

* Note that A4 is a control dataset where no stimuli where applied and where thus no difference should be detected.

S5 Sensitivity comparison between model-based decoding and Hotelling's T^2-test

Since model-based feature construction greatly reduces the dimensionality of the feature space, one may ask whether the two trial types can be discriminated without invoking a cross-validation scheme and using a conventional encoding model instead (see section 'Dimensionality of the parameter space' in the main text). Specifically, we compared the significance of above-chance decoding accuracies to the outcome of Hotelling's T^2-test, the multivariate generalization of Student's t-test. In our context, the null hypothesis states the absence of any difference between class-conditional means of model parameter estimates. In the case of decoding, we computed p-values as the probability of obtaining the observed balanced accuracy under the null hypothesis that the classifier operates at chance. In the case of Hotelling's T^2-test, we computed p-values as

the probability of the T^2 statistic being equal or greater than the observed value under the null hypothesis of the between-condition Mahalanobis distance being zero (see Table S5).

Given that our data represent averages and should conform to parametric assumptions by the central limit theorem, the Neyman-Pearson lemma states that Hotelling's T^2-test should provide the most powerful test. However, it can be only be applied when there are fewer features than examples, which means that the decoding scheme described in the main text has a greater domain of application.

For the first dataset, p-values were numerically indistinguishable from zero in all experimental cases (A1–A3); in the control case where no stimuli were applied (A4) and where no significant p-value is expected, neither method yielded a false positive result. For the second dataset, there was no meaningful difference between decoding-based p-values and Hotelling's p-values in two out of three cases, while only Hotelling's p-value was significant for the third animal. These anecdotal results are consistent with the notion that Hotelling's T^2-test provides the most powerful test when applicable.

Animal	**Hotelling's T^2-test**		**Model-based decoding**	**Comment**
A1	0	≈	0	indistinguishable
A2	0	≈	0	indistinguishable
A3	0	≈	0	indistinguishable
A4*	0.17	<	0.31	decoding more specific
B1	6.8×10^{-6}	≈	3.1×10^{-6}	indistinguishable
B2	4.5×10^{-4}	≈	1.2×10^{-4}	indistinguishable
B3	0.001	<	0.18	Hotelling's more sensitive

Table S5 – Comparison of p-values

* Note that A4 is a control dataset where no stimuli where applied and where thus no difference should be detected.

SUPPLEMENTARY FIGURE LEGENDS

Fig. 11 Evoked responses

Separately for each trial type, the plot shows averaged responses from the channel that was used for model-based decoding of dataset 1 (channel 3). Each row represents one of the four experiments. The left column presents the data on a wide-interval [-100, 800] ms peristimulus time window, while the right column shows the same data with a focus on a shorter time window just after the stimulus. Each response is given as mean ± 2 standard errors of the mean, in μV. While the main recordings (A1–A3) show clear and differential responses to the two types of stimuli, the control recording (A4) is diffuse and does not deviate significantly from its baseline when other traces do (note that the *y*-axes are scaled individually to show the full amplitude of the response).

Fig. 12 Scatter plot of two exemplary informative features

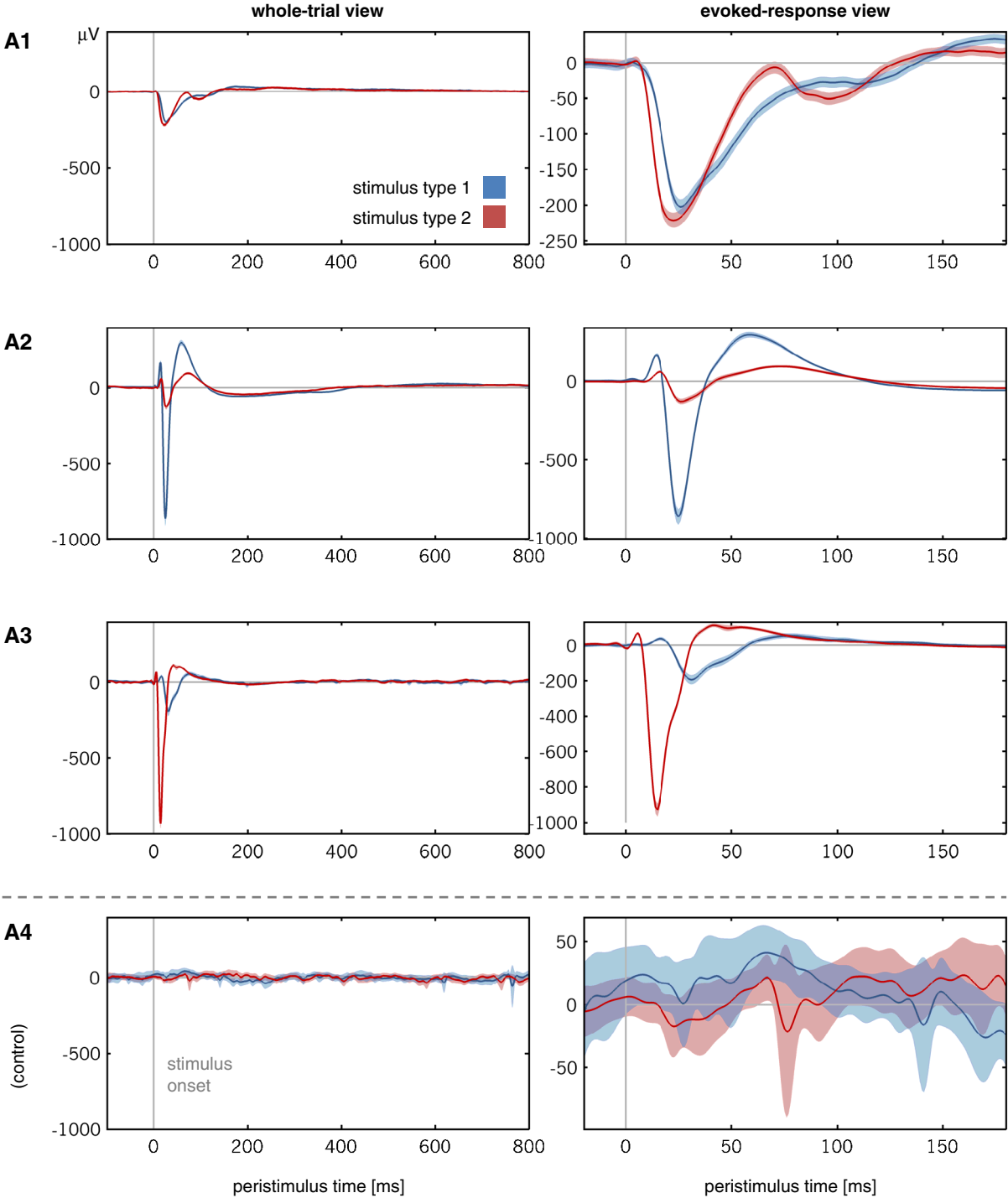
The plot shows the distribution of trials in the two classes (blue and red), separately for each experiment of dataset 1 (i.e., corresponding exactly to the data shown in Fig. 11). Each trial is expressed in terms of its model parameters R_1 and R_2. These two parameters were found to be particularly informative in dataset A2, while only R_1 was of noteable importance in datasets A1 and A3. Taken together, the plots confirm the notion indicated by Figure 6: the higher the feature weight of a particular model parameter, the easier it is to distinguish the two experimental conditions along the corresponding axis. In dataset 3, for example, Figure 6 (rightmost plot) shows that the parameter R_1 (stimulus onset) has the highest discriminative power. Consistent with this, Figure 12 (rightmost plot) shows that a hyperplane orthogonal to the *x*-axis can comfortably separate red and blue points

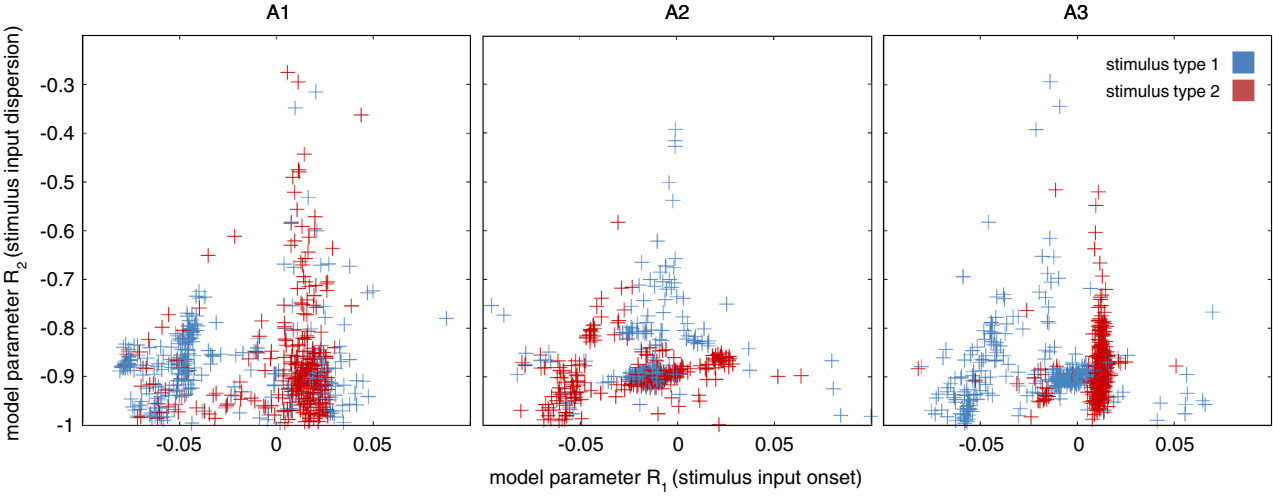
to a reasonable degree of accuracy, whereas a hyperplane orthogonal to the *y*-axis would fail to do so.

SUPPLEMENTARY REFERENCES

Doron, N. N., Ledoux, J. E., & Semple, M. N. (2002). Redefining the tonotopic core of rat auditory cortex: physiological evidence for a posterior field. *The Journal of Comparative Neurology*, *453*(4), 345-360. doi: 10.1002/cne.10412.

Paxinos, G., & Watson, C. (2007). *The rat brain in stereotaxic coordinates*. Academic Press.


whole-trial view
evoked-response view
A1
A2
A3
A4
μV
stimulus type 1
stimulus type 2
stimulus onset
(control)
peristimulus time [ms]
peristimulus time [ms]


A1
A2
A3
stimulus type 1
stimulus type 2
model parameter R2 (stimulus input dispersion)
model parameter R1 (stimulus input onset)
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1
-0.05
0
0.05