Research Methods

Conclusio

Dr. Sven Magg, Prof. Dr. Stefan Wermter

Knowledge Technology

http://www.informatik.uni-hamburg.de/WTM/

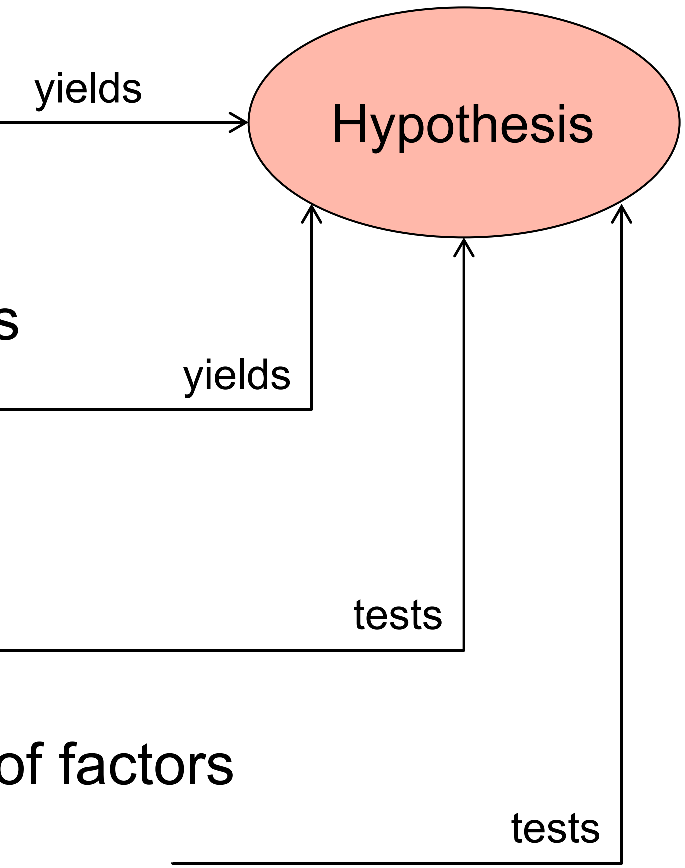
Plan for today!

1. Some loose ends
   a) Observation Experiments
   b) Measurements often used in HRI
   c) Examples for data visualisation
2. Summary
3. Q & A

Types of studies

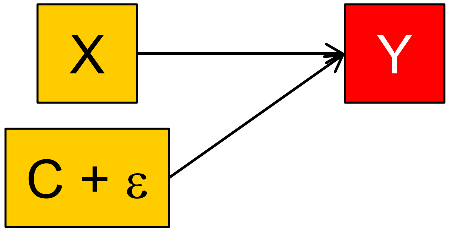
We have a system,….

- Exploratory Study
  - what can/does it do?
  - Yields hypotheses for other studies
- Assessment Study
  - where are its limits and defaults?
- Manipulation Experiment
  - what happens if….?
  - Test hypotheses about influences of factors
- Observation Experiments
  - how correct is my model of what should happen?



Effects

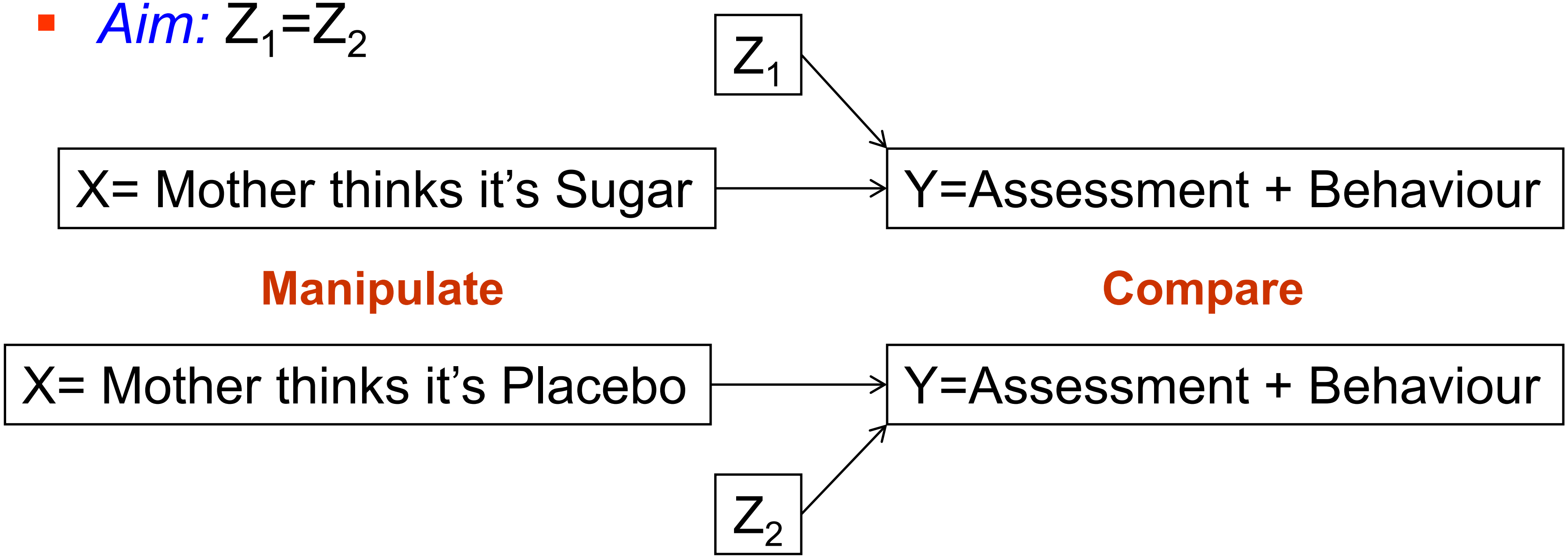
- Experiments are conducted to
  - correctly attribute the cause of a change (or lack of change) in a dependent variable
  - correctly attribute the causes of effects



- If X is a cause of Y, then X should produce and effect on Y
- “Effect” usually the variance in Y explained by X

Manipulation Experiments

- We manipulate X and measure effects on Y for each value of X
- “each value of X” = *Condition*
- *Variability in Y through X and Zs in each condition*
- *Aim:* $Z_1 = Z_2$



Observation Experiments

- What if we can't manipulate X?
  - **Sometimes impossible or not feasible to manipulate X**
    - internal chemistry of ants determines decision to go left or right at crossing
    - composition of comets affects length of vapour trail
  - **Sometimes not ethical to manipulate X**
    - Smoking causes lung cancer
    - Nocebo effect as strong as Placebo effect

Observation Experiments

- In order to test for effects, we need different conditions
- Instead of manipulating X, we can classify by X to create sample groups
- Example:

Tverdal, Aage, et al. "Coffee consumption and death from coronary heart disease in middle aged Norwegian men and women." BMJ: British Medical Journal 300.6724 (1990): 566.

  - *19398 men and 19166 women aged 35-54 years*
  - *Examination with follow up 4-6 years later*
  - *"How many cups of coffee do you usually drink per day?"*
  - *6 groups (<1, 1-2, 3-4, 5-6, 7-8, >8)*
  - *Mortality (deaths per 100,000) reported for these 6 groups*
  - *Results with age, cigarette consumption, cholesterol,… adjusted mortality rate also reported*

Observation Experiments

- **How is this different to exploratory studies?**
  - There is a very thin line between them
  - Sampling usually didn’t take place with grouping in mind
  - Conceptually you have a model beforehand, predict the outcome according to the model and then compare

- **Difference between manipulation and observation**
  - Some argue that observation studies can not prove effect
  - Difficult to detect biases and hidden factors compared to randomised experiments
  - The larger the set of recorded factors, the higher the likelihood that one factor is correlated purely by chance

Training and Performance

- Common procedure in HRI or neural network studies
  - Training of parameters or weights of a system
  - detection or recognition systems, e.g. for face detection, speaker recognition, etc.

- How to measure performance of such systems?
- What are training, test and validation sets?

- Let’s assume the system works and can be trained
  - We are not interested in the details of the system and its purpose for now

Training a system

- Training phase
  - A number of training examples is fed into the system and the system parameters adjusted
  - This is done for a number of *epochs*
    1 epoch ≙ 1 run over the whole training set

- With increasing number of epochs, the output error will decrease
  - The system learns to classify/recognise/etc. the training data

- The training set should be small to increase the speed of training and large enough to be a representative sample

Training a system

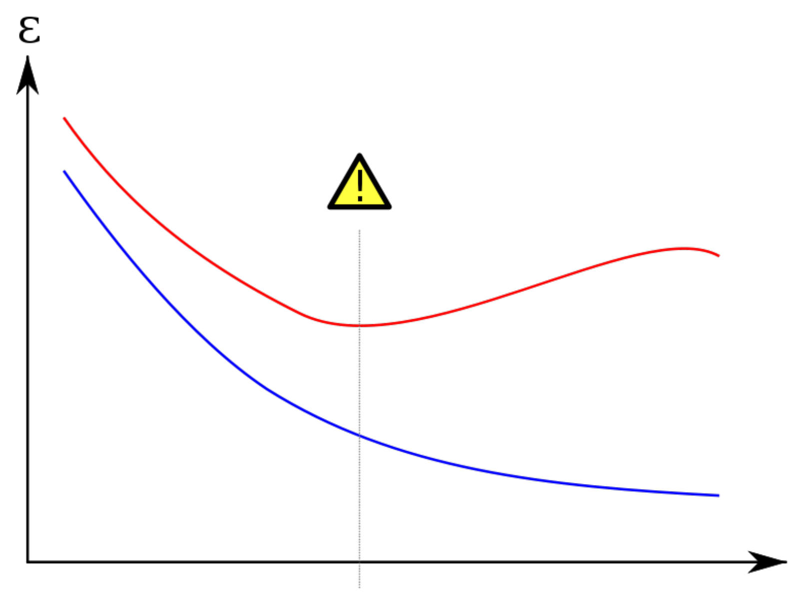
- How well has the system trained?
- When can I stop?

- Main problem: Overfitting
  - Example:
    System to detect faces in pictures
    Training set: 100 pictures, 50 of them with 10 different faces
    Aim: System should be able to detect any face
  - You train for a specific time and reach an error of 0
  - The system has learned to detect the 10 faces, NOT any face
- How to detect overfitting?

Training and test set

- You divide the whole data set into training and test set
  - Training set: Used for training
  - Test set: Used to test system on so far unseen data
    - How good are the generalisation capabilities of the system?

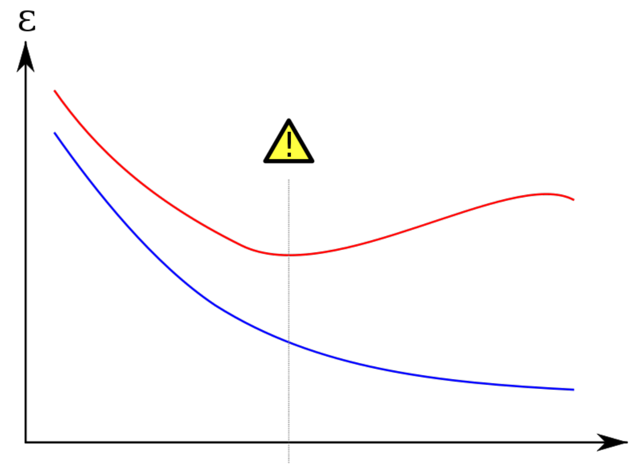
- If you run the training for a different number of epochs and then test, how would you be able to detect overfitting?



Validation set

- Can we include the test for overfitting in the training?

- Validation set
  - Different to training AND test set



- Approach:

1. For each epoch
   a) for each training data instance adjust parameters according to error
   b) calculate the accuracy A_t over training data
   c) calculate the accuracy A_v over the validation data
   d) if A_t increased and A_v stayed constant or decreased exit training else continue
2. Calculate accuracy on test set to quantify generalisation capabilities

How to divide the data set?

- Training set
  - Small to decrease training times
  - Large to decrease variation in parameter estimation
- Test and validation set
  - Large to decrease variation in performance statistic
  - Small since every test item is a lost training item
- Rule of thumb:
  - Training, Validation, Test: 60% - 20% - 20%
  - Training, Test : 80% - 20%
- Split depends on the amount of variation you want / need and the amount of data available (see also cross-validation)

Types of measures

- The sets usually contain positive and negative samples
  - e.g. pictures with faces vs. pictures without

- We get the following contingency table:

	Positive Sample P	**Negative Sample N**
Result Positive	Correct Outcome True Positive (TP)	Wrong Outcome False Positive (FP) Type I (α-)Error
Result Negative	Wrong Outcome False Negative (FN) Type II (β-)Error	Correct Outcome True Negative (TN)

Performance measures

	Positive Sample P	Negative Sample N
Result Positive	True Positive	False Positive
Result Negative	False Negative	True Negative

- Precision
  - How many of the positive results were really correct?
  - Precision = TP / (TP + FP)
  - Also: Positive Predictive Value
- Recall
  - How many of the available positives were found?
  - Recall = TP / P = TP / (TP + FN)
  - Also: Sensitivity, hit rate, True Positive Rate (TPR)
- Both usually not sufficient on their own (100% recall by always returning positive result)

Performance measures

- Negative Predictive Value
  NPV = TN / (TN + FN)
- True Negative Rate
  TNR = TN / (TN + FP)
  - Also: Specificity

	Positive Sample P	Negative Sample N
Result Positive	True Positive	False Positive
Result Negative	False Negative	True Negative

- We would like a measure to include both errors

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Not that common in our field

F-Score

- F-Measure combines precision and recall
- F_1-Score
  - Harmonic mean between precision and recall

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

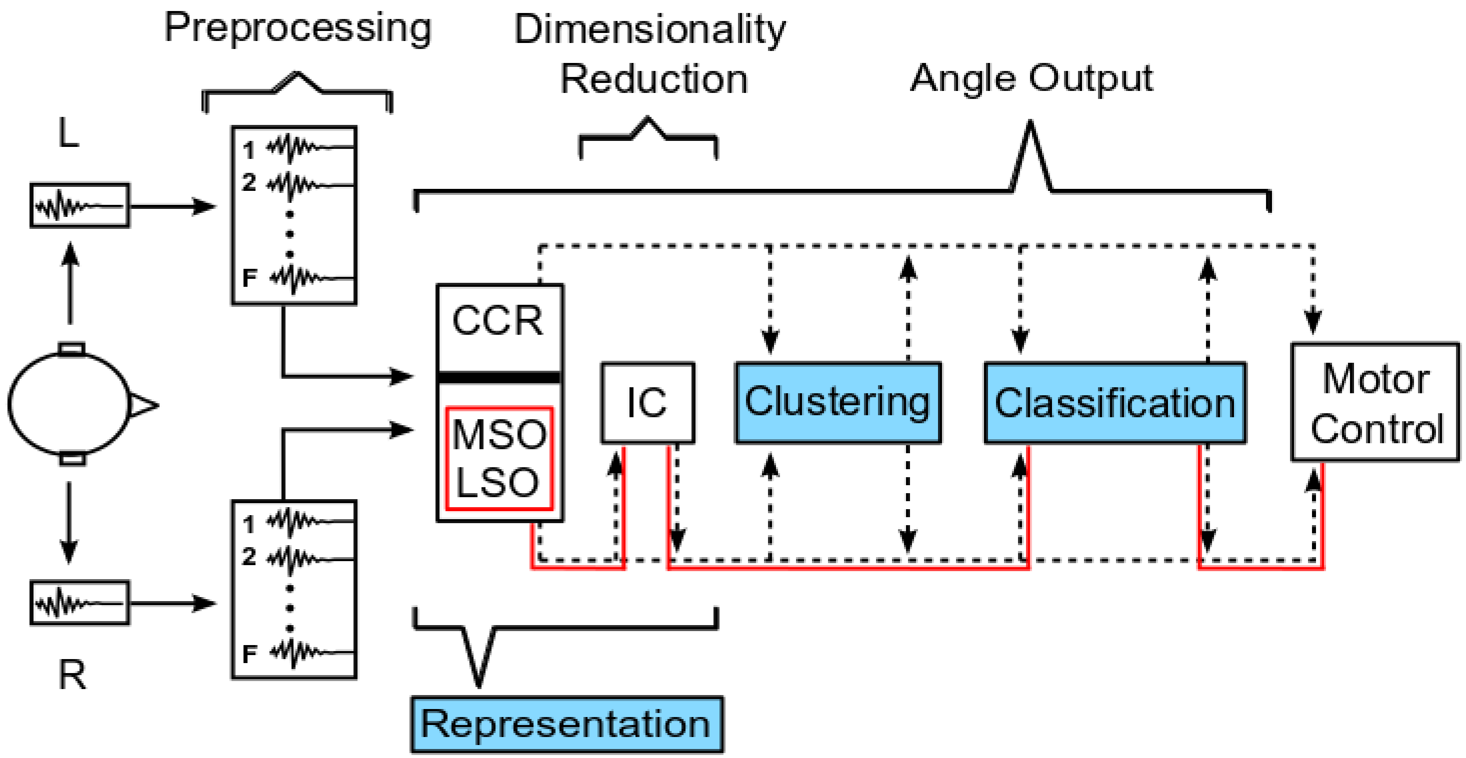
- General: F_β-Score

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

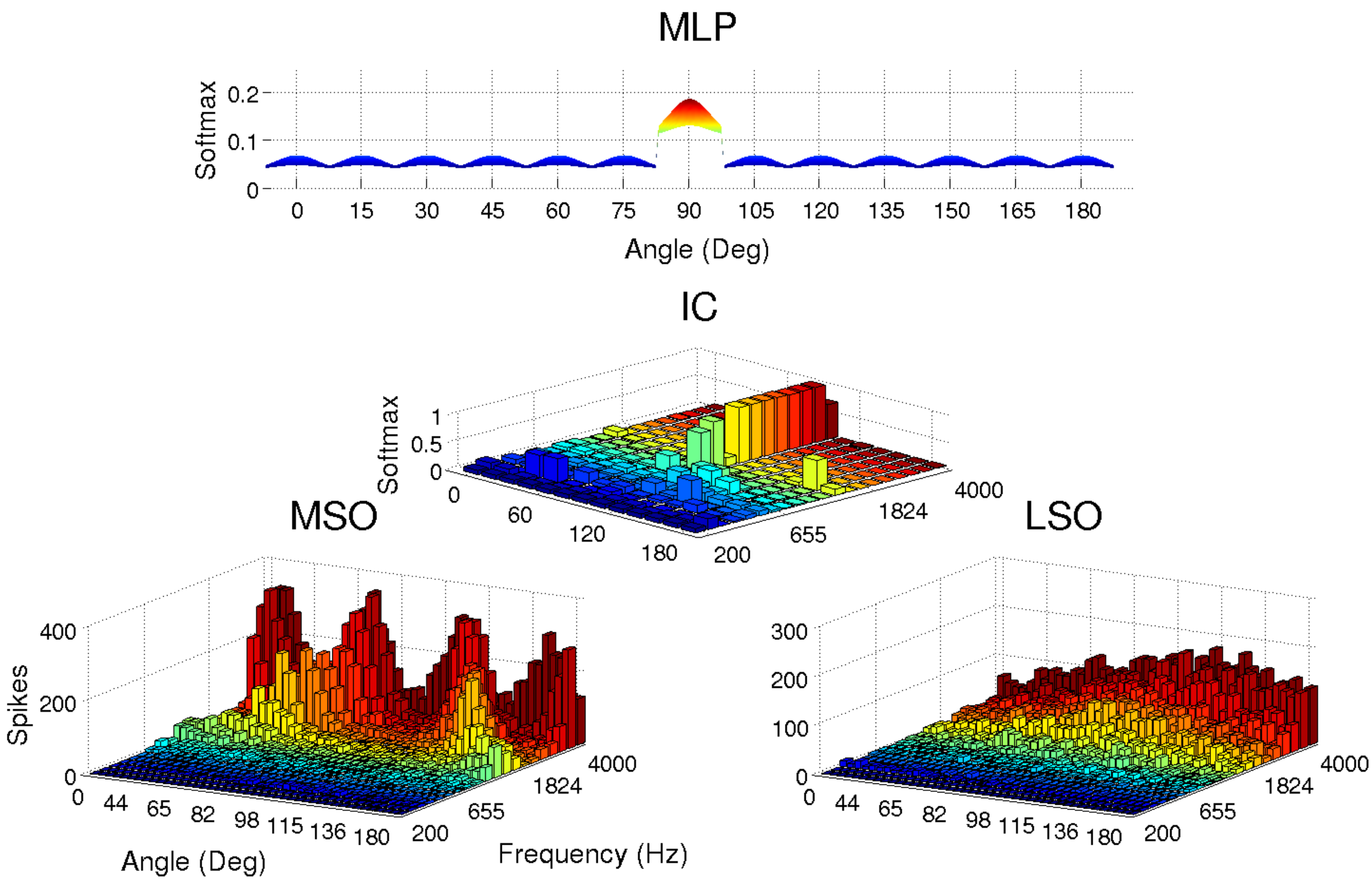
  - van Rijsbergen (1979): "[…] effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision"

Examples for visualisation

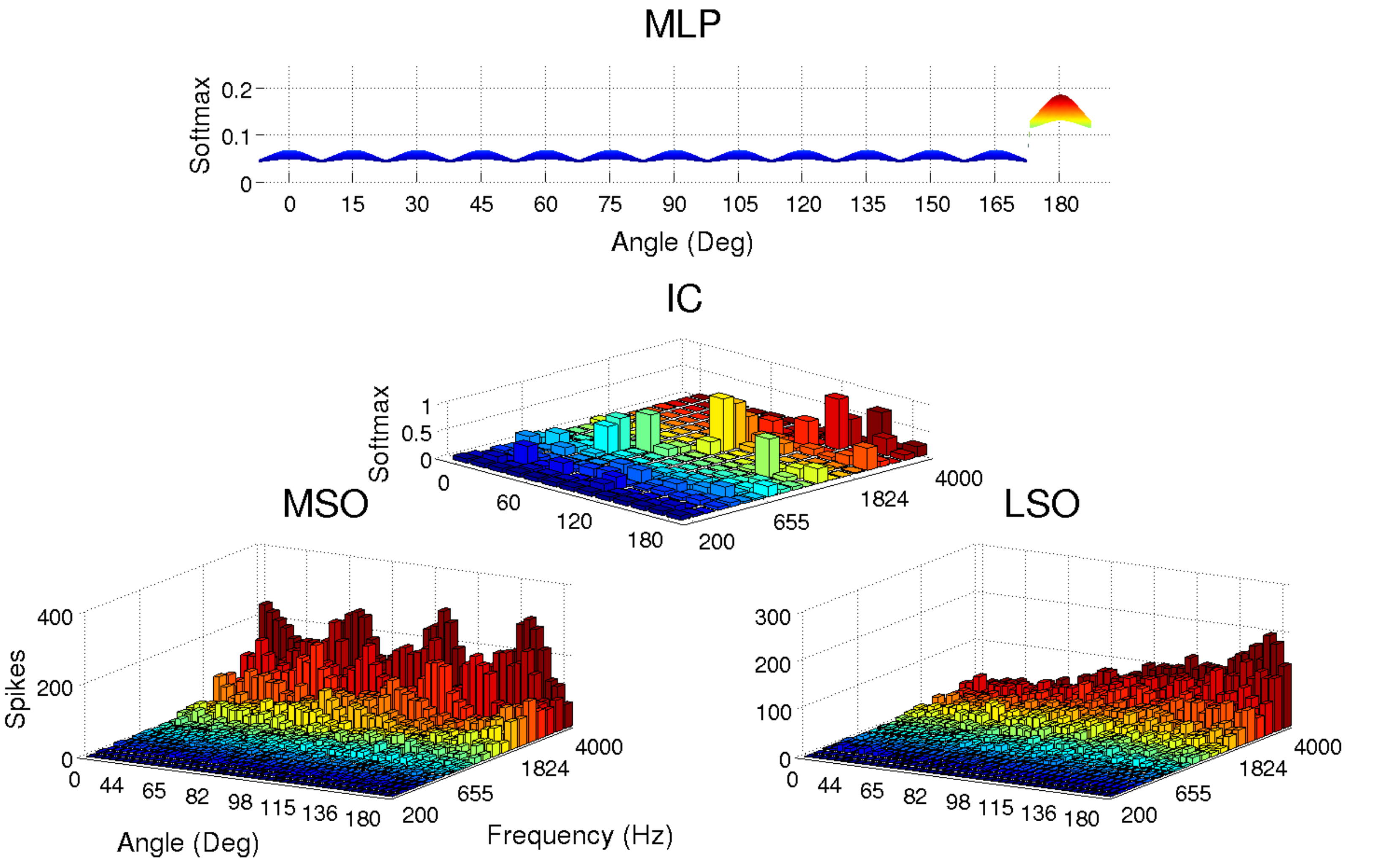
- Jorge: Bio-inspired sound source localisation



Examples for visualisation

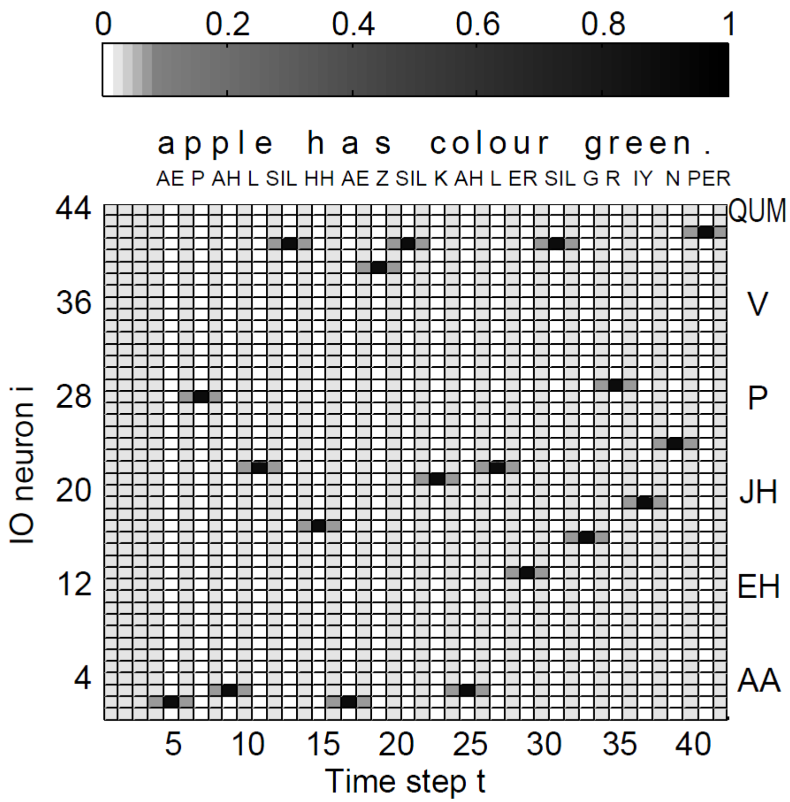


Examples for visualisation

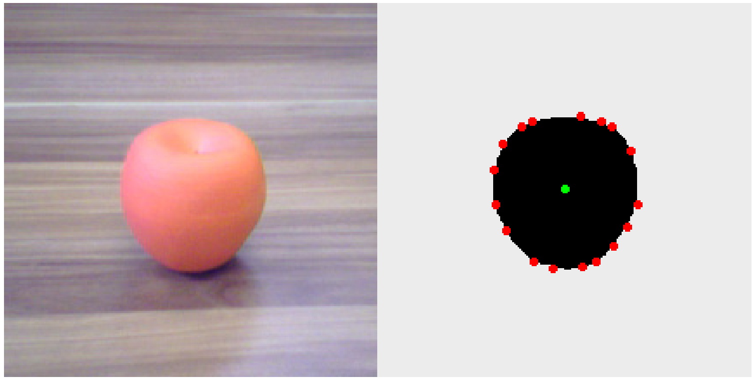


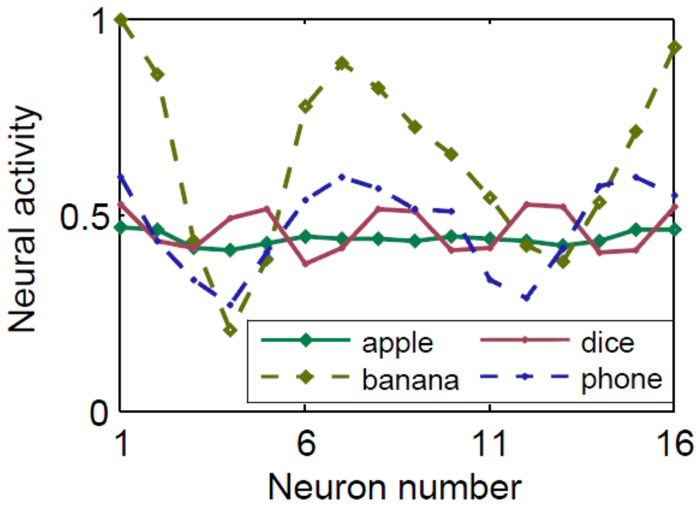
Examples for visualisation

Stefan: Embodied Language Understanding with an MTRNN

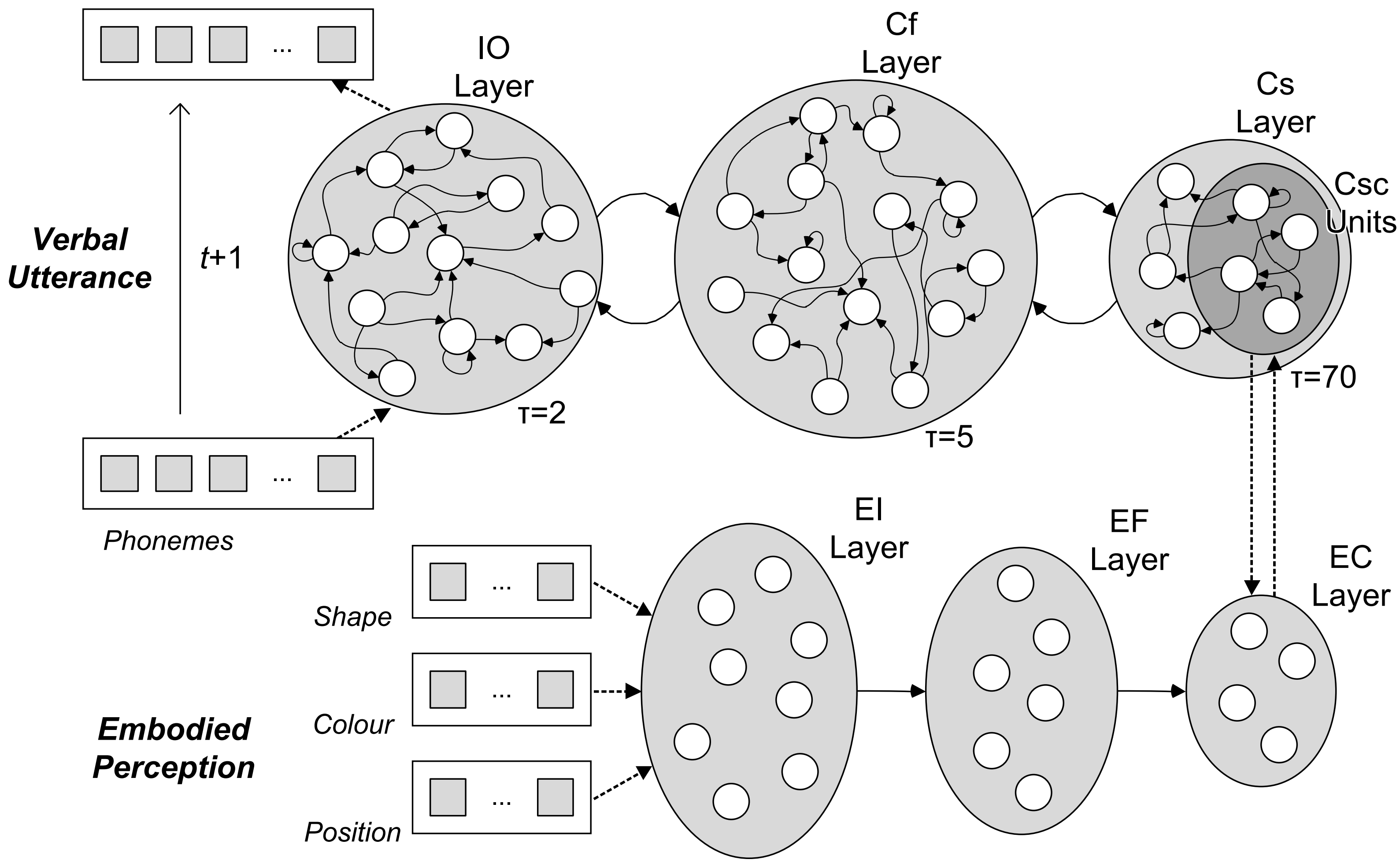


Sample encoded utterance



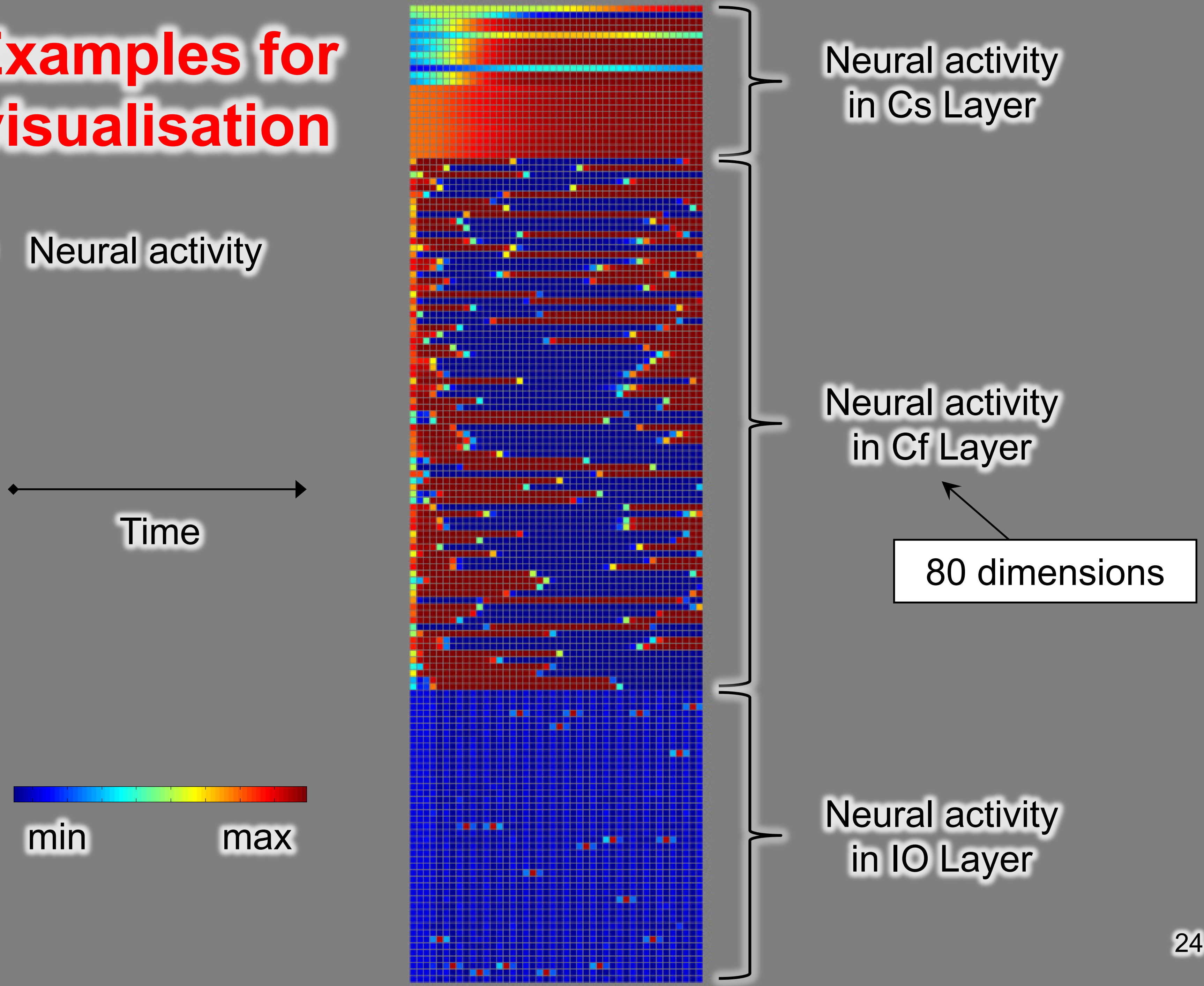


Examples for visualisation



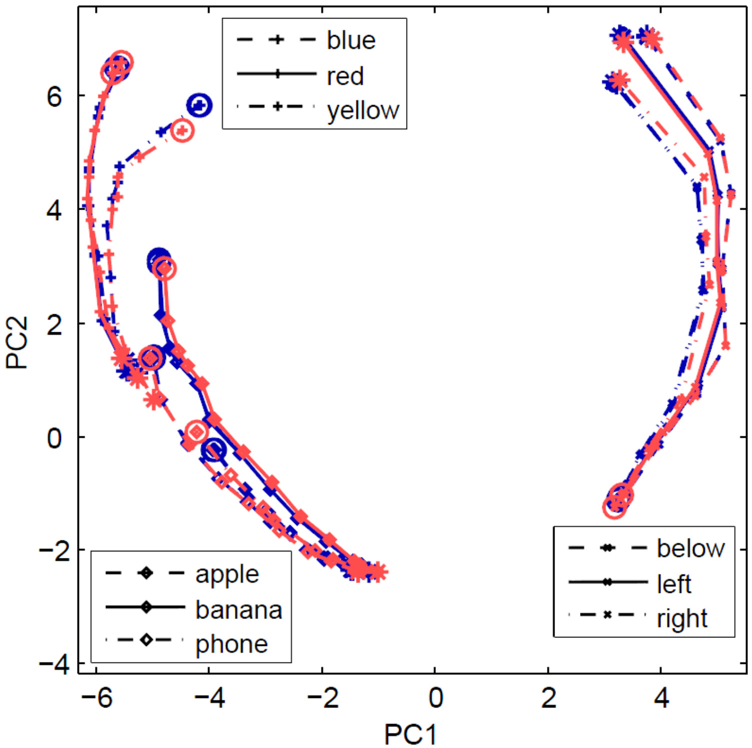
Examples for visualisation

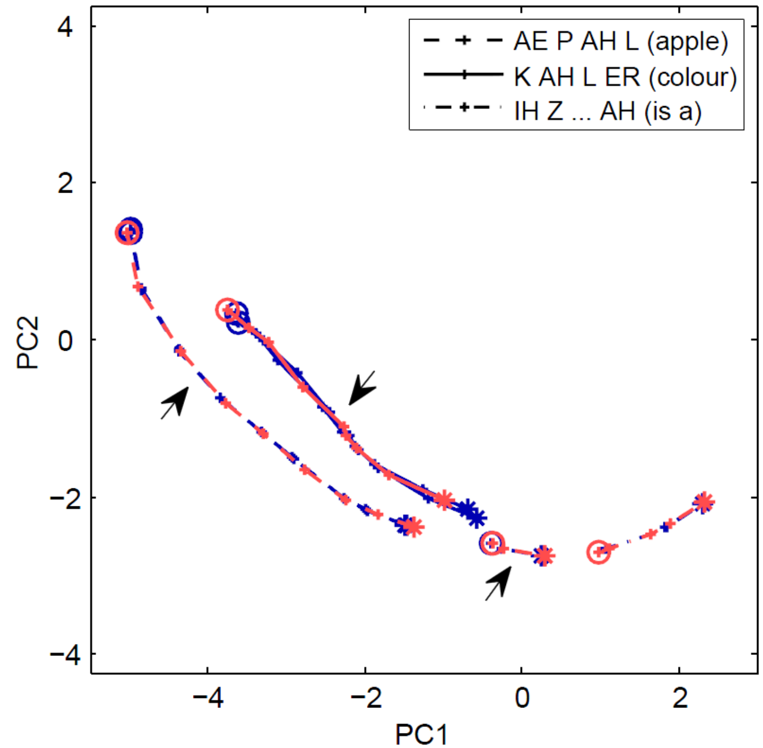
- Neural activity



Examples for visualisation

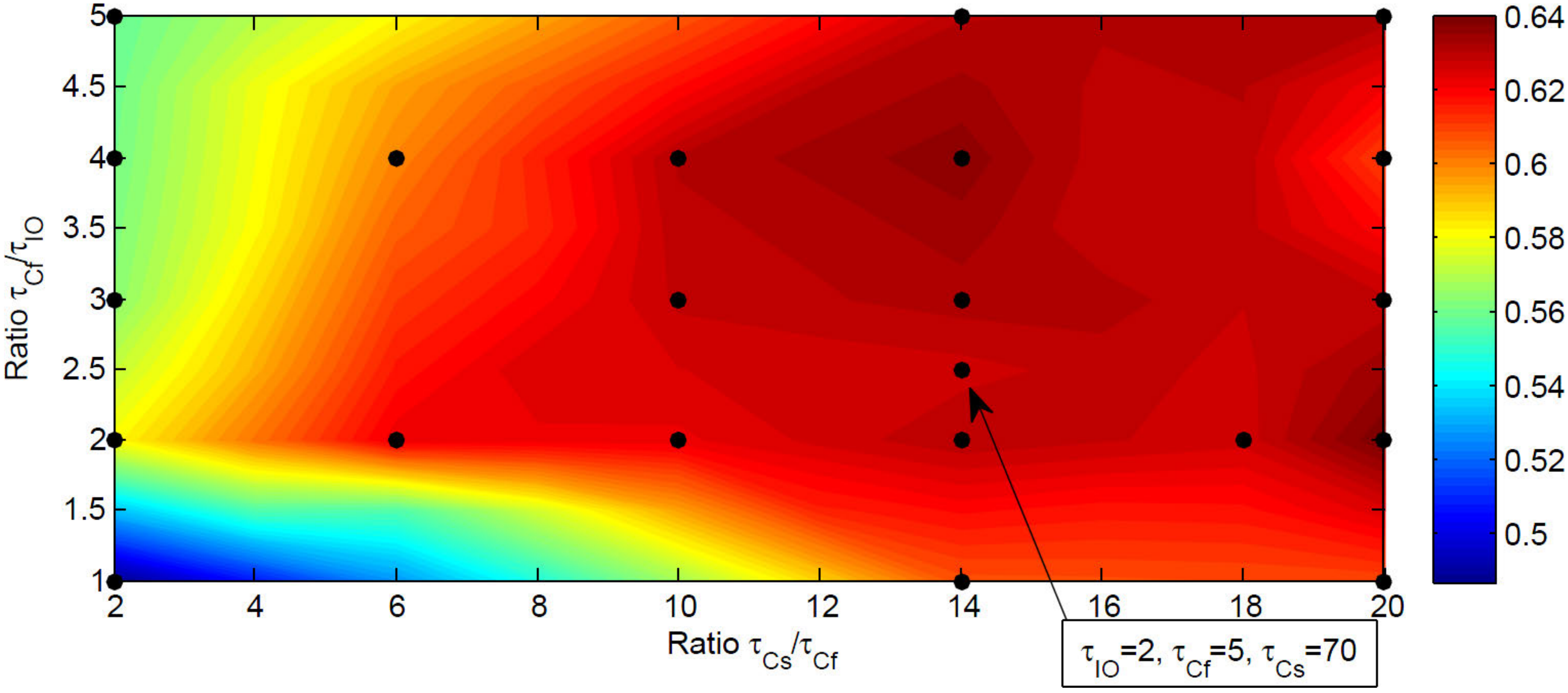
- Principle component analysis (PCA)





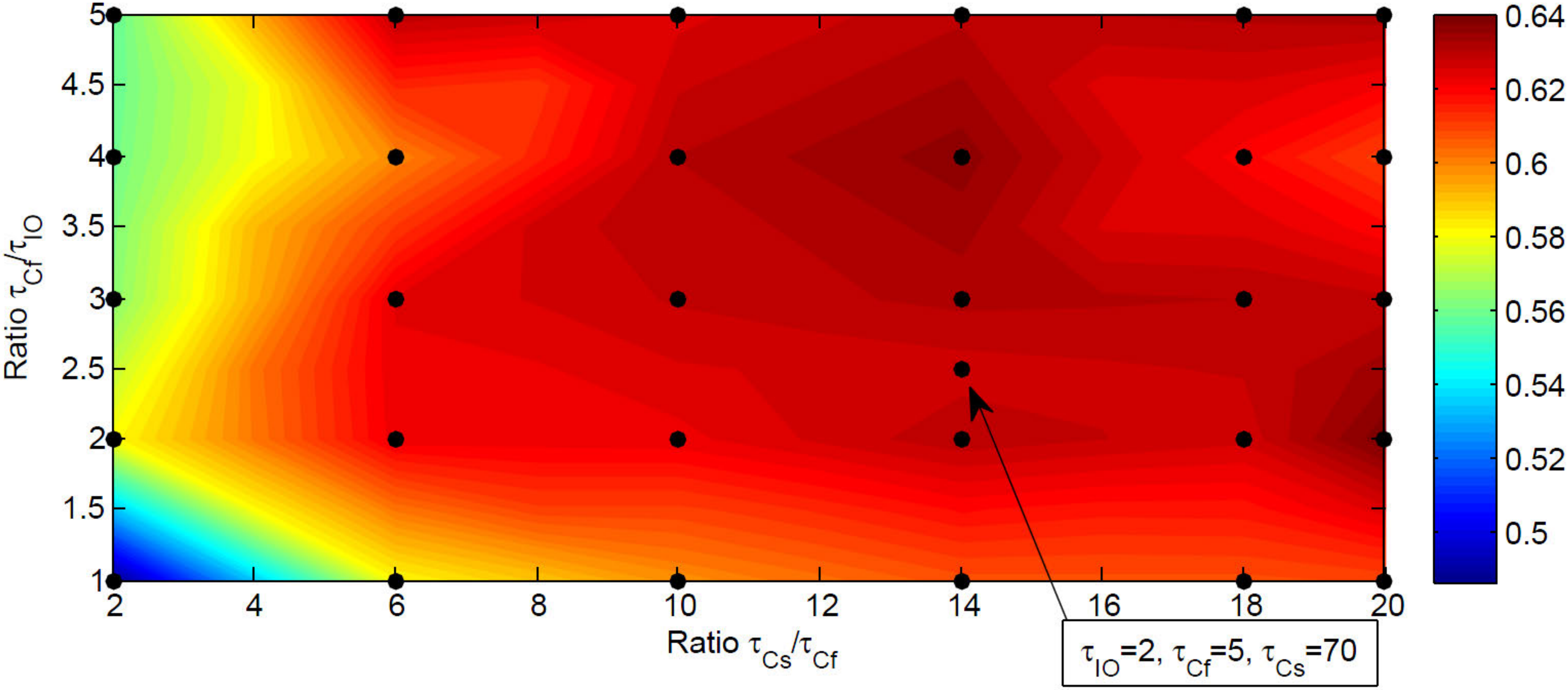
Examples for visualisation

- Parameter tuning: Timescales



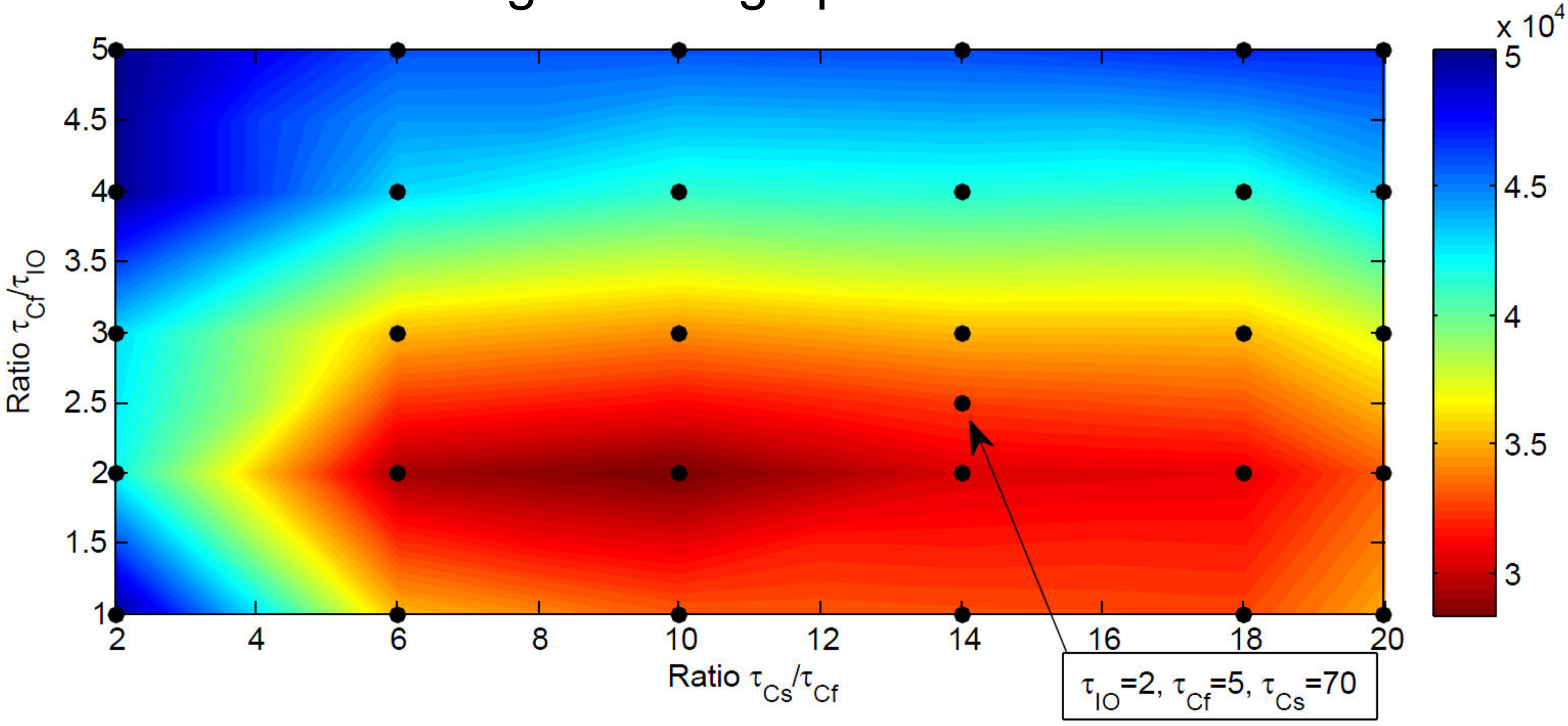
Examples for visualisation

- Parameter tuning: More data points



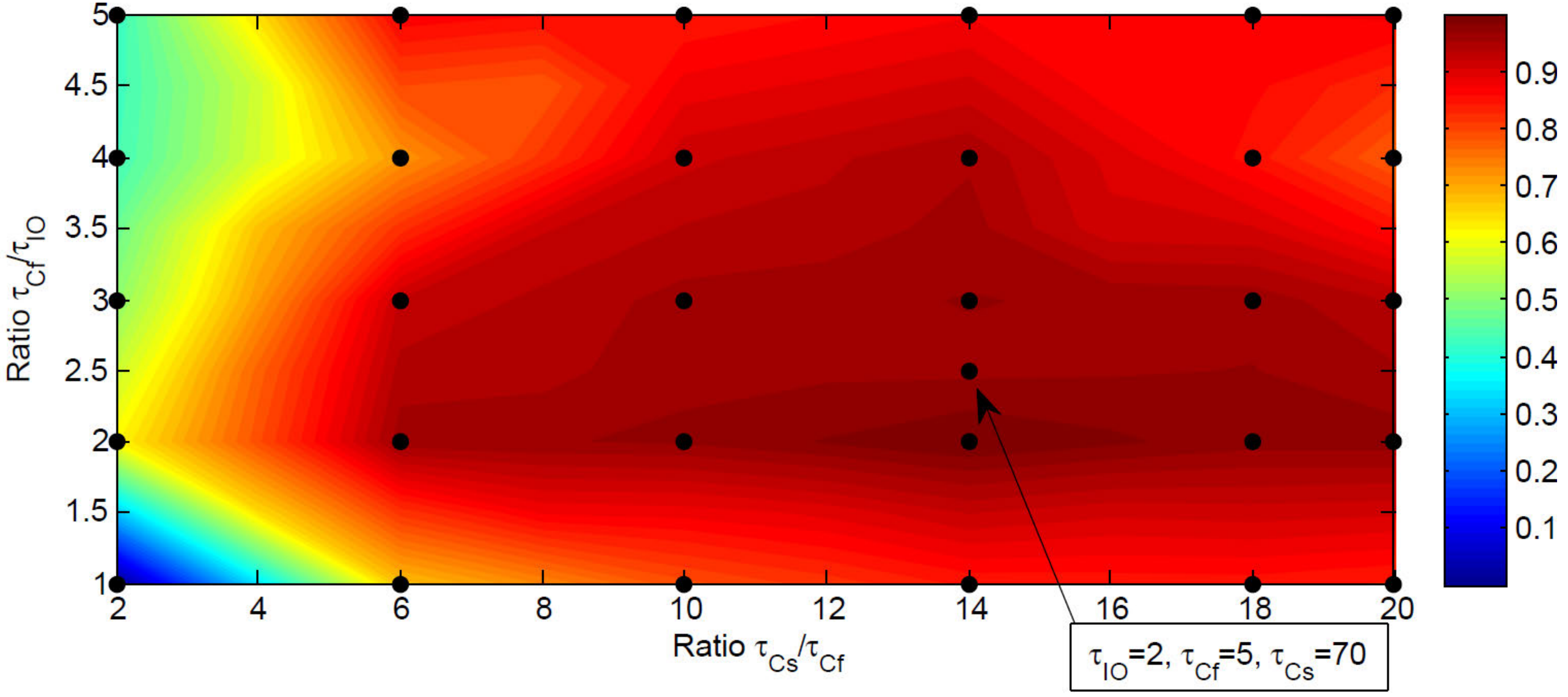
Examples for visualisation

- Parameter tuning: Training epochs



Examples for visualisation

- Parameter tuning: Training time



Summary

- What have we covered?
  - Scientific process and its components
  - Data and variability
    - Data scales, factor vs. variable, samples
    - Variability within and between, measuring variability
    - Difference between sample and population
  - Descriptive statistics and exploratory studies
    - Central tendency, shape of distributions, dispersion measures
    - Visualisations for uni- and multi-variate EDA
    - Joint distributions, contingency table, χ^2
    - Covariance, correlation coefficients
    - Time series, trend, smoothing, differencing

Summary

- Hypothesis testing and parameter estimation
  - Hypothesis, p-values, general form of statistical tests
  - Sampling distributions and how to get them
  - Statistical tests (Z-, t-, Fishers r-to-z, …)
  - confidence intervals
- Experiments design
  - Effects, independent/dependent variables
  - Control, placebo/blinding, randomization
  - Biases, spurious effects
  - Sample size, large or small?
- Human participants & data collection
  - data collection, questionnaire design
- Publishing & Peer review

Oral Exam

- First date: 14. 2
- Second date: 28.3

- Time slots as given by the exams office
- Around 25min examination
- Room: F-210
- Content:
  - Lecture
  - Homework and discussion

Questions?