Statistical Limits of Machine Learning

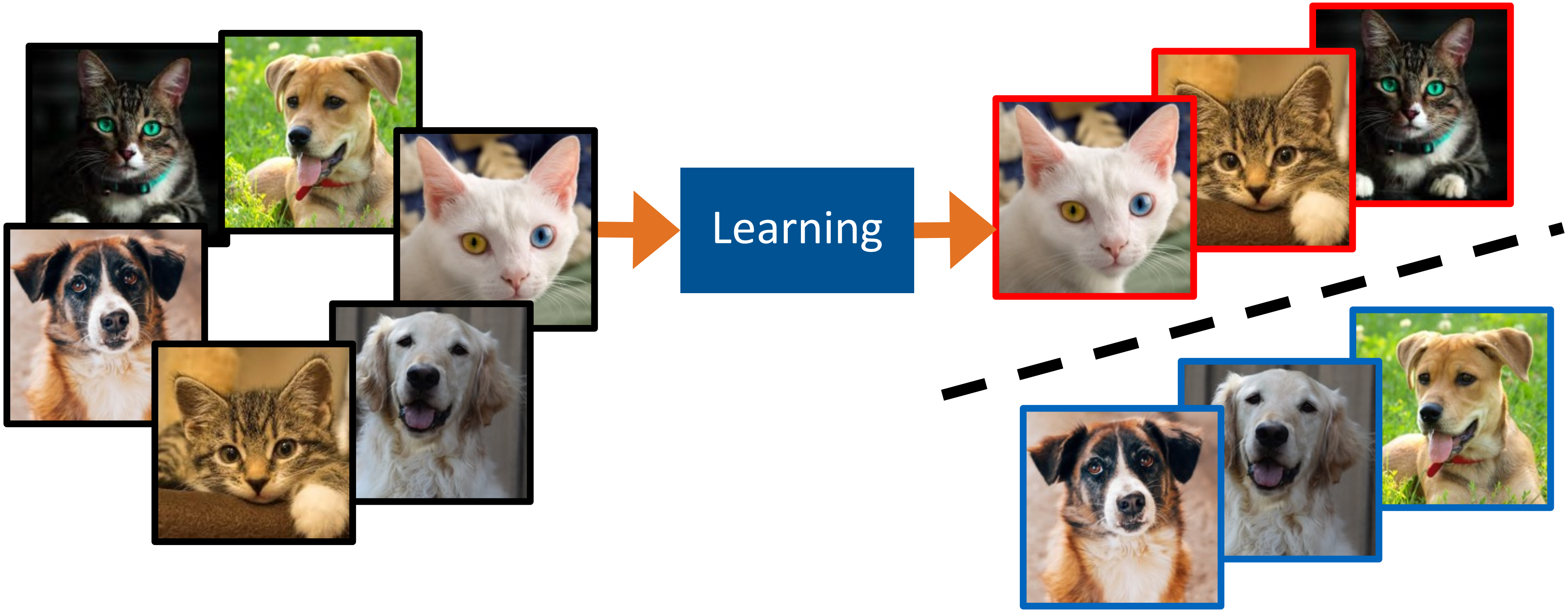
Debarghya Ghoshdastidar

Assistant Professor

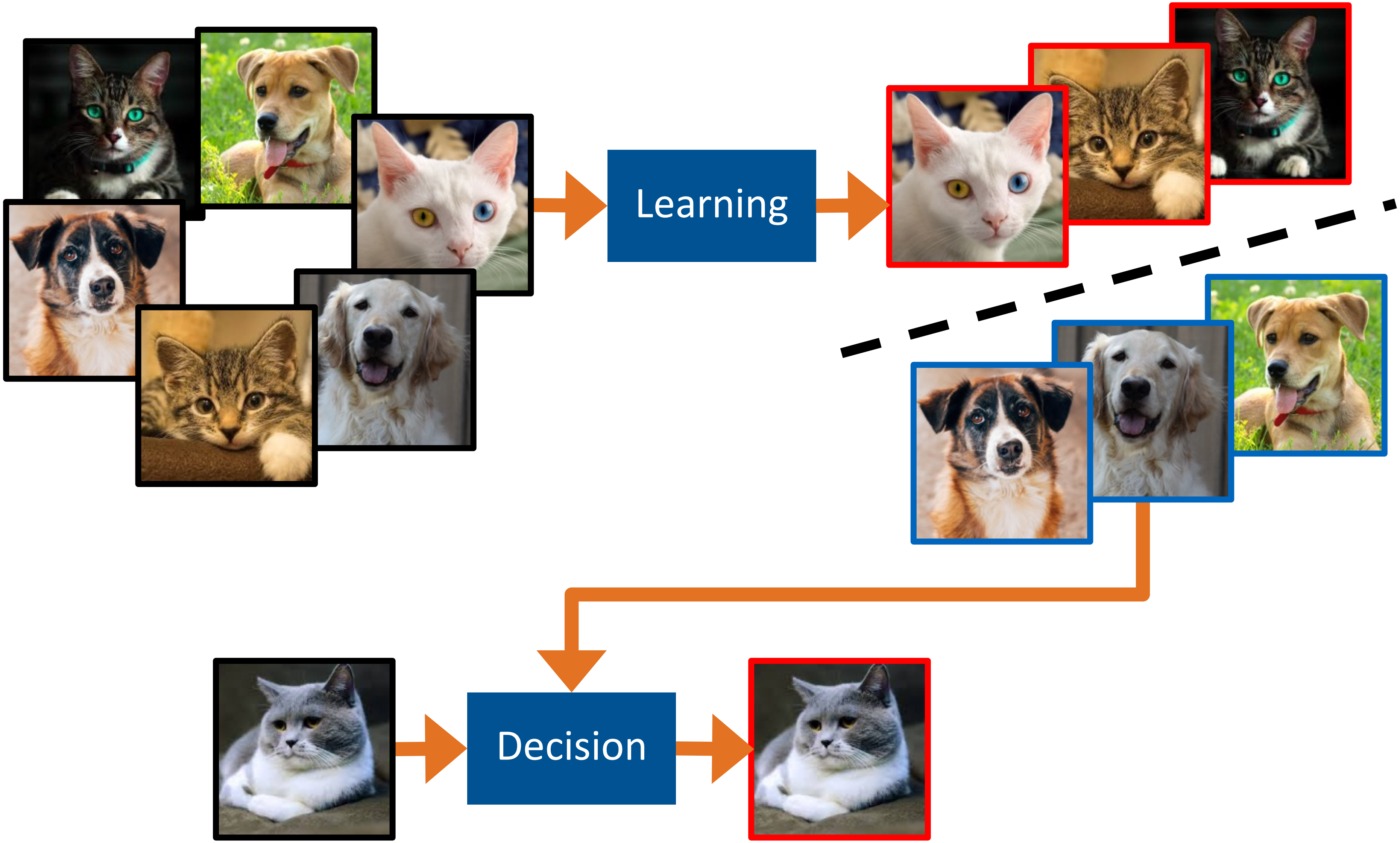
Theoretical Foundations for Artificial Intelligence

TUM Informatik

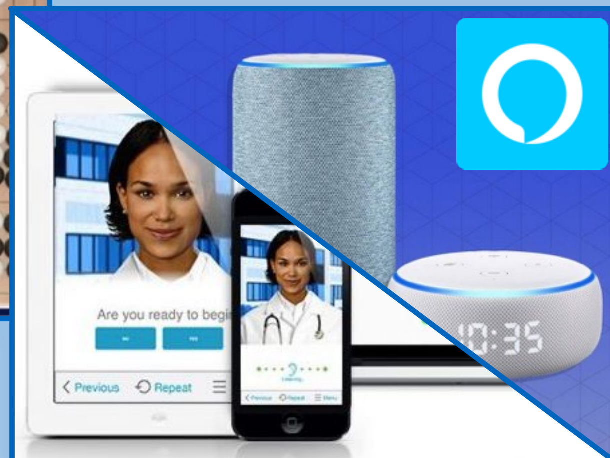
Machine Learning



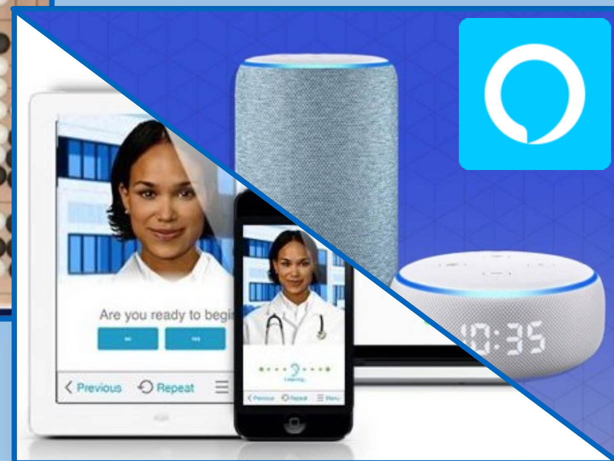
Machine Learning



Machine Learning works



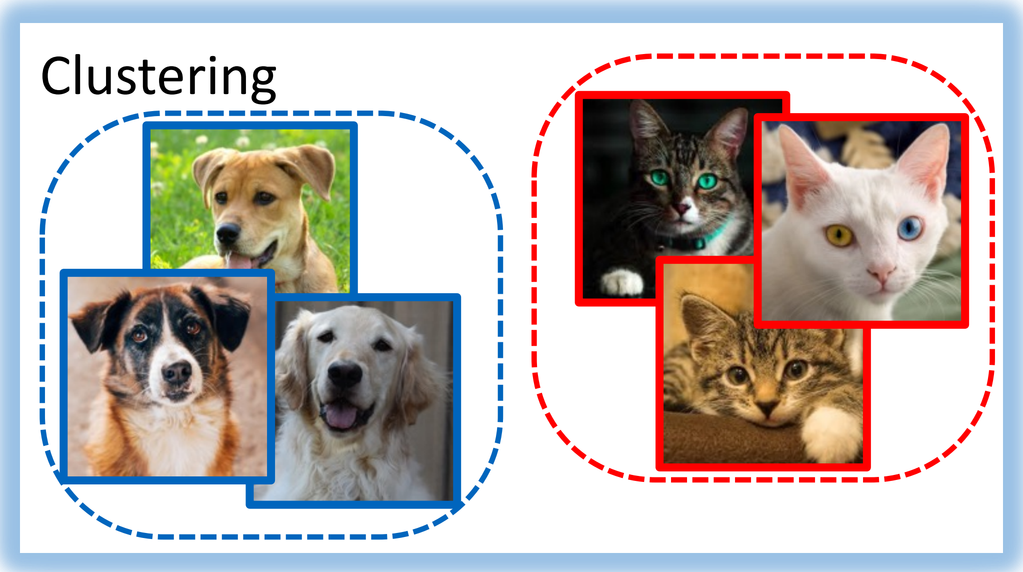
Machine Learning works



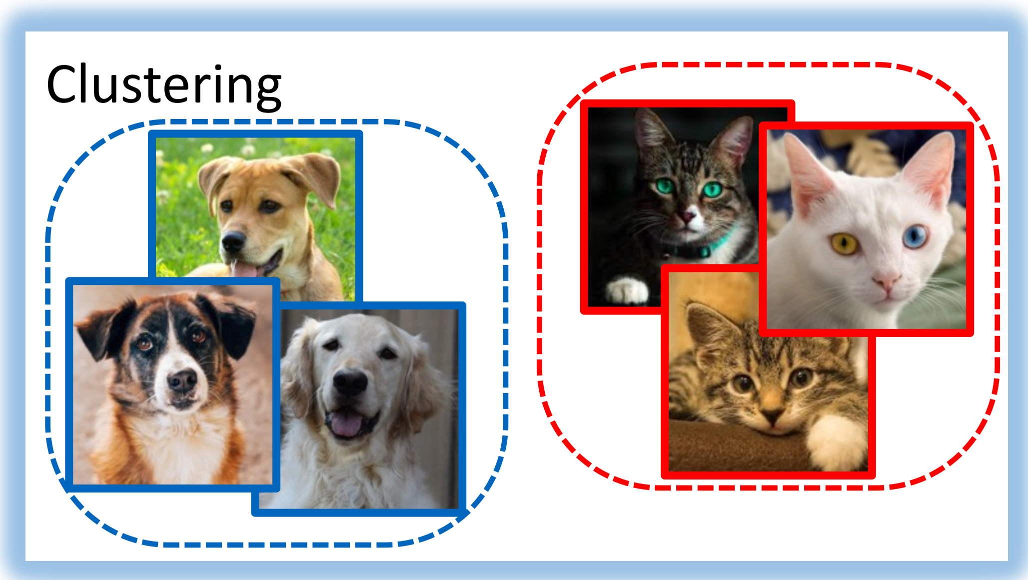
- What do ML algorithms learn?
- Why do ML algorithms work?
- When can we find patterns?

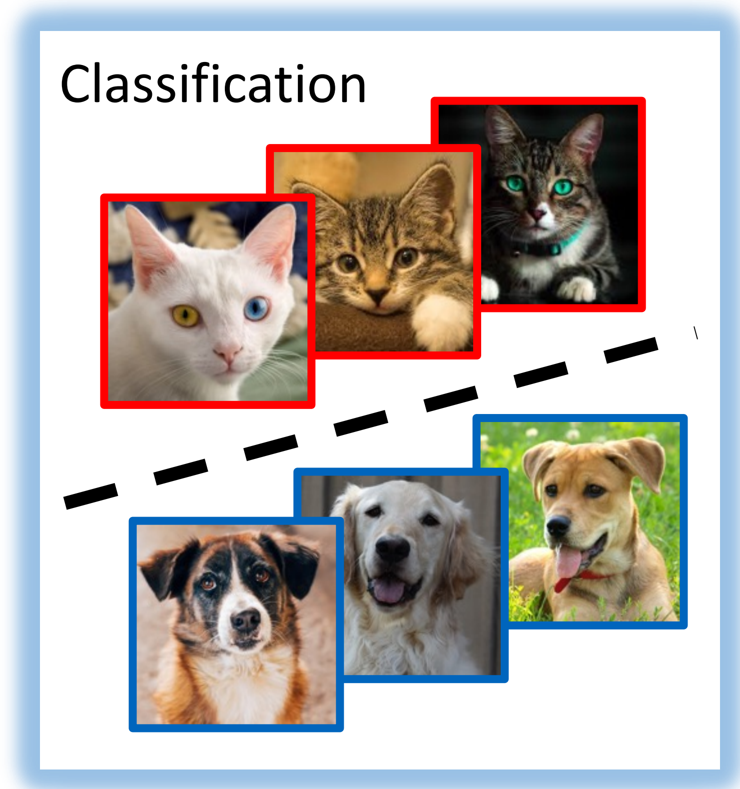
Diversity of Machine Learning

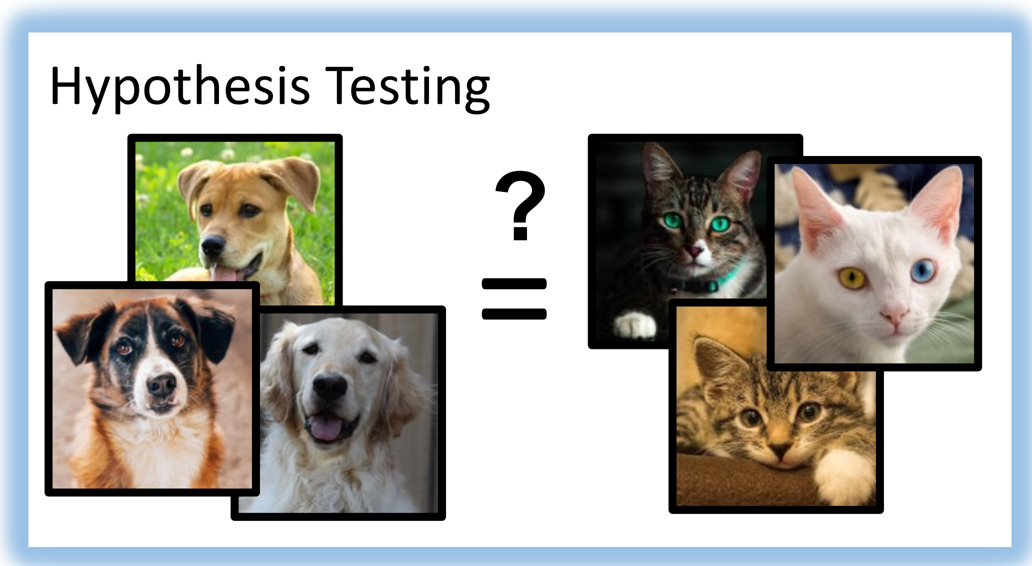
Diversity of Machine Learning



Diversity of Machine Learning







... and many more

We know a lot, but not everything

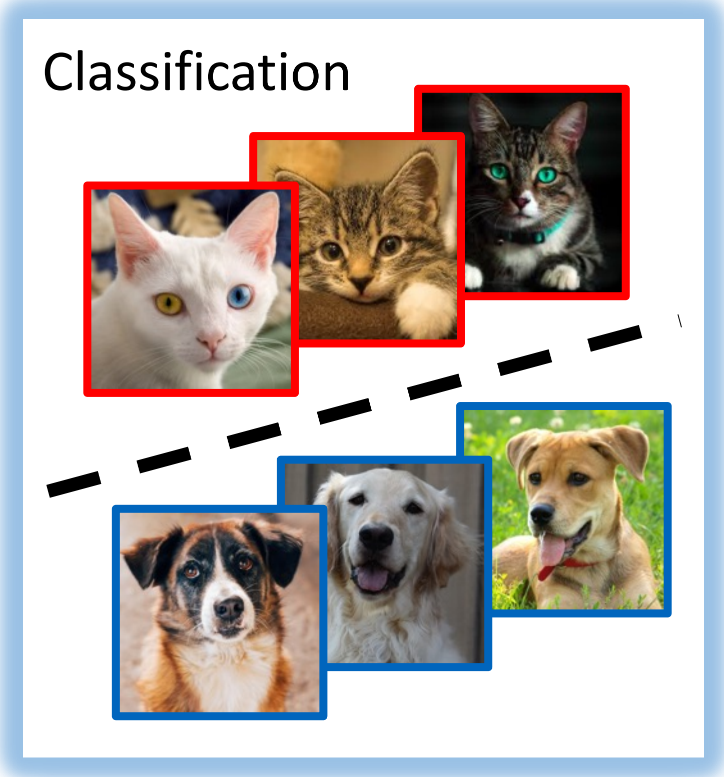
- Machine learning works because we have huge amount of data

  ... limited data in many problems

- Sample complexity:

  How much training data needed to learn good classifier?

  ... Statistical limit of learning !!



- Theoretical foundation since 1980s

  PAC learning, Generalisation ...

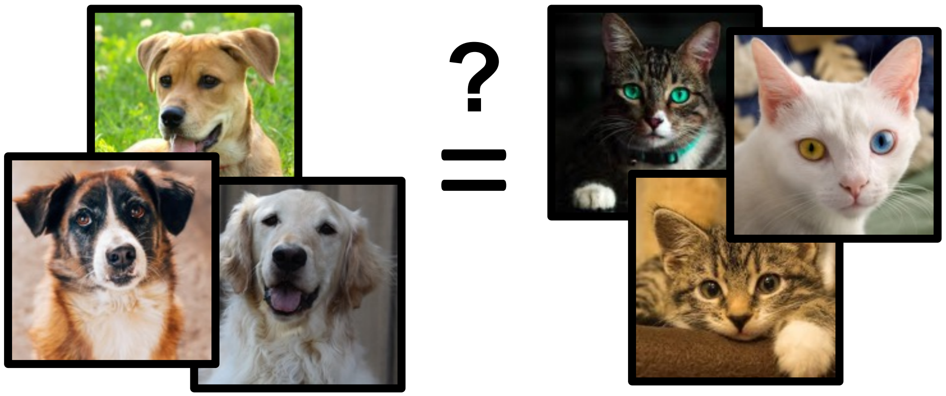
We know something, but not enough

- Studied in statistics since 1900s, but theory for smaller data
- Less understood in high-dimensional setting:

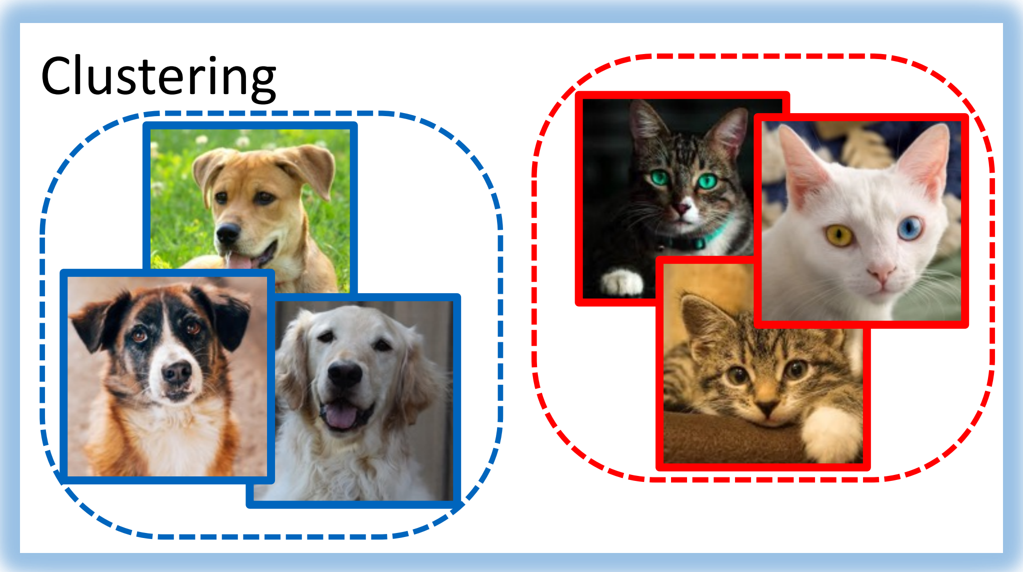
number of samples < data dimension

- Statistical limit of testing:
  - When can we detect differences?

Hypothesis Testing



We know very little



- Still no consensus on the definition of *good* cluster
- Many theoretical results, but limited understanding
- Statistical limit of clustering:
  - When does data reveal clusters?
  - When can we find clusters?

Focus of our group

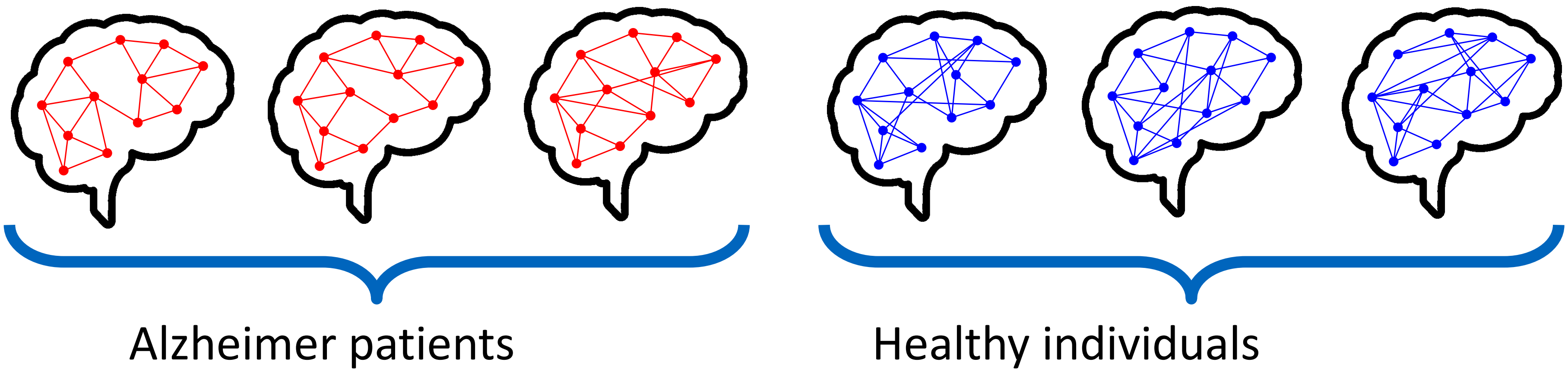
- **Statistical limits of planted clustering**
  - Find clusters hidden in random data
  - High-dimensional data / Large graphs

- **Statistical limits of hypothesis testing for graphs**
  - Two-sample testing

Two-sample testing of graphs

- Given two populations,

  test if both come from same distribution

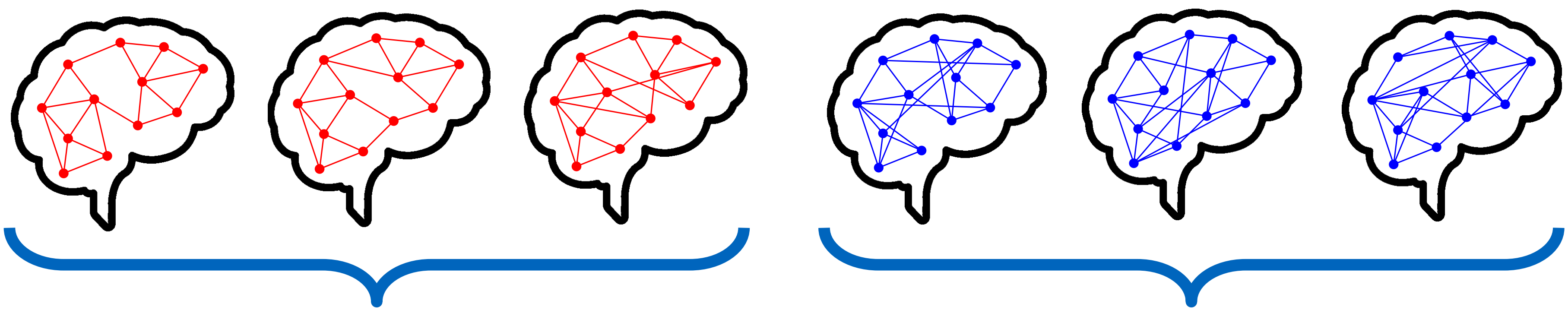


- Do brain networks reveal neurological disorders?

- Do we interact in the same way on various social networks?

Two-sample testing of graphs

- All graphs on common set of n vertices
- Two models (distributions) $\mathcal{P}$ and $\mathcal{Q}$
- Observe m graphs from each model



$$G_1, \ldots, G_m \sim \text{model-}\mathcal{P}$$

$$H_1, \ldots, H_m \sim \text{model-}\mathcal{Q}$$

- **Problem:**

$\mathcal{P} = \mathcal{Q}$ vs. $\mathcal{P} \neq \mathcal{Q}$

models identical — models different

Testing with few samples

Study on Alzheimer (ADNI) [Zajac et al. Brain Sci. 2017]

- Structural brain networks with 68 vertices (ROIs)
  - 10 Alzheimer patients
  - 10 Control (healthy) subjects … $m = 10$
- **Conclusion:** Alzheimer affects brain network

Oregon network data [Leskovec et al. KDD 2005]

- Peering networks among 11806
  - 2 networks generated per week … $m = 2$
  - Data for 9 weeks (9 different groups)
- **Conclusion:** Networks change significantly over time

Theoretical concerns

- Classical tests typically work when $m \to \infty$
  - Often use asymptotic null distributions

- Any test /algorithm returns a result
  - Not necessarily correct for small m
  - Need methods with guarantees for small m

- Small changes cannot be detected for small m
  - Modify the problem:

$$\mathcal{P} = \mathcal{Q} \qquad \text{vs} \qquad d(\mathcal{P}, \mathcal{Q}) > \rho$$

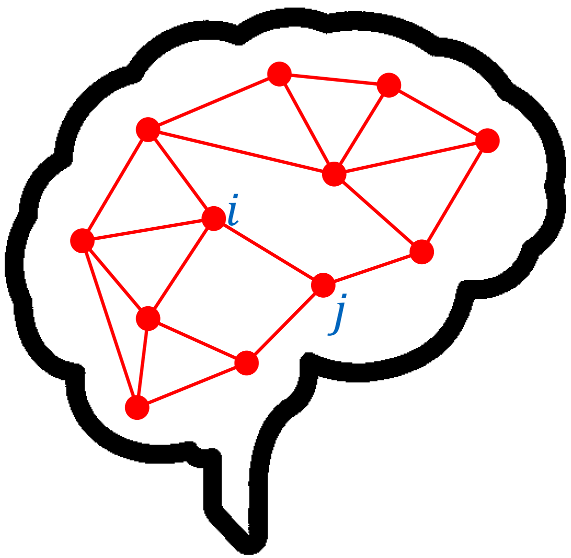
models identical   models highly different

  - Which distance should we use?

A simple graph model

- Common vertex set $\{1, 2, \ldots, n\}$
- Inhomogeneous Erdös-Rényi (IER) model
  - All edges are independent
- Model $\boldsymbol{\mathcal{P}}$ characterised by $n \times n$ matrix P
  - Edge (i, j) added with probability P_{ij}
- Model $\boldsymbol{Q}$ characterised by $n \times n$ matrix Q
- Given graphs $G_1, \ldots, G_m \sim \mathrm{IER}(P)$ and $H_1, \ldots, H_m \sim \mathrm{IER}(Q)$
- Test hypotheses:

$$P = Q \qquad \text{vs} \qquad d(P, Q) > \rho$$



Typical two-sample test

- Given: $G_1, \ldots, G_m$ (1^{st} population) and $H_1, \ldots, H_m$ (2^{nd} population)

- Let $\hat{P}_{ij} =$ fraction of graphs in 1^{st} population with edge (i, j)

  $\hat{Q}_{ij} =$ same for 2^{nd} population

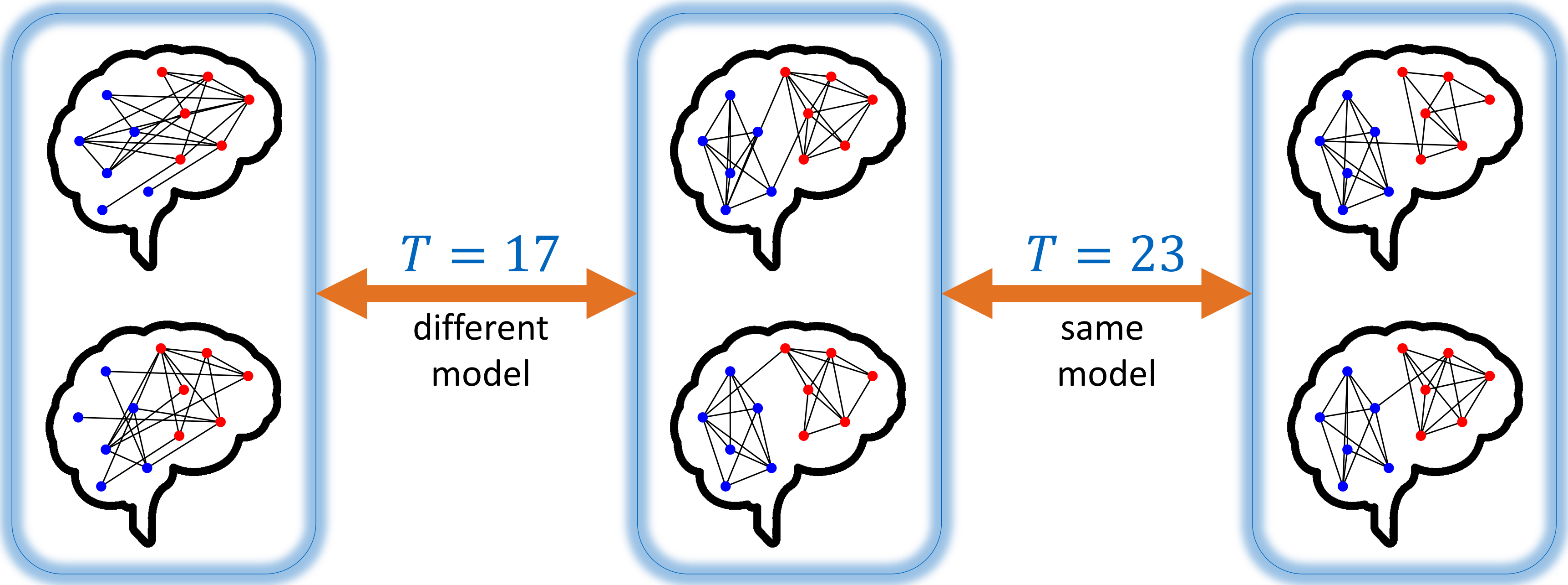
- Statistic $T = \sum_{(i,j)} w_{ij} \left(\hat{P}_{ij} - \hat{Q}_{ij}\right)^2$ $\qquad w_{ij} =$ suitable weights

- Theory: $\lim_{m \to \infty} T = \begin{cases} \chi^2\text{−random variable} & \text{for } P = Q \\ \infty & \text{for } P \neq Q \end{cases}$

χ^2-test: Say models are different if T large

- Result: Test has high accuracy for large m

Performance for small m



- **Theoretical result:**
  - Test detects difference in total variation (TV) distance
  - No high accuracy test for TV-distance for $m \ll n$

New two-sample tests

- Aim to detect difference in matrix norms (Frobenius, spectral)
- Statistic $T =$ unbiased estimate of $\|P - Q\|$
  – Different norms lead to different new tests

– Ghoshdastidar & Luxburg. Practical methods for graph two-sample testing. *Neurips 2018.*

– Ghoshdastidar et al. Two-sample hypothesis testing for inhomogeneous random graphs. *The Annals of Statistics (in press).*

New two-sample tests

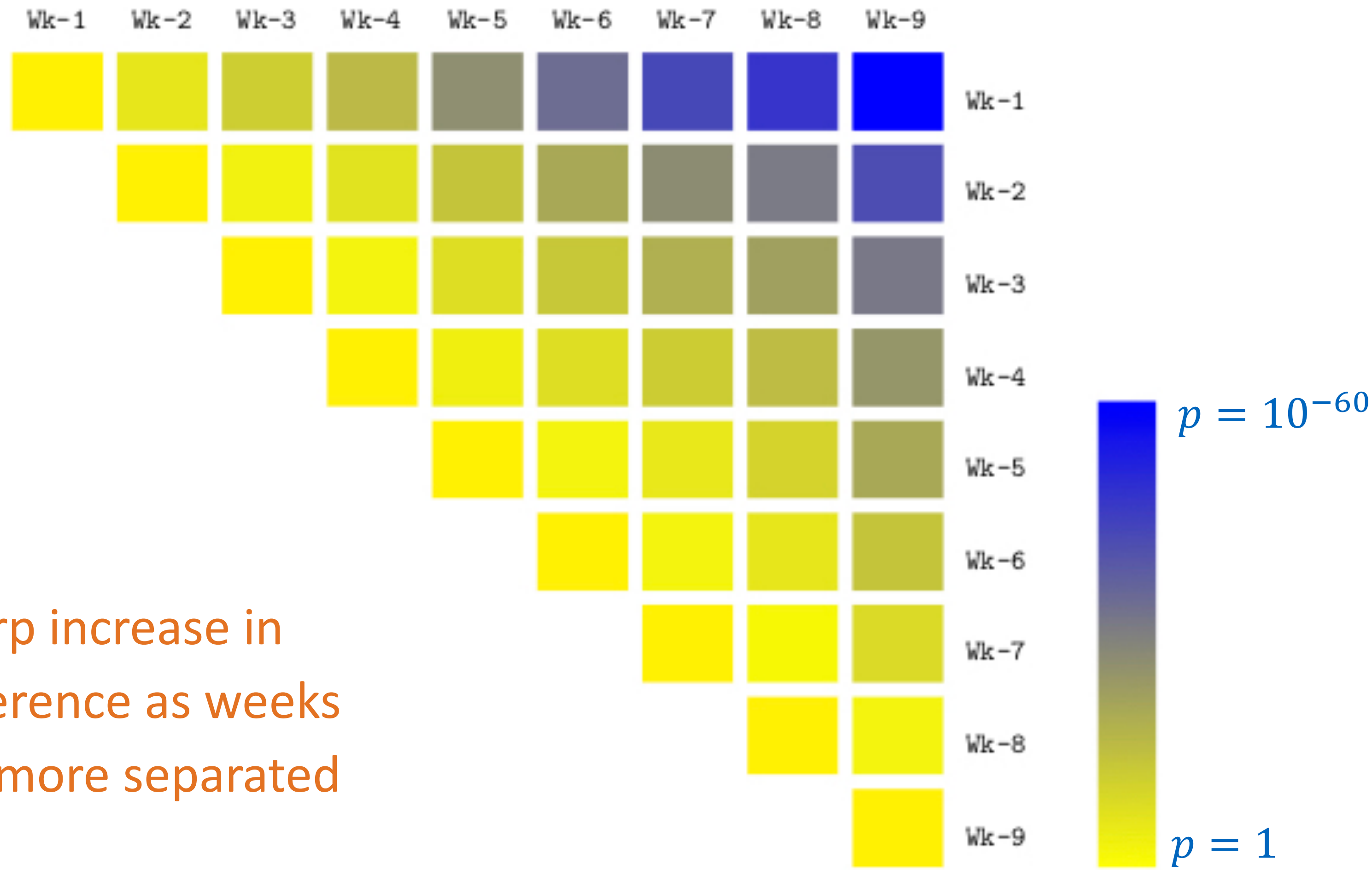
- Aim to detect difference in matrix norms (Frobenius, spectral)
- Statistic $T =$ unbiased estimate of $\|P - Q\|$
  - Different norms lead to different new tests
- **Theoretical results:**
  - Tests have high accuracy as $n \to \infty$ for every $m \geq 1$
  - No test can detect small separation $\|P - Q\| \lesssim \sqrt{\frac{n}{m}}$
  - Tests are optimal: Accurate whenever $\|P - Q\| \gtrsim \sqrt{\frac{n}{m}}$

– Ghoshdastidar & Luxburg. Practical methods for graph two-sample testing. *Neurips 2018.*

– Ghoshdastidar et al. Two-sample hypothesis testing for inhomogeneous random graphs. *The Annals of Statistics (in press).*

Testing Oregon networks

- Peering information of $n = 11806$ routers over 9 weeks
- $m = 2$ networks for each week (classical tests do not work)
- Colour plot for p-values (lower p indicates more difference)



Sharp increase in difference as weeks are more separated

Conclusion: Statistical limits of testing

- Difficult to infer from few samples / graphs
  - Problem may become unsolvable (in minimax sense)
- Should not *blindly* apply classical tests
  - Need new techniques / new perspectives
- Better understanding of tests / algorithms needed
- General recommendation:

Look before you leap (into conclusion)

Research Group

Leena Vankadara

Ph.D. Student

University of Tübingen

Pascal Esser

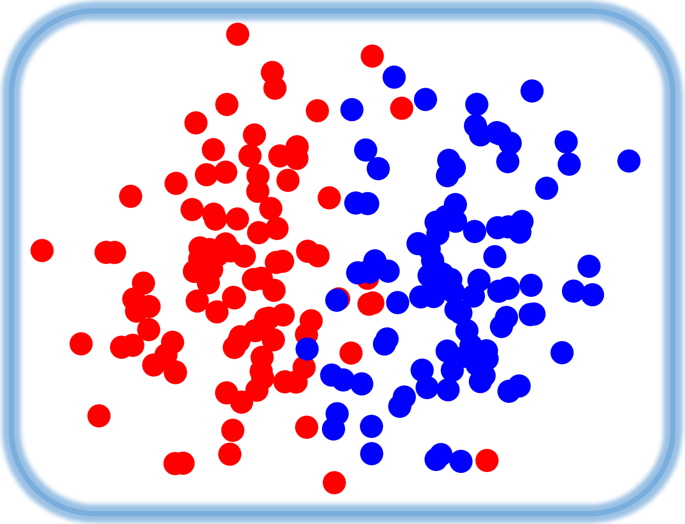
Ph.D. Student

joined in December

Statistical limits of planted clustering

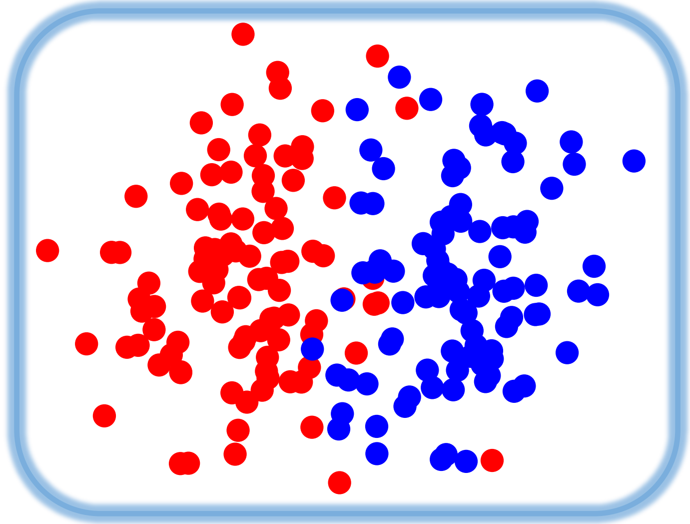
- Clustered data + random noise
  - High dim data / large graphs
- When can we say that there are clusters?

Information theoretic limit



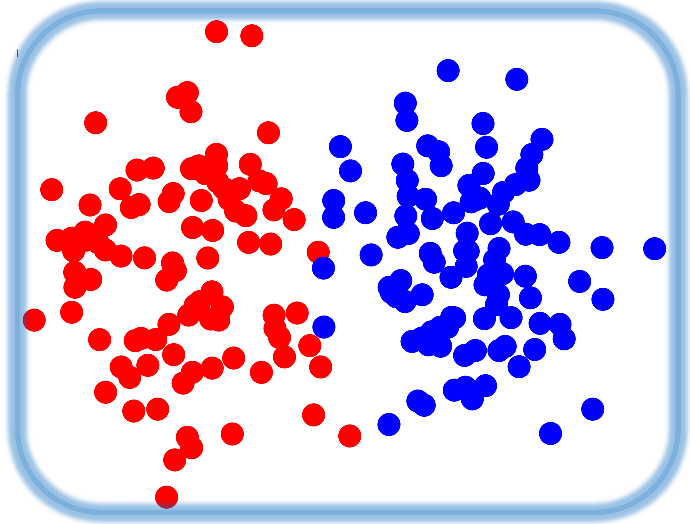
Statistical limits of planted clustering

- Clustered data + random noise
  - High dim data / large graphs
- When can we say that there are clusters?



Information theoretic limit

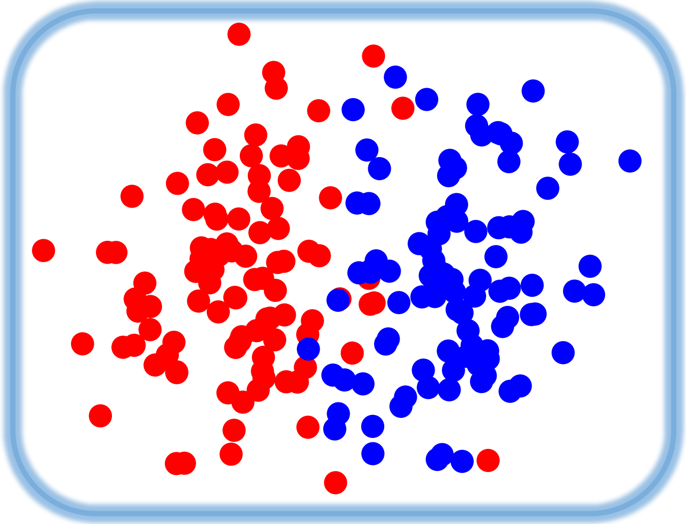
Computational limit



- When can algorithms find clusters?
  - Spectral algorithms
  - Kernel methods
  - Neural networks

Statistical limits of planted clustering

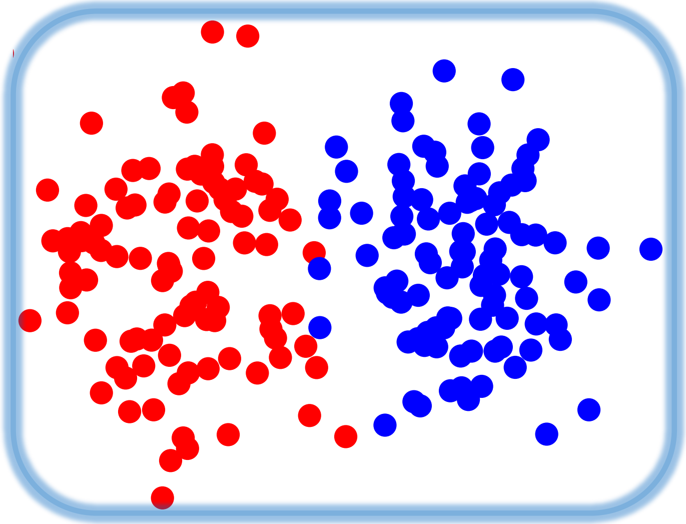
- Clustered data + random noise
  - High dim data / large graphs
- When can we say that there are clusters?



Information theoretic limit

≠

Computational limit



- When can algorithms find clusters?
  - Spectral algorithms
  - Kernel methods
  - Neural networks

Two-sample test for $m = 1$

- Given: G (with adjacency matrix A_G) and H (adjacency matrix A_H)

- Statistic $T = \|w \circ (A_G - A_H)\|_{spectral}$

$w \circ A =$ rescale matrix entries with suitable weights

$\|A\|_{spectral} =$ matrix spectral norm

- $$\lim_{n\to\infty} T = \begin{cases} \text{Tracy Widom variable} & \text{for } P = Q \\ \infty & \text{for large } \|P - Q\|_{spectral} \end{cases}.$$

Tracy-Widom-test: Say models are different if T large

- Result: Test has high accuracy for large n

- Variant of the test has high accuracy for large n or large m

Two-sample test for $m = 2$

- Given: G, G' (adjacency $A_G, A_{G'}$) and H, H' (adjacency $A_H, A_{H'}$)

- Statistic $T = w \sum_{(i,j)} \left((A_G)_{ij} - (A_H)_{ij} \right) \left((A_{G'})_{ij} - (A_{H'})_{ij} \right)$

- $\lim_{n \to \infty} T = \begin{cases} \text{standard normal} & \text{for } P = Q \\ \infty & \text{for large } \|P - Q\|_{Frobenius} \end{cases}$

Normal-test: Say models are different if $|T|$ large

- Result: Test has high accuracy for large n

- Variant of the test has high accuracy for large n or large m