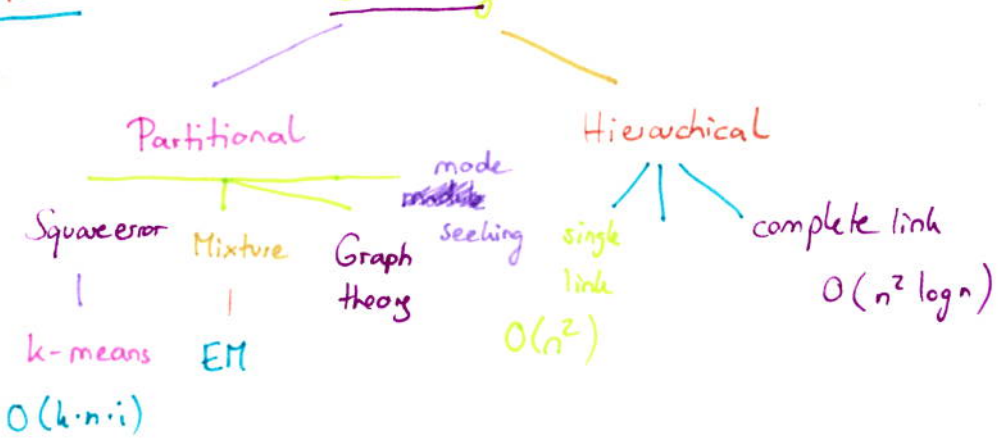
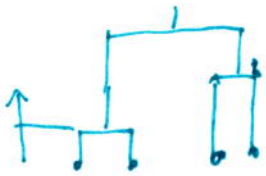
Model choices

Clustering



- Hard vs soft / fuzzy

- Det. vs Stoch.
- Monothetic vs Polythothic
- Applomeative vs divisive



:-)

Pattern repr

object $\longrightarrow \mathbb{R}^d$

1 Remove stopwords (in some way)

2 tf - idf

tf = indoc.

idf = $\log\left(\frac{\#\text{doc}}{\#\text{doc with the word}}\right)$

$\#\text{occ word}$ / $\#$ words in the doc.

3 $\text{dist}^{(d_1,d_2)} = \arccos\left(\frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}\right)$