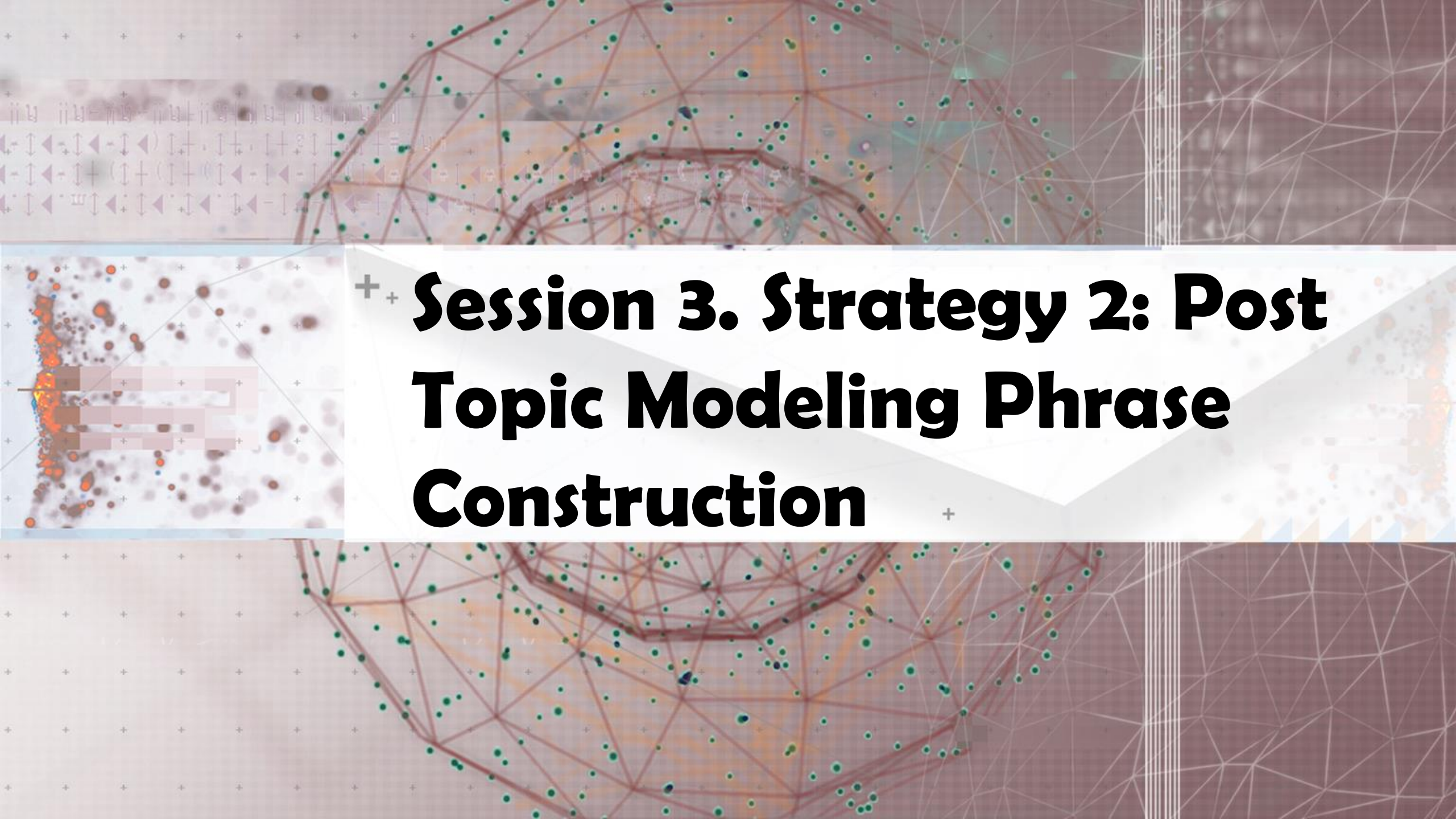


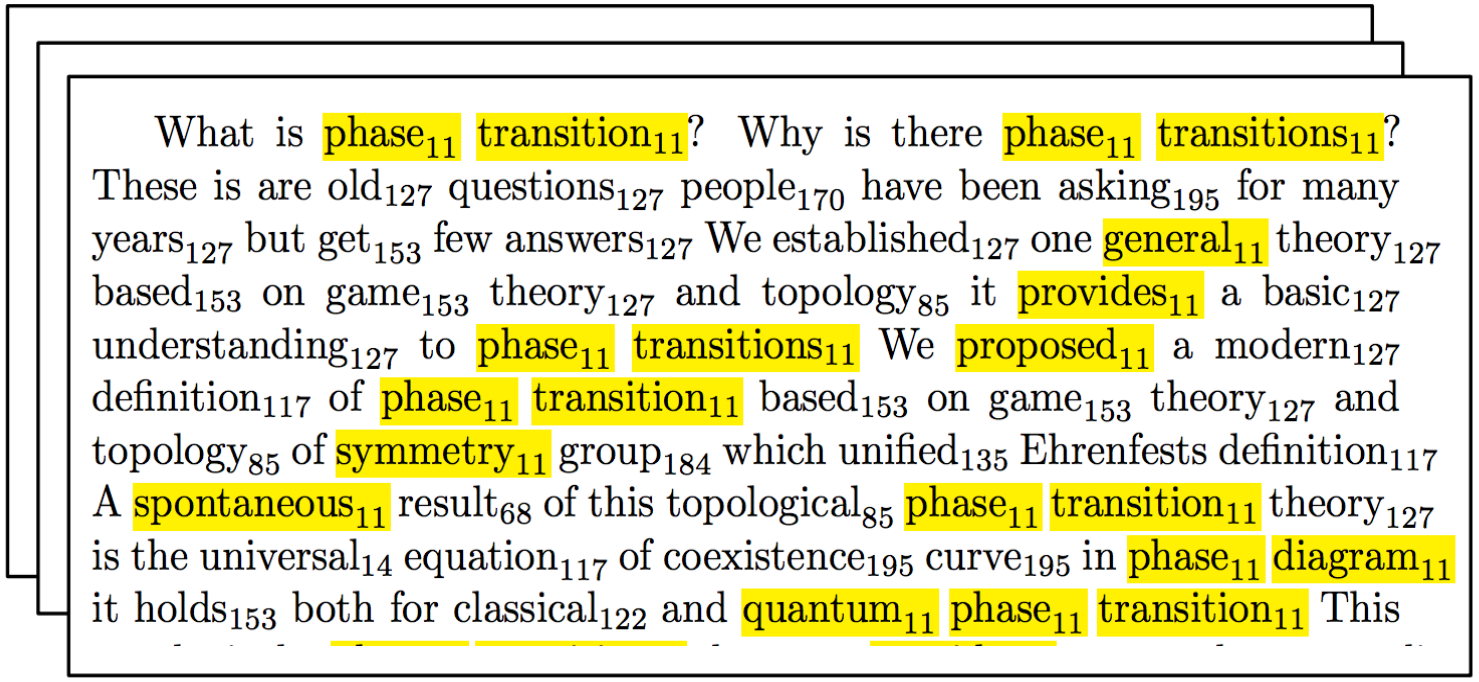
Session 3. Strategy 2: Post Topic Modeling Phrase Construction

Strategy 2: Post Topic Modeling Phrase Construction

- **TurboTopics** [Blei & Lafferty'09] – Phrase construction as a post-processing step to Latent Dirichlet Allocation
  - Perform Latent Dirichlet Allocation on corpus to assign each token a topic label
  - Merge adjacent unigrams with the same topic label by a distribution-free permutation test on arbitrary-length back-off model
  - End recursive merging when all significant adjacent unigrams have been merged
- **KERT** [Danilevsky et al.'14] – Phrase construction as a post-processing step to Latent Dirichlet Allocation
  - Perform **frequent pattern mining** on each topic
  - Perform **phrase ranking** based on four different criteria

Example of TurboTopics

Annotated documents



What is phase$_{11}$ transition$_{11}$? Why is there phase$_{11}$ transitions$_{11}$? These is are old$_{127}$ questions$_{127}$ people$_{170}$ have been asking$_{195}$ for many years$_{127}$ but get$_{153}$ few answers$_{127}$ We established$_{127}$ one general$_{11}$ theory$_{127}$ based$_{153}$ on game$_{153}$ theory$_{127}$ and topology$_{85}$ it provides$_{11}$ a basic$_{127}$ understanding$_{127}$ to phase$_{11}$ transitions$_{11}$ We proposed$_{11}$ a modern$_{127}$ definition$_{117}$ of phase$_{11}$ transition$_{11}$ based$_{153}$ on game$_{153}$ theory$_{127}$ and topology$_{85}$ of symmetry$_{11}$ group$_{184}$ which unified$_{135}$ Ehrenfests definition$_{117}$ A spontaneous$_{11}$ result$_{68}$ of this topological$_{85}$ phase$_{11}$ transition$_{11}$ theory$_{127}$ is the universal$_{14}$ equation$_{117}$ of coexistence$_{195}$ curve$_{195}$ in phase$_{11}$ diagram$_{11}$ it holds$_{153}$ both for classical$_{122}$ and quantum$_{11}$ phase$_{11}$ transition$_{11}$ This

LDA topic #11

phase, transitions, phases, transition, quantum, critical, symmetry, field, point, model, order, diagram, systems, two, theory, system, study, breaking, spin, first

Turbo topic #11

phase transitions, model, symmetry, point, quantum, systems, phase transition, phase diagram, system, order, field, order, parameter, critical, two, transitions in, models, different, symmetry breaking, first order, phenomena

- Perform LDA on corpus to assign each token a topic label
  - E.g., … phase$_{11}$ transition$_{11}$ …. game$_{153}$ theory$_{127}$ …
- Then merge adjacent unigrams with same topic label

KERT: Topical Keyphrase Extraction & Ranking

[Danilevsky, et al. 2014]

knowledge discovery using least squares support vector machine classifiers

support vectors for reinforcement learning

a hybrid approach to feature selection

pseudo conditional random fields

automatic web page classification in a dynamic and hierarchical way

inverse time dependency in convex regularized learning

postprocessing decision trees to extract actionable knowledge

variance minimization least squares support vector machines

…

Unigram topic assignment: Topic 1 & Topic 2

learning
support vector machines
reinforcement learning
feature selection
conditional random fields
classification
decision trees
:

Topical keyphrase extraction & ranking

Framework of KERT

1. Run bag-of-words model inference and assign topic label to each token
2. Extract candidate keyphrases within each topic

Frequent pattern mining

3. Rank the keyphrases in each topic
   - Popularity: 'information retrieval' vs. 'cross-language information retrieval'
   - Discriminativeness: only frequent in documents about topic t
   - Concordance: 'active learning' vs. 'learning classification'
   - Completeness: 'vector machine' vs. 'support vector machine'

Comparability property: directly compare phrases of mixed lengths

KERT: Topical Phrases on Machine Learning

Top-Ranked Phrases by Mining Paper Titles in DBLP

kpRel [Zhao et al. 11]	KERT (-popularity)	KERT (-discriminativeness)	KERT (-concordance)	KERT [Danilevsky et al. 14]
learning	effective	support vector machines	learning	learning
classification	text	feature selection	classification	support vector machines
selection	probabilistic	reinforcement learning	selection	reinforcement learning
models	identification	conditional random fields	feature	feature selection
algorithm	mapping	constraint satisfaction	decision	conditional random fields
features	task	decision trees	bayesian	classification
decision	planning	dimensionality reduction	trees	decision trees
:	:	:	:	:

The topic that represents the area of Machine Learning