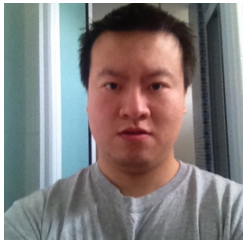
Stony Brook University
Computer Science

External Evaluation of Topic Models: A Graph Mining Approach

Hau Chan*



Leman Akoglu

Stony Brook University

ICDM 2013 December 7-10, 2013

Topic Models

Stony Brook University
Computer Science

Document Corpora

T (topics)

Topic Model (e.g. LDA)

Topic 1: $w_{11}, \ldots, w_{1N}$
Topic 2: $w_{21}, \ldots, w_{2N}$
Topic 3: $w_{31}, \ldots, w_{3N}$
...
...
...
Topic T: $w_{T1}, \ldots, w_{TN}$

EXAMPLE TOPICS T1 (HIGH-QUALITY) AND T2 (LOW-QUALITY) OF A TOPIC MODEL.

T1:	steam, engine, valve, piston, cylinder, pressure, boiler, air, pump, pipe
T2:	cut, system, capital, pointed, opening, building, character, round, france, paris

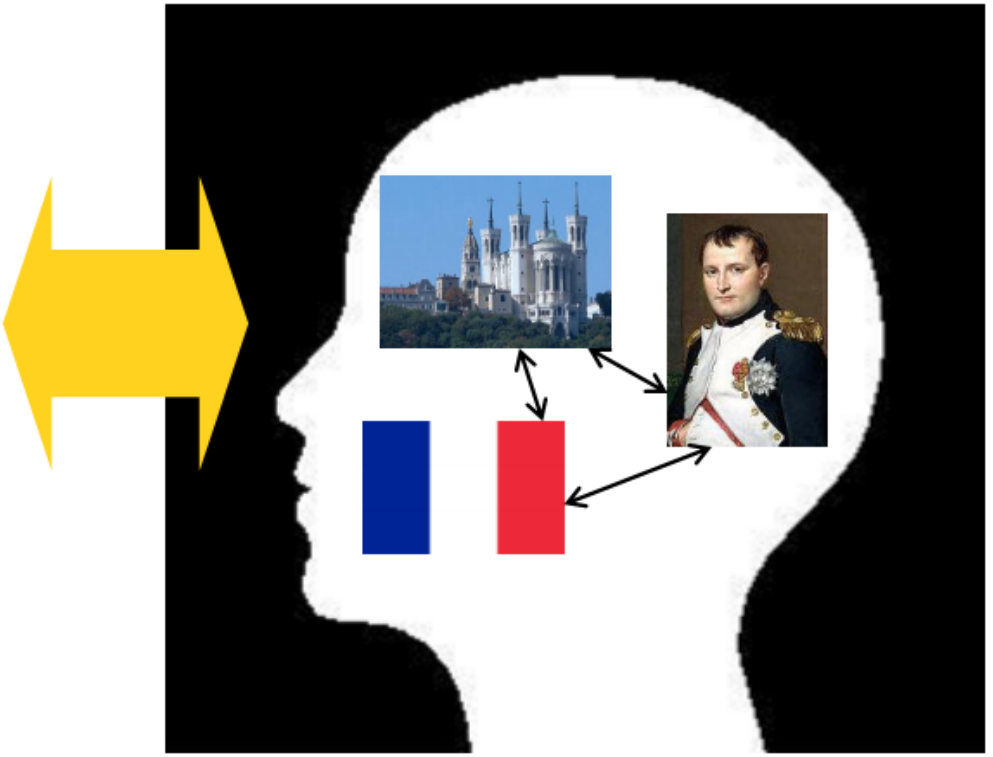
Our goal: distinguish **good** topics from **poor** topics!

Motivation

Stony Brook University
Computer Science

- Negative correlation (!) between
  - Human evaluation & Statistical evaluation [Chang+ NIPS'09]
- Applications where human-perceived quality essential:
  - doc-doc similarity (via topic distribution)
  - word-sense disambiguation
  - multi-doc summarization

Main idea: exploit Wikipedia

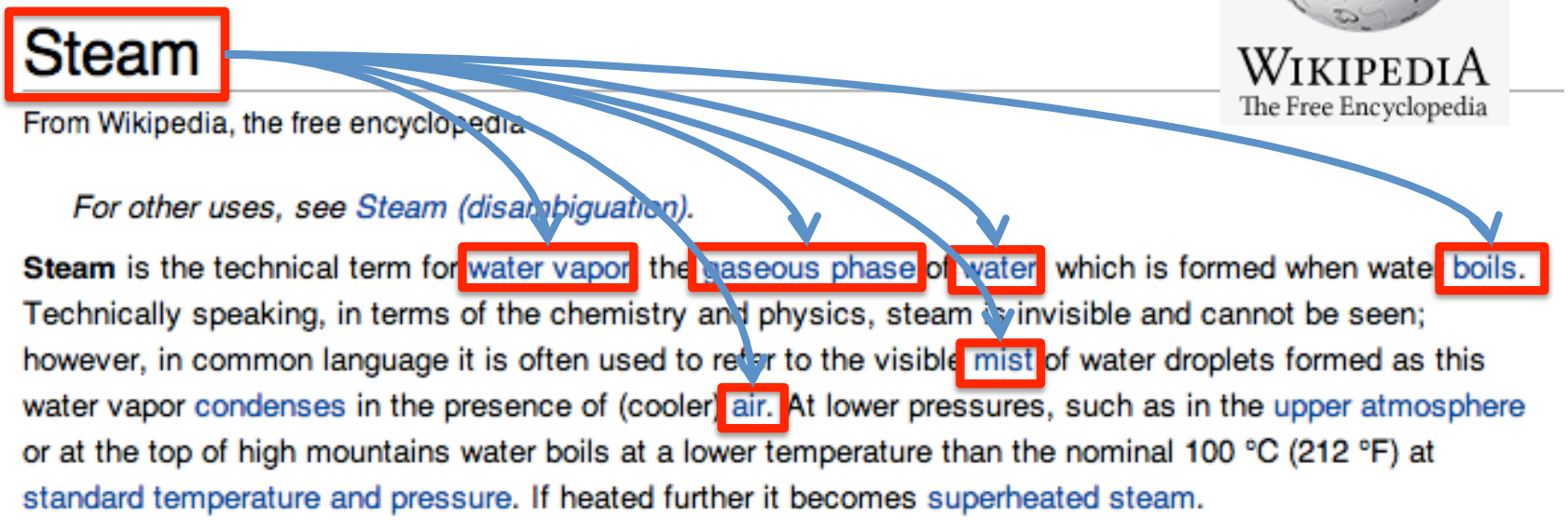


Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

[West-Leskovec, 2012]

Main idea: exploit Wikipedia



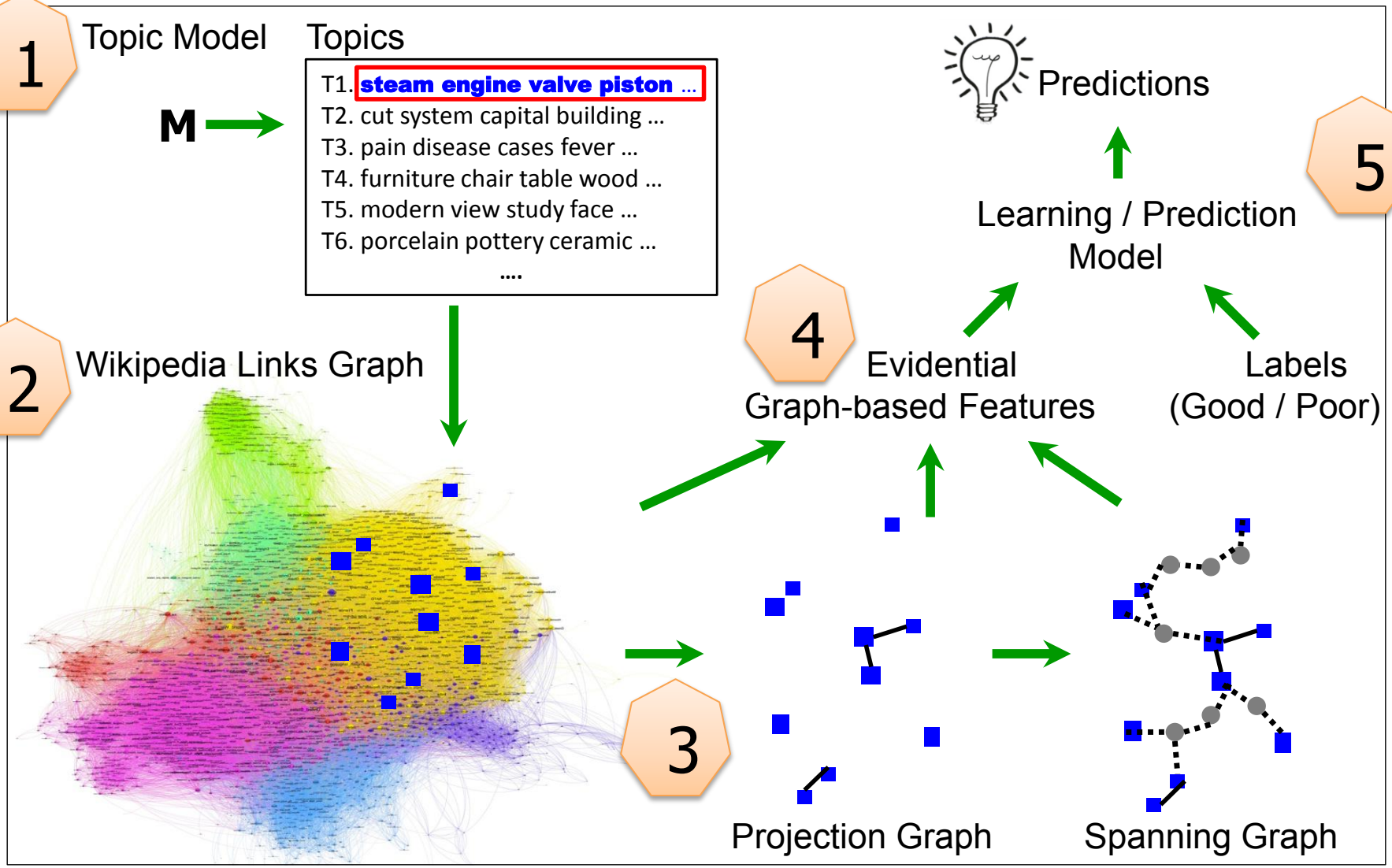
Wikipedia is a graph!

node = entity & (directed) edge = pagelink

Intuition: Good topic words are conceptually "coherent" → "close-by" in WikiLinks

Our Approach

Stony Brook University
Computer Science



Projection Graph

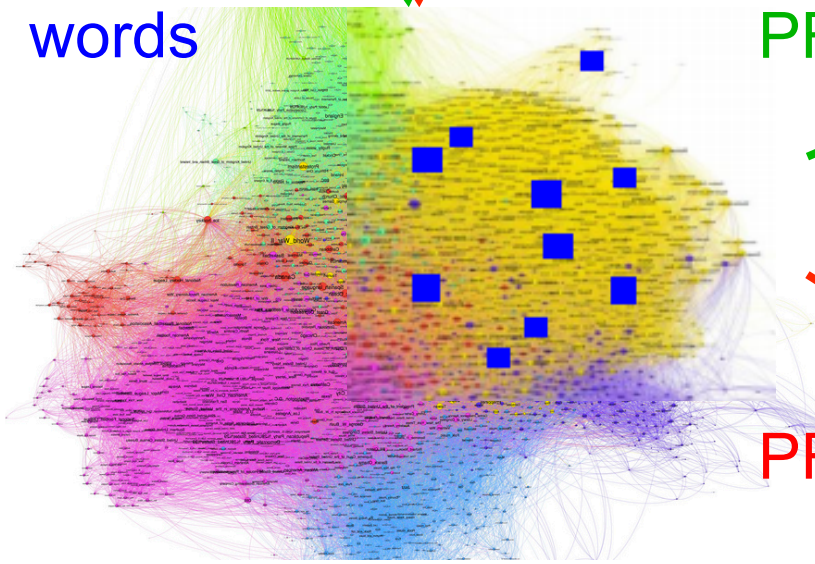
Stony Brook University
Computer Science

EXAMPLE TOPICS T1 (HIGH-QUALITY) AND T2 (LOW-QUALITY) OF A TOPIC MODEL.

T1:	steam, engine, valve, piston, cylinder, pressure, boiler, air, pump, pipe
T2:	cut, system, capital, pointed, opening, building, character, round, france, paris

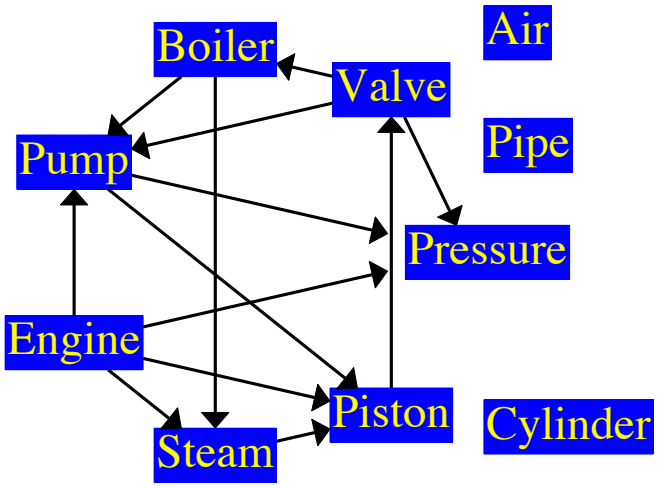
match T2 words

match T1 words

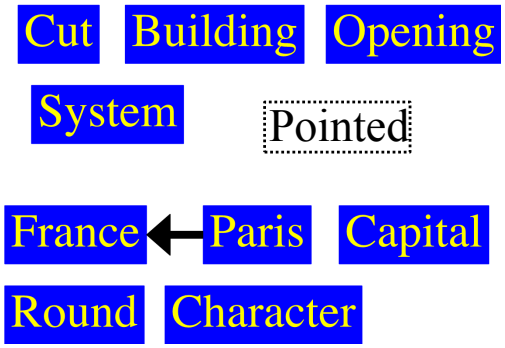


WikiLinks Graph

PROJ T1

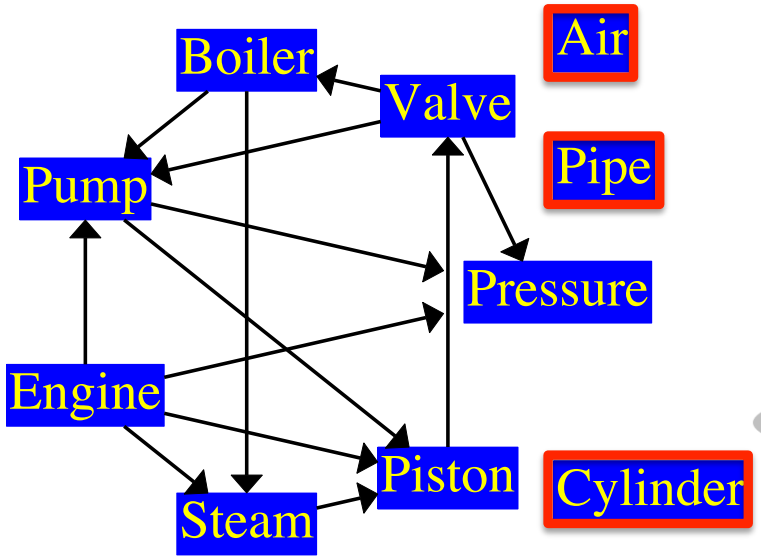


PROJ T2



Spanning Graph

Stony Brook University
Computer Science

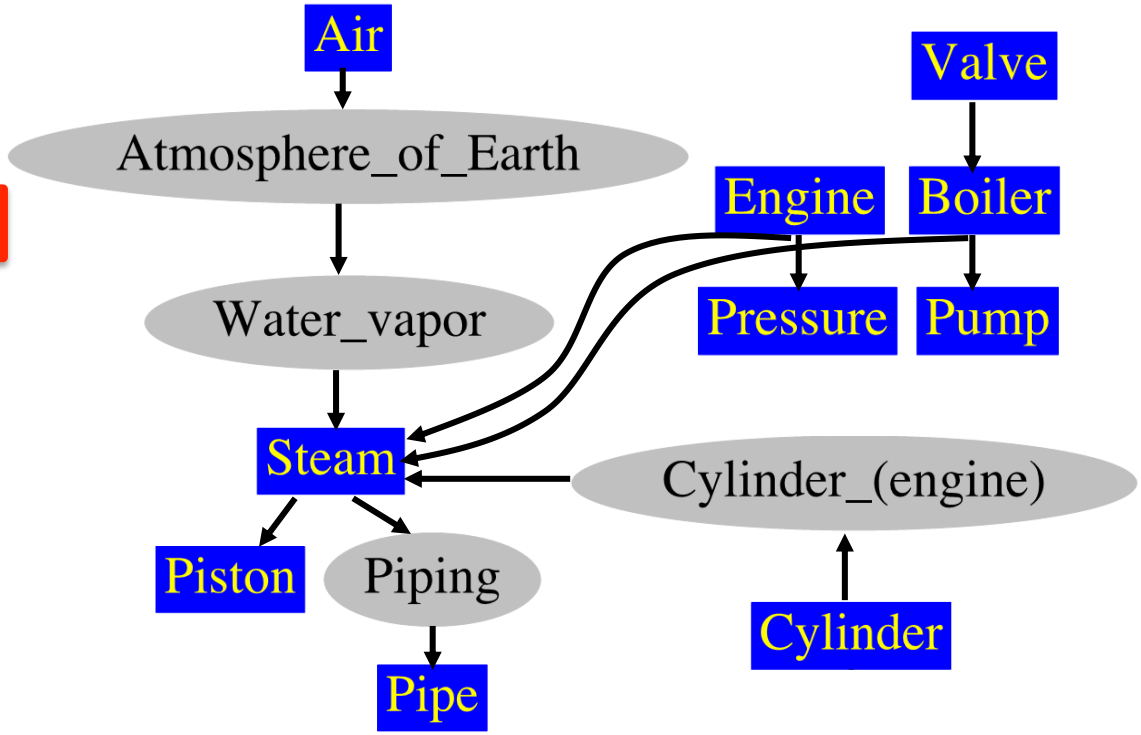


PROJ T1

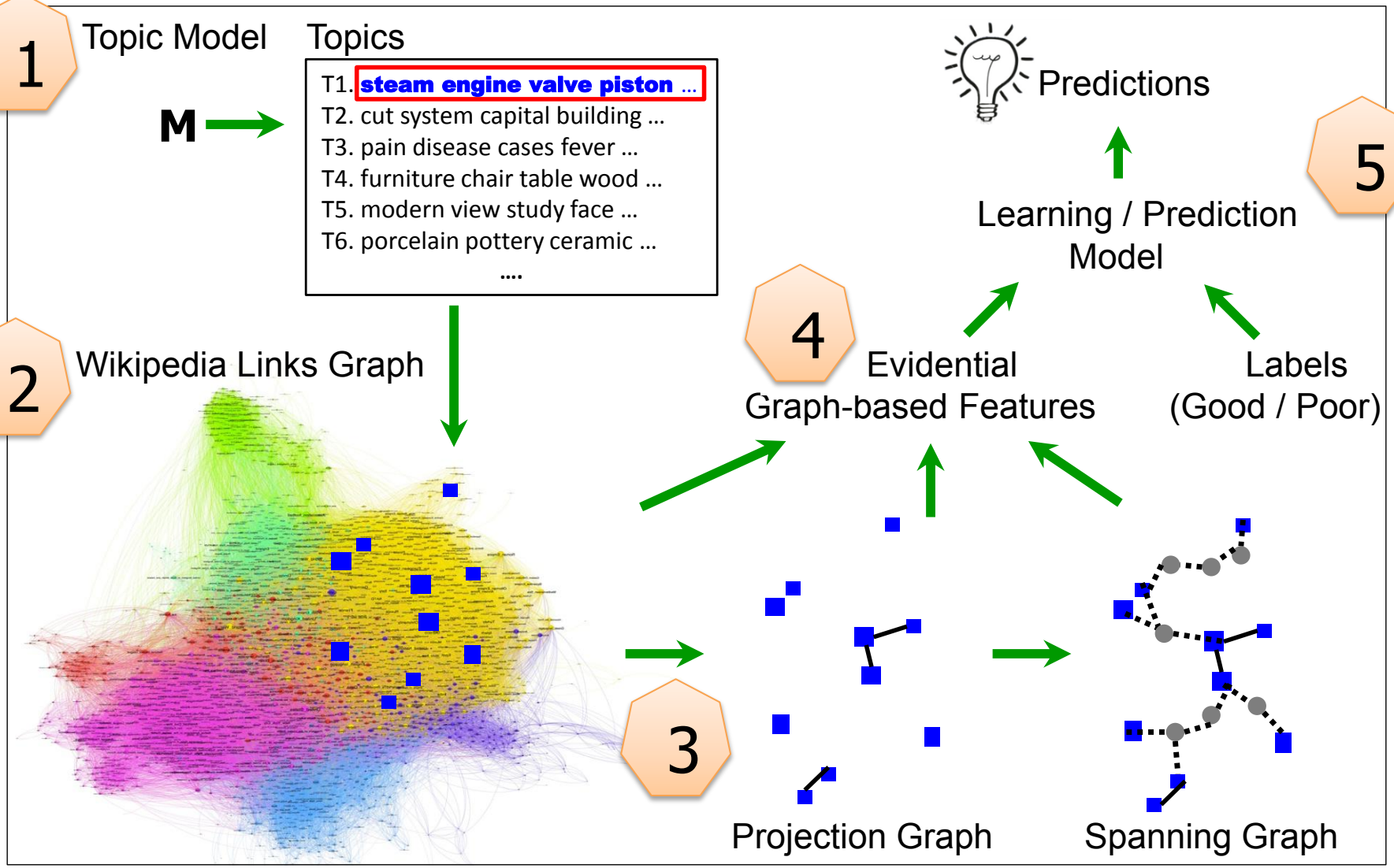
WikiLinks induced on topic-words

SPAN T1

Connection subgraph of topic-words



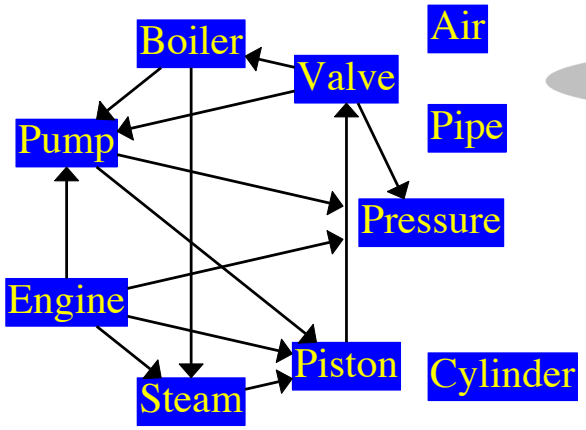
Our Approach



All Features: 3 groups

Stony Brook University
Computer Science

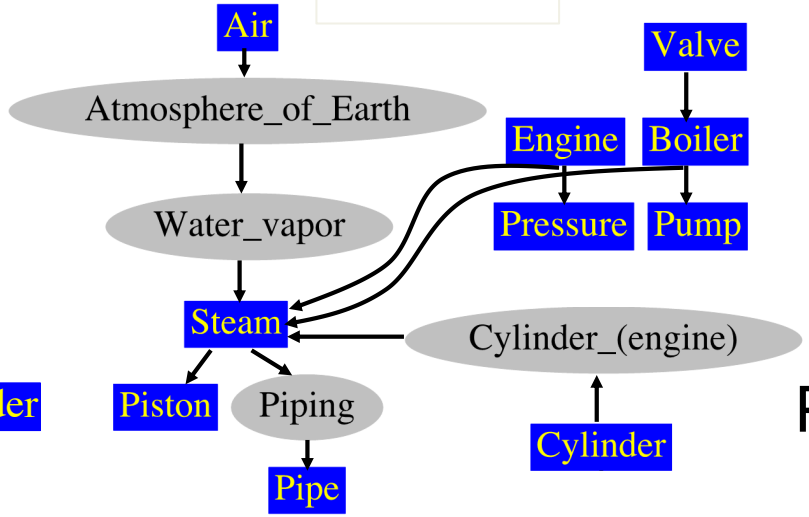
PROJ



4 features

e.g
Largest CC Size

SPAN



9 features

e.g
#Connector Nodes

SP

	Air	…	Pipe
Air	1		5
	Shortest Path		
Pipe	NP		1

9 features

e.g
Average SP

Labels: Good vs. Poor Topics

Stony Brook University
Computer Science

Dataset	# Documents	# Topics	Labels
Press	2,246	100	No
Brain Injury	10,000	200	No
Books	12,000	120	Yes
News	55,000	117	Yes

Relative Quality Prediction (Labels – Generated)

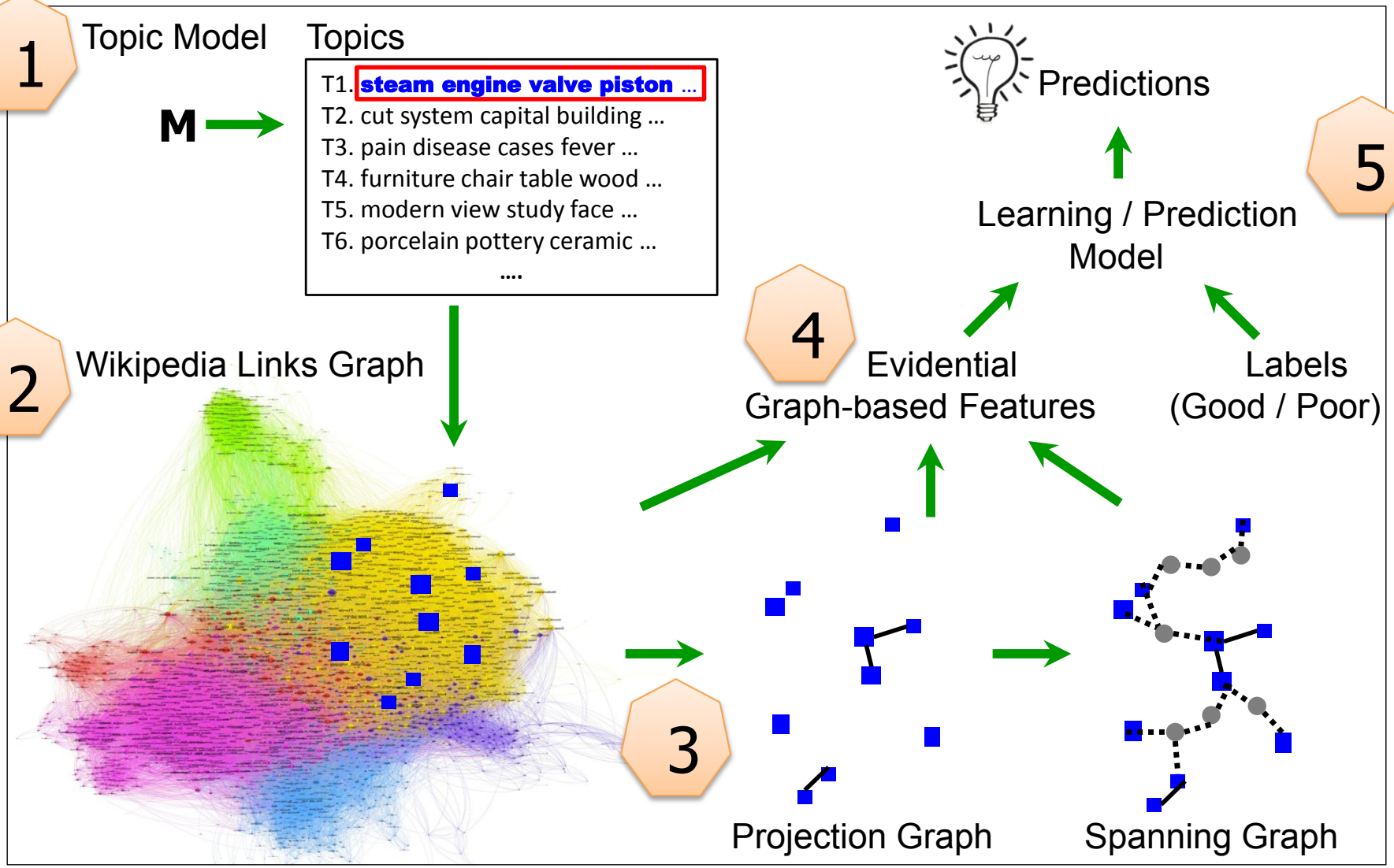
T topics; W words; Top 10 vs Top X-Y

$w_1, w_2, \ldots, w_{10}, \ldots, w_{11},\ldots, w_{20}, \ldots, w_{31},\ldots, w_{40}, \ldots, w_{91},\ldots, w_{100}$

Absolute Quality Prediction (Labels – Human Annotators)

*We thank David Newman and his group for sharing these data

Our Approach



Relative Quality Prediction

Stony Brook University
Computer Science

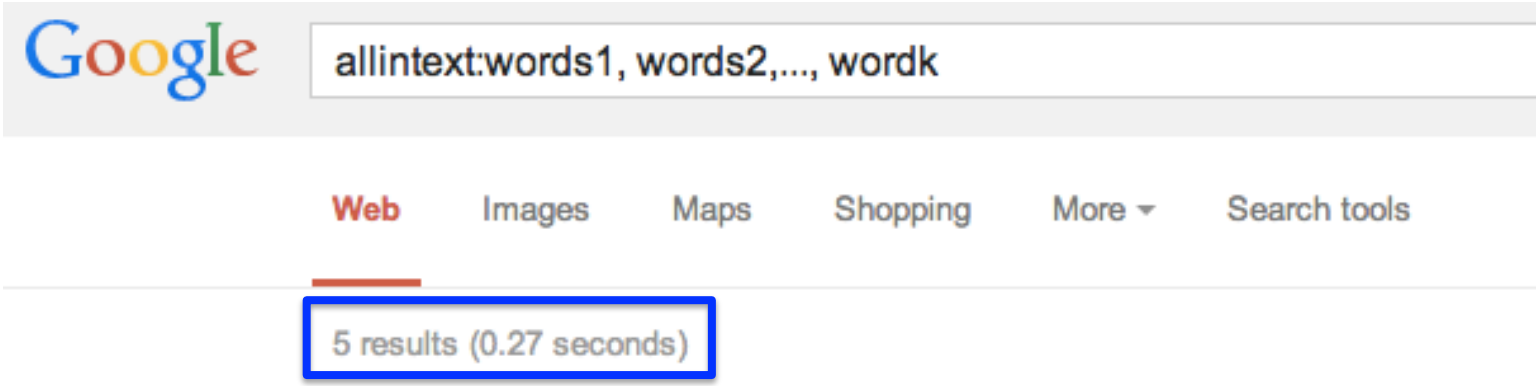
Feature set	top-10 vs. $w_1, w_2, \ldots, w_{10}$	top-[11-20] $w_{11},\ldots, w_{20}$		top-[31-40] $w_{31},\ldots, w_{40}$		top-[91-100] $w_{91},\ldots, w_{100}$	
		PRESS	BRAIN	PRESS	BRAIN	PRESS	BRAIN
BASELINE-MAJORITY		0.500	0.500	0.500	0.500	0.500	0.500
PROJ		0.505	0.622	0.715	0.705	0.765	0.725
D-SPAN		0.650	0.687	0.760	0.740	0.805	0.762
D-SP		0.605	0.665	0.710	0.760	0.750	0.790
PROJ+D-SPAN		0.650	0.687	0.745	0.722	0.790	0.777
PROJ+D-SP		0.650	0.672	0.710	0.752	0.815	0.800
PROJ+D-SPAN+D-SP		0.660	0.687	0.735	0.752	0.810	0.807

>15% >23% >30%

improvement over random baseline

Consistent higher accuracy for easier relative tasks

Absolute Quality Prediction – Baselines



Baseline (1)

Google Features:
(1) allintext
(2) intitle
(3) inanchor
(4) inurl

Baseline (2)

Personalize PageRank:
(1) average pairwise score
(2) median pairwise score
(3) average pairwise rank
(4) median pairwise rank

Absolute Quality Prediction

Stony Brook University
Computer Science

Feature set	BOOKS	NEWS	BOOKS +NEWS
BASELINE-MAJORITY	0.610	0.521	0.549
BASELINE-GOOGLE	0.642	0.624	0.629
BASELINE-PPR	0.842	0.735	0.785
PROJ	0.875	0.812	0.848
D-SPAN	0.892	0.769	0.844
D-SP	0.883	0.786	0.852
PROJ+D-SPAN	0.883	0.795	0.844
PROJ+D-SP	0.892	0.795	0.848
PROJ+D-SPAN+D-SP	0.900	0.821	0.831

6% – 9%
improvement over ALL baselines

Cross-Domain Prediction

Stony Brook University
Computer Science

Train \ Test	BOOKS	NEWS
BOOKS	0.900	0.769
NEWS	0.867	0.821

Our graph-centric features are domain-independent
(only based on "graph closeness")

Learned Coefficients

Stony Brook University
Computer Science

Selected Feature	Coef: BOOKS	Coef: NEWS
$g_M NumMiss$	0.0626	0.0918
$g_S RatioC$	0.2940	0.5909
$g_M MaxDeg$	−0.2921	−0.4541
$g_M SizeMaxComp$	−0.8667	– – ––
$g_S AvgMSTWeight$	– – ––	0.2598
$NumSP2$	−0.9685	– – ––

Good topics:

- Fewer missing (matched) words on WikiLinks
- Fewer connector nodes (in spanning graph)
- Higher maximum degree (in projection graph)

Thank you!

- SBU Office of the Vice President for Research
- NSF Graduate Research Fellowship
- ICDM 2013 Travel Grants

hauchan@cs.stonybrook.edu

http://www.cs.sunysb.edu/~hauchan

4. Prediction

- For relative quality prediction, we consider topics with 20, 40, and 100 words (ordered by descending probability of describing the topics) where the topics are generated by LDA
  - Positive Class: top 10 words for each topics
  - Negative Class: Bottom 11-20, Bottom 31-40, and Bottom 91-100 words for each topics
- For absolute quality prediction
  - Positive Class: Good topics
  - Negative Class: Bad topics

5. Learning Model

- Logistics Regression Classifier
  - L1-Norm Regularization
- We report Leave-One-Out Cross-Validation (LOOCV)