Data Search using Deep Learning
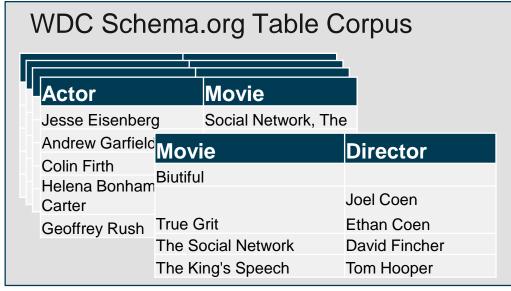
The Web is a rich source for tables:

- Many tables describe the same real world entities

- Most tables contain only partial information

- Tables are scattered across different websites

Table Augmentation

- Augment an input table with information from a table corpus

- Task not trivial due to heterogeneity and size of table corpus

Movie	Premiere	Director	Actors
Biutiful	05/17/2010	?	?
True Grit	12/22/2010	?	?
The Social Network	10/01/2010	?	?
The King's Speech	1/07/2010	?	?
127 Hours	09/04/2010	?	?
The Fighter	12/10/2010	?	?
?	?	?	?

WDC Schema.org Table Corpus

Actor	Movie
Jesse Eisenberg	Social Network, The
Andrew Garfield	
Colin Firth	
Helena Bonham Carter	
Geoffrey Rush	

Movie	Director
Biutiful	
	Joel Coen
True Grit	Ethan Coen
The Social Network	David Fincher
The King's Speech	Tom Hooper

How can we use transformer-based models for Table Augmentation?

Data Search using Deep Learning

Project Goal
Experiment with State-of-the-Art NLP Transformer Models
and use them to search for Tabular Data

Involves
– Data Profiling, Data Preprocessing

– Model Training, Evaluation, Selection

Learning Targets
– Gain technical experience with

    State-of-the-Art Data Search Technologies

– Gain work experience as Data Scientist

Requirements
– Data Science & Engineering Skills, Programming Experience (Python)

– Relevant Courses: Web Data Integration, Data Mining I & II, Information Retrieval & Web Search

Organization: 4-6 people, 6 months, work as a complete team and in subgroups
Instructors: Alexander Brinkmann, Christian Bizer