pdf2gtfs: Timetable Extraction from PDF Files

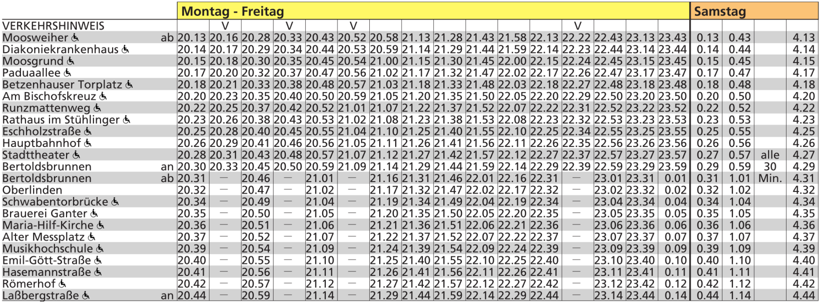
Bachelor's Thesis Presentation

Julius Heinzinger

Faculty of Engineering
University of Freiburg

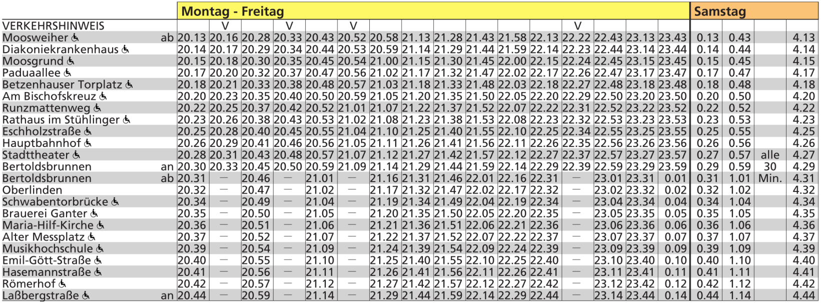
July 2023

Input: PDF Timetable



		Montag - Freitag																Samstag			
VERKEHRSHINWEIS			V		V		V							V							
Moosweiher ♿	ab	20.13	20.16	20.28	20.33	20.43	20.52	20.58	21.13	21.28	21.43	21.58	22.13	22.22	22.43	23.13	23.43	0.13	0.43		4.13
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34	20.44	20.53	20.59	21.14	21.29	21.44	21.59	22.14	22.23	22.44	23.14	23.44	0.14	0.44		4.14
Moosgrund ♿		20.15	20.18	20.30	20.35	20.45	20.54	21.00	21.15	21.30	21.45	22.00	22.15	22.24	22.45	23.15	23.45	0.15	0.45		4.15
Paduaallee ♿		20.17	20.20	20.32	20.37	20.47	20.56	21.02	21.17	21.32	21.47	22.02	22.17	22.26	22.47	23.17	23.47	0.17	0.47		4.17
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38	20.48	20.57	21.03	21.18	21.33	21.48	22.03	22.18	22.27	22.48	23.18	23.48	0.18	0.48		4.18
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40	20.50	20.59	21.05	21.20	21.35	21.50	22.05	22.20	22.29	22.50	23.20	23.50	0.20	0.50		4.20
Runzmattenweg ♿		20.22	20.25	20.37	20.42	20.52	21.01	21.07	21.22	21.37	21.52	22.07	22.22	22.31	22.52	23.22	23.52	0.22	0.52		4.22
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43	20.53	21.02	21.08	21.23	21.38	21.53	22.08	22.23	22.32	22.53	23.23	23.53	0.23	0.53		4.23
Eschholzstraße ♿		20.25	20.28	20.40	20.45	20.55	21.04	21.10	21.25	21.40	21.55	22.10	22.25	22.34	22.55	23.25	23.55	0.25	0.55		4.25
Hauptbahnhof ♿		20.26	20.29	20.41	20.46	20.56	21.05	21.11	21.26	21.41	21.56	22.11	22.26	22.35	22.56	23.26	23.56	0.26	0.56		4.26
Stadttheater ♿		20.28	20.31	20.43	20.48	20.57	21.07	21.12	21.27	21.42	21.57	22.12	22.27	22.37	22.57	23.27	23.57	0.27	0.57	alle	4.27
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50	20.59	21.09	21.14	21.29	21.44	21.59	22.14	22.29	22.39	22.59	23.29	23.59	0.29	0.59	30	4.29
Bertoldsbrunnen	ab	20.31	—	20.46	—	21.01	—	21.16	21.31	21.46	22.01	22.16	22.31	—	23.01	23.31	0.01	0.31	1.01	Min.	4.31
Oberlinden		20.32	—	20.47	—	21.02	—	21.17	21.32	21.47	22.02	22.17	22.32	—	23.02	23.32	0.02	0.32	1.02		4.32
Schwabentorbrücke ♿		20.34	—	20.49	—	21.04	—	21.19	21.34	21.49	22.04	22.19	22.34	—	23.04	23.34	0.04	0.34	1.04		4.34
Brauerei Ganter ♿		20.35	—	20.50	—	21.05	—	21.20	21.35	21.50	22.05	22.20	22.35	—	23.05	23.35	0.05	0.35	1.05		4.35
Maria-Hilf-Kirche ♿		20.36	—	20.51	—	21.06	—	21.21	21.36	21.51	22.06	22.21	22.36	—	23.06	23.36	0.06	0.36	1.06		4.36
Alter Messplatz ♿		20.37	—	20.52	—	21.07	—	21.22	21.37	21.52	22.07	22.22	22.37	—	23.07	23.37	0.07	0.37	1.07		4.37
Musikhochschule ♿		20.39	—	20.54	—	21.09	—	21.24	21.39	21.54	22.09	22.24	22.39	—	23.09	23.39	0.09	0.39	1.09		4.39
Emil-Gött-Straße ♿		20.40	—	20.55	—	21.10	—	21.25	21.40	21.55	22.10	22.25	22.40	—	23.10	23.40	0.10	0.40	1.10		4.40
Hasemannstraße ♿		20.41	—	20.56	—	21.11	—	21.26	21.41	21.56	22.11	22.26	22.41	—	23.11	23.41	0.11	0.41	1.11		4.41
Römerhof ♿		20.42	—	20.57	—	21.12	—	21.27	21.42	21.57	22.12	22.27	22.42	—	23.12	23.42	0.12	0.42	1.12		4.42
Laßbergstraße ♿	an	20.44	—	20.59	—	21.14	—	21.29	21.44	21.59	22.14	22.29	22.44	—	23.14	23.44	0.14	0.44	1.14		4.44

Input: PDF Timetable



		Montag - Freitag																Samstag			
VERKEHRSHINWEIS			V		V		V							V							
Moosweiher ♿	ab	20.13	20.16	20.28	20.33	20.43	20.52	20.58	21.13	21.28	21.43	21.58	22.13	22.22	22.43	23.13	23.43	0.13	0.43		4.13
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34	20.44	20.53	20.59	21.14	21.29	21.44	21.59	22.14	22.23	22.44	23.14	23.44	0.14	0.44		4.14
Moosgrund ♿		20.15	20.18	20.30	20.35	20.45	20.54	21.00	21.15	21.30	21.45	22.00	22.15	22.24	22.45	23.15	23.45	0.15	0.45		4.15
Paduaallee ♿		20.17	20.20	20.32	20.37	20.47	20.56	21.02	21.17	21.32	21.47	22.02	22.17	22.26	22.47	23.17	23.47	0.17	0.47		4.17
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38	20.48	20.57	21.03	21.18	21.33	21.48	22.03	22.18	22.27	22.48	23.18	23.48	0.18	0.48		4.18
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40	20.50	20.59	21.05	21.20	21.35	21.50	22.05	22.20	22.29	22.50	23.20	23.50	0.20	0.50		4.20
Runzmattenweg ♿		20.22	20.25	20.37	20.42	20.52	21.01	21.07	21.22	21.37	21.52	22.07	22.22	22.31	22.52	23.22	23.52	0.22	0.52		4.22
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43	20.53	21.02	21.08	21.23	21.38	21.53	22.08	22.23	22.32	22.53	23.23	23.53	0.23	0.53		4.23
Eschholzstraße ♿		20.25	20.28	20.40	20.45	20.55	21.04	21.10	21.25	21.40	21.55	22.10	22.25	22.34	22.55	23.25	23.55	0.25	0.55		4.25
Hauptbahnhof ♿		20.26	20.29	20.41	20.46	20.56	21.05	21.11	21.26	21.41	21.56	22.11	22.26	22.35	22.56	23.26	23.56	0.26	0.56		4.26
Stadttheater ♿		20.28	20.31	20.43	20.48	20.57	21.07	21.12	21.27	21.42	21.57	22.12	22.27	22.37	22.57	23.27	23.57	0.27	0.57	alle	4.27
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50	20.59	21.09	21.14	21.29	21.44	21.59	22.14	22.29	22.39	22.59	23.29	23.59	0.29	0.59	30	4.29
Bertoldsbrunnen	ab	20.31	–	20.46	–	21.01	–	21.16	21.31	21.46	22.01	22.16	22.31	–	23.01	23.31	0.01	0.31	1.01	Min.	4.31
Oberlinden		20.32	–	20.47	–	21.02	–	21.17	21.32	21.47	22.02	22.17	22.32	–	23.02	23.32	0.02	0.32	1.02		4.32
Schwabentorbrücke ♿		20.34	–	20.49	–	21.04	–	21.19	21.34	21.49	22.04	22.19	22.34	–	23.04	23.34	0.04	0.34	1.04		4.34
Brauerei Ganter ♿		20.35	–	20.50	–	21.05	–	21.20	21.35	21.50	22.05	22.20	22.35	–	23.05	23.35	0.05	0.35	1.05		4.35
Maria-Hilf-Kirche ♿		20.36	–	20.51	–	21.06	–	21.21	21.36	21.51	22.06	22.21	22.36	–	23.06	23.36	0.06	0.36	1.06		4.36
Alter Messplatz ♿		20.37	–	20.52	–	21.07	–	21.22	21.37	21.52	22.07	22.22	22.37	–	23.07	23.37	0.07	0.37	1.07		4.37
Musikhochschule ♿		20.39	–	20.54	–	21.09	–	21.24	21.39	21.54	22.09	22.24	22.39	–	23.09	23.39	0.09	0.39	1.09		4.39
Emil-Gött-Straße ♿		20.40	–	20.55	–	21.10	–	21.25	21.40	21.55	22.10	22.25	22.40	–	23.10	23.40	0.10	0.40	1.10		4.40
Hasemannstraße ♿		20.41	–	20.56	–	21.11	–	21.26	21.41	21.56	22.11	22.26	22.41	–	23.11	23.41	0.11	0.41	1.11		4.41
Römerhof ♿		20.42	–	20.57	–	21.12	–	21.27	21.42	21.57	22.12	22.27	22.42	–	23.12	23.42	0.12	0.42	1.12		4.42
Laßbergstraße ♿	an	20.44	–	20.59	–	21.14	–	21.29	21.44	21.59	22.14	22.29	22.44	–	23.14	23.44	0.14	0.44	1.14		4.44

- First problem: Table extraction from a PDF

Output: GTFS

- Output format: GTFS (= General Transit Feed Specification)
  - de-facto standard for transit data
  - GTFS feed: `.zip`-archive of different files
  - each file contains a specific part of the transit information

Output: GTFS

- Output format: GTFS (= General Transit Feed Specification)
  - de-facto standard for transit data
  - GTFS feed: `.zip`-archive of different files
  - each file contains a specific part of the transit information
- Excerpt of a `stops.txt`

stop_id	stop_name	stop_lat	stop_lon
de:08311:30800:0:1	Moosweiher	48.0288	7.8089
this_is_an_id_as_well	Hauptbahnhof	47.9967	7.8399
de:08311:30300:0:1	Laßbergstraße	47.9846	7.8937
...	...	...	...

  - `stop_id` is used to reference a stop in other files
  - location is required

Output: GTFS

- Output format: GTFS (= General Transit Feed Specification)
  - de-facto standard for transit data
  - GTFS feed: `.zip`-archive of different files
  - each file contains a specific part of the transit information
- Excerpt of a `stops.txt`

stop_id	stop_name	stop_lat	stop_lon
de:08311:30800:0:1	Moosweiher	48.0288	7.8089
this_is_an_id_as_well	Hauptbahnhof	47.9967	7.8399
de:08311:30300:0:1	Laßbergstraße	47.9846	7.8937
...	...	...	...

  - `stop_id` is used to reference a stop in other files
  - location is required
- Second problem: Location detection

Table Extraction

Background & Approach

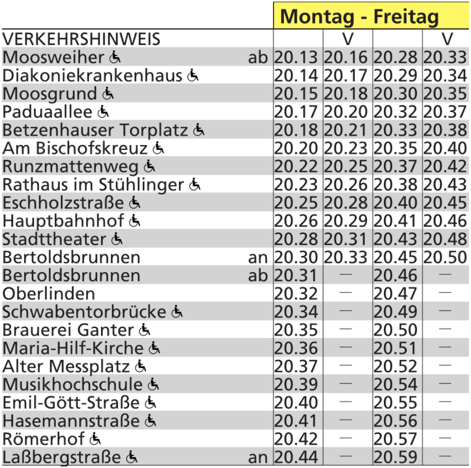
Table Extraction: Background 1/2

- A PDF file does not store plain text
  - stores position and other properties of text pieces
- Relation between different text pieces is lost
- Relevance of text is unclear
- We can extract characters or text fragments from a PDF with e.g., `pdfminer.six`

Table Extraction: Background 2/2

- Table consists of cells
- Cells contain one or more characters
- We define a celltype using content and other cells (e.g., Time, Stop, Day)
- Time cells easy to detect
  - simple, restrictive format

▸ More on cell types



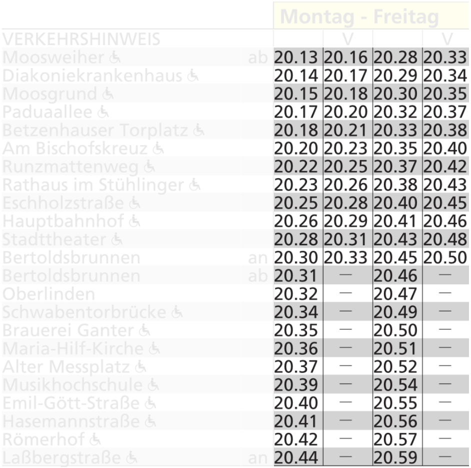
		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

- Idea: Use body (i.e., times) to detect the table
- Run basic type detection

Table Extraction: Approach

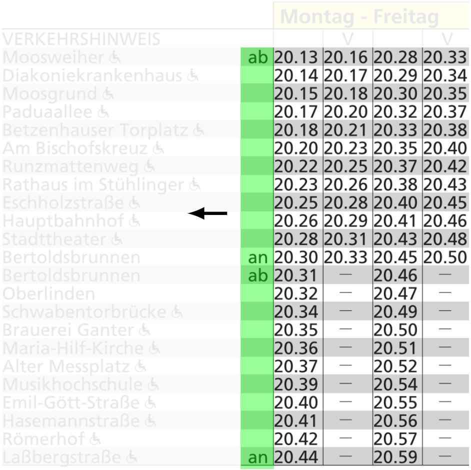
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

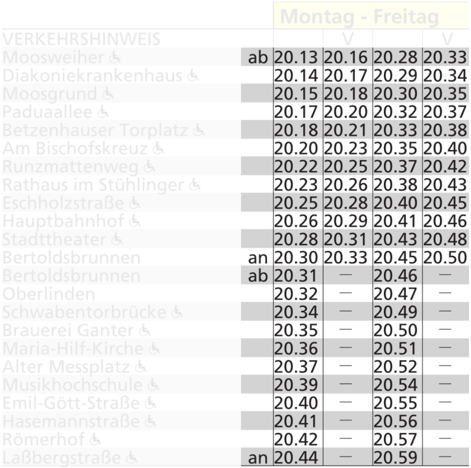
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

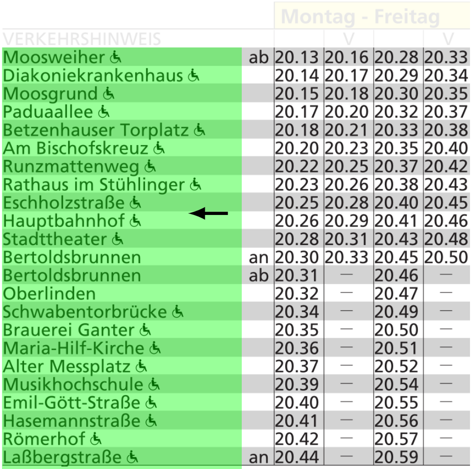
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

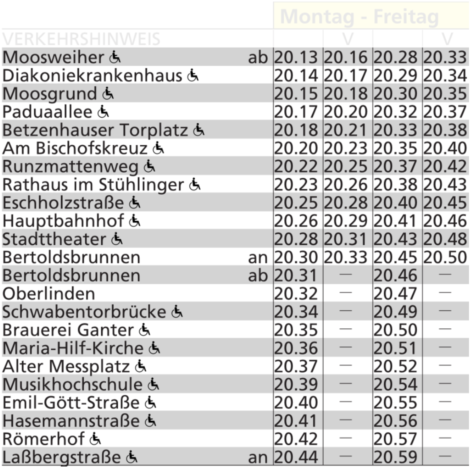
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿ ←		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

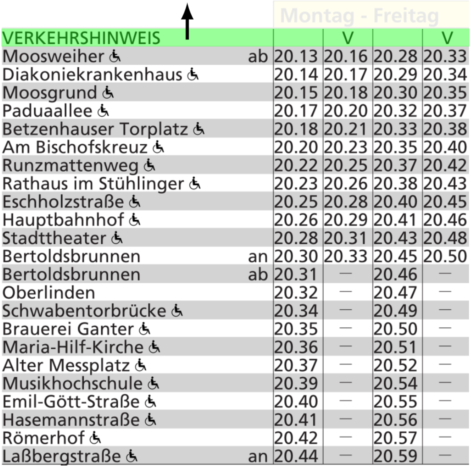
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

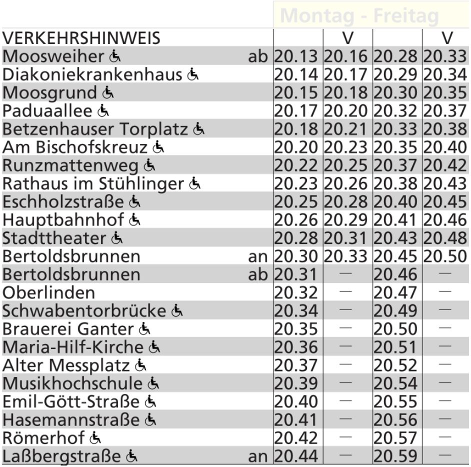
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

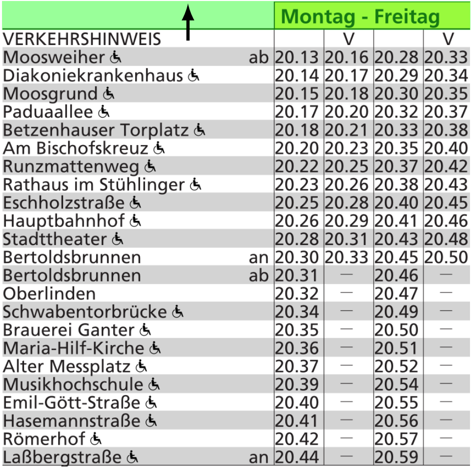
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

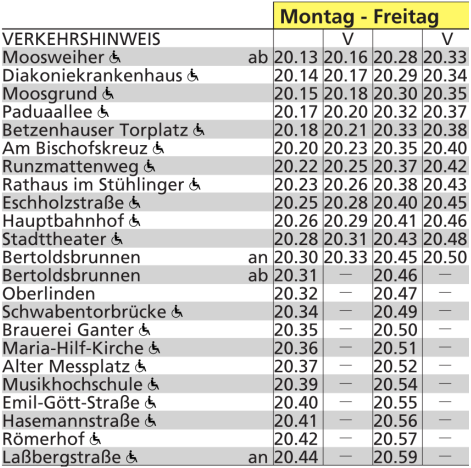
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

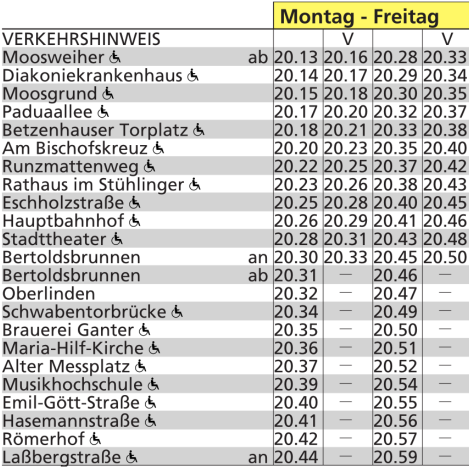
- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction: Approach

- Idea: Use body (i.e., times) to detect the table
- Run basic type detection
- Expand the table until no more cells can be added
  1. Select adjacent cells in a single direction
  2. Add adjacent cell, if it overlaps with row/column
- Run advanced type detection using other cells of the table



		Montag - Freitag			
VERKEHRSHINWEIS			V		V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34
Moosgrund ♿		20.15	20.18	20.30	20.35
Paduaallee ♿		20.17	20.20	20.32	20.37
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40
Runzmattenweg ♿		20.22	20.25	20.37	20.42
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43
Eschholzstraße ♿		20.25	20.28	20.40	20.45
Hauptbahnhof ♿		20.26	20.29	20.41	20.46
Stadttheater ♿		20.28	20.31	20.43	20.48
Bertoldsbrunnen	an	20.30	20.33	20.45	20.50
Bertoldsbrunnen	ab	20.31	—	20.46	—
Oberlinden		20.32	—	20.47	—
Schwabentorbrücke ♿		20.34	—	20.49	—
Brauerei Ganter ♿		20.35	—	20.50	—
Maria-Hilf-Kirche ♿		20.36	—	20.51	—
Alter Messplatz ♿		20.37	—	20.52	—
Musikhochschule ♿		20.39	—	20.54	—
Emil-Gött-Straße ♿		20.40	—	20.55	—
Hasemannstraße ♿		20.41	—	20.56	—
Römerhof ♿		20.42	—	20.57	—
Laßbergstraße ♿	an	20.44	—	20.59	—

Table Extraction

Evaluation

Table Extraction: Evaluation 1/5

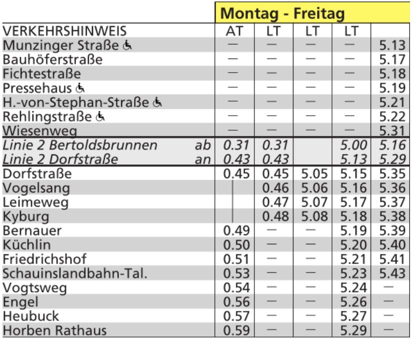
- Three datasets

VAG	Verkehrs AG Freiburg 4 PDFs
RMV	Rhein-Main-Verkehrsverbund 3 PDFs
TTT	Transposed timetables different US transit-agencies 4 PDFs

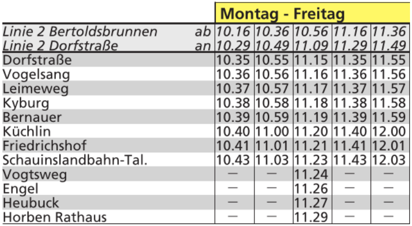
▸ More on transposed timetables

- PDFs selected based on table features

Table Extraction: Evaluation 2/5



		Montag - Freitag				
VERKEHRSHINWEIS		AT	LT	LT	LT	
Munzinger Straße ♿		—	—	—	—	5.13
Bauhöferstraße		—	—	—	—	5.17
Fichtestraße		—	—	—	—	5.18
Pressehaus ♿		—	—	—	—	5.19
H.-von-Stephan-Straße ♿		—	—	—	—	5.21
Rehlingstraße ♿		—	—	—	—	5.22
Wiesenweg		—	—	—	—	5.31
Linie 2 Bertoldsbrunnen	*ab*	*0.31*	*0.31*		*5.00*	*5.16*
Linie 2 Dorfstraße	*an*	*0.43*	*0.43*		*5.13*	*5.29*
Dorfstraße		0.45	0.45	5.05	5.15	5.35
Vogelsang		│	0.46	5.06	5.16	5.36
Leimeweg		│	0.47	5.07	5.17	5.37
Kyburg		│	0.48	5.08	5.18	5.38
Bernauer		0.49	—	—	5.19	5.39
Küchlin		0.50	—	—	5.20	5.40
Friedrichshof		0.51	—	—	5.21	5.41
Schauinslandbahn-Tal.		0.53	—	—	5.23	5.43
Vogtsweg		0.54	—	—	5.24	—
Engel		0.56	—	—	5.26	—
Heubuck		0.57	—	—	5.27	—
Horben Rathaus		0.59	—	—	5.29	—



		Montag - Freitag				
Linie 2 Bertoldsbrunnen	*ab*	*10.16*	*10.36*	*10.56*	*11.16*	*11.36*
Linie 2 Dorfstraße	*an*	*10.29*	*10.49*	*11.09*	*11.29*	*11.49*
Dorfstraße		10.35	10.55	11.15	11.35	11.55
Vogelsang		10.36	10.56	11.16	11.36	11.56
Leimeweg		10.37	10.57	11.17	11.37	11.57
Kyburg		10.38	10.58	11.18	11.38	11.58
Bernauer		10.39	10.59	11.19	11.39	11.59
Küchlin		10.40	11.00	11.20	11.40	12.00
Friedrichshof		10.41	11.01	11.21	11.41	12.01
Schauinslandbahn-Tal.		10.43	11.03	11.23	11.43	12.03
Vogtsweg		—	—	11.24	—	—
Engel		—	—	11.26	—	—
Heubuck		—	—	11.27	—	—
Horben Rathaus		—	—	11.29	—	—

- Left: More features
  (Connections between normal stops, has route annotations)

- Right: Less features (Connections at the start, has no route annotations)

▸ More on connections

Table Extraction: Evaluation 3/5

- No ground truth exists
  - manually create `.csv` files for each table
  - two tables per PDF for VAG/RMV, one table per PDF for TTT
- Three table extraction methods:
  - **PDFTables** Online solution for (general) table extraction
  - **pdf2gtfs-old** previous table extraction algorithm of pdf2gtfs
  - **pdf2gtfs-new** new algorithm using the shown approach
- Comparison between extracted `.csv` and ground truth by hand

Table Extraction: Evaluation 4/5

- Three measures: *Precision*, *Recall*, and F_1-*score*
- Compare extracted cells to cells in ground truth (GT)
  - **True Positive (TP)**
    Correctly extracted cells (content and relative position)
  - **False Positive (FP)**
    All cells that do not exist in GT or with different content/position
  - **True Negative (TN)**
    All empty extracted cells that are empty in GT
  - **False Negative (FN)**
    All cells that exist in GT but were not extracted

Table Extraction: Evaluation 5/5

- Precision: $P = \frac{\text{TP}}{\text{TP}+\text{FP}}$
  - relative amount of relevant cells that were extracted
- Recall: $R = \frac{\text{TP}}{\text{TP}+\text{FN}}$
  - relative amount of correct cells of all extracted cells
- F_1-score: $F_1 = \frac{2PR}{P+R}$
  - Harmonic mean between precision and recall

Table Extraction

Results

Table Extraction: Results 1/2

VAG	Precision	Recall	F_1-score
PDFTables	86.84%	57.63%	69.28%
pdf2gtfs-old	99.83%	88.84%	94.01%
pdf2gtfs-new	93.40%	97.78%	95.54%
RMV	Precision	Recall	F_1-score
PDFTables	94.03%	85.34%	89.78%
pdf2gtfs-old	98.82%	95.94%	97.36%
pdf2gtfs-new	98.97%	91.05%	94.84%

- Similar results for pdf2gtfs’ algorithms
- PDFTables (expectedly) worse

Table Extraction: Results 2/2

TTT	Precision	Recall	F_1-score
PDFTables	61.36%	43.12%	50.65%
pdf2gtfs-old	22.87%	8.48%	12.37%
pdf2gtfs-new	49.83%	96.76%	65.79%

- Clearly worse results than for “normal” timetables
- Low precision of pdf2gtfs-new mainly due to “difficult” time format (e.g., “09.42 A”)

  ▸ Show Example

Location Detection

Background & Approach

Location Detection: Background & Approach

- Timetable does not contain locations
  - we only have the names and order of stops
- First: We need the possible locations of each stop
  → OpenStreetMap (OSM)

  ▸ More on OSM
- Idea: Build a graph using these locations

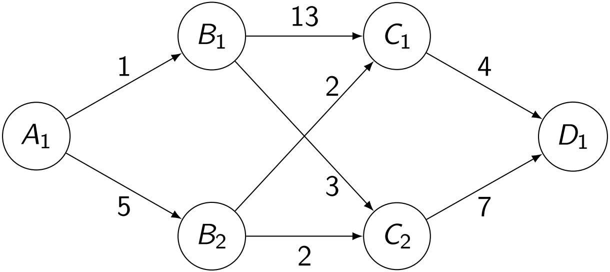
  ▸ More on graphs
  - each location is a node
  - each node has an edge to every node of the next stop

Location Detection: Background & Approach

- Timetable does not contain locations
  - we only have the names and order of stops
- First: We need the possible locations of each stop
  → OpenStreetMap (OSM) ▸ More on OSM
- Idea: Build a graph using these locations ▸ More on graphs
  - each location is a node
  - each node has an edge to every node of the next stop

  → shortest-path between a start and an end node
  (should) give the correct location for each stop
- Implementation detail:
  we use Dijkstra's algorithm for the shortest-path search

Location Detection: Caveats



- Weight of edges is the sum of
  - difference in stop name vs. node name
  - available OSM-tags
  - point-to-point distance to parent node (= previous stop)
- interpolate locations if we can not find one for a stop

Location Detection

Evaluation

Location Detection: Evaluation 1/2

- Three datasets with different transit agencies

  VAG Verkehrs AG Freiburg
  5 PDFs: one for each tram line

  RMV Rhein-Main-Verkehrsverbund
  2 PDFs: one bus line and one metro line

  VGN Verkehrsverbund Großraum Nürnberg GmbH
  4 PDFs: one bus, one S-Bahn, and two train lines

- Each agency provides the true locations

- Problem: GTFS feeds use different IDs
  → need a mapping between the feeds

Location Detection: Evaluation 2/2

- Create the mappings between the stop_ids of the feeds manually
  - search the ground truth for each stop
  - if there are multiple locations for a stop,
    use the station/first location
- Create p2g-eval to automatically evaluate a feed
  - Takes two feeds and the mapping between them
  - Calculate the distance of the mapped stops

Location Detection

Results

Results: Location Detection 1/4

VAG	both	detected	missing
count	100	98	2
min	2	2	129
max	175	123	175
mean	34	32	152
std	30	25	32

- Very close to true location
- Almost all stops detected

Results: Location Detection 2/4

RMV	both	detected	missing
count	27	18	9
min	6	6	40
max	1 012	83	1 012
mean	231	39	616
std	319	24	282

- $\sim$ 33% missing locations
- Similar results for detected stops

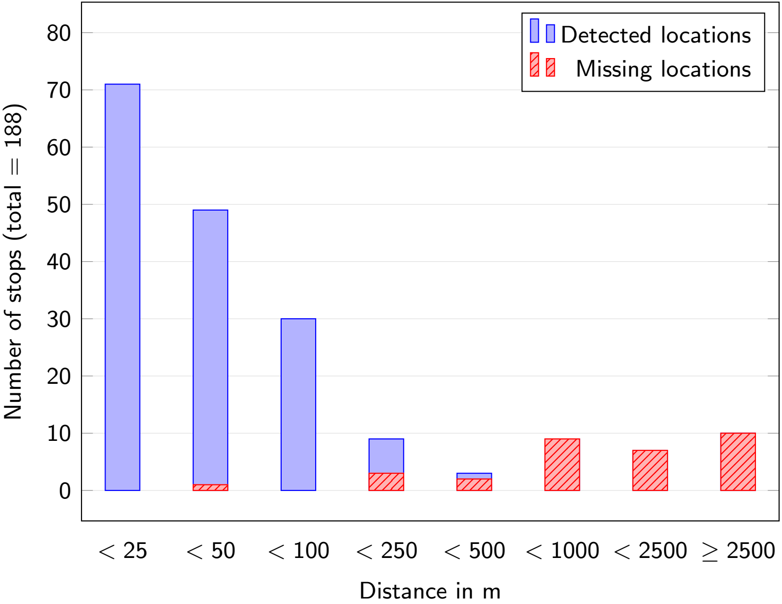
Results: Location Detection 3/4

VGN	both	detected	missing
count	61	40	21
min	4	5	107
max	87 317	260	87 317
mean	3 743	44	10 788
std	14 043	49	22 630

- $\sim$ 33% missing locations
- Similar results for detected stops with some outliers
- High distance for some missing stops
  (Reason: Stops of connections)

▸ More on connections

Results: Location Detection 4/4



Future Work

Future Work

- Location detection:
    - Automate the stop-mapping creation for p2g-eval using the stop-times

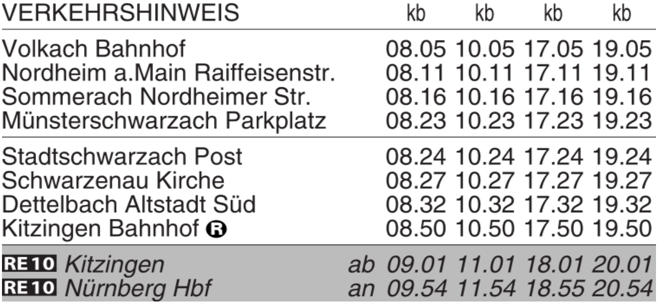
Future Work

- Location detection:
  - Automate the stop-mapping creation for p2g-eval using the stop-times
- Table extraction:
  - Overall stability
  - Main problem: Type detection and detection of multi-word cells

Future Work

- Location detection:
  - Automate the stop-mapping creation for p2g-eval using the stop-times
- Table extraction:
  - Overall stability
  - Main problem: Type detection and detection of multi-word cells
- **Questions?**

Appendix: Connections



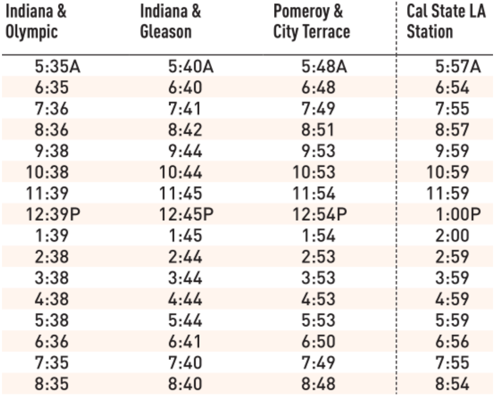
VERKEHRSHINWEIS		kb	kb	kb	kb
Volkach Bahnhof		08.05	10.05	17.05	19.05
Nordheim a.Main Raiffeisenstr.		08.11	10.11	17.11	19.11
Sommerach Nordheimer Str.		08.16	10.16	17.16	19.16
Münsterschwarzach Parkplatz		08.23	10.23	17.23	19.23
Stadtschwarzach Post		08.24	10.24	17.24	19.24
Schwarzenau Kirche		08.27	10.27	17.27	19.27
Dettelbach Altstadt Süd		08.32	10.32	17.32	19.32
Kitzingen Bahnhof 🅡		08.50	10.50	17.50	19.50
RE 10 *Kitzingen*	*ab*	*09.01*	*11.01*	*18.01*	*20.01*
RE 10 *Nürnberg Hbf*	*an*	*09.54*	*11.54*	*18.55*	*20.54*

- not part of the route
- usually serviced by fast(er) trains
- difficult to detect

◂ Return to table extraction

◂ Return to location detection

Appendix: Transposed Timetable

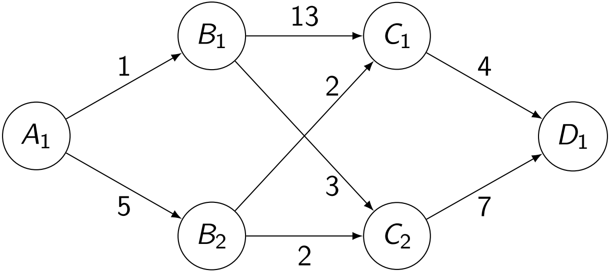


Indiana & Olympic	Indiana & Gleason	Pomeroy & City Terrace	Cal State LA Station
5:35A	5:40A	5:48A	5:57A
6:35	6:40	6:48	6:54
7:36	7:41	7:49	7:55
8:36	8:42	8:51	8:57
9:38	9:44	9:53	9:59
10:38	10:44	10:53	10:59
11:39	11:45	11:54	11:59
12:39P	12:45P	12:54P	1:00P
1:39	1:45	1:54	2:00
2:38	2:44	2:53	2:59
3:38	3:44	3:53	3:59
4:38	4:44	4:53	4:59
5:38	5:44	5:53	5:59
6:36	6:41	6:50	6:56
7:35	7:40	7:49	7:55
8:35	8:40	8:48	8:54

- Stops in the first row
- Each row contains a trip

Return

Appendix: Graph

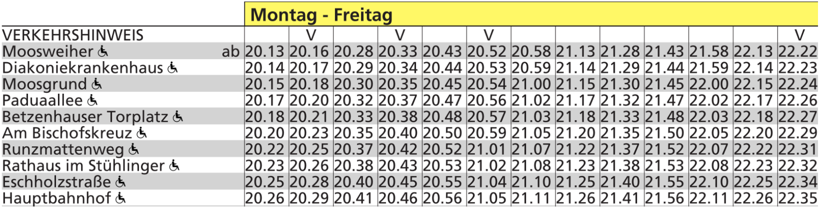


- consists of vertices (or nodes) and edges
- directed: edges have a direction
- weighted: edges have some weight
- Path: list of vertices that are connected by edges

Return

Appendix: Cell Types

- Types for route data, e.g., Time, Stop, Days
- Types for metadata, all annotation and indicator types
  - Indicator types (e.g., `RouteAnnotationIdentifier`):
    Indicates cell type of other cells
    Detected using user-defined keywords, e.g., ‘Verkehrshinweis’
  - Annotation types (e.g., `StopAnnotation`):
    Additional info about the data of other cells



		Montag - Freitag												
VERKEHRSHINWEIS			V		V		V							V
Moosweiher ♿	ab	20.13	20.16	20.28	20.33	20.43	20.52	20.58	21.13	21.28	21.43	21.58	22.13	22.22
Diakoniekrankenhaus ♿		20.14	20.17	20.29	20.34	20.44	20.53	20.59	21.14	21.29	21.44	21.59	22.14	22.23
Moosgrund ♿		20.15	20.18	20.30	20.35	20.45	20.54	21.00	21.15	21.30	21.45	22.00	22.15	22.24
Paduaallee ♿		20.17	20.20	20.32	20.37	20.47	20.56	21.02	21.17	21.32	21.47	22.02	22.17	22.26
Betzenhauser Torplatz ♿		20.18	20.21	20.33	20.38	20.48	20.57	21.03	21.18	21.33	21.48	22.03	22.18	22.27
Am Bischofskreuz ♿		20.20	20.23	20.35	20.40	20.50	20.59	21.05	21.20	21.35	21.50	22.05	22.20	22.29
Runzmattenweg ♿		20.22	20.25	20.37	20.42	20.52	21.01	21.07	21.22	21.37	21.52	22.07	22.22	22.31
Rathaus im Stühlinger ♿		20.23	20.26	20.38	20.43	20.53	21.02	21.08	21.23	21.38	21.53	22.08	22.23	22.32
Eschholzstraße ♿		20.25	20.28	20.40	20.45	20.55	21.04	21.10	21.25	21.40	21.55	22.10	22.25	22.34
Hauptbahnhof ♿		20.26	20.29	20.41	20.46	20.56	21.05	21.11	21.26	21.41	21.56	22.11	22.26	22.35

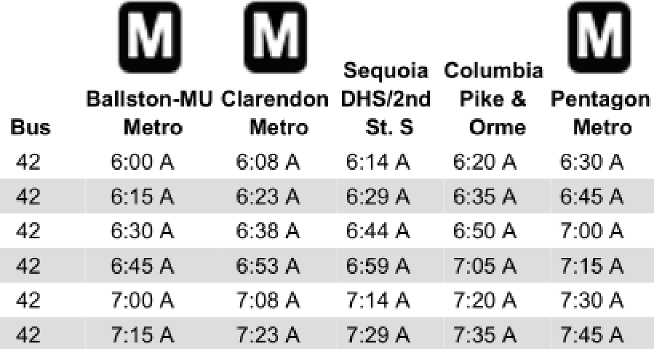
◂ Return

Appendix: OpenStreetMap

- OpenStreetMap (OSM) provides open map data, supplied by its users
- Information is stored in different types of objects
  - For us: only Nodes (henceforth OSMNodes) are relevant
- OSMNode contains
  - location of a point of interest (POI)
  - additional information about that POI using tags: simple key-value pairs (e.g., 'railway'='tram_stop')
- OSMNodes and their tags can be queried using, e.g., QLever

◂ Return

Appendix: Difficult Time Format



	M	M			M
Bus	Ballston-MU Metro	Clarendon Metro	Sequoia DHS/2nd St. S	Columbia Pike & Orme	Pentagon Metro
42	6:00 A	6:08 A	6:14 A	6:20 A	6:30 A
42	6:15 A	6:23 A	6:29 A	6:35 A	6:45 A
42	6:30 A	6:38 A	6:44 A	6:50 A	7:00 A
42	6:45 A	6:53 A	6:59 A	7:05 A	7:15 A
42	7:00 A	7:08 A	7:14 A	7:20 A	7:30 A
42	7:15 A	7:23 A	7:29 A	7:35 A	7:45 A

- time contains space
- no valid `strpformat()` format code (%p requires AM or PM)

◂ Return