Itinerary Recognition: Travelers, like Traveling Salesmen, Prefer Efficient Routes

Marco D. Adelfio Hanan Samet

Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA

GIR 2014

Itineraries: Introduction

- **Itinerary**:
  - Schedule of travel
  - Informal description of stops along travel route and related data
- Encapsulates experience of prior travelers
  - Useful for anyone traveling to a new part of the world
  - Useful for businesses that cater to tourists
- Evidence of demand for browsing itineraries
  - Commonly found in travel sections of magazines/newspapers
  - Travel companies and travel agents create custom itineraries for clients
  - Travel sites provide curated lists of itineraries (users paid to contribute, e.g., Tripoto.com)
- Question:
  - Can we recognize itineraries that occur on the Web?

Itineraries: Introduction

- Variety of representations
  - Graphical
  - Textual
  - Tabular

Itineraries: Introduction

- Variety of representations
  - Graphical
  - Textual
  - Tabular

Itineraries: Introduction

- Variety of representations
  - Graphical
  - Textual
  - Tabular

China In Two Weeks

An extra week allows visitors to experience some of the incredible contrasts that China has to offer; you'll see imperial sights, ultra-modern cities like Hong Kong and Shanghai, and idyllic Chinese countryside.

Days 1-4: Beijing

Start your time in Beijing with a visit to the **Forbidden City** and make sure you have plenty of time to explore all of its nooks and crannies. The following day head to the **Temple of Heaven** and in the afternoon visit the **Lama Temple**, then stroll through the lakes and *hutong* alleys of the **Back Lakes**. Dine looking out over it all, aboard one of **Kaorouji's** flat-bottomed boats. On the third day, head out early to the **Great Wall** at **Jinshanling** or **Simatai** and take a steep, spectacular hike. By the time you get back to Beijing you will certainly have the appetite for a hearty and delicious Beijing Duck dinner. On day 4 visit the **Summer Palace**, or whichever of Beijing's astounding range of attractions appeals, and then take an overnight train to Xi'an.

Days 5-6: Xi'an

Check into your hotel and freshen up before heading out to explore Xi'an's historic sights. Get some insight about that which you'll see in the coming days, by heading to the first-rate **Shanxi History Museum**. Later in the day go for a wander through the artist's quarter, maybe looking in on the **Forest of Stelae Museum**, or taking a stroll along the **City Wall** which gives a good overview of the city. In the evening dine on dumplings at **De Fa Chang**. Early the following day, head out to the **Terra-Cotta Warriors**. In the afternoon visit the **Great Mosque** and then as the light fades enjoy an alfresco dinner of kabobs in the Muslim markets.

Day 7: Home Time

Take a morning flight back to Beijing and connect with your international flight home.

Days 7-9: Shanghai

On day 7 make your way down to the **Big Goose Pagoda**, and in the afternoon take a flight from Xi'an to **Shanghai.** Speed into town on the maglev and settle into your hotel. In the evening head to **Cloud 9** in the Jin Mao Tower for a cocktail while you gaze out over the city. From here it's not far to **Jade on 36**, where you enjoy yet more views over dinner.

The following day visit the **Shanghai Museum**, one of the finest in the country. After exploring the museum, wander around **Yu Garden** (**Yu Yuan**) which makes for an interesting afternoon, especially when combined with souvenir shopping in the surrounding bazaar, and then a relaxing pot of green tea in the floating teahouse. In the evening head for a sumptuous dinner at one of the eateries in **Three on the Bund**, and then take a wander along the river. If you have any shoe leather left, cross over and continue your stroll along one of Shanghai's main shopping arteries, **Nanjing Road**, where very few places close before 10pm.

Itineraries: Introduction

- Variety of representations
  - Graphical
  - Textual
  - Tabular

China In Two Weeks

An extra week allows visitors to experience some of the incredible contrasts that China has to offer; you'll see imperial sights, ultra-modern cities like Hong Kong and Shanghai, and idyllic Chinese countryside.

Days 1-4: Beijing

Start your time in Beijing with a visit to the **Forbidden City** and make sure you have plenty of time to explore all of its nooks and crannies. The following day head to the **Temple of Heaven** and in the afternoon visit the **Lama Temple**, then stroll through the lakes and *hutong* alleys of the **Back Lakes**. Dine looking out over it all, aboard one of **Kaorouji's** flat-bottomed boats. On the third day, head out early to the **Great Wall** at **Jinshanling** or **Simatai** and take a steep, spectacular hike. By the time you get back to Beijing you will certainly have the appetite for a hearty and delicious Beijing Duck dinner. On day 4 visit the **Summer Palace**, or whichever of Beijing's astounding range of attractions appeals, and then take an overnight train to Xi'an.

Days 5-6: Xi'an

Check into your hotel and freshen up before heading out to explore Xi'an's historic sights. Get some insight about that which you'll see in the coming days, by heading to the first-rate **Shanxi History Museum**. Later in the day go for a wander through the artist's quarter, maybe looking in on the **Forest of Stelae Museum**, or taking a stroll along the **City Wall** which gives a good overview of the city. In the evening dine on dumplings at **De Fa Chang**. Early the following day, head out to the **Terra-Cotta Warriors**. In the afternoon visit the **Great Mosque** and then as the light fades enjoy an alfresco dinner of kabobs in the Muslim markets.

Day 7: Home Time

Take a morning flight back to Beijing and connect with your international flight home.

Days 7-9: Shanghai

On day 7 make your way down to the **Big Goose Pagoda**, and in the afternoon take a flight from Xi'an to **Shanghai.** Speed into town on the maglev and settle into your hotel. In the evening head to **Cloud 9** in the Jin Mao Tower for a cocktail while you gaze out over the city. From here it's not far to **Jade on 36**, where you enjoy yet more views over dinner.

The following day visit the **Shanghai Museum**, one of the finest in the country. After exploring the museum, wander around **Yu Garden** (**Yu Yuan**) which makes for an interesting afternoon, especially when combined with souvenir shopping in the surrounding bazaar, and then a relaxing pot of green tea in the floating teahouse. In the evening head for a sumptuous dinner at one of the eateries in **Three on the Bund**, and then take a wander along the river. If you have any shoe leather left, cross over and continue your stroll along one of Shanghai's main shopping arteries, **Nanjing Road**, where very few places close before 10pm.

	A	B	C
1	PERU – June 2012		
2	*Date*	*Location*	*Accommodations*
3	06/06/12	Lima	Allpa Hotel
4	06/07/12	Arequipa	Casa Andina Jerusalén
5	06/08/12	Arequipa	Casa Andina Jerusalén
6	06/09/12	Colca	Casa Andina Colca
7	06/10/12	Colca	Casa Andina Colca
8	06/11/12	Puno	La Hacienda
9	06/12/12	Puno	La Hacienda
10	06/13/12	Cuzco	Casa Andina San Blas Hotel
11	06/14/12	Cuzco	Casa Andina San Blas Hotel
12	06/15/12	Sacred Valley	La Casona de Yucay
13	06/16/12	Inca Trail	Campsite
14	06/17/12	Inca Trail	Campsite
15	06/18/12	Inca Trail	Campsite
16	06/19/12	Machu Picchu	Campsite
17	06/20/12	Cuzco	Casa Andina San Blas Hotel
18	06/21/12	Amazon	Sandoval Lake Lodge
19	06/22/12	Amazon	Sandoval Lake Lodge

Itineraries: Introduction

- Variety of representations
  - Graphical
  - Textual
  - Tabular

China In Two Weeks

An extra week allows visitors to experience some of the incredible contrasts that China has to offer; you'll see imperial sights, ultra-modern cities like Hong Kong and Shanghai, and idyllic Chinese countryside.

Days 1-4: Beijing

Start your time in Beijing with a visit to the **Forbidden City** and make sure you have plenty of time to explore all of its nooks and crannies. The following day head to the **Temple of Heaven** and in the afternoon visit the **Lama Temple**, then stroll through the lakes and *hutong* alleys of the **Back Lakes**. Dine looking out over it all, aboard one of **Kaorouji's** flat-bottomed boats. On the third day, head out early to the **Great Wall** at **Jinshanling** or **Simatai** and take a steep, spectacular hike. By the time you get back to Beijing you will certainly have the appetite for a hearty and delicious Beijing Duck dinner. On day 4 visit the **Summer Palace**, or whichever of Beijing's astounding range of attractions appeals, and then take an overnight train to Xi'an.

Days 5-6: Xi'an

Check into your hotel and freshen up before heading out to explore Xi'an's historic sights. Get some insight about that which you'll see in the coming days, by heading to the first-rate **Shanxi History Museum**. Later in the day go for a wander through the artist's quarter, maybe looking in on the **Forest of Stelae Museum**, or taking a stroll along the **City Wall** which gives a good overview of the city. In the evening dine on dumplings at **De Fa Chang**. Early the following day, head out to the **Terra-Cotta Warriors**. In the afternoon visit the **Great Mosque** and then as the light fades enjoy an alfresco dinner of kabobs in the Muslim markets.

Day 7: Home Time

Take a morning flight back to Beijing and connect with your international flight home.

Days 7-9: Shanghai

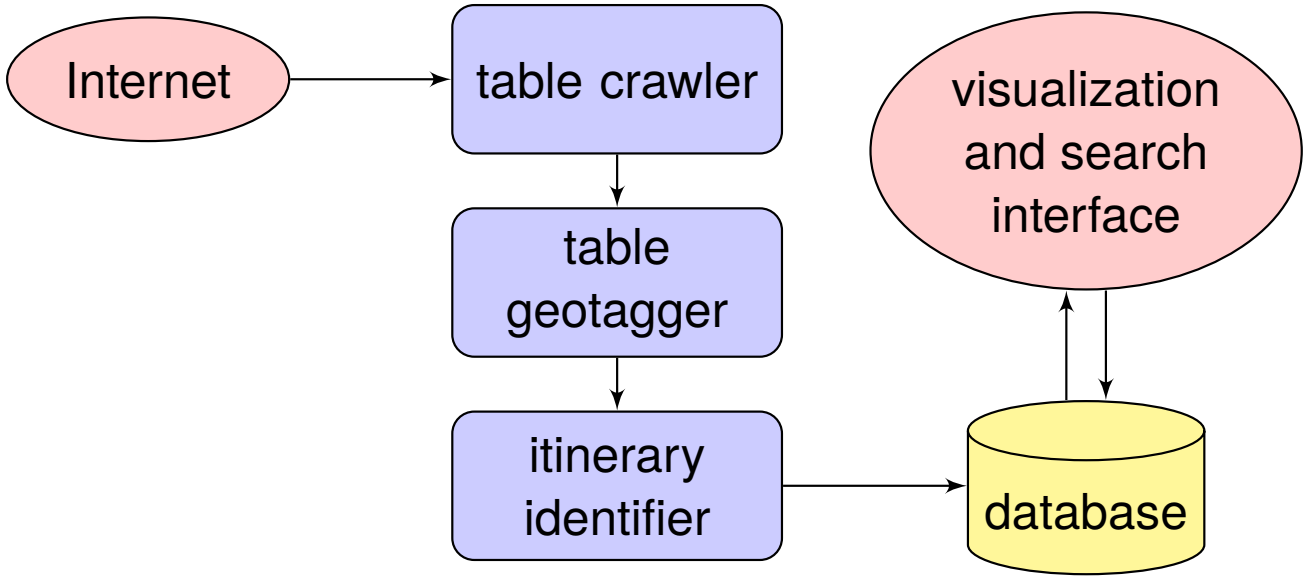
On day 7 make your way down to the **Big Goose Pagoda**, and in the afternoon take a flight from Xi'an to **Shanghai.** Speed into town on the maglev and settle into your hotel. In the evening head to **Cloud 9** in the Jin Mao Tower for a cocktail while you gaze out over the city. From here it's not far to **Jade on 36**, where you enjoy yet more views over dinner.

The following day visit the **Shanghai Museum**, one of the finest in the country. After exploring the museum, wander around **Yu Garden** (**Yu Yuan**) which makes for an interesting afternoon, especially when combined with souvenir shopping in the surrounding bazaar, and then a relaxing pot of green tea in the floating teahouse. In the evening head for a sumptuous dinner at one of the eateries in **Three on the Bund**, and then take a wander along the river. If you have any shoe leather left, cross over and continue your stroll along one of Shanghai's main shopping arteries, **Nanjing Road**, where very few places close before 10pm.

	A	B	C
1	PERU – June 2012		
2	*Date*	*Location*	*Accommodations*
3	06/06/12	Lima	Allpa Hotel
4	06/07/12	Arequipa	Casa Andina Jerusalén
5	06/08/12	Arequipa	Casa Andina Jerusalén
6	06/09/12	Colca	Casa Andina Colca
7	06/10/12	Colca	Casa Andina Colca
8	06/11/12	Puno	La Hacienda
9	06/12/12	Puno	La Hacienda
10	06/13/12	Cuzco	Casa Andina San Blas Hotel
11	06/14/12	Cuzco	Casa Andina San Blas Hotel
12	06/15/12	Sacred Valley	La Casona de Yucay
13	06/16/12	Inca Trail	Campsite
14	06/17/12	Inca Trail	Campsite
15	06/18/12	Inca Trail	Campsite
16	06/19/12	Machu Picchu	Campsite
17	06/20/12	Cuzco	Casa Andina San Blas Hotel
18	06/21/12	Amazon	Sandoval Lake Lodge
19	06/22/12	Amazon	Sandoval Lake Lodge

- Focus on tabular itineraries
  - Clearer delineation of stops
  - Associated stop metadata
- Goal:
  - Create large collection of itineraries found on the Web
  - Harness *volunteered geographic data*

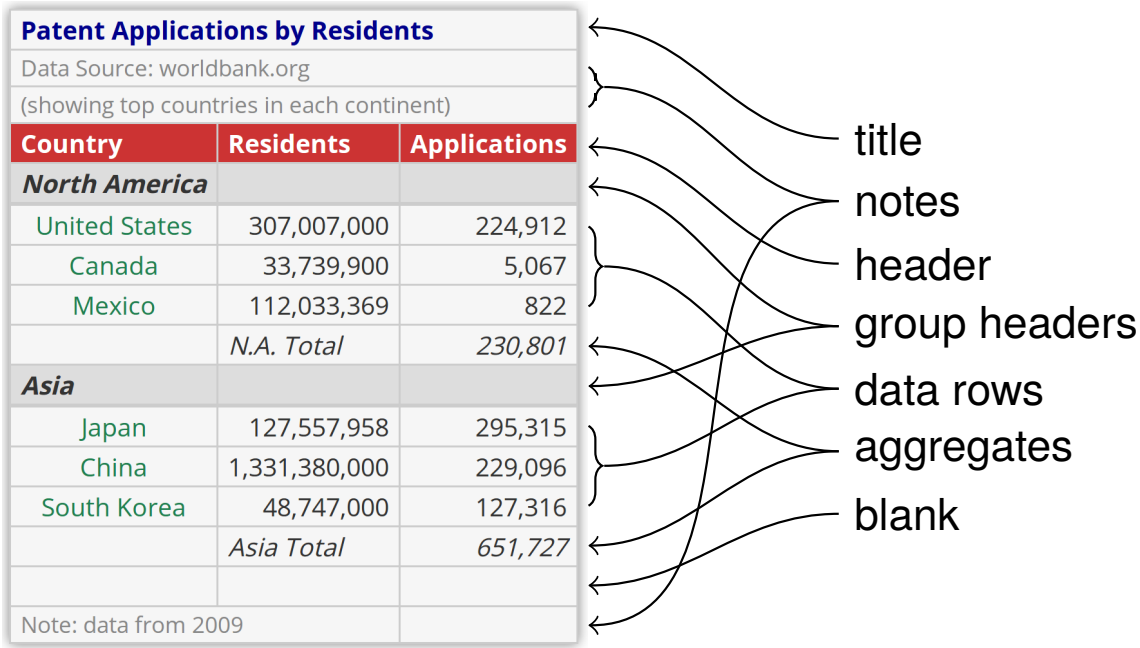
Pipeline



1. Obtain documents containing tables from Internet
2. Extract tables to generic datagrid format
3. Identify and geotag geographic tables
4. Classify geographic tables as itineraries or non-itineraries

Table Extraction

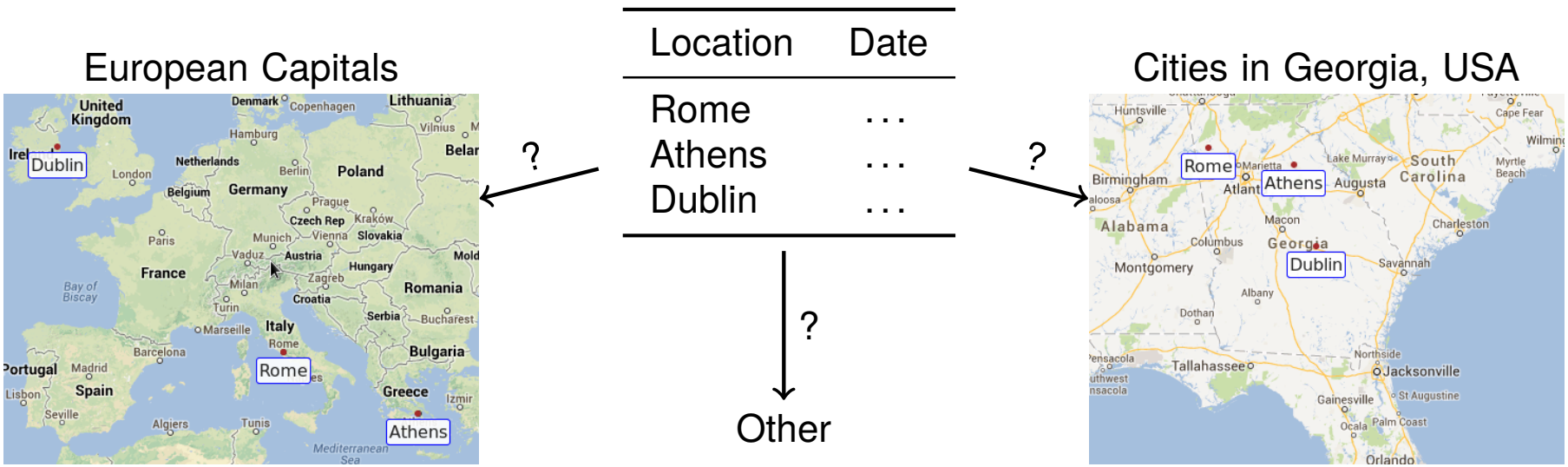
- Extract data portion of specified tables
- Generic extractor (not itinerary-specific)
- Purpose: Identify role of each row in table to support downstream processing



- Based on *Schema Extraction for Tabular Data on the Web*, Adelfio and Samet, VLDB 2013.

Table Geotagger

- Associate each row with geographic entity (from GeoNames gazetteer)
  1. Given table, determine whether table contains any geographic columns
  2. If so, identify a *category* that can be used to disambiguate geographic references
  3. Assign geographic interpretation to each row from within category



- Based on *Structured Toponym Resolution Using Combined Hierarchical Place Categories*, Adelfio and Samet, GIR 2013.

Itinerary Identifier

- Let T be a table containing an ordered set of relations $r_1, r_2, \ldots, r_n$, where each relation r_i has an associated location l_i. The **table itinerary decision problem** (TIDP) is to determine whether T represents an itinerary.

Date	Location	Delivery #
12/16/04	Oestrich-Winkel, DE	20031
03/17/05	Lavera, FR	20053
03/17/05	Lavera, FR	20054
04/27/05	Marl, DE	20065
05/25/05	Beringen, BE	20104
06/23/05	Schwechat-Mannswörth, AT	20112
09/08/05	Dordrecht, NL	20131
11/21/06	Litvinov, CZ	20142
11/10/05	Pasir Gudang, Johor, MY	20152
11/10/05	Pasir Gudang, Johor, MY	20153
12/14/05	Antwerpen, BE	20177
11/16/05	Tehran, IR	20179
12/19/05	Brüssel, BE	20183
01/19/06	Torre Boldone (BG), IT	20186
01/19/06	Torre Boldone (BG), IT	20187
...	...	...

Day	Dest	Activities
1	Vienna	Hotel check-in
2	Vienna	City tour
3	Vienna	Transfer to Budapest
	Budapest	City tour
4	Mohacs	Pecs excursion
	Villany	Wine tasting
5	Vukovar	Yugoslav Civil War tour
	Novi Sad	Walking tour
6	Belgrade	City tour
7	Iron Gates	Full day cruising
8	Vidin	Belogradchik excursion
9	Giurgiu	Palace of Parliament
10	Rousse	Disembarkation
	Plovdiv	Walking tour
11	Erdine	Lunch stop
12	Istanbul	City tour
13	Istanbul	Tour Topkapi Palace
14	Istanbul	Return flight home

Date	ETA	Location	Notes
9/19/07	8:00	Splendora FBC	Depart
	10:11	Nacogdoches, TX	Gas Stop
	12:09	Marshall, TX	Gas Stop & Lunch
	14:51	Texarkana, AR	
	15:22	Hope, AR	Gas Stop
	15:57	Gum Springs, AR	
	16:23	Arkadelphia, AR	Stop
9/20/07	7:30	Arkadelphia, AR	Depart
	7:39	Caddo Valley	Gas
	11:16	Dardanelle, AR	Gas Stop
	13:06	Jasper, AR	Lunch
	14:26	Dogpatch USA	Scenic/Photos
	14:42	Harrison, AR	Gas Stop & Lunch
	16:33	Francis, AR	
	16:49	Eureka Springs, AR	Stop & Gas
9/21/07	9:00	Eureka Springs, AR	Depart
	10:48	Ozark, AR	
	11:17	Van Buren, AR	Gas & Lunch
	12:53	Fort Smith, AR	
	12:55	Entering Oklahoma	
	15:10	Sunset Corner, OK	
	16:04	Entering Arkansas	
...	...	...	...

Itinerary indicators

- Dates
- Keywords
- Order

Itinerary Identifier

- Let T be a table containing an ordered set of relations $r_1, r_2, \ldots, r_n$, where each relation r_i has an associated location l_i. The **table itinerary decision problem** (TIDP) is to determine whether T represents an itinerary.

Date	Location	Delivery #
12/16/04	Oestrich-Winkel, DE	20031
03/17/05	Lavera, FR	20053
03/17/05	Lavera, FR	20054
04/27/05	Marl, DE	20065
05/25/05	Beringen, BE	20104
06/23/05	Schwechat-Mannswörth, AT	20112
09/08/05	Dordrecht, NL	20131
11/21/06	Litvinov, CZ	20142
11/10/05	Pasir Gudang, Johor, MY	20152
11/10/05	Pasir Gudang, Johor, MY	20153
12/14/05	Antwerpen, BE	20177
11/16/05	Tehran, IR	20179
12/19/05	Brüssel, BE	20183
01/19/06	Torre Boldone (BG), IT	20186
01/19/06	Torre Boldone (BG), IT	20187
...	...	...

Day	Dest	Activities
1	Vienna	Hotel check-in
2	Vienna	City tour
3	Vienna	Transfer to Budapest
	Budapest	City tour
4	Mohacs	Pecs excursion
	Villany	Wine tasting
5	Vukovar	Yugoslav Civil War tour
	Novi Sad	Walking tour
6	Belgrade	City tour
7	Iron Gates	Full day cruising
8	Vidin	Belogradchik excursion
9	Giurgiu	Palace of Parliament
10	Rousse	Disembarkation
	Plovdiv	Walking tour
11	Erdine	Lunch stop
12	Istanbul	City tour
13	Istanbul	Tour Topkapi Palace
14	Istanbul	Return flight home

Date	ETA	Location	Notes
9/19/07	8:00	Splendora FBC	Depart
	10:11	Nacogdoches, TX	Gas Stop
	12:09	Marshall, TX	Gas Stop & Lunch
	14:51	Texarkana, AR	
	15:22	Hope, AR	Gas Stop
	15:57	Gum Springs, AR	
	16:23	Arkadelphia, AR	Stop
9/20/07	7:30	Arkadelphia, AR	Depart
	7:39	Caddo Valley	Gas
	11:16	Dardanelle, AR	Gas Stop
	13:06	Jasper, AR	Lunch
	14:26	Dogpatch USA	Scenic/Photos
	14:42	Harrison, AR	Gas Stop & Lunch
	16:33	Francis, AR	
	16:49	Eureka Springs, AR	Stop & Gas
9/21/07	9:00	Eureka Springs, AR	Depart
	10:48	Ozark, AR	
	11:17	Van Buren, AR	Gas & Lunch
	12:53	Fort Smith, AR	
	12:55	Entering Oklahoma	
	15:10	Sunset Corner, OK	
	16:04	Entering Arkansas	
...	...	...	...

Itinerary indicators

- Dates
- Keywords
- Order

Itinerary Identifier

- Let T be a table containing an ordered set of relations $r_1, r_2, \ldots, r_n$, where each relation r_i has an associated location l_i. The **table itinerary decision problem** (TIDP) is to determine whether T represents an itinerary.

Date	Location	Delivery #
12/16/04	Oestrich-Winkel, DE	20031
03/17/05	Lavera, FR	20053
03/17/05	Lavera, FR	20054
04/27/05	Marl, DE	20065
05/25/05	Beringen, BE	20104
06/23/05	Schwechat-Mannswörth, AT	20112
09/08/05	Dordrecht, NL	20131
11/21/06	Litvinov, CZ	20142
11/10/05	Pasir Gudang, Johor, MY	20152
11/10/05	Pasir Gudang, Johor, MY	20153
12/14/05	Antwerpen, BE	20177
11/16/05	Tehran, IR	20179
12/19/05	Brüssel, BE	20183
01/19/06	Torre Boldone (BG), IT	20186
01/19/06	Torre Boldone (BG), IT	20187
...	...	...

Day	Dest	Activities
1	Vienna	Hotel check-in
2	Vienna	City tour
3	Vienna	Transfer to Budapest
	Budapest	City tour
4	Mohacs	Pecs excursion
	Villany	Wine tasting
5	Vukovar	Yugoslav Civil War tour
	Novi Sad	Walking tour
6	Belgrade	City tour
7	Iron Gates	Full day cruising
8	Vidin	Belogradchik excursion
9	Giurgiu	Palace of Parliament
10	Rousse	Disembarkation
	Plovdiv	Walking tour
11	Erdine	Lunch stop
12	Istanbul	City tour
13	Istanbul	Tour Topkapi Palace
14	Istanbul	Return flight home

Date	ETA	Location	Notes
9/19/07	8:00	Splendora FBC	Depart
	10:11	Nacogdoches, TX	Gas Stop
	12:09	Marshall, TX	Gas Stop & Lunch
	14:51	Texarkana, AR	
	15:22	Hope, AR	Gas Stop
	15:57	Gum Springs, AR	
	16:23	Arkadelphia, AR	Stop
9/20/07	7:30	Arkadelphia, AR	Depart
	7:39	Caddo Valley	Gas
	11:16	Dardanelle, AR	Gas Stop
	13:06	Jasper, AR	Lunch
	14:26	Dogpatch USA	Scenic/Photos
	14:42	Harrison, AR	Gas Stop & Lunch
	16:33	Francis, AR	
	16:49	Eureka Springs, AR	Stop & Gas
9/21/07	9:00	Eureka Springs, AR	Depart
	10:48	Ozark, AR	
	11:17	Van Buren, AR	Gas & Lunch
	12:53	Fort Smith, AR	
	12:55	Entering Oklahoma	
	15:10	Sunset Corner, OK	
	16:04	Entering Arkansas	
...	...	...	...

Itinerary indicators

- Dates
- Keywords
- Order

Itinerary Identifier

- Let T be a table containing an ordered set of relations $r_1, r_2, \ldots, r_n$, where each relation r_i has an associated location l_i. The **table itinerary decision problem** (TIDP) is to determine whether T represents an itinerary.

Date	Location	Delivery #
12/16/04	Oestrich-Winkel, DE	20031
03/17/05	Lavera, FR	20053
03/17/05	Lavera, FR	20054
04/27/05	Marl, DE	20065
05/25/05	Beringen, BE	20104
06/23/05	Schwechat-Mannswörth, AT	20112
09/08/05	Dordrecht, NL	20131
11/21/06	Litvinov, CZ	20142
11/10/05	Pasir Gudang, Johor, MY	20152
11/10/05	Pasir Gudang, Johor, MY	20153
12/14/05	Antwerpen, BE	20177
11/16/05	Tehran, IR	20179
12/19/05	Brüssel, BE	20183
01/19/06	Torre Boldone (BG), IT	20186
01/19/06	Torre Boldone (BG), IT	20187
...	...	...

Day	Dest	Activities
1	Vienna	Hotel check-in
2	Vienna	City tour
3	Vienna	Transfer to Budapest
	Budapest	City tour
4	Mohacs	Pecs excursion
	Villany	Wine tasting
5	Vukovar	Yugoslav Civil War tour
	Novi Sad	Walking tour
6	Belgrade	City tour
7	Iron Gates	Full day cruising
8	Vidin	Belogradchik excursion
9	Giurgiu	Palace of Parliament
10	Rousse	Disembarkation
	Plovdiv	Walking tour
11	Erdine	Lunch stop
12	Istanbul	City tour
13	Istanbul	Tour Topkapi Palace
14	Istanbul	Return flight home

Date	ETA	Location	Notes
9/19/07	8:00	Splendora FBC	Depart
	10:11	Nacogdoches, TX	Gas Stop
	12:09	Marshall, TX	Gas Stop & Lunch
	14:51	Texarkana, AR	
	15:22	Hope, AR	Gas Stop
	15:57	Gum Springs, AR	
	16:23	Arkadelphia, AR	Stop
9/20/07	7:30	Arkadelphia, AR	Depart
	7:39	Caddo Valley	Gas
	11:16	Dardanelle, AR	Gas Stop
	13:06	Jasper, AR	Lunch
	14:26	Dogpatch USA	Scenic/Photos
	14:42	Harrison, AR	Gas Stop & Lunch
	16:33	Francis, AR	
	16:49	Eureka Springs, AR	Stop & Gas
9/21/07	9:00	Eureka Springs, AR	Depart
	10:48	Ozark, AR	
	11:17	Van Buren, AR	Gas & Lunch
	12:53	Fort Smith, AR	
	12:55	Entering Oklahoma	
	15:10	Sunset Corner, OK	
	16:04	Entering Arkansas	
...	...	...	...

Itinerary indicators

- Dates
- Keywords
- Order

Itinerary Identifier

- Let T be a table containing an ordered set of relations $r_1, r_2, \ldots, r_n$, where each relation r_i has an associated location l_i. The **table itinerary decision problem** (TIDP) is to determine whether T represents an itinerary.

Date	Location	Delivery #
12/16/04	Oestrich-Winkel, DE	20031
03/17/05	Lavera, FR	20053
03/17/05	Lavera, FR	20054
04/27/05	Marl, DE	20065
05/25/05	Beringen, BE	20104
06/23/05	Schwechat-Mannswörth, AT	20112
09/08/05	Dordrecht, NL	20131
11/21/06	Litvinov, CZ	20142
11/10/05	Pasir Gudang, Johor, MY	20152
11/10/05	Pasir Gudang, Johor, MY	20153
12/14/05	Antwerpen, BE	20177
11/16/05	Tehran, IR	20179
12/19/05	Brüssel, BE	20183
01/19/06	Torre Boldone (BG), IT	20186
01/19/06	Torre Boldone (BG), IT	20187
...	...	...

Day	Dest	Activities
1	Vienna	Hotel check-in
2	Vienna	City tour
3	Vienna	Transfer to Budapest
	Budapest	City tour
4	Mohacs	Pecs excursion
	Villany	Wine tasting
5	Vukovar	Yugoslav Civil War tour
	Novi Sad	Walking tour
6	Belgrade	City tour
7	Iron Gates	Full day cruising
8	Vidin	Belogradchik excursion
9	Giurgiu	Palace of Parliament
10	Rousse	Disembarkation
	Plovdiv	Walking tour
11	Erdine	Lunch stop
12	Istanbul	City tour
13	Istanbul	Tour Topkapi Palace
14	Istanbul	Return flight home

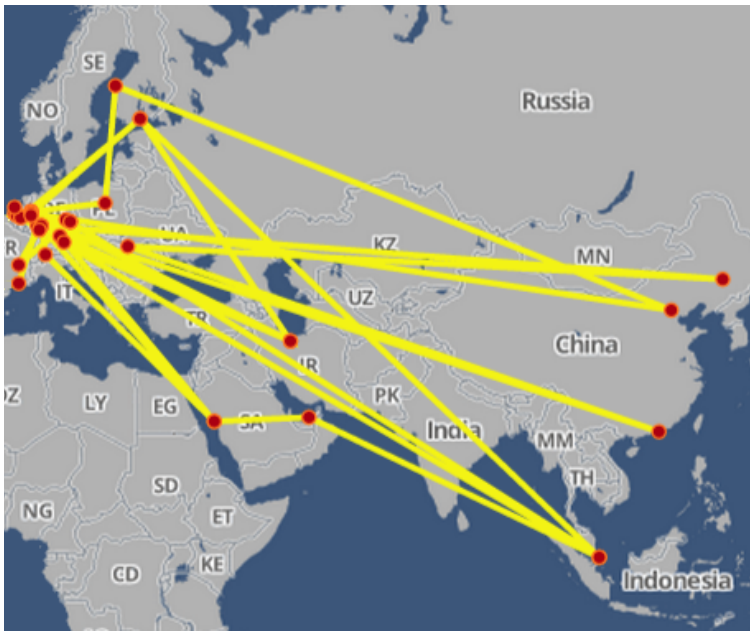
Date	ETA	Location	Notes
9/19/07	8:00	Splendora FBC	Depart
	10:11	Nacogdoches, TX	Gas Stop
	12:09	Marshall, TX	Gas Stop & Lunch
	14:51	Texarkana, AR	
	15:22	Hope, AR	Gas Stop
	15:57	Gum Springs, AR	
	16:23	Arkadelphia, AR	Stop
9/20/07	7:30	Arkadelphia, AR	Depart
	7:39	Caddo Valley	Gas
	11:16	Dardanelle, AR	Gas Stop
	13:06	Jasper, AR	Lunch
	14:26	Dogpatch USA	Scenic/Photos
	14:42	Harrison, AR	Gas Stop & Lunch
	16:33	Francis, AR	
	16:49	Eureka Springs, AR	Stop & Gas
9/21/07	9:00	Eureka Springs, AR	Depart
	10:48	Ozark, AR	
	11:17	Van Buren, AR	Gas & Lunch
	12:53	Fort Smith, AR	
	12:55	Entering Oklahoma	
	15:10	Sunset Corner, OK	
	16:04	Entering Arkansas	
...	...	...	...

Itinerary indicators

- Dates
- Keywords
- Order
- Geography?

Itinerary Identifier

- Utilizing geography
  - Examine induced route
  - Itineraries ordered based on spatial relationships (by definition), while non-itineraries usually are spatially “random”

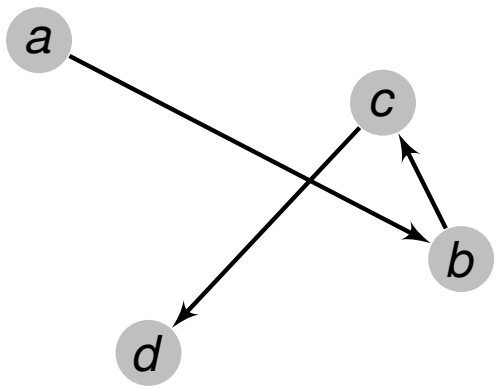


Itinerary Identifier

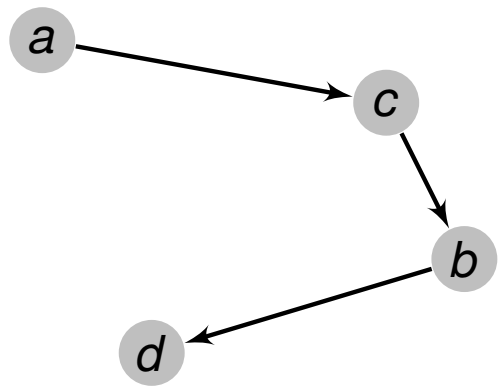
- Hypothesis: Humans use mixture of geographic knowledge and textual clues
  - Geographic Knowledge
    - Does route “look like” an itinerary?
  - Textual Clues
    - Ordered date columns
    - Terms related to travel
    - (Absence of) alphabetized columns

Traveling Salesman Optimizations

- Traveling Salesman Problem
  - Find Hamiltonian Cycle visiting all nodes in shortest total distance.
  - NP-hard problem, approximations are used for efficient solutions
- 2-Opt optimization procedure [Croes 1958]
  - Given one path connecting all nodes, select two edges at random. If swapping endpoints leads to shorter total path length, switch and continue.



(a) path *abcd*



(b) path *acbd*

- Reversing subpath *bc* results in shorter overall path
  - Subpath *bc* is not *reasonably ordered*

Efficiency Measures

- Intuition: count fraction of subpaths that are reasonably ordered
- Choice of measures with different granularity
  - Unclear at what granularity humans tend to optimize their travel plans
  - So: try multiple granularities (*local* and *general* efficiency)
- For ordered set of locations $L = l_1 l_2 \ldots l_n$ and $d(l_i, l_j) =$ distance between l_i and l_j, let

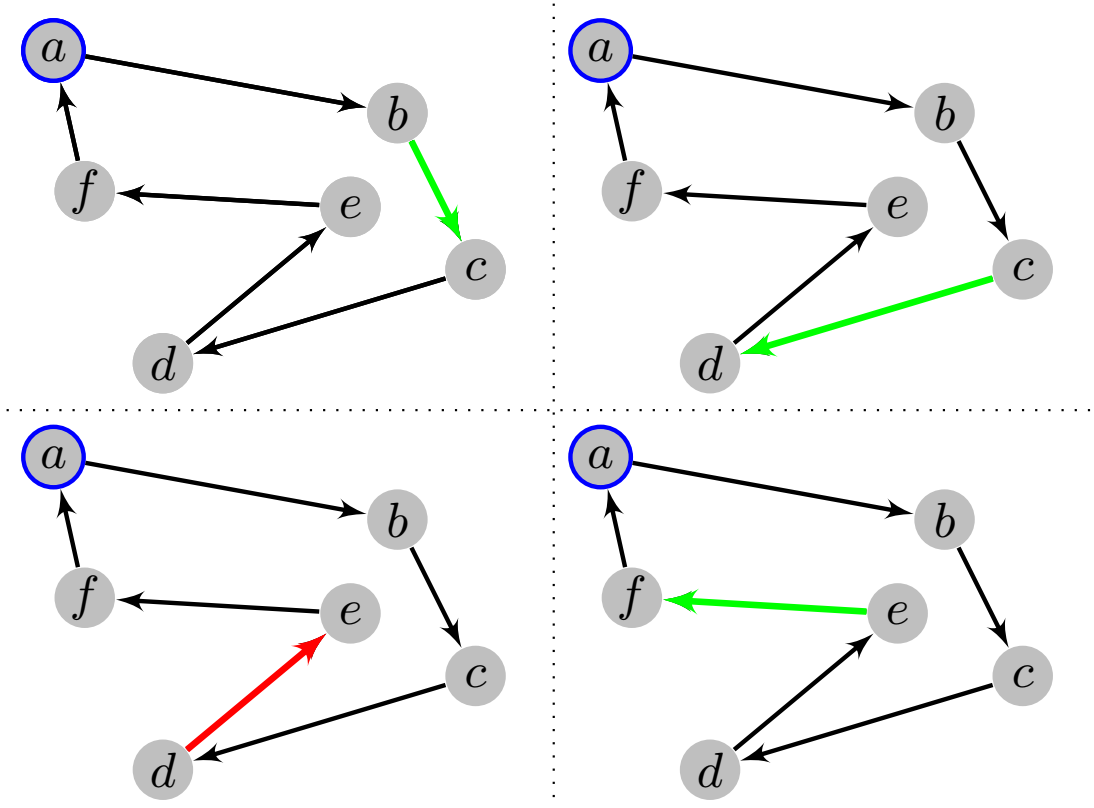
$$\delta_{i,j}(L) = \begin{cases} 1 & \text{if } (d(l_i, l_{i+1}) + d(l_j, l_{j+1})) \leq \\ & \quad (d(l_i, l_j) + d(l_{i+1}, l_{j+1})) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

- Subpath between stops i and j is reasonably ordered *iff* $\delta_{i,j} = 1$

Local Efficiency

- **Local efficiency** (ϵ_1) is the fraction of consecutive stop pairs whose reversal would lead to a longer total route distance. For locations $L = l_1 l_2 \ldots l_n$,

$$\epsilon_1(L) = \frac{1}{n-3} \sum_{i=1}^{n-3} \delta_{i,i+2}(L). \tag{2}$$



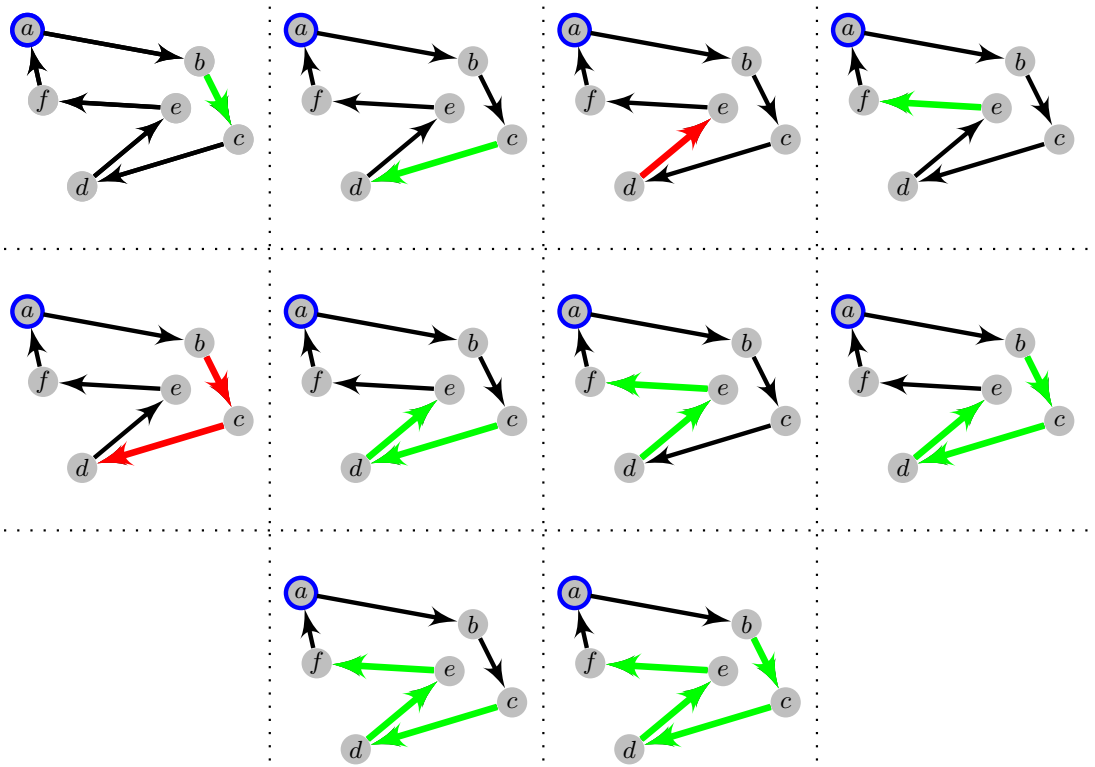
- Find reasonably ordered subpaths of length 2 (1 edge)
- Don't consider subpaths connected to first and last stop
- Count fraction of subpaths that are reasonably ordered
- This example:
  - Swapping order of *d* and *e* would reduce length
  - 3 of 4 are reasonably ordered
  - $\epsilon_1 = 0.75$

reasonably ordered
not reasonably ordered
first/last stop

General Efficiency

- **General efficiency** is the fraction of all unique, non-consecutive edge pairs that would result in a longer total route if their endpoints were swapped. For locations $L = l_1 l_2 \ldots l_n$,

$$\epsilon_2(L) = \frac{1}{\binom{n-2}{2}} \sum_{i=1}^{n-3} \sum_{j=i+2}^{n-1} \delta_{i,j}(L). \tag{3}$$



- Similar to local efficiency, but allow subpaths with > 1 edge
- Captures notion that travelers avoid excess travel over long subpaths
- This example:
  - Subpaths *de* and *bcd* are not reasonably ordered
  - 8 of 10 are reasonably ordered
  - $\epsilon_2 = 0.8$

Context Features

- Round trip table
  - $f_r(t) = 1$ *iff the primary location column of the table includes the same location in the first and last positions, 0 otherwise*.
- Temporal table
  - $f_{od}(t) =$ *# of ordered date/time columns found in the table*. Since itineraries are temporal objects, itineraries in tables commonly include a date/time column.
- Numeric table
  - $f_{on}(t) =$ *# of ordered numeric columns found in the table*. While ordered numeric columns are a component of some itinerary tables, they are also common in non-itineraries.
- Alphabetic table
  - $f_a(t) =$ *# of text columns found in the table that are sorted alphabetically*. It is rare for tables to be arranged both spatially and alphabetically.
- Term vector
  - $\vec{f_t}(t) =$ *a term vector of words commonly found in itineraries*. We use a list of 40 words and phrases that have a substantial difference in their frequency of use in itineraries versus non-itineraries. Such terms include “itinerary”, “trip”, “travel”, “airport”, “hotel”, “cruise”, month names, and others.

Classifiers

- Identified several indicators of itineraries
- Machine learning models for TIDP classification:
  1. Naive Bayes classifier
  2. Decision Tree classifier
  3. Support Vector Machine classifier
- Features pre-processed based on expected feature format of classifier
  1. Binarized for Naive Bayes
  2. Raw numeric values for Decision Tree
  3. Standardized (mean- and variance-adjusted) for SVM

Dataset

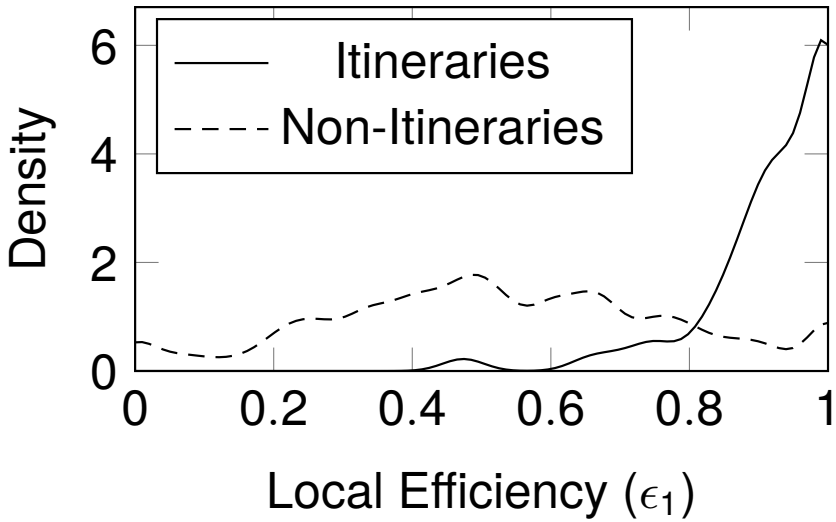
- Started with two million tables from the Web
  - Found documents containing tables using keyword search

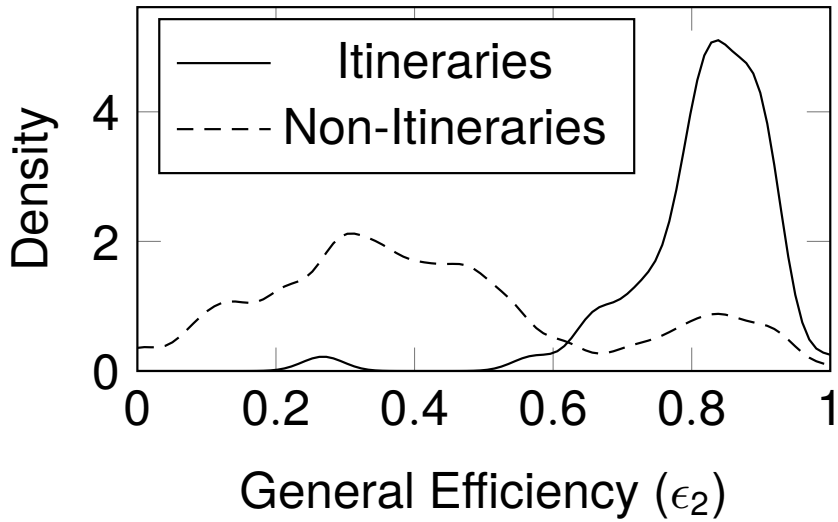
Full Dataset	
Documents	2,000,000
containing data tables	662,511
Data tables	2,128,032
Columns	10,142,785
Cells	280,170,694
After removing non-geographic tables	
Documents	130,294
Data Tables	235,433
Columns	1,527,890
Cells	80,432,927

- Selected 300 tables for annotation and evaluation
  - 200 selected randomly from geographic tables (3 itineraries)
  - Itinerary collection is too sparse.
  - Added 100 selected randomly with $\epsilon_1 > 0.8$ and $n \geq 10$ (57 itineraries)
  - Total: 60 itineraries, 240 non-itineraries

Efficiency Measures

- Examined discriminatory power of efficiency measures
- Look at density of values for itineraries and non-itineraries
- Densities scaled by relative frequency of efficiency values in test set vs. full corpus

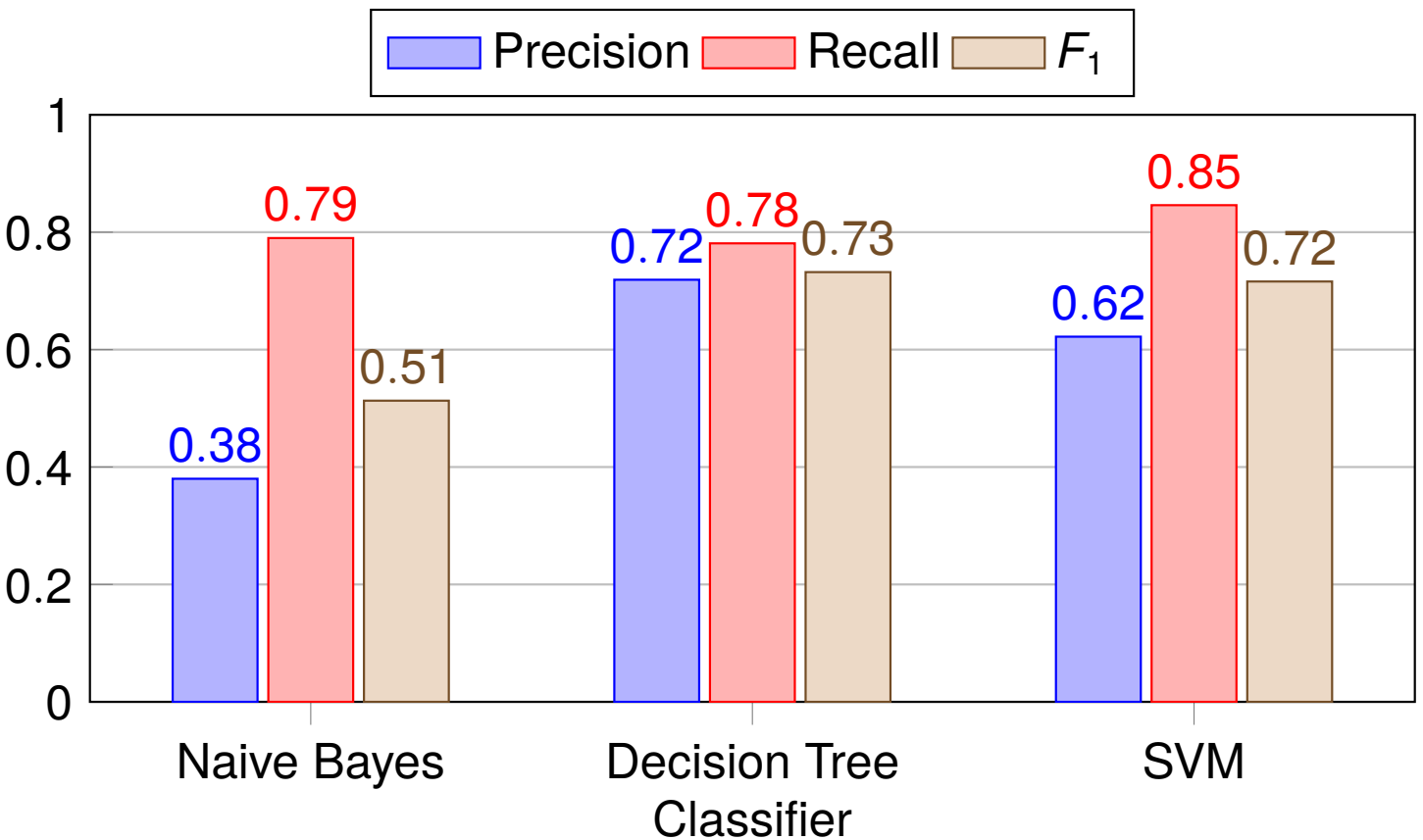




- Both are highly discriminatory (itineraries and non-itineraries have very different densities)
- Appears that ϵ_1 is more skewed towards 1.0 than ϵ_2
  - May be due to mechanism for creating test set, but informally appears to be true

Classification Accuracy

- Tested three classification models at itinerary recognition (TIDP)
- Measured precision, recall, and F_1 measure for recognizing *true itineraries*
- Decision tree achieves highest F_1 measure, ahead of SVM and Naive Bayes
- F_1 measure of 0.73: many itineraries retrieved accurately, room for improvement



Contribution of Individual Features

- Compared accuracy of decision tree with some features removed during training
- Shows marginal contribution of individual features
- Highest contributions from local efficiency measure

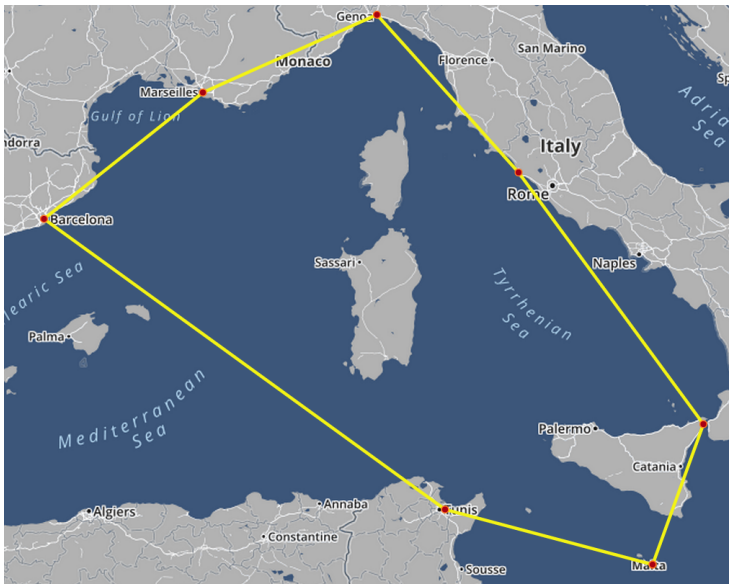
Feature	F_1 Without Feature	Marginal Contribution to F_1 score
ϵ_1	0.62	+0.11
ϵ_2	0.70	+0.03
ϵ_1 and ϵ_2	0.44	+0.29
f_r	0.71	+0.02
f_{od}	0.69	+0.04
f_{on}	0.72	+0.01
f_a	0.69	+0.04
$\vec{f_t}$	0.72	+0.01
non-efficiency	0.66	+0.07

Itinerary Identification

- Applied to our full collection of Web tables
- Result: 1206 itineraries, (0.5% of the 235,433 geographic tables)
- Examined results for trends

- Many cruises

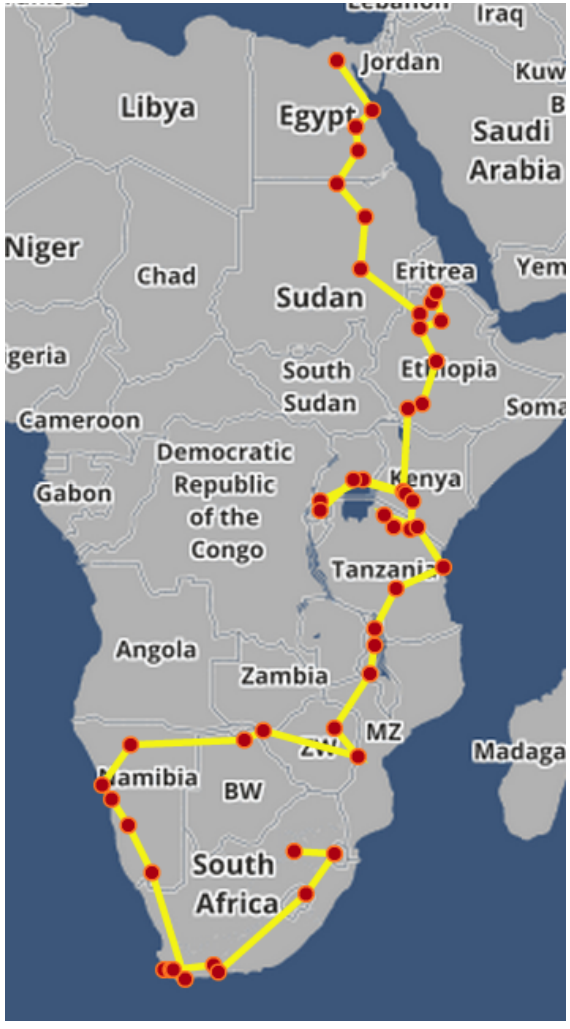
Day	Port	Arrival	Departure
Nov 30, 1999	Valletta, Malta	---	6:00 p.m.
Dec 01, 1999	Messina, Italy	7:00 a.m.	2:00 p.m.
Dec 02, 1999	Civitavecchia, Italy	8:00 a.m.	7:00 p.m.
Dec 03, 1999	Genoa, Italy	9:00 a.m.	6:00 p.m.
Dec 04, 1999	Marseilles, France	8:00 a.m.	6:00 p.m.
Dec 05, 1999	Barcellona, Spain	7:00 a.m.	2:00 p.m.
Dec 06, 1999	La Goulette, Tunisia	8:00 a.m.	6:00 p.m.
Dec 07, 1999	Valletta, Malta	9:00 a.m.	---



Sample Itineraries

- Some extended road trips

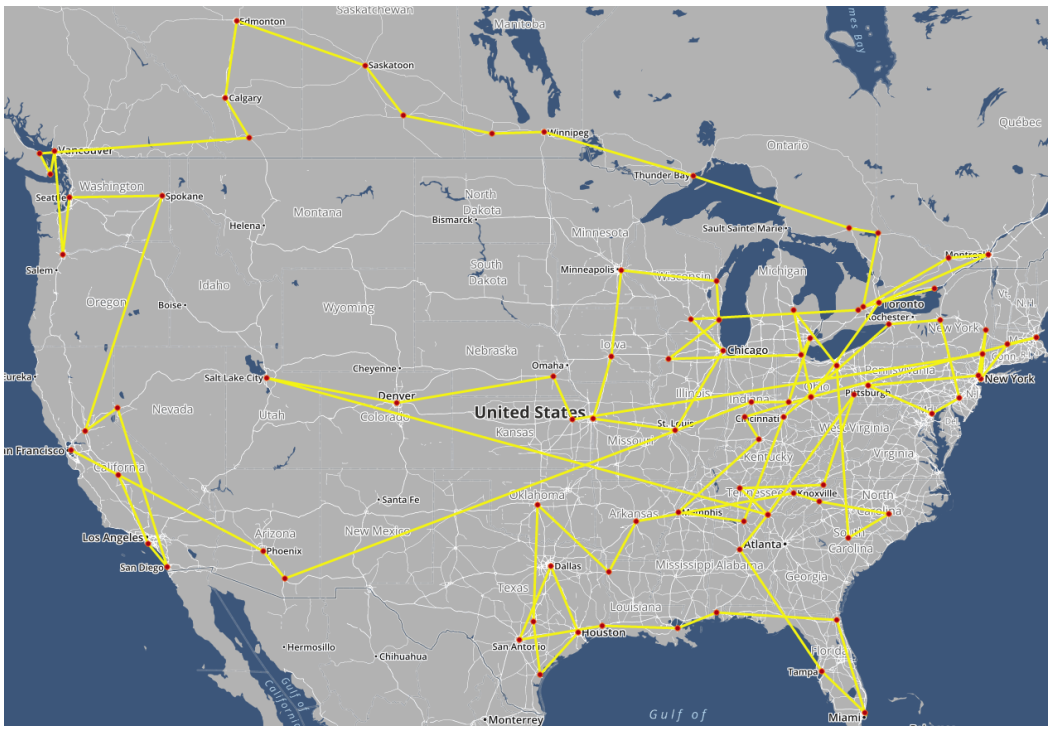
	Destination	Arrival
1	Cairo, Egypt	November 18, 2012
2	Safaga, Egypt	November 29, 2012
3	Luxor, Egypt	December 1, 2012
4	Aswan, Egypt	December 4, 2012
5	Wadi Halfa, Sudan	December 11, 2012
6	Abu Hamed, Sudan	December 14, 2012
7	Khartoum, Sudan	December 15, 2012
8	Gondar, Ethiopia	December 18, 2012
9	Simien Mountains, Ethiopia	December 20, 2012
10	Aksum, Ethiopia	December 23, 2012
11	Lalibela, Ethiopia	December 25, 2012
12	Bahir Dar, Ethiopia	December 27, 2012
13	Addis Ababa, Ethiopia	December 30, 2012
14	Arba Minch, Ethiopia	January 2, 2013
15	Jinka, Ethiopia	January 3, 2013
16	Nakuru, Kenya	January 8, 2013
17	Jinja, Uganda	January 11, 2013
18	Kabale, Uganda	January 13, 2013
19	Kigali, Rwanda	January 15, 2013
20	Volcans National Park, Rwanda	January 15, 2013
21	Kampala, Uganda	January 19, 2013
22	Lake Naivasha, Kenya	January 22, 2013
23	Nairobi, Kenya	January 24, 2013
24	Arusha, Tanzania	January 26, 2013
25	Ngorongoro Crater, Tanzania	January 28, 2013
26	Serengeti, Tanzania	January 29, 2013
27	Mount Kilimanjaro, Tanzania	January 30, 2013



Sample Itineraries

- Several concert tours

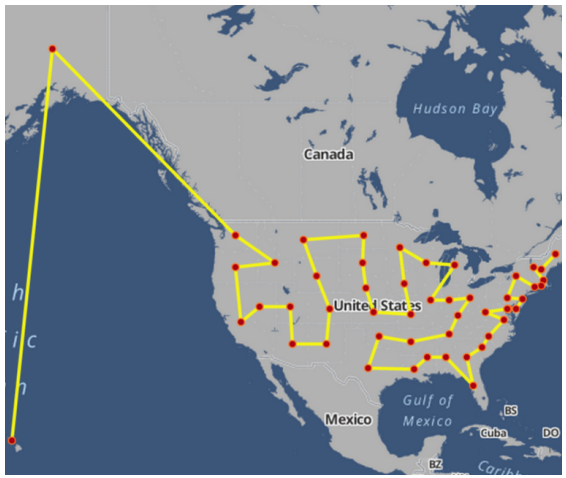
	A	B	C	
1	**Historical Rush Tour Dates Listing from Cygnus-X1.Net**			
2	Updated:	Thursday, October 23, 2014		
3				
4	**Date**	**Venue**	**Location**	**Opening / Supporting Band**
5				
6	**Tour Supporting 'Hemispheres' (10/14/1978 through 6/4/1979)**			
7	October 14, 1978	Kingston Memorial Centre	Kingston, Ontario	
8	October 15, 1978	War Memorial Hall, University of Guelph	Guelph, Ontario	
9	October 17, 1978	North Bay Memorial Gardens	North Bay, Ontario	
10	October 18, 1978	Sudbury Arena	Sudbury, Ontario	
11	October 20, 1978	Fort William Gardens	Thunder Bay, Ontario	
12	October 21, 1978	Winnipeg Arena	Winnipeg, Manitoba	(Streetheart)
13	October 22, 1978	Keystone Centre	Brandon, Manitoba	
14	October 24, 1978	Agridome	Regina, Saskatchewan	
15	October 25, 1978	Saskatoon Arena	Saskatoon, Saskatchewan	(Streetheart)
16	October 27, 1978	Northlands Coliseum	Edmonton, Alberta	
17	October 28, 1978	Calgary Corral	Calgary, Alberta	
18	October 29, 1978	Lethbridge Sportsplex	Lethbridge, Alberta	
19	October 31, 1978	K.X.A	Auditorium	Kamloops, British Columbia



Errors

- Misclassification due to table extraction and geotagging errors
- Tables with high efficiency that are not itineraries
  - Dewey Decimal Numbers

Dewey Decimal Number	State
974.1	Maine
974.2	New Hampshire
974.3	Vermont
974.4	Massachusetts
974.5	Rhode Island
…	…



  - Italian Coastal Regions

Region	Coastline
Imperia	62.7 km
Savona	80.5 km
Genova	109.2 km
Massa Carrara	13.0 km
Lucca	20.5 km
…	…

Conclusions

- Developed method for identifying itineraries from geographic tables
- Utilized efficiency measure based on TSP optimization

- Observations
  - Table format may be better suited to some types of itineraries than others (better for cruises than safaris, which tend to use text representations).
  - Some itineraries combine multiple stops

- Many interesting problems for future work
  - Do different types of itineraries exhibit different types of efficiency?
  - Extending recognition efforts to textual itineraries
  - Identifying effective search interfaces
  - Applying method to full Web crawl

Acknowledgements

- Thanks to our sponsors:
  - Google Research
  - National Science Foundation
  - NVIDIA