

Constraint-based Information Integration



Craig Knoblock

University of Southern California

Thanks to Jose Luis for some of these slides



Constraint Satisfaction and Propagation for Integration

- Integrating data from multiple sources often involves reasoning about the information
- Constraints provide a approach to expressing relationships and filtering data



Outline

- Part I: Integration Frameworks
 - Constraint satisfaction in SmartClients
 - Constraint propagation in Heracles
- Part II: CSPs for Integrating Data
 - Constraint satisfaction for building identification from open source data



SmartClients [Torrens et al, 2002]

- Cast an integration problem as a Constraint Satisfaction Problem (CSP)
- Given a request, the server retrieves the required data and sends the data and the CSP to the client
- Client solves the CSP locally
 - Large complex problem transmitted in small amount of space
 - Provides fine-grained user interaction with the data



Example Problem

- I live in Bern, Switzerland, and would like to visit colleagues in Princeton and London. I would like to spend at least two days in each place, and will need to travel in the first two weeks of February.
 - 1st leg from Bern to Princeton: flights from ZRH/BSL/GVA to JFK/EWR/PHL on the dates from 1st to 10th February
 - 2nd leg from Princeton to London: flights from JFK/EWR/PHL to LGW/LHR/LCY on the dates from 4th to 12th February, and
 - 3rd leg from London to Bern: flights from LGW/LHR/LCY to ZRH/BSL/GVA on the dates from 6th to 14th February



Constraint Satisfaction Problem

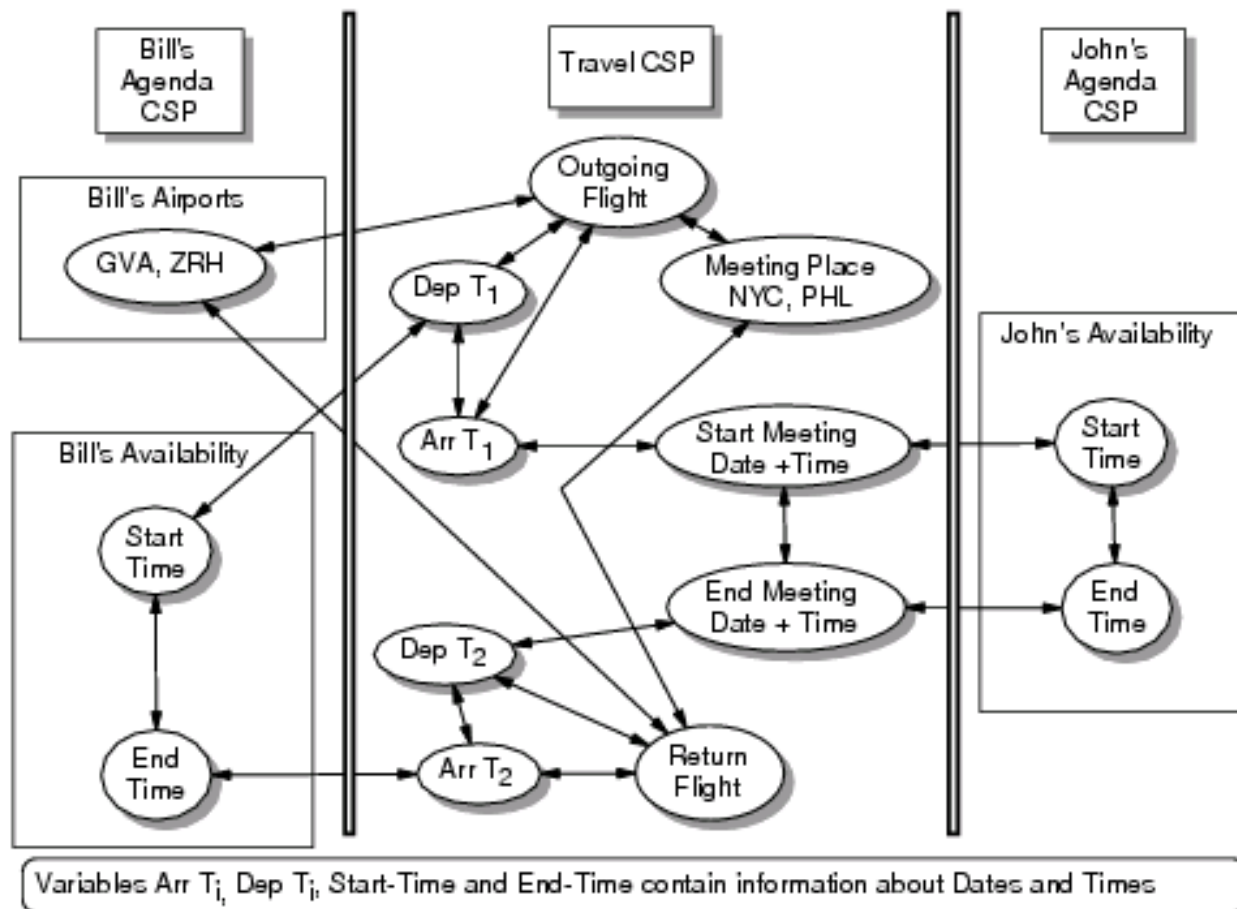
- A Constraint Satisfaction Problem (CSP) is:
 - a set of variables each with
 - a set of domain of values, and
 - a set of constraints that define which combinations of variable values are allowed
 - the task is to find a value for each variable such that all constraints are simultaneously satisfied
- Algorithms and techniques for solving CSPs have been widely studied



Conjunctive Normal Form: Example

- As a CSP:
 - Each clause is a constraint, each literal is a variable with a {true, false} domain
- $\{\neg X \vee \neg y \vee \neg Z\} \wedge \{X \vee y \vee \neg Z\} \wedge \{\neg X \vee y\} \wedge \{\neg y \vee Z\}$
- To Solve:
 - All clauses evaluate true
 - e.g.: $x = \text{false}, y = \text{true}, z = \text{true}$

Example CSP Graph for Travel





Formalization of Example: Variables

- $X = \{DT_0, \dots, DT_{n-1}, AT_0, \dots, AT_{n-1}, Airports_0, \dots, Airports_n, Flights_0, \dots, Flights_{n-1}, AirCrafts, Fares, Airlines, \dots\}$ is a set of variables
 - DT_i and AT_i represent the dates and times on which the traveler could depart and arrive respectively
 - $Airports_i$ represents the possible airports near the departure for leg of the itinerary
 - $Flights_i$ stands for the possible flights between the airports of $Airports_i$ and $Airports_{i+1}$



Formalization of Example: Domains

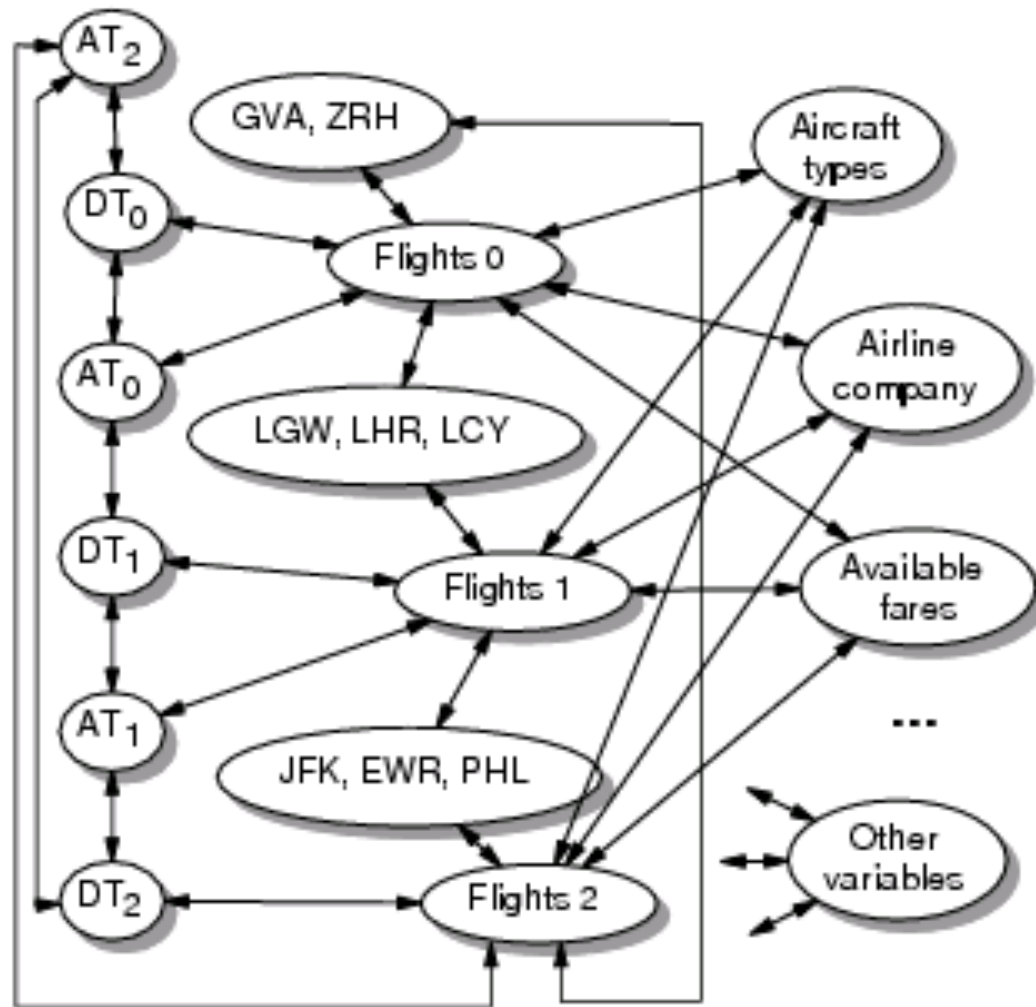
- $D = \{D_1, \dots, D_n\}$ is the set of domains
 - For variables DT_i or AT_i : the domain contains all possible departure and arrival times for the leg $_i$
 - For variables $Airports_i$: the domain is a set of airports for the departure of the leg $_i$
 - For variables $Flights_i$: the domain is the set of possible flights from $Airports_i$ to $Airports_{i+1}$
 - For variables $AirCrafts$, $Fares$ and $Airlines$: the domain is the set of different aircrafts, the set of available fares or the set of airline companies respectively



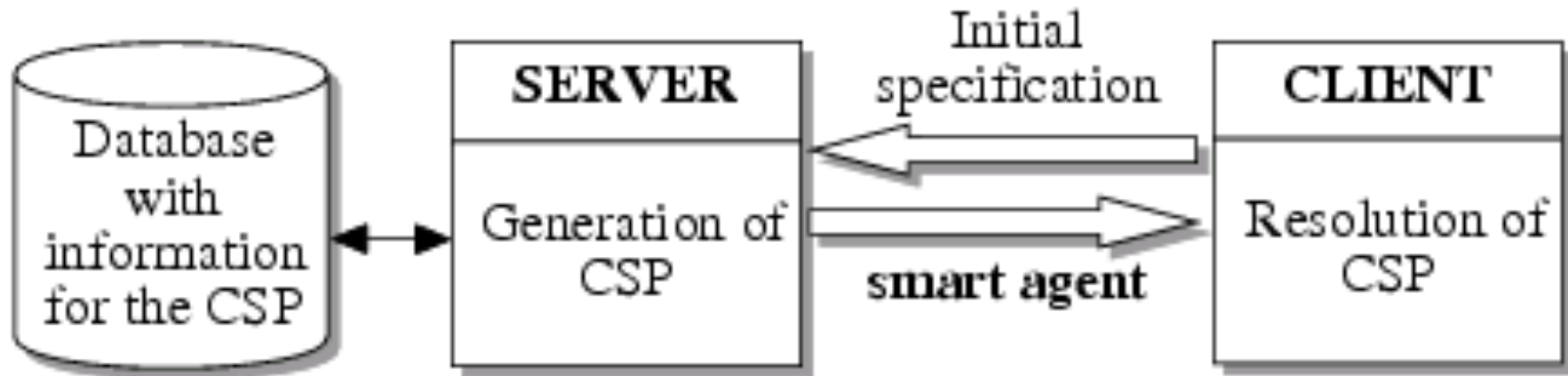
Formalization of Example: Constraints

- $C = \{C_1, \dots, C_n\}$ is the set of constraints
 - Two types of constraints:
 - Those imposed by the user's preferences
 - Those imposed by flight schedules
 - There are constraints on the variables Flights_i , Airports_i , DT_i and AT_i that guarantee the the flight is compatible with the airports, departure times and arrival times
 - A binary constraint in between AT_i and DT_{i+1} takes into consideration that the flight for leg_{i+1} departs after the flight for leg_i arrives
 - User preferences are expressed by means of constraints between Flight_i variables and Aircrafts, Airlines, Fares, and other variables

Constraint Graph for Flights



Architecture for SmartClients





Pros and Cons

- Pros

- Elegant approach that exploits past work on CSPs
- Minimizes the data retrieval and supports complex reasoning and integration of the data

- Cons

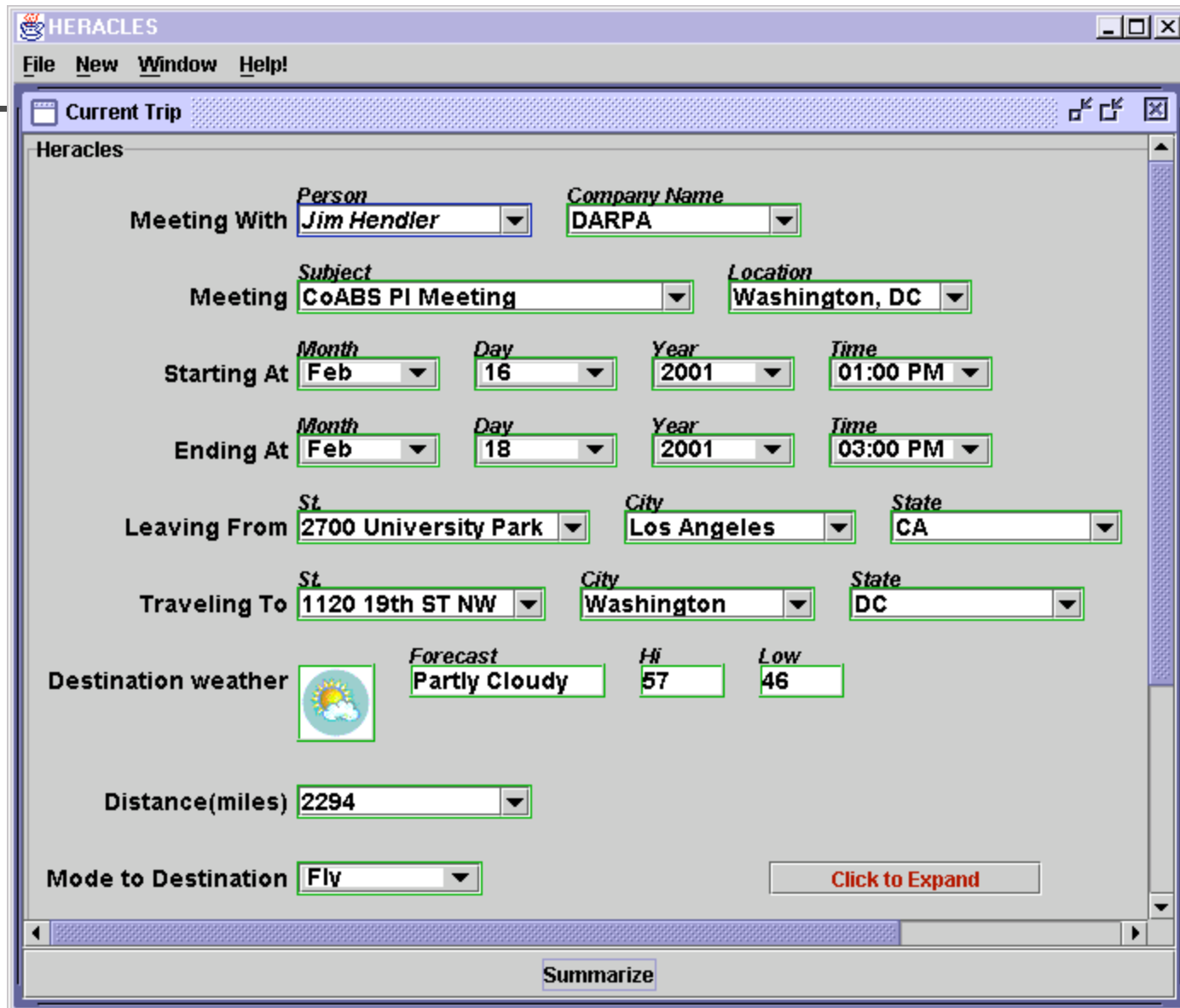
- Assumes that all data can be retrieved before any reasoning about the data
- In the travel planning, assumes that prices are the same on any date and there are no issues with flight availability




Heracles Constraint Integration Framework

- Framework for building integrated applications
- Interleaves planning and information gathering
- Uses a constraint reasoner to decide what sources to query and to integrate the results

The Travel Assistant



The image shows a screenshot of the HERACLES Travel Assistant software interface. The window title is "HERACLES" and it has a menu bar with "File", "New", "Window", and "Help!". The main area is titled "Current Trip" and contains the following fields:

- Meeting With:** Person: ; Company Name:
- Meeting:** Subject: ; Location:
- Starting At:** Month: ; Day: ; Year: ; Time:
- Ending At:** Month: ; Day: ; Year: ; Time:
- Leaving From:** St: ; City: ; State:
- Traveling To:** St: ; City: ; State:
- Destination weather:**  Forecast: ; Hi: ; Low:
- Distance(miles):**
- Mode to Destination:**

At the bottom right, there is a button labeled "Click to Expand". At the bottom center, there is a button labeled "Summarize".

Dynamically Updates Slots as Information Becomes Available

HERACLES

File New Window Help!

Current Trip

Fly

Choose Flights Based on: **BLACK** Lowest Price

Departing From: Code GREEN LAX Name GREEN LOS ANGELES INTL

Arriving In: Code GREEN DCA Name GREEN NATIONAL APT

Airline: IAD RoundTrip Fare GREEN 389

Flight: BWI Warning GREEN 1 long layover

HGR

Departure: SBY Day GREEN 15

MDT Day GREEN 15

Arrival: CHO

LNS

Summarize

HERACLES

File New Window Help!

Current Trip

Fly

Choose Flights Based on: **BLACK** Lowest Price

Departing From: Code GREEN LAX Name GREEN LOS ANGELES INTL

Arriving In: Code BLUE IAD Name GREEN WASHINGTON DULLES

Airline: Continental BLUE RoundTrip Fare RED 389

Flight: CLE RED Warning RED 1 long layover Class RED Coach

HGR

Departure: Month RED Mar Day RED 15 Time RED 6:30 am

Arrival: Month RED Mar Day RED 15 Time RED 6:46 pm

Summarize

Supports Informed Choices

HERACLES
File New Window Help!

Current Trip

Airline: **Continental** RoundTrip Fare: **389**

Flight: Stops: **CLE** Warning: **2 prop plane segments** Class: **Coach**

Departure: Month: **Mar** Day: **15** Time: **6:30 am**

Arrival: Month: **Mar** Day: **15** Time: **4:19 pm**

Parking: Lot: **Terminal Parking** Daily Rate(dollars): **16.00** Duration(days): **4** Total(dollars): **64**

Taxi: Dist2Airport: **15.1** Taxi fare(dollars): **23.00**

Mode to Departure Airport: **Take a Taxi** [Click to hide details](#)

Take a Taxi

Leaving From: St: **2700 University Park** City: **Los Angeles** State: **CA**

Driving To: St: **LOS ANGELES INTL** City: **Los Angeles** State: **CA**

Suggested Departure: Month: **Mar** Day: **15** Year: **2001** Time: **05:08 AM**

Predicted Arrival: Month: **Mar** Day: **15** Year: **2001** Time: **05:30 AM**

Taxi fare(dollars): **23.00**

Summarize

HERACLES
File New Window Help!

Current Trip

Airline: **Continental** RoundTrip Fare: **389**

Flight: Stops: **CLE** Warning: **2 prop plane segments** Class: **Coach**

Departure: Month: **Mar** Day: **15** Time: **6:30 am**

Arrival: Month: **Mar** Day: **15** Time: **4:19 pm**

Parking: Lot: **Economy Lot C *** Daily Rate(dollars): **7.00** Duration(days): **4** Total(dollars): **28**

Taxi: **Economy Lot B *** Taxi fare(dollars): **23.00**

Mode to Departure Airport: **Economy Lot C *** [Click to hide details](#)

Drive

Leaving From: St: **2700 University Park** City: **Los Angeles** State: **CA**

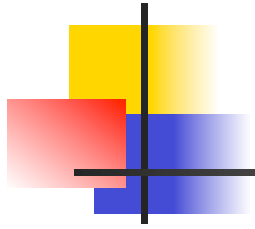
Driving To: St: **LOS ANGELES INTL** City: **Los Angeles** State: **CA**

Suggested Departure: Month: **Mar** Day: **15** Year: **2001** Time: **05:08 AM**

Predicted Arrival: Month: **Mar** Day: **15** Year: **2001** Time: **05:30 AM**

Summarize

Changes Propagate Throughout



HERACLES
File New Window Help!

Current Trip

Fly

Choose Flights Based on: **Lowest Price**

Departing From: Code **LAX** Name **LOS ANGELES INTL**

Arriving In: Code **BUR** Name **NATIONAL APT**

Airline: **LGB** Round Trip Fare: **371**

Flight: **SNA** Warning: **1 long layover** Class: **Coach**

Departure: **ONT** Day: **15**

Arrival: **OXR** Day: **15**

Departure: **SBD**

Arrival: **CLD**

HERACLES
File New Window Help!

Current Trip

Take a Taxi

Leaving From: St. **2700 University Park** City: **Los Angeles** State: **CA**

Driving To: **LOS ANGELES INTL** City: **Los Angeles** State: **CA**

Suggested Departure: Month **Mar** Day **15** Year **2001** Time **05:08 AM**

Predicted Arrival: Month **Mar** Day **15** Year **2001** Time **05:30 AM**

Taxi fare(dollars) **23.00**

Total Drive: Distance(miles) **15.1** Hrs: **0** Mins: **22**

Maps

© 2000 MapQuest.com, Inc. © 2000 Navigation Technologies

Summarize

HERACLES
File New Window Help!

Current Trip

Take a Taxi

Leaving From: St. **2700 University Park** City: **Los Angeles** State: **CA**

Driving To: **LONG BEACH** City: **Los Angeles** State: **CA**

Suggested Departure: Month **Mar** Day **14** Year **2001** Time **04:30 PM**

Predicted Arrival: Month **Mar** Day **14** Year **2001** Time **05:04 PM**

Taxi fare(dollars) **34.20**

Total Drive: Distance(miles) **23.5** Hrs: **0** Mins: **34**

Maps

© 2000 MapQuest.com, Inc. © 2000 Navigation Technologies

Summarize

User Can Specify High-Level Preferences

HERACLES

File New Window Help!

Current Trip

Hotel

Preference **Choose Hotels Based on:** **Closest to Meeting** **Preferred Type:** **Normal** **Preferred Amenities:** **Business Center**

Location **Washington**

Check in **Mon** **Mar** **15** **2001**

Check out **Mon** **Mar** **19** **2001**

Name **Quality Inn Iwo Jima**

Address **St. 1501 ARLINGTON BLVD.** **City ARLINGTON** **State VA**

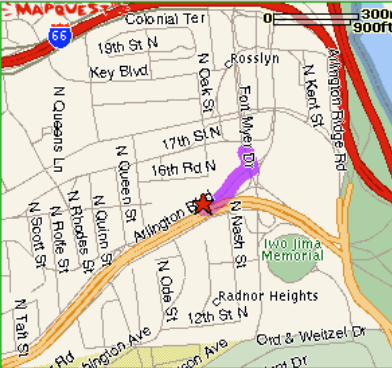
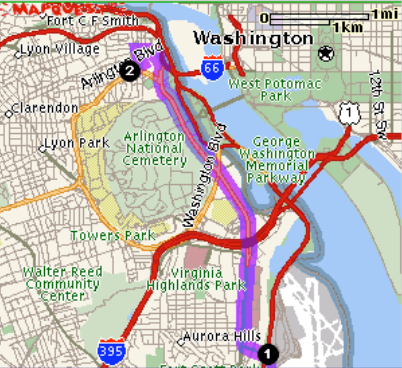
PHONE **703 524-5000**

FAX **703 522-5484**

Price **Daily Rate 82.00** **# of Days 4** **Total 328**

Driving **Distance(miles) 4.1** **Hrs 0** **Mins 9**

Maps



Summarize

HERACLES

File New Window Help!

Current Trip

Hotel

Preference **Choose Hotels Based on:** **Closest to Airport** **Preferred Type:** **Normal** **Preferred Amenities:** **Business Center**

Location **Closest to Airport**

Check in **Mon** **Mar** **5** **2001**

Check out **Mon** **Mar** **9** **2001**

Name **Econo Lodge National Airport**

Address **St. 2485 S. GLEBE RD.** **City ARLINGTON** **State VA**

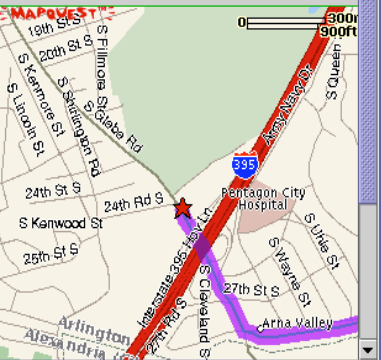
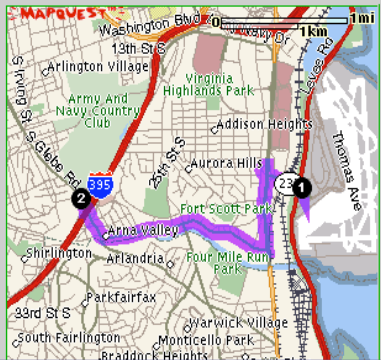
PHONE **703 979-4100**

FAX **703 979-6120**

Price **Daily Rate 64.00** **# of Days 4** **Total 256**

Driving **Distance(miles) 2.7** **Hrs 0** **Mins 6**

Maps



Summarize



Constraint Networks for Integrating Information

- Components:
 - Representation of the variables
 - Representation of constraints
 - Hierarchical template representation
 - Constraint propagation and cycle detection

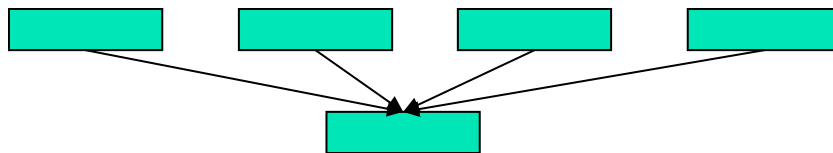


Constraint Networks for Integrating Information

- Components:
 - Representation of the variables
 - Representation of constraints
 - Hierarchical template representation
 - Constraint propagation and cycle detection

Constraint Variables

- Constraint network consists of a set of variables such as:
 - MeetingStartTime
 - MeetingLocation
- Each variable depends on a set of ancestors.



- Variables are related by constraints that determine the possible values of a solution



Constraint Networks for Integrating Information

- Components:
 - Representation of the variables
 - Representation of constraints
 - Hierarchical template representation
 - Constraint propagation and cycle detection



Constraint Representation

- Constraints are computable components:
 - Local calculations based on Xquery
 - MeetingStartTime + MeetingDuration --> MeetingEndTime
 - Web and Database Wrappers
 - ITN: DepartureAirport, ArrivalAirport, Date --> Flights
 - Yahoo Weather: City, Date --> Weather predication
 - External Programs (Outlook, Planners, etc)
 - Outlook Calendar: Date --> Meetings

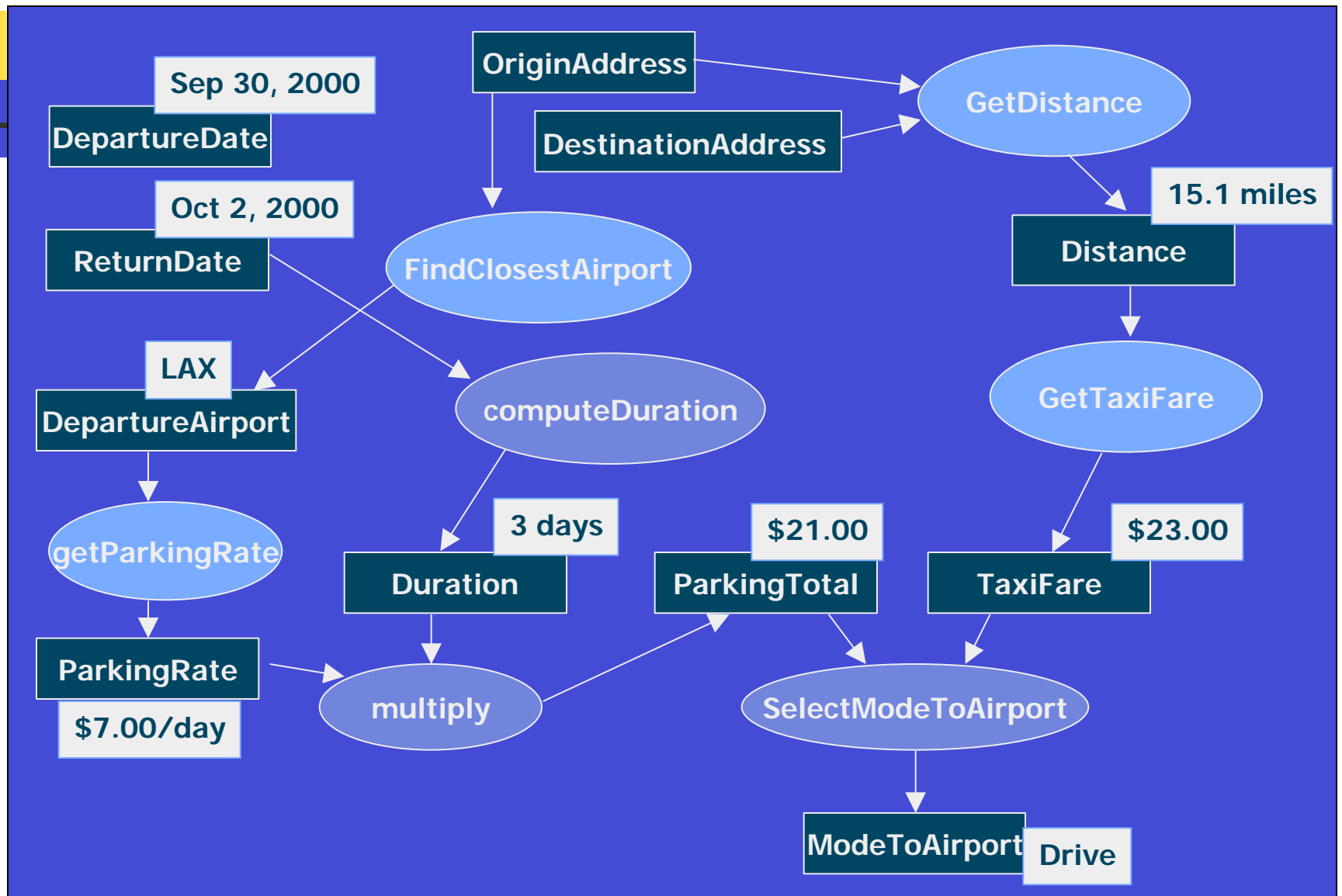


Constraint Structure

Constraint

- Arguments: input and output variables
- Call:
 - Construct table with inputs and corresponding calls (http requests, SQL queries, etc) to sources (wrappers, DBs, etc) [using XML Query]
 - Calls are executed and results stored in an table
- Output
 - Restructure source results into desired output [using XML Query]

Drive or Take a Taxi?





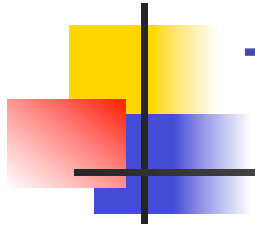
Constraint Networks for Integrating Information

- Components:
 - Representation of the variables
 - Representation of constraints
 - Hierarchical template representation
 - Constraint propagation and cycle detection



Hierarchically-Partitioned Constraint Networks

- Template:
 - Groups related variables and constraints
 - Organizes information for computation and presentation to user
- Templates organized hierarchically
 - Template decomposed into subtemplates
 - Choose among alternative subtemplates

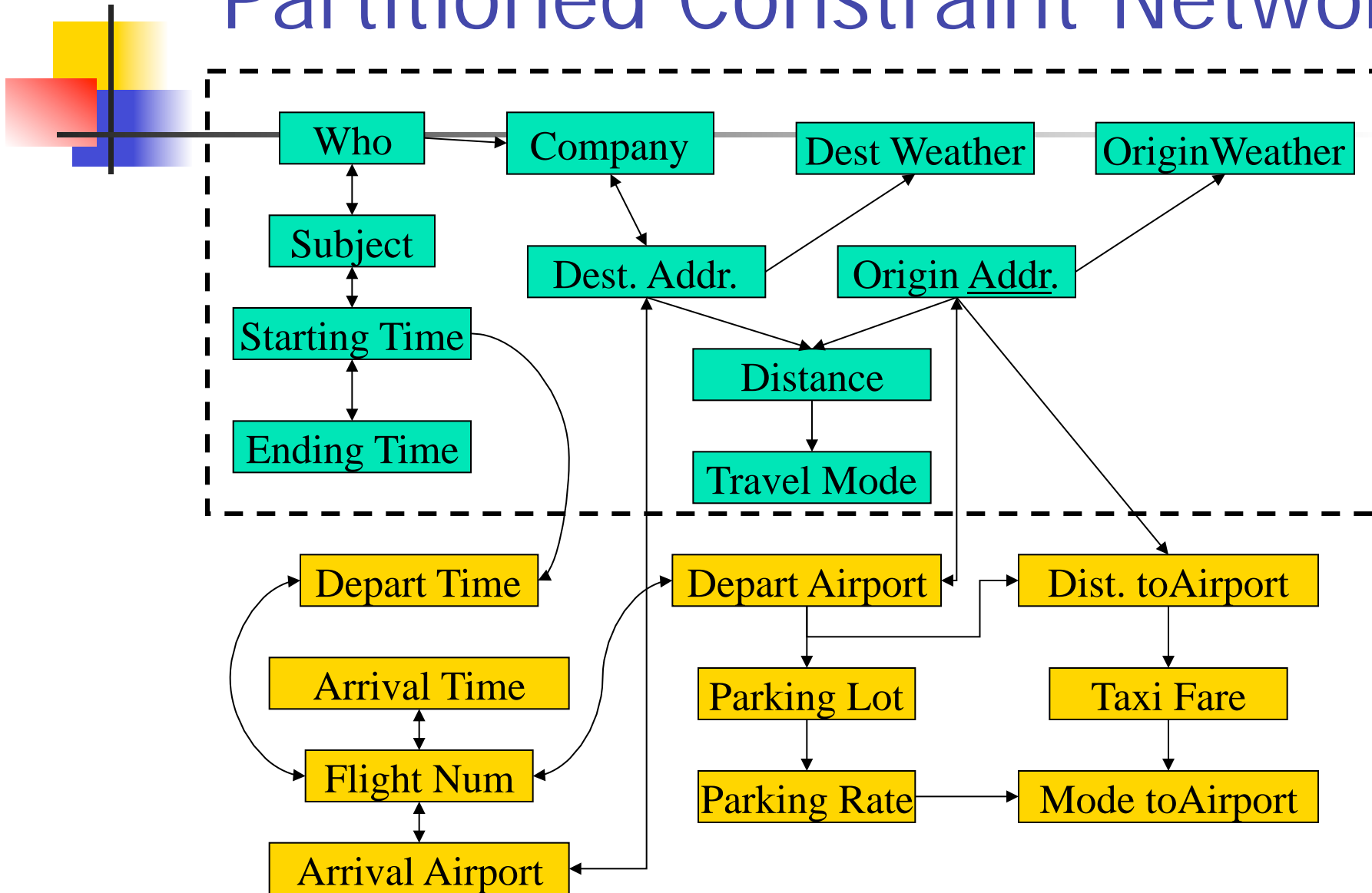


Template Structure

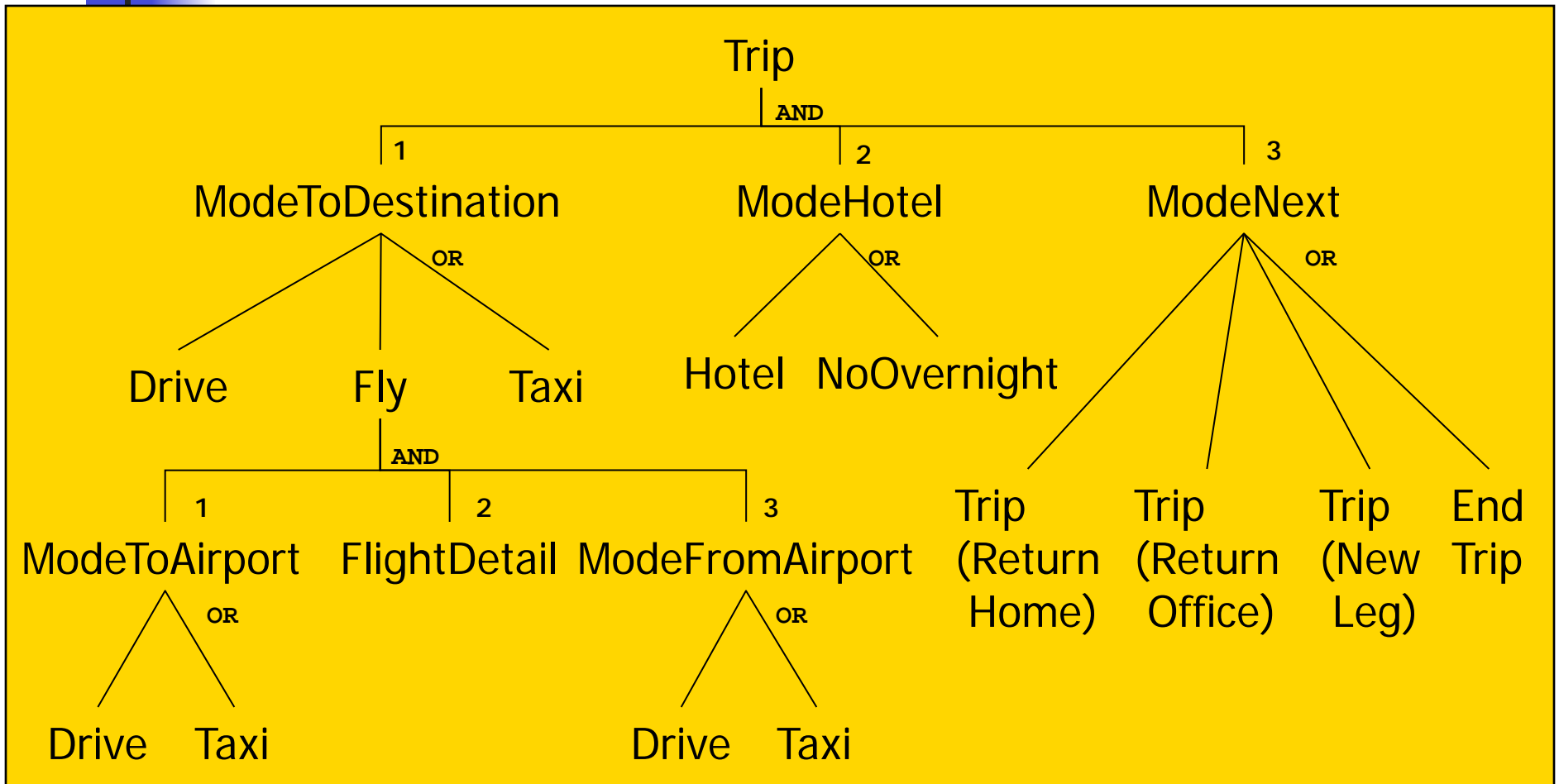
Template

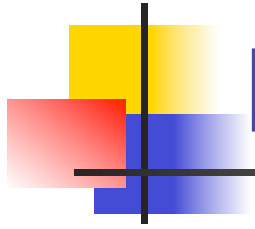
- Arguments: input and output variables
- Variables: name, type, default values
- Constraints
- Expansions: alternative subtemplate calls
- GUI specification

Partitioned Constraint Network



Template Hierarchy for the Travel Assistant





Dynamic Constraint Networks

Generalization of Constraint Networks

- Variables can be active or inactive
- Normal Constraints

$$x_1 = k_1 \wedge \dots \wedge x_m = k_m \rightarrow x_n = k_n$$

- Activity constraints:

$$x_1 = k_1 \wedge \dots \wedge x_m = k_m \rightarrow \text{active}(x_n)$$

- Inactive variables do not participate in the network, i.e., do not propagate constraints



Heracles: Template Selection

- Core network
 - Computes values of template selection vars
 - Always active
- Template selection variables
 - Inputs to activity constraints: determine the choice of subtemplates, i.e., which additional variables are active



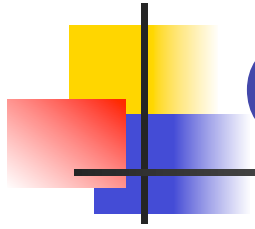
Constraint Networks for Integrating Information

- Components:
 - Representation of the variables
 - Representation of constraints
 - Hierarchical template representation
 - Constraint propagation and cycle detection



Constraint Propagation

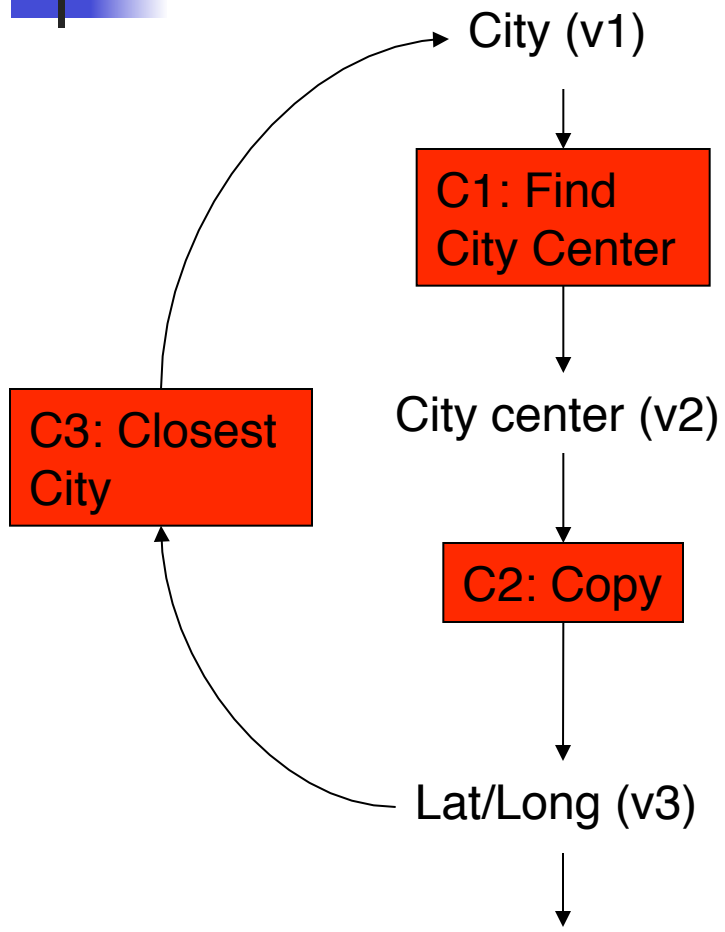
- Approach
 - When a variable is assigned a value, re-compute the value sets and assigned values of all dependent variables
 - Proceeds recursively until no values are changed or a cycle is detected
- Core network
 - Propagates all variables through the core network
 - Remaining variables are computing when a template is opened
- Does not perform full CSP
 - Less costly
 - Does not require all information in advance
 - Makes choices locally, so may fail to find optimal assignment



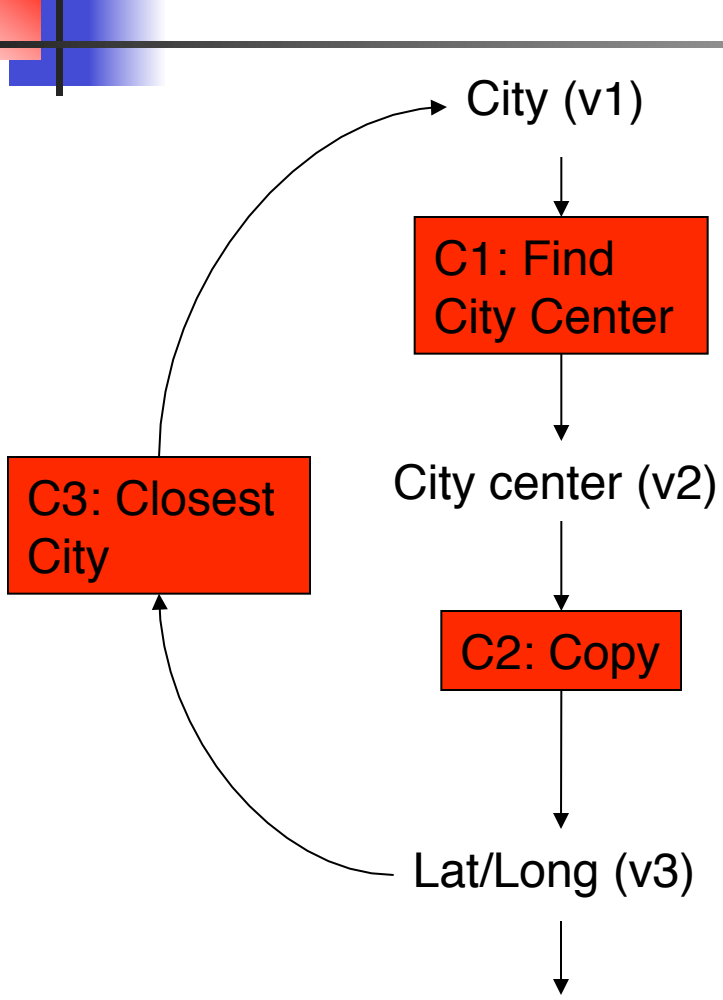
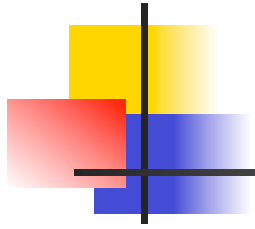
Cycle Detection

- Address cyclic *interactive* networks
 - Multiple input paths:
 - region/country/city vs. lat/long
 - Conversion rounding errors:
 - lat/long, temperature, ...
- => Cycle detection in constraint propagation

Interactive Cyclic Network

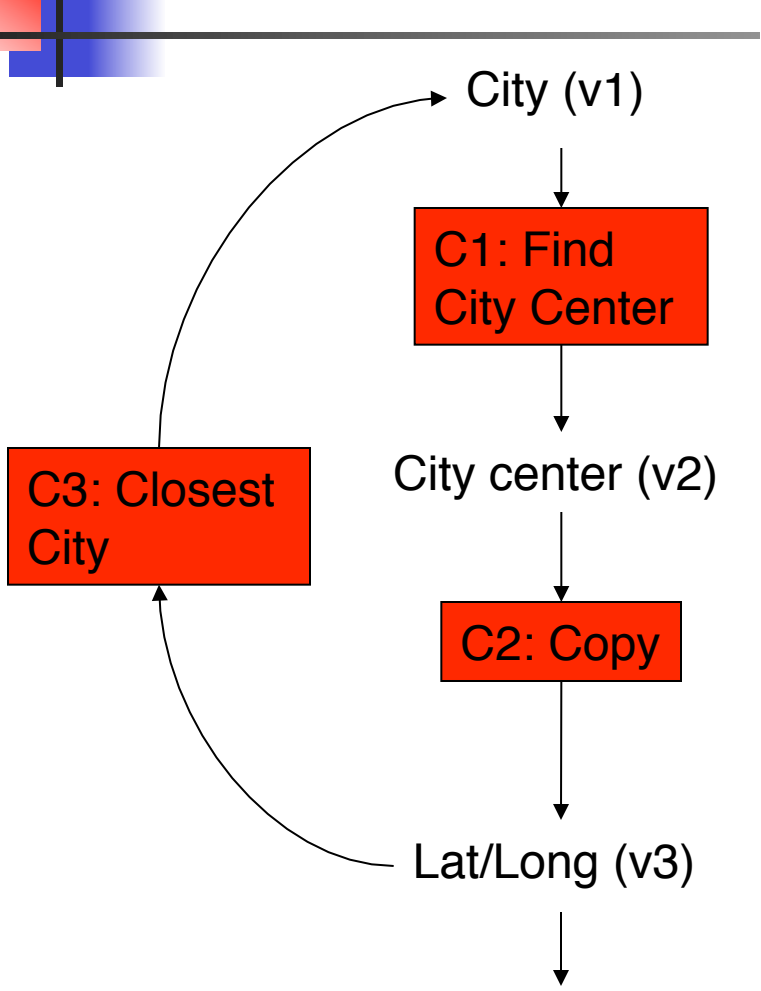
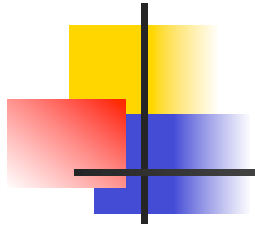


Interactive Cyclic Network



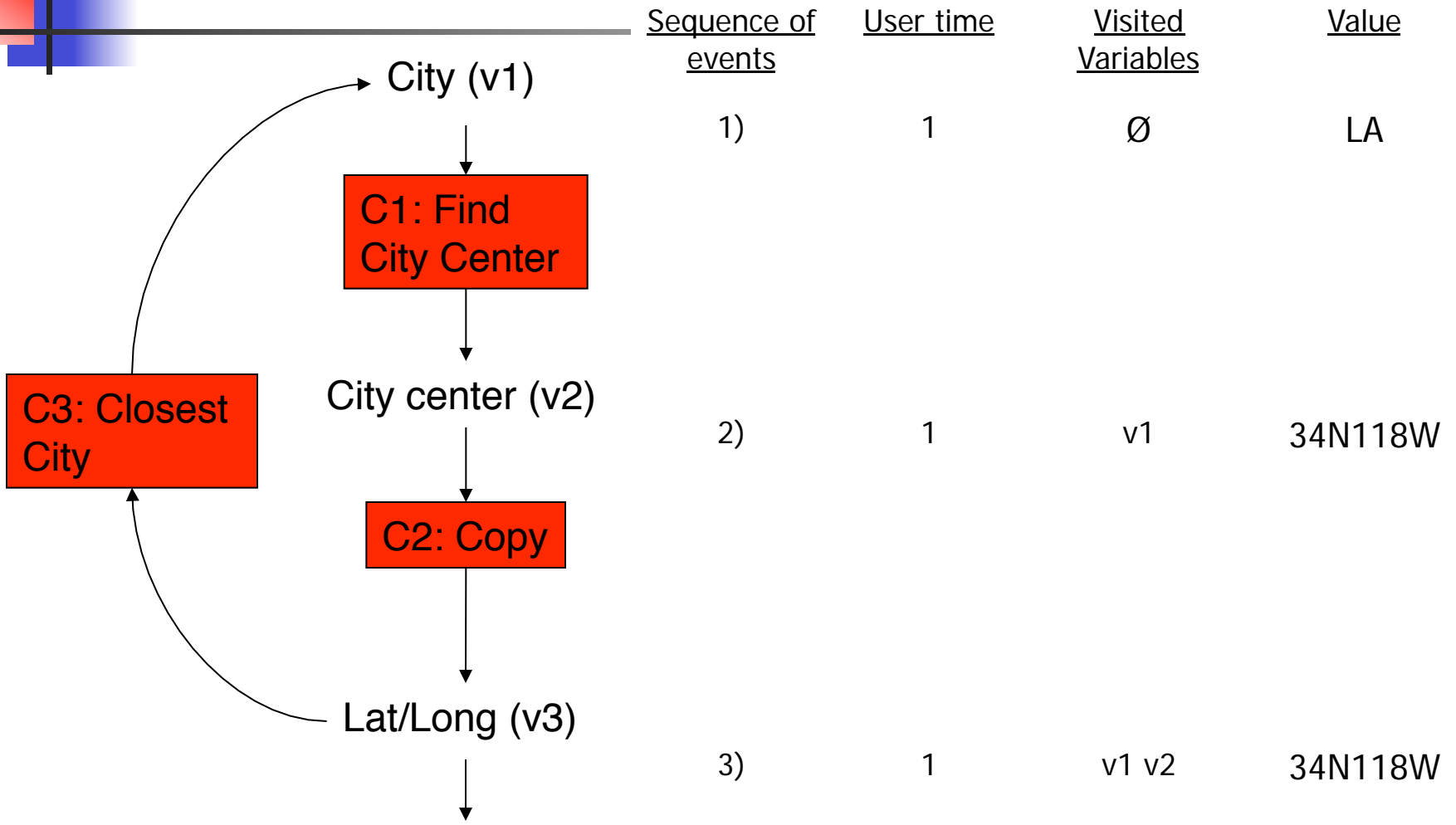
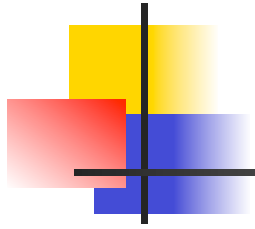
<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	∅	LA

Interactive Cyclic Network



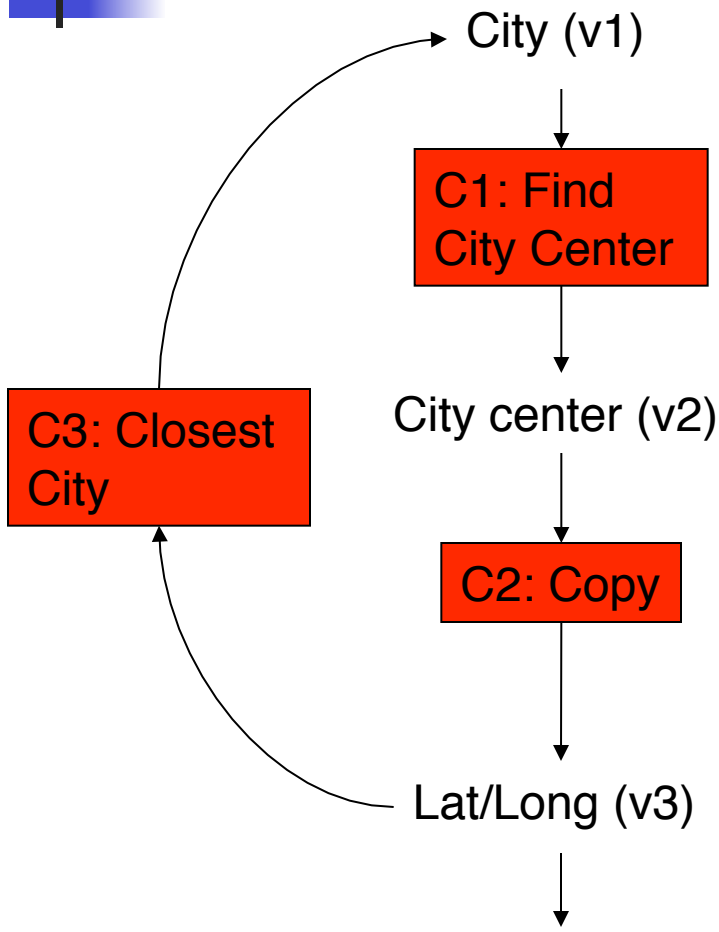
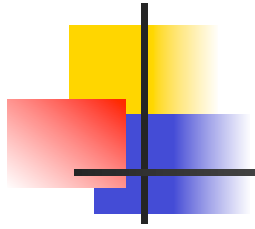
<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	∅	LA
2)	1	v1	34N118W

Interactive Cyclic Network



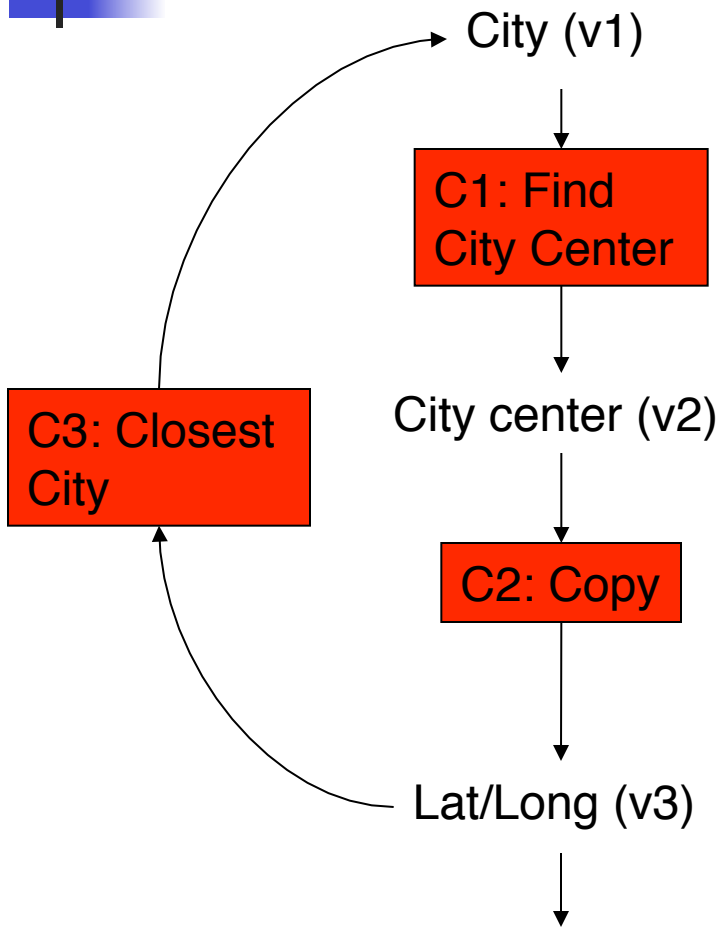
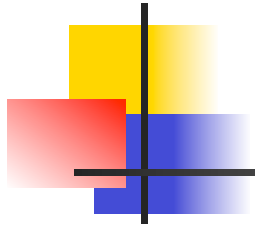
<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	∅	LA
2)	1	v1	34N118W
3)	1	v1 v2	34N118W

Interactive Cyclic Network



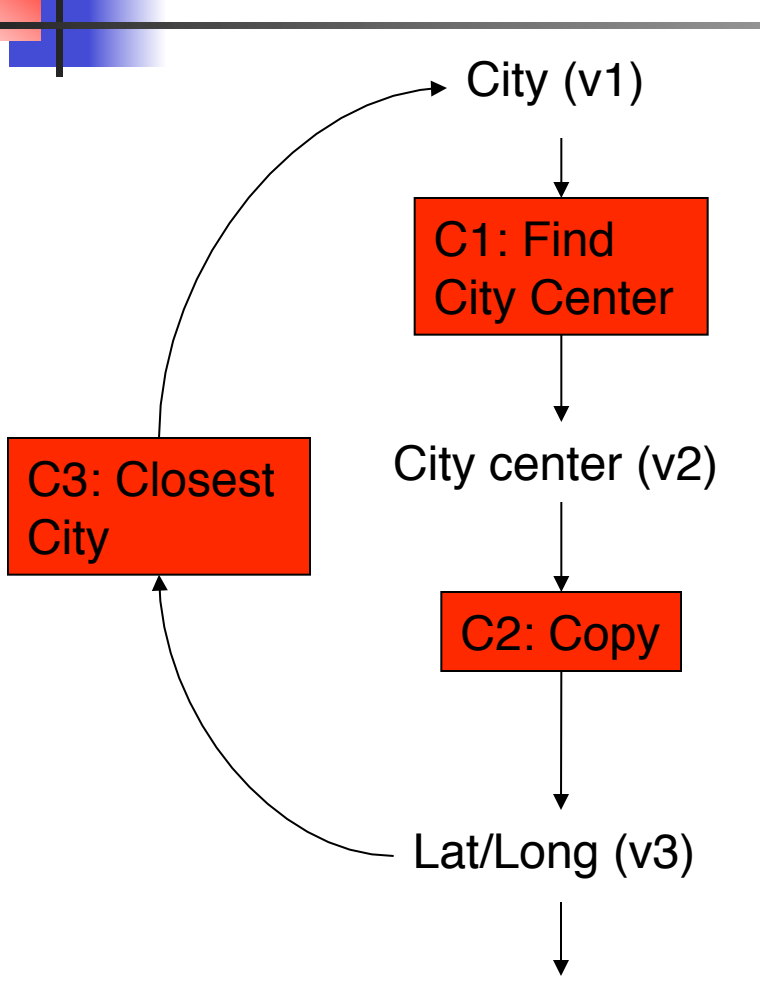
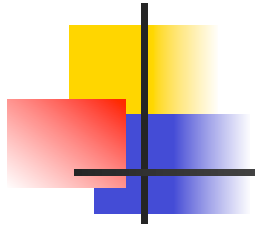
<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	\emptyset	LA
4)	Blocked!	$t(v3) = t(v1) \wedge v1 \in vis(v3)$	
2)	1	v1	34N118W
3)	1	v1 v2	34N118W

Interactive Cyclic Network



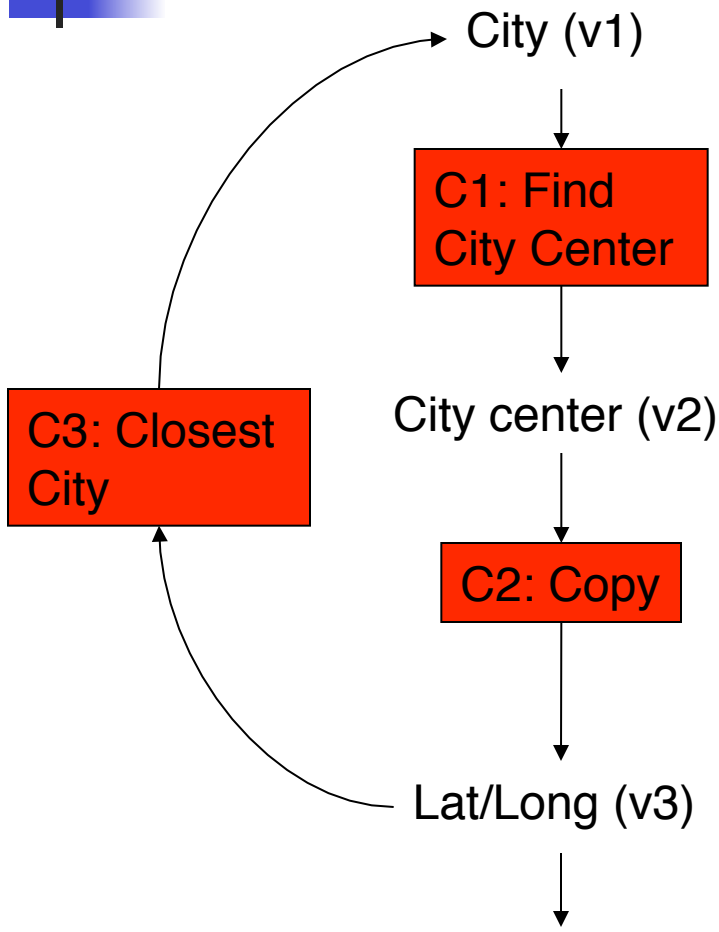
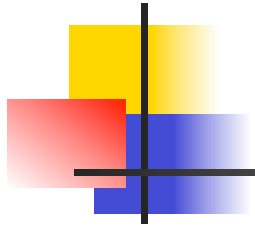
<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	\emptyset	LA
4)	Blocked!	$t(v3) = t(v1) \wedge v1 \in vis(v3)$	
2)	1	v1	34N118W
3)	1	v1 v2	34N118W
5)	2	\emptyset	40N70W

Interactive Cyclic Network



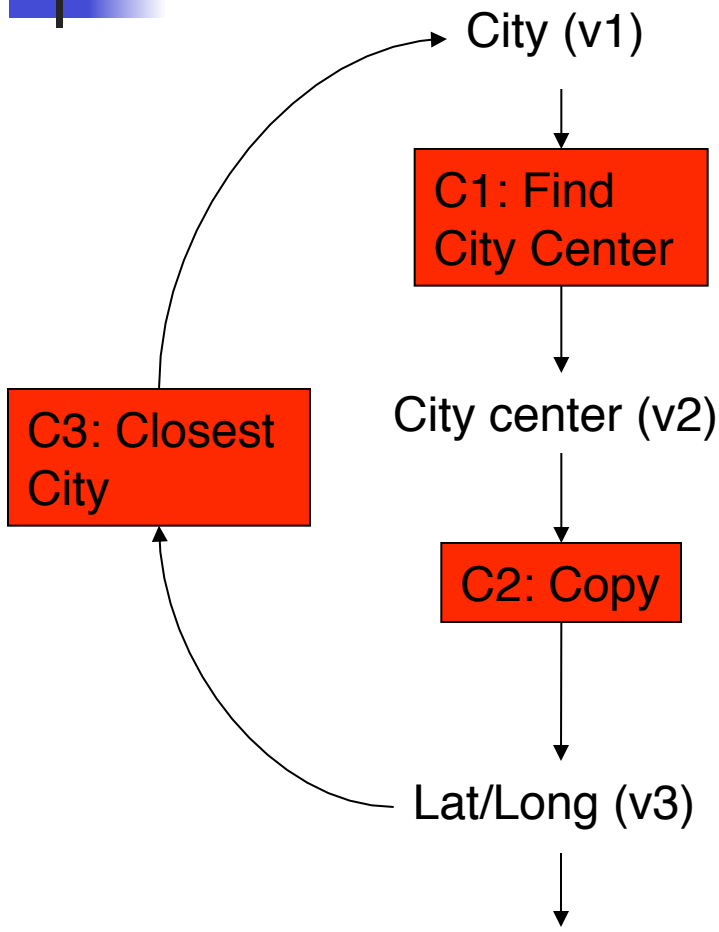
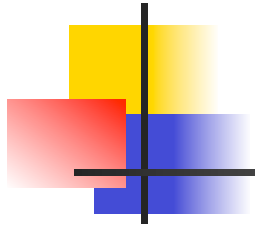
<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	∅	LA
4)	Blocked!	$t(v3) = t(v1) \wedge v1 \in vis(v3)$	
6)	2	v3	NY
2)	1	v1	34N118W
3)	1	v1 v2	34N118W
5)	2	∅	40N70W

Interactive Cyclic Network



<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	\emptyset	LA
4)	Blocked!	$t(v3) = t(v1) \wedge v1 \in vis(v3)$	
6)	2	v3	NY
2)	1	v1	34N118W
7)	2	v3 v1	40N73W
3)	1	v1 v2	34N118W
5)	2	\emptyset	40N70W

Interactive Cyclic Network



<u>Sequence of events</u>	<u>User time</u>	<u>Visited Variables</u>	<u>Value</u>
1)	1	\emptyset	LA
4)	Blocked!	$t(v3) = t(v1) \wedge v1 \in vis(v3)$	
6)	2	v3	NY
2)	1	v1	34N118W
7)	2	v3 v1	40N73W
3)	1	v1 v2	34N118W
5)	2	\emptyset	40N70W
8)	Blocked!	$t(v3) = t(v2) \wedge v3 \in vis(v2)$	



Discussion

- General framework for interleaving planning and information gathering
 - Retrieves information as needed
 - Gathers and integrates data in a uniform framework
 - Evaluates tradeoffs and selects among alternatives
 - Allows the users to explore alternatives
 - Supports a wide variety of information types: databases, web pages, images, video, etc.