UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

CS411 - Extreme Indexing

illinois.edu

Announcements

- Get MP2 finished
- Enjoy your break!

Extreme Indexing

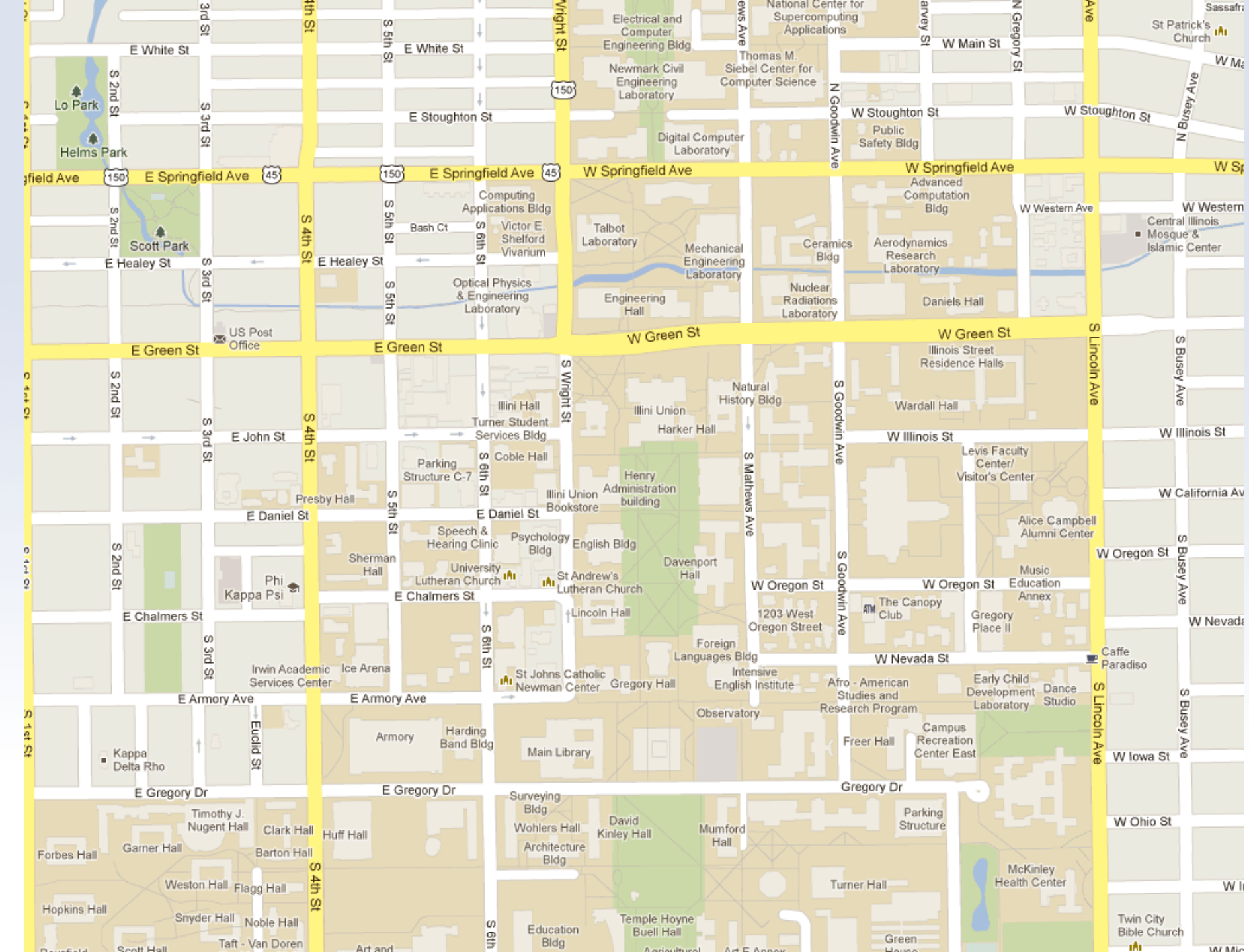
- Indexing tricks we've used so far are great for relational data.
- What if we wanted to query other kinds of data quickly?

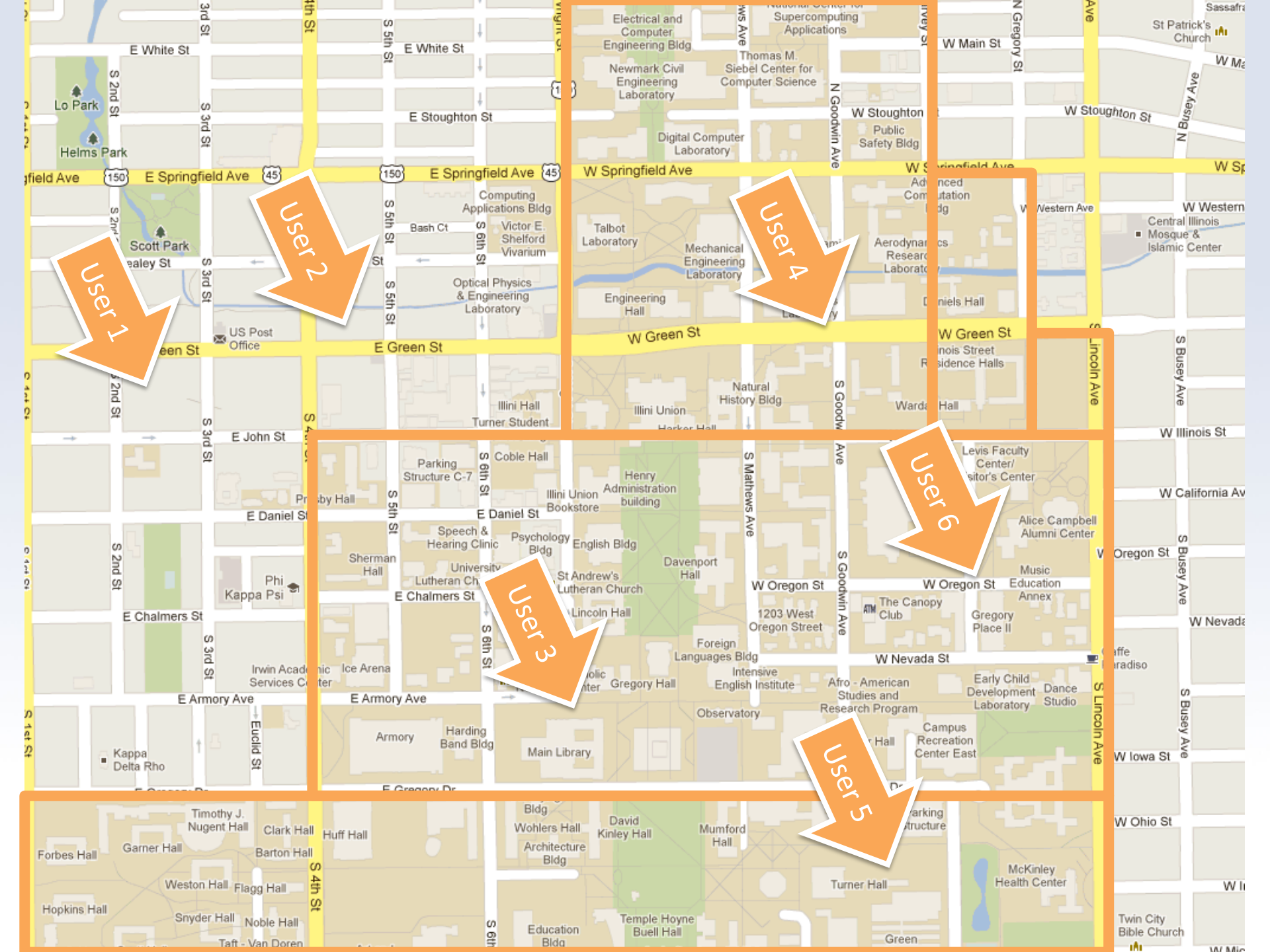
Examples

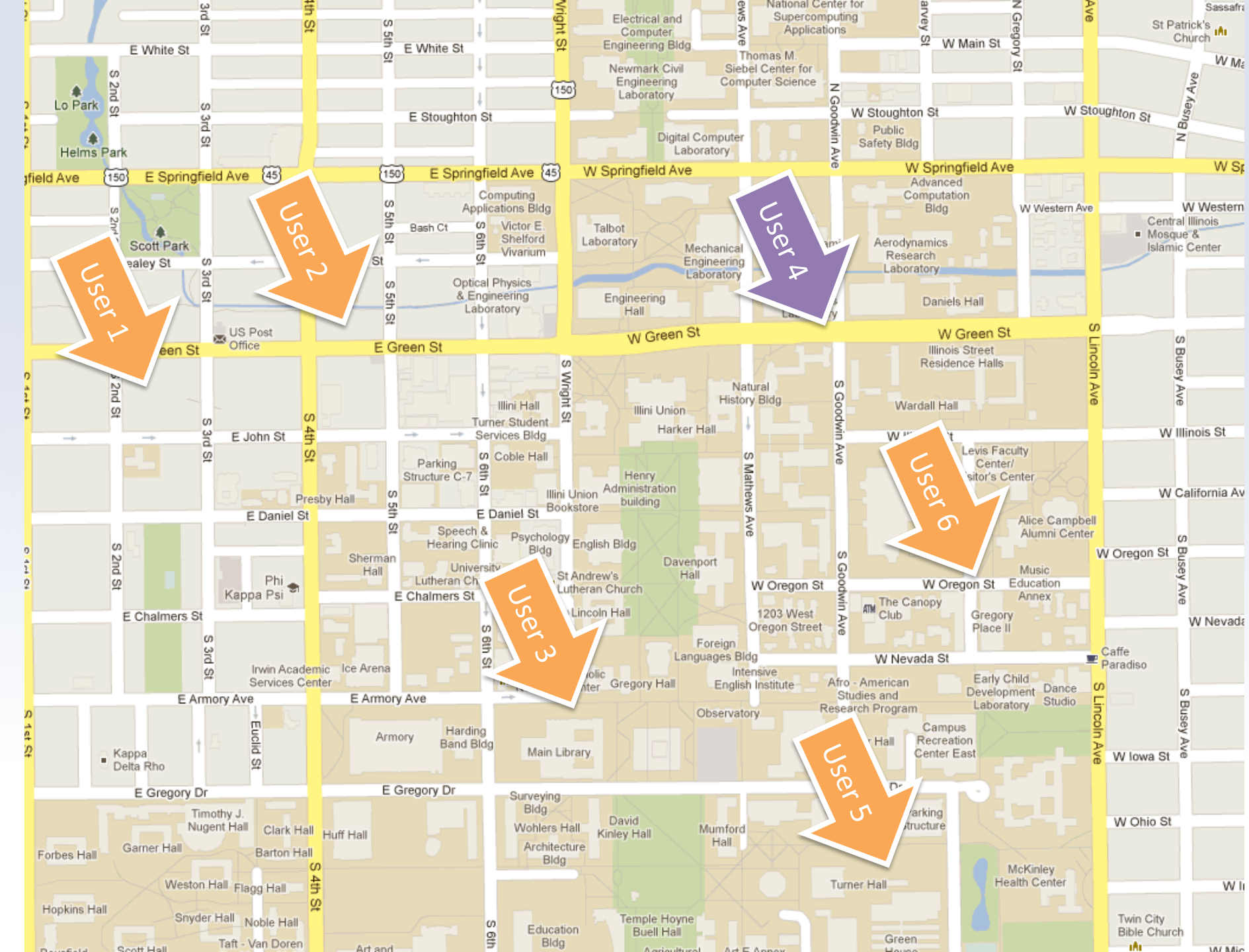
- Where is the nearest grocery store?
- What coffee shop am I in now?
- What web pages contain the words “Belle” and “Sebastian”?
- Where in the human genome (a string of 3 billion characters) does the string “GATTACA” appear?

Multidimensional Data

- We're focus on spatial data
- Queries involve points, lines, and polygons
  - Range queries: find points/shapes in a region
  - Nearest-neighbor: find the nearest point
  - Location queries: find shapes containing a point
- Our examples will be two dimensions: x,y


E White St
S 2nd St
S 3rd St
S 4th St
S 5th St
S Wright St
Lo Park
Helms Park
E Stoughton St
Electrical and Computer Engineering Bldg
National Center for Supercomputing Applications
Thomas M. Siebel Center for Computer Science
Newmark Civil Engineering Laboratory
W Main St
N Gregory St
St Patrick's Church
N Goodwin Ave
W Stoughton St
Public Safety Bldg
Digital Computer Laboratory
E Springfield Ave
W Springfield Ave
Advanced Computation Bldg
W Western Ave
Central Illinois Mosque & Islamic Center
Computing Applications Bldg
Victor E. Shelford Vivarium
Bash Ct
Scott Park
E Healey St
Talbot Laboratory
Mechanical Engineering Laboratory
Ceramics Bldg
Aerodynamics Research Laboratory
Optical Physics & Engineering Laboratory
Engineering Hall
Nuclear Radiations Laboratory
Daniels Hall
US Post Office
E Green St
W Green St
S Lincoln Ave
Illinois Street Residence Halls
Natural History Bldg
Wardall Hall
Illini Hall
Turner Student Services Bldg
Illini Union
Harker Hall
S Busey Ave
E John St
W Illinois St
Parking Structure C-7
Coble Hall
Henry Administration building
Levis Faculty Center/ Visitor's Center
Presby Hall
Illini Union Bookstore
E Daniel St
W California Av
Speech & Hearing Clinic
Psychology Bldg
English Bldg
Alice Campbell Alumni Center
Sherman Hall
University Lutheran Church
St Andrew's Lutheran Church
Davenport Hall
W Oregon St
Music Education Annex
Phi Kappa Psi
E Chalmers St
Lincoln Hall
1203 West Oregon Street
The Canopy Club
Gregory Place II
W Nevada
Foreign Languages Bldg
W Nevada St
Caffe Paradiso
Irwin Academic Services Center
Ice Arena
St Johns Catholic Newman Center
Gregory Hall
Intensive English Institute
Afro - American Studies and Research Program
Early Child Development Laboratory
Dance Studio
E Armory Ave
Observatory
Armory
Harding Band Bldg
Main Library
Freer Hall
Campus Recreation Center East
W Iowa St
Kappa Delta Rho
Euclid St
E Gregory Dr
Gregory Dr
Surveying Bldg
Timothy J. Nugent Hall
Clark Hall
Huff Hall
Wohlers Hall
David Kinley Hall
Mumford Hall
Parking Structure
W Ohio St
Forbes Hall
Garner Hall
Barton Hall
Architecture Bldg
McKinley Health Center
Weston Hall
Flagg Hall
Turner Hall
Hopkins Hall
Snyder Hall
Noble Hall
Taft - Van Doren
Education Bldg
Temple Hoyne Buell Hall
Twin City Bible Church


User 1
User 2
User 3
User 4
User 5
User 6
E White St
E Springfield Ave
E Green St
E John St
E Daniel St
E Chalmers St
E Armory Ave
E Stoughton St
W Springfield Ave
W Green St
W Main St
W Stoughton St
W Oregon St
W Nevada St
W Illinois St
W California Ave
W Iowa St
W Ohio St
W Western Ave
S 2nd St
S 3rd St
S 4th St
S 5th St
S 6th St
S Mathews Ave
N Goodwin Ave
S Goodwin Ave
S Lincoln Ave
S Busey Ave
Lo Park
Helms Park
Scott Park
Bash Ct
Euclid St
Electrical and Computer Engineering Bldg
Supercomputing Applications
Thomas M. Siebel Center for Computer Science
Newmark Civil Engineering Laboratory
Digital Computer Laboratory
Public Safety Bldg
Computing Applications Bldg
Victor E. Shelford Vivarium
Talbot Laboratory
Mechanical Engineering Laboratory
Optical Physics & Engineering Laboratory
Engineering Hall
US Post Office
Illini Hall
Illini Union
Natural History Bldg
Parking Structure C-7
Coble Hall
Henry Administration building
Illini Union Bookstore
Speech & Hearing Clinic
Psychology Bldg
English Bldg
Sherman Hall
Davenport Hall
St Andrew's Lutheran Church
Lincoln Hall
Phi Kappa Psi
Kappa Delta Rho
Irwin Academic Services Center
Ice Arena
Armory
Harding Band Bldg
Main Library
Gregory Hall
Foreign Languages Bldg
Intensive English Institute
Observatory
Afro - American Studies and Research Program
Early Child Development Laboratory
Dance Studio
Campus Recreation Center East
1203 West Oregon Street
The Canopy Club
Gregory Place II
Music Education Annex
Alice Campbell Alumni Center
Levis Faculty Center/ Visitor's Center
Central Illinois Mosque & Islamic Center
St Patrick's Church
Timothy J. Nugent Hall
Clark Hall
Huff Hall
Garner Hall
Barton Hall
Forbes Hall
Weston Hall
Flagg Hall
Hopkins Hall
Snyder Hall
Noble Hall
Taft - Van Doren
Wohlers Hall
Architecture Bldg
David Kinley Hall
Mumford Hall
Turner Hall
McKinley Health Center
Education Bldg
Temple Hoyne Buell Hall
Twin City Bible Church


User 1
User 2
User 3
User 4
User 5
User 6
E White St
E Springfield Ave
E Green St
W Green St
W Springfield Ave
W Stoughton St
E Stoughton St
E John St
E Daniel St
E Chalmers St
E Armory Ave
E Gregory Dr
W Nevada St
W Oregon St
W Illinois St
W Main St
W Western Ave
W California Ave
W Iowa St
W Ohio St
S 1st St
S 2nd St
S 3rd St
S 4th St
S 5th St
S 6th St
S Wright St
S Goodwin Ave
S Mathews Ave
S Lincoln Ave
S Busey Ave
N Goodwin Ave
N Gregory St
N Busey Ave
Lo Park
Helms Park
Scott Park
US Post Office
Illini Hall
Turner Student Services Bldg
Coble Hall
Parking Structure C-7
Presby Hall
Illini Union Bookstore
Henry Administration building
Illini Union
Harker Hall
Natural History Bldg
Wardall Hall
Illinois Street Residence Halls
Levis Faculty Center/ Visitor's Center
Alice Campbell Alumni Center
Music Education Annex
Davenport Hall
English Bldg
Psychology Bldg
Speech & Hearing Clinic
Sherman Hall
St Andrew's Lutheran Church
Lincoln Hall
1203 West Oregon Street
The Canopy Club
Gregory Place II
Caffe Paradiso
Foreign Languages Bldg
Intensive English Institute
Afro - American Studies and Research Program
Early Child Development Laboratory
Dance Studio
Observatory
Gregory Hall
Irwin Academic Services Center
Ice Arena
Armory
Harding Band Bldg
Main Library
Campus Recreation Center East
Kappa Delta Rho
Phi Kappa Psi
Timothy J. Nugent Hall
Clark Hall
Huff Hall
Garner Hall
Barton Hall
Forbes Hall
Weston Hall
Flagg Hall
Hopkins Hall
Snyder Hall
Noble Hall
Taft - Van Doren
Surveying Bldg
Wohlers Hall
Architecture Bldg
David Kinley Hall
Mumford Hall
Turner Hall
McKinley Health Center
Education Bldg
Temple Hoyne Buell Hall
Green
Twin City Bible Church
Electrical and Computer Engineering Bldg
Newmark Civil Engineering Laboratory
Thomas M. Siebel Center for Computer Science
National Center for Supercomputing Applications
Digital Computer Laboratory
Public Safety Bldg
Advanced Computation Bldg
Computing Applications Bldg
Victor E. Shelford Vivarium
Bash Ct
Talbot Laboratory
Mechanical Engineering Laboratory
Aerodynamics Research Laboratory
Optical Physics & Engineering Laboratory
Engineering Hall
Daniels Hall
Central Illinois Mosque & Islamic Center
St Patrick's Church


W University Ave
W Springfield Ave
W Main St
W Clark St
W Stoughton St
E Clark St
E Healey St
W Western Ave
S Wright St
N Mathews Ave
N Goodwin Ave
N Harvey St
N Gregory St
N Lincoln Ave
S Lincoln Ave
S 6th St
N Romine St
Goodwin Ave
Mathews Ave
Beckman Institute for Advanced Science and Technology
Engineering Parking Garage
Chiller Plant
Electrical and Computer Engineering Bldg
Coordinated Science Laboratory
Oval Allée
Newmark Civil Engineering Laboratory
Thomas M. Siebel Center for Computer Science
Richmond Studio
University High School Gymnasium
Public Safety Bldg
Digital Computer Laboratory
University Place Christian Church
Shlens Edmund N CPA
Niro's Gyros
Mulberry Alley
Dos Reales
Hot Wok Express
Computing Applications Bldg
Grainger Engineering Library Information Center
Oil Chemistry Building
Victor E. Shelford Vivarium
Talbot Laboratory
Mechanical Engineering Laboratory
Transportation Bldg
Aeronautical Laboratory A
Ceramics Kiln House
Ceramics Bldg
Advanced Computation Bldg
Environmental Health and Safety
Aerodynamics Research Laboratory
Plant Services Bldg North
Engineering Senior Design Studio & Annex Plant Services Storage Bldg
Optical Physics & Engineering Laboratory
Mechanical Engineering Bldg
Nuclear Radiations Laboratory
Everitt Laboratory
Engineering Hall
Materials Science & Engineering
Atmospheric Sciences Annex2
Daniels Hall
Astronomy Bldg
Sushi Ichiban
Follett
User 4

MySQL

```
create table Points (
name VARCHAR(20) PRIMARY KEY,
location Point NOT NULL,
description VARCHAR(200),
SPATIAL INDEX(location)
);
```

MySQL

```
SELECT name, AsText(location)
FROM Points
WHERE Intersects(location, GeomFromText(@bbox) );
+------------+----------------------+
| name       | AsText(location)     |
+------------+----------------------+
| name_243.0 | POINT(6.1908 3.1375) |
| name_429.0 | POINT(6.5194 0.4023) |
| name_533.0 | POINT(7.7479 3.0894) |
| name_808.0 | POINT(4.6818 7.618)  |
+------------+----------------------+
```

Multidimensional Indexing

- The book covers a lot of structures for this:
  - Grid Files
  - Partitioned Hash Tables
  - Multiple-Key Indexes
  - Quad trees
- We'll only cover two

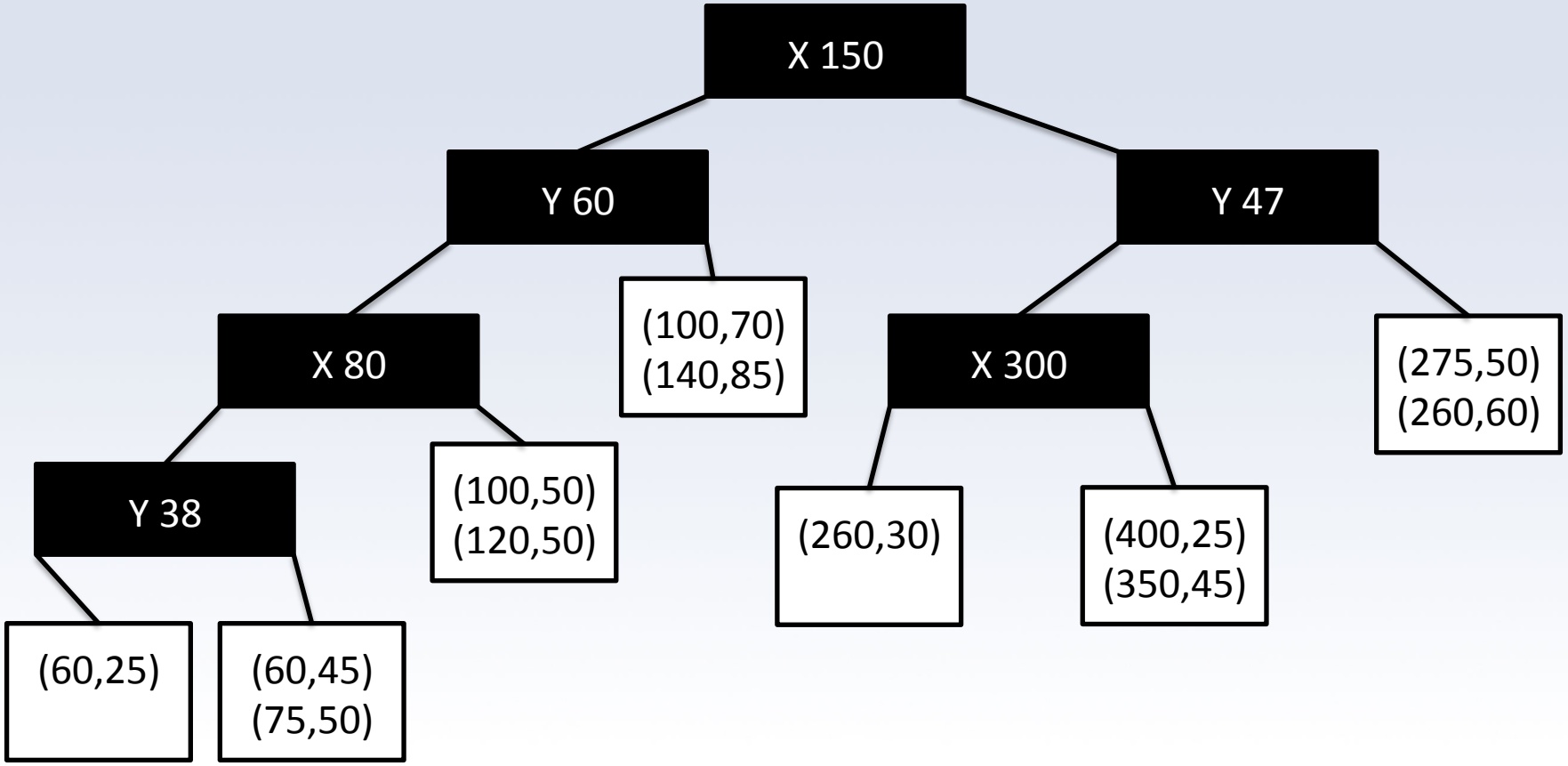
k-d trees

- Useful for range and nearest neighbor queries
- Similar to binary search tree
- Internal nodes alternate dimensions
  - Root splits x
  - Child splits y
  - Grandchild splits x
  - Great grandchild splits y…

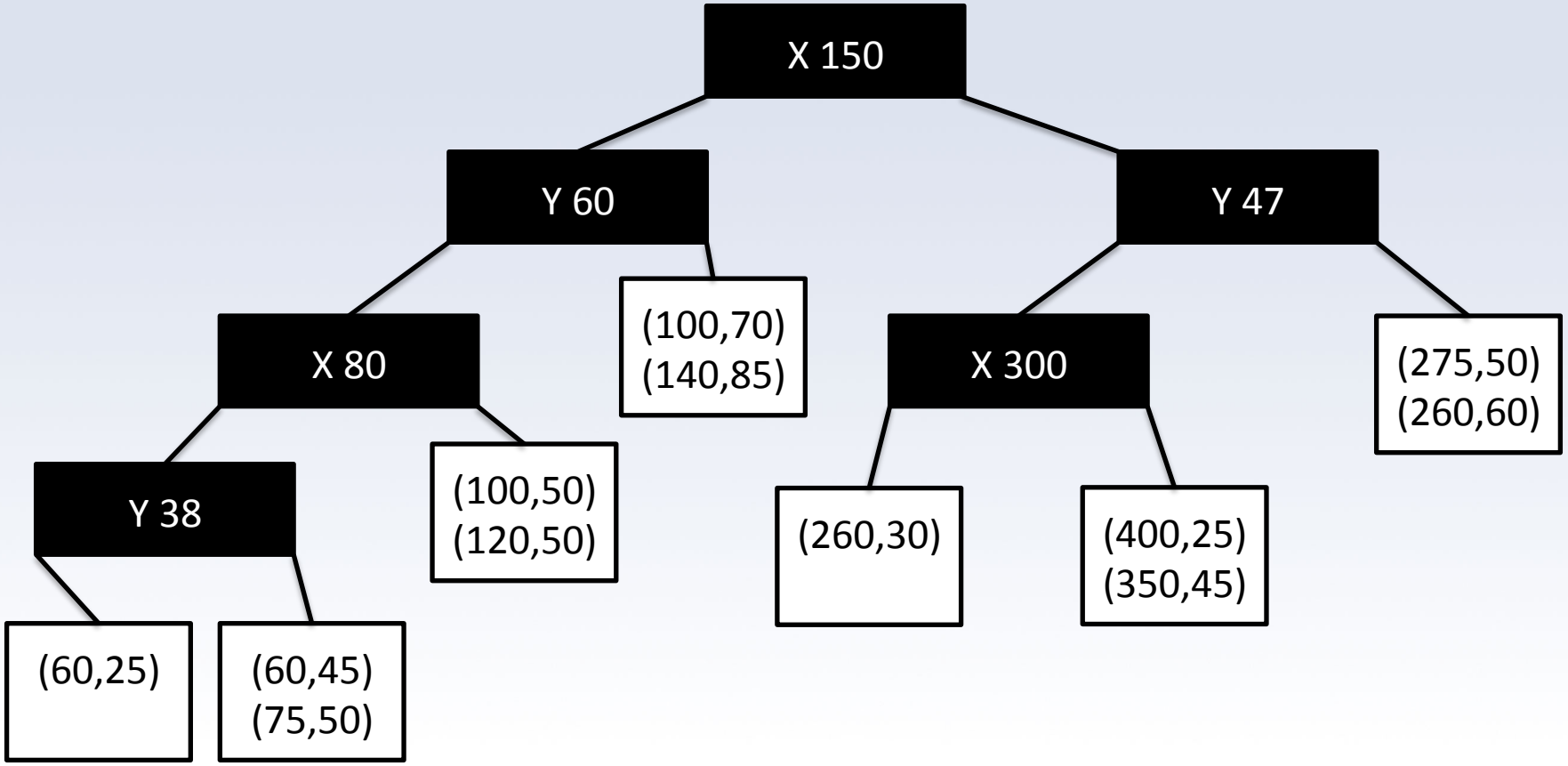
k-d trees

- Each internal node has a dimension D and value V associated with it
  - If R.D < V, point in left subtree
  - If R.D $\geq$ V, point in right subtree
- Leaves contain blocks
  - not typical for memory-based k-d trees

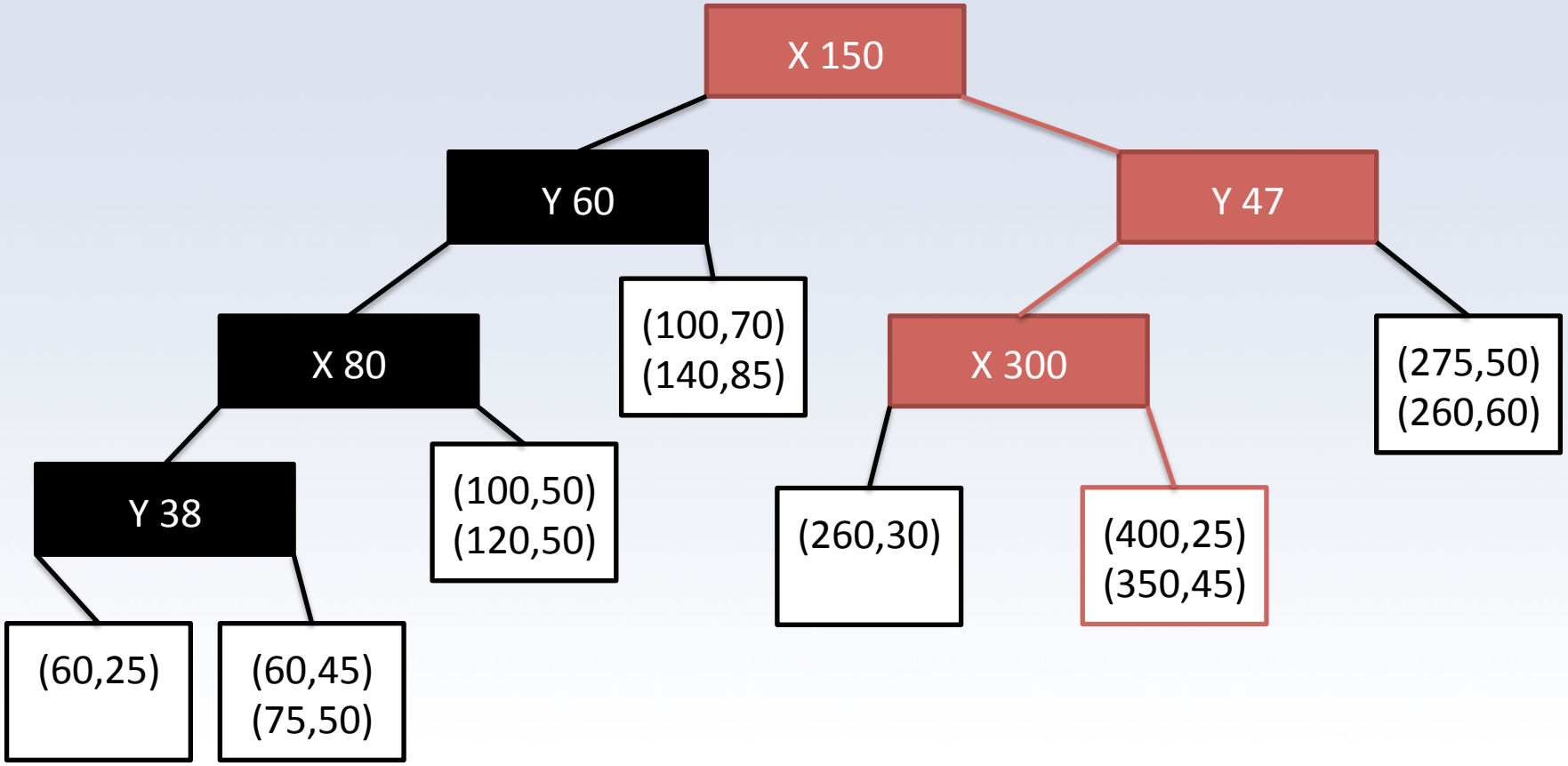
Example


X 150
Y 60
Y 47
X 80
(100,70)
(140,85)
X 300
(275,50)
(260,60)
Y 38
(100,50)
(120,50)
(260,30)
(400,25)
(350,45)
(60,25)
(60,45)
(75,50)

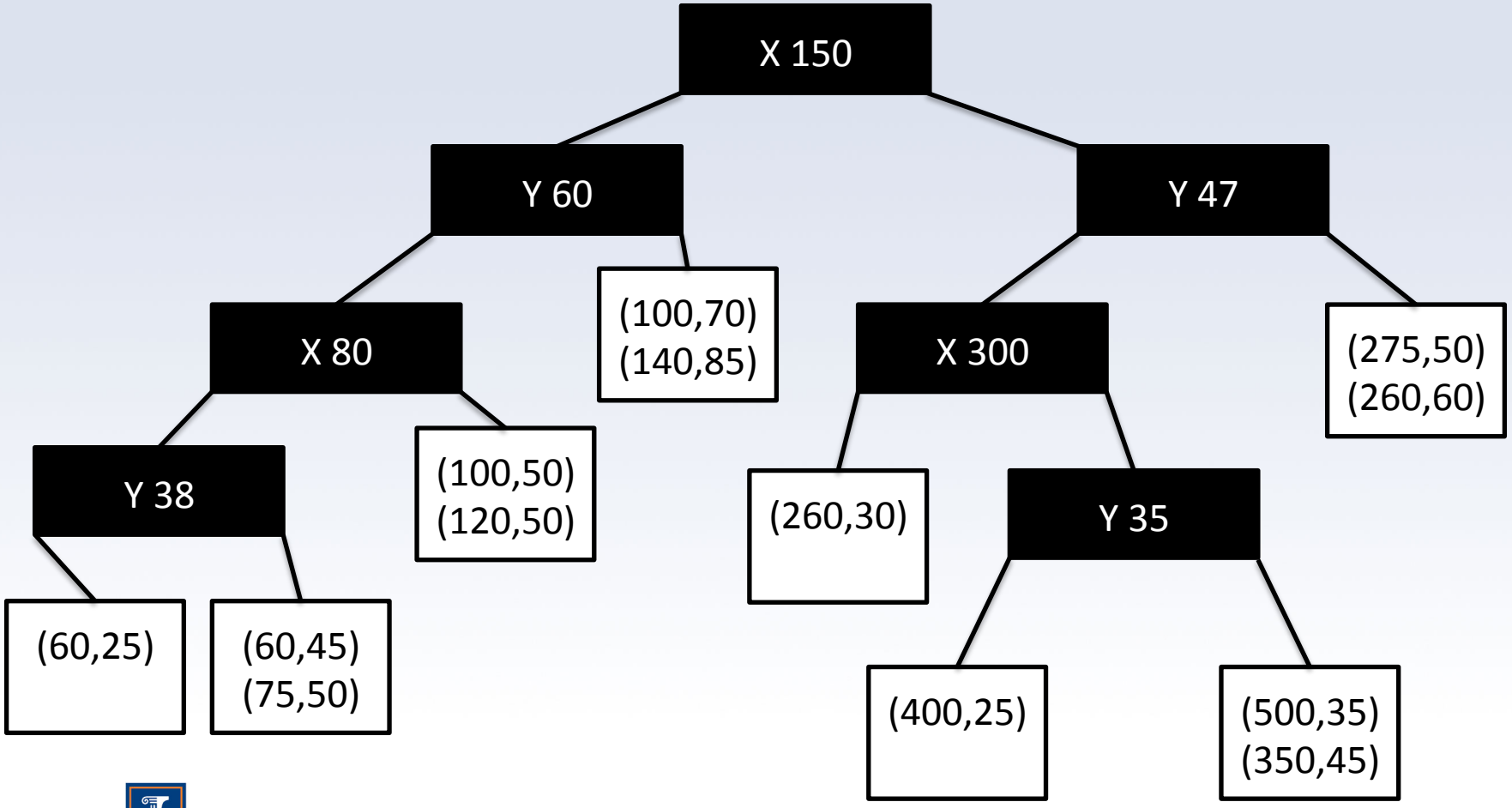
Insert (500,35)



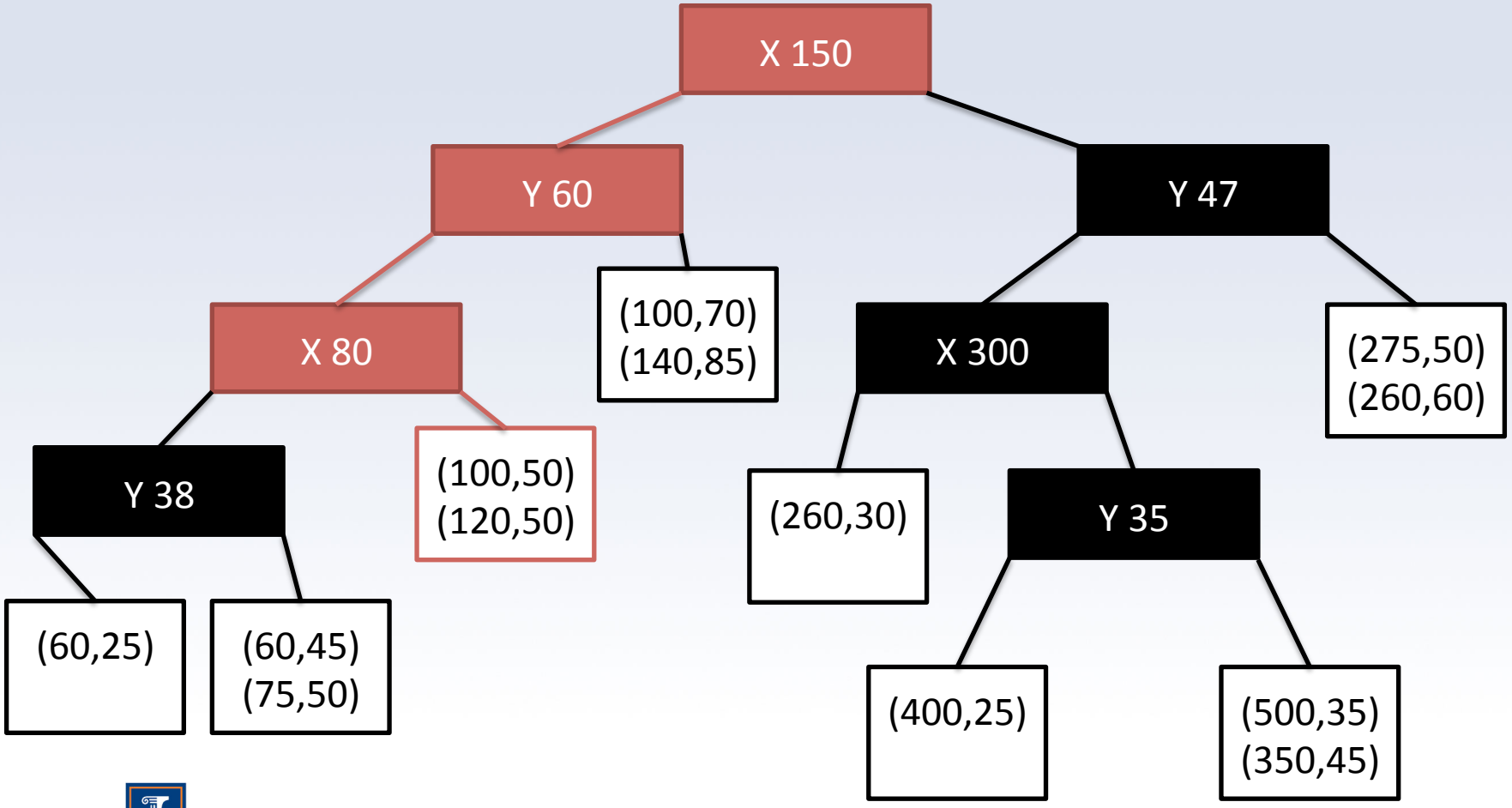
Insert (500,35)



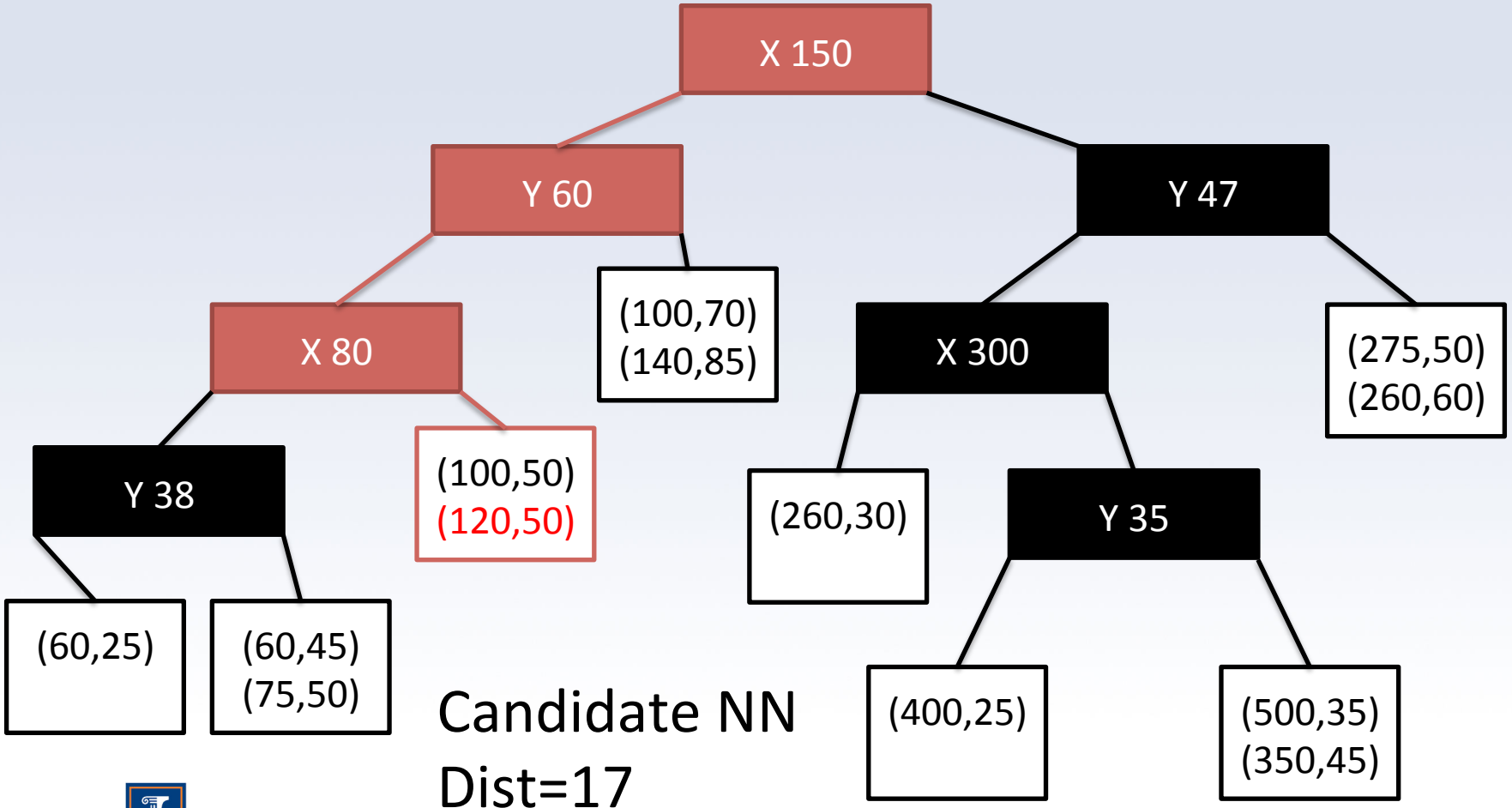
Insert (500,35)



Find NN (120,33)



Find NN (120,33)



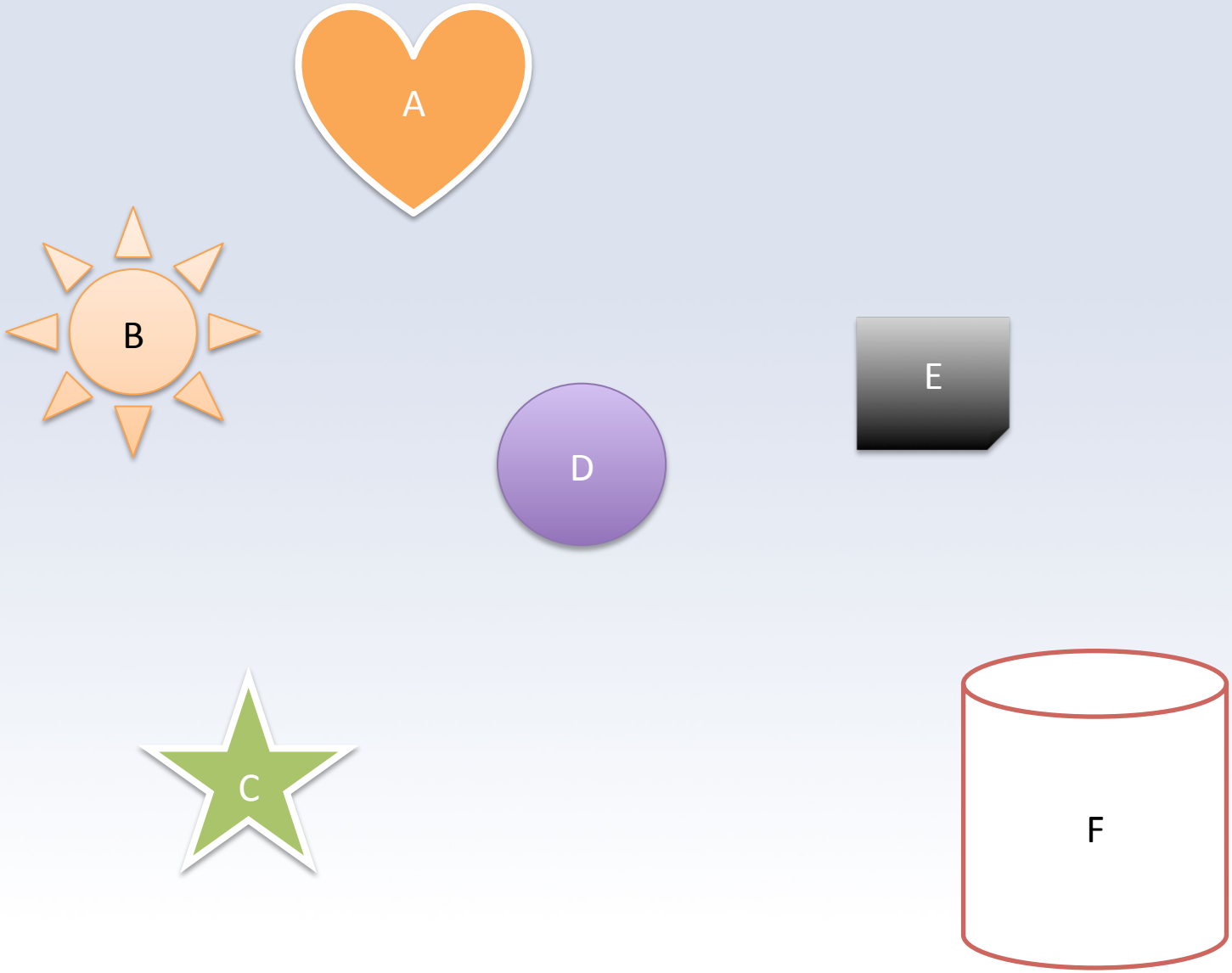
Candidate NN
Dist=17

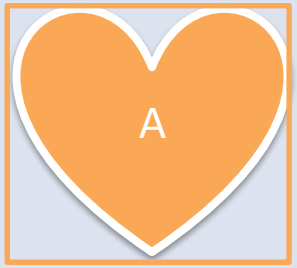
Find NN (120,33)

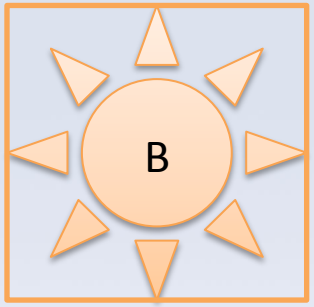
- After finding a candidate, we have to "unwind" back up the search tree to check if any points are closer
- Requires more data at internal nodes
- Not going to cover this

R-trees

- Groups nearby objects by *minimum bounding rectangles*
  - Smallest rectangle that contains them
- Groups the groups with more bounding rectangles
- Creates a tree of bounding regions


A
B
D
E
C
F

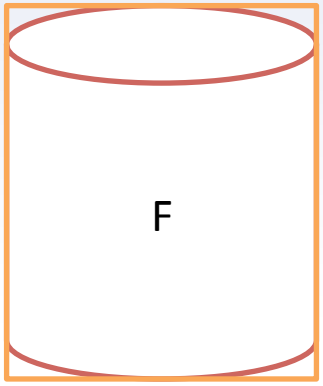

A

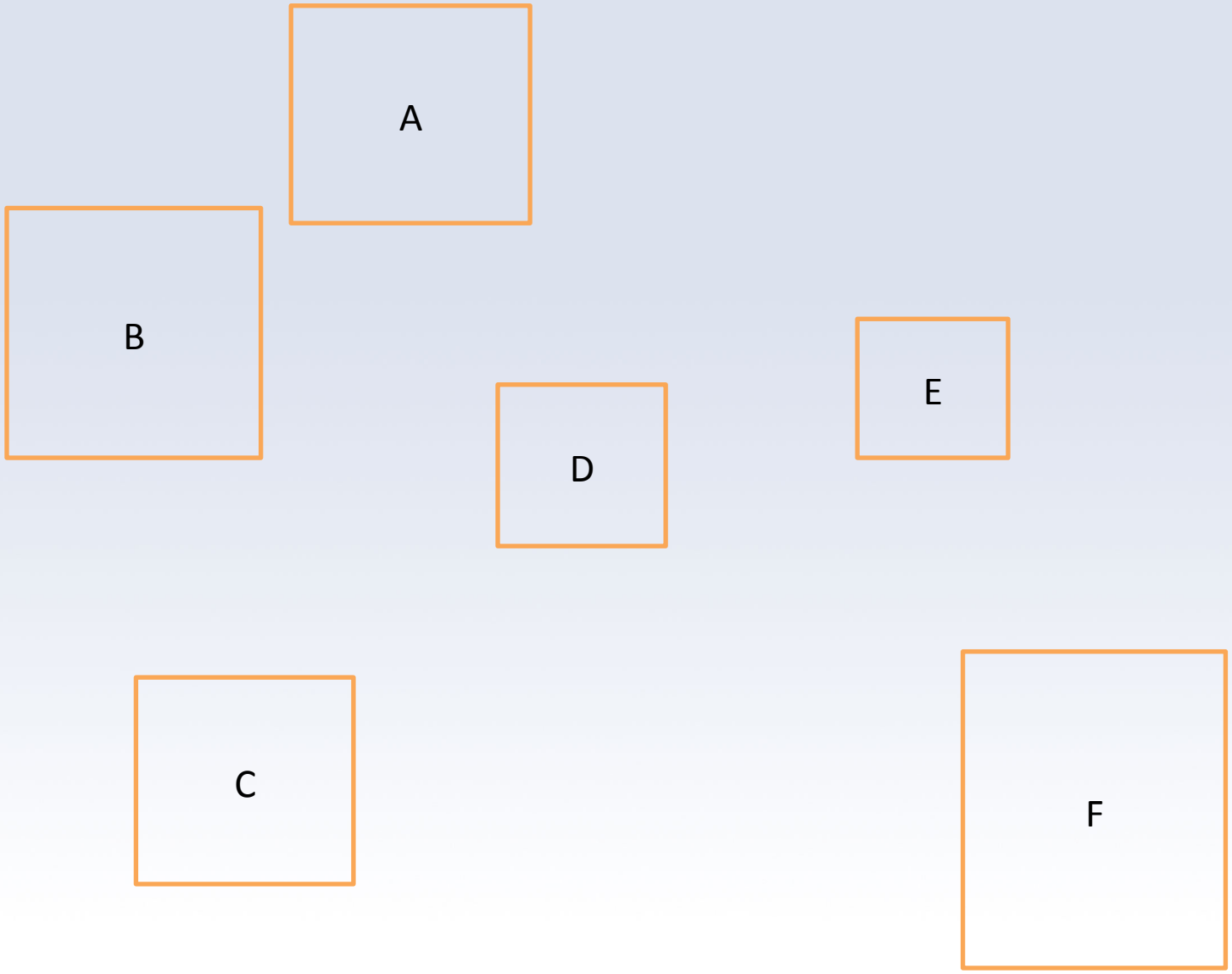

B

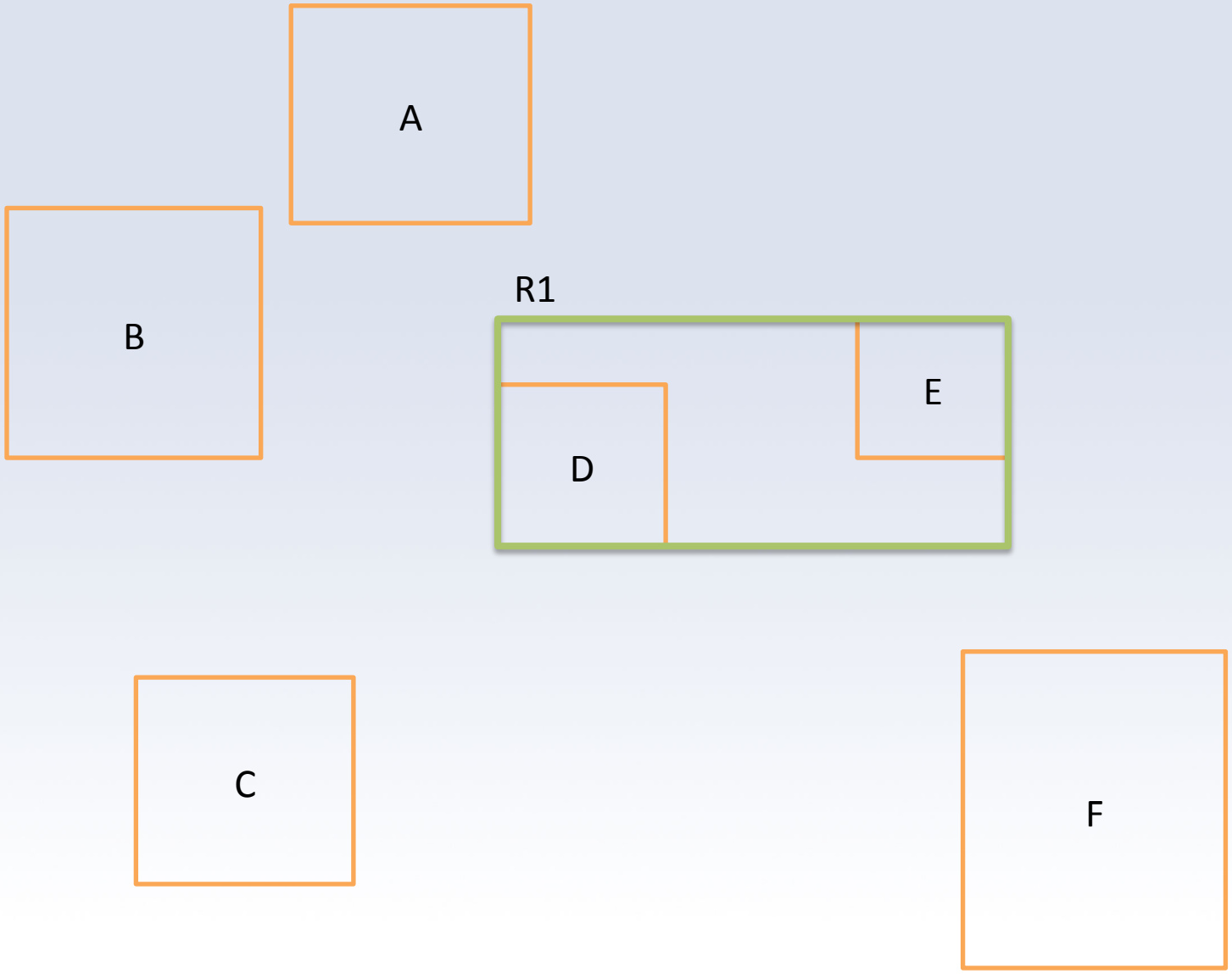

D

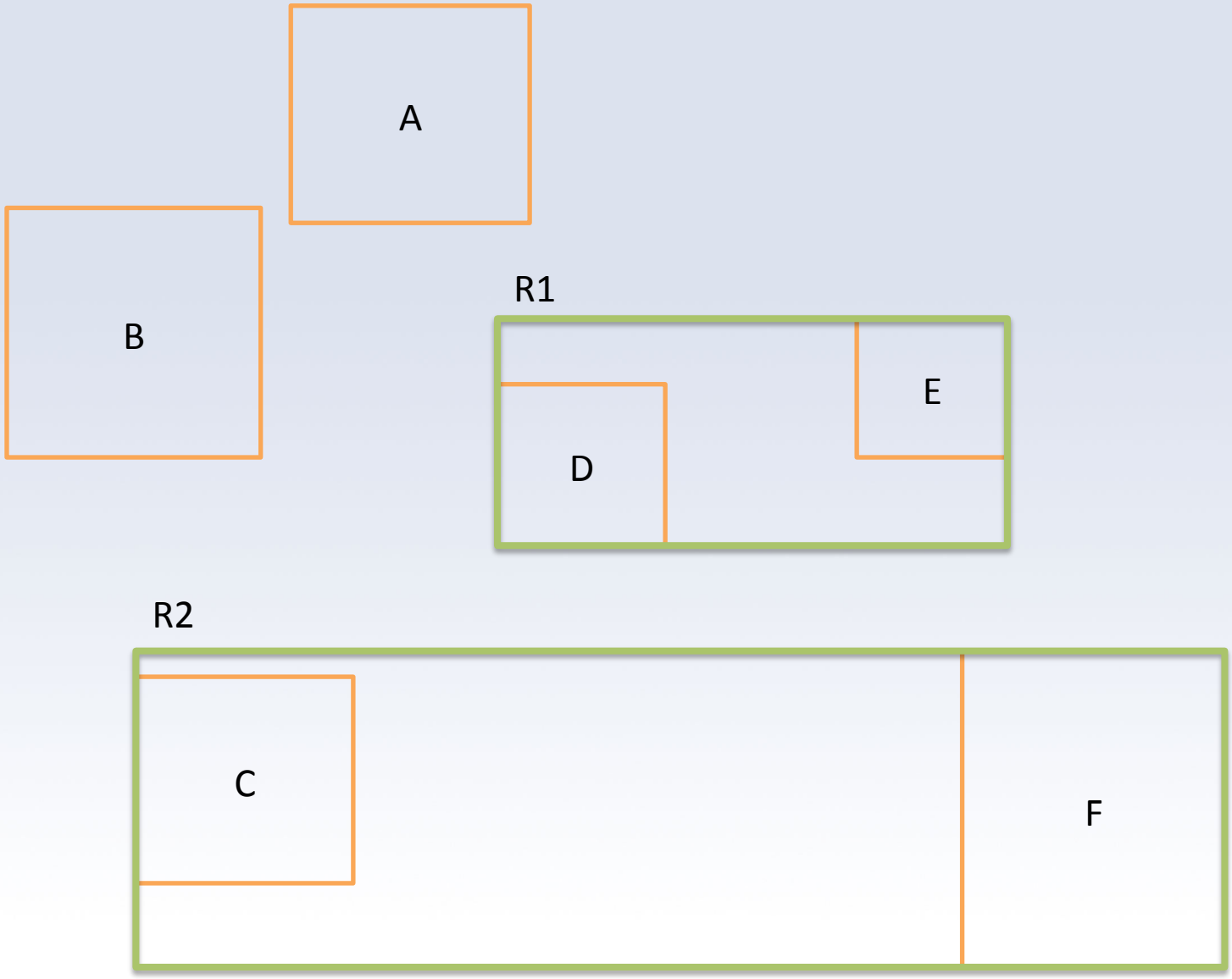

E

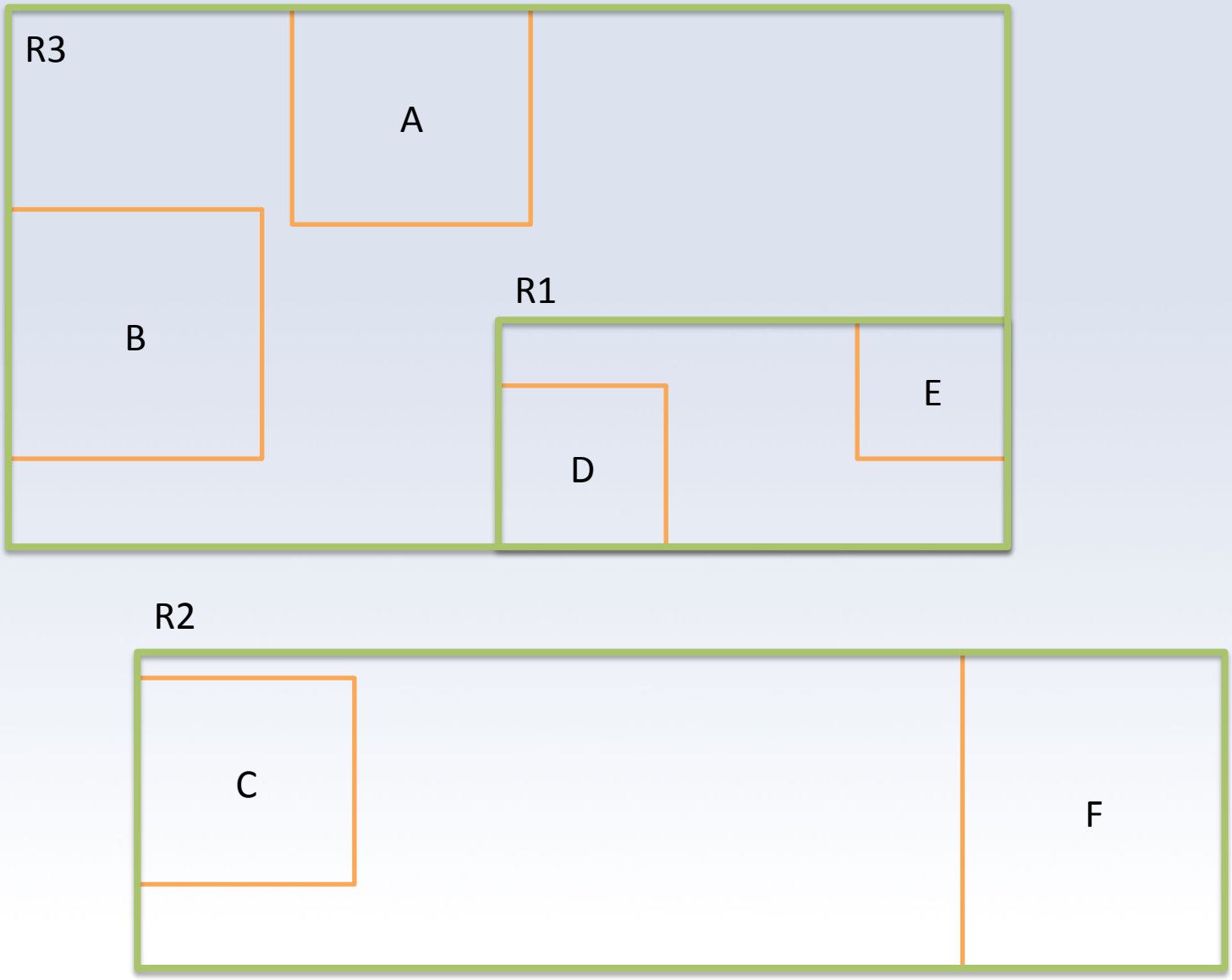

C

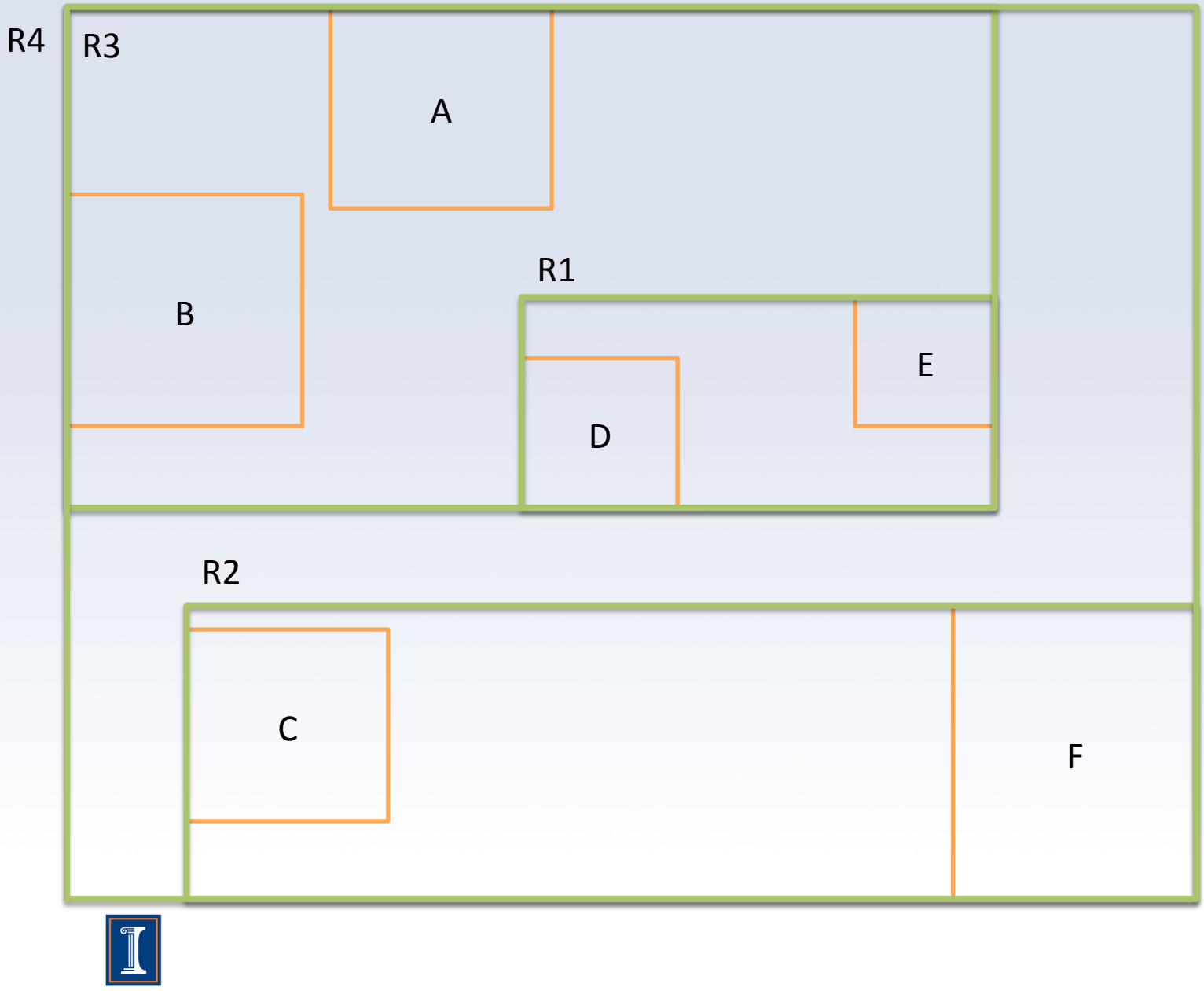

F


A
B
D
E
C
F

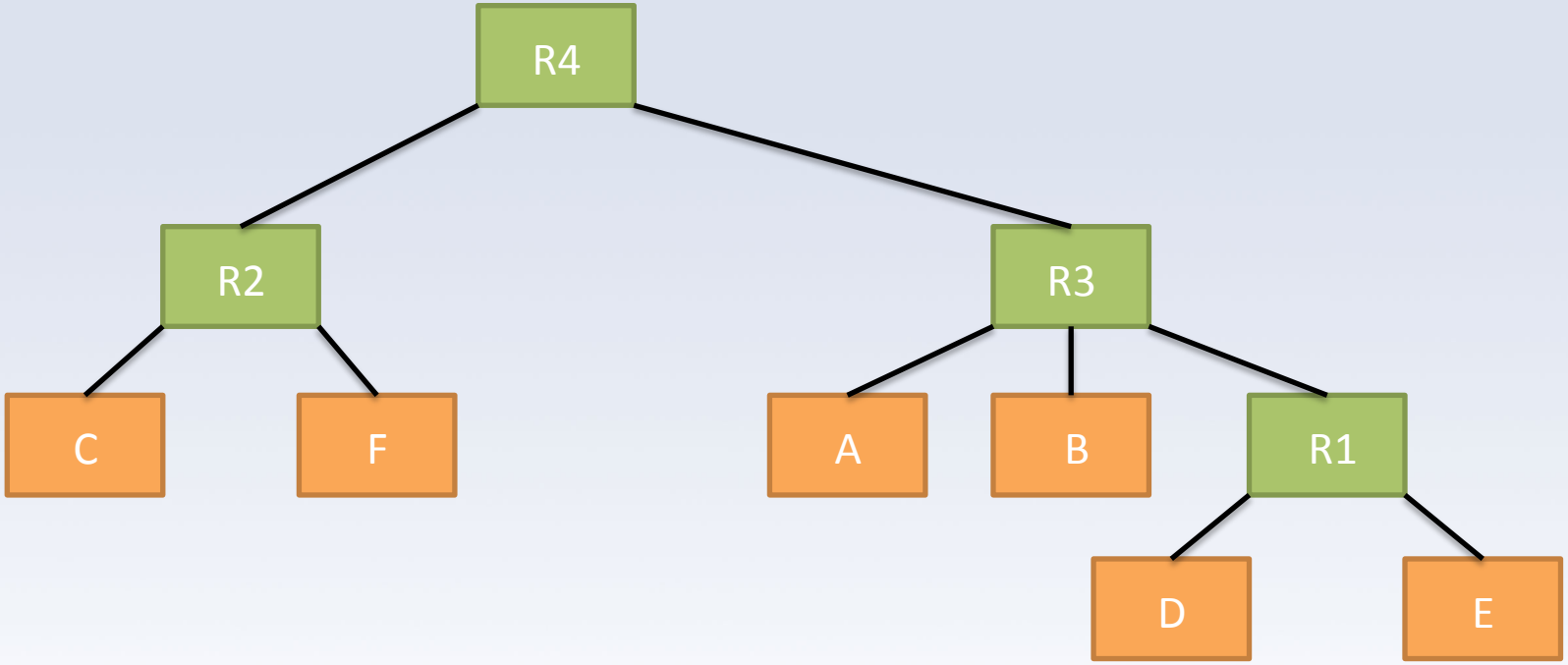

A
B
R1
E
D
C
F


A
B
R1
E
D
R2
C
F


R3
A
B
R1
E
D
R2
C
F

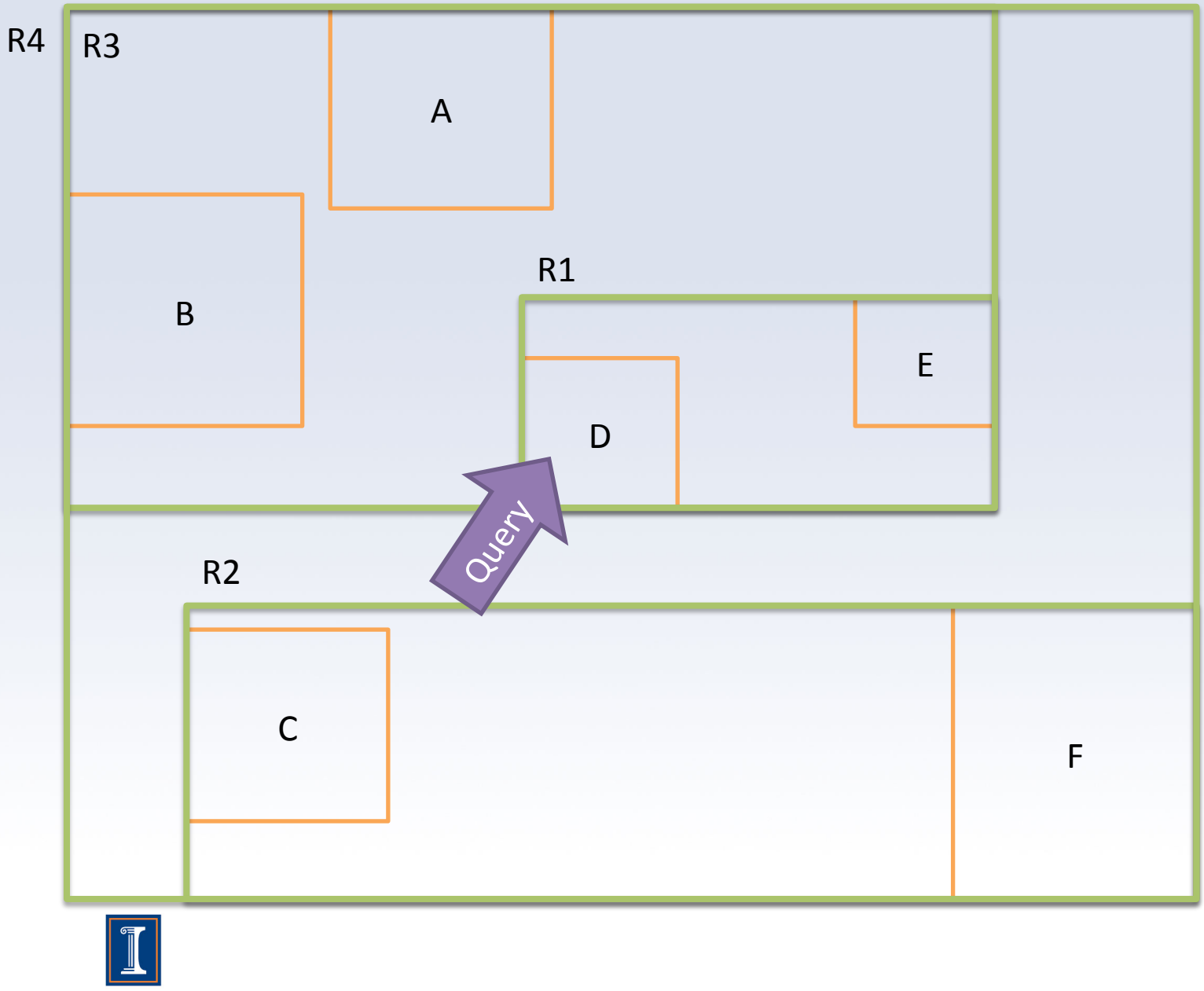

R4
R3
A
B
R1
E
D
R2
C
F

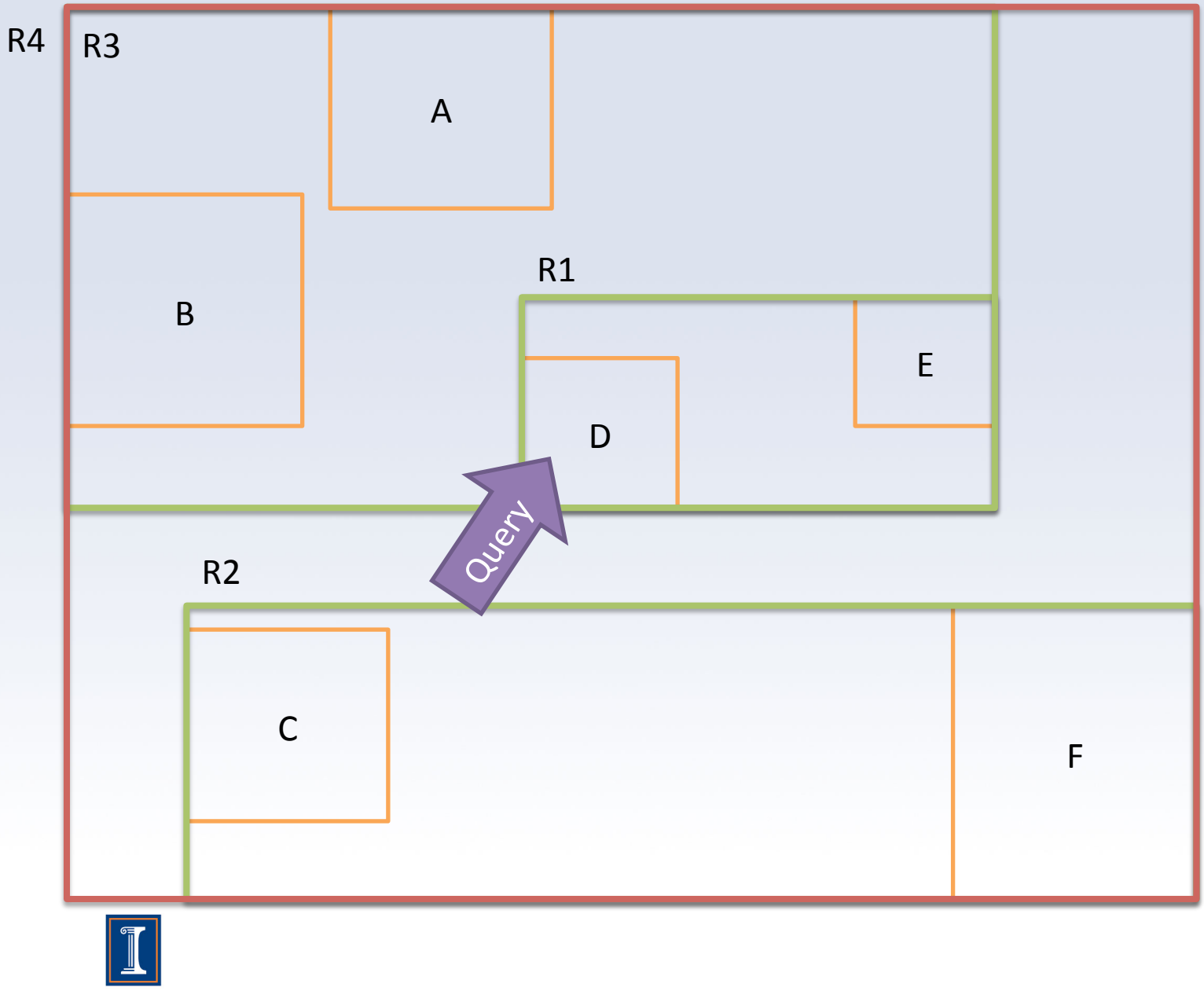
R-tree

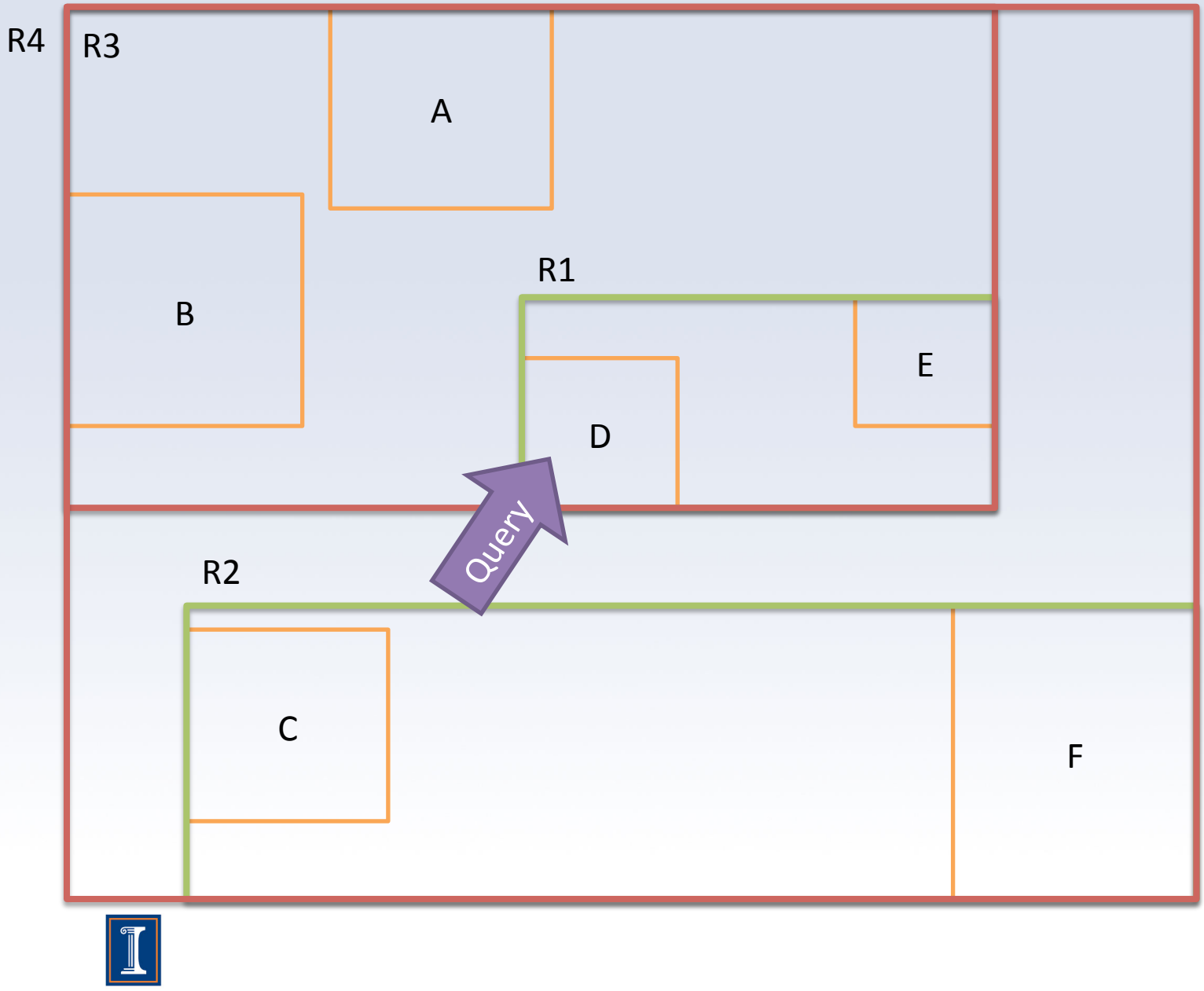


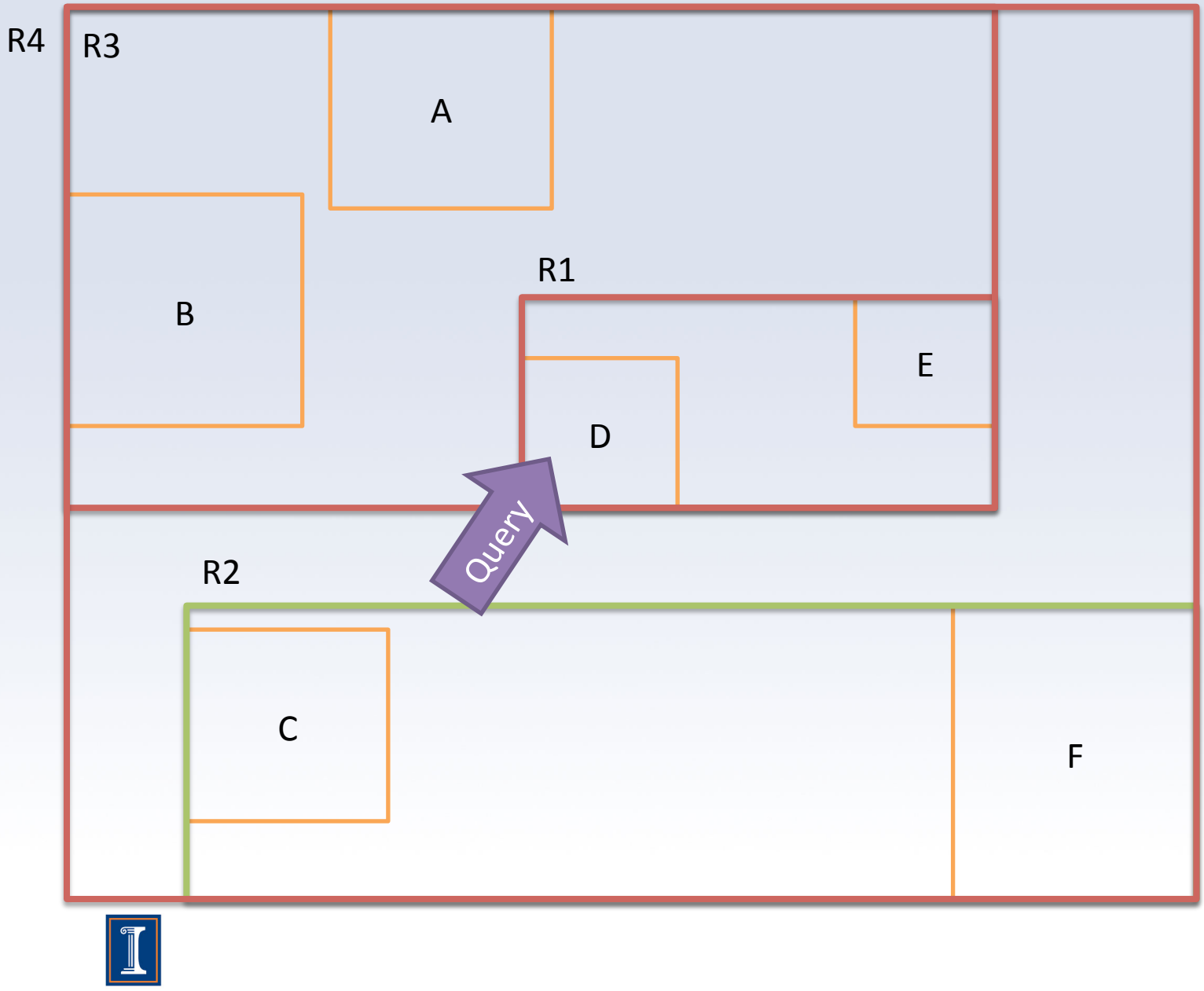
Location query

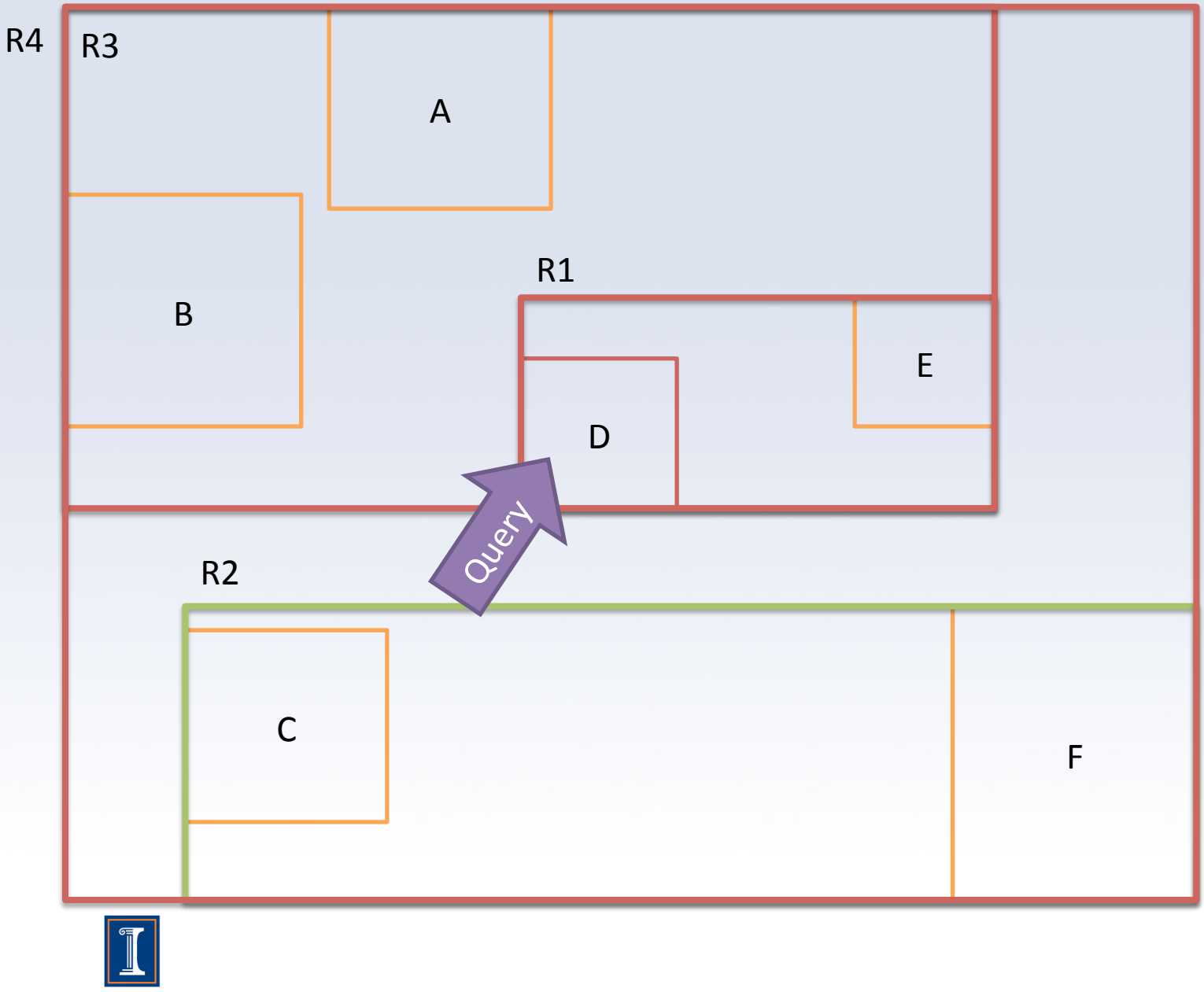
- Recursively check if point is in each rectangle


R4
R3
A
B
R1
E
D
Query
R2
C
F

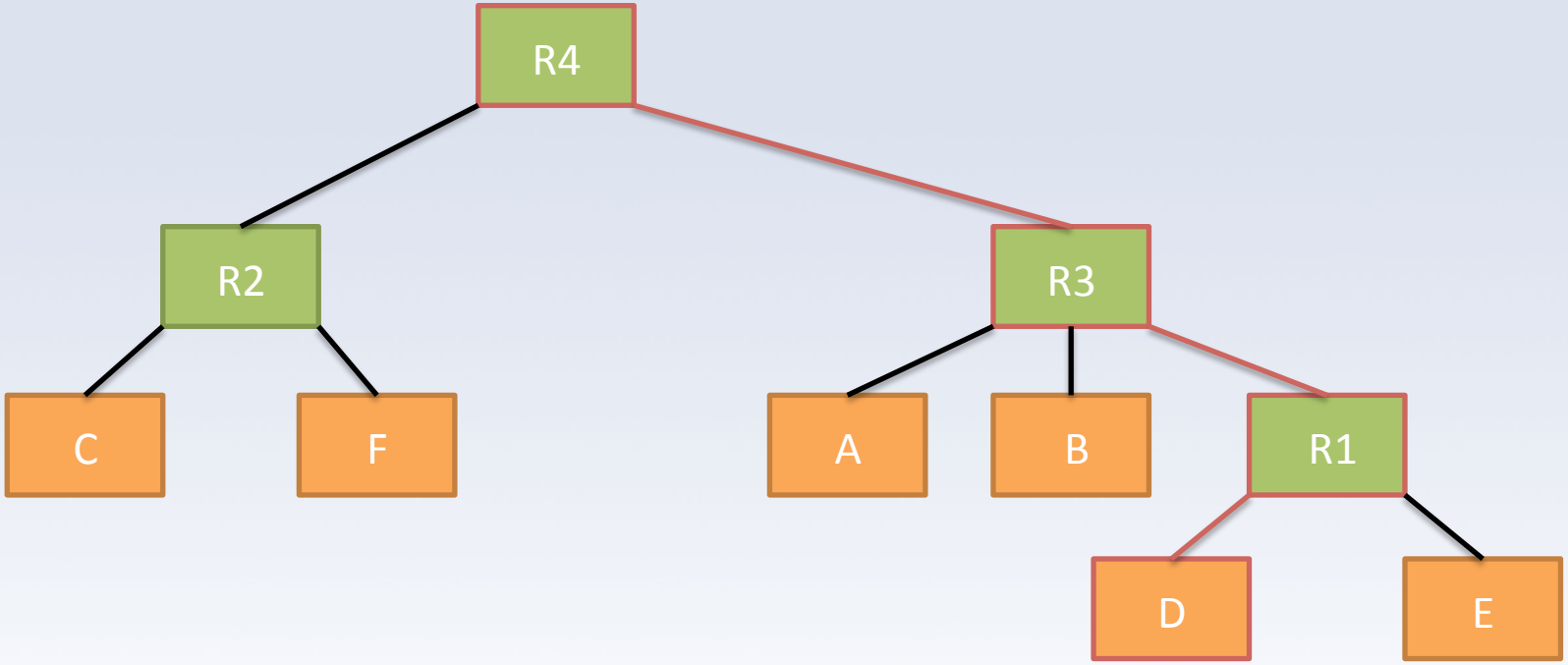

R4
R3
A
B
R1
E
D
Query
R2
C
F


R4
R3
A
B
R1
E
D
Query
R2
C
F


R4
R3
A
B
R1
E
D
Query
R2
C
F


R4
R3
A
B
R1
E
D
Query
R2
C
F

Location Query



R-trees

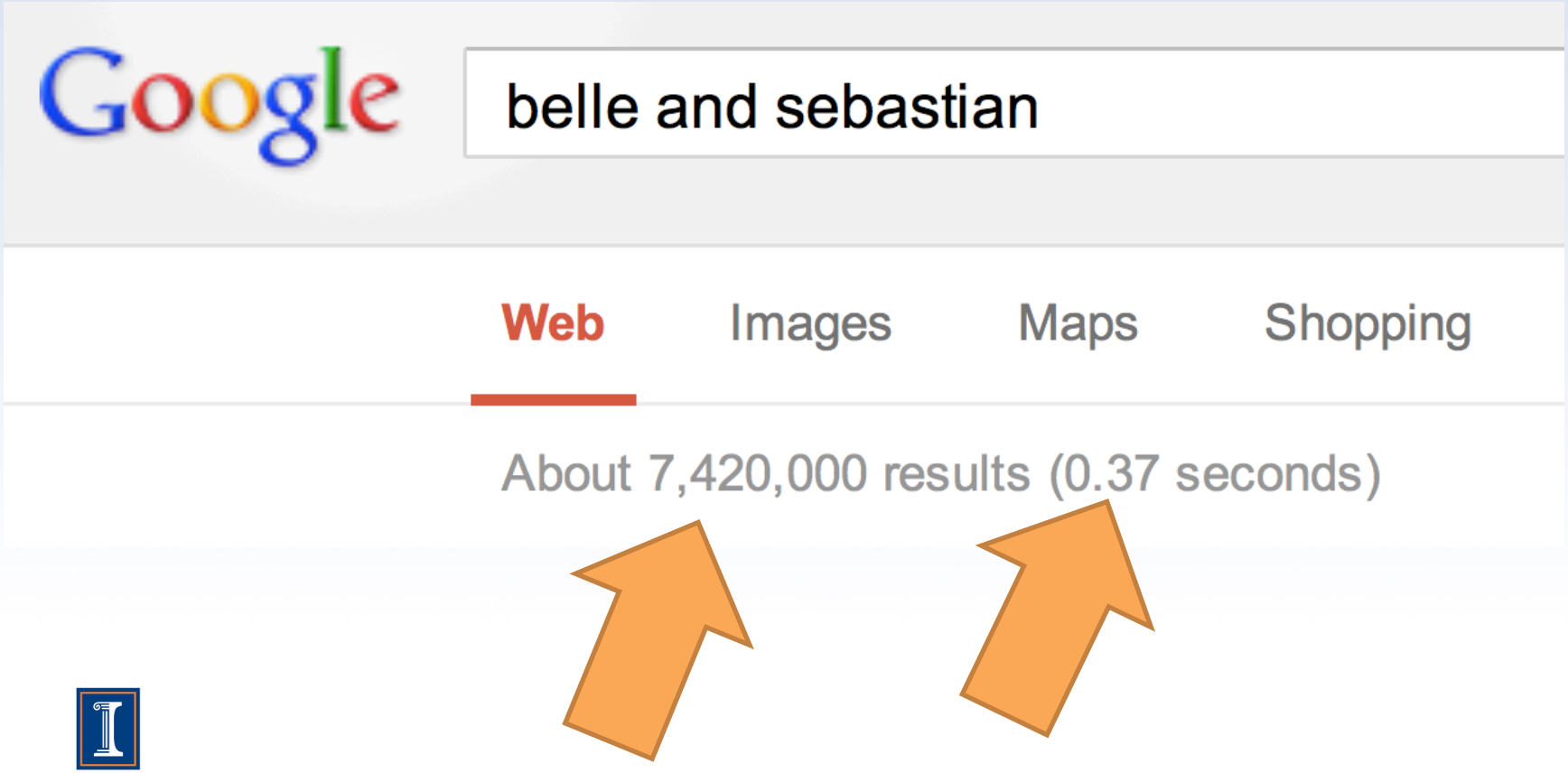
- Actually implemented in MySQL

```
CREATE SPATIAL INDEX spatial_ind
ON Points(location);
```

Document Indexing

- Given a collection of text documents, how can we quickly find all the documents containing a word

Document search



Inverted Index

DOCUMENT 1
Don't you love her madly?

DOCUMENT 2
Say you love me.

DOCUMENT 3
Come as you are.

DOCUMENT 4
All you need is love.

don't	{1}
you	{1,2,3,4}
love	{1,2,4}
her	{1}
madly	{1}
say	{1}

me	{2}
come	{3}
as	{3}
are	{3}
all	{4}
need	{4}

is	{4}

Query: you love

DOCUMENT 1
Don't you love her madly?

DOCUMENT 2
Say you love me.

DOCUMENT 3
Come as you are.

DOCUMENT 4
All you need is love.

don't	{1}
you	{1,2,3,4}
love	{1,2,4}
her	{1}
madly	{1}
say	{1}

me	{2}
come	{3}
as	{3}
are	{3}
all	{4}
need	{4}

is	{4}

Query: you love

DOCUMENT 1
Don't you love her madly?

DOCUMENT 2
Say you love me.

DOCUMENT 3
Come as you are.

DOCUMENT 4
All you need is love.

don't	{1}
you	{1,2,3,4}
love	{1,2,4}
her	{1}
madly	{1}
say	{1}

me	{2}
come	{3}
as	{3}
are	{3}
all	{4}
need	{4}

is	{4}

$$\{1,2,3,4\} \cap \{1,2,4\} = \{1,2,4\}$$

Query: you love

DOCUMENT 1
Don't you love her madly?

DOCUMENT 2
Say you love me.

DOCUMENT 3
Come as you are.

DOCUMENT 4
All you need is love.

don't	{1}
you	{1,2,3,4}
love	{1,2,4}
her	{1}
madly	{1}
say	{1}

me	{2}
come	{3}
as	{3}
are	{3}
all	{4}
need	{4}

is	{4}

$$\{1,2,3,4\} \cap \{1,2,4\} = \{1,2,4\}$$

Inverted Indexing

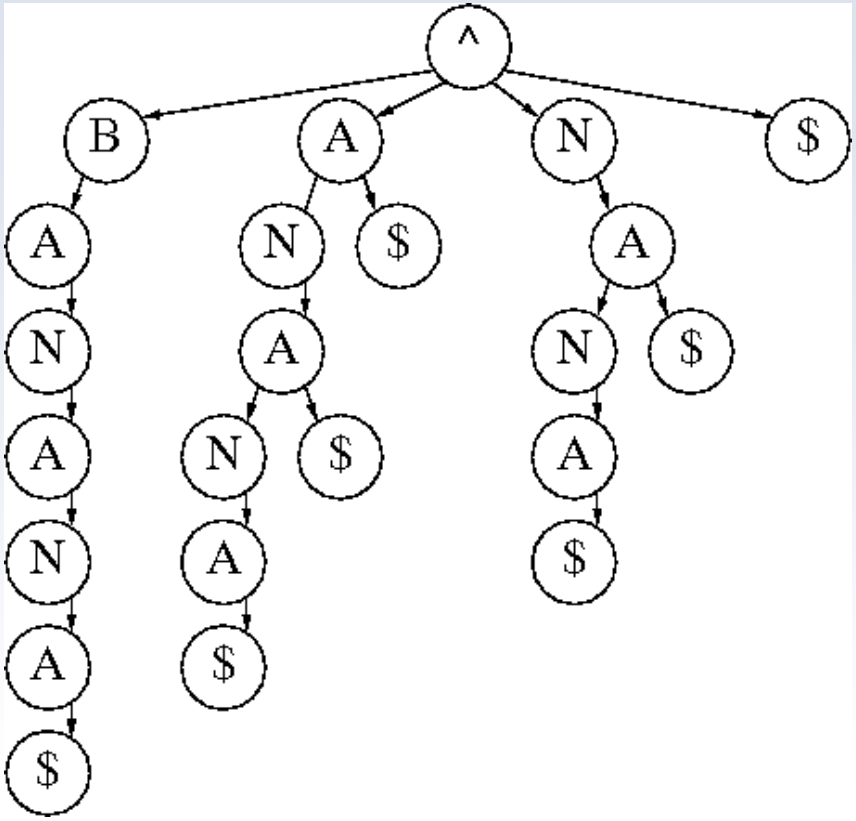
- Real inverted indexes also keep track of the location of the word in each page
- This allows exact phrases to be looked up

String indexing

- How can we index an extremely long string for search?

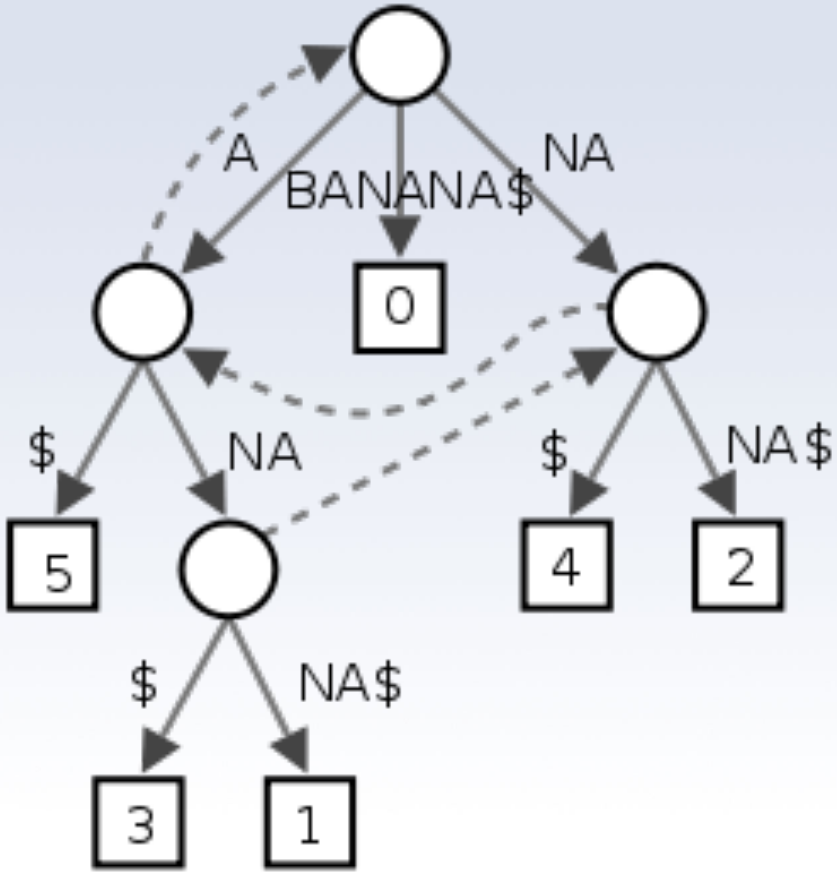
Suffix Trie

- A tree where each path from root to leaf corresponds to a suffix
- O(n) to search
- O(n²) memory



Suffix Tree

- Compressed version of the suffix trie
- Can be searched in O(n)
- Takes O(n) memory
- Can be constructed in O(n) ?!?!?!
- Ukkonen's algorithm



Suffix Array

- SA:
  - Given a text T, add a string terminator character '$'
  - Compute all n rotations (cyclic rotations)
  - Sort the rotations
  - Suffix array is the index of the rotation
- Can be computed in O(n) from a suffix tree

Suffix Array

Suffix	i
banana$	1
anana$	2
nana$	3
ana$	4
na$	5
a$	6
$	7

Suffix Array

Suffix	i
$	7
a$	6
ana$	4
anana$	2
banana$	1
na$	5
nana$	3

Burrows-Wheeler Transform

- BWT:
  - Given a text T, add a string terminator character '$'
  - Compute all n rotations (cyclic rotations)
  - Sort the rotations
  - BWT is last column
- Can be computed in O(n) from a suffix array

A	B	R	A	C	A	D	A	B	R	A

A	B	R	A	C	A	D	A	B	R	A	$

0	A	B	R	A	C	A	D	A	B	R	A	$

0	A	B	R	A	C	A	D	A	B	R	A	$
1	B	R	A	C	A	D	A	B	R	A	$	A

0	A	B	R	A	C	A	D	A	B	R	A	$
1	B	R	A	C	A	D	A	B	R	A	$	A
2	R	A	C	A	D	A	B	R	A	$	A	B

0	A	B	R	A	C	A	D	A	B	R	A	$
1	B	R	A	C	A	D	A	B	R	A	$	A
2	R	A	C	A	D	A	B	R	A	$	A	B
3	A	C	A	D	A	B	R	A	$	A	B	R

0	A	B	R	A	C	A	D	A	B	R	A	$
1	B	R	A	C	A	D	A	B	R	A	$	A
2	R	A	C	A	D	A	B	R	A	$	A	B
3	A	C	A	D	A	B	R	A	$	A	B	R
4	C	A	D	A	B	R	A	$	A	B	R	A

0	A	B	R	A	C	A	D	A	B	R	A	$
1	B	R	A	C	A	D	A	B	R	A	$	A
2	R	A	C	A	D	A	B	R	A	$	A	B
3	A	C	A	D	A	B	R	A	$	A	B	R
4	C	A	D	A	B	R	A	$	A	B	R	A
5	A	D	A	B	R	A	$	A	B	R	A	C
6	D	A	B	R	A	$	A	B	R	A	C	A
7	A	B	R	A	$	A	B	R	A	C	A	D
8	B	R	A	$	A	B	R	A	C	A	D	A
9	R	A	$	A	B	R	A	C	A	D	A	B
10	A	$	A	B	R	A	C	A	D	A	B	R
11	$	A	B	R	A	C	A	D	A	B	R	A

11	$	A	B	R	A	C	A	D	A	B	R	A
10	A	$	A	B	R	A	C	A	D	A	B	R
7	A	B	R	A	$	A	B	R	A	C	A	D
0	A	B	R	A	C	A	D	A	B	R	A	$
3	A	C	A	D	A	B	R	A	$	A	B	R
5	A	D	A	B	R	A	$	A	B	R	A	C
8	B	R	A	$	A	B	R	A	C	A	D	A
1	B	R	A	C	A	D	A	B	R	A	$	A
4	C	A	D	A	B	R	A	$	A	B	R	A
6	D	A	B	R	A	$	A	B	R	A	C	A
9	R	A	$	A	B	R	A	C	A	D	A	B
2	R	A	C	A	D	A	B	R	A	$	A	B

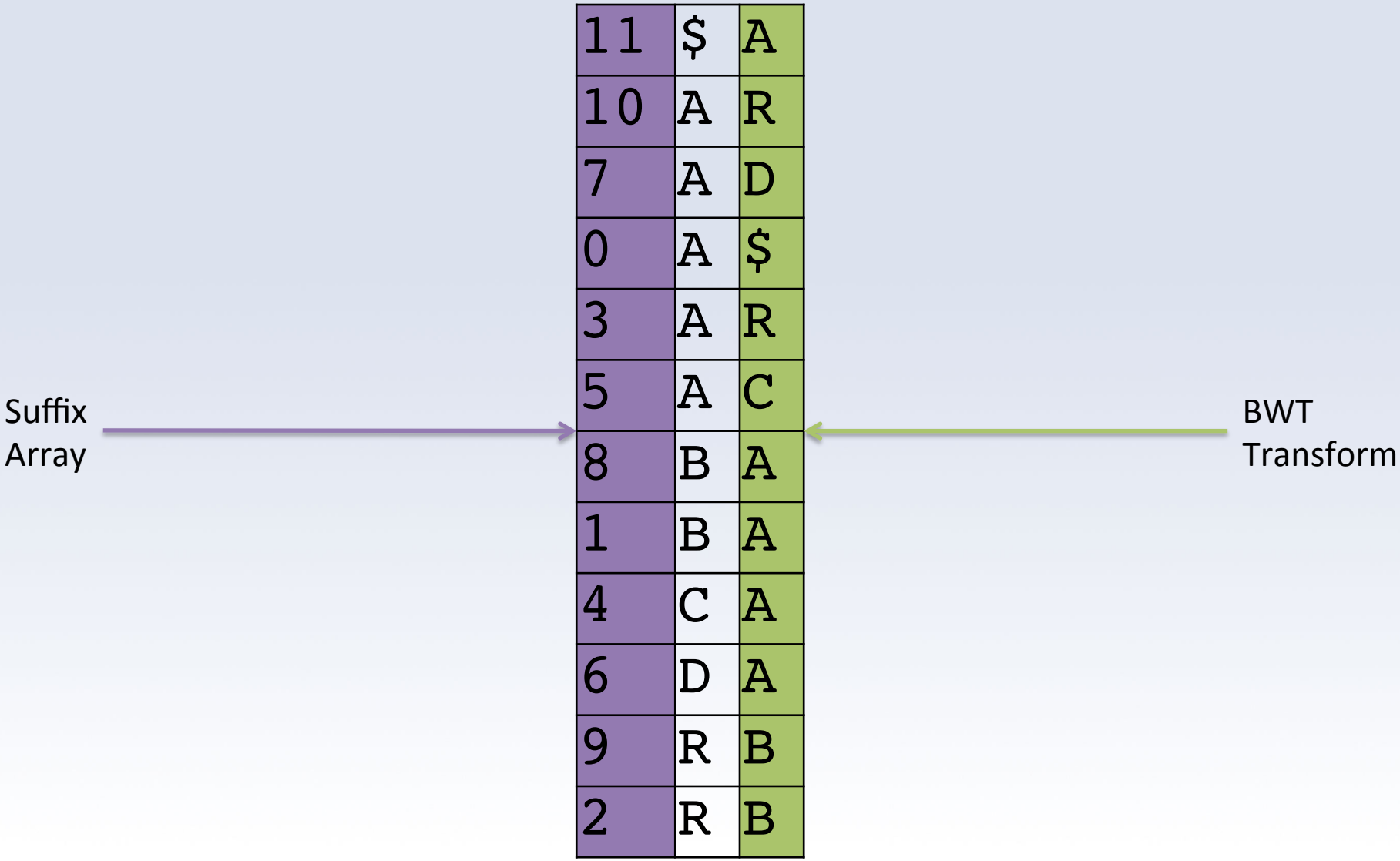
Suffix Array

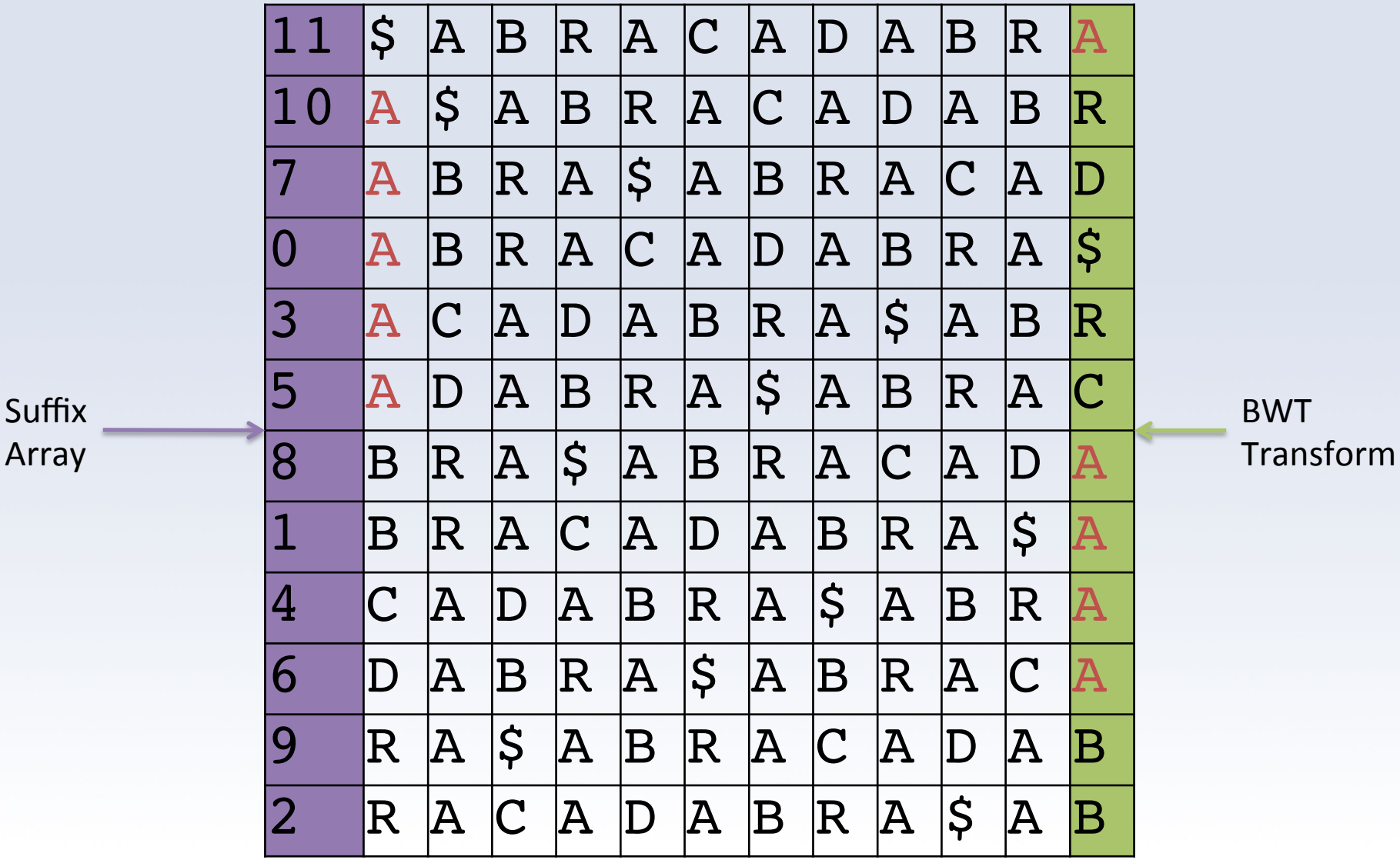
BWT Transform

illinois.edu

ABRA

11	$	A	B	R	A	C	A	D	A	B	R	A
10	A	$	A	B	R	A	C	A	D	A	B	R
7	A	B	R	A	$	A	B	R	A	C	A	D
0	A	B	R	A	C	A	D	A	B	R	A	$
3	A	C	A	D	A	B	R	A	$	A	B	R
5	A	D	A	B	R	A	$	A	B	R	A	C
8	B	R	A	$	A	B	R	A	C	A	D	A
1	B	R	A	C	A	D	A	B	R	A	$	A
4	C	A	D	A	B	R	A	$	A	B	R	A
6	D	A	B	R	A	$	A	B	R	A	C	A
9	R	A	$	A	B	R	A	C	A	D	A	B
2	R	A	C	A	D	A	B	R	A	$	A	B


Suffix
Array
11 $ A
10 A R
7 A D
0 A $
3 A R
5 A C
8 B A
1 B A
4 C A
6 D A
9 R B
2 R B
BWT
Transform


Suffix Array
BWT Transform
11 $ A B R A C A D A B R A
10 A $ A B R A C A D A B R
7 A B R A $ A B R A C A D
0 A B R A C A D A B R A $
3 A C A D A B R A $ A B R
5 A D A B R A $ A B R A C
8 B R A $ A B R A C A D A
1 B R A C A D A B R A $ A
4 C A D A B R A $ A B R A
6 D A B R A $ A B R A C A
9 R A $ A B R A C A D A B
2 R A C A D A B R A $ A B

	$	A	B	R	A	C	A	D	A	B	R	A	1
1	A	$	A	B	R	A	C	A	D	A	B	R	
2	A	B	R	A	$	A	B	R	A	C	A	D	
3	A	B	R	A	C	A	D	A	B	R	A	$	
4	A	C	A	D	A	B	R	A	$	A	B	R	
5	A	D	A	B	R	A	$	A	B	R	A	C	
	B	R	A	$	A	B	R	A	C	A	D	A	2
	B	R	A	C	A	D	A	B	R	A	$	A	3
	C	A	D	A	B	R	A	$	A	B	R	A	4
	D	A	B	R	A	$	A	B	R	A	C	A	5
	R	A	$	A	B	R	A	C	A	D	A	B	
	R	A	C	A	D	A	B	R	A	$	A	B	

Search Interval: (0,11)

Search String: ABRA

11	$	A
10	A	R
7	A	D
0	A	$
3	A	R
5	A	C
8	B	A
1	B	A
4	C	A
6	D	A
9	R	B
2	R	B

Search Interval: (0,11)

Search String: ABRA

11	$	A
10	A	R
7	A	D
0	A	$
3	A	R
5	A	C
8	B	A
1	B	A
4	C	A
6	D	A
9	R	B
2	R	B

Search Interval: (0,11)

Search String: ABRA

11	$	A
10	A	R
7	A	D
0	A	$
3	A	R
5	A	C
8	B	A
1	B	A
4	C	A
6	D	A
9	R	B
2	R	B

Search Interval: (0,11)

Search String: ABRA

11	$	A	
10	A	R	1st R
7	A	D	
0	A	$	
3	A	R	2nd R
5	A	C	
8	B	A	
1	B	A	
4	C	A	
6	D	A	
9	R	B	
2	R	B	

Search Interval: (10,11)

Search String: ABRA

11	$	A	
10	A	R	1st R
7	A	D	
0	A	$	
3	A	R	2nd R
5	A	C	
8	B	A	
1	B	A	
4	C	A	
6	D	A	
9	R	B	
2	R	B	

Search Interval: (10,11)

Search String: ABRA

11	$	A	
10	A	R	
7	A	D	
0	A	$	
3	A	R	
5	A	C	
8	B	A	
1	B	A	
4	C	A	
6	D	A	
9	R	B	1st B
2	R	B	2nd B

Search Interval: (6,7)

Search String: ABRA

11	$	A
10	A	R
7	A	D
0	A	$
3	A	R
5	A	C
8	B	A
1	B	A
4	C	A
6	D	A
9	R	B
2	R	B

Search Interval: (6,7)

Search String: ABRA

11	$	A	
10	A	R	
7	A	D	
0	A	$	
3	A	R	
5	A	C	
8	B	A	2nd A
1	B	A	3rd A
4	C	A	
6	D	A	
9	R	B	
2	R	B	

Search Interval: (2,3)

Search String: ABRA

11	$	A
10	A	R
7	A	D
0	A	$
3	A	R
5	A	C
8	B	A
1	B	A
4	C	A
6	D	A
9	R	B
2	R	B

Two occurrences found
At indexes 0 and 7

01234567
ABRACADABRA

Improvements

- This algorithm can be improved:
  - We can make lookup tables for the “rank” of each symbol in BWT
  - We can make lookup tables for the location of each symbol in the middle array
  - Lookup is O(n) (with a very small constant)
  - Memory footprint is just 3 arrays for each character in the original string!

Improvements

- This algorithm can be expanded
  - Allow for mismatches

HAVE A GOOD BREAK!

- GO HAVE FUN!!!!!!