Querying

Introduction to Information Retrieval
INF 141
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org

Querying

Overview

- Boolean Retrieval
- Weighted Boolean Retrieval
- Zone Indices
- Term Frequency Metrics
- The full vector space model

Querying

From the bottom

Querying

From the bottom

- "Grep"
  - Querying without an index or a crawl
  - Whenever you want to find something you look through the entire document for it.
  - Example:
    - You have the collected works of Shakespeare on disk
    - You want to know which play contains the words
      - "Brutus AND Caesar"

Querying

Querying

- "Grep"
  - "Brutus AND Caesar" is the query.
  - This is a boolean query. Why?
  - What other operators could be used?
  - The grep solution:
    - Read all the files and all the text and output the intersection of the files

Querying

Querying

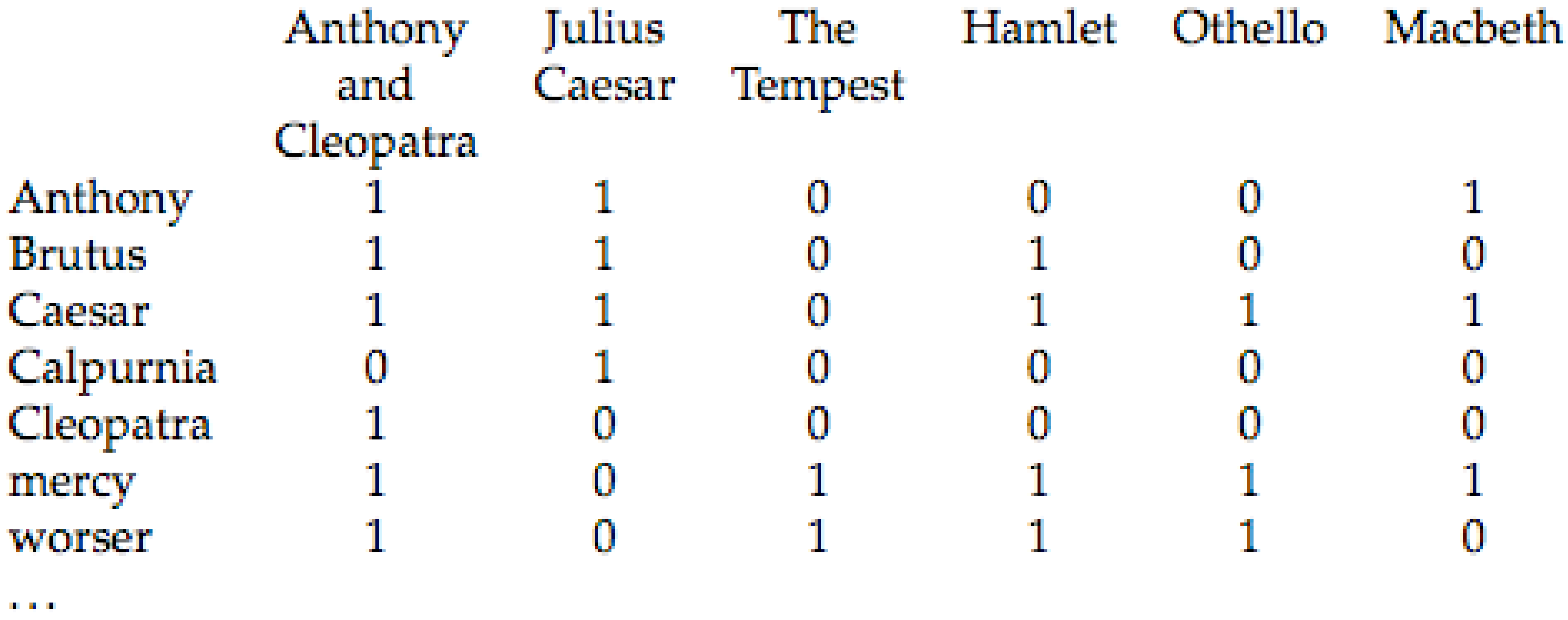
- "Grep"
    - Slow for large corpora
    - Calculating "NOT" requires exhaustive scanning
    - Some operations not feasible
        - Query: "Romans NEAR Countrymen"
    - Doesn't support ranked retrieval
- Moving beyond grep is the motivation for the inverted index.

Querying

Our inverted index is a 2-D array or Matrix

A Column For Each Document

A Row for Each Word (or "Term")



	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

Querying

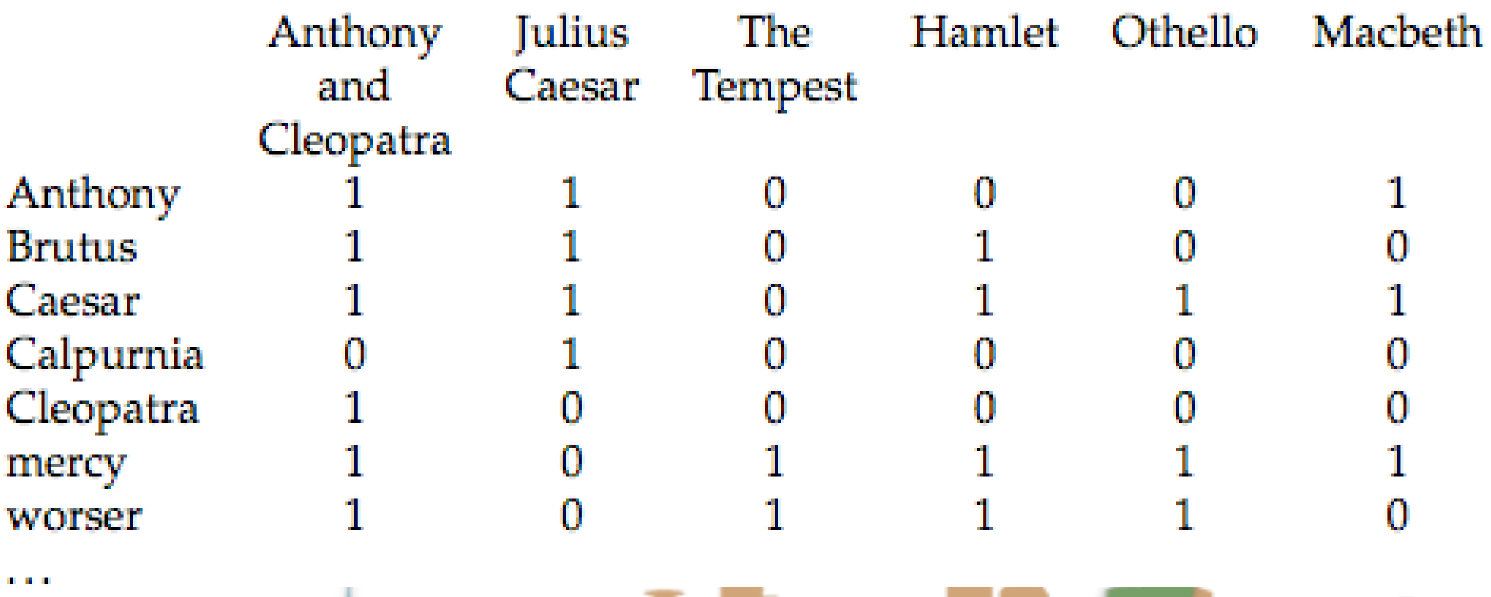
- Boolean Query
  - Queries are boolean expressions
  - Search returns all documents which satisfy the expression
  - Does Google use the Boolean model?

Querying

- Boolean Query
  - Straightforward application of inverted index
  - where cells of inverted index are (0,1)
    - indicating presence or absence of a term

Document

Term



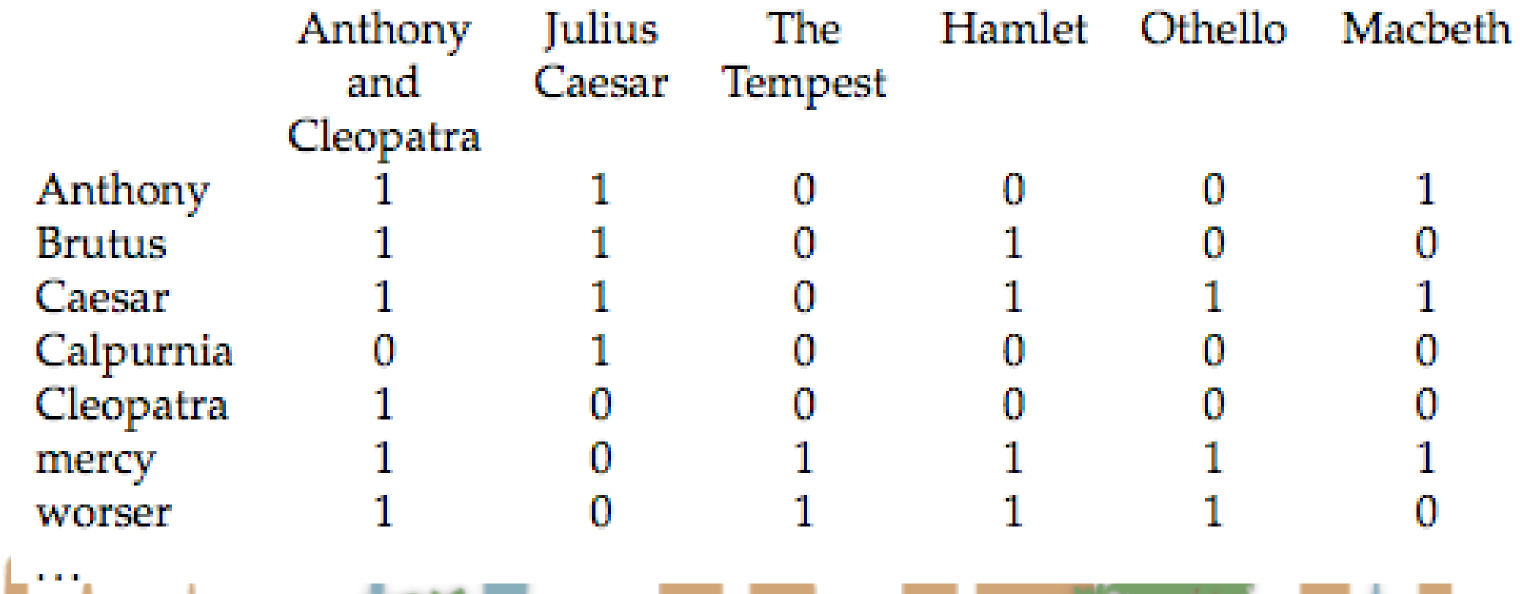
	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

Querying

- Boolean Query
  - 0/1 vector for each term
  - “Brutus AND Caesar AND NOT Calpurnia =
  - Perform bitwise Boolean operation on each row:
    - 110100 AND 110111 AND !(010000) = 100100

Document

Term



	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

Querying

- Boolean Query
  - A big corpus means a sparse matrix
  - A sparse matrix motivates the introduction of the posting
    - Much less space to store
    - Only recording the “1” positions

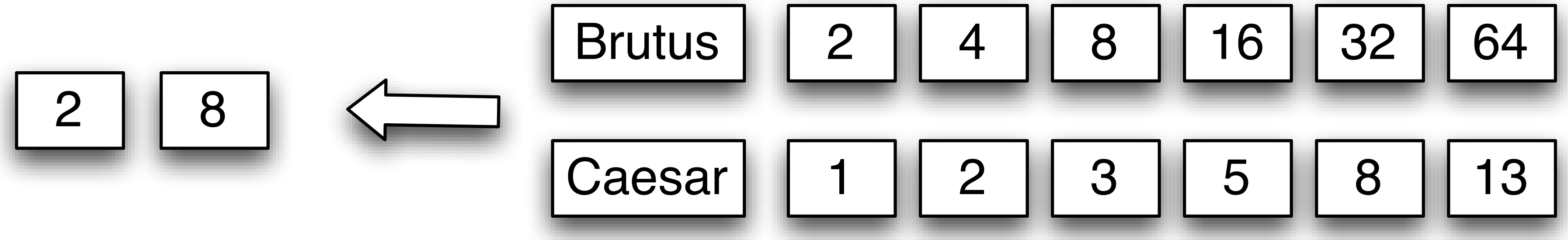
Querying

- Boolean Query
  - Query processing on postings
  - Brutus AND Caesar
    - Locate the postings for Brutus
    - Locate the postings for Caesar
    - Merge the postings

Brutus	2	4	8	16	32	64
Caesar	1	2	3	5	8	13

Querying

- Boolean Query
  - Merging -> walk through the two posting simultaneously
  - postings sorted by doc ID



2 8

Brutus 2 4 8 16 32 64

Caesar 1 2 3 5 8 13

Querying

- Boolean Query
  - An algorithm based on postings
  - Linear in the size of the postings

```
INTERSECT($p_1, p_2$)
  $answer \leftarrow <>$
  while $p_1 \neq nil$ and $p_2 \neq nil$
      do if $docID(p_1) = docID(p_2)$
            then ADD($answer, docID(p_1)$)
                 $p_1 \leftarrow next(p_1)$
                 $p_2 \leftarrow next(p_2)$
            else if $docID(p_1) < docID(p_2)$
                   then $p_1 \leftarrow next(p_1)$
                   else $p_2 \leftarrow next(p_2)$
  return $answer$
```

Querying

- Boolean Query
  - Is the algorithmic complexity better than scanning?
  - Where would you put more complex formulae?

```
INTERSECT($p_1, p_2$)
  $answer \leftarrow <>$
  while $p_1 \neq nil$ and $p_2 \neq nil$
      do if $docID(p_1) = docID(p_2)$
            then ADD($answer, docID(p_1)$)
                 $p_1 \leftarrow next(p_1)$
                 $p_2 \leftarrow next(p_2)$
            else if $docID(p_1) < docID(p_2)$
                   then $p_1 \leftarrow next(p_1)$
                   else $p_2 \leftarrow next(p_2)$
  return $answer$
```

Querying

- Boolean Queries
  - Exact match
  - Views each document as a "bag of words"
  - Precise: a document matches or it doesn't
  - Primary commercial retrieval tool for 3 decades
  - Professional searchers (e.g., lawyers) still like Boolean queries
    - No question about what you are getting

Building up our query technology

- Linear on-demand retrieval (aka grep)
- 0/1 Vector-Based Boolean Queries
- Posting-Based Boolean Queries

Building up our query technology

- Linear on-demand retrieval (aka grep)
- 0/1 Vector-Based Boolean Queries
- Posting-Based Boolean Queries

- How would it apply to
  - http://www.rhymezone.com/shakespeare/

Boolean Model vs. Ranked Retrieval Methods

* Only game for 30 years

* uses precise queries

* user decides relevance

* stayed current with proximity queries

* precise controlled queries

* transparent queries

* controlled queries

* Appeared with www

* uses “free-text” queries

* system decides relevance

* works with enormous corpora

* “no guarantees” in queries

Querying - Boolean Search Example

- Westlaw
  - Largest commercial (paying subscribers) legal search service (started in 1975, ranking added in 1992)
  - Tens of terabytes of data
  - 700,000 users
  - Majority of users still use boolean queries (default in 2005)
    - Example:
      - What is the status of limitations in cases involving federal tort claims act?
      - LIMIT! /3 STATUTE ACTION /S FEDERAL /2 TORT /3 CLAIM
      - /3 = within 3 words. /S same sentence

Querying - Boolean Search Example

- Westlaw
  - Example:
    - Requirements for disabled people to be able to access a workplace
    - disabl! /p access! /s work-site work-place employment /3 place
    - space is a disjunction not a conjunction
    - long precise queries, proximity operators, incrementally developed, not like web search
    - preferred by professionals, but not necessarily better

Building up our query technology

- “Matching” search
    - Linear on-demand retrieval (aka grep)
    - 0/1 Vector-Based Boolean Queries
    - Posting-Based Boolean Queries
- Ranked search
    - Parametric Search

Ranked Search

- Rather than saying
  - (query, document) matches or not (0,1)
    - (“Capulet”,”Romeo and Juliet) = 1
- Now we are going to assign rankings
  - (query, document) in {0,1}
    - (“capulet”,”Romeo and Juliet”) = 0.7

Querying

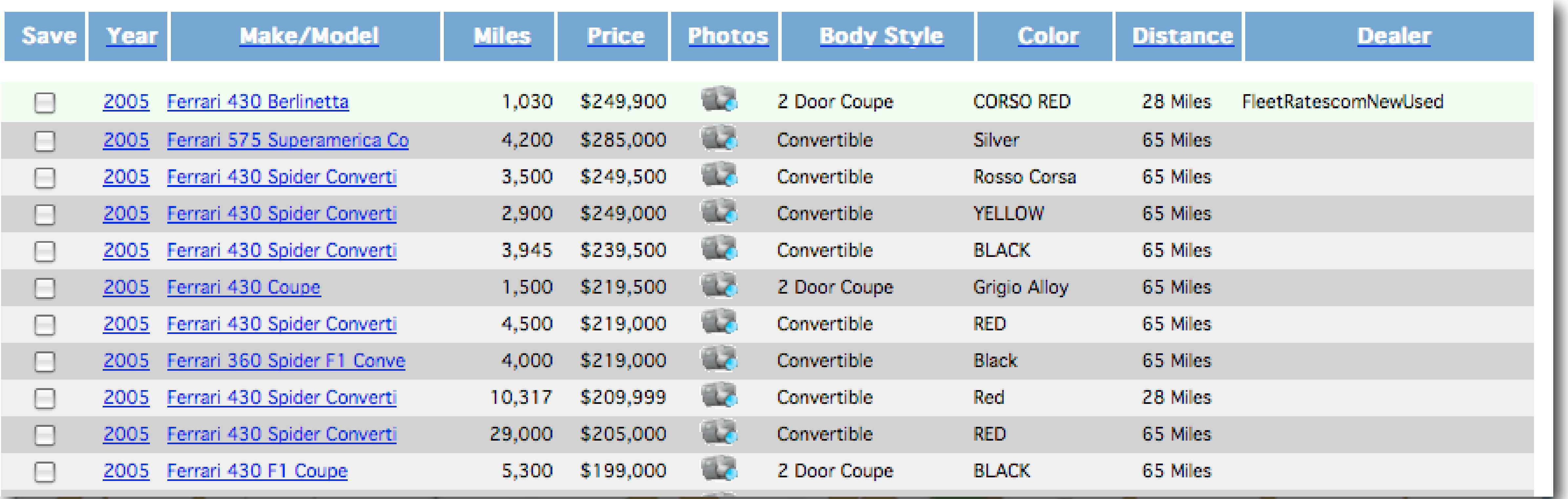
- Metadata = structured additional information about a document.
  - Examples:
    - The author of a document
    - The creation date of a document
    - The title of a document
    - The location where a document was created
  - author, creation date, title, location are fields
  - searching for “William Shakespeare” in a doc differs from
  - searching for “William Shakespeare” in the author of a doc

Querying

- Parametric Search
  - supports searching on meta-data explicitly
  - a parametric search interface allows a mix of full-text query and meta-data queries
  - Example:
    - www.carfinder.com

Querying

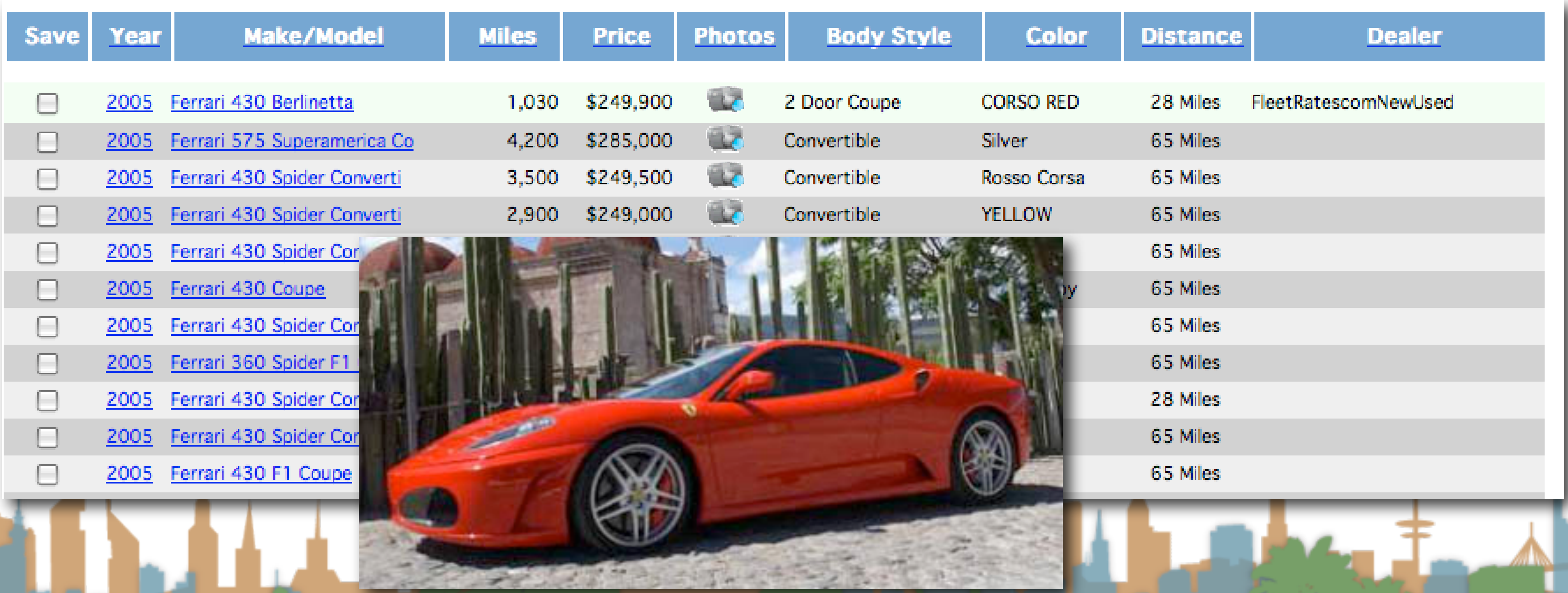
- Parametric Search
  - Example:
    - Result is a large table
    - Columns are fields
    - Searching for “2005” only applied to year field



Save	Year	Make/Model	Miles	Price	Photos	Body Style	Color	Distance	Dealer
□	2005	Ferrari 430 Berlinetta	1,030	$249,900	📷	2 Door Coupe	CORSO RED	28 Miles	FleetRatescomNewUsed
□	2005	Ferrari 575 Superamerica Co	4,200	$285,000	📷	Convertible	Silver	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,500	$249,500	📷	Convertible	Rosso Corsa	65 Miles	
□	2005	Ferrari 430 Spider Converti	2,900	$249,000	📷	Convertible	YELLOW	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,945	$239,500	📷	Convertible	BLACK	65 Miles	
□	2005	Ferrari 430 Coupe	1,500	$219,500	📷	2 Door Coupe	Grigio Alloy	65 Miles	
□	2005	Ferrari 430 Spider Converti	4,500	$219,000	📷	Convertible	RED	65 Miles	
□	2005	Ferrari 360 Spider F1 Conve	4,000	$219,000	📷	Convertible	Black	65 Miles	
□	2005	Ferrari 430 Spider Converti	10,317	$209,999	📷	Convertible	Red	28 Miles	
□	2005	Ferrari 430 Spider Converti	29,000	$205,000	📷	Convertible	RED	65 Miles	
□	2005	Ferrari 430 F1 Coupe	5,300	$199,000	📷	2 Door Coupe	BLACK	65 Miles	

Querying

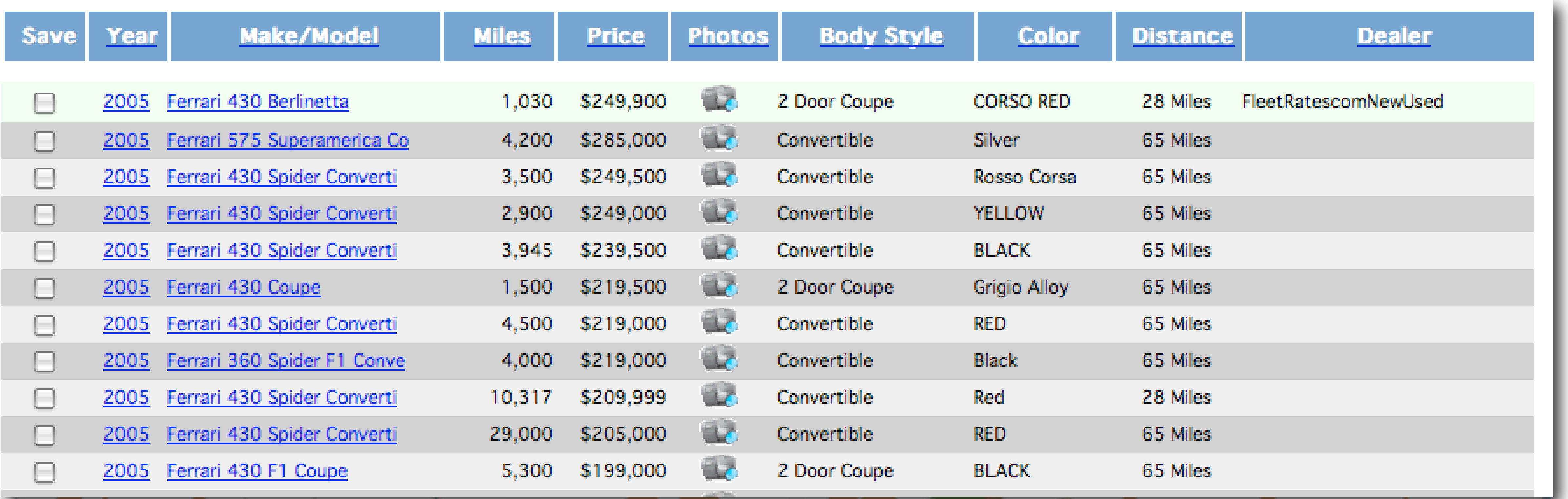
- Parametric Search
  - Example:
    - Result is a large table
    - Columns are fields
    - Searching for “2005” only applied to year field



Save	Year	Make/Model	Miles	Price	Photos	Body Style	Color	Distance	Dealer
□	2005	Ferrari 430 Berlinetta	1,030	$249,900	📷	2 Door Coupe	CORSO RED	28 Miles	FleetRatescomNewUsed
□	2005	Ferrari 575 Superamerica Co	4,200	$285,000	📷	Convertible	Silver	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,500	$249,500	📷	Convertible	Rosso Corsa	65 Miles	
□	2005	Ferrari 430 Spider Converti	2,900	$249,000	📷	Convertible	YELLOW	65 Miles	
□	2005	Ferrari 430 Spider Cor						65 Miles	
□	2005	Ferrari 430 Coupe					by	65 Miles	
□	2005	Ferrari 430 Spider Cor						65 Miles	
□	2005	Ferrari 360 Spider F1						65 Miles	
□	2005	Ferrari 430 Spider Cor						28 Miles	
□	2005	Ferrari 430 Spider Cor						65 Miles	
□	2005	Ferrari 430 F1 Coupe						65 Miles	

Querying

- Parametric Search
  - Example:
    - Result is a large table
    - Columns are fields
    - Searching for “2005” only applied to year field



Save	Year	Make/Model	Miles	Price	Photos	Body Style	Color	Distance	Dealer
□	2005	Ferrari 430 Berlinetta	1,030	$249,900	📷	2 Door Coupe	CORSO RED	28 Miles	FleetRatescomNewUsed
□	2005	Ferrari 575 Superamerica Co	4,200	$285,000	📷	Convertible	Silver	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,500	$249,500	📷	Convertible	Rosso Corsa	65 Miles	
□	2005	Ferrari 430 Spider Converti	2,900	$249,000	📷	Convertible	YELLOW	65 Miles	
□	2005	Ferrari 430 Spider Converti	3,945	$239,500	📷	Convertible	BLACK	65 Miles	
□	2005	Ferrari 430 Coupe	1,500	$219,500	📷	2 Door Coupe	Grigio Alloy	65 Miles	
□	2005	Ferrari 430 Spider Converti	4,500	$219,000	📷	Convertible	RED	65 Miles	
□	2005	Ferrari 360 Spider F1 Conve	4,000	$219,000	📷	Convertible	Black	65 Miles	
□	2005	Ferrari 430 Spider Converti	10,317	$209,999	📷	Convertible	Red	28 Miles	
□	2005	Ferrari 430 Spider Converti	29,000	$205,000	📷	Convertible	RED	65 Miles	
□	2005	Ferrari 430 F1 Coupe	5,300	$199,000	📷	2 Door Coupe	BLACK	65 Miles	

Querying

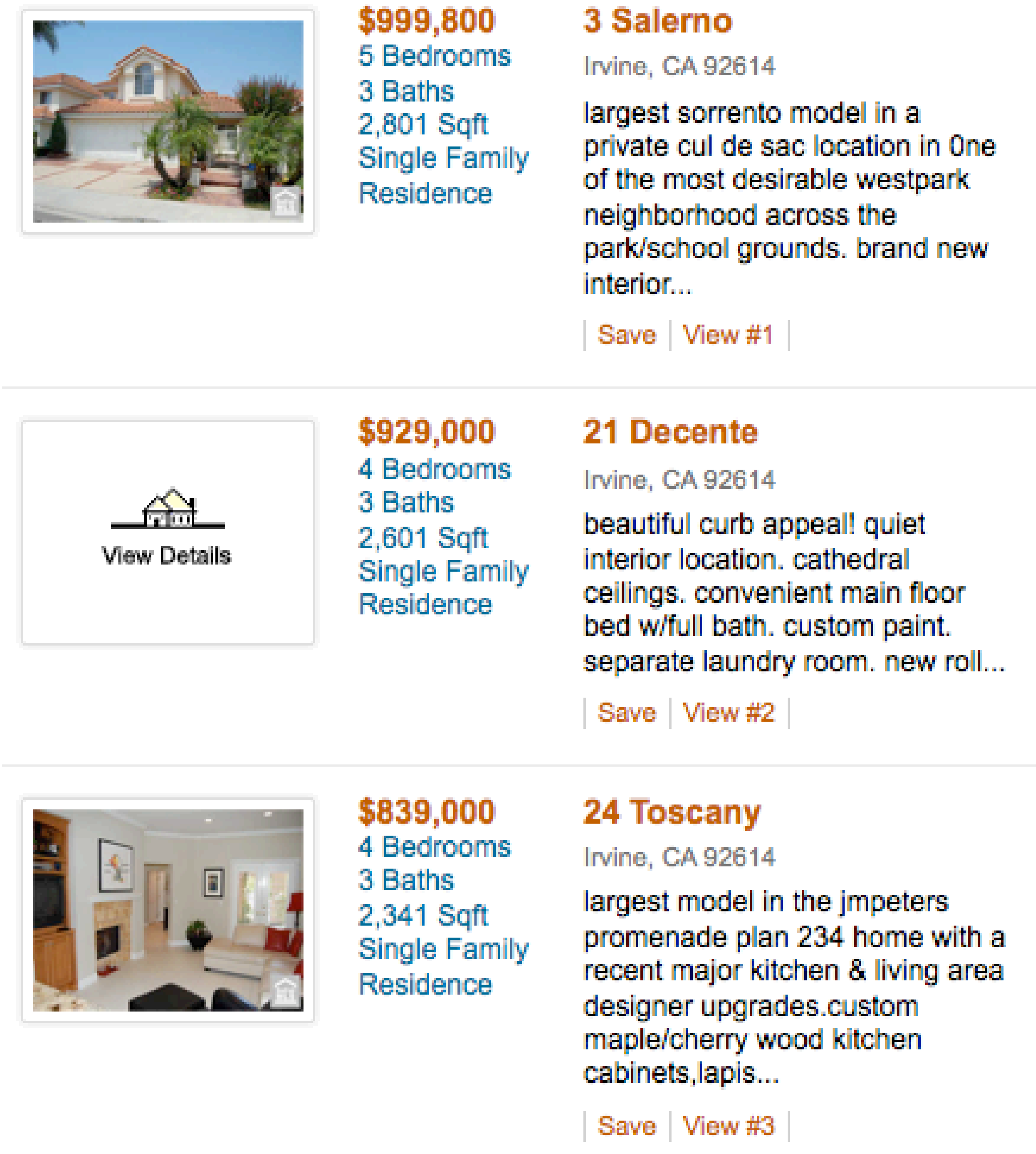
- Parametric Search
  - Example:
    - http://www.ocregister.com/realestate/

Querying

- Parametric Search
  - Example:
    - http://www.ocregister.com/realestate/

Querying

- Parametric Search
  - Example:
    - http://www.ocregister.com/realestate/
    - 92614: 77 results



$999,800
5 Bedrooms
3 Baths
2,801 Sqft
Single Family Residence

3 Salerno
Irvine, CA 92614
largest sorrento model in a private cul de sac location in 0ne of the most desirable westpark neighborhood across the park/school grounds. brand new interior...
| Save | View #1 |

$929,000
4 Bedrooms
3 Baths
2,601 Sqft
Single Family Residence

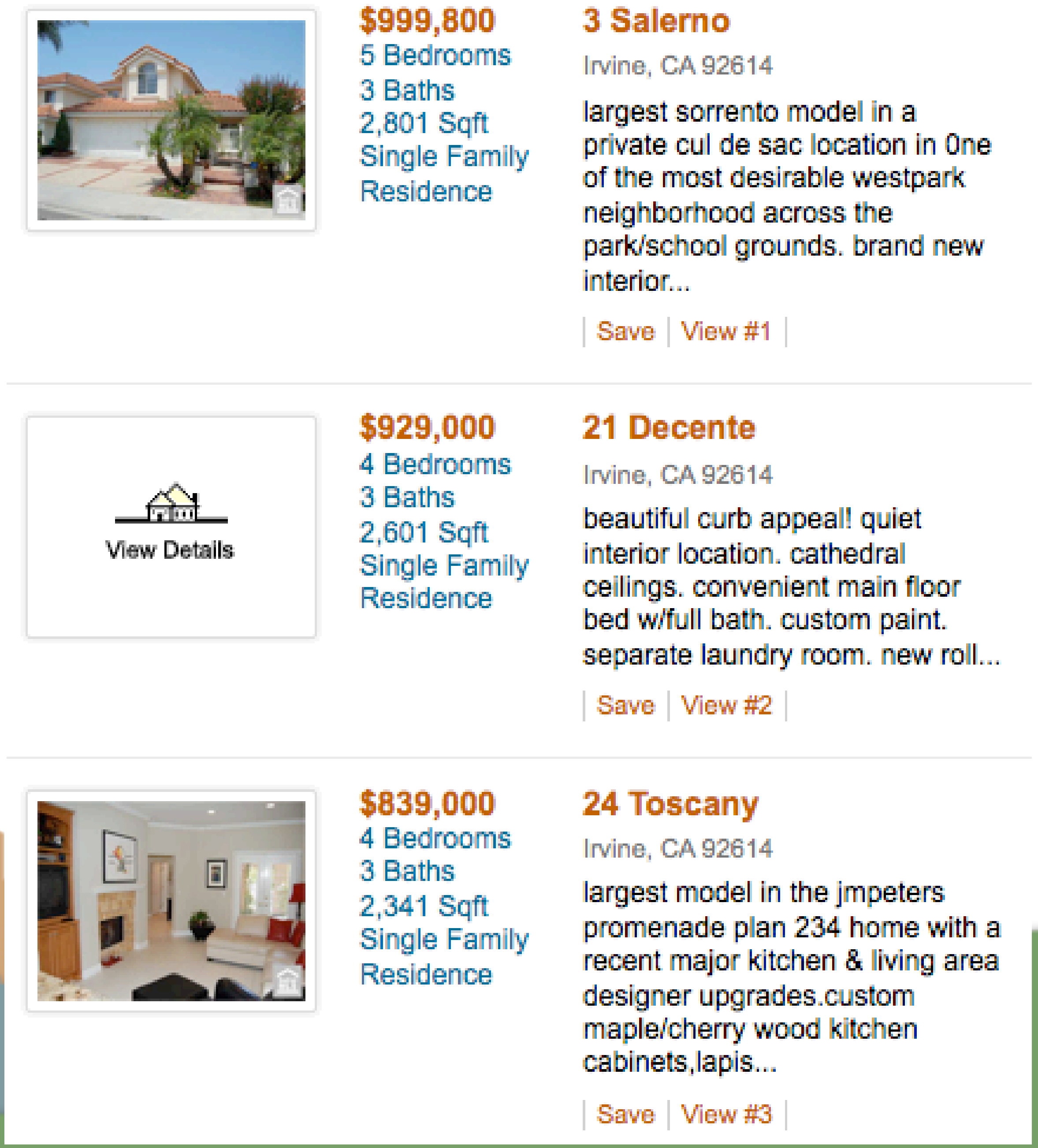
21 Decente
Irvine, CA 92614
beautiful curb appeal! quiet interior location. cathedral ceilings. convenient main floor bed w/full bath. custom paint. separate laundry room. new roll...
| Save | View #2 |

$839,000
4 Bedrooms
3 Baths
2,341 Sqft
Single Family Residence

24 Toscany
Irvine, CA 92614
largest model in the jmpeters promenade plan 234 home with a recent major kitchen & living area designer upgrades.custom maple/cherry wood kitchen cabinets,lapis...
| Save | View #3 |

Querying

- Parametric Search
  - Example:
    - http://www.ocregister.com/realestate/
    - 92614: 77 results



$999,800
5 Bedrooms
3 Baths
2,801 Sqft
Single Family Residence

3 Salerno
Irvine, CA 92614
largest sorrento model in a private cul de sac location in One of the most desirable westpark neighborhood across the park/school grounds. brand new interior...
| Save | View #1 |

$929,000
4 Bedrooms
3 Baths
2,601 Sqft
Single Family Residence

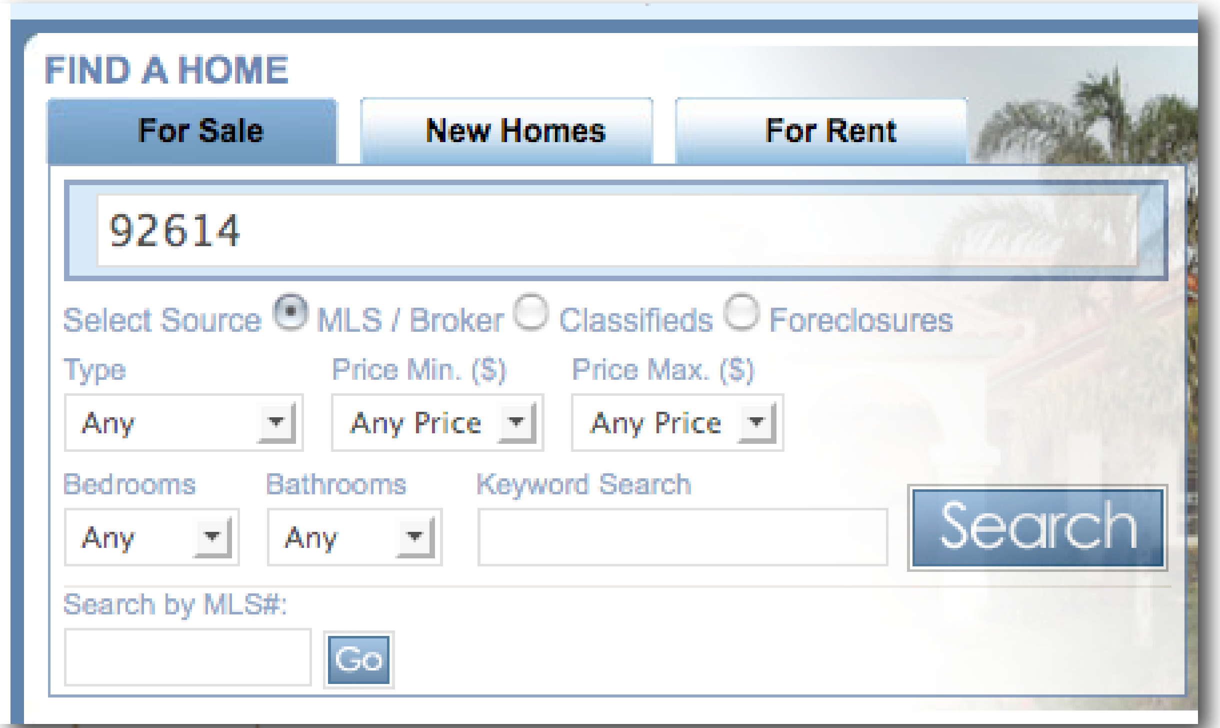
21 Decente
Irvine, CA 92614
beautiful curb appeal! quiet interior location. cathedral ceilings. convenient main floor bed w/full bath. custom paint. separate laundry room. new roll...
| Save | View #2 |

$839,000
4 Bedrooms
3 Baths
2,341 Sqft
Single Family Residence

24 Toscany
Irvine, CA 92614
largest model in the jmpeters promenade plan 234 home with a recent major kitchen & living area designer upgrades.custom maple/cherry wood kitchen cabinets,lapis...
| Save | View #3 |

Querying

- Parametric Search
  - Example:
    - http://www.ocregister.com/realestate/
    - 92614: 77 results



FIND A HOME

For Sale | New Homes | For Rent

92614

Select Source ● MLS / Broker ○ Classifieds ○ Foreclosures

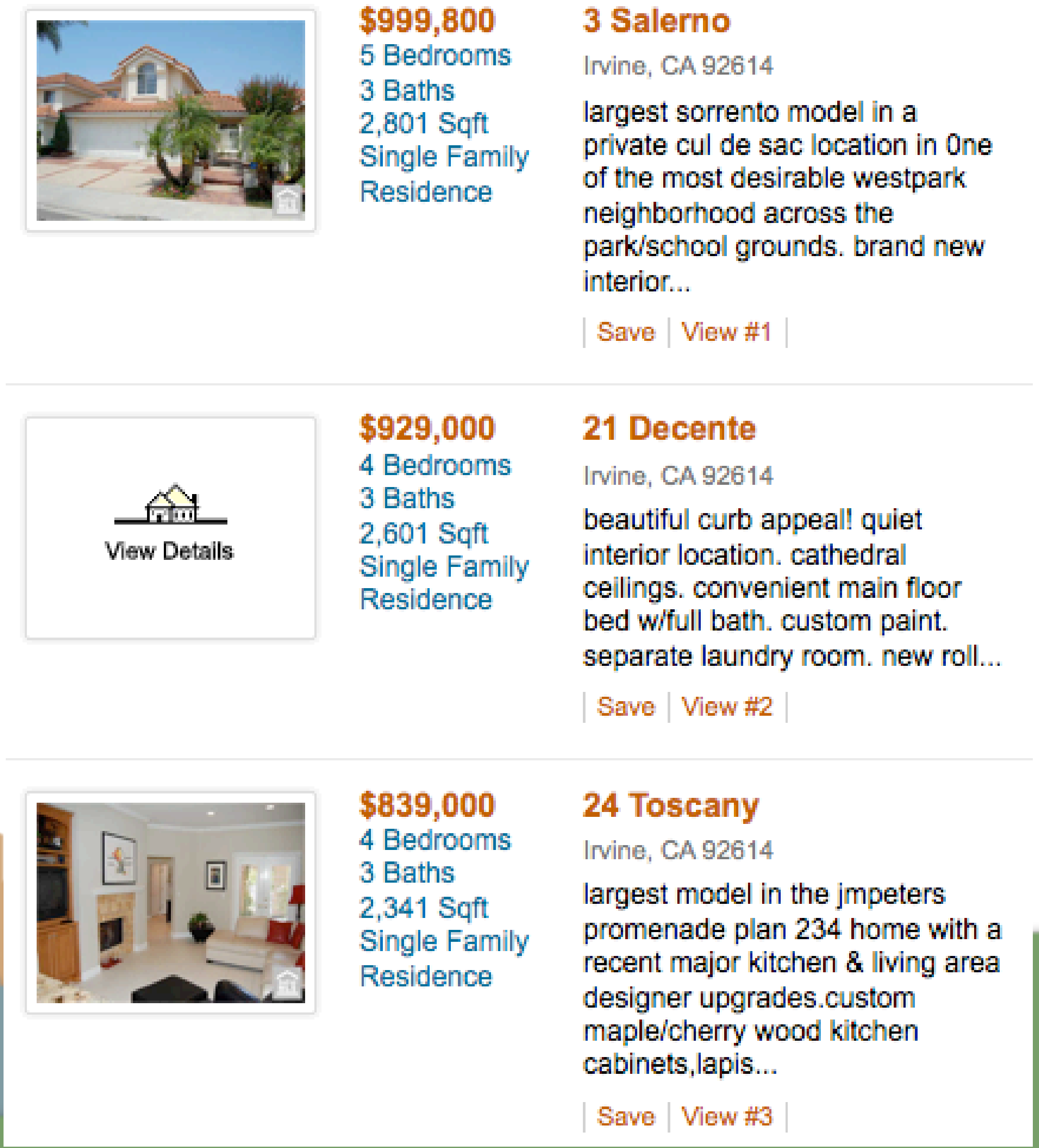
Type: Any
Price Min. ($): Any Price
Price Max. ($): Any Price

Bedrooms: Any
Bathrooms: Any
Keyword Search:

Search

Search by MLS#:

Go



$999,800
5 Bedrooms
3 Baths
2,801 Sqft
Single Family Residence

3 Salerno
Irvine, CA 92614

largest sorrento model in a private cul de sac location in 0ne of the most desirable westpark neighborhood across the park/school grounds. brand new interior...

| Save | View #1 |

View Details

$929,000
4 Bedrooms
3 Baths
2,601 Sqft
Single Family Residence

21 Decente
Irvine, CA 92614

beautiful curb appeal! quiet interior location. cathedral ceilings. convenient main floor bed w/full bath. custom paint. separate laundry room. new roll...

| Save | View #2 |

$839,000
4 Bedrooms
3 Baths
2,341 Sqft
Single Family Residence

24 Toscany
Irvine, CA 92614

largest model in the jmpeters promenade plan 234 home with a recent major kitchen & living area designer upgrades.custom maple/cherry wood kitchen cabinets,lapis...

| Save | View #3 |