Solr for newbies

https://hectorcorrea.com/solr-for-newbies

code{4}lib

Hector Correa
hector_correa@princeton.edu
Princeton University

Useful Links

Workshop links

http://hectorcorrea.com/solr-for-newbies

Code4Lib code of conduct

https://2023.code4lib.org/conduct/

Workshop Outline

1. Introduction
(concepts, quick tour, installation)

2. Schema
(fields, field types, query/index analyzers, tokenizers)

3. Searching
(query parsers, search params, facets, highlighting)

4. Miscellaneous
(directories, configuration, synonyms, spellcheck)

1. Introduction
(concepts, quick tour, installation)

2. Schema
(fields, field types, query/index analyzers, tokenizers)

3. Searching
(query parsers, search params, facets, highlighting)

4. Miscellaneous
(directories, configuration, synonyms, spellcheck)

What is Solr

“Solr is the popular, blazing-fast, open source enterprise **search platform** built on Apache Lucene.”

- Solr’s Home Page

“Solr is a scalable, ready-to-deploy enterprise **search engine** that’s optimized to search large volumes of text-centric data and return results sorted by relevance.”

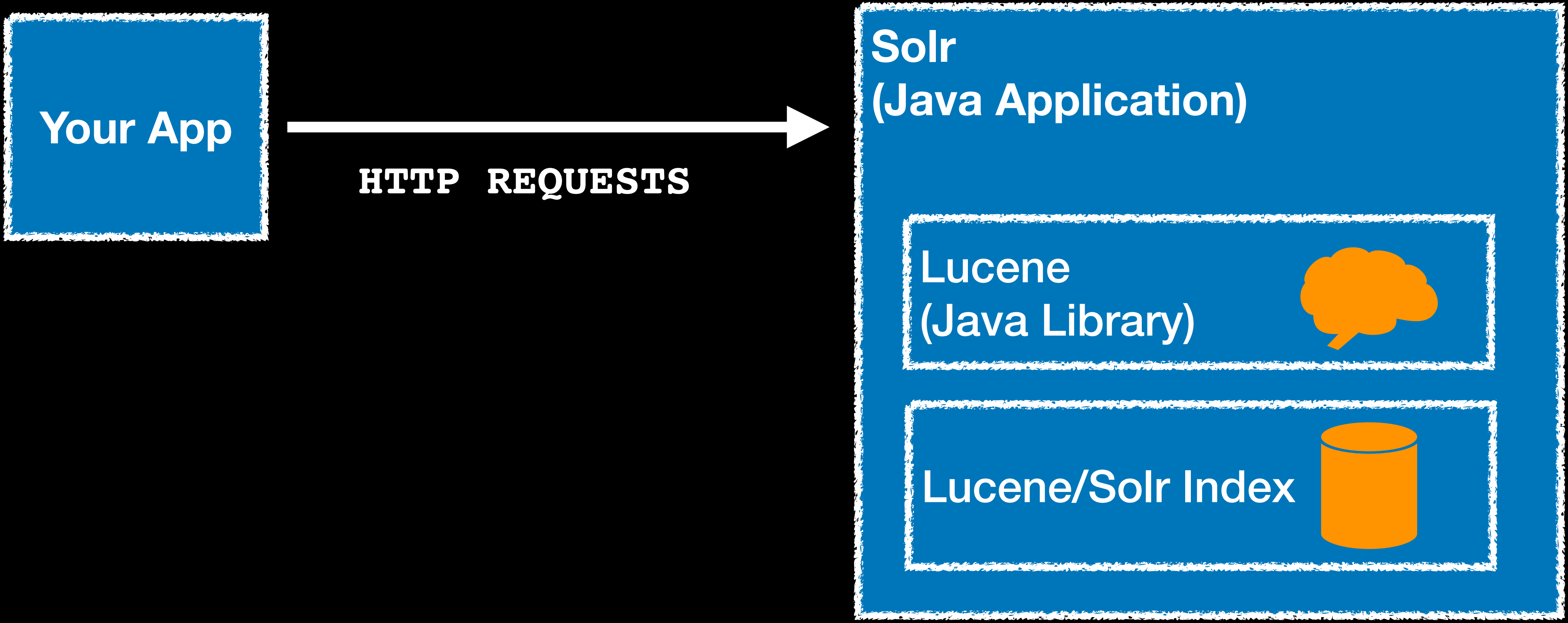
- Solr in Action [p. 4]

What is Solr

“Solr is the popular, blazing-fast,
open source enterprise **search platform**
built on Apache Lucene.”
- Solr’s Home Page

“Solr is a scalable, ready-to-deploy
enterprise **search engine** that’s optimized
to search large volumes of **text-centric data**
and return **results sorted by relevance**.”
- Solr in Action [p. 4]

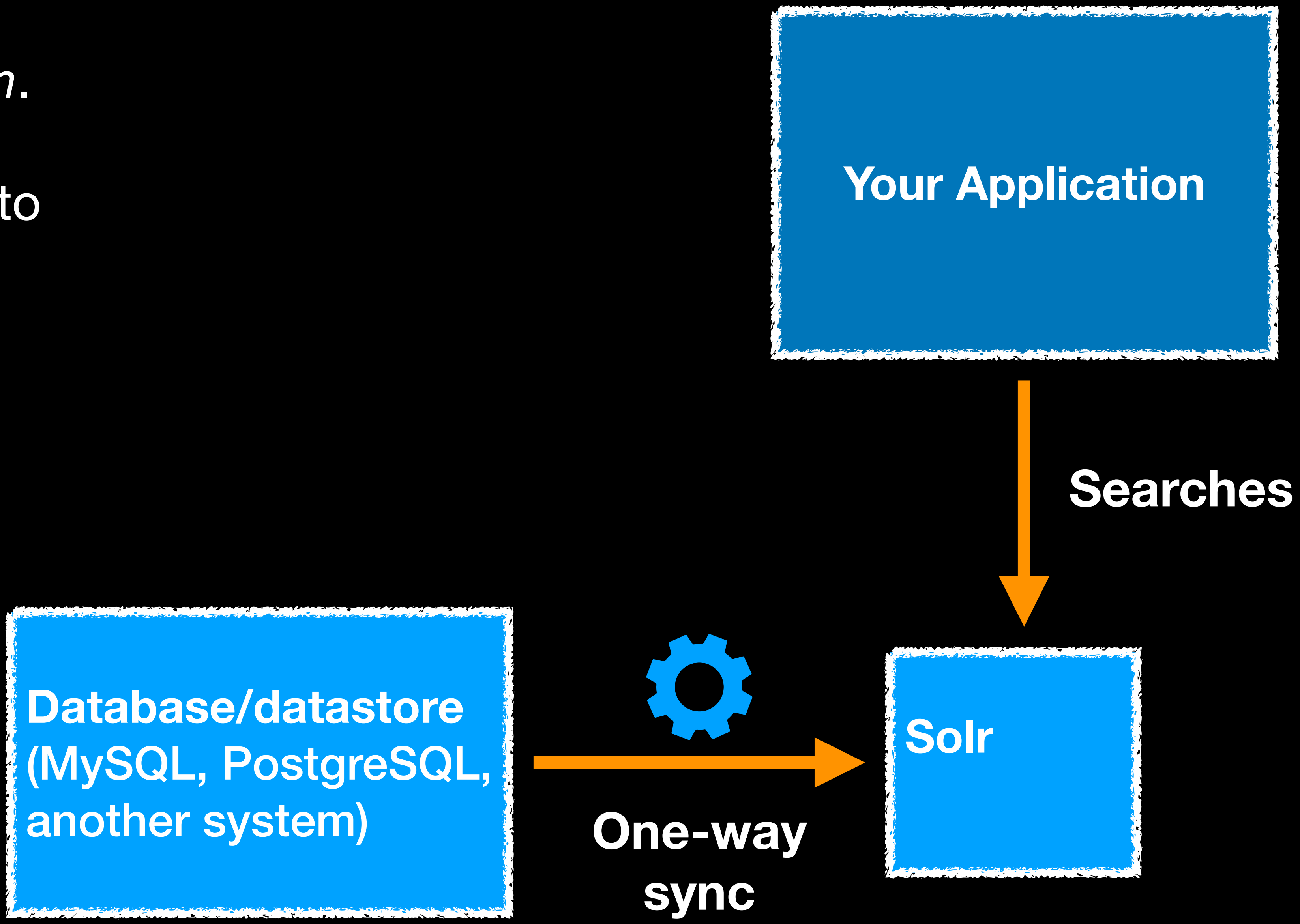
Your App, Solr, and Lucene



Typical Architectures I

Your application searches via Solr, *but the data is maintained in another system*.

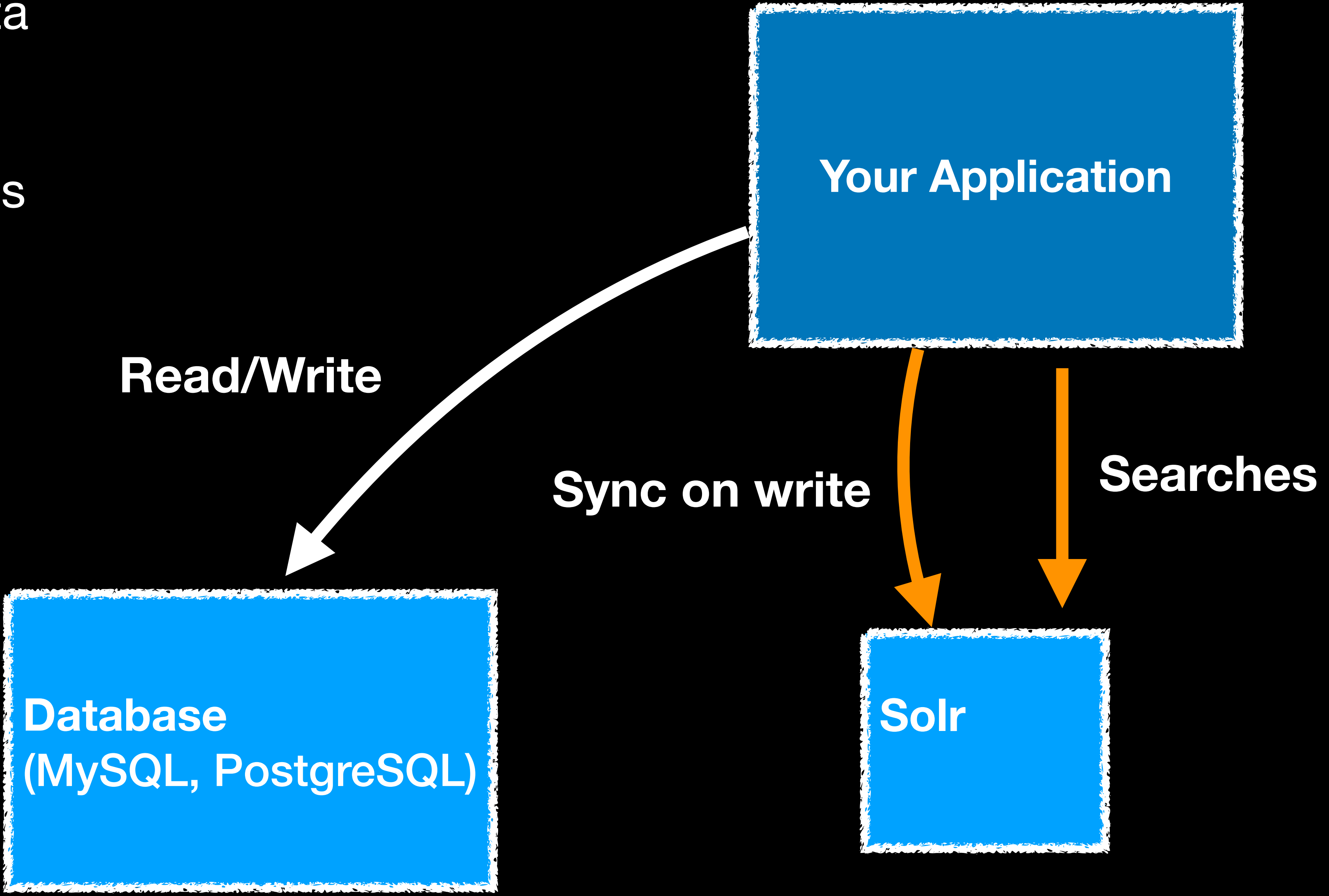
Blacklight applications tend to follow this pattern.



Typical Architectures II

Your application uses a database to maintain the data and Solr for searches.

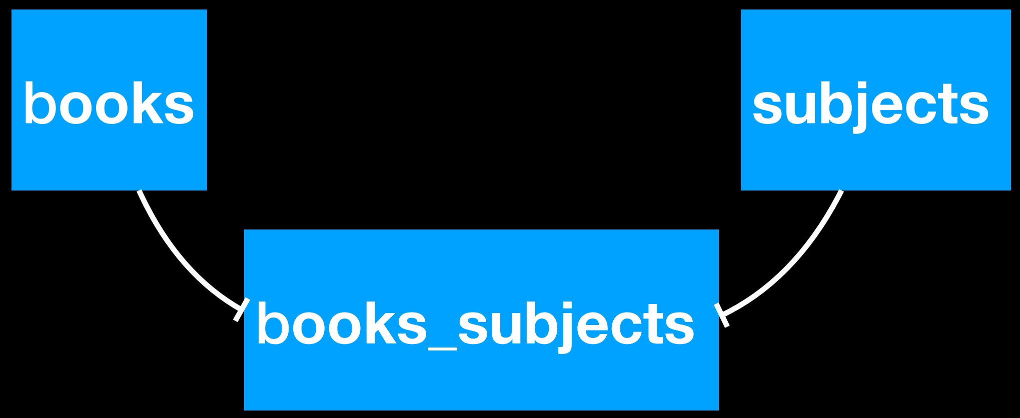
VIVO and SamVera follow this pattern.



Document Model
(how Solr *stores* your data)

id	book_title	subjects
1	Princeton guide for dog owners	animals, guides
2	Princeton tour guide	guides
3	Cats and dogs	animals

Relational Model



Document Model

```
solr_doc: {
  id:”1”,
  title:”Princeton guide for dog owners”,
  subjects: [“animals”, “guides”]
}
```

Solr Documents are flat*

```
your_data:
{
  id:”9041”,
  title:”Using Qualitative Inquiry to Promote…”,
  authors: [
    {uri:”http://somebody/51", name: “Loya, Karla”},
    {uri:”http://somebody/82", name: “Kimball, Ezekiel”}
  ],
  subjects: [“higher education”, “org theory”]
}
```

data in Solr is flatten

```
solr_doc: {
  id:”9041”,
  title:”Using Qualitative Inquiry to Promote…”,
  authors_uri: [”http://somebody/51”, ”http://somebody/82”],
  authors_name: [“Kimball, Ezekiel”, “Loya, Karla”],
  subjects: [“higher education”, “org theory”]
}
```

** in newer versions of Solr there is support for complex/nested objects but that is outside the scope of this workshop*

Inverted Index

(how Solr *indexes* your data)

id	title	subjects
1	Princeton guide for dog owners	animals, guides
2	Princeton tour guide	guides
3	Cats and dogs	animals

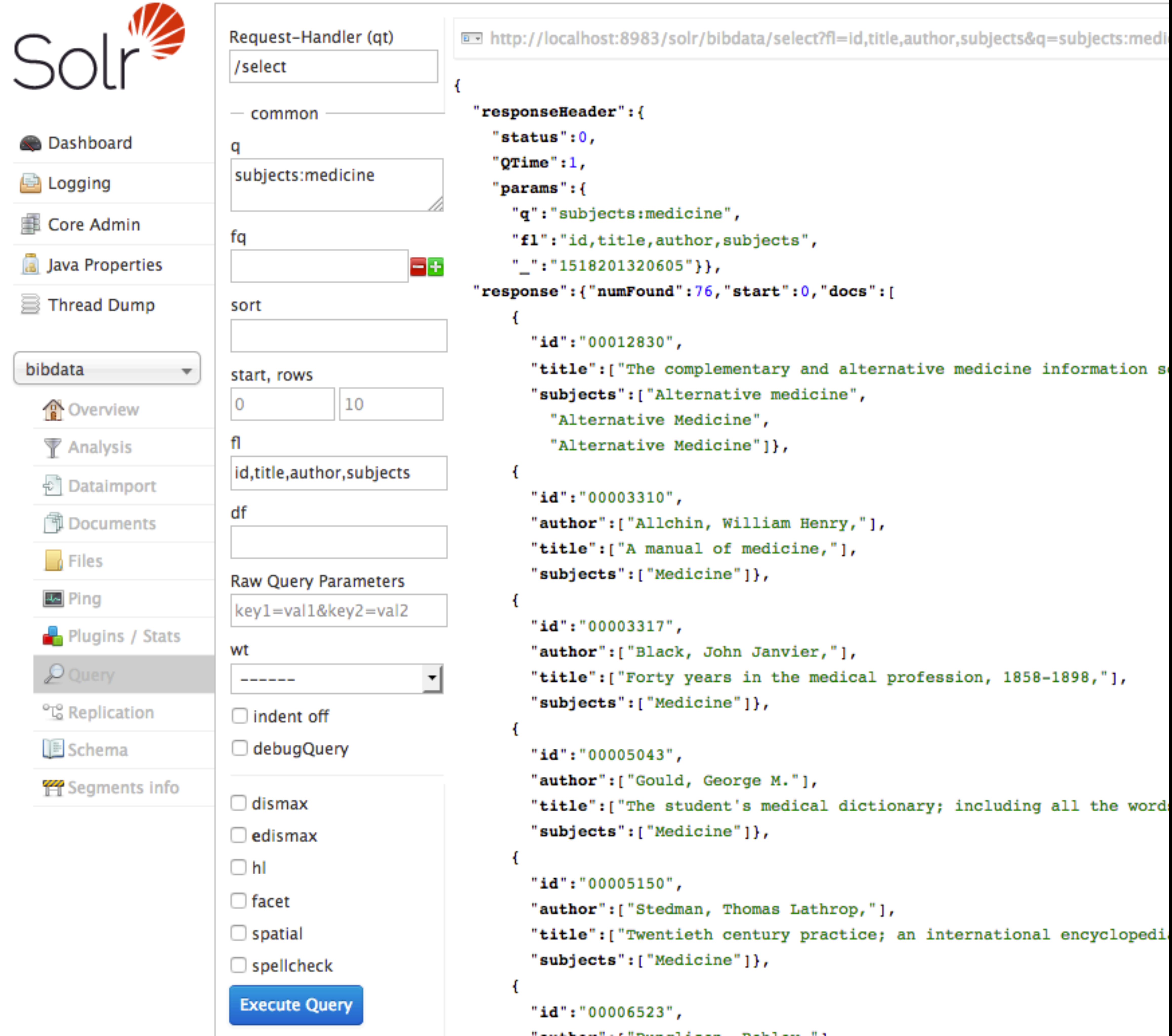
Traditional Index

id	title
3	Cats and dogs
1	Princeton guide for dog owners
2	Princeton tour guide

Inverted Index

key	ids
princeton	1, 2
owners	1
dogs	1, 3
guide	1, 2
tour	2
cats	3

Solr Admin - Quick Tour



http://localhost:8983/solr/bibdata/select?fl=id,title,author,subjects&q=subjects:medi

Request-Handler (qt): /select

common

q: subjects:medicine

fq:

sort:

start, rows: 0 10

fl: id,title,author,subjects

df:

Raw Query Parameters: key1=val1&key2=val2

wt: ------

□ indent off

□ debugQuery

□ dismax

□ edismax

□ hl

□ facet

□ spatial

□ spellcheck

```
{
  "responseHeader":{
    "status":0,
    "QTime":1,
    "params":{
      "q":"subjects:medicine",
      "fl":"id,title,author,subjects",
      "_":"1518201320605"}},
  "response":{"numFound":76,"start":0,"docs":[
      {
        "id":"00012830",
        "title":["The complementary and alternative medicine information s
        "subjects":["Alternative medicine",
          "Alternative Medicine",
          "Alternative Medicine"]},
      {
        "id":"00003310",
        "author":["Allchin, William Henry,"],
        "title":["A manual of medicine,"],
        "subjects":["Medicine"]},
      {
        "id":"00003317",
        "author":["Black, John Janvier,"],
        "title":["Forty years in the medical profession, 1858-1898,"],
        "subjects":["Medicine"]},
      {
        "id":"00005043",
        "author":["Gould, George M."],
        "title":["The student's medical dictionary; including all the word
        "subjects":["Medicine"]},
      {
        "id":"00005150",
        "author":["Stedman, Thomas Lathrop,"],
        "title":["Twentieth century practice; an international encyclopedi
        "subjects":["Medicine"]},
      {
        "id":"00006523",
```
]

1. Introduction
(concepts, quick tour, installation)

2. Schema
(fields, field types, query/ index analyzers, tokenizers)

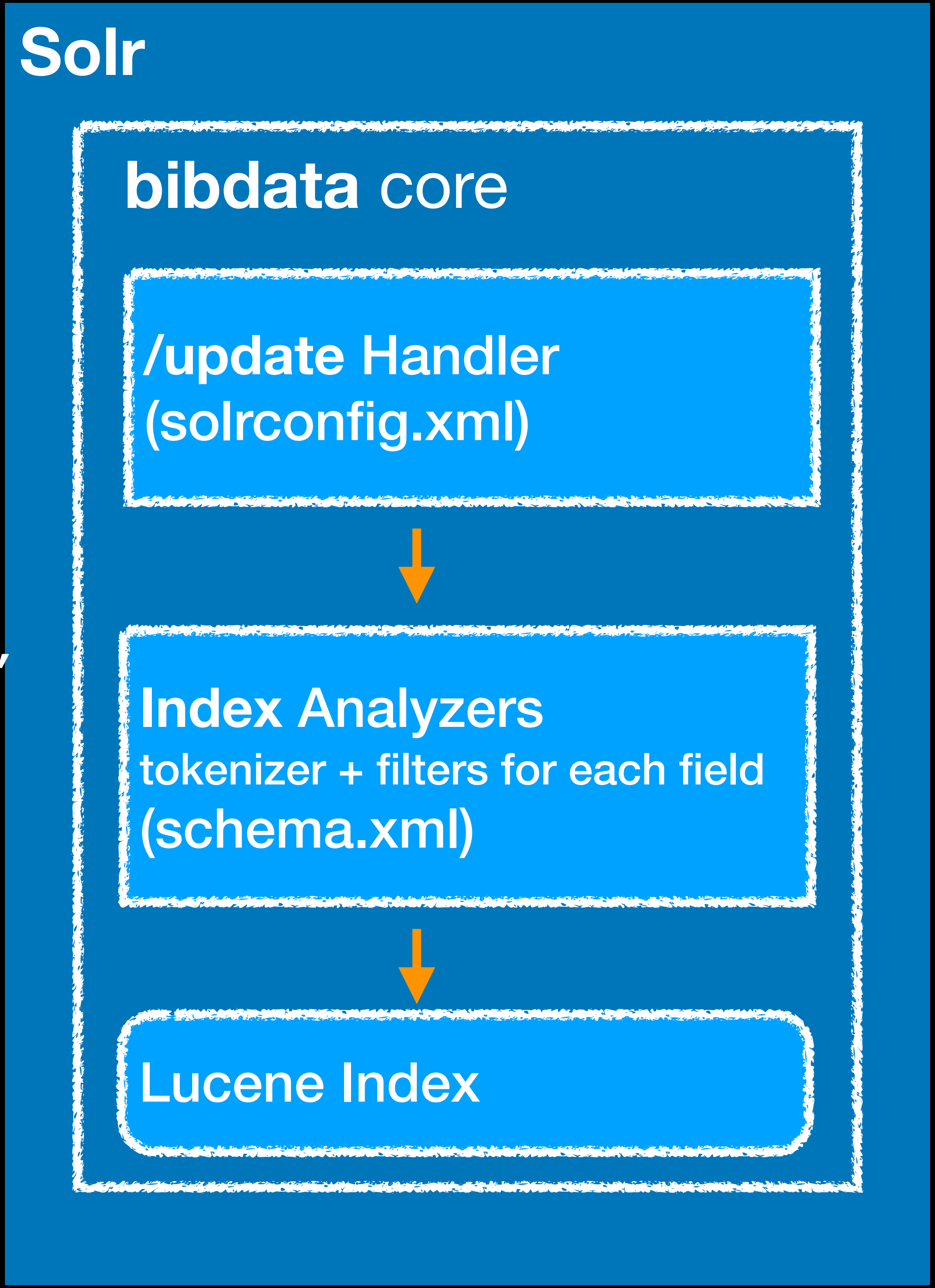
3. Searching
(query parsers, search params, facets, highlighting)

4. Miscellaneous
(directories, configuration, synonyms, spellcheck)

Adding a document to Solr

```
HTTP POST
http://localhost/solr/bibdata/update

{
 id:”1”,
 title_txt_en:”history of medicine”,
 subject_s: “medicine”,
 abstract_txt: “this book is about...”
}
```



Solr

bibdata core

/update Handler
(solrconfig.xml)

Index Analyzers
tokenizer + filters for each field
(schema.xml)

Lucene Index

Adding a document to Solr

Your data

```
{
  id:”1”,
  title_txt_en:”history of medicine”,
  subject_s: “medicine”,
  abstract_txt_en: “this book is about...”
}
```

+ Solr’s Schema

```
<field name="id" type=“string" multiValued=“false” />

<dynamicField name=“*_s” type=“string" />
<dynamicField name=“*_txt_en” type=“text_en" />
<dynamicField name="*_txt" type="text_general" />
```

Gives

id and subject will be handled as a string
title and abstract will be handled as text_en

Adding a document to Solr

id and subject will be handled as a string

```
$ curl localhost:8983/solr/bibdata/schema/fieldtypes/string

“fieldType”:{
    "name":"string",
    "class":"solr.StrField",
    "sortMissingLast":true,
    “docValues”:true
}
```

Adding a document to Solr

Title and abstract will be handled as a text_en

```
$ curl localhost:8983/solr/bibdata/schema/fieldtypes/text_en
```

```
“fieldType":{
    “name":"text_en",
    "class":"solr.TextField",
    "positionIncrementGap":"100",
    “multiValued":true,
    "indexAnalyzer":{
      “tokenizer":{StandardTokenizerFactory},
      “filters”:[StopFilterFactory, LowerCaseFilterFactory,
EnglishPossessiveFilterFactory, PorterStemFilterFactory}]},
    "queryAnalyzer":{
      “tokenizer”:…,
      “filters”:[…]}}
}
```

Workshop Outline

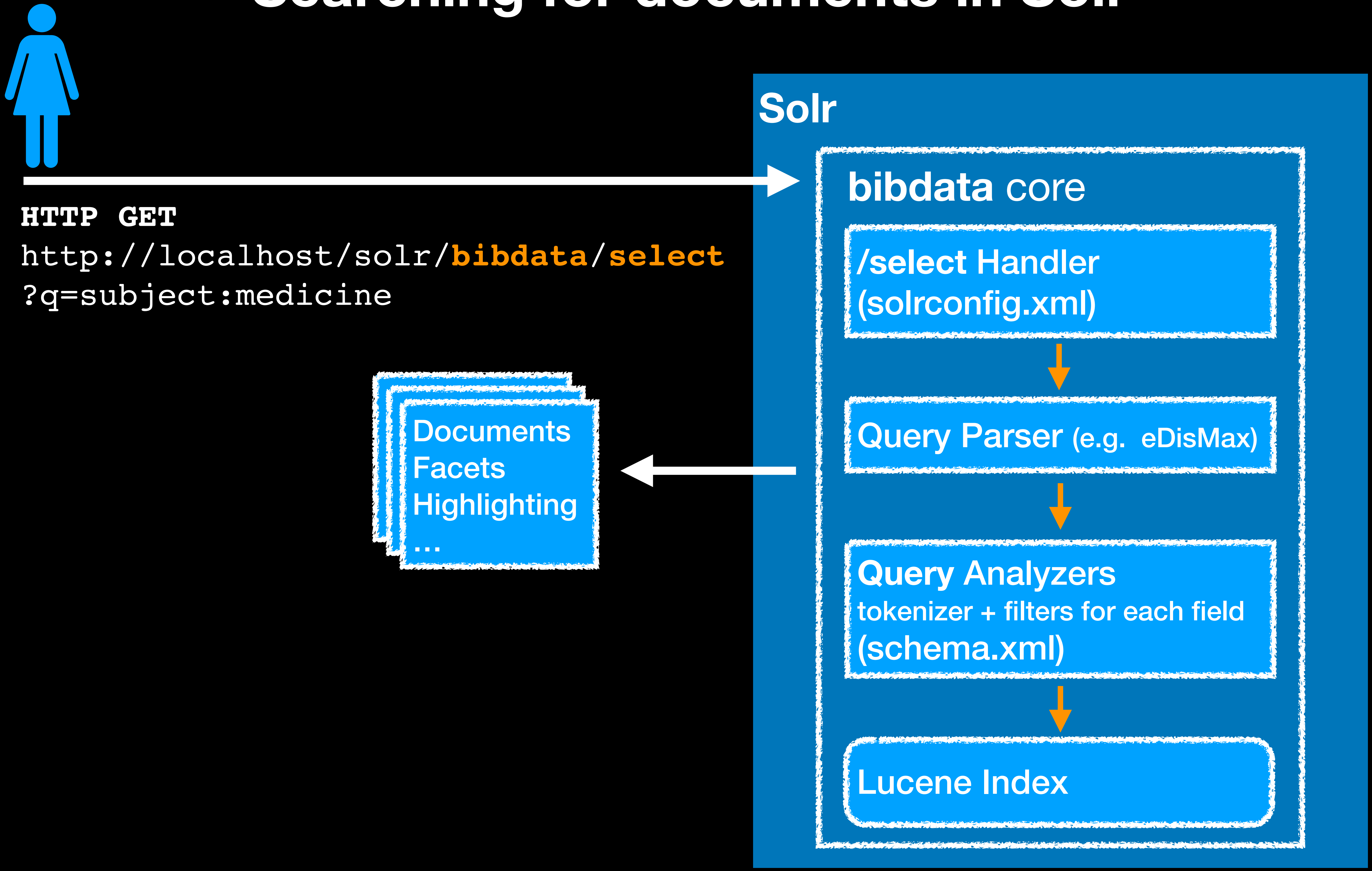
1. Introduction
(concepts, quick tour, installation)

2. Schema
(fields, field types, query/ index analyzers)

3. Searching
(query parsers, search params, facets, highlighting)

4. Miscellaneous
(directories, configuration, synonyms, spellcheck)

Searching for documents in Solr



HTTP GET

```
http://localhost/solr/bibdata/select
?q=subject:medicine
```

Solr

bibdata core

/select Handler
(solrconfig.xml)

Query Parser (e.g. eDisMax)

Query Analyzers
tokenizer + filters for each field
(schema.xml)

Lucene Index

Documents
Facets
Highlighting
…

1. Introduction
(concepts, quick tour, installation)

2. Schema
(fields, field types, query/ index analyzers)

3. Searching
(query parsers, search params, facets, highlighting)

4. Miscellaneous
(directories, configuration, synonyms, spellcheck)

Thanks and good luck

Stay in touch

hector_correa@princeton.edu
`https://hectorcorrea.com/solr-for-newbies`