Searching for Statistical Diagrams

Michael Cafarella
University of Michigan

Joint work with
Shirley Zhe Chen and Eytan Adar

U.S. Frontiers of Engineering Symposium
2011

Statistical Diagrams

- Everywhere in serious academic, governmental, scientific documents
- Our only peek into data behind docs
- Previously rare and precious, Web gives us a *flood*
  - In small Web crawl, found 319K diagrams in 153K academic papers
- Google makes it easy to find docs, images; very hard to find diagrams
- Searching for diagrams part of larger semantic processing trends

U.S. Population 1900 - 2100

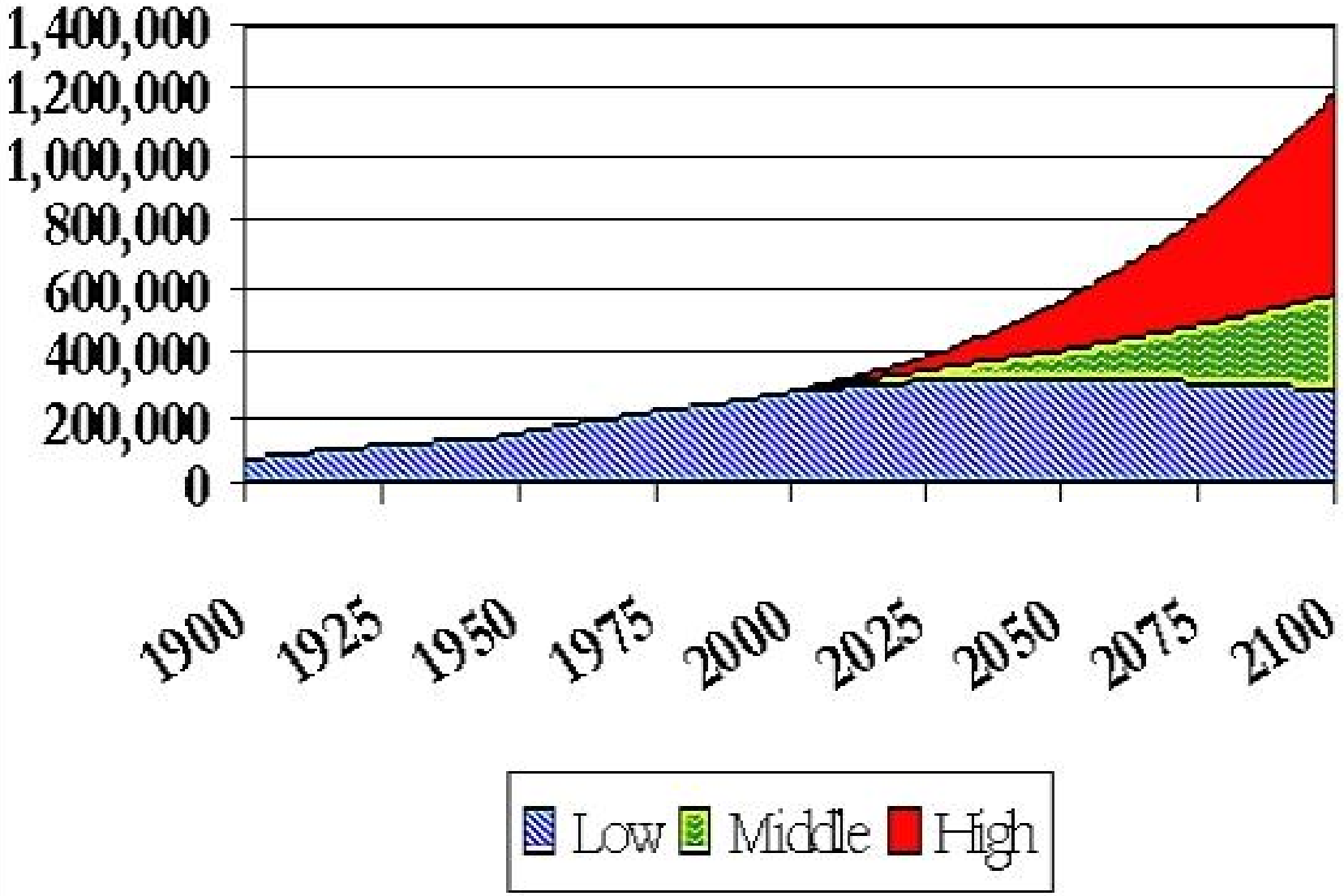
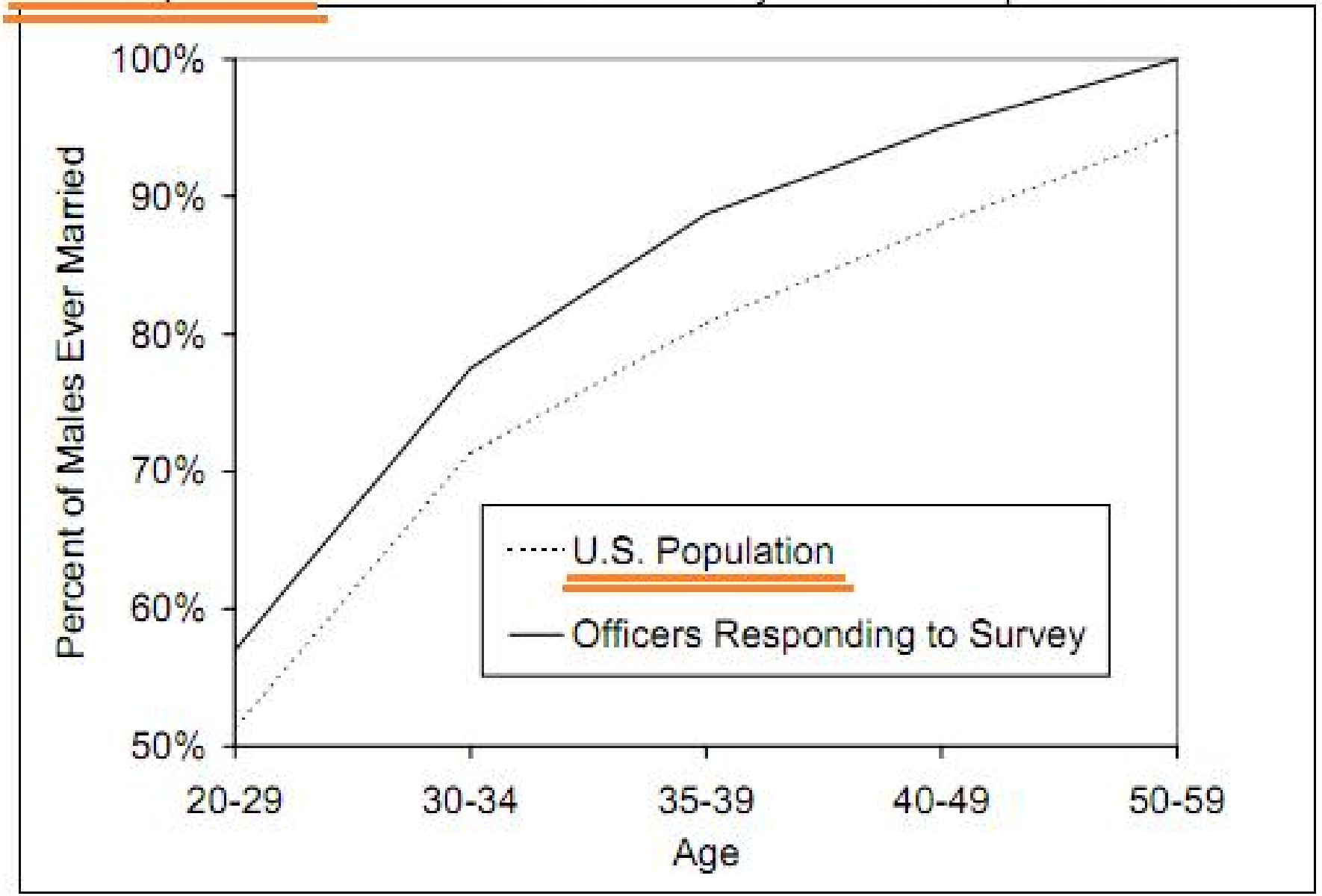
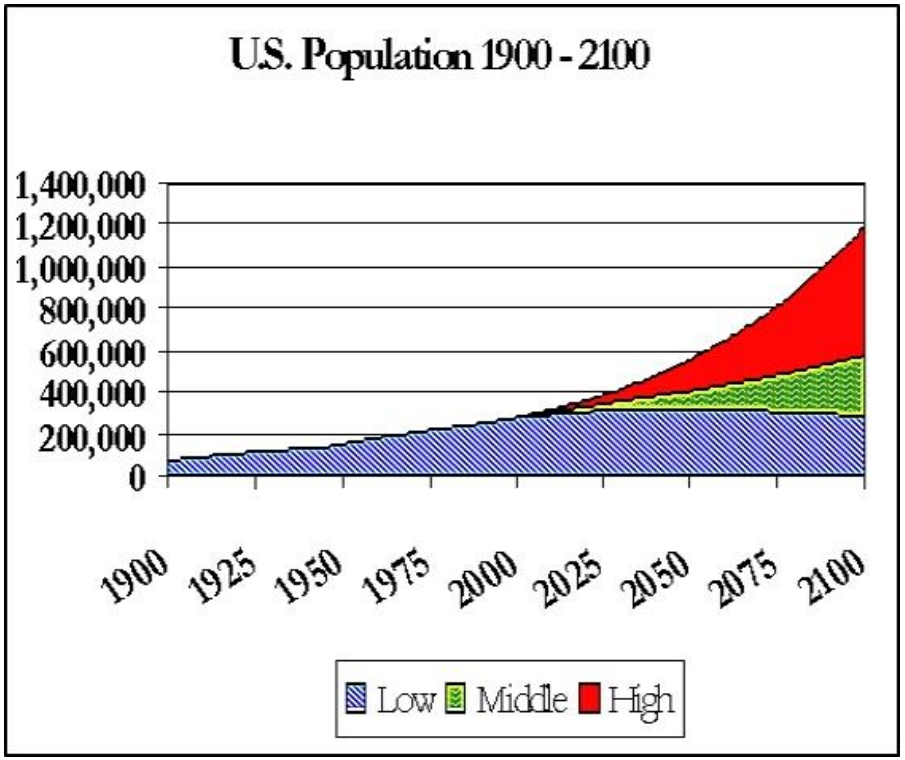
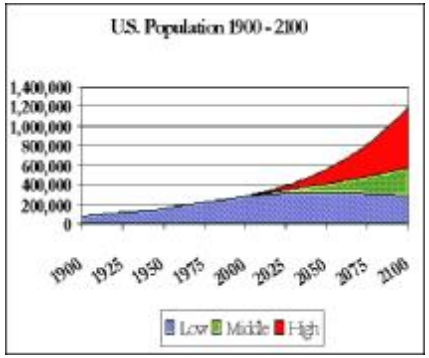
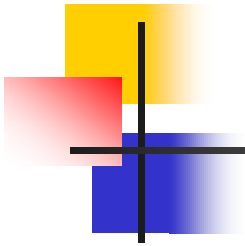


Figure 7.1 Comparison of the Percent of Males Ever Married, U.S. Population and Sacramento County Sheriff's Department Officers



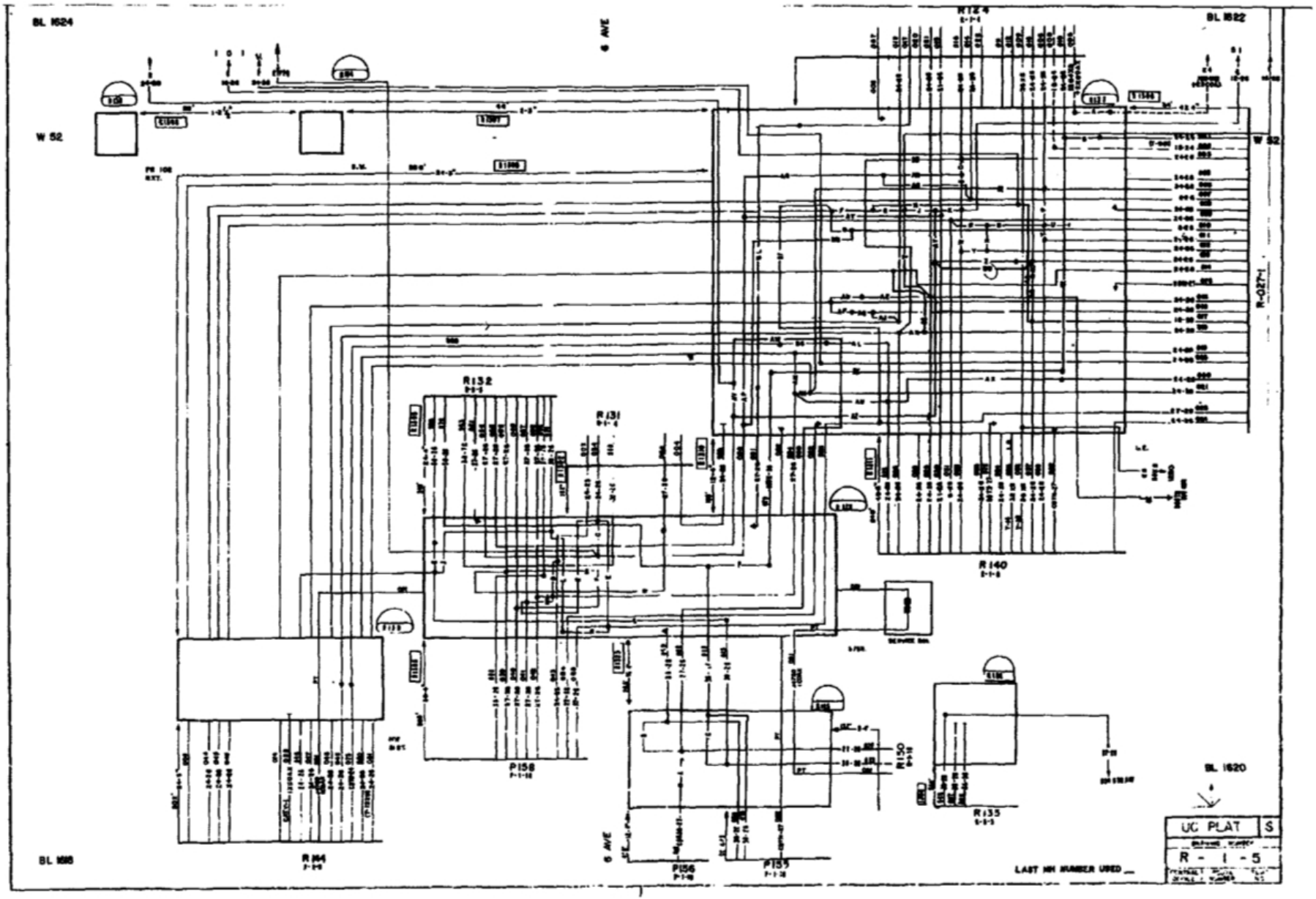

U.S. Population 1900 - 2100
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
1900
1925
1950
1975
2000
2025
2050
2075
2100
Low
Middle
High


U.S. Population 1900 - 2100
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
1900
1925
1950
1975
2000
2025
2050
2075
2100
Low
Middle
High

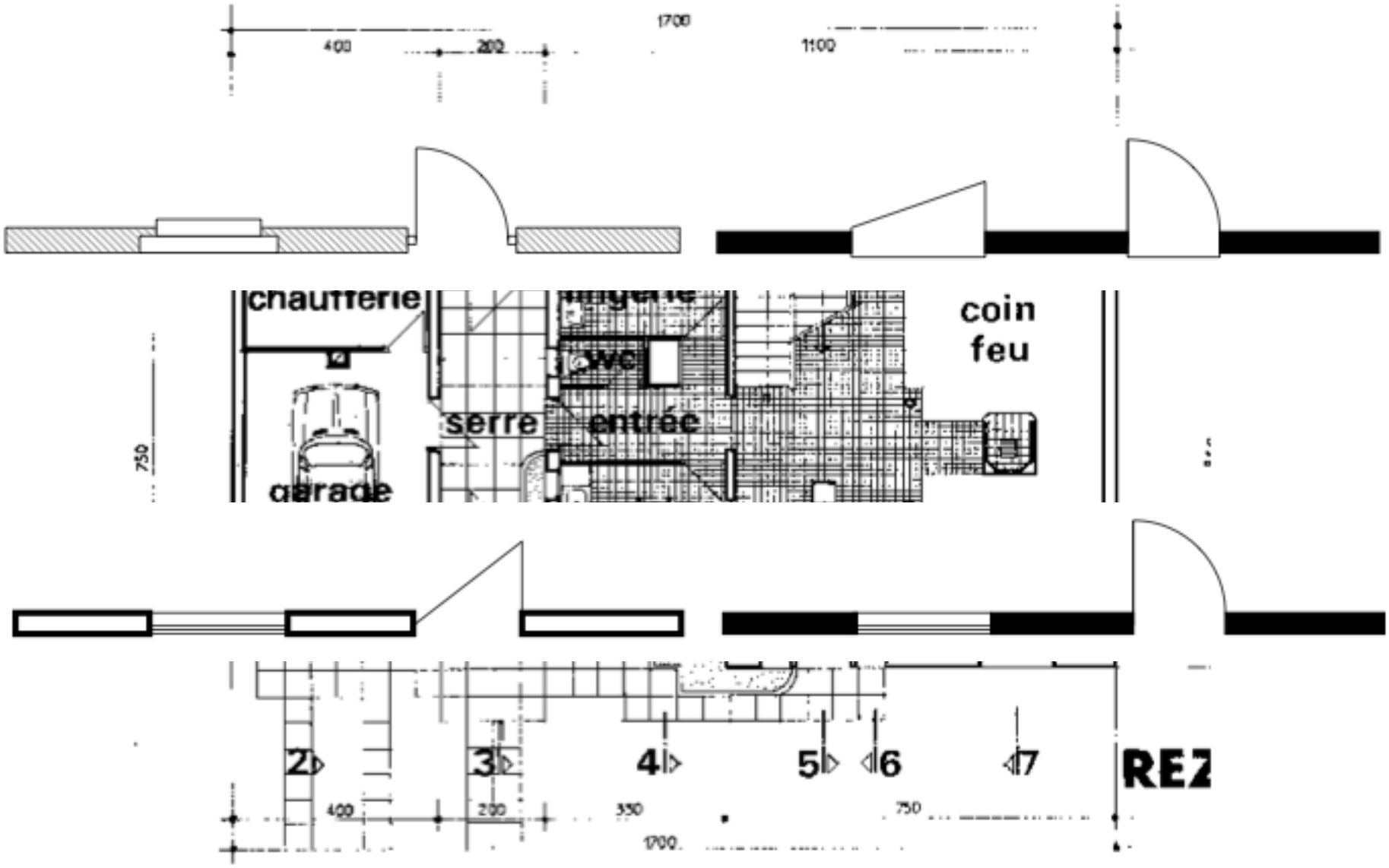


Previous Work

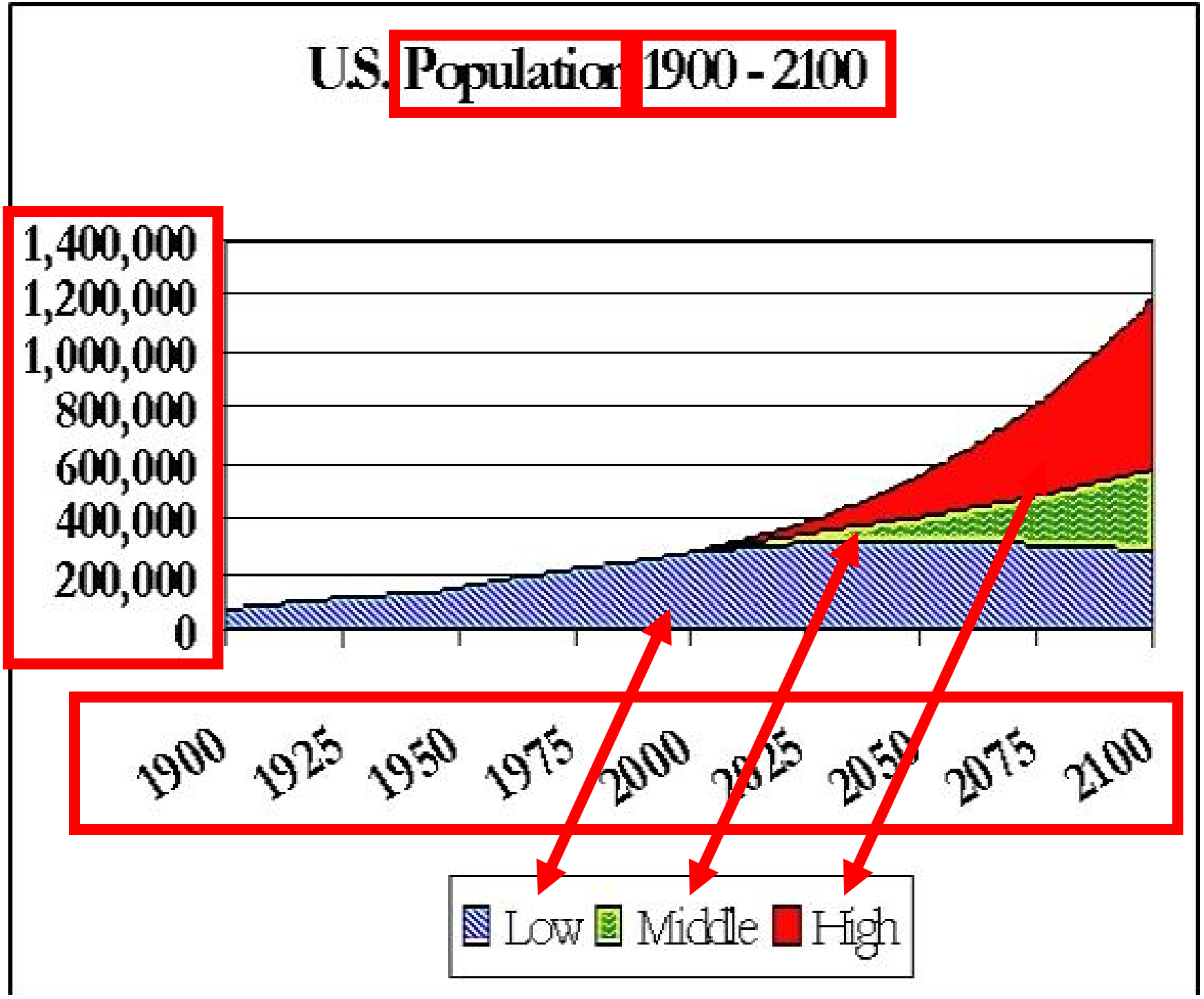
- Searching for diagrams requires some amount of understanding
- Lots of work in image search, most inapplicable to diagrams
- But even understanding diagrams isn't new



Telephone System Manhole Drawing, from
Arias, *et al*, Pattern Recognition Letters 16, 1995

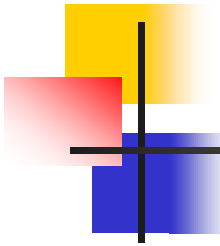


Sample Architectural Drawing, from Ah-Soon and Tombre,
Proc'ds of Fourth Int'l Conf on Document Analysis and Recognition, 1997.


U.S. Population 1900 - 2100
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
1900
1925
1950
1975
2000
2025
2050
2075
2100
Low
Middle
High

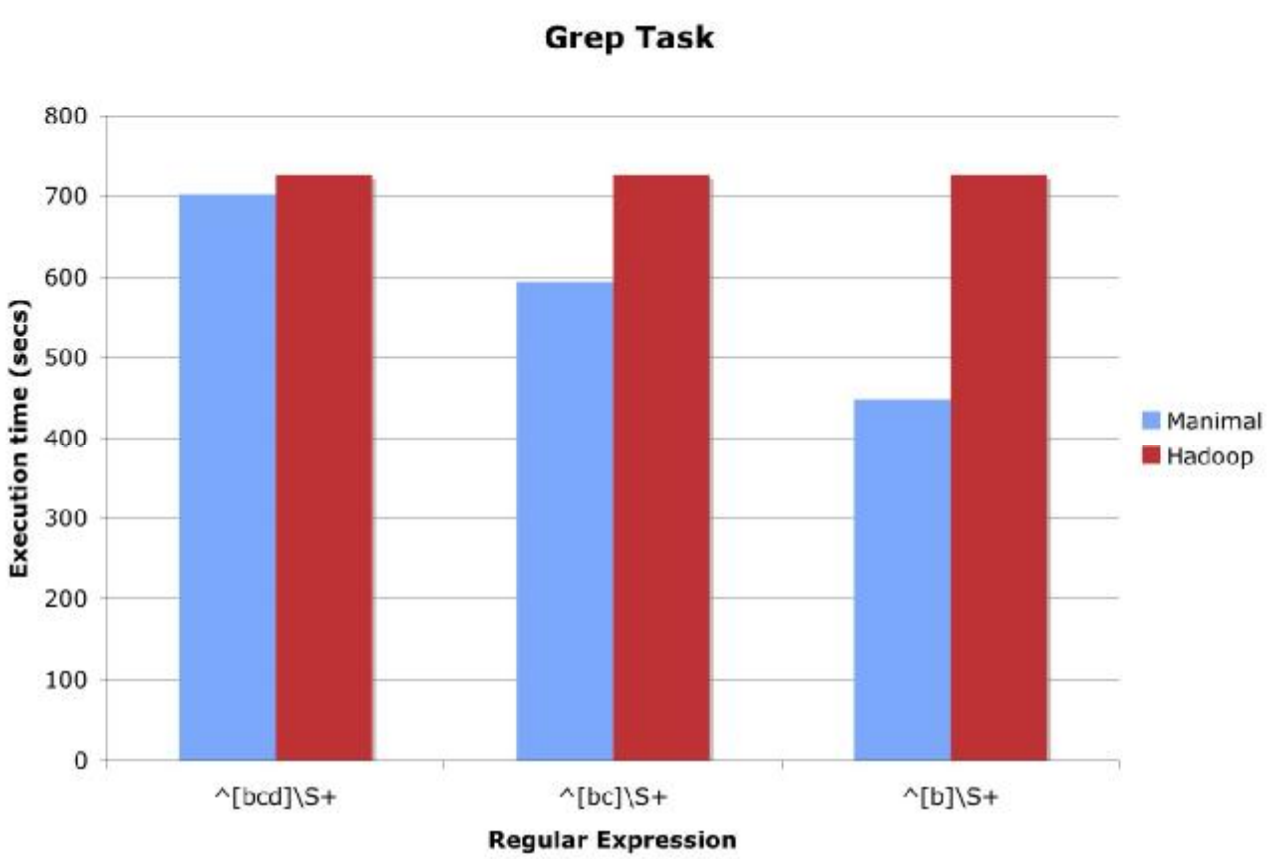
Previous Work

- Understanding diagrams isn't new
- Understanding a *Web's worth* of diagrams is new
  - Need to search statistical diagrams in medicine, economics, biology, physics, *etc*
- The phone company can afford a system tailored for manhole diagrams, but we can't
- Effective scaling with # of topics is central goal of *topic-independent information extraction*



Topic-Independent IE

- Information extraction topic since early 1990s
- Goal is to obtain structured information from unstructured raw documents
  - [Title, Price] from online bookstores
  - [Director , Film] from discussion boards
  - [Scientist, Birthday] from biographies
- Traditional solutions require topic-specific code, features, data
- Costs of TI IE do not grow with # topics


Grep Task
800
700
600
500
400
300
200
100
0
Execution time (secs)
^[bcd]\S+
^[bc]\S+
^[b]\S+
Regular Expression
Manimal
Hadoop

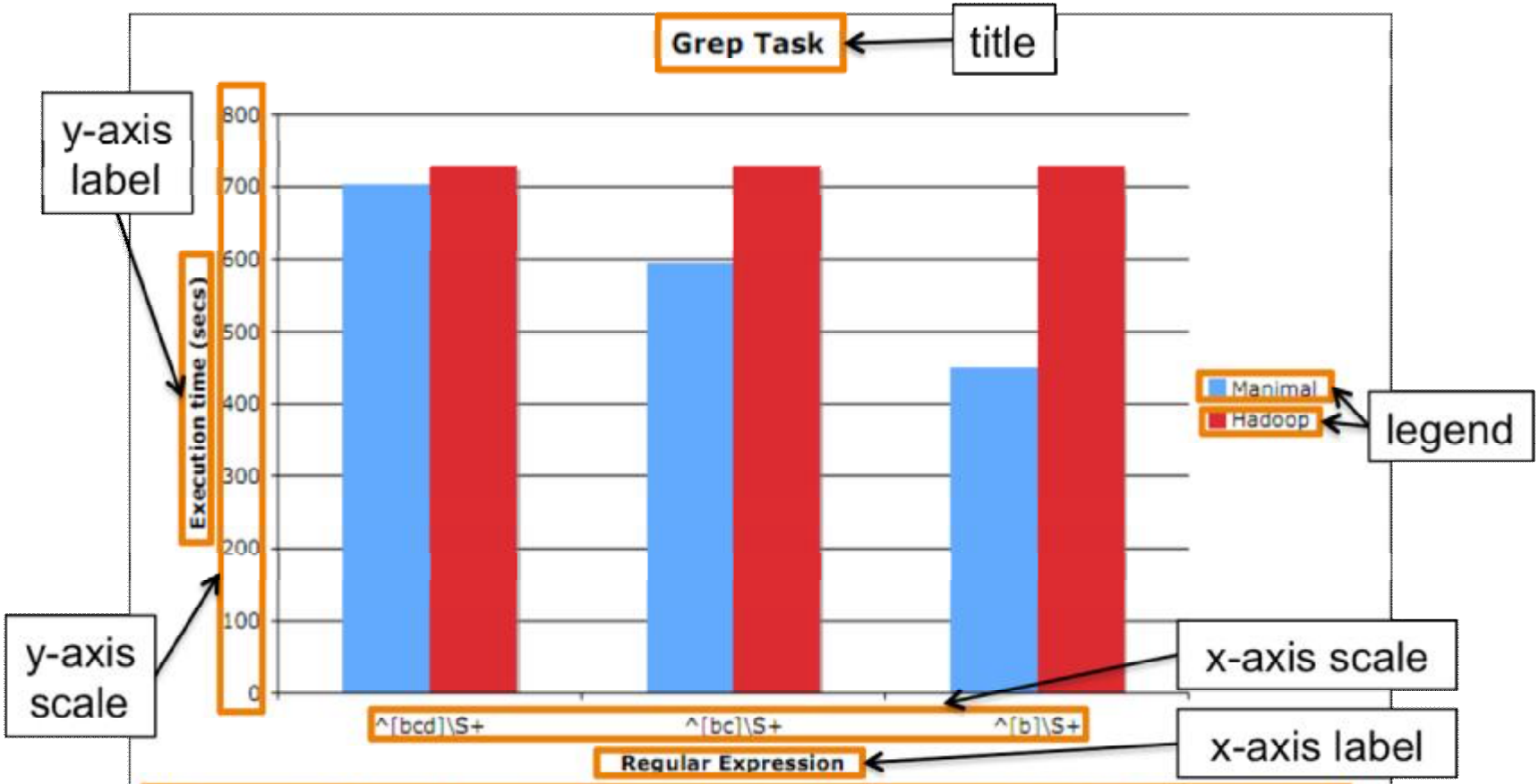
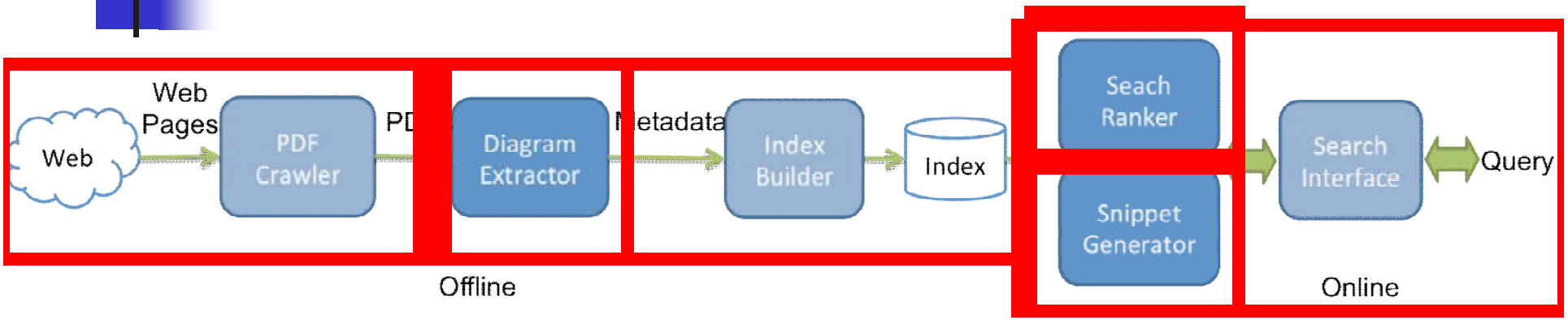


Figure 2 shows the results. As expected, Hadoop's conventional MapReduce execution time is almost wholly insensitive to the selectivity of the program. MANIMAL execution time, in contrast, decreases as the conditional test becomes more restrictive, dropping to 63% of Hadoop in the case of ^[b]\S+. There is nothing about the MANIMAL approach

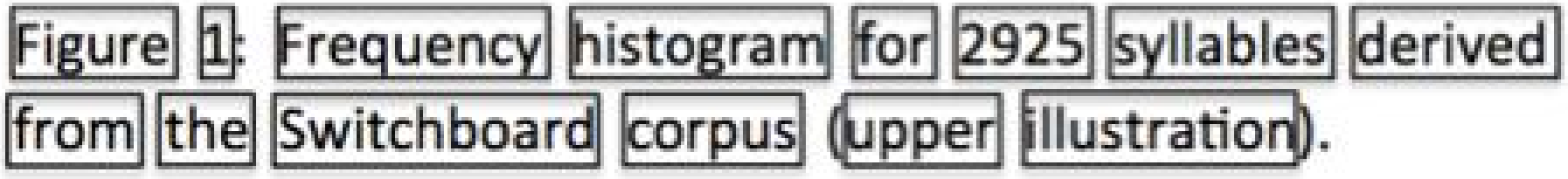
Our Approach



- Typical Web search pipeline
  - Crawl Web for documents
  - Obtain and index text
  - Make index queryable
- Our novel components
  - Diagram metadata extraction
  - Custom search ranker
  - Snippet generator

Metadata Extraction

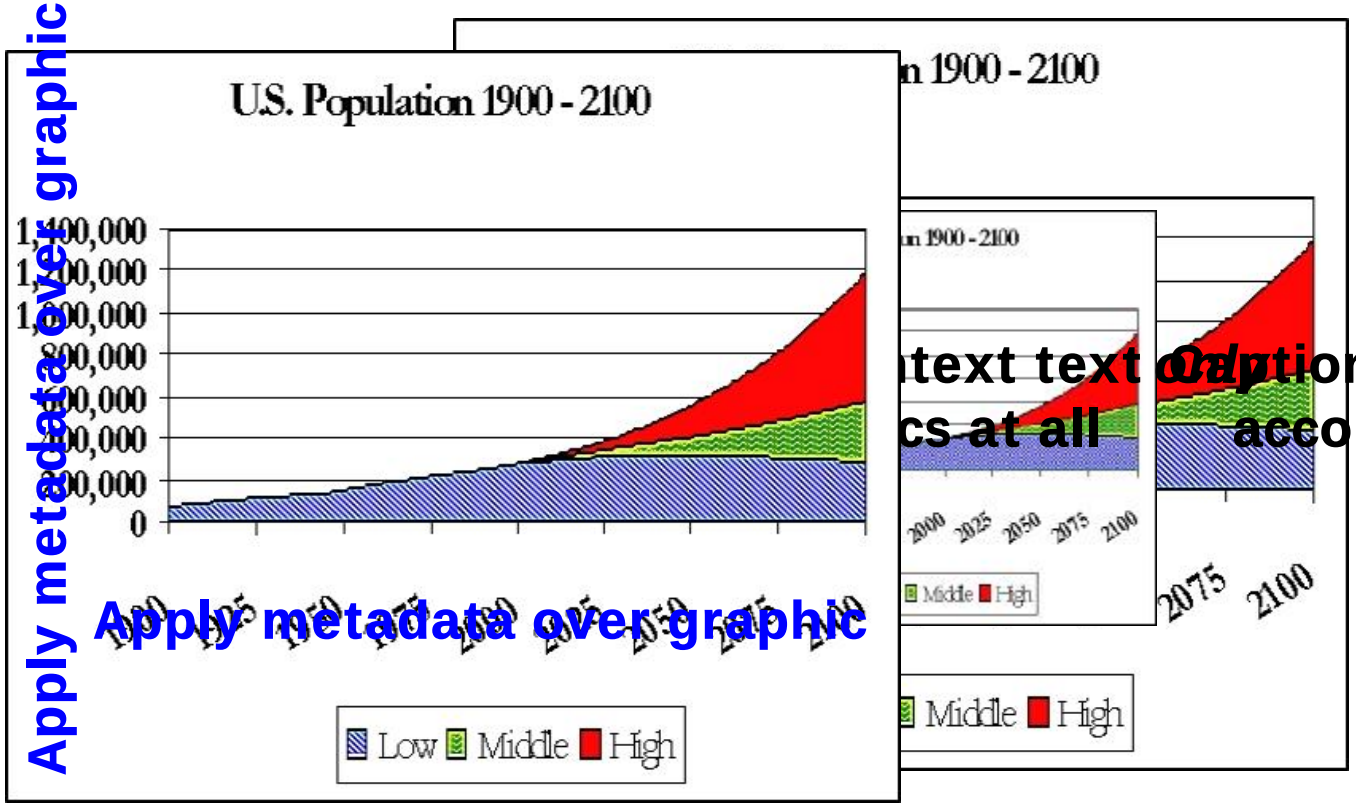
1. Recover good (text, x, y) from PDFs
2. Apply simple role label: *title*, *legend*, etc



3. Group texts into “model diagram” candidates, throw away unlikely ones
   - E.g., must include *something* on x scale
4. Relabel text using geometric relationships
   - Distance, angle to diagram’s origin?
   - Leftmost in diagram? Under a caption?

Snippet Generation

- Tested five versions



Apply metadata over graphic

Apply metadata over graphic

text text Caption and Context text
cs at all accompany graphics

1. [illegible]

Experiments

- Crawled Web for scientific papers
  - From ClueWeb09
  - Any URL ending in **.pdf** from **.edu** URL
  - 319K diagrams
- Fed data to prototype search engine
- Evaluated
  - Metadata extraction
  - Rank quality
  - Snippet effectiveness
- All results compared against human judgments

DiagramFlyer

population Search

Human Population Billions

compared with ~1 Tg only 50 years ago (The Fertilizer Institute, 2000; International Fertilizer Industry Association (IFA), 2004).

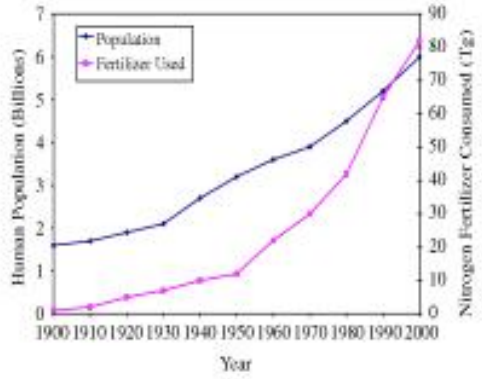


Fig. 1. Graph showing population increase and use of nitrogen fertilizer from 1900 to 2000.

While NH from agricul tural air poll concern, inc [e.g., nitrog nitrous oxid (VOCs) (e.g. organic aci particulates particle conv (e.g., hydrog emissions o practices inc tions (CAFI manure and biomass bun

In many a managed cro parallelly to for food. In

Year

Tags: fertilizer, century, population, billions, yr,

Caption: Fig 1 Graph showing **population** increase and use of nitrogen fertilizer from 1900 to 2000

X-axis Label: Year

Y-axis Label: Human **population** Billions Nitro g en Fertilizer Consumed T g

Legend: population Fertilizer Used

Title:

Context: Fig 1 shows the parallel increase in human **population** and fertilizer usage over the past century Currently the global production of fertilizer is more than 80 Tg of N yr 1 compared with 1 Tg only 50 y

uired Sample

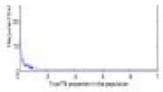


Figure 1: Minimal sample size for estimating within relative error bounds.

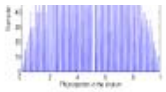


Figure 2: Minimal sample size for estimating within absolute error bounds.

Examp data with By discov strata, wh On the o estimate, be obtain

Y-axis Label: Required Sample Size [Percent]

Legend: Relative Error of 20 * **population** 100 000

Title: Sample size required to estimate the FN proportion in the **population**

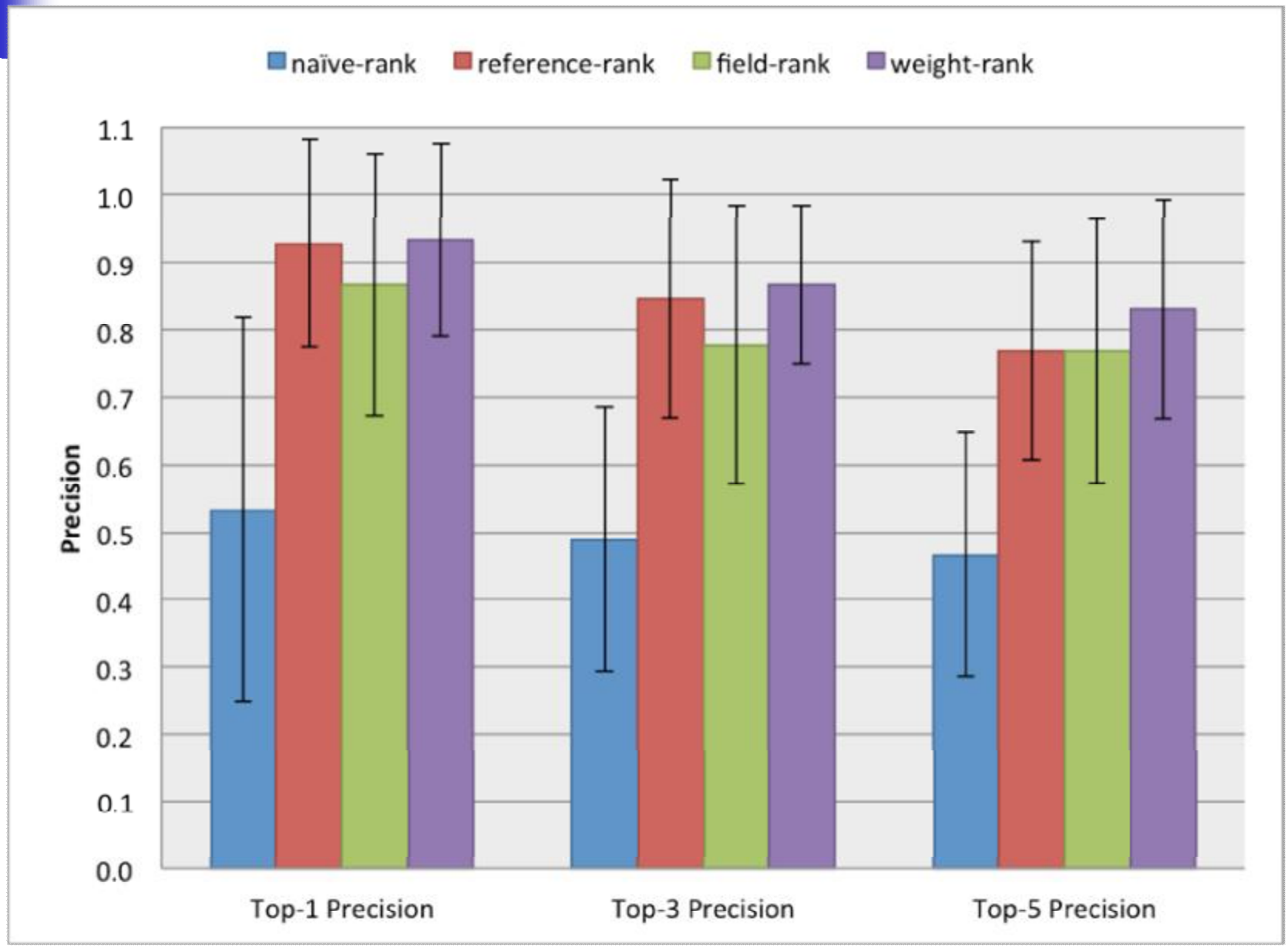
1. Experiments - Extraction

	Recall			**Precision**		
	Text	*All*	*Full*	*Text*	*All*	*Full*
title	0.256	0.651	0.674	0.344	0.609	0.617
Y-scale	0.782	0.796	0.754	0.899	0.843	0.900
Y-label	0.835	0.864	0.874	0.775	0.752	0.797
X-scale	0.903	0.835	0.835	0.616	0.915	0.896
X-label	0.241	0.681	0.681	0.340	0.842	0.835
legend	0.520	0.623	0.656	0.349	0.615	0.631
caption	0.952	0.887	0.839	0.450	0.887	0.929
nondiag	0.768	0.924	0.313	0.850	0.909	0.838

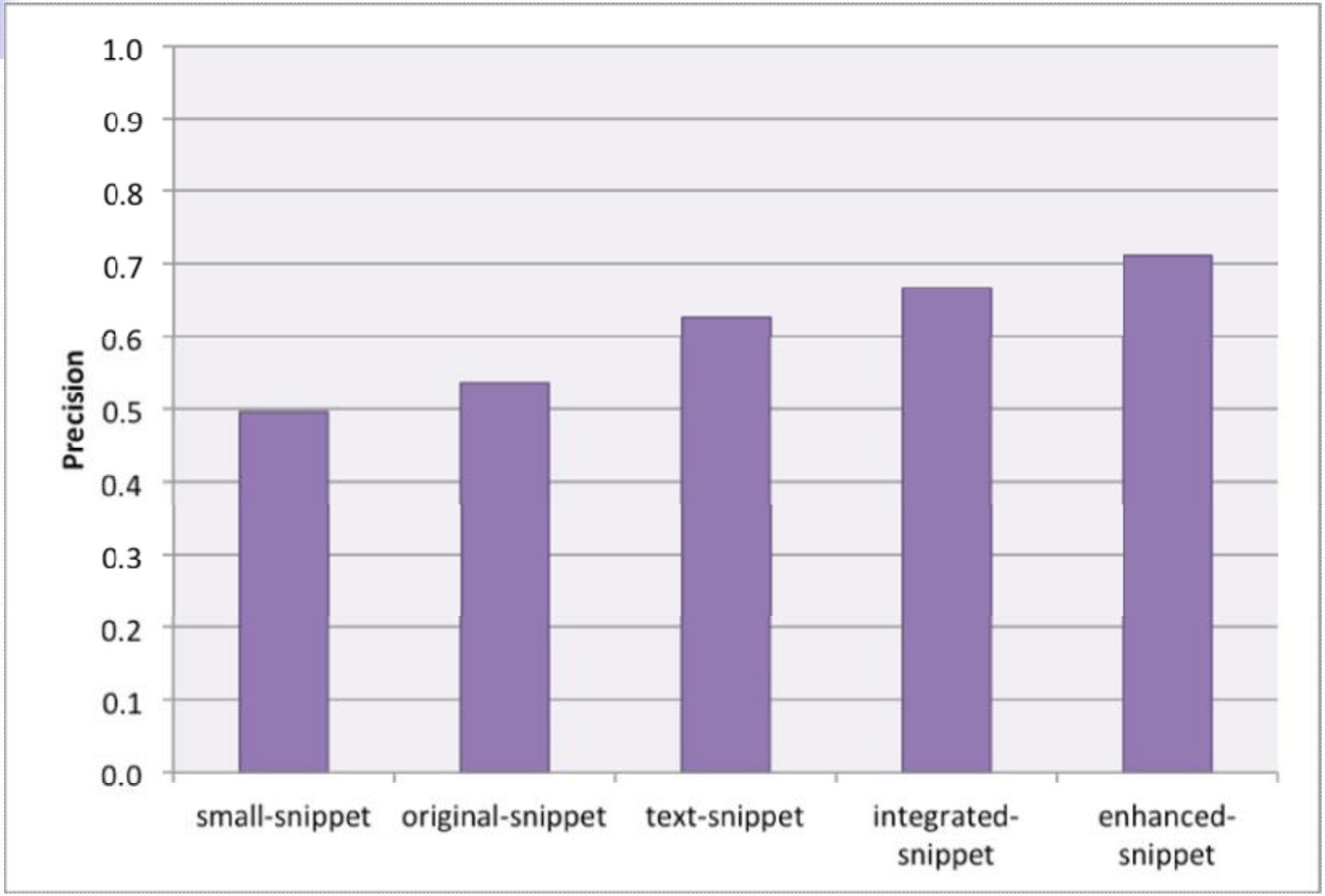
1. Experiments - Extraction

	Recall			**Precision**		
	Text	*All*	*Full*	*Text*	*All*	*Full*
title	0.256	0.651	**0.674**	0.344	0.609	**0.617**
Y-scale	0.782	**0.796**	0.754	0.899	0.843	**0.900**
Y-label	0.835	0.864	**0.874**	0.775	0.752	**0.797**
X-scale	**0.903**	0.835	0.835	0.616	**0.915**	0.896
X-label	0.241	**0.681**	**0.681**	0.340	**0.842**	0.835
legend	0.520	0.623	**0.656**	0.349	0.615	**0.631**
caption	0.952	**0.887**	0.839	0.450	0.887	**0.929**
nondiag	0.768	**0.924**	0.313	0.850	**0.909**	0.838

2. Experiments - Ranking

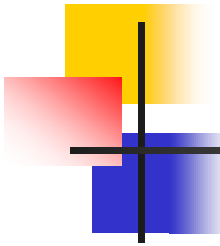


3. Experiments - Snippets



Other Applications

- Working now
  - Search by axis label
  - Search by range
  - Given a query diagram (or paper), find related papers
- In future:
  - Improved academic paper search
  - Show plots that support my hypothesis



Future Work

- Spreadsheets
  - Has experiment X ever been run before?
  - WY GDP vs coal production in 2002
  - Preemptively compute good diagrams
- Deeper questions for messy data
  - HTML tables, data files, spreadsheets
  - Lots of structured data lives outside DBMS
- Structured search

Conclusions

- Metadata extraction enables *52%* better search ranking
- Extraction-enhanced snippets allow users to choose *33% more accurately*
- We rely on open information extraction, but extracted data not the main product
  - Can be successful even with imperfect extractors

Thanks

- Academy of Engineering
- FOE sponsors
- Google
- You!

Related Work

- Suitable for Web search settings
  - Huang *et al*, "Associating text and graphics…", ICDAR 2005
  - Huang *et al*, "Model-based chart image recognition", GREC 2003
  - Kaiser *et al*, "Automatic extraction…", AAAI 2008
  - Liu *et al*, "Automated analysis…", IJDAR 2009
- Diagram parsing
  - E.g., Futrelle, "Summarization…", 1999
- Visually-impaired access
  - E.g., Demir *et al*, "Generating textual…", INLG 2008.