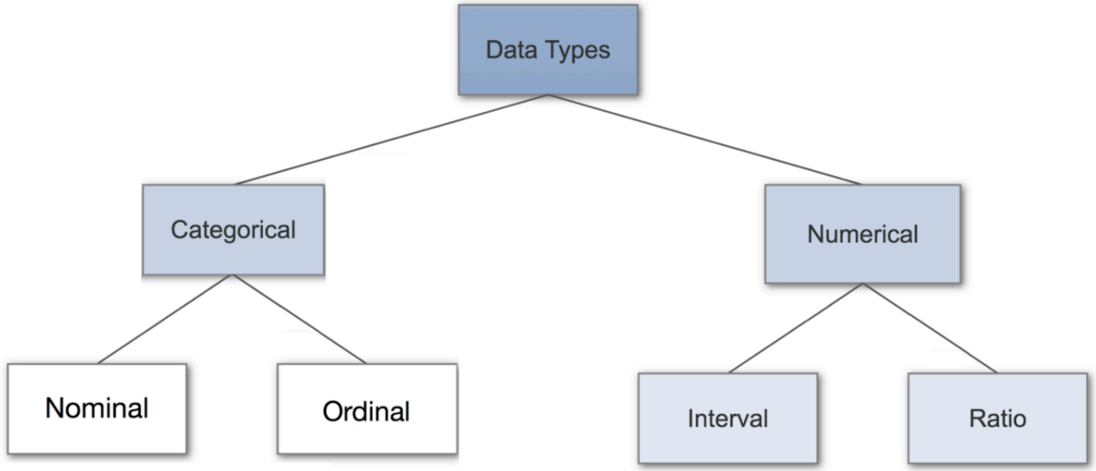
ISE 599
Special Topics Applied Predictive Analytics

Mayank Kejriwal
Research Assistant Professor/Research Lead
Department of Industrial and Systems Engineering
Information Sciences Institute
USC Viterbi School of Engineering
kejriwal@isi.edu

Data: how we 'make sense' of it (them?)

- First, important to understand the different types of data
  - An 'ontology' of data types
- We've already (kind of) seen one example!



In practice, such formal definitions are rarely useful (except in labs)

- Article in Forbes shows 13 different types of data
- It’s really broad, and some of the categories overlap, but useful as a framework

1 - Big data

2 - Structured, unstructured, semi-structured data

3 - Time-stamped data

4 - Machine data

5 - Spatiotemporal data

6 - Open data

7 - Dark data

- Real time data

9 - Genomics data

10 - Operational data

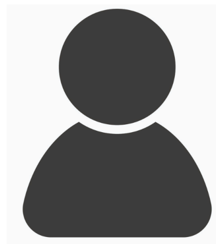
11 - High-dimensional data

12 - Unverified outdated data

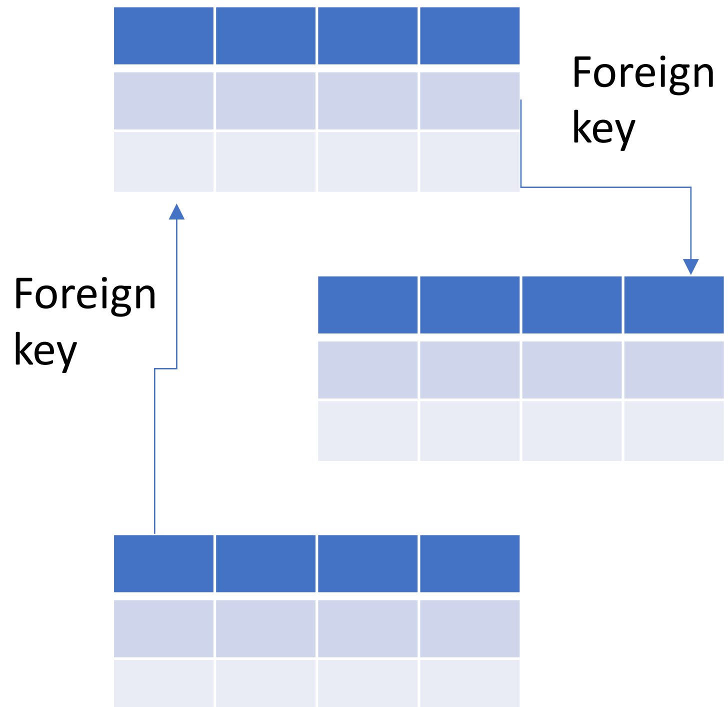
13 - Translytic Data

Structured vs. ‘unstructured’ data

Two ‘extremes’?



Natural language, social media data



Is data ever *really* unstructured?

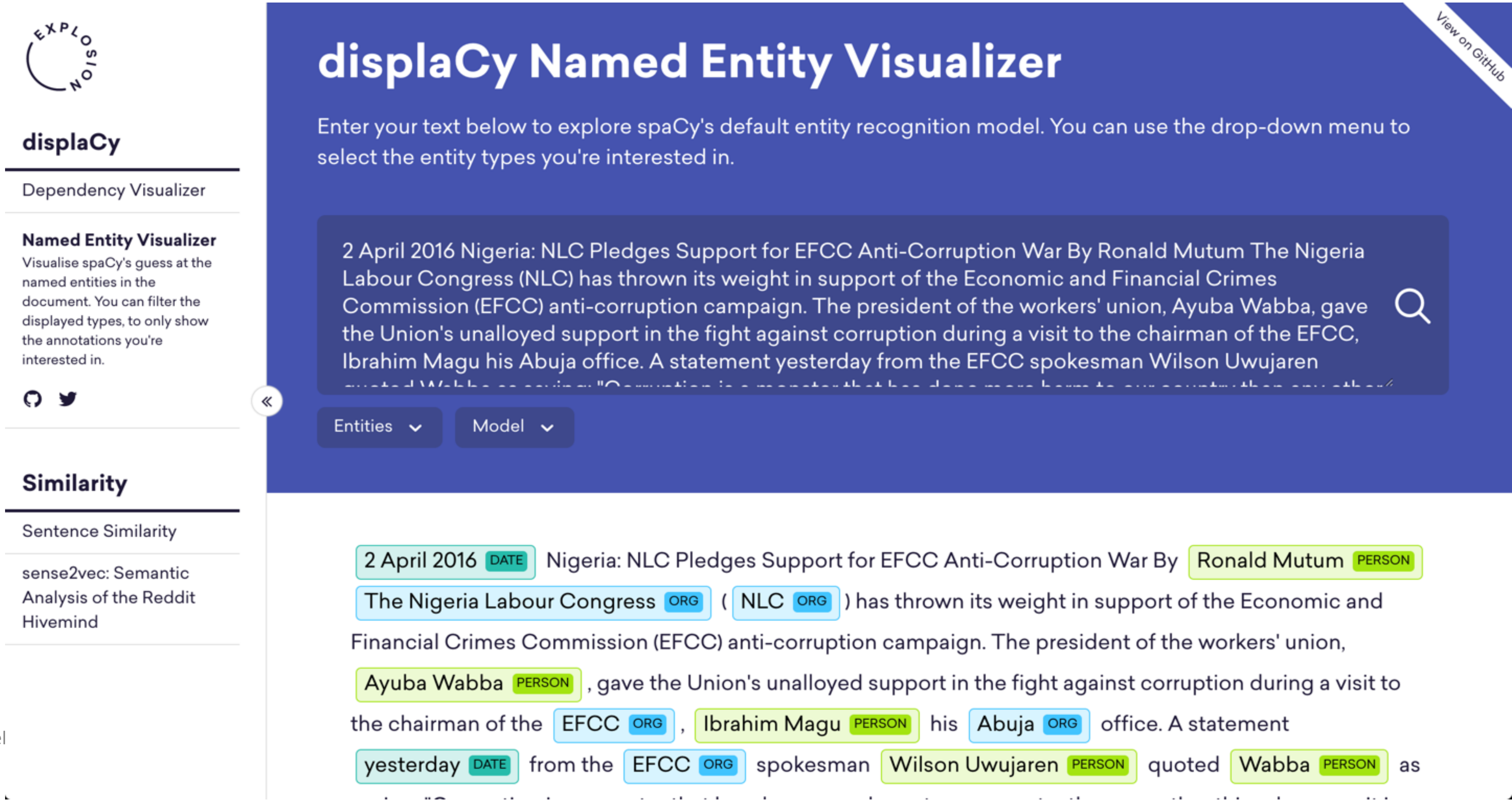
- Many computer scientists would call English ‘unstructured’
  - You can substitute English for any ‘natural’ language spoken by humans in society
- The great philosopher Gottlieb Frege, like so many others, felt English was woefully imprecise
- Thought of logic as one way to address these difficulties
- Never panned out in the AI community, too many irregularities in English and other languages!

Structure (beauty) is in the eye of the application (beholder)

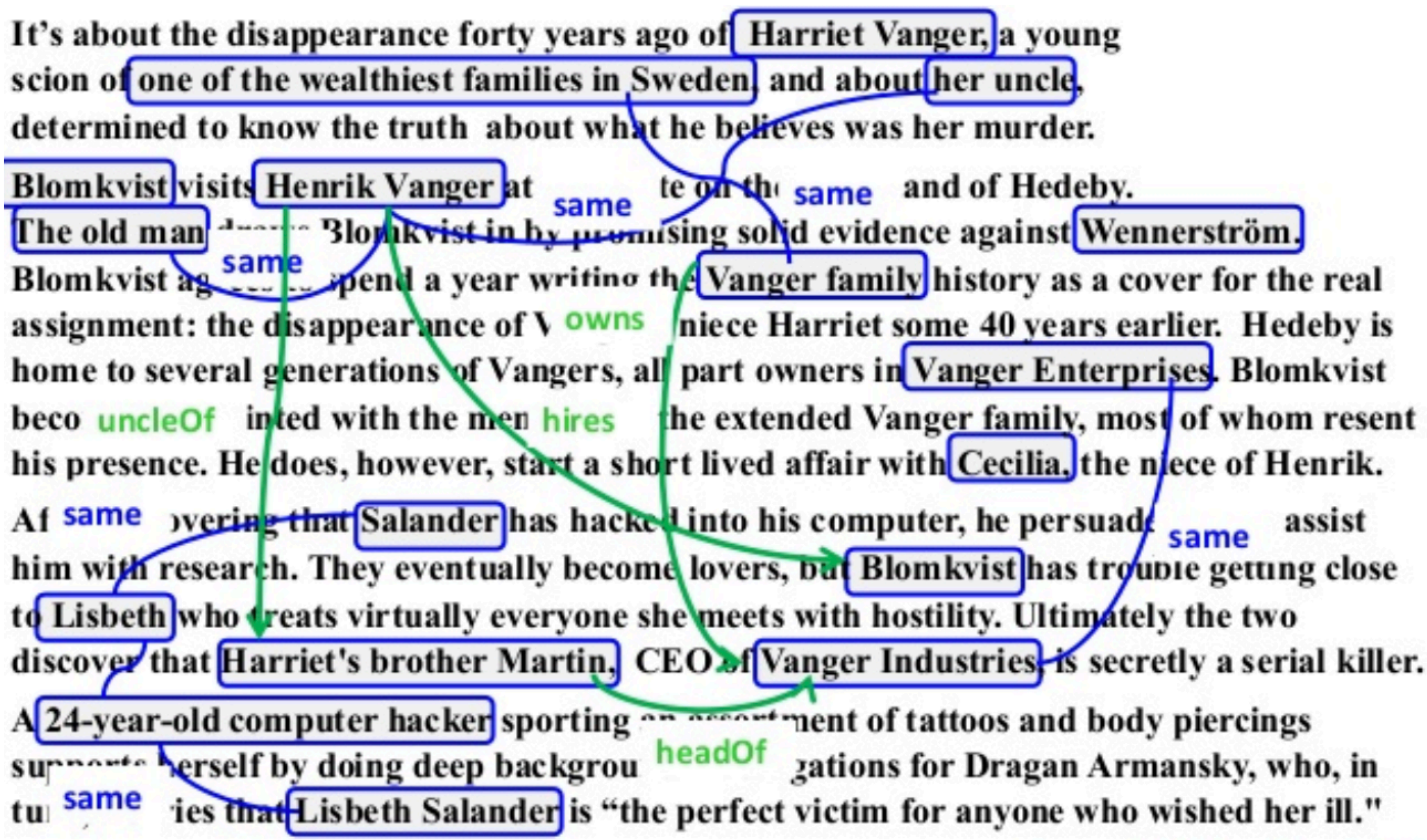
- Unfortunately, ‘natural language’ data is often called unstructured data by many practitioners
  - I encourage the phrase ‘natural language’ vs. unstructured, since it has an impact on how you think about the data

Demo: spacy

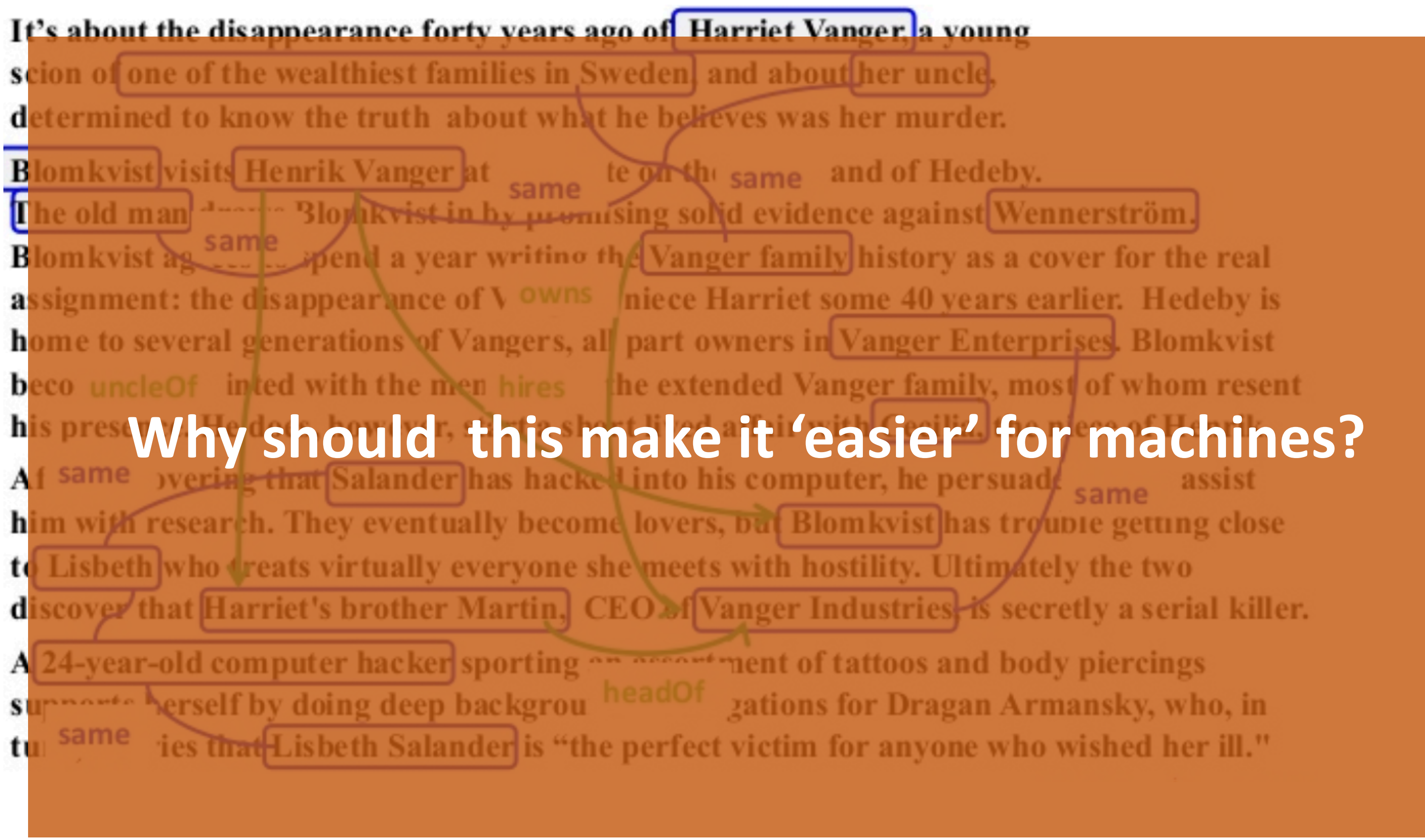
https://demos.explosion.ai/displacy-ent



In practice, IE is usually more complex



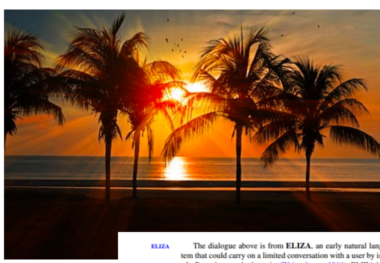
In practice, IE is usually more complex

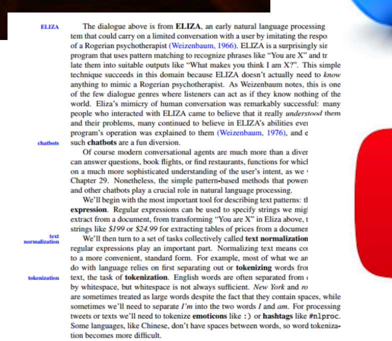


Why should this make it 'easier' for machines?

Machines like certain kinds of structure and technologies like IE can ‘parse’ that structure from human-centric structure

Humans like…





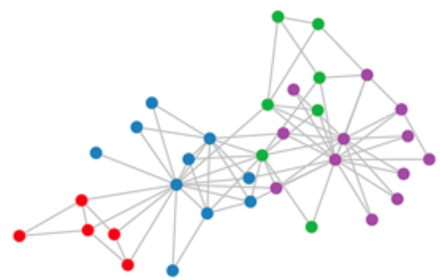
ELIZA The dialogue above is from **ELIZA**, an early natural language processing
tem that could carry on a limited conversation with a user by imitating the respo
of a Rogerian psychotherapist (Weizenbaum, 1966). ELIZA is a surprisingly sir
program that uses pattern matching to recognize phrases like “You are X” and tr
late them into suitable outputs like “What makes you think I am X?”. This simple
technique succeeds in this domain because ELIZA doesn’t actually need to *know*
anything to mimic a Rogerian psychotherapist. As Weizenbaum notes, this is one
of the few dialogue genres where listeners can act as if they know nothing of the
world. Eliza’s mimicry of human conversation was remarkably successful: many
people who interacted with ELIZA came to believe that it really *understood* them
and their problems, many continued to believe in ELIZA’s abilities eve
program’s operation was explained to them (Weizenbaum, 1976), and e
chatbots such **chatbots** are a fun diversion.

Of course modern conversational agents are much more than a diver
can answer questions, book flights, or find restaurants, functions for whicl
on a much more sophisticated understanding of the user’s intent, as we
Chapter 29. Nonetheless, the simple pattern-based methods that power
and other chatbots play a crucial role in natural language processing.

We’ll begin with the most important tool for describing text patterns: tl
expression. Regular expressions can be used to specify strings we mig
extract from a document, from transforming “You are X” in Eliza above, t
strings like *$199* or *$24.99* for extracting tables of prices from a documen

text normalization We’ll then turn to a set of tasks collectively called **text normalization**
regular expressions play an important part. Normalizing text means co
to a more convenient, standard form. For example, most of what we ar
do with language relies on first separating out or **tokenizing** words fro
tokenization text, the task of **tokenization**. English words are often separated from
by whitespace, but whitespace is not always sufficient. *New York* and *ro*
are sometimes treated as large words despite the fact that they contain spaces, while
sometimes we’ll need to separate *I’m* into the two words *I* and *am*. For processing
tweets or texts we’ll need to tokenize **emoticons** like `:)` or **hashtags** like `#nlproc`.
Some languages, like Chinese, don’t have spaces between words, so word tokeniza-
tion becomes more difficult.

Machines like…

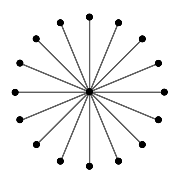


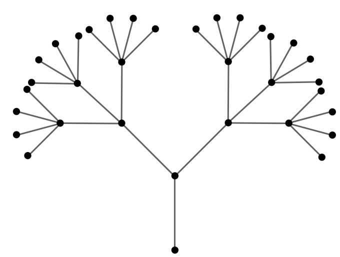
XML

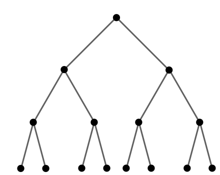

```
<empinfo>
  <employees>
    <employee>
      <name>James Kirk</name>
      <age>40></age>
    </employee>
    <employee>
      <name>Jean-Luc Picard</name>
      <age>45</age>
    </employee>
    <employee>
      <name>Wesley Crusher</name>
      <age>27</age>
    </employee>
  </employees>
</empinfo>
```


JSON


```
{  "empinfo" :
     {
       "employees" : [
         {
           "name" : "James Kirk",
           "age" : 40,
         },
         {
           "name" : "Jean-Luc Picard",
           "age" : 45,
         },
         {
           "name" : "Wesley Crusher",
           "age" : 27,
         }
                     ]
     }
}
```

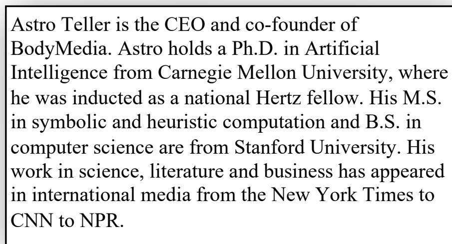







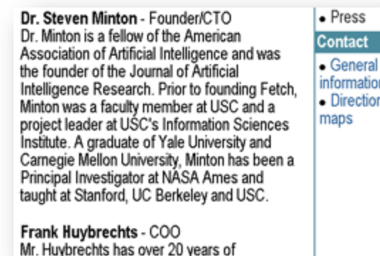
Many other kinds of ‘structure’ out there…

Text paragraphs without formatting



Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences plus some formatting & links



Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

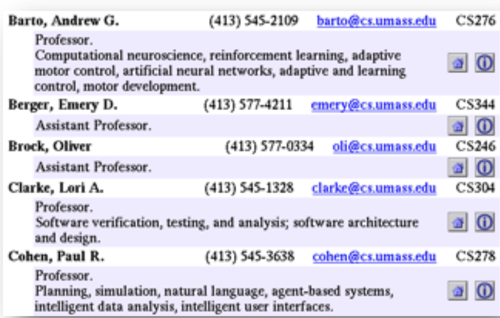
Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links



Barto, Andrew G. (413) 545-2109 barto@cs.umass.edu CS276
Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

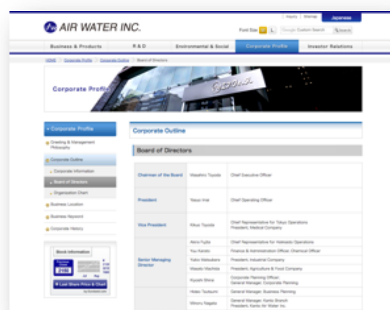
Berger, Emery D. (413) 577-4211 emery@cs.umass.edu CS344
Assistant Professor.

Brock, Oliver (413) 577-0334 oli@cs.umass.edu CS246
Assistant Professor.

Clarke, Lori A. (413) 545-1328 clarke@cs.umass.edu CS304
Professor.
Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

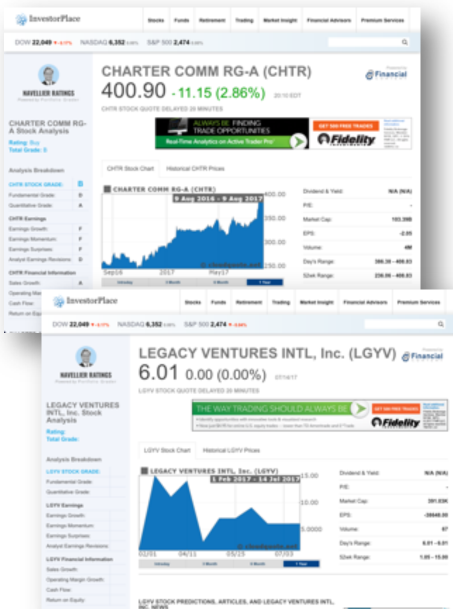
Tables



Charts

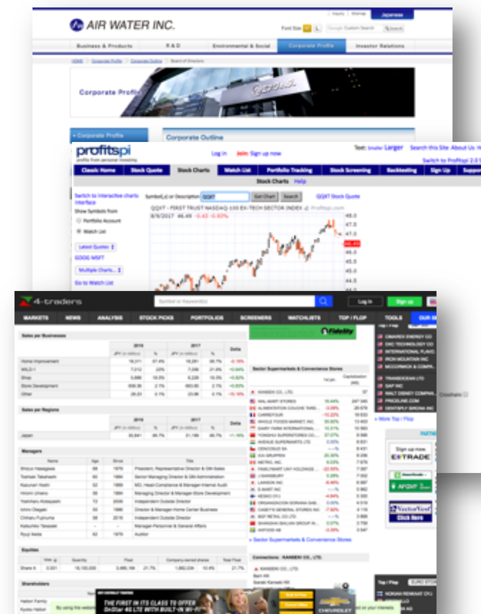
Structure (and IEs) can depend on **scope**

Web site specific



Genre specific (e.g., forums)

Wide, non-specific



Structure (and IEs) can depend on **pattern complexity**

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.