

ISE 599

Special Topics Applied Predictive Analytics

Mayank Kejriwal

Research Assistant Professor/Research Lead

Department of Industrial and Systems Engineering

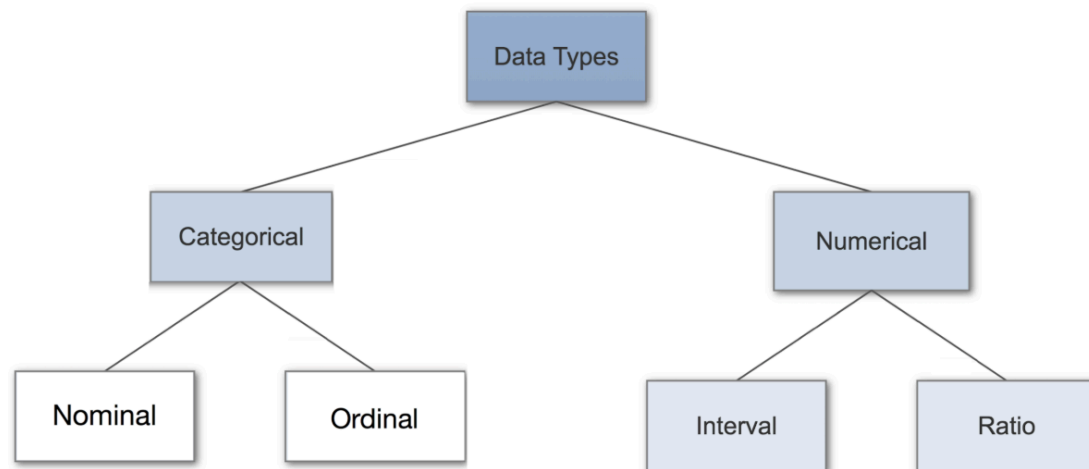
Information Sciences Institute

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

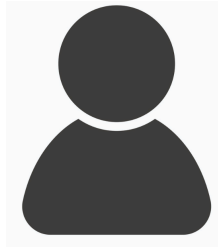
11 - High-dimensional data

12 - Unverified outdated data

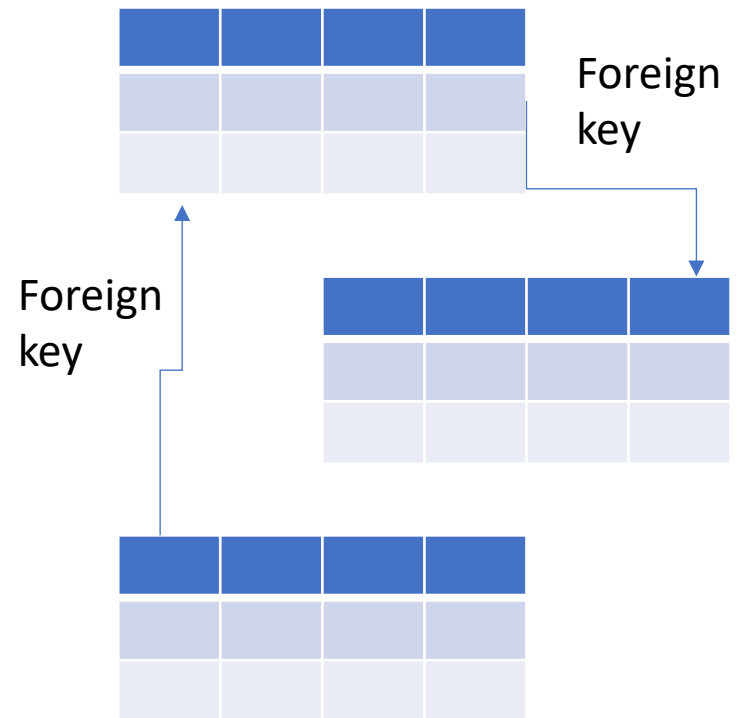
13 - Translytic Data

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>



displaCy

Dependency Visualizer

Named Entity Visualizer

Visualise spaCy's guess at the named entities in the document. You can filter the displayed types, to only show the annotations you're interested in.



Similarity

Sentence Similarity

sense2vec: Semantic Analysis of the Reddit Hivemind

displaCy Named Entity Visualizer

Enter your text below to explore spaCy's default entity recognition model. You can use the drop-down menu to select the entity types you're interested in.

View on GitHub

2 April 2016 Nigeria: NLC Pledges Support for EFCC Anti-Corruption War By Ronald Mutum The Nigeria Labour Congress (NLC) has thrown its weight in support of the Economic and Financial Crimes Commission (EFCC) anti-corruption campaign. The president of the workers' union, Ayuba Wabba, gave the Union's unalloyed support in the fight against corruption during a visit to the chairman of the EFCC, Ibrahim Magu his Abuja office. A statement yesterday from the EFCC spokesman Wilson Uwujaren quoted Wabba as saying "Corruption is a monster that has done more harm to our country than any other"



Entities ▾

Model ▾

2 April 2016 DATE Nigeria: NLC Pledges Support for EFCC Anti-Corruption War By Ronald Mutum PERSON
The Nigeria Labour Congress ORG (NLC ORG) has thrown its weight in support of the Economic and
Financial Crimes Commission (EFCC) anti-corruption campaign. The president of the workers' union,
Ayuba Wabba PERSON , gave the Union's unalloyed support in the fight against corruption during a visit to
the chairman of the EFCC ORG , Ibrahim Magu PERSON his Abuja ORG office. A statement
yesterday DATE from the EFCC ORG spokesman Wilson Uwujaren PERSON quoted Wabba PERSON as

In practice, IE is usually more complex

It's about the disappearance forty years ago of **Harriet Vanger**, a young scion of **one of the wealthiest families in Sweden**, and about **her uncle**, determined to know the truth about what he believes was her murder.

Blomkvist visits **Henrik Vanger** at **same** te on the **same** and of Hedeby.

The old man **Blomkvist** in by promising solid evidence against **Wennerström**.

Blomkvist ag **same** spend a year writing the **Vanger family** history as a cover for the real assignment: the disappearance of **V owns** niece **Harriet** some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in **Vanger Enterprises**. **Blomkvist** beco **uncleOf** inted with the men **hires** the extended **Vanger family**, most of whom resent his presence. He does, however, start a short lived affair with **Cecilia**, the niece of **Henrik**.

At **same** overing that **Salander** has hacked into his computer, he persuades **same** assist him with research. They eventually become lovers, but **Blomkvist** has trouble getting close to **Lisbeth** who treats virtually everyone she meets with hostility. Ultimately the two discover that **Harriet's brother Martin**, CEO of **Vanger Industries**, is secretly a serial killer.

A **24-year-old computer hacker** sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou **headOf** gations for **Dragan Armansky**, who, in tu **same** ies that **Lisbeth Salander** is "the perfect victim for anyone who wished her ill."

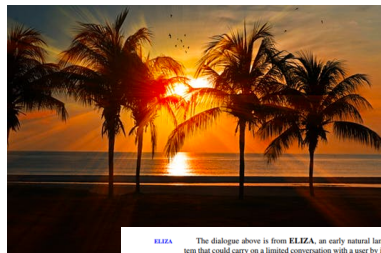
In practice, IE is usually more complex

It's about the disappearance forty years ago of **Harriet Vanger**, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder. **Blomkvist** visits **Henrik Vanger** at the same time on the same island of Hedeby. The old man convinces **Blomkvist** in by promising solid evidence against **Wennerström**. **Blomkvist** agrees to spend a year writing the **Vanger family** history as a cover for the real assignment: the disappearance of **Vanger's** niece **Harriet** some 40 years earlier. Hedeby is home to several generations of **Vangers**, all part owners in **Vanger Enterprises**. **Blomkvist** becomes acquainted with the men who run the extended **Vanger family**, most of whom resent his presence. He does, however, establish a rapport with **Cecilia**, the niece of **Harriet**. After discovering that **Salander** has hacked into his computer, he persuades **Salander** to assist him with research. They eventually become lovers, but **Blomkvist** has trouble getting close to **Lisbeth** who treats virtually everyone she meets with hostility. Ultimately the two discover that **Harriet's brother Martin**, CEO of **Vanger Industries**, is secretly a serial killer. A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for **Dragan Armansky**, who, in turn, convinces **Lisbeth Salander** is "the perfect victim for anyone who wished her ill."

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 system that could carry on a limited conversation with a user by imitating the responses of a Rogerian psychotherapist (**Weizenbaum, 1966**). ELIZA is a surprisingly simple program that uses patterns matching to recognize phrases like "You are X" and it later 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 even when the program's operation was explained to them (**Weizenbaum, 1976**), and such **chatbots** are a fun diversion.

chatbot Of course modern conversational agents are much more than a diversion can answer questions, book flights, or find restaurants, functions for which 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.

text normalization We'll begin with the most important tool for describing text patterns: **expression**. Regular expressions can be used to specify strings we might extract from a document, from transforming "You are X" in ELIZA above, to strings like \$199 or \$24.99 for extracting tables of prices from a document.

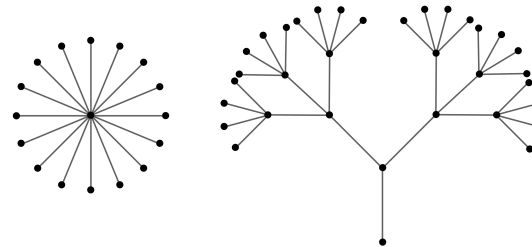
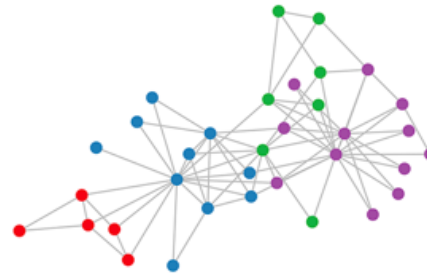
tokenization We'll then turn to a set of tasks collectively called **text normalization**. Regular expressions play an important part. Normalizing text means to to a more convenient, standard form. For example, most of what we do with language relies on first separating out or **tokenizing** words from text, the task of **tokenization**. English words are often separated from by whitespace, but whitespace is not always sufficient. *New York and me* 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 **emojis** like :) or **hashtags** like #Eliza. Some languages, like Chinese, don't have spaces between words, so word tokenization becomes more difficult.



ComputerHope.com



Machines like...



XML

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

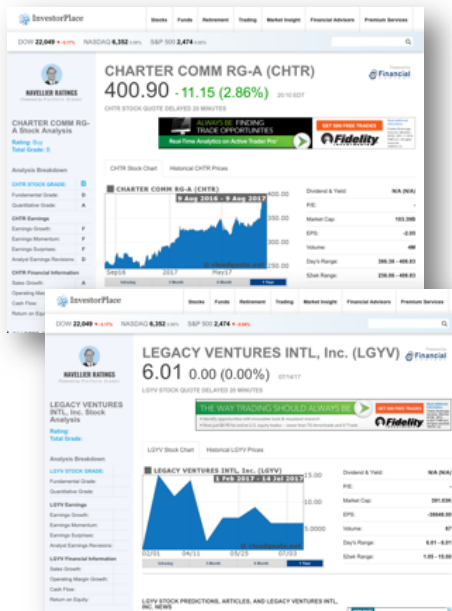
AIR WATER INC.	
Business & Products	Environmental & Social
Corporate Profile	Corporate Overview
Board of Directors	Board of Directors
Chairman of the Board	Chairman of the Board
President	President
Chief Executive Officer	Chief Executive Officer
Chief Financial Officer	Chief Financial Officer
Chief Operating Officer	Chief Operating Officer
Chief Technology Officer	Chief Technology Officer
Chief Marketing Officer	Chief Marketing Officer
Chief Human Resources Officer	Chief Human Resources Officer
Chief Information Officer	Chief Information Officer
Chief Legal Officer	Chief Legal Officer
Chief Compliance Officer	Chief Compliance Officer
Chief Sustainability Officer	Chief Sustainability Officer
Chief Security Officer	Chief Security Officer
Chief Privacy Officer	Chief Privacy Officer
Chief Risk Officer	Chief Risk Officer
Chief Ethics Officer	Chief Ethics Officer
Chief Diversity Officer	Chief Diversity Officer
Chief Inclusion Officer	Chief Inclusion Officer
Chief Accessibility Officer	Chief Accessibility Officer
Chief Digital Officer	Chief Digital Officer
Chief Innovation Officer	Chief Innovation Officer
Chief Future Officer	Chief Future Officer
Chief Impact Officer	Chief Impact Officer
Chief Social Officer	Chief Social Officer
Chief Community Officer	Chief Community Officer
Chief Government Affairs Officer	Chief Government Affairs Officer
Chief Public Affairs Officer	Chief Public Affairs Officer
Chief Communications Officer	Chief Communications Officer
Chief Content Officer	Chief Content Officer
Chief Creative Officer	Chief Creative Officer
Chief Design Officer	Chief Design Officer
Chief Development Officer	Chief Development Officer
Chief Engineering Officer	Chief Engineering Officer
Chief Finance Officer	Chief Finance Officer
Chief General Counsel	Chief General Counsel
Chief Human Resources	Chief Human Resources
Chief Information Technology	Chief Information Technology
Chief Legal	Chief Legal
Chief Marketing	Chief Marketing
Chief Operations	Chief Operations
Chief Product	Chief Product
Chief Project	Chief Project
Chief Quality	Chief Quality
Chief Research & Development	Chief Research & Development
Chief Sales	Chief Sales
Chief Strategy	Chief Strategy
Chief Support	Chief Support
Chief Training	Chief Training
Chief User Experience	Chief User Experience
Chief Vendor Management	Chief Vendor Management
Chief Workplace	Chief Workplace
Chief Writing	Chief Writing

Charts



Structure (and IEs) can depend on scope

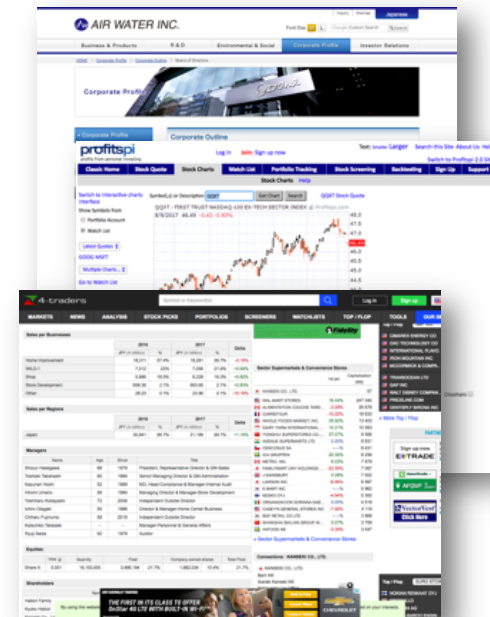
Web site specific



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.