COLLECTING DATA AND WORKING WITH THEM... ☺

CARINA F. DORNELES

CARINA.DORNELES@UFSC.BR

WWW.INF.UFSC.BR/~CARINA.DORNELES

SHORT BIO

- Carina F. Dorneles is a Professor at the Department of Informatics and Statistics (INE) at the Federal University of Santa Catarina (UFSC), Brazil. She received her Master Science Degree in Computer Science (2000), working on a strategy of ontology-based extraction of Semi-Structured Data from the Web, and her Ph.D. Degree in Computer Science (2000), when she worked on a strategy for allowing meaningful and comparable scores in approximate matching, both at the University Rio Grande do Sul, RS, Brazil.

  She has worked as a member of several committees in Brazil, such as the Steering Committee Committee of the Special Database Commission of the Brazilian Computer Society from 2018 to 2021; the Capes Quadrennial Evaluation Committee in 2017 and this year 2022; the Education Committee of the Brazilian Computer Society during 2013-2015. She has also been the Coordinator of Research Support at the Prorectorate of Research at UFSC (PROPESQ /UFSC) from 2012 to 2013; and Coordinator of the Graduate Program in Computer Science at UFSC from 2015 to 2017. Her research interests include Data Engineering tasks and Data Management, Information Retrieval, Mining of Data with an emphasis on the Web, Knowledge Discovery, and Information Extraction and Matching. She coordinates and participates in research projects in the area, and international collaboration projects, including the VIDAS project, with France, within the CAPES / COFECUB program. She also participates as a member of technical committees for conferences and workshops held in Brazil and abroad; works as ad hoc reviewer for funding national agencies such as CNPq, Capes, FAPESC, FAPERGS and FAPESP, as well as CTIC / RNP. She contributes as a reviewer of articles in national and international journals and events.

SCHEDULE

- BigData
- *Web crawling and scraping*
- Data Extraction
- Named Entity Resolution

WHAT DOES “BIG DATA” MEANS?

(1) **Collecting** large amounts of data:

- via computers, sensors, people, events, etc.

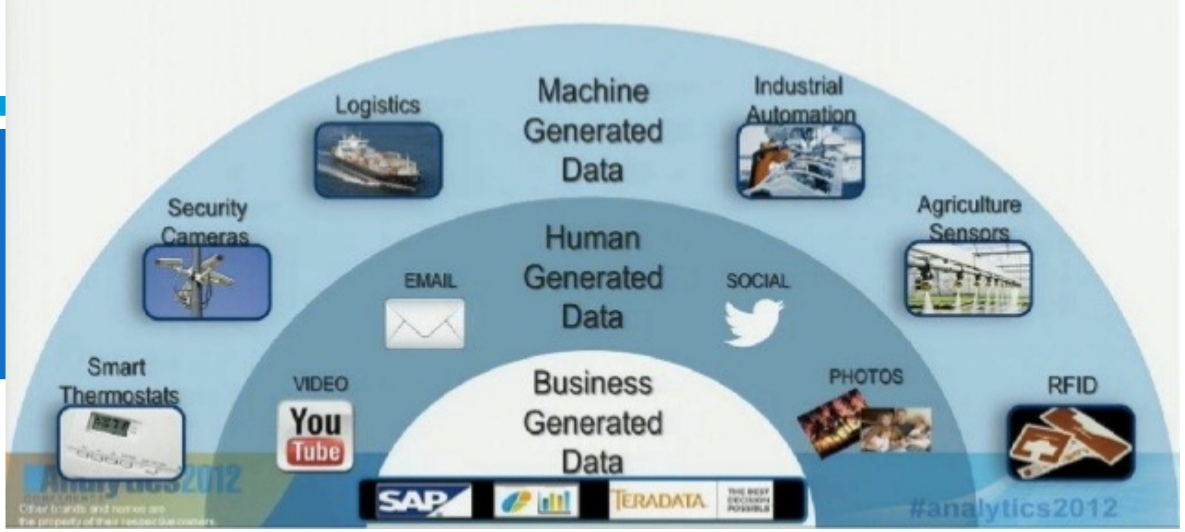
© Can Stock Photo

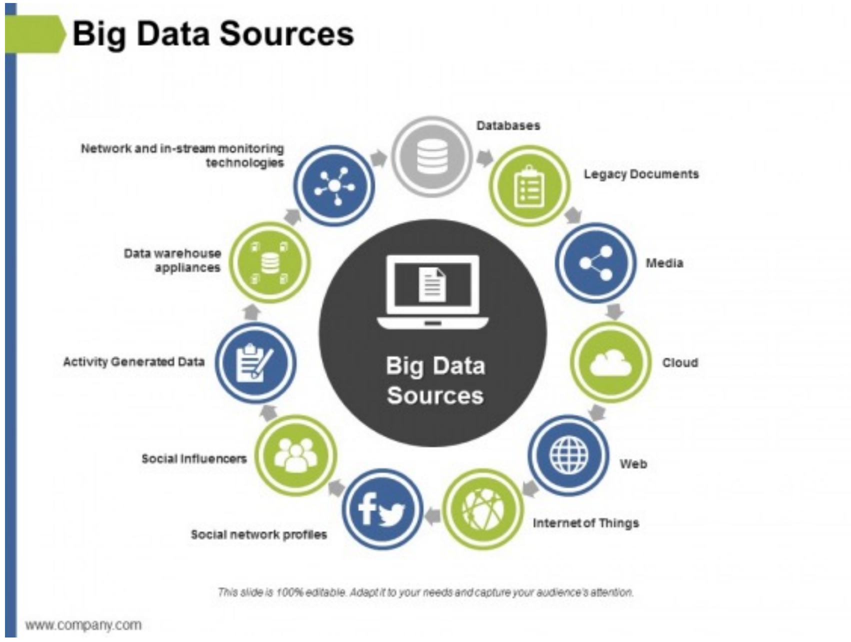
(2) **Doing something** with it:

- making decisions, confirming hypotheses, gaining insights, predicting future, etc.

Jennifer Widom, Stanford University

WHERE IS THE BIG DATA???

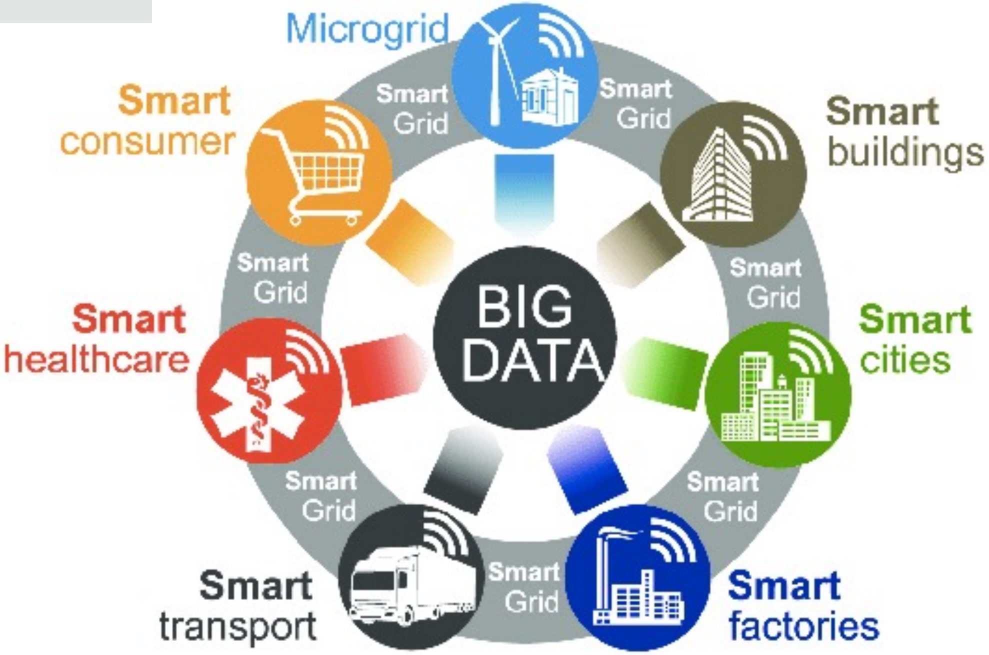




Big Data Sources

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

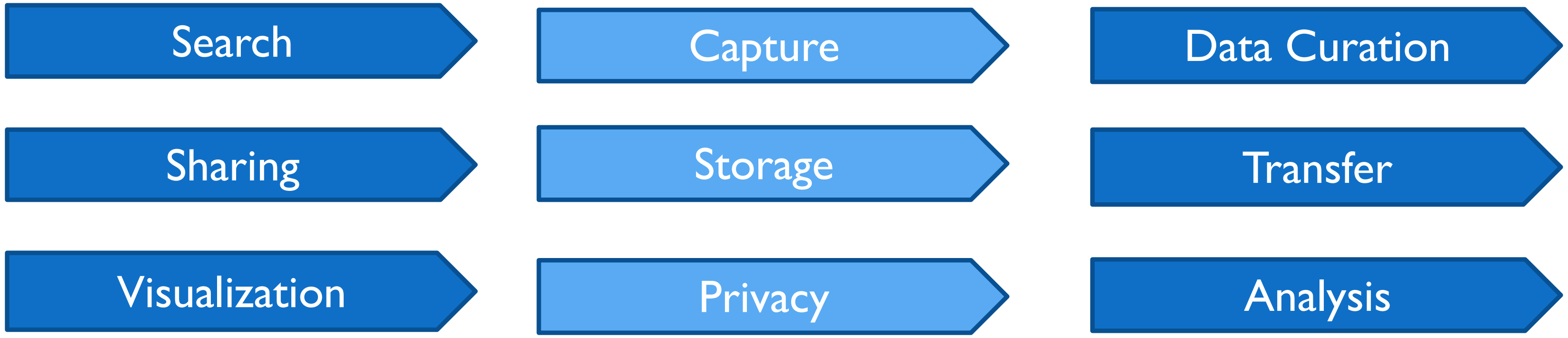
www.company.com



From Google Image

BIG DATA (SOME) CHALLENGES

- A broad term for such **large or complex data sets** that **traditional data processing** applications **are inadequate**.



Search

Capture

Data Curation

Sharing

Storage

Transfer

Visualization

Privacy

Analysis

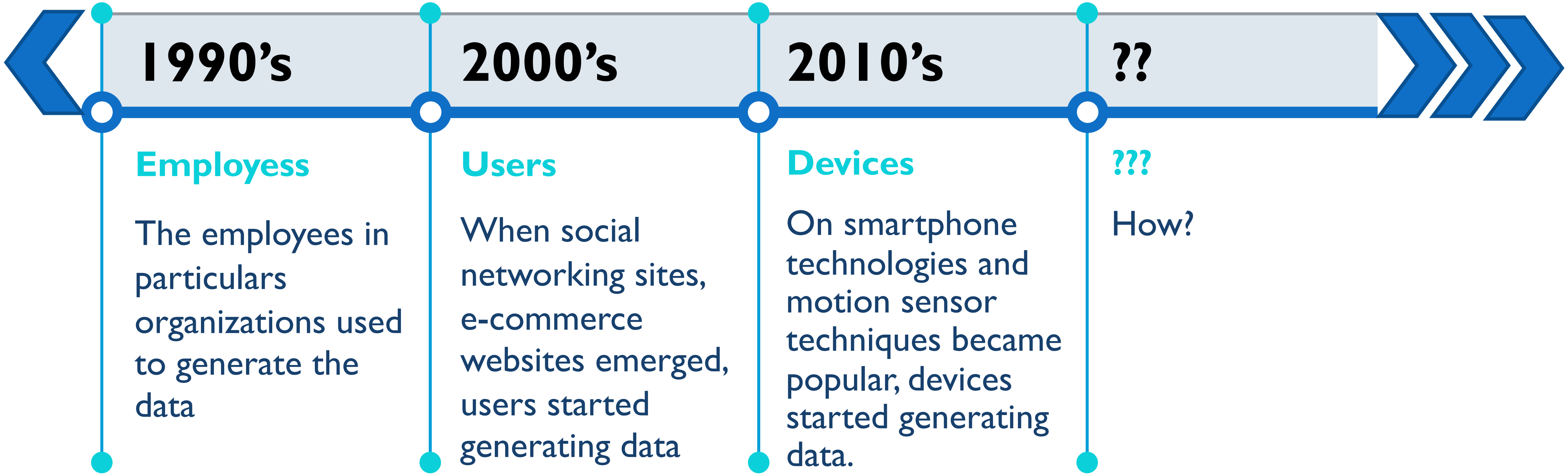
HOW MUCH DATA IS GENERATED EVERY DAY?

- *“There are about 2.5 quintillion bytes* of data created each day”*
- Every **minute**:
  - **Facebook**: there are 510,000 comments posted and 293,000 statuses updated
  - **Twitter**: 456,000 tweets are sent
  - **Snapchat**: users share 527,760 photos
  - **LinkedIn**: more than 120 professionals join it
  - **Instagram**: 46,740 photos are posted

Bernard Marr. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Forbes, 2018.

* https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1

HOW IS THE DATA GENERATED?



1990's

Employess

The employees in particulars organizations used to generate the data

2000's

Users

When social networking sites, e-commerce websites emerged, users started generating data

Orkut/Facebook – 2004
MySpace – 2005

2010's

Devices

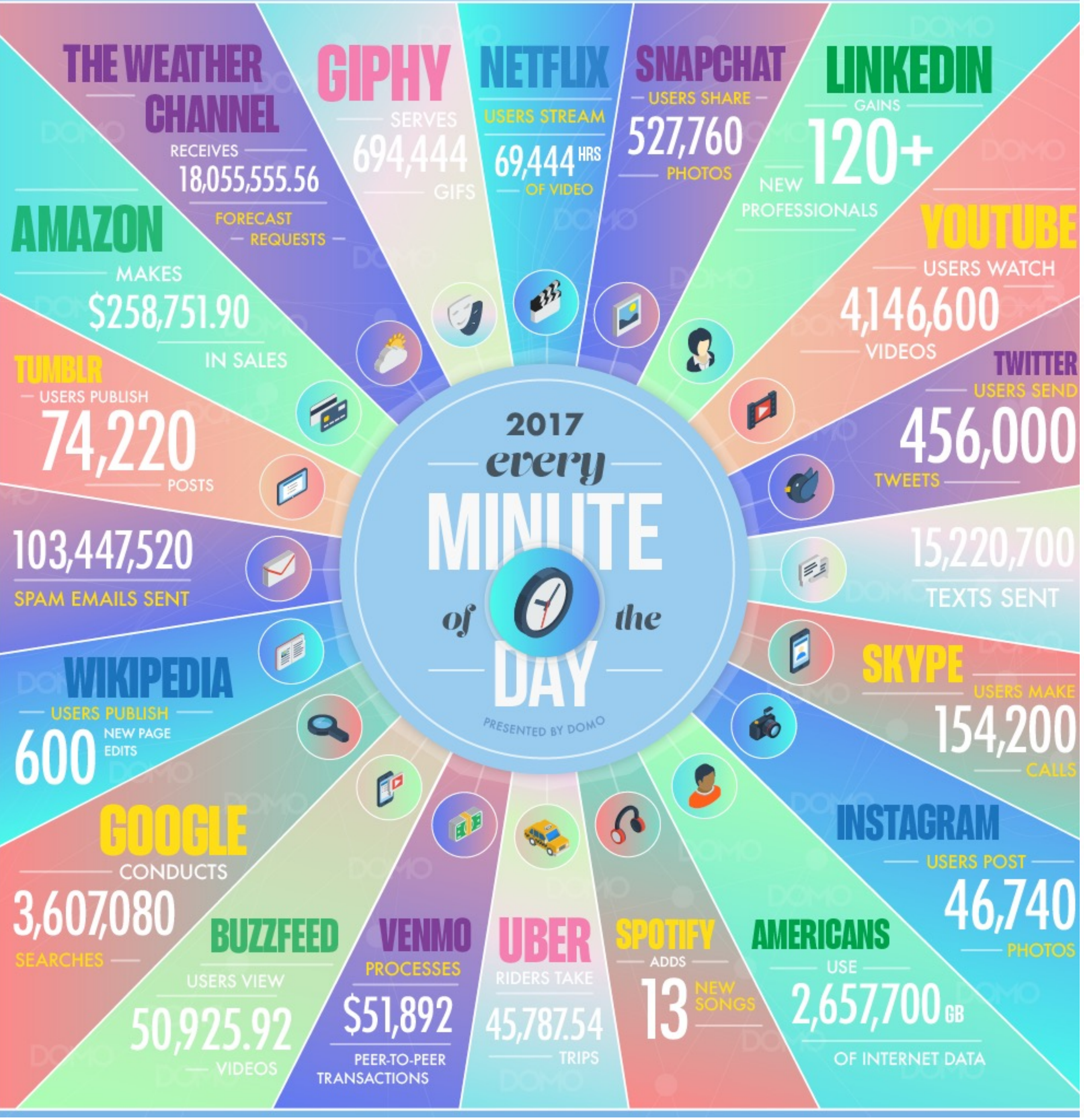
On smartphone technologies and motion sensor techniques became popular, devices started generating data.

??

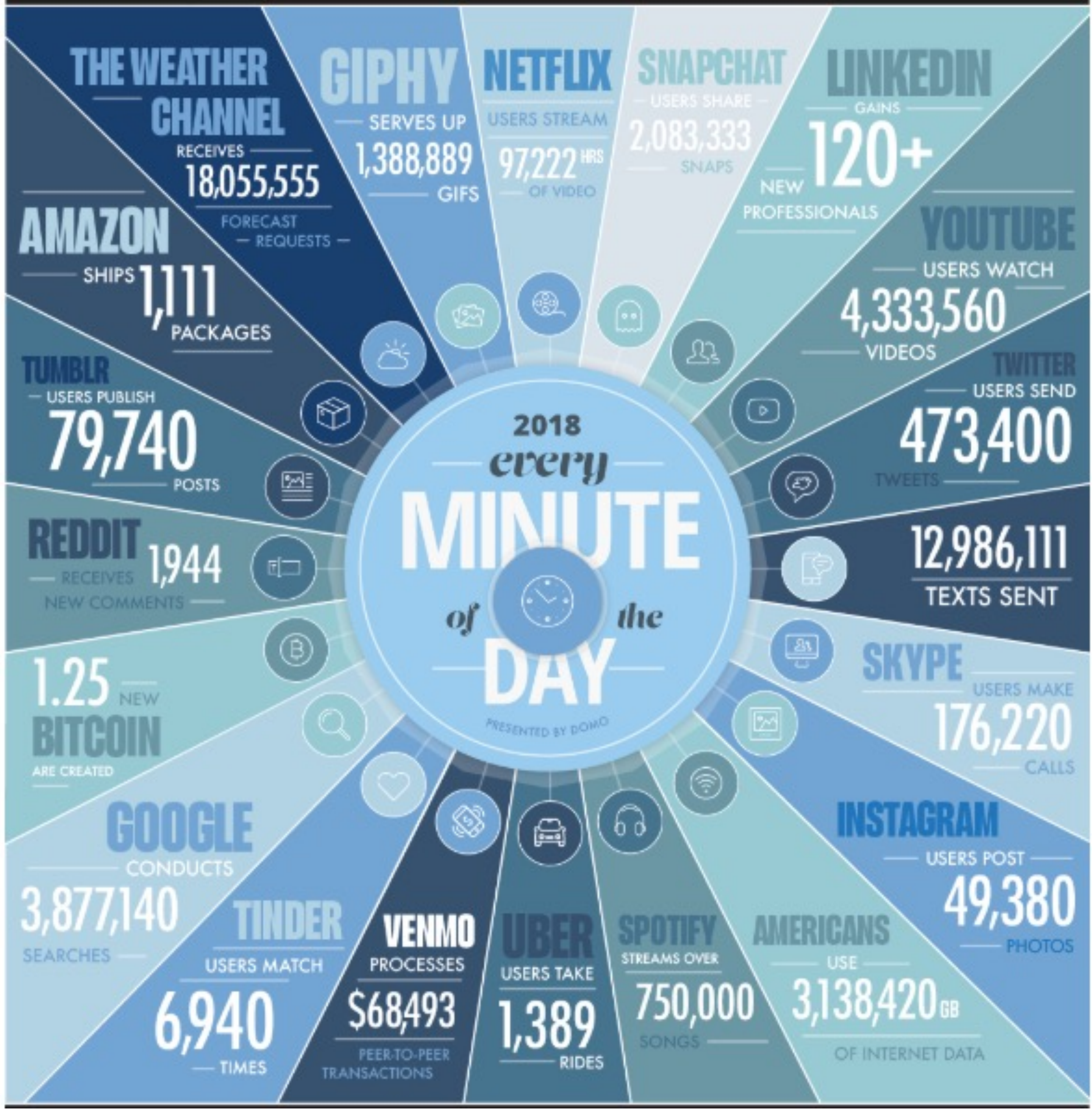
???

How?

Over the **last two years** alone **90%** of the data in the world was generated


2017 every MINUTE of the DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL RECEIVES 18,055,555.56 FORECAST REQUESTS
GIPHY SERVES 694,444 GIFS
NETFLIX USERS STREAM 69,444 HRS OF VIDEO
SNAPCHAT USERS SHARE 527,760 PHOTOS
LINKEDIN GAINS 120+ NEW PROFESSIONALS
YOUTUBE USERS WATCH 4,146,600 VIDEOS
TWITTER USERS SEND 456,000 TWEETS
15,220,700 TEXTS SENT
SKYPE USERS MAKE 154,200 CALLS
INSTAGRAM USERS POST 46,740 PHOTOS
AMERICANS USE 2,657,700 GB OF INTERNET DATA
SPOTIFY ADDS 13 NEW SONGS
UBER RIDERS TAKE 45,787.54 TRIPS
VENMO PROCESSES $51,892 PEER-TO-PEER TRANSACTIONS
BUZZFEED USERS VIEW 50,925.92 VIDEOS
GOOGLE CONDUCTS 3,607,080 SEARCHES
WIKIPEDIA USERS PUBLISH 600 NEW PAGE EDITS
103,447,520 SPAM EMAILS SENT
TUMBLR USERS PUBLISH 74,220 POSTS
AMAZON MAKES $258,751.90 IN SALES

2017


2018 every MINUTE of the DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL RECEIVES 18,055,555 FORECAST REQUESTS
GIPHY SERVES UP 1,388,889 GIFS
NETFLIX USERS STREAM 97,222 HRS OF VIDEO
SNAPCHAT USERS SHARE 2,083,333 SNAPS
LINKEDIN GAINS 120+ NEW PROFESSIONALS
YOUTUBE USERS WATCH 4,333,560 VIDEOS
TWITTER USERS SEND 473,400 TWEETS
12,986,111 TEXTS SENT
SKYPE USERS MAKE 176,220 CALLS
INSTAGRAM USERS POST 49,380 PHOTOS
AMERICANS USE 3,138,420 GB OF INTERNET DATA
SPOTIFY STREAMS OVER 750,000 SONGS
UBER USERS TAKE 1,389 RIDES
VENMO PROCESSES $68,493 PEER-TO-PEER TRANSACTIONS
TINDER USERS MATCH 6,940 TIMES
GOOGLE CONDUCTS 3,877,140 SEARCHES
1.25 NEW BITCOIN ARE CREATED
REDDIT RECEIVES 1,944 NEW COMMENTS
TUMBLR USERS PUBLISH 79,740 POSTS
AMAZON SHIPS 1,111 PACKAGES

2018

Source: https://www.domo.com

COMING BACK

WHAT DOES "BIG DATA" MEANS?

- (1) **Collecting** large amounts of data
  - Via computers, sensors, people, events …

© Can Stock Photo

- (2) **Doing something** with it
  - Making decisions, confirming hypotheses, gaining insights, predicting future …

Jennifer Widom, Stanford University

WHAT DOES "BIG DATA" MEANS?

- (1) **Collecting** large amounts of data
  - Via computers, sensors, people, events …

© Can Stock Photo

- (2) **Doing something** with it
  - Making decisions, confirming hypotheses, gaining insights, predicting future …

"Data Science" = Going from (1) to (2)

Jennifer Widom, Stanford University

ACTUALLY, DATA SCIENCE IS…

- Science:
  - the careful study of the structure and behavior of **something**, especially by watching, measuring, and doing experiments, and the development of theories to describe the results of these activities
- Data:
  - information, especially facts or numbers, in an electronic format that can be stored and processed by a computer

Data Science can be defined as: "*the careful study of the structure and behavior of **data**, especially by watching, measuring, and doing experiments, and the development of theories to describe the results of these activities*"

Jennifer Widom, Stanford University

BIG DATA

- Ability to collect data will only increase
- Ability to analyze data will only improve

Web Crawling. WEB scraping. Data Extraction. Dark Data

CONTENT DETECTION

"Understanding of digital file formats, their detection and data extraction from them"

Chris Mattmann
University of Southern California

CONTENT DETECTION

*“Understanding of digital **file formats**, their detection and data extraction from them”*

Chris Mattmann
University of Southern California

CONTENT DETECTION

*"Understanding of digital **file formats**, their detection and data extraction from them"*

Chris Mattmann
University of Southern California

- How is its structure?
- Is it structured, semistructured or unstructured?
- Is there noise content on it?
- If image: what are the features will be extracted?
- ????

CONTENT DETECTION

*“Understanding of digital file formats, **their detection** and data extraction from them”*

Chris Mattmann
University of Southern California

Identify, on a very large data set, which are the desired formats

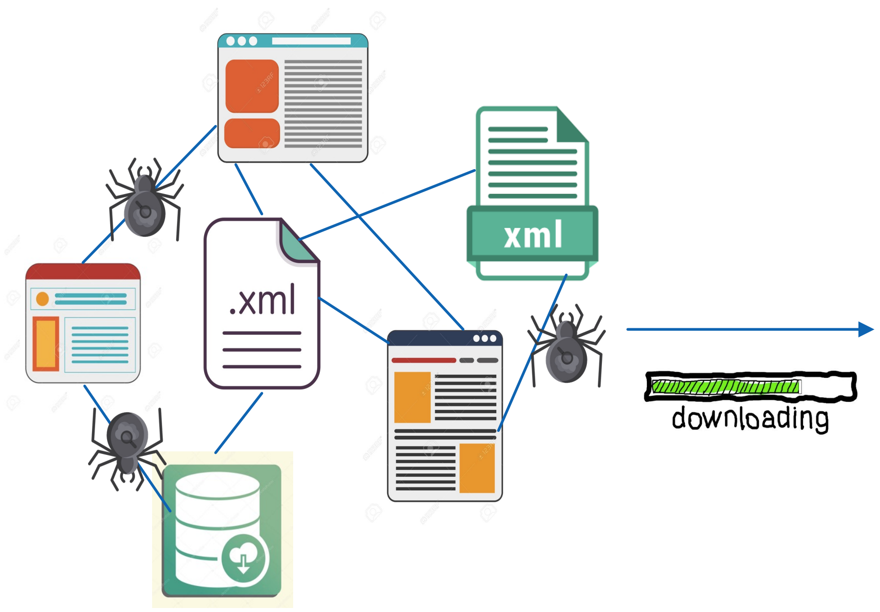
CONTENT DETECTION

"Understanding of digital file formats, their detection and ***data extraction from them****"*

Chris Mattmann
University of Southern California

Identify relevant data/information from documents, or texts aggregating them into a homogeneous format

WEB SCRAPING



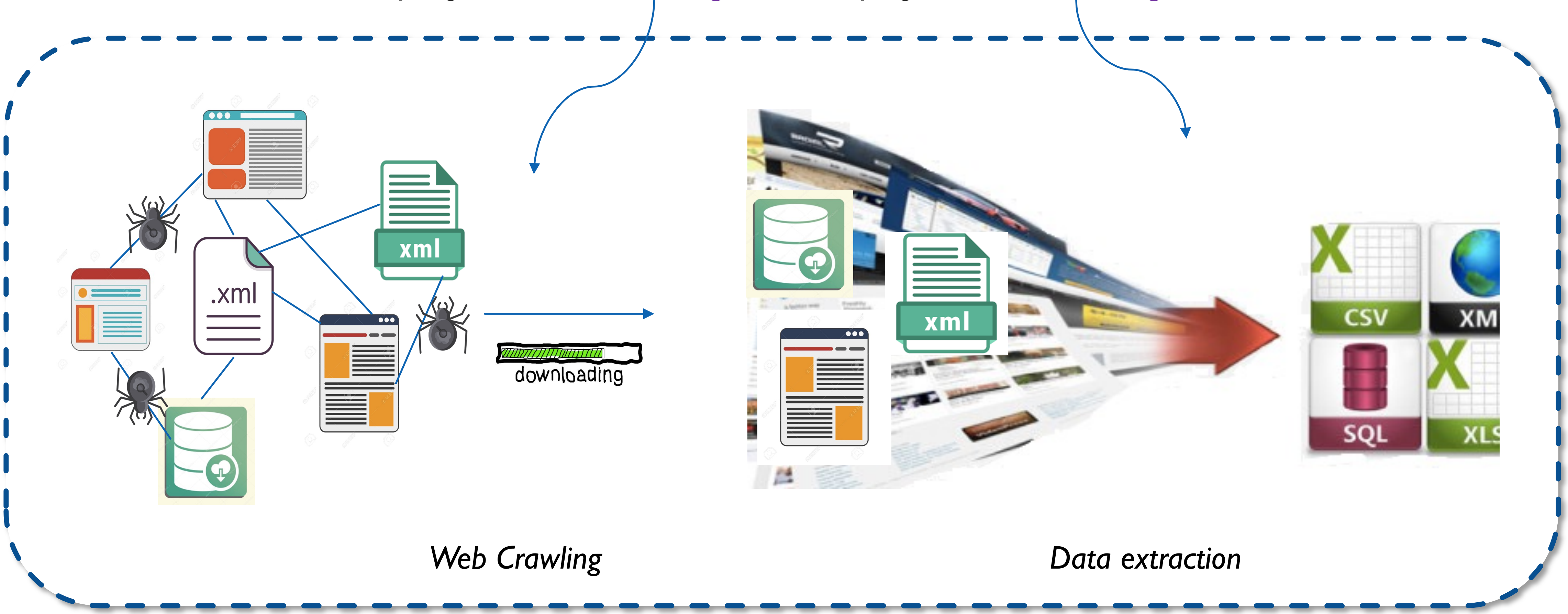
Web Crawling

xml
CSV
XM
SQL
XLS

Data extraction

WEB SCRAPING

Web scraping involves ***fetching*** *the web page and* ***extracting*** *from it*



Web Crawling

Data extraction

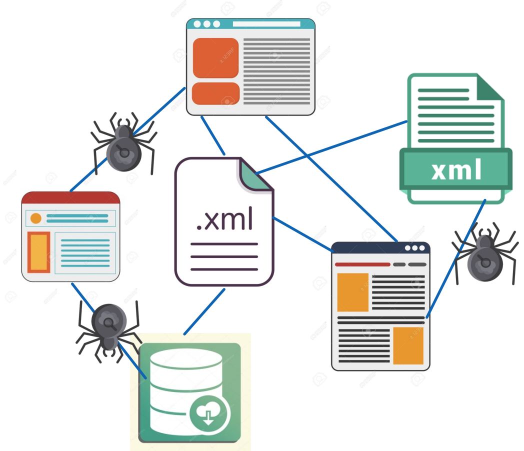
Web Crawling

PPGCC
Programa de Pós-Graduação
em Ciência da Computação

UFSC

INTRODUCTION

- A Web Crawler is a software/algorithm for **downloading** pages/datasets from the **Web**
- Also known as **Web Spider**, **Web Robot**, or simply **Bot**
- Web crawling steps
  1. Downloading a set of seed pages, that are parsed and scanned for new links
  2. Added to a central queue the links that have not yet been downloaded (for download later)
  3. Select a new page for download and the process is repeated until a stop criterion is met



APPLICATIONS

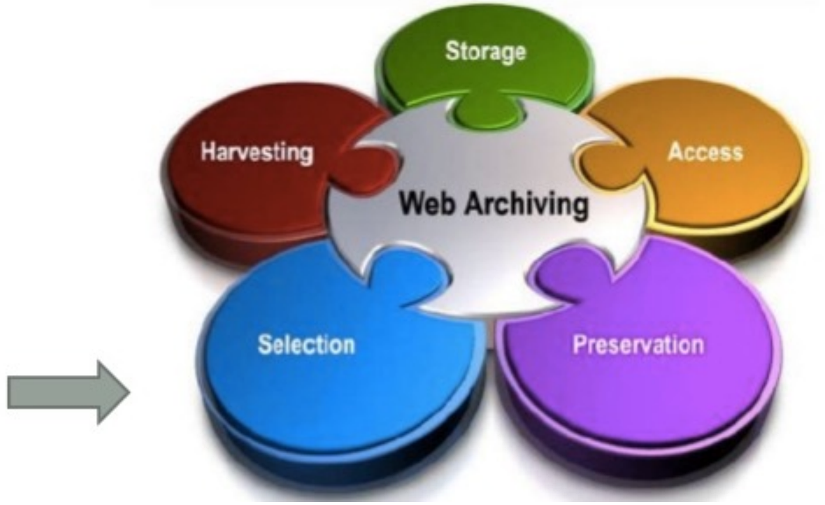
- Create indexes
  - Covering **broad** topics (general Web search)
  - Covering **specific** topics (vertical Web search)
- Archive content (Web archival)
- Analyze Web sites for extracting aggregate statistics (Web characterization)
- Keep copies or replicate Web sites (Web mirroring)
- Web sites analysis
- Knowledge bases building/enrichment

APPLICATIONS

- Create indexes
  - Covering **broad** topics (general Web search)
  - Covering **specific** topics (vertical Web search)
- Archive content (Web archival)
- Analyze Web sites for extracting aggregate statistics (Web characterization)
- Keep copies or replicate Web sites (Web mirroring)
- Web sites analysis
- Knowledge bases building/enrichment

Process of collecting portions of WWW to ensure the information is preserved for future researchers and the public

Web Archive Life Cycle



APPLICATIONS

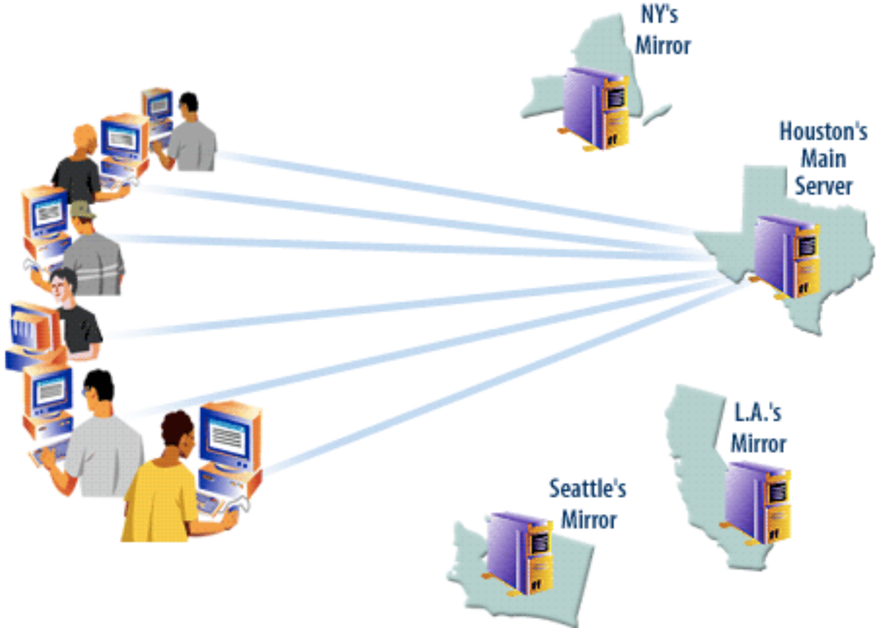
- Create indexes
  - Covering **broad** topics (general Web search)
  - Covering **specific** topics (vertical Web search)
- Archive content (Web archival)
- Analyze Web sites for extracting aggregate statistics (Web characterization)
- Keep copies or replicate Web sites (Web mirroring)
- Web sites analysis
- Knowledge bases building/enrichment

Answer the question:
What does the Web look like?

- Number of public sites
- Websites by country
- Popular Websites
- Websites language
- HTML vs. non-HTML contente
- How dynamic is the web?
- Downloads? Uploads? New Pages?
- ....

APPLICATIONS

- Create indexes
  - Covering **broad** topics (general Web search)
  - Covering **specific** topics (vertical Web search)
- Archive content (Web archival)
- Analyze Web sites for extracting aggregate statistics (Web characterization)
- Keep copies or replicate Web sites (Web mirroring)
- Web sites analysis
- Knowledge bases building/enrichment



Share Web Site activity during of high visitation or servers problems

APPLICATIONS

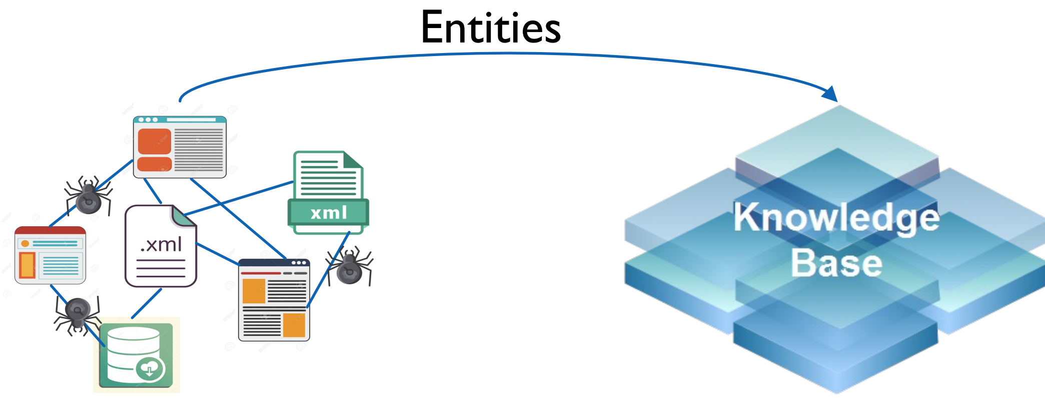
- Create indexes
  - Covering **broad** topics (general Web search)
  - Covering **specific** topics (vertical Web search)
- Archive content (Web archival)
- Analyze Web sites for extracting aggregate statistics (Web characterization)
- Keep copies or replicate Web sites (Web mirroring)
- Web sites analysis
- Knowledge bases building/enrichment

1. The website/page rank in the search results
2. Total number of visitors-daily/monthly
3. Number of visitors that were generated from advertisements
4. New visitors
5. …

Tarefa feita pelo Google Analytics

APPLICATIONS

- Create indexes
  - Covering **broad** topics (general Web search)
  - Covering **specific** topics (vertical Web search)
- Archive content (Web archival)
- Analyze Web sites for extracting aggregate statistics (Web characterization)
- Keep copies or replicate Web sites (Web mirroring)
- Web sites analysis
- Knowledge bases building/enrichment



APPLICATIONS

- Create indexes
  - Covering **broad** topics (general Web search)
  - Covering **specific** topics (vertical Web sea
- Archive content (Web archival)
- Analyze Web sites for extracting
- Keep copies or replicate Web sites
- Web sites analysis
- Knowledge bases building/enrichment

TYPES OF CRAWLER

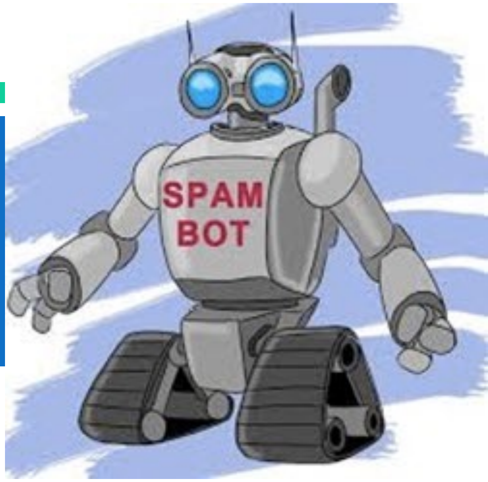
- General Web search

- Vertical Web search

GENERAL CRAWLER

- General Web search
  - Done by large search engines (Google, Yahoo!, Bing)
  - Must balance coverage and quality
    - **Coverage**: It must scan pages that can be used to answer many different queries
    - **Quality**: The pages should have high quality

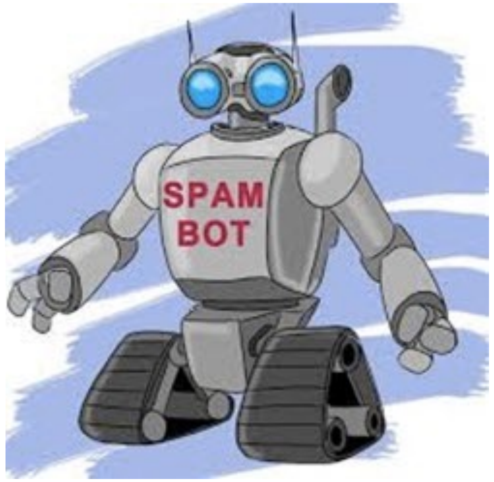
VERTICAL CRAWLER



- Vertical Web search
  - Focus on a particular subset of the Web, defined geographically, linguistically, topically, etc.
  - Examples
    - Shopbot: designed to download information from on-line shopping catalogs and provide an interface for comparing prices in a centralized way
    - News crawler: gathers news items from a set of pre-defined sources
    - Spambot: crawler aimed at harvesting e-mail addresses inserted on Web pages

VERTICAL CRAWLER

- Also includes segmentation by a data format or structure
  - Format: collect only objects of a specific type, as image, audio, or video objects
  - Structure: collect objects of a specific structure (Web forms, deep web data)
  - Example
    - Feed crawler: checks for updates in RSS/RDF files in Web sites

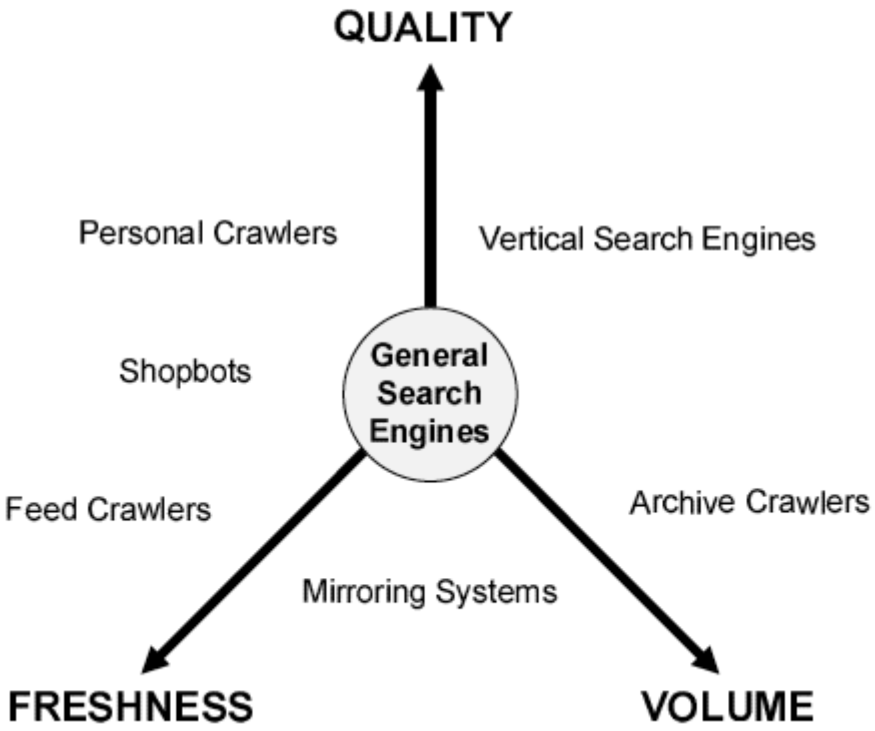


FOCUSED CRAWLER

- Vertical crawler that focus on a specific topic
- A more efficient strategy to avoid collecting more pages than necessary
  - Main problem of focused crawling: to predict the relevance of a page before downloading the page
- The input is the description of a topic and usually is
  - A driving query
  - A set of example documents
- It can operate in
  - Batch mode, collecting pages periodically
  - On-demand, collecting pages driven by a user query

CRAWLER CLASSIFICATION

- The crawlers can be classified according to three axes



POLITENESS

- Crawlers should fulfill politeness
  - A crawler cannot overload a Web site with HTTP requests
    - It implies that a crawler should wait a small delay between two requests to the same Web site

POLITENESS POLICY

- Robots are useful for a number of tasks, but with a price for the general community
  - Web crawlers require considerable bandwidth
  - Server overload, specially if the frequency of access to a given server is high, and/or if the robot is poorly written
- A set of guidelines is also important for the continued operation of a Web crawler
- A crawler that is impolite with a Web site may be banned by the hosting provider

- The three basic rules for Web crawler operation are:
  - Must **identify** itself as a robot, and must not pretend to be a regular Web user
  - Must obey the robots **exclusion protocol**
  - Must keep a **low bandwidth usage** in a given Web site

POLITENESS POLICY - MUST **IDENTIFY** ITSELF AS A ROBOT

- Web servers detect the navigational pattern of a crawler
- Detection is more effective if the crawler identifies itself
  - HTTP protocol includes a user-agent field that can be used to identify who is issuing a request
  - The Web crawler should include an address in this field containing information on the crawler, as well as contact information

- Must **identify** itself as a robot, and must not pretend to be a regular Web user
- Must obey the robots **exclusion protocol**
- Must keep a low **bandwidth usage** in a given Web site

POLITENESS POLICY - MUST OBEY THE **EXCLUSION PROTOCOL**

- Types: server-wide, page-wise exclusions, and cache exclusions
  - Server-wide exclusion instructs the crawler about directories that should not be crawled (via a robots.txt file located in the root directory of a Web site)

```
User-agent: *
Disallow: /data/private
Disallow: /cgi-bin
```

  - Page-wise exclusion is done by the inclusion of meta-tags in the pages themselves (HTML source)

```
<meta name="robots” content="noindex,nofollow"/>
```

  - Cache exclusion is used by publishers that sell access to their information

```
<meta name="robots" content="nocache"/>
```

- Must obey the **robots exclusion protocol**
- Must keep a low **bandwidth usage** in a given Web site

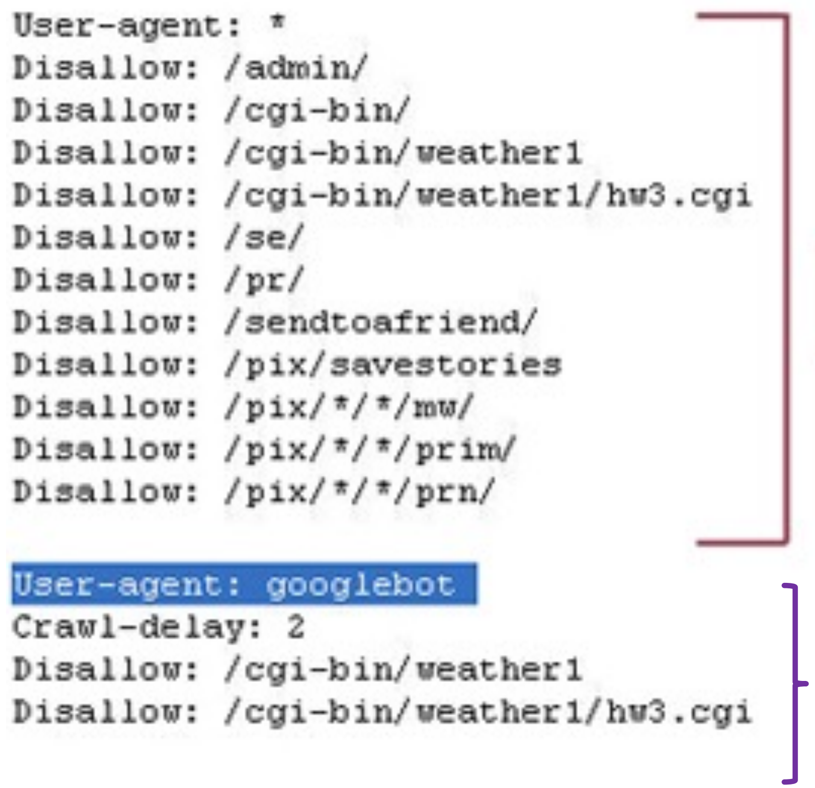
POLITENESS POLICY - MUST KEEP A **LOW BANDWIDTH USAGE**

- The use of Web robots, useful for a number of tasks, but with a price for the general community
  - Web crawlers require considerable bandwidth

- A Web crawler might easily overload a Web server, specially a smaller one
- To avoid this:
  - to open only one connection to a given Web server at a time
  - to take a delay between two consecutive accesses
    - Some authors suggest adopting 10 seconds as the interval between consecutive accesses, others 15 or 30 seconds
    - Some Web site operators decide which is the delay that should be used

  - Must keep a **low bandwidth usage** in a given Web site

ROBOTS.TXT

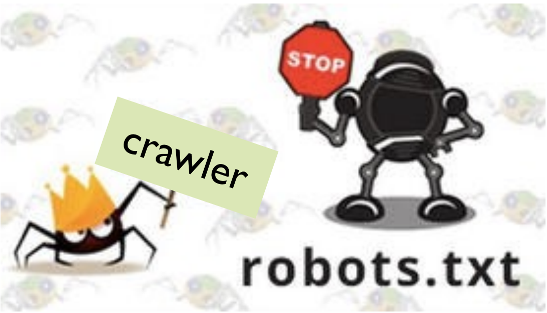


```
User-agent: *
Disallow: /admin/
Disallow: /cgi-bin/
Disallow: /cgi-bin/weather1
Disallow: /cgi-bin/weather1/hw3.cgi
Disallow: /se/
Disallow: /pr/
Disallow: /sendtoafriend/
Disallow: /pix/savestories
Disallow: /pix/*/*/mw/
Disallow: /pix/*/*/prim/
Disallow: /pix/*/*/prn/

User-agent: googlebot
Crawl-delay: 2
Disallow: /cgi-bin/weather1
Disallow: /cgi-bin/weather1/hw3.cgi
```

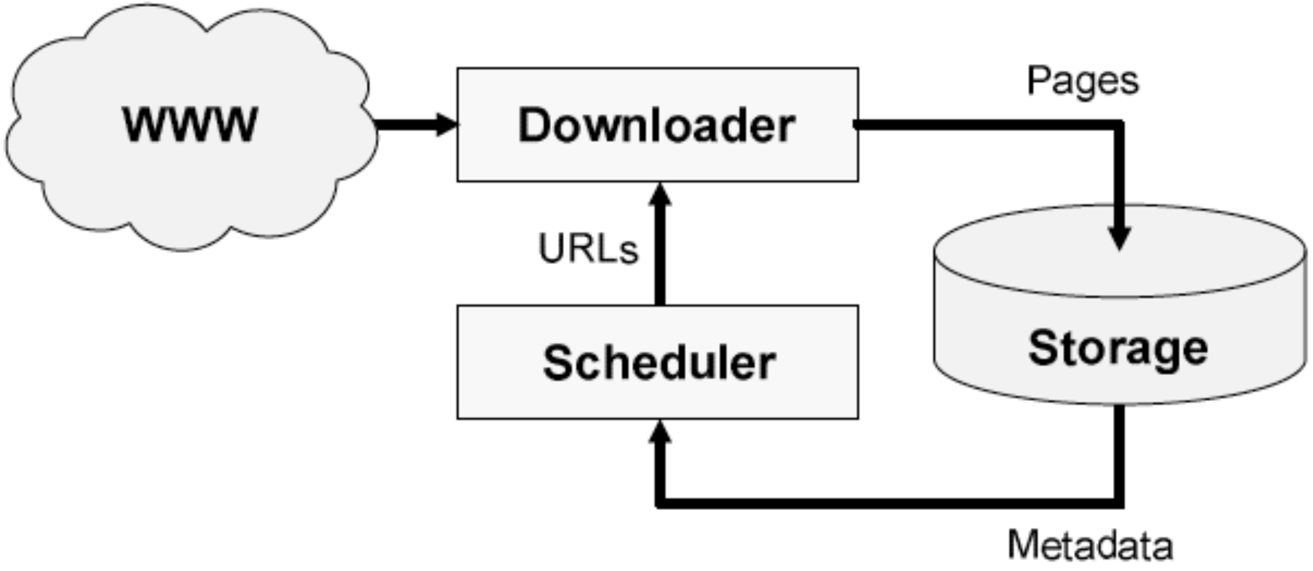
Will be negleted by bots

Instructions to GoogleBot



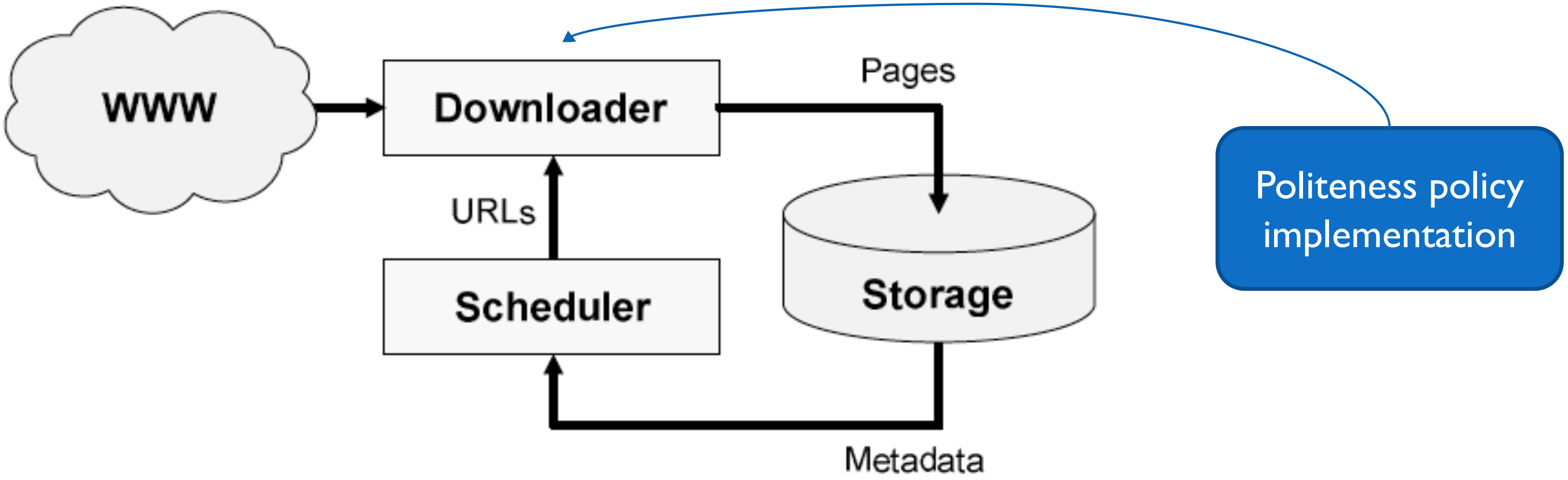
CRAWLER ARCHITECTURE

- Scheduler: maintains a queue of URLs to visit
- Downloader: downloads the pages
- Storage: makes the indexing of the pages, and provides the scheduler with metadata on the pages retrieved



CRAWLER ARCHITECTURE

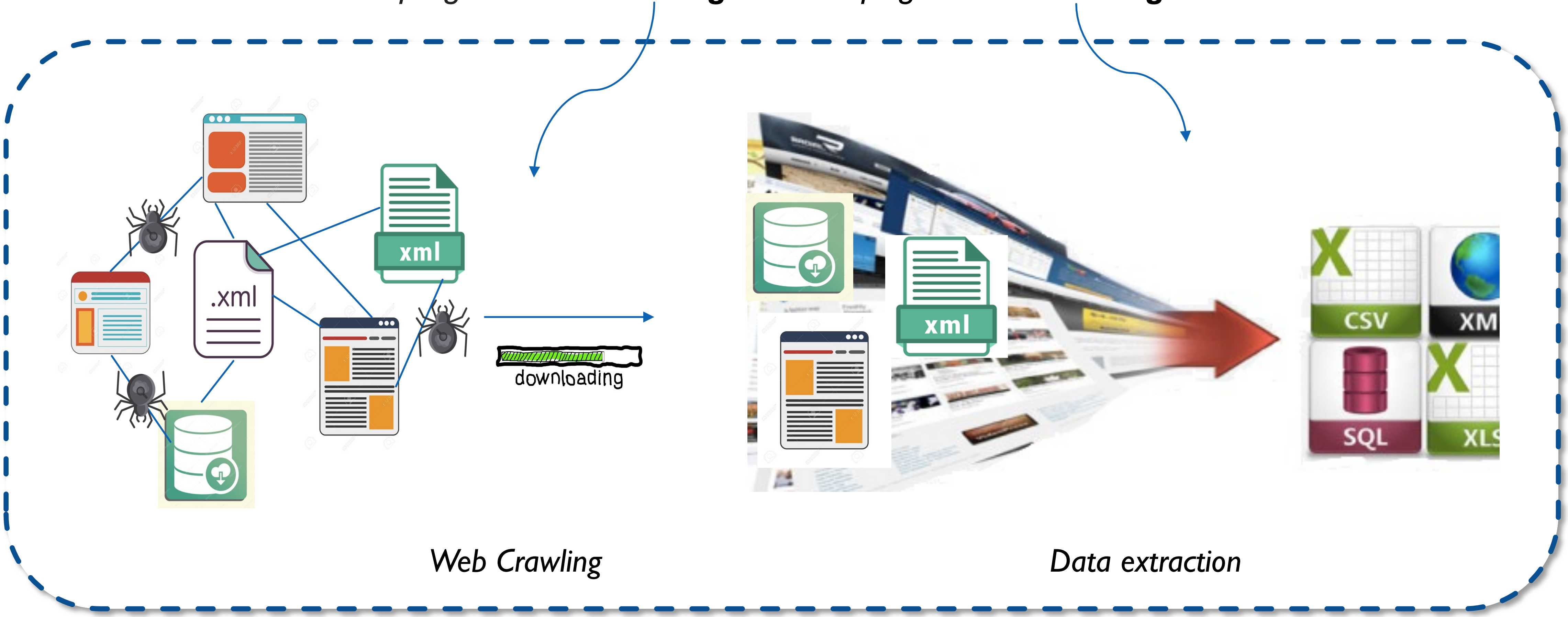
- Scheduler: maintains a queue of URLs to visit
- Downloader: downloads the pages
- Storage: makes the indexing of the pages, and provides the scheduler with metadata on the pages retrieved



CRAWLER EXAMPLES

- Heritrix: Internet Archieve
- GoogleBot
- Java:
  - WebSPHINX
  - NUTCH (part of the Lucene search engine)
  - Crawler4j
- C:
  - WIRE
  - Dig
- Python
  - Scrapy
  - Beautifulsoup

*Web scraping involves **fetching** the web page and **extracting** from it*



Web Crawling

Data extraction

Data Extraction

DATA EXTRACTION

- Data Extraction
  - Process executed by Information Extraction (IE) systems
  - Find and understand relevant parts of texts
  - Join information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a *knowledge base*…
  - Goals:
    1. Organize information so that it is useful to people
    2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

INFORMATION EXTRACTION (IE)

- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- Example:
  - Join earnings, profits, board members, headquarters, etc. from company reports
    - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
    - headquarters("BHP Biliton Limited", "Melbourne, Australia")
  - Learn drug-gene product interactions from medical research literature

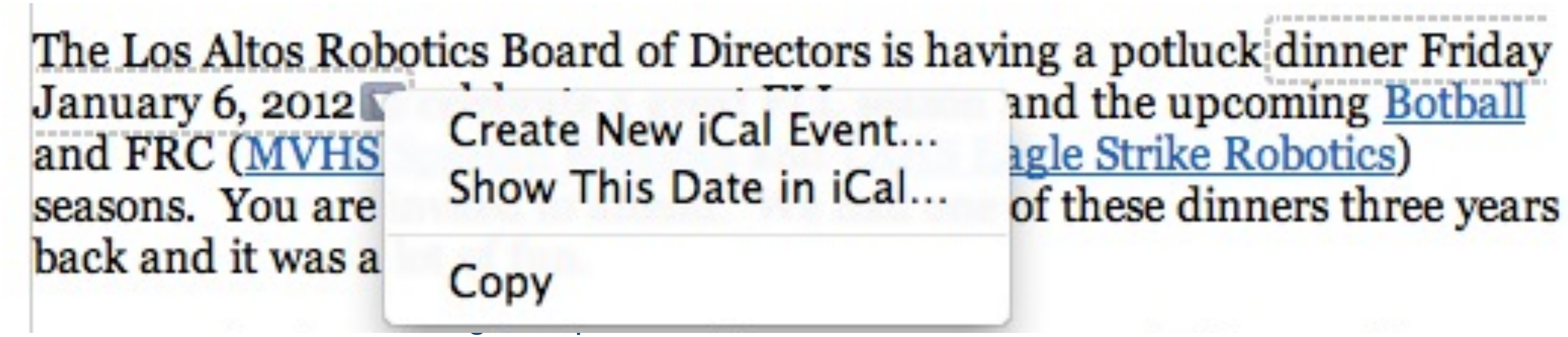
INFORMATION EXTRACTION (IE)

- High-level
  - to determine the high level structure, that is where the sections are with their headings, which part is the reference section, what is a table, etc.

- Low-level
  - to determine the low level structure, that is, given that you know a piece of text contains an affiliation, determine the individual elements of the affiliation like, for example, institute, street address, post box, city, zipcode, state, and country.

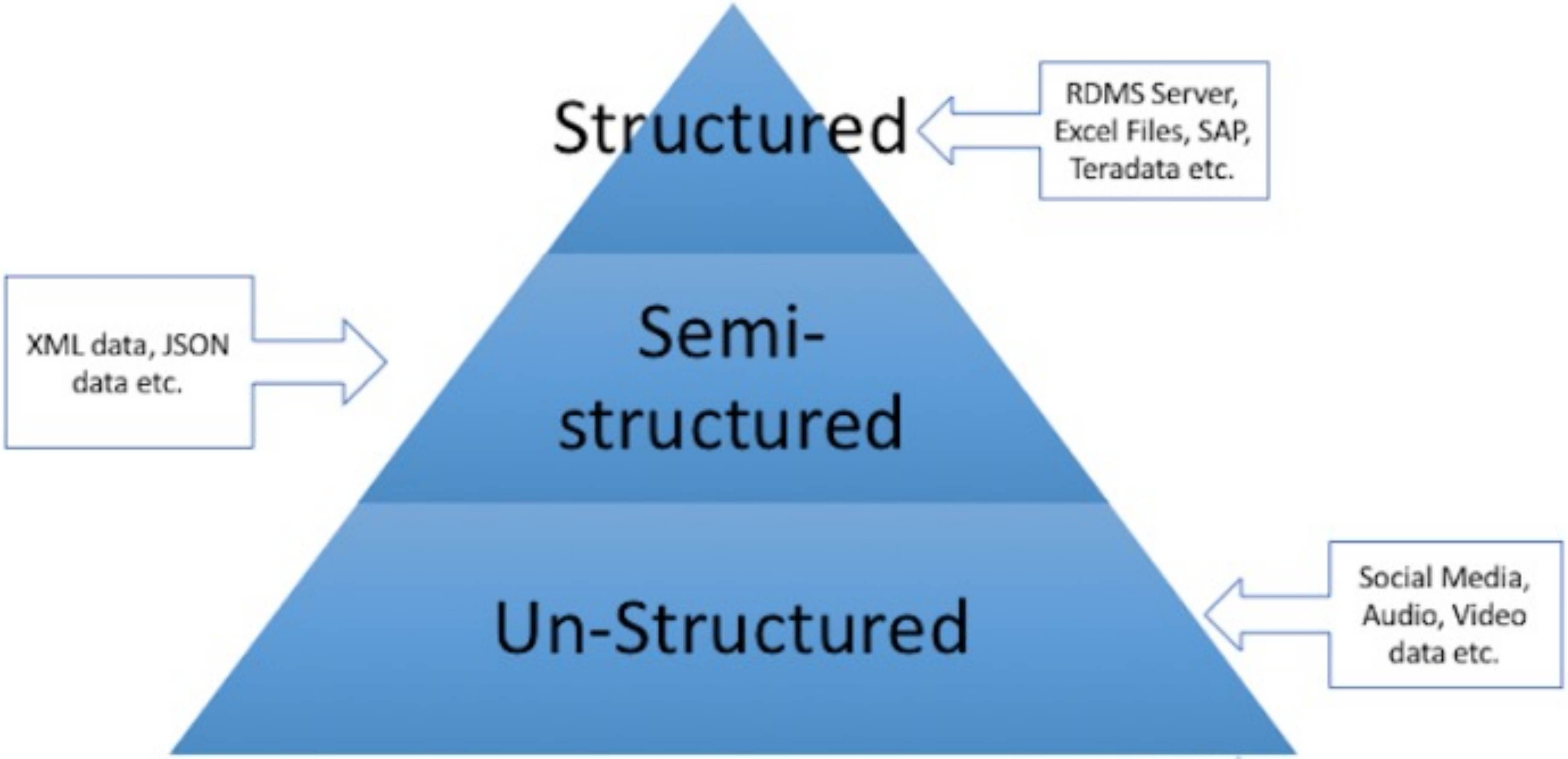
LOW-LEVEL INFORMATION EXTRACTION

- Is now available in applications like Apple or Google mail, and web indexing
  - Specialized kinds of relations done using regular expressions



The Los Altos Robotics Board of Directors is having a potluck dinner Friday
January 6, 2012 and the upcoming Botball
and FRC (MVHS agle Strike Robotics)
seasons. You are of these dinners three years
back and it was a

DATA SOURCE STRUCTURE



Structured

Semi-structured

Un-Structured

RDMS Server, Excel Files, SAP, Teradata etc.

XML data, JSON data etc.

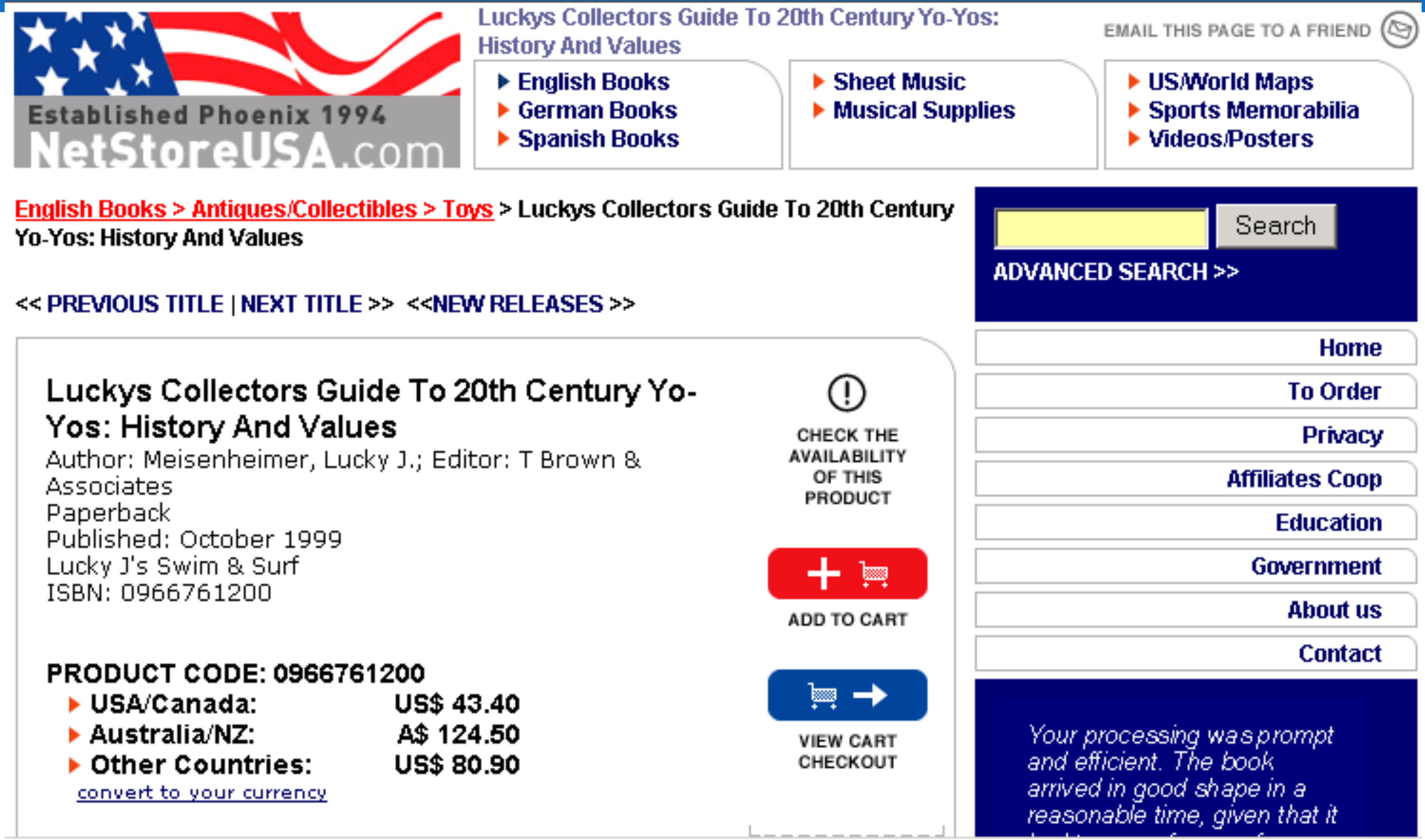
Social Media, Audio, Video data etc.

WHY IS IE HARD ON THE WEB?

A book,
Not a toy

Title

Need this price



Established Phoenix 1994
NetStoreUSA.com

Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

EMAIL THIS PAGE TO A FRIEND

- English Books
- German Books
- Spanish Books

- Sheet Music
- Musical Supplies

- US/World Maps
- Sports Memorabilia
- Videos/Posters

English Books > Antiques/Collectibles > Toys > Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

<< PREVIOUS TITLE | NEXT TITLE >> <<NEW RELEASES >>

Search

ADVANCED SEARCH >>

Home
To Order
Privacy
Affiliates Coop
Education
Government
About us
Contact

Luckys Collectors Guide To 20th Century Yo-Yos: History And Values
Author: Meisenheimer, Lucky J.; Editor: T Brown & Associates
Paperback
Published: October 1999
Lucky J's Swim & Surf
ISBN: 0966761200

CHECK THE AVAILABILITY OF THIS PRODUCT

ADD TO CART

VIEW CART CHECKOUT

PRODUCT CODE: 0966761200

USA/Canada:	US$ 43.40
Australia/NZ:	A$ 124.50
Other Countries:	US$ 80.90

convert to your currency

Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it

NAMED ENTITY RECOGNITION (NER)
A BRIEF OVERVIEW

CARINA DORNELES

CARINA.DORNELES@UFSC.BR

INE410136 - Content Detection and Analysis on Big Web Data

PPGCC
Programa de Pós-Graduação
em Ciência da Computação

UFSC

INTRODUCTION

- Named Entity Recognition - NER
  - A process where an algorithm
    - Input: a **string** of text (sentence or paragraph)
    - Process: identifies **relevant nouns** (people, places, and organizations) that are mentioned in that string.
    - Output: **named entities**

NAMED ENTITY RECOGNITION (NER)

- Named Entity Recognition (NER)
- A **data extraction sub-task**

.

NAMED ENTITY RECOGNITION (NER)

- Named Entity Recognition (NER)
- A **data extraction sub-task**
  - find and classify names in text, for example:

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

NAMED ENTITY RECOGNITION (NER)

- Named Entity Recognition (NER)
- A **data extraction sub-task**
  - **find** and classify names in text, for example:

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Christopher Manning – Stanford University

NAMED ENTITY RECOGNITION (NER)

- Named Entity Recognition (NER)
- A **data extraction sub-task**
  - **find** and classify names in text, for example:

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

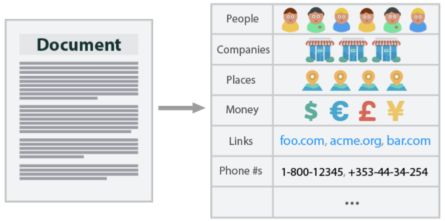
Person: Andrew Wilkie, Wilkie, Rob Oakeshott, Tomy Windsor
Date: 2010
Location: ---
Organization: Greens, Labor

Christopher Manning – Stanford University

APPLICATIONS AND USE CASES

- Classifying content
  - news providers – to categorize news
  - Powering Content Recommendations to recommend similar products/articles using named entities
  - Customer Support – To categorize the complaint and assign it to the relevant department within the organization

- Locate entity in a given document
  - “That person always appears in the context of some violence event”



- More Efficient Search Algorithms
  - relevant entities associated with each of those articles could speed up the search process considerably.

GENERAL VS. DOMAIN SPECIFIC NAMED ENTITIES

- For **general** entity such as name, location and organization
  - we can use pre-trained library which are Stanford NER, spaCy and NLTK NE_Chunk to tackle it.

- For **domain specific** entity, such as animals, trees, stars and so on
  - spend time on labeling so that we can recognize those entity.

METHODS FOR DOING NER

- Hand-written regular expressions - REGEX
- Classifiers methods, such as
    - Neural Networks
    - Decision Trees
    - Naïve Bayes and Bayesian Networks
    - Support Vector Machine
    - kNN (k-nearest-neighbor)
- Rule-based method
- Sequence models
    - Hidden Markov Model - HMM
    - Conditional Markov Model - CMM
    - Conditional Random Fields – CRF
- Deep Learning