

# Informatics 225

## Computer Science 221

### Information Retrieval

#### Lecture 6

*Duplication of course material for any commercial purpose without the explicit written permission of the professor is prohibited.*

*These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Prof. Alberto Krone-Martins, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.*

# Information Retrieval

Search Engine Optimization

*Coercion Techniques*

# Keyword Stuffing

---

- 1<sup>st</sup>- gen search engines relied **heavily** on textual content and frequency of words

# Keyword Stuffing

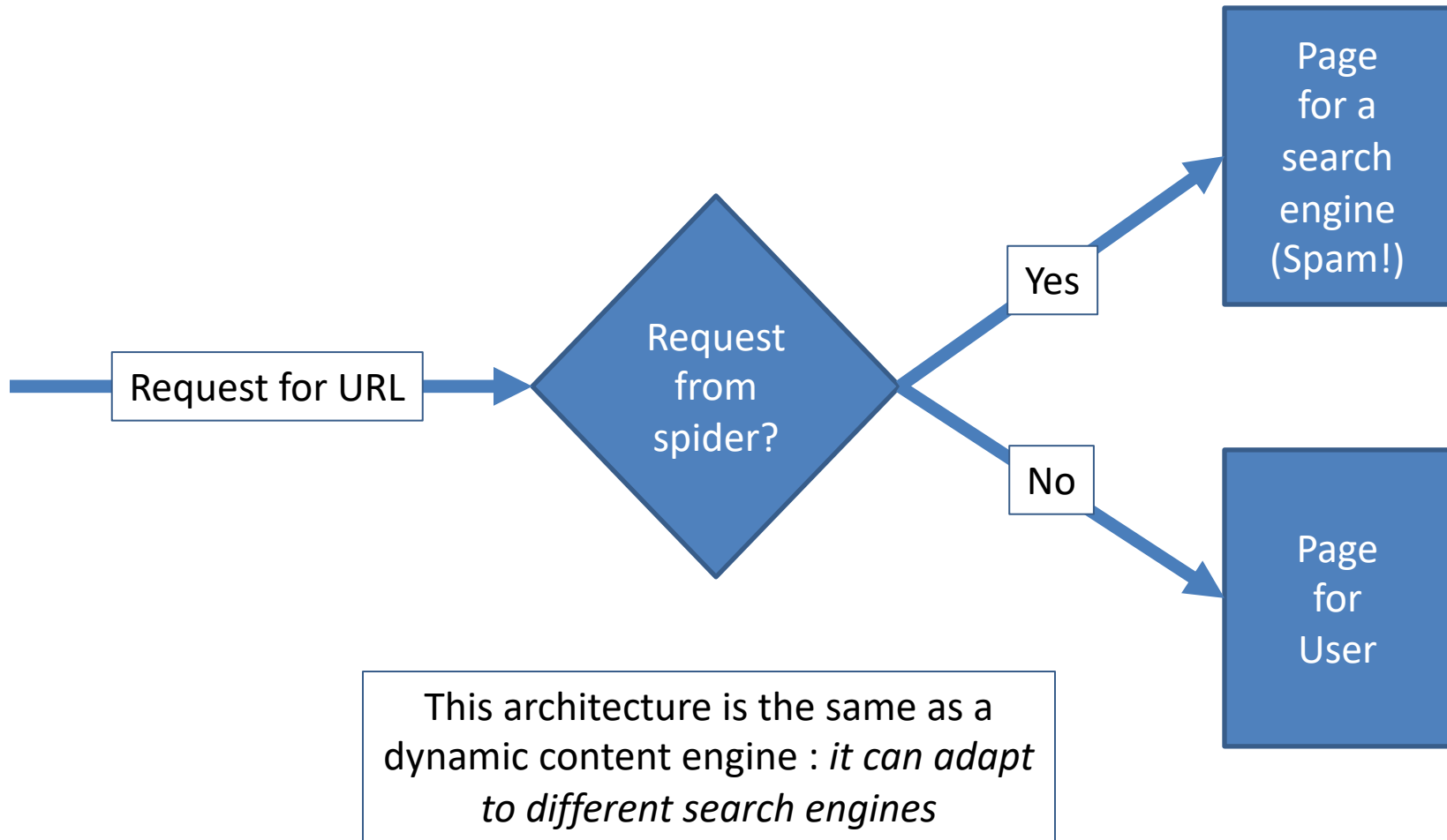
- 1<sup>st</sup>- gen search engines relied heavily on textual content and frequency of words
- SEO moved to play around with keywords
  - Misleading meta-tags

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.d
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<COMMENT TITLE="MONITOR"></COMMENT>
<meta http-equiv="Content-Language" content="en-us" />
<meta http-equiv="Content-type" content="text/html; charset=iso-8859-1"/>
<META NAME="ROBOTS" CONTENT="NOODP"><meta name="verify-v1" content="aeVxP6zTHeQzT620ipj5+ikXd/VXcdlKoYUJ/C6vVdY=" />
<META NAME="keywords" content="Expedia, Travel, Cheap Airfare, Car, Hotels, Vacations, Airfare, Car Rental, Cruises,
<META NAME="description" content="Purchase airline tickets, make hotel reservations, find vacation packages, car rent
```

- Repeating words over and over
- Playing games with colors (white on white)
  - visible to spiders/crawlers and indexers, but not users

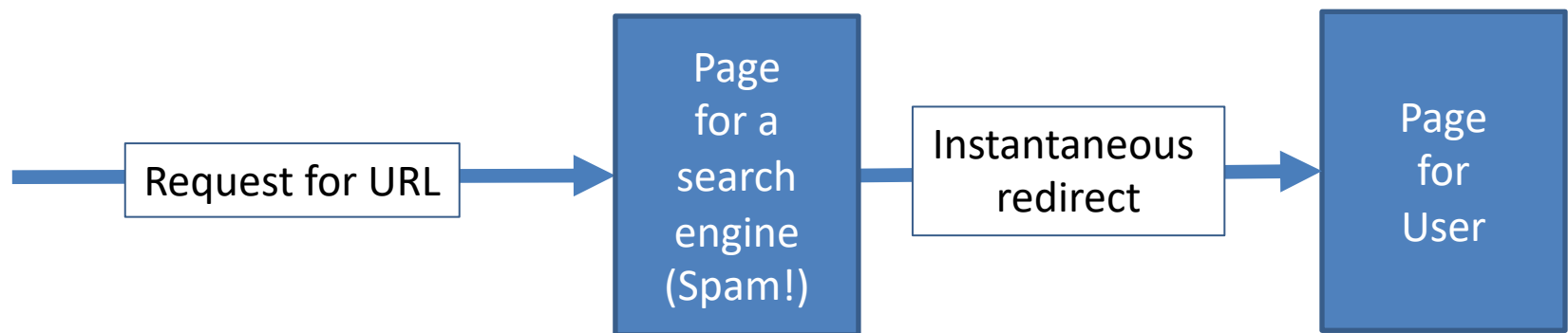
# Cloaking

- Serving different content to a spider than to a browser



# Doorway Pages

- Like cloaking but using a redirect (302)
  - Code 302 : “The requested resource resides **temporarily** under a different URI. Since the redirection might be altered on occasion, the client SHOULD continue to use the Request-URI for future requests”
  - **Initial page is optimized for spider**, then redirect takes user to actual content; **the user sees the final page.**



# Link exchanges

---

- I link to you, you link to me
- “Translations”
- Universities and professors are targets
  - Essentially, due to “trust” transfer (important at the link analysis level)

# Information Retrieval

Search Engine Optimization

*SPAM*



# Two major types

---

- Link spamming
  - Bots that search for blogs and leave comments with links
- Clicker bots
  - Bots that issue queries and “click” on targeted query results

# Spam industry

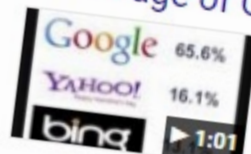
## The First Page of Google and How I Get my Clients There

[www.malleeblue.com/1st-page-google-optimization-tips/](http://www.malleeblue.com/1st-page-google-optimization-tips/)

★★★★★ Rating: 10/10 - 275 reviews

Nov 8, 2016 - Can I get your website or page to rank on the first page of Google? ... If you want to get 1st page of Google search engine results listing, make ...

## First Page of Google GUARANTEED !! | How To Get Your Business At ...



[https://www.youtube.com/watch?v=qjASS\\_NsIL8](https://www.youtube.com/watch?v=qjASS_NsIL8)

Dec 26, 2013 - Uploaded by Spotlight Ventures

Business on The First Page of Google GUARANTEED !! or Goto: ... video, blog, or Business Listing on the ...

## How To Get On The First Page of Google In 24 hours - YouTube



[https://www.youtube.com/watch?v=2O\\_pdMbJR3k](https://www.youtube.com/watch?v=2O_pdMbJR3k)

May 15, 2016 - Uploaded by Amazing Tricks World

how to get first page rank on google how to get first rank in google search how to rank first on google how ...

# Spamming Web contests

---

- Do you want to win a web contest?
  - <https://www.wholewhale.com/tips/how-to-win-or-cheat-any-online-voting-contest/>

# The war on spam

---

- Quality indicators
  - Statistical analysis of links
  - Statistical analysis of votes

# The war on spam

---

- Quality indicators
  - Statistical analysis of links
  - Statistical analysis of votes
- Usage indicators
  - Analytics

# The war on spam

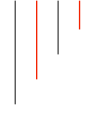
- Quality indicators
  - Statistical analysis of links
  - Statistical analysis of votes
- Usage indicators
  - Analytics
- Anti-bot mechanisms
  - Captchas



# The war on spam

---

- Limits on meta keywords
- Family-friendly filters
- Robust link analysis
  - Ignore statistically improbable links:
    - A page in Portuguese from Portugal links to a page in Chinese from China
  - Detect cycles
  - Use link analysis to detect spammers
    - Very simple rule : Guilt by association
- Editorial (human!) intervention
  - Black-listing and creation of training sets for supervised ML
- Spam recognition by machine learning



- Search engines have SEO policies
  - What is allowed and not allowed
- **Must not be ignored**
  - Once a site is **blacklisted by a search engine**, it will virtually **disappear** from the Web



# SEO “tricks” are volatile

---

- Dependent on how the web search engines work
- The methods behind the engines change
  - Google: PageRank (1996 - ?), Panda (2011), Penguin (2012), Hummingbird (2013), Pigeon (2014), ...
    - **Words and links.**

# SEO “tricks” are volatile

---

- Dependent on how the web search engines work
- The methods behind the engines change
  - Google: PageRank (1996 - ?), Panda (2011), Penguin (2012), Hummingbird (2013), Pigeon (2014), ...
    - **Words and links.**
  - Since 2016, it was using RankBrain: machine learning based.
    - **Concepts.**

# SEO “tricks” are volatile

- Dependent on how the web search engines work
- The methods behind the engines change
  - Google: PageRank (1996 - ?), Panda (2011), Penguin (2012), Hummingbird (2013), Pigeon (2014), ...
    - **Words and links.**
  - Since 2016, it was using RankBrain: machine learning based.
    - **Concepts.**
  - Since October 2019, Google is using BERT: NLP, machine learning.
  - Now Google Search uses : RankBrain + BERT + ... > 200 methods&signals (!)
  - Can you optimize for BERT? Google says no (but too early to tell).

# SEO “tricks” are volatile

---

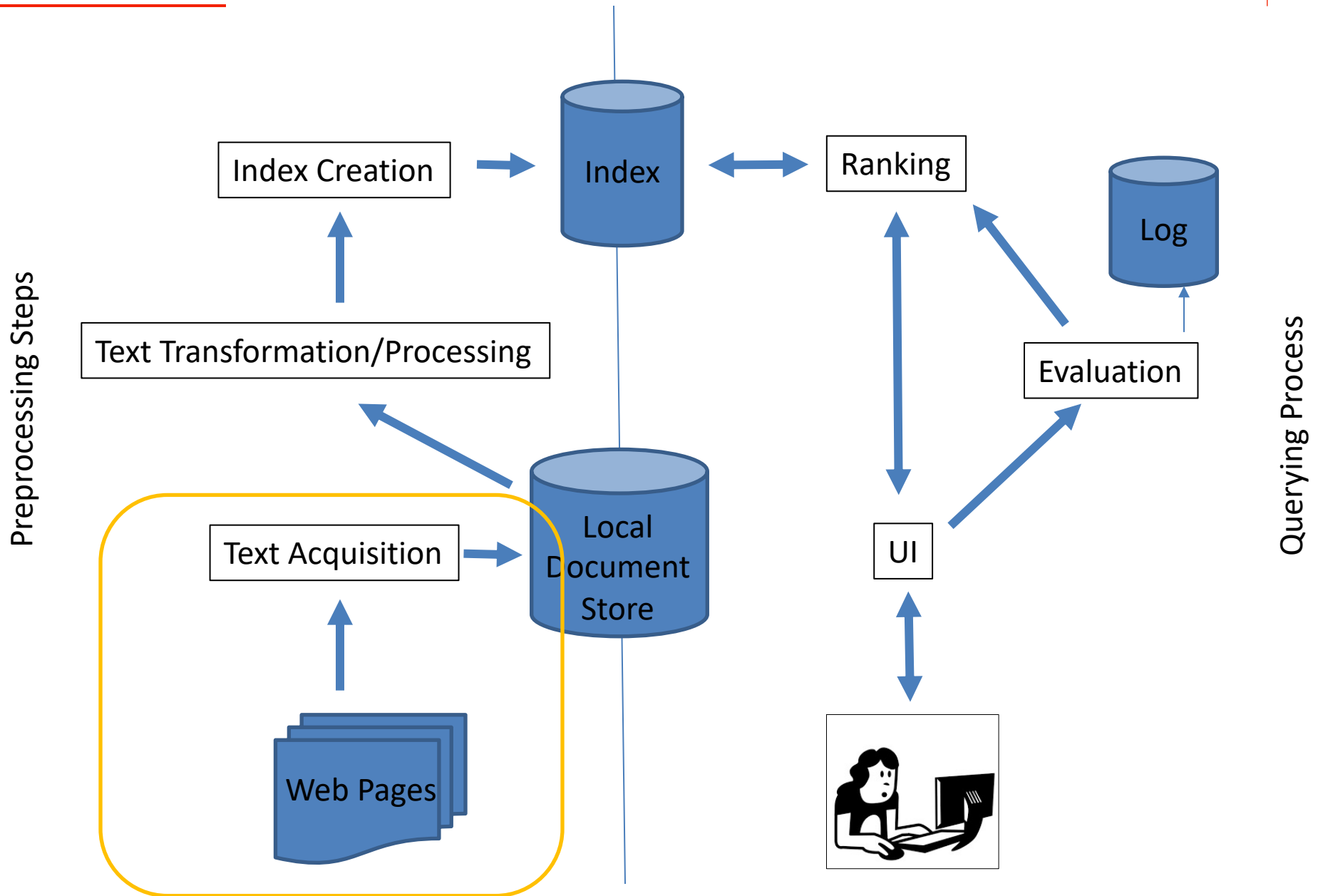
- Solution :

**Optimizing for users will guarantee to optimize for good search engines**

# Acquiring data from the Web

Information Retrieval

# Architecture



# Universal Resource Identifiers


- Universal Resource Identifier (URI)
  - DEF: A string of characters used to identify a resource
- Examples of URIs:
  - <http://www.ics.uci.edu> (URL)
  - ISBN 0-486-27777-3 (URN)
  - <ftp://ftp.ics.uci.edu> (URL)

# Universal Resource Identifiers

- Universal Resource Identifier (URI)
  - DEF: A string of characters used to identify a resource
- Examples of URIs:
  - <http://www.ics.uci.edu> (URL)
  - ISBN 0-486-27777-3 (URN)
  - <ftp://ftp.ics.uci.edu> (URL)
- URL (locator) vs URN (name)
  - Locator: must specify *where* the resource is. Name: just *what* it is.
- We are going to focus on URLs
  - But “URI” might slip in as synonym



# Anatomy of a URL

- Syntax:
  - scheme://domain:port/path?query\_string#fragment\_id
    - authority*
  - *(the entire picture is slightly more complicated than this)*
- Full spec:
  - <http://www.w3.org/Addressing/URL/url-spec.txt>

# Anatomy of a URL

- Syntax: **mandatory** **optional**

- `scheme://domain:port/path?query_string#fragment_id`  
*authority*

- (slightly more complicated than this)

- Full spec:

- <http://www.w3.org/Addressing/URL/url-spec.txt>

# Anatomy of a URL

- `http://www.ics.uci.edu/~lopes`

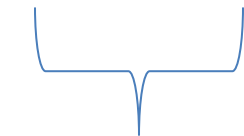
*on a web  
server*

*no port!  
just domain*

*path*

- Domains and subdomains:

– `calendar.ics.uci.edu`



Domain name

# Anatomy of a URL

- `http://www.ics.uci.edu/~lopes`

*on a web  
server*

*no port!  
just domain*

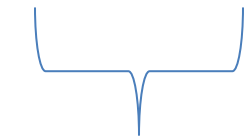
*path*

*query*

- `http://calendar.ics.uci.edu/calendar.php?type=month&calendar=1&category=&month=02&year=2013`

- Domains and subdomains:

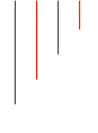
– `calendar.ics.uci.edu`



Domain name

# Different Flavors of Web Data Collection

---



- How to acquire data?
  - Data dumps
  - URL downloads
  - Web APIs
  - Web Crawling

# Data dumps

---

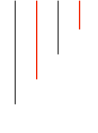
- Sites may package their data periodically and provide it as a “dump”
  - Example: [Wikipedia](#) (it suggests you to Torrent)
  - arXiv Bulk Full-Text Access: [https://arxiv.org/help/bulk\\_data\\_s3](https://arxiv.org/help/bulk_data_s3)

# URL Downloads

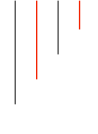
- Two step process:
  1. Find out the URLs of specific resources
  2. Run a downloader that takes that list and downloads the resources
- Example: “crawling” (!) sourceforge / github for source code
- Doesn't need to be source code; can be papers, pages, etc.
  - [http://link.springer.com/chapter/10.1007/978-3-642-34213-4\\_1](http://link.springer.com/chapter/10.1007/978-3-642-34213-4_1)
  - [http://link.springer.com/chapter/10.1007/978-3-642-34213-4\\_2](http://link.springer.com/chapter/10.1007/978-3-642-34213-4_2)
  - ...
- Some sites use regular URLs. E.g. Google Code
  - <http://code.google.com/p/python-for-android/downloads/list>
  - ...

- Sites may provide (REST) interfaces for getting their data
  - Usually higher-level: avoids having to parse HTML
  - Usually restrictive: only part of the data
- Examples:
  - [Facebook Graph API](#)
    - [My data in facebook api](#)
    - [More examples](#)
  - [Youtube API](#)
  - [Twitter API](#)
  - [arXiv API](#)
  - ...





- Like people, getting HTML pages and other documents and discovering new URLs as it goes
  - Good for **changing** collections
  - Good for **unknown** documents



- Like people, getting HTML pages and other documents and discovering new URLs as it goes
  - Good for changing collections
  - Good for unknown documents
- Web admins don't like crawlers
  - Crawlers consume resources that are meant for people
  - More on this later...

Next class!

# Algorithm?

---

- Knuth's definition of algorithm
  - **Finiteness** : *does it ends after a finite number of steps?*
  - **Definiteness** : *are the steps rigorously and unambiguously specified?*
  - **Input** : *does it has zero or more inputs?*
  - **Output** : *does it has one or more outputs related to the inputs?*
  - **Effectiveness** : *are the operations basic enough for someone to perform them using pencil and paper only?*