

Data **collection** some techniques

Genoveva Vargas-Solar

CR1, CNRS, LIG-LAFMIA

Genoveva.Vargas@imag.fr

<http://vargas-solar.com>, Montevideo, 21st November, 2014



HADAS
GROUP



Web scraping

<http://slides.com/myasoobkhalid/web-scraping#/32>

Web scraper

3

- Any program that retrieves structured data from the web, and then transforms it to conform with a different structure
- Wait, isn't that just ETL? (extract, transform, load)
- Well, sort of, but I don't want to call it that...

Web scraper

4

- “Scraping” applies to web pages, getting data from a CSV or JSON
- Why not ETL?
 - ETL implies that there are rules and expectations
 - These two things don’t exist in the world of the Web
 - They can change the structure of their dataset without telling you, or even take the dataset down on a whim.
- A program that pulls down data is often going to be a bit hacky by necessity, so “scraper” seems like a good term for that

Web scraping

- Web scraping (web harvesting or web data extraction) is a computer software technique of extracting information from websites
- Usually, such software programs simulate human exploration of the World Wide Web by either
 - Implementing low-level Hypertext Transfer Protocol (HTTP)
 - Embedding a fully-fledged web browser, such as Internet Explorer or Mozilla Firefox

Wikipedia

- *Method to extract data from a website that does not have an API or we want to extract a LOT of data which we can not do through an API due to rate limiting*
- *Through web scraping we can extract any data which we can see while browsing the web.*

What for?

- Extract product information
- Extract job postings and internships
- Extract offers and discounts from deal-of-the-day websites
- Crawl forums and social websites
- Extract data to make a search engine
- Gathering weather data

Web scraping vs. API

7

- Web Scraping is not rate limited
- Anonymously access the website and gather data
- Some websites do not have an API
- Some data is not accessible through an API

Web scraping workflow

8

- Get the website - using HTTP library
- Parse the html document - using any parsing library
- Store the results - either a db, csv, text file, etc

Libraries for parsing

- Some of the most widely known libraries used for web scraping are:
 - BeautifulSoup
 - lxml
 - re
 - Scrapy (a complete framework)

Parsing libraries

- BeautifulSoup

- `tree = BeautifulSoup(html_doc)`
- `tree.title`

- lxml

- `tree = lxml.html.fromstring(html_doc)`
- `title = tree.xpath('/title/text()')`

- re

- `title = re.findall('<title>(.*?)</title>', html_doc)`

BeautifulSoup

11

- A beautiful API
 - `soup = BeautifulSoup(html_doc)`
 - `last_a_tag = soup.find("a", id="link3")`
 - `all_b_tags = soup.find_all("b")`
- very easy to use
- can handle broken markup
- purely in Python
- slow :(

lxml

12

The lxml XML toolkit provides Pythonic bindings for the C libraries libxml2 and libxslt without sacrificing speed

- very fast
- not purely in Python
- If you have no "pure Python" requirement use lxml
- lxml works with all python versions from 2.4 to 3.3

- re is the regex library for Python. It is used only to extract minute amount of text
- Entire HTML parsing is not possible with regular expressions
- However it is
 - purely baked in Python
 - a part of standard library
 - very fast - I will show later
 - supports every Python version

Steps to writing a scraper

- Find the data source
- Find the metadata
- Analysis (verify the primary key): should also include noting which fields should be lookup fields
- Develop
- Test: is always done on real data and has three phases:
 - dry run (nothing added or updated),
 - dry run with lookups (only lookups are added),
 - production run: run all three phases on a local instance before deploying to production
- Fix (repeat ∞ times)

Storing scraped data

15

- Do not create tables before you understand how you want to use the data
- Consider using a non-relational DB
- See Adrian Holovaty's talk on how EveryBlock avoided creating new tables for each dataset
 - <http://bit.ly/YI6VAZ> (relevant part starts at 7:10)

Components of a scraping system

16

- Downloader
- Cacher
 - Caching is essential when scraping a dataset that involves a large number of HTML pages
 - Test runs can take hours if you're making requests over the network
 - A good caching system pretty prints the files it downloads so you can more easily inspect them
- Raw item retriever
- Existing item detector
- Item transformer
- Status reporter:
 - Reporting is essential if you're managing a group of scrapers.
 - Since you KNOW that at least one of your scrapers will be broken at any time, you might as well know which ones are broken.
 - A good reporting mechanism shows when your scrapers break, as well as when the dataset itself has issues (determined heuristically)

Scraping at scale

17

- You want to scrape millions of web pages everyday
- You want to make a broad scale web scraper
- You want to use something that is thoroughly tested
- Is there any solution ?

Scrapy (<http://scrapy.org>)

18

- Application framework for writing web spiders that crawl web sites and extract data from them
 - Scrapy only supports Python 2.7 and NOT 3.x
 - It's a tested framework
 - It's asynchronous
 - It's easy to use
 - It has everything you need to start scraping



Types of scrapers according to sources

Some tools

Main types of scrapers

- CSV
- RSS/Atom
- JSON
- XML
- HTML crawler
- Web browser
- PDF
- Database dump
- GIS
- Mixed

CSV

21

- Import csv
- You should usually use `csv.DictReader`
- If the column names are all caps, consider making them lowercase.
- Watch out for CSV datasets that do not have the same number of elements on each row

CSV

```
def get_rows(csv_file):  
    reader = csv.reader(open(csv_file))  
    # Get the column names, lowercased.  
    column_names = tuple(k.lower() for k in  
        reader.next())  
    for row in reader:  
        yield dict(zip(column_names, row))
```

XML

23

- `import lxml.etree`
- Get rid of namespaces in the input document. <http://bit.ly/L05x7H>
- A lot of XML datasets have a fairly flat structure
- In these cases, convert the elements to dictionaries

XML

24

```
<root>
<items>
  <item>
    <id>3930277-ac</id>
    <name>Frodo Samwise</name>
    <age>56</age>
    <occupation>Tolkien scholar</occupation>
    <description>Short, with hairy feet</description>
  </item>
... </items>
</root>
```

```
import lxml.etree
tree = lxml.etree.fromstring(SOME_XML_STRING)
for el in tree.findall('items/item'):
    children = el.getchildren()
    # Keys are element names.
    keys = (c.tag for c in children)
    # Values are element text contents.
    values = (c.text for c in children)
    yield dict(zip(keys, values))
```


HTML

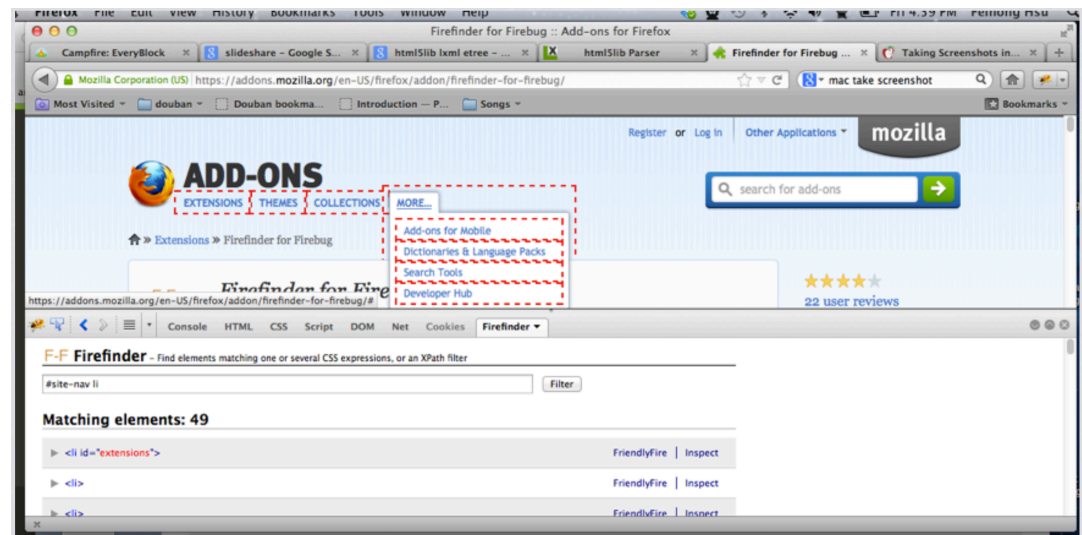
25

- `import requests`
- `import lxml.html`
- Use XPath, but pyquery seems fine too
- If the HTML is very funky, use `html5lib` as the parser
- Sometimes data can be scraped from a chunk of JavaScript embedded in the page

HTML

26

- Firefinder (<http://bit.ly/kr0UOY>) Extension for Firebug
- Allows you to test CSS and XPath expressions on any page, and visually inspect the results.



HTTP libraries

- Requests

- `r = requests.get('https://www.google.com').html`

- urllib and urllib2

- `html = urllib2.urlopen('http://python.org/').read()`

- httplib and httplib2

- `h = httplib2.Http(".cache")`

- `(resp_headers, content) = h.request("http://example.org/", "GET")`

PDF

28

- There are no Python libraries that handle all kinds of PDF documents in the wild
- Use the `pdftohtml` command to convert the PDF to XML
- When debugging, use `pdftohtml` to generate HTML that you can inspect in the browser
- If the text in the PDF is in tabular format, you can group text cells by proximity

The “group by proximity” strategy works like this:

- 1. Find a text cell that has a very distinct pattern (probably a date cell)
This is your “anchor”
- 2. Find all cells that have the same row position as the anchor
(possibly off by a few pixels)
- 3. Figure out which grouped cells belong to which fields based on
column position

RSS/Atom

30

- `import feedparser`
- Sometimes `feedparser` can't handle custom fields, and you'll have to fall back to `lxml.etree`
- Unfortunately, plenty of RSS feeds are not compliant XML
 - Either do some custom munging or try `html5lib`

youtube-dl (<http://rg3.github.io/youtube-dl/>)

31

- Python script that allows you to download videos and music from various websites like :
 - Facebook,
 - YouTube
 - Vimeo
 - Dailymotion
 - Metacafen and almost 300 more !

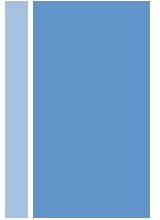
Design patterns

32

- If a field contains a finite number of possible values, use a lookup table instead of storing each value
- Make a scraper superclass that incorporates common scraper logic
- The scraper superclass will probably have convenience methods for converting dates/times, cleaning HTML, looking for existing items, etc. It should also incorporate the caching and reporting logic

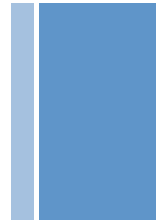
Web crawling

Motivation



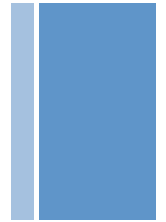
A key motivation for designing Web crawlers has been to retrieve Web pages and add their representations to a local repository

Web Crawling



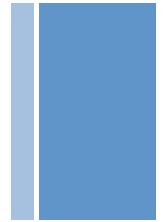
- A Web crawler (*also known as a Web spider, Web robot, or—especially in the FOAF community—Web scutter*) is a program or automated script that browses the World Wide Web in a
 - methodical
 - automated manner
- Other less frequently used names for Web crawlers are ants, automatic indexers, bots, and worms.

Crawlers



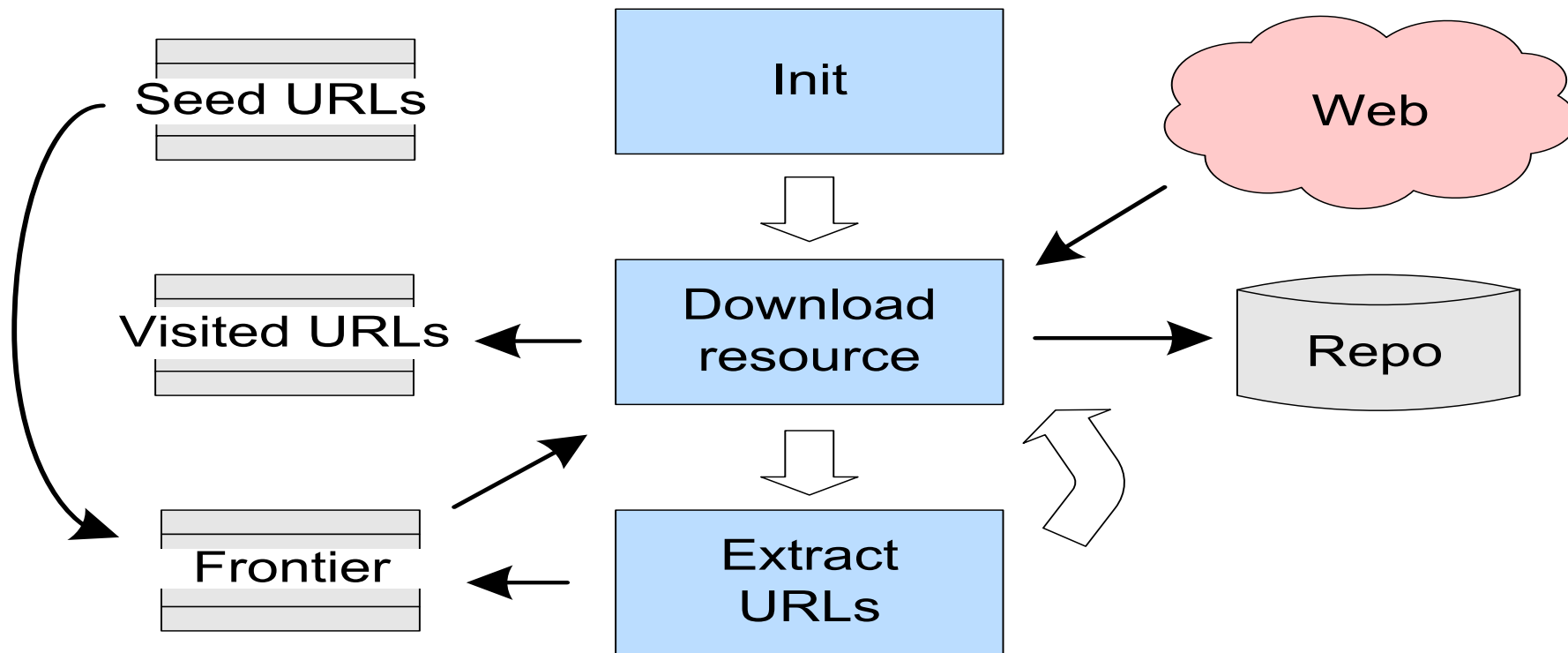
- Computer programs that roam the Web with the goal of automating specific tasks related to the Web
- The role of Crawlers is to collect Web Content

Basic crawler operation

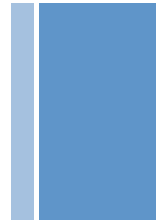


- Begin with known “seed” pages
- Fetch and parse them
- Extract URLs they point to
- Place the extracted URLs on a Queue
- Fetch each URL on the queue and repeat

Traditional Web Crawler



Web crawler: basic algorithm



```
{  
    Pick up the next URL  
    Connect to the server  
    GET the URL  
    When the page arrives, get its links  
    (optionally do other stuff)  
    REPEAT  
}
```

Uses

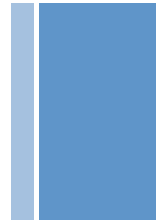


- Complete web search engine

Search Engine = **Crawler** + Indexer/Searcher /(Lucene)
+ GUI

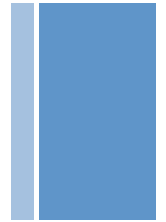
- Find stuff
- Gather stuff
- Check stuff

Types of Crawlers



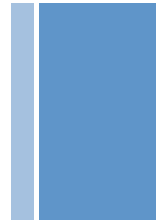
- **Batch** : Crawl a snapshot of their crawl space, until reaching a certain size or time limit
- **Incremental** : Continuously crawl their crawl space, revisiting URL to ensure freshness
- **Focused**: Attempt to crawl pages pertaining to some topic/theme, while minimizing number of off topic pages that are collected

URL normalization



- Crawlers usually perform some type of URL normalization in order to avoid crawling the same resource more than once.
- The term *URL normalization* refers to the process of
 - modifying
 - standardizing
 - a URL in a consistent manner

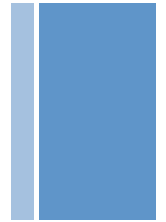
The challenges of « Web Crawling »



Three characteristics of the Web that make crawling it very difficult:

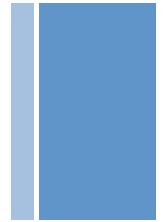
- Its large volume
- Its fast rate of change
- Dynamic page generation

Examples of Web crawlers



- RBSE
- World Wide Web Worm
- Google Crawler
- WebFountain
- WebRACE

Web 3.0 Crawling



Web 3.0 defines advanced technologies and new principles for the next generation search technologies that is summarized in

- Semantic Web

- Website Parse Template concepts

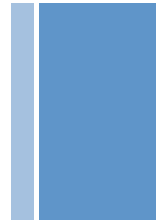
Web 3.0 crawling and indexing technologies will be based on

- Human-machine clever associations

How Web API are used ?

- Series or collection of web services
- Sometimes used interchangeably with “web services”
- Examples: Google API, Amazon.Com APIs

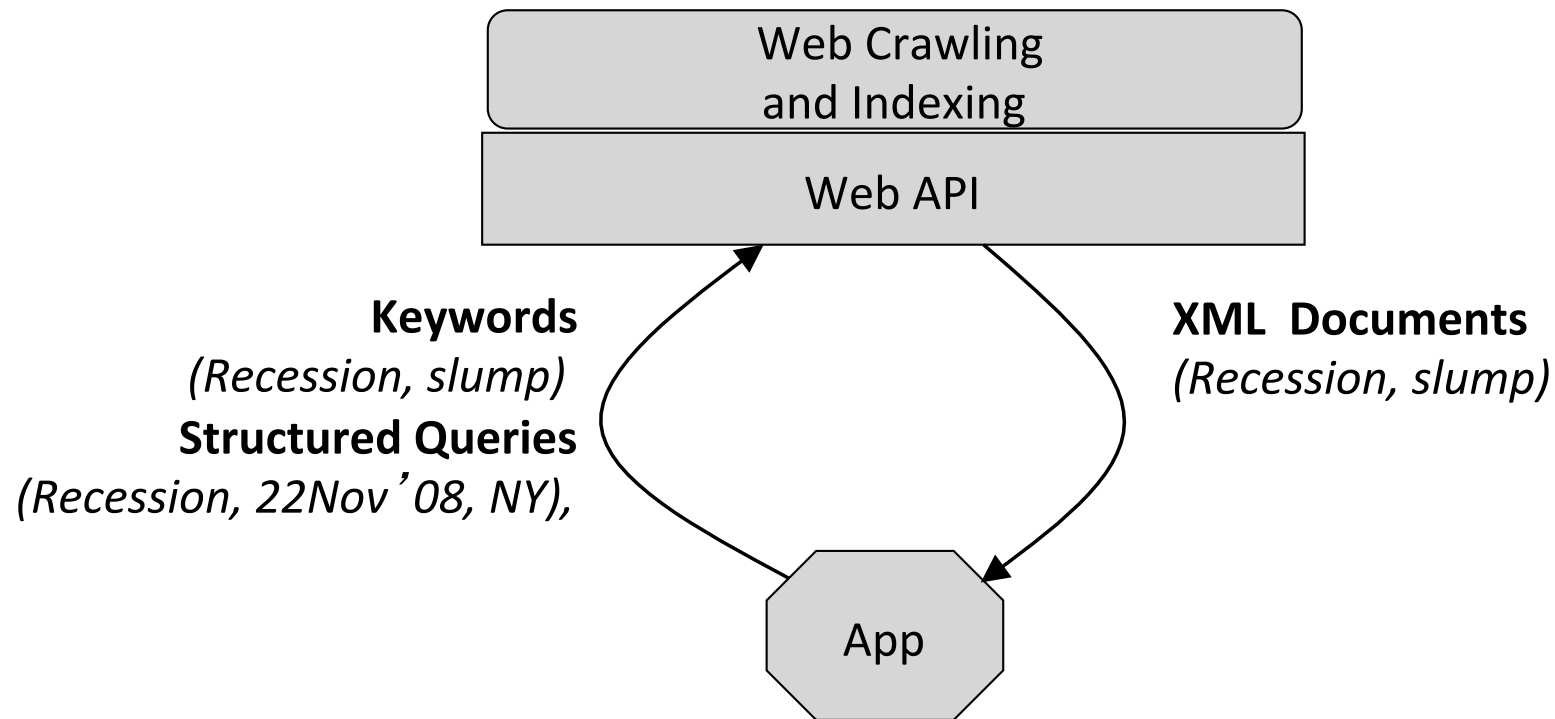
How Do You Call a Web API?



XML web services can be invoked in one of three ways:

- Using REST (HTTP-GET)
 - URL includes parameters
 - Example: “ `http://search.twitter.com/search.atom?q=` “
- Using HTTP-POST
 - You post an XML document
 - XML document returned
- Using SOAP
 - More complex, allows structured and type information

APIs that deliver information



References



- http://en.wikipedia.org/wiki/Web_crawling
- www.cs.cmu.edu/~spandey
- www.cs.odu.edu/~fmccown/research/lazy/crawling-policies-ht06.ppt
- <http://java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler/>
- www.grub.org
- www.filesland.com/companies/Shettysoft-com/web-crawler.html
- www.ciw.cl/recursos/webCrawling.pdf
- www.openldap.org/conf/odd-wien-2003/peter.pdf

Fansourcing

Open Sourcing

Crowdcasting

Wikinomics

Crowdsourcing

Collective Intelligence

Mass Collaboration

Collective Customer Commitment

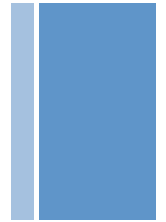
Open Innovation



Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call

"Crowdsourcing" - The term was coined by Jeff Howe in Wired Magazine in 2006 [3](#)

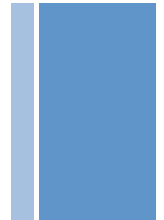
Wisdom of the Crowds



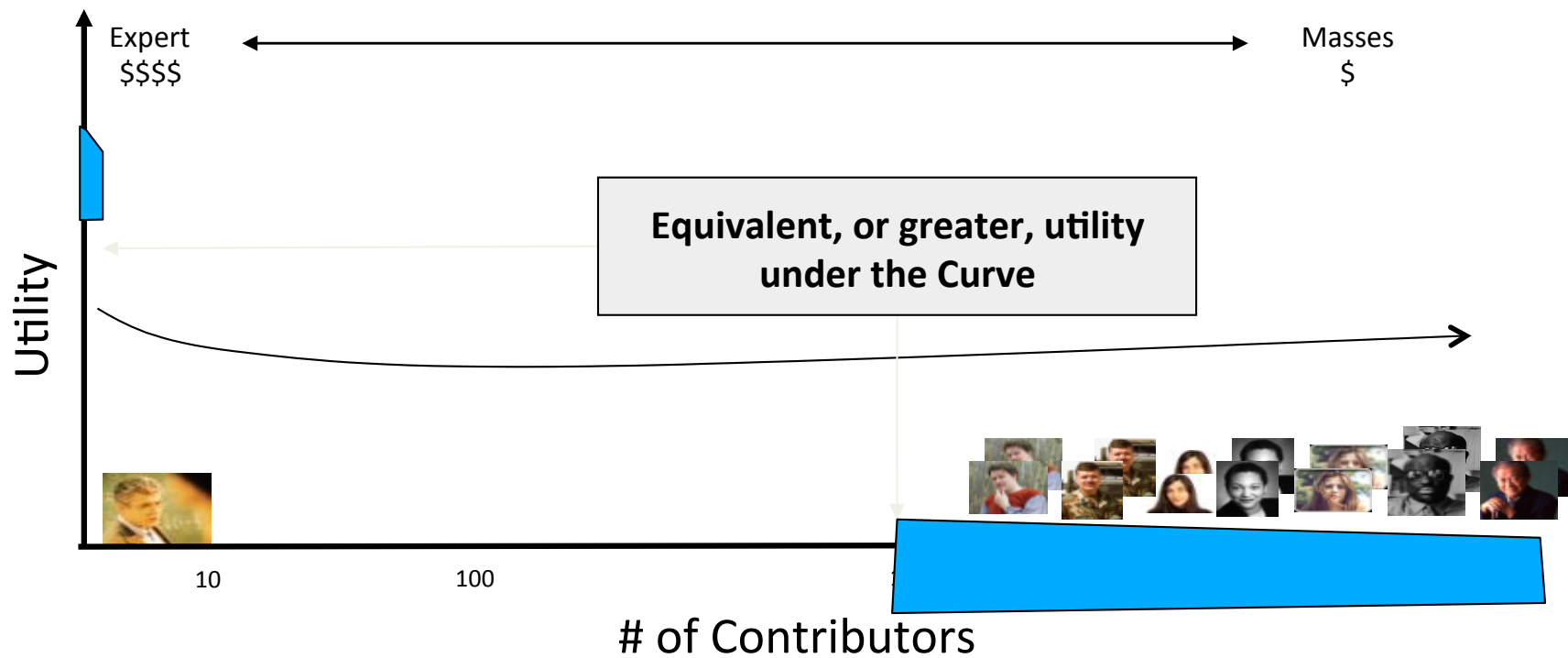
- The crowd at a county fair accurately guessed the weight of an ox when their individual guesses were averaged
- Average
 - Closer to the ox's true butchered weight than the estimates of most crowd members, and also
 - Closer than any of the separate estimates made by cattle experts

Wikinomics 101

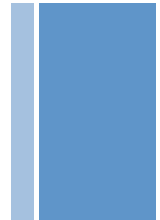
Wisdom of the Crowds



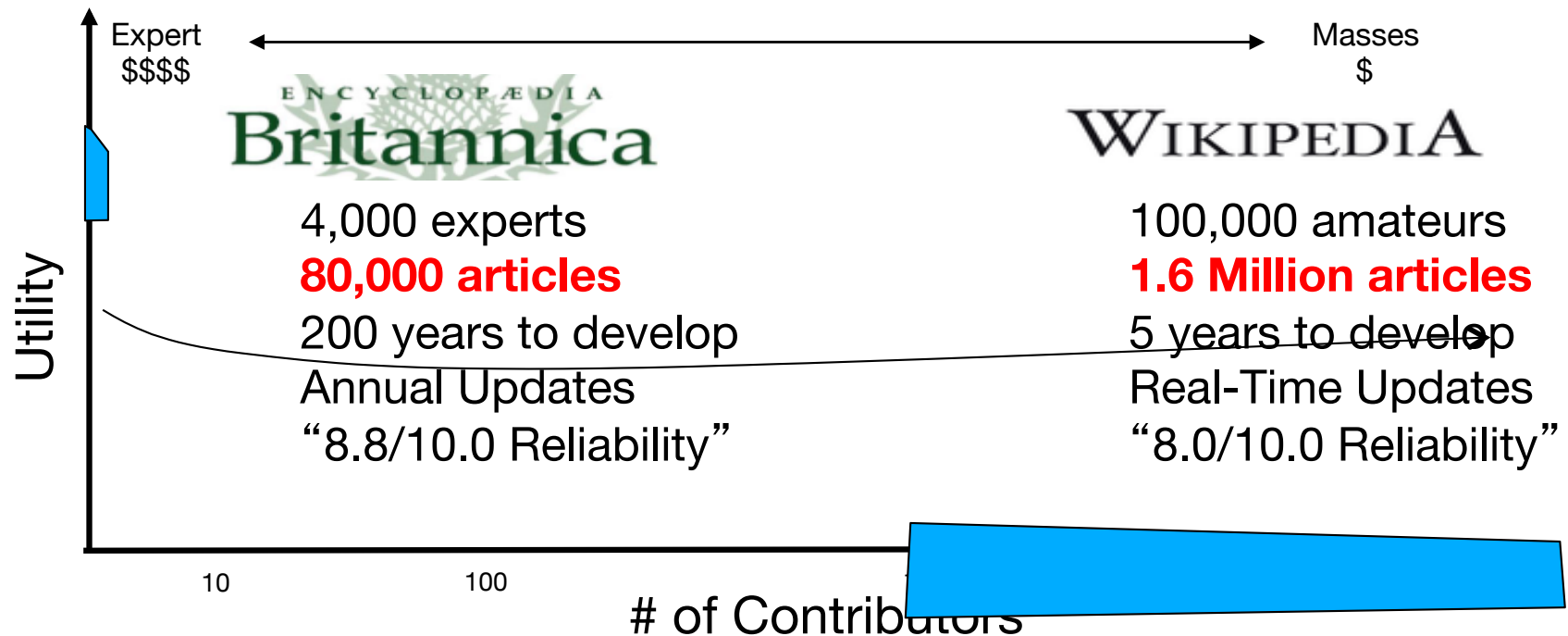
Enterprise Web 2.0



Economics & Wikinomics



Enterprise Web 2.0



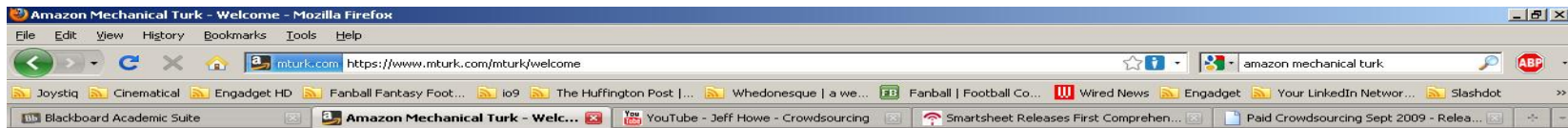
What is crowdsourcing?

- Crowdsourcing is an online, distributed problem solving and production model
 - **Users**--also known as the crowd--typically form into online communities based on the Web site, and the crowd submits solutions to the site or produce its contents
 - The crowd can also sort through the solutions, finding the best ones
 - These best solutions are then owned by the entity that broadcast the problem in the first place--the **crowdsourcer**
- The winning individuals in the crowd are sometimes rewarded
- Many individuals in the crowd participate just for intellectual stimulation or because of emotional ties to product or service

Benefits of Crowdsourcing to Companies!



- Problems can be explored at comparatively little cost
- Payment is by results
- The organization can tap a wider range of talent than might be present in its own organization
- Turn customers into designers
- Turn customers into marketers



amazonmechanicalturk
Artificial Artificial Intelligence

[Your Account](#) [HITS](#) [Qualifications](#)

[Introduction](#) | [Dashboard](#) | [Status](#) | [Account Settings](#)

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.
46,185 HITS available. [View them now.](#)

Make Money by working on HITS

HITS - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITS now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITS - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Register Now](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITS completed in minutes
- Pay only when you're satisfied with the results



or [learn more about being a Requester](#)

Amazon Mechanical Turk - Mozilla Firefox

File Edit View History Bookmarks Tools Help

mturk.comhttps://www.mturk.com/mturk/searchbar?selectedSearchType=hitgroups&requesterId=AO1R9DU8M1ACHamazon mechanical turk

Joystiq Cinematical Engadget HD Fanball Fantasy Foot... io9 The Huffington Post |... Whedonesque | a we... Fanball | Football Co... Wired News Engadget Your LinkedIn Networ... Slashdot

Blackboard Academic Suite Amazon Mechanical Turk YouTube - Jeff Howe - Crowdsourcing Smartsheet Releases First Comprehen... Paid Crowdsourcing Sept 2009 - Relea...Sign In

amazonmechanicalturk
Artificial Intelligence

Your AccountHITSQualifications46,048 HITS available now

All HITS | HITS Available To You | HITS Assigned To You

Search for containing that pay at least \$ for which you are qualified

HITS Created by 'AboutUs Mechanical Turk Operators'

1-2 of 2 Results

Sort by: [Show all details](#) | [Hide all details](#)

Summarize a website in a sentence - Earn BONUS of 5 centsView a HIT in this group

Requester: [AboutUs Mechanical Turk Operators](#)

HIT Expiration Date: Oct 14, 2009 (3 weeks 5 days)

Reward: \$0.05

Time Allotted: 30 minutes

HITS Available: 5913

Description: Visit the given website and write a short summary that uniquely describes its purpose

Keywords: [summary](#), [websites](#)

Qualifications Required:
AUQualifiedSummaryWriter is 1
HIT approval rate (%) is not less than 85
Adult Content Qualification is 1

Review short summary of a websiteView a HIT in this group

Requester: [AboutUs Mechanical Turk Operators](#)

HIT Expiration Date: Oct 2, 2009 (1 week 6 days)

Reward: \$0.05

Time Allotted: 30 minutes

HITS Available: 1

Description: Visit the given website and review its short summary. Fix the summary if it needs improvement

Keywords: [review](#), [websites](#)

Qualifications Required:
AUQualifiedSummaryWriter is 1
HIT approval rate (%) is not less than 85
Adult Content Qualification is 1
AUTrusted is 1

FAQ | [Contact Us](#) | [Careers at Amazon](#) | [Developers](#) | [Press](#) | [Policies](#)

©2005-2009 Amazon.com, Inc. or its Affiliates

An [amazon.com](#) company

Done

RentACoder



Software Buyers

Need custom software? Receive bids from our pool of 268,624 registered coders. Review work histories and resumes online, and then conduct business stress-free using our "[Safe Project Escrow](#)"(tm).

[How Does It Work for Buyers?](#)

[Request bids on your project or problem](#)

Software Coders

Earn cash with your high tech skills on 2,502 currently open bid requests. Then subscribe to our newsletter and receive daily bid requests from our 126,220 registered buyers.

[How Does It Work for Coders?](#)

[Register and be notified of new projects!](#)

Newest Bid Requests

From [2502 open bid requests](#).

[PSD to Xhtml](#)

By Art Studios Online
on Sep 20
Max Bid: \$20.00

[GDSN database using Altova](#)

By Creative launch on
Sep 20

[Put this ticker on your site](#)

Quotes

"After a visit to
RentACoder.com...I got
something I needed for a
fifth of the price I would
otherwise have had to pay."

the guardian

[...Other quotes...](#)

Returning Users:

[Login](#)

Buyers: [My Bid Requests](#) | [My buyer financials](#)

Coders: [My Bids](#) | [My coder financials](#)

Mode: Hi-resolution ([Show low-resolution version of this page](#))

2007, 2008 winner of:

Inc. 5000

Written about in:

FAST COMPANY

And [elsewhere...](#)

Click here to put this
ticker on your own site
and/or get live RSS
newsfeeds

I need someone in the Philippines create account at a certain website.

**adding functionality to
xhtml site(repost)**

[uoielh](#)
(29 ratings)

Small Business Project: \$100(USD) and above

Web, Database, Language Specific, PHP, MySQL, Javascript, Software
Related (Includes Websites)

 21 since
Bidding open Sep 18, 2009
Max bid: Open to fair
suggestions 9:22:58 PM EDT


Attached are the requirements and copy for the functionality and the xhtml. build it on your dev. site, and then i'll have you
deploy to the final host site Would like the site to be built in php the xhtml (which is at ... (see bid request for full description)

**FLASH CODER TO EMBED
FLASH PLAYER INTO MY
VIDEO**

[Nick O](#)
(15 ratings)

Very Small Business Project: under \$100(USD)

Graphic Design / Art / Music, Video Editing

 49 since
Bidding open Sep 18, 2009
Max bid: \$30.00 (USD)
(3)

I need a FLV player embeded with my short video. The video player must fit around the small video properly. I might need help
with installing the video onto my website. I use XSitePro Version 2 to build my website. When the web ... (see bid request for full
description)

**Convert Simple C++ COM
examples to use Vole
COM library**

[infiniteidea](#)
(43 ratings)

Very Small Business Project: under \$100(USD)

C++ / C

 56 since
Bidding open Sep 18, 2009
Max bid: \$70.00 (USD)
(3)

Attached are simple C++ examples that call a COM library/ The COM library is the Order2Go library from FXCM
(http://www.fxprogrammers.com) and will be provided to the winning bidder.We want all COM calls to use VOLE at
http://vole.sourceforge ... (see bid request for full description)

PHP coding

[forrestrunden](#)
(28 ratings)

Small Business Project: \$100(USD) and above

Web, Flash, Page / Site Design, Database, Operating Systems /
Platforms, Web Services, Linux, SQL Server, MySQL, Other
(Database), Other (Web), XML / XHTML, Software Related (Includes
Websites), Other, SQLite

 91 since
Bidding open Sep 18, 2009
Max bid: \$100.00 (USD)
(3)

I need a php script to do the following:I do not need design work.Here is how the site will work.Click here to see the concept
page.http://exampreview.com/1.html1. Step 1 (One) A person will enter their email and be ... (see bid request for full description)

**php, mysql, css and ajax
expert**

[inat22](#)
(250 ratings)

Very Small Business Project: under \$100(USD)

Software Related (Includes Websites)

 70 since
Bidding open Sep 18, 2009
Max bid: \$50.00 (USD)
(3)

I'm looking for someone that can install 2 existing php scripts in my server and have them work with my existing code and
database.I will also need you to verify my existing code as there appears to be a small bug in the updating of data (I don' ...
(see bid request for full description)

UP NEXT: Please click on the image of Project Price or Beginner Assistance

 56 since

[View All Bidders by
ranking](#)

[Top Expert Rating Exam
Scorers](#)

and Spun for Article
Submit...

By Success Systems
on Sep 18
Max Bid: \$40

Create account for
certain website
By Zigzagzilar on
Sep 18
..

Add to Active Desktop

Click here to put this
ticker on your own site
and/or get live RSS
newsfeeds

Description:

- Rent A Coder reminder: You MAY NOT post the final solution for this (and any) project before your bid is accepted and funds are fully escrowed. Anyone who does may have their account permanently suspended. However, you CAN post:
 - On programming projects: A prototype or functional demo...as long as source code is not provided.
 - On graphics projects: A watermarked and low-resolution version of the work.

I need someone to create a custom thread manager / thread pool class (or tell me how to do this correctly) that will suit my application requirements. Read below for my problem. If you can help, and are available for IM discussion while working - please bid.

I have a vb.net application that is having some problems when stopping threads and creating new threads.

I have an array of 100 objects (client() as ClientClass). This is created in the main form. The object class has a timer that does web requests every 3 seconds.

I have created an array of 100 threads also created in the main form to begin each instance of these objects above (t() as Thread).

When I want to start the instances, I create a new object and a new thread with the address of the "start" sub in the class.

```
client(x) = New ClientClass  
t(x) = New Thread(AddressOf client(x).Start)  
t.Start()
```

My problem is that when trying to STOP and START any specific index (x), the application, including the IDE hang completely with NO error information. However, this is random.. I cannot always reproduce it. Sometimes it takes 8 hours sometimes it takes 2 minutes.

When I stop the process (which takes a long time) I get the error "Unable to break execution" from VS2008.

I THINK (only think!) the problem happens when I am restarting threads so I am assuming it is a threading issue and I need a better way to manage threads.

Otherwise it must be something else and I do not know how to locate this error.



Platform:

vb.net



Deliverables:

- 1) Complete and fully-functional working program(s) in executable form as well as complete source code of all work done.
- 2) Deliverables must be in ready-to-run condition, as follows (depending on the nature of the deliverables):

Learn
9.79 avg. over 278
jobs.

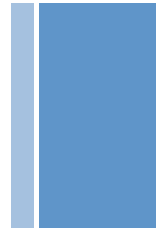
9) Atik
9.75 avg. over 217
jobs.

10) Small Software
Consultant
9.94 avg. over 458
jobs.

...See ALL coders by
ranking

Top Expert Rating Exam
Scorers

Crowdsourcing: the benefits



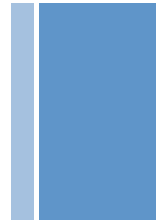
■ Companies Get ⁵

- Improved quality and productivity
- Feedback
- Good Exposure
- Minimum of Cost

■ People Get ⁶

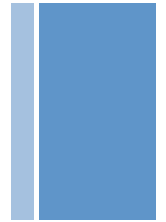
- Incentive
 - Cash Cash Cash
- Recognition
 - Sense of accomplishment among peers
- Make Life Better
 - Linux
 - Obama Campaign

Problems with Crowdsourcing



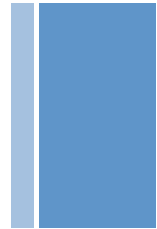
- Quality
- Intellectual property leakage
- No time constraint
- Not much control over development or ultimate product
- Ill-will with own employees
- Choosing what to crowdsource & what to keep in-house

Type of problems to outsource



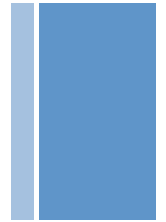
- No internal expertise
- Non-essential and non-critical
- One that has no time constraint
- One that benefits from crowd involvement
- One-time problems

Some Applications of Crowdsourcing



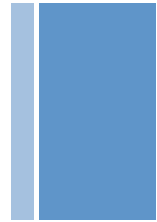
- Testing & Refining a Product
 - [Netflix](#)
 - [SellaBand](#)
- Market Research
 - [Threadless](#)
- Knowledge Management
 - [Accenture](#)
 - [Wikipedia](#)
- Customer Service
 - [My Starbucks ideas](#)
- R & D
 - [InnoCentive](#)
 - [P&G Connect & Develop](#)
- Polling and Voting
 - [InTrade](#)
 - [Building a new city](#)

Elements for a Wise Crowd!



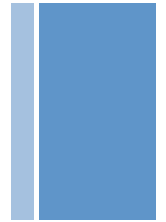
- **Diversity of opinion:** Each person should have private information even if it's just an eccentric interpretation of the known facts
- **Independence:** People's opinions aren't determined by the opinions of those around them
- **Decentralization:** People are able to specialize and draw on local knowledge
- **Aggregation:** Some mechanism exists for turning private judgments into a collective decision

Reasons to fear Crowd Intelligence



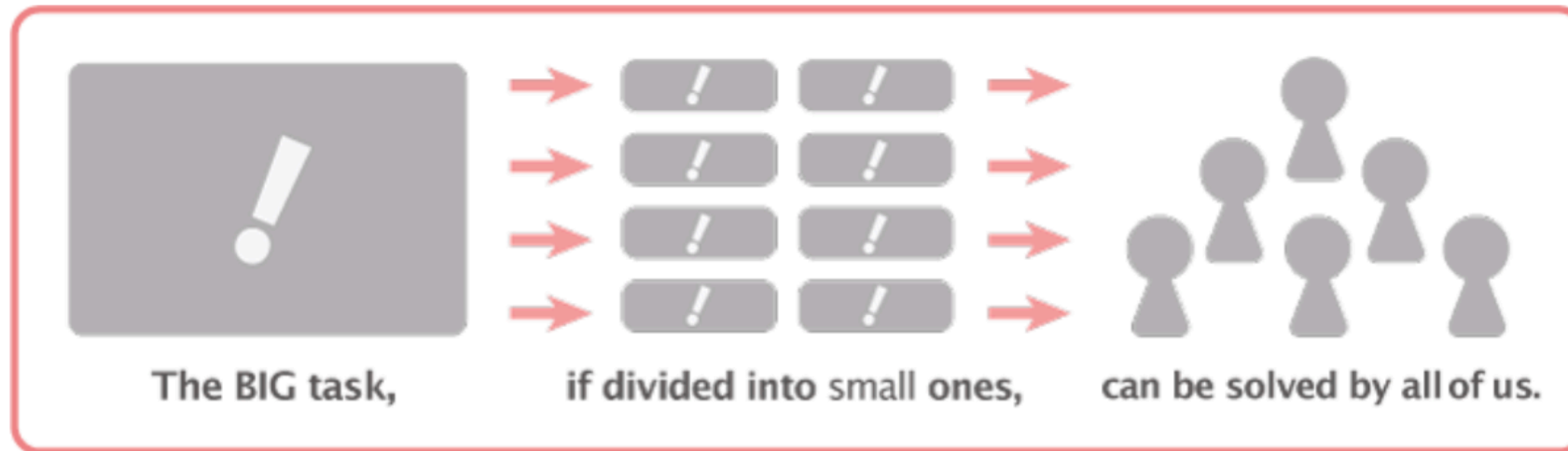
- **Too homogeneous:** The need for diversity within a crowd to ensure enough variance in approach, thought process, and private information.
- **Too centralized:** The Columbia Shuttle Disaster, hierarchical NASA management bureaucracy decision making was totally closed to the wisdom of low-level engineers
- **Too divided:** The US Intelligence community failed to prevent the September 11 attacks partly because information held by one subdivision was not accessible by another. Crowds work best when they choose for themselves what to work on and what information they need
- **Too imitative:** Where choices are visible and made in sequence, an information cascade can form in which only the first few decision makers gain anything by contemplating the choices available
- **Too emotional:** Emotional factors, such as a feeling of belonging, can lead to peer pressure and herd mentality

Conclusion:



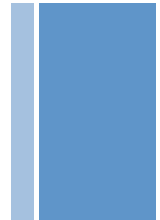
- Crowdsourcing used properly
 - Generates New Ideas
 - Cuts Development Costs
 - Creates a Direct, Emotional, bond with customers
- Used Improperly
 - Can Produce Useless Wasteful Results
 - Beware of Mob Rule

“Crowds can be wise, but they can also be stupid. “



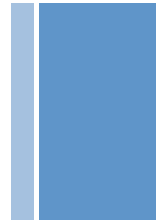
<https://crowd4u.org>

Want More Information?



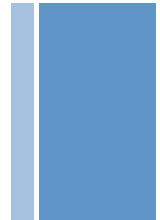
- About Crowdsourcing
 - Jeff Lowe Blog
 - www.crowdsourcing.com
 - The Rise of Crowdsourcing
 - www.wired.com/wired/archive/14.06/crowds.html
 - Paid Crowdsourcing: Current: State and Progress towards Mainstream Business Use
 - <http://www.marketwire.com/press-release/SmartsheetCom-1045951.html>

Bibliography



- Alsever, Jennifer, “What is Crowdsourcing?” www.bnet.com Mar 7th, 2007 Reliability = Good, Article summarized a lot of need to know information about Crowdsourcing as it was just becoming a topic for business.
- Lowe, Jeff Crowdsourcing Definition <http://www.crowdsourcing.com> Checked Apr 18th 2009 Reliability = Blog site of Jeff Lowe who coined the term Crowdsourcing. Site contains links and thoughts on articles in the news and feedback from speaking events.
- Lowe, Jeff “The Rise of Crowdsourcing” www.wired.com 06-Sep Reliability = Great, The original Article where the Term “Crowdsourcing” was born and talks about a few companies that are using it.
- Frei, Brent “Paid Crowdsourcing: Current State & Progress toward Mainstream Business Use” www.marketwire.com 09/16/2009 Source = Decent Whitepaper on Crowdsourcing includes timelines of adoption as well as companies that are using it and how they are using it.
- Hempel, Jessi “Crowdsourcing: Milking the Masses for Inspiration” www.businessweek.com 09/25/2006 Reliability = Good, Article talking about how to reign in the Crowdsourced Crowds.
- Abrahamson, Shaun, “What do Crowds Get from Crowdsourcing” www.mutopo.com 04/12/2009 Reliability = Decent, Article about the motivation of Crowds in Crowdsourcing
- Netflix “Frequently Asked Questions” www.netflixprize.com 10/01/2006 Reliability = Great, Official Website for Netflix Prize.
- Copeland, Michael, “Box office boffo for brainiacs: The Netflix Prize” <http://brainstormtech.blogs.fortune.cnn.com> 09/21/2009 Reliability = Good, A brief news article about the winning Netflix Prize team and some statistics.

Bibliography (Continued)



- Charles, Dan, “Internet Users Join Search For Steve Fossett” www.npr.org 09.12.07 Reliability = Great, Article talking about how the internet search for Steve Fossett started and how it was sent out to the crowds
- Barbalace, Kenneth, “Internet search for Steve Fossett eight weeks later” blog.environmentalchemistry.com 10/31/2007 Reliability = Decent, Blog Entry about the Internet Search for Steve Fossett and some future applications of the technology used.
- National Academy of Public Administration, <http://opengov.ideascale.com/> Sep 18th, 2009 Reliability = Good, The Website that was opened up for public to submit and vote on policy issues for President Obama
- Hansell, Saul, "Ideas Online, Yes, but Some Not So Presidential" www.nytimes.com 06/22/2009 Reliability = Great, News Article Talking about Policy Issues Website and Results
- Various Sources “Just Some Thoughts on the Contest” www.netflixprize.com 07/05/2009 Reliability = Good, Some feedback from the participants on why they thought the Netflix Prize was such a successful contest.
- Waltner, Charles, “I-Prize Contest Proving a Winning Approach to Discovering Billion-Dollar Business Ideas” newsroom.cisco.com 07/14/2008, Reliability = Great, Information about what the I-prize is and a small amount of information on the winning team

