Evaluation in IR

Introduction to Information Retrieval
CS 221
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org

Exercise

- If my system returns A,C,D,E to query q....

$Document$	$Relevant(q)$	$Not\ Relevant(q)$
A	$\surd$	
B		$\surd$
C		$\surd$
D	$\surd$	
E		$\surd$
F	$\surd$	

$Precision$	$\frac{1}{2}$
$Recall$	$\frac{2}{3}$
$Accuracy$	$\frac{1}{2}$

- What do I want Accuracy to be?

Exercise

- If my system returns A,C,D,E to query q....

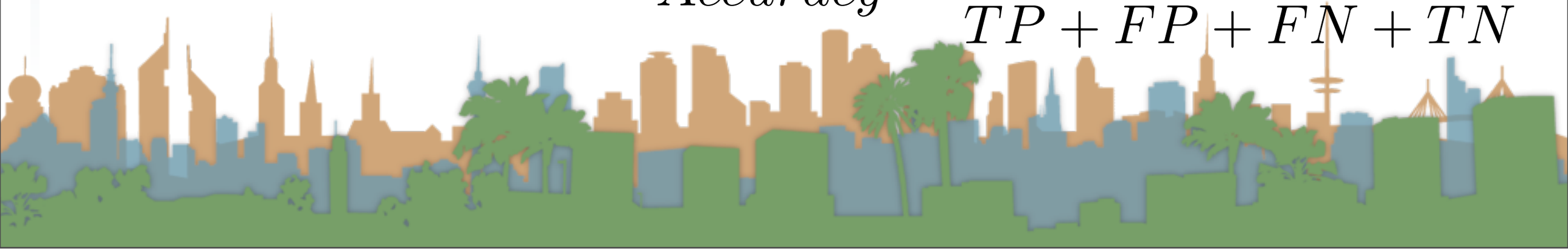
Document	*Relevant*(q)	*Not Relevant*(q)
A	√	
B		√
C		√
D	√	
E		√
F	√	

Precision	$\frac{1}{2}$
Recall	$\frac{2}{3}$
Accuracy	$\frac{1}{2}$

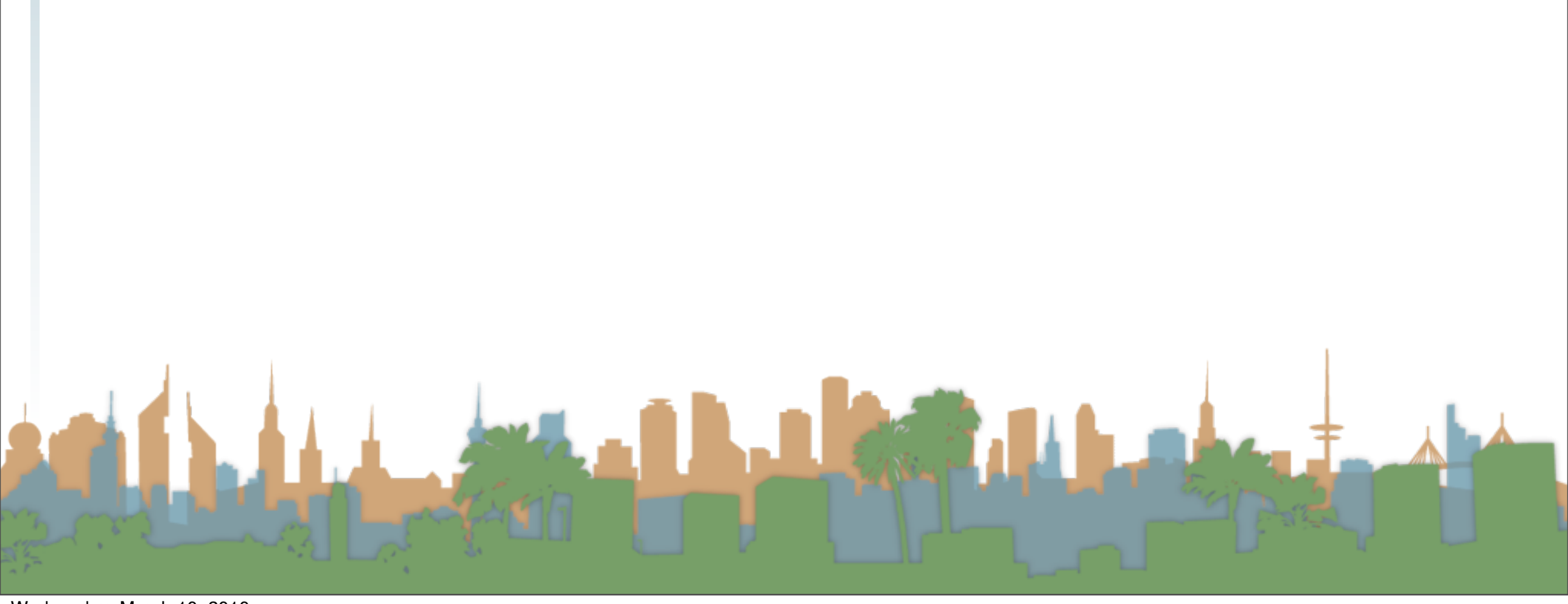
- What do I want Accuracy to be?

	Relevant	*Not Relevant*
Retrieved	*TP*	*FP*
Not Retrieved	*FN*	*TN*

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

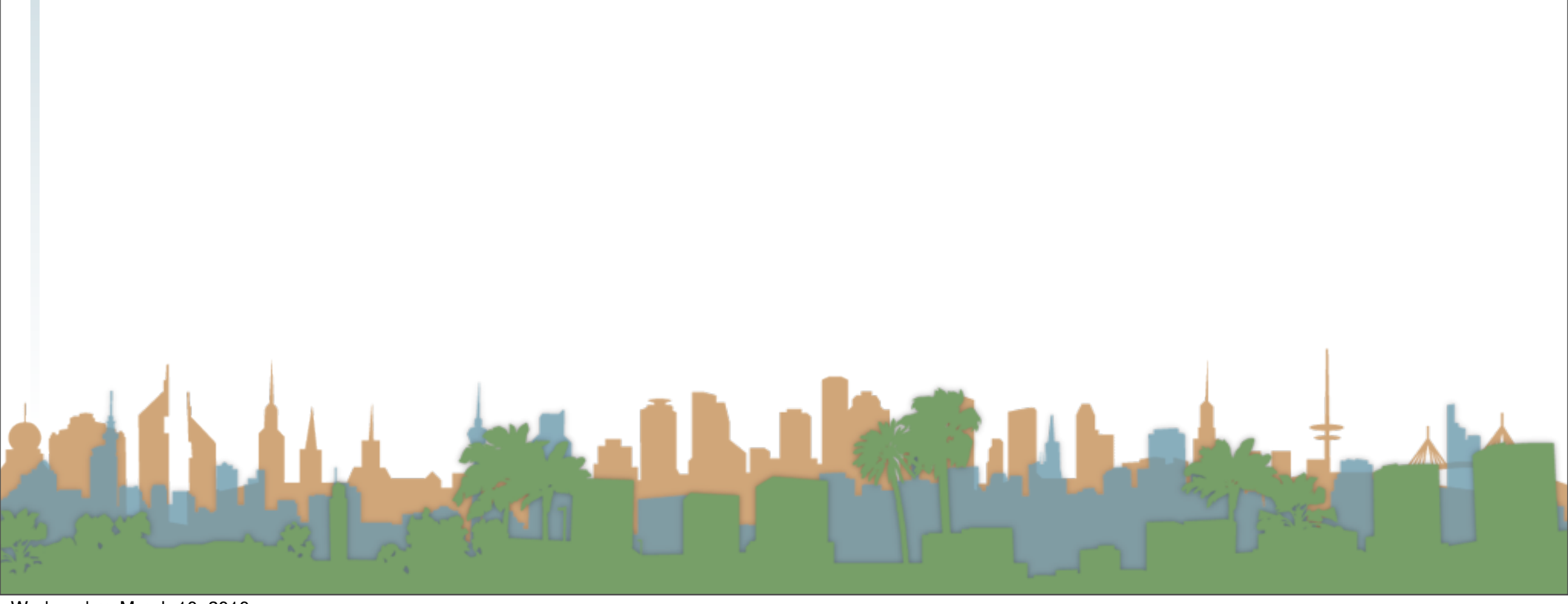


Unranked retrieval - Accuracy



Unranked retrieval - Accuracy

- Welcome to my search engine

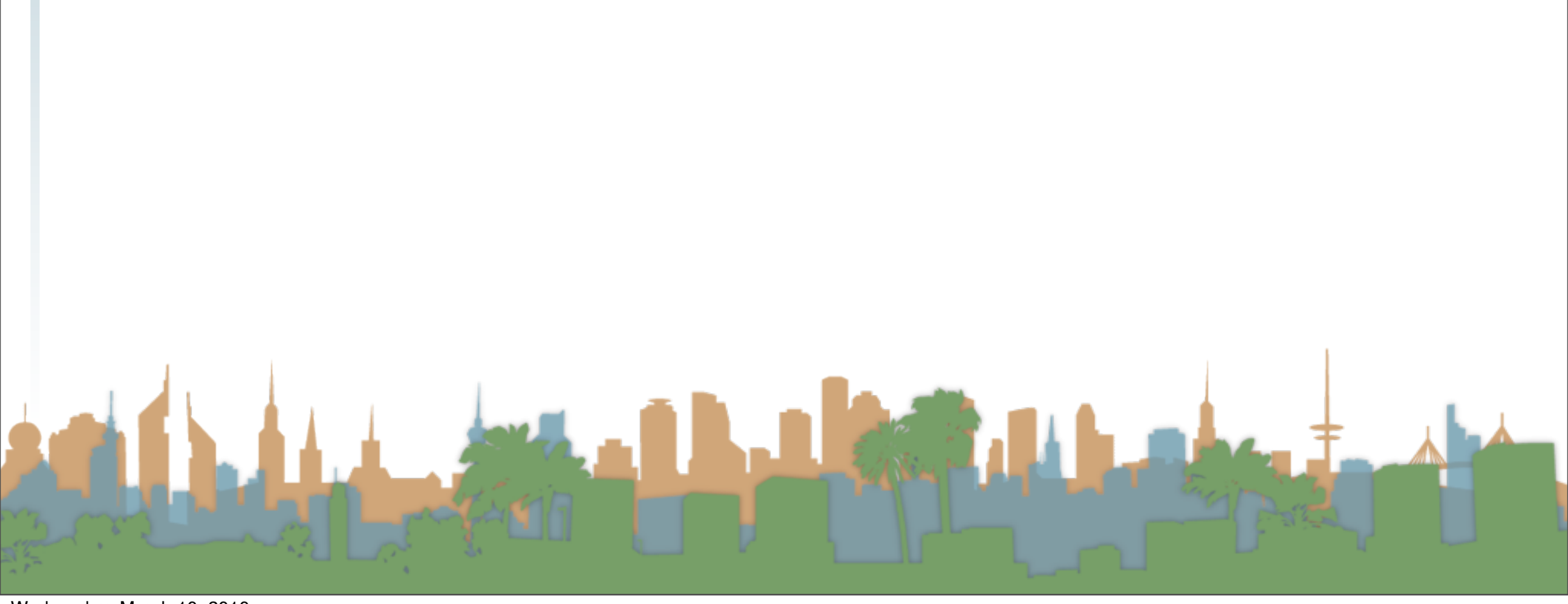


Unranked retrieval - Accuracy

- Welcome to my search engine
  - I guarantee a 99.9999% accuracy.

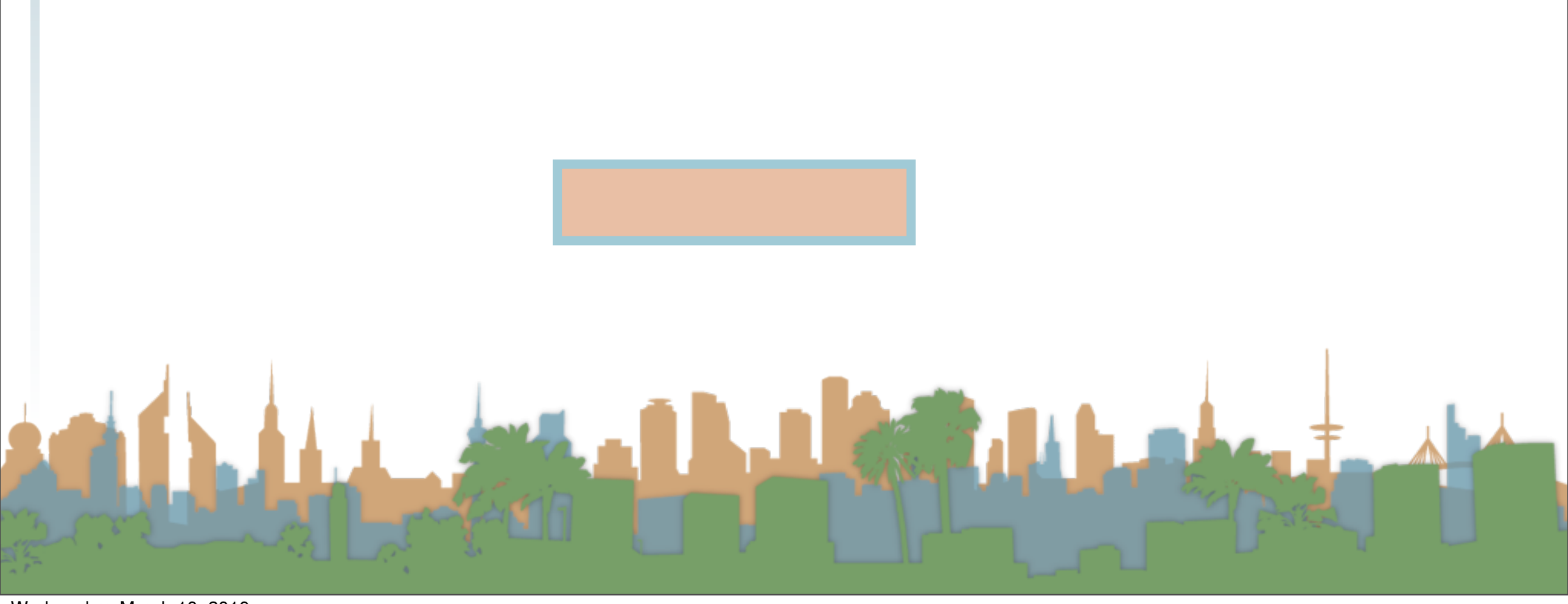
Unranked retrieval - Accuracy

- Welcome to my search engine
  - I guarantee a 99.9999% accuracy.
  - Bring on the venture capital



Unranked retrieval - Accuracy

- Welcome to my search engine
  - I guarantee a 99.9999% accuracy.
  - Bring on the venture capital



Unranked retrieval - Accuracy

- Welcome to my search engine
  - I guarantee a 99.9999% accuracy.
  - Bring on the venture capital

Unranked retrieval - Accuracy

- Welcome to my search engine
  - I guarantee a 99.9999% accuracy.
  - Bring on the venture capital

Unranked retrieval - Accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Accuracy = \frac{0 + \uparrow}{0 + 0 + \epsilon + \uparrow}$$

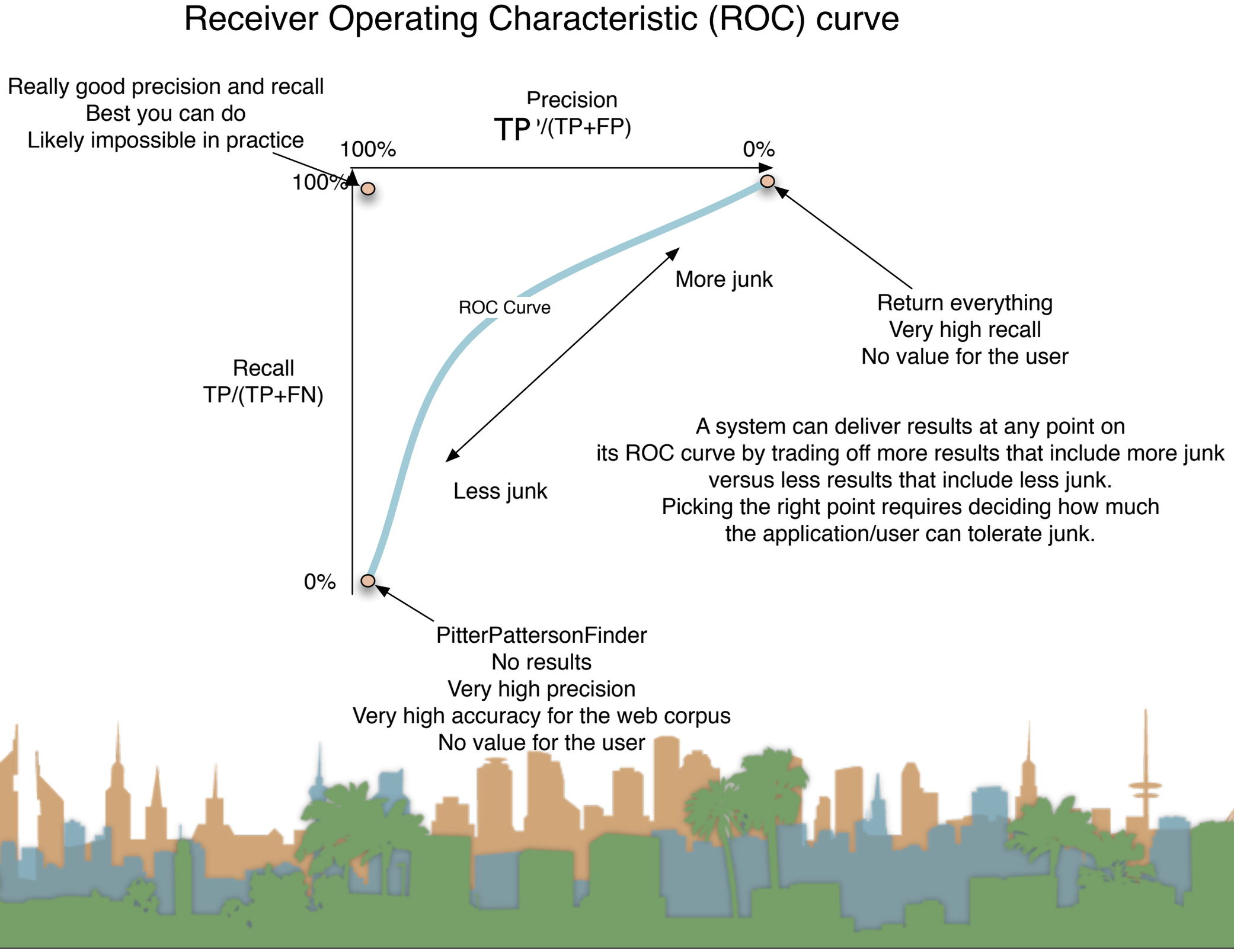
Unranked retrieval - Accuracy

- Most people want to find something and can tolerate some junk

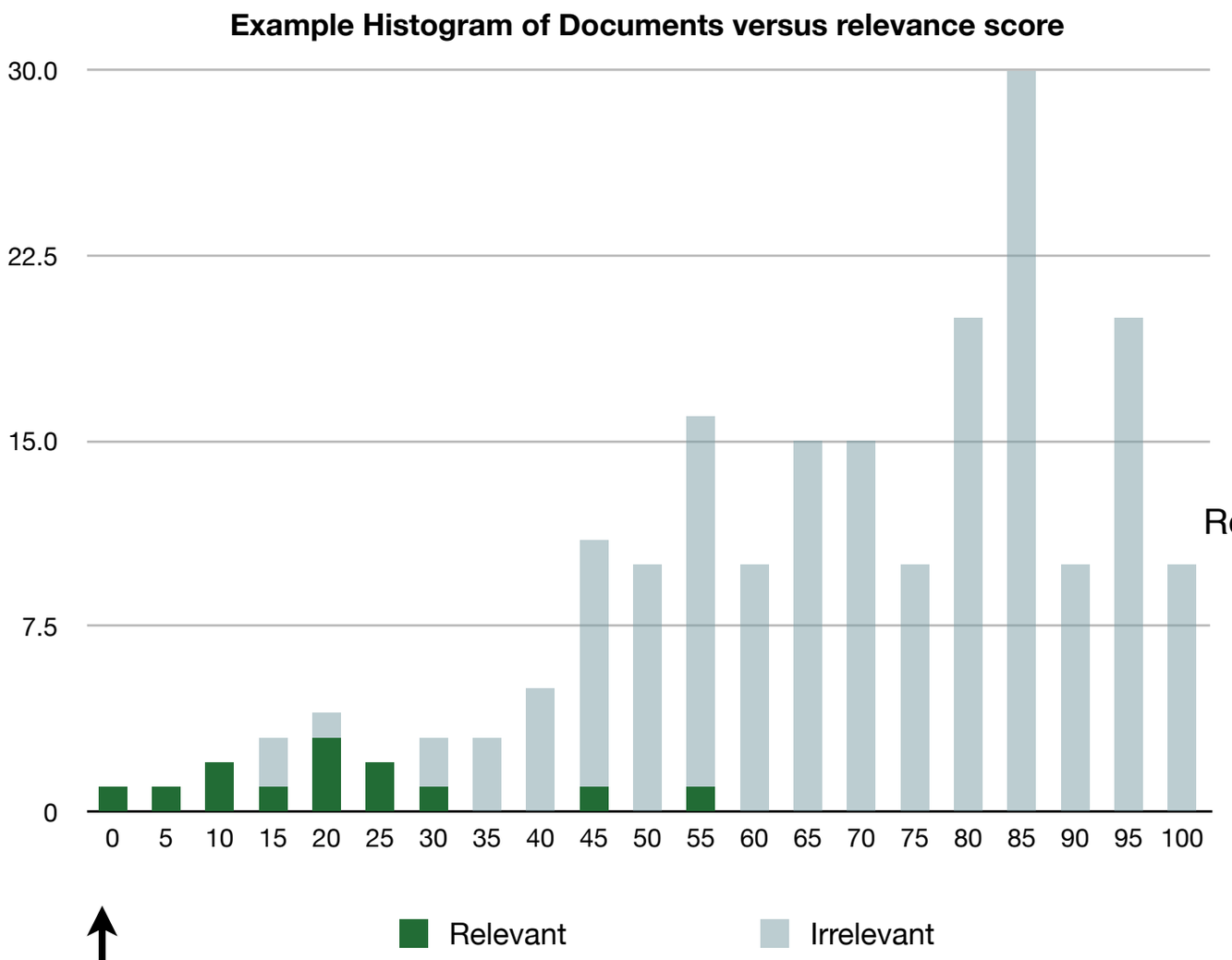
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

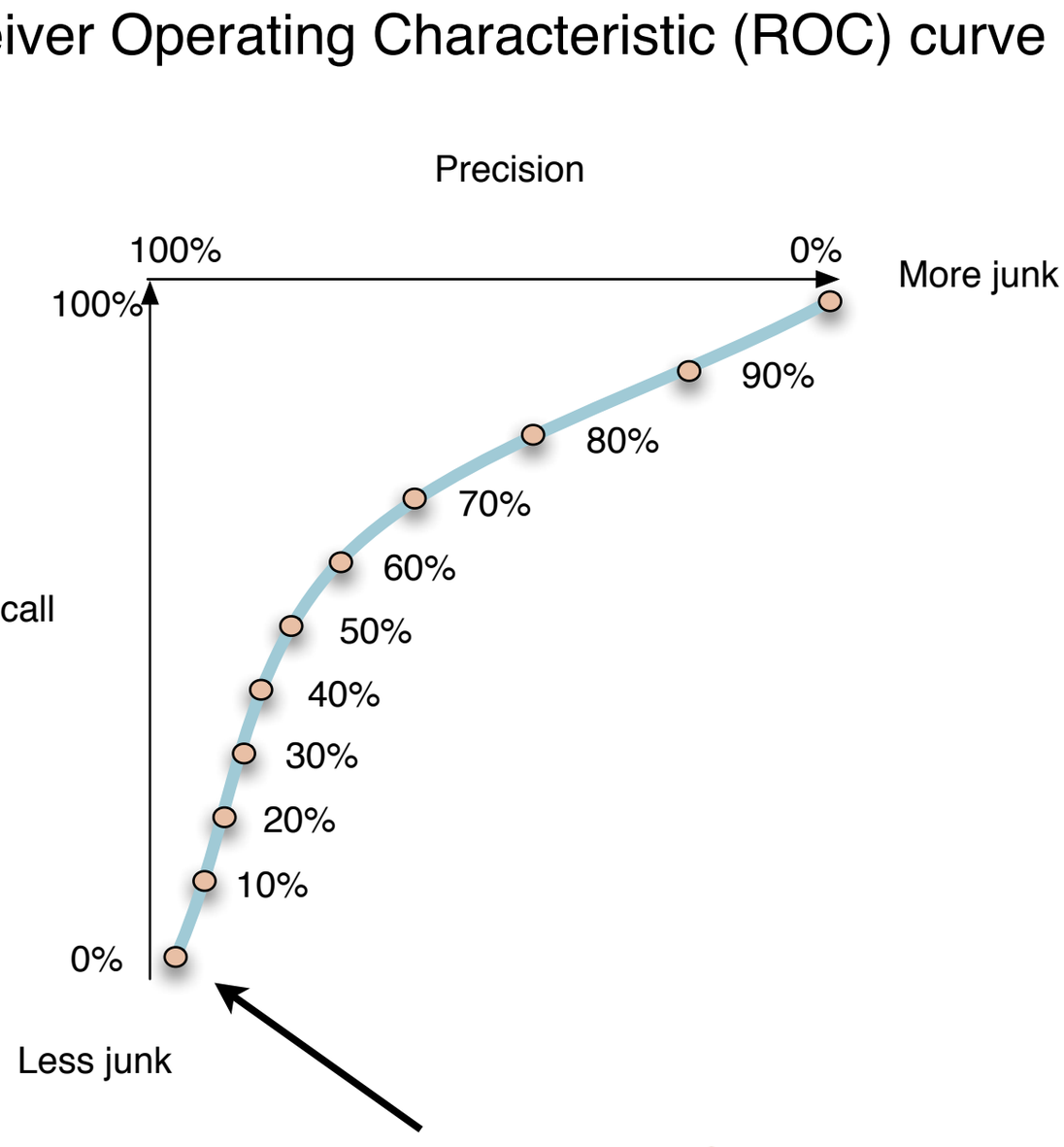
$$Accuracy = \frac{0 + \uparrow}{0 + 0 + \epsilon + \uparrow}$$

Unranked retrieval - ROC curve

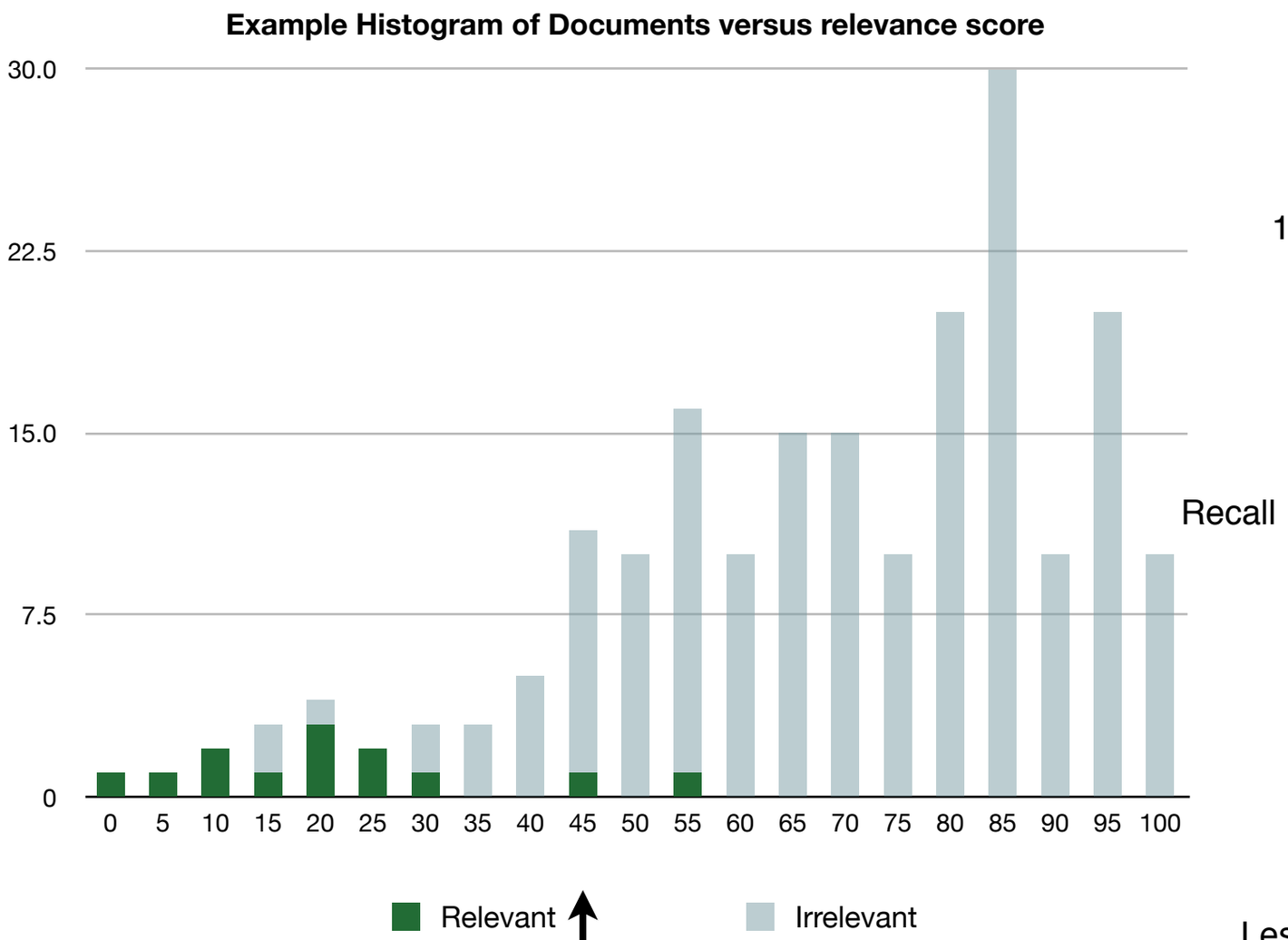


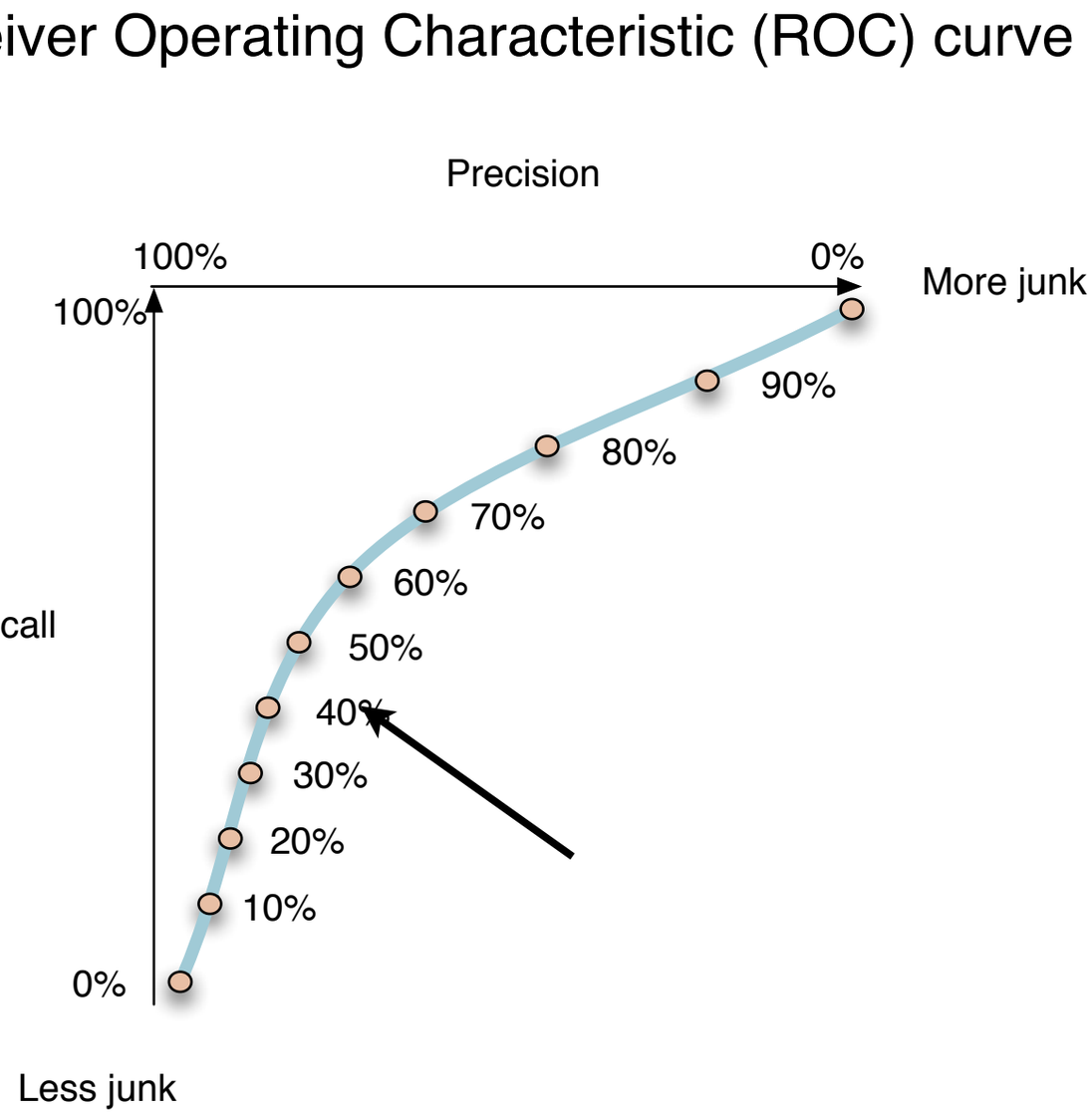
Unranked retrieval - ROC curve



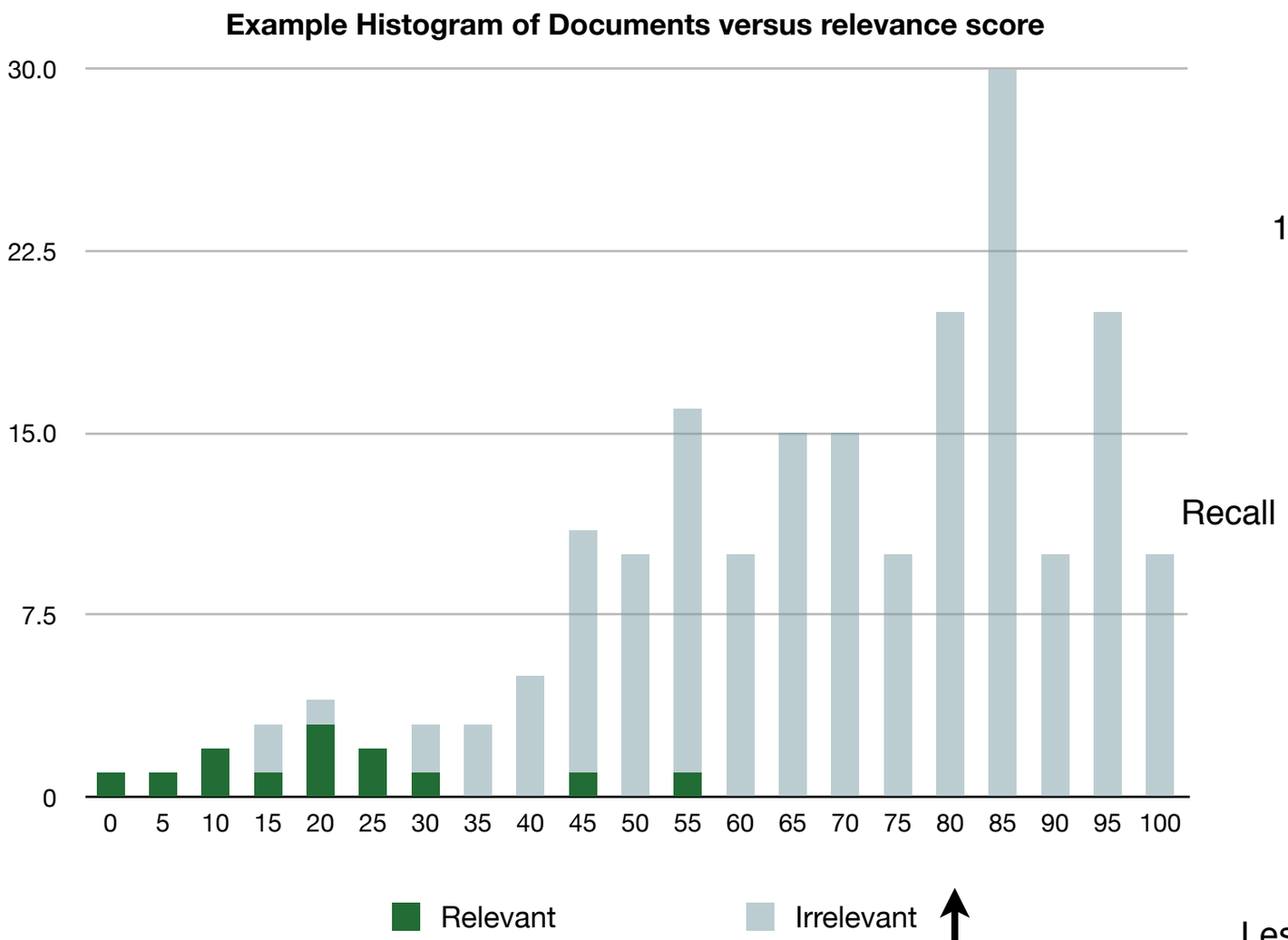


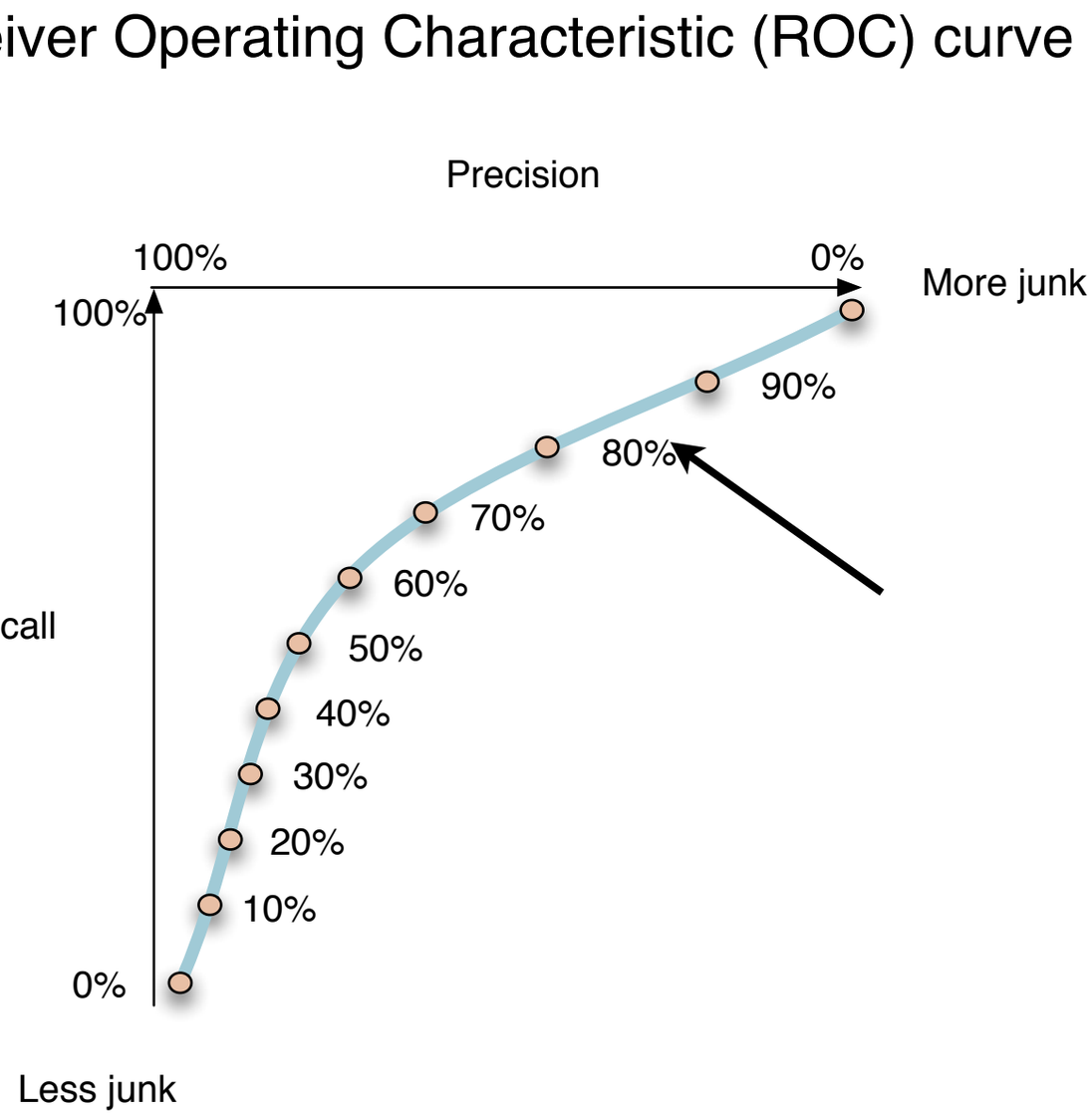
Unranked retrieval - ROC curve

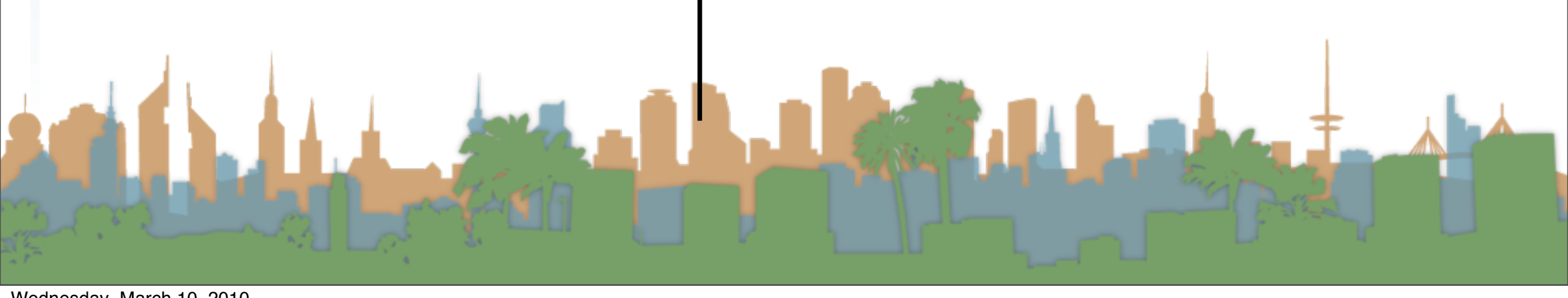




Unranked retrieval - ROC curve







Ranked Retrieval

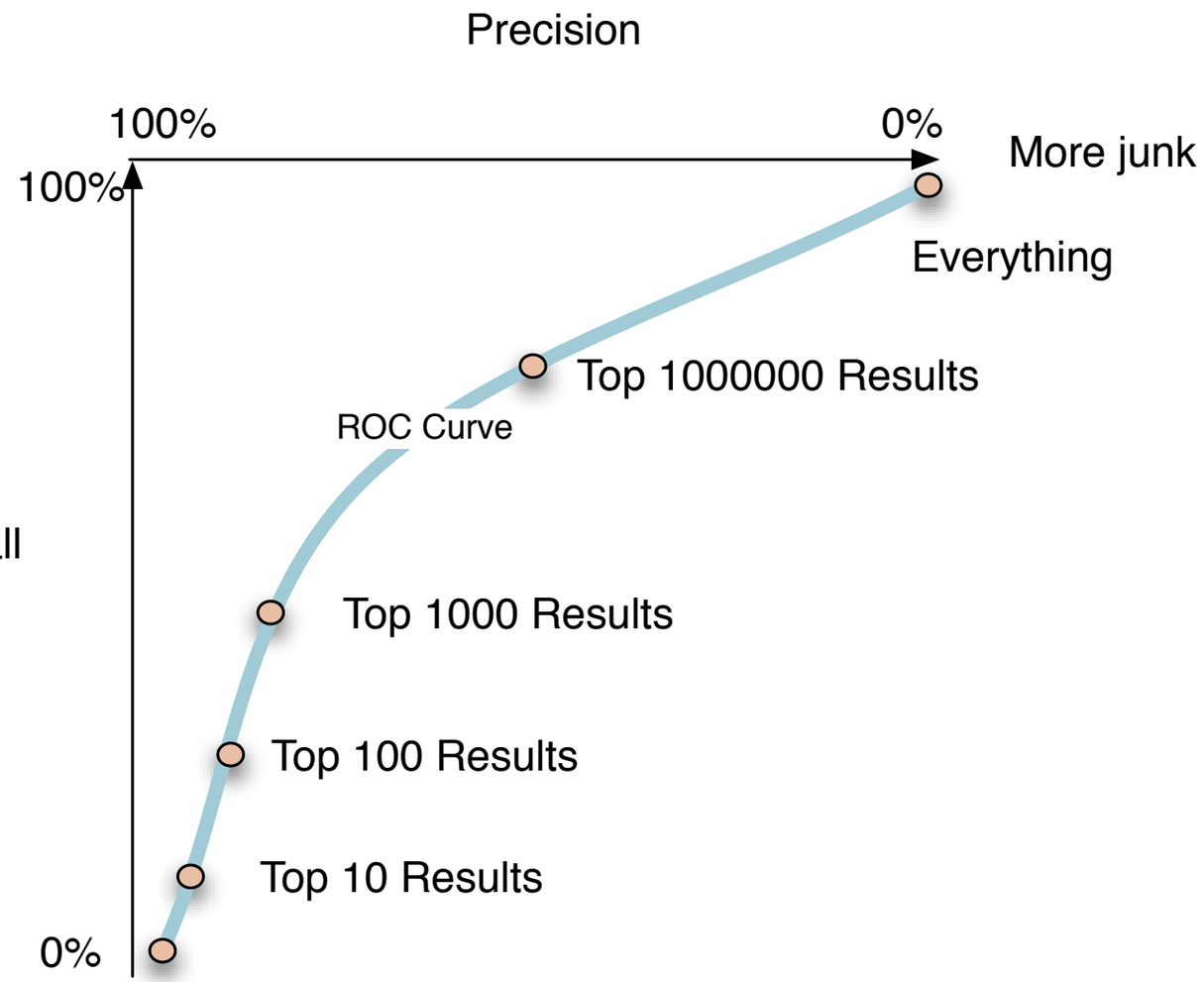
- Precision and Recall are set-based measures
  - They are computed independent of order
  - But, web search return things in lists
  - Lists have order.
  - A better metric of user happiness/relevance is warranted

Ranked Retrieval

- Let's use our existing metrics and extend them to ranked retrieval
  - In one system we can get many samples
  - We can get the top X results:
    - X= 10, 20, 30, 40, etc...
- Each one of those sets has a precision and recall value
- Each of those sets corresponds to a point on the ROC curve.

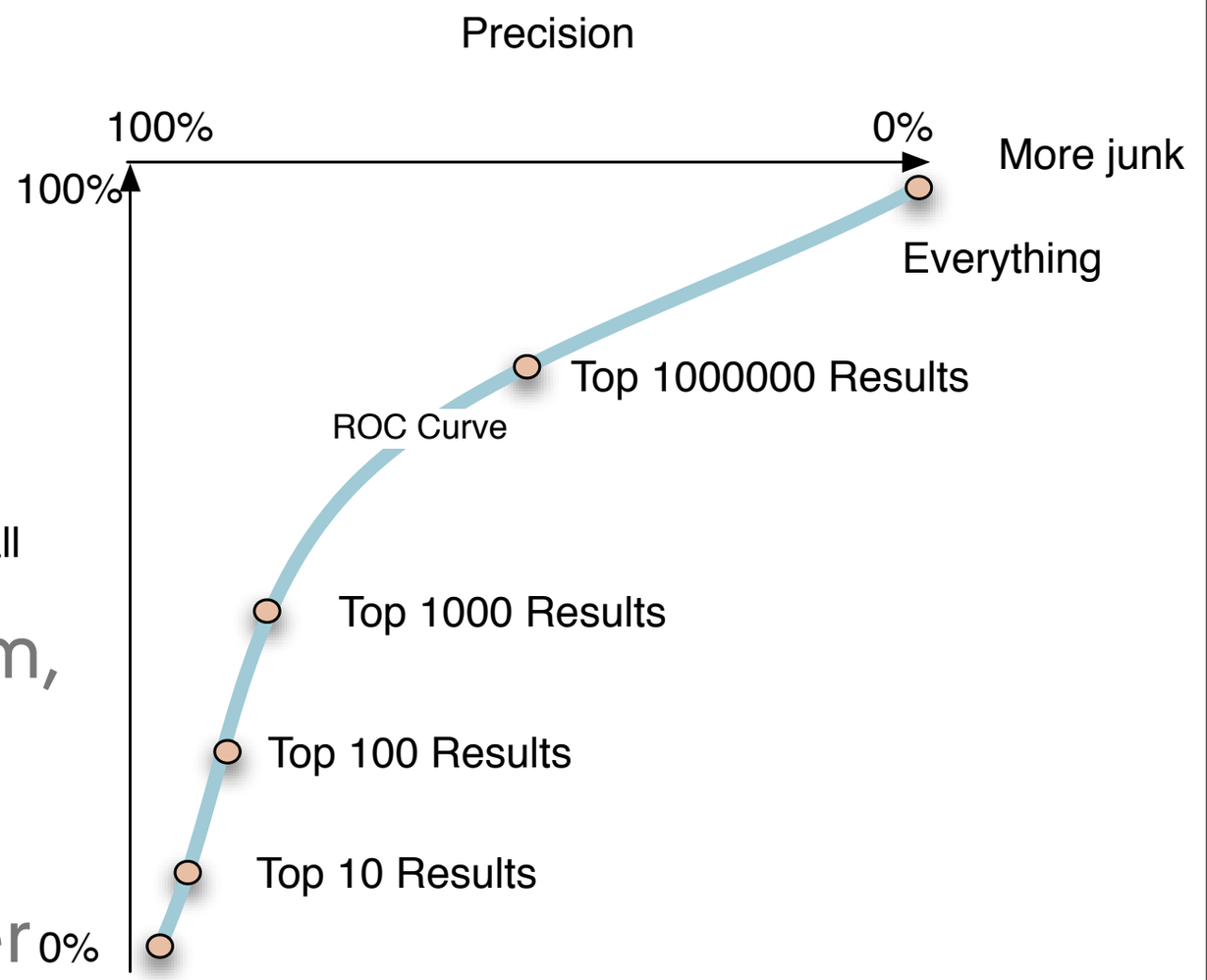
Ranked Retrieval

- Each of those sets corresponds to a point on the ROC curve.



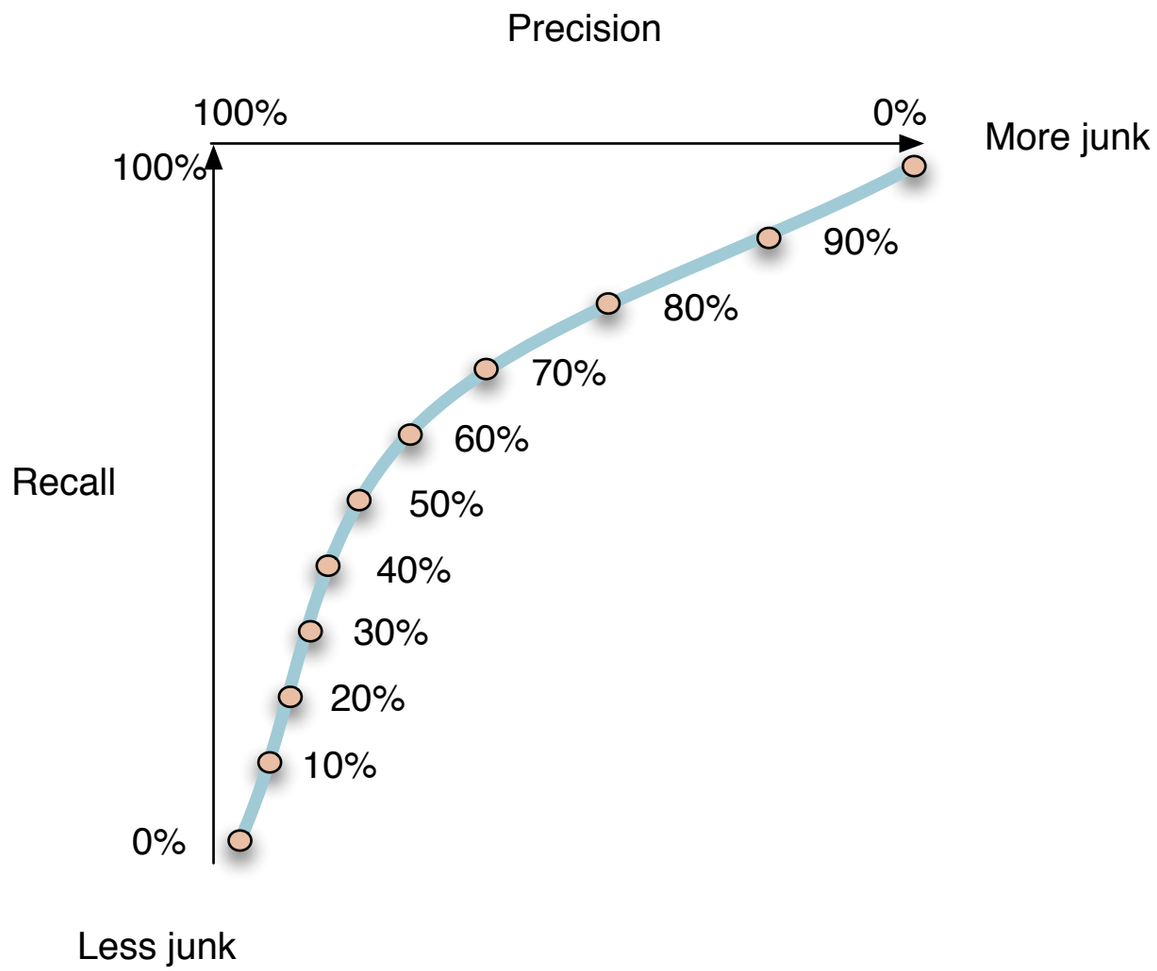
Ranked Retrieval

- One option is to average the precision scores at discrete points on the ROC curve
- But which points?
- We want to evaluate the system, not the corpus
- So it can't be based on number of documents returned



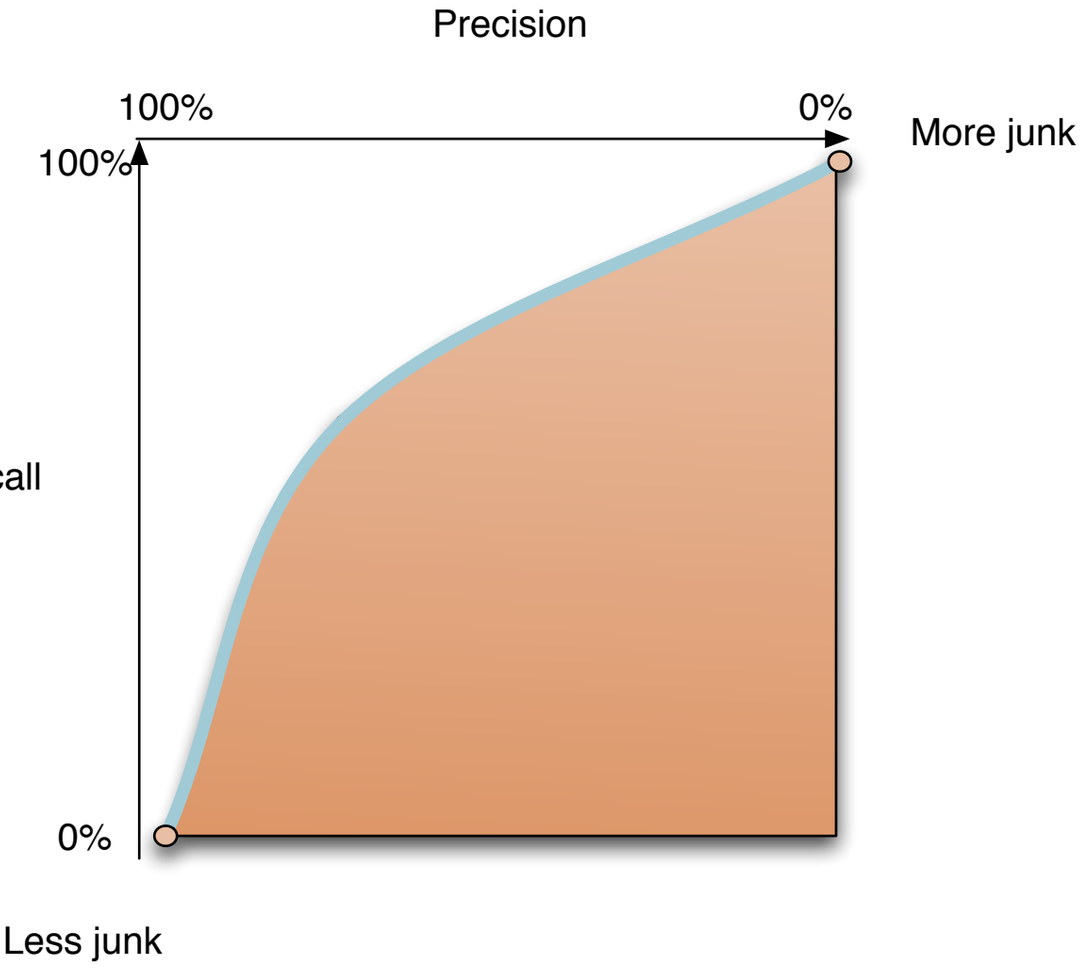
Ranked Retrieval - 11 point precision

- Evaluate based on precision at defined recall points
- Average the precision at 11 points
- This can be compared across corpora
  - because it isn't based on corpus size or number of results returned



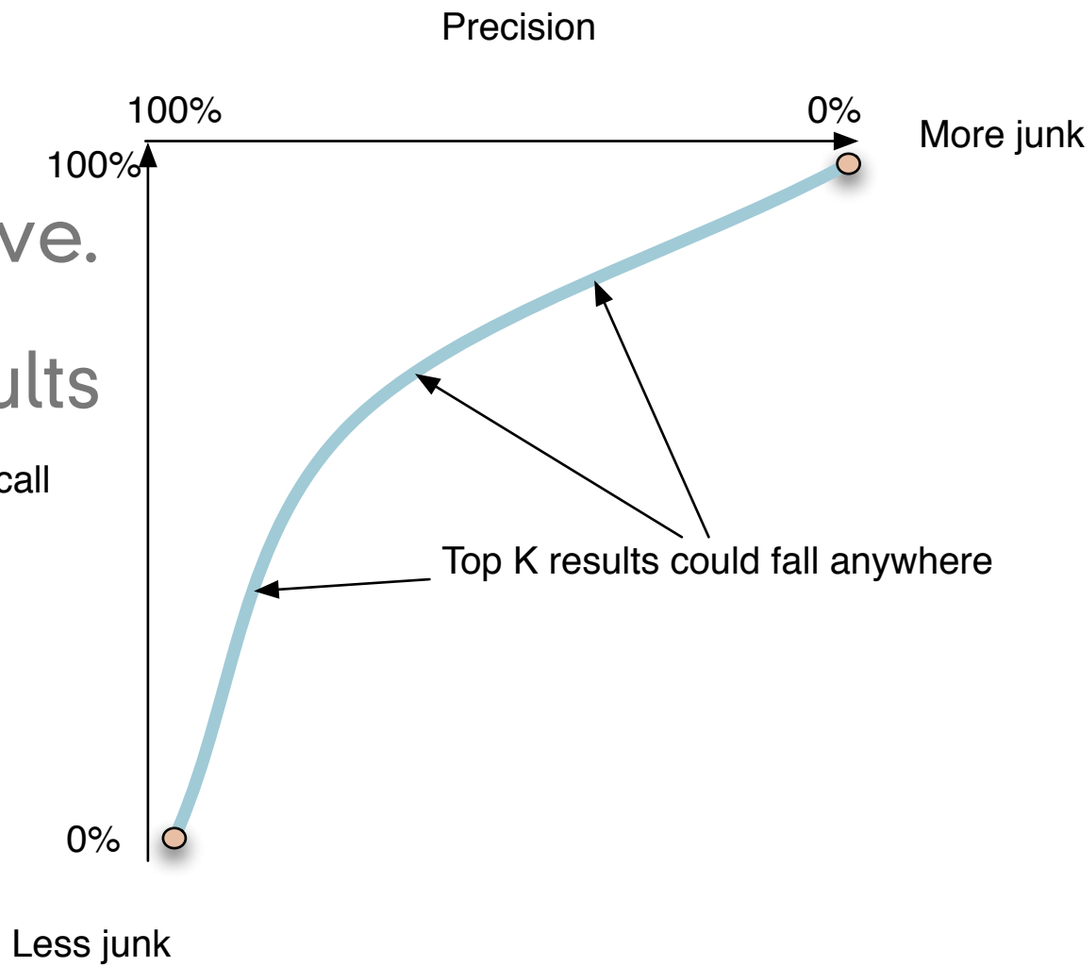
Ranked Retrieval - Mean Average Precision

- Why just 11 points?
- Why not average over all points?
- This is roughly equivalent to measuring the area under the curve.



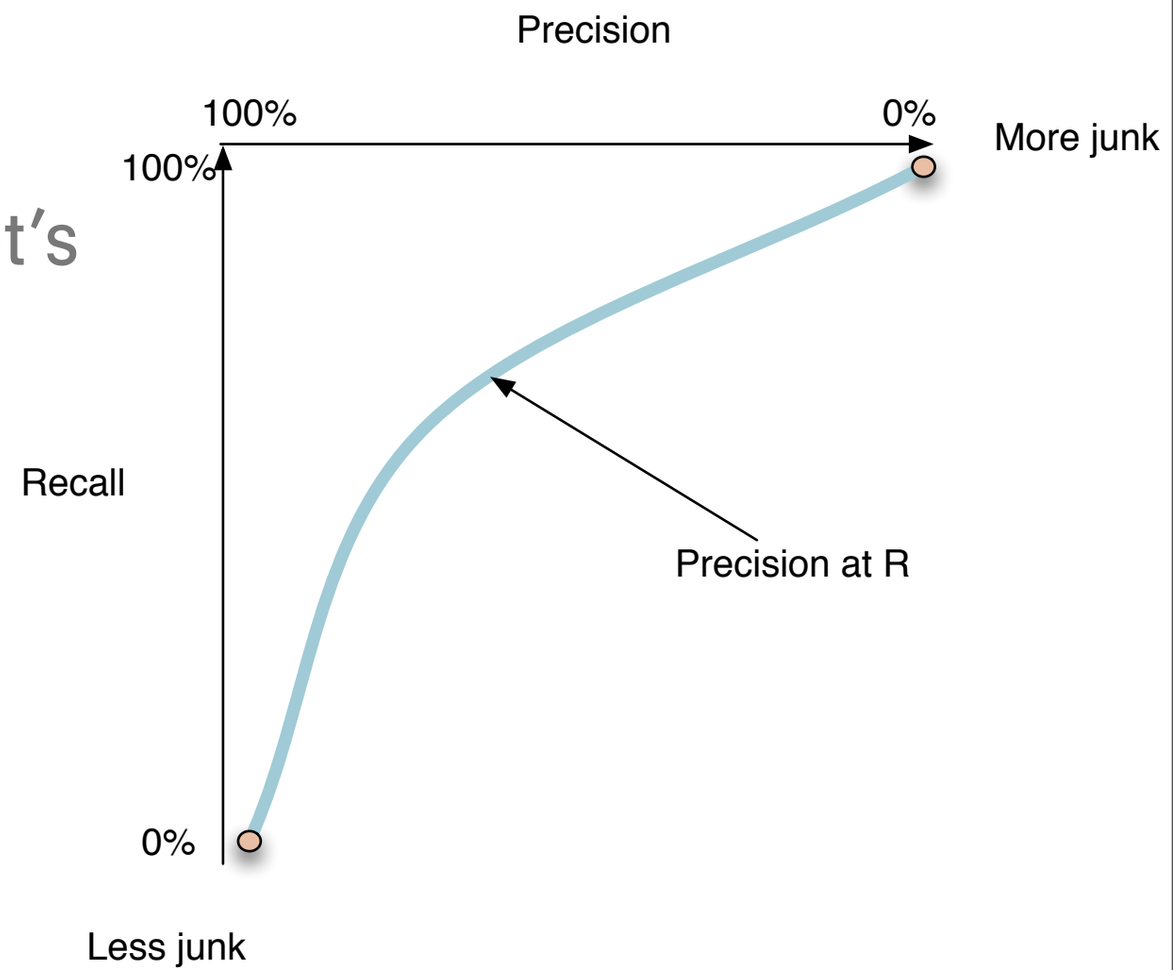
Ranked Retrieval - Precision at k

- Users don't care about results past a page or two
- So area under the curve is too naive.
- Let's evaluate precision with k results instead.
- Highly dependent on number of relevant documents
- If k is 20 and relevant docs is 8
  - best score is 8/(8+12) =0.4



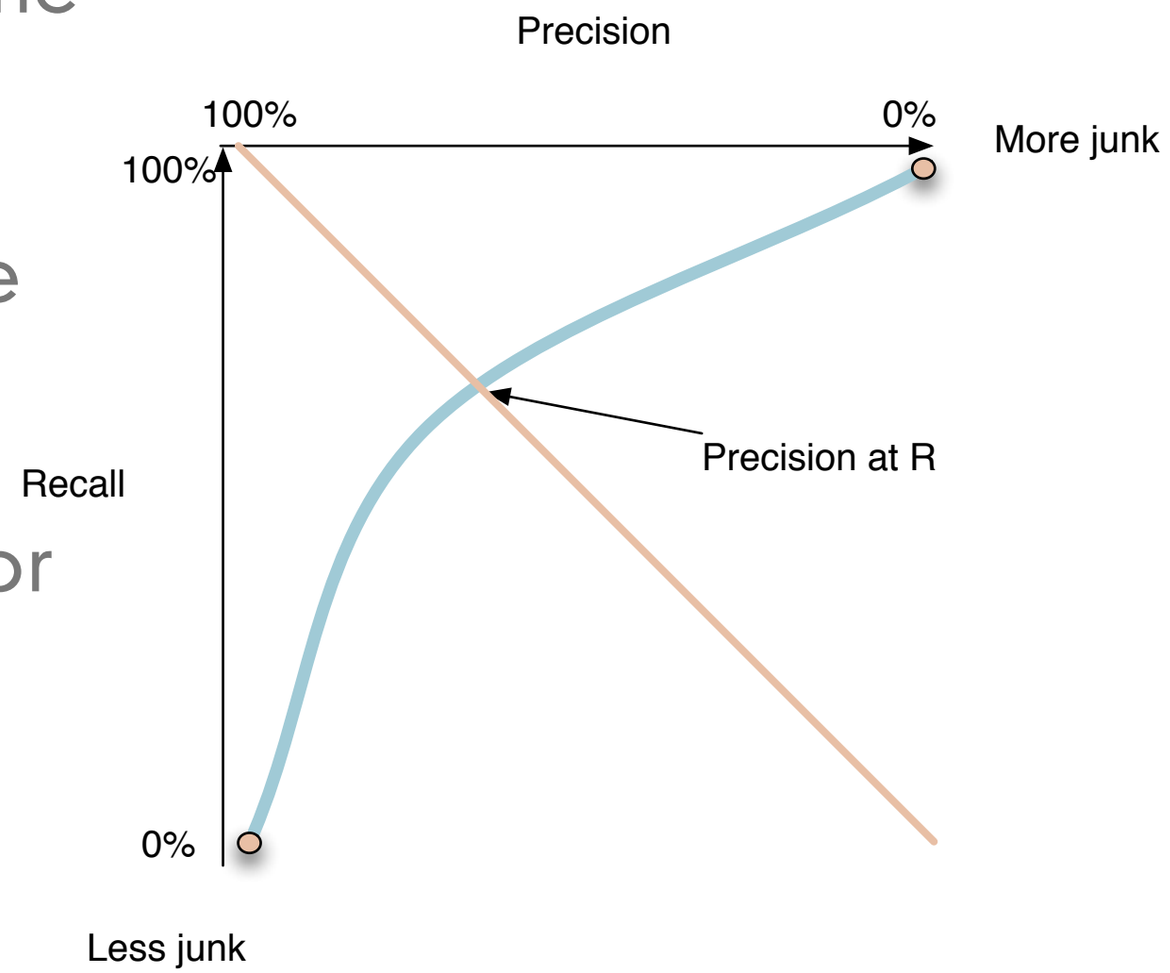
Ranked Retrieval - Precision at R

- We know the number of relevant documents, r, so
- rather than looking at k results let's look at the top r results
- If r is 20
- best score is 20/(20) =1.0
- best score is always 1.0



Ranked Retrieval - Precision at R

- It turns out that Precision at R is the break-even point
  - When Precision and Recall are equal
  - Do we care about this point for any rational reason?



Critiques of relevance

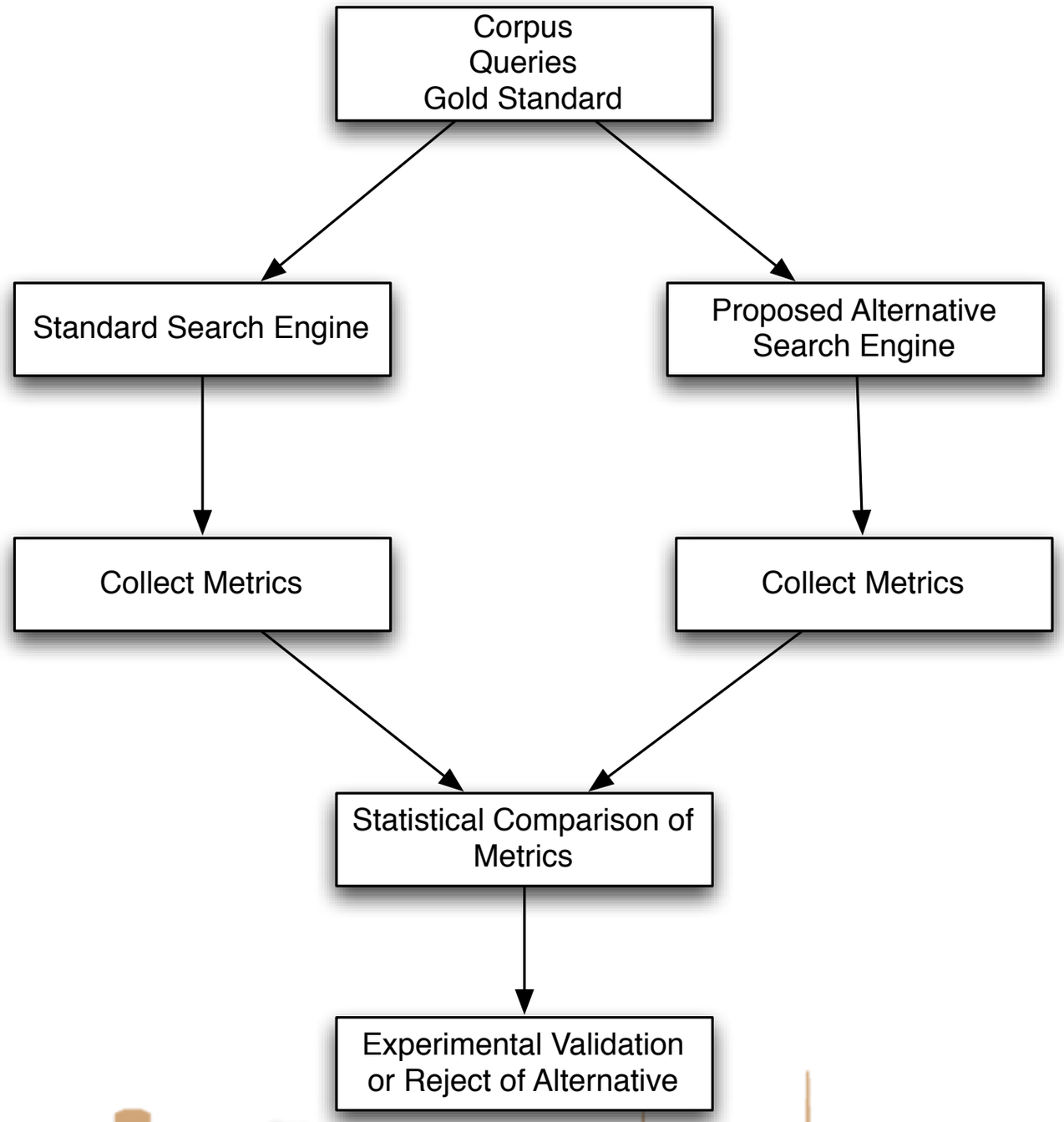
- Is the relevance of one document independent of another?
- Is a gold standard possible?
  - Is a gold standard static?
  - Uniform?
  - Binary?
- Perhaps relevance as a ranking is better.
- Relevance versus marginal relevance
  - what does another document add?

Refining a deployed system

- Once you have a system, with metrics, how do you consider changing the system to improve the metrics?
- A common approach is A/B testing.
  - This is done by Google for clients and Amazon for itself and probably many others.
- The idea:
  - Treat a small number of your users as experiments.
  - Have them use the different system.
  - Evaluate metrics on experimental group.

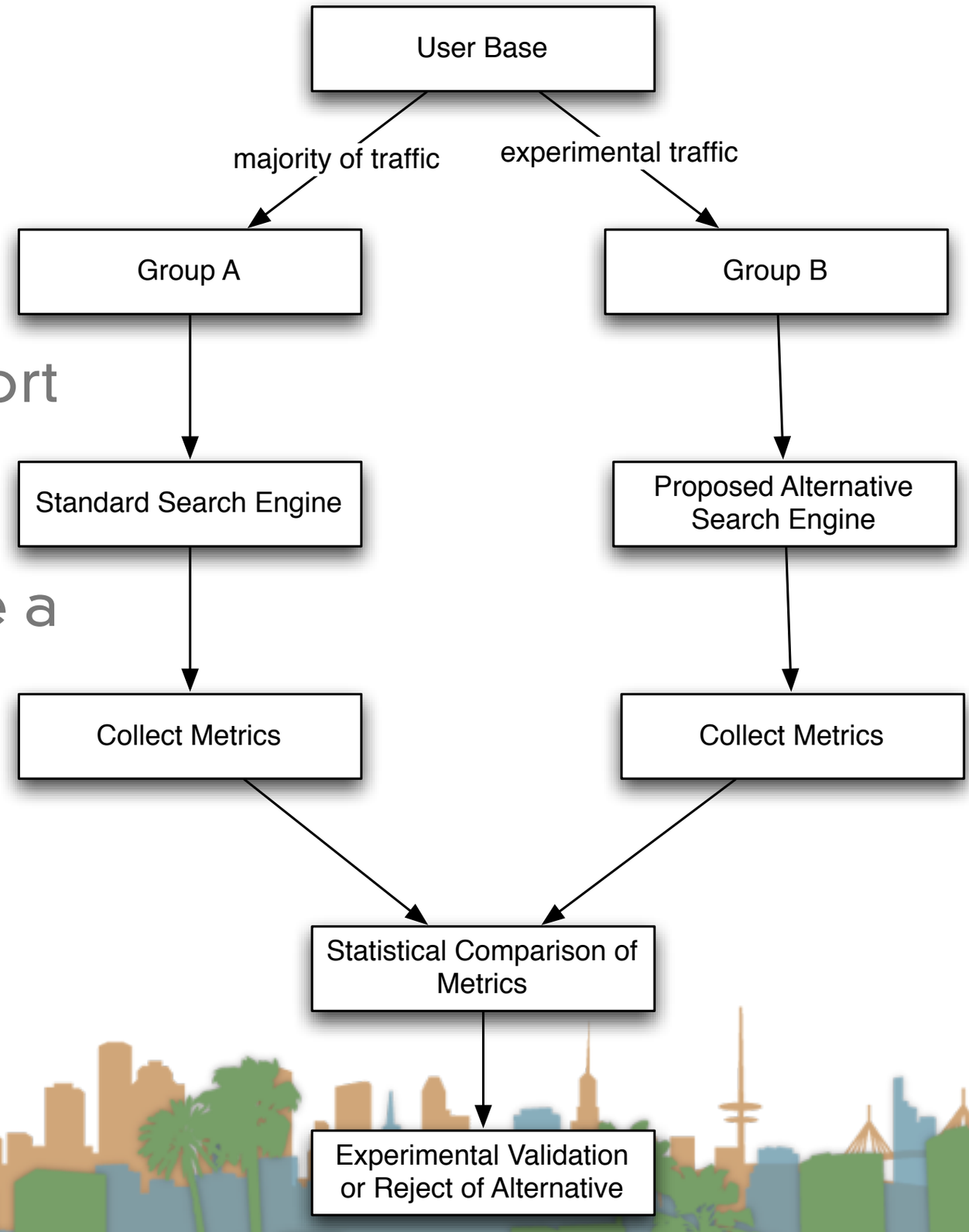
Evaluation in IR

- Gold standard approach

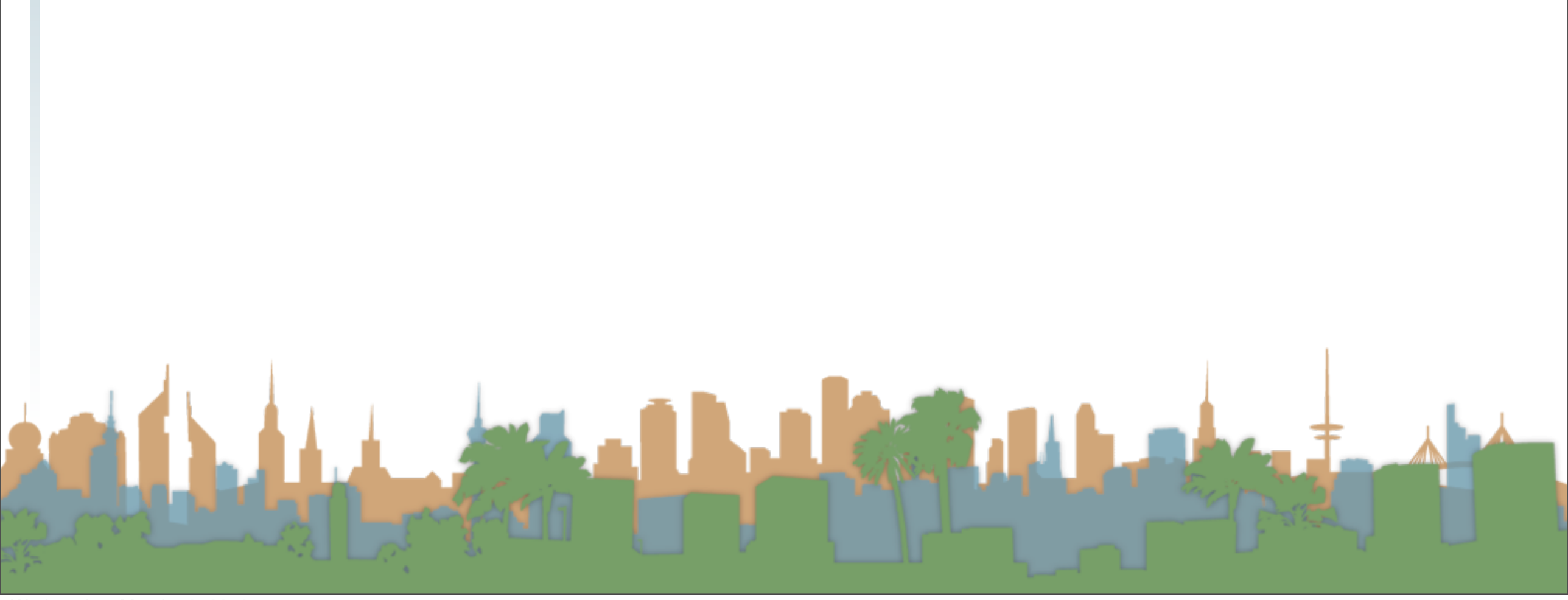


Online A/B approach

- Requires
  - users
  - an infrastructure to support testing
  - metrics that don't require a gold standard



Amazon



amazon.com

Hello, Sandra S. Patterson. We have recommendations for you. (Not Sandra?)

Get FREE Two-Day Shipping Now

Sandra's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards | Your Account | Help

Search: Amazon.com | GO | Cart | Your Lists

Shop All Departments

- Books
- Movies, Music & Games
- Digital Downloads
- Electronics & Computers
- Home & Garden
- Grocery
- Toys, Kids & Baby
- Apparel, Shoes & Jewelry
- Health & Beauty
- Sports & Outdoors
- Tools, Auto & Industrial

Introducing Kindle: Amazon's Revolutionary Wireless Reading Device

Amazon is excited to introduce Kindle—a wireless, portable reading device with instant access to more than 100,000 books, blogs, newspapers, and magazines. Whether you're in bed or on the train, Kindle lets you think of a book and get it in less than a minute.

› Learn more

amazonkindle

Amazon Daily BLOG — 7 posts since yesterday — Posts for Sandra

Treat Yourself

Design Methods

Since its initial publication in 1970, Design Methods has been considered the seminal work on... Read more

› See more in your Wish List

Denim Dilemma?

Jeans from Levi's, Lee, Diesel, True Religion, GUESS, 7 for all Mankind, Citizens of Humanity, and more

Get Yourself a Little Something

◄ ►

› See more in your Wish List

Add to Your Collection

Captain Underpants And...

In this worthy sequel to The Adventures of Captain Underpants, Pilkey maintains the... Read more

› See more items in the Captain Underpants series

Check This Out

Amazon Breakthrough Novel Award
Vote for the winner now.

J.K. Rowling's Fairy Tales
Find out more about this rare book.

High-Def 101
Learn. Shop. Discuss.

Selling on Amazon
List items for free and sell to millions.

Save $50 on Select Toshiba Laptops

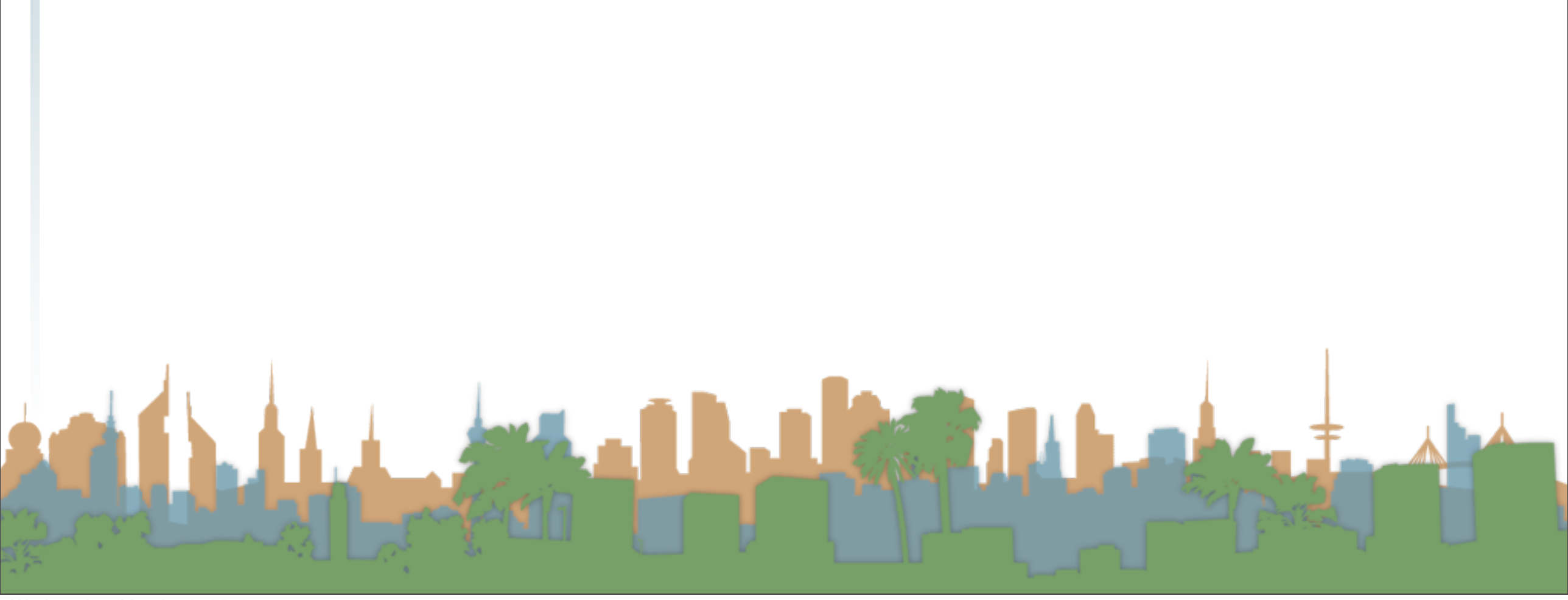
For a limited time save $50 on select Toshiba laptops offered by Amazon.com, with power Intel or AMD processors, large hard drives, and 13- or 15.4-inch display sizes. Hurry--savings end March 17.

Features & Services

Selling on Amazon

- Publish on Kindle
- Sell Your Stuff
- Fulfillment by Amazon
- WebStore by Amazon

Amazon



amazon.com

Hello. Sign in to get personalized recommendations. New customer? Start here.

Get FREE Two-Day Shipping Now

Your Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards | Your Account | Help

Search: Amazon.com | GO | Cart | Your Lists

Shop All Departments

- Books
- Movies, Music & Games
- Digital Downloads
- Electronics & Computers
- Home & Garden
- Grocery
- Toys, Kids & Baby
- Apparel, Shoes & Jewelry
- Health & Beauty
- Sports & Outdoors
- Tools, Auto & Industrial

Introducing Kindle: Amazon's Revolutionary Wireless Reading Device

Amazon is excited to introduce Kindle—a wireless, portable reading device with instant access to more than 100,000 books, blogs, newspapers, and magazines. Whether you're in bed or on the train, Kindle lets you think of a book and get it in less than a minute.

› Learn more

amazonkindle

Amazon Daily BLOG — 8 posts since yesterday — Read posts

Time to Quit?

Here's an extra incentive: Save an additional 15% on select smoking cessation products from Commit, NicoDerm, and more.

› Shop now

Check This Out

Amazon Breakthrough Novel Award
Vote for the winner now.

J.K. Rowling's Fairy Tales
Find out more about this rare book.

High-Def 101
Learn. Shop. Discuss.

Selling on Amazon
List items for free and sell to millions.

Shop Irish Jewelry at Amazon.com

› Claddagh

› Clovers & Shamrocks

› Celtic

› Shop all Irish jewelry

Save up to 30%

Save on Xyron's helpful and handy craft products and expand your crafting repertoire. Shop now.

Save $50 on Select Toshiba Laptops

For a limited time save $50 on select Toshiba laptops offered by Amazon.com, with power Intel or AMD processors, large hard drives, and 13- or 15.4-inch display sizes. Hurry--savings end March 17.

$5 or $10 Bonus on *Southern Living*

Southern Living

Subscribe to *Southern Living*

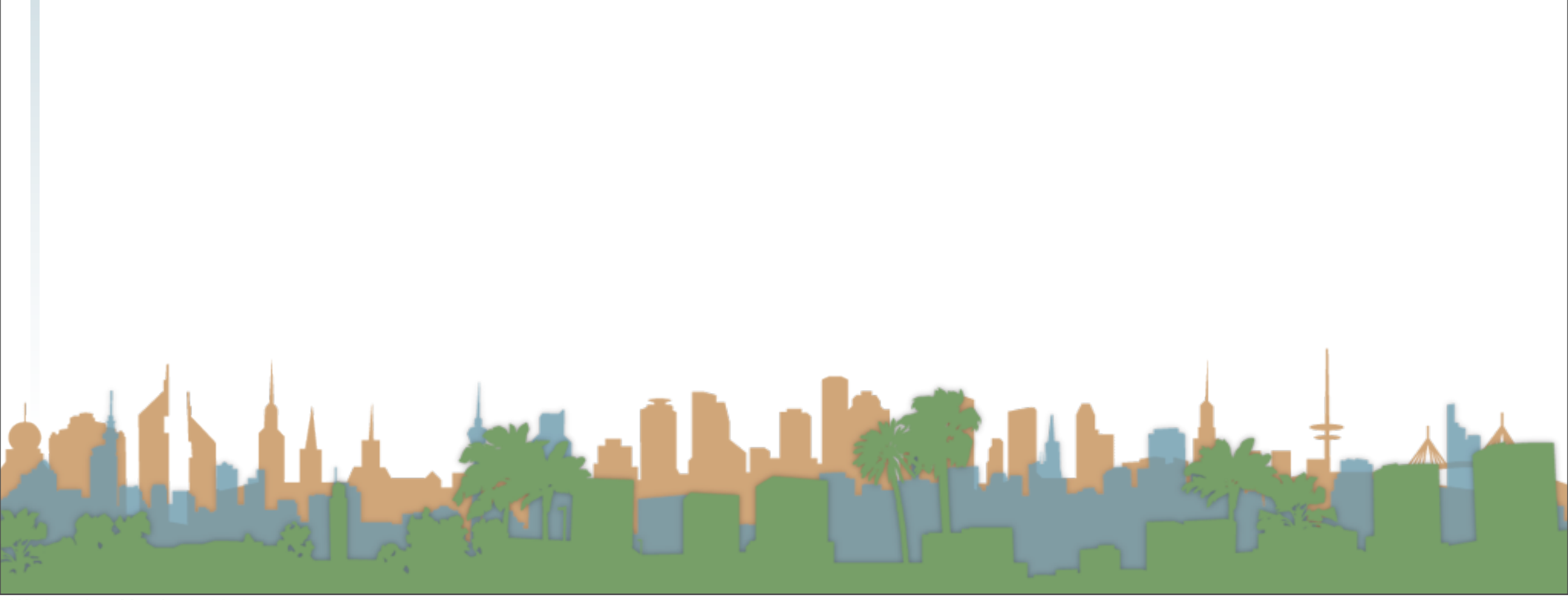
Features & Services

Selling on Amazon

- Publish on Kindle
- Sell Your Stuff
- Fulfillment by Amazon
- WebStore by Amazon

Evaluation in IR

Google



Evaluation in IR

G

Google AdWords

donald.j.patterson.iii@gmail.com | Help | Contact Us | Sign Out
Customer ID: 475-906-9660

Campaign Management | Reports | Analytics | My Account

Campaign Summary | Tools | Conversion Tracking | Website Optimizer

Search

Tools > Website Optimizer

Website Optimizer (beta): Experiment List

✓ **Congratulations!** You've signed up for Website Optimizer and are ready to create your experiment. As you move through the steps, your information will be saved.

+ Create a new experiment

Delete [?] View: All experiments

Experiment	Status	Page Visitors	Conv.	Conv. Rate	Finish Time
No experiments					
Click Create experiment to get started.					

My Client Center User Access [?]

Tell us about Website Optimizer. Take a survey

Common Questions

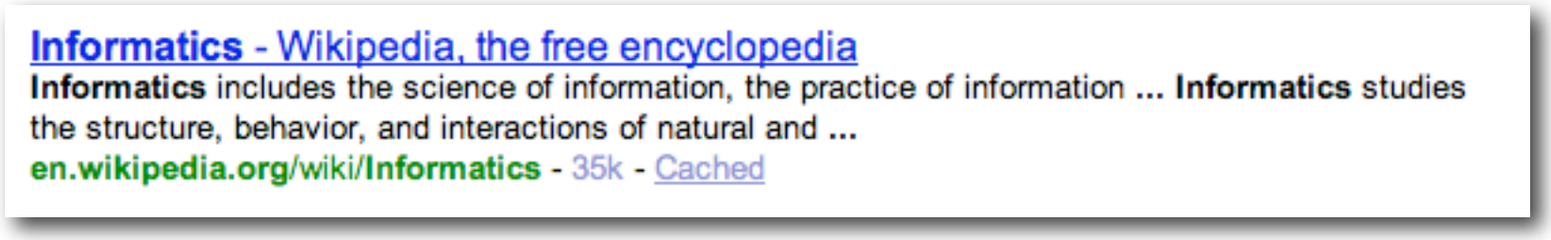
- How do I create an experiment?
- Do I need my webmaster's help?

Helpful Links

- Website Optimizer Help
- Website Optimizer Demo (Flash - English Only)
- Website Optimizer User Guide (Flash - English Only)
- Quick Start Guide
- Installation Guide
- Testing Guides and Strategies
- Discussion Forum
- Website Optimizer Home Page
- Send Us Feedback

Snippets

- Little bits of text that summarize the page



- They function as an implicit tool for users to rank the results on their own (among those visible)
- The user does the final ranking
- Users are still biased by presented order though.

Snippets

- The goal of snippet generation is
  - present the most informative bit of a document in light of the query
  - present something which is self-contained
    - i.e., a clause or a sentence
  - present something short enough to fit in output
  - be fast, accurate (where are the snippets stored?)
- Challenges
  - Multiple occurrences of keyword in document
  - Poor English (or other language) grammar

Snippets

- Snippets can be static
  - A snippet for a web page is precompiled and always the same.
- Snippets can be dynamic
  - Depends on the query
  - "informatics"

Informatics - Wikipedia, the free encyclopedia
Informatics includes the science of information, the practice of information ... **Informatics** studies the structure, behavior, and interactions of natural and ...
en.wikipedia.org/wiki/**Informatics** - 35k - Cached

  - "informatics definition"

Informatics - Wikipedia, the free encyclopedia
Usage has since modified this **definition** in three ways. ... broader **definition** at the launch of its School of Computing and **Informatics** in September 2006. ...
en.wikipedia.org/wiki/**Informatics** - 35k - Cached

Snippets

- Snippets may contain
  - A few sentences from the web page
  - Meta data about the page
    - Author, Date, Title
  - Output of a text-summarization algorithm
    - Advanced technology that attempts to write snippets
- Images from the document