The Data Divide

Luke Segars, Google

4/15/2013 @ CS10

Statement

Google's advantage is not in writing drastically better software; it's in having more data.

Question

Can any problem be solved by computers if enough data is available?

One major change that's come about from the digital revolution is the fact that *MUCH more data is available* for consumption now than ever before.

Internet
~4.5M URLs / month

Twitter
~5.5B tweets / month

Blogs
lots / month

Google Maps
>= 5M miles of road

Project Gutenberg
40,000 free books

And that's just the visible stuff...

iClicker Question

Approximately how many web pages did Google have in its search index earlier this year*?

a. 200 million
b. 10 billion
c. 30 billion
d. 47 billion
e. 100 billion

* according to public estimates

That's a lot to search through, but it's also a lot to learn from.

All of the text, images, video and other media generated on the Internet aren't entirely independent, and finding regular occurrences generally implies a correlation between concepts.

Let's consider an example:

*Say you have **10,000** news articles from diverse sources about Hurricane Sandy.*

***7,000** of them contain the phrase "New York."*

***3** of them contain the phrase "Arizona."*

Assuming that your news sources are actually telling a story, it is reasonable to assume that Hurricane Sandy is more closely related to New York than Arizona.

The core idea is based in statistics:

1 Many people, places, things, and ideas are somehow related to each other.

2 Ideas that are more closely related to each other are more likely to co-occur.

3 Co-occurring once means nothing, but co-occurring millions of times suggests that the two ideas are related.

Let's look at three places where Google uses this principle to make great things.

PageRank

How do we objectively measure a site's reputation?

Spell checking

How can we build a system that automatically learns new words in any language?

Web ranking

How do we know which pages best answer a particular query?

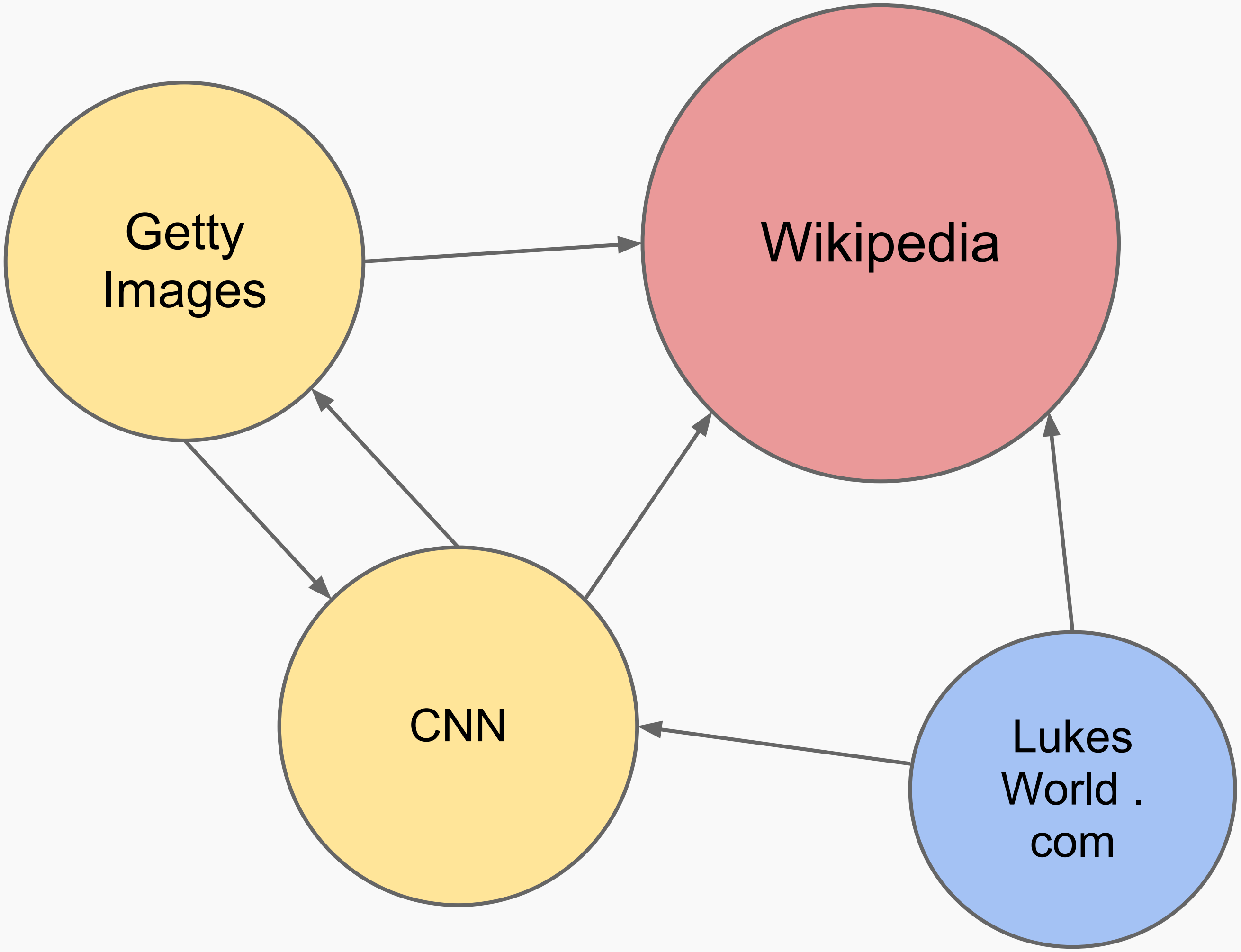
Example #1: PageRank

Article on CNN › my blog post

The Internet is full of data sources, some more reputable and dependable than others.

Pagerank is an algorithm that estimates the reputation of a website by looking at the reputation of websites that provide links to it.

Example #1: PageRank



Example #2: Spell checker

Splel checking is a fairly well-understood problem if you've got a reasonably static vocabulary. The Internet does not have a static vocabulary; new phrases are emerging all the time in all different languages.

How can Google keep its spell checker up to date in every language without a ton of work?

Example #2: Spell checker

Statistics! Or counting, really…

Key observations:

- A particular misspelling will likely be uncommon across the web, especially on reputable sites.
- The context that a misspelling occurs in will be similar to the context of a similar but more common spelling.

Example #3: Web ranking

Web ranking is concerned with maximizing ranking at the head, i.e. it's much more important to have #1 and #2 ranked correctly than #234, 401 and #234,402.

In many cases it's hard to accurately rank documents at the head by analyzing the content of the pages.

Example #3: Web ranking

If only there was some other way to know that a particular page was a good answer for a particular query...

There are many other significant applications of this principle, even without looking outside of Google.

Identifying synonyms and acronyms

How else can a particular term be phrased yet have the same meaning?

Major event detection

How can we know when a significant event is occurring?

Email spam detection

How can we identify legitimate messages?

Why do we have so many unsolved problems if all it takes is lots of data?

...either because you can't actually solve all problems with lots of data or because we don't have the right data.

Summary: until very recently, people often assumed that our ability to compute was based on our ability to write effective algorithms.

It turns out that many interesting problems can be "solved" with gigantic inputs to rather simple algorithms.

Considering the vastness of the Internet and the billions of people who still aren't connected, we've still got a lot of learning left to do.

Consider some problems that are currently unsolved but could potentially be solved with more data:

Identifying signatures for all genetic diseases.

What parts of a person's genome sequence is responsible for particular characteristics?

Traffic.

How can we adapt traffic lights, navigation, and tolls to minimize traffic?

Natural disaster prediction.

Can we predict natural disasters like earthquakes earlier to save more lives?