

---

# Data and Big Data

**Welcome to the Age of Information**

---

Notes for CSC 100 - The Beauty and Joy of Computing  
The University of North Carolina at Greensboro

---

# Reminders

---

## Project:

- Written proposals due on Wednesday

## Homework 3

- Due in on Monday, November 6

## Reading:

- The work of Luis von Ahn: Discussion contribution by Monday
-

# Data...

---

What is data? Is it the same as information?

“You can have data without information, but you cannot have information without data.” - Daniel Keys Moran

Data is being collected, generated, and stored far, far faster than ever before. How much?

“In 2012, every day **2.5 quintillion bytes of data** (1 followed by 18 zeros) are created, with 90% of the world’s data created in the last two years alone.” <http://marciaconner.com/blog/data-on-big-data/>

The result is a flood of data...

... or “Big Data”

---

# Everyone is talking about “Big Data”

“Big Data” is Big News...



**WIRED** GEAR SCIENCE ENTERTAINMENT

INNOVATION INSIGHTS contributor content

## Big Data: Not Just for Nerds Any

BY YANIV MOR, XPLENTY 10.02.13 4:02 PM



**THE WALL STREET JOURNAL.** | OPINION

INFORMATION AGE

## Why 'Big Data' Is a Big Deal

Analysis showed that prematurely born infants with unusually stable vital signs correlates with serious fevers 24 hours later—enabling physicians to take preventive measures.

March 24, 2013 5:13 p.m. ET

Analog-era minds have a hard time processing a key product of the digital era: the staggering amount of information being created, collected and correlated. What's called "big data" already identifies flu outbreaks and treatments for premature babies, and it predicts apartment overcrowding and airline delays.



**Forbes** News

CMO NETWORK | 10/22/2013 @ 9:31AM | 6,012 v

## Four Steps To Turn Big Data Into Action

McKinsey & Company

+ Comment Now + Follow

By Chris Meyer, Maher Masri, Ab

Data is meaningful make decisions th

Image: StuffEyeSee/Flickr

Peanut butter and jelly. Peas and and data? Of course.



**The New York Times** Technology | Personal Tech | Business Day

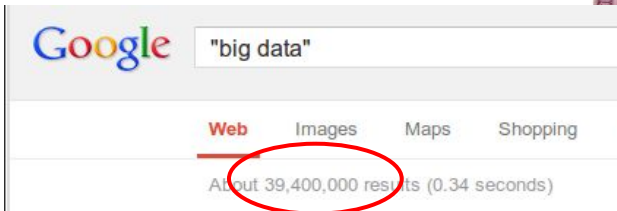
## Bits

SEPTEMBER 17, 2013, 2:21 PM | 1 Comment

## Improving the Big Data Toolkit

By STEVE LOHR

Open source software tends to march into the marketplace step by step, a quiet but steady strategy compared with the grand marketing events of the commercial software world. And Hadoop, the bedrock software of the fast-growing Big Data business, is on the march.



Google "big data"

Web Images Maps Shopping

About 39,400,000 results (0.34 seconds)

# What does “Big Data” mean?

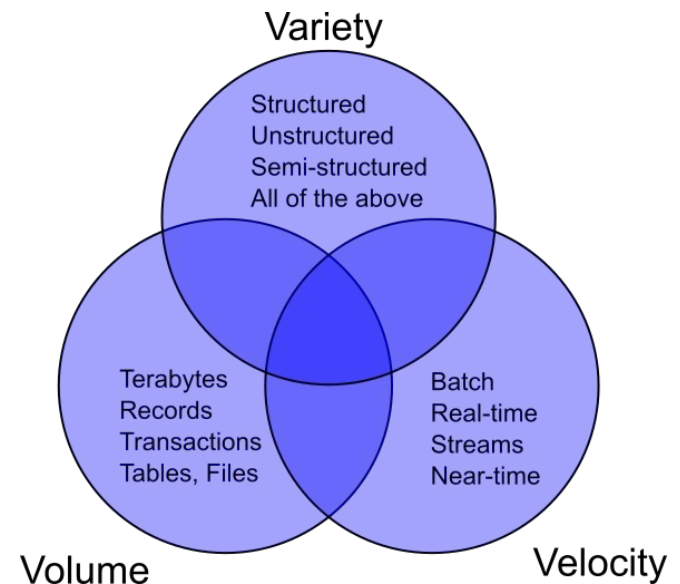
---

“Big Data” characterized by “three V’s”:

- Volume: How much data (lots!)
- Velocity: How fast does it arrive (and get processed)
- Variety: Different sources, different types - text, images, ...

Some people add a fourth “V”:

- Veracity: Quality of data



# Volume: Measuring Data

---

The basic unit of information is a bit

... or a byte (usually 8 bits)

↑  
“usually”?!? Really? Really!!  
Almost always 8 bits, but not always  
When it’s important that we refer to 8 bits, the term used is “octet”

What is a kilobyte (kB)?

Memory (RAM) sizes must be a power of 2, so 1 kB was traditionally  $2^{10}=1024$  bytes

Different from SI units version of “kilo” (1000, as in kilometer, kilogram, ...)

But it’s close!

So traditionally, 1 MB =  $2^{20}$  bytes = 1,048,576 bytes ; 1 GB =  $2^{30}$  B = 1,073,741,824 B

So off by over 7% for 1GB

Hard drive manufacturers revolted!

Wanted to advertise a  $2^{29}$  B drive as 537 MB rather than 512 MB - back to SI units!

Now: RAM typically in power-of-two units (some suggest KiB/MiB/GiB for this), and persistent storage in SI units. What about flash drives? If it’s important, get clarification!

---

# How much is...

1kB?

- Paragraph of text

1MB?

- 4 megapixel JPEG-compressed image

1GB?

- 30 minutes of DVD-quality SD TV
- 3.5 minutes at Blu-ray HD rate

1TB?

- 2,000 hours of audio (uncompressed)
- 17,000 hours as MP3s (255,000 4-minute songs)

1PB?

- DNA of the entire population of the US - three times over!
- Two months data from the planned Large Synoptic Survey Telescope



WD Gold 12TB Enterprise Class Hard Disk Drive - 7200 RPM Class SATA 6Gb/s 256MB Cache 3.5 Inch - WD121KRYZ

Be the first to review this product... Ask Or Answer A Question

SHARE

In stock. Limit 1 per customer. Ships from United States.

Sold and Shipped by Newegg

Capacity: 12TB

1TB 2TB 4TB 6TB 8TB 10TB 12TB

- Designed with a workload rating up to 550TB per year
- Up to 2.5M hours MTBF with a 5 year limited warranty
- 4th generation HelioSeal technology (12TB)

See also <http://www.jamesshuggins.com/h/tek1/how-big.htm>

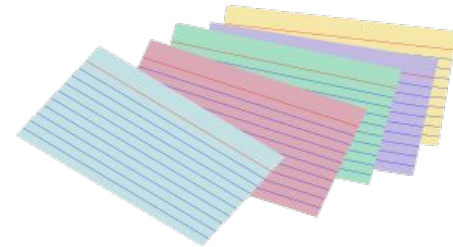
# A Flood of Data

## Recall from “Organizing Data” lecture

---

Consider the amount of data we deal with:

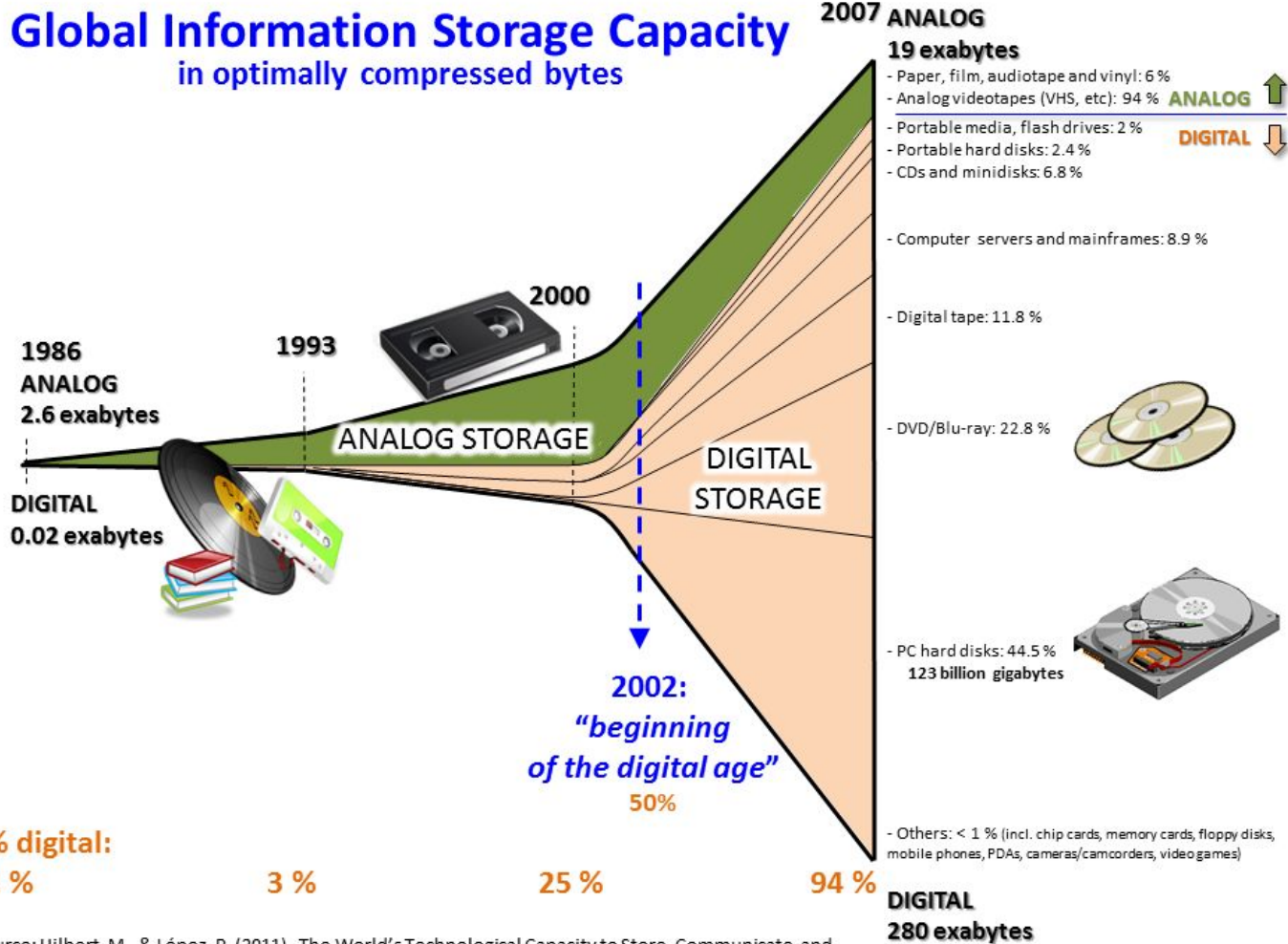
- Human genome: Just over 3 billion base pairs
  - *Typing in 12pt on 8.5x11 paper fits 2880 characters*
  - *So the human genome would be over a million pages (printed two-sided, an 86 foot high stack of paper)*
- Facebook - <http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/>
  - *2.01 billion monthly active users (1.32 billion daily!)*
  - *Messenger+WhatsApp: Over 60 billion messages/day*
    - *On index cards, would be a stack 7500 miles high!*
  - *... or end-to-end would stretch around the world 180 times*
- Large Synoptic Survey Telescope
  - *16 terabytes (16,000,000,000,000 bytes) will be captured per day*
  - *Most of this data will never be seen by a human being*
- Walmart customer transaction database
  - *Estimated to be approximately 2.5 petabytes*





# A Flood of Data

## Another view...



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

# So much data available

## *Some publicly-available big datasets*

---

### Some examples of available data:

- data.gov
    - Over 130,000 datasets on Oct 26, 2014
    - Census data, USGS Topo maps, house price indexes, NOAA Geophysical Data Center, ...
  - Amazon web services public data sets (<http://aws.amazon.com/datasets>)
    - Web crawl of over 5 billion web pages
    - “1000 Genomes Project”
    - Japan Census Data
    - Google Books Ngrams
  - UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets.html>)
    - All sorts of data for machine learning experiments
  - NCBI (<http://www.ncbi.nlm.nih.gov/genome>)
    - Genome information including sequences of many organisms
    - From the U.S. National Library of Medicine
-

# Making sense out of data - processing

## *Data mining: Finding patterns*

---

What do you do with lots of data - find patterns!

*An old urban legend: A supermarket analyzed purchasing data and found a correlation between purchases of diapers and beer that no one knew about. They put the two closer together in the store and..... profit!*

*The real story: Almost true... It was Osco Drug stores, not a supermarket. And while they found the correlation (between 5:00 and 7:00pm) they didn't actually change anything as a result - that was just a "what if..." comment that became legend.*

There are certainly real stories that are even more astounding:

<http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>

*From the story: "As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy."*

See also: Video at the end of "Big Data" lectures.

---

# Making sense out of data - processing

*Data mining: Finding patterns*

## THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate

JOURNAL REPORTS: LEADERSHIP

### Big Data Uncovers Some Weird Correlations

There's a Link Between Sales and Phases of the Moon, Among Other Things

By *Deborah Gage*

Updated March 23, 2014 4:36 p.m. ET

Are sales deals affected by the cycles of the moon? Is it possible to determine credit risk by the way a person types?

JOURNAL REPORT

- Insights from [The Experts](#)
- Read more at

Fast new data-crunching software combined with a flood of public and private data is allowing companies to test these and other seemingly far-fetched theories, asking

#### The Human Factor

The online lender ZestFinance Inc. found that people who fill out their loan applications using all capital letters default more often than people who use all lowercase letters, and more often still than people who use uppercase and lowercase letters correctly.

Privacy

- [Companies Embrace Real-Time](#)

businesses an advantage in an increasingly

# Making sense out of data - processing

## *Tools and Technologies - Software*

---

All the major tools are free / open source



Initial release in 2006

Based on Google technology

Main pieces:

- HDFS: Filesystem
- MapReduce: Processing
- YARN: Resource mgmt



“Data warehouse” solution

Originally from Facebook (2008)

Facebook’s implementation

- 150,000 tables
- 100 PB of storage



Cluster computing framework

Orig from UC Berkeley (2012)

10x-100x faster on some apps

# Making sense out of data - processing

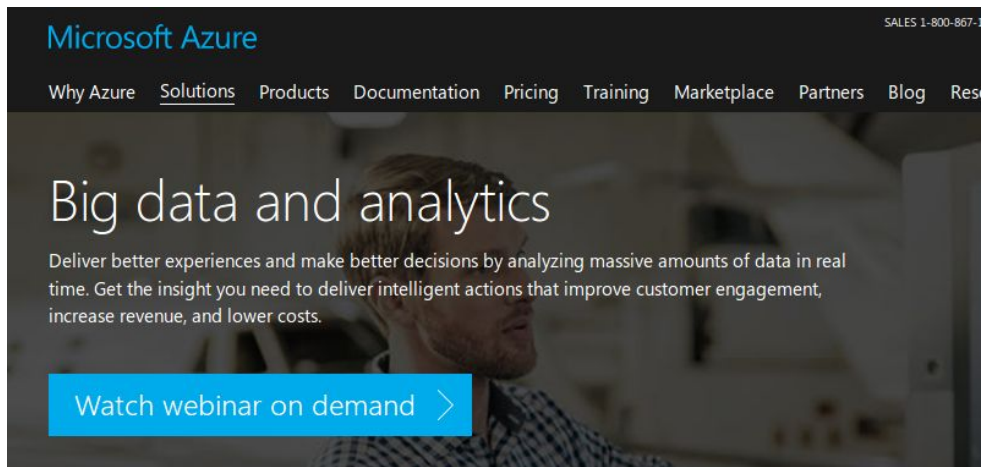
## *Tools and Technologies - Hardware*

---

Where to get such large computing resources? Lots of “cloud” resources to rent:



A blue banner for Amazon EMR. At the top center is a white icon of a network of nodes. Below it, the text "Amazon EMR" is written in white. Underneath that, a thin white line separates the text from the tagline: "Easily Run and Scale Apache Hadoop, Spark, HBase, Presto, Hive, and other Big Data Frameworks". At the bottom center is a white button with the text "Get started with Amazon EMR".



A dark banner for Microsoft Azure. The top left corner has the "Microsoft Azure" logo in light blue. The top right corner has the text "SALES 1-800-867-1344" in small white font. Below the logo is a navigation menu with items: "Why Azure", "Solutions", "Products", "Documentation", "Pricing", "Training", "Marketplace", "Partners", "Blog", and "Resources". The main text reads "Big data and analytics" in large white font. Below it is a paragraph: "Deliver better experiences and make better decisions by analyzing massive amounts of data in real time. Get the insight you need to deliver intelligent actions that improve customer engagement, increase revenue, and lower costs." At the bottom left is a blue button with white text: "Watch webinar on demand >".



A light blue banner for Cloudera SDX. The main text reads "Be one with your data" in large black font, followed by "Introducing SDX. Experience your data whenever, wherever, and however you'd like." in smaller black font. Below this is a blue link: "Learn more >". At the bottom is the Cloudera logo (the word "cloudera" in small black font) above the large "sdx" logo in blue, with the tagline "shared data experience" in small black font below it.

# Making sense out of data - viewing

## *Visualization*

---

How can we view data? A couple of on-line examples:

### D3: Data-Drive Documents

Technology allows interactive data presentation

Example: <http://benschmidt.org/Degrees/>

D3 toolkit: <http://d3js.org/>

### Dataseed: View data along several dimensions

Demos: <https://getdataseed.com/demo>

### Other viz tools:

<http://selection.datavisualization.ch/>

---

# Utilizing Data

## Why Companies Care....

---



### Big data in action: UPS

As a company with many pieces and parts constantly in motion, UPS stores a large amount of data – much of which comes from sensors in its vehicles. That data not only monitors daily performance, but also triggered a major redesign of UPS drivers' route structures. The initiative was called ORION (On-Road Integration Optimization and Navigation), and was arguably the world's largest operations research project. It relied heavily on online map data to reconfigure a driver's pickups and drop-offs in real time.

The project led to savings of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes. UPS estimates that saving only one daily mile per driver saves the company \$30 million, so the overall dollar savings are substantial.

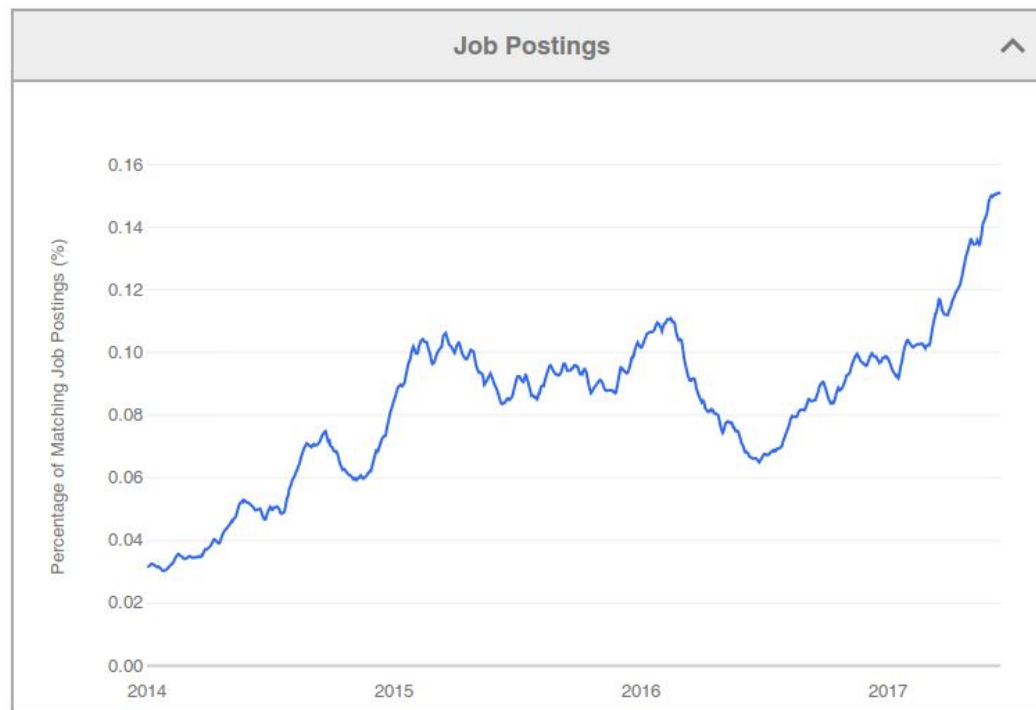


# Data Analytics / Data Science

*A valuable and growth-area skill*

---

Job postings for “Data Scientist” (5x increase in past 3 years):



Indeed searches millions of jobs from thousands of job sites. The jobseeker interest graph shows the percentage of jobseekers who have searched for "Data Scientist" jobs. The job postings graph shows trends for jobs containing "Data Scientist".

Valuable contests to demonstrate skills and develop techniques:

<http://www.kaggle.com/>

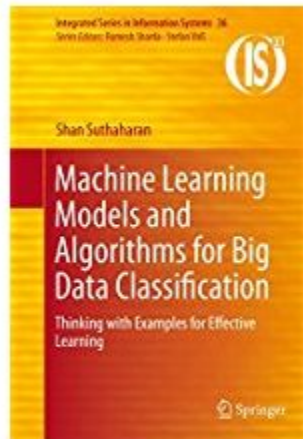
---

# What's happening at UNCG?

---

## Dr. Suthaharan

- Over a decade of work in cluster computing / big data
- Wrote a very popular book on Big Data classification (over 60,000 downloads of e-book)



**Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning (Integrated Series in Information Systems)**  
by Shan Suthaharan

Kindle Edition  
**\$109<sup>99</sup>**



Paperback  
**\$141<sup>55</sup>** ~~\$149.00~~ ✓prime

FREE Shipping on eligible orders  
Available for Pre-order. This item will be released on  
November 4, 2017.

Other Formats: [Hardcover](#)

- Has supervised >30 master's student project/theses in past 5 years
  - Teaches CSC 510 (Big Data and Machine Learning)
-

# What's happening at UNCG?

---

Dr. Mohanty

Before UNCG: Was lead at the “Innovative Data Laboratory” at Mississippi State

Some past projects:

- Anomaly Detection in High Velocity Streaming Data
- Social Media Tracking and Analysis System
- Organic Social Media During Severe Weather Events
- Networks of Research

Teaches CSC 495 (Data Science)



Lambda “DevBox” - Deep Learning Machine

- 14,336 GPU computing cores
- > 40 Gbps memory throughput
- Machine learning software...

# Summary

---

## Main take-aways:

- More data than ever before being collected and used
- Must be able to manage the data
- Making sense out of the data is a very valuable skill
  - Analysis, mining, and visualization are all parts of this

## What you should have gotten from this lecture:

- A sense for data sizes
  - An idea of what data is out there: available and private
  - Some ideas and pointers for how data is used
-

# And finally... a video

---

## Relevant information:

- *NY Times Magazine* story on data mining  
<http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>
- Forbes story “How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did”:  
<http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

## The entertaining take:

<http://www.cc.com/video-clips/dv9iqc/the-colbert-report-the-word---surrender-to-a-buyer-power>

---