2018 CDSE Days

April 9 – 13, 2018 | University at Buffalo | Buffalo, NY

DATA FROM & FOR EVERYBODY

The Role of Spreadsheet Software in the Current Academic and Non-Academic Environments:

A First Line of Defense Against the Deluge of Big, Fast, Varied, Dirty and Valuable Data

Thursday, April 12th, 11:30 am - 12:30 pm, UB Student Union 201

Joaquin Carbonara, SUNY Buffalo State

2018 CDSE Days

April 9 – 13, 2018 | University at Buffalo | Buffalo, NY

DATA FROM & FOR EVERYBODY

Thanks to

CDSE days organizers --Abani Patra in particular—

UB and SUNY institutions for leading the way into making Western New York a powerhouse in DS&A

my students Harsha Kankanamge and

Michel Schultz for help preparing the demos.

John Ringland for his leadership in advancing the use of computing tools in Applied Mathematics

Sara for interesting discussions about data and life.

Agenda

1. Software and the information revolution
   - The internet provides access to files for everybody. The internet of (indexed) data will bring access to raw data for everybody.
   - Spreadsheet software has been the first line of defense for the data deluge for many years now
2. Comments about the role of computing inside and outside academia
3. Data Science and Analytics
4. Spreadsheet software demos within the DS&A context

Introduction – DATA FOR ALL and the INFORMATION REVOLUTION

- Data Analytics Across Disciplines (Sciences, Art,…)
  - Programming and coding languages (MAX, python, Excel,…)
- Data Analytics Across Gender
- Data Analytics Across Cultural and Social Silos
- Data Analytics Across Devices (Edge computing, IoT, …)
  - Video surveillance – Router – Network - Cell phone/computer

Introduction – DATA FOR ALL and the INFORMATION REVOLUTION

- Data needs to have the following properties:

Fluidity (across all media and languages**)**
Integration (components should work together**)**
Transparency (all components should update each other**)**
Availability (components should be reachable at all times**)**
Security (RSA encryption, data and identity integrity**)**

Software, coding and Data Science

- There are many reasons for using software depending on the field:
  - **Teaching** related (k-12 and 13-16+): Illustrating concepts, preparing students for industry, discipline support in general, testing new technologies…
  - **Government**: transparency, integrity, governance…
  - **Industry**: improved control of processes, accounting…
  - **Research**: design, implement, simulate…
  - **Personal:** fitness, gaming, time management

Software, coding and Data Science

- But more important than any intellectual, political and social reason AND for the same reason the Internet took off in the 90's…

DATA IS THE NEW OIL

&

INFORMATION IS THE NEWEST HOTTEST COMMODITY

- Data fuels the new economy.
- Raw Data provides a "NEW" kind of information to the general user ("REAL" and "PURE" information). Different than we got from the Internet.

Why software, coding and Data Science?

- Tim-Berners Lee video on Ted.com (2009)
  - TBL invented the internet
  - TBL and Raw Data
  - TBL and what we can do with Raw Data

Data reports that may help understand the role of different software

- The following reports were created using data from Burning Glass Technologies (Real-Time Job Market Analytics Software)
- (1) A team in Rutgers analyzed trends in skill requirements for Data Analytics jobs. (c. 2015)
- (2) A well know higher education analyst looked at skills requirements very broadly in the US job market. (Presented Aug 2017)

Burning Glass data 1

- https://mbs.rutgers.edu/articles/numbers-making-sense-job-titles-analytics [file]
- Using a tool called Labor Insight from Burning Glass (a real-time labor market data products and analysis company), which de-duplicates, parses, and allows you to mine data from on-line job postings, this survey examined the skills most frequently sought by employers across thousands of on-line job ads.

Burning Glass data 1

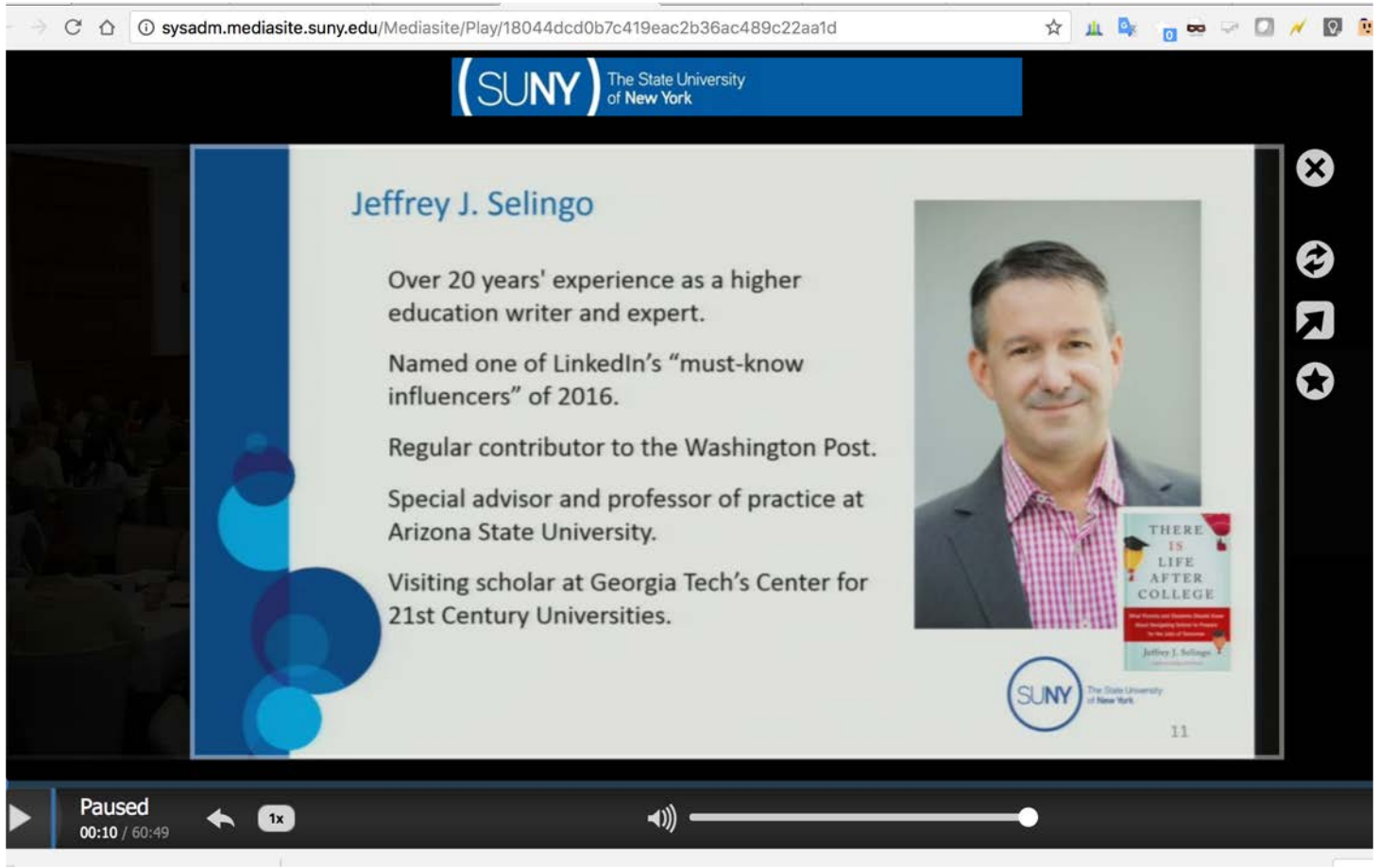
- **Data Scientist**
  Data Scientists are one of the most frequently advertised job titles in the Analytics field. There were 7,753 on-line job postings for Data Scientists across the U.S. In the last 12 months, according to our analysis of Burning Glass data. Python was the most commonly requested programming language, though still only about half of all job ads for Data Scientists included this. About half of all ads also focused on Machine Learning and SQL, while 43% required Mathematics. About 1/3 of the ads for this title were seeking people who had R, data mining, Hadoop, SAS, and/or JAVA skills.

Burning Glass data 1

- Other job titles in this report included**:**
  - ***Title:*** **Data Engineer**
  - ***Title:*** **Data Analyst (Analytics Specialty)**
  - ***Title:*** **Business Intelligence Analyst**
  - ***Title:*** **Consultant (Analytics)**
  - ***Title:*** **Big Data Software Developer**

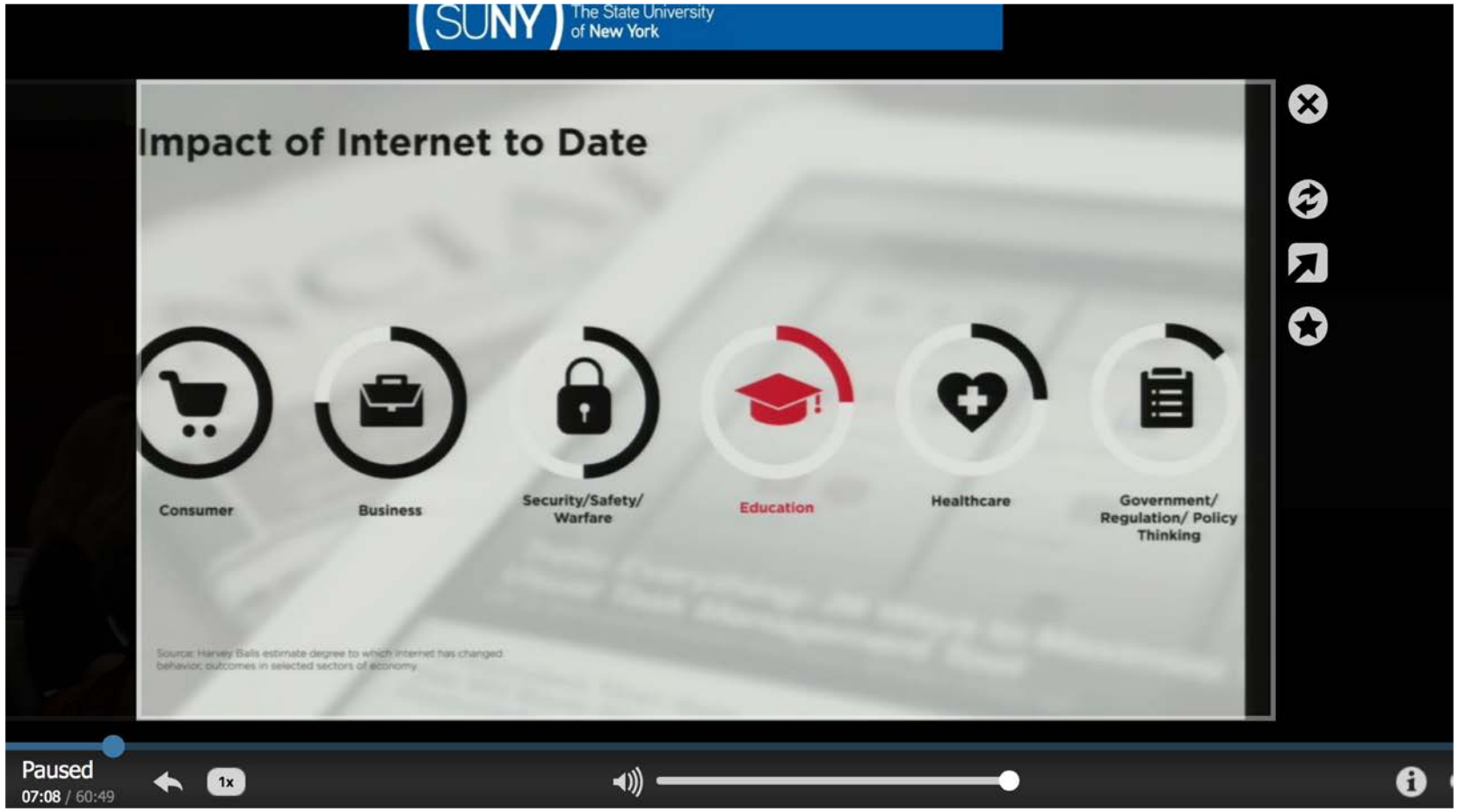
- None of the jobs here required knowledge of Excel. Tableau was mentioned in three places (also a spreadsheet software).

Burning Glass data 2

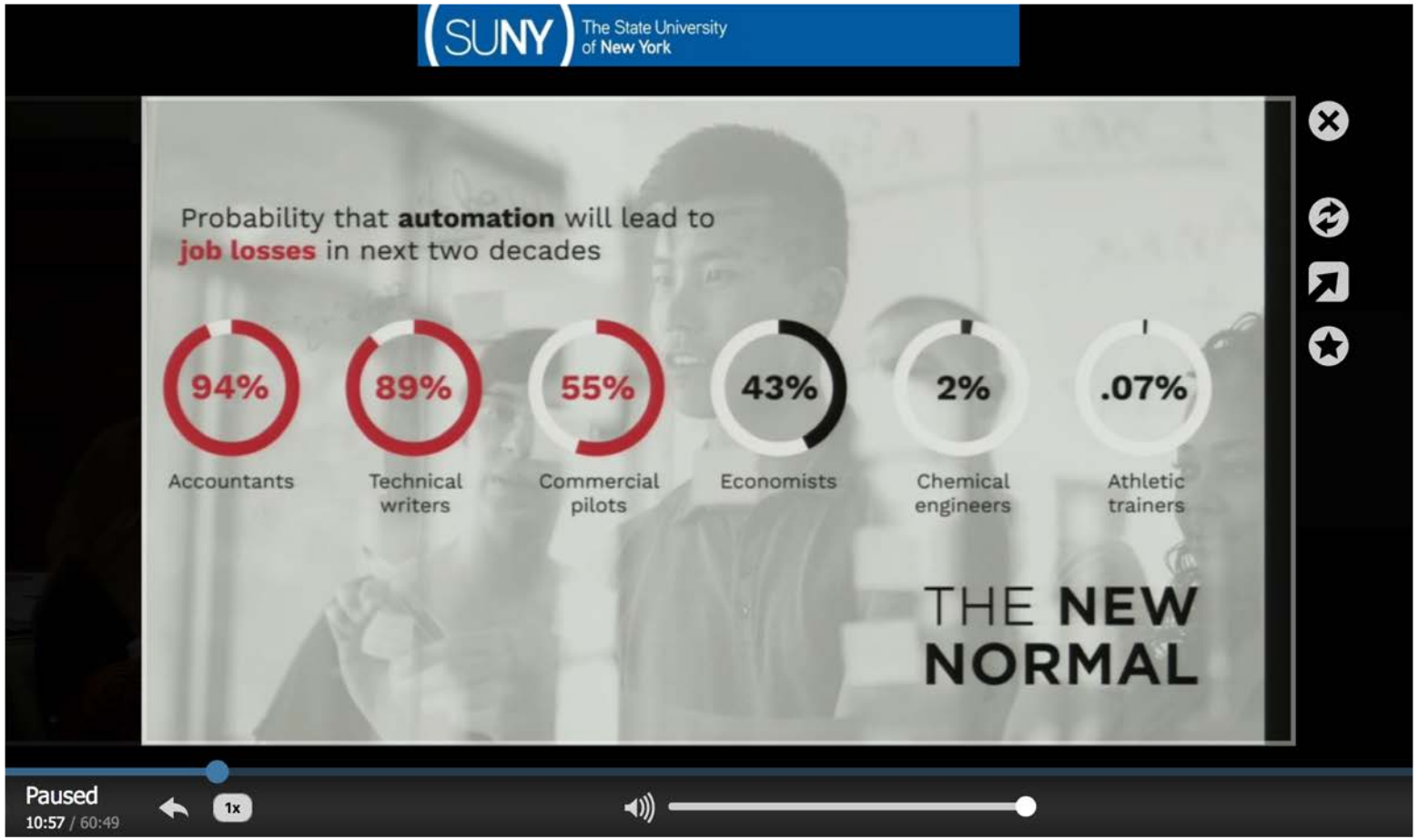


Burning Glass data 2

Burning Glass data 2



Burning Glass data 2



Burning Glass data 2

- This is a 1 minute cut from the presentation given on August 2017 in SUNY central by Jeffrey Salingo:

Video

Burning Glass data 2

WHAT DOES DATA TELL YOU? – A joke about immediacy…

Sherlock Holmes and **Dr. Watson** decide to go on a camping trip. After dinner and a bottle of wine, they lay down for the night, and go to sleep.

Some hours later, Holmes awoke and nudged his faithful friend.

"Watson, look up at the sky and tell me what you see."

Watson replied,

"I see millions of stars."

"What does that tell you?"

Watson pondered for a minute.

"Astronomically, it tells me that there are millions of galaxies and potentially billions of planets."
"Astrologically, I observe that Saturn is in Leo."
"Horologically, I deduce that the time is approximately a quarter past three."
"Theologically, I can see that God is all powerful and that we are small and insignificant."
"Meteorologically, I suspect that we will have a beautiful day tomorrow."
"What does it tell you, Holmes?"

WHAT DOES DATA TELL YOU? – A joke about immediacy…

Sherlock Holmes and **Dr. Watson** decide to go on a camping trip. After dinner and a bottle of wine, they lay down for the night, and go to sleep.

Some hours later, Holmes awoke and nudged his faithful friend.

"Watson, look up at the sky and tell me what you see."

Watson replied,

"I see millions of stars."

"What does that tell you?"

Watson pondered for a minute.

"Astronomically, it tells me that there are millions of galaxies and potentially billions of planets."
"Astrologically, I observe that Saturn is in Leo."
"Horologically, I deduce that the time is approximately a quarter past three."
"Theologically, I can see that God is all powerful and that we are small and insignificant."
"Meteorologically, I suspect that we will have a beautiful day tomorrow."
"What does it tell you, Holmes?"

Holmes was silent for a minute, then spoke:

"Watson, you idiot. Someone has stolen our tent!"

Quiz of the day for extra credit

- How do you know a wanna-be-Comedian is really a professor?

Answer

- The Jokes are on SLIDES

Answer

- You get a quiz after the joke

Stages in the evolution of Data Analysis

- Anecdotal information collected during 10 years of intense involvement with companies through the US because of being part of the National Professional Science Master's Association:
  - Most small companies are just now starting to migrate from Excel to more robust software (e.g. MySQL, SAS and Python). Even then, Excel is all over the place.

- Why?…
  - is it clear to companies what Data is all about?
  - It took a decade for companies to learn how to make money from the advent of the Internet.

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

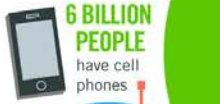
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

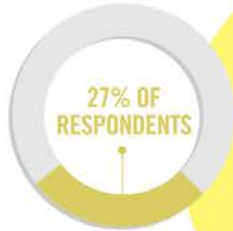
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



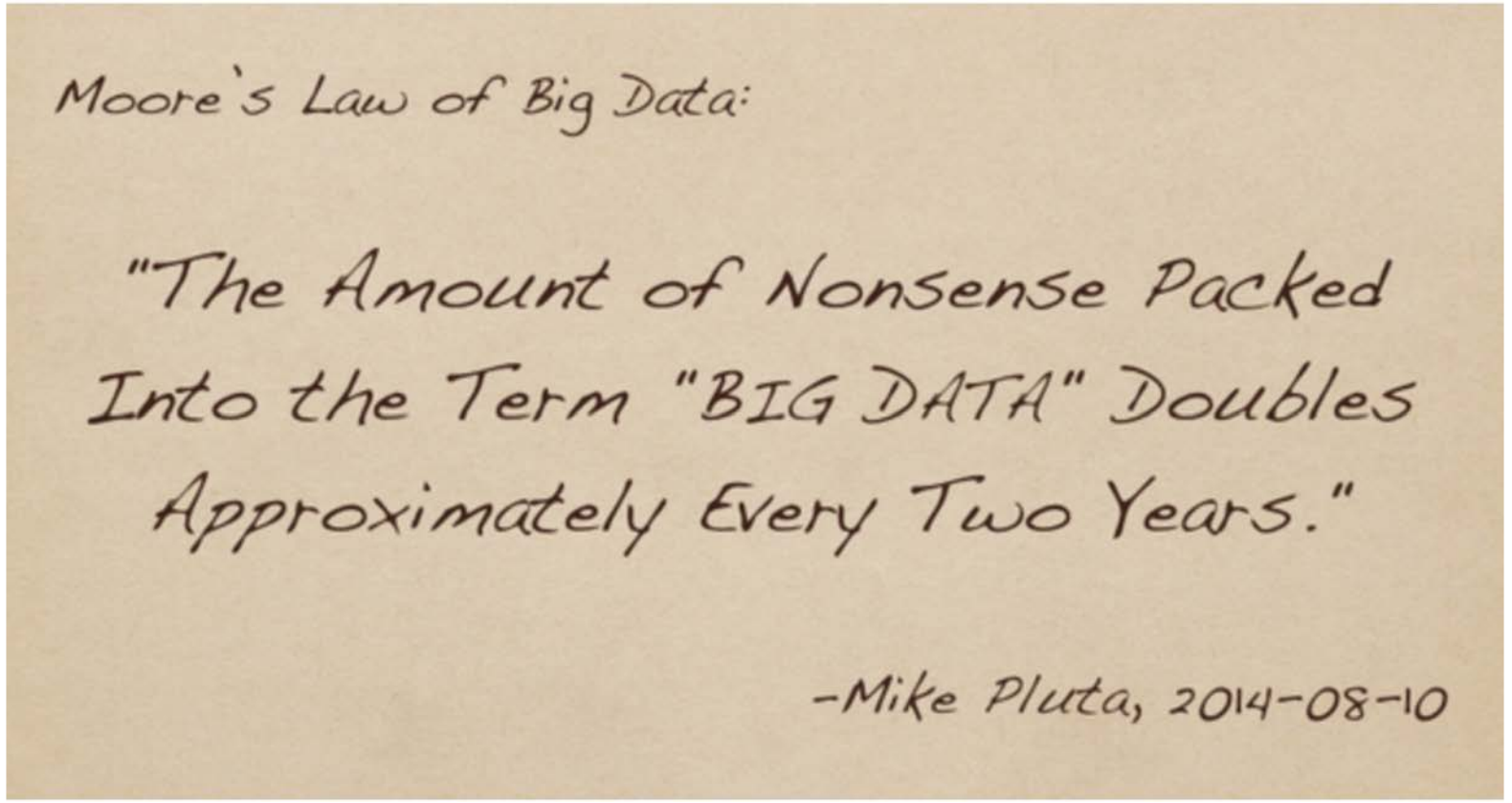
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Big Data? Or just Data?



How many V’s can you come up with to describe what data is?

- Take a look at the record holder:

Big Data by Tom Shafer, Elder Research, Inc.

Dozens of spreadsheet software are available

- The most popular SS available include
  - Excel (proprietary economic model)
  - LibreOffice (open source economic model)
  - Google Sheets (omnipresent, omnipotent, beyond economic modeling)

- What are their advantages and limitations?

Can Spreadsheet Software handle processing of data…

Using software for data processing creates a cycle

- 1) Collection.
- 2) Preparation (Raw data cannot be processed).
- 3) Input (verified data is coded or converted into machine readable).
- 4) Processing (when the data is subjected to various means and methods of manipulation).
- 5) Output and interpretation (the stage where processed information is now transmitted to the user).
- 6) Storage.

Can Spreadsheet Software handle processing the data…

Using software for data processing creates a cycle

- 1) Collection.
  - Can read any CSV file collected from any data source
  - Live data from the web (video)
  - More to come in other slides…

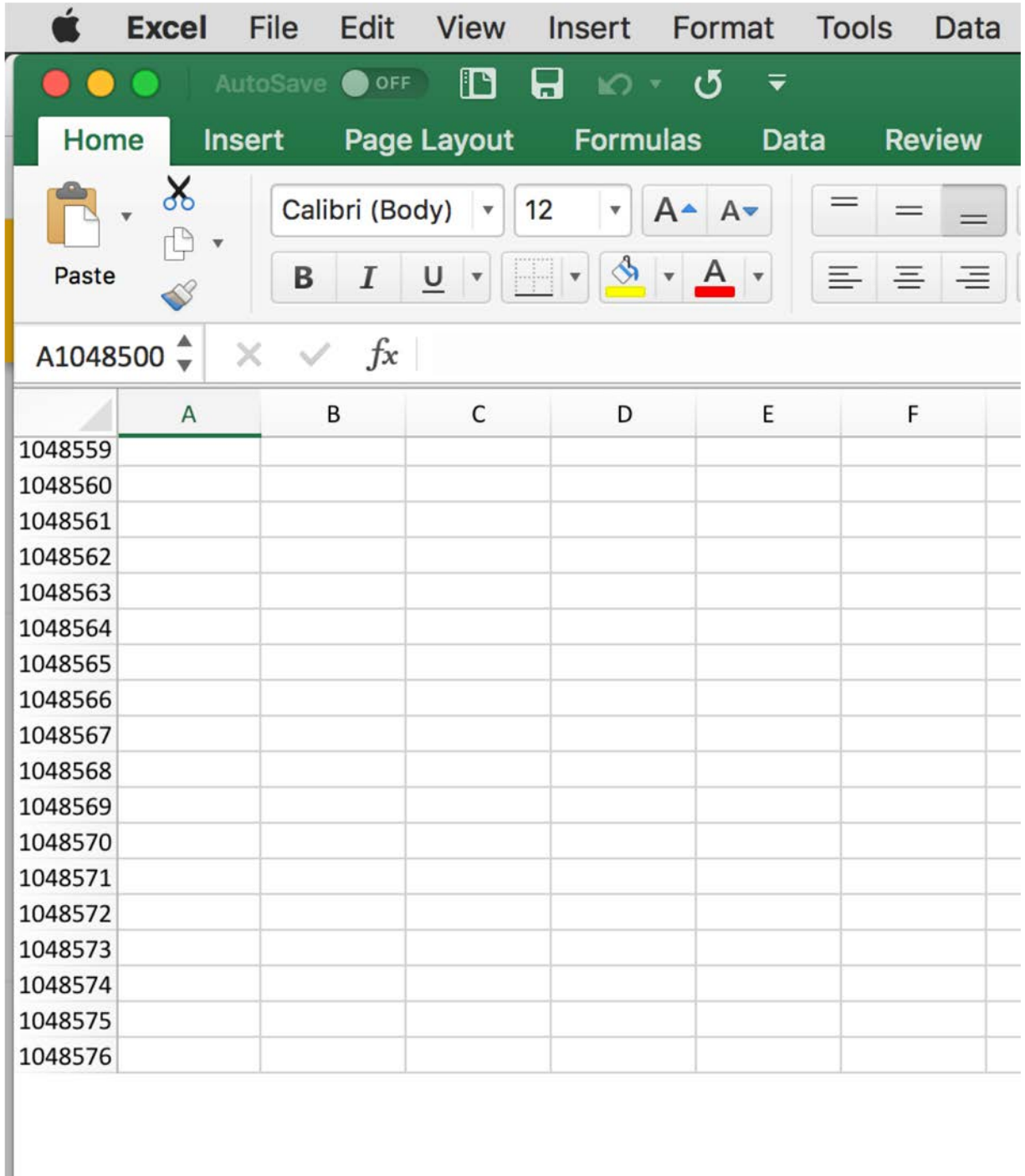
Can Spreadsheet Software handle processing the data…

Using software for data processing creates a cycle

- 6) Storage.
  - Software versions as well as bells and whistles of different versions cause problems.
  - Size is limited

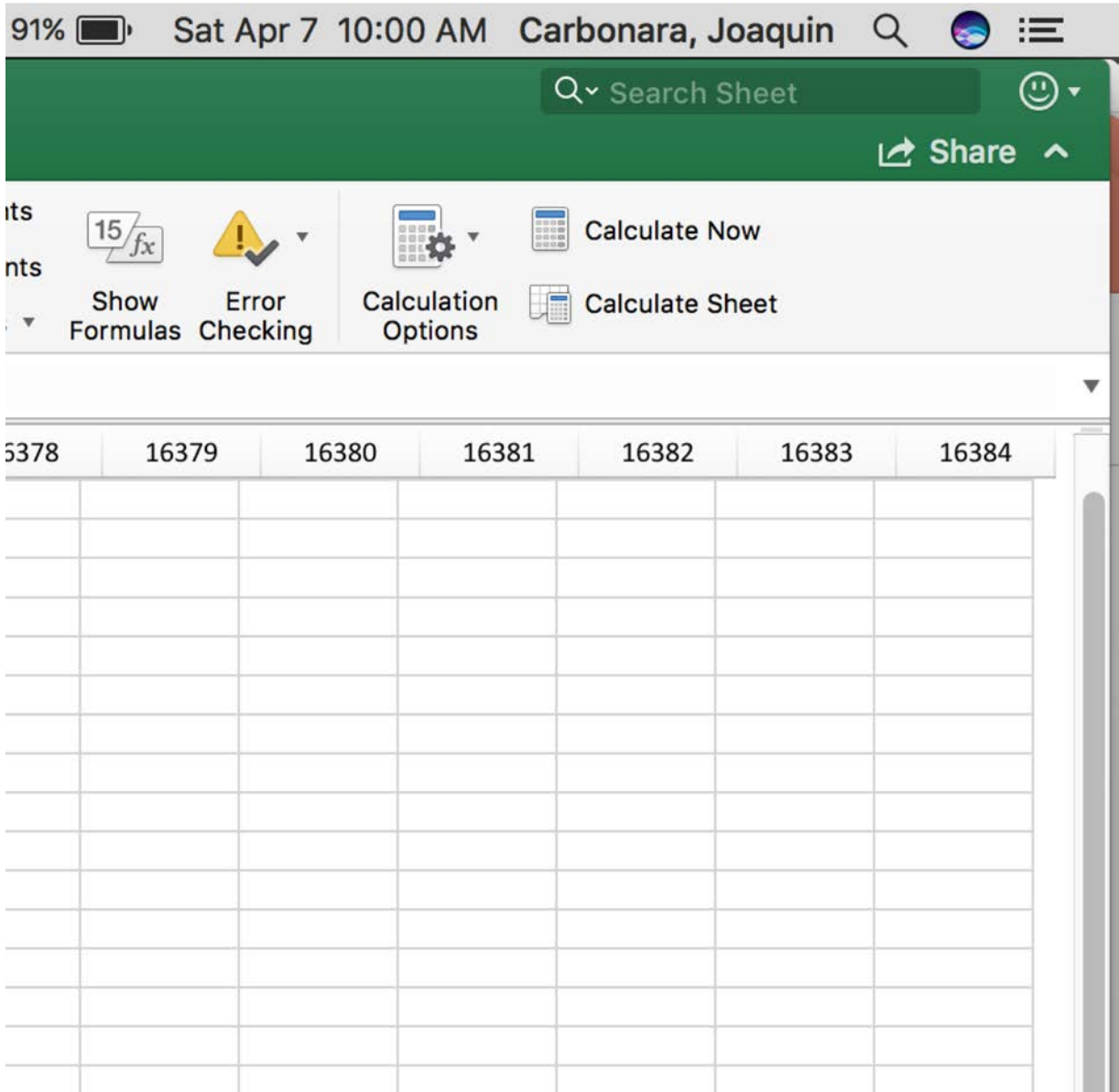
Data Table size

- Rows in Excel



Data Table size

- Columns in Excel



Having data on a grid can be appealing

- Data output in 2 dimensions: Usual output from computations are 1 dimensional (i.e. they output a single data structure – a number or an array). Having a grid of cells where output data can go can be very appealing.
- Xlwings is a package in Python that allows interaction between python and Excel.

Can Spreadsheet Software handle processing the data…

Using software for data processing creates a cycle

- 2,3) Preparation, input, processing
  - Functions and Array-functions
    - Demo 1 (Reference: Walkenbach, John. Excel 2013 Formulas)
  - Replacing VBA with python
    - Demo 2

Can Spreadsheet Software handle processing the data…

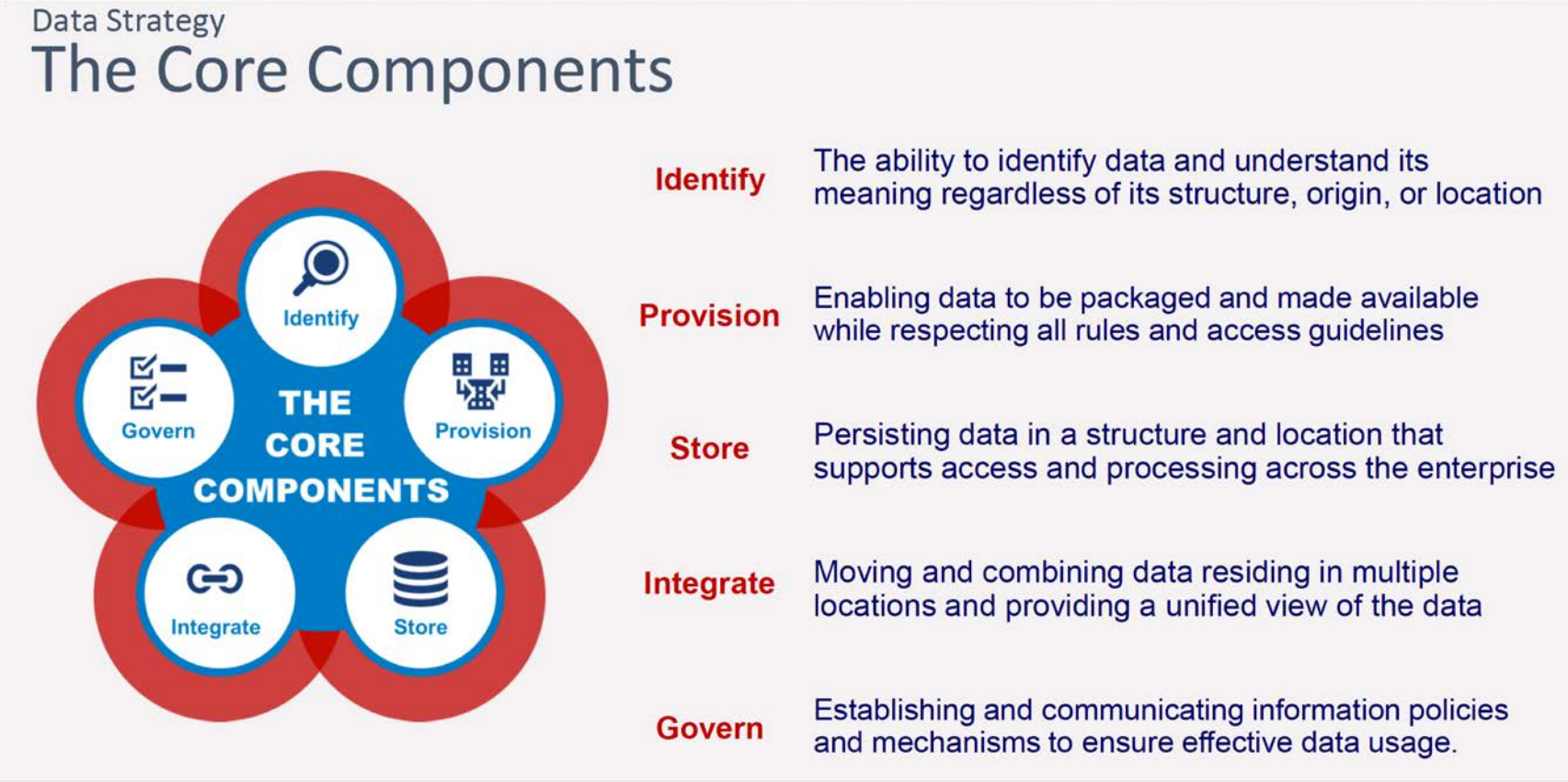
Using software for data processing creates a cycle

- 5) Visualization

  Storytelling With Data: A Data Visualization Guide for Business Professionals by Cole Nussbaumer Knaflic

What about everything else…

E. Levy (SAS)



What about everything else…

- A Bitcoin is data that only exists if a distributed network approves it. So, what kind of data are Blockchains?
- Cryptographic signed hashes only are a proof of existence
- Artificial intelligence and Artificial General Intelligence
- Deep learning

Conclusion…

- Choosing software will depend a lot on what will happen to Data in the future.
- Tables, databases (relational, graph based, ontologies) and other ways to store and recall data are constantly evolving.
- Ways to transform bits (0-1) into organized text is also changing (XML, JSON).
- There is evidence that the equivalent to verbal language (storage and recall), mathematical language (perception and abstraction) and art (creation, intuition, decision) encompass all we want Data to handle in the future.

Thanks for being here….

- Questions