BlackComputeHER

A Taste of Data Science [Python]

Brandeis Marshall, Founder/CEO
Dataedx Group, LLC
April 5, 2019

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Master Class Outline

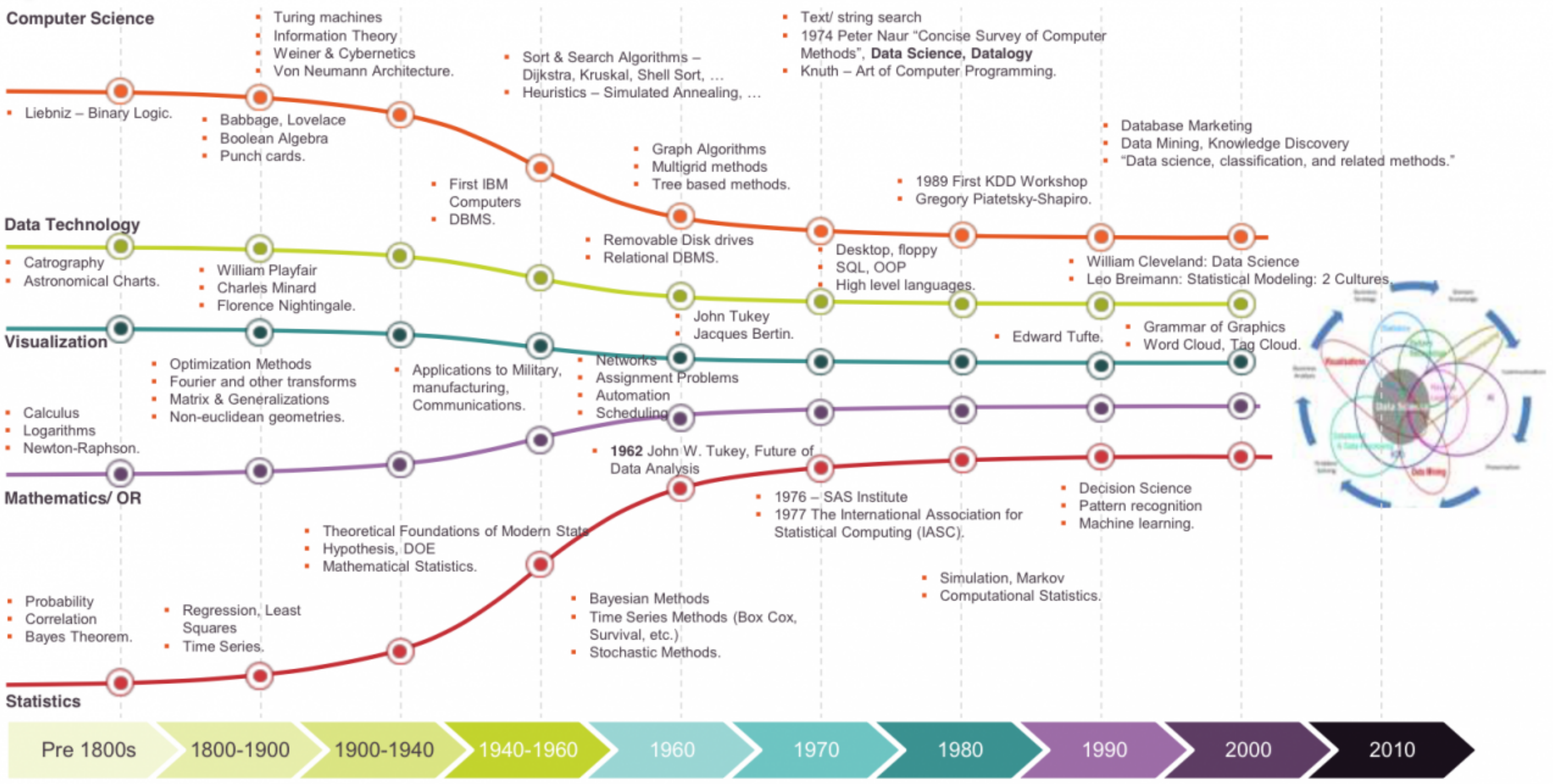
- Background
- Race & Data
- What data science is
- The misconceptions of data science
- Hands-On with Jupyter Notebook

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Background

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

History of Data Science



Computer Science

Data Technology

Visualization

Mathematics/ OR

Statistics

- Liebniz – Binary Logic.
- Catrography
- Astronomical Charts.
- Calculus
- Logarithms
- Newton-Raphson.
- Probability
- Correlation
- Bayes Theorem.

- Turing machines
- Information Theory
- Weiner & Cybernetics
- Von Neumann Architecture.

- Babbage, Lovelace
- Boolean Algebra
- Punch cards.

- William Playfair
- Charles Minard
- Florence Nightingale.

- Optimization Methods
- Fourier and other transforms
- Matrix & Generalizations
- Non-euclidean geometries.

- Regression, Least Squares
- Time Series.

- First IBM Computers
- DBMS.

- Applications to Military, manufacturing, Communications.

- Theoretical Foundations of Modern Stats
- Hypothesis, DOE
- Mathematical Statistics.

- Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, …
- Heuristics – Simulated Annealing, …

- Graph Algorithms
- Multigrid methods
- Tree based methods.

- Removable Disk drives
- Relational DBMS.

- John Tukey
- Jacques Bertin.

- Networks
- Assignment Problems
- Automation
- Scheduling

- **1962** John W. Tukey, Future of Data Analysis

- Bayesian Methods
- Time Series Methods (Box Cox, Survival, etc.)
- Stochastic Methods.

- Text/ string search
- 1974 Peter Naur "Concise Survey of Computer Methods", **Data Science, Datalogy**
- Knuth – Art of Computer Programming.

- Desktop, floppy
- SQL, OOP
- High level languages.

- 1976 – SAS Institute
- 1977 The International Association for Statistical Computing (IASC).

- 1989 First KDD Workshop
- Gregory Piatetsky-Shapiro.

- Edward Tufte.

- Simulation, Markov
- Computational Statistics.

- Database Marketing
- Data Mining, Knowledge Discovery
- "Data science, classification, and related methods."

- William Cleveland: Data Science
- Leo Breimann: Statistical Modeling: 2 Cultures.

- Grammar of Graphics
- Word Cloud, Tag Cloud.

- Decision Science
- Pattern recognition
- Machine learning.

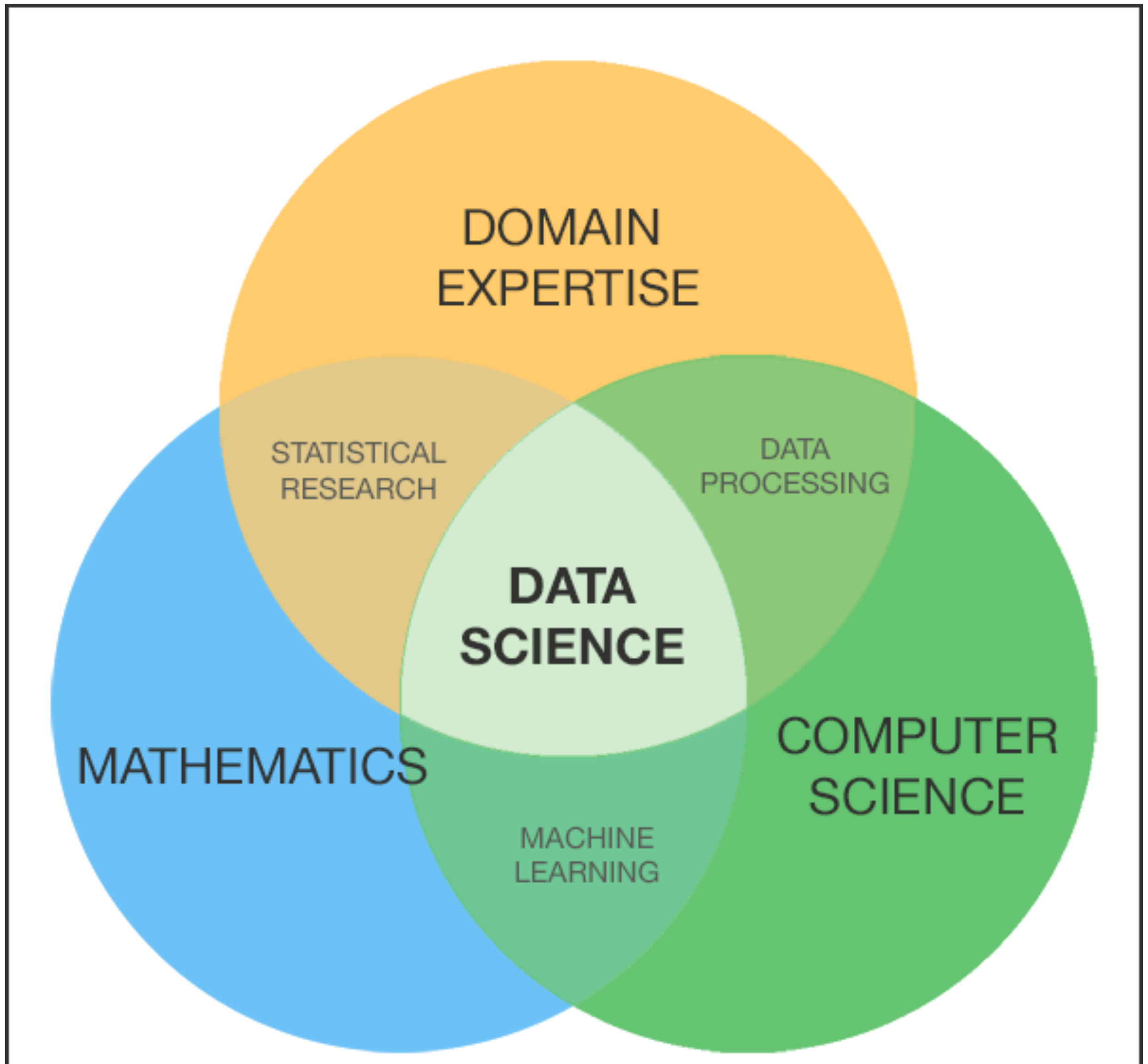
Pre 1800s | 1800-1900 | 1900-1940 | 1940-1960 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010

Impact of Big Data on Analytics by Mamatha Upadhyaya

https://dataedxgroup.co

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

History of Data Science

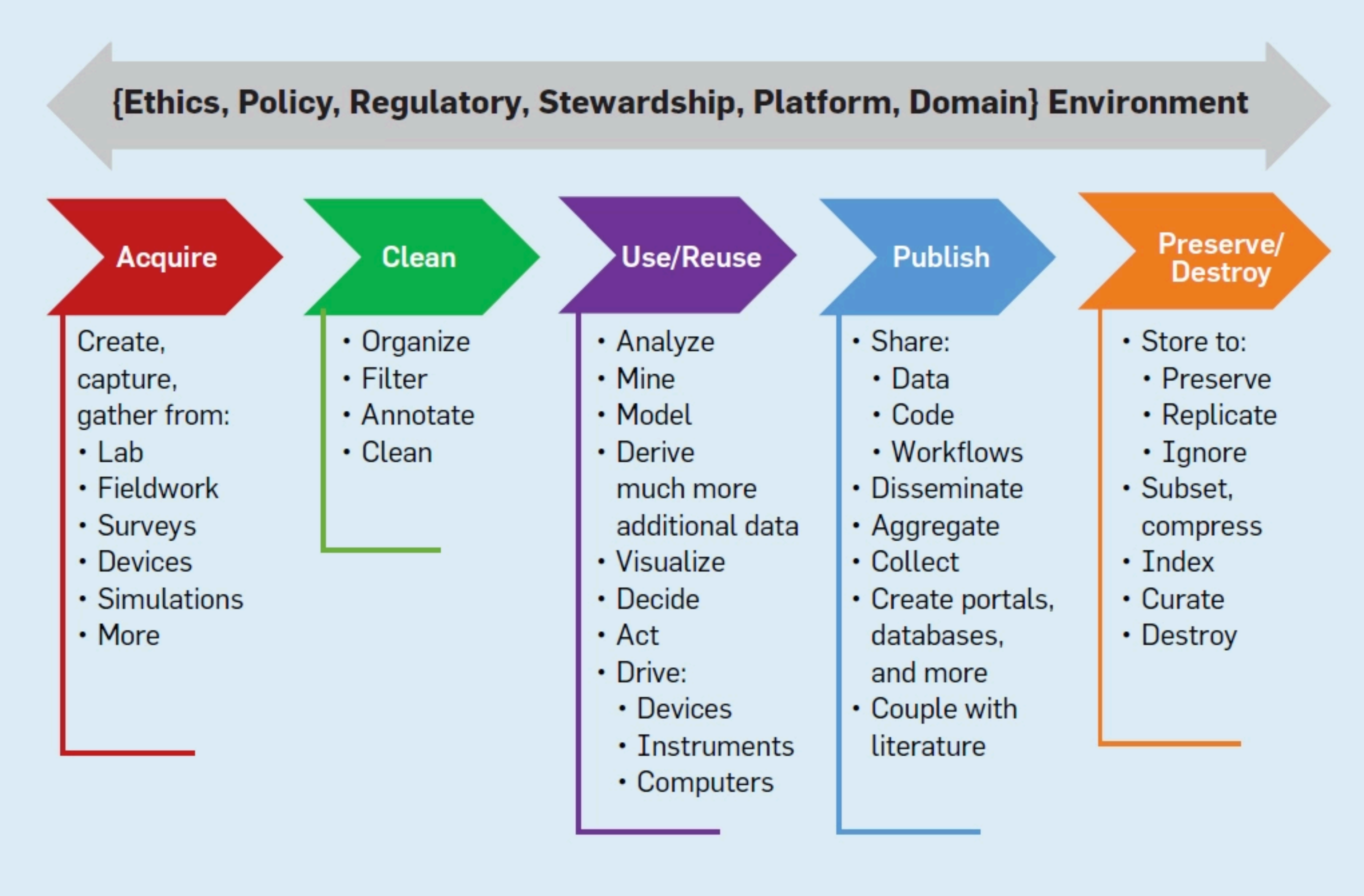


Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

- Data Science is represented as a blend of AT LEAST 2 existing disciplines + domain-specific applications

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

History of Data Science



The Data Life Cycle

Berman et al. (2016) Realizing the Potential of Data Science

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Race & Data

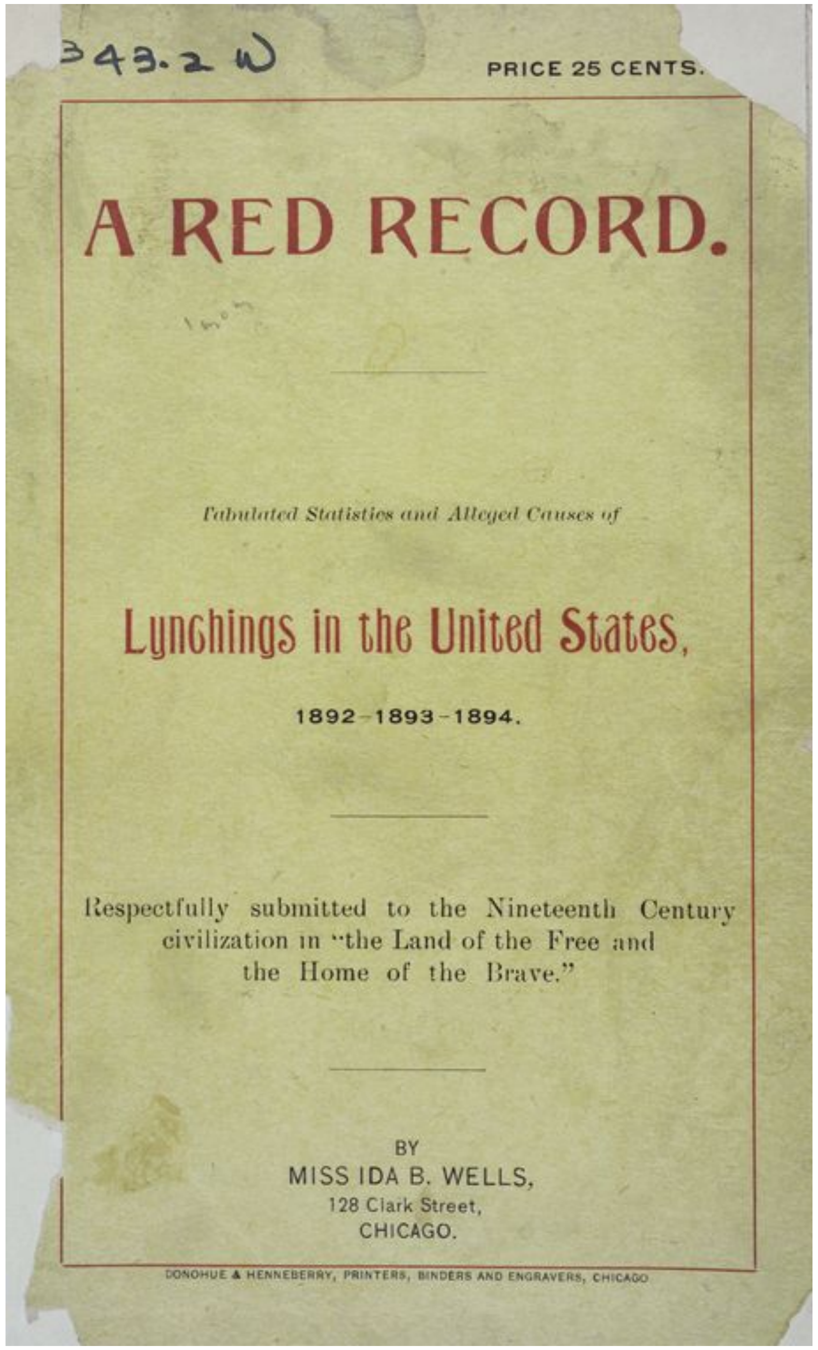
Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Race & Data

- On December 19, 2018, the U.S. Senate passed the Justice for Victims of Lynching Act of 2018, that amended title 18, United States Code, to specify lynching as a deprivation of civil rights, and for other purposes.
- The legislation was introduced in June 2018 by the three African American members of the U.S. Senate: Senators Kamala Harris (D-Calif.), Tim Scott (R-SC) and Cory Booker (D-NJ).
- EJI — Montgomery, Alabama
  - The Legacy Museum: From Enslavement to Mass Incarceration (April 2018)
  - The National Memorial for Peace and Justice (April 2018)
- NAACP (1909)
- Ida B. Wells Barnett (1895)

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Race & Data



343.2 W

PRICE 25 CENTS.

A RED RECORD.

Tabulated Statistics and Alleged Causes of

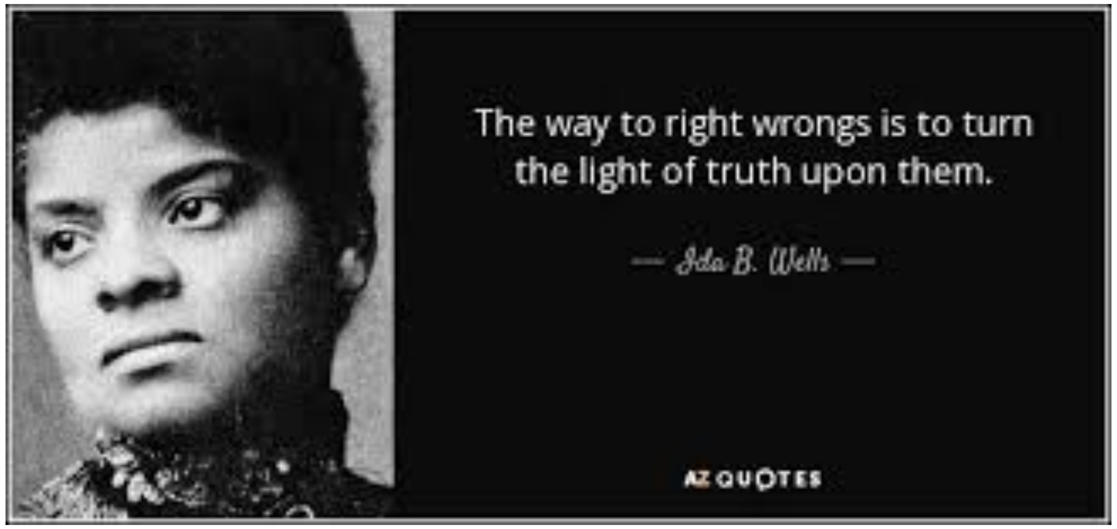
Lynchings in the United States,

1892-1893-1894.

Respectfully submitted to the Nineteenth Century civilization in "the Land of the Free and the Home of the Brave."

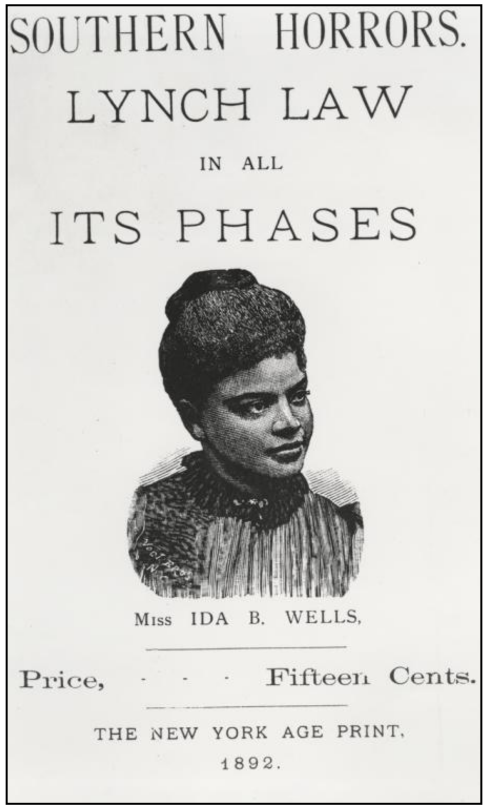
BY
MISS IDA B. WELLS,
128 Clark Street,
CHICAGO.

DONOHUE & HENNEBERRY, PRINTERS, BINDERS AND ENGRAVERS, CHICAGO



The way to right wrongs is to turn the light of truth upon them.

— Ida B. Wells —



SOUTHERN HORRORS.

LYNCH LAW

IN ALL

ITS PHASES

Miss IDA B. WELLS,

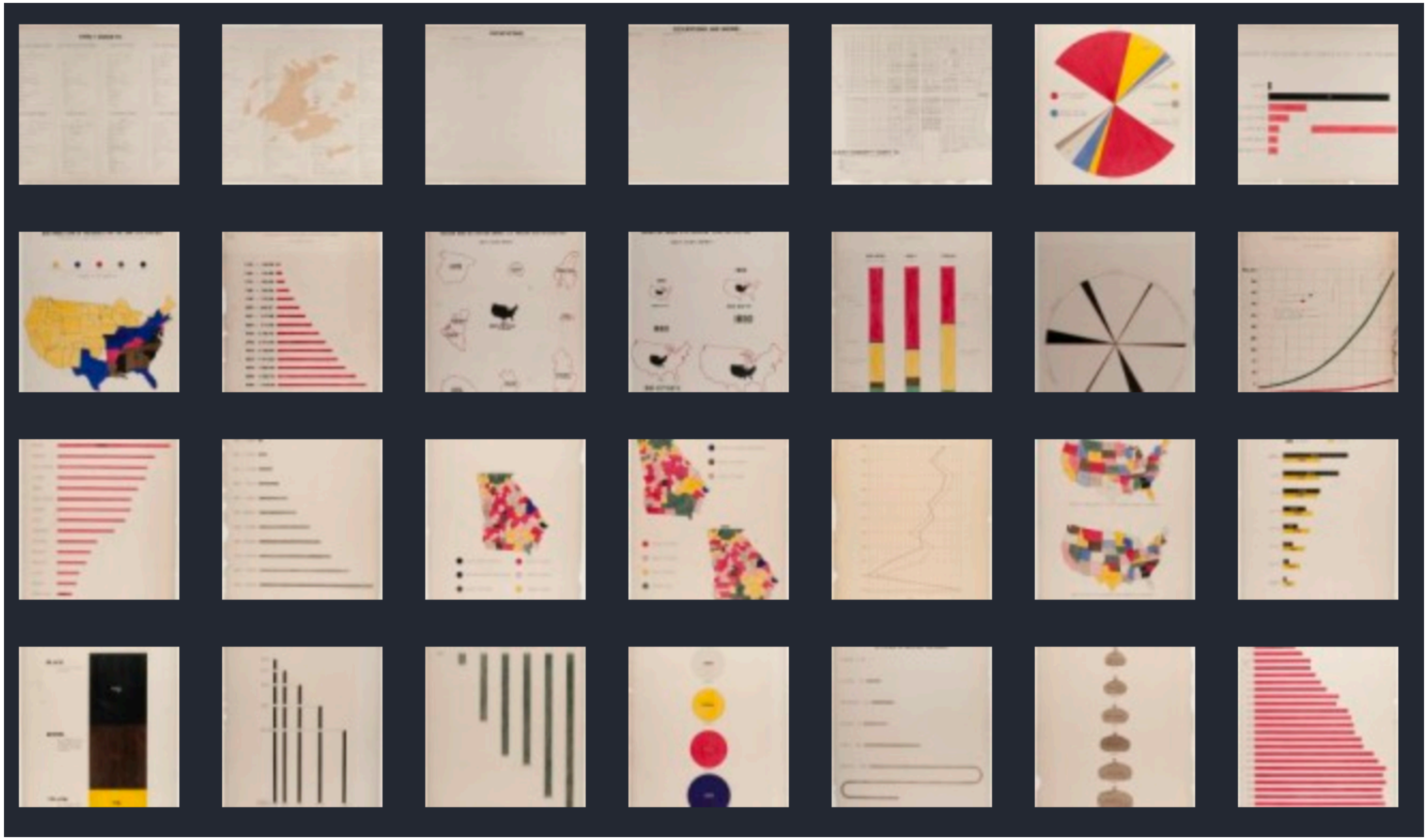
Price, - - - Fifteen Cents.

THE NEW YORK AGE PRINT,
1892.

A chronicle of lynchings in the United States from 1892-1894

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

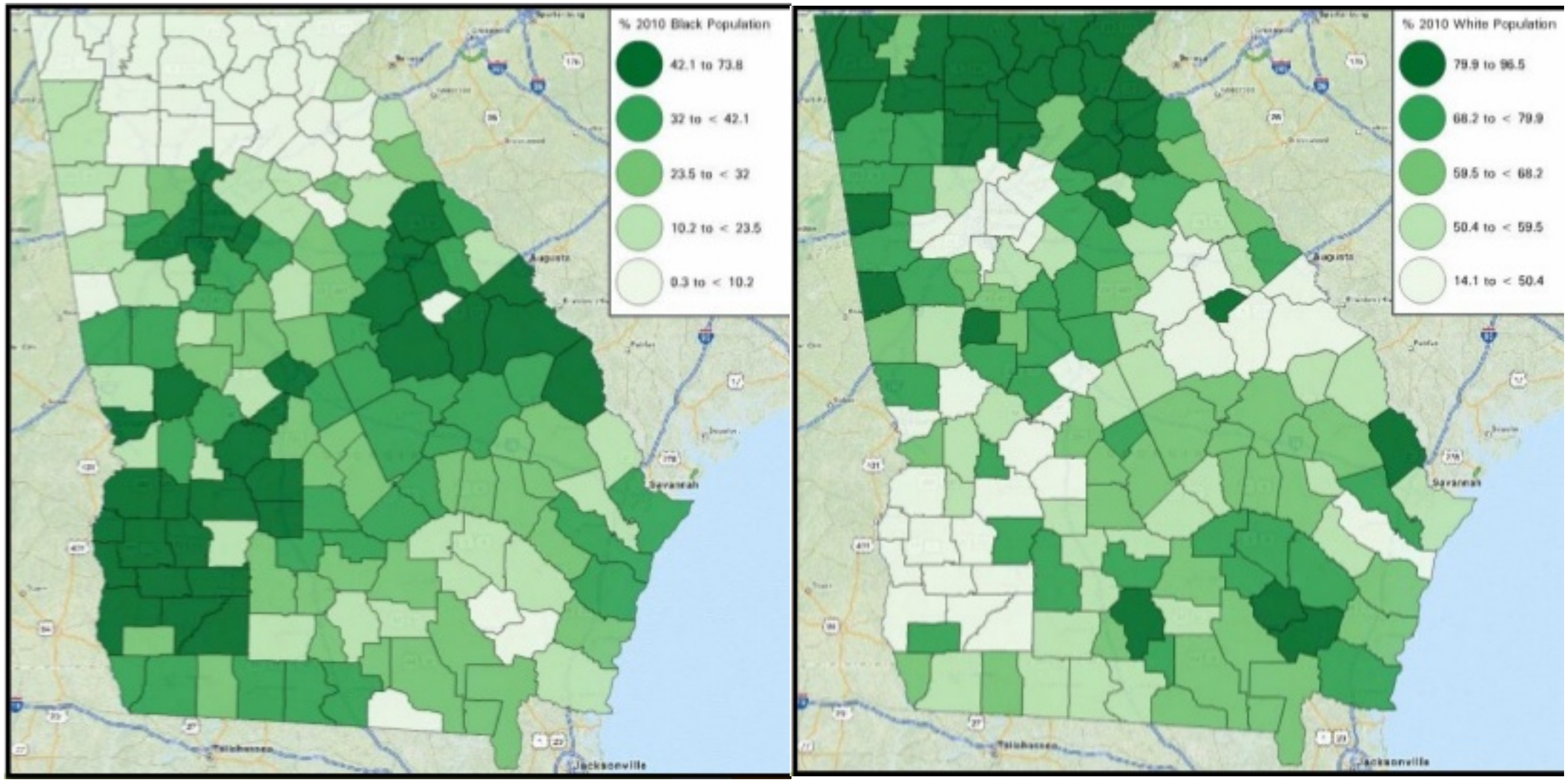
Race in Georgia circa 1900



W. E. B. Du Bois's Data Portraits: Visualizing Black America
By Whitney Battle-Baptiste, Britt Rusert

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Race in Georgia circa 2010



% Black, 2010

% White, 2010

Race Distribution in Georgia

Source: Atlanta Regional Commission

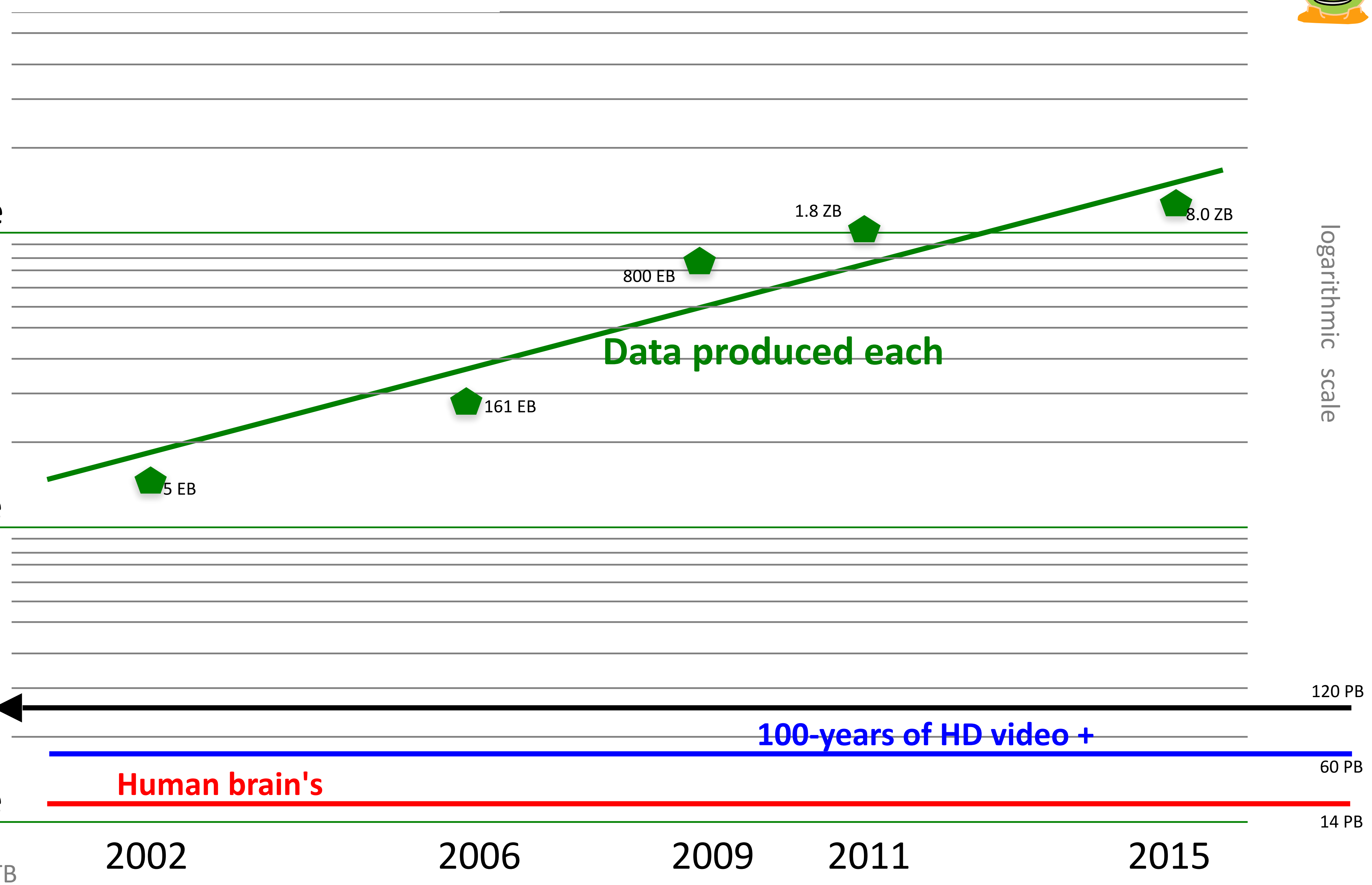
Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

What data science is

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Data, data everywhere...

There's certainly a lot of it!



IBM builds 120 petabyte cluster out of 200,000 hard drives

By Sebastian Anthony on August 26, 2011 at 6:18 am | 16 Comments

Smashing all known records by a multiple of 10, IBM Research Almaden, California, has developed hardware and software technologies that will allow it to strap together 200,000 hard drives to create a single storage cluster of 120 petabytes — or 120 million gigabytes. The drive collective, when it is complete, is expected to store one trillion files — or to put it in Apple terms, two billion hours of MP3 music.

Share This Article

1 Petabyte == 1000 TB
1 TB = 1000 GB

Reference

(2015) 8 ZB: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

(2011) 1.8 ZB: http://www.emc.com/leadership/programs/digital-universe.htm

(2009) 800 EB: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf

(2006) 161 EB: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

(2002) 5 EB: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm

(life in video) 60 PB: in 4320p resolution, extrapolated from 16MB for 1:21 of 640x480 video (w/ sound) – almost certainly a gross overestimate, as sleep can be compressed significantly!

(brain) 14 PB: http://www.quora.com/Neuroscience-1/How-much-data-can-the-human-brain-store

Data Science in Context

How much data are we talking about?

- This laptop stores 500GB…
- Your brain stores 28,000 laptops (or 14 Petabytes).
- 100 years of HD video is equivalent to 120,000 laptops (or 60 Petabytes).

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

A Data Science Definition

- The data science ecosystem is dedicated to the **systematic collection, management, analysis, visualization, explanation and preservation** of both structured and unstructured data. Through scientific methods and processes, the field of data science intends to extract impactful knowledge and insights to better the human condition.

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

The Data Science "Superhighway"

Acquisition & Cleanup	Storage & Management	Analysis	Visualization	Storytelling
Raw data is messy. Locating good data is challenging.	Housing and retrieval data is not trivial.	Input: data Output: information	Representing analyses outcomes via plots becomes a skill.	Who is your audience? What do you intend to communicate?
Where is it? How much quantity? Do you need to reformat?	MS Excel, MySQL Workbench, Oracle, NoSQL, Cassandra, MongoDB, etc.	Python, R, SAS SPSS, MS Excel	Python, R, Tableau, SPSS, MS Excel	text, video, audio, and/or images narrative
		Clustering, Classification, Regression, Statistics, etc	ggplot, scatter, bar, histogram, pie, box, line	GitHub, The Binder Project, MS Word

http://berkeleysciencereview.com/article/first-rule-data-science/

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

The misconceptions of data science

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

The misconceptions of data science

1. Access to more data translates to higher accuracy

2. Data science and business intelligence (or data analytics, or machine learning, or artificial intelligence) are the same

3. You must have access to a lot of data

4. Qualifications trump talent and experience

5. Data scientists know how to code

https://insidebigdata.com/2017/12/28/5-misconceptions-data-science/

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

The misconceptions of data science

6. Machine Learning is not a branch of Data science. Machine Learning originated from Artificial Intelligence. Data science is only using ML as a tool. The reason is that it produces amazing and autonomous results for specific tasks

7. It's not the salvation of companies that never measured anything and now want to get insights from their data. "Garbage in, garbage out" Data science will be as good as the data generated on the following years.

8. Just present data using some Excel charts without any insight about the data

9. …

10. …

What is data science and what is it not?

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

The misconceptions of data science

- Data science is **BROAD**
- No one person be proficient at all data-driven skills
  - What are the expectations of a “data scientist” in practice?
  - Most organizations now have a data division or data science teams that spans the data science superhighway

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Hands-On Jupyter Notebook

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Getting Started

- **Project Jupyter** supplies the open-source software, open-standards, and services for interactive computing across dozens of programming languages. Hopefully you all have installed ***Jupyter using Anaconda*.** If not, partner with a sister

  - It's a must-have on your personal laptop. Data science learning/training environments use Jupyter. You'll need ~600MB available to download, then another ~500MB to install & run.

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.

Now, let's get into Jupyter…

Poetic WordClouds

Copyright 2019. Dataedx Group, LLC. All Rights Reserved.