Principles and Practices of Data Science

Lecture 1

Melvin Ayala

Tentative Course Contents and Activities

Activity	Contents
Lecture 1	**Introduction to Data Science:** Objectives of the course, relationship between data science and artificial intelligence, characteristics of data, classes of data, concept of dark data, evolution of data science.
Lecture 2	**Data Science Domains, Categories, and Roles:** Data science definitions, data science team.
Lecture 3	**Data Science Analytics and Methodology:** Data science methodology, data analytics and lifecycle, data analytics methodologies, data collection, exploration, preparation and cleaning, data representation.
Lecture 4	**Data Collection, Requirements, and Visualization:** Data collection and requirements, exploratory data analysis, descriptive statistics, data transformation, data visualization.
Lab 1/IBM1	**IBM Skills Academy - Explore and Understand Data:** Obtain an IBM cloud account, Watson Studio, explore and visualize data.
Lecture 5	**Vector Calculus and Optimization:** Calculus, partial differentiation and gradients, gradients of matrices, convex optimization.
Lecture 6	**Geometry:** Norms, inner products, distances, angles, orthogonality, rotations, similarity.
Lecture 7	**Bayesian Decision Theory:** Naive Bayes Classifiers, Bayes theorem.
Lecture 8	**Probability, Information, and Uncertainty:** Concepts of probability, distributions, probabilistic systems, information theory.
Lecture 9	**Linear Algebra:** Systems of linear equations, matrix computation, vector space, linear models.
Lecture 10	**Sampling and Hypothesis Testing:** Sample distribution, central limit theorem, statistical tests (t-tests, Chi-squared test).
Lecture 11	**Clustering Techniques:** Similarity distances, k-means, hierarchical clustering.
Lab 2/IBM2	**IBM Skills Academy - Explore Insurance Claims Data:** Cleansing data, run first job, prepare and transform data, hypothesis 1 (loss claim after expired policy), hypothesis 2 (loss claim after expired license), hypothesis 3 (excessive claim amount) (IBM-based labs are subject to change and are to be run online by the students).
Lecture 12	**Analysis of Variance:** Foundations and examples.
Lecture 13	**Fundamentals of Statistics and Regression:** Fundamentals of statistics, linear and non-linear regression analysis.
Lecture 14	**Principal Component Analysis:** Foundations, dimensionality reduction, examples.
Lecture 15	**Feature Engineering:** feature spaces, feature engineering and selection, logistic function, kernels.
Lecture 16	**Fisher's Linear Discriminant:** Foundations, linear separability.
Lab 3	**Introduction to Python:** Python syntax and Jupyter Notebooks.
Lab 4	**Introduction to PyCharm:** Installation, run code and debugging.
Lecture 17	**Fundamentals of Support Vector Machines:** Linear separability revisited, margin and support vectors, primal and dual problem, bias.
Lecture 18	**Fundamentals of Artificial Neural Networks, Part 1:** History, McCullogh/Pitts model, Hebbian network and learning rule, Rosenblatt's perceptron.

no class in Spring break week → that will take three encounters out
class materials will be posted in blackboard after class

Activity	Contents
Lecture 19	**Fundamentals of Artificial Neural Networks, Part 2:** Supervised vs. unsupervised learning, self-organizing maps, stochastic gradient descent method, activation functions, simulation of logical operators, loss function, ADALINE/MADALINE networks (this lecture might be covered in two encounters).
Lecture 20	**Fundamentals of Artificial Neural Networks, Part 3:** Multilayer networks, backpropagation algorithm (this lecture might be covered in two encounters).
Lecture 21	**Fundamentals of Artificial Neural Networks, Part 4:** Recurrent neural networks for time series predictions, limitations, LSTM, GRU (this lecture might be covered in two encounters).
Lecture 22	**Image Representation and Processing:** Image representations, histogram techniques, basic transformations.
Lecture 23	**Convolutional neural networks for image processing/classification:** convolution, padding, strides, architecture, training (this lecture might be covered in two to three encounters).
Lecture 24	**Fundamentals of Decision Trees:** Structure, types, splitting and pruning, algorithms.
Lecture 25	**Natural Language Processing, Part I:** Introduction, text handling techniques, topic modeling, sequence models
Lecture 26	**Natural Language Processing, Part II:** Word Embeddings (Word2Vec), encoder-decoders, autoencoders
Lab 5	**Guide to Data Preprocessing - Steps and Coding Examples:** How to deal with missing data, plot histograms, display box/violin/scatter plots, calculate and display correlation, handle imbalanced data.
Lecture 27	**Performance Evaluation of Classifiers:** Confusion matrix, Receiver Operating Characteristics Analysis (ROC), metrics, comparison of classifiers.
Lab 6	**Coding Practical Regression Examples with Python:** Train a regression model using multivariate data.
Lab 7	**Coding Practical Classification Examples with Python:** Train a classification model to predict diabetes.
Lab 8/IBM3	**IBM Skills Academy - Discovering Fraudulent Claims with Data Transformation:** Data refinery, visualization output (IBM-based labs are subject to change and are to be run online by the students).
Lab 9/IBM4	**IBM Skills Academy - Fraud Diagnostic Analytics:** Fraud diagnostics analytics, data visualization, data presentation, create the Jupyter notebook (IBM-based labs are subject to change and are to be run online by the students).
Lab 10/IBM5	**IBM Skills Academy - Using AutoAI:** Predicting fraud with AutoAI, data model augmentation (IBM-based labs are subject to change and are to be run online by the students).

Lecture 1: Introduction to Data Science

Sections:

1. Introduction
2. Objectives of this Course
3. Characteristics of Data
4. Classes of Data
5. Concept of Dark Data
6. Evolution of Data Science

1.1. Introduction

About this Course

Why study Data Science?

Definition of Data Science:

- a multidisciplinary approach to extracting info from volumes of data

volumes of data: large + increasing

What do we do with Data Science?

- prepare data for analysis: collection, storage, formatting
- pre-process data: filtering, cleaning, simplification, removal of unwanted data
- process data: apply different algorithms according to the subject (correlation, classification, prediction, AI, …). Application of software tools.
- Use existing tools to get expedite results
- Use existing programming environments to personalize the research
- Goals = find patterns, classify, predict etc. to support business-decision making
- But we need to test/validate the results (using scientific methods)
- In the process: use visualization techniques

Data Analyst vs. Data Scientist

Data Analyst:

- gathers data to identify patterns
- performs statistical analysis
- uses pre-existing tools like:
    - SQL to query the data, Excel
    - Data mining or integration methods
    - Programming languages (at a basic level) to perform rudimentary operations
    - Visualization tools to represent the data

Data Scientist:

- More involved with the design of the data modeling process.
- Creates algorithms
- Applies mathematics, statistics and scientific methods
- Uses wide range of tools and techniques
- programming languages (python, C#, etc.)
- Designs application to automate data processing and calculations
- Creates machine learning models using AI

Relationship between Data Science and Artificial Intelligence (AI)

AI performs better when we train our systems with more data.
Massiveness of data comes with a challenge (to be overcome by Data Science)
The quality of the data processing affects the quality of the AI output (prediction/classification/control action)

Data Analysts and Data Scientists are currently in high demand.

1.2. Objectives of this Course

Objectives of this Course

1. Understand the evolution and relevance of data science in the world today.
2. Understand the scientific method for science projects and the data science team's key role.
3. Explore data engineering and data modeling practices using machine learning.
4. Understand the basics of regression and classification models.
5. Understand the fundamentals of machine learning, including model training and evaluation
6. Learn the Python programming language and how to use programming platforms (Jupyter notebooks and PyCharm)
7. Learn how to design basic machine learning models using Python.

How completing this course could benefit you?

Rapid growth of artificial intelligence (AI) → demand for data scientists (higher than availability)

Applications of AI are found in many fields of Medicine.

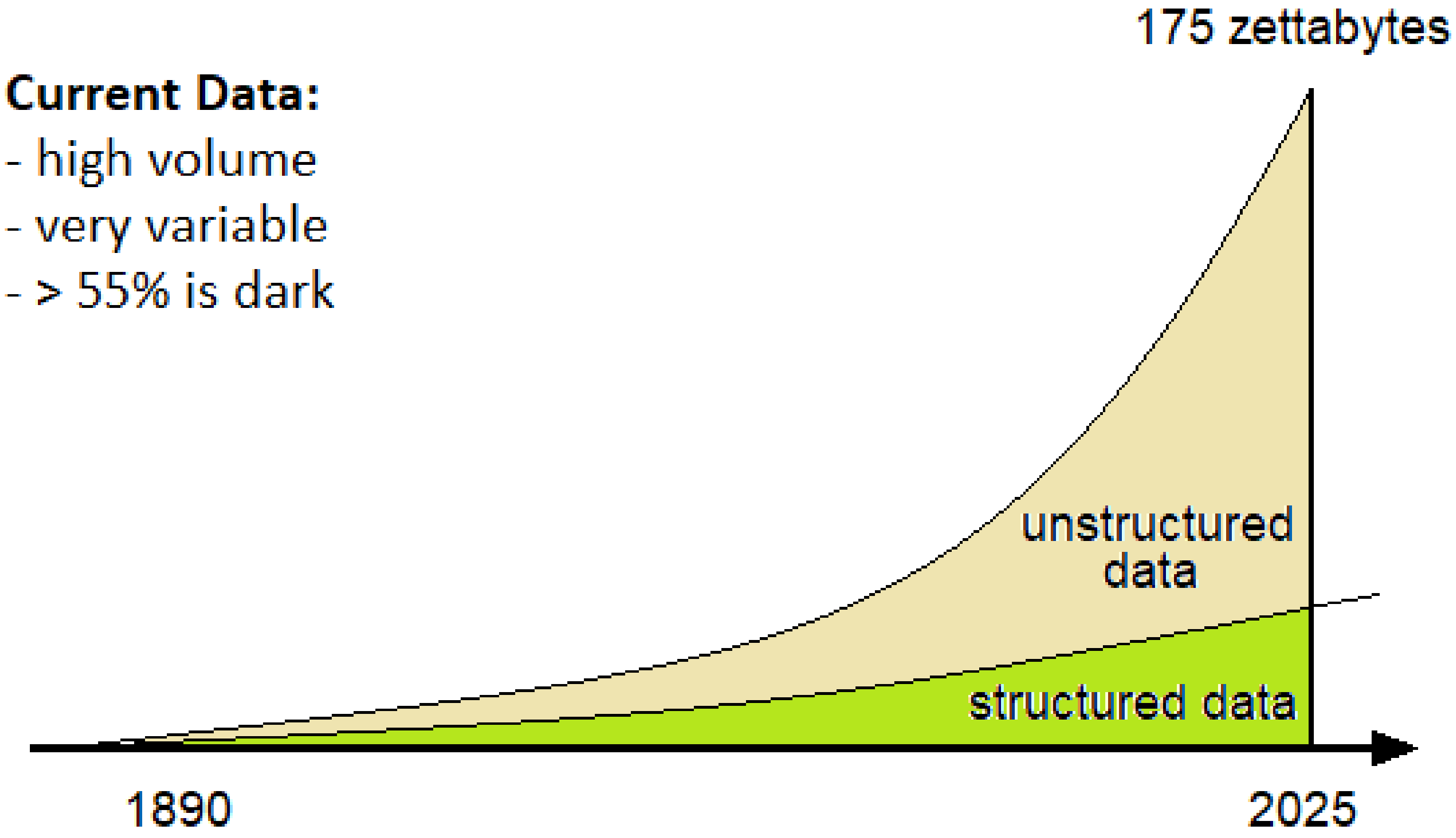
- Extraction of information from medical images.
- Extraction of information from medical reports (radiology, pathology, …)
- Medical professionals can benefit from it

1.3. Characteristics of Data

Data In the World Today

Data Grow and Projection

Current Data:
- high volume
- very variable
- > 55% is dark



Characteristics of Data

The 5 Vs

Volume
Large amounts of data generated every second

Velocity
Speed at which new data is generated and moves around

Variety
Different types of data we can use

Value
Ability to get value out of data

Veracity
Trustworthiness, unreliability of the data

Cognitive Computing:

- subfield of artificial intelligence
- simulates human thought processes in machines using self-learning algorithms
- applies through data mining, pattern recognition, and natural language processing.
- mimics human thought processes (→ help people make better and easier decisions)

The Price of Not Knowing

- what is the price of not curing cancer?
- what is the price of not discovering alternative energy sources?

Cognitive Computing (cont'd)

Impact of Cognitive Computing:

- **In oil & gas industry:**
  - supply chain has more than 80,000 sensors in place
  - one single reservoir we can produce immense amounts of data per day
  - preventing drilling in the wrong place
  - help is with the flow of oil or petroleum through all the pipelines
- **In retail industry:**
  - tweets and Facebook posts made by billions of people with cellphones
  - data can be mined to better understand customer needs and demands
- **In healthcare**
  - areas of electronic medical records,
  - patient population
  - medical imaging
  - one person generates a huge amount of data

Cognitive Computing (cont'd)

Other areas impacted by Cognitive Computing:

- **Smart digital meters:**
  - Most of the data is dark, but can help in better understanding demands
- **Transportation:**
  - In the next decades, most transportation systems will be interconnected.

Goal of Cognitive Computing:

- Changing the world
- Changing entire industries
- Getting insights in data the way we have never been able to do before.

1.4. Classes of Data

Categories of Data:

1. Structured data: data bases, formatted files
2. Semi-structured data: json files, xml, emails, web files (html)
3. Unstructured data: audio, video, images, documents (MS Word, text)

Example of formatted data: Medical data extracted from radiology reports

#	Age	Gender	Medication	Days administered	Tumor size
1	…	…	…	…	…
2	…	…	…	…	…
3	…	…	…	…	…
…	…	…	…	…	…

Notes:

- multidimensional data points 1,2, 3, …
- features with different categories
  - age: incremental
  - gender: categorical data
  - etc.

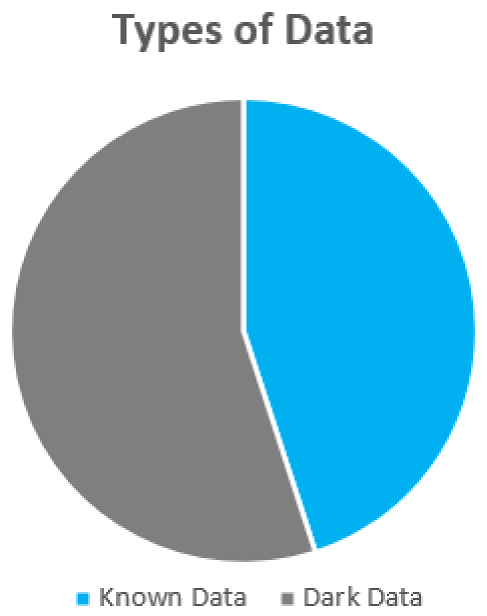
1.5. Concept of Dark Data

Dark data:

- all of the unused, unknown and untapped data across an organization
- generated as a result of users' daily interactions with devices and systems
- may be considered too old to provide value, incomplete or redundant,
- often limited by a format that can't be accessed with available tools
- all too often, they don't even know it exists.

Importance of Dark Data:

- may be one of an organization's biggest untapped resources.
- data is increasingly a major organizational asset
- competitive organizations will need to tap into its full value.
- stringent data regulations may necessitate complete management of an organization's data.

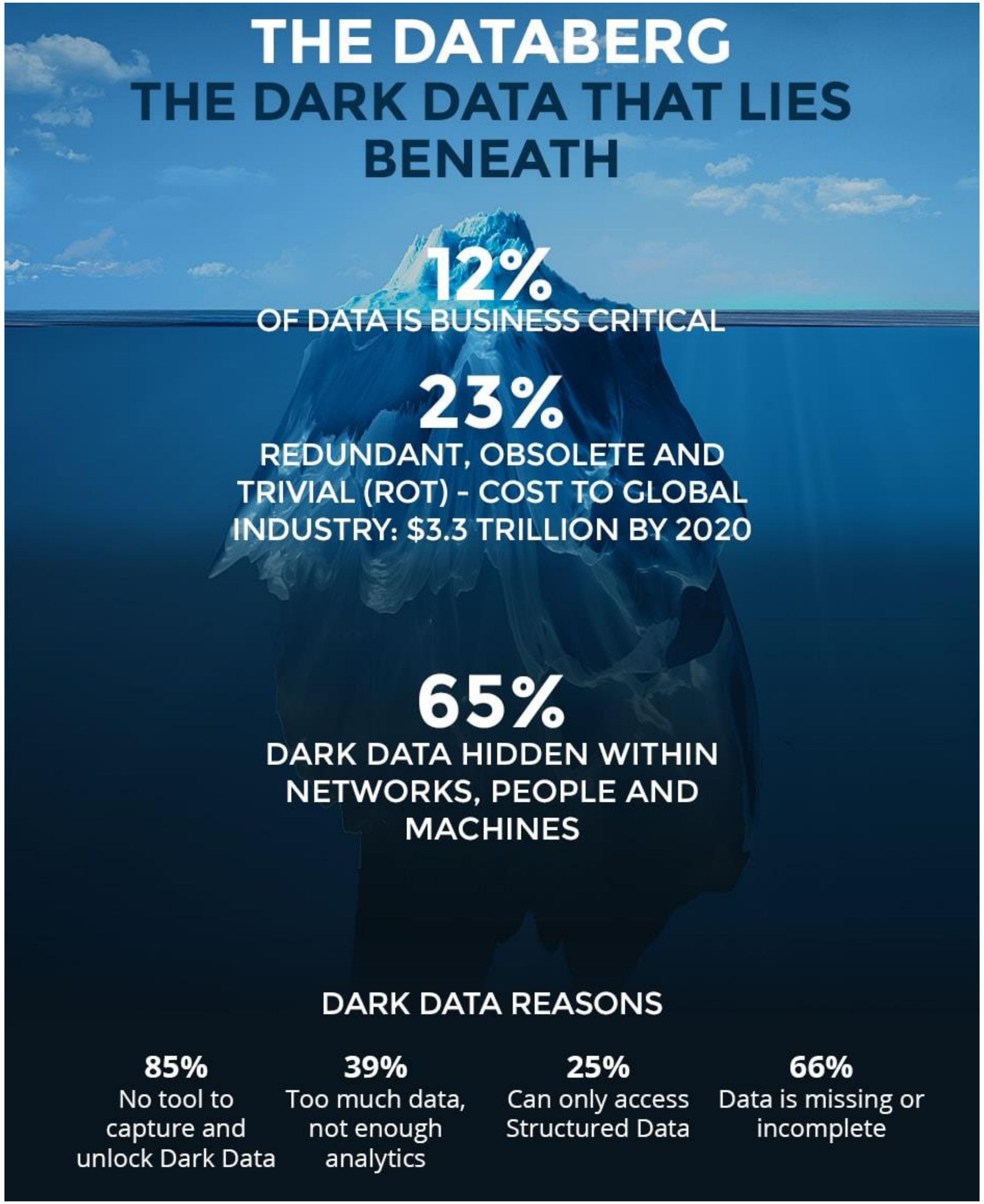


Globally:
More than 50% of an organization's data is considered "dark" (1).

Survey Estimate:	Reported by:
"More than 75% of their data was dark"	33% of respondents
"Less than a quarter of their data was dark"	11% of respondents
"At least half their data was dark"	44% of respondents from China
	65% of respondents from France and Japan

(global average: 60%)

(1) According to a recent State of Dark Data report by TRUE Global Research.


THE DATABERG
THE DARK DATA THAT LIES BENEATH
12%
OF DATA IS BUSINESS CRITICAL
23%
REDUNDANT, OBSOLETE AND TRIVIAL (ROT) - COST TO GLOBAL INDUSTRY: $3.3 TRILLION BY 2020
65%
DARK DATA HIDDEN WITHIN NETWORKS, PEOPLE AND MACHINES
DARK DATA REASONS
85%
No tool to capture and unlock Dark Data
39%
Too much data, not enough analytics
25%
Can only access Structured Data
66%
Data is missing or incomplete

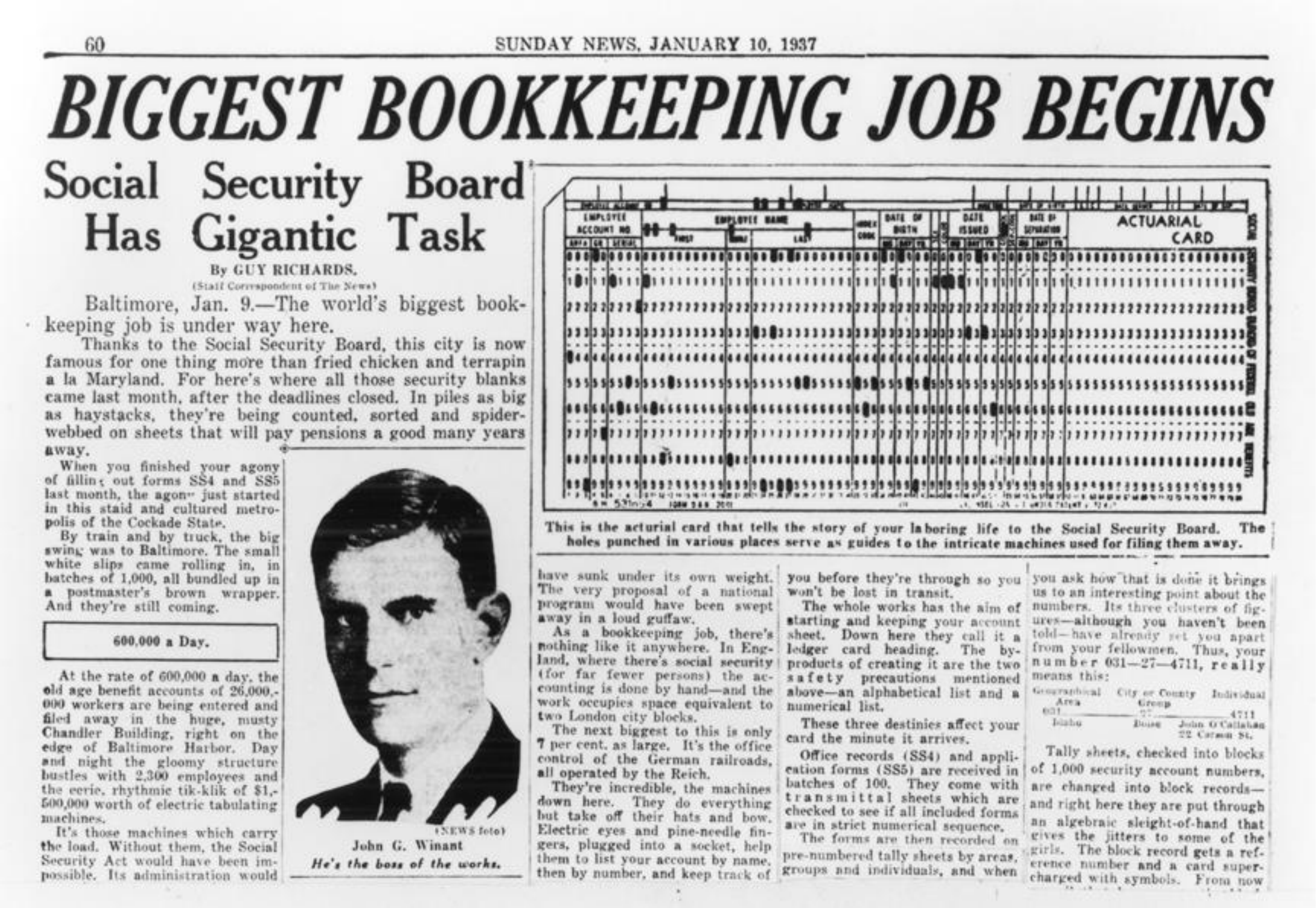
1.6. Evolution of Data Science

Brief History of Data Science

Statistical Sciences	1890 **US Census** From 10 years to just 2.5 years	1935 **Social Security Act** Payroll, reports, statistics milestone with 26M SSNs	1952-1956 **Eisenhower Election** Machine predicts voting results	1962 **Future Data Analysis** John Tukey	1974 **Neural Networks** NPIS Conference	1999 **Data Mining** Large Data Sets Jacob Zahavi	2001 **Data Science Plan** Cleveland	2005 **Big Data** Web 2.0 User-driven data trend	2006 **Deep Learning** Algorithms G. Hinton	Today **Data Science** Global Mainstream	
Computer Applications	1890 **Tabulating Machines** H. Hollerith	1936 **Universal Machine** Alan Turing	1944 **Mark I** 1st computer to perform long automatic computations	1954 **Large Scale Adoption** Computers in governemtn and corporations	1970 **Relational Databases and SQL** Edgar Codd	1981 **Personal Computers** Distributed computing	1996 **DeepBlue** Chess: Kasparov, 200 M moves/second	2007-Today **Cloud Computing** Global IT Infrastructure	2011-Today **Machine Learning** Watson Jeopardy		
Digital Data	80 Bytes Punch Cards 1890	62.4 KB Magnetic Drum 1932-1950	2 MB Magnetic Tape Drive 1948	5 MB Hard Disk RAMAC 1956	400 MB Mainframe System 360 1964-1974	3 1/2" Floppy Affordable Personal Data 1985	Internet Global Data Sharing 2000-Today	Social Media Democratizing Digital Data 2005-Today			
	1890	1930	1940	1945	1950	1960	1970	1980	1990	2000	2010	2020

From 1890 to 1936, punch cards helped:

- performing repetitive computing operations (Social security, payroll, etc.)
- supporting statistical analysis



60 SUNDAY NEWS, JANUARY 10, 1937

BIGGEST BOOKKEEPING JOB BEGINS

Social Security Board Has Gigantic Task

By GUY RICHARDS.
(Staff Correspondent of The News)

Baltimore, Jan. 9.—The world's biggest bookkeeping job is under way here.

Thanks to the Social Security Board, this city is now famous for one thing more than fried chicken and terrapin a la Maryland. For here's where all those security blanks came last month, after the deadlines closed. In piles as big as haystacks, they're being counted, sorted and spider-webbed on sheets that will pay pensions a good many years away.

When you finished your agony of fillin; out forms SS4 and SS5 last month, the agon" just started in this staid and cultured metropolis of the Cockade State.

By train and by truck, the big swing was to Baltimore. The small white slips came rolling in, in batches of 1,000, all bundled up in a postmaster's brown wrapper. And they're still coming.

600,000 a Day.

At the rate of 600,000 a day, the old age benefit accounts of 26,000,000 workers are being entered and filed away in the huge, musty Chandler Building, right on the edge of Baltimore Harbor. Day and night the gloomy structure bustles with 2,300 employees and the eerie, rhythmic tik-klik of $1,500,000 worth of electric tabulating machines.

It's those machines which carry the load. Without them, the Social Security Act would have been impossible. Its administration would have sunk under its own weight. The very proposal of a national program would have been swept away in a loud guffaw.

As a bookkeeping job, there's nothing like it anywhere. In England, where there's social security (for far fewer persons) the accounting is done by hand—and the work occupies space equivalent to two London city blocks.

The next biggest to this is only 7 per cent. as large. It's the office control of the German railroads, all operated by the Reich.

They're incredible, the machines down here. They do everything but take off their hats and bow. Electric eyes and pine-needle fingers, plugged into a socket, help them to list your account by name, then by number, and keep track of you before they're through so you won't be lost in transit.

The whole works has the aim of starting and keeping your account sheet. Down here they call it a ledger card heading. The by-products of creating it are the two safety precautions mentioned above—an alphabetical list and a numerical list.

These three destinies affect your card the minute it arrives.

Office records (SS4) and application forms (SS5) are received in batches of 100. They come with transmittal sheets which are checked to see if all included forms are in strict numerical sequence.

The forms are then recorded on pre-numbered tally sheets by areas, groups and individuals, and when you ask how that is done it brings us to an interesting point about the numbers. Its three clusters of figures—although you haven't been told—have already set you apart from your fellowmen. Thus, your number 031—27—4711, really means this:

Geographical Area	City or County Group	Individual
031	27	4711
Idaho	Boise	John O'Callahan 22 Carmen St.

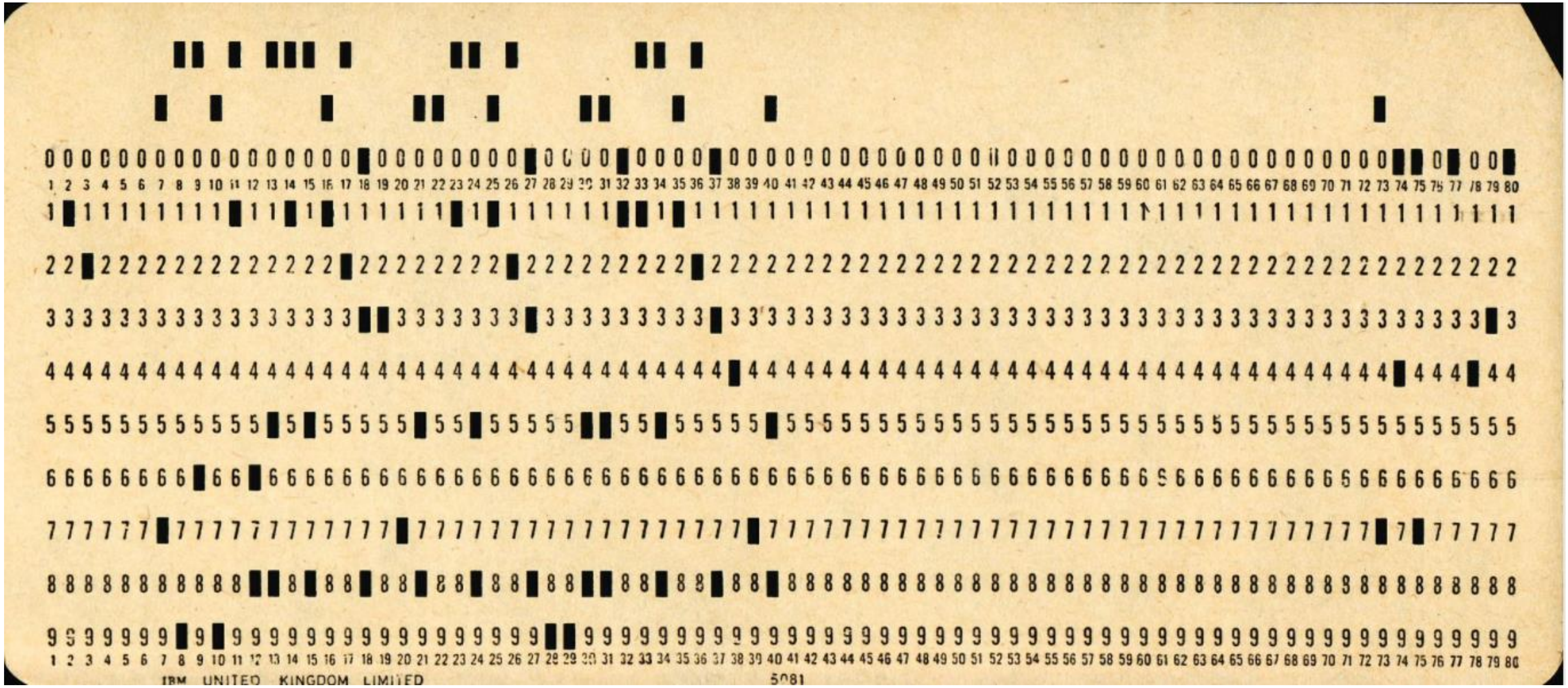
Tally sheets, checked into blocks of 1,000 security account numbers, are changed into block records—and right here they are put through an algebraic sleight-of-hand that gives the jitters to some of the girls. The block record gets a reference number and a card supercharged with symbols. From now

(NEWS foto)
John G. Winant
He's the boss of the works.

ACTUARIAL CARD

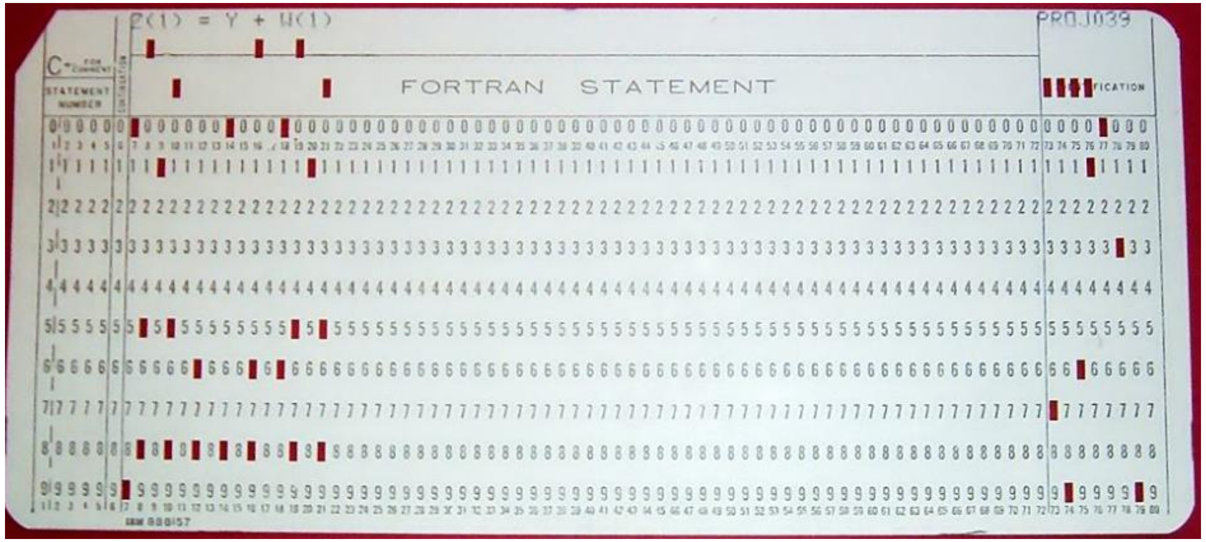
This is the acturial card that tells the story of your laboring life to the Social Security Board. The holes punched in various places serve as guides to the intricate machines used for filing them away.

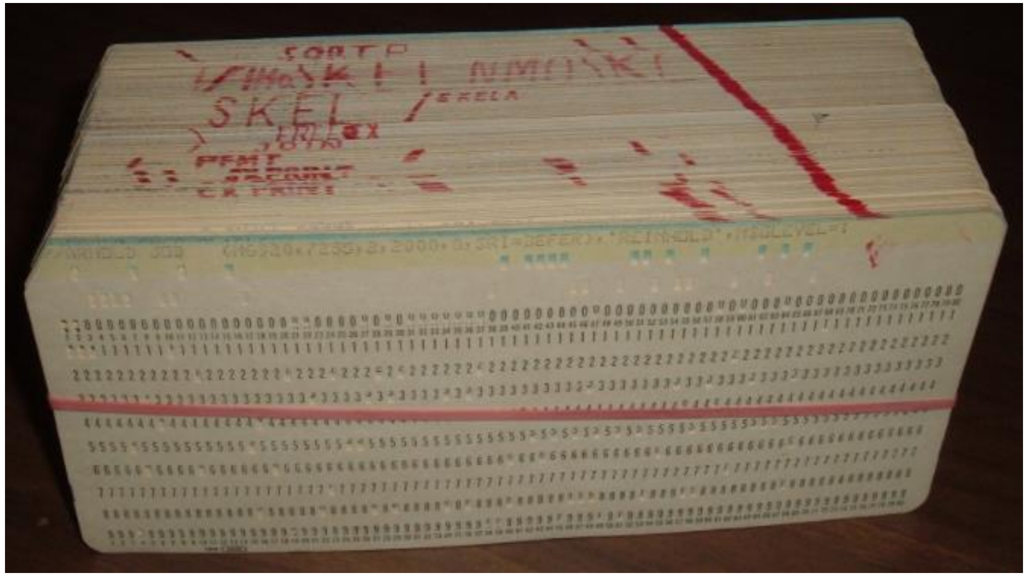
A Punched Card (Hollerith)



A 12-row/80-column IBM punched card from the mid-twentieth century
Also used in some countries until 1990s.

Punched Cards

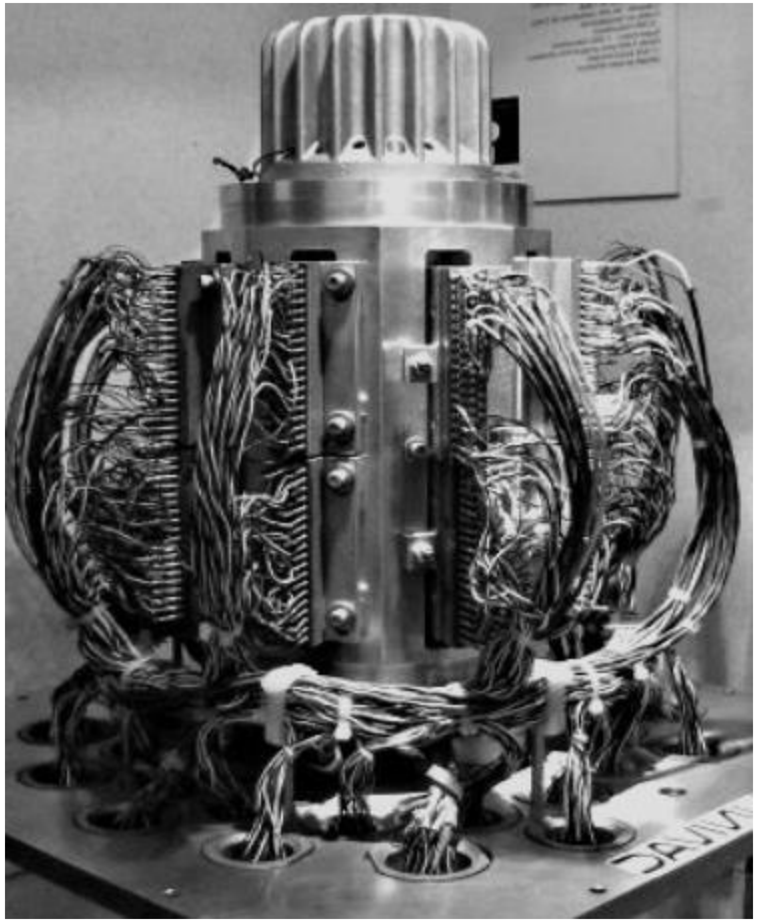




Punch cards from the 1950s SAGE air defense system

62,500 punched cards (around 5 MB of data).

Magnetic Drums



- drum memory was a magnetic data storage device
- about 62 kilobytes

End of Lecture