Big Data Cloud Platform in Industry For Data Scientist

Ming Li

Who am I?

amazon

OUTLINE

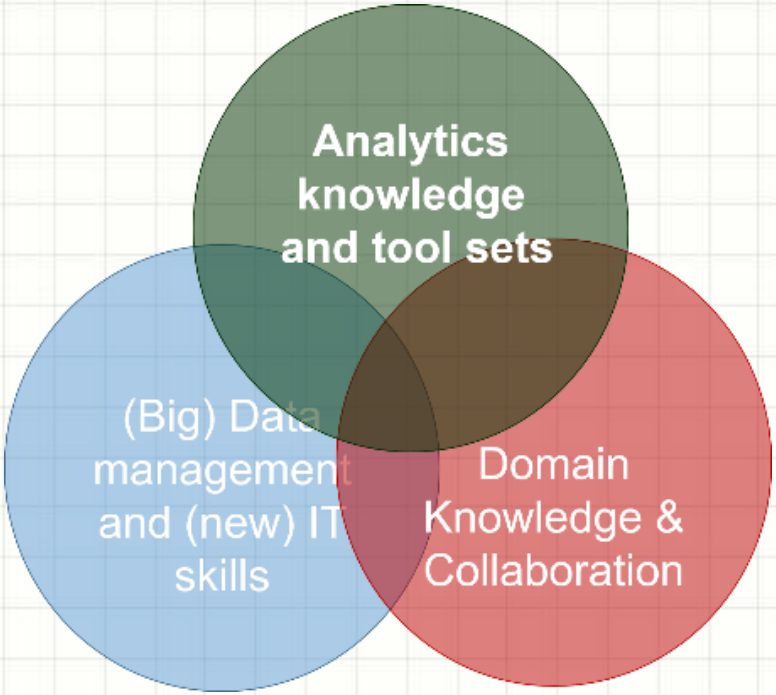
- ✓Challenge of Big Data Modeling
- ✓Cloud Environment
- ✓Data warehouse and Database
- ✓Summary

Challenge of Big Data Modeling & Opportunity for Statistician

Different Practice Between Statistician and Data Scientist

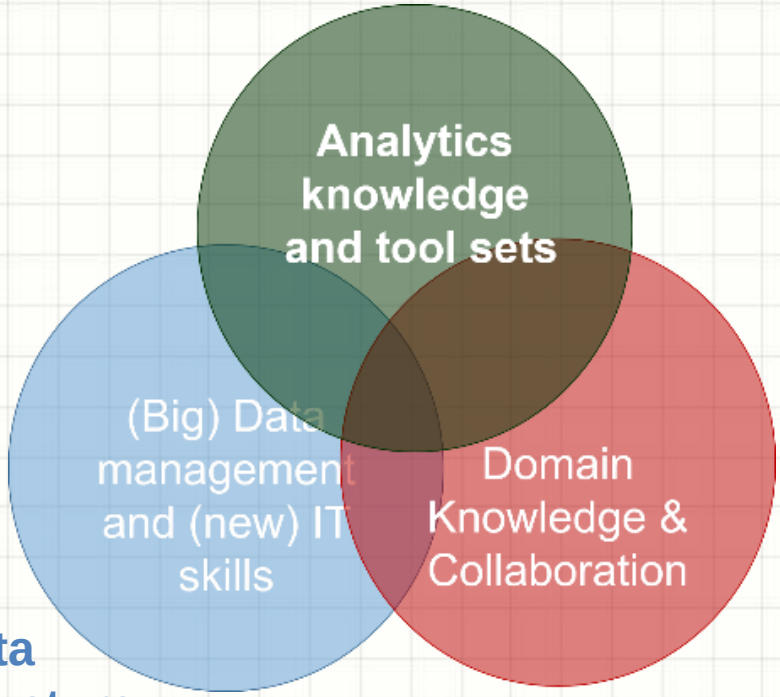
Statistician	Data Scientist
Data is given (i.e. in text files)	Need to get data (i.e. for database)
Focus on modeling	Focus on results
Bring data to model	Bring model to data
Further away from production system	Embedded in production system

Three Pillars of Knowledge for Success in Data Science



Three Pillars of Knowledge for Success in Data Science

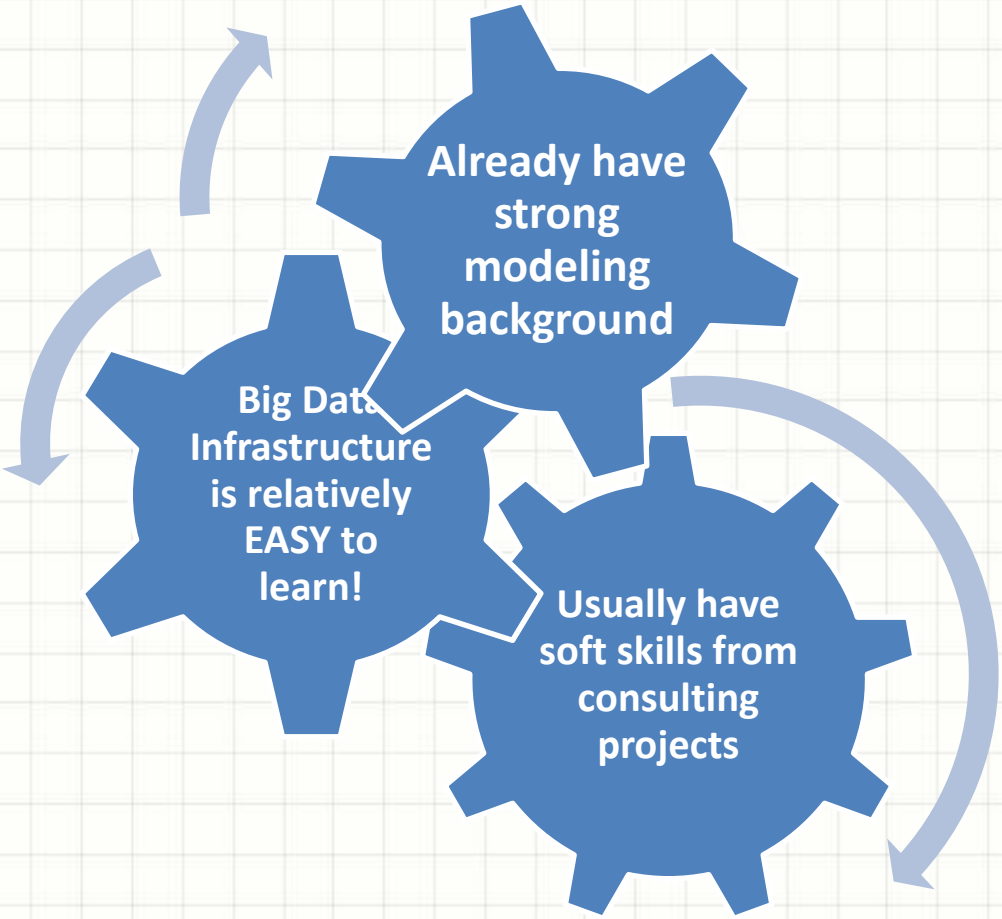
- **Understand and prepare data**
- **Statistical methods and problem solving**
- **Machine learning and data mining experience**



- **Unstructured data**
- **Big data infrastructure**
- **Database and data retrieval**
- **Software: Hadoop, SQL, R, Python …**
- **"Cloud" Solution**

- **Teamwork**
- **Problem definition**
- **Communication skill**
- **Strategic planning & execution**

Opportunity for Statistician: not too different from what we do!



Data

Information

Knowledge

Insight

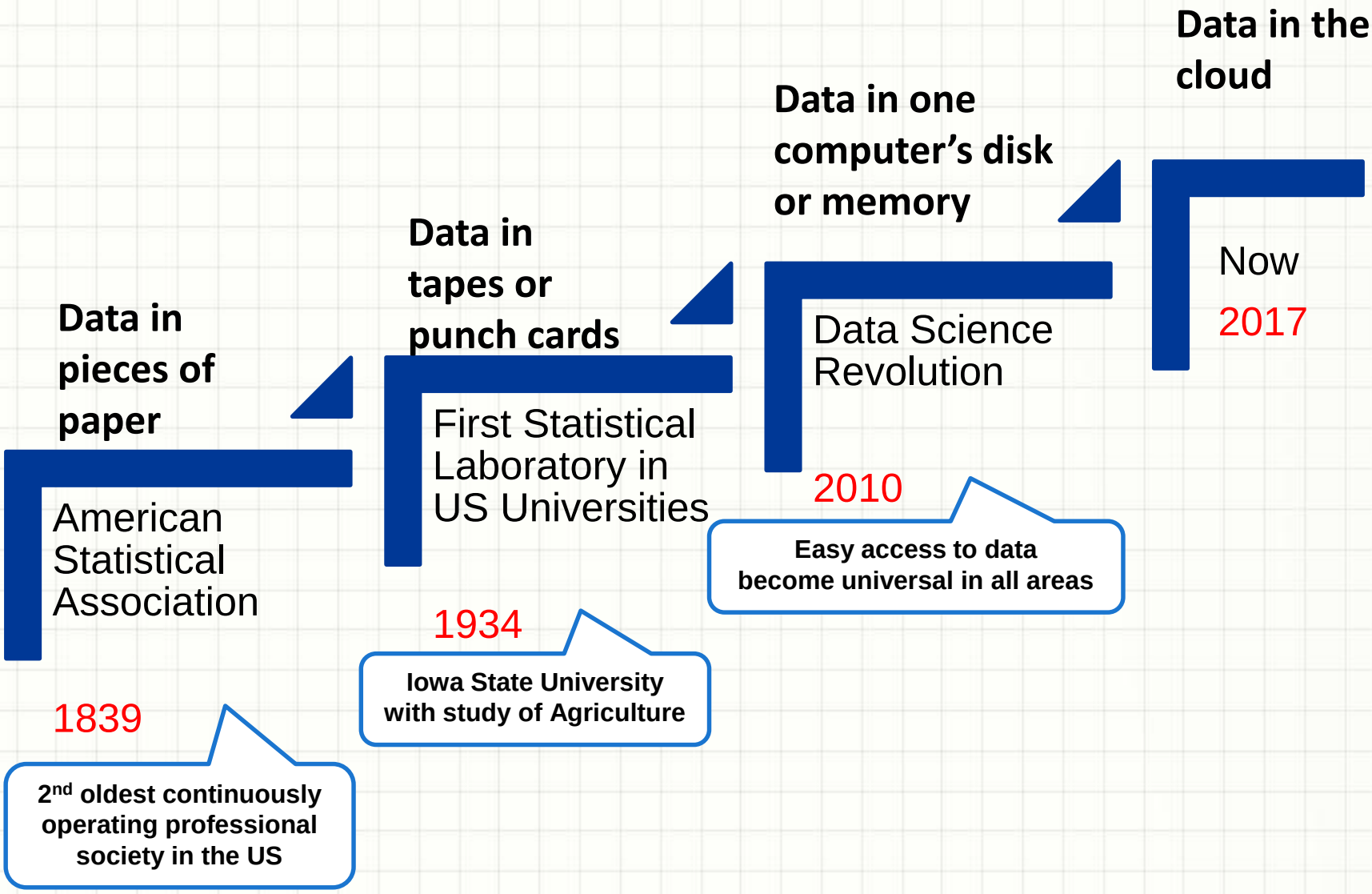
Automatic Decision & Action

Big Data Infrastructure, Integration, Automation & Execution

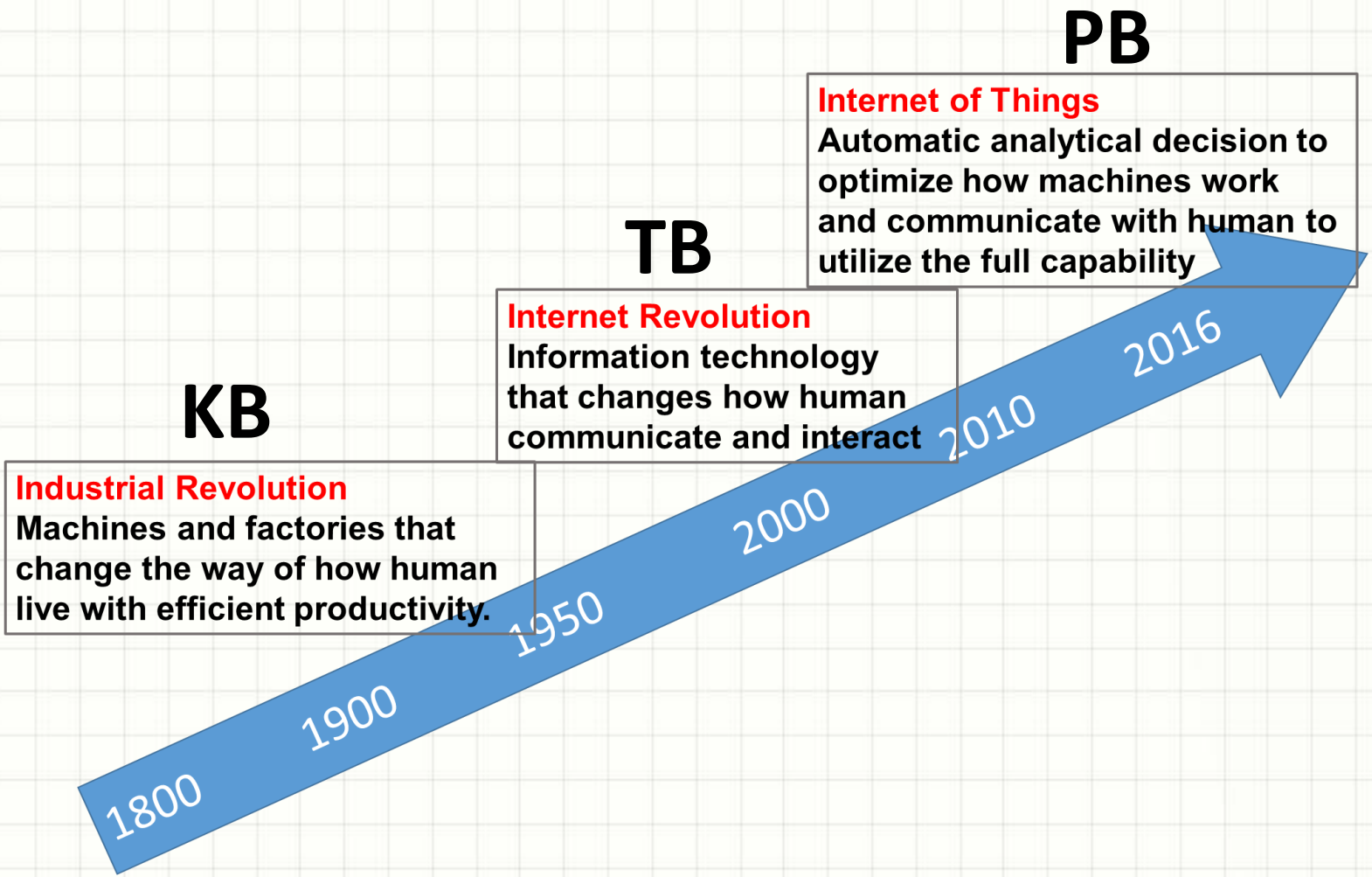
Partially solved by "Cloud" Platform

Cloud Environment

History of Statistician in the America



New Wave of Industrial Revolution



Advantages of Cloud Environments

- Scalability and maintenance becomes behind scenes and usually taken care of by the infrastructure provider
- Massive data set is easier to get for analysis
- Efficient model training on larger amount data become possible
- Model deployment to production environment is much easier
- Model refresh becomes relative automatic tasks

A few free cloud environments providers for learning purposes

- Databrick Community Edition
- Amazon AWS cloud environment
- Microsoft Azure cloud environment
- Google cloud platform

- Two short videos for how to use Databrick community edition which can run R/Python/SQL/Scala:
  - https://youtu.be/vx-3-htFvrg
  - https://youtu.be/C7uCNwoF9h0

Database Management (SQL)

One of the key elements that many Statistician lack, but really easy to learn!

Why Database Management System

- When the data is beyond computer's memory or hard disk, a database management system is need to management data.
- Even data can be fit into computer's hard disk, database management systems provides a much better way to manage data such as Extract /Transform/Load (i.e. ETL) and ensure data integrity
- SQL standards make it easy to deal with different database management system including the modern big data infrastructure (such as Hadoop/Hive)

Main Database Concepts

- Database -> table -> rows (records) & columns (fields) -> one data cell
- Mechanism for fast data retrieval: Key / Partition / Index
- Row-oriented vs. column-oriented databases
- Hard disk vs. in memory
- Distributed and parallel
- A lot of work is done behind sense, and users can focus on writing SQL

Main SQL functions

- Simple data retrieval using SELECT statement
- Combine data from multiple tables using JOIN and UNION
- Sort data using ORDER BY
- Aggregating using GROUP BY
- Subset selection using WHERE and HAVING
- Change data using UPDATE, INSERT, and DELETE
- A few videos that describe the basic SQL functions through my YouTube channel: https://www.youtube.com/channel/UCiQRHcIhhqeiUXKpJCeQnpg

Linux System Command

Another element that some (theoretical) Statistician lack, but really easy to start!

Why Linux Operation System

- Most production system and big data system requires Linux operation system knowledge
- Multiple user system that allow many user to use at the same time
- High availability that the system can run days or months without restarts
- It is easy to start and obtain the basic knowledge for every day usage
- It is indeed take time to be an expert in using it

SUMMARY

- Big data modeling is challenging for statistician in data retrieval, model scalability, and model implementation and integration
- Statistician do have the advantage of in-depth understanding of the models and can go beyond just calling functions from packages
- Cloud environment lowers the barrier for statistician to become successful data scientist

Questions?

MLI@alumni.iastate.edu

Thank You!