

Big Data Analytics

Kohei Arai
Emeritus Prof. of Saga University
Adjunct Prof. of Brawijaya University

1

Introduction to Big Data

- Why do we tend to click on ads on search pages and SNS?
- Or will there be advertisements for things I want?
- **Online shops use customer purchase data, member information, and website information.**
- We store various data such as click information within the website as big data.
- Using that big data, artificial intelligence can help customers or agents
- Decide which products are likely to be purchased and display them as recommended products to customers. This will lead to sales.

2

Just an example of Big Data

- Search history
 - looking at car
 - looked for test values
 - looked for a campsite
- As a result of the identification prediction, you will be directed to the desired site and click on it by Search Engines or agents



3

Toward Data-Driven Society

- A data-driven society is an attempt to create **new social value** by analyzing data from the real world (physical space) using digital technology.
- The amount of data is increasing explosively due to the evolution and spread of the Internet, mobile, IoT, and AI (becomes Big Data). **A data-driven society utilizes this amount of data to generate knowledge and information and create social value.**
- **A feature of a data-driven society is that the real world and cyberspace are closely linked through CPS (Cyber Physical System), and the knowledge created is put to use in the real world.**

4

Decision making

- Until now, decisions related to company management have relied on experience and intuition, but from now on we will be based on information backed by data and knowledge generated.
- **By utilizing data, the cycle of identifying strategies can be completed in a short lead time. The reasons for this as follows,**
 - •Diversification and complexity of **consumer behavior and needs**
 - •Diversification and complexity of **operations**
 - •**Market uncertainty** and rapid changes in the business environment
 - •Need for data utilization

5



6

Data-driven management

- The two main benefits of data-driven management are **"improving the productivity of existing businesses"** and **"creating new businesses,"** and being able to quickly make the necessary management decisions.
 - •Improve **operational efficiency through visualization**
 - •Improve **productivity through automation**
 - •Optimization using **prediction**
 - •Create **new value**

7

McKinsey report

- In the report, McKinsey defines the new data-driven enterprise of 2025 as having seven characteristics:
 - ① **Data embedded in every decision, interaction, and process**
 - ② **Process and distribute data in real time**
 - ③ **Flexible data stores enable integrated, ready-to-use data**
 - ④ **Data operation model that treats data like a product**
 - ⑤ **The role of the chief data officer (CDO) will expand to create value**
 - ⑥ **Data ecosystem membership becomes mainstream**
 - ⑦ **Prioritize and automate data management with privacy, security, and resilience in mind**

8

Internal and external data including social media

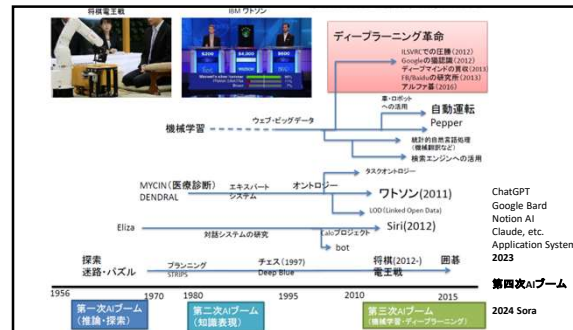
- Business analysis is the analysis of financial statements, and the use of the obtained information for management. Specifically, it refers to the process of objectively viewing the **business situation based on numbers**, understanding problems, and checking the effectiveness of measures.
- Financial statements include **balance sheets, income statements, and cash flow statements**. The specific content of each will be explained in detail later, but all contain important information for corporate management and are useful materials for diagnosing the current situation from various angles.
- Basically, analysis is performed using **quantitative numbers from these information sources**, but individual circumstances that are not reflected in the numbers can also have a significant impact on business results.

9

AI Fours Era

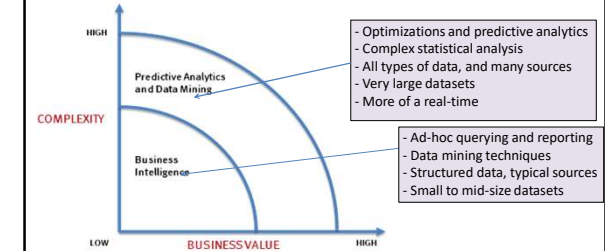
- 2022~: The era of large-scale models
- Introducing ChatGPT (3.0): **175 billion parameters** (number that refers to data input from outside) → By applying parameters to calculation formulas and learning them, it is now possible to derive **more than 40 trillion pieces of knowledge** → **This is 1000 times more knowledge than a person can gain in a lifetime.**
- The difference from previous AI is that it generates from judgment.
- **ChatGPT (4.0) expands to over 1 trillion parameters (6000 times)**
- Application using AI as DX → Collaboration of ChatGPT with cloud computing and API

10

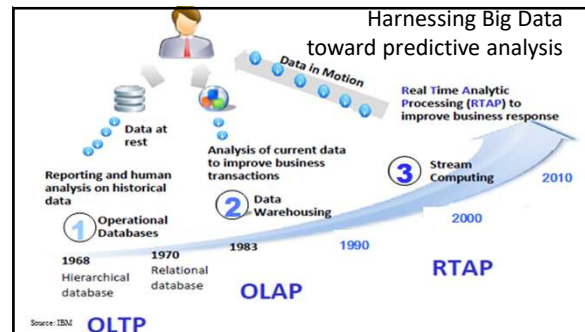


11

What's driving Big Data



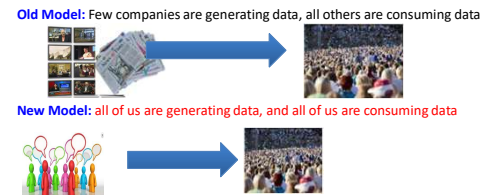
12



13

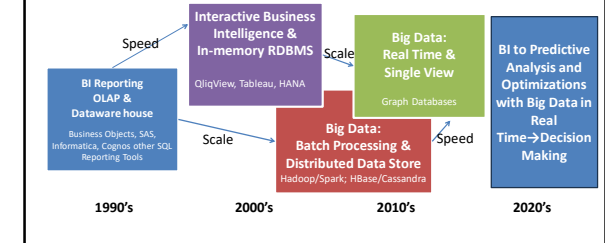
The Model Has Changed...

- **The Model of Generating/Consuming Data has Changed**



14

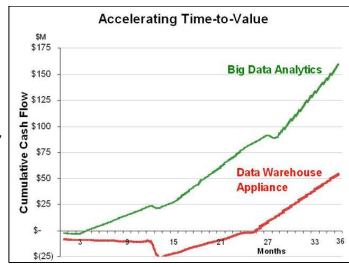
THE EVOLUTION OF BUSINESS INTELLIGENCE



15

Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



16

Required big data analysis tools

- **Data cleansing** is the process of removing or correcting inaccurate, incomplete, or invalid data from datasets to improve data quality.
- **Data cleaning** is the process of detecting and correcting errors and inaccurate values in datasets and improving data quality.
- **Company/competitor analysis** includes not only sales but also many aspects such as market share, customer satisfaction, comparison of product features, comparison of advertising effectiveness, etc.

17

Google Analytics, Heat-map, access

- Get real-time data from specified properties (such as sites and apps) and views (specific view IDs) using the **Google Analytics Reporting API**
- **Heat map analysis** is a method to visually understand patterns and correlations by expressing the values corresponding to each cell in a matrix using colors.
- **Access analytics** is the analysis of data about user access to a website or application, typically including user sessions, page views, user attributes, etc.

18

Page-view, keyword, EDA, financial

- **Page view analysis** is part of the analysis done to understand user behavior on digital platforms such as web pages.
- **Keyword analysis** is some of the techniques used to extract important words and phrases from text data
- **Exploratory Data Analysis (EDA)** code is the initial analysis performed to understand and characterize a dataset.
- In **financial analysis**, appropriate indicators differ depending on the state of the company and industry, so it is important to consider various indicators as necessary. Visualize the transition of using Matplotlib

19

Visitor prediction, sales prediction

- There are various methods for **visitor forecasting**, and the appropriate method differs depending on the specific scenario. → It is important to select a model according to the nature of the data and the forecast period. Convert time-series data into a format suitable for Prophet, train the Prophet model, and predict future visitor numbers. Requires data preprocessing and model adjustment to match actual data and requirements
- That is same thing for **sales prediction**.

20

3C analysis: Company

- 3C analysis (**Company, Customer, Competitor Analysis**) is a method in which companies conduct analysis that focuses on themselves, their customers, and their competitors. It is difficult to give a concrete example of performing 3C analysis using Python code: 3C analysis involves subjective aspects and is difficult to express with simple code based on numerical data and algorithms.
- 1. Company Analysis: Analytical methods for evaluating the company itself include **financial analysis, SWOT analysis, value chain analysis, etc.**

21

Customer, competitor analysis

- 2. **Customer Analysis: Market research, customer feedback analysis, segmentation, etc.** are effective ways to understand customer needs and behavior.
- 3. **Competitor Analysis:** In competitive analysis, it is important to understand **the strengths and weaknesses of competitors and compare their market shares and strategies.**

22

PEST, STP analysis

- **PEST analysis** (Political, Economic, Social, Technological Analysis) is a method of analyzing factors in the external environment, and mainly focuses on four factors: politics, economy, society, and technology → Rather than a code for PEST analysis, The process of collecting and analyzing information about each factor is common.
- **STP analysis** (Segmentation, Targeting, Positioning Analysis) is a marketing method for segmenting the market (Segmentation), focusing on specific segments (Targeting), and appropriately positioning products and services in each segment (Positioning). strategy method

23

- There are various approaches to **customer analysis**, but one example uses K-Means clustering, which uses customer data to cluster and identify customer segments.
- **Customer Lifetime Value (LTV)** is a metric that evaluates the future value that a particular customer is expected to bring to a company. Calculate LTV using data.
- The cost to acquire one customer, CAC, is an abbreviation for "Customer Acquisition Cost" and is the unit cost of customer acquisition, and is calculated as follows, $CAC = \text{Cost required to acquire customers} \div \text{Number of new customers acquired}$. The cost required to acquire customers refers to the cost of sales activities and advertising/marketing costs: By understanding LTV and CAC, you can implement marketing measures that maximize cost effectiveness.

24

Decile, CS portfolio analysis

- **Decile Analysis** is a method that divides data into 10 equal parts and analyzes the characteristics of each group. In an example for performing decile analysis using Python, random data is generated and divided into deciles to calculate the mean and median.
- **CS Portfolio Analysis** (Customer Segmentation Portfolio Analysis) is a method for classifying customers into different segments and determining strategies and actions for each segment.

25

Conjoint, cohort analysis

- **Conjoint analysis** is a method for understanding **customer preferences** for the features and characteristics of products and services.
- **Cohort analysis** is a method of **grouping customers** who have registered or purchased during a specific period into the same group (cohort), and then tracking and comparing their subsequent behavior and transactions.

26

SWOT, ABC analysis

- **SWOT analysis** (Strengths, Weaknesses, Opportunities, Threats) is a framework for evaluating an organization or project's **internal strengths and weaknesses and external opportunities and threats**. SWOT analysis is primarily used to organize non-quantitative information and understand the strategic direction of an organization.
- **ABC analysis** is a method of classifying the importance of items used in inventory management, customer analysis, etc.

27

- **Five forces analysis** is useful for **understanding the competitive situation of an industry** and considering business strategies. View ratings for each element based on business environment factors. In specific businesses and industries, it is necessary to collect data and information and conduct more detailed and objective analysis.
- **Compare your market share, sales, and customer satisfaction with your competitors.** In actual **competitive analysis**, it is necessary to collect more detailed information and data on market trends, and to analyze in detail the differences between you and your competitors. Understand market positioning and points of differentiation through competitive analysis to establish competitive advantage.

28

Access, value chain analysis

- To **analyze the number of accesses**, it is necessary to use appropriate libraries and tools according to the specific data and purpose. Typical **access analysis** tools include Google Analytics.
- **Value chain analysis** is a method for **understanding a company's value** creation process and evaluating which activities contribute to customer value. Value chain analysis generally has no dedicated code and is primarily a method for understanding and evaluating business processes and value flows.

29

4P, RFM analysis

- **4P analysis** is a method for considering **marketing strategies**, and evaluates the four elements of **Product, Price, Place, and Promotion**.
- **RFM analysis (Recency, Frequency, Monetary)** is one of the methods useful for **customer segmentation** and marketing strategy planning. RFM consists of the following three elements. Recency: Time elapsed since the customer's last purchase. Frequency: How often a customer purchases. Monetary: Total amount of money the customer has purchased in the past.

30

MA, customer journey map

- **Marketing automation** is the use of automated tools and processes to streamline marketing efforts. Popular marketing automation tools offer features such as **email marketing, social media posting, website tracking, and lead management**.
- A **customer journey map** is a visual representation of how a customer uses a product or service and can help you understand **the customer experience for marketing and service improvement**.

31

Lead extraction, generation

- Generate hypothetical data about leads and **extract leads with high interest scores** → Interest scores are randomly generated, and in real projects, real data and marketing automation tools are used to analyze customer behavior and Rate your interest. Actual **lead generation** involves a combination of marketing strategies, digital advertising, SEO (search engine optimization), etc. → Specific methods and tools vary depending on business needs and industry.

32

Lead nurturing, qualification

- The lead data has an interest score and an engagement score, and if the interest score is high, send a white paper and update the engagement score: Once the engagement score is above a certain level, the **lead nurturing** process will proceed by **inviting you to a webinar**.
- **Lead qualification** processes vary depending on business requirements and sales department policies: It is necessary to clarify the conditions and judgment criteria for qualification and build a process to automatically make decisions based on them. → This typically involves **a combination of data analysis and machine learning techniques**

33

SFA

- **Salesforce Automation (SFA)** is a solution for **efficiently managing customer interactions**. Mainly used as part of CRM (**Customer Relationship Management**) tools. When using proprietary tools such as Salesforce, it is common to use APIs to retrieve and update data and perform actions. A simple example of creating data from Python to Salesforce using Salesforce's REST API (in practice, you need to be careful about Salesforce credentials and security)

34

CRM, funnel analysis

- CRM is a process or tool for effectively **managing customer information and strengthening customer relationships**. There is a code example for a simple CRM with CRUD (Create, Read, Update, Delete) operations. When using it in actual business, it is necessary to consider data security, access control, etc.
- **Funnel Analysis** is a method that visualizes the process a user takes toward a specific action or goal, and **analyzes the conversion rate** at each stage of that process.

35

SCM, sentiment analysis

- **SCM (Supply Chain Management)** is a process and tool for efficiently managing the entire supply chain from **production to sales**.
- **Sentiment Analysis** is the task of **extracting emotions and opinions from text data**. A basic example for doing sentiment analysis in Python uses the Natural Language Toolkit (NLTK) and VADER Sentiment Analyzer.

36

Summery

- Python code as a tool for business analysis, financial analysis, marketing analysis, etc. can be created relatively easily by using libraries and frameworks.
- Executing Python code requires environment settings, which is a tedious task. The analysis method introduced here does not require heavy resources, so we recommend running it on [Google Colaboratory](#), which does not require any environment settings.
- It takes a lot of time to prepare **big data** to be input into analysis tools → Data cleaning and cleansing are also important

What's Big Data?

No single definition; here is from Wikipedia:

- **Big data** is the term for a **collection of data sets so large and complex that it becomes difficult to process** using on-hand database management tools or traditional data processing applications.
- The challenges include **capture, curation, storage, search, sharing, transfer, analysis, and visualization as well as integration.**
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to **"spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."**

Topics

- Scope: Big Data & Analytics
- Topics:
 - Foundation of Data Analytics and Data Mining
 - Hadoop/Map-Reduce Programming and Data Processing & BigTable/Hbase/Cassandra
 - Graph Database and Graph Analytics

Big Data

- So large data that it becomes difficult to process it using a 'traditional' system

Big Data

Volume - examples

- Facebook processes 500 TB per day
- Walmart handles 1 million customer transactions per hour
- Airbus generates 640 TB in one flight (10 TB per 30 minutes)
- 72 hours of video uploaded to youtube every minute
- SMS, e-mail, internet, social media

[illegible]

The amount of data is getting large almost twice every year

1.75 ZB of Big data around the world in 2023

Platform/Activity	Amount per Minute
Facebook	1.9 Million
Google	4.1 Million Search Queries
Netflix	764,000 Hours Watched
Amazon	1.7 Million Items Sold
WhatsApp	2.5 Million Messages Sent
Instagram	59 Million Photos Taken
Imur	3.5 Million Images Taken
Spotify	105 Smart Speakers Shipped
Amazon	1,400 Amazon Echoes Sold
Twitter	1.2 Million Tweets Sent
Tinder	1.6 Million Swipes
Twitter	194,444 People Tweeting
Instagram	694,444 Stories Uploaded
Instagram	400,000 Apps Downloaded
YouTube	10 Million Hours Viewed
YouTube	1.9 Million Views

Created by: @davidmiller
@PollyDataHub

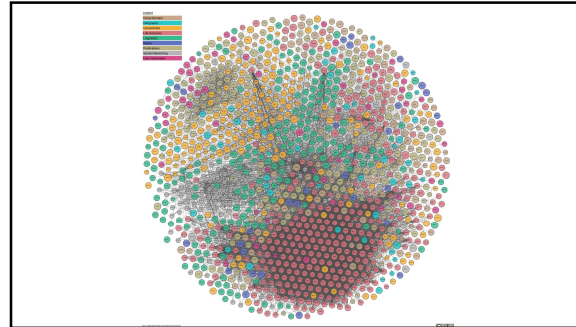
<https://www.allaccess.com/merge/archive/31294/infographic-what-happens-in-an-internet-minute/>

45

Subclouds by domain

Cross-Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User-Generated

46



47

Linked open data of US government

Format (# Datasets) <http://catalog.data.gov/>

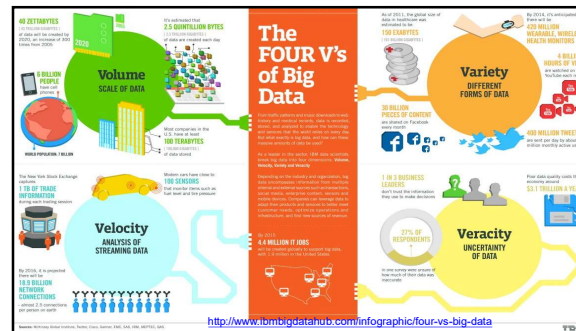
- HTML (101140)
- PDF (45026)
- XML (36564)
- ZIP (22376)
- CSV (21190)
- Originator data format (21023)
- WMS (16531)
- JSON (16196)
- TIFF (14180)
- SID (11874)
- RDF (11798)

48

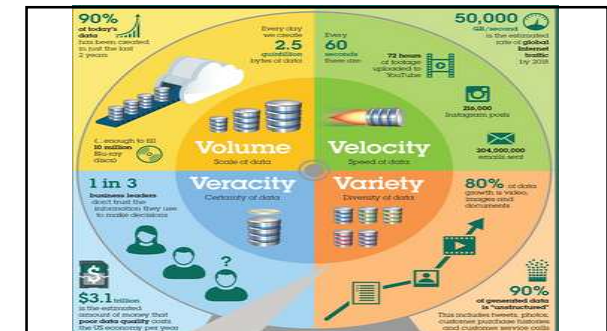
Velocity - examples

- Traffic data
- Financial market
- Social networks

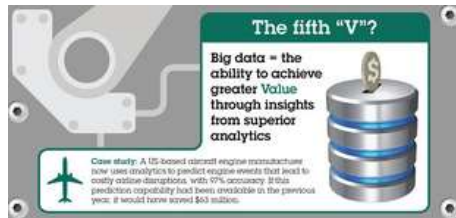
49



50



51

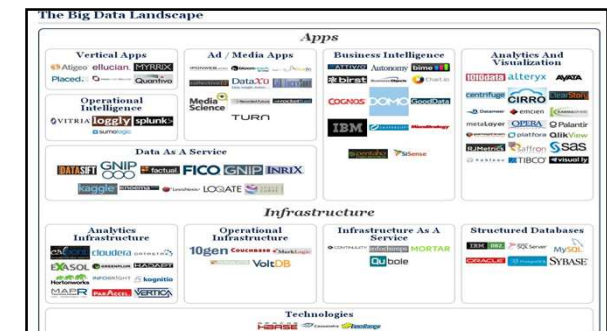


52

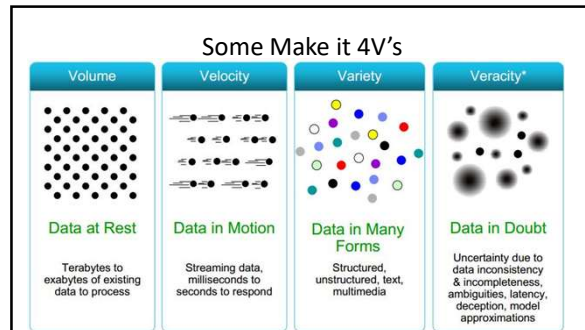
Real-Time Analytics/Decision Requirement



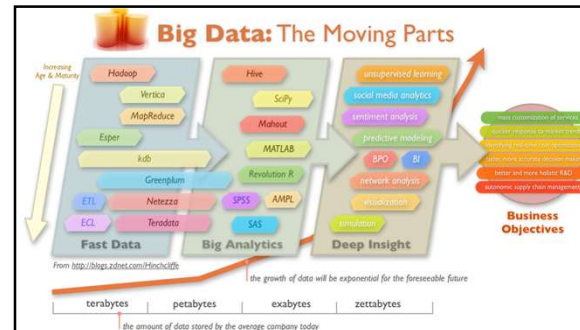
53



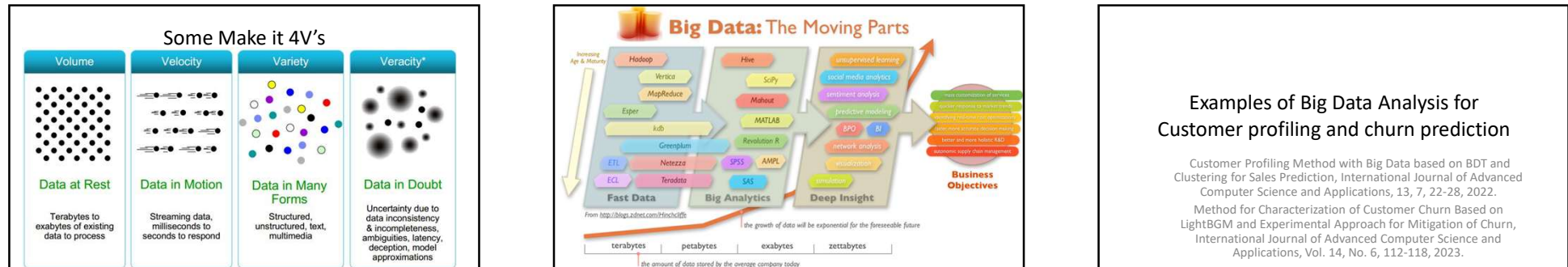
54



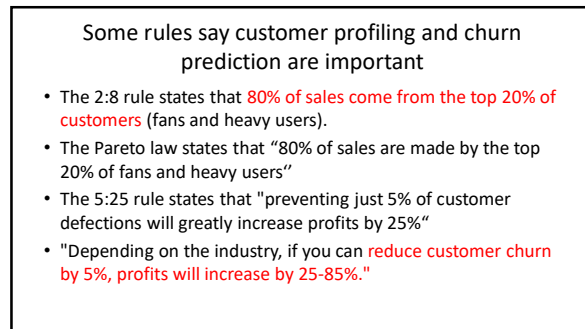
64



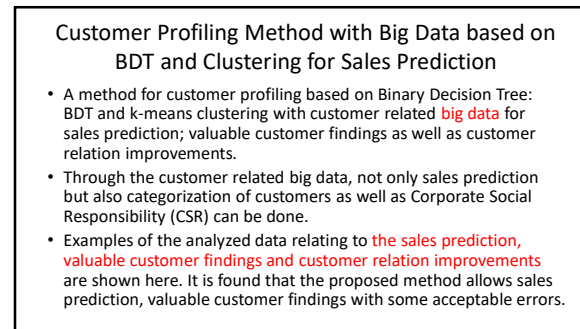
65



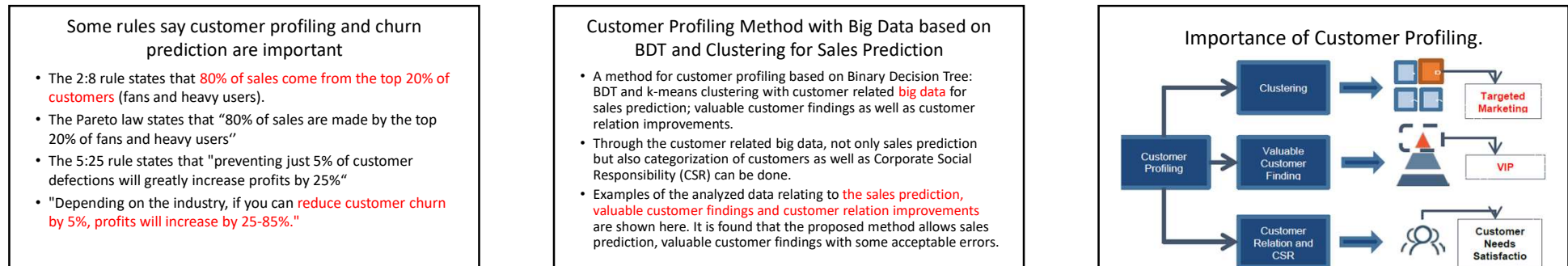
66



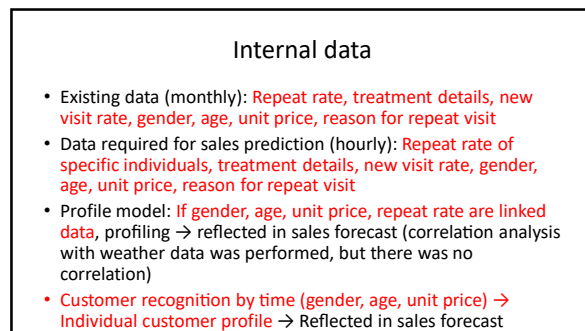
67



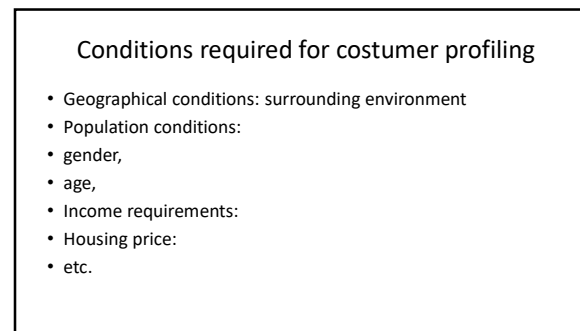
68



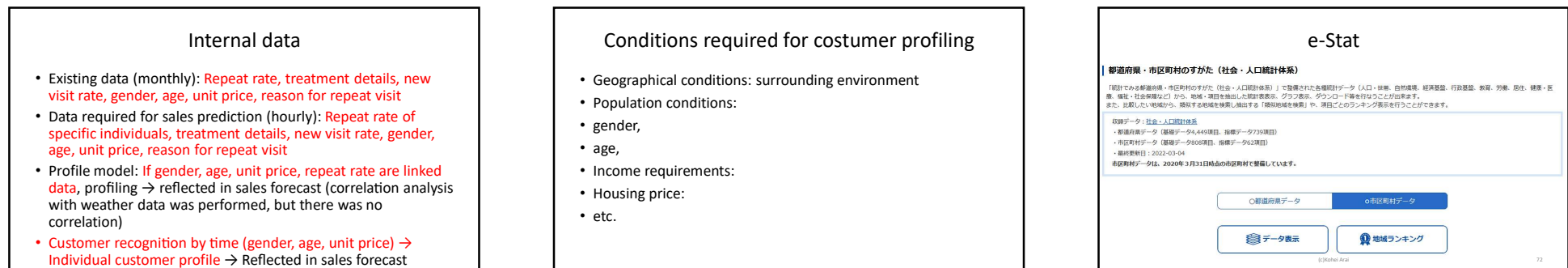
69



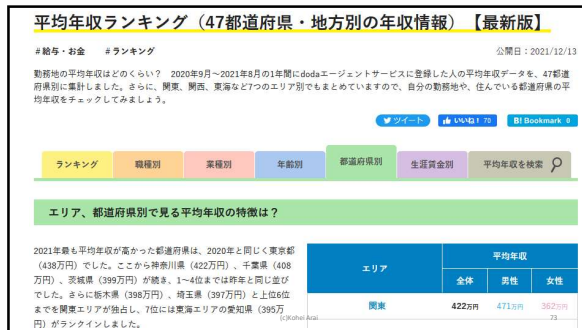
70



71



72

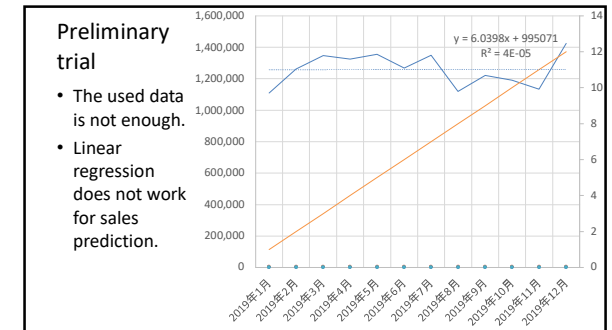


73

Variables

- **Objective variable: Sales,**
- Explanatory variables: Gender (no correlation), age (no correlation), seasonal change (some correlation with number of months: 0.492), number of holidays (0.412)
- Trend analysis: Trend is not fixed → Nonlinear regression is required (make the prediction function nonlinear or DL) → **prophet of non-linear regressive model is attempted.**

74



75

Insufficient explanatory variables

- Linear regression is difficult → Prediction using DL (practice Bitcoin price prediction): <https://www.codexa.net/keras-lstm-cryptos-forecast/Insufficient-explanatory-variables> → Introducing personal profiles
- It is necessary to reacquire POS data to understand time zone changes.
- At that time, perform face registration and facial recognition to identify the individual and profile the individual customer → DL learning

76

The second trial: PARAMETERS FOR SALES PREDICTION

- **Tier 7. Prediction**
- **Tier 6. Brand Affinity**
- **Tier 5. Shopping Pattern**
- **Tier 4. Preference & Interests**
- **Tier 3. Geographic**
- **Tier 2. Demo-graphics**
- **Tier 1. Recognition & Contact**

77

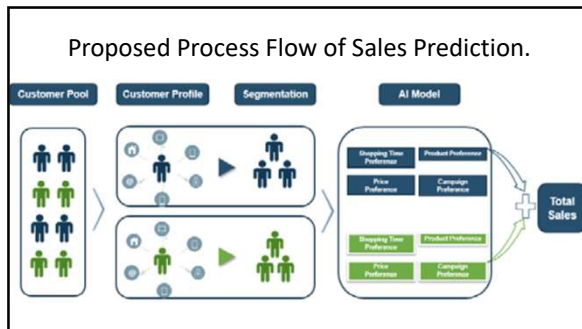
Data gathering

• Data Labels (Parameters) in the Assigned Layers.

• It takes time.

Data Labels			
Risk to lose	Propensity to buy	Predicted date to come	
Membership Tenure	First Ten Date	Last Ten Date	Total Membership Points
Total Spending	Total Tens	Campaign Response Rate	
Annual Tens	Item Per Ten	ATV Average Ten Value	
Favorite Category		Favorite Item	
EM/DM	Weekday/Weekend/Holiday	Weather	Specific Date
Province	City	Address	Post Code
Age	Gender	Job/Occupation	Annual Revenue
Customer ID	Phone	Name	Email

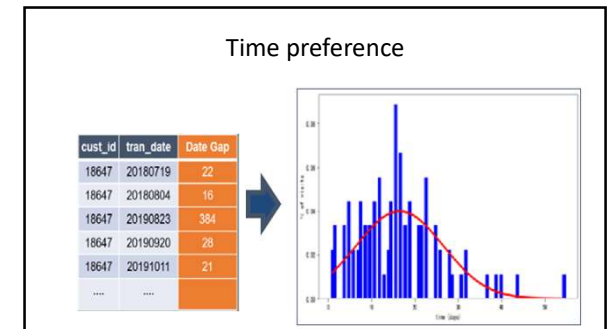
78



79

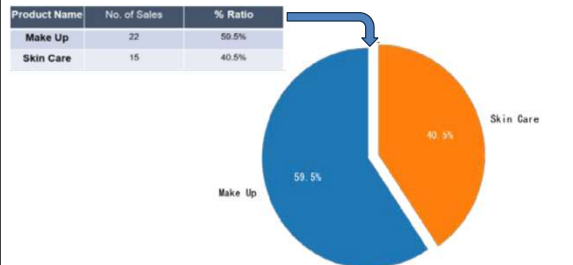


80



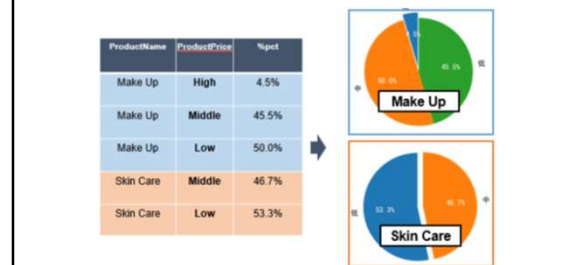
81

Product Preference.



82

Price Preference.



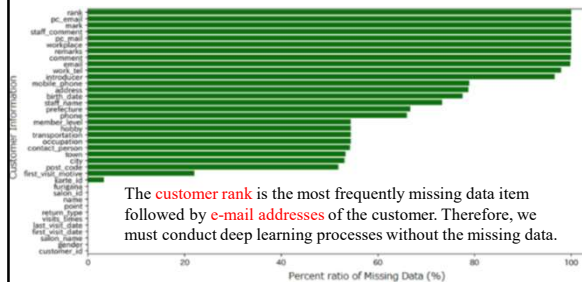
83

SALES DATA FOR THE SPECIFIC CUSTOMER

Customer ID	Product ID	Product Price	No. of Sales
133034	32	Skin Care	1
133034	48	Skin Care	1
133034	80	Skin Care	1
133034	25	Make Up	1
133034	33	Make Up	1
133034	42	Make Up	1
133034	79	Make Up	2
133034	82	Make Up	3
133034	1	Make Up	2
133034	4	Make Up	2
133034	7	Make Up	1
133034	12	Skin Care	1
133034	19	Make Up	1
133034	24	Skin Care	1
133034	29	Make Up	1
133034	36	Make Up	1
133034	37	Skin Care	1
133034	40	Make Up	2
133034	52	Skin Care	1
133034	54	Make Up	1

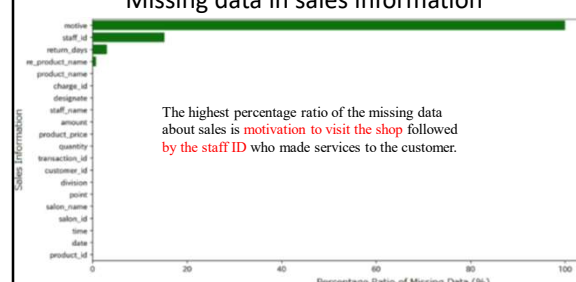
84

Missing data in customer information



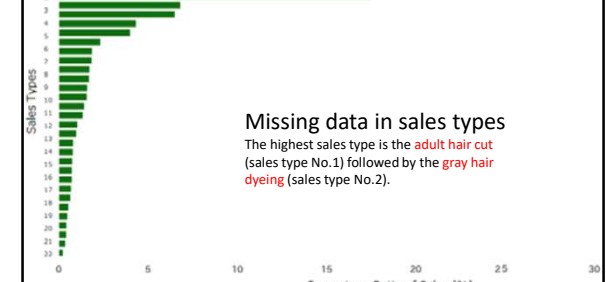
85

Missing data in sales information



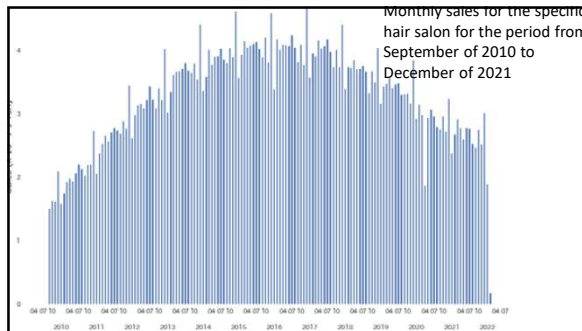
86

Missing data in sales types



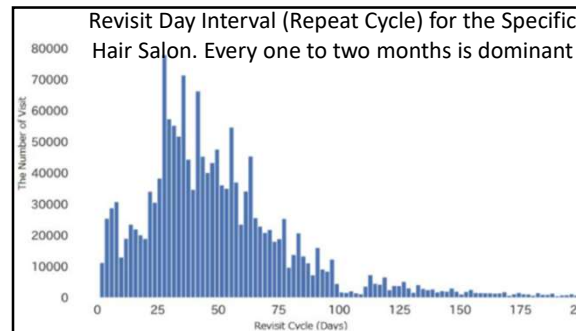
87

Monthly Sales for the specific hair salon for the period from September of 2010 to December of 2021



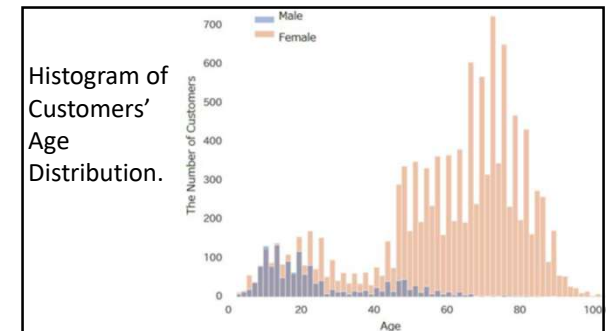
88

Revisit Day Interval (Repeat Cycle) for the Specific Hair Salon. Every one to two months is dominant



89

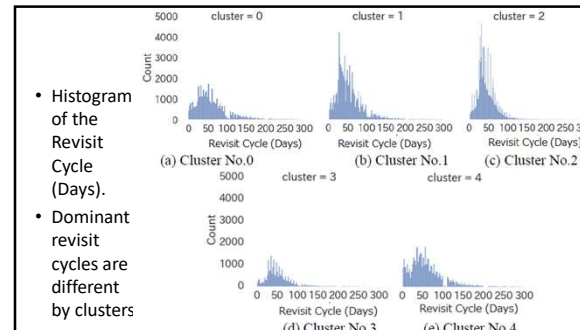
Histogram of Customers' Age Distribution.



90



91

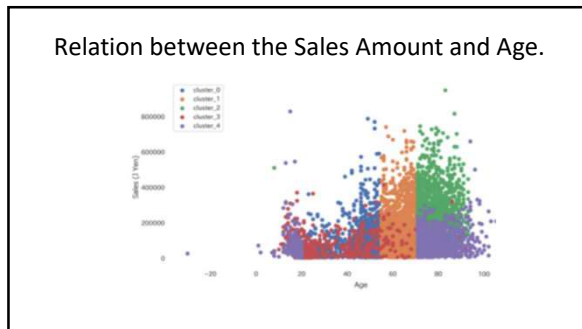


92

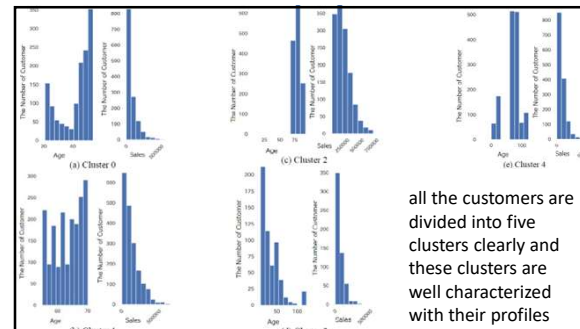
NUMBER OF CUSTOMERS FOR EACH CLUSTER

Cluster No.	The _Number_of_ Customers
1	1833
4	1439
2	1360
0	1316
3	564

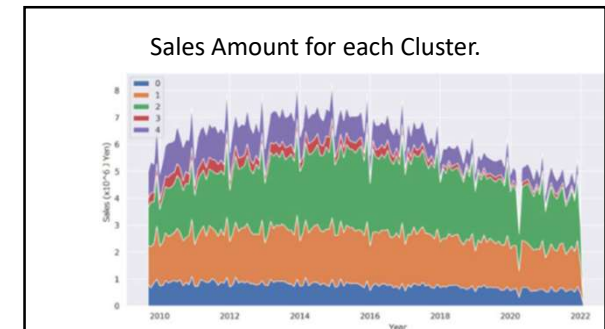
93



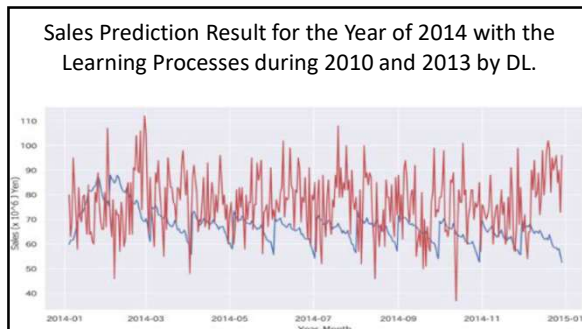
94



95



96



97

Summery

- The method for **customer profiling works well**. All the customer characteristics are clarified.
- Examples of the analyzed data relating to **the sales prediction, valuable customer findings and customer relation improvements shows marginal accuracy**.
- It is found that the proposed method allows sales prediction, valuable customer findings with some acceptable errors (**21.25% of RMSE**) based on DL.

98

Churn Prediction

- By predicting customers who are likely to defect using decision tree models such as **LightGBM**, which is a machine learning method, and logistic regression, we will discover important feature values in prediction and utilize the knowledge obtained through Exploratory Data Analysis (EDA).
- As results for experiments, it is found that the proposed method allows **estimation and prediction of churn customers as well as characteristics and behavior of churn customers**. Also, it is found that the proposed method is superior to the conventional method, **GradientBoostingClassifier (GBC)** by around 10%.

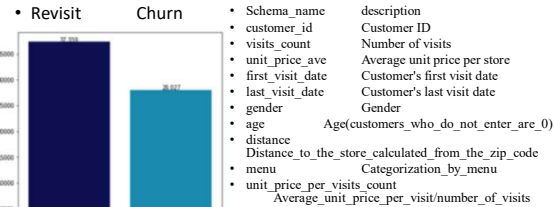
99

The Data Used

- The data used in the experiment targeted approximately **65,000 customers of POS customer data(Big data) from 1 September 2009 to December 31, 2021.**
 - 1) **Total number of customers (persons) 878,181** Number of unique customer IDs
 - 2) **Total number of cases (cases) 8,857,257** Number of sales item IDs (cut and color are counted as 2 cases, discounts are also counted as 1 case)
 - 3) **Cancellation of sales (number of cases) 350,017** Number of sales cancellation details

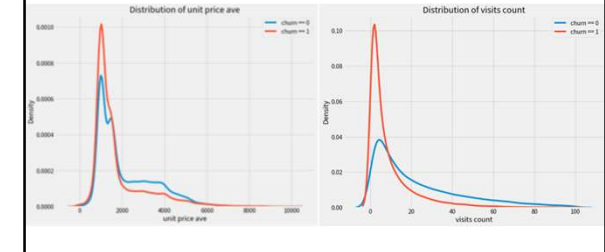
100

Data types



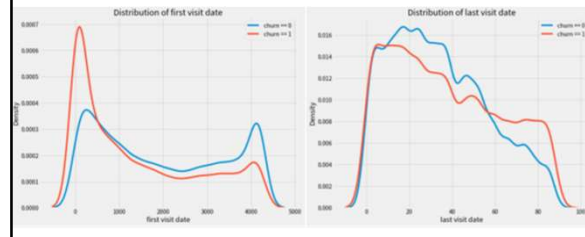
101

Price and the number of visitors



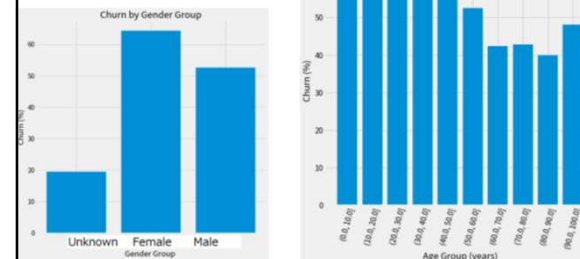
102

First visit date and last visit date



103

Gender and age



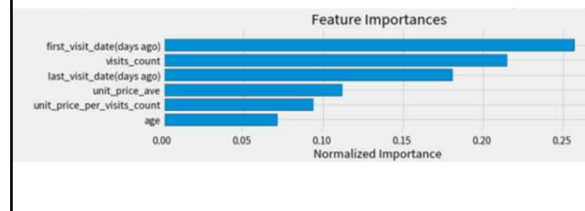
104

Menu and price



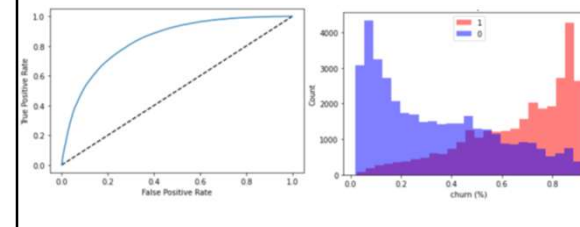
105

Importance factor



106

ROC curve



107

Summery

- 1) Age: Younger customers have a higher attrition rate, while those in their 60s to 80s have a lower attrition rate.
- 2) Gender: The churn rate is higher for female customers.
- 3) Number of visits: Customers with a lower number of visits have a higher churn rate.
- 4) Unit price (per hair salon visit): Customers with lower unit prices have a higher withdrawal rate.
- 5) Date of first visit: Customers who have recently visited for the first time have a higher churn rate.
- 6) Date of last visit: The churn rate is lower for customers who have recently visited.
- 7) Menu: Gray hair dye customers have a low attrition rate, while school and child cuts have a high attrition rate.
- 8) Distance to hair salons: The presence or absence of parking lots seems to have little impact on the churn rate.

108

Conclusion

- Data driven society requires big data analytics for predictive analysis and decision making in real time basis
- There are so many analysis methods. Most of those can be created with Python code (library, frameworks)
- Examples of AI utilized big data analysis, customer profiling and churn prediction as well as sales prediction can be done in online basis for quick decision making

109

Related papers

- Kohei Arai, Ikuya Fujikawa, Yusuke Nakagawa, Ryoya Momozaki, Sayuri Ogawa, **Modified Prophet+Optuna Prediction Method for Sales Estimations**, International Journal of Advanced Computer Science and Applications, 13, 8, 58-63, 2022.
- Kohei Arai, Ikuya Fujikawa, Yusuke Nakagawa, Ryuya Momozaki, Sayuri Ogawa, **Churn Customer Estimation Method based on LightGBM for Improving Sales**, International Journal of Advanced Computer Science and Applications, 14, 2, 119-125, 2023.

110

- Kohei Arai, Ikuya Fujikawa, Yusuke Nakagawa, Ryoya Momozaki, Sayuri Ogawa, **Method for Characterization of Customer Churn Based on LightGBM and Experimental Approach for Mitigation of Churn**, International Journal of Advanced Computer Science and Applications, Vol. 14, No. 6, 112-118, 2023.
- Kohei Arai, Ikuya Fujikawa, Yusuke Nakagawa, Sayuri Ogawa, **Method for Predictive Trend Analytics with SNS Information for Marketing**, International Journal of Advanced Computer Science and Applications, Vol. 15, No. 2, 419-425, 2024.

111

Required tool 1: Data Analytics & Data Mining

- Exploratory Data Analysis
- Linear Classification (Perceptron & Logistic Regression)
- Linear Regression
- C4.5 Decision Tree
- Apriori
- K-means Clustering
- EM Algorithm
- PageRank & HITS
- Collaborative Filtering

112

Required tool 2: Hadoop/MapReduce Programming & Data Processing

- Architecture of Hadoop, HDFS, and Yarn
- Programming on Hadoop
- Basic Data Processing: Sort and Join
- Information Retrieval using Hadoop
- Data Mining using Hadoop (Kmeans+Histograms)
- Machine Learning on Hadoop (EM)
- Hive/Pig
- HBase and Cassandra

113

Required tool 3: Graph Database and Graph Analytics

- Graph Database (http://en.wikipedia.org/wiki/Graph_database)
 - Native Graph Database (Neo4j)
 - Pregel/Giraph (Distributed Graph Processing Engine)
- Neo4j/Titan/GraphLab/GraphQL

114

Textbooks

- No Official Textbooks
- References:
 - Hadoop: The Definitive Guide, Tom White, O'Reilly
 - Hadoop In Action, Chuck Lam, Manning
 - Data-Intensive Text Processing with MapReduce, Jimmy Lin and Chris Dyer (www.umiacs.umd.edu/~jimnylin/MapReduce-book-final.pdf)
 - Data Mining: Concepts and Techniques, Third Edition, by Jiawei Han et al.
- Many Online Tutorials and Papers

115

Cloud Resources

- Hadoop on your local machine
- Hadoop in a virtual machine on your local machine (Pseudo-Distributed on **Ubuntu**)
- Hadoop in the clouds with Amazon EC2

116

Prerequisite

- Prerequisite:
 - Java Programming / C++
 - Data Structures and Algorithm
 - Computer Architecture
 - Basic Statistics and Probability
 - Database and Data Mining (preferred)
 - Python

117

This is not for you...

- If you do not have a strong Java programming background
 - This course is not about only programming (on Hadoop).
 - Focus on “thinking at scale” and algorithm design
 - Focus on how to manage and process Big Data!
- No previous experience necessary in
 - MapReduce
 - Parallel and distributed programming