BIBA: Business Intelligence and Big Data

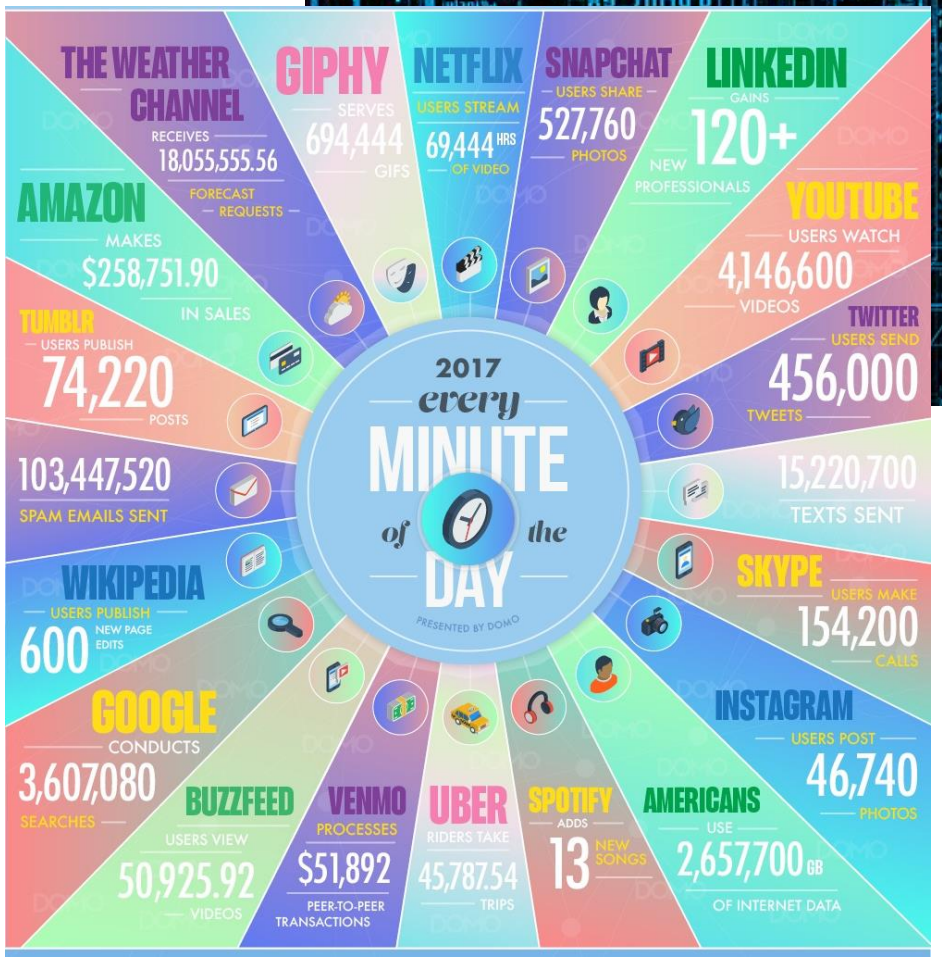
2018-09-12

Jens Ulrik Hansen

Today’s program

- Introduction to the course
- Introduction to the exam and the synopsis
- Introduction to Business Intelligence
- How to find example data
- Introduction to R
- Hands-on exercises i R

Introduction to the course


2017
every
MINUTE
of the
DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL
RECEIVES
18,055,555.56
FORECAST REQUESTS
GIPHY
SERVES
694,444
GIFS
NETFLIX
USERS STREAM
69,444 HRS
OF VIDEO
SNAPCHAT
USERS SHARE
527,760
PHOTOS
LINKEDIN
GAINS
NEW 120+
PROFESSIONALS
AMAZON
MAKES
$258,751.90
IN SALES
YOUTUBE
USERS WATCH
4,146,600
VIDEOS
TUMBLR
USERS PUBLISH
74,220
POSTS
TWITTER
USERS SEND
456,000
TWEETS
103,447,520
SPAM EMAILS SENT
15,220,700
TEXTS SENT
WIKIPEDIA
USERS PUBLISH
600
NEW PAGE EDITS
SKYPE
USERS MAKE
154,200
CALLS
GOOGLE
CONDUCTS
3,607,080
SEARCHES
INSTAGRAM
USERS POST
46,740
PHOTOS
BUZZFEED
USERS VIEW
50,925.92
VIDEOS
VENMO
PROCESSES
$51,892
PEER-TO-PEER TRANSACTIONS
UBER
RIDERS TAKE
45,787.54
TRIPS
SPOTIFY
ADDS
13
NEW SONGS
AMERICANS
USE
2,657,700 GB
OF INTERNET DATA

Introduction to the course

- Using data (and data analysis) to solve business problems
- What are the computer science challenges related to this (especially concerning large amount of data)

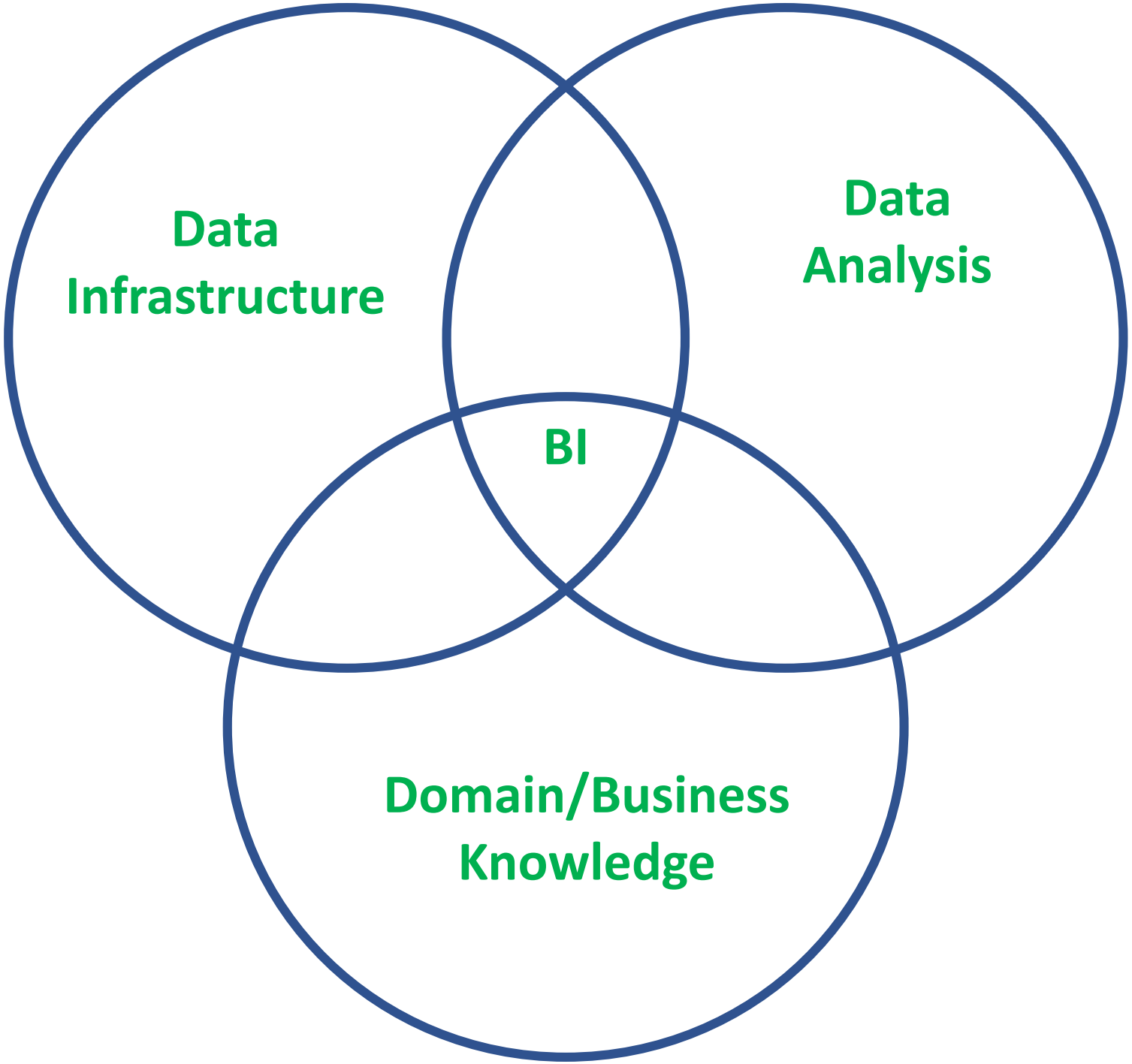
Introduction to the course

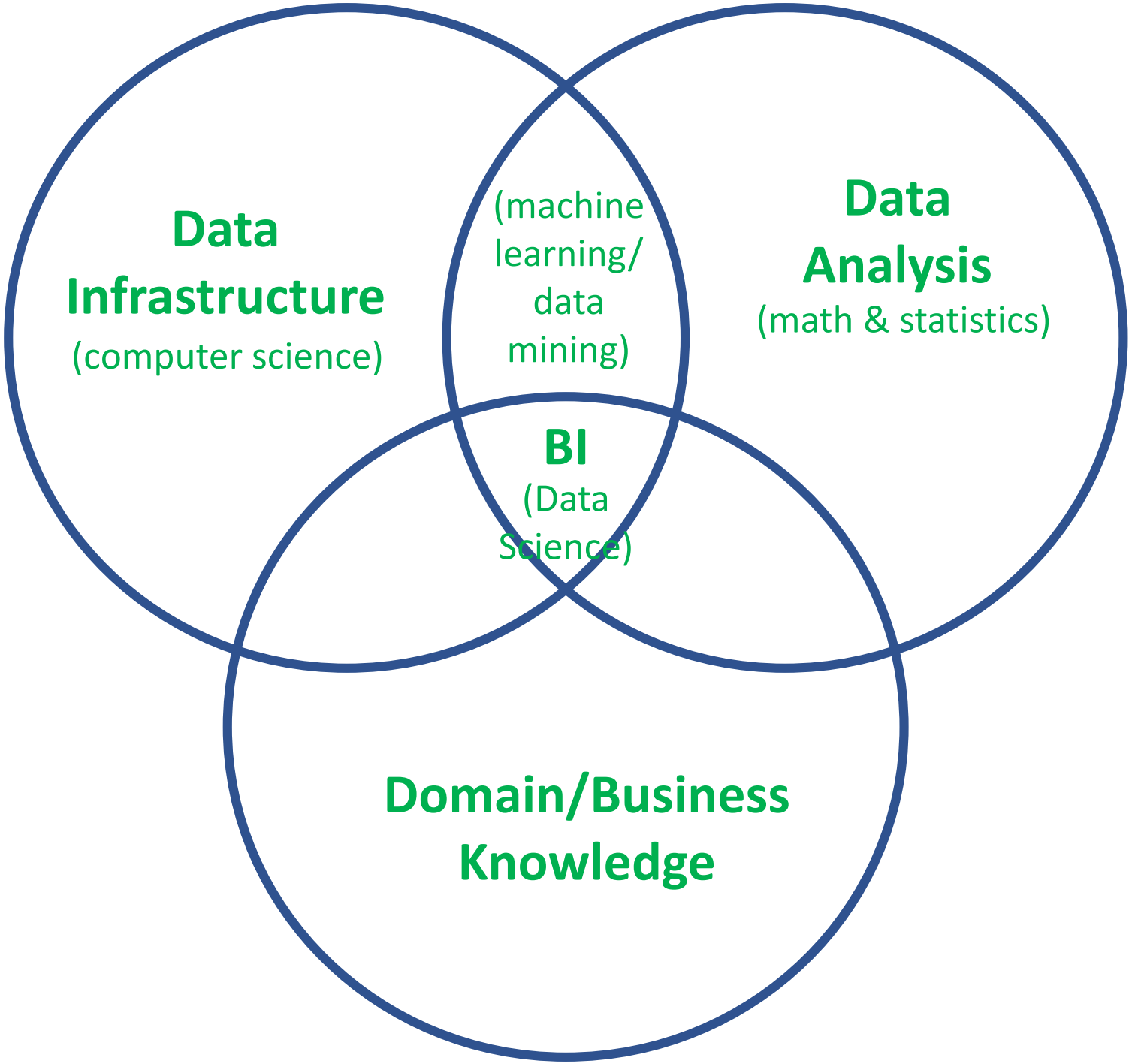
- Using data (and data analysis) to solve business problems
  1. identify business problem
  2. collect data
  3. prepare data
  4. analyze data
  5. conclude and communicate
- What are the computer science challenges related to this (especially concerning large amount of data)
  - Storage of big data
  - Computation on big data
- Final reflections: GDPR, Ethics, etc.

Introduction to the course

- Using data (and data analysis) to solve business problems
  1. identify business problem
  2. collect data
  3. prepare data
  4. analyze data
  5. conclude and communicate

- What are the computer science challenges related to this (especially concerning large amount of data)
  - Storage of big data
  - Computation on big data
- Final reflections: GDPR, Ethics, etc.


Data
Infrastructure
Data
Analysis
BI
Domain/Business
Knowledge


Data
Infrastructure
(computer science)
(machine
learning/
data
mining)
Data
Analysis
(math & statistics)
BI
(Data
Science)
Domain/Business
Knowledge

After this course you will be able to

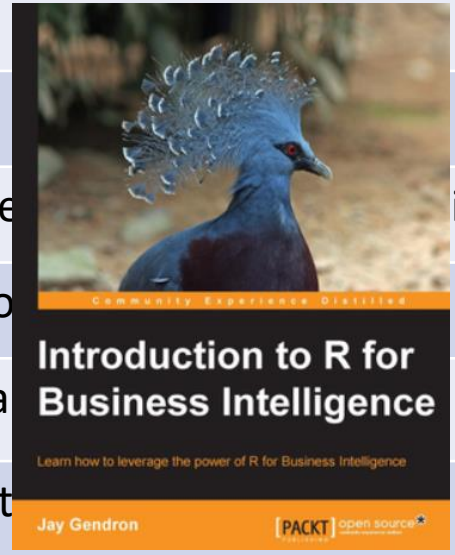
- Identify business cases for use of BI/Data Science/Machine Learning/AI
- Understand the entire data analysis process from identifying problem to presenting results
  - Implement such a process yourself in R
- Understand the challenges of the necessary data infrastructure in our current data world

Course outline

Date	Title	Content
12/9	Introduction	Intro to course, intro to BI, intro to R, example datasets
19/9	Data	Data formats, tidy data, ETL, data cleaning
26/9	Explorative data analysis	EDA, descriptive statistics, basic plotting
3/10	Correlation and causation	Basic correlation, correlation vs causation, Linear regression w. BI applications
10/10	Clustering and Classification	Clustering and classification models w. BI applications
17/10	More modelling…	Association rule mining, time series analysis w. BI applications, modelling
24/10	Visualization and storytelling with data	Visualization, storytelling with data, dashborads, other BI tools
31/10	Big Data	Big Data, SQL, No-SQL, JSON, map-reduce
7/11	Data Storage	Data warehouse, Hadoop, cloud computing
14/11	Final reflections	GDPR, ethics, related topics, rounding off

A book plus papers and chapters etc.

Date	Title	Content
12/9	Introduction	Intro to course, intro to BI, intro to R, example datasets
19/9	Data	Data formats, tidy data, ETL, data cleaning
26/9	Explorative data analysis	EDA, descriptive statistics, basic plotting
3/10	Correlation and causation	Basic correlation, correlation vs causation, Linear re ions
10/10	Clustering and Classification	Clustering and classification models w. BI applicatio
17/10	More modelling…	Association rule mining, time series analysis w. BI a
24/10	Visualization and storytelling with data	Visualization, storytelling with data, dashborads, ot
31/10	Big Data	Big Data, SQL, No-SQL, JSON, map-reduce
7/11	Data Storage	Data warehouse, Hadoop, cloud computing
14/11	Final reflections	GDPR, ethics, related topics, rounding off



Introduction to the exam and the synopsis

The exam and the synopsis

- Individual oral exam (20 min) based on a small written synopsis
  - Present the synopsis briefly
  - We will ask question to the synopsis and the content of the course in general
- The synopsis
  - 5-10 pages (plus graphs) – Officially: max 48.000 characters including spaces
  - Cover all elements from the course: Identify a business problem (a case), get data, prepare data, analyze data, create models, explain how the analysis and models help solve the business problem, reflect on big data technologies and ethical an regulatory aspects
  - 3 hand-ins of the synopsis during the semester (only general feedback will be given):
    - Business case and selection of data – September 16
    - Explorative data analysis and data cleaning – September 30
    - Modeling – October 21
  - (You are allowed to change these three parts later before the final hand-in)
  - Final hand-in of the synopsis on eksamen.ruc.dk – November 26

Business case description hand-in this week!

- Deadline for hand-in: September 16, 23:55
- Length: half a page
- Your business case description should contain information about what type of business or organization it is (it can be both a real or fake business) and what business problems you aim to solve by analyzing data. You should also briefly describe you data set, where to find it, and how it relate to the business.
- See Moodle: https://moodle.ruc.dk/course/view.php?id=10876
- We will put all cases into a single sheet such that you can get inspired by others
- You are allowed to change you business case before next hand-in or before the final submission of your final synopsis

Business Intelligence

Some definitions of BI

- *"BI is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance."* - Gartner
- *"BI is about delivering relevant and reliable information to the right people at the right time with the goal of achieving better decisions faster"* - Hitachi Solutions
- *"BI refers to a group of tools and techniques that collects and organize your data and presents it in a way that is useful and make sense"* - Hitachi Solutions
- *"Applications that transform data into meaningful information which helps business make better decisions"* - TechnologyAdvice
- *"Business Intelligence is essentially timely, accurate, high-value, and actionable business insights, and the work processes and technologies used to obtain them."* - S Scheps "Business Intelligence for Dummies"

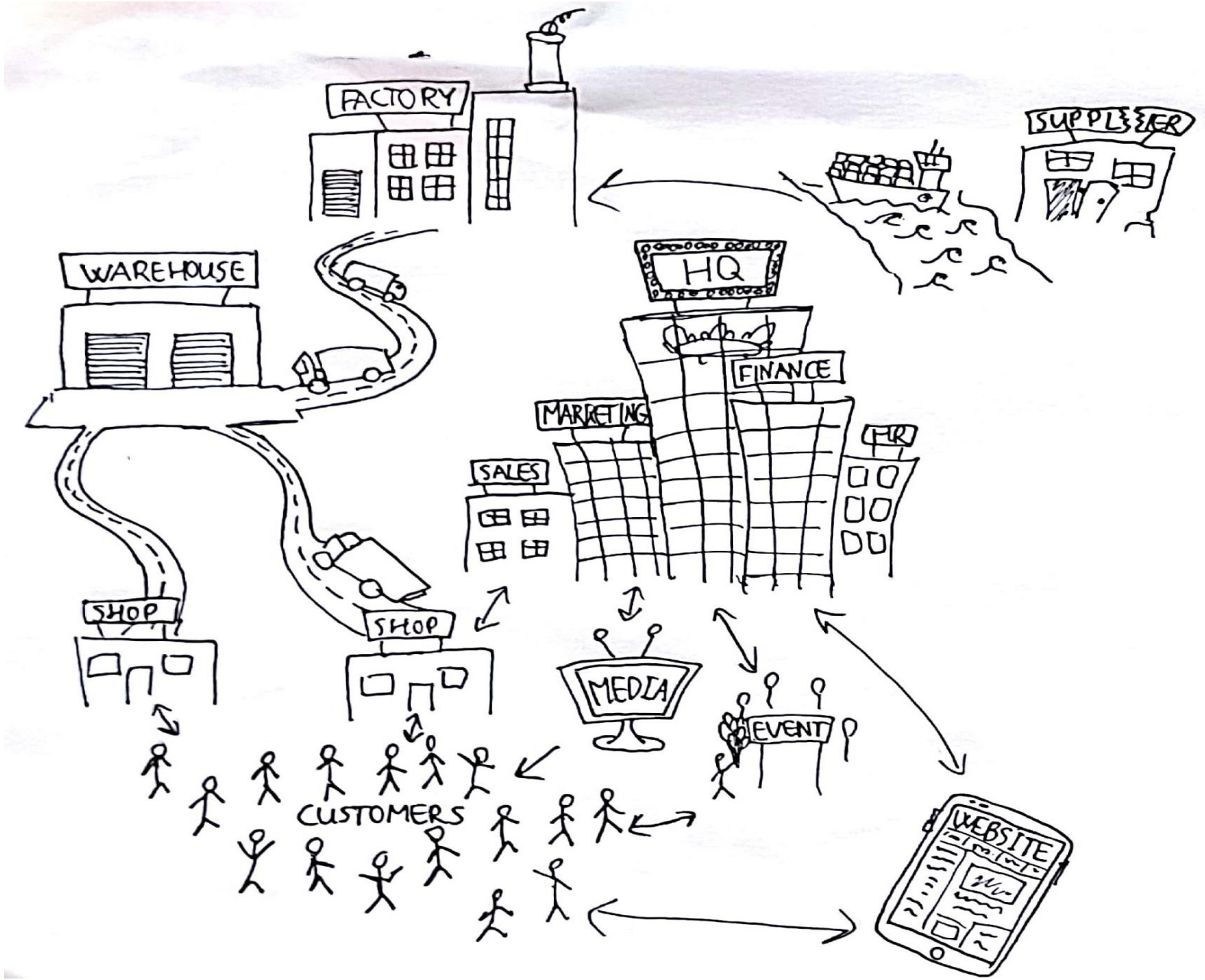
What they have in common

- ***They all seem to agree on BI involving:***
  - Tools and techniques
  - Data into information and business insights
  - Better decisions
  - Timing
- ***They all highlight the importance of:***
  - Quality of information (reliable and accurate)
    - GIGO-rule: Garbage in, garbage out
  - The right information (relevant, high-value, and meaningful)
    - It needs to actionable
  - Analysis of data
    - The right analysis can make a big difference


Business Intelligence
by examples

BI examples

- BI can be applied in:
  - *Marketing*
  - *Sales*
  - *Supply-Chain Management*
  - *Financial Management*
  - *Production and supply*
  - *Customer Relation Management*
  - *Customer support*
  - *HR*
  - *Web appearance, marketing, and sales*



Data Science at Booking.com

- Getting customers
  - *"web marketing, attribution models and ROI predictions help bring customers to our site"*
- Improve product
  - *"… recommendation systems help us show more relevant destinations, hotels and content to our users"*
- Customer service
  - *"… call volume predictions and scheduling algorithms help staff our call centers and connect customers to the right agent as quickly as possible"*
- *"In fact, I honestly struggle to think of a single department that is not using predictive analytics in one way or another."*
- *"… we try to validate almost all changes we make to our product through in vivo randomised controlled trials"*
- *" We need all of our product development teams to understand the basics of hypothesis testing and the statistics behind it."*
- https://predictiveanalyticsworld.co.uk/interviews/interview-with-lukas-vermeer-data-scientist-booking-com/

Predicting customer behavior

- Based on historical data one can make models to:
  - Predict how much of a product will be sold
  - Suggest other related products to a customer
  - Recommend what product to place together in a physical store
  - Predict how a customer will rate a product
  - Predict what product a costumer will be interested in at a given moment of time
  - Suggest other customers to link to
- Example companies
  - Amazon, Netflix, ...
  - Wal-Mart, Target, ...
  - Facebook, LinkedIn, …
- Note, this does not only apply to online business!
  - https://www.nytimes.com/video/business/100000002206849/big-data-hits-real-life.html

Marking

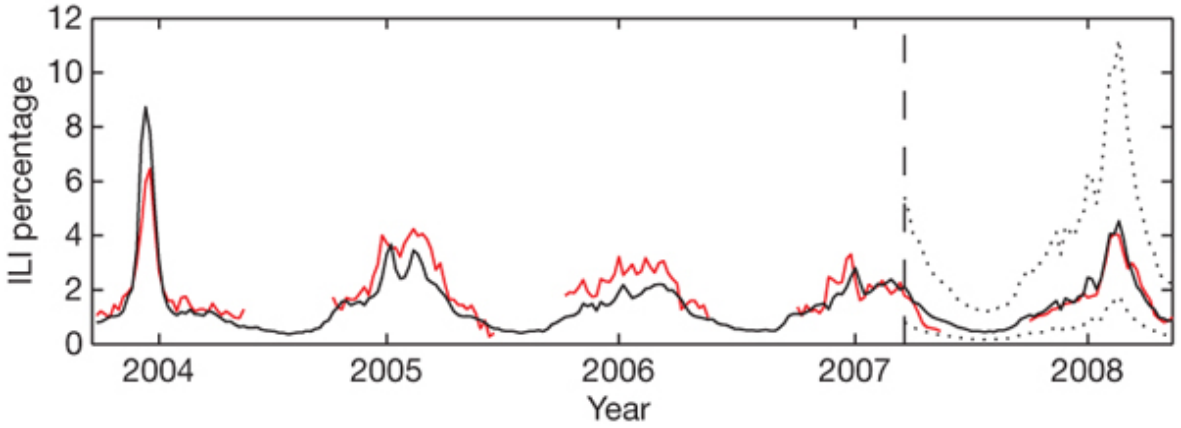
- Predict the effect of different marketing campaigns
- Predict the effect of different medias
  - TV adds vs online banners
- Targeted marketing
  - In emails and on facebook
- Social media effects
- Online add placement
- Digital marketing/Data Driven Markting
  - https://www.youtube.com/watch?v=_PWqIMQux-g
  - https://www.youtube.com/watch?v=ar4rR9FWBqg

People Analytics – BI examples in HR

- *“Now a growing number of businesses are applying analytics to processes such as recruiting and retention, uncovering surprising sources of talent and counterintuitive insights about what drives employee performance.”*

- *https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/using-people-analytics-to-drive-business-performance-a-case-study*

Google Flue Trend (USA 2008)



- Traditionally, doctors in the US report cases of flue to the CDC (Centers for Disease Control and Prevention) on a regular basis and CDC updates their statistics once a week.
- For that reason, there can be a delay of up to 2 weeks in the flue statistics
- Google Flue Trend used historical data from the CDC together with search data (people searching for flue symptoms, cough, etc.) to make predictions of flue cases in each state in real time.
- In 2008 Google Flue Trend predicted flue cases with a 97% precision
- Since then, there have been great variation in the precision of Google Flue Trend and the predictions are no longer made.
- A good example of data from one domain is used to make predictions in another domain – combining data is where the real value lies
- http://static.googleusercontent.com/media/research.google.com/en/us/archive/papers/detecting-influenza-epidemics.pdf

BI related concepts

- Business Intelligence, Business Analytics
- Recommendations and Decision Support Systems
- Machine Learning, Data Mining, Statistics, Data Science
- Big Data, Big Data Analytics
- Advanced Analytics, Predictive Analytics
- Data Warehouse, ETL (Extract transform load)
- OLAP (Online Analytical Processing)
- Data Engineering
- Artificial Intelligence, Robotics

Key Performance Indicators (KPIs)

- Measures/metrics of business performance and success
  - One can measure a quantity, progress or change, for instance
- Examples
  - Revenue/turnover/sales
  - Number of costumers
  - Number of potential costumers
  - Number of new customers
  - Number of website visit/number visitors to a physical store
  - Active users
  - Customer Lifetime value
  - Churn (Leaving customers)
  - ROI (Return on investment) (profit – cost) / cost
  - CPS (cost per sale)
- BI is also about calculating, presenting and predicting KPIs

Big Data and BI in organizations

Big Data MBA – a few Chapter 1 points

- ”[organizations] do not need a big data strategy as much as they need a business strategy that incorporates big data”
- BI/Data Analysis is not another tool, it is at the center of doing business
- The Data Analysis needs to be business centric
- You need to be willing to act on the knowledge gained from data
- ”[Business leaders must ask] How effective is our organization at integrating data and analytics into our business model?”
- There is a difference between:
  - Using data to improve an existing business
  - Building a business on data
  - Disrupting an industry by being a data company – disruption happened not as much because of big data technologies, but because of the way they are used by businesses

Big Data MBA – think differently

- ”Don’t Think Big Data Technology, Think Business Transformation”
  - Business first. Set business goals and then figure out what analysis, data and technology you need
- ”Don’t Think Business Intelligence, Think Data Science”
  - A useless distinction, I think: Just use the descriptive, predictive, and prescriptive data analysis tools and models that help you improve your business – if it makes you money it doesn’t matter what you call it”
- ”Don’t Think Data Warehouse, Think Data Lake”
  - Data is used for different things that might require different storage technologies (Henrik will return to the issue of storage)
  - Simple sales data reporting for the C-suite can be done on Data Warehouse infrastructure, while R&D needs all (unstructured) data to create and improve new products
- ”Don’t Think ’What happened’ Think, ’What Will Happen’”
  - To gain true business value, you need to go beyond ”what happened” (and ”What will happen”) to ”What should I do!” - see Table 1-2. page 13.
- ”Don’t Think HIPPO, Think Collaboration”
  - You need to be willing to act on the insights gained by the data analyses

Descriptive, Predictive, and Prescriptive analytics

Table 1-2: Evolution of the Business Questions

WHAT HAPPENED? (DESCRIPTIVE/BI)	WHAT WILL HAPPEN? (PREDICTIVE ANALYTICS)	WHAT SHOULD I DO? (PRESCRIPTIVE ANALYTICS)
How many widgets did I sell last month?	How many widgets will I sell next month?	Order [5,0000] units of Component Z to support widget sales for next month
What were sales by zip code for Christmas last year?	What will be sales by zip code over this Christmas season?	Hire [Y] new sales reps by these zip codes to handle projected Christmas sales
How many of Product X were returned last month?	How many of Product X will be returned next month?	Set aside [$125K] in financial reserve to cover Product X returns
What were company revenues and profits for the past quarter?	What are projected company revenues and profits for next quarter?	Sell the following product mix to achieve quarterly revenue and margin goals
How many employees did I hire last year?	How many employees will I need to hire next year?	Increase hiring pipeline by 35 percent to achieve hiring goals

The Big Data Business Model Maturity Index

- Do I really need to transform my business models as a consequence of the Big Data era?
- Where to begin and where can I end up?
- "How effective is my organization at integrating data and analysis into our business models?"
- The Big Data Business Model Maturity Index can be used to measure how far organizations are in their adoption of Big Data
- Critique
  - Do all businesses need to strive after achieving the top level?
  - There is also a dimension of automation that is not mentioned

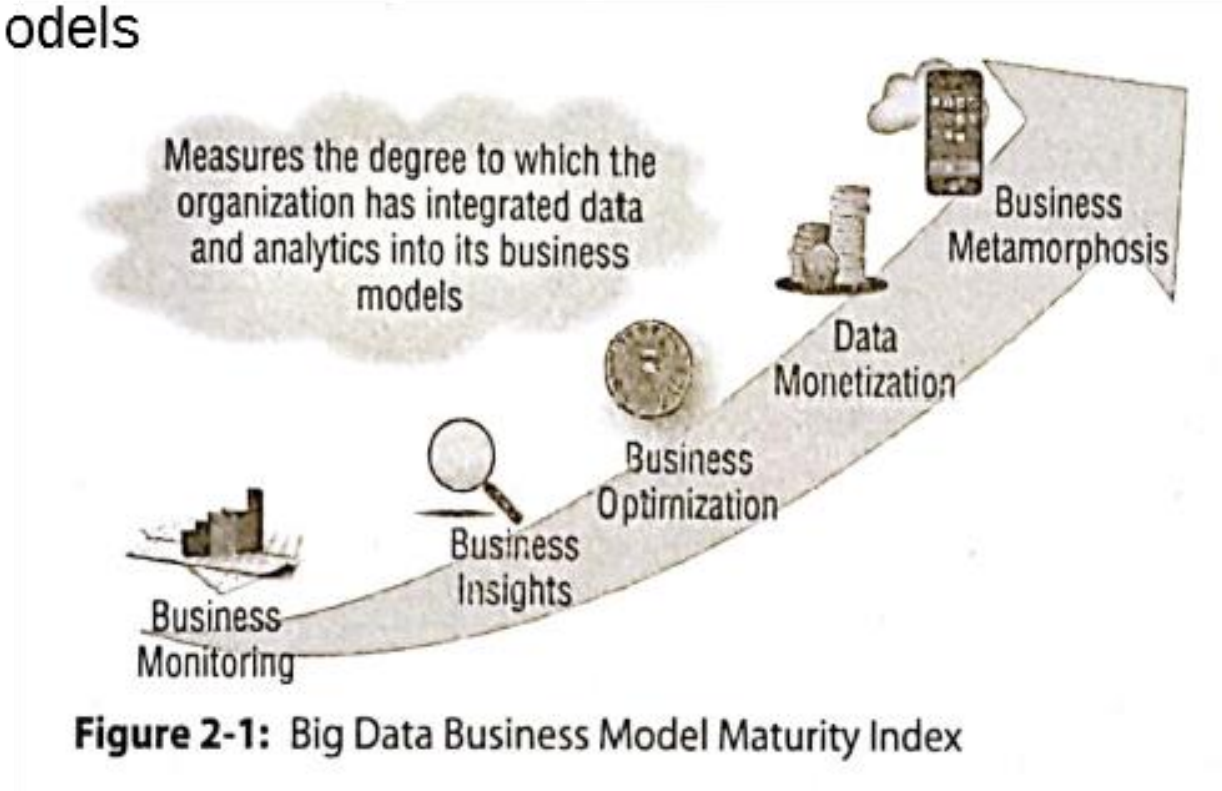


Figure 2-1: Big Data Business Model Maturity Index

Phase 1: Business Monitoring

- Monitoring ongoing business performance
- Use reports and Dashboards
- Document key business processes and indicators
- Note: It is not completely trivial to achieve this state and once it is achieved it can provide great value
- Moving beyond this state might require a complete rework of technologies and data pipelines

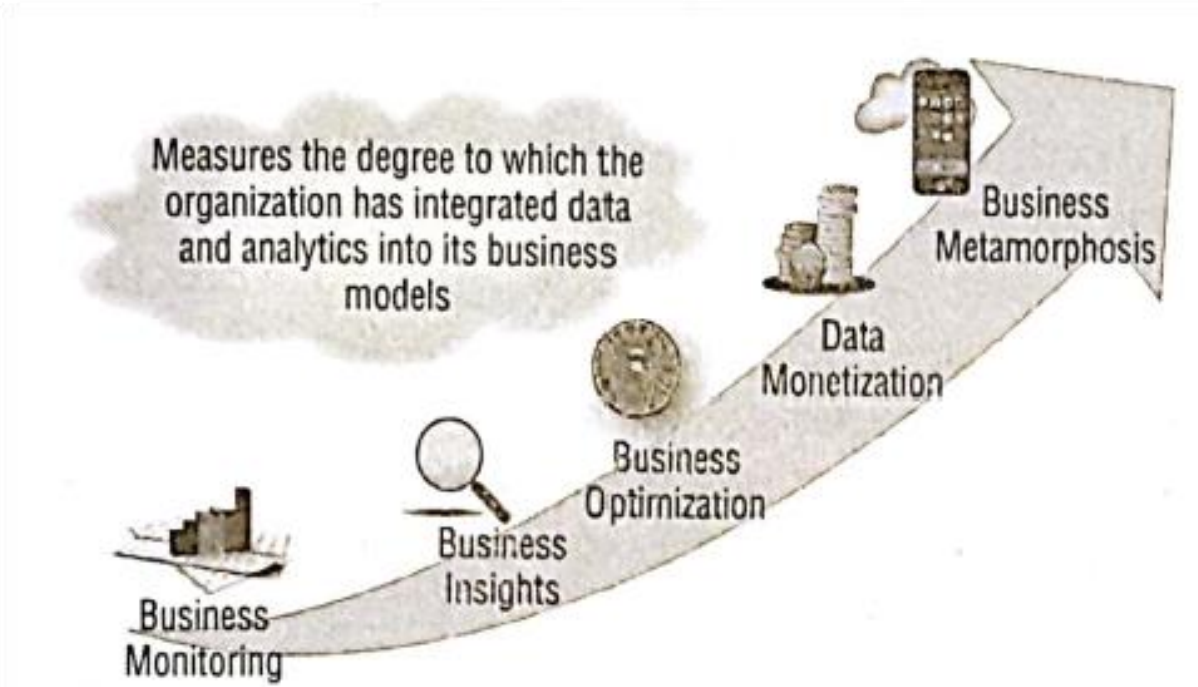


Figure 2-1: Big Data Business Model Maturity Index

Phase 2: Business Insights

- Deriving insights from the data
- Break down data silos
- Work with transaction level and individual (customer) level data instead of aggregated data
- Use unstructured data such as social media data, emails, customer service calls etc.
- Exploiting real-time analysis
- Using predictive analytics

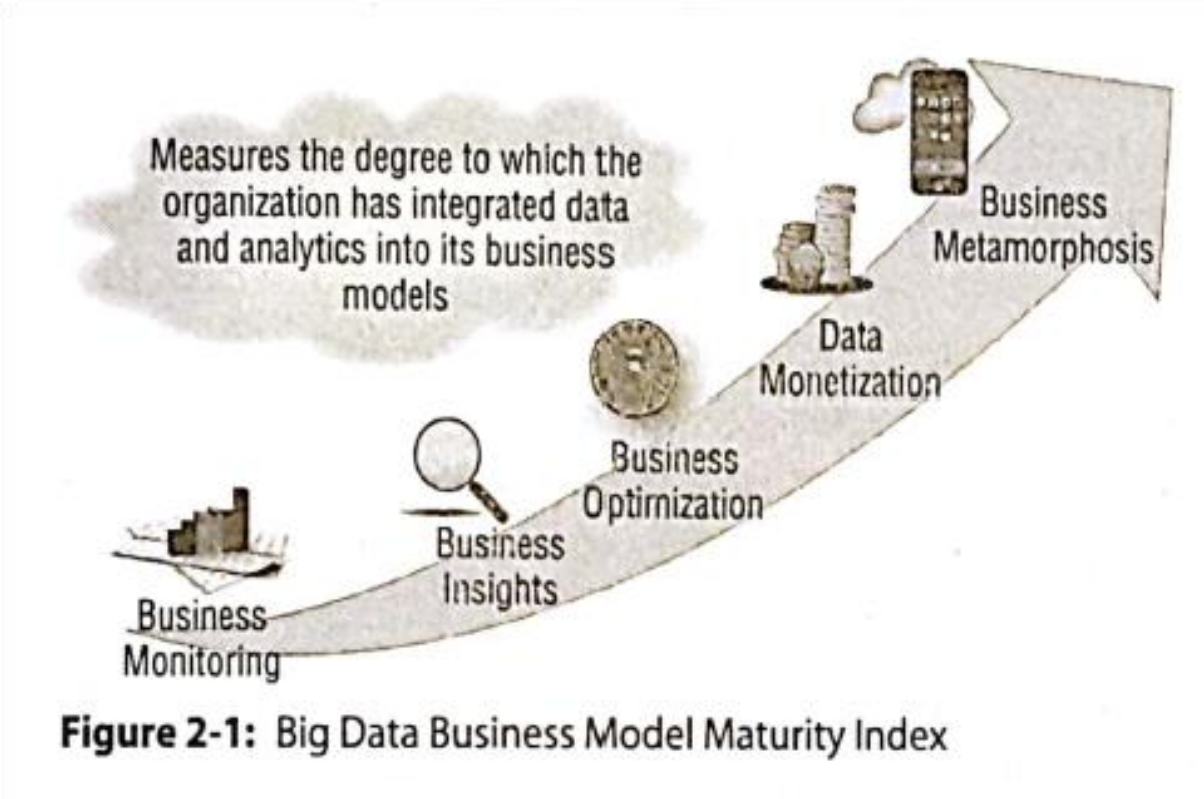


Figure 2-1: Big Data Business Model Maturity Index

Phase 3: Business Optimization

- Predictive and prescriptive analysis is used to optimize business decisions
- Examples
  - Deliver distribution and inventory recommendations given predicted sales and local traffic, weather and demographic data
  - Deliver pricing recommendations based on competitors pricing and demands etc.
  - Deliver media investment plan for marketing campaigns
  - Deliver expected number of returned goods
  - Recommend to customers what to buy

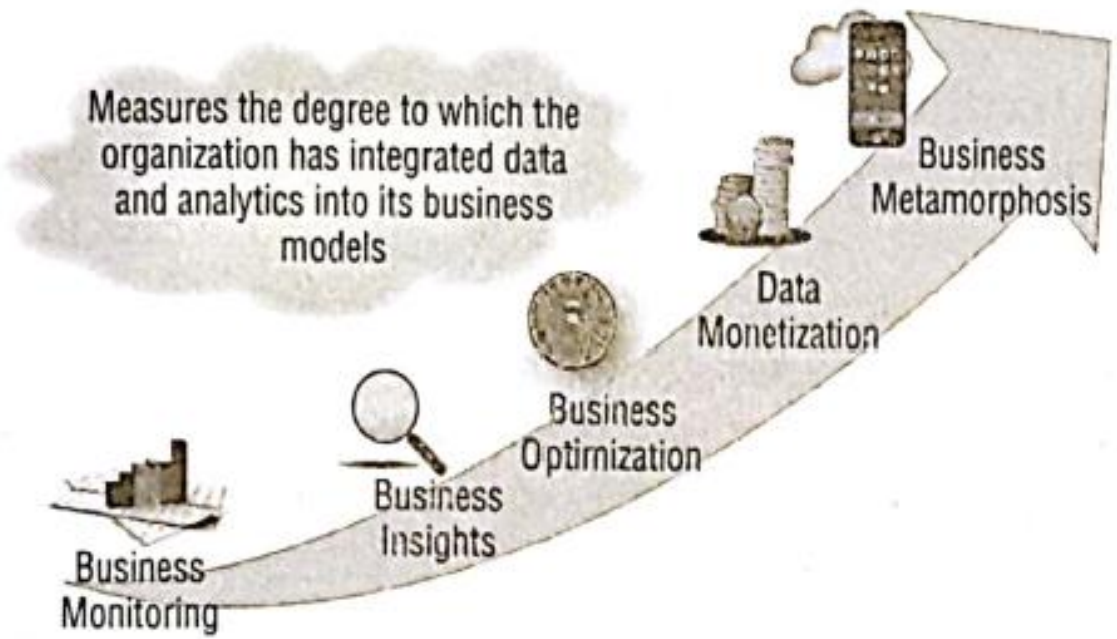


Figure 2-1: Big Data Business Model Maturity Index

Phase 4: Data Monetization

- Selling on collected data to third party
- Customizing products
- Make money out of the data
- Example:
  - How facebook profit from its users data – it is not the individual profile users that are the true customer of facebook, they are the product!

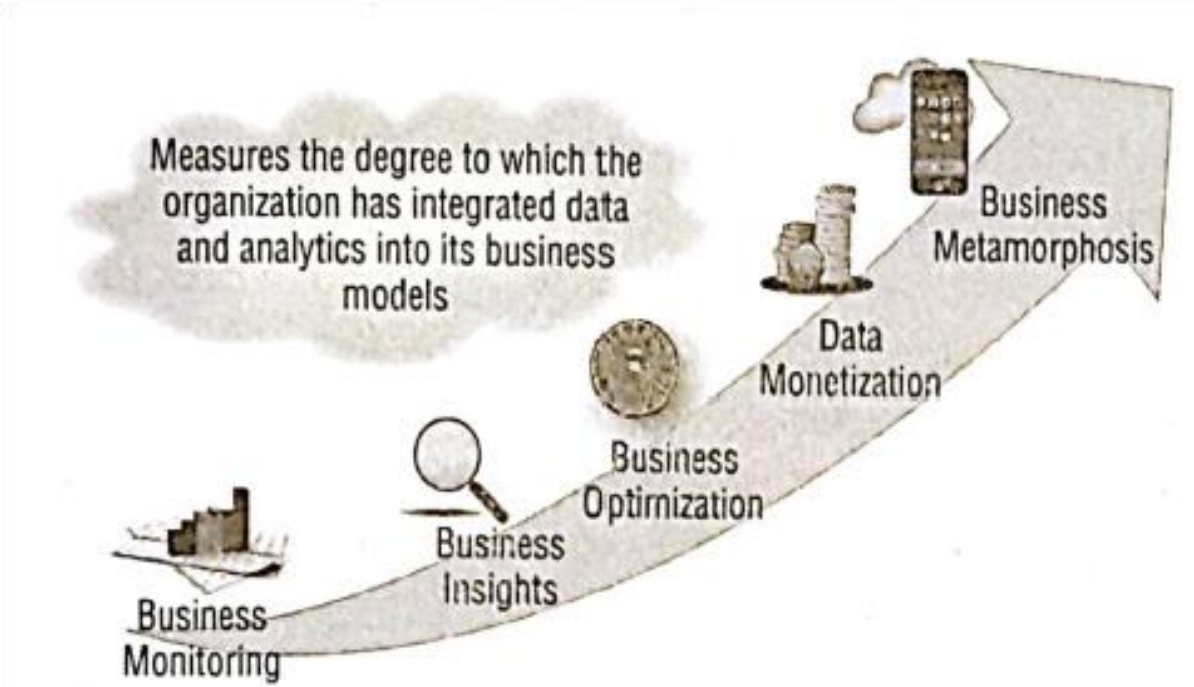


Figure 2-1: Big Data Business Model Maturity Index

Phase 4: Data Monetization

- Selling on collected data to third party
- Customizing products
- Make money out of the data
- Example:
  - How facebook profit from its users data – it is not the individual profile users that are the true customer of facebook, they are the product!

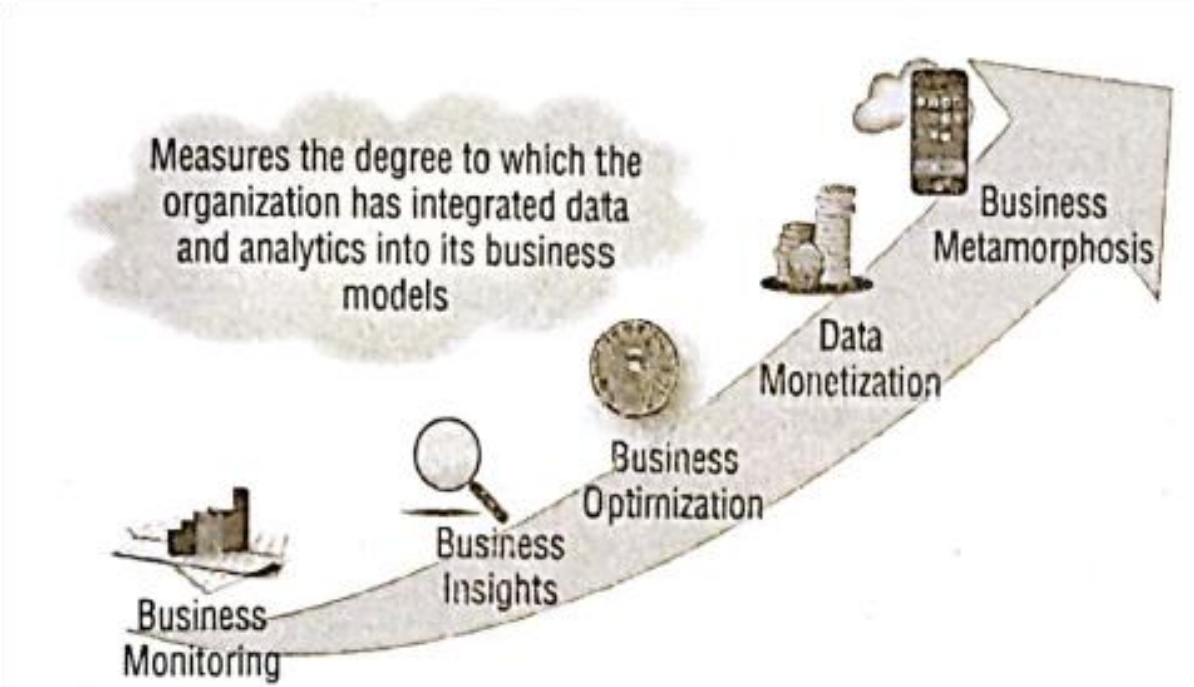


Figure 2-1: Big Data Business Model Maturity Index

How to find sample data

Some places to find sample data

- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
- OPEN DATA DK: http://www.opendata.dk/
- Statistics Denmark: https://www.dst.dk/en#
- The Municipality of Copenhagen: https://data.kk.dk/
- Google Trends: https://trends.google.com/
- Frequent Itemset Mining Dataset Repository: http://fimi.ua.ac.be/data/
- The U.S. Government's open data: https://www.data.gov/
- Kaggle: https://www.kaggle.com/
- Azure AI Gallery: https://gallery.azure.ai/
- Google Analytics (Google Merchandise Store): https://analytics.google.com
- Google launch search engine for scientist to find data: https://www.theverge.com/2018/9/5/17822562/google-dataset-search-service-scholar-scientific-journal-open-data-access
- Find some sample or real data yourself…

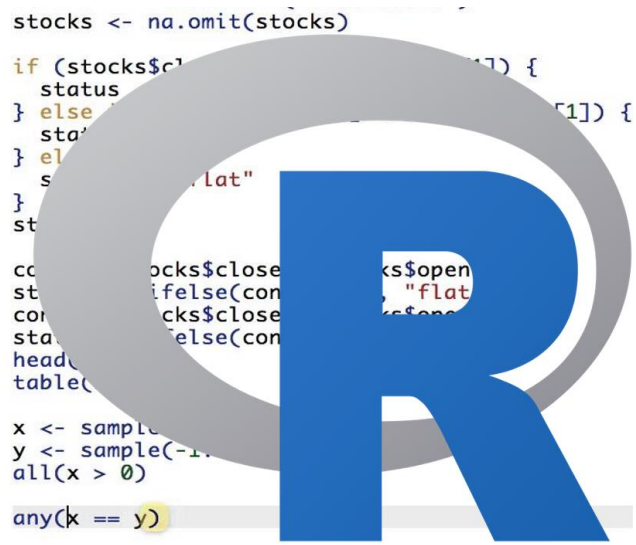
Exercise

- Try and find some data that can be used for you business case

Introduction to R

Why use R?

- It is developed for statistical purposes
  - Ideal for data manipulation and data analysis
- It is open source
  - It is free
- It has a large and active community
  - There are packages for all the newest algorithms and techniques before most commercial software
- Together with Python it is one of the most used languages within Data Science
  - There is a lot of online courses etc.
- It has useful extensions and IDEs
  - RStudio, Shiny, RMarkdown, etc.



Jupyter notebook and Azure Notebooks

- We will use R through Jupyter notebooks which we will run on Azure Notebooks:
  - https://notebooks.azure.com/
  - Sign into Azure Notebooks (create a Microsoft account if needed)
  - Clone my Library for the course: https://notebooks.azure.com/jensuh/libraries/BIBA-2018
  - Open the notebook *"My first R notebook.ipynb"*