

# Introduction to Business Intelligence

Tatiana Zolotareva

March 20, 2019

# Origin of the Term “Business Intelligence”

**Business:** “A collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, etc.”

**Intelligence:** “The ability to apprehend the interrelationships of presented facts in such way as to guide action towards a desired goal ...”

(Webster’s Dictionary)

# Notions associated with the term Business

- ▶ **Size of business:** *size of the enterprise*
- ▶ **Scope of business:** *complexity of activities*
- ▶ **Business strategy:** *how the organization intends to succeed*
- ▶ **Business model:** *the strategy to create value*
- ▶ **Business process:** *A collection of related and structured activities for delivering a certain good or service to customers together with possible response activities from customers*
- ▶ **Business process instance:** *An observable realization of business process*

# Origin of the Term “Business Intelligence”

1958: H. P. Luhn (IBM): “A Business Intelligence System”

Challenges:

- ▶ “Information generated and utilized at an ever-increasing rate”
- ▶ “Growing need for more prompt decisions at the level of responsibility”
- ▶ “Applying machines to process information retrieval”

# Origin of the Term “Business Intelligence”

1958: H. P. Luhn (IBM): “A Business Intelligence System”

Techniques:

- ▶ **Document input:** photograph a document to microfilm, transcript the microfilm to magnetic tape
- ▶ **Auto-abstracting:** the document: statistical analysis of the text based on the word frequency and distribution
- ▶ **Auto-encoding** the document: creating document profiles or deriving information patterns that can characterize a document

# BI in Relation to OR

1920s-... **Operational research** as a research area: wide range of problem solving techniques and methods applied in pursuit of improved decision making and efficiency (Wikipedia).

# BI in Relation to OR

- ▶ **Decision theory** (sub-branch of game theory): defining a strategy that maximizes a utility function under uncertainty

- ▶ Should I bring the umbrella today?

dry/wet cloths:  $\pm 5$ , light/heavy suitcase:  $\pm 3$ .

	rain ( $p=0.5$ )	no rain ( $p=0.5$ )
umbrella	dry cloths, heavy suitcase	dry cloths, heavy suitcase
no umbrella	wet cloths, light suitcase	dry cloths, light suitcase

- ▶ I am looking to buy a house. Should I buy this one? (I can find a better house if I keep searching, but searching costs time)
- ▶ Am I going to smoke another cigarette? One cigarette is no problem, but if I am going to make the decision sufficiently many times it may kill me.

# BI in Relation to OR

- ▶ **Project planning:** designing models that comprise all different paths along which a business process can evolve with corresponding cost and utility
- ▶ **Supply-chain management:** minimizing costs with respect to conflicts between different parts of the chain (sales department desiring to have higher inventory level to fulfill the demand and warehouse department desiring to have lower inventory to reduce the holding costs)
- ▶ **Assignment** problem (discrete optimization). Example: taxi-cabs at different locations and customers who want to be picked up as soon as possible.



# BI in Relation to Data Mining

Change in focus due to the availability of the data

1989: H. Dresner (Gartner Group): *“An umbrella term to describe concepts and methods to improve business decision making by using fact-based support systems.”*

2004 Negash: *“Business Intelligence systems combine operational data with analytical tools to represent complex and competitive information to planners and decision makers. The objective is to improve the timeliness and quality of inputs to the decision process.”*

# BI in Relation to Data Mining

1960s: DBMS, CODASYL (“Database Task Group”), IMS (IBM): direct access storage in contrast with the tape based storage. Data is stored in files in a navigational form: objects are found by following references from other objects:

```
get department with name='Sales'  
get first employee in set department-employees  
until end-of-set do  
    get next employee in set department-employees  
    process employee
```

1970, Edgar Codd: RDMS: data is stored in tables in an Entity-Relationship form.

# BI in Relation to Data Mining

Central repositories of data integrated from various sources/systems and organized under a unified view that supports information retrieval.

1970s: Express (product with OLAP functionality)

1980s: IBM “Business Data Warehouse”

1993: E. Code OLAP

# BI in Relation to Data Mining

- ▶ Staging layer: stores raw data extracted from each source
- ▶ Integration layer: integrates data from different sources (transforms and cleans it if needed) and stores the result in an “operational database”
- ▶ Warehouse database layer: arranging data into hierarchical groups (dimensions) and into facts and aggregated facts.

**facts:** *number of products ordered, total price payed;*

**dimensions:** *order date, customer name, product name*

facts + dimensions = star schema

# BI in Relation to Statistical Learning

- ▶ 1805, A.-M. Legendre: Least Squares Linear Regression
- ▶ 1812, P.-S. Laplace: Bayes Theorem
- ▶ 1950, A. Turing: Turing's Learning Machine
- ▶ 1957, F. Rosenblatt: Perceptron
- ▶ 1963, D. Michie: Reinforcement Learning in tic-tac-toe
- ▶ 1967, KNN
- ▶ 1970, S. Linnainmaa: Backpropagation
- ▶ 1982, J. Hopfield: RNN
- ▶ 1984, L. Breiman: Decision Trees
- ▶ 1989, C. Watkins: Q-Learning
- ▶ 1995, C. Cortes, V. Vapnik: Support Vector Machines
- ▶ 1997, IBM Deep Blue beats Kasparov
- ▶ 1997, S. Hochreiter, J. Schmidhuber: LSTM
- ▶ ...

# Decision Support Systems

1970s-....: DSS as area of research:

Interactive computer-based systems which help decision-makers utilize databases and models to solve ill-structured problems

1987, Texas Instruments: DSS tool for United Airlines

- ▶ *Passive*: provide information that aids in decision making (reporting), view on business defined from the data
- ▶ *Active*: suggest decisions
- ▶ *Collaborative*: decision-makers can refine solutions proposed by the system, the system can refine solutions proposed by the decision-maker

# Passive DSS

What kind of information should a passive DSS provide?

# Passive DSS

What kind of information should a passive DSS provide?

- ▶ Expressive **data model** (ER-model, UML, process model)
- ▶ **Conformance** of observations with the data model/**anomalies**
- ▶ View at data on different **aggregation** levels (data warehousing model)
- ▶ **Data abstraction**: matching data with domain knowledge (temperature 39 ° is high, paracetamol is a drug)



# Passive DSS

What kind of information should a passive DSS provide?

- ▶ Evaluation of the past or prediction of the future **KPIs** (Key Performance Indicators)
- ▶ Analysis of **influential factors**: parameters defined from the data that have an influence on KPIs of interest
- ▶ **Data segmentation** (clustering and community detection): *customer profiles, “basket analysis”, sensor measurements*
- ▶ **Reporting, visualization**

# Active DSS

How can an active DSS suggest decisions?

# Active DSS

How can an active DSS suggest decisions?

- ▶ **Logical rules** that map the values of KPIs, or outputs of analytical models to decisions
- ▶ Optionally: **automated execution** of the decisions

*Examples:*

- ▶ *Automated trading*
  - ▶ *Automated scheduling of machine maintenance*
  - ▶ *Automated inventory reordering*
- 
- ▶ **Reinforcement learning**
  - ▶ Could be used as a **black box** if the output of an analytical model belongs directly to the set of possible decisions

# Decision Support Systems

What is the idea behind collaborative DSS?

# Decision Support Systems

What is the idea behind collaborative DSS?

- ▶ **Ontology**: system of entities and relationships between them
- ▶ Users can define and maintain rules themselves
- ▶ Ability to explain how the system came to a conclusion
- ▶ **Reasoner**: ability to deduce new rules from the data, logical formalization that support some form of inference.

*$pizza(x) \Rightarrow Italian\ dish(x)$ . If we have a new entry “Margarita” and we know that it is a pizza, the system should be able to deduce that it is an Italian dish. Or if we want to know if “Margarita” is an Italian dish, the system should be able to look through all the rules that would allow to deduce if it is a pizza.*

# Decision Support Systems

What tools do we need to design a DSS?

# Decision Support Systems

What tools do we need to design a DSS?

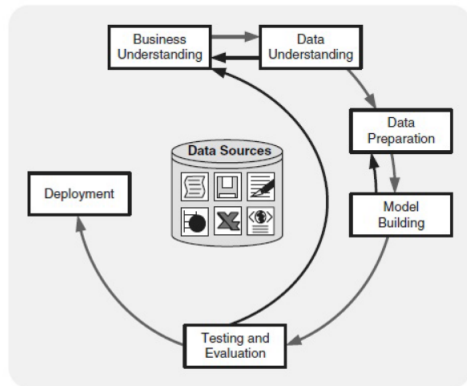
- ▶ Data model configuration (*schema, diagram, graph*)
- ▶ Database (*MySQL, MongoDB, GraphDB, XML*)
- ▶ Query language (*SQL, JavaScript, GraphQL, XQuery*)
- ▶ Programming language/software to implement analytical models that provide answers to business questions
- ▶ Logical language to describe/infer relationships in the data or/and map data properties to business decisions
- ▶ Programming language/software for reporting, visualization, and user interface

## Related topics

- ▶ **Business Analytics:** finding new insights and understanding of the business
- ▶ **Predictive Analytics:** prediction of future business performance/events
- ▶ **Data Mining:** extracting information of interest from large data sets
- ▶ **Process Mining:** finding structure in instances of business processes
- ▶ **Data Warehousing**
- ▶ **Machine Learning:** computer programs with the ability to learn how to solve a task

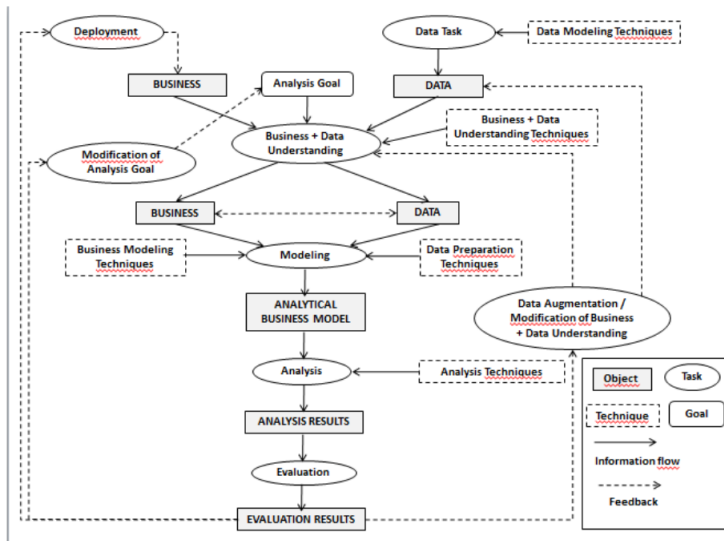


# Cross Industry Standard Process for Data Mining

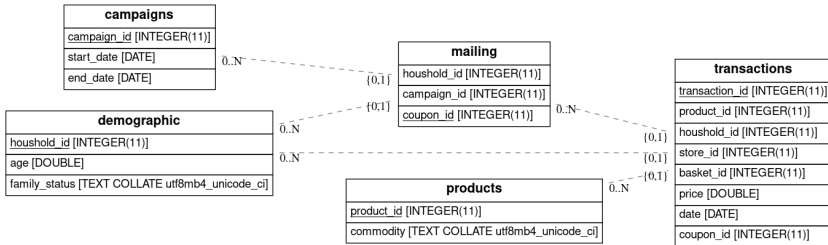


# Analysis formats

iMine  
Format  
Combining  
CRISP  
and L\*



# Example: supermarket data



# Analytical Tasks

- ▶ Business/Data/Process understanding task
- ▶ Descriptive task: formulation of KPIs, segmentation (detection of clusters, communities, profiles)
- ▶ Predictive/modeling task: formulating the predictive goals, implementation of the models
- ▶ Evaluation and report task: present the results of the analysis in the context of the business

# Business and Data Understanding Task

- ▶ Explore **application environment**: *size and scope of the business, project strategy, resources and time horizon of the project*
- ▶ Which **business perspective** is of main interest?
  - ▶ **Production perspective**: *What should be offered to customers? How should the offer be produced?*
  - ▶ **Customer perspective**: *How do the customers react to the offers?*
  - ▶ **Organizational perspective**: *What organizational structure is behind production?*

# Business and Data Understanding Task

- ▶ **Production perspective:** What products should be sold in the store? What is the sales dynamics for each kind of product? What products should be promoted? Are promoted products sold better than the ones that were not promoted?
- ▶ **Customer perspective:** What customers should be mailed with coupons? How fast do the customers react to the campaigns? How often do the customers use coupons? Do customers that use coupons have bigger baskets? What are the customer profiles that react to campaigns the best?
- ▶ **Organizational perspective:** How often should be campaigns scheduled? How long should the campaigns last? How many coupons should we send for each product?

# Process Understanding Task

- ▶ **Identify (formalize) process:** *finding rules that determine the relationships between the events of the process*
- ▶ **Process analysis:** *investigate conformance of the process instances with a defined process*
- ▶ Identify **involved parties:**
  - ▶ **Process owner:** *responsible for setting up the rules behind the process*
  - ▶ **Process subjects:** *identifiers for process instances*
  - ▶ **Process actors:** *people or organizational units that participate in the process*

# Process Understanding Task

- ▶ **Process owner:** supermarket
- ▶ **Process subjects:** customers
- ▶ **Process actors:** customers, campaign-department of the supermarket

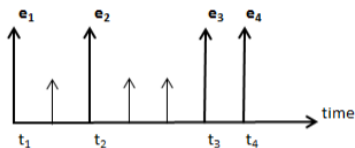


# Process Understanding Task

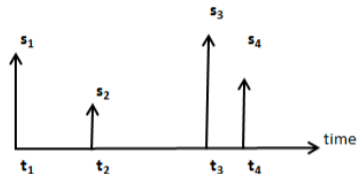
The business process can be viewed from different angles:

- ▶ Event view: an event is characterized by its start time, end time, and possibly interruption times
- ▶ State view: in connection with events attributes are measured that characterize the state of the process at a given time
- ▶ Cross-sectional view: look at the history of many process instances at a given time

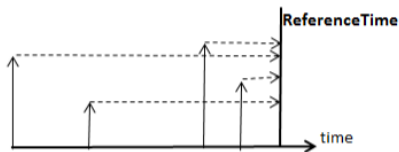
# Process Understanding Task



a) Foreground process, event view



b) Foreground process, state view



c) Foreground process, cross-sectional view

## Descriptive Task: KPIs

Average basket value for people at the age between 35 and 45 years old per store.

## Descriptive Task: KPIs

Average basket value for people at the age between 35 and 45 years old per store.

SQL:

```
SELECT t.store_id,  
       SUM(t.price)/COUNT(DISTINCT t.basket_id)  
       as average_basket_value  
FROM transactions as t  
LEFT JOIN demographic as d  
ON t.household_id = d.household_id  
WHERE d.age >= 35 AND d.age <= 45  
GROUP BY t.store_id;
```

# Descriptive Task: KPIs

Average basket value for people at the age between 35 and 45 years old per store.

pandas:

```
transactions.merge(demographic, how="left", on="houshold_id")\
    .query("age >= 35 and age<=45")\
    .groupby(["store_id", "basket_id"])["price"].sum()\
    .reset_index()\
    .groupby("store_id")["price"].mean()\
    .reset_index()\
    .rename(columns={"price": "avarage_basket_value"})
```

## Examples KPIs

Ratio of fruit-product purchases (commodity fruits) with a coupon redeemed.

## Examples KPIs

Ratio of fruit-product purchases (commodity fruits) with a coupon redeemed.

SQL:

```
SELECT AVG(CASE WHEN t.coupon_id IS NOT NULL
                  THEN 1 ELSE 0 END)
FROM transactions as t
LEFT JOIN products as p
ON t.product_id = p.product_id
WHERE p.commodity="fruits";
```

# Descriptive Task: KPIs

Ratio of fruit-products (commodity fruits) that were bought with a coupon.

pandas

```
transactions.merge(products, how="left", on="product_id")\
    .query("commodity=='fruits'")\
    .assign(coupon_used=lambda x: x["coupon_id"].notnull())\
    ["coupon_used"].mean()
```



## Descriptive Task: KPIs

Average number of customers per campaign that were mailed but did not use the coupon.

## Descriptive Task: KPIs

Average number of customers per campaign that were mailed but did not use the coupon.

```
WITH counts AS (  
    SELECT COUNT(DISTINCT t.household_id) as n, m.campaign_id  
    FROM transactions as t  
    LEFT JOIN mailing as m  
    ON t.household_id = m.household_id  
    WHERE t.coupon_id IS NULL  
    AND m.coupon_id IS NOT NULL  
    GROUP BY m.campaign_id)  
SELECT AVG(n) FROM counts;
```

# Descriptive Task

Average number of customers per campaign that were mailed but did not use the coupon.

pandas:

```
transactions.merge(mailing, how="left", on="houshold_id")\
    .query("coupon_id_x.isnull() and coupon_id_y.notnull()")\
    .groupby("campaign_id")["houshold_id"].nunique()\
    .mean()
```

# Predictive Task

# Predictive Task

- ▶ Predict performance of future campaigns
- ▶ Predict the improvement in sales for a product after a campaign
- ▶ Predict how well a given customer will participate in a campaign

# Modeling Task

**Model:** formalized representation of some part of the business process allowing precise formulation of the questions of interest.

A model can be decomposed into basic elements and rules how to compose them.

Semantic defines the meaning of the elements in a language independent from the domain.

Generic questions: questions formulated in the semantic of the model language about the properties of the model.

# Modeling Task

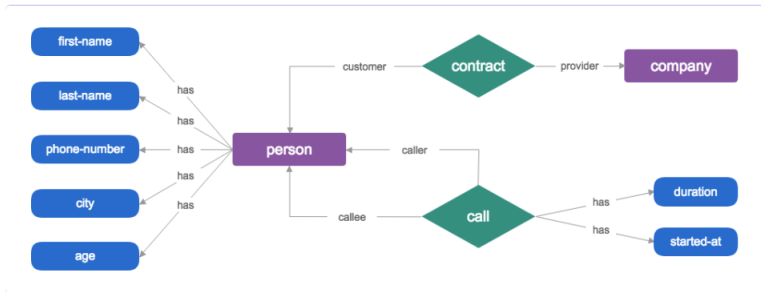
- ▶ **ER models:** model elements are entities are relationships represented by tables; generic questions are formulated in a query language.
- ▶ **Logical models:** (ontologies) model elements are entities and relationships between them, generic questions are formulated in a logical language (OWL or other)
- ▶ **Mathematical models:** elements of the model and relationships between them are formulated via structures like vector spaces (Euclidean space), functional spaces, probability spaces, fields, algebras
- ▶ **Graph models:** elements and relationships are represented by nodes, edges, and weights. Generic questions: shortest path, best matching nodes, etc...
- ▶ **Statistical models:** random variables and mathematical relationships between them

# Modeling Example: GRAKN.AI

- ▶ Knowledge graph engine to organize complex networks
- ▶ ER model: has entities, relationships, and attributes described in the graql language
- ▶ Model configuration happens in a schema
- ▶ Implements a reasoner: new rules about the data can be inferred



# Defining the ontologie



# Modeling Example: GRAKN.AI

Configuring the schema: *start the file with the keyword define*

## **define new attribute:**

```
new_attribute sub attribute,  
    datatype string;
```

## **define new relationship:**

```
new_relationship sub relationship,  
    relates role1,  
    relates role2,  
    has attribute1;
```

## **define a new entity:**

```
new_entity sub entity,  
    plays role1,  
    plays role2,  
    has attribute1,  
    has attribute2;
```

# Modeling Example: GRAKIN.AI

**Load the schema:**

**start (stop) server:** `/.grakn server start (stop)`

**load the schema:** `./grql console --keyspace phone_calls  
--file path/to/the/schema.gql`

**open the console in an interactive mode:** `./grql console  
--keyspace phone_calls`

**make sure that the schema is properly defined:** `match $x  
sub thing; get;`

## Modeling Example: GRAKIN.AI

**Business value:** The person with phone number +86 921 547 9004 has been identified as a lead. We (company “Telecom”) would like to know which of our customers have been in contact with this person since September 14th. This helps us in converting this lead into a customer.

### query:

```
match
$customer isa person, has phone-number $phone-number;
$company isa company, has name "Telecom";
(customer: $customer, provider: $company) isa contract;
$target isa person, has phone-number "+86 921 547 9004";
(caller: $customer, callee: $target) isa call,
  has started-at $started-at;
$min-date == 2018-09-14T17:18:49; $started-at > $min-date;
get $phone-number;
```

# Modeling Example: GRAKIN.AI

**Business value:** The customer with phone number +7 171 898 0853 and +370 351 224 5176 have been identified as friends. We (company “Telecom”) like to know who their common contacts are in order to offer them a group promotion.

# Modeling Example: GRAKIN.AI

**Business value:** The customer with phone number +7 171 898 0853 and +370 351 224 5176 have been identified as friends. We (company “Telecom”) like to know who their common contacts are in order to offer them a group promotion.

To calculate count, mean, max, min, median, sum of a set of values, use pattern:

```
match
$sch isa school, has ranking $ran;
aggregate max $ran;
```

# Modeling Example: GRAKIN.AI

Rules: Grakn uses rule-based reasoning to perform reasoning over data and to dynamically create relationships.

## **syntax as part of the schema::**

```
rule-label sub rule,  
when {  
    ##  
    ## conditions go here  
    ##  
},  
then {  
    ## concluded fact  
};
```

# Modeling Example: GRAKIN.AI

## Example rule:

```
people-with-same-parents-are-siblings sub rule,  
when { (mother: $m, $x) isa parentship;  
  (mother: $m, $y) isa parentship;  
  (father: $f, $x) isa parentship;  
  (father: $f, $y) isa parentship;  
  $x != $y;  
}, then {  
  ($x, $y) isa siblings;  
};
```



# Modeling Example: GRAKIN.AI









## Define relationships inheriting from existing relationships:

```
location sub relationship is abstract,  
  relates located-subject,  
  relates subject-location;
```

```
location-of-birth sub location-of-everything,  
  relates located-birth as located-subject,  
  relates birth-location as subject-location;
```

**Exercise:** Add a new relationship “long-distance-call” to the schema that inherits from the relationship call. Add a rule: if the city of the caller is different from the city of the callee, then the call is a long distance call.

# References

-  Fundamentals of Business Intelligence - W. Grossmann, S. Rinderle-Ma - 2015
-  Intelligence I slides - W. Grossmann, S. Rinderle-Ma - 2018
-  Decision Theory – A Brief Introduction - S. O. Hansson - 1994
-  A Business Intelligence System - H.P. Luhn - 1958
-  <https://dev.grakn.ai/docs/examples/>
-  <https://www.dunnhumby.com/careers/engineering/sourcefiles>
-  [https://en.wikipedia.org/wiki/Timeline\\_of\\_machine\\_learning](https://en.wikipedia.org/wiki/Timeline_of_machine_learning)
-  [https://en.wikipedia.org/wiki/Decision\\_support\\_system](https://en.wikipedia.org/wiki/Decision_support_system)