Plateforme décisionnelle SAS

Grégoire de Lassence

- **Grégoire de Lassence**
  Responsable Pédagogie et Recherche
  Département Académique

  Tel : +33 1 60 62 12 19

  gregoire.delassence@fra.sas.com
  http://www.sas.com/france/academic

SAS dans le monde

1976 : Création en Caroline du Nord
Société privée
CA 2010 : 2.43 milliards $
11 000 employés
24 % du CA réinvesti en R&D

SAS en France

280 collaborateurs

Services SAS Academic

academic@fra.sas.com
http://www.sas.com/offices/europe/france/academic/index.html

- Cours
  - Développement de programmes, études de cas, support pédagogique
  - Experts SAS,
  - e-learning,
  - Certification SAS

- Club SAS Academic
  - SAS pendant le stage : « CPPS »
  - Licence Gratuite à Domicile
  - Offres de stage et d'embauche de nos clients

- Divers
  - Newsletter Internationale
  - Student Ambassador Competition / Papiers SFF
  - Recherche & Chaires
  - Evènements & Sponsoring
  - …

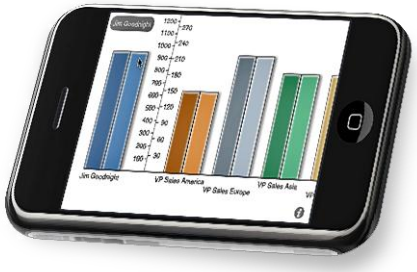
Plateforme décisionnelle

Quel projet décisionnel ?

Descriptif

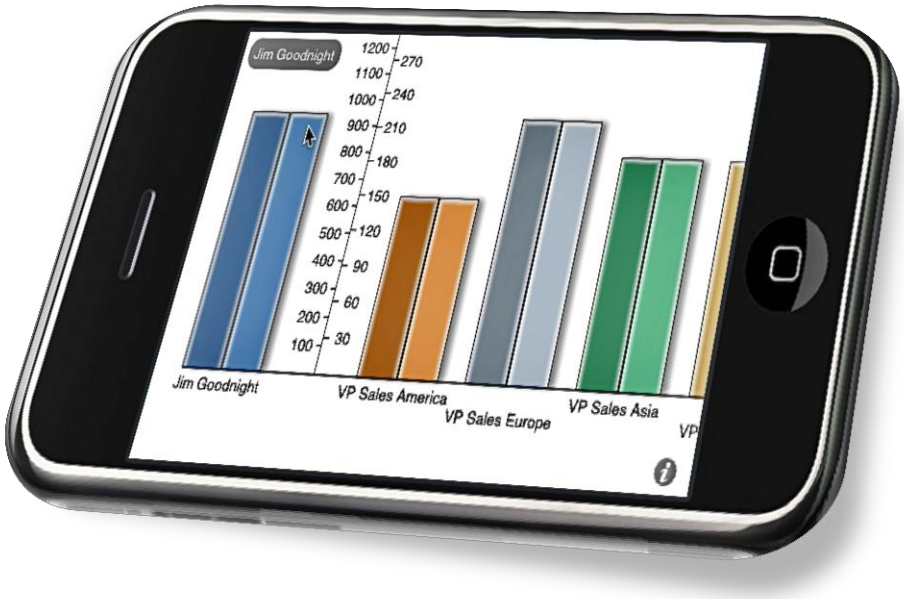
Business Intelligence ?

SAS® Mobile Business Intelligence



SAS® Mobile Business Analytics

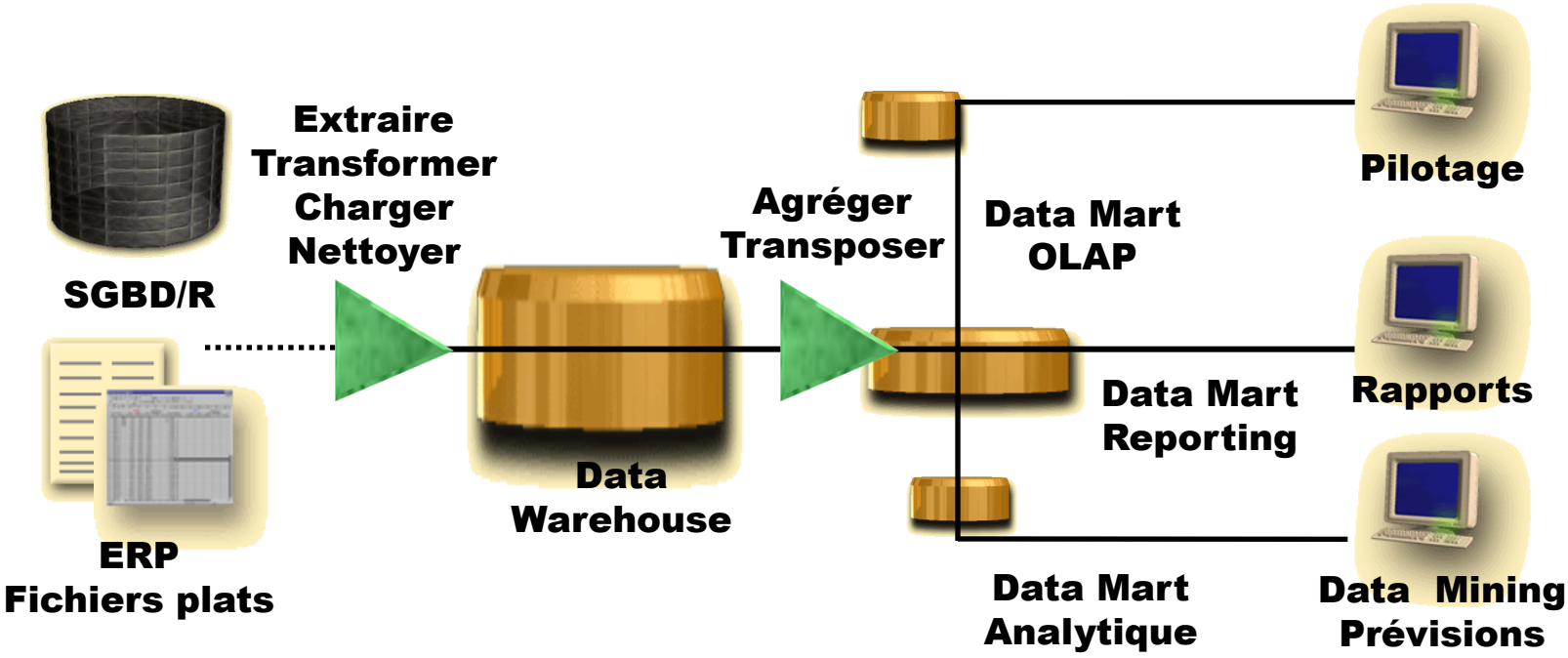
iPhone

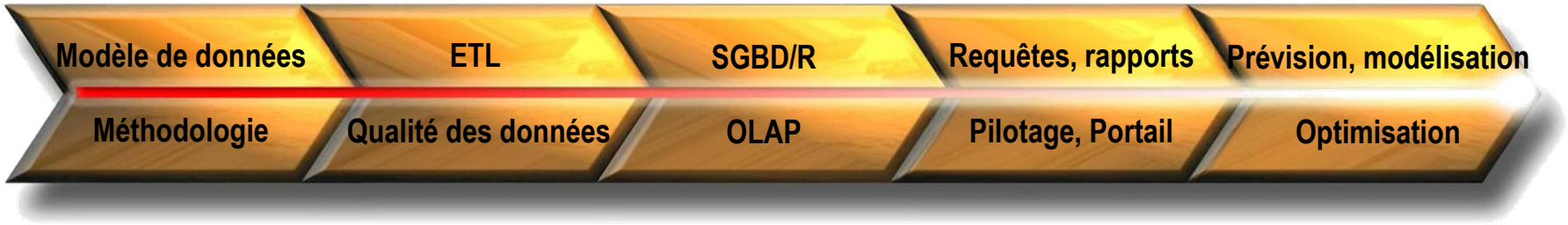



Prédictif
Analytique

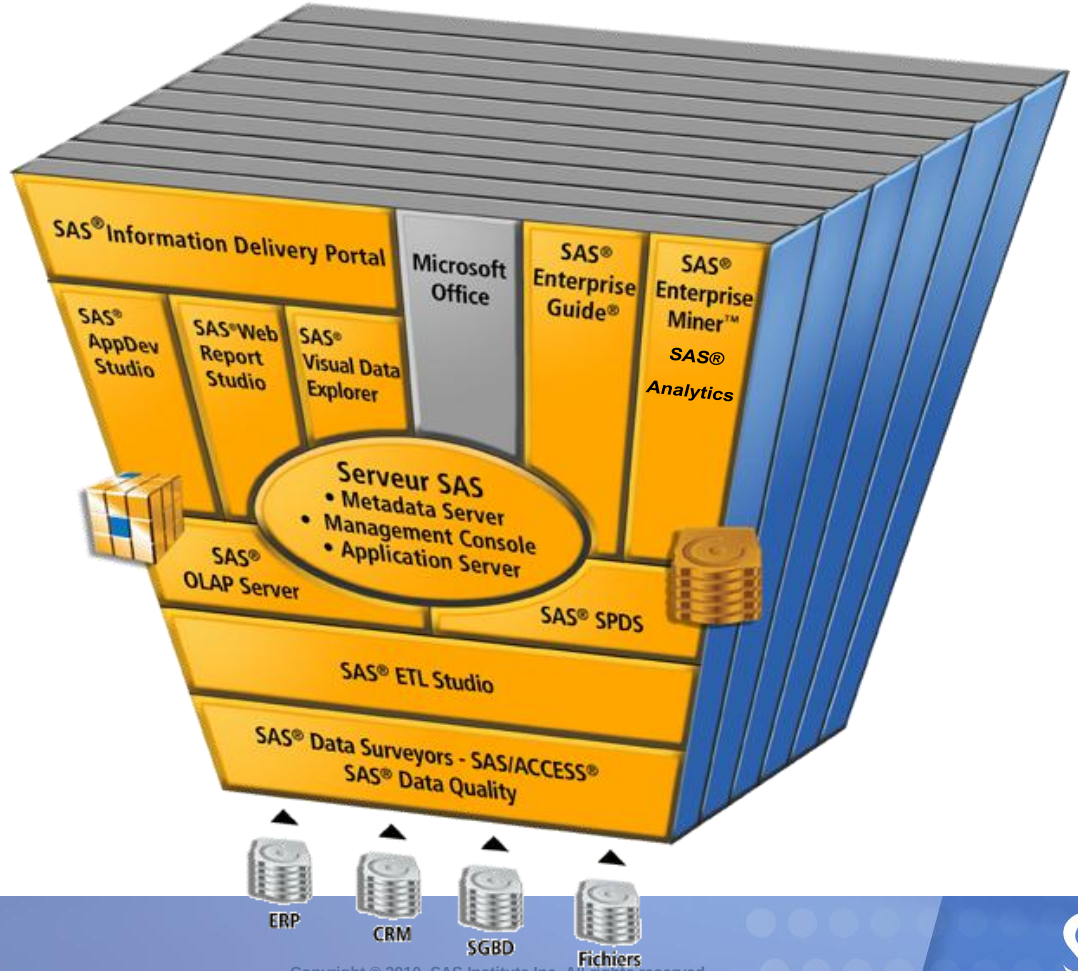

Plan
ETL^Q
Intelligent Storage
Business Intelligence
Analytic Intelligence

Maîtriser la chaîne « décisionnelle »



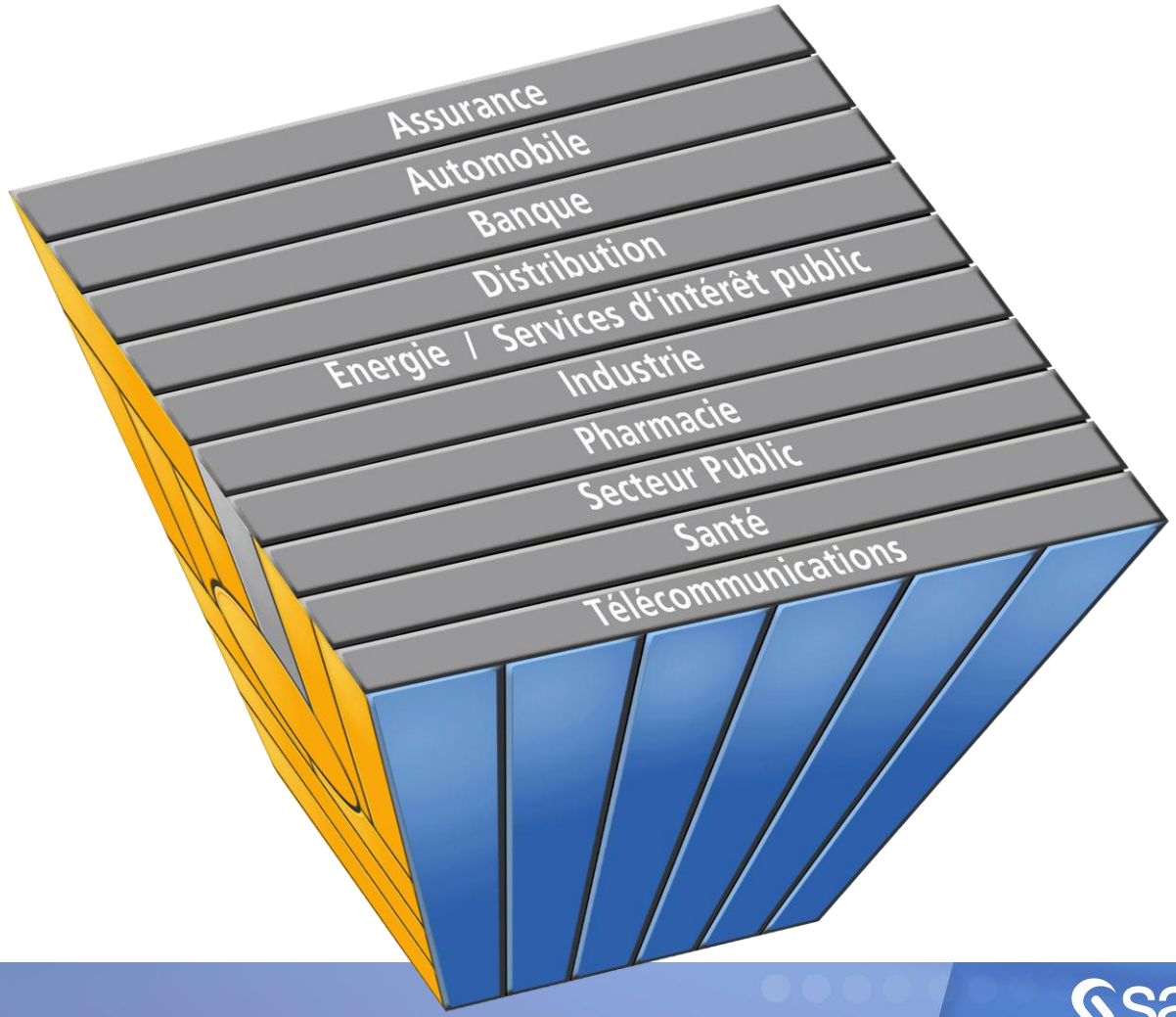


La plate forme décisionnelle
Enterprise Intelligence Platform

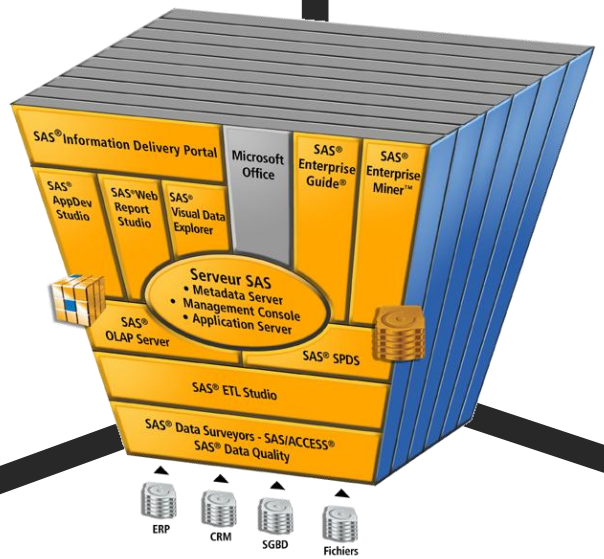


Enterprise Intelligence Platform
Une déclinaison métier

Enterprise Intelligence Platform
Une déclinaison sectorielle



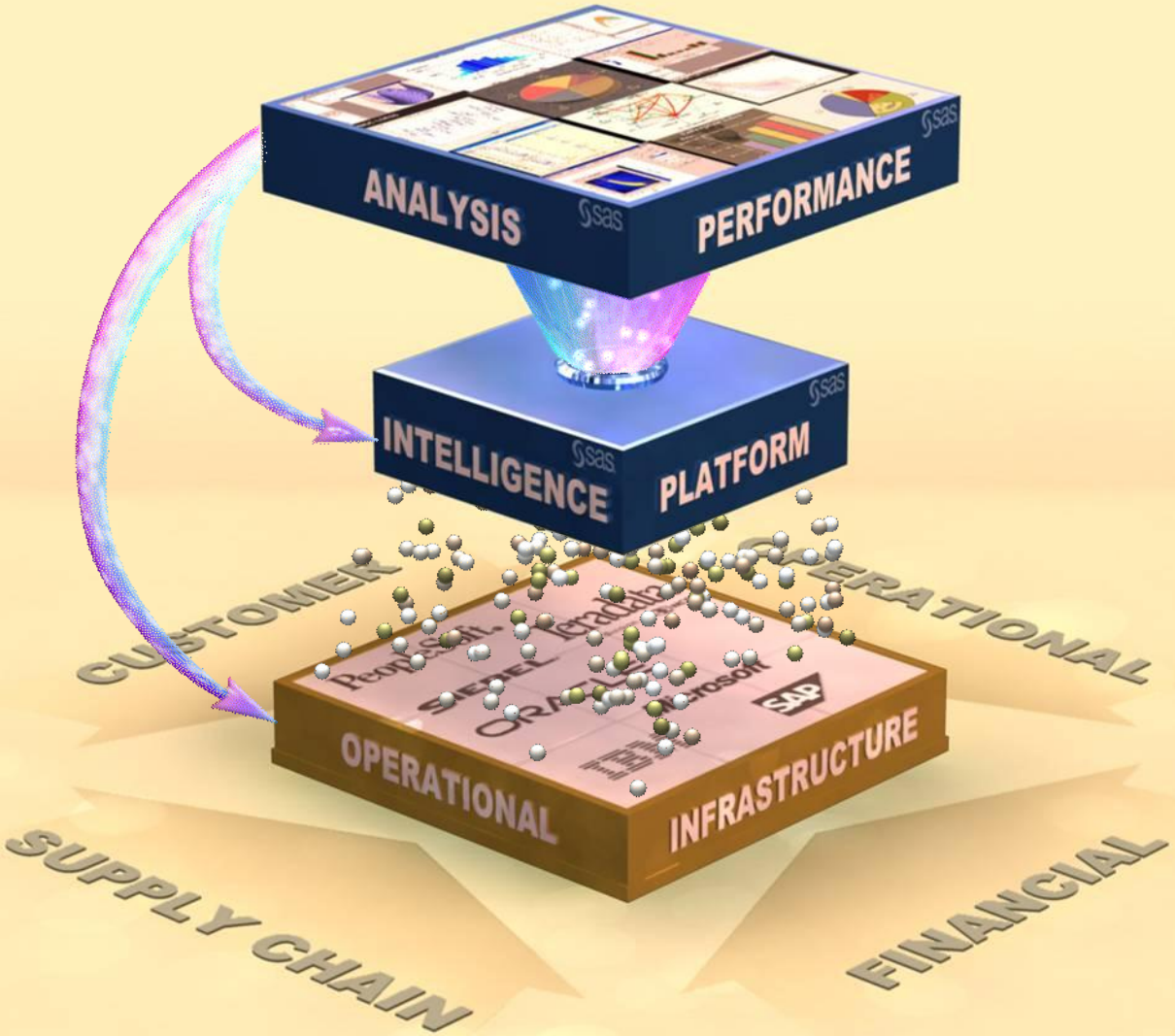
Secteurs d'activité

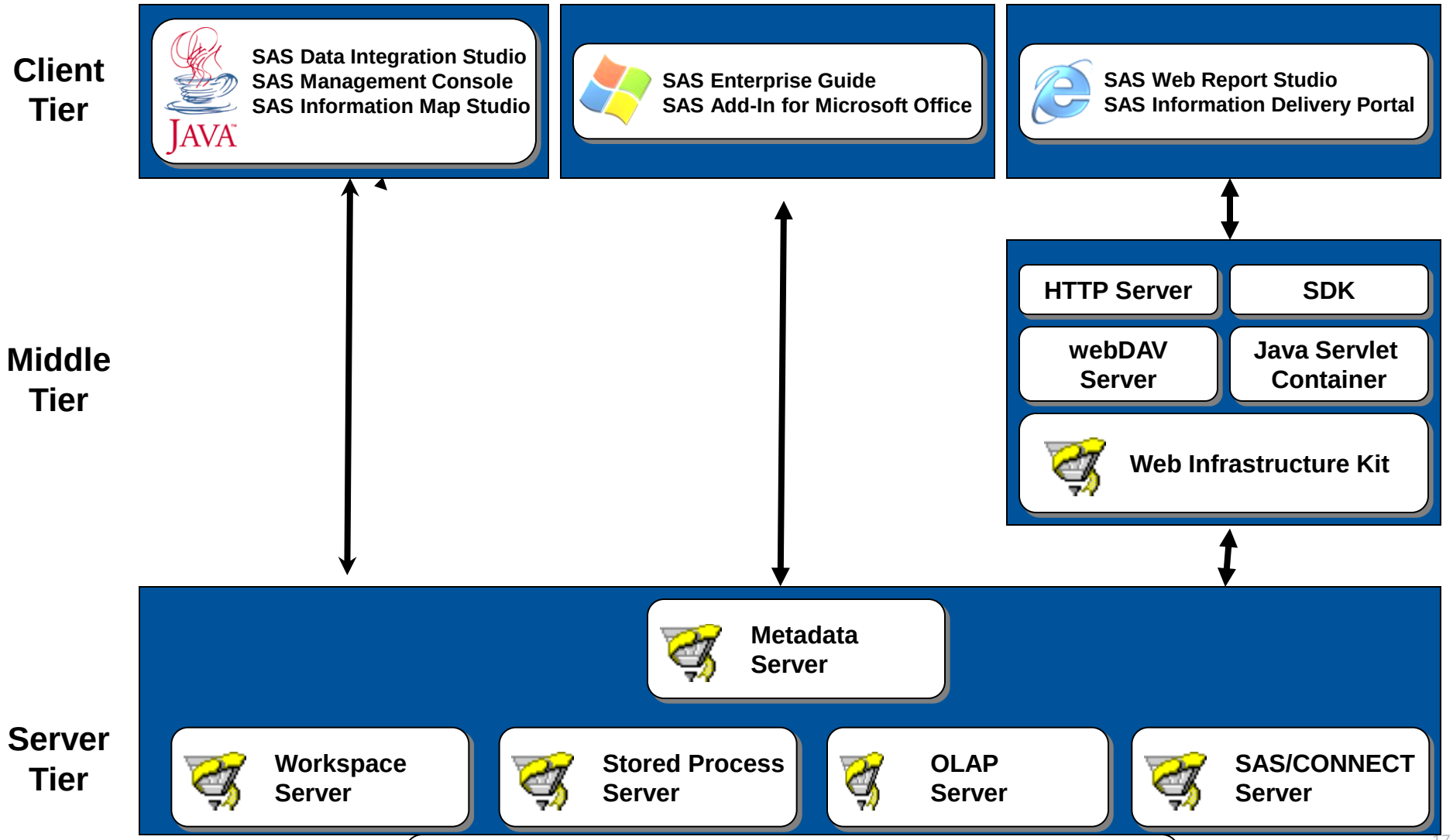


Système d'information

Métiers

Le décisionnel au cœur des processus de l'entreprise





SAS Business Intelligence Personas

IT Support

IT Admin (Ahmed)

- User administration
- Reporting administration
- Software administration

Data Modeler (Marcel)

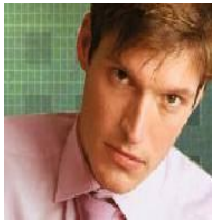
- Business View manager
- Understands physical data model
- SQL programmer

Report Administrator (Robert)

- Report Builder
- Schedule reports
- Monitor queue

Power Users

Business Analyst (Jacques)



- No DBMS or programming
- Strong Excel
- Ad hoc queries
- OLAP
- Create reports
- Publish reports
- Custom reports
- Understands business metrics

Power User (Gloria)

- Some DBMS and programming
- Strong Excel
- Ad hoc queries
- Custom reports
- Modeling
- Analytics
- Detail data
- Understands business domain

Information Consumers

C-level Execs

- Annotation
- E-mail

Middle Management (Henri)

- Drill down
- Manipulation
- Annotation

Operational Consumers

- Sales
- Marketing
- Customer Service
- Finance
- Technical Support

Le Cas Orion

La société : Orion

Cette société fictive, présente au niveau mondial, est spécialisée dans la commercialisation d'articles de sport et d'extérieur

Le siège sociale aux États-Unis, gère des filiales en Belgique, Pays Bas, Allemagne, Royaumes Unis, Danemark, France, Italie, Espagne et Australie.

Les produits sont vendu en magasin, par catalogue et par Internet.

Il y a 5 ans de transaction, depuis le 1er janvier 2003. Nous somme aujourd'hui le 1er janvier 2008.

Structure de l'organisation : (suite)

- Les employés sont enregistrés dans la base de données selon cinq niveaux:
  - Pays
  - Compagnie
  - Département
  - Section
  - Groupe

- Les informations complémentaires sur les employés sont notamment:
  - Date d'entrée et de départ de l'employé
  - Date de début et de fin de contrat (pour certain contrat)
  - Adresse
  - Sexe
  - Salaire
  - Responsable hiérarchique

L’offre

- La société Orion propose environ 5500 références. Certains ne sont pas vendus dans tous les pays, d’autres, de part les volumes commercialisés, reflètent certaines particularités régionales, certains sports nationaux.
- Tous les noms sont fictifs.
- Les produits sont organisés selon 4 niveaux:
  - Ligne de produit
  - Catégorie de produit
  - Groupe de produit
  - Produit
- Chaque produit a un coût et un prix de vente. Le système informatique gère tous les prix en dollars. En utilisant les dates de début et de fin, ces prix varient en fonction du temps. Cet historique est sauvegardé. Le système gère aussi les remises pour certains produits, à certaines périodes. Les prix sont généralement uniques de part le monde.

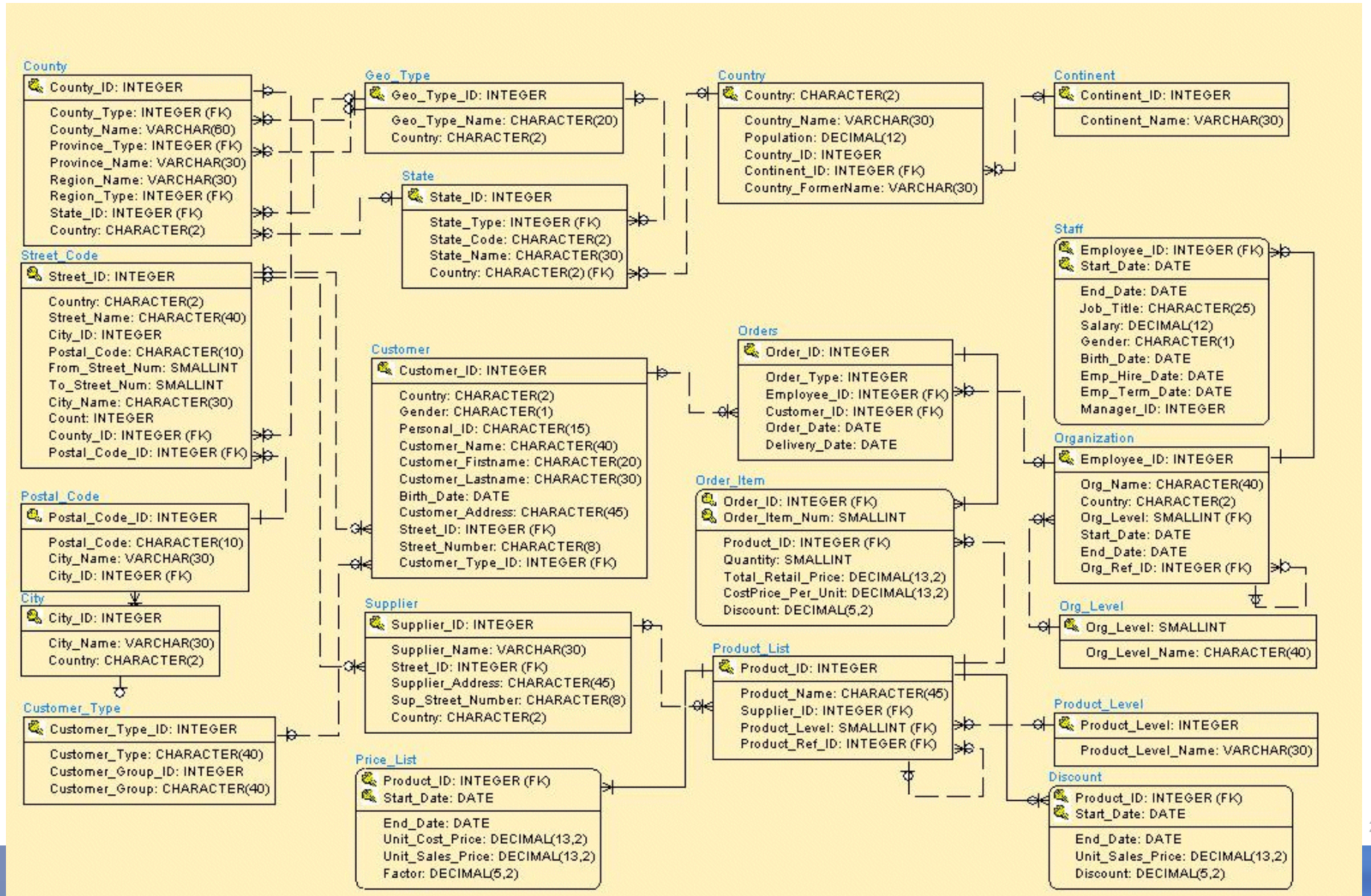
Les clients

- Les clients d'Orion Star sont repartis à travers le monde, notamment dans les pays où se trouvent des filiales, mais pas uniquement.
- Les noms et adresses sont fictifs, même si les villes, régions/comtés et pays, sont réels.
- La base de données enregistre environ 90 000 clients, pas tous actifs.
- L'adresse des clients comprend tout ou partie des informations:
  - Rue
  - Code postal
  - Ville
  - Région / département / conté
  - Etat
  - Pays
  - Continent
- La gestion des adresses est contrôlée par des pointeurs (identifiant de colonnes), ce qui facilite le changement d'adresse.
- Les clients sont classés dans des groupes en fonction de leur activité d'achat.

Les commandes

- La plupart des commandes de cette étude de cas sont pour des clients détenteurs de la carte de fidélité ‘Orion Star’, clients pour lesquels les informations sont enregistrées.
- Chaque commande pointe vers le commercial qui a enregistré la vente. Environ 980 000 commandes sont enregistrées dans cette étude de cas, commandes qui reflètent notamment les saisonnalités.
- Chaque commande comprend une ou plusieurs lignes, une ligne par produit.

Schéma relationnel normalisé des données de production



Cahier des Charges !

Suite à un audit interne, voici les principales questions recensées :

- Quelle est la tendance des ventes :
  - Quels produits sont disponibles en stock ? Où ?
  - Quels sont les produits qui se vendent le mieux ?
  - Y a-t-il une relation entre le temps, l'espace et la vente de produit ?
  - Qui a fait le plus de vente ?
- Quels sont les produits en perte de vitesse :
  - Quels sont les produits les moins vendus?
  - Est-ce que ces mauvaises ventes sont corrélées à l'espace ou au temps?
  - Quels sont les produits qui contribuent à moins de 0.05% du CA pour un Pays/une année donné(e)?
  - Est-ce que ces produits peuvent être remisés?

Cahier des Charges : (suite)

- Marge
  - Quelle est la marge générée par ce produit, ce groupe de produit, catégorie de produit et ligne de produit ?
  - Est-ce que la marge dépend de la quantité vendue ?
- Remise
  - Est-ce que les remises font augmenter les ventes ?
  - Est-ce que les remises font augmenter la marge ?
- Clients
  - Quels groupes de clients sont identifiés ?
  - Quel client achète par quel canal ?
  - Quels sont les clients les plus rentables ?
- Fournisseur
  - Quel fournisseur me propose des produits rentables?

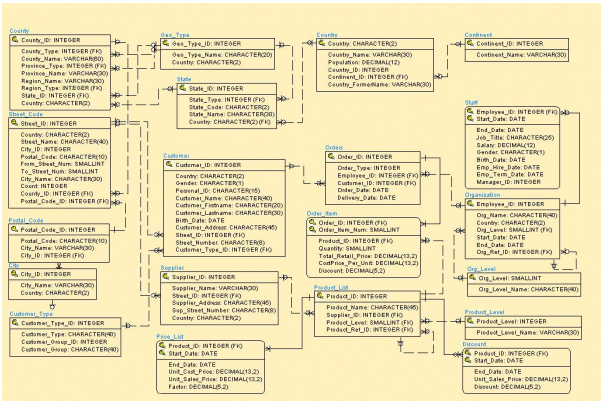
Mission :

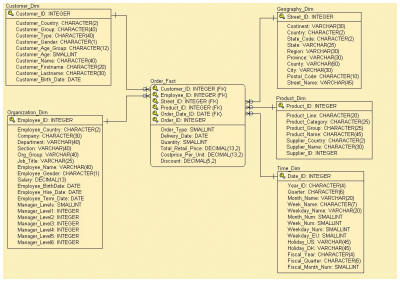
- L’objectif de cette étude de cas est de présenter un POC au comité de direction de la société Orion, répondant au cahier des charges et présentant l’intérêt de l’intégration d’un système décisionnel dans cette société.

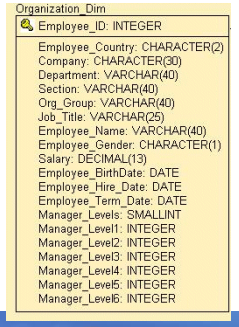
Élément de solution du cas Orion :

Supports Physiques Variés !

- Relationnel normalisé
- Relationnel dé-normalisé
- Fichiers plats
- Multidimensionnel
- Virtuel

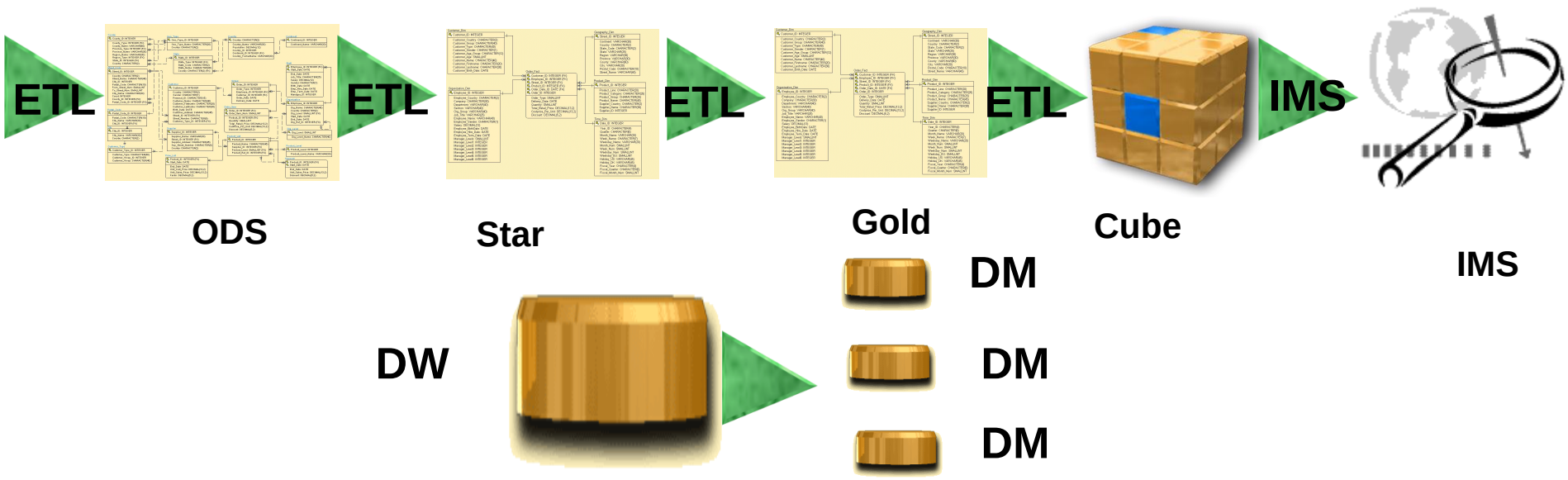


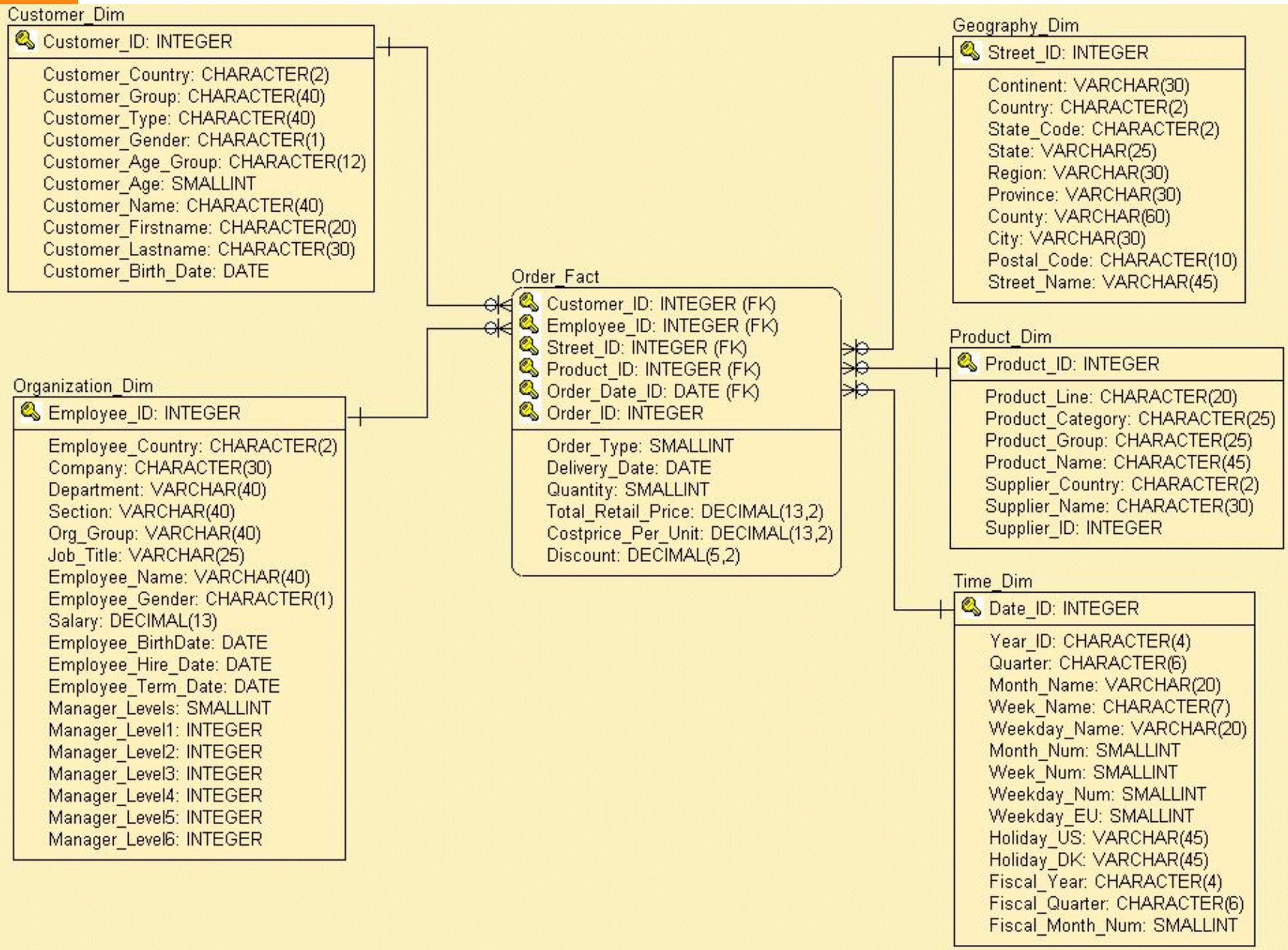




- » Index
- » Partition

Orion DW, DM ?



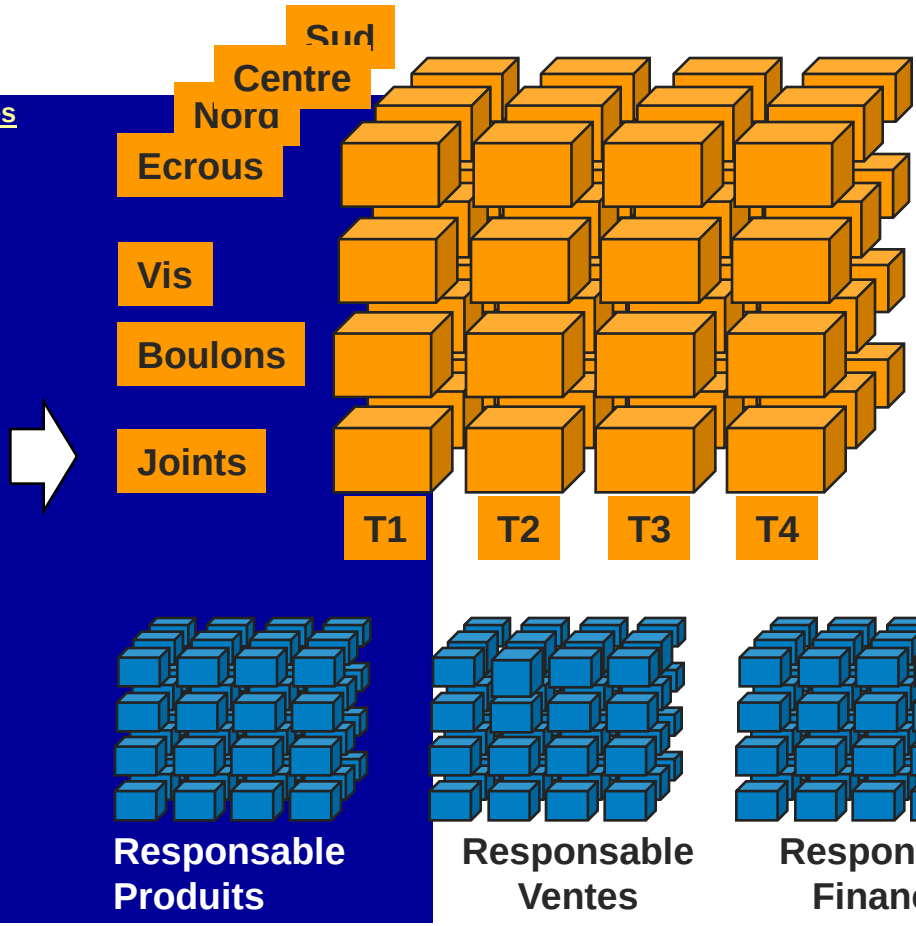

Customer_Dim
Customer_ID: INTEGER
Customer_Country: CHARACTER(2)
Customer_Group: CHARACTER(40)
Customer_Type: CHARACTER(40)
Customer_Gender: CHARACTER(1)
Customer_Age_Group: CHARACTER(12)
Customer_Age: SMALLINT
Customer_Name: CHARACTER(40)
Customer_Firstname: CHARACTER(20)
Customer_Lastname: CHARACTER(30)
Customer_Birth_Date: DATE
Geography_Dim
Street_ID: INTEGER
Continent: VARCHAR(30)
Country: CHARACTER(2)
State_Code: CHARACTER(2)
State: VARCHAR(25)
Region: VARCHAR(30)
Province: VARCHAR(30)
County: VARCHAR(60)
City: VARCHAR(30)
Postal_Code: CHARACTER(10)
Street_Name: VARCHAR(45)
Order_Fact
Customer_ID: INTEGER (FK)
Employee_ID: INTEGER (FK)
Street_ID: INTEGER (FK)
Product_ID: INTEGER (FK)
Order_Date_ID: DATE (FK)
Order_ID: INTEGER
Order_Type: SMALLINT
Delivery_Date: DATE
Quantity: SMALLINT
Total_Retail_Price: DECIMAL(13,2)
Costprice_Per_Unit: DECIMAL(13,2)
Discount: DECIMAL(5,2)
Product_Dim
Product_ID: INTEGER
Product_Line: CHARACTER(20)
Product_Category: CHARACTER(25)
Product_Group: CHARACTER(25)
Product_Name: CHARACTER(45)
Supplier_Country: CHARACTER(2)
Supplier_Name: CHARACTER(30)
Supplier_ID: INTEGER
Organization_Dim
Employee_ID: INTEGER
Employee_Country: CHARACTER(2)
Company: CHARACTER(30)
Department: VARCHAR(40)
Section: VARCHAR(40)
Org_Group: VARCHAR(40)
Job_Title: VARCHAR(25)
Employee_Name: VARCHAR(40)
Employee_Gender: CHARACTER(1)
Salary: DECIMAL(13)
Employee_BirthDate: DATE
Employee_Hire_Date: DATE
Employee_Term_Date: DATE
Manager_Levels: SMALLINT
Manager_Level1: INTEGER
Manager_Level2: INTEGER
Manager_Level3: INTEGER
Manager_Level4: INTEGER
Manager_Level5: INTEGER
Manager_Level6: INTEGER
Time_Dim
Date_ID: INTEGER
Year_ID: CHARACTER(4)
Quarter: CHARACTER(6)
Month_Name: VARCHAR(20)
Week_Name: CHARACTER(7)
Weekday_Name: VARCHAR(20)
Month_Num: SMALLINT
Week_Num: SMALLINT
Weekday_Num: SMALLINT
Weekday_EU: SMALLINT
Holiday_US: VARCHAR(45)
Holiday_DK: VARCHAR(45)
Fiscal_Year: CHARACTER(4)
Fiscal_Quarter: CHARACTER(6)
Fiscal_Month_Num: SMALLINT

OLAP

Structure multidimensionnelle
Base à n-dimensions

Données Relationnelles

Produit	Region	Temps	Ventes
Ecrous	Nord	T1	100
Ecrous	Sud	T1	70
Ecrous	Centre	T1	50
Vis	Nord	T1	80
Vis	Sud	T1	70
Vis	Centre	T1	40
Boulons	Nord	T1	50
Boulons	Sud	T1	40
Boulons	Central	T1	10
Joints	Nord	T1	40
Joints	Sud	T1	40
Joints	Central	T1	30
Ecrous	Nord	T2	90
Ecrous	Sud	T2	70
Ecrous	Centre	T2	40
Vis	Nord	T2	90
Vis	Sud	T2	60
Vis	Centre	T2	35
Boulons	Nord	T2	45
Boulons	Sud	T2	45
Boulons	Centre	T2	20
Joints	Nord	T2	30
Joints	Sud	T2	35
Joints	Centre	T2	30



Structure multidimensionnelle
Qu’est-ce que la consolidation??

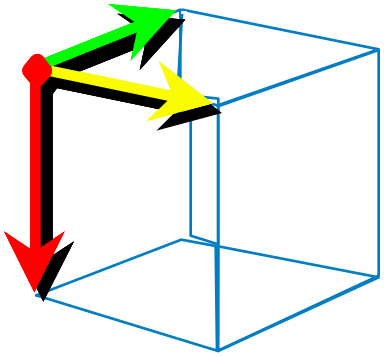
- Le seul moyen d’obtenir des temps de réponse performants consiste à «pré-calculer» tous les totaux logiques

Structure multidimensionnelle
On Line Analytical Processing

- **R**olap : Relational olap
- **M**ultidimensional olap
- **H**ybrid olap

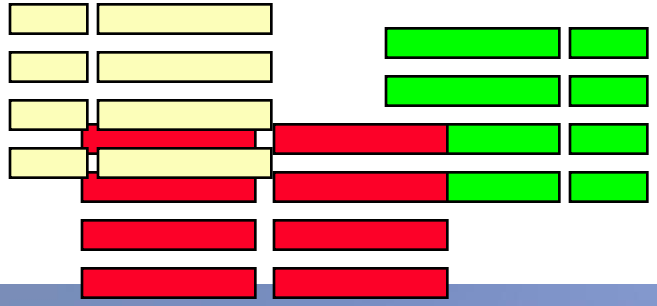
Structure multidimensionnelle

Rolap



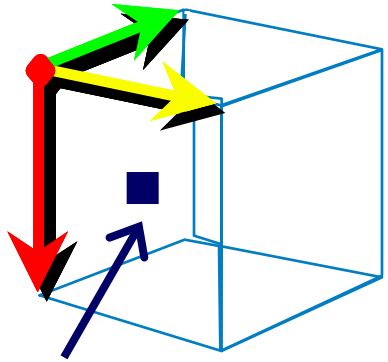
- Repose sur une structure relationnelle,
- Pas de structure de stockage dédiée,
- Calcul à la volée des agrégats demandés par l ’utilisateur.

Requêtes SQL



Structure multidimensionnelle
Molap

Indicateur

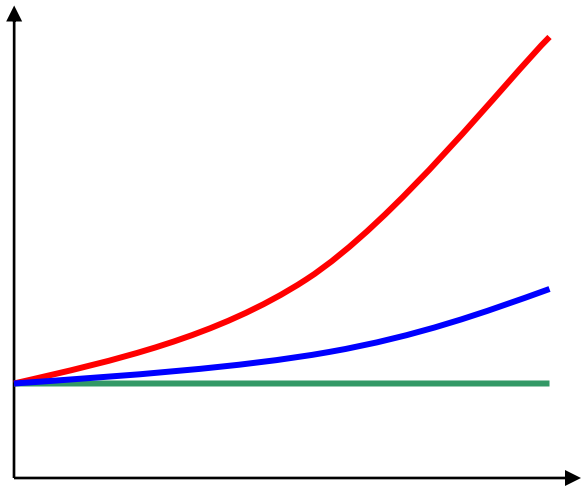


- accès direct

- Forme OLAP la plus « pure »
- Repose sur une structure multidimensionnelle.
- Accès immédiat à l'agrégat désiré.

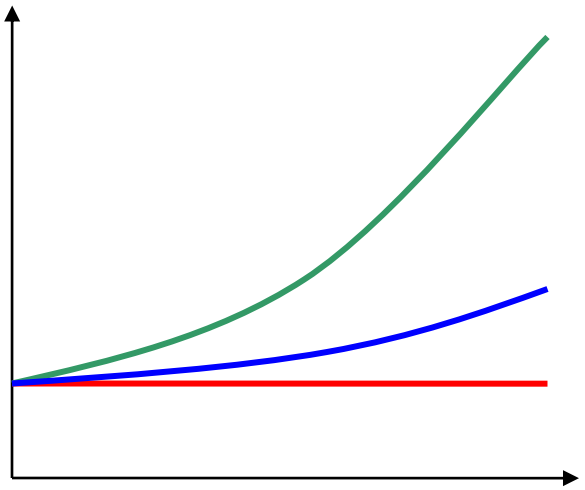
MOLAP ROLAP HOLAP

Espace disque



Complexité

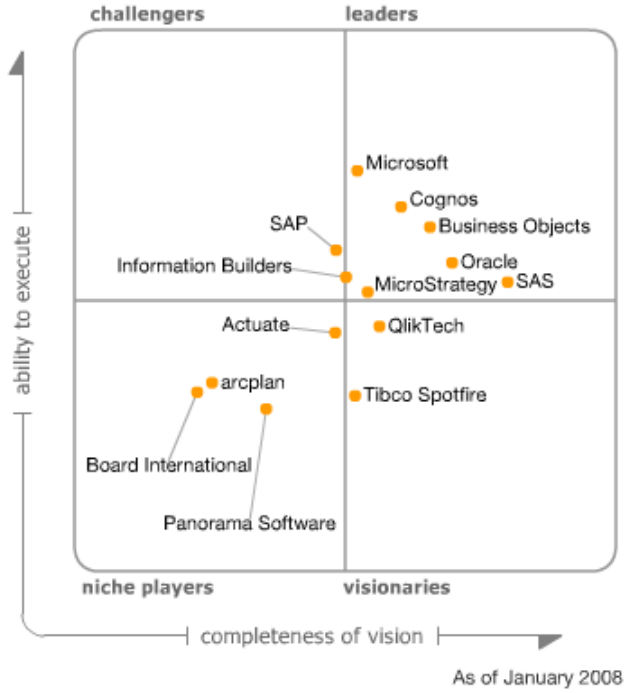
Temps de réponse



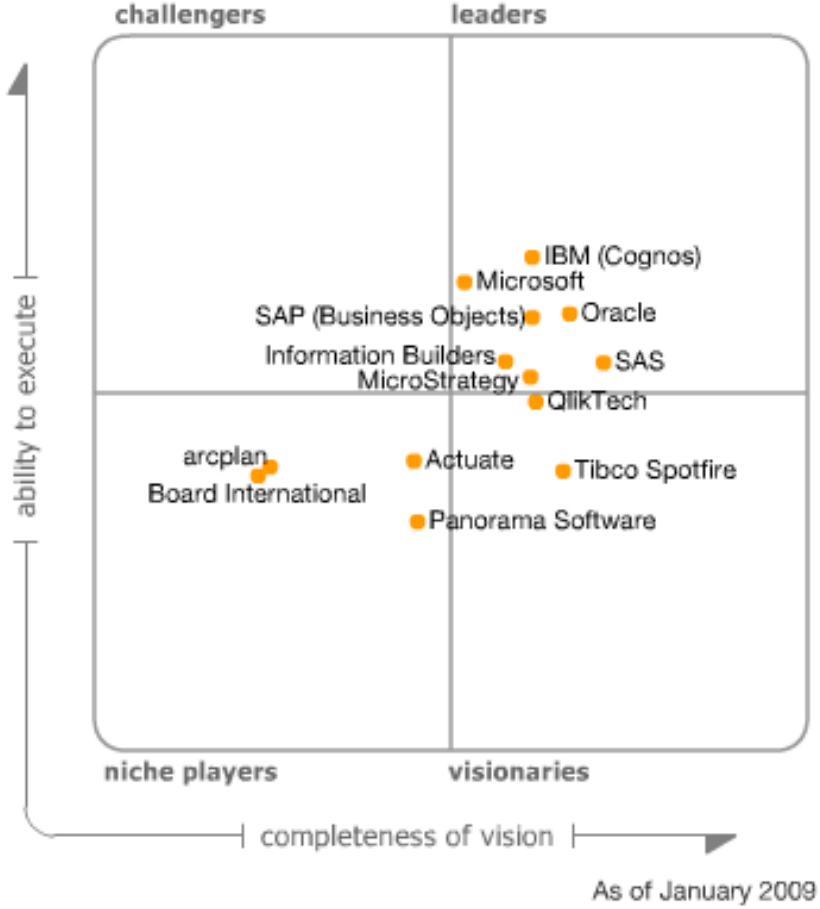
Complexité

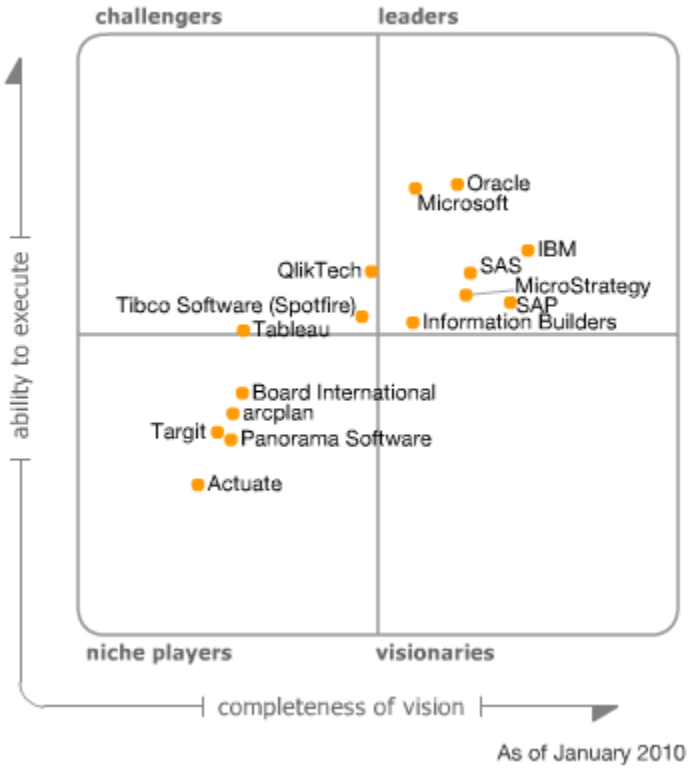
Le marché de la BI

Figure 1. Magic Quadrant for Business Intelligence Platforms, 2008



Source: Gartner (January 2008)





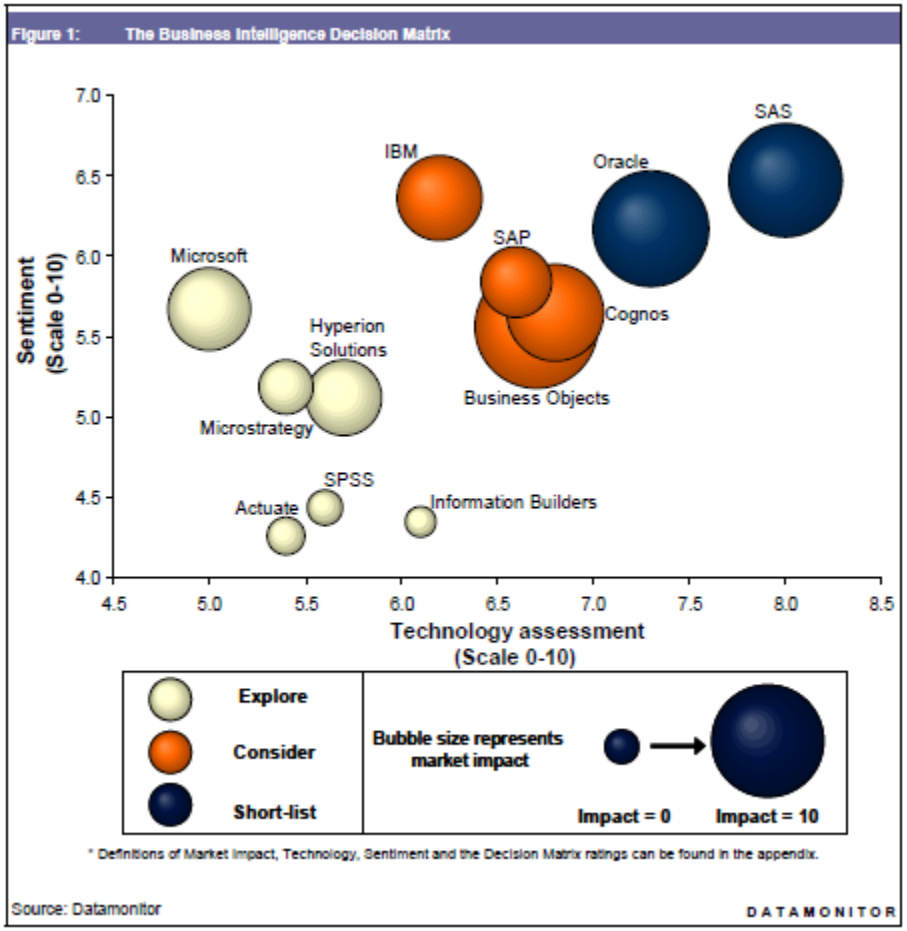


TABLE 3

(Top 5) Worldwide Business Intelligence Tools Revenue by Vendor, 2007–2009

	Revenue ($M)			Share (%)				
Company	2007	2008	2009	2007	2008	2009	2007–2008 Growth (%)	2008–2009 Growth (%)
SAP	1,356.7	1,574.6	1,557.1	19.0	20.2	19.5	16.1	-1.1
IBM	1,153.3	1,145.6	1,224.3	16.1	14.7	15.3	-0.7	6.9
SAS	785.4	870.5	909.5	11.0	11.1	11.4	10.8	4.5
Oracle	596.7	701.1	719.5	8.3	9.0	9.0	17.5	2.6
Microsoft	554.9	648.7	701.3	7.8	8.3	8.8	16.9	8.1
Other	2,706.2	2872.7	2893.6	37.8	36.8	36.1	6.2	0.7
Total	7,153.2	7,813.4	8,005.3	100.0	100.0	100.0	9.2	2.5

Note: In cases where acquisitions were completed in 2009, revenue from acquired companies has been appended to the current and past years for the given vendor.

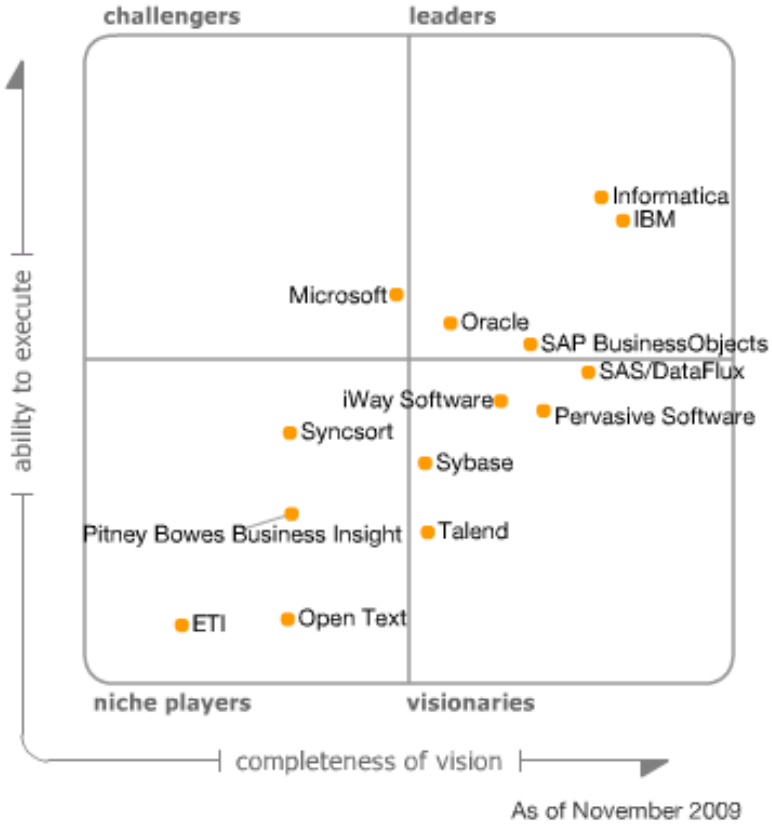
Source: IDC, June 2010

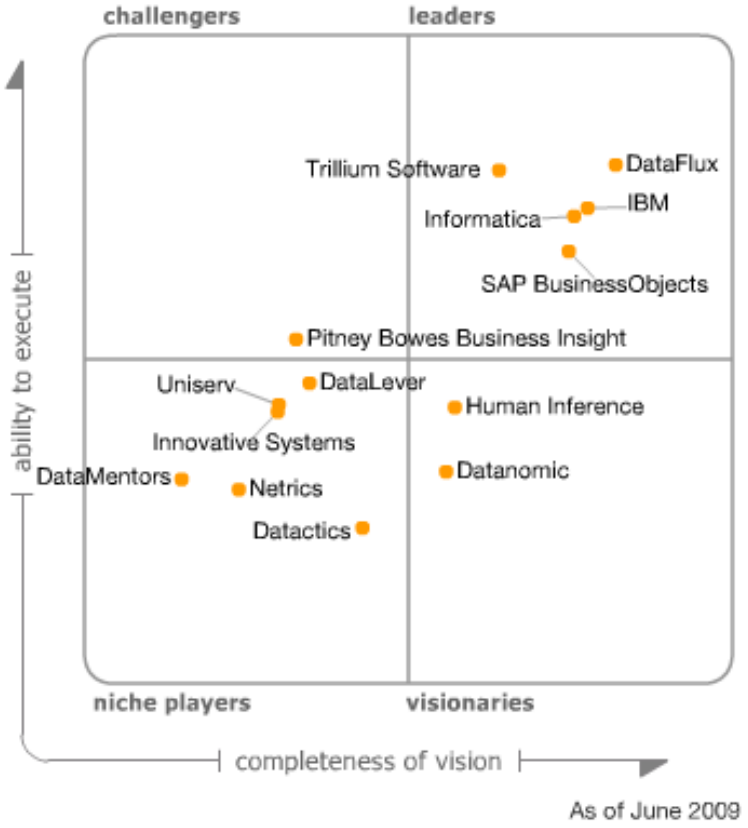
TABLE 5

(Top 7) Worldwide Advanced Analytics Tools Revenue by Vendor, 2007–2009

	Revenue ($M)			Share (%)				
Company	2007	2008	2009	2007	2008	2009	2007–2008 Growth (%)	2008–2009 Growth (%)
SAS	433.8	497.8	529.0	32.0	32.9	34.7	14.8	6.3
IBM	228.2	242.0	228.0	16.8	16.0	15.0	6.0	-5.8
Microsoft	22.0	25.3	27.3	1.6	1.7	1.8	15.0	8.0
KXEN	11.7	16.3	15.4	0.9	1.1	1.0	39.3	-5.5
TIBCO	15.5	20.5	12.8	1.1	1.4	0.8	32.3	-37.6
FICO	13.3	18.2	12.1	1.0	1.2	0.8	36.8	-33.5
Oracle	9.1	9.4	9.5	0.7	0.6	0.6	3.0	1.1
Other	621.3	683.2	688.3	45.9	45.2	45.2	10.0	0.7
Total	1,355.0	1,512.7	1,522.4	100.0	100.0	100.0	11.6	0.6

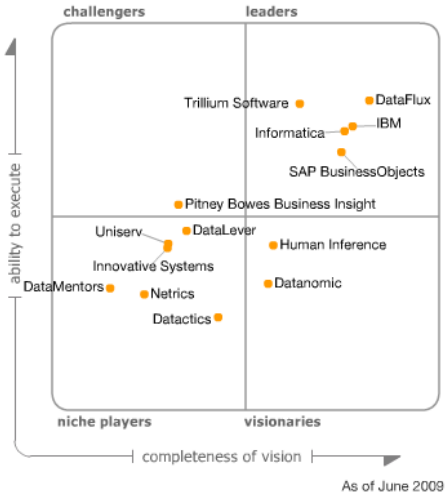
Note: In cases where acquisitions were completed in 2009, revenue from acquired companies has been appended to the current and past years for the given vendor. Source: IDC, June 2010



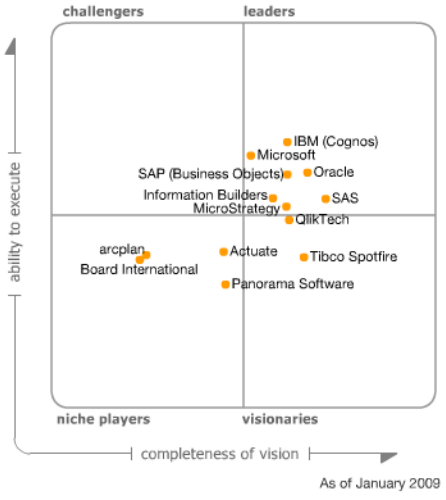


Enterprise Marketing Management

Data Quality Tools



BI Platforms



Operational Risk Manag[mt] Software for Financial Services : **Leaders**' Quadrant

Enterprise Marketing Management : **Visionaries** Quadrant

Data Quality Tools : **Leaders**' Quadrant

CRM Multichannel Campaign Management : **Leaders**' Quadrant

Energy Trading and Risk Management Platforms : **Visionaries** Quadrant

Marketing Resource Management : **Visionaries**' Quadrant

Business Intelligence Platforms : **Leaders**' Quadrant

Data Integration Tools : **Visionaries**' Quadrant

Figure 2 The Forrester Wave: Predictive Analytics And Data Mining Solutions, Q1 '10

Source: Forrester Research, Inc.

Total Cost of Ownership

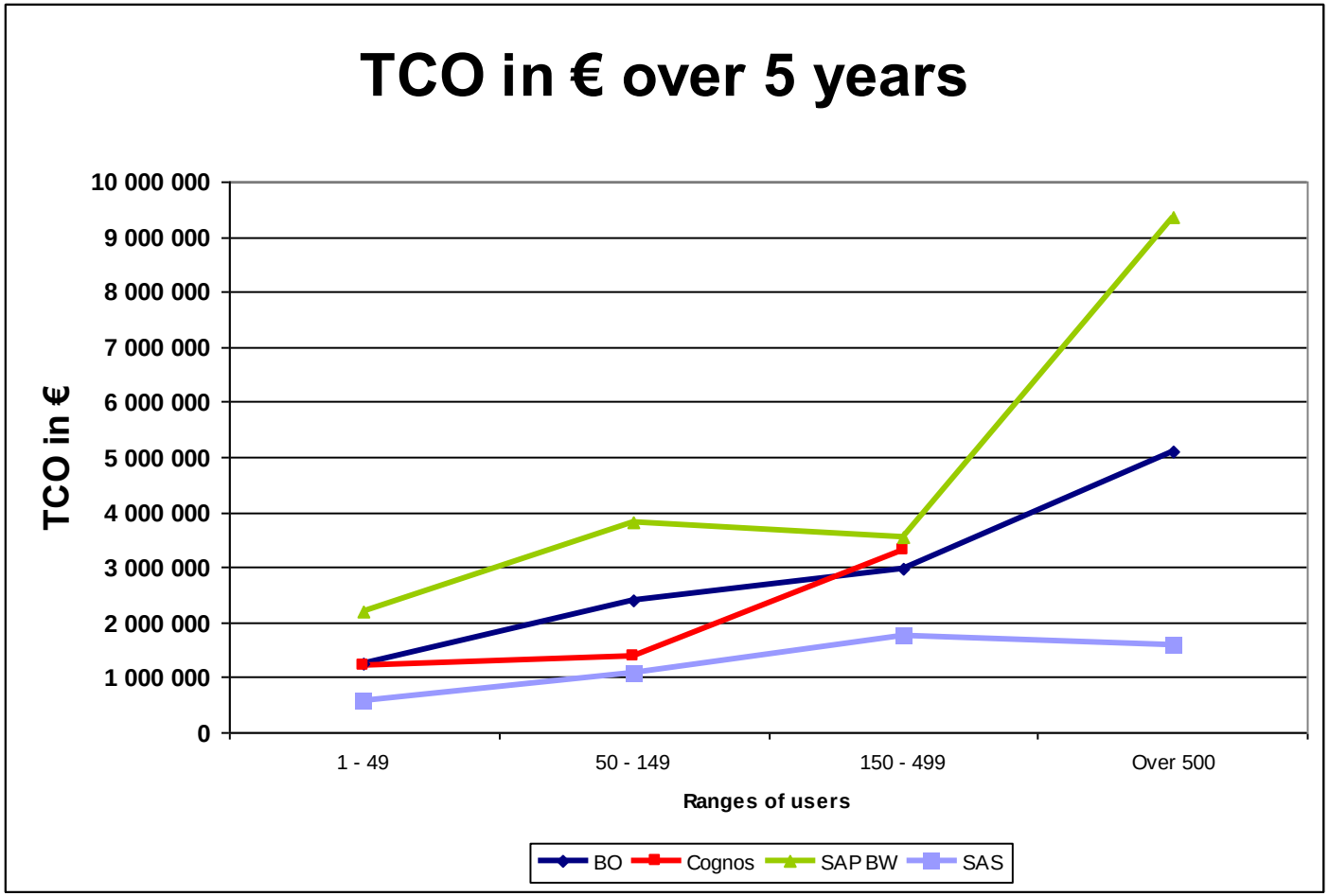
Components of the Total Cost of Ownership are:

- Software costs
- Hardware costs
- Manpower costs to deploy and maintain the solution

Many companies based their investment decisions mainly based on the software acquisition costs

This is only a minor component of the TCO

Manpower costs are by far more important and this is the focus of this presentation



Le Data Mining avec Enterprise Miner

Définition

- Les 2 familles de techniques de DM
  - Les techniques descriptives :
    - segmentation (« clustering »)
    - Recherche d'associations (séquences)
    - Algorithmes génétiques (SAS OR)
  - Les techniques prédictives :
    - régression logistique
    - arbres de décision
    - réseaux de neurones
    - Raisonnement à base de cas
    - SVM
  - Autres choses
- Traitement de gros volumes et intégration du DM dans les processus de production

Les 10 étapes d'un projet

- Choix du sujet - Définition des objectifs
- Inventaire des données existantes
- Collecte, nettoyage et mise en forme des données
- Constitution de la base d'analyse
- Mise en œuvre des algorithmes (segmentation, scoring…) - Elaboration des modèles
- Validation et choix d'un modèle
- Déclaration à la CNIL
- Déploiement du modèle
- Formation des utilisateurs
- Analyse des retours de l'action et suivi des outils

Source

http://data.mining.free.fr/

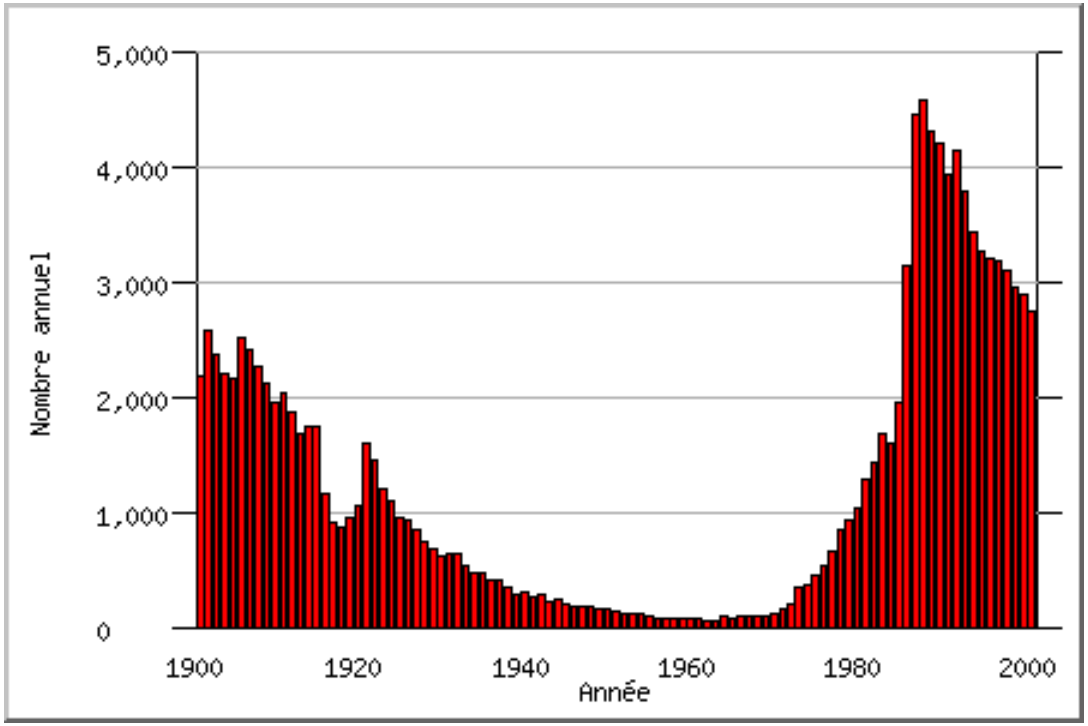
Segmentation RFM

Nombre de commandes Récence	4	3	2	1
T – 1	1111	1110 1101 1011	1100 1010 1001	1000
T – 2		0111	0110 0101	0100
T – 3			0011	0010
T - 4				0001

Les données utilisées en Data Mining

A partir des données opérationnelles :

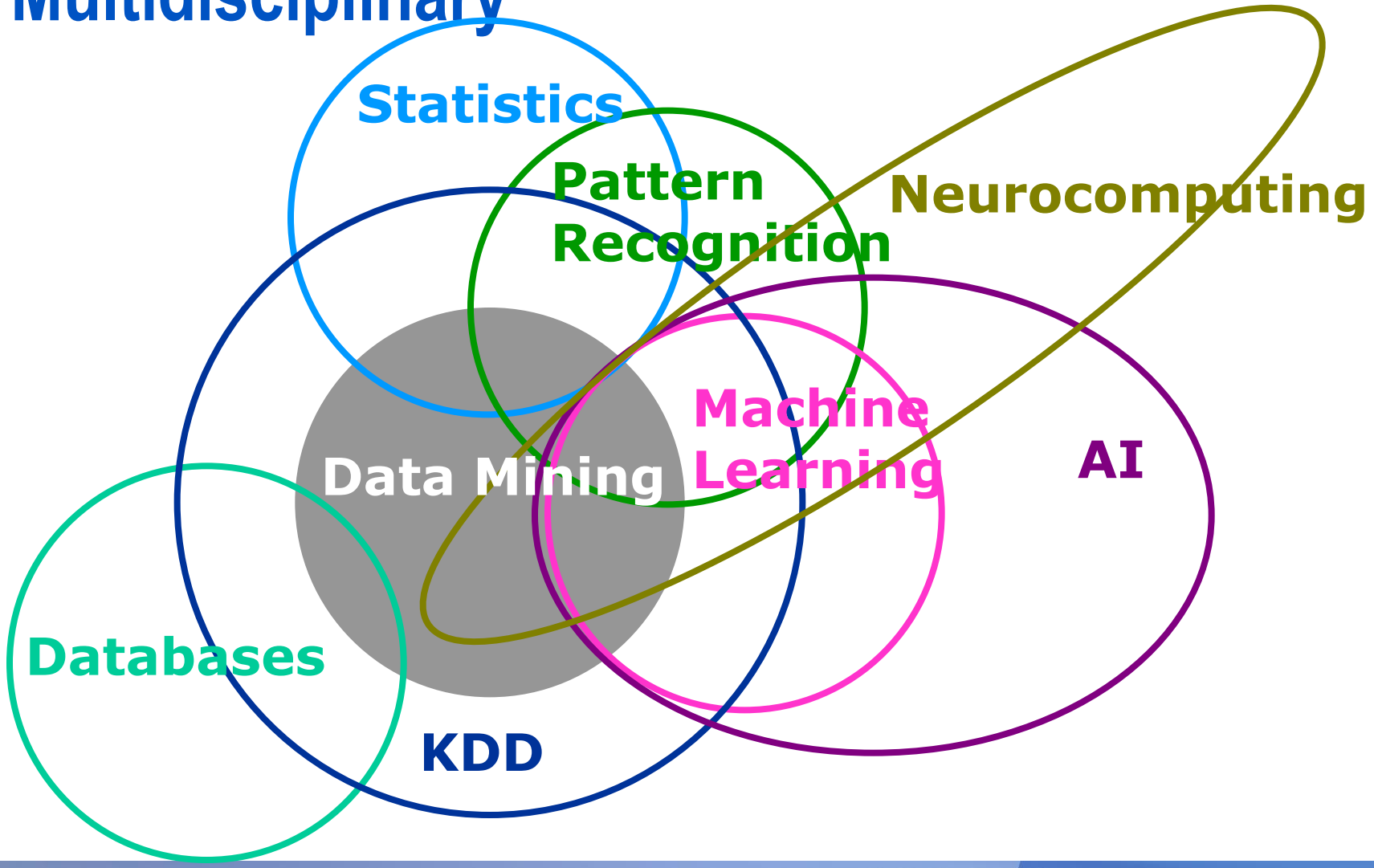
- Où (lieux géographiques, Internet,)
- Quand (Fréquence, récence,)
- Comment (mode de payement,)
- Combien (nombre de TE,)
- Quoi (Produit,)
-
-
-



Distribution du prénom Charlotte

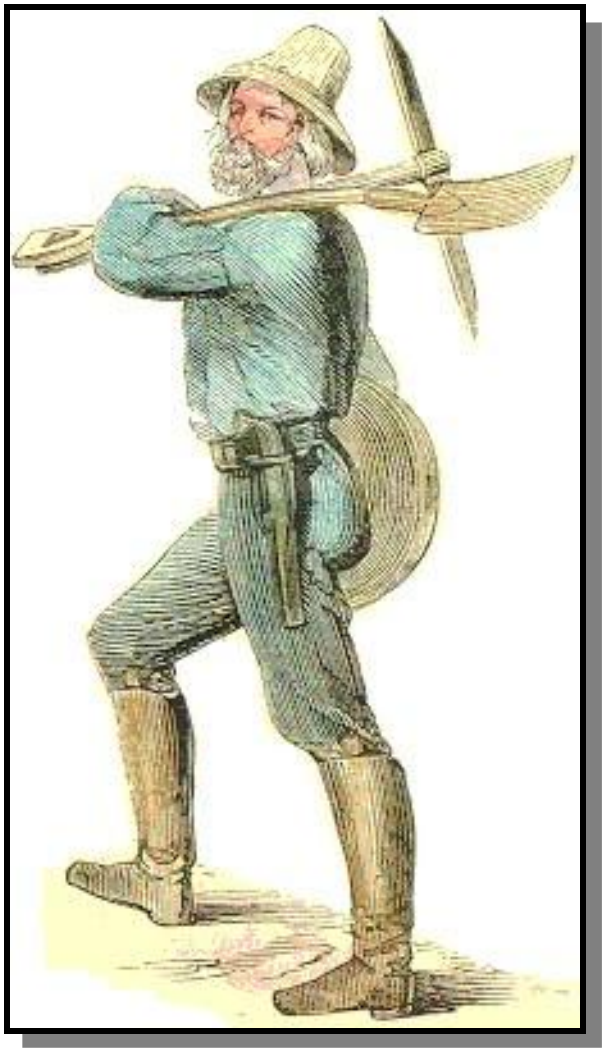
http://www.meilleursprenoms.com

Multidisciplinary

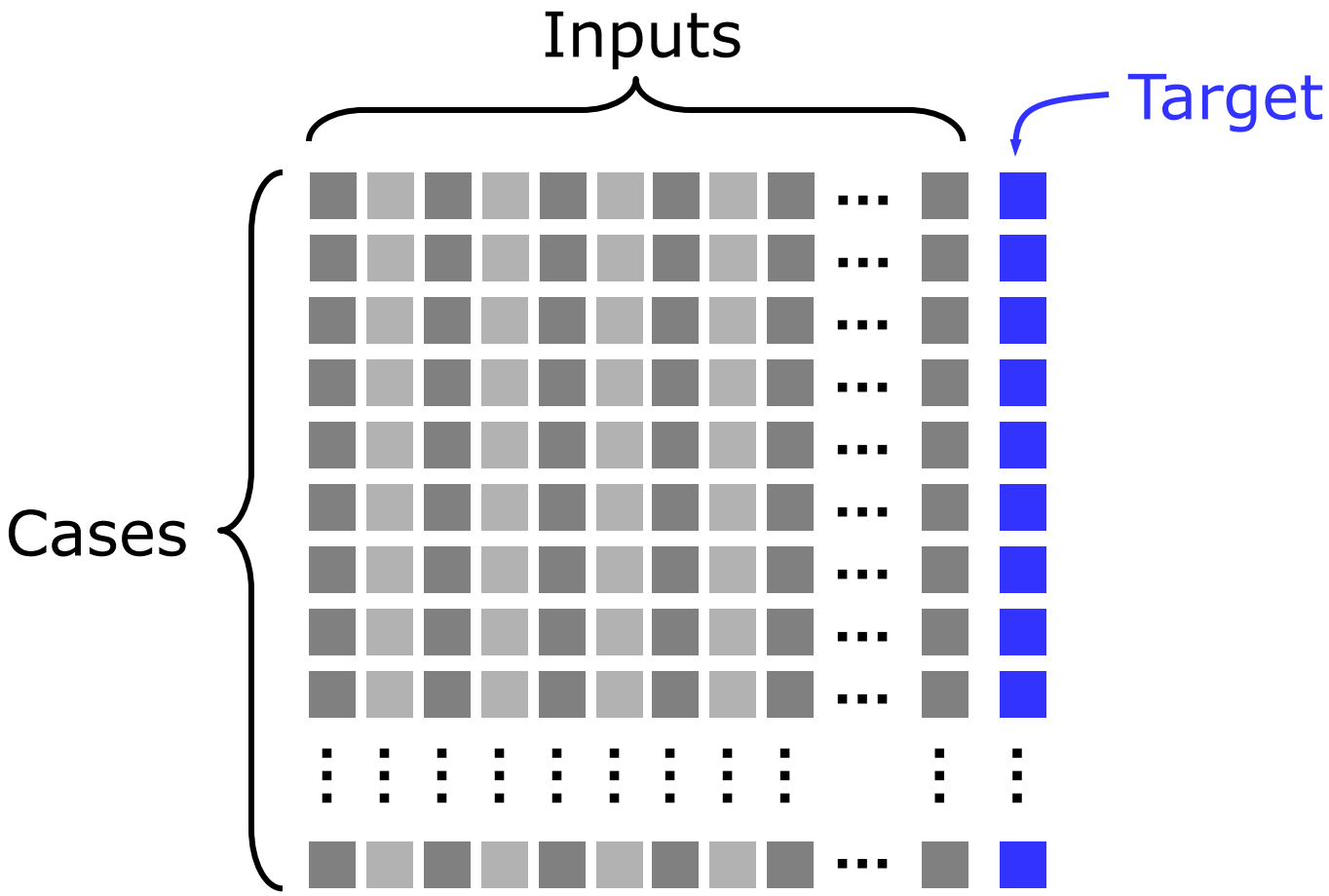


Required Expertise

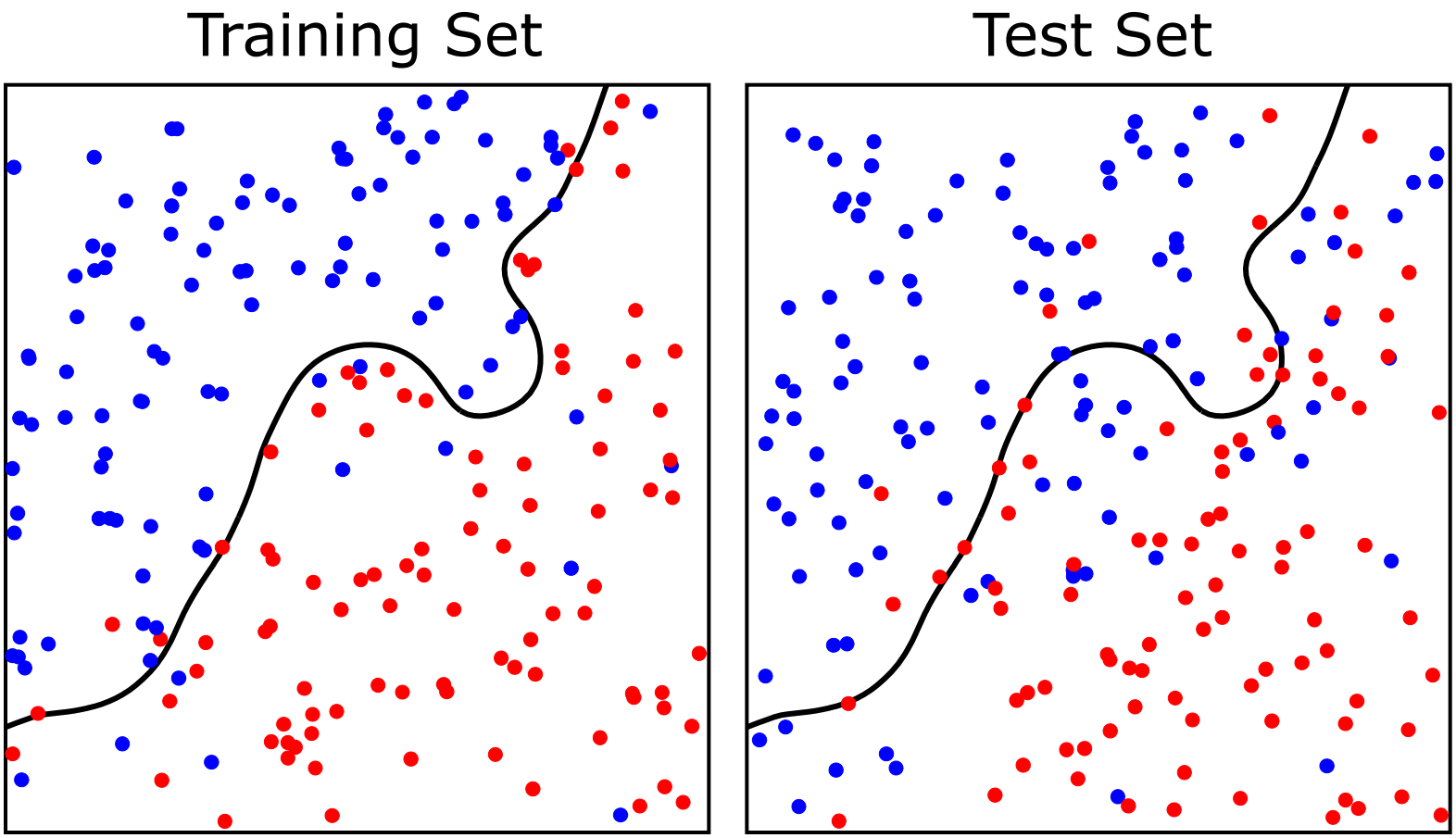
- Domain
- Data
- Analytical Methods



Predictive Modeling



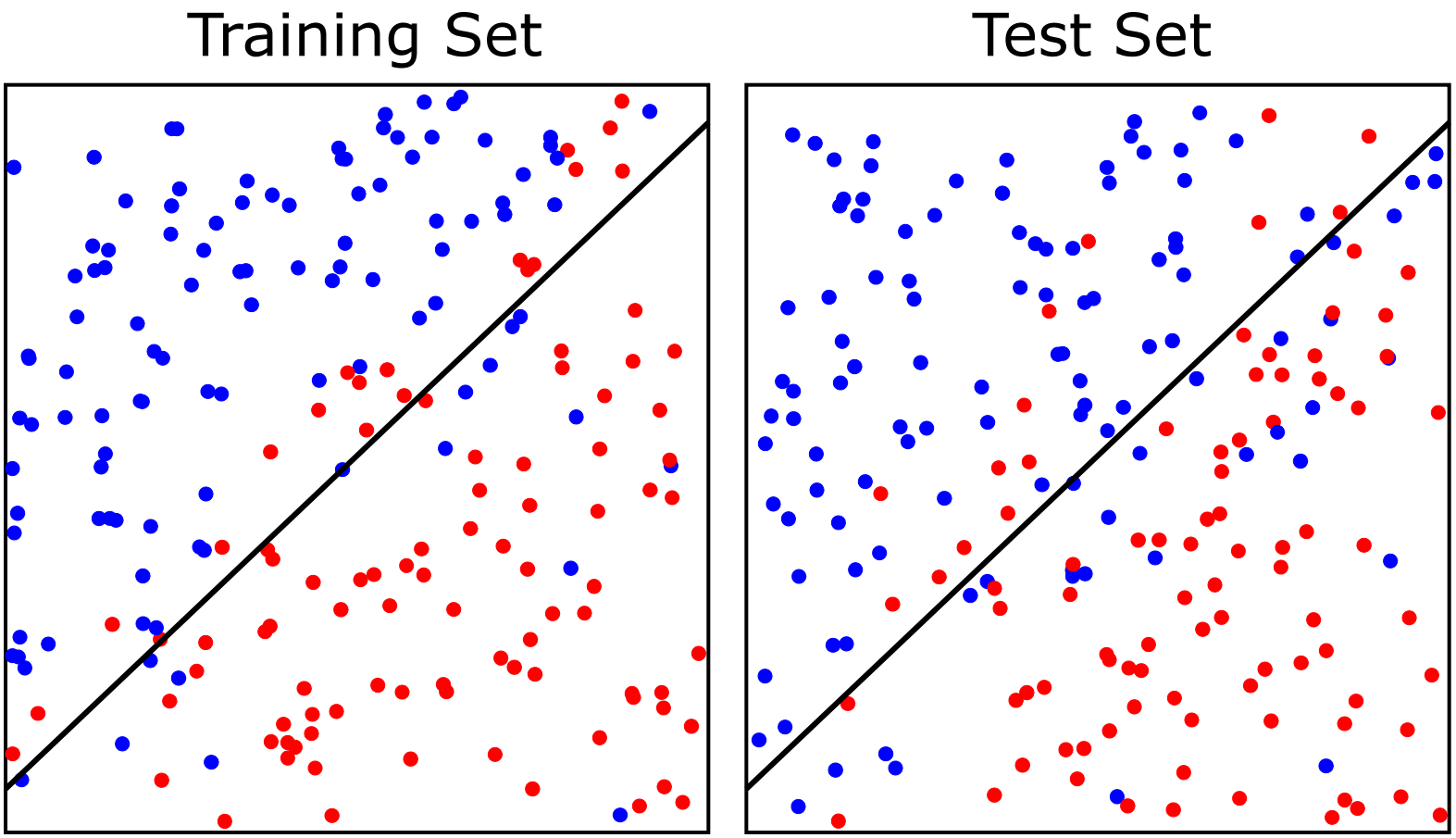
Overfitting



19 e = 90 %

49 e = 75 %

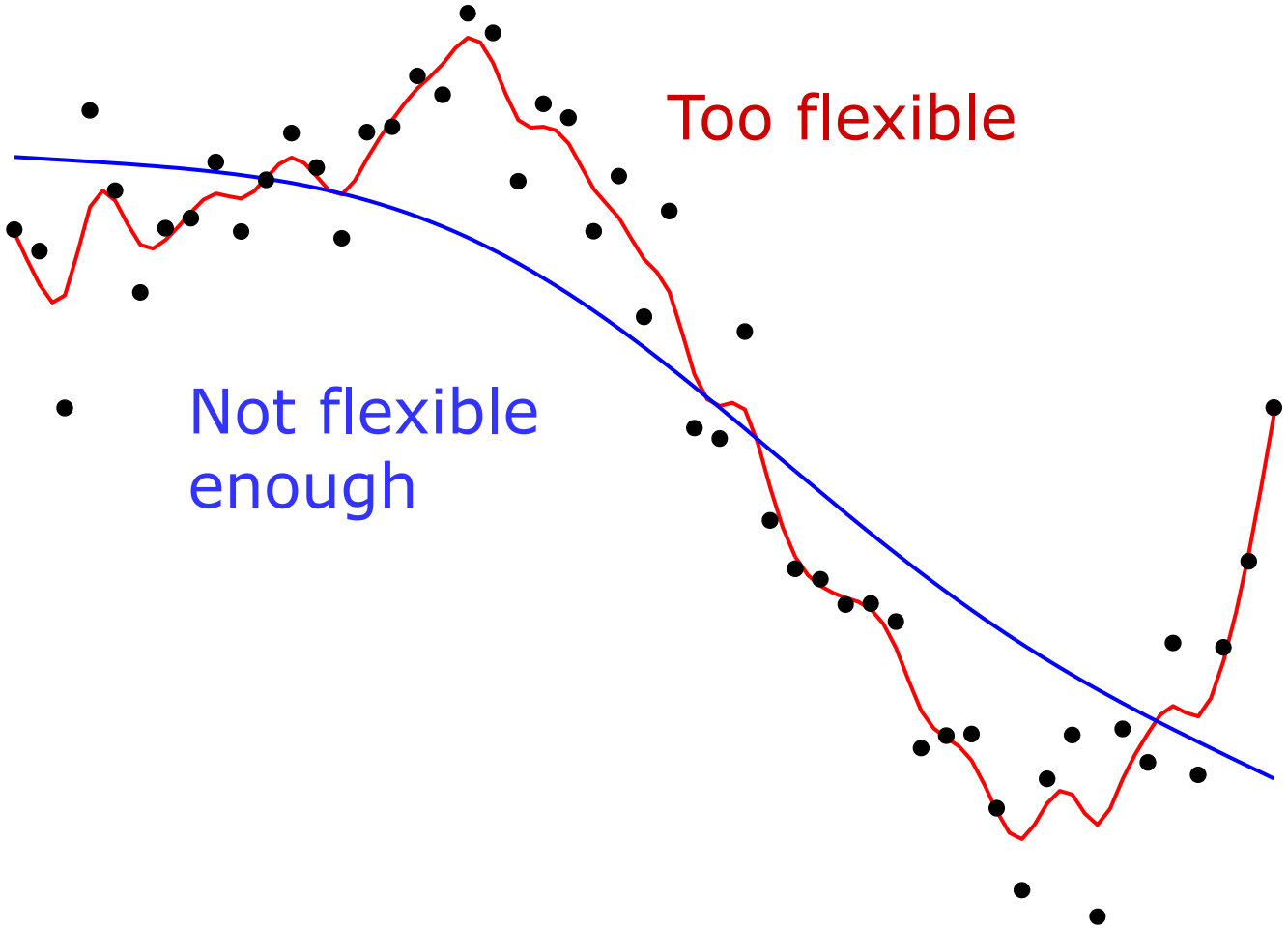
Better Fitting



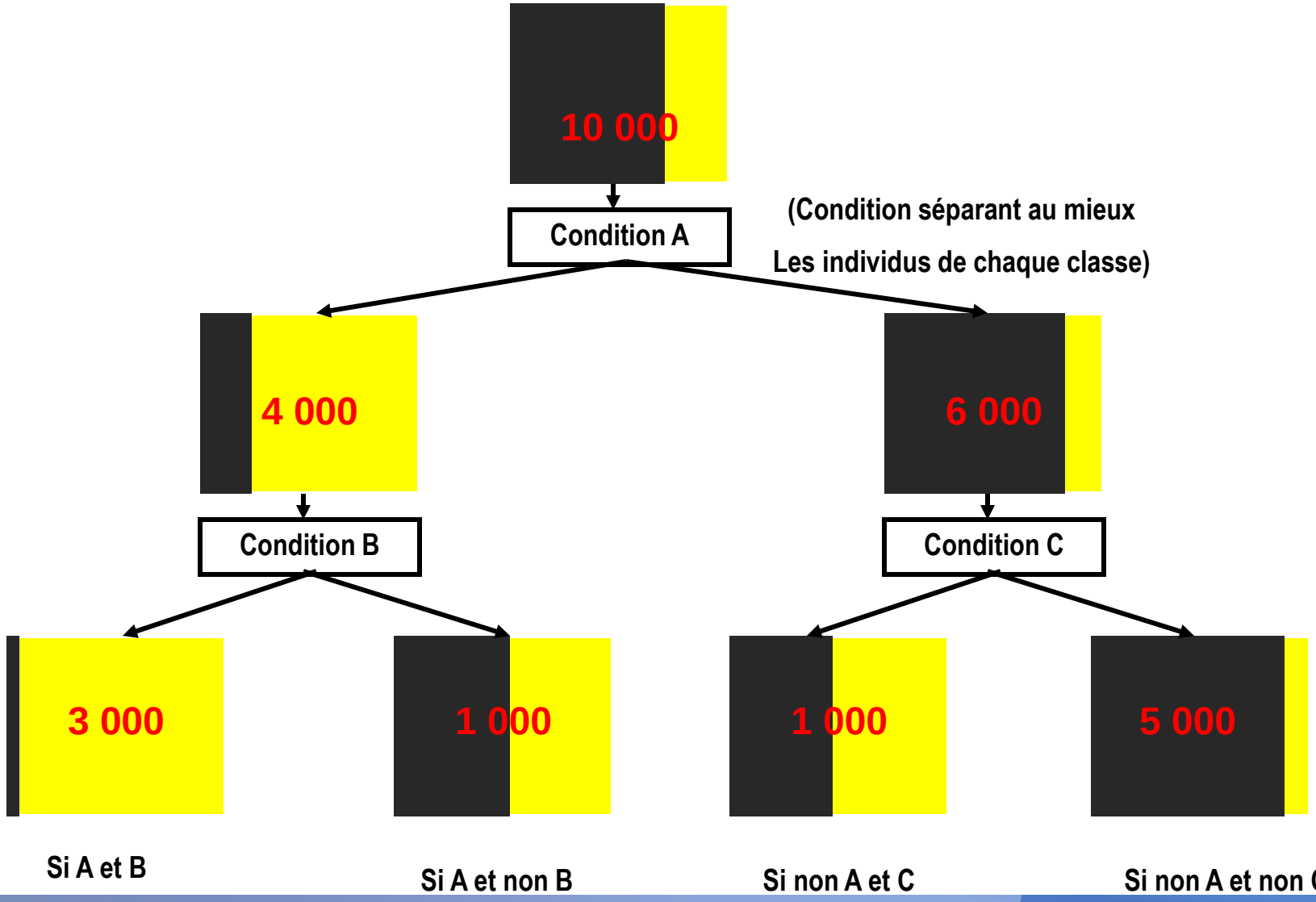
34 e = 83%

43 e =78%

Model Complexity

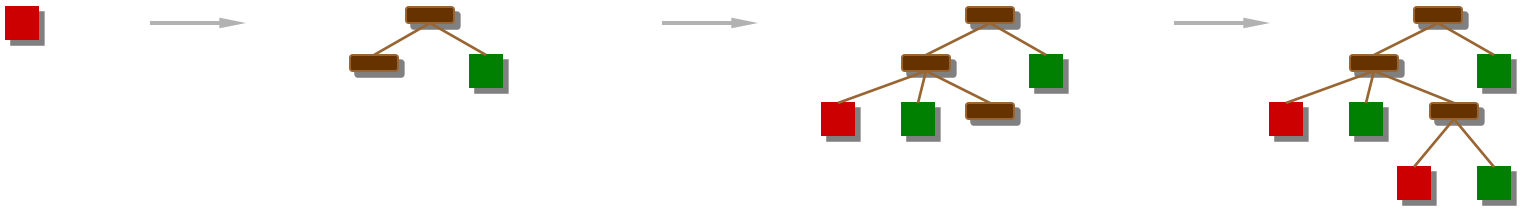


Arbre de décision

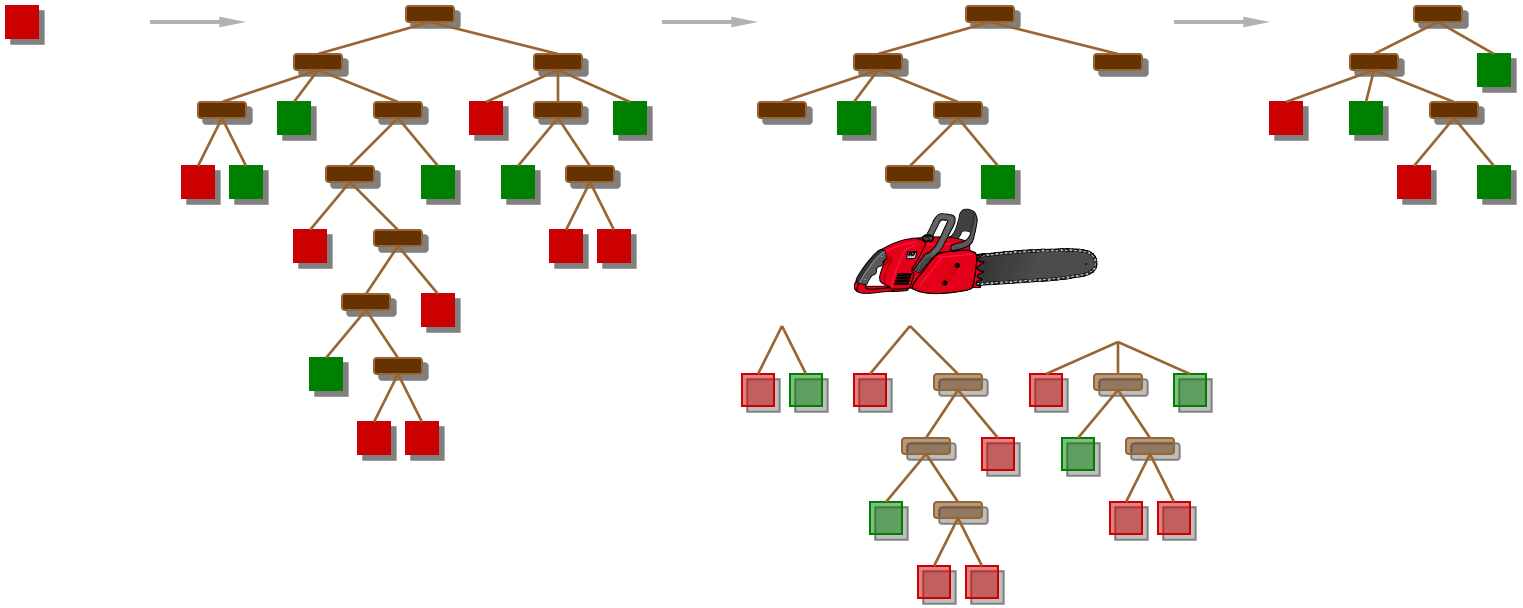


The Right-Sized Tree

Stunting



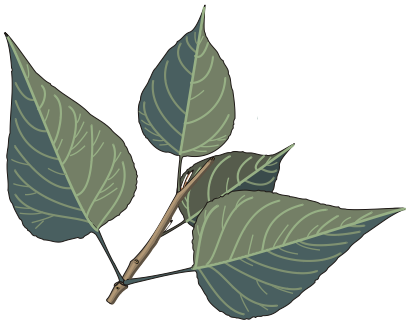
Pruning



A Field Guide to Tree Algorithms

AID
THAID
CHAID

CART



ID3
C4.5
C5.0

Measurement:

unary - one value
for example, a variable with a particular value that was used to create a data subset

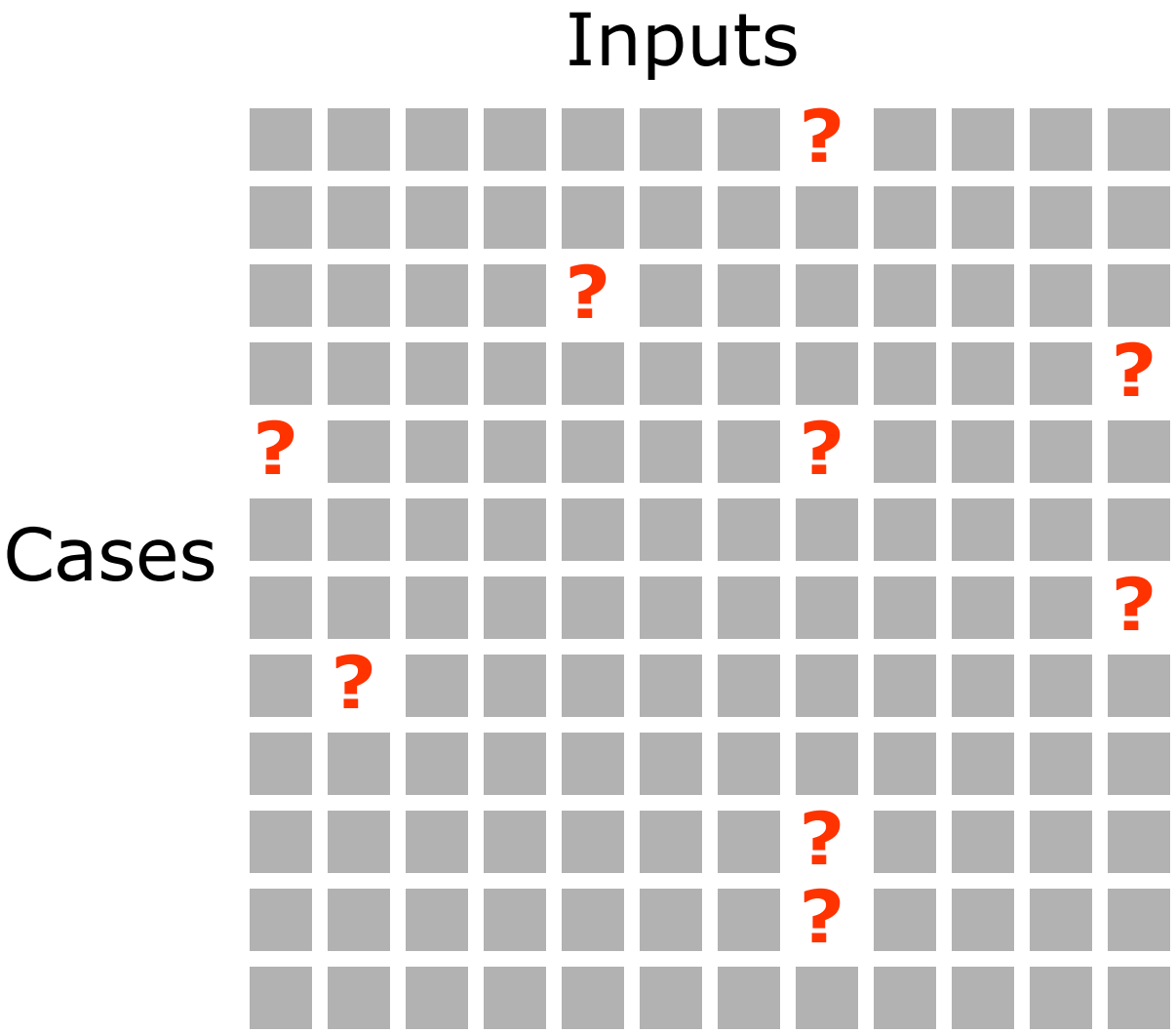
binary - two values
for example, the variable **MARITAL** that contains No or Yes

nominal - more than two non-numeric values, but no implied order
for example, **STATECOD** that contains AK, AL, AR, AZ, etc.

ordinal - more than two but not more than ten numeric values, with implied order
for example, **NUMCARS** that contains values from 0 to 3

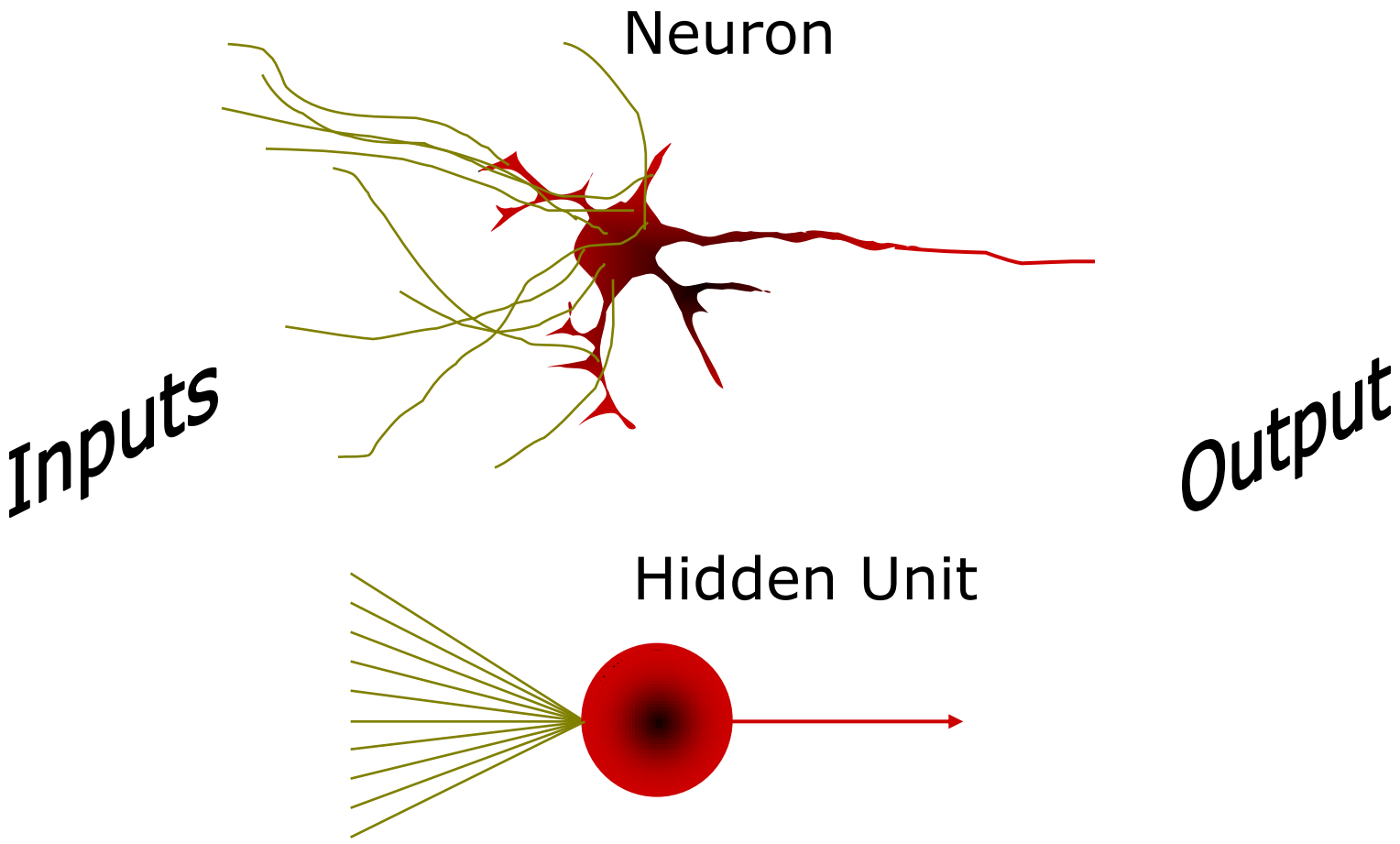
interval - more than ten numeric values
for example, **AMOUNT** that contains many different dollar values

Missing Value Imputation

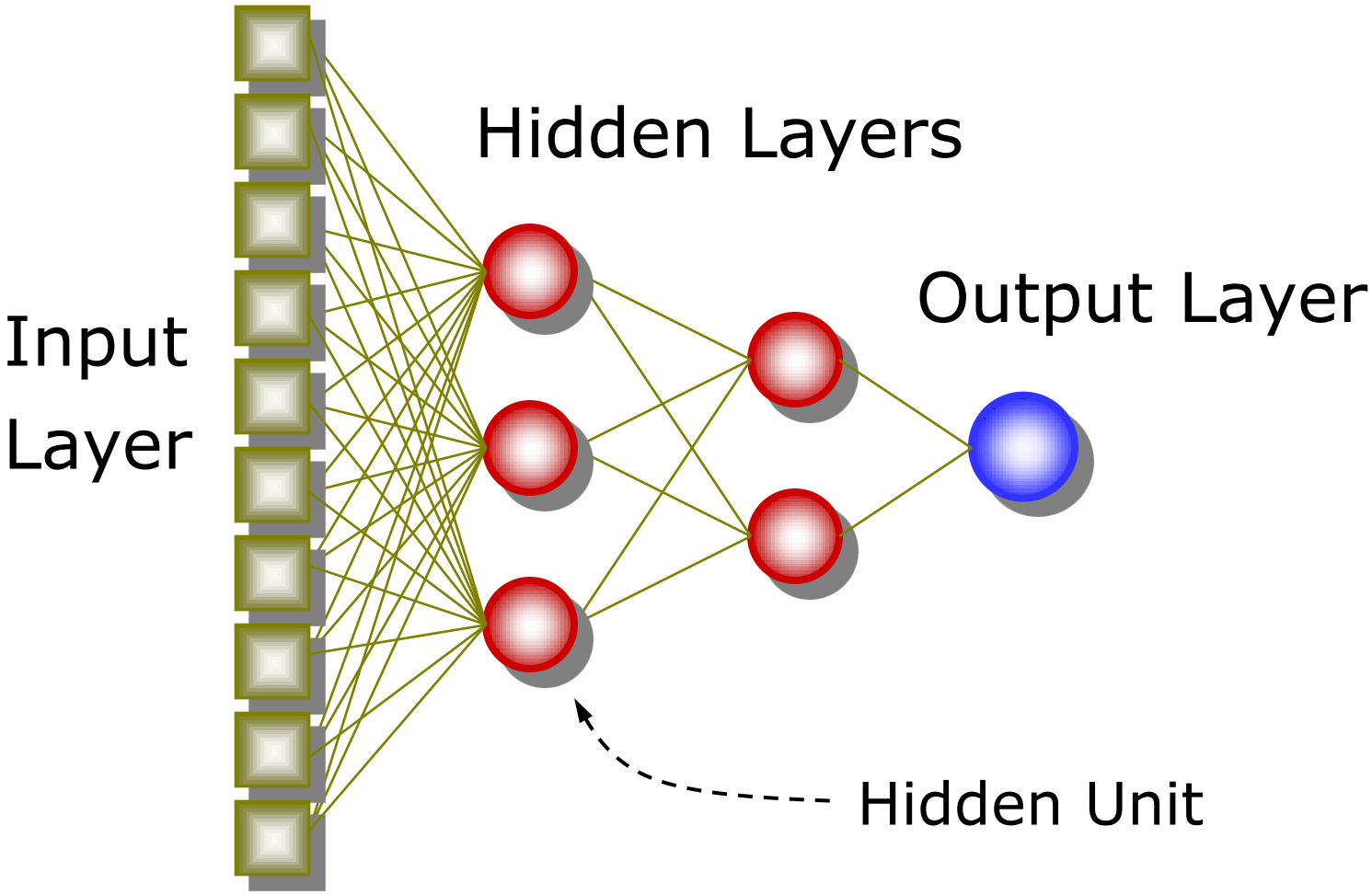


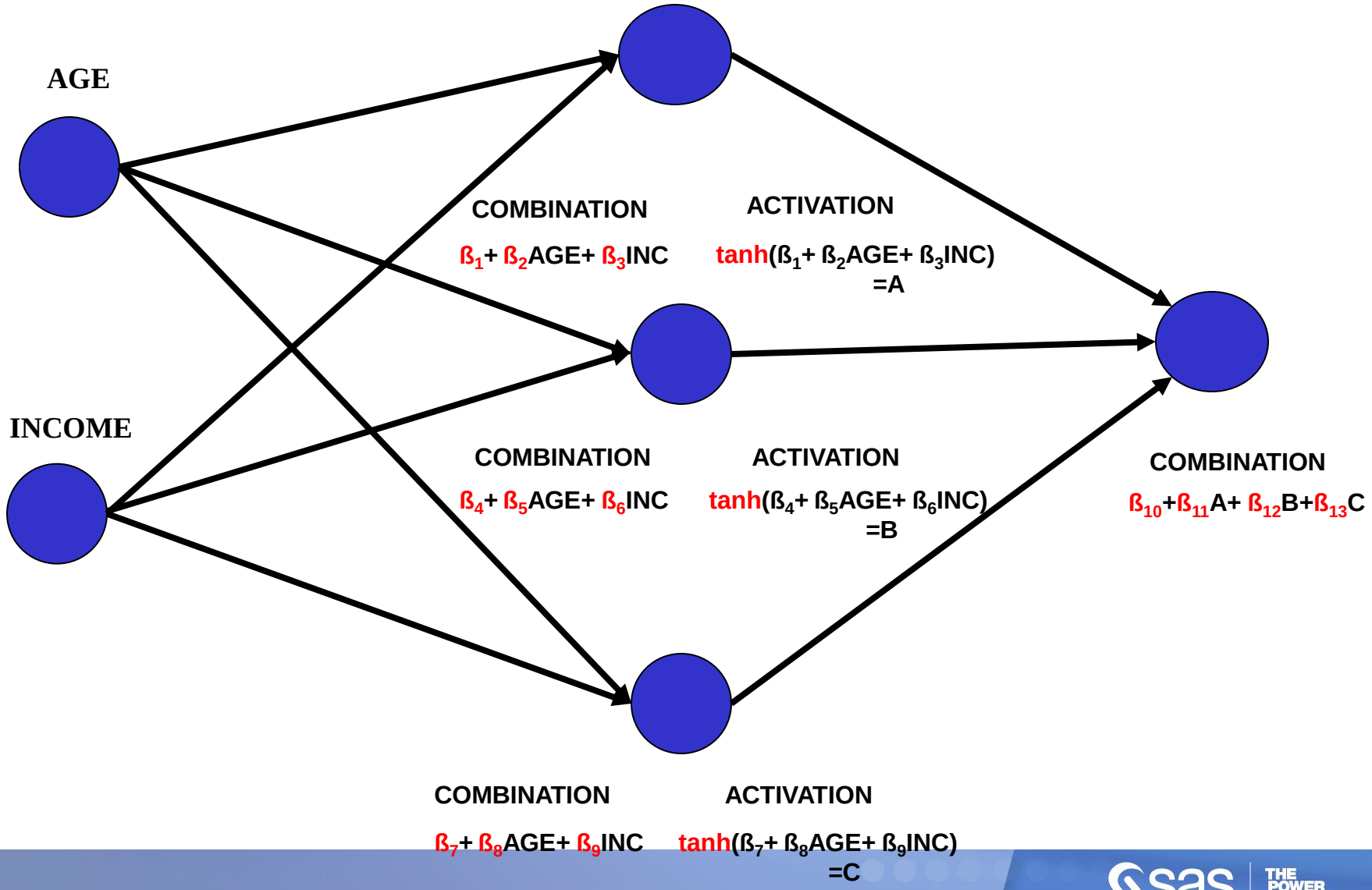
Réseau Neuronaux

Artificial Neural Networks

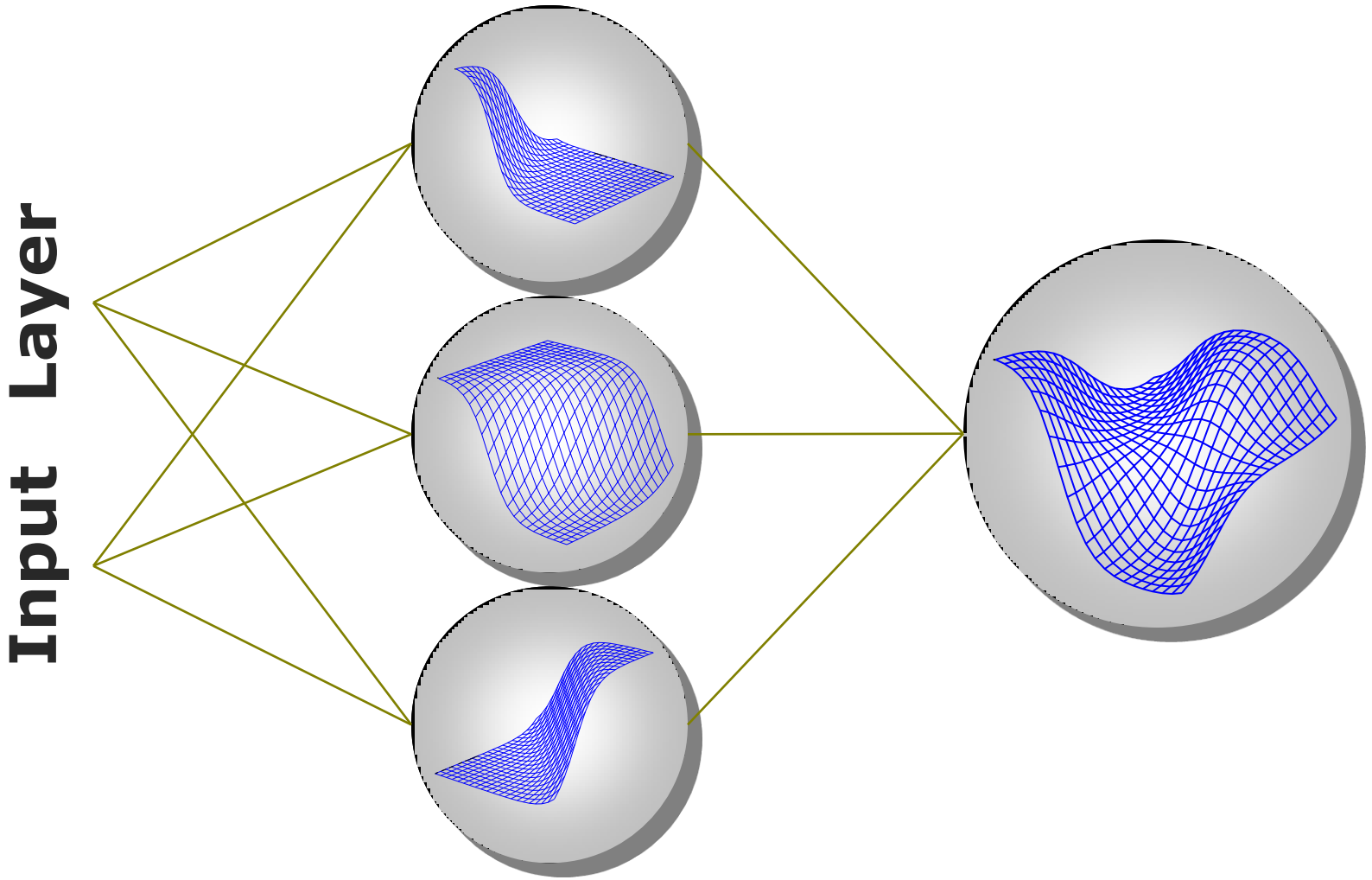


Multilayer Perceptron

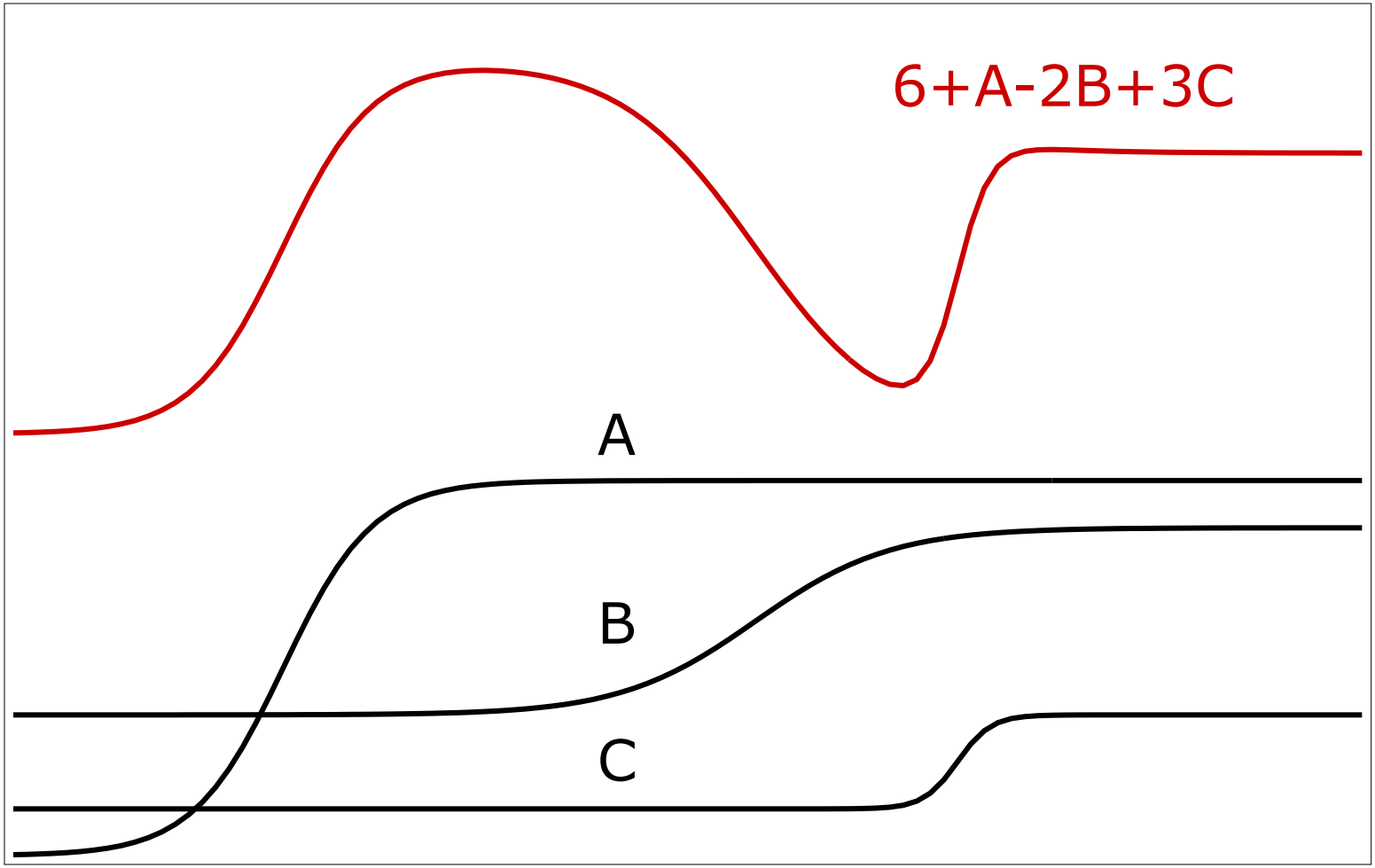




Activation Function



Universal Approximator



Association Rules

Rule	Support	Confidence
A ⇒ D	2/5	2/3
C ⇒ A	2/5	2/4
A ⇒ C	2/5	2/3
B & C ⇒ D	1/5	1/3