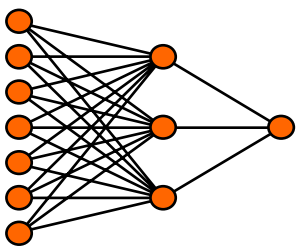
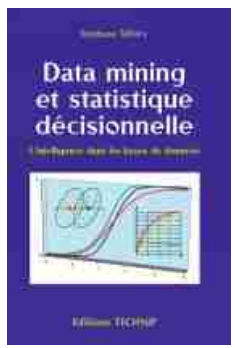
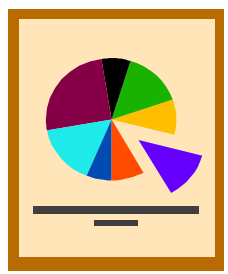
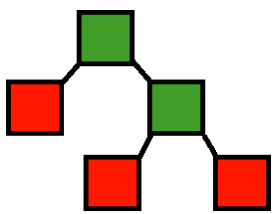
Stéphane Tufféry

DATA MINING
& STATISTIQUE DÉCISIONNELLE









Plan du cours

- Qu'est-ce que le data mining ?
- A quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs - Consulting
- L'analyse et la préparation des données
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- *Logiciels de statistique et de data mining*
- Informatique décisionnelle et de gestion
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

Logiciels de statistique et de data mining

Les logiciels de data mining et statistique

- Il existe de nombreux logiciels de statistique et data mining sur PC :
  - faciles à installer et pas très chers
  - avec des algorithmes de bonne qualité
  - généralement conviviaux
  - bons pour des PME car pouvant gérer plusieurs dizaines de milliers voire plusieurs centaines de milliers d'individus
  - **S-PLUS**™ de Insight, **Alice**™ de Isoft, **Predict**™ de Neuralware, **R** (version gratuite de S-PLUS) et les freewares **Weka** et **TANAGRA**...
- Cependant :
  - ils ne permettent pas de traiter exhaustivement de très grandes bases de données
  - ils ne mettent souvent en œuvre qu'une ou deux techniques (sauf quelques produits tels S-PLUS, R, Tanagra et Weka)

Zoom sur les gros logiciels

- Certains logiciels sont conçus :
  - pour exploiter de grands volumes de données
  - pour couvrir une large palette de techniques
- Ils existent parfois en version « statistique » ou « data mining » (le 2nd étant parfois une sur-couche du 1er)
- Ils peuvent fonctionner en mode client-serveur

- Il s'agit de *SPSS*™ et *Clementine*™ de SPSS
- Et de *SAS/STAT*™ et *Enterprise Miner*™ de SAS
- Et de *Statistica Data Miner*™ de StatSoft
- Et de *S-PLUS*™ et *Insightful Miner*™ de Insightful
- On peut ajouter *KXEN*™

Statistique vs Data mining

Appellation commerciale	**Logiciel de statistique**	**Logiciel de data mining**
Plate-forme	PC ou client-serveur	PC ou client-serveur
Interface	Fenêtre de programmation ou menus déroulants	Icônes à déplacer et relier par des flèches
Algorithmes	Ce qui sert couramment : moins les arbres de décision (existent parfois en logiciels spécifiques) sauf SPSS qui depuis sa version 13 contient un module d'arbres	Comme le « logiciel statistique » : moins certains algos de statistique (ex : tests non paramétriques) et d'analyse des données (ex : analyse factorielle discriminante) plus les arbres de décision, réseaux de neurones, détection d'associations plus parfois une gestion + performante des grosses bases de données
Prix	Un certain prix	Un prix certain !

Cartographie des logiciels

Logiciels multi-techniques	**Insightful – S-PLUS** **R (version gratuite de S-PLUS)** **Weka** **Univers. Lyon – TANAGRA**	**SAS/STAT** **SAS – Entreprise Miner** **SPSS** **SPSS – Clementine** **SPAD** **Statsoft – Statistica Data Miner** **Insightful – Insightful Miner** **KXEN** **IBM – Intelligent Miner**
Logiciels mono-techniques	**Salford Systems – CART** **Neuralware – Predict**	**Isoft – Alice** **(SPSS – Answer Tree)**
↑ Possibilités **Puissance →**	**Logiciels micros**	**Logiciels clients/serveurs**

Logiciels de data mining 1/3
(poids légers : dizaines de milliers de lignes)

Produit	Spécialité (le cas échéant)	Éditeur
Stat Lab		SLP InfoWare (Gemplus)
StartMiner	Réseaux de neurones – Arbres de décision	Grimmersoft
Alice	Arbres de décision	Isoft
Predict	Réseaux de neurones	Neuralware
NeuroOne	Réseaux de neurones	Netral
Wizwhy	Associations	Wizsoft
WEKA		« open source » (logiciel gratuit)
R		« open source » (initialement développé à l'Université d'Auckland, Nelle-Zélande)
DATALAB	Prétraitement des données	Complex Systems

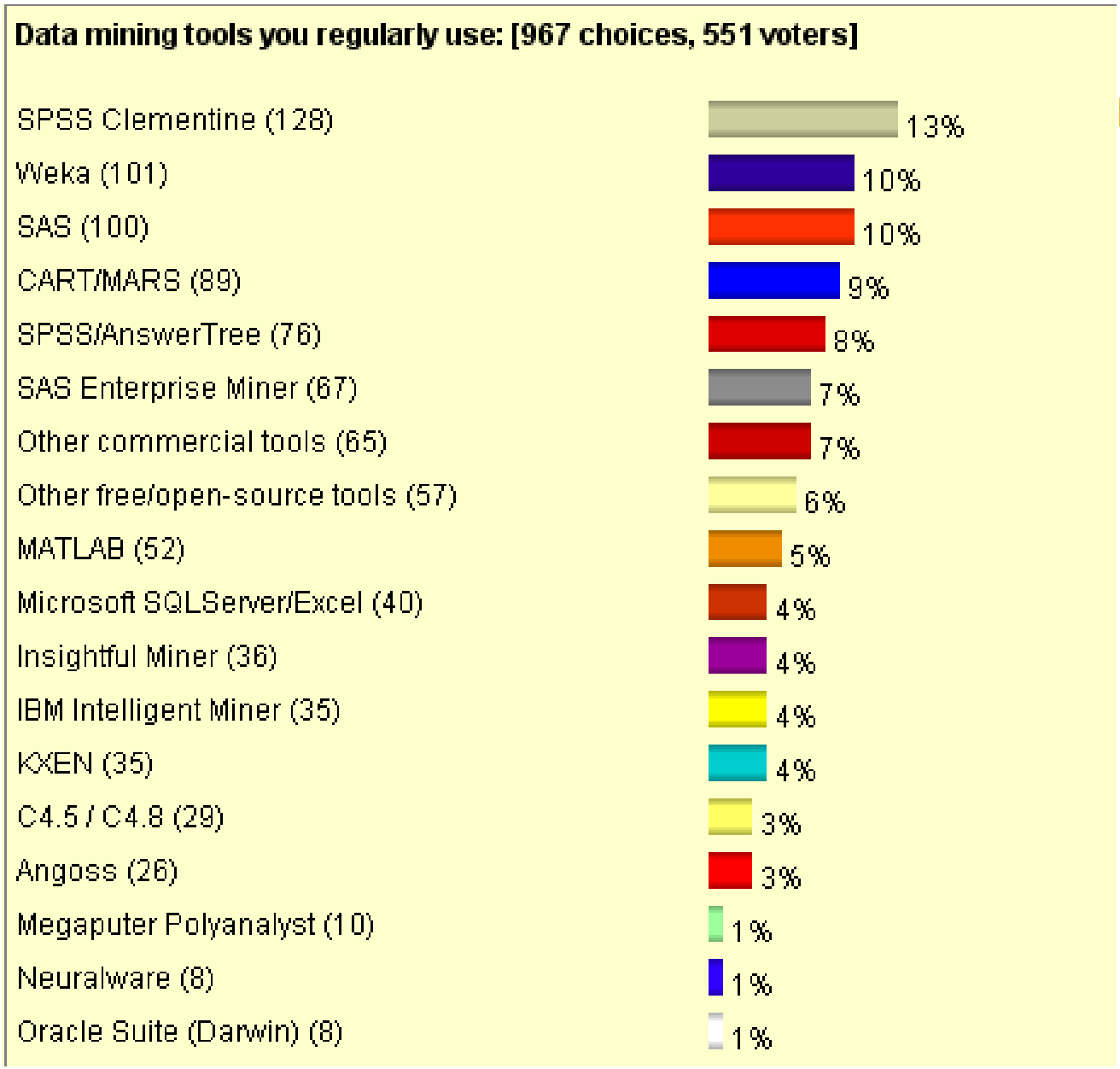
Logiciels de data mining 2/3
(poids moyens : centaines de milliers de lignes)

Produit	Spécialité (le cas échéant)	Éditeur
4Thought	Réseaux de neurones	Cognos
KnowledgeSEEKER	Arbres de décision	Angoss
KnowledgeSTUDIO		Angoss
C5.0 (Unix) See5 (Windows)	Arbres de décision	RuleQuest Research
Data Mining Suite		Salford Systems
CART	Arbres de décision	Salford Systems
Polyanalyst		Megaputer
S-PLUS		Insightful
TANAGRA		Laboratoire ERIC de l'Université de Lyon

Logiciels de data mining 3/3
(poids lourds : millions de lignes)

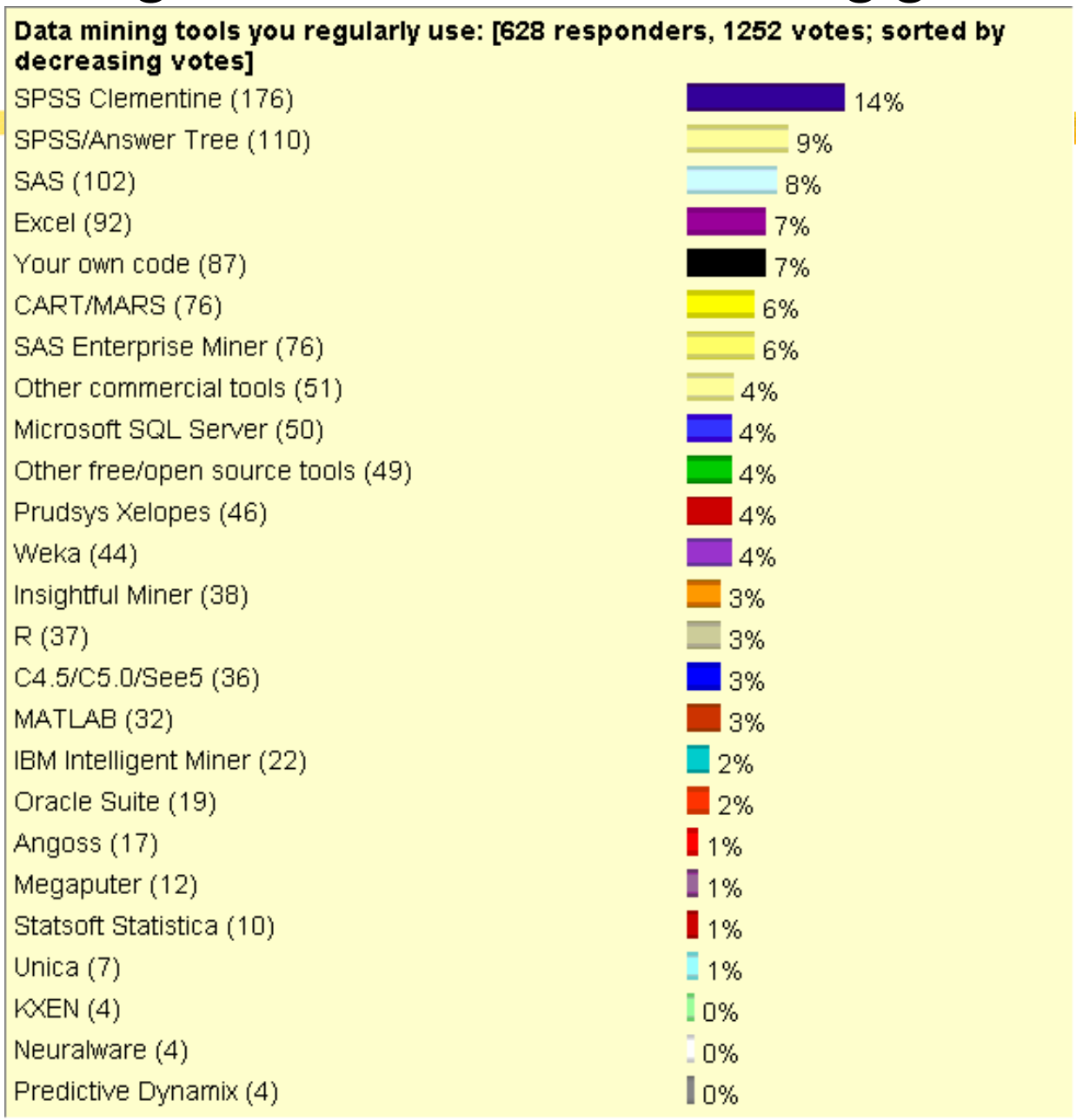
Produit	Spécialité (le cas échéant)	Éditeur
KXEN	Théorie de l'apprentissage de Vapnik	KXEN
Intelligent Miner	Classification relationnelle – Réseaux de neurones	IBM
Microsoft Analysis Services	Arbres de décision – clustering	Microsoft
Oracle Data Mining		Oracle
SPAD		SPAD
SPSS		SPSS
Clementine		SPSS
Statistica Data Miner		Statsoft
Insightful Miner		Insightful
SAS/STAT		SAS
Entreprise Miner		SAS

Sondage sur www.kdnuggets.com



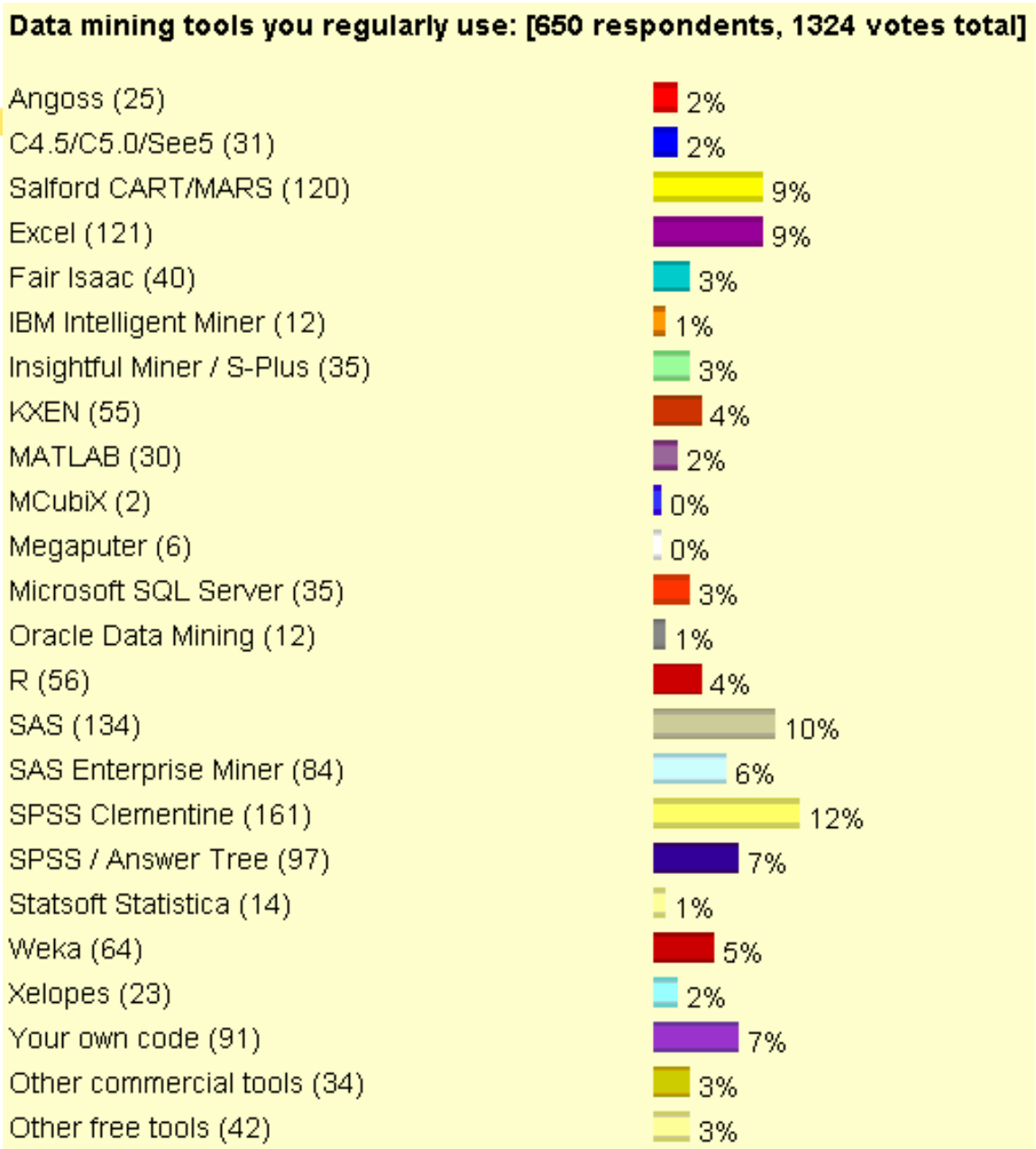
Sondage effectué en juin 2002

Sondage sur www.kdnuggets.com



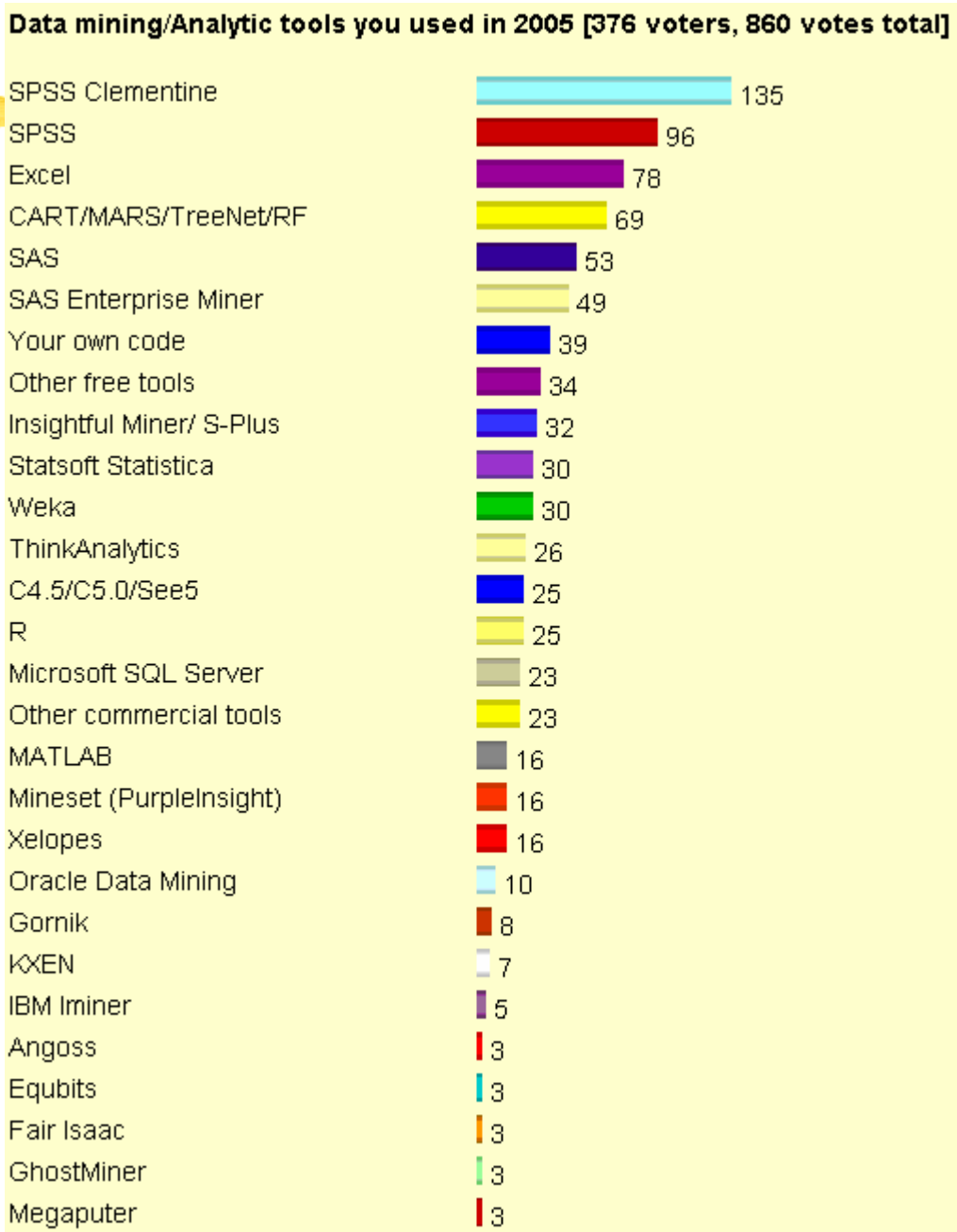
Sondage effectué en mai 2003

Sondage sur www.kdnuggets.com



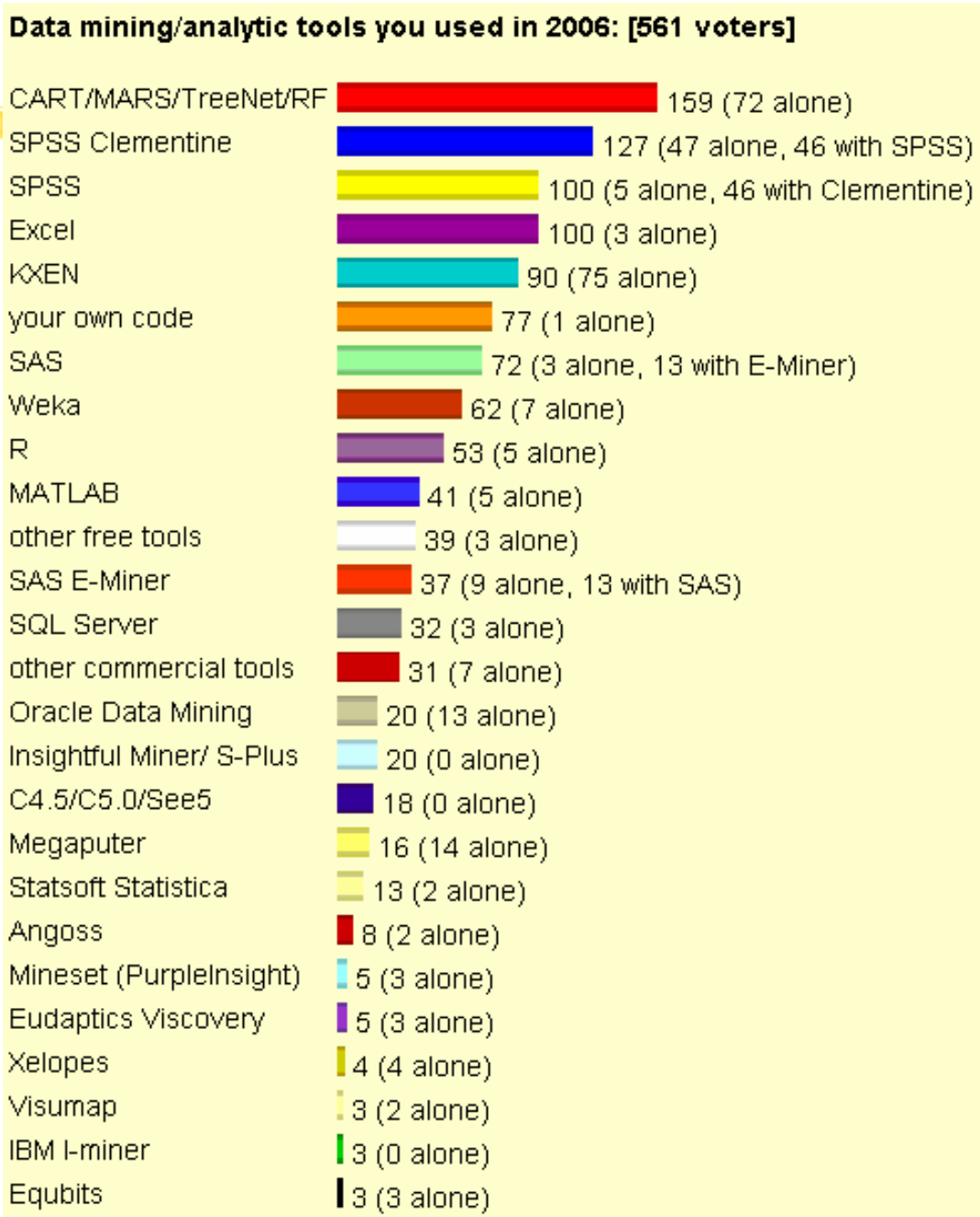
Sondage effectué en mai 2004

Sondage sur www.kdnuggets.com



Sondage effectué en mai 2005

Sondage sur www.kdnuggets.com



Sondage effectué en mai 2006

Critères de choix d'un logiciel

- Variété des algorithmes de data mining, de statistique et de préparation des données
  - + simple d'avoir tout dans un seul outil
- Qualité des algorithmes implémentés
  - documentation éditeur pas toujours accessible
- Capacité à traiter de grands volumes de données
  - peut être cruciale à partir de plusieurs centaines de milliers d'individus à traiter
- Types de données gérés
  - exemple : choix influencé si l'entreprise possède déjà un infocentre SAS...
- Existence d'un langage de programmation évolué
- Convivialité du logiciel et facilités à produire des rapports
- Prix !

Ce qu'on peut attendre d'un logiciel 1/5

- Algorithmes de statistique et de data mining :
  - classement (analyse discriminante linéaire, régression logistique binaire ou polytomique, modèle linéaire généralisé, régression logistique PLS, arbres de décision, réseaux de neurones, k-plus proches voisins...)
  - prédiction (régression linéaire, modèle linéaire général, régression robuste, régression non-linéaire, régression PLS, arbres de décision, réseaux de neurones, + proches voisins...)
  - classification (« clustering ») (centres mobiles, nuées dynamiques, k-means, classification hiérarchique, méthode mixte, réseaux de Kohonen...)
  - analyse des séries temporelles
  - analyse de survie
  - détection des associations

Ce qu'on peut attendre d'un logiciel 2/5

- Fonctions de préparation des données
  - manipulation de fichiers (fusion, agrégation, transposition...)
  - visualisation des individus, coloriage selon critère
  - détection, filtrage et winsorisation des extrêmes
  - analyse et imputation des valeurs manquantes
  - transformation de variables (recodage, standardisation, normalisation automatique, discrétisation...)
  - création de nouvelles variables (fonctions logiques, chaînes, statistiques, mathématiques...)
  - sélection des discrétisations, des interactions et des variables les plus explicatives

Ce qu'on peut attendre d'un logiciel 3/5

- Fonctions statistiques
  - détermination des caractéristiques de tendance centrale, de dispersion, de forme...
  - tests statistiques de moyenne, de variance, de distribution, d'indépendance, d'hétéroscédasticité, de multicolinéarité...
- Fonctions d'échantillonnage et de partition des données
  - pour créer des échantillons d'apprentissage, de test et de validation (l'échantillonnage stratifié doit être possible)
  - bootstrap, jackknife (validation croisée)
- Fonctions d'analyse exploratoire des données et d'analyse factorielle
  - ACP, ACP avec rotation, AFC, ACM
- Langage avancé de programmation
  - macros

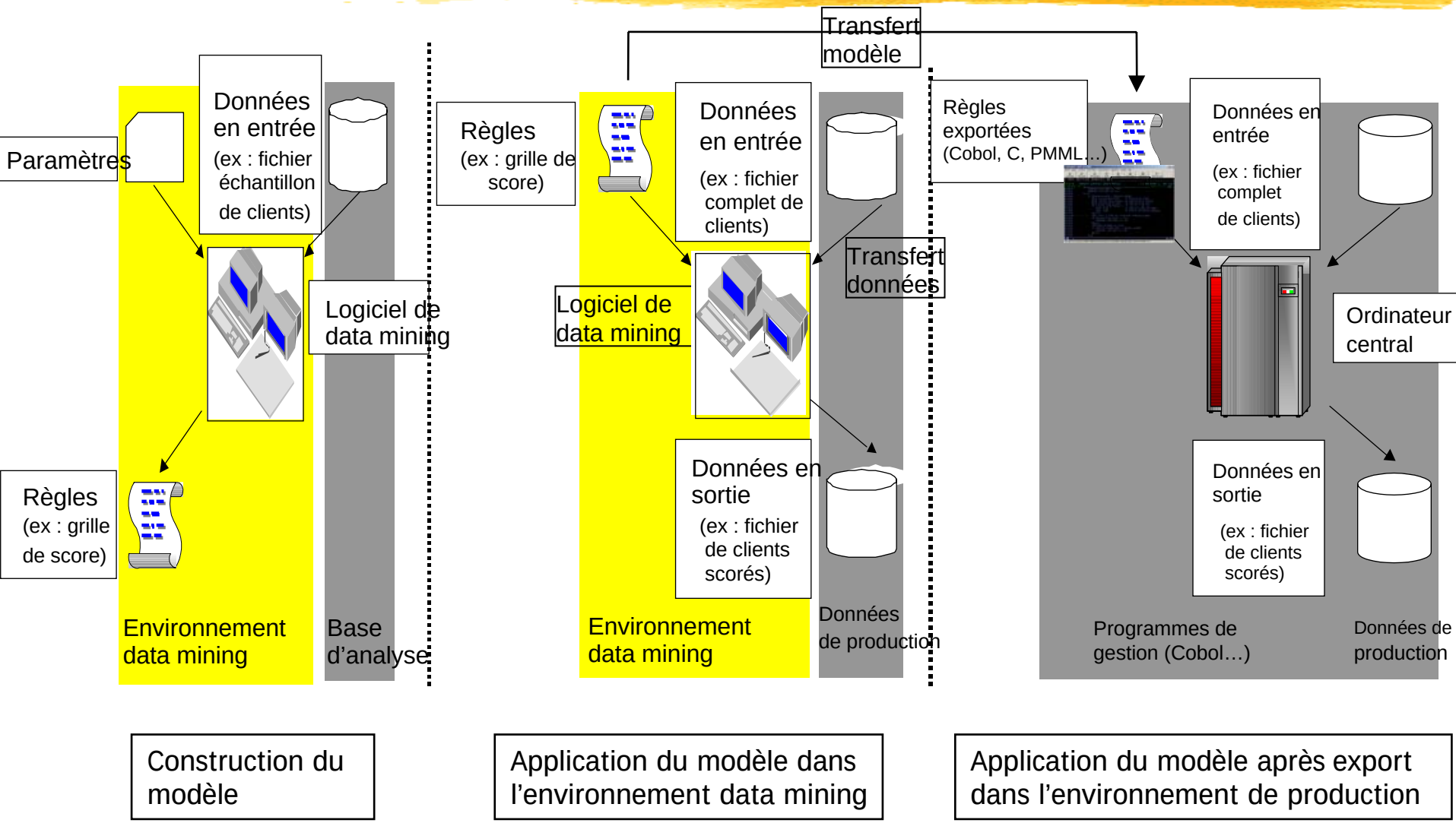
Ce qu'on peut attendre d'un logiciel 4/5

- Présentation des résultats
  - visualisation des résultats
  - manipulation des tableaux
  - bibliothèque de graphiques (2D, 3D, interactifs...)
  - navigation dans les arbres de décision
  - affichage des courbes de performances (ROC, lift, gain...)
  - indice de Gini, aire sous la courbe ROC
  - facilité d'incorporation de ces éléments dans un rapport
- Gestion des métadonnées
  - variables définies identiquement pour tous les fichiers du projet (identifiant, cible, exclusions...)
  - définition de groupes de variables

Ce qu'on peut attendre d'un logiciel 5/5

- Plates-formes supportées (Windows, Unix, Sun, IBM MVS...)
- Formats d'entrée/sortie des données gérés :
  - tables Oracle, Sybase, DB2, SAS, fichiers Excel, à plat...
- Enchaînements programmés de plusieurs algorithmes
- Volume de données pouvant être raisonnablement traité
- Pour plus de puissance
  - architecture client-serveur : calculs sur le serveur et visualisation des résultats sur le client
  - algorithmes parallélisés
- Exécution en mode interactif ou différé
- Portabilité des modèles construits (C, XML, Java, SQL...)

Utilisation d'un logiciel



Construction du modèle

Application du modèle dans l'environnement data mining

Application du modèle après export dans l'environnement de production

Critères de performance des techniques 1/5

- Nous récapitulons dans les tableaux suivants les principales techniques de data mining avec leurs conditions d'utilisation sur 3 points essentiels :
  - l'absence d'hypothèses restrictives fortes préalables à la recherche
  - la capacité de traiter exhaustivement les données, en un temps raisonnable
  - la possibilité de manier des données lacunaires et de types hétérogènes (numériques ou non)

Critères de performance des techniques 2/5 (classification)

Techniques	Absence d'hypothèses sur le problème à résoudre	Traitement exhaustif des bases de données	Traitement des données hétérogènes ou lacunaires
méthode des centres mobiles et ses variantes	non (nombre de classes et centres initiaux fixés)	oui	variables numériques et sans valeurs manquantes
classification hiérarchique	oui, mais les classes au niveau n sont déterminées par ceux au niveau n–1	non (algorithme non linéaire), on ne peut traiter plus de quelques milliers d'observations	oui (possibilité de traiter des variables non numériques avec une distance *ad hoc*)
classification neuronale (Kohonen)	non (nombre de classes fixé)	oui	les variables $\notin$ [0,1] doivent être transformées
classification relationnelle	oui	oui	variables qualitatives

Critères de performance des techniques 3/5 (classement/prédiction)

Techniques	Absence d'hypothèses sur le problème à résoudre	Traitement exhaustif des bases de données	Traitement des données hétérogènes ou lacunaires
analyse discriminante	non (relations linéaires entre les variables et hypothèses sur les lois conditionnelles X_i/Y)	oui	variables numériques et sans valeurs manquantes
analyse discriminante sur coordonnées factorielles d'une ACM (méthode DISQUAL)	oui (permet de s'affranchir largement des hypothèses sur les lois conditionnelles X_i/Y)	oui	oui (les valeurs manquantes sont traitées comme des valeurs à part entière)
régression linéaire, régression PLS	non (relations linéaires entre les variables + autres hypothèses)	oui	variables numériques et sans valeurs manquantes

Critères de performance des techniques 4/5 (classement/prédiction)

Techniques	Absence d'hypothèses sur le problème à résoudre	Traitement exhaustif des bases de données	Traitement des données hétérogènes ou lacunaires
régression logistique, modèle linéaire généralisé	oui	oui	oui (découper en classes les variables continues avec des valeurs manquantes)
arbres de décision	comme la classification hiérarchique (sorte « d'arbre à l'envers »)	non (mais moins vite limité que la classification hiérarchique)	certains arbres comme CHAID doivent discrétiser les variables continues
réseaux de neurones perceptrons	oui (mais il faut fixer le nombre de neurones cachés)	non (pas d'apprentissage sur plusieurs centaines de variables)	les variables $\notin$ [0,1] doivent être transformées
réseaux à fonction radiale de base	comme les perceptrons	oui	les variables $\notin$ [0,1] doivent être transformées

Critères de performance des techniques 5/5 (associations)

Techniques	Absence d'hypothèses sur le problème à résoudre	Traitement exhaustif des bases de données	Traitement des données hétérogènes ou lacunaires
recherche d'associations	oui	oui (sur une machine puissante)	oui
séquences similaires	oui	oui (idem)	oui

Diminuer les temps de traitement 1/3

- Travailler sur des fichiers structurés (SAS, SPSS, DB2...) plutôt que des fichiers à plat
- Limiter le fichier analysé aux variables utiles au traitement en cours (par des sélections judicieuses)
- Recoder les variables pour diminuer leur taille
- Bien définir la longueur des variables utilisées en la limitant au strict minimum (utilisation de l'ordre LENGTH en SAS)
- Avoir sur le disque dur un espace disponible > 4 fois la taille du fichier analysé
- Défragmenter au besoin le disque dur
- Augmenter la mémoire vive
- Attention si on accède à un réseau distant pour le fichier analysé ou pour l'espace temporaire de travail

Diminuer les temps de traitement 2/3

- (SPSS) Éviter les ordres EXECUTE inutiles
- (SPSS) Utiliser l'option PRESORTED dans les agrégations
- (SPSS) Épurer régulièrement le fichier « journal » (log)
- (SAS) Préférer BY à CLASS dans la proc MEANS
- (SAS) Si une requête utilise au moins 3 fois une variable dans un filtre WHERE ou un BY, créer un index sur cette variable qui optimisera le WHERE (rendant inutile la lecture complète de la table) et évitera de faire précéder le BY d'une procédure SORT plus coûteuse en temps
- (SAS) Utiliser l'option TAGSORT pour trier une grande table sur une petite clé (le gain peut dépasser 40%). L'option NOEQUALS stipule qu'il est inutile de conserver le même ordre dans la table pour les observations qui ont les mêmes valeurs pour toutes les variables du BY. L'option THREADS permet de paralléliser les calculs de tri sur une machine multiprocesseurs.

Diminuer les temps de traitement 3/3

- (SAS) Utiliser les formats, qui peuvent être définis de façon globale dans un fichier séparé et non définis pour chaque table comme dans SPSS, et qui permettent de remplacer de nombreuses modalités de variables par des codes plus compacts et économes d'espace disque
- (SAS) Faire le ménage dans le répertoire temporaire WORK aussi souvent que possible (PROC DATASETS LIB= WORK KILL NOLIST), car il n'est purgé automatiquement qu'à la fin de la session SAS
- Utiliser la compression pour diminuer l'espace disque occupé par un fichier
  - (SPSS) option par défaut
  - (SAS) s'écrit COMPRESS = YES si les variables sont majoritairement alphanumériques, et COMPRESS = BINARY si les variables sont majoritairement numériques