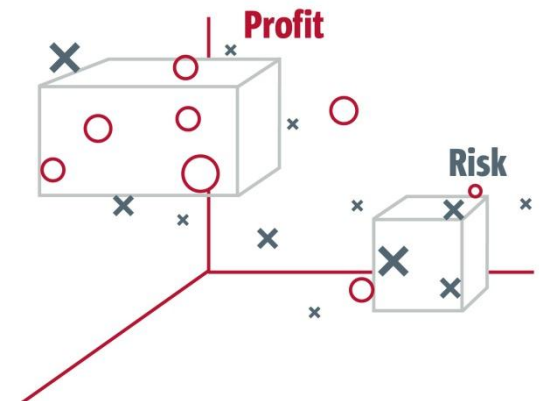


HyperCube

Complex Event Intelligence

Biologie et Big Data: comprendre la complexité du vivant

Yann Gaston-Mathé, Marc Damez-Fontaine
Conférence X-Biotech, 6 déc 2012



Plan

Vers une transformation du système de santé, portée par une triple révolution :

- Economique
- Technologique
- Techno-scientifique

L'émergence de l'analyse des données « big data » comme une compétence stratégique dans le monde de la santé

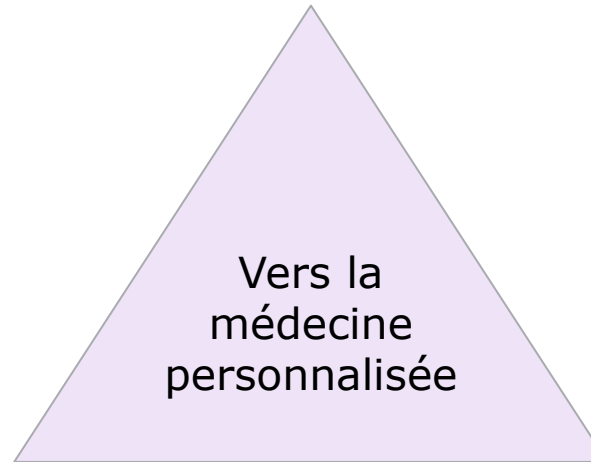
Panorama (rapide) des technologies de data mining

HyperCube, une technologie de data mining originale : principes et applications

Une époque charnière pour le monde la santé

Tendance économique :

- Explosion des coûts de la santé
- Pression de plus en plus forte sur les prix, les coûts, les remboursements
- Vers un système de plus en plus « dirigiste »



Tendance « techno-sociologique » :

- Informatisation, dossiers électroniques, interopérabilité
- Multiplication des données personnelles
- Vers le partage et l'intégration

Tendance « techno-scientifique » :

- Technologies « omiques »
- Vers le génome à 1000\$
- Vers la généralisation de l'utilisation des données génétiques dans la pratique clinique

L'évolution tendancielle des coûts de la santé n'est pas soutenable et va entraîner des changements majeurs dans l'organisation des soins

Le dilemme de la santé

Progress in science and technology brings improvement in clinical end point and quality of life...

... but at the expense of ever-increasing costs...

... and of persistent inefficiencies and sub-optimal situations in terms of public health

The cost-inflation effect is multiplied by the growing healthcare need

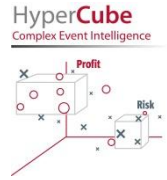
- Ageing population
- Social security / insurance systems
- Cultural change

Health costs are becoming progressively a major challenge for our societies

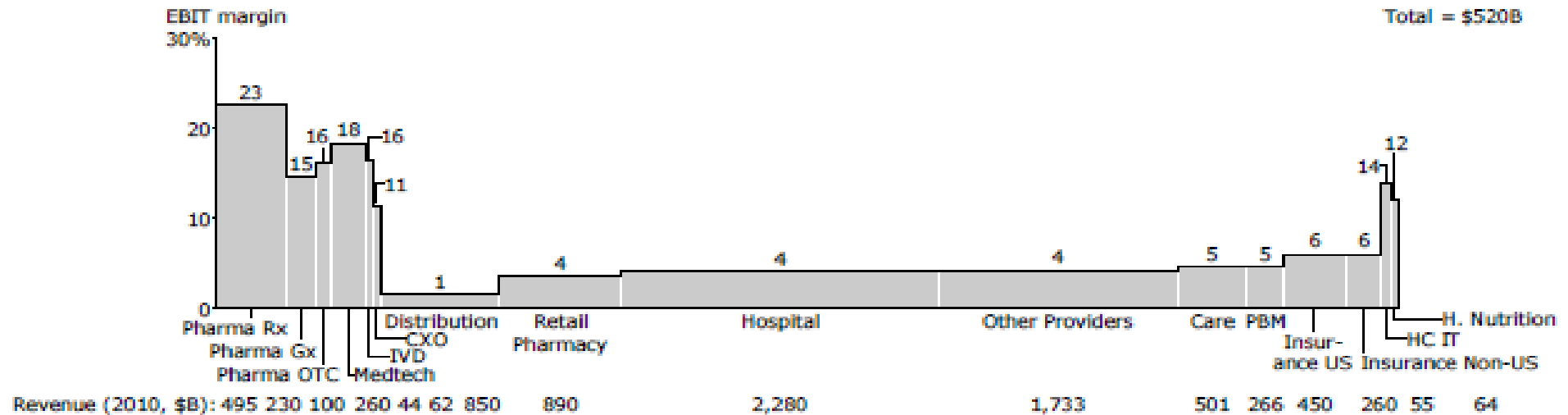
Unless healthcare productivity improves, our current social security model is threatened

The challenge is now to be able to combine progress in patient care with progress in overall productivity, either through new technologies or new way to manage patient care

La maîtrise des coûts de la santé est un enjeu majeur des années à venir



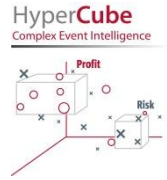
GLOBAL HEALTHCARE PROFIT POOL (2010)



Healthcare costs:

- \$B9,600 healthcare spending world wide
- 8-17% of GDP in OECD countries
- Growing fast in emerging countries
- Cost growth driven by demographics and medical/technological progress
- Publicly funded social security systems struggling to keep costs under control

Des transformations majeures vont avoir lieu, et sont déjà à l'oeuvre



L'augmentation des coûts de la santé est portée par :

- Population vieillissante
- Progrès médical synonyme d'augmentation des coûts

La tendance n'est plus soutenable

- Ralentissement structurel de la croissance
- Systèmes sociaux et comptes publics et privés en déficit

Les leviers existent

- De nombreuses inefficiences dans la gestion des soins résultant de trois facteurs majeurs :
 - Coordination imparfaite entre les acteurs
 - Mauvaise utilisation de la connaissance
 - Connaissance toujours parcellaire de la réalité biologique et clinique

Les tendances pour l'avenir :

- pression sur les acteurs du système de santé, réduction de leur marge de manoeuvre (« protocolisation »)
- Meilleure exploitation des données disponibles pour améliorer les décisions et les recommandations (Evidence-based medicine)
- Génération et exploitation des données biologiques (omiques) pour franchir un pas vers la connaissance et la médecine personnalisée

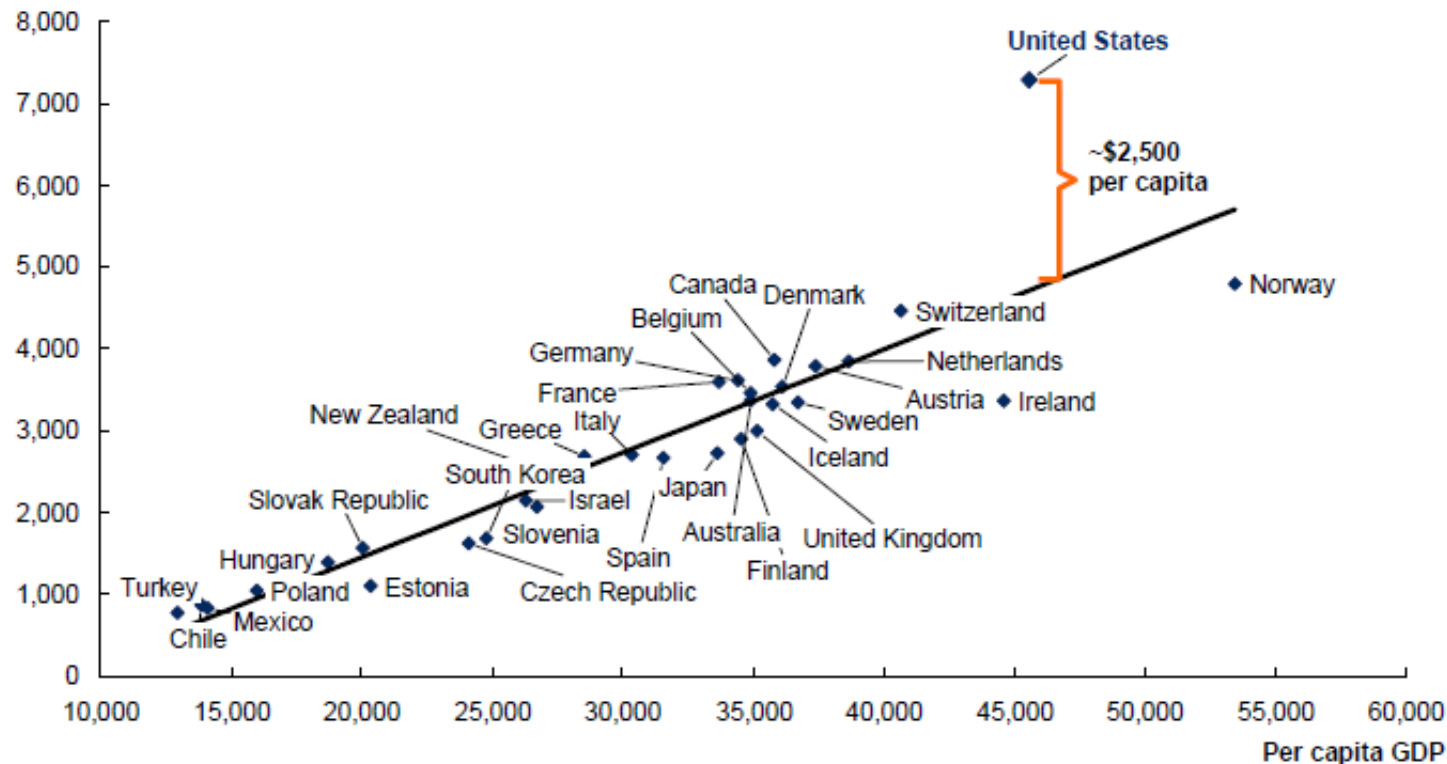
Exemple : potentiel d'économie sur les coûts de la santé aux US

Exhibit 13

A comparison with OECD countries suggests that the total economic potential for efficiency improvements is about \$750 billion

Per capita health expenditure and per capita GDP, OECD countries, 2007
\$ purchasing power parity (PPP)

Per capita health expenditure



SOURCE: Organisation for Economic Co-operation and Development (OECD)

Exemple : Alzheimer, une épidémie en devenir

Every 68 seconds, another American is diagnosed with Alzheimer's disease.

The coming wave of older people with dementia will place an unprecedented emotional and economic burden on society (see "[The Dementia Plague](#)").

We're living longer thanks in part to advances in medical technology, but we're not necessarily living better.

Approximately 80 million baby boomers are beginning to approach the years when symptoms of Alzheimer's disease emerge.

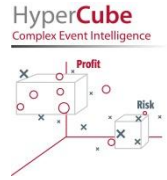
The risk for those 65 and older is 10 percent

It nears 50 percent by age 85.

A report on the global economic impact of dementia by Alzheimer's Disease International estimated the 2010 costs of care at \$604 billion.

Quelle médecine demain ?

La convergence de la révolution génomique et de l'optimisation de la gestion des soins



The revolution of disease management

- Ever-growing healthcare costs are not affordable anymore
- We need to optimize the cost of healthcare
- Towards systematic prevention and detection based on individualized risk assessment
- Towards systematic, real-life assessment of treatment effectiveness, and dynamic evolution of healthcare policies based on real life data

The revolution of genomics

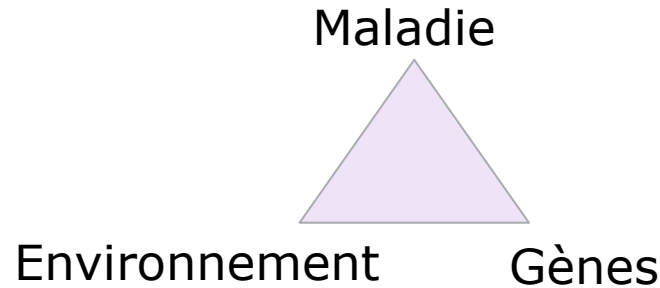
- The 1000\$ genome is at hand
- Emergence of related technologies: transcriptomics, proteomics, etc.
- Medical research is accelerating
- Towards a personalized patient care based on individual genome data

From standardized protocols to personalized medicine

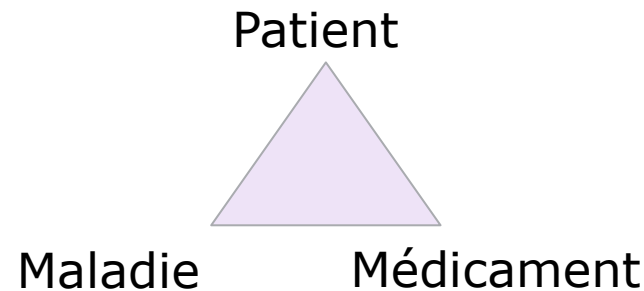
- The era of “one medicine for all” and major blockbusters is over
- Towards “The right treatment for the right patient”

La transformation de la santé sera portée par l'amélioration de la compréhension des phénomènes à 3 niveaux

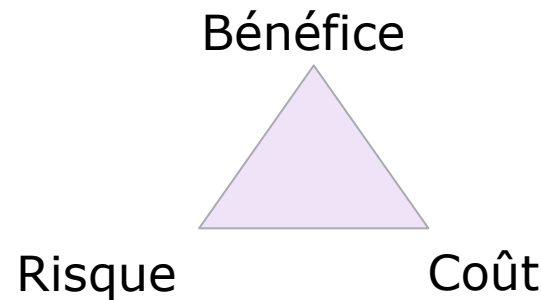
Science / Biologie :



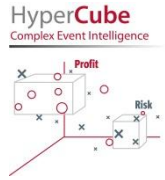
Clinique :



Système de santé :



Le paradigme de la médecine de demain proposé par Leroy Hood : P4 Medicine



Predictive

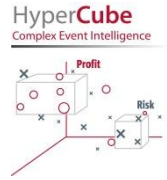
Preventive

Personalized

Participatory

Des changements importants sont
déjà à l'oeuvre

La pression politique, réglementaire et des payeurs sur le système de santé



Pression sur la gestion du risque

- Inflation des exigences sur la safety au moment de l'AMM
- Exigences pour la surveillance post-AMM :
 - FDA – Risk Evaluation and Mitigation Strategies (REMS)
 - EMA - Risk Management Plans (RMP)
 - France : Loi Bertrand

Pression sur les prix et les remboursements

- US: Health Technology Assessment, Pay for Performance
- UK: NICE
- France: HAS/ANSM : réévaluation des SMR, des taux de remboursement, des indications, retraits d'AMM, etc.

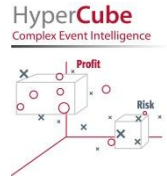
⇒ **Vers une réévaluation continue de la balance bénéfice-risque-coût**

L'efficacité clinique ne suffit plus :

Les Anglais ont émis un avis négatif pour l'abiraterone dans le cancer de la prostate
[Medscape France > Actualités ; 6 février 2012]

Cancer colorectal : 3 thérapies ciblées menacées de déremboursement au R-U
[Medscape France > Actualités ; 21 septembre 2011]

Des données supplémentaires demandées à l'industrie de la santé...



Renforcement de l'obligation de surveillance post-AMM du médicament via les RMPs et les REMS

- Etudes post-AMM
- Etudes non-interventionnelles / registres
- Pharmacovigilance

Maintien à jour en permanence du « core dossier » du médicament

- L'ensemble des données disponibles
- Les analyses correspondantes (notamment safety)

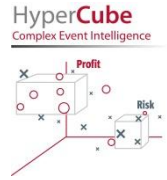
Reporting annuel vis-à-vis des autorités

- Safety : (PSUR)
- Résultats des études complémentaires

Réévaluation du rapport bénéfices-risques dans chacune des indications

VERS UNE CULTURE DE L'EVALUATION CONTINUE

...touchant de plus en plus à la question de l'impact pharmacoéconomique



La loi Bertrand est focalisée sur la gestion du risque

Mais par ailleurs, la pression sur les prix et le remboursement est de plus en plus forte

Répondre aux interrogations sur le rapport bénéfice-risque n'est pas suffisant

Il est nécessaire d'apporter en permanence des réponses concernant le positionnement du médicament dans la stratégie thérapeutique et le service médical rendu par rapport à la compétition

Il faut maintenant parler de rapport bénéfice-risque-coût

Mieux comprendre la réalité clinique et biologique : une nécessité pour optimiser la balance bénéfice-risque-coût

Au départ, une réalité clinique et biologique complexe...

- Hétérogénéité des situations et des populations
- Multi-causalité des phénomènes : génétiques, environnementales, individuelles, collectives...

...Qui est ignorée par les essais cliniques...

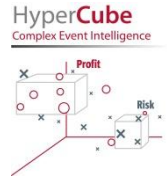
- Qui sélectionnent de petites populations homogènes dans un cadre très contrôlé
- Dont les analyses « post-hoc » sont souvent limitées

...Conduisant fréquemment à l'observation d'écarts importants avec les résultats observés dans la pratique clinique...

- Certains médicaments « déçoivent »
- Certains médicaments « marchent mieux »

...Et à une remise en cause par les Autorités ou les payeurs de la balance bénéfice-risque-coût

Les labos doivent maintenant pouvoir apporter des réponses à des questions clés :



Quelles sont les combinaisons de facteurs qui favorisent une efficacité optimale du médicament ?

Y-a-t-il des sous-populations qui répondent mieux (moins bien) ?

Y-a-t-il des associations de médicaments/traitements défavorables à l'efficacité ?

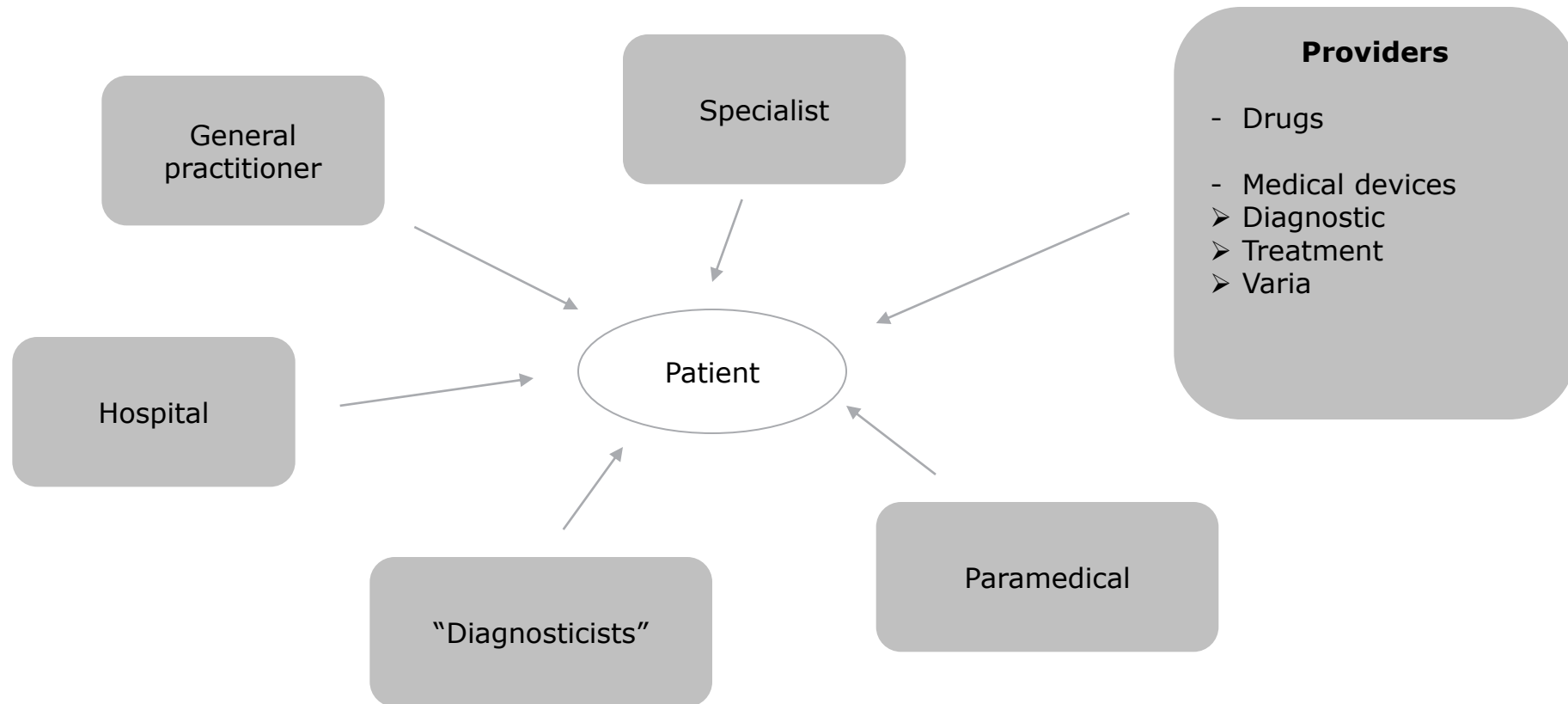
Dans le cas de survenue d'événements indésirables préoccupants, quels sont les profils des patients concernés ?

Dans quelles conditions le médicament agit-il favorablement sur le coût de prise en charge ?

Sur quels leviers agir pour optimiser la balance bénéfice-risques ?

L'informatisation des données de
santé, un enjeu majeur

Le système actuel est caractérisé par une faible coordination entre les différents acteurs de la santé

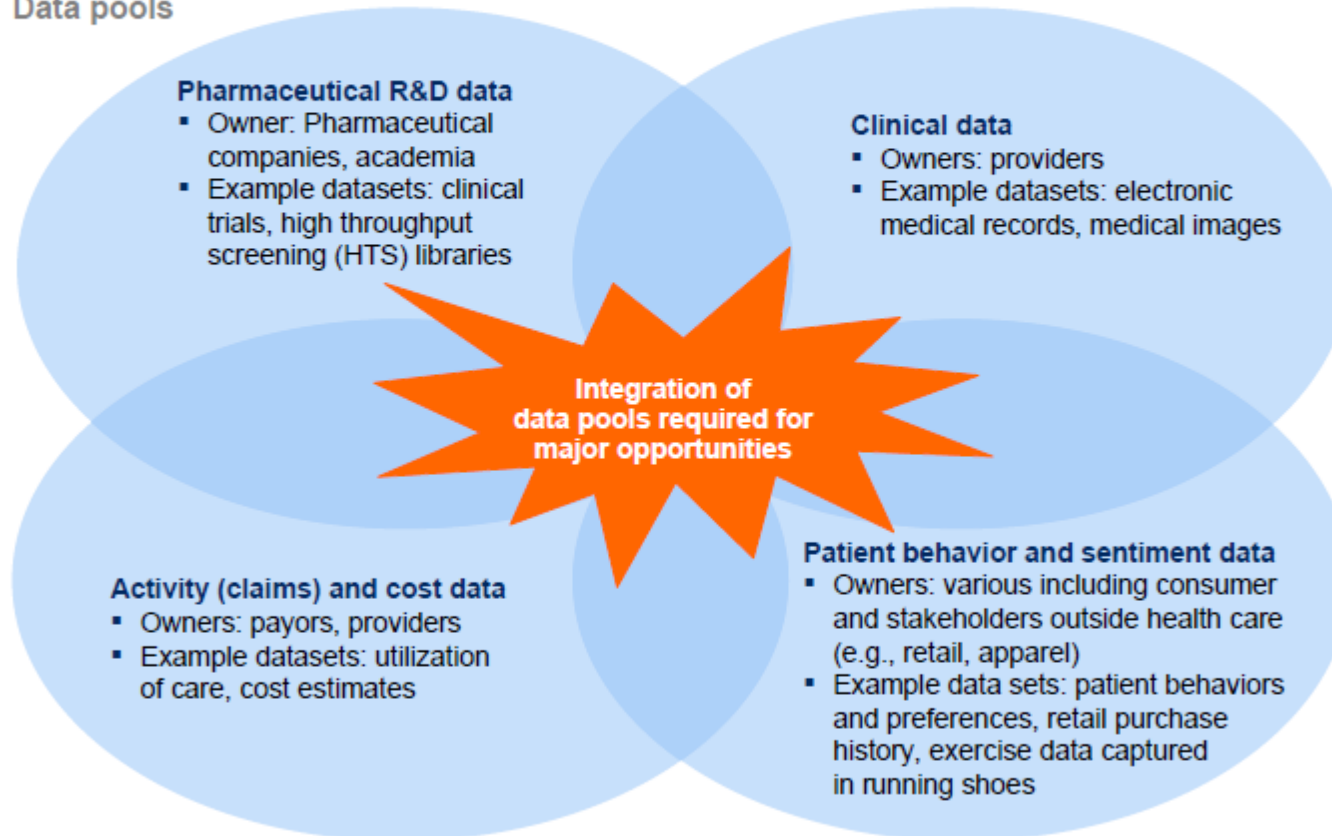


Des données éclatées (Exemple des US)

Exhibit 14

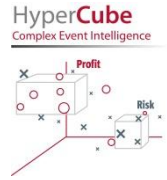
Four distinct big data pools exist in the US health care domain today with little overlap in ownership and low integration

Data pools

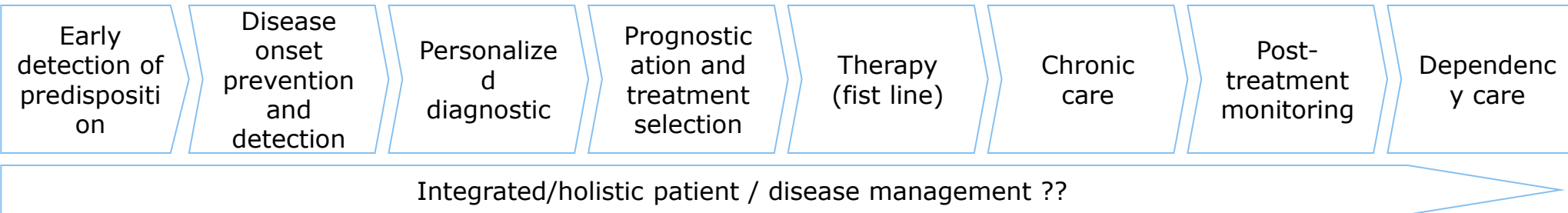


SOURCE: McKinsey Global Institute analysis

L'enjeu : optimiser les décisions tout au long de la vie du patient



Patient care value chain



Ce que cela suppose :

- En amont (recherche), la production de connaissances à partir des données consolidées et mutualisées
- En amont : la mise à jour et la diffusion permanente des connaissances
- En aval : la maîtrise des flux de données et la coordination effective des acteurs du système de santé

La consolidation des données devient un enjeu majeur

L'émergence des bases de données patients

Geisinger

Kayser

L'Islande

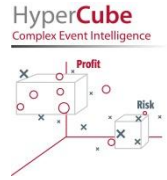
23andme

...

L'exemple de Geisinger

- **Geisinger : fournisseur de soins intégré aux US (Boston area)**
- **Fédère 800 médecins, 40 cliniques, labos d'analyse biologiques, laboratoires de recherche et enseignement**
- **Couvre 225,000 patients avec un « health plan »**
- **3,5 millions de patients dans un système EHR (Electronic Health Record)**
- **Base de données avec 14.000 génomes, utilisables pour de la recherche clinique**

Consulter votre dossier médical centralisé sur votre mobile



Dans le cadre du programme américain visant à donner accès au **dossier médical électronique en ligne** :

Humetrix, une société américaine, a mis au point **iBluebutton**, un outil de **communication mobile à double-sens** entre les patients et les médecins.

Cette plate-forme mobile comprend deux applications : l'une pour le patient et l'autre pour le médecin. **Elle donne accès au dossier médical complet à tout moment au patient** en utilisant l'application depuis un smartphone ou une tablette. **Quand le patient rencontre son médecin, il peut transférer en temps réel son dossier médical électronique sur la tablette de son médecin**

Ce système fonctionne également dans l'autre sens et permet au médecin ou à l'hôpital de transférer au patient examens, comptes-rendus ou images radiographies.

Cette technologie vise à la fois à **améliorer la qualité et le suivi des soins** et à **maîtriser le coût collectif des dépenses de santé** qui sont deux fois plus élevées par habitant aux Etats-Unis qu'en France.

L'émergence de la médecine moléculaire

L'évolution spectaculaire de la technologie du séquençage

1953 : découverte de l'ADN

1983 : invention de la PCR

- Début du séquençage gène à gène

1990 : Human Genome Project

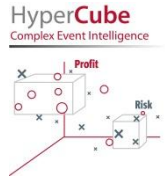
- 13 ans
- 3 milliards de \$



Demain : le next-generation sequencing :

- Une société anglaise, « Oxford Nanopore » a présenté un séquenceur d'ADN qui peut décoder un génome humain complet en moins d'une demi-heure

La croissance sans fin du volume et de la variété des données de biologie moléculaire (1)



Le génome : la séquence de l'ADN

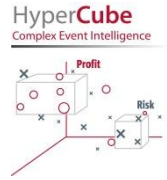
- Du séquençage gène à gène
- Au génotypage (Genome-Wide Association Studies)
 - De quelques dizaines de milliers de Single Nucleotide Polymorphisms
 - A des puces de plusieurs millions de SNPs
- A l'analyse des nombres de copies de gènes (CNV)
- Au séquençage d'exome
- Au séquençage de génome complet (2,3 milliards de bases)

Le transcriptome : la caractérisation et la quantification des ARN messagers (23.000 gènes, une variabilité inter-organe et type cellulaire...)

L'épigénome : analyse de la méthylation de l'ADN et des histones

Les micro-RNAs, les si-RNA

La croissance sans fin du volume et de la variété des données de biologie moléculaire (2)



Le protéome : analyse, caractérisation, quantification des protéines de l'organisme

- Circulantes
- Membranaires
- Intra-cellulaires

Le microbiome : analyse, caractérisation, quantification des milliards de micro-organismes de la flore intestinale

Le métabolome : l'analyse des métabolites de l'organismes

- The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes.↓

L'analyse des variations inter-individuelles du répertoire immunitaire

Le connectome : l'analyse des connexions inter-neuronales

Etc...

Vers la médecine personnalisée : quelques exemples

A detailed study predicts the onset of diabetes and shows that treatment works

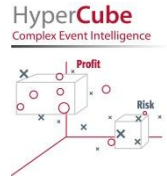
Forecasting Heart Attacks - A new test could tell doctors when danger is imminent

One startup, H3 Biomedicines, uses genome data to design drugs aimed at small groups of patients.

Foundation Medicine is offering a test that helps oncologists choose drugs targeted to the genetic profile of a patient's tumor cells.

Essai WINTHER : traitements ant-cancéreux fondé sur des biomarqueurs moléculaires

L'oncologie, le laboratoire de la médecine moléculaire personnalisée



Cancer du sein : l'analyse du génome des métastases peut guider le traitement
[Medscape France > Actualité ; 5 octobre 2012]

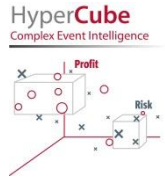
La thérapie ciblée du cancer à l'épreuve dans l'essai SHIVA
[Medscape France > Actualités ; 14 septembre 2012]

Thérapies ciblées : 61 000 cancéreux ont eu des tests moléculaires en 2010
[Medscape France > Actualités ; 12 janvier 2012]

L'oncologie personnalisée, un concept confronté à la dure réalité
[Medscape France > Point de vue ; 26 octobre 2012]

Scientists Identify 10 Breast Cancer Sub-types with DNA, RNA and micro-RNA analysis

Cancer : vers des traitements personnalisés grâce à la bioinformatique



WINTHER : essai clinique international pour proposer des traitements personnalisés à des patients qui, jusqu'à présent, ne pouvaient bénéficier que de traitements standards.

« A l'heure actuelle, pour l'immense majorité des patients le choix thérapeutique n'est pas encore guidé par les analyses génétiques mais repose sur des protocoles standards ayant peu évolués depuis des décennies »

200 patients atteints de différents types de tumeurs métastatiques pour lesquelles le dernier traitement n'a pas fonctionné.

Des biopsies de la tumeur et du tissu sain seront réalisées pour mener des analyses des ADN et ARN

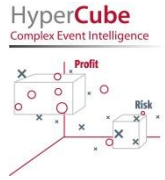
A partir de ces résultats, une liste de traitements sera proposée avec un "score prédictif d'efficacité pour les différentes thérapies".

Un système de score permet de prédire la réponse aux traitements existants sur le marché (standards ou ciblés, coûteux ou non) pour un patient donné.

Cet outil bioinformatique complexe intègre toutes les données connues sur les biomarqueurs, les anomalies moléculaires, les thérapies ciblées ou non, et des algorithmes prédictifs d'efficacité.

L'essai va comparer "la survie sans progression de la maladie" avec les traitements guidés par l'analyse bioinformatique à la survie avec les traitements standards.

Explosion des capacités de séquençage et du volume de données générées



DNA Sequencing Caught in Deluge of Data (NY times, Nov 30, 2011)

BGI

- “BGI, based in China, is the world’s largest genomics research institute, with 167 DNA sequencers producing the equivalent of 2,000 human genomes a day.”
 - BGI’s 500-node supercomputer processes 10 terabytes of raw sequencing data every 24 hours
 - BGI churns out so much data that it often cannot transmit its results to clients or collaborators over the Internet or other communications lines

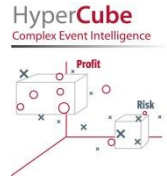
“Data handling is now the bottleneck,” said David Haussler, director of the center for biomolecular science and engineering at the University of California, Santa Cruz. **“It costs more to analyze a genome than to sequence a genome.”**

World capacity is now 13 quadrillion DNA bases a year

The amount stored by federal online archive of raw sequencing data has more than tripled just since the beginning of the year, reaching 300 trillion DNA bases and taking up nearly 700 trillion bytes of computer memory.

Le rôle clé de l'informatique dans
l'évolution de la médecine

La médecine assistée par ordinateur devient une réalité – Exemple :



Aux Etats-Unis, l'hôpital Cedars-Sinai de Los Angeles va utiliser le **supercalculateur Watson d'IBM pour aider les médecins à choisir les meilleures options thérapeutiques** pour les malades.

Watson va être **alimenté en informations par l'énorme base de données** de cet établissement hospitalier réputé dans le traitement du cancer.

Il va recouper et comparer cette masse immense de données et proposer aux médecins un diagnostic et une première approche thérapeutique.

Grâce à sa **mémoire gigantesque et à sa capacité d'apprentissage et d'inférence**, Watson devrait être capable de poser des diagnostics de maladies rares et peu rencontrées par les médecins

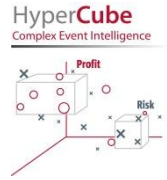
Il pourra également **intégrer une multitude de paramètres personnels liés au patients pour affiner ses propositions de traitements**, en tenant compte bien sûr des dernières avancées médicales et des interactions entre les médicaments envisagés.

Une avalanche de nouvelles données biologiques à classer, à recouper et à traiter :

- dans le domaine de la recherche : inventaire de tous les gènes, de toutes les protéines existantes et toutes les bactéries connues,
- dans le champ thérapeutique : nécessité de maîtriser toutes les interactions entre des milliers de molécules médicamenteuses

La médecine est confrontée à un nouveau défi : celui de la gestion et de l'utilisation intelligente de cette masse croissante d'informations.

La montée en puissance de la puissance de calcul continue



L'avènement, d'ici 2020 des superordinateurs exaflopiques.

- Puces CMOS combinant électronique et optique
- Connexions photoniques
- puissance de calcul phénoménale d'un milliard de milliard d'opérations par seconde, c'est-à-dire 1000 fois plus que les machines pétaflopiques les plus puissantes actuelles.

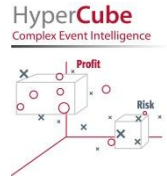
Une telle puissance de calcul permettra :

- de concevoir très rapidement de nouvelles molécules
- de simuler des interactions biologiques extrêmement complexes
- D'analyser les centaines de milliards de données biologiques individuelles

Cela aura un impact majeur dans la connaissance des systèmes biologiques et dans la mise au point de nouvelles approches thérapeutiques.

Analyser les données pour
comprendre les phénomènes, un
enjeu majeur

Exemple : 23andMe Expands its Data Mining Operations



The personal genetics company purchased disease-community website CureTogether earlier this week.

SUSAN YOUNG

Wednesday, July 11, 2012

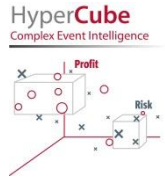
Personal genetics company 23andMe bought a patient social networking site called CureTogether earlier this week, a move that will expand the Silicon Valley-based company's ability to mine data from its patients for its own research.

CureTogether is an online forum with some 25,000 members who share information about their illness, and personal anecdotes about the effects of treatments and lifestyle choices on their symptoms.

23andMe sells DNA tests directly to consumers who send a spit sample to the company for genomic screening of up to 1 million genetic variants, which are a mix of both recreational (e.g. the genetic basis for tongue rolling) and serious medical traits (e.g. genetic factors associated with Parkinson's Disease) as well as variants that indicate a customer's ancestry.

The amount of the purchase was not reported and is 23andMe's first acquisition, according to the San Francisco Business Times.

Les phénomènes humains et biologiques sont complexes



Grande diversité de situations

Populations hétérogènes

Multi-causalité des phénomènes

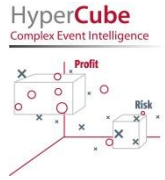
- Génétiques vs Environnementales
- Liées à la maladie vs Liées au traitement
- Effets combinés

Nombre de variables explicatives potentiellement énorme :

- Génome : 23.000 gènes, 2 milliards de bases
- Transcriptome
- Proteome
- Epigénome
- Microbiome
- ...

De nombreuses données pouvant contribuer à l'explication ne sont pas collectées

La matière première est hétérogène et la qualité est parfois faible



Différentes sources

Problèmes d'accessibilité

Formats hétérogènes

Données manquantes

Données mal renseignées

Taux d'erreur y compris sur les technologies omiques

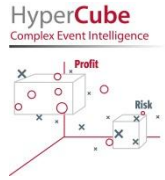
1/1000 sur les séquences d'ADN

Beaucoup plus sur de nombreux autres paramètres

Influence du contexte sur les données

« On ne parle pas des trains qui arrivent à l'heure » : prédominance des données portant sur les « cas remarquables »

Maîtriser cette complexité est l'enjeu de la recherche de demain



Les contraintes :

- Hétérogénéité des populations et des causes des phénomènes
- Phénomènes liées à des interactions
- Phénomènes probabilistes
- Beaucoup de variables, peu d'observations (high dimensionality data)
- Des données de qualité hétérogènes, des données manquantes, du bruit
- Des données de type et de natures différentes

Difficultés et enjeux :

- Volumétrie, puissance de calcul nécessaire
- Détection des « signaux faibles », surtout les effets résultants d'interactions entre variables
- Capacité à généraliser les résultats (surapprentissage)
- Interprétabilité des résultats
- Capacité à prédire

Les technologies de data mining et
d'intelligence artificielle

Quelques définitions

Machine learning

L'**apprentissage automatique** (*machine learning* en anglais), un des champs d'étude de l'intelligence artificielle, est la discipline scientifique concernée par le développement, l'analyse et l'implémentation de méthodes automatisables qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques.

(définition Wikipedia)

Data mining

L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, *data mining*, ou encore extraction de connaissances à partir de données, « ECD » en français, « KDD » en anglais, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

(définition Wikipedia)

L'apprentissage

Apprentissage humain

raisonnement logique, inductif

Apprentissage cognitif

mémorisation, comparaison, association

Apprentissage automatique

algorithmes, système expert

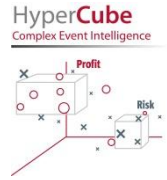
Apprentissage statistique

reconnaissance de formes

Prédiction vs extraction de connaissance

(modèle pour reproduire) vs (modèle pour expliquer)

Les grandes familles de technologies de fouille et d'analyse de données



L'apprentissage non supervisé (clustering)

Les méthodes fondées sur la régression statistique

Les méthodes graphiques / connexionnistes

- Les réseaux de neurones

- Les réseaux bayesiens

Les modèles informationnels

- Les arbres de décision

- Random forests

Les approches vectorielles

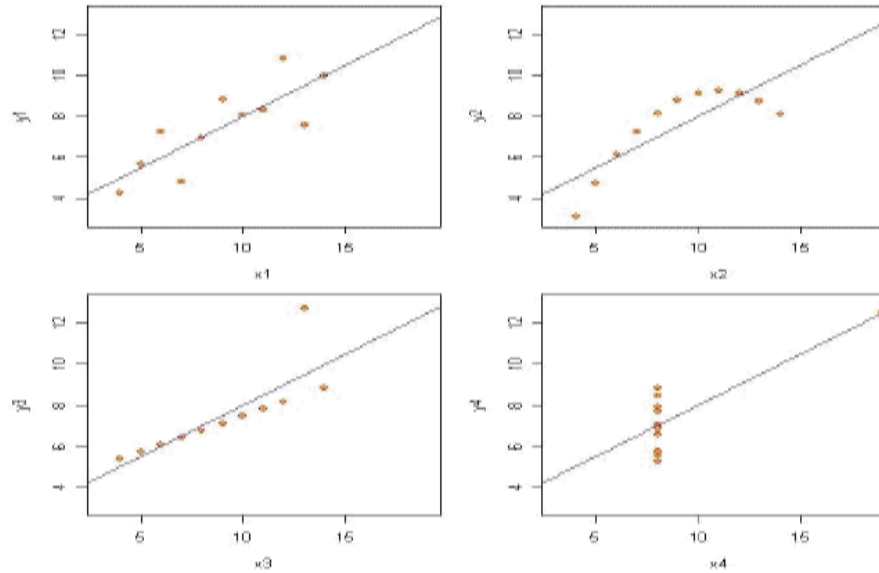
- Support Vector Machines

Les règles d'association (formal concept analysis, treillis de Gallois)

L'exploration hypercubique

Modèles de régression statistique

Concepts



Avantages

- **Globalement convergeant**
- **Simple à appliquer**

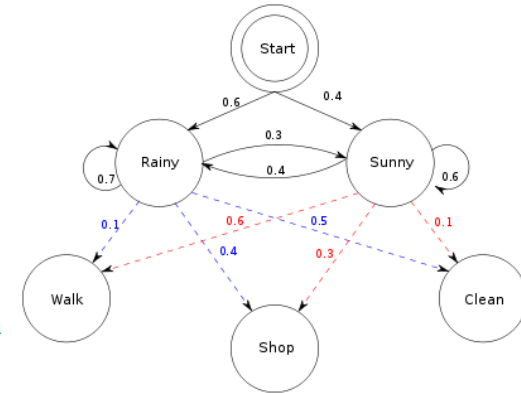
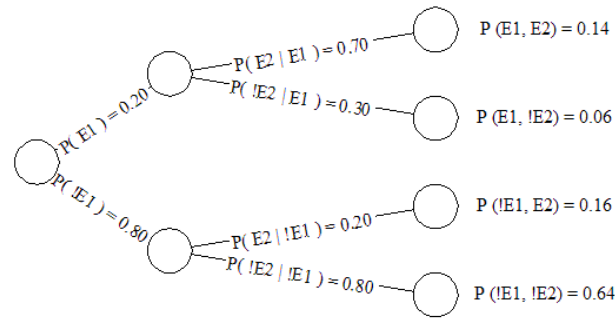
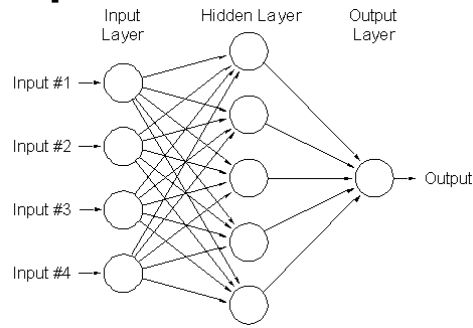
Inconvénients

- **Comportement des données reste inconnu**
- **L'expression d'une décision est une probabilité**

Méthodes connexionnistes / graphiques

Réseau de neurones, Réseau bayésien, Modèle markovien

■ Concept



■ Avantages

- Modélisation de séquences
- Traitement des phénomènes séquentiels
- Décisions multi-classes

■ Inconvénients

- Boite noire
- Apprentissage long
- Paramétrage difficile
- Convergence aux minima locaux

Méthodes vectorielles

SVM

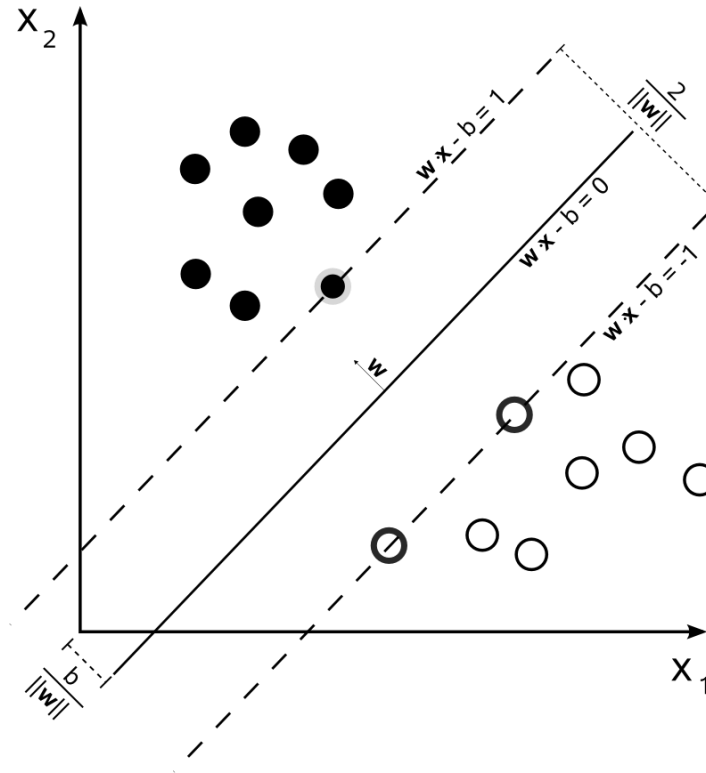
Concept

Avantages

- Heuristique intuitive
- Pas de minimum local
- Peu de paramétrage

Inconvénients

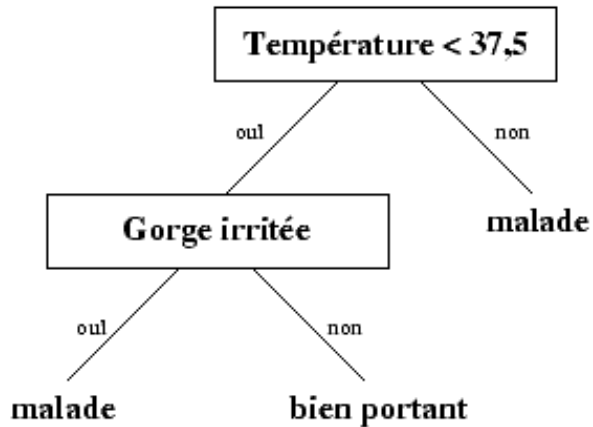
- Approche globale
- Boite noire
- Choix du noyau



Modèles informationnels

Concept

$$-\sum_{i=1}^N p_i \log_2(p_i)$$



Avantages

- Interprétables
- Facile à utiliser

Inconvénients

- Performances limitées

Apports et limites du data mining

Apport : capacité à « dénicher » des explications aux phénomènes qui sont inaccessibles aux approches statistiques classiques

Limites :

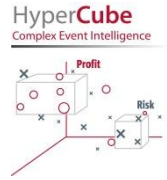
- Certaines lois de causalité échappent...
- Risque de « faux positifs » (sur-apprentissage, multiple testing)
- On ne trouve que ce qu'il y a dans les données que l'on a à disposition
- Problèmes liés à la qualité des données
- Nombreux cas de contradictions entre des résultats trouvés dans différents contextes (cas des études sur le rôle préventif de l'aspirine dans le cancer par exemple)

En pratique :

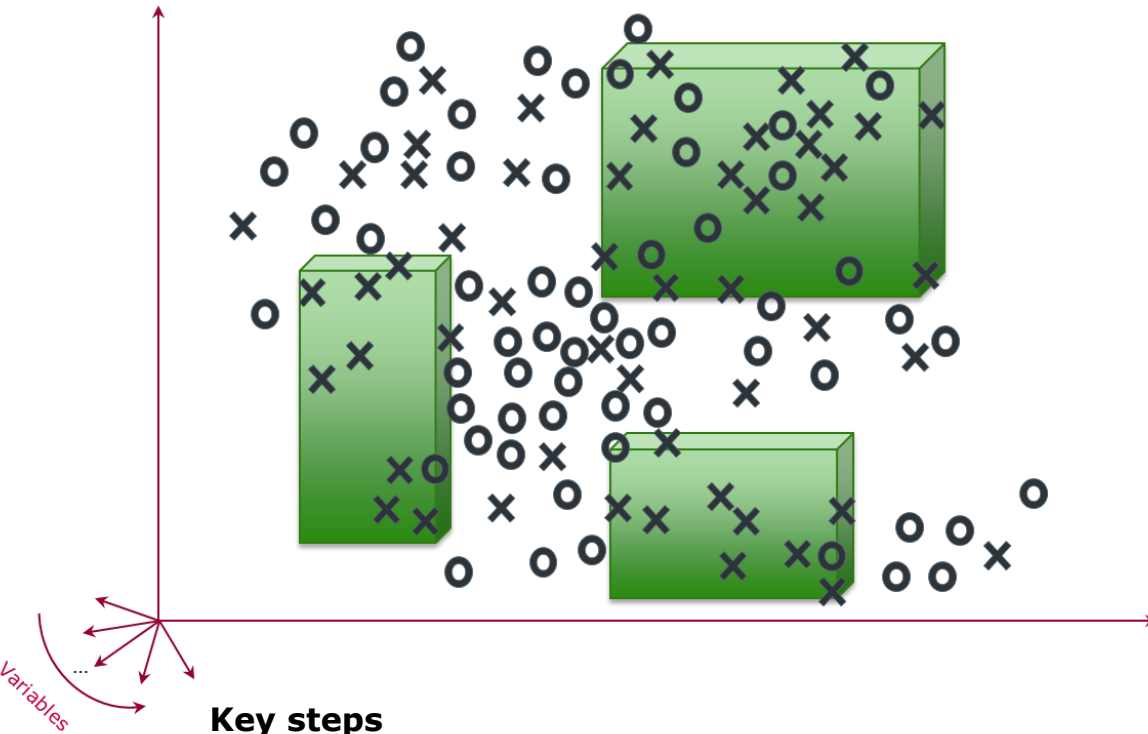
- Les autorités et les comités de revue sont encore suspicieux vis-à-vis des résultats trouvés par l'utilisation d'outils de data mining lorsque ces résultats ne sont pas « répliqués » dans un échantillon indépendant
- Ce qui est rarement possible en pratique dans le domaine médical
- L'utilisation du data mining reste insuffisamment développée par rapport à son potentiel de création de valeur

HyperCube, une approche originale
parmi les technologies de data mining
et d'apprentissage supervisé

HyperCube est une technologie de data mining profondément originale



Multi-dimensional space explored by Hypercube



Key steps

1. Explore space to identify local areas of over-density of phenomenon
2. For each identified area (hypercube):
 1. Select influent variables
 2. Find optimal limits
 3. Test statistical significance
 4. Outputs scientific rule

Hypercube key principle:

- Explore the multi-dimensional data space to identify local areas (hypercubes) of over-density of the phenomenon studied

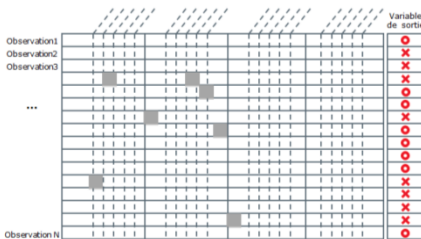
Key features:

- **No pre-established model or assumptions** on the distribution of variables or relationships between input and output
- **Systematic and exhaustive** exploration of the space, which provides **several strong local correlations** rather one weak global relationship
- Uncovers the **interactions between variables** which cause the phenomenon
- Uses **non-euclidean** mathematics i.e., can work with both continuous and discrete variables and is **tolerant to missing data**

L'approche Hypercube

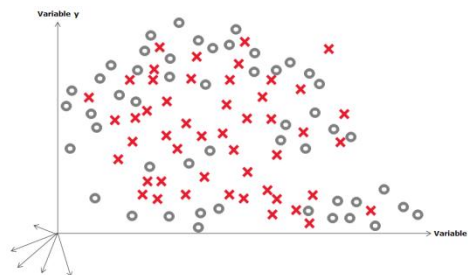
1

Data collection and structuring in a formatted matrix



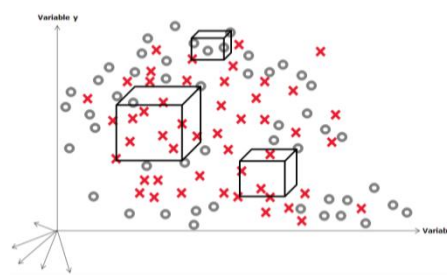
2

Creation of the n-dimensional space corresponding to the matrix



3

Hypercube definition and group selection



4

Analysis of the selected groups

1. Why?
2. Limits?
3. Stability?
4. Rules generation

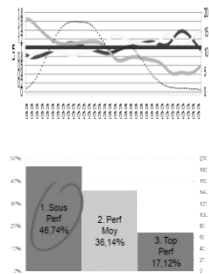
5

Rules expression

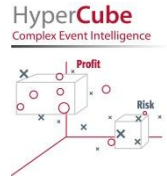
Rule 3: Patients with the following characteristics are 6.5 times more often "Good responders"

Size: 250 cases / 750
P-value: 5.10^{-6}

Variable	Limits
Age	[55;67]
Med_antecedent_Angor	[No]
Syst_Blood_Pressure_BL	[8;10]
Gene_XXXX	[1]
Gene_XXX	[2]
Conc_trt_Beta_blocker	[yes]



Les caractéristiques d'Hypercube en font un outil pertinent pour la recherche biomédicale



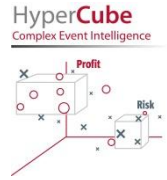
Characteristics of biological and clinical phenomena

- Multi-factorial causality (genetic, epigenetic, environmental...)
- Redundancy of pathways
- Heterogeneity of patient populations
- High number of variables especially with the emergence of genomics and proteomics
- Variables are of different types (continuous, binary, discrete)
- Data quality and homogeneity often insufficient (missing data, heterogeneous data format e.g., meta-analysis)

Characteristics of Hypercube technology

- Finds interactions between variables that are linked to the outcome
- Identifies hard-to-find segments with high-impact behaviour better than standard statistical methods
- Finds real patterns in data without making any assumptions
- Can handle almost limitless number of variables
- Can handle both continuous and discrete variables
- Finds patterns even on incomplete data – handles missing data very well

Un exemple d'application : l'analyse des facteurs influant sur les épisodes de malaria (collaboration avec l'Institut Pasteur)



OPEN ACCESS Freely available online



An Exhaustive, Non-Euclidean, Non-Parametric Data Mining Tool for Unraveling the Complexity of Biological Systems – Novel Insights into Malaria

Cheikh Loucoubar^{1,2,3}, Richard Paul¹, Avner Bar-Hen^{2,4}, Augustin Huret⁵, Adama Tall³, Cheikh Sokhna⁶, Jean-François Trape⁶, Alioune Badara Ly³, Joseph Faye³, Abdoulaye Badiane³, Gaoussou Diakhaby³, Fatoumata Diène Sarr³, Aliou Diop⁷, Anavaj Sakuntabhai^{1,8*}, Jean-François Bureau¹

Objectives: Using Hypercube, detect unknown risk factors of malaria events. Validate findings. Compare Hypercube explanatory and predictive performance with statistical method and regression tree technology

Data set: cohort of 46,837 outcome events from 1,653 individuals with 34 explanatory variables

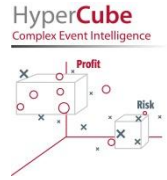
Results:

- A 4-variable rule generated by Hypercube enabled to identify a high-risk population with relative risk of malaria episode 3.71 higher than general population.
- All rules generated by Hypercube were successfully replicated in a second, independent cohort
- Hypercube enabled to identify a previously unknown risk factor, i.e., number of previous *P. malariae* episodes
- No other tool used enabled to generate results with as good level of relative risk as the rule generated by Hypercube

Authors conclusion: "Search of local over density in m-dimensional space, explained by easily interpretable rules, is thus seemingly ideal for generating hypotheses for large datasets to unravel the complexity inherent in biological systems."

Loucoubar C, Paul R, Bar-Hen A, Huret A, Tall A, et al. (2011) An Exhaustive, Non-Euclidean, Non-Parametric Data Mining Tool for Unraveling the Complexity of Biological Systems – Novel Insights into Malaria. PLoS ONE 6(9): e24085. doi:10.1371/journal.pone.00240

HyperCube pour la détection précoce de la maladie d'Alzheimer (coll. EPFL / CHUV)

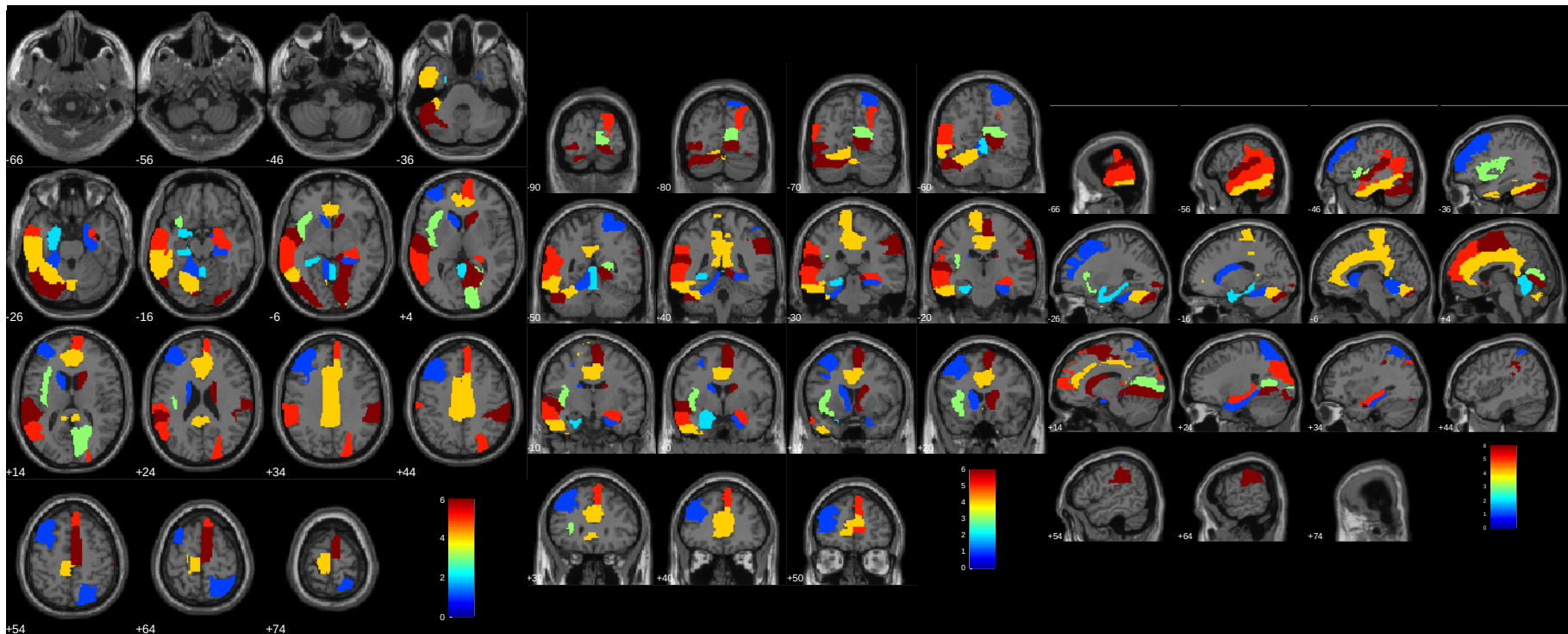


Le cas :

- une cohorte de patients AD et Contrôle
- Des données d'imagerie cérébrale normalisée

Application HyperCube : trouver des règles permettant de discriminer les patients malades des contrôles

Résultats : 6 règles identifiant à 100% les patients malades des contrôles



En résumé

Une révolution économique, technologique, scientifique est à l'œuvre dans le domaine de la médecine et de la santé

Cette révolution sera portée par la mise à disposition d'un volume et d'une variété de données sans commune mesure avec la situation actuelle

Etre capable de tirer de la valeur de ces données est un enjeu majeur

L'analyse à valeur ajoutée des données biologiques et médicales reste un challenge

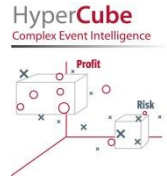
- Multi-causalité
- Populations hétérogènes
- Données de qualité hétérogène et parfois faible

Les technologies de data mining telles qu'HyperCube joueront un rôle clé dans les progrès à venir

- Identification des causes locales
- Appréhension des interactions entre variables
- Traitement performant des données manquantes

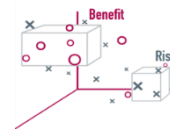
Les progrès dans la recherche sur l'intelligence artificielle et l'explosion des puissances de calcul permettront à ces technologies de donner leur plein potentiel

Contact



Yann Gaston-Mathé
Director
Life Sciences

HyperCube
Complex Event Intelligence



T +33 (0) 1 58 86 30 00
M + 33 (0) 6 23 74 14 14

Immeuble Galilée
51, esplanade du Général de Gaulle
92000 Puteaux Cedex
France

yann.gastonmathe@hypercube-research.com