Artificial intelligence, Data reuse, big data in healthcare and hospital databases

Pr Emmanuel Chazard

https://www.youtube.com/user/emmanuelchazard/videos

CHU LILLE

Université de Lille

CERIM

Presentation

- Lille University:
  - Professor of medicine (public health)
  - Researcher in biostatistics and medical informatics, notably artificial intelligence and data reuse
  - Head of the CERIM

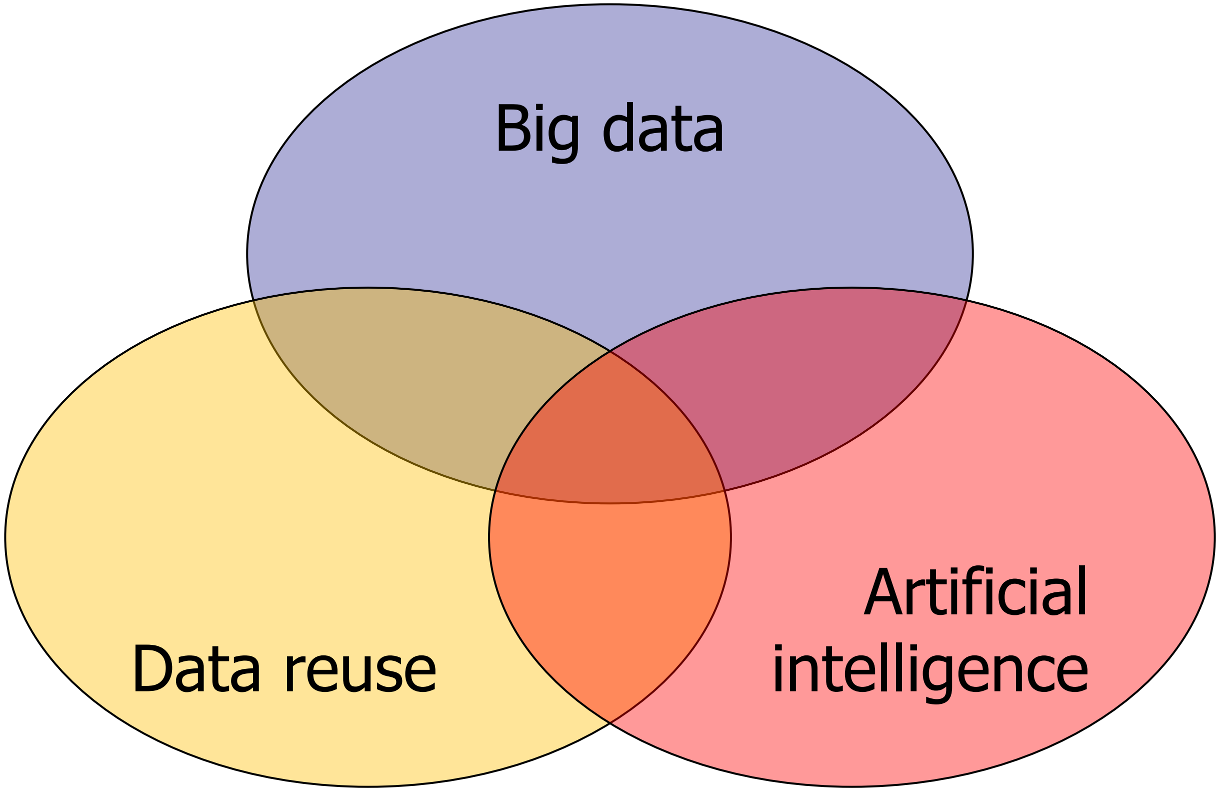
Université de Lille

- Lille University Hospital
  - Hospital practitioner: quality of care, methodological and statistical support for clinical researchers

CHU LILLE

The ideas expressed in this presentation do not necessarily represent those of the organisms nor of the persons who are cited.

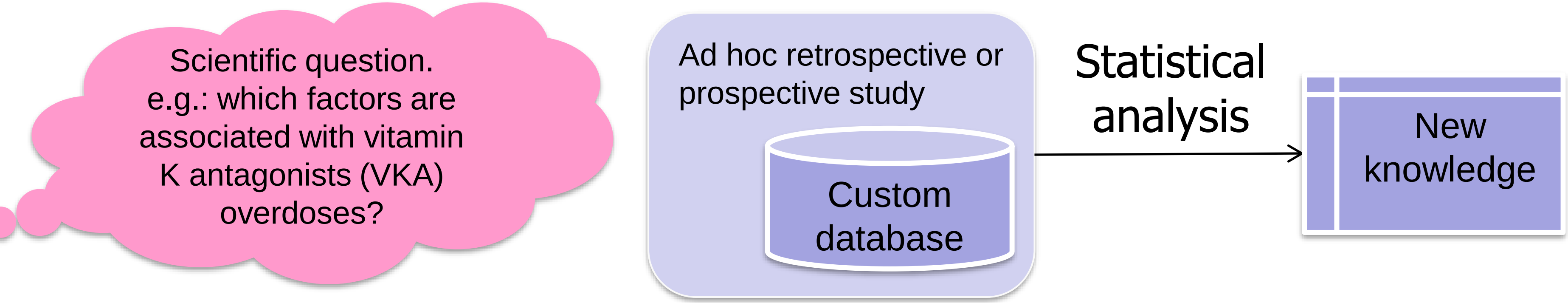
Overlap, frequent confusion…



Data reuse of inpatient hospital records, data mining.

Definition of *Data reuse* (or secondary use of data)

- Traditional approaches (before data reuse):



Scientific question. e.g.: which factors are associated with vitamin K antagonists (VKA) overdoses?

Ad hoc retrospective or prospective study

Custom database

Statistical analysis

New knowledge

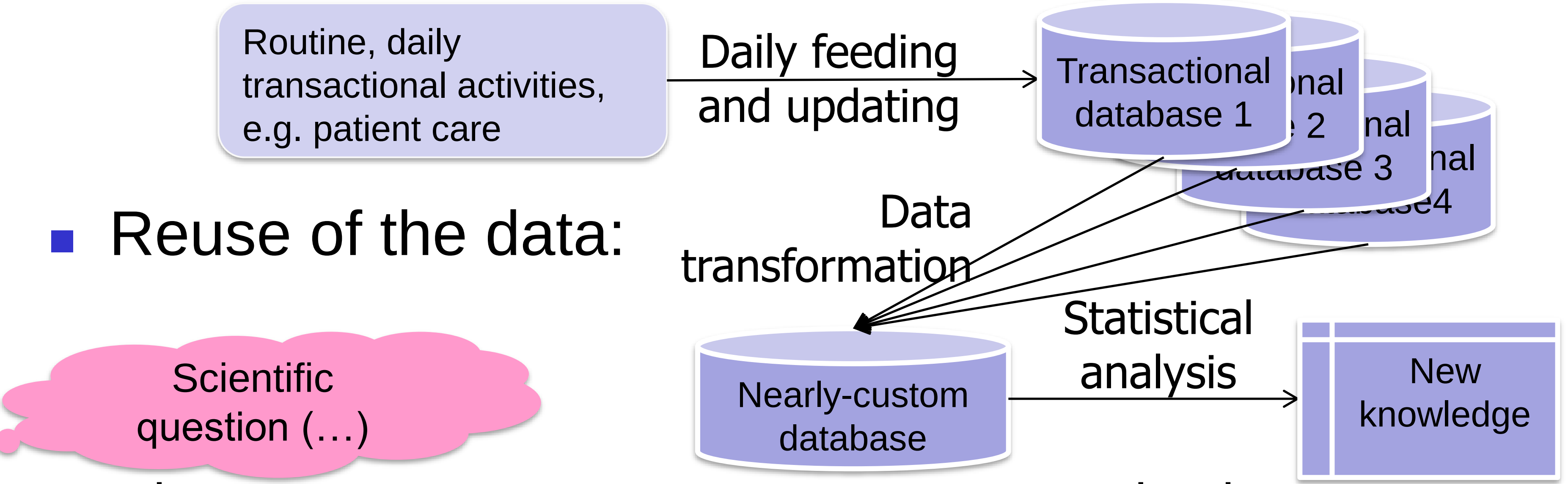
- Advantages:
  - Simple and specific data collection
  - Simple data analysis
  - Answers accurately the initial question

- Drawbacks:
  - Time-consuming
  - Expensive
  - Late results
  - Often few records (low power)
  - Data cemeteries

Definition of *Data reuse* (or data re-use)

- Routine collection of transactional data:

- Reuse of the data:

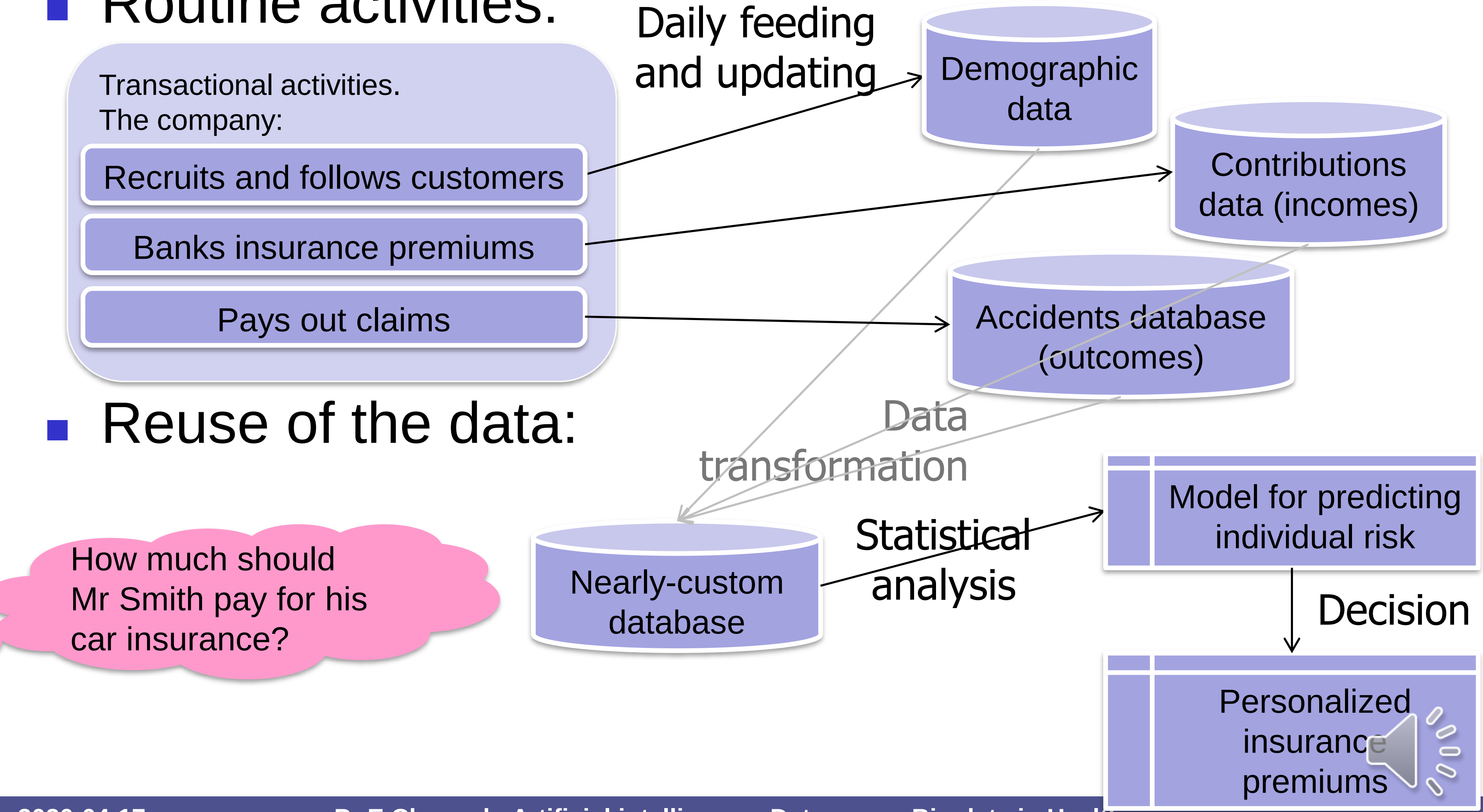


- Advantages:
  - Low-cost
  - Fast results
  - Data enhancement
  - Amount of records => high statistical power

- Drawbacks:
  - Often approximately answers the question
  - Not easy, methodological issues

Data reuse in insurance companies…

- Routine activities:



Transactional activities.
The company:

- Recruits and follows customers
- Banks insurance premiums
- Pays out claims

Daily feeding and updating

Demographic data

Contributions data (incomes)

Accidents database (outcomes)

- Reuse of the data:

Data transformation

Nearly-custom database

Statistical analysis

Model for predicting individual risk

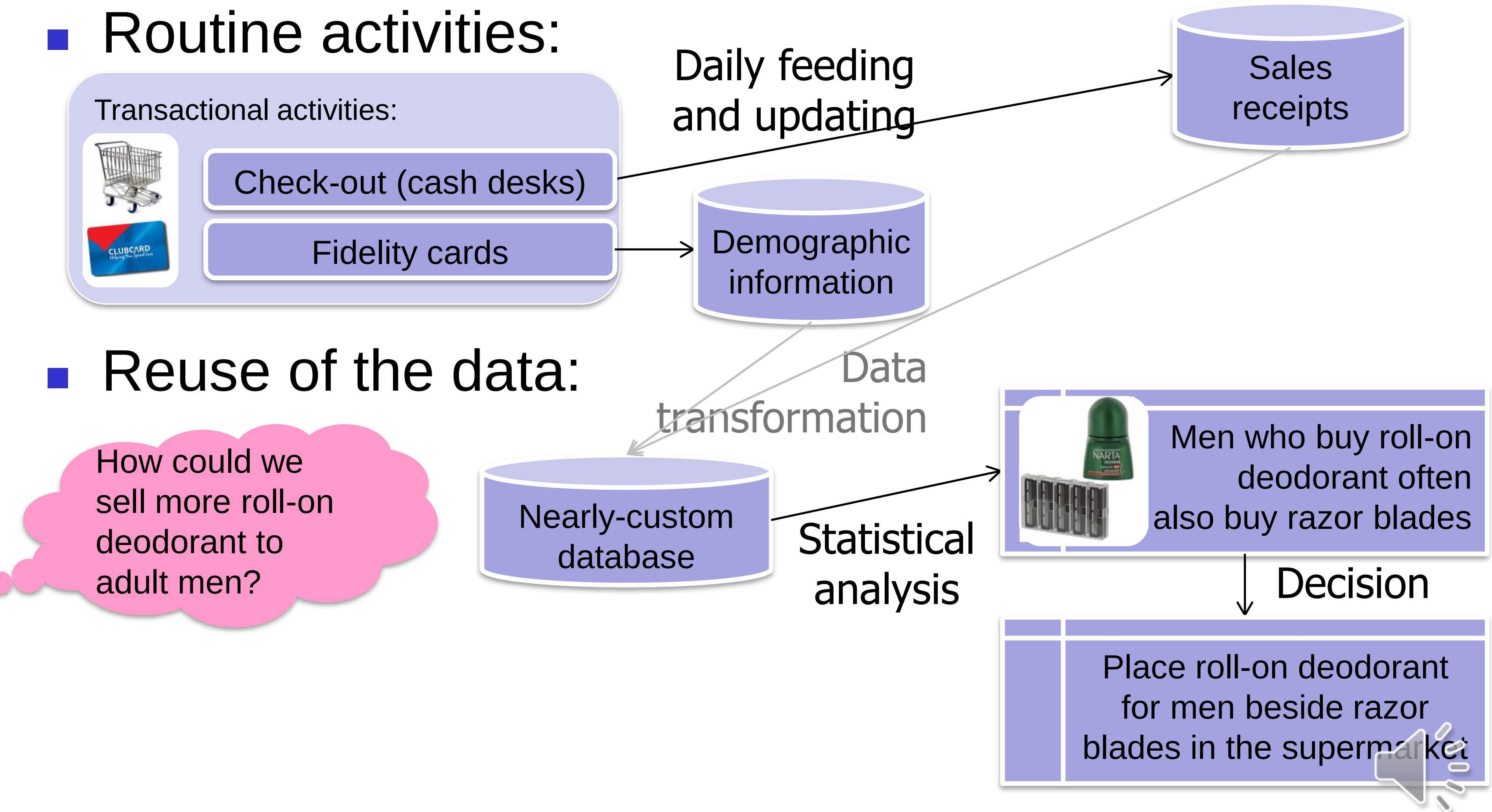
Decision

Personalized insurance premiums

How much should Mr Smith pay for his car insurance?

Data reuse in supermarket…

- Routine activities:
- Reuse of the data:



Transactional activities:

Check-out (cash desks)

Fidelity cards

Daily feeding and updating

Sales receipts

Demographic information

Data transformation

Nearly-custom database

How could we sell more roll-on deodorant to adult men?

Statistical analysis

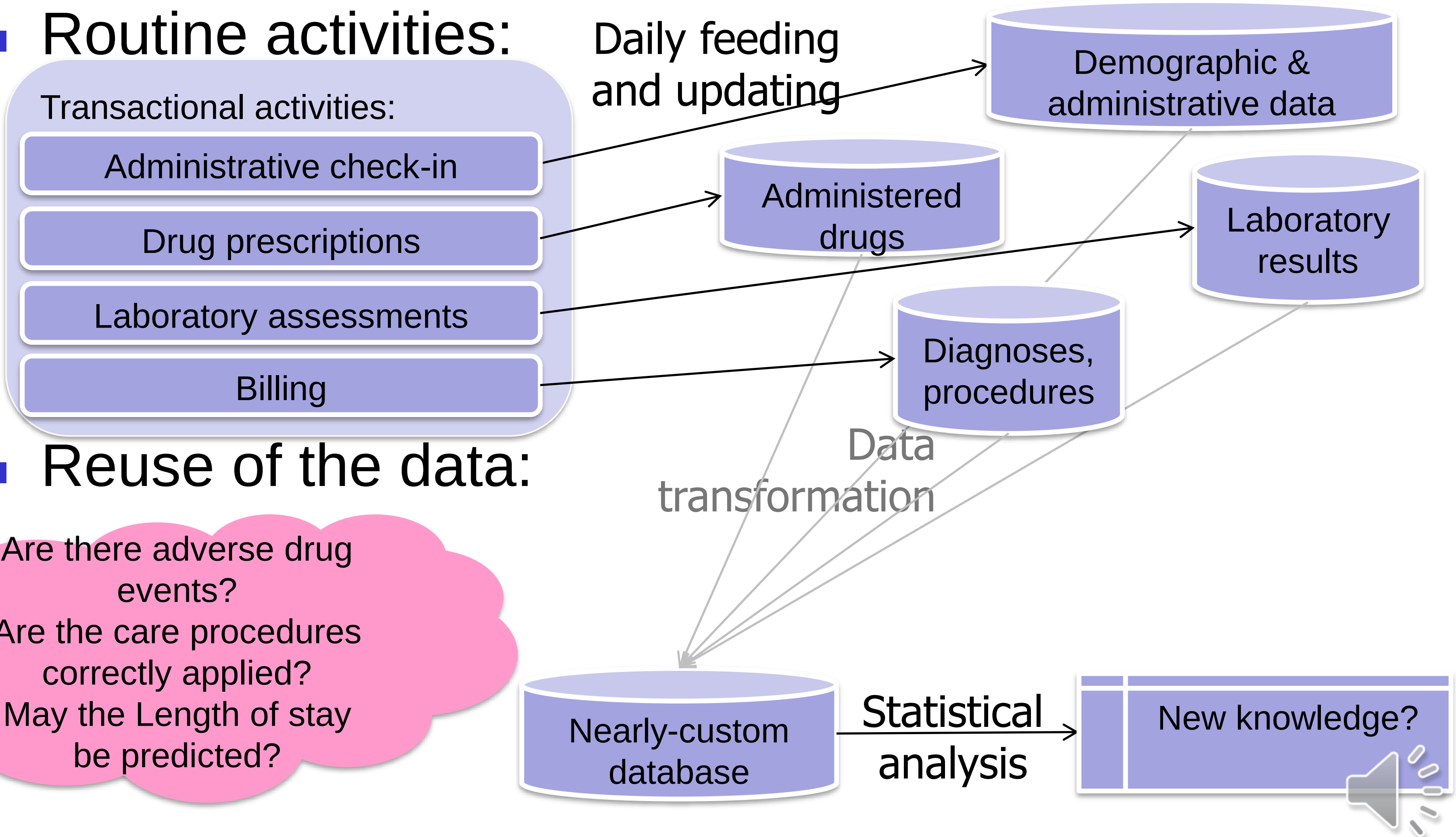
Men who buy roll-on deodorant often also buy razor blades

Decision

Place roll-on deodorant for men beside razor blades in the supermarket

Data reuse in health?
Probably under-realized today

- Routine activities:



Transactional activities:

- Administrative check-in
- Drug prescriptions
- Laboratory assessments
- Billing

Daily feeding and updating

- Demographic & administrative data
- Administered drugs
- Laboratory results
- Diagnoses, procedures

Data transformation

Nearly-custom database

Statistical analysis

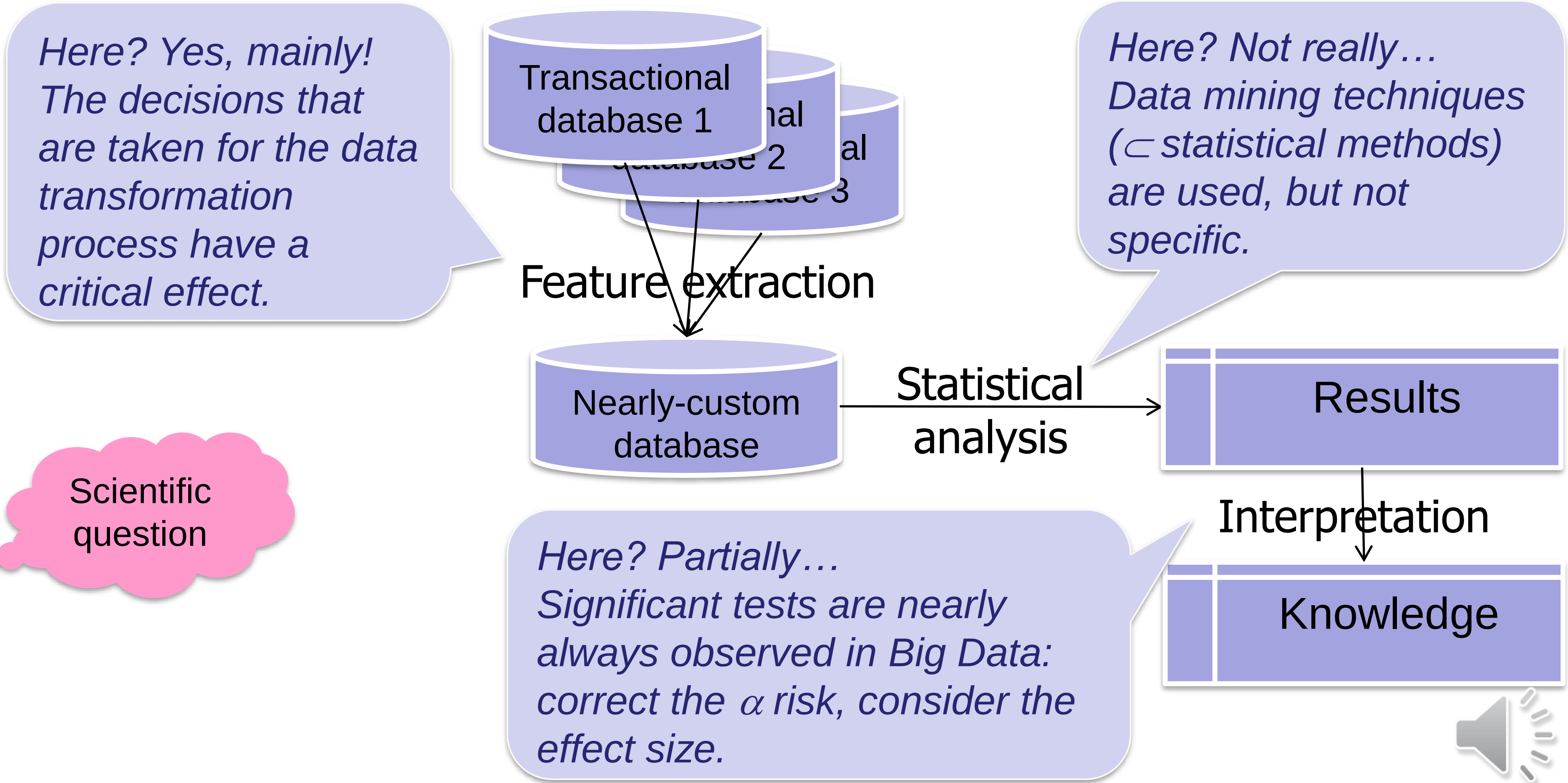
New knowledge?

- Reuse of the data:

Are there adverse drug events?
Are the care procedures correctly applied?
May the Length of stay be predicted?

Challenges in *data reuse*

- Where is <u>the</u> secret of a successful data reuse?



Here? Yes, mainly! The decisions that are taken for the data transformation process have a critical effect.

Here? Not really… Data mining techniques ($\subset$ statistical methods) are used, but not specific.

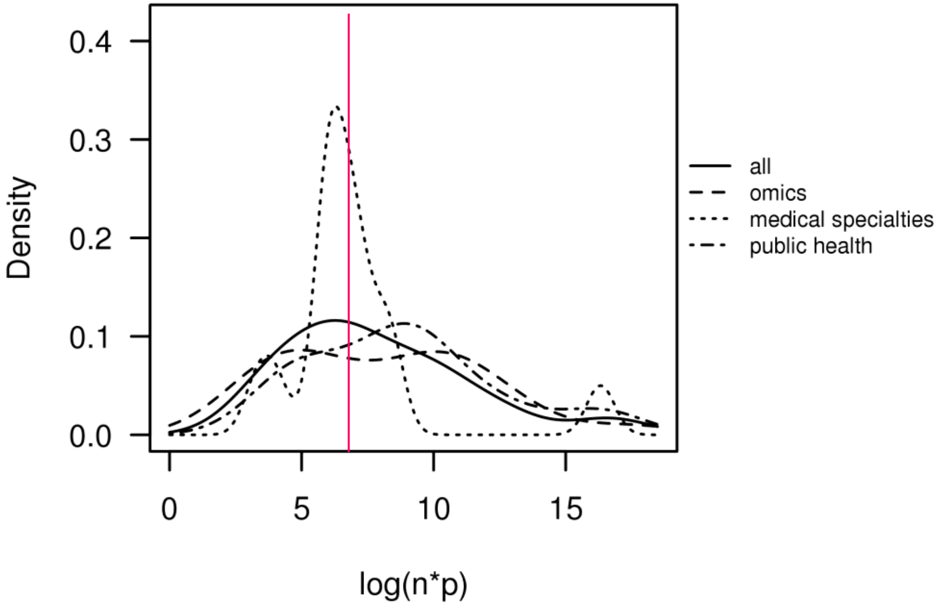
Here? Partially… Significant tests are nearly always observed in Big Data: correct the α risk, consider the effect size.

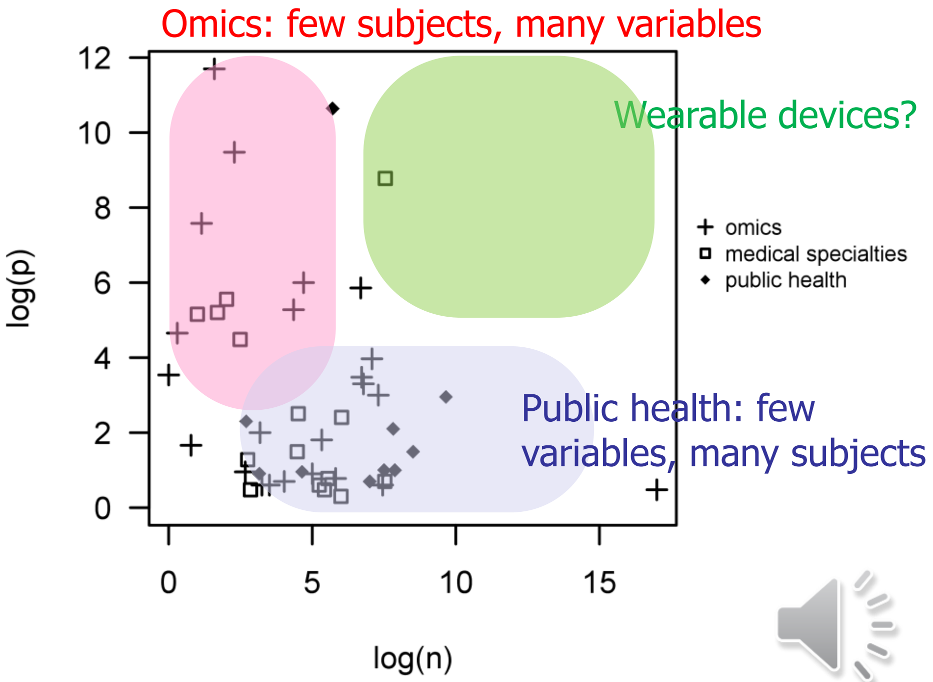
Big data

Definition of Big Data in Healthcare

Baro 2014, Toward a Literature-Driven Definition of Big Data in Healthcare

- Bibliographic review of 330 international papers with the “big data” keyword (among which 48 describing a dataset)
- Literature-driven definition of big data:
  - high volume Log(n*p)≥7 (n=nb of subjects, p=nb of variables)





Properties of “big data”

- Definition : “big” only
- Properties
  - Data storage (…)
  - Data analysis
    - Computer treatment: computational power, RAM overload
    - Data management: too complex relationships, difficulty to handle associations in data driven approaches, data visualization
    - Statistical analysis: increase of type I error, overfitting of multivariate models
- Wrong properties:
  - Knowledge extraction, low cost, etc.
  - Frequent confusion with “data reuse”

Healthcare big data

- Big data sources:
  - Medical records (electronic health records)
  - Laboratory results (~100 lines per inpatient stay)
  - Medical imaging
  - Drug prescriptions (~20 lines per inpatient stay)
  - Medical insurance
  - Lille University hospital: 2 million patient records
- Some examples in France:
  - 1.5 millions diabetic patients
  - 60 millions records for Social Security
  - PMSI: ~27 million acute care inpatient stays per year

Definition of “big data”

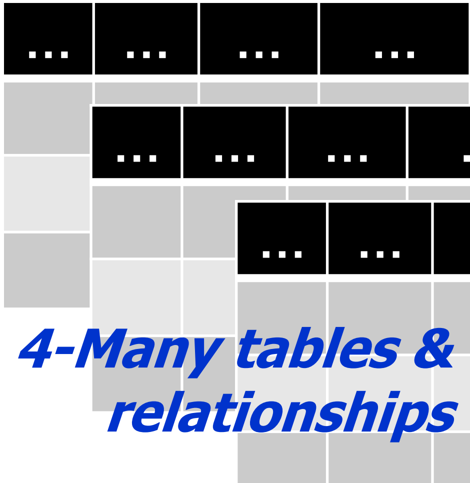
- “big data” is generally a property of the routinely collected data that can be *reused*
- “Big” can be understood through 5 dimensions:

2-Many variables

id	age	gender	diagnosis	…
123	23	M	I10	…
125	78	M	K37	…
245	13	F	M61.2	…
278	24	M	I41	…
324	65	F	I48	…
350	34	F	F20.2	…
…	…		…	…

1-Many records

3-Many possible values for qualitative variables



4-Many tables & relationships

5-Variables with repeated measurements

Id	Par	Val
123	K+	4.5
123	K+	4.8
123	K+	5.2

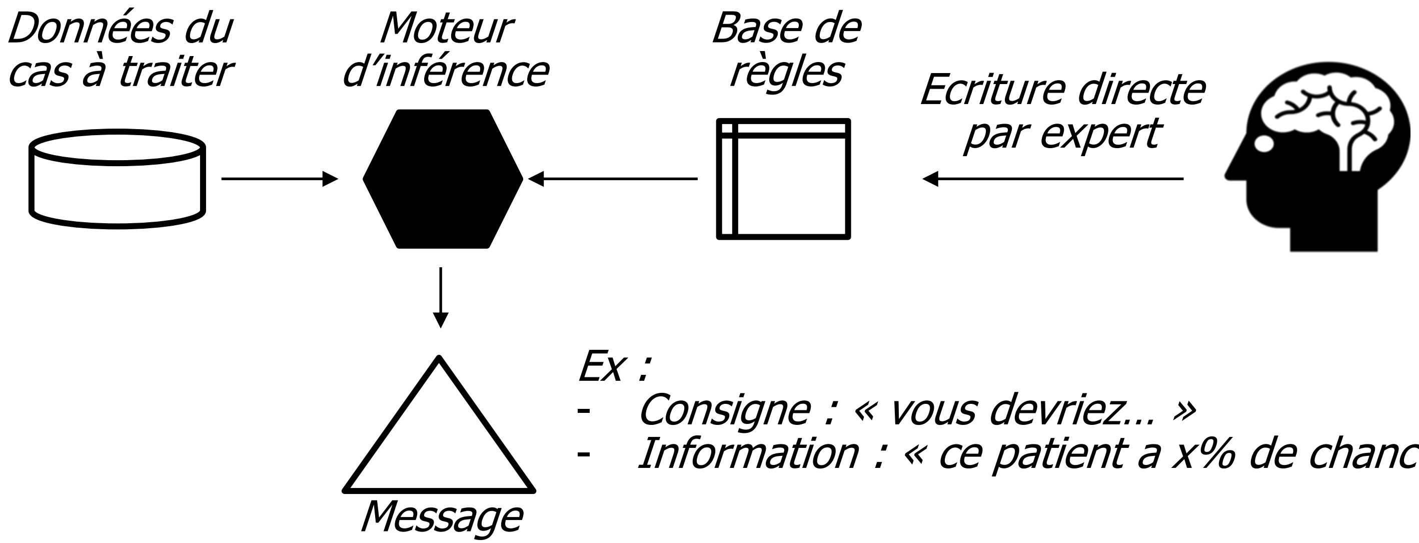
Artificial intelligence

Some definitions

- Intelligence :
  - ensemble des facultés mentales permettant de
    - comprendre les choses et les faits
    - découvrir les relations entre eux
    - d'aboutir à la connaissance conceptuelle et rationnelle
  - se perçoit dans l'aptitude à comprendre (rétrospectif)
  - et à s'adapter facilement à des situations nouvelles (prospectif)
- Intelligence artificielle :
  - recherche de moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles comparables à celles des êtres humains, ou en tout cas mimant ces capacités
  - analyse (rétrospectif) et décision (prospectif)
  - le gazon artificiel n’est pas du gazon…

Three levels of artificial intelligence

- Niveau 1 : exécuter
  - Répertoire de règles déjà disponibles
  - Identifier simplement quand les conditions des règles s'appliquent
  - *=> permet d'appliquer un traitement tel qu'il a été prévu*

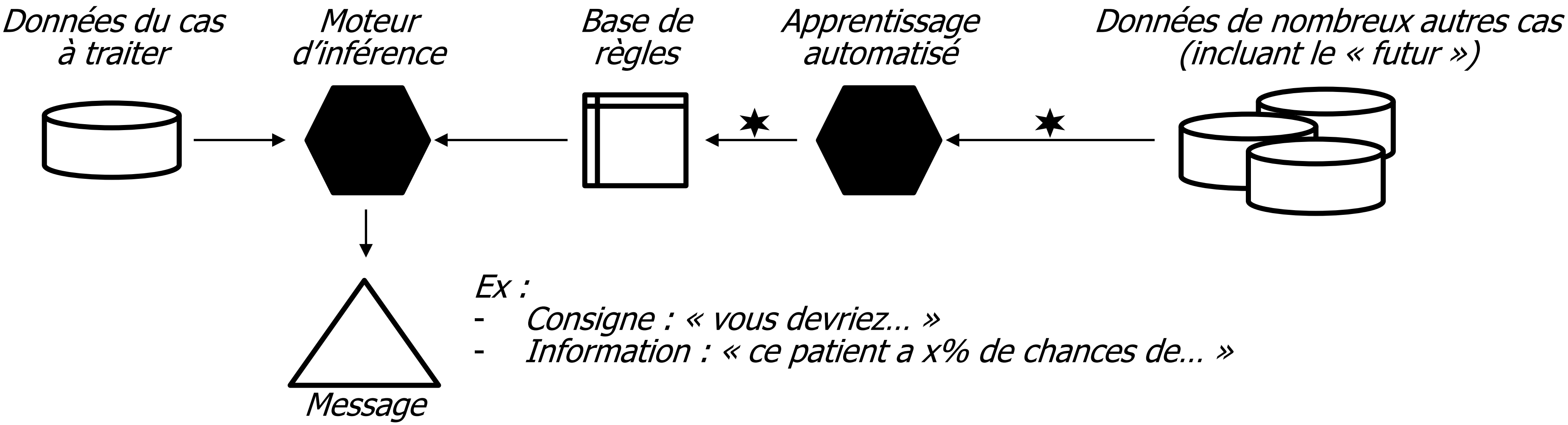


Ex :
- *Consigne : « vous devriez… »*
- *Information : « ce patient a x% de chances de… »*

- Niveau 2 : apprendre et réappliquer
- Niveau 3 : s'adapter intelligemment

Three levels of artificial intelligence

- Niveau 1 : exécuter
- Niveau 2 : apprendre et réappliquer
  - Observer des expériences
  - En déduire des associations, relations (démarche empirique, machine learning, data mining supervisé)
  - Appliquer ces règles apprises par l’expérience, mais pas formalisées
  - *=> permet de traiter une situation déjà rencontrée*


Données du cas à traiter
Moteur d’inférence
Base de règles
Apprentissage automatisé
Données de nombreux autres cas (incluant le « futur »)
Message

Ex :
- *Consigne : « vous devriez… »*
- *Information : « ce patient a x% de chances de… »*

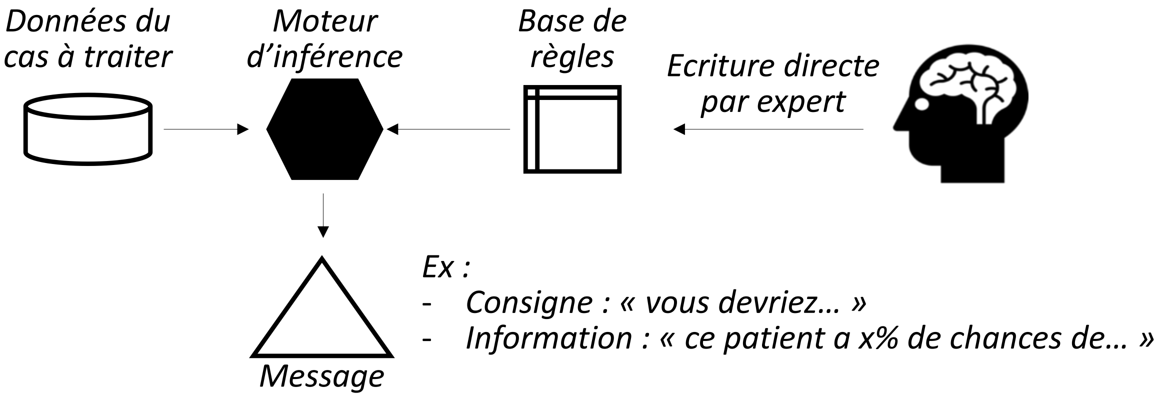
- Niveau 3 : s’adapter intelligemment

Three levels of artificial intelligence

- Niveau 1 : exécuter
- Niveau 2 : apprendre et réappliquer
- Niveau 3 : s’adapter intelligemment
    - Raisonnement mélangeant les règles prédéfinies, de l’expérience et du « bon sens »
    - Logique floue, intégration de conclusions contradictoires (ex : traiter un patient diabétique et insuffisant coronarien)
    - Permet de traiter une circonstance partiellement ou totalement inédite
    - Notion d’intelligence artificielle « forte », machine capable de :
        - adopter des comportements intelligents
        - éprouver une conscience de soi
        - comprendre ses propres raisonnements
    - Serait possible sur un argument purement quantitatif (en nombre d’opérations par seconde) :
        - Balance de Roberval : 1
        - Ordinateur 1970 : 1*107
        - Ordinateur 2005 : 1*1011
        - Cerveau humain : 2*1014 … atteint en 2019 selon loi de Moore
    - Mais la puissance de calcul ne suffit pas :
        - Limite actuelle : ce que les humains savent « enseigner » aux machines
        - Impossible avec notre représentation actuelle des connaissances et de la logique
        - Supposerait intégration de la logique floue, manipulation d’autres formalismes et symbolismes, et une « initiative » de la machine…
    - => n’existe actuellement pas

Three levels of artificial intelligence

- Niveau 1 : exécuter

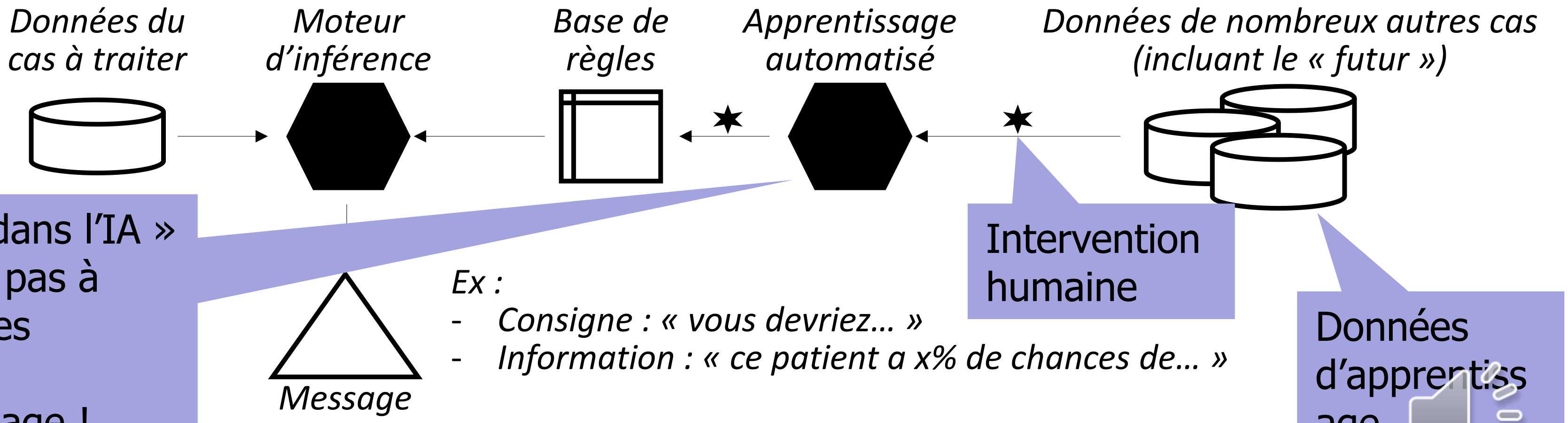


Ex :

- *Consigne : « vous devriez… »*
- *Information : « ce patient a x% de chances de… »*

Pas nouveau dans l'aide à la décision médicale. Ex : MYCIN de Shortcliffe (1974)

- Niveau 2 : apprendre et réappliquer



Ex :

- *Consigne : « vous devriez… »*
- *Information : « ce patient a x% de chances de… »*

« travailler dans l'IA » ne se limite pas à concevoir des méthodes d'apprentissage !

Intervention humaine

Données d'apprentissage

Brief history in Healthcare

- Definition of fundaments
  - Theoretical definition in the 1950s
  - Perceptron and artificial neural network in the 1960s (computer modelling of neural function)
- 1970s: first winter of the AI (disappointment)
- Late 1970s: Generalization of expert systems (CDSS clinical decision support systems), AI level 1
- Late 1980s: Second winter of the AI
  - Deletion of the term AI, but continued development of CDSS
- 1990s: Level 2 AI
  - concept of data mining, machine learning, knowledge discovery in databases
- 2000's: New AI boom
  - New machine learning methods: deep learning, multi-layer neural networks, support vector machine
  - No new concept, but ubiquitous AI (smartphones...)
  - 2015: Politicians and journalists discover the AI and talk about it in the future...

Let's make it clear: AI and medical liability

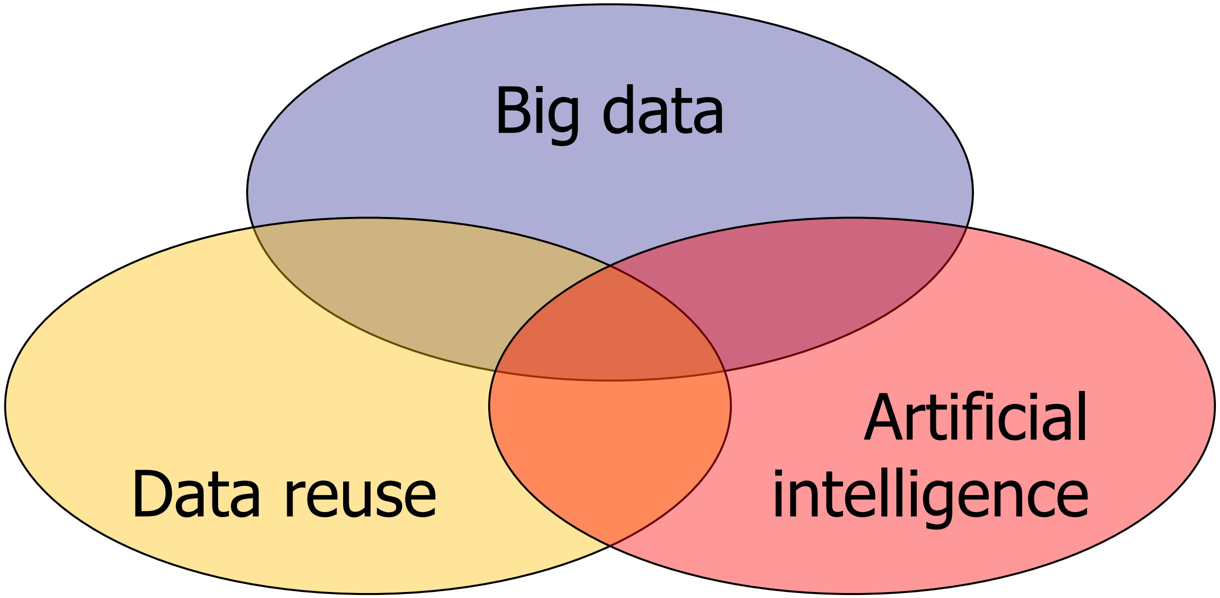
- Physician before AI:
  - Takes a medical decision, mitigating: risk related to abstention / risk related to the intervention / cost / urgency / patient experience...
  - All the information is more or less false, more or less outdated (biological test, imaging, interrogation, etc.)
  - Doctor trained to mitigate uncertain information
  - Only responsible for his decision in this context
  - Cannot invoke a single false examination result to clear himself of customs
- AI as decision support for physicians:
  - One more information, also more or less false
  - Does not change the doctor's responsibility
  - Fee for service = fee for the intellectual act of interpretation + cost of the act
  - Absolutely nothing new, problem solved since always!!!!!
- AI in replacement of physicians
  - Does not exist today (no parallel with the autonomous car)
  - This problem is currently purely theoretical

Data reuse, big data, and artificial intelligence…

- Input of a data reuse study:
  - Often from big data
  - Often from data of average size
- Purpose of data reuse study:
  - Sometimes to generate level 2 artificial intelligence (learning set)
  - Sometimes to evaluate artificial intelligence
  - Often for other studies
- Summary:
  - Data reuse: a set of processes, the goal of data collection is not the same as data analysis
  - Big data: the size of the data set
  - Artificial intelligence: a field of interest, that sometimes benefits from data reuse and big data

Data reuse, big data, and artificial intelligence…

E.g.: the French Ministry of Education disseminates the results of the baccalaureate of 743,000 candidates



Ex : a journalist team shows that:
P(success / FirstName='Joséphine')=97%

Ex : an online form asks your first name and predicts your success probability

Examples of artificial intelligence

- Other part of this course:
  - Adverse drug events detection and prevention
  - ECG automated interpretation
  - Prediction of the inpatient length of stay
  - Increase of hospital incomes
  - Automated computation of process quality indicators

May the machine outperform the Human?

« la peau de l'Ours », Pierre Soulages

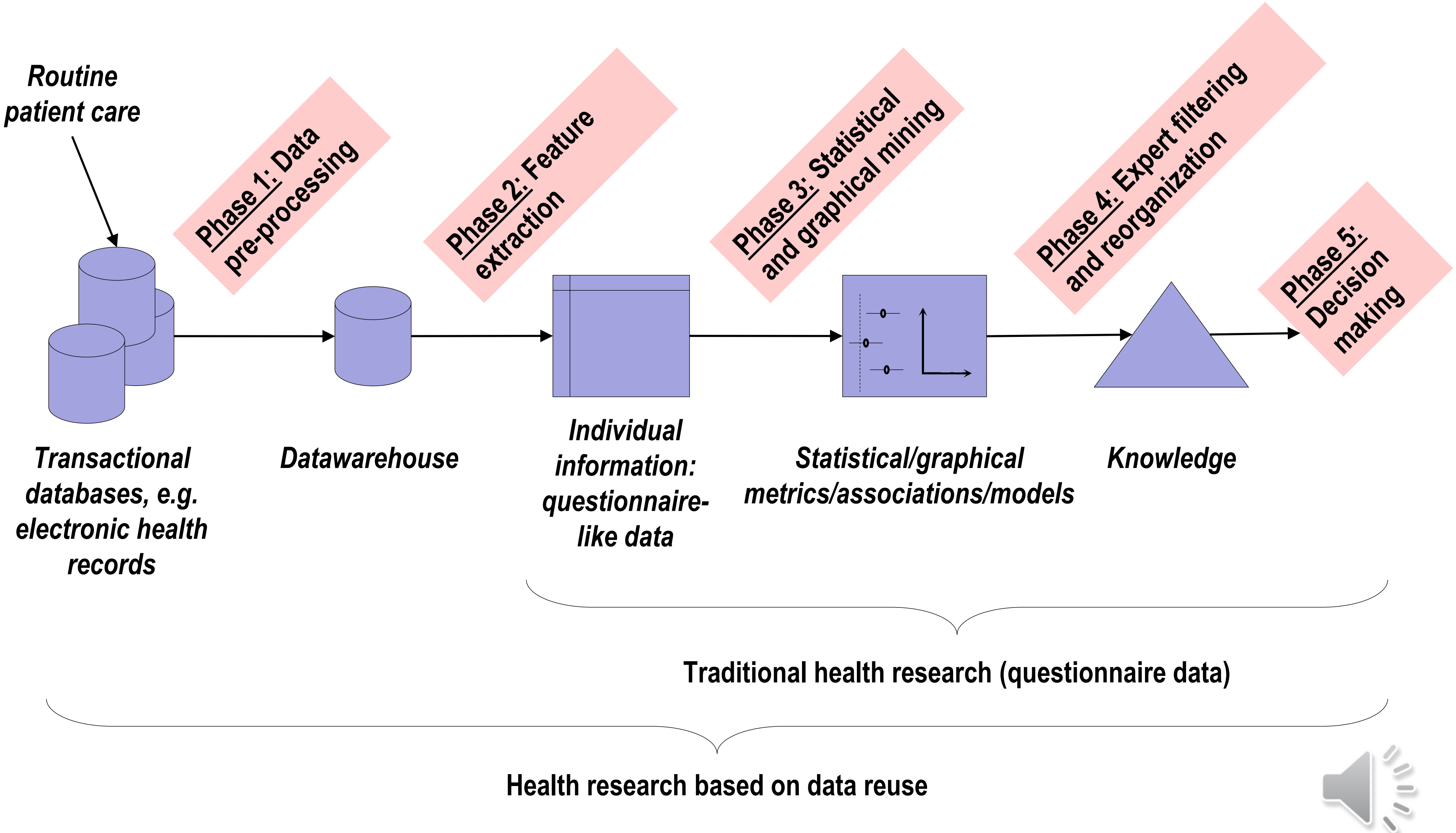
Maybe other intelligent creatures will do it before ;-)

Feature extraction

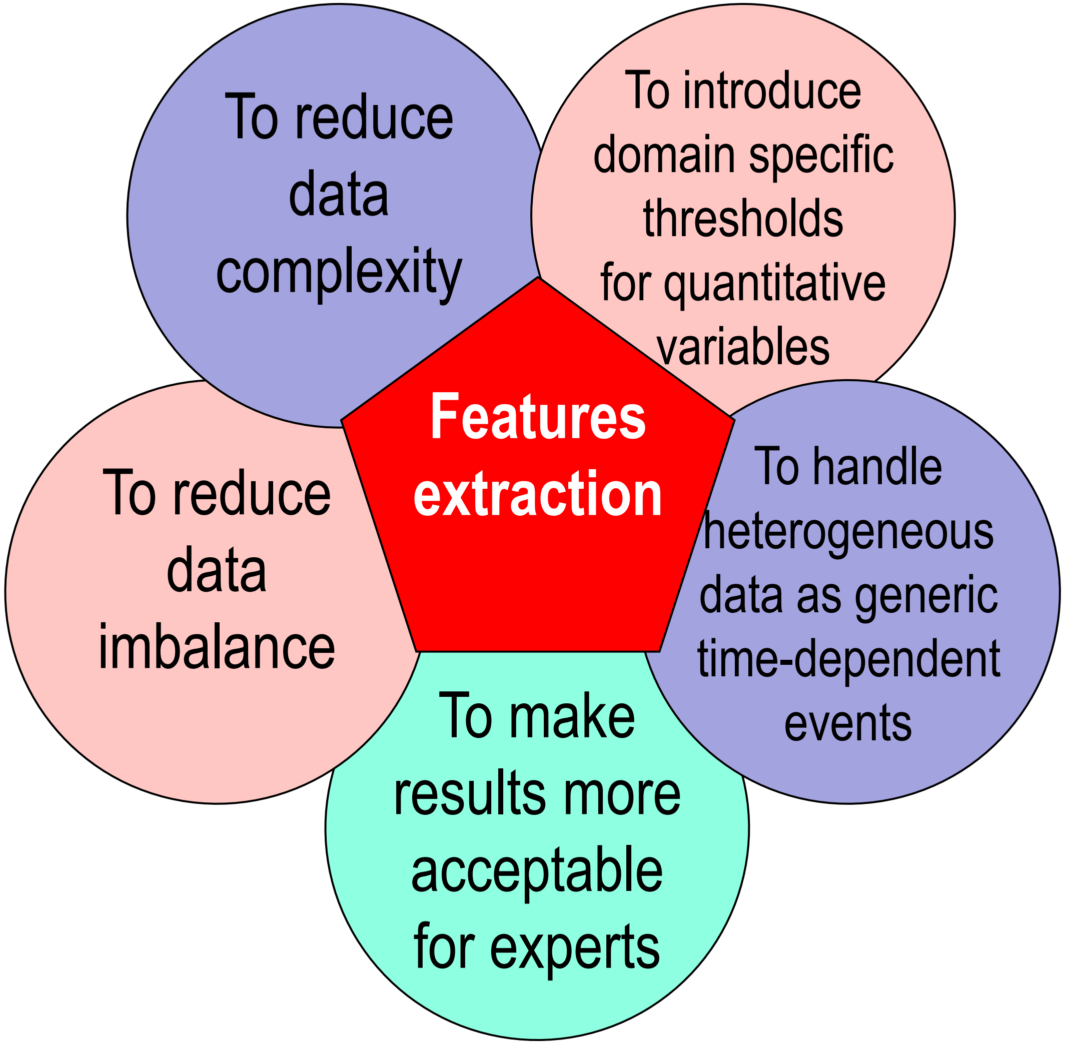
“Feature extraction”

- Has been discussed for signal / images / free-text, not for secondary use of structured data
- Feature extraction = transformation:
  - From native data, unsuitable for statistical analysis
  - To “questionnaire-like” data, that can be analyzed
- From our experience, is the most critical part of secondary use of healthcare structured data
- Other terms: data transformation, data pre-processing, data aggregation”, etc.

Structured data reuse process in healthcare


Routine patient care
Phase 1: Data pre-processing
Phase 2: Feature extraction
Phase 3: Statistical and graphical mining
Phase 4: Expert filtering and reorganization
Phase 5: Decision making
Transactional databases, e.g. electronic health records
Datawarehouse
Individual information: questionnaire-like data
Statistical/graphical metrics/associations/models
Knowledge
Traditional health research (questionnaire data)
Health research based on data reuse

The objectives of feature extraction



Features extraction

To reduce data complexity

To introduce domain specific thresholds for quantitative variables

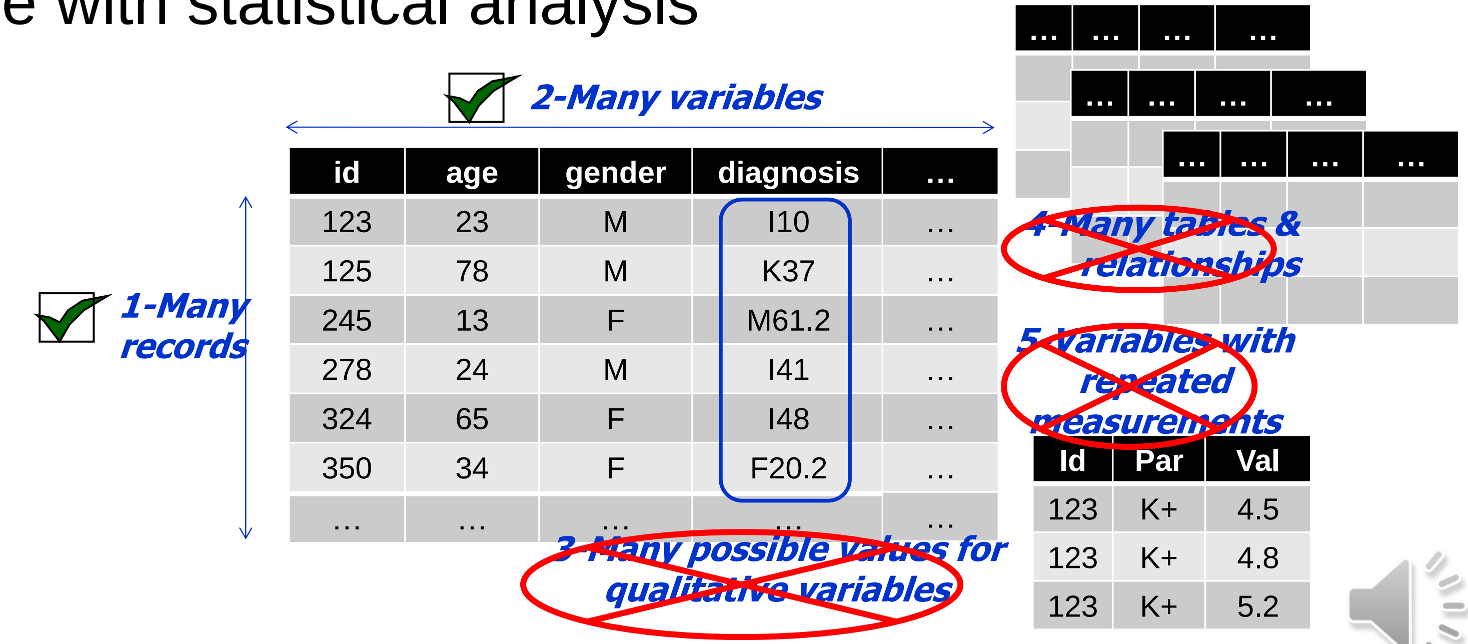
To handle heterogeneous data as generic time-dependent events

To make results more acceptable for experts

To reduce data imbalance

Feature extraction to reduce data complexity

- Mandatory before using statistical methods
- Enables to transform data into information: how would a human comment/summarize the raw data?
- Suppresses 3 over 5 dimensions of “bigness” that are not compatible with statistical analysis



☑ *2-Many variables*

☑ *1-Many records*

id	age	gender	diagnosis	…
123	23	M	I10	…
125	78	M	K37	…
245	13	F	M61.2	…
278	24	M	I41	…
324	65	F	I48	…
350	34	F	F20.2	…
…	…	…	…	…

~~*3-Many possible values for qualitative variables*~~

…	…	…	…

~~*4-Many tables & relationships*~~

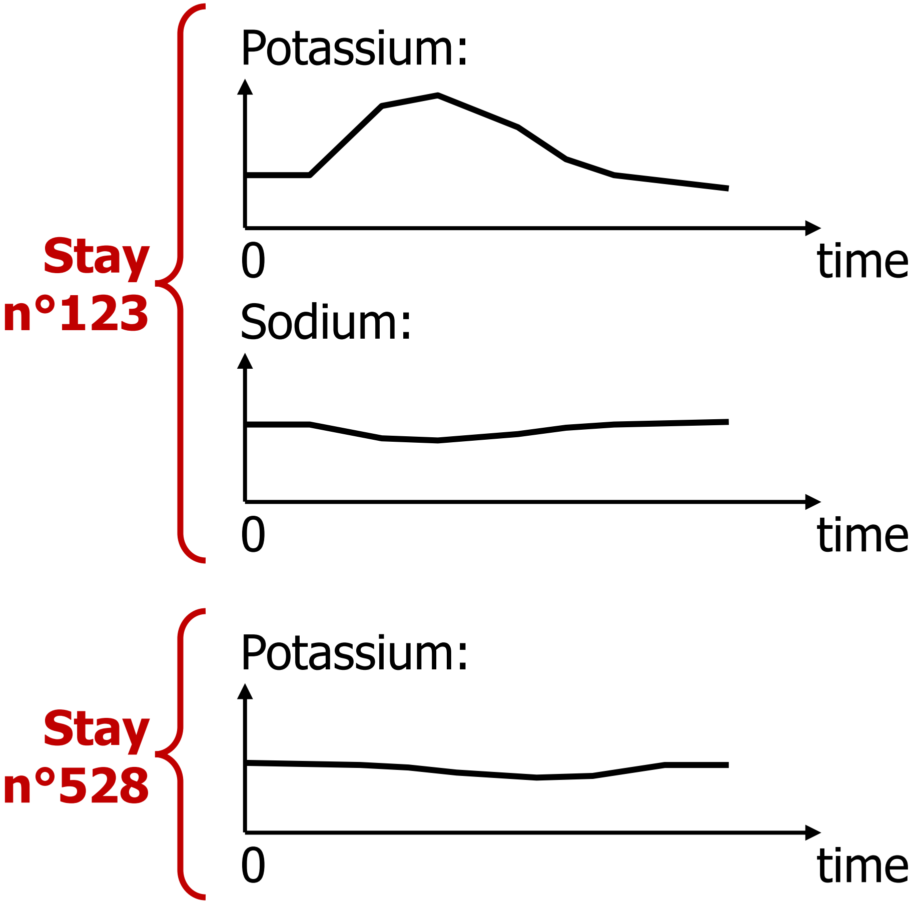
~~*5-Variables with repeated measurements*~~

Id	Par	Val
123	K+	4.5
123	K+	4.8
123	K+	5.2

Feature extraction, example 1: suppression of repeated measurements (1)

Id stay	Date	Parameter	Value
123	0	Potassium	4
123	1	Potassium	4
123	…	…	…
123	0	Sodium	140
123	…	…	…
528	0	Potassium	3.2
528	…	…	…

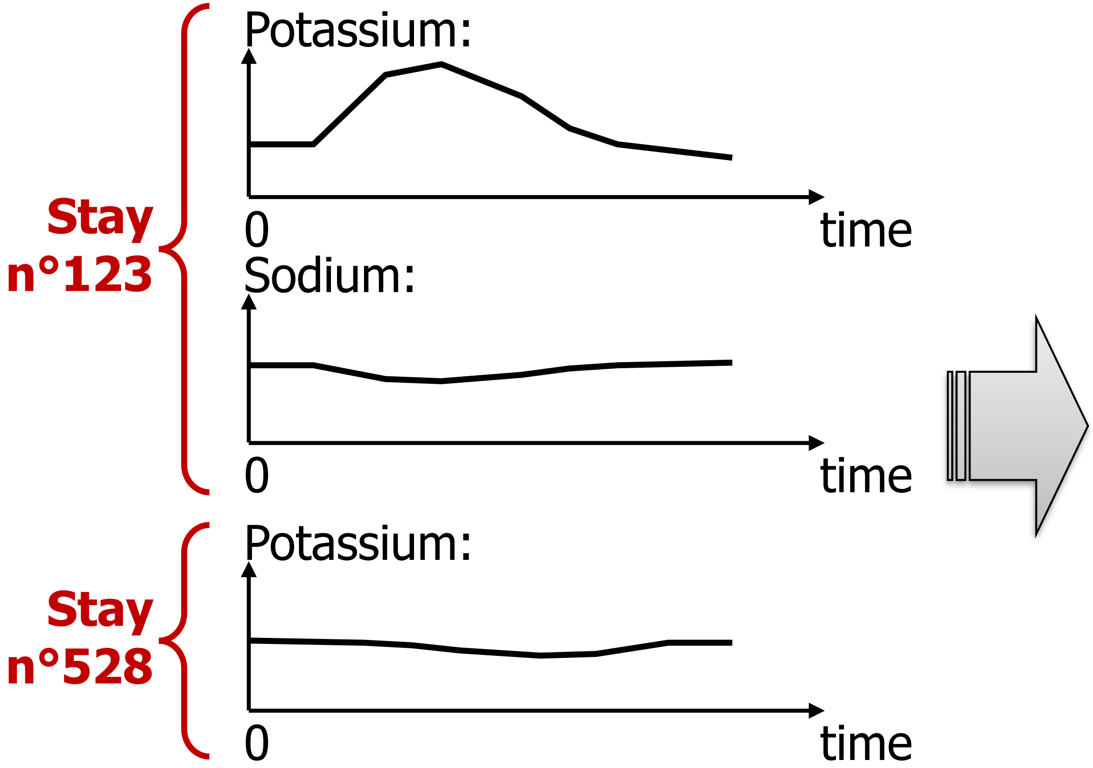
Several parameters may be measured several times during a given inpatient stay.
=> One curve per {id_patient*parameter}


Stay n°123

Stay n°528

Feature extraction, example 1: suppression of repeated measurements (2)

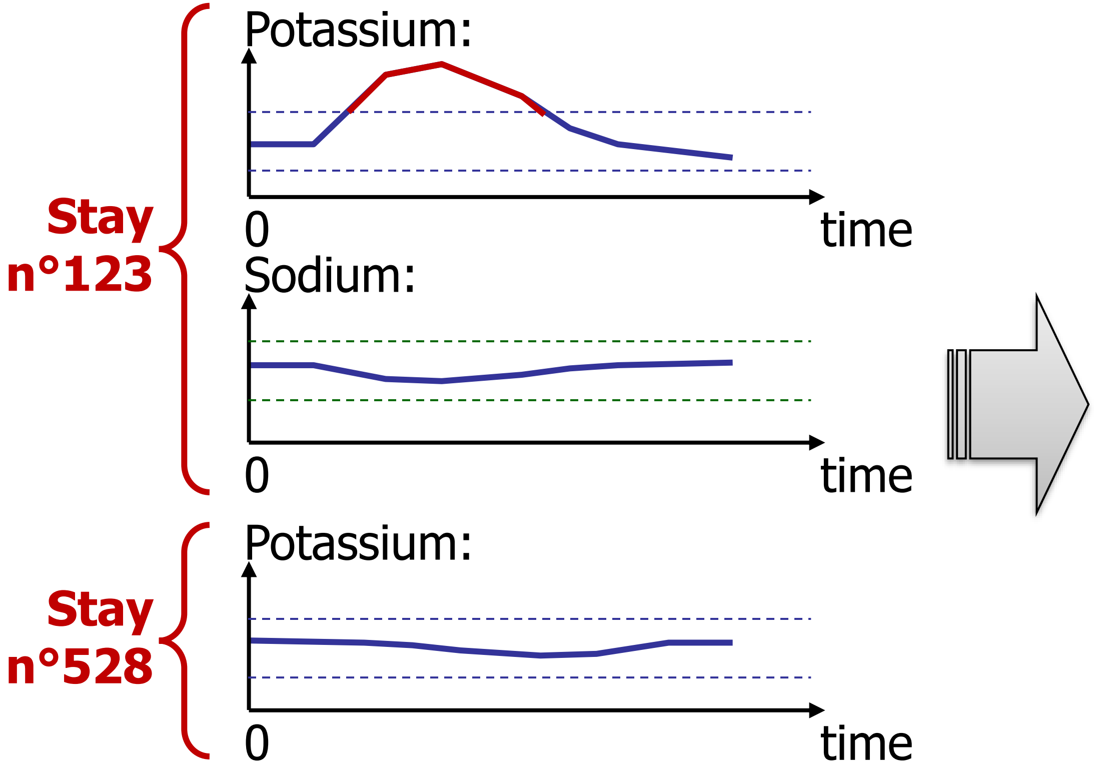
- Objective: 1 table with 1 row per stay
- Example of simple transformation (without *a priori* knowledge):



Id stay	Potassium.Min	Potassium.Max	Sodium.Min	Sodium.Max
123	3.8	6.2	136	142
528	2.8	3.2	NA	NA

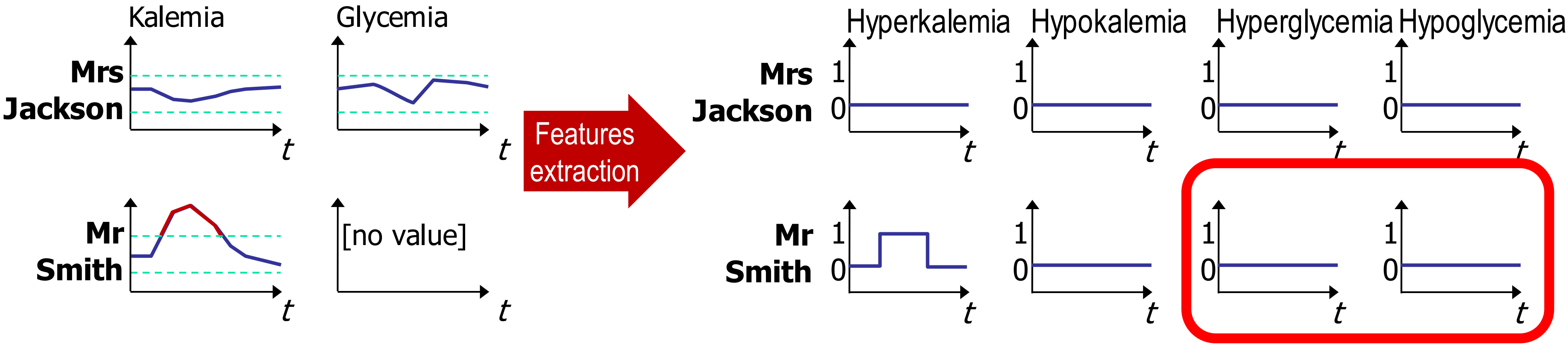
Feature extraction, example 1: suppression repeated measurements (3)

- Another example of transformation, with knowledge
- Uses the range of normal values according to the parameters, summarizes the abnormalities



Id stay	Hyperkalemia	Hypokalemia	Hypernatremia	Hyponatremia
123	1	0	0	0
528	0	0	NA	NA

Feature extraction, example 1: taking time into account (4)



Example of missing data handling

- Formally:
  - Example of 2 patients, 2 parameters measured 5 times
  - Before: 1 table with 2 lines + 1 table with 10 lines
  - After: 1 table with 2 lines

Feature extraction, example 2: hernia surgery (1)

Id stay	Procedure.code	Procedure.wording
123	LMMC004	Bilateral treatment of inguinal hernia without prosthesis, by video surgery
123	GLHF001	Arterial blood collection for blood gas and pH sampling
528	LMMC020	Treatment of abdominal hernia with prosthesis, by laparoscopy
528	ZZBQ002	Thorax radiography

Each stay may have 0, 1 or several procedures. The terminology used (CCAM) has more than 5,000 possible codes. In this case, we only interest on hernia treatment.

Feature extraction, example 2: hernia surgery (2)

Id stay	Procedure.code
123	LMMC004
~~123~~	~~GLHF001~~
528	LMMC020
~~123~~	~~ZZBQ002~~

Id stay	H.type	H.prosthesis	H.approach	H.bilateral
123	Inguinal	0	Videosurgery	1
528	Abdominal	1	Laparoscopy	0

New qualitative variables are created:

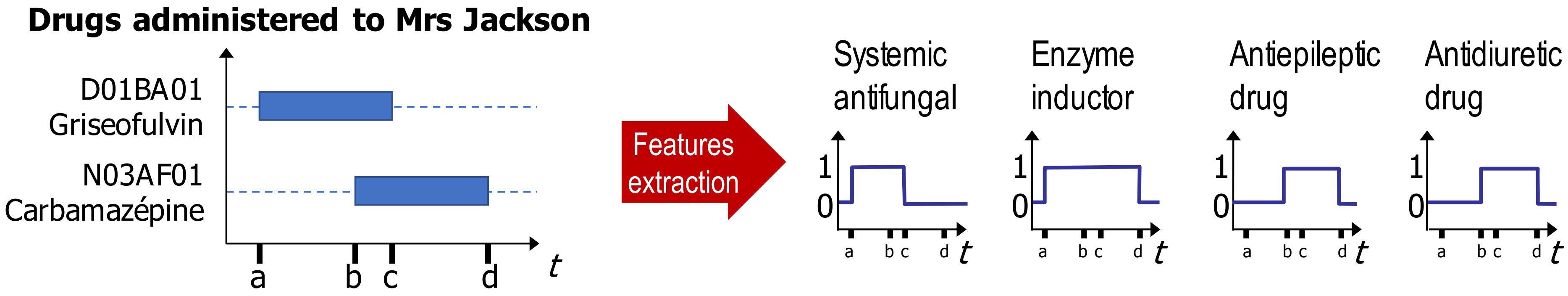
- Each variable has few possible values
- A mapping is necessary (with overlaps)
- Requires a strong medical knowledge about codes

Feature extraction, example 2: hernia surgery (3)

- Example of mapping used for hernia
- FROM (*raw data*):
  - 18 codes
- TO (*information*):
  - Type (n=3)
  - Prosthesis (n=2)
  - Approach (n=6)
  - Bilateral (n=2)

Hernia code	type	prosthesis	approach	bilateral
LMMA009	abdominal	0	direct	0
LMMA014	abdominal	0	direct	0
LMMC020	abdominal	1	laparoscopy	0
LMMA019	inguinal	0	inguinal	1
LMMA018	inguinal	0	inguinal	1
LMMA016	inguinal	0	inguinal	0
LMMA017	inguinal	0	inguinal	0
LMMC003	inguinal	0	videosurgery	0
LMMC004	inguinal	0	videosurgery	1
LMMA006	inguinal	1	direct	0
LMMA001	inguinal	1	inguinal	1
LMMA012	inguinal	1	inguinal	0
LMMA002	inguinal	1	other	1
LMMA008	inguinal	1	other	0
LMMC002	inguinal	1	videosurgery	0
LMMC001	inguinal	1	videosurgery	1
LMMA011	other	0	inguinal	0
LLMA007	other	0	laparotomy	0

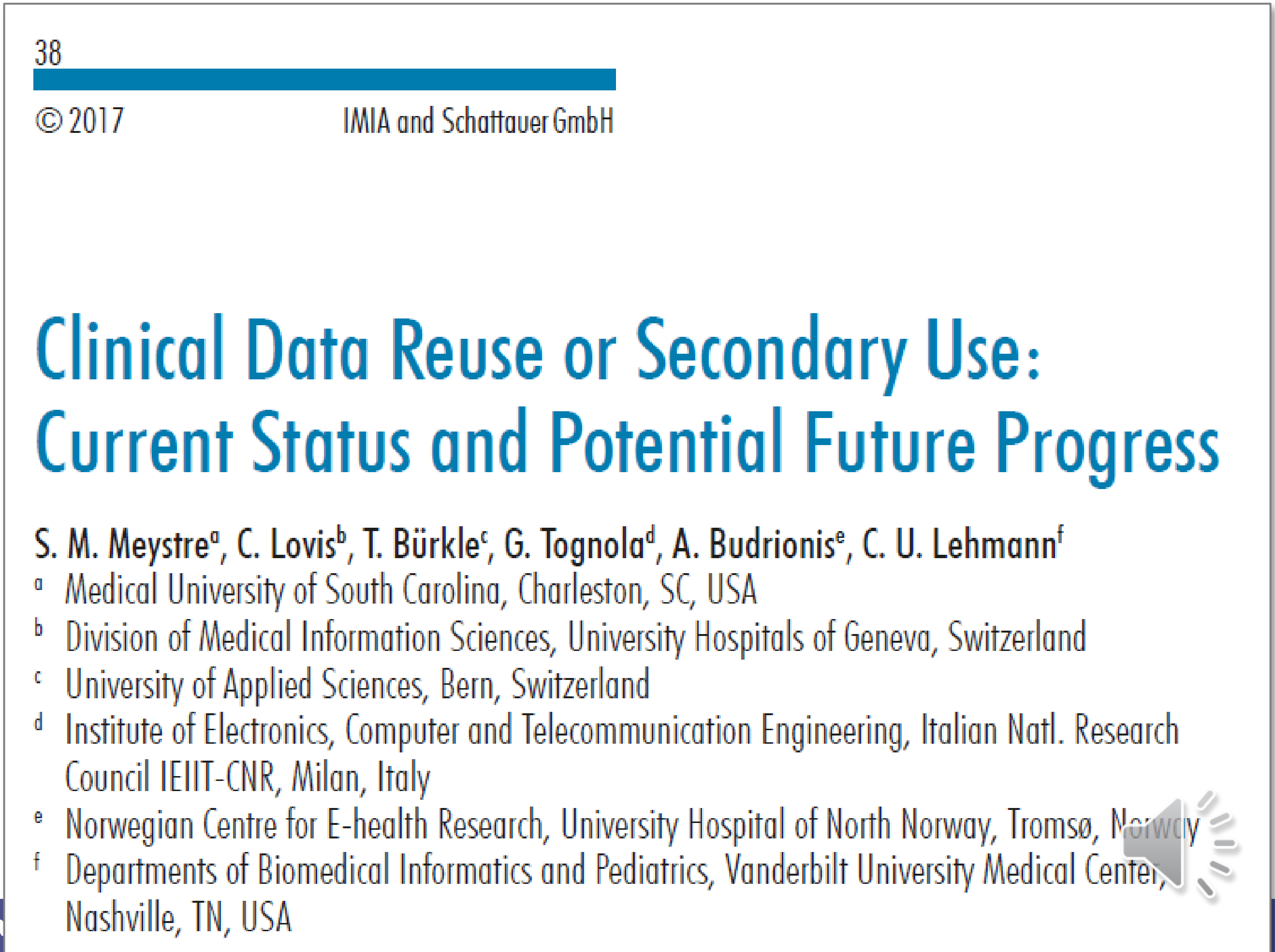
Feature extraction, example 3: administered drugs



- Formally:
  - Example of 1 patient, 2 administered drugs
  - Before: 1 table with 1 line + 1 table with 2 lines
  - After: 1 table with 1 line

The place of domain-specific knowledge

- Literature review published par Meystre et al. in 2017
- Classifications performed by Arnaud Dezetrée & Adrien Lecoeuvre



38

© 2017 IMIA and Schattauer GmbH

Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress

S. M. Meystre[a], C. Lovis[b], T. Bürkle[c], G. Tognola[d], A. Budrionis[e], C. U. Lehmann[f]

[a] Medical University of South Carolina, Charleston, SC, USA

[b] Division of Medical Information Sciences, University Hospitals of Geneva, Switzerland

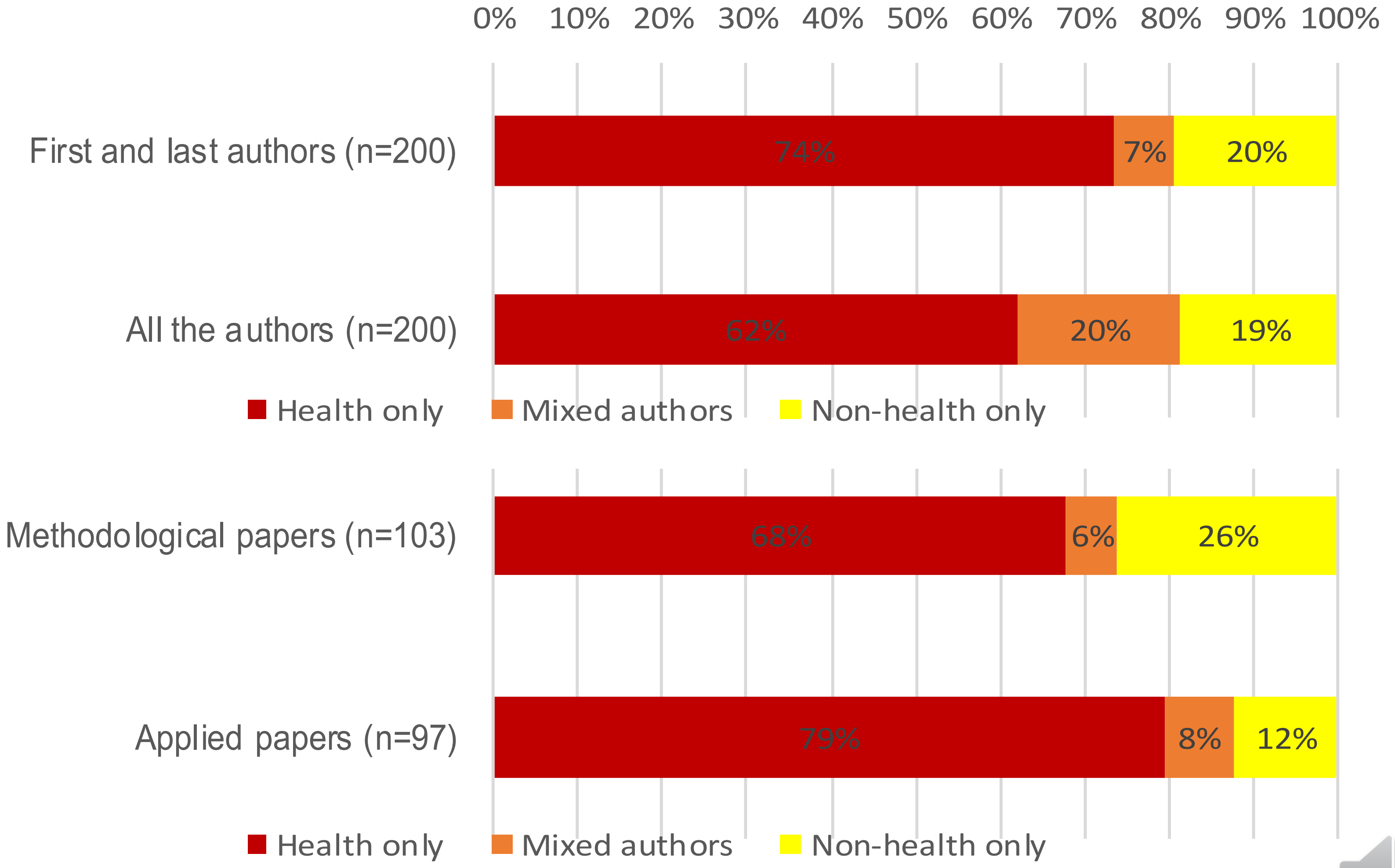
[c] University of Applied Sciences, Bern, Switzerland

[d] Institute of Electronics, Computer and Telecommunication Engineering, Italian Natl. Research Council IEIIT-CNR, Milan, Italy

[e] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway

[f] Departments of Biomedical Informatics and Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

The place of domain-specific knowledge

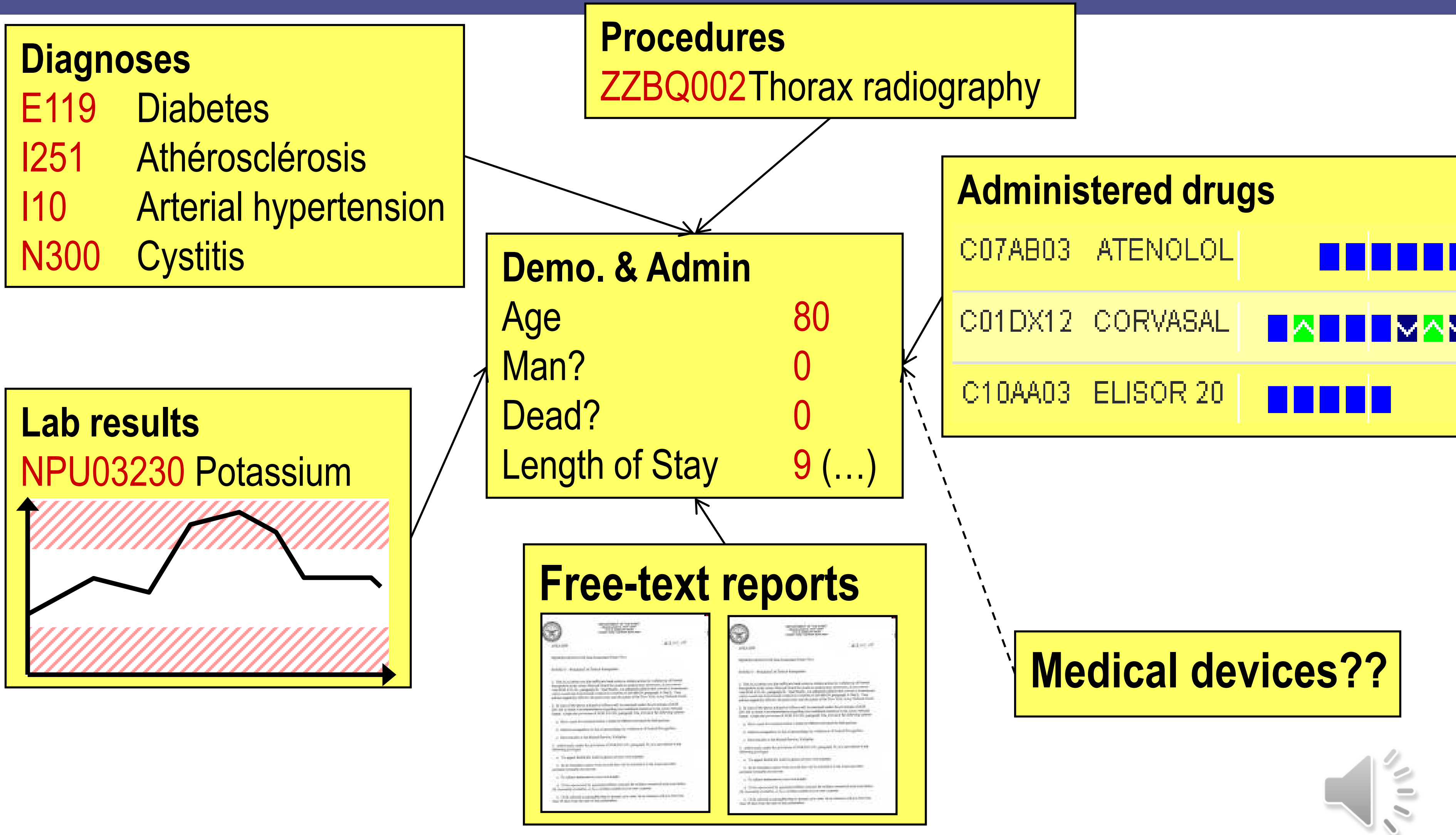


Data quality assessment

- Getting reliable data is a challenge => an iterative quality control is mandatory
- Extraction format, basic requirements (tables, fields)
- Single value validity. E.g.:
  - Incorrect type: Age="old"
  - Impossible value: age=141, diagnosis="HHFA001"
  - Out-of-terminology value: diagnosis="B99.0"
- Univariate validity. E.g.:
  - Each value is possible, but Mean(age)=85
  - Contextual: mean(age)=21 in Pediatrics
- Bivariate validity. E.g.:
  - Length of stay=2, admission="2013-05-14", discharge="2013-05-14"
  - Age=21 in a Geriatrics Unit
  - Age=21 with diagnosis="Alzheimer disease"

Data from electronic health records (EHR)

Data reuse of EHR



Diagnoses

E119	Diabetes
I251	Athérosclérosis
I10	Arterial hypertension
N300	Cystitis

Procedures

ZZBQ002Thorax radiography

Administered drugs

C07AB03	ATENOLOL
C01DX12	CORVASAL
C10AA03	ELISOR 20

Demo. & Admin

Age	80
Man?	0
Dead?	0
Length of Stay	9 (…)

Lab results

NPU03230 Potassium

Free-text reports

Medical devices??

Data reuse of EHR
Terminologies

- Vocabulary, support of semantic interoperability:
  - Couples of codes and wordings
  - Codes are sometimes included in a hierarchy
  - Enable to associate each concept to a code
  - Contrary to free-text, unambiguous and usable for statistical analysis
- Example :
  - Érysipèle = érésipèle = dermo-hypodermite = A46 in ICD10

Laboratory results

- Biochemistry, hematology, bacteriology, virology, immunology
- Examples terminologies:
  - LOINC (recommended)
  - IUPAC
  - …
- But most of the time, each laboratory produces its own terminology:
  - No semantic interoperability:
    - Impossible to pool 2 hospital databases
    - In the same hospital, data are sometimes heterogeneous according to the date of the measure
  - Need for producing *ad hoc* mappings

Data reuse of HER
Drug terminologies

- Commercial name
  - Ex : Dafalgan®
- CID Common international denomination
  - Ex : Paracétamol / Acetaminophen
- Various terminology
  - Among them, the ATC classification
  - ATC = Anatomical Therapeutic and Chemical classification
  - Codes, wordings and hierarchy (principally based on the therapeutic indication)

Example of the Aspirin in the ATC classification

A alimentary tract and metabolism
↳ A01AD other agents for local treatment
↳ A01AD05 …aspirin…

B blood and blood forming organs
↳ B01AC platelet aggregation inhibitors
↳ B01AC06 …aspirin…

C cardiovascular system
↳ C10BX (…) other combinations
↳ C10BX01 & C10BX02 …aspirin…

M musculo-skeletal system
↳ M01BA anti-inflammatory
↳ M01BA03 …aspirin…

N nervous system
↳ N02BA salicylic acid and derivatives
↳ N02BA01, N02BA51, N02BA71 …aspirin…

Data reuse of EHR
Some data sources…

- Drug (distinguish drug prescription and drug administration). Don't forget to add some sources:
  - Some procedures to map, e.g.:
    Scanner with iodine => "iodine"
    Surgery with anesthesia => "anesthetic drug"
  - Perfusion, that are often provided without terminology code
  - The drugs the patient brings with him…
- Medical devices (implanted or not):
  - Raw data: huge amount
  - Aggregated data
  - Results of the interpretation automatically done by the device

Definition of medical devices

- A medical device is:
  - an instrument,
  - apparatus,
  - implant,
  - in vitro reagent,
  - or similar or related article
- that is used to diagnose, prevent, or treat disease or other conditions,
- and does not achieve its purposes through chemical action within or on the body (≠drug)

Summarized from the FDA's definition."Is The Product A Medical Device?". U.S. Department of Health and Human Services -. U.S. Food and Drug Administration. 10 June 2014

Definition of medical devices

- This definition notably includes:
  - External devices that can be used for diagnosis, e.g. electrocardiograph
  - External devices that can support the medical decision, such as software
    e.g. clinical decision support systems
  - Internal devices that can be used to supply a physiological function, e.g. hip replacement implant
  - … and many others

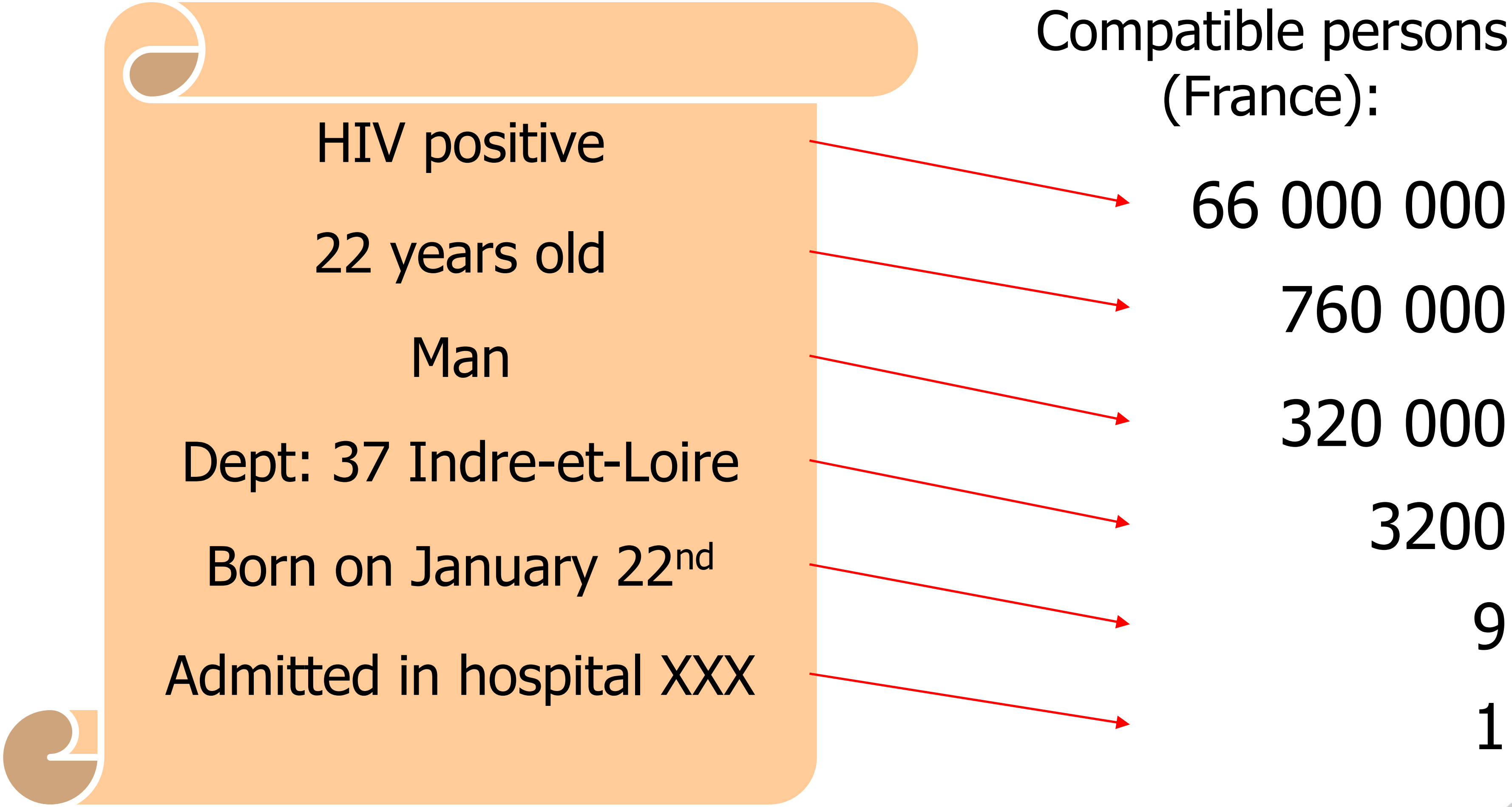
Issues regarding patient confidentiality

According to journalist Matt Lauer, M.S. has paid 10 millions dollars to silence those who wanted to spread the information (November 17th 2015)

Why, and how?

- Purposes:
  - Blackmail to get money, action or abstention
  - Slander, gossip
  - Estate transactions, life lease
  - Lapse of insurance contract
  - Etc.
- How? Differentiate:
  - Anonymous database
  - Nominative database
  - Indirectly nominative database (very frequent!)

A piece of paper found in the street…



HIV positive

22 years old

Man

Dept: 37 Indre-et-Loire

Born on January 22nd

Admitted in hospital XXX

Compatible persons (France):

66 000 000

760 000

320 000

3200

9

1

We did not even need the Zip code, the admission date, etc.

Some solutions

- Limit the available information
  - E.g.: Mister X's age = 50-60 years
  - Impedes Research
  - Limited efficiency
- Only spread aggregated information
  - E.g.: 430,000 men are 61 years old
  - False "open data": not data at all!
  - Nearly useless information
  - Claude Le Pen: "scientists are fond of individual data, but actually they don't mind about individuals".
- Data perturbation
  - E.g.: Mister X's age = 53.567 years (actually it was 55.249…)
  - Random, unpredictable and irreversible perturbation
  - Preserves statistical associations (and then usable for research)
  - Make all the information suspicious, even the unaltered one!
  - E.g.: the perturbed data from the Framingham cohort is available for teaching purposes