# Open & Big Data for Life Imaging

*Technical aspects : existing solutions, main difficulties*

Pierre Mouillard MD

Vigisys

CAMPING Toulouse by TIC Valley Kick-off for Start-ups

TIC valley concentré d'entreprises innovantes

TiC valley

concentré d'entreprises innovantes

FRENCH TECH TOULOUSE

SO START'UP !

# What is Big Data?

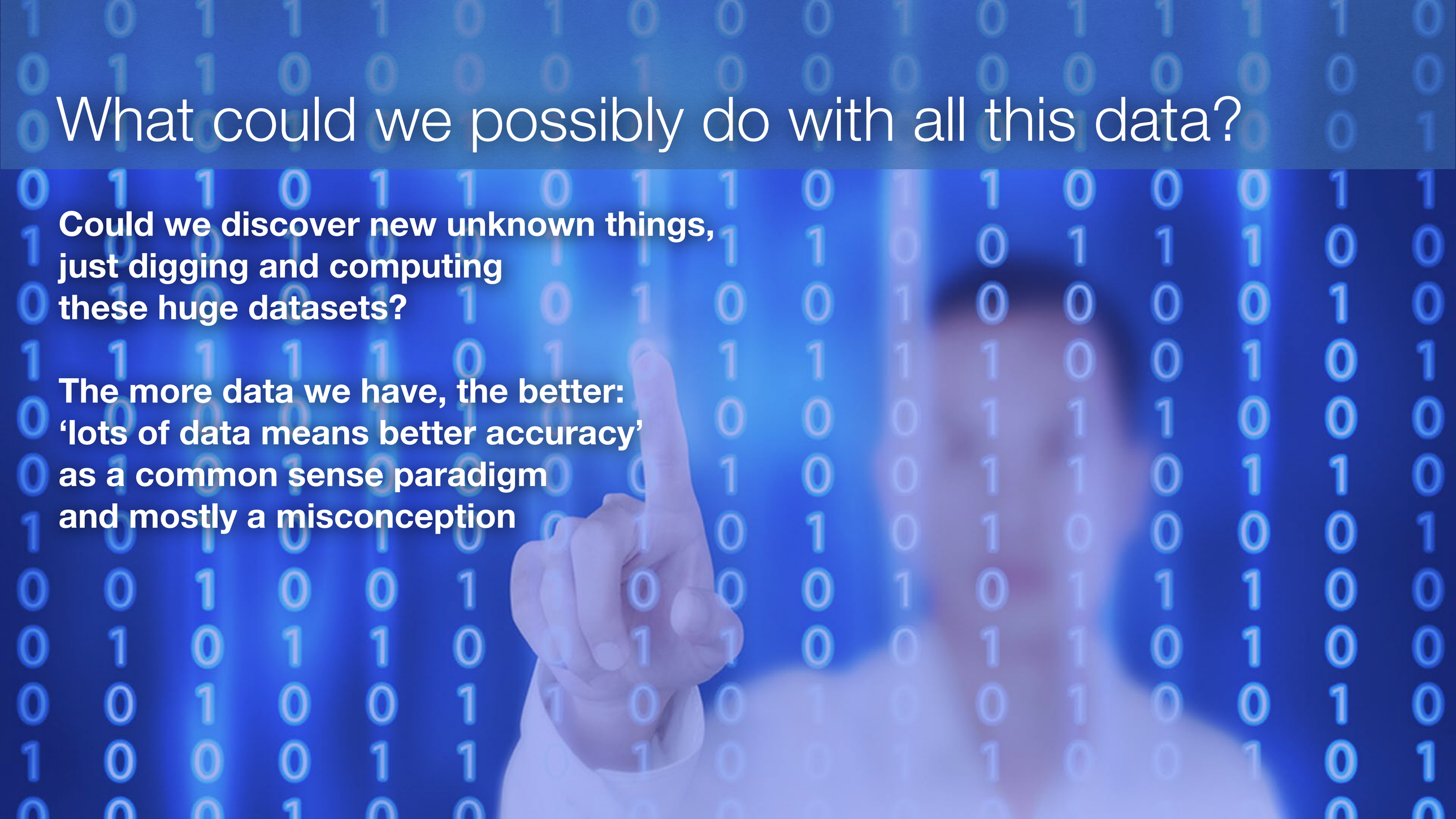
- lots of data

  *more than you can process using common database software and standard computers*

- complex data
- dataflows and time series

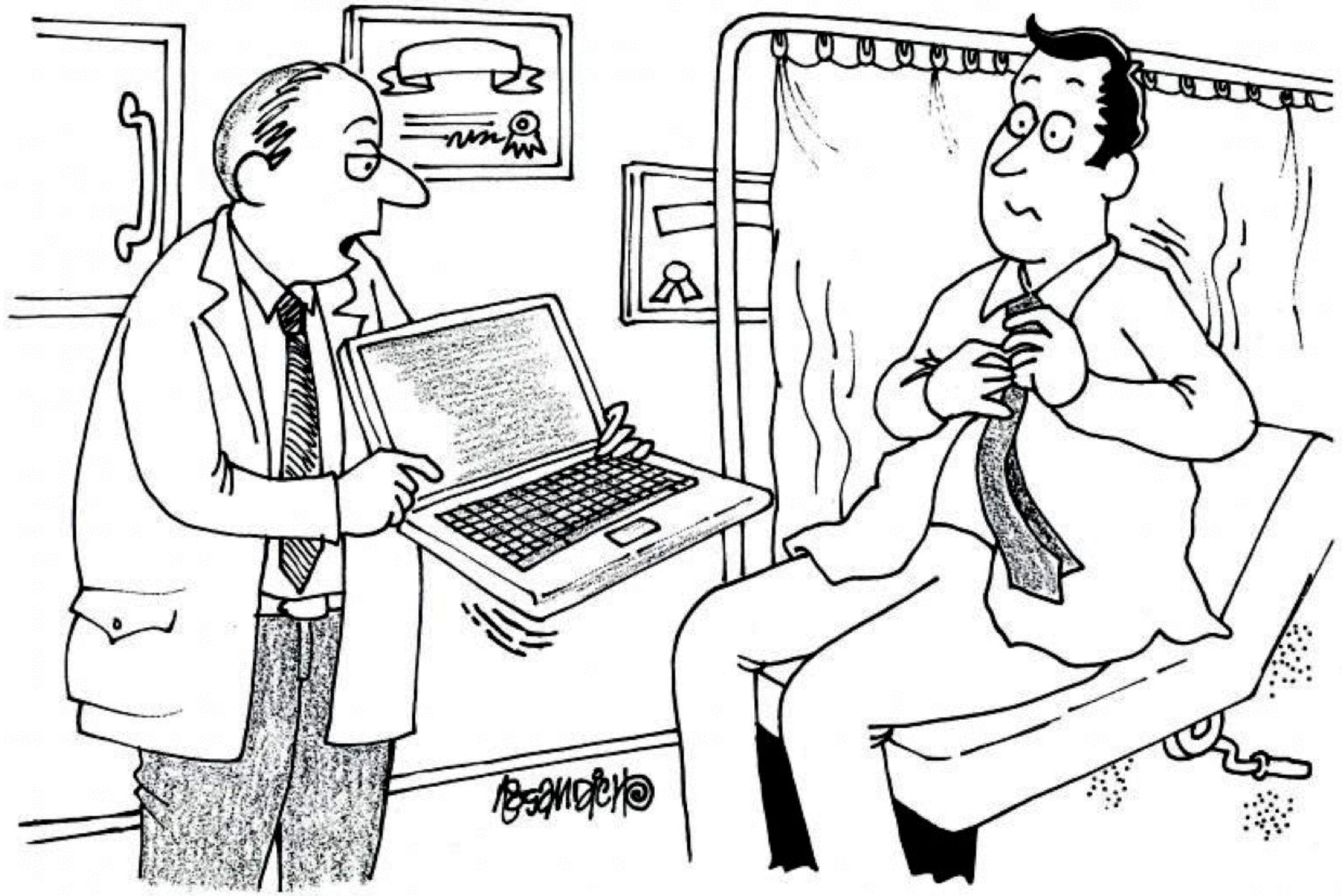
the origins: big science, demographics, economics
the next generation origins: e-commerce

# What could we possibly do with all this data?



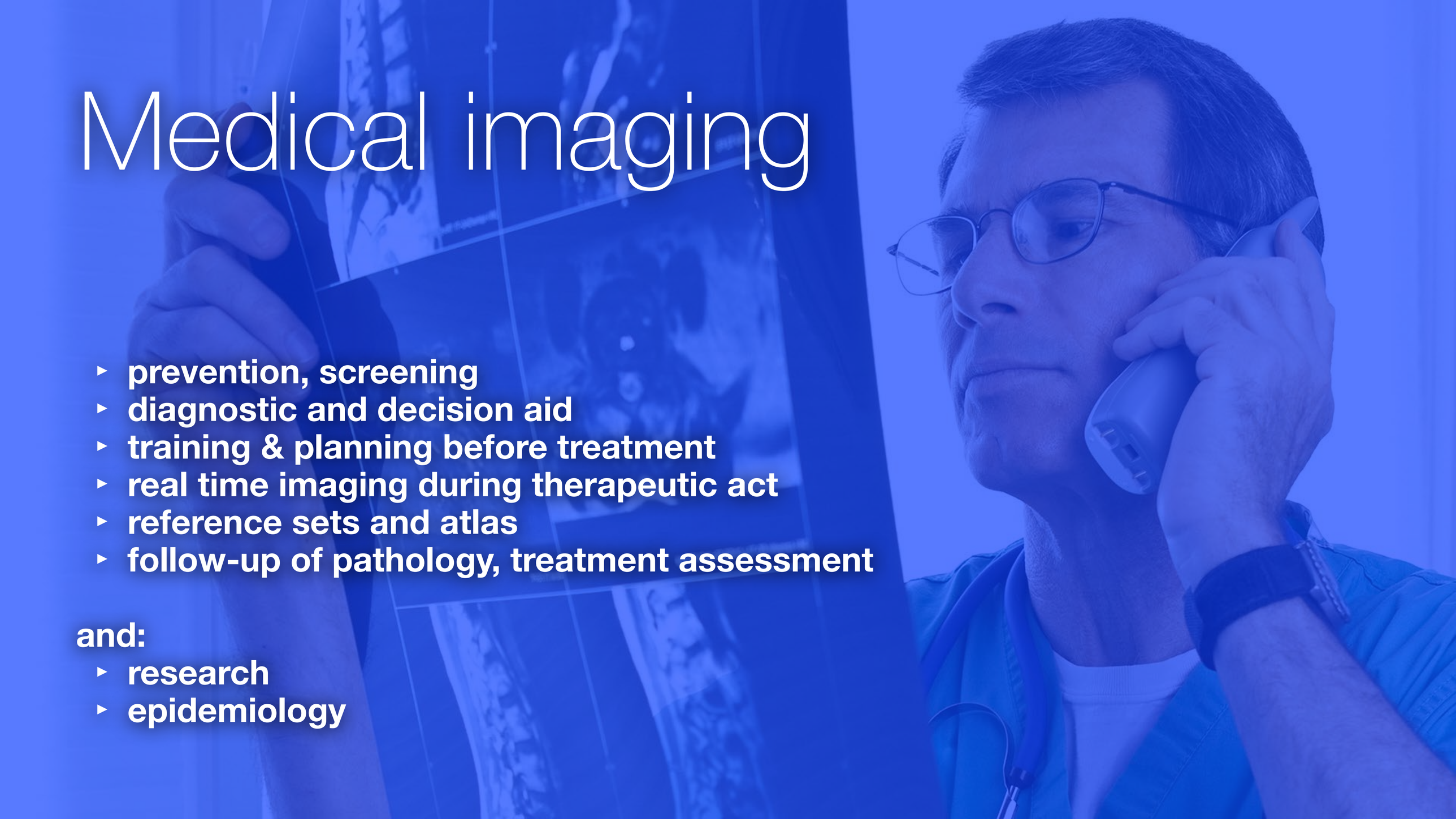
**Could we discover new unknown things,
just digging and computing
these huge datasets?**

**The more data we have, the better:
‘lots of data means better accuracy’
as a common sense paradigm
and mostly a misconception**



If you don't mind, I'll grab some big data from you.

# Medical imaging



- prevention, screening
- diagnostic and decision aid
- training & planning before treatment
- real time imaging during therapeutic act
- reference sets and atlas
- follow-up of pathology, treatment assessment

and:

- research
- epidemiology

# Big data & medical imaging

- **medical images are big:**
  **dynamic 3D CT scan = 1Gb, digitized XR = 100Mb**
- **an average hospital generate about 10-300 Tb / year**
- **mostly unstructured data (60-80%)**
- **medical image archives are increasing by 20-40% / year**
- **for one patient, average of 3 imaging modalities per medical act**

very different from e-commerce!
‘fewer’ instances, more data *per capita*

# Sharing medical images

- **medical practice becomes increasingly collective**
- **multi-modality (CT, MRI, Usound,  XR, PET…) means numerous experts involved**
- **tele-medecine, tele-diagnostics means remote access to images & medical data**
- **nosology and semiotics have to be redefined and expanded because of continuous advances in imaging**
- **research is more focused on specific diseases**

all this means we need to improve and facilitate medical image sharing

# Big data: what for?

- **numeric reference images databases**
- **patient similarity searching**
- **disease progression monitoring, clinical follow-up**
- **cases studies, training and learning, expertise sharing**
- **nosology and semiotics redefinition**
- **new algorithms and image editing tools testing**
- **shared archives**
- **epidemiology**

big & open data in medical imaging is a challenge, on a global scale

# Big data: 2 concepts

## STATIC BIG DATA

- **retrospective and prospective studies on finite sets of images**
- **statistical analysis at some point and conclusions**
- **= ‘knowledge extraction’ and rules definition**

## DYNAMIC BIG DATA

- **‘forever’ ongoing studies, with always expanding image sets**
- **auto-adaptative and machine learning systems (neural networks, genetic algorithms…)**
- **= ‘always improving’ and automatic optimization**

digital expertise and AI is rising!

# Imaging data

- images are just big packs of digits
- each pixel (voxel) is a measure (each image is a rich dataset)
- images are just raw data matrix, unstructured data
- to be used as big data, image processing is needed

- image has to be normalized, segmented, and expertized
- ROI definition, measures of volumes & distances, characterization of structures, 3D reconstruction, dynamic analysis…
- automatic processing or manual (assisted)
- metadata is the key for relevant retrieval and selection

and clinical context data is always needed!

# And then came the Internet…

- **Internet technologies are becoming the *de facto* standards for data sharing, collaboration, and global access to information**
- **Internet is now the strongest incentive for technological innovation in IT**

# Technological trends

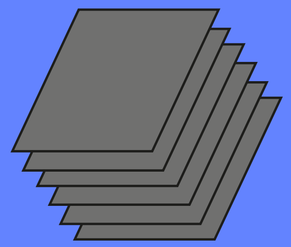
- **data encapsulation**
- **distributed storage**
- **processing and storage virtualization**
- **data centric architectures**
- **APIs, front / back dissociation**
- **new databases systems**
- **security**
- **open formats, open source, normalization**

innovation is coming from high demand Internet application projects:
social networks, large e-commerce platforms, data sharing clouds

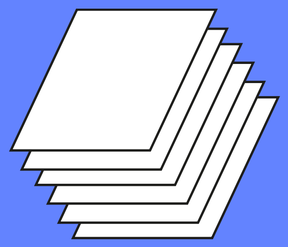
# Imaging data object

encapsulation

images



overlays



3D infos

settings

measures

tags

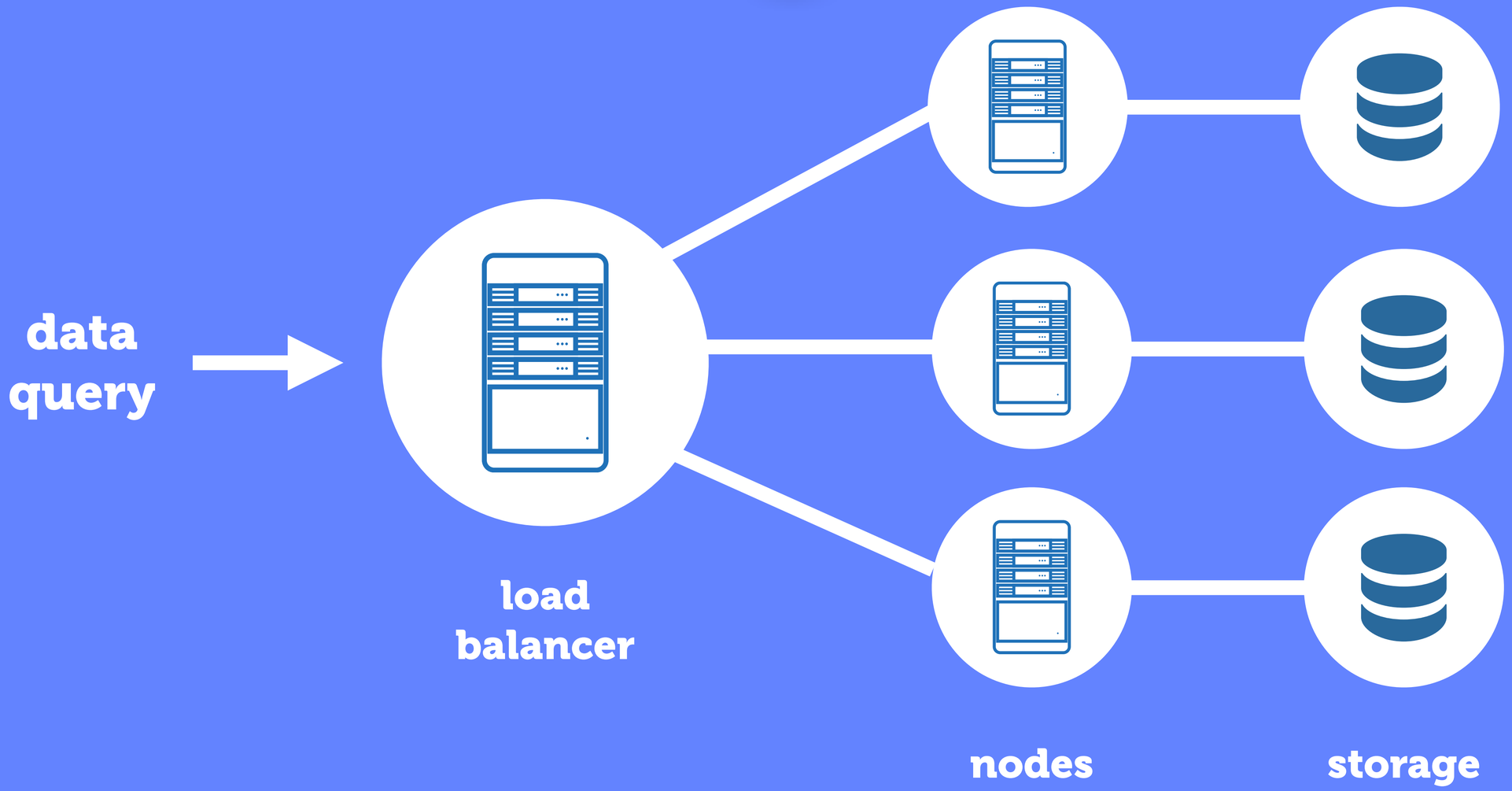
patient ID

report

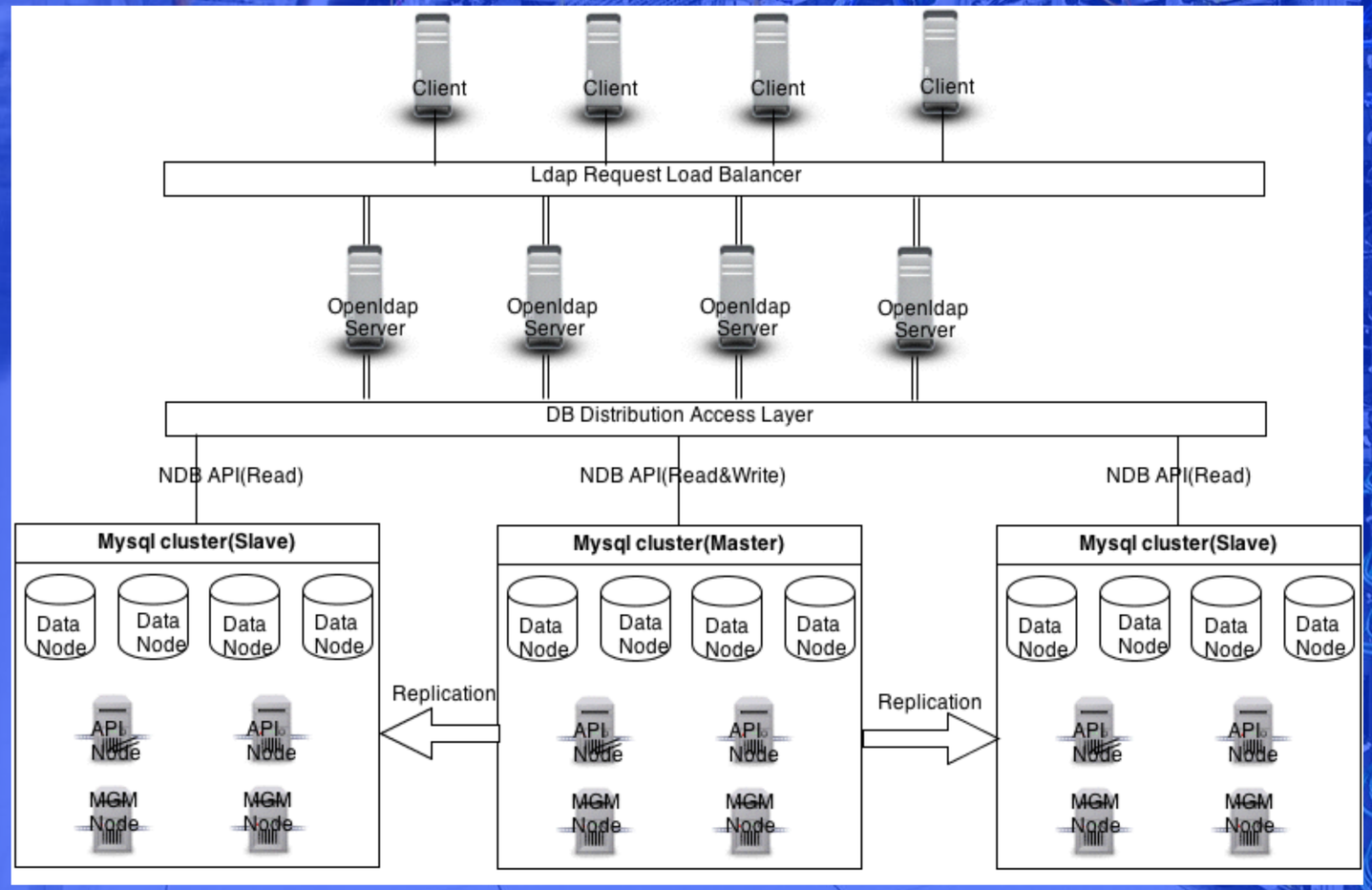
clinical data

DICOM

# Distributed storage



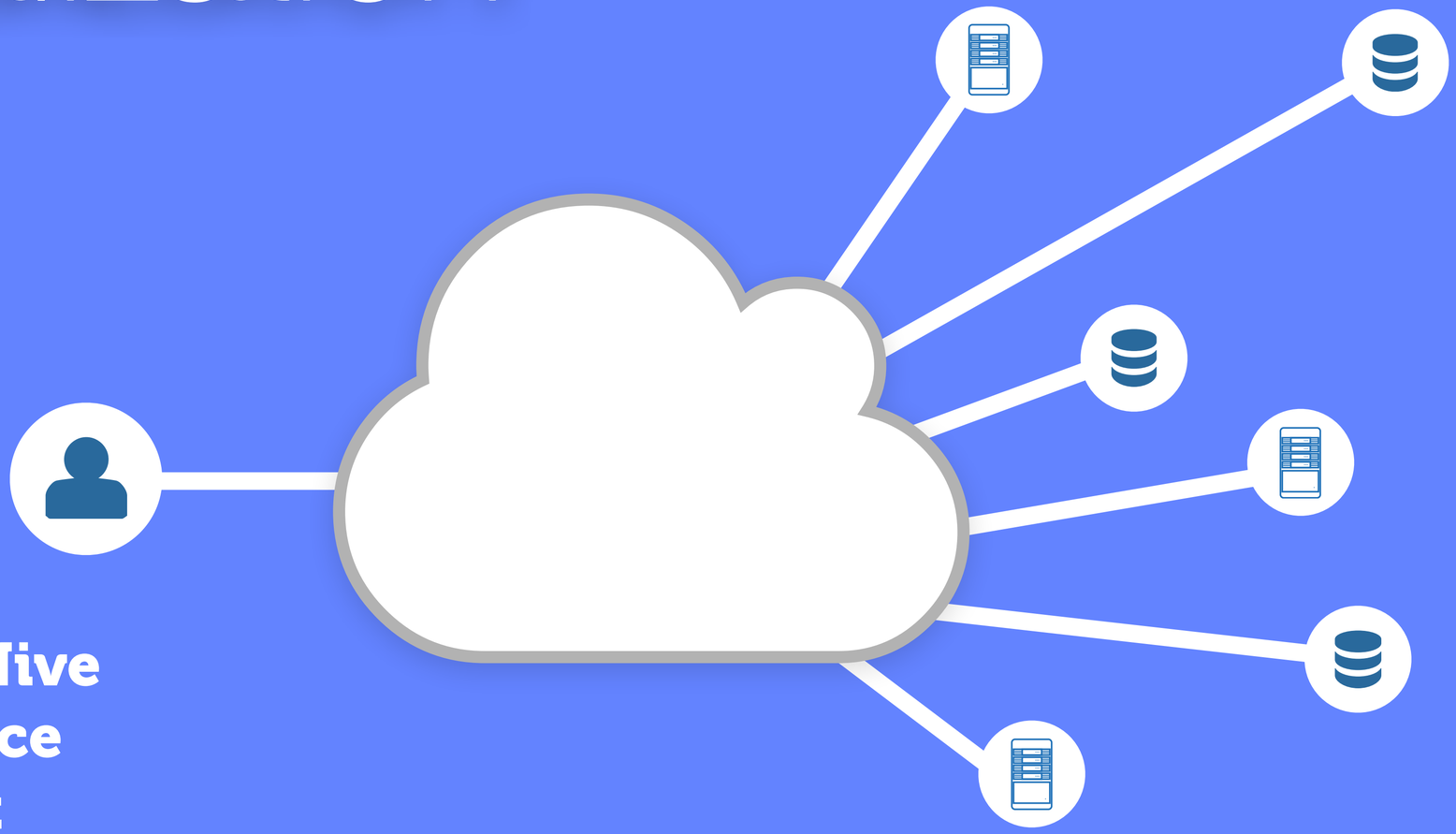
# Distributed storage




**outsourced hosting**
**health data agreement**

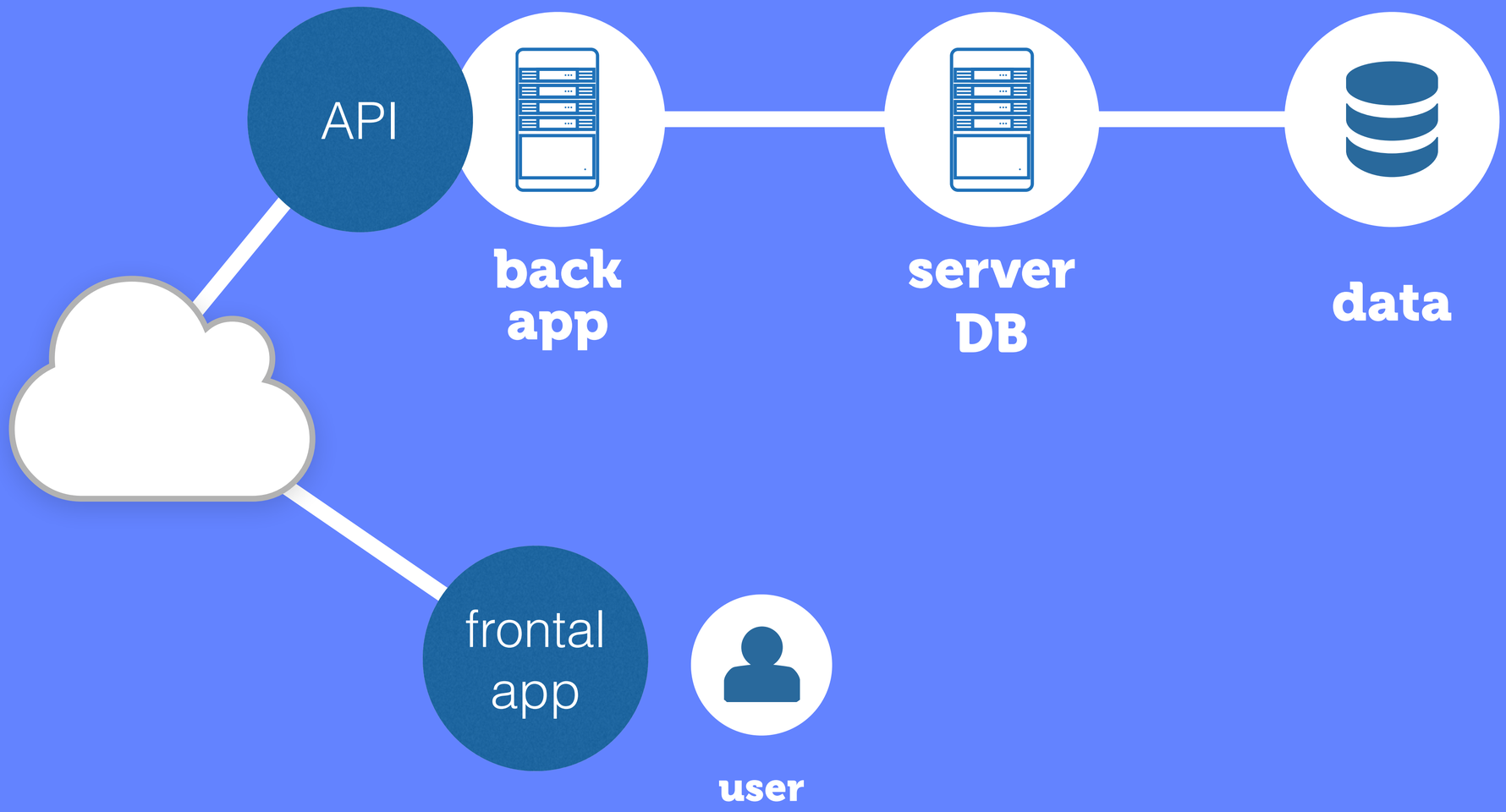

Google
twoden
049

# Processing & storage virtualization



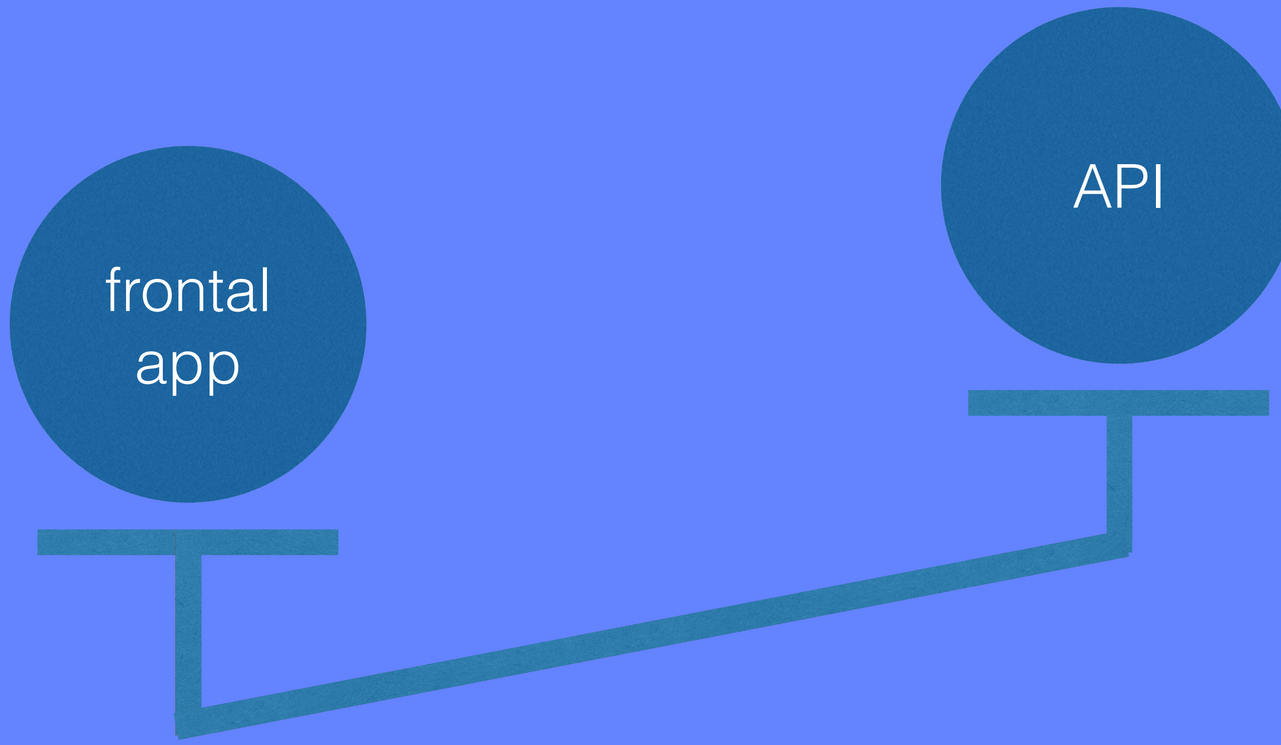
**Hadoop / Hive**
**Map reduce**
**Mahout**

# APIs - front / back



# APIs - front / back

**store-search-retrieval**
**batch processing**
**distributed computing**

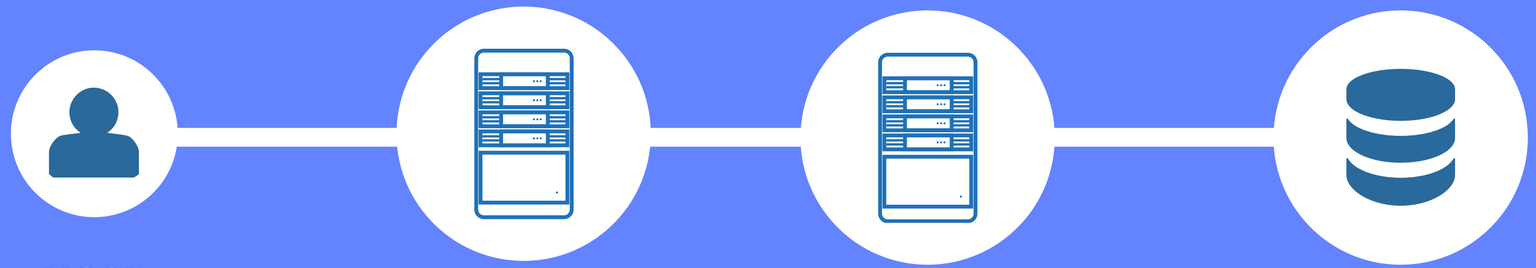



**server**

**user computer**

**pre-processing**
**normalization**
**individual analytics**
**interface**

# Flat database systems



user

app server

SGDB

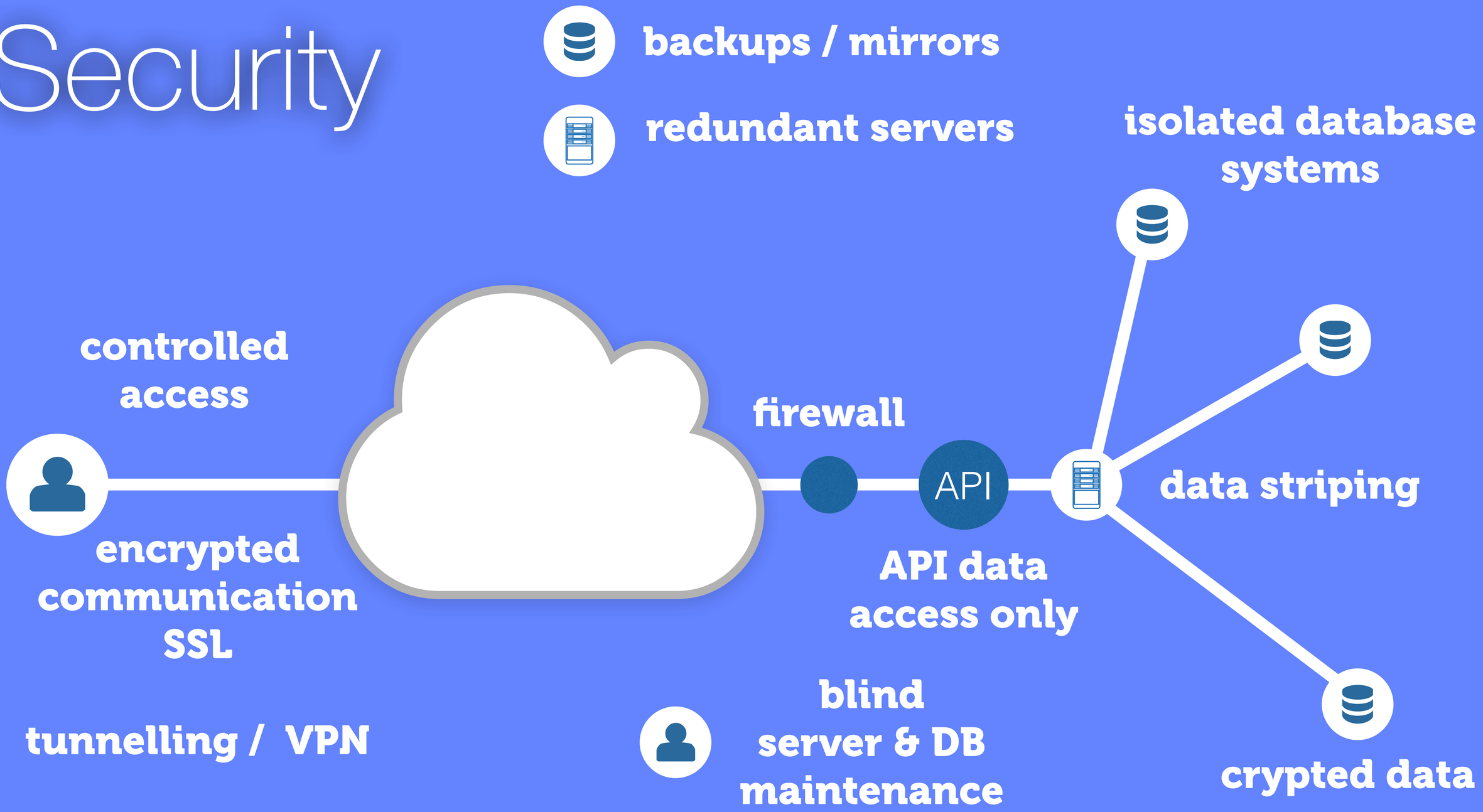
data in proprietary format
(no direct access)

user

app server

flat data (file system access)
NoSQL - Cassandra, MongoDB

# Security


backups / mirrors
redundant servers
isolated database systems
controlled access
firewall
API
data striping
encrypted communication SSL
API data access only
tunnelling / VPN
blind server & DB maintenance
crypted data

# SaaS


scalability
virtualization
cloud
cluster

hosting
outsourcing

log / data supervisor

WEB APP (SaaS)

logs

CRUD APP +interface

controlled access

imagery system

linked data

open data

data providers

other SaaS

API sandbox

reports boards stats

Queries

data replication recordsets

anonymization

database backups

snapshots ghost server

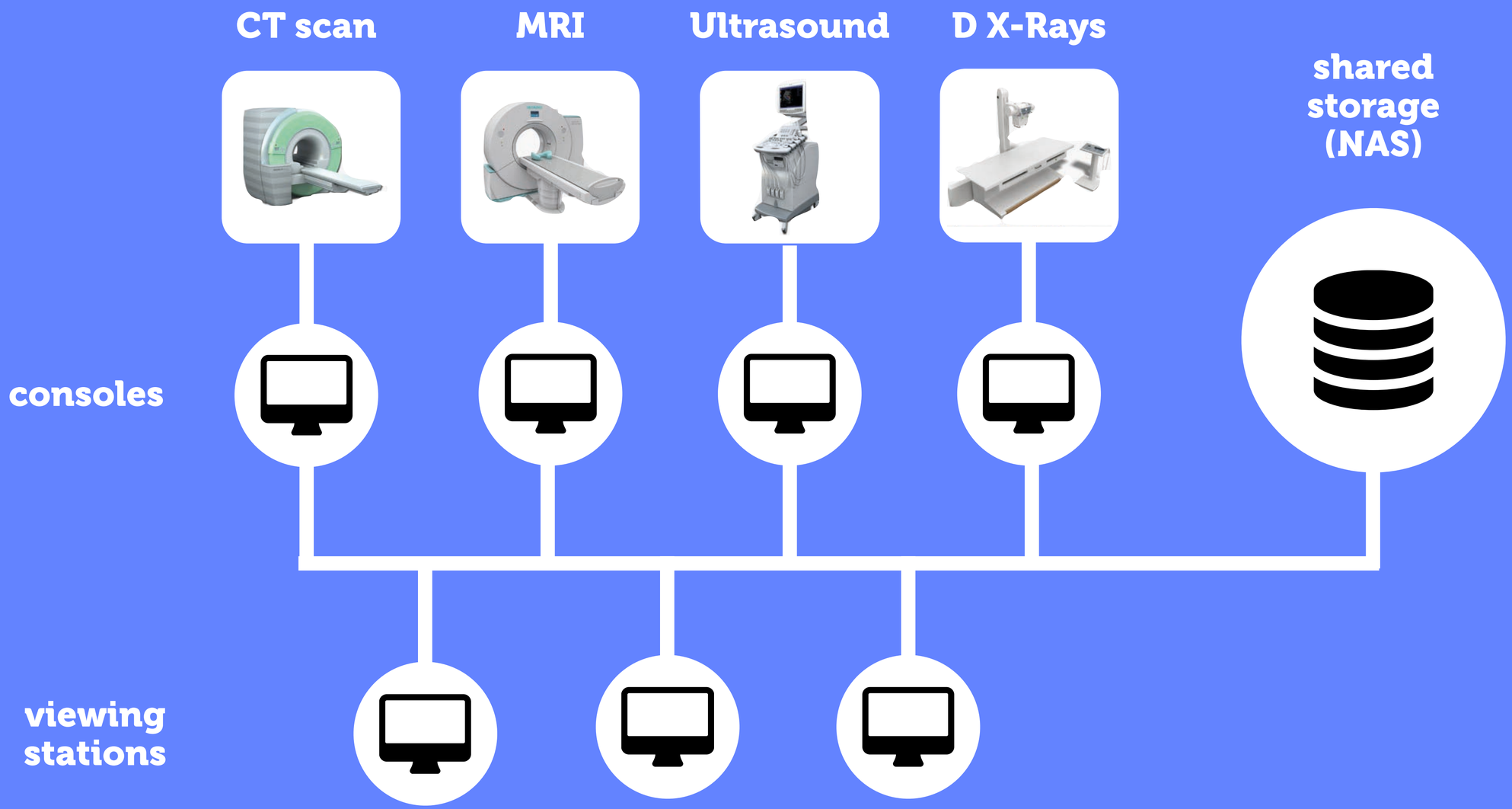
# Imaging data architecture

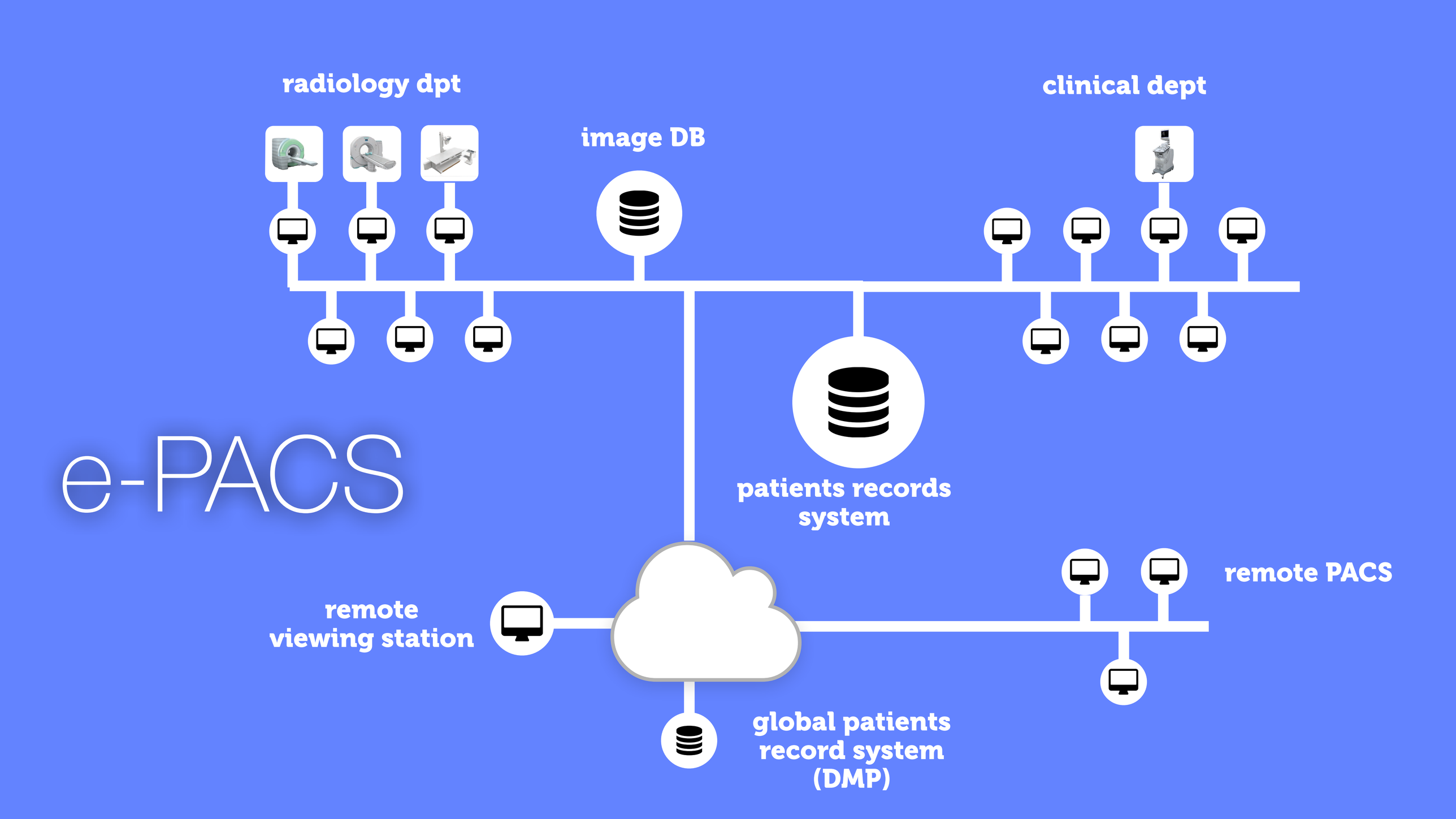
- **non vendors PACS (open PACS) - DCM4CHEE, Osirix server, PACSOne**
- **open source visualization software :  Osirix, Weasis (webapp)**
- **BYOD and scaling down: standard desktop computer, laptop, even tablets as viewing stations**
- **accessing the PACS at the patient bed or from remote location**
- **linking with in house medical record system and global medical record system**

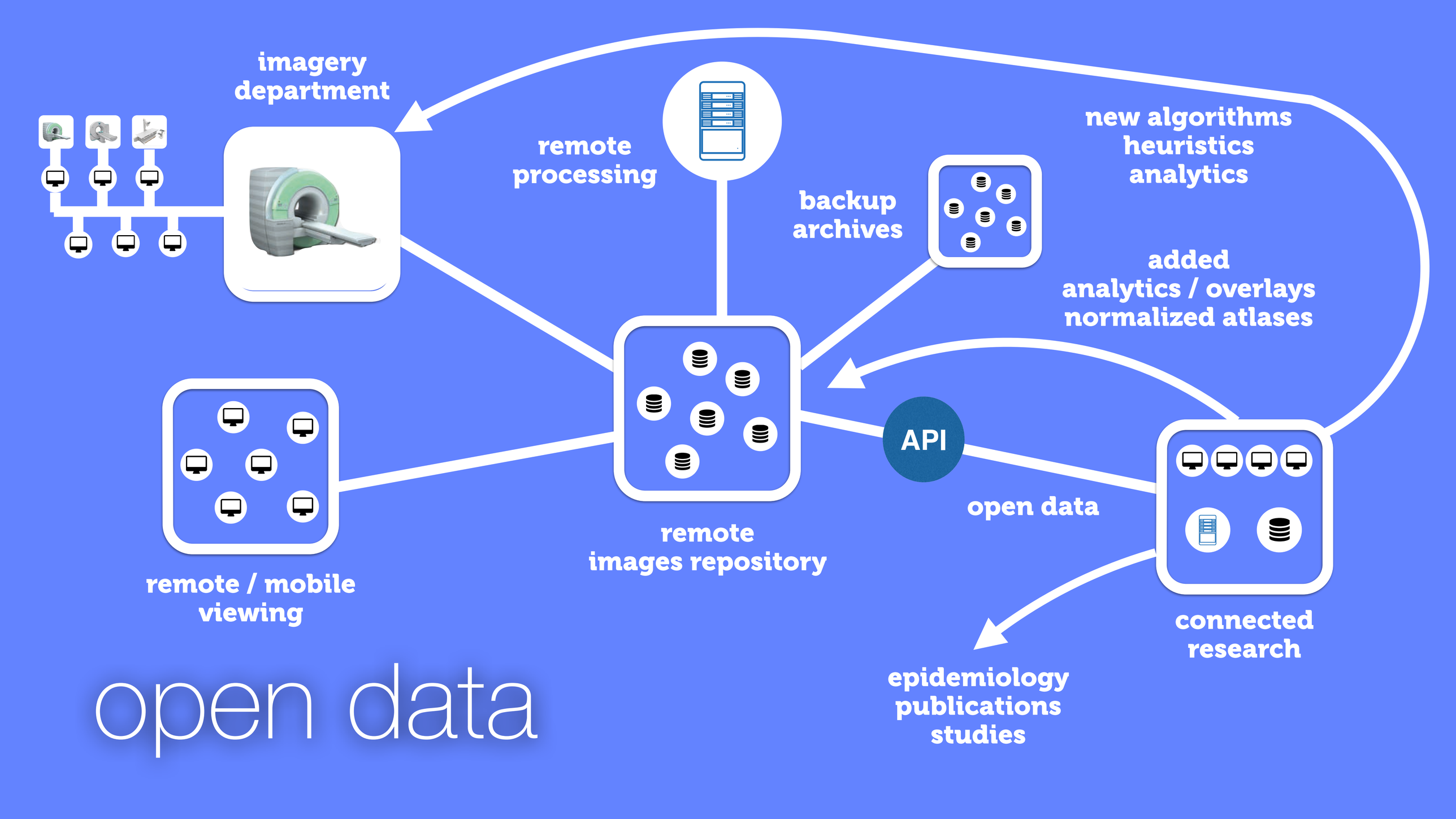
- **research connexion: direct access to dynamic big data**
- **integration of new algorithms as add-ons modules (local processing)**

objective : promote imaging as actionable data,
for research and benefit of the other patients

# standard PACS




radiology dpt
image DB
clinical dept
e-PACS
patients records system
remote viewing station
remote PACS
global patients record system (DMP)


imagery department
remote processing
backup archives
new algorithms heuristics analytics
added analytics / overlays normalized atlases
API
open data
remote / mobile viewing
remote images repository
connected research
epidemiology publications studies

# open data

# difficulties

- pooling images from different sources is difficult: different resolutions, different angles, registration, quality of image…
- image analysis is computing intensive: but radiologists cannot wait too long for results in an everyday use
- who wants to share? medical conservatism… too much privacy regulation?
- extraction of the important and relevant features in images is a daunting task
- image formats, clinical records are not always coherent
- incompatible proprietary systems - we need better open source software
- what do we do with old data? heterogeneous data?
- ethics: who owns the data? what about patient consent?
- security concerns: black hat hackers are everywhere
- the IT people should not access to patient's health data
- how to integrate into PACS new algorithms modules?
- who pays for this?

# perspectives

- new imaging sources: mobile ultrasound, mobile MRI
- connected objects: calibrated images taken with smartphone
- dynamic images and time series
- robotic surgery
- in vivo image multimodal overlay (guided surgery)
- PACS app for iPhone?

# thank you!


**pierre.mouillard@vigisys.fr**

@pierremouillard

pierre.mouillard

pierre mouillard