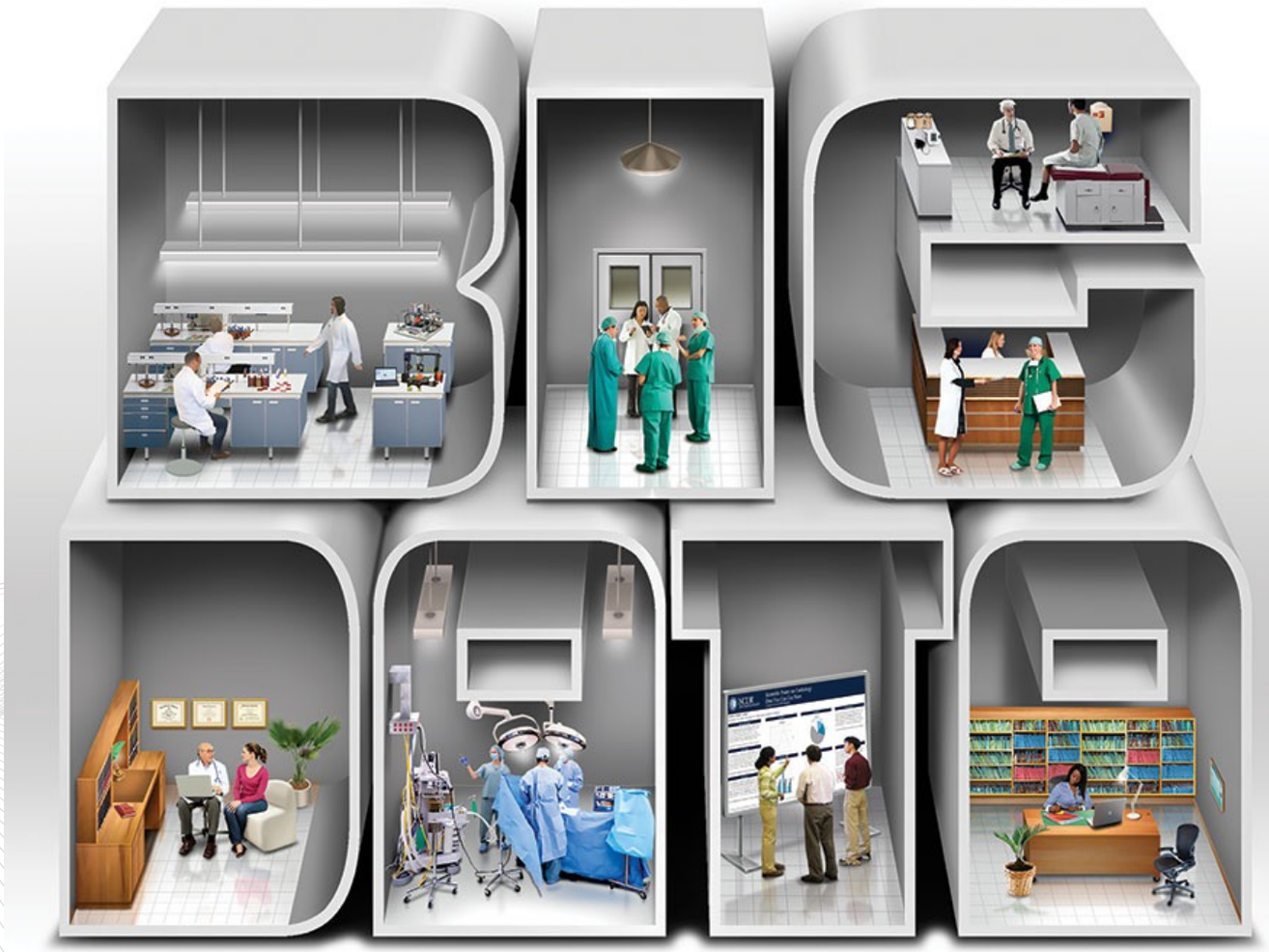
From Big Data to Big Insights



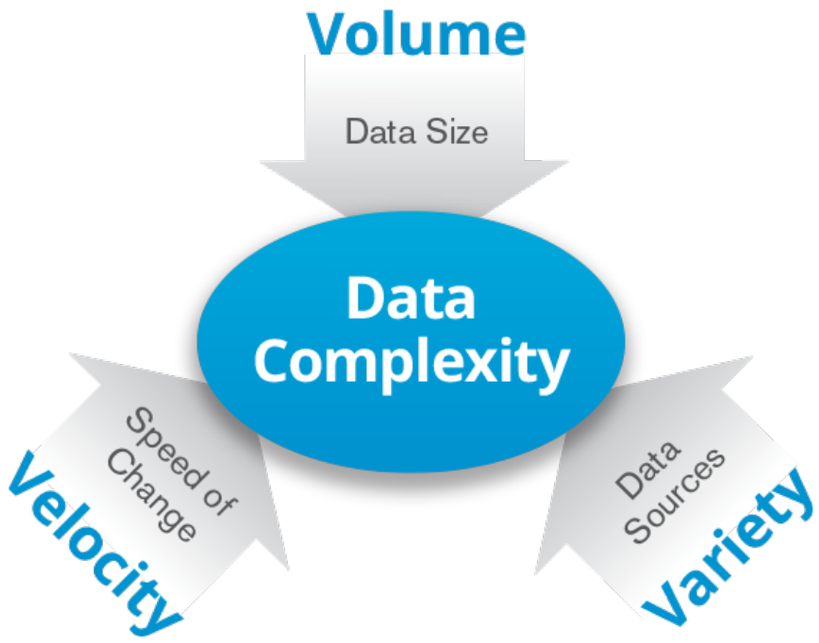
ComPh modeling week

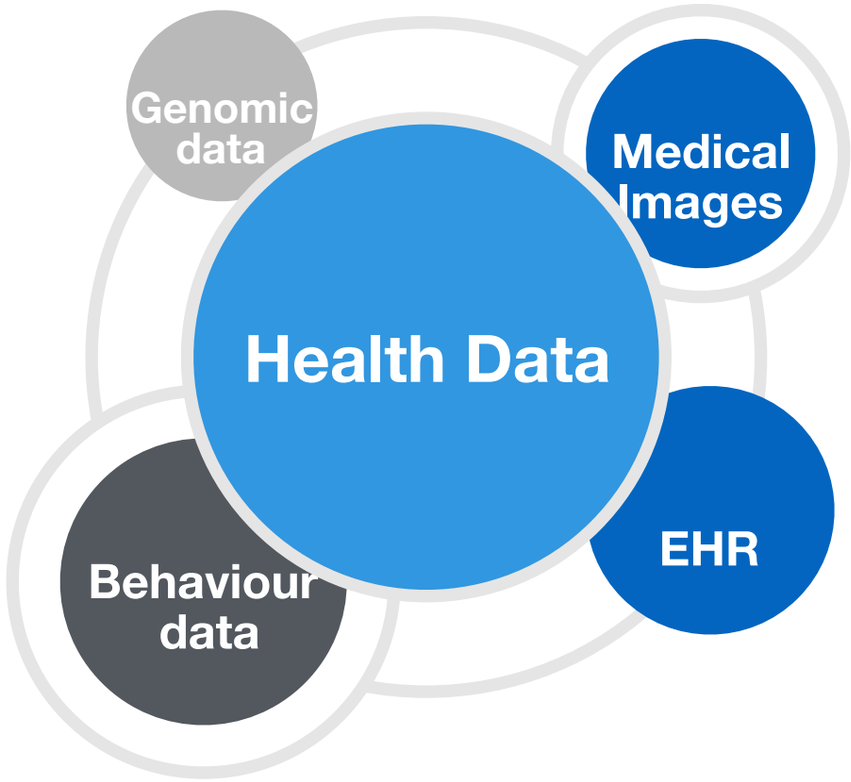
March 20-24, 2017

Valeriya Naumova

Simula Research Laboratory AS

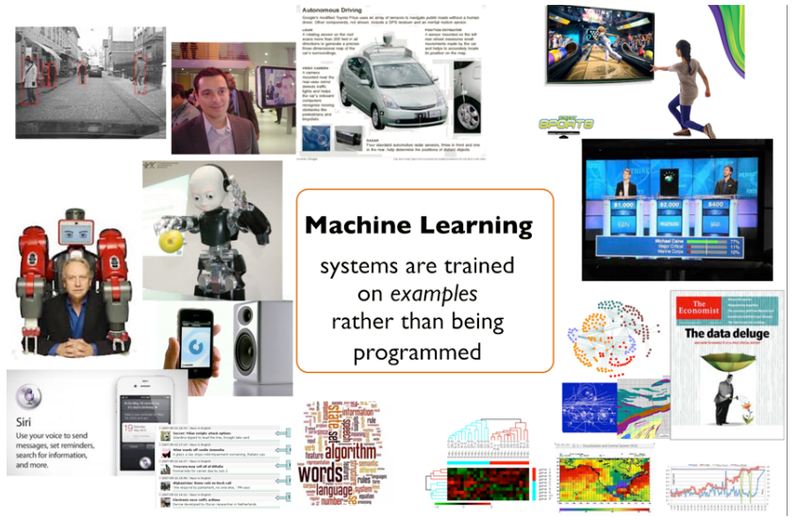
Table of Contents





What is Big Data?

Sources and Techniques for Big Data in Healthcare



Big Data Analysis - How?

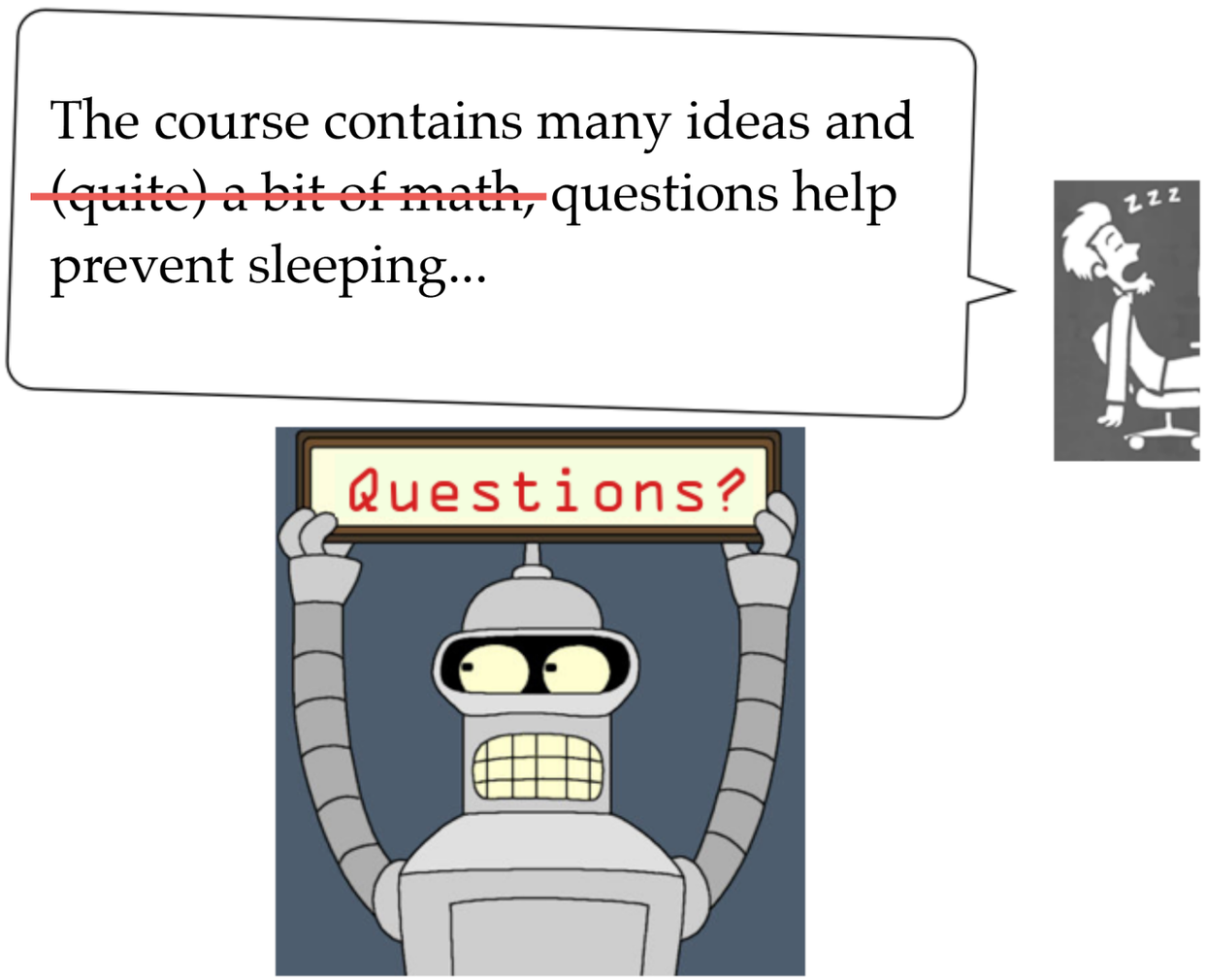
Colour Coding / Feedback

Really important stuff

Important stuff

Regular stuff

Let us know if you have comments, concerns, suggestions!



Short Curriculum Vitae

Relevant positions

- **01/2016-present**: Senior Research Scientist, SSRI, Norway.
- **01/2014-12/2015**: EU Coordinator in Computational Biomedicine, Simula, Norway.
- **06/2010-12/2013**: Research Scientist, Johann Radon Institute for Computational and Applied Mathematics (RICAM) Austrian Academy of Sciences, Austria.
- **06/2010-08/2012:** Collaborator and Participant of the large-scale EU-FP7 project DIAdvisor: Personal Glucose Predictive Diabetes Advisor, Austria.

Education and academic qualification

- **2012:** Doctoral degree in Natural Sciences (Applied and Computational Mathematics), Johannes Kepler University Linz, Austria.
- **2010:** Master degree in Intellectual Decision-Making Systems, National University of Kyiv-Mohyla Academy, Ukraine.
- **2008**: Bachelor degree in Computer Science, National University of Kyiv-Mohyla Academy, Ukraine.

Data scientist ... 'sexiest job of the 21st Century'.

Harvard Business Review

In God we trust, all others bring data.

William Deming, Engineer & Statistician

The ~~four~~ five dimensions (V's) of Big Data

- A collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications.

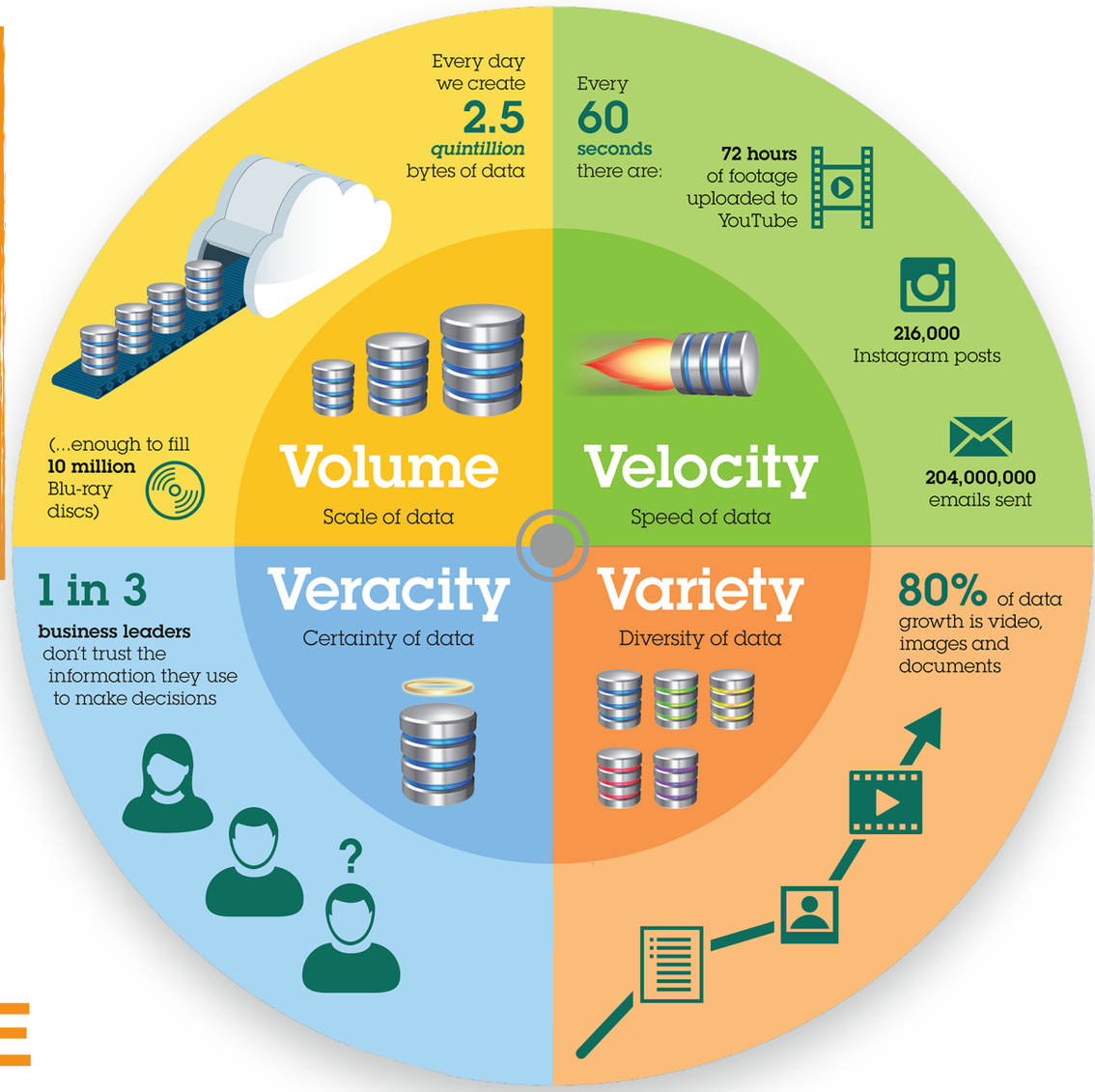
- "Big data refers to the tools, processes and procedures allowing an organisation to create, manipulate, and manage very large data sets and storage facilities".

according to zdnet.com

Big data is not just about size.

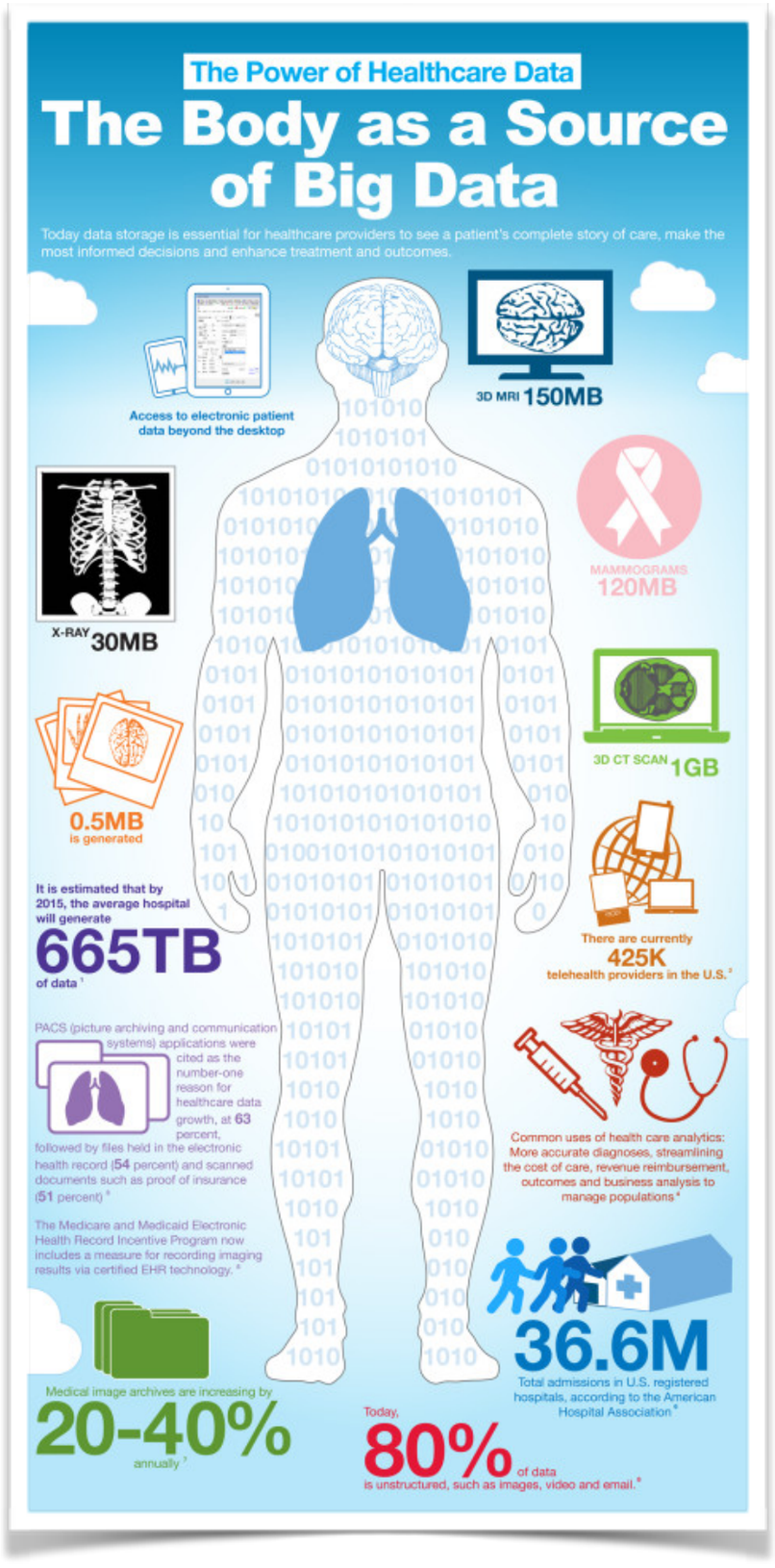
- Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data.
- It aims to answer questions that were previously unanswered.

The challenges include capturing, storing, searching, sharing & analysing.



+ VALUE

Digital Medical Data Growth

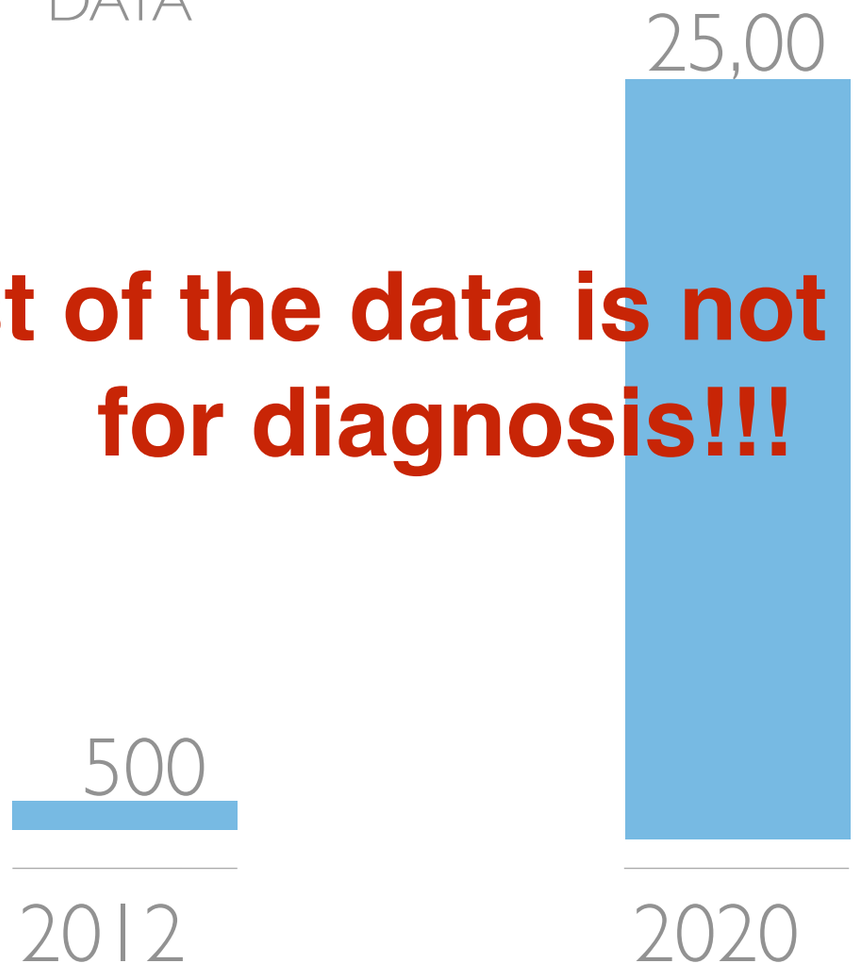


12,000 +

Possible medical diagnoses

WHO ICD-10 codes

PETABYTES OF DATA



Most of the data is not used for diagnosis!!!

Medical Imaging / EHR / Genomics data / Wearables

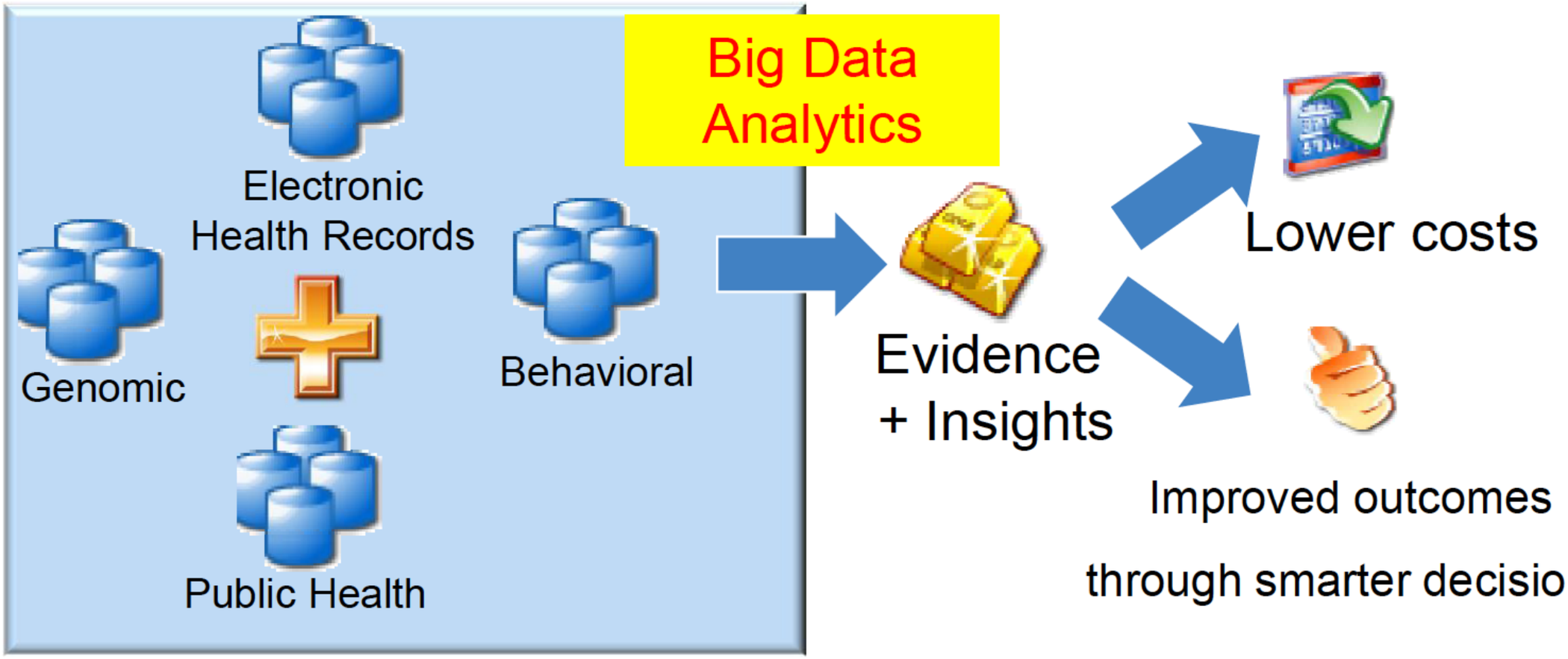
Big Data Challenges in Healthcare

“We need to learn how to leverage these new data to prevent disease, identify disease earlier, and recognize patterns of care and trends in outcomes.”

Richard Kovacs, chair of ACC’s Clinical Quality Committee.

- Inferring knowledge from complex heterogeneous patient sources. Leveraging the patient/data correlations in longitudinal records.
- Understanding unstructured clinical notes in the right context.
- Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.
- Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity.
- Capturing the patient’s behavioral data through several sensors; their various social interactions and communications.

Big Data Analytics in Healthcare

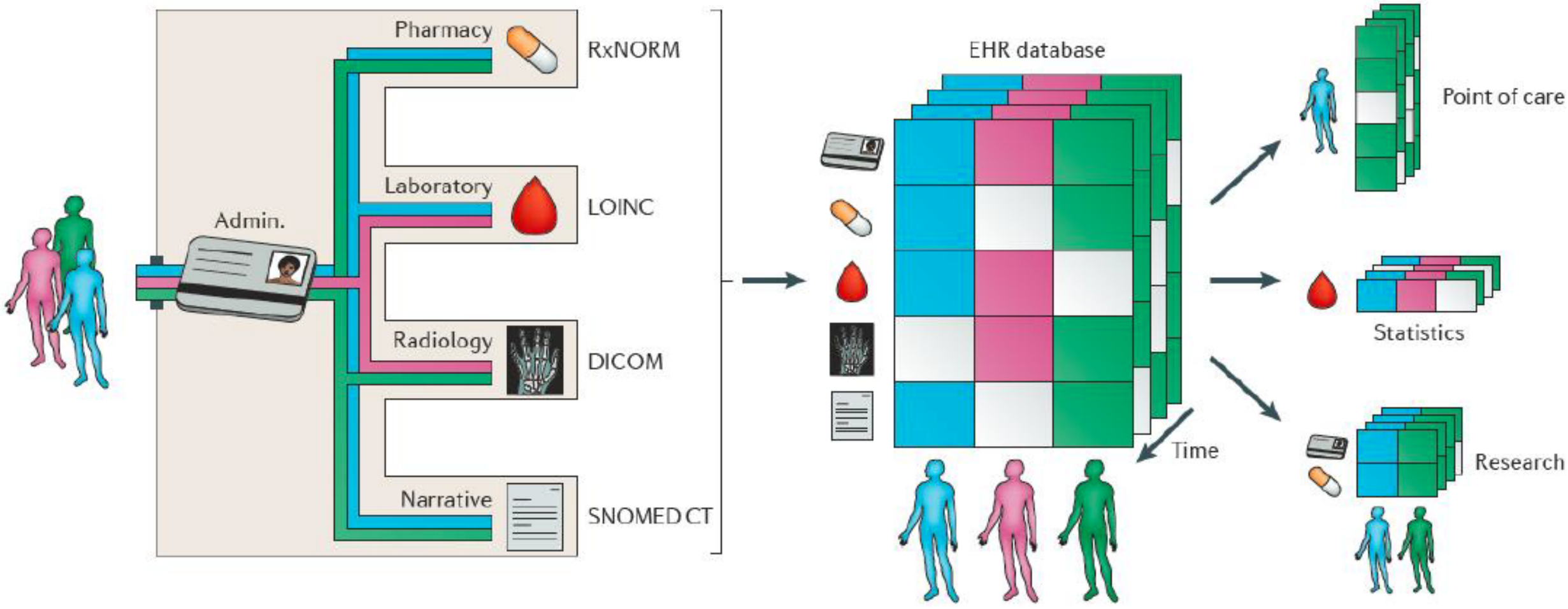


- Take advantage of the massive amounts of data and provide right intervention to the right patient at the right time.
- Personalized care to the patient.
- Potentially benefit all the components of a healthcare system i.e., provider, payer, patient, and management.

Motivating Example: Penalties for Poor Care

- Hospitalizations account for more than 30% of the 2 trillion annual cost of healthcare in the United States. Around 20% of all hospital admissions occur within 30 days of a previous discharge.
  - not only expensive but are also potentially harmful, and most importantly, they are often preventable.
- Medicare penalizes hospitals that have high rates of readmissions among patients with heart failure, heart attack, and pneumonia.
- Identifying patients at risk of readmission can guide efficient resource utilization and can potentially save millions of healthcare dollars each
- Effectively making predictions from such complex hospitalization data will require the development of novel advanced analytical models.

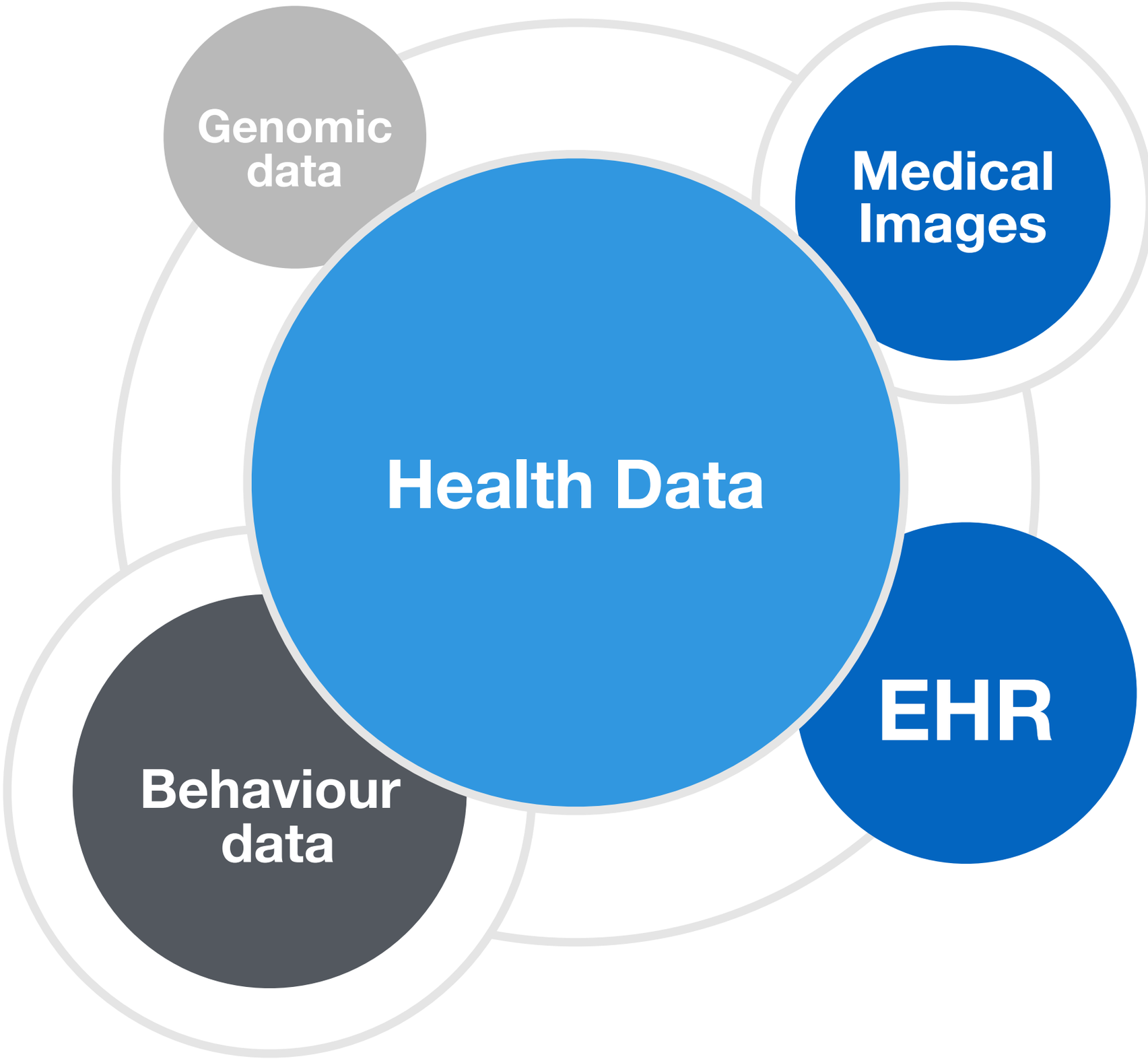
Data Collection and Analysis



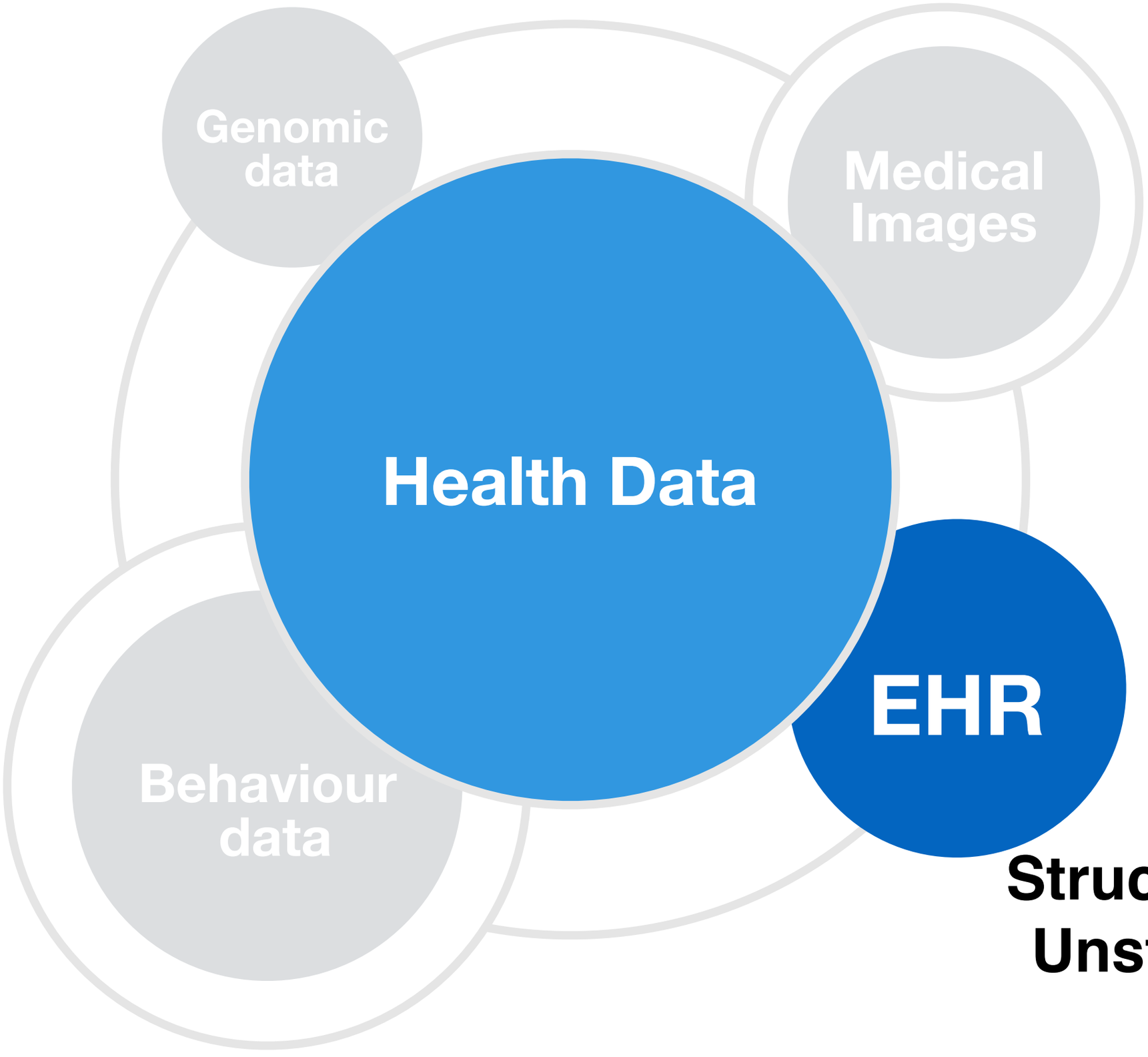
Effectively integrating and efficiently analysing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

P. Jensen et al., Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics (2012)

Sources and Techniques for Big Data in Healthcare



Electronic Health Records



Structured and Unstructured Data

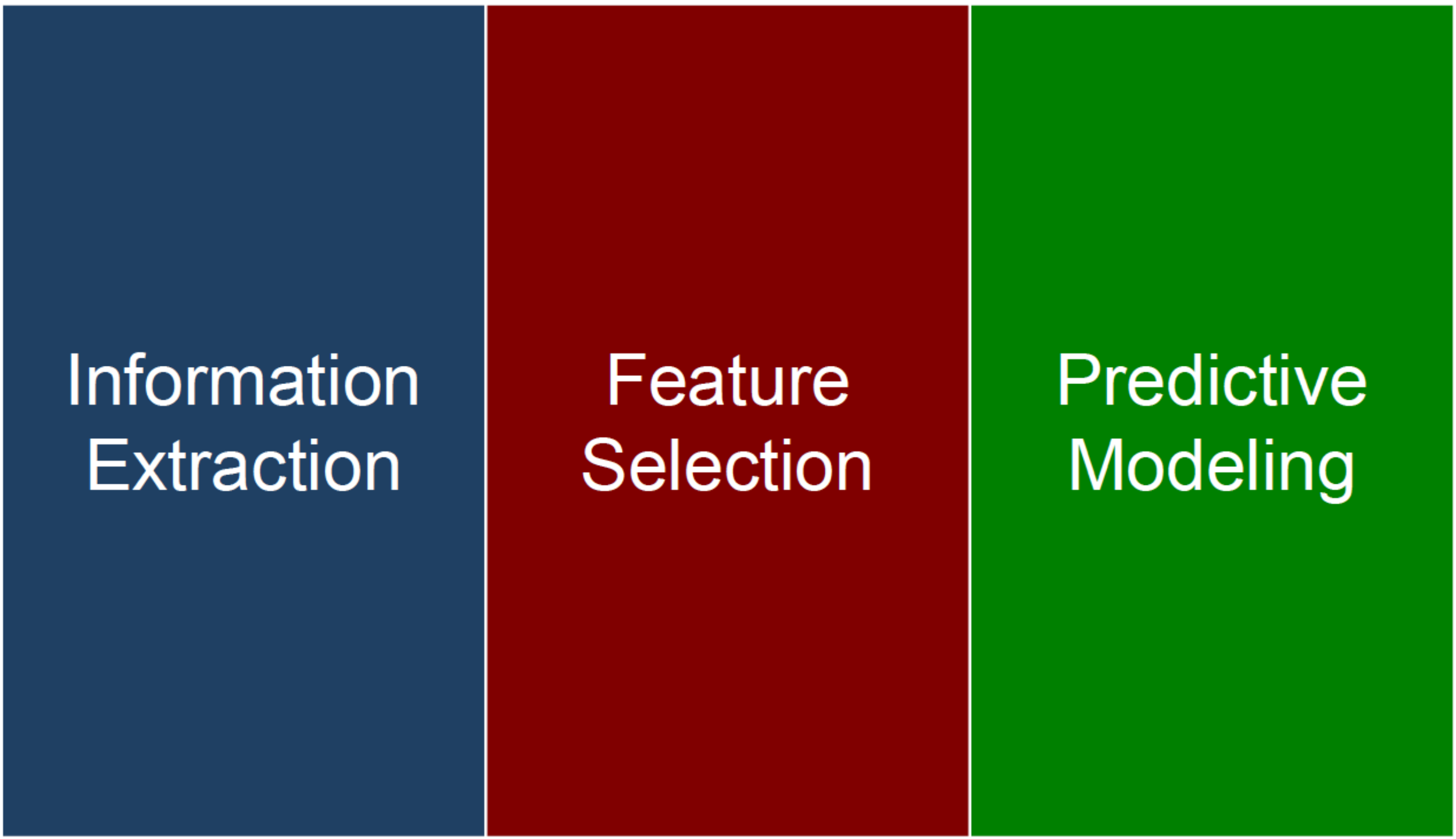
Electronic Health Records

	ICD	CPT	Lab	Medication	Clinical notes
Availability	High	High	High	Medium	Medium
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: Variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: Variable	Medium high
Format	Structured	Structured	Mostly structured	Structured and unstructured	Unstructured
Pros	Easy to work with, a good approximation of disease status	Easy to work with, high precision	High data validity	High data validity	More details about doctors' thoughts
Cons	Disease code often used for screening, therefore disease might not be there	Missing data	Data normalization and ranges	Prescribed not necessary taken	Difficult to process

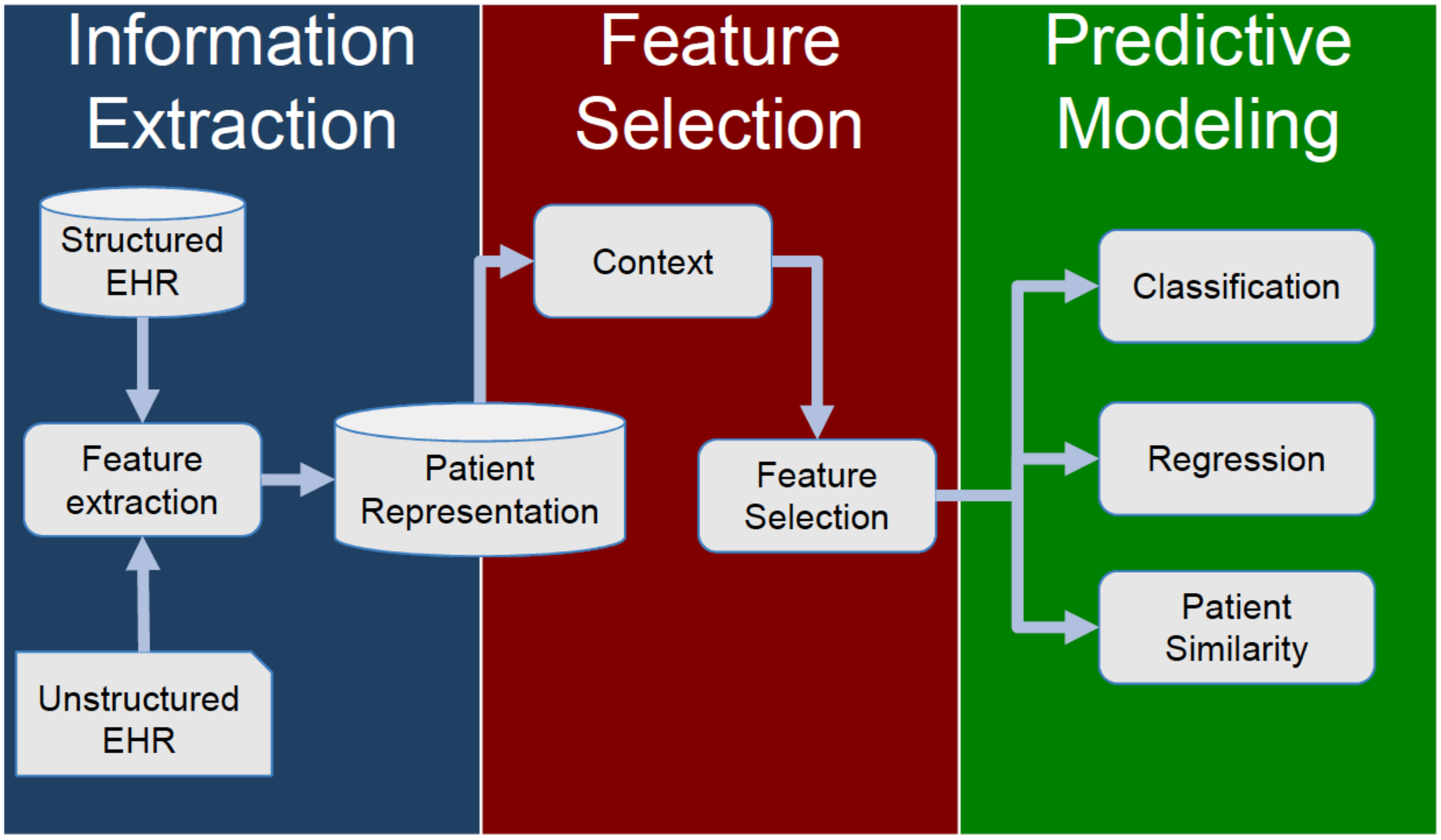
Electronic Health Records: Example of clinical notes

Subjective:	Objective:
ANXIETY STATE NOS 300.00 DEPRESSIVE DISORDER NEC 311 ATRIAL FIBRILLATION 427.31 OLD MYOCARDIAL INFARCT 412 CONGESTIVE HEART FAILURE 428.0 Current outpatient prescriptions ** LOPRESSOR 50 MG PO TABS 1 tab two times a day 60 5	250.00 DM, CONTROLLED, TYPE II (primary encounter diagnosis) 428.0 CONGESTIVE HEART FAILURE 585.3 KIDNEY DZ,CHRONIC (GFR>30-59) STAGE III 412 OLD MYOCARDIAL INFARCT 715.09 GENERAL OSTEOARTHROSIS 427.31 ATRIAL FIBRILLATION
Assessment:	**Plan:**
BP 122/68 | Pulse 78 | Temp (Src) 98.1 (Oral) | Resp 22 | Wt 227 lbs Abdomen: abdomen soft, non-tender, obese and no masses or organomegaly Back: No CVA tenderness Extremities: No edema	Continue present medication(s): Referral(s) to: eye Injection(s) ordered: b12 Schedule labs: Labs on return.

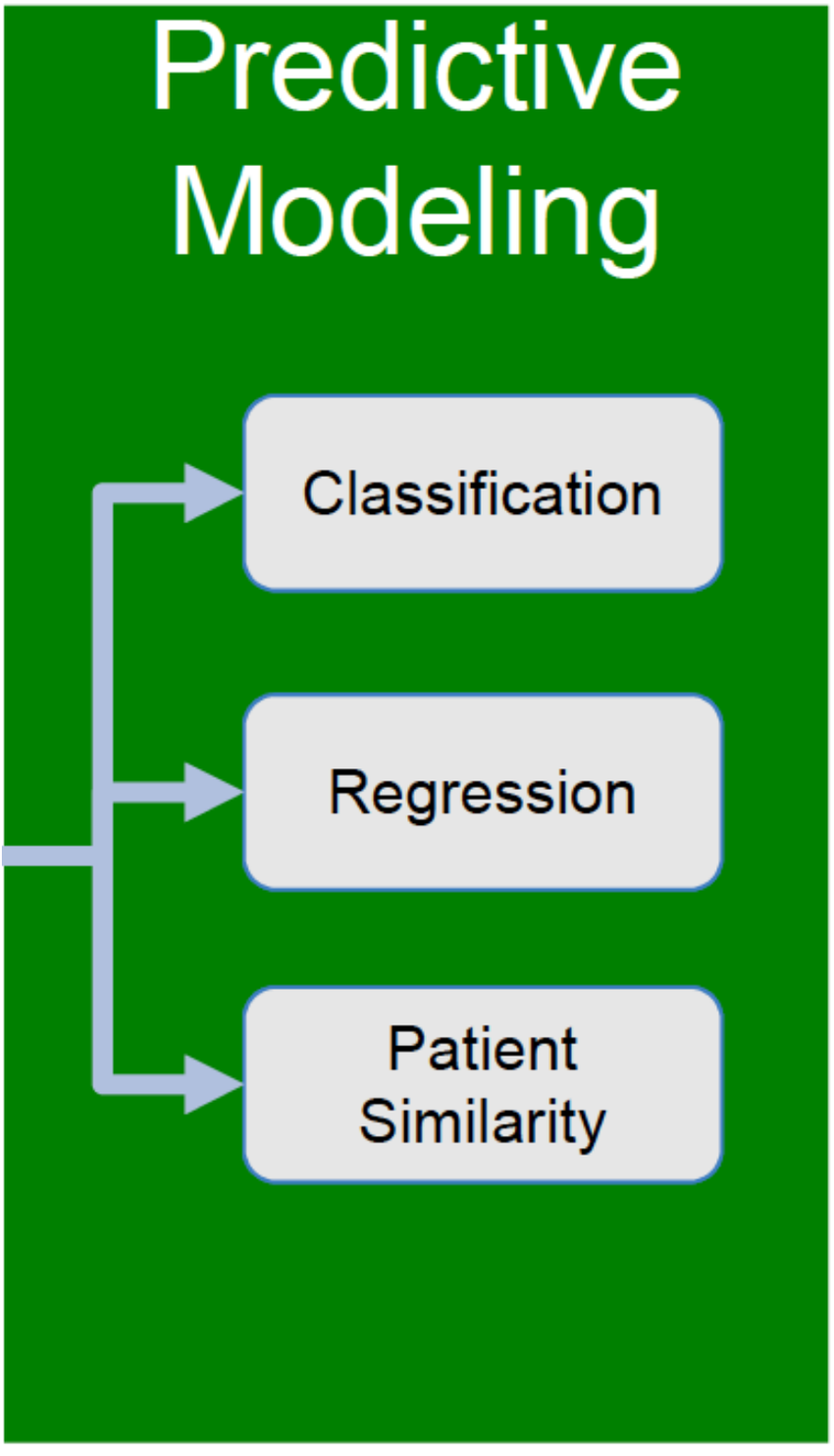
Analytic Platform

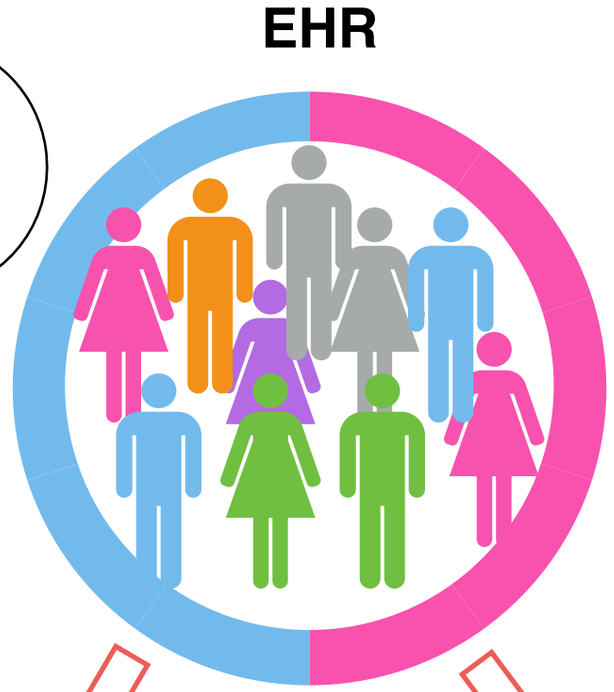


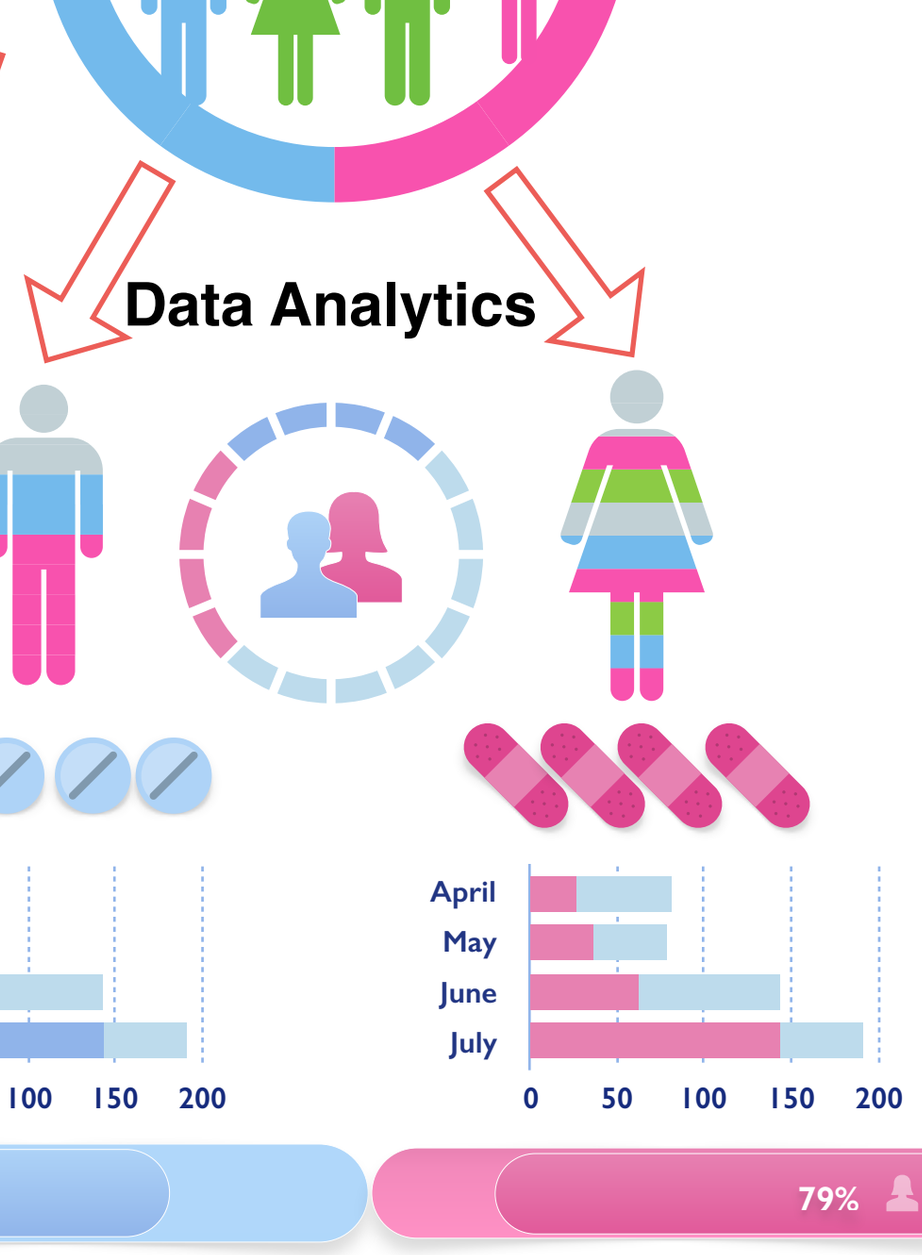
Analytic Platform



Analytic Platform







Patient Similarity

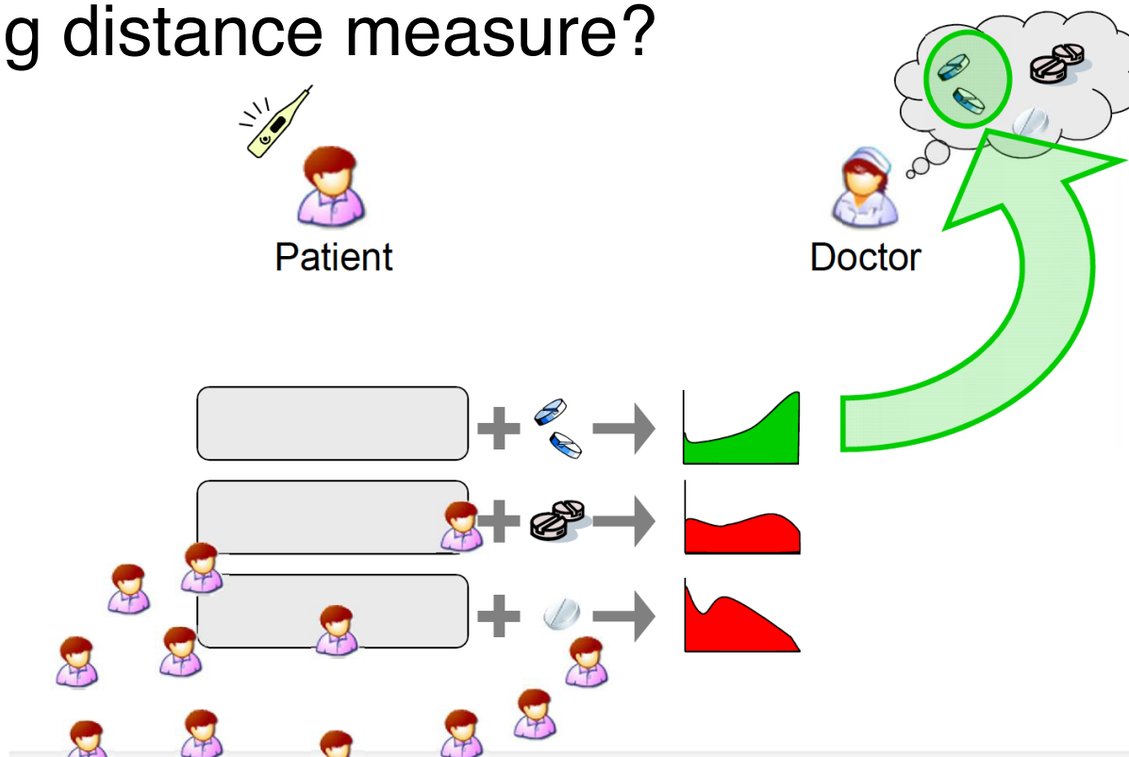
Patient similarity learns a customized distance metric for a specific clinical context.

Extension 1: Composite distance integration (Comdi)

– How to jointly learn a distance by multiple parties without data sharing?

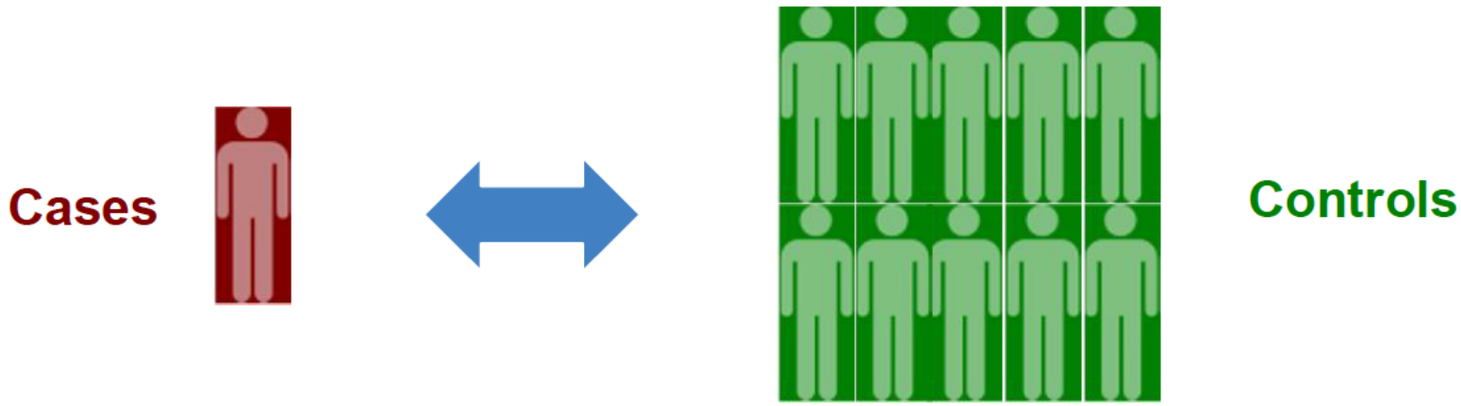
Extension 2: Interactive metric update (iMet)

– How to interactively update an existing distance measure?

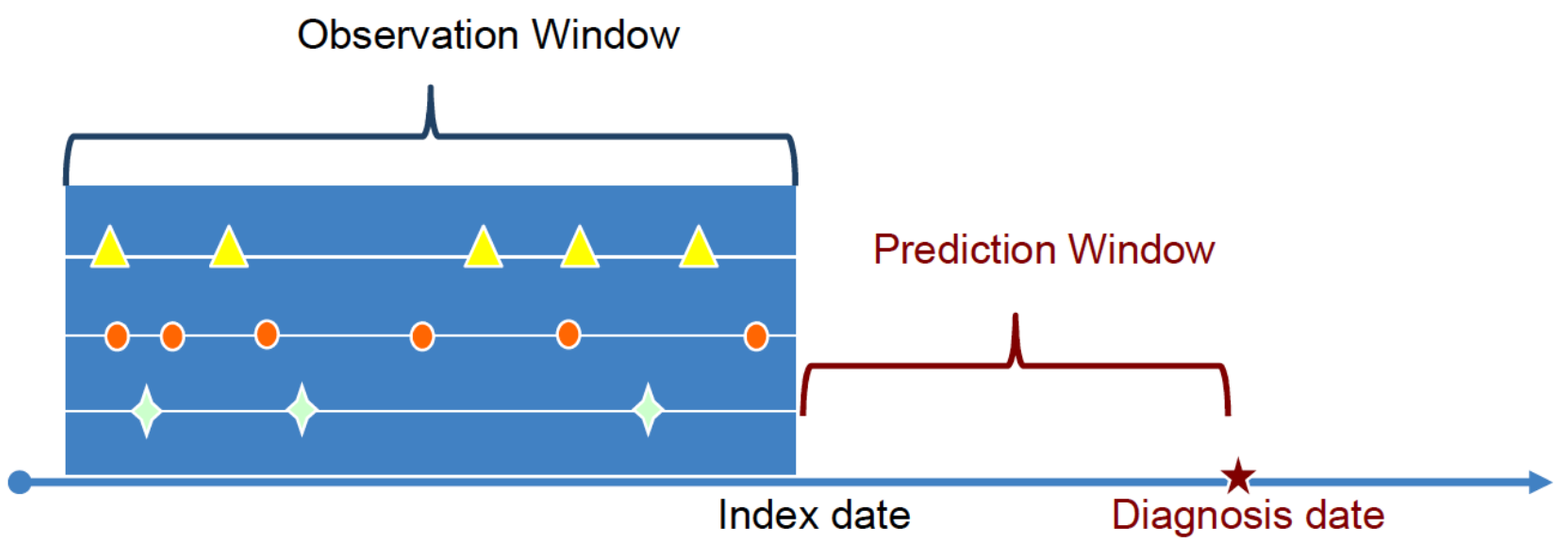


Predictive Models

Classify disease cases against control patients / Predict disease progression.



Use the training data (from observation window) to select features and predict onset after the prediction window.



Philosophy for Data Analytics

- Improve patient outcomes, cost effectiveness, and clinical workflows.
- Integration into the daily workflow (e.g., EHR).
- Support tools.
- Self-learning and improvement.
- Real-time and Retrospective Insights.
- Accuracy.

Resources and Software

1. **i2b2 Informatics for Integrating Biology & the Bedside:** Clinical notes used for clinical NL

   challenges

2. **Computational Medicine Centre:** Classifying Clinical Free Text Using Natural Language

   Processing

3……

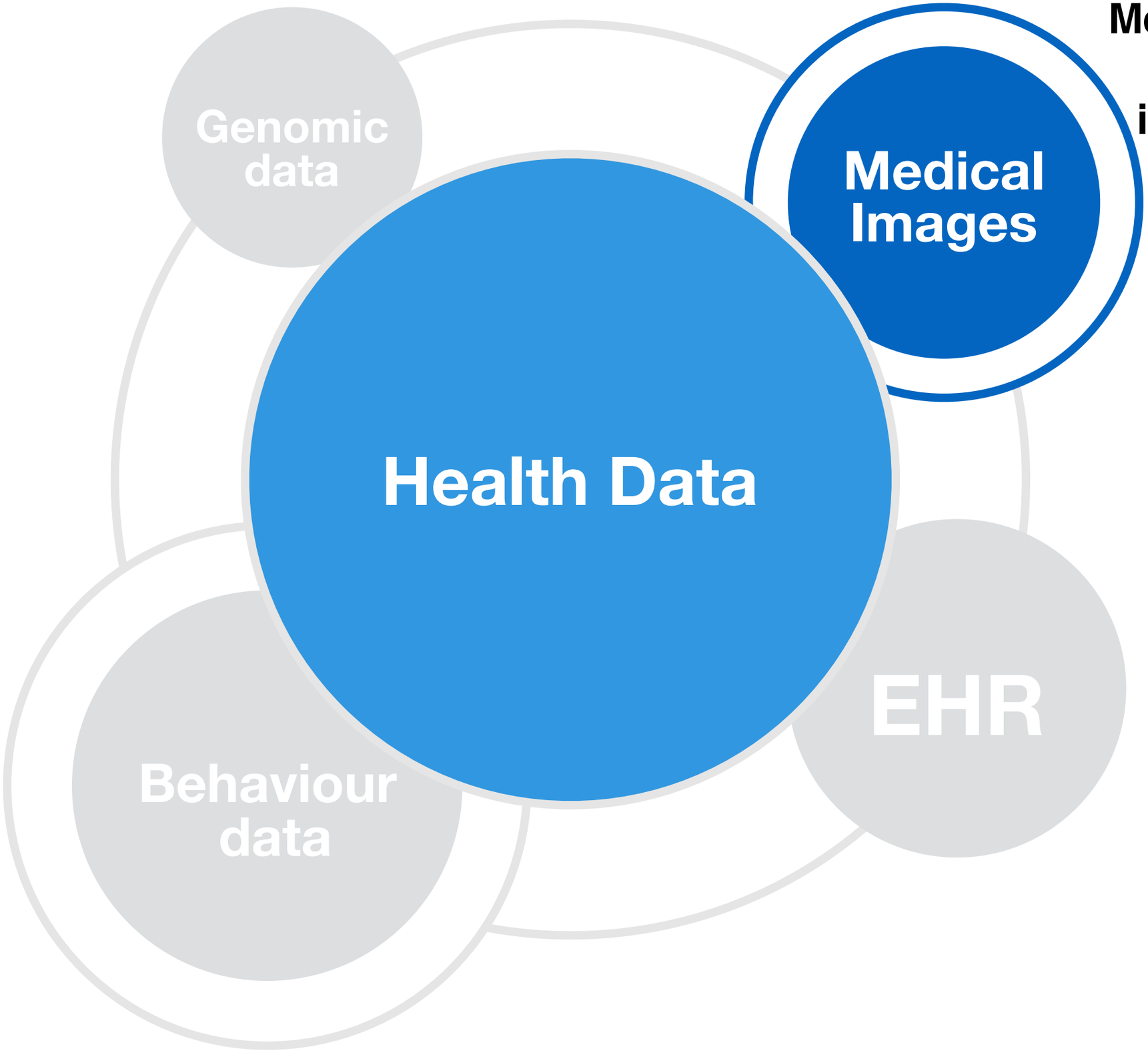
Software:

1. **MetaMap** maps biomedical text to UMLS metathesaurus. Developed for parsing

   medical article not clinical notes.

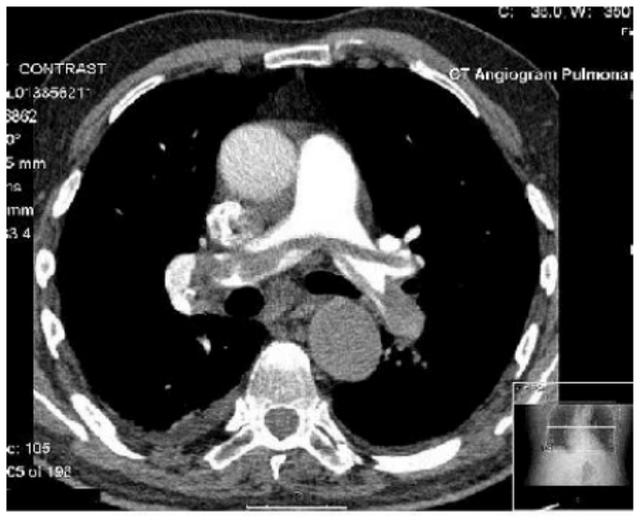
2.cTAKES allows for clinical text analysis and knowledge extraction system.

Medical Image Data

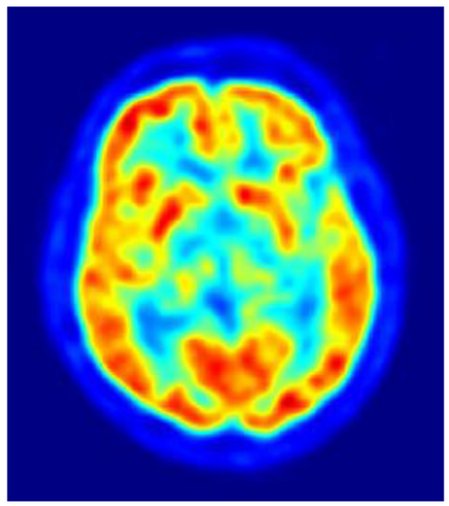


Medical Imaging archives are increasing by 20%-40%

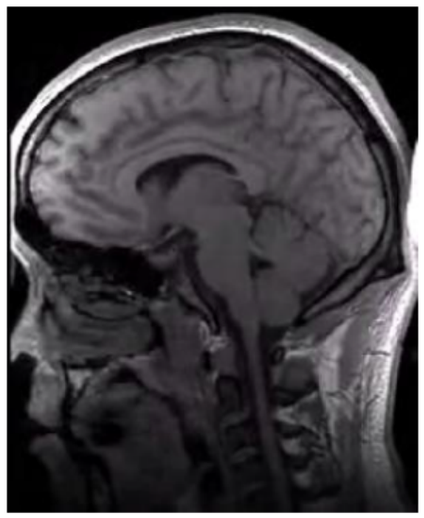
Medical Image Modalities and Challenges



Computed Tomography (CT)



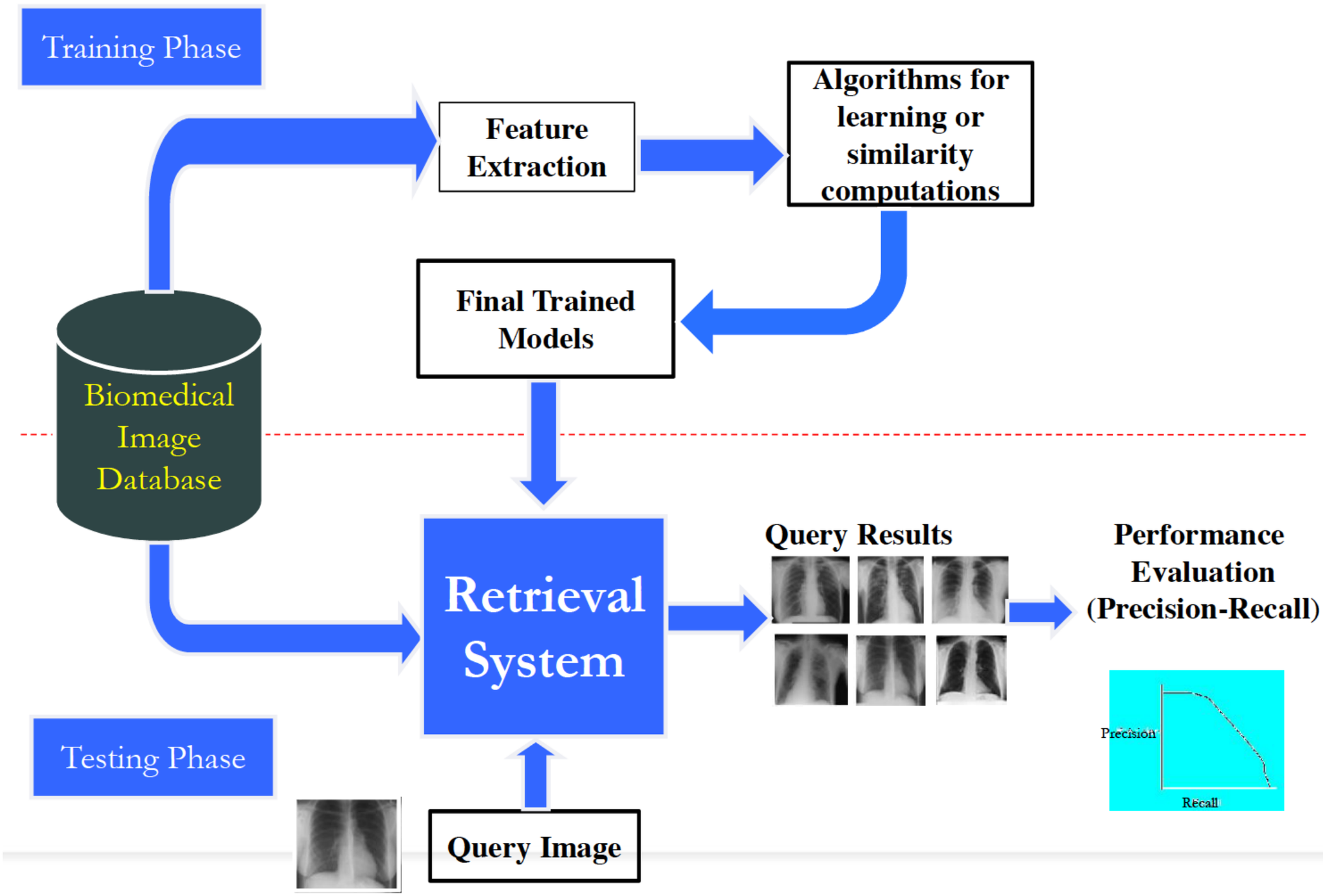
Positron Emission Tomography (PET)



Magnetic Resonance Imaging (MRI)

- The main challenge with the image data is that it is not only huge, but is also high-dimensional and complex.
- Extraction of the important and relevant features is a daunting task.
- Many research works applied image features to extract the most relevant images for a given query.

Medical Image Retrieval System



Challenges with Image Data

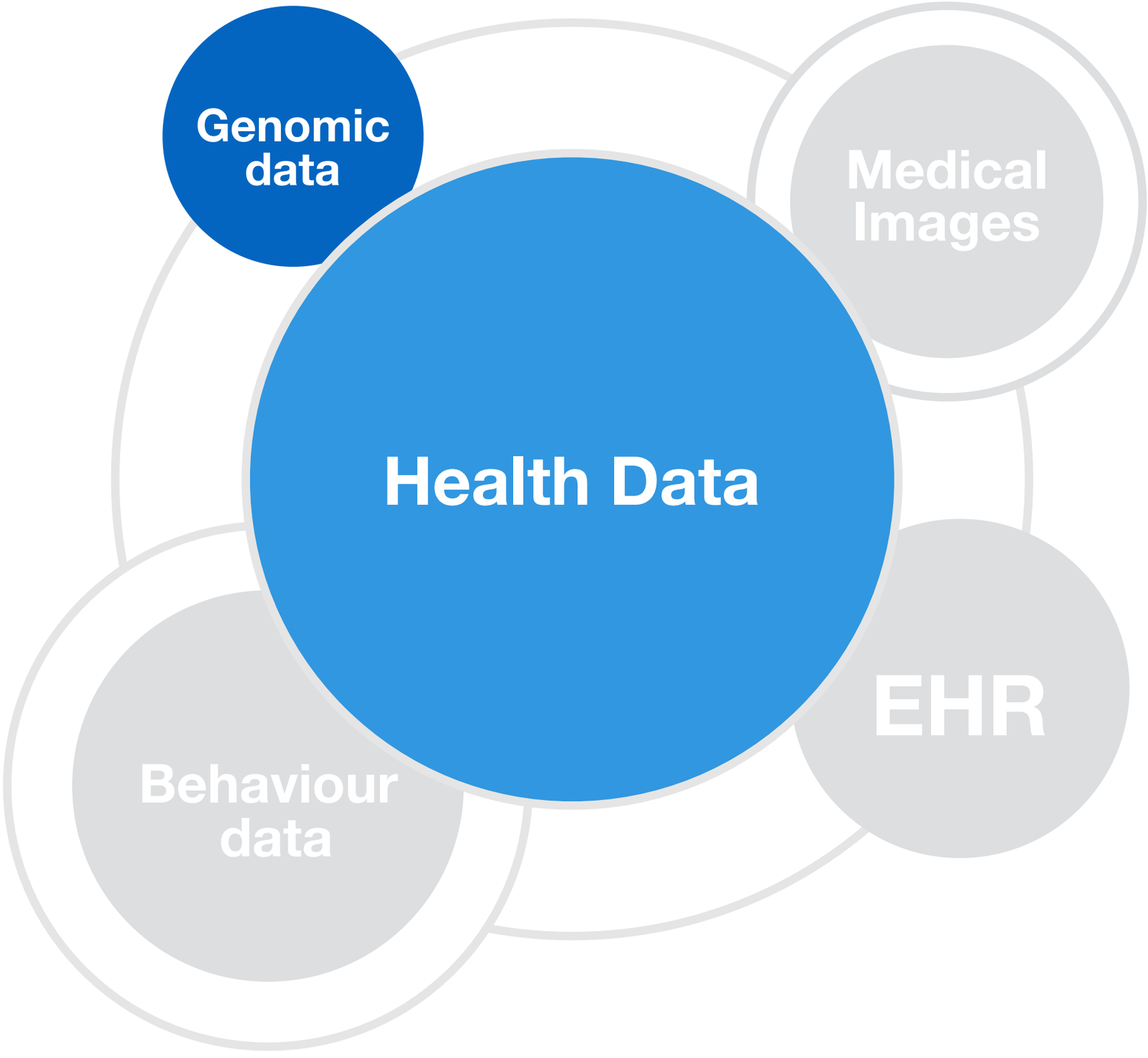
- **Extracting informative features.**
- **Selection of relevant features.**
  - Sparse methods and dimensionality reducing techniques
- **Integration of Image data with other data available**
  - Early Fusion
    - Vector-based Integration
  - Intermediate Fusion
    - Multiple Kernel Learning
  - Late Fusion
    - Ensembling results from individual modalities

Resources

1. **Cancer Imaging Archive Database (CT, DX, CR):** lesion detection and classification, accelerated diagnostic image decision, quantitative image assessment of drug response.
2. **Image CLEF Database (PET, CT, MRI, US):** modality classification, visual image annotation, scientific data management.

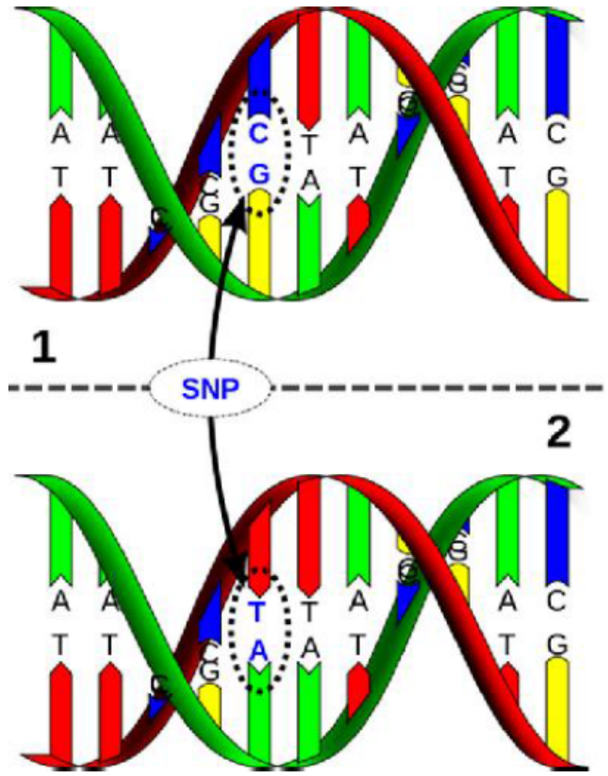
3.....

Genomic Data



Genomic Data: GWAS

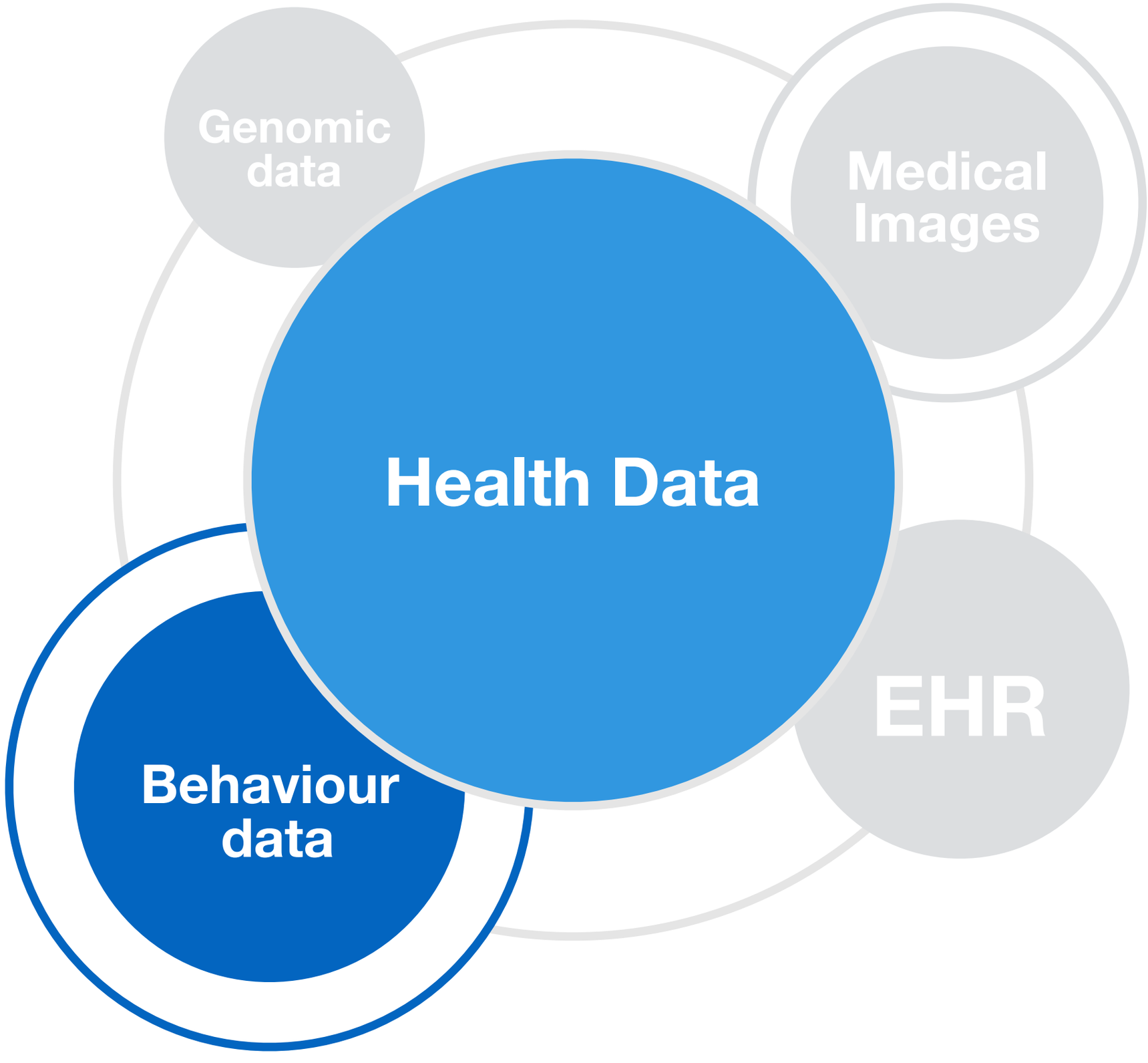
- Genome-wide association studies (GWAS) are used to identify **common genetic factors that influence health and disease.**
- These studies normally compare the DNA of two groups of participants: people with the disease (cases) and similar people without (controls). (One million Loci)
- Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence differs between individuals.
- SNPs occur every 100 to 300 bases along the 3-billion-base human genome.



Heavy computational burden!!!

- New sparse methods and software for SNP data analysis.
- Public Databases (e.g., WTCCC).

Behaviour Data

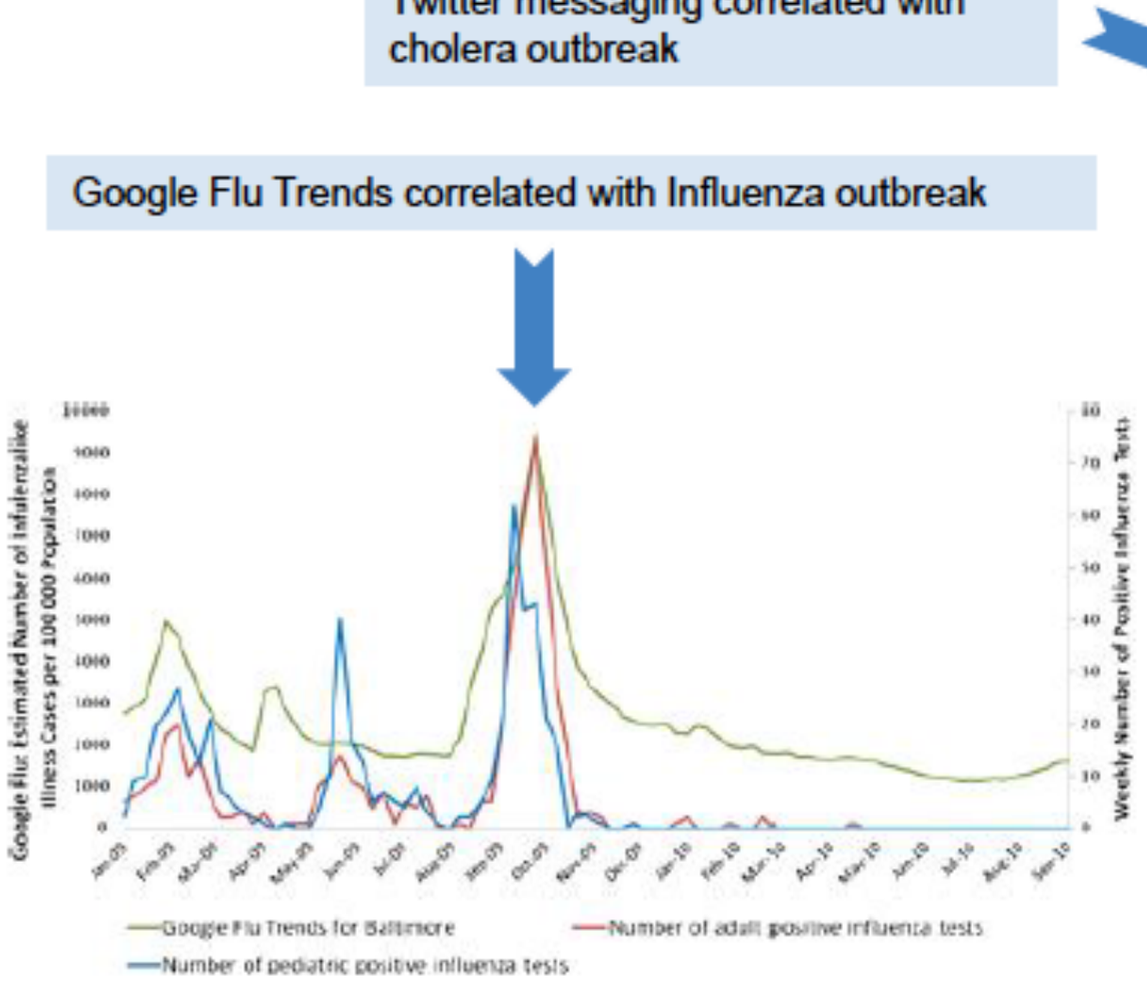


Social Media can Sense Public Health

During infectious disease outbreaks, data collected through health institutions and official reporting structures may not be available for weeks, hindering early epidemiologic assessment. Social media can get it in near real-time.

Twitter messaging correlated with cholera outbreak

Google Flu Trends correlated with Influenza outbreak



Dugas, Andrea Freyer, et al. "Google Flu Trends: correlation with emergency department influenza rates and crowding metrics." *Clinical infectious diseases* 54.4 (2012): 463-469.

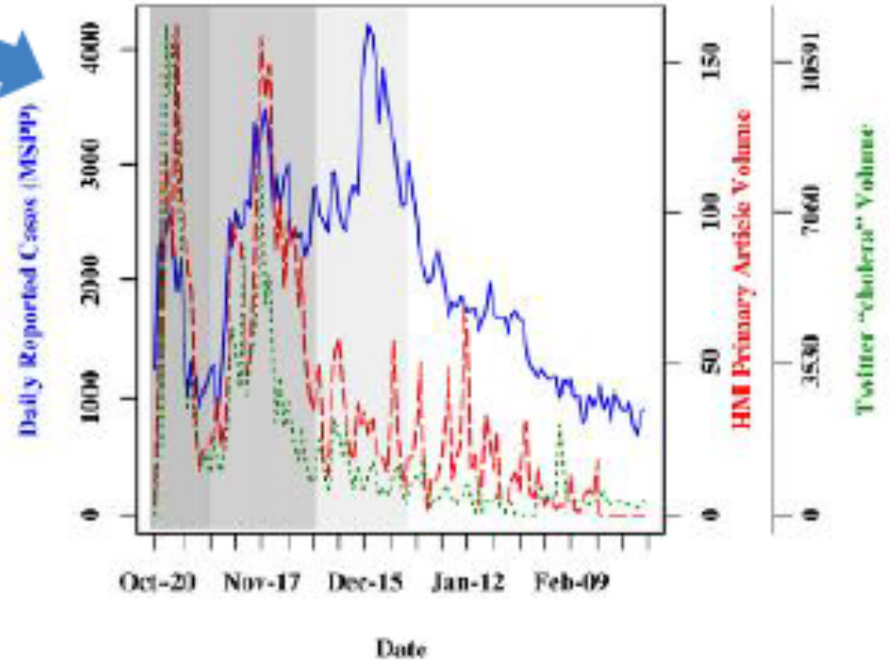
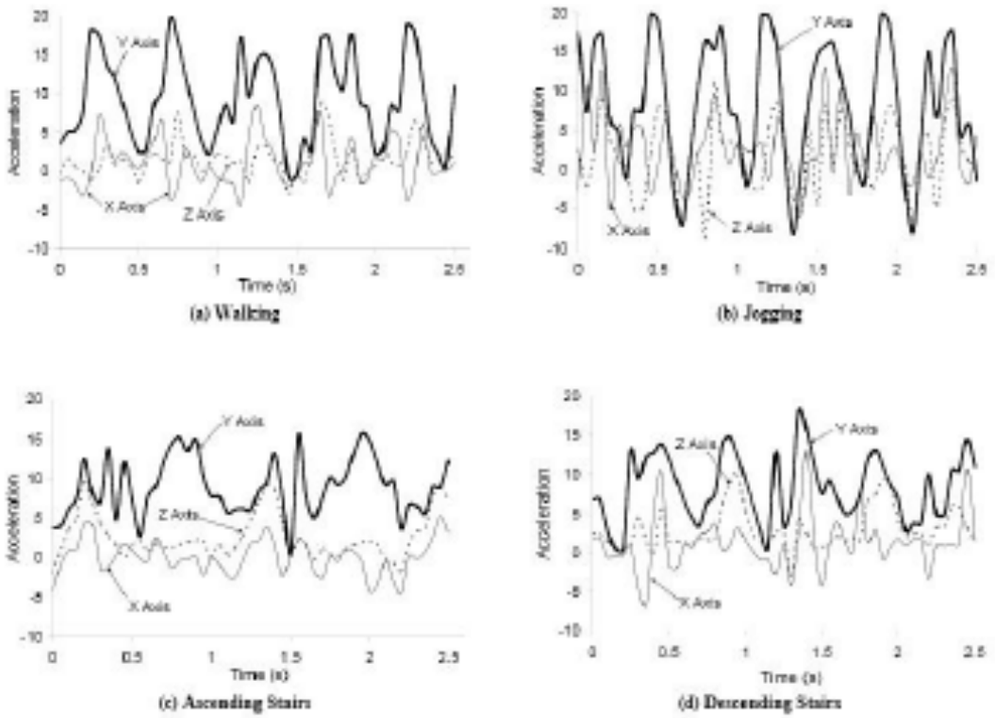


FIGURE 2. Daily reported case data for all departments from the Haiti Ministry of Health (solid), daily volume of primary HealthMap alerts (dashed), and daily volume of Twitter posts containing the word "cholera" or "#cholera" (dotted). Each curve has an initial peak at the onset of the outbreak (dark grey), and a peak during the time that Hurricane Tomas affected Haiti (medium grey). The first 100 days of the outbreak are shaded in light grey. Ministère de la Santé Publique et de la Population (MSPP) case counts peak again in late December, although HealthMap and Twitter volume only have daily variations during this time.

Chunara, Rumi, Jason R. Andrews, and John S. Brownstein. "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak." *American Journal of Tropical Medicine and Hygiene* 86.1 (2012): 39.

Home Monitoring and Sensing Technologies

- Advancements in sensing technology are critical for developing effective and efficient home-monitoring systems
- Sensing devices can provide several types of data in real-time.
- Activity Recognition using Cell Phone Accelerometers



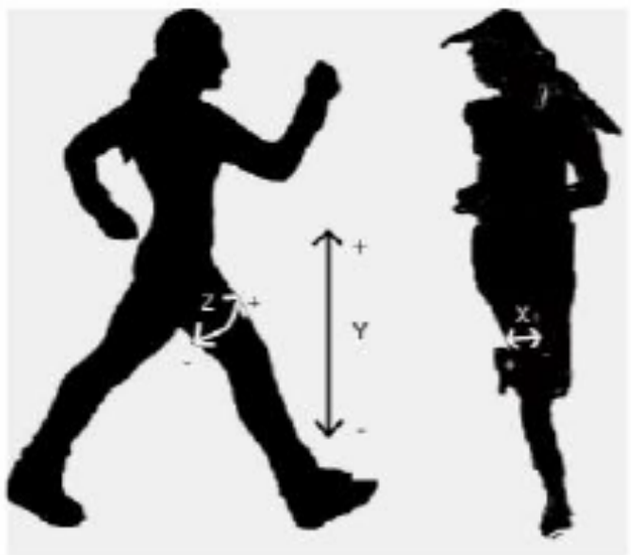


Figure 1: Axes of Motion Relative to User

Kwapisz, Jennifer R., Gary M. Weiss, and Samuel A. Moore. "Activity recognition using cell phone accelerometers." *ACM SIGKDD Explorations Newsletter* 12.2 (2011): 74-82.

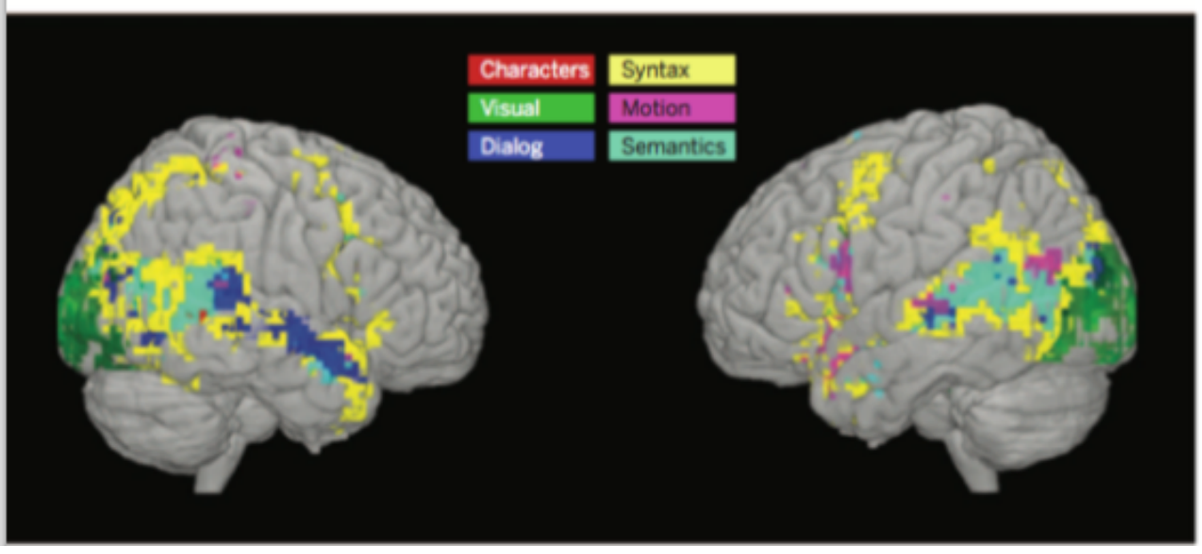
Rashidi, Parisa, et al. "Discovering activities to recognize and track in a smart environment." *Knowledge and Data Engineering, IEEE Transactions on* 23.4 (2011): 527-539.

Big Data Analysis - How?

→ machine / statistical learning

A breakthrough in machine learning would be worth ten Microsofts.

Bill Gates, Microsoft



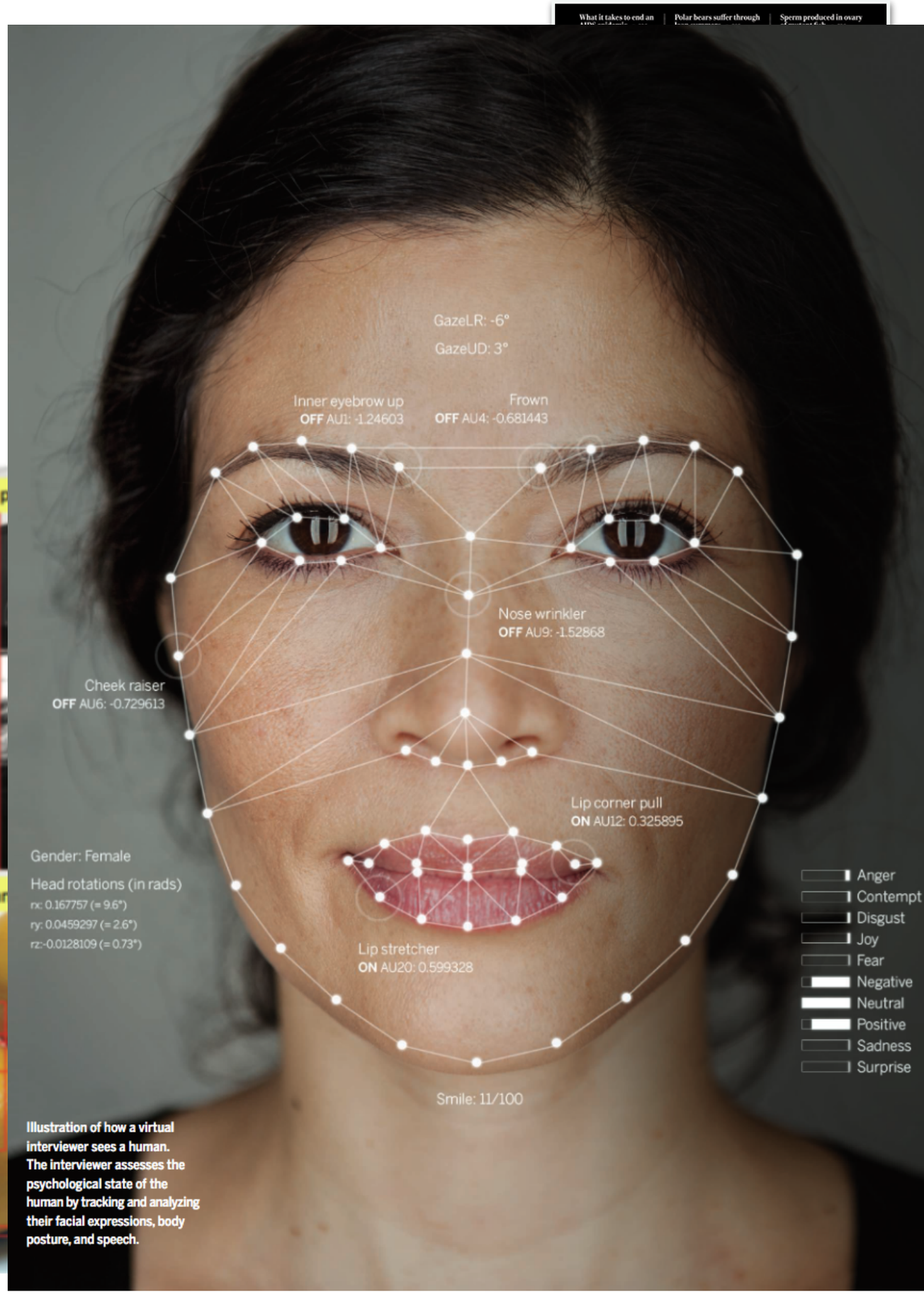
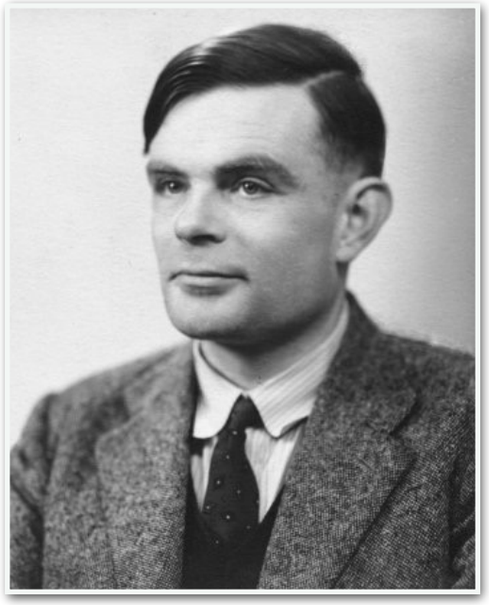
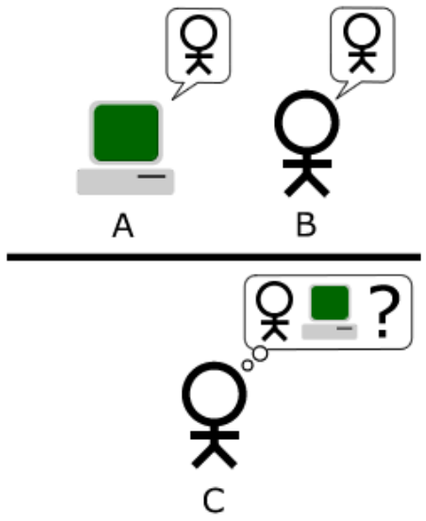


Illustration of how a virtual interviewer sees a human. The interviewer assesses the psychological state of the human by tracking and analyzing their facial expressions, body posture, and speech.

Science vol 349, 2015

The Quest for AI. Birth of Dream





Intelligence: a Working Definition

- *abstract reasoning, knowledge acquisition, decision making.*
- *knowledge acquisition: memorisation vs learning.*

Alan Turing (1912–1954) proposed a test that calls for a panel of judges to review typed answers to any question that has been addressed to both a computer and a human. If the judges can make no distinctions between the two answers, the machine may be considered **intelligent.**

Ingredients of AI

- natural language processing;
- knowledge representation;
- automated reasoning;
- machine learning;
- computer vision;
- robotics to manipulate.

The Quest for AI. Birth of Dream

1950
Turing Test to measure machine's intelligence

1950
Computer chess by C.Shannon

1956
AI founded as an academic discipline

1955
Dartmouth College Summer AI Conference
(by C. Shannon, J. McCarthy, M. Minsky, N. Rochester)

1948
'If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that!'
J. Neumann

1955
"Logic Theorist"
for math proofs
(by A. Newell and H. Simon)

New paradigm
"reasoning as search"

Computers communicate in natural languages

The Optimism by AI researchers

Paradigm
"Micro-worlds"

First problems:
- limited computer power;
- intractability;
- commonsense knowledge.

The revival of connectionism

$850 million
Fifth generation computer (Japan)

Nouvelle approach to AI, based on robotics

'Rise of expert systems'
1980
XCON saved
$40 million annually

AI became a more rigorous scientific discipline

Intelligent Agents

1997
Deep Blue chess machine (IBM) beats Garry Kasparov

2005
DARPA Grand Challenge

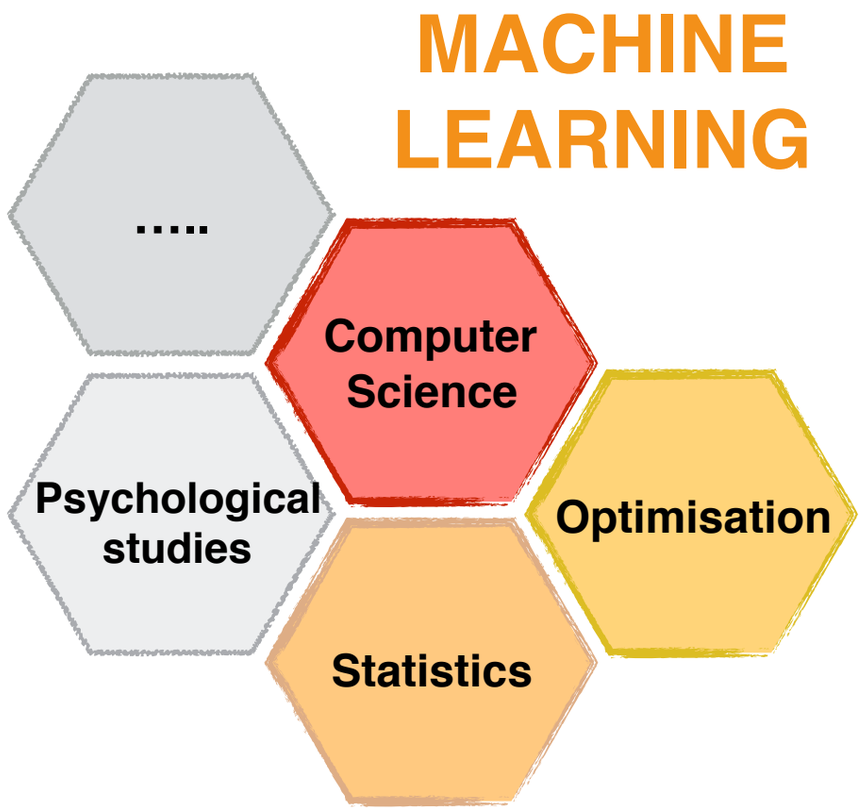
2009
Google self-driving car

2011
Apple's Siri, Google's Google Now…

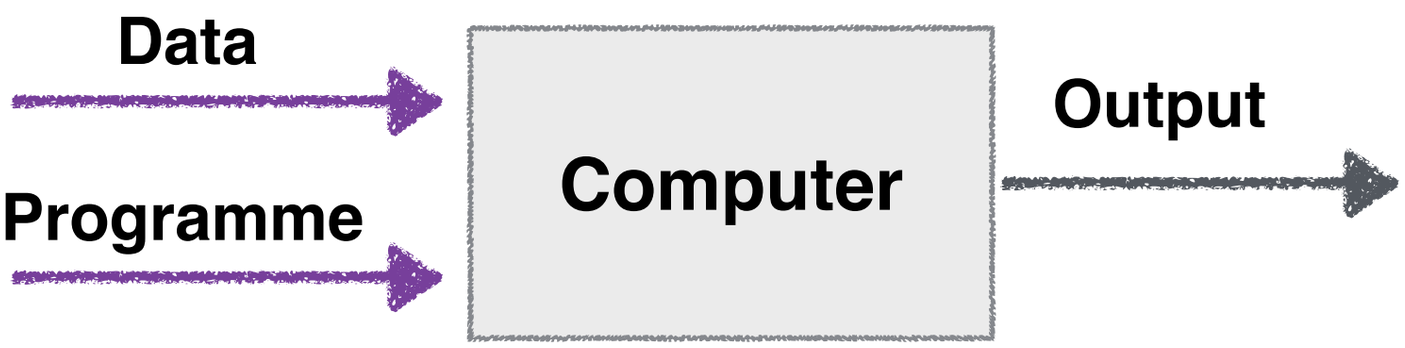
~1950 | **1952-1956** The birth of AI | **1956-1974** The golden years | **1974-1980** First AI Winter | **1980-1987** Boom | **1987-1993** Second AI Winter | **>1993**

So What is Machine Learning?

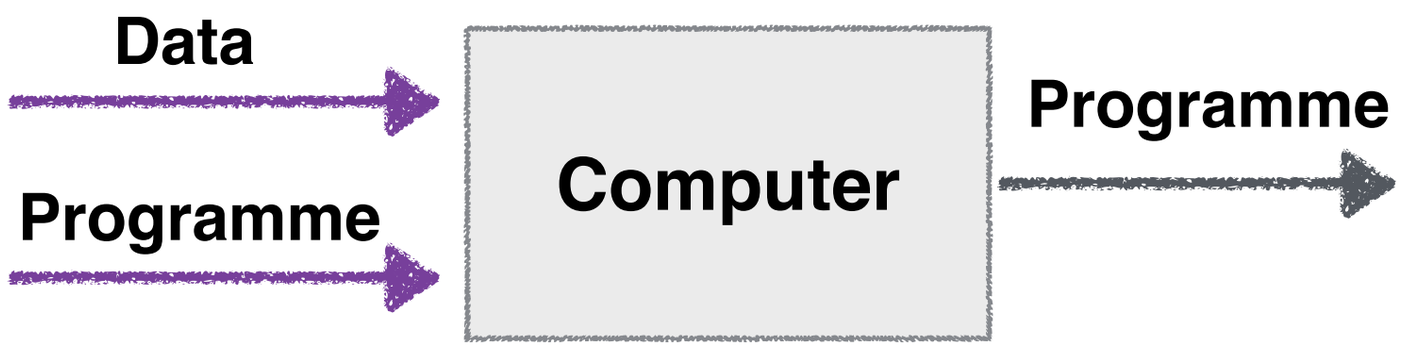
- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!



Traditional Programming



Machine Learning



Is Machine Learning Magic?

No, it is more like gardening…

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

> Machine learning is not a magic; it can't get something from nothing. What it does it gets more from less.
>
> *Pedro Domingos, UC Washington*

Sample Applications: Prediction

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

Sample Applications: Prediction

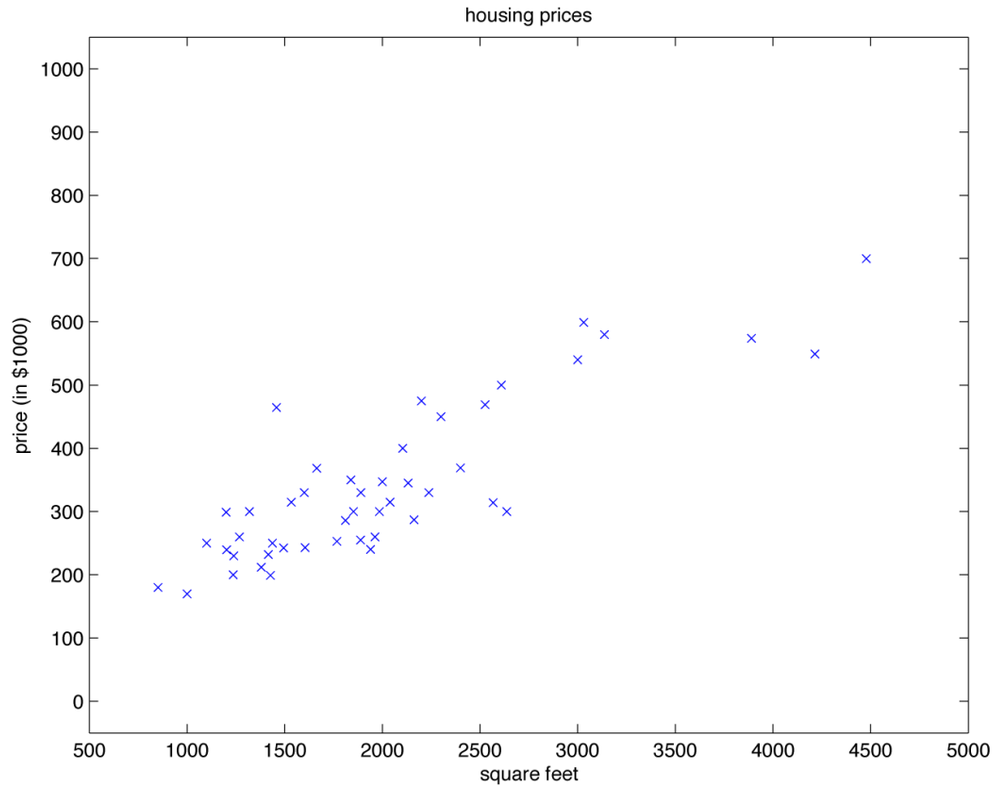
Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
$\vdots$	$\vdots$

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

Sample Applications: Prediction

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

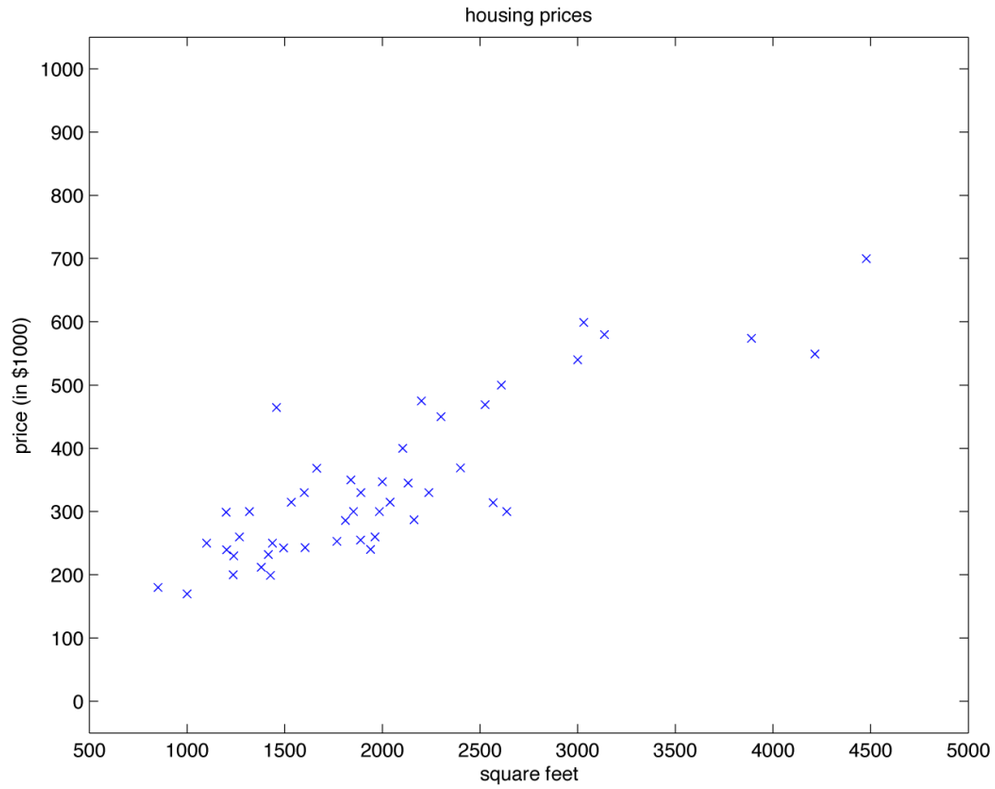


Sample Applications: Prediction

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮



$$y_i = f(x_i) + \sigma\varepsilon_i, \;\; \sigma > 0$$

$$\text{e.g.} \quad f(x) = w^T x, \;\; \varepsilon_i \sim N(0,1)$$

Sample Applications: Spam Filtering

True loneliness is when you don't even receive spam

Ham

[simula]

Mail — More — 1–50 of 2,489

COMPOSE

Inbox (2)
Starred
Important
Sent Mail
Drafts (233)
Administration/Apa…
Administration/Boo…
Administration/Har…
Administration/Lan…
Administration/VN
Administration SIm…
Administration SIm…
Valeriya

Kristin McLeod	Hiking/cabin weekend August 20-21 - Hey gang Hermenegild, Johannes, and I were thinking of going for a weekend hiking/c	10:21
Airbnb	Reservation reminder - August 13, 2016 - Airbnb Pack your bags! It's almost time for your trip to \ Modify reservation	10:18
H2020 - IKT	**Don't miss this autumn's interesting events - H2020 ICT 10/2016** - H2020 ICT - National Contact Points - NCP Hvis du ik	**5 Aug**
Andy Edwards	Fwd: Your reservation: Aug 7 through 12 - VRBO.com #857936 - Google maps, what a tremendous piece of software. Forwa	4 Aug
Sigurd, me (2)	Report: Wednesday and Thursday (08.03 and 08.04) - Hi Sigurd, Thanks for the update and your hard work (so late!)! It look	4 Aug
Viviane, me (2)	Awesome Summer School Pool Party! - Hi Viviane, Thanks a lot for the invitation, which I gladly accept :) Looking forward to	4 Aug
Massimo, me (6)	Funny. - I fully agree :) I think I need at least one year to decide for myself, finish current work, see how it works with the grou	4 Aug
me, Kereta (6)	Projects/FRIPRO/VN Reimbursement - Dear Zeljko, Great, many thanks for the response and your wishes! I am very happy	4 Aug
Inger, me (2)	EU news item - Thanks Karoline for your kind email and changing the news! Best Valeriya On 04 Aug 2016, at 15:22, Inger H	4 Aug
Jan, me, Julio (7)	**[Reminder] CanPathPro - WP5 Skype meeting** - Dear all, this is a brief reminder regarding our regular Skype meeting. As	**4 Aug**
Lars, me (2)	Crash Course on Machine Learning - Dear Lars, Yes for sure! You are more than welcome to join. Best Valeriya Dr. Valeriya	4 Aug

Spam

[simula] in:spam

Mail — More — 1–12 of 12

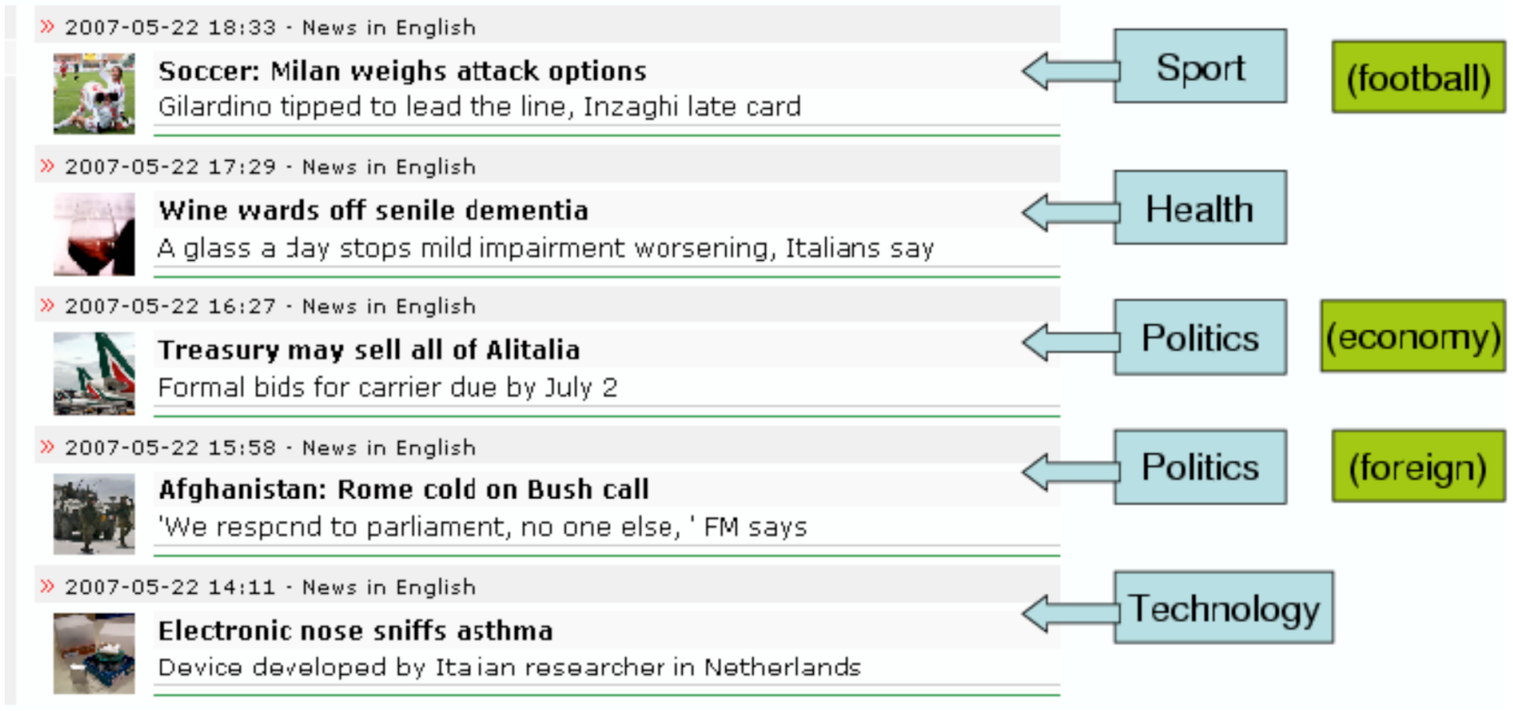
Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

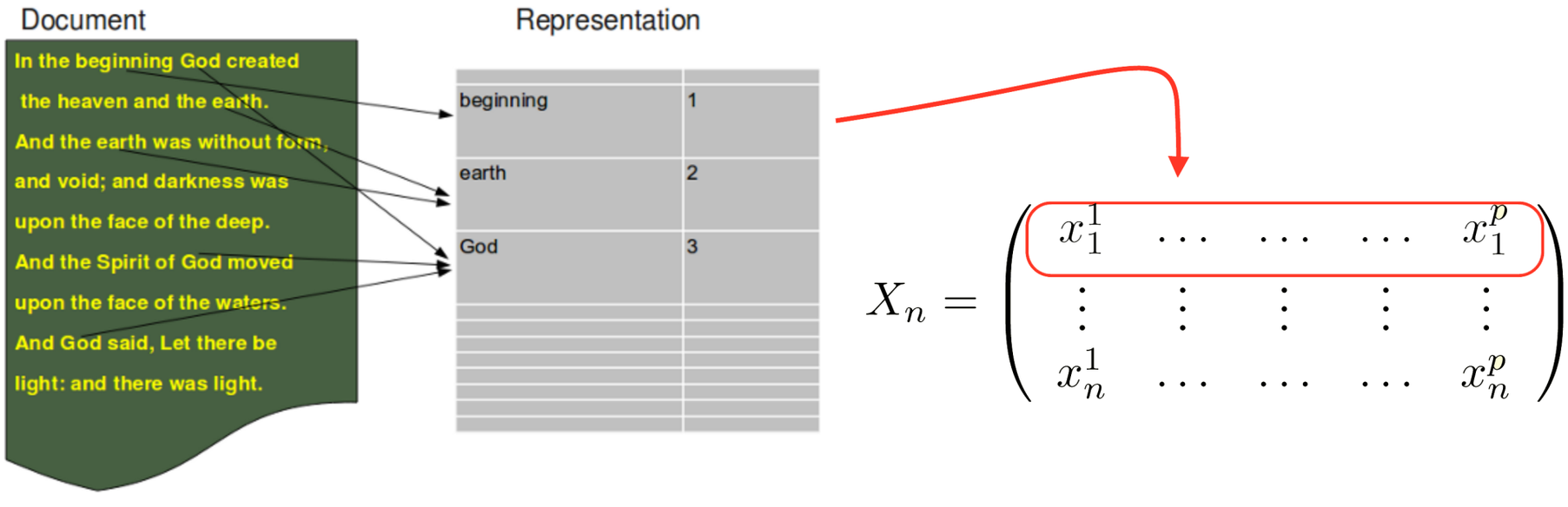
COMPOSE

Inbox
Starred
Important
Sent Mail
Drafts (233)
Administration/Apa…
Administration/Boo…
Administration/Har…
Administration/Lan…
Administration/VN
Administration SIm…
Administration SIm…
Valeriya

Markus Stahl	**Mehr als 448 Euro pro Tag verdienen? HEUTE noch möglich** - Sehr geehrter Interessent, wir möchten Sie mit dieser Em	**03:39**
Laura Schubert (4)	**RE: Kundendienst schuldet dir Geld** - Hallo valeriya@simula.no! Das Customer Service Team sendete mir deine Details ü	**5 Aug**
Dr. Matthias Reich	**Der unerwartete Geldsegen** - Guten Tag, Unverhofft kommt oft? Naja, in diesem Fall stimmt es zumindest. Klicken Sie hier	**5 Aug**
Google Trading Inc. (5)	**TAG 1 - Wettbewerb einladen for valeriya@simula.no!** - Hallo valeriya@simula.no, TAG 1 competition.jpeg Ich habe eine	**5 Aug**
Linnea	**Symposium 2016 - Invitation Letter from China** - Symposium 2016 - Invitation Letter from China Dear Professor/Researcl	**5 Aug**
Mr.Brian William.	**Attention.** - BG Group PLC Thames Valley Park, Reading, RG6 1PT, Berkshire, United Kingdom. Attention. I am Mr. Brian V	**5 Aug**
Sven Lindholm	**Recently posted academic job vacancies at Educaloxy** - Dear Colleague, We are pleased to present you our specialised	**5 Aug**
Google Benachrichtigung	**Wettbewerb einladen for valeriya@simula.no!** - IEMB16051 Hallo valeriya@simula.no, Ich habe eine Woche gewartet oh	**4 Aug**
DNB Bank ASA	**Varsling!** - Kjære kunde, Du har mottatt en intern melding. Klikk her Vennlig hilsen DNB ASA	4 Aug
giuseppelunardi	RE: valeriya	2 Aug

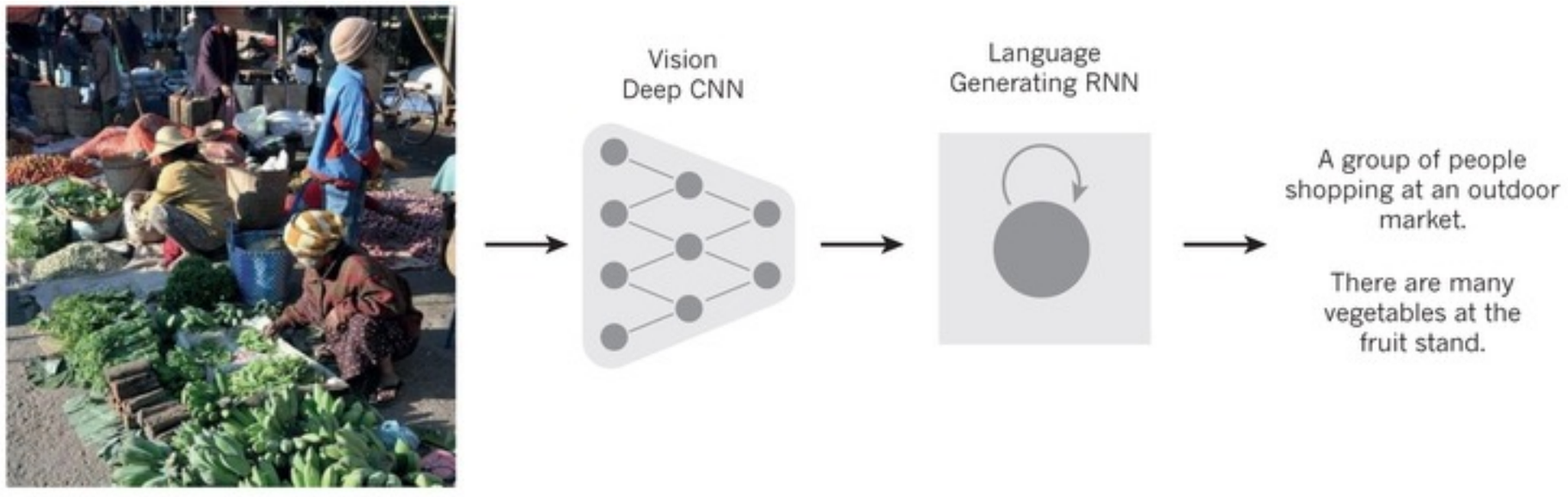
Sample Applications: Text Classification

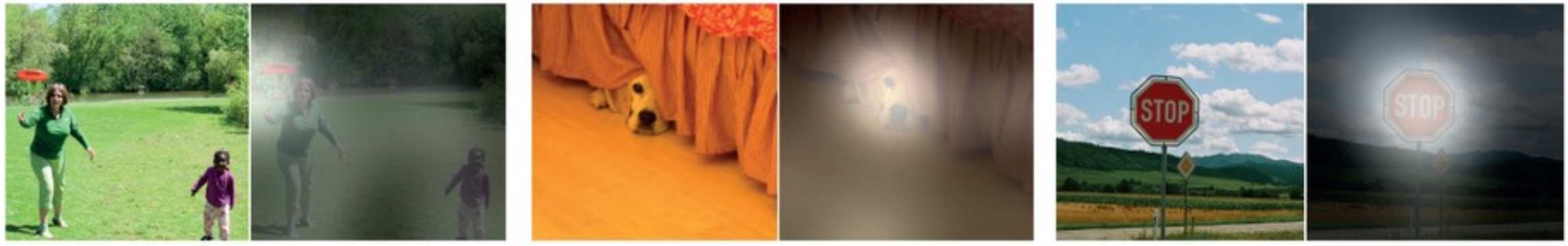




Sample Applications: Object Recognition

Sample Applications: from Image to Text





A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background

A little **girl** sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

A giraffe standing in a forest with **trees** in the background.

Nature vol 521, 2015

Sample Applications

Web ranking

Social Networks

Computational Biology

Debugging

Finance

E-Commerce

Mobile technology

Robotics

Space exploration

Navigation

…[Your favourite area]

Tens of thousands of machine learning algorithms

Hundreds new every year

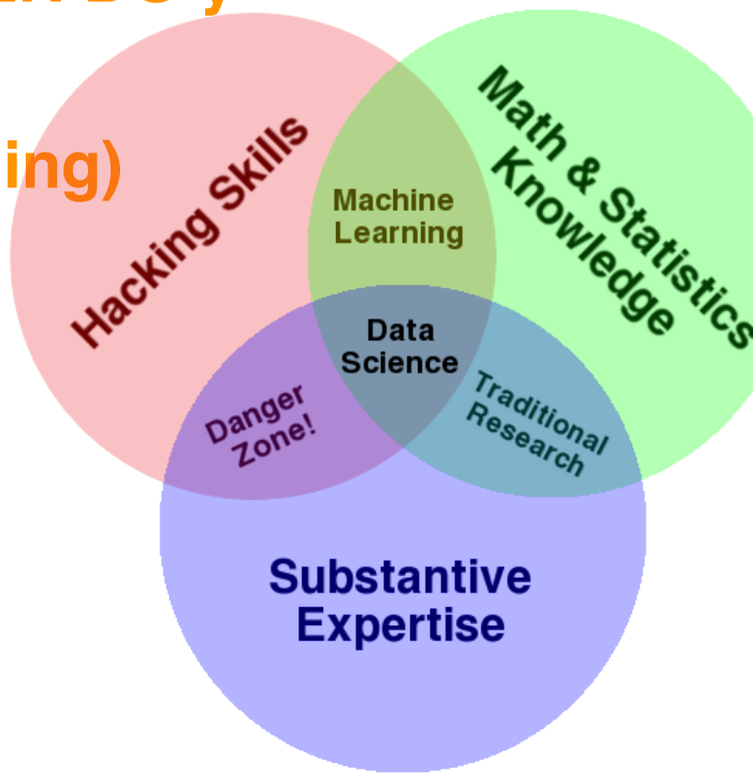
Programming with Data / Data Analysis

- Want adaptive robust and fault tolerant systems
- Rule-based implementation is (often)
  - difficult (for the programmer)
  - brittle (can miss many edge-cases)
  - becomes a nightmare to maintain explicitly
  - often doesn't work too well (e.g. OCR)

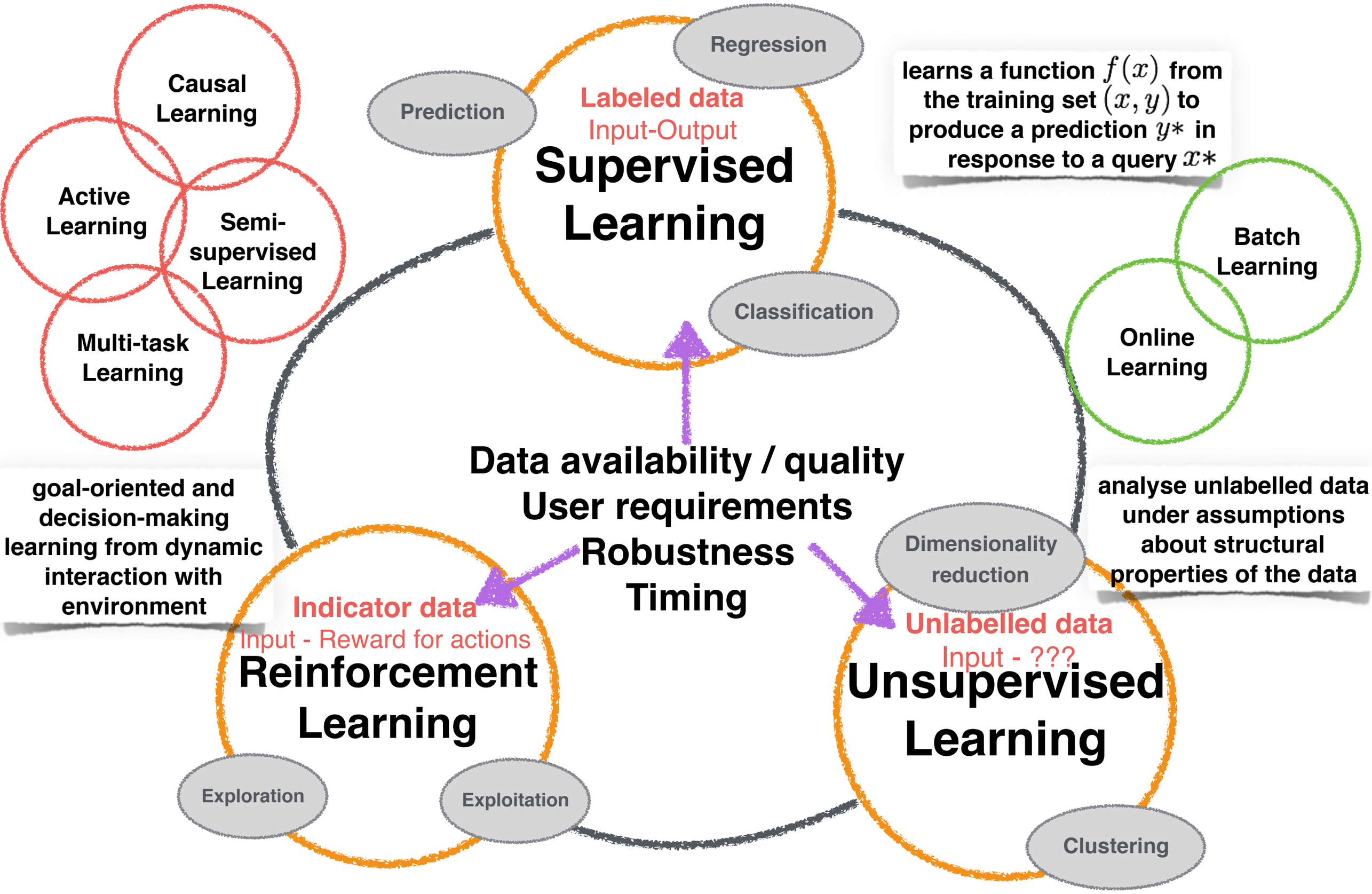
> We say that a program for performing a task has been acquired by learning if it has been acquired by any means ofter than explicit programming.
>
> *Valiant, 1984*

- Usually easy to obtain examples of what we want **IF x THEN DO y**
- Collect many pairs $\{(x_i, y_i)\}_{i=1}^{N}$
- Estimate function such that $f(x_i) = y_i$ **(supervised learning)**
- Detect patterns in data **(unsupervised learning)**

> . . . learning from examples, refers to systems that are trained instead of programmed with a set of examples, that is, a set of input/output pairs…
>
> *Poggio and Smale, 2003*



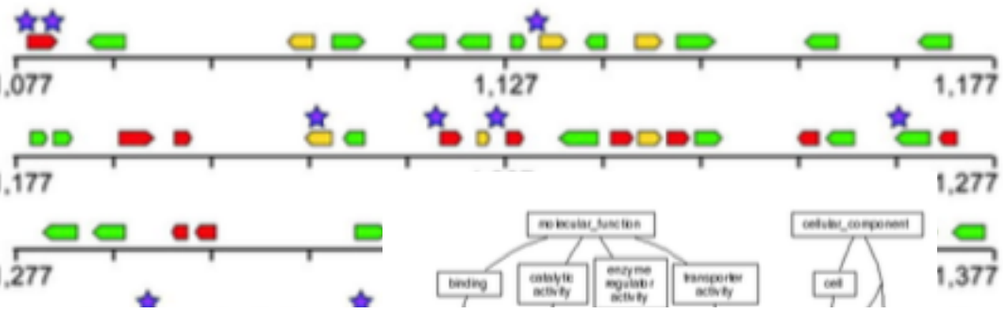
Three paradigms of learning



Problem Prototypes: Supervised Learning $y = f(x)$

We are given a set of input-output pairs (training set) $\{(x_i, y_i)\}_{i=1}^{n}$

- **Binary classification:** Given x find y in {-1,1}
- **Multi-category (Multi-class) classification:** Given x find y in {1,. . . ,K}
- **Regression:** Given x find y in $\mathbb{R}$ (or $\mathbb{R}^d$)
- **Sequence annotation:** Given sequence $x_1, \ldots, x_k$ find $y_1, \ldots, y_l$
- **Hierarchical categorisation (Ontology):**



Given x find a point in

- **Prediction:** Given x_t and $y_{t-1}, \ldots, y_1$ find

Forms of mapping f:

- decision trees / forests,
- logistic regression,
- support vector machine,
- neural networks,
- kernel machines,
- Bayesian classifiers.

Map image to category

given sequence

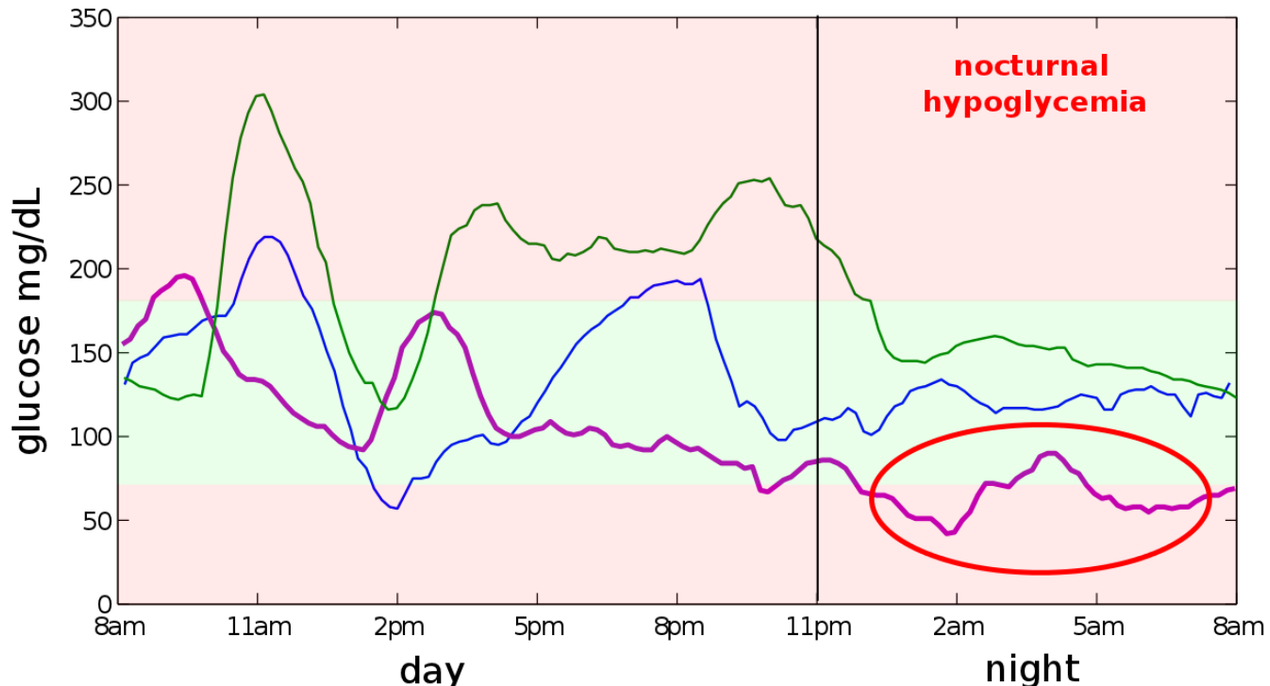
gene finding

speech recognition

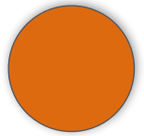
activity segmentation

Often with loss function

$$V(y, f(x))$$



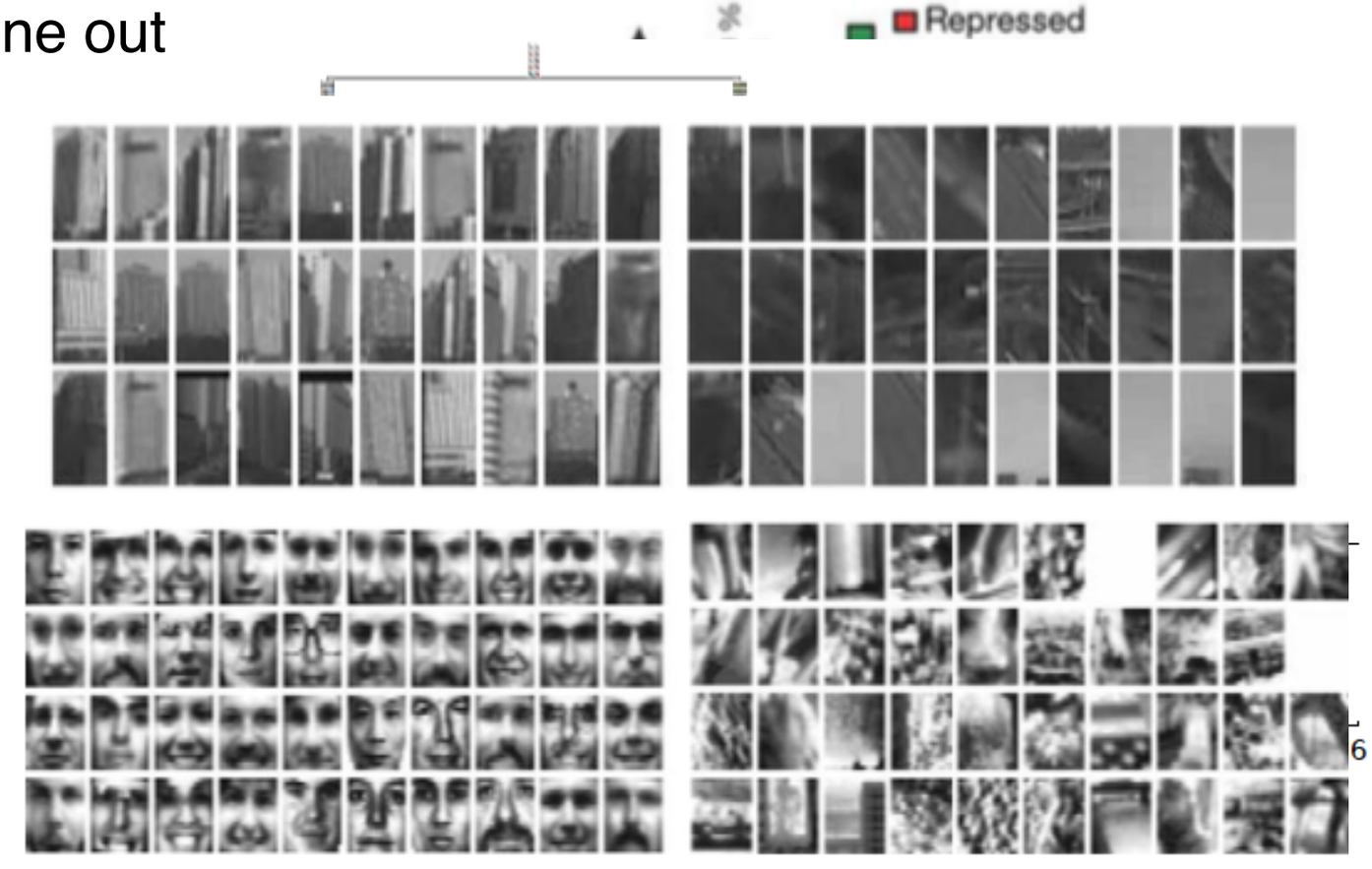
Problem Prototypes: Unsupervised Learning



Given data x , ask a good question…. about x or about model for x

- **Clustering:** Find a set of prototypes representing the data
- **Principal components:** Find a subspace representing the data
- **Sequence analysis:** Find a latent causal sequence for observations

  (Kalman Filter, Hidden Markov Model)
- **Hierarchical representation**
- **Dictionary learning:** Find (small) set of factors for observation
- **Novelty detection:** Find the odd one out



Problem Prototypes: Reinforcement Learning

Examples of desired behaviour are not available but it is possible to score examples of behaviour according to some performance measure

Sequential decision tasks:

- Take action
- Environment responds
- Observe stuff
- Update model
- Repeat

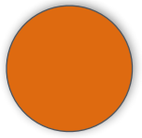
Bandits: Pick arm, get reward, pick new arm

- Choose an option
- See what happens (get reward)
- Update model
- Choose next option

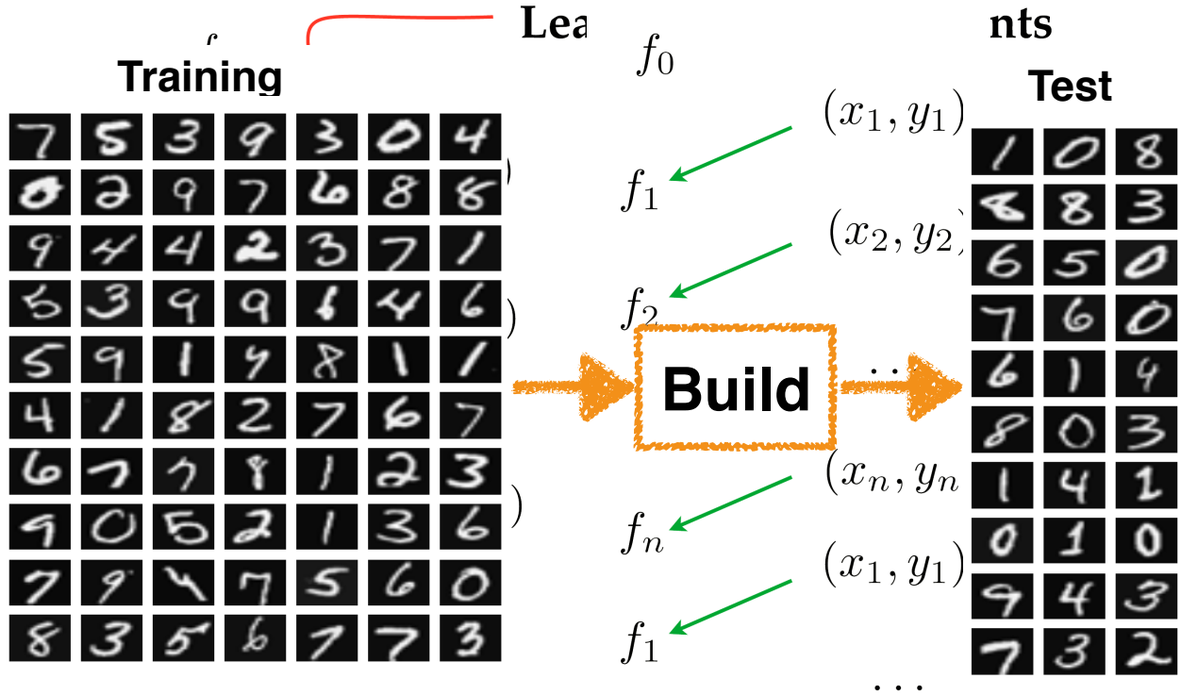
Reinforcement learning combines search and long-term memory

Differences to Supervised learning?

Problem Prototypes: Other types



- **Induction:** Reasoning from observed training cases to general rules, which are then applied to the test cases
- **Transductive inference:** Reasoning from observed, specific (training) cases to specific (test) cases



- **Multi-task learning:** Learns a problem together with other related problems at the same time, using a shared representation.
- **Active learning:** Interactive querying of the user to obtain the desired outputs at anew data points
- **Online learning:** Analyse each training example as it is presented
- **Batch learning:** Collect training examples, analyse them, output an hypothesis

Key Issues in Machine Learning

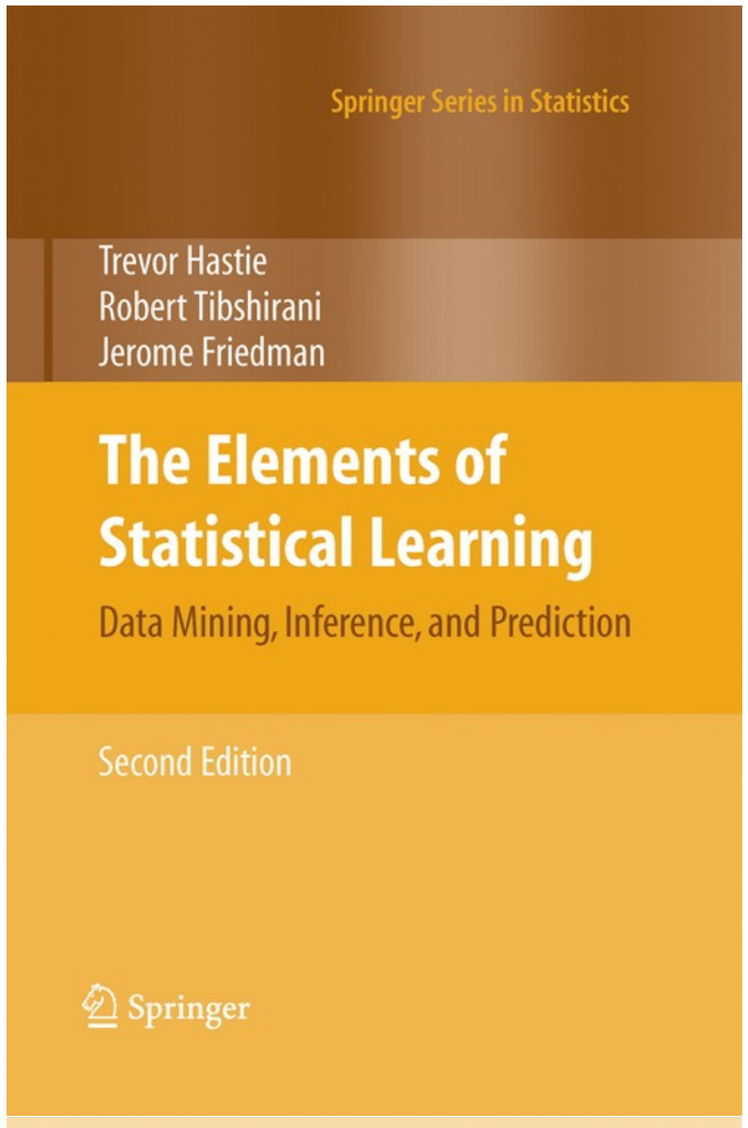
- **What are good hypothesis spaces?** Which spaces have been useful in practical applications and why?
- **What algorithms can work with these spaces?** Are there general design principles for machine learning algorithms?
- **How can we optimise accuracy on future data points?** This is sometimes referred to “generalisation” ability or problem of overfitting.
- **How can we have confidence in the results?** *(the statistical question)*
- **Are some learning problems computationally intractable?**

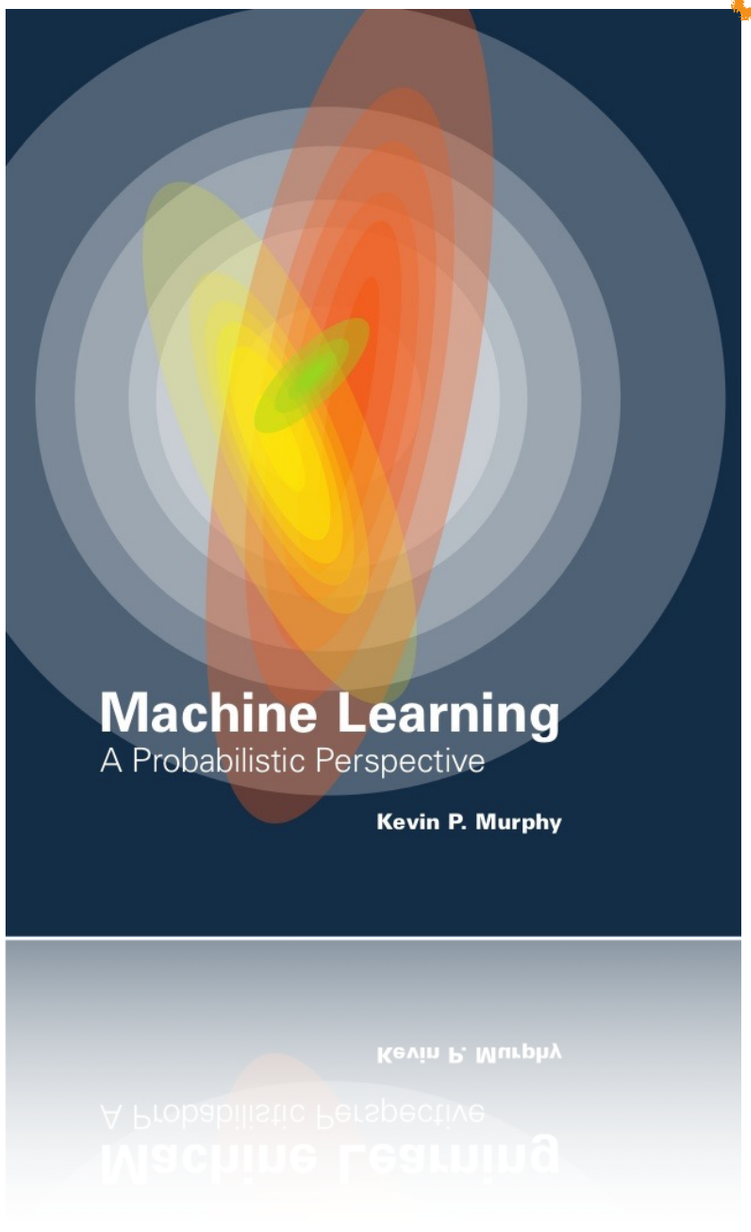
  (the computational question)
- **How can we formulate application problems as machine learning problems?** *(the engineering question)*

Hypothesis: a function produced by a learning algorithm, which is believed to be close to the true function.

Hypotheses space: the space of all hypotheses that can be an output by a learning algorithm.

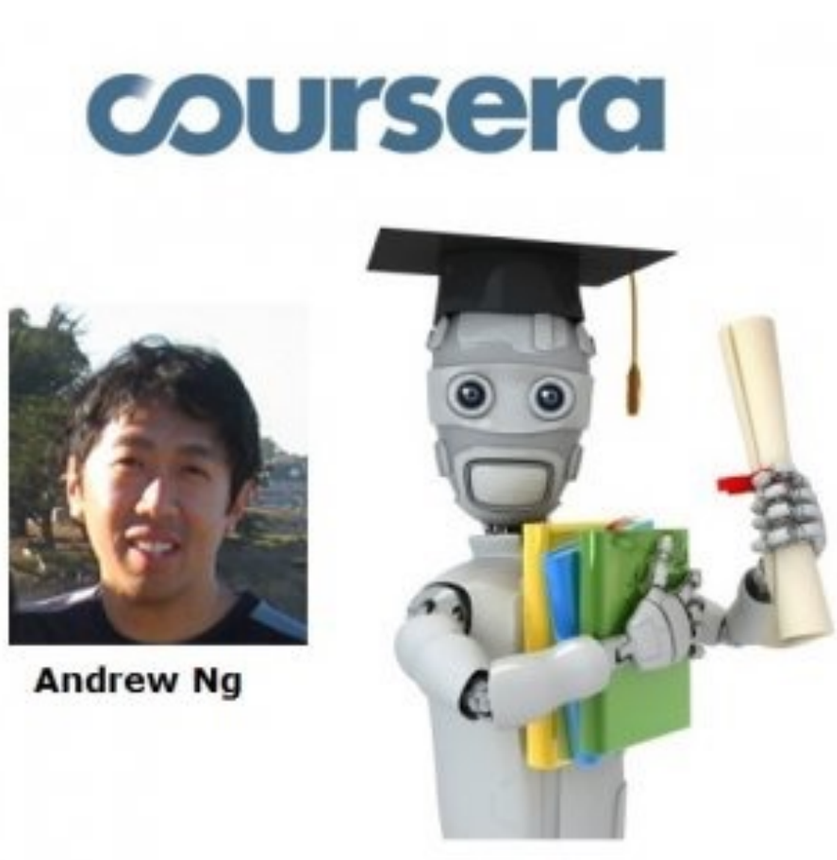
Literature





We are drowning in information and starving for knowledge.

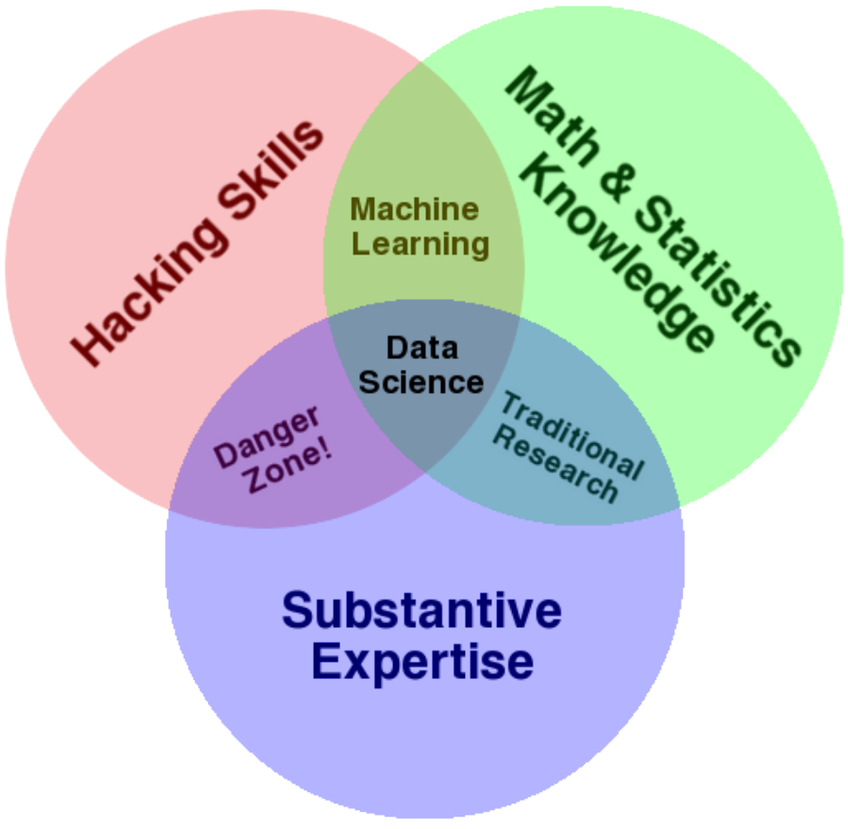
Rutherford D. Roger, American Librarian



Online Resources

1. **Python ML Library**: scikit-library
2. **Machine Learning Open Source Software**: http://mloss.org/software/
3. **Kaggle** your home for data science

Conclusions



- Big data analytics is a promising right direction which is in **its infancy** for the healthcare domain.
- Healthcare is a data-rich domain. As more and more data is being collected, there will be **increasing demand for big data analytics.**
- Unraveling the “Big Data” related complexities can provide many insights about making the **right decisions at the right time** for the patients.
- Efficiently utilizing the colossal healthcare data repositories can yield some immediate returns in terms of **patient outcomes and lowering care costs.**
- **Data with more complexities** keep evolving in healthcare thus leading to more opportunities for big data analytics.