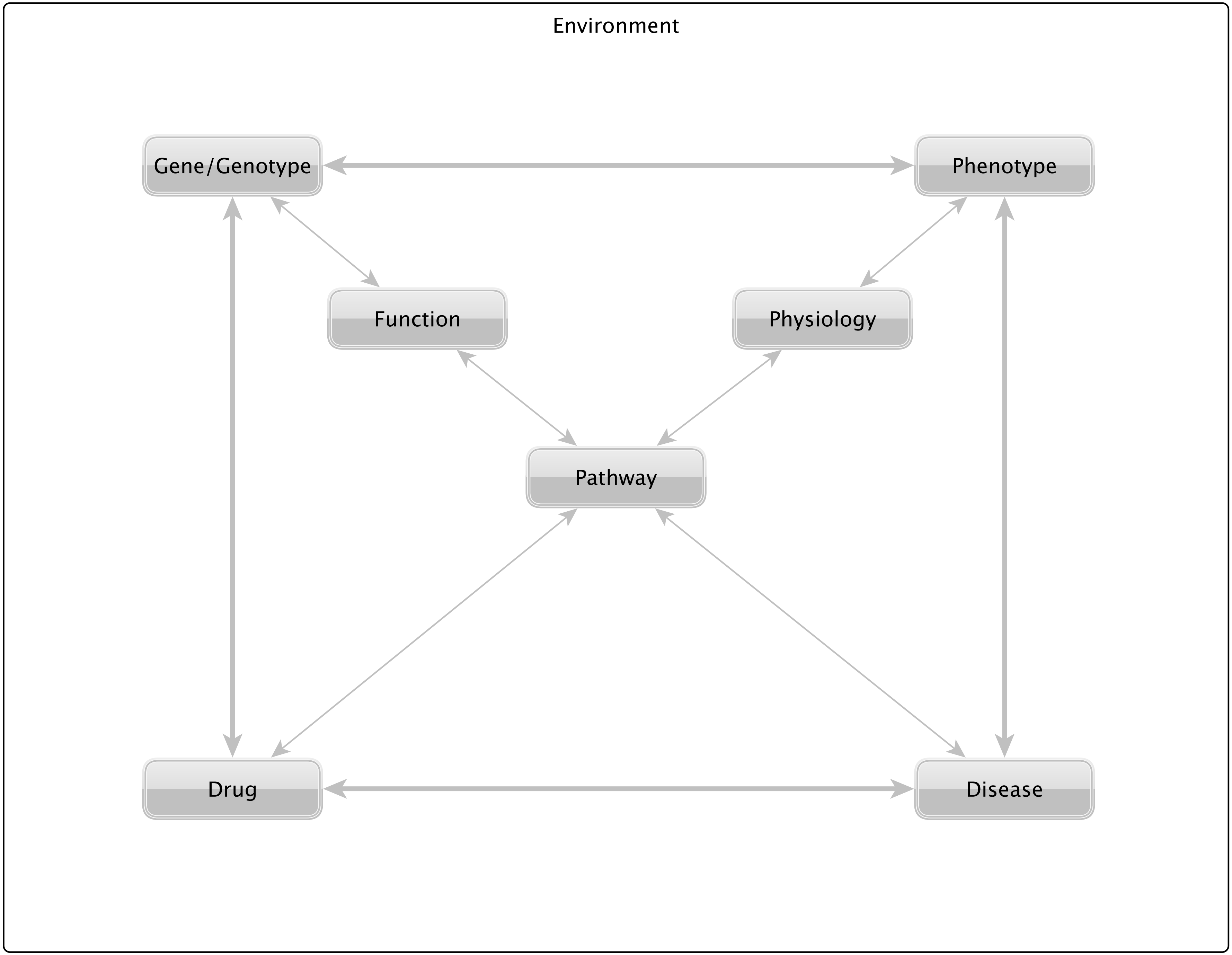
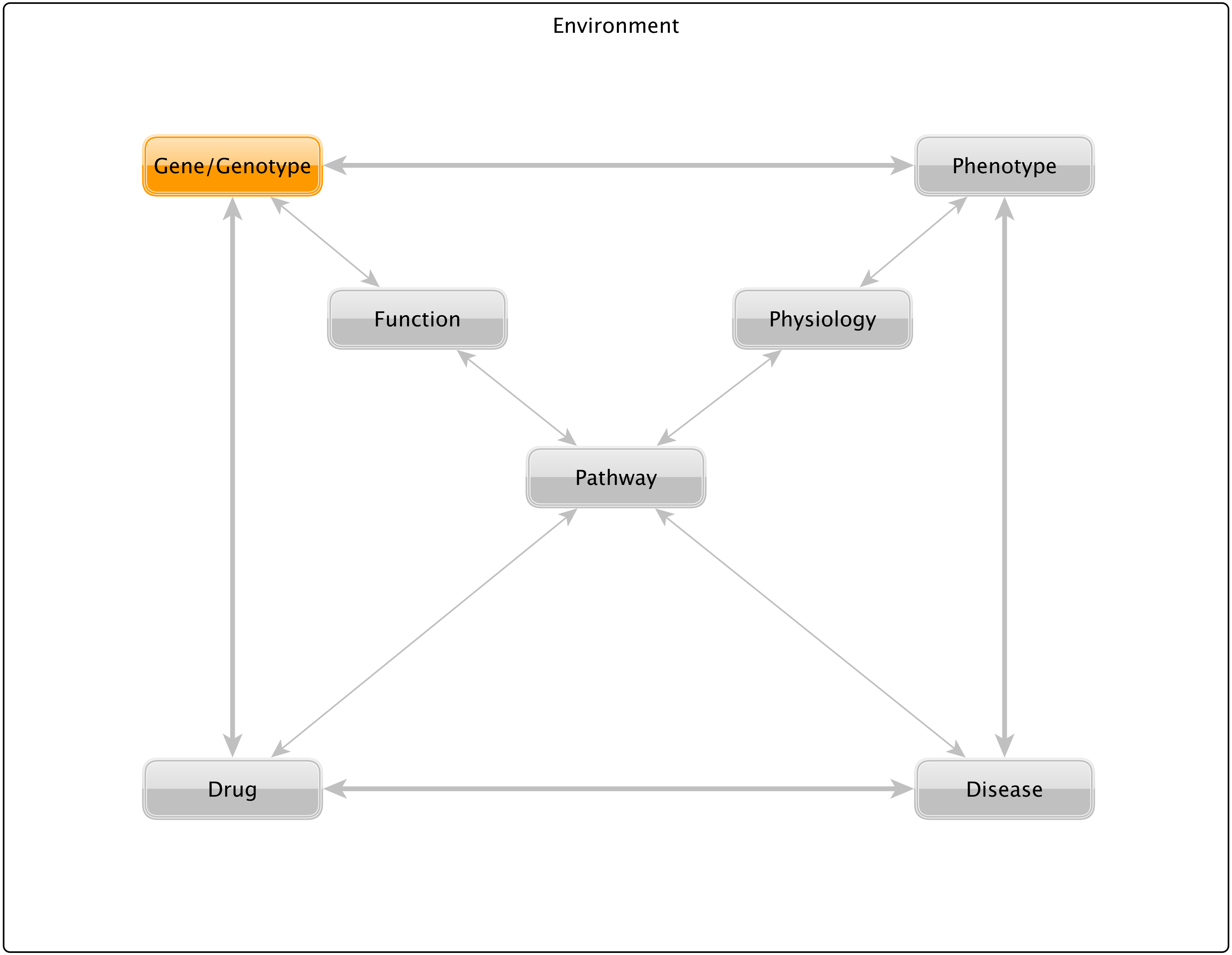
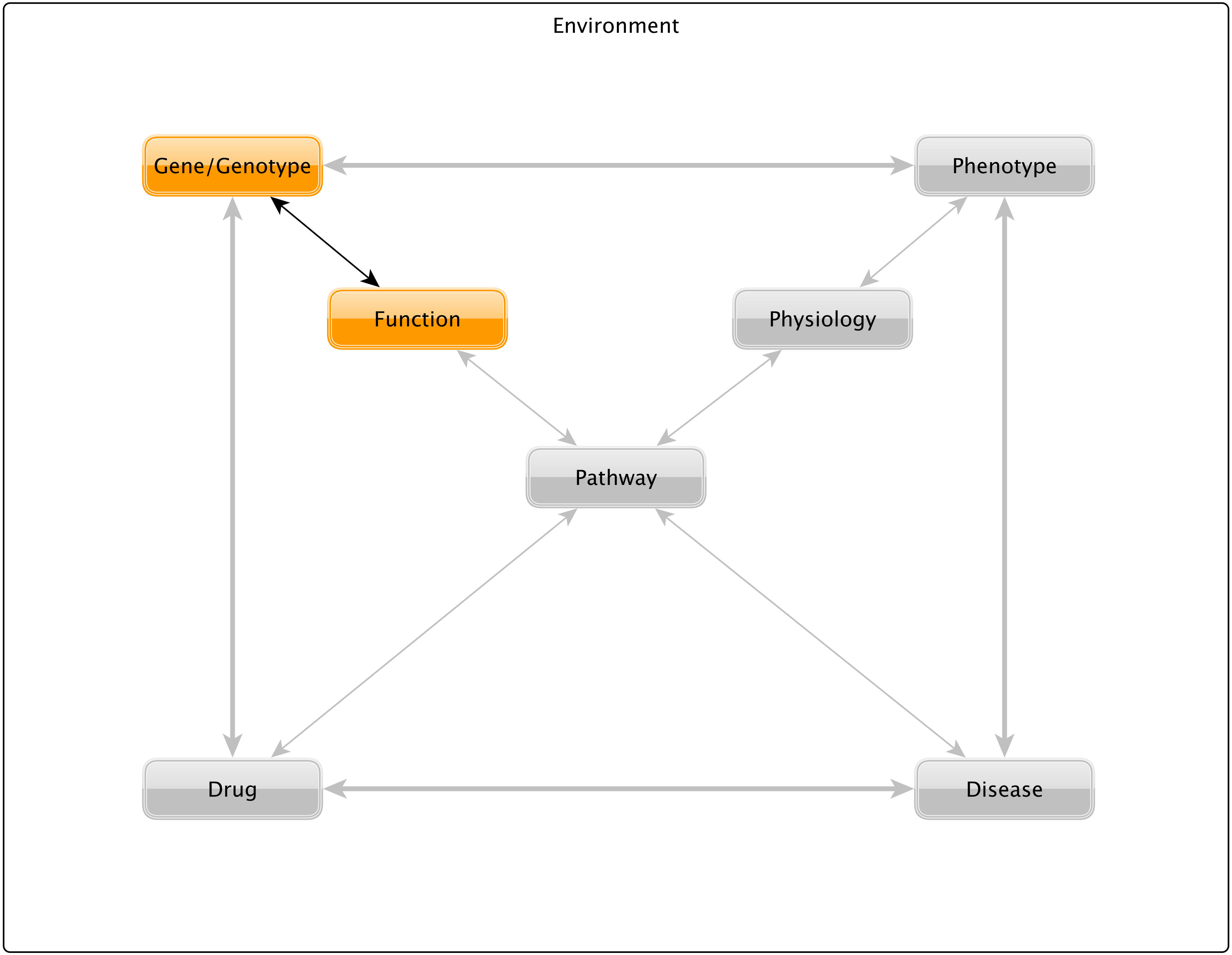
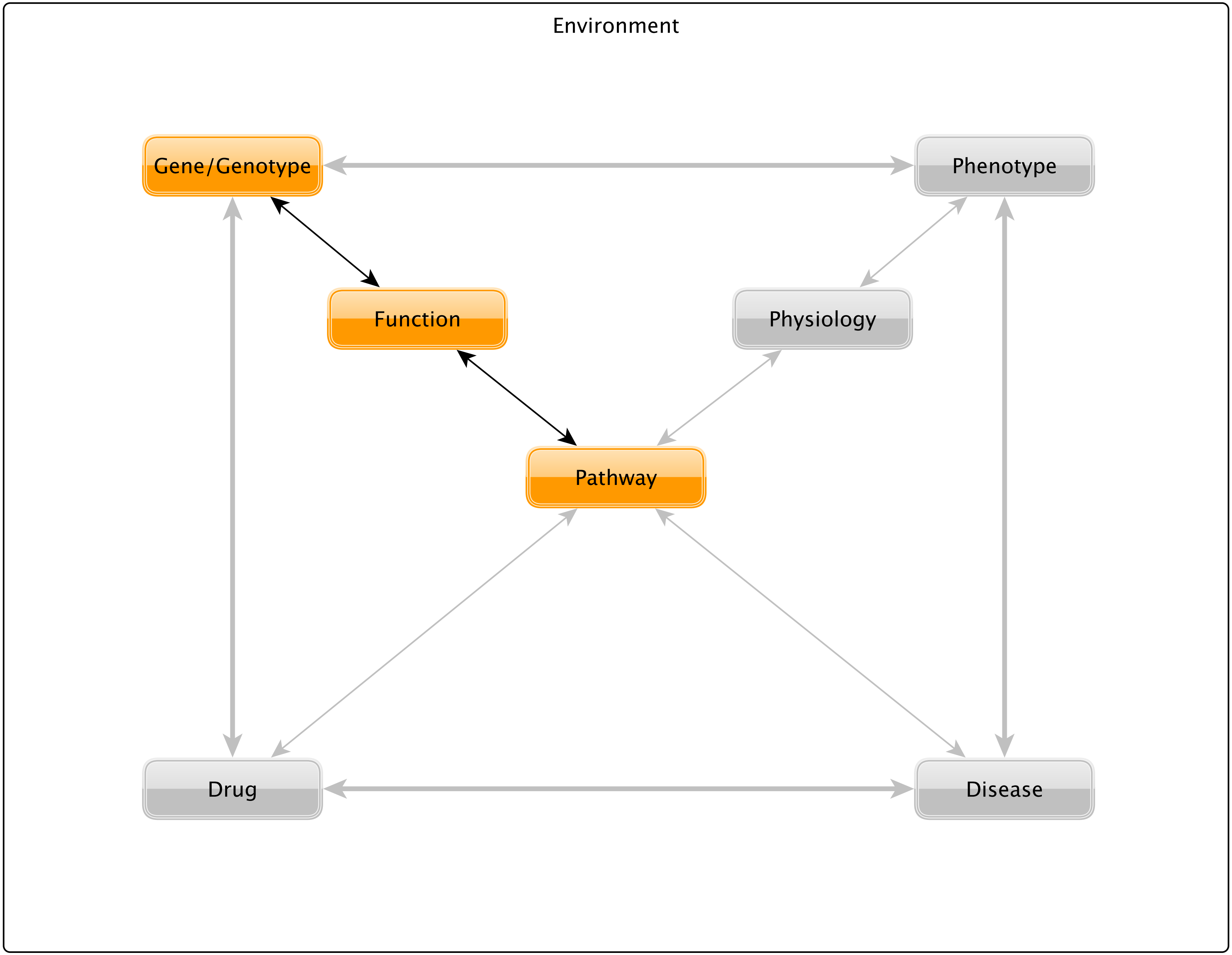
Semantic Representation & Biomedical research


Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease

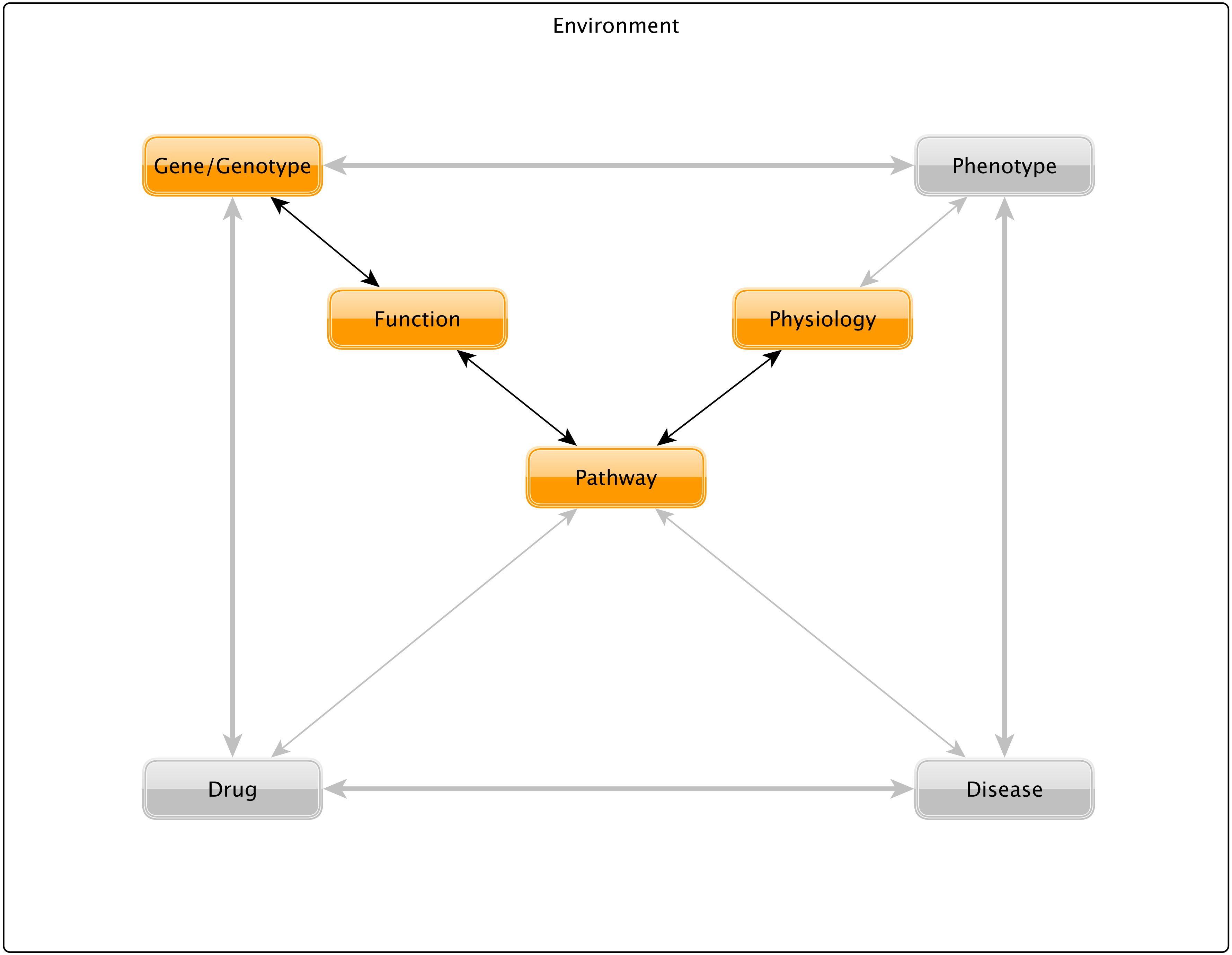


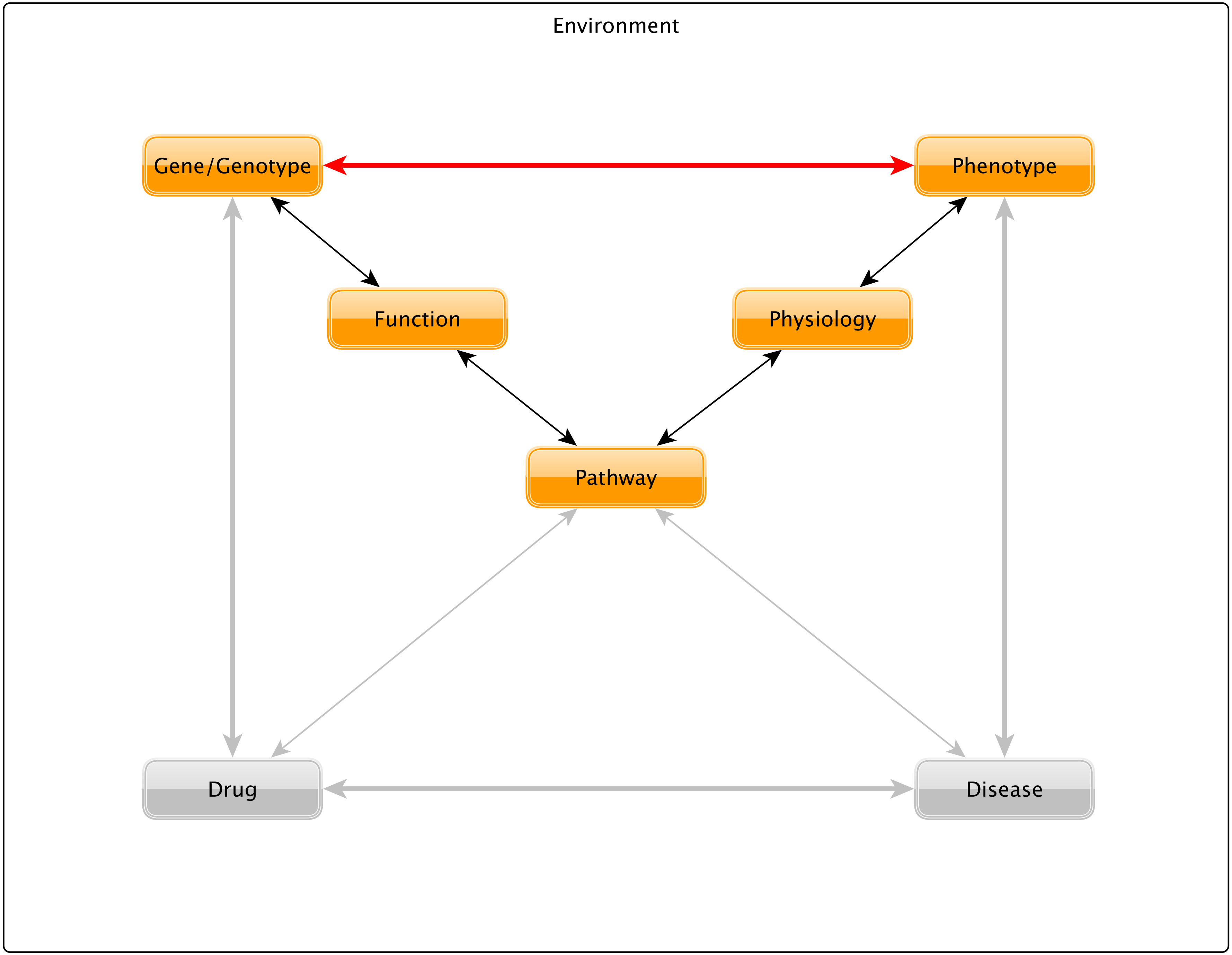
Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease

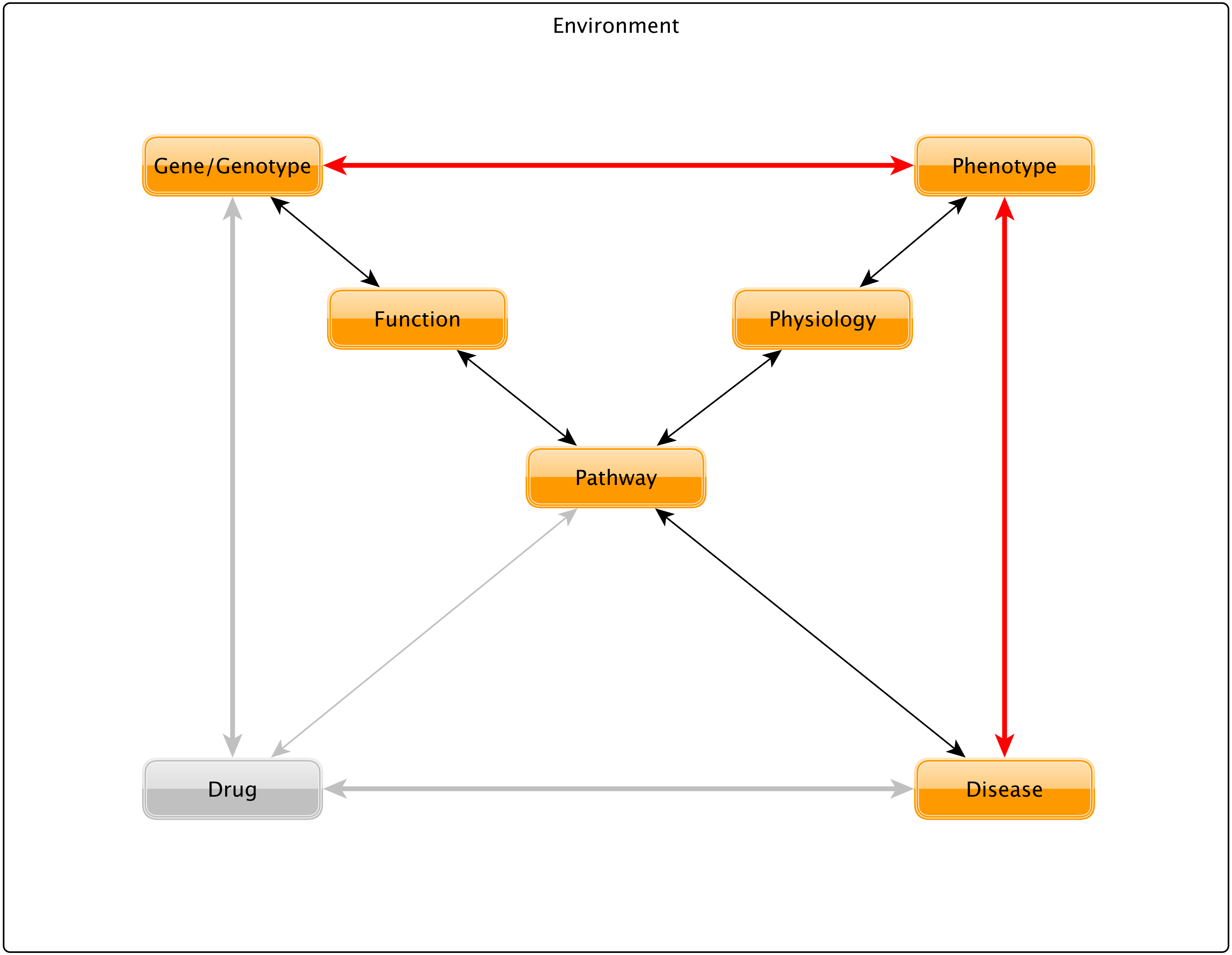

Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease

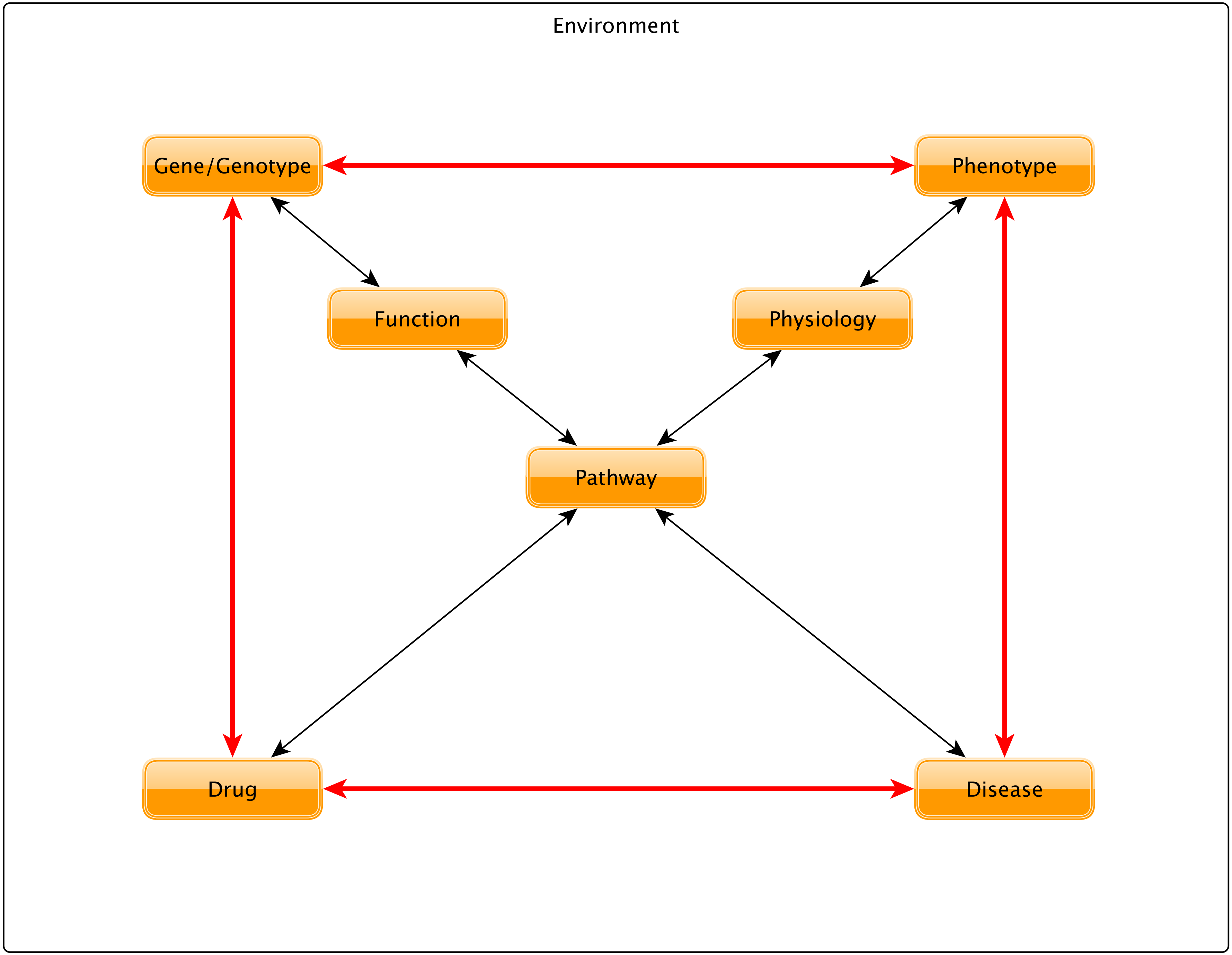


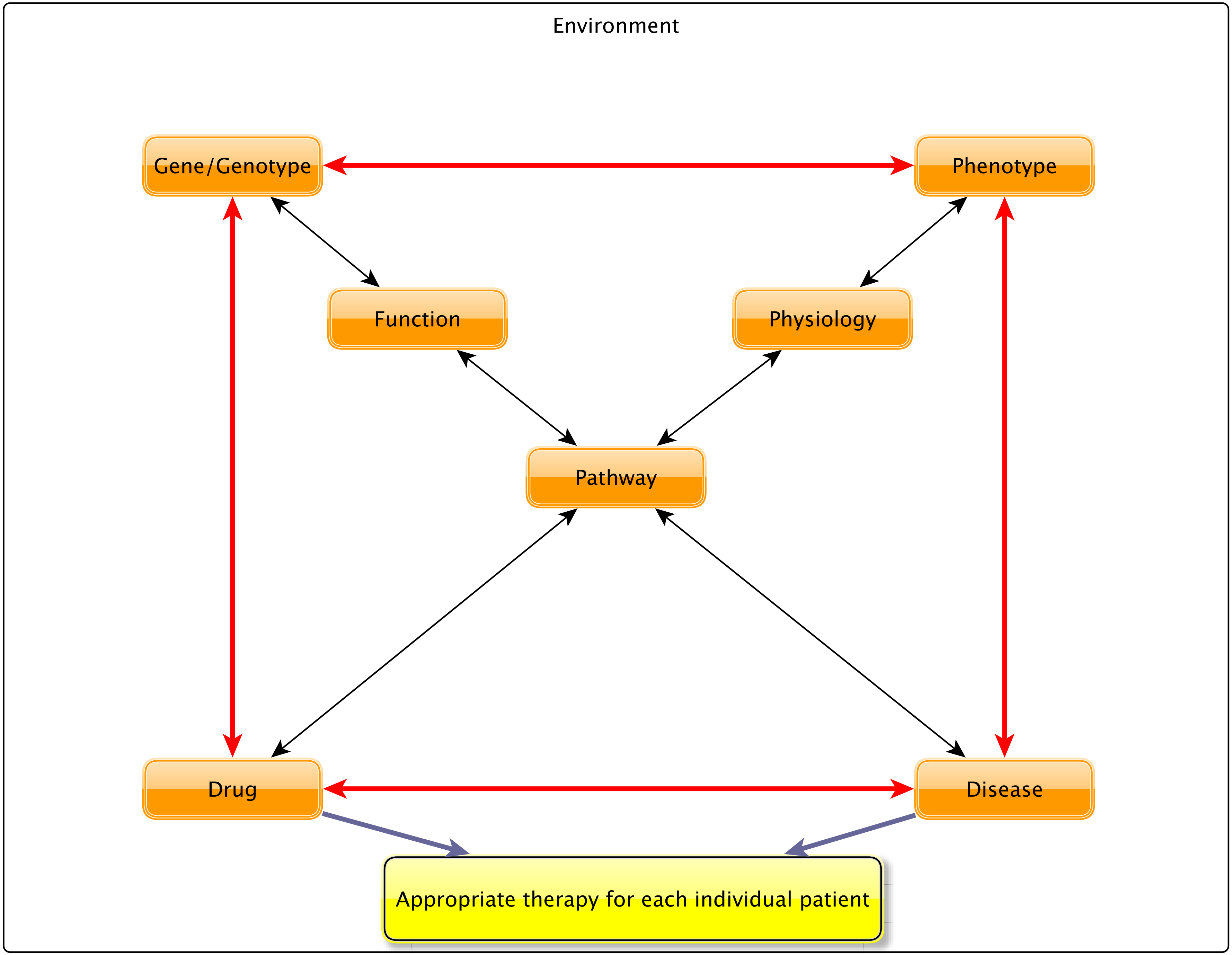
Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease


Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease


Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease


Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease


Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease


Environment
Gene/Genotype
Phenotype
Function
Physiology
Pathway
Drug
Disease
Appropriate therapy for each individual patient

Gain an understanding of the molecular basis of human disease

↓

cause and affect

biological context in which the illness occurs

fundamental biological processes causing a disease

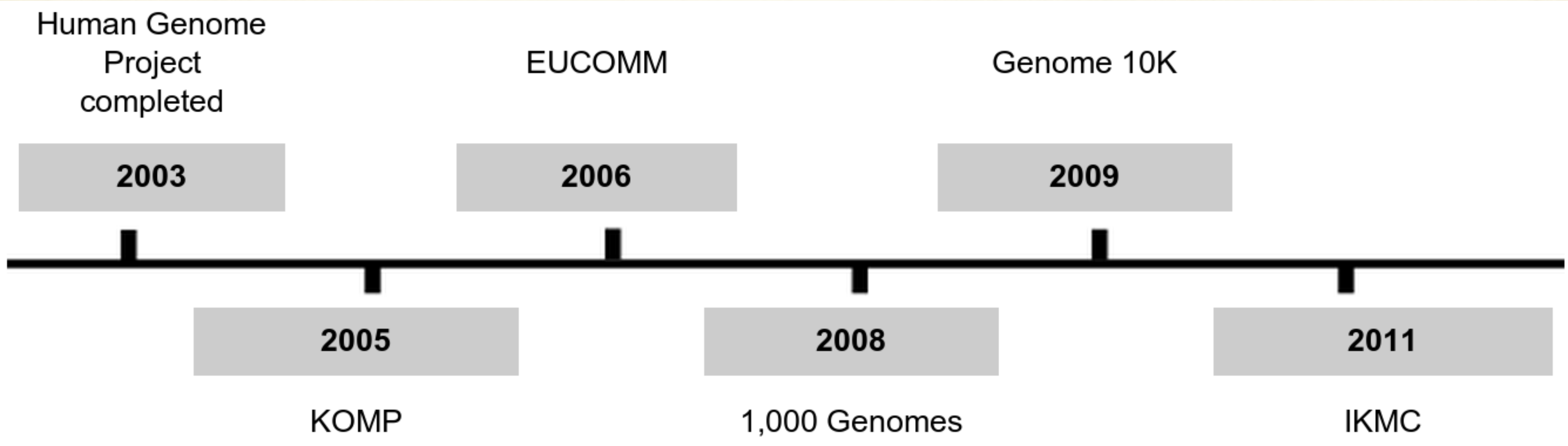
↓

appropriate therapy for each individual patient

genotype of an individual

experimental data from model organisms

Exploring the Phenome



Key EU/NIH missions:

– integration and analysis of disease data within and across species → diagnostic and therapeutic advances at the clinical level

– identification of causative genes for Mendelian orphan diseases

NATURE | NEWS

Data-crunch highlights potential transplant drugs

Widely prescribed statin could have alternative applications.

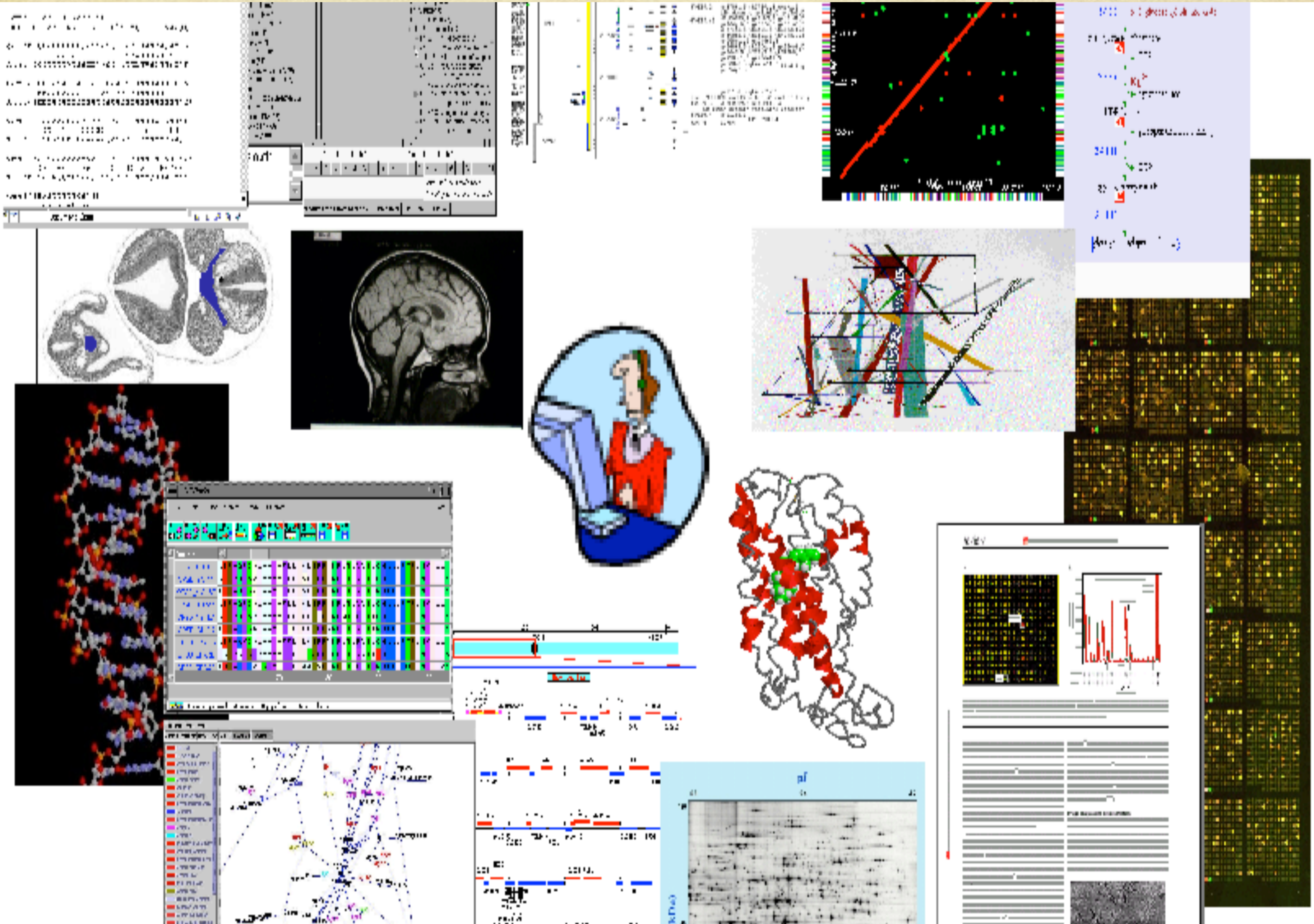
Monya Baker

14 October 2013

““It took me about thirty minutes. Honestly, it is scary how easy it seems now, in retrospect.””

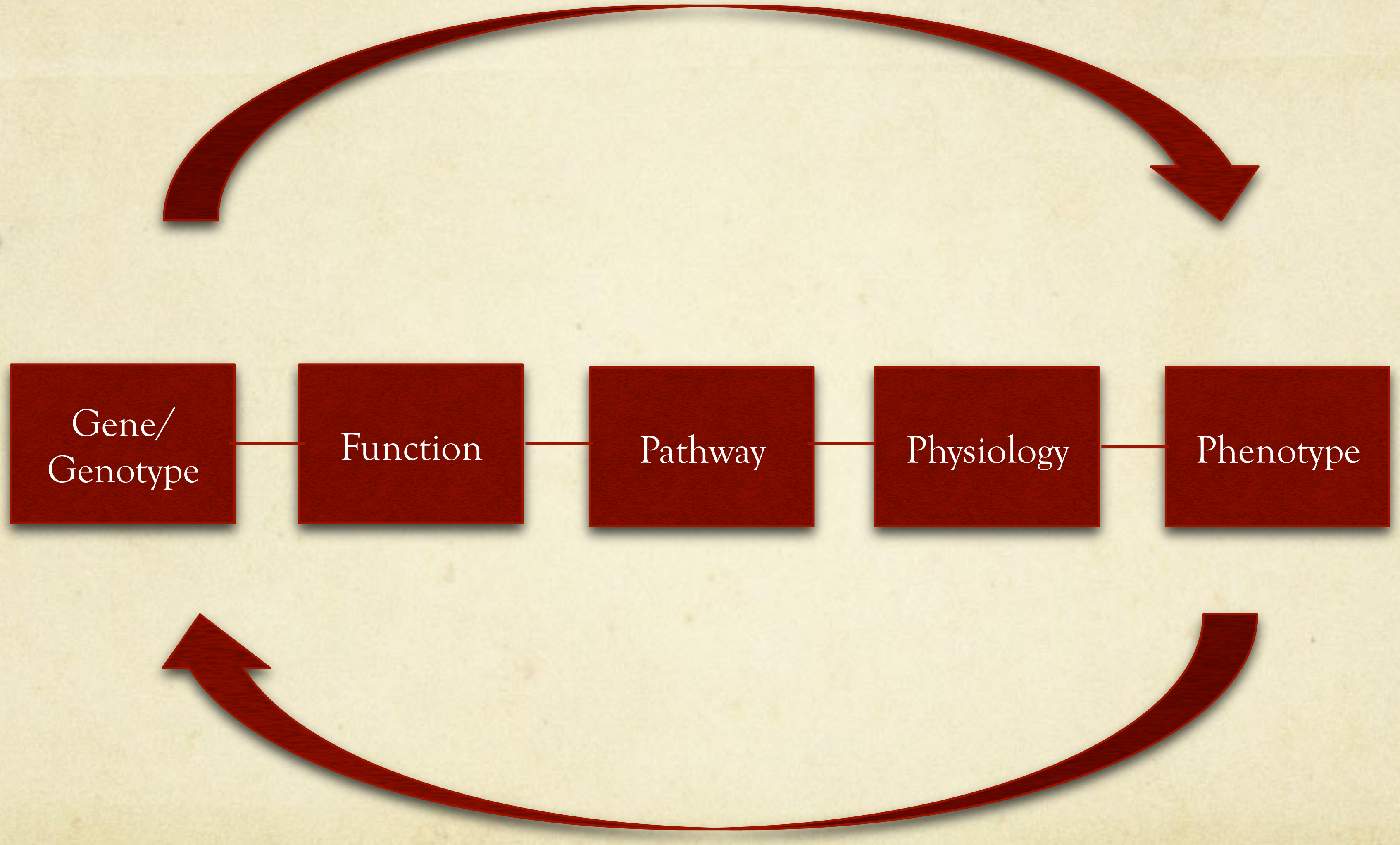
“This is a good story, and there is some promise for future directions,” Suthanthiran adds. “It will be nice to see these drugs evaluated in a prospective clinical trial.”

Nature doi:10.1038/nature.2013.13944



Reverse Genetics

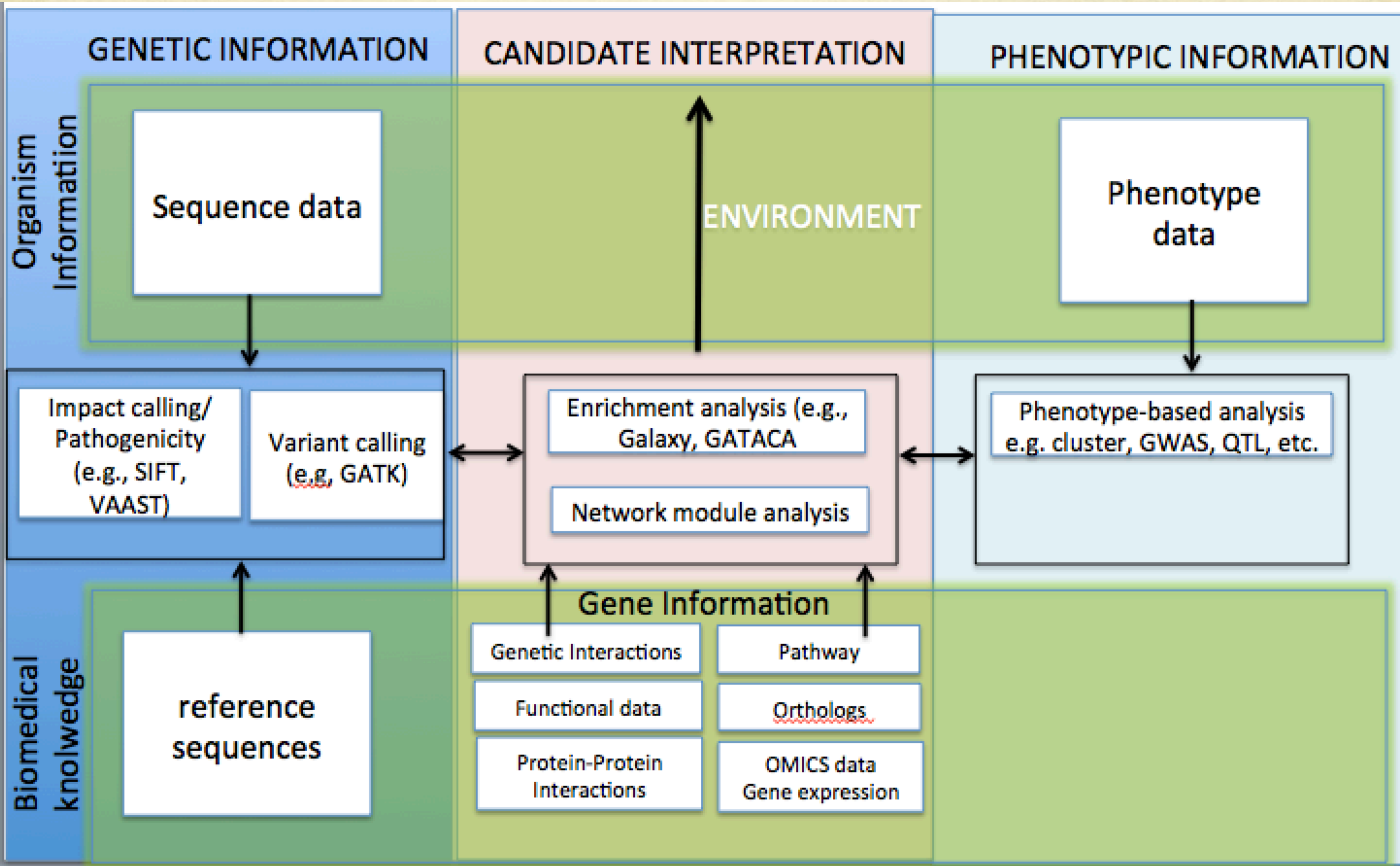
Functional Analysis



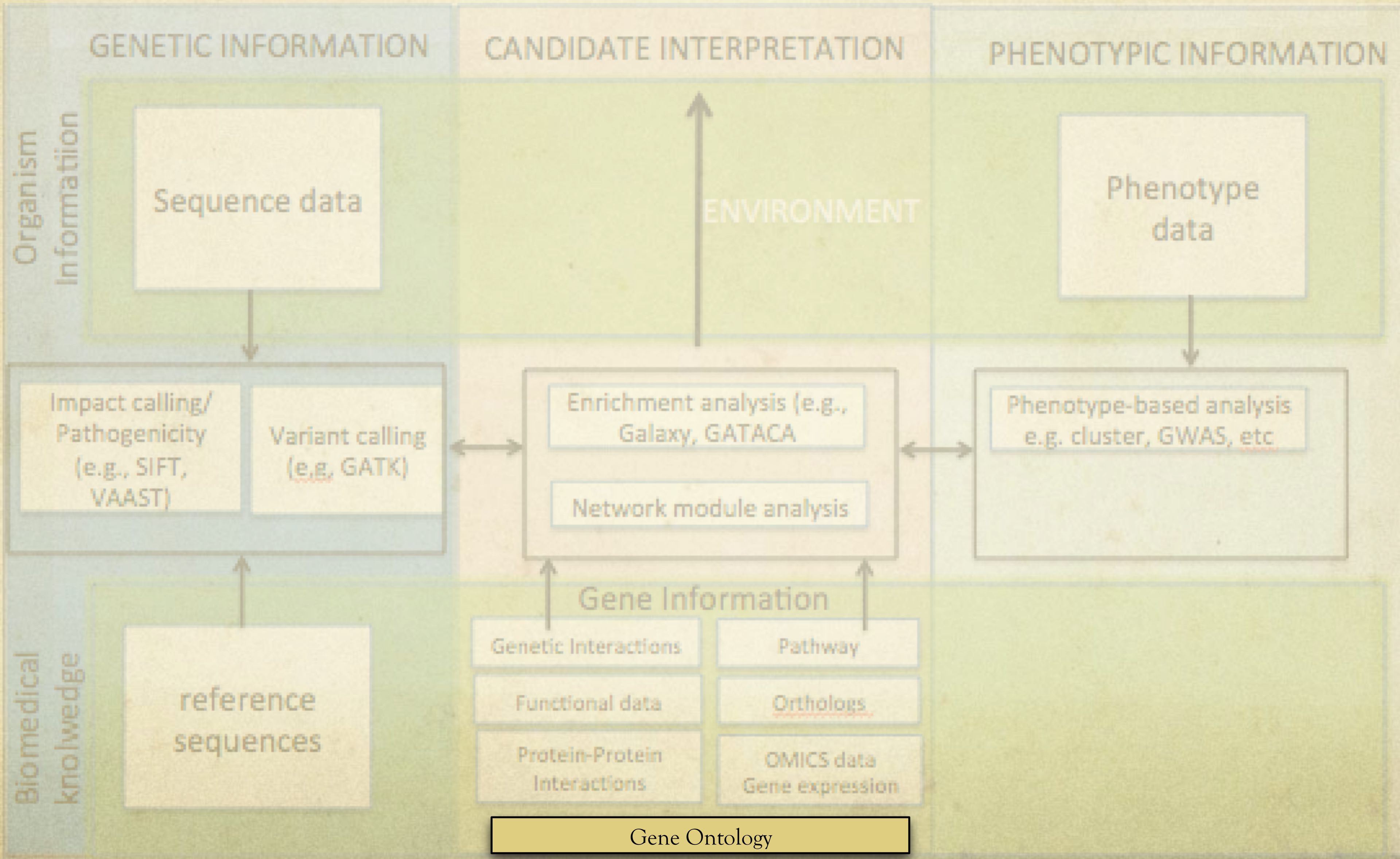
Forward Genetics

Positional Cloning

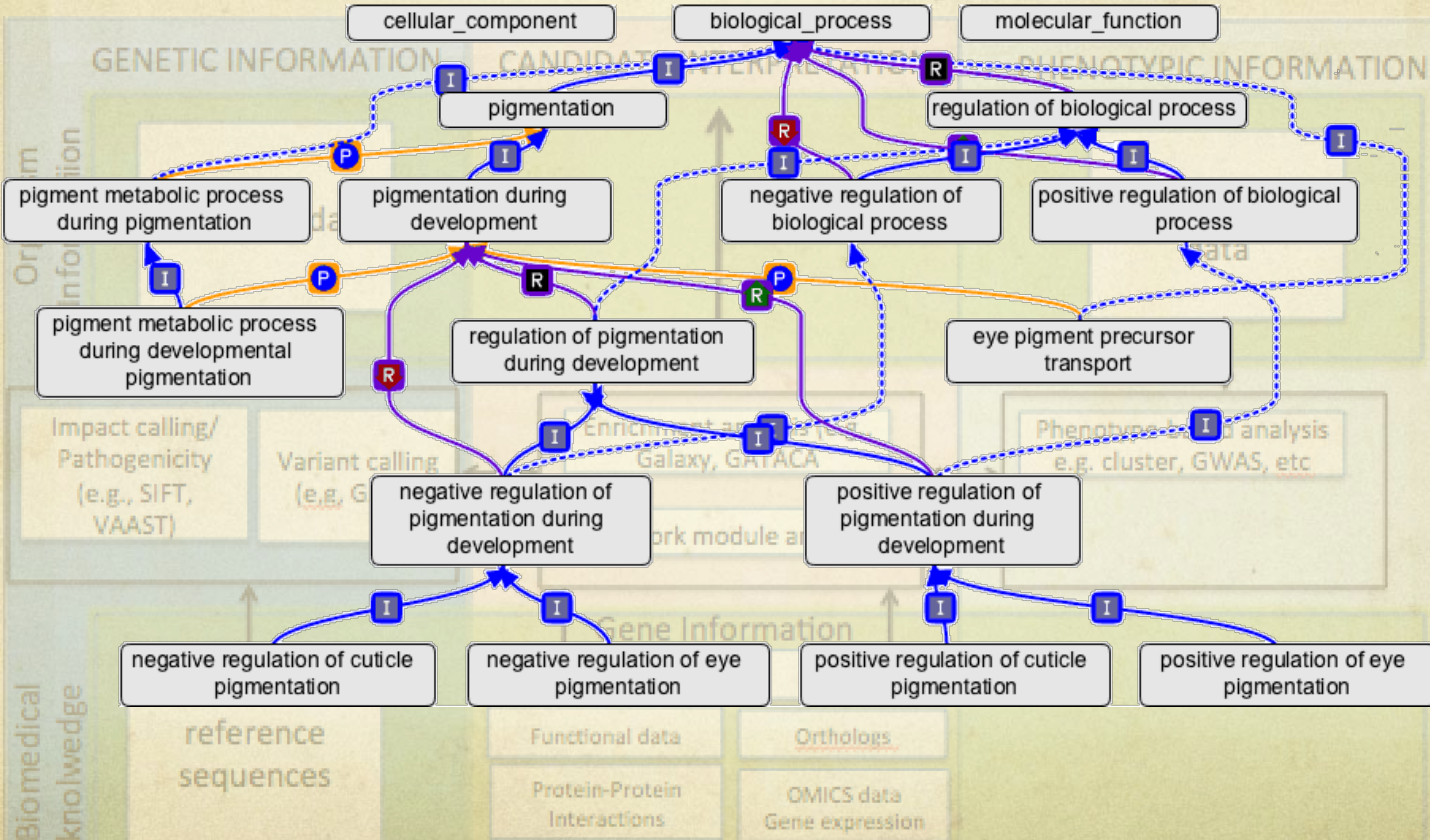
Candidate gene prioritization



Gene Ontology



Gene Ontology



Gene Ontology

How much data?

- Our ability to identify causative variants/variants of interest *depends* on the layer of biological knowledge

- More genetic data will *increase our ability* to prioritise gene candidates (GWAS, QTL, etc.)

- More phenotype data will *alter our potential* for revealing gene candidates

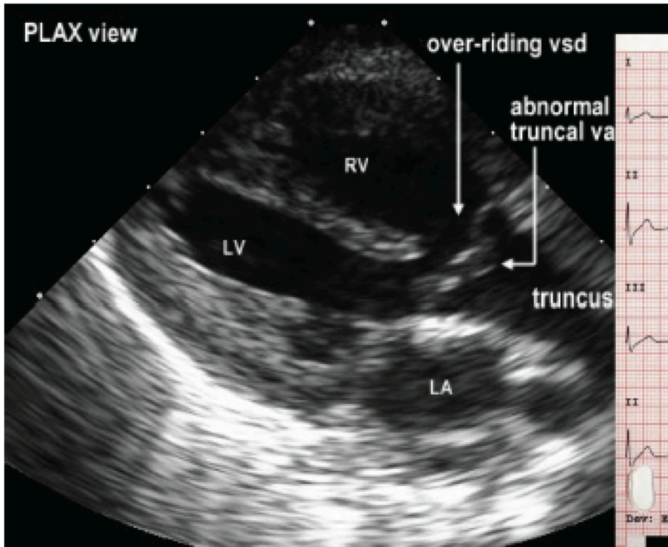
The next step

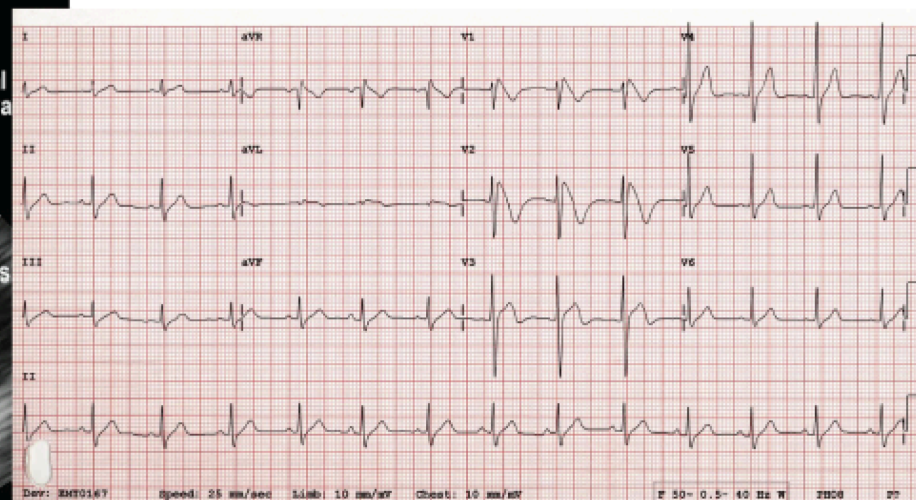
We have become increasingly adept at comparing genetic sequences

Can we compare phenotypes/traits in a similar fashion we compare genetic sequences?

What can we learn about the molecular mechanisms underlying phenotypes through this analysis?

Phenotypes in the clinic



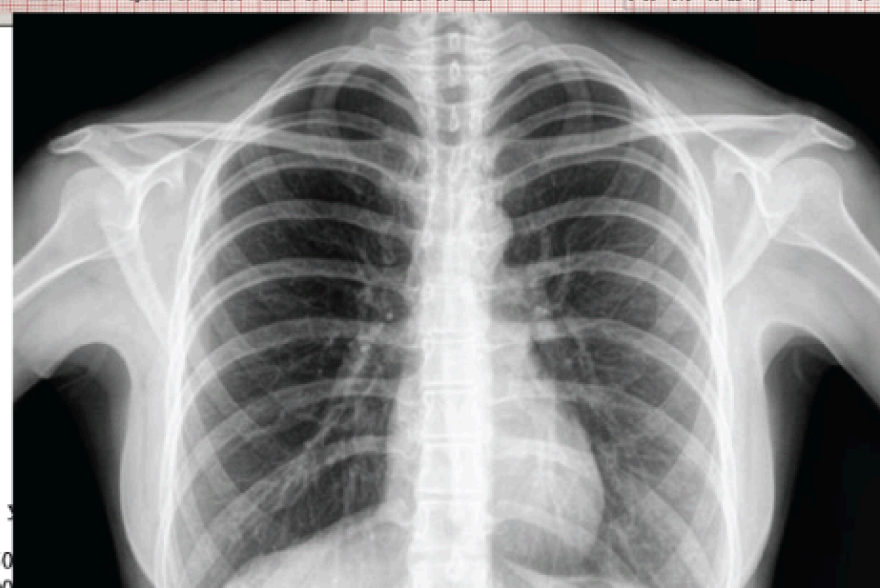


Complete Blood Count:

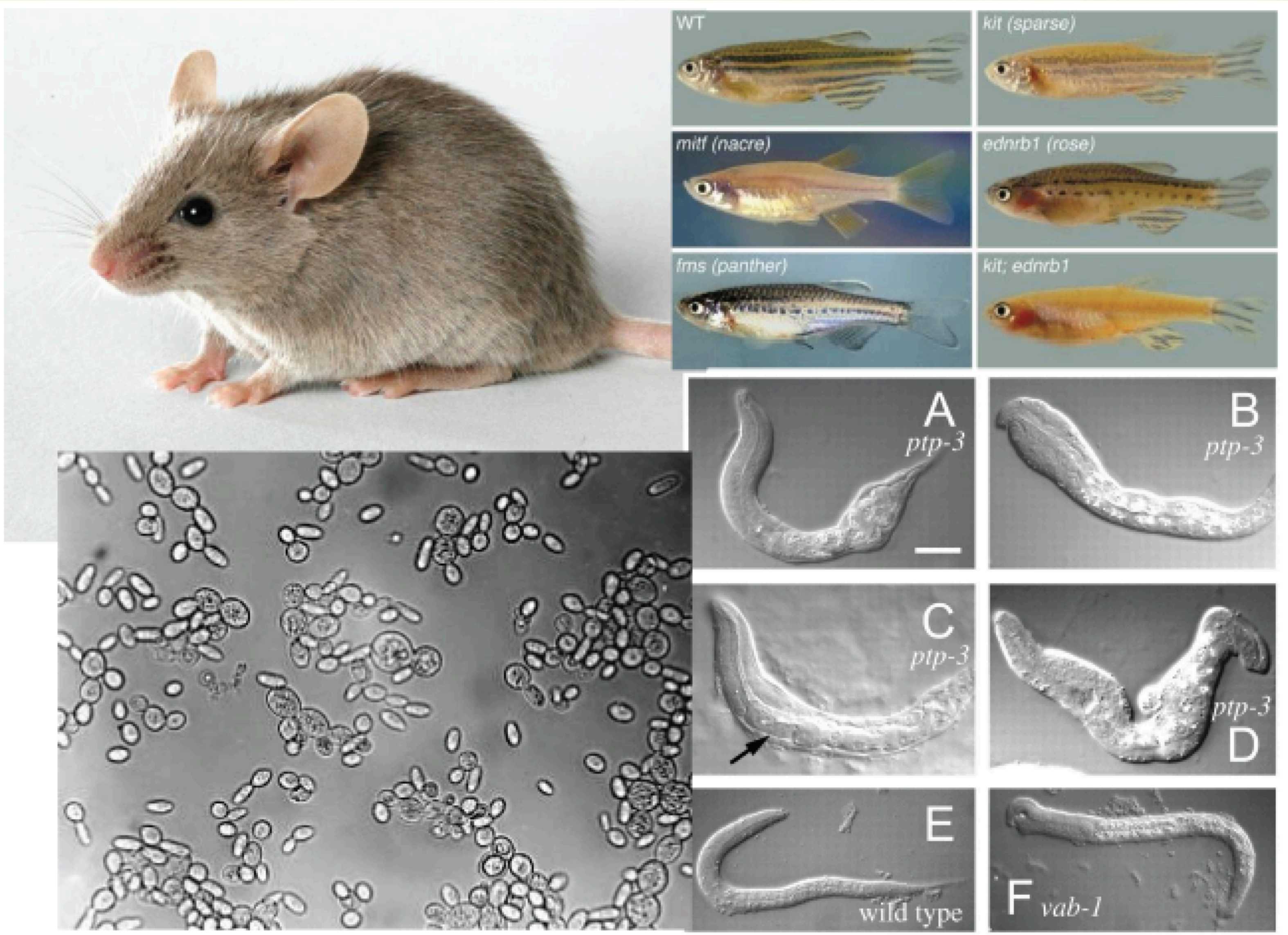
	Patient Value	Normal Range 2 years – 6 years
WBC	8.4 x 10^9 / L	(5.0 – 17.0)
RBC	2.77 x 10^{12} / L	(3.90 – 5.30)
Hgb	7.5 g/dl	(11.5 – 13.5)
Hct	21.8 %	(34.0 – 40.0)
MCV	78.6 fl	(75.0 – 87.0)
MCH	26.9 pg	(25.0 – 31.0)
MCHC	34.2 gm/dl	(31.0 – 36.0)
RDW	17.3 %	(11.5 – 15.0)
PLT	192 x 10^9 / L	(150 – 450)

Differential:

	Absolute	Normal Range Number	2 years – 6 y
Neutrophils	43 %	(3.61)	(1.50 – 8.50
Bands	6 %	(0.50)	(0.00 – 1.00



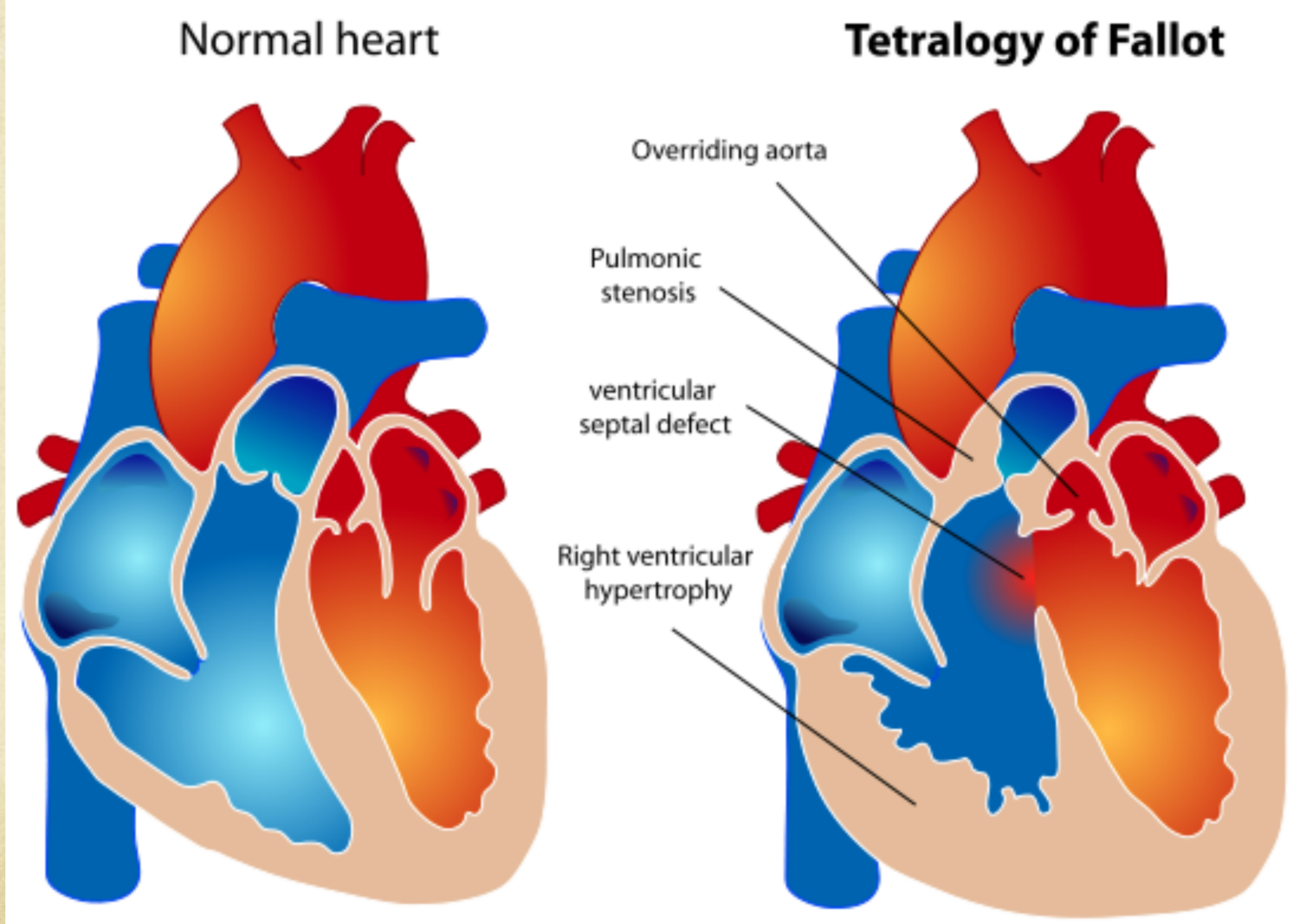
Animal Model Phenotypes



Plant Phenotypes

- Comparative Phenomics
- Gene function determination
- Systematic Genome-Wide Phenotyping
- Translational Research
- Rare and orphan diseases
- Clinical diagnostics decision support systems
- Novel drug discovery and repurposing
- Disease and drug pathways
- Genotype to Phenotype

Candidate disease gene prioritization



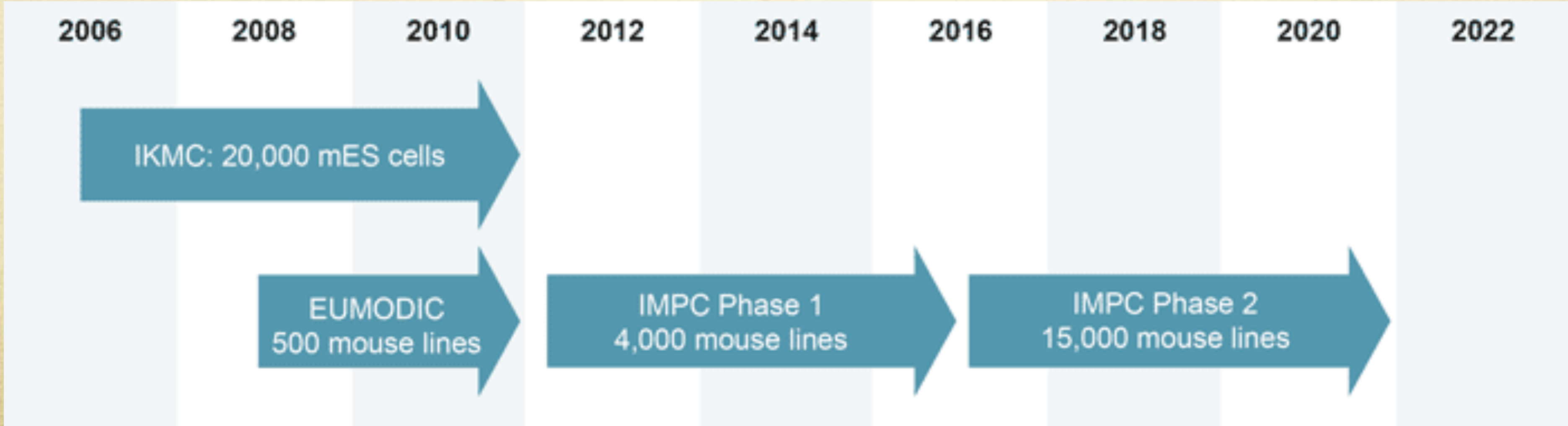
The promise of animal models

- Forward and reverse genetics (e.g. Collaborative Cross panel, IKMC/IMPC)

  → **understanding of gene function by taking a pan-genomic and pan-phenomic approach**

- EU invested to date > €700 million

- IMPC - systematic, agnostic phenotyping of the mouse genome



Gene-Disease associations based on minimal phenotype information

- The nature of a phenotyping pipeline is breadth whilst depth will rely on secondary and tertiary phenotyping carried out by domain experts

- Results to date has not revealed any significant associations

- Challenge – select genes based on primary screens to look at for secondary phenotyping

- Aim - platform for the identification of possible gene disease associations based on minimal phenotype information

Nsun2

Gene Details

Marker Name(s):	NOL1/NOP2/Sun domain family member 2 view this gene in MGI
Marker Type:	protein coding gene
Synonyms:	D13Wsu123e, Misu
Location:	Chr13:69672624-69774658(+)

Related Human Conditions (from OMIM) – *no related Human Condition*

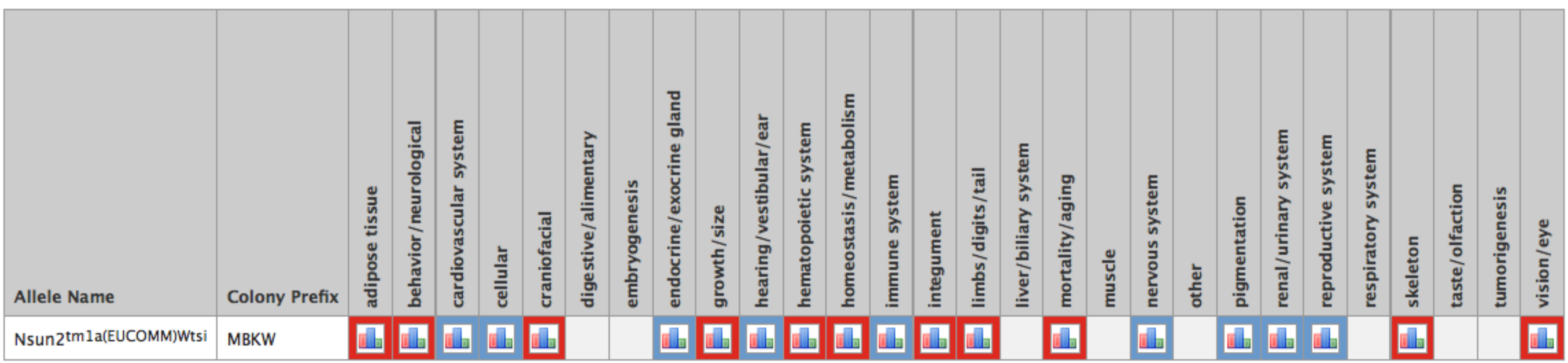
More Information

Data provided by Mouse Genome Informatics (MGI), Ensembl

Hyperactivity
Glucose homoeostasis
Decreased body fat
Decreased grip strength
Decreased body weight
Increased erythrocytes
Decreased blood lipids
Abnormal skeletal morphology and mineralisation
Cataracts
Abnormal cornea

WTSI Phenotyping

MP Ontology Based Heatmap



Legend: No Raw Data | No Significant Annotations | Significant Annotation Present | Link to a test report page

Rare and orphan diseases

Number of Entries:

Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
* Gene description	12,750	627	48	35	13,460
+ Gene and phenotype, combined	250	14	0	2	266
# Phenotype description, molecular basis known	2,836	240	4	28	3,108
% Phenotype description or locus, molecular basis unknown	1,628	135	5	0	1,768
Other, mainly phenotypes with suspected mendelian basis	1,819	130	2	0	1,951
Totals	19,283	1,146	59	65	20,553

- at least 3000 diseases without known molecular basis
- disease-gene identification methods have a limited focal range that may include up to 300 genes
- necessary to suggest possible causative geves

Bassoe Syndrome

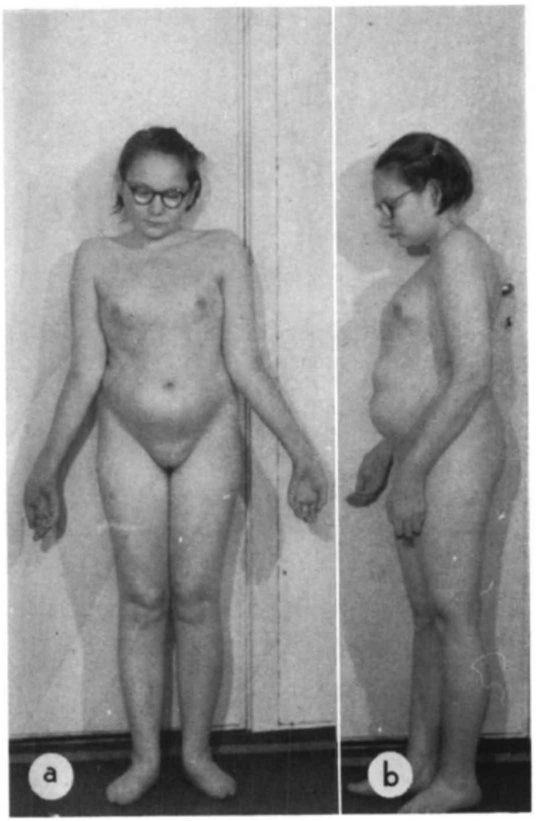


FIG. 2. *Case 1*. **a.** Showing cubitus valgus, elongated extremities and lack of pubic hair. **b.** Showing protruding abdomen, thoracic kyphosis and underdeveloped breasts.

FAMILIAL CONGENITAL MUSCULAR DYSTROPHY WITH GONADAL DYSGENESIS

HANS H. BASSÖE, M.D.*

Hammerfest Hospital, Hammerfest, Norway
(*Medical Section—Head, H. Schartum-Hansen*)

IN A family living in a small isolated village in Finnmark county, Norway, we have observed 7 persons suffering from congenital muscular dystrophy. Their symptoms were similar to those seen in congenital amyotonia (Oppenheim). Several children in the family had died very early in life; 3 others had been still born. In the third generation (III, Fig. 1) the child mortality was 33.3 per cent, whereas the average child mortality in Finnmark in the period 1926–1930 was 9.17 per cent. In the same generation, 2 siblings (Cases 1 and 2) were affected. In addition to the muscular dystrophy there was gonadal dysgenesis—ovarian agenesis and testicular

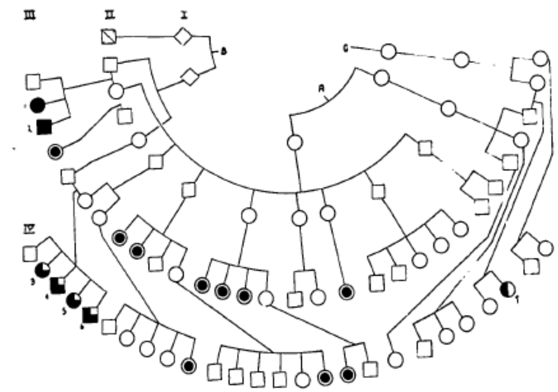


FIG. 1. *Family history*. Upright white squares and circles indicate normal men and women; slanted white squares indicate unknown sex. Black squares and circles indicate men with Klinefelter's syndrome and women with ovarian agenesis, both with cataract and muscular dystrophy. Black squares and circles with $\frac{1}{4}$ white area indicate men and women with muscular dystrophy, but without endocrine disorder. Black circle with $\frac{1}{2}$ white area indicates woman with muscular dystrophy and epicanthus. Black double circles indicate stillbirths.

Received for publication February 9, 1956.

* Present address: Medical Department B, University Clinic of Bergen, Bergen, Norway.

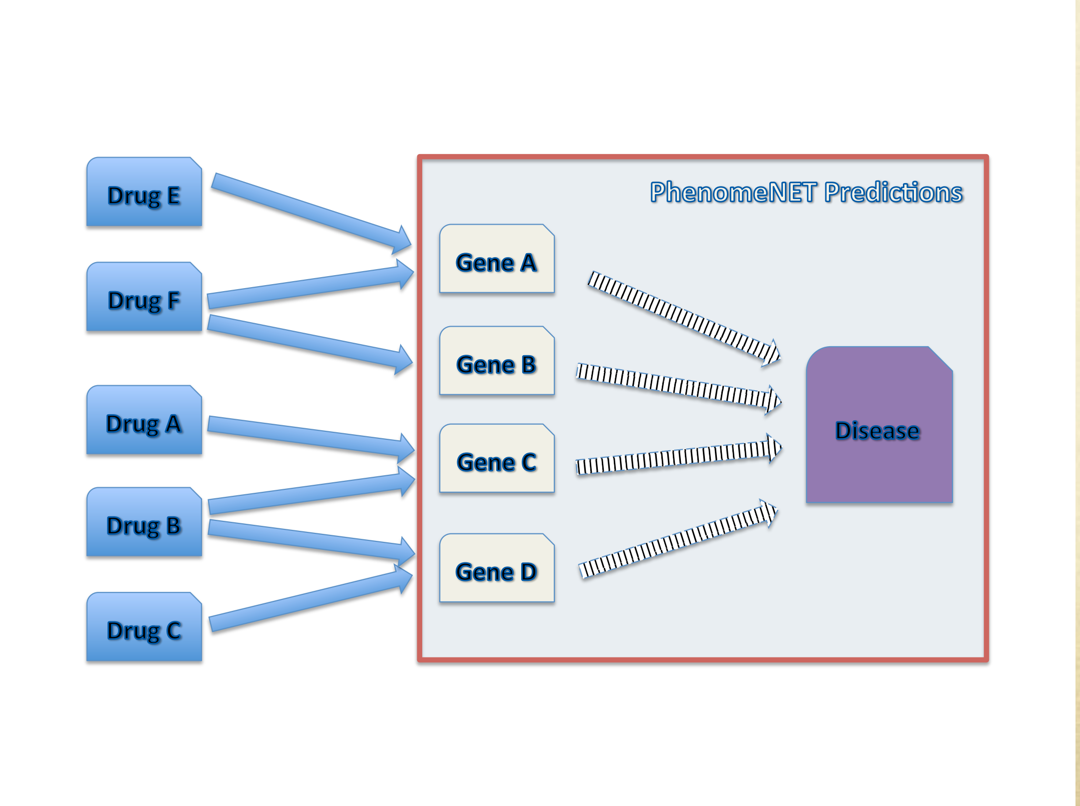
Clinical diagnostics decision support systems

- “show me all genes involved in degenerative processes of the brain or heart for which no evidence of cerebellar degeneration is available in mouse models”
- “show me all genes associated with a particular process that are also associated with mental retardation”
- “prioritize these genes in order with their relevance to a particular set of phenotypes or a particular syndrome”

Novel drug discovery and repurposing

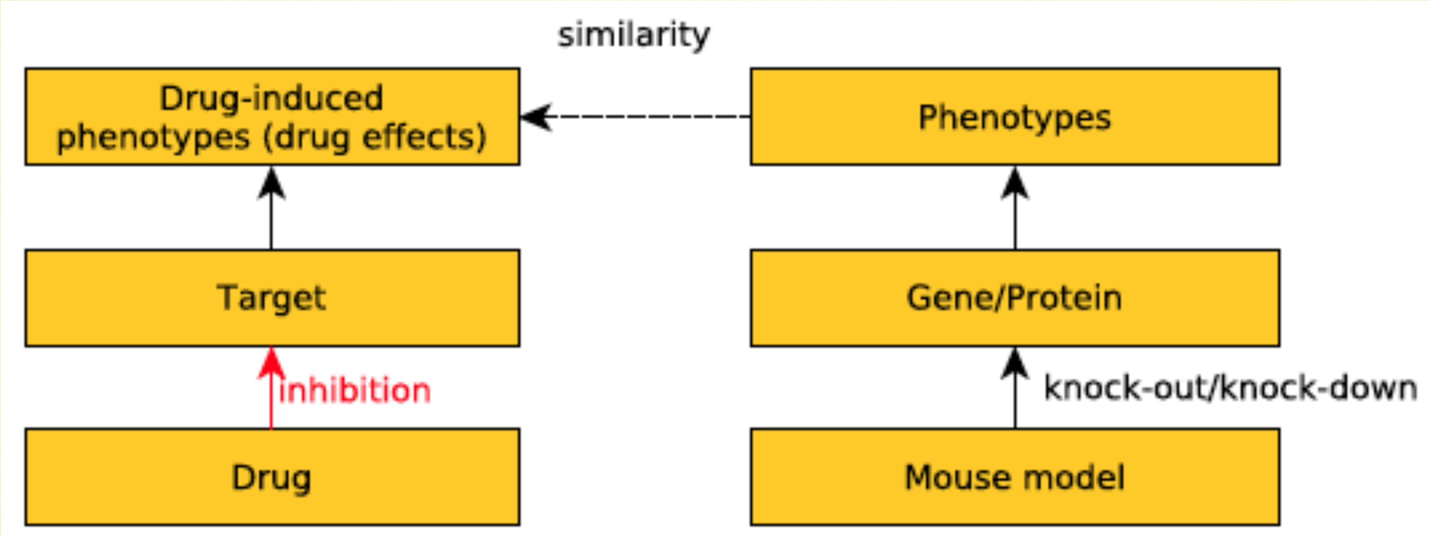
Variety of methods successfully being applied for drug repositioning and the suggestions of potentially novel drugs

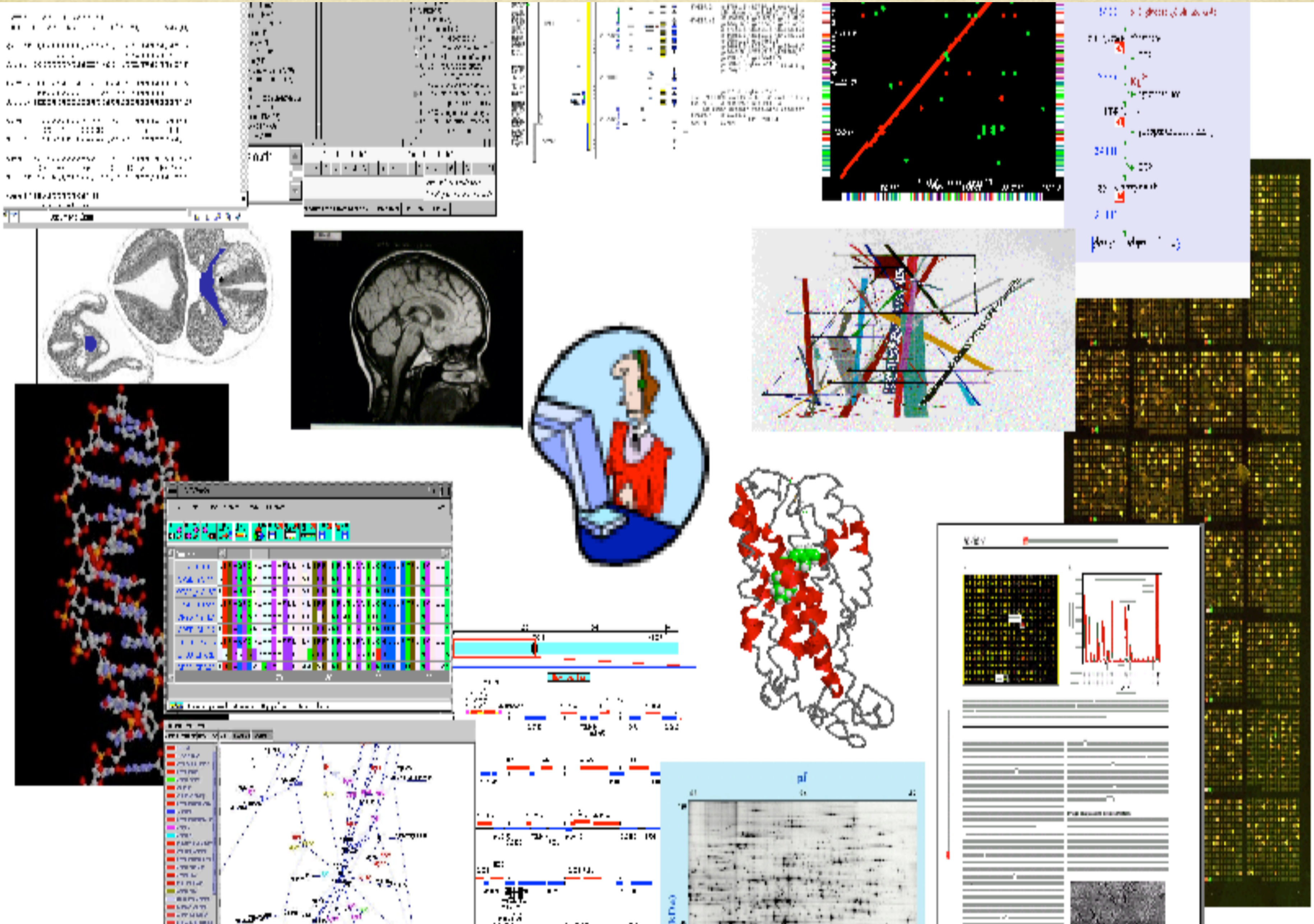
Can a phenotype of gene which the drug interacts be used to predict diseases in which the drug is active?



Drug targets and indications identification

A similarity between drug D's effects and phenotypes resulting from *knock-out/knock-down* of a gene/protein in an animal model may indicate that D *inhibits* the gene/protein (or its human ortholog).





What is big data?

- "Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

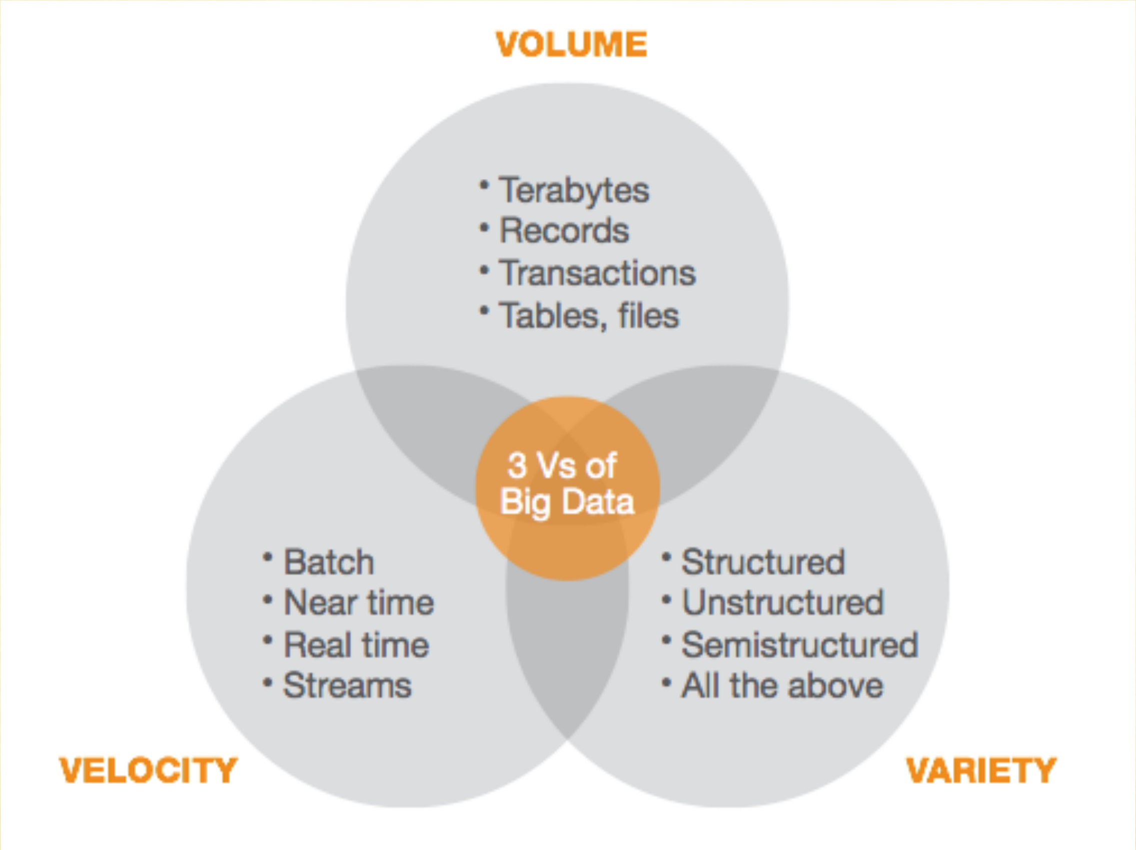
  This data is "big data."

Huge amount of data

- There are huge volumes of data in the world:
  - From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data.
  - In 2011, the same amount was created every two days
  - In 2013, the same amount of data is created every 10 minutes.

3 Vs of Big Data

- The "BIG" in big data isn't just about volume



How Is Big Data Different?

1) Automatically generated by a machine

(e.g. Sensor embedded in an engine)

2) Typically an entirely new source of data

(e.g. Use of the internet)

3) Not designed to be friendly

(e.g. Text streams)

4) May not have much value

- Need to focus on the important part

Examples

Healthcare

The average amount of data per hospital will increase from 167TB to 665TB in 2015, driven by the enormous growth of medical images and electronic medical records.[1]

With Big Data

Medical professionals can improve patient care and reduce costs by extracting relevant clinical information from vast amounts of data to better understand the past and predict future outcomes.

Customer Service

Today, 86% of consumers quit doing business with a company because of a bad customer experience, up from 59% four years ago.[2]

With Big Data

Service representatives can use data to gain a more holistic view of their customers, understanding their likes and dislikes in real-time in order to resolve a problem or capitalize on happy clients faster.

Insurance

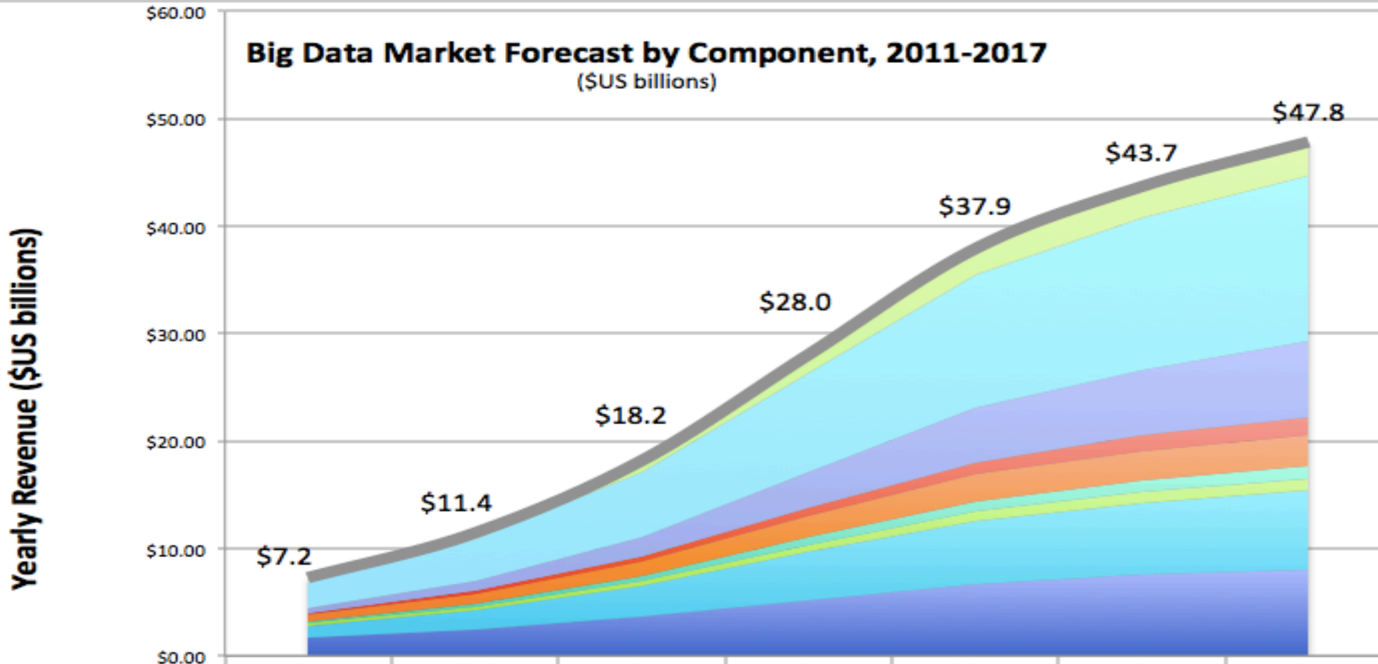
Insurance companies and government agencies each gather fraud data related to their own individual missions. But the kind, quality and volume of data compiled varies widely.[3]

With Big Data

An insurance or citizen services provider can apply advanced analytics to data and detect fraud quickly, before funds are paid out.

How are revenues looking like….

Wikibon

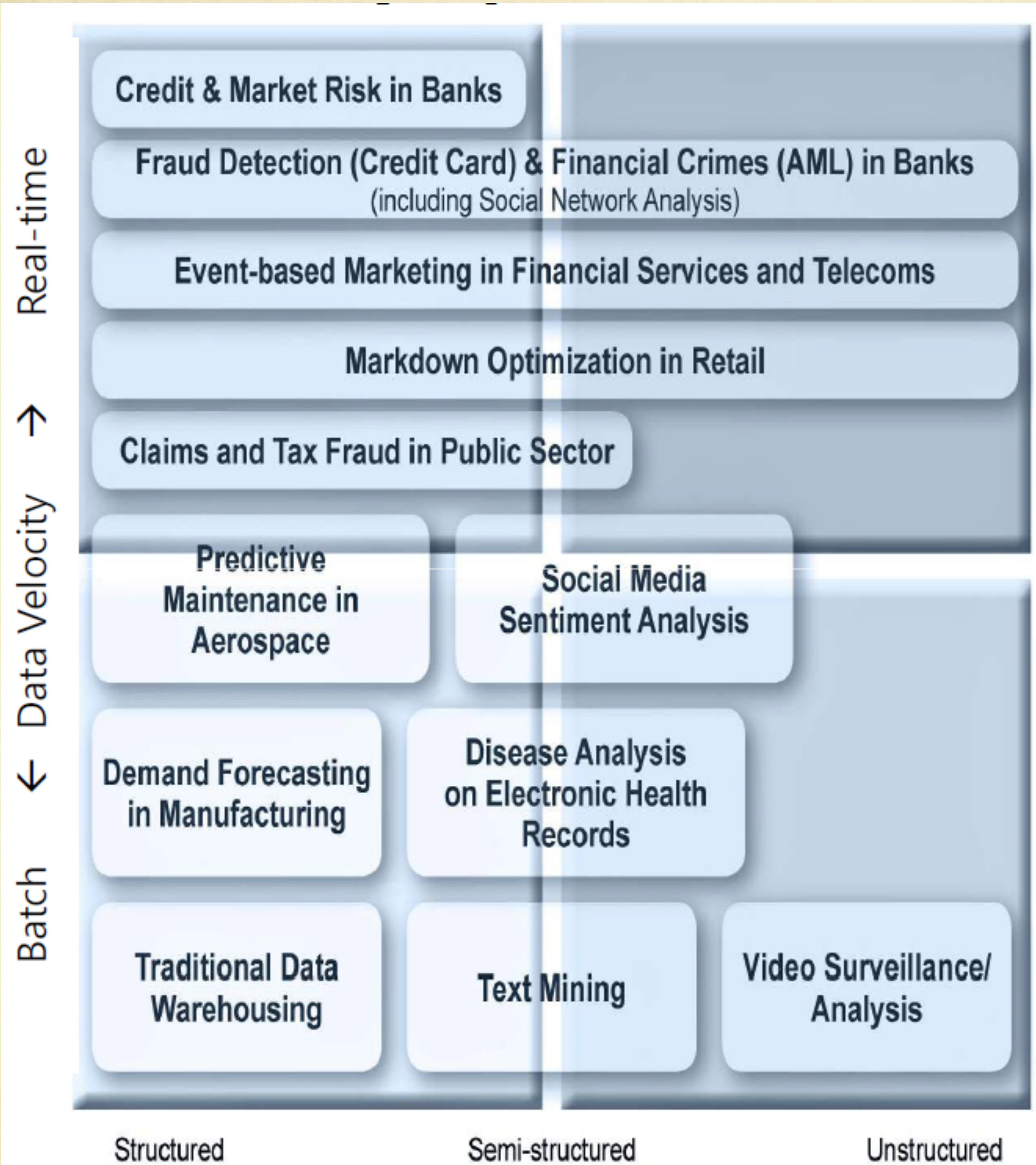


	2011	2012	2013	2014	2015	2016	2017
Big Data XaaS Revenue	$0.35	$0.61	$1.05	$1.74	$2.47	$2.91	$3.24
Big Data Professional Services Revenue	$2.45	$3.87	$6.10	$9.29	$12.37	$14.14	$15.38
Big Data Application (Analytic and Transactional) Software	$0.49	$0.94	$1.80	$3.29	$5.02	$6.15	$7.00
Big Data NoSQL Database Software	$0.10	$0.19	$0.39	$0.73	$1.14	$1.41	$1.62
Big Data SQL Database Software	$0.72	$1.02	$1.45	$1.99	$2.47	$2.73	$2.90
Big Data Infrastructure Software	$0.16	$0.26	$0.43	$0.70	$0.96	$1.12	$1.24
Big Data Networking Revenue	$0.18	$0.28	$0.44	$0.67	$0.89	$1.02	$1.11
Big Data Storage Revenue	$1.16	$1.83	$2.89	$4.40	$5.86	$6.70	$7.28
Big Data Compute Revenue	$1.64	$2.45	$3.64	$5.23	$6.70	$7.50	$8.06
Total Big Data Revenue	$7.2	$11.4	$18.2	$28.0	$37.9	$43.7	$47.8

Types of tools typically used in Big Data Scenario

- Where is the processing hosted?
  - Distributed server/cloud
- Where data is stored?
  - Distributed Storage (eg: Amazon s3)
- Where is the programming model?
  - Distributed processing (Map Reduce)
- How data is stored and indexed?
  - High performance schema database
- What operations are performed on the data?
  - Analytic/Semantic Processing (Eg. RDF/OWL)

The Structure of Big Data



- Structured
  - Most traditional data sources
- Semi-structured
  - Many sources of big data
- Unstructured
  - Image data, audio data etc

Structured data

- Structured data tends to refer to information in "tables"

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (for text) queries, e.g.,

Salary < *60000 AND Manager = Smith.*

Unstructured data

- Typically refers to free text
- Allows
  - Keyword queries including operators
  - More sophisticated "concept" queries e.g.,
    - find all web pages dealing with *drug abuse*
- Classic model for searching text documents

Semi-structured data

- In fact almost no data is "unstructured"
- E.g., this slide has distinctly identified zones such as the *Title* and *Bullets*
    - … to say nothing of linguistic structure
- Facilitates "semi-structured" search such as
    - *Title* contains data AND *Bullets* contain search
- Or even
    - *Title* is about Data AND *Author* something like stro*rup
    - where * is the wild-card operator

unstructured data

Informal system

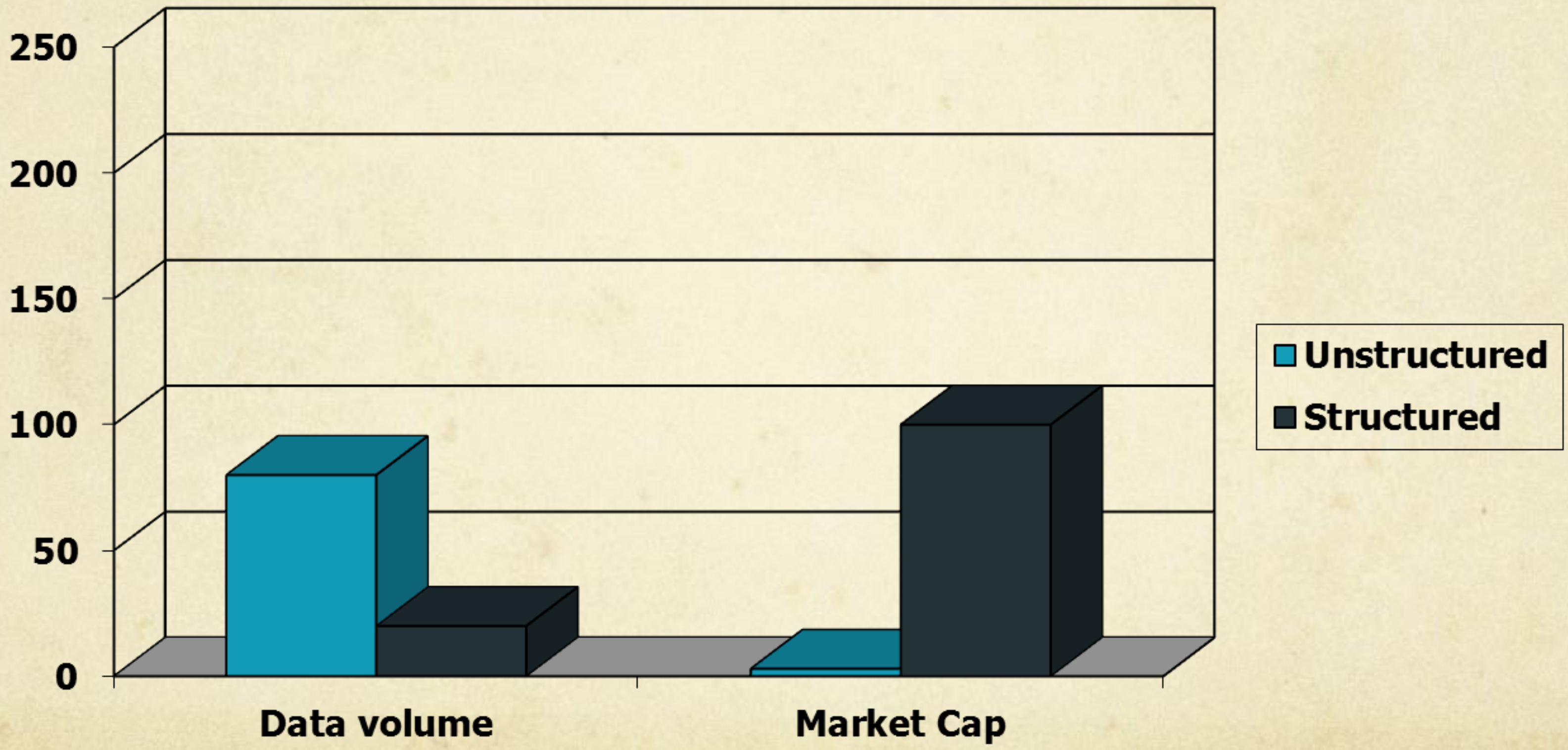
- .doc files
- .txt files
- email
- transcripted telephone
- Books/journal
- GP's note
- etc.

structured data

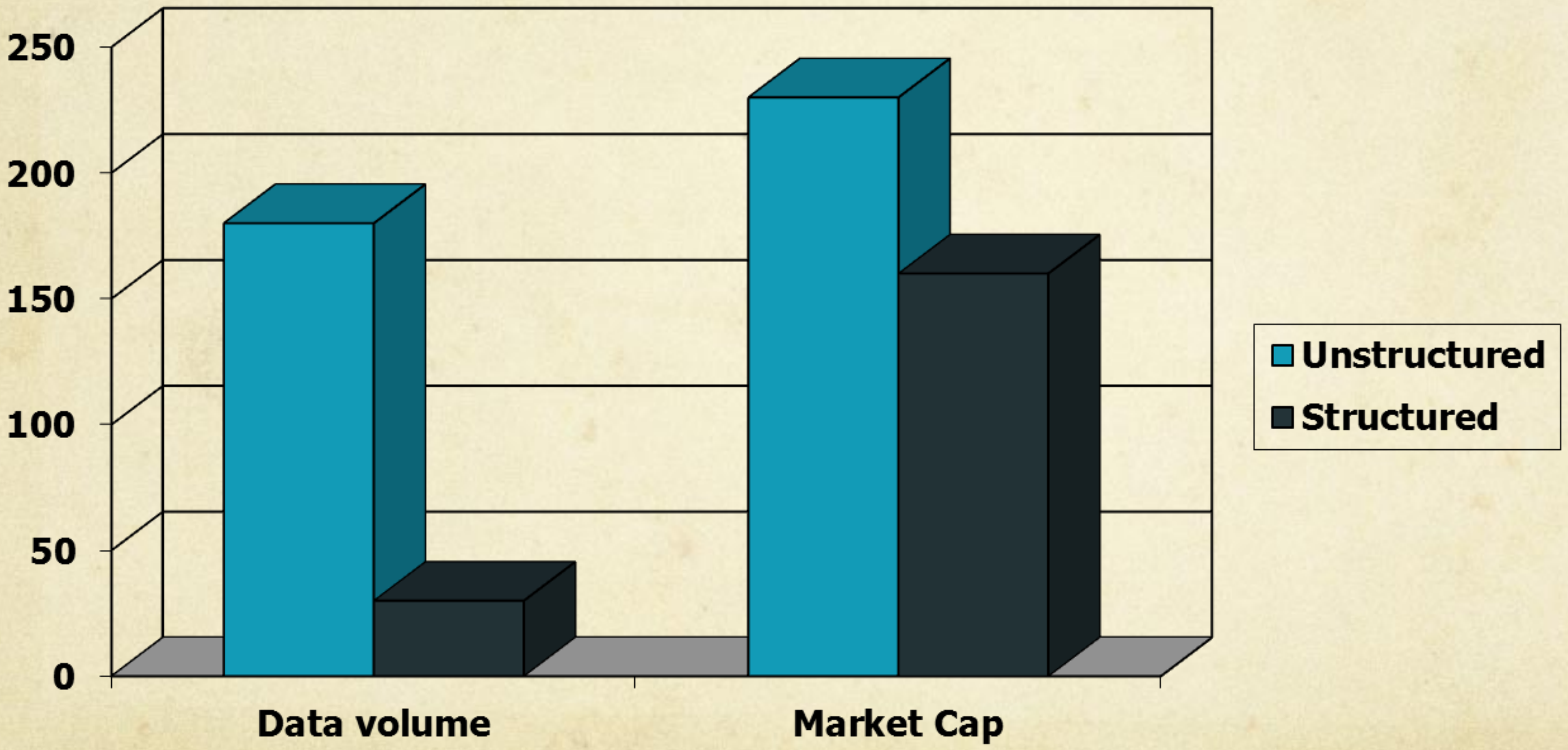
formal system

- corporate transactions
- corporate reports
- corporate databases
- customer files
- audit reports
- EHR
- etc

Unstructured (text) vs. structured (database) data in the mid-nineties

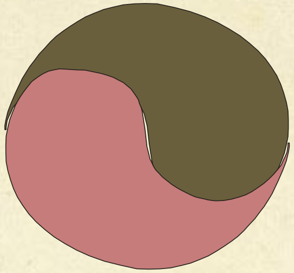


Unstructured (text) vs. structured (database) data today



Next big challenge in science

Unstructured data



structured data

imagine what would happen if the
two worlds could be integrated.......

The next frontier for innovation, competition, and productivity

It will revolutionize all sectors :

- healthcare
- public sector
- retail sector
- manufacturing
- etc.
- etc.

Unstructured data

structured data

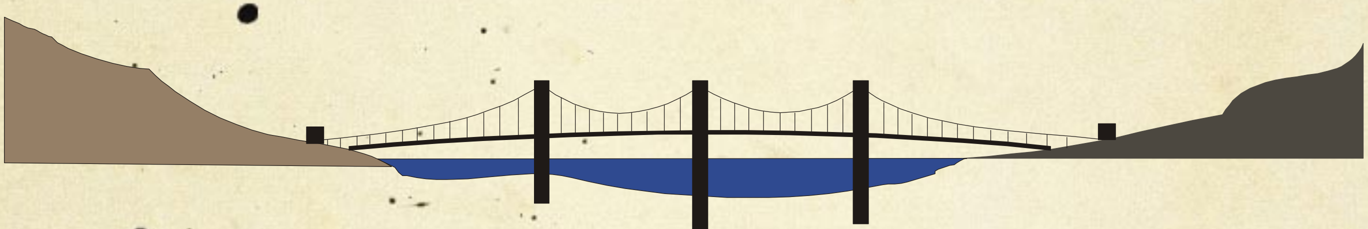
There is a gulf between the two worlds:

- technology
- business practice
- organizational
- historical

etc.

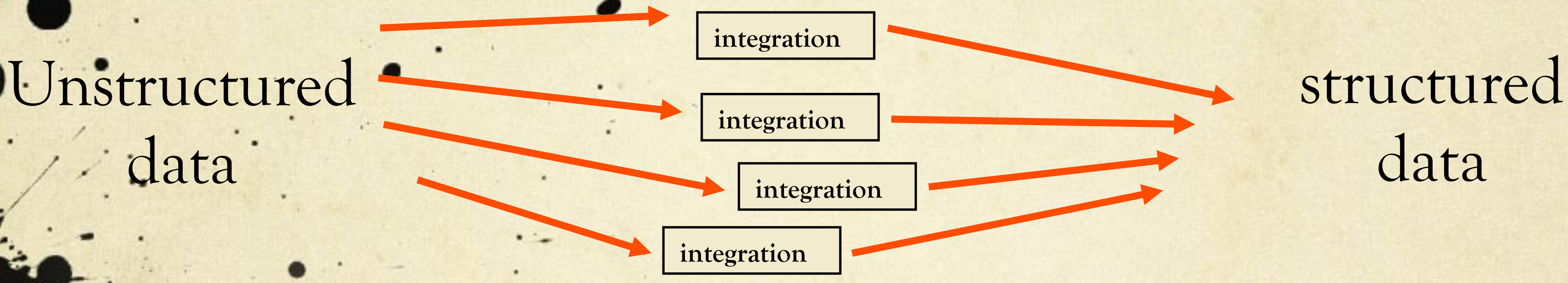
Unstructured data

structured data

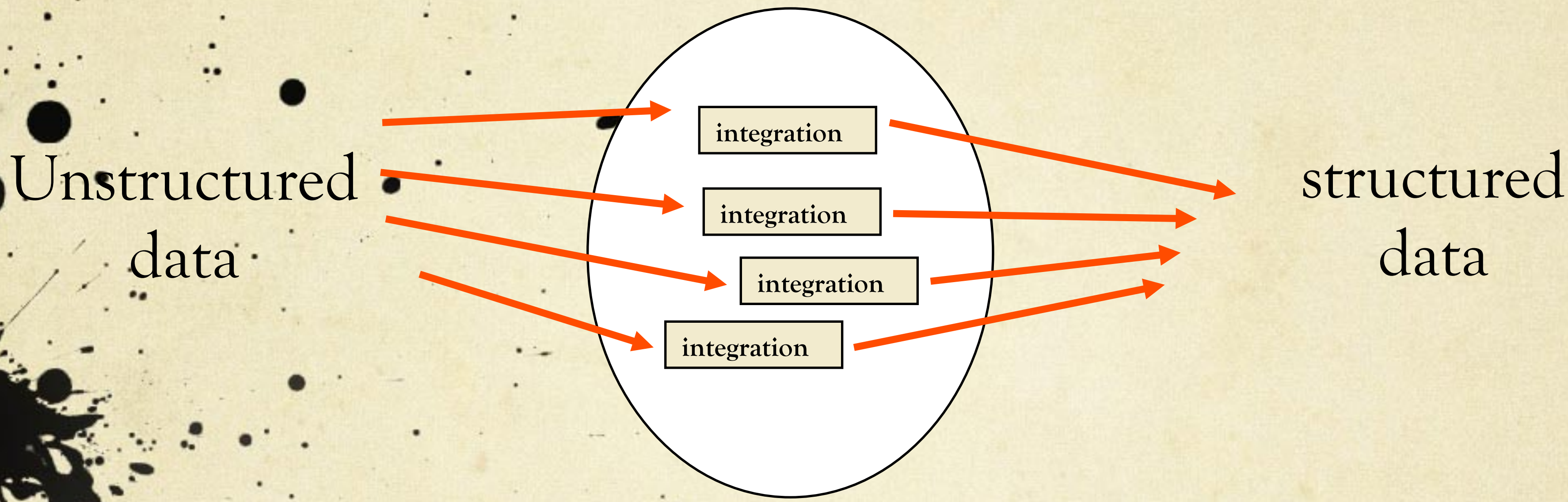


Think of the possibilities!x

Vision implementation

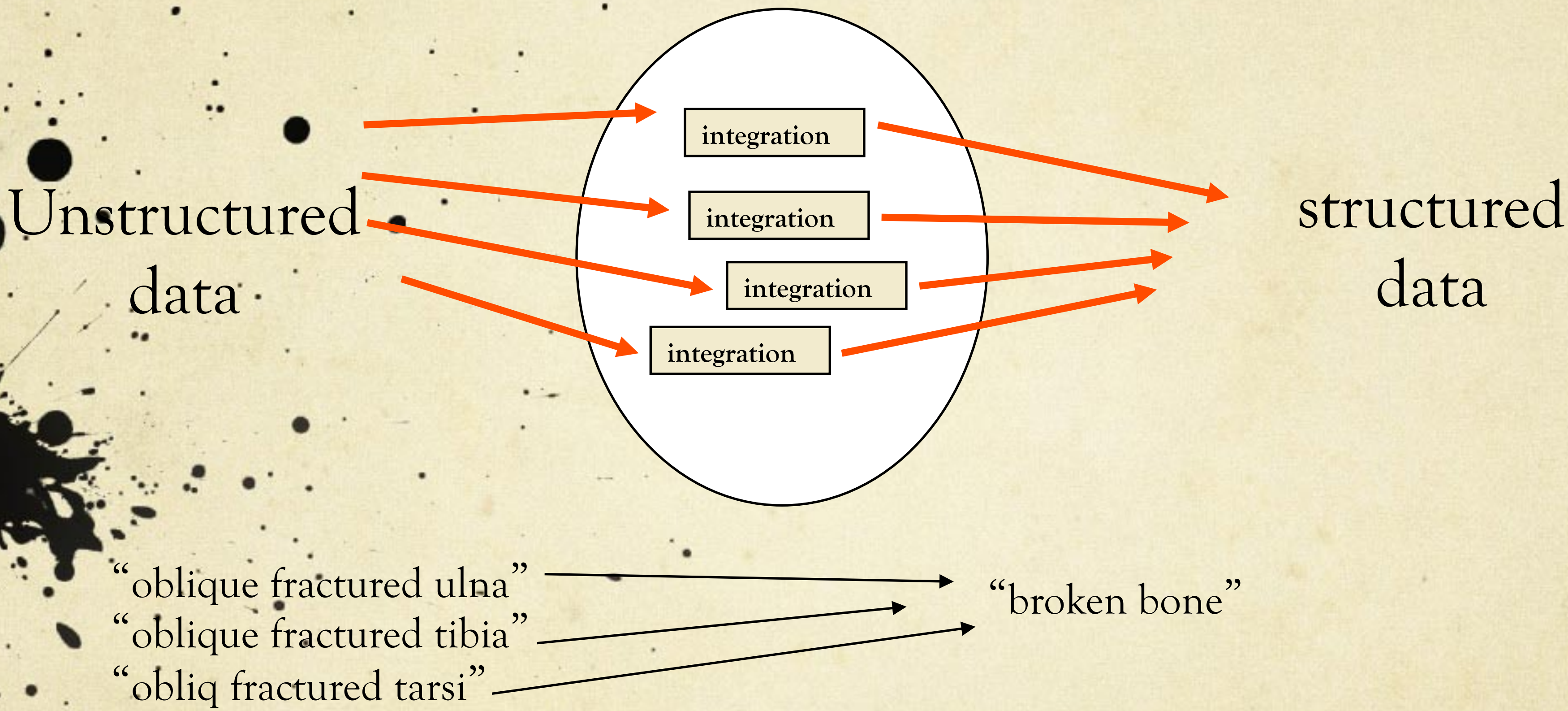


Challenge


Unstructured data
integration
integration
integration
integration
structured data

Not new....

- NLP (Natural Language Processing)
- Data scientists
- Impressive applications
  - E.g MapReduce (supercharged mapreduce)
- A lot of scientists, resources, enterprises, departments, consortia aim to achieve this goal


Unstructured data
integration
integration
integration
integration
structured data
"oblique fractured ulna"
"oblique fractured tibia"
"obliq fractured tarsi"
"broken bone"

Data and Variables

Data are often discussed in terms of variables, where a **variable** is:

> Any characteristic that *varies* from one member of a population to another.

A simple example is height in centimeters, which varies from person to person.

Types of Variables

There are two basic types of variables: *numerical* and *categorical* variables.

Numerical Variables: variables to which a number is assigned as a quantitative value.

Categorical Variables: variables defined by the classes or categories into which an individual member falls.

Ratio Scale

	SYSTOLIC			DIASTOLIC			5th PHASE
11.	1	2	0	1	0	0	First reading (R1)
12.		2	0		1	5	RZ1 (ranges from 0 - 40)
13.	1	0	0		8	5	First corrected (R1 - RZ1)
14.	2	0	0	1	2	0	Second reading (R2)

Categorical Variables

Defined by the classes or categories into which an individual member falls.

- **Nominal Scale:** Name only--Gender, hair color, ethnicity

- **Ordinal Scale:** Nominal categories with an implied order--Low, medium, high.

Nominal Scale

b. Appearance of plasma:	b.
1. Clear..........................	1.
2. Turbid.......................	2.
9. Not done....................	9.

Ordinal Scale

81.Urine protein (dipstick reading):	81.
1. Negative…………………	1.
2. Trace……………………	2.
3. 30 mg% or +……………	3.
4. 100 mg% or ++…………	4.
5. 300 mg% or +++…………	5.
6. 1000 mg% or ++++………	6.
If urine protein is 3+ or above, be sure subject gets a 24 hour urine collection container and instruction	

Disease Stage

Disease progression/stage
Chronic Kidney Disease (CDK)

1	2	3	4	5
GFR 90+	GFR 60-89	GFR 45-59	GFR 15-29	GFR <15 or in dialysis

GFR (glomerular filtration rate)

Datasets and Data Tables

Dataset: Data for a group of variables

Data Table: A dataset organized into a table, with one column for each variable

Typical Data Table

OBS	AGE	BMI	FFNUM	TEMP(^{0}F)	GENDER	EXERCISE LEVEL	QUESTION
1	26	23.2	0	61.0	0	1	1
2	30	30.2	9	65.5	1	3	2
3	32	28.9	17	59.6	1	3	4
4	37	22.4	1	68.4	1	2	3
5	33	25.5	7	64.5	0	3	5
6	29	22.3	1	70.2	0	2	2
7	32	23.0	0	67.3	0	1	1
8	33	26.3	1	72.8	0	3	1
9	32	22.2	3	71.5	0	1	4
10	33	29.1	5	63.2	1	1	4
11	26	20.8	2	69.1	0	1	3
12	34	20.9	4	73.6	0	2	3
13	31	36.3	1	66.3	0	2	5
14	31	36.4	0	66.9	1	1	5
15	27	28.6	2	70.2	1	2	2
16	36	27.5	2	68.5	1	3	3
17	35	25.6	143	67.8	1	3	4
18	31	21.2	11	70.7	1	1	2
19	36	22.7	8	69.8	0	2	1
20	33	28.1	3	67.8	0	2	1

Definitions for Variables

- AGE: Age in years
- BMI: Body mass index, weight/height^2 in kg/m^2
- FFNUM: The average number of times eating "fast food" in a week
- TEMP: High temperature for the day
- GENDER: 1- Female 0- Male
- EXERCISE LEVEL: 1- Low 2- Medium 3- High

CDK

Stage	GFR*	Description	Treatment stage
1	**90+**	Normal kidney function but urine findings or structural abnormalities or genetic trait point to kidney disease	Observation, control of blood pressure. More on management of Stages 1 and 2 CKD.
2	**60-89**	Mildly reduced kidney function, and other findings (as for stage 1) point to kidney disease	Observation, control of blood pressure and risk factors. More on management of Stages 1 and 2 CKD.
3A **3B**	**45-59** **30-44**	Moderately reduced kidney function	Observation, control of blood pressure and risk factors. More on management of Stage 3 CKD.
4	**15-29**	Severely reduced kidney function	Planning for endstage renal failure. More on management of Stages 4 and 5 CKD.
5	**<15 or on dialysis**	Very severe, or **endstage** kidney failure (sometimes call **established renal failure**)	Treatment choices. More on management of Stages 4 and 5 CKD.

Is structured data enough ?

Phenotypic information captured differently within the same domain (OMIM)

Query	**# of records**
“large bone”	713
"enlarged bone"	136
"big bones"	16
"huge bones"	4
"massive bones"	28
"hyperplastic bones"	8
"hyperplastic bone"	34
"bone hyperplasia"	122
"increased bone growth"	543

The Need for Standards

- Become more structured over time
- Fine-tune to be friendlier for analysis
- Standardize enough to make life much easier

Facilitate interoperability

Interoperability – the great challenge

Semantic resource interoperability

↓

Interpretation of the meaning of data

Interoperability

ability of two or more systems or components to **exchange information** and to **use the information** that has been exchanged

Source: IEEE Standard Computer Dictionary, 1990

exchange information → syntactic interoperability (e.g. XML)

use the information → semantic interoperability (e.g. metadata)

How interoperability is achieved ?

resources need to be able to:

- exchange data and services in a consistent and effective way
- provide universal access capacities independent of platforms

focus on data visibility and accessibility and enable flexibility in data exchange

Central goal:

- Increase the amount of *useful* data available
- Ensure that data is understandable

Standardization – a vital ingredient of interoperability

standardization is the process to obtain and apply a set of rules and agreements in order to create clarity and unity in areas where diversity is unwanted

(Aalders, 1998)

developing and agreeing upon documents that establish uniform engineering or technical specifications, criteria, methods, processes, or practices.

vital for achieving interoperability, interchangeability and functionality.

Type of Standards

- Two types: "de jure" & "de facto"
  - De jure standards that are approved or endorsed by an authoritative body
  - De facto standards are what everybody uses since they have achieved a dominant position, by tradition, enforcement, or market dominance
- Frequently defined in a form of specification by a Standards Development Organisation
- "Open" to "Proprietary"

Examples of different kind of standards

- Data exchange standards (e.g. XML)
- Data security standards (e.g. digital signatures)
- Data representation standards (e.g. RDF)
- Information management standards (e.g. EHRs)
- Terminology standards (e.g. medical terminologies)

Example of standards in the biomedical domain

SDO	Organizational Title
ANSI	American National Standards Institute
HL7	Health Level 7
ASTM	American Society for Testing and Materials
HISB	Healthcare Informatics Standards Board

Example of standards in the biomedical domain

SDO	Area
SNOMED	Standard Nomenclature for Medicine
DICOM	Digital Imaging and Communications in Medicine
NCPDP	National Council for Prescription Drug Programs
MedDRA	Medical Directory of Regulatory Affairs

Biomedical Terminology and interoperability

Interoperability requires means of standardizing the **encoding and semantic representation** data for exchange, comparison or aggregation among systems.

- biomedical terminologies are systematic representation of terms with the goal of enabling information exchange
- a prerequisite for interoperability

What is “terminology” ?

“The lexicon of a special subject language reflects the organisational characteristics of the discipline by tending to provide as many lexical units as there are concepts...”

Juan C. Sager

Terminologies define items which are characterised by **special reference within a specific discipline** whilst a vocabulary defines items that are characterised by **general reference in a language system**

Related Terms

- Nomenclature
  - A system of terms which is elaborated according to pre-established naming rules as used by a community. For example a nomenclature of plants, chemicals, animals etc
- Coding System
  - a combination of: a system of concepts; a terminology; a set of code values; at least one coding scheme to relate the codes to the concepts or the terms

- Thesaurus
  - A controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge
- Taxonomy
  - A terminological system whose system of concepts is structured by generic relations only, in other words a classification typically organised by subtype-supertype relationships

Ideal Terminology

An Ideal Terminology should be:

Complete, Formal, Universal, Translatable

- Completeness
  - cover as much of the domain of interest as possible
- Boundary
  - allow the growth within the domain of choice in order to strive to achieve completeness.
- Organization
  - defined relations between terms, offering related terms, synonyms etc
- Absence of ambiguity
  - all terms being not only textual but also semantically defined

Evolution of biomedical terminology systems

- First Generation
  - paper-based based of information
  - simple hierarchies and organization
  - lack any computational support → quite expensive to maintain and reuse.
- Second Generation
  - compositional systems
  - *categorial* structure with semantic links
  - limited semantic based processing
- Third Generation
  - formal models with dynamic inferred hierarchies
  - semantic based processing

Biomedical Terminologies Categories

Intended application

- Classification (e.g. disease, drug, epidemiology)
- Billing, Auditing and Reimbursement
- Phenotype (symptoms, diseases, lab results, progress etc)
- Reference - linking different kinds of terminologies
- Etc.

Biomedical Terminologies Categories

area of coverage

- diseases
- drugs
- medical equipment
- surgical procedures
- different domains (e.g. anatomy, pathology etc)

Biomedical Terminologies Categories

technical approach

- pre-coordinated (precomposed, enumerative)
  - Exhaustive classification
- post-coordinated (postcomposed, compositional)
  - minimum of pre-defined terms which are combined to created more complicated terms.

- Lexical (a number of techniques, including natural language processing, for mapping terms to natural language, free text, literature etc.)

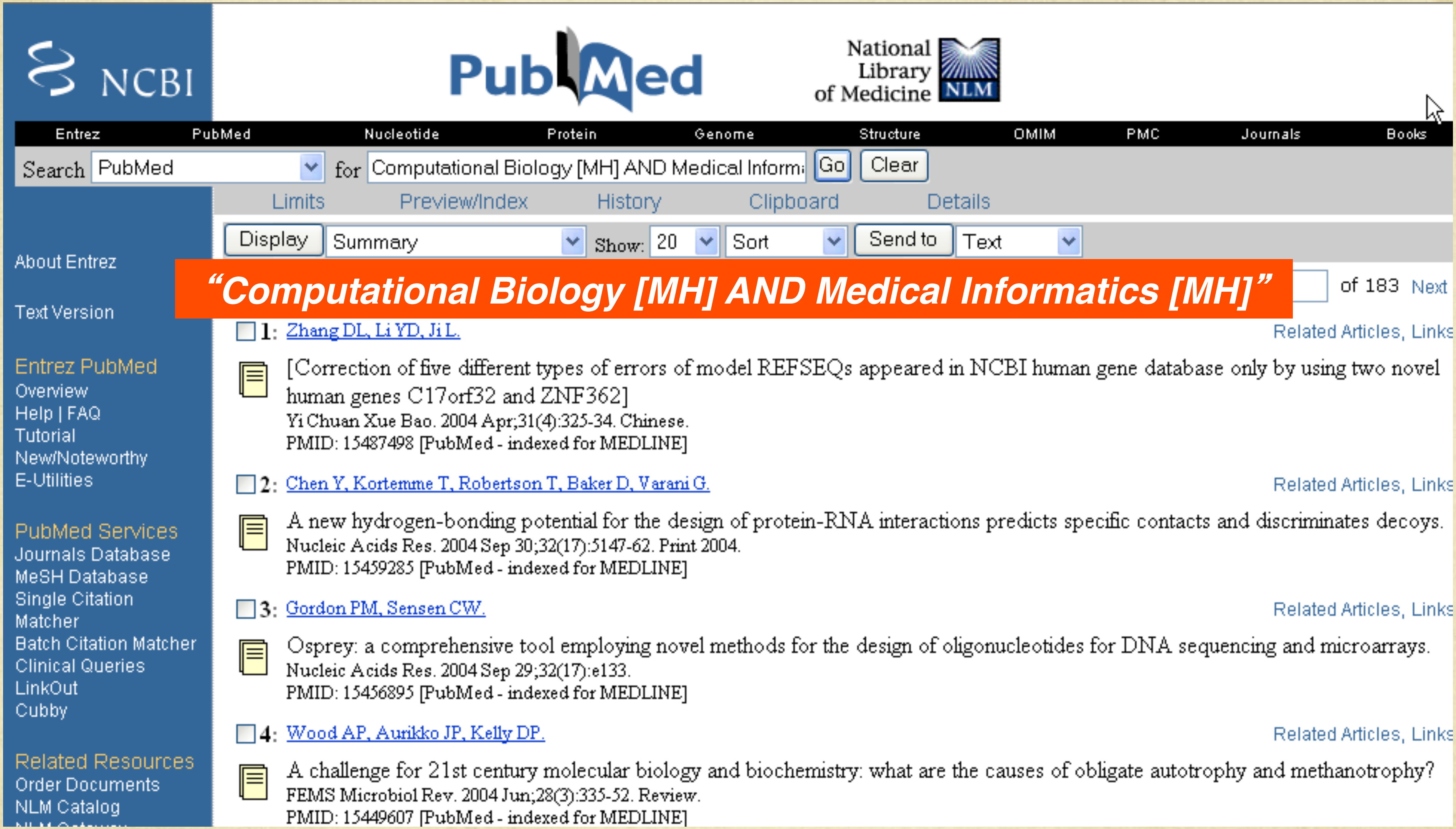
Some examples

(pre-coordinated, post-coordinated and lexical)

Medical Subject Headings (MeSH)

- from *Index Medicus* to MeSH (NLM)
- purpose is to index medical literature
- used for MEDLINE/PubMed
- gained wide acceptance and adopted for a wide range of applications
- widely used by both health sciences libraries and abstracting and indexing services in the health sciences.

“Computational Biology [MH] AND Medical Informatics [MH]”



MeSH structure

- 19,000 MeSH Heading are arranged in a hierarchy of about 15 categories (for example anatomy, organism, diseases etc.)
- MeSH Headings have unique IDs (multiple synonyms etc.)
- MeSH Headings may have multiple locations in multiple trees, known as multiple inheritance.
- Standard qualifiers
- Example: “saliva secretion”

International Classification of Diseases (ICD)

- ICD published by WHO provides a classifcation of disease and other health problems
- enables the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes
- provides the basis for the compilation of national mortality and morbidity statistics by WHO Member States.
- insurance, statistics and epidemiology

ICD structure

- Divided into categories based on 5-digit numeric code
- code is both the concept and the unique identifier, whilst the last 2 digits are called modifiers
- Multiple terms linked to the same code
- patients are "coded" with as many terms as possible

ICD limitations

Unsuitable for medical information interoperability and for medical knowledge representation

'V32.22' Occupant of three-wheeled motor vehicle injured in collision with two- or three-wheeled motor vehicle, person on outside of vehicle, nontraffic accident, while working for income

Current Procedural Terminology (CPT)

- published by the American Medical Association (AMA) in 1998
- developed as a method of communication between medical personnel
- CPT describes a uniform language that accurately describes medical, surgical, and diagnostic services
- Intended to be used for insurance, and reimbursement purposes

CPT structure

- Standard terms and descriptors using a 5-digit classification system
- codes are grouped by medical specialty (i.e., surgery, medicine, etc.). A
- Every code includes a therapeutic function (e.g. thermal stimulation) and optionally a time (e.g. assessment of aphasia, per hour) and body part component (e.g. thermal stimulation of a particular muscle group)

From pre-coordinated to post-coordinated medical terminologies

- expressivity and domain cover → demand for new terms
- poor biomedical knowledge representation
- Combinatorial explosion

SNOMED (Systematized Nomenclature of Medicine)

- Systematized Nomenclature of Pathology (SNOP) → SNOMED
- standardize clinical information and interoperability by providing a common language of sufficient capturing, sharing and aggregating health data across clinical specialties and sites of care.
- SNOMED-RT (Reference Terminology)
- SNOMED-CT (Clinical terminology)

SNOMED-CT (Clinical Terms)

- systematically organized computer processable collection of medical terminology for most areas of clinical information, including diseases, observation, procedures etc.
- SNOMED Reference Terminology (SNOMED RT + Clinical Terms Version 3 (Read Codes)
- divided into 11 hierarchies such as Clinical findings, Procedures, Observable entities, Body structure and so on.

SNOMED-CT Structure

- Concept
  - Unique identifier (ConceptID)
  - Fully Specified Name (FSN), Preferred Term or synonym (DescriptionID)
- Relationships
  - Four types (Defining, Qualifying, Historical, Additional)
  - formally defined in terms of their relationships with other concepts

GALEN (Generalised Architecture for Language Encyclopaedias and Nomenclature in Medicine)

- reconcile diversity of needs for terminology & information sharing
- avoid costs for harmonisation of variants
- facilitate clinical applications
- detail vs abstraction
- multilanguage systems
- GALEN -- > OpenGALEN

Hybrid Medical Terminologies

- Logical Observation Identifiers, Names and Codes (LOINC)
  - around 7000 universal codes and names to identify laboratory and other clinical observations
  - collaboration with SNOMED-CT
- International Classification of Nursing Procedures (ICNP)
  - started as pre-coordinated terminology of nursery practice but now forms a unified nursing language system that facilitates the development of and the cross-mapping among local terms and existing terminologies

post-coordinated medical terminologies problems

- meaningless terms
- redundancy
- classification
- intractability

Unified Medical Language System (UMLS)

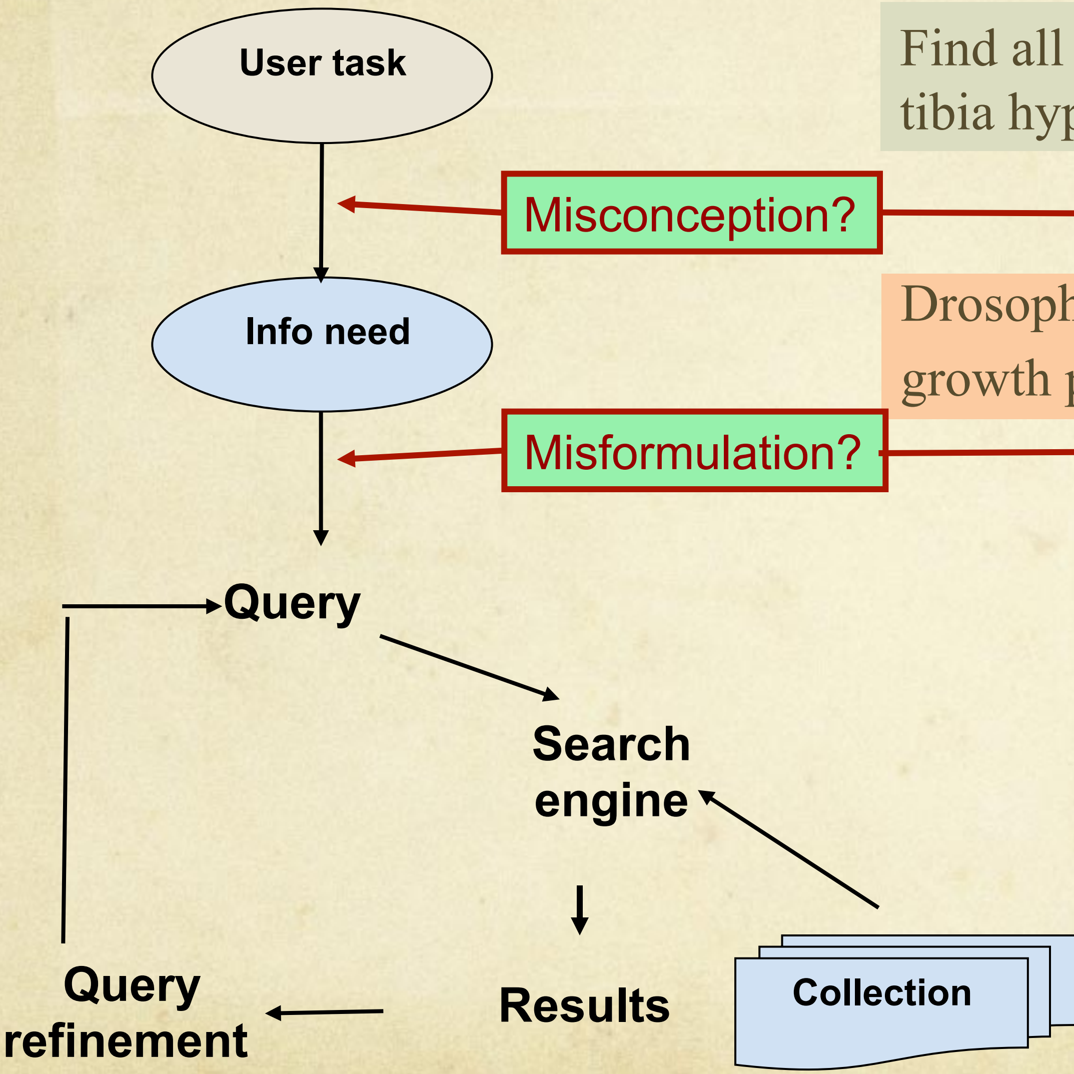
- computer systems that "understand" the meaning of the language of biomedicine and health
- Metathesaurus
  - vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them
- Semantic Network
  - consistent categorization of all concepts& relations represented in the UMLS Metathesaurus
- SPECIALIST Lexicon

Challenges

"… inadequate due to lack of content coverage at the desired level of granularity, lack of consistent meanings for concepts and their relationships, and lack of explicit, formal concept-representation principles"

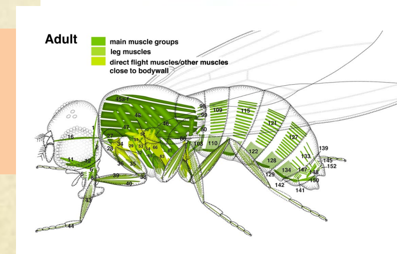
Cho, I., Park, H. (2003).

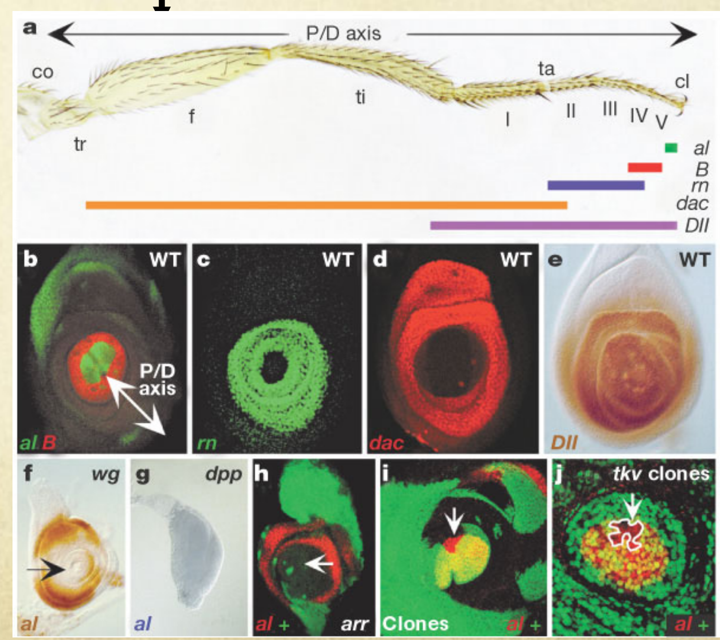
The classic search model



Find all genes related to tibia hypoplasia processes

Drosophila genes related to growth processes





Defining Disease – are terminologies enough ?

- pathological manifestation of the tissue response to an underlying lesion or set of lesions
- genetic lesion either somatic or inherited and subject to genetic background

Disease description

Description of pathology as part of disease representation

- Anatomical manifestation
- Physiological manifestation

Disease description challenge

a detailed description of patho-anatomical and patho-physiological manifestations based on a consistent representation model for genotypic and phenotypic information

Sufficiency of biomedical terminologies

- Useful for classification and information retrieval
- Lack of meaningful relationships between the terms for logical reasoning or inference