Introduction to ‘Omics’ and Big Data

jodie.lord@kcl.ac.uk

@JodieLord5

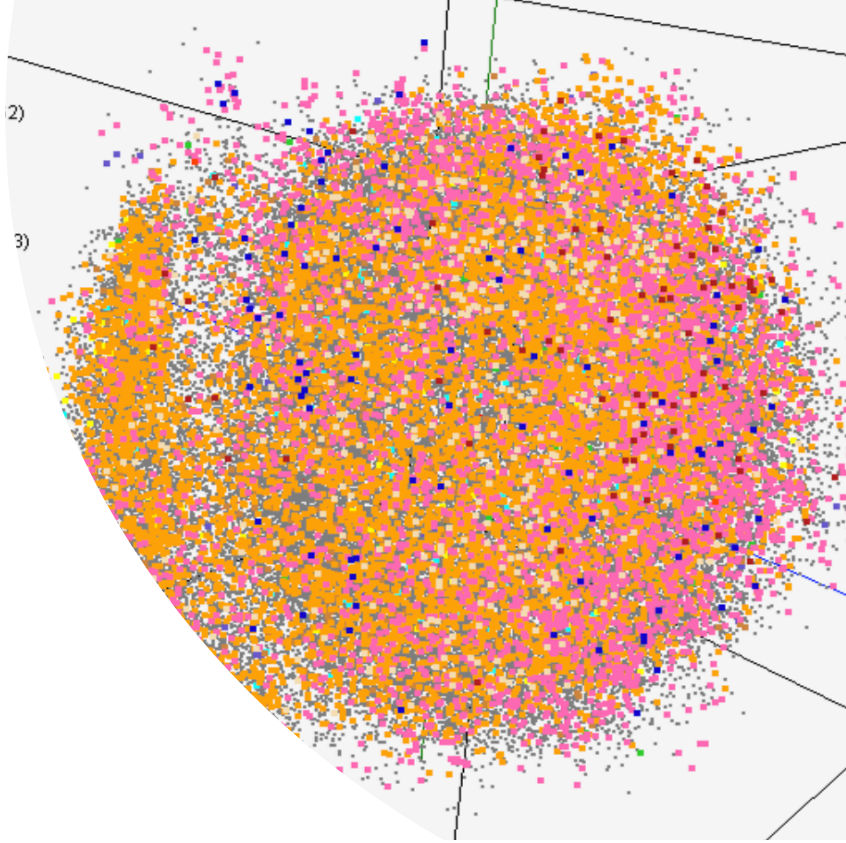
Learning Objectives

- Understand the concept of “high dimensional data” (HDD)
- Understand the concept of “omics”.
- Understand where “omics” fit in the realms of HDD
- Grasp possibilities and challenges that accompany HDD

Learning Objectives

- Understand the concept of “high dimensional data” (HDD)
- Understand the concept of “omics”.
- Understand where “omics” fit in the realms of HDD
- Grasp possibilities and challenges that accompany HDD

High Dimensional Data - What is it?

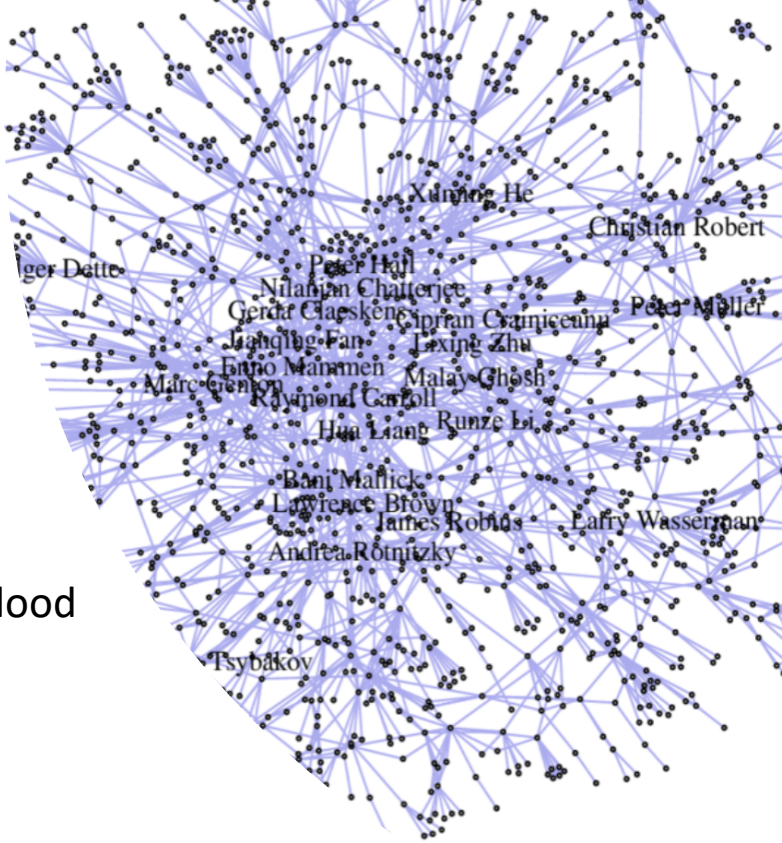


- HDD = another name for “Big Data”.
- Typically tens of / hundreds of thousands / millions of samples
- Attributes (variables) >> samples
- Variables often interlinked / correlated.
- Require large computational effort and typically sophisticated algorithms to make sense of

High Dimensional Data – Where did it come from?

- Exponential advancements in technology now allow us to collect information on a mass scale.
- “Super computers” and “super servers” allow us to store information on a grand scale.
- The computing era now means (nearly) every click, purchase, message, event and step we take is traceable, trackable, and consequently, analyzable.

Some examples of High Dimensional Data

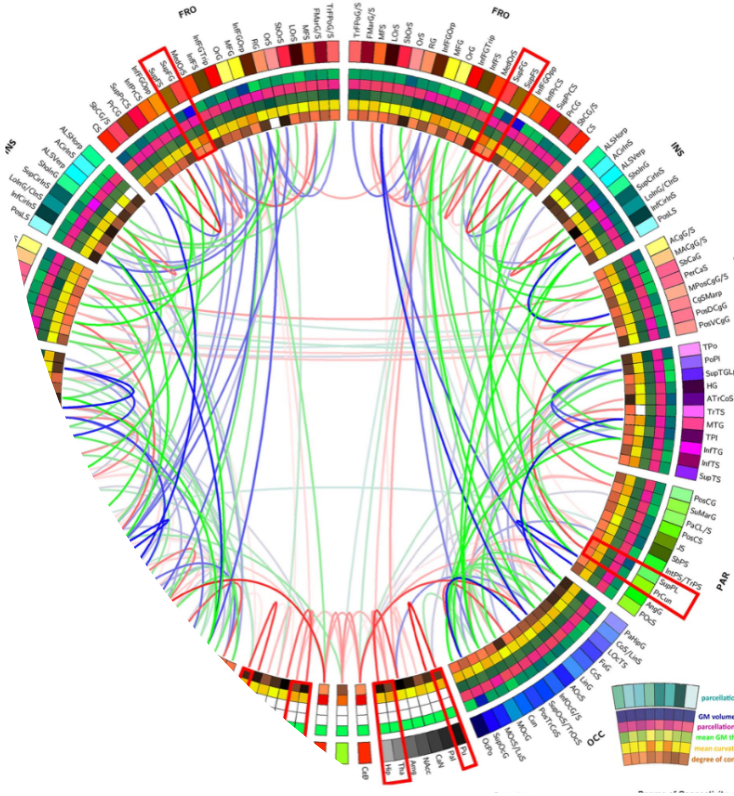


- **Dense medical records**
  (hundreds-thousands patients, hundreds – thousands health related variables (BMI, blood pressure, allergies, alcohol intake etc).
  → How do we use this to better target patient treatment?

- **Web activity logs**
  (thousands-millions(?) of web history hits).
  → How do we use this for advertisement / customer targeting / understanding online behavior?

- **Financial data**
  (hundreds of thousands of spenders, hundreds-thousands spending related data (savings, investments, shopping habits, etc).
  → How do we use this to monitor risk / forecast future spending habits / explain current state of market?

Some examples of High Dimensional Data

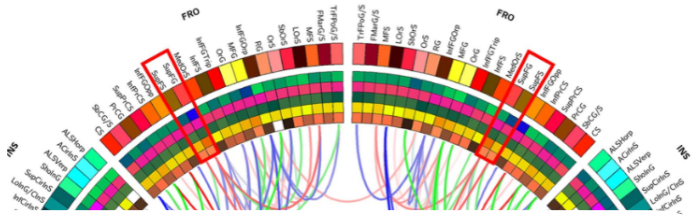


- **Information across the genome**
  (>25,000 genes, $3x10^9$ single basepairs!!)
  → How do we make sense of this information to inform genetic risk?

- **Information across all of the proteins within the human body**
  (>100,000 proteins each with different "signaling" jobs, and which interact with each other, impact each other, have different downstream effects).
  → How do we use this information to inform treatment, health behaviours, biological risk?

- **Information across all biological systems within the human body**
  Genes, proteins, metabolites, transcripts, methylation etc etc
  → How do these influence / interact with each other. What is the order of relationship? How large is the risk / what influences this risk / what mediation processes are involved / any suppressors etc etc etc

Some examples of High Dimensional Data



- **Information across the genome**
  (>25,000 genes, $3x10^9$ single basepairs!!)
  → How do we make sense of this information to inform genetic risk?

- **Information across all of the proteins within the human body**
  (>100,000 proteins each with different "signaling" jobs, and which interact with each other, impact each other, have different downstream effects).
  → How do we use this information to inform treatment, health behaviours, biological risk?

- **Information across all biological systems within the human body**
  Genes, proteins, metabolites, transcripts, methylation etc etc
  → How do these influence / interact with each other. What is the order of relationship? How large is the risk / what influences this risk / what mediation processes are involved / any suppressors etc etc etc

OMICS!!

Some examples of High Dimensional Data

- **Dense medical records**
  (hundreds-thousands patients, hundreds
  related variables (BMI, blood press
  → How do we use this to be

- **Wel**

- ndreds-thousands spending
  , shopping habits, etc).
  → onitor risk / forecast future spending
  ha ent state of market?

HIGH DIMENSIONAL DATA ≠ OMICS

Some examples of High Dimensional Data

BUT OMICS = HIGH DEMENSIONAL DATA!

- **Information across the genome**
  (>25,000 genes, 3x10^9 single basepair
  → How do we make sense of

  ach

  behaviours, biological

- **In iological systems within the human body**
  Ge metabolites, transcripts, methylation etc etc
  → H do these influence / interact with each other. What is the order of relationship? How large is the risk / what influences this risk / what mediation processes are involved / any suppressors etc etc etc

Learning Objectives

- Understand the concept of “high dimensional data” (HDD)

- Understand the concept of “omics”.

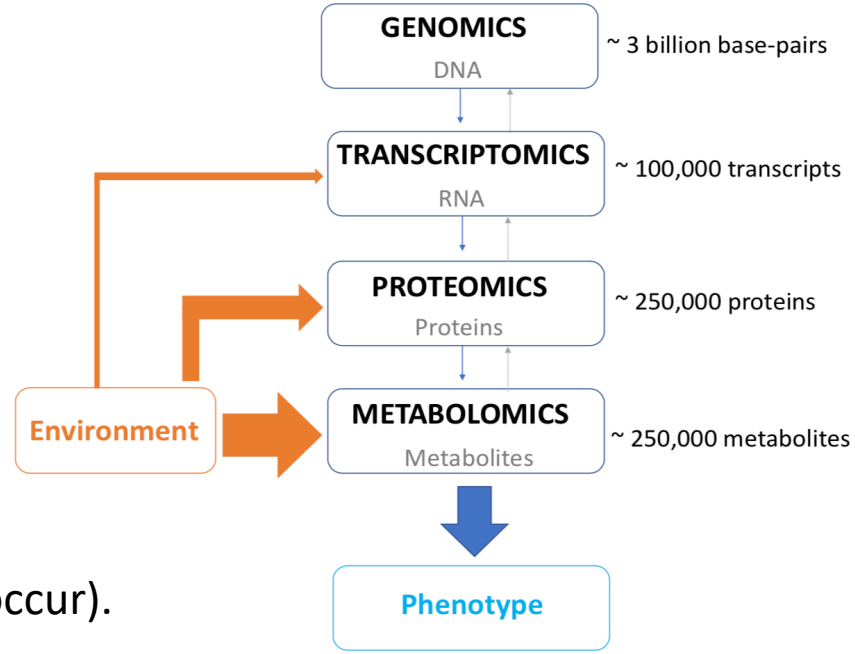
- Understand where “omics” fit in the realms of HDD

- Grasp possibilities and challenges that accompany HDD

What is Omics?

- **The science of trying to capture a biological system(s) in its entirety.**

- '**Omics**" fields are split across different biological strata, e.g:
  - Gen**omics** = the study of all the genes across the genome.
  - Epigen**omics** = the study of all reversable epigenetic modification points throughout the genome (which impact how much a gene is or is not expressed).
  - Prote**omics** = the study of all the proteins throughout the human body (proteins = key signaling molecules which allow biological processes to occur).
  - Metabol**omics** = the study of all metabolites throughout the human body (metabolites = small molecular weight compounds which are the bi-product of biological processes → fats, cholesterols, amino acids etc).



What is Omics?

- **The science of trying to capture a biological system(s) in its**

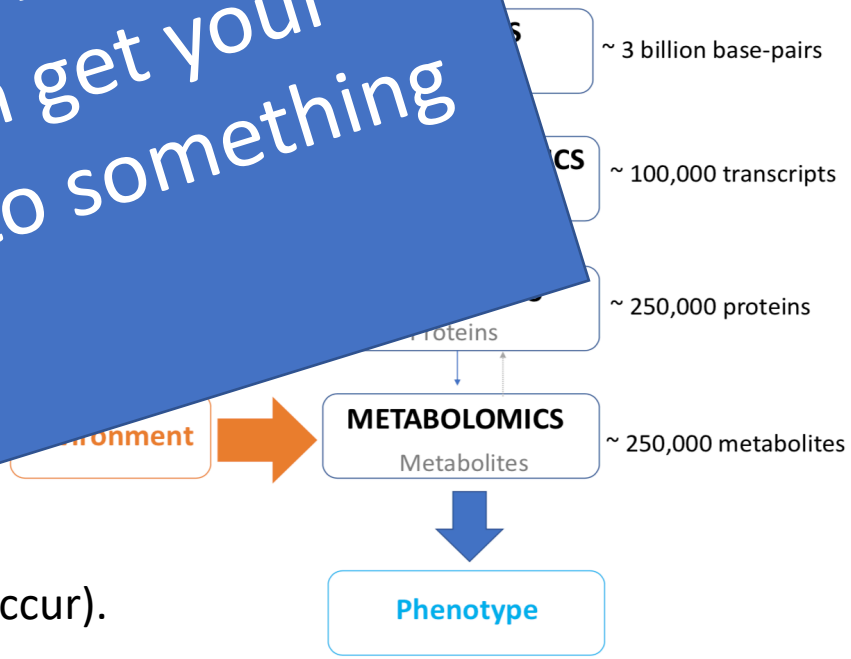
- '**Omics**" fields are split across different biol

  - Gen**omics** = the st

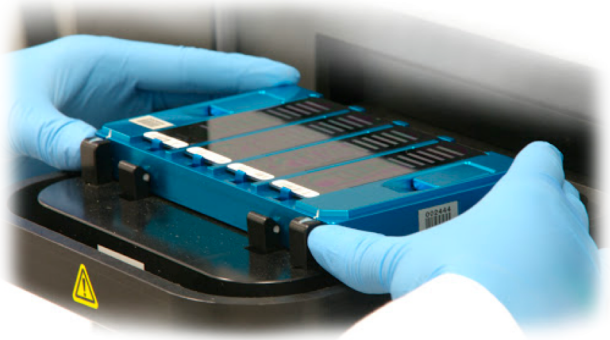
  - human body
    biological processes to occur).

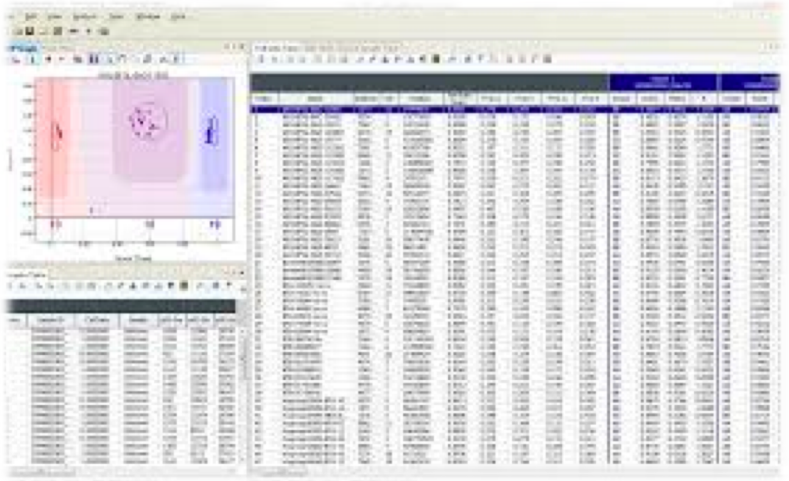
  - all metabolites throughout the human body
    molecular weight compounds which are the bi-product of
    processes → fats, cholesterols, amino acids etc).

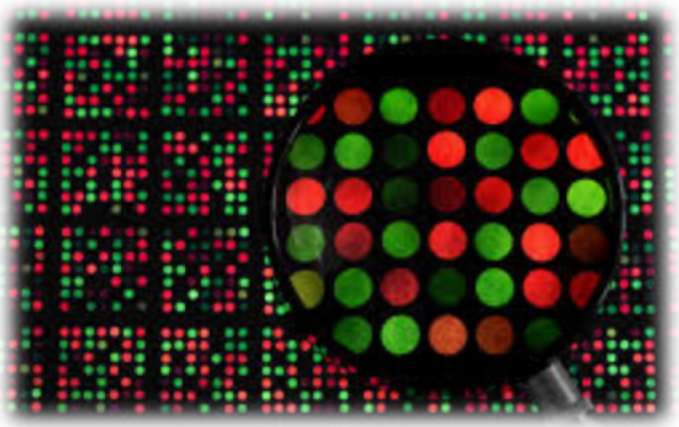
Taking the entirety of the information within a biological strata (or as much as you can get your hands on) and trying to translate it into something meaningful.



What is Omics?



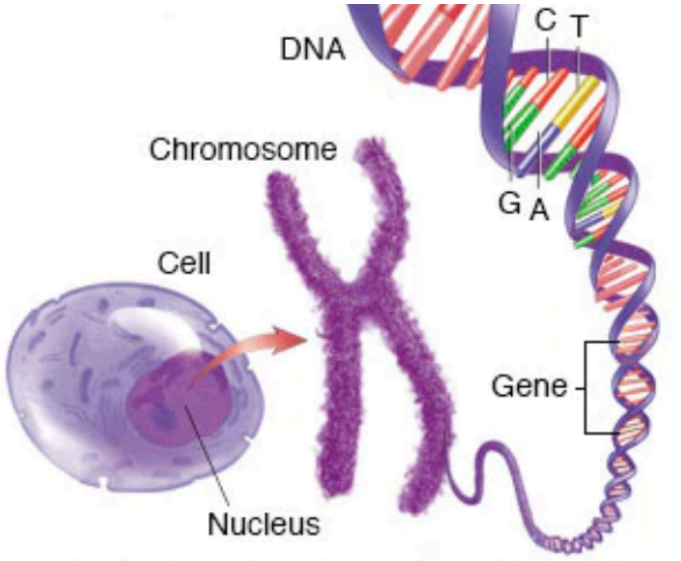




What is Omics?

- A (very) quick biology refresher….

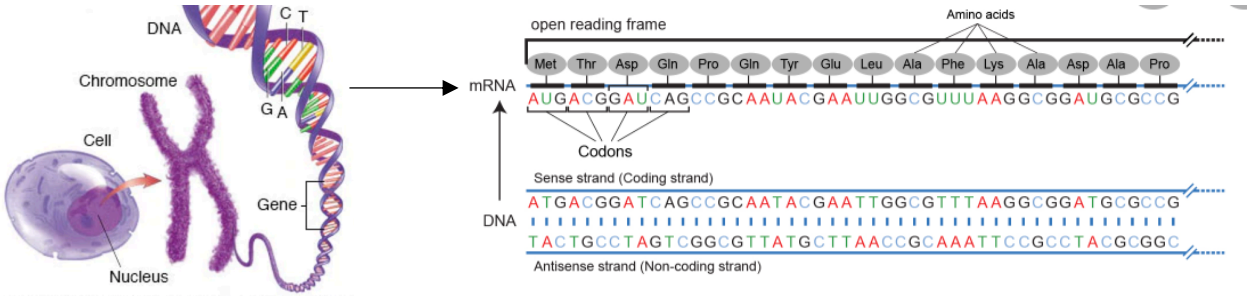
GENES

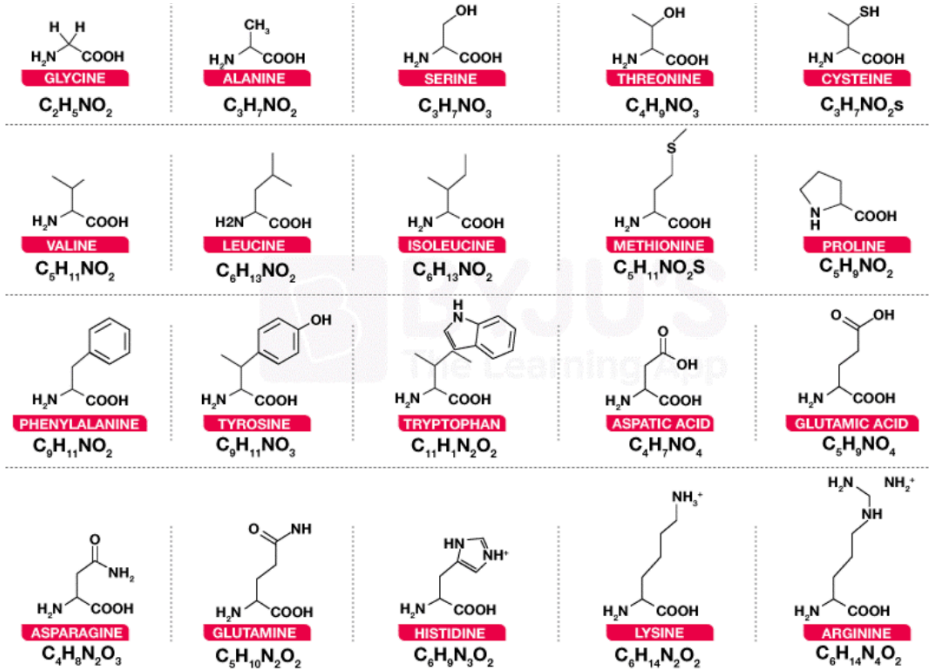


What is Omics?

- A (very) quick biology refresher….

GENES MAKE AMINO ACIDS

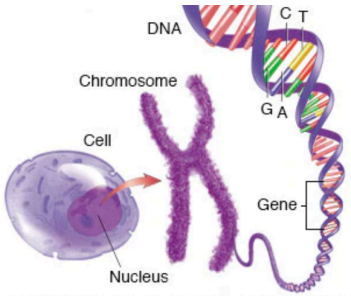


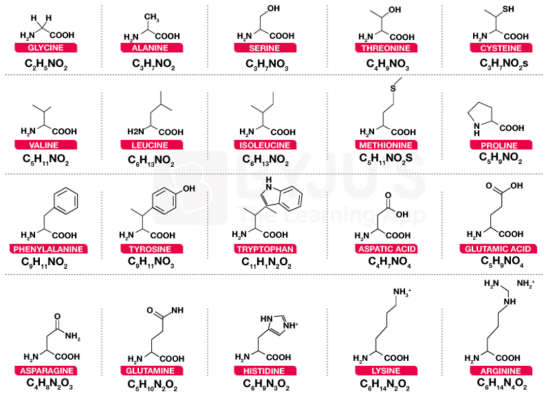


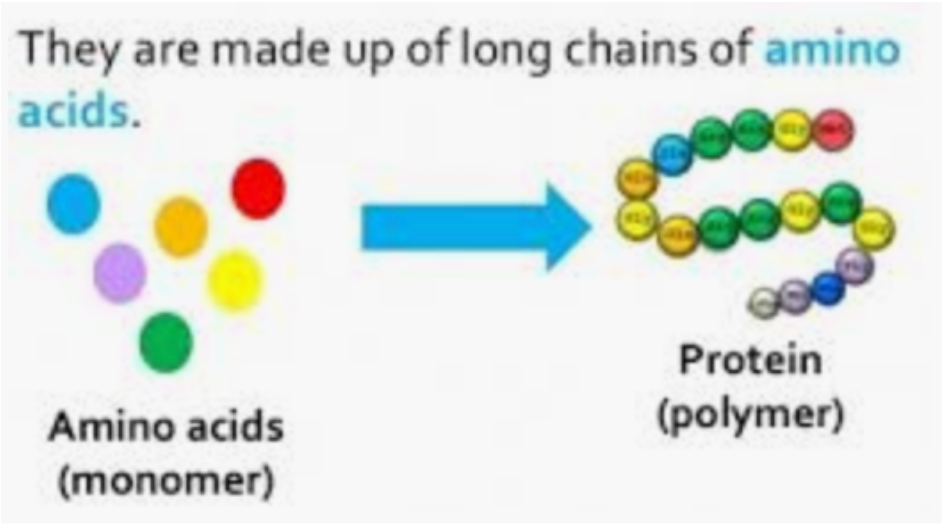
What is Omics?

- A (very) quick biology refresher….

GENES MAKE AMINO ACIDS FORM PROTEINS



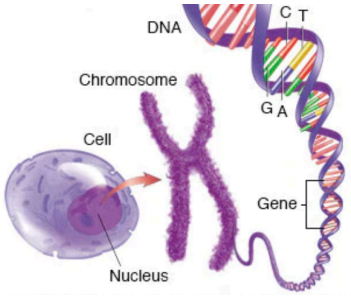


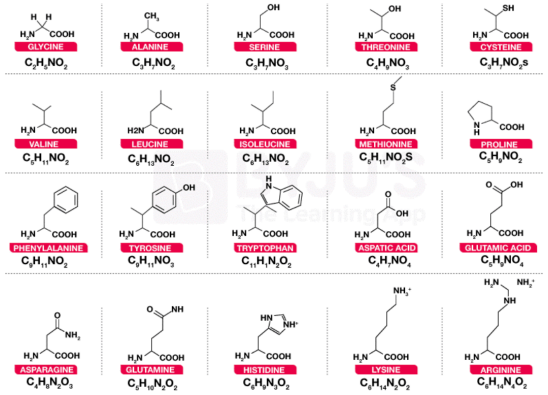


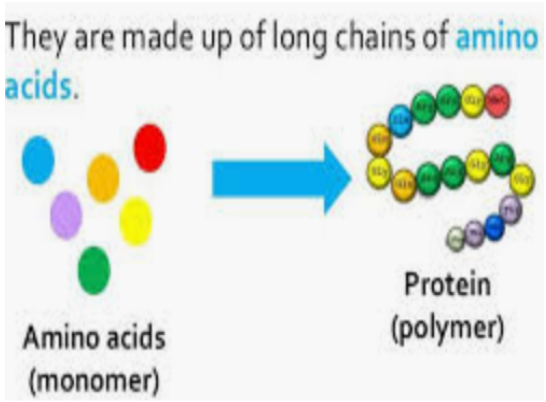
What is Omics?

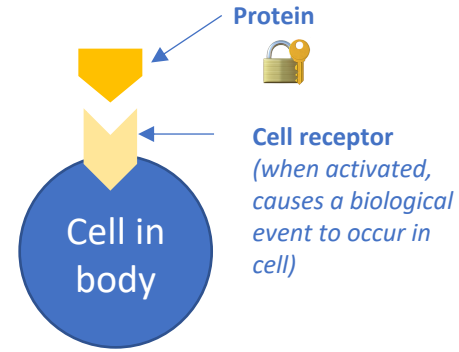
- A (very) quick biology refresher….

GENES **MAKE** AMINO ACIDS **FORM** PROTEINS **ENABLE** BIOLOGICAL PROCESSES



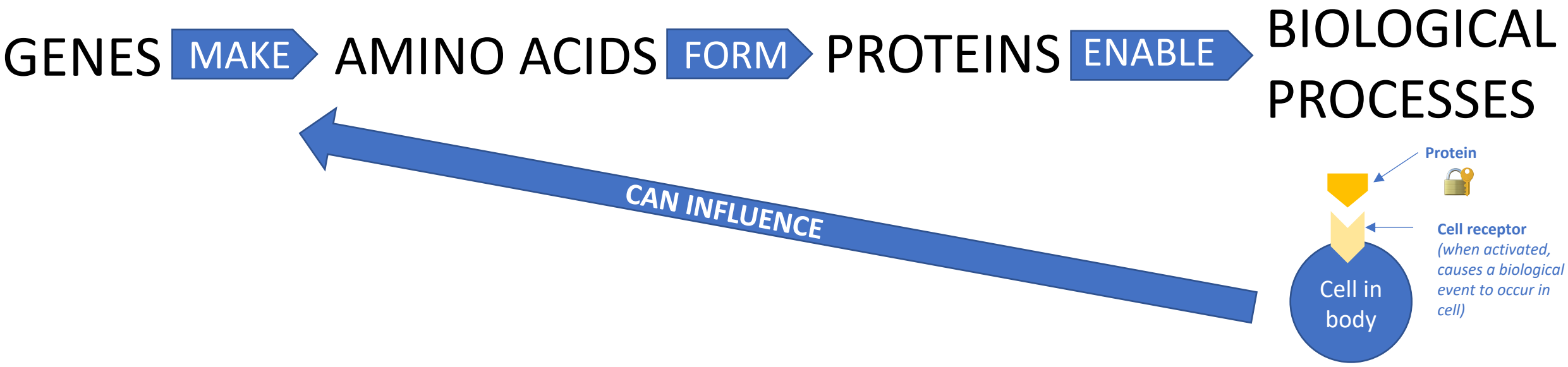






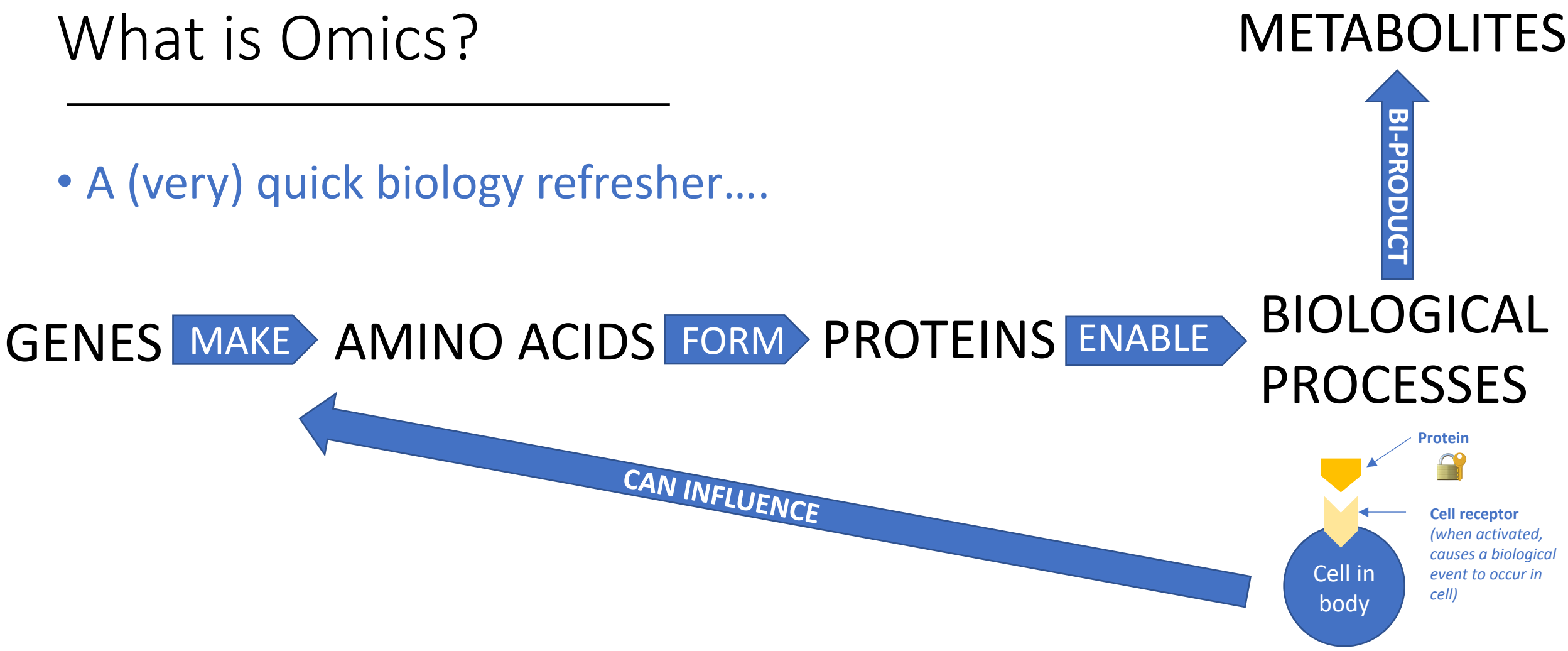
What is Omics?

- A (very) quick biology refresher….



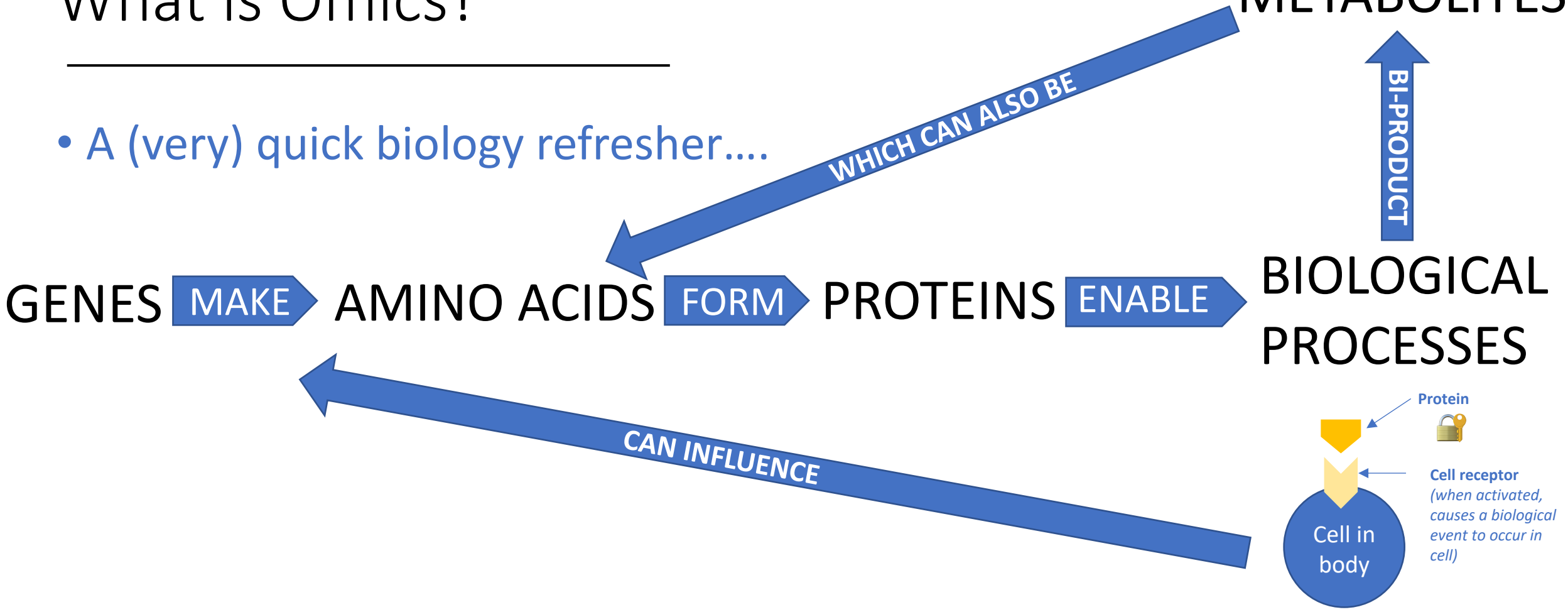
What is Omics?

- A (very) quick biology refresher….



What is Omics?

- A (very) quick biology refresher….



What is Omics?

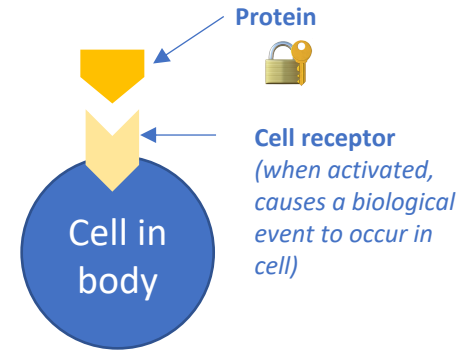
- A (very) quick biology refresher…

GENES MAKE

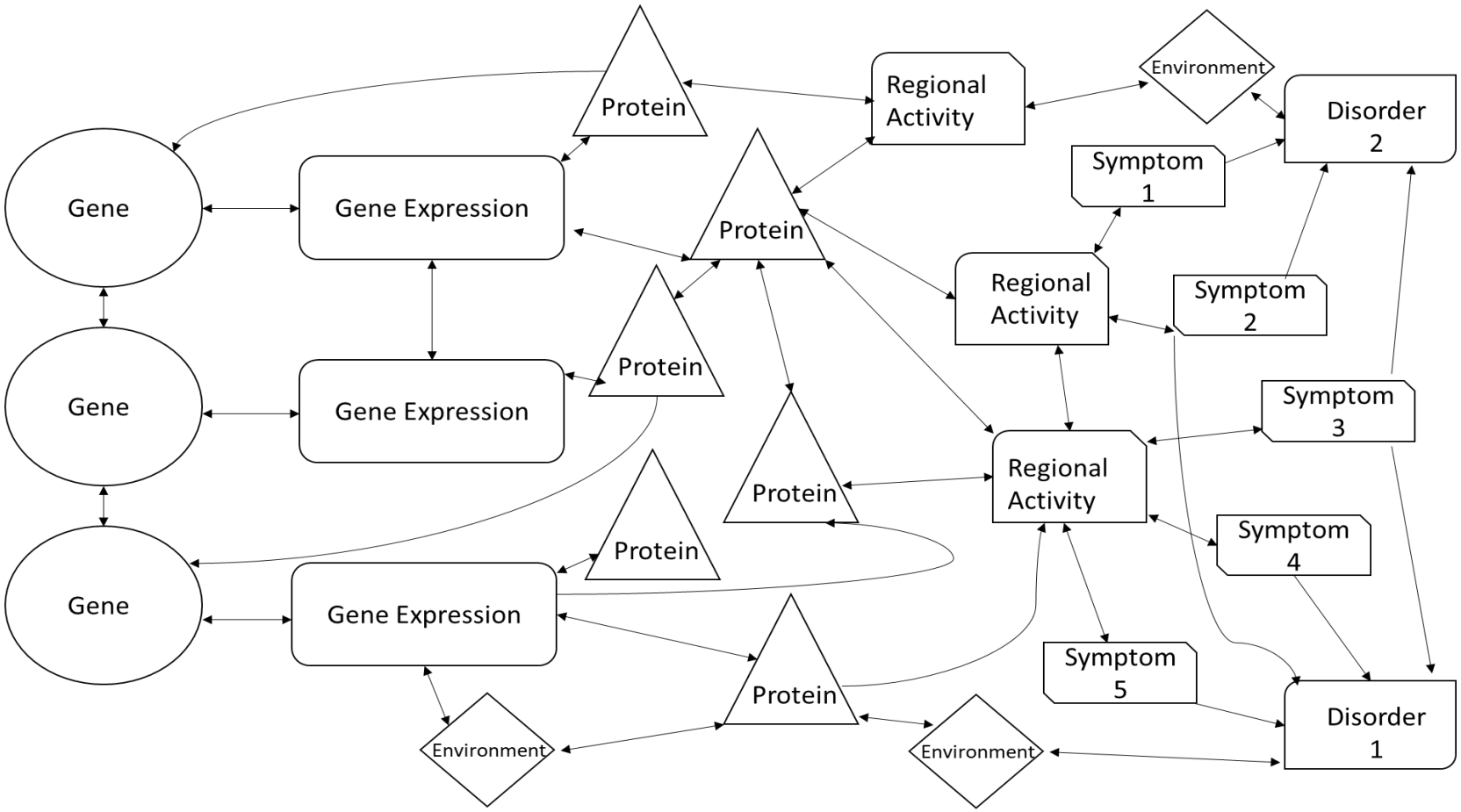
ETC ETC ETC…

…UENCE

…OLOGICAL PROCESSES

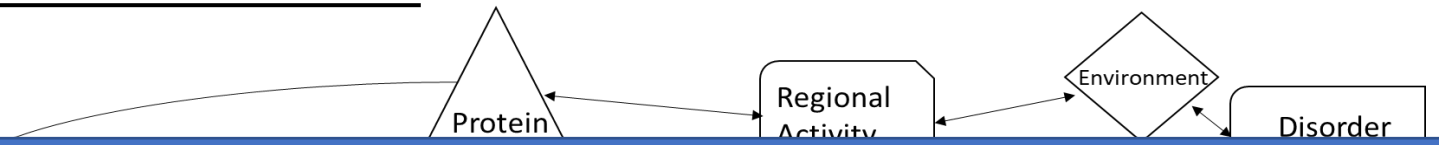


What is Omics?



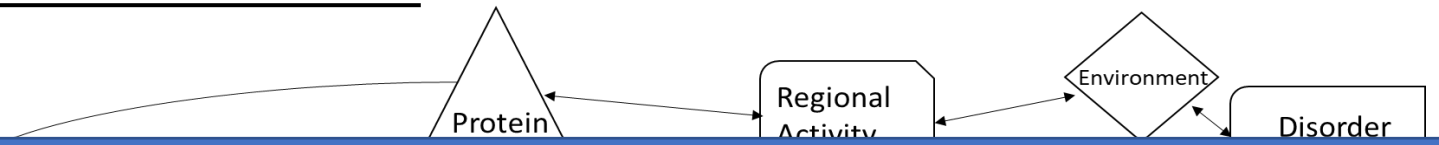
** EXTREMELY simplified

What is Omics?



- DIFFERENT BIOLOGICAL STRATA DO NOT WORK IN SILOS…
- **MULTI-OMICS** = A RELATIVELY NEW FIELD OF OMICS WHICH ATTEMPT TO UTILIZE INFORMATION **ACROSS** THE **BIOLOGICAL STRATA** TO BETTER INFORM OUTCOMES…

What is Omics?



- THE DIFFERENT STRATA DO NOT WORK IN SILOS…
- **MULTI-OMICS** = A RELATIVELY NEW FIELD OF OMICS WHICH ATTEMPT TO UTILIZE INFORMATION **ACROSS** THE **BIOLOGICAL STRATA** TO BETTER INFORM OUTCOMES…

THIS IS ME 🙃

Learning Objectives

- Understand the concept of “high dimensional data” (HDD)
- Understand the concept of “omics”.
- Understand where “omics” fit in the realms of HDD
- Grasp possibilities and challenges that accompany HDD

Possibilities & Challenges of HDD / Omics

- High dimensional datasets contain an incredible wealth of information which have the potential to unlock previously unknown patterns and relationships between variables.
- Allow for advanced “networks” of relationships to be modelled.
- Lend themselves to modern machine learning techniques and and emerging sophisticated algorithms which can ingest and learn from the information passed through to un
- Allow for data-driven / hypothesis free learning.
- Personalized medicine.
- Exponential insights.

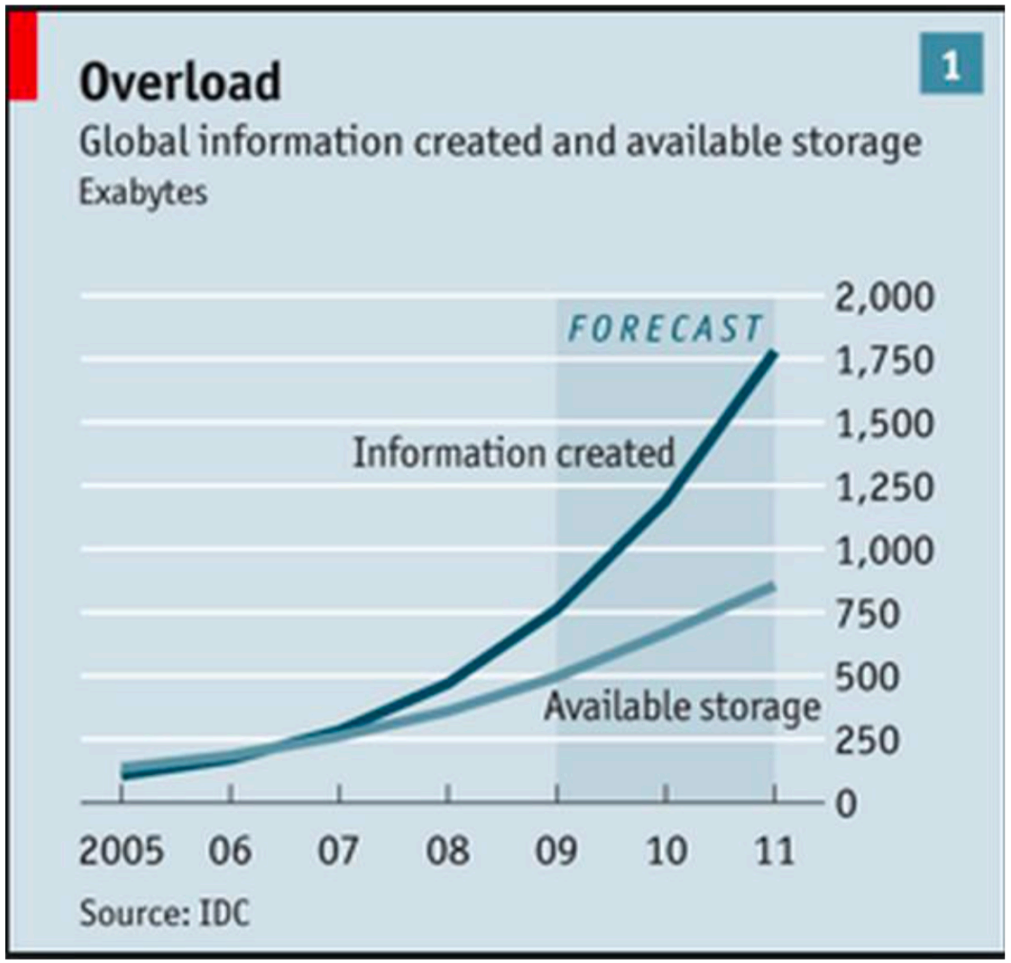
Possibilities & Challenges of HDD / Omics

- High dimensional datasets contain an incredible wealth of information which have the potential to unlock previously unknown patterns and relationships between variab
- Allow for advanced “
- ticated algorithms
- ypothesis free learning.
- Personalized medicine.
- Exponential insights.

HOWEVER…

Possibilities & Challenges of HDD / Omics

- A computational nightmare…



Possibilities & Challenges of HDD / Omics

- Different processing platforms / quality control processes / out of date builds.
- Mapping strata across different samples (multi-omics).
- Mapping phenotypes across different samples.
- Biology and behavior are not static → temporal changes / cause vs effect / feedback loops etc.
- Heterogeneity
- Where to begin? – what to focus on? – how to start unpicking the numbers?

Takeaways…

- We have entered the “data driven” era.
- This offers huge potential for uncovering novel insights to healthcare and drive personalized approaches to medicine.
- **BUT** the amount of data out there is (already) on an overwhelming scale, and mapping this into something meaningful presents a major challenge.
- Different “omic” modalities have already demonstrated promise in unlocking clues to some of the biological underpinnings of disease and behavior – but the majority of current approaches concentrate on single biological strata.
- Methods which can integrate data across different modalities and disentangle heterogeneous networks are crucial if we hope to unravel aetiological complexities of disease and offer plausible diagnostic and treatment strategies.