BIG DATA E DADOS DE PESQUISA EM SAÚDE

Benilton S Carvalho, Ph.D.

Departamento de Estatística

UNICAMP

’Big Data’ é um conjunto de recursos de grande **volume**, alta **velocidade** e ampla **variedade**, que demanda formas de processamento de informação que sejam custo-efetivas, além de inovadoras, para a obtenção de conhecimento melhorado e para a tomada de decisão. (Gartner – 3 V’s de Big Data)

’Big Data’ é um conjunto de recursos de grande **volume**, alta **velocidade** e ampla **variedade**, que demanda formas de processamento de informação que sejam custo-efetivas, além de inovadoras, para a obtenção de conhecimento melhorado e para a tomada de decisão. (Gartner – 3 V’s de Big Data)

VARIEDADE – MAIS IMPORTANTE!

- Imagens, imagens com escalas temporais;
- Testes/exames realizados, componente temporal;
- -ômicas em suas diversas facetas;
- Tratamentos aplicados e seus resultados (fornecedores? Asparaginase);
- Complicações e desfechos;
- Wearables;
- Histórico pessoal de exposição (rotina cotidiana);
- Dados geo-populacionais relevantes;
- Histórico familiar (Inception!);

VELOCIDADE – MAIS MAL-COMPREENDIDO

- Não se refere à rapidez de execução analítica;
- Refere-se à rapidez na conexão das informações coletadas;
- Registro eletrônico de entrada/saída de medicamentos já é realidade:
  - Realidade, habitualmente, para fins contábeis;
  - Habilitar conexão a este serviço também para fins de pesquisa;
  - Empregar estas informações para monitoramento real-time de índices de interesse;

VOLUME – MAIS SIMPLES DE SE ENTENDER

- Ocupa muito espaço!
- Nem sempre, tratam-se de dados estruturados;
- Dificuldade em organização lógica dos dados;
- Se, em um hospital, explorarmos todos estes dados para todos os pacientes atendidos ao longo de um ano → ainda mais volumoso!
- Mudança de paradigma analítico:
  - os dados não seguem para análise;
  - a análise é enviada para os dados;
  - resultados são necessários imediatamente (dashboards, por exemplo);
- Fazemos parte de uma sociedade que separa o com da unidade:
  - "Meu estudo ..."; "Meus pacientes ..."

+7 V'S PARA BIG DATA

- **Variabilidade**: diferentes tipos para pacientes diferentes;
- **Veracidade**: como os dados foram gerados? Podem ter sido adulterados?
- **Validade**: os dados são acurados para o uso que terão?
- **Vulnerabilidade**: os dados estão seguros? O que pode acontecer em caso de vazamento de informação?
- **Volatilidade**: qual a idade de "aposentadoria" dos dados?
- **Visualização**: como representar a multitude de dados de relevância?
- **Valor**: consigo melhorar processos? Prover melhores serviços de saúde?

Zuckerberg admite: Facebook coleta dados mesmo de quem não é usuário

Executivo alega motivos de segurança para justificar a prática

POR **BLOOMBERG NEWS**

11/04/2018 17:29 / atualizado 11/04/2018 19:48

FORMAS CUSTO-EFETIVAS E INOVADORAS PARA PROCESSAMENTO DE DADOS

- Custo-Efetivo é diferente de barato;
- Mire nas respostas apropriadas para o problema em questão;
- Exemplos:
  - Como armazenar e processar dados não-estruturados?
  - Como conectar os diferentes provedores de dados?

CONHECIMENTO APRIMORADO E TOMADA DE DECISÃO

- Este é o objetivo principal!
- O que sabemos agora que não estava disponível antes?
  - Uma métrica de adequacidade para um certo tratamento?
  - Detecção prematura de riscos?
- Fase mais complexa de todo o processo

ESTRATÉGIAS PARA FACILITAR A DINÂMICA DE COLETA DE DADOS

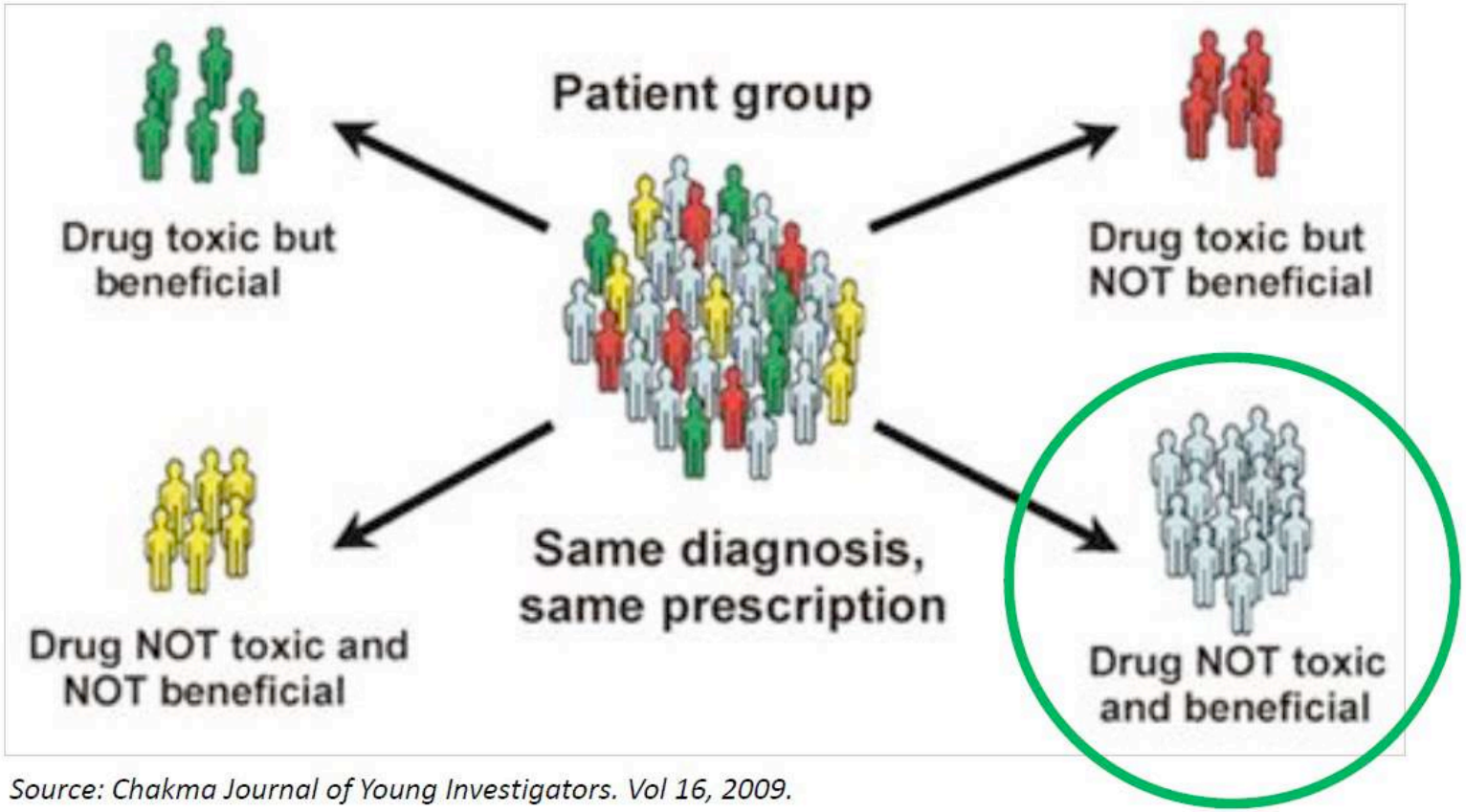
- Vocabulário padronizado: CID, por exemplo;
- Formalização de protocolos;
  - Identificação de pontos de coletas de dados;
  - Conectividade com sistemas existentes;
- Para dados que sejam coletados manualmente, buscar estratégias que busquem minimizar a possibilidade de erros;
- Vale ter seu estatístico-de-bolso para consultar que dados são necessários para cada análise e o melhor formato para os mesmos.

FERRAMENTAS ÚTEIS

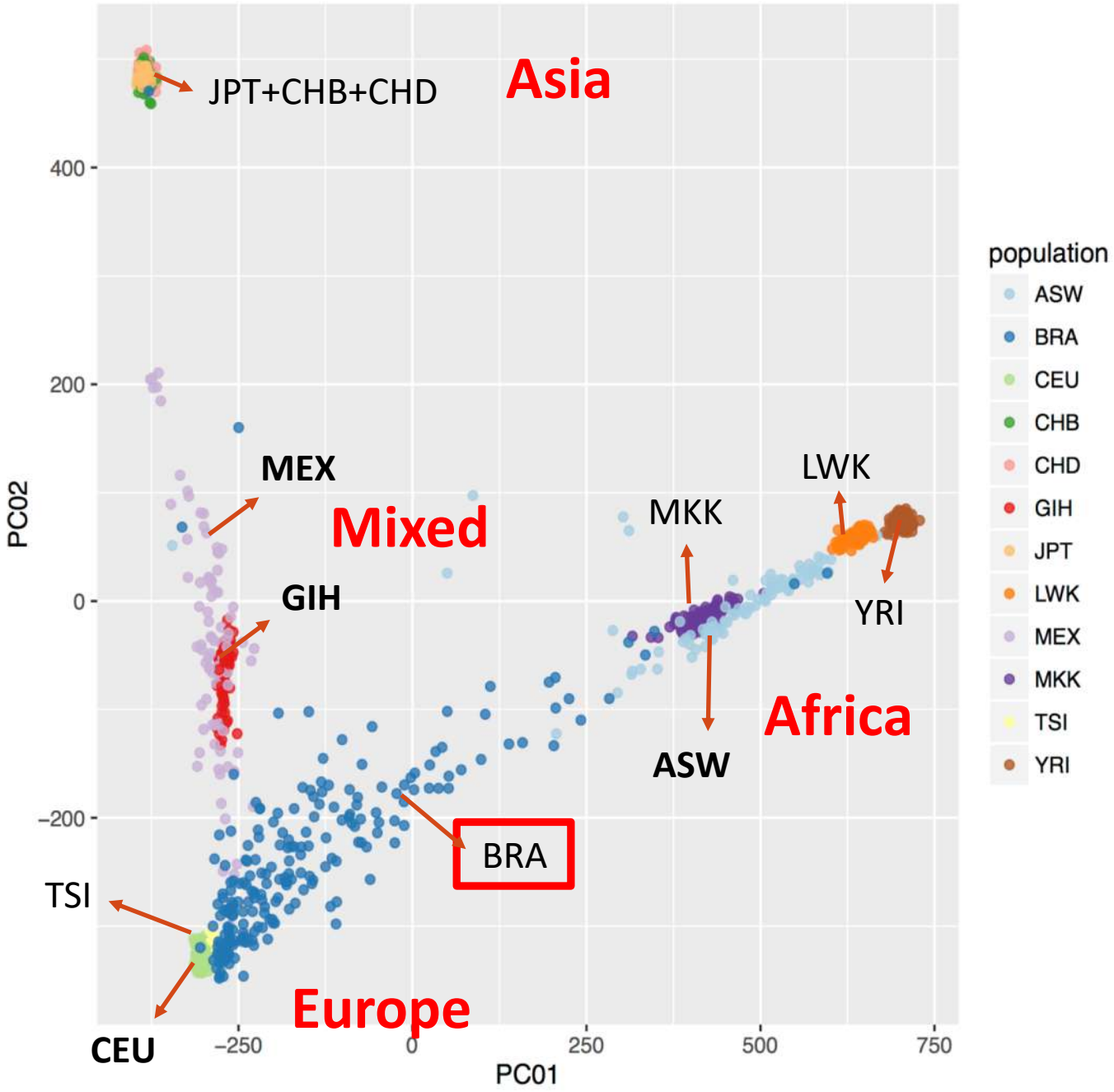
- Processamento de Linguagem Natural (NLP);
- Reconhecimento Ótico de Caracter (OCR);
- Aprendizado (Estatístico) de Máquina;
- Sistemas de armazenamento de dados de alta performance;
- A nuvem, mesmo que interna, agrega valor a estratégias assim, pois melhora a mobilidade;

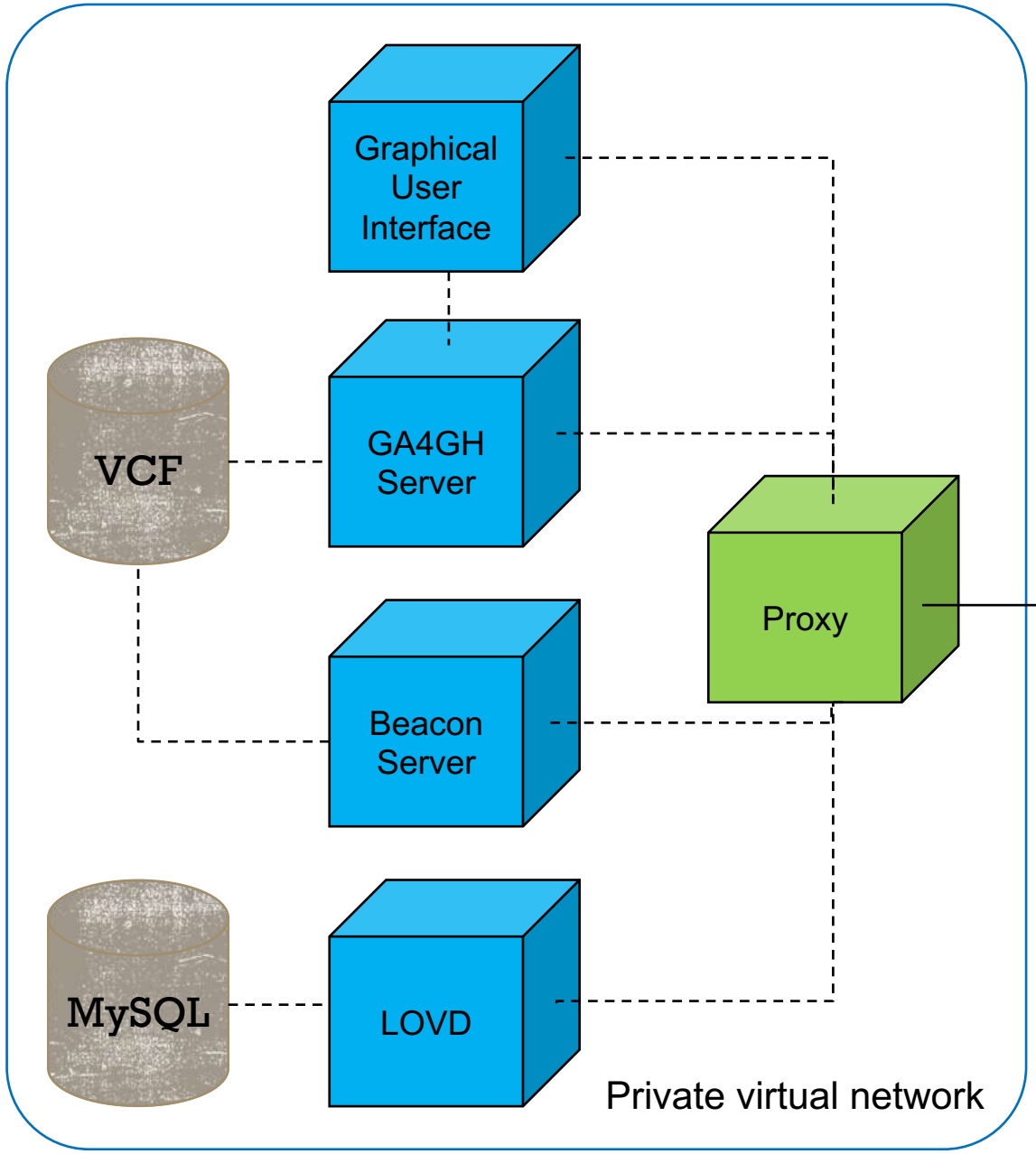
Grupo	Droga	Beneficia
Verde	T	S
Vermelho	T	N
Azul	NT	S
Amarelo	NT	N

Principle of Personalized/Precision/Targeted Medicine



Source: Chakma Journal of Young Investigators. Vol 16, 2009.


JPT+CHB+CHD
Asia
400
200
0
−200
PC02
MEX
Mixed
GIH
MKK
LWK
YRI
Africa
ASW
BRA
TSI
Europe
CEU
−250
0
250
500
750
PC01
population
ASW
BRA
CEU
CHB
CHD
GIH
JPT
LWK
MEX
MKK
TSI
YRI



https://labbcb.github.io/federated-genomic-databases

Dataset
bipmed-exome

Variant Set
bipmed-exome

Gene Symbols File
Browse... No file

Gene Symbol
SCN1A

Genomic Feature
- (•) Genes
- ( ) Transcripts
- ( ) Exons
- ( ) CDS
- ( ) Promoters

Reference Name
chr2

Start
165989160

End

Variants | Beacon Network | Help

Show 10 entries Search:

	DNA change	dbSNP ID	Reference Name	Start	End	Reference Bases	Alternate Bases	ExcessHet
1	g.165991411A>G	chr2:165991411_A/G	chr2	165991411	165991411	A	G	3.03090000152588
2	g.165991493G>C	chr2:165991493_G/C	chr2	165991493	165991493	G	C	3.01029992103577
3	g.165991857C>T	chr2:165991857_C/T	chr2	165991857	165991857	C	T	3.03090000152588
4	g.165992330G>A	chr2:165992330_G/A	chr2	165992330	165992330	G	A	3.01029992103577
5	g.165992388G>A	chr2:165992388_G/A	chr2	165992388	165992388	G	A	3.03090000152588
6	g.165992564T>A	chr2:165992564_T/A	chr2	165992564	165992564	T	A	0.105499997735023
7	g.165995836_165995840delinsT	chr2:165995836_TTTAA/T	chr2	165995836	165995840	TTTAA	T	3.03150010108948
8	g.165995889G>C	chr2:165995889_G/C	chr2	165995889	165995889	G	C	3.07249999046326
9	g.165998005C>T	chr2:165998005_C/T	chr2	165998005	165998005	C	T	3.07369995117188
10	g.165999649G>A	chr2:165999649_G/A	chr2	165999649	165999649	G	A	3.2195999622345

Showing 1 to 10 of 66 entries

Previous 1 2 3 4 5 6 7 Next

Download Results

Dataset
bipmed-exome

Variant Set
bipmed-exome

Gene Symbols File
Browse... No file

Gene Symbol
SCN1A

Genomic Feature
- (•) Genes
- ( ) Transcripts
- ( ) Exons
- ( ) CDS
- ( ) Promoters

Reference Name
chr2

Start
165989160

End

Variants | Beacon Network | Help

Beacon Network

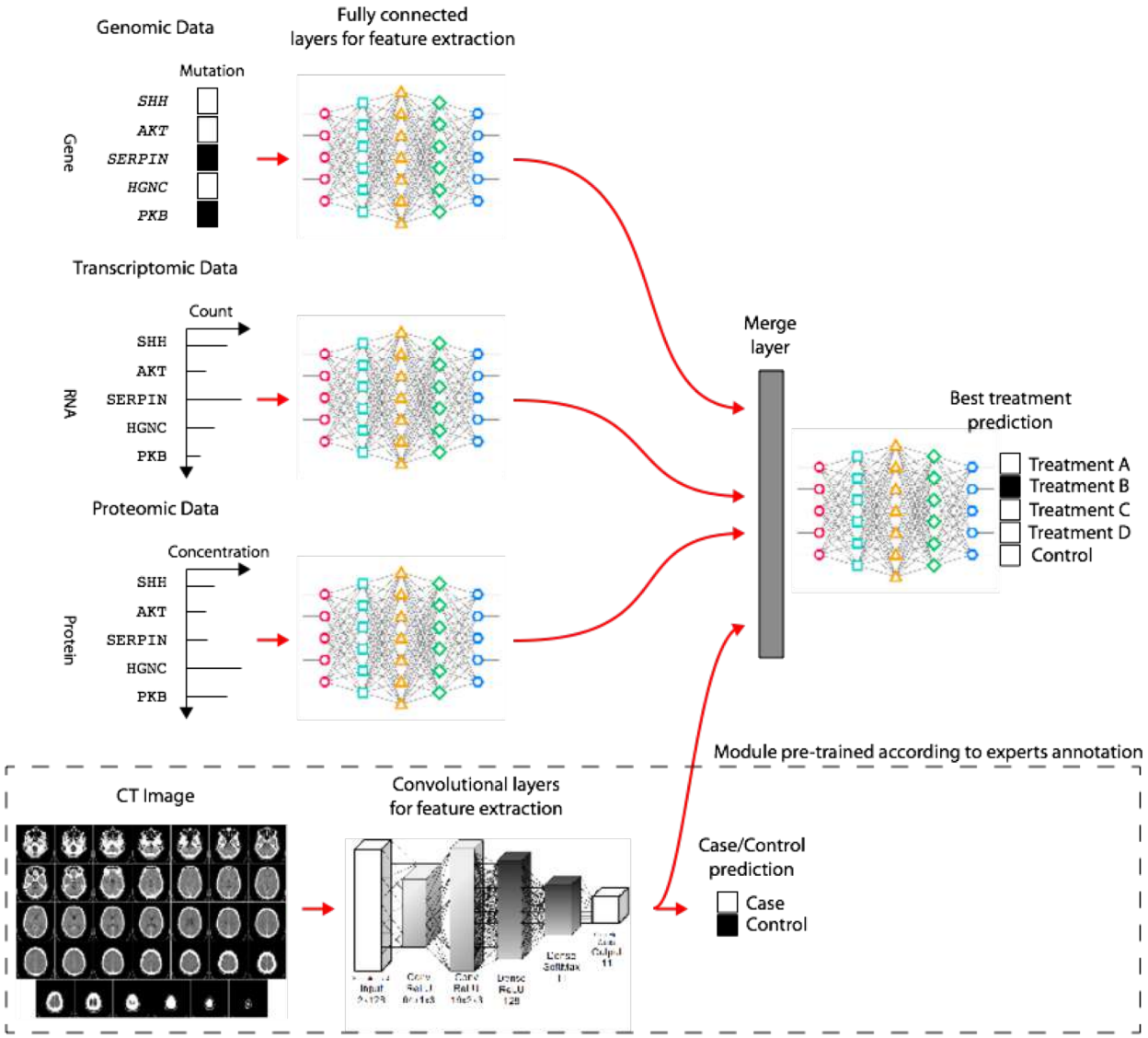
Response All None

- [x] Found 5
- [ ] Not Found 7
- [ ] Not Applicable 50

Organization All None

- [x] AMPLab, UC Berkeley
- [x] Australian Genomics Health Alliance
- [x] Belgian Medical Genomics Initiative
- [x] BGI
- [x] BioReference Laboratories
- [x] Brazilian Initiative on Precision Medicine
- [x] BRCA Exchange
- [x] Broad Institute
- [x] Centre for Genomic Regulation
- [x] Centro Nacional de Analisis Genomico
- [x] Children's Mercy Hospital
- [x] Curoverse
- [x] DNAstack
- [x] ELIXIR
- [x] EMBL-European Bioinformatics Institute

TREINAMENTO DE REDES NEURAIS



ANÁLISES INTEGRATIVAS

- Combinar dados de diferentes origens para explicação de um ou mais fatores de interesse;
- Múltiplas fontes de dados (confiáveis) incrementa o poder de detecção de eventos, identificação de grupos, etc.

ACHADOS EM ANÁLISES INTEGRATIVAS

- Precisamos de muito de tudo;
  - Muitas amostras;
  - Muito poder computacional (GPU);
  - Muito espaço de armazenamento;
- Cresce o valor de comunidade de dados e compartilhamento responsável de informações;
- Precisamos de legislação!

OBRIGADO!

- Benilton Carvalho
- benilton@unicamp.br

UNICAMP