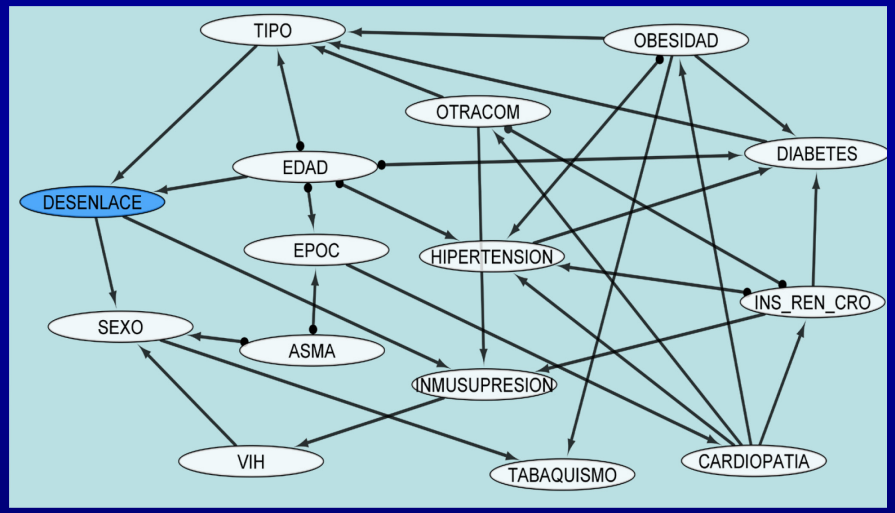
CONFERENCIA INTERNACIONAL BIREDIAL-ISTEC '21
10 años 2011 - 2021
25-29 de octubre 2021 TONANTZINTLA, PUEBLA • MÉXICO

Descubrimiento Causal en la Base de Datos Mexicana de COVID-19



L. Enrique Sucar

Instituto Nacional de Astrofísica, Óptica y Electrónica

Contenido

- Introducción
- Base de Datos Mexicana COVID-19
- Curación de la Base de Datos
- Estadísticas
- Modelos Causales y Descubrimiento Causal
- Descubrimiento causal en BD de COVID-19
- Conclusiones

Introducción

- Actualmente las bases de datos son el “oro” de la investigación científica en muchos campos, por lo que es importante hacerlas accesibles a la comunidad
- En algunos dominios, como medicina, es difícil encontrar bases de datos abiertas, por cuestiones de privacidad, seguridad y/o intereses comerciales
- Sin embargo, es posible pre-procesar estas BD para acerlas accesibles y a la vez proteger los datos delicados, como es el caso que veremos …

La Base de Datos Mexicana de COVID 19

- La BD Mexicana de COVID-19 es el resultado de una colaboración de la Secretaría de Salud y la Universidad Nacional Autónoma de México.
- Cuenta con más de 6.5 millones de individuos con 97 variables, y continúan recabándose datos diariamente.
- La Base de Datos está públicamente accesible para fines de investigación:

http://covid-19.iimas.unam.mx

La Base de Datos Mexicana de COVID 19

- Se presenta la información de carácter público recopilada a nivel nacional por la Dirección General de Epidemiología (DGE) de la Secretaría de Salud la cual se provee a la UNAM para su curación y divulgación

- La base de datos SISVER incluye todas las pruebas procesadas por la red de laboratorios del sistema de salud pública nacional, la cual incluye el Instituto Nacional de Diagnóstico Epidemiológico y de Referencia así como los 32 laboratorios estatales de salud pública para el monitoreo y apoyo epidemiológico.

La Base de Datos Mexicana de COVID 19

- SISVER contiene información de pruebas, hospitalización y decesos de 5,186 unidades, que incluyen a las 475 unidades del sistema Centinela, además de otras 4,281 públicas y 430 privadas distribuidas en los tres niveles del sistema de salud.

- El sistema es de amplio espectro pero no contempla registrar todo caso o deceso en el país

La Base de Datos Mexicana de COVID 19

- Una vez que se eliminó toda la información sensible respecto a la identidad de los individuos, se procedió a hacer un proceso de "curación" de la Base de Datos por expertos de la UNAM (Dr. Zian Fanti, Técnico Académico del Departamento de Ciencias de la Computación del IIMAS).

Proceso de Curación

1. Cambio de codificación de texto a un formato universal UTF8 y transformación de formato txt a CSV.
2. Se eliminan caracteres especiales y se transforma el texto a mayúsculas.
3. Se asigna un tipo de dato, para cada columna del archivo CSV.
4. Las columnas correspondientes a fechas se transforman al formato universal de representación de fechas.
5. Se estandariza el número de caracteres en el campo ID_REGISTRO.
6. Se valida el campo CURP con el formato oficial.

Proceso de Curación

7. Se crean campos independientes para cada una de las partes que forman la dirección del paciente (calle, número, colonia, CP, etc.).
8. Se validan los datos de LONGITUD y LATITUD.
9. Se corrigen anomalías de errores de escritura y faltas de ortografía detectadas para las columnas que representan catálogos del archivo CSV.
10. Se actualizan los catálogos, si es que hubiera nuevos datos.
11. Se eliminan los campos que contienen información privada o sensible, para crear un archivo anónimo.
12. Se remplazan cadenas de caracteres por enteros basado en los catálogos actualizados.

Información de la Base de Datos

- La Base de Datos contiene 97 campos que se pueden dividir en:
  - Información General del Paciente
  - Síntomas
  - Comorbilidades
  - Diagnóstico y Tratamiento
  - Resultados (positivos, defunciones)

Variables para Análisis

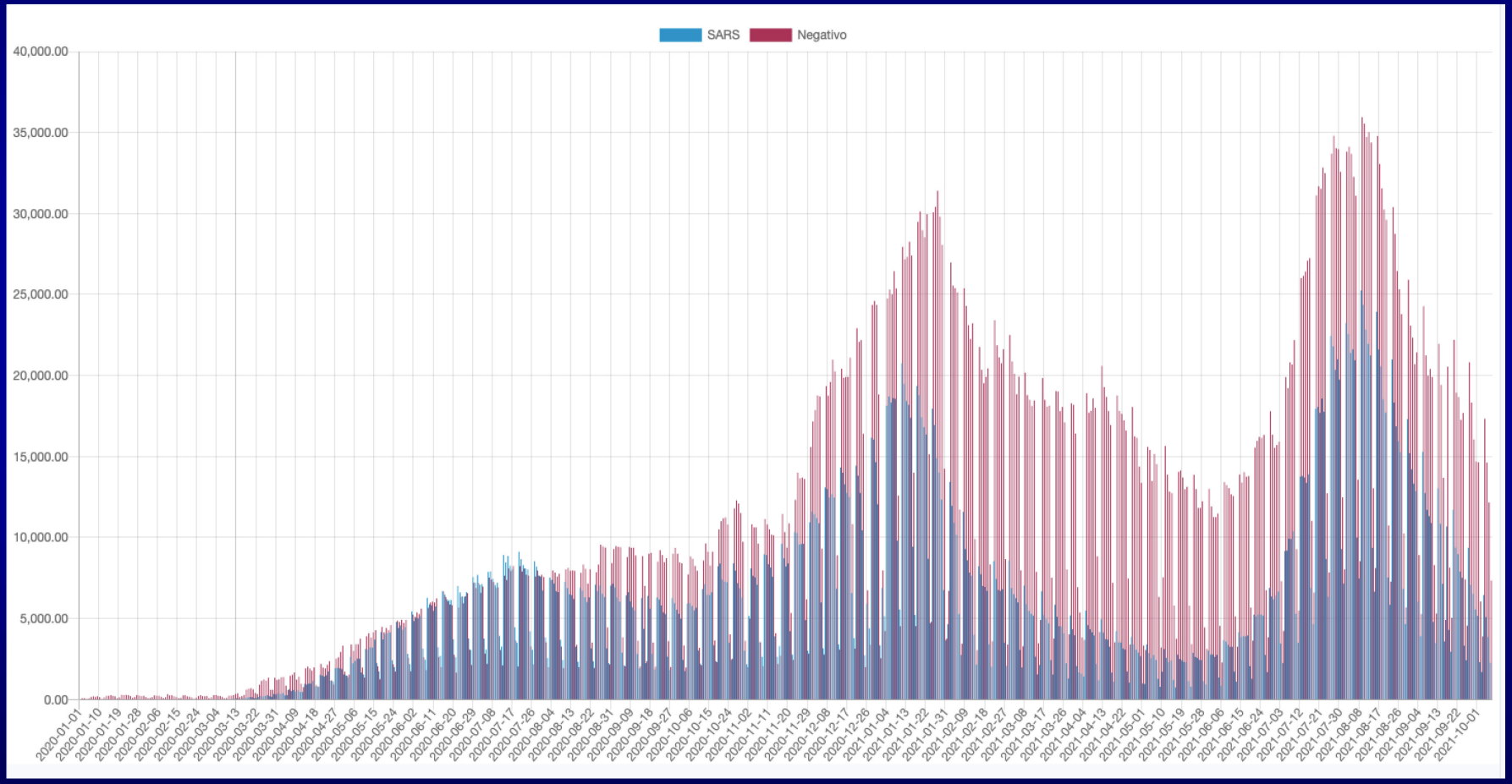
Para un análisis inicial se seleccionaron 47 variables

Category	Variables
patient-data	GENDER, AGE, CITY, NATIONALITY, PATIENT TYPE, INDIGENOUS, JOB, HOSPITAL SERVICE, CONTACT BIRDS, CONTACT PIGS, CONTACT COVID
symptoms	FEVER, COUGH, ODINOGY, DYSPNOEA, IRRITABILITY, DIARRHEA, CHEST PAIN, CHILL, HEADACHE, MYALGIA, ARTHRALGIA, DISCOMFORT, RHINORRHEA, POLYPNEA, VOMITING, ABDOMINAL PAIN, CONJUNCTIVITIS, CYANOSIS, SUDDEN SYMPTOMS, ANOSMIA, DYSGEUSIA
comorbidities	DIABETES, COPD, ASTHMA, IMMUNOSUPPRESSION, HYPERTENSION, HIV-AIDS, OTHER COMORBIDITIES, ENDO-CARDITIS, OBESITY, CHRONIC KIDNEY, SMOKING
diagnosis and treatment	ANALGESIC, ANTIVIRAL, ANTIPYRETICS, VACCINATED
objective class	COVID19, MORTALITY

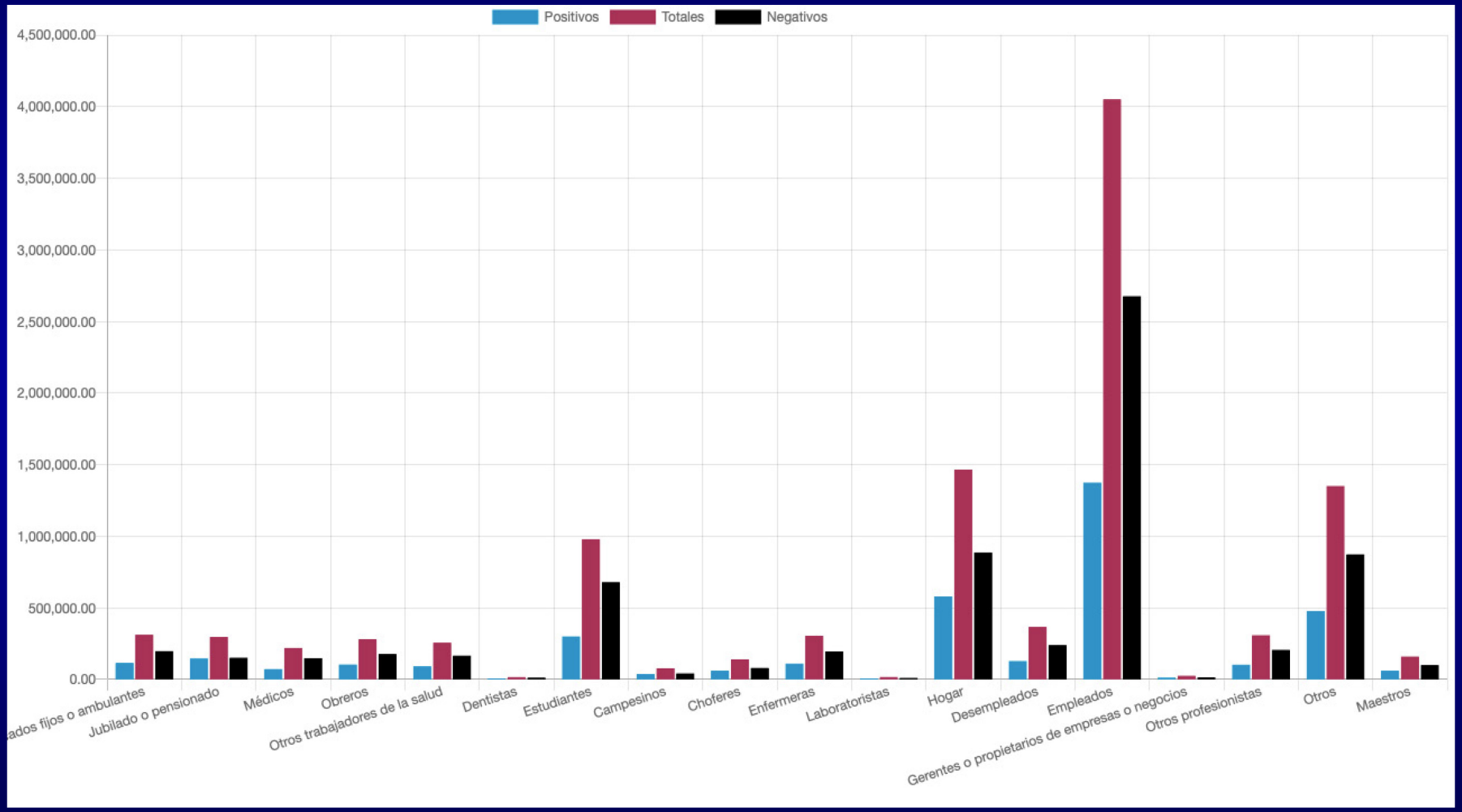
Estadísticas

- A partir de los datos se pueden obtener diversos análisis estadísticos
- A continuación se ilustran algunos ejemplos que se generan directamente en la página donde se encuentra alojada la base de datos Mexicana de Covid 19

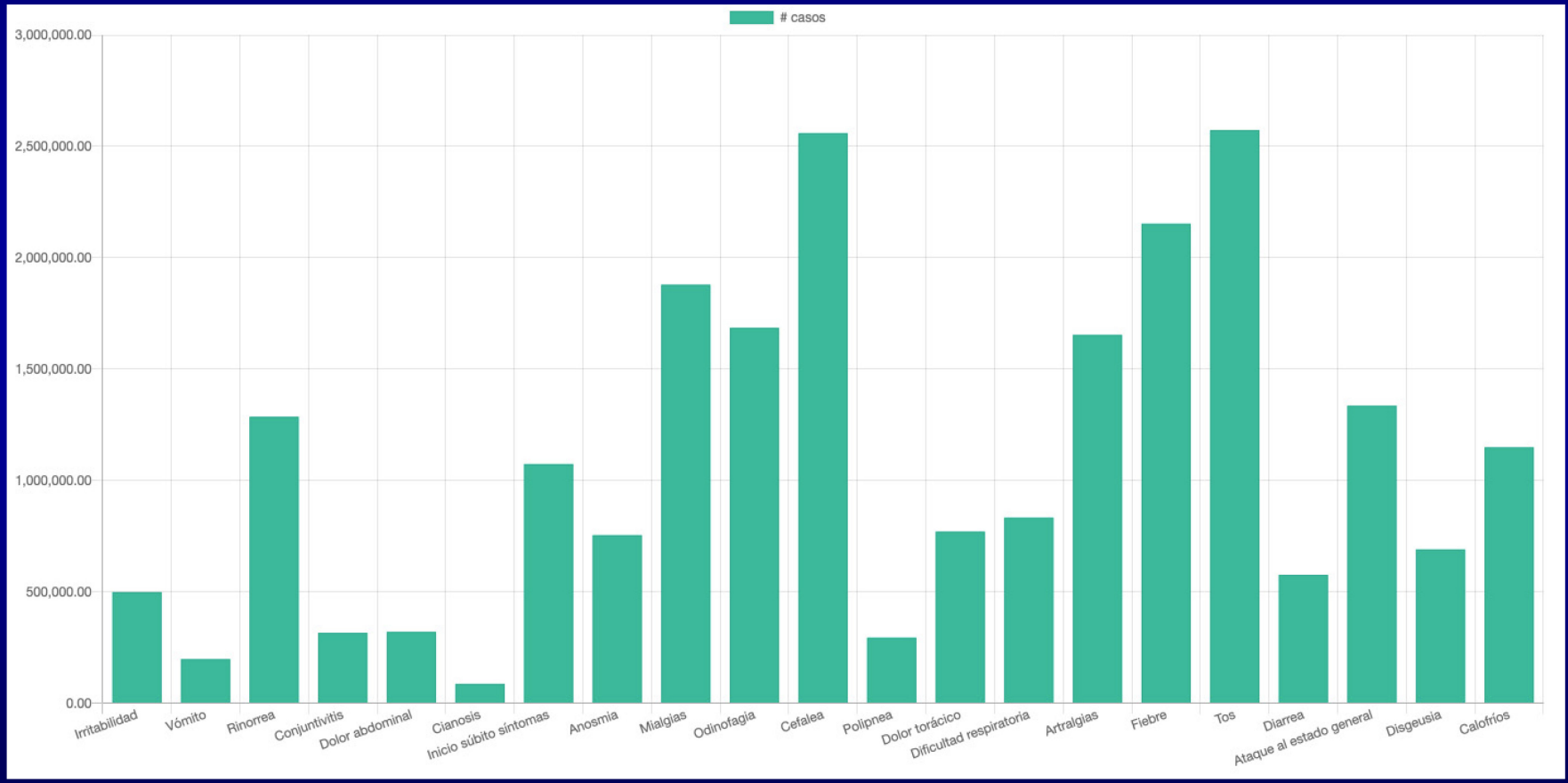
Evolución de la pandemia



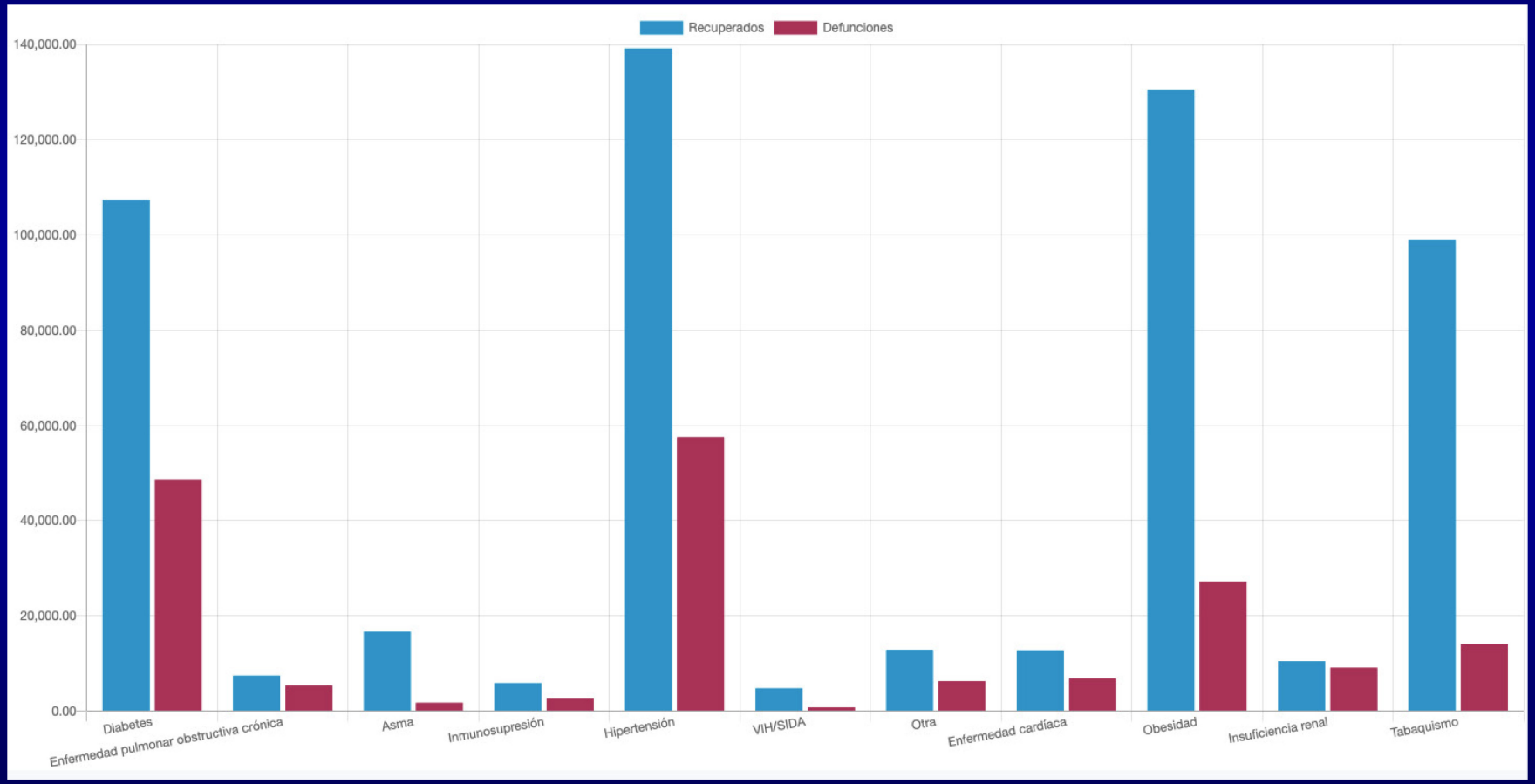
Número de Casos por Ocupación



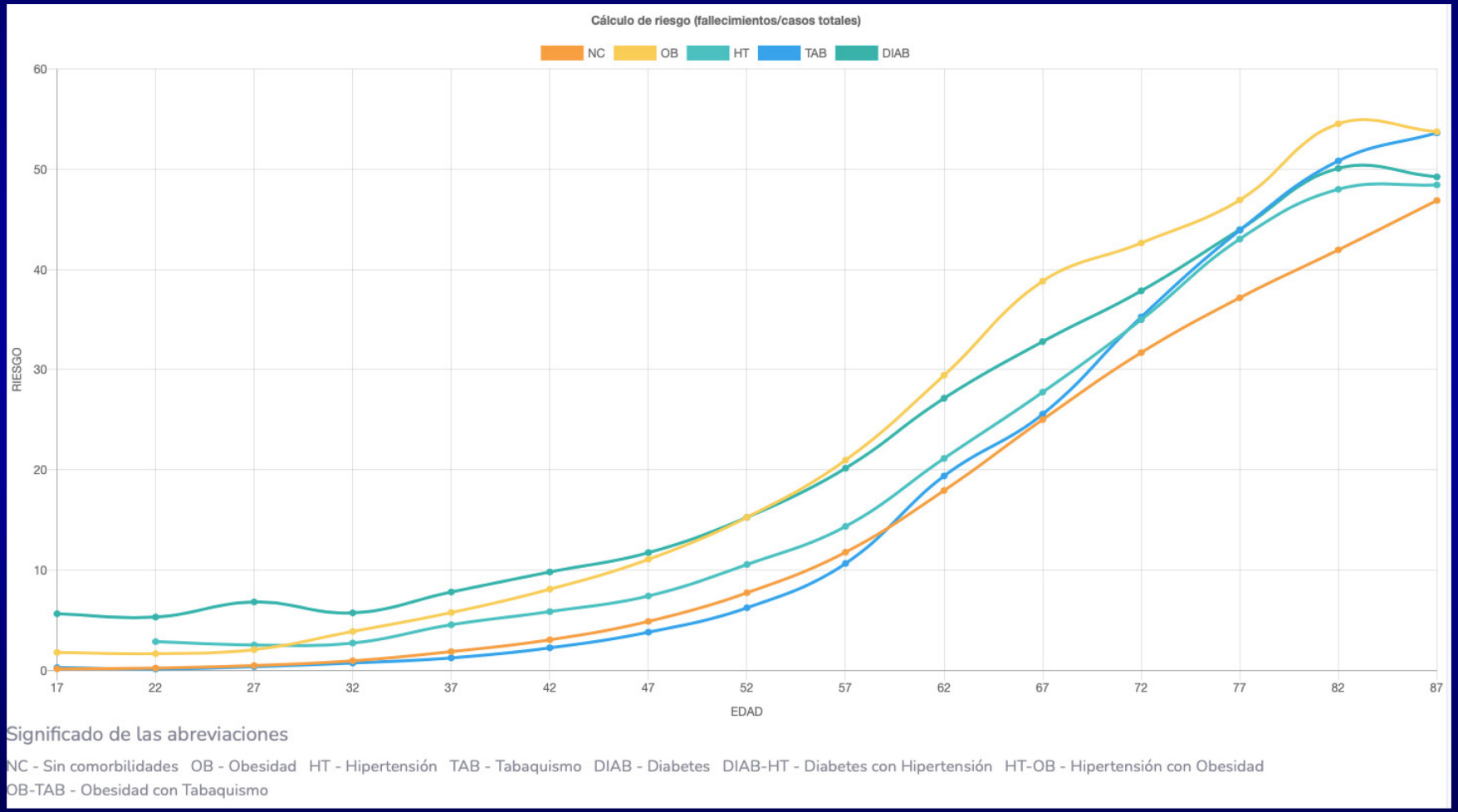
Número de Casos por Síntoma



Recuperados y defunciones por Comorbilidad



Riesgo (fallecimiento) por edad y algunas comorbilidades



Modelos Causales

Minería de Datos

- Las técnicas tradicionales de minería de datos obtienen asociaciones / correlaciones entre las variables, que pueden ser engañosas
- Para ir más allá de simplemente aprender asociaciones necesitamos aprender relaciones causales, lo que se conoce como descubrimiento causal

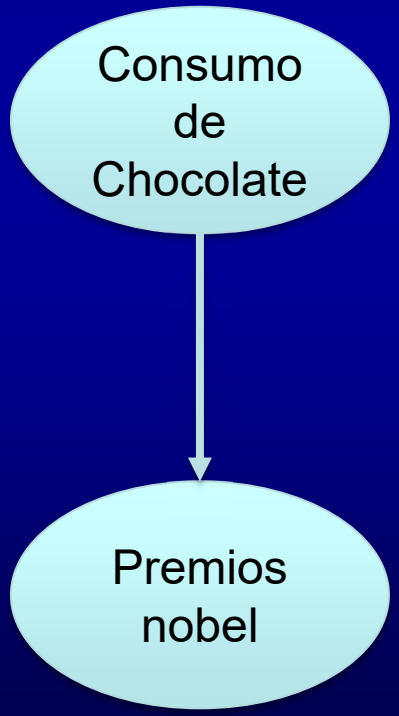
Descubrimiento Causal

- Que podríamos inferir de estos datos …

Consumo anual de chocolate per capita	Número de premios nobel
10	13
2	0
5	2
15	17
20	29
2	1
…	…

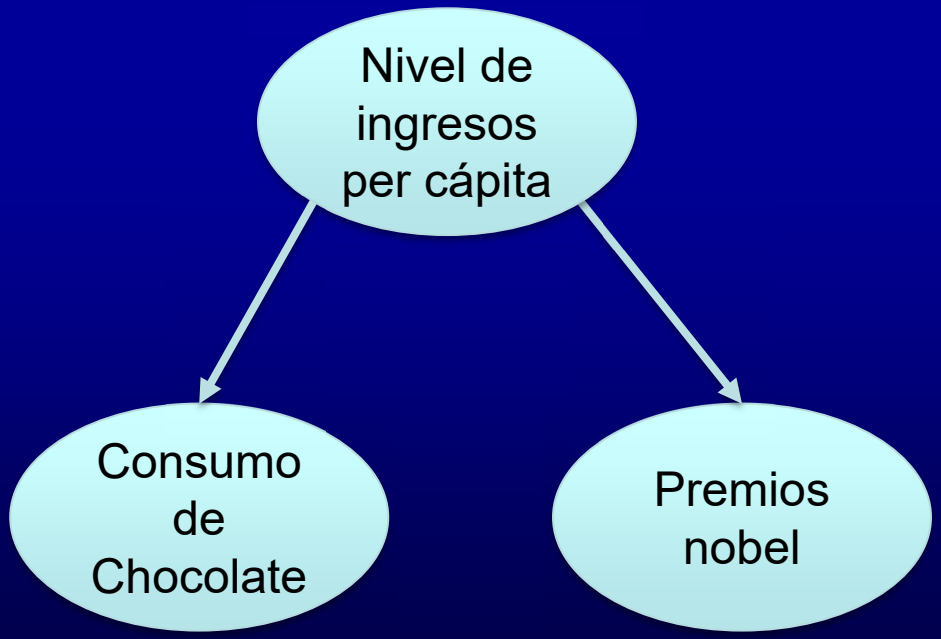
Posible modelo “causal”

- ¡Comer chocolate aumenta la probabilidad de lograr un premio nobel!



Modelo alternativo

- Hay un *cofactor* (causa común) que explica dicha asociación – países de mayores ingresos tienden a comer más chocolate y tienen más premios nobel



¿Porqué?

- Lo humanos pensamos en términos causales … fumar causa cáncer, manejar borracho pueda causar accidentes, un virus causa el COVID, …
- Nos preguntamos ¿porqué?
- La ciencia de la causalidad trata de contestarlo formalmente – entender el razonamiento causal y emularlo en las computadoras

Modelos Causales

- En los últimos años hay importantes avances en la representación y razonamiento causal, en particular en modelos gráficos causales:
  - Representación de conocimiento causal
  - Inferencia causal – *predecir* el efecto de intervenciones (¿cuál será el efecto de cierta política?) e *imaginar* escenarios alternativos (¿estaría vivo si no hubiera tomado cierta droga?)

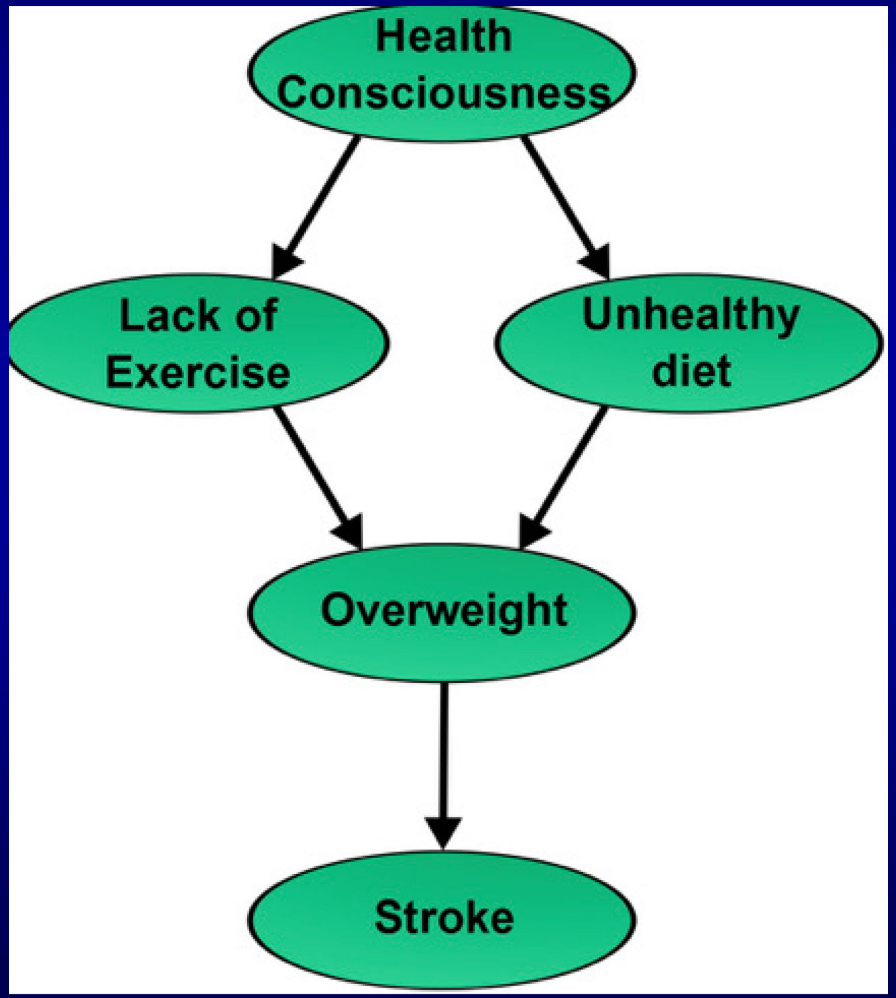
Niveles de Causalidad [The Book of Why]

- Asociación – detectar regularidades (animales, computadoras) $P(Y \mid X)$
  - Probabilidad de ataque cardiaco dada obesidad
- Intervención – predecir el efecto de acciones (pocas especies) $P(Y \mid DO(X=x1))$
  - Probabilidad de ataque cardiaco si se le administra cierta medicina
- Contrafactuales – imaginar, introspección (humanos) $X=x1,\ P(Y \mid DO(X=x2))$
  - Murió de un ataque cardiaco, hubiera muerto si le hubieramos administrado la medicina

Redes Bayesianas Causales

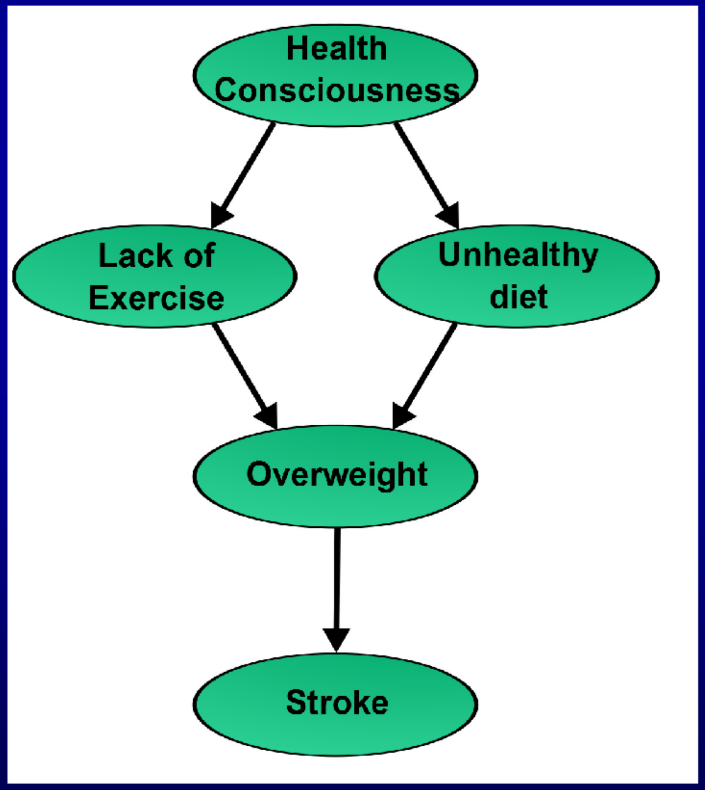
- Modelo gráfico que representa relaciones causales (arcos) entre variables (nodos)
- Implica suposiciones más fuertes que las redes bayesianas (causalidad vs. dependencias o asociaciones)
- Permiten hacer razonamiento causal – intervenciones y cointrafactuales

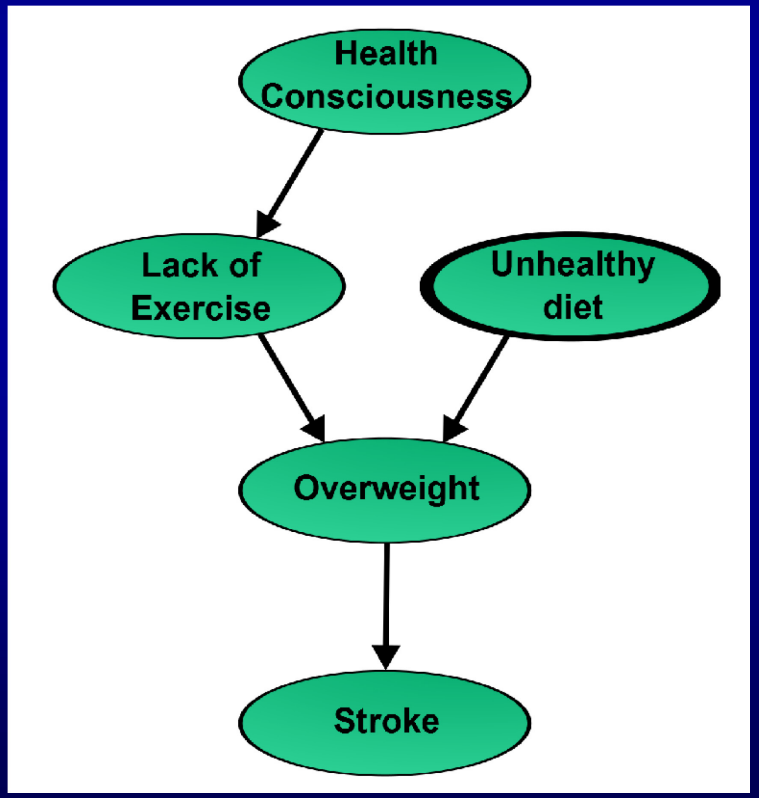
Ejemplo


Health Consciousness
Lack of Exercise
Unhealthy diet
Overweight
Stroke

Predicción - ejemplo

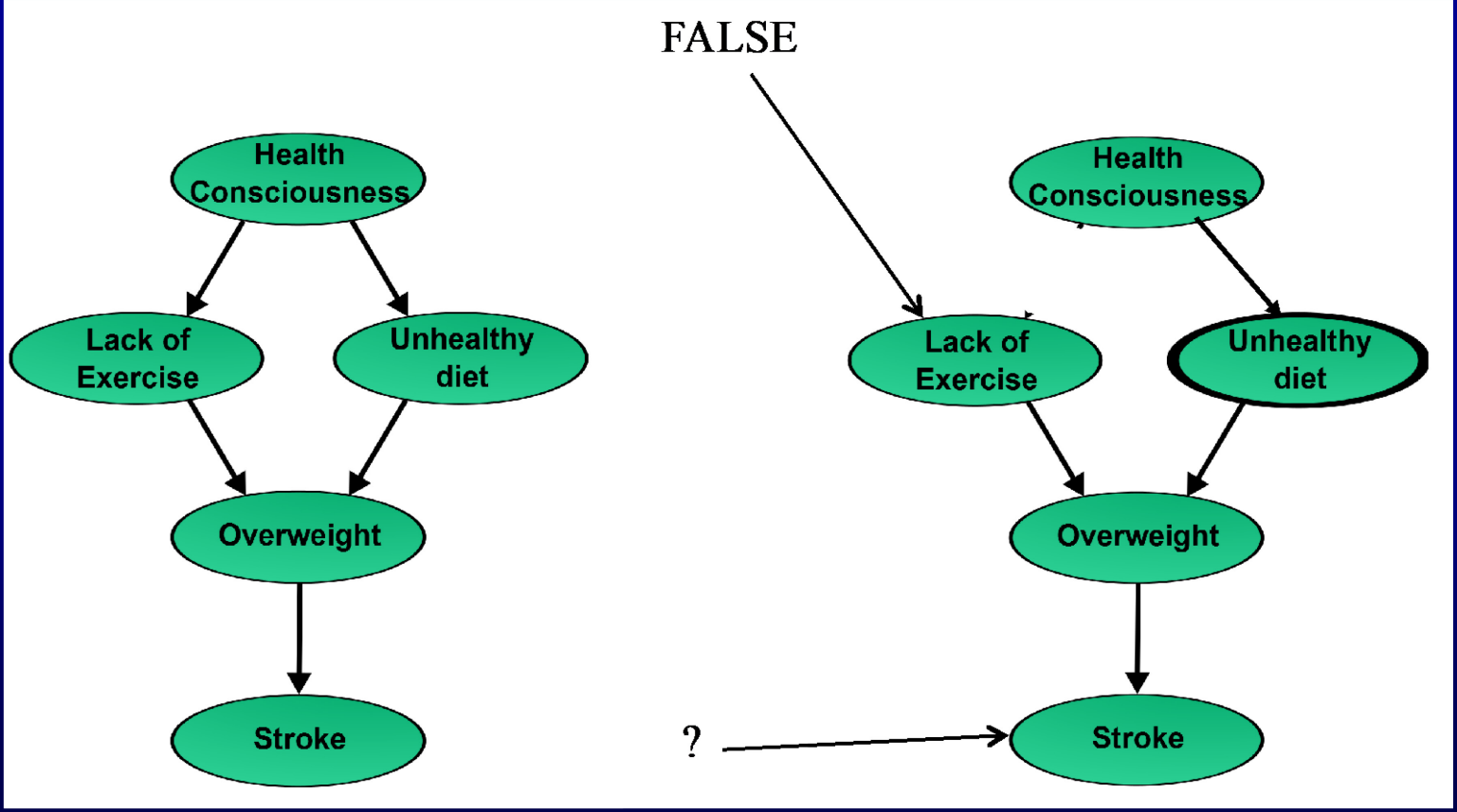
- Predecir el efecto de una dieta no saludable





Contrafactuales - ejemplo

- ¿Hubiera sufrido el EVC si hubiera hecho más ejercicio?

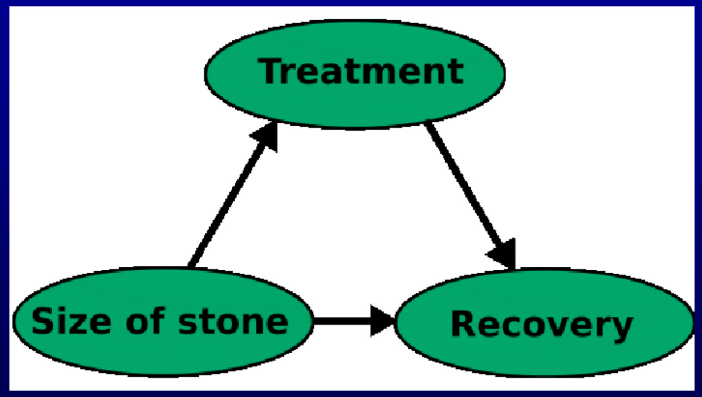


Comentarios …

- Es diferente el resultado de intervenir que el de observar (el modelo se modifica y por lo tanto la distribución de probabilidad)
- Puede haber cofactores que afecten la predicción, estos se deben tomar en cuenta de otra forma el resultado puede ser engañoso

Paradoja de Sympson

Stone Size	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Both	78% (273/350	83% (289/350)



Descubrimiento Casual

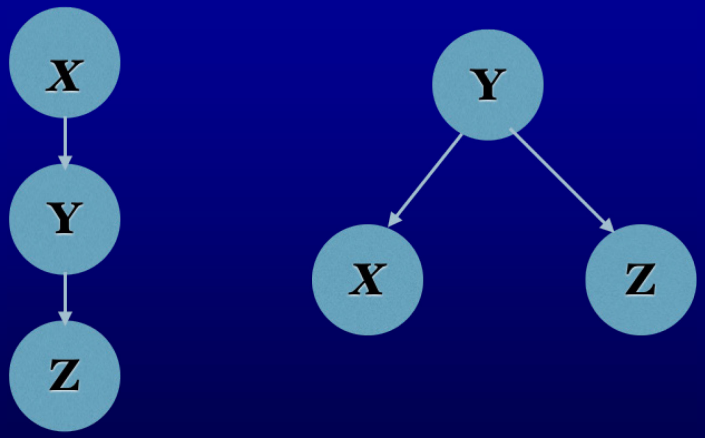
- Idealmente para aprender relaciones causales debemos hacer experimentos – intervenciones
- Por ejemplo para saber si una vacuna *causa* inmunidad al Covid, se tienen que aplicar a muchos individuos (cuidando de los demás co-factores – género, edad, …) y comparar contra el no aplicarla (placebo)

Descubrimiento Casual

- No siempre es posible hacer experimentos – costos, cuestiones éticas, factibilidad (*¿fumar causa cáncer?*)
- En cambio los datos observacionales son más fáciles de obtener y abundantes en muchos dominios – ¿podemos aprender modelos causales a partir de datos observacionales?

Descubrimiento de datos observacionales

- Correlación no implica causalidad (ejemplo del chocolate)
- En base a estadísticas en general no obtenemos un modelo único – varios modelos equivalentes en cuanto a las relaciones de independencia que representan:

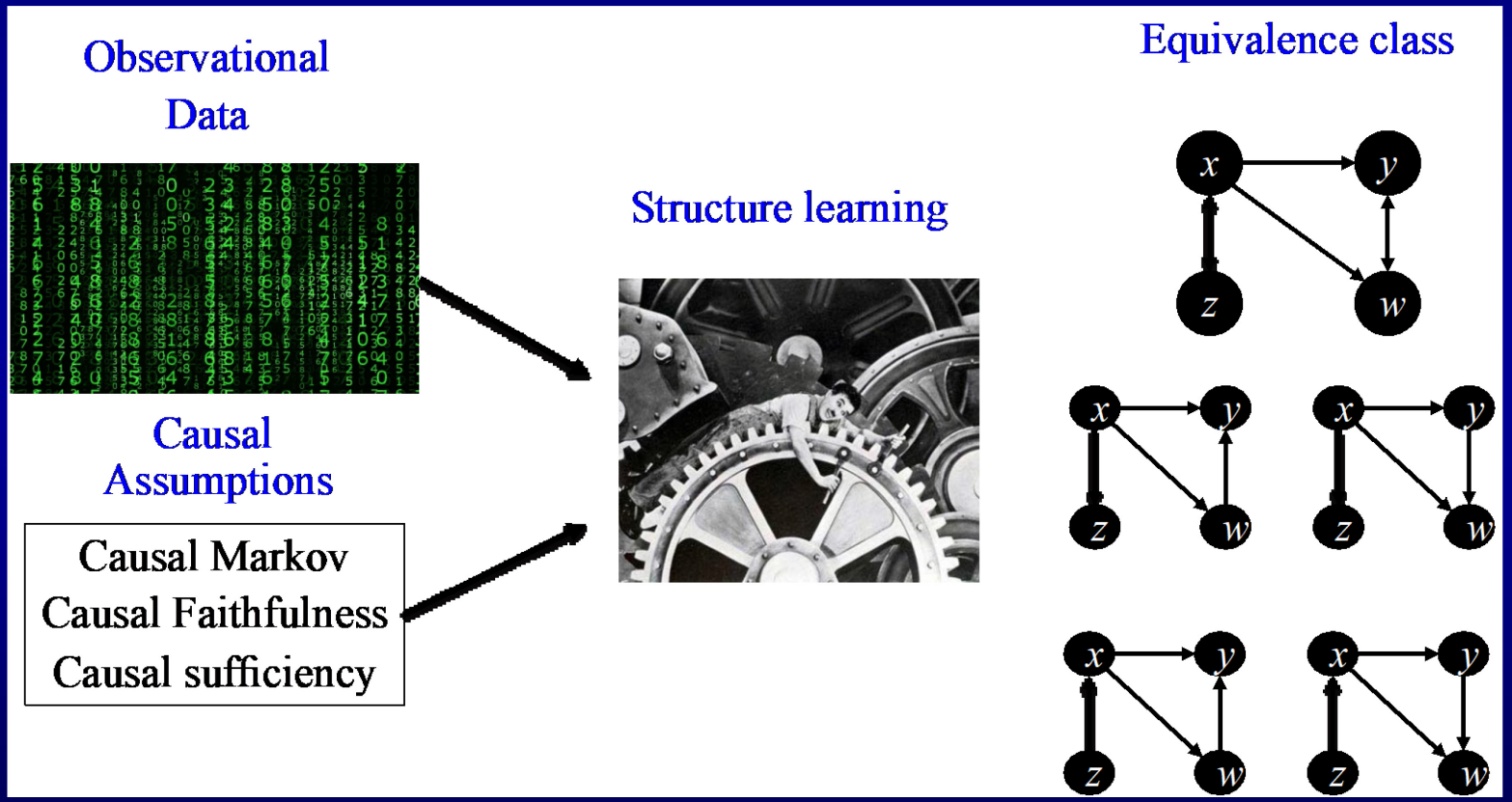


x

Aprendizaje causal

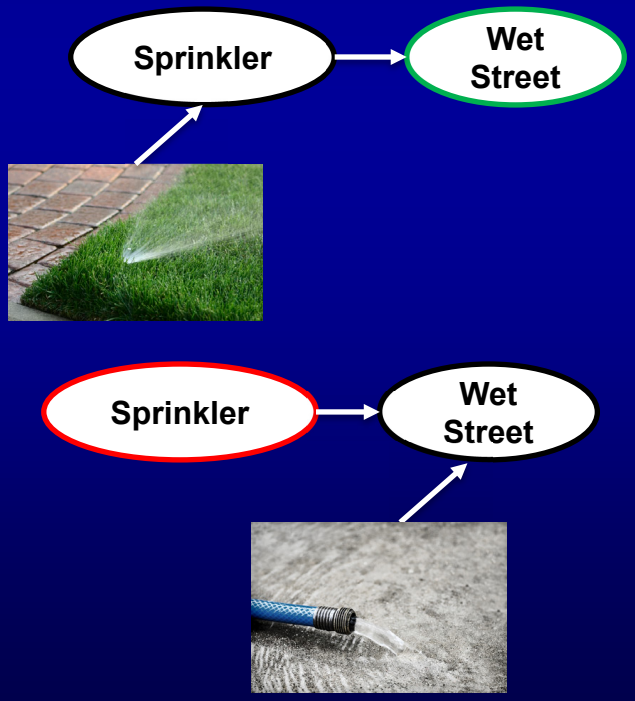
- Para aprender de datos observacionales:
  - Incluir una serie de suposiciones (suficiencia causal, modelo markoviano, …)
  - Asumir cierto tipo de distribuciones (modelos lineales gaussianos)
  - Conocimiento previo
  - Realizar algunas intervenciones

Aprendizaje de una “clase de equivalencia”



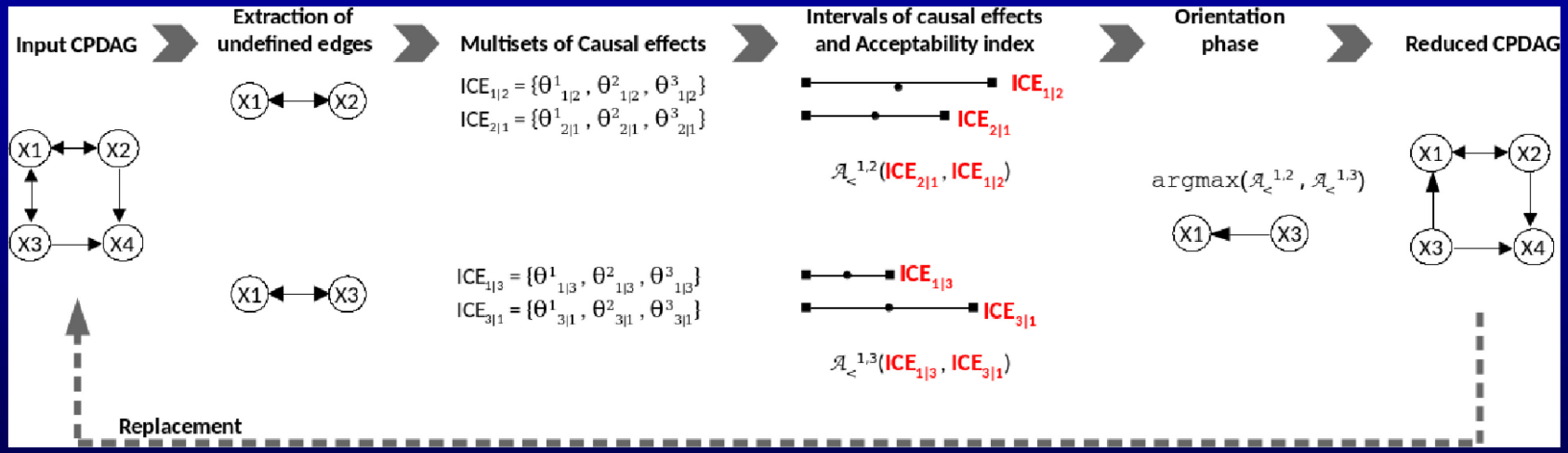
Cálculo de Efectos Causales

- Intervenir una variable y estimar el efecto (cambios de valor) en otras variables

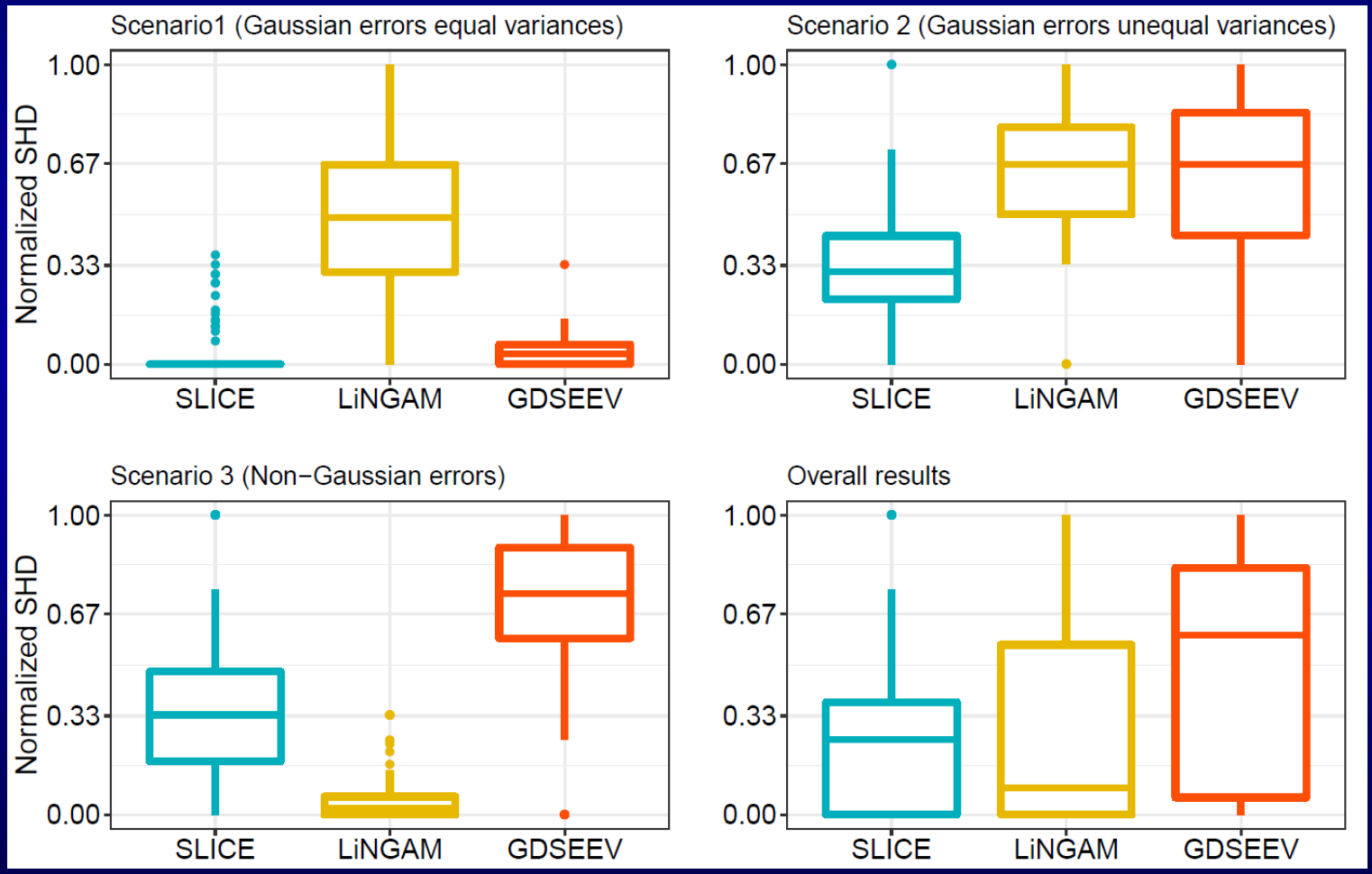


SLICE

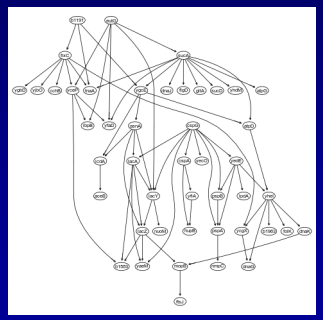
- Partiendo de un modelo parcial, calcula los efectos causales en los diferentes modelos alternativos (obtiene un intervalo de efectos)
- Comparando los intervalos se pueden estimar las relaciones indefinidas



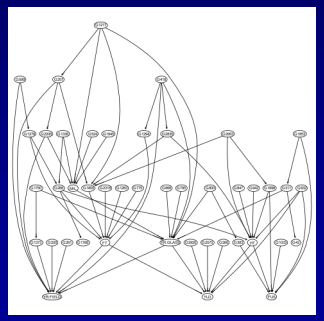
Resultados con modelo sintéticos (distancia al modelo base)



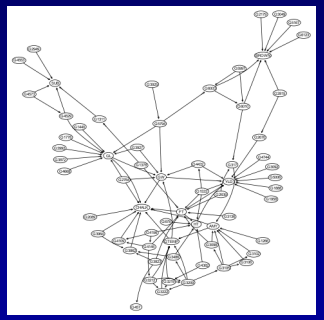
Resultados con modelos “reales”



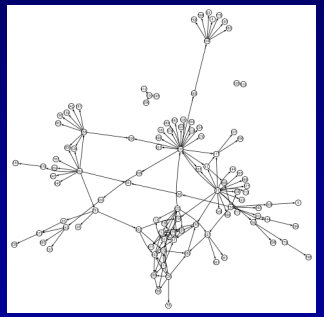
Ecoli70



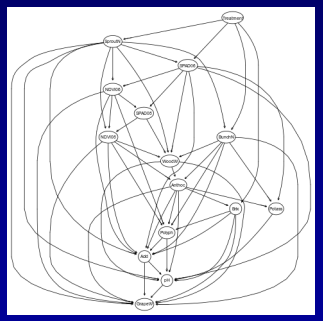
Magic-niab



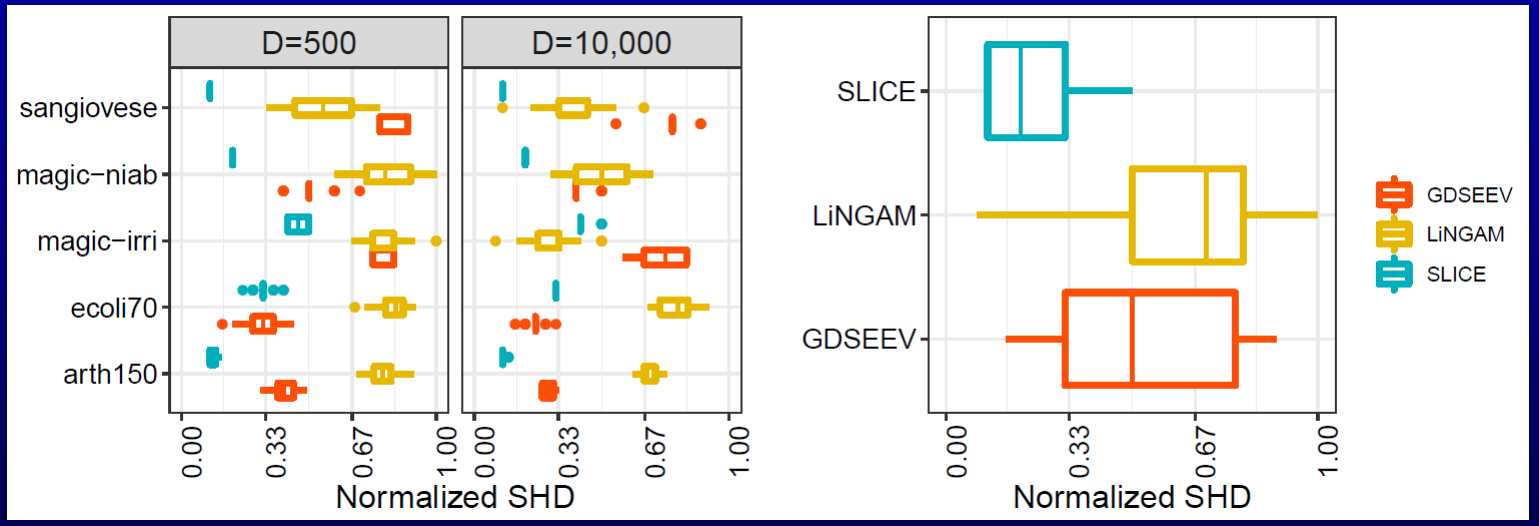
Magic-irri



Arth150



Sangiovese

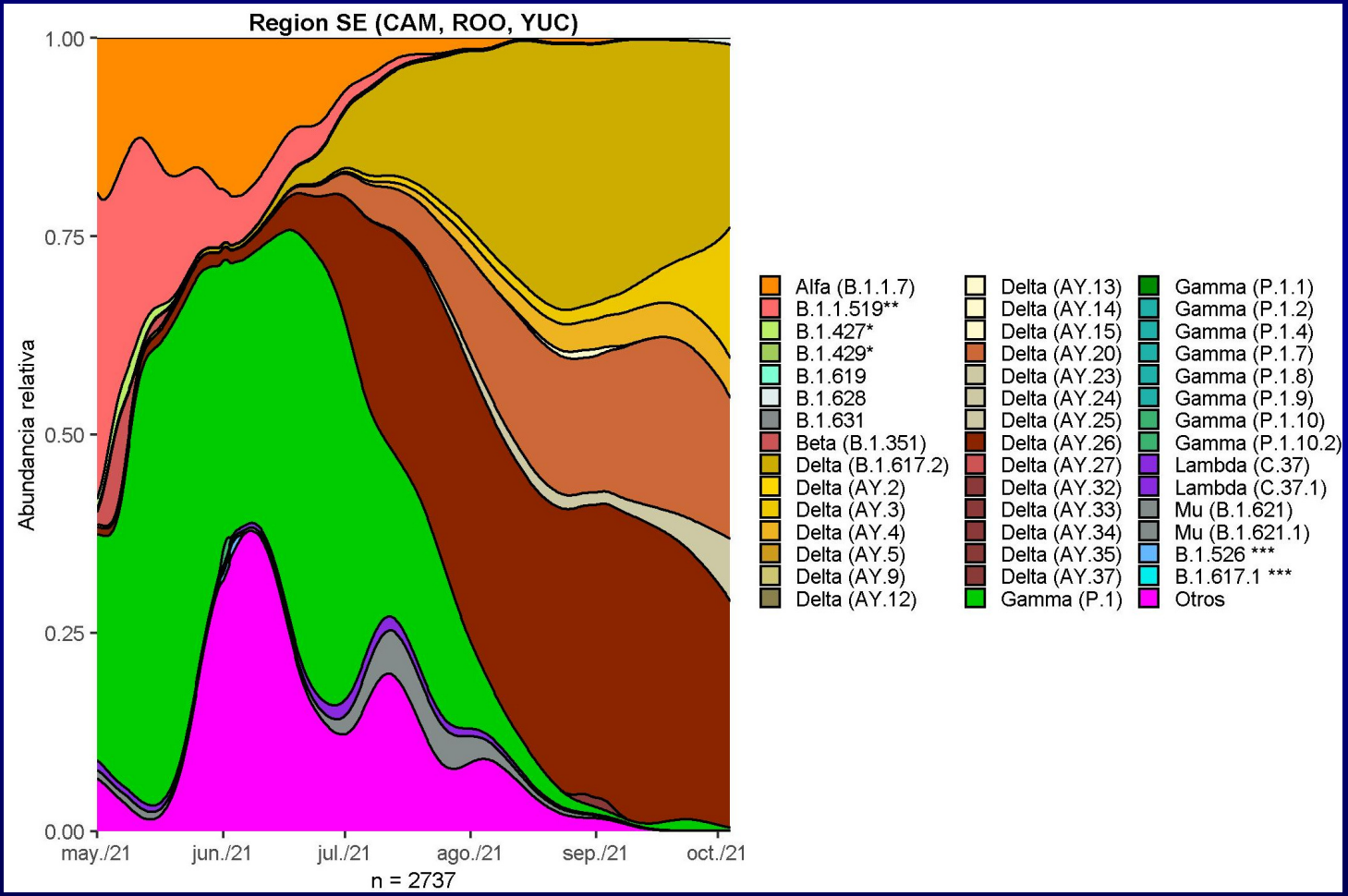


Descubrimiento causal en datos del COVID

Aprendizaje de Modelos Causales

- Aplicamos diversos algoritmos de aprendizaje causal a la BD COVID-19
- Nos enfocamos en la relación de ciertas variables con mortalidad (desenlace) en los datos de CDMX y Yucatán
- Analizamos los datos de las 3 principales etapas (olas) de la pandemia en México, para tratar de entender mejor el fenómeno y las diferencias entre las etapas

Evolución de las variantes de COVID en el sureste de México



Cortesía de IBT-UNAM (Rosa Ma. Gutérrez, Antonio Loza)

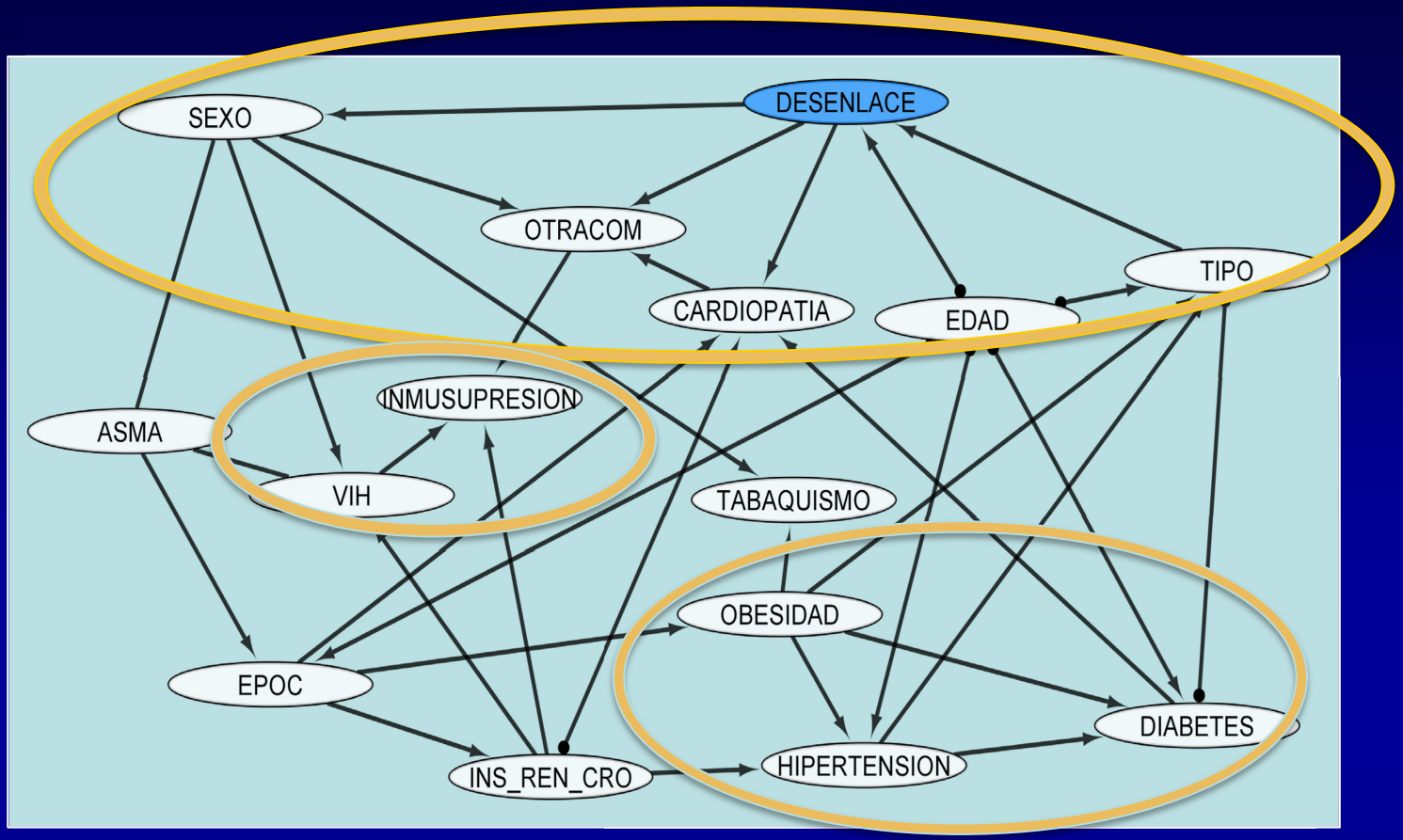
CdMx

VARIABLES:

1. PERIODO={per1, per2, per3}
   - per1: 273517
   - per2 : 80496
   - per3 : 250412
2. SEXO ={MASCULINO, FEMENINO}
3. EDAD={EDAD < 60; EDAD >=60}
4. TIPO ={AMBULATORIO, HOSPITALIZADO}
5. DESENLACE={RECUPERADO, FALLECIDO}
6. DIABETES={0,1}
7. EPOC={0,1}
8. ASMA={0,1}
9. INMUSUPRESION={0,1}
10. HIPERTENSION={0,1}
11. VIH={0,1}
12. OTRACOM={0,1}
13. CARDIOPATIA={0,1}
14. OBESIDAD={0,1}
15. INS_REN_CRO={0,1}
16. TABAQUISMO={0,1}

NUM CASOS : 604425

GFCI



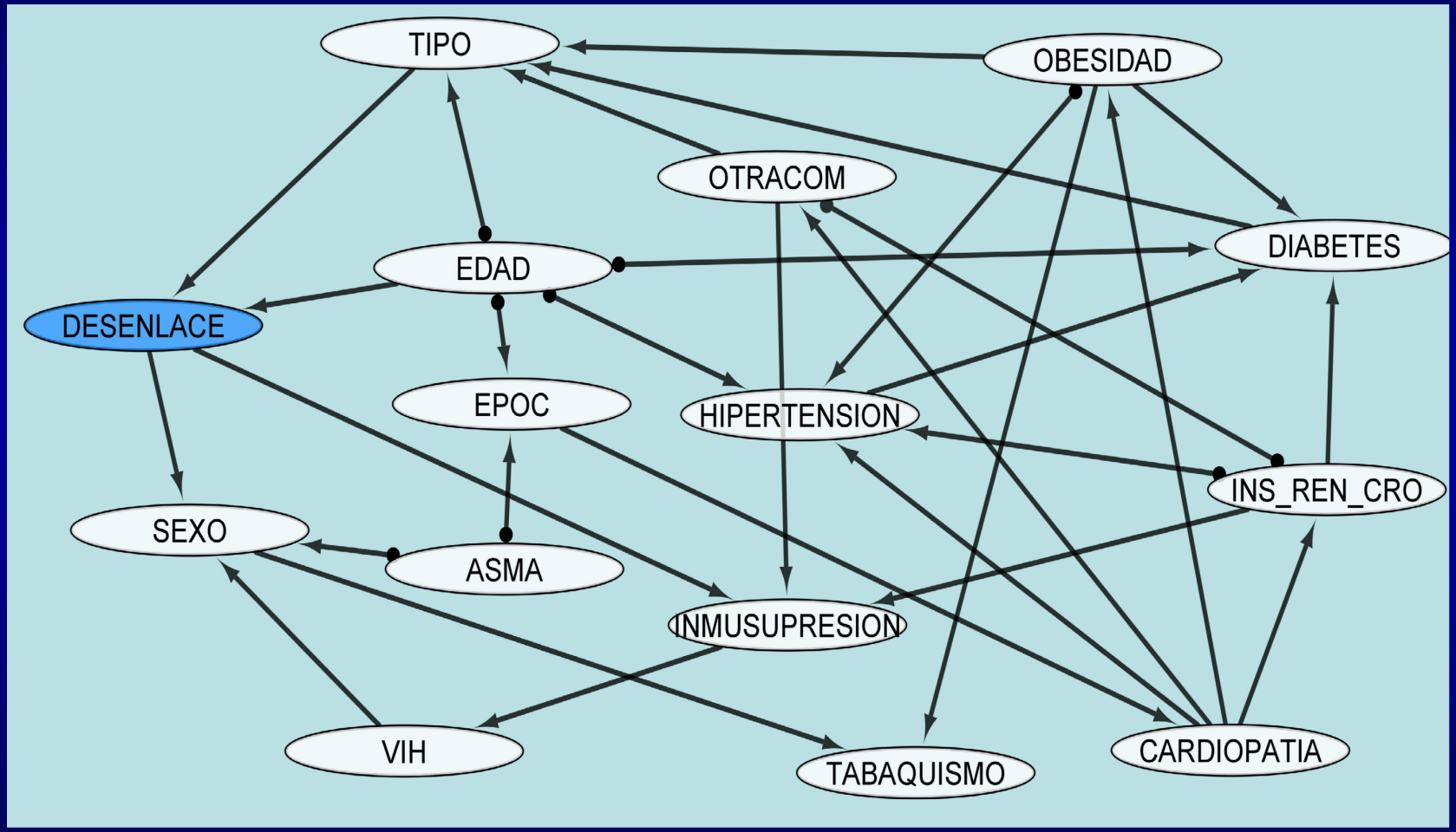
GFCI

- X→Y: X is cause of Y
- X•->Y:
  - X is cause of Y, or
  - hidden common cause of X, Y.
- X•-•Y:
  - X is cause of Y, or
  - Y is cause of X, or
  - hidden common cause of X, Y.

PERIODO 1

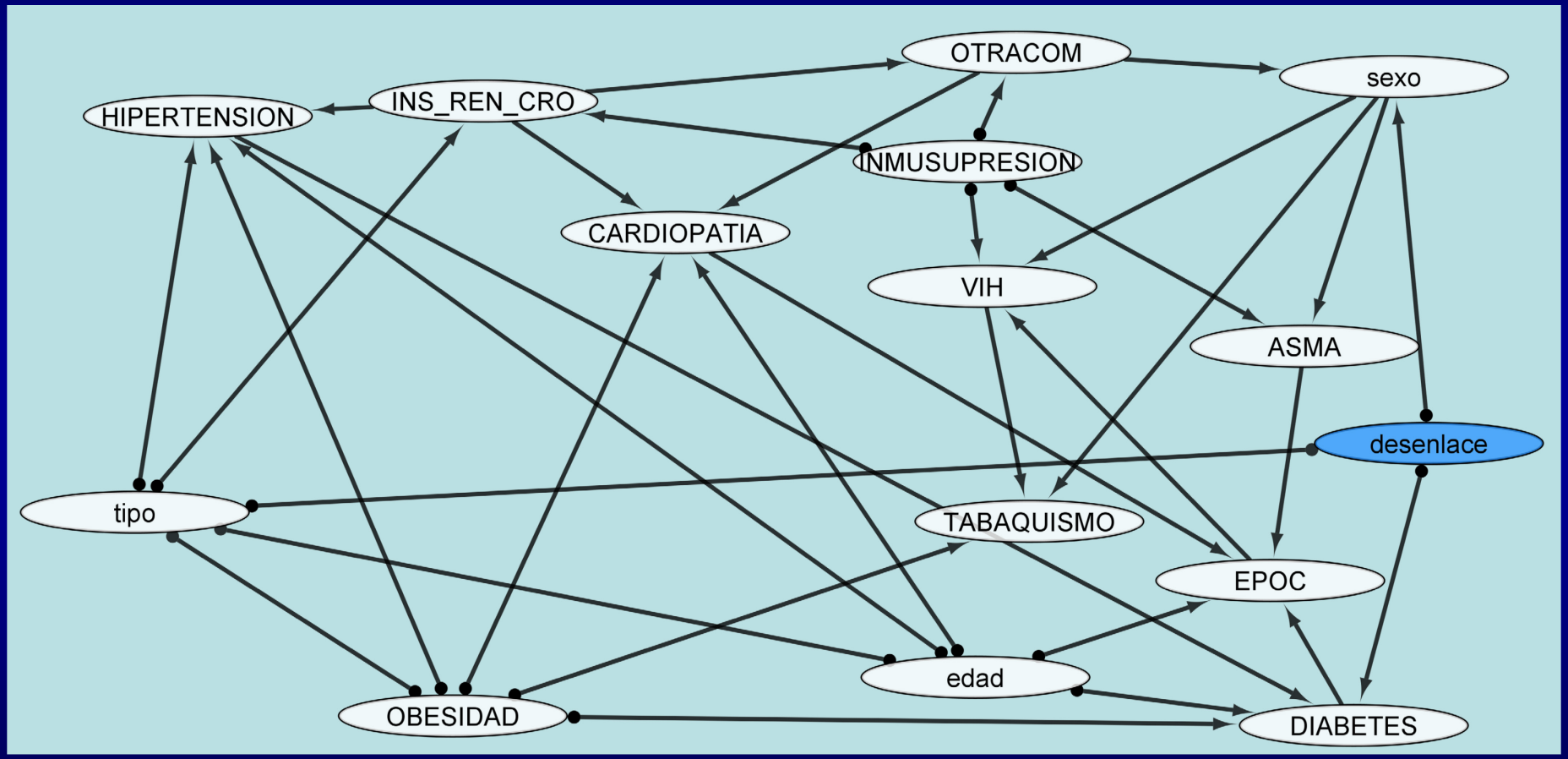
GFCI

PERIODO 2



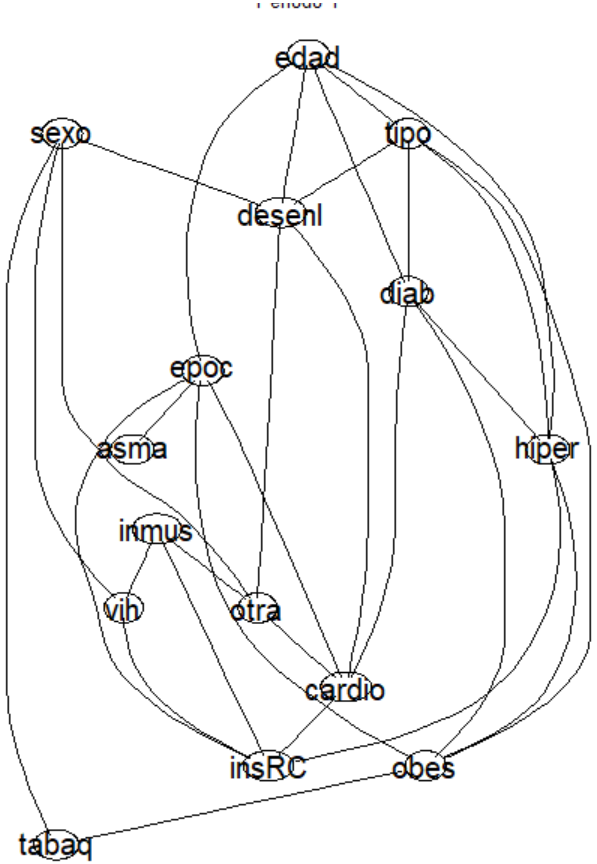
GFCI

PERIODO 3

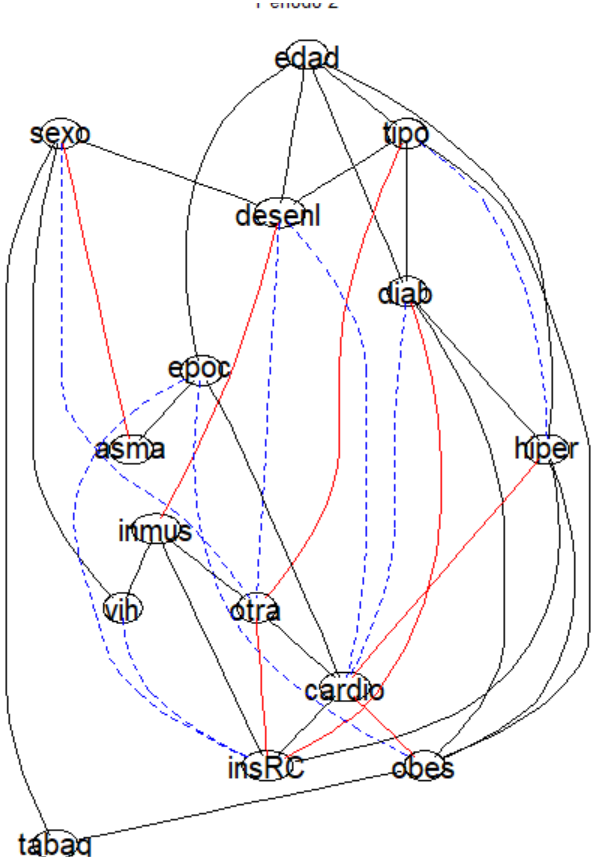


Diferencias entre periodos

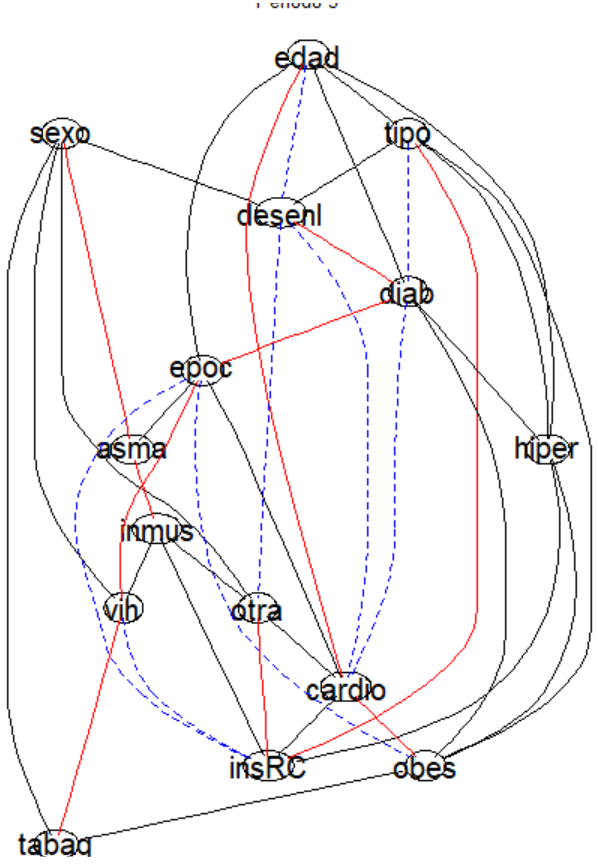
PERÍODO 1



PERÍODO 2



PERÍODO 3



→, —: Enlace que no aparece en el modelo del período 1

---→, ----: Enlace del modelo del período 1 que no aparece en el modelo del período 2 o período 3

Conclusiones

- Las bases de datos son críticas para la investigación, para ello es necesario que se proteja la información sensible y que se realice una buena curación y documentación de la Base de Datos
- La BD Mexicana de COVID-19 es un buen ejemplo – una mina de oro abierta a todos los investigadores

Conclusiones

- Los modelos causales permiten razonar sobre intervenciones y contrafactuales – una herramienta para desarrollar sistemas inteligentes más robustos y explicables
- El descubrimiento causal obtiene relaciones causales y no sólo asociaciones – importante para generar conocimiento útil y para la toma de decisiones

Trabajo Actual y Futuro

- Continuar analizando la BD de Covid para tratar de entender mejor el fenómeno, y eventualmente ayudar a la toma de decisiones
- Incluir la información de la genética del virus
- Aprendizaje de modelos de sujeto / grupo específico (transferencia de conocimiento)

Referencias

- Pearl, J., Mackienze, D.: *The Book of Why*. Basic Books, New York (2018)
- Pearl, J.: *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York (2009)
- Spirtes, P., Glymour, C., Scheines, R.: *Causation, Prediction, and Search*, MIT Press (2000)
- Sucar, L. E, *Probabilistic Graphical Models*, 2nd Edition, Springer Nature, Switzerland (2021) – Chapters 14, 15
- Montero, S.A., Orihuela-Espina, F., Herrera-Vega, J., Sucar, L.E.: *Causal probabilistic graphical models for decoding effective connectivity in functional near infrared spectroscopy*. Flairs, AAAI Press (2016)
- Montero, S.A., Orihuela-Espina, F., Sucar, L.E.: *Intervals of Causal Effects for Learning Causal Graphical Models*, PGM, pp. 296–307, JMLR, 2018.
- Sucar, L.E., Serrano, J., Rodríguez, V., Gutiérrez, R.M., Pineda, L.A., *Prediction and Analysis of COVID-19 with Graphical Models*, Bayesian Modelling Applications Workshop, UAI (2021).

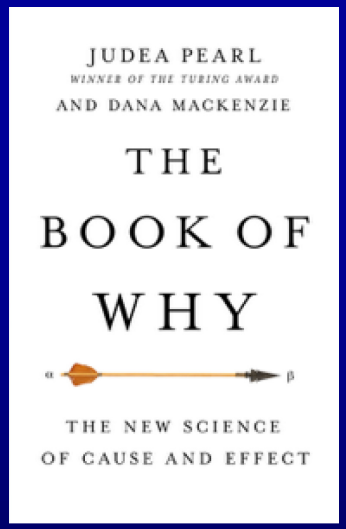
Agradecimientos

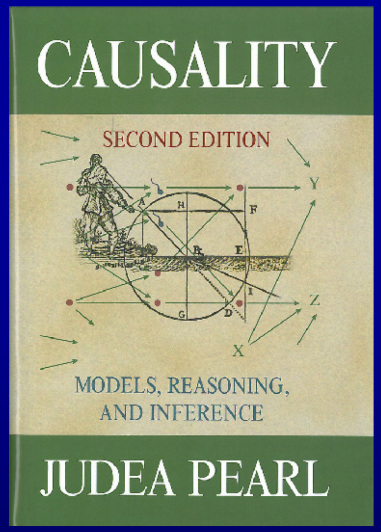
- Rosa María Gutiérrez, Antonio Loza, IBT-UNAM
- Luis Alberto Pineda, Zian Fanti, IIMAS-UNAM

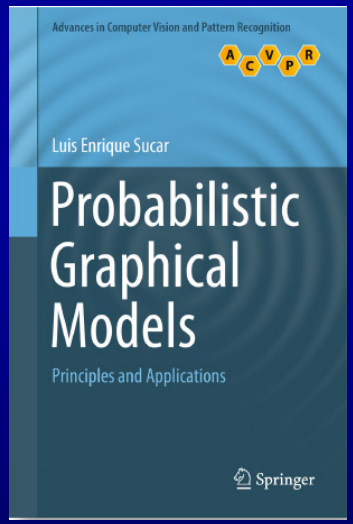
- Felipe Orihuela-Espina, INAOE & U. of Birmingham
- Javier Vega, INAOE & BUAP
- Samuel Montero, University of Houston
- Verónica Rodríguez, PhD student, INAOE

Hasta la vista!

Mi página personal: http://ccc.inaoep.mx/~esucar/







BD COVID-19: http://covid-19.iimas.unam.mx