

Data Science: First things first

FTI UII
13 Januari 2020

Dr. Ing. Ridho Rahmadi, S.Kom., M.Sc.



UNIVERSITAS
ISLAM
INDONESIA

- Head of Center of Data Science, UII
- Head of Research Laboratory, Informatics, UII
- Education
 - Universitas Islam Indonesia (S1)
 - Czech Technical University in Prague (S2)
 - Johannes Kepler University, Austria (S2)
 - Radboud University Nijmegen, the Netherlands (S3)
 - Carnegie Mellon University, USA (Visiting scholar)
- Research interest
 - Machine learning, deep learning, causal modeling, stability selection, multi-objective evolutionary algorithms
- ridho.rahmadi@uii.ac.id
- 081129513045

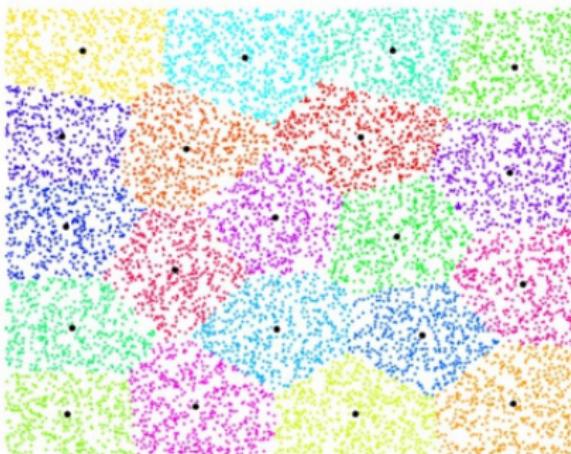
Data science

Data science is **to seek** values from data, and **to orchestrate** them into **meaningful models**

It is applied to many domains

Values of data

4 | 45



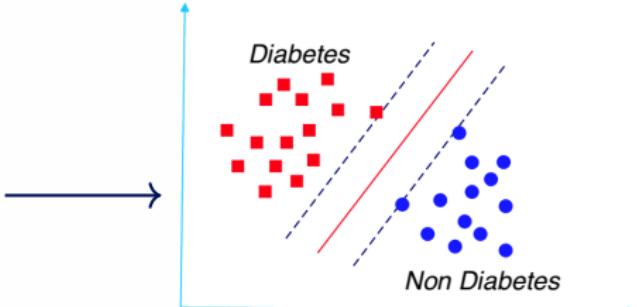
- patterns
- clusters
- classes
- principal components
- independent components

Orchestrated values yield models

5 | 45



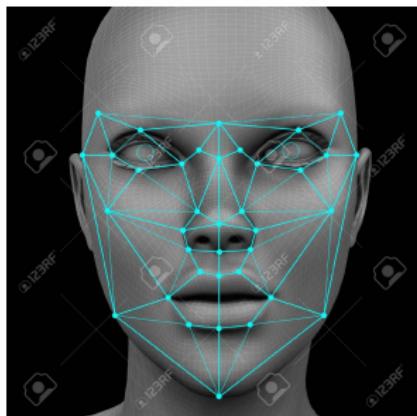
Clinical data



diagnosis models

Orchestrated values yield models

6 | 45



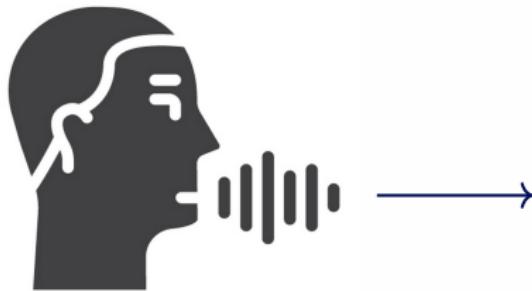
Biometric data



Phone unlocking models

Orchestrated values yield models

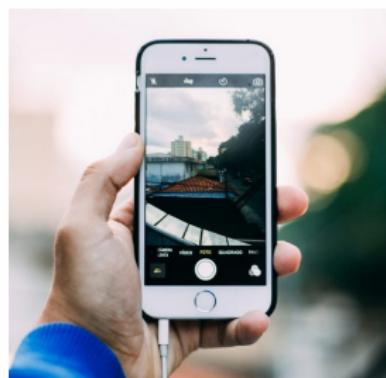
7 | 45



VectorStock®

VectorStock.com/15514684

Speech data



Taking picture models

Orchestrated values yield models

8 | 45



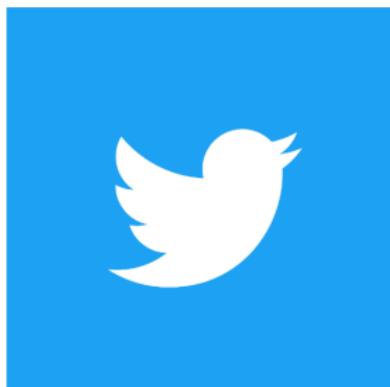
Retail data



Inventory models

Orchestrated values yield models

9 | 45



Tweets data



VectorStock®

VectorStock.com/T160321#

Hoax models

Orchestrated values yield models

10 | 45



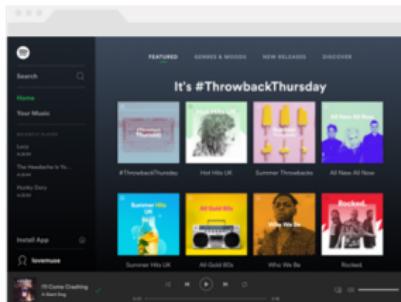
Network data



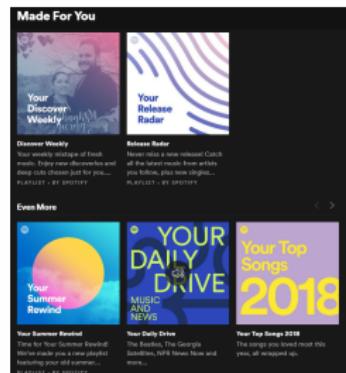
Intruder models

Orchestrated values yield models

11 | 45



Spotify browsing data



Playlist models



- Redefining data...
- What are data now and then

source: <http://partnersinmeded.com>

Data have been about
observations

Digitization **transforms** the world
into **computer-readable** data

Redefining Data

15 | 45



- 49K photos in Instagram
- 3.9M Google searches
- 4.3M Youtube watch
- 473K Twitter tweets
- 12.9M text sent
- 750K Spotify stream
- 156M emails sent
- 154K Skype calls



Estimated there are
> 2.500.000.000.000.000.000 bytes
generated per day

How to data science?

- **Problems understanding ←**
- Data understanding (Exploration & Visualization)
- **Modeling methods ←**
- Result interpretation

Problem understanding: Association

19 | 45

Problem	Activity	Questions	Examples
Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?

What is ...	the price of a house the status of a patient the market sentiment	based on its area seeing her/his clinical examination considering current political situation
-------------	---	---

P

y

x

Problem understanding: Intervention

21 | 45

Problem	Activity	Questions	Examples
Intervention $P(y do(x), z)$	Doing, Intervening	What if? What if I do $X = x$?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?

Problems of Intervention $P(y|do(x), z)$ 22 | 45

What if ...	will their headache be cured?	we give patients paracetamol
	will their disease be healed?	we give patients a treatment
	will I loose some weight?	I run 3km per day

P

y

$do(x)$

Note: read in the order of P , $do(x)$, and then y .

Problem understanding: Counterfactual 23 | 45

Problem	Activity	Questions	Examples
Counterfactual $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? What if I had not been smoking the past 2 years?

Problems of Counterfactual $P(y_x|x', y')$ 24 | 45

What ...

would Adam's status be, had he didn't get the treatment?

would be Hawa's score, had she doubled the study hours?

given that his status is cured and that he underwent a treatment.

given that her score is 60 and that she studied for 3 hours

P

y_x

x', y'

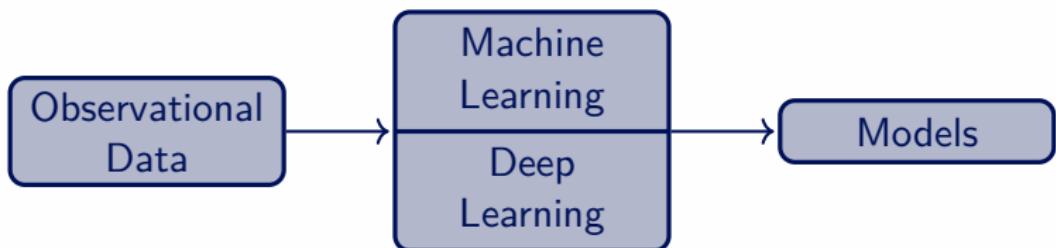
Modeling Methods

Problem	Data	Methods
Association	Observational	Statistics, Machine learning, Deep learning
Intervention	Experimental	Causal Model

Note that we skip the counterfactual case at this moment

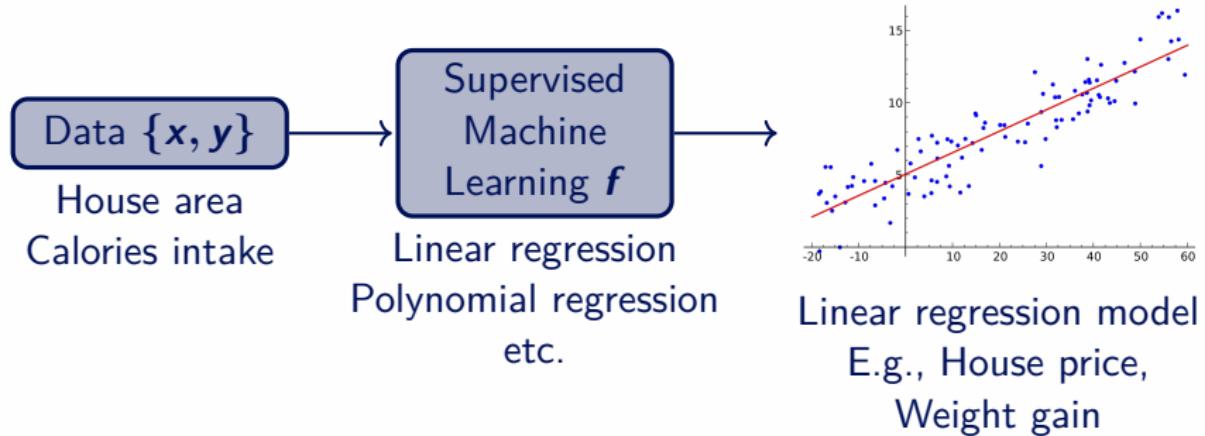
Modeling methods for association

27 | 45



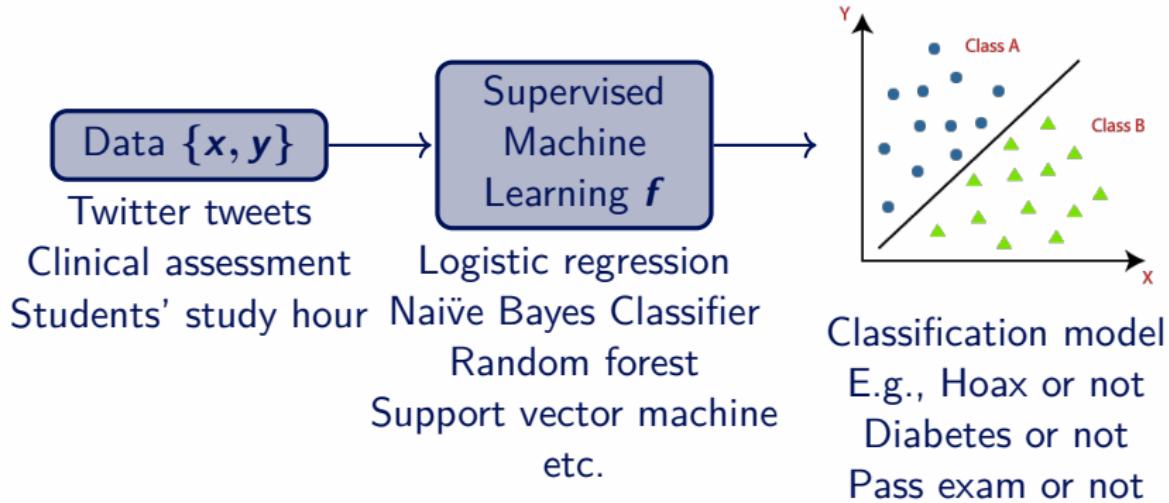


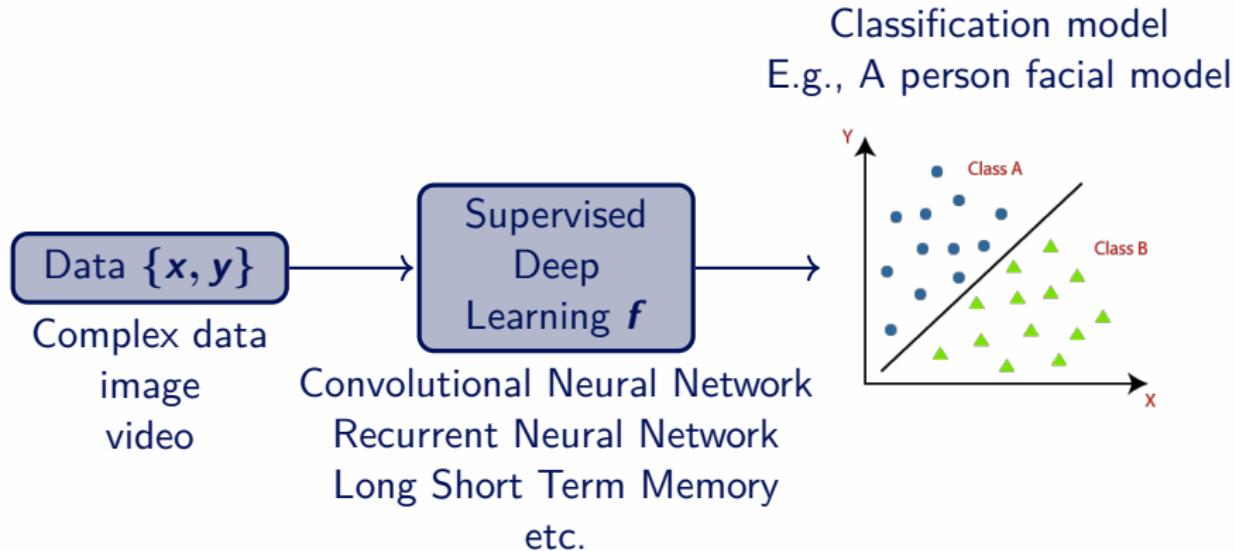
ML or DL is (mostly) a **mathematical function f** that transforms data into a model of interest.

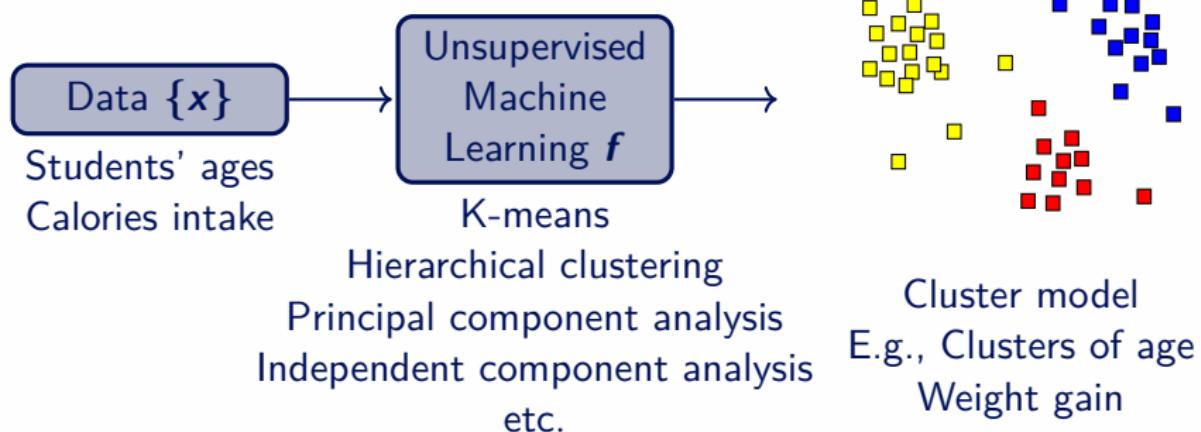


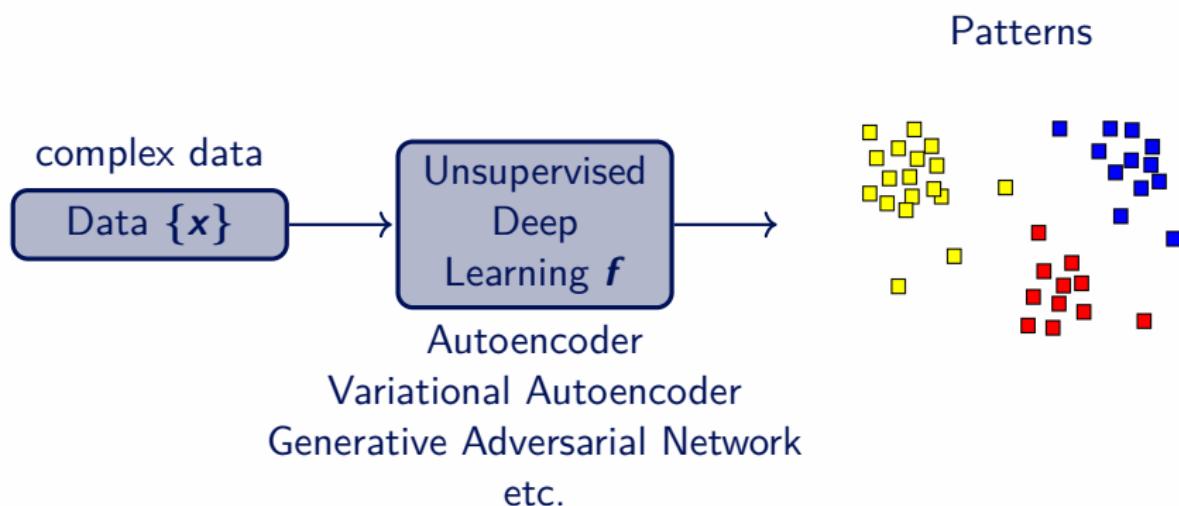
Machine Learning & Deep Learning

30 | 45



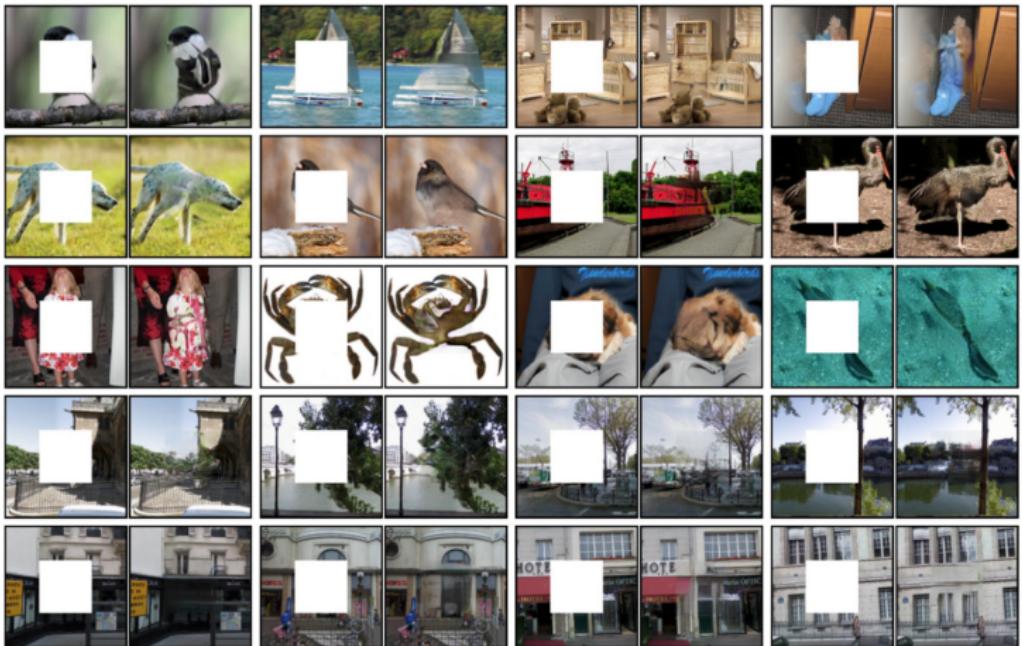






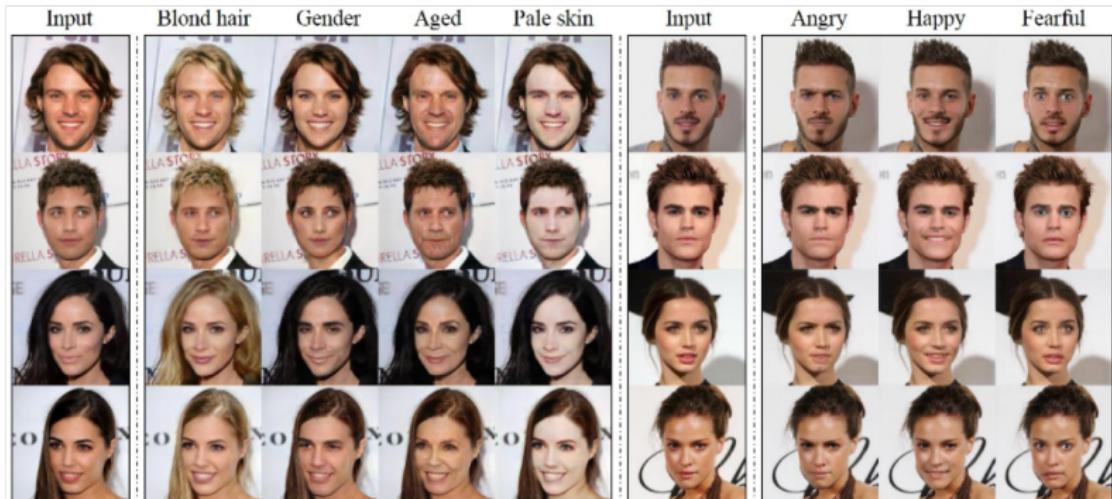
Deep Learning Application

34 | 45



Deep Learning Application

35 | 45



Deep Learning Application

36 | 45

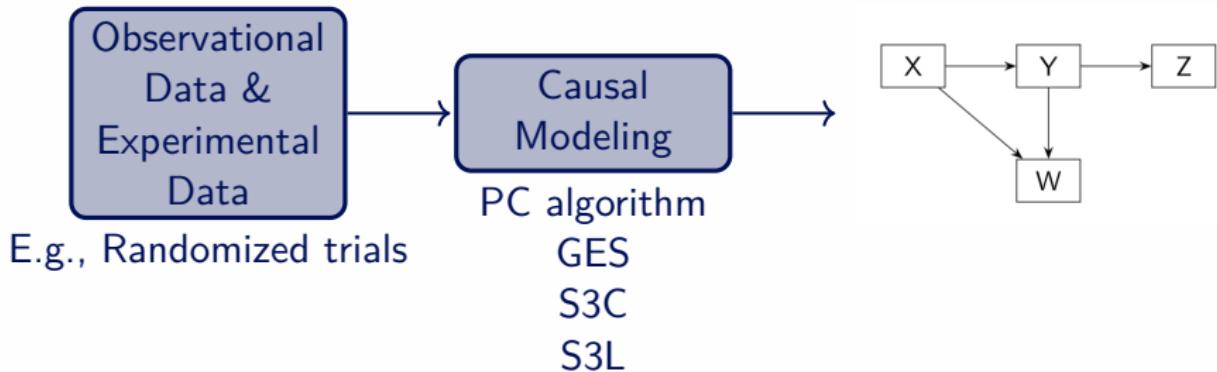


Applications to video (clickable examples):

- Natural environment *reading*
- Face recognition
- Behavior recognition
- Action recognition
- Environmental behavior recognition
- Crowd Learning

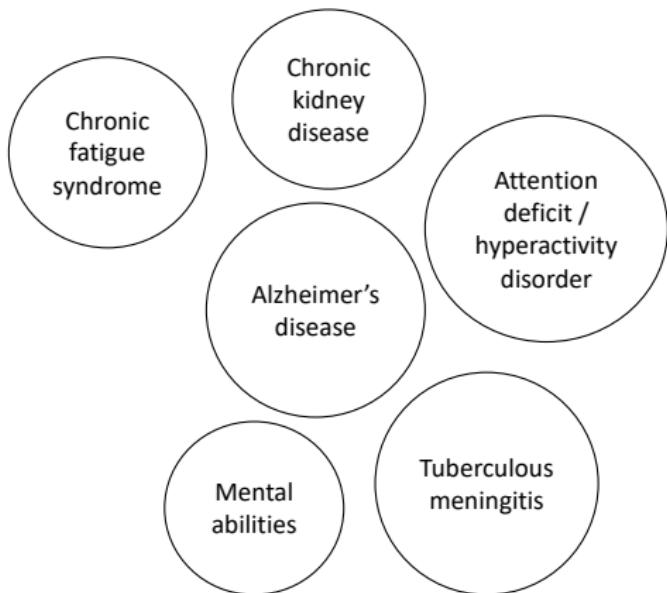
Modeling methods for Intervention

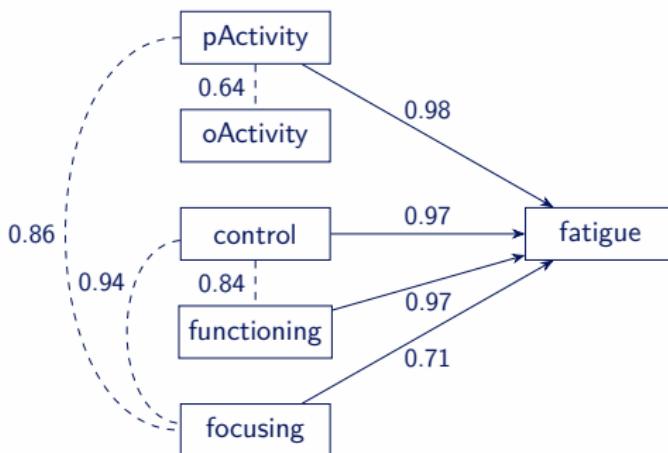
38 | 45



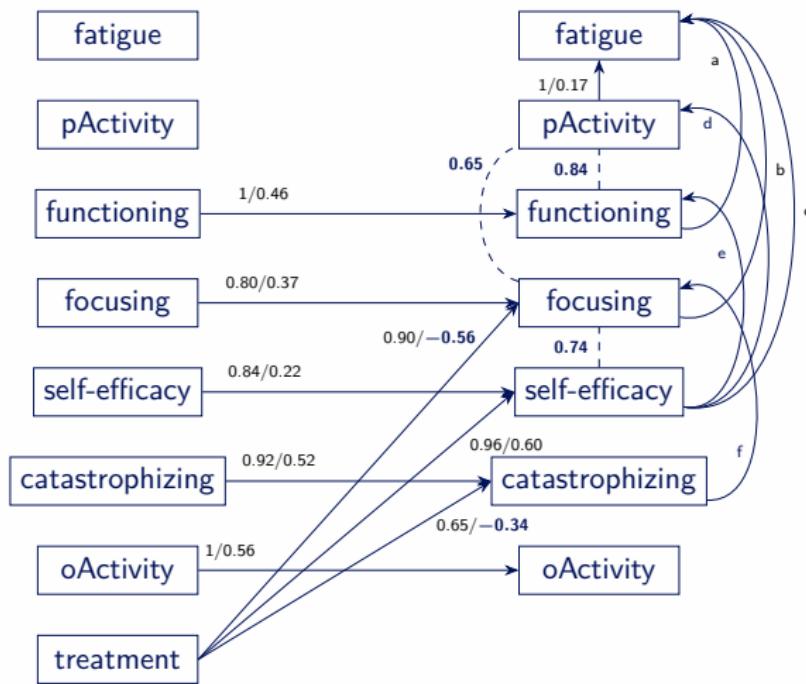
Real-world clinical examples

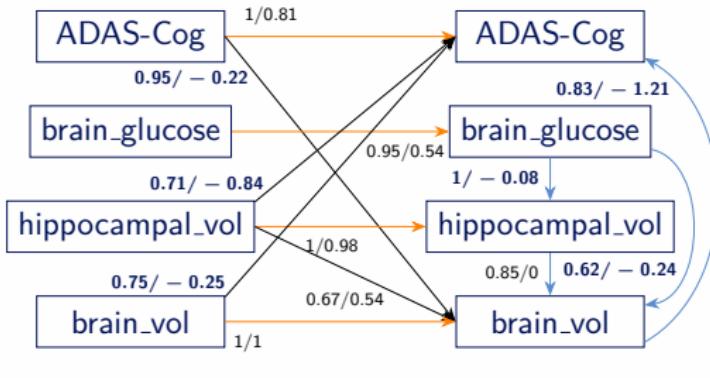
39 | 45



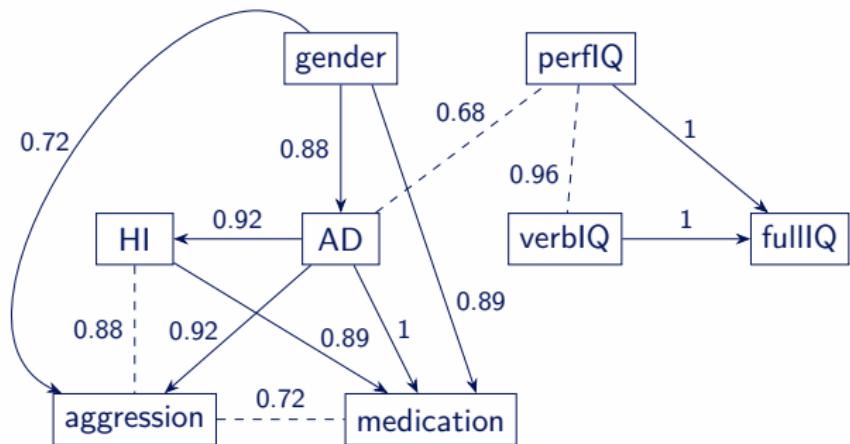


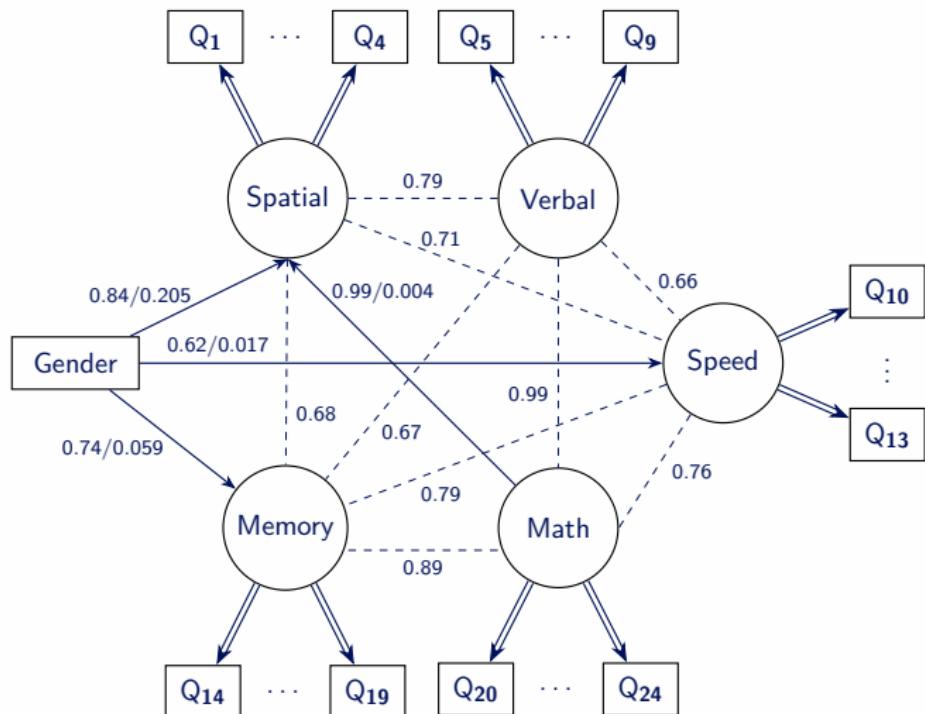
Result on
cross-sectional
CFS data

**Result on
longitudinal CFS
data**



Result on
longitudinal
Alzheimer data

Result on
ADHD data



Result on
Holzinger data

Terima Kasih!
ridho.rahmadi@uii.ac.id
08112951304