Healthcare AI Datathon and MIMIC Dataset An Introduction

Asst. Prof. Mengling 'Mornin' Feng
National University of Singapore
Senior Assistant Director
National University Hospital

mornin@gmail.com
www.mornin-feng.com

Rule 1: Every question is appreciated!

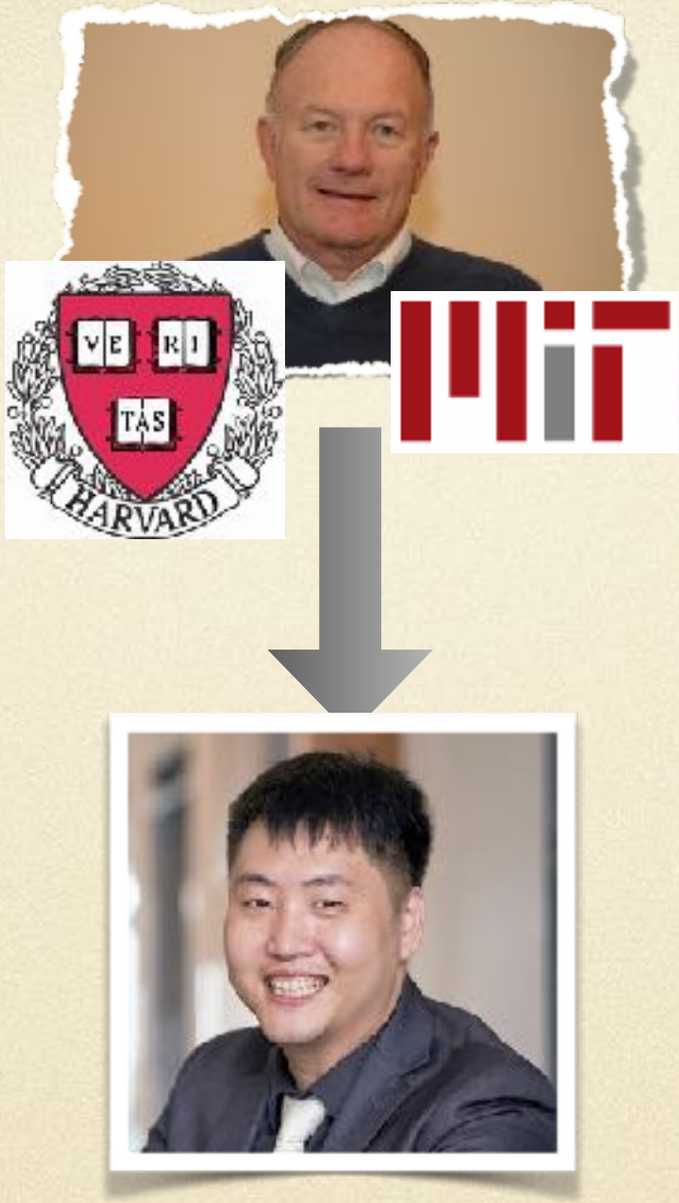
Rule 2: Welcome to stop me anytime if you may have a question!

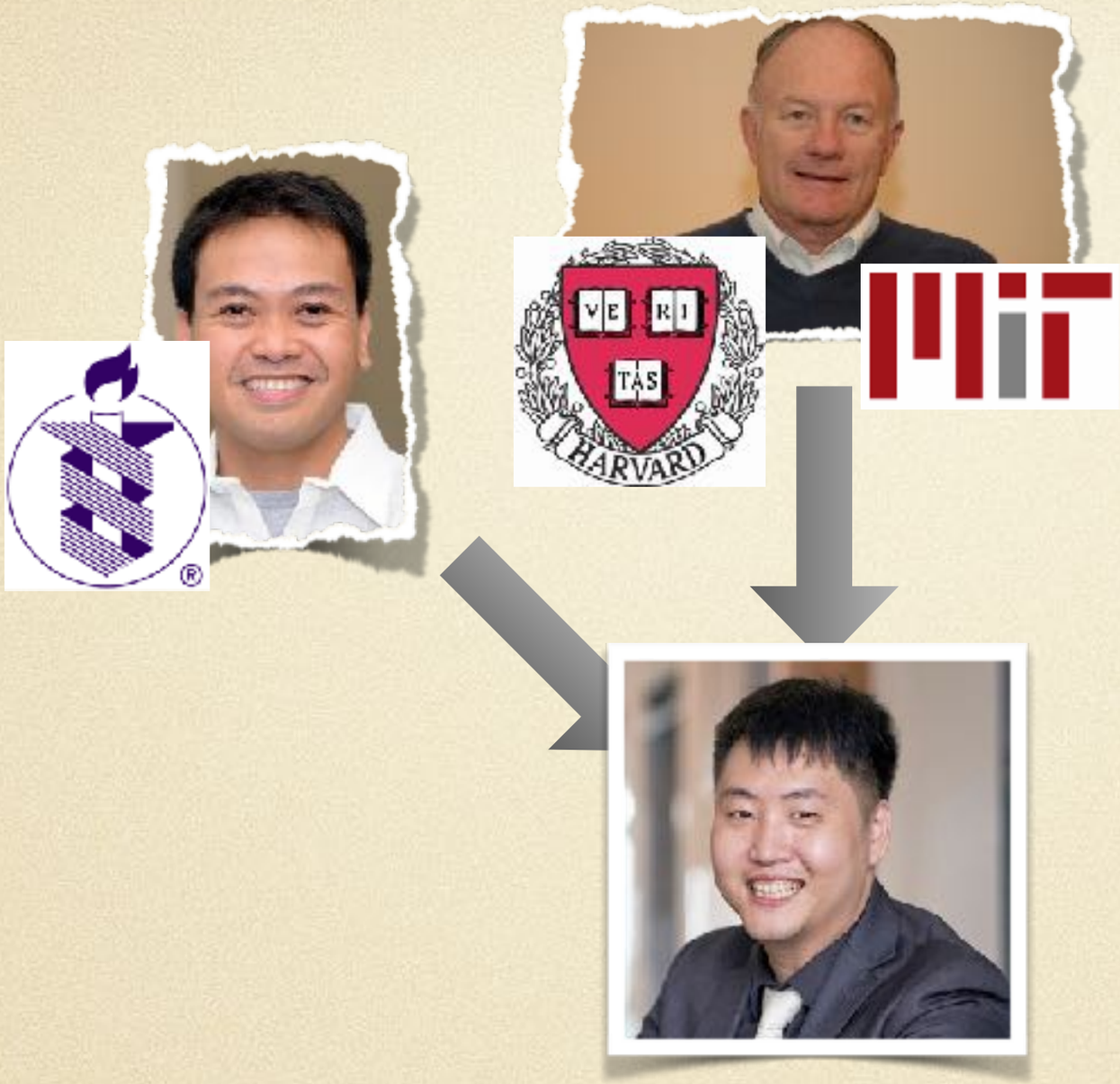
Rule 3: Please practice rule 1 & 2!

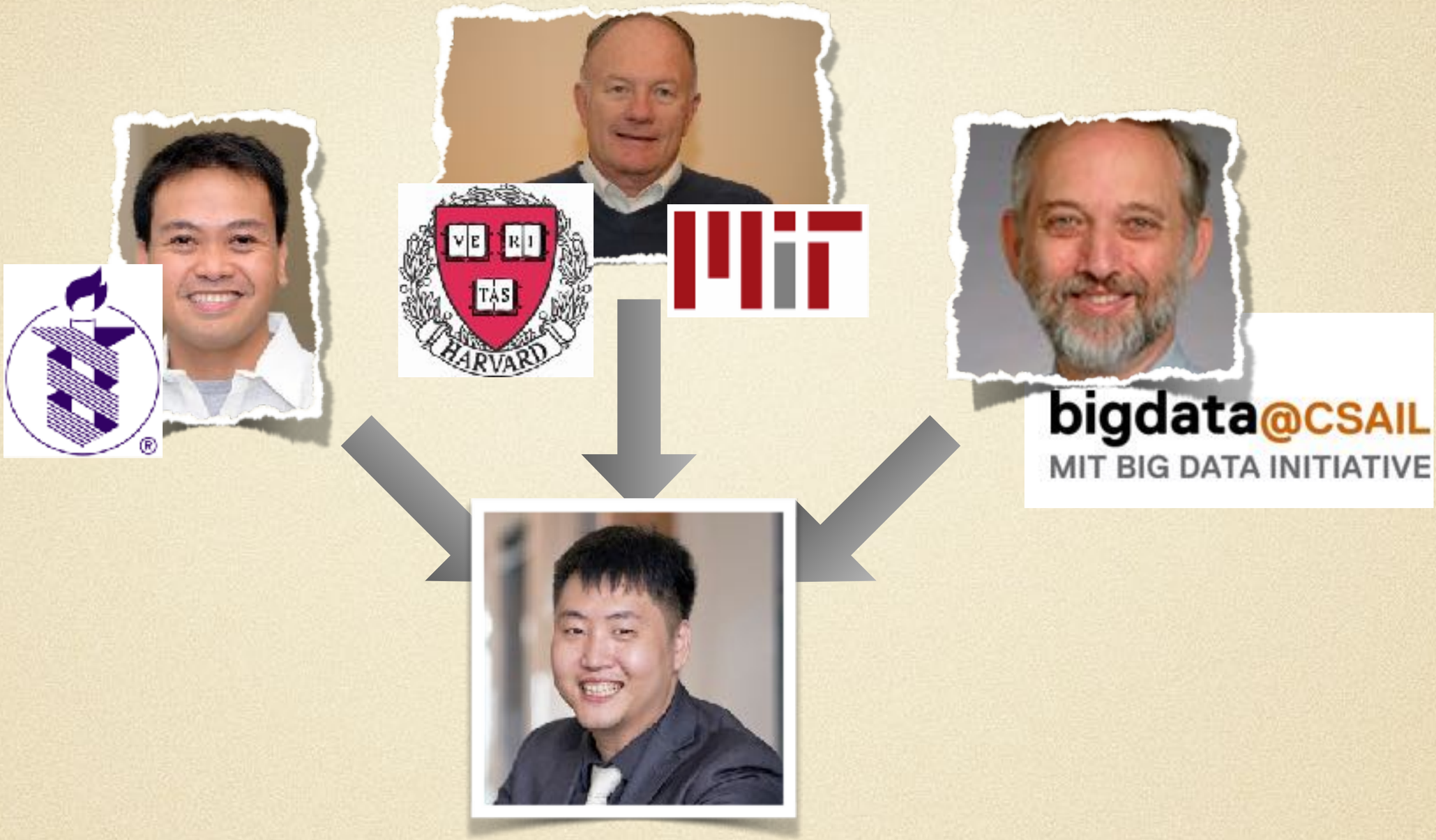
NANYANG TECHNOLOGICAL UNIVERSITY

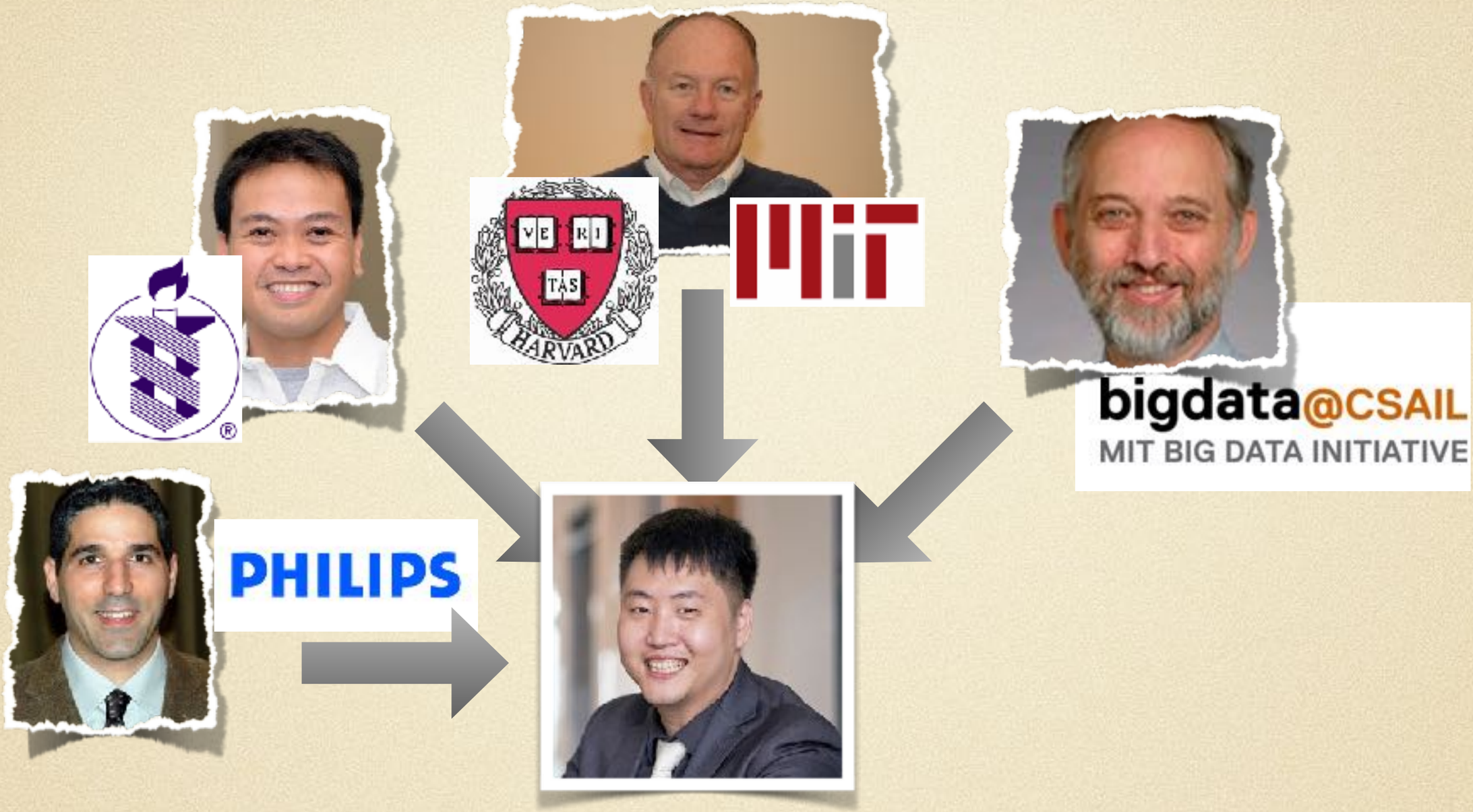
NANYANG TECHNOLOGICAL UNIVERSITY

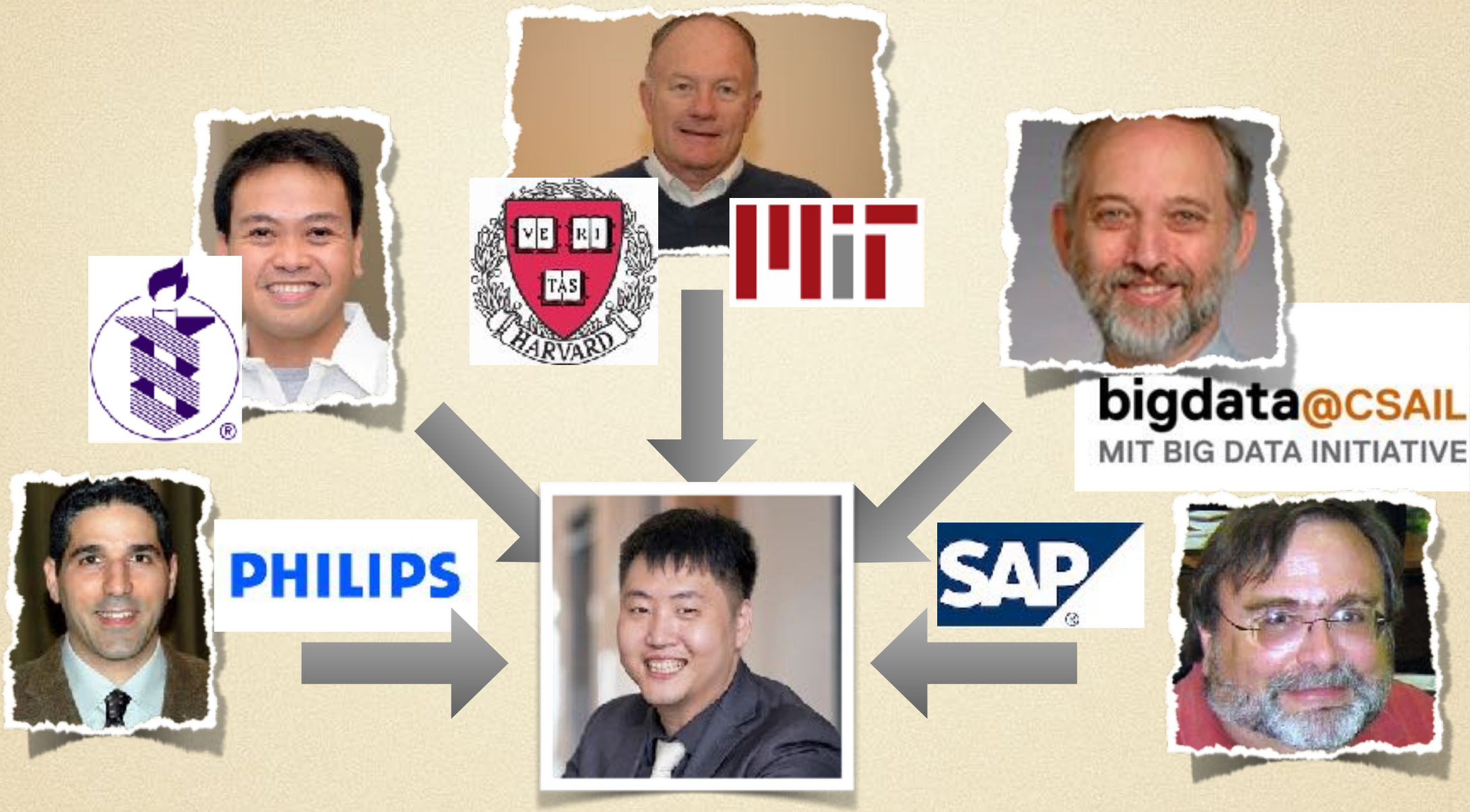
NUS National University of Singapore

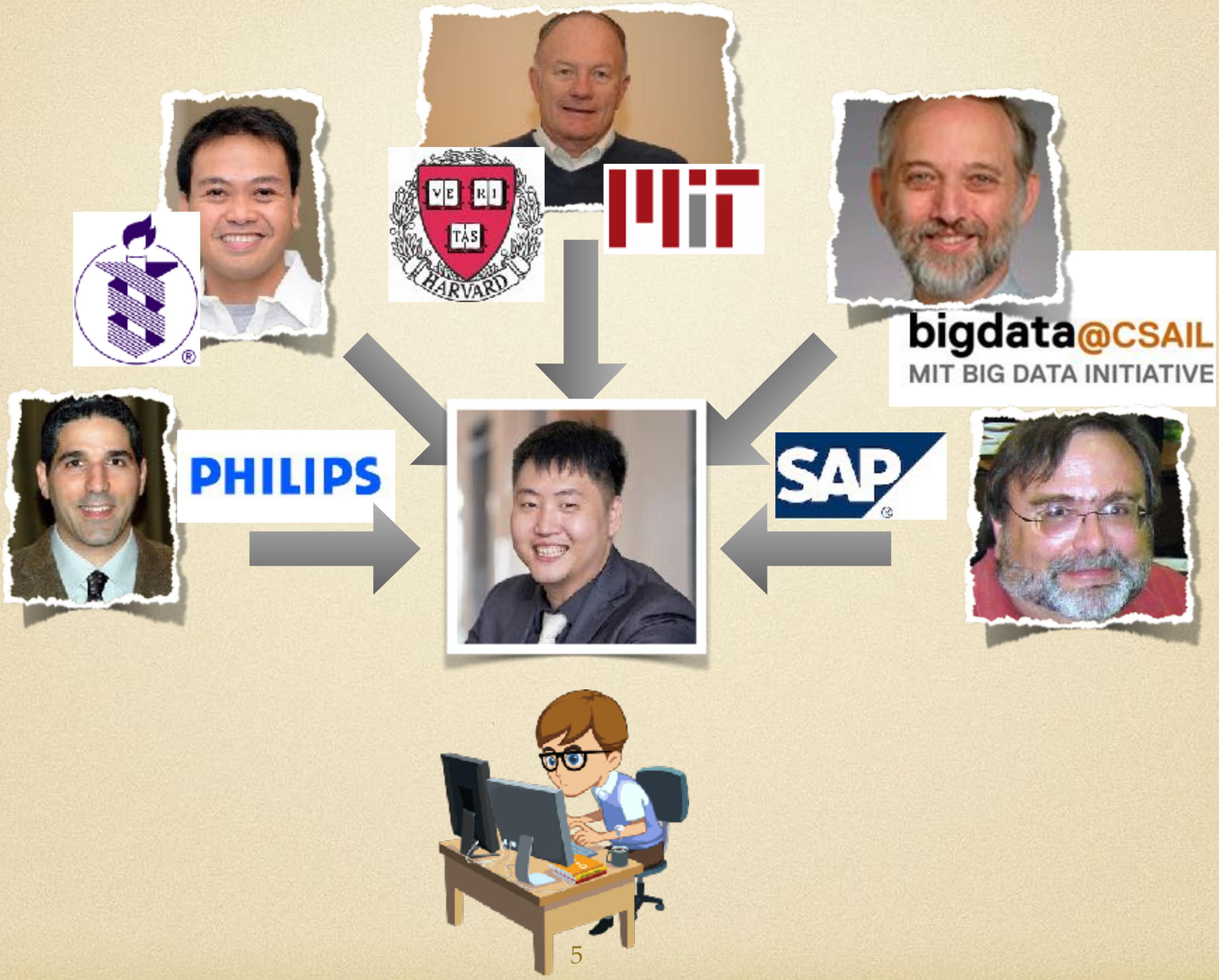

VE RI
TAS
HARVARD
MIT

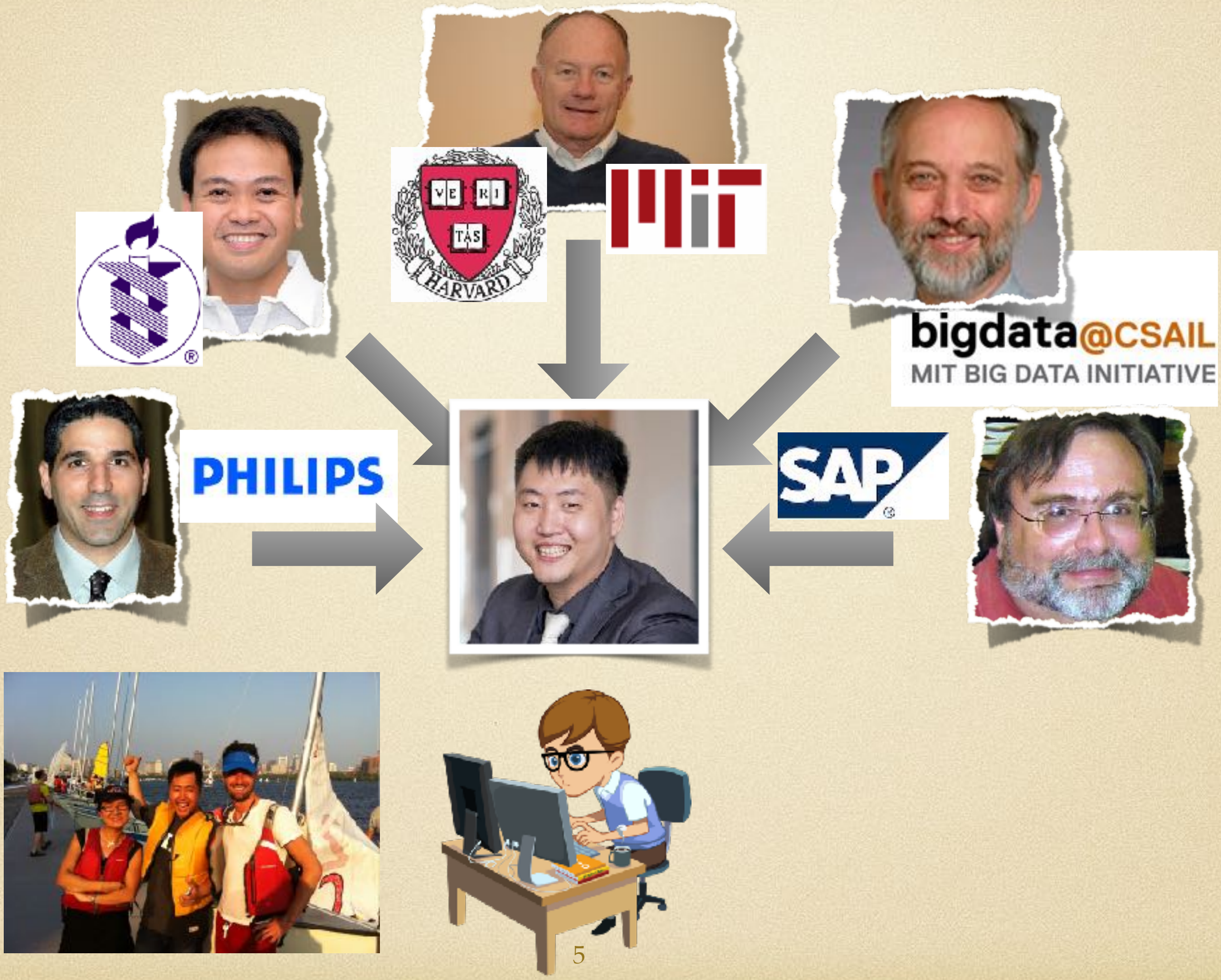

VE RI
TAS
HARVARD


VE RI TAS
HARVARD
bigdata@CSAIL
MIT BIG DATA INITIATIVE


VE RI TAS
HARVARD
MIT
bigdata@CSAIL
MIT BIG DATA INITIATIVE
PHILIPS


VE RI TAS
HARVARD
bigdata@CSAIL
MIT BIG DATA INITIATIVE
PHILIPS
SAP

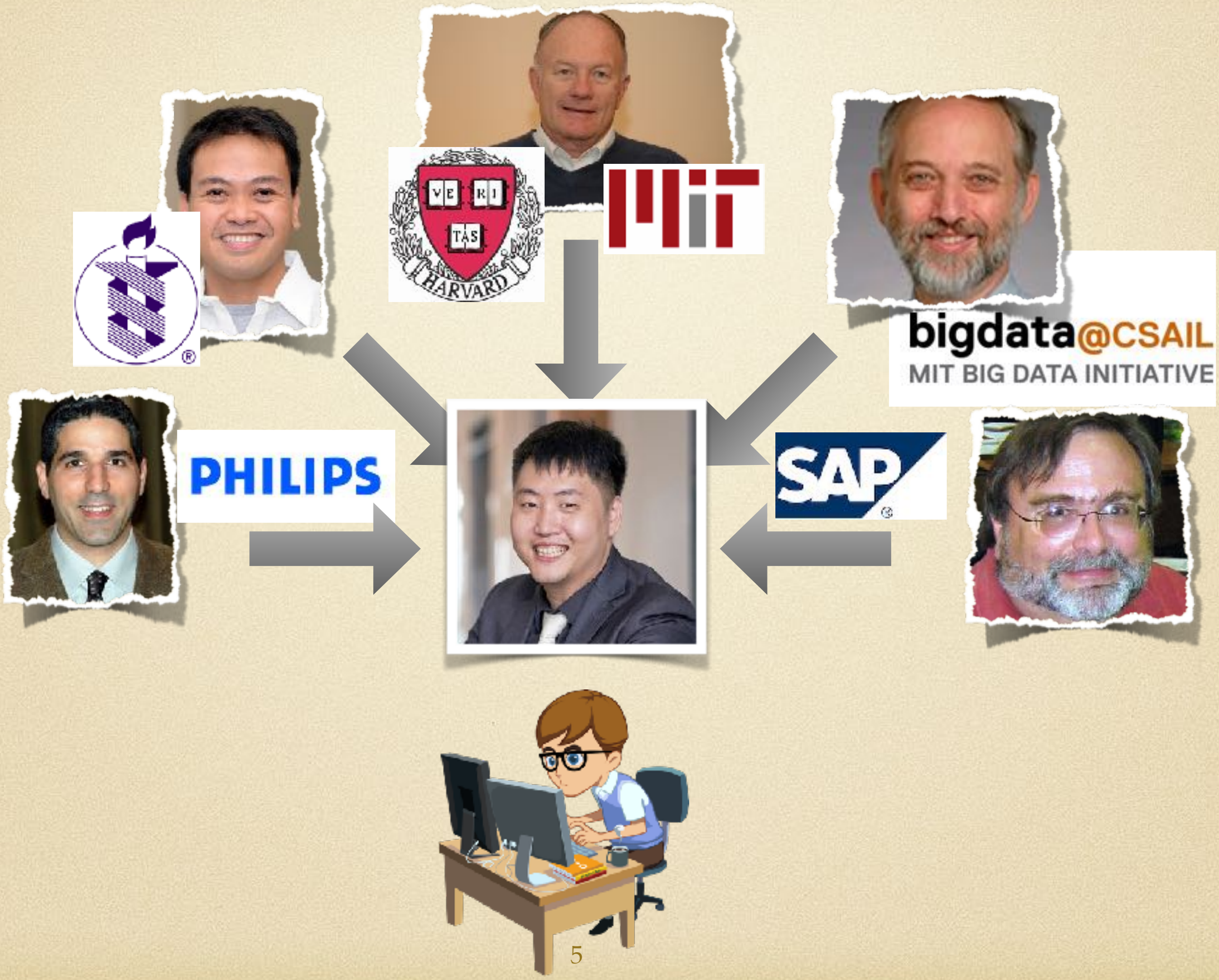

VE RI TAS
HARVARD
bigdata@CSAIL
MIT BIG DATA INITIATIVE
PHILIPS
SAP


VE RI TAS
HARVARD
MIT
bigdata@CSAIL
MIT BIG DATA INITIATIVE
PHILIPS
SAP

bigdata@CSAIL
MIT BIG DATA INITIATIVE

PHILIPS

SAP


VE RI TAS
HARVARD
bigdata@CSAIL
MIT BIG DATA INITIATIVE
PHILIPS
SAP

NUS
National University
of Singapore

NUS
National University of Singapore

NUS
National University of Singapore

School of Computing

NUS National University of Singapore

NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore

School of Computing

NUS National University of Singapore

NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore

School of Computing

NUS National University of Singapore

NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore
School of Computing

NUS National University of Singapore
Yong Loo Lin School of Medicine

NUS National University of Singapore

NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore
School of Computing

NUS National University of Singapore
Yong Loo Lin School of Medicine

NUS National University of Singapore

NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore

School of Computing

NUS National University of Singapore

Yong Loo Lin School of Medicine

NUS National University of Singapore

Saw Swee Hock School of Public Health

NUS National University of Singapore

NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore

School of Computing

NUS National University of Singapore

Yong Loo Lin School of Medicine

NUS National University of Singapore

Saw Swee Hock School of Public Health

NUS National University of Singapore

NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore
School of Computing

NUS National University of Singapore
Yong Loo Lin School of Medicine

NUS National University of Singapore
Saw Swee Hock School of Public Health

NUS National University of Singapore

NUH National University Hospital

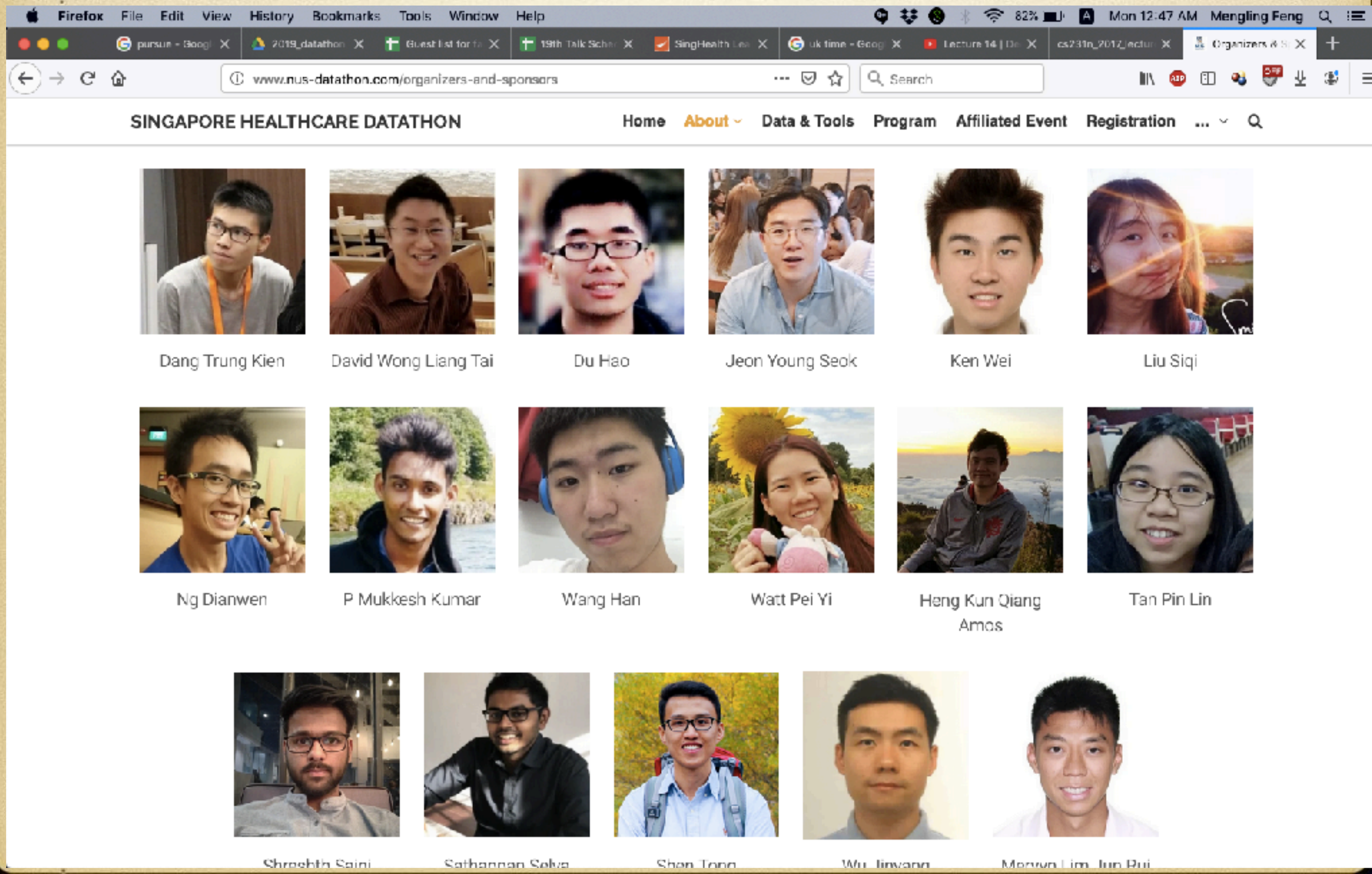
NUS National University of Singapore | Institute of Data Science

NUS National University of Singapore School of Computing

NUS National University of Singapore Yong Loo Lin School of Medicine

NUS National University of Singapore Saw Swee Hock School of Public Health

My AI Lab


Firefox File Edit View History Bookmarks Tools Window Help
Mon 12:47 AM Mengling Feng
www.nus-datathon.com/organizers-and-sponsors
Search
SINGAPORE HEALTHCARE DATATHON
Home About Data & Tools Program Affiliated Event Registration
Dang Trung Kien
David Wong Liang Tai
Du Hao
Jeon Young Seok
Ken Wei
Liu Siqi
Ng Dianwen
P Mukkesh Kumar
Wang Han
Watt Pei Yi
Heng Kun Qiang Amos
Tan Pin Lin

My AI Lab: Group Mornin www.group-mornin.com

- AI solutions for medical image analysis
- Reinforcement learning for treatment optimisation
- Natural Language Processing Algorithm for clinical text understanding

Part I: Introduction to Datathon

Datathon: An Effective Way to Cultivate Collaboration

Science Translational Medicine

Home | News | Journals | Topics | Careers

Science | Science Advances | Science Immunology | Science Robotics | Science Signaling | Science Translational Medicine

SHARE

0

0

PERSPECTIVE | REPRODUCIBILITY

A "datathon" model to support cross-disciplinary collaboration

Jerôme Aboab[1,*], Leo Anthony Celi[1], Peter Charlton[1], Mengling Feng[1], Mohammad Ghassemi[1], Dominic C. Marshall[1,†], Louis Mayaud[1], Tristan Naumann[1], Ned McCague[1], Kenneth E. Paik[1], Tom J. Pollard[1], Matthieu Resche-Rigon[1], Justin D. Salciccioli[1] and David J. Stone[2,3]

+ Author Affiliations

↵†Corresponding author. E-mail: dominic.marshall12@imperial.ac.uk

↵* All authors contributed equally to this manuscript.

Science Translational Medicine 06 Apr 2016:
Vol. 8, Issue 333, pp. 333ps8
DOI: 10.1126/scitranslmed.aad9072

How did it happen?

It first happened in MIT in the middle of a now storm @2014!

It first happened in MIT in the middle of a now storm @2014!

It first happened in MIT in the middle of a now storm @2014!

It first happened in MIT in the middle of a now storm @2014!

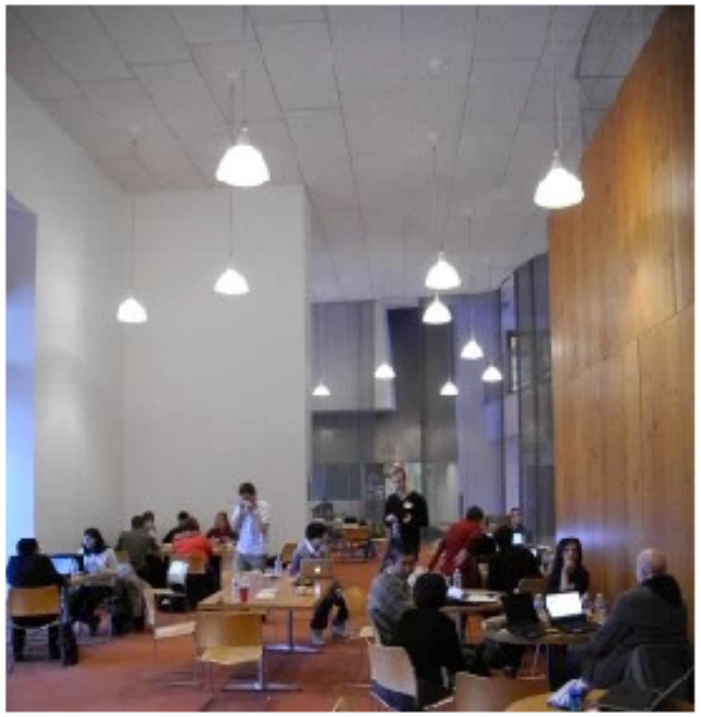
https://medtechboston.medstro.com/blog/2014/01/05/mit-critical-data-marathon-intro-to-the-mimic-ii-datab

Apps | Bookmarks | Elizabeth Stuart's... | Survival analysis | VISA | AR drones | R | AMIA 2013 Year in... | Year-

MIT CRITICAL DATA MARATHON: INTRO TO THE MIMIC-II DATABASE AND MEET THE TEAMS!

JENNIFER M. JOE, MD • 5 JAN, 2014 •

Today was the first full day of the Massachusetts Institute of Technology (MIT) Critical Data Marathon–a weekend hackathon designed to highlight the incredible resource of the MIMIC-II database.



Brilliant Minds Gathered in the Iconic MIT Stata Center

The First Datathon in 2014

- We had 100+ participants:
  - Physicians
  - Nurses
  - Computer scientists
- 10 teams
- Published 3 Journals
- Published 5 abstracts

Datathon =

Data + Hackathon

Key Elements in Datathon

- Preparation workshop
- **Clinical question pitching**

Key Elements in Datathon

- Preparation workshop
- **Clinical question pitching**

Key Elements in Datathon

- Preparation workshop
- **Clinical question pitching**

Key Elements in Datathon

- Preparation workshop
- **Clinical question pitching**
  - **Prepare your pitch for computer scientists**
  - **Be concise and clear**
  - **Show your passion**

Key Elements in Datathon

- Preparation workshop
- Clinical question pitching
- **Team Forming**

Key Elements in Datathon

- Preparation workshop
- Clinical question pitching
- **Team Forming**
  - **We need 1 clinical lead, 1 clinical co-lead**
  - **Need ideally 2 SQL data scientists**
  - **Need at least 1 statistician**

Key Elements in Datathon

- Preparation workshop
- Clinical question pitching
- Team Forming
- Hacking

Key Elements in Datathon

- Preparation workshop
- Clinical question pitching
- Team Forming
- **Hacking**
  - **Be realistic**
  - **Set multiple small milestones**
  - **Prepare your final presentation as earl as possible**

Key Elements in Datathon

- Preparation workshop
- Clinical question pitching
- Team Forming
- Hacking
- **Final Presentation**

Key Elements in Datathon

- Preparation workshop
- Clinical question pitching
- Team Forming
- Hacking
- **Final Presentation**
  - **Manage your time well**
  - **Emphasise what you have done**
  - **Elaborate the future plan**

Key Elements in Datathon

- Preparation workshop
- Clinical question pitching
- Team Forming
- Hacking
- Final Presentation
- **Friendship and collaboration beyond the datathon!**

Outcomes from Datathon

Outcomes from Datathon

Let's Have Fun Working Together!

Part II:
The MIMIC-III Dataset

The MIMIC-III Dataset
https://mimic.physionet.org/

MENU

SCIENTIFIC DATA

Data Descriptor | OPEN | Published: 24 May 2016

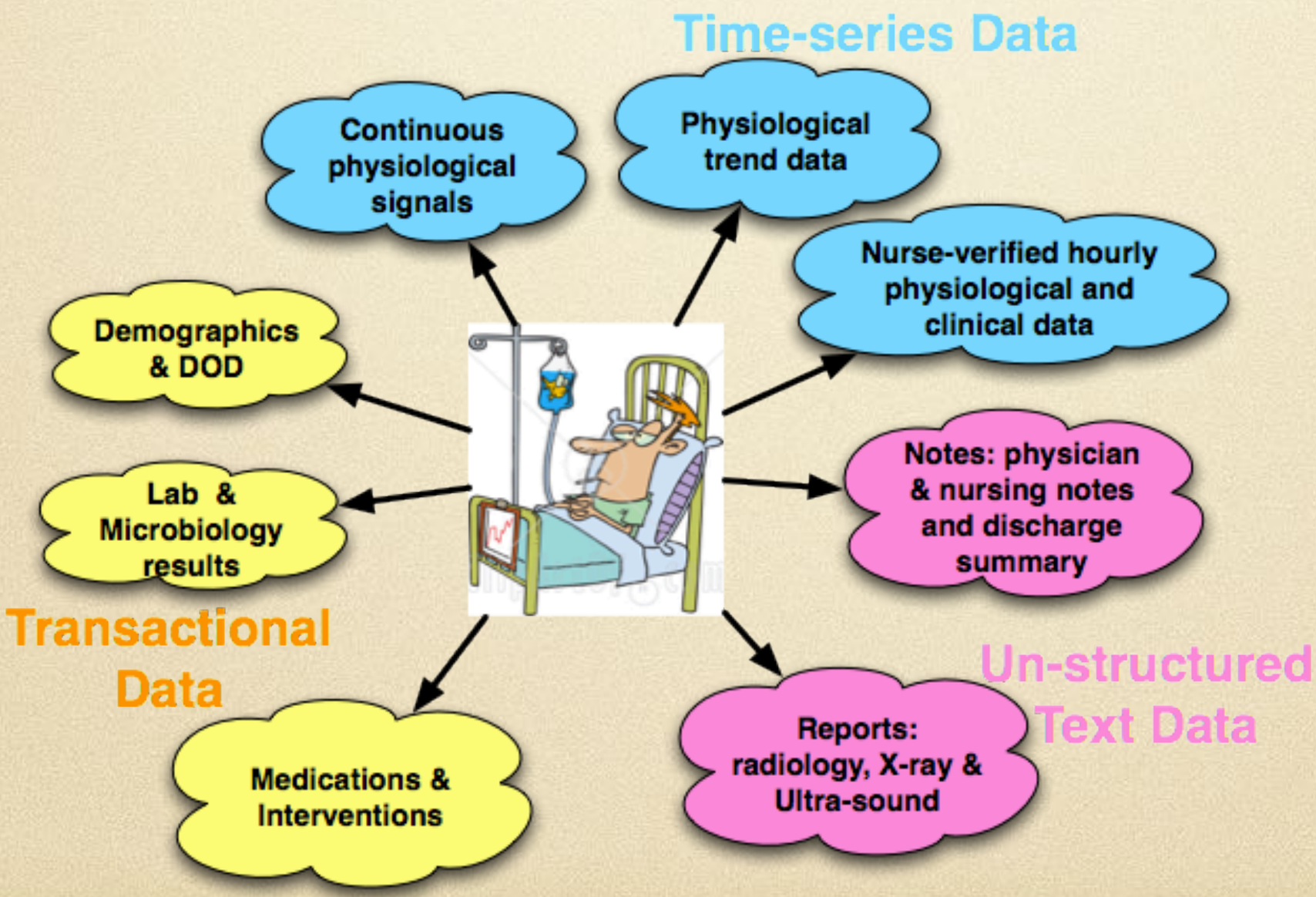
MIMIC-III, a freely accessible critical care database

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi & Roger G. Mark

Scientific Data **3**, Article number: 160035 (2016) | Download Citation

MIMIC-III Dataset

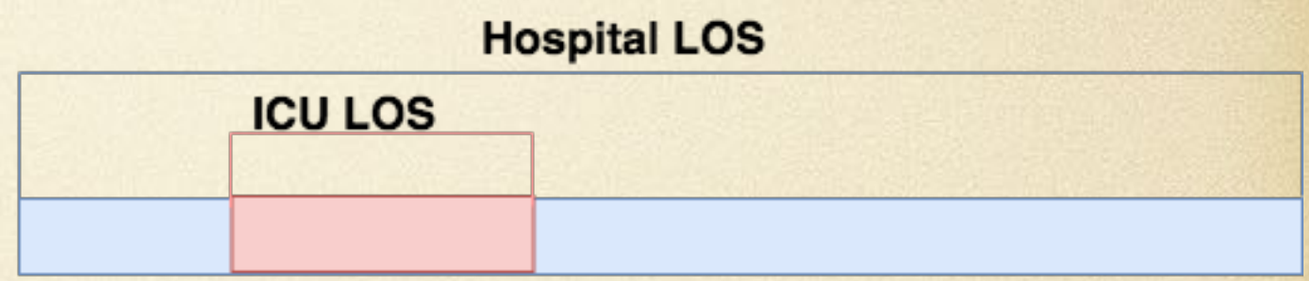
- Open Electronic Health Record (EHR) data for critical care
- 53,423 ICU admissions 2001~2012
- 26 tables



Organization of Data

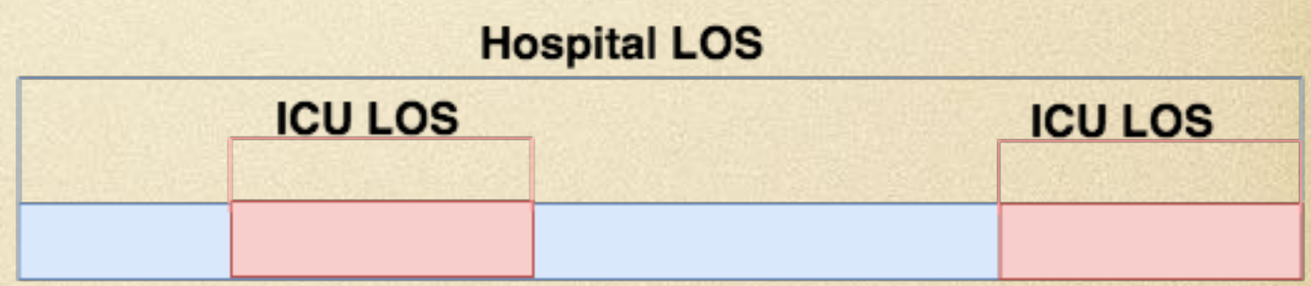
Subject_ID
= 2212

HADM_ID
= 300100



ICUSTAY_ID = 111222

HADM_ID
= 512612



ICUSTAY_ID
= 222111

ICUSTAY_ID
= 222112

De-identification

- Health Insurance Portability and Accountability Act (HIPAA 2003)
  - Names
  - Geographic locations
  - Dates
  - Ages over 89
  - Telephone/Pager #
  - SSN
  - Medical record #
  - Device ID & Serial #

Date Shifting

- Dates all shifted to random future
- Shifts are consistent for any individual patient
- Seasons were maintained
- Day of the week were maintained
- Time were maintained
- Years and dates were all wrong

Age Shifting

- For any patients older than 90
  - Age were shifted to 200 for privacy protection
  - Replace the age with 91.4 (median of the group)

Example of De-identified Text

Example 1: **[**Known patient name**]** admitted in transfer from **[**Hospital 1**]** Hospital on **[**3012-09-02**]** following a clinic visit where he c/o sob, ruq pain, and possible r/o chf. pt admitted to ccu on **[**09-03**]** for increased aggitation, twitching, periods of unresponsiveness, arf illustrated by a creatnine of 4 (1.2 1 month ago), and elevated liver enzymes. pt of dr. **[**First Name 2**] [**Doctor Last Name 3**]**.

Example 2: pt claims that he is compliant at home on meds, amiodarone 200mg po qd, dig .125mg po qd, colace 100mg po qd, methadone (methadone in house has been dc'd) his wife, **[**Name2 4**] [**Telephone/Fax 5**]**) confirms the story.

https://mimic.physionet.org/

Part III:
Study Design

The PICO System

- **P**opulation
  - Study cohort: inclusion and exclusion criteria
- **I**ntervention/ **E**xposure
  - The focus treatment of the study
- **C**ovariate/co-founding factors/variables
  - Factors that are associated with both the Exposure and the Outcome
- **O**utcome

Co-variates/Co-founding factors

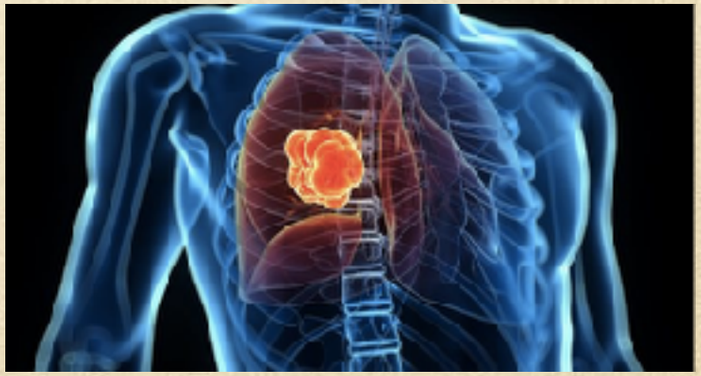
- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome

Co-variates/Co-founding factors

- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome

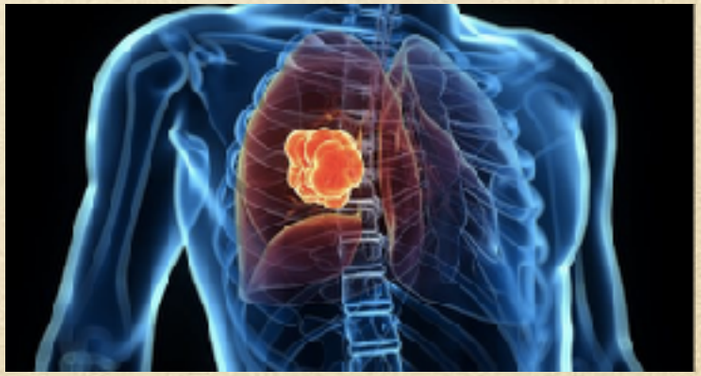
Co-variates/Co-founding factors

- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome



Co-variates/Co-founding factors

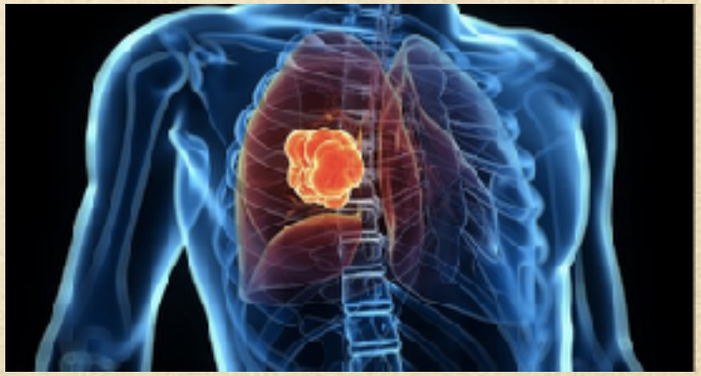
- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome



Co-variates/Co-founding factors

- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome

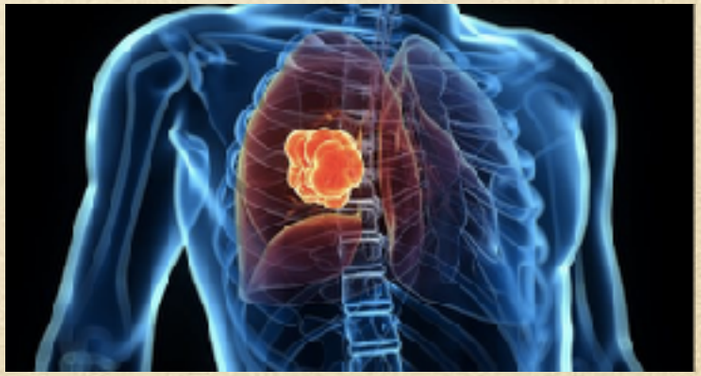
OR =5



Co-variates/Co-founding factors

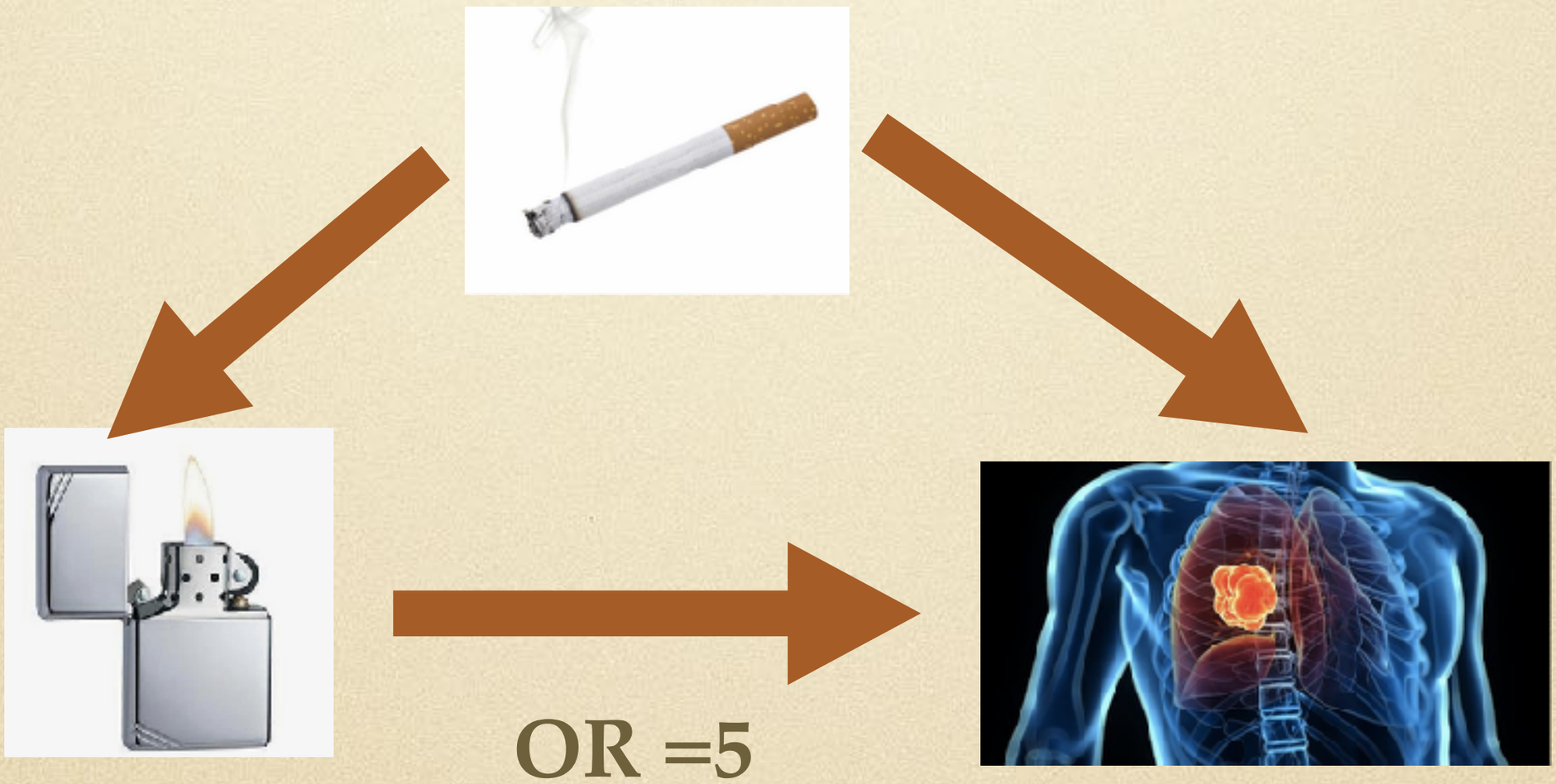
- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome

OR =5



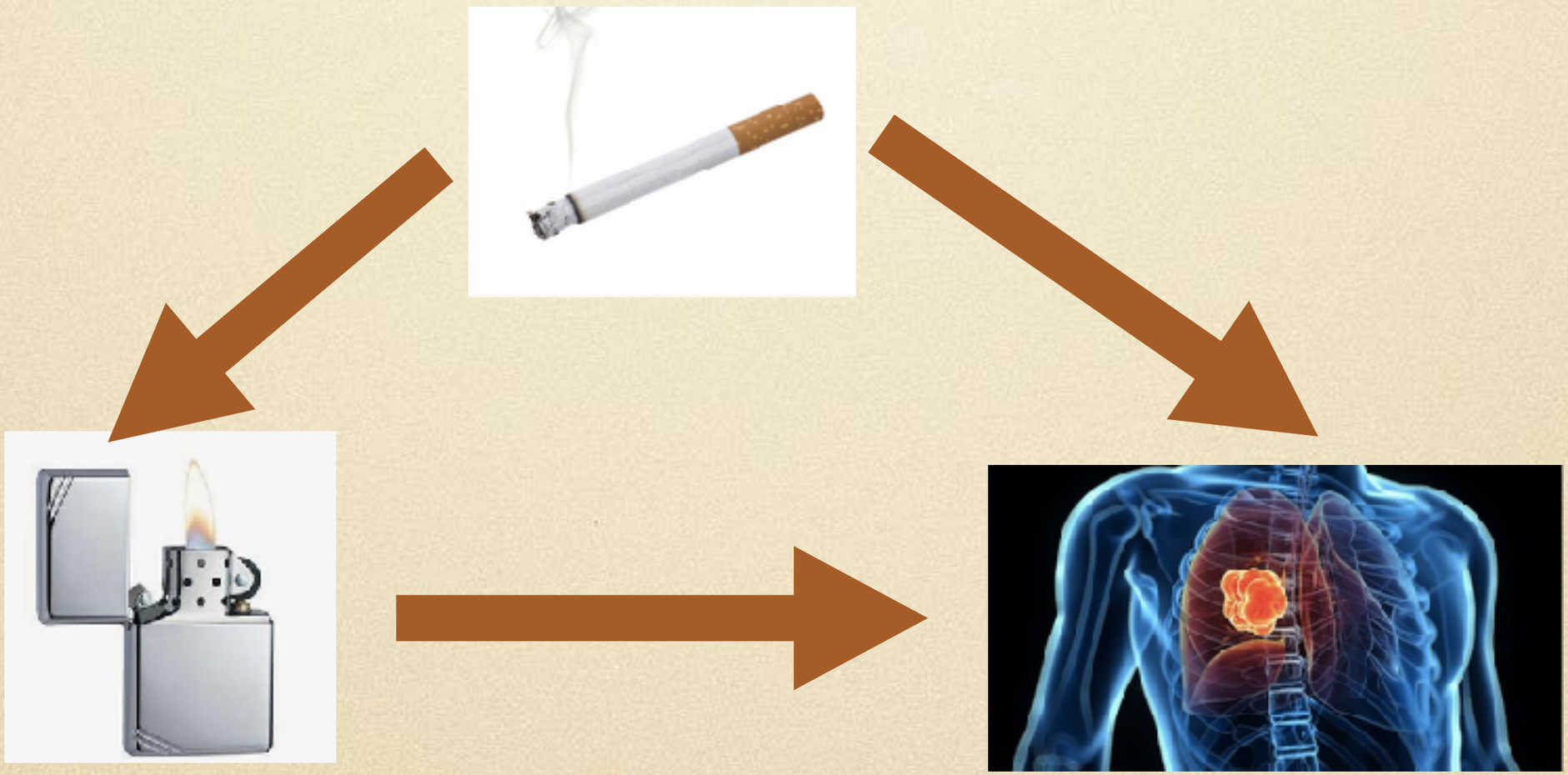
Co-variates/Co-founding factors

- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome



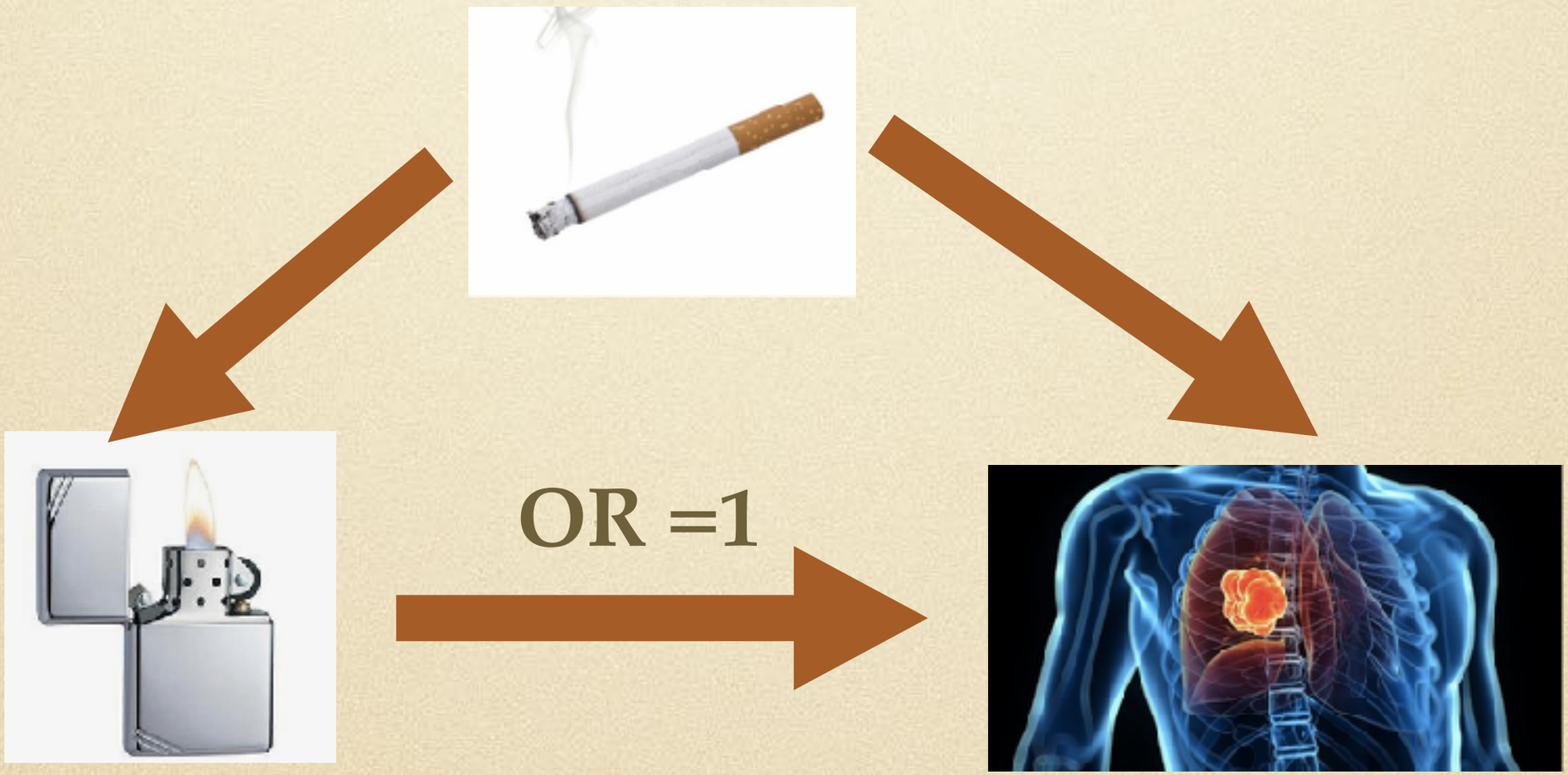
Co-variates/Co-founding factors

- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome



Co-variates/Co-founding factors

- Crucial for Causal Inference studies
- Factors that are associated with both the Exposure and the Outcome



Outcomes

- Mortality
  - Hospital mortality, ICU mortality
  - 28/30 days mortality
  - Up to 1 year of follow-up period
- LOS
  - Hospital/ICU LOS
- Ventilation Free Days (to counter the competitive effect of mortality)
  - E.g. 30-day ventilation free day
- Vasopressor free days
- Primary and Secondary outcomes

Case Study

Intensive Care Medicine
June 2018, Volume 44, Issue 6, pp 884–892 | Cite as

Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database

Authors | Authors and affiliations

Mengling Feng, Jakob I. McSparron, Dang Trung Kien, David J. Stone, David H. Roberts, Richard M. Schwartzstein, Antoine Vieillard-Baron, Leo Anthony Celi

Original
First Online: 28 May 2018

7 Shares | 1.1k Downloads | 1 Citations

Case Study

While the use of transthoracic echocardiography (TTE) in the ICU is rapidly expanding, the contribution of TTE to altering patient outcomes among ICU patients with sepsis has not been examined. This study was designed to examine the association of TTE with 28-day mortality specifically in that population.

Part IV:
Data Extraction & Analysis

Data Extration

https://github.com/MIT-LCP/mimic-code

Summarise Basic Characteristics of Cohort

- Table 1
  - Basic descriptive statistics and univariate test results for all co-variates and some common clinical variables
  - P-values may not be that important
  - Check the quality and reliability of data — Does it make sense?

Covariate	Original Cohort			Matched Cohort			Missing
	Non-TTE	TTE	SSMD	Non-TTE	TTE	SSMD	Data %
n	3099	3262		1626	1626		NA
Age	**66.69 (17.21)**	**65.74 (16.55)**	**0.056**	66.58 (16.54)	65.84 (17.07)	0.044	0.0%
Gender (Female)	**50.80%**	**47.80%**	**0.061**	48.00%	49.00%	0.021	0.0%
Service Unit (MICU %)	**77.60%**	**79.50%**	**0.047**	78.40%	78.90%	0.014	0.0%
Weight (Kg)	**78.56 (23.58)**	**83.17 (26.87)**	**0.182**	81.35 (24.69)	79.50 (23.11)	0.077	9.4%
SAPS Score	**19.63 (5.79)**	**20.76 (5.44)**	**0.202**	20.29 (5.42)	19.80 (5.17)	0.092	0.0%
SOFA Score	**5.31 (3.62)**	**6.33 (3.79)**	**0.277**	**5.99 (3.74)**	**5.55 (3.50)**	**0.123**	0.0%
Elixhauser Score	**8.51 (7.45)**	**10.07 (7.67)**	**0.207**	9.54 (7.67)	8.82 (7.52)	0.096	0.0%
Interventions							
Mechanical Ventilation Use (1st 24 Hours)	**47.10%**	**58.30%**	**0.226**	**52.30%**	**49.60%**	**0.055**	0.0%
Vasopressor Use (1st 24 Hours)	**27.10%**	**37.60%**	**0.227**	**32.70%**	**31.20%**	**0.032**	0.0%
Sedative Use (1st 24 Hours)	**40.20%**	**49.80%**	**0.192**	43.80%	42.80%	0.021	0.0%
Co-morbidities							
CHF	**18.20%**	**40.00%**	**0.495**	**28.80%**	**19.40%**	**0.221**	0.0%
AFIB	**20.10%**	**32.40%**	**0.282**	**27.60%**	**24.80%**	**0.062**	0.0%
RENAL	**14.10%**	**16.40%**	**0.066**	15.40%	14.30%	0.031	0.0%
LIVER	10.10%	11.20%	0.034	11.40%	11.90%	0.013	0.0%
COPD	**15.20%**	**17.50%**	**0.063**	17.80%	15.60%	0.061	0.0%
CAD	**11.90%**	**15.80%**	**0.113**	13.20%	13.20%	0.002	0.0%
STROKE	**7.70%**	**10.80%**	**0.106**	8.80%	8.70%	0.002	0.0%
MALIGNANCY	**25.20%**	**22.30%**	**0.068**	25.00%	25.20%	0.003	0.0%
Vital Signs							
MAP	79.91 (19.44)	80.03 (20.48)	0.006	80.13 (19.95)	80.31 (20.05)	0.009	1.2%
Heart Rate	**93.01 (19.81)**	**95.07 (21.79)**	**0.099**	94.88 (20.77)	93.88 (20.04)	0.049	1.2%
Temperature (C)	**36.75 (1.05)**	**36.85 (1.90)**	**0.07**	36.81 (1.06)	36.86 (1.85)	0.031	1.6%
CVP	**11.88 (17.09)**	**13.74 (20.45)**	**0.099**	12.41 (19.95)	11.86 (11.63)	0.033	64.5%
Lab tests							
WBC	13.48 (14.03)	13.76 (12.41)	0.021	13.71 (15.17)	13.70 (13.84)	0.001	4.1%
HEMOGLOBIN	**10.55 (1.98)**	**10.60 (2.04)**	**0.021**	10.56 (2.01)	10.57 (2.02)	0.009	4.0%
PLATELET	**223.96 (134.79)**	**211.55 (128.46)**	**0.094**	215.31 (130.72)	216.43 (129.90)	0.009	4.1%
SODIUM	**139.09 (6.45)**	**138.56 (5.89)**	**0.086**	138.75 (5.88)	138.52 (6.08)	0.039	2.6%
POTASSIUM	**4.10 (0.79)**	**4.15 (0.84)**	**0.061**	4.12 (0.79)	4.09 (0.76)	0.033	2.4%
BICARBONATE	22.28 (5.40)	22.35 (5.71)	0.012	22.39 (5.67)	22.30 (5.33)	0.018	2.9%
CHLORIDE	**106.21 (7.45)**	**105.14 (7.20)**	**0.145**	**105.50 (7.14)**	**105.47 (7.21)**	**0.004**	2.6%
BUN	**32.02 (26.34)**	**36.01 (27.38)**	**0.149**	34.05 (26.40)	32.41 (25.57)	0.063	2.8%
LACTATE	2.62 (2.35)	2.51 (2.27)	0.047	2.58 (2.21)	2.47 (2.14)	0.049	38.2%
CREATININE	**1.60 (1.66)**	**1.83 (1.81)**	**0.132**	1.72 (1.68)	1.63 (1.64)	0.053	2.8%
PH	**7.35 (0.11)**	**7.34 (0.11)**	**0.043**	7.34 (0.11)	7.35 (0.11)	0.086	30.5%
PO2	**147.61 (107.49)**	**133.24 (93.89)**	**0.142**	135.33 (98.90)	139.78 (100.24)	0.045	33.1%
PCO2	**41.48 (13.34)**	**42.77 (14.83)**	**0.091**	42.18 (13.69)	41.46 (13.81)	0.053	33.1%
BNP (tested)	**1.20%**	**4.40%**	**0.196**	**2.20%**	**1.50%**	**0.05**	0.0%
TROPONIN (tested)	**20.40%**	**40.70%**	**0.451**	**31.90%**	**24.10%**	**0.173**	0.0%
CREATININE KINASE (tested)	**37.20%**	**59.30%**	**0.452**	**51.80%**	**44.00%**	**0.157**	0.0%
Day of ICU Admission			**0.148**			0.093	
SUNDAY	12.70%	14.20%		14.40%	12.90%		
MONDAY	13.40%	14.60%		13.00%	13.70%		
TUESDAY	13.80%	15.10%		13.70%	15.70%		
WEDNESDAY	13.60%	15.80%		15.40%	13.80%		
THURSDAY	15.00%	15.30%		14.60%	14.50%		
FRIDAY	17.40%	13.70%		15.70%	17.10%		
SATURDAY	14.10%	11.40%		13.20%	12.20%		0.0%

Statistical Models

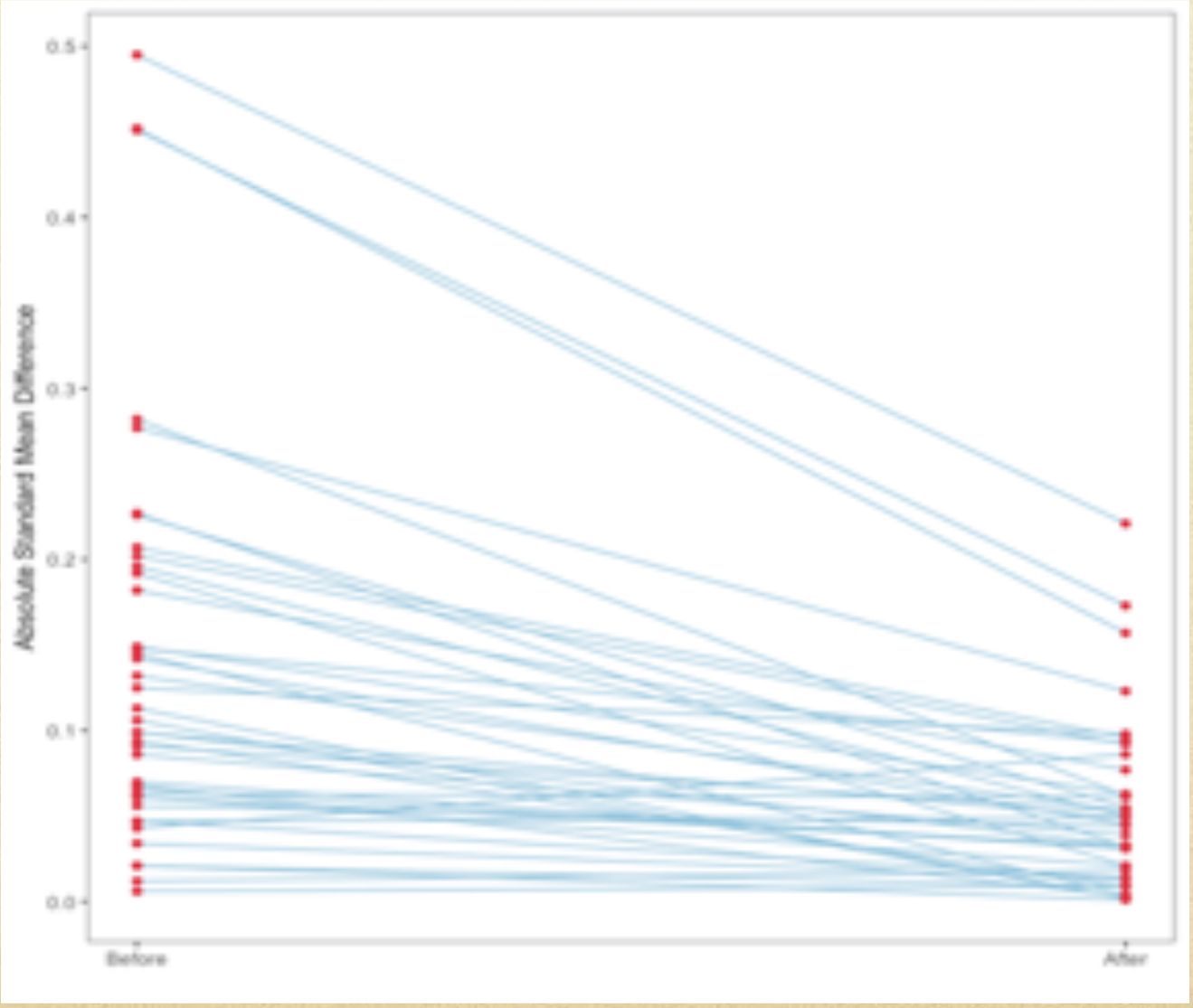
- Multi-variate Study with logistic Regression
- Propensity Score Based Model
- Doubly Robust Model

Propensity Score

- A predictive model with all the covariate to estimate the likelihood of Intervention or Exposure
- The predicted likelihood = Propensity Score

Propensity Score

- A predictive model with all the covariate to estimate the likelihood of Intervention or Exposure
- The predicted likelihood = Propensity Score



Doubly Robust Model

- Propensity Score based IPW Model + Multivariate model

mornin@gmail.com
www.mornin-feng.com

Now is your turn!