THE UNIVERSITY OF MELBOURNE

Introduction to Data Science

Daniel Capurro, MD, PhD
Computing and Information Systems
Melbourne School of Engineering
Centre for Digital Transformation of Health - MDHS

Short introduction

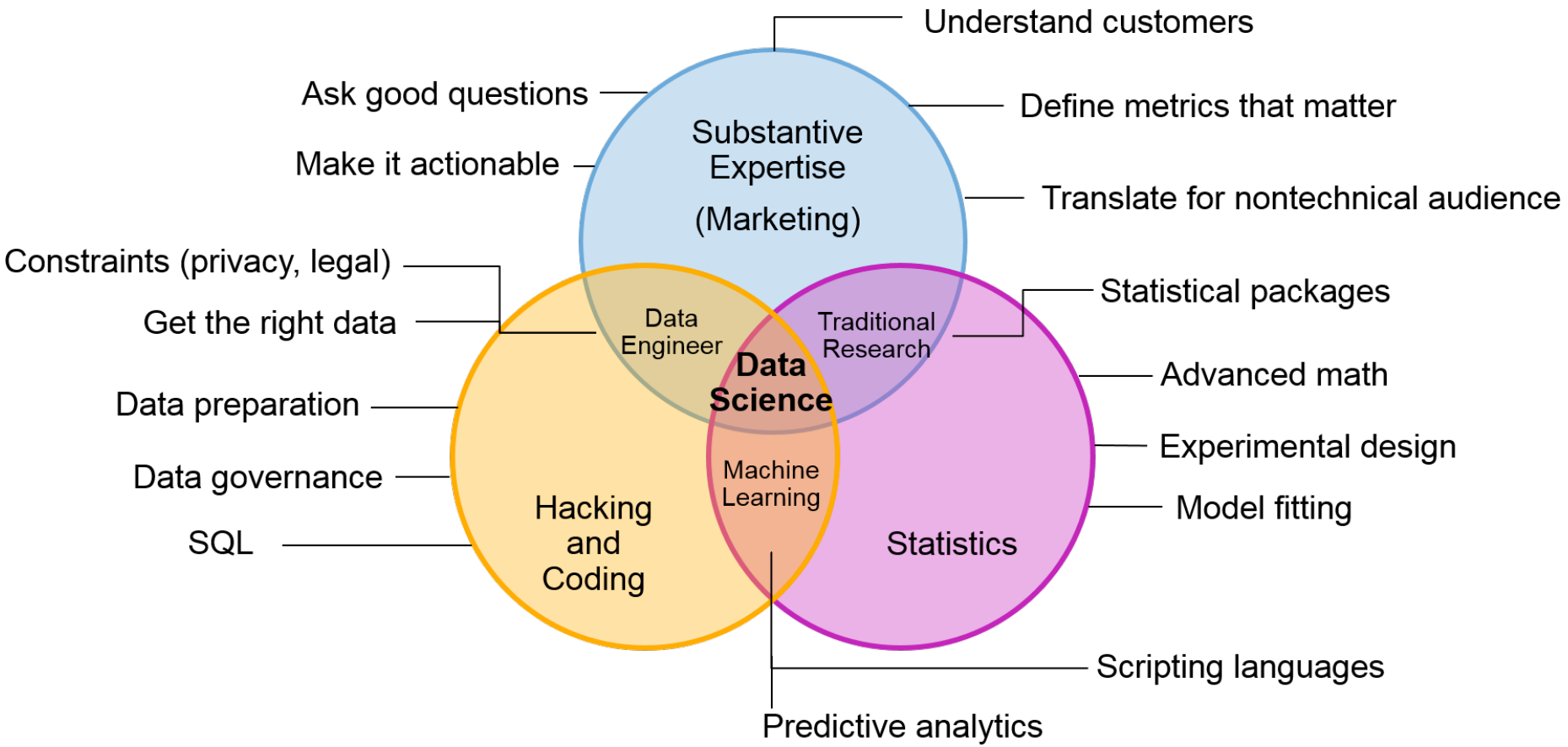
- Medical Doctor, practiced Internal Medicine up to a year ago
- PhD in Biomedical Informatics
- Senior Lecturer in Digital Health, School of Computing and Information Systems
- Centre for the Digital Transformation of Health

Definition…

- Not really a consensus definition…

- Intersection between
  - Statistics
  - Computer Science (Machine Learning, Big Data)
  - Business Analytics (in our case, healthcare analytics)

- Definition will probably evolve over time


Understand customers
Ask good questions
Define metrics that matter
Substantive Expertise
(Marketing)
Make it actionable
Translate for nontechnical audience
Constraints (privacy, legal)
Statistical packages
Get the right data
Data Engineer
Traditional Research
Data Science
Advanced math
Data preparation
Experimental design
Machine Learning
Data governance
Model fitting
Hacking and Coding
Statistics
SQL
Scripting languages
Predictive analytics

Data Science a paradigm shift

Traditional research:

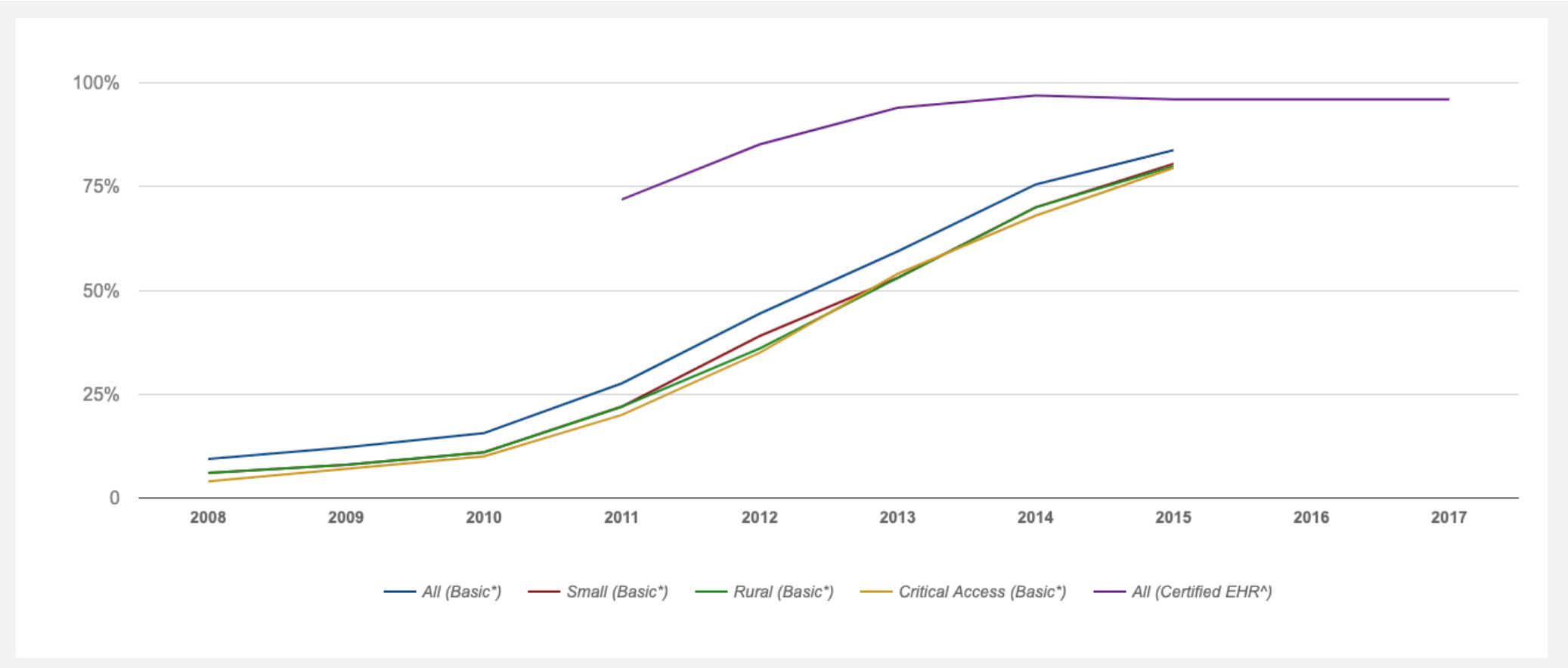
1. Hypothesis
2. Study design to answer that hypothesis
3. Collect data
4. Analyze data
5. Reject or accept your hypothesis, generate new hypothesis, etc.

What enables the paradigm shift?

- Data is being routinely collected in information systems
- We continuously interact with ‘digital environments’

In healthcare this (mostly) means EHR adoption

US hospital adoption of EHRs



Health IT Dashboard. The Office of the National Coordinator for Health Information Technology. Accessed on 08/08/2019, available from https://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php

But also other sources of health/behaviour data

AJPH RESEARCH

Twitter as a Tool for Health Research: A Systematic Review

Lauren Sinnenberg, BA, Alison M. Buttenheim, PhD, MBA, Kevin Padrez, MD, Christina Mancheno, BA, Lyle Ungar, PhD, and Raina M. Merchant, MD, MSHP

Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. American journal of public health. 2017 Jan;107(1):e1-8.

Smartphone sensors

iPhone 7 Teardown. iFixit. Accessed on 08/08/2019, available from https://www.ifixit.com/Teardown/iPhone+7+Teardown/67382

Special Collection: Smart Wellness Services for Lifestyles of Health and Sustainability (LOHAS)

INTERNATIONAL JOURNAL OF ENGINEERING BUSINESS MANAGEMENT

International Journal of Engineering Business Management
Volume 10: 1–8

DOI: 10.1177/1847979017750669
journals.sagepub.com/home/enb
SAGE

A real-time fall detection system based on the acceleration sensor of smartphone

Youngmin Lee[1], Hongjin Yeh[2], Ki-Hyung Kim[2], and Okkyung Choi[3]

Lee Y, Yeh H, Kim KH, Choi O. A real-time fall detection system based on the acceleration sensor of smartphone. International Journal of Engineering Business Management. 2018 Jan 5;10:1847979017750669.

New paradigm: 2 options

1. Hypothesis-driven:

    Hypothesis

    Study design using previously collected data

    Data linkage, data preparation, data cleaning

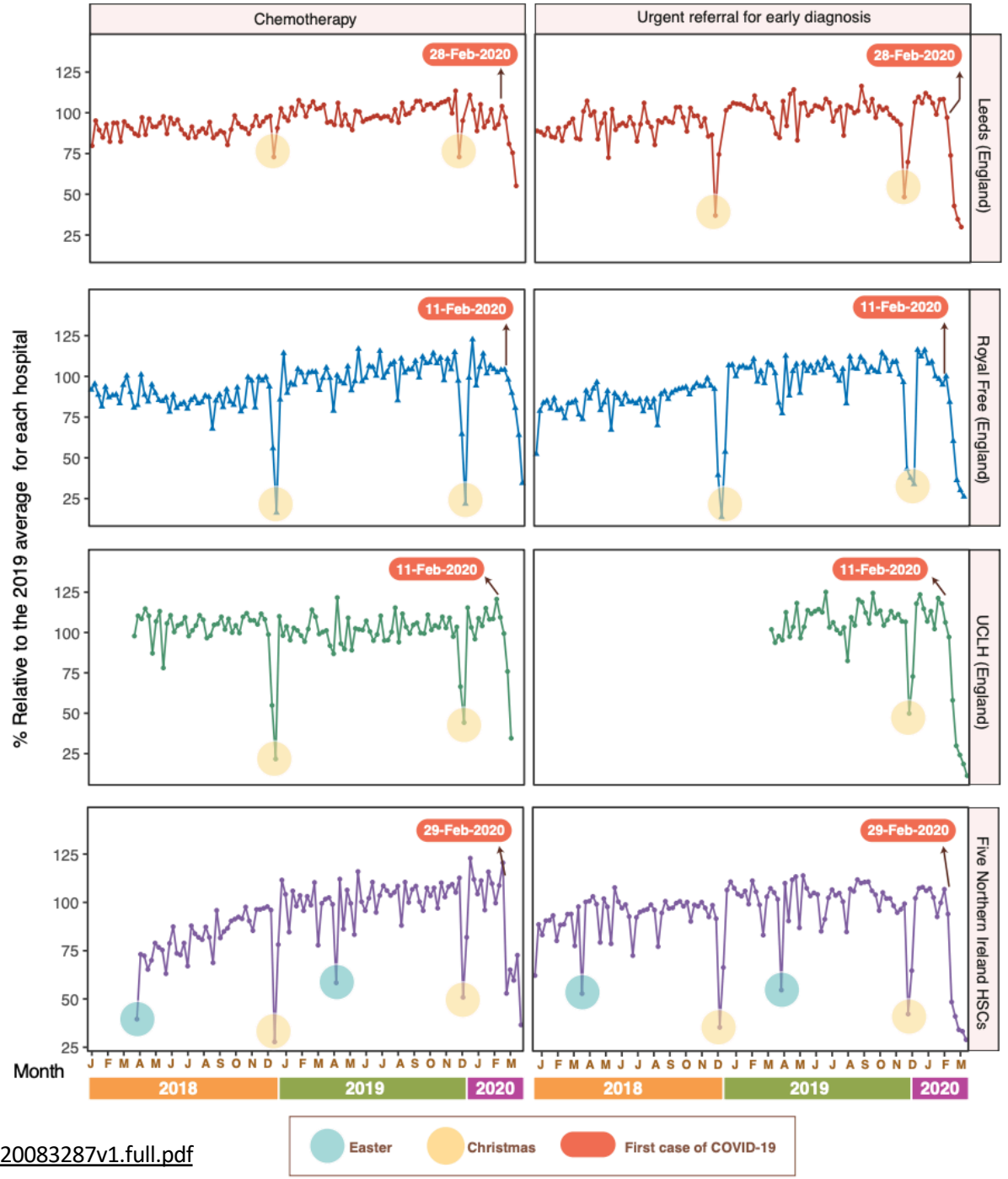
    Analyze data

    Reject or accept your hypothesis, generate new hypothesis, etc.

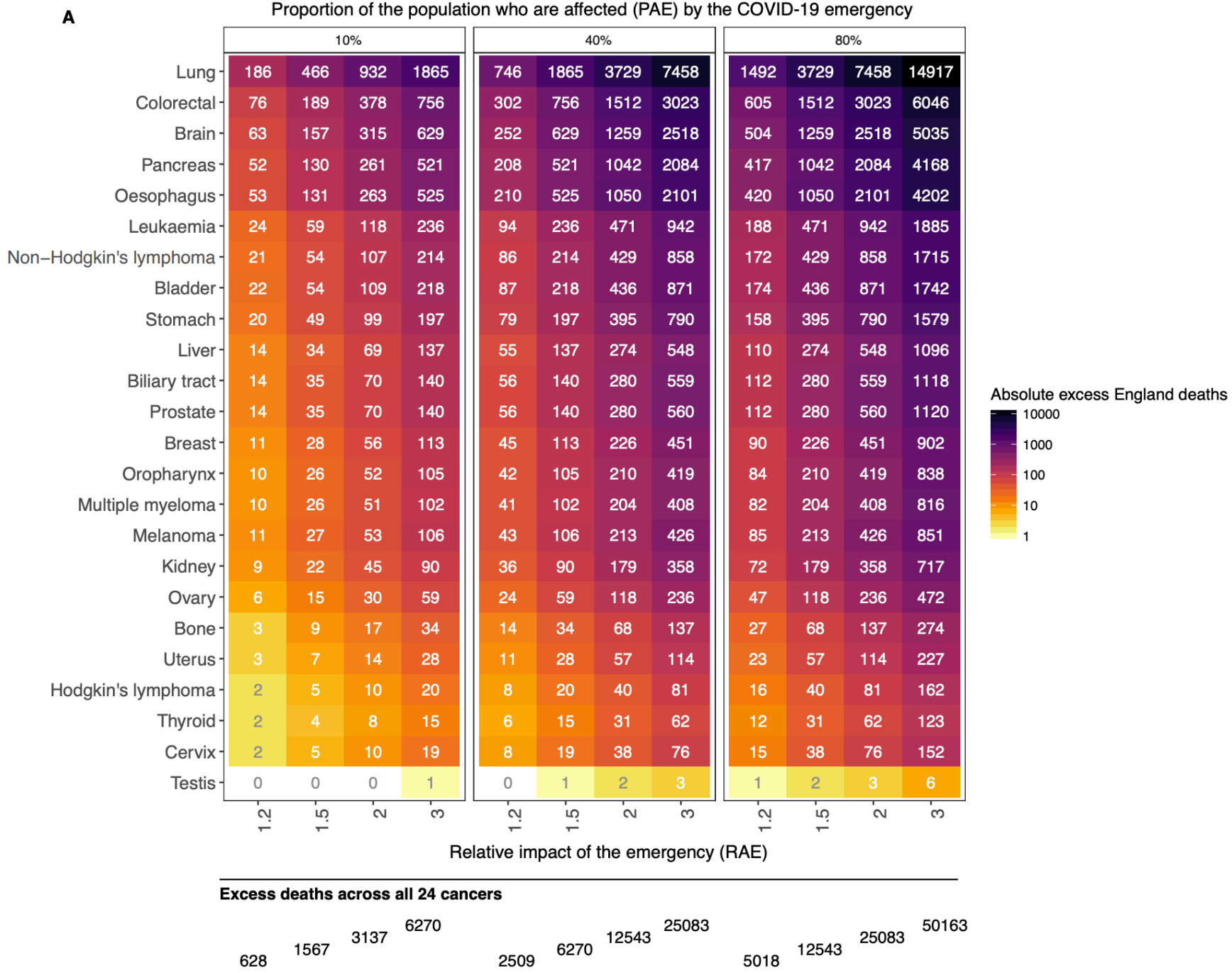
Estimating excess mortality in people with cancer and multimorbidity in the COVID-19 emergency

Alvina G. Lai, Ph.D.[1,2,✉], Laura Pasea, Ph.D.[1,2*], Amitava Banerjee, DPhil[1,2,3*], Spiros Denaxas, Ph.D.[1,2,6,7], Michail Katsoulis, Ph.D.[1,2], Wai Hoong Chang, MSc[1,2], Bryan Williams, Ph.D.[4,5,6], Deenan Pillay, Ph.D.[8], Mahdad Noursadeghi, Ph.D.[8], David Linch, FMedSci[6,9], Derralynn Hughes, FRCPath[10,11], Martin D. Forster, Ph.D.[4,10], Clare Turnbull, Ph.D.[12], Natalie K. Fitzpatrick, MSc[1,2], Kathryn Boyd, MD[13], Graham R. Foster, Ph.D.[14], DATA-CAN[15], Matt Cooper, Ph.D.[15], Monica Jones, PGDip[15], Kathy Pritchard-Jones, FMedSci[15,16,17,18], Richard Sullivan, Ph.D.[19], Geoff Hall, Ph.D.[15,20,21], Charlie Davie, FRCP[11,15,16], Mark Lawler, Ph.D.[15,22], and Harry Hemingway, FMedSci[1,2,6,✉]

https://www.medrxiv.org/content/10.1101/2020.05.27.20083287v1.full.pdf



https://www.medrxiv.org/content/10.1101/2020.05.27.20083287v1.full.pdf



https://www.medrxiv.org/content/10.1101/2020.05.27.20083287v1.full.pdf

New paradigm: 2 options

2. Data-driven

    (Usually) No hypothesis

    Data linkage, data preparation, data cleaning

    Identify patterns (knowledge) in databases

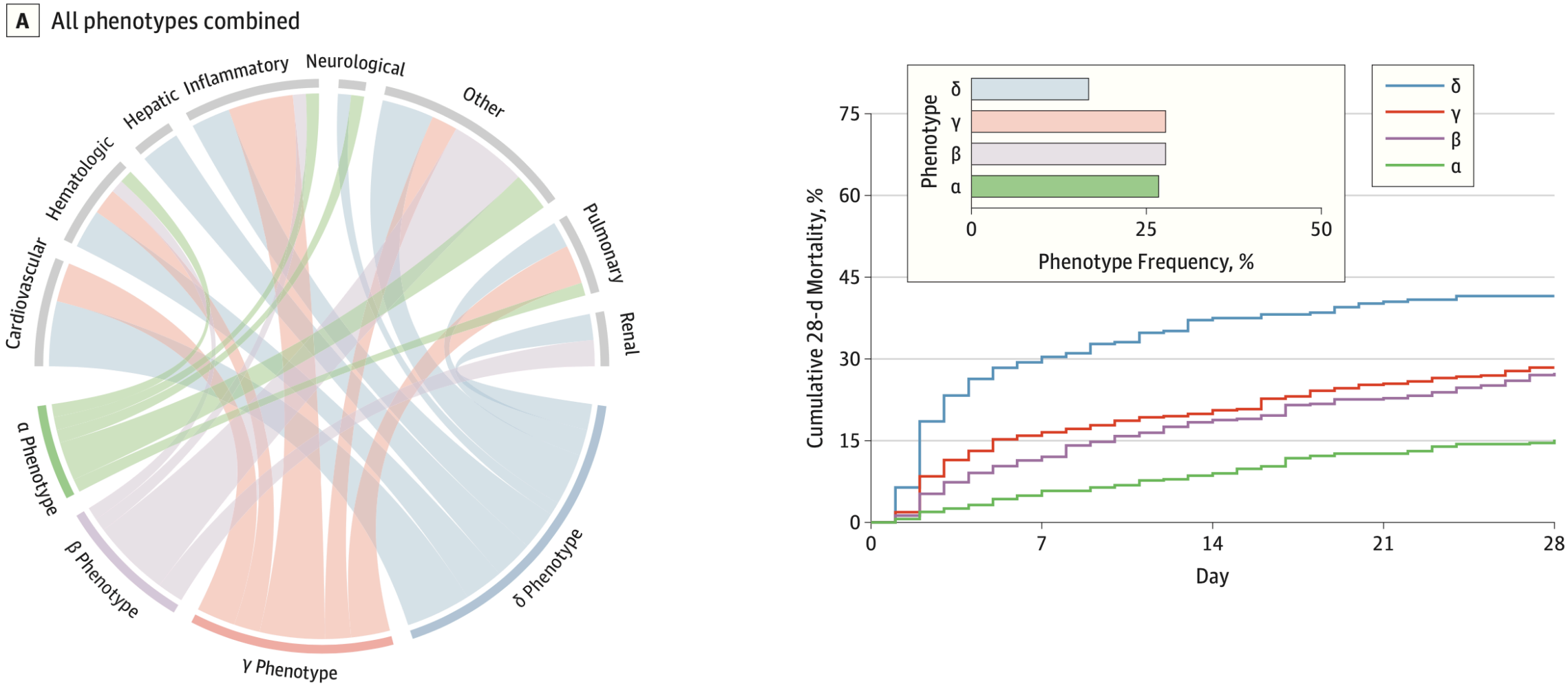
        Supervised and unsupervised methods

    New hypothesis that <u>needs to be tested</u>

JAMA | Original Investigation | CARING FOR THE CRITICALLY ILL PATIENT

Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis

Christopher W. Seymour, MD, MSc; Jason N. Kennedy, MS; Shu Wang, MS; Chung-Chou H. Chang, PhD; Corrine F. Elliott, MS; Zhongying Xu, MS; Scott Berry, PhD; Gilles Clermont, MD, MSc; Gregory Cooper, MD, PhD; Hernando Gomez, MD, MPH; David T. Huang, MD, MPH; John A. Kellum, MD, FACP, MCCM; Qi Mi, PhD; Steven M. Opal, MD; Victor Talisa, MS; Tom van der Poll, MD, PhD; Shyam Visweswaran, MD, PhD; Yoram Vodovotz, PhD; Jeremy C. Weiss, MD, PhD; Donald M. Yealy, MD, FACEP; Sachin Yende, MD, MS; Derek C. Angus, MD, MPH



Seymour CW, Kennedy JN, Wang S, Chang CC, Elliott CF, Xu Z, Berry S, Clermont G, Cooper G, Gomez H, Huang DT. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019 May 28;321(20):2003-17.

THE UNIVERSITY OF MELBOURNE

How would that work?

(one) Data Science Pipeline

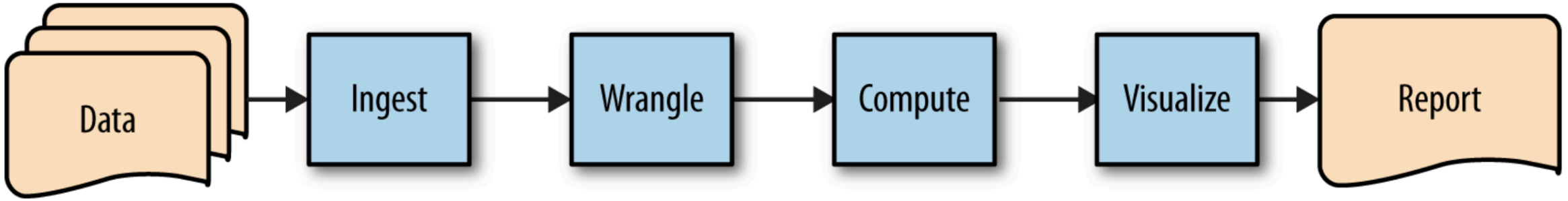
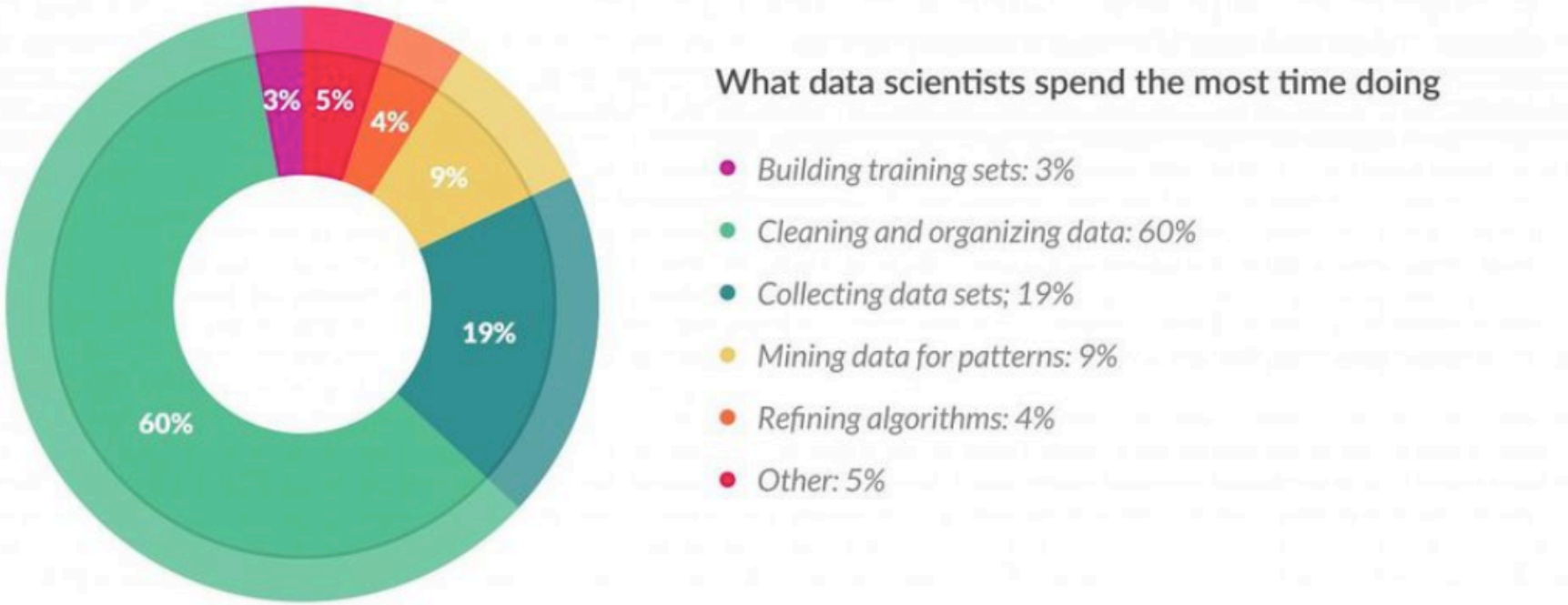


Figure 1-1. The data science pipeline

Bengfort B, Kim J. Data analytics with Hadoop: an introduction for data scientists. " O'Reilly Media, Inc."; 2016.

Data preparation

Data preparation *accounts for about 80% of the work of data scientists*



Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. Accessed on 8/8/2019, available from https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#65e413be6f63

Some steps involved in data preparation

- Record linkage
- Variable Standardization - Labelling
- Out-of-range values
- Logical inconsistencies
- Variable transformation
- Discretization
- Missing Data
- Data reduction

Holmes J. Knowledge Discovery in Biomedical Data: Theory and Methods. In: Sarkar IN, Methods in Biomedical Informatics. A Pragmatic Approach. Academic Press 2013.

Labelling clinical data is complex

Pt is 87 yo woman, highschool teacher with past medical history that includes

- status post cardiac catheterization in April 2019.

She presents today with palpitations and chest pressure.

HPI : Sleeping trouble on present dosage of Clonidine. Severe Rash on face and leg, slightly itchy

Meds : Vyvanse 50 mgs po at breakfast daily,

Clonidine 0.2 mgs -- 1 and 1 / 2 tabs po qhs

HEENT : Boggy inferior turbinates, No oropharyngeal lesion

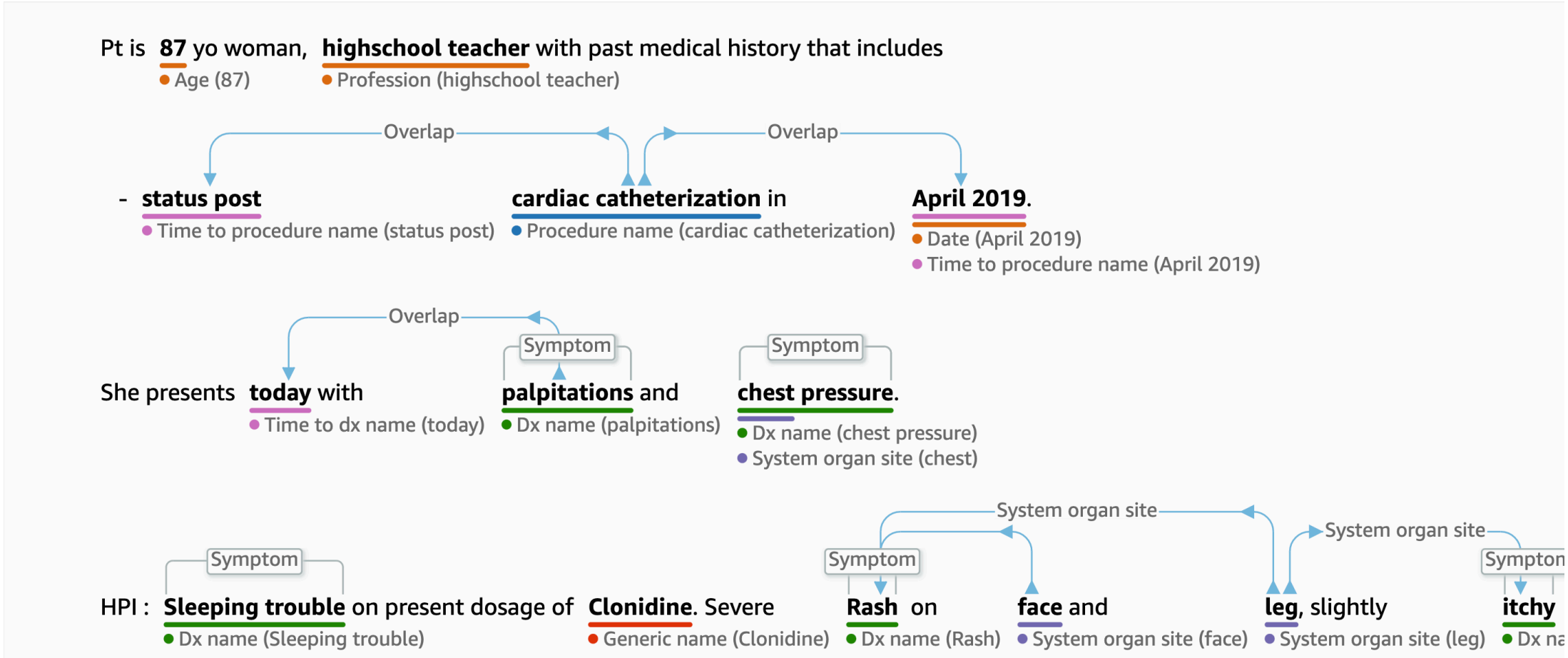
Lungs : clear

Heart : Regular rhythm

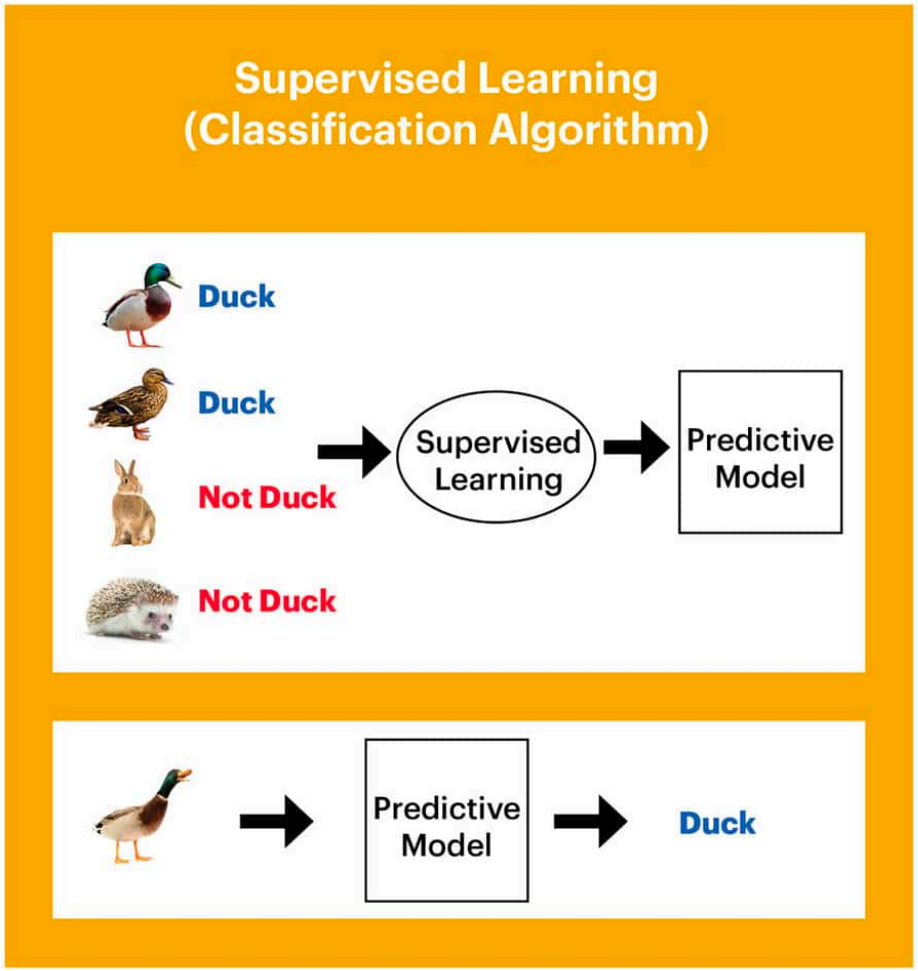
Skin : Mild erythematous eruption to hairline

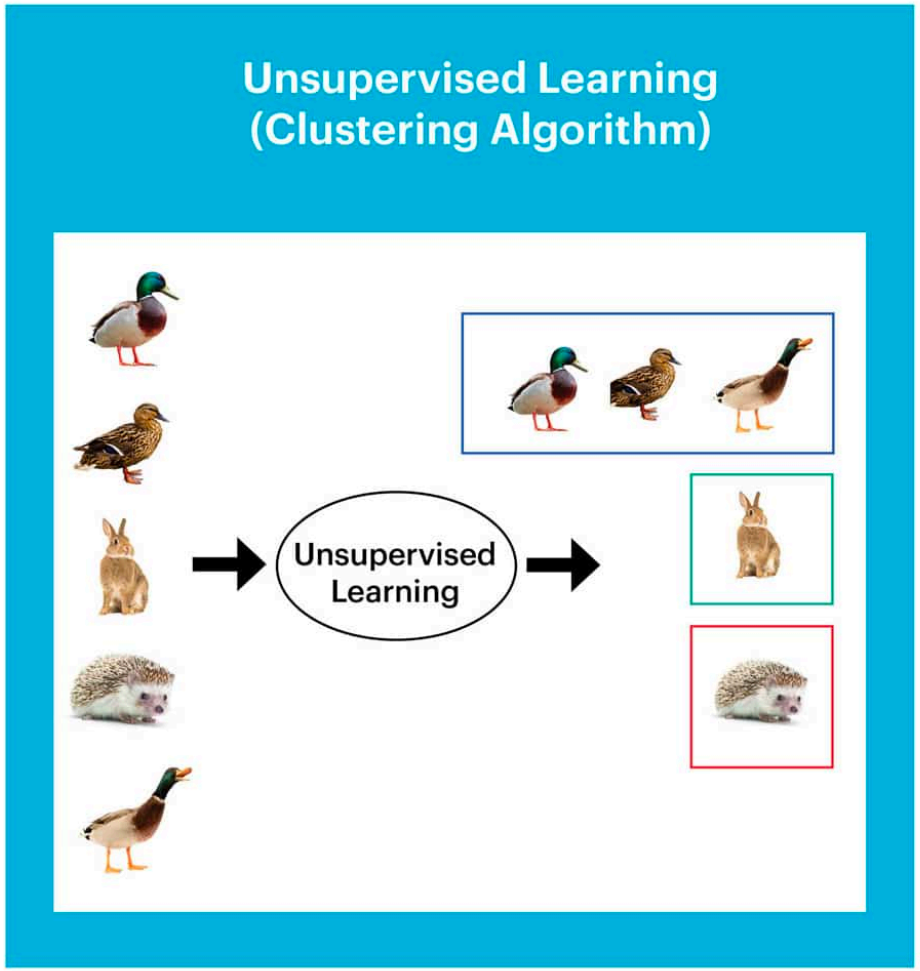
Follow-up as scheduled

Clinical Natural Language Processing



Supervised vs. Unsupervised Machine Learning





https://blog.westerndigital.com/wp-content/uploads/2018/05/supervised-learning-diagram.jpg

Opportunities and Challenges

Some opportunities

- Accelerate discoveries
- Discovery of non-apparent patterns
- Rare diseases
- New data sources
- Real-time analysis
- Machine learning and Artificial Intelligence

Some challenges

- Technical
  - Data quality
  - Data standards
  - Big Data
  - Analytical methods
- Non-technical
  - Validation
  - Privacy
  - Bias
  - Unintended consequences