2020-2021学年秋季学期

数据科学导论

The Introduction of Data Science

授课团队：沙瀛　周川

助　　教：梁棋

数据科学导论

The Introduction of Data Science

[第2章] 数据科学生命周期

授课教师：沙瀛

授课时间：2020/9/25

概 要

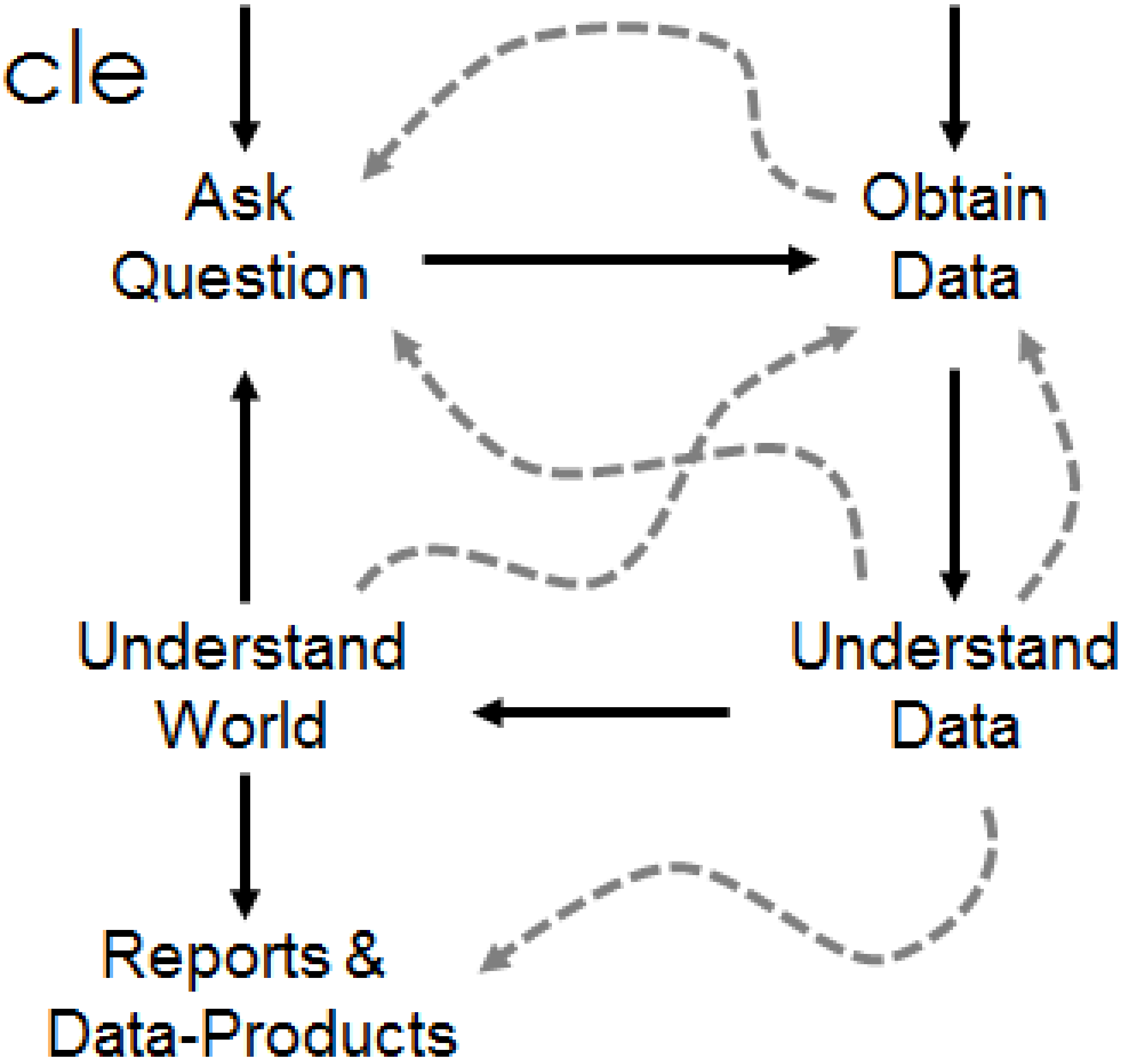
- 数据科学生命周期
- 问题的提出
- 实验设计
- 数据与取样
- 验证

Data Science Lifecycle

High-level description of the data science workflow

- Frame questions & design experiments
- Obtain and clean data
- Summarize and visualize data
- Inference and prediction

continuous process …



DATA SCIENCE LIFECYCLE: AN ALTERNATE VIEW

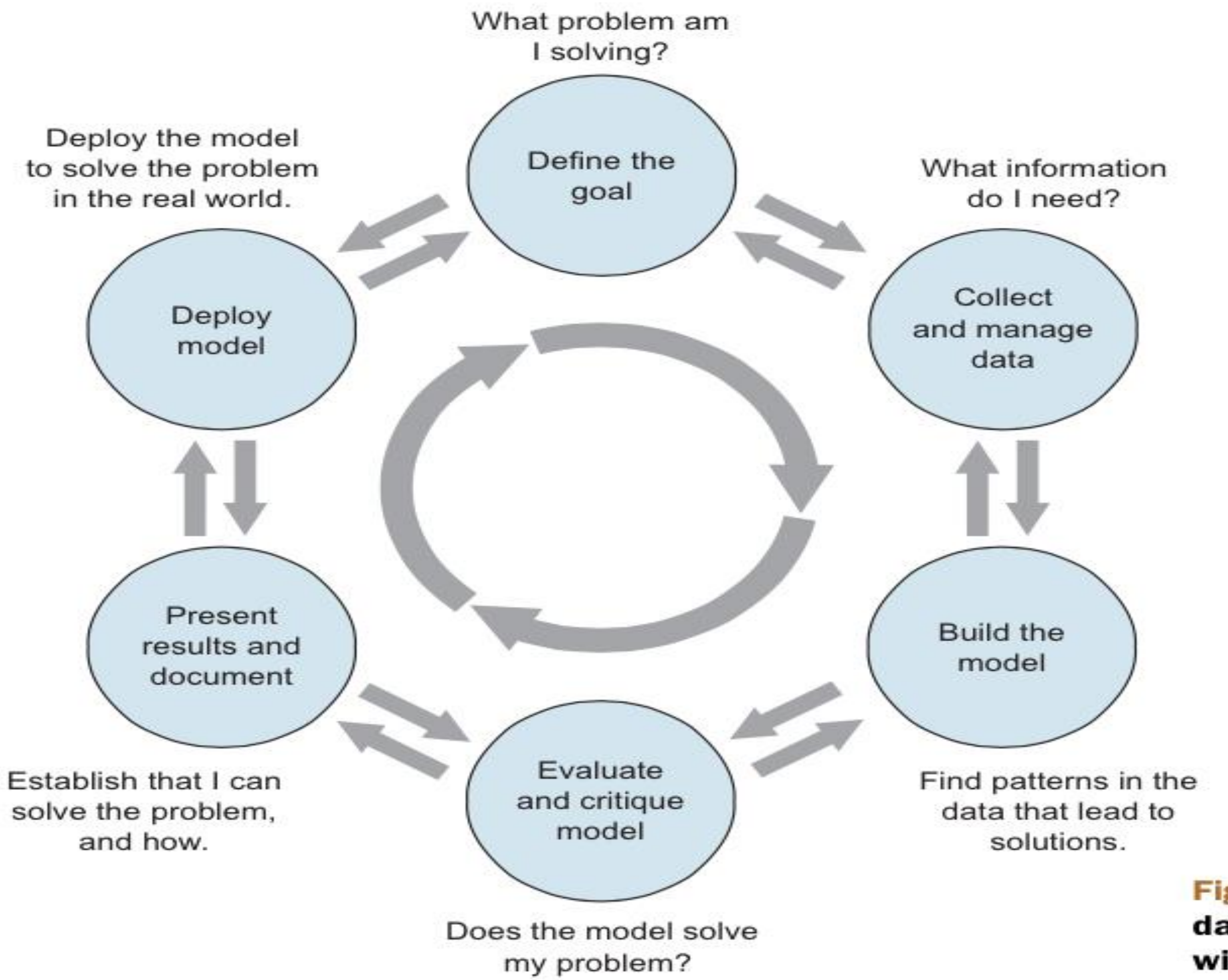
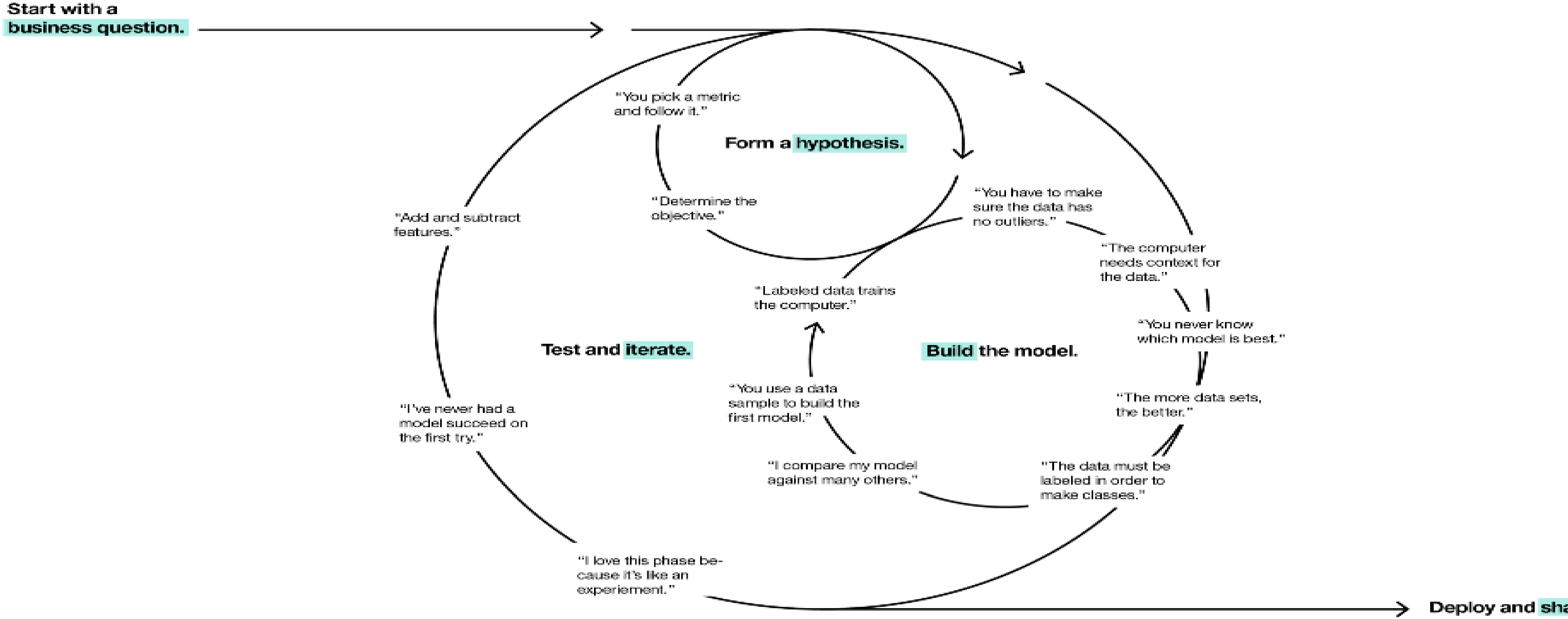

What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Figure 1.1 The lifecycle of a data science project: loops within loops

Cyclical process of data science



Start with a business question.
"You pick a metric and follow it."
Form a hypothesis.
"Determine the objective."
"Add and subtract features."
"You have to make sure the data has no outliers."
"The computer needs context for the data."
"Labeled data trains the computer."
Test and iterate.
Build the model.
"You never know which model is best."
"You use a data sample to build the first model."
"I've never had a model succeed on the first try."
"The more data sets, the better."
"I compare my model against many others."
"The data must be labeled in order to make classes."
"I love this phase because it's like an experiement."
Deploy and sha

Data science solution building/development

1. Formulation of business question（商业问题表述）
2. Mapping of the business question into a technical problem（将业务问题映射为技术问题）
3. Set up acceptable performance/accuracy parameters for the answers（设置可接受的性能、准确性参数）
4. Determination of data availability（确定可获得的数据）
   a. Data sourcing （数据源）
   b. Data cataloguing （数据类别）
5. Preliminary verification that the data available can be used for answering the business question（初步验证获得的数据可以回答该问题）
6. Identify target environments（确定目标环境）
   a. Development （研发）
   b. Production/deployment （生产/部署）

Contd…

Contd…

7. Data Sourcing/Acquisition （数据采集）
8. Data preparation （数据准备）
   a. Data cleaning （数据清洗）
   b. Data transformation/ feature extraction （数据转换/特征提取）
   c. Data sampling and partitioning （数据取样和划分）
   d. Data fusion （数据融合）
9. Data exploration and Understanding （数据探索和理解）
   a. Data visualization （数据可视化）
   b. Data statistics （数据统计）
   c. Data reconciliation （数据对账）
   d. Data preprocessing （数据预处理）

Contd…

Contd…

10. Design solution architecture（设计解决方案架构）
11. Select relevant/possible/available modeling/analysis techniques for solution components （为解决方案组件选择 相关/可能/可用的建模/分析 技术）
    a. Perform modeling/analysis using each of the selected techniques （利用所选技术进行建模/分析）
    b. Evaluation and comparison of performance （性能评估与比较）
        I. Applying trained model on test data （将训练好的模型应用于测试数据）
        II. Analyze the results （分析结果）
            - Result statistics （结果统计分析）
            - Result visualization （结果可视化分析）
    c. Select and/or combine best performing model(s) （选择或组合性能最佳的模型）
12. Build the solution by putting together the components （集成组件形成解决方案）
13. Analyze and validate the solution （分析和验证解决方案）
14. Translate the output in business language （翻译成商业语言）
15. Create visualization and/or data products (e.g., dashboard) comprehensible to the business user （创建业务用户易于理解的可视化或数据产品，如仪表板）

Fundamental Concepts & Exemplary Techniques 1

1. 作为战略资产的数据和数据科学能力
   - 示范技术: Signet Bank to Capitol One.
2. 一组规范的数据挖掘任务：数据挖掘过程；监督与非监督数据挖掘
   - 示范技术: 用于数据挖掘的跨行业标准流程
3. 识别有信息量的属性：属性递进式选择:
   - 示范技术：寻找相关性；属性选择；归纳树
4. 根据数据寻找“最优”模型参数：数据挖掘目标的选择，目标函数，损失函数:
   - 示范技术：线性回归; 逻辑斯蒂回归；SVM

Fundamental Concepts & Exemplary Techniques 2

5. 泛化、拟合、过拟合、复杂度控制:
   - 示范技术: Cross-validation; Attribute selection; Tree pruning; Regularization.

6.计算数据描述对象的相似度;利用相似度进行预测;聚类即基于相似度的分割::
   - 示范技术:搜索相似的实体;最近邻方法;聚类方法;用于计算相似度的距离度量.

7.仔细考虑需要从数据科学的结果中得到什么;期望价值作为关键的评价框架;考虑适当的比较基线::
   - 示范技术:各种评估指标;估算成本和效益，计算预期利润;创建用于比较的基线方法

8.基本概念:各种不确定性下模型性能的可视化;进一步考虑从数据挖掘结果中期望得到什么::
   - 示范技术: Profit curves; Cumulative response curves; Lift curves; ROC curves.（利润曲线;累积响应曲线;升力曲线;ROC曲线）

Source: Data Science for Business (2013) pages 111, 141, 187, & 209.

Fundamental Concepts & Exemplary Techniques 3

9.基于贝叶斯规则的显式证据组合、基于条件独立性假设的概率推理::

- 示范技术: Naive Bayes classification; Evidence lift（证据移除）.

10.构建数据挖掘友好的数据表示的重要性、用于数据挖掘的文本表示::

- 示范技术: Bag of words representation; TFIDF calculation; N-grams; Stemming; Named entity extraction; Topic models.

11.用数据科学解决业务问题，从分析工程开始:基于可用的数据、工具和技术设计分析解决方案::

- 示范技术: Expected value as a framework for data science solution design.

12.基本概念是许多常见数据科学技术的基础;熟悉数据科学构建模块的重要性:

- 示范技术: 相关与共生; 行为分析；链接预测；数据还原；隐式信息挖掘；推荐；误差的偏差-方差；模型的集成；因果推理

Source: Data Science for Business (2013) pages 233, 251, 279, & 291.

Fundamental Concepts & Exemplary Techniques 4

13.数据驱动是业务成功的基础;通过数据科学获取和保持竞争优势;谨慎管理数据科学能力的重要性。

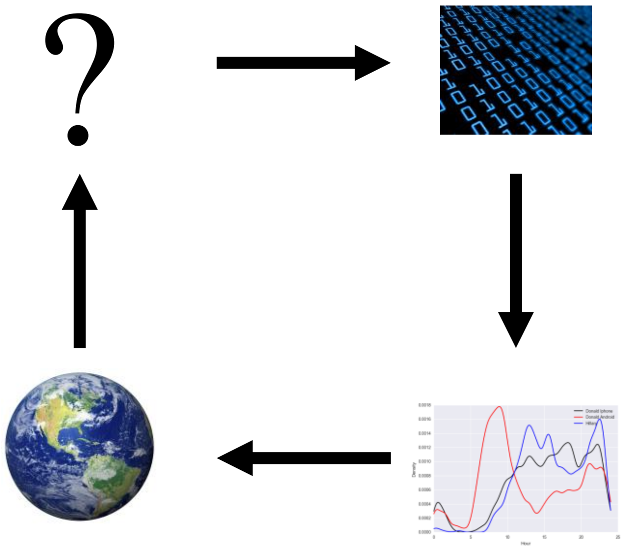
- 示范技术:检查数据科学的案例研究和所需的数据科学建议.

14. 结论:

- 示范技术: If you can’t explain it simply, you don’t understand it well enough（如果你不能简单说清楚，就是你还没有完全明白。）— Albert Einstein
- 提案评审指南:有效的数据分析思维应该允许你系统地评估潜在的数据挖掘项目.

Source: Data Science for Business (2013) pages 315, 333, 349, 353, 357, & 361.

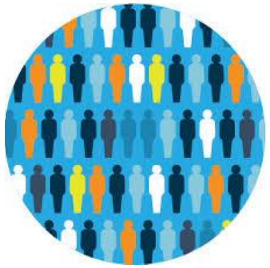
I: Problem Formulation and Experimental Design



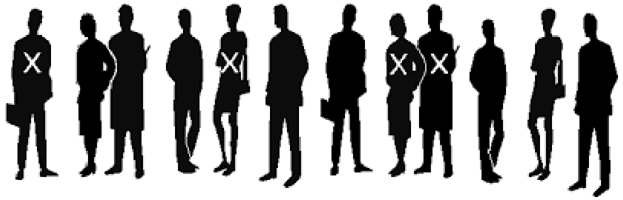
I. Data Analytics QPR-V

- **Q**uestion

- **P**opulation



- **R**epresentative data collection (data neutral)



- **-** (to be filled in throughout the course)

- **V**etting（审查） or validation of answers

ADOPT
Validate
ASSESS
Validate
Impact
Validate
IMPLEMENT

Q ? Question, question, question

- Domain question to answer

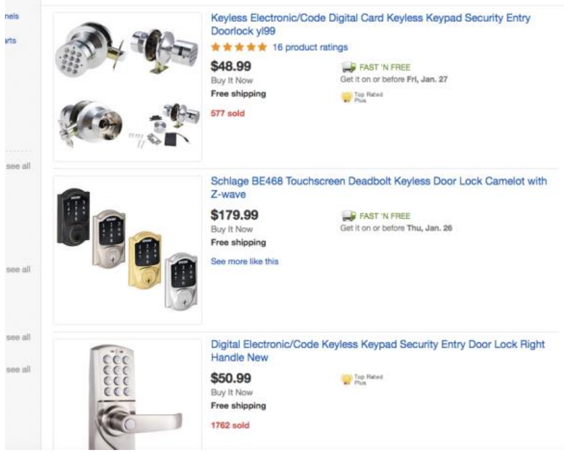
- Examples
  - ○ Why didn’t the polls work well?
  - ○ Does smoking cause cancer?
  - ○ Does BrainPlus IQ work?

Write down question as record keeping

More questions:

- Which restaurant to go to in Berkeley?

- Should eBay update their door lock ranking?



nels

arts

see all

see all

see all

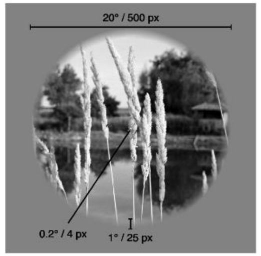
see all

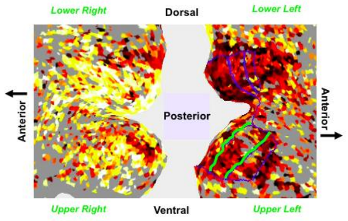
Keyless Electronic/Code Digital Card Keyless Keypad Security Entry Doorlock yl99
★★★★★ 16 product ratings
$48.99
Buy It Now
Free shipping
577 sold
FAST 'N FREE
Get it on or before Fri, Jan. 27
Top Rated Plus

Schlage BE468 Touchscreen Deadbolt Keyless Door Lock Camelot with Z-wave
$179.99
Buy It Now
Free shipping
See more like this
FAST 'N FREE
Get it on or before Thu, Jan. 26

Digital Electronic/Code Keyless Keypad Security Entry Door Lock Right Handle New
$50.99
Buy It Now
Free shipping
1762 sold
Top Rated Plus

- How to “read mind”? Or how to reconstruct a movie based on brain fMRI signals?

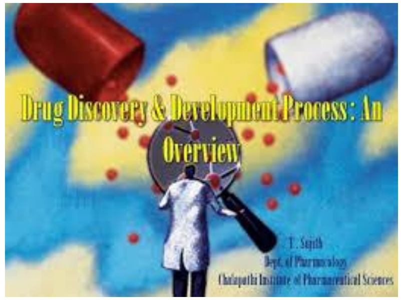




Two types of questions

- Hypothesis driven: whether a new drug works

- Discovery-driven: search for new diseases that an existing drug could treat



Drug Discovery & Development Process: An Overview

T. Sujith
Dept. of Pharmacology
Chalapathi Institute of Pharmaceutical Sciences

- Separate hypothesis generation and decision on hypothesis: discovery phase vs. validation phase

Sample-split is recommended:

Half of the data

Half of the data

Validation: what makes sense below?

Feasibility of question

- Is the question feasible to answer using data?

- We need to translate the questions into a more precise question:

Why didn’t polls work? → how did we predict the result of the 2016 election in Oct. 2016? → how did we predict the popular vote?

- Did Gallup Poll have the resources (energy, expertise, relevant data) to do the prediction?

P. “Population” in the question

- Population is the relevant group of people (objects, units) that the data-driven answer to the question will be applied to

- Write the population down as a record for any DS project

“Population” for the 2016 election

- Oracle (future) population

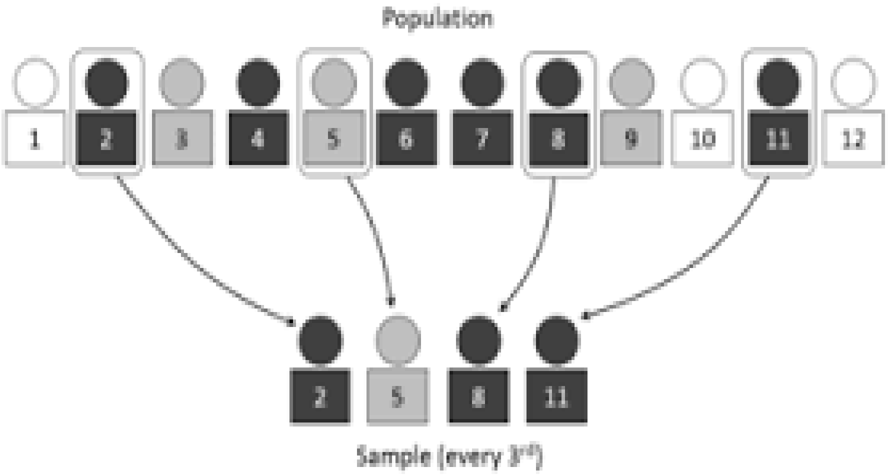
  all the votes casted on election day

- Short of magic, we want to predict or guess these votes ahead of time

- How to go about this?

R. Representative data collection

- Data to answer the question should be "representative" of the relevant group (or population), or data neutral



Never a brand new problem…

- There is always prior knowledge or data
- Qualitative: literature, human co-workers
- Quantitative: previous data, new pilot study

Election example: data collection

- Population: all the votes to be casted on election day in CA -- impossible since we can’t time travel

  Even if we can, too much money to ask everyone

- How about a representative sample of the election day votes, also for the sake of money -- can’t time travel so possible only under assumptions

Two layers of uncertainty on Nov. 8

- Who are going to vote on election day?

- How is a voter going to vote?

- We can’t ask all voters, even if we could, people change their minds – for 2016, about 12% undecided voters, compared with 4-5% in the past

Assumptions in the election problem

1. Today’s votes are the same as on the election day: People are asked what they would vote if they are voting today –

2. People are telling the truth

3. Undecided voters vote similarly as decided voters

4. The polled group is representative of the voter population

How to make the sample representative or data neutral?

- A whole field of statistics, **Survey Sampling**, has been devoted to this. It recommends random sampling of different kinds.

- Simplest: **simple random sampling (SRS)** without (or with) replacement – putting all entities of the population in a hat and randomly drawing one by one

One particular poll: Gallup Poll

- Simplest form: 1000 SRS samples from numbers of phones (landline and cell phones) (one landline can map to many voters).

- What is the population? Is it the same as the voter population?

- What if people without phones voted very differently from people with phones?

Gallup Poll: some quick calculations

- Suppose Gallup poll is an SRS of the voter population on election day:

= , and previous assumptions

139 million votes, 68 for Trump, 71 for Clinton – less than 2% diff.

- Margin of error is about 3%, one Gallup Poll is not enough to predict such a close race **even if all the assumptions hold.**

Vetting results: election

- Votes on election day

- Prediction on popular votes held (by luck), but not for election outcome.

Vetting data results in general

- Prediction on test data
- Stability analysis
- Post-modeling EDA or visualization
- Domain knowledge verification
- …

Longer time scale

- Down-stream consequences
- Further studies
- …

Data science lessons

- Analytical algorithms CAN NOT automatically detect non-representative or biased samples

- Shoe leather work needs to be taken seriously and with analytical algorithms:

Information about undecided voters and people who do not respond to polls can be obtained only through ground operatives in their talking and interactions with such people.

II. Experimental design

The science and subfield of statistics about how to collect data effectively…

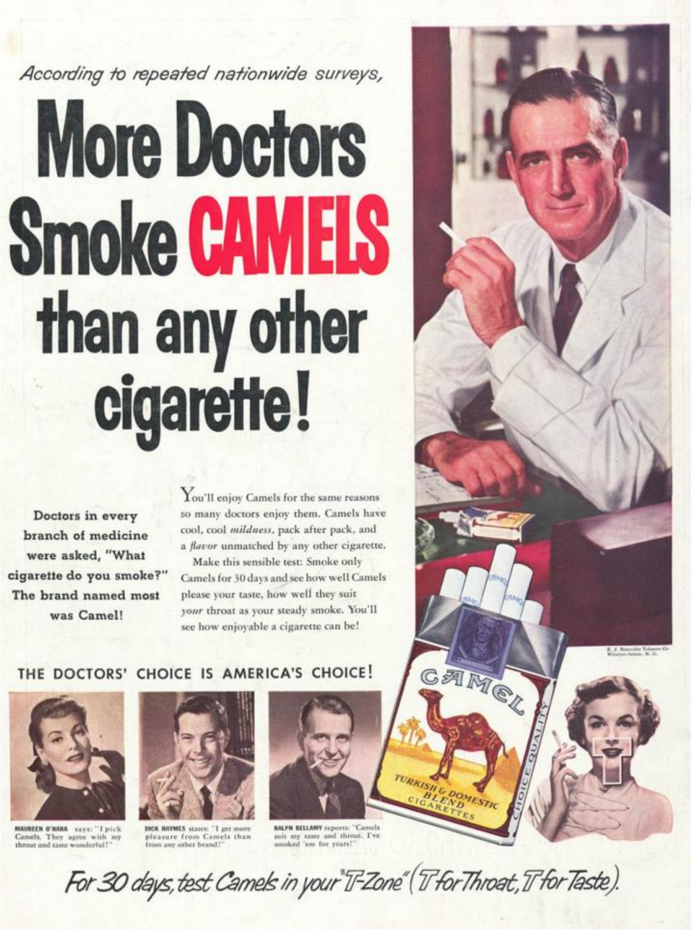
R. A. Fisher (1890-1962)
Founding father of Modern Statistics
Geneticist

"There's a flaw in your experimental design. All the mice are scorpios."

Smoking causes lung cancer?

You can smoke fewer cigarettes by smoking longer ones.
"It's wacky, but it works."

Population?



According to repeated nationwide surveys,

More Doctors Smoke CAMELS than any other cigarette!

Doctors in every branch of medicine were asked, "What cigarette do you smoke?" The brand named most was Camel!

You'll enjoy Camels for the same reasons so many doctors enjoy them. Camels have cool, cool *mildness*, pack after pack, and a *flavor* unmatched by any other cigarette.

Make this sensible test: Smoke only Camels for 30 days and see how well Camels please your taste, how well they suit *your* throat as your steady smoke. You'll see how enjoyable a cigarette can be!

THE DOCTORS' CHOICE IS AMERICA'S CHOICE!

MAUREEN O'HARA says: "I pick Camels. They agree with my throat and taste wonderful!"

DICK HAYMES states: "I get more pleasure from Camels than from any other brand!"

RALPH BELLAMY reports: "Camels suit my taste and throat. I've smoked 'em for years!"

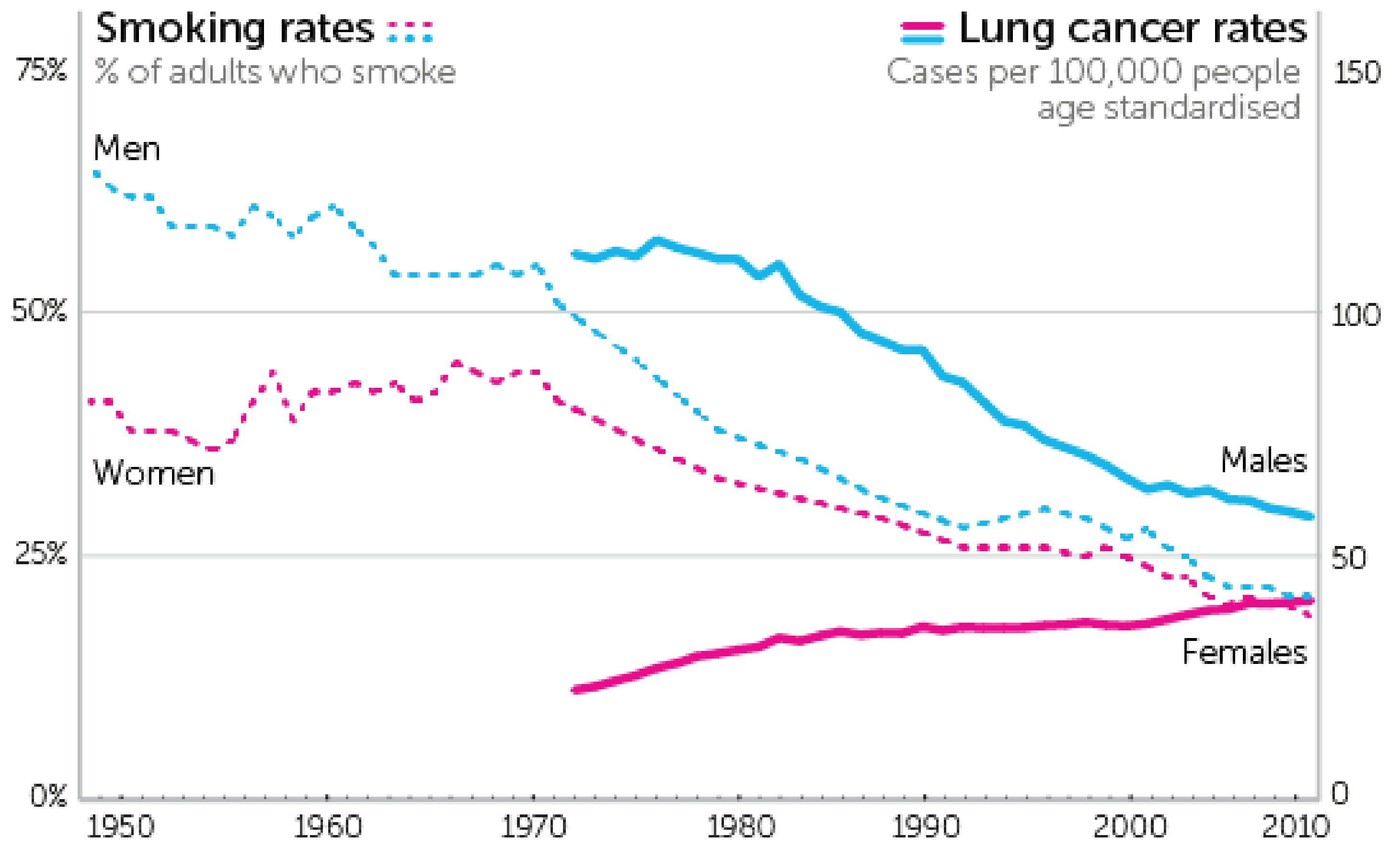
For 30 days, test Camels in your "T-Zone" (T for Throat, T for Taste).

Thanks to R. Barter for help on some smoking-cancer slides

Does smoking cause cancer?

UK Cancer Research

SMOKING RATES AND LUNG CANCER RATES IN THE UK



Can we conclude that

smoking causes cancer for men;

but is good for women??

Possible explanations based on "colleague-sourcing"

- Lung cancer is not all the same （肺癌并不都是相同的）
- Only 10-15% is closely associated to smoking
- Another type is not, but occurs more among women
- Women started smoking later than men and the cohort（同龄人）with peak smoking is still working its way through the population
- Women are more susceptible to tobacco toxins（毒素）
- …

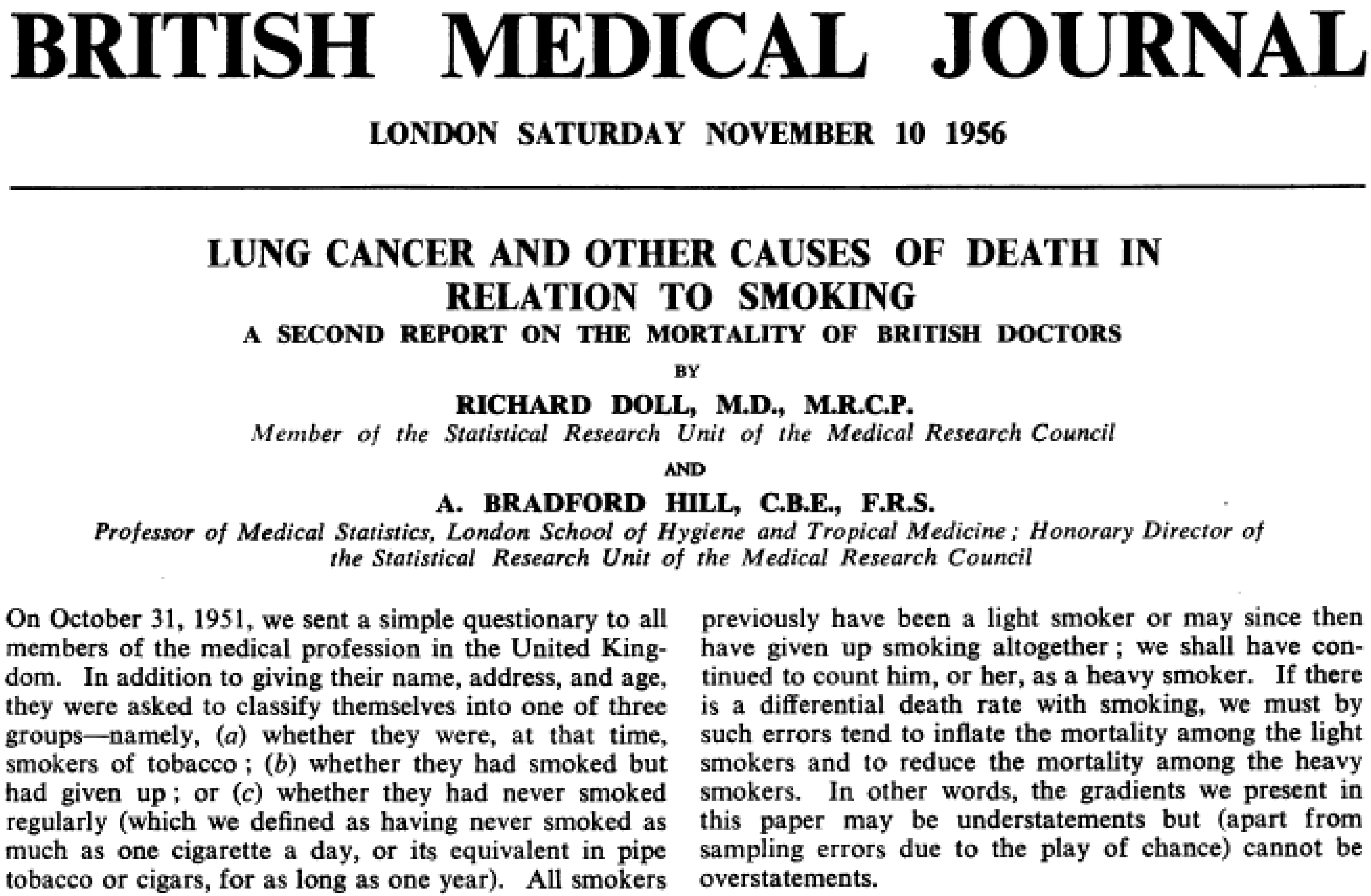
Thanks to P. Stark, P. Ding, J. Sekhon

Lessons

- Ecological（生态学的） correlation does not imply correlation at person level (ecological correlation = correlation between rates)
- Association (correlation) is not causation
- Confounding（混杂） factors are always lurking in the back

(confounding factor: a possible driver for both smoking and lung cancer, e.g. genetics)

The first solid epidemiological（流行病学）evidence, or observational study



BRITISH MEDICAL JOURNAL

LONDON SATURDAY NOVEMBER 10 1956

LUNG CANCER AND OTHER CAUSES OF DEATH IN RELATION TO SMOKING

A SECOND REPORT ON THE MORTALITY OF BRITISH DOCTORS

BY

RICHARD DOLL, M.D., M.R.C.P.
Member of the Statistical Research Unit of the Medical Research Council

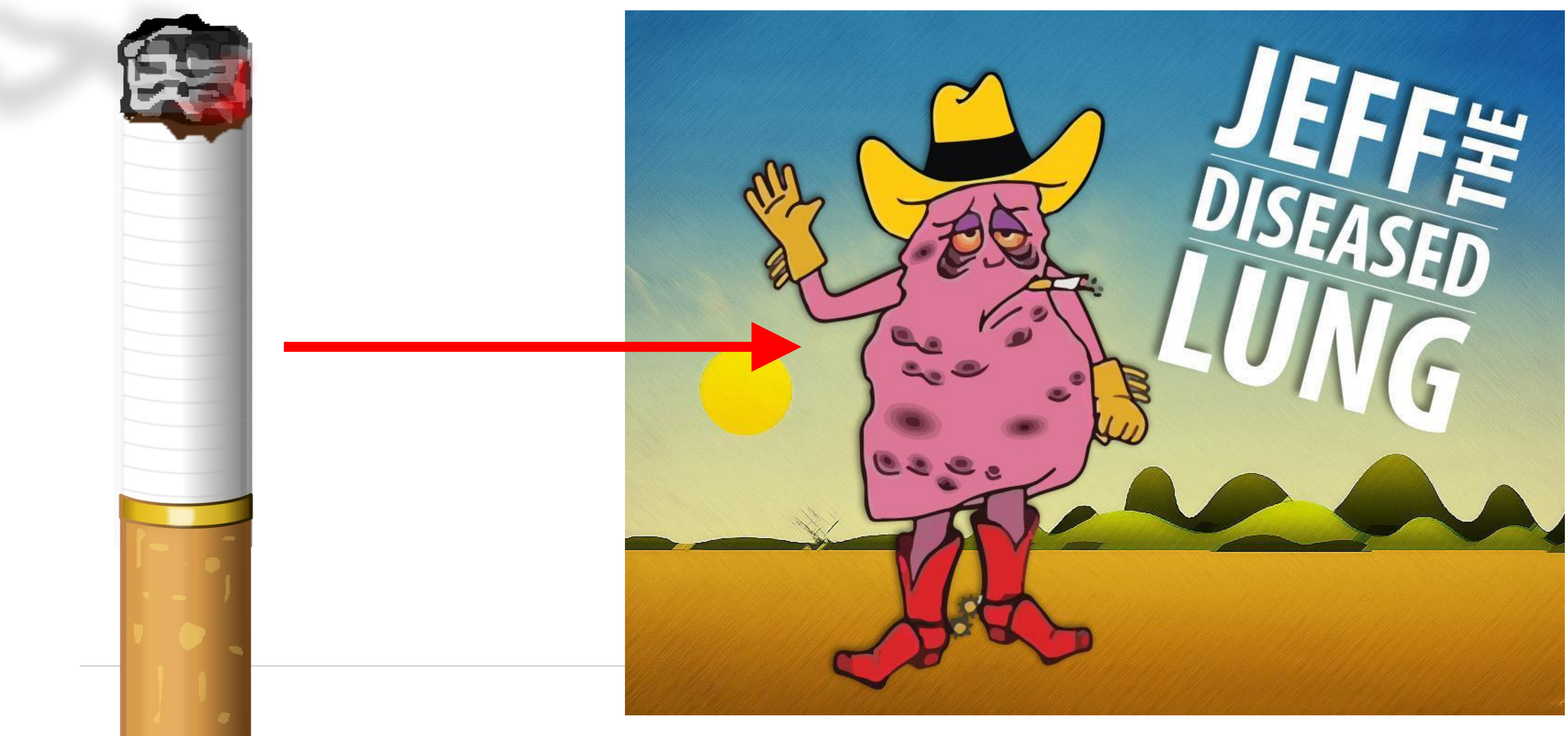
AND

A. BRADFORD HILL, C.B.E., F.R.S.
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

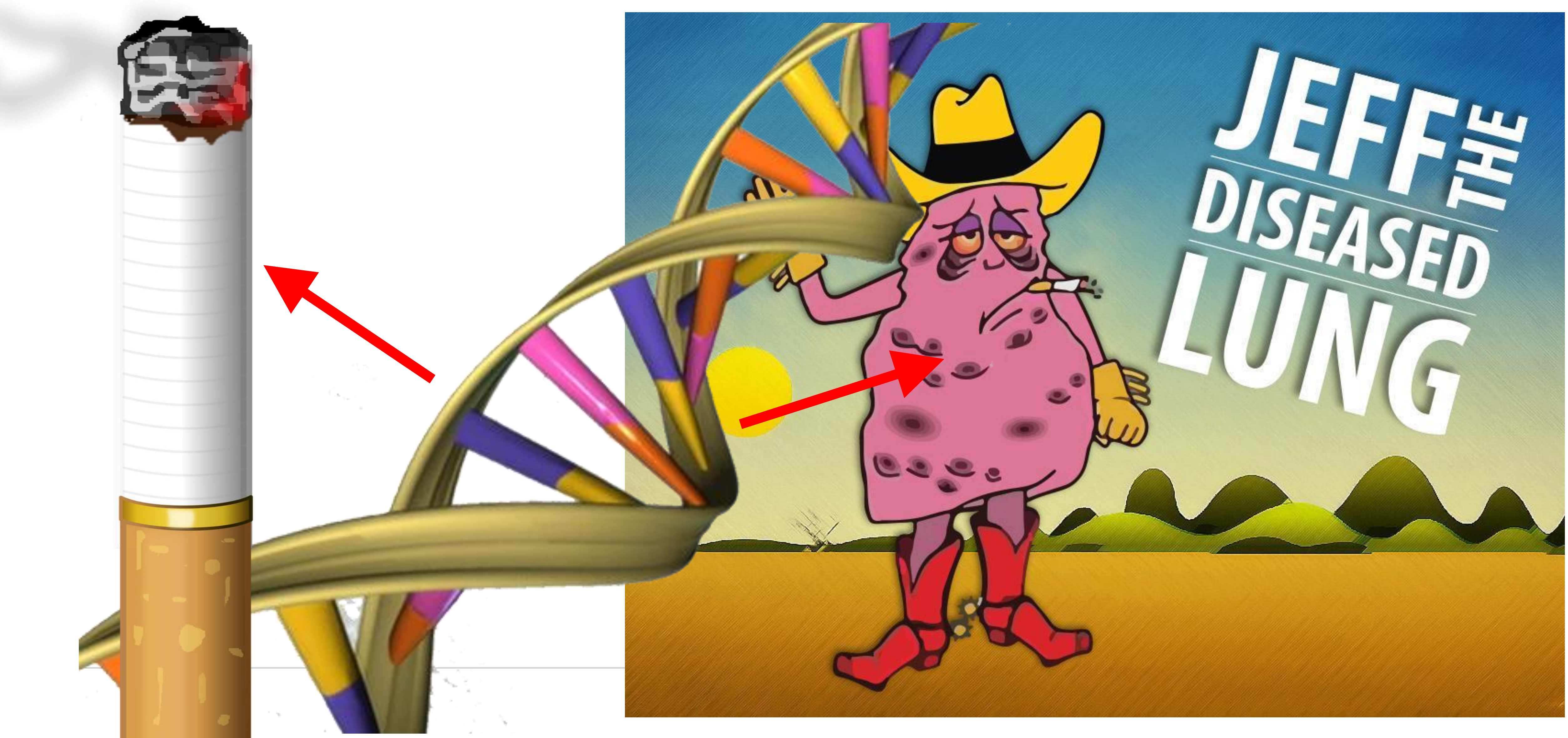
On October 31, 1951, we sent a simple questionary to all members of the medical profession in the United Kingdom. In addition to giving their name, address, and age, they were asked to classify themselves into one of three groups—namely, (*a*) whether they were, at that time, smokers of tobacco; (*b*) whether they had smoked but had given up; or (*c*) whether they had never smoked regularly (which we defined as having never smoked as much as one cigarette a day, or its equivalent in pipe tobacco or cigars, for as long as one year). All smokers

previously have been a light smoker or may since then have given up smoking altogether; we shall have continued to count him, or her, as a heavy smoker. If there is a differential death rate with smoking, we must by such errors tend to inflate the mortality among the light smokers and to reduce the mortality among the heavy smokers. In other words, the gradients we present in this paper may be understatements but (apart from sampling errors due to the play of chance) cannot be overstatements.

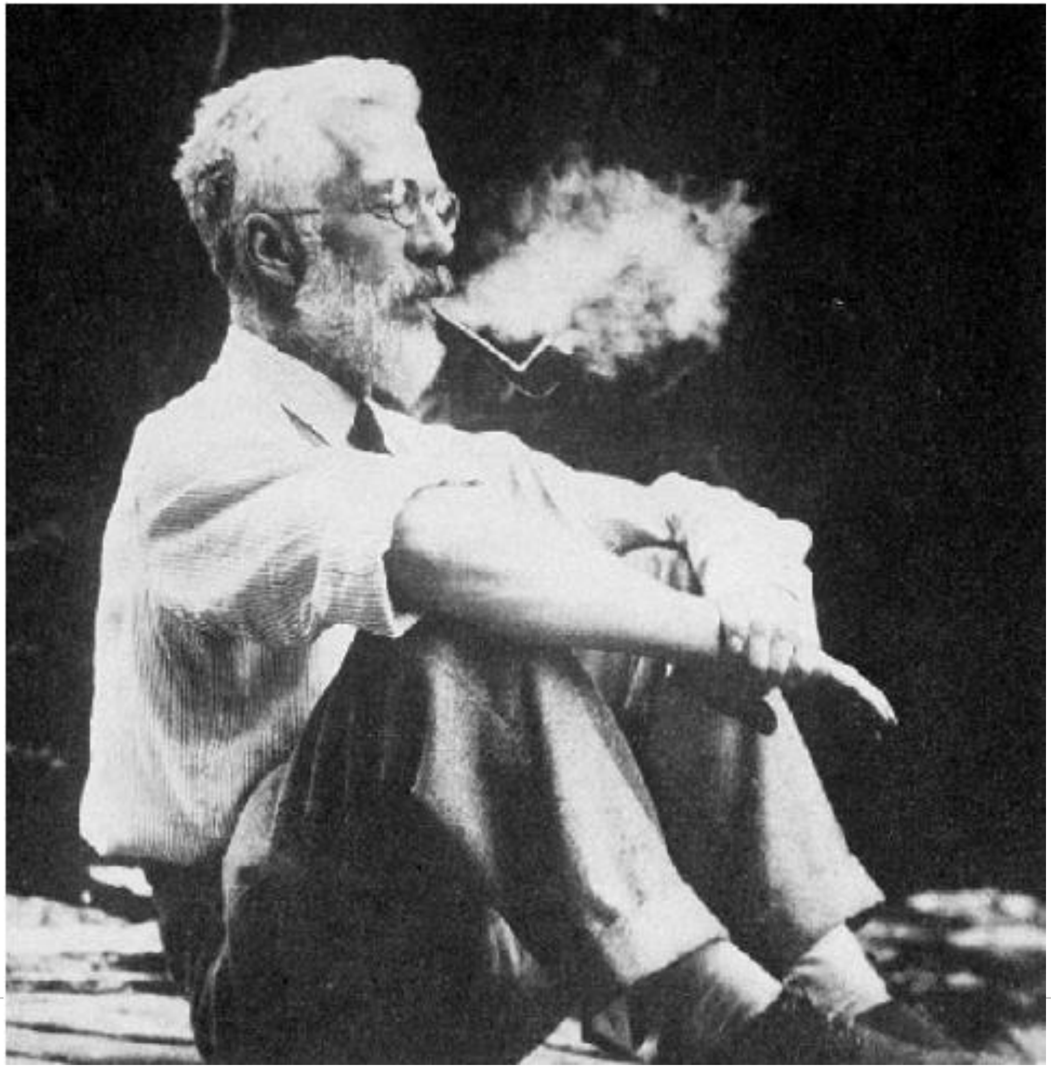
Cigarettes cause lung cancer!



Genetics cause both smoking and lung cancer? Or genetics could be a confounding factor

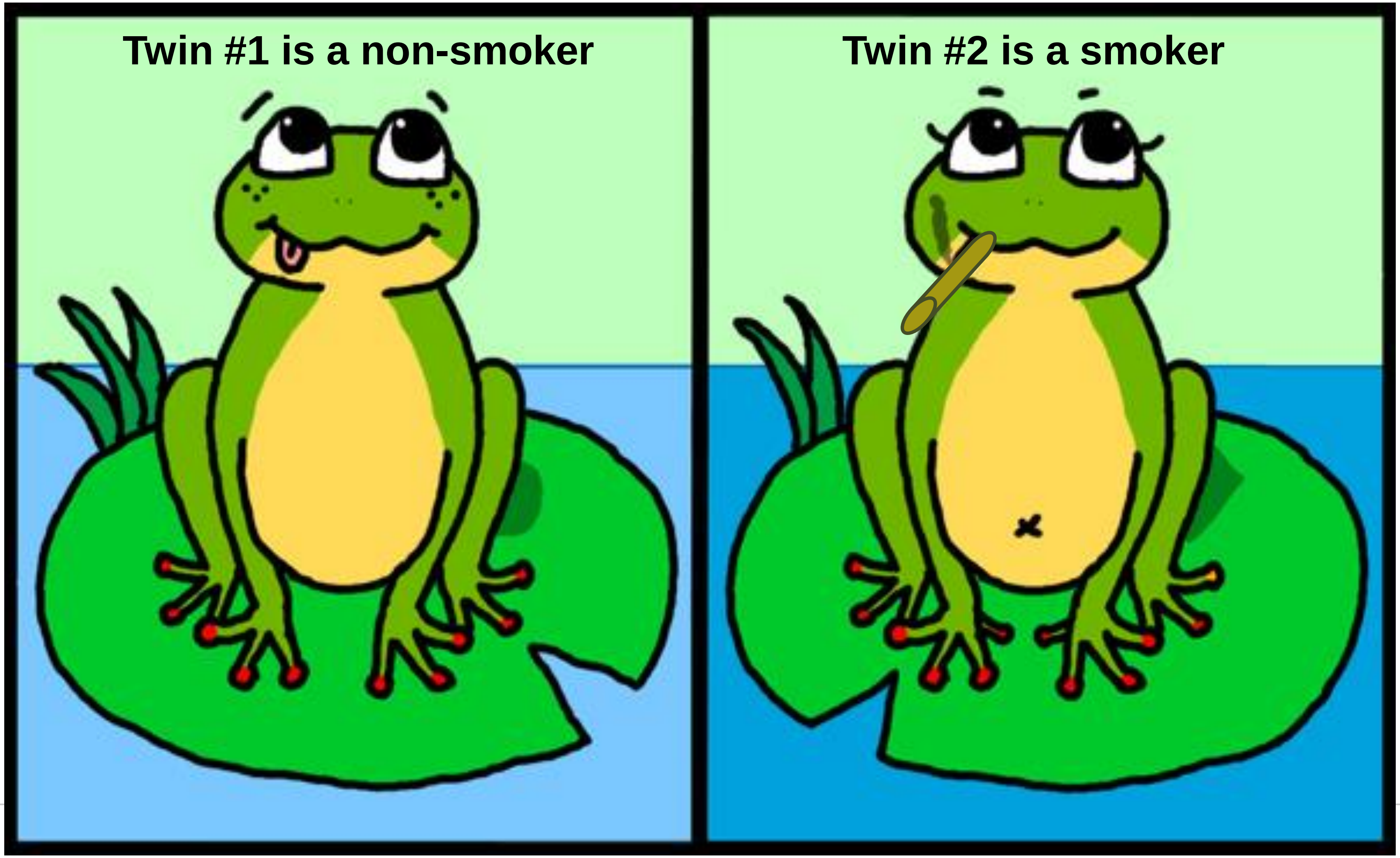


R. A. Fisher strongly believed in a common genetic cause!

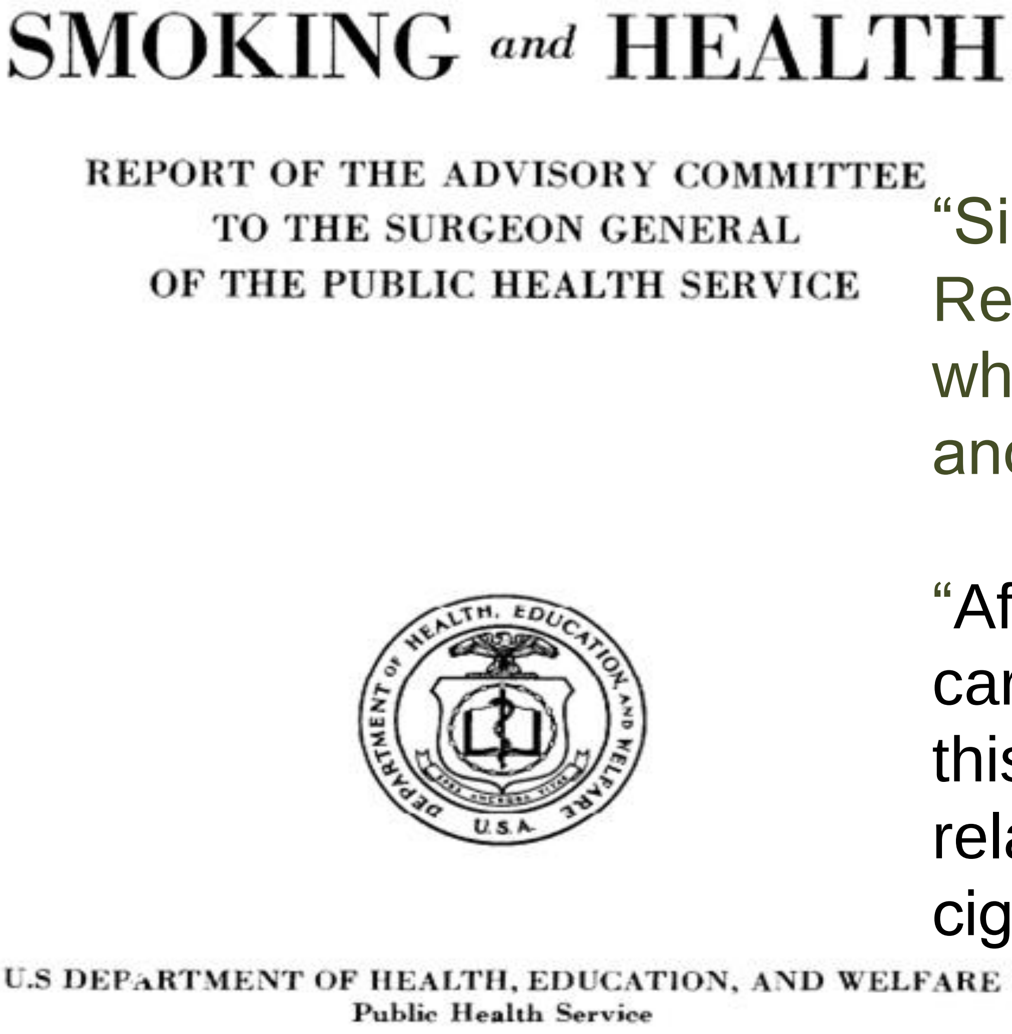


17 February 1890 –
29 July 1962

Control for genetic differences: compare identical twins!



The evidence kept piling in from **observational studies**… Until finally in 1964



SMOKING *and* HEALTH

REPORT OF THE ADVISORY COMMITTEE
TO THE SURGEON GENERAL
OF THE PUBLIC HEALTH SERVICE

U.S DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

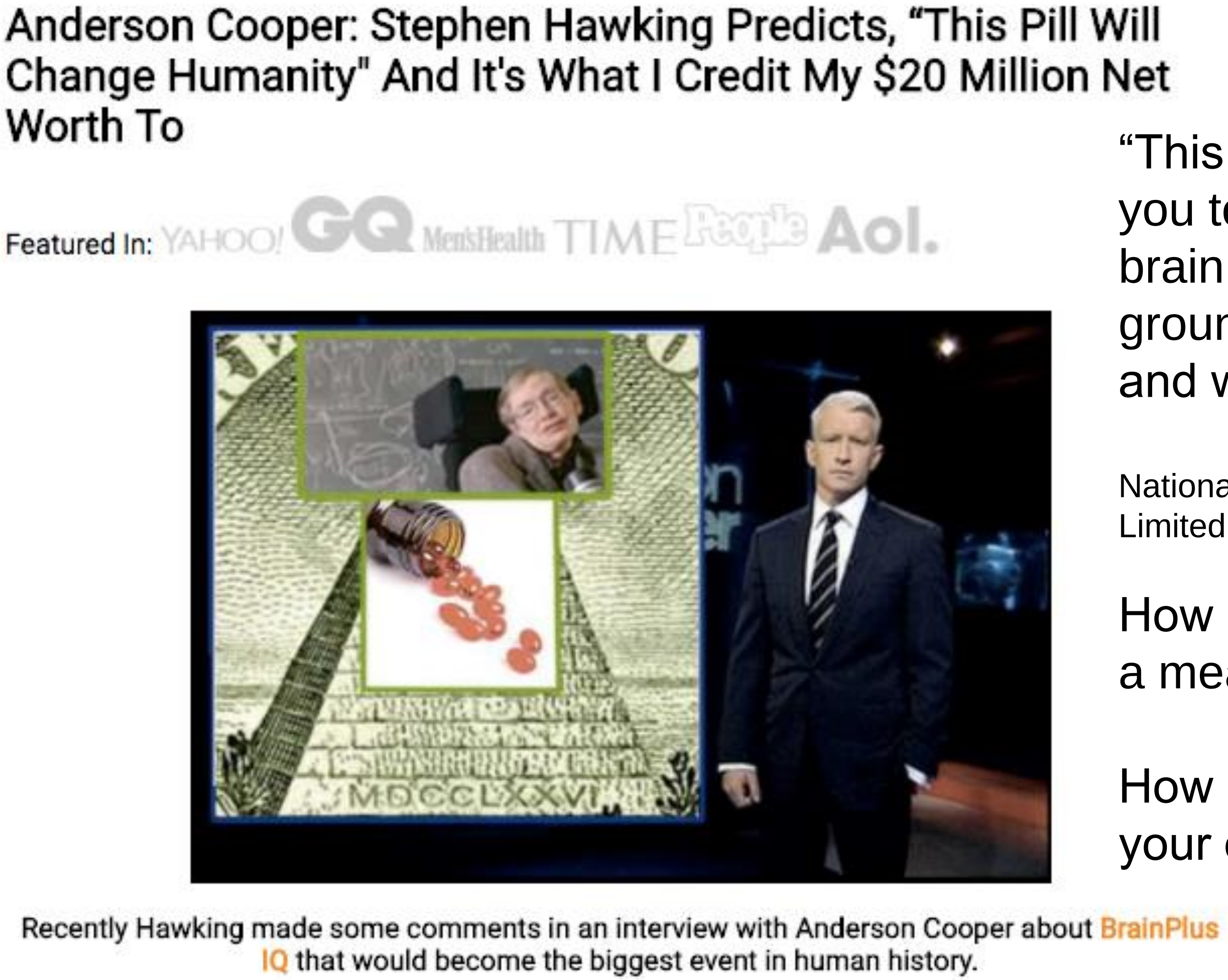
"Since 1939 there have been 29 Retrospective(回顾性的) studies of lung cancer alone which have varying degrees of completeness and validity."

"After appraising(鉴定) 16 independent studies carried on in five countries over a period of 18 years, this group concluded that there is a causal relationship between excessive smoking of cigarettes and lung cancer."

What is an ideal experiment to see whether smoking causes cancer?

- What is the population?
- Are you going to ask some people to smoke? What is the difference from people choose to smoke?
- And then what?

Does BrainPlus IQ work?



Anderson Cooper: Stephen Hawking Predicts, “This Pill Will Change Humanity" And It's What I Credit My $20 Million Net Worth To

Featured In:

Recently Hawking made some comments in an interview with Anderson Cooper about BrainPlus IQ that would become the biggest event in human history.

“This pill unlocks your brain power, allowing you to adventure further inside your own brain than ever before. This is the most groundbreaking BrainPlus IQ ever created, and we had to showcase it to the world!”

National Geographic
Limited Edition Cover Page

How do we translate “change humanity” into a measureable outcome?

How do we measure “adventure further inside your own brain”?

Does BrainPlus IQ work?

- What is the population? All people who want to take BrainPlus IQ (many are unborn yet)…

- An easier question: is BrainPlus IQ better on average for people who have signed up for a study conducted by the company?

  Translate “work” into IQ measure…

Population: two potential outcomes (IQ test score) while taking BrainPlus IQ and taking a placebo, respectively

Data collection by observing the population

- Administer BrainPlus IQ on Monday
- Administer Placebo on Tuesday

What assumption are we making?

If want to remove the assumption

- Administer BrainPlus IQ to half of the group and placebo to the other half

How would you choose the half?

Gold Standard of Causal Evidence

Randomization

which often relies on pseduo-random number generators (PRNGs) – deterministic processes that can be flawed …

It randomly assigns each subject to the treatment or control group – to take BrainPlus IQ or Placebo – or a random split.

Virtues of Randomization

- Reduce or combat confounding (often not to zero)
- Ground probabilistic reasoning and calculation

Three principles of experimental design

- Replication
- **Randomization**
- Blocking (reducing variability by using extra information – highly needed for the election situation)

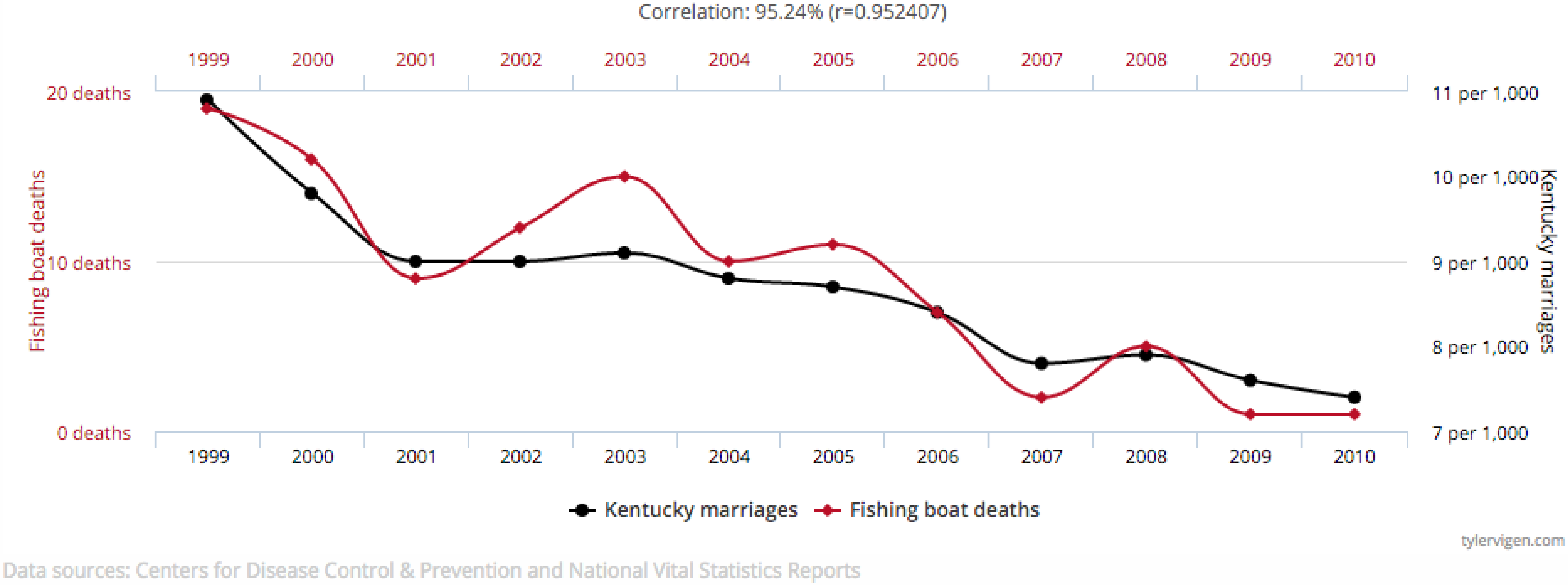
Summary

- Q (not unique: translating Q in English into a Q about data…)
- P
- R
- -
- V
  - Association is not causation
  - SRS, Randomization (randomized experiment vs. observational study)
  - Confounding factors
  - Ecological correlation

Last but not least:

a data-driven claim

Marriage “causes” drowning...



Data sources: Centers for Disease Control & Prevention and National Vital Statistics Reports

Q&A