Data Mining

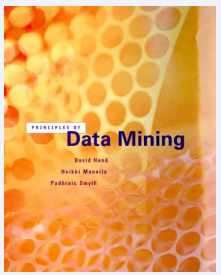
What is it?

Ch. 1

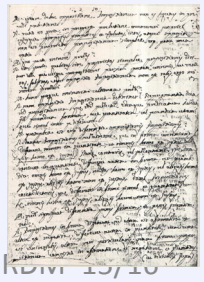
What is Data Mining?

"Data mining is the process of extracting hidden patterns from data."

"An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results."



"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

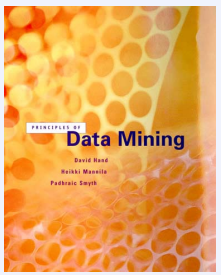


"Data mining, in a broad sense, is the set of techniques for analyzing and understanding data."

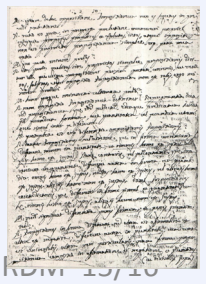
What is Data Mining?

"Data mining is the process of extracting hidden patterns from data."

~~"An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results."~~

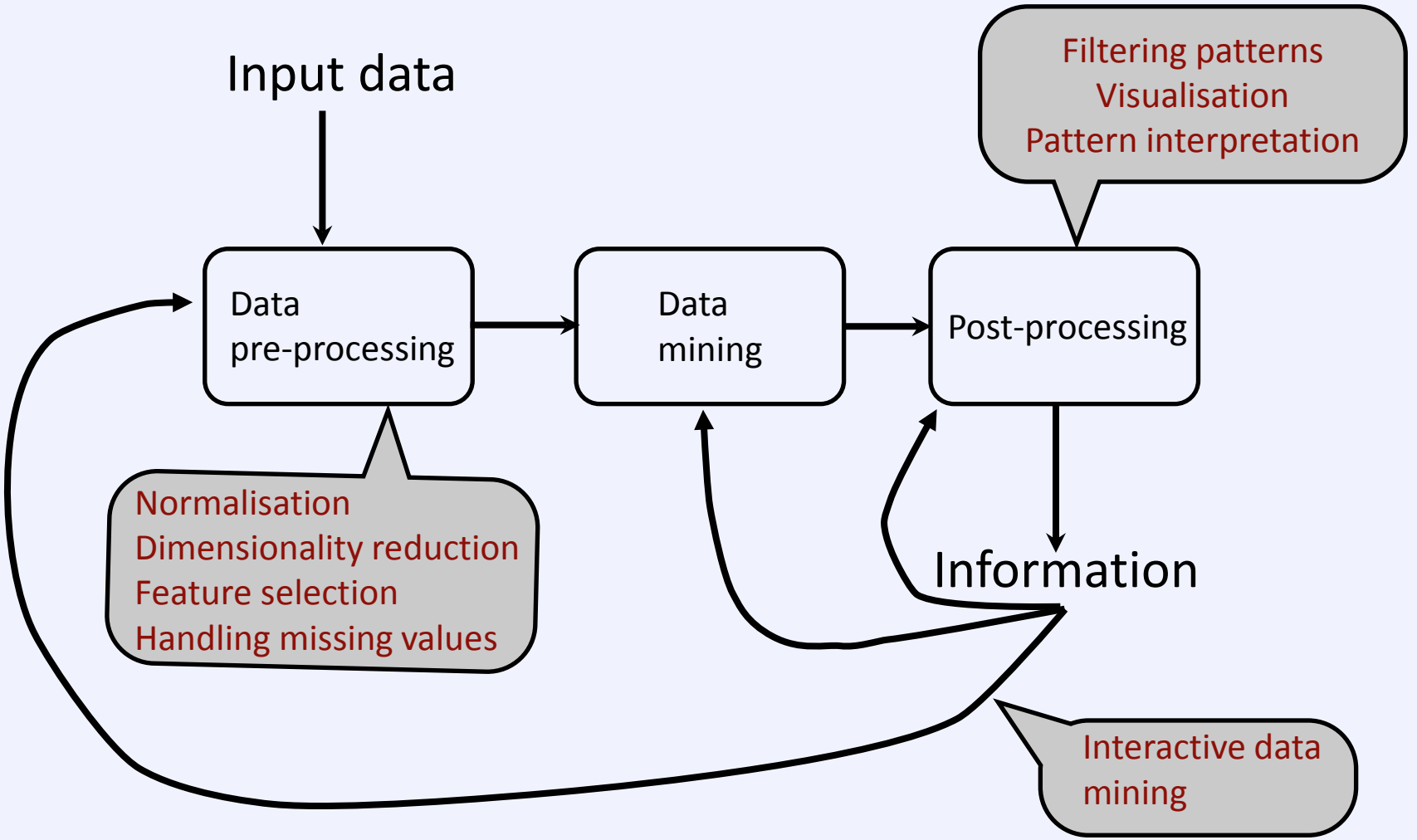


"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."



"Data mining, in a broad sense, is the set of techniques for analyzing and understanding data."

The KDD (Knowledge Discovery from Data) Process



Data Mining vs. Information Retrieval

IR is answering questions the user asked
DM is answering questions the user **didn't** ask

"Show me the web pages relevant to this query"
vs.
"Show me **the interesting patterns** in
the contents of these web pages"

Vague problem… How to define interestingness?
How to evaluate results?

Data Mining's position in Science

Data mining uses statistics to infer from data

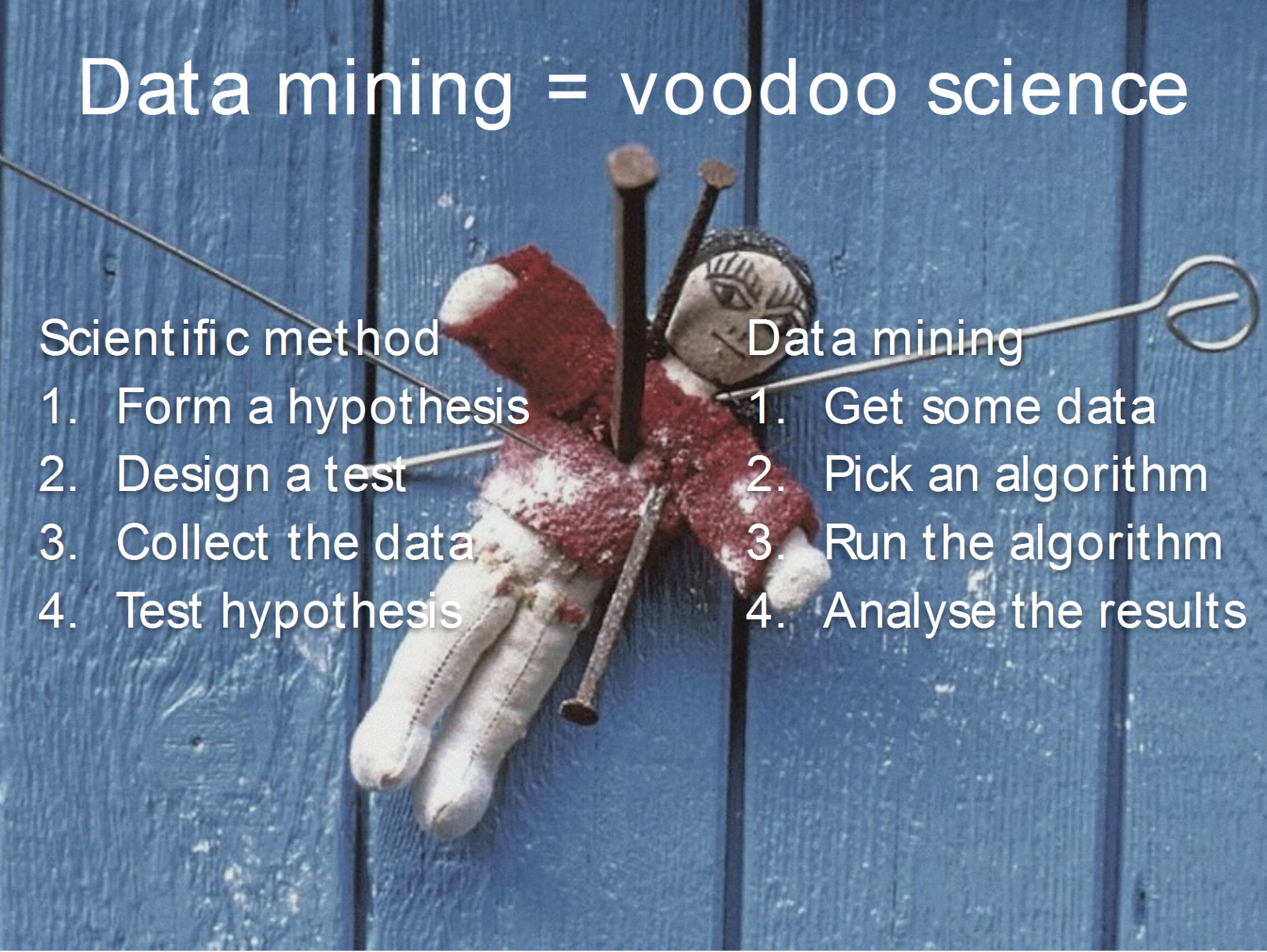
- is data mining just a fancy name for statistics?

Data mining uses methods to learn unseen patterns

- is data mining just a boring name for machine learning?

Is data mining voodoo science?

Data mining = voodoo science



Scientific method

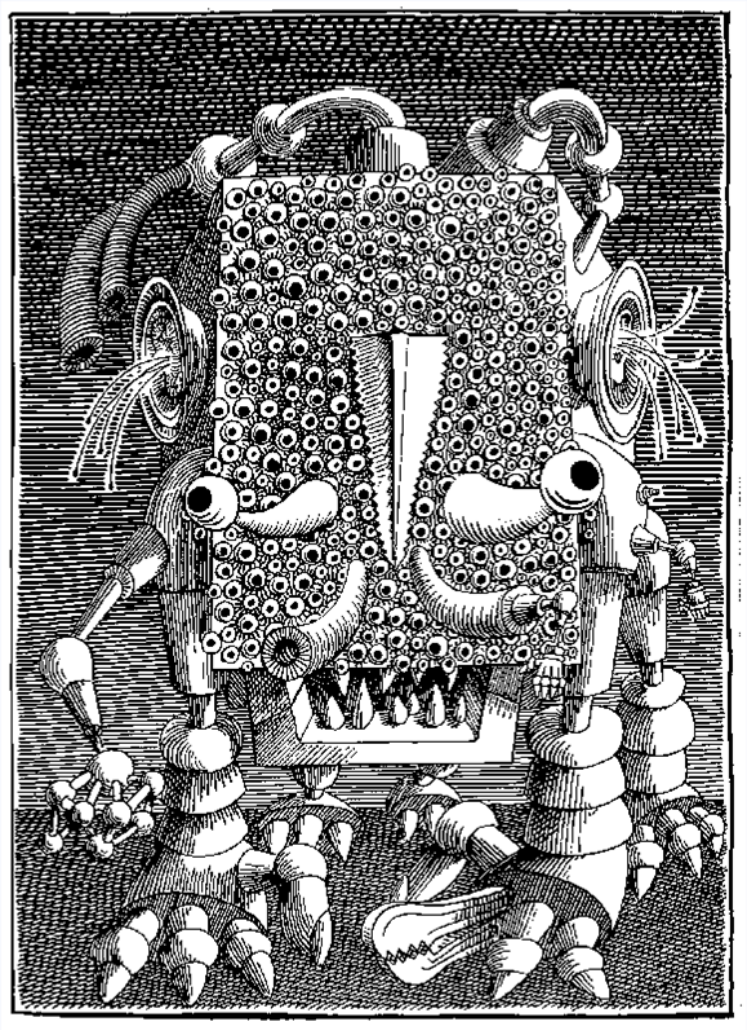
1. Form a hypothesis
2. Design a test
3. Collect the data
4. Test hypothesis

Data mining

1. Get some data
2. Pick an algorithm
3. Run the algorithm
4. Analyse the results

Why Data Mining?

Why Data Mining?

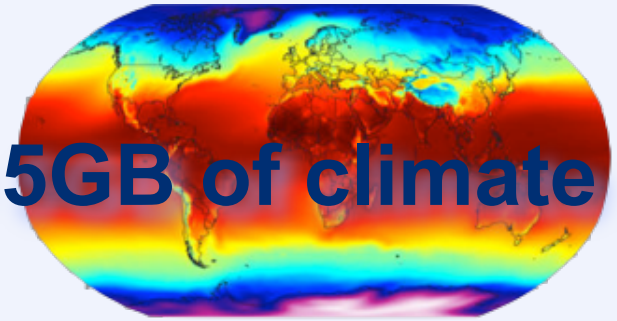


The ”PHT” Pirate wanted all information of the world. But before he realized most of it was useless, he was already buried under it.

—Stanisław Lem, *The Cyberiad*


BIG
DATA

Data, data, data



≈ 5GB of climate data

500 000 000 tweets per day

Walmart

1 250 000 transactions per hour

350 000 000 photos. Per day.

Data, data, data

1 250 000 transactions per hour

≈ 5G

To use this data, we need tools to analyse and understand it.

We need data mining.

5

ay.

tweets per day

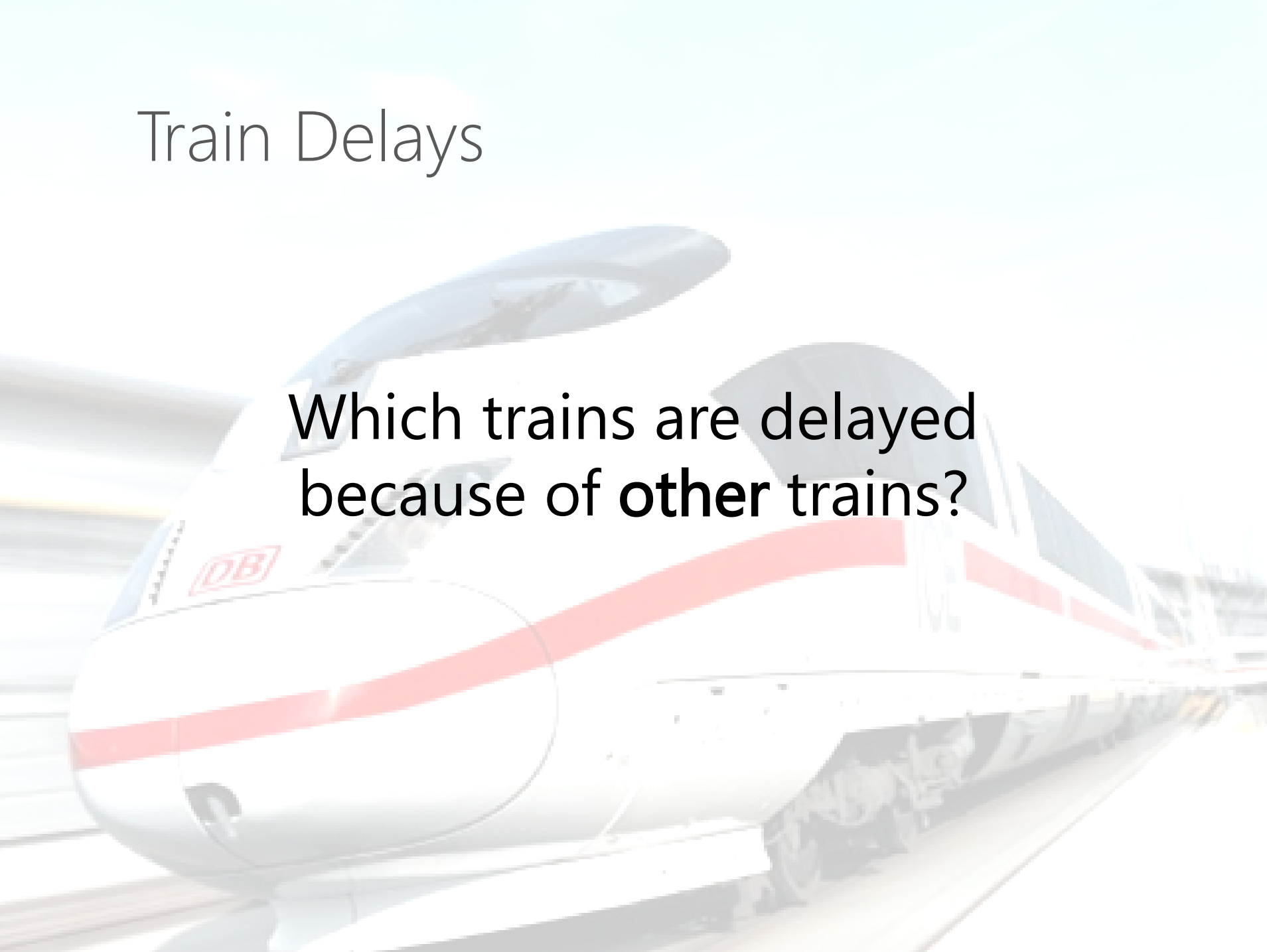
Data Mining Applications

Business Intelligence

Shopping Data

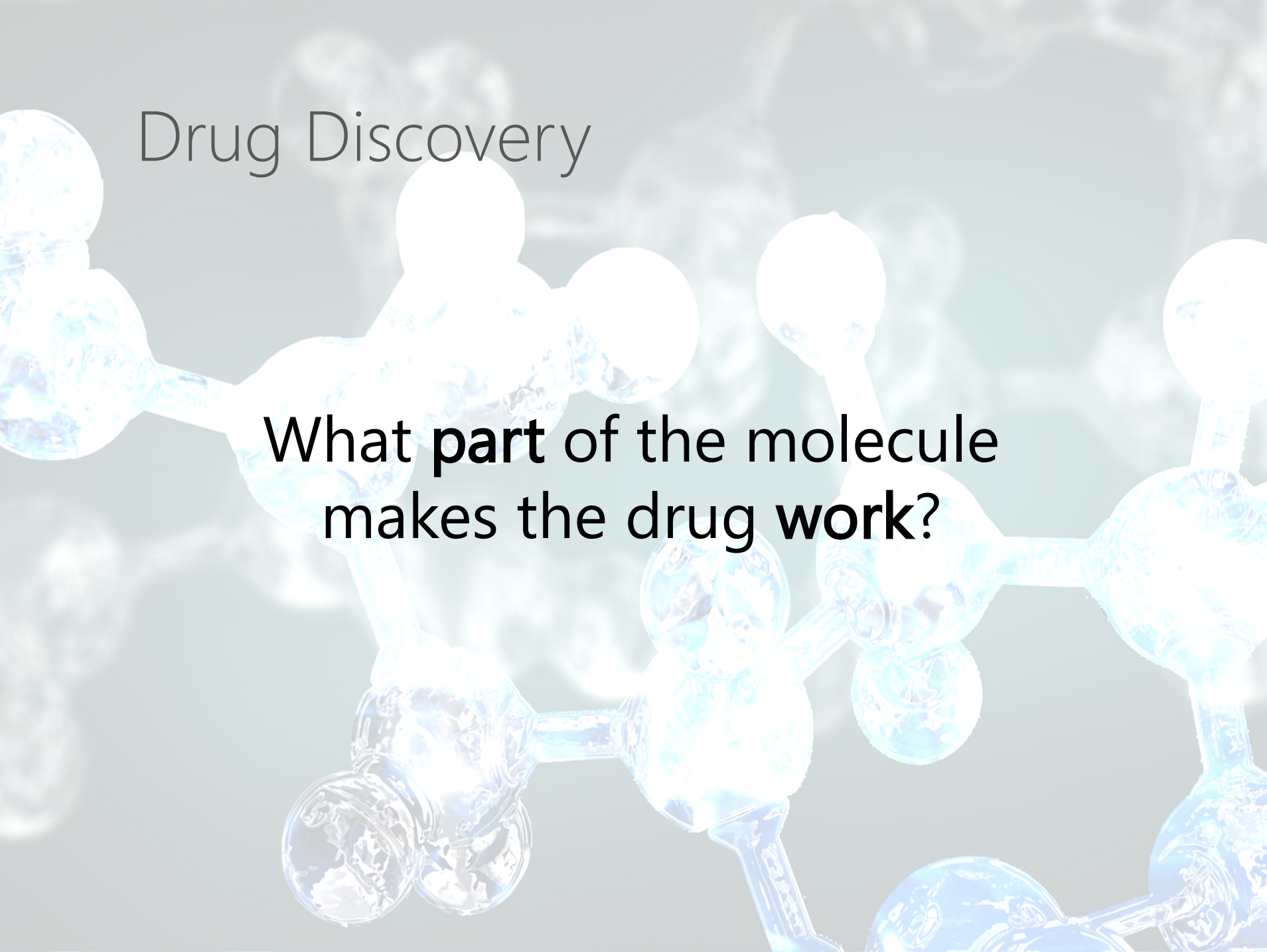
Which products are often bought **together**?

Train Delays



Which trains are delayed because of **other** trains?

Drug Discovery



What **part** of the molecule makes the drug **work**?

Data Mining Applica

Business Intelligence

- what do customers buy together?
- what are the seasonal trends?

Scientific Data Analysis

- what genes cause diseases?
- what are the differences between languages?

And anything else where you have data…

- who should Hillary Clinton try to persuade to vote?
- is everything alright with my space object?

Faster! Do it faster!

Nobody likes exponential time algorithms.

In data mining, we don't even like polynomial time

- Your solution is $O(n^3)$?
  Great… my data is only 10M records!

(Sub-)Linear runtime is what we strive for

- this means cutting corners: **good enough is good enough**
- often search space is so complicated there is no guarantee:
  hopefully good enough is good enough, hopefully

Sampling from Static Data

Can we trim Big Data to Reasonably Sized Data?

Without bias

- by sampling **uniformly,** every row has the same probability
- often without replacement: duplicate rows may be a nuisance

With bias to recent records

- by sampling with exponential decay, $p(\bar{X}) \propto e^{-\lambda \cdot \delta t}$
- where λ is the decay rate, and δt the age of element $\bar{X}$

With bias to certain (e.g. rare) classes

- by stratified sampling, often uniform with a probability per class

How much should we ask for?

How **much** data should we sample?

- depends on the **sample complexity** of your problem space

No Free Lunch Theorem: number of samples needed for error ϵ depends on the **actual** distribution p of the data, and there always exists some p with **arbitrarily high** sample complexity

Vapnik-Chervonenkis (VC) dimensionality and Rademacher complexity instead show how rich a set of hypotheses $\mathcal{H}$ is for your data. Promising, but often difficult to use in practice.

So, for many practical problems, we simply don't know, and just sample **as much as we can handle**

Streaming Data

Lots of data comes in over time, as a **data stream**

- e.g. sensor networks, telemetry data, CERN

Often, more data comes in than we can/want to store

- to analyse this data, we need specialised algorithms, that have a memory complexity $m \ll |S|$

Static databases are also streams

- streaming data is simply non-random access, e.g. we allow only one pass (or n) over your data

How can we sample from a stream?

- without bias?

Sampling from Streams

How can we get a **uniform sample** R **of** k **elements over a stream** S?

- that is, how do we make sure that after n elements of S, each of those have the **same probability** to be in R?

Reservoir Sampling, The Key Idea:

- initialise reservoir R with first k elements of S
- insert nth element into R with probability $\frac{k}{n}$
- if successful, remove one of the k old points uniformly at random

Now, every element of S has the probability $\frac{k}{n}$ to be in R (!)

(Aggarwal Ch. 2.4.1)

Conclusions

We're collecting more and more data

- most of it is boring — how to find out what part is interesting?

Scientific Method

- form hypothesis, collect data, test hypothesis

Data Mining

- collect data first, ask questions later;
  let the computer find what (interesting) hypotheses hold in it

Efficiency is very important

- a good answer now is much better than
  the perfect answer when we're all dead.

Thank you!

We're collecting more and more data

- most of it is boring — how to find out what part is interesting?

Scientific Method

- form hypothesis, collect data, test hypothesis

Data Mining

- collect data first, ask questions later;
  let the computer find what (interesting) hypotheses hold in it

Efficiency is very important

- a good answer now is much better than
  the perfect answer when we're all dead.