

*Towards automation of
chromatographic data treatment:
a Bayesian approach*

*Gabriel Vivó Truyols, Michael Woldegebriel
Van 't Hoff institute for molecular sciences
University of Amsterdam*

Automation is our challenge now...

Why Bayesian?

Evolution of the instrumentation



1 bit



GC: 1 KB



HPLC: 1KB



GC-MS; ~100 KB



HPLC-DAD; ~100 KB



HPLC-MS;
1 MB



GCxGC: ~100 KB



GCxGC-MS;
LCxLC-MS,
1 GB/hour



GCxGC-HRMS;
LCxLC-HRMS,
15 GB/hour

Automation using a frequentist approach

Why Bayesian?

Data



Algorithm

+

$$\sum \theta \sin \epsilon$$

=

Information

There is a chromatographic peak at $t_R=12$ min.

These 10 peaks in these 10 chromatograms belong to the same compound.

... etc.

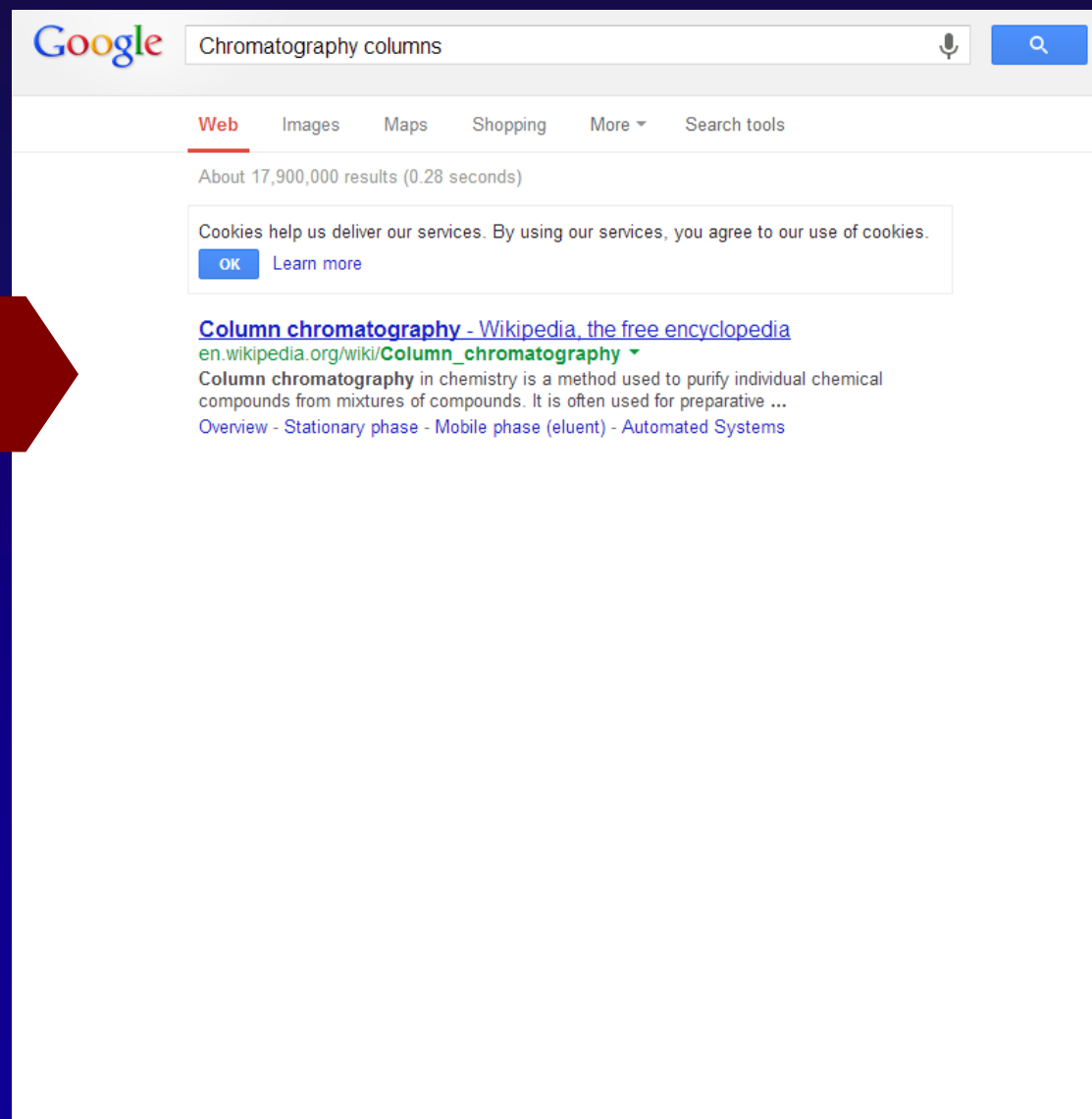
In a sense, the machines are “taking responsibility” on the decision...

... and only the final result is shown.

Automation using a frequentist approach

Why Bayesian?

Final result



Google Chromatography columns

Web Images Maps Shopping More Search tools

About 17,900,000 results (0.28 seconds)

Cookies help us deliver our services. By using our services, you agree to our use of cookies. [OK](#) [Learn more](#)

[Column chromatography - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/[Column_chromatography](#)

Column chromatography in chemistry is a method used to purify individual chemical compounds from mixtures of compounds. It is often used for preparative ...

[Overview](#) - [Stationary phase](#) - [Mobile phase \(eluent\)](#) - [Automated Systems](#)

Automation using a "Bayesian" approach

Why Bayesian?

Most likely result

Less likely result (but still possible)

...

Google Chromatography columns

Web Images Maps Shopping More Search tools

About 17,900,000 results (0.28 seconds)

Cookies help us deliver our services. By using our services, you agree to our use of cookies. [Learn more](#)

[Column chromatography - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Column_chromatography
Column chromatography in chemistry is a method used to purify individual chemical compounds from mixtures of compounds. It is often used for preparative ...
Overview - Stationary phase - Mobile phase (eluent) - Automated Systems

[Images for chromatography columns - Report images](#)

[Glass Chromatography Columns - Glassware | Sigma-Aldrich](#)
www.sigmaaldrich.com > Product Directory > Labware > Glassware
Flash-Chromatography columns are available with ball joints or threaded standard taper joints and EZSafe™ connections and are useful for rapid, preparative ...

[Empty Columns & Collection Plates for Chromatography](#)
www.gelifsciences.com/site/.../en/.../AlternativeProductStructure_17146...
From lab-scale to pilot-scale and production, we have an extensive portfolio of chromatography columns.

[Chromatography Columns and Supplies Thermo Scientific](#)
www.thermoscientific.com/columns
Chromatography consumables including premier HPLC, GC and SPE phases.

[Chromatography Columns - Low & Medium Pressure - HPLC:Life ...](#)
www.bio-rad.com > ... > Products > Chromatography
Discover Bio-Rad's array of empty and prepacked chromatography columns - gravity and spin columns, low- to medium-pressure, and HPLC columns.

Automation using a Bayesian approach

Why Bayesian?

Data



87% chance: There is a peak centred at $t_R=12$ min.

13% chance: There is no peak centred at $t_R=12$ min.

20% chance: Peaks 1-9 belong to the same compound. Peak 10 is different.

15% chance: Peaks 1-5,7-9 belong to the same compound. Peaks 6 and 10 are different.

... etc.

Information

There is a chromatographic peak at $t_R=12$ min.
These 10 peaks in these 10 chromatograms belong to the same compound.

a collection of all possibilities (ranked by their probability)

In a sense, the machines are ^{not} taking responsibility" on the decision...

... and ~~only~~ the final result is shown.

It is up to the chromatographer to take the final decision

Bayesian statistics in court

Why Bayesian?

Posterior odds

$$\frac{p(H_0 | D)}{p(H_1 | D)}$$

=

Likelihood ratio

$$\frac{p(D | H_0)}{p(D | H_1)}$$

x

Prior odds

$$\frac{p(H_0)}{p(H_1)}$$

The suspect is innocent

The suspect is guilty



The scientist does not decide upon the validity of H_0 or H_1 , only calculates the likelihood ratio (the value of the evidence)

Bayesian statistics in data automation

Why Bayesian?

Posterior odds

$$\frac{p(H_0|D)}{p(H_1|D)}$$

=

Likelihood ratio

$$\frac{p(D|H_0)}{p(D|H_1)}$$

x

Prior odds

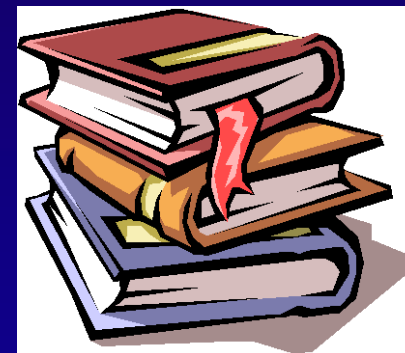
$$\frac{p(H_0)}{p(H_1)}$$



I take the decision!!



The data is only used to “update” our prior probability on a situation (**but the decision is not “taken” by the algorithm**)

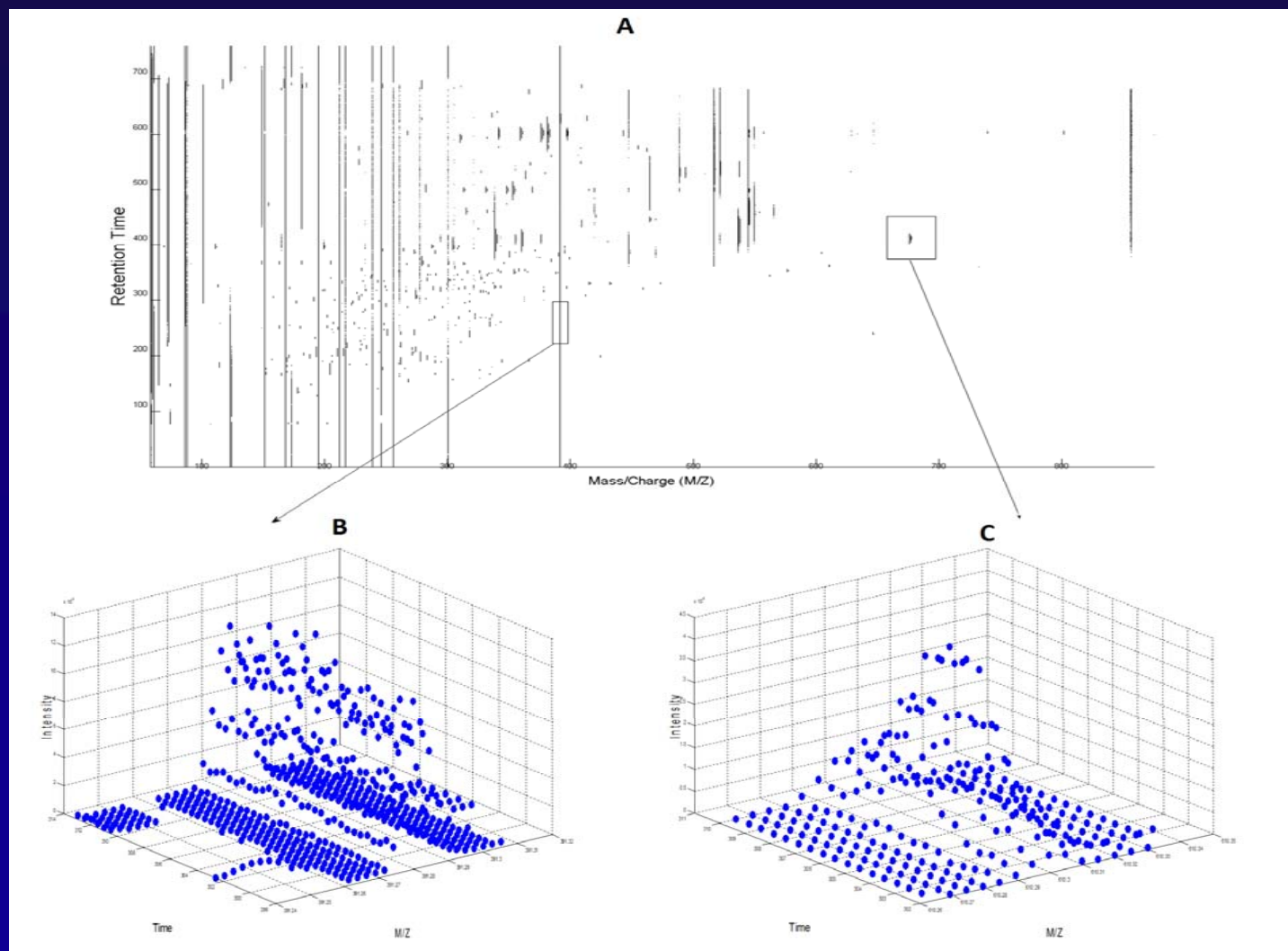


Prior experiments
Bibliographic information

Some practical applications in chromatography

Example I: peak detection in LC-MS

Example I



Example I: peak detection in LC-MS

Example I

What is the probability
that a certain point is affected
by a chromatographic peak?

The point IS affected by a
chromatographic peak

H_1

The point IS NOT affected
by a chromatographic peak

H_2

Posterior odds

$$\frac{p(H_1|D)}{p(H_2|D)}$$

=

Likelihood ratio

$$\frac{p(D|H_1)}{p(D|H_2)}$$

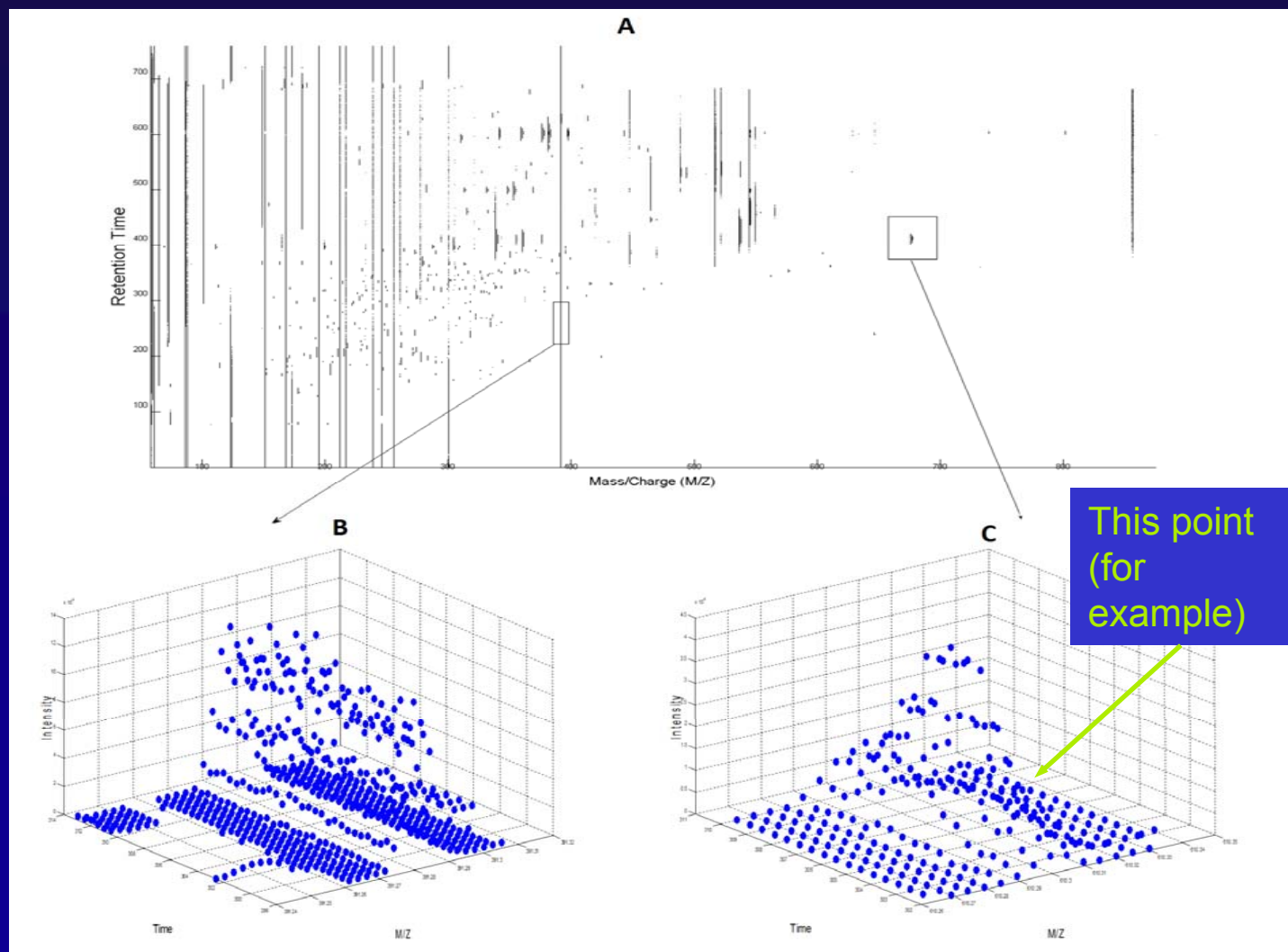
x

Prior odds

$$\frac{p(H_1)}{p(H_2)}$$

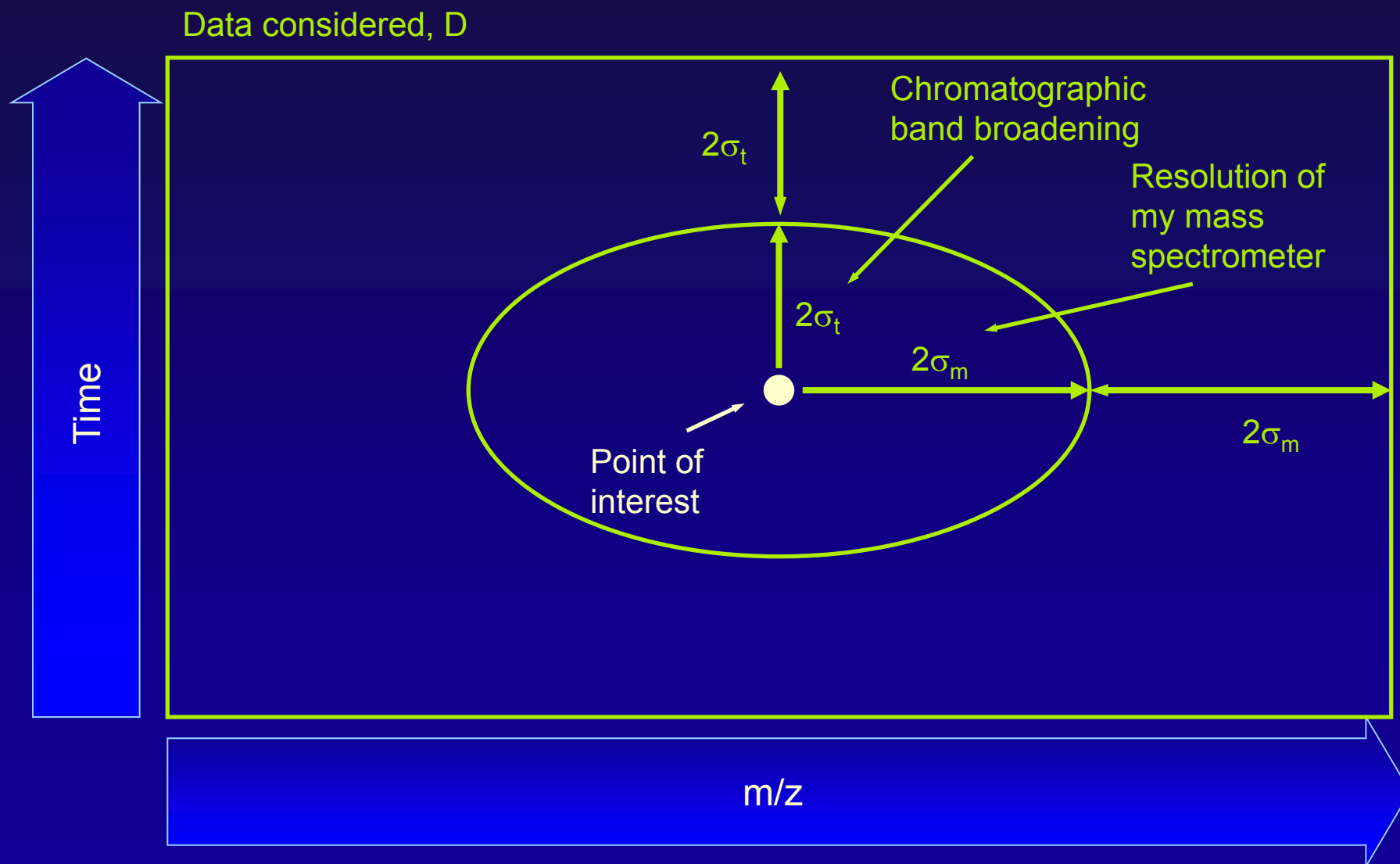
Example I: peak detection in LC-MS

Example I



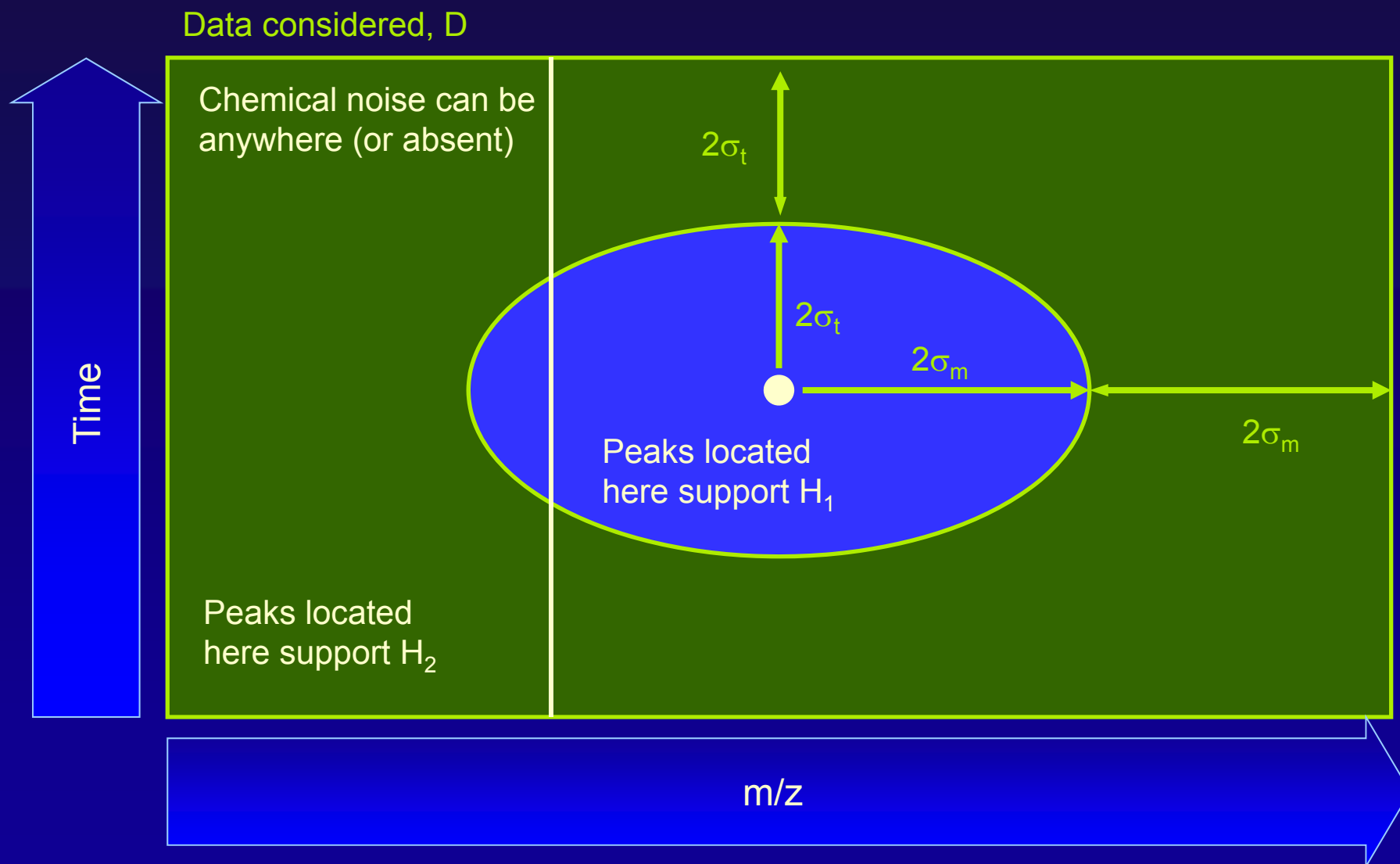
Example I: peak detection in LC-MS

Example I



Example I: peak detection in LC-MS

Example I



Example I: peak detection in LC-MS

Example I

Likelihood ratio

$$\frac{p(D|H_1)}{p(D|H_2)}$$

The point is affected by a peak

The point is not affected

Number of models that support H_1

$$p(D|H_1) = \sum_{w \in H_1}^{n1} P(D, w|H_1) = \sum_{w \in H_1}^{n1} P(D|w, H_1) P(w|H_1)$$

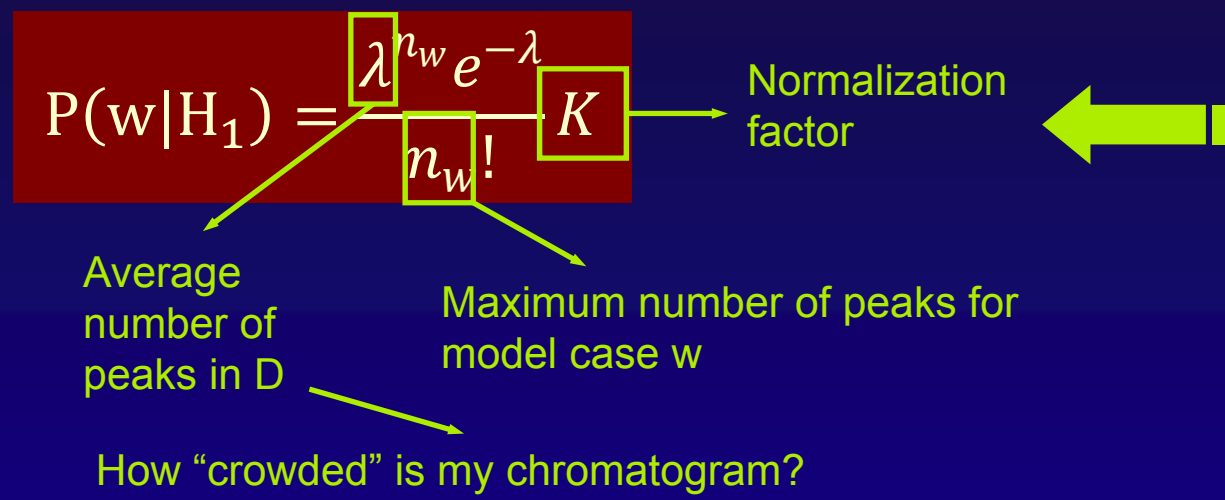
Model case

Number of models that support H_2

$$p(D|H_2) = \sum_{w \in H_2}^{n2} P(D, w|H_2) = \sum_{w \in H_2}^{n2} P(D|w, H_2) P(w|H_2)$$

Example I: peak detection in LC-MS

Example I



J.M. Davis and J.C. Giddings, Statistical overlap theory, Anal. Chem., 55 (1983).

Total number of peaks / peak capacity

$$P(H_2) = e^{(-M/n_c)} K'$$

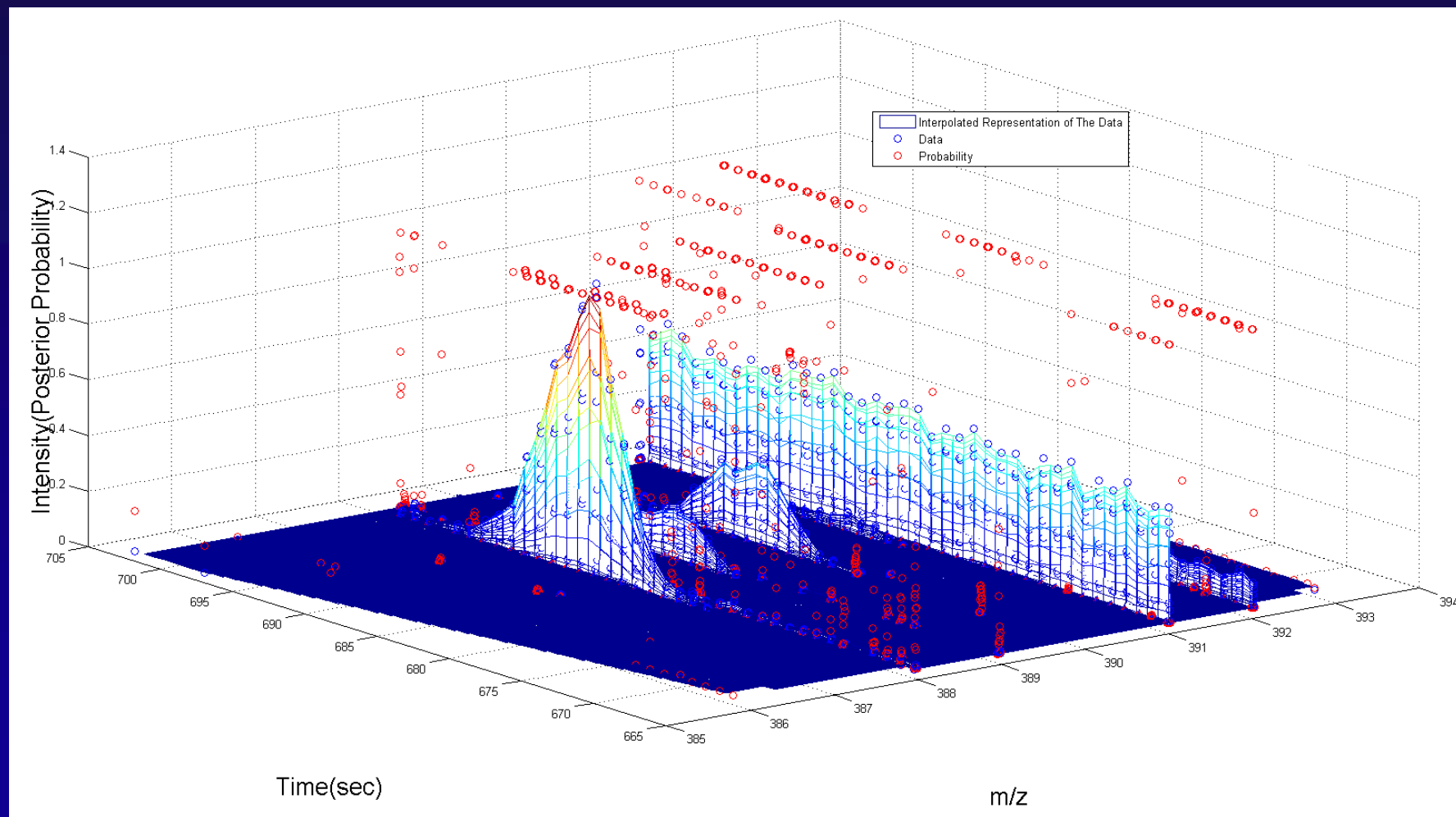
Normalization factor

Priors

$$P(H_1) = 1 - P(H_2)$$

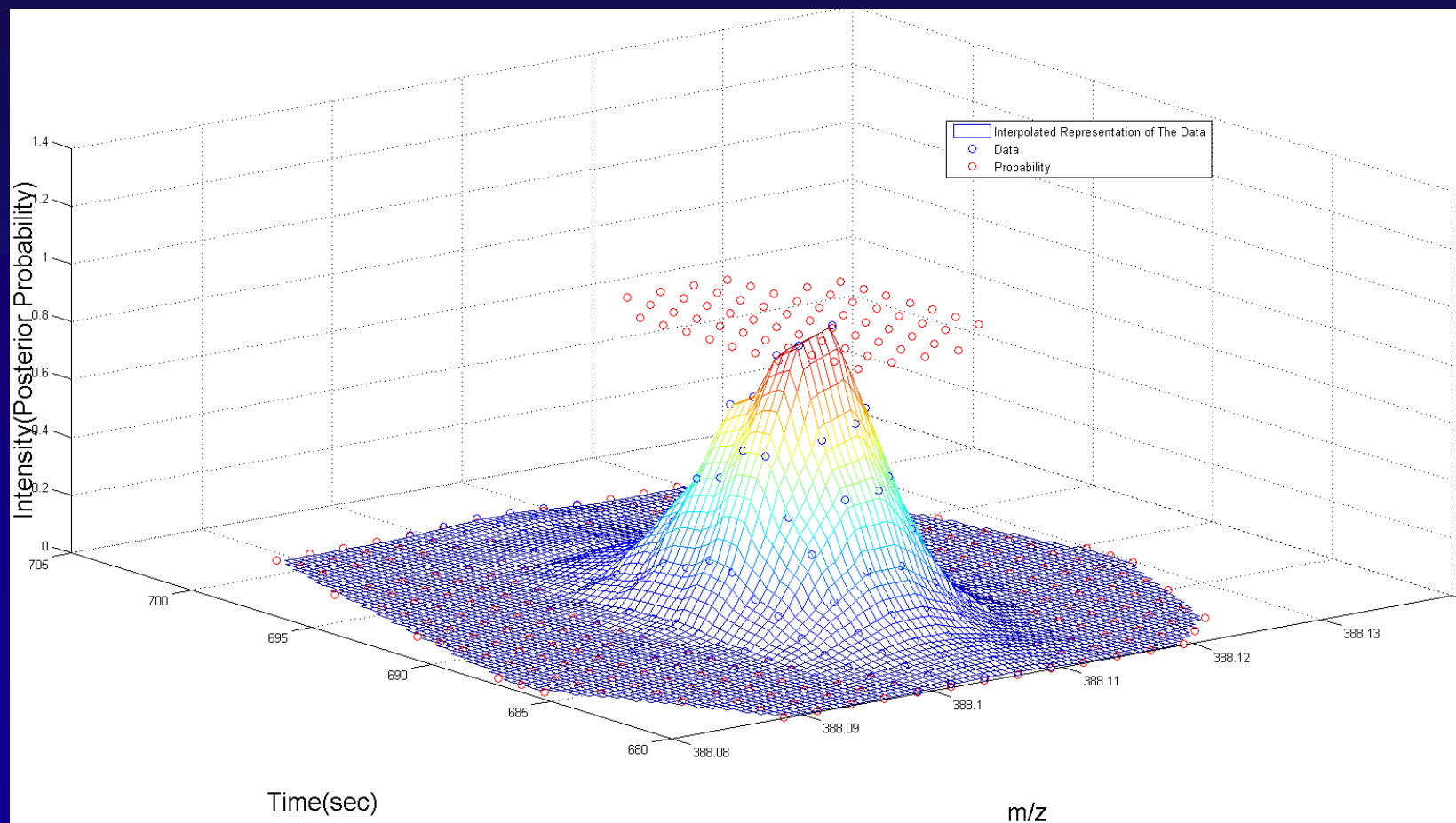
Example I: peak detection in LC-MS. Results

Example I



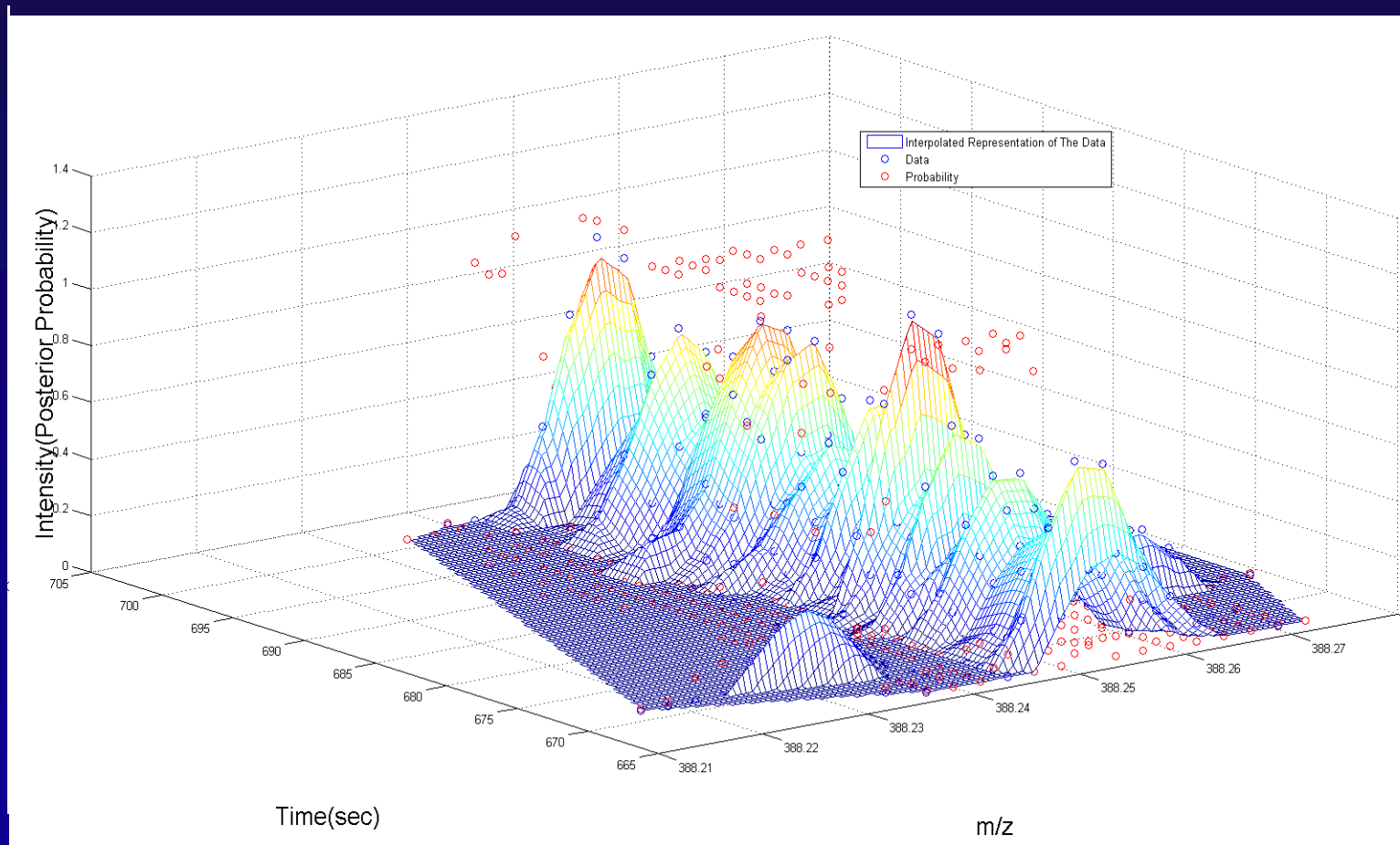
Example I: peak detection in LC-MS. Results

Example I



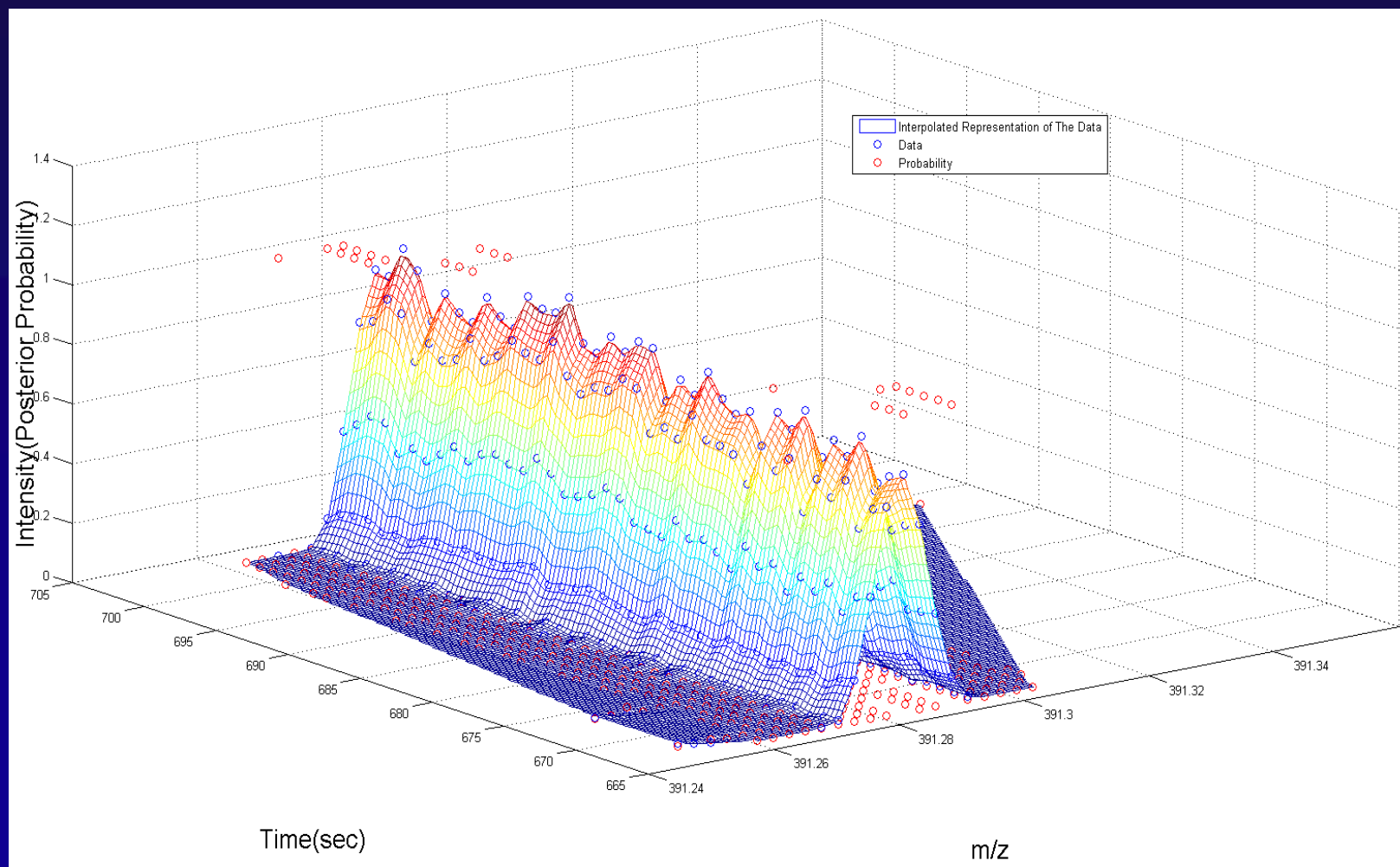
Example I: peak detection in LC-MS. Results

Example I



Example I: peak detection in LC-MS. Results

Example I

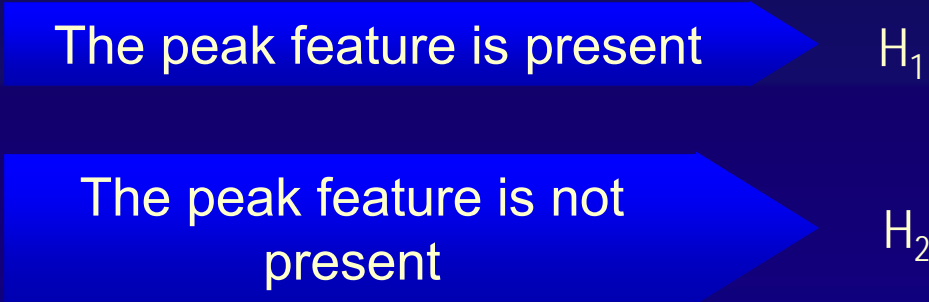


Example II: Targeted screening

Example II

In forensic toxicology, we want to know the probability of a compound (out of a list of ~500) being present (so we should submit the sample for confirmation)

What is the probability that the peak feature of a certain compound is present?



Posterior odds

$$\frac{p(H_1|D)}{p(H_2|D)}$$

=

Likelihood ratio

$$\frac{p(D|H_1)}{p(D|H_2)}$$

x

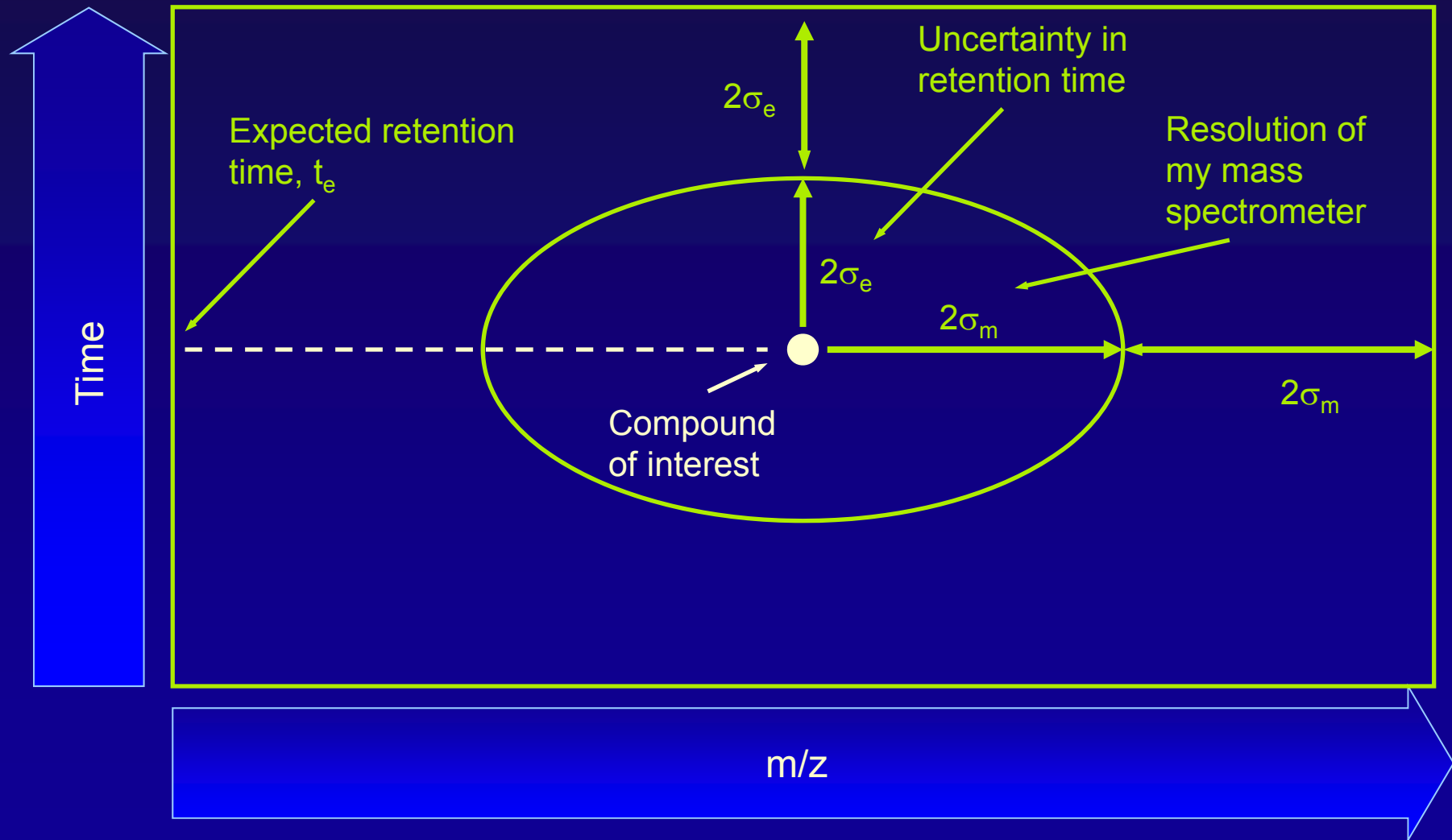
Prior odds

$$\frac{p(H_1)}{p(H_2)}$$

Example II: Targeted screening

Example II

Data considered, D



Example II: Targeted screening

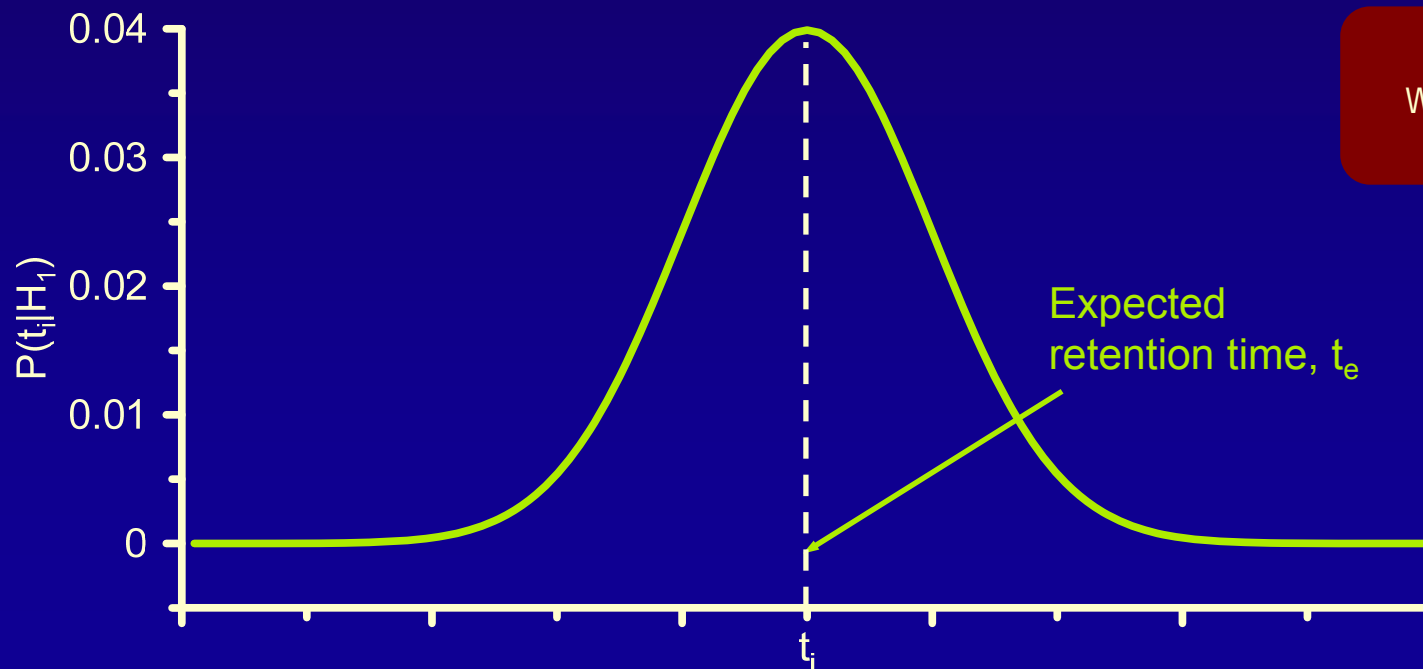
Example II

$$p(D|H_1) = \sum_{i=1}^k \sum_{w \in H_1} P(D|w, t_i, H_1) P(w|H_1) P(t_i|H_1)$$

Model case
(excluding
change in t_R)

My peak can elute at
various t_i values

What is the prior probability
distribution for the elution time?



I am not sure
where my peak
elutes...



Example II: Targeted screening

Example II

Evidence D_1 ... data at the most abundant isotope

$$\frac{p(H_1|D_1)}{p(H_2|D_1)} = \frac{p(D_1|H_1)}{p(D_1|H_2)} \times \frac{p(H_1)}{p(H_2)}$$

Evidence D_2 ... isotope ratios

$$\frac{p(H_1|D_1, D_2)}{p(H_2|D_1, D_2)} = \frac{p(D_2|H_1, D_1)}{p(D_2|H_2, D_1)} \times \frac{p(H_1|D_1)}{p(H_2|D_1)}$$

Example II: Targeted screening

Example II

Testing the method with 500 compounds

Bayesian vs. frequentist are not comparable!

We set up a threshold on the posterior probability to “declare” a match...

... and then we can compare with existing software

Bayesian screening

Sensitivity: 96.5%
Specificity 95.8%

Mass Hunter

Sensitivity: 93.5%
Specificity 95.5%

Example II: Targeted screening

Example II

In forensic toxicology, we want to know the probability of a compound (out of a list of ~500) being present (so we should submit the sample for confirmation)

What is the probability that the peak feature of a certain compound is present?

compound

The peak feature is present

H₁

compound

The peak feature is not present

H₂

Posterior odds

$$\frac{p(H_1|D)}{p(H_2|D)}$$

=

Likelihood ratio

$$\frac{p(D|H_1)}{p(D|H_2)}$$

x

Prior odds

$$\frac{p(H_1)}{p(H_2)}$$

Conclusions

- Automation: the data-analysis part doesn't involve a decision → It is just informing the scientist about the probabilities of the different hypothesis being true...
- Weight the data, do not discard anything!
- The good (old) theories work!: e.g. Statistical Overlap Theory (Davies, Giddings)
- Toxicology screening (including MS). Combining evidence: e.g. different number of isotopes, as well as uncertainty in the retention time.
- Bayesian screening gives probabilities (not decisions). When we convert it into a decision... we are beating the "Mass Hunter" software (Agilent).

Special thanks to...

Acknowledgements

- Analytical-chemistry group (University of Amsterdam).
- M. Sjerps (dept. of mathematics, univesrity of Amsterdam).
- COAST consortium (NWO + DSM + NFI + Rikilt) for funding.

Both the tutorial (Monday) and this presentation
available at www.tecnometrix.com

Feel free to download and provide me feedback!

Thanks for your attention!