The necessity of human insight for solving some problems supposedly solved by AI

Christian Hennig
Dipartimento di Scienze Statistiche "Paolo Fortunati"
christian.hennig@unibo.it

1. Some problems

Some problems that some claim AI can solve,
or even has solved already.

These highlight limitations of AI, and how
human decision is needed to see and (occasionally) solve them.

1. Some problems

Some problems that some claim AI can solve,
or even has solved already.

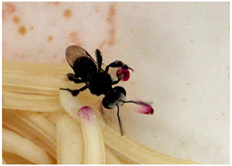
These highlight limitations of AI, and how
human decision is needed to see and (occasionally) solve them.

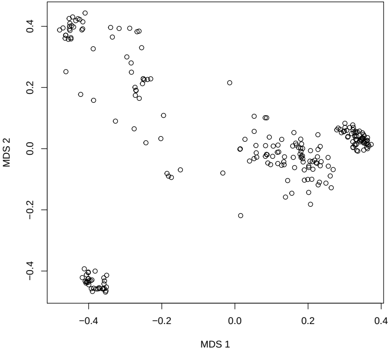
1.1 Unsupervised classification (cluster analysis)

Supervised classification: assigning observations
to already known groups (e.g., recognising animals on images).

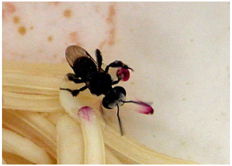
Unsupervised classification: How are our observations grouped?

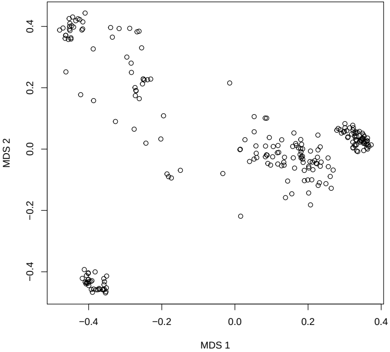
Genetic information on tetragonula bees - what are the species?



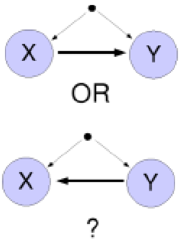


Unsupervised classification is “creative” -
far more difficult than supervised.





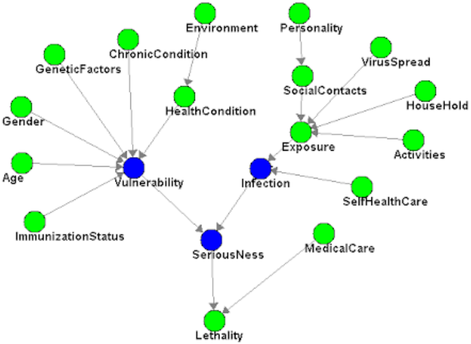
1.2 What is the direction of causality?



Criminality is often higher in poor city districts.
Does poverty cause criminality,
or does criminality produce poverty and deter wealthy people?

What do the data say?

More generally discover causality relations in systems from data.



1.3 Why do we know so little about Covid-19?

There are lots of data regarding amount of infections,
courses of disease, deaths.

There are hardly any reliable predictions
(some modelling in beginning was far off)
there is little knowledge about
how rules such as wearing masks outside affect it.

Shouldn't AI based on "big data" do better?

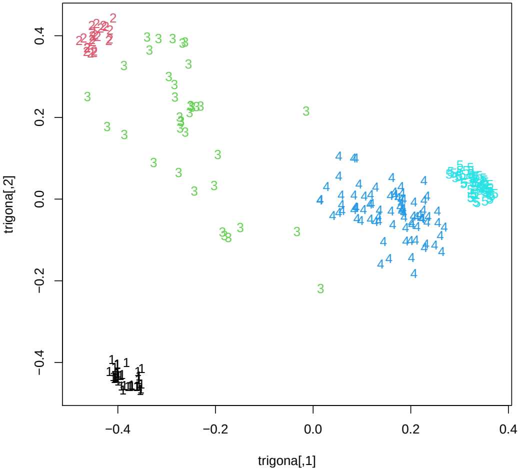
Key issues with AI learning from data:

- Problem definition and identifiability
- Data quality
- Issues with independence and identity

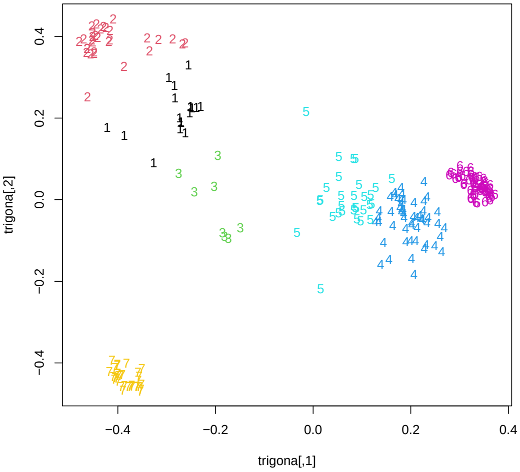
2. Problem definition and identifiability

What information is actually in the data,
what problems can be solved?

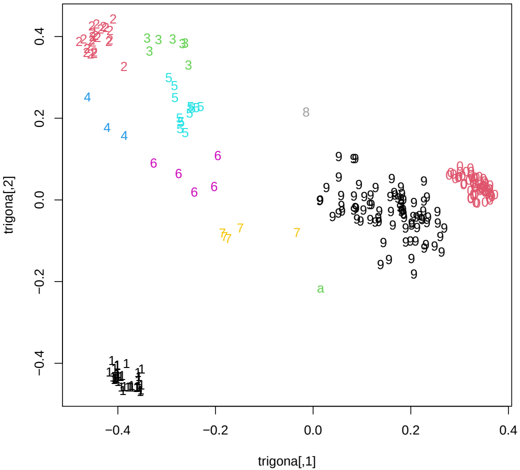
Four different clustering solutions for tetragonula bees
by four different algorithms:



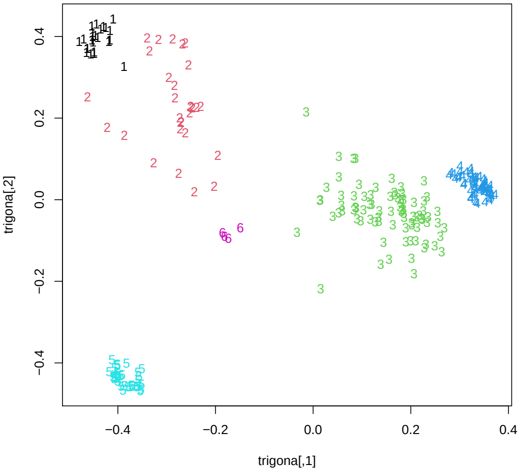
May not want to allow too much heterogeneity within clusters.



Whatever subset is separated could form a cluster.



Allow more heterogeneous clusters if they are well separated.



Which of these is best?

Which of these is best?

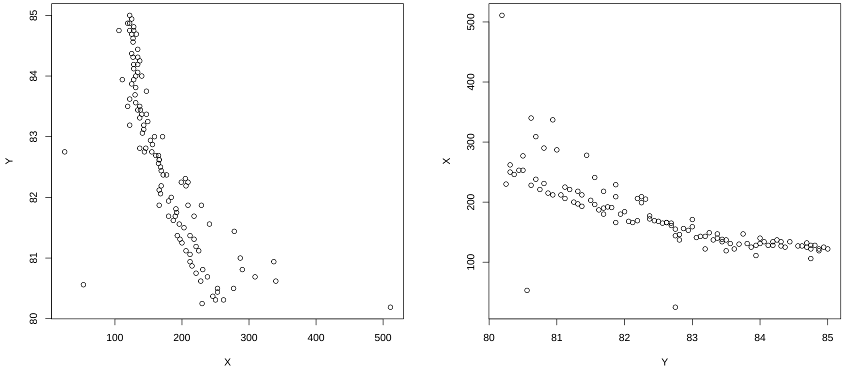
Different "cluster concepts" $\Rightarrow$ different solutions.

Is separation or homogeneity more important?
Can within-cluster variation differ strongly between clusters?
Can very small clusters be tolerated?

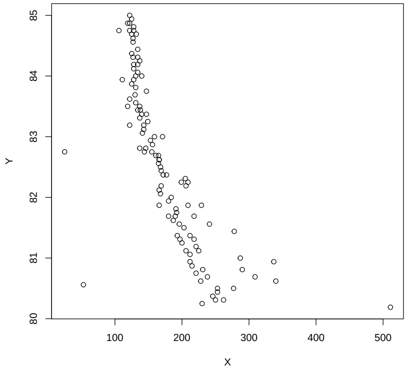
The data do not hold information on which one is "correct";
the researcher needs to decide what features clusters need to have.

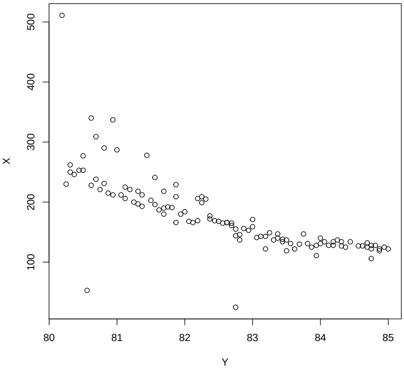
This is ignored in much AI and statistical literature.

Direction of causality from data?



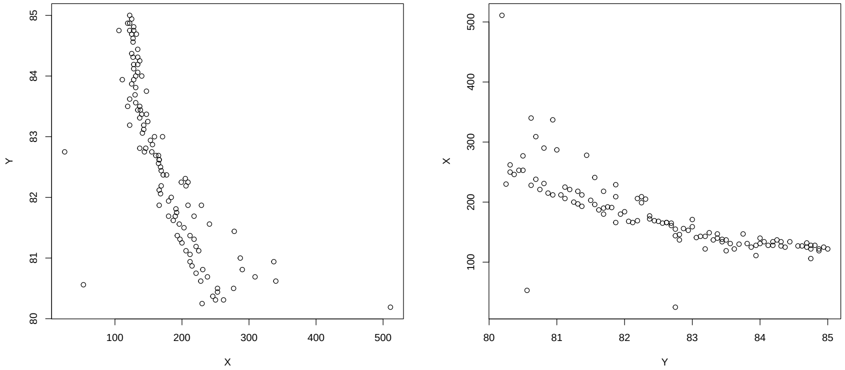
Direction of causality from data?





Not identifiable! Causal models can be set up both ways.

Direction of causality from data?

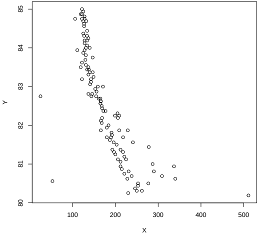


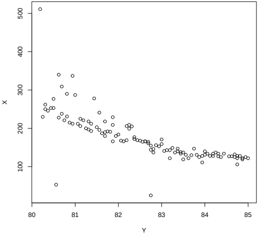
Not identifiable! Causal models can be set up both ways.
But if we had to guess. . . ?

Guyon et al. (2010, 2019) set up causality challenges
for the machine learning community.

Reframing of problem:
They made it a supervised classification problem:
From database of datasets with known (or semi-simulated) ground truth
predict $X \to Y$ or $Y \to X$.

Classify causality based on looking "similar"
to data with known causality direction.





Y water temperature, X rotation time for Stirling engine.

Can data on Stirling engine teach us something about causality between poverty and criminality?

3. Data quality

Reasoning from data relies on quality data.
Algorithms have a hard time detecting issues.
Background knowledge required.

- Measurement issues
- Relevance of data
- Representativity of data

3.1 Measurement issues

Covid-19 infection counts:
very different test and reporting practices
in different countries and even regions;
hard to compare.

3.1 Measurement issues

Covid-19 infection counts:
very different test and reporting practices
in different countries and even regions;
hard to compare.

Similar issues for criminality and poverty measurement.

3.2 Relevance of data

Filtration effectivity of masks
$\Rightarrow$ effectivity of mask prescription policies?

3.2 Relevance of data

Filtration effectivity of masks
$\Rightarrow$ effectivity of mask prescription policies?

Causality is about consequences of changing X or Y;
may need experimental data with interventions changing them.

3.3 Representativity of data

Percentage of positive tests representative
for percentage of Covid infections in population?
Not really; persons with symptoms and contact tested first.

Causality challenge collection of data with known “ground truth” is quite lopsided.

Most known causalities are physical or with clear time order.

Nothing like “poverty and criminality” in training collection; not really informative for that kind of problem.

Causality challenge collection of data with known "ground truth" is quite lopsided.

Most known causalities are physical or with clear time order.

Nothing like "poverty and criminality" in training collection; not really informative for that kind of problem.

Situations with known causality may be systematically different from situations with unknown causality of interest.

Great results in challenge

Causality challenge collection of data with known “ground truth” is quite lopsided.

Most known causalities are physical or with clear time order.

Nothing like “poverty and criminality” in training collection; not really informative for that kind of problem.

Situations with known causality may be systematically different from situations with unknown causality of interest.

Great results in challenge
. . . but only for situation with known causality
(otherwise they couldn’t even evaluate success).

4. Issues with independence and identity

Most statistical and ML methods assume data to be identically and independently distributed ("*iid*").

There are models for time and spatial dependence, non-identity based on explanatory variables (regression).

4. Issues with independence and identity

Most statistical and ML methods assume data to be identically and independently distributed ("*iid*").

There are models for time and spatial dependence, non-identity based on explanatory variables (regression).

Most of these assume *iid* "residuals" or "innovations".

Some aspects of data need be *iid*, in order to allow learning from training data about future.

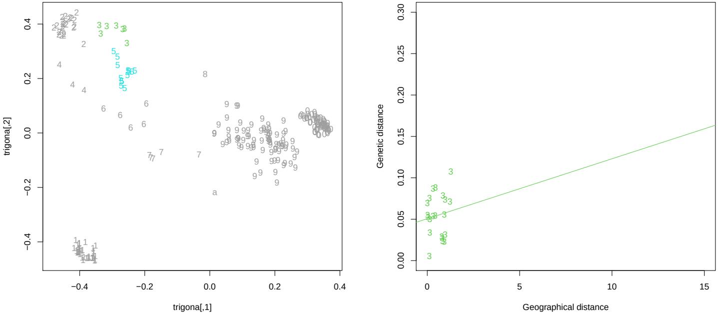
4.1 Geographical dependence of genetic setup

Genetic setup of bees, even of the same species,
can be quite different if they live far away from each other;
more similar if closer (because of dependence).

4.1 Geographical dependence of genetic setup

Genetic setup of bees, even of the same species,
can be quite different if they live far away from each other;
more similar if closer (because of dependence).

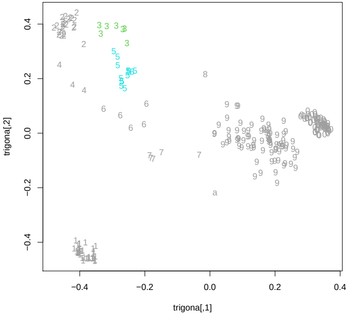
Model relating genetic distance to geographical distance (need new data):

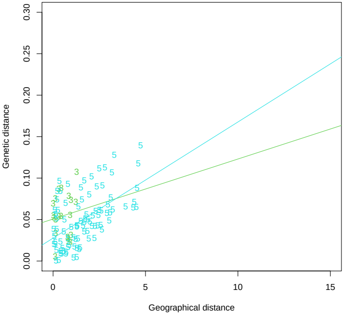


4.1 Geographical dependence of genetic setup

Genetic setup of bees, even of the same species,
can be quite different if they live far away from each other;
more similar if closer (because of dependence).

Model relating genetic distance to geographical distance (need new data):

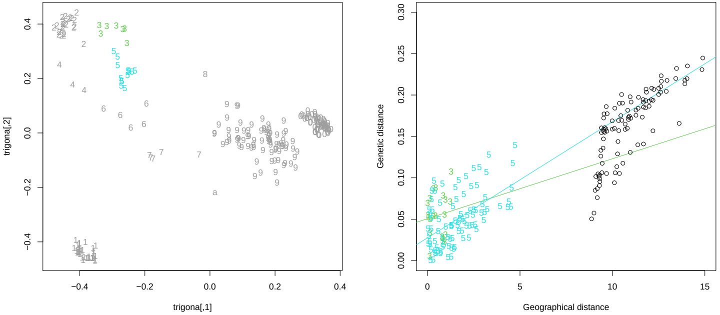




4.1 Geographical dependence of genetic setup

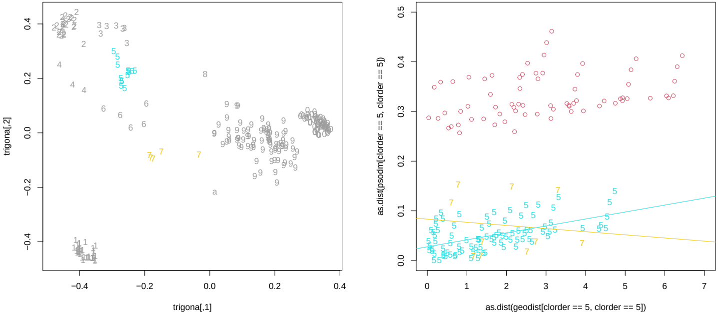
Genetic setup of bees, even of the same species,
can be quite different if they live far away from each other;
more similar if closer (because of dependence).

Model relating genetic distance to geographical distance (need new data):



4.1 Geographical dependence of genetic setup

Genetic setup of bees, even of the same species,
can be quite different if they live far away from each other;
more similar if closer (because of dependence).



For testing, still need iid bees *given geographical distance*.

4.2 Dependence and irregularity of infections

Covid-19: Infections are highly dependent;
hot spots can be without cases, then have many cases in few days.

Politicians react on data; irregular changes:
lockdowns, restrictions, testing regime changes, new treatments,...
iid at *any* level is very dubious;
prediction hardly possible at any acceptable precision.

Conclusion

Issues listed here require background knowledge
that is not represented in data.

Easily forgotten or ignored in Machine Learning/AI community.

AI won't make us humans superfluous any time soon.

Conclusion

Issues listed here require background knowledge
that is not represented in data.

Easily forgotten or ignored in Machine Learning/AI community.

AI won't make us humans superfluous any time soon.

Message: Use models and algorithms
with awareness of how they get things wrong
- best way to improve!

C. Anderson (Wired, 2008): *"The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"*

Can AI make the decisions on its own given just enough data?

C. Anderson (Wired, 2008): *"The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"*
Can AI make the decisions on its own given just enough data?

Without "ground truth data" (unsupervised learning, causality directions, future of irregular processes)
AI cannot even know whether it gets better with more data.

C. Anderson (Wired, 2008): *"The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"*
Can AI make the decisions on its own given just enough data?

Without "ground truth data" (unsupervised learning, causality directions, future of irregular processes)
AI cannot even know whether it gets better with more data.

Tempting to think: What cannot be represented as data is irrelevant; good prediction results on the database imply it works well.
This can be very wrong.

References

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired https://www.wired.com/2008/06/pb-theory/

Gelman, A. and Hennig, C. (2017) Beyond subjective and objective in statistics (with discussion). Journal of the Royal Statistical Society Series A 180: 967-1033.

Guyon, I., Janzing, D., Schölkopf, B. (2010) Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008, PMLR 6: 1-42.

Guyon, I., Statnikov, A., Batu, B. B. (2019) Cause Effect Pairs in Machine Learning. Springer, NY.

Hausdorf, B. and Hennig, C. (2010) Species Delimitation Using Dominant and Codominant Multilocus Markers. Systematic Biology 59: 491-503.

Hausdorf, B. and Hennig, C. (2020) Species and geography. Molecular Ecology Resources 20: 950-960.

Hennig, C. (2015) What are the true clusters? Pattern Recognition Letters 64: 53-62