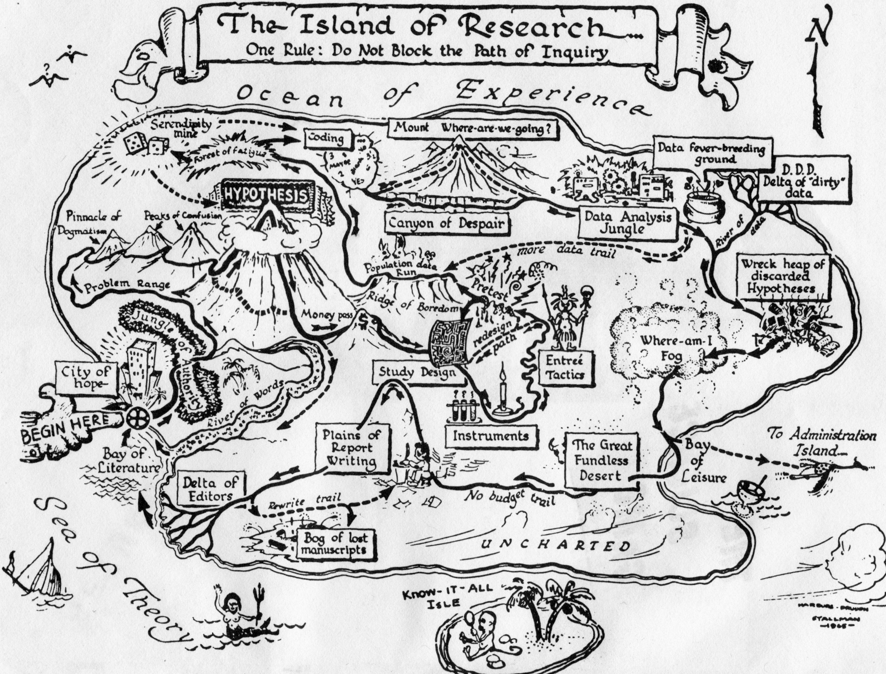
# Machine Learning for Complex Data Analysis

Michael Gutmann

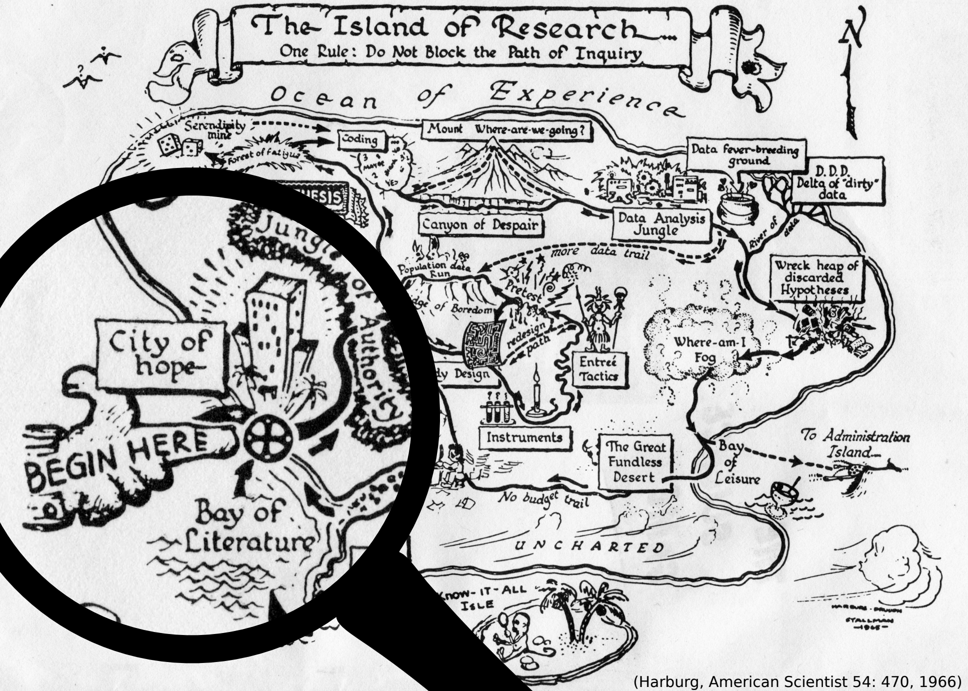
michael.gutmann@ed.ac.uk

School of Informatics, University of Edinburgh

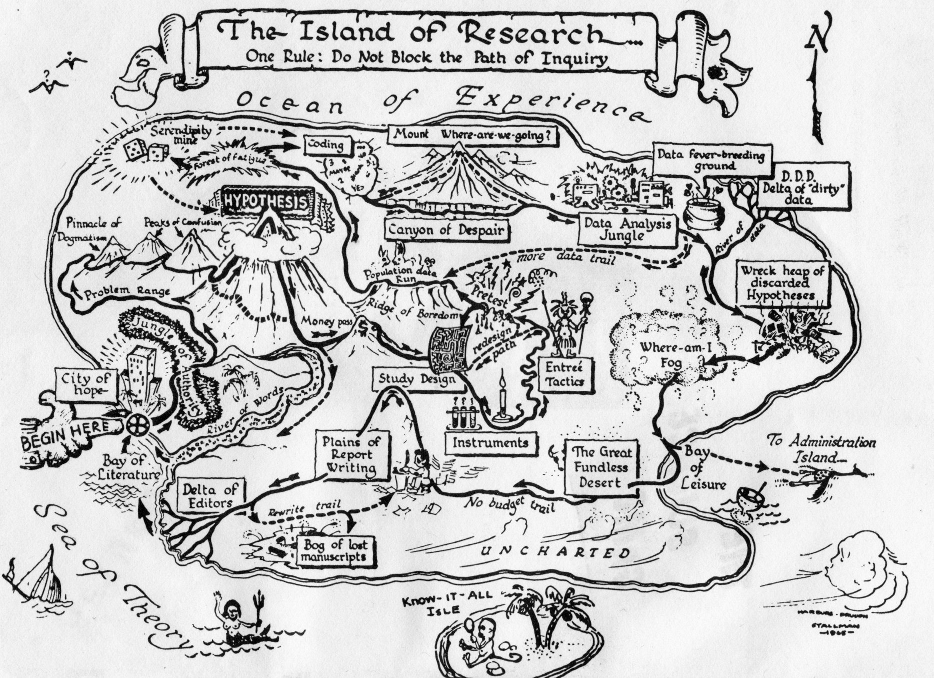
13 December 2017



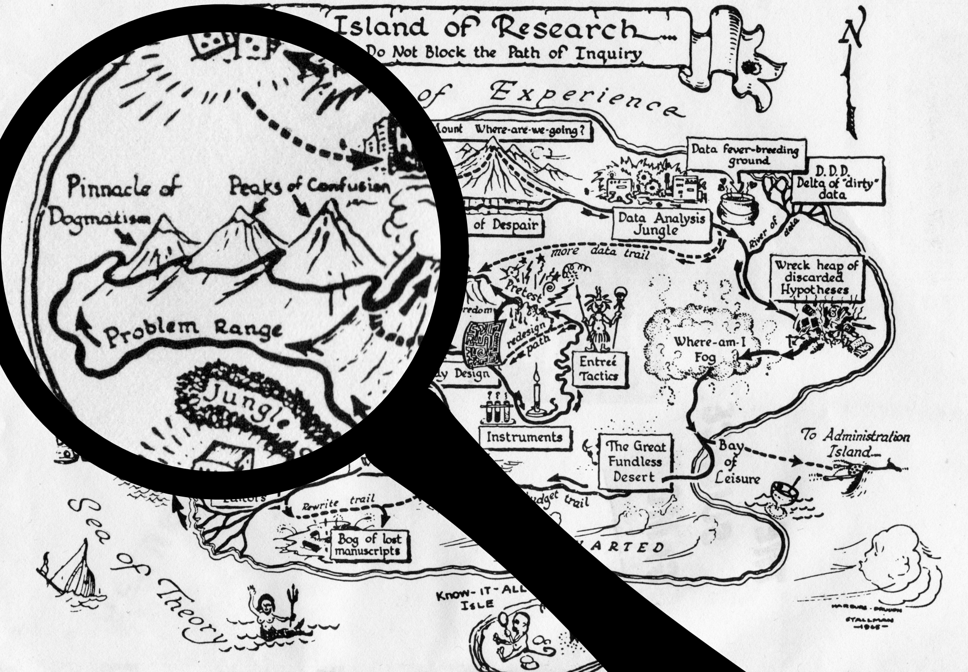

(Harburg, American Scientist 54: 470, 1966)



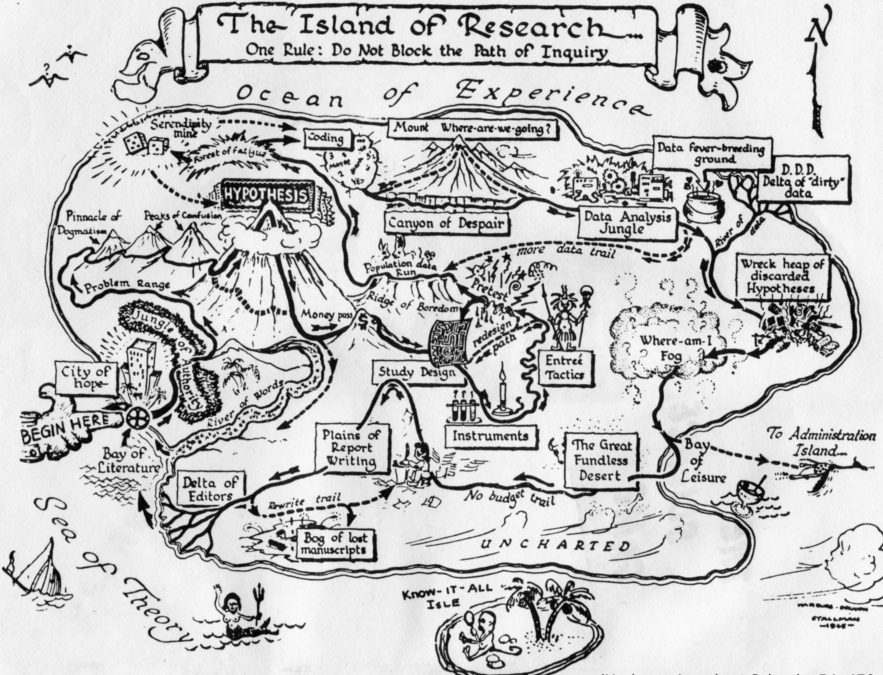

(Harburg, American Scientist 54: 470, 1966)




(Harburg, American Scientist 54: 470, 1966)

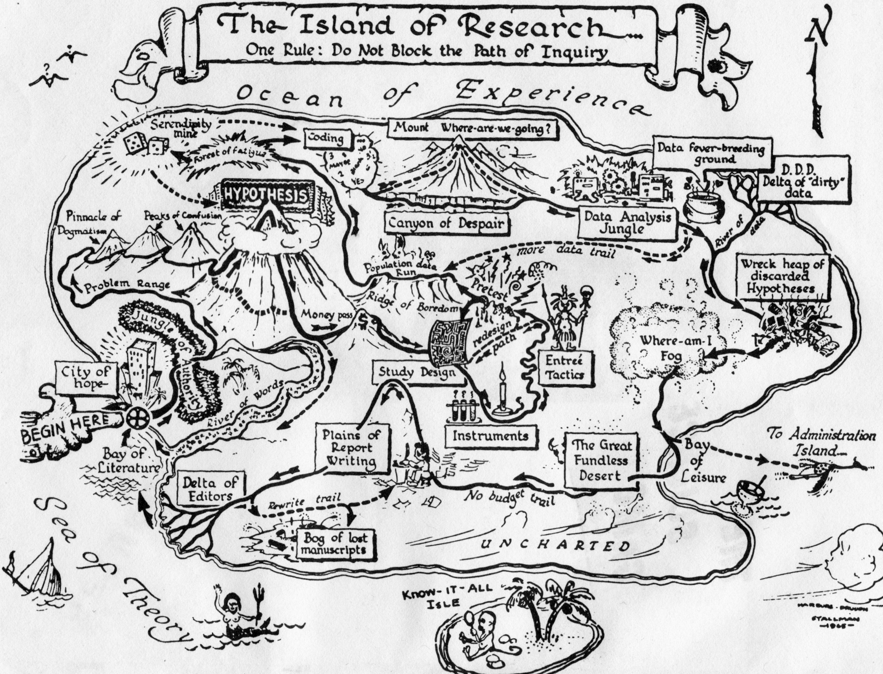



can Scientist 54: 470, 1966)

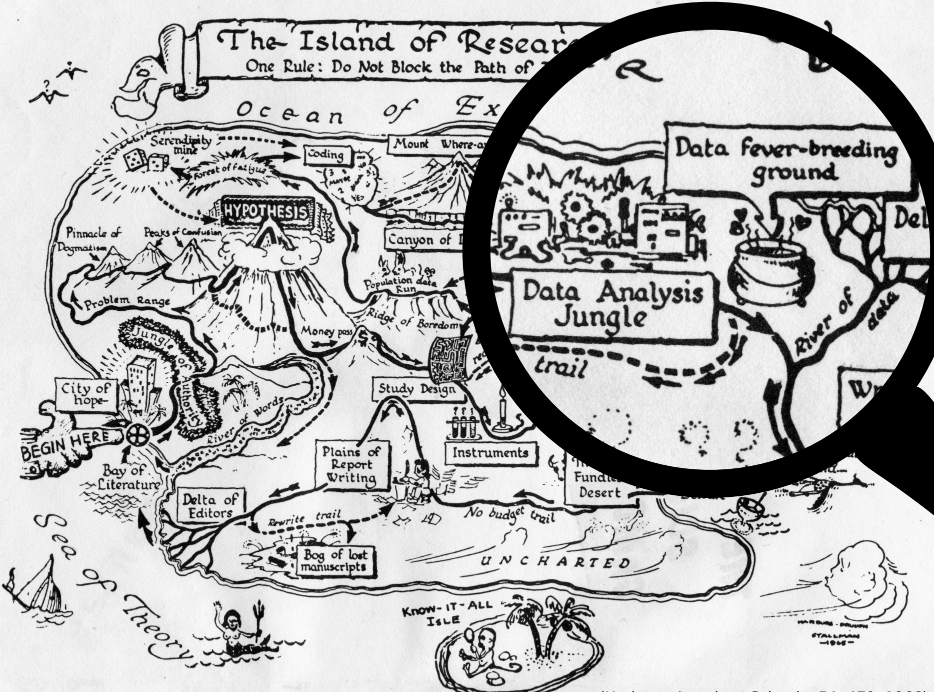



(Harburg, American Scientist 54: 470, 1966)

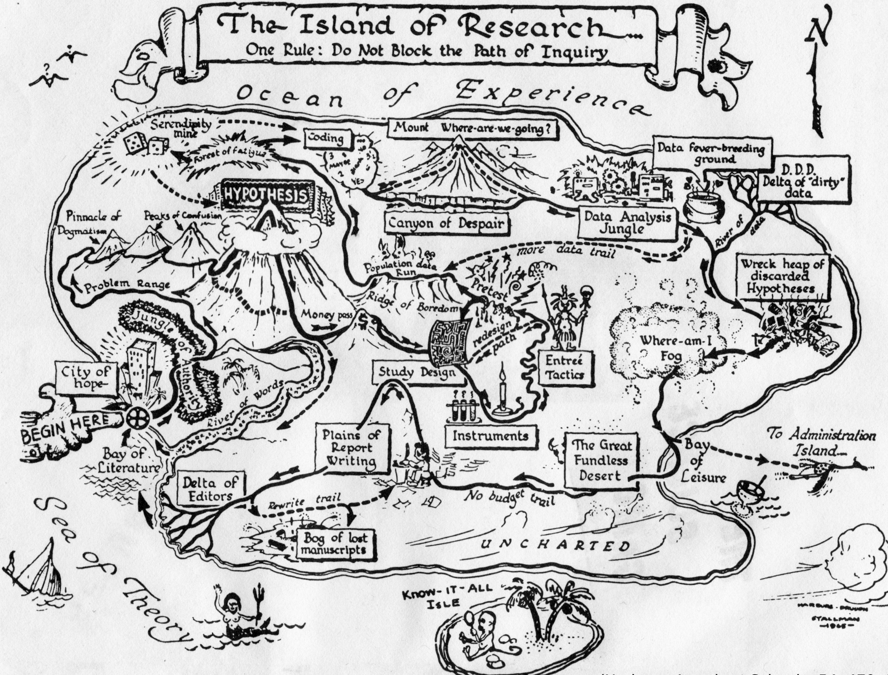

(Harburg, American Scientist 54: 470, 1966)



(Harburg, American Scientist 54: 470, 1966)




(Harburg, American Scientist 54: 470, 1966)




(Harburg, American Scientist 54: 470, 1966)

# Progress in data science

- In the 60's, data analysis was no picnic.
- Today it's easier. We have
  - databases to store and access large amounts of data
  - high-performance computing
  - sound data analysis principles from probability & statistics

# Progress in data science

- In the 60's, data analysis was no picnic.
- Today it's easier. We have
  - databases to store and access large amounts of data
  - high-performance computing
  - sound data analysis principles from probability & statistics
- Challenge to further progress:
  - The basic principles do not consider the computational cost
  - For complex problems, exact solutions are computationally impossible
  - Textbook approximate methods too slow or too approximate

# Progress in data science

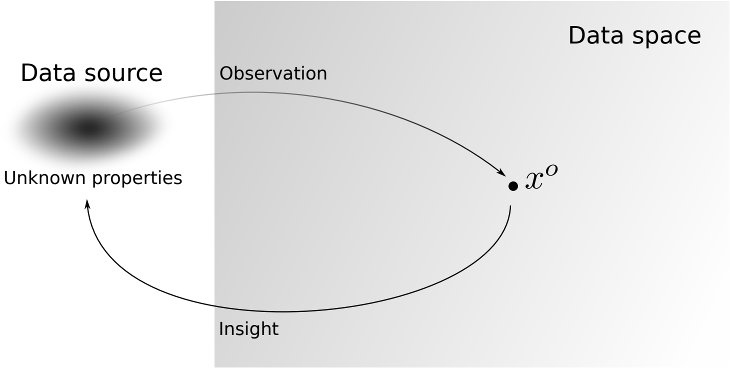
- In the 60's, data analysis was no picnic.
- Today it's easier. We have
  - databases to store and access large amounts of data
  - high-performance computing
  - sound data analysis principles from probability & statistics
- Challenge to further progress:
  - The basic principles do not consider the computational cost
  - For complex problems, exact solutions are computationally impossible
  - Textbook approximate methods too slow or too approximate
- Need for new data analysis methods with a good trade-off between speed and accuracy

# Message of the talk

AI and machine learning greatly improve the trade-off between speed and accuracy in data analysis.

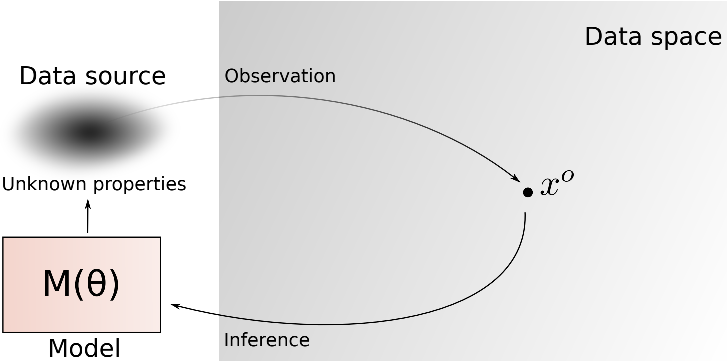
# Overall goal of data analysis

- Use observed data $\boldsymbol{x}^o$ to learn about their source
- Enables decision making, predictions, ...



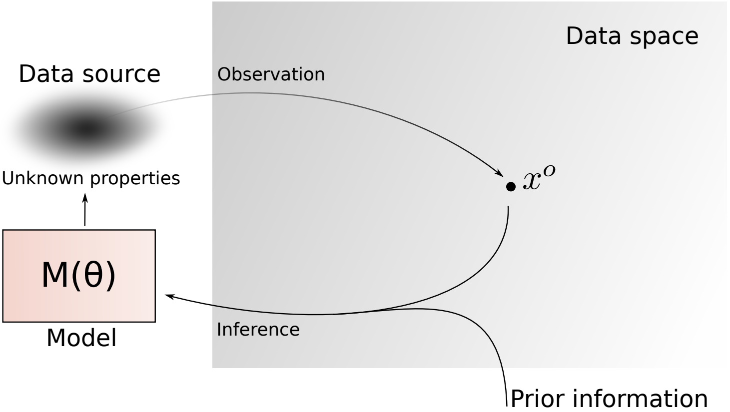
# General approach

- Set up a model with potential properties $\boldsymbol{\theta}$ (parameters)
- See which $\boldsymbol{\theta}$ are in line with the observed data $\boldsymbol{x}^o$



# General approach

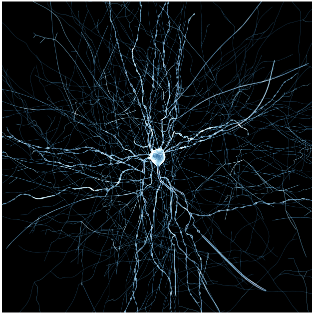
- Set up a model with potential properties $\boldsymbol{\theta}$ (parameters)
- See which $\boldsymbol{\theta}$ are in line with the observed data $\boldsymbol{x}^o$



# Simulator-based models

- Models specified by a data generating mechanism
  - e.g. emulators / simulators of some complex physical or biological process
  - aka: generative models, implicit models

- Widely used in science & engineering
  - Neuroscience: Simulating neural activity
  - Evolutionary biology: Simulating evolution
  - Robotics: Simulating actions
  - ...

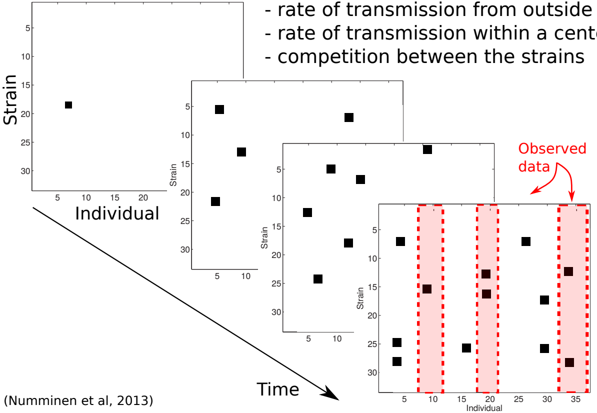


Simulated neural activity in rat somatosensory cortex (Figure from `https://bbp.epfl.ch/nmc-portal`)

# Example: Bacterial transmissions in child care centres

- Model: latent continuous-time Markov chain for the transmission dynamics and an observation model
- What can we say about the parameters of interest?

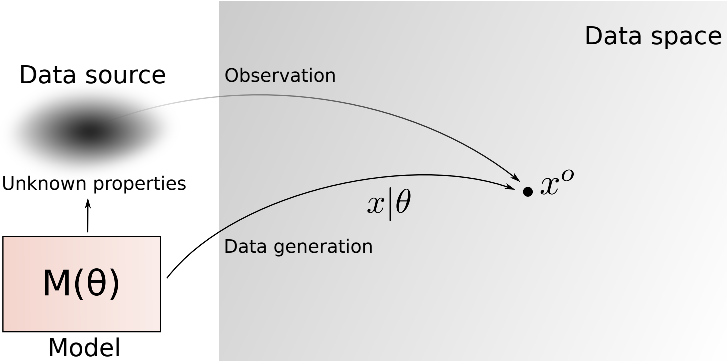
Parameters of interest:
- rate of transmission from outside
- rate of transmission within a center
- competition between the strains




(Numminen et al, 2013)

# The likelihood function

- Measures agreement between $\boldsymbol{\theta}$ and the observed data $\boldsymbol{x}^o$
- Probability to generate data like $\boldsymbol{x}^o$ if hypothesis $\boldsymbol{\theta}$ holds



# Research question

- For child care centre and other simulator-based models: likelihood function is too expensive to evaluate.
- Research question:

  How to efficiently perform (Bayesian) inference when
  - the likelihood function cannot be evaluated
  - but sampling from the model is possible

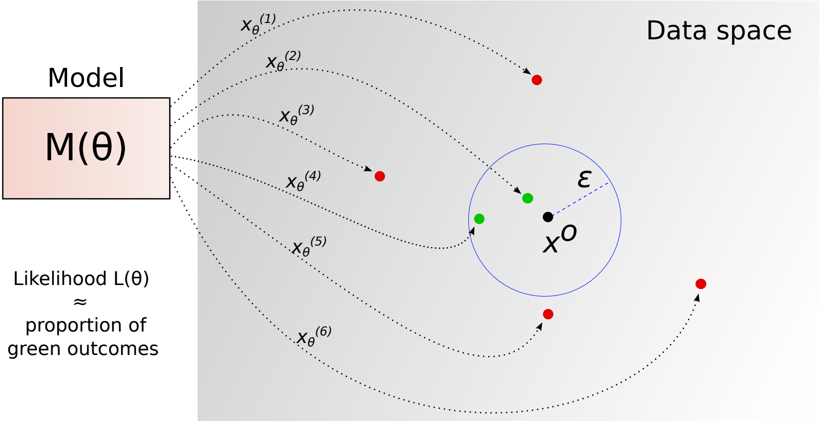
# Research question

- For child care centre and other simulator-based models: likelihood function is too expensive to evaluate.
- Research question:

  How to efficiently perform (Bayesian) inference when
  - the likelihood function cannot be evaluated
  - but sampling from the model is possible
- Area of research called "likelihood-free inference" or "approximate Bayesian computation"

# Simple approach: approximate by counting

Likelihood: Probability to generate data like $\boldsymbol{x}^o$ for parameter value $\boldsymbol{\theta}$



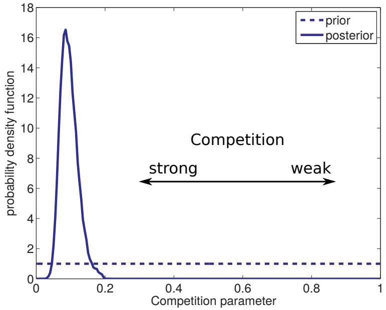

Likelihood L(θ)
≈
proportion of
green outcomes

# Example: Bacterial transmissions in child care centres

- Data: *Streptococcus pneumoniae* colonisation for 29 centres
- Inference with a smarter version of the counting-based approach (Markov chain Monte Carlo ABC)
- Reveals strong competition between different bacterial strains

Expensive:

- 4.5 days on a cluster with 200 cores
- More than one million simulated data sets



# Fast Bayesian inference using machine learning

- We developed a fast inference algorithm using machine learning (Bayesian optimisation).
- Roughly equal results using 1000 times fewer simulations.

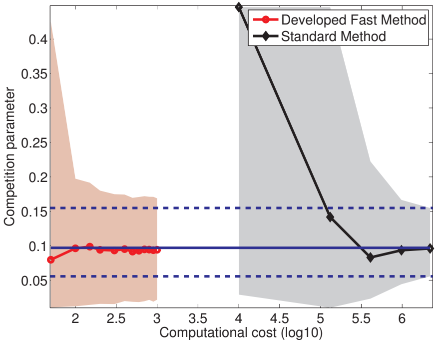
4.5 days with 200 cores

↓

90 minutes with seven cores

Posterior means: solid lines,

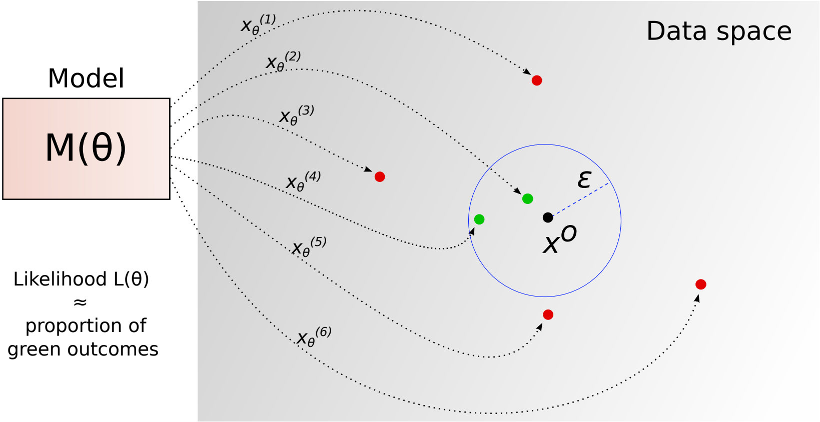
credibility intervals: shaded areas or dashed lines.




(Gutmann and Corander, JMLR, 2016)

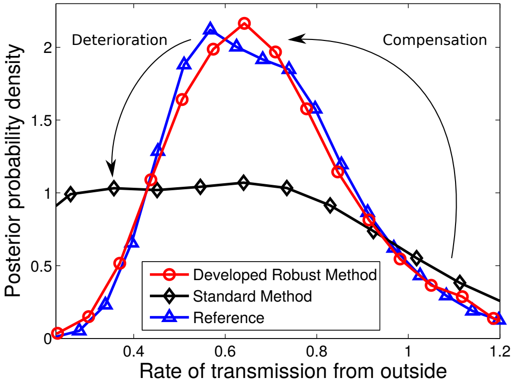
# Simple approach: approximate by counting

Likelihood: Probability to generate data like $\boldsymbol{x}^o$ for parameter value $\boldsymbol{\theta}$



# Robust Bayesian inference using machine learning

- Traditionally, expert knowledge is used to judge whether the simulated and observed data are close
- But experts make mistakes too
- Robustify using machine learning (Gutmann et al, 2014, 2017)



# Conclusions

- Complex data analysis problems in science and engineering
- Inference for models where the likelihood is intractable but sampling is possible (likelihood-free inference)
- Machine learning to accelerate and robustify the inference

$\Rightarrow$ Improved trade-off between speed and accuracy

# Conclusions

- Complex data analysis problems in science and engineering
- Inference for models where the likelihood is intractable but sampling is possible (likelihood-free inference)
- Machine learning to accelerate and robustify the inference

⇒ Improved trade-off between speed and accuracy

Further information:

- Review paper: Lintusaari et al, Systematic Biology, 2017
- My homepage: `http://homepages.inf.ed.ac.uk/mgutmann`
- Software: ELFI – Engine for Likelihood-Free Inference
  `http://elfi.readthedocs.io`