Reducing dimensions and cost for UQ in complex systems

Monday 5th March 2018 to Friday 9th March 2018
Newton Institute, Cambridge

Model selection, model frames, and scientific interpretation

Julia Brettschneider

WARWICK
THE UNIVERSITY OF WARWICK

Warwick Statistics

BIG DATA INSTITUTE

Outline

- Models and sampling
- Berkson's bias and related phenomena
- Length bias and related phenomena
- Throughout the talk there will are examples (admissions, cancer screening, individual trader behaviour, quality inspection, microscopy)

Berkson's bias

Two independent events become conditionally dependent (negatively) given that at least one of them occurs.

$$P(A|B) = P(A), \text{ but } P(A|\, B, A \cup B) < P(A|A \cup B)$$

- Example of a selection bias
- Selection leads to correlation between previously unassociated variables
- Aka Berkson's paradox, Berkson's fallacy, conditioning on a collider
- Version with a priori positive dependency:

$$P(A|B) > P(A), \text{ but } P(A|\, B, A \cup B) < P(A|A \cup B)$$

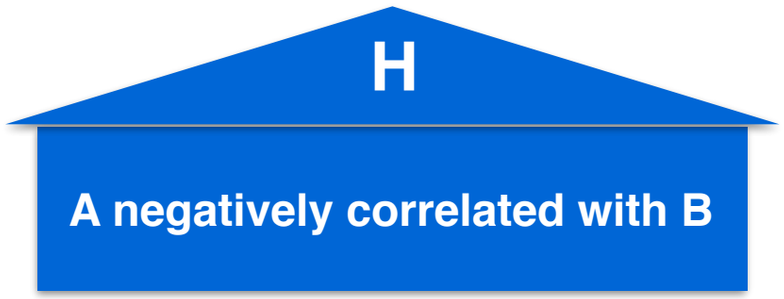
- (Continuous) random variable version: $P(X \geq x|Y \geq y) \geq P(X \geq x)$,

$$\text{but } P(X \geq x|Y \geq y, X + Y \geq x + y) < P(X \geq x|X + Y \geq x + y)$$

Berkson's bias: the original

Two independent events become conditionally dependent (negatively) given that at least one of them occurs.

A not correlated with B



$$P(A|\, B, A \cup B) < P(A|A \cup B)$$

- Observed (spurious) negative correlation between risk factor (B) and disease (A) in hospital in-patient population
- In other words: Hospitalised patient w/o risk factor has increased likelihood for disease (compared to person in the general population)
- Explanation: Patients w/o diabetes may have had some non-diabetes cholecystitis-causing reason to enter the hospital
- In original example (*) A=cholecystitis B=diabetes

(*) Berkson, Joseph (June 1946). "Limitations of the Application of Fourfold Table Analysis to Hospital Data". *Biometrics Bulletin*. **2** (3): 47–53.

Berkson's bias: the maths

Two independent events become conditionally dependent (negatively) given that at least one of them occurs.

$$P(A|B) = P(A), \text{ but } P(A|\, B, A \cup B) < P(A|A \cup B)$$

How is this possible?

$$P(A \cap B) = P(A)P(B) \not\Leftrightarrow P(A|C)P(B|C) = P(A \cap B|C)$$

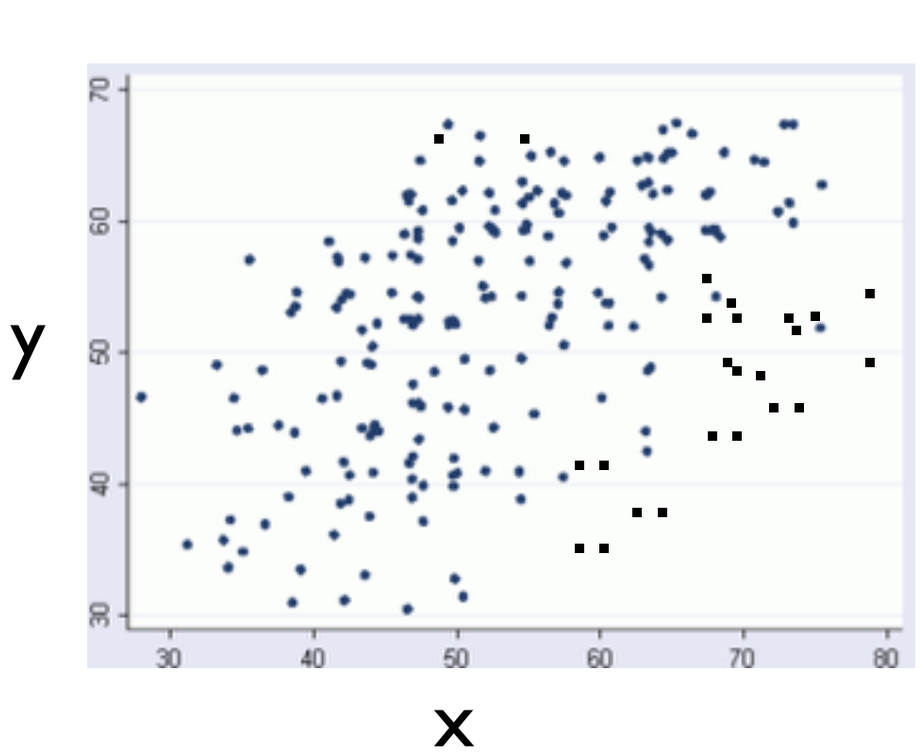
Possible mechanism leading to Berkson's bias:

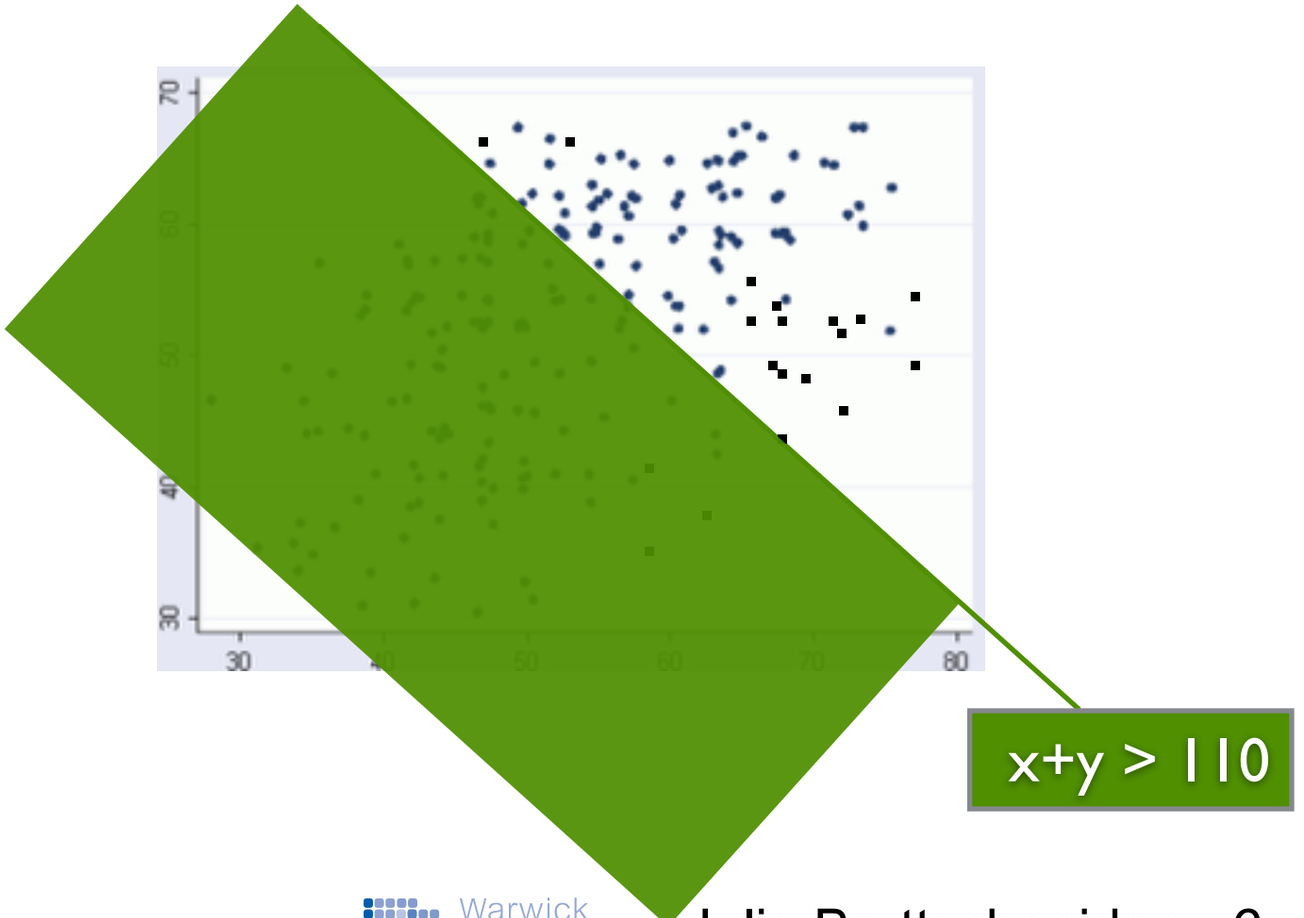
occurrence of A or B inflates the chance of A

$$P(A|\, B, A \cup B) = P(A|\, B \cap (A \cup B)) = P(A|B) = P(A) < P(A|A \cup B)$$

Berkson's bias: a cartoon version

Two independent (or positively dependent) random variables become conditionally dependent (negatively) given that at least one of them is above a threshold.

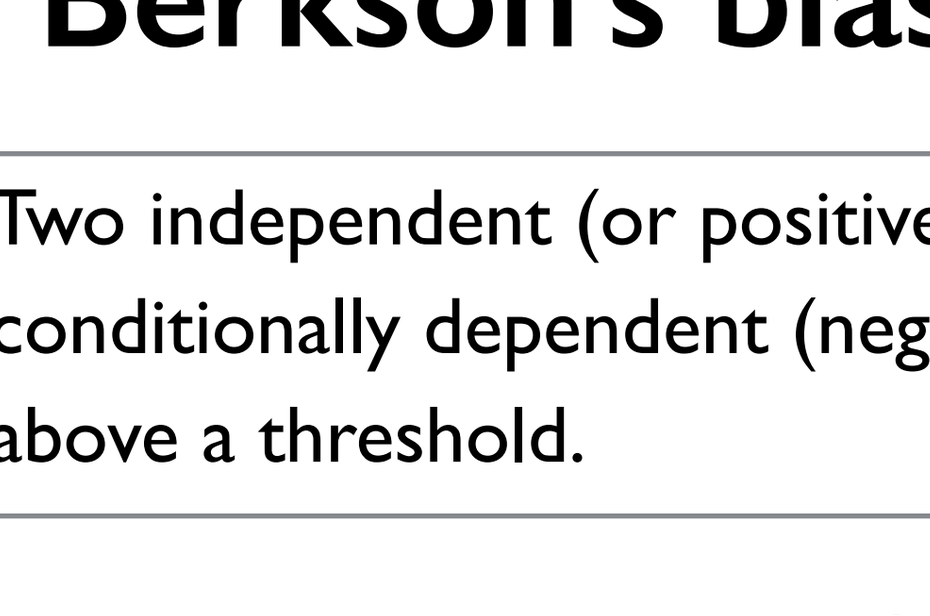


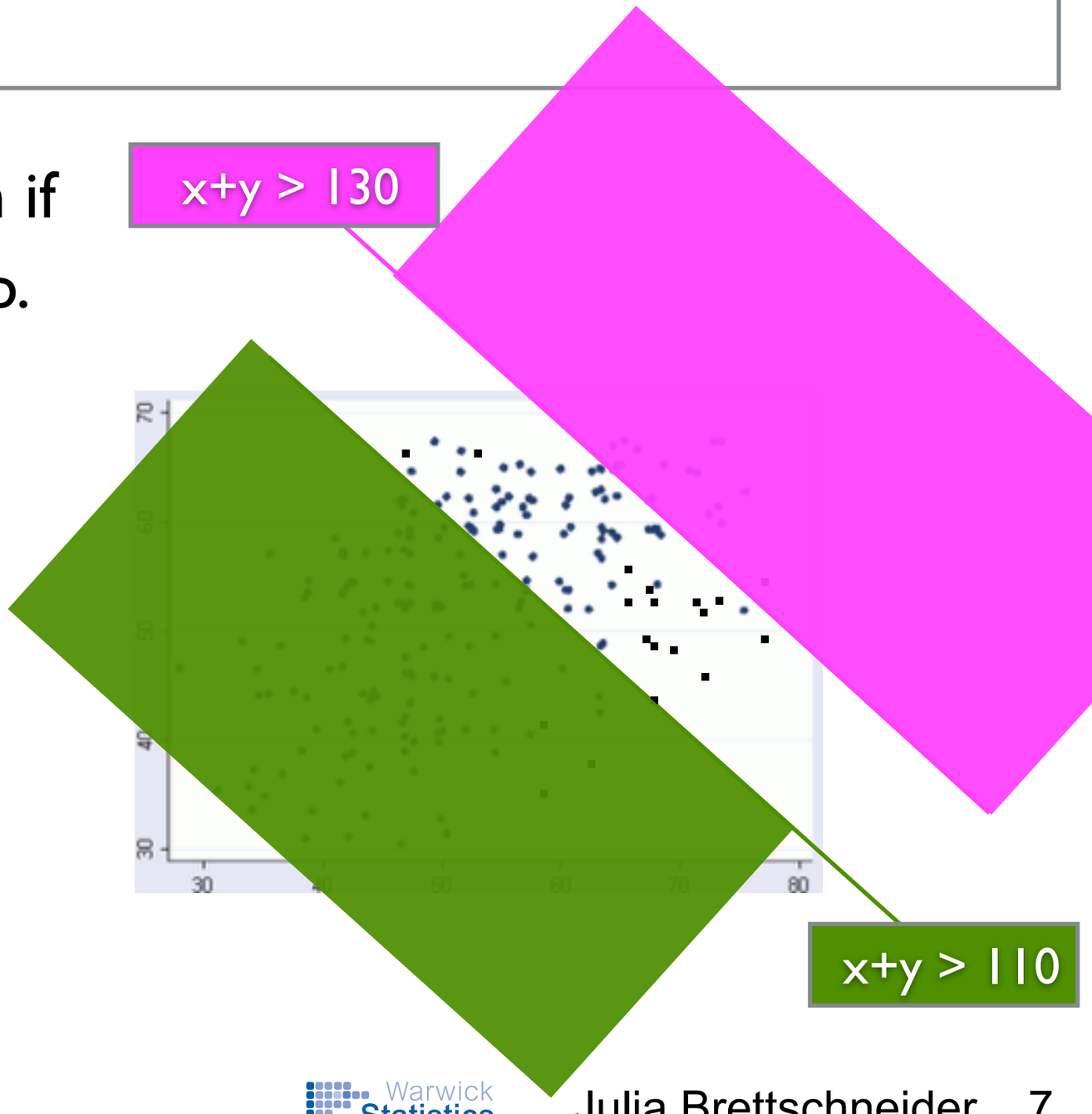


Berkson's bias: a cartoon version

Two independent (or positively dependent) random variables become conditionally dependent (negatively) given that at least one of them is above a threshold.

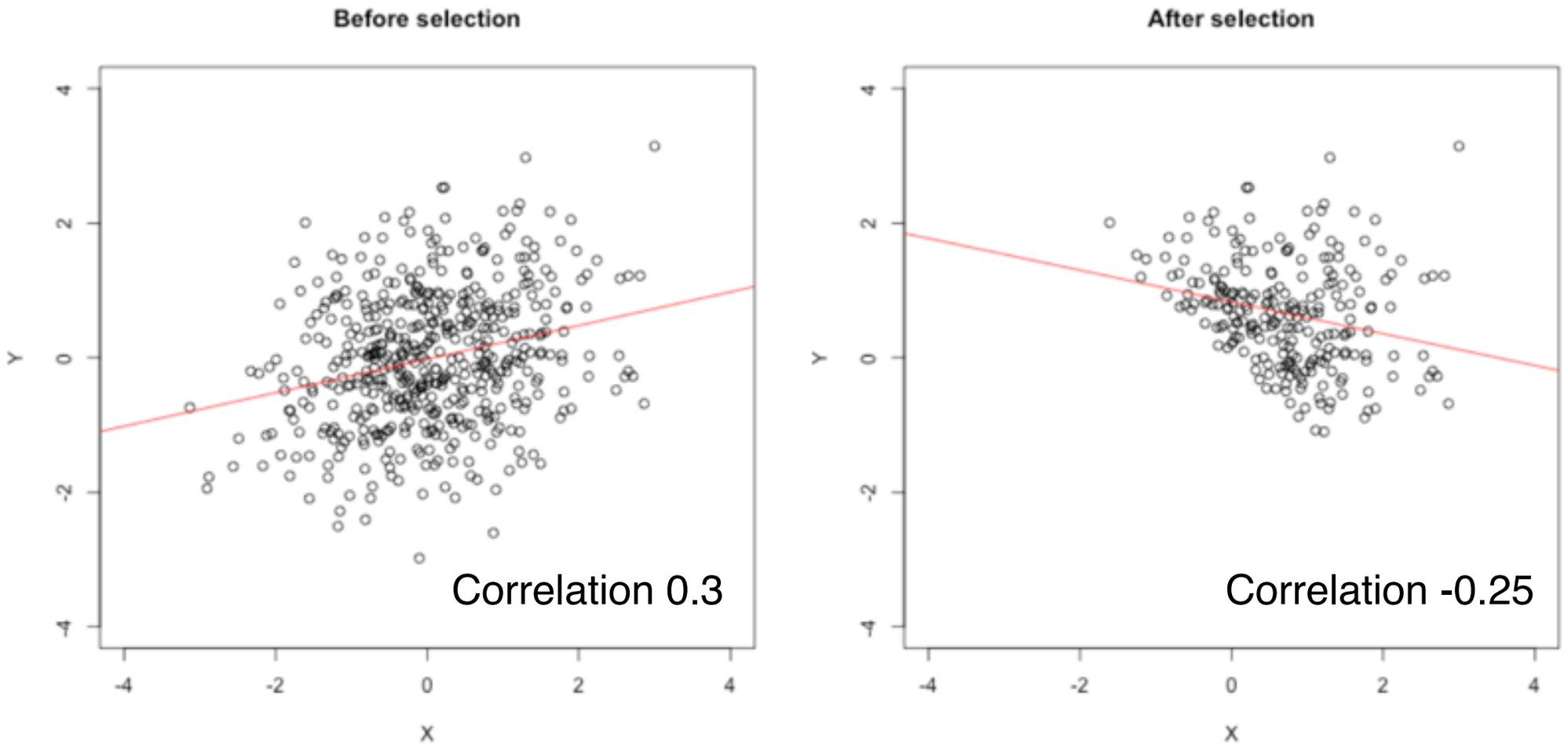
Even stronger negative correlation if additional selection at the very top.





Berkson's bias: grad school

Grad school admission in the US using GRE and GPA: trend reversal. GRE and GPA negative among *admitted* students but positive among *applicants* (*). Illustration with simulated data (**).



(*) Robyn Dawes, Graduate Admission Variables and Future Success, *Science* 28 Feb 1975: Vol. 187, Issue 4178, pp. 721-723

(**) Illustration from https://hardsci.wordpress.com/2014/08/04/

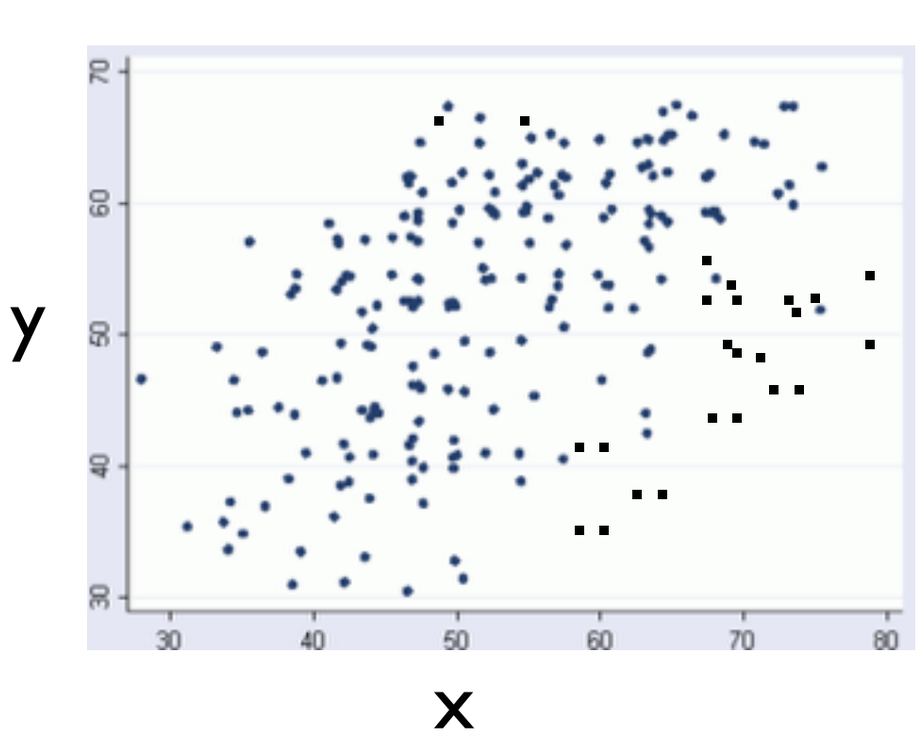
Berkson's bias in UG admissions?

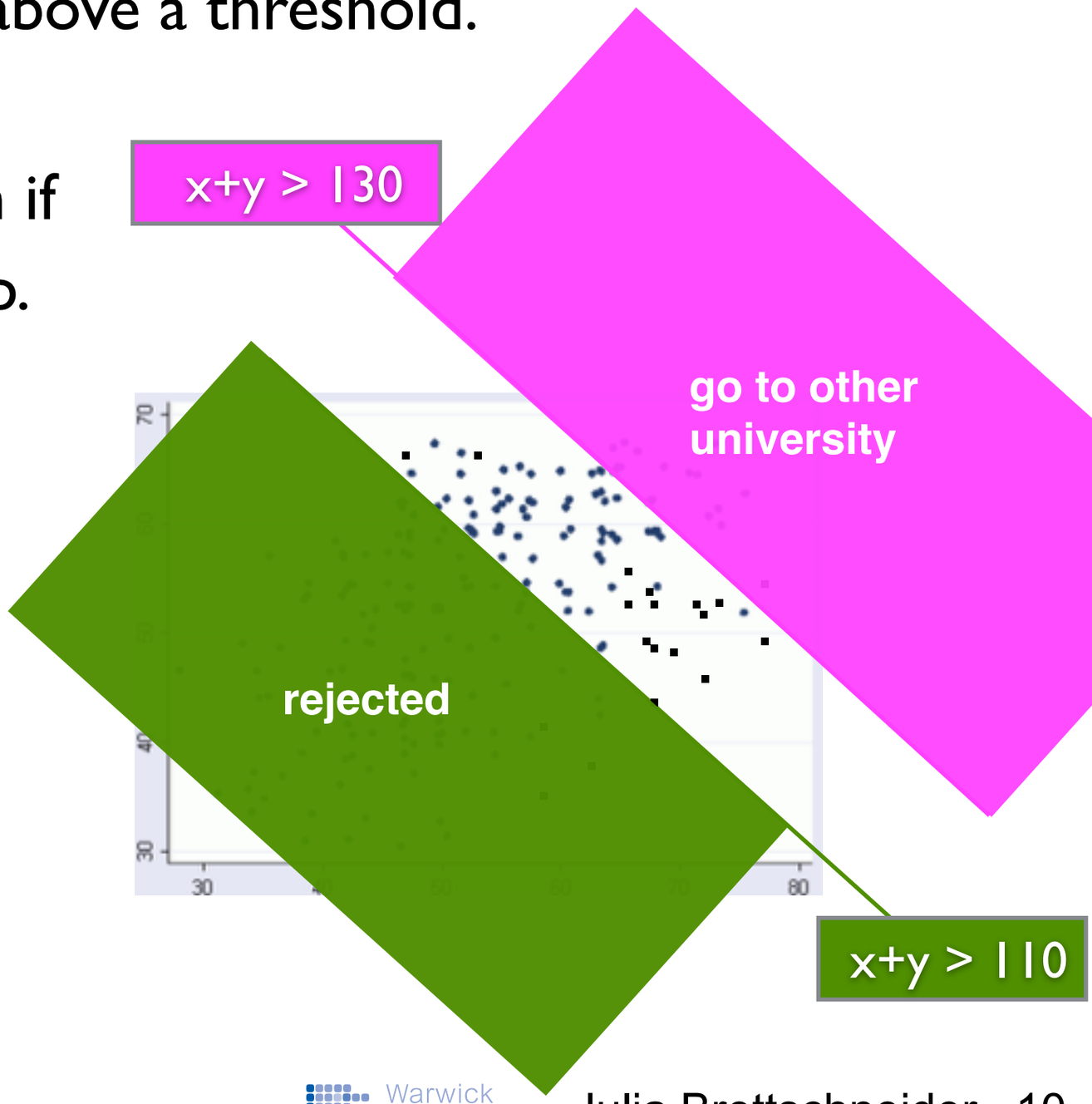
- UG admissions may involve uncorrelated or weakly correlated criteria (school grades, interview, additional papers (STEP), reference letters, statements etc).
- Criteria typically weakly positively correlated or independent.
- How predictive are admission criteria for success at university?
- In UK also relevant: How predictive are predicted (by schools) A-level grades for achieved A-level grades?
- Observations at Warwick Statistics Department (preliminary analysis of 5 years of UG admissions data) shows evidence of Berkson's bias for Further Maths grade and other criteria (STEP etc).

Admissions (hypothetical data)

Two independent (or positively dependent) criteria used for admissions may become conditionally dependent (negatively) given that at least one of them is above a threshold.

Even stronger negative correlation if additional selection at the very top.

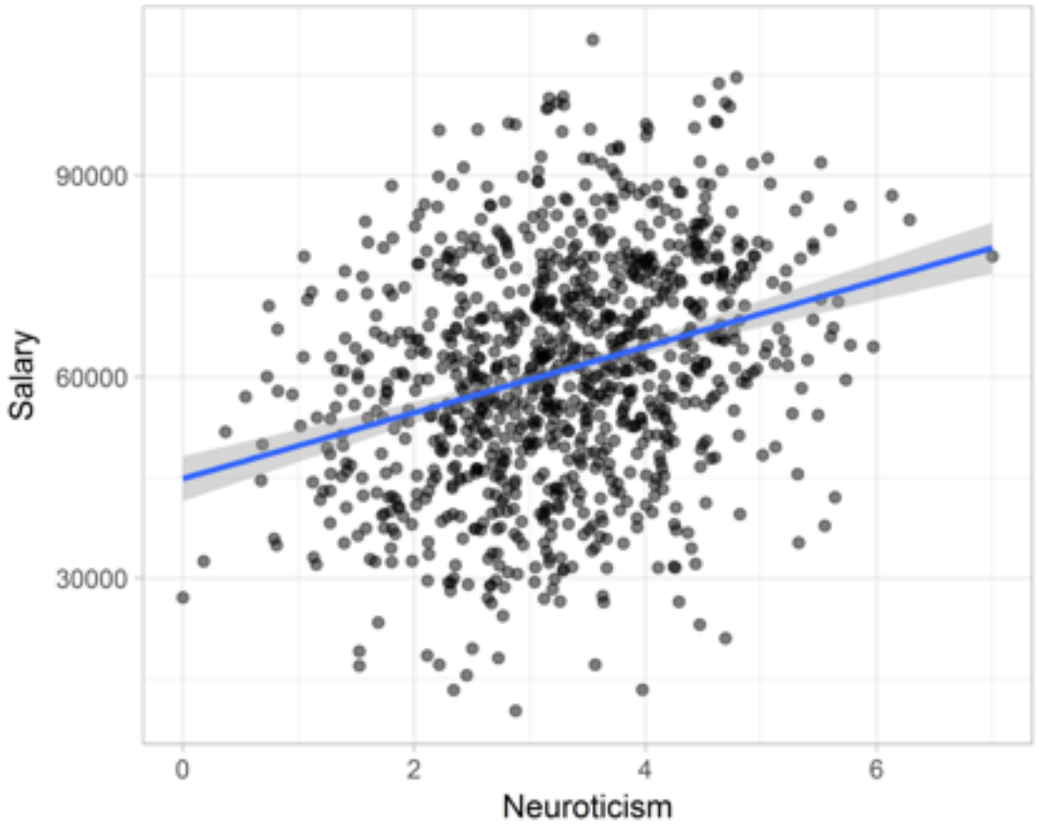




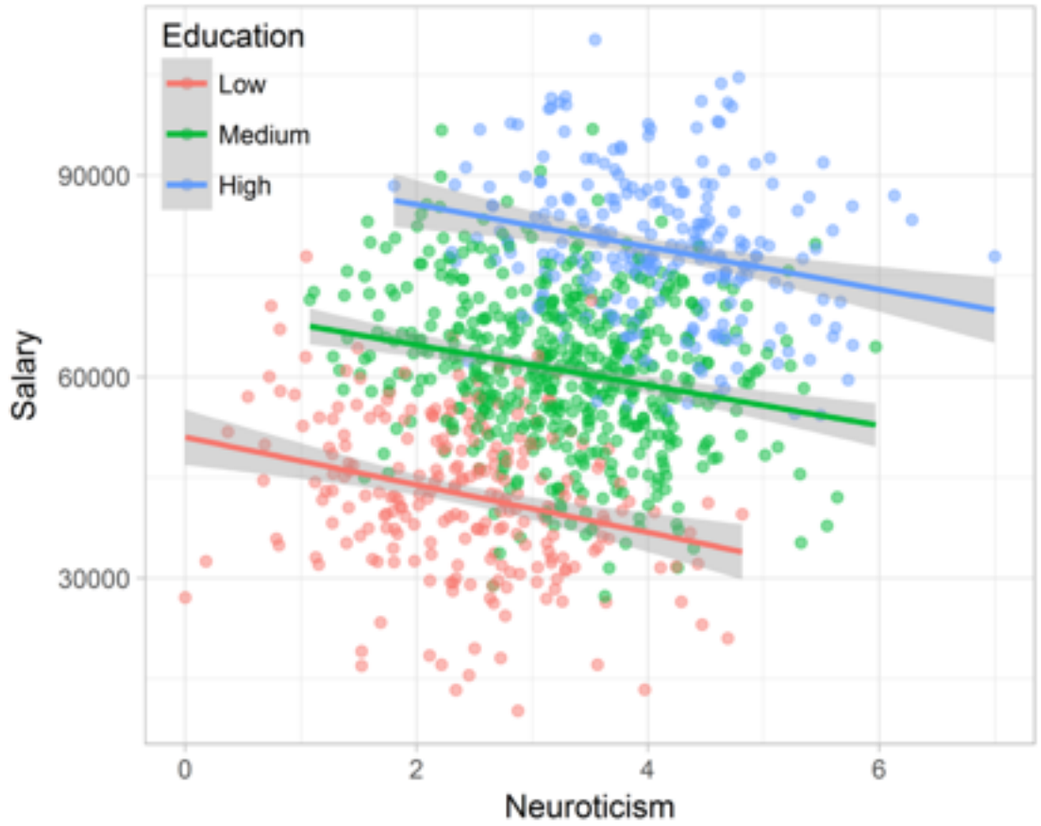
Simpson's paradox

Trend appears in several different groups of data but disappears or reverses when these groups are combined (aggregation).

Continous case (aka ecological regression):



positive correlation overall



negative correlation for each group

https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/

Simpson's paradox (ct'ed)

Discrete case

Example: Which language group in Canada has more children?

Mean Number of Children 1971-1976
Source: Canadian Census 1976, N. Keyfitz,
Applied Mathematical Demography

	French	English	
Quebec	1.8	1.64	**French have more**
Other Provinces	2.14	1.97	**French have more**
Total Canada	1.85	1.95	**English have more**

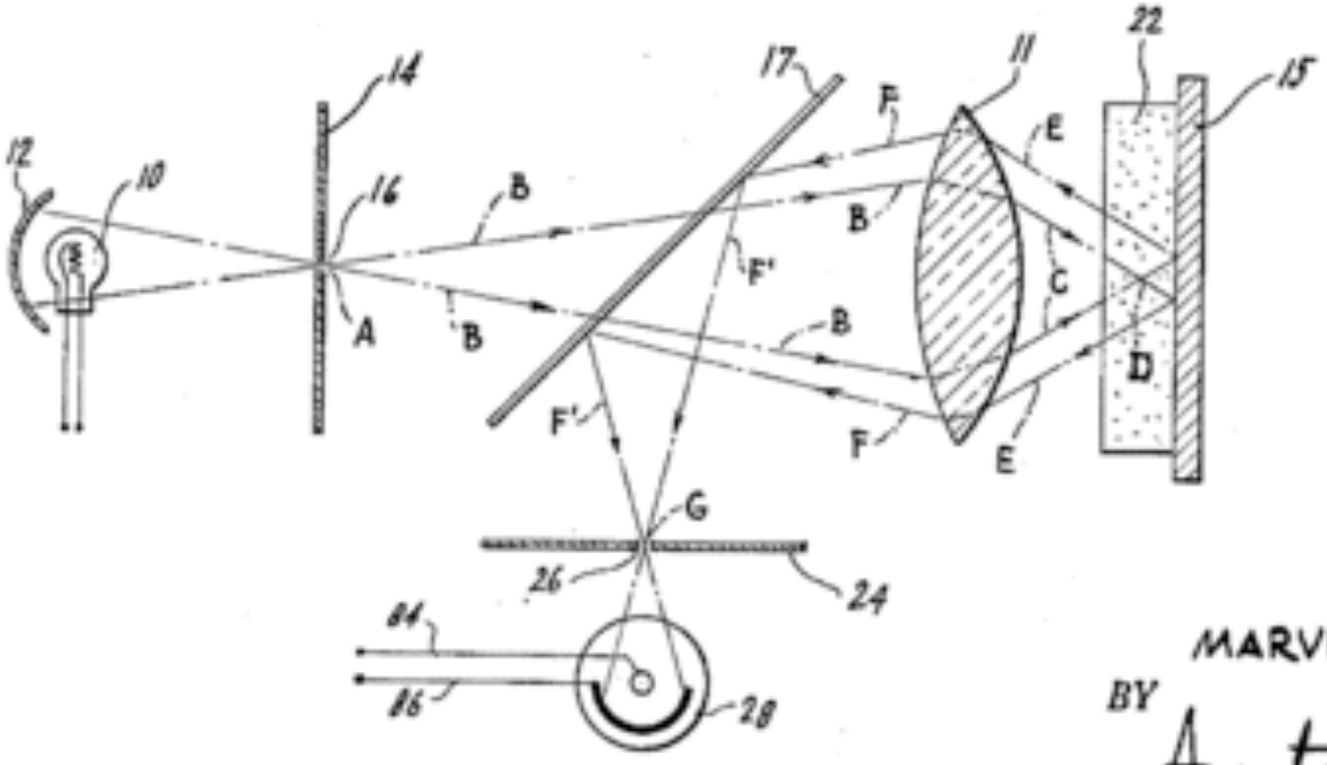
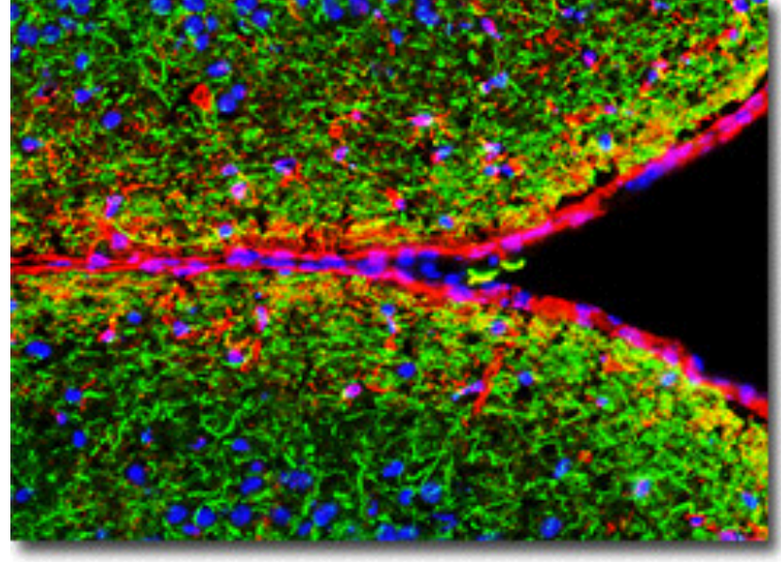
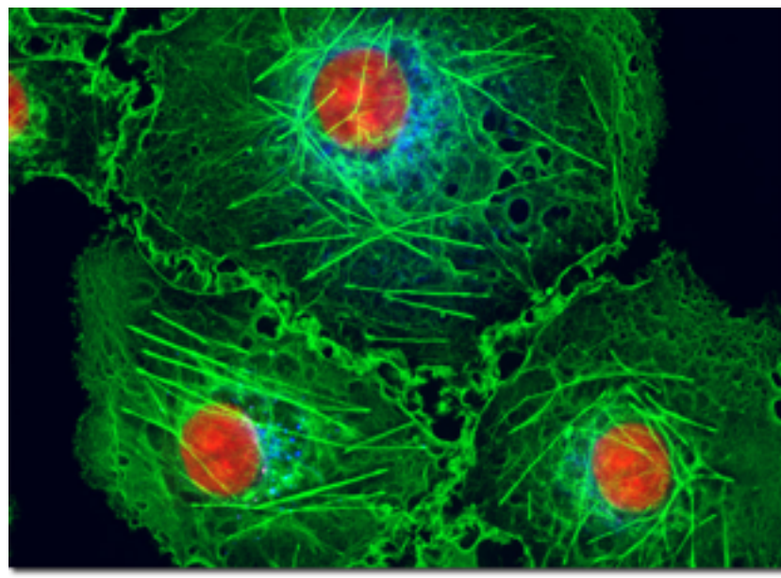


FIG. 3.

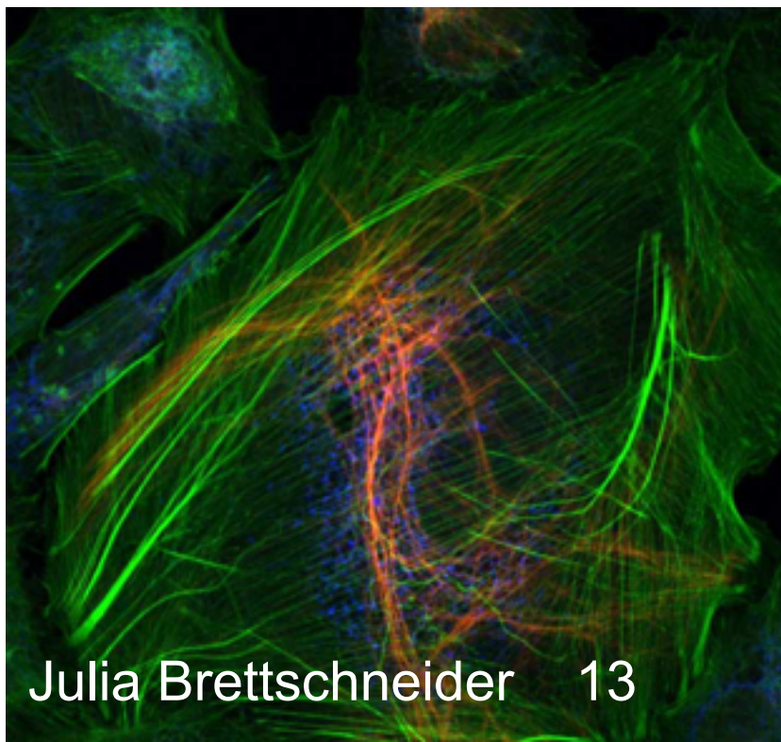
INVENTOR.
MARVIN MINSKY
BY Amster & Levy
ATTORNEYS

Confocal fluorescent laster microscopy





Fluorescent microscopy for observing intensity & location of proteins



Collaboration with Steve Royle Lab

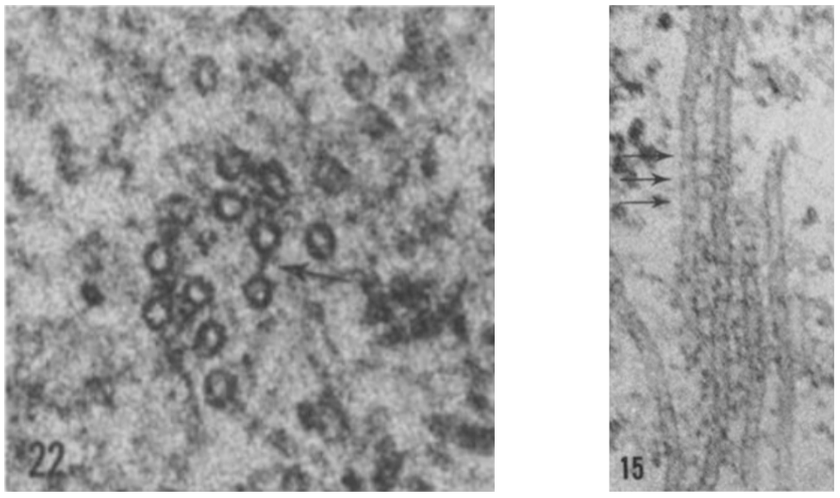


Figure 2: Sample microscope images taken perpendicular to the microtubule axis, left, and parallel to the microtubule axis, right (Hepler et al., 1970). Arrows indicate the location intermicrotubule bridges formed by mesh.

centre for mechanochemical cell biology

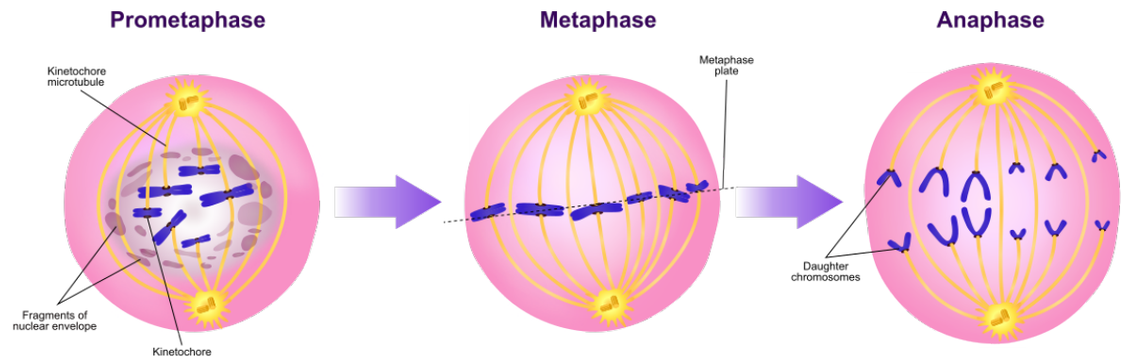


Figure 1: Diagram of the stages of mitosis (Ali Zifan).

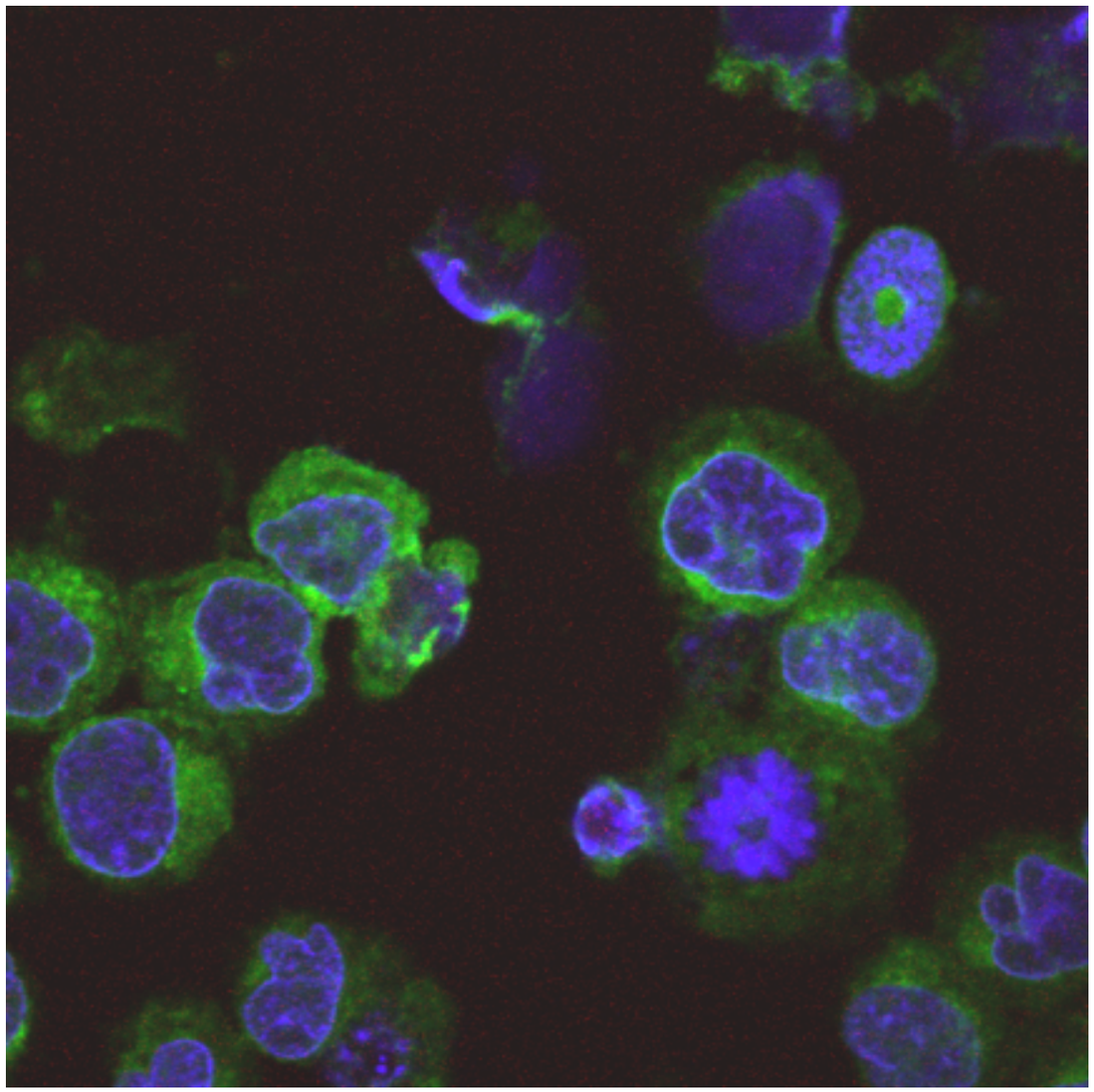
Nixon*, F.M., Honnor*, T.R., Starling, G.P., Beckett, A.J., Johansen, A.M., Brettschneider, J.A., Prior, I.A. & Royle, S.J. **Microtubule organization within mitotic spindles revealed by serial block face scanning EM and image analysis,** J Cell Science, April 2017

TR Honnor, JA Brettschneider, AM Johansen
Differences in spatial point patterns with applications to subcellular biological structures
CRiSM Working Paper Series No. 17-01, 2017

TR Honnor, AM Johansen, JA Brettschneider
A nonparametric test for dependency between between estimated local bulk movement patterns
CRiSM Working Paper Series No. 17-03, 2017

Example: Quantifying protein abundance in their actual locations in cells

Sub-cellular localisation of tumour antigen SSX2IP in leukemia cells



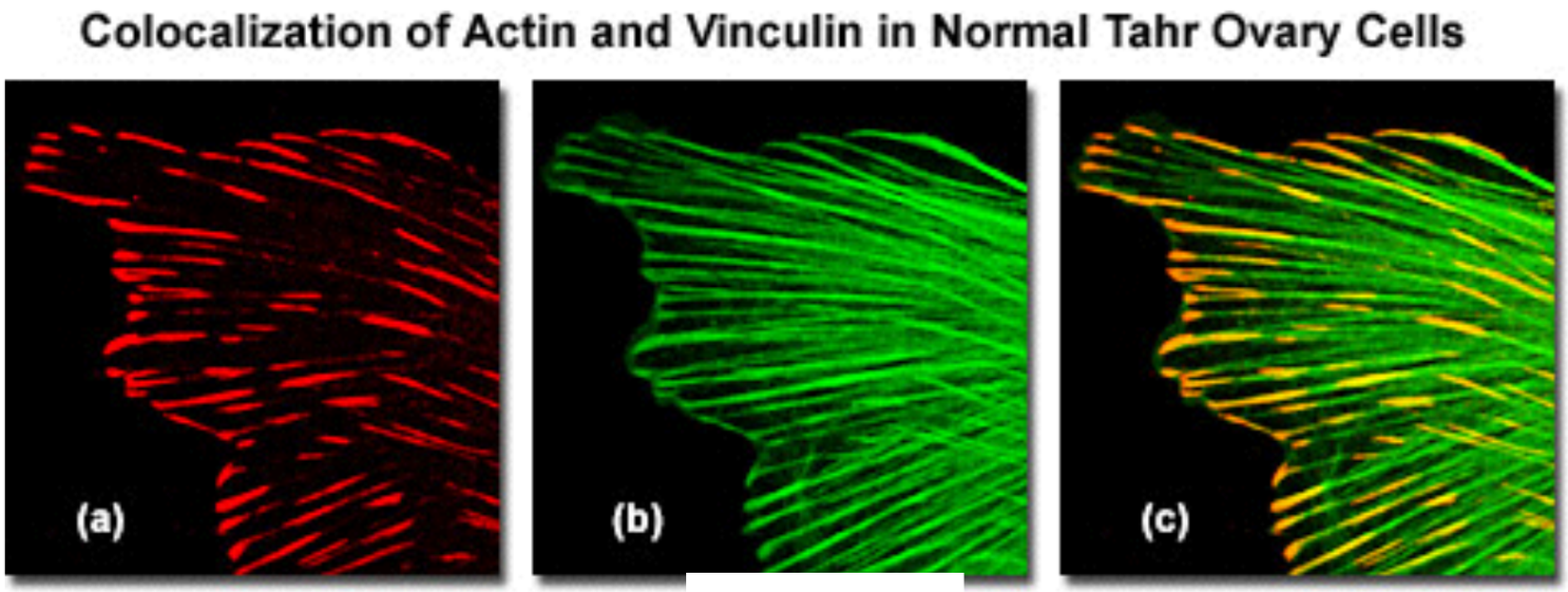
Green: SSX2IP expression visualised by anti-SSX2IP-fluorescein isothiocyanate on the cell's surface.

Blue: Stained Cell nuclei using 4,6'-diamino-2-phenylindole (DAPI).

Protocol of the experiment:
Leukaemia cell line K562 air dried for 4-18hours onto glass microscope slides, stored at -20°C wrapped in saranwrap, defrosted, stained with antigen specific primary, and fluorescently labelled secondary antibodies.

http://www.intechopen.com/books/novel-gene-therapy-approaches/identification-and-validation-of-targets-for-cancer-immunotherapy-from-the-bench-to-bedside

Example: Colocalisation of two proteins in cancer cells



Colocalization in the lateral optical plane of the cytoskeletal protein **actin** with **vinculin**, a protein associated with focal adhesion and adherens junctions.

Applications:

- Detect physical location within cell
- Uncover functions of proteins based on location
- Unravel interactions, build networks, infer function

Need: quantification, inferential methods for colocalisation

http://www.olympusmicro.com/primer/techniques/confocal/applications/colocalization.html

Quantifying colocalisation

Pearson correlation

$$r_p = \frac{\sum_i (A_i - a)(B_i - b)}{\sqrt{\sum_i (A_i - a)^2 (B_i - b)^2}}$$

where A_i and B_i are the voxel or pixel intensities (also called grey values) of channels A and B, respectively, and a and b are the corresponding average intensities over the entire image.

- Scaling invariant
- Does not fully take into account spatial information, noise

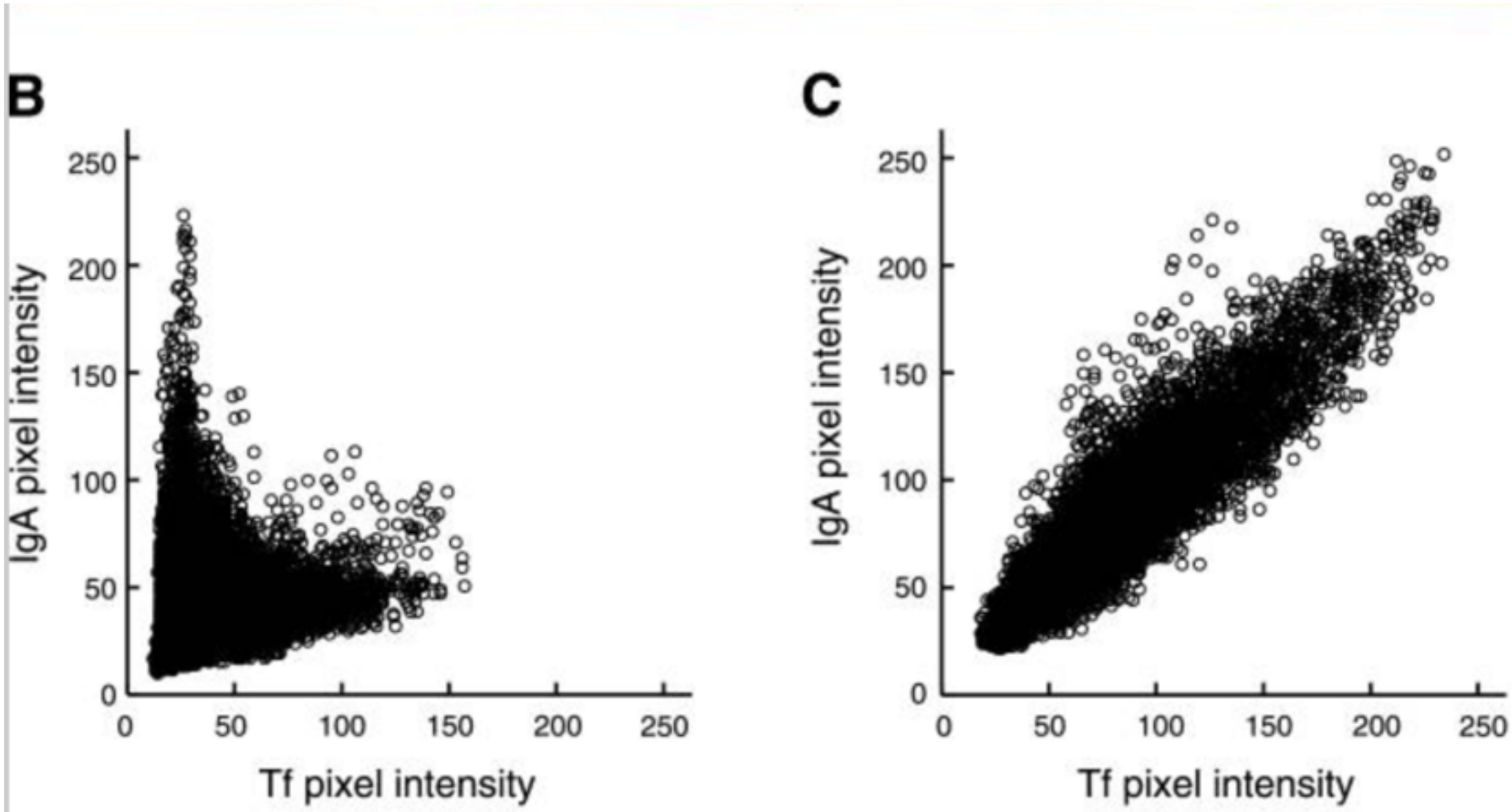
Manders coefficients

$$M_1 = \frac{\sum_i A_i I_{B_i}}{\sum_i A_i}, M_2 = \frac{\sum_i B_i I_{A_i}}{\sum_i B_i}$$

where $I_{A_i} = 0$ if $A_i = 0$ and $I_{A_i} = 1$ if $A_i > 0$ (analogously for B_i). Thus, M_1 and M_2 can be interpreted as the amount of signal intensities of colocalizing objects in each channel, relative to the total signal intensity in that channel.

- Two measures, not scaling invariant
- Problem with all measures: depends on selection of regions. Automatic selection addressed by Wang et al (2016)

Detecting correlation

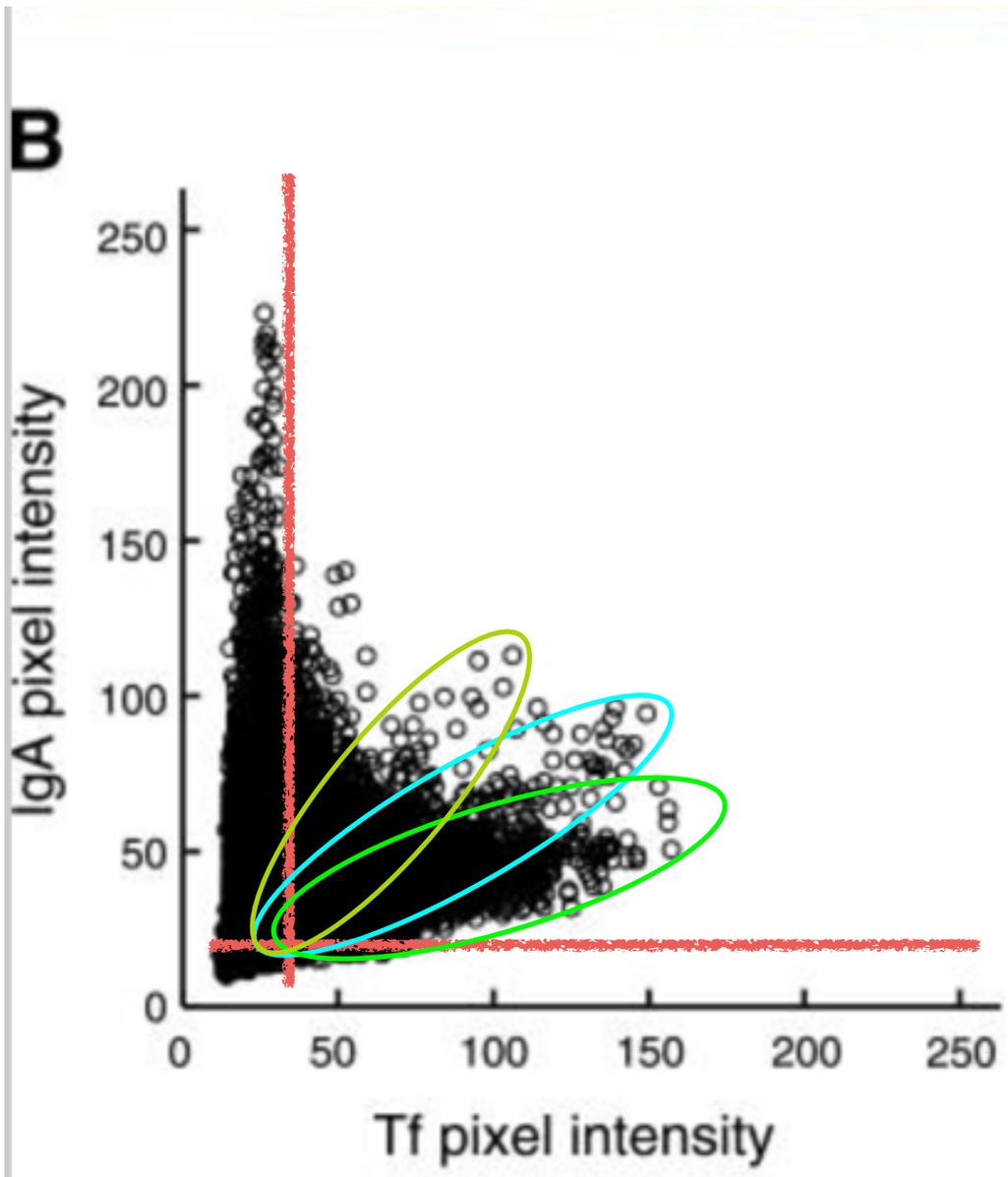


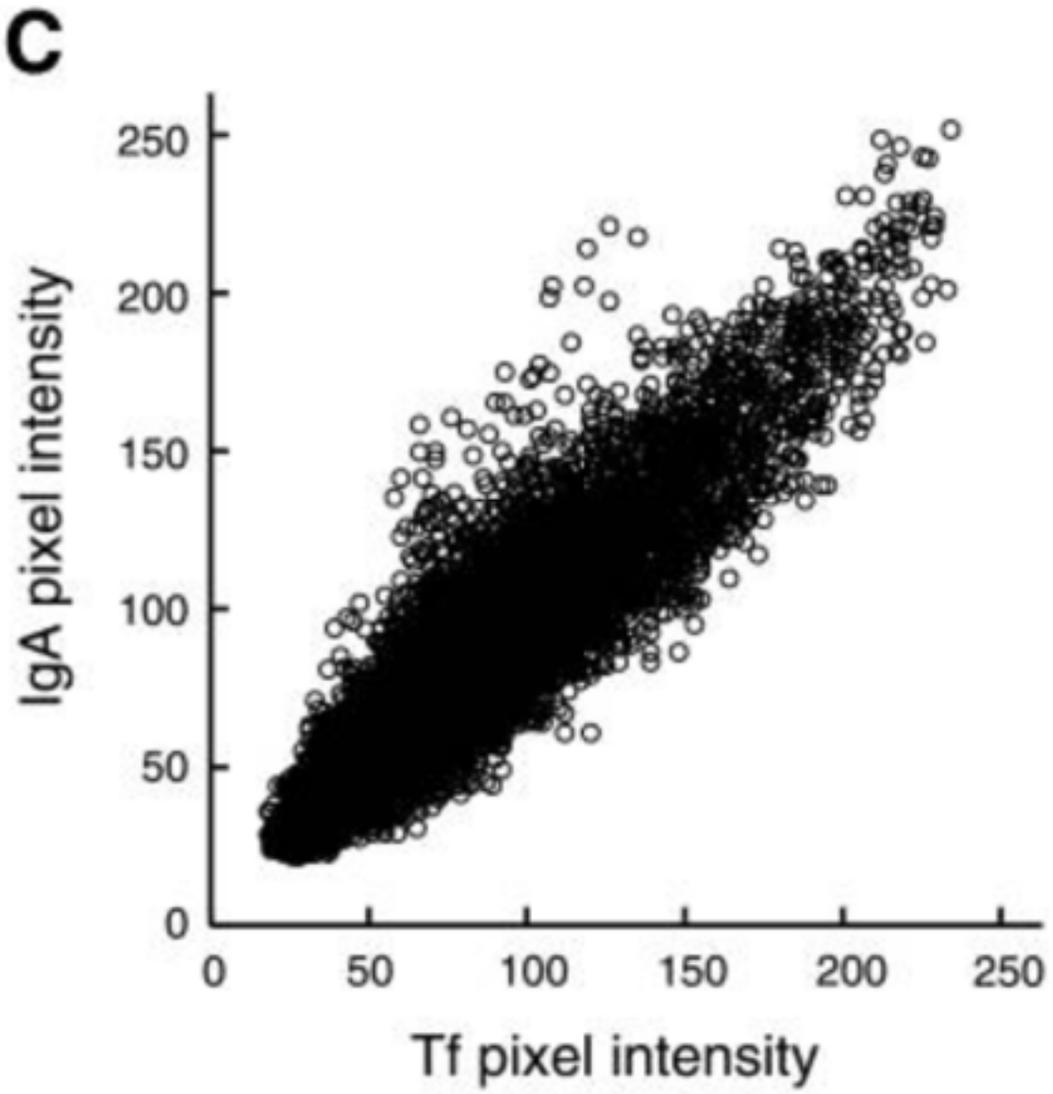
- Another case of induced negative correlation (after removing noise)?
- Correlation hidden by ambiguity due to subgroups?

Easy case

Image source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074624/

Background thresholds





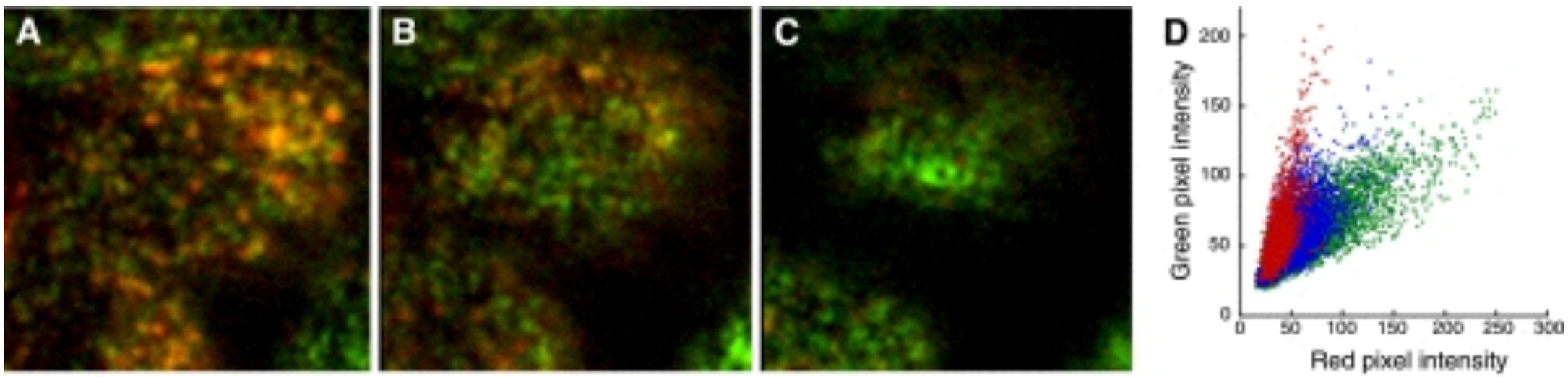
Costes: Identify background

Partial collocalisation

Easy case

Image source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074624/

Meaningful subgroups



Colocalization without a simple linear relationship.

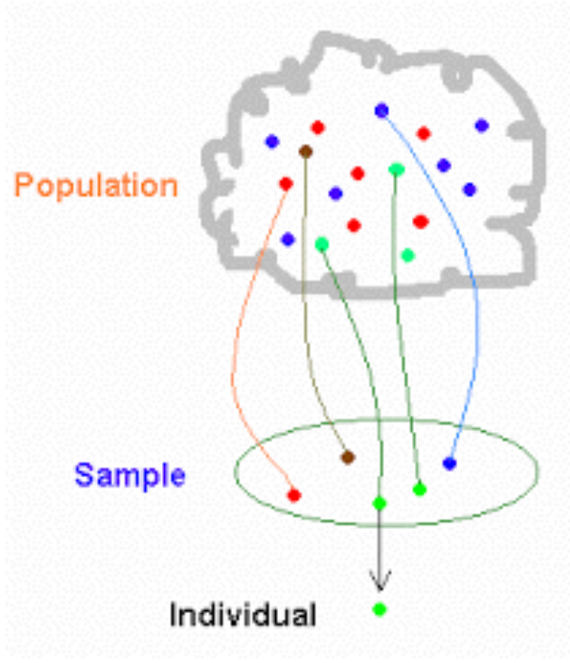
- *A*: medial focal plane of MDCK cells incubated with Texas Red-transferrin (red) and Oregon Green IgA (green).
- *B*: as in *A* but collected 1.2 µm higher.
- *C*: as in *A* but collected 2.4 µm higher.
- *D*: scatterplots of red and green pixel intensities of the top cell collected from the focal plane shown in *A* (green), *B* (blue), or *C* (red).

Image source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074624/

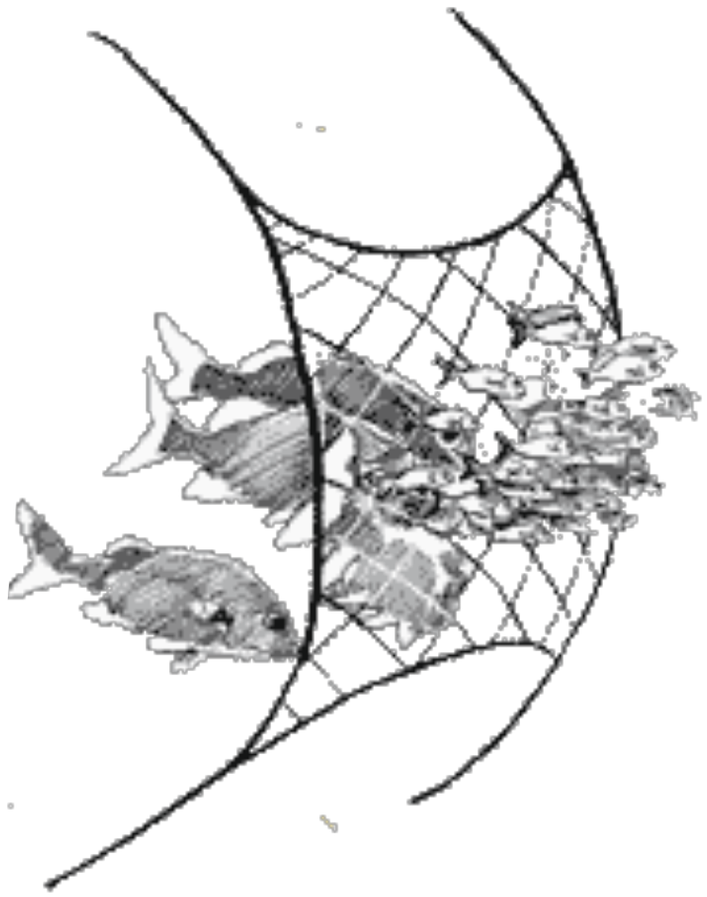
Sampling bias

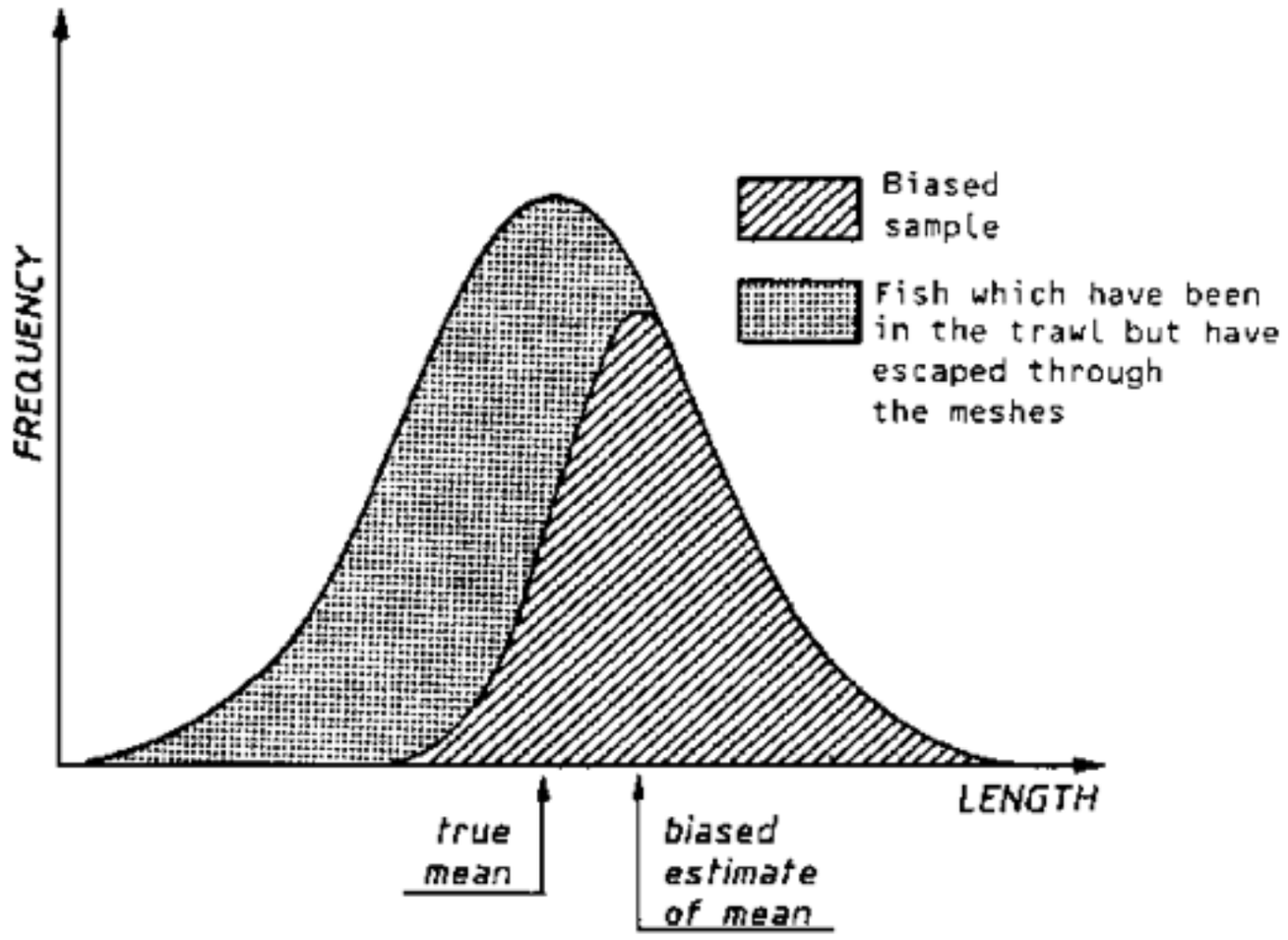
Systematic error due to non-random sample of a population, causing some members of the population to be less likely to be included than others

- Time interval (e.g. early termination of experiment)
- Selection mechanism (e.g. Chicago Tribune vs Gallup pool, 1936 USA)
- Malmquist bias (astronomy: intrinsically bright objects preference)
- Self-selection/non-response
- Symptom based
- Cavemen bias (preservation)
- Cherry picking/confirmation
- Censoring

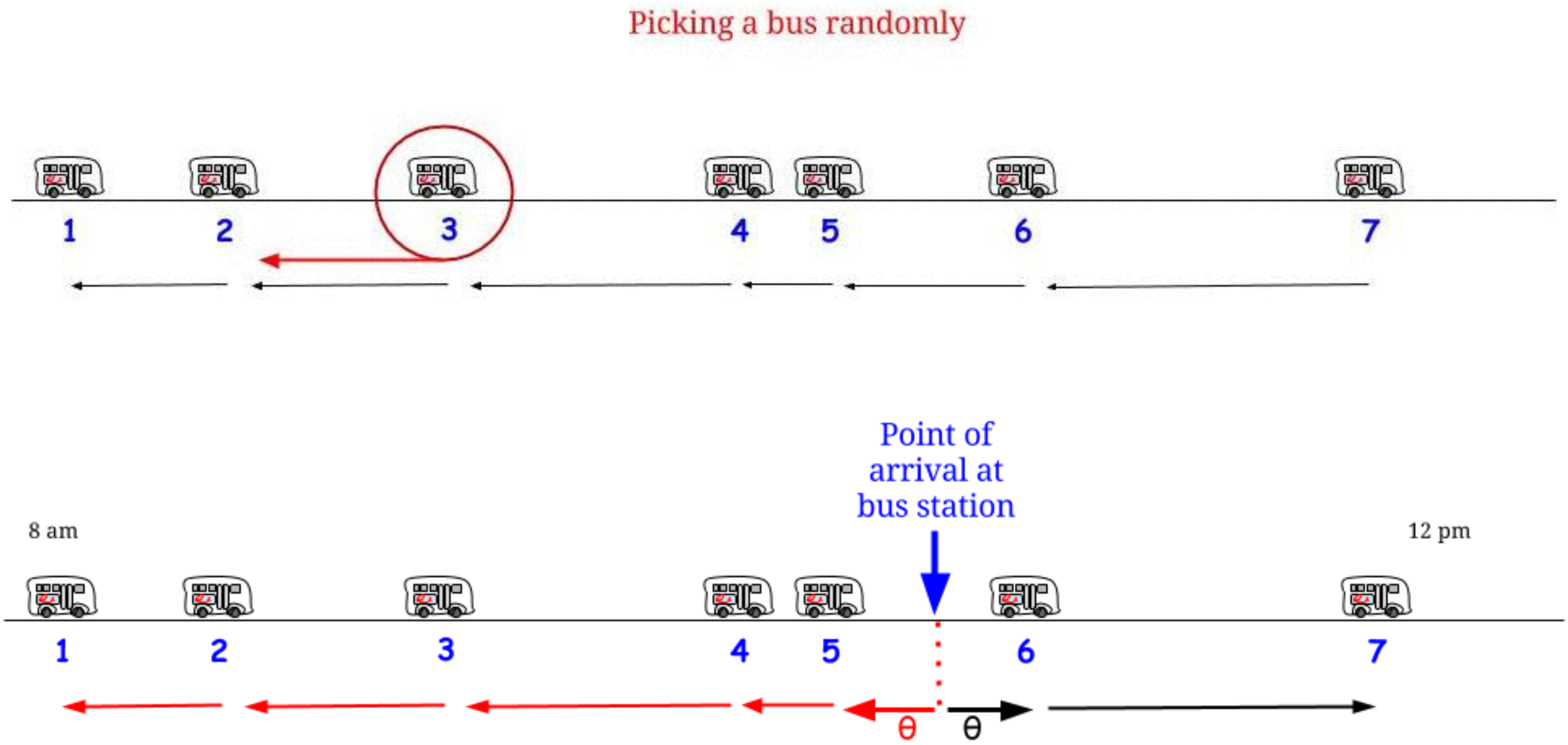


Size bias: cartoon version





Length bias: cartoon version

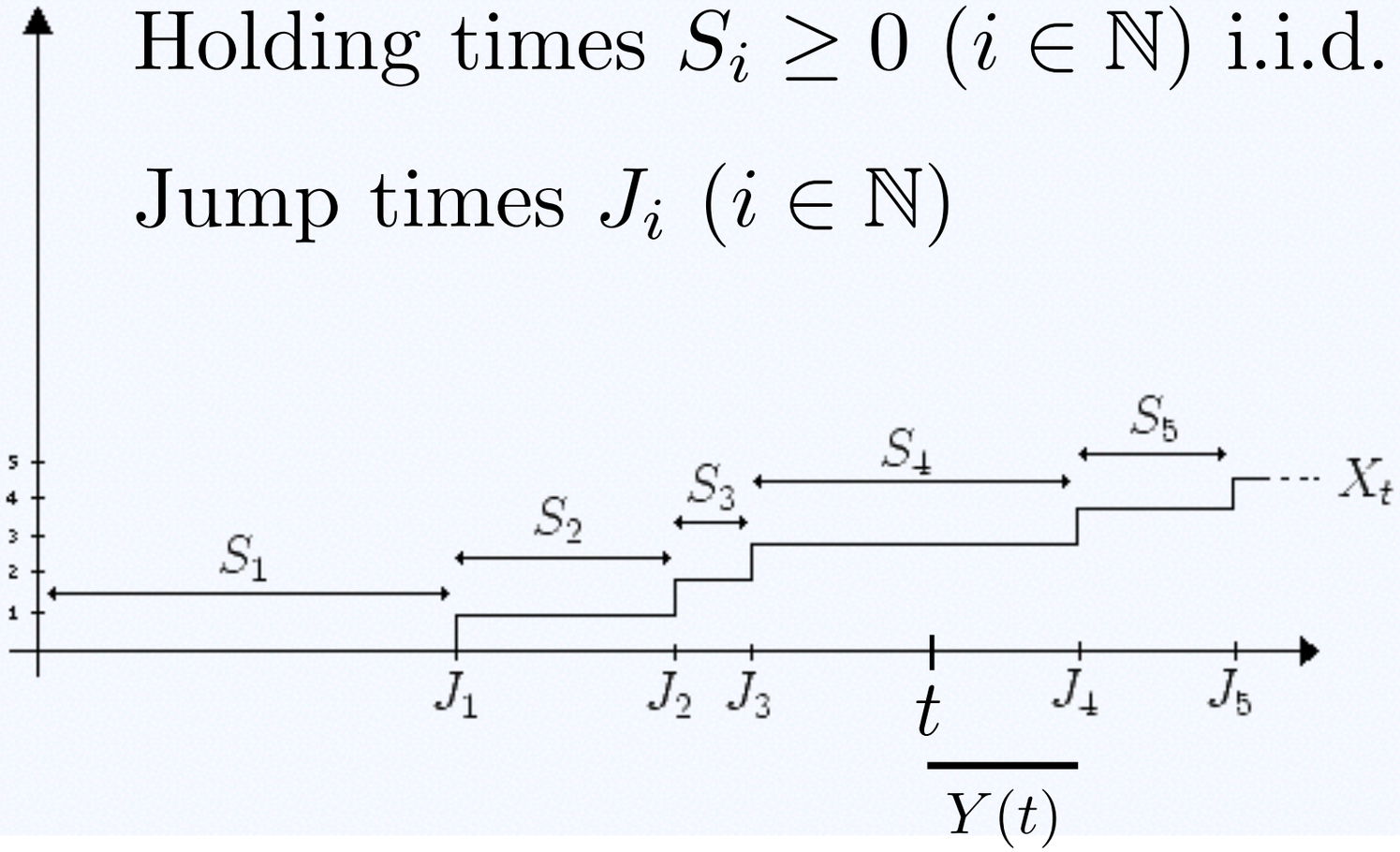


Source: https://stats.stackexchange.com/questions/122722/please-explain-the-waiting-paradox

Length bias: the maths

Holding times $S_i \geq 0$ $(i \in \mathbb{N})$ i.i.d. with CDF F

Jump times J_i $(i \in \mathbb{N})$



Renewal process $N(t) = \sup\{n : J_n \leq t\}$ $(t > 0)$

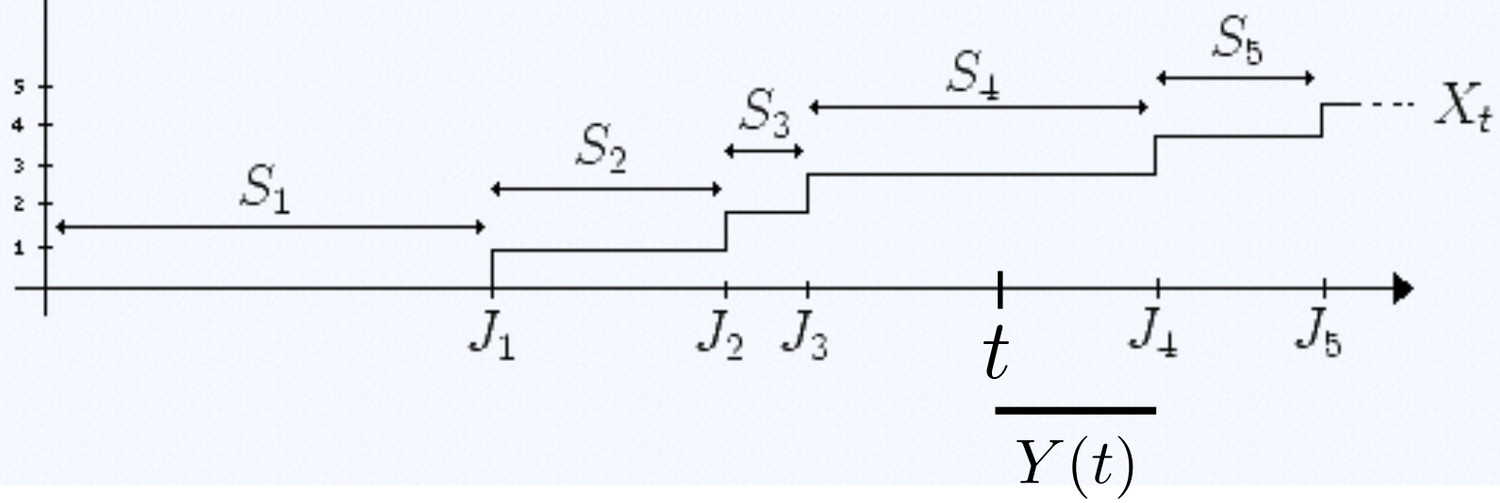
$\forall t \; \exists! \; N(t) : J_{N(t)} \leq t < J_{N(t)+1}$

Residual time $Y(t) = J_{N(t)+1} - t$

Length bias: the maths

Holding times $S_i \geq 0$ $(i \in \mathbb{N})$ i.i.d. with CDF F
mean μ, SD σ

Jump times J_i $(i \in \mathbb{N})$



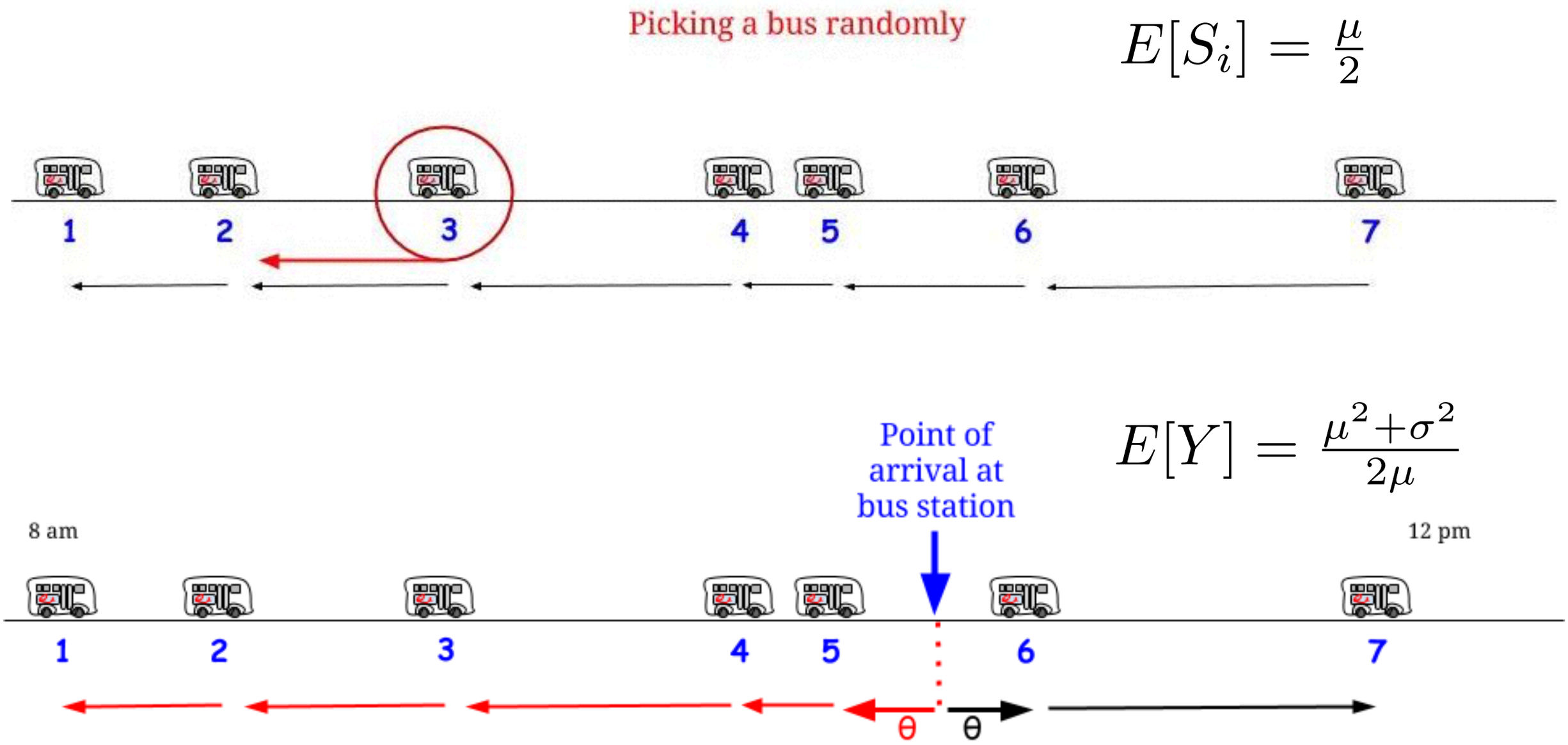
For large t, distribution of Y becomes independent of t.

$$E[Y] = \frac{\mu^2+\sigma^2}{2\mu} > \frac{\mu}{2} = E[S_i]$$

whenever $\sigma^2 > 0$

Waiting time paradox

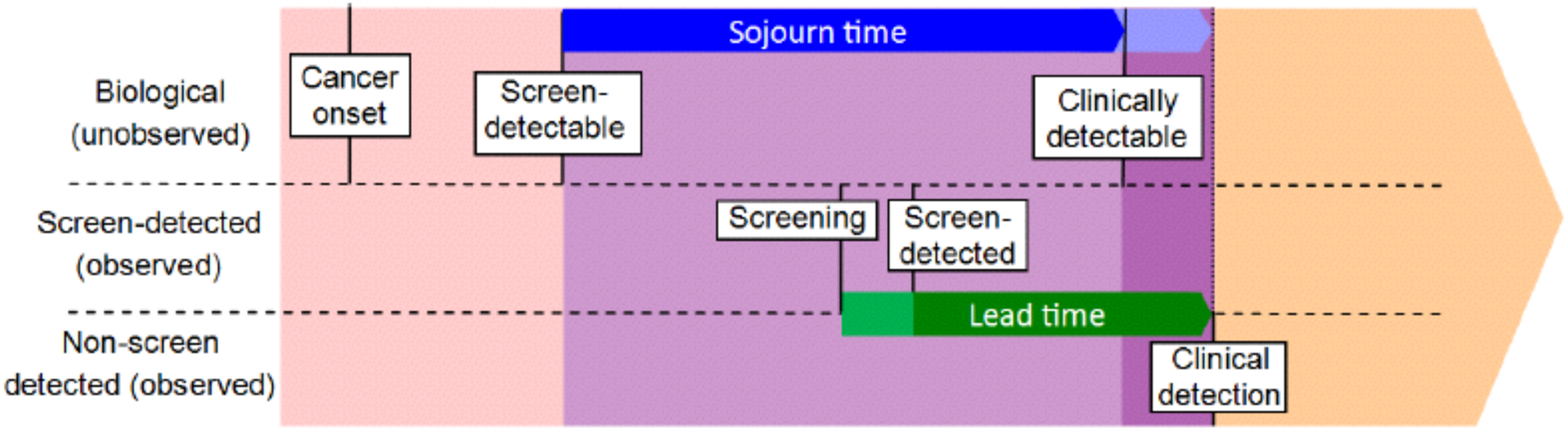
Length bias: cartoon version



Source: https://stats.stackexchange.com/questions/122722/please-explain-the-waiting-paradox

Length bias in cancer screening: Terminology

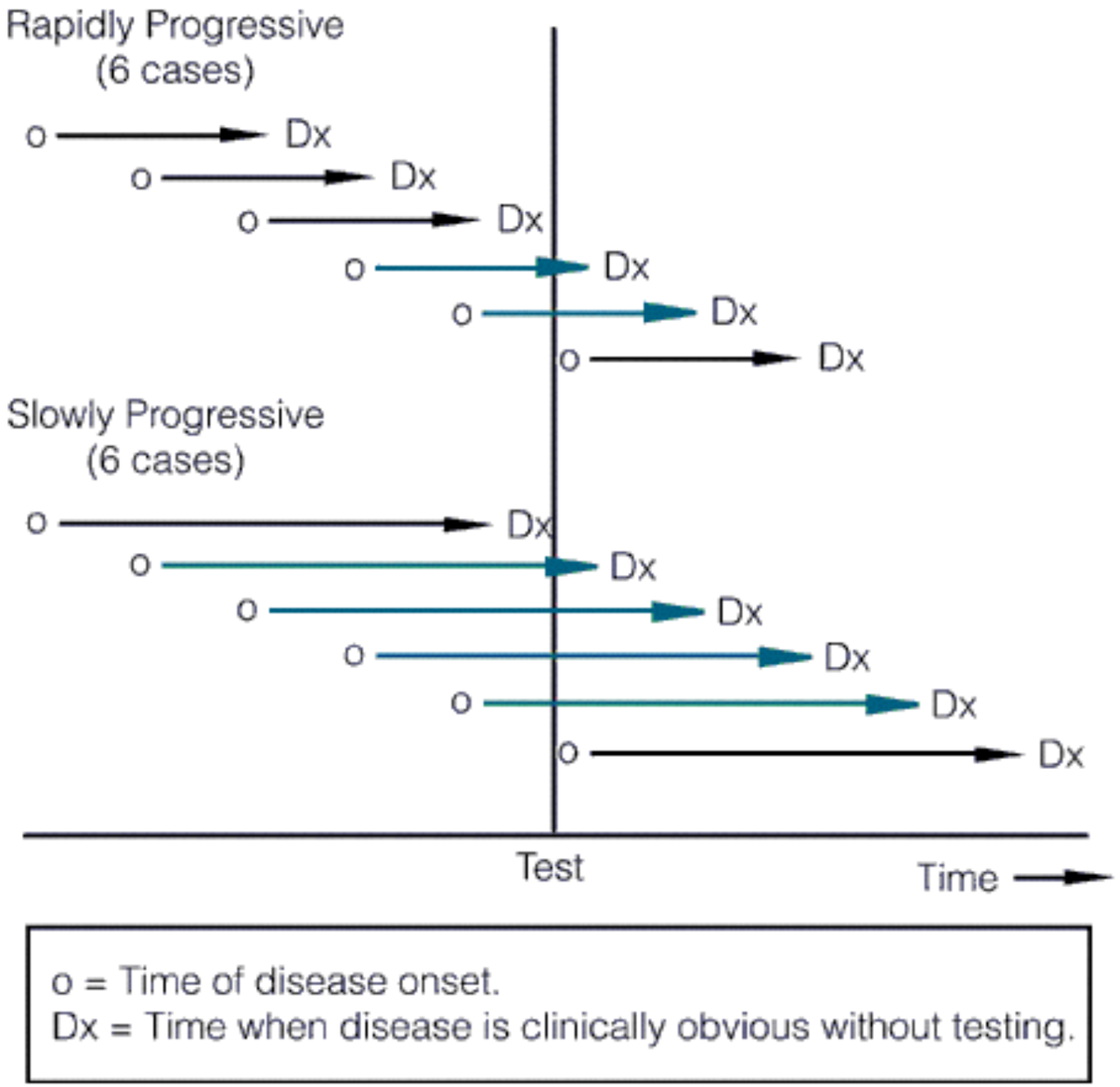
Context: Population based cancer screening programme at regular intervals



Sojourn time: Length of time between screen detectable and clinically detectable

Lead time: Length of time by which diagnosis is advanced by screening

Length bias in cancer screening:
Connection with tumour growth



Rapidly growing cancers:

Screen detected 2 out of 6

Length bias!

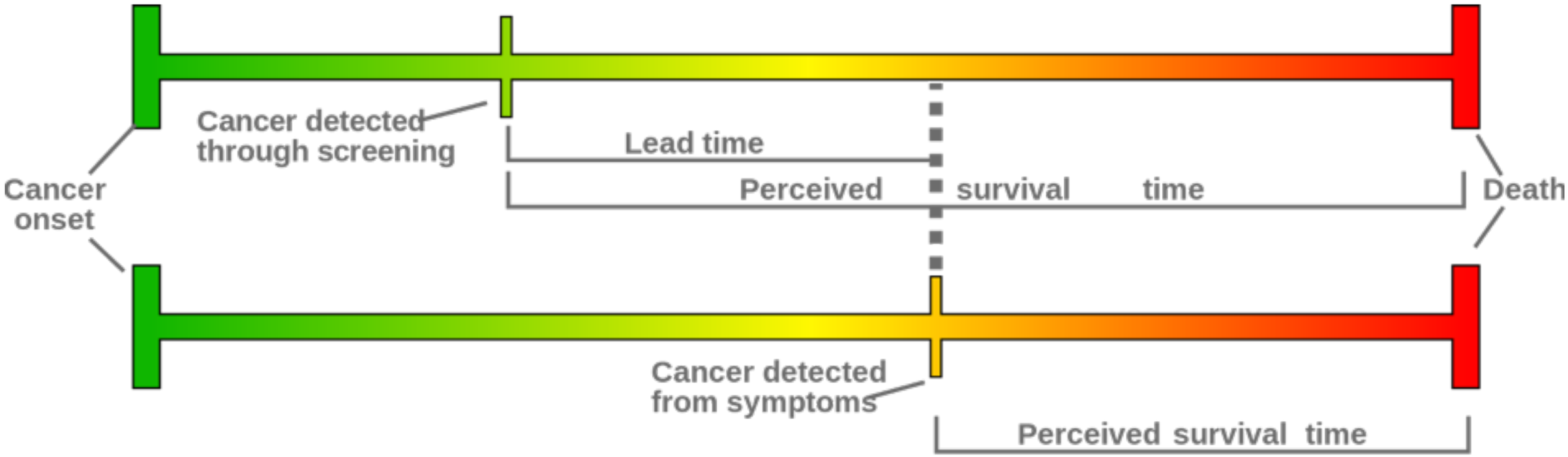
(Relative oversampling of slowly vs rapidly growing tumours)

Slowly growing cancers:

Screen detected 4 out of 6

Screening oversamples slowly growing tumours relative to rapidly growing ones.

Length bias in cancer screening:
Lead time and survival time



- Screening oversamples slowly growing tumours relative to rapidly growing ones.
- Lead time is perceived as additional survival time.
- Lead time allows early intervention, which may increase survival.
- If no early treatment option is available, early diagnosis is questionable.

Estimation tasks in cancer screening

Context: Population based cancer screening programme at regular intervals

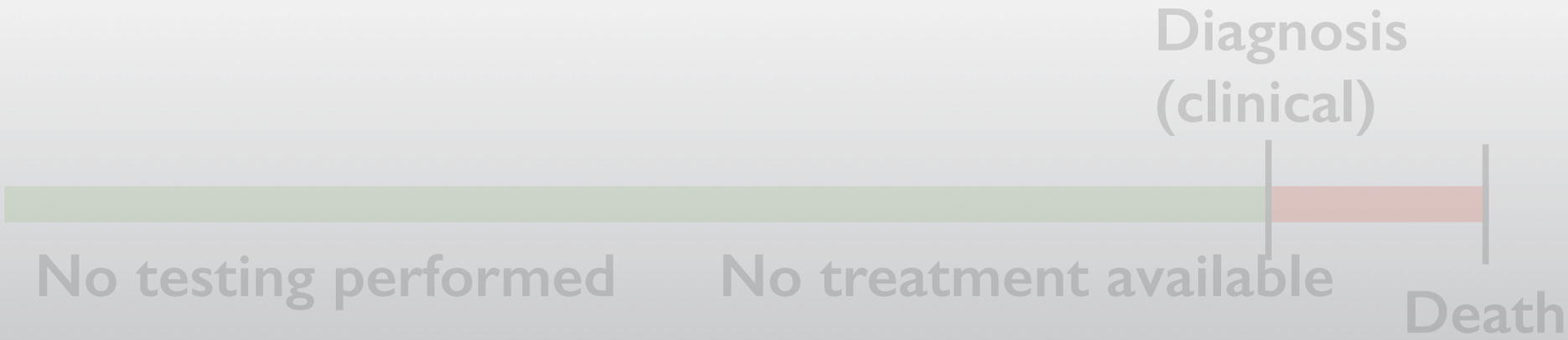
Sojourn time: Length of time between screen detectable and clinically detectable
Lead time: Length of time by which diagnosis is advanced by screening

- Both these times are unobservable
- Mathematical models for estimation
- Sojourn time: recurrence models, Markov chains under progressive assumption
- NHSBSP data: round length 3 years, women ages 50-70
- Maximum likelihood estimate: mean sojourn time=2.91 years

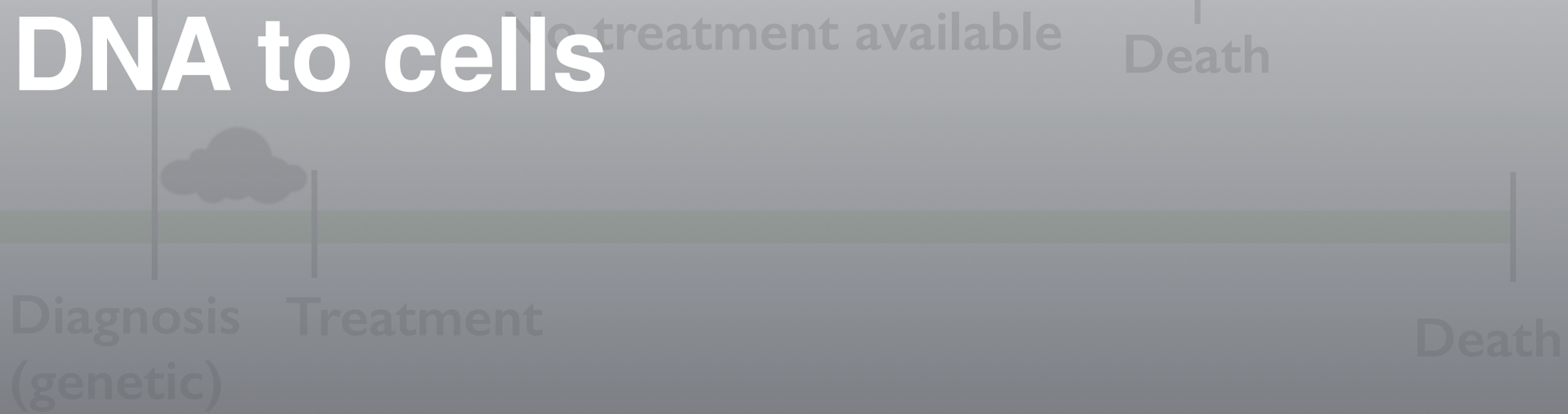
S Cheung, JL Hutton, JA Brettschneider
Review of sojourn time calculation models used in breast cancer screening
CRiSM Working Paper Series No. 17-04, 2017

S Cheung MPhil Dissertation 2016 (Supervisors: JL Hutton, JA Brettschneider)

Lead time and survival time:
Genomic testing under three scenarios

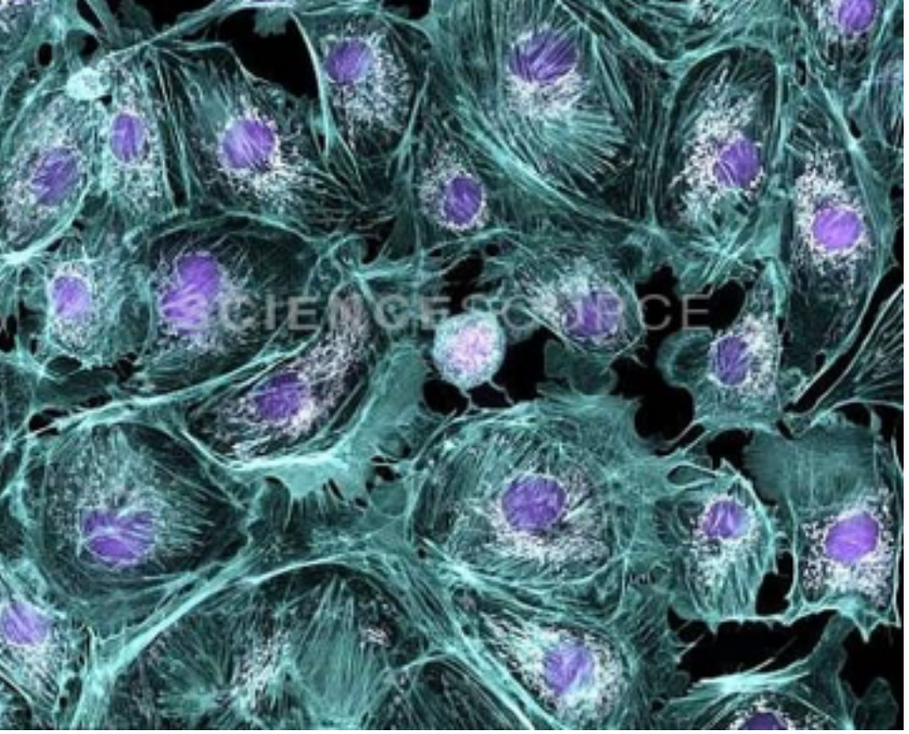


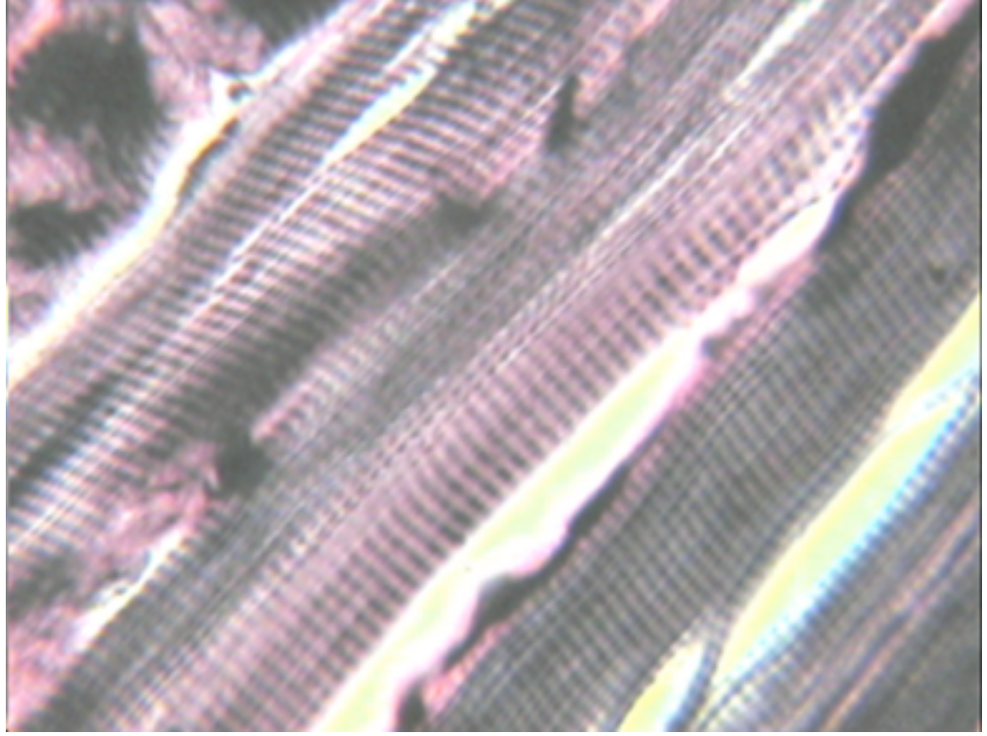
Basic molecular biology:
From DNA to cells

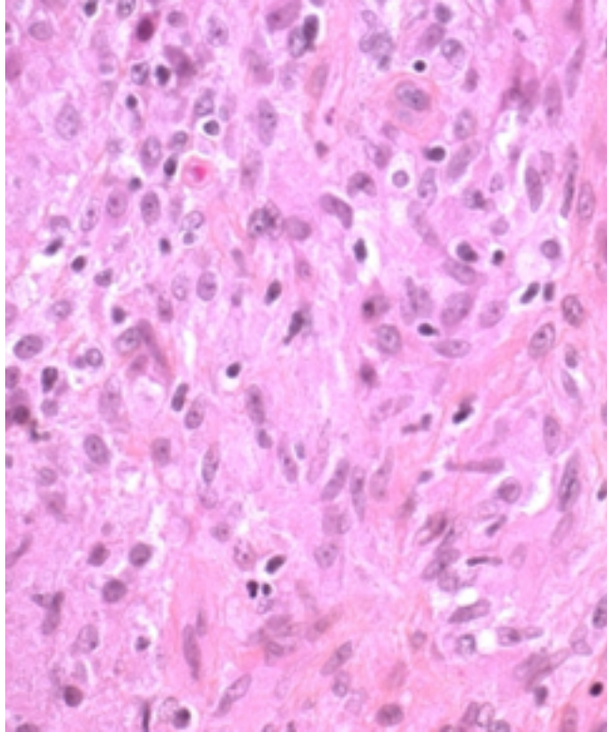


Test result may cause anxiety and apathy during lead time (could be 50 years!)

- If treatment is available, testing may increase survival time

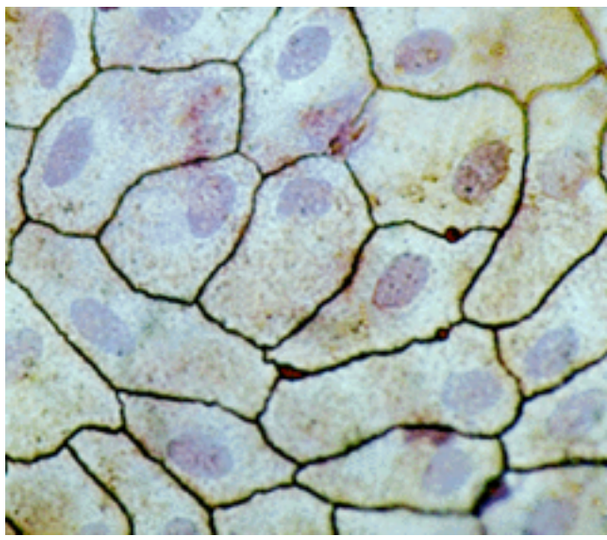


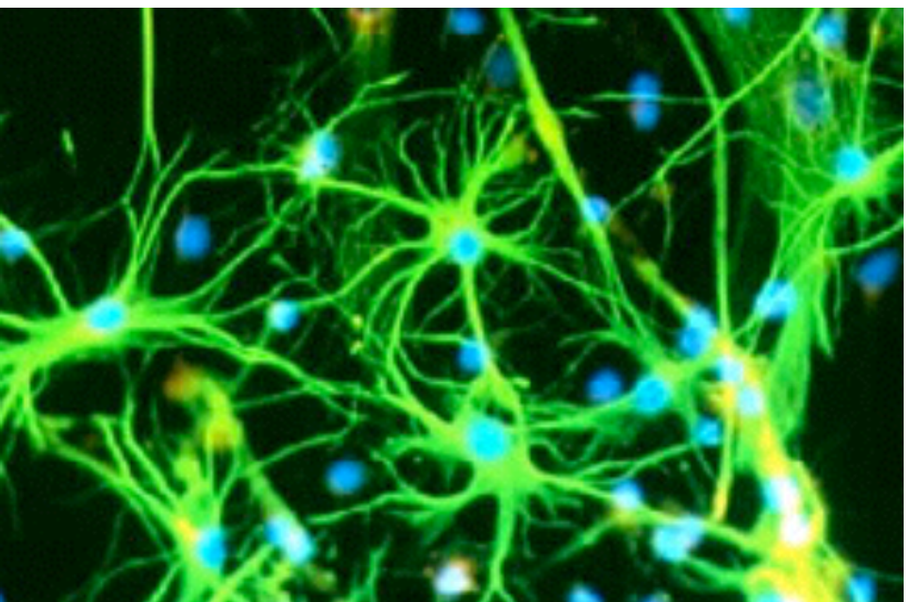


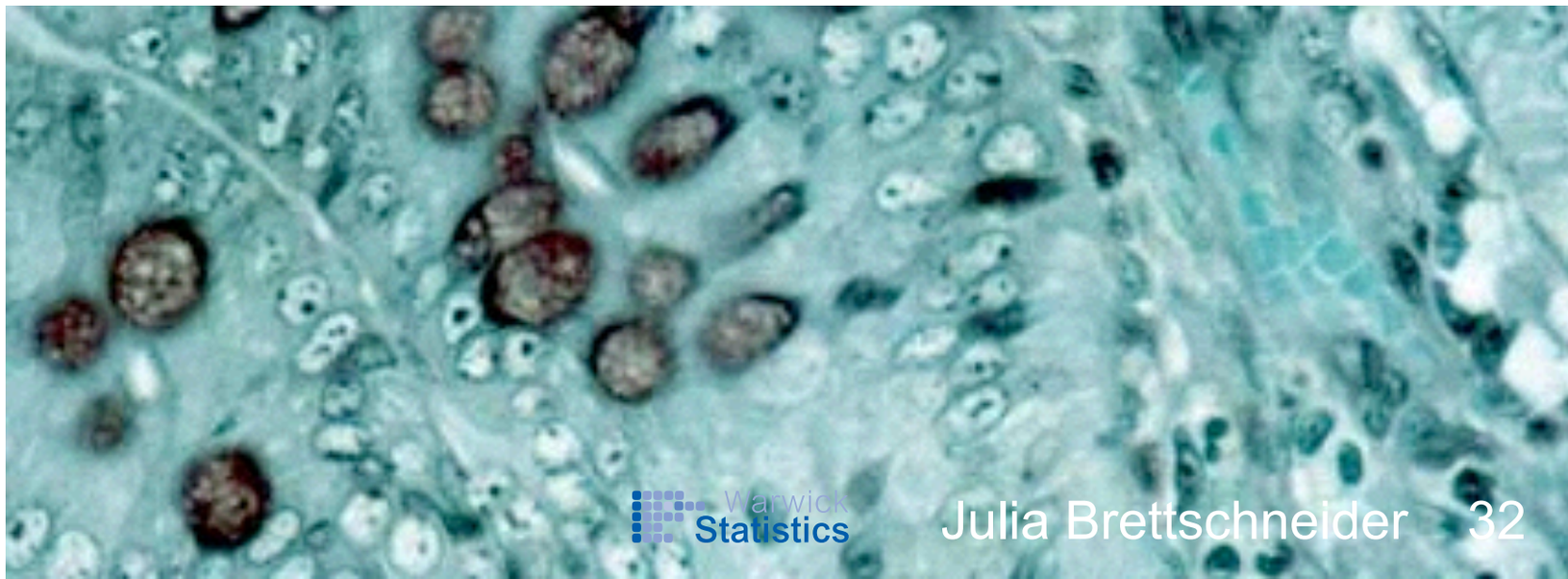


Basic molecular biology:

From DNA to cells

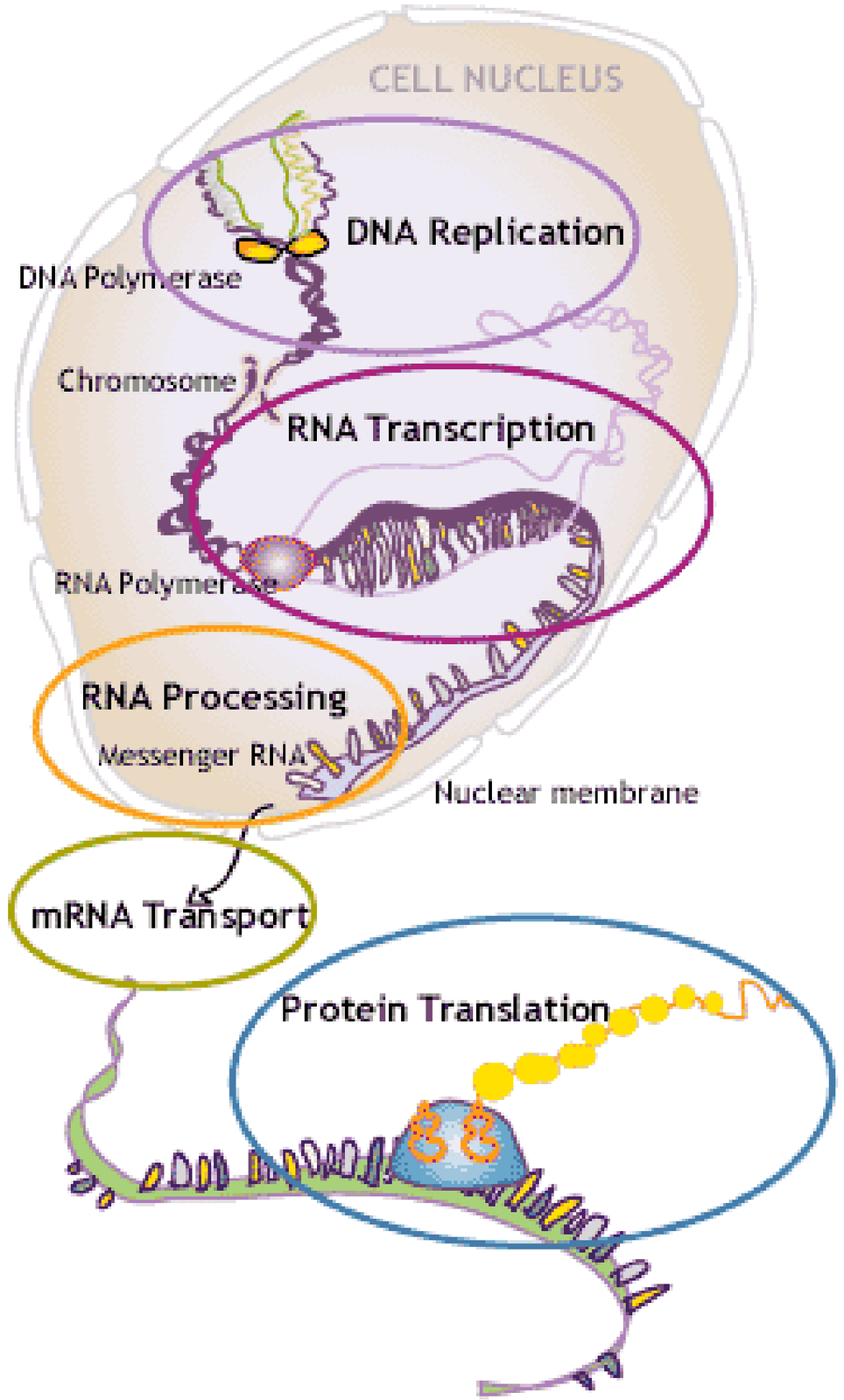




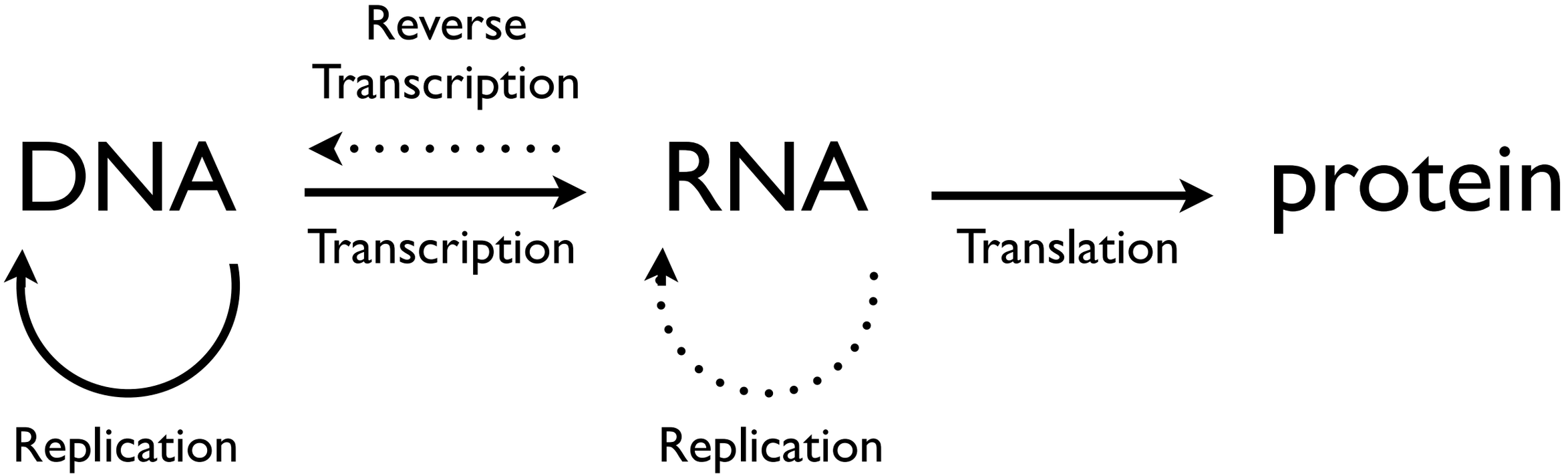


Construction of organisms *(very simplified)*

- Proteins are built according to genetic information, in a multi-step process.
- Proteins are the building blocks for cells.
- Genes are the blue print of the cells.



Biological information flow (*)



(*) very simplified

Gene expression

Gene expression =

the gene's degree of biochemical activity

(to a molecular biologist: amount of RNA

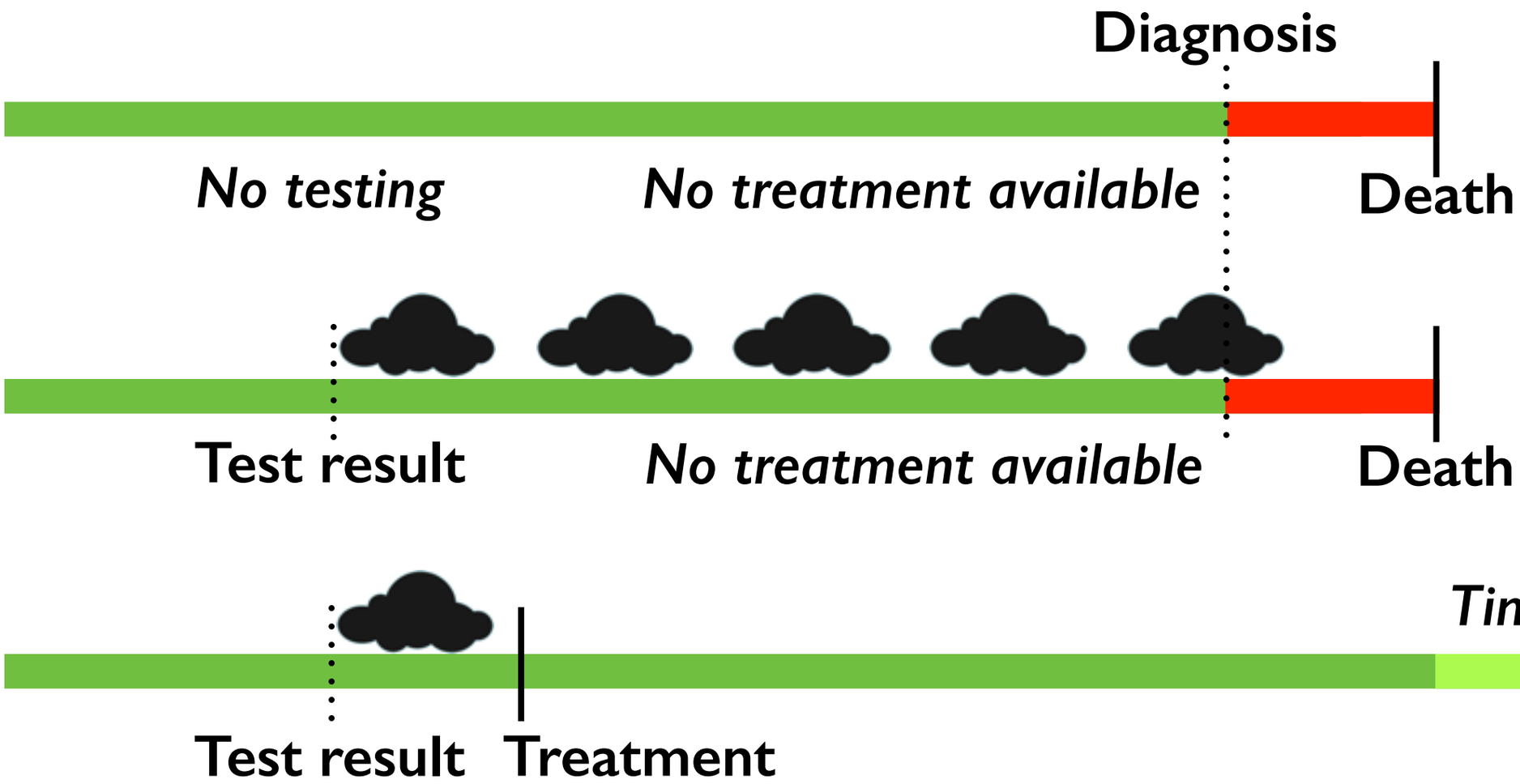
to a biochemist: amount of protein)

Depends on **factors** such as:

- Type of the cell
- State of cell (developmental, diseased/healthy etc)
- Cellulaire structure

Measuring expression levels helps understanding cellular processes, diseases, development etc.

Lead time and survival time: Genomic testing under three scenarios



- Test result may cause anxiety and apathy during lead time (could be 50 years!)
- Test result may be wrong (e.g. immature research, multiple testing)
- If correct and if treatment is available testing may increase survival time

Two tails of the City:
Trades & traders

Basics psychology:

Homo ~~economicus~~ sapiens

decisions theory

Normative theory versus descriptive theory

Normative theories of decision making:

- How idealised (rational) world behave when taking decisions
- Based on an "idealised" form of human being:
  e.g. *homo economics* (rational utility maximiser), *homme moyen*
- Methods: Mathematical axioms and optimisation

Descriptive theories of decision making

- How *people actually make* decisions
- Based on observation (empirical studies)
- Methods: empirical studies, revised models

Why is normative theory not enough?

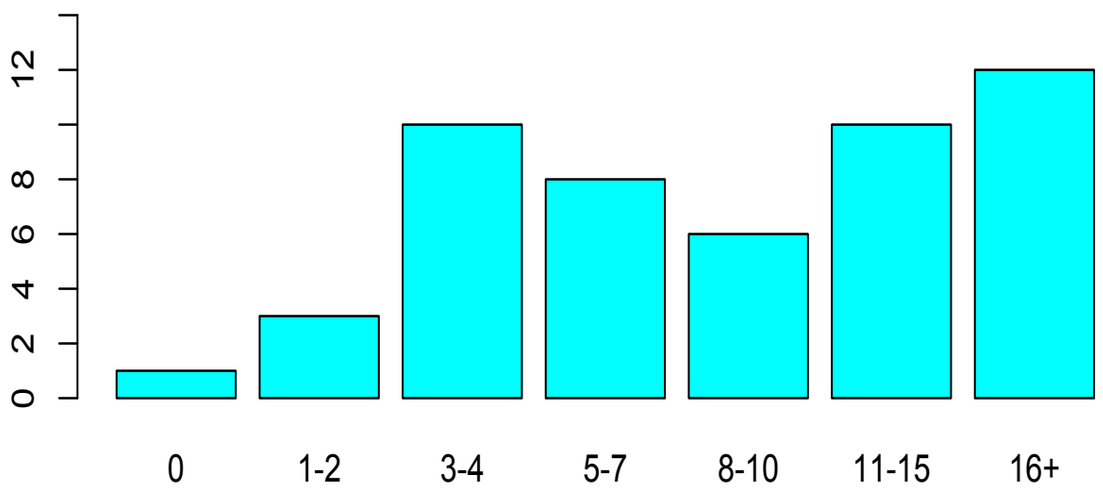
Empirical studies have demonstrated that people do not always follow the axioms of probability (biases, fallacies, heuristics).

Empirically shown deviations from normative theory

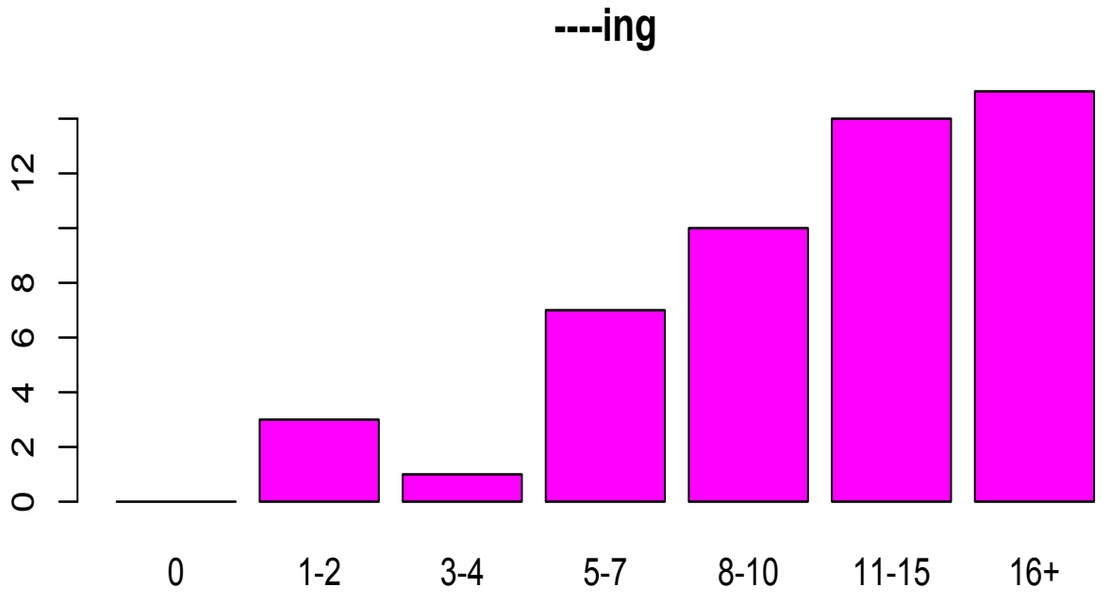
- Gambler's fallacy, inverse gambler's fallacy, belief in hot hand
- Random sequences generation biases (starting value, runs)
- Clustering illusion
- Certainty effect (Allais paradox)
- Anchoring bias (with related and unrelated information)
- Framing effect (Kahneman & Tversky)
- Availability bias (Kahneman & Tversky)
- Conjunction fallacy / "Linda problem" (Kahneman & Tversky)
- Disjunction effect (Shafir & Tversky)
- Base rate neglect
- Disposition effect (Odean, Weber & Camerer)

Example: Availability bias

----n- group 51 students



----ing group 49 students



The *less* restrictive condition creates fewer words!

Violates normative rules of probability:

For A subset of B,

P(A) < P(B)

Reason: Increased efficiency of memory search in "- - - ing"-condition

Heuristics & biases

Heuristic strategies:

- Shortcuts and approximations, especially if correct answers are not known or would take too long to construct.
- Using heuristics is useful and necessary, but may lead to biases in judgement.

Context here: Probability (risk & prospect)
Perception, estimation and judgement

Resulting behavioural bias:

Differences arising from the use of heuristics or other non-normative strategies.

Empirical study of selling behaviour

EUT would suggest a threshold strategy, i.e. stop at random time

$$S(b, X) = min\{t \geq 0 : X_t \geq\}$$

where X is current price level and b is an e*x ante* optimal threshold.

Individual trader data from LDB (Odean, 1998):

- Discount brokerage house use by primarily non-professional investors
- Trading activities of 78,000 American individual households
- Period 1991-1996
- Characteristics of individuals

JA Brettschneider, M Burgess
Using a frailty model to measure the effect of covariates on the disposition effect
CRiSM Working Paper Series No. 17-05, 2017

Giovanni Burro, PhD project (Supervised by JA Brettschneider, Vicky Henderson)

Two tails of the City: Trades & traders

Two alternative choices for observational unit.

- **Trades' perspective:** round trips
- **Trader's perspective:** series of round trips

Do the two perspectives lead to different results?
Answer depends on the study question.

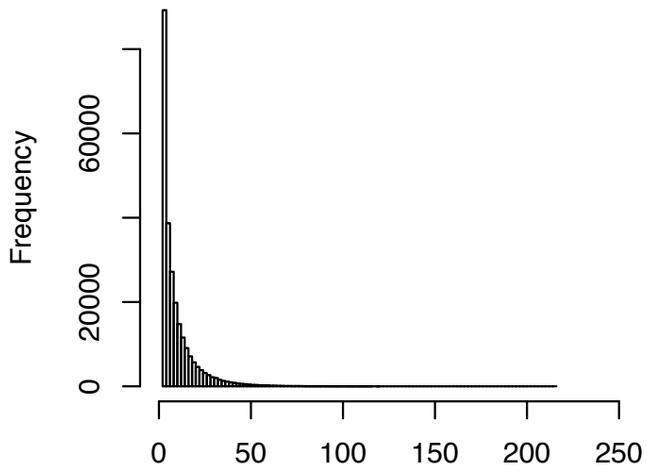
In the trade's perspective, faster traders are oversampled. Hence the distribution of observations will be dominated by these oversampled fast traders. If the outcome variable is associated with trading speed then this oversampling can create a bias.

This can be overcome by using the average of a series of roundtrips

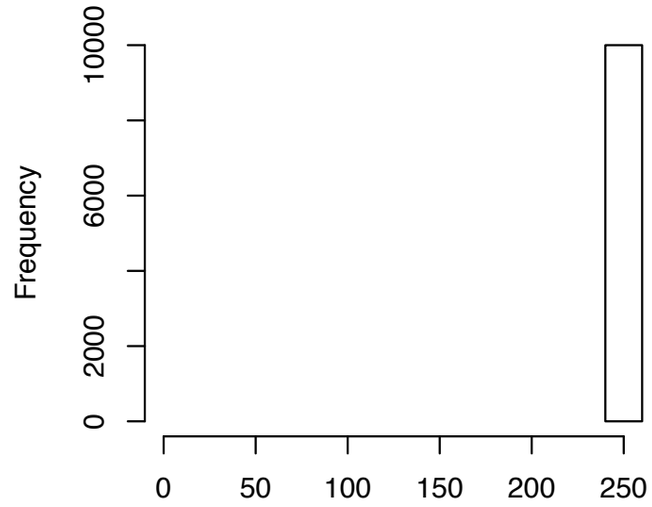
Simulated data:

Stopping times

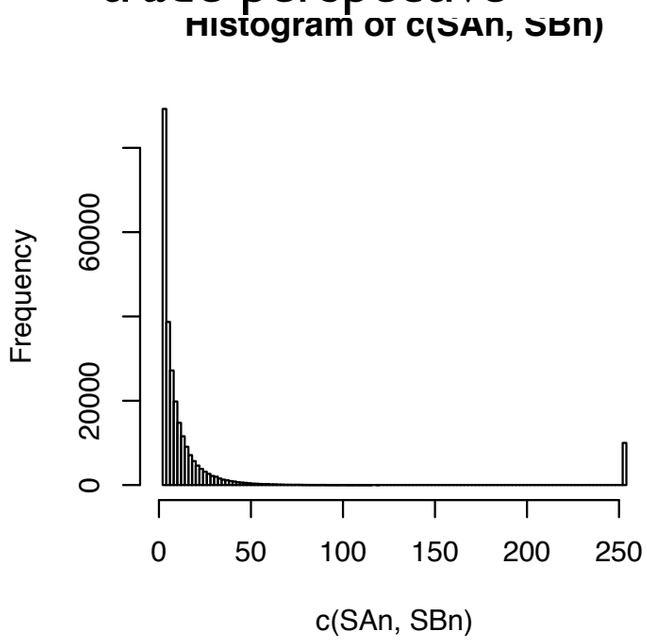
threshold (fast) group



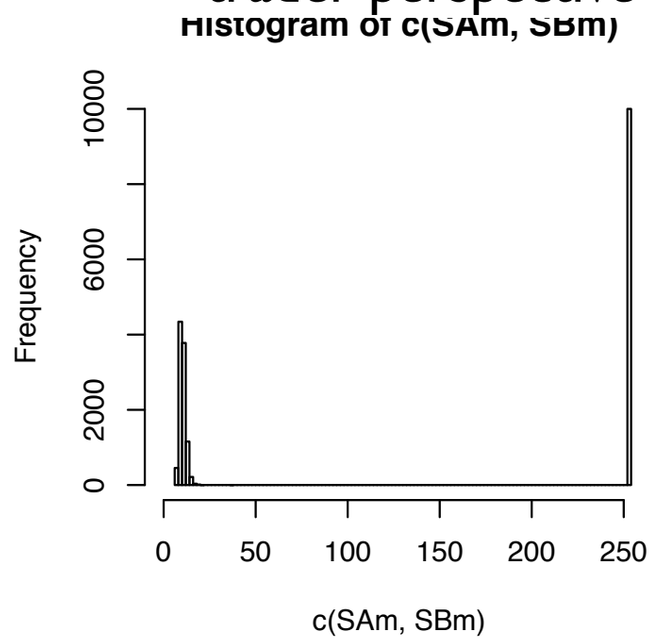
deterministic (slow) group



both groups - trade perspective

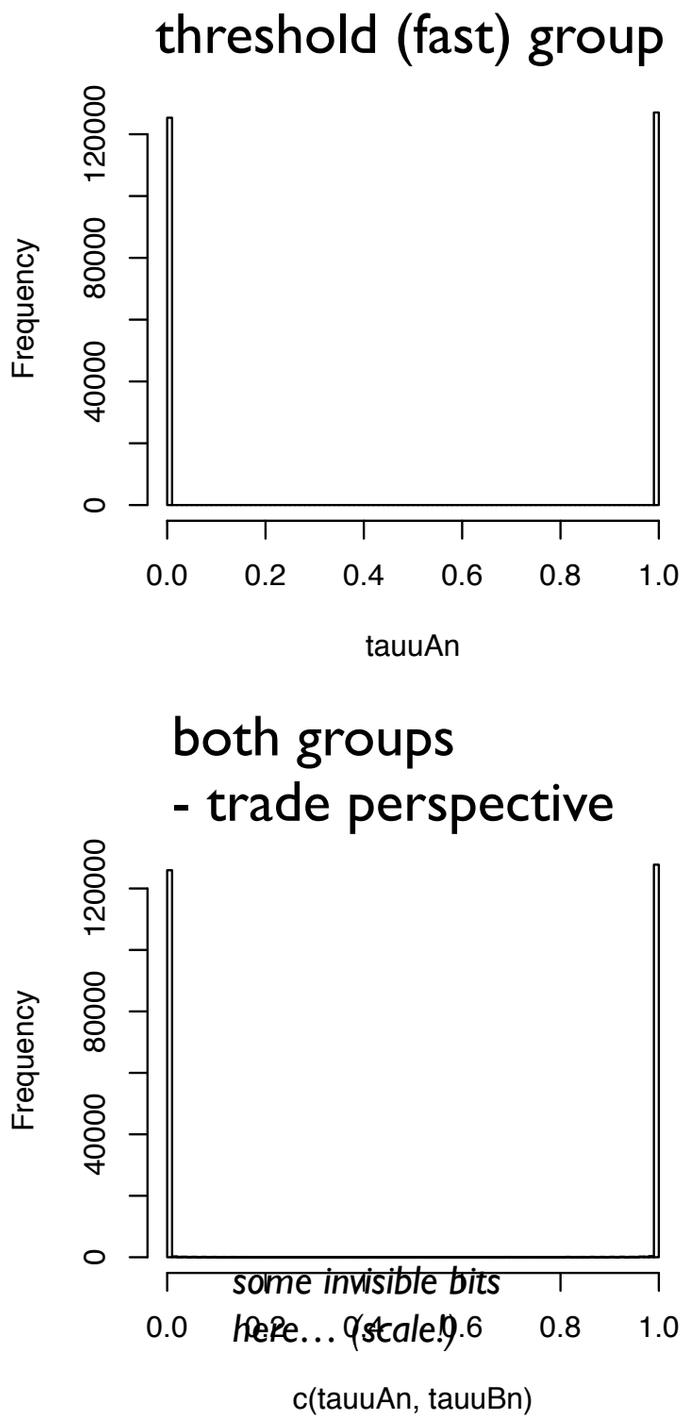


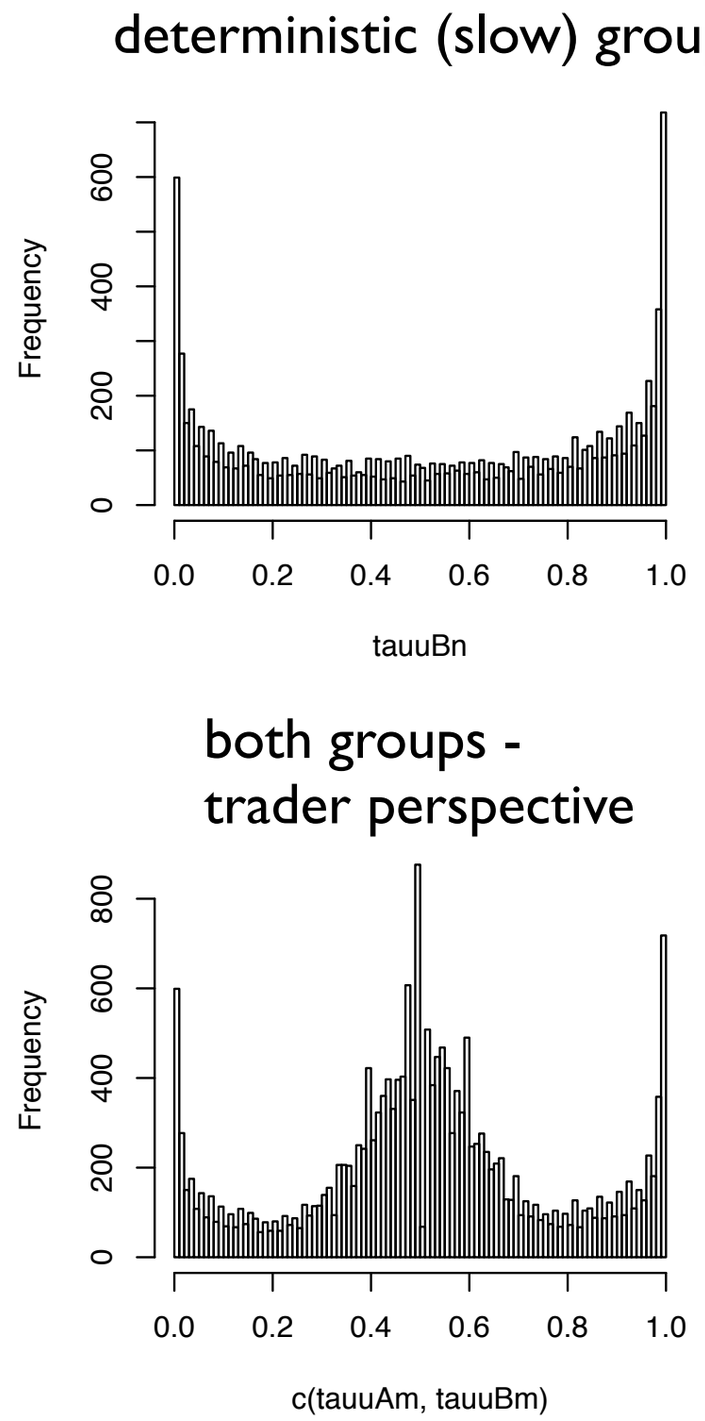
both groups - trader perspective



Simulated data:

Relative #days below selling price





Further applications

- Quality assessment of items
- Quality of data in crowd sourced repositories
- X-ray detector damage monitoring

Thanks!