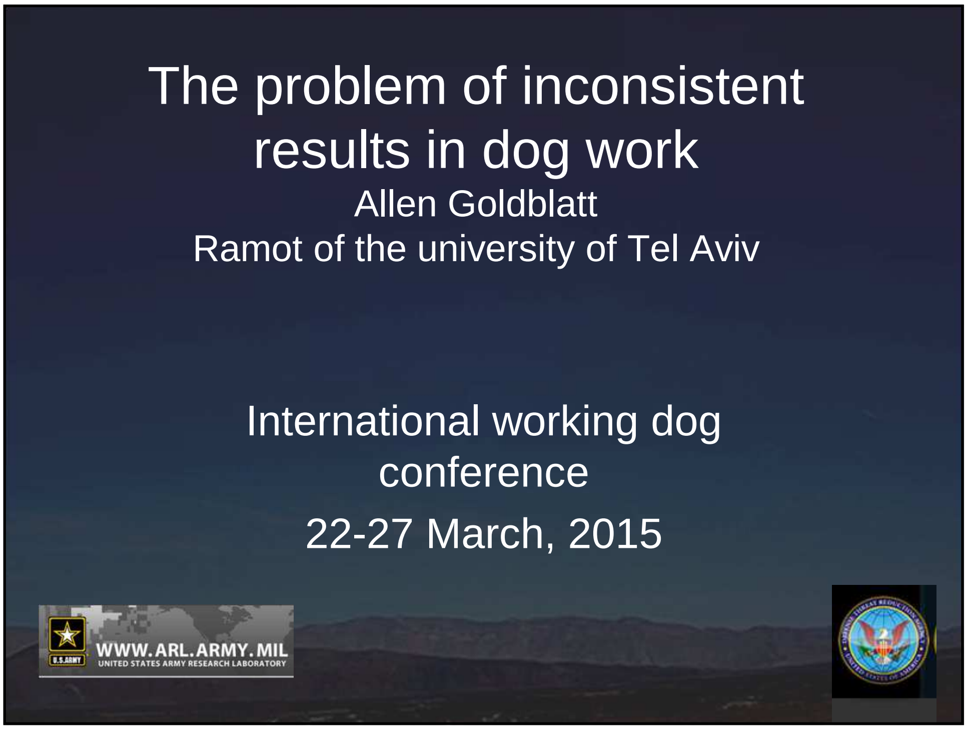


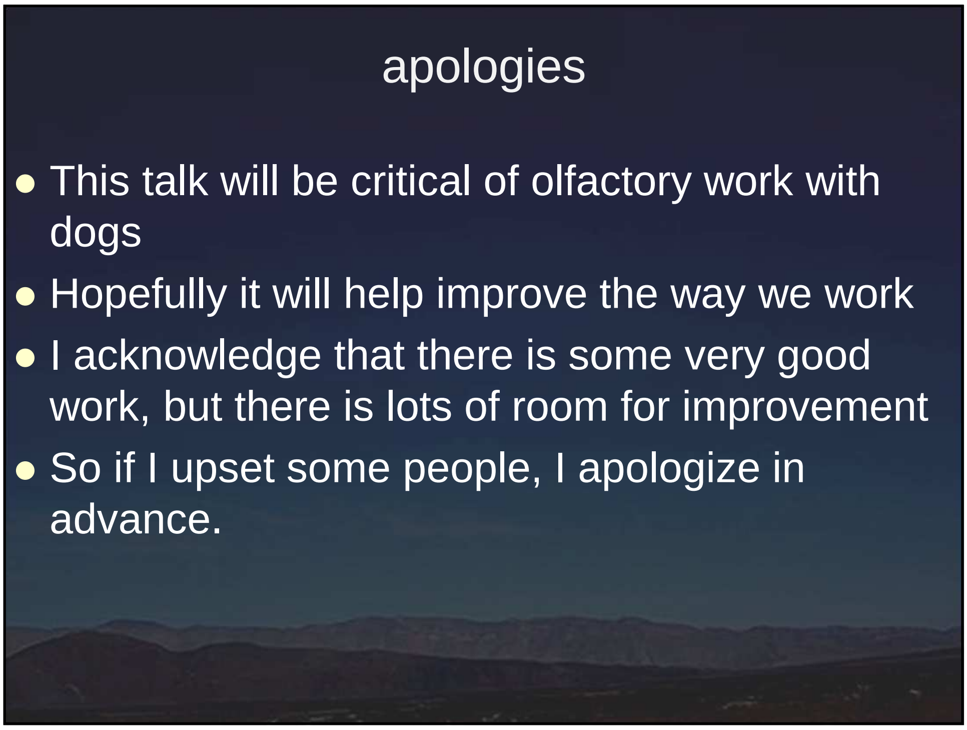
The problem of inconsistent results in dog work

Allen Goldblatt
Ramot of the university of Tel Aviv

International working dog conference

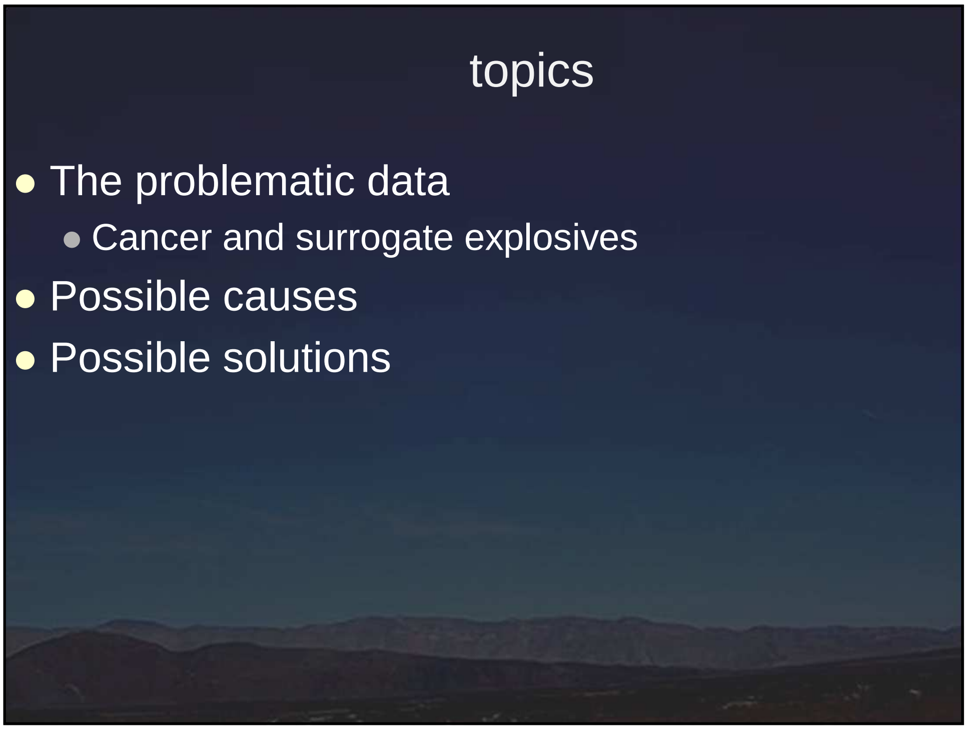
22-27 March, 2015

WWW.ARL.ARMY.MIL
UNITED STATES ARMY RESEARCH LABORATORY



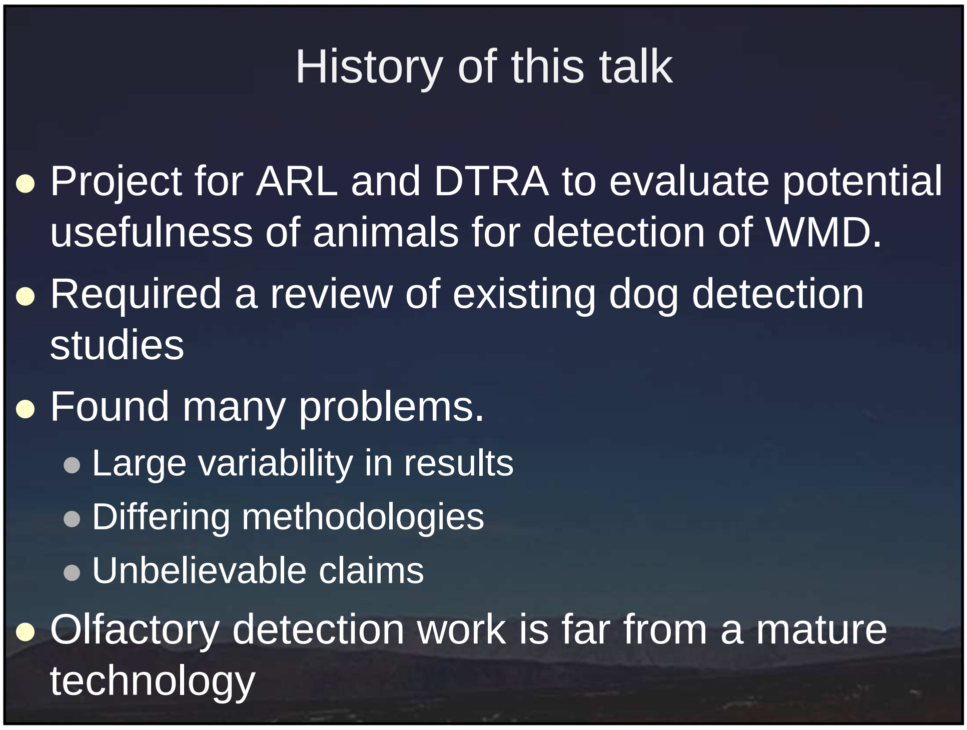
apologies

- This talk will be critical of olfactory work with dogs
- Hopefully it will help improve the way we work
- I acknowledge that there is some very good work, but there is lots of room for improvement
- So if I upset some people, I apologize in advance.



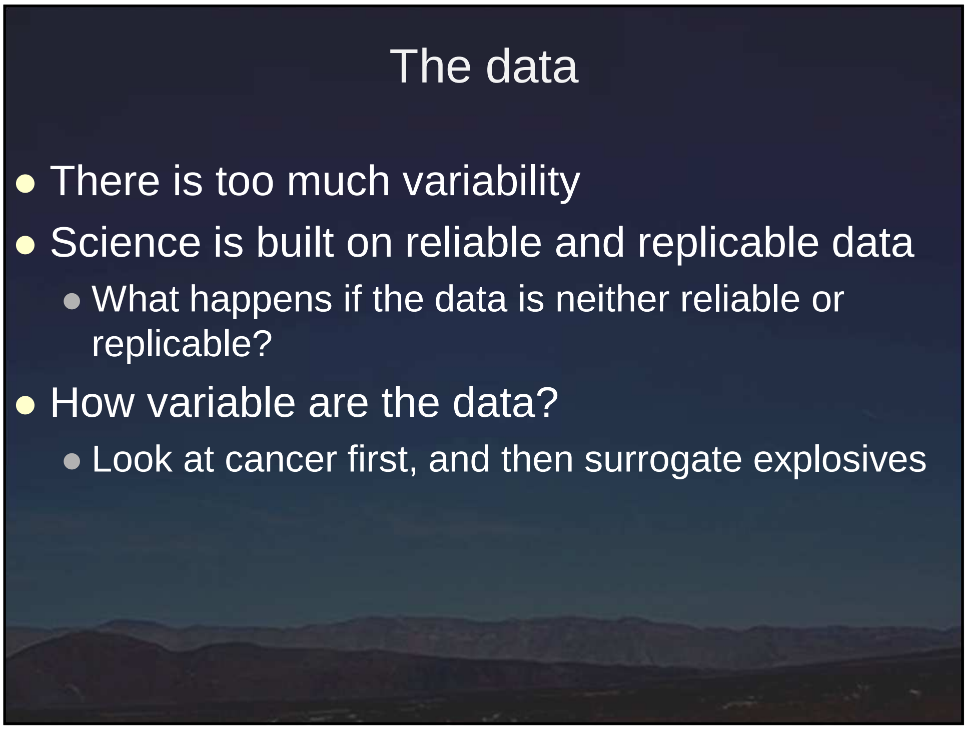
topics

- The problematic data
  - Cancer and surrogate explosives
- Possible causes
- Possible solutions



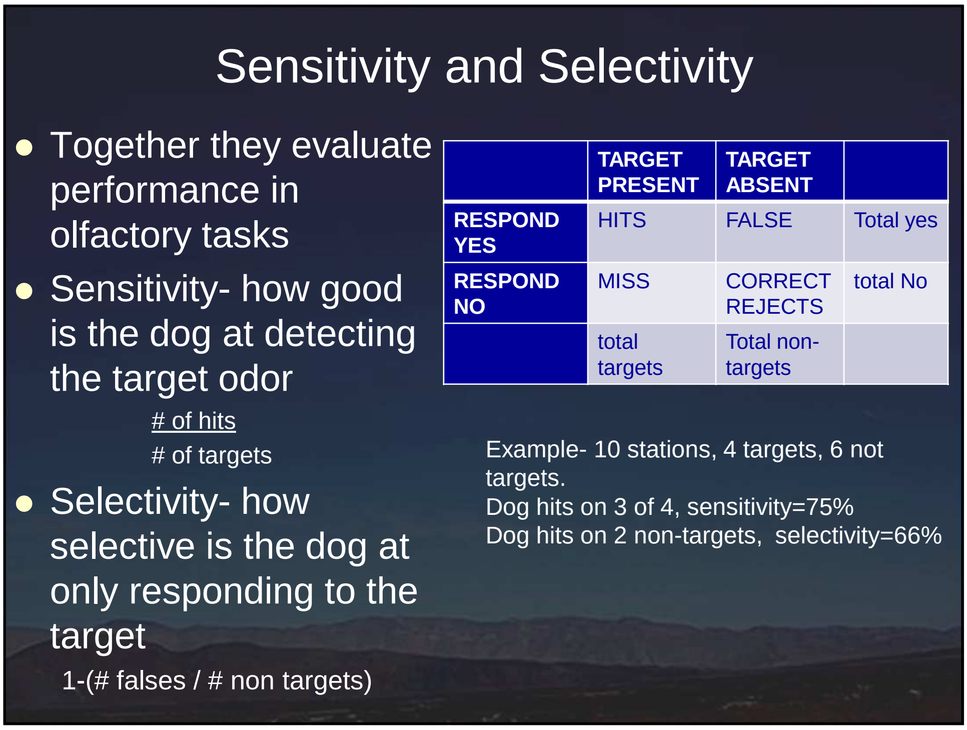
History of this talk

- Project for ARL and DTRA to evaluate potential usefulness of animals for detection of WMD.
- Required a review of existing dog detection studies
- Found many problems.
  - Large variability in results
  - Differing methodologies
  - Unbelievable claims
- Olfactory detection work is far from a mature technology



The data

- There is too much variability
- Science is built on reliable and replicable data
  - What happens if the data is neither reliable or replicable?
- How variable are the data?
  - Look at cancer first, and then surrogate explosives



Sensitivity and Selectivity

- Together they evaluate performance in olfactory tasks
- Sensitivity- how good is the dog at detecting the target odor

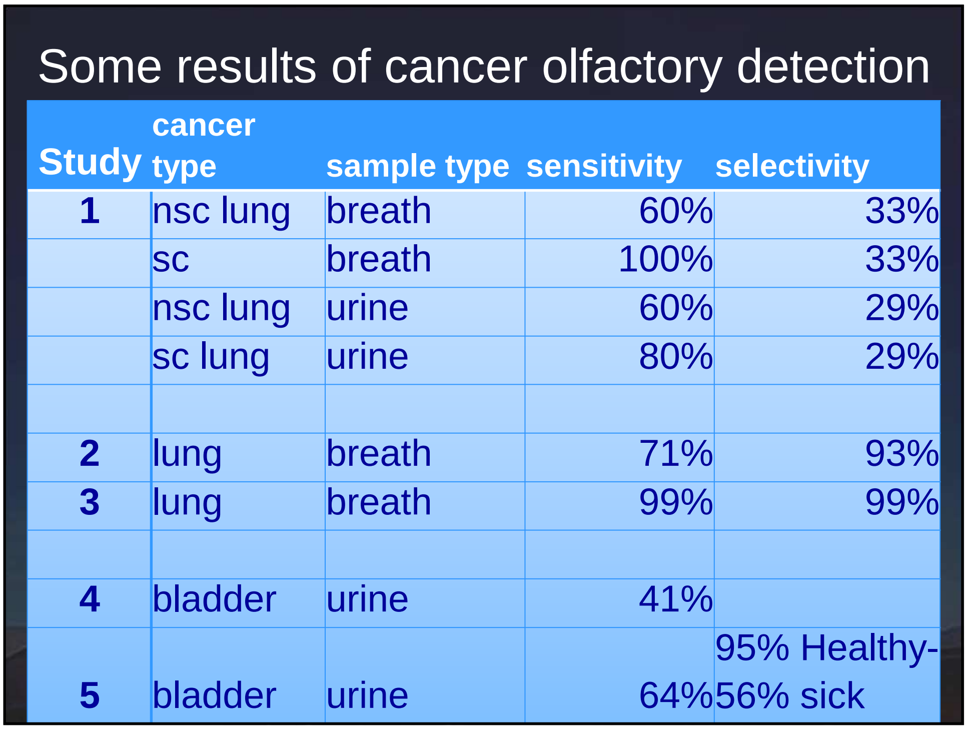
  # of hits

  # of targets
- Selectivity- how selective is the dog at only responding to the target

  1-(# falses / # non targets)

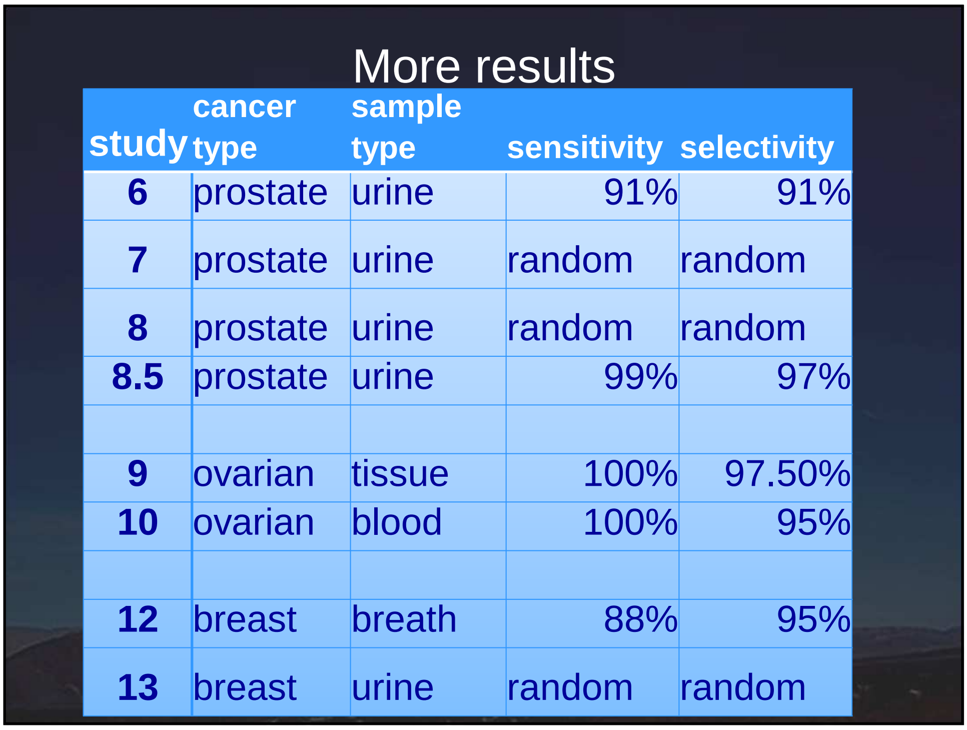
	TARGET PRESENT	TARGET ABSENT	
RESPOND YES	HITS	FALSE	Total yes
RESPOND NO	MISS	CORRECT REJECTS	total No
	total targets	Total non-targets	

Example- 10 stations, 4 targets, 6 not targets.
Dog hits on 3 of 4, sensitivity=75%
Dog hits on 2 non-targets, selectivity=66%



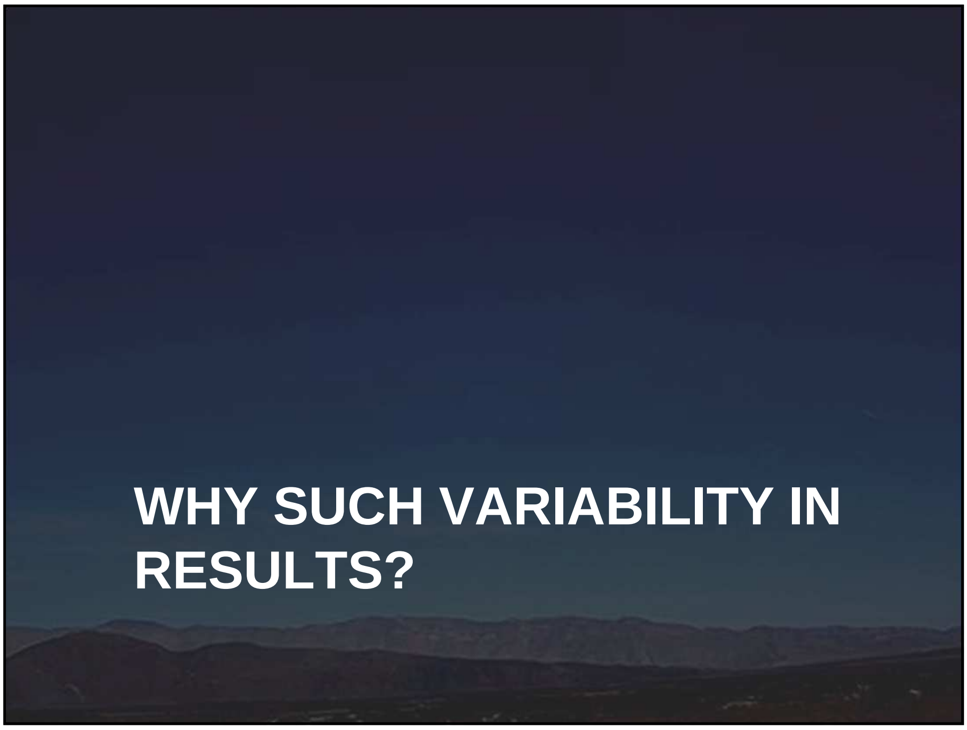
Some results of cancer olfactory detection

Study	cancer type	sample type	sensitivity	selectivity
1	nsc lung	breath	60%	33%
	sc	breath	100%	33%
	nsc lung	urine	60%	29%
	sc lung	urine	80%	29%
2	lung	breath	71%	93%
3	lung	breath	99%	99%
4	bladder	urine	41%	
5	bladder	urine	64%	95% Healthy- 56% sick

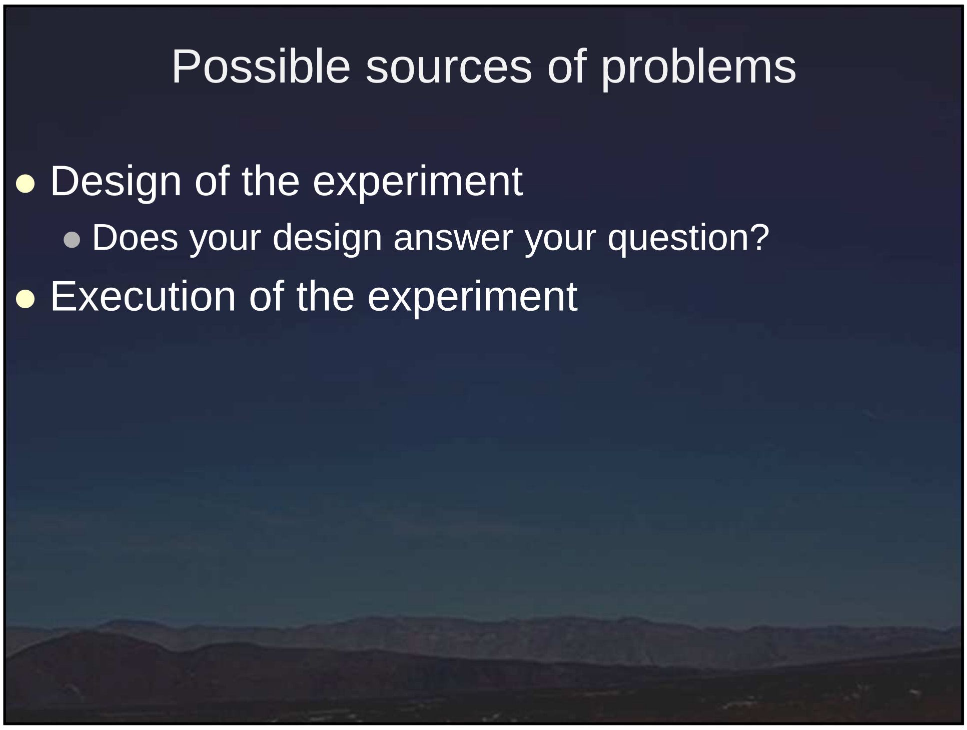


More results

study	cancer type	sample type	sensitivity	selectivity
6	prostate	urine	91%	91%
7	prostate	urine	random	random
8	prostate	urine	random	random
8.5	prostate	urine	99%	97%
9	ovarian	tissue	100%	97.50%
10	ovarian	blood	100%	95%
12	breast	breath	88%	95%
13	breast	urine	random	random

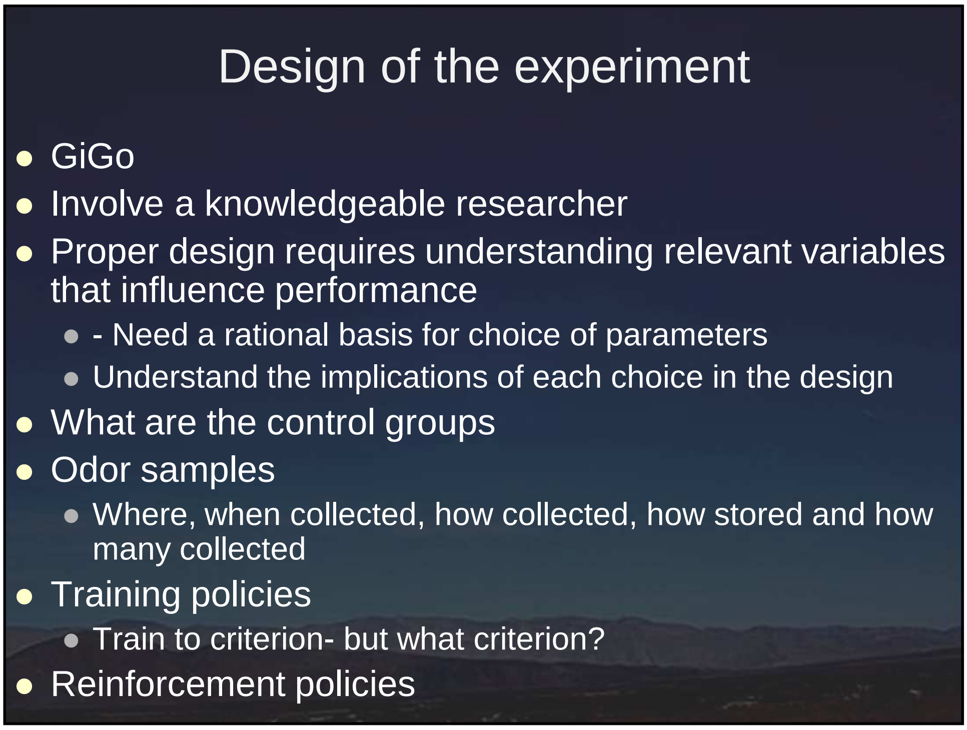


WHY SUCH VARIABILITY IN RESULTS?



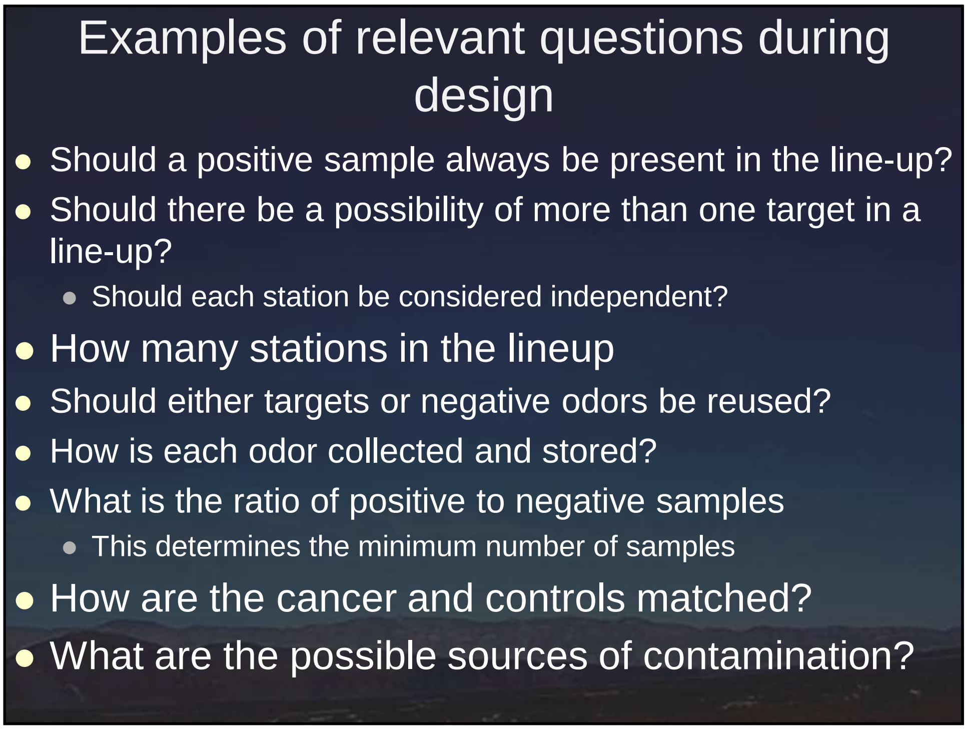
Possible sources of problems

- Design of the experiment
  - Does your design answer your question?
- Execution of the experiment



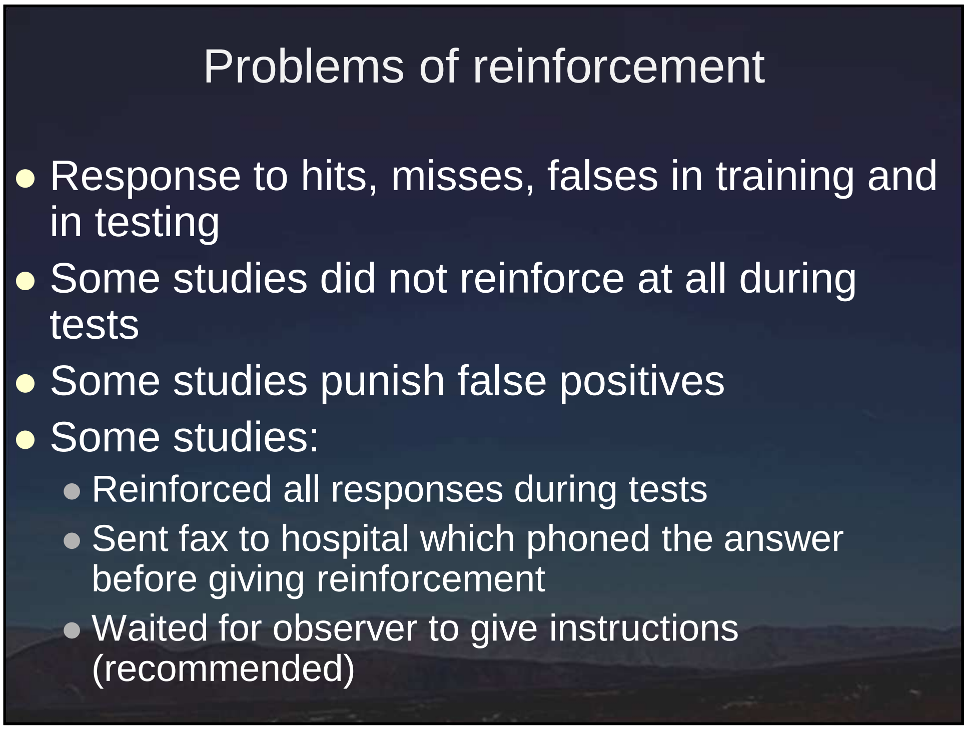
Design of the experiment

- GiGo
- Involve a knowledgeable researcher
- Proper design requires understanding relevant variables that influence performance
  - - Need a rational basis for choice of parameters
  - Understand the implications of each choice in the design
- What are the control groups
- Odor samples
  - Where, when collected, how collected, how stored and how many collected
- Training policies
  - Train to criterion- but what criterion?
- Reinforcement policies



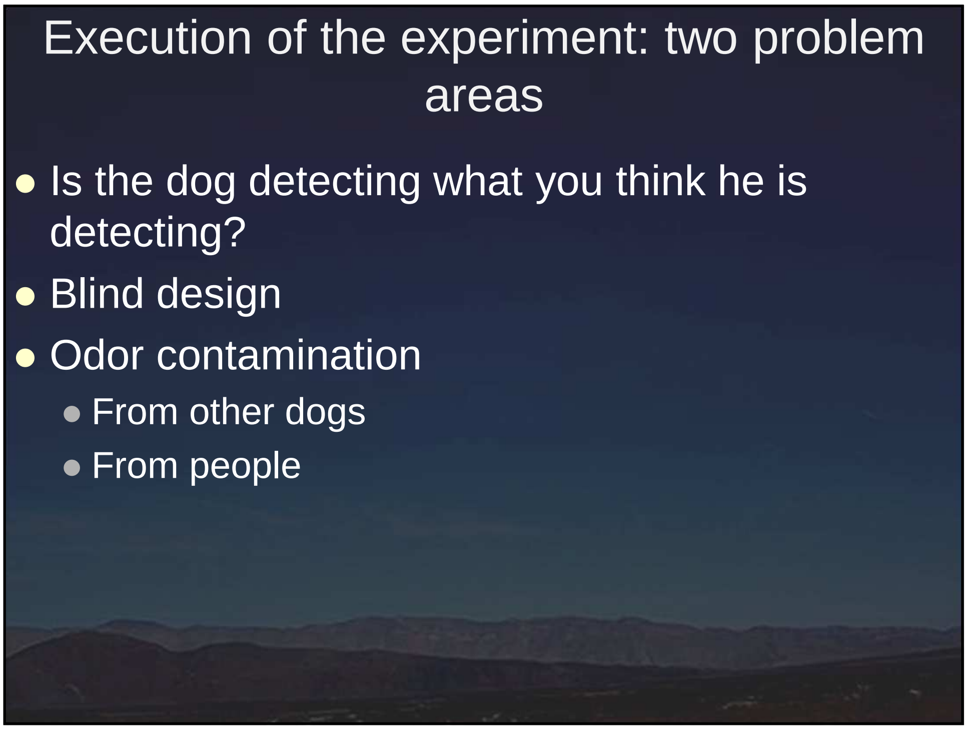
Examples of relevant questions during design

- Should a positive sample always be present in the line-up?
- Should there be a possibility of more than one target in a line-up?
  - Should each station be considered independent?
- How many stations in the lineup
- Should either targets or negative odors be reused?
- How is each odor collected and stored?
- What is the ratio of positive to negative samples
  - This determines the minimum number of samples
- How are the cancer and controls matched?
- What are the possible sources of contamination?



Problems of reinforcement

- Response to hits, misses, falses in training and in testing
- Some studies did not reinforce at all during tests
- Some studies punish false positives
- Some studies:
  - Reinforced all responses during tests
  - Sent fax to hospital which phoned the answer before giving reinforcement
  - Waited for observer to give instructions (recommended)



Execution of the experiment: two problem areas

- Is the dog detecting what you think he is detecting?
- Blind design
- Odor contamination
  - From other dogs
  - From people

Essential need for blind experiments

- If the experiment is not blind, it is worthless
- Robert Hinde " the moment you begin to observe, you abstract"
  - One cannot be an objective observer
- Neither the handler nor anyone else in the room should know the correct container
- You cannot fail to give cues-
  - Even when measures are "objective"

Bias in vet students in rating social behavior of pigs

Observer bias in animal behaviour research: can we believe what we score, if we score what we believe?

F. A. M. Tuyttens et al. Animal Behaivor (2014).

students were trained to code pig social behaviour
Saw two videos, one of a group of pigs selectively bred for pro-social behaviour, one control
. They were told which group was specially bred. And then scored each video
As expected, the pro-social group was more social.

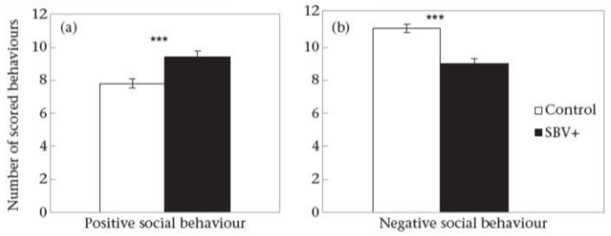


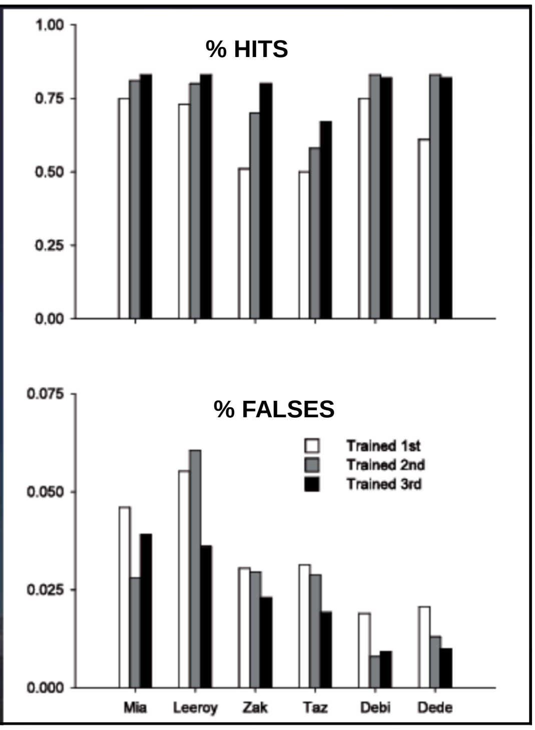
Figure 2. The mean number ± SE of (a) positive and (b) negative behaviours scored during the 5 min video clip of the control group and the video clip where the students were told that the animals were selected for high social breeding value (SBV+). $^{***}P < 0.001$.

Olfactory contamination: some examples

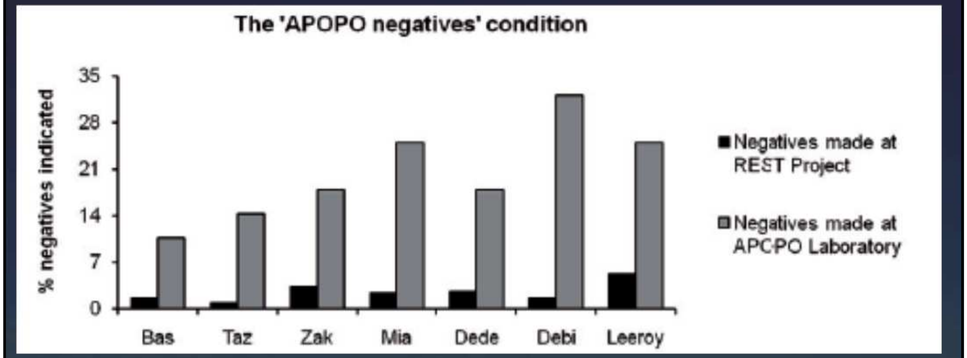
- Dogs respond to ambient air of hospital
  - Over 30% positive response to the hospital air

Hits and falses as a function of preceding dog on same stimuli

- The first dog always had fewer hits and usually more false positives.
- The second and third dogs had more hits and fewer false positives.



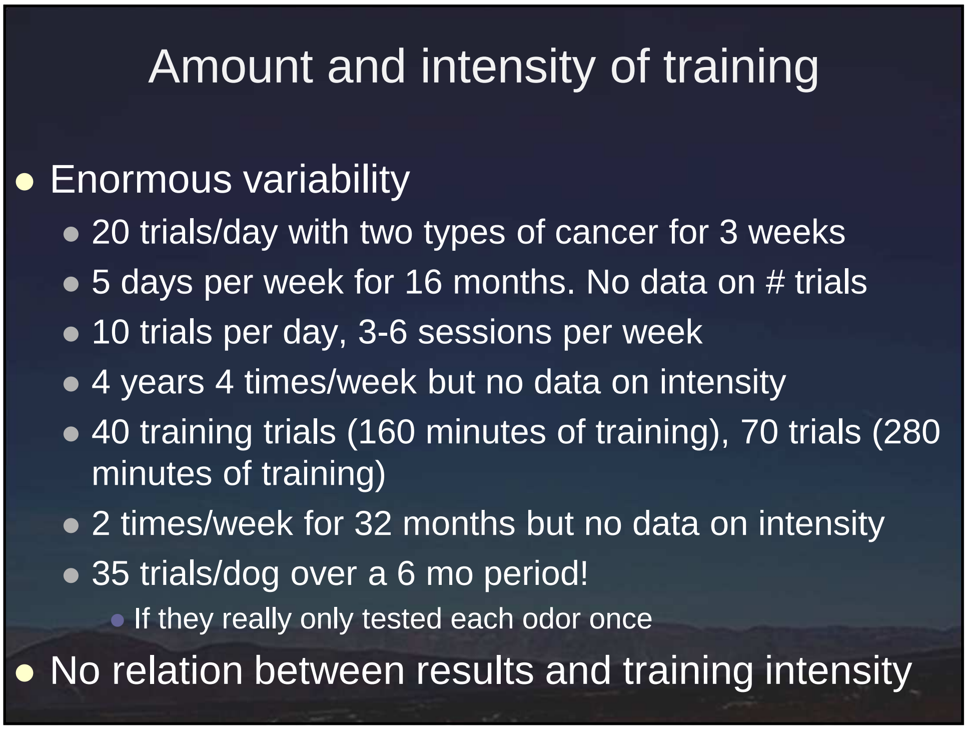
Number falses as a function of preparation room.



Could mean either that there was cross-contamination or, more likely, The two areas had a different ambient odor.

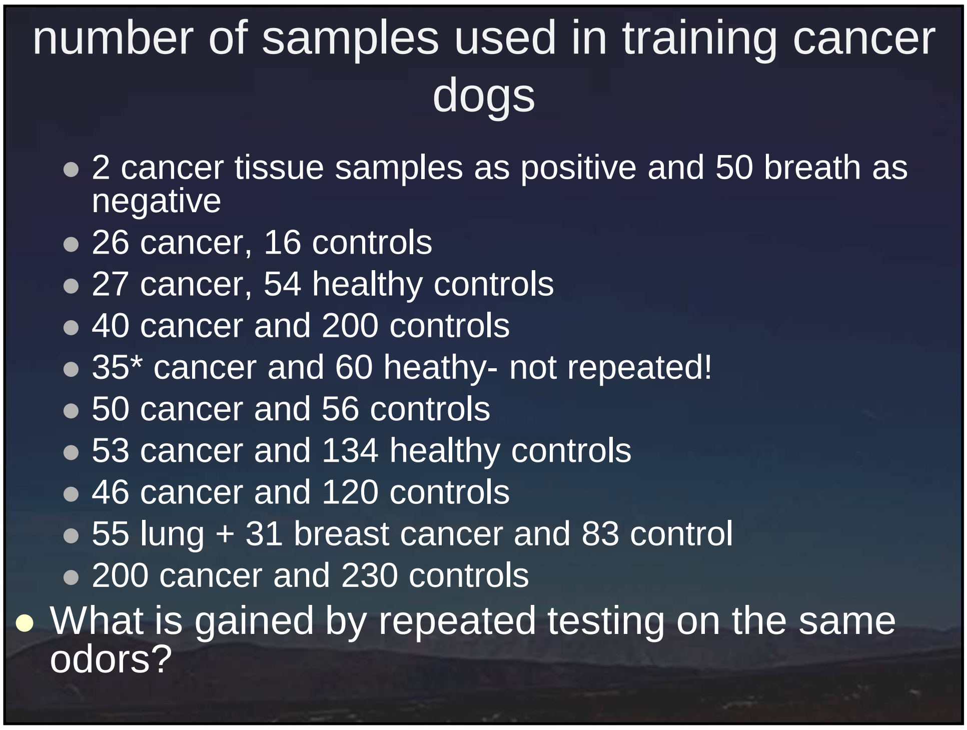
Two critical parameters determine validity

- Amount of training
- Number of exemplars
- They are inter-related



Amount and intensity of training

- Enormous variability
  - 20 trials/day with two types of cancer for 3 weeks
  - 5 days per week for 16 months. No data on # trials
  - 10 trials per day, 3-6 sessions per week
  - 4 years 4 times/week but no data on intensity
  - 40 training trials (160 minutes of training), 70 trials (280 minutes of training)
  - 2 times/week for 32 months but no data on intensity
  - 35 trials/dog over a 6 mo period!
    - If they really only tested each odor once
- No relation between results and training intensity



number of samples used in training cancer dogs

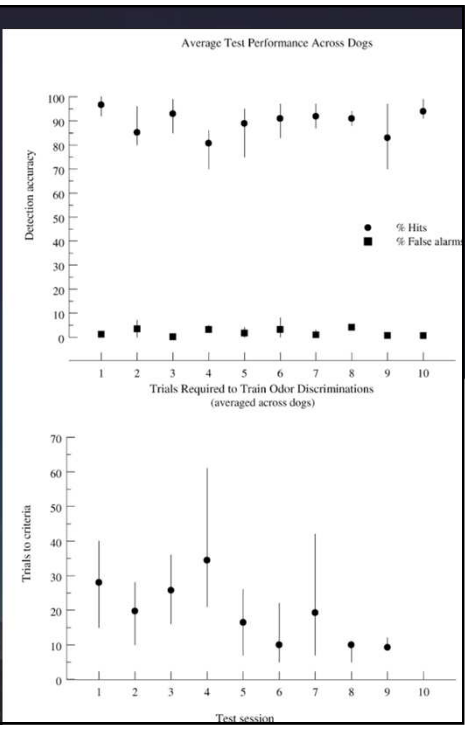
  - 2 cancer tissue samples as positive and 50 breath as negative
  - 26 cancer, 16 controls
  - 27 cancer, 54 healthy controls
  - 40 cancer and 200 controls
  - 35* cancer and 60 heathy- not repeated!
  - 50 cancer and 56 controls
  - 53 cancer and 134 healthy controls
  - 46 cancer and 120 controls
  - 55 lung + 31 breast cancer and 83 control
  - 200 cancer and 230 controls
- What is gained by repeated testing on the same odors?

Problems of small sample sizes

- the number of samples used in training is insufficient
- In many experiments at least some odor samples are used repeatedly
- Dog can memorize a large number of odors very rapidly
  - If the positive AND negative samples are not always new, the dog can learn to recognize the individual odors
    - One experiment found that dogs that discriminated the training samples could not discriminate new samples in the test
    - Another found a learning curve with repeated testing with the same odors

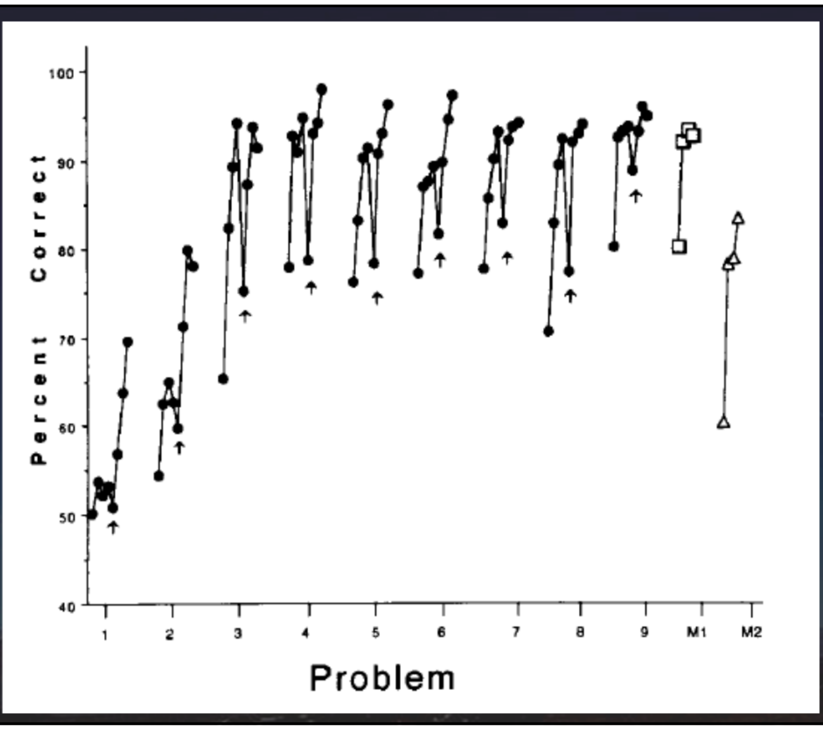
Odor memory in dogs

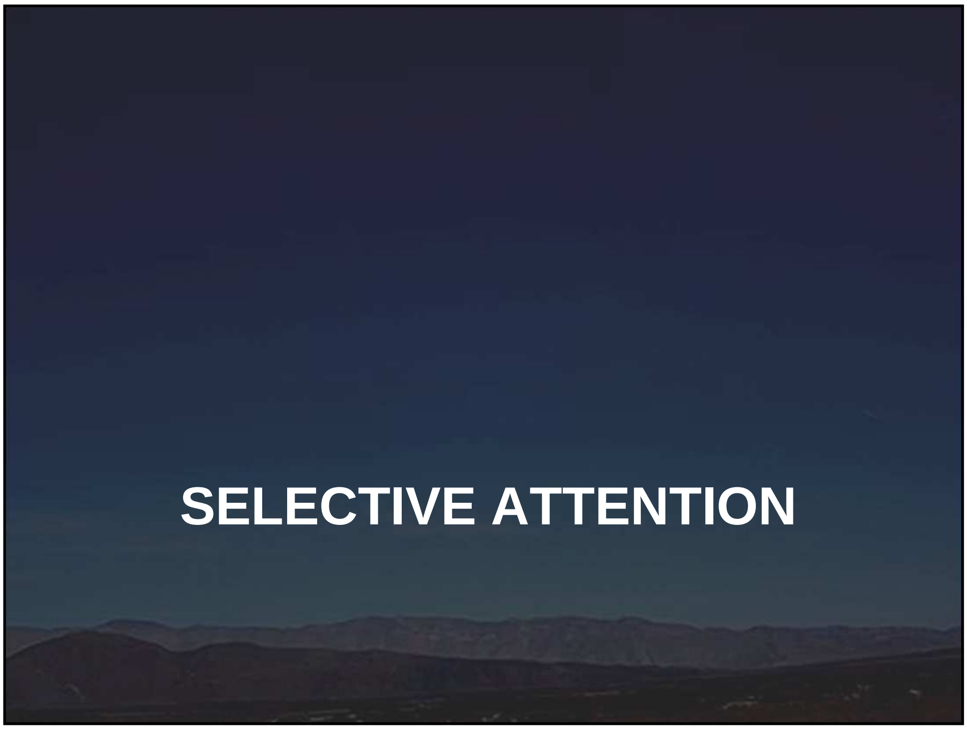
- Dogs have good memory
  - Rico knew 200 names
  - Chaser knew 1000 names
- No one has tested the limits of olfactory memory
- But it is long lasting- even years!
- Williams trained dogs to detect 10 different odors. He found no problems



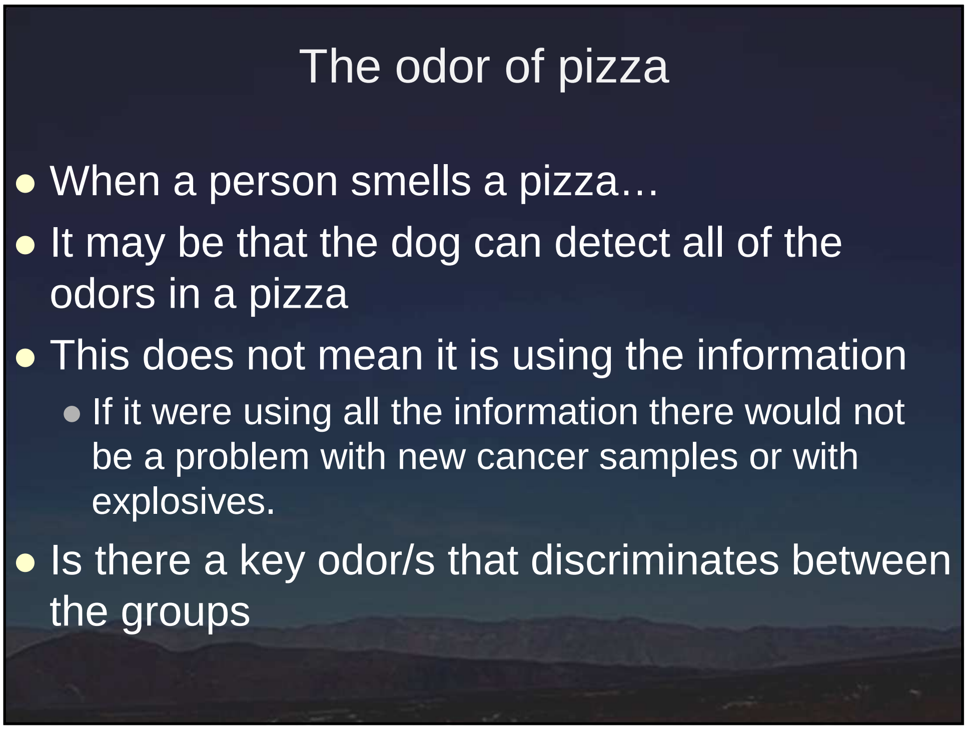
Olfactory memory in rats

- Slotnick: 9 sets of 8 odors, 4+ and 4- (36+ 36-)
  - 160 trials/day/set for two days
  - Two memory probes
    - Rerun of set 3
    - Rerun of set 6- but reversed.
- Results: the rats rapidly learned the odors- and remembered them!
- Dogs can easily learn all the positive and negative odors used in experiments with small samples.
  - If the same samples are used repeatedly.





SELECTIVE ATTENTION



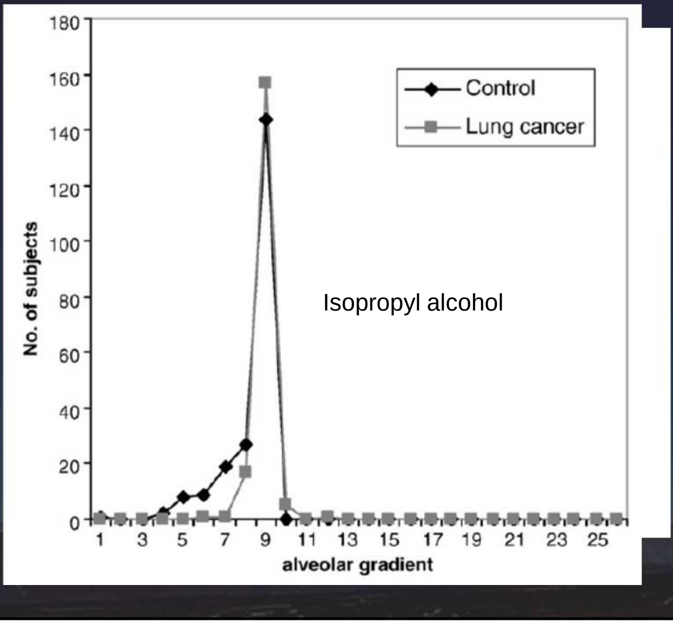
The odor of pizza

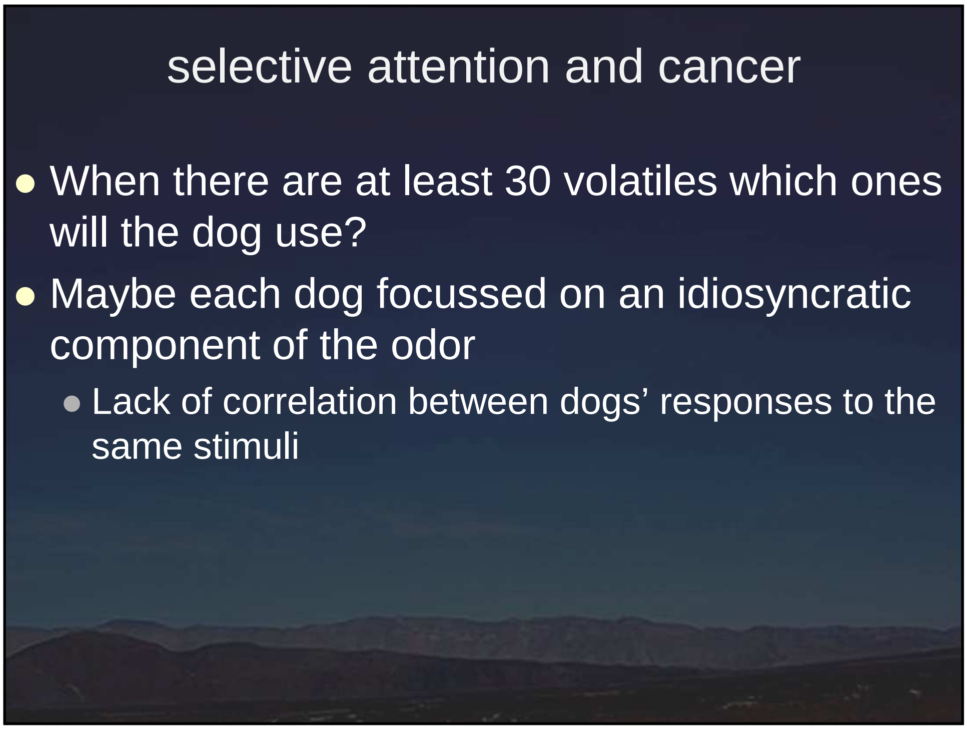
- When a person smells a pizza…
- It may be that the dog can detect all of the odors in a pizza
- This does not mean it is using the information
  - If it were using all the information there would not be a problem with new cancer samples or with explosives.
- Is there a key odor/s that discriminates between the groups

Is there a characteristic scent for a specific cancer?

- Do all cancers smell the same?
- Do all examples of a given cancer type smell the same?
- There are thousands of VOCs in the breath and in urine.
  - Phillips et al (1999) found:
    - An average of 204 VOCs in breath of healthy people
    - 3481 different VOCs were found in his subjects
    - Only 27 VOCs were common for all 50 subjects
- 30 VOCs distinguished lung cancer patients from other sick patients (Phillips, 2007).
  - Quantitative, not qualitative
- Does the dog detect ONE VOC or a combination of VOCs?

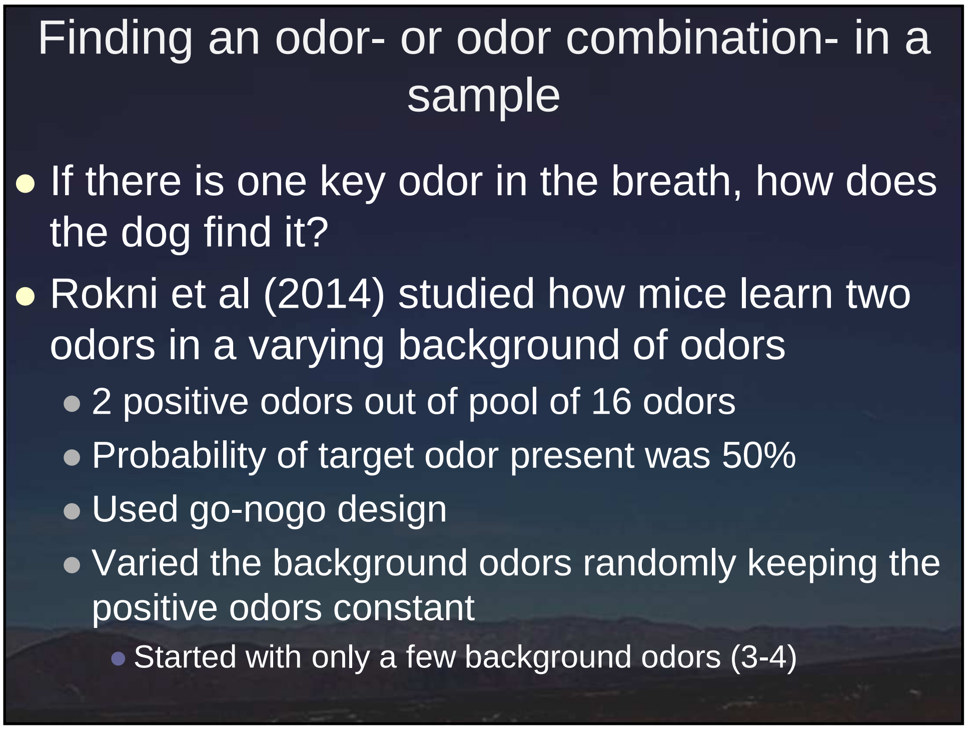
One voc versus several





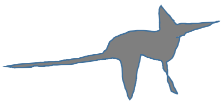
selective attention and cancer

- When there are at least 30 volatiles which ones will the dog use?
- Maybe each dog focussed on an idiosyncratic component of the odor
  - Lack of correlation between dogs’ responses to the same stimuli

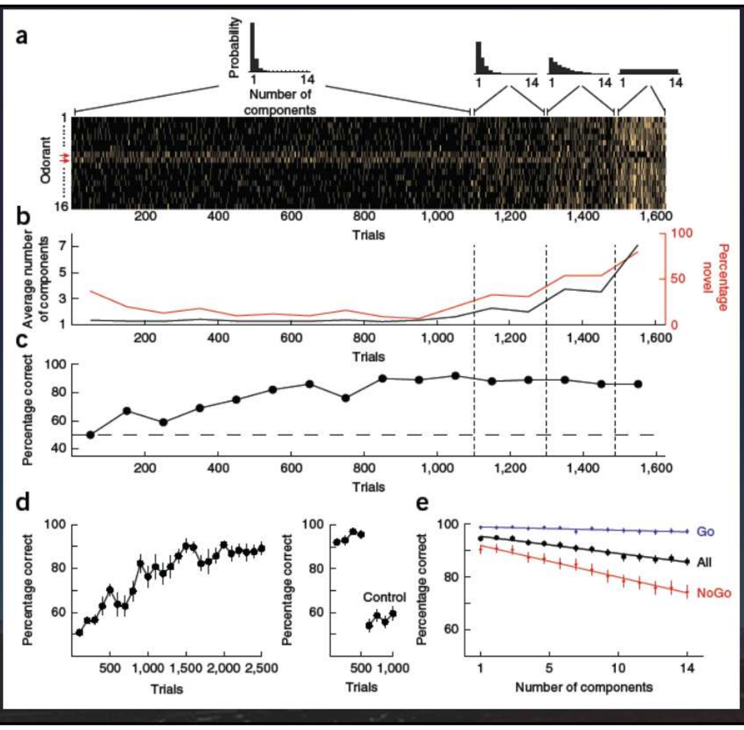


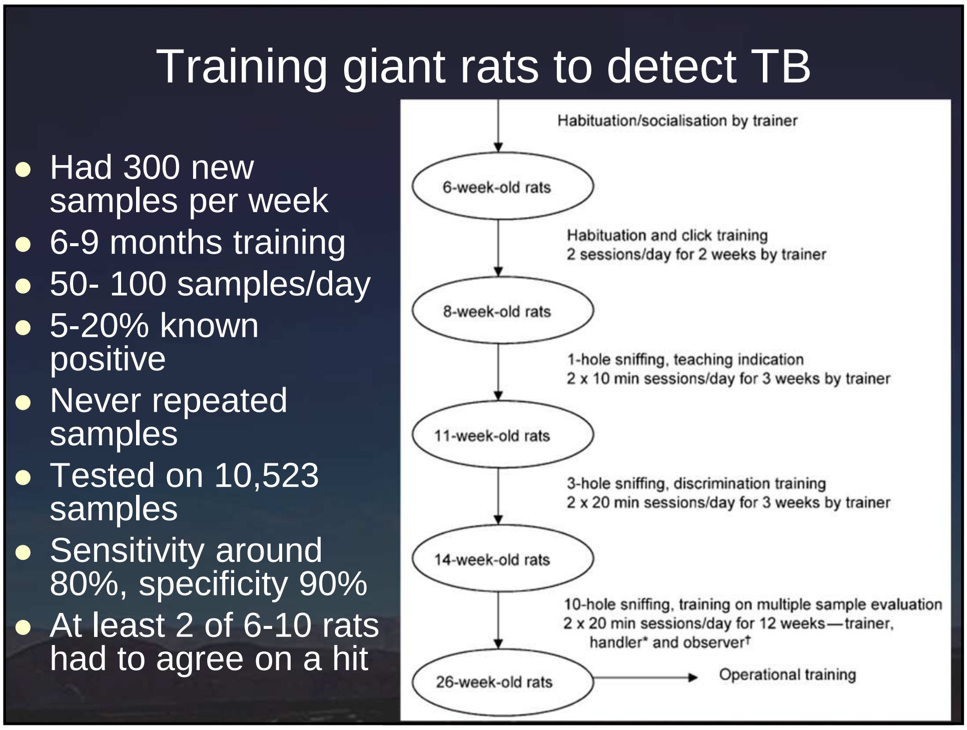
Finding an odor- or odor combination- in a sample

- If there is one key odor in the breath, how does the dog find it?
- Rokni et al (2014) studied how mice learn two odors in a varying background of odors
  - 2 positive odors out of pool of 16 odors
  - Probability of target odor present was 50%
  - Used go-nogo design
  - Varied the background odors randomly keeping the positive odors constant
    - Started with only a few background odors (3-4)



Reached criterion of 80% after about 1000 trials!
Reached plateau after 2400 trials
Continued with an additional 34,000 trials





Training giant rats to detect TB

- Had 300 new samples per week
- 6-9 months training
- 50- 100 samples/day
- 5-20% known positive
- Never repeated samples
- Tested on 10,523 samples
- Sensitivity around 80%, specificity 90%
- At least 2 of 6-10 rats had to agree on a hit

Habituation/socialisation by trainer

6-week-old rats

Habituation and click training
2 sessions/day for 2 weeks by trainer

8-week-old rats

1-hole sniffing, teaching indication
2 x 10 min sessions/day for 3 weeks by trainer

11-week-old rats

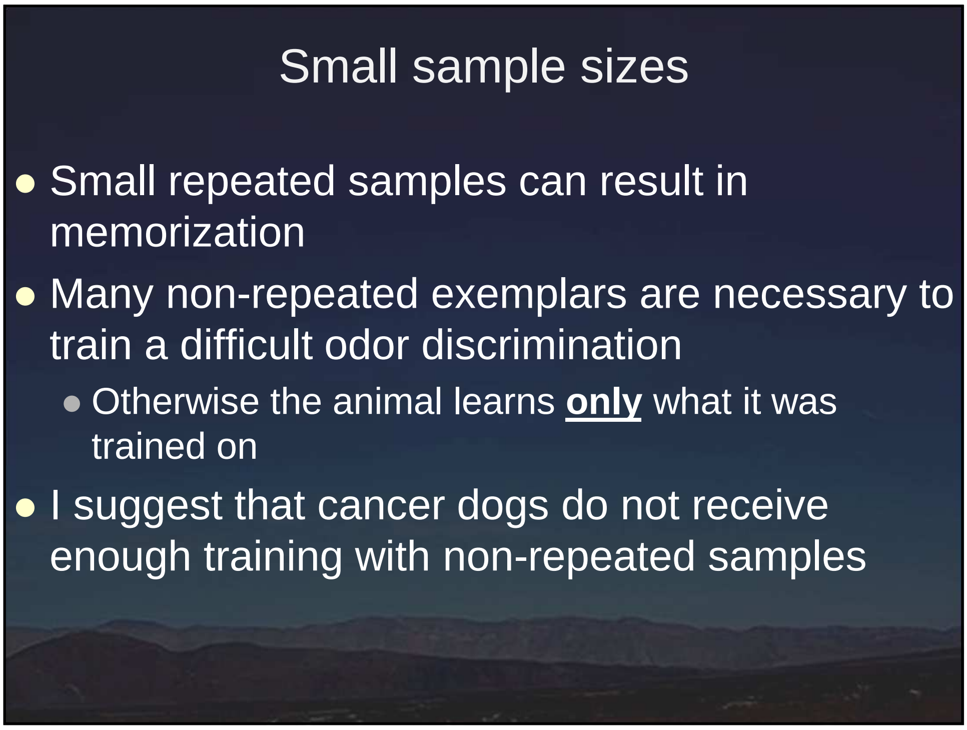
3-hole sniffing, discrimination training
2 x 20 min sessions/day for 3 weeks by trainer

14-week-old rats

10-hole sniffing, training on multiple sample evaluation
2 x 20 min sessions/day for 12 weeks—trainer, handler* and observer†

26-week-old rats

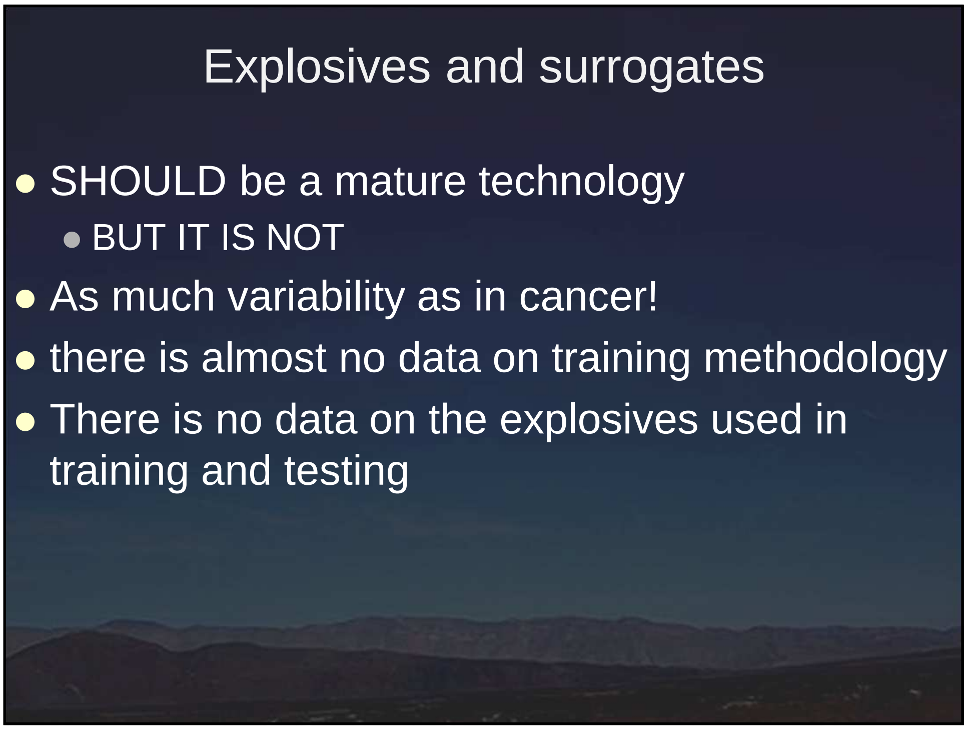
Operational training



Small sample sizes

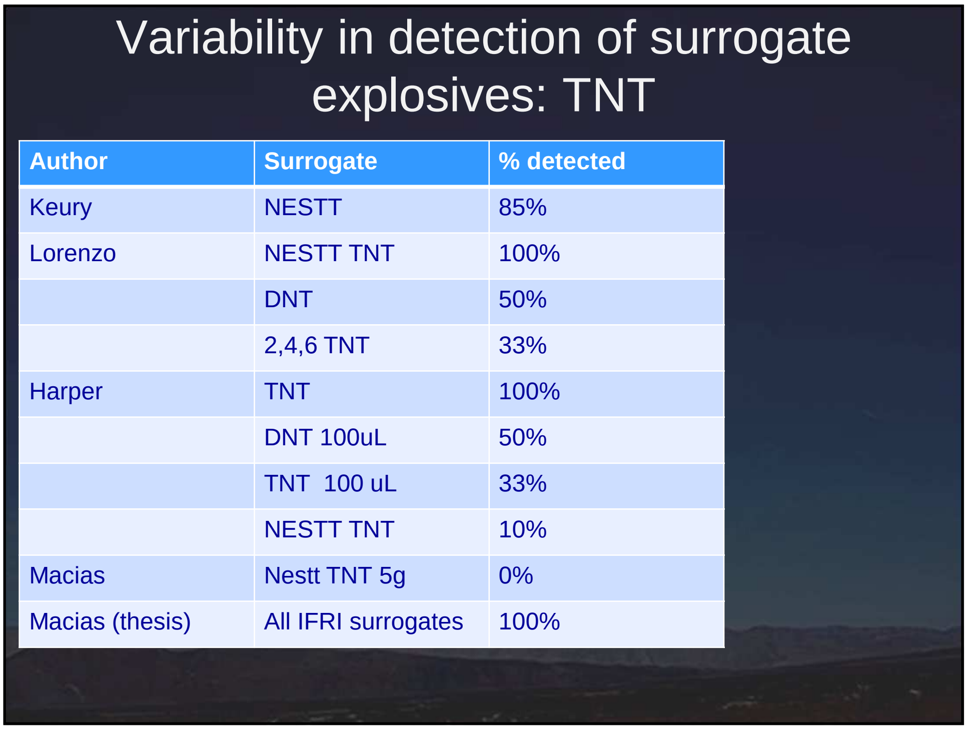
- Small repeated samples can result in memorization
- Many non-repeated exemplars are necessary to train a difficult odor discrimination
  - Otherwise the animal learns **only** what it was trained on
- I suggest that cancer dogs do not receive enough training with non-repeated samples

DOG DETECTION OF SURROGATE EXPLOSIVES



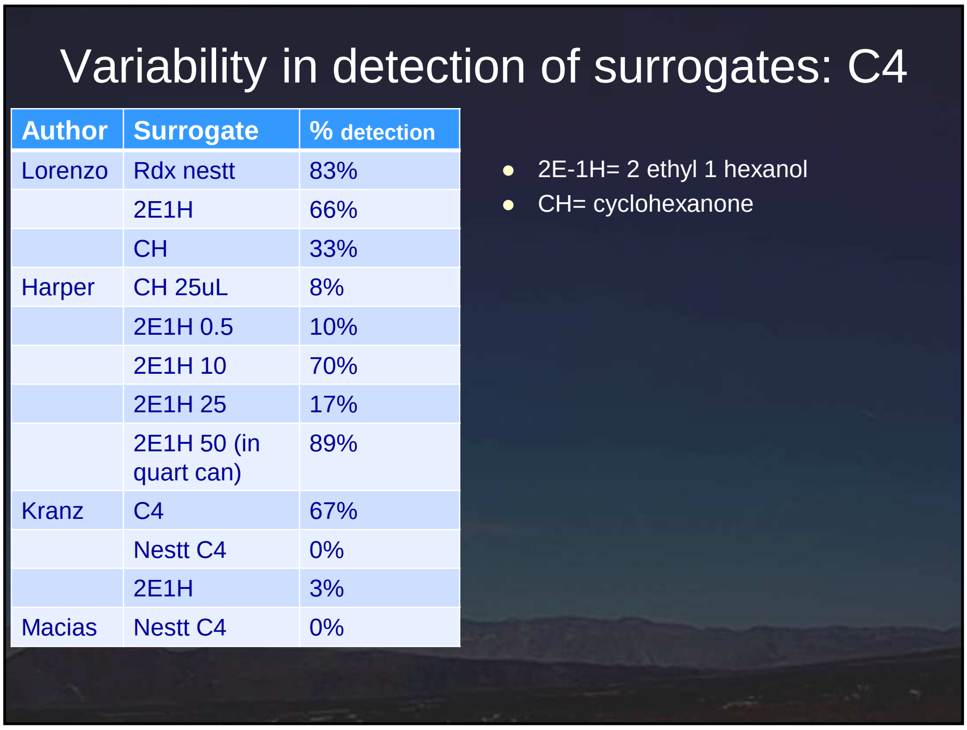
Explosives and surrogates

- SHOULD be a mature technology
  - BUT IT IS NOT
- As much variability as in cancer!
- there is almost no data on training methodology
- There is no data on the explosives used in training and testing



Variability in detection of surrogate explosives: TNT

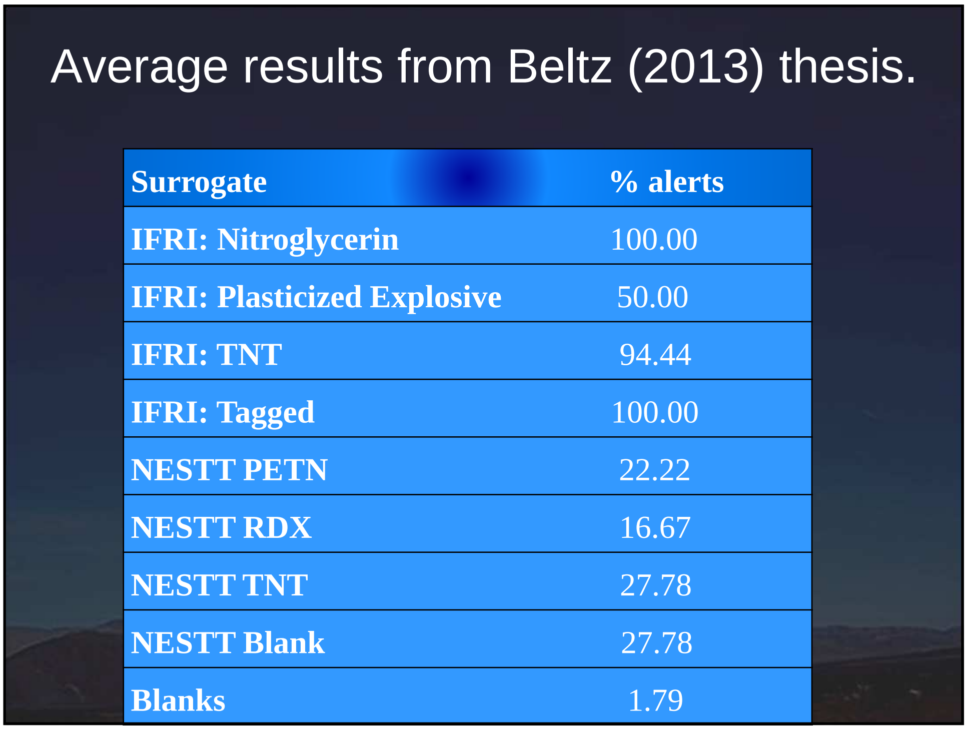
Author	Surrogate	% detected
Keury	NESTT	85%
Lorenzo	NESTT TNT	100%
	DNT	50%
	2,4,6 TNT	33%
Harper	TNT	100%
	DNT 100uL	50%
	TNT 100 uL	33%
	NESTT TNT	10%
Macias	Nestt TNT 5g	0%
Macias (thesis)	All IFRI surrogates	100%



Variability in detection of surrogates: C4

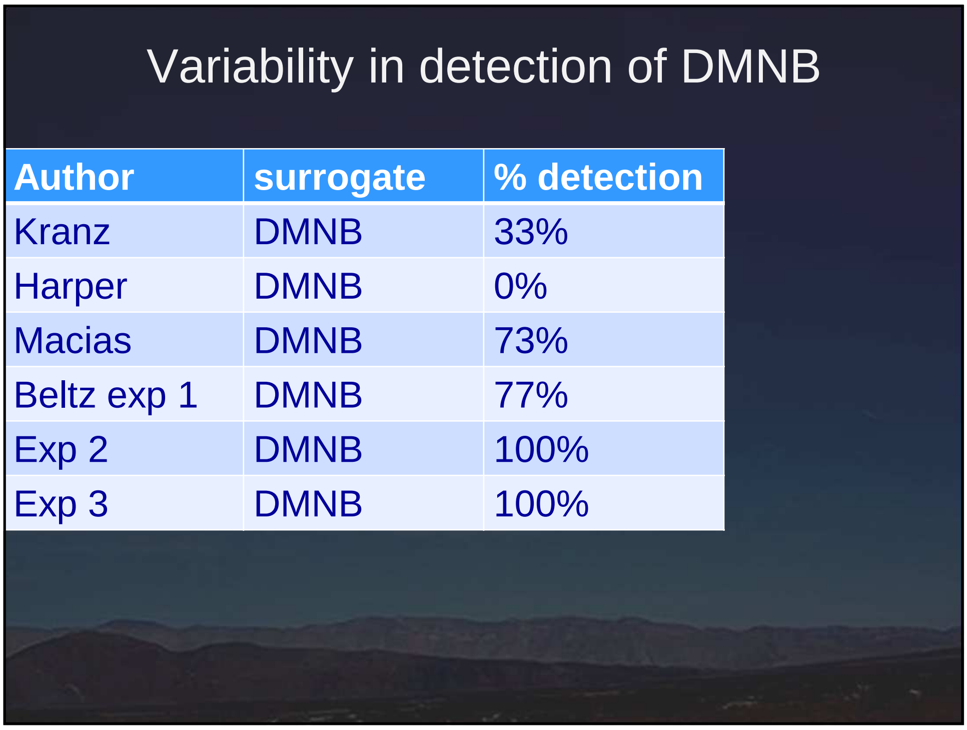
Author	Surrogate	% detection
Lorenzo	Rdx nestt	83%
	2E1H	66%
	CH	33%
Harper	CH 25uL	8%
	2E1H 0.5	10%
	2E1H 10	70%
	2E1H 25	17%
	2E1H 50 (in quart can)	89%
Kranz	C4	67%
	Nestt C4	0%
	2E1H	3%
Macias	Nestt C4	0%

- 2E-1H= 2 ethyl 1 hexanol
- CH= cyclohexanone



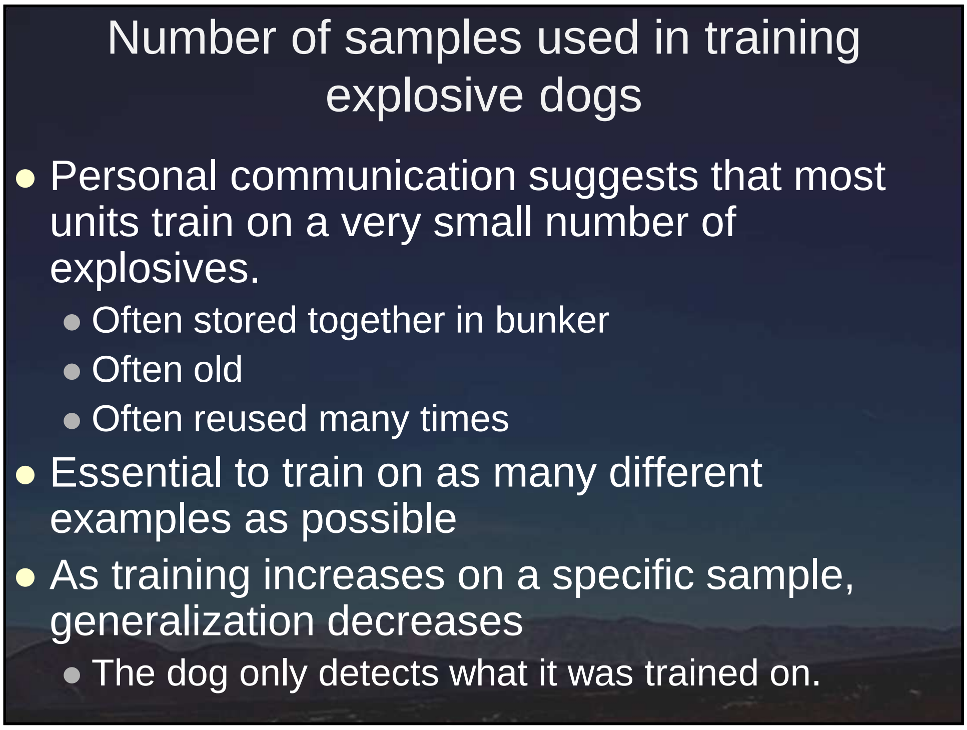
Average results from Beltz (2013) thesis.

Surrogate	% alerts
IFRI: Nitroglycerin	100.00
IFRI: Plasticized Explosive	50.00
IFRI: TNT	94.44
IFRI: Tagged	100.00
NESTT PETN	22.22
NESTT RDX	16.67
NESTT TNT	27.78
NESTT Blank	27.78
Blanks	1.79



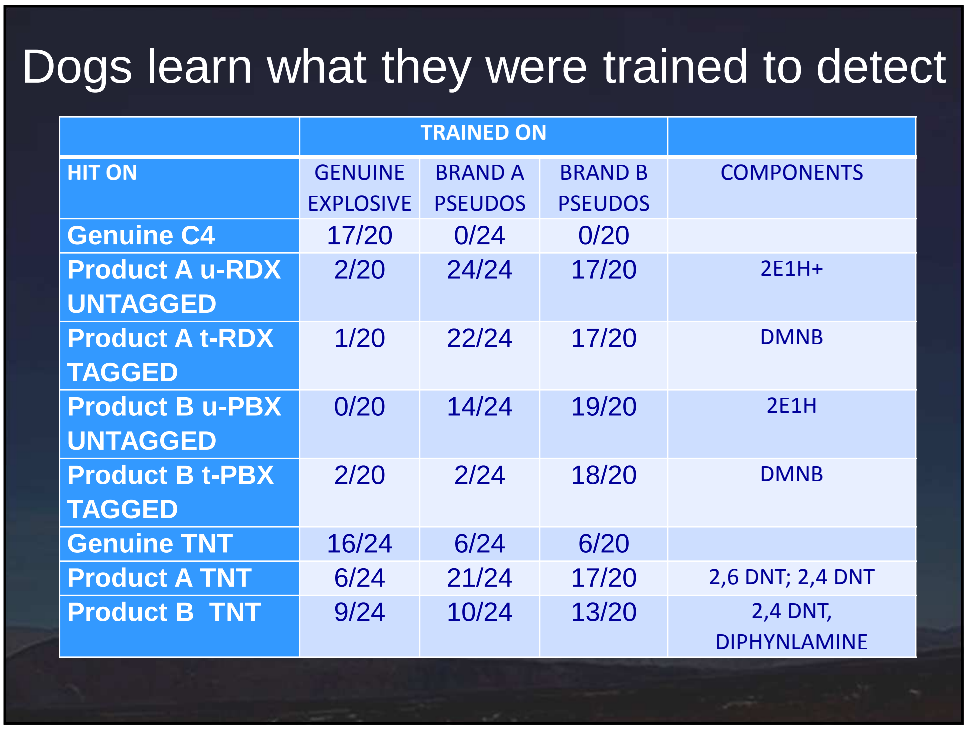
Variability in detection of DMNB

Author	surrogate	% detection
Kranz	DMNB	33%
Harper	DMNB	0%
Macias	DMNB	73%
Beltz exp 1	DMNB	77%
Exp 2	DMNB	100%
Exp 3	DMNB	100%



Number of samples used in training explosive dogs

- Personal communication suggests that most units train on a very small number of explosives.
  - Often stored together in bunker
  - Often old
  - Often reused many times
- Essential to train on as many different examples as possible
- As training increases on a specific sample, generalization decreases
  - The dog only detects what it was trained on.



Dogs learn what they were trained to detect

	TRAINED ON			
HIT ON	GENUINE EXPLOSIVE	BRAND A PSEUDOS	BRAND B PSEUDOS	COMPONENTS
Genuine C4	17/20	0/24	0/20	
Product A u-RDX UNTAGGED	2/20	24/24	17/20	2E1H+
Product A t-RDX TAGGED	1/20	22/24	17/20	DMNB
Product B u-PBX UNTAGGED	0/20	14/24	19/20	2E1H
Product B t-PBX TAGGED	2/20	2/24	18/20	DMNB
Genuine TNT	16/24	6/24	6/20	
Product A TNT	6/24	21/24	17/20	2,6 DNT; 2,4 DNT
Product B TNT	9/24	10/24	13/20	2,4 DNT, DIPHYNLAMINE

Selective attention

- Kranz- although relatively poor results but:
  - One dog trained on surrogate TNT did hit on all examples of real TNT
  - One dog trained on surrogate gunpowder did hit on all examples of real gunpowder

Is there a characteristic odor for a given explosive?

Williams and what dogs detect in C4

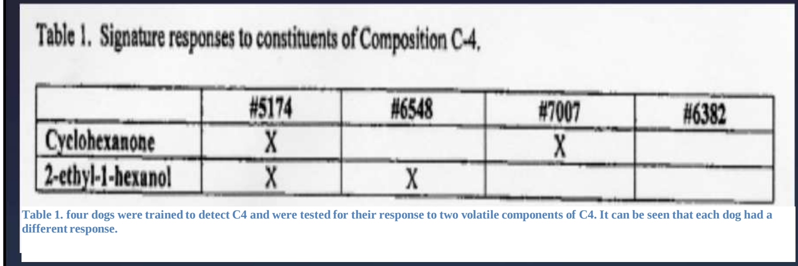


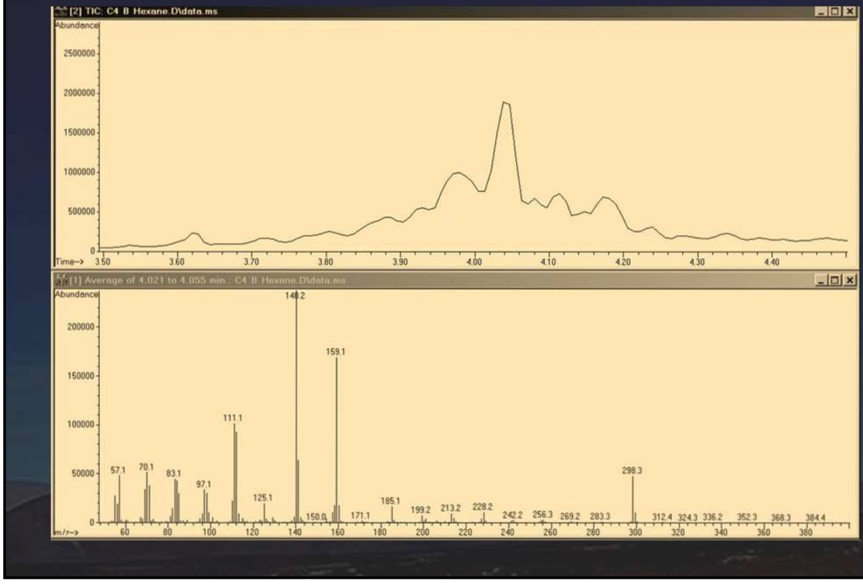
Table 1. Signature responses to constituents of Composition C-4.

	#5174	#6548	#7007	#6382
Cyclohexanone	X		X	
2-ethyl-1-hexanol	X	X		

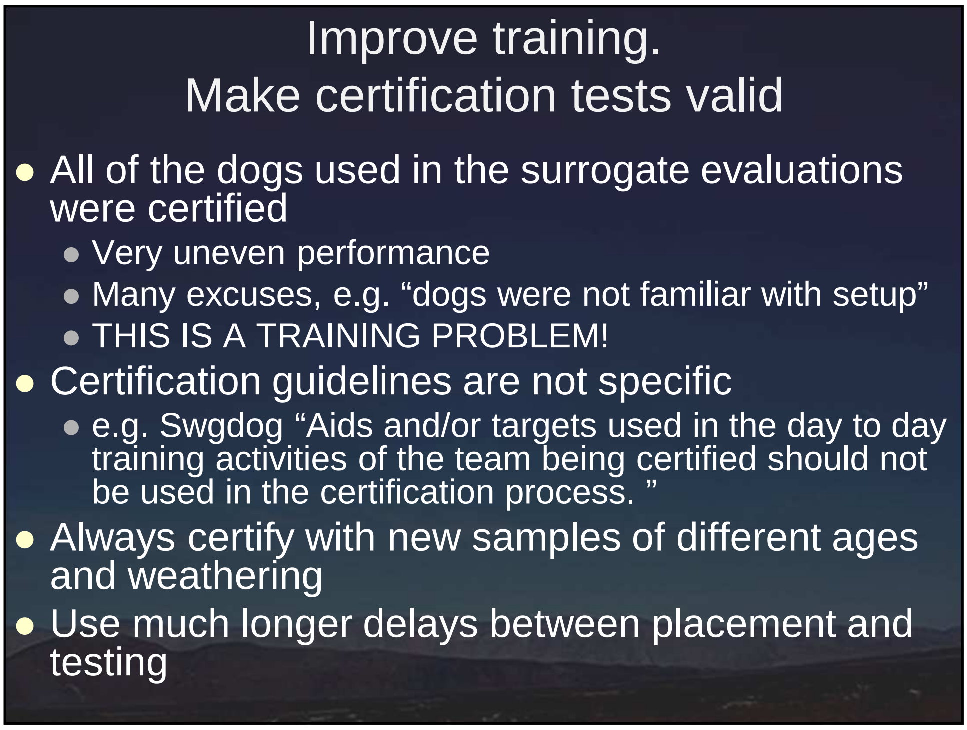
Table 1. four dogs were trained to detect C4 and were tested for their response to two volatile components of C4. It can be seen that each dog had a different response.

Dogs do not necessarily select the one component we think they should select! We have to help them make the choice.

g.c. of C4

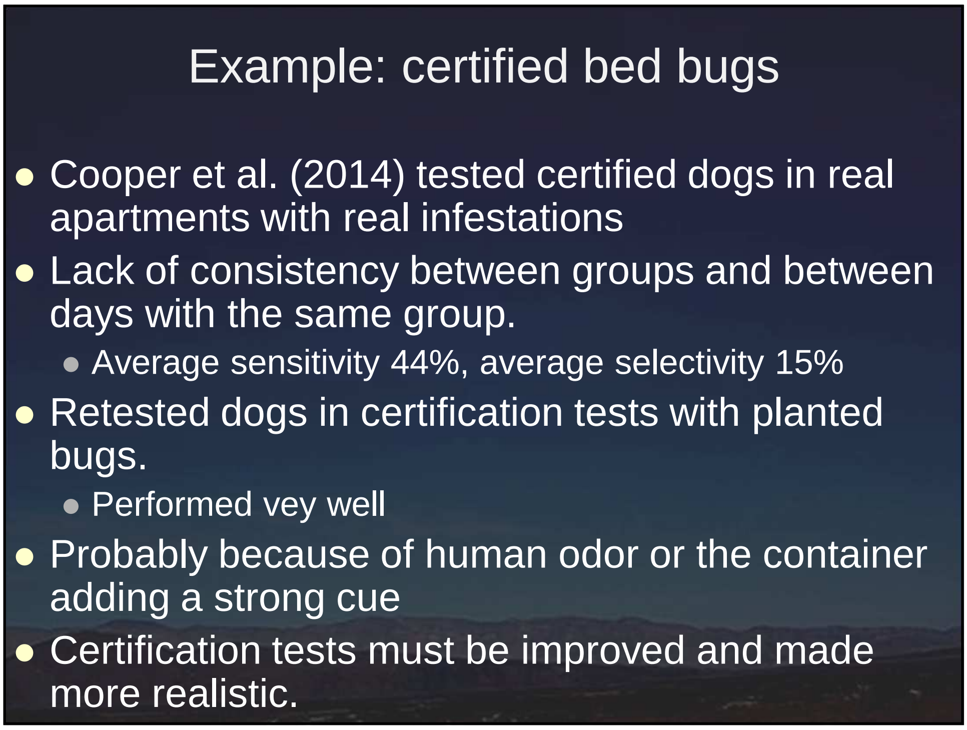


A SUGGESTED SOLUTION



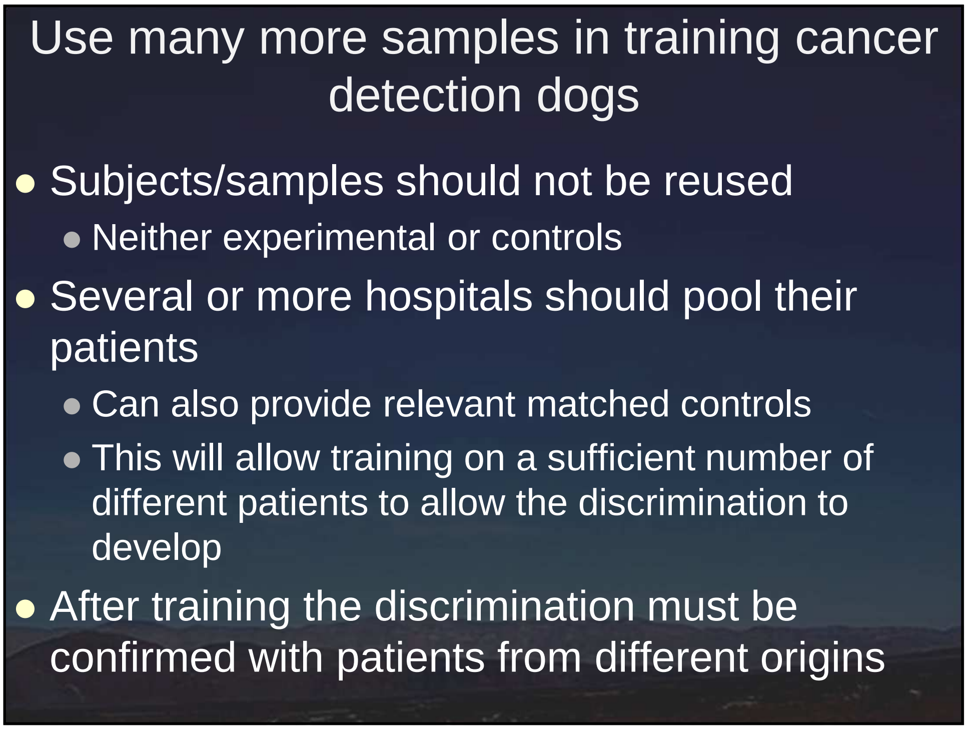
Improve training. Make certification tests valid

- All of the dogs used in the surrogate evaluations were certified
  - Very uneven performance
  - Many excuses, e.g. “dogs were not familiar with setup”
  - THIS IS A TRAINING PROBLEM!
- Certification guidelines are not specific
  - e.g. Swgdog “Aids and/or targets used in the day to day training activities of the team being certified should not be used in the certification process. ”
- Always certify with new samples of different ages and weathering
- Use much longer delays between placement and testing



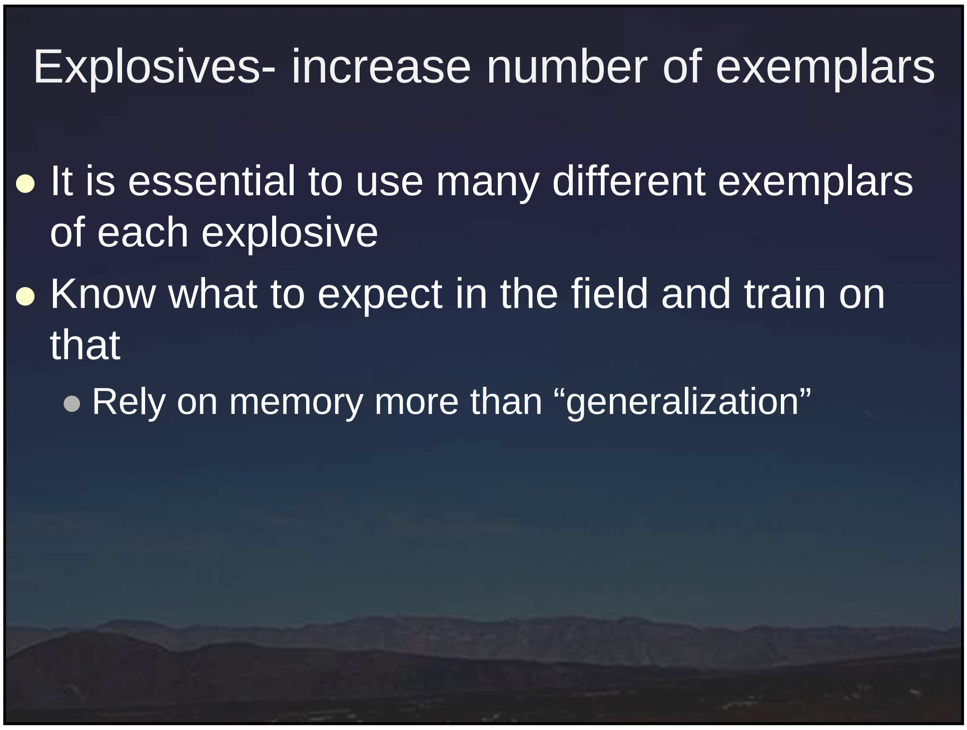
Example: certified bed bugs

- Cooper et al. (2014) tested certified dogs in real apartments with real infestations
- Lack of consistency between groups and between days with the same group.
  - Average sensitivity 44%, average selectivity 15%
- Retested dogs in certification tests with planted bugs.
  - Performed vey well
- Probably because of human odor or the container adding a strong cue
- Certification tests must be improved and made more realistic.



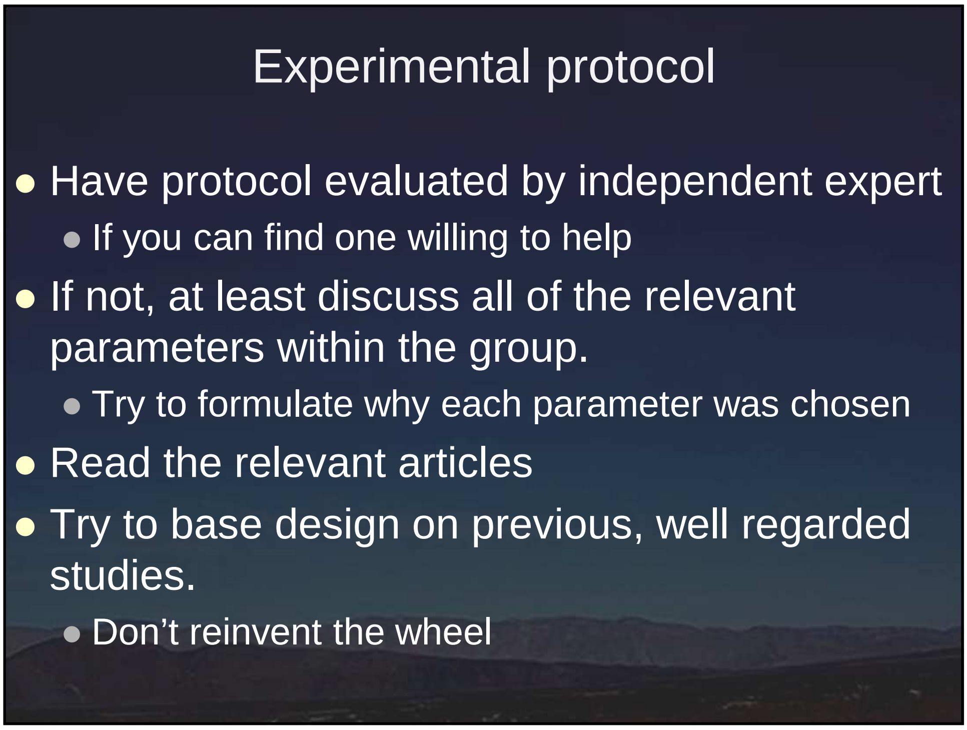
Use many more samples in training cancer detection dogs

- Subjects/samples should not be reused
  - Neither experimental or controls
- Several or more hospitals should pool their patients
  - Can also provide relevant matched controls
  - This will allow training on a sufficient number of different patients to allow the discrimination to develop
- After training the discrimination must be confirmed with patients from different origins



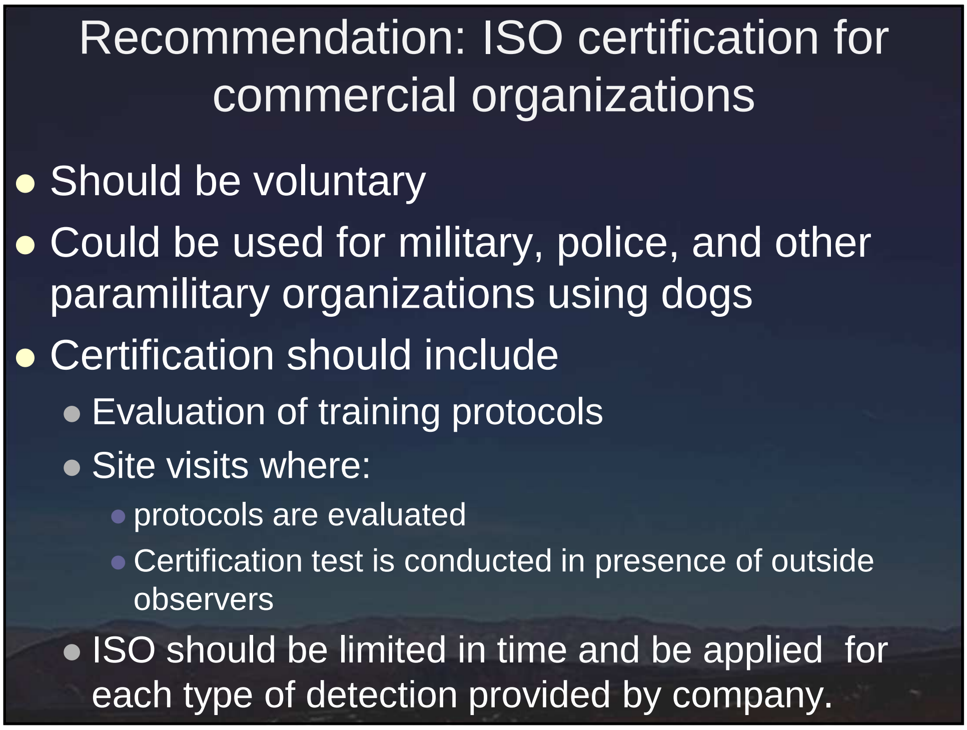
Explosives- increase number of exemplars

- It is essential to use many different exemplars of each explosive
- Know what to expect in the field and train on that
  - Rely on memory more than “generalization”



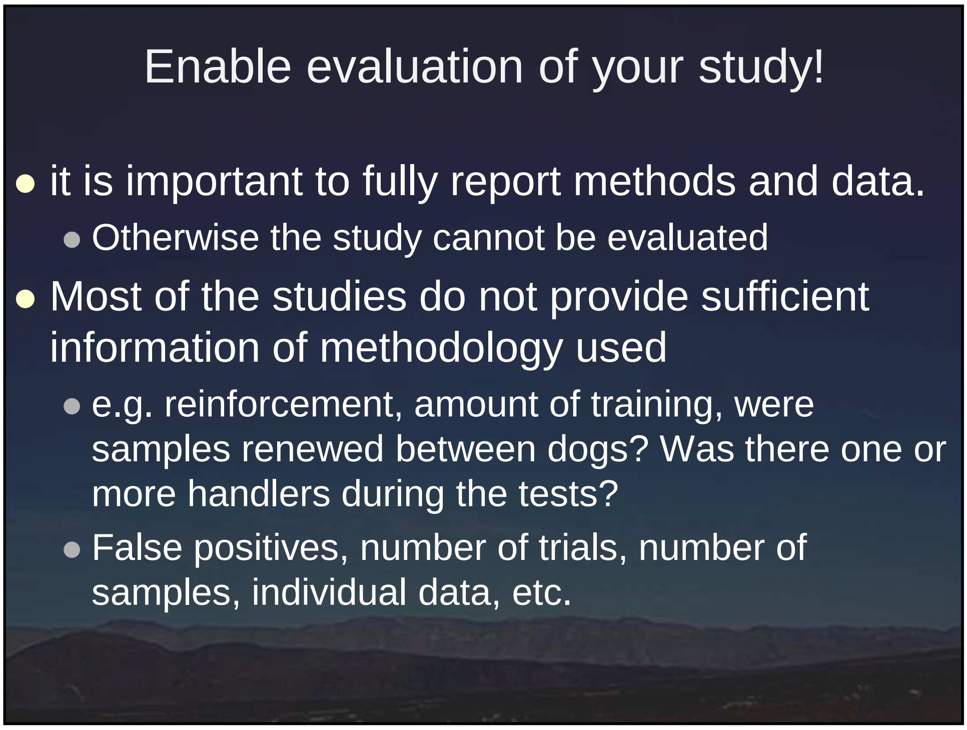
Experimental protocol

- Have protocol evaluated by independent expert
  - If you can find one willing to help
- If not, at least discuss all of the relevant parameters within the group.
  - Try to formulate why each parameter was chosen
- Read the relevant articles
- Try to base design on previous, well regarded studies.
  - Don’t reinvent the wheel



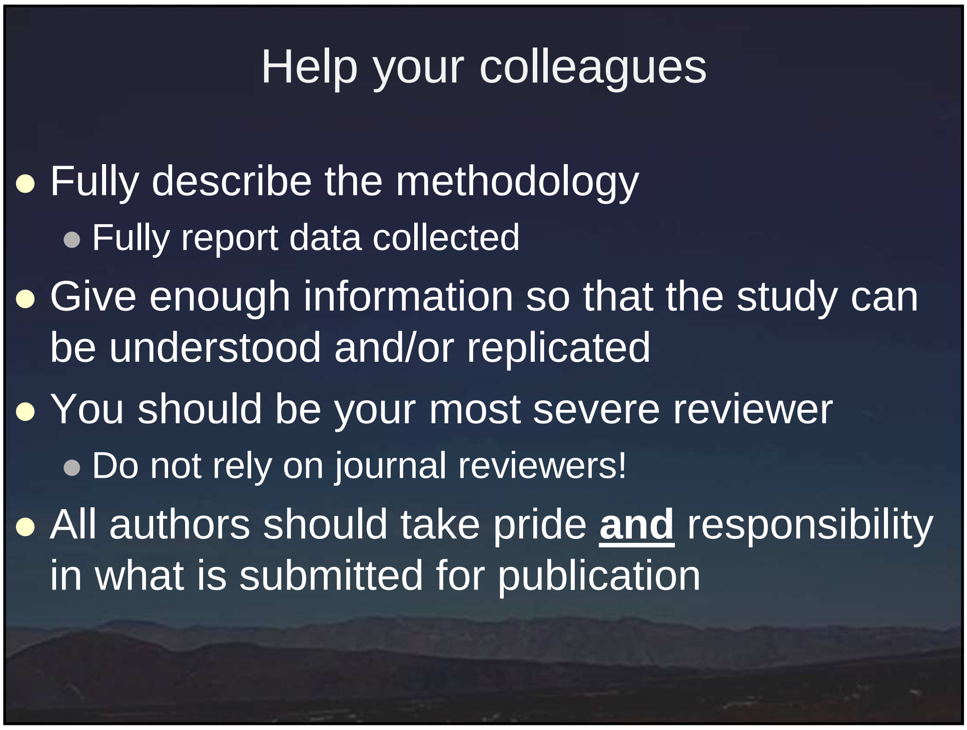
Recommendation: ISO certification for commercial organizations

- Should be voluntary
- Could be used for military, police, and other paramilitary organizations using dogs
- Certification should include
  - Evaluation of training protocols
  - Site visits where:
    - protocols are evaluated
    - Certification test is conducted in presence of outside observers
  - ISO should be limited in time and be applied for each type of detection provided by company.



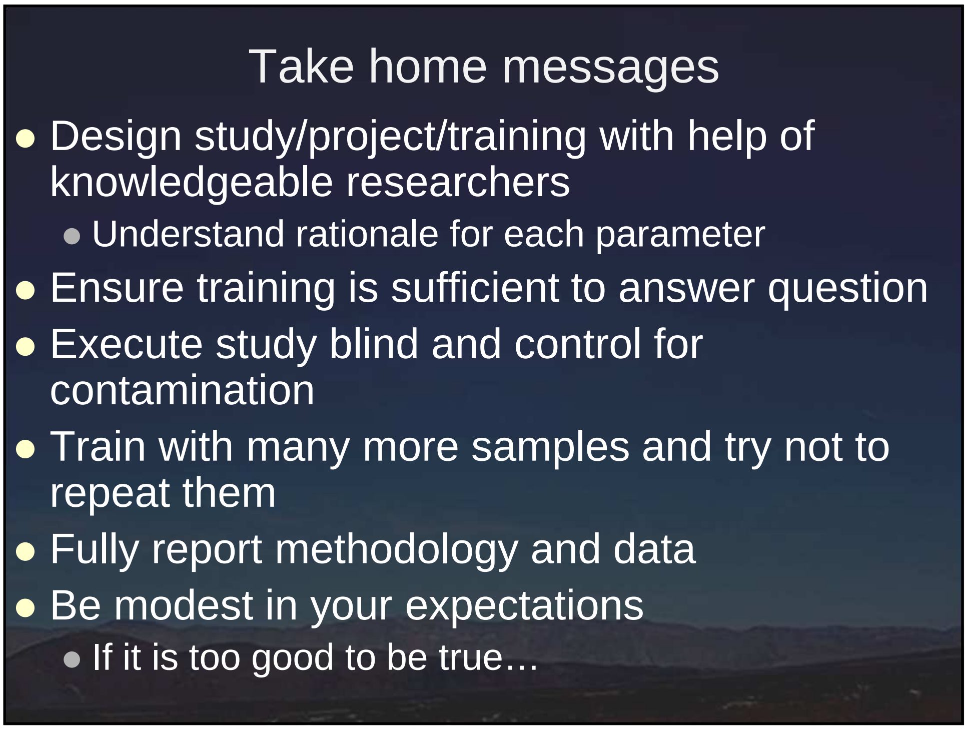
Enable evaluation of your study!

- it is important to fully report methods and data.
  - Otherwise the study cannot be evaluated
- Most of the studies do not provide sufficient information of methodology used
  - e.g. reinforcement, amount of training, were samples renewed between dogs? Was there one or more handlers during the tests?
  - False positives, number of trials, number of samples, individual data, etc.



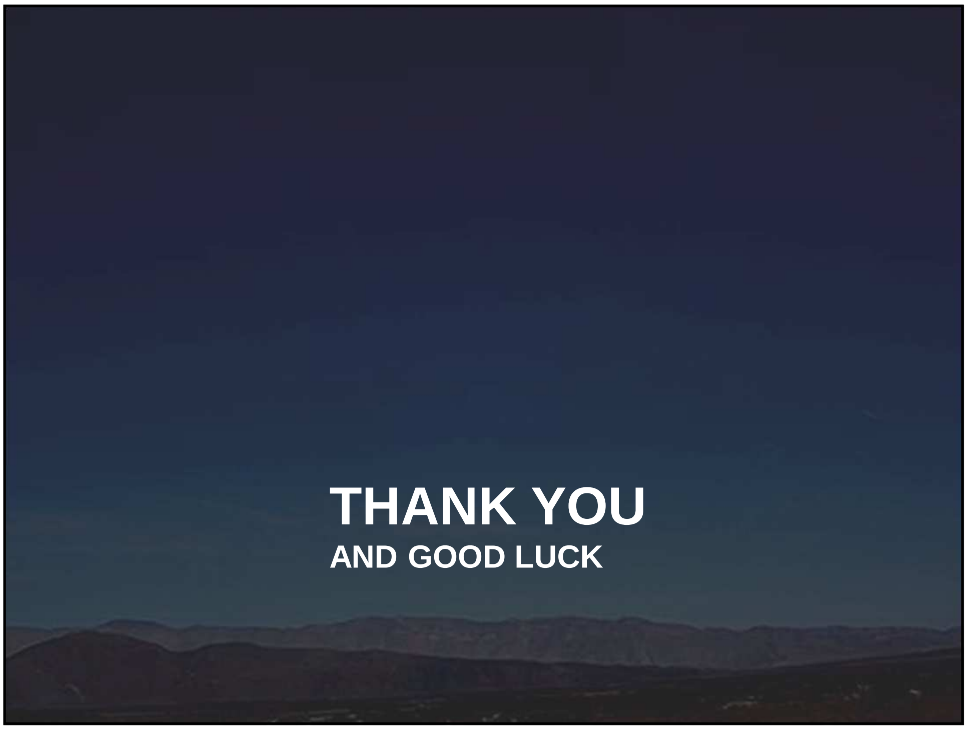
Help your colleagues

- Fully describe the methodology
  - Fully report data collected
- Give enough information so that the study can be understood and/or replicated
- You should be your most severe reviewer
  - Do not rely on journal reviewers!
- All authors should take pride **and** responsibility in what is submitted for publication



Take home messages

- Design study/project/training with help of knowledgeable researchers
  - Understand rationale for each parameter
- Ensure training is sufficient to answer question
- Execute study blind and control for contamination
- Train with many more samples and try not to repeat them
- Fully report methodology and data
- Be modest in your expectations
  - If it is too good to be true…



THANK YOU

AND GOOD LUCK