Lecture plan

- Fundamentals of Statistics
  - Summarizing Data
  - Understanding and modeling Random Variables
  - The Issue: Unknown 'Population' Vs known 'Sample'
  - The Idea: Central Limit Theorem (CLT)
- Hypothesis testing
  - A/B Testing
  - Other tests
- Anomaly Detection

Assume: $H_0: \mu = \square$ $\quad H_a: \mu > \square$

Evidence: $\bar{x} = \_\_\_$

$s = \_\_\_$

$n = \_\_\_$

p-value

Question: what are the chances of observing this evidence or worse if H_0 is true

$p < \alpha$ $\quad p \geq \alpha$

Unlikely → Reject H_0, Accept H_a

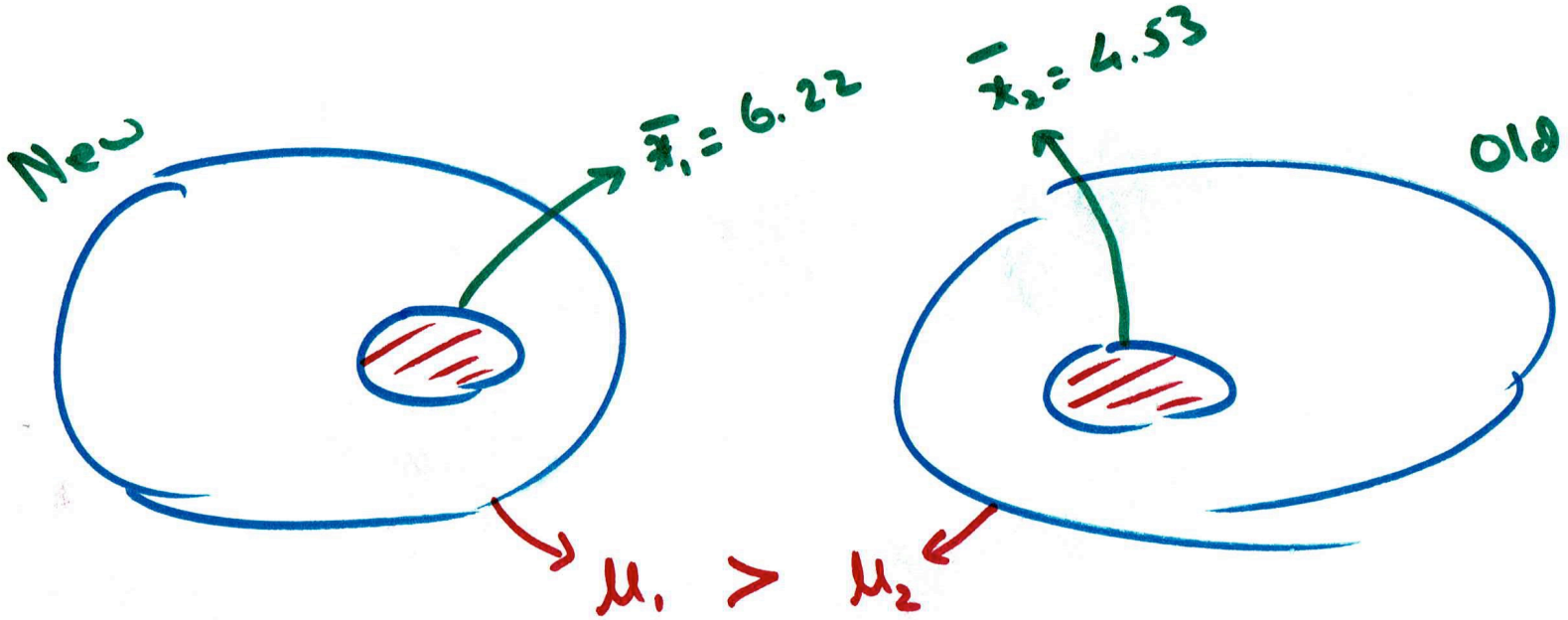
likely → Accept H_0

A/B Testing

- Online news portal aims to expand its business by acquiring new subscribers.
- A team analyzes user behaviors and comes up with a new feature.
- Does this feature really work or help?

user_id	group	landing_page	time_spent_on_the_page	converted	language
546592	control	old	3.48	no	Spanish
546468	treatment	new	7.13	yes	English
546462	treatment	new	4.4	no	Spanish
546567	control	old	3.02	no	French
546459	treatment	new	4.75	yes	Spanish
546558	control	old	5.28	yes	English
546448	treatment	new	5.25	yes	French
546581	control	old	6.53	yes	Spanish
546461	treatment	new	10.71	yes	French

- Why is this a difficult problem?



Assume: $H_0: \mu_1 = \mu_2$

$H_a: \mu_1 > \mu_2$

evidence: $\bar{x}_1 = 6.22$ $s_1 = 1.82$ $n_1 = 50$

$\bar{x}_2 = 4.53$ $s_2 = 2.58$ $n_2 = 50$

Question: what are the chances of seeing $\bar{x}_1 - \bar{x}_2 > 1.69$ if $\mu_1 - \mu_2 = 0$ is true

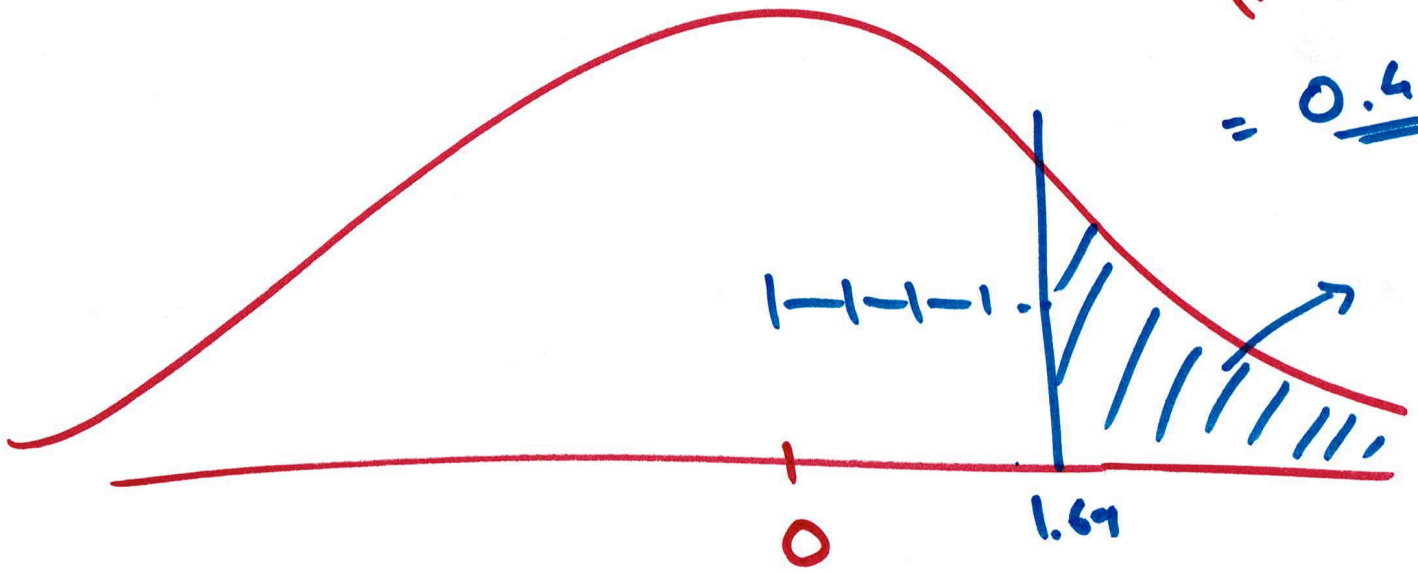
dist of $\bar{x}_1 - \bar{x}_2$

↓

t-dist = $n_1 + n_2 - 2$: dof

$$s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= 0.4465$$



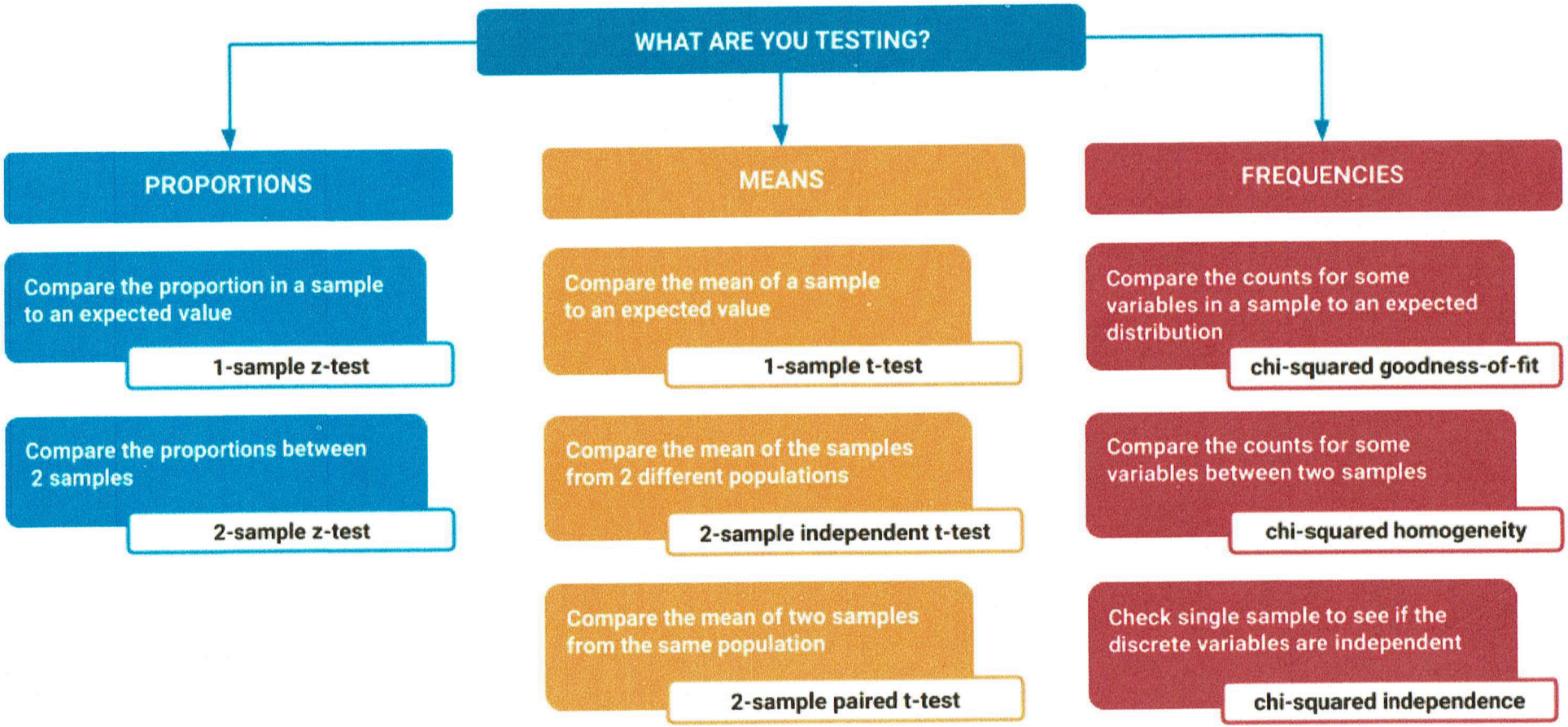
dist of $\bar{x} = \frac{s}{\sqrt{n}}$

$$t = \frac{1.69}{0.4465} = 3.786$$

P-value = 0.000278

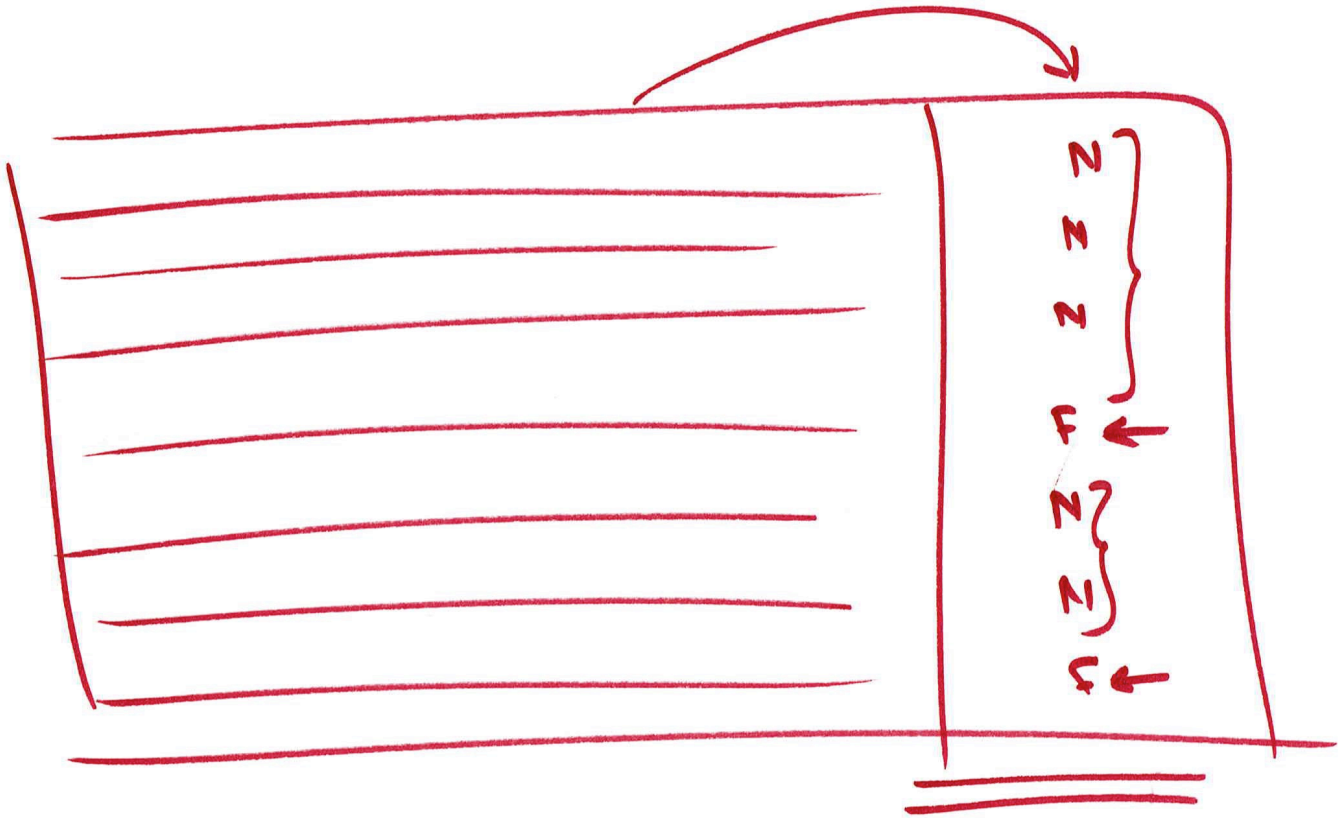
Simple Hypothesis Testing

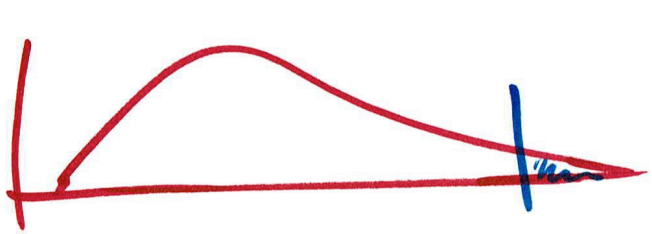
Choosing a simple test for comparing differences in populations

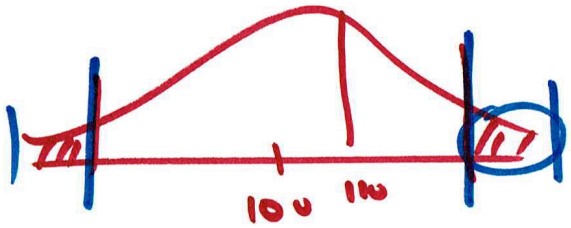


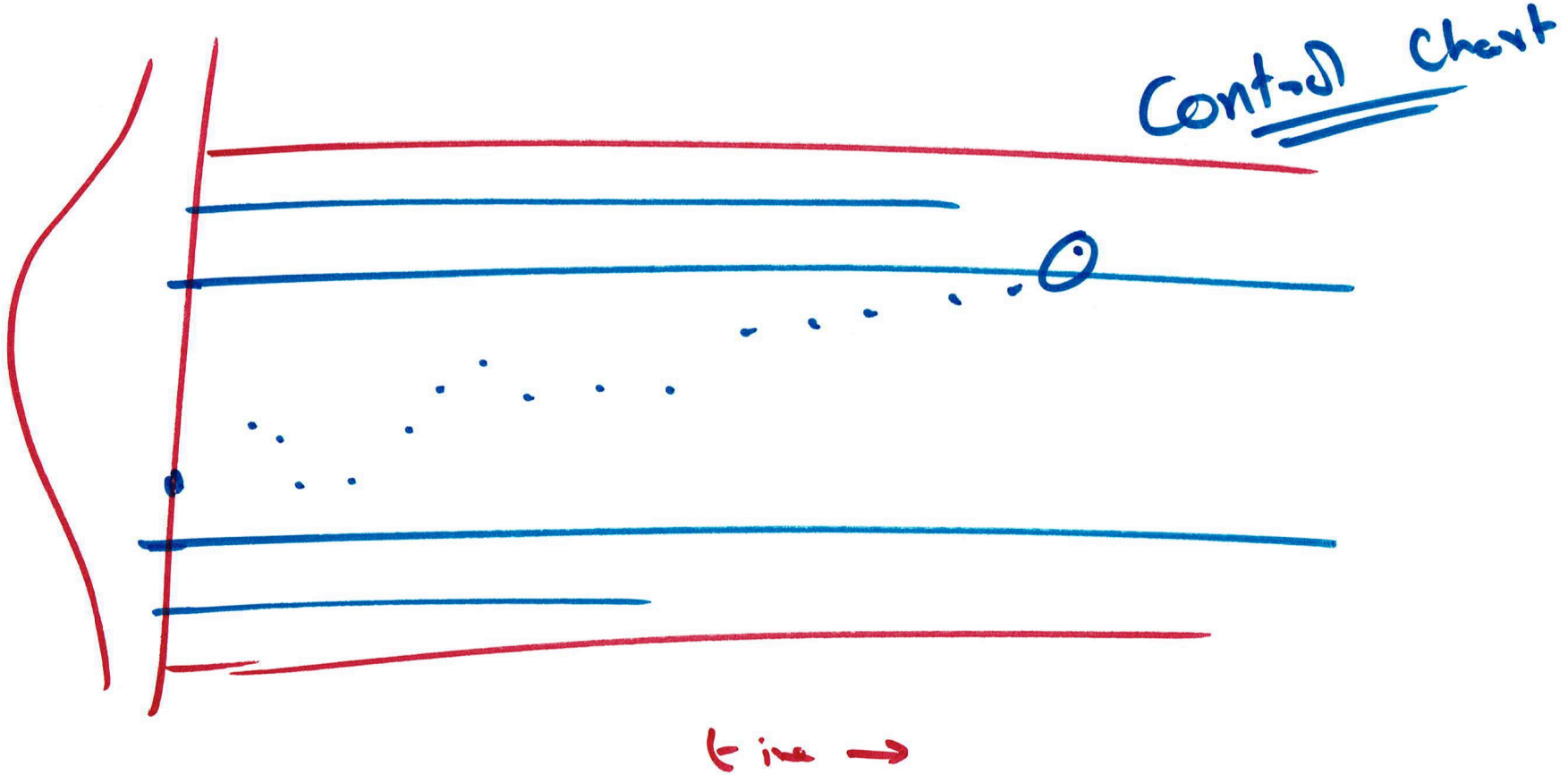
Anomaly Detection

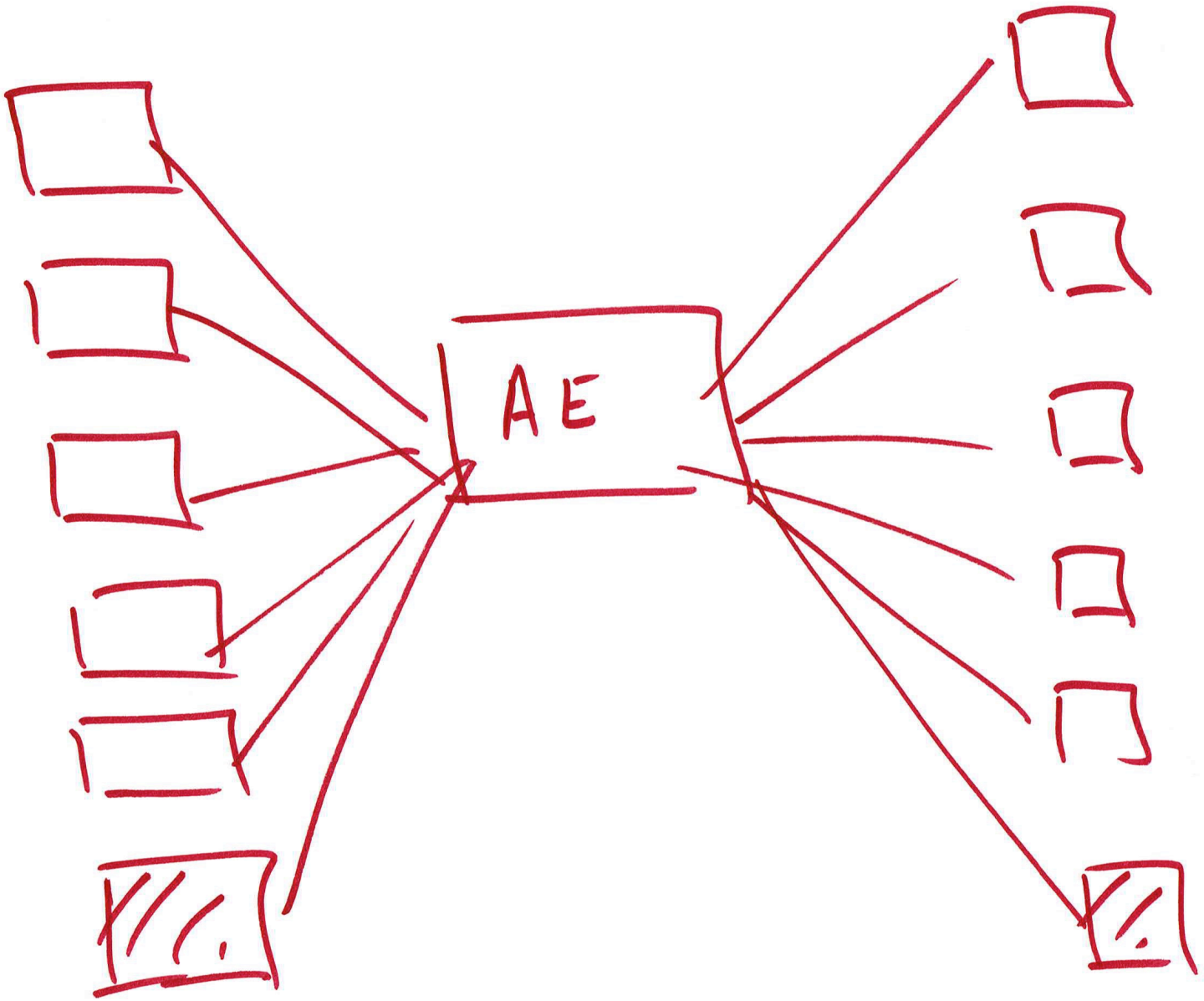
- Identify rare events
- Several applications: intrusion, fraud, fault, health, computer vision, etc
- Change detection is a closely related problem
- Methods:
  - Hypothesis testing based methods
    - Control charts
  - Most machine learning techniques can be adapted
    - Vision examples use autoencoders

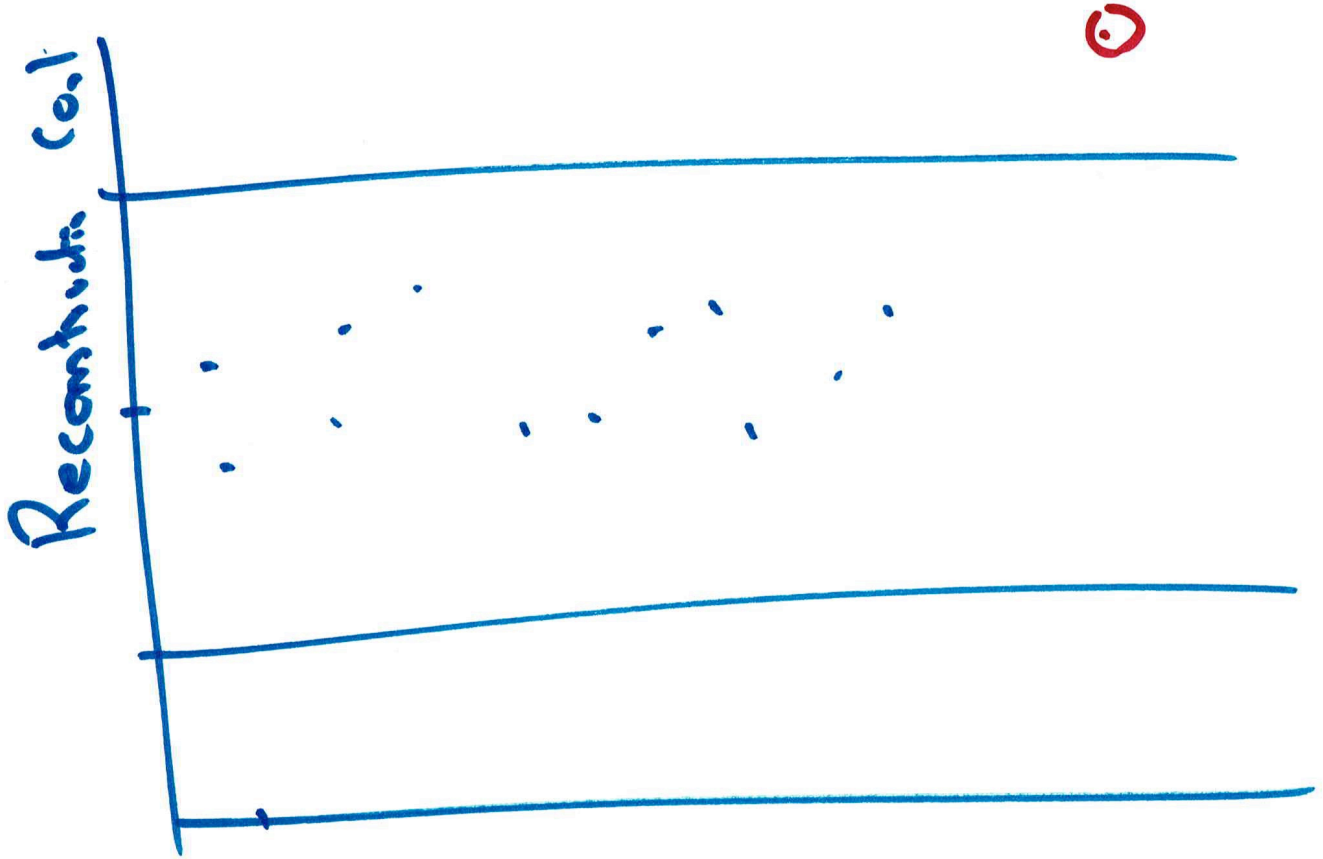

N
N
N
F
N
N
F




LSL
USL


Cont-d Chart
time


AE


Recombination