What's up with our datasets? What's the way forward?

Spurious cues

Not really learning target task

Toxic content

Representational harms, social biases

Management issues

Legal issues, poor consent

Filter data

Spot issues algorithmically

Reformulate tasks

Careful collection & documentation

Datasheets, participatory approaches