Goals

Introduction

We are analysing a data set derived from an RNAi screen. Here are some facts:

- genome-wide knockdown screen
- multiple pathogens
- imaging data with roughly 500 – 700 features

As usual with knockdown screens we should not hope to get too much signal here, so what we could answer is explained in the following.

Questions related to knockdown

- What do knockdowns influence? What kind of change do they induce?
  - Numbers of cells?
  - Singleton ration? Density if cells?
  - Sizes of cells?
  - Formation of clusters? Characterization of clusters

None of these resonates some interest. We could try to model these features generatively:

- estimation of full joint distribution? With this we could ask questions like:

$$P(\text{infection} \mid \text{sirna} = s)P(\text{transfection} \mid \text{sirna} = s)P(\text{size} \mid \text{sirna} = s)\dots$$

* Generative models using GANs: with this we could try to estimate a model to create these pictures. Would the noise model be able to reveal information about the biological processes?

Questions related to pathogens

Since the data is largely crappy, we should probably focus on the *B. henselae* screens, Turns our that *B. henselae* is the most interesting pathogen anyway due to its capacity of two different routes of infection (see wikipedia).

- What are Bartonella's drivers for infection?
- Is there a specific way how Bartonellas enter cells? Does it depend on the cells?
  - Cell size,
  - cell density,
  - cell shape,
  - other cell features :)
- Do cell features influence secondary/primary ways of infection?
- How can we model this casually?
- Can we model this spatially? This would be a perfect opportunity for Gaussian processes, we however need to clearly we we are after here. McEarlth should be a good guide :)