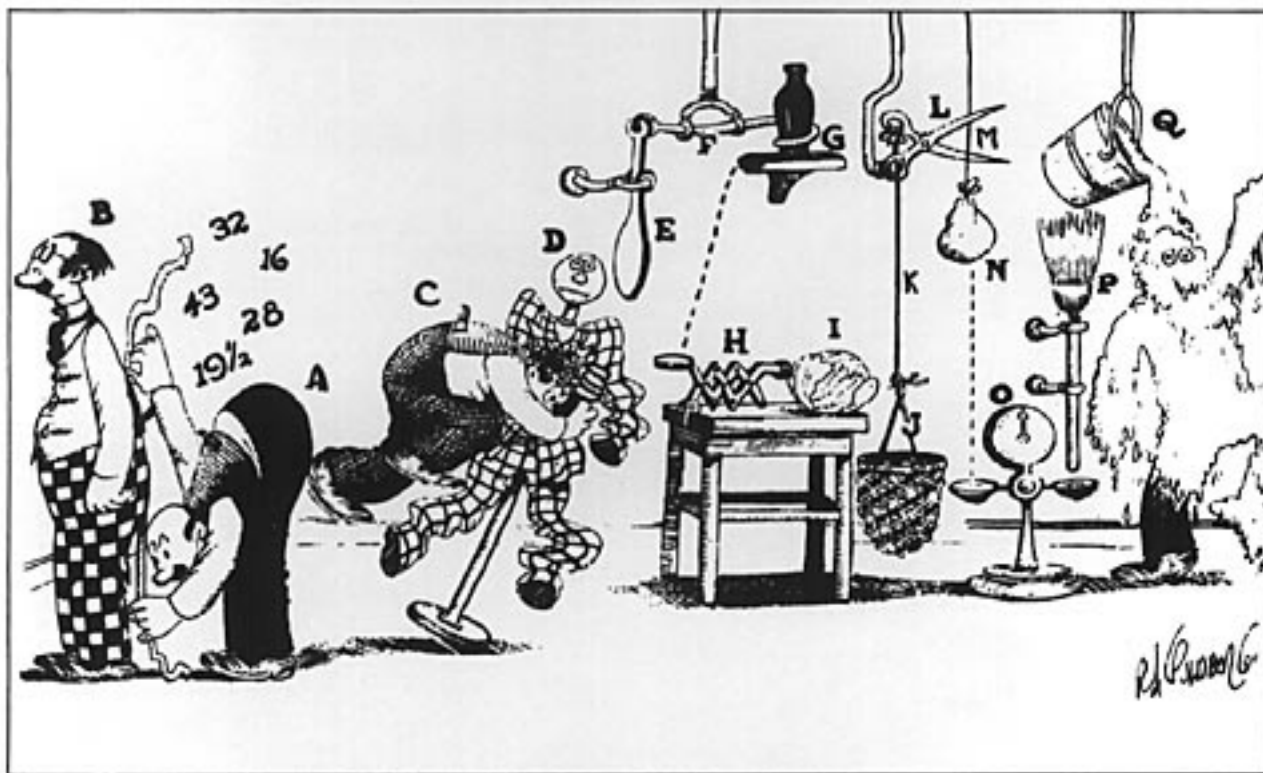


## Working on Cancer



Idea For Dodging Bill Collectors RUBE GOLDBERG (tm) RGI 046

Ola Bini

computational metalinguist & paranoia principal

[ola.bini@gmail.com](mailto:ola.bini@gmail.com)

<http://olabini.com/blog>

698E 2885 C1DE 74E3 2CD5 03AD 295C 7469 84AF 7FoC

# The problem

# Terms

## DNA

A string of bases, usually organized in a double helix

## Nucleobases / bases

One of four molecules, called A, C, G and T

## RNA

A DNA like molecule. Uses A, C, G and U instead.

# Terms

## Protein / Polypeptides / Amino Acids

A protein is a chain of amino acids. It can be enzymes or other types of proteins

## Codon

A set of three bases that gets translated to an amino acid

## Gene

An inherited unit under selection

## Variant

A mutation

# Sequencing

Taking DNA and turning it into bits

## Steps

- Prepare the analyte

- Shred the DNA into 200bp long segments (called *reads*)

- Sequence all the reads separately

- Find overlapping reads (*assembly*)

- Find where the reads belong by comparing to a reference (*alignment*)

- Optional: compare against another genome and output the results (*variant calling*)

The \$1000 genome

# Cancer

Not one disease - at least 10 000 diseases

Organ of origin less interesting than molecular make up

Cancer is modifications of DNA in various ways

- Stops apoptosis

- Enhances G cell cycle (growth)

- Removes error correcting mechanisms

Through genetic modifications of various kinds

Driver mutations vs passenger mutations

Lots of noise

# The treatment problem

Standard of care is based on organ

Ovarian cancer has ca 3 first level chemo's

If one doesn't work, try the next

But they're expensive: \$100 000 for a round

And 3 months of time

And severe pain and damage to the body

The information is out there

In research papers

In clinical trial data

# The team

# The process

# Our solution

# Our solution

Suck in data from lots of resources

Unify and normalize

Types of data

Patient

Reference

Experience

Put everything in a graph

Model biology

Enhance raw information with deduced information

Connect up treatments in relationships with biomarkers

# Tech stack

Clojure

Neo4J

JRuby

CoffeeScript

Sinatra

Compojure

Jetty

# Graph database

# Infrastructure

# Infrastructure

AWS

Puppet

Boto & Fabric

Custom provisioning code

Ca 12 repositories, all with “go” scripts

Self-installing, using setuptools & virtualenv

# Go

Started with Jenkins

Switched to Go for easy deployment pipelines

Master have to be built from a dev machine

Agents can be added on the fly

# On every deploy...

Provision a new EC2 instance

Copy necessary keys to the new instance

Attach a data volume cloned from snapshots

Install puppet

Copy all puppet manifests to machine

Apply puppet

Install all deployment RPMs

Start servers (Apache, Jetty, etc...)

Associate elastic IP with new box

Terminate old instance

# Monitoring etc

Piwik

Statsd & Graphite

Monit for notifications

Fairly standard logging setup

Status checking using Go

# Data ingestion

# One-page app

# Polyglot architecture

# Internal DSLs

# Conclusions

Small teams win over large teams

Use the right language

Molecular biology is very cool and needs more work

Continuous Delivery is a must

This approach to cancer will likely work for the next 10 years

# Questions?

**OLA BINI**

**ThoughtWorks®**

<http://olabini.com>  
[obini@thoughtworks.com](mailto:obini@thoughtworks.com)

@olabini