KOH Chuan Hock
許泉福

kohchuanhock.com

NUS National University of Singapore

東京大学 THE UNIVERSITY OF TOKYO

Koh and Wong *BMC Systems Biology* 2012, **6**(Suppl 2):S3
http://www.biomedcentral.com/1752-0509/6/S2/S3

BMC Systems Biology

PROCEEDINGS Open Access

Embracing noise to improve cross-batch prediction accuracy

Chuan Hock Koh[1,2*], Limsoon Wong[2]

From 23rd International Conference on Genome Informatics (GIW 2012)
Tainan, Taiwan. 12-14 December 2012

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 27 no. 5 2011, pages 734–735*
doi:10.1093/bioinformatics/btq727

Systems biology Advance Access publication January 5, 2011

MIRACH: efficient model checker for quantitative biological pathway models

Chuan Hock Koh[1,2,3], Masao Nagasaki[3,*], Ayumu Saito[3], Chen Li[3], Limsoon Wong[2] and Satoru Miyano[3]

[1]NUS Graduate School for Integrative Sciences and Engineering, Singapore 117597, [2]School of Computing, National University of Singapore, Computing Drive, Singapore 117417 and [3]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Associate Editor: Trey Ideker

PNAS

Intrinsically disordered proteins aggregate at fungal cell-to-cell channels and regulate intercellular connectivity

Julian Lai[a,b], Chuan Hock Koh[c,d], Monika Tjota[a,b], Laurent Pieuchot[a,b], Vignesh Raman[a,b], Karthik Balakrishna Chandrababu[b], Daiwen Yang[b], Limsoon Wong[c], and Gregory Jedd[a,b,1]

[a]Temasek Life Sciences Laboratory and [b]Department of Biological Sciences, National University of Singapore, Singapore 117604; [c]School of Computing, National University of Singapore, Singapore 117417; and [d]Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 117597

Journal of Bioinformatics and Computational Biology
Vol. 7, No. 6 (2009) 973–990

ICP Imperial College Press
www.icpress.co.uk

SIRIUS PSB: A GENERIC SYSTEM FOR ANALYSIS OF BIOLOGICAL SEQUENCES

CHUAN HOCK KOH*,†,§, SHARENE LIN*,¶, GREGORY JEDD‡,‖ and LIMSOON WONG*,**

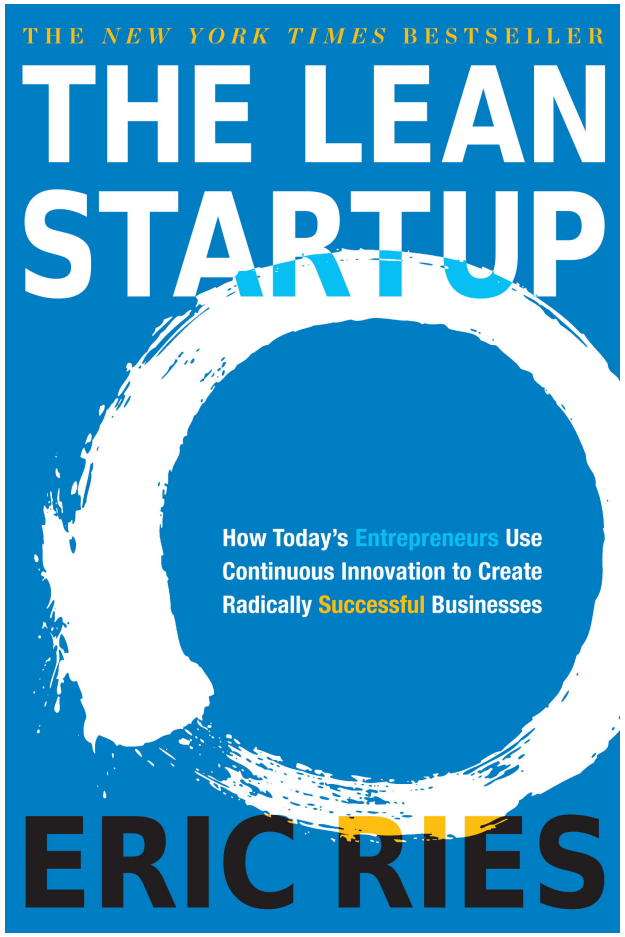
楽天

Rakuten

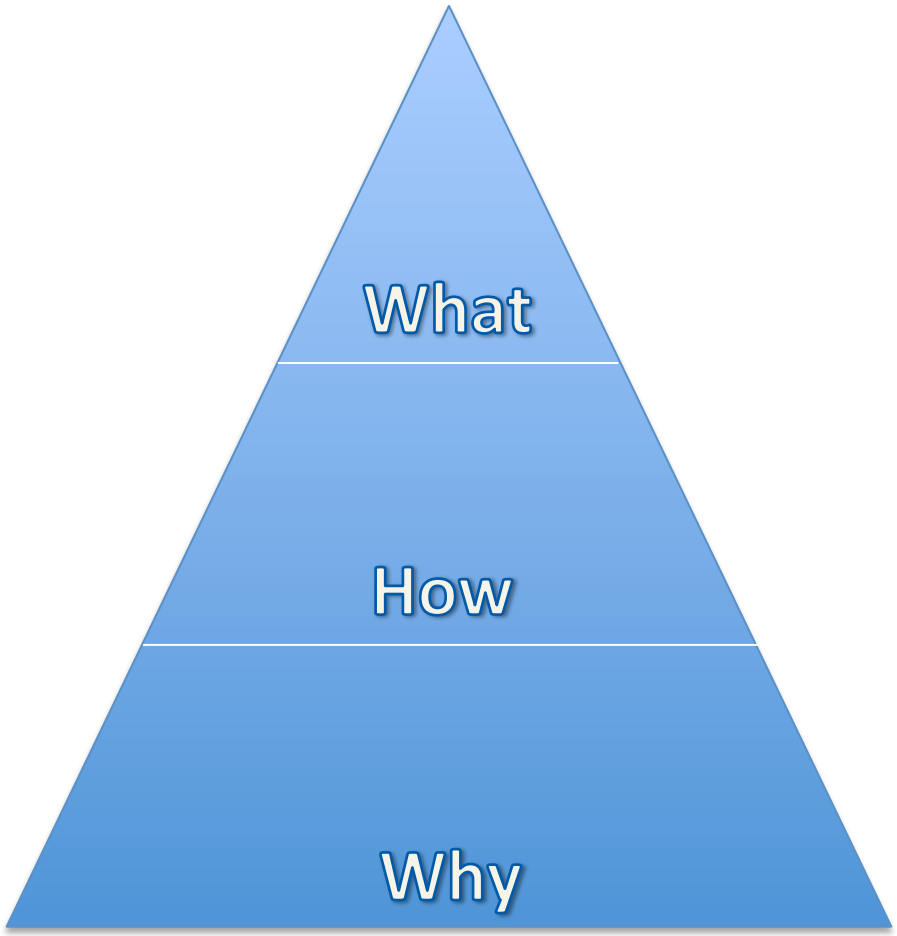
:DeNA

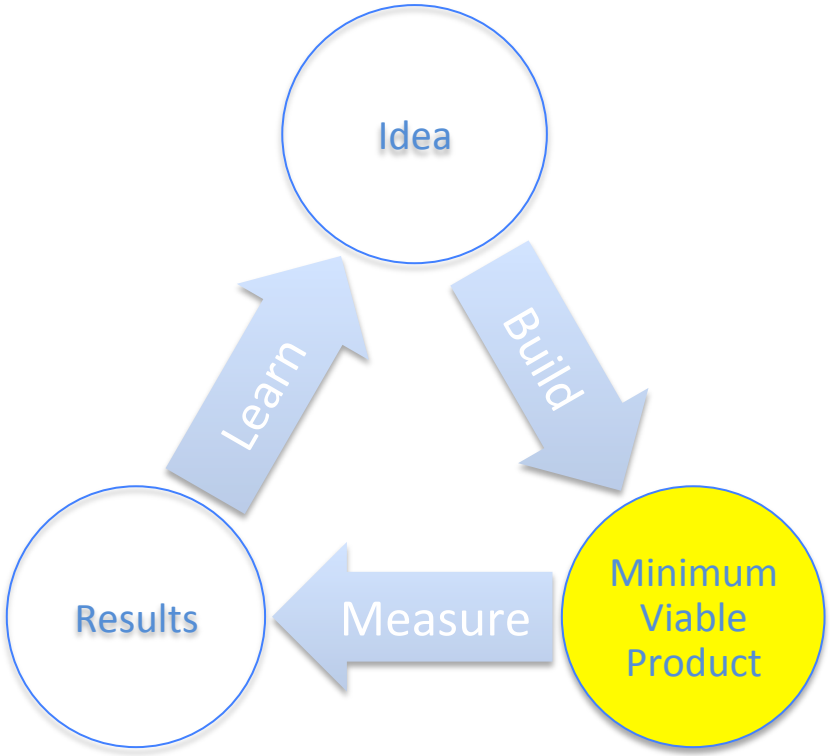
WWW.TRAVELCAFE.ME
Travel. Share. Discover.

Lean Research



The ~~Lazy~~ Lean Startup





Embracing Noise in Bioinformatics

A Thesis submitted for the degree of

Doctor of Philosophy

2012

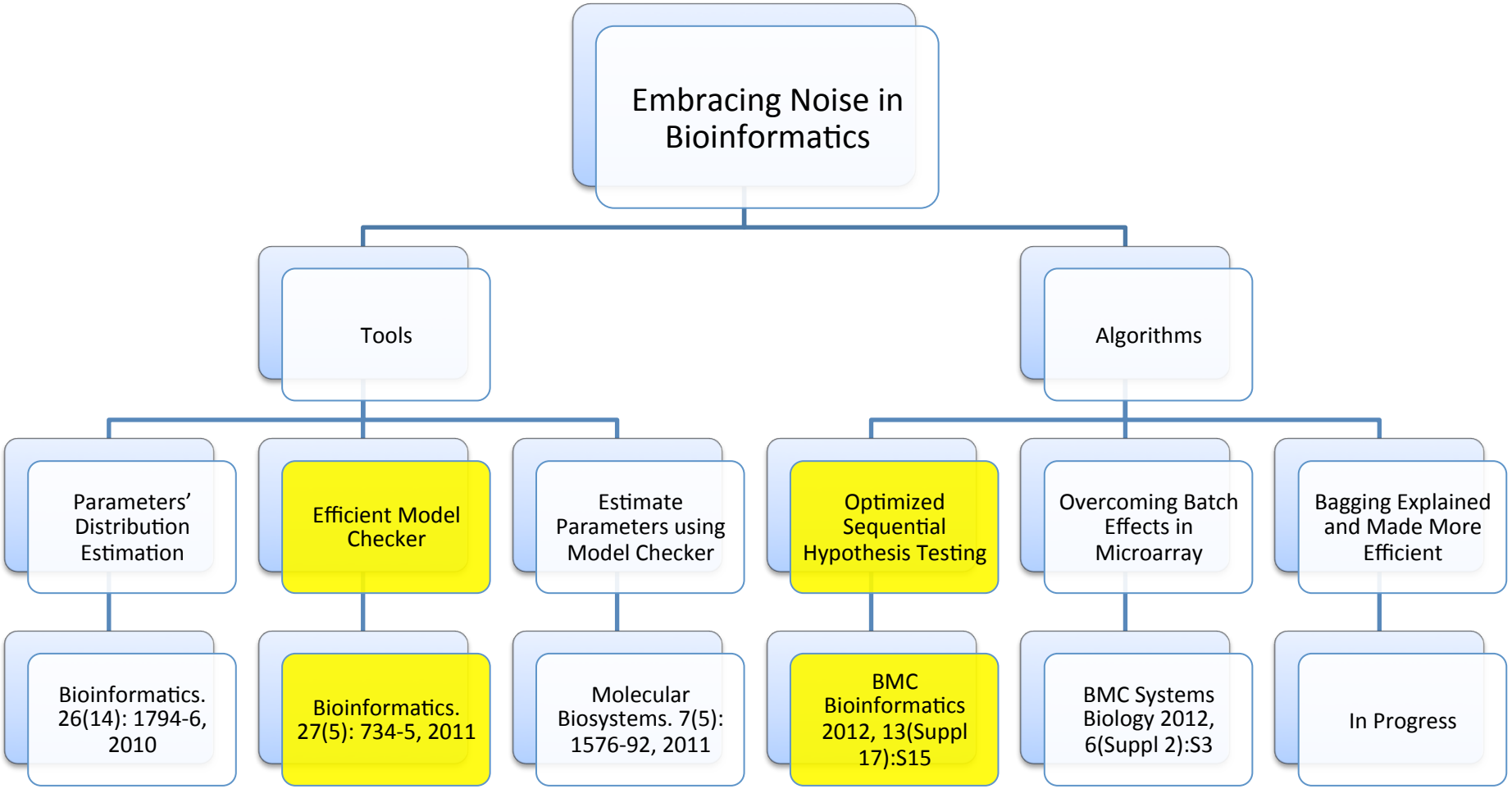
Why?

- Because it's a noisy world
  - Experimental Noise
    - Random
    - Systematic (Batch effect)
  - Inherent Noise
    - Intrinsic (Within Cell)
    - Extrinsic (Between Cell)

How (to handle noise)?

- Measure and remove
  - Increase sample size
  - Better algorithms
- Embrace
  - Recognize and accept noise

What?


Embracing Noise in Bioinformatics
Tools
Algorithms
Parameters' Distribution Estimation
Efficient Model Checker
Estimate Parameters using Model Checker
Optimized Sequential Hypothesis Testing
Overcoming Batch Effects in Microarray
Bagging Explained and Made More Efficient
Bioinformatics. 26(14): 1794-6, 2010
Bioinformatics. 27(5): 734-5, 2011
Molecular Biosystems. 7(5): 1576-92, 2011
BMC Bioinformatics 2012, 13(Suppl 17):S15
BMC Systems Biology 2012, 6(Suppl 2):S3
In Progress

Model Checking

- Why
  - Complexity of simulation models are increasing
  - Need to validate the model
    - Does the simulation model exhibit certain behaviors
    - E.g. P (Survive) > 0.5
  - Current model checker are inadequate
    - Not scaleable
- How
  - Integrate it with a simulation engine
- What
  - Implement a model checker + simulation engine

Current Model Checkers

- MC2 (Donaldson and Gilbert, 2008b)
  - An offline model checker
  - Independent of the simulation model
    - Only needs simulation results

Comparison

- MC2
  - Checks after simulation completes
  - Only needs simulation results
    - Able to do checking on existing traces and biological experiments results
- MIRACH
  - Checks as simulation runs
  - More efficient in terms of running time

Minimum Viable Product

Comparison

- Using Levchenko *et al.* (2000) model
  - 22 entities (nodes)
  - 30 reactions (edges)

*in seconds	**100 Samples**	**1000 Samples**
MC2 (Donaldson and Gilbert 2008a)		
Initialization	12.14 (0.40)	107.95 (1.52)
Checking	10.13 (0.29)	88.58 (1.11)
Total Time	22.27	**196.53**
MIRACH		
Initialization	6.85 (0.24)	6.86 (0.31)
Checking	5.34 (0.20)	40.74 (0.90)
Total Time	12.19	**47.6**

Bioinformatics. 27(5): 734-5, 2011

How many samples?

How many samples?

- Exact or Approximative
  - Exact explores all possible states
  - Approximative does sampling

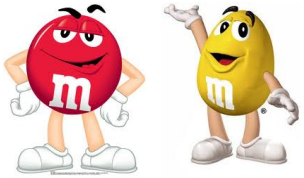
- Biological systems are inherently noisy
  - Have infinite possible states
  - Requires approximative approach

How many samples?

Can you design a vending machine for students that will..

- Randomly dispenses Red / Yellow M&M
- But half the time, red M&M should be dispensed (i.e. probability of red M&M = 0.5)

School management



Sure. No Problem!

Supplier

How many samples?

Great! Let me test it first ……

School management

Here it is! The vending machine as you wanted it!

Supplier

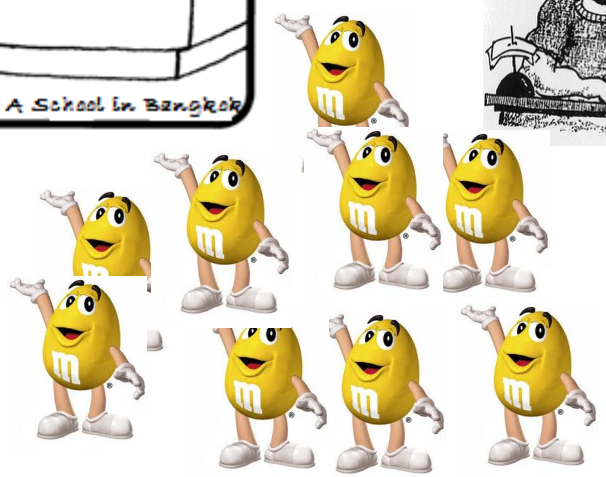
How many samples?

Argh! This is not what I wanted! I put 10 coins in and only one of the candy is red…. There is a problem!

School management

Oops! Sorry there was a problem with our algorithm..

Supplier

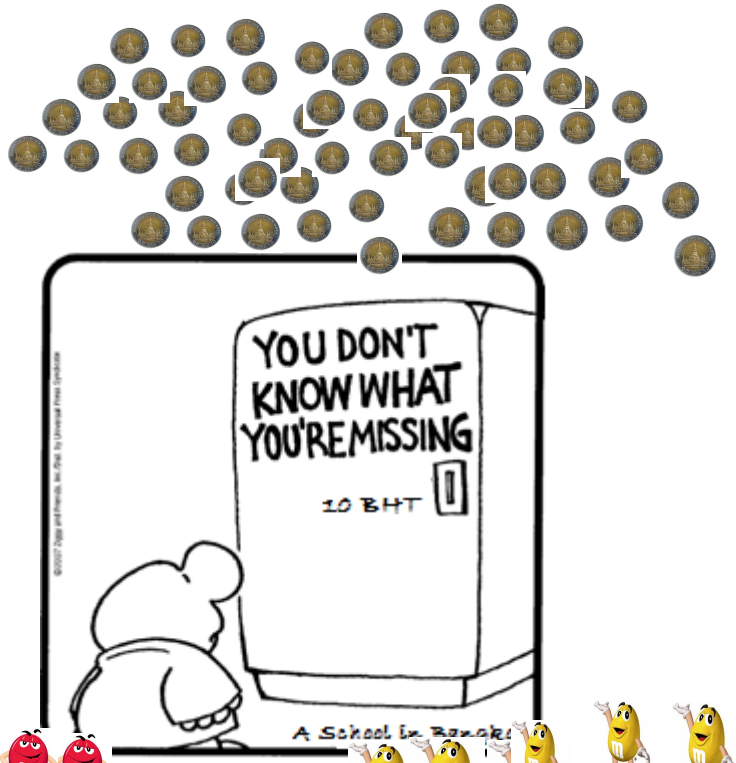


How many samples?

Seems good to me.. I put about 100 coins and 49 were red and 51 was yellow… most likely it is fine

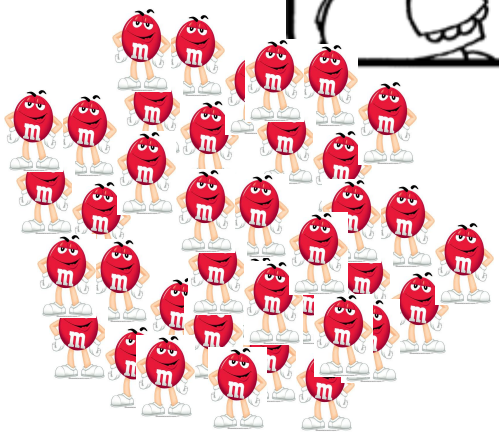
School management

But how do I know for sure? How many samples should I take?



Here is the new vending machine.. Should be good now..

Supplier



How many samples?

- Donaldson and Gilbert (2008)
  - Just sample a fixed number that is assumed to be large enough (10,000)

- Clarke *et al.* (2008)
  - Based on sequential hypothesis testing
  - Sample until enough (with some error bound)
    - by Younes and Simmons (2002)

Sampling Algorithm

- Why
  - Sampling is required to understand stochastic systems
    - Ask probabilistic questions such as P (KOSPI Increase > 0.7)?
  - Current approaches have practical limitations
- How
  - Leverage on current approaches
- What
  - A sampling algorithm that works in all situations

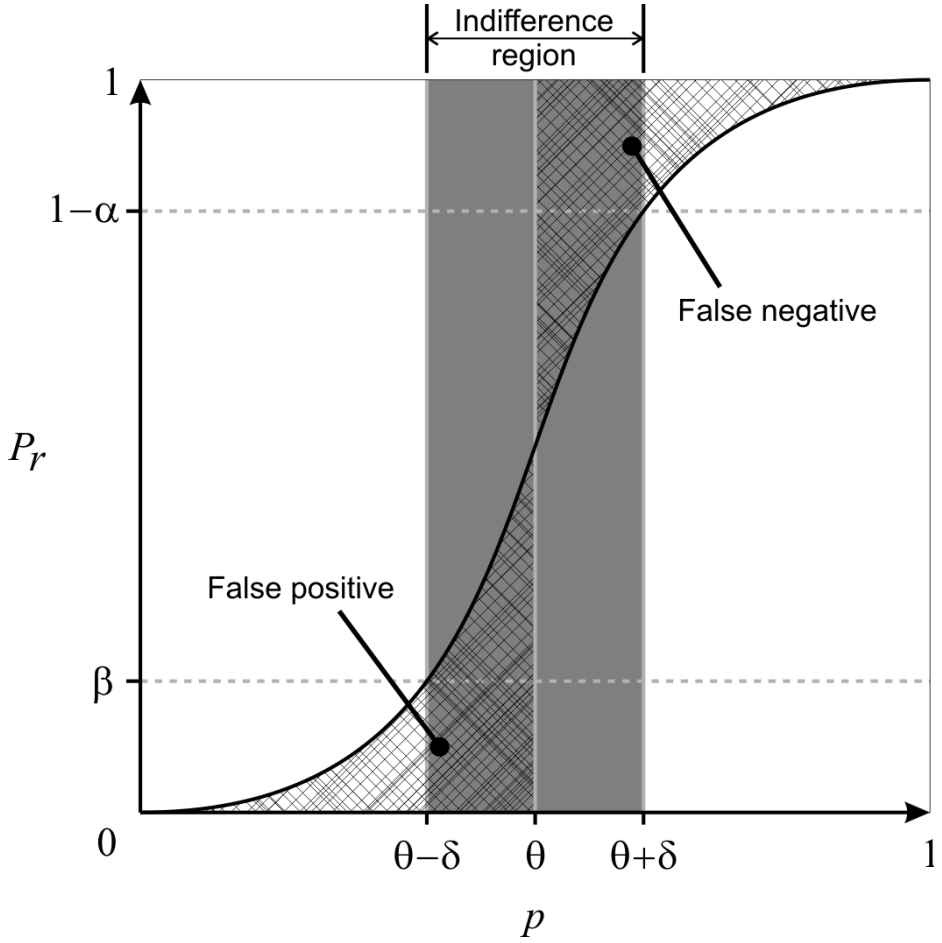
Younes and Simmons (2002)

- Algorithm:
  1. Sample (Simulate)
  2. After each sample, determine if another sample is required or a decision can be made

- Relaxed the standard hypothesis testing from
  - H_0: $p \geq \theta$ vs. H_1: $p < \theta$ to
  - H_0: $p \geq \theta + \delta$ vs. H_1: $p \leq \theta - \delta$
    - $(\theta - \delta, \theta + \delta)$ is known as the indifference region

Younes and Simmons (2002)

- Plus points
  - Guaranteed error rates when p is outside indifference region
- Limitation
  - Error rates are not bounded if p is within indifference region
  - Choice of δ is critical
    - Too small, samples required increases significantly
    - Too large, higher chance for p to be inside indifference region



Minimum Viable Product

Proposed algorithm

- Dynamically select the indifference region
  - Initialize δ to 1.0
  - Half δ based on conditions below
  - Stop when a definite result is returned

- Uses two acceptance tests
  - $H_0: p > \theta$ vs. $H_1: p \leq \theta - \delta$ with $<\alpha, \gamma>$
  - $H'_0: p > \theta + \delta$ vs. $H'_1: p \leq \theta$ with $<\gamma, \beta>$.

  $p \geq \theta$ is accepted as true iff H_0 and H'_0
  $p \geq \theta$ is accepted as false iff H_1 and H'_1
  else half δ

BMC Bioinformatics 2012, 13(Suppl 17):S15

Proposed algorithm

- A sampling algorithm that..
  - Can ask probabilistic questions
    - E.g. P (KOSPI Increase > 0.7)?
  - And obtain "good" decisions
    - Decision with *N* samples = Decision with infinite *samples*
      - *With statistical guarantees on the error rates*

- What can we do with it?

Bagging

1. Given a test instance *T* and a training set *S*
2. Train *N* number of classifiers
   – from bags of *M* instances randomly drawn with repetitions from *S*
3. Predict T to be positive if >N/2 of these classifiers predict T to be positive

- Standard bagging
  – *N* would be arbitrarily fixed at 10, 100 or so

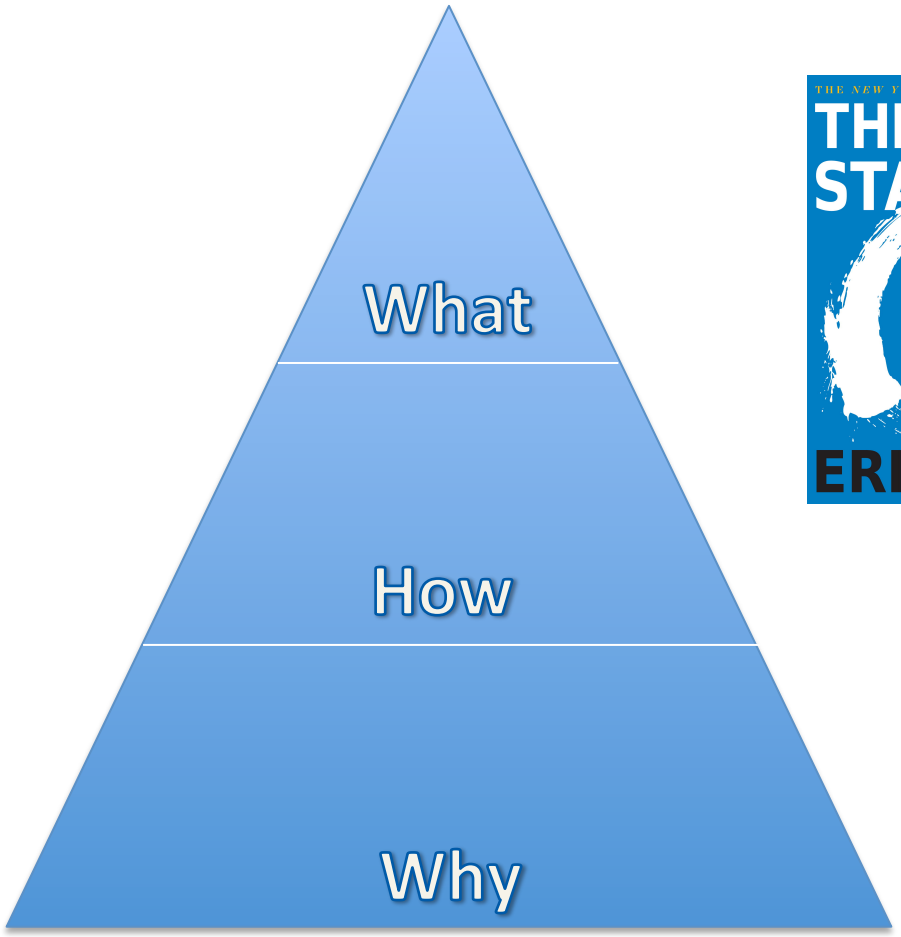
Minimum Viable Product

Dynamic Bagging

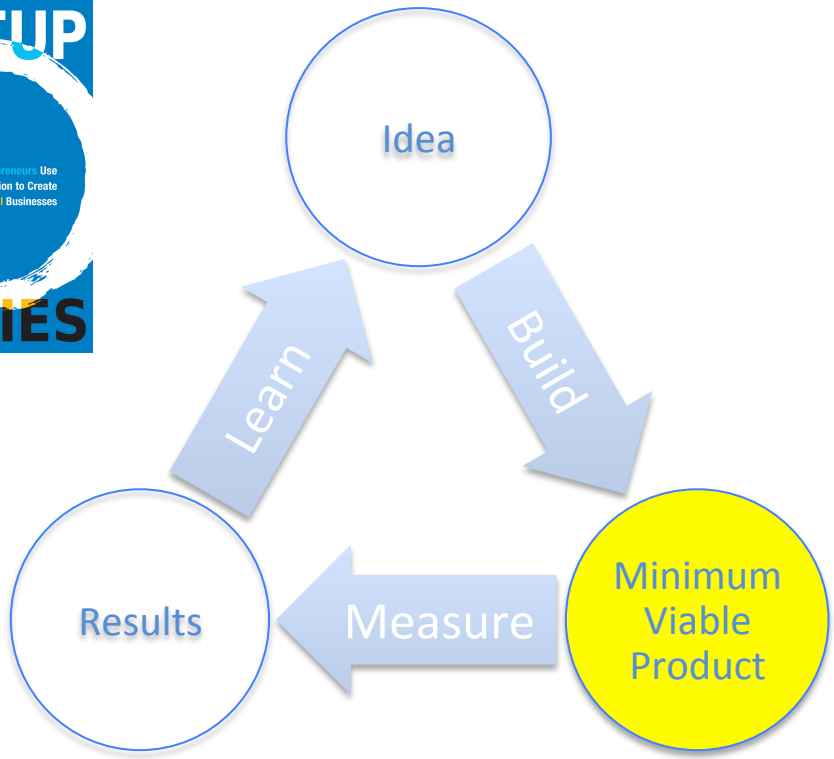
- Will >50% of classifiers predict T to be positive?
  - P (T to be predicted as Positive > 0.5)?
- Advantages over standard bagging
  - No need to a priori and arbitrarily fix *N*
  - Statistical guarantees on the error rates
    - Decision with N samples = Decision with infinite samples

BMC Systems Biology 2012, 6(Suppl 2):S3

Recap


What
How
Why


THE NEW YORK TIMES BESTSELLER
THE LEAN STARTUP
How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses
ERIC RIES


Idea
Build
Minimum Viable Product
Measure
Results
Learn

Things I will do differently (maybe..)

- Serve the correct “customers”
  - Biologists and medical doctors instead of reviewers

- Create what is needed, not what I can

Thank you!