Lecture #1. January 28 2014.

Administrivia.

Social Network X
CSCI-GA.3033-004.

Rm 1002, 715 Broadway, NY. 10003.
Phone: 212.998-3464
Email: rmishra@nyu.edu.
Office Hrs: By appt. +

Class Hrs. + Class Room:
Tuesdays, 5^10 pm - 7^00 pm. EST.
CIWW (251 Mercer Street). Room 317.

Text Book: John Hopcroft & Ravi Kannan.
Foundations of Data Science.
Unpublished Notes, 2013.
Available on Course Home page.

+ Mark. Newman
Networks: An Introduction
Oxford University Press, 2010.

◇ Introduction

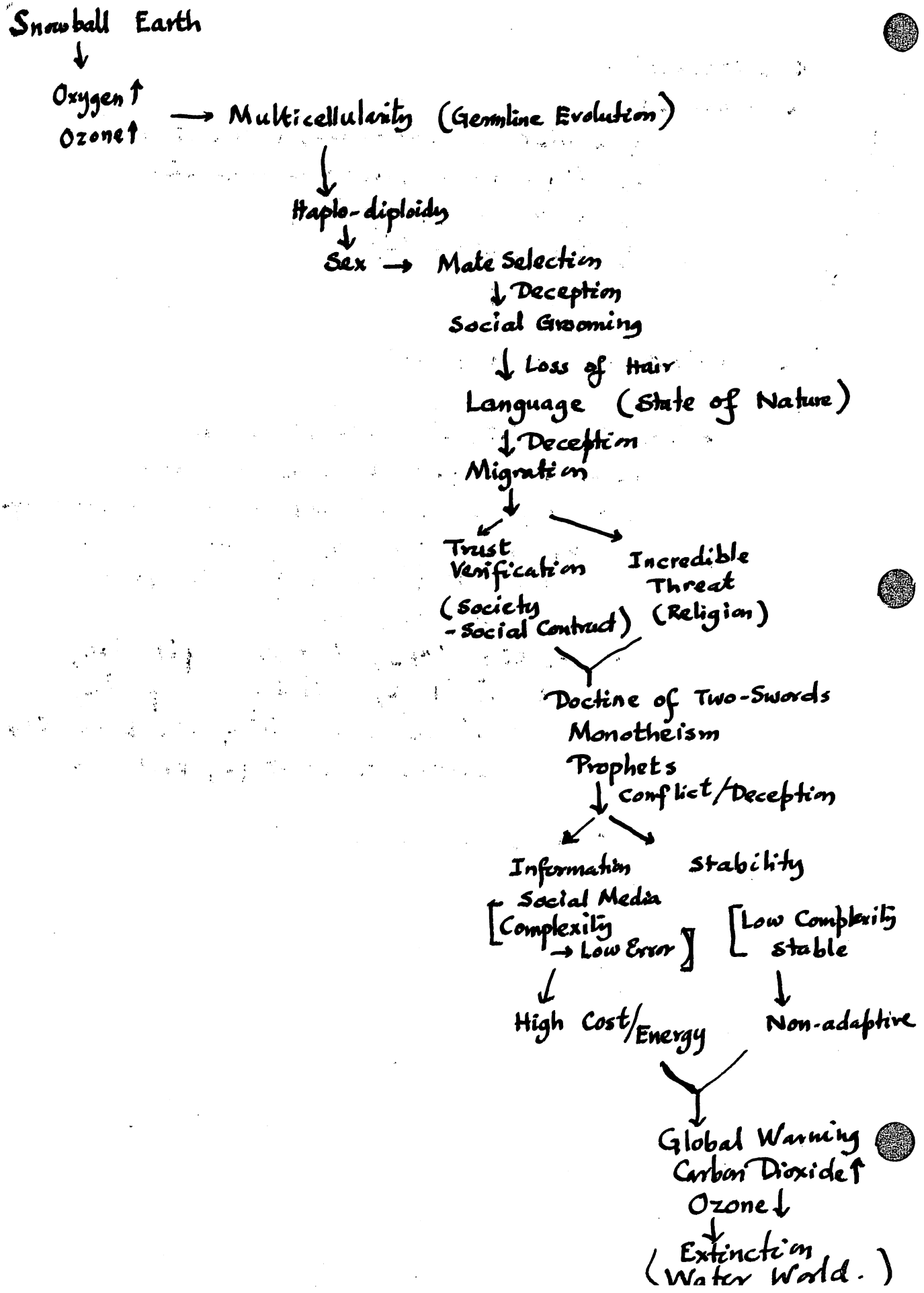
◇ Syllabus:

a) Statistical Inference / Data Science
b) Graph Theory / Network Analysis.
c) Game Theory / Signaling
d) Enterpreneurship.

◇ Motivation: Startup 2.0.

• The Economist, January 18 2014.
Tech Startups: A Cambrian Moment.
(A 14-page Special Report on Tech Startups)
Coming to an office near you...
What today's technology will do to tomorrow's jobs.
- Creating jobs / Venture Capitalists / Accelerators / Building Companies / Business Communities / The Dark Side / Hardware Startups / Platforms

- A Cambrian Moment.


Snowball Earth
Oxygen↑
Ozone↑
Multicellularity (Germline Evolution)
Haplo-diploidy
Sex
Mate Selection
Deception
Social Grooming
Loss of Hair
Language (State of Nature)
Deception
Migration
Trust Verification (Society - Social Contract)
Incredible Threat (Religion)
Doctine of Two-Swords
Monotheism
Prophets
Conflict/Deception
Information
Social Media
Complexity → Low Error
Stability
Low Complexity stable
High Cost/Energy
Non-adaptive
Global Warming
Carbon Dioxide↑
Ozone↓
Extinction
(Water World.)

CAMBRIAN RADIATION:

◇ The first geological period of the Paleozoic Era
542 mya ~ 486 mya.

◇ Evolution of Multicellularity:
- Animals, Plants, Fungi, Brown; Red- and Green-Algae.
- Signaling: Tyrosine Kinases.
MicroRNA...
- Symbiosis: Mitochondria, Chloroplasts, ...
- Cytophagy:
- Sex: Haplo-diploidy.
- Morphology: Organelle, Organs, Organisms...
- Eco-Systems: Eco-System Engineering.

◇ Causes:
- Environmental Changes
~ Snowball Earth (Metastable Climate)
~ Increase in Oxygen Level (Result of photosynthesis)
Area of oxygen absorbing surface
~ Ozone Formation
~ Volcanic Activity of Midocean Ridges
- Calcium + Phosphorus.
- Exoskeleton, skeleton, Teeth, Morphology
~ Ecology
- Food Chain, Arms race among Prey-Predato

⇒ Signaling
⇒ Ecological Networks
⇒ Competition & Cooperation..

ANALOGY.

1) Signaling → Data + Platforms
2) Ecological Networks → Social Networks, P2P, B2B networks, Communities
3) Coopetition → Game Theory, Deception, Security, Privacy. Desiderata

"Why Software is Eating the World?"
Marc Andreessen. WSJ, Aug 20, 2011.

a) Understanding Data Science (Hopcroft + Kannan)
b) Understanding Technology Communities (Silicon Valley vs. Silicon Alley)*
c) Understanding Deception (Cyber Security, Bitcoins, Market Microstructure)

Data

Datum vs Desideratum.
Dare → Desiderare
Give → Desire

Example: Omne Datum Optimum
Pope Innocent II 1139
Order of the Poor Knights of Christ and of the Temple of Solomon (Knights Templar)

"As for the things that you will receive from the spoils, you can ~~put~~ put confidently to your own use, and we prohibit that you be coerced against your will to give anyone a portion of these."
Every Perfect Gift.

Datum ≡ Given, Gift.
→ Input, what is given, known, ...

Gift and Social Contract:
Marcel Mauss (1872-1950) "The Gift"
"Essai sur le Don"
→ Open Source

Gifts → Reciprocal Exchange
Inalienable Bond between Giver & Gift.
Failure to reciprocate →
Results in loss of Social Status.
Social Grooming

Data Economy → Based on Social Contracts
→ Social Networks.

- Social Status / Social Rank / Reputation
- Signaling / Reciprocal Exchange / Trust
- Properties (Alienability) / Valuation, Recommendation
  - ~ Freemium, Open Source, Potlatch Economy, Crowd-Sourcing
  - ~ Deception.

Mapping: Data → Desiderata { Machine Learning from Data, MetaData, Provenance Data.

↓

Information
(Asymmetric Information Game
Between an Informed Sender
and an Uninformed Receiver)

↓

Deception.

Two Examples for Data Science:

a) House Price : Data: a) Lot Size, x_i
(amount of land being conveyed from a buyer to seller)
b) House Price, y_i
(Transaction deed data).

$\mathcal{D} = \{\langle x_i, y_i \rangle \mid i = 1, \ldots, n\}$
← Training Set.

Desiderata: Given an unpriced/new house with lot size x_0, estimate y_0.
y_0 = Your bid price.

b) Cancer Survival: Data: a) Tumor Size, x_i
(E.g. Tis (DCIS), Breast cancer, Ductal Carcinoma in situ Tisx_i, code records only greater dimension.)
→ At detection.
b) Survival Time, y_i
(Time to relapse)
Right-censoring Problem.

$\mathcal{D} = \{\langle x_i, y_i \rangle \mid i = 1, \ldots, n\}$
← Training set

Desiderata: Given a newly diagnosed breast cancer (DCIS) with size x_0, estimate y_0.
y_0 → Determines whether to select mastectomy ?

Simple Models. (Linear Regression)

→ Find the equation of a straight line that provides the best fit for the data point.

$$y = \alpha + \beta x + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

$$y | \alpha, \beta, x, \sigma^2 \sim \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\alpha-\beta x)^2/2\sigma^2}$$

Maximum Loglikelihood Estimator:

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Heuristic Soln:

$$y_i = \alpha + \beta x_i$$

$$\left.\begin{aligned} \Sigma y_i &= \alpha \Sigma 1 + \beta \Sigma x_i \rightarrow \bar{y} = \alpha + \beta \bar{x} \\ \Sigma x_i y_i &= \alpha \Sigma x_i + \beta \Sigma x_i^2 \rightarrow \overline{xy} = \alpha \bar{x} + \beta \overline{x^2} \end{aligned}\right\}$$

$$\therefore \quad \bar{x}\bar{y} = \alpha \bar{x} + \beta (\bar{x})^2$$

$$\overline{xy} = \alpha \bar{x} + \beta \overline{x^2}$$

$$\beta [\overline{x^2} - (\bar{x})^2] = \overline{xy} - \bar{x}\bar{y}$$

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{Cov[x,y]}{Var[x]}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \bar{y} - \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \cdot \bar{x}$$

$$= \frac{\bar{y}\overline{x^2} - \bar{y}(\bar{x})^2 - \bar{x}\,\overline{xy} + (\bar{x})^2\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$= \frac{\bar{y}\overline{x^2} - \bar{x}\,\overline{xy}}{\overline{x^2} - (\bar{x})^2}$$

Are these models good?

◇ Linear relationship between a scalar dependent variable y and a single independent/explanatory variable x may not suffice.

- More explatory variable?
- How many more?

$$y = \theta^T x + \epsilon \quad \left\{ \begin{array}{l} \vec{y} = X\vec{\theta} \rightarrow X^T\vec{y} = (X^TX)\vec{\theta} \\ \hat{\vec{\theta}} = (X^TX)^{-1}X^T\vec{y} \end{array} \right.$$

Problem with underfitting / overfitting.

$$y = \theta^T x + \epsilon \rightarrow \text{Sparse Solution}$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \frac{1}{2}\|\vec{y} - X\vec{\theta}\|_2^2 + \lambda\|\vec{\theta}\|_0$$

Note $\|x\|_p = \sqrt[p]{\sum x_i^p}$

$\|x\|_0 = \sqrt[0]{\sum x_i^0} = \sum \mathbb{1}_{x_i \neq 0} = \#(i \mid x_i \neq 0)$

$\|x\|_1 = \sum |x_i|$

$\|x\|_2 = \left(\sum x_i^2\right)^{1/2}$

$\vdots$

Another Formulation:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \frac{1}{2}\|\vec{y} - X\vec{\theta}\|_2^2 + \lambda\|\vec{\theta}\|_1$$

Lasso Regression
→ Shrinkage + Continuous Model Selection.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \frac{1}{2}\|\vec{y} - X\vec{\theta}\|_2^2 + \lambda\|\vec{\theta}\|_2$$

Ridge Regression
→ Shrinkage to origin
No sparsity is forced.

Computational Complexity: NPC, LP, Linear Algebra...

Are these models sound?

◇ Gauss-Markov Theorem:

In a linear regression model in which errors have
{ expectation zero,
{ no correlation, and
{ equal variances,
the best linear unbiased estimator (BLUE) of the coefficients is given by ordinary least square (OLS)?

- GLS (Generalized Least Squares)

$$\vec{y} = X\vec{\theta} + \epsilon \qquad \epsilon \sim \mathcal{N}(\vec{0}, \Sigma)$$

$$\hat{\theta} = \arg\min_{\theta} \; (\vec{y} - X\vec{\theta})^T \Sigma^{-1} (\vec{y} - X\vec{\theta})$$

$$\hat{\theta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \vec{y}$$

- Knowledge about Relationship among
  explanatory variables (e.g. gene network)
  or dependent variables (e.g. neighbour's prices)

  ~ Graph Laplacians
  ~ Kriging / Cokriging
  ⋮

Model Selection?

◇ Model Complexity (Rate) vs Sample Error (Distortion)

Connection to Rate Distortion Theory (RDT)
Information Theory.

- Regression, Logistic Regression, Ridge-, Lasso, ...
- Support Vector Machines
- Ensemble Methods
- PCA, ICA, ...
  ⋮

There will still be some issues with our housing and cancer models...

Housing Model:

◇ Least square error → ?
Need to model mortgage loans
→ Default, Prepayment
→ Credit rating, Interest Rates
→ Subprime, Regular, Jumbo
→ Derivative Market
Risks, Liquidity. hedonic

◇ Do houses have an intrinsic ~~hedonistic~~ price?

◇ Can it be manipulated?
Gentrification.
Securitization ...

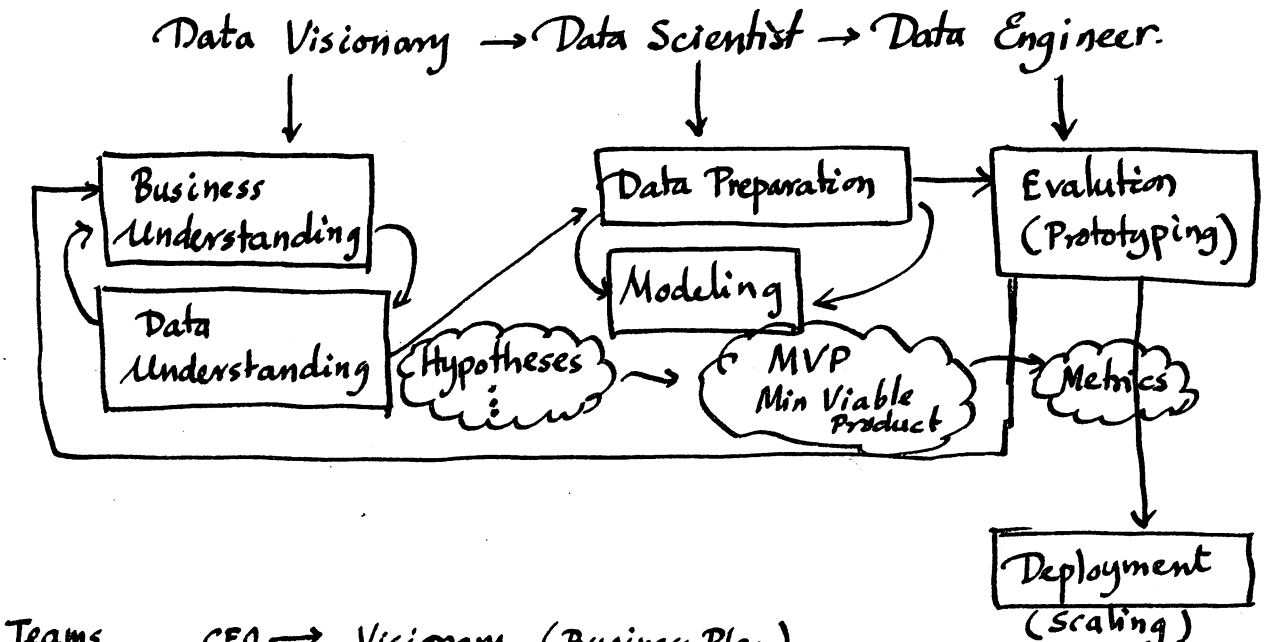
Cancer Model ?

- ◇ Gene Network, gene-gene interaction.
- ◇ Genomic Instability, copy number, syntenic correlations.
- ◇ Multiple Hypothesis Testing
  - Driver mutations.
    - Oncogene vs Tumor Suppressor Genes.
    - Dominant vs Recessive.

Examples: BRCA1
Age effect
LOH, 2hit.

- ◇ Progression Models. Causality.
- ◇ Population Stratification / Age / covariates
  - P73 vs P53.

Projects : Structure :



Teams
- CEO → Visionary (Business Plan)
- CSO → Scientist (IP/Design)
- CTO → Engineer
- CIO → Deployment
- CFO → Deployment.