New Data Science – research methods studying innovation and development

AFRICALICS 2018 Academy, MArrakesh, Morocco

October 1, 2018

Daniel S. Hain

dsh@business.aau.dk

Assistant Professor
Department of Business and Management

AALBORG UNIVERSITY
DENMARK

About me

- Assistant Professor at the IKE group
- Main fields of interest:
  - Economic Complexity, Network Theory
  - Applied Econometric„ Datascience, ML&AI

A brief CV

- Diploma in “Industrial Engineering” (University of Stuttgart)
- Followed by some experience in consulting (Mercedes Benz) in process and supply chain optimization, supplier auditing
- Ad-Hoc joint Master IBE (Hohenheim) + MIKE (Aalborg) → Afterwards PhD. at the IKE group
- Visiting Research Scholar at Stanford → Obsession with Complex System & Networks research.
- Awarded PhD. in late 2015. Thesis: “The Network Dynamics of Financing Technological (R-) Evolution: The Case of Technological Change in the Renewable Energy Area”
- Since: Various research projects and publications around the development and application of novel methods in Innovation Studies.
- Recently: Winner of the 1^{st} OECD IDPSM “Big Data Analytics Challange”
- Recently: Co-founder and coordinator of the “Social Data Science” program at AAU

Agenda

New Problems and Dynamics

New Data Sources (and Types)

New Methods
- ML&AI 101
- Supervised ML
- Unsupervised ML
- Some own examples

New Ways of Teaching Data Science

New Problems and Dynamics

SSA's emerging entrepreneurial ecosystems in the digital economy

OKHi: App that overcomes lack of physical address

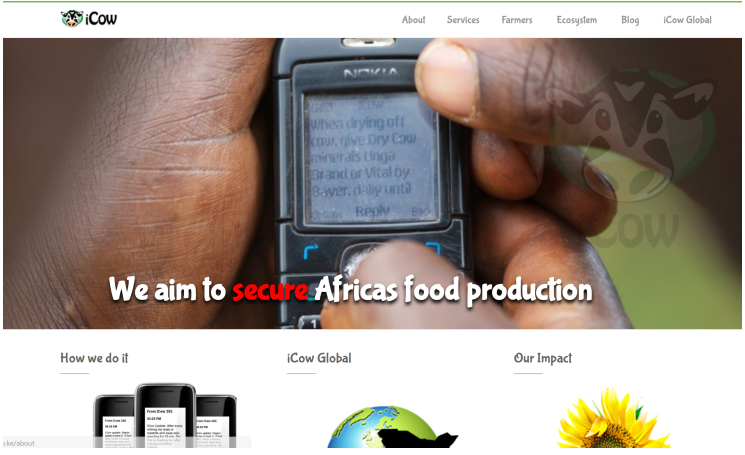
BRCK: "backup generator" for the internet

Zipline: Drone-delivery of medicine

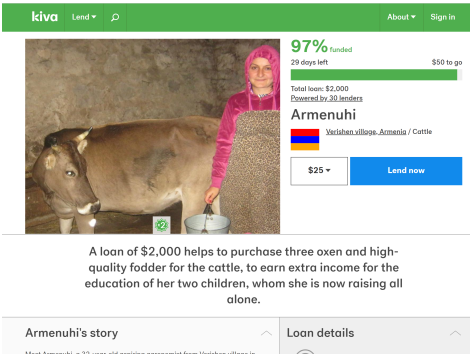
Sendy: Uber-like on demand delivery

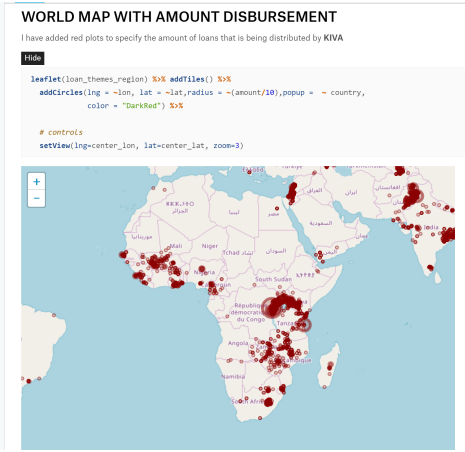
Ushahadi: Open source software for information collection, visualisation, and interactive mapping

SSA's inclusive innovation start-ups

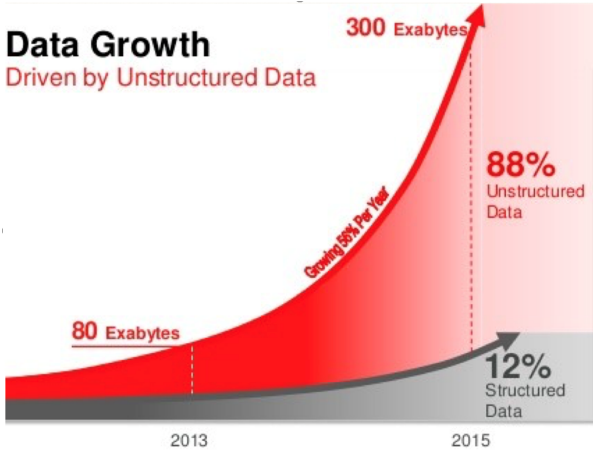


Crowdfunding & Social Lending





New Data Sources (and Types)



New Data Sources
Unstructured Data

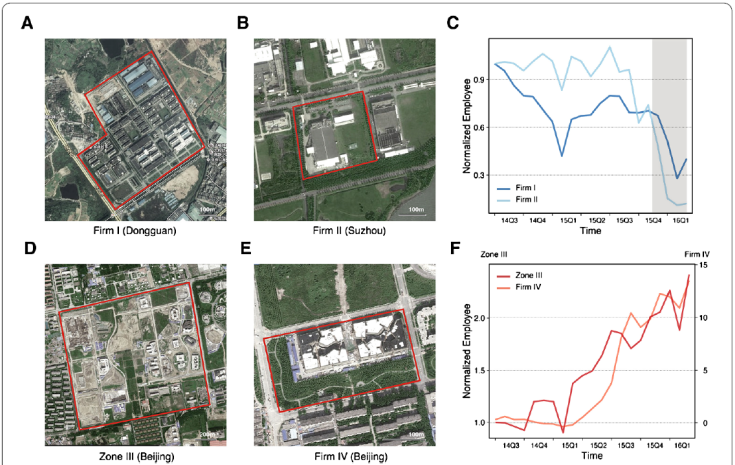
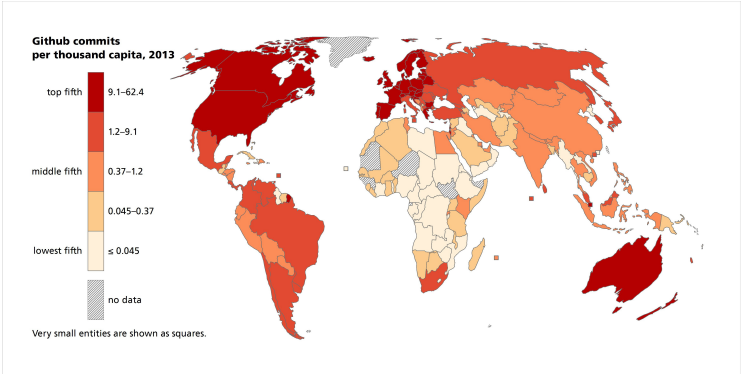


Figure 2 Measuring employment changes at the company/zone level. We study four typical cases, Firm I **(A)** and Firm II **(B)** announced mass layoffs, and Zone III **(D)** and Firm IV **(E)** experienced rapid growth. The normalized employee data are shown in **(C)** and **(F)**. The closing times for Firms I and II are marked by the grey bar in **(C)**. Remote sensing images were derived from Baidu Maps.

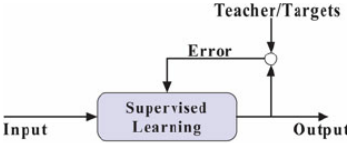


New Methods

"*Field of study that gives computers the ability to learn without being explicitly programmed.*" – Arthur Samuel, 1959

Supervised ML

- Input-Output (A→B) mapping
- Input and output are known.
- Computer task: Learn how A causes B

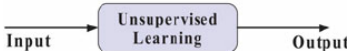


Semi-Supervised ML

- Actions A cause feedback B (reward)
- Computer task: Try different A, learn which brings highest reward

Unsupervised ML

- A is known, B not (maybe there doesn't exist right/wrong B)
- Computer task: Find meaningful pattern in A



The Econometric Approach

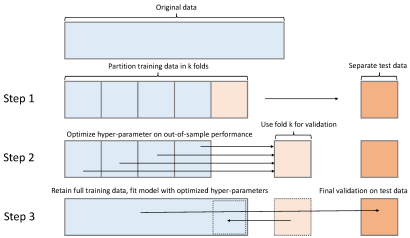
- Mostly interested in producing good *parameter estimates*: Construct models with unbiased estimates of β, capturing the relationship x and y.
- Supposedly "structural" models: Causal effect of directionality $x \rightarrow y$, robust across a variety of observed as well as up to now unobserved settings.
- How: Carefully draw from theories and empirical findings, apply logical reasoning to formulate hypotheses.
- Typically, multivariate testing, *cetris paribus*.
- Main concern: Minimize standard errors ϵ of β estimates.
- Not overly concerned with overall predictive power (eg. R^2) of those models, but about various type of endogeneity issues, leading us to develop sophisticated *identification strategies*

The ML&AI Approach

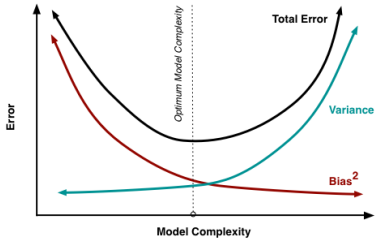
- To large extend driven by the needs of the private sector → data analysis is gear towards producing good *predictions* of outcomes.
  - Recommender systems: Amazon, Netflix, Sportify ect.
  - "Risk scores": Eg.g likelihood that a particular person has an accident, turns sick, or defaults on their credit.
  - Image classification: Finding Cats & Dogs online
- Often rely on big data (N,x_i)
- Not overly concerned with the properties of parameter estimates, but very rigorous in optimizing the overall prediction accuracy.
- Often more flexibility wrt. the functional form, and non-parametric approaches.
- Yet: No "build in" causality guarantee → verification techniques.

Partial Solution: Out-Of-Sample Validation

1. Split the dataset in a training and a test sample.
2. Fit you regression (train your model) on one dataset
3. Optimal: Tune hyperparameters by minimizing loss in a validation set.
4. Optimal: Retrain final model configuration on whole training set
5. Finally, evaluate predictive power on test sample, on which model is not fitted.



Overfitting: The main challenge

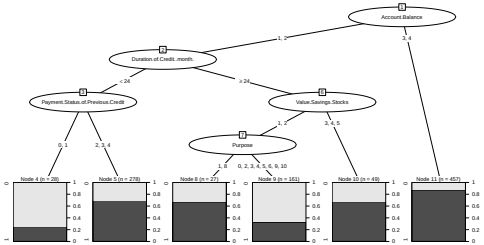


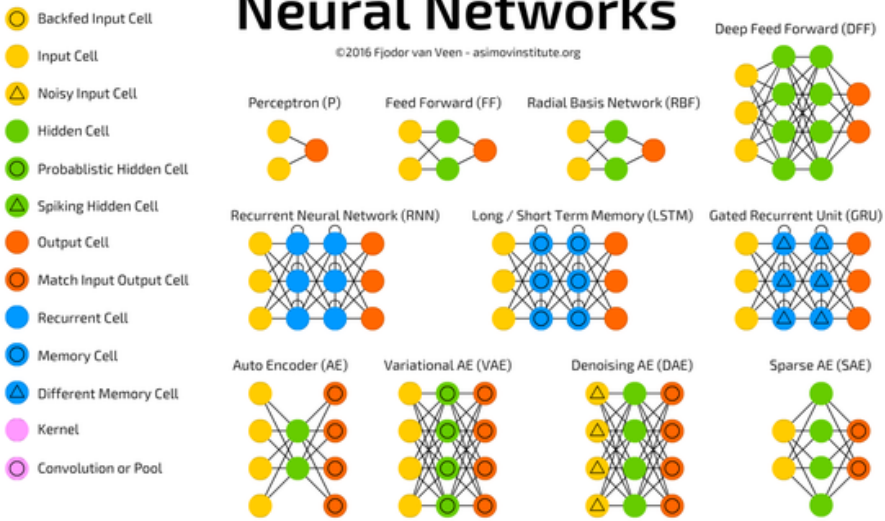
$$\underbrace{\text{minimize} \sum_{i=1}^{n} L(f(x_i), y_i),}_{\text{in-sample loss}} \text{ over } \overbrace{f \in F}^{\text{function class}} \text{ subject to } \underbrace{R(f) \leq c.}_{\text{complexity restriction}} \quad (1)$$

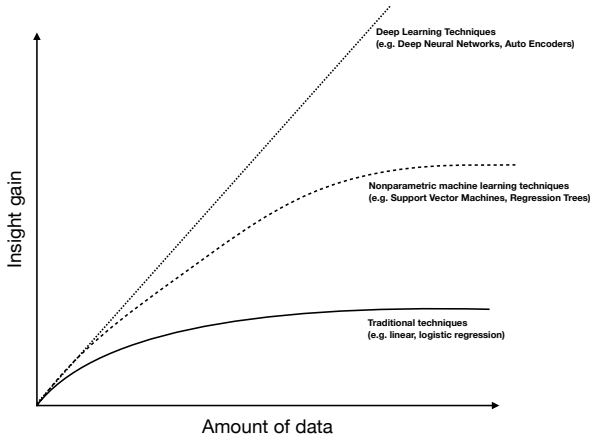
Regression & Classification Trees

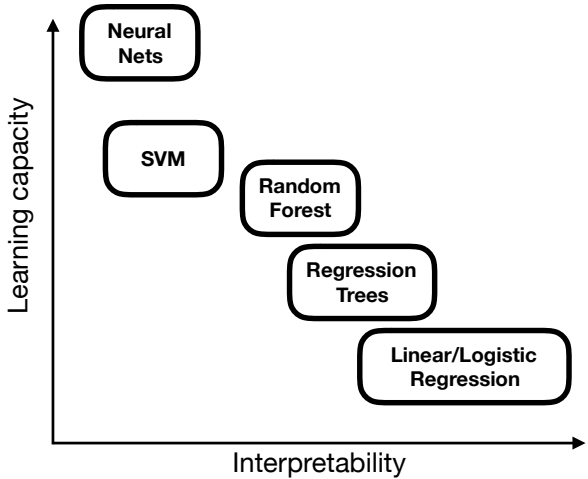
- Mostly used in classification problems on continuous or categorical variables.
- Idea: split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.
- Repeat till stop criterium reached. Leads to a tree-like structure.

Figure: A regression-tree example on credit defaults



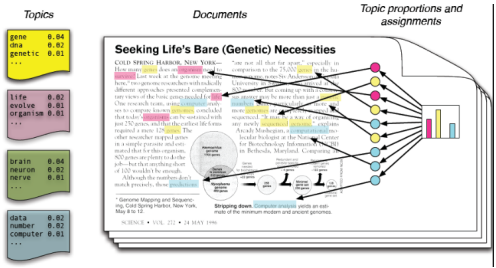




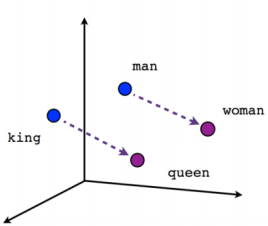


How would you turn text into numbers, preserving semantic features?

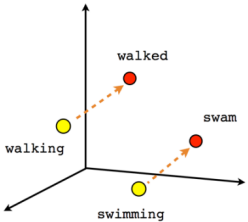
- One-hot encoding $0, 1, 0, 1, 0, 0$
- Term Frequency-inverse document frequency TF-IDF $0, 0.7, 0, 2, 0, 0$
- Latent semantic indexing (LSA)
- Sidenote: Recently more focus on word embeddings (e.g. Word2Vec, GloVe, FastText)



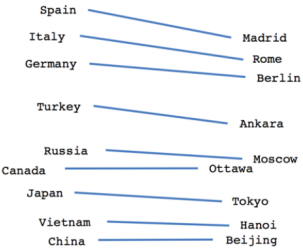
- Text data
- Visual Data
- Everything2Vec: NLP & Computer Vision



Male-Female



Verb tense



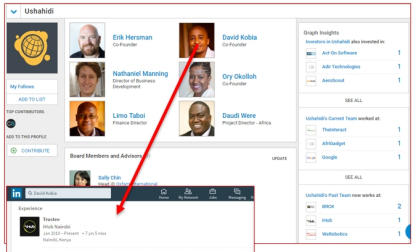
Country-Capital

Opportunities

- Better Estimates
- Understanding human behavior
- Rare event prediction
- MApping and understanding complex & dynamic systems
- Analysis of traditionally data-sparse environments

Challenges

- Handling Big Data
- Prediction vs. Deriving desired statistical properties
- Approaching new tools
- Institutional Inertia



Hain, D. S., Christensen, J.L. & Jurowetzki, R.: The Value of Human Capital Signals for Investment Decision Making under Uncertainty, **2017**, tbs @ DRUID NYC

Table 3: GLM regression results (link=Logit): Investments (Foreign only, Post-PSM)

	Dependent variable: Investor participates in round					
	EU	KINGS	EU	KINGS	EU	KINGS
	(1)	(2)	(3)	(4)	(5)	(6)
	Deal characteristics (controls)					
round.seed	−0.349	−0.660*	−0.292	−0.687*	−0.375	−0.650
	(0.658)	(0.394)	(0.666)	(0.396)	(0.666)	(0.399)
round.venture	0.021	−0.563	0.021	−0.637	0.080	−0.561
	(0.722)	(0.437)	(0.727)	(0.441)	(0.725)	(0.444)
inv.sum	−0.042	0.039**	−0.067	0.041**	−0.069	0.043**
	(0.072)	(0.016)	(0.076)	(0.017)	(0.076)	(0.017)
log(round.usd)	−0.126	0.0002	−0.063	0.008	−0.117	−0.001
	VC characteristic					
inv.exp	−0.010***	−0.0005	−0.009***	−0.006	−0.011***	−0.0005
	(0.003)	(0.001)	(0.003)	(0.004)	(0.003)	(0.001)
inv.exp.dc	0.171***	0.254***	0.182***	0.274***	0.181***	0.210***
	(0.021)	(0.076)	(0.023)	(0.078)	(0.024)	(0.080)
	PC human capital signals					
	(0.104)	(0.059)	(0.106)	(0.060)	(0.105)	(0.060)
comp.press	0.001	−0.0004	0.004*	−0.001	0.003	0.001
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)
per.entrp	−0.141	0.275	−0.061	0.444**	−0.295	0.392**
	(0.208)	(0.195)	(0.247)	(0.215)	(0.303)	(0.221)
per.edu	0.109	−0.096	0.097	−0.114	0.187	0.110**
	(0.151)	(0.095)	(0.196)	(0.092)	(0.211)	(0.064)
per.mng	−0.022	−0.001	0.034	−0.071	0.012	−0.025
	(0.090)	(0.076)	(0.132)	(0.083)	(0.134)	(0.082)
	*VC * PC interaction terms*					
per.entrp:inv.exp			−0.004	0.015*		
			(0.007)	(0.008)		
comp.press:inv.exp			−0.0001*	0.0001*		
			(0.0001)	(0.0001)		
per.edu:inv.exp			−0.001	−0.0004		
			(0.006)	(0.003)		
per.mng:inv.exp			−0.001	0.000**		
			(0.004)	(0.003)		
per.entrp:inv.exp.dc					0.023	0.108**
					(0.028)	(0.047)
comp.press:inv.exp.dc					−0.0003	−0.009
					(0.0002)	(0.007)
per.edu:inv.exp.dc					0.012	−0.149**
					(0.021)	(0.157)
per.mng:inv.exp.dc					−0.002	0.173*
					(0.012)	(0.081)
comp.country	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	1,002	894	1,002	894	1,002	894
Log Likelihood	-229.100	-376.900	-224.700	-342.600	-227.900	-314.100

Note: *p<0.1; **p<0.05; ***p<0.01 , standard errors in parentheses

AALBORG UNIVERSITY DENMARK

The Emerging Entrepreneurial Ecosystem in SSA, cont'd

- Bottom up emerging ecosystem
- Vivid mix of foreign / local support institutions, investors ect.
- "Link" function of VCs
- But that's another story…

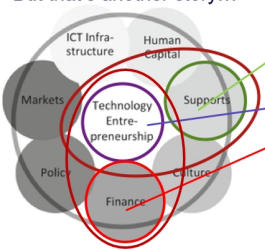
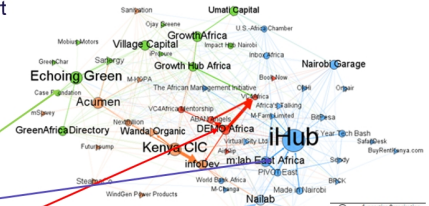


Figure 3 Direct interactions (mentions) between start-ups and supporting organizations. Colors denote statistical communities detected by the Louvain algorithm.



Legend

Metrics on the whole network

Density	13.90%	Average path	2.38

Community detection and metrics

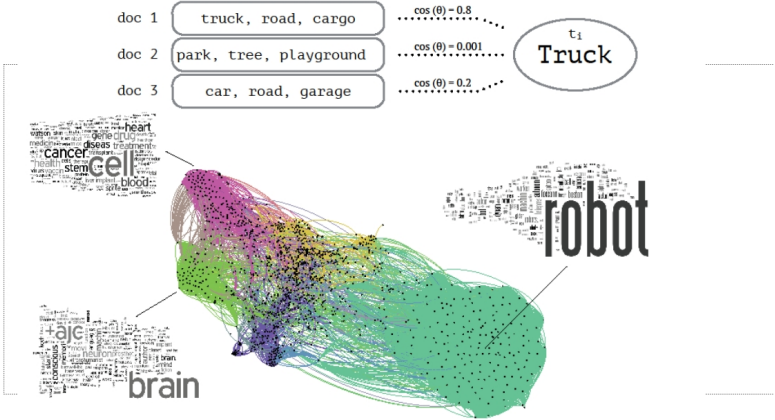
Community	Activity	Sector	Density 1	Density 2*	Key Players
●	Investment	App based	40%		Demo Africa
●	Technology; Networking; Incubator/Accelerator/Co-working space	App/Computer Software and hardware; Internet; Services	21.90%	20.50%	iHub/Nairobi Garage/m:lab East Africa; Sendy/DHH/Virtual City
●	Promoting Technology and entrepreneurship	Internet	100%		Nailab
●	Development Goals; Social entrepreneurship	Agribusiness; Clean	27.30%		Acument; Wanda Organic
●	Social Innovation; Impact investment	Renewable energy; Agribusiness; Clean Technology; Low-cost mobility	23.10%	15.50%	Eccoing Green; Sanergy

A

*Communities are grouped into two main clusters which will serve as reference to compare the results of the different networks.

Park, E.; Martins, R.M; **Hain, D. S.** & Jurowetzki, R.2017: Entrepreneurial Ecosystem for Technology Start-ups in Nairobi: Empirical analysis of Twitter networks of Start-ups and Support organizations, 2017, IKE Working Paper Series

1st Results: Static technology network (semantic similarity)



New Methods
Some examples

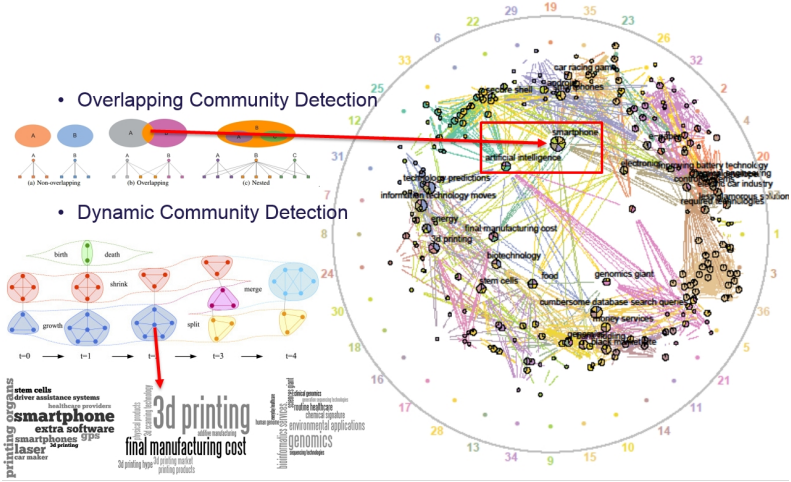
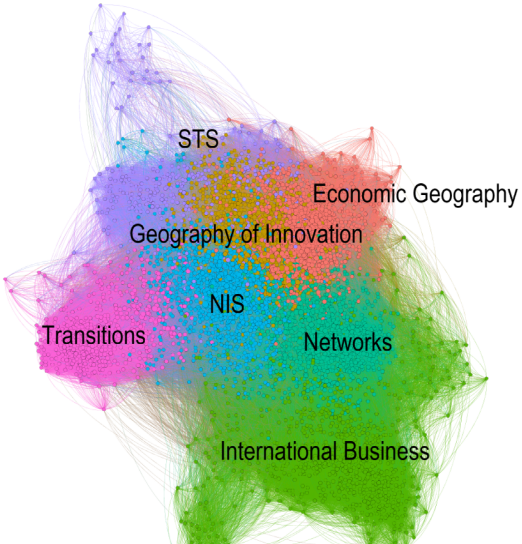


Figure 4: Bibliographic coupling network of the IS literature



New Ways of Teaching Data Science

- exponential growth of data in all domains
- availability of data
- exponential growth of computer performance
- industry makes enormous efforts to develop analysis techniques and capitalise on data
- OpenSourcing of easy-to learn, professional tools:
  - R & Python
  - NLTK, Spacy, Gensim and other language modelling libraries
  - recent developments in deep learning: Google's Tensorflow, Theano, Keras (and R alternatives), Open AI Gym
  - pre-trained deep learning models: Inception, Word2Vec
  - picking up in research

Old paradigm in instruction

- (naïve) Assumption of an **existing, clean dataset** in a world full of unstructured data
- Data preparation (i.e. collecting, mining, cleaning & organising) makes up 80% of the time use of professional data scientist (Forbes, 2016).
- No or very limited training in relational data (networks), while these concepts receive enormous attention otherwise
- Lock-in in (soon to be) **obsolete proprietary software** (e.g. SAS, Stata, SPSS), and Office/Administration software that is not suited for serious data analysis (e.g. Excel)

- Extreme demand for analytical skills in Finance, Insurance, Consulting, and other sectors
- growing awareness in the public sector
- growing demand for "data-driven decision makers"
- → not only tech-skills but also domain insight

ARTWORK: TAMAR COHEN, ANDREW J. BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 11 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see possibilities. He began forming theories, testing hunches, and

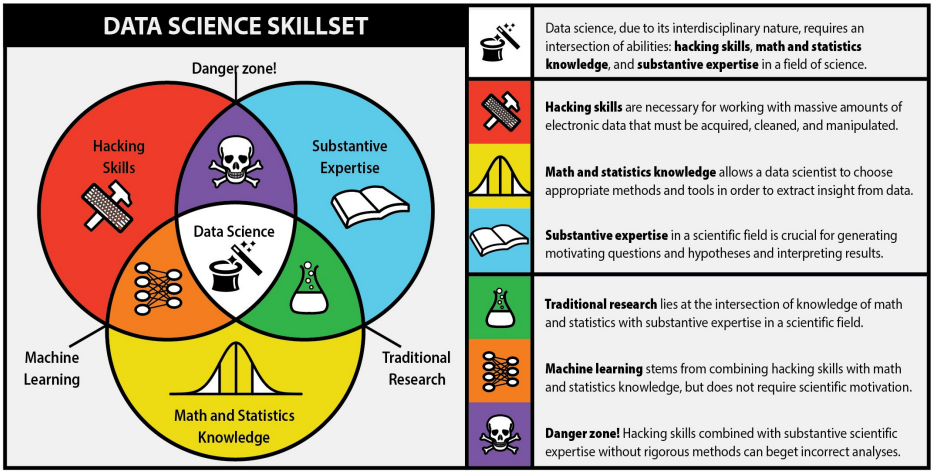
WHAT TO READ NEXT

Big Data: The Management Revolution

VIEW MORE FROM THE

October 2012 Issue

The Data Science Skillset



New Content

- Dynamics of technological & economic opportunities.

New Data

- Rise of unstructured data.

New Methods

- Extremely rapid progress in predictive methods, alignment of opportunities in research, policy, business.

New Ways of Teaching

- Critical point in time, where new approaches get accessible for non computer-science PhDs
- Drawing largely from open source and open access.
- Enormous potential for self study.

Thank you for your attention. Any questions?

AALBORG UNIVERSITY
DENMARK