Statistical learning from data as the ultimate agile development tool

Peter Norvig
Google
Agile Development Principles

- Customer satisfaction by rapid delivery
- Working software is delivered frequently
- Working software is the measure of progress
- Even late changes in requirements are welcomed
- Daily cooperation business people / developers
- Face-to-face conversation
- Projects built around motivated individuals
- Continuous attention to excellence of design
  - Simplicity
  - Self-organizing teams
  - Regular adaptation to changing circumstances

see also: “Good Agile / Bad Agile,” Steve Yegge
Keep it Simple: More Data vs. Better Algorithms

Figure 2. Learning Curves for Confusable Disambiguation

Banko & Brill, 2001
Keep it Simple: More Data vs. Better Algorithms

Figure 2. Learning Curves for Confusable Disambiguation
Keep it Simple: More Data vs. Better Algorithms

Figure 2. Learning Curves for Confusable Disambiguation

Banko & Brill, 2001

Friday, October 3, 2008
Keep it Simple: More Data vs. Better Algorithms

Figure 2. Learning Curves for Confusible Disambiguation
Rational Programming
(or what to do when you don’t know what to do)

Maximize Expected Utility:

\[ action = \arg\max_{a \in \text{actions}} \text{EU}(a) \]

\[ \text{EU}(a) = \sum_{s \in \text{Results}(a)} P(s) \times U(s) \]

Learn/Approximate \text{Results}(a), P(s), U(s)
from big data sources
Segmentation

靳羽西中国新锐画家大奖
Segmentation

nowisthetimeforalldementocometothe
Segmentation
Segmentation

Probability of a segmentation = $P(\text{first word}) \times P(\text{rest})$
Probability of a segmentation = $P(\text{first word}) \times P(\text{rest})$

Best segmentation = one with highest probability
Segmentation

Probability of a segmentation = $P(\text{first word}) \times P(\text{rest})$

Best segmentation = one with highest probability

$P(\text{word}) =$ estimated by counting
Segmentation

\[
\text{segment(“nowisthetime…”)} \\
P_f(“n”) \times P_r(“owisthetime…”) \\
P_f(“no”) \times P_r(“wisthetime…”) \\
P_f(“now”) \times P_r(“isthetime…”) \\
P_f(“nowi”) \times P_r(“sthetime…”) \\
\]

...
Segmentation

segment(“nowisthetime…”)

<table>
<thead>
<tr>
<th>f</th>
<th>P(f)</th>
<th>P(r)</th>
<th>= P(f)P(r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“n”</td>
<td>.003%</td>
<td>$10^{-30%}$</td>
<td>$10^{-34%}$</td>
</tr>
<tr>
<td>“no”</td>
<td>.26%</td>
<td>$10^{-26%}$</td>
<td>$10^{-29%}$</td>
</tr>
<tr>
<td>“now”</td>
<td>.23%</td>
<td>$10^{-21%}$</td>
<td>$10^{-24%}$</td>
</tr>
<tr>
<td>“nowi”</td>
<td>$10^{-7%}$</td>
<td>$10^{-21%}$</td>
<td>$10^{-30%}$</td>
</tr>
</tbody>
</table>

...
Segmentation

from utils import Pw, product, memo

def splits(characters, longest=12):
    "All ways to split chars into a first word and remainder."
    return [(characters[:i], characters[i:])
            for i in range(1, 1+min(longest, len(characters)))]

def Pwords(words): return product(words, key=Pw)

@memo
def segment(text):
    "Best segmentation of text into words, by probability."
    return [] if (text==")" else (
        max([[first]+segment(rest) for first,rest in splits(text)],
             key=Pwords))
Trained on 1.7B words English
Trained on 1.7B words English

• 98% word accuracy. Typical errors:
Trained on 1.7B words English

- 98% word accuracy. Typical errors:
- baseratesoughtto
Trained on 1.7B words English

- 98% word accuracy. Typical errors:
- baseratesoughtto
  base rate__sought to
Trained on 1.7B words English

- 98% word accuracy. Typical errors:
  - base rates ought to
  - base rate sought to
  - small and insignificant
Trained on 1.7B words English

- 98% word accuracy. Typical errors:
  - baseratesoughtto
    - base rate sought to
  - smallandinsignificant
    - small and insignificant
Trained on 1.7B words English

- 98% word accuracy. Typical errors:
  - baseratesoughtto
    - base rate e sought to
  - small and insignificant
    - small and in significant
  - ginormousego
Trained on 1.7B words English

- 98% word accuracy. Typical errors:
  - baseratesoughtto: base rate sought to
  - smallandinsignificant: small and insignificant
  - ginormousego: gin or mouse go
Mistakes were made
Mistakes were made

whorepresents.com ⇒[“who”, “represents”]
Mistakes were made

whorepresents.com ⇒[“who”, “represents”]
therapistfinder.com ⇒[“therapist”, “finder”]
Mistakes were made

whorepresents.com ⇒[“who”, “represents”]
therapistfinder.com ⇒[“therapist”, “finder”]
expertsexchange.com ⇒[“experts”, “exchange”]
Mistakes were made

whorepresents.com ⇒[“who”, “represents”]
therapistfinder.com ⇒[“therapist”, “finder”]
expertsexchange.com ⇒[“experts”, “exchange”]
penisland.com ⇒ error: expected [“pen”, “island”]
Mistakes were made

whorepresents.com ⇒[“who”, “represents”]
therapistfinder.com ⇒[“therapist”, “finder”]
expertsexchange.com ⇒[“experts”, “exchange”]
penisland.com ⇒ error: expected [“pen”, “island”]
speedofart.net ⇒[“speed”, “of”, “art”]
Spelling

Mehran Sahami
Spelling

Typical word processor:

Mehran Sahami
Spelling

Typical word processor:

Tehran Salami
```cpp
for (; *n && key.length() < MAXPHONEMELEN; n++)
{
    /* Drop duplicates except for CC */
    if (*(n - 1) == *n && *n != 'C')
        continue;
    /* Check for F J L M N R or first letter vowel */
    if (same(*n) || *(n - 1) == '0' && vowel(*n))
        key << *n;
    else
    {
        switch (*n)
        {
            case 'B':
                /*
                 * B unless in -MB
                 */
                if (*(n + 1) || *(n - 1) != 'M')
                    key << *n;
                break;
            case 'C':
                /*
                 * X if in -CIA-, -CH- else S if in
                 * -CI-, -CE-, -CY- else dropped if
                 * in -SCI-, -SCE-, -SCY- else K
                 */
                if (*(n - 1) != 'S' || !frontv(*(n + 1)))
                {
                    if (*(n + 1) == 'I' && *(n + 2) == 'A')
                        key << 'X';
                    else if (frontv(*(n + 1)))
                        key << 'S';
                    else if (*(n + 1) == 'H')
                        key << ((((n - 1) == '0' && !vowel
                                    || *(n - 1) == 'S')
                                    ? 'K' : 'X');
                    else
                        key << 'K';
                }
        }
    }
```
Spelling with Statistical Learning
Spelling with Statistical Learning

Probability of a spelling correction, \( c = \)
\[ P(\text{c as a word}) \times P(\text{original is a typo for c}) \]
Spelling with Statistical Learning

Probability of a spelling correction, $c = P(c \text{ as a word}) \times P(\text{original is a typo for } c)$

Best correction = one with highest probability
Spelling with Statistical Learning

Probability of a spelling correction, \( c = \)
\[ P(c \text{ as a word}) \times \]
\[ P(\text{original is a typo for } c) \]

Best correction =
one with highest probability

\[ P(c \text{ as a word}) = \]
estimated by counting
Probability of a spelling correction, \( c = \)
\[ P(c \text{ as a word}) \times P(\text{original is a typo for } c) \]
Best correction =
one with highest probability
\[ P(c \text{ as a word}) = \]
estimated by counting
\[ P(\text{original is a typo for } c) = \]
proportional to number of changes
from utils import Pw, alphabet

def edits1(word):
    n = len(word)
    ## deletion, transposition, alteration, insertion
    return set([word[0:i]+word[i+1:] for i in range(n)] +
                [word[0:i]+word[i+1]+word[i]+word[i+2:] for i in range(n-1)] +
                [word[0:i]+c+word[i+1:] for i in range(n) for c in alphabet] +
                [word[0:i]+c+word[i:] for i in range(n+1) for c in alphabet])

def known_edits2(word):
    return set(e2 for e1 in edits1(word) for e2 in edits1(e1) if e2 in Pw)

def known(words):
    return set(w for w in words if w in Pw)

def correct(word):
    candidates = (known([word]) or known(edits1(word)) or
                  known_edits2(word) or [word]
    return max(candidates, key=Pw)
Google Sets

Automatically create sets of items from a few examples.

Enter a few items from a set of things. (example)
Next, press Large Set or Small Set and we'll try to predict other items in the set.

- pablo picasso
- henri matisse

(clear all)

Large Set  Small Set (15 items or fewer)
Google Sets

Given “henri matisse,” “pablo picasso” find:

henri matisse, pablo picasso, vincent van gogh, claude monet, pablo picasso, salvador dali, edgar degas, paul cezanne, andy warhol, pierre auguste renoir, marc chagall, paul gauguin, mary cassatt, michelangelo, jackson pollock, camille pissarro, paul klee, georgia o'keeffe, norman rockwell, leonardo da vinci, edward hopper, frida kahlo, rembrandt, wassily kandinsky, rene magritte, francisco goya, raphael, georges seurat, giotto, henri rousseau, diego rivera, van gogh, botticelli, johannes vermeer, ansel adams, monet, titian
Google Sets

Given “lion, tiger, bear” find:

- bear, tiger, lion, elephant, monkey, giraffe, dog, cat, snake, horse, zebra, rabbit, wolf, dolphin, dragon, pig, frog, duck, cheetah, bird, cow, cotton, hippo, turtle, penguin, rat, gorilla, leopard, sheep, mouse, puppy, ox, rooster, fish, lamb, panda, wood, musical, toddler, fox, goat, deer, squirrel, koala, crocodile, hamster
Google Sets
Google Sets

Given “cat, man” find:
Google Sets

Given “cat, man” find:
cat, man, ls, cp, rm, mkdir, mv, cd,
pwd, rmdir, chmod, ln, grep, touch,
find, ps, chown, df, less, tar, chgrp, du,
sort, date, echo, kill, tail, wc, mount,
sed, passwd, dd, head, vi, which, gzip,
lp, who, file, su, umount, diff, cut,
uname, exit, basename, clear, sleep,
whoami, mknod
Google Sets
**Google Sets**

**Tin Woodsman**: Some, but mostly lions and tigers and bears.

**Dorothy**: Lions?

**Scarecrow**: And tigers?

**Tin Woodsman**: And bears.
Google Sets

Tin Woodsman: Some, but mostly lions and tigers and bears.
Dorothy: Lions?
Scarecrow: And tigers?
Tin Woodsman: And bears.

African cats:
• **Lions**
• **Leopards**
• **Cheetahs**
Tin Woodsman: Some, but mostly lions and tigers and bears.
Dorothy: Lions?
Scarecrow: And tigers?
Tin Woodsman: And bears.

African cats:
- Lions
- Leopards
- Cheetahs

Animal Toys (Ages 2-6):
- Bears
- Puppies
- Dolphins
- Wooden
Google Sets

Tin Woodsman: Some, but mostly lions and tigers and bears. Dorothy: Lions? Scarecrow: And tigers?
Tin Woodsman: And bears.

African cats:
• Lions
• Leopards
• Cheetahs

Animal Toys (Ages 2-6):
• Bears
• Puppies
• Dolphins
• Wooden

10:03:37 [cheetah pics]
10:05:51 [leopard pics]
Google Sets

Tin Woodsman: Some, but mostly lions and tigers and bears.
Dorothy: Lions?
Scarecrow: And tigers?
Tin Woodsman: And bears.

African cats:
- Lions
- Leopards
- Cheetahs

U.S. Senate Committee on Environment and Public Works
Exotic animals, such as lions, tigers, servals, monkeys, bears, snakes, iguanas, wolves, prairie dogs, and binturongs are being privately possessed as ...
epw.senate.gov/hearing_statements.cfm?id=213174 - 13k -
• Collect parallel texts

SEHR GEEHRTER GAST!
KUNST, KULTUR UND
KOMFORT IM HERZEN
BERLIN.

DEAR GUESTS,
ART, CULTURE AND
LUXURY IN THE HEART
OF BERLIN.

DIE ÖRTLICHE
NETZSPANNUNG
BETRÄGT 220/240 VOLT
BEI 50 HERTZ.

THE LOCAL VOLTAGE
IS 220/240 VOLTS 50 HZ.
• Align

KUNST, KULTUR UND KOMFORT IM HERZEN BERLINS.

ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.
Statistical Machine Translation

• Align

KUNST, KULTUR UND KOMFORT IM HERZEN BERLINS.

ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.
• Align

KUNST, KULTUR UND KOMFORT IM HERZEN BERLINS.

ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.
Statistical Machine Translation

- Align

KUNST, KULTUR UND KOMFORT IM HERZEN BERLINS.

ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.
• Align

KUNST, KULTUR UND KOMFORT IM HERZEN BERLINS.

ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.
Statistical Machine Translation

- The northern Korean threat has declared that the Security Council is the only legitimate body that can impose sanctions. 

- The army attacked in 1974, which was divided into two sections, and was accompanied by a coup in the third section of the northern island.

- Eight years of war in Iraq, Iran, and the western part of the island were mentioned, which was attacked by theERE military forces on the basis of the Iraq-Iran war (1988-1980), and the explosion of the incident of the destruction of the city of Arbil and the adoption of the arms treaty.
It is noteworthy that the Security Council is the only authorized to impose sanctions against North Korea, which warned that it would consider sanctions a declaration of war.

Cyprus has been divided into two parts since the year 1974 when the Turkish army invaded the northern third of the island in response to a coup by Greek nationalists with the aim of annexing the island to Greece.

It is worth mentioning that Iran, which fought a war against Iraq lasted eight years (1980 – 1988) opposes American military attack on Iraq, which Washington accuses of possessing weapons of mass destruction and that it was linked to terrorism.
Statistical Machine Translation

• 新华社大马士革4月15日电（记者拱振喜）叙利亚总统巴沙尔·阿萨德15日在此间与来访的美国国务卿鲍威尔举行了会谈，双方讨论了中东局势的最新发展，特别是巴勒斯坦的严重局势以及黎以边界地区的紧张局势等问题。

• 叙利亚通讯社报道，巴沙尔总统在会谈中说：“在巴勒斯坦发生的事件使（中东）和平进程走进了死胡同，如果不能认识到这一点，事情的发展有可能达到无法挽回的程度，那时，我们只能再等待一代人的时间。”

• 他指出，只有在以色列从它占领的巴勒斯坦领土撤军，停止屠杀巴勒斯坦人以后，才可以谈和平进程的问题。
Xinhua News Agency, Damascus, April 15 (Reporter Gong Zhenxi) Syrian President Bashar Assad 15th here with visiting US Secretary of State Colin Powell held talks, the two sides discussed the latest development of the situation in the Middle East, especially the serious situation in Palestine and the tension in the border region between Lebanon and Israel and other issues.

According to the Syrian News Agency, President Bashar during the talks, said: "In the incident to the Palestinian (Middle East) peace process into a dead end, if not realize that this is happening may not be able to restore to the extent that time, we can only wait for the generation of time."

He pointed out that only in the Israeli withdrawal from the occupied Palestinian territories, stop massacre of Palestinians, can talk about the peace process.
他 信 也 说 自 己 仍 然 是 总 理 ， 拒 绝 辞 职。
他 信 也 说 自 己 仍 然 是 总 理，拒 绝 辞 职。
他 信 也 说 自 己 仍 然 是 总 理 ， 拒 绝 辞 职。
他信也说自己仍然是总理，拒绝辞职。

Thaksin Thaksin Chinnawat and Joint Communique
Dr Thaksin
Joint Communique
Mr Thaksin
in his letter
his letter
others

he has
of his
Thaksin also said in his letter that he has confidence in his letter of trust, believe that he is of his others. Dr Thaksin Chinnawat and Joint Communique, Mr Thaksin in his letter of confidence, believe that he was his letter to him. He has said himself is still the Prime Minister, and he has refused to resign.
他信也说自己仍然是总理，拒绝辞职。

Thaksin Chinnawat and Joint Communique
Dr Thaksin also.

in his letter

Thaksin also

saying that he would say that

said that she had saying that he has
Continues to be the main
themselves would still be
said that he says they
continued to be
say that remains one of
said he remains one
say that continues to be the
said they still is
themself remains an area
saying still viewed by
would say are always one of
said that she had
saying that he has

Friday, October 3, 2008
Thaksin also said that he and Joint Communique Dr Thaksin and Joint Communique, Mr Thaksin in his letter his letter others his has of his

Prime Minister the Prime Minister is the Prime Minister

continue to be the continues to be the

remains one of remains one

say that continues to be the

remains an area remains an area

saying that still viewed by

would are always one of

said that she had saying that he has
Prime Minister Thaksin Chinnawat and Joint Communique Dr Thaksin in his letter his letter others he has of his

He said that he would say that he has saying that he has

He also said that he remains one of the Prime Minister is the Prime Minister

He said that he remains one of the Prime Minister is the Prime Minister

He said that he continues to be the Prime Minister is the Prime Minister

He is the Prime Minister is the Prime Minister

He said that he continues to be the Prime Minister is the Prime Minister

He said that he continues to be the Prime Minister is the Prime Minister

He said that he continues to be the Prime Minister is the Prime Minister

He said that he remains one of the Prime Minister is the Prime Minister

He said that he remains one of the Prime Minister is the Prime Minister

He said that he remains one of the Prime Minister is the Prime Minister

He said that he remains one of the Prime Minister is the Prime Minister

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has

He said that she had said that she had saying that he has
Thaksin also said in his letter that he has continued to be the Prime Minister, the Prime Minister remains one of the main areas of leaving their service, resigned as counsel. He has said that she had said that she had continued to be the Prime Minister, the Prime Minister remains one of the main areas of leaving their service, resigned as counsel. He has said that the remain one of the main areas of leaving their service, resigned as counsel. He has said that the remain one of the main areas of leaving their service, resigned as counsel. He has said that the remain one of the main areas of leaving their service, resigned as counsel. He has said that the remain one of the main areas of leaving their service, resigned as counsel.
More Data Still Helps

AE BLEU [%]

Friday, October 3, 2008
Image Data
Compare low-level features

(a) A vs. B
(b) A vs. C
(c) A vs. D
(d) B vs. C
(e) B vs. D
(f) C vs. D
SIFT Features
Jay Yagnik, Atiq Islam, Google: Learning People Annotation
George Bush, Angela Merkel and a barrel of Bismarck herrings. Photograph: Heribert Proepper/AP

Here's a strange set of things that come together more often than you would think: George Bush, Germany and fish. A few
Jay Yagnik, Atiq Islam, Google: Learning People Annotation

George Bush

Alan Alda
In Conclusion
“Measuring programming progress by lines of code is like measuring aircraft building progress by weight.”

- Bill Gates
Data is the Ultimate Asset

- Test–Driven Development
- Easy to generate a new version
- Compositional
- Results oriented
- Simple
- Easy to update for new circumstances
- Works with or without understanding
- Faster and better (in many cases)
Questions?